Highly-oscillatory evolution equations with multiple frequencies: averaging and numerics

Philippe Chartier, Mohammed Lemou, Florian Méhats

Highly-oscillatory evolution equations with multiple frequencies: averaging and numerics

Philippe Chartier[1], Mohammed Lemou [2] and Florian Méhats[3]

October 10, 2016

Abstract

In this paper, we are concerned with the application of the recently introduced *multi-revolution composition methods*, on the one hand, and *two-scale methods*, on the other hand, to a class of highly-oscillatory evolution equations with multiple frequencies. The main idea relies on a well-balanced reformulation of the problem as an equivalent mono-frequency equation which allows for the use of the two aforementioned techniques.

Keywords: highly-oscillatory evolution equation, multi-revolution composition method, two-scale method, geometric integration, asymptotic preserving scheme, uniform accuracy.

MSC numbers: 34K33, 37L05, 35Q55.

1 Introduction

This article is devoted to the numerical solution of *highly-oscillatory problems (HOPs)* by *multiscale methods*. We consider the situation where a *finite* -strictly more than one- number $d > 1$ of constant frequencies $\omega_1 < \ldots < \omega_d = 1$, occur in the problem, and assume that these frequencies are scaled with the inverse of a small parameter ε and are not all rational, thus introducing simultaneous high-oscillations in the equations. More specifically, we shall consider evolution equations of the form (with $d \geq 2$)

$$\dot{u}(t) = \frac{1}{\varepsilon}\left(\sum_{i=1}^{d} \omega_i A_i\right) u(t) + g(u(t)), \quad u(0) = u_0 \in X, \quad t \in [0,1], \tag{1}$$

where the linear operators A_i, $i = 1, \ldots, d$, commute with each other and generate 2π-periodic propagators $\tau \mapsto e^{\tau A_i}$, and where the function g is either a linear or a nonlinear map from X to itself. Since we wish to focus on the obstacles induced by the presence of several frequencies, we shall content ourselves here with *ordinary differential equations* posed in $X = \mathbb{R}^n$, though more general evolution equations could also be considered[1] and will be

[1]INRIA, ENS Rennes, IRMAR, Campus de Beaulieu, 35042 Rennes Cedex, France. Philippe.Chartier@inria.fr

[3]IRMAR University of Rennes 1 and INRIA, Campus de Beaulieu, 35042 Rennes Cedex, France.

[2]CNRS, IRMAR University of Rennes 1 and INRIA, Campus de Beaulieu, 35042 Rennes Cedex, France.

[1]Let us note however that in the application to infinite-dimensional problems, the technique we introduce here may raise difficulties that we will not comment on in this paper.

indeed used as test case in the numerical experiment section 5. A fundamental assumption throughout the paper is that the *scaled vector of frequencies* $\omega = (\omega_1, \ldots, \omega_{d-1}, 1)$ has not all its components in $\mathbb{Q}$, that is to say that $\omega \notin \mathbb{Q}^d_+$. In the sequel, we shall assume in addition the following

Assumption 1. *Equation (1) admits a uniquely defined solution for all* $0 < |\varepsilon| < \varepsilon_0 \leq 1$ *and this solution remains in an open bounded set* $\mathcal{K} \subset X$ *for all* $(t, |\varepsilon|) \in [0, 1] \times]0, \varepsilon_0[$.

Problem (1) is notoriously difficult to solve numerically: in order to achieve some accuracy, usual numerical schemes are forced to follow more and more oscillations as ε becomes smaller and smaller, whereas the *averaged* dynamics is often what only matters in applications. Standard methods such as Lie-Trotter and Strang splittings, or compositions thereof, suffer from severe step size restrictions, rendering them useless in practice for very small values of ε. More elaborate schemes of Gautschi type overcome some of the limitations of splitting techniques, but certainly not all of them (see [12], Chapter XII) and in particular are subject to resonances. It is thus of paramount importance to design effective methods.

The mono-frequency problem (i.e. equation (1) with only one operator) has drawn much attention in recent years and one has witnessed the introduction of several multiscale methods able to produce outputs with equal accuracy and cost [4, 5, 6, 7, 9], irrespect of the stiffness parameter $1/\varepsilon$. For instance, *two-scale methods (TSMs)* [5], on the one hand, and *multi-revolution composition methods (MRCMs)* [6], on the other hand, both permit to filter out the oscillations in the solution and to capture the behavior of the underlying smooth equation. These methods have been applied successfully in various contexts (ODEs but also PDEs such as kinetic equations and Schrödinger equations) and have demonstrated their ability to deliver uniformly good results in a wide range of ε-values, a property referred to as *uniform accuracy*. In this work, our goal is to rewrite the original equation (1), which is multi-frequency in essence, in such a way that the two aforementioned methods can be employed. To this aim, we shall approximate all the frequencies simultaneously by rational numbers with the same denominator. This strategy has already been successfully used in the context of homogenisation methods by several authors [1, 19] and control of PDEs [14]. In our context, it is fundamental for the diophantine approximation error to remain small -as compared to the parameter ε- and simultaneously that this common denominator also remains small. The strategy we use to balance these contradictory requirements is rather simple and will be described in Section 2. However, it requires an ad-hoc estimate which falls within number theory: its proof indeed requires continued fractions approximation for $d = 2$ and more elaborate results from [17, 18] for $d > 2$. At this point, let us emphasize that the error estimates we establish here are obviously not claimed to be a major breakthrough in the field of best diophantine approximations. Nevertheless, they appear to be novel as drawn by the specific point of view adopted here. This new formulation of the problem is then amenable to mono-frequency averaging techniques and associated numerical methods.

In Section 2, we shall present the rationale of our technique and state an averaging result in Section 3, which allows to consider problem (1) as a mono-frequency problem with a rescaled parameter ε^β for some[2] $0 < \beta < 1$. Strikingly, the estimates obtained in this non-fully-resonant scenario are qualitatively similar to those obtained by standard techniques (e.g. by filtering with the flow of the harmonic oscillators, and applying the averaging estimates

[2]The exponent β explicitly depends on the dimension d of the frequency vector and will greatly influence the efficiency of the numerical methods presented in Section 5.

of C. Simo [21]). Section 4 will deal with the required error estimates for the simultaneous approximation of the frequencies ω_i. Special attention will be paid to the dimension $d = 2$, as larger values of β can be obtained for specific values of ω_1 (namely those which can be written as a continued fraction with a bounded sequence of *coefficients*). The general situation with $d > 2$ frequencies will also be explored in this section. Finally, Section 5 will present numerical experiments for both MRCMs and TSMs. Their use in the present context will also be explained. Note that we have added an Appendix which recalls the results of [8, 11] used in this paper.

2 Motivations and method rationale

Since efficient numerical methods for mono-frequency HOPs are close at hand, the idea at the core of this work consists in reformulating equation (1) as a one-frequency HOP. Note that approximating simultaneously real numbers by rational ones with a common denominator in highly-oscillatory problems is reminiscent of previous works in the literature on homogenisation methods [1, 19] and on control of PDEs [14]. Moreover, simultaneous diophantine approximation is *per se* a thoroughly studied problem and one may find in the literature several famous related statements. However and up to our knowledge, none of them perfectly meets our requirements. In this section, we expose how this can be done appropriately in our situation and then examine the overall expected computational gain, before further commenting on existing classical results from the literature.

2.1 Rewriting the d-frequency system as a one-frequency system

We first notice that by rescaling the time (or equivalently ε), we may suppose that $\omega_d = 1$. Anticipating its proof in next section, we now use the following statement: for almost all $\omega \in]0,1]^d$ with $\omega_d = 1$ and all $0 < \alpha < 1/(d-1)$, there exists a positive constant C_ω^α such that

$$\forall P \in [1, +\infty[, \exists \mathbf{p} \in \mathbb{N}^d, \text{ s.t. } p_d \leq P,\ p_1 \wedge \ldots \wedge p_d = 1 \text{ and } \max_{i=1,\ldots,d} \left| \omega_i - \frac{p_i}{p_d} \right| \leq \frac{C_\omega^\alpha}{P^{1+\alpha}} \qquad (2)$$

where we have denoted $\mathbf{p} = (p_1, \ldots, p_d)$. The main idea of this work now consists in replacing the frequencies ω_i, $i = 1, \ldots, d$ by approximations

$$\omega_i \approx \frac{p_i}{p_d}, \quad i = 1, \ldots, d,$$

with the *same denominator* p_d as in (2). Equation (1) can then be written in a -strictly-equivalent form

$$\dot{u}(t) = \frac{1}{\varepsilon p_d} \left(\sum_{i=1}^{d} p_i A_i \right) u(t) + \frac{1}{\varepsilon} \sum_{i=1}^{d} \left(\omega_i - \frac{p_i}{p_d} \right) A_i u(t) + g\left(u(t)\right)$$

with

$$\left\| \frac{1}{\varepsilon} \sum_{i=1}^{d} \left(\omega_i - \frac{p_i}{p_d} \right) A_i u \right\|_X \leq \frac{C_\omega^\alpha}{\varepsilon P^{1+\alpha}} \left(\sum_{i=1}^{d} \|A_i\|_{\mathcal{L}(X)} \right) \|u\|_X. \qquad (3)$$

In order to get a mono-frequency highly-oscillatory problem of the form considered in [5, 6], namely

$$\dot{u}(t) = \frac{1}{\mu} A u(t) + \tilde{g}(u(t)), \quad t \in [0,1], \tag{4}$$

where $t \mapsto e^{tA}$ is 2π-periodic and $\tilde{g}$ is *uniformly* bounded for all sufficiently small ε, it thus suffices to consider the rational approximations provided by (2) for $P = \varepsilon^{-\beta}$ with $\beta := \frac{1}{1+\alpha}$, so that

$$A = \sum_{i=1}^{d} p_i A_i, \qquad \tilde{g}(u) = g(u) + \frac{1}{\varepsilon} \sum_{i=1}^{d} (\omega_i - \frac{p_i}{p_d}) A_i u \quad \text{and} \quad \mu = \varepsilon p_d.$$

