UNIVERSITÉ DE GENÈVE
Section de mathématiques

FACULTÉ DES SCIENCES
Professeur E. Hairer

Analysis and Numerical Treatment of Highly Oscillatory Differential Equations

THÈSE

présentée à la Faculté des sciences
de l'Université de Genève
pour obtenir le grade de Docteur ès sciences,
mention mathématiques

par

David COHEN
de
Gontenschwil (AG)

Thèse N^o 3524

Genève
Atelier de reproduction de la Section de physique
2004

En équilibre instable entre le royaume des abstractions
et celui de la matière,
il essayait désespérément de se raccrocher à
quelque chose.

(David Brin 1995, *Rédemption*)

Remerciements

Je tiens à remercier chaleureusement les personnes qui ont contribué à l'élaboration de ce travail ; en particulier

• Ma famille pour son soutien et la patience qu'elle a eue. Merci "petit frère", c'est toi l'exemple.

• Ernst Hairer, directeur de thèse, pour sa disponibilité et surtout sa patience, ce n'était pas tous les jours faciles avec moi... Ce fut un honneur de travailler avec un BOSS pareil !

• Gerhard "Papa Noël" Wanner, les membres du GANG et les invités du séminaire d'analyse numérique.

• Arieh Iserles et Christian Lubich, membres du jury, pour l'intérêt qu'ils ont porté à ce travail.

• En vrac, pour tout ce qu'ils ont fait: Eugenio Rodriguez, Jeremy Blanc et sa dame de coeur, Luc "l'homme qui parlait plus vite que son ombre" Guyot et la French Connection, le "petit" Nicolas Bartholdi, Harald Wendler, Sonja Hairer, Paola Argentin, Christian Wuthrich, Emmanuel Zabey, les étudiants de la section, NTM, Google et les trois petits points.

• Les enseignants de la section de mathématiques pour la qualité de leurs cours.

• Les bibliothécaires et le personnel administratif de la section pour leur disponibilité, leur compétence et leur gentillesse.

• Je l'oublie toujours : ..... (← écris ton nom).

When worst come to worst
my peoples come first.

(Dilated Peoples, *Worst Comes To Worst*)

Contents

Introduction

"By carefully tracing the dependence on N of the constants in the $\mathcal{O}(\omega^{-N})$-terms, near-conservation of I over exponentially long time intervals can be shown also within the present framework of modulated Fourier expansions" [HLW02, p.443], it is with this sentence that all began ...

But what exactly ?

The study of highly oscillatory second-order differential equations and numerical methods to solve them.

What stands these words for ?

Basically, an ordinary differential equation (ODE) is an equation involving an unknown function and its derivatives, we say that it has order 2 when the second derivative (at most) of this function appears in this relation. They are often met in Physics, but those that are our subject appear in Molecular Dynamics (MD). The definition of the term "highly oscillatory" is difficult to give: "[...] it does not seem possible to give a precise mathematical definition which would include most of the problems that scientists, engineers and numerical analysts have described as highly oscillatory" [PJY97, p.438]. However, we hope that this term would be clear after reading this work. The ODE studied in this thesis is a mathematical model for MD problems, as a simple example of an application, we can describe the motion of a linear diatomic molecule (see Figure 1) with this model. The bond length between the two atoms can be modelled with the help of a stiff spring (with a large stiffness constant). This makes the molecule vibrate very rapidly, with high frequency.

Figure 1: A linear diatomic molecule.

We can also describe the planar motion of such a molecule, but it's a little bit more complicated (see Chapter 5) ...

What can we expect for those systems ?

For some physical or MD systems described by the so called Hamiltonian equations, we know that the total energy (we sometimes call this quantity the Hamiltonian) is conserved. But, systems that we are looking at have another near-preserved quantity: the oscillatory energy, denoted by I in the first sentence. For our simple diatomic problem, this adiabatic invariant correspond to the energy of the stiff spring. The tool we use to explain the behaviour of the solution of the second-order oscillatory differential equation studied here is called the *modulated Fourier expansion.* This expansion was developed in the paper [HL00], it expresses the solution of the ODE as an expansion involving smooth functions.

And how do you solve this kind of problems ?

Analytically it's impossible. This is why we develop numerical methods. They are designed to solve ODEs on a computer. Basically, starting with an initial value, a numerical method gives a sequence of points, sometimes after horrible calculations ...each point approximating the solution at a certain time. We finally just have to plot the desired results with the help of this sequence.

Moreover, numerical methods allow us to carry out simulations under extreme temperature, pressure or other quantities such as they would not be possible in a laboratory. For example, on a computer you can recreate the weather in Mars, this is (nowadays) not possible in a laboratory.

Because the solution of the oscillatory differential equations studied have special properties (like the oscillations, the preservation of some quantities, ...), we would like our numerical methods to preserve these properties too. Beside these requirements, to avoid lots of computations, we want that the product of the step size of the numerical method with the highest frequency of the system is not small.

Hum ... It seems interesting. Where can I read more about this subject?

I suggest you, beside all the papers present in the Bibliography, the document that you hold in your hand. This thesis is organized as follows: in the first chapter, we will explain in more details the kind of problems studied, recall the methods used to solve them and review theoretical results obtained so far. Chapter 2 is devoted to a rigorous proof of the near-conservation of the oscillatory energy on exponentially long time intervals. To do this, we use the so-called *modulated Fourier expansion.* In the third chapter, we will develop numerical methods to solve these kinds of problems. Chapter 4 will be a generalization of Chapter 2 to the multi-frequency case. Finally, the last chapter will analyse a new class of oscillatory differential equations (who includes those of Chapter 2) and a new kind of numerical methods to solve these differential equations.

Chapter 1

The problem and some results

In this chapter we describe the problem considered in this work. As an example, we analyze the modified Fermi-Pasta-Ulam (FPU) model treated in [HLW02, Chap. XIII]. We recall some theoretical properties of the differential equation considered and its numerical treatement.

1.1 Description of the problem

We consider Hamiltonian problems

$$\begin{aligned} \dot{p} &= -\nabla_q H(p,q) \\ \dot{q} &= \nabla_p H(p,q), \end{aligned} \tag{1.1}$$

where the variables p and q are in $\mathbb{R}^d$ and the scalar function $H(p,q)$ is sufficiently differentiable. The formulation (1.1) is a way to describe the dynamics of general mechanical systems with d degrees of freedom, the variables q and p are called the *generalized coordinates* and the *conjugate momenta*. The Hamiltonian H represents the total energy of the system and is conserved along the solution of (1.1). This kind of problems often appears in Molecular Dynamics (MD) or in physics (see below). In these cases, the Hamiltonian consists of the sum of the *kinetic energy* $T(p,q)$ and the *potential energy* $V(q)$.

More precisely, we are interested in highly oscillatory solutions of Hamiltonian problems. In particular, we consider the special form of H where the kinetic and potential energies are given by

$$T(p,q) = \frac{1}{2} p_1^T M(q)^{-1} p_1 + \frac{1}{2}(p_2^T p_2 + \ldots + p_n^T p_n), \quad V(q) = \frac{1}{2}(\omega_2^2 q_2^T q_2 + \ldots + \omega_n^2 q_n^T q_n) + U(q), \tag{1.2}$$

where $M(q)$ is a mass matrix, $\Omega = diag(0, \omega_2 I, \ldots, \omega_n I)$ is a square matrix with blocks of arbitrary dimension, and $\omega_i \gg 1$ for $i = 2, \ldots, n$. We partition the variables $p = (p_1, \ldots, p_n)$ and $q = (q_1, \ldots, q_n)$ according to the partition of the matrix Ω. The vector p_i (for $i = 1, \ldots, n$) is of the same dimension as its corresponding block in the matrix

Ω. We suppose that the nonlinearity gradient $\nabla U(q)$ is analytic with derivatives bounded independently of $\omega_2, \ldots, \omega_n$ and the initial values of (1.1) are assumed to satisfy

$$\frac{1}{2}\Big(\|p(0)\|^2 + \|\Omega q(0)\|^2\Big) \leq E, \tag{1.3}$$

where E is independent of ω_i, for $i = 2, \ldots, n$.

In Chapter 5, we will take one frequency ω in (1.2). An example for such a Hamiltonian system is the motion of the planar diatomic molecule seen in the introduction.

Sometimes the mass matrix $M(q)$ is the identity and we write $x, \dot{x}$ instead of q, p and $x_j, \dot{x}_j$ instead of q_j, p_j. For this kind of problems, the Hamiltonian reads (see Chapter 4 and the end of first example below)

$$H(x, \dot{x}) = \frac{1}{2}\Big(\|\dot{x}\|^2 + \|\Omega x\|^2\Big) + U(x). \tag{1.4}$$

In the first example and in Chapter 2, we will consider the particular case of a one frequency highly oscillatory differential equation with Hamiltonian of the form

$$H(x, \dot{x}) = \frac{1}{2}\Big(\|\dot{x}\|^2 + \|\hat{\Omega} x\|^2\Big) + U(x). \tag{1.5}$$

We sometimes consider the more general second-order differential equation

$$\ddot{x} + \hat{\Omega}^2 x = g(x) \qquad \text{with} \qquad \hat{\Omega} = \begin{pmatrix} 0 & 0 \\ 0 & \omega I \end{pmatrix}, \tag{1.6}$$

with arbitrary nonlinear smooth function $g(x)$ with derivatives bounded independently of ω. We remark that taking for the function g the negative gradient of a potential U, this problem is Hamiltonian with (1.5).

For all this kind of problems, we are interested in explaining the near-conservation of the oscillatory energy

$$I(x, \dot{x}) = I_2(x, \dot{x}) + \ldots + I_n(x, \dot{x}), \quad \text{where} \quad I_j(x, \dot{x}) = \frac{1}{2}\Big(\|\dot{x}_j\|^2 + \omega_j^2 \|x_j\|^2\Big), \tag{1.7}$$

over long time intervals. Moreover, we are also interested in using numerical methods that conserve very well this quantity and the total energy H. These numerical methods should be able to use a step size h for which the product with the highest frequency ω is not small.

Example (FPU). As a first example, we consider a nonlinear mass-spring model (as in [HL00]). This model is a variation of the classical FPU model ([FPU55],[Wei97],[AT87]) and consists of a chain of $2n$ mass points connected with alternating soft nonlinear and stiff linear springs with fixed end points (see Figure 1.1).

For this problem, the Hamiltonian reads

$$H(q, \dot{q}) = \frac{1}{2}\sum_{i=1}^{n}\big(\dot{q}_{2i-1}^2 + \dot{q}_{2i}^2\big) + \frac{\omega^2}{4}\sum_{i=1}^{n}(q_{2i} - q_{2i-1})^2 + \sum_{i=0}^{n}(q_{2i+1} - q_{2i})^4, \tag{1.8}$$

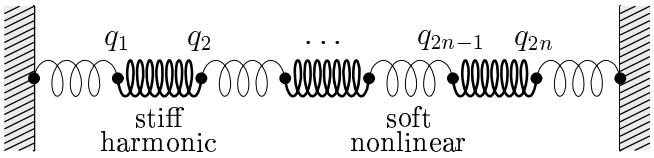


Figure 1.1: Chain with alternating soft nonlinear and stiff linear springs. (@[HLW02])

with $q_0 = q_{2n+1} = 0$, where the variables q_i stand for the displacements of the mass points. This is not exactly an Hamiltonian of the form (1.5), however, using the symplectic change of variables (which represent a scaled displacement of the stiff springs and a scaled expansion (or compression) of the ith stiff spring)

$$x_i = (q_{2i} + q_{2i-1})/\sqrt{2}, \qquad x_{n+i} = (q_{2i} - q_{2i-1})/\sqrt{2},$$

this Hamiltonian takes the desired form (1.5). To illustrate the conservation of the total energy (1.5) and the near-conservation of the oscillatory energy (1.7), we use a numerical method called DOP853 with high precision (for a definition of this method, see [HNW93]). We plot the different energies (here, $I_j = \frac{1}{2}(\dot{x}_{2,j}^2 + \omega^2 x_{2,j}^2)$) for the modified FPU problem with $n = 3$ and $\omega = 50$.

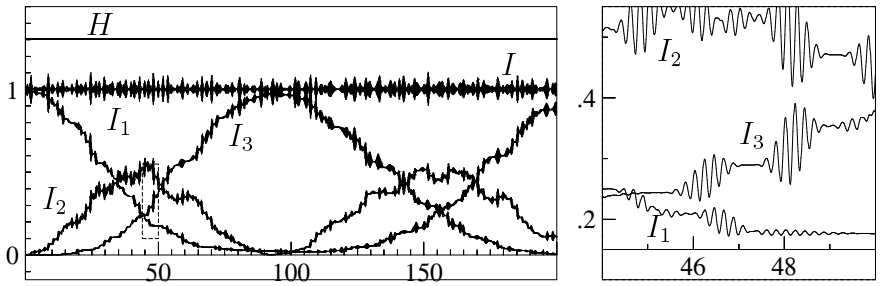


Figure 1.2: Scaled total energy and oscillatory energy using DOP853.

This shows that the oscillatory energy I (which corresponds to the total energy of the stiff springs) is preserved up to $\mathcal{O}(\omega^{-1})$. We can also see the energy exchange among the stiff spring, going from the first excited spring to the second and finally the third one.

We remark that if we consider different large elasticity constants for each stiff springs, we obtain a Hamiltonian (1.4) with more than one high frequency.

Example (MD). In classical MD (see for example [GKZC04] and [Mic01]), the motion of n atoms is described by Newton's law

$$M\ddot{x} = -\nabla U(x),$$

where the vector x contains the Cartesian coordinates of the atoms. These equations are Hamiltonian systems with Hamiltonian

$$H(x, \dot{x}) = \frac{1}{2}\dot{x}M^{-1}\dot{x} + U(x). \tag{1.9}$$

The potential $U(x)$ takes the form

$$U(x) = \sum_{k=1}^{n} U_1(x_k) + \sum_{k<l}^{n} U_2(x_k, x_l) + \sum_{k<l<m}^{n} U_3(x_k, x_l, x_m) + \dots,$$

with $x_k, x_l, \dots \in \mathbb{R}^3$ the coordinates of each atoms. For example, $U_2(x_k, x_l)$ is typically a sum of potentials $V(||x_k - x_l||)$ depending only on the distance of the atoms x_k and x_l, where $V(r)$ may take one of the following forms

- The Coulomb electrostatical potential $V(r) = \frac{q_1 q_2}{4\pi\varepsilon r}$, where q_1, q_2 and ε are given constants and stand for the electric charge of the atoms and for the electrical permittivity of space.
- The Lennard-Jones potential $V(r) = 4\varepsilon\big((\frac{\sigma}{r})^{12} - (\frac{\sigma}{r})^{6}\big)$, where ε and σ are suitable constants depending on the atoms.
- Harmonic potential (or Hook's law) $V(r) = \frac{k}{2}(r - r_0)^2$, where k is the bond constant of the spring and r_0 is the reference bond length.

The first two terms are non-bonded terms, while the last one is a bonded term. If we consider more atoms, other types of potential appear, such as torsion or angle bond...

We remark that if we consider a diatomic molecule with large bond stretches constant k and for a mass matrix $M = I$, the Hamiltonian (1.9) gives the desired second-order differential equation (1.6). Typically bond length constants and bond length for diatomic molecules can be found in the following table (see [Lea96] and the web ...)

Diatomic molecules	bond length $r_0[\mathring{A}]$	bond constant $k[kcal\ mol^{-1}]$
CC	1.337	690
CO	1.203	777
CN	1.345	719
HO	0.9572	595
HH	0.74	436

Table 1.1: Bond length and bond constants in MD.

1.2 Theoretical results

In this section, we give some references concerning theoretical results for our class of problems with Hamiltonian (1.5) and for two other classes of Hamiltonian systems that are related to our problem.

Let's first begin with the theoretical results concerning problem (1.6), using the modulated Fourier expansion, Hairer and Lubich (in [HL00]) showed the existence of two invariants of the modulation functions appearing in this expansion. They are related to

the Hamiltonian (1.5) and to the oscillatory energy (1.7). Moreover, they showed the near-conservation of the oscillatory energy over long time intervals. These results are also available in [HLW02, Chap. XIII]. Based on the same ideas, the oscillatory energy is shown, in [CHL03], to be nearly conserved over exponentially long time intervals. Finally, extension to the multi frequency case (1.4) is analysed in [CHL].

In [BGG87], Benettin et al. studied the Hamiltonian problem

$$H_\omega(p,x,\pi,\zeta) = \frac{1}{2}(\pi^2 + \omega^2\zeta^2) + \hat{h}(p,x) + f(p,x,\pi,\zeta), \tag{2.10}$$

with analytic $\hat{h}, f$, and the function f vanishes for vanishing ζ. The variables π and ζ are in $\mathbb{R}$, p and x in $\mathbb{R}^n$. Using an action-angle transformation and some canonical transformations, they also proved the near-conservation of the oscillatory energy over exponentially long time intervals. We can remark that Hamiltonian problem (2.10) contains our class of Hamiltonian problems if the second component of the Hamiltonian (1.5) is scalar.

In the second part ([BGG89]), they considered the multi frequency Hamiltonian case

$$H(p,x,\pi,\zeta) = h_\omega(\pi,\zeta) + \hat{h}(p,x) + f(p,x,\pi,\zeta), \tag{2.11}$$

here $h_\omega(\pi,\zeta)$, with $(\pi,\zeta) = (\pi_1,\ldots,\pi_\nu,\zeta_1,\ldots,\zeta_\nu) \in \mathbb{R}^{2\nu}$, is the Hamiltonian of a set of ν uncoupled harmonic oscillators of angular frequency $\omega = (\omega_1,\ldots,\omega_\nu)$. $\hat{h}(p,x)$, with $(p,x) = (p_1,\ldots,p_n,x_1,\ldots,x_n) \in \mathbb{R}^{2n}$ represents any dynamical system with n degrees of freedom. The coupling function f is assumed to vanish for $\zeta = 0$ and to be a polynomial of order 2 in π. Using perturbation theory they proved similar results than those given in [CHL], like for example the near-conservation of the total oscillatory energy.

1.3 Numerical methods

In this section, we recall the definition of a numerical method, give some examples and properties of numerical methods (we refer for example to [HLW02]). Let's consider the ordinary differential equation (ODE)

$$\dot{y} = f(y), \tag{3.12}$$

where the vector y is in $\mathbb{R}^n$ and the function $f : \mathbb{R}^n \to \mathbb{R}^n$ is sufficiently differentiable. We denote the initial value of (3.12) by $y(0) = y_0$. We define the flow φ_t to be the mapping which, to the point y_0 in the phase space, associates the value $y(t)$ of the solution of the ODE with initial value $y(0) = y_0$.

Definition 1.3.1 *A* numerical one-step method $\Phi_h : \mathbb{R}^n \to \mathbb{R}^n$ *is a mapping that approximates the time h flow of the differential equation (3.12). The recursion $y_{n+1} = \Phi_h(y_n)$ yields an approximation to the solution on the grid $\{t_n\}$ where $t_n = nh$, for a constant step size h.*

The simplest of all numerical methods is the method proposed by Euler,

$$y_{n+1} = y_n + hf(y_n). \tag{3.13}$$

The *implicit Euler method* reads

$$y_{n+1} = y_n + hf(y_{n+1}). \tag{3.14}$$

Here, we see that the approximation y_{n+1} is defined implicitly by relation (3.14). Thus, one has to solve a nonlinear system of equations in each step of the numerical method.

Taking the mean of y_n and y_{n+1} in the argument of the function f, we obtain another implicit method: the *implicit midpoint rule*

$$y_{n+1} = y_n + hf(\frac{y_n + y_{n+1}}{2}). \tag{3.15}$$

For *partitioned systems*

$$\begin{aligned} \dot{p} &= f(p,q) \\ \dot{q} &= g(p,q), \end{aligned} \tag{3.16}$$

such as the Hamiltonian problem (1.1), we consider the *symplectic Euler method*

$$\begin{aligned} p_{n+1} &= p_n + hf(p_{n+1}, q_n) \\ q_{n+1} &= q_n + hg(p_{n+1}, q_n), \end{aligned} \tag{3.17}$$

which treats variables p, q by the implicit (resp. explicit) Euler method.

A widely used numerical method in MD is the Störmer-Verlet scheme. The one-step formulation of this method applied to the Hamiltonian problem (1.5) reads (see for example [PJY97],[HLW03] or [HLW02])

$$\begin{aligned} \dot{x}_{n+1/2} &= \dot{x}_n - \frac{h}{2}(\Omega^2 x_n + \nabla U(x_n)) \\ x_{n+1} &= x_n + h\dot{x}_{n+1/2} \\ \dot{x}_{n+1} &= \dot{x}_{n+1/2} - \frac{h}{2}(\Omega^2 x_{n+1} + \nabla U(x_{n+1})), \end{aligned} \tag{3.18}$$

where the new positions and momenta $x_{n+1}, \dot{x}_{n+1}$ at time t_{n+1} are computed from x_n and $\dot{x}_n$ in an explicit way.

By comparing the Taylor series of the Euler method (3.13) and the exact solution of (3.12), we obtain

$$\varphi_h(y) - \Phi_h(y) = \mathcal{O}(h^2).$$

Methods satisfying this relation for all sufficiently regular problems (3.12) are said to be of order 1. Generalizing this idea, one gets the following definition of the order of a numerical one step method.

Definition 1.3.2 *A one-step numerical method Φ_h has* order *p, if for all sufficiently regular problems (3.12) the* local error *$\varphi_h(y) - \Phi_h(y)$ satisfies*

$$\varphi_h(y) - \Phi_h(y) = \mathcal{O}(h^{p+1}).$$

For the numerical methods considered above, we can check (by comparing the Taylor series as before) that the implicit Euler method and the symplectic Euler have order 1, the midpoint scheme and the Störmer-Verlet method are of order 2.

The flow φ_t of problem (3.12) satisfies $\varphi_{-t}^{-1} = \varphi_t$, we want to know if this property is also shared by a numerical method.

Definition 1.3.3 *A one-step numerical method* $y_1 = \Phi_h(y_0)$ *is* symmetric *if* $\Phi_h \circ \Phi_{-h} = id$.

We can check (by exchanging $y_n \leftrightarrow y_{n+1}$ and $h \leftrightarrow -h$) that the midpoint and the Störmer-Verlet methods are symmetric.

Symmetric integrators are related to a property of the differential equation (3.12), namely:

Definition 1.3.4 *Let ρ be an invertible linear transformation in the phase space of problem (3.12). The problem is* ρ-reversible *if*

$$\rho(f(y)) = -f(\rho(y)) \quad \textit{for all} \quad y. \tag{3.19}$$

As an example, we consider partitioned system (3.16) where $f(-p,q) = f(p,q)$ and $g(-p,q) = -g(p,q)$. Here, the linear transformation ρ is given by $\rho(p,q) = (-p,q)$. All Hamiltonian system for which $H(p,q)$ is an even function of p are ρ-reversible with respect to this particular transformation.

For ρ-reversible differential equations, the exact flow satisfies

$$\rho \circ \varphi_t = \varphi_{-t} \circ \rho = \varphi_t^{-1} \circ \rho,$$

this motivates the following definition.

Definition 1.3.5 *A map $\Phi_h(y)$ is called* ρ-reversible *if*

$$\rho(\Phi_h(y)) = \Phi_h^{-1}(\rho(y)) \quad \textit{for all} \quad y. \tag{3.20}$$

We can show that the Störmer-Verlet method is ρ-reversible for partitioned system evoked just before this definition.

We finally mention a characteristic property of Hamiltonian systems. To do this, we first give the definition of this property and then give a theorem due to Poincaré (for a proof, see [HLW02]).

Definition 1.3.6 *A differentiable map $g : U \to \mathbb{R}^{2d}$, where $U \subset \mathbb{R}^{2d}$ is an open subset, is called* symplectic *if*

$$g'(p,q)^T J g'(p,q) = J, \tag{3.21}$$

for the structural matrix

$$J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}.$$

Theorem 1.3.7 *Let $H(p,q)$ be a twice continuously differentiable function on $U \subset \mathbb{R}^{2d}$. Then, for each fixed t, the flow φ_t is a symplectic transformation wherever it is defined.*

It is natural to extend this definition to numerical methods.

Definition 1.3.8 *The numerical one-step method $\Phi_h : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ is* symplectic *if it satisfies condition (3.21) whenever it is applied to a smooth Hamiltonian system.*

For the numerical methods defined above, we can show that the midpoint scheme, the Störmer-Verlet and the symplectic Euler methods are symplectic.

The Störmer-Verlet method (3.18) is commonly used in MD because of two important geometric properties: the symplecticity and the reversiblity under the application $\dot{x} \to -\dot{x}$. However, a major disadvantage is the restriction $h\omega < 2$ on the step size. For step sizes that do not satisfy this inequality, the numerical solution is unstable and explodes after a few steps. Due to this restriction on the step size h, for highly oscillatory second-order differential equations (1.6), this method is very costly (a lot of force evaluations). This motivates to search for methods that permits bigger step sizes. In the next section, we present the ideas that lead to the definition of *trigonometric methods*.

1.4 Trigonometric methods

In this section, we give a brief survey of a class of numerical methods designed for Hamiltonian problems (1.5).

The variation-of-constant formula gives for the exact solution of (1.6)

$$\begin{pmatrix} x(t) \\ \dot{x}(t) \end{pmatrix} = \begin{pmatrix} \cos(t\Omega) & \Omega^{-1}\sin(t\Omega) \\ -\Omega\sin(t\Omega) & \cos(t\Omega) \end{pmatrix} \begin{pmatrix} x_0 \\ \dot{x}_0 \end{pmatrix} + \int_0^t \begin{pmatrix} \Omega^{-1}\sin((t-s)\Omega) \\ \cos((t-s)\Omega) \end{pmatrix} g(x(s))ds.$$

Taking an approximation for the integral, one is led to the following definition.

Definition 1.4.1 Trigonometric methods *are given by the scheme*

$$\begin{aligned} x_{n+1} &= \cos(h\Omega)x_n + \Omega^{-1}\sin(h\Omega)\dot{x}_n + \frac{h^2}{2}\Psi g_n \\ \dot{x}_{n+1} &= -\Omega\sin(h\Omega)x_n + \cos(h\Omega)\dot{x}_n + \frac{h}{2}(\Psi_0 g_n + \Psi_1 g_{n+1}), \end{aligned} \tag{4.22}$$

where $g_n = g(\Phi x_n)$ and $\Phi = \phi(h\Omega)$, $\Psi = \psi(h\Omega)$, $\Psi_0 = \psi_0(h\Omega)$, $\Psi_1 = \psi_1(h\Omega)$ with even real-valued functions $\phi, \psi, \psi_0, \psi_1$ with $\phi(0) = \psi(0) = \psi_0(0) = \psi_1(0) = 1$. These functions are called filter functions.

Depending on the choice of these filter functions, we obtain different numerical methods. For example, taking $\psi(\zeta) = \mathrm{sinc}^2(\zeta), \phi(\zeta) = \mathrm{sinc}(\zeta), \psi_0 = \cos(\zeta)\phi(\zeta), \psi_1 = \phi(\zeta)$ (where $\mathrm{sinc}(\zeta) = \sin(\zeta)/\zeta$) one gets the mollified impulse method of Garcìa-Archilla et al. ([GASSS98]). For other choices of the filter functions, we obtain Gautschi-type methods (see [HL99]).

Trigonometric methods are well analysed in [HLW02, Chap. XIII], we just recall the conditions on the filter functions to obtain symplectic or symmetric methods. Exchanging $n \leftrightarrow n+1$ and $h \leftrightarrow -h$ in the definition of a trigonometric method, it is seen that the method is symmetric if and only if

$$\psi(\zeta) = \operatorname{sinc}(\zeta)\psi_1(\zeta), \qquad \psi_0(\zeta) = \cos(\zeta)\psi_1(\zeta). \tag{4.23}$$

The method is symplectic if and only if

$$\psi(\zeta) = \operatorname{sinc}(\zeta)\phi(\zeta). \tag{4.24}$$

Interesting properties for these methods is nearly conservation of the total and oscillatory energies over long time intervals under conditions on the filter functions. Moreover, a detailed analysis of the stability of the impulse methods can be found in [GASSS98] and one can see in [HL99] that, on intervals of length $\mathcal{O}(1)$ independent of ω, the order of convergence of Gautschi-types methods is two independently of how large ω is.

1.5 Other methods

We finally say a few words on two other types of methods: implicit methods and methods consisting by replacing the oscillations by constraints.

What happens when we use implicit methods, like for example the midpoint scheme, for the problem (1.6)? Unfortunately, if we want to use these methods, we have to solve nonlinear systems (this is usually very costly). Moreover, in [AR99a] and [AR99b]), the authors investigate, in particular, the use of the implicit midpoint rule to highly oscillatory problems. Beside resonance instability, other difficulties arise: one must require the step size to be small enough or else errors in the energy and other slowly varying quantities may grow undesirably. Or, worse, the computation may yield misleading information.

Another possibility to solve highly oscillatory differential equation is to freeze out the system ([Lea96]). This method is widely used in chemistry because "[...] molecular bond vibrations would occur so rapidly than an extremely short time step would be required to solve the equations of motion." [AT87, p.84]. It consists of considering the limit $\omega \to \infty$ so that the high oscillations disappear and are replaced by an algebraic constraint for the system. However, we can remark that for "[...]realistic molecular dynamics: bond length constraints are permissible but bond angle constraints not." [AT87, p.98]. Once this is done, one can use a numerical method designed for differential algebraic equations like SHAKE and RATTLE, see [Lea96], [AT87], [PJY97] or [LRS96].

As a conclusion, we would just keep in mind a sentence appearing in [PJY97]:“The best method to use is strongly dependent on the application.”

Chapter 2

Highly oscillatory differential equation

This chapter introduces the modulated Fourier expansion, the fundamental tool to study second-order differential equations (1.1). We also explain the near-conservation of the oscillatory energy over exponentially long time intervals. This Chapter is identical to the publication [CHL03].

2.1 Introduction

We study the system of differential equations

$$\ddot{x} + \Omega^2 x = g(x) \qquad \text{with} \qquad \Omega = \begin{pmatrix} 0 & 0 \\ 0 & \omega I \end{pmatrix} \tag{1.1}$$

where $\omega \gg 1$ and the nonlinearity is $g(x) = -\nabla U(x)$, so that the problem is Hamiltonian with

$$H(x, \dot{x}) = \frac{1}{2}\Big(\|\dot{x}\|^2 + \|\Omega x\|^2 \Big) + U(x). \tag{1.2}$$

An important property of such systems is the near-conservation over long times of the *oscillatory energy*

$$I(x, \dot{x}) \;=\; \frac{1}{2}\Big(\|\dot{x}_2\|^2 + \omega^2 \, \|x_2\|^2 \Big). \tag{1.3}$$

Here, the vectors $x = (x_1, x_2)$ and $\dot{x} = (\dot{x}_1, \dot{x}_2)$ are partitioned according to the partitioning of the matrix Ω in (1.1). A possible way of studying problems of the type (1.1) is via averaging techniques and Lindstedt series, see for example Neishtadt [Neĭ84], Murdock [Mur91], Pronin and Treschev [PT00]. The very problem (1.1) was thoroughly studied in Benettin, Galgani and Giorgilli [BGG89], Fassò [Fas90], and Bambusi and Giorgilli [BG93], using coordinate transformations of Hamiltonian perturbation theory. In this chapter we give a variant of their result, obtained with a completely different proof. It is based on writing the solution of (1.1) as a *modulated Fourier expansion*

$$x(t) = y(t) + \sum_{k \neq 0} \mathrm{e}^{\mathrm{i}k\omega t} z^k(t), \tag{1.4}$$

where $y(t)$ and $z^k(t)$ are smoothly varying functions (i.e., their derivatives are bounded independently of ω).

Such a representation of the solution has first been proposed by Miranker and van Veldhuizen [MvV78], who derived a scheme for constructing the "envelopes" $z^k(t)$. They suggested to compute numerically these envelopes and used them for approximating the solution $x(t)$. In [HL00] and [HLW02, Chap. XIII] this technique of modulated Fourier expansions has been further developed and used in the analysis of the long-time behaviour of numerical integrators when the time step is not small compared to ω^{-1}. Standard backward error analysis (see for example [HLW02, Chap. IX]) requires $\Delta t \cdot \omega$ to be small and therefore cannot be applied. In this situation, modulated Fourier expansions provide much insight into the long-time behaviour of numerical integrators. In this chapter, they are used to obtain rigorous long-time results for the exact solution of the differential equation.

The following result states the near-conservation of the oscillatory energy over time intervals that are exponentially long in ω. Here we assume that the initial values satisfy

$$\frac{1}{2}\Big(\|\dot x(0)\|^2 + \|\Omega x(0)\|^2\Big) \;\le\; E, \tag{1.5}$$

where E is independent of ω. (We do not require E to be small.)

Theorem 2.1.1 *Assume that $g(x) = -\nabla U(x)$ is analytic and bounded by M in the complex neighbourbood $D = \{x \in \mathbb{C}^n\,;\, \|x-\xi\| \le R$ for some ξ with $H(\xi,0) \le H(x(0),\dot x(0))\}$ of the set of energetically admissible positions. Furthermore, let the initial values $x(0), \dot x(0)$ satisfy (1.5). Then there exist positive constants $\gamma, C, \widehat C, \omega_0$ depending on E, M, and R (but not on ω) such that for $\omega \ge \omega_0$*

$$\big\|I\big(x(t),\dot x(t)\big) - I\big(x(0),\dot x(0)\big)\big\| \;\le\; C\,\omega^{-1} \qquad \text{for} \quad 0 \le t \le \widehat C\,\mathrm{e}^{\gamma\omega}.$$

The proof of this theorem will be given in the last section of this chapter. We first discuss the modulated Fourier expansion in Section 2.2, and we show that the coefficient functions of (1.4) are given by asymptotic differential and algebraic equations. The effect of truncating the asymptotic series is studied in Section 2.3. Whereas these two sections treat the general problem (1.1), the final Section 2.4 assumes that $g(x) = -\nabla U(x)$. It is shown that the coefficient functions of the modulated Fourier expansion are then exponentially close to the solution of a Hamiltonian system in an infinite dimensional space, which has two invariants: one is close to the Hamiltonian (1.2) and the other is close to the oscillatory energy (1.3).

Let us mention that the dominating fluctuation terms in the oscillatory energy can be given explicitly. Writing down the $\mathcal{O}(\omega^{-1})$ terms in $\mathcal{I}$ of (4.4) below we find that

$$J(x,\dot x) = \frac{1}{2}\Big(\|\dot x_2\|^2 + \omega^2\,\|x_2\|^2\Big) - x_2^T g_2(x_1,0) \tag{1.6}$$

satisfies

$$\big\|J\big(x(t),\dot x(t)\big) - J\big(x(0),\dot x(0)\big)\big\| \;\le\; C\,\omega^{-2}$$

on exponentially long time intervals. Since $x_2 = \mathcal{O}(\omega^{-1})$, this implies that the fluctuations in $I(x, \dot{x})$ are of size $\mathcal{O}(\omega^{-2})$ when $g_2(x_1, 0) = \mathcal{O}(\omega^{-1})$.

The techniques of this chapter can also be applied to the slightly more general situation where the potential $U(x)$ contains expressions of the form $\varphi_1(x_1, x_2) + \omega\varphi_2(x_1/\omega, x_2)$, such that the differential equation becomes

$$\begin{aligned} \ddot{x}_1 &= g_1(x_1, x_2) \\ \ddot{x}_2 + \omega^2 x_2 &= \omega g_2(x_1, x_2) \end{aligned}$$

with $g(x)$ depending smoothly on ω^{-1}. In this case, the quantity

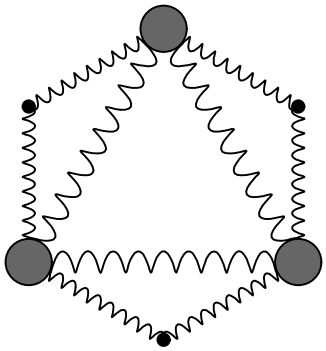
$$K(x, \dot{x}) = \frac{1}{2}\Big(\|\dot{x}_2\|^2 + \omega^2 \|x_2\|^2\Big) - \omega x_2^T g_2(x_1, 0) + \frac{1}{2}\|g_2(x_1, 0)\|^2 \tag{1.7}$$

satisfies

$$\big\|K\big(x(t), \dot{x}(t)\big) - K\big(x(0), \dot{x}(0)\big)\big\| \le C\,\omega^{-1}$$

on exponentially long time intervals. Notice that the additional terms in (1.7) are in general of size $\mathcal{O}(1)$, so that the oscillatory energy exhibits fluctuations that can be large independent of the size of ω.

Example. Inspired by an example of Bambusi and Giorgilli [BG93] we consider a closed chain of an even number of particles with alternate light and heavy masses. They interact through springs which are harmonic up to small perturbations, and neighbouring heavy particles interact also through arbitrary anharmonic springs (see the picture to the right). More precisely, we consider the Hamiltonian system with



$$\begin{aligned} H(\xi, \dot{\xi}) = \sum_{i=1}^{2N} \frac{\dot{\xi}_i^2}{2m_i} + \frac{1}{2}\sum_{i=1}^{2N} (\xi_i - \xi_{i-1})^2 + \sum_{j=1}^{N} \varphi_j(\xi_{2j} - \xi_{2j-2}) \\ + \sum_{i=1}^{2N} \psi_i\big(\sqrt{m}\,(\xi_i - \xi_{i-1})\big), \end{aligned}$$

where $m_{2j-1} = m \ll 1$ and $m_{2j} = 1$ for $j = 1, \dots, N$, and $\xi_0 = \xi_{2N}$. Applying the symplectic change of coordinates $\xi_i \mapsto \sqrt{m_i}\,\xi_i$, $\dot{\xi}_i \mapsto \dot{\xi}_i/\sqrt{m_i}$, and using the notation $\omega = 1/\sqrt{m}$, the Hamiltonian becomes

$$\begin{aligned} H(\xi, \dot{\xi}) \;=\; & \frac{1}{2}\sum_{i=1}^{2N} \dot{\xi}_i^2 + \frac{1}{2}\sum_{j=1}^{N} \Big((\xi_{2j} - \omega\xi_{2j-1})^2 + (\omega\xi_{2j-1} - \xi_{2j-2})^2\Big) \\ & + \sum_{j=1}^{N} \varphi_j(\xi_{2j} - \xi_{2j-2}) + \sum_{j=1}^{N} \Big(\psi_{2j}\Big(\frac{\xi_{2j}}{\omega} - \xi_{2j-1}\Big) + \psi_{2j-1}\Big(\xi_{2j-1} - \frac{\xi_{2j-2}}{\omega}\Big)\Big). \end{aligned}$$

We then consider an orthogonal linear transformation $\xi^* = Q\xi$ that takes the harmonic part of the Hamiltonian to diagonal form. It is given by

$$\xi^*_{2j-1} = \xi_{2j-1} - \frac{1}{2\omega}\big(\xi_{2j} + \xi_{2j-2}\big) + \mathcal{O}(\omega^{-2}),$$
$$\xi^*_{2j} = \xi_{2j} + \frac{1}{2\omega}\big(\xi_{2j+1} + \xi_{2j-1}\big) + \mathcal{O}(\omega^{-2}).$$

Omitting the stars, the Hamiltonian becomes (in the new variables)

$$H(\xi,\dot\xi) = \frac{1}{2}\sum_{i=1}^{2N}\dot\xi_i^2 + \omega^2\sum_{j=1}^{N}\xi_{2j-1}^2 + \Phi_1(\xi) + \Phi_2(\xi_1, \xi_2/\omega, \xi_3, \xi_4/\omega, \ldots),$$

which is of the form treated above.

Numerical Experiment. For a concrete example we put $N = 3$, $\omega = 50$, we let $\varphi_j(s) = \chi(\sqrt[6]{2} - s/\omega)$ with $\chi(s) = s^{-12} - s^{-6}$ be the Lennard-Jones potential, we take $\psi_{2j}(s) = s^2/2 + s^4/4$ for $j = 1, \ldots, N-1$, and $\psi_i(s) = 0$ else.

Figure 2.1 shows the components ξ_2, ξ_4, ξ_6, and $10\xi_5$ on the interval $0 \le t \le 10$. The factor 10 multiplying ξ_5 is included to show more clearly the oscillations of size $\mathcal{O}(\omega^{-1})$ in the numerical solution.

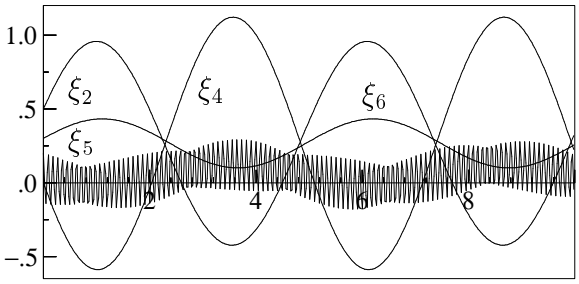


Figure 2.1: Solution components, where the non-zero initial positions are $\xi_2(0) = 0.5, \xi_3(0) = (2\omega)^{-1}, \xi_5(0) = \omega^{-1}, \xi_6(0) = 0.3$ and the non-zero initial velocities are $\dot\xi_1(0) = -\dot\xi_3(0) = \omega^{-1}, \dot\xi_2(0) = 0.8, \dot\xi_4(0) = -1, \dot\xi_6(0) = 0.2$.

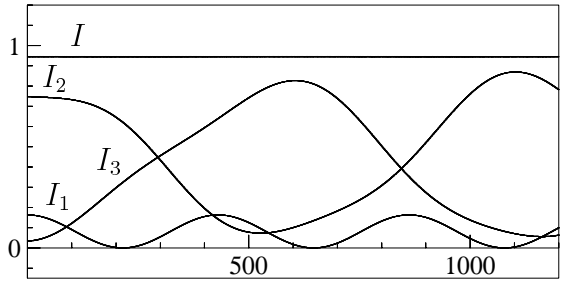


Figure 2.2: Oscillatory energy for the solution with initial values as in Fig. 2.1.

In Fig. 2.2 we plot the energies $I_j(\xi^*, \dot\xi^*) = \frac{1}{2}(\dot\xi^*_{2j-1})^2 + \omega^2(\xi^*_{2j-1})^2$ together with the oscillatory energy $I = I_1 + I_2 + I_3$ (cf. (1.3)) along the numerical solution on the interval $0 \le t \le 1200$. For this example, the expression $g_2(x_1, 0)$ is of size ω^{-1}, so that the oscillatory energy is conserved up to terms of size ω^{-2} (see (1.6)). Therefore, the oscillations cannot be observed in Fig. 2.2.

2.2 The modulated Fourier expansion

We write the system (1.1) in the equivalent form

$$\begin{aligned} \ddot x_1 &= g_1(x_1, x_2) \\ \ddot x_2 + \omega^2 x_2 &= g_2(x_1, x_2), \end{aligned} \tag{2.1}$$

where $\omega \gg 1$ represents the dominant frequency of the system. In this section we do not assume that $g(x)$ is the gradient of a potential. Our aim is to present a technique that allows us to separate the smooth and the oscillating parts of the solution of (2.1) and to write it in the form

$$\begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} = \begin{pmatrix} y_1(t) \\ y_2(t) \end{pmatrix} + \sum_{k \neq 0} \mathrm{e}^{\mathrm{i}k\omega t} \begin{pmatrix} z_1^k(t) \\ z_2^k(t) \end{pmatrix}, \tag{2.2}$$

where $y_i(t)$ and $z_i^k(t)$ are smoothly varying functions (i.e., their derivatives are bounded independently of ω). The functions $y_i(t)$ are real-valued and $z_i^k(t)$ are complex-valued. Since the solution $x_i(t)$ is real-valued, we have to require that $z_i^{-k} = \overline{z_i^k}$. We also use the notations $z_2 := z_2^1$ and $z_2^0 := y_2$.

Inserting (2.2) into (1.1), expanding the nonlinearity into a Taylor series around $\big(y_1(t), 0\big)$, and comparing the coefficients of $\mathrm{e}^{\mathrm{i}k\omega t}$ yields differential equations for the coefficient functions $y_i(t)$ and $z_i^k(t)$. With the exception of $y_1(t)$ they are of singular perturbation type. We have to find smooth solutions of these equations. As explained in [HL00], the functions y_1 and z_2 are seen to be given by differential equations of the form [1]

$$\ddot y_1 = \sum_{l \ge 0} \omega^{-l} F_{1l}(y_1, \dot y_1, z_2), \qquad \dot z_2 = \sum_{l \ge 1} \omega^{-l} F_{2l}(y_1, \dot y_1, z_2), \tag{2.3}$$

and the remaining functions by algebraic relations

$$z_i^k = \sum_{l \ge 0} \omega^{-l} G_{il}^k(y_1, \dot y_1, z_2). \tag{2.4}$$

Observe that $y_2 = z_2^0$, so that we also have an algebraic relation for y_2. Furthermore, for $i = 2$ and $k = 1$, we have the trivial identity $z_2^1 = z_2$ which implies

$$G_{20}^1(y_1, \dot y_1, z_2) = z_2, \qquad G_{2l}^1(y_1, \dot y_1, z_2) = 0 \quad \text{for} \quad l \ge 1. \tag{2.5}$$

Remember that z_i^{-k} is the complex conjugate of z_i^k, so that also G_{il}^{-k} is the complex conjugate of G_{il}^k.

[1] The series (2.3) and (2.4) are asymptotic expansions and do not converge in general. For convenience we nevertheless use the symbol $=$. Later, we shall truncate them suitably in order to get rigorous statements.

2.2.1 Recurrence relations for the coefficient functions

For a computation of the functions F_{il} and G_{il}^k in (2.3) and (2.4) it is convenient to introduce the Lie operator $\mathcal{L}_l$. It can be applied to smooth functions $G(y_1, \dot{y}_1, z_2)$ and it is defined for $l \geq 0$ by

$$\mathcal{L}_l G = D_2 G \cdot F_{1l} + D_3 G \cdot F_{2l} + \begin{cases} D_1 G \cdot \dot{y}_1 & \text{if } l = 0 \\ 0 & \text{if } l \geq 1, \end{cases} \tag{2.6}$$

where D_j denotes the partial derivative with respect to the jth argument of $G(y_1, \dot{y}_1, z_2)$. This definition is motivated by the fact that, whenever $y_1(t)$ and $z_2(t)$ are a solution of the differential equation (2.3), then we have

$$\frac{d}{dt} G\big(y_1(t), \dot{y}_1(t), z_2(t)\big) = \sum_{l \geq 0} \omega^{-l} \mathcal{L}_l G\big(y_1(t), \dot{y}_1(t), z_2(t)\big). \tag{2.7}$$

Lemma 2.2.1 *The function $\big(x_1(t), x_2(t)\big)$ of (2.2) with $y_i(t)$ and $z_i^k(t)$ given by (2.3) and (2.4) represents a formal solution of (2.1) if the coefficient functions F_{il} and G_{il}^k satisfy the following recurrence relations (for $l \geq 0$):*

$$\begin{aligned}
F_{1l} &= S_1(0,l) \\
G_{1l}^k &= \frac{1}{k^2}\Big(\sum_{m+n+j=l-2} \mathcal{L}_m \mathcal{L}_n G_{1j}^k + 2\mathrm{i}k \sum_{m+j=l-1} \mathcal{L}_m G_{1j}^k - S_1(k, l-2)\Big) \\
F_{2l} &= \frac{1}{2\mathrm{i}}\Big(S_2(1, l-1) - \sum_{m+j=l-1} \mathcal{L}_m F_{2j}\Big) \\
G_{2l}^k &= \frac{1}{1-k^2}\Big(S_2(k, l-2) - \sum_{m+n+j=l-2} \mathcal{L}_m \mathcal{L}_n G_{2j}^k - 2\mathrm{i}k \sum_{m+j=l-1} \mathcal{L}_m G_{2j}^k\Big).
\end{aligned}$$

The sums are over $m \geq 0, n \geq 0, j \geq 0$, and we have used the abbreviation

$$S_i(k,l) = \sum_{m,n \geq 0} \frac{1}{m!\, n!} \sum_{\substack{\alpha,\beta \\ s(\alpha)+s(\beta)=k}} \sum_{\substack{e,f \\ s(e)+s(f)=l}} D_1^m D_2^n g_i(y_1, 0)(G_{1e}^\alpha, G_{2f}^\beta).$$

Here, $\alpha = (\alpha_1, \ldots, \alpha_m), \beta = (\beta_1, \ldots, \beta_n), e = (e_1, \ldots, e_m), f = (f_1, \ldots, f_n)$ are multi-indices with $\alpha_i \neq 0$, β_i arbitrary, $e_i \geq 0$, $f_i \geq 0$, and $(G_{1e}^\alpha, G_{2f}^\beta) = (G_{1,e_1}^{\alpha_1}, \ldots, G_{1,e_m}^{\alpha_m}, G_{2,f_1}^{\beta_1}, \ldots, G_{2,f_n}^{\beta_n})$. We use the abbreviation $s(\alpha) = \sum_{i=1}^m \alpha_i$ and similarly for the other multi-indices.

Proof. Inserting the relation (2.2) into the first equation of the system (2.1), and expanding the nonlinearity into a Taylor series around $(y_1, 0)$, we obtain

$$\begin{aligned}
\ddot{y}_1 &+ \sum_{k \neq 0} \mathrm{e}^{\mathrm{i}k\omega t}\big(\ddot{z}_1^k + 2\mathrm{i}k\omega \dot{z}_1^k - k^2\omega^2 z_1^k\big) \\
&= \sum_{m,n \geq 0} \frac{1}{m!\, n!} \sum_{\alpha,\beta} \mathrm{e}^{\mathrm{i}\omega t(s(\alpha)+s(\beta))} D_1^m D_2^n g_1(y_1, 0)(z_1^\alpha, z_2^\beta),
\end{aligned}$$

where $(z_1^\alpha, z_2^\beta) = (z_1^{\alpha_1}, .., z_1^{\alpha_m}, z_2^{\beta_1}, .., z_2^{\beta_n})$, and the last sum is over all multi-indices α, β with $\alpha_i \neq 0$. We now insert our ansatz (2.3) for $\ddot{y}_1$ and (2.4) for z_i^k, we use the Lie derivative for expressing the derivatives of z_1^k, and thus obtain

$$\begin{aligned}\sum_{l\geq 0} \omega^{-l} F_{1l} \;+\; \sum_{k\neq 0} \mathrm{e}^{\mathrm{i}k\omega t}\Big(\sum_{m,n,j\geq 0} \omega^{-m-n-j} \mathcal{L}_m \mathcal{L}_n G_{1j}^k \\ + 2\mathrm{i}k \sum_{m,j\geq 0} \omega^{-m-j+1} \mathcal{L}_m G_{1j}^k - k^2 \sum_{j\geq 0} \omega^{-j+2} G_{1j}^k \Big) \\ = \sum_{m,n\geq 0} \frac{1}{m!\,n!} \sum_{\alpha,\beta} \mathrm{e}^{\mathrm{i}\omega t(s(\alpha)+s(\beta))} D_1^m D_2^n g_1(y_1,0) \\ \Big(\sum_{e\geq 0} \omega^{-s(e)} G_{1e}^\alpha, \sum_{f\geq 0} \omega^{-s(f)} G_{2f}^\beta \Big).\end{aligned}$$

We just have to compare the coefficients of $\mathrm{e}^{\mathrm{i}k\omega t}$ and ω^{-l} (resp. ω^{-l+2}) to obtain the recurrence relations for the functions F_{1l} and G_{1l}^k. This implies

$$G_{10}^k = 0, \quad G_{11}^k = 0 \qquad \text{for all} \;\; k \neq 0, \tag{2.8}$$

so that the series expansions (2.4) for all z_1^k start with the ω^{-2}-term.

Looking at the second equation of the system (2.1), we obtain

$$\begin{aligned}\ddot{y}_2 + \omega^2 y_2 \;+\; \sum_{k\neq 0} \mathrm{e}^{\mathrm{i}k\omega t}\big(\ddot{z}_2^k + 2\mathrm{i}k\omega \dot{z}_2^k + (1-k^2)\omega^2 z_2^k\big) \\ = \sum_{m,n\geq 0} \frac{1}{m!\,n!} \sum_{\alpha,\beta} \mathrm{e}^{\mathrm{i}\omega t(s(\alpha)+s(\beta))} D_1^m D_2^n g_2(y_1,0)(z_1^\alpha, z_2^\beta).\end{aligned}$$

We insert the ansatz (2.3) for $\dot{z}_2$ and (2.4) for z_i^k, and in the same way as above we get the recurrence relations for the functions F_{2l} and G_{2l}^k. They imply

$$G_{20}^k = 0, \quad G_{21}^k = 0 \qquad \text{for} \;\; k \neq \pm 1, \tag{2.9}$$

so that also the expansions (2.4) for z_2^k ($k \neq \pm 1$) start with the ω^{-2}-term. □

2.2.2 Estimates for the functions F_{ij} and G_{ij}^k

Our next aim is to get upper bounds for the coefficient functions F_{ij} and G_{ij}^k of (2.3) and (2.4). Since they depend on the derivatives of $g_i(x_1, x_2)$, it is natural to require $g(x)$ to be analytic and bounded (by M) in a suitable complex domain, say in $\{(x_1, x_2)\,;\, \|x_1 - y_{10}\| \leq 4R, \|x_2\| \leq 3R\}$. Cauchy's estimates then imply

$$\|D_1^m D_2^n g_i(y_1, 0)\| \leq m!\, n!\, M\, (3R)^{-m-n} \qquad \text{for} \quad \|y_1 - y_{10}\| \leq R \tag{2.10}$$

and for all $n, m \geq 0$. This is our main assumption of this section. To obtain the desired estimates for the coefficient functions we combine and adapt the techniques of [BG94] and [HLW02, Sect. IX.5].

We fix a value $\mathcal{Y}_0 = (y_{10}, \dot{y}_{10}, 0)$, and consider the complex ball

$$B_\rho(\mathcal{Y}_0) = \big\{(y_1, \dot{y}_1, z_2)\,;\, \|y_1 - y_{10}\| \le \rho R, \|\dot{y}_1 - \dot{y}_{10}\| \le \rho M, \|z_2\| \le \rho R\big\}. \tag{2.11}$$

For a function $G(y_1, \dot{y}_1, z_2)$ defined on $B_\rho(\mathcal{Y}_0)$ we let

$$\|G\|_\rho = \max\big\{\|G(y_1, \dot{y}_1, z_2)\|\,;\, (y_1, \dot{y}_1, z_2) \in B_\rho(\mathcal{Y}_0)\big\}. \tag{2.12}$$

Since the coefficient functions are defined via expressions of the form $\mathcal{L}_l G$, the following lemma will be useful.

Lemma 2.2.2 *Let G be analytic and bounded on $B_\rho(\mathcal{Y}_0)$, and let F_{1l} and F_{2l} be bounded on $B_\sigma(\mathcal{Y}_0)$ with $0 \le \sigma < \rho$. Then we have*

$$\begin{aligned}
\|\mathcal{L}_0 G\|_\sigma &\le \tfrac{1}{\rho-\sigma} \cdot \|G\|_\rho \cdot \max\big(\|F_{10}\|_\sigma/M, \|\dot{y}_1\|_\sigma/R\big),\\
\|\mathcal{L}_l G\|_\sigma &\le \tfrac{1}{\rho-\sigma} \cdot \|G\|_\rho \cdot \max\big(\|F_{1l}\|_\sigma/M, \|F_{2l}\|_\sigma/R\big) \qquad \textit{for} \quad l \ge 1.
\end{aligned}$$

Proof. Consider $\alpha(\zeta) = G\big(y_1, \dot{y}_1 + \zeta F_{1l}(y_1, \dot{y}_1, z_2), z_2 + \zeta F_{2l}(y_1, \dot{y}_1, z_2)\big)$, where $(y_1, \dot{y}_1, z_2) \in B_\sigma(\mathcal{Y}_0)$. This function is analytic for $|\zeta| \le \varepsilon$ with $\varepsilon := (\rho-\sigma)\big/ \max\big(\|F_{1l}\|_\sigma/M, \|F_{2l}\|_\sigma/R\big)$. Since $\alpha'(0) = (\mathcal{L}_l G)(y_1, \dot{y}_1, z_2)$, Cauchy's estimate yields

$$\|(\mathcal{L}_l G)(y_1, \dot{y}_1, z_2)\| = \|\alpha'(0)\| \le \frac{1}{\varepsilon} \sup_{|\zeta| \le \varepsilon} \|\alpha(\zeta)\| \le \frac{1}{\varepsilon} \|G\|_\rho\,,$$

which proves the statement for $l \ge 1$. For $l = 0$ we have to consider the function $\alpha(\zeta) = G\big(y_1 + \zeta \dot{y}_1, \dot{y}_1 + \zeta F_{10}(y_1, \dot{y}_1, z_2), z_2\big)$, because $F_{20} = 0$ by Lemma 2.2.1. □

The use of Lemma 2.2.2 implies that we cannot work with only one norm $\|\cdot\|_\rho$ for finding estimates of the coefficient functions. We therefore fix a positive integer L, put $\delta = 1/(2L)$, and consider the norms corresponding to balls with shrinking radius $\rho = 1 - l\delta$ $(0 \le l \le L)$.

Lemma 2.2.3 *Let $\mathcal{Y}_0 = (y_{10}, \dot{y}_{10}, 0)$ be given, and assume that (2.10) holds. The functions F_{ij} and G_{ij}^k of Lemma 2.2.1 satisfy*

$$\begin{aligned}
&\|F_{10}\|_1 \le a_0 M, \qquad \|\dot{y}_1\|_1 \le a_0 R\\
&\|F_{1l}\|_{1-l\delta} \le a_l M, \qquad \|F_{2l}\|_{1-l\delta} \le a_l R, \qquad 1 \le l \le L\\
&\|G_{20}^{-1}\|_1 + \|G_{20}^{1}\|_1 \le b_0 R\\
&\max\Big(\sum_{k\neq 0} k^2 \|G_{1l}^k\|_{1-l\delta}\ ,\ \sum_{k\in\mathbb{Z}} |1-k^2|\,\|G_{2l}^k\|_{1-l\delta}\Big) \le b_l R, \qquad 1 \le l \le L,
\end{aligned}$$

where $a_0 = \max(9, (\|\dot{y}_{10}\|_1 + M)/R)$, $b_0 = 2$, and the generating functions $a(\zeta) = \sum_{l\ge1} a_l \zeta^l$ and $b(\zeta) = \sum_{l\ge1} b_l \zeta^l$ are implicitly given by

$$\begin{aligned}
a(\zeta) &= -9 + 9\big(1 + \tfrac{M\zeta}{2R}\big)\big(1 - b(\zeta)\big)^{-2} + \tfrac{\zeta}{2\delta}\big(a_0 + a(\zeta)\big)a(\zeta),\\
b(\zeta) &= \tfrac{9M\zeta^2}{R}\big(1 - b(\zeta)\big)^{-2} + \tfrac{2\zeta}{\delta}\big(a_0 + a(\zeta)\big)\big(b_0 + b(\zeta)\big)\\
&\qquad + \tfrac{\zeta^2}{\delta^2}\big(a_0 + a(\zeta)\big)^2\big(b_0 + b(\zeta)\big).
\end{aligned} \tag{2.13}$$

Proof. (a) In this proof we shall use the shorthand notation

$$\|G\|_l := \|G\|_{1-l\delta} = \max\big\{\|G(y_1,\dot y_1,z_2)\|\,;\,(y_1,\dot y_1,z_2)\in B_{1-l\delta}(\mathcal{Y}_0)\big\}. \tag{2.14}$$

Observe that $\|G\|_l$ is a decreasing function of l.

To obtain the desired statement, we begin with some estimations and then we prove the result of this Lemma by induction on l.

(b) Because of (2.8), (2.9) and (2.5), the above estimates for G^k_{il} also imply

$$\sum_{k\neq 0}\|G^k_{1l}\|_l \,\le\, b_l R, \qquad \sum_{k\in\mathbb{Z}}\|G^k_{2l}\|_l \,\le\, b_l R \qquad \text{for} \quad l\ge 0. \tag{2.15}$$

Using these relations and the analyticity assumption (2.10), we are able to majorize the $S_i(k,l)$ as follows:

$$\begin{aligned}
\sum_{k\in\mathbb{Z}}\|S_i(k,l)\|_l \;&\le \sum_{m,n\ge 0}\frac{m!\,n!}{m!\,n!}\sum_{\substack{\alpha,\beta\\ \alpha_i\neq 0}}\sum_{\substack{s(e)+s(f)\\ =l}} M(3R)^{-m-n}\|G^{\alpha_1}_{1e_1}\|_l\ldots\|G^{\beta_1}_{2f_1}\|_l\ldots\\
&\le\; M\sum_{m,n\ge 0}\;\sum_{s(e)+s(f)=l} 3^{-m-n}b_{e_1}\ldots b_{e_m}b_{f_1}\ldots b_{f_n}\\
&\le\; M\sum_{j\ge 0}(j+1)\sum_{d_1+\ldots+d_j=l} 3^{-j}b_{d_1}\ldots b_{d_j} \;=\; Mc_l,
\end{aligned}$$

where c_l $(l\ge 0)$ are the coefficients of the generating function

$$\sum_{l\ge 0}c_l\zeta^l = c(\zeta) = \frac{1}{\big(1-\frac{b_0+b(\zeta)}{3}\big)^2} = \frac{9}{\big(1-b(\zeta)\big)^2}.$$

The second equality follows from the derivative of the geometric series. We have used $\|G^{\alpha_1}_{1e_1}\|_l \le \|G^{\alpha_1}_{1e_1}\|_{e_1}$ and $\|G^{\beta_1}_{2f_1}\|_l \le \|G^{\beta_1}_{2f_1}\|_{f_1}$, which are a consequence of $e_1\le l$ and $f_1\le l$.

(c) For $m+n+j=l-2$ a twofold application of Lemma 2.2.2 yields

$$\|\mathcal{L}_m\mathcal{L}_nG^k_{ij}\|_l \le \frac{1}{\delta^2}\,\|G^k_{ij}\|_j\,a_m\,a_n \quad\text{and}\quad \sum_{k\neq 0}\|\mathcal{L}_m\mathcal{L}_nG^k_{1j}\|_l \le \frac{R}{\delta^2}\,b_j\,a_m\,a_n.$$

This implies

$$\sum_{k\neq 0}\sum_{m+n+j=l-2}\|\mathcal{L}_m\mathcal{L}_nG^k_{1j}\|_l \le \frac{R}{\delta^2}\,d_{l-2},$$

where the generating function of the d_l is

$$d(\zeta) = \sum_{l\ge 0}d_l\zeta^l = \big(b_0+b(\zeta)\big)\big(a_0+a(\zeta)\big)^2.$$

The same estimate is obtained for $\sum_{k\in\mathbb{Z}}\sum_{m+n+j=l-2}\|\mathcal{L}_m\mathcal{L}_nG^k_{2j}\|_l$.

(d) In order to estimate $|k|\,\|\mathcal{L}_m G_{ij}^k\|_l$ for $m+j=l-1$, we observe that similarly to (2.15) also

$$\sum_{k\in\mathbb{Z}} |k|\,\|G_{1l}^k\|_l \;\le\; b_l R, \qquad \sum_{k\in\mathbb{Z}} |k|\,\|G_{2l}^k\|_l \;\le\; b_l R \qquad \text{for} \quad l\ge 0 \tag{2.16}$$

holds. As in part (b) we thus obtain

$$\sum_{k\in\mathbb{Z}} |k| \sum_{m+j=l-1} \|\mathcal{L}_m G_{ij}^k\|_l \;\le\; \frac{R}{\delta}\, q_{l-1},$$

where the generating function for the q_l is

$$q(\zeta) = \sum_{l\ge 0} q_l \zeta^l = \big(b_0 + b(\zeta)\big)\big(a_0 + a(\zeta)\big).$$

(e) After these preparations the statement can be proved by induction on l. The bounds a_0 and b_0 are defined just to satisfy the estimates for $l=0$. The form of the generating functions for a_l and b_l are a consequence of the recurrence relations of Lemma 2.2.1 and of parts (b), (c) and (d) of this proof. □

To get bounds on the expressions of Lemma 2.2.3, we have to majorize a_l and b_l. This can be done with the help of Cauchy's inequalities, because the generating functions $a(\zeta)$ and $b(\zeta)$ are analytic in a neighbourhood of the origin. Since the equations (2.13) depend on δ, R and M, we have to be careful in determining the radius of the disc of analyticity. In the following we assume $M \ge R$. This can be done without loss of generality, because we can always increase M without violating (2.10) or, even better, we can rescale time in the differential equation and thus multiply $g(x)$ by a scalar factor.

Theorem 2.2.4 *We fix $\mathcal{Y}_0 = (y_{10}, \dot y_{10}, 0)$, and we assume that the nonlinearity $g(x)$ satisfies (2.10) with $M \ge R$, and that $\|\dot y_{10}\| \le M$. The coefficient functions of Lemma 2.2.1 then satisfy for $l \ge 1$*

$$\|F_{1l}\|_{1/2} \le \mu M\Big(\frac{\nu l M}{R}\Big)^l, \qquad \|F_{2l}\|_{1/2} \le \mu R\Big(\frac{\nu l M}{R}\Big)^l,$$
$$\max\Big(\sum_{k\ne 0} k^2 \|G_{1l}^k\|_{1/2}, \sum_{k\in\mathbb{Z}} |1-k^2|\,\|G_{2l}^k\|_{1/2}\Big) \le \mu R\Big(\frac{\nu l M}{R}\Big)^l,$$

where μ and ν only depend on an upper bound of M/R but not on the other data of the differential equation. The norm is that of (2.12).

Proof. We multiply the ζ in (2.13) either by $\frac{1}{\delta} \ge 1$ or by $\frac{M}{R} \ge 1$ so that the relations only depend on $\frac{\zeta M}{\delta R}$, $a(\zeta)$, and $b(\zeta)$. This makes the coefficients a_l and b_l at worst larger, so that

the estimates of Lemma 2.2.3 still hold. We then introduce the new variables $\hat\zeta = \zeta M/\delta R$, $\widehat a(\hat\zeta) = a(\zeta)$, and $\widehat b(\hat\zeta) = b(\zeta)$, so that (2.13) becomes

$$\begin{aligned} \widehat a(\hat\zeta) &= -9 + 9\big(1+\tfrac{\hat\zeta}{2}\big)\big(1-\widehat b(\hat\zeta)\big)^{-2} + \tfrac{\hat\zeta}{2}\big(a_0+\widehat a(\hat\zeta)\big)\widehat a(\hat\zeta), \\ \widehat b(\hat\zeta) &= 9\,\hat\zeta^2\big(1-\widehat b(\hat\zeta)\big)^{-2} + 2\hat\zeta\big(a_0+\widehat a(\hat\zeta)\big)\big(2+\widehat b(\hat\zeta)\big) \\ &\qquad + \hat\zeta^2\big(a_0+\widehat a(\hat\zeta)\big)^2\big(2+\widehat b(\hat\zeta)\big). \end{aligned} \tag{2.17}$$

Observe that $a_0 \le \max(9, 2M/R)$, which is a consequence of $\|\dot y_{10}\| \le M$.

In the equations (2.17) we obtain $\widehat a = 0$, $\widehat b = 0$ for $\hat\zeta = 0$. The Jacobian matrix at $\widehat a = \widehat b = 0$ is invertible and the Implicit Function Theorem can be applied. This proves the existence of constants μ and ν, such that $\widehat a(\hat\zeta)$ and $\widehat b(\hat\zeta)$ are analytic in the disc $|\hat\zeta| \le 2/\nu$ and bounded by μ. Cauchy's inequalities thus prove that the lth coefficient of these generating functions is bounded by $\mu(\nu/2)^l$. This yields

$$a_l\Big(\frac{\delta R}{M}\Big)^l \le \mu\Big(\frac{\nu}{2}\Big)^l, \qquad b_l\Big(\frac{\delta R}{M}\Big)^l \le \mu\Big(\frac{\nu}{2}\Big)^l.$$

Putting $l = L$ in the estimates of Lemma 2.2.3 and inserting the just obtained upper bounds for a_L and b_L, proves the theorem. We use the fact that $1 - L\delta = 1/2$. □

2.3 Exponentially small error estimates

In general, the series expansions in (2.3) and (2.4) diverge, even for arbitrarily large ω. For obtaining rigorous statements we have to truncate these series. We thus consider

$$\ddot y_1 = \sum_{0\le l\le N} \omega^{-l} F_{1l}(y_1, \dot y_1, z_2), \quad \dot z_2 = \sum_{1\le l\le N} \omega^{-l} F_{2l}(y_1, \dot y_1, z_2), \tag{3.1}$$

$$z_i^k = \sum_{2\le l\le N} \omega^{-l} G_{il}^k(y_1, \dot y_1, z_2). \tag{3.2}$$

The choice of the truncation index will be made on the basis of the estimates of Theorem 2.2.4. The lth term in the expansions (2.3) and (2.4) is majorized by Const $(\nu l M/\omega R)^l$, which is minimal for $\nu l M/\omega R = 1/\mathrm{e}$. We therefore choose the integer truncation index N such that

$$N \le \frac{\omega R}{\mathrm{e}\nu M} < N + 1. \tag{3.3}$$

Using the inequality

$$\sum_{2\le l\le N} l^2\Big(\frac{\nu l M}{\omega R}\Big)^{l-2} \le \sum_{2\le l\le N} l^2\left(\frac{l}{\mathrm{e}N}\right)^{l-2} \le 8.65,$$

which can be checked numerically for small N, and the left-hand expression of which is a decreasing function of N for large N, it immediately follows from Theorem 2.2.4 that

$$\sum_{k\neq 0} k^2 \sum_{2\le l\le N} \omega^{-l}\|G^k_{1l}\|_{1/2} \;\le\; 8.65\,\mu\,R\Big(\frac{\nu M}{\omega R}\Big)^2 \;\le\; \text{Const}\cdot R\Big(\frac{M}{\omega R}\Big)^2. \tag{3.4}$$

The remaining bounds of Theorem 2.2.4 yield similar estimates also for G^k_{2l}, F_{1l}, and F_{2l}.

2.3.1 Initial values for the modulated Fourier expansion

In this section we consider the function

$$\begin{pmatrix}\widetilde{x}_1(t)\\ \widetilde{x}_2(t)\end{pmatrix} = \begin{pmatrix}y_1(t)\\ y_2(t)\end{pmatrix} + \sum_{k\neq 0} \mathrm{e}^{\mathrm{i}k\omega t}\begin{pmatrix}z^k_1(t)\\ z^k_2(t)\end{pmatrix}, \tag{3.5}$$

where $y_i(t)$ and $z^k_i(t)$ are solutions of the truncated system (3.1)–(3.2). The sum over k is still infinite.

In the following we consider the differential equation (2.1) with initial values $x_1(0) = x_{10}$, $\dot{x}_1(0) = \dot{x}_{10}$, $x_2(0) = x_{20}$, $\dot{x}_2(0) = \dot{x}_{20}$, and we assume that the harmonic energy of these initial values is bounded by E independent of ω, see (1.5). We first show that to these initial values there correspond (locally) unique initial values for the system (3.1), such that $\widetilde{x}(0) = x(0)$ and $\dot{\widetilde{x}}(0) = \dot{x}(0)$. We then show that the function (3.5), obtained with these initial values for y_1, $\dot{y}_1$ and z_2, has an exponentially small defect when it is inserted into (2.1).

Lemma 2.3.1 *Consider the differential equation (2.1) with initial values $x(0) = (x_{10}, x_{20})$, $\dot{x}(0) = (\dot{x}_{10}, \dot{x}_{20})$ satisfying (1.5). Assume that the nonlinearity $g(x)$ is analytic in a ball $\{(x_1, x_2)\,|\,\|x_1 - x_{10}\| \le 4R, \|x_2\| \le 3R\}$ and bounded by M, with $M \ge R$. For sufficiently large ω ($M/\omega R \le \gamma$, where γ does not depend on ω) there exist (locally) unique initial values $y_1(0) = y_{10}$, $\dot{y}_1(0) = \dot{y}_{10}$, $z_2(0) = z_{20}$ for the system (3.1), such that*

$$x(0) = \widetilde{x}(0), \qquad \dot{x}(0) = \dot{\widetilde{x}}(0) \tag{3.6}$$

with $\widetilde{x}(t)$ from (3.5). These initial values satisfy

$$\begin{aligned} x_{10} &= y_{10} + \mathcal{O}(R\omega^{-2}), & x_{20} &= z_{20} + \overline{z}_{20} + \mathcal{O}(R\omega^{-2}),\\ \dot{x}_{10} &= \dot{y}_{10} + \mathcal{O}(R\omega^{-1}), & \dot{x}_{20} &= \mathrm{i}\omega z_{20} - \mathrm{i}\omega\overline{z}_{20} + \mathcal{O}(R\omega^{-1}), \end{aligned}$$

where the constant symbolizing the $\mathcal{O}(\cdot)$ can depend on M/R and on the harmonic energy E, but not on ω.

Proof. Using the truncated relations (3.2) and the Lie operator $\mathcal{L}_k$, the condition (3.6) becomes

$$\begin{aligned}
x_{10} &= y_{10} + \sum_{k\neq 0} \sum_{2\le l\le N} \omega^{-l}\, G^k_{1l}(y_{10}, \dot y_{10}, z_{20}, \overline{z}_{20}) \\
x_{20} &= z_{20} + \overline{z}_{20} + \sum_{|k|\neq 1} \sum_{2\le l\le N} \omega^{-l}\, G^k_{2l}(y_{10}, \dot y_{10}, z_{20}, \overline{z}_{20}) \\
\dot x_{10} &= \dot y_{10} + \sum_{k\neq 0} \sum_{2\le l\le N} \omega^{-l}\Big((\mathrm{i}k\omega) G^k_{1l}(y_{10}, \dot y_{10}, z_{20}, \overline{z}_{20}) \\
&\qquad\qquad + \sum_{0\le s\le N} \omega^{-s}\, (\mathcal{L}_s G^k_{1l})(y_{10}, \dot y_{10}, z_{20}, \overline{z}_{20})\Big) \\
(\mathrm{i}\omega)^{-1}\dot x_{20} &= z_{20} - \overline{z}_{20} + (\mathrm{i}\omega)^{-1} \sum_{|k|\neq 1} \sum_{2\le l\le N} \omega^{-l}\Big((\mathrm{i}k\omega) G^k_{2l}(y_{10}, \dot y_{10}, z_{20}, \overline{z}_{20}) \\
&\qquad\qquad + \sum_{0\le s\le N} \omega^{-s}\, (\mathcal{L}_s G^k_{2l})(y_{10}, \dot y_{10}, z_{20}, \overline{z}_{20})\Big),
\end{aligned}$$

Collecting the unknown variables into a vector $\mathcal{Y}_0 = (y_{10}, \dot y_{10}, z_{20}, \overline{z}_{20})$, this system can be readily brought to the form $\mathcal{Y}_0 = \mathcal{F}(\mathcal{Y}_0)$. Using Cauchy's inequalities and (3.4), we have $\|\mathcal{F}'(\mathcal{Y})\| \le \mathrm{Const}\cdot(\frac{M}{\omega R}) < 1$ if $M/\omega R$ is sufficiently small. This implies, by the Mean Value Theorem, that $\mathcal{F}$ is a contraction on the closed ball

$$B = \{(y_1, \dot y_1, z_2) \,|\, \|y_1 - x_{10}\| \le R/4, \|\dot y_1 - \dot x_{10}\| \le M/4, \|z_2\| \le R/4\}.$$

Furthermore, by (1.5), (3.4) and using the fact that $M/\omega R$ is sufficiently small, we have $\mathcal{F}(B) \subset B$. To conclude the proof, we apply the Banach Fixed Point Theorem to solve the nonlinear system $\mathcal{Y} = \mathcal{F}(\mathcal{Y})$. □

2.3.2 Estimation of the defect

After having found suitable initial values for the differential equation (3.1), wich exist for $\omega \ge \omega_0$ with a sufficiently large ω_0, we investigate the length of the time-interval such that the solution exists and remains in the ball

$$B = \{(y_1, \dot y_1, z_2) \,|\, \|y_1 - y_{10}\| \le R/2, \|\dot y_1 - \dot y_{10}\| \le M/2, \|z_2\| \le R/2\}.$$

We assume that the nonlinearity $g(x)$ satisfies (2.10) with $M \ge R$ and that $\|\dot y_{10}\| \le M$ (this assumption is essentially a definition of M and R). Similar to (3.4), the estimates of Theorem 2.2.4 then yield

$$\begin{aligned}
\sum_{0\le l\le N} \omega^{-l}\, \|F_{1l}(y_1, \dot y_1, z_2)\|_{1/2} &\le \mathrm{Const}\cdot M \\
\sum_{1\le l\le N} \omega^{-l}\, \|F_{2l}(y_1, \dot y_1, z_2)\|_{1/2} &\le \mathrm{Const}\cdot R\Big(\frac{M}{\omega R}\Big) \le \mathrm{Const}\cdot M\cdot\omega^{-1}
\end{aligned} \tag{3.7}$$

for $(y_1, \dot{y}_1, z_2) \in B$. As long as the solution of (3.1) remains in B, we thus have the estimates

$$\begin{aligned} \|y_1(t) - y_{10}\| &\le t\,\|\dot{y}_{10}\| + t^2 M\text{Const} \\ \|\dot{y}_1(t) - \dot{y}_{10}\| &\le t\,M\text{Const} \\ \|z_2(t) - z_{20}\| &\le t\,M\omega^{-1}\text{Const}. \end{aligned} \tag{3.8}$$

This proves the existence of a $T > 0$ such that $\big(y_1(t), \dot{y}_1(y), z_2(t)\big) \in B$ for $0 \le t \le T$. As the generic constant Const, also T only depends on an upper bound of M/R.

In the following we denote

$$y^0(t) = \begin{pmatrix} y_1(t) \\ y_2(t) \end{pmatrix}, \qquad y^k(t) = \mathrm{e}^{\mathrm{i}k\omega t} \begin{pmatrix} z_1^k(t) \\ z_2^k(t) \end{pmatrix}, \tag{3.9}$$

where $y_i(t)$ and $z_i^k(t)$ are the solution of the system (3.1)–(3.2). The approximate solution $\widetilde{x}(t)$ of (3.5) is thus equal to $\sum_k y^k(t)$. Without any truncation of the series in (3.1)–(3.2), the functions $y^k(t)$ are formally a solution of

$$\ddot{y}^k + \Omega^2 y^k = \sum_{m\ge 0} \frac{1}{m!} \sum_{s(\alpha)=k, \alpha_i \neq 0} g^{(m)}(y^0)\big(y^{\alpha_1}, \dots, y^{\alpha_m}\big), \tag{3.10}$$

because they are obtained by comparing the coefficients of $\mathrm{e}^{\mathrm{i}k\omega t}$ (see the proof of Lemma 2.2.1). Let us study here the effect of the truncation.

Theorem 2.3.2 *Consider the differential equation (2.1) with initial values $x(0)$ and $\dot{x}(0)$ satisfying (1.5). Assume that the nonlinearity $g(x)$ is analytic in the complex ball $\{(x_1, x_2)|\ \|x_1 - x_1(0)\| \le 4R, \|x_2\| \le 4R\}$ and bounded by M with $M \ge R$ and let $\|\dot{y}_{10}\| \le M$. Let the truncation index N in (3.1) and (3.2) be determined by (3.3). Then, there exist $\gamma > 0, T > 0$ and $\omega_0 > 0$ such that the defect*

$$\delta_k(t) = \ddot{y}^k(t) + \Omega^2 y^k(t) - \sum_{m\ge 0} \frac{1}{m!} \sum_{s(\alpha)=k, \alpha_i \neq 0} g^{(m)}(y^0(t))\big(y^{\alpha_1}(t), \dots, y^{\alpha_m}(t)\big)$$

satisfies for $0 \le t \le T$ and for $\omega \ge \omega_0$

$$\sum_{k\in\mathbb{Z}} \|\delta_k(t)\| \le C\,M\,\mathrm{e}^{-\gamma\omega}.$$

The constants C, γ, T, ω_0 only depend on an upper bound of M/R but not on ω.

Proof. First we let N and ω be independent variables (for the time being not related by (3.3)), and we consider the defect as a function of t, N, and ω^{-1}, i.e., $\delta_k(t) = \delta_k(t, N, \omega^{-1})$. By the construction of the coefficient functions y^k, the defect δ_k is an analytic function of $\zeta = \omega^{-1}$ in a neighbourhood of the origin, and moreover, $\delta_k = \mathcal{O}(\omega^{-N-1})$. Therefore, the following function is analytic in a neighbourhood of the origin:

$$F(\zeta) = \sum_{|k|\le m} u_k^*\,\delta_k(t, N, \zeta)\,\zeta^{-(N+1)},$$

where m is an arbitrary integer, and the u_k are arbitrary vectors of unit norm. For $t \le T$, with T sufficiently small (see (3.8)), the function $F(\omega^{-1})$ is well defined for $|\omega^{-1}| \le \varepsilon_N$, where

$$\varepsilon_N := \frac{R}{2\nu MN} ,$$

so that the Maximum Principle can be applied on this disk. For $|\omega^{-1}| = \varepsilon_N$, i.e., for $|\omega|$ and N related like in (3.3) but with 2 instead of e in the denominator, the bounds (3.4) and (3.7) are still valid (except that the constant 8.65 increases to 12.4).

For $t \le T$, we have $\|y^0(t) - x(0)\| \le R$ and Cauchy's estimates yield

$$\begin{aligned}\sum_{k\in\mathbb{Z}} \Big\| \sum_{m\ge 0} \frac{1}{m!} \sum_{s(\alpha)=k, \alpha_i\neq 0} & g^{(m)}(y^0(t))\big(y^{\alpha_1}(t),\dots,y^{\alpha_m}(t)\big)\Big\| \\ &\le M \sum_{m\ge 0} \frac{1}{m!} \sum_{\alpha_i\neq 0} \dots \sum_{\alpha_m\neq 0} m!(3R)^{-m} \|y^{\alpha_1}\| \dots \|y^{\alpha_m}\| \le \text{Const}\cdot M.\end{aligned}$$

The last inequality is a consequence of (3.4) and (3.7), which yield

$$\sum_{\alpha\neq 0} \|y^\alpha\| \le \text{Const}\cdot M\cdot \omega^{-1}$$

which is smaller than $2R$ for $\omega \ge \omega_0$ (take ω_0 greater if necessary). Again by (3.4) and (3.7), we obtain

$$\sum_{k\in\mathbb{Z}} \|\ddot{y}^k + \Omega^2 y^k\| = \sum_{k\in\mathbb{Z}} \|\ddot{z}^k + 2ik\omega\dot{z}^k - k^2\omega^2 z^k + \Omega^2 z^k\| \le \text{Const}\cdot M.$$

Putting this together, we obtain the bound

$$\sum_{k\in\mathbb{Z}} \|\delta_k(t,N,\zeta)\| \le \text{Const}\cdot M \qquad \text{for} \quad |\zeta| = \varepsilon_N.$$

With the Maximum Principle, this gives for $|\omega^{-1}| \le \varepsilon_N$

$$\begin{aligned}|F(\omega^{-1})| &\le \max_{|\zeta|=\varepsilon_N} |F(\zeta)| \\ &\le \max_{|\zeta|=\varepsilon_N} \sum_{k\in\mathbb{Z}} \|\delta_k(t,N,\zeta)\| \cdot \varepsilon_N^{-(N+1)} \le \text{Const}\cdot M\cdot \varepsilon_N^{-(N+1)}.\end{aligned}$$

Choosing now $u_k = \delta_k(t,N,\omega^{-1})/\|\delta_k(t,N,\omega^{-1})\|$ in the definition of $F(\zeta)$ and letting $m \to \infty$ gives

$$\sum_{k\in\mathbb{Z}} \|\delta_k(t,N,\omega^{-1})\| \le \text{Const}\cdot M\cdot (\omega\varepsilon_N)^{-(N+1)}.$$

For ω and N related by (3.3) we have $(\omega\varepsilon_N)^{-1} \le 2/\mathrm{e} = \mathrm{e}^{-\alpha}$ with $\alpha = 1 - \ln 2 > 0$, so that in this case

$$\sum_{k\in\mathbb{Z}} \|\delta_k(t)\| \le \text{Const}\cdot M\cdot \mathrm{e}^{-\alpha(N+1)} \le \text{Const}\cdot M\cdot \mathrm{e}^{-\gamma\omega}$$

holds with the exponent $\gamma = \frac{\alpha R}{\nu M \mathrm{e}}$. □

2.4 The Hamiltonian case

Sections 2.2 and 2.3 treated general second order differential equations with rapid oscillations. Our main interest is in Hamiltonian systems, where $g(x) = -\nabla U(x)$ and $U(x)$ is an analytic potential. The Hamiltonian $H(x, \dot{x})$ of the system (2.1) is then given by (1.2).

2.4.1 Hamiltonian of the modulated Fourier expansion

It is interesting to note that the Hamiltonian structure passes over to the differential equation for the coefficients of the modulated Fourier expansion. To see this, we let

$$\mathbf{y} = (..., y^{-2}, y^{-1}, y^0, y^1, y^2, ...)$$

be a two-sided infinite sequence, and we define

$$\mathcal{U}(\mathbf{y}) = U(y^0) + \sum_{m\geq 0} \frac{1}{m!} \sum_{s(\alpha)=0, \alpha_i \neq 0} U^{(m)}(y^0)\big(y^{\alpha_1}, \dots, y^{\alpha_m}\big). \tag{4.1}$$

This function is well-defined as long as $\sum_{k\neq 0} \|y^k\| \leq R$. The system (3.10) then becomes

$$\ddot{y}^k + \Omega^2 y^k \;=\; -\nabla_{y^{-k}} \mathcal{U}(\mathbf{y}) \tag{4.2}$$

and is Hamiltonian with

$$\mathcal{H}(\mathbf{y}, \dot{\mathbf{y}}) \;=\; \frac{1}{2} \sum_{k\in\mathbb{Z}} \Big((\dot{y}^{-k})^T \dot{y}^k + (y^{-k})^T \Omega^2 y^k \Big) + \mathcal{U}(\mathbf{y}). \tag{4.3}$$

2.4.2 An almost-invariant close to the oscillatory energy

It turns out that, besides the Hamiltonian $\mathcal{H}(\mathbf{y}, \dot{\mathbf{y}})$ (see [HL00]), the system (4.2) also has

$$\mathcal{I}(\mathbf{y}, \dot{\mathbf{y}}) \;=\; -\mathrm{i}\,\omega \sum_{k\neq 0} k (y^{-k})^T \dot{y}^k \tag{4.4}$$

as a conserved quantity. This series converges if $\sum_{k\neq 0} |k| \, \|y^k\| < \infty$ and $\max_{k\neq 0} \|\dot{y}^k\| < \infty$. For the functions $y^k(t)$ of (3.9), where $y_i(t)$ and $z_i^k(t)$ are the solution of the truncated system (3.1)–(3.2), this is a consequence of (3.4).

We shall prove here that the expression $\mathcal{I}\big(\mathbf{y}(t), \dot{\mathbf{y}}(t)\big)$ is conserved up to exponentially small terms. Moreover it turns out that this expression is close to the oscillatory energy

$$I(x, \dot{x}) = \frac{1}{2} \|\dot{x}_2\|^2 + \frac{\omega^2}{2} \|x_2\|^2 \tag{4.5}$$

of the system (2.1) with $g(x) = -\nabla U(x)$.

Theorem 2.4.1 *Let $\mathbf{y}(t)$ be the infinite vector with components $y^k(t)$ given by (3.9) and corresponding to initial values given by Lemma 2.3.1. Under the assumption of Theorem 2.3.2 we then have*

$$\begin{aligned}\mathcal{I}\big(\mathbf{y}(t),\dot{\mathbf{y}}(t)\big) &= \mathcal{I}\big(\mathbf{y}(0),\dot{\mathbf{y}}(0)\big) + \mathcal{O}(\mathrm{e}^{-\gamma\omega})\\ \mathcal{I}\big(\mathbf{y}(t),\dot{\mathbf{y}}(t)\big) &= I\big(x(t),\dot{x}(t)\big) + \mathcal{O}(\omega^{-1})\end{aligned}$$

for $0 \le t \le T$ and $\omega \ge \omega_0$, where the constants symbolizing the $\mathcal{O}(\cdot)$ depend on E, M, and R, but not on ω.

Proof. We use the algebraic identity

$$\sum_{k\neq 0} \mathrm{i}\, k\, (y^k)^T \nabla \mathcal{U}_{y^k}(\mathbf{y}) \;=\; 0, \tag{4.6}$$

which holds for $\sum_{k\neq 0} |k|\, \|y^k\| < \infty$. For a proof we refer to [HL00] and [HLW02, Sect. XIII.6.2].

We then compute the time-derivative of $\mathcal{I}\big(\mathbf{y}(t),\dot{\mathbf{y}}(t)\big)$ with $y(t)$ of (3.9):

$$\begin{aligned}\frac{d}{dt}\mathcal{I}\big(\mathbf{y}(t),\dot{\mathbf{y}}(t)\big) &= -\mathrm{i}\,\omega \sum_{k\neq 0} k\, \dot{y}^{-k}(t)^T \dot{y}^k(t) - \mathrm{i}\,\omega \sum_{k\neq 0} k\, y^{-k}(t)^T \ddot{y}^k(t)\\ &= -\mathrm{i}\,\omega \sum_{k\neq 0} k\, y^{-k}(t)^T \Big(\ddot{y}^k(t) + \Omega^2 y^k(t) + \nabla \mathcal{U}_{y^{-k}}\big(\mathbf{y}(t)\big)\Big)\\ &= -\mathrm{i}\,\omega \sum_{k\neq 0} k\, y^{-k}(t)^T \delta_k(t).\end{aligned}$$

We have used that the terms $k\,(\dot{y}^{-k})^T \dot{y}^k$ as well as $k\,(y^{-k})^T \Omega^2 y^k$ cancel with the corresponding terms for $-k$. Furthermore, we have added the expression (4.6) to let the defect appear in the right-hand expression. The first statement now follows from Theorem 2.3.2, and by an integration on the interval $[0,t]$.

The second statement is obtained as in the proof of Theorem 4.3 in [HL00]. □

2.4.3 Proof of Theorem 2.1.1

To prove the main theorem of this chapter, which states that (4.5) is nearly conserved over exponentially long time, we only have to use Theorem 2.4.1 and change the $\mathcal{O}(\omega^{-N})$ remainders by $\mathcal{O}(\mathrm{e}^{-\gamma\omega})$ in the proof of Corollary 4.4 in [HL00].

Chapter 3

Numerical methods

In this chapter we develop some numerical methods to solve oscillatory differential equations of the form (1.1) encountered in Chapter 2. We first give the general numerical method, then in Section 3.2 we discuss some geometric properties of this method. We then analyse our numerical methods in Section 3.3. Numerical comparisons and examples are left for the last section of this chapter.