We thus proceed as follows: given $\varepsilon > 0$, define $P = P^\varepsilon = \varepsilon^{-\beta}$ and choose an integer $p_d = p_d^\varepsilon \leq P$ and $d-1$ integers $p_i = p_i^\varepsilon$ satisfying estimate (2) and such that p_d is minimum. The parameter $\mu = \mu^\varepsilon = \varepsilon p_d^\varepsilon$ is then bounded by $\varepsilon P^\varepsilon = \varepsilon^{1-\beta}$ which is small as soon as ε is (given that $0 < \beta < 1$). It is then straightforward that (using that $\bar{\mathcal{K}} \subset X$ is compact, see Assumption 1)

$$\sup_{u \in \bar{\mathcal{K}}} \|\tilde{g}\| \leq \sup_{u \in \bar{\mathcal{K}}} \|g\| + C_\omega^\alpha \left(\sum_{i=1}^{d} \|A_i\|_{\mathcal{L}(X)} \right) \sup_{u \in \bar{\mathcal{K}}} \|u\| = \mathcal{O}(\varepsilon^0).$$

2.2 Expected computational speed-up

The "price to be paid", in going from equation (1) to equation (4), stems from the fact that the *averaging parameter* ε, intended to be small as it appears in (1), has been multiplied by p_d in (4). In the worst case, p_d can be of size P, so that εp_d is then of size $\varepsilon^{\frac{\alpha}{1+\alpha}}$. So to say, in passing from (1) to (4), the highly-oscillatory character of the problem has slightly faded away and the potential gain expected from multiscale methods has been reduced accordingly. However, it appears that the use of MRCMs or TSMs still allows for a significant overall gain. This can be seen as follows:

(i) on the one hand, if one solves the original equation (1)

$$\dot{u}(t) = \frac{1}{\varepsilon} \sum_{i=1}^{d} \omega_i A_i u(t) + g(u(t)), \quad t \in [0,1],$$

by a direct method (say for instance a splitting method), then the smallest period of intrinsic oscillations (that is to say $2\pi\varepsilon$, the period of $e^{\frac{t}{\varepsilon} A_d}$ given that $\omega_d = 1$ and $\omega_i < 1$, $i = 1, \ldots, d-1$) needs to be meshed with a fixed number of steps (independent of ε), say m. Altogether, the integration of (1) over the interval $[0,1]$ thus requires $m/(2\pi\varepsilon)$ steps.

(ii) on the other hand, if one first reformulates equation (1) as

$$\dot{u}(t) = \frac{1}{\mu} \sum_{i=1}^{d} p_i A_i u(t) + \tilde{g}(u(t)) = \frac{1}{\mu} A u(t) + \tilde{g}(u(t)), \quad t \in [0,1],$$

and then solves it by MRCMs or TSMs, then the solution has to be computed over a fixed number, say M (independent of μ owing to the design of these methods), of intervals of length $2\pi\mu$ (the period of $e^{\frac{t}{\mu}A}$). The integration over one period uses $p_d \times m$ steps for the p_d oscillations to be resolved as accurately as in the first case, so that computing the solution requires $M\, p_d\, m$ steps.

The computational gain is thus the ratio

$$\frac{m/(2\pi\varepsilon)}{M\ p_d\ m} = \text{Const}/(\varepsilon p_d)) \geq \text{Const}\ \varepsilon^{-\frac{\alpha}{1+\alpha}}.$$

Since $0 < \alpha < 1/(d-1)$, it clearly depends on the number d of frequencies. The expected gain is essentially of size $\text{Const}/\sqrt{\varepsilon}$ for two frequencies and deteriorates with increasing d.

2.3 Further comments on diophantine estimates from the literature

The famous Dirichlet's theorem on Diophantine approximation states that, given any vector $\omega \in]0,1]^d$ with $\omega_d = 1$, as in Subsection 2.1, and any natural number $P \geq 1$, there exists $\mathbf{p} \in \mathbb{N}^{\mathbf{d}}$ with $p_d \leq P$ such that

$$\max_{i=1,\ldots,d} \left| \omega_i - \frac{p_i}{p_d} \right| \leq \frac{1}{p_d P^\alpha} \tag{5}$$

with $\alpha = 1/(d-1)$. Its proof is a consequence of the pigeonhole principle and may be found in textbooks on arithmetic [15, 3]. However, estimate (5) is not sufficient for our purpose: as a matter of fact, the upper bound (3) is then weakened to

$$\left\| \frac{1}{\varepsilon} \sum_{i=1}^{d} \left(\omega_i - \frac{p_i}{p_d} \right) A_i u \right\|_X \quad \leq \quad \frac{1}{\varepsilon p_d P^\alpha} \left(\sum_{i=1}^{d} \|A_i\|_{\mathcal{L}(X)} \right) \|u\|_X,$$

and thus requires in essence that (i) $\varepsilon p_d P^\alpha \geq 1$ while (ii) keeping $\mu = \varepsilon p_d$ small w.r.t. ε, say of size ε^β for some $0 < \beta < 1$. In order to ensure the second condition, one has no option but to choose $\varepsilon P = \varepsilon^\beta$, since no information is provided by Dirichlet's theorem on the actual size of p_d, which may be close to 1 or quite the opposite, close to P. The inequality $\varepsilon p_d P^\alpha \geq 1$ then becomes $p_d \geq \varepsilon^{\alpha\beta - 1}$, a condition impossible to guarantee given that $\alpha\beta - 1 < 0$.

Another well-known result for the $d = 2$ case, namely the Borel-Hurwitz theorem (see for instance [15] or [19]), states that, given the irrationality of ω_1, there exists an sequence of fractions $(p_{n,1}/p_{n,2})_{n\in\mathbb{N}}$ with increasing denominators such that

$$\left| \omega_1 - \frac{p_{n,1}}{p_{2,n}} \right| \leq \frac{1}{\sqrt{5} p_{2,n}^2}.$$

At first glance, it appears to refine estimate (5) in this case. However, it does not provide estimates on the growth of $p_{2,n}$ with n. In particular, it may happen that the sequence $(p_{2,n+1} - p_{2,n})_{n\in\mathbb{N}}$ be unbounded, a scenario in which conditions (i) $\varepsilon p_{2,n}^2 \geq 1$ and (ii) $\mu = \varepsilon\, p_{2,n}$ small may be impossible to satisfy simultaneously.

The necessity of controlling the difference between consecutive common denominators was precisely the driving motivation for using and deriving estimate (2), whose proof follows mostly from standard results in arithmetic (see Subsection 4.2).

3 An averaging result for multi-frequency HOPs

In this subsection, we now establish an averaging result similar to the early paper [21] or to [11] which uses B-series.

3.1 Statement of the result

According to the discussion of Subsection 2.1, we henceforth explicitly indicate the dependence on ε of $\mathbf{p}^\varepsilon = (p_1^\varepsilon, \dots, p_d^\varepsilon)$, P^ε and $\mu^\varepsilon = \varepsilon\, p_d^\varepsilon$, by upper indices and consider the change of variables from X to itself

$$u \mapsto \chi_\theta(u) = \exp\Big(\sum_{i=1}^d \theta_i A_i\Big) u,$$

parametrized by $\theta = (\theta_1, \dots, \theta_d) \in \mathbb{T}^d \equiv [0, 2\pi]^d$. Introducing

$$G_\theta(v) = \chi_{-\theta}\Big(g\left(\chi_\theta(v)\right)\Big)$$

and performing the change of variables $u = \chi_{\frac{t}{\mu^\varepsilon}\mathbf{p}^\varepsilon}(v)$, the differential equation for v can be written

$$\dot v(t) = G_{\frac{t}{\mu^\varepsilon}\mathbf{p}^\varepsilon}\left(v(t)\right) + \frac{1}{\varepsilon}\sum_{i=1}^d \left(\omega_i - \frac{p_i^\varepsilon}{p_d^\varepsilon}\right) A_i\, v(t) := f^\varepsilon_{\frac{t}{\mu^\varepsilon}}(v(t)), \quad v(0) = v_0, \tag{6}$$

where we have used the commutation of χ_θ and the A_i's, and denoted

$$f^\varepsilon_\tau(v) = G_{\tau\mathbf{p}^\varepsilon}(v) + \frac{1}{\varepsilon}\sum_{i=1}^d \left(\omega_i - \frac{p_i^\varepsilon}{p_d^\varepsilon}\right) A_i\, v.$$

Note that since 1 is the greatest common divisor of $p_1^\varepsilon, \dots, p_d^\varepsilon$, 2π is the smallest period of the function $\tau \mapsto G_{\tau\mathbf{p}^\varepsilon}$.