3.1 The general method

Let's first recall the problem that we want to solve; it is a second-order differential equation of the form

$$\ddot{x} + \Omega^2 x = g(x) \qquad \text{with} \qquad \Omega = \begin{pmatrix} 0 & 0 \\ 0 & \omega I \end{pmatrix}, \ \omega \gg 1, \tag{1.1}$$

and with initial values satisfying

$$\frac{1}{2}\Big(\|\dot{x}(0)\|^2 + \|\Omega x(0)\|^2\Big) \ \leq \ E, \tag{1.2}$$

where E is independent of ω. We do not require E to be small. As usual, the block matrices in Ω are of arbitrary dimension.

In the preceding chapter, we have seen that the solution of (1.1) admits an expansion of the following form

$$x(t) = y(t) + \sum_{k \neq 0} \mathrm{e}^{\mathrm{i}k\omega t} z^k(t). \tag{1.3}$$

The numerical methods proposed in this chapter are based on approximations of the first two terms of this expansion. This seems to be reasonable by the bounds obtained in Theorem 5.1 of [HLW02, Sect. XIII.5.1]. Indeed, for $|k| \geq 2$, the functions z^k are of order $\mathcal{O}(\omega^{-k-2})$. Thus, we search for a smooth real valued function y and a smooth complex valued function $z := z^1$ such that

$$x_*(t) = y(t) + \mathrm{e}^{\mathrm{i}\omega t} z(t) + \mathrm{e}^{-\mathrm{i}\omega t} \bar{z}(t) \tag{1.4}$$

is a good approximation of the solution of (1.1) (see [HLW02, Sect. XIII.3.1]). How do we find the functions in (1.4)? We insert x_* into the differential equation (1.1) and expand the function g around y.

$$\begin{aligned}\begin{pmatrix} \ddot{y}_1 \\ \ddot{y}_2 + \omega^2 y_2 \end{pmatrix} &+ \mathrm{e}^{\mathrm{i}\omega t} \begin{pmatrix} -\omega^2 z_1 + 2\mathrm{i}\omega \dot{z}_1 + \ddot{z}_1 \\ 2\mathrm{i}\omega \dot{z}_2 + \ddot{z}_2 \end{pmatrix} + \mathrm{e}^{-\mathrm{i}\omega t} \begin{pmatrix} -\omega^2 \bar{z}_1 - 2\mathrm{i}\omega \dot{\bar{z}}_1 + \ddot{\bar{z}}_1 \\ -2\mathrm{i}\omega \dot{\bar{z}}_2 + \ddot{\bar{z}}_2 \end{pmatrix} \\ &= g(y) + \mathrm{e}^{\mathrm{i}\omega t} g'(y) z + \mathrm{e}^{-\mathrm{i}\omega t} g'(y) \bar{z} + g''(y)(z, \bar{z}) + \dots \end{aligned} \tag{1.5}$$

We now compare the coefficients of $1, \mathrm{e}^{\mathrm{i}\omega t}, \mathrm{e}^{-\mathrm{i}\omega t}$, solve for the term with highest power of ω, and remove terms of order $\mathcal{O}(\omega^{-2})$ on the left of the equality. Depending and how we truncate the Taylor series on the right of the equality, we obtain the following two systems that determine the functions y and z.

- $$\begin{aligned} \ddot{y}_1 &= g_1(y_1, y_2) + g_1''(y_1, y_2)(z, \bar{z}) \\ \dot{z}_2 &= -\tfrac{\mathrm{i}}{2\omega} g_2'(y_1, y_2) z \\ y_2 &= \tfrac{1}{\omega^2}(g_2(y_1, y_2) + g_2''(y_1, y_2)(z, \bar{z})) \\ z_1 &= -\tfrac{1}{\omega^2} g_1'(y_1, y_2) z. \end{aligned} \tag{1.6}$$

- Neglecting further the second order term in the Taylor series for g_2, we get

$$\begin{aligned} \ddot{y}_1 &= g_1(y_1, y_2) + g_1''(y_1, y_2)(z, \bar{z}) \\ \dot{z}_2 &= -\tfrac{\mathrm{i}}{2\omega} g_2'(y_1, y_2) z \\ y_2 &= \tfrac{1}{\omega^2} g_2(y_1, y_2) \\ z_1 &= -\tfrac{1}{\omega^2} g_1'(y_1, y_2) z. \end{aligned} \tag{1.7}$$

Methods 3 and 4 (see below) are designed to solve these systems.

Initial values. The initial values for the differential equation (1.1) permit us to find initial values for the system of differential equations in (1.6) or (1.7). In fact, we want $x_*(0) = x(0)$ and $\dot{x}_*(0) = \dot{x}(0)$. This means that we have to solve

$$\begin{aligned} x_1(0) &= y_1(0) + z_1(0) + \bar{z}_1(0) \\ x_2(0) &= y_2(0) + z_2(0) + \bar{z}_2(0) \\ \dot{x}_1(0) &= \dot{y}_1(0) + \mathrm{i}\omega(z_1(0) - \bar{z}_1(0)) + \dot{z}_1(0) + \dot{\bar{z}}_1(0) \\ \dot{x}_2(0) &= \dot{y}_2(0) + \mathrm{i}\omega(z_2(0) - \bar{z}_2(0)) + \dot{z}_2(0) + \dot{\bar{z}}_2(0), \end{aligned}$$

which can be reformulated as

$$\begin{aligned} y_1(0) &= x_1(0) - 2z_{1r}(0) \\ \dot{y}_1(0) &= \dot{x}_1(0) + 2\omega z_{1i}(0) - 2\dot{z}_{1r}(0) \\ z_{2r}(0) &= \tfrac{x_2(0) - y_2(0)}{2} \\ z_{2i}(0) &= \tfrac{\dot{y}_2(0) + 2\dot{z}_{2r}(0) - \dot{x}_2(0)}{2\omega}, \end{aligned} \tag{1.8}$$

where z_{2r} and z_{2i} stand for the real and imaginary part of z_2 (same definition for z_1). This system can now be solved by fixed point iterations to yield $y_1(0), \dot{y}_1(0), z_{2r}(0), z_{2i}(0)$, the

initial values required for systems (1.6) or (1.7). Here, we use the algebraic relations for y_2 and z_1 from (1.6) or (1.7) and the fact that if ω is sufficiently large, the iterations converge.

Defect. As long as $z_2(t) = \mathcal{O}(\omega^{-1})$, on a bounded interval, from (1.6) and (1.7), we obtain the following bounds

$$y_2(t) = \mathcal{O}(\omega^{-2}),\ z_1(t) = \mathcal{O}(\omega^{-3}),\ \dot{z}_2(t) = \mathcal{O}(\omega^{-2}). \tag{1.9}$$

Inserting (1.4), with y and z given either by (1.6) or (1.7), into the second order differential equation (1.1), the defect

$$d(t) = \ddot{x}_*(t) + \Omega^2 x_*(t) - g(x_*(t))$$

is seen to be of size $\mathcal{O}(\omega^{-2})$.

Error. The Fundamental Lemma (see for example [Car77, p.116]) shows that, on a bounded interval, the error $x(t) - x_*(t)$ is of the same magnitude as the defect.

Going back to (1.5), we try to simplify even more the systems that determine the functions y and z. We put $y_2 = z_1 = 0$ and obtain the following two systems

- $$\begin{array}{rcl} \ddot{y}_1 &=& g_1(y_1,0) + g_1''(y_1,0)(z,\bar{z}) \\ \dot{z}_2 &=& -\frac{\mathrm{i}}{2\omega} D_2 g_2(y_1,0) z_2. \end{array} \tag{1.10}$$

- Neglecting the term of size $\mathcal{O}(\omega^{-2})$ in the Taylor series for g_1, we get

$$\begin{array}{rcl} \ddot{y}_1 &=& g_1(y_1,0) \\ \dot{z}_2 &=& -\frac{\mathrm{i}}{2\omega} D_2 g_2(y_1,0) z_2. \end{array} \tag{1.11}$$

Methods 1 and 2 (see below) are designed to solve these systems.

Initial values. Putting $z_1 = y_2 = 0$ in (1.8), the initial values for (1.10) or (1.11) are given by

$$\begin{array}{rcl} y_1(0) &=& x_1(0) \\ \dot{y}_1(0) &=& \dot{x}_1(0) \\ z_{2r}(0) &=& \frac{x_2(0)}{2} \\ z_{2i}(0) &=& \frac{2\dot{z}_{2r}(0) - \dot{x}_2(0)}{2\omega}, \end{array} \tag{1.12}$$

where $\dot{z}_{2r}(0)$ is given by the differential equation (1.10) or (1.11).

Defect. Unfortunatly, this time, the defect is of size $\mathcal{O}(1)$. This is because its second component contains the term $g_2(y)$.

Error. However, looking at system (4.2) in Chapter 2 that determine the modulated functions, we can show that

$$x_1(t) - x_{*1}(t) = y_1(t) - \tilde{y}_1(t) + \mathcal{O}(\omega^{-2}) = \mathcal{O}(\omega^{-2}),$$

where $y_1(t)$ is a term of the modulated expansion $x_1(t)$ and $\tilde{y}_1(t)$ is a solution of (1.10) or (1.11). For the second component of the error, we obtain

$$x_2(t) - x_{*2}(t) = \mathrm{e}^{\mathrm{i}\omega t}(z_2(t) - \tilde{z}_2(t)) + \mathrm{e}^{-\mathrm{i}\omega t}(\bar{z}_2(t) - \bar{\tilde{z}}_2) + \mathcal{O}(\omega^{-2}) = \mathcal{O}(\omega^{-2}).$$

We thus get $x(t) - x_*(t) = \mathcal{O}(\omega^{-2})$.

All the systems (1.6), (1.7), (1.10) and (1.11) can be written in the following form:

$$\begin{array}{rcl} \ddot{y}_1 &=& f_1(y,z) \\ \dot{z}_2 &=& f_2(y)z \\ y_2 &=& \omega^{-2} f_3(y,z) \\ z_1 &=& \omega^{-2} f_4(y)z. \end{array} \tag{1.13}$$

This system consists of one second-order differential equation, one first-order differential equation and two algebraic equations. We remark that for sufficiently large ω, we can use the Implicit Function Theorem for the algebraic equations of problem (1.13), and show that y_2 and z_1 are functions of y_1 and z_2. Namely, $y_2 = \omega^{-2}a(y_1, z_2)$ and $z_1 = \omega^{-2}b(y_1, z_2)z_2$.

We point out that this technique was also used in [MvV78], but no attention was paid to energy conservation.

To solve the system (1.13), we first have to find initial values (this was done above), then we solve the second-order differential equation with a Störmer-Verlet's type method and the first-order differential equation with a method of the kind of the midpoint scheme, within that we solve the algebraic equations, and finally we give an approximation of the solution of (1.1) with (1.4).

The numerical method gives, for a step size h and approximations v_1^n, y_1^n, z_2^n of $\dot{y}_1(nh)$, $y_1(nh), z_2(nh)$, and y_2^n,z_1^n satisfying $y_2^n = \omega^{-2}f_3(y^n, z^n)$, resp. $z_1^n = \omega^{-2}f_4(y^n)z^n$,

$$\begin{array}{rcl} v_1^{n+1/2} &=& v_1^n + \frac{h}{2} f_1(y^n, z^n) \\ y_1^{n+1} &=& y_1^n + h v_1^{n+1/2} \\ y_2^{n+1} &=& \omega^{-2} f_3(y^{n+1}, z^{n+1}) \\ z_1^{n+1} &=& \omega^{-2} f_4(y^{n+1}) z^{n+1} \\ z_2^{n+1} &=& z_2^n + h f_2(\frac{y^n + y^{n+1}}{2})(\frac{z^n + z^{n+1}}{2}) \\ v_1^{n+1} &=& v_1^{n+1/2} + \frac{h}{2} f_1(y^{n+1}, z^{n+1}). \end{array} \tag{1.14}$$

With these values, we can now give an approximation of the solution of (1.1). Indeed, the approximation of the first terms of the modulated Fourier expansion is:

$$\begin{array}{rcl} x^n &=& y^n + \mathrm{e}^{\mathrm{i}\omega t_n} z^n + \mathrm{e}^{-\mathrm{i}\omega t_n} \bar{z}^n \\ \dot{x}^n &=& \dot{y}^n + \mathrm{e}^{\mathrm{i}\omega t_n}(\mathrm{i}\omega z^n + \dot{z}^n) + \mathrm{e}^{-\mathrm{i}\omega t_n}(-\mathrm{i}\omega \bar{z}^n + \dot{\bar{z}}^n), \end{array} \tag{1.15}$$

where $t_n = nh$, $\dot{y}_1^n = v_1^n$, and $\dot{z}_2^n$ is given by (1.13). The remaining values $\dot{y}_2^n$ and $\dot{z}_1^n$ are computed depending on the numerical methods considered. For the two first methods (see below), designed to solve (1.11) and (1.10), we put $\dot{y}_2^n = \dot{z}_1^n = 0$ along the numerical solution. For the third and fourth methods, which solves (1.7) and (1.6), we solve the following system

$$\dot{y}_2 = \omega^{-2}\frac{\partial f_3}{\partial y}(y,z)\dot{y}, \qquad \dot{z}_1 = \omega^{-2}\Big(\frac{\partial f_4}{\partial y}(y)(\dot{y}, z) + f_4(y)\dot{z}\Big).$$

3.2 Numerical properties

In this section, we analyse geometric properties of the numerical method given in the first section and of the flow of the differential equations contained in (1.13). We first consider the symmetry of the numerical method and then we look at the reversibility property of the problem.

3.2.1 Symmetry

The general method (1.14) is symmetric. To see this, we exchange $n \leftrightarrow n+1$ and $h \leftrightarrow -h$. The method is then symmetric if and only if

$$\begin{aligned} v_1^{n+1/2} &= v_1^{n+1} - \tfrac{h}{2} f_1(y^{n+1}, z^{n+1}) \\ y_1^n &= y_1^{n+1} - h v_1^{n+1/2} \\ z_2^n &= z_2^{n+1} - h f_2(\tfrac{y^{n+1}+y^n}{2})(\tfrac{z^{n+1}+z^n}{2}) \\ v_1^n &= v_1^{n+1/2} - \tfrac{h}{2} f_1(y^n, z^n), \end{aligned}$$

this is the case by definition of the numerical method (1.14).

The method (1.14) is consistent with problem (1.13) and is symmetric, thus it has order 2.

3.2.2 ρ-reversibility

As shown in the preceding section, we know that y_2 and z_1 are functions of y_1 and z_2. We thus rewrite the differential equations in (1.13) as a system of differential equations of order one:

$$\begin{aligned} \dot{y}_1 &= v_1 \\ \dot{v}_1 &= \tilde{f}_1(y_1, z_2) \\ \dot{z}_2 &= \tilde{f}_2(y_1, z_2) z_2, \end{aligned} \tag{2.1}$$

where $\tilde{f}_1$ is real-valued and $\tilde{f}_2$ is complex-valued function defined by

$$\tilde{f}_1(y_1, z_2) = f_1(y_1, \omega^{-2} a(y_1, z_2), \omega^{-2} b(y_1, z_2) z_2, z_2)$$

and

$$\tilde{f}_2(y_1, z_2) z_2 = f_2(y_1, \omega^{-2} a(y_1, z_2))(\omega^{-2} b(y_1, z_2) z_2, z_2).$$

Remembering (see Chapter 1) the definitions of ρ-reversibility for a problem $\dot{y} = f(y)$ and for a numerical method Φ_h, we have the following proposition:

Proposition 3.2.1 *If $\tilde{f}_1(y_1, z_2) = \tilde{f}_1(y_1, \bar{z}_2)$ and $\bar{\tilde{f}}_2(y_1, z_2) = -\tilde{f}_2(y_1, \bar{z}_2)$, the problem (2.1) is ρ-reversible for the map $\rho_1(Y) = (y_1, -v_1, \bar{z}_2)$, where $Y = (y_1, v_1, z_2)$. Here, one should interpret the third component of vector $\rho_1(Y)$ in terms of real and imaginary part of z_2. Thus one obtains a linear map $\rho(Y) = (y_1, -v_1, z_{2r}, -z_{2i})$.*

We also have ρ-reversibility for $\rho_2(Y) = (y_1, -v_1, -\bar{z}_2)$ if $\tilde{f}_1(y_1, z_2) = \tilde{f}_1(y_1, -\bar{z}_2)$ and $\tilde{f}_2(y_1, -\bar{z}_2) = -\bar{\tilde{f}}_2(y_1, z_2)$.

Systems (1.6), (1.7), (1.10), and (1.11) are ρ-reversible for these two maps. As an example, we show that the conditions $\tilde{f}_1(y_1, z_2) = \tilde{f}_1(y_1, \bar{z}_2)$ and $\bar{\tilde{f}}_2(y_1, z_2) = -\tilde{f}_2(y_1, \bar{z}_2)$ are fulfilled for problem (1.11), thus it is ρ_1-reversible. Indeed, we have $\tilde{f}_1(y_1, z_2) = g_1(y_1, 0) = \tilde{f}_1(y_1, \bar{z}_2)$ and $\bar{\tilde{f}}_2(y_1, z_2) = \frac{\mathrm{i}}{2\omega} D_2 g_2(y_1, 0) = -\tilde{f}_2(y_1, \bar{z}_2)$.

Proof: We prove the proposition only for the first application ρ_1. We first have to compute

$$\rho_1(F(Y)) \;=\; \begin{pmatrix} v_1 \\ -\tilde{f}_1(y_1, z_2) \\ \bar{\tilde{f}}_2(y_1, z_2)\bar{z}_2 \end{pmatrix},$$

where $F(Y) = (v_1, \tilde{f}_1(y_1, z_2), \tilde{f}_2(y_1, z_2)z_2)^T$ of (2.1). We have ρ-reversibility if this quantity is equal to

$$-F(\rho_1(Y)) \;=\; -F(y_1, -v_1, \bar{z}_2) \;=\; \begin{pmatrix} v_1 \\ -\tilde{f}_1(y_1, \bar{z}_2) \\ -\tilde{f}_2(y_1, \bar{z}_2)\bar{z}_2 \end{pmatrix}.$$

Using the hypothesis on functions $\tilde{f}_1$ and $\tilde{f}_2$ permits us to show that the ρ-reversibility condition is verified and this concludes the proof of the proposition. □

By Theorem 1.5. in [HLW02, Sect. V.1], for a symmetric method Φ_h, it is sufficient to prove the ρ-compatibility condition $\rho \circ \Phi_h = \Phi_{-h} \circ \rho$ to show ρ-reversibility. This permits us to prove that the numerical method (1.14) is ρ-reversible for the two applications defined in the last proposition.

Proposition 3.2.2 *The numerical method (1.14) applied to problem (2.1) is ρ-reversible for the two applications given in Proposition 3.2.1.*

Proof: We just show the condition $\rho \circ \Phi_h = \Phi_{-h} \circ \rho$ for ρ_1, the proof for ρ_2 is very similar. Under the conditions on the functions $\tilde{f}_1$ and $\tilde{f}_2$ of the last proposition, this follows from the fact that replacing h, v_1^n, z_2^n by $-h, -v_1^n, \bar{z}_2^n$ yields $y^{n+1}, -v_1^{n+1}, \bar{z}_2^{n+1}$ as numerical solution. □

3.3 Four numerical methods

In this section we define and analyse four numerical methods based on systems (1.10), (1.11) and (1.6), (1.7). They are presented from the easiest to compute to the more complicated.

3.3.1 Method 1

Let's recall the problem that we solve; for the first method, we fix $y_2 = z_1 = \dot{y}_2 = \dot{z}_1 = 0$ (which is a reasonable first approximation of these quantities) and we solve the following

differential equations (see (1.11))

$$\begin{aligned} \ddot{y}_1 &= g_1(y_1, 0), \\ \dot{z}_2 &= -\tfrac{\mathrm{i}}{2\omega} D_2 g_2(y_1, 0) z_2. \end{aligned} \tag{3.1}$$

As explained in Section 3.1, Method 1 reads

$$\begin{aligned} v_1^{n+1/2} &= v_1^n + \frac{h}{2} g_1(y_1^n, 0) \\ y_1^{n+1} &= y_1^n + h v_1^{n+1/2} \\ z_2^{n+1} &= z_2^n - h\frac{\mathrm{i}}{2\omega} D_2 g_2(\frac{y_1^n + y_1^{n+1}}{2}, 0)(\frac{z_2^n + z_2^{n+1}}{2}) \\ v_1^{n+1} &= v_1^{n+1/2} + \frac{h}{2} g_1(y_1^{n+1}, 0). \end{aligned} \tag{3.2}$$

The computaion of $v_1^{n+1/2}, y_2^{n+1}$, and v_1^{n+1} is explicit, that of z_2^{n+1} is linearly implicit.

Let's first recall that this problem and the numerical method (3.2) associated to it are ρ-reversible (for a proof, see Section 3.2).

Lemma 3.3.1 *The problem (3.1) and the numerical method (3.2) are ρ-reversible for the two applications defined in Proposition 3.2.1.*

What can we show for the conservation of the Hamiltonian H and of the oscillatory energy I if the function $g(x)$ is a smooth gradient $g(x) = -\nabla U(x)$?

First of all, the second-order differential equation in (3.1), namely,

$$\begin{aligned} \dot{y}_1 &= v_1, \\ \dot{v}_1 &= g_1(y_1, 0) = -\nabla U(y_1, 0), \end{aligned}$$

is a Hamiltonian system with $\hat{H}(y_1, v_1) = \frac{1}{2} v_1^T v_1 + U(y_1, 0)$. The first part of the method reduces to the Störmer-Verlet scheme, which is symplectic, so that we have for the numerical solution $\hat{H}(y_1^n, v_1^n) = \hat{H}(y_1^0, v_1^0) + \mathcal{O}(h^2)$ for $nh \leq T$ on exponentially long time intervals T.

Writing the first order differential equation of (3.1) as a real part and a complex part, we get

$$\begin{pmatrix} \dot{z}_{2R} \\ \dot{z}_{2I} \end{pmatrix} = A(y_1) \begin{pmatrix} z_{2R} \\ z_{2I} \end{pmatrix},$$

with a real skew-symmetric matrix $A(y_1)$. Therefore, the following quadratic quantity $\hat{I}(z_2) = z_{2R}^T z_{2R} + z_{2I}^T z_{2I} = z_2^T \bar{z}_2$ is an invariant for the problem considered. Our method preserves this invariant. Indeed, the method reads $\mathrm{z}^1 = \mathrm{z}^0 + \frac{h}{2} A\mathrm{z}^0 + \frac{h}{2} A\mathrm{z}^1$ where $\mathrm{z} = (z_{2R}, z_{2I})$ (we do not write the argument in the matrix A). We thus have

$$\begin{aligned} \hat{I}(z_2^1) = \mathrm{z}^{1T}\mathrm{z}^1 &= \mathrm{z}^{1T}\mathrm{z}^0 + \frac{h}{2}\mathrm{z}^{1T} A\mathrm{z}^0 + \frac{h}{2}\mathrm{z}^{1T} A\mathrm{z}^1 \\ &= \mathrm{z}^{1T}\mathrm{z}^0 + \frac{h}{2}\mathrm{z}^{1T} A\mathrm{z}^0 = \mathrm{z}^{0T}\mathrm{z}^0 + \frac{h}{2}\mathrm{z}^{0T} A\mathrm{z}^1 + \frac{h}{2}\mathrm{z}^{1T} A\mathrm{z}^0 = \mathrm{z}^{0T}\mathrm{z}^0 = \hat{I}(z_2^0), \end{aligned}$$

where we have used the skew-symmetric structure of matrix A and the fact that $\hat{I}$ is an invariant. This is exactly the proof that the mid-point rule preserves quadratic first integrals.

Let's come back to the Hamiltonian $H(x,\dot{x}) = \frac{1}{2}\dot{x}^T\dot{x} + \frac{1}{2}x^T\Omega^2 x + U(x)$ and to the oscillatory energy $I(x,\dot{x}) = \frac{1}{2}||\dot{x}_2||^2 + \frac{\omega^2}{2}||x_2||^2$ of problem (1.1). The next proposition tells something more about the almost preservation of these quantities.

Proposition 3.3.2 *The numerical method (3.2) applied to problem (3.1) yields for the second-order differential equation (1.1) with initial values satisfying (1.2),*

$$\begin{aligned} H(x^n,\dot{x}^n) &= H(x(0),\dot{x}(0)) + \mathcal{O}(\omega^{-1}) + \mathcal{O}(h^2), \\ I(x^n,\dot{x}^n) &= I(x(0),\dot{x}(0)) + \mathcal{O}(\omega^{-2}), \end{aligned}$$

on exponentially long time intervals.

Proof: Let's start the proof by computing $H(x^n,\dot{x}^n)$, we have:

$$\begin{aligned} H(x^n,\dot{x}^n) &= \tfrac{1}{2}(v_1^n)^T(v_1^n) + 2\omega^2(z_2^n)^T(z_2^n) + U(y_1,0) + \mathcal{O}(\omega^{-1}) \\ &= \hat{H}(y_1^n,v_1^n) + 2\omega^2\hat{I}(z_2^n) + \mathcal{O}(\omega^{-1}), \end{aligned}$$

if $z_2(0) = \mathcal{O}(\omega^{-1})$. Using the symplecticity of the method, we get $\hat{H}(y_1^n,v_1^n) = \hat{H}(y_1^0,v_1^0) + \mathcal{O}(h^2)$ for $nh \leq T$ on exponentially long time intervals T, thus

$$\begin{aligned} H(x^n,\dot{x}^n) &= \hat{H}(y_1(0),v_1(0)) + \mathcal{O}(h^2) + 2\omega^2\hat{I}(z_2(0)) + \mathcal{O}(\omega^{-1}) \\ &= H(x(0),\dot{x}(0)) + \mathcal{O}(\omega^{-1}) + \mathcal{O}(h^2). \end{aligned}$$

For the oscillatory energy, we obtain

$$\begin{aligned} I(x^n,\dot{x}^n) &= 2\omega^2\hat{I}(z_2^n) + \mathcal{O}(\omega^{-2}) = 2\omega^2\hat{I}(z_2(0)) + \mathcal{O}(\omega^{-2}) \\ &= I(x(0),\dot{x}(0)) + \mathcal{O}(\omega^{-2}). \end{aligned}$$

□

3.3.2 Method 2

As for the first method proposed, we fix $y_2 = z_1 = \dot{y}_2 = \dot{z}_1 = 0$. This time, we go further in the computation of the Taylor series of function g. We solve the system, see (1.10),

$$\begin{aligned} \ddot{y}_1 &= g_1(y_1,0) + g_1''(y_1,0)(z,\bar{z}) \\ \dot{z}_2 &= -\tfrac{\mathrm{i}}{2\omega}D_2g_2(y_1,0)z_2. \end{aligned} \tag{3.3}$$

The numerical scheme reads

$$\begin{aligned} v_1^{n+1/2} &= v_1^n + \frac{h}{2}(g_1(y_1^n,0) + g_1''(y_1^n,0)(z^n,\bar{z}^n)) \\ y_1^{n+1} &= y_1^n + hv_1^{n+1/2} \\ z_2^{n+1} &= z_2^n - h\frac{\mathrm{i}}{2\omega}D_2g_2(\frac{y_1^n+y_1^{n+1}}{2},0)(\frac{z_2^n+z_2^{n+1}}{2}) \\ v_1^{n+1} &= v_1^{n+1/2} + \frac{h}{2}(g_1(y_1^{n+1},0) + g_1''(y_1^{n+1},0)(z^{n+1},\bar{z}^{n+1})). \end{aligned} \tag{3.4}$$

The numerical method is also ρ-reversible (see the first method for a proof), the quantity $\hat{I}$ is still an invariant of problem (3.3) and we also have $\hat{I}(z_2^n) = \hat{I}(z_2^0)$ for the numerical solution. Moreover, the same proof for the near conservation of I (see Proposition 3.3.2) can be applied to yield:

Proposition 3.3.3 *The numerical method (3.4) applied to problem (3.3) yields for the second-order differential equation (1.1) with initial values satisfying (1.2),*

$$I(x^n, \dot{x}^n) \;=\; I(x(0), \dot{x}(0)) + \mathcal{O}(\omega^{-2}). \tag{3.5}$$

3.3.3 Method 3

We now add the algebraic equations for the variables y_2, z_1. Thus, the third method consists of solving

$$\begin{array}{rcl} \ddot{y}_1 &=& g_1(y_1, y_2) + g_1''(y_1, y_2)(z, \bar{z}) \\ \dot{z}_2 &=& -\frac{\mathrm{i}}{2\omega} g_2'(y_1, y_2) z \\ y_2 &=& \frac{1}{\omega^2} g_2(y_1, y_2) \\ z_1 &=& -\frac{1}{\omega^2} g_1'(y_1, y_2) z, \end{array} \tag{3.6}$$

as explained in Section 3.1. The method reads

$$\begin{array}{rcl} v_1^{n+1/2} &=& v_1^n + \frac{h}{2}(g_1(y^n) + g_1''(y^n)(z^n, \bar{z}^n)) \\ y_1^{n+1} &=& y_1^n + h v_1^{n+1/2} \\ y_2^{n+1} &=& \omega^{-2} g_2(y^{n+1}) \\ z_1^{n+1} &=& -\omega^{-2} g_1'(y^{n+1}) z^{n+1} \\ z_2^{n+1} &=& z_2^n - h\frac{\mathrm{i}}{2\omega} g_2'(\frac{y^n + y^{n+1}}{2})(\frac{z^n + z^{n+1}}{2}) \\ v_1^{n+1} &=& v_1^{n+1/2} + \frac{h}{2}(g_1(y^{n+1}) + g_1''(y^{n+1})(z^{n+1}, \bar{z}^{n+1})). \end{array} \tag{3.7}$$

The same ideas as given in the preceding section allow us to show that problem (3.6) and the numerical method associated to it are ρ-reversible.

To conclude the analysis of the third method, we mention the result concerning the oscillatory energy:

Proposition 3.3.4 *The numerical method (3.7) applied to problem (3.6) yields for the second-order differential equation (1.1) with initial values satisfying (1.2),*

$$I(x^n, \dot{x}^n) \;=\; I(x(0), \dot{x}(0)) + \mathcal{O}(\omega^{-2}). \tag{3.8}$$

3.3.4 Method 4

The last method is very similar to the third one, but we take one term further in the equation for y_2. We consider, see (1.6),

$$\begin{array}{rcl} \ddot{y}_1 &=& g_1(y_1, y_2) + g_1''(y_1, y_2)(z, \bar{z}) \\ \dot{z}_2 &=& -\frac{\mathrm{i}}{2\omega} g_2'(y_1, y_2) z \\ y_2 &=& \frac{1}{\omega^2}(g_2(y_1, y_2) + g_2''(y_1, y_2)(z, \bar{z})) \\ z_1 &=& -\frac{1}{\omega^2} g_1'(y_1, y_2) z. \end{array} \tag{3.9}$$

Next, we solve this system using the numerical method

$$\begin{aligned} v_1^{n+1/2} &= v_1^n + \frac{h}{2}(g_1(y^n) + g_1''(y^n)(z^n, \bar{z}^n)) \\ y_1^{n+1} &= y_1^n + h v_1^{n+1/2} \\ y_2^{n+1} &= \omega^{-2}(g_2(y^{n+1}) + g_2''(y^{n+1})(z^{n+1}, \bar{z}^{n+1})) \\ z_1^{n+1} &= -\omega^{-2} g_1'(y^{n+1}) z^{n+1} \\ z_2^{n+1} &= z_2^n - h\frac{\mathrm{i}}{2\omega} g_2'(\frac{y^n + y^{n+1}}{2})(\frac{z^n + z^{n+1}}{2}) \\ v_1^{n+1} &= v_1^{n+1/2} + \frac{h}{2}(g_1(y^{n+1}) + g_1''(y^{n+1})(z^{n+1}, \bar{z}^{n+1})). \end{aligned} \tag{3.10}$$

This method is also symmetric, ρ-reversible and satisfies (3.8).

We do not have results concerning the conservation of the total energy by the last three methods. However, a look at the forthcoming numerical experiments (see Section 3.4) shows a good behaviour for these methods. This is perhaps due to the fact that all these methods are symmetric and ρ-reversible.

Two major disadvantages of these methods is that they are implicit and they require the computation of the first two derivatives of the function g in (1.1). However, as we will see in the next section, they preserve the two energies $I(x, \dot{x}) = \frac{1}{2}||\dot{x}_2||^2 + \frac{\omega^2}{2}||x_2||^2$ and $H(x, \dot{x}) = \frac{1}{2}\Big(\|\dot{x}\|^2 + \|\omega x_2\|^2\Big) + U(x)$ uniformly for all values of $h\omega$ (this was a disadvantage of the trigonometric methods).

3.4 Examples

We illustrate our methods with two examples. The first one is an FPU type problem consisting of six alternating stiff/soft springs (see Chapter 1). The datas comes from [HLW02, Sect. I.4.1] and [HLW02, Sect. XIII.3]. We first plot the solution given by DOP853 (for a definition of this numerical method, see [HNW93]) and then apply our methods with different step sizes. We make an experiment with larger ω, and a last one on long time intervals.

Let us first solve the problem on a time interval of length 220 with $\omega = 50$ using DOP853 and our methods with step sizes $h = 0.22$.

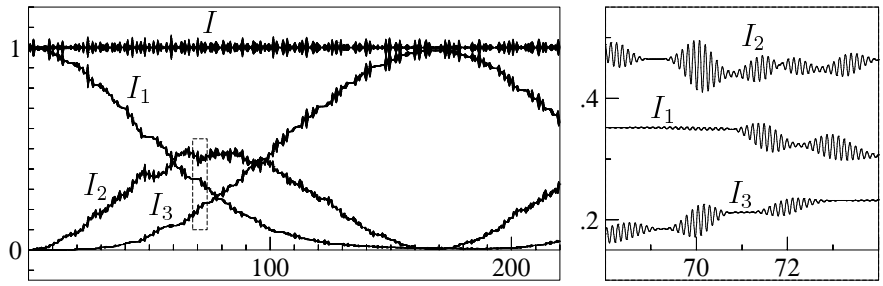


Figure 3.1: Oscillatory energies using DOP853.

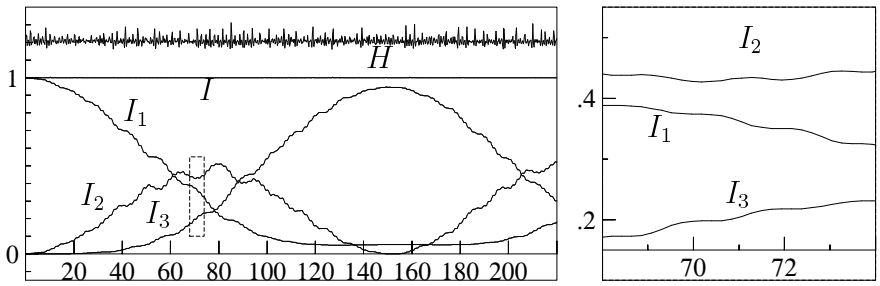


Figure 3.2: Total and oscillatory energies using Method 1 with $h = 0.22$.

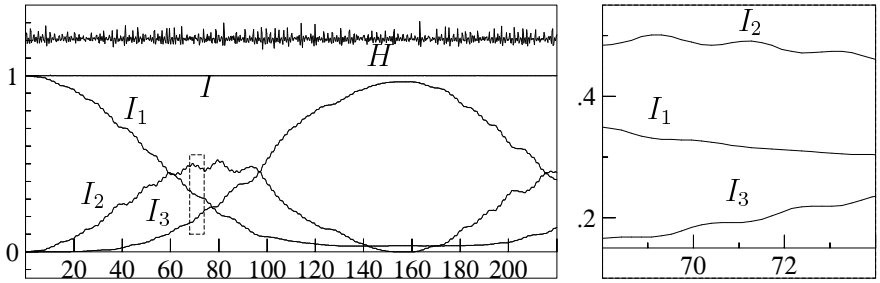


Figure 3.3: Total and oscillatory energies using Method 2 with $h = 0.22$.

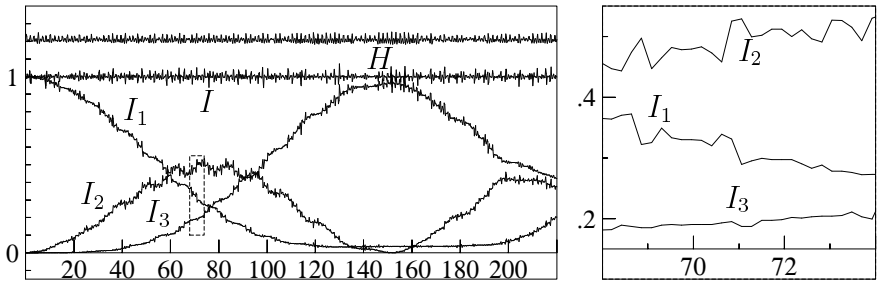


Figure 3.4: Total and oscillatory energies using Method 3 with $h = 0.22$.

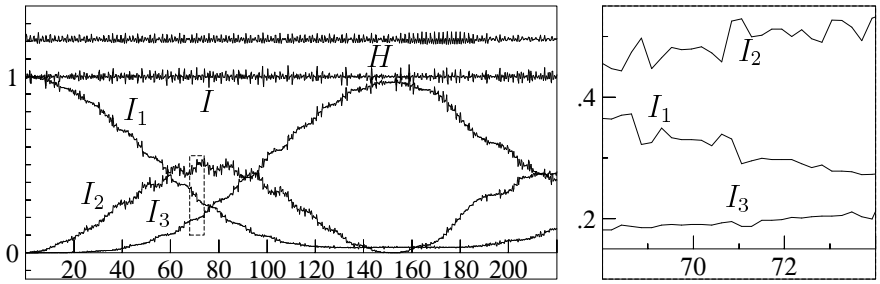


Figure 3.5: Total and oscillatory energies using Method 4 with $h = 0.22$.

We see that, as predicted in the preceding section, the oscillatory energy I is well preserved during the experiment. The energy exchange between the stiff components is quite well modelled by our numerical methods. We however see a bad behaviour in the end

of the interval of integration. This is due to the fact that the FPU problem is very sensitive to perturbations in the initial data. All numerical methods described in this section nead the computation of the initial datas by iterations. We can also see in the close-ups of the pictures that the oscillations in the components of the oscillatory energy are not present for the two first numerical solutions. The close-ups for Method 3 and 4 do not represent the oscillations, this is because we take h and ω such that $h\omega = 11$ (thus h is greater than the period of the oscillations) and link $I_j(x^n, \dot{x}^n)$ and $I_j(x^{n+1}, \dot{x}^{n+1})$ by a straight line.

Let us take a smaller step size, say $h = 0.022$, and see what we get:

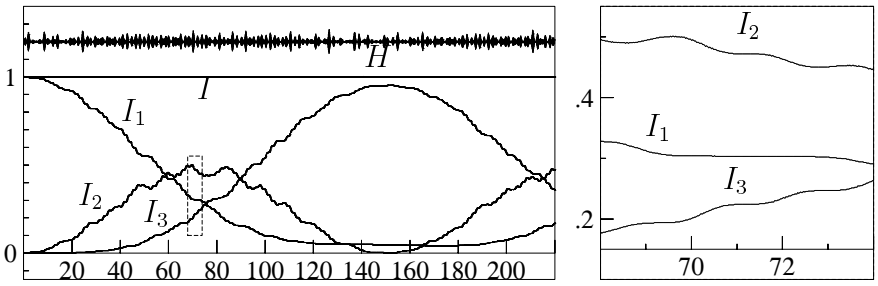


Figure 3.6: Total and oscillatory energies using Method 1 with $h = 0.022$.

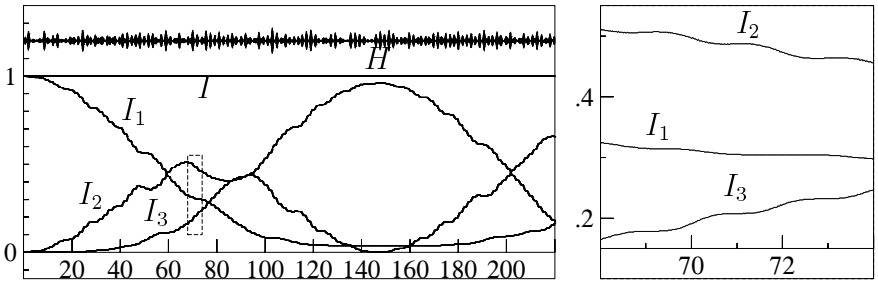


Figure 3.7: Total and oscillatory energies using Method 2 with $h = 0.022$.

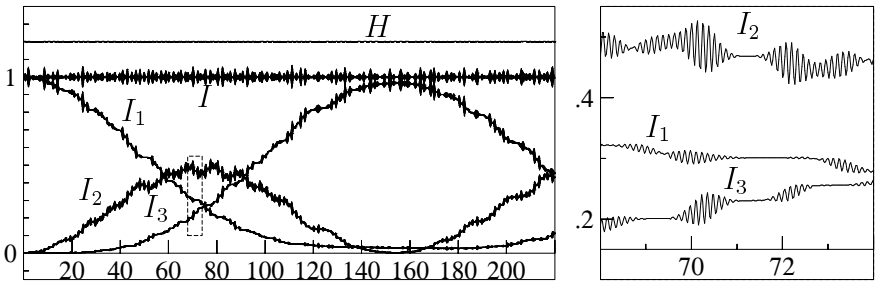


Figure 3.8: Total and oscillatory energies using Method 3 with $h = 0.022$.

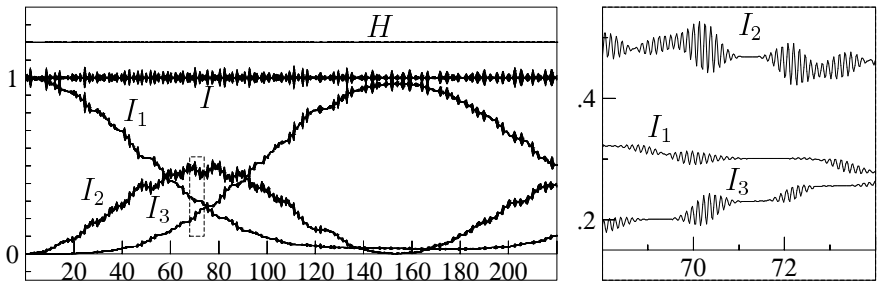


Figure 3.9: Total and oscillatory energies using Method 4 with $h = 0.022$.

This time too, the oscillatory energy is well preserved. We can also see a better behaviour in the energy exchange for the third and fourth methods. For these two methods, the oscillations in the components of I are now present and correspond quite well to the one given by DOP853 (see Figures 3.8 and 3.9).

Let's look at the numerical solution given by Deuflhard's method, for a definition, see Chapter 1. For this method, the filter functions are given by $\phi(\zeta) = 1, \psi(\zeta) = \text{sinc}(\zeta), \psi_0(\zeta) = \cos(\zeta), \psi_1(\zeta) = 1$.

It is known (see [HLW02]) that the trigonometric methods of Chapter 1, don't conserve H and I uniformly for all values of $h\omega$ (bad energy conservations for multiples of π). But what about Method 3? Let's plot the maximum deviation (close to π) of the total (resp. the oscillatory) energy on the interval $[0, 1000]$ as a function of $h\omega$ for $h = 0.02$.

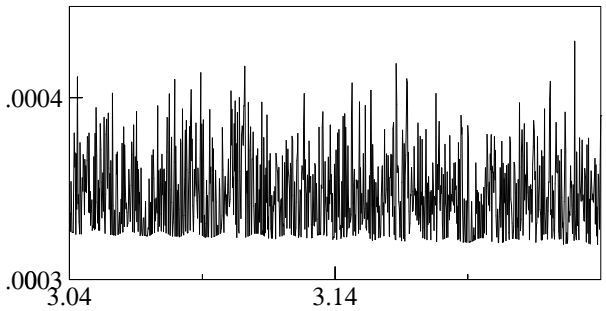


Figure 3.10: Maximum deviation of the total energy as a function of $h\omega$, for $h = 0.02$.

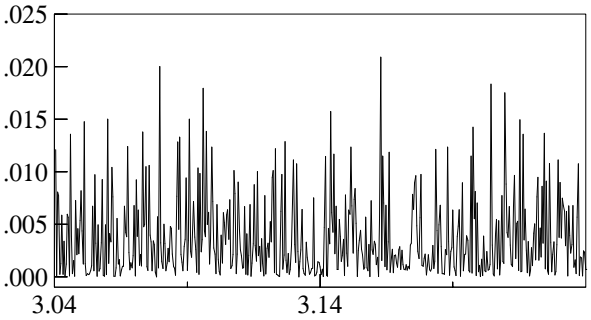


Figure 3.11: Maximum deviation of the oscillatory energy as a function of $h\omega$, for $h = 0.02$.

We see that for this method we have a uniform conservation of the two energies for all values of $h\omega$. Similar observations are also made for the other numerical methods given in the preceding section or for $h\omega$ near 2π.

We now take $\omega = 500$ and $h = 0.025$, we thus have $h\omega = 12.5$. We only plot the numerical solution obtained by DOP853, Method 1 and Method 3 on the interval $[0, 1000]$.

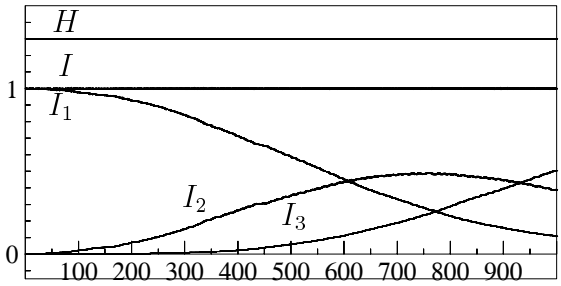


Figure 3.12: Large omega experiment using DOP853.