We wish to study the differential equation (1) in the open bounded $\mathcal{K} \subset \mathbb{R}^n$, as defined in Introduction. Since the derivation of exponentially small error estimates in the averaging procedure requires some analyticity assumptions, we further introduce, for $\rho \geq 0$, the extended set

$$\mathcal{K}_\rho = \{v + w \in \mathbb{C}^n : v \in \bar{\mathcal{K}},\ \|w\| \leq \rho\}$$

where $\|\cdot\|$ denotes the *euclidean norm* on $\mathbb{C}^n$ as well as the induced subordinated norm for matrices of $\mathcal{M}_n(\mathbb{C})$. Finally, we denote by $\|f\|_\rho = \sup_{u\in\mathcal{K}_\rho} \|f(u)\|$ the maximum norm on the compact set $\mathcal{K}_\rho$. We are now ready to state the main hypothesis on the map $(\theta, v) \mapsto G_\theta(v)$:

Assumption 2. *There exist $R > 0$ and an open set $\mathcal{U}$ containing $\mathcal{K}_R$ such that, for all $\theta \in \mathbb{T}^d$ the function $v \mapsto G_\theta(v)$ can be extended to a map from $\mathcal{U}$ to $\mathbb{C}^n$ which is analytic at each point $v \in \mathcal{K}_R$. Furthermore, the sum of the norms of the Fourier coefficients $\hat G_{\mathbf{k}}$, $\mathbf{k} \in \mathbb{Z}^d$, of G, is bounded, i.e.*

$$M := \sum_{\mathbf{k}\in\mathbb{Z}^d} \|\hat G_{\mathbf{k}}\|_R < +\infty.$$

Note that we have

$$G_{\tau\mathbf{p}^\varepsilon}(v) = \sum_{l\in\mathbb{Z}} e^{il\tau} \left(\sum_{\mathbf{k}\in\mathbb{Z}^d,\, \mathbf{k}\cdot\mathbf{p}^\varepsilon = l} \hat{G}_{\mathbf{k}}(v) \right) \tag{7}$$

where the multi-indices $\mathbf{k} \in \mathbb{Z}^d$ in the inner-sum can be expressed under the form

$$\mathbf{k} = \mathbf{x} + S\mathbf{y}, \quad \mathbf{x} \in \mathbb{Z}^d, \quad S \in \mathcal{M}_{d,d-1}(\mathbb{Z}), \quad \mathbf{y} \in \mathbb{Z}^{d-1},$$

$\mathbf{x}$ and S being fixed values depending on l and $\mathbf{p}^\varepsilon$, while $\mathbf{y}$ takes all values in $\mathbb{Z}^{d-1}$. The series $(\|\hat{G}_{\mathbf{k}}\|_R)_{\mathbf{k}\in\mathbb{Z}^d}$ being summable, the inner series in $G_{\tau\mathbf{p}^\varepsilon}(v)$ are also convergent so that the Fourier coefficients of $G_{\tau\mathbf{p}^\varepsilon}(v)$ can be expanded as the inner series in (7). To sum up, we have

$$G_{\tau\mathbf{p}^\varepsilon}(v) = \sum_{l\in\mathbb{Z}} e^{il\tau} \hat{G}^\varepsilon_l(v) \quad \text{where} \quad \hat{G}^\varepsilon_l(v) = \sum_{\mathbf{k}\in\mathbb{Z}^d,\, \mathbf{k}\cdot\mathbf{p}^\varepsilon = l} \hat{G}_k(v). \tag{8}$$

Remark 3.1. *For instance, for $d = 2$, we have*

$$G_{\tau\mathbf{p}^\varepsilon}(v) = \sum_{l\in\mathbb{Z}} e^{il\tau} \left(\sum_{m\in\mathbb{Z}} \hat{G}_{(la^\varepsilon + mp^\varepsilon_2, lb^\varepsilon - mp^\varepsilon_1)}(v) \right)$$

where a^ε and b^ε are two integers such that $a^\varepsilon p^\varepsilon_1 + b^\varepsilon p^\varepsilon_2 = p^\varepsilon_1 \wedge p^\varepsilon_2 = 1$. All Fourier coefficients $\hat{G}_{\mathbf{k}}$, $\mathbf{k} \in \mathbb{Z}^2$, of $G_\theta(v)$ appear in this sum, but are gathered by blocks to form the Fourier coefficients of $G_{\tau\mathbf{p}^\varepsilon}(v)$.

Theorem 3.2. *Consider $\omega \in]0,1]^d$ with $\omega_d = 1$, $\omega_i < 1$ for $i = 1, \dots, d-1$ and $0 < \alpha < 1/(d-1)$ such that (2) holds for some constant C^α_ω. Suppose that G satisfies Assumption 2 and denote*

$$\tilde{M} := M + C^\alpha_\omega \sum_{i=1}^{d-1} \|A_i\| R.$$

Then, for any $0 < \varepsilon < \varepsilon_0$ and for any $N \in \mathbb{N}^*$ such that

$$\varepsilon^{\frac{\alpha}{1+\alpha}} N \le \tilde{c} := \frac{R}{8\tilde{M}},$$

there exists a near-identity (and periodic) change of variables

$$v = \Phi^{[\varepsilon,N]}_{t/\mu^\varepsilon}(V) \quad \text{with} \quad \Phi^{[\varepsilon,N]} : \mathbb{T} \times \mathcal{K}_{R/2} \to \mathcal{K}_R, \qquad \mu^\varepsilon = \varepsilon\, p^\varepsilon_d,$$

transforming equation (6) into the equation

$$\dot{V} = F^{[\varepsilon,N]}(V) + R^{[\varepsilon,N]}_{t/\mu^\varepsilon}(V), \quad V(0) = v_0,$$

with averaged vector field $F^{[\varepsilon,N]} : \mathcal{K}_{R/2} \to \mathbb{C}^n$ and remainder $R^{[\varepsilon,N]} : \mathbb{T} \times \mathcal{K}_{R/2} \to \mathbb{C}^n$ satisfying the following bounds

$$\|F^{[\varepsilon,N]} - \hat{f}^\varepsilon_0\|_{R/2} \le \frac{\tilde{M}}{2} \varepsilon^{\frac{\alpha}{1+\alpha}} \quad \text{and} \quad \forall \tau \in \mathbb{T}, \quad \|R^{[\varepsilon,N]}_\tau\|_{R/2} \le \frac{5 \left(\frac{\varepsilon^{\frac{\alpha}{1+\alpha}} N}{\tilde{c}} \right)^N}{1 - \frac{\varepsilon^{\frac{\alpha}{1+\alpha}} N}{\tilde{c}}} \tilde{M}. \tag{9}$$

In particular, taking $N = N^\varepsilon$ as the integer part of $\tilde{c}/(e\varepsilon^{\frac{\alpha}{1+\alpha}}) \geq 1$, one has

$$\forall \theta \in \mathbb{T}, \quad \|R_\theta^{[\varepsilon,N^\varepsilon]}\|_{R/2} \leq \frac{5e^2}{e-1}\tilde{M}\exp\Big(-\frac{\tilde{c}}{e}\varepsilon^{\frac{-\alpha}{1+\alpha}}\Big). \tag{10}$$

Proof. Let $0 < \varepsilon < \varepsilon_0$ and consider $\mathbf{p}^\varepsilon$ satisfying (2) for $P^\varepsilon = \varepsilon^{-\frac{1}{1+\alpha}}$. We start from equation (6)

$$\dot{v}(t) = G_{\frac{t}{\mu^\varepsilon}\mathbf{p}^\varepsilon}(v(t)) + \frac{1}{\varepsilon}\sum_{i=1}^d \left(\omega_i - \frac{p_i^\varepsilon}{p_d^\varepsilon}\right) A_i\, v(t), \quad v(0) = v_0,$$

and consider for the time being $\mu = \mu^\varepsilon$ as a small parameter varying independently of ε, while keeping ε fixed, i.e.

$$\dot{v}(t) = f_{t/\mu}^\varepsilon(v(t)), \quad v(0) = v_0, \quad t \in [0,1], \tag{11}$$

where

$$f_\tau^\varepsilon(v) = G_{\tau\mathbf{p}^\varepsilon}(v) + \frac{1}{\varepsilon}\sum_{i=1}^d \left(\omega_i - \frac{p_i^\varepsilon}{p_d^\varepsilon}\right) A_i\, v$$

is 2π-periodic owing to the choice of $\mathbf{p}^\varepsilon$. In virtue of Assumption 2, the function f_τ^ε has Fourier coefficients

$$\hat{f}_l^\varepsilon(v) = \sum_{\mathbf{k}\cdot\mathbf{p}^\varepsilon = l} \hat{G}_k(v) \quad \text{for } l \neq 0 \quad \text{and} \quad \hat{f}_0^\varepsilon(v) = \sum_{\mathbf{k}\cdot\mathbf{p}^\varepsilon = 0} \hat{G}_k(v) + \frac{1}{\varepsilon}\sum_{i=1}^d \left(\omega_i - \frac{p_i^\varepsilon}{p_d^\varepsilon}\right) A_i\, v$$

where $\mathbf{k}$ runs in $\mathbb{Z}^d$, so that

$$\sum_{l\in\mathbb{Z}} \|\hat{f}_l^\varepsilon\|_R \leq C_\omega^\alpha \sum_{i=1}^{d-1} \|A_i\|\, R + \sum_{l\in\mathbb{Z}}\sum_{\mathbf{k}\cdot\mathbf{p}^\varepsilon = l} \|\hat{G}_\mathbf{k}\|_R \leq \tilde{M}$$

where $\tilde{M}$ is **independent** of ε. Theorem 5.1 thus applies: For any $N \in \mathbb{N}^*$ and any $\mu \in \mathbb{C}$ such that $|\mu|N \leq \tilde{c} := \frac{R}{8\tilde{M}}$, there exist a vector field $V \in \mathcal{K}_{R/2} \mapsto F^{[\varepsilon,\mu,N]}(V)$, a 2π-periodic-in-time change of variables $(\tau, V) \in \mathbb{T}\times\mathcal{K}_{R/2} \mapsto \Phi_\tau^{[\varepsilon,\mu,N]}(V)$, and a 2π-periodic-in-time remainder $(\tau, V) \in \mathbb{T}\times\mathcal{K}_{R/2} \mapsto R_\tau^{[\varepsilon,\mu,N]}(V)$, such that the solution of (6) reads

$$v(t) = \Phi_{t/\mu}^{[\varepsilon,\mu,N]}(V(t))$$

where V satisfies a differential equation of the form

$$\dot{V}(t) = F^{[\varepsilon,\mu,N]}(V(t)) + R_{t/\mu}^{[\varepsilon,\mu,N]}(V(t)), \quad V(0) = v_0,$$

with the following bounds

$$\|F^{[\varepsilon,\mu,N]} - \hat{f}_0^\varepsilon\|_{R/2} \leq \frac{\tilde{M}}{2}\mu, \quad \|R_\tau^{[\varepsilon,\mu,N]}\|_{R/2} \leq \frac{5(\mu N/\tilde{c})^N}{1-(\mu N/\tilde{c})}\tilde{M}.$$