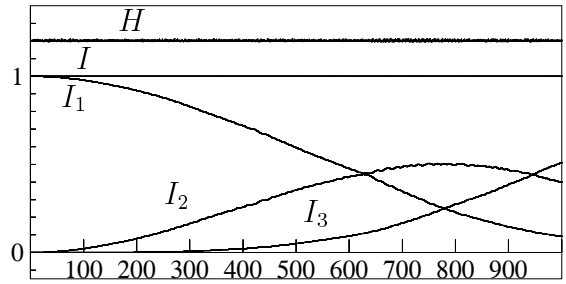


Figure 3.13: Large omega experiment using Method 1.

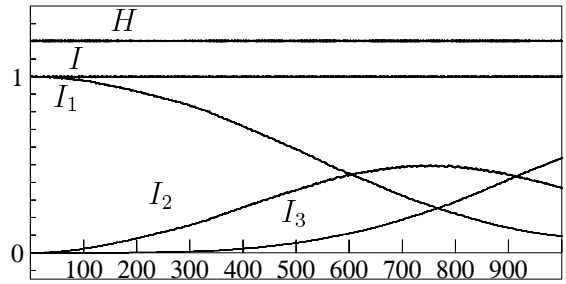


Figure 3.14: Large omega experiment using Method 3.

This illustrates the fact that for large $h\omega$ too, the total and oscillatory energies are well preserved by the numerical methods.

The last figures show the numerical solution obtained by the third method and DOP853 on an interval of length 10000 with $\omega = 50$, we take $h = 0.2$ for the third method. For long time intervals, the preservation of the energies is still valid.(For graphical reason, we have ploted a shiffted Hamiltonian).

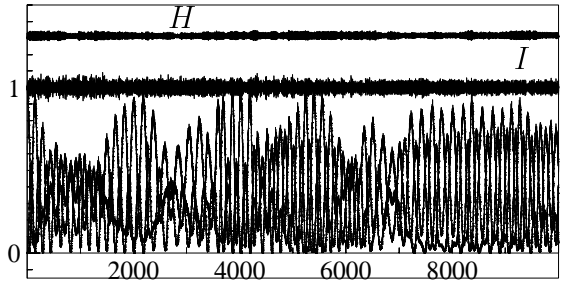


Figure 3.15: Long time intervals experiment using Method 3.

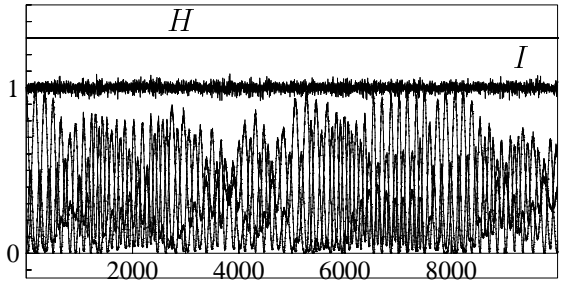


Figure 3.16: Long time intervals experiment using DOP853.

The second example describes a one dimensional model of a diatomic gas with short-range interaction forces. This model is described in [BGG94], for four diatomic molecules, the Hamiltonian reads

$$H(x,\zeta,p,\pi) = \sum_{l=1}^{2} \frac{1}{2}(\pi_l^2 + \omega^2\zeta_l^2 + p_l^2) + \sum_{l=1}^{3} U(x_l + \zeta_l - x_{l-1} - \zeta_{l-1}),$$

where the fixed end particles satisfies $x_0 = \zeta_0 = \zeta_3 = 0$ and $x_3 = 12$. We take two different potentials: a Lennard-Jones one $U(s) = s^{-12} - s^{-6}$ and $U(s) = \frac{\mathrm{e}^{-s^2}}{s}$. For each potential, we use DOP853, Method 1 and Method 3.

For the Lennard-Jones potential, we plot the oscillatory energies and the Hamiltonian obtained by the numerical methods on a time interval of length 200. For the initial values, we take $x = (4, 4+2^{1/6}), \zeta = (1/\omega, 1/100000), p = (0, 1.1), \pi = (1.2, 0.1)$. For the parameter ω, we take $\omega = 150$.

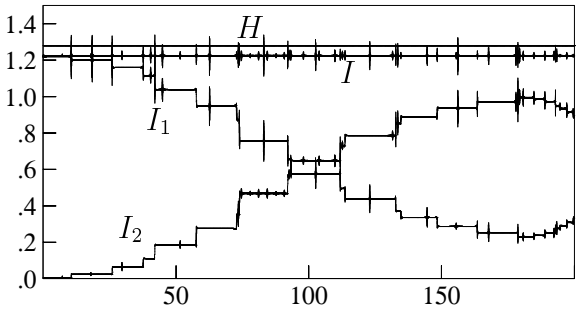


Figure 3.17: Total and oscillatory energies for a Lennard-Jones potential using DOP853.

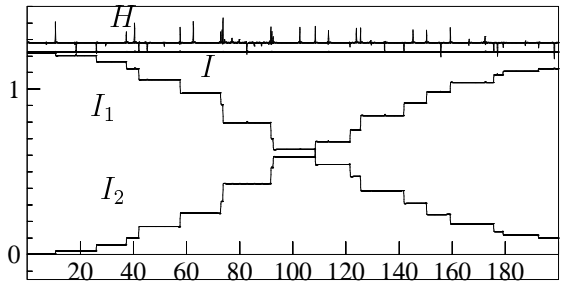


Figure 3.18: Total and oscillatory energies for a Lennard-Jones potential using Method 1.

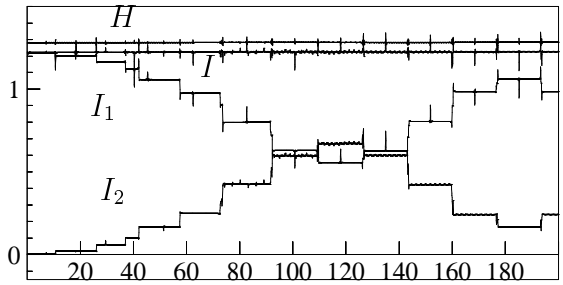


Figure 3.19: Total and oscillatory energies for a Lennard-Jones potential using Method 3.

For the other potential, we plot the same quantities as before on the same interval. The initial values are $x = (1, 5), \zeta = (1/\omega, 1/10000), p = (1, 1.1), \pi = (0.2, 1.1)$ and $\omega = 50$.

We see that for both potentials, the numerical behaviour discounted is present.

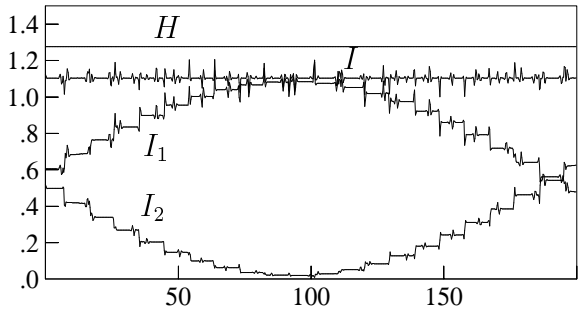


Figure 3.20: Total and oscillatory energies for a short-range potential using DOP853.

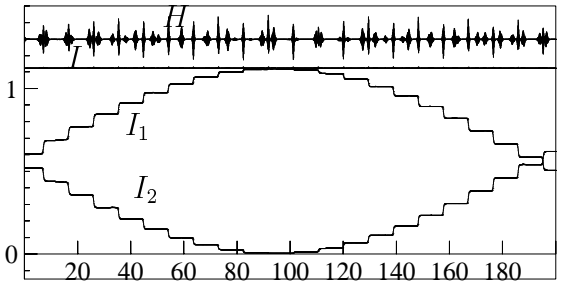


Figure 3.21: Total and oscillatory energies for a short-range potential using Method 1.

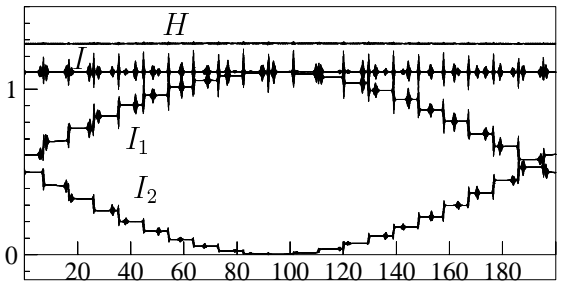


Figure 3.22: Total and oscillatory energies for a short-range potential using Method 3.

Chapter 4

Multi-frequency oscillatory differential equations

In this chapter we extend the techniques of Chapter 2 to the multi-frequency case. We first study the non-resonant case in Section 4.1. Sections 4.2 and 4.3 are devoted to the two frequencies resonant case. The general case is analysed in Section 4.4. Finally, we say a few words on the numerical treatment of these differential equations. This chapter is based on ideas of the texts [CHL03] and [HL00]. Similar results are obtained in [CHL] but with a different approach (see the end of this chapter for more details).

4.1 Non-resonant case

We generalize the single high frequency problem studied in Chapter 2 by adding other frequencies to equation (1.1). More precisely, we study the system of differential equations

$$\begin{aligned} \ddot{x}_1 &= g_1(x_1,\dots,x_n) \\ \ddot{x}_j + \omega_j^2 x_j &= g_j(x_1,\dots,x_n), \quad \text{for} \quad j=2,\dots,n, \end{aligned}$$

or shortly

$$\ddot{x} + \Omega^2 x = g(x) \qquad \text{with} \qquad \Omega = \mathrm{diag}(0,\omega_2 I,\omega_3 I,\dots,\omega_n I), \tag{1.1}$$

where we use the notation $\omega_i = a_i\,\lambda$ for $i=2,\dots,n$. The real numbers a_i are fixed, distinct and are such that $\min_i(a_i) \geq 1$ (for normalization). Each block of the matrix Ω has arbitrary dimension. We only consider this problem for $\lambda \gg 1$ and take $g(x) = -\nabla U(x)$ so that the problem is Hamiltonian.

To obtain near-conservation of

$$I_j(x,\dot{x}) \,=\, \frac{1}{2}\Big(\|\dot{x}_j\|^2 + \omega_j^2\,\|x_j\|^2\Big) \text{ , with } j=2,\dots,n, \tag{1.2}$$

in the n-frequencies case (1.1) we need some additional hypothesis, namely, that there exist some positive constants ν,γ such that

$$|k\cdot a| \geq \gamma\cdot|k|^{-\nu} \text{ for all } \; k\in\mathbb{Z}^{n-1}\,,\, k\neq 0. \tag{1.3}$$

In this section, we first analyse the case $n = 3$ in (1.1), the general case will be discused by the end of Section 4.1. We remark that (1.3) is automatically satisfied if $a_2 = \sqrt{2}$ and $a_3 = 1$. This is also the case if, for example, $\frac{a_2}{a_3}$ is algebraic [1]. The main result of this section is to prove the near-conservation over exponentially long time intervals of the following two oscillatory energies :

$$I_2(x, \dot{x}) = \frac{1}{2}\Big(\|\dot{x}_2\|^2 + \omega_2^2 \|x_2\|^2\Big) \tag{1.4}$$

$$I_3(x, \dot{x}) = \frac{1}{2}\Big(\|\dot{x}_3\|^2 + \omega_3^2 \|x_3\|^2\Big). \tag{1.5}$$

We assume again that the initial values of system (1.1) satisfy

$$\frac{1}{2}\Big(\|\dot{x}(0)\|^2 + \|\Omega x(0)\|^2\Big) \leq E, \tag{1.6}$$

where E is independent of λ. We do not require E to be small.

Before we begin to treat the non-resonant case, let us mention some preparations which will be also useful for the general case. We use the technique of modulated Fourier expansions to write the solution of (1.1) as a (formal) series of the form

$$x(t) = y(t) + \sum_{|k|\neq 0} \mathrm{e}^{\mathrm{i}k\cdot\omega t} z^k(t), \tag{1.7}$$

where $k = (k_2, k_3)$ (the k_i are integers), $\omega = (\omega_2, \omega_3)$, $k \cdot \omega$ is the usual scalar product of two vectors and $|k| = |k_2| + |k_3|$. We also use the notation $z^{-k} = \bar{z}^k$ for the index k.

As explained in [HL00], inserting (1.7) into the differential equation (1.1), expanding the nonlinearity $g(x(t))$ around $y(t)$ and comparing coefficients of $\mathrm{e}^{\mathrm{i}k\cdot\omega t}$ gives relations that allow us to determine the real function $y(t)$ and the complex functions $z^k(t)$.

This gives for the first component of the system (1.1):

$$\begin{aligned} \ddot{y}_1 \; + \; & \sum_{|k|\neq 0} \mathrm{e}^{\mathrm{i}k\cdot\omega t}\big(\ddot{z}_1^k + 2\mathrm{i}(k\cdot\omega)\dot{z}_1^k - (k\cdot\omega)^2 z_1^k\big) \\ = \; & g_1(y) + \sum_{m\geq 1} \frac{1}{m!} g_1^{(m)}(y)\Big(\sum_{|k|\neq 0} \mathrm{e}^{\mathrm{i}k\cdot\omega t} z^k\Big)^m. \end{aligned}$$

Comparing coefficients of $\mathrm{e}^{\mathrm{i}k\cdot\omega t}$ on each side of the equality yields the following systems (for λ sufficiently large):

$$\begin{aligned} &\text{for } k = (0,0): \quad \ddot{y}_1 = g_1(y) + \sum_{s(\alpha)=(0,0)} \frac{1}{m!} g_1^{(m)}(y) z^\alpha, \\ &\text{if } |k\cdot\omega| = 0: \quad \ddot{z}_1^k = \sum_{s(\alpha)=k} \frac{1}{m!} g_1^{(m)}(y) z^\alpha\,, \\ &\text{if } |k\cdot\omega| \neq 0: \quad -(k\cdot\omega)^2 z_1^k + 2\mathrm{i}(k\cdot\omega)\dot{z}_1^k + \ddot{z}_1^k = \sum_{s(\alpha)=k} \frac{1}{m!} g_1^{(m)}(y) z^\alpha, \end{aligned} \tag{1.8}$$

[1] We thank N. Bartholdi and J. Steinig for an interesting discussion on the subject. For a reference, see [Cas72, Chap. 4]

where $g_1^{(m)}(y)z^\alpha$ stands for $g_1^{(m)}(y)(z^{\alpha_2},\dots,z^{\alpha_m})$ and $s(\alpha)=(\sum_{i=2}^m \alpha_i)$ for a multi-index $\alpha=(\alpha_2,\dots,\alpha_m)$ with $\alpha_i\in\mathbb{Z}^2$.

Exactly the same calculations for the second and third component gives the equations:

$$\begin{aligned}
\text{for } k=(0,0): \qquad & \omega_j^2 y_j+\ddot{y}_j = g_j(y)+\sum_{s(\alpha)=(0,0)}\frac{1}{m!}g_j^{(m)}(y)z^\alpha,\\
\text{if } |\omega_j^2-(k\cdot\omega)^2|=0: \quad & 2\mathrm{i}(k\cdot\omega)\dot{z}_j^k+\ddot{z}_j^k=\sum_{s(\alpha)=k}\frac{1}{m!}g_j^{(m)}(y)z^\alpha\,, \qquad (1.9)\\
\text{if } |\omega_j^2-(k\cdot\omega)^2|\neq 0: \quad & (\omega_j^2-(k\cdot\omega)^2)z_j^k+2\mathrm{i}(k\cdot\omega)\dot{z}_j^k+\ddot{z}_j^k=\\
& \sum_{s(\alpha)=k}\frac{1}{m!}g_j^{(m)}(y)z^\alpha,
\end{aligned}$$

with $j=2,3$.

By using iteratively equations (1.8) and (1.9), we remove the higher derivatives and obtain algebraic relations and differential equations, like those in Section 2.2.

The non-resonant hypothesis (1.3) allows us to express all the modulation functions y and z^k of (1.7) as functions of the variable $\mathcal{Y}=(y_1,\dot{y}_1,z_2,z_3)$, where $z_2=z_2^{(1,0)}$ and $z_3=z_3^{(0,1)}$. Indeed, the case $|\omega_2^2-(k\cdot\omega)^2|=0$ of (1.9) (which leads to a differential equation for z_2^k) only appears if $a_2^2=(a_2\,k_2+a_3\,k_3)^2$, i.e. if and only if $k_2=\pm1$ and $k_3=0$. The same arguments occur for the third component: we have $|\omega_3^2-(k\cdot\omega)^2|=0$ if and only if $k_2=0$ and $k_3=\pm1$. Thus, this gives a system of differential equations for y_1, z_2, z_3, namely:

$$\ddot{y}_1=\sum_{l\geq 0}\lambda^{-l}F_{1l}(\mathcal{Y}),\quad \dot{z}_2=\sum_{l\geq 1}\lambda^{-l}F_{2l}(\mathcal{Y}),\quad \dot{z}_3=\sum_{l\geq 1}\lambda^{-l}F_{3l}(\mathcal{Y}), \qquad (1.10)$$

and algebraic relations for the other variables (with the notation $z_i^{(0,0)}=y_i$, for $i=2,3$)

$$z_i^k=\sum_{l\geq 0}\lambda^{-l}G_{il}^k(\mathcal{Y}). \qquad (1.11)$$

4.1.1 Formal analysis

In this subsection, we follow ideas of [HL00] and give a formal analysis that leads to the near-conservation of the oscillatory energies over long time intervals. In the next subsection, we do this analysis rigorously and show the near-conservation of this quantities over exponentially long time intervals.

The series in (1.10) and (1.11) usually diverge, so we have to truncate them

$$\ddot{y}_1=\sum_{l=0}^{N}\lambda^{-l}F_{1l}(\mathcal{Y}),\quad \dot{z}_2=\sum_{l=1}^{N}\lambda^{-l}F_{2l}(\mathcal{Y}),\quad \dot{z}_3=\sum_{l=1}^{N}\lambda^{-l}F_{3l}(\mathcal{Y}), \qquad (1.12)$$

$$z_i^k=\sum_{l=0}^{N}\lambda^{-l}G_{il}^k(\mathcal{Y}), \qquad (1.13)$$

at an arbitrary integer $N \geq 2$. We insert these sums into (1.8) and (1.9) to determine recursively the functions F_{ij} and G_{ij}^k. The initial values given for the differential equation (1.1) permit us to find initial values for the system (1.12). This gives the desired modulated Fourier expansion. The proofs of [HL00] extend in a straightforward way and give the following result.

Theorem 4.1.1 *Under hypothesis (1.6), if $x(t)$ stays in a compact set for $0 \leq t \leq T$, and for λ sufficiently large, the solution of (1.1) has, for an arbitrary integer $N \geq 2$, a modulated Fourier expansion*

$$x(t) = y(t) + \sum_{0<|k|<N} \mathrm{e}^{\mathrm{i}k\cdot\omega t} z^k(t) + R_N(t), \tag{1.14}$$

where the remainder satisfies

$$R_N(t) = \mathcal{O}(\lambda^{-N}) \ . \tag{1.15}$$

The real functions $y = (y_1, y_2, y_3)$ and the complex functions $z^k = (z_1^k, z_2^k, z_3^k)$ defined above are bounded, together with their derivatives, by

$$\begin{array}{llll} y_1 = \mathcal{O}(1), & & z_1^k = \mathcal{O}(\lambda^{-2-|k|}), & k \neq (0,0), \\ z_2 = \mathcal{O}(\lambda^{-1}), & & z_3 = \mathcal{O}(\lambda^{-1}), & \\ z_2^k = \mathcal{O}(\lambda^{-2-|k|}), & k \neq (\pm 1, 0), & z_3^k = \mathcal{O}(\lambda^{-2-|k|}), & k \neq (0, \pm 1), \end{array} \tag{1.16}$$

where we have used the notation $|k| = |k_2| + |k_3|$ and $z^0 = y$. The constants symbolized by $\mathcal{O}$ are independent of λ and t. □

We turn now to the Hamiltonian case, where $g(x) = -\nabla U(x)$ with $U(x)$ an analytic potential. The Hamiltonian of system (1.1) is then given by

$$H(x, \dot{x}) = \frac{1}{2}\Big(\|\dot{x}\|^2 + \|\Omega x\|^2 \Big) + U(x). \tag{1.17}$$

Note that the Hamiltonian structure passes over to the differential equation for the coefficients of the modulated Fourier expansion:

$$y^0(t) = y^{(0,0)}(t) = \begin{pmatrix} y_1(t) \\ y_2(t) \\ y_3(t) \end{pmatrix} \quad \text{and} \quad y^k(t) = \mathrm{e}^{\mathrm{i}k\omega t} \begin{pmatrix} z_1^k(t) \\ z_2^k(t) \\ z_3^k(t) \end{pmatrix} . \tag{1.18}$$

To see this, let $\mathbf{y} = (..., y^{(-1,0)}, y^0, y^{(1,0)}, ...)$ be the sequence formed by the vectors in (1.18) with $|k| < N$ and define

$$\mathcal{U}(\mathbf{y}) = U(y^0) + \sum_{m\geq 0} \frac{1}{m!} \sum_{s(\alpha)=(0,0), |\alpha_i|\neq 0} U^{(m)}(y^0)\big(y^{\alpha_1}, \ldots, y^{\alpha_m}\big). \tag{1.19}$$

By definition (see (1.8)–(1.9)), the modulation functions satisfy the system

$$\ddot{y}^k + \Omega^2 y^k \ = \ -\nabla_{y^{-k}} \mathcal{U}(\mathbf{y}) + \mathcal{O}(\lambda^{-N}) \tag{1.20}$$

which is Hamiltonian (neglecting the $\mathcal{O}(\cdot)$ term) with

$$\mathcal{H}(\mathbf{y},\dot{\mathbf{y}}) \;=\; \frac{1}{2}\sum_{0\le|k|<N}\Big((\dot{y}^{-k})^T\dot{y}^k + (y^{-k})^T\Omega^2 y^k\Big) + \mathcal{U}(\mathbf{y}). \tag{1.21}$$

As in [HL00], we can prove that (1.21) is well conserved and very close to our Hamiltonian (1.17):

Theorem 4.1.2 *Under the assumptions of Theorem 4.1.1 we have, for* $0 \le t \le T$ *and* $\mathbf{y}(t)$ *solution of (1.20),*

$$\mathcal{H}(\mathbf{y}(t),\dot{\mathbf{y}}(t)) \;=\; \mathcal{H}(\mathbf{y}(0),\dot{\mathbf{y}}(0)) + \mathcal{O}(t\lambda^{-N}), \tag{1.22}$$
$$\mathcal{H}(\mathbf{y}(t),\dot{\mathbf{y}}(t)) \;=\; H(x(t),\dot{x}(t)) + \mathcal{O}(\lambda^{-1})\;. \tag{1.23}$$

The constants symbolized by $\mathcal{O}(\cdot)$ *are independent of* λ *and* t*.*

We are now prepared to study the adiabatic invariants (1.4) and (1.5) mentioned in the beginning of this chapter. We first prove that the expressions

$$\mathcal{I}_2(\mathbf{y},\dot{\mathbf{y}}) = \; -\mathrm{i}\,\omega_2 \sum_{0<|k|<N} k_2 (y^{-k})^T \dot{y}^k$$

and

$$\mathcal{I}_3(\mathbf{y},\dot{\mathbf{y}}) = \; -\mathrm{i}\,\omega_3 \sum_{0<|k|<N} k_3 (y^{-k})^T \dot{y}^k$$

are well conserved by the solution of (1.20) and that they are close to (1.4), (1.5) respectively.

Theorem 4.1.3 *Let* $\mathbf{y}(t)$ *be the vector with components* $y^k(t)$ *of (1.18). Under the assumptions of Theorem 4.1.1 we have, for* $j = 2, 3$ *and for* $0 \le t \le T$*,*

$$\mathcal{I}_j\big(\mathbf{y}(t),\dot{\mathbf{y}}(t)\big) \;=\; \mathcal{I}_j\big(\mathbf{y}(0),\dot{\mathbf{y}}(0)\big) + \mathcal{O}(t\lambda^{-N}),$$
$$\mathcal{I}_j\big(\mathbf{y}(t),\dot{\mathbf{y}}(t)\big) \;=\; I_j\big(x(t),\dot{x}(t)\big) + \mathcal{O}(\lambda^{-1}),$$

where $I_j\big(x(t),\dot{x}(t)\big)$ *are defined by (1.4)–(1.5) and where the constants symbolizing the* $\mathcal{O}(\cdot)$ *are independent of* λ *and* t*.*

Proof. Looking at the proofs given in [HL00] and [HLW02, Sect. XIII.6.2], we want to use a relation of the form

$$\sum_{0<|k|<N} \mathrm{i}\,k\,(y^k)^T \nabla\mathcal{U}_k(\mathbf{y}) \;=\; 0.$$

The equivalant algebraic identities for the two frequencies case are now given by

$$\sum_{0<|k|<N} \mathrm{i}\,k_j\,(y^k)^T \nabla\mathcal{U}_{y^k}(\mathbf{y}) \;=\; 0 \quad \text{for} \quad j = 2, 3. \tag{1.24}$$

To prove these identities, let $a(\mu) = (\ldots, \mathrm{e}^{\mathrm{i}\,(-1,0)\cdot\mu} y^{(-1,0)}, y^{(0,0)}, \mathrm{e}^{\mathrm{i}\,(1,0)\cdot\mu} y^{(1,0)}, \ldots)$ where $\mu = (\mu_2, \mu_3)$ is a couple of real numbers. By definition of $\mathcal{U}$, the function $\mathcal{U}(a(\mu))$ does not depend on μ so that the partial derivatives $\dfrac{\partial \mathcal{U}}{\partial \mu_j}$ are equal to zero. Evaluating these partial derivatives at $\mu = (0,0)$ yields (1.24). Once this is done, we just have to follow the ideas of the proof of Theorem 4.3 in [HL00] to conclude the proof. □

To prove the fact that the oscillatory energies (1.4) and (1.5) are nearly conserved over long time intervals, we use repeatedly Theorem 4.1.3 as explained in [HL00, Cor. 4.4].

As in Chapter 2, it is possible to obtain, for the non-resonant case, conservation of the two quantities (1.4) and (1.5) over exponentially long time intervals. This will be discussed in more details in the next subsection.

4.1.2 Rigorous estimates

To obtain the desired near-conservation of oscillatory energies (1.2), we adapt the proofs and ideas of Sections 2.2 and 2.3 of Chapter 2 to the case of three components in (1.1), the n-component case follows by similar arguments (see the end of this section).

To get recurrence relations for the coefficient functions of (1.10) and (1.11), it is convenient to use the following Lie operators $\mathcal{L}_l$

$$\mathcal{L}_l G = D_2 G \cdot F_{1l} + D_3 G \cdot F_{2l} + D_4 G \cdot F_{3l} + \begin{cases} D_1 G \cdot \dot{y}_1 & \text{if } \ l = 0 \\ 0 & \text{if } \ l \geq 1, \end{cases} \tag{1.25}$$

where D_j denotes the partial derivative with respect to the jth argument of a smooth function $G(y_1, \dot{y}_1, z_2, z_3)$. This definition of the Lie operators is made to satisfy

$$\frac{\mathrm{d}}{\mathrm{d}t} G(y_1(t), \dot{y}_1(t), z_2(t), z_3(t)) = \sum_{l \geq 0} \lambda^{-l} \mathcal{L}_l G(y_1(t), \dot{y}_1(t), z_2(t), z_3(t))$$

for $y_1(t), z_2(t)$ and $z_3(t)$ solutions of the differential equations (1.10). A straightforward but laborious extension of the proof of Lemma 2.2.1 gives

Lemma 4.1.4 *The function $\big(x_1(t), x_2(t), x_3(t)\big)$ of (1.7), with $y_i(t)$ and $z_i^k(t)$ given by (1.10) and (1.11), represents a formal solution of (1.1) if the coefficient functions F_{il} and*

G^k_{il} satisfy the following recurrence relations (for $l \geq 0$):

$$\begin{aligned}
F_{1l} &= S_1(0,l),\\
G^k_{1l} &= \frac{1}{(k_2a_2+k_3a_3)^2}\Big(\sum_{m+n+j=l-2}\mathcal{L}_m\mathcal{L}_nG^k_{1j}\\
&+ 2\mathrm{i}(k_2a_2+k_3a_3)\sum_{m+j=l-1}\mathcal{L}_mG^k_{1j} - S_1(k,l-2)\Big),\\
F_{il} &= \frac{1}{2\mathrm{i}}\Big(S_i(1,l-1) - \sum_{m+j=l-1}\mathcal{L}_mF_{ij}\Big),\\
G^k_{il} &= \frac{1}{a_i^2-(k_2a_2+k_3a_3)^2}\Big(S_i(k,l-2) - \sum_{m+n+j=l-2}\mathcal{L}_m\mathcal{L}_nG^k_{ij}\\
&- 2\mathrm{i}(k_2a_2+k_3a_3)\sum_{m+j=l-1}\mathcal{L}_mG^k_{ij}\Big).
\end{aligned}$$

Here, $i=2,3$ and the sums are over $m\geq 0, n\geq 0, j\geq 0$; we have used the abbreviation

$$S_i(k,l) = \sum_{m,n,p\geq 0}\frac{1}{m!\,n!\,p!}\sum_{\substack{\alpha,\beta,\gamma\\ s(\alpha)+s(\beta)+s(\gamma)=k}}\sum_{\substack{e,f,h\\ s(e)+s(f)+s(h)=l}} D_1^mD_2^nD_3^pg_i(y_1,0,0)(G^\alpha_{1e},G^\beta_{2f},G^\gamma_{3h}),$$

with $\alpha=(\alpha_1,\ldots,\alpha_m), \beta=(\beta_1,\ldots,\beta_n), \gamma=(\gamma_1,\ldots,\gamma_p), e=(e_1,\ldots,e_m), f=(f_1,\ldots,f_n)$, $h=(h_1,\ldots,h_p)$ are multi-indices with $\alpha_i\neq(0,0)$, β_i and γ_i arbitrary in $\mathbb{Z}\times\mathbb{Z}$, $e_i,f_i,h_i\geq 0$, and $(G^\alpha_{1e},G^\beta_{2f},G^\gamma_{3h}) = (G^{\alpha_1}_{1,e_1},\ldots,G^{\alpha_m}_{1,e_m},G^{\beta_1}_{2,f_1},\ldots,G^{\beta_n}_{2,f_n},G^{\gamma_1}_{3,h_1},\ldots,G^{\gamma_p}_{3,h_p})$. We use the abbreviation $s(\alpha)=\sum_{i=1}^m\alpha_i$ and similarly for the other multi-indices. □

Keeping in mind that we want to get upper bounds for the functions appearing in the last lemma, we require the function $g(x)$ to be analytic and bounded by M in a complex domain $\{(x_1,x_2,x_3) : \|x_1-y_{10}\|\leq 4R, \|x_2\|\leq 3R, \|x_3\|\leq 3R\}$. Cauchy's estimates then imply

$$\|D_1^mD_2^nD_3^pg_i(y_1,0,0)\| \leq m!\,n!\,p!\,M\,(3R)^{-m-n-p}, \qquad \text{for} \quad \|y_1-y_{10}\|\leq R \tag{1.26}$$

and for all $m,n,p\geq 0$.

We fix a value $\mathcal{Y}_0=(y_{10},\dot y_{10},0,0)$, and we consider the complex ball

$$B_\rho(\mathcal{Y}_0) = \big\{(y_1,\dot y_1,z_2,z_3) : \|y_1-y_{10}\|\leq\rho R, \|\dot y_1-\dot y_{10}\|\leq\rho M, \|z_2\| \text{ and } \|z_3\|\leq\rho\frac{R}{4}\big\}. \tag{1.27}$$

For a function $G(y_1,\dot y_1,z_2,z_3)$ defined on $B_\rho(\mathcal{Y}_0)$ we let

$$\|G\|_\rho = \max\big\{\|G(y_1,\dot y_1,z_2,z_3)\| : (y_1,\dot y_1,z_2,z_3)\in B_\rho(\mathcal{Y}_0)\big\}. \tag{1.28}$$

Since the coefficient functions are defined via expressions of the form $\mathcal{L}_lG$, the following lemma (based on Cauchy's estimate) will be useful.

Lemma 4.1.5 *Let G be analytic and bounded on $B_\rho(\mathcal{Y}_0)$, and let F_{1l}, F_{2l} and F_{3l} be bounded on $B_\sigma(\mathcal{Y}_0)$ with $0 \le \sigma < \rho$. Then we have*

$$\begin{aligned}
\|\mathcal{L}_0 G\|_\sigma &\le \tfrac{1}{\rho-\sigma} \cdot \|G\|_\rho \cdot \max\big(\|F_{10}\|_\sigma/M, \|\dot y_1\|_\sigma/R\big),\\
\|\mathcal{L}_l G\|_\sigma &\le \tfrac{1}{\rho-\sigma} \cdot \|G\|_\rho \cdot \max\big(\|F_{1l}\|_\sigma/M, , 4\|F_{2l}\|_\sigma/R, 4\|F_{3l}\|_\sigma/R\big) \quad \textit{for} \quad l \ge 1.
\end{aligned}$$

With these tools in hand, we can now bound the coefficient functions F_{ij} and G_{ij}^k. Similarly to [CHL03] we fix a positive integer L, we put $\delta = 1/(2L)$, and we consider the norms corresponding to balls with shrinking radius $\rho = 1 - l\delta$ $(0 \le l \le L)$. We also consider a sequence μ_l defined by $\mu_0 = \frac{1}{2}$ and $\mu_l = \mu_0 + l\delta$ $(1 \le l \le L)$.

Lemma 4.1.6 *Let $\mathcal{Y}_0 = (y_{10}, \dot y_{10}, 0, 0)$ be given, and assume that (1.26) holds. The functions F_{ij} and G_{ij}^k of Lemma 4.1.4 satisfy*

$$\begin{aligned}
&\|F_{10}\|_1 \le a_0 M, \qquad \|\dot y_1\|_1 \le a_0 R/4,\\
&\|F_{1l}\|_{1-l\delta} \le a_l M, \qquad \|F_{2l}\|_{1-l\delta} \le a_l R/4, \qquad \|F_{3l}\|_{1-l\delta} \le a_l R/4, \qquad 1 \le l \le L,\\
&\tfrac{1}{\mu_0}\Big(\|G_{20}^{(1,0)}\|_1 + \|G_{20}^{(-1,0)}\|_1 + \|G_{30}^{(0,1)}\|_1 + \|G_{30}^{(0,-1)}\|_1\Big) \le b_0 R,\\
&\max\Big(\sum_{k\in\mathbb{Z}\times\mathbb{Z}} \frac{|k|^2}{\mu_l^{|k|}}\|G_{1l}^k\|_{1-l\delta}\ ,\ \sum_{k\in\mathbb{Z}\times\mathbb{Z}} \frac{1+|k|^2}{\mu_l^{|k|}}\|G_{il}^k\|_{1-l\delta}\Big) \le b_l R, \quad 1 \le l \le L \text{ and } i = 2,3,
\end{aligned}$$

where $a_0 = \max(27, 4(\|\dot y_{10}\|_1 + M)/R)$, $b_0 = 2$, and the generating functions $a(\zeta) = \sum_{l\ge1} a_l\zeta^l$ and $b(\zeta) = \sum_{l\ge1} b_l\zeta^l$ are implicitly given by

$$\begin{aligned}
a(\zeta) &= -27 + 27\big(1 + \tfrac{2M\zeta}{R}\big)\big(1 - b(\zeta)\big)^{-3} + \tfrac{2\zeta}{\delta}\big(a_0 + a(\zeta)\big)a(\zeta),\\
b(\zeta) &= \tfrac{\alpha!}{\gamma^2}(2L)^{\alpha+1}\Big(\tfrac{27M\zeta^2}{R}\big(3 - b_0 - b(\zeta)\big)^{-3}\\
&+ \tfrac{2\max(a_2,a_3)\zeta}{\delta}\big(a_0 + a(\zeta)\big)\big(b_0 + b(\zeta)\big)\\
&+ \tfrac{\zeta^2}{\delta^2}\big(a_0 + a(\zeta)\big)^2\big(b_0 + b(\zeta)\big)\Big).
\end{aligned} \tag{1.29}$$

With $\alpha = [3 + 2\nu] + 1$, γ and ν taken from (1.3).

Proof. (a) In this proof we shall use the shorthand notation

$$\|G\|_l := \|G\|_{1-l\delta} = \max\big\{\|G(y_1, \dot y_1, z_2, z_3)\| \ :\ (y_1, \dot y_1, z_2, z_3) \in B_{1-l\delta}(\mathcal{Y}_0)\big\}.$$

Observe that $\|G\|_l$ is a decreasing function of l.

To obtain the desired statement, we begin with some estimations and then prove the result of this lemma by induction on l.

(b) Because of Lemma 4.1.4, the above estimates for G_{il}^k also imply

$$\sum_{|k|\ne0} \frac{\|G_{1l}^k\|_l}{\mu_l^{|k|}} \le b_l R, \quad \sum_{k\in\mathbb{Z}\times\mathbb{Z}} \frac{\|G_{il}^k\|_l}{\mu_l^{|k|}} \le b_l R, \quad \text{for} \quad l \ge 0 \quad \text{and} \quad i = 2,3. \tag{1.30}$$

Using these relations and the analyticity assumption (1.26), we are able to majorize $S_i(k,l)$ as follows (for $i=1,2,3$):

$$\begin{aligned}\|S_i(k,l)\|_l \;\le\; & \sum_{m,n,p\ge 0} \frac{m!\,n!\,p!}{m!\,n!\,p!} \sum_{\substack{\alpha,\beta,\gamma\\ |\alpha_i|\neq 0}} \sum_{\substack{s(e)+s(f)+s(h)\\ =l}} M(3R)^{-m-n-p}\cdot\\ & \cdot\mu_{e_1}^{|\alpha_1|}\frac{\|G_{1e_1}^{\alpha_1}\|_{e_1}}{\mu_{e_1}^{|\alpha_1|}}\ldots\mu_{h_1}^{|\gamma_1|}\frac{\|G_{3h_1}^{\gamma_1}\|_{h_1}}{\mu_{h_1}^{|\gamma_1|}}\ldots\\ \le\; & M\mu_l^{|k|}\sum_{m,n,p\ge 0}\sum_{s(e)+s(f)+s(h)=l} 3^{-m-n-p}\,b_{e_1}\ldots b_{f_1}\ldots b_{h_p}\\ \le\; & M\mu_l^{|k|}\sum_{j\ge 0}\frac{(j+1)(j+2)}{2}\sum_{d_1+\ldots+d_j=l} 3^{-j}\,b_{d_1}\ldots b_{d_j} \;=\; M\mu_l^{|k|}c_l,\end{aligned}$$

where c_l $(l\ge 0)$ are the coefficients of the generating function

$$\sum_{l\ge 0} c_l\zeta^l = c(\zeta) = \frac{1}{\big(1-\frac{b_0+b(\zeta)}{3}\big)^3} = \frac{27}{\big(1-b(\zeta)\big)^3}.$$

We used $\|G_{1e_1}^{\alpha_1}\|_l \le \|G_{1e_1}^{\alpha_1}\|_{e_1}$, which is a consequence of $e_1\le l$ (the same estimates hold for the other multi-indices). We also used the fact that $|k|\le|\alpha_1|+|\beta_1|+|\gamma_1|+\ldots$ for α,β and γ such that $s(\alpha)+s(\beta)+s(\gamma)=k$.

(c) A twofold application of Lemma 4.1.5 and the fact that $\|G_{1j}^k\|\le\mu_j^{|k|}b_jR$ yields

$$\begin{aligned}\sum_{m+n+j=l-2}\|\mathcal{L}_m\mathcal{L}_nG_{1j}^k\|_l \;&\le \frac{1}{\delta^2}\sum_{m+n+j=l-2}\|G_{1j}^k\|_j\,a_m\,a_n\\ &\le \frac{R}{\delta^2}\sum_{m+n+j=l-2}\mu_j^{|k|}\,b_j\,a_m\,a_n.\end{aligned}$$

This implies

$$\sum_{m+n+j=l-2}\|\mathcal{L}_m\mathcal{L}_nG_{1j}^k\|_l \le \frac{R\mu_{l-2}^{|k|}}{\delta^2}\,d_{l-2},$$

where the generating function of the d_l is

$$d(\zeta)=\sum_{l\ge 0}d_l\zeta^l = \big(b_0+b(\zeta)\big)\big(a_0+a(\zeta)\big)^2.$$

The same estimate is obtained for $\sum_{m+n+j=l-2}\|\mathcal{L}_m\mathcal{L}_nG_{ij}^k\|_l$ where $i=2,3$.

(d) In order to estimate $|k|\sum_{m+j=l-1}\|\mathcal{L}_mG_{ij}^k\|_l$ for $i=1,2,3$, we observe that similarly to (1.30) also

$$|k|\,\|G_{il}^k\|_l \;\le\; R\mu_l^{|k|}b_l, \quad \text{for} \quad l\ge 0, \tag{1.31}$$

holds. As in part (c), we thus obtain

$$|k| \sum_{m+j=l-1} \|\mathcal{L}_m G_{ij}^k\|_l \;\le\; \frac{R\mu_{l-1}^{|k|}}{\delta}\, q_{l-1},$$

where the generating function for the q_l is

$$q(\zeta) = \sum_{l\ge 0} q_l \zeta^l = \big(b_0 + b(\zeta)\big)\big(a_0 + a(\zeta)\big).$$

(e) We can now take the sum over k and use the Diophantine estimate (1.3). Because the quantity $\dfrac{\mu_{l-1}}{\mu_l}$ is smaller than 1, we obtain the following estimate using the geometric series

$$\begin{aligned}\sum_{k\in\mathbb{Z}\times\mathbb{Z}} \frac{|k|^2}{\mu_l^{|k|}} \|G_{1l}^k\|_l \;&\le\; \frac{1}{\gamma^2} \sum_{k\in\mathbb{Z}\times\mathbb{Z}} |k|^{2+2\nu}\cdot \\ &\quad\cdot\big(\tfrac{\mu_{l-1}}{\mu_l}\big)^{|k|} \Big(\tfrac{R}{\delta^2} d_{l-2} + 2\max(a_2,a_3)\tfrac{R}{\delta} q_{l-1} + M c_{l-2}\Big) \\ &\le\; \tfrac{\alpha!}{\gamma^2}(2L)^{\alpha+1} \Big(\tfrac{R}{\delta^2} d_{l-2} + 2\max(a_2,a_3)\tfrac{R}{\delta} q_{l-1} + M c_{l-2}\Big).\end{aligned}$$

Similar estimations are obtained for the other components.

(f) After these preparations the statement can be proved by induction on l. The bounds a_0 and b_0 are defined just to satisfy the estimates for $l = 0$. The form of the generating functions for a_l and b_l are a consequence of the recurrence relations of Lemma 4.1.4 and of parts (b), (c), (d) and (e) of this proof. □

To get bounds on the expressions of Lemma 4.1.6, we have to majorize a_l and b_l. This can be done with the help of Cauchy's inequalities, because the generating functions $a(\zeta)$ and $b(\zeta)$ are analytic in a neighborhood of the origin. Since equations (1.29) depend on δ, R, L and M, we have to be careful in determining the radius of the disc of analyticity. In the following we assume $M \ge R$. This can be done without loss of generality, because we can always increase M without violating (1.26) or, even better, we can rescale time in the differential equation and thus multiply $g(x)$ by a scalar factor. Similar arguments given for the proof of Theorem 2.2.4 yield the following result.

Theorem 4.1.7 *We fix $\mathcal{Y}_0 = (y_{10}, \dot y_{10}, 0, 0)$, and we assume that the nonlinearity $g(x)$ satisfies (1.26) with $M \ge R$, and that $\|\dot y_{10}\| \le M$. The coefficient functions of Lemma 4.1.4 then satisfy for $l \ge 1$*

$$\begin{aligned}&\|F_{1l}\|_{1/2} \le \mu M \Big(\frac{\eta l^{\alpha+2} M}{R}\Big)^l, \qquad \|F_{il}\|_{1/2} \le \frac{\mu R}{4}\Big(\frac{\eta l^{\alpha+2} M}{R}\Big)^l, \\ &\max\Big(\sum_{k\in\mathbb{Z}\times\mathbb{Z}} \frac{|k|^2}{\mu_l^{|k|}} \|G_{1l}^k\|_{1-l\delta}\;,\;\sum_{k\in\mathbb{Z}\times\mathbb{Z}} \frac{1+|k|^2}{\mu_l^{|k|}} \|G_{il}^k\|_{1-l\delta}\Big) \;\le\; \mu R\Big(\frac{\eta l^{\alpha+2} M}{R}\Big)^l,\end{aligned}$$

with $i = 2, 3$ and where μ and η only depend on an upper bound of M/R but not on the other data of the differential equation. The norm is that of (1.28). □

Following [CHL03], we choose the truncation index N such that

$$N^{\alpha+2} \leq \frac{\lambda R}{\mathrm{e}^{\alpha+2}\eta M} < N^{\alpha+2}+1. \tag{1.32}$$

With this choice for the integer N, we obtain the following estimations :

$$\begin{aligned} \sum_{k\in\mathbb{Z}\times\mathbb{Z}} k^2 \sum_{2\leq l\leq N} \lambda^{-l}\|G^k_{1l}\|_{1/2} &\leq \text{Const}\cdot R\left(\frac{M}{\lambda R}\right)^2, \\ \sum_{k\in\mathbb{Z}\times\mathbb{Z}} (1+|k|^2) \sum_{2\leq l\leq N} \lambda^{-l}\|G^k_{il}\|_{1/2} &\leq \text{Const}\cdot R\left(\frac{M}{\lambda R}\right)^2, \end{aligned} \tag{1.33}$$

and for the coefficient functions F_{ij} :

$$\begin{aligned} \sum_{0\leq l\leq N} \lambda^{-l}\,\|F_{1l}\|_{1/2} &\leq \text{Const}\cdot M, \\ \sum_{1\leq l\leq N} \lambda^{-l}\,\|F_{il}\|_{1/2} &\leq \text{Const}\cdot M\cdot\lambda^{-1}, \end{aligned} \tag{1.34}$$

where $i=2,3$ and the constant only depend on an upper bound of M/R and ν. We have similar estimates as in Chapter 2, so that nothing has to be changed in the discussion of the initial values and in the existence of a T such that the solution stays in the ball

$$B=\{(y_1,\dot y_1,z_2,z_3)\,:\,\|y_1-y_{10}\|\leq R/2, \|\dot y_1-\dot y_{10}\|\leq M/2, \|z_2\|\leq R/8, \|z_3\|\leq R/8\}$$

for $0\leq t\leq T$.