This result holds for all μ such that $|\mu|N \leq \tilde{c}$, so in particular for $\mu = \mu^\varepsilon = \varepsilon p_d^\varepsilon$ provided $\varepsilon P^\varepsilon N = \varepsilon^{\frac{\alpha}{1+\alpha}} N \leq \tilde{c}$, thus leading to the bounds given in (9). Estimate (10) is then obtained as in Theorem 5.1. □

3.2 Conserved quantities in autonomous Hamitonian systems

In this section, we consider the situation of Section 3 in [11], that is to say the case of Hamiltonian systems

$$\dot{u} = J^{-1}\nabla_u \mathcal{H}^\varepsilon(u) \tag{12}$$

where J is the *canonical* matrix

$$J = \begin{pmatrix} 0 & \mathrm{Id} \\ -\mathrm{Id} & 0 \end{pmatrix}, \quad \mathrm{Id} \in \mathcal{M}(\mathbb{R}^m),$$

and where the Hamiltonian is of the form

$$\mathcal{H}^\varepsilon(u) = \frac{1}{\varepsilon}\Big(\sum_{j=1}^{d} \omega_j I_j(u)\Big) + K(u) \tag{13}$$

with ω a vector of frequencies as considered in this paper. Furthermore, the following assumptions are satisfied:

(i) The functions I_j are in *involution*, i.e. for all $i, j = 1, \ldots, d$, one has $\{I_i, I_j\} = 0$ where the bracket used here is the Poisson bracket (see for instance [11]).

(ii) For all $j = 1, \ldots, d$, the flow $\chi_\tau^{[j]}$ of the differential system

$$\frac{d}{d\tau}\chi_\tau^{[j]}(u) = J^{-1}\nabla_u I_j(\chi_\tau^{[j]}(u))$$

is 2π-periodic.

We then denote, for $\theta \in \mathbb{T}^d$

$$\chi_\theta = \chi_{\theta_1}^{[1]} \circ \chi_{\theta_2}^{[2]} \circ \ldots \circ \chi_{\theta_d}^{[d]}$$

where the composition is commutative by virtue of the first assumption (i). In accordance with [11] again, we shall work under the following hypothesis:

Assumption 3. *There exist $R > 0$ and an open set $\mathcal{U}$ containing $\mathcal{K}_R$ such that:*

(i) for all $j = 1, \ldots, d$, I_j can be extended to an analytic map on $\mathcal{U}$;

(ii) for each $\theta \in \mathbb{T}^d$, $K \circ \chi_\theta$ can be extended to a map from $\mathcal{U}$ to $\mathbb{C}$ which is analytic at each point in $\mathcal{K}_R$.

Furthermore, the Fourier coefficients $\hat{H}_{\mathbf{k}}$, $\mathbf{k} \in \mathbb{Z}^d$, of $K \circ \chi_\theta$ satisfy the following bound

$$M := \sum_{\mathbf{k} \in \mathbb{Z}^d} \|\hat{H}_{\mathbf{k}}\|_R < +\infty.$$

We can now decompose the Hamiltonian just as we did for the vector field in previous section and write

$$\mathcal{H}^\varepsilon(u) = \frac{1}{\varepsilon p_d^\varepsilon}\Big(\sum_{j=1}^{d} p_j^\varepsilon I_j(u)\Big) + K(u) + \sum_{j=1}^{d} \frac{1}{\varepsilon}\left(\omega_j - \frac{p_j^\varepsilon}{p_d^\varepsilon}\right) I_j(u) = \frac{1}{\mu^\varepsilon} I^\varepsilon(u) + K^\varepsilon(u)$$

with

$$I^\varepsilon := \sum_{j=1}^{d} p_j^\varepsilon I_j, \qquad K^\varepsilon := K + \sum_{j=1}^{d} \frac{1}{\varepsilon}\left(\omega_j - \frac{p_j^\varepsilon}{p_d^\varepsilon}\right) I_j,$$

and where $\mathbf{p}^\varepsilon$ is chosen so as to satisfy (2) with $P^\varepsilon = \varepsilon^{-\frac{1}{1+\alpha}}$. Noticing that

$$K^\varepsilon \circ \chi_\theta = K \circ \chi_\theta + \sum_{j=1}^{d} \frac{1}{\varepsilon}\left(\omega_j - \frac{p_j^\varepsilon}{p_d^\varepsilon}\right) I_j$$

owing to assumption (i), it is clear that the Fourier coefficients $\hat{H}_l^\varepsilon$ of $K_\tau^\varepsilon := K^\varepsilon \circ \chi_{\tau \mathbf{p}^\varepsilon}$ can be written as follows

$$\hat{H}_l^\varepsilon = \sum_{\mathbf{k}\cdot\mathbf{p}^\varepsilon = l} \hat{H}_{\mathbf{k}} \quad \text{for } l \neq 0 \quad \text{and} \quad \hat{H}_0^\varepsilon = \sum_{\mathbf{k}\cdot\mathbf{p}^\varepsilon = 0} \hat{H}_{\mathbf{k}} + \sum_{j=1}^{d} \frac{1}{\varepsilon}\left(\omega_j - \frac{p_j^\varepsilon}{p_d^\varepsilon}\right) I_j$$

where $\mathbf{k}$ runs in $\mathbb{Z}^d$. Under Assumption 3, we thus have

$$\sum_{l\in\mathbb{Z}} \|\hat{H}_l^\varepsilon\|_R \leq C_\omega^\alpha \sum_{j=1}^{d-1} \|I_j\|_R + M := \tilde{M}$$

where $\tilde{M}$ is independent of ε. We can thus state the following theorem:

Theorem 3.3. *Consider $\omega \in [0,1]^d$ with $\omega_d = 1$, $\omega_i < 1$ for $i = 1, \ldots, d-1$ and $0 < \alpha < 1/(d-1)$ such that (2) holds for some constant C_ω^α. Suppose that $K \circ \chi_\theta$ satisfies Assumption 3 and let $\tilde{M}$ denote the quantity $\tilde{M} := M + C_\omega^\alpha \sum_{j=1}^{d-1} \|I_j\|_R$. Then for any $(\varepsilon, N) \in]-\varepsilon_0, \varepsilon_0[\times \mathbb{N}^*$ such that $0 < \varepsilon^{\frac{\alpha}{1+\alpha}}(N+1) \leq \frac{1}{\tilde{L}} := \frac{R^2}{8e\tilde{M}}$, the vector field $F^{[\varepsilon,N]}$ of Theorem 3.2 is Hamiltonian with Hamiltonian $\tilde{K}^{[\varepsilon,N]}$ and there exists a modified invariant $\tilde{I}^{[\varepsilon,N]}$ such*

$$\mathcal{H}^\varepsilon = \frac{1}{\varepsilon p_d^\varepsilon} \tilde{I}^{[\varepsilon,N]} + \tilde{K}^{[\varepsilon,N]}$$

where the three terms are "almost in involution" in the sense that

1. *For all $u \in \mathcal{K}$,*

$$|\{\mathcal{H}^\varepsilon(u), \tilde{I}^{[\varepsilon,N]}(u)\}| \leq \left(\frac{R}{8e}\right)^2 \left(\tilde{L}\, \varepsilon^{\frac{\alpha}{1+\alpha}}(N+1)\right)^{(N+1)}. \tag{14}$$

2. *Assume that $\tilde{L}\varepsilon^{\frac{\alpha}{1+\alpha}} \leq 1/(2e)$ and choose $N = N^{[\varepsilon]}$ as the integer part of $\tilde{L}^{-1}\varepsilon^{-\frac{\alpha}{1+\alpha}}e^{-1} - 1$. Then for all $u \in \mathcal{K}$,*

$$|\{\mathcal{H}^\varepsilon(u), \tilde{I}^{[\varepsilon,N^{[\varepsilon]}]}(u)\}| \leq \frac{\tilde{M}}{8\tilde{L}} \exp\left(-\frac{1}{e\tilde{L}\varepsilon^{\frac{\alpha}{1+\alpha}}}\right). \tag{15}$$

3.3 An illustrative example

As illustration, we consider the version of Fermi-Pasta Ulam problem discussed in [12] and used as a test problem in [10], which is of the form (12) considered in Section 3.2. It concerns a 10-dimensional Hamiltonian system with Hamiltonian function

$$\mathcal{H}^\varepsilon(u) = \frac{\lambda}{\varepsilon} I_1(u) + \frac{1}{\varepsilon} I_2(u) + K(u)$$

where $\lambda = \sqrt{2}$ and with

$$\begin{aligned}
I_1(u) &= \frac{1}{2}\left(\frac{u_4^2}{\lambda} + \lambda\, u_9^2\right), \\
I_2(u) &= \frac{1}{2}(u_1^2 + u_6^2) + \frac{1}{2}(u_2^2 + u_7^2) + \frac{1}{2}(u_3^2 + 4u_8^2), \\
K(u) &= \frac{1}{2}(u_5^2 + u_{10}^2) + \frac{1}{560} u_6^2 u_{10}^2 + \frac{1}{4900}\left(\frac{\sqrt{70}}{20} + u_6 + u_7 + \frac{5}{2}u_8 + u_9\right)^4.
\end{aligned}$$