We can now estimate the effect of the truncation made in (1.33) and (1.34).

Theorem 4.1.8 *Consider the differential equation (1.1) with initial values $x(0)$ and $\dot x(0)$ satisfying (1.6). Assume that the nonlinearity $g(x)$ is analytic in a complex ball and bounded by M. Then, there exists $\beta>0, T>0$ and $\lambda_0>0$ such that the defect*

$$\delta_k(t)=\ddot y^k(t)+\Omega^2 y^k(t)-\sum_{m\geq 0}\frac{1}{m!}\sum_{s(\alpha)=k,\alpha_i\neq 0} g^{(m)}(y^0(t))\big(y^{\alpha_1}(t),\dots,y^{\alpha_m}(t)\big)$$

(here $\alpha_i,k,s(\alpha)$ are in $\mathbb{Z}\times\mathbb{Z}$) satisfies for $0\leq t\leq T$ and for $\lambda\geq\lambda_0$

$$\sum_{k\in\mathbb{Z}\times\mathbb{Z}}\|\delta_k(t)\| \leq C\,M\,\mathrm{e}^{-\beta\lambda^{1/\alpha+2}}.$$

The constants C,γ,T,λ_0 only depend on an upper bound of M/R, on a_i, and on the Diophantine constants γ,ν (like the constant α, see Lemma 4.1.6) but not on λ.

Proof. To prove this theorem, we adapt the proof of Theorem 2.3.2 given in Chapter 2 to our case. First we let N and λ be independent variables, and we consider the defect as a function of t, N, and λ^{-1}. By the construction of the coefficient functions y^k, the

defect δ_k is an analytic function of $\zeta = \lambda^{-1}$ in a neighborhood of the origin and, moreover, $\delta_k = \mathcal{O}(\lambda^{-N-1})$. Therefore, the following function is analytic in a neighborhood of the origin:

$$F(\zeta) = \sum_{|k|\le m} u_k^* \, \delta_k(t, N, \zeta) \, \zeta^{-(N+1)},$$

where m is an arbitrary integer, and the u_k are arbitrary vectors of unit norm. For $t \le T$, the function $F(\lambda^{-1})$ is well defined for $|\lambda^{-1}| \le \varepsilon_N$, where

$$\varepsilon_N := \frac{R}{2\eta M N^{\alpha+2}} \ ,$$

so that the Maximum Principle can be applied on this disk. For $|\lambda^{-1}| = \varepsilon_N$, i.e., for $|\lambda|$ and N related like in (1.32) but with 2 instead of $\mathrm{e}^{\alpha+2}$ in the denominator, the bounds (1.33) and (1.34) are still valid.

For $t \le T$, we have $\|y^0(t) - x(0)\| \le R$ and Cauchy's estimates yield

$$\sum_{k\in\mathbb{Z}\times\mathbb{Z}} \Big\| \sum_{m\ge 0} \frac{1}{m!} \sum_{s(\alpha)=k, \alpha_i\neq 0} g^{(m)}(y^0(t)) \big(y^{\alpha_1}(t), \ldots, y^{\alpha_m}(t)\big) \Big\| \le \text{Const} \cdot M.$$

Again by (1.33) and (1.34), we obtain

$$\sum_{k\in\mathbb{Z}\times\mathbb{Z}} \|\ddot{y}^k + \Omega^2 y^k\| = \sum_{k\in\mathbb{Z}\times\mathbb{Z}} \|\ddot{z}^k + 2ik\omega\dot{z}^k - (k\omega)^2 z^k + \Omega^2 z^k\| \le \text{Const} \cdot M.$$

Putting this together, we obtain the bound

$$\sum_{k\in\mathbb{Z}\times\mathbb{Z}} \|\delta_k(t, N, \zeta)\| \le \text{Const} \cdot M \qquad \text{for} \quad |\zeta| = \varepsilon_N.$$

With the Maximum Principle, this gives for $|\lambda^{-1}| \le \varepsilon_N$

$$\begin{aligned} |F(\lambda^{-1})| &\le \max_{|\zeta|=\varepsilon_N} |F(\zeta)| \\ &\le \max_{|\zeta|=\varepsilon_N} \sum_{k\in\mathbb{Z}\times\mathbb{Z}} \|\delta_k(t, N, \zeta)\| \cdot \varepsilon_N^{-(N+1)} \le \text{Const} \cdot M \cdot \varepsilon_N^{-(N+1)}. \end{aligned}$$

Choosing now $u_k = \delta_k(t, N, \lambda^{-1}) / \|\delta_k(t, N, \lambda^{-1})\|$ in the definition of $F(\zeta)$ and letting $m \to \infty$ gives

$$\sum_{k\in\mathbb{Z}\times\mathbb{Z}} \|\delta_k(t, N, \lambda^{-1})\| \le \text{Const} \cdot M \cdot (\lambda\varepsilon_N)^{-(N+1)}.$$

For λ and N related by (1.32) we have $(\lambda\varepsilon_N)^{-1} \le 2/\mathrm{e}^{\alpha+2} = \mathrm{e}^{-\beta}$ with $\beta = \alpha + 2 - \ln 2 > 0$, so that in this case

$$\sum_{k\in\mathbb{Z}\times\mathbb{Z}} \|\delta_k(t)\| \le \text{Const} \cdot M \cdot \mathrm{e}^{-\beta(N+1)} \le \text{Const} \cdot M \cdot \mathrm{e}^{-\beta\lambda^{1/\alpha+2}}$$

holds with λ sufficiently large. □

We remark that, in the non-resonant case, we also have an exponentially small defect when the truncated solution is inserted into (1.1) so that the techniques of Chapter 2 can be applied to prove the near conservation of the two quantities (1.4) and (1.5) over exponentially long time intervals.

n-frequencies case. The n-frequencies case (1.1) follows with similar arguments. Here, we give the results concerning the bounds of the coefficient functions, the estimation of the defect and finally the result concerning the near conservation of the oscillatory energies. Like for (1.10) and (1.11), we obtain a system of differential equations for the variables $y_1, z_2, z_3, \ldots, z_n$ (where $z_i = z_i^{(0,\ldots,0,\underset{i}{1},0,\ldots,0)}$)

$$\ddot{y}_1 = \sum_{l\geq 0} \lambda^{-l} F_{1l}(\mathcal{Y}),\ \dot{z}_2^{(1,0,\ldots,0)} = \sum_{l\geq 1} \lambda^{-l} F_{2l}(\mathcal{Y}), \ldots,\ \dot{z}_n^{(0,\ldots,0,1)} = \sum_{l\geq 1} \lambda^{-l} F_{nl}(\mathcal{Y}) \tag{1.35}$$

and algebraic relations

$$z_i^k = \sum_{l\geq 0} \lambda^{-l} G_{il}^k(\mathcal{Y}), \tag{1.36}$$

where $\mathcal{Y} = (y_1, \dot{y}_1, z_2^{(1,0,\ldots,0)}, \ldots, z_n^{(0,\ldots,0,1)})$ and $k \in \mathbb{Z}^{n-1}$.

Like before, we fix a value $\mathcal{Y}_0 = (y_{10}, \dot{y}_{10}, 0, \ldots, 0)$, and we consider the complex ball

$$B_\rho(\mathcal{Y}_0) = \big\{(y_1, \dot{y}_1, z_2, \ldots, z_n) \,:\, \|y_1 - y_{10}\| \leq \rho R, \|\dot{y}_1 - \dot{y}_{10}\| \leq \rho M, \|z_i\| \leq \frac{\rho R}{2(n-1)}\big\}, \tag{1.37}$$

this permits us to bound the functions F_{ij} and G_{ij}^k.

Lemma 4.1.9 *Let $\mathcal{Y}_0 = (y_{10}, \dot{y}_{10}, 0, \ldots, 0)$ be given, and assume that (1.26) holds (for the n-dimensional case). The functions F_{ij} and G_{ij}^k satisfy*

$$\begin{aligned}
&\|F_{10}\|_1 \leq a_0 M, \qquad \|\dot{y}_1\|_1 \leq a_0 \tfrac{R}{2(n-1)},\\
&\|F_{1l}\|_{1-l\delta} \leq a_l M, \qquad \|F_{il}\|_{1-l\delta} \leq a_l \tfrac{R}{2(n-1)}, \qquad i = 2, \ldots, n\\
&\tfrac{1}{\mu_0}\Big(\|G_{20}^{(1,0,\ldots,0)}\|_1 + \|G_{20}^{(-1,0,\ldots,0)}\|_1 + \ldots + \|G_{n0}^{(0,\ldots,0,1)}\|_1 + \|G_{n0}^{(0,\ldots,0,-1)}\|_1\Big) \leq b_0 R,\\
&\max\Big(\sum_{k\in\mathbb{Z}^{n-1}} \frac{|k|^2}{\mu_l^{|k|}} \|G_{1l}^k\|_{1-l\delta}\ ,\ \sum_{k\in\mathbb{Z}^{n-1}} \frac{1+|k|^2}{\mu_l^{|k|}} \|G_{il}^k\|_{1-l\delta}\Big) \leq b_l R,\ 1 \leq l \leq L \text{ and } i = 2, \ldots, n
\end{aligned}$$

where $a_0 = \max(3^n, 2(n-1)(\|\dot{y}_{10}\|_1 + M)/R)$, $b_0 = 2$, and the generating functions $a(\zeta) = \sum_{l\geq 1} a_l \zeta^l$ and $b(\zeta) = \sum_{l\geq 1} b_l \zeta^l$ are implicitly given by

$$\begin{aligned}
a(\zeta) &= -3^n + 3^n\big(1 + \tfrac{(n-1)M\zeta}{R}\big)\big(1 - b(\zeta)\big)^{-n} + \tfrac{(n-1)\zeta}{\delta}\big(a_0 + a(\zeta)\big)a(\zeta),\\
b(\zeta) &= \tfrac{\alpha!}{\gamma^2}(2L)^{\alpha+1}\Big(\tfrac{3^n M\zeta^2}{R}\big(3 - b_0 - b(\zeta)\big)^{-n}\\
&+ \tfrac{2\max_{i=2,n}(a_i)\zeta}{\delta}\big(a_0 + a(\zeta)\big)\big(b_0 + b(\zeta)\big)\\
&+ \tfrac{\zeta^2}{\delta^2}\big(a_0 + a(\zeta)\big)^2\big(b_0 + b(\zeta)\big)\Big).
\end{aligned} \tag{1.38}$$

With $\alpha = [n + 2\nu] + 1$, γ and ν taken from (1.3).

These estimates are very similar to those given in Lemma 4.1.6 so we can also choose N, verifying (1.32), as the index of truncation for the series (1.35) and (1.36). We can now estimate the defect:

Theorem 4.1.10 *Consider the differential equation (1.1) with initial values $x(0)$ and $\dot{x}(0)$ satisfying (1.6). Assume that the nonlinearity $g(x)$ is analytic in a complex ball and bounded by M. Then, there exists $\beta > 0, T > 0$ and $\lambda_0 > 0$ such that the defect*

$$\delta_k(t) = \ddot{y}^k(t) + \Omega^2 y^k(t) - \sum_{m\geq 0} \frac{1}{m!} \sum_{s(\alpha)=k, \alpha_i \neq 0} g^{(m)}(y^0(t)) \big(y^{\alpha_1}(t), \ldots, y^{\alpha_m}(t)\big)$$

(here $\alpha_i, k, s(\alpha)$ are in $\mathbb{Z}^{n-1}$) satisfies for $0 \leq t \leq T$ and for $\lambda \geq \lambda_0$

$$\sum_{k\in\mathbb{Z}^{n-1}} \|\delta_k(t)\| \;\leq\; C\, M\, \mathrm{e}^{-\beta\lambda^{1/\alpha+2}}.$$

The constants C, γ, T, λ_0 only depend on an upper bound of M/R, on the a_i, and on the Diophantine constants γ, ν (like the constant α see Lemma 4.1.6) but not on λ.

We finish this section by mentioning the result concerning the conservation of the oscillatory energies (1.2).

Theorem 4.1.11 *Under the assumptions of Theorem 4.1.10, there exist positive constants γ, λ_0, C such that we have, for $j = 2, .., n$, $\lambda \geq \lambda_0$ and $0 \leq t \leq C\mathrm{e}^{\gamma\lambda}$*

$$I_j(x(t), \dot{x}(t)) = I_j(x(0), \dot{x}(0)) + \mathcal{O}(\lambda^{-1}). \tag{1.39}$$

4.1.3 Some nice pictures

To illustrate the near conservation of (1.4) and (1.5) for Hamiltonian problems of the form (1.1), we consider the following Hamiltonian (which is a modification of the Fermi-Pasta-Ulam problem encountered in the third chapter)

$$\begin{array}{rl} H(x_1, x_2, x_3, \dot{x}_1, \dot{x}_2, \dot{x}_3) = & \frac{1}{2}\sum_{i=1}^{5} \dot{x}_{1i}^2 + \sum_{i=1}^{3} \dot{x}_{2i}^2 + \sum_{i=1}^{2} \dot{x}_{3i}^2 \\ + & \frac{\omega_2^2}{2}\sum_{i=1}^{3} x_{2i}^2 + \frac{\omega_3^2}{2}\sum_{i=1}^{2} x_{3i}^2 + \frac{1}{4}(x_{11} - x_{21})^4 \\ + & \frac{1}{4}(-x_{15} - x_{32})^4 + \frac{1}{4}\sum_{i=1}^{2}(x_{1i+1} - x_{2i+1} - x_{1i} - x_{2i})^4 \\ + & \frac{1}{4}(x_{14} - x_{31} - x_{13} - x_{23})^4 + \frac{1}{4}(x_{15} - x_{32} - x_{14} - x_{31})^4, \end{array} \tag{1.40}$$

where $\omega_2 = \sqrt{2} \cdot 70$, $\omega_3 = 1 \cdot 70$ and the initial values are given by

$$\begin{array}{lllllllllll} x_{11} &=& 0.2, & x_{12} &=& 0.1, & x_{13} &=& 0.1, & x_{14} &=& 0.1, & x_{15} &=& 0.1, \\ x_{21} &=& \frac{0.3}{\omega_2}, & x_{22} &=& \frac{0.1}{\omega_2}, & x_{23} &=& \frac{0.2}{\omega_2}, & x_{31} &=& \frac{0.2}{\omega_3}, & x_{32} &=& \frac{0.1}{\omega_3}, \\ \dot{x}_{11} &=& 0.5, & \dot{x}_{12} &=& 1.5, & \dot{x}_{13} &=& 0.6, & \dot{x}_{14} &=& 0.7, & \dot{x}_{15} &=& 1., \\ \dot{x}_{21} &=& 0.6, & \dot{x}_{22} &=& 0.7, & \dot{x}_{23} &=& 1.1, & \dot{x}_{31} &=& 1.4, & \dot{x}_{32} &=& 0.7. \end{array}$$

This system consists in fact of 12 mass points joined by alternating soft nonlinear and stiff linear springs. The two end masses are fixed and the chain begins and ends with soft springs. Among the five stiff springs, the first three springs have a stiffness constant ω_2 while the two last have a constant ω_3.

To solve this problem, we use a trigonometric method (see Chapter 1 and [HLW02, Chap. XIII]) with a constant step size $h = 0.005$. For the filter functions, we take $\psi(\zeta) = \operatorname{sinc}(\zeta)^2, \psi_0(\zeta) = \cos(\zeta)\operatorname{sinc}(\zeta), \psi_1(\zeta) = \operatorname{sinc}(\zeta)$ and $\phi(\zeta) = 1$. The next figure shows the shifted Hamiltonian and both oscillatory energies (1.4)–(1.5) of the numerical solution.

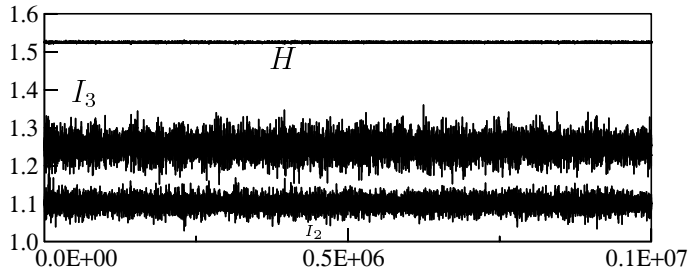


Figure 4.1: Experiment without resonance.

For the second experiment, we adapt Hamiltonian (1.40) to the 18 dimension vector

$$\Omega = \operatorname{diag}(0, \ldots, 0, \sqrt{2}\lambda, \sqrt{2}\lambda, \sqrt{5}\lambda, \sqrt{5}\lambda, \sqrt{5}\lambda, 3\lambda, 3\lambda, 3\lambda, 3\lambda)$$

where $\lambda = 100$. The initial values satisfy (1.6) and are given by

$$\begin{array}{lllllllllll}
x_{11} & = & 0.2, & x_{19} & = & 0.1, & x_{21} & = & \dfrac{0.98}{\omega_2}, & x_{22} & = & \dfrac{0.9}{\omega_2}, & x_{31} & = & \dfrac{1}{\omega_3}, \\
x_{32} & = & \dfrac{0.5}{\omega_3}, & x_{33} & = & \dfrac{0.4}{\omega_3}, & x_{41} & = & \dfrac{0.99}{\omega_4}, & x_{44} & = & \dfrac{1.6}{\omega_4}, & \dot{x}_{11} & = & 0.5, \\
\dot{x}_{12} & = & 0.1, & \dot{x}_{15} & = & 0.2, & \dot{x}_{19} & = & 0.01, & \dot{x}_{21} & = & 1, & \dot{x}_{31} & = & 1, \\
\dot{x}_{33} & = & 2, & \dot{x}_{41} & = & 1, & \dot{x}_{43} & = & 0.5, & \dot{x}_{44} & = & 0.1, & & & \\
\end{array}$$

and zero for the remaining initial values. Again we plot the three corresponding oscillatory energies and the shifted Hamiltonian, with the same method as above, over an interval of length one million.

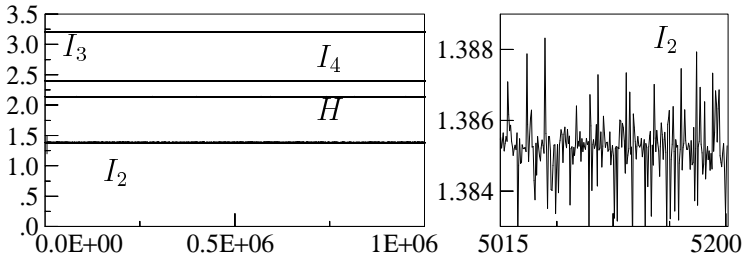


Figure 4.2: Near conservation of the oscillatory energies, with a zoom in the second component of the oscillatory energy I.

4.2 $(1,2)$-Case

This section is devoted to the 2-frequencies resonant case, more precisely we consider the matrix Ω in (1.1) with $a_2 = 1$, $a_3 = 2$. Our goal is to show that the quantities (1.4) and (1.5) are nearly conserved over long time, but not exponentially long time. To prove this fact we follow the same approach as in the non-resonant case, namely we search for the dominating terms in (1.8)–(1.9), give bounds and initial values for the functions of the truncated system and finally give an analogue of Theorem 4.1.11 (where we proved the near conservation of the two quantities mentioned above).

4.2.1 The dominating terms

Let us examine which couples of integers (k_2, k_3) in (1.8)–(1.9) give a differential equation for their corresponding function in the modulated Fourier expansion (1.7) :

- $k \cdot \omega = 0 \Leftrightarrow k_2\lambda + 2\,k_3\lambda = 0$. So if k_2 is even and $k_3 = -\frac{k_2}{2}$, we obtain second order differential equation for $z_1^{(k_2,k_3)}$.

- For the second component, we obtain a first order differential equation for $z_2^{(k_2,k_3)}$ if k_2 is odd and $k_3 = \frac{1-k_2}{2}$ or $k_3 = -\frac{1}{2}(1+k_2)$. Indeed, $|\omega_2^2 - (k\cdot\omega)^2| = 0 \Leftrightarrow 1 - k_2^2 - 4\,k_3^2 - 4\,k_2\,k_3 = 0 \Leftrightarrow k_3 = \frac{1-k_2}{2}$ or $k_3 = -\frac{1}{2}(1+k_2)$.

- The same calculations yield for the third component differential equations for $z_3^{(k_2,k_3)}$ if k_2 is even and $k_3 = 1 - \frac{k_2}{2}$ or $k_3 = -1 - \frac{k_2}{2}$.

Remark: We now have to choose the initial values for the systems of differential equation (1.8)–(1.9). The fact that we have too much freedom naturally motivate us to fix, for the k's mentioned above (but not for $y_1(0)$,$\dot{y}_1(0)$, $z_2(0)$ and $z_3(0)$), initial values $\dot{z}_1^k(0) = z_1^k(0) = 0$

and $z_j^k(0) = 0$ for $j = 2, 3$. Here we give a list of the couples of integers who give raise to the differential equations mentioned above.

1. For the first component

$k_2 =$	$\cdots$	-4	-2	0	2	4	$\cdots$
$k_3 =$	$\cdots$	2	1	0	-1	-2	$\cdots$

($+2$ over 0 and 2; $+1$ under 1 and 0)

Table 4.1: Couples of integers that lead to a second order differential equation for $z_1^{(k_2,k_3)}$.

2. For the second

$k_2 =$	$\cdots$	-3	-1	1	3	5	$\cdots$
$k_3 =$	$\cdots$	2	1	0	-1	-2	$\cdots$

($+2$ over 1 and 3; $+1$ under 1 and 0)

Table 4.2: Couples of integers that lead to a first order differential equation for $z_2^{(k_2,k_3)}$.

3. And finally for the third component

$k_2 =$	$\cdots$	-4	-2	0	2	4	$\cdots$
$k_3 =$	$\cdots$	3	2	1	0	-1	$\cdots$

($+2$ over 0 and 2; $+1$ under 2 and 1)

Table 4.3: Couples of integers that lead to a first order differential equation for $z_3^{(k_2,k_3)}$.

Thus we can express the modulation functions of the expansion (1.7) as functions of $\mathcal{Y} = (z_1^{k_1}, \dot{z}_1^{k_1}, z_2^{k_2}, z_3^{k_3})$ with k_j coming from the lists of the previous tables.

We remark that $|k_2 + 2\,k_3| = 0, 1$ or 2 for the k's of the previous lists, so that if we want to express a given couple of integers (n_2, n_3) with elements of this list, we need at least $\dfrac{|n_2 + 2\,n_3|}{2}$ terms of this list.

4.2.2 Bounds for the modulation functions

We have now the tools to estimate the functions appearing in the modulated Fourier expansions (1.7). To avoid an infinite system of differential equations for the functions of Table 4.1 to Table 4.3, we have to truncate this expansion. We therefore consider

$$\widetilde{x}(t) = y(t) + \sum_{0<|k|<N} \mathrm{e}^{\mathrm{i}k\cdot\omega t} z^k(t). \tag{2.1}$$

For the moment N is an arbitrary integer, we discuss the choice of this integer in the last theorem of this section.

Theorem 4.2.1 *Under the assumptions of Theorem 4.1.1, there exists a T such that, for $0 \le t \le T$ and λ sufficiently large, the defect $x(t) - \widetilde{x}(t)$ is of size $\mathcal{O}(\lambda^{-N})$. The functions $y = (y_1, y_2, y_3)$ and $z^k = (z_1^k, z_2^k, z_3^k)$ are bounded by*

$$\begin{aligned}
&y_1 = \mathcal{O}(1),\\
&z_2 = \mathcal{O}(\lambda^{-1}),\ z_3 = \mathcal{O}(\lambda^{-1}),\\
&\text{for } k \text{ in Table 4.1 to Table 4.3 : } z_1^k = \mathcal{O}(\lambda^{-|k|}),\ z_2^k = \mathcal{O}(\lambda^{-1-|k|}),\ z_3^k = \mathcal{O}(\lambda^{-1-|k|}),\\
&\text{for the other values of } k : z_j^k = \mathcal{O}(\lambda^{-2-|k|}),
\end{aligned} \tag{2.2}$$

where we have used the notation $|k| = |k_2| + |k_3|$ and the constants symbolized by $\mathcal{O}$ are independent of λ and t.

Proof. We follow the approach of the proof of Theorem 4.1. in [HL00]. As mentioned in the previous subsection, we have a system of differential equations

$$\begin{aligned}
\ddot{z}_1^k(t) &= \sum_{l=0}^{N} \lambda^{-l} F_{1l}^k(\mathcal{Y}), \quad \dot{z}_1^k(0) = z_1^k(0) = 0, \ \text{for } k \neq (0,0)\\
\dot{z}_j^k(t) &= \sum_{l=1}^{N} \lambda^{-l} F_{jl}^k(\mathcal{Y}), \quad z_j^k(0) = 0, \ \text{for } k \neq (\pm 1, 0), \text{ if } j = 2 \text{ and } k \neq (0, \pm 1), \text{ if } j = 3\\
\ddot{y}_1(t) &= \sum_{l=0}^{N} \lambda^{-l} F_{1l}^{(0,0)}(\mathcal{Y}),\\
\dot{z}_2^{(\pm 1,0)}(t) &= \sum_{l=1}^{N} \lambda^{-l} F_{2l}^{(\pm 1,0)}(\mathcal{Y}),\\
\dot{z}_3^{(0,\pm 1)}(t) &= \sum_{l=1}^{N} \lambda^{-l} F_{3l}^{(0,\pm 1)}(\mathcal{Y}),
\end{aligned} \tag{2.3}$$

for the couples of integers $k = (k_2, k_3)$ appearing in Tables 4.1 -4.3. We determine the remaining initial values for the system (2.3) such that $\widetilde{x}(0) = x(0)$ and $\dot{\widetilde{x}}(0) = \dot{x}(0)$. This

yields a system

$$\begin{aligned} x_1(0) &= y_1(0) + \mathcal{O}(\lambda^{-2}) \\ x_2(0) &= z_2(0) + \overline{z_2}(0) + \mathcal{O}(\lambda^{-2}) \\ x_3(0) &= z_3(0) + \overline{z_3}(0) + \mathcal{O}(\lambda^{-2}) \\ \dot{x}_1(0) &= \dot{y}_1(0) + \mathcal{O}(\lambda^{-1}) \\ \dot{x}_2(0) &= \mathrm{i}\omega_1 z_2(0) - \mathrm{i}\omega_1 \overline{z_2}(0) + \mathcal{O}(\lambda^{-1}) \\ \dot{x}_3(0) &= \mathrm{i}\omega_2 z_3(0) - \mathrm{i}\omega_2 \overline{z_3}(0) + \mathcal{O}(\lambda^{-1}), \end{aligned} \tag{2.4}$$

which gives, by the Implicit Function Theorem, locally unique initial values $y_1(0), \dot{y}_1(0)$, $z_2(0), z_3(0)$. Assumption (1.6) and the variation of constant formula imply that $z_2(t) = z_3(t) = \mathcal{O}(\lambda^{-1})$ on a compact interval $0 \leq t \leq T$. Looking closer at the last remark of the previous section, we see that k_2 factors $z_2^{(\pm 1,0)}$ and k_3 factors $z_3^{(0,\pm 1)}$ are contained in each terms of the right-hand side of the differential equations for z^k with $k = (k_2, k_3)$. This implies the bounds for the other functions in (2.3). The rest of the proof is now very similar to the one given in Theorem 4.1. of [HL00]. □

4.2.3 Near invariants

In this subsection we deal with the almost-invariants

$$\mathcal{I}_j(\mathbf{y}, \dot{\mathbf{y}}) \;=\; -\mathrm{i}\,\omega_j \sum_{0<|k|<N} k_j (y^{-k})^T \dot{y}^k, \text{ with } j = 2, 3,$$

where $\mathbf{y} = (..., y^{(-1,0)}, y^0, y^{(1,0)}, ...)$ is the sequence formed by the vectors $y^k = \mathrm{e}^{\mathrm{i}k\omega t} z^k$, with $|k| < N$, of (2.1) with initial values given by Theorem 4.2.1. Like in the non-resonant case, we have the following theorem.

Theorem 4.2.2 *Under the assumptions of Theorem 4.2.1 we have, for $j = 2, 3$, $0 \leq t \leq T$ and λ sufficiently large*

$$\begin{aligned} \mathcal{I}_j\big(\mathbf{y}(t), \dot{\mathbf{y}}(t)\big) &= \mathcal{I}_j\big(\mathbf{y}(0), \dot{\mathbf{y}}(0)\big) + \mathcal{O}(t\lambda^{-N}), \\ \mathcal{I}_j\big(\mathbf{y}(t), \dot{\mathbf{y}}(t)\big) &= I_j\big(x(t), \dot{x}(t)\big) + \mathcal{O}(\lambda^{-1}), \end{aligned}$$

where the constants symbolizing the $\mathcal{O}(\cdot)$ are independent of λ and t.

We are now able to prove the main result

Theorem 4.2.3 *Consider the differential equation (1.1) with $a_2 = 1, a_3 = 2$ and initial values $x(0)$ and $\dot{x}(0)$ satisfying (1.6). If the solution stays in a compact set, and under the assumptions of Theorem 4.2.1, if the integer N is larger than $a_2 + a_3 - 1$ then*

$$I_j(x(t), \dot{x}(t)) = I_j(x(0), \dot{x}(0)) + \mathcal{O}(\lambda^{-1}) + \mathcal{O}(t\lambda^{-a_2-a_3+1})$$

for $j = 2, 3$ and $0 \leq t \leq \mathrm{Const} \cdot \lambda^{a_2+a_3-1}$.

Proof. The beginning of the proof is exactly the same as the one given in [HL00, Cor. 4.4]. But now we have (with our notation)
$\mathcal{I}_j(\mathbf{y}_{l+1}(0), \dot{\mathbf{y}}_{l+1}(0)) - \mathcal{I}_j(\mathbf{y}_l(0), \dot{\mathbf{y}}_l(0)) = \mathcal{O}(\lambda^{-a_2-a_3+1})$ instead of $\mathcal{O}(\lambda^{-N})$ where $\mathbf{y}_l(t)$ denotes the vector of the modulated Fourier expansion terms that correspond to starting values $(x(lT), \dot{x}(lT))$.

This is due to the fact that our initial values for z_j^k, with the k taken from Table 4.1 to Table 4.3, are zero (see Remark 4.2.1) except for z_2, z_3 and y_1. Thus, the transition from $\mathbf{y}_l$ to $\mathbf{y}_{l+1}$ at the point $(l+1)T$ is not smooth and has a jump discontinuity of size $\mathcal{O}(\lambda^{-a_2-a_3})$. Moreover, one looses a factor λ because of the term $\dot{y}^k$ present in the definition of $\mathcal{I}_j(\mathbf{y}(t), \dot{\mathbf{y}}(t))$. Using repeatedly Theorem 4.2.2 yields the result. □

4.3 (a_2, a_3)-Case

The ideas described in the preceding section can be applied to the more general case of two integers a_2, a_3 with $1 \leq a_2 < a_3$ and $gcd(a_2, a_3) = 1$ (if there is a common factor in a_2 and a_3, we can put it in λ).

The list of integers k_2, k_3 who leads to differential equations (2.3) takes now the form:

1. For the first component

$$\begin{array}{lcccccccc} k_2 = & \cdots & -2a_3 & -a_3 & 0 & \overset{+a_3}{\frown} & a_3 & 2a_3 & \cdots \\ k_3 = & \cdots & 2a_2 & a_2 & \underset{+a_2}{\smile} & 0 & -a_2 & -2a_2 & \cdots \end{array}$$

Table 4.4: Couples of integers that lead to a second order differential equation for $z_1^{(k_2,k_3)}$.

2. For the second

$$k_2 = \cdots 1 - a_3 \quad \overset{+a_3}{\frown} 1 \quad 1 + a_3 \cdots$$

$$k_3 = \cdots \quad a_2 \quad \underset{+a_2}{\smile} 0 \quad -a_2 \quad \cdots$$

or

$$k_2 = \cdots -1 - a_3 \quad \overset{+a_3}{\frown} -1 \quad -1 + a_3 \quad \cdots$$

$$k_3 = \cdots \quad a_2 \quad \underset{+a_2}{\smile} 0 \quad -a_2 \quad \cdots$$

Table 4.5: Couples of integers that lead to a first order differential equation for $z_2^{(k_2,k_3)}$.

3. And finally for the third component

$$k_2 = \cdots \quad -a_3 \quad 0 \quad \overset{+a_3}{\frown} a_3 \quad \cdots$$

$$k_3 = \cdots \quad 1 + a_2 \quad \underset{+a_2}{\smile} 1 \quad 1 - a_2 \quad \cdots$$

or

$$k_2 = \cdots \quad -a_3 \quad 0 \quad \overset{+a_3}{\frown} a_3 \quad \cdots$$

$$k_3 = \cdots -1 + a_2 \quad \underset{+a_2}{\smile} -1 \quad -1 - a_2 \quad \cdots$$

Table 4.6: Couples of integers that lead to a first order differential equation for $z_3^{(k_2,k_3)}$.

We can remark that the bounds obtained in Theorem 4.2.1 are still valid for this case, and so the main result can now be stated

Theorem 4.3.1 *Consider the differential equation (1.1) with a_2, a_3 two prime integers with $1 \leq a_2 < a_3$ and initial values $x(0)$ and $\dot{x}(0)$ satisfying (1.6). Under the assumptions of Theorem 4.2.1, if the index of truncation N is larger than $a_2 + a_3 - 1$ then*

$$I_j(x(t), \dot{x}(t)) = I_j(x(0), \dot{x}(0)) + \mathcal{O}(\lambda^{-1}) + \mathcal{O}(t\lambda^{-a_2-a_3+1})$$

for $j = 2, 3$ and $0 \leq t \leq \text{Const} \cdot \lambda^{a_2+a_3-1}$.

Proof. We first prove that the minimum of $\{|k|; k \cdot a = 0, |k| \neq 0\}$ is obtained for $k_2 = a_3$ and $k_3 = -a_2$ and so $|k| = a_2 + a_3$. Indeed, if $k \cdot a = 0$ with $a_2 a_3 \neq 0$, we have $k_2 k_3 \neq 0$ and $\frac{a_2}{a_3} = -\frac{k_3}{k_2}$. The fraction $\frac{a_2}{a_3}$ is irreducible, so that $-k_3 = \mu a_2$ and $k_2 = \mu a_3$. The minimum is now obtained for $\mu = 1$ and we have $|k| = a_2 + a_3$. Like in the proof of Theorem 4.2.3, we have $\mathcal{I}_j(\mathbf{y}_{l+1}(0), \dot{\mathbf{y}}_{l+1}(0)) - \mathcal{I}_j(\mathbf{y}_l(0), \dot{\mathbf{y}}_l(0)) = \mathcal{O}(\lambda^{-a_2-a_3+1})$ and the rest of the proof is similar to the one given in Theorem 4.2.3. □

4.3.1 Some more pictures

To compare the results of Theorem 4.3.1 with the one given in the non-resonant case, we take the Hamiltonian (1.40) of Figure 4.1 Section 4.1.3, this time with $\omega_2 = 1 \cdot 70$, $\omega_3 = 2 \cdot 70$ and with the same initial values. To plot the various energies, we use the same method and same step size h than before. We see in Figure 4.3 that the oscillatory energies I_2 and I_3 are well conserved over long times, much longer than those predicted. It is not clear if the drift observed in Figure 4.3 is due to the $(1, 2)$-resonance or to round off errors.

Unfortunately, the drift in the oscillatory energies evoked in Theorem 4.2.3 does not appear in the time predicted. This is perhaps due to round off errors.

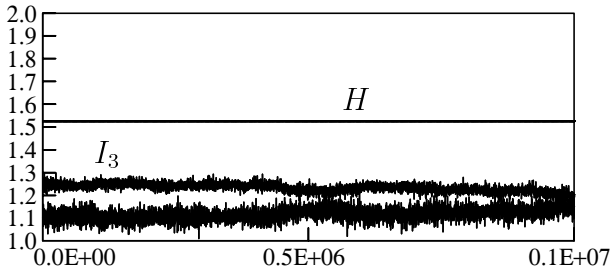


Figure 4.3: $(1, 2)$-resonant case

The second example, inspired by [CHL], is more explicit. Let's consider the Hamiltonian

$$H(x_1, x_2, x_3, \dot{x}_1, \dot{x}_2, \dot{x}_3) = \frac{1}{2} \sum_{i=1}^{3} \dot{x}_i^2 + \frac{\omega_2^2}{2} x_2^2 + \frac{\omega_3^2}{2} x_3^2 + (0.001 x_1 + x_2 + x_3)^4,$$

where $\omega_2 = 1 \cdot 70$, $\omega_3 = 2 \cdot 70$. For the initial values, we take

$$\begin{array}{rclrclrcl} x_1 & = & 1.0, & x_2 & = & \frac{0.8}{70}, & x_3 & = & \frac{0.7}{70}, \\ \dot{x}_1 & = & -0.75, & \dot{x}_2 & = & 0.6, & \dot{x}_3 & = & 0.85. \end{array}$$

This time, we use a very precise numerical method called DOP853 (for a definition, see [HNW93]) and plot all the oscillatory energies. This time a clear drift can be observed already for $t \geq 2000$. We can also see the good conservation of the total oscillatory energy.

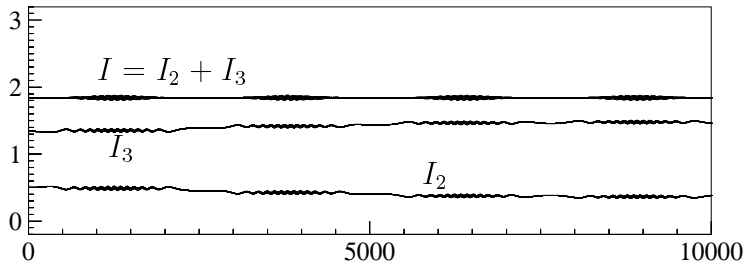


Figure 4.4: $(1, 2)$-resonant case, with drift.

4.4 $(a_2, a_3, \ldots, a_n)$-Case

In this section we consider the most general case $(a_2, \ldots, a_n)$ where the real numbers a_j for $j = 2, \ldots, n$ satisfy $1 \leq a_2 < \ldots < a_n$. The goal is to search for integer vectors k with minimal norm, such that $k \cdot a = 0$ or $k \cdot a = \pm a_i$ with $i = 2, \ldots, n$. To do this, we define the positive integer $m(a) = \min\{|k|; k \cdot a = 0, |k| \neq 0\}$.

For this integer, we have the bound $3 \leq m(a)$.

Indeed, this is obtained by the fact that if $|k| = 1$ or 2 with $k \cdot a = 0$, we have $a_i = 0$ or $a_i = a_j$ for some i, j, this contradicts our hypothesis on the vector a.

In the special case where the a_i are integers with $gcd(a_2, \ldots, a_n) = 1$, we also obtain an upper bound $m(a) \leq \min_{ij} \frac{a_i + a_j}{gcd(a_i, a_j)}$. To show this upper bound, we take k with all its components equal to zero except two components, we now are in the case discussed in the last section and we have $m(a) \leq \min_{ij} \frac{a_i + a_j}{gcd(a_i, a_j)}$.

Let's look at some examples :

Example: If $a = (1, 5, 6)$ then $m(a) = 3$ with for example $k = (1, 1, -1)$. Taking the third component greater $a = (1, 5, 21)$ we have $m(a) = 6$ with $k = (-5, 1, 0)$. For real a_i, one could take for example $a = (1, \sqrt{2}, 1 + \sqrt{2})$ and obtain $m(a) = 3$ with the choice $k = (1, -1, 1)$. For the vector $a = (4, 52, 2807, 2902, 6005)$ in $\mathbb{R}^5$, we have the $m(a) = 13$ obtained with $k = (-4, 6, 1, 1, -1)$.

We know that the integer $m(a)$ exists and that $3 \leq m(a)$, so that for a non trivial solution of $k \cdot a = 0$, we have a non trivial solution in the other cases $k \cdot a = a_i$ with $|k| = m(a) - 1$ at worst.

Like in the preceding section, we obtain a system of differential equation for the values of k such that $k \cdot a = 0$, $k \cdot a = \pm a_i$, $i = 2, \ldots, n$. We take all the initial values equal to zero except the ones for $y_1, z_2^{(\pm 1, 0, \ldots, 0)}, \ldots, z_n^{(0, \ldots, 0, \pm 1)}$. As before, we define

$$\widetilde{x}(t) = y(t) + \sum_{0 < |k| < N} \mathrm{e}^{\mathrm{i} k \cdot \omega t} z^k(t), \tag{4.1}$$

and the proof of Theorem 4.2.1 can be adapted to the more general case to yield

Theorem 4.4.1 *Under the assumptions of Theorem 4.1.1, there exists a T such that, for $0 \leq t \leq T$ and λ sufficiently large, the error $x(t) - \widetilde{x}(t)$ is of size $\mathcal{O}(\lambda^{-N})$. The functions*

$y = (y_1, \ldots, y_n)$ and $z^k = (z_1^k, \ldots, z_n^k)$ are bounded by

$$\begin{array}{l} y_1 = \mathcal{O}(1), \\ z_2^{(1,0,\ldots,0)} = \mathcal{O}(\lambda^{-1}), \ \ldots, \ z_n^{(0,\ldots,0,1)} = \mathcal{O}(\lambda^{-1}), \\ \text{for } k \text{ such that } k \cdot a = 0, \ k \cdot a = a_i : z_1^k = \mathcal{O}(\lambda^{-|k|}), \ z_i^k = \mathcal{O}(\lambda^{-1-|k|}), \ i = 2, \ldots, n, \\ \text{for the other values of } k : z_j^k = \mathcal{O}(\lambda^{-2-|k|}), \end{array} \tag{4.2}$$

where we have used the notation $|k| = |k_2| + \ldots + |k_n|$ and the constants symbolized by $\mathcal{O}$ are independent of λ and t.

We can now show that the quantities (1.2) are well conserved on time interval of length $\mathcal{O}(\lambda^{m(a)-1})$.

Theorem 4.4.2 *Consider the differential equation (1.1) with $1 \leq a_2 < \ldots < a_n$ and initial values $x(0)$ and $\dot{x}(0)$ satisfying (1.6). If the solution stays in a compact set, and under the assumptions of Theorem 4.4.1, if the index of truncation N is larger than $m(a) - 1$ then*

$$I_j(x(t), \dot{x}(t)) = I_j(x(0), \dot{x}(0)) + \mathcal{O}(\lambda^{-1}) + \mathcal{O}(t\lambda^{-m(a)+1})$$

for $j = 2, \ldots, n$ and $0 \leq t \leq \text{Const} \cdot \lambda^{m(a)-1}$.

Proof. With our choice for the initial values of the system of differential equation, the transition from the functions $\mathbf{y}_l$ to $\mathbf{y}_{l+1}$ (defined in Theorem 4.2.3) at the point $(l+1)T$ is not smooth and as a jump discontinuity of size $\mathcal{O}(\lambda^{-m(a)})$. So that we have $\mathcal{I}_j(\mathbf{y}_{l+1}(0), \dot{\mathbf{y}}_{l+1}(0)) - \mathcal{I}_j(\mathbf{y}_l(0), \dot{\mathbf{y}}_l(0)) = \mathcal{O}(\lambda^{-m(a)+1})$. We now use repeatedly the results of Theorem 4.2.2 adapted to the n dimensional case to conclude this proof. □

To conclude with the analysis of the exact solution of (1.1), we want to mention that this approach is different from the one given in [CHL]. Let us see what are the advantages and disadvantages of these approaches. The major disadvantage of our approach is that we have too much freedom in the choice of the initial values of the system that determines the modulation functions. We arbitrarily set these initial values to zero. On the other hand, we have adiabatic invariants for every frequency in this system. The advantage of the approach given in [CHL] is that we do not have this freedom for the choice of the initial values by considering the frequency modulus. However, we have less adiabatic invariants in the system of the modulation functions.

4.5 Numerical solution

The techniques used during this chapter for the analysis of the exact solution of (1.1) are also applicable to the treatment of the numerical solution of Hamiltonian problems with Hamiltonian (1.17). We just give the main result without proof. For a proof with the other techniques, see [CHL].