Note that λ appearing in I_1 is considered later on as a *fixed value*: it will not be replaced in I_1 by its approximation p_1/p_2. As a result, the resulting Hamiltonian system is clearly also of the form (1) as can be seen by writing

$$\dot{u} = \frac{\lambda}{\varepsilon} A_1 u + \frac{1}{\varepsilon} A_2 u + J^{-1}\nabla_u K, \quad A_1 = J^{-1}\nabla_u^2 I_1, \quad A_2 = J^{-1}\nabla_u^2 I_2$$

where the maps $t \mapsto e^{tA_1}$ and $t \mapsto e^{tA_2}$ are 2π-periodic. According to previous section, we then split $\mathcal{H}^\varepsilon$ into two parts

$$\begin{aligned}
\mathcal{H}^\varepsilon(u) &= \frac{1}{\mu^\varepsilon}\Big(p_1^\varepsilon I_1(u) + p_2^\varepsilon I_2(u)\Big) + \Big(K(u) + \frac{(\lambda - p_1^\varepsilon/p_2^\varepsilon)}{\varepsilon} I_1(u)\Big), \\
&= \frac{1}{\mu^\varepsilon} I^\varepsilon(u) + K^\varepsilon(u),
\end{aligned}$$

with $\mu^\varepsilon = \varepsilon p_2^\varepsilon$. The change of coordinates $u = \chi_{\frac{t\mathbf{p}^\varepsilon}{\mu^\varepsilon}}(v)$ leads to the new Hamiltonian system

$$\dot{v} = \frac{(\lambda - p_1^\varepsilon/p_2^\varepsilon)}{\varepsilon} J^{-1}\nabla_v I_1(v) + J^{-1}\nabla_v K_{\frac{t\mathbf{p}^\varepsilon}{\mu^\varepsilon}}(v) \quad \text{where} \quad K_\theta = K \circ \chi_\theta.$$

Since the solution of an elementary 2-dimensional Hamiltonian system with Hamiltonian $\frac{1}{2}(\frac{x^2}{\nu} + \nu y^2)$ is given by $x(t) = \cos(t)\, x_0 - \nu \sin(t)\, y_0$ and $y(t) = (\sin(t)/\nu)\, x_0 + \cos(t)\, y_0$, the expression of $K_{(\theta_1,\theta_2)}(v)$ is obtained by replacing in $K(v)$, the coordinates as follows

$$\begin{aligned}
v_6 &\mapsto \sin(\theta_1)\, v_1 + \cos(\theta_1)\, v_6 & v_7 &\mapsto \sin(\theta_1)\, v_2 + \cos(\theta_1)\, v_7 \\
v_8 &\mapsto (\sin(2\theta_1)/2)\, v_3 + \cos(2\theta_1)\, v_8 & v_9 &\mapsto (\sin(\theta_2)/\sqrt{2})\, v_4 + \cos(\theta_2)\, v_9
\end{aligned}$$

leading to

$$\begin{aligned}
K_\theta(v) &= \frac{1}{2}(v_5^2 + v_{10}^2) + \frac{1}{560}(\sin(\theta_1)\, v_1 + \cos(\theta_1)\, v_6)^2 v_{10}^2 + \frac{1}{4900} \times \\
&\Big(\frac{\sqrt{70}}{20} + \sin(\theta_1)\, v_1 + \cos(\theta_1)\, v_6 + \sin(\theta_1)\, v_2 + \cos(\theta_1)\, v_7 + \frac{5}{2}(\sin(2\theta_1)/2\, v_3 + \cos(2\theta_1)\, v_8) \\
&+ (\sin(\theta_2)/\sqrt{2})\, v_4 + \cos(\theta_2)\, v_9\Big)^4.
\end{aligned}$$

Now, let us note that, according to Proposition 4.1 of Section 4 below, estimate (2) holds with $\alpha = 1$. Given that $\varepsilon = 1/70$, we can approximate $\sqrt{2}$ by $17/12 = \frac{p_1^\varepsilon}{p_2^\varepsilon}$ as inferred from the sequence of so-called convergents

$$1, \frac{3}{2}, \frac{7}{5}, \frac{17}{12}, \frac{41}{29}, \frac{99}{70}, \frac{239}{169}, \dots$$

for $\sqrt{2}$. We have indeed $5 < P^\varepsilon = 1/\sqrt{\varepsilon} \approx 8.366 < 12$. Regarding the error

$$\left|\sqrt{2} - \frac{7}{5}\right| \approx 0.0142 < 0.0143 \approx \frac{1}{70} = \varepsilon,$$

resulting from the rational approximation we picked up, it is clearly less than $\varepsilon C^1_{\sqrt{2}}$ (indeed, $C^1_{\sqrt{2}} \leq \frac{5}{2}$, see below).

4 Some useful error estimates on diophantine approximation

4.1 The case $d = 2$: rational approximation of a single irrational

We start by showing that, for some irrationals ω, there exists $\mathbf{p} = (p_1, p_2) \in (\mathbb{N}^*)^2$ with $p_2 \leq P$, such that an estimate of the following form

$$\left|\omega - \frac{p_1}{p_2}\right| \leq \frac{C^1_\omega}{P^2} \tag{16}$$

holds true for some positive constant C^1_ω depending on ω but not on P. If we consider the continued fraction representations $[a_0; a_1, a_2, \dots, a_n]$ of a real ω for $n = 0, 1, \dots$, two situations occur:

1. if $\omega \in \mathbb{Q}$, then there exists a finite representation, i.e.

   $$\omega = [a_0; a_1, a_2, \dots, a_j]$$

   for some $j \in \mathbb{N}$. Note that if $j > 0$ then for all $1 \leq i \leq j$, $a_i \geq 1$. Conversely, it is clear that any finite continued fraction is rational.

2. if $\omega \in \mathbb{R}\backslash\mathbb{Q}$, then ω is obtained as the limit

   $$\omega = \lim_{n\to\infty} [a_0; a_1, a_2, \dots, a_n]$$

   and for all $i \geq 1$, $a_i \geq 1$. Conversely, any infinite sequence $(a_n)_{n\in\mathbb{N}}$ with $a_i \geq 1$ for all $i \geq 1$, defines an element of $\mathbb{R}\backslash\mathbb{Q}$.

The bound (16) holds true either if $\omega \in \mathbb{Q}$ or if $\omega \in \mathbb{R}\backslash\mathbb{Q}$ and $(a_n)_{n\in\mathbb{N}}$ is bounded.

Proposition 4.1. *If either $\omega \in \mathbb{Q}_+$ or $\omega \in \mathbb{R}_+\backslash\mathbb{Q}_+$ and $(a_n)_{n\in\mathbb{N}}$ is bounded, then there exists a positive constant C^1_ω such that*

$$\forall P \in \mathbb{N}^*,\ \exists \mathbf{p} \in \mathbb{N}^2, s.t.\ p_2 \leq P,\ p_1 \wedge p_2 = 1 \quad \text{and} \quad \left|\omega - \frac{p_1}{p_2}\right| \leq \frac{C^1_\omega}{P^2}.$$

Proof. If $\omega \in \mathbb{Q}_+$ then the estimate is trivially satisfied for a sufficiently large constant C^1_ω. Otherwise, the continued fraction $[a_0; a_1, a_2, \dots,]$ defines for all $n \in \mathbb{N}$ two sequences of positive integers $(h_n)_{n\in\mathbb{N}}$ and $(k_n)_{n\in\mathbb{N}}$ such that

$$\forall n \in \mathbb{N}^*, \quad [a_0; a_1, a_2, \dots, a_n] = \frac{h_n}{k_n} \text{ with } h_n \wedge k_n = 1. \tag{17}$$

It is known that $(h_n)_{n\in\mathbb{N}}$ and $(k_n)_{n\in\mathbb{N}}$ satisfy the recurrence relations (see for instance [15])

$$\begin{aligned} h_n &= a_n h_{n-1} + h_{n-2}, \quad h_{-1} = 1, \quad h_{-2} = 0, \\ k_n &= a_n k_{n-1} + k_{n-2}, \quad k_{-1} = 0, \quad k_{-2} = 1, \end{aligned}$$

and the error estimates

$$\frac{1}{k_n(k_n + k_{n+1})} < \left|\omega - \frac{h_n}{k_n}\right| < \frac{1}{k_n k_{n+1}}. \tag{18}$$

Since $\omega \in \mathbb{R}_+\backslash\mathbb{Q}_+$, then $(k_n)_{n\in\mathbb{N}}$ is strictly increasing (owing to $a_n \geq 1$) and for all $P \in \mathbb{N}^*$, there exists k_n such that $k_n \leq P < k_{n+1}$. For this value of P, we have on the one hand

$$\frac{1}{k_n k_{n+1}} \leq \frac{1}{P^2}\frac{k_{n+1}}{k_n},$$

and on the other hand

$$\frac{k_1}{k_0} = a_1 \leq C_{max} \quad \text{and} \quad \frac{k_{n+1}}{k_n} \leq a_{n+1} + \frac{1}{a_n} \leq C_{max} + 1/C_{min} \quad \text{for} \quad n \geq 1,$$

so that one can take $C^1_\omega = C_{max} + 1/C_{min}$ where C_{max} and $C_{min} \geq 1$ are upper and lower bounds of $(a_n)_{n\in\mathbb{N}^*}$. □

Since for irrational solutions of quadratic polynomials with rational coefficients, the sequence $(a_n)_{n\in\mathbb{N}^*}$ is periodic, it is in particular bounded and (16) holds. For instance, we have $C^1_{\sqrt{2}} \leq 2 + 1/2 = 5/2$ and $C^1_{\frac{1+\sqrt{5}}{2}} \leq 1 + 1 = 2$. In contrast, e has a continued fraction with coefficients $a_{2+3n} = 2n + 2$, so that we cannot establish the existence of C^1_e with this technique. Moreover, the existence of C^1_ω can not be assumed for all reals ω, as the following proposition shows:

Proposition 4.2. *For any $0 < \alpha \leq 1$, there exist real numbers $\omega \in \mathbb{R}_+/\mathbb{Q}_+$ such that*

$$\limsup_{P\to+\infty} \left(P^{1+\alpha} \min_{\substack{(p_1,p_2)\in(\mathbb{N}^*)^2 \\ p_2 \leq P}} \left|\omega - \frac{p_1}{p_2}\right| \right) = +\infty$$

Proof. Let $(a_n)_{n\in\mathbb{N}}$ be a sequence of integers satisfying $a_n \geq 2$ for all $n \in \mathbb{N}^*$, define $(h_n)_{n\in\mathbb{N}}$ and $(k_n)_{n\in\mathbb{N}}$ by the recurrence relations

$$\begin{aligned} h_n &= a_n h_{n-1} + h_{n-2}, \quad h_{-1} = 1, \quad h_{-2} = 0, \\ k_n &= a_n k_{n-1} + k_{n-2}, \quad k_{-1} = 0, \quad k_{-2} = 1, \end{aligned}$$

and consider the corresponding irrational defined by the continued fraction

$$\omega = \lim_{n\to+\infty}[a_0; a_1, a_2, a_3, \ldots, a_n] = \lim_{n\to+\infty}\frac{h_n}{k_n}.$$

Since $(k_n)_{n\in\mathbb{N}}$ is strictly increasing, for any $P \in \mathbb{N}^*$, there exists n such that $k_n \leq P < k_{n+1}$. It is known that the best rational approximation of ω with a denominator less or equal to P is either h_n/k_n or a rational of the form

$$\frac{h_{n-1} + ah_n}{k_{n-1} + ak_n}$$

with a satisfying $a_{n+1} \geq a \geq \lfloor a_{n+1}/2 \rfloor$ and $k_{n-1} + ak_n \leq P$. If $r_{n+1} = \lfloor a_{n+1}/2 \rfloor \geq 1$ and $P = k_{n-1} + r_{n+1}k_n - 1 \geq k_n$, then the best rational approximation p_1/p_2 with $p_2 \leq P$ is $\frac{h_n}{k_n}$. For this value of P, we thus have

$$P^{1+\alpha}\left|\omega - \frac{h_n}{k_n}\right| > \frac{P^{1+\alpha}}{k_n(k_n + k_{n+1})} = \frac{(k_{n-1} + r_{n+1}k_n - 1)^{1+\alpha}}{k_n(k_n + k_{n+1})}$$

and

$$\frac{(k_{n-1} + r_{n+1}k_n - 1)^{1+\alpha}}{k_n(k_n + k_{n+1})} = \frac{(k_{n-1} + r_{n+1}k_n - 1)^{1+\alpha}}{k_n(k_{n-1} + (a_{n+1} + 1)k_n)} \geq ca_{n+1}^{\alpha}k_n^{\alpha-1}$$

for some $c > 0$ and for sufficiently large n. Taking $\delta + 1 = \lfloor 1/\alpha \rfloor + 1 > 1/\alpha$ and $a_{n+1} = k_n^{\delta}$, this gives

$$a_{n+1}^{\alpha}k_n^{\alpha-1} = k_n^{\alpha(\delta+1)-1},$$

a sequence that tends to infinity when n tends to infinity. This completes the proof. □

Now, since inequality (16) is not satisfied for all reals, the question arises whether it is true for almost all reals. Again, the answer is negative and one can additionally assert that for almost every real, (16) is not satisfied[3]. However, the following proposition holds true:

Proposition 4.3. *Let $0 < \alpha < 1$ be given. For almost every real $\omega \in [0, 1]$, there exists a positive constant C_ω^α, such that*

$$\forall P \in [0, +\infty[,\ \exists \mathbf{p} \in \mathbb{N}^2,\ p_2 \leq P,\ p_1 \wedge p_2 = 1 \quad \textit{and} \quad \left|\omega - \frac{p_1}{p_2}\right| \leq \frac{C_\omega^\alpha}{P^{1+\alpha}}. \tag{19}$$

Proof. Consider $(h_n(\omega)/k_n(\omega))_{n\in\mathbb{N}}$ the series of convergents associated with $\omega \in (\mathbb{R}_+\backslash\mathbb{Q}_+) \cap [0, 1]$, i.e. $h_n(\omega)/k_n(\omega) = [0; a_1(\omega), \ldots, a_n(\omega)]$ with

$$\omega = \lim_{n\to\infty}[0; a_1(\omega), \ldots, a_n(\omega)].$$

Given $\eta = \frac{1-\alpha}{\alpha} > 0$, define the sets $(S_n)_{n\in\mathbb{N}^*}$ by

$$S_n = \{\omega \in]0, 1[,\ k_{n+1}(\omega) > k_n(\omega)^{1+\eta}\}.$$

[3]Since this is not the main focus of this paper, we shall not further elaborate on this question.

If ω belongs to S_n, then there exists p_1 $(= h_n(\omega))$ such that $1 \leq p_1 \leq k_n(\omega)$ and satisfying

$$\left|\omega - \frac{p_1}{k_n(\omega)}\right| < \frac{1}{k_n(\omega)k_{n+1}(\omega)} < \frac{1}{k_n(\omega)^{2+\eta}},$$

so that

$$\mu(S_n) \leq \sum_{p_2 \geq \delta r^n} \sum_{1 \leq p_1 \leq p_2} \frac{2}{p_2^{2+\eta}} \leq 2 \sum_{p_2 \geq \delta r^n} \frac{1}{p_2^{1+\eta}}$$

where μ is the Lebesgue measure on $\mathbb{R}$. As a matter of fact, for $\omega \in S_n$, the strict inequality $k_{n+1}(\omega) > k_n(\omega)$ implies that $a_{n+1}(\omega) \geq 1$, so that for all $1 \leq j \leq n$, $a_j(\omega) \geq 1$ and $k_n(\omega) \geq \delta r^n$ with $\delta = \frac{1}{\sqrt{5}}$ and $r = \frac{\sqrt{5}+1}{2} > 1$. Now, since $\eta > 0$, we thus have

$$\sum_{n \geq 1} \mu(S_n) < +\infty,$$

and owing to Borel-Cantelli's theorem

$$\mu(\limsup_n S_n) = 0.$$

As a consequence, for almost every $\omega \in [0,1]$, $\omega \notin \limsup_n S_n$, that is to say, for almost every $\omega \in]0,1[$, there is only a finite number of indices $n \in \mathbb{N}^*$ such that $\omega \in S_n$. In other terms, for almost every $\omega \in]0,1[$, there exists $j(\omega) \in \mathbb{N}^*$ such that

$$\forall n \geq j(\omega), \quad k_{n+1}(\omega) \leq k_n(\omega)^{1+\eta}. \tag{20}$$

Eventually, given ω satisfying (20), and $P \geq k_{j(\omega)}$, consider $n \geq j(\omega)$ such that $k_n(\omega) \leq P < k_{n+1}(\omega)$. Then we have

$$\left|\omega - \frac{h_n(\omega)}{k_n(\omega)}\right| < \frac{1}{k_n(\omega)k_{n+1}(\omega)} \leq \frac{1}{P^{1+\frac{1}{1+\eta}}} = \frac{1}{P^{1+\alpha}}.$$

□

4.2 Simultaneous approximation of a vector of irrationals

Whenever more than one frequency have to be approximated, the situation is getting more involved. The so-called problem of *simultaneous rational approximation* is notoriously more difficult in dimension $d \geq 3$ for essentially one key-aspect, namely the absence of a continued fraction algorithm and of its associated relations. However, the result obtained in previous proposition can be generalized without too much difficulty if we content ourselves with a non-constructive sequence of best approximations, defined as follows (in the sequel, we write $r = d - 1 > 1$ and $\tilde{\omega} = (\omega_1, \ldots, \omega_{d-1})$ for a better readability) :

Definition 4.4. *Let $\tilde{\omega} \in [0,1]^r$. The strictly positive integer q is said to be a best approximation of $\tilde{\omega}$ if and only if*

$$\forall 0 < k < q, \quad \min_{\mathbf{p} \in \mathbb{Z}^r} \|q\,\tilde{\omega} - \mathbf{p}\|_\infty < \min_{\mathbf{p} \in \mathbb{Z}^r} \|k\,\tilde{\omega} - \mathbf{p}\|_\infty$$

where $\mathbf{p} = (p_1, \ldots, p_r)$.