We consider the class of trigonometric methods (see Chapter 1 and [HLW02, Chap. XIII]). For a step size h, a two-step formulation of these methods is given by

$$x_{m+1} - 2\cos(h\Omega)x_m + x_{m-1} = h^2\Psi g(\Phi x_m), \tag{5.1}$$

where subscripts refer to the time step. Here, $\Psi = \psi(h\Omega)$ and $\Phi = \phi(h\Omega)$, where the filter functions ψ and ϕ are real-valued bounded functions with $\psi(0) = \phi(0) = 1$. We have a velocity approximation

$$2h\ \mathrm{sinc}(h\Omega)\dot{x}_m = x_{m+1} - x_{m-1} \tag{5.2}$$

if $\mathrm{sinc}(h\Omega)$ (where $\mathrm{sinc}(\xi) = \sin(\xi)/\xi$ is invertible. We make the following assumptions (see [CHL]):

- The energy of the initial values is bounded independently of λ,

$$\frac{1}{2}\|\dot{x}(0)\|^2 + \frac{1}{2}\|\Omega x(0)\|^2 \le E\ . \tag{5.3}$$

- The numerical solution values Φx_m stay in a compact subset of a domain on which the potential U is smooth.

- We impose a lower bound on the step size: $h\lambda \ge c_0 > 0$.

- We assume the numerical non-resonance condition

$$\Big|\sin\Big(\frac{h}{2}\,k\cdot\omega\Big)\Big| \ \ge\ c\sqrt{h} \quad \text{for all } k \in \mathbb{Z}^{n-1} \setminus \mathcal{M} \text{ with } |k| \le N, \tag{5.4}$$

  for some $N \ge 2$ and $c > 0$. Where $\mathcal{M} = \{k \in \mathbb{Z}^{n-1} : k_2a_2 + \ldots + k_na_n = 0\}$.

- The filter functions $\psi(\xi)$ satisfies, with $\xi_j = h\omega_j = ha_j\lambda$,

$$|\psi(\xi_j)| \ \le\ C\,\Big|\mathrm{sinc}\big(\tfrac{1}{2}\,\xi_j\big)\Big| \quad \text{for} \ \ j = 2,\ldots,n, \tag{5.5}$$

  and

$$|\psi(\xi_j)| \ \le\ C\ \mathrm{sinc}^2\big(\tfrac{1}{2}\,\xi_j\big), \tag{5.6}$$

$$|\psi(\xi_j)| \ \le\ C\,|\phi(\xi_j)| \qquad \text{for} \ \ j = 2,\ldots,n. \tag{5.7}$$

It is then possible to give the numerical analogue of Theorem 4.1.1, the numerical solution also has a modulated Fourier expansion, namely

$$x_m = y_h(t) + \sum_{0<|k|<N} \mathrm{e}^{\mathrm{i}k\cdot\omega t} z_h^k(t), \tag{5.8}$$

where $t = mh$ and $k \in \mathbb{Z}^{n-1}$. Once again, the modulation functions z_h^k (we note $z_h^0 = y_h$) have formal invariants that are related to the Hamiltonian H (see (1.17)) and to the oscillatory energies I_j (see (1.2)). We mention the results in the following theorem

Theorem 4.5.1 *Under the above conditions (5.3)–(5.7), the numerical solution obtained by the method (5.1)–(5.2) satisfies*

$$\begin{aligned} H(x_m, \dot{x}_m) &= H(x_0, \dot{x}_0) + \mathcal{O}(h) \\ I_j(x_m, \dot{x}_m) &= I_j(x_0, \dot{x}_0) + \mathcal{O}(h) \quad \textit{for} \quad j = 2, \ldots, n, \end{aligned}$$

for $0 \leq mh \leq \text{Const} \cdot \min(\lambda^{m(a)-1}, h^N)$. *The constants symbolized by* $\mathcal{O}$ *are independent of* n, h, λ, a_j *satisfying the above conditions, but depend on* N *and the constants in the conditions.*

Chapter 5

Another type of oscillatory Hamiltonian systems

In this chapter, we enlarge the class of highly oscillatory Hamiltonian systems considered by generalizing the Hamiltonian encountered in the second chapter. We begin this chapter by studying an interesting example, then we give the modulated Fourier expansion of the exact solution for the class of problems considered. We then look at certain invariants. The last four sections deal with numerical methods and the modulated Fourier expansion for the numerical solution. This allows us to prove near-energy conservation also in this setting.

5.1 Introduction

We adapt the technique of the modulated Fourier expansion to the more general Hamiltonian

$$H(p,q) = K(p_1,q) + \frac{1}{2}p_2^T p_2 + \frac{\omega^2}{2}q_2^T q_2. \tag{1.1}$$

We denote the variables $p = (p_1, p_2)$ and $q = (q_1, q_2)$ according to the partitioning of the square matrix

$$\Omega = \begin{pmatrix} 0 & 0 \\ 0 & \omega I \end{pmatrix}, \quad \omega \gg 1,$$

with blocks of arbitrary dimension.

Again, the initial values are assumed to satisfy

$$\frac{1}{2}||p(0)||^2 + \frac{1}{2}||\Omega q(0)||^2 \leq E. \tag{1.2}$$

By taking the function K to be $\frac{1}{2}p_1^T p_1 + U(q)$, we get the Hamiltonian (1.1) of the second chapter. With function (1.1) it is even possible to consider coupling between the

position q and the momenta p_1. For example, one can take $K(p_1, q) = \frac{1}{2} p_1^T M(q) p_1$, with a mass matrix $M(q)$. Such problems often occur in physics.

As a concrete example, we consider the stiff spring pendulum in polar coordinates as described in [AR99a]. In this case, the Hamiltonian reads

$$H(p, q) = \frac{1}{2}(p_2^2 + (q_2 + 1)^{-2} p_1^2 + q_1^2 + \omega^2 q_2^2), \tag{1.3}$$

with a large parameter ω. Here, the fast component q_2 represents the displacement of the mass connected to the spring around the equilibrium circle of radius 1. The slow component q_1 corresponds to the angle of the pendulum.

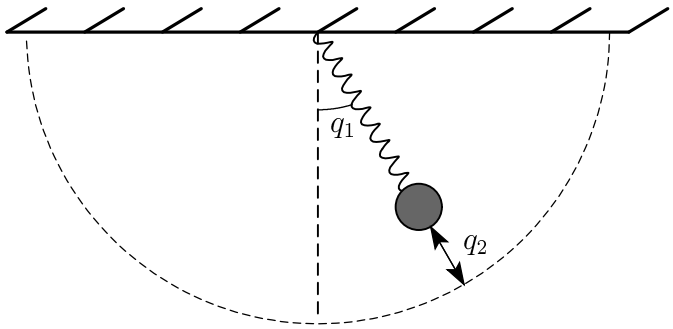


Figure 5.1: Stiff spring pendulum.

Let's use DOP853 (for a definition of this numerical method, see [HNW93]) and plot the Hamiltonian and the oscillatory energy

$$I(p, q) = \frac{1}{2}||p_2||^2 + \frac{\omega^2}{2}||q_2||^2 \tag{1.4}$$

of problem (1.3) over a time interval of length 100. For this experiment, we take for the initial values $p_1(0) = -\frac{1}{\sqrt{2}}, p_2(0) = \frac{1}{\sqrt{2}}, q_1(0) = 0, q_2(0) = 0$, and $\omega = 80$.

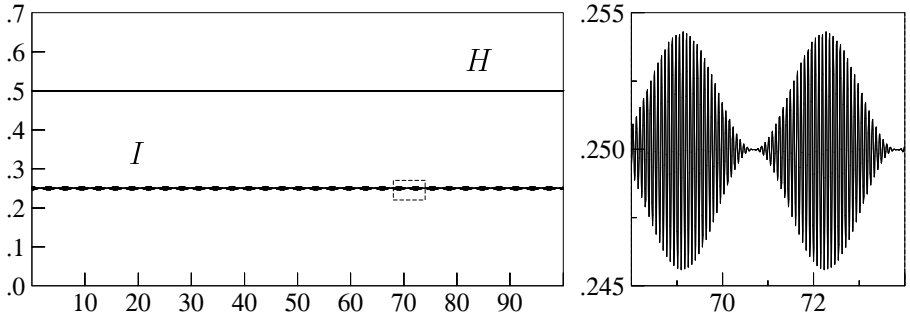


Figure 5.2: Hamiltonian and oscillatory energies for (1.3).

This example illustrates that for a Hamiltonian problem of the form (1.1), the oscillatory energy is still a near-invariant. Our goal in the next sections is to explain this behaviour.

5.2 Expansion of the exact solution

We first decompose the exact solution of (1.1) into a modulated Fourier expansion and estimate the remainder. Then, we give two almost-invariants for the coefficient functions of this expansion. They are related to the Hamiltonian (1.1) and to the oscillatory energy (1.4).

To do this, we follow the lines of [HLW02, Sect. XIII.5.1] and write the equations of motion for the Hamiltonian (1.1)

$$\begin{aligned} \dot{p}_1 &= -\nabla_{q_1} K(p_1, q) \\ \dot{p}_2 &= -\omega^2 q_2 - \nabla_{q_2} K(p_1, q) \\ \dot{q}_1 &= \nabla_{p_1} K(p_1, q) \\ \dot{q}_2 &= p_2, \end{aligned} \tag{2.1}$$

or briefly

$$\begin{aligned} \dot{p} &= -\Omega^2 q + g(p_1, q) \\ \dot{q} &= \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix} p + h(p_1, q), \end{aligned} \tag{2.2}$$

with analytical smooth functions g and h. We then expand p and q as an asymptotic series and estimate the remainder when this expansion is inserted into (2.2). This will be discussed in the following theorem

Theorem 5.2.1 *If the solution $(p(t), q(t))$ of (1.1) satisfies the condition (1.2) and stays in a compact set K for $0 \le t \le T$, then the solution admits an expansion*

$$\begin{aligned} p(t) &= \sum_{|k|<N} \mathrm{e}^{\mathrm{i}k\omega t} \eta^k(t) + R_N(t), \\ q(t) &= \sum_{|k|<N} \mathrm{e}^{\mathrm{i}k\omega t} \zeta^k(t) + S_N(t), \end{aligned} \tag{2.3}$$

for arbitrary $N \ge 2$, where the remainder terms are bounded by

$$R_N(t) = \mathcal{O}(\omega^{-N}), \quad S_N(t) = \mathcal{O}(\omega^{-N}), \qquad \text{for} \qquad 0 \le t \le T\ . \tag{2.4}$$

The real functions $\eta = \eta^0 = (\eta_1, \eta_2), \zeta = \zeta^0 = (\zeta_1, \zeta_2)$ and the complex functions $\eta^k = (\eta_1^k, \eta_2^k), \zeta^k = (\zeta_1^k, \zeta_2^k)$ are bounded, together with all their derivatives, by

$$\begin{aligned} &\zeta_1 = \mathcal{O}(1), && \eta_1 = \mathcal{O}(1), && \zeta_2 = \mathcal{O}(\omega^{-2}), && \eta_2 = \mathcal{O}(\omega^{-2}), \\ &\zeta_1^1 = \mathcal{O}(\omega^{-2}), && \eta_1^1 = \mathcal{O}(\omega^{-2}), && \zeta_2^1 = \mathcal{O}(\omega^{-1}), && \dot{\eta}_2^1 = \mathcal{O}(\omega^{-1}), \\ &\zeta_1^k = \mathcal{O}(\omega^{-k-1}), && \eta_1^k = \mathcal{O}(\omega^{-k-1}), && \zeta_2^k = \mathcal{O}(\omega^{-k-2}), && \eta_2^k = \mathcal{O}(\omega^{-k-1}), \end{aligned} \tag{2.5}$$

for $k = 2, \dots, N-1$. Moreover, we have $\eta^{-k} = \overline{\eta^k}$ and $\zeta^{-k} = \overline{\zeta^k}$. These functions are unique up to terms of size $\mathcal{O}(\omega^{-N})$. The constants symbolized by the $\mathcal{O}$-notation are independent of ω and t with $0 \le t \le T$ but depend on N, T and E.

Proof. To determine the smooth functions $\eta, \eta^1, \dots, \eta^{N-1}$ and $\zeta, \zeta^1, \dots, \zeta^{N-1}$, we put

$$\begin{aligned} p_*(t) &= \eta(t) + \sum_{0<|k|<N} \mathrm{e}^{\mathrm{i}k\omega t}\eta^k(t) \\ q_*(t) &= \zeta(t) + \sum_{0<|k|<N} \mathrm{e}^{\mathrm{i}k\omega t}\zeta^k(t), \end{aligned} \tag{2.6}$$

insert these functions into (2.2), expand the nonlinearity functions g and h around $(\eta_1(t), \zeta(t))$ and compare the coefficients of $\mathrm{e}^{\mathrm{i}k\omega t}$.

With the notation $D_1^m D_2^n g(\eta_1, \zeta)(\eta_1^\alpha, \zeta^\beta) = D_1^m D_2^n g(\eta_1, \zeta)(\eta_1^{\alpha_1}, \dots, \eta_1^{\alpha_m}, \zeta^{\beta_1}, \dots, \zeta^{\beta_n})$ for multi-indices $\alpha = (\alpha_1, \dots, \alpha_m)$ and $\beta = (\beta_1, \dots, \beta_n)$ (we adopt a similar notation for the function h), we obtain the following system of differential equations:

$$\begin{pmatrix} \dot\eta_1 \\ \omega^2\zeta_2 \end{pmatrix} + \begin{pmatrix} 0 \\ \dot\eta_2 \end{pmatrix} = g(\eta_1, \zeta) + \sum_{s(\alpha)+s(\beta)=0} \frac{1}{m!n!} D_1^m D_2^n g(\eta_1, \zeta)(\eta_1^\alpha, \zeta^\beta) \tag{2.7}$$

$$\begin{pmatrix} \dot\zeta_1 \\ \dot\zeta_2 \end{pmatrix} = \begin{pmatrix} h_1(\eta_1, \zeta) + \sum_{s(\alpha)+s(\beta)=0} \frac{1}{m!n!} D_1^m D_2^n h_1(\eta_1, \zeta)(\eta_1^\alpha, \zeta^\beta) \\ \eta_2 \end{pmatrix} \tag{2.8}$$

$$\begin{pmatrix} \mathrm{i}k\omega\eta_1^k \\ \omega^2\zeta_2^k \end{pmatrix} + \begin{pmatrix} \dot\eta_1^k \\ \mathrm{i}k\omega\eta_2^k + \dot\eta_2^k \end{pmatrix} = \sum_{s(\alpha)+s(\beta)=k} \frac{1}{m!n!} D_1^m D_2^n g(\eta_1, \zeta)(\eta_1^\alpha, \zeta^\beta) \tag{2.9}$$

$$\begin{pmatrix} \mathrm{i}k\omega\zeta_1^k \\ \mathrm{i}k\omega\zeta_2^k \end{pmatrix} + \begin{pmatrix} \dot\zeta_1^k \\ \dot\zeta_2^k \end{pmatrix} = \begin{pmatrix} 0 \\ \eta_2^k \end{pmatrix} + \sum_{s(\alpha)+s(\beta)=k} \frac{1}{m!n!} D_1^m D_2^n h(\eta_1, \zeta)(\eta_1^\alpha, \zeta^\beta). \tag{2.10}$$

Here the sums range over all $m, n \ge 0$ and all multi-indices $\alpha = (\alpha_1, \dots, \alpha_m)$, $\beta = (\beta_1, \dots, \beta_n)$ with integers α_i, β_i satisfying $0 < |\alpha_i|, |\beta_i| < N$, which have a given sum $s(\alpha) + s(\beta) = \sum_{j=1}^m \alpha_j + \sum_{j=1}^n \beta_j$.

For large ω, the dominating terms in these differential equations are given by the left-most expressions. We are interested in smooth functions $\eta, \zeta, \eta^k, \zeta^k$ that satisfy the system up to a defect of size $\mathcal{O}(\omega^{-N})$. However, their higher derivatives make difficulties, we remove these terms by using iteratively the differential equations (2.7)–(2.10). This leads

to a system

$$\begin{aligned}
\dot\zeta_1 &= \mathcal{F}_1(\eta,\zeta,\eta^k,\zeta^k,\omega^{-1}), & \dot\eta_1 &= \tilde{\mathcal{F}}_1(\eta,\zeta,\eta^k,\zeta^k,\omega^{-1}),\\
\zeta_2 &= \omega^{-2}\mathcal{G}_2(\eta,\zeta,\eta^k,\zeta^k,\omega^{-1}), & \eta_2 &= \omega^{-2}\tilde{\mathcal{G}}_2(\eta,\zeta,\eta^k,\zeta^k,\omega^{-1}),\\
\zeta_1^k &= \omega^{-1}\mathcal{G}_1^k(\eta,\zeta,\eta^k,\zeta^k,\omega^{-1}), & \eta_1^k &= \omega^{-1}\tilde{\mathcal{G}}_1^k(\eta,\zeta,\eta^k,\zeta^k,\omega^{-1}),\\
\zeta_2^1 &= \omega^{-1}\mathcal{G}_2^1(\eta,\zeta,\eta^k,\zeta^k,\omega^{-1}), & \dot\eta_2^1 &= \omega^{-1}\tilde{\mathcal{F}}_2^1(\eta,\zeta,\eta^k,\zeta^k,\omega^{-1}),\\
\zeta_2^k &= \omega^{-2}\mathcal{G}_2^k(\eta,\zeta,\eta^k,\zeta^k,\omega^{-1}), & \eta_2^k &= \omega^{-2}\tilde{\mathcal{G}}_2^k(\eta,\zeta,\eta^k,\zeta^k,\omega^{-1}),
\end{aligned} \tag{2.11}$$

where $\mathcal{F}_j, \mathcal{G}_j, \mathcal{G}_j^k, \tilde{\mathcal{F}}_j, \tilde{\mathcal{G}}_j, \tilde{\mathcal{G}}_j^k$ are formal series in powers of ω^{-1}. Since we get formal algebraic relations for $\zeta_2, \eta_2, \zeta_1^k, \eta_1^k, \zeta_2^1, \zeta_2^k, \eta_2^k$, we can further eliminate these variables in the series (2.11). We finally obtain for $\eta_2, \zeta_2, \zeta_1^k, \eta_1^k, \zeta_2^1, \zeta_2^k, \eta_2^k$ the algebraic relations

$$\begin{aligned}
\zeta_1^k &= \omega^{-1}\big(G_{10}^k(\eta_1,\zeta_1,\eta_2^1) + \omega^{-1}G_{11}^k(\eta_1,\zeta_1,\eta_2^1) + \ldots\big)\\
\eta_1^k &= \omega^{-1}\big(\tilde G_{10}^k(\eta_1,\zeta_1,\eta_2^1) + \omega^{-1}\tilde G_{11}^k(\eta_1,\zeta_1,\eta_2^1) + \ldots\big)\\
\zeta_2 &= \omega^{-2}\big(G_{20}(\eta_1,\zeta_1,\eta_2^1) + \omega^{-1}G_{21}(\eta_1,\zeta_1,\eta_2^1) + \ldots\big)\\
\eta_2 &= \omega^{-2}\big(\tilde G_{20}(\eta_1,\zeta_1,\eta_2^1) + \omega^{-1}\tilde G_{21}(\eta_1,\zeta_1,\eta_2^1) + \ldots\big)\\
\zeta_2^1 &= \omega^{-1}\big(G_{20}^1(\eta_1,\zeta_1,\eta_2^1) + \omega^{-1}G_{21}^1(\eta_1,\zeta_1,\eta_2^1) + \ldots\big)\\
\zeta_2^k &= \omega^{-2}\big(G_{20}^k(\eta_1,\zeta_1,\eta_2^1) + \omega^{-1}G_{21}^k(\eta_1,\zeta_1,\eta_2^1) + \ldots\big)\\
\eta_2^k &= \omega^{-1}\big(\tilde G_{20}^k(\eta_1,\zeta_1,\eta_2^1) + \omega^{-1}\tilde G_{21}^k(\eta_1,\zeta_1,\eta_2^1) + \ldots\big),
\end{aligned} \tag{2.12}$$

and a system of real first-order differential equations for η_1, ζ_1 and complex first-order differential equations as follows for η_2^1:

$$\begin{aligned}
\dot\zeta_1 &= F_{10}(\eta_1,\zeta_1,\eta_2^1) + \omega^{-1}F_{11}(\eta_1,\zeta_1,\eta_2^1) + \ldots\\
\dot\eta_1 &= \tilde F_{10}(\eta_1,\zeta_1,\eta_2^1) + \omega^{-1}\tilde F_{11}(\eta_1,\zeta_1,\eta_2^1) + \ldots\\
\dot\eta_2^1 &= \omega^{-1}\big(\tilde F_{20}^1(\eta_1,\zeta_1,\eta_2^1) + \omega^{-1}\tilde F_{21}^1(\eta_1,\zeta_1,\eta_2^1) + \ldots\big).
\end{aligned} \tag{2.13}$$

The series present in the ansatz (2.12)–(2.13) usually diverge, we thus truncate this ansatz after the $\mathcal{O}(\omega^{-N})$ terms. Inserting this ansatz and its first derivative into equations (2.7)–(2.10) and comparing like powers of ω^{-1} yields recurrence relations for the functions $F_{jl}^k, G_{jl}^k, \tilde F_{jl}^k, \tilde G_{jl}^k$. This shows that these functions together with their derivatives are all bounded on compact sets.

Next, we determine initial values for the differential equations (2.13) such that the functions $p_*(t), q_*(t)$ of (2.6) satisfy $p_*(0) = p(0)$ and $q_*(0) = q(0)$ (where $p(0)$ and $q(0)$

are the initial values for the system (1.1)). Because of the special structure of the ansatz (2.12)–(2.13), this gives a system

$$\begin{aligned} p_1(0) &= \eta_1(0) + \mathcal{O}(\omega^{-2}) \\ p_2(0) &= 2Re(\eta_2^1(0)) + \mathcal{O}(\omega^{-2}) \\ q_1(0) &= \zeta_1(0) + \mathcal{O}(\omega^{-2}) \\ q_2(0) &= 2\omega^{-1} Im(\eta_2^1(0)) + \mathcal{O}(\omega^{-2}), \end{aligned} \tag{2.14}$$

which, by the implicit function theorem, yields (locally) unique initial values $\eta_1(0)$, $\zeta_1(0)$, $\eta_2^1(0)$. The assumption (1.2) on the initial values implies that $\eta_2^1(0) = \mathcal{O}(1)$. It further follows from the boundedness of $\tilde{F}_{2l}^1$ that $\eta_2^1(t) = \mathcal{O}(1)$ for $0 \le t \le T$. By looking closer at the structure of the function $F_{jl}^k, \tilde{F}_{jl}^k, G_{jl}^k, \tilde{G}_{jl}^k$ it can be seen that it contains at least k times the factors ζ^1 or η_1^1. This implies the stated bounds for all other functions.

We still have to estimate the remainders $R_N(t) = p(t) - p_*(t)$ and $S_N(t) = q(t) - q_*(t)$. To do this, we consider the solution of (2.12)–(2.13) with initial values (2.14). By construction, these functions satisfy the system (2.7–(2.10) up to a defect of $\mathcal{O}(\omega^{-N})$. This gives a defect of size $\mathcal{O}(\omega^{-N})$, when the functions $p_*(t)$ and $q_*(t)$ of (2.6) are inserted into (2.2). Hence on a finite time interval $0 \le t \le T$, we obtain $R_N(t) = \mathcal{O}(\omega^{-N})$ and $S_N(t) = \mathcal{O}(\omega^{-N})$. □

5.3 Two almost-invariants of the modulated Fourier expansion

In this section, we show that the system for the modulation functions of the expansion of the exact solution has two formal invariants. As we said before, one is related to the Hamiltonian (1.1). This almost-invariant is defined by

$$\mathcal{H}(\mathbf{p}, \mathbf{q}) = \frac{1}{2} \sum_{|k|<N} \left((q^{-k})^T \Omega^2 q^k + (p_2^{-k})^T p_2^k \right) + \mathcal{K}(\mathbf{p}_1, \mathbf{q}), \tag{3.1}$$

where $\mathbf{p} = (p^{-N+1}, \dots, p^0, \dots, p^{N-1})$ and $p^k = \mathrm{e}^{\mathrm{i}k\omega t}\eta^k(t)$ (the same notation is used for $\mathbf{q}$), and

$$\mathcal{K}(\mathbf{p}_1, \mathbf{q}) = K(p_1^0, q^0) + \sum_{s(\alpha)+s(\beta)=0} \frac{1}{m!n!} D_1^m D_2^n K(p_1^0, q^0)(\mathbf{p}_1^\alpha, \mathbf{q}^\beta). \tag{3.2}$$

Here, the sum is over all m and n greater or equal to zero and all multi-indices $\alpha = (\alpha_1, \dots, \alpha_m)$, $\beta = (\beta_1, \dots, \beta_n)$ with integers $0 < |\alpha_j|, |\beta_j| < N$ which have a given sum $s(\alpha)$, resp. $s(\beta)$.

The other almost-invariant is related to the oscillatory energy (1.4) and is given by

$$\mathcal{I}(\mathbf{p}, \mathbf{q}) = -\mathrm{i}\omega \sum_{0<|k|<N} k (q^{-k})^T p^k. \tag{3.3}$$

By (2.7)–(2.10), the functions $\mathbf{p}, \mathbf{q}$ satisfy the system

$$\begin{aligned} \dot{p}^k + \Omega^2 q^k &= -\nabla_{q^{-k}} \mathcal{K}(\mathbf{p}_1, \mathbf{q}) + \mathcal{O}(\omega^{-N}) && (3.4) \\ \dot{q}^k &= \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix} p^k + \nabla_{p^{-k}} \mathcal{K}(\mathbf{p}_1, \mathbf{q}) + \mathcal{O}(\omega^{-N}), && (3.5) \end{aligned}$$

which, neglecting the $\mathcal{O}(\omega^{-N})$ terms, is a Hamiltonian system with (3.1).

Theorem 5.3.1 *Under the assumptions of Theorem (5.2.1), the Hamiltonian (3.1) satisfies*

$$\begin{aligned} \mathcal{H}(\mathbf{p}(t), \mathbf{q}(t)) &= \mathcal{H}(\mathbf{p}(0), \mathbf{q}(0)) + \mathcal{O}(\omega^{-N}), && (3.6) \\ \mathcal{H}(\mathbf{p}(t), \mathbf{q}(t)) &= H(p(t), q(t)) + \mathcal{O}(\omega^{-1}), && (3.7) \end{aligned}$$

with $0 \le t \le T$.

Proof. Multiplying (3.4) with $(\dot{q}^{-k})^T$ and (3.5) (but with $-k$ instead of k) with $(\dot{p}^k)^T$ and summing up yields

$$\begin{aligned} \sum_{|k|<N} (\dot{q}^{-k})^T (\dot{p}^k + \Omega^2 q^k) &= -\sum_{|k|<N} (\dot{q}^{-k})^T \nabla_{q^{-k}} \mathcal{K}(\mathbf{p}_1, \mathbf{q}) + \mathcal{O}(\omega^{-N}) \\ \sum_{|k|<N} (\dot{p}^k)^T \dot{q}^{-k} = \sum_{|k|<N} (\dot{p}^k)^T &\left(\begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix} p^{-k} + \nabla_{p^k} \mathcal{K}(\mathbf{p}_1, \mathbf{q}) \right) + \mathcal{O}(\omega^{-N}). \end{aligned}$$

Substracting these two equations gives $\frac{d}{dt}\mathcal{H}(\mathbf{p}(t), \mathbf{q}(t)) = \mathcal{O}(\omega^{-N})$, integrating from 0 to T yields (3.6).

By the bounds obtained in Theorem 5.2.1, we have

$$\begin{aligned} H(p,q) &= K(p_1^0, q^0) + \frac{1}{2}||p_2^1 + p_2^{-1}||^2 \\ &+ \frac{\omega^2}{2}||q_2^1 + q_2^{-1}||^2 + \mathcal{O}(\omega^{-1}), \\ \mathcal{H}(\mathbf{p}, \mathbf{q}) &= K(p_1^0, q^0) + ||p_2^1||^2 + \omega^2 ||q_2^1||^2 + \mathcal{O}(\omega^{-1}). \end{aligned}$$

Using $p_2^1 + p_2^{-1} = \mathrm{i}\omega(q_2^1 - q_2^{-1}) + \mathcal{O}(\omega^{-1})$, $||p_2^1|| = \omega||q_2^1|| + \mathcal{O}(\omega^{-1})$ and developing these terms shows (3.7). □

Concerning the other formal invariant, we have a similar result

Theorem 5.3.2 *Under the assumptions of Theorem (5.2.1), the almost invariant (3.3) satisfies*

$$\begin{aligned} \mathcal{I}(\mathbf{p}(t), \mathbf{q}(t)) &= \mathcal{I}(\mathbf{p}(0), \mathbf{q}(0)) + \mathcal{O}(\omega^{-N}), && (3.8) \\ \mathcal{I}(\mathbf{p}(t), \mathbf{q}(t)) &= I(p(t), q(t)) + \mathcal{O}(\omega^{-1}), && (3.9) \end{aligned}$$

with $0 \le t \le T$.

Proof. For a real number λ, let us define the two vectors

$$\begin{aligned}\mathbf{p}(\lambda) &= (\mathrm{e}^{\mathrm{i}(-N+1)\lambda}p^{-N+1},\dots,p^0,\dots,\mathrm{e}^{\mathrm{i}(N-1)\lambda}p^{N-1}),\\ \mathbf{q}(\lambda) &= (\mathrm{e}^{\mathrm{i}(-N+1)\lambda}q^{-N+1},\dots,q^0,\dots,\mathrm{e}^{\mathrm{i}(N-1)\lambda}q^{N-1}).\end{aligned}$$

The definition (3.2) of $\mathcal{K}$ shows that $\mathcal{K}(\mathbf{p}_1(\lambda),\mathbf{q}(\lambda))$ does not depend on λ, hence its derivative with respect to λ yields

$$\begin{aligned}0 &= \tfrac{d}{d\lambda}\mathcal{K}(\mathbf{p}_1(\lambda),\mathbf{q}(\lambda))\\ &= \sum_{0<|k|<N}\mathrm{i}k\mathrm{e}^{\mathrm{i}k\lambda}\big((p^k)^T\nabla_{p^k}\mathcal{K}(\mathbf{p}_1(\lambda),\mathbf{q}(\lambda))+(q^k)^T\nabla_{q^k}\mathcal{K}(\mathbf{p}_1(\lambda),\mathbf{q}(\lambda))\big),\end{aligned}$$

and putting $\lambda=0$ we obtain

$$0=\sum_{0<|k|<N}\mathrm{i}k\big((p^k)^T\nabla_{p^k}\mathcal{K}(\mathbf{p}_1,\mathbf{q})+(q^k)^T\nabla_{q^k}\mathcal{K}(\mathbf{p}_1,\mathbf{q})\big). \tag{3.10}$$

We now multiply (3.4) with $-\mathrm{i}\omega k(q^{-k})^T$ and the transpose of (3.5) with $-\mathrm{i}\omega k(p^k)$, taking the sum over all k, with $0<|k|<N$, yields

$$\begin{aligned}-\mathrm{i}\omega\sum_{0<|k|<N}k(q^{-k})^T(\dot p^k+\Omega^2q^k) =\\ \mathrm{i}\omega\sum_{0<|k|<N}k(q^{-k})^T\nabla_{q^{-k}}\mathcal{K}(\mathbf{p}_1,\mathbf{q})+\mathcal{O}(\omega^{-N}),\end{aligned}$$

and

$$\begin{aligned}-\mathrm{i}\omega\sum_{0<|k|<N}k(\dot q^{-k})^Tp^k =\\ -\mathrm{i}\omega\sum_{0<|k|<N}k(p^{-k})^T\Big(\begin{pmatrix}0&0\\0&I\end{pmatrix}p^k+\nabla_{p^k}\mathcal{K}(\mathbf{p}_1,\mathbf{q})\Big)+\mathcal{O}(\omega^{-N}).\end{aligned}$$

Exchanging k with $-k$ in the last sum, adding these two quantities, using (3.10) and the symmetry of the terms $\sum_k k(q^{-k})^T\Omega q^k$ and $\sum_k k(p^{-k})^T\begin{pmatrix}0&0\\0&I\end{pmatrix}p^k$ yields

$$-\mathrm{i}\omega\sum_{0<|k|<N}k(\dot q^{-k})^Tp^k+k(q^{-k})^T\dot p^k=\mathcal{O}(\omega^{-N}).$$

This is nothing else than the time derivative of (3.3) which, after integration, implies (3.8).

Using the bounds from Theorem 5.2.1, we obtain

$$\begin{aligned}
\mathcal{I}(\mathbf{p}(t),\mathbf{q}(t)) &= -\mathrm{i}\omega(q_2^{-1})^T p_2^1 + \mathrm{i}\omega(q_2^1)^T p_2^{-1} + \mathcal{O}(\omega^{-1}) \\
&= -\mathrm{i}\omega(\zeta_2^{-1})^T(\mathrm{i}\omega\zeta_2^1) + \mathrm{i}\omega(\zeta_2^1)^T(-\mathrm{i}\omega\zeta_2^{-1}) + \mathcal{O}(\omega^{-1}) \\
&= 2\omega^2||\zeta_2^1||^2 + \mathcal{O}(\omega^{-1}), \\
I(p(t),q(t)) &= \tfrac{1}{2}||p_2^1 + p_2^{-1}||^2 + \tfrac{\omega^2}{2}||q_2^1 + q_2^{-1}||^2 + \mathcal{O}(\omega^{-1}) \\
&= 2\omega^2||q_2^1||^2 + \mathcal{O}(\omega^{-1}).
\end{aligned}$$

Result (3.9) follows using $q^k = \mathrm{e}^{\mathrm{i}k\omega t}\zeta^k$. □

We have obtained the same estimates for the two almost-invariants $\mathcal{H}$ and $\mathcal{I}$ as in [HLW02, Chap. XIII]. We thus can show that the oscillatory energy (1.4) is nearly conserved over long time intervals:

Theorem 5.3.3 *If the solution $(p(t),q(t))$ of the Hamiltonian problem (2.2), with initial values satisfying (1.2), stays in a compact set for $0 \le t \le \omega^N$, then*

$$I(p(t),q(t)) = I(p(0),q(0)) + \mathcal{O}(\omega^{-1}) + \mathcal{O}(t\omega^{-N}).$$

The constants symbolized by $\mathcal{O}$ are independent of ω and t, but depend on E and N.

5.4 Numerical methods

In this section, we adapt the numerical methods given in [HLW02, Chap. XIII] to our problem (1.1). Then we analyse the method, and present some geometric properties. Next, we illustrate the method on the stiff spring pendulum problem (see Section 5.1) and on the motion of a diatomic molecule. Finally we look at the expansion and at the almost invariants of the numerical methods.

5.4.1 New Trigonometric Methods (NTM)

Treating the second components of the variables p and q with a trigonometric method like those given in [HLW02, Chap. XIII] and the first components with the Störmer-Verlet scheme, the numerical method NTM reads

$$\begin{aligned}
p_1^{n+1/2} &= p_1^n - \frac{h}{2}\nabla_{q_1}K(p_1^{n+1/2},\Phi q^n) \\
q_1^{n+1} &= q_1^n + \frac{h}{2}\big(\nabla_{p_1}K(p_1^{n+1/2},\Phi q^n) + \nabla_{p_1}K(p_1^{n+1/2},\Phi q^{n+1})\big) \\
q_2^{n+1} &= \cos(h\omega)q_2^n + \omega^{-1}\sin(h\omega)p_2^n - \frac{h^2}{2}\Psi_2\nabla_{q_2}K(p_1^{n+1/2},\Phi q^n) \\
p_1^{n+1} &= p_1^{n+1/2} - \frac{h}{2}\nabla_{q_1}K(p_1^{n+1/2},\Phi q^{n+1}) \\
p_2^{n+1} &= -\omega\sin(h\omega)q_2^n + \cos(h\omega)p_2^n - \frac{h}{2}\big(\tilde{\Psi}_2\nabla_{q_2}K(p_1^{n+1/2},\Phi q^n) \\
&+ \hat{\Psi}_2\nabla_{q_2}K(p_1^{n+1/2},\Phi q^{n+1})\big),
\end{aligned} \tag{4.1}$$

where here and in the following $\Psi = \psi(h\Omega), \hat{\Psi} = \hat{\psi}(h\Omega), \tilde{\Psi} = \tilde{\psi}(h\Omega)$ and $\Phi = \phi(h\Omega)$ and $\Psi_2 = \psi(h\omega), \hat{\Psi}_2 = \hat{\psi}(h\omega), \tilde{\Psi}_2 = \tilde{\psi}(h\omega)$ and $\Phi = \phi(h\Omega)$. The filter functions $\psi, \hat{\psi}, \tilde{\psi}, \phi$ are even real-valued functions with $\psi(0) = \hat{\psi}(0) = \tilde{\psi}(0) = \phi(0) = 1$.

Because of the special structure of the Hamiltonian $H(p,q)$, the derivatives $\nabla_p K(p_1, q)$ and $\nabla_q K(p_1, q)$ do not depend on p_2, so that this component need not to be computed.

Moreover, we can remark that the method is explicit if the function $K(p_1, q)$ takes the form $K(p_1, q) = \frac{1}{2} p_1^T M(q_2) p_1 + U(q)$.

5.4.2 Numerical properties

We first present some conditions which determine the symmetry and the symplecticity of method (4.1), then we give its order and finally we mention a result concerning the conservation of the Hamiltonian (1.1).

Proposition 5.4.1 *The numerical method (4.1) is symmetric if and only if*

$$\psi(\zeta) = \operatorname{sinc}(\zeta)\hat{\psi}(\zeta), \qquad \tilde{\psi}(\zeta) = \cos(\zeta)\hat{\psi}(\zeta), \tag{4.2}$$

where $\operatorname{sinc}(\zeta) = \sin(\zeta)/\zeta$.

Proof: Exchanging $n \leftrightarrow n+1$ and $h \leftrightarrow -h$ in the definition of the numerical method (4.1), we have

$$\begin{aligned}
p_1^{n+1/2} &= p_1^{n+1} + \frac{h}{2}\nabla_{q_1} K(p_1^{n+1/2}, \Phi q^{n+1}) \\
q_1^n &= q_1^{n+1} - \frac{h}{2}\big(\nabla_{p_1} K(p_1^{n+1/2}, \Phi q^{n+1}) + \nabla_{p_1} K(p_1^{n+1/2}, \Phi q^n)\big) \\
q_2^n &= \cos(h\omega) q_2^{n+1} - \omega^{-1}\sin(h\omega) p_2^{n+1} - \frac{h^2}{2}\Psi_2 \nabla_{q_2} K(p_1^{n+1/2}, \Phi q^{n+1}) \\
p_1^n &= p_1^{n+1/2} + \frac{h}{2}\nabla_{q_1} K(p_1^{n+1/2}, \Phi q^n) \\
p_2^n &= \omega \sin(h\omega) q_2^{n+1} + \cos(h\omega) p_2^{n+1} + \frac{h}{2}\big(\tilde{\Psi}_2 \nabla_{q_2} K(p_1^{n+1/2}, \Phi q^{n+1}) \\
&+ \hat{\Psi}_2 \nabla_{q_2} K(p_1^{n+1/2}, \Phi q^n)\big).
\end{aligned}$$

The equations for $p_1^{n+1/2}, p_1^n, q_1^n$ coincide with those of (4.1). Solving the remaining two equations for q_2^{n+1}, p_2^{n+1} yields

$$\begin{aligned}
q_2^{n+1} &= \cos(h\omega)\big(q_2^n + \frac{h^2}{2}\Psi_2 \nabla_{q_2} K(p_1^{n+1/2}, \Phi q^{n+1})\big) \\
&+ \omega^{-1}\sin(h\omega)\big(p_2^n - \frac{h}{2}\big(\tilde{\Psi}_2 \nabla_{q_2} K(p_1^{n+1/2}, \Phi q^{n+1}) \\
&+ \hat{\Psi}_2 \nabla_{q_2} K(p_1^{n+1/2}, \Phi q^n)\big)\big), \\
p_2^{n+1} &= -\omega \sin(h\omega)\big(q_2^n + \frac{h^2}{2}\Psi_2 \nabla_{q_2} K(p_1^{n+1/2}, \Phi q^{n+1})\big) \\
&+ \cos(h\omega)\big(p_2^n - \frac{h}{2}(\tilde{\Psi}_2 \nabla_{q_2} K(p_1^{n+1/2}, \Phi q^{n+1}) \\
&+ \hat{\Psi}_2 \nabla_{q_2} K(p_1^{n+1/2}, \Phi q^n))\big).
\end{aligned}$$

Comparing these equations with those in the definition of the method, we thus have symmetry if and only if (4.2) holds. □

For symmetric methods, we can now give the condition for symplecticity

Theorem 5.4.2 *Under the symmetry conditions (4.2), if*

$$\hat{\psi}(\zeta) = \phi(\zeta) \tag{4.3}$$

holds then method (4.1) is symplectic.

Proof: Under the symmetry conditions, the numerical method reads

$$\begin{aligned}
p^{n+1/2} &= p^n - \frac{h}{2}\hat{\Psi}\nabla_q K(p_1^{n+1/2}, \Phi q^n) \\
q_1^{n+1} &= q_1^n + \frac{h}{2}\big(\nabla_{p_1} K(p_1^{n+1/2}, \Phi q^n) + \nabla_{p_1} K(p_1^{n+1/2}, \Phi q^{n+1})\big) \\
q_2^{n+1} &= \cos(h\omega) q_2^n + h\ \mathrm{sinc}(h\omega) p_2^{n+1/2} \\
p_1^{n+1} &= p_1^{n+1/2} - \frac{h}{2}\nabla_{q_1} K(p_1^{n+1/2}, \Phi q^{n+1}) \\
p_2^{n+1} &= -\omega \sin(h\omega) q_2^n + \cos(h\omega) p_2^{n+1/2} - \frac{h}{2}\hat{\Psi}_2 \nabla_{q_2} K(p_1^{n+1/2}, \Phi q^{n+1}).
\end{aligned} \tag{4.4}$$

We write this method as a composition of three methods

$$\begin{pmatrix} p^n \\ q^n \end{pmatrix} \to \begin{pmatrix} p^{n+1/2} \\ q^{n+1/2} \end{pmatrix} = \begin{pmatrix} p^n - \frac{h}{2}\hat{\Psi}\nabla_q K(p_1^{n+1/2}, \Phi q^n) \\ q^n + \frac{h}{2}\nabla_p K(p_1^{n+1/2}, \Phi q^n) \end{pmatrix}, \tag{4.5}$$

$$\begin{pmatrix} p^{n+1/2} \\ q^{n+1/2} \end{pmatrix} \to \begin{pmatrix} \hat{p}^{n+1/2} \\ \hat{q}^{n+1/2} \end{pmatrix} = \begin{pmatrix} p_1^{n+1/2} \\ -\omega \sin(h\omega) q_2^{n+1/2} + \cos(h\omega) p_2^{n+1/2} \\ q_1^{n+1/2} \\ \cos(h\omega) q_2^{n+1/2} + h\ \mathrm{sinc}(h\omega) p_2^{n+1/2} \end{pmatrix}, \tag{4.6}$$

$$\begin{pmatrix} \hat{p}^{n+1/2} \\ \hat{q}^{n+1/2} \end{pmatrix} \to \begin{pmatrix} p^{n+1} \\ q^{n+1} \end{pmatrix} = \begin{pmatrix} \hat{p}^{n+1/2} - \frac{h}{2}\hat{\Psi}\nabla_q K(\hat{p}_1^{n+1/2}, \Phi q^{n+1}) \\ \hat{q}^{n+1/2} + \frac{h}{2}\nabla_p K(\hat{p}_1^{n+1/2}, \Phi q^{n+1}) \end{pmatrix}, \tag{4.7}$$

where we have defined the second method so that the third method is the adjoint of the first. We have also used that $q_2^n = q_2^{n+1/2}$ which is due to the fact that the function K does not depend of p_2.

We now prove that (4.5) is symplectic if $\hat{\psi}(\zeta) = \phi(\zeta)$. We also show that (4.6) is symplectic. The proof for the symplecticity of the third composition method is similar to the one given for the symplecticity of the first method. Let us define the Hamiltonian $\hat{K}(p,q) = K(p_1, \Phi q)$. If we apply the symplectic Euler method to the system

$$\begin{aligned}
\dot{p} &= -\nabla_q \hat{K}(p,q) \\
\dot{q} &= \nabla_p \hat{K}(p,q),
\end{aligned}$$

we obtain

$$\begin{aligned} p^{n+1/2} &= p^n - \frac{h}{2}\nabla_q \hat{K}(p^{n+1/2}, q^n) \\ q^{n+1/2} &= q^n + \frac{h}{2}\nabla_p \hat{K}(p^{n+1/2}, q^n), \end{aligned}$$

which is nothing else than method (4.5) if $\hat{\psi}(\zeta) = \phi(\zeta)$. Thus the first composition method is symplectic if the condition (4.3) is satisfied.

To show that the second method is symplectic, we verify the condition $\Phi_h'^T J \Phi_h' = J$ where Φ_h stands for the second method (4.6) and J for the structure matrix. We have

$$\Phi_h' = \begin{pmatrix} I & 0 & 0 & 0 \\ 0 & \cos(h\omega)I & 0 & -\omega\sin(h\omega)I \\ 0 & 0 & I & 0 \\ 0 & h\ \mathrm{sinc}(h\omega)I & 0 & \cos(h\omega)I \end{pmatrix}.$$

This permits us to check the symplecticity condition for the second method and concludes the proof because the numerical method (4.4) is then a composition of three symplectic methods. Thus it is symplectic. □

We finally mention that we have, for fixed ω and $h \to 0$,

$$|H(p^n, q^n) - H(p^0, q^0)| \leq Ch^2 + C_N h^N t, \quad \text{for} \quad 0 \leq t = nh \leq h^{-N}, \tag{4.8}$$

for arbitrary positive integer N. This is due to the fact that our numerical method is consistent with the problem, symmetric, symplectic and has order two. However, the constants C, C_N depend on ω and this result is useless for $h\omega \gg 1$.

5.4.3 Numerical examples

We consider the stiff spring pendulum encountered in Section 5.1, and plot the Hamiltonian H and the oscillatory energy I along the numerical solution obtained with (4.1), first on an interval of length 200 and then 10000. For the filter functions, we choose $\psi_2(\zeta) = \mathrm{sinc}^2(\frac{\zeta}{2}), \hat{\psi}_2(\zeta) = \psi_2(\zeta)/\mathrm{sinc}(\zeta), \bar{\psi}_2(\zeta) = \cos(\zeta)\hat{\psi}_2(\zeta)$ and $\phi_2(\zeta) = \hat{\psi}_2(\zeta)$. With this choice, the numerical method NTM is symmetric and symplectic.

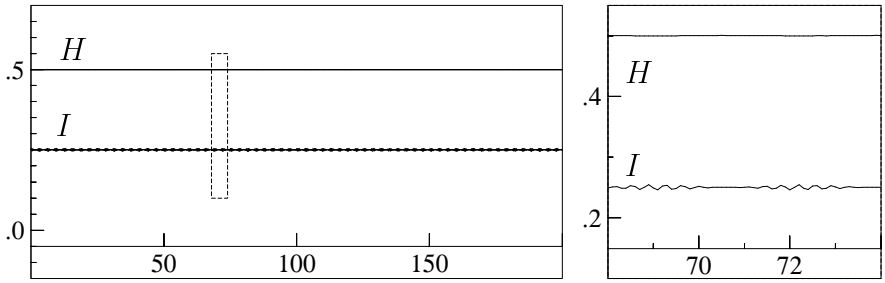


Figure 5.3: Total and oscillatory energies for problem (1.3) at xend= 200, with 2000 steps.

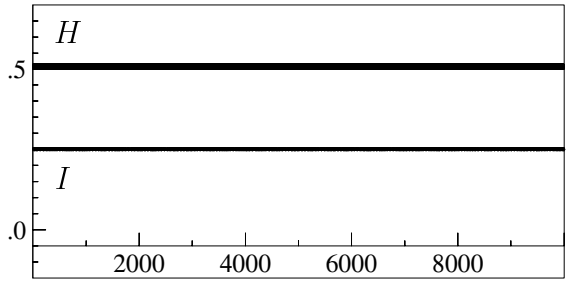


Figure 5.4: Total and oscillatory energies for problem (1.3) at xend= 10000, with 20000 steps.

As a second example, we consider the motion, in $\mathbb{R}^2$, of a diatomic molecule as the one encountered in the introduction. As in [AR99b], we use the local coordinates: $q_c \in \mathbb{R}^2$ for the center of mass of the two atoms, r for the bond length and ϕ for the angle of rotation. The Hamiltonian is

$$H(q_c, \phi, r, p_c, p_\phi, p_r) = \frac{1}{8} p_c^T p_c + \frac{1}{2} p_r^2 + \frac{1}{2}(r + r_0)^{-2} p_\phi^2 + \frac{\omega^2}{2} r^2,$$

where $r_0 = 1$ is the equilibrium length and ω is the stiffness of the spring. Let us use the same method used for the previous example and plot the total and oscillatory energies obtained with initial values $q_{c,1} = 0.1, p_\phi = p_r = 1$ and zero for the remaining initial values.

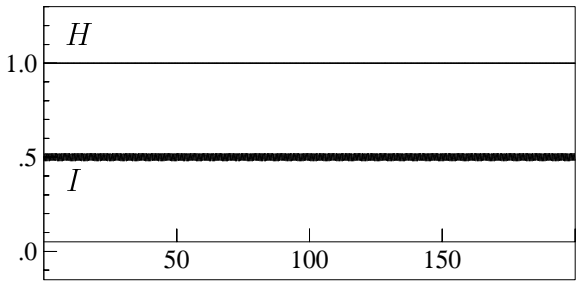


Figure 5.5: Total and oscillatory energies of the diatomic molecule at xend= 200, with 2000 steps and $\omega = 50$.

Once again, we obtain the desired behaviour.

A slight modification of the method (4.1) gives another type of numerical method for the problem (1.1)

$$\begin{aligned} p_1^{n+1/2} &= p_1^n - \frac{h}{2}\nabla_{q_1}K(p_1^n, \Phi q^n) \\ q_1^{n+1} &= q_1^n + \frac{h}{2}(\nabla_{p_1}K(p_1^{n+1/2}, \Phi q^n) + \nabla_{p_1}K(p_1^{n+1/2}, \Phi q^{n+1})) \\ q_2^{n+1} &= \cos(h\omega)q_2^n + \omega^{-1}\sin(h\omega)p_2^n - \frac{h^2}{2}\psi_2(h\omega)\nabla_{q_2}K(p_1^n, \Phi q^n) \\ p_1^{n+1} &= p_1^{n+1/2} - \frac{h}{2}\nabla q_1 K(p_1^{n+1}, \Phi q^{n+1}) \\ p_2^{n+1} &= -\omega\sin(h\omega)q_2^n + \cos(h\omega)p_2^n - \frac{h}{2}\big(\tilde{\psi}_2(h\omega)\nabla_{q_2}K(p_1^n, \Phi q^n) \\ &+ \hat{\psi}_2(h\omega)\nabla_{q_2}K(p_1^{n+1}, \Phi q^{n+1})\big). \end{aligned} \tag{4.9}$$

This method is also symmetric (with same conditions, see above) and gives similar results. It is explicit if the function K takes one of the following forms $K(p_1, q) = \tilde{K}(p_1, q_2)$ or $K(p_1, q) = \frac{1}{2}p_1^T M(q_2)p_1 + U(q)$.

5.5 Expansion of the numerical solution

In this section, we show that the symmetric numerical method (4.4) also admits a modulated Fourier expansion. As in [HLW02, Chap. XIII], we confine our discussion to the case where $h\omega \geq c_0 > 0$ and to the non-resonant case, and assume that h and ω^{-1} lie in a subregion of the (h, ω^{-1})-plan of small parameters for which there exists a positive constant c such that

$$|\sin(\frac{1}{2}kh\omega)| \geq c\sqrt{h}, \quad \text{for} \quad k = 1, \ldots, N, \ \text{with} \ N \geq 2. \tag{5.1}$$

For a given h and ω, this condition imposes a restriction on N. In the following, N is a fixed integer such that (5.1) holds.

We first recall the symmetric numerical method

$$\begin{aligned} p^{n+1/2} &= p^n - \frac{h}{2}\hat{\Psi}\nabla_q K(p_1^{n+1/2}, \Phi q^n) \\ q_1^{n+1} &= q_1^n + \frac{h}{2}\big(\nabla_{p_1}K(p_1^{n+1/2}, \Phi q^n) + \nabla_{p_1}K(p_1^{n+1/2}, \Phi q^{n+1})\big) \\ q_2^{n+1} &= \cos(h\omega)q_2^n + h\,\mathrm{sinc}(h\omega)p_2^{n+1/2} \\ p_1^{n+1} &= p_1^{n+1/2} - \frac{h}{2}\nabla_{q_1}K(p_1^{n+1/2}, \Phi q^{n+1}) \\ p_2^{n+1} &= -\omega\sin(h\omega)q_2^n + \cos(h\omega)p_2^{n+1/2} - \frac{h}{2}\hat{\psi}_2(h\omega)\nabla_{q_2}K(p_1^{n+1/2}, \Phi q^{n+1}), \end{aligned} \tag{5.2}$$

to prove the numerical analogue of Theorem 5.2.1, we need the additional requirements on the filter functions

$$\begin{aligned} |\hat{\psi}(h\omega)| &\leq C_1\,\mathrm{sinc}^2(\tfrac{1}{2}h\omega), \\ |\hat{\psi}(h\omega)| &\leq C_2|\,\mathrm{sinc}(h\omega)|. \end{aligned} \tag{5.3}$$

Theorem 5.5.1 *If the numerical solution (p^n, q^n) of (1.1) with initial values verifying (1.2) stay in a compact set, satisfies the non-resonance condition (5.1), the hypothesis (5.3) and if $h\omega \geq c_0 > 0$ then it admits, for $0 \leq t = nh \leq T$, the expansion*

$$\begin{aligned} p^n &= \sum_{|k|<N} \mathrm{e}^{\mathrm{i}k\omega t}\eta_h^k(t) + R_{h,N}(t), \\ q^n &= \sum_{|k|<N} \mathrm{e}^{\mathrm{i}k\omega t}\zeta_h^k(t) + S_{h,N}(t), \end{aligned} \tag{5.4}$$

where the remainder terms are bounded by

$$R_{h,N}(t) = \mathcal{O}(th^{N-2}), \quad S_{h,N}(t) = \mathcal{O}(th^{N-2}). \tag{5.5}$$

The coefficient functions are bounded, together with all their derivatives, by

$$\begin{array}{llll} \zeta_{h,1} = \mathcal{O}(1), & \eta_{h,1} = \mathcal{O}(1), & \zeta_{h,2} = \mathcal{O}(\omega^{-2}), & \eta_{h,2} = \mathcal{O}(\omega^{-1}), \\ \zeta^1_{h,1} = \mathcal{O}(\omega^{-2}), & \eta^1_{h,1} = \mathcal{O}(\omega^{-2}), & \zeta^1_{h,2} = \mathcal{O}(\omega^{-1}), & \dot\eta^1_{h,2} = \mathcal{O}(\omega^{-1}), \\ \zeta^k_{h,1} = \mathcal{O}(\omega^{-k-1}), & \eta^k_{h,1} = \mathcal{O}(\omega^{-k-1}), & \zeta^k_{h,2} = \mathcal{O}(\omega^{-k-2}), & \eta^k_{h,2} = \mathcal{O}(\omega^{-k-1}), \end{array} \tag{5.6}$$

for $k = 2, \ldots, N-1$. Moreover, we have $\eta^{-k} = \overline{\eta^k}$ and $\zeta^{-k} = \overline{\zeta^k}$. The constants symbolized by the $\mathcal{O}$-notation are independent of ω and h, but depend on E, N, c_0 and T.

Proof. The proof of this theorem is very similar to the one given for the expansion of the exact solution. We look for two functions

$$\begin{aligned} p_h(t) &= \eta_h(t) + \sum_{0<|k|<N} \mathrm{e}^{\mathrm{i}k\omega t}\eta_h^k(t) \\ q_h(t) &= \zeta_h(t) + \sum_{0<|k|<N} \mathrm{e}^{\mathrm{i}k\omega t}\zeta_h^k(t), \end{aligned} \tag{5.7}$$

with smooth (in the sense that all their derivatives are bounded independently of h and ω) coefficients $\zeta_h, \zeta_h^k, \eta_h$ and η_h^k such that, with $t = nh$,

$$\begin{aligned} p^n &= p_h(t) + \mathcal{O}(h^{N-1}) \\ q^n &= q_h(t) + \mathcal{O}(h^{N-1}). \end{aligned} \tag{5.8}$$

Construction of the coefficient functions. As in Theorem 5.2.1, we insert (5.7) in the numerical method (5.2), expand the nonlinearity functions $\nabla_p K$ and $\nabla_q K$ around $(\eta_{h,1}(t), \Phi(h\omega)\zeta_h(t))$ and compare the coefficients of $\mathrm{e}^{\mathrm{i}k\omega t}$. To motivate the ansatz (5.12) below, we compare the dominant terms.

- Firstly, we implicitly define, for $t = nh + \frac{h}{2}$,

$$\hat{p}_h(t) = p_h(t - \tfrac{h}{2}) - \tfrac{h}{2}\hat{\Psi}\nabla_q K(\hat{p}_{h,1}(t), \Phi q_h(t - \tfrac{h}{2})). \tag{5.9}$$

As for (5.7), we also define

$$p^{n+1/2} = \hat{p}(t) = \xi_h(t) + \sum_{0<|k|<N} \mathrm{e}^{\mathrm{i}h\omega t}\xi_h^k(t). \tag{5.10}$$

The coefficient functions of (5.10) satisfy $\xi_h^k(t) = \eta_h^k(t) + \mathcal{O}(h)$.

- For the second term of the numerical method (5.2), we have

$$q_{h,1}(t + \tfrac{h}{2}) - q_{h,1}(t - \tfrac{h}{2}) = \tfrac{h}{2}\big(\nabla_{p_1} K(\hat{p}_{h,1}(t), \Phi q_h(t - \tfrac{h}{2})) + \nabla_{p_1} K(\hat{p}_{h,1}(t), \Phi q_h(t + \tfrac{h}{2}))\big).$$

Using (5.7), we get

$$\begin{aligned}\sum_{|k|<N} \mathrm{e}^{\mathrm{i}k\omega(t+h/2)}\zeta_h^k(t + \tfrac{h}{2}) - \sum_{|k|<N} \mathrm{e}^{\mathrm{i}k\omega(t-h/2)}\zeta_h^k(t - \tfrac{h}{2}) &= \tfrac{h}{2}\big(\nabla_{p_1} K(\hat{p}_{h,1}(t), \Phi q_h(t - \tfrac{h}{2})) \\ &+ \nabla_{p_1} K(\hat{p}_{h,1}(t), \Phi q_h(t + \tfrac{h}{2}))\big).\end{aligned}$$

Expanding the smooth functions η_h and ζ_h around $h = 0$ and the function $\nabla_{p_1} K$ into its Taylor series, and comparing the coefficient of $\mathrm{e}^{\mathrm{i}k\omega t}$ yields for $k = 0$

$$\begin{aligned}\zeta_{h,1}(t) &+ \tfrac{h}{2}\dot{\zeta}_{h,1}(t) - \zeta_{h,1}(t) + \tfrac{h}{2}\dot{\zeta}_{h,1}(t) + \mathcal{O}(h^3) = h\nabla_{p_1} K(\eta_{h,1}(t), \Phi\zeta_h(t)) \\ &+ \tfrac{h}{2} \sum_{s(\alpha)+s(\beta)=0} \frac{1}{m!n!} D_1^{m+1} D_2^n K(\eta_{h,1}(t), \Phi\zeta_h(t))(\eta_{h,1}^\alpha(t), \Phi\zeta_h^\beta(t)) \\ &+ \tfrac{h}{2} \sum_{s(\alpha)+s(\beta)=0} \frac{1}{m!n!} \mathrm{e}^{\mathrm{i}\omega h/2(s(\beta)-s(\alpha))} D_1^{m+1} D_2^n K(\eta_{h,1}(t), \Phi\zeta_h(t))(\eta_{h,1}^\alpha(t), \Phi\zeta_h^\beta(t)) \\ &+ \mathcal{O}(h),\end{aligned}$$

where we used the same notations as in Theorem 5.2.1 for multi-indices α and β. This yields a relation for $\dot{\zeta}_{h,1}(t)$. Similarly, for $k \neq 0$, we obtain

$$\begin{aligned}\zeta_{h,1}^k(t) &= \frac{h}{4\mathrm{i}\sin(k\omega\frac{h}{2})}\Big(\sum_{s(\alpha)+s(\beta)=k} \frac{1}{m!n!} \mathrm{e}^{-\mathrm{i}k\omega h/2} D_1^{m+1} D_2^n K(\eta_{h,1}(t), \Phi\zeta_h(t))(\eta_{h,1}^\alpha(t), \Phi\zeta_h^\beta(t)) \\ &+ \sum_{s(\alpha)+s(\beta)=k} \frac{1}{m!n!} \mathrm{e}^{\mathrm{i}\omega h/2(s(\beta)-s(\alpha))} D_1^{m+1} D_2^n K(\eta_{h,1}(t), \Phi\zeta_h(t))(\eta_{h,1}^\alpha(t), \Phi\zeta_h^\beta(t))\Big) \\ &+ \mathcal{O}(h^2).\end{aligned}$$

- For the third equation of (5.2), we obtain

$$q_{h,2}(t+\frac{h}{2}) = \cos(h\omega)q_{h,2}(t-\frac{h}{2}) + h\,\mathrm{sinc}(h\omega)\hat{p}_{h,2}(t).$$

Again, we expand the smooth functions η_h^k and ζ_h^k into their Taylor series. Comparing the coefficient of $\mathrm{e}^{\mathrm{i}k\omega t}$ yields for $k=0$

$$(1-\cos(h\omega))\zeta_{h,2}(t) = \frac{h}{2}(-1-\cos(h\omega))\dot{\zeta}_{h,2}(t) + h\,\mathrm{sinc}(h\omega)\eta_{h,2}(t) + \mathcal{O}(h^2).$$

For $k=1$, we obtain

$$\mathrm{i}\sin(h\omega)\zeta^1_{h,2}(t) = (-h\cos(h\omega) - \mathrm{i}\frac{h}{2}\sin(h\omega))\dot{\zeta}^1_{h,2}(t) + h\,\mathrm{sinc}(h\omega)\eta^1_{h,2}(t) + \mathcal{O}(h^2).$$

And finally, for the remaining k

$$\begin{aligned}(\cos(kh\omega)+\mathrm{i}\sin(kh\omega))(\zeta^k_{h,2}(t)+\frac{h}{2}\dot{\zeta}^k_{h,2}(t)) = \cos(h\omega)(\zeta^k_{h,2}(t)-\frac{h}{2}\dot{\zeta}^k_{h,2}(t))\\ + h\,\mathrm{sinc}(h\omega)\eta^k_{h,2}(t) + \mathcal{O}(h^2).\end{aligned}$$

Here, we can remark that these equations also depend on the derivative of the coefficient functions. We can remove them by using iteratively these equations.

- From the fourth equation of (5.2), we get similar relations as for the second equation (see above) for the coefficient functions $\dot{\eta}_{h,1}, \eta^k_{h,1}$ but with $-\nabla_{q_1}K$ instead of $\nabla_{p_1}K$.

- For the last formula of (5.2), we use the symmetry of the method, exchanging $n \leftrightarrow n+1$ and $h \leftrightarrow -h$, we get $p^n_{h,2} = \omega\sin(h\omega)q^{n+1}_{h,2} + \cos(h\omega)p^{n+1/2}_{h,2} + \frac{h}{2}\hat{\psi}_2(h\omega)\nabla_{q_2}K(p^{n+1/2}_{h,1}, \Phi q^n_h)$. Taking $n-1$ in place of n in this last expression and adding this quantity to $p^{n+1}_{h,2}$ yields

$$\begin{aligned}p_2^{n+1} &+ p_2^{n-1} = \cos(h\omega)(p_2^{n+1/2} + p_2^{n-1/2})\\ &+ \frac{h}{2}\hat{\psi}_2(h\omega)\big(\nabla_{q_2}K(p_1^{n-1/2}, \Phi q^{n-1}) - \nabla_{q_2}K(p_1^{n+1/2}, \Phi q^{n+1})\big).\end{aligned}$$

Inserting (5.7) and using the fact that $p_1^{n-1/2} = p_1^{n-1} + \mathcal{O}(h)$ and $p_1^{n+1/2} = p_1^n + \mathcal{O}(h)$, we obtain

$$\begin{aligned}p_{h,2}(t+\frac{h}{2}) + p_{h,2}(t-\frac{3h}{2}) &= 2\cos(h\omega)p_{h,2}(t-\frac{h}{2})\\ &+ \frac{h}{2}\cos(h\omega)\hat{\psi}_2(h\omega)\big(\nabla_{q_2}K(p_{h,1}(t-\frac{3h}{2}), \Phi q_h(t-\frac{h}{2}))\\ &- \nabla_{q_2}K(p_{h,1}(t-\frac{h}{2}), \Phi q_h(t-\frac{h}{2}))\big)\\ &+ \frac{h}{2}\hat{\psi}_2(h\omega)\big(\nabla_{q_2}K(p_{h,1}(t-\frac{3h}{2}), \Phi q_h(t-\frac{3h}{2}))\\ &- \nabla_{q_2}K(p_{h,1}(t-\frac{h}{2}), \Phi q_h(t+\frac{h}{2}))\big) + \mathcal{O}(h^2).\end{aligned} \tag{5.11}$$

This relation is true for every t, so we can exchange t with $t+\frac{h}{2}$. Using the operator

$$\mathcal{L}(hD) = \mathrm{e}^{hD} - 2\cos(h\Omega) + \mathrm{e}^{-hD} = 4\sin(\tfrac{h}{2}h\Omega + \tfrac{1}{2}\mathrm{i}hD)\sin(\tfrac{h}{2}h\Omega - \tfrac{1}{2}\mathrm{i}hD)$$

defined in [HLW02, Chap. XIII], we can rewrite formula (5.11) as

$$\begin{aligned}\mathcal{L}(hD)p_{h,2}(t) = {} & \frac{h}{2}\cos(h\omega)\hat{\psi}_2(h\omega)\Big(\nabla_{q_2}K(p_{h,1}(t-h),\Phi q_h(t)) \\ & - \nabla_{q_2}K(p_{h,1}(t),\Phi q_h(t))\Big) + \frac{h}{2}\hat{\psi}_2(h\omega)\Big(\nabla_{q_2}K(p_{h,1}(t-h),\Phi q_h(t-h)) \\ & - \nabla_{q_2}K(p_{h,1}(t),\Phi q_h(t+h))\Big) + \mathcal{O}(h^2).\end{aligned}$$

Now, by the hypothesis (5.1) on N, the dominating terms in the Taylor expansions of $\mathcal{L}(hD)$ and $\mathcal{L}(hD+\mathrm{i}hk\omega)$ give the desired first terms for the series of the coefficient functions $\eta^k_{h,2}$. Indeed, we have

$$\begin{aligned}\eta_{h,2}(t) &= \frac{h}{8\sin^2(\frac{h}{2}\omega)}\hat{\psi}_2(h\omega)\cos(h\omega)(\ldots) + \frac{h}{8\sin^2(\frac{h}{2}\omega)}\hat{\psi}_2(h\omega)(\ldots) + \mathcal{O}(h^2) \\ \dot{\eta}^1_{h,2}(t) &= \frac{1}{4\mathrm{i}\sin(h\omega)}\hat{\psi}_2(h\omega)\cos(h\omega)(\ldots) + \frac{1}{4\mathrm{i}\sin(h\omega)}\hat{\psi}_2(h\omega)(\ldots) + \mathcal{O}(h) \\ \eta^k_{h,2}(t) &= -\frac{h}{8\sin(\frac{(k-1)h}{2}\omega)\sin(\frac{(k+1)h}{2}\omega)}\hat{\psi}_2(h\omega)\cos(h\omega)(\ldots) \\ &\quad - \frac{h}{8\sin(\frac{(k-1)h}{2}\omega)\sin(\frac{(k+1)h}{2}\omega)}\hat{\psi}_2(h\omega)(\ldots) + \mathcal{O}(h^2),\end{aligned}$$

where the $(\ldots)$ terms are big expressions involving sums like those encountered in the formulas for $\zeta^k_{h,1}$ (see above).

This motivates the ansatz

$$\begin{aligned}\dot{\zeta}_{h,1} &= f_{10}(\cdot) + \ldots \\ \dot{\eta}_{h,1} &= g_{10}(\cdot) + \ldots \\ \dot{\eta}^1_{h,2} &= \frac{\hat{\Psi}_2}{\sin(h\omega)}(g^1_{20}(\cdot) + \ldots) \\ \zeta^k_{h,1} &= \frac{h}{\sin(k\frac{h}{2}\omega)}(f^k_{10}(\cdot) + \ldots) \\ \eta^k_{h,1} &= \frac{h}{\sin(k\frac{h}{2}\omega)}(g^k_{10}(\cdot) + \ldots) \\ \zeta^k_{h,2} &= f^k_{20}(\cdot) + \ldots \\ \eta_{h,2} &= \frac{\hat{\Psi}_2}{\sin^2(\frac{h}{2}\omega)}(g_{20}(\cdot) + \ldots) \\ \eta^k_{h,2} &= \frac{h\hat{\Psi}_2}{\sin(\frac{k-1}{2}h\omega)\sin(\frac{k+1}{2}h\omega)}(g^k_{20}(\cdot) + \ldots),\end{aligned} \tag{5.12}$$

where the dots stands for power series in h with coefficient functions f^k_{mn} and g^k_{mn} depending on the variables $\zeta_{h,1}, \eta_{h,1}, \eta^1_{h,2}$ and $h\omega$. Using the hypothesis on the filter functions and a closer look at the functions f^k_{mn} and g^k_{mn} gives the bounds (5.6) on the coefficient functions of the modulated Fourier expansion.

Initial values. The conditions $p_{h,1}(0) = p_1(0), p_{h,2}(0) = p_2(0), q_{h,1}(0) = q_1(0)$ and $p_{h,2}(h) = p_2(h)$, give a system

$$\begin{aligned} p_1(0) &= \eta_{h,1}(0) + \mathcal{O}(\omega^{-2}) \\ p_2(0) &= 2\mathrm{Re}(\eta^1_{h,2}(0)) + \mathcal{O}(\omega^{-1}) \\ q_1(0) &= \zeta_{h,1}(0) + \mathcal{O}(\omega^{-2}) \\ \omega q_2(0) &= 2\mathrm{Im}(\eta^1_{h,2}(0)) + \mathcal{O}(\omega^{-1}), \end{aligned}$$

that can be solved using the implicit function theorem to yield locally the desired initial values $\eta_{h,1}(0), \zeta_{h,1}(0), \eta^1_{h,2}(0)$ for the differential equations appearing in the ansatz (here we used condition (1.2)).

Defect. Let's define the components of the defect, for $t = nh$,

$$\begin{aligned} d_1(t) &= q_{h,1}(t+h) - q_{h,1}(t) - \frac{h}{2}\big(\nabla_{p_1} K(\hat{p}_{h,1}(t+\tfrac{h}{2}), \Phi q_h(t)) + \nabla_{p_1} K(\hat{p}_{h,1}(t+\tfrac{h}{2}), \Phi q_h(t+h))\big) \\ d_2(t) &= p_{h,1}(t+h) - p_{h,1}(t) + \frac{h}{2}\big(\nabla_{q_1} K(\hat{p}_{h,1}(t+\tfrac{h}{2}), \Phi q_h(t)) + \nabla_{q_1} K(\hat{p}_{h,1}(t+\tfrac{h}{2}), \Phi q_h(t+h))\big) \\ d_3(t) &= q_{h,2}(t+h) - \cos(h\omega) q_{h,2}(t) - h\,\mathrm{sinc}(h\omega) p_{h,2}(t) \\ &\quad + \frac{h^2}{2}\,\mathrm{sinc}(h\omega)\hat{\psi}_2(h\omega)\nabla_{q_2} K(\hat{p}_{h,1}(t+\tfrac{h}{2}), \Phi q_h(t)) \\ d_4(t) &= p_{h,2}(t+h) + \omega \sin(h\omega) q_{h,2}(t) - \cos(h\omega) p_{h,2}(t) \\ &\quad + \frac{h}{2}\,\mathrm{sinc}(h\omega)\hat{\psi}_2(h\omega)\nabla_{q_2} K(\hat{p}_{h,1}(t+\tfrac{h}{2}), \Phi q_h(t)). \end{aligned}$$

By definition of the coefficient functions ζ^k_h, η^k_h, we have $d_1(t) = d_2(t) = \mathcal{O}(h^{N+1})$ and $d_3(t) = \mathcal{O}(h^N)$. For the fourth component of the defect, we have to use the two-step formulation for $p_{h,2}$, this gives $d_4(t+h) + d_4(t-h) = \mathcal{O}(h^N)$. With our choice for the initial values, the defect at $t = 0$ is $d_4(0) = \mathcal{O}(h^N)$, so that we have $d_4(t) = \mathcal{O}(h^N) + \mathcal{O}(th^{N-1})$.

We still have to estimate the remainders (5.5). To do this, we define $R^n = ||p^n - p_h(t)||$, $S^n = ||q^n - q_h(t)||$, and the norm $||(S_1, R_1, S_2, R_2)||_* = ||(S_1, R_1, \omega S_2, R_2)||$. Let's begin with the difference $p_1^{n+1/2} - \hat{p}_{h,1}(t + \frac{h}{2})$. For the difference of the first equation in (5.2) and of (5.10), if the gradient of K satisfies a Lipschitz condition, we have, by a triangle inequality

$$||p_1^{n+1/2} - \hat{p}_{h,1}(t+\tfrac{h}{2})|| \le ||R_1^n|| + C_1 h||p_1^{n+1/2} - \hat{p}_{h,1}(t+\tfrac{h}{2})|| + C_2 h||S^n||.$$

This gives

$$||p_1^{n+1/2} - \hat{p}_{h,1}(t+\tfrac{h}{2})|| \le \alpha, \qquad \text{where} \qquad \alpha = \frac{1}{1 - C_1 h}(||R_1^n|| + C_2 h||S^n||).$$

Similarly, for the remainders, we then have

$$\begin{aligned} ||(S_1, R_1, S_2, R_2)^{n+1}||_* &\leq ||(S_1, R_1, S_2, R_2)^n||_* + h\kappa_1\alpha \\ &+ h\kappa_2||(S_1, R_1, S_2, R_2)^{n+1}||_* \\ &+ h\kappa_3||(S_1, R_1, S_2, R_2)^n||_* + \kappa_4 h^{N-1}. \end{aligned}$$

Using this relation repeatedly and the fact that $||(S_1, R_1, S_2, R_2)^0||_* = \mathcal{O}(h^N)$ is given by the definition of the initial values, we obtain the following estimate for the remainders

$$\begin{aligned} ||(S_1, R_1, S_2, R_2)^n||_* &\leq (\frac{1+h\tilde{\kappa}_1}{1-h\tilde{\kappa}_2})^n||(S_1, R_1, S_2, R_2)^0||_* + \tilde{\kappa}_3(n+1)h^{N-1} \\ &\leq \tilde{C}nh^{N-1}. \end{aligned}$$

□

5.6 Almost-invariants of the numerical method

We want to show that the coefficient functions of the modulated Fourier expansion of the numerical method (5.2) also have two almost-invariants. By the last theorem of the preceding section, we have (with a different N than the one given in that theorem)

$$\begin{aligned} &\hat{p}_h(t) - p_h(t-\tfrac{h}{2}) = -\frac{h}{2}\hat{\Psi}\nabla_q K(\hat{p}_{h,1}(t), \Phi q_h(t-\tfrac{h}{2})) \\ &q_{h,1}(t+\tfrac{h}{2}) - q_{h,1}(t-\tfrac{h}{2}) = \frac{h}{2}\big(\nabla_{p_1} K(\hat{p}_{h,1}(t), \Phi q_h(t-\tfrac{h}{2})) \\ &\qquad + \nabla_{p_1} K(\hat{p}_{h,1}(t), \Phi q_h(t+\tfrac{h}{2}))\big) + \mathcal{O}(h^N) \\ &p_{h,1}(t+\tfrac{h}{2}) - \hat{p}_{h,1}(t) = -\frac{h}{2}\nabla_{q_1} K(\hat{p}_{h,1}(t), \Phi q_h(t+\tfrac{h}{2}) + \mathcal{O}(h^N) \\ &p_{h,2}(t+\tfrac{h}{2}) + \omega\sin(h\omega)q_{h,2}(t-\tfrac{h}{2}) - \cos(h\omega)\hat{p}_{h,2}(t) = \\ &\qquad - \frac{h}{2}\hat{\psi}_2(h\omega)\nabla_{q_2} K(\hat{p}_{h,1}(t), \Phi q_h(t+\tfrac{h}{2}) + \mathcal{O}(h^N) \\ &q_{h,2}(t+\tfrac{h}{2}) - \cos(h\omega)q_{h,2}(t-\tfrac{h}{2}) = h\,\mathrm{sinc}(h\omega)\hat{p}_{h,2} + \mathcal{O}(h^N), \end{aligned}$$

where we define $q_h(t) = \sum_{|k|<N} q_h^k(t)$, $p_h(t) = \sum_{|k|<N} p_h^k(t)$ and $\hat{p}_h(t) = \sum_{|k|<N} \hat{p}_h^k(t)$ with $q_h^k(t) = \mathrm{e}^{\mathrm{i}k\omega t}\zeta_h^k(t), p_h^k(t) = \mathrm{e}^{\mathrm{i}k\omega t}\eta_h^k(t)$ and $\hat{p}_h^k(t) = \mathrm{e}^{\mathrm{i}k\omega t}\varsigma_h^k(t)$. Comparing the coefficient of $\mathrm{e}^{\mathrm{i}k\omega t}$, we

get, writting the resulting equations in terms of $\hat{p}_h^k, p_h^k$ and q_h^k,

$$\begin{aligned}
&\hat{p}_h^k(t) - p_h^k(t-\tfrac{h}{2}) = -\tfrac{h}{2}\hat{\Psi}\Phi^{-1}\nabla_{q^{-k}}\mathcal{K}_h(\hat{\mathbf{p}}_1(t), \mathbf{q}(t-\tfrac{h}{2}))\\
&q_{h,1}^k(t+\tfrac{h}{2}) - q_{h,1}^k(t-\tfrac{h}{2}) = \tfrac{h}{2}\big(\nabla_{p_1^{-k}}\mathcal{K}_h(\hat{\mathbf{p}}_1(t), \mathbf{q}(t-\tfrac{h}{2}))\\
&\qquad\qquad + \nabla_{p_1^{-k}}\mathcal{K}_h(\hat{\mathbf{p}}_1(t), \mathbf{q}(t+\tfrac{h}{2}))\big) + \mathcal{O}(h^N)\\
&p_{h,1}^k(t+\tfrac{h}{2}) - \hat{p}_{h,1}^k(t) = -\tfrac{h}{2}\nabla_{q_1^{-k}}\mathcal{K}_h(\hat{\mathbf{p}}_1(t), \mathbf{q}(t+\tfrac{h}{2})) + \mathcal{O}(h^N)\\
&p_{h,2}^k(t+\tfrac{h}{2}) + \omega\sin(h\omega)q_{h,2}^k(t-\tfrac{h}{2}) - \cos(h\omega)\hat{p}_{h,2}^k(t) =\\
&\qquad\qquad - \tfrac{h}{2}\hat{\psi}_2(h\omega)\phi_2^{-1}(h\omega)\nabla_{q_2^{-k}}\mathcal{K}_h(\hat{\mathbf{p}}_1(t), \mathbf{q}(t+\tfrac{h}{2})) + \mathcal{O}(h^N)\\
&q_{h,2}^k(t+\tfrac{h}{2}) - \cos(h\omega)q_{h,2}^k(t-\tfrac{h}{2}) = h\,\mathrm{sinc}(h\omega)\hat{p}_{h,2}^k(t) + \mathcal{O}(h^N),
\end{aligned} \tag{6.1}$$

where, similarly to (3.2), we define

$$\mathcal{K}_h(\hat{\mathbf{p}}_1, \mathbf{q}) = K(\hat{p}_1^0, \Phi q^0) + \sum_{s(\alpha)+s(\beta)=0} \frac{1}{m!n!} D_1^m D_2^n K(\hat{p}_1^0, \Phi q^0)(\hat{\mathbf{p}}_1^\alpha, (\Phi\mathbf{q})^\beta), \tag{6.2}$$

for a vector $\hat{\mathbf{p}}_1 = (\hat{p}_{h,1}^{-N+1}, \ldots, \hat{p}_{h,1}^0, \ldots, \hat{p}_{h,1}^{N-1})$ and $\hat{p}_{h,1}^k = \mathrm{e}^{\mathrm{i}k\omega t}\xi_{h,1}^k(t)$, where $\xi_{h,1}^k(t)$ are the modulated functions (5.10) (see Theorem 5.5.1). The same notation is used for $\mathbf{q}$. From here, we do not write the index h in the modulation functions.

Lemma 5.6.1 *Under the assumptions of Theorem 5.5.1, the coefficient functions of the modulated Fourier expansion of the numerical solution satisfy*

$$\hat{\mathcal{H}}_h[\eta^k, \zeta^k](t) = \hat{\mathcal{H}}_h[\eta^k, \zeta^k](0) + \mathcal{O}(th^N), \tag{6.3}$$

for $0 \le t \le T$. Moreover, we have

$$\hat{\mathcal{H}}_h[\eta^k, \zeta^k](t) = 2\omega^2\mu(h\omega)(\zeta_2^{-1})^T\zeta_2^1 + K(\eta_1, \Phi\zeta) + \mathcal{O}(h), \tag{6.4}$$

where $\mu(h\omega) = \phi_2(h\omega)\hat{\psi}_2^{-1}(h\omega)$.

Proof. The idea of the proof is to multiply the relations in (6.1) by a derivative of some coefficient functions, then we take the sum over all k with $|k| < N$ and show that the resulting formula is in fact a total derivative of a function, say, $\hat{\mathcal{H}}_h[\eta^k, \zeta^k](t)$. After integration, we obtain the desired statement (6.3).

After multiplications and summations, (6.1) becomes

$$\begin{aligned}
&\sum_{|k|<N}\Big(-(\dot{q}^{-k}(t-\tfrac{h}{2}))^T\Phi\hat{\Psi}^{-1}\big(\hat{p}^k(t) - p^k(t-\tfrac{h}{2})\big) + \dot{\hat{p}}_1^{-k}(t)^T(q_1^k(t+\tfrac{h}{2}) - q_1^k(t-\tfrac{h}{2}))\\
&\quad - (\dot{q}_1^{-k}(t+\tfrac{h}{2}))^T(p_1^k(t+\tfrac{h}{2}) - \hat{p}_1^k(t)) - (\dot{q}_2^{-k}(t+\tfrac{h}{2}))^T\phi_2(h\omega)\hat{\psi}_2^{-1}(h\omega)\big(p_2^k(t+\tfrac{h}{2})\\
&\quad + \omega\sin(h\omega)q_2^k(t-\tfrac{h}{2}) - \cos(h\omega)\hat{p}_2^k(t)\big)\Big) =\\
&= \tfrac{h}{2}\tfrac{\mathrm{d}}{\mathrm{d}t}\Big(\mathcal{K}_h(\hat{\mathbf{p}}_1(t), \mathbf{q}(t+\tfrac{h}{2})) + \mathcal{K}_h(\hat{\mathbf{p}}_1(t), \mathbf{q}(t-\tfrac{h}{2}))\Big) + \mathcal{O}(h^N).
\end{aligned}$$

Expanding the functions $\zeta^k(t \pm \frac{h}{2})$ and $\eta^k(t \pm \frac{h}{2})$ around t and replacing $\hat{p}_2^k$ by the last formula of (6.1) shows that the left side of this equality is a total derivative.

Moving the terms from the left to the right side of the equality, we get

$$\frac{\mathrm{d}}{\mathrm{dt}}\hat{\mathcal{H}}_h[\eta^k, \zeta^k](t) = \mathcal{O}(h^N),$$

and an integration yields statement (6.3) of the theorem.

This construction of $\hat{\mathcal{H}}_h[\eta^k, \zeta^k](t)$, the bounds of Theorem 5.5.1, hypothesis (5.3) on the filter functions and the fact that we have $\eta_2^1 = \mathrm{i}\omega\zeta_2^1 + \mathcal{O}(h^2)$ yields (6.4) and conclude the proof. □

For the second near invariant, similarly to (3.10), we obtain

$$\omega \sum_{0<|k|<N} \mathrm{i}k\Big((\hat{p}^k)^T\nabla_{p^k}\mathcal{K}_h(\hat{\mathbf{p}}_1, \mathbf{q}) + (q^k)^T\nabla_{q^k}\mathcal{K}_h(\hat{\mathbf{p}}_1, \mathbf{q})\Big) = 0, \tag{6.5}$$

for $\mathcal{K}_h(\hat{\mathbf{p}}_1(t), \mathbf{q}(t))$ of (6.2). The same tricks used in the proof of the last lemma permit to prove the following lemma

Lemma 5.6.2 *Under the assumptions of Theorem 5.5.1, the coefficient functions of the modulated Fourier expansion of the numerical solution satisfy*

$$\hat{\mathcal{I}}_h[\eta^k, \zeta^k](t) = \hat{\mathcal{I}}_h[\eta^k, \zeta^k](0) + \mathcal{O}(th^N), \tag{6.6}$$

for $0 \le t \le T$. *Moreover, we have*

$$\hat{\mathcal{I}}_h[\eta^k, \zeta^k](t) = 2\omega^2\mu(h\omega)(\zeta_2^{-1})^T\zeta_2^1 + \mathcal{O}(h^2), \tag{6.7}$$

where $\mu(h\omega) = \phi_2(h\omega)\hat{\psi}_2^{-1}(h\omega)$.

Proof. This time, we multiply and sum the equalities in (6.1) too make (6.5) appear. We get

$$\begin{aligned}
&\mathrm{i}\omega \sum_{0<|k|<N} k\big(-(q^{-k}(t-\tfrac{h}{2}))^T\Phi\hat{\Psi}^{-1}(\hat{p}^k - p^k(t-\tfrac{h}{2})) + (\hat{p}_2^{-k})^T(q_1^k(t+\tfrac{h}{2}) - q_1^k(t-\tfrac{h}{2}))\\
&\quad - (q_1^{-k}(t+\tfrac{h}{2}))^T(p_1^k(t+\tfrac{h}{2}) - \hat{p}_1^k) - (q_2^{-k}(t+\tfrac{h}{2}))^T\phi_2(h\omega)\hat{\psi}_2^{-1}(h\omega)(p_2^k(t+\tfrac{h}{2})\\
&\quad + \omega\sin(h\omega)q_2^k(t-\tfrac{h}{2}) - \cos(h\omega)\hat{p}_2^k\big) =\\
&\quad = \frac{h\omega}{2}\sum_{0<|k|<N} \mathrm{i}k\Big((\hat{p}^k(t))^T\nabla_{p^k}\mathcal{K}_h(\hat{\mathbf{p}}_1(t), \mathbf{q}(t-\tfrac{h}{2})) + (q^k(t-\tfrac{h}{2}))^T\nabla_{q^k}\mathcal{K}_h(\hat{\mathbf{p}}_1(t), \mathbf{q}(t-\tfrac{h}{2}))\\
&\quad + (\hat{p}^k(t))^T\nabla_{p^k}\mathcal{K}_h(\hat{\mathbf{p}}_1(t), \mathbf{q}(t+\tfrac{h}{2})) + (q^k(t+\tfrac{h}{2}))^T\nabla_{q^k}\mathcal{K}_h(\hat{\mathbf{p}}_1(t), \mathbf{q}(t+\tfrac{h}{2}))\Big) + \mathcal{O}(h^N).
\end{aligned}$$

The left side of this equality is again a total derivative. For the right side, we have, using (6.5), $0 + \mathcal{O}(h^N)$. Thus, we get

$$\frac{\mathrm{d}}{\mathrm{dt}}\hat{\mathcal{I}}_h[\eta^k, \zeta^k](t) = \mathcal{O}(h^N),$$

and an integration from 0 to t yields the result (6.6). Like before, statement (6.7) follows from the bounds obtained in Theorem 5.5.1. □

We see that for symplectic numerical methods, we have $\mu(h\omega) = 1$ and hence $\hat{\mathcal{I}}_h[\eta_h^k, \zeta_h^k](nh) = I(p_n, q_n) + \mathcal{O}(h)$ and $\hat{\mathcal{H}}_h[\eta_h^k, \zeta_h^k](nh) = H(p_n, q_n) + \mathcal{O}(h)$. Under the additional hypothesis on the function μ

$$\mu(h\omega) \geq c_0 > 0, \tag{6.8}$$

we have the following result concerning the near conservation of (1.4) and (1.1) over long time intervals. The proof of this result is similar to the proof of Theorem 7.1. in [HLW02, Sect. XIII.7].

Theorem 5.6.3 *Under the assumptions of Theorem 5.5.1 and the additional hypothesis (6.8), if the numerical solution (p_n, q_n) stays in a compact set, we have*

$$\begin{aligned} H(p_n, q_n) &= H(p_0, q_0) + \mathcal{O}(h) \\ I(p_n, q_n) &= I(p_0, q_0) + \mathcal{O}(h), \end{aligned}$$

for $0 \leq nh \leq h^{-N+1}$.

To be able to treat numerical methods where $\mu(h\omega)$ can be small, one should look closer at the equations that determine the modulated functions and follows the approach given in [HLW02, p. 446].

Appendix A

Résumé de la thèse en français

A.1 Introduction

Ce mémoire traite de la résolution (numérique et exacte) d'équations différentielles d'ordre 2 à grandes oscillations. Derrière ce nom barbare se cache en fait l'équation suivante

$$\begin{aligned} \ddot{x}_1 &= g_1(x_1, x_2) \\ \ddot{x}_2 + \omega^2 x_2 &= g_2(x_1, x_2), \end{aligned}$$

ou sous forme matricielle

$$\ddot{x} + \Omega^2 x = g(x) \qquad \text{où} \qquad \Omega = \begin{pmatrix} 0 & 0 \\ 0 & \omega I \end{pmatrix}, \tag{1.1}$$

et $\omega \gg 1$. De plus, nous supposerons que la fonction (non-linéaire) $g : \mathbb{R}^n \to \mathbb{R}^n$ soit analytique avec toute ses dérivées bornées indépendamment de ω. Nous supposons également que les blocs de la matrice carrée Ω soient de taille arbitraire. En ce qui concerne les conditions initiales du système (1.1), nous demandons qu'elles satisfassent l'hypothèse suivante

$$\frac{1}{2}\Big(\|\dot{x}(0)\|^2 + \|\Omega x(0)\|^2\Big) \;\leq\; E, \tag{1.2}$$

où E est une constante indépendante de ω.