The proof of Proposition 4.5 uses three results that we now quote separately in anticipation:

- The so-called *fundamental inequality*, obtained by several authors (see for instance [17, 18] for a slightly more general version than the one exposed here), which generalizes the error estimate (18) for k_n, states that, for $\tilde{\omega} \notin \mathbb{Q}_+^r$, *there exists a strictly growing sequence of integers* $(q_n)_{n\in\mathbb{N}}$ *(the sequence of best approximations) such that*

$$\min_{\mathbf{p}\in\mathbb{Z}^r} \|q_n\tilde{\omega} - \mathbf{p}\|_\infty^r \leq \frac{1}{q_{n+1}}.$$

- For any $\tilde{\omega} \notin \mathbb{Q}_+^r$, there exists a constant $\lambda > 1$ such that

$$\forall n \geq 0, q_n \geq \lambda^n. \tag{21}$$

  This result is a consequence of the stronger estimate derived in [17]: For any $\tilde{\omega} \notin \mathbb{Q}_+^r$,

$$\lim_{n\to+\infty} \inf(q_n)^{1/n} \geq 1 + \frac{1}{2^{r+1}}.$$

- The Borel-Cantelli's theorem which states that for any sequence of sets $A_n \subset \mathbb{R}^d$ such that

$$\sum_{n\geq 0} \mu(A_n) < +\infty$$

  one has $\mu(\limsup_{n\to+\infty} A_n) = 0$ where μ denotes here the Lebesgue measure on $\mathbb{R}^d$.

Proposition 4.5. *Let $0 < \alpha < 1/r$ be given. For almost every real $\tilde{\omega}$ in $]0,1]^r$, there exists a positive constant $C_{\tilde{\omega}}^\alpha$, such that*

$$\forall Q \in [1, +\infty[,\ \exists q \leq Q,\ \exists \mathbf{p} \in \mathbb{N}^r \text{ s.t. } p_1 \wedge \ldots \wedge p_r \wedge q = 1,\ \max_{i=1,\ldots,r} \left|\tilde{\omega}_i - \frac{p_i}{q}\right| \leq \frac{C_{\tilde{\omega}}^\alpha}{Q^{1+\alpha}}.$$

Proof. For $\tilde{\omega} \notin \mathbb{Q}_+^r$, consider the sequence $(q_n(\tilde{\omega}))_{n\in\mathbb{N}}$ of best approximations and define for $\eta = \frac{1/r-\alpha}{1-1/r+\alpha} > 0$, the sets $(A_n)_{n\in\mathbb{N}}$ by

$$A_n = \{\tilde{\omega} \in [0,1]^r,\ \tilde{\omega} \notin \mathbb{Q}_+^r,\ q_{n+1}(\tilde{\omega}) > q_n(\tilde{\omega})^{1+\eta}\}.$$

If $\tilde{\omega}$ belongs to A_n, then there exists $\mathbf{p} = (p_1, \ldots, p_r) \in \mathbb{N}^r$ such that for all $i = 1, \ldots, r$, $0 \leq p_i \leq q_n(\tilde{\omega}) - 1$ and satisfying

$$\min_{i=1,\ldots,r} \left|\tilde{\omega}_i - \frac{p_i}{q_n(\tilde{\omega})}\right| \leq \frac{1}{q_n(\tilde{\omega}) q_{n+1}^{1/r}(\tilde{\omega})} < \frac{1}{q_n(\tilde{\omega})^{\frac{r+1+\eta}{r}}}.$$

Any such $\tilde{\omega}$ belongs to a ball (w.r.t. to the $\|\cdot\|_\infty$-norm) $B(\mathbf{p}/q, \rho)$ with radius $\rho \leq 1/q_n(\tilde{\omega})^{\frac{r+1+\eta}{r}}$ and center $\mathbf{p}/q$ such that $\mathbf{p} = (p_1, \ldots, p_r)$, $1 \leq p_i \leq q$ $(i = 1, \ldots, r)$, and, owing to (21), $q \geq \lambda^n$. Hence, we have

$$\mu(A_n) \leq \sum_{q\geq\lambda^n} \sum_{1\leq p_1\leq q} \cdots \sum_{1\leq p_r\leq q} \left(\frac{2}{q^{\frac{r+1+\eta}{r}}}\right)^r \leq \sum_{q\geq\lambda^n} q^r \left(\frac{2}{q^{\frac{r+1+\eta}{r}}}\right)^r \leq 2^r \sum_{q\geq\lambda^n} \frac{1}{q^{1+\eta}}$$

where μ is the Lebesgue measure on $\mathbb{R}^r$. Now, since $\eta > 0$ and $\lambda > 1$, we have

$$\sum_{n\geq 0} \mu(A_n) < +\infty,$$

and by Borel-Cantelli's theorem

$$\mu(\limsup_n A_n) = 0.$$

As in Proposition 4.3, for almost every $\tilde{\omega} \in]0,1]^r$, there exists $j(\tilde{\omega}) \in \mathbb{N}$ such that

$$\forall n \geq j(\tilde{\omega}), \quad q_{n+1}(\tilde{\omega}) \leq q_n(\tilde{\omega})^{1+\eta}. \tag{22}$$

Eventually, given $\tilde{\omega}$ satisfying (22), and $Q \geq q_{j(\tilde{\omega})}$, consider $n \geq j(\tilde{\omega})$ such that $q_n(\tilde{\omega}) \leq Q < q_{n+1}(\tilde{\omega})$. Then we have

$$\min_{\mathbf{p}\in\mathbb{N}^r} \left\| \tilde{\omega} - \frac{\mathbf{p}}{q_n(\tilde{\omega})} \right\|_\infty < \frac{1}{q_n(\tilde{\omega}) q_{n+1}^{1/r}(\tilde{\omega})} \leq \frac{1}{Q^{\frac{1}{r}+\frac{1}{1+\eta}}} = \frac{1}{Q^{1+\alpha}}.$$

□

Taking into account the shift $r = d-1$ and the fact that ω_d is assumed to be 1, this proposition proves our estimate (2).

5 Numerical experiments

In this section, we present some numerical experiments that show the efficiency of our strategy. We shall consider two different problems and two different methods. The problems are, on the one hand, the Fermi-Pasta-Ulam described in [12] and exposed in Section 5.1, and on the other hand, a multi-component Schrödinger equation. As for the methods, we shall use, on the one hand, the multi-revolution composition method (MRCM), introduced in [6], and on the other hand, the two-scale method (TSM) introduced in [5]. Both methods have been originally designed for mono-frequency problems and, in order to handle the two aforementioned test-cases, we apply the strategy exposed in this paper. Let us briefly present the main ideas underlying the two techniques:

(i) MCRMs: the flow corresponding to the integration over **one period of time** of a differential equation of the form $\dot{u}^\varepsilon = f_{t/\varepsilon}(u^\varepsilon)$ (with f periodic in t/ε) is a near-identity map $\varphi_\varepsilon : \mathbb{R}^m \to \mathbb{R}^m$. Computing the exact solution over N periods thus amounts to computing the N-th iterate φ_ε^N of φ_ε. The idea of MRCMs consists in approximating φ_ε^N by a composition of the form

$$\varphi_\varepsilon^N = \varphi_{\alpha_1 H} \circ \varphi^*_{\beta_1 H} \circ \cdots \circ \varphi_{\alpha_s H} \circ \varphi^*_{\beta_s H} + \mathcal{O}(\varepsilon^{p+1}), \qquad H = N\varepsilon,$$

where $\varphi_\varepsilon^* := (\varphi_{-\varepsilon})^{-1}$ and where p is made as high as possible by choosing appropriate coefficients α's and β's (and letting them depend on N). Whenever $s \ll N$, the computational effort is considerably reduced. In fact, a careful analysis shows that for ε small enough, the overall cost is independent of ε, whereas it typically grows like $1/\varepsilon$ for standard integration methods. Here, we shall use the fourth-order ($p = 4$) MRCM of [6], where φ_ε itself is approximated by a Strang splitting method.

(ii) TSMs: in two-scale methods for $\ddot{u}^\varepsilon = f_{t/\varepsilon}(u^\varepsilon)$, the solution is sought as the diagonal $\tau = t/\varepsilon$ of an approximation of $U^\varepsilon(t,\tau)$ satisfying the transport equation

$$\partial_t U^\varepsilon(t,\tau) + \frac{1}{\varepsilon}\partial_\tau U^\varepsilon(t,\tau) = f_\tau(U^\varepsilon(t,\tau)).$$

The main idea is then that the initial condition $U^\varepsilon(0,\tau)$ can be chosen in such a way that all derivatives of $U^\varepsilon(t,\tau)$ remain bounded w.r.t. to ε up to some arbitrary order p, thus allowing for the construction of **uniformly accurate methods** of order $p-1$. In this section, we shall consider the uniformly second-order method obtained in [5].

Both techniques have been designed for mono-frequency highly-oscillatory problems, the first one (MRCM) in the context of ordinary differential equations and the second one (TSM) originally for kinetic equations and later on for the Schrödinger equation. In the aforementioned situations, they are capable of delivering numerical approximations with constant accuracy and constant cost w.r.t. ε in the limit where ε tends to zero. MRCMs are in addition provably *geometric*, while TSMs are not, even though they often behave likewise. The situation is reversed as far as *uniform accuracy* is concerned: MRCMs are strictly speaking not uniformly accurate, while TSMs are. It is thus enlightening to study whether MRCMs preserve, as predicted in this paper, the energy of Hamiltonian systems, and similarly to test whether TCMs behave correctly. The other part of our tests aims at assessing the extent to which TCMs remain uniformly accurate. The Strang method with tiny step-sizes is used here to obtain a very accurate reference solution in all experiments. In comparison, both MRCMs and TSMs become competitive for $\varepsilon \leq 10^{-4}$ with the FPU problem and $\varepsilon \leq 10^{-3}$ with the system of coupled Schrödinger equations.