Ces équations différentielles apparaissent parfois en physique ou en dynamique moléculaire (nous en verrons des exemples plus précis plus loin). Si la non linéarité est sous la forme $g(x) = -\nabla U(x)$ avec une fonction $U : \mathbb{R}^n \to \mathbb{R}$, le système est Hamiltonien avec

$$H(x, \dot{x}) = \frac{1}{2}\Big(\|\dot{x}\|^2 + \|\Omega x\|^2\Big) + U(x). \tag{1.3}$$

Pour ce type de problème, nous nous sommes intéressés à la presque conservation de l'énergie oscillatoire

$$I(x, \dot{x}) = \frac{1}{2}\Big(\|\dot{x}_2\|^2 + \omega^2 \|x_2\|^2\Big) \tag{1.4}$$

pour des temps exponentiellement longs.

Une fois les propriétés de l'équation différentielle (1.1) et de sa solution connues, ils nous faut trouver une manière efficace pour résoudre ce genre de problème à l'aide de méthodes numériques.

Nous rappelons qu'une méthode numérique pour résoudre des équations différentielles $\dot{y} = f(y), y(0) = y_0$ consiste en un schéma itératif, $y_{n+1} = \Phi_h(y_n)$, nous donnant une suite de valeurs y_n, pour $n \geq 0$. Ces valeurs sont des approximations de la solution $y(t_n)$ où les $t_n = nh$ sont des points équidistants d'une subdivision de l'intervalle de temps considéré.

Par "manière efficace" de résoudre ce genre de problème, nous entendons une méthode numérique suffisamment simple à programmer mais précise et ayant de bonnes propriétés géométriques (symétrie, symplecticité, bonne conservation de l'énergie totale H et de l'énergie oscillatoire I, ...). Au lecteur n'ayant pas eu la chance de suivre le cours "Intégration géométrique" de E. Hairer, nous remémorons ces notions. Une méthode numérique $y_1 = \Phi_h(y_0)$ est symétrique si quand on échange $y_0 \leftrightarrow y_1$ et $h \leftrightarrow -h$ la méthode reste la même, i.e. $\Phi_h = \Phi_{-h}^{-1}$. Pour la symplecticité, c'est un peu plus laborieux, il faut que la méthode satisfasse

$$\Phi_h'(y)^T J \Phi_h'(y) = J \quad \text{où} \quad J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}.$$

Nous mentionons encore le fait que la symplecticité est une charactérisation typique des systèmes Hamiltoniens. Des exemples de méthodes numériques vérifiant ces propriétés sont disponibles, entre autre, dans [HLW02].

D'un point de vu théorique, Hairer et Lubich, voir [HL00], ont développé un outil, la *modulated Fourier expansion*, pour montrer la presque conservation de l'énergie oscillatoire pour des temps longs. Cet outil nous permettra de montrer, dans la prochaine section, la presque conservation de cette même énergie mais pour des temps exponentiellement longs.

Des systèmes Hamiltoniens similaires au problème Hamiltonien avec pour fonction hamiltonienne (1.3), quoique plus géneraux dans certains cas, ont été étudiés par Benettin et al. dans les deux articles [BGG87] et [BGG89]. Ces auteurs ont utilisés d'autres techniques pour montrer la presque conservation de I pour des temps exponentiellement longs.

En ce qui concerne la résolution pratique de telles équations, plusieurs types de méthodes numériques ont été élaborés pour ce genre d'équations différentielles, nous en citerons quelques unes. Commençons par une méthode très utilisée en dynamique moléculaire: la méthode de Störmer-Verlet (voir par exemple [PJY97],[HLW03] or [HLW02]). Appliquons cette méthode au problème Hamiltonien avec fonction hamiltonienne (1.3), un pas du schéma numérique s'écrit

$$\begin{aligned} \dot{x}_{n+1/2} &= \dot{x}_n - \frac{h}{2}(\Omega^2 x_n + \nabla U(x_n)) \\ x_{n+1} &= x_n + h\dot{x}_{n+1/2} \\ \dot{x}_{n+1} &= \dot{x}_{n+1/2} - \frac{h}{2}(\Omega^2 x_{n+1} + \nabla U(x_{n+1})), \end{aligned}$$

où h est la longueur du pas. Bien qu'ayant de nombreux avantages (facile à programmer, symétrie et symplecticité), cette méthode est extrêmement onéreuse lorsqu'elle est utilisée pour résoudre des équations différentielles à grandes oscillations sur de longs intervalles de temps (nous avons une restriction sur la longueur du pas h). Une alternative à la méthode de Störmer-Verlet a été proposée par plusieurs auteurs. Une majorité de ces méthodes numériques peuvent être englobées dans une seule classe de méthodes. Cette classe a pour nom: les méthodes trigonométriques (méthodes analysées dans [HLW02, Chap. XIII]). Elle englobe entre autre les méthodes de type Gautschi [HL99] et la *mollified impulse method* de Garcìa-Archilla et al. [GASSS98].

Comme première application du modèle (1.3), nous considérons le mouvement linéaire d'une molécule diatomique, p.ex. $C-C, H-H$ ou $H-O$, soumise à un champ de force externe. L'expression de l'énergie de liaison entre les deux atomes peut être décrite par la loi de Hook. Cette loi décrit le mouvement d'un ressort et s'exprime de la façon suivante: $E = \frac{\omega^2}{2}(x_2 - x_1 - r_0)^2$, où ω est la constante du ressort (grande dans notre cas), x_j, pour $j = 1, 2$, les positions des deux atomes et r_0 la distance au repos entre les deux atomes (voir Figure A.1). Pour ce problème, l'Hamiltonien est donc donné par

$$H(x_1, x_2, \dot{x}_1, \dot{x}_2) = \frac{1}{2}(\dot{x}_1^2 + \dot{x}_2^2) + \frac{\omega^2}{2}(x_2 - x_1 - x_0)^2 + U(x_1, x_2),$$

où U est un potentiel externe à la molécule.

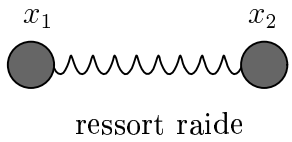


Figure A.1: Molécule diatomique.

Un changement de variables approprié ramène cet Hamiltonien sous la forme désirée (1.3). Pour ce problème, l'énergie oscillatoire I correspond en fait à l'énergie du ressort qui lie les deux atomes.

Le prochain exemple est un modèle un peu plus physique, c'est en fait une variante du problème de Fermi-Pasta-Ulam (FPU) (voir [FPU55],[Wei97],[AT87]). Nous considérons une chaîne de $2n$ points masses reliés alternativement par des ressorts durs et mous, fixée aux extrémités.

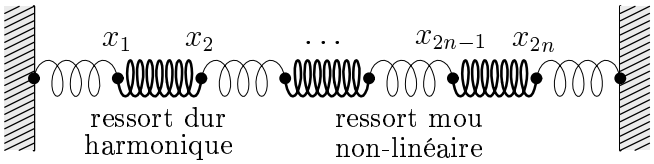


Figure A.2: Chaîne de ressorts. (@[HLW02])

Les équations du mouvement sont données (voir [HLW02, p. 17]) par l'Hamiltonien

$$\begin{aligned} H(x,\dot{x}) &= \frac{1}{2}\sum_{i=1}^{2n}\dot{x}_i^2 + \frac{\omega^2}{2}\sum_{i=1}^{n}x_{n+i}^2 + \frac{1}{4}\Big((x_1 - x_{n+1})^4 \\ &+ \sum_{i=1}^{n-1}(x_{i+1} - x_{n+i+1} - x_i - x_{n+i})^4 + (x_n + x_{2n})\Big), \end{aligned} \tag{1.5}$$

où x_i, pour $i = 1, \ldots, n$, représentent le déplacement du i-ème ressort dur, x_{i+n} la compréssion (ou décompression) de ce même ressort et ω la constante des ressorts durs. Nous allons considérer une chaîne de trois ressorts durs, dans ce problème, I correspond à l'énergie totale des trois ressorts durs, H à l'énergie totale du système et $I_j(x_{j+n}, \dot{x}_{j+n}) = \frac{1}{2}(\dot{x}_{j+n}^2 + \omega^2 x_{j+n}^2)$ à l'énergie du j-ème ressort dur. Pour ce système, nous avons utilisé une méthode très précise, appelée DOP853 (pour une définition, voir [HNW93]), pour résoudre les équations du mouvement des six points masses. La Figure A.3 montre les différentes énergies du système pour la solution numérique avec $\omega = 50$ et les valeurs initiales $x_1(0) = \dot{x}_1(0) = \dot{x}_4(0) = 1, x_4(0) = \omega^{-1}$ (les autres valeurs initiales sont nulles).

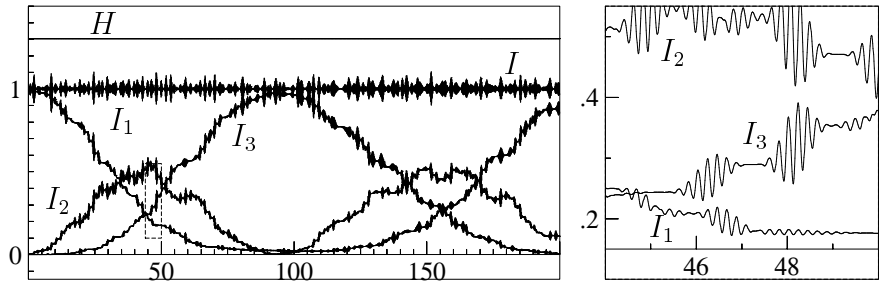


Figure A.3: Énergie totale et énergie oscillatoire du problème FPU modifié.

Nous remarquons la bonne conservation de I et de H, ainsi que l'échange des énergies entre les trois ressorts durs. Toutes ces oscillations, aussi bien dans I et les I_j que dans la solution exacte (voir Figure A.4), nous "expliquent" le terme "équation différentielle à grandes oscillations".

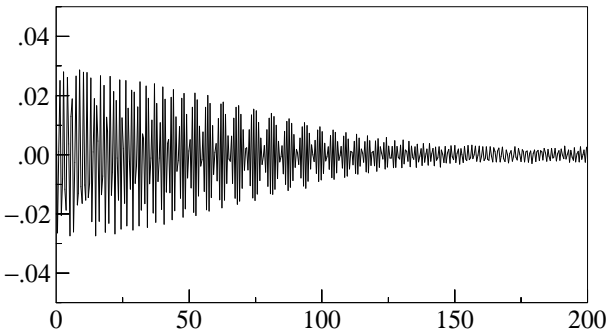


Figure A.4: Première composante rapide (x_4) du problème FPU modifié.

A.2 Equations différentielles à grandes oscillations

Dans cette section, nous présentons un résultat concernant la presque conservation de l'énergie oscillatoire I (voir (1.4)) sur des temps exponentiellement longs. Nous donnerons aussi quelques idées afin d'arriver à ce résultat.

Théorème A.2.1 *Si la solution $x(t)$ de (1.1) reste dans un compact K et si $g(x)$ est analytique et bornée par M dans un voisinage complexe*

$$D = \{x \in \mathbb{C}^n ; \|x - \xi\| \leq R \ \ \textit{pour un} \ \ \xi \in K\}.$$

De plus, si les valeurs initiales $x(0), \dot{x}(0)$ satisfont (1.2), alors, il existe des constantes positives $\gamma, C, \widehat{C}, \omega_0$ dépendantes de E, M et R (mais pas de ω) telles que, pour $\omega \geq \omega_0$, on a

$$\big|I\big(x(t), \dot{x}(t)\big) - I\big(x(0), \dot{x}(0)\big)\big| \ \leq \ C\,\omega^{-1} \qquad \textit{pour} \quad 0 \leq t \leq \widehat{C}\,\mathrm{e}^{\gamma\omega}.$$

L'outil principal pour montrer ce résultat est la *modulated Fourier expansion*. Sans rentrer trop dans les détails, il s'agit d'écrire la solution de (1.1) comme une série formelle

$$x(t) = y(t) + \sum_{k \neq 0} \mathrm{e}^{\mathrm{i}k\omega t} z^k(t), \tag{2.6}$$

où $y(t)$ et $z^k(t)$ sont des fonctions lisses (i.e., de dérivées bornées indépendament de ω). Nous rendons le lecteur attentif au fait que le k des $z^k(t)$ est un indice et que nous utilisons la notation $z^{-k} = \bar{z}_k$. Ces fonctions sont données par une équation différentielle d'ordre 2 pour y_1, par une équation différentielle d'ordre 1 pour z_2^1, par des relations algébriques pour les autres. Pour trouver ces équations qui déterminent ces fonctions, il "suffit" d'insérer (2.6) dans (1.1), de développer la fonction g en série de Taylor autour de y puis de comparer les coefficients de $\mathrm{e}^{\mathrm{i}k\omega t}$. Cela nous donne le système différentiel suivant

$$\ddot{y}_1 = \sum_{l \geq 0} \omega^{-l} F_{1l}(y_1, \dot{y}_1, z_2^1), \qquad \dot{z}_2^1 = \sum_{l \geq 1} \omega^{-l} F_{2l}(y_1, \dot{y}_1, z_2^1), \tag{2.7}$$

ainsi que les équations algébriques (pour $i = 1$, $k = 1, 2, \ldots$ et pour $i = 2$, $k = 0, 2, \ldots$ avec la convention $z^0 = y$)

$$z_i^k = \sum_{l \geq 0} \omega^{-l} G_{il}^k(y_1, \dot{y}_1, z_2^1). \tag{2.8}$$

Malheureusement ces séries divergent, il nous faut les tronquer et estimer l'erreur ainsi faite. Une fois cette tâche accomplie, nous constatons que le système qui détermine ces fonctions a deux invariants formels. Ces deux invariants sont en fait reliés à (1.3) et (1.4), et une comparaison de tous ces termes nous amène au théorème.

A.3 Méthodes numériques

La section précédente analysait la solution exacte de (1.1), il s'agit maintenant de trouver des méthodes numériques pour résoudre ce genre de problème. Nous proposons des méthodes basées sur les premiers termes de la série (2.6), nous cherchons donc des fonctions lisses $y(t)$ et $z(t)$ telles que

$$x_*(t) = y(t) + \mathrm{e}^{\mathrm{i}\omega t} z(t) + \mathrm{e}^{-\mathrm{i}\omega t} \bar{z}(t) \tag{3.9}$$

donne un petit défaut quand on l'insère dans (1.1) et satisfasse les conditions initiales

$$x_*(0) = x_0, \qquad \dot{x}_*(0) = \dot{x}_0. \tag{3.10}$$

Insérons donc (3.9) dans l'équation différentielle (1.1), un développement en série de Taylor de g autour de y et une comparaison des coefficients de $1, \mathrm{e}^{\mathrm{i}\omega t}, \mathrm{e}^{-\mathrm{i}\omega t}$ nous donne (pour les termes dominants)

$$\begin{aligned} \ddot{y}_1 &= g_1(y) + g_1''(y)(z, \bar{z}) \\ 2\mathrm{i}\omega \dot{z}_2 &= g_2'(y) z \\ \omega^2 y_2 &= g_2(y) \\ -\omega^2 z_1 &= g_1'(y) z. \end{aligned} \tag{3.11}$$

Pour la méthode numérique, on commence par trouver des valeurs initiales pour les équations différentielles contenues dans (3.11). Ensuite, on résout l'équation différentielle d'ordre 2 (par une méthode ressemblant à la méthode de Störmer-Verlet), puis celle d'ordre 1 (par une méthode ressemblant à la méthode du point milieu) et enfin on résout les équations algébriques (par des itérations successives). Une fois les fonctions y_1, y_2, z_1 et z_2 trouvées, on calcule une approximation de la solution de (1.1) par la formule (3.9).

Quatre méthodes numériques ont été développées selon la façon de tronquer la série de Taylor de la fonction g ou de négliger des termes. Toutes les méthodes numériques proposées sont d'ordre 2, symétriques, ρ-réversibles (pour une définition, voir [HLW02, Chap. V]) pour les deux applications $\rho_1(y_1, \dot{y}_1, z_2) = (y_1, -\dot{y}_1, z_{2r}, -z_{2i})$ et $\rho_2(y_1, \dot{y}_1, z_2) = (y_1, -\dot{y}_1, -z_{2r}, z_{2i})$ où z_{2r} et z_{2i} désignent la partie réelle et imaginaire de z_2. De plus,

elles conservent bien l'Hamiltonien et l'énergie oscillatoire I. Un avantage pour ces méthodes numériques par rapport aux autres méthodes est le fait que nous avons conservation uniforme des énergies H et I pour toutes valeurs de $h\omega$ où h est le pas de la méthode.

Malheureusement, ces méthodes sont implicites et nécessitent le calcul de la deuxième dérivées de la fonction g (pour certaines), elles sont donc coûteuses.

A.4 Multi-fréquences

Une fois le cas (1.1) étudié, il est naturel de regarder ce qu'il se passe pour plusieurs fréquences ω. Considérons le système

$$\begin{aligned} \ddot{x}_1 &= g_1(x_1, x_2, \ldots, x_n) \\ \ddot{x}_j + \omega_j^2 x_j &= g_j(x_1, x_2, \ldots, x_n), \quad \text{pour} \quad j = 2, \ldots, n, \end{aligned}$$

ou en notation matricielle

$$\ddot{x} + \Omega^2 x = g(x) \qquad \text{avec} \qquad \Omega = diag(0, \omega_2 I, \ldots, \omega_n I), \tag{4.12}$$

où $\omega_i = a_i \lambda$, $\lambda \gg 1$, les a_i sont des réels plus grands que un et les blocs dans la matrice Ω sont de taille arbitraire. Nous supposerons toujours que les valeurs initiales satisfassent l'hypothèse (1.2) et que la fonction g soit analytique.

L'introduction de fréquences supplémentaires et différentes dans la matrice Ω peut entraîner de la résonance dans le système (4.12). Nous commencerons par étudier le cas de non-résonance:

$$\text{il existe des constantes positives } \nu, \gamma \text{ telles que} \tag{4.13}$$

$$|k \cdot a| \geq \gamma \cdot |k|^{-\nu} \text{ pour tout } \quad k \in \mathbb{Z}^{n-1}, \, k \neq 0. \tag{4.14}$$

Ici $k \cdot a$ est le produit scalaire $(k_2 a_2 + \ldots + k_n a_n)$ et $|k| = |k_2| + \ldots + |k_n|$. Si $n = 3$, un exemple est donné en prenant $a_2 = 1$ et $a_3 = \sqrt{2}$ dans (4.12). Dans ce cas, il est possible de montrer la presque conservation des énergies oscillatoires

$$I_j(x, \dot{x}) = \frac{1}{2}\Big(\|\dot{x}_j\|^2 + \omega_j^2 \|x_j\|^2 \Big) \text{ , pour } j = 2, \ldots, n \tag{4.15}$$

sur des temps exponentiellement longs.

En ce qui concerne le cas de résonance (prenons par exemple $a_2 = 1$ et $a_3 = 2$), la durée de la presque conservation des énergies oscillatoires (4.15) dépend des valeurs que nous prenons pour les a_j. Dans ce résumé, nous ne mentionnons que le pire des cas possible (il est réalisé en prenant $a_2 = 1$ et $a_3 = 2$):

Théorème A.4.1 *Considérons l'équation différentielle (4.12) avec $1 \leq a_2 < \ldots < a_n$ et valeurs initiales $x(0)$ et $\dot{x}(0)$ satisfaisant (1.2). Si la solution $x(t)$ reste dans un ensemble compact et si λ est suffisament grand, alors, nous avons*

$$I_j(x(t), \dot{x}(t)) = I_j(x(0), \dot{x}(0)) + \mathcal{O}(\lambda^{-1}) + \mathcal{O}(t\lambda^{-2})$$

pour $j = 2, \ldots, n$ et $0 \leq t \leq \text{Const} \cdot \lambda^2$.

Pour bien voir la différence entre les résultats pour le cas de résonance et celui sans, considérons l'Hamiltonien

$$H(x_1, x_2, x_3, \dot{x}_1, \dot{x}_2, \dot{x}_3) = \frac{1}{2}\sum_{i=1}^{3} \dot{x}_i^2 + \frac{\omega_2^2}{2} x_2^2 + \frac{\omega_3^2}{2} x_3^2 + (0.001x_1 + x_2 + x_3)^4, \qquad (4.16)$$

où $\omega_2 = 1 \cdot 70$, $\omega_3 = 2 \cdot 70$, pour le cas de résonance, et $\omega_2 = 1 \cdot 70$, $\omega_3 = \sqrt{2} \cdot 70$, pour le cas de non-résonance. À nouveau, nous utilisons une méthode très précise (DOP853) et dessinons dans chaque cas les différentes énergies du problème Hamiltonien (4.16).

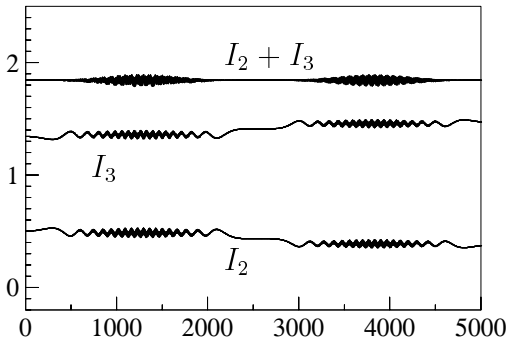


Figure A.5: Énergies oscillatoires du problème (4.16): cas de résonance.

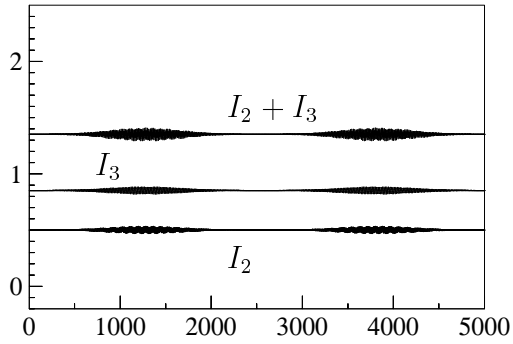


Figure A.6: Énergies oscillatoires du problème (4.16): cas de non-résonance.

Les figures A.5 et A.6 montrent bien un *drift* des énergies I_2 et I_3 dans le cas de résonance, alors que pour l'autre cas tout est bien conservé.

Les idées développées dans cette section permettent aussi de mieux comprendre la solution numérique du problème (4.12). Pour des méthodes trigonométriques, sous certaines conditions, la solution numérique admet aussi une *modulated Fourier expansion*, nous avons

$$x_n = y_h(t) + \sum_{0<|k|<N} \mathrm{e}^{\mathrm{i}k\cdot\omega t} z_h^k(t),$$

où $t = nh$ et $k \in \mathbb{Z}^n$. Comme pour la solution exacte, le système qui détermine les fonctions lisses $y_h(t)$ et $z_h^k(t)$ admet, dans le cas où $g(x) = -\nabla U(x)$, des invariants

formels. Ces derniers sont reliés à l'Hamiltonien (1.3) et aux énergies oscillatoires (4.15). Ceci nous permettra d'expliquer la presque conservation de ces énergies par les méthodes trigonométriques.

A.5 Une nouvelle classe d'Hamiltonien hautement oscillatoire

Pour certains problèmes en physique, il est possible que dans l'Hamiltonien les vitesses et positions ne soient pas aussi distinctement séparées que dans (1.3). Dans cette section, nous considérons l'Hamiltonien suivant

$$H(p,q) = K(p_1,q) + \frac{1}{2}p_2^T p_2 + \frac{\omega^2}{2}q_2^T q_2. \tag{5.17}$$

Nous supposerons bien entendu que la fonction $K(p_1,q)$ soit lisse et analytique. On dénote les variables $p=(p_1,p_2)$ et $q=(q_1,q_2)$ conformément à la partition de la matrice carrée

$$\Omega = \begin{pmatrix} 0 & 0 \\ 0 & \omega I \end{pmatrix}, \quad \omega \gg 1,$$

où les blocs de la matrice sont de dimensions arbitraires. De nouveau, les valeurs initiales du système Hamiltonien découlant de (5.17) satisfont l'hypothèse

$$\frac{1}{2}\Big(\|p(0)\|^2 + \|\Omega q(0)\|^2\Big) \leq E, \tag{5.18}$$

où E est une constante indépendante de ω.

Nous remarquons que cet Hamiltonien englobe l'Hamiltonien (1.3), en effet en prenant pour $K(p_1,q)$ la fonction $K(p_1,q) = \frac{1}{2}p_1^T p_1 + U(q)$, nous retombons sur (1.3). Il est même un peu plus général car des couplages du style $K(p_1,q) = \frac{1}{2}p_1^T M(q) p_1$, où $M(q)$ est une matrice de masse, sont aussi possibles.

Comme illustration du problème Hamiltonien (5.17), nous considérons le mouvement d'un pendule à ressort. Ce problème a été étudié par Ascher et Reich dans [AR99a] et la fonction hamiltonienne s'écrit

$$H(p,q) = \frac{1}{2}(p_2^2 + (q_2+1)^{-2}p_1^2 + q_1^2 + \omega^2 q_2^2), \tag{5.19}$$

où ω représente la constante du ressort. Ici, la composante rapide q_2 représente le déplacement de la masse connectée au ressort autour du cercle d'équilibre de rayon 1. La composante lente q_1 correspond à l'angle du pendule (voir Figure A.7).

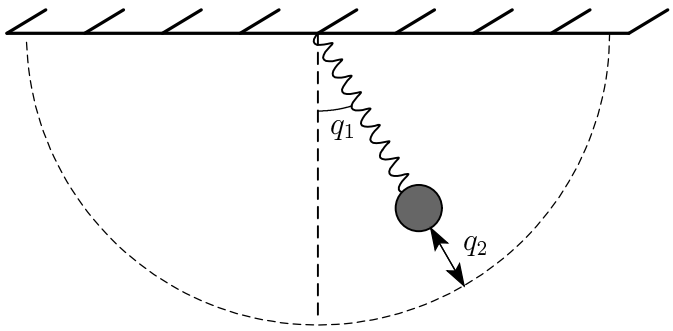


Figure A.7: Pendule harmonique.

Nous allons réutiliser la méthode numérique DOP853 pour voir ce qu'il se passe pour les différentes énergies du problème (5.19) avec valeurs initiales $p_1(0) = -\frac{1}{\sqrt{2}}, p_2(0) = \frac{1}{\sqrt{2}}, q_1(0) = 0, q_2(0) = 0$. Pour le paramètre ω, nous prenons $\omega = 80$.

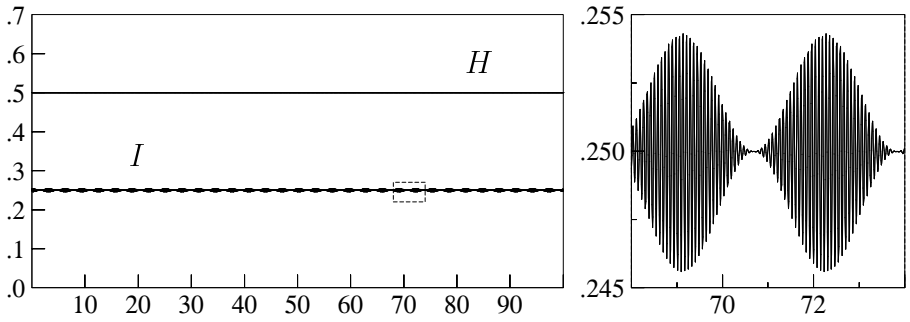


Figure A.8: Énergie totale et oscillatoire de (5.19).

De nouveau, nous constatons que l'énergie oscillatoire

$$I(p,q) = \frac{1}{2}||p_2||^2 + \frac{\omega^2}{2}||q_2||^2 \tag{5.20}$$

est presque conservée. Le zoom sur I nous montre les petites oscillations d'amplitude $\mathcal{O}(\omega^{-1})$. Une fois de plus, la *modulated Fourier* nous permettra de résoudre ce mystère.

En effet, les équations du mouvement pour l'Hamiltonien (5.17) sont données par

$$\begin{aligned} \dot{p}_1 &= -\nabla_{q_1} K(p_1, q) \\ \dot{p}_2 &= -\omega^2 q_2 - \nabla_{q_2} K(p_1, q) \\ \dot{q}_1 &= \nabla_{p_1} K(p_1, q) \\ \dot{q}_2 &= p_2, \end{aligned} \tag{5.21}$$

ou en notation matricielle

$$\begin{aligned} \dot{p} &= -\Omega^2 q + g(p_1, q) \\ \dot{q} &= \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix} p + h(p_1, q), \end{aligned} \tag{5.22}$$

avec g et h des fonctions lisses et analytiques. Quelques adaptations aux idées de la Section A.2 nous permettent de montrer le résultat suivant

Théorème A.5.1 *Si la solution $(p(t), q(t))$ de (5.17) satisfait la condition (1.2) et reste dans un ensemble compact $\tilde{K}$ pour $0 \le t \le T$, alors elle admet une* modulated Fourier expansion *de la forme*

$$\begin{aligned} p(t) &= \sum_{|k|<N} \mathrm{e}^{\mathrm{i}k\omega t} \eta^k(t) + R_N(t), \\ q(t) &= \sum_{|k|<N} \mathrm{e}^{\mathrm{i}k\omega t} \zeta^k(t) + S_N(t), \end{aligned} \tag{5.23}$$

pour un entier $N \ge 2$ arbitraire, et des fonctions lisses η^k et ζ^k (i.e. ces fonctions et toutes leurs dérivées sont bornées indépendammant de ω). De plus, les termes de reste sont bornés par

$$R_N(t) = \mathcal{O}(\omega^{-N}), \quad \text{et} \quad S_N(t) = \mathcal{O}(\omega^{-N}) \qquad \text{pour} \qquad 0 \le t \le T\ . \tag{5.24}$$

Dans ce théorème, nous pouvons aussi trouver des formules de récurrence pour déterminer les fonctions lisses $\eta^k(t)$ et $\zeta^k(t)$ et même trouver des bornes pour ces fonctions ainsi que toutes leurs dérivées. Comme dans le premier cas analysé (voir (1.3)), le système qui détermine les fonctions $\eta^k(t)$ et $\zeta^k(t)$ possède deux invariants formels, disons $\mathcal{H}(\mathbf{p}(t), \mathbf{q}(t))$ et $\mathcal{I}(\mathbf{p}(t), \mathbf{q}(t))$ (voir plus bas), qui sont reliés à l'Hamiltonien (5.17) et à l'énergie oscillatoire (5.20). Nous avons

$$\begin{aligned} \mathcal{H}(\mathbf{p}(t), \mathbf{q}(t)) &= H(p(t), q(t)) + \mathcal{O}(\omega^{-1}), \\ \mathcal{I}(\mathbf{p}(t), \mathbf{q}(t)) &= I(p(t), q(t)) + \mathcal{O}(\omega^{-1}), \end{aligned}$$

pour $0 \le t \le T$. et le vecteur $\mathbf{p} = (p^{-N+1}, \ldots, p^0, \ldots, p^{N-1})$ où $p^k = \mathrm{e}^{\mathrm{i}k\omega t}\eta^k(t)$ avec $k = 0, \ldots, N-1$. Ceci nous permet de montrer la presque conservation de I, pour le problème Hamiltonien (5.17), sur des temps longs. Une analyse plus détaillée des estimations obtenues plus haut nous permettra sûrement de montrer la presque conservation de cette énergie oscillatoire pour des temps exponentiellement longs.

Dans ce travail, nous avons aussi développé des méthodes numériques pour résoudre suffisamment bien des systèmes Hamiltoniens donnés par (5.17). Ces méthodes sont une extension assez naturelle des méthodes trigonométriques et ont pour nom les New Trigonometric Methods (NTM). Comme pour les méthodes trigonométriques, ces méthodes dépendent de fonctions filtres, elles sont données par

Définition A.5.2 *Un pas de la méthode numérique s'écrit*

$$\begin{aligned}
p_1^{n+1/2} &= p_1^n - \frac{h}{2}\nabla_{q_1} K(p_1^{n+1/2}, \Phi q^n) \\
q_1^{n+1} &= q_1^n + \frac{h}{2}\big(\nabla_{p_1} K(p_1^{n+1/2}, \Phi q^n) + \nabla_{p_1} K(p_1^{n+1/2}, \Phi q^{n+1})\big) \\
q_2^{n+1} &= \cos(h\omega) q_2^n + \omega^{-1}\sin(h\omega) p_2^n - \frac{h^2}{2}\psi_2(h\omega)\nabla_{q_2} K(p_1^{n+1/2}, \Phi q^n) \\
p_1^{n+1} &= p_1^{n+1/2} - \frac{h}{2}\nabla_{q_1} K(p_1^{n+1/2}, \Phi q^{n+1}) \\
p_2^{n+1} &= -\omega\sin(h\omega) q_2^n + \cos(h\omega) p_2^n - \frac{h}{2}\big(\tilde{\psi}_2(h\omega)\nabla_{q_2} K(p_1^{n+1/2}, \Phi q^n) \\
&+ \hat{\psi}_2(h\omega)\nabla_{q_2} K(p_1^{n+1/2}, \Phi q^{n+1})\big),
\end{aligned} \tag{5.25}$$

où $\Psi = \psi(h\Omega), \hat{\Psi} = \hat{\psi}(h\Omega), \tilde{\Psi} = \tilde{\psi}(h\Omega)$ *and* $\Phi = \phi(h\Omega)$. *L'indice* 2 *dans ces matrices correspond à la partition de la matrice* Ω. *Les fonctions filtres* $\psi, \hat{\psi}, \tilde{\psi}, \phi$ *sont réelles et paires avec* $\psi(0) = \hat{\psi}(0) = \tilde{\psi}(0) = \phi(0) = 1$.

Sans entrer trop dans les détails, le schéma numérique pour cette classe de méthodes consiste en fait à utiliser une méthode de Störmer-Verlet pour les premières composantes de p et q et une méthode trigonométrique pour les composantes rapides de ces variables. Comme pour la solution analytique, nous avons le résultat suivant

Théorème A.5.3 *Si la solution numérique* (p^n, q^n) *de (5.21) avec valeurs initiales vérifiant (1.2) satisfait la condition de non résonance*

$$|\sin(\frac{1}{2}kh\omega)| \geq c\sqrt{h}, \quad \textit{pour} \quad k = 1, \ldots, N, \quad \textit{avec} \quad N \geq 2,$$

et que les fonctions filtres de la méthode numérique vérifient

$$\begin{aligned}
|\hat{\psi}(h\omega)| &\leq C_1 \operatorname{sinc}^2(\tfrac{1}{2}h\omega), \\
|\hat{\psi}(h\omega)| &\leq C_2 |\operatorname{sinc}(h\omega)|.
\end{aligned}$$

De plus, si la solution reste dans un compact et si $h\omega \geq c_0 > 0$ *alors, la solution numérique admet, pour* $0 \leq t = nh \leq T$, *un développement*

$$\begin{aligned}
p^n &= \sum_{|k|<N} \mathrm{e}^{\mathrm{i}k\omega t}\eta_h^k(t) + R_{h,N}(t), \\
q^n &= \sum_{|k|<N} \mathrm{e}^{\mathrm{i}k\omega t}\zeta_h^k(t) + S_{h,N}(t),
\end{aligned}$$

où les termes de reste sont bornés par

$$R_{h,N}(t) = \mathcal{O}(nh^{N-1}), \quad S_{h,N}(t) = \mathcal{O}(nh^{N-1}).$$

Les coefficients sont bornés, ainsi que toutes leures dérivées, par

$$\begin{array}{llll} \zeta_{h,1} = \mathcal{O}(1), & \eta_{h,1} = \mathcal{O}(1), & \zeta_{h,2} = \mathcal{O}(\omega^{-2}), & \eta_{h,2} = \mathcal{O}(\omega^{-1}), \\ \zeta^1_{h,1} = \mathcal{O}(\omega^{-2}), & \eta^1_{h,1} = \mathcal{O}(\omega^{-2}), & \zeta^1_{h,2} = \mathcal{O}(\omega^{-1}), & \dot\eta^1_{h,2} = \mathcal{O}(\omega^{-1}), \\ \zeta^k_{h,1} = \mathcal{O}(\omega^{-k-1}), & \eta^k_{h,1} = \mathcal{O}(\omega^{-k-1}), & \zeta^k_{h,2} = \mathcal{O}(\omega^{-k-2}), & \eta^k_{h,2} = \mathcal{O}(\omega^{-k-1}), \end{array}$$

pour $k = 2, \ldots, N-1$. *De plus, nous avons* $\eta^{-k} = \overline{\eta^k}$ *et* $\zeta^{-k} = \overline{\zeta^k}$. *Les constantes symbolisées par* $\mathcal{O}$ *sont indépendantes de* ω *et* h, *mais pas de* E, N, c_0 *et* T.

Nous donnons des conditions sur les fonctions filtres pour que la méthode numérique ait des propriétés géométriques intéressantes. Suivant les cas, ces méthodes peuvent être symétriques, symplectiques, d'ordre 2 et elles conservent bien les deux quantités habituelles H et I. De plus, ces méthodes sont explicites si la fonction $K(p_1, q)$ prend la forme $K(p_1, q) = \frac{1}{2} p_1^T M(q_2) p_1 + U(q)$ (ce qui est le cas par exemple pour le pendule harmonique du début de section).

Une analyse des méthodes NTM symmétriques à l'aide de la *modulated Fourier expansion* a été effectuée. Nous montrons qu'à nouveau le système qui détermine les fonctions apparaîssant dans ce développement possède deux presques invariants notés $\hat{\mathcal{H}}_h[\eta_h^k, \zeta_h^k]$ et $\hat{\mathcal{I}}_h[\eta_h^k, \zeta_h^k]$.

Si la méthode numérique est symplectique, nous avons le résultat suivant pour la presque conservation de l'énergie oscillatoire (5.20) et de l'Hamiltonien (5.17) sur des temps longs.

Théorème A.5.4 *Sous les conditions du théorème précédent, nous avons*

$$\begin{aligned} \hat{\mathcal{H}}_h[\eta_h^k, \zeta_h^k](t) &= \hat{\mathcal{H}}_h[\eta_h^k, \zeta_h^k](0) + \mathcal{O}(th^N) \\ \hat{\mathcal{I}}_h[\eta_h^k, \zeta_h^k](t) &= \hat{\mathcal{I}}_h[\eta_h^k, \zeta_h^k](0) + \mathcal{O}(th^N) \\ \hat{\mathcal{H}}_h[\eta_h^k, \zeta_h^k](t) &= H(p_n, q_n) + \mathcal{O}(h) \\ \hat{\mathcal{I}}_h[\eta_h^k, \zeta_h^k](t) &= I(p_n, q_n) + \mathcal{O}(h), \end{aligned}$$

pour $0 \le t = nh \le T$. *Il vient donc, si la solution numérique reste dans un compact,*

$$\begin{aligned} H(p_n, q_n) &= H(p_0, q_0) + \mathcal{O}(h) \\ I(p_n, q_n) &= I(p_0, q_0) + \mathcal{O}(h), \end{aligned}$$

pour $0 \le nh \le h^{-N+1}$. *Les constantes symbolisées par* $\mathcal{O}$ *dépendent de* E, N *et* T.

Pour conclure, nous voudrions remercier le lecteur d'avoir pris du temps pour lire ce mémoire et d'être arrivé jusqu'à la dernière page ...

Laisse-toi porter par le son,
ne te pose surtout pas de question.

(NTM, *De Best*)

Bibliography

[AR99a] Uri M. Ascher and Sebastian Reich. The midpoint scheme and variants for Hamiltonian systems: advantages and pitfalls. *SIAM J. Sci. Comput.*, 21(3):1045–1065 (electronic), 1999.

[AR99b] Uri M. Ascher and Sebastian Reich. On some difficulties in integrating highly oscillatory Hamiltonian systems. In *Computational molecular dynamics: challenges, methods, ideas (Berlin, 1997)*, volume 4 of *Lect. Notes Comput. Sci. Eng.*, pages 281–296. Springer, Berlin, 1999.

[AT87] M.P. Allen and D.J. Tildesley. *Computer simulation of liquids.* Clarendon Press, Oxford, 1987.

[BG93] Dario Bambusi and Antonio Giorgilli. Exponential stability of states close to resonance in infinite-dimensional Hamiltonian systems. *J. Statist. Phys.*, 71(3-4):569–606, 1993.

[BG94] Giancarlo Benettin and Antonio Giorgilli. On the Hamiltonian interpolation of near-to-the-identity symplectic mappings with application to symplectic integration algorithms. *J. Statist. Phys.*, 74(5-6):1117–1143, 1994.

[BGG87] Giancarlo Benettin, Luigi Galgani, and Antonio Giorgilli. Realization of holonomic constraints and freezing of high frequency degrees of freedom in the light of classical perturbation theory. I. *Comm. Math. Phys.*, 113(1):87–103, 1987.

[BGG89] Giancarlo Benettin, Luigi Galgani, and Antonio Giorgilli. Realization of holonomic constraints and freezing of high frequency degrees of freedom in the light of classical perturbation theory. II. *Comm. Math. Phys.*, 121(4):557–601, 1989.

[BGG94] Giancarlo Benettin, Luigi Galgani, and Antonio Giorgilli. The dynamical foundations of classical statistical mechanics and the Boltzmann-Jeans conjecture. In *Seminar on Dynamical Systems (St. Petersburg, 1991)*, volume 12 of *Progr. Nonlinear Differential Equations Appl.*, pages 3–14. Birkhäuser, Basel, 1994.

[Car77] Henri Cartan. *Cours de calcul différentiel.* Hermann, Paris, 1977.

[Cas72] J. W. S. Cassels. *An introduction to Diophantine approximation*. Hafner Publishing Co., New York, 1972. Facsimile reprint of the 1957 edition, Cambridge Tracts in Mathematics and Mathematical Physics, No. 45.

[CHL] David Cohen, Ernst Hairer, and Christian Lubich. Numerical energy conservation for multi-frequency oscillatory differential equations. *submitted for publication*.

[CHL03] David Cohen, Ernst Hairer, and Christian Lubich. Modulated Fourier expansions of highly oscillatory differential equations. *Found. Comput. Math.*, 3(4):327–345, 2003.

[Fas90] Francesco Fassò. Lie series method for vector fields and Hamiltonian perturbation theory. *Z. Angew. Math. Phys.*, 41(6):843–864, 1990.

[FPU55] E. Fermi, J. Pasta, and S. Ulam. Studies of nonlinear problems. *Los Alamos Report No. LA-1940*, 1955.

[GASSS98] B. García-Archilla, J. M. Sanz-Serna, and R. D. Skeel. Long-time step methods for oscillatory differential equations. In *Numerical analysis 1997 (Dundee)*, volume 380 of *Pitman Res. Notes Math. Ser.*, pages 111–123. Longman, Harlow, 1998.

[GKZC04] M. Griebel, S. Knape, G. Zumbusch, and A. Caglar. *Numerische Simulation in der Moleküldynamik*. Springer-Verlag, Berlin Heidelberg, 2004.

[HL99] Marlis Hochbruck and Christian Lubich. A Gautschi-type method for oscillatory second-order differential equations. *Numer. Math.*, 83(3):403–426, 1999.

[HL00] Ernst Hairer and Christian Lubich. Long-time energy conservation of numerical methods for oscillatory differential equations. *SIAM J. Numer. Anal.*, 38(2):414–441 (electronic), 2000.

[HLW02] Ernst Hairer, Christian Lubich, and Gerhard Wanner. *Geometric numerical integration*, volume 31 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2002. Structure-preserving algorithms for ordinary differential equations.

[HLW03] Ernst Hairer, Christian Lubich, and Gerhard Wanner. Geometric numerical integration illustrated by the störmer-verlet method. *Acta Numerica*, pages 399–450, 2003.

[HNW93] E. Hairer, S. P. Nørsett, and G. Wanner. *Solving ordinary differential equations. I*, volume 8 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, second edition, 1993. Nonstiff problems.

[Lea96] Andrew R. Leach. *Molecular Modelling: Principles and Applications.* Addison Wesley Longman, Essex, England, 1996.

[LRS96] Benedict J. Leimkuhler, Sebastian Reich, and Robert D. Skeel. Integration methods for molecular dynamics. In *Mathematical approaches to biomolecular structure and dynamics (Minneapolis, MN, 1994)*, volume 82 of *IMA Vol. Math. Appl.*, pages 161–185. Springer, New York, 1996.

[Mic01] Olivier Michielin. Molecular modeling of the tcr-peptide-mhc complex: application to melanoma immunotherapy. In *MD-PHD Thesis.* Université de Lausanne, 2001.

[Mur91] James A. Murdock. *Perturbations.* A Wiley-Interscience Publication. John Wiley & Sons Inc., New York, 1991. Theory and methods.

[MvV78] W. L. Miranker and M. van Veldhuizen. The method of envelopes. *Math. Comp.*, 32(142):453–496, 1978.

[Neĭ84] A. I. Neĭshtadt. The separation of motions in systems with rapidly rotating phase. *Prikl. Mat. Mekh.*, 48(2):197–204, 1984.

[PJY97] Linda R. Petzold, Laurent O. Jay, and Jeng Yen. Numerical solution of highly oscillatory ordinary differential equations. In *Acta numerica, 1997*, volume 6 of *Acta Numer.*, pages 437–483. Cambridge Univ. Press, Cambridge, 1997.

[PT00] A. V. Pronin and D. V. Treschev. Continuous averaging in multi-frequency slow-fast systems. *Regul. Chaotic Dyn.*, 5(2):157–170, 2000.

[Wei97] Thomas P. Weissert. *The genesis of simulation in dynamics.* Springer-Verlag, New York, 1997. Pursuing the Fermi-Pasta-Ulam problem.