5.1 A Fermi-Pasta-Ulam system with two frequencies

In this subsection, we consider the Hamiltonian system with a finite degrees of freedom $q \in \mathbb{R}^5$, $p \in \mathbb{R}^5$, borrowed from [12] and used in [10]:

$$H(p,q) = \lambda_1 \left(\frac{p_1^2}{2} + \frac{q_1^2}{2}\right) + \sum_{j=2}^{5} \left(\frac{p_j^2}{2\varepsilon} + \frac{\lambda_j^2 q_j^2}{2\varepsilon}\right) + U(q),$$

$$U(q) = \frac{\delta^2}{8} q_1^2 q_2^2 + \delta^4 \left(\frac{\sqrt{70}}{20} + q_2 + q_3 + \frac{5}{2} q_4 + q_5\right)^4,$$

with

$$\lambda_1 = 1, \qquad \lambda_2 = \lambda_3 = 1, \qquad \lambda_4 = 2, \qquad \lambda_5 = \sqrt{2}$$

and

$$\delta = 1/70, \quad q(0) = (1, 0.3\delta, 0.8\delta, 0.7\delta, -1.1\delta), \qquad p(0) = (-0.2, 0.6\delta, 0.7\delta, 0.8\delta, -0.9\delta).$$

Energy exchanges

We observe the evolution over a long time of the quantities

$$I_1 = \lambda_1 \left(\frac{p_1^2}{2} + \frac{q_1^2}{2}\right), \qquad I_j = \frac{p_1^2}{2} + \frac{\lambda_j^2 q_1^2}{2}, \quad j = 2, \ldots, 5,$$

computed with three methods:

– the Strang splitting method (Figure 1) with the time-step $\Delta t = \frac{2\pi}{16}\varepsilon$; such a step makes the approximation accurate enough to regard the solution as 'exact'. This is indeed our reference solution.

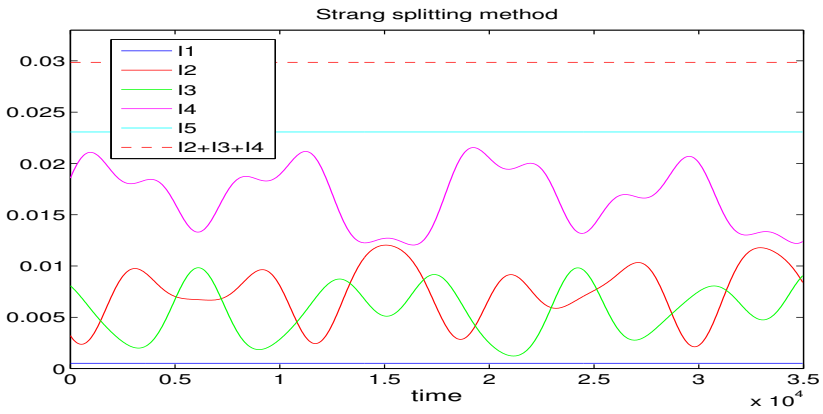

Figure 1: (FPU, $d = 1$) Energy exchanges, Strang splitting method

– the MCRM [6] of order 4, with $N = 60$ and with the micro time-step $\Delta t = \frac{2\pi}{16}q$ for the micro-integrator over one period $[0, 2\pi]$, which represents a computational gain of a factor 10 compared to the direct Strang splitting method (Figure 2);

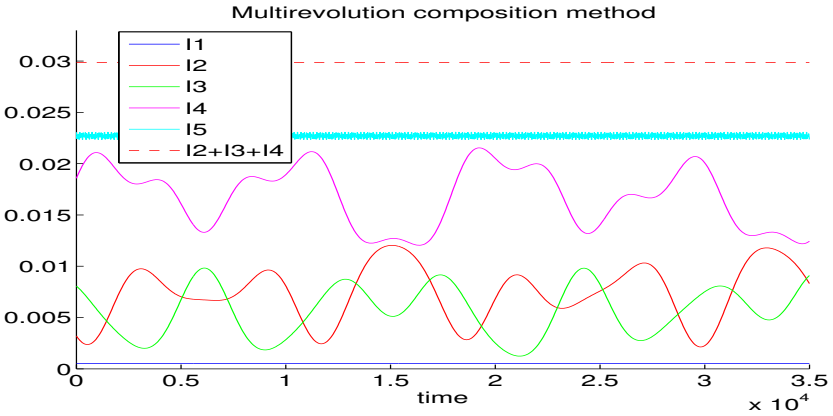

Figure 2: (FPU, $d = 1$) Energy exchanges, multirevolution composition method

– the TSM [5] of second-order, implemented with the implicit mid-point scheme (Figure 3), with the time-step $\Delta t = \frac{2\pi}{16}$ and 32 discretization points in the variable τ.

We take $\varepsilon = 10^{-3}$ and, for the two latter methods, $p = 41$ and $q = 29$. It is apparent from Figure 2 and Figure 3 that the energy exchanges are well reproduced by both the MRCM and the TSM. In next section, we now investigate the accuracy of both methods.

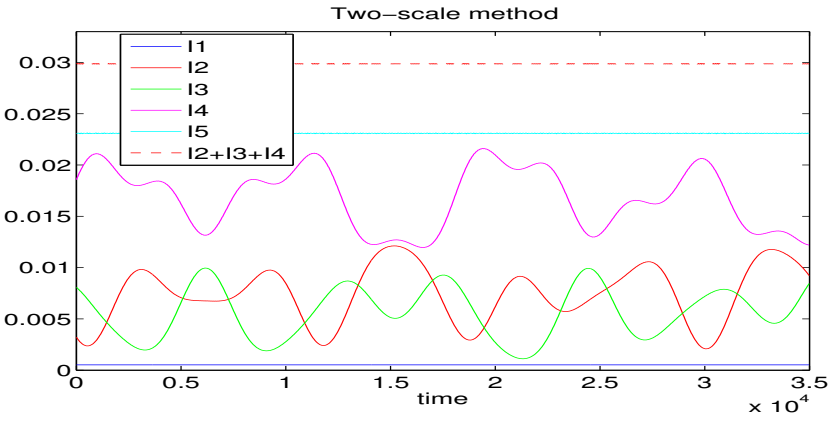

Figure 3: (FPU, $d = 1$) Energy exchanges, two-scale method

Accuracy of the MRCM

On Figure 4, we plot the error for the MRCMs of order 1, 2 and 4 of [6], as a function of the macrostep $H = q\varepsilon N$ for two values of ε ($\varepsilon = 4 \times 10^{-5}$ and $\varepsilon = 10^{-5}$). For varying ε, note

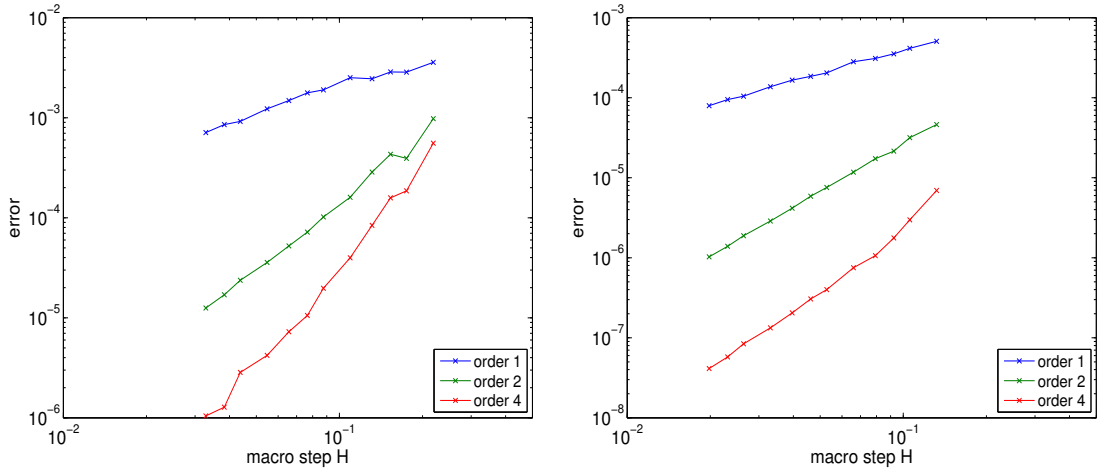

Figure 4: (FPU, $d = 1$) Error versus macrostep for $\epsilon = 4 \times 10^{-5}$ (left) and $\epsilon = 10^{-5}$ (right)

that $q\varepsilon \approx \sqrt{\varepsilon}$ and that by choosing $N \approx 1/\sqrt{\varepsilon}$ the error remains essentially constant while the computational cost grows like $1/\sqrt{\varepsilon}$. This, of course, compares favorably with the $1/\varepsilon$ increase observed for standard methods such as Strang.

Uniform accuracy of the two-scale method

The goal is here to observe the uniform second order accuracy of the TSM. The final time is taken equal to 2π and the number of discretization points for the τ-variable is $\max(64, 16p)$. The values of p and q in the approximation $\frac{p}{q}$ of $\lambda_5 = \sqrt{2}$, as well as the value of the remainder

$Err := \frac{1}{\varepsilon}|\sqrt{2} - \frac{p}{q}|$, are given in the following table: they are obtained by a continued fraction algorithm.

ε	0.64	0.32	0.16	0.08	0.04	0.02	0.01	0.005	0.0025	0.00125	0.000625
p	1	1	3	3	7	7	7	17	17	17	41
q	1	1	2	2	5	5	5	12	12	12	29
Err	0.64	1.29	0.54	1.07	0.36	0.71	1.42	0.49	0.98	1.96	0.67

ε	$3.12 * 10^{-4}$	$1.56 * 10^{-4}$	$7.81 * 10^{-5}$	$3.91 * 10^{-5}$	$1.95 * 10^{-5}$	$9.77 * 10^{-6}$
p	41	99	99	99	239	239
q	29	70	70	70	169	169
Err	1.34	0.46	0.92	1.85	0.63	1.27

On the left picture of Figure 5, we plot the error as a function of the time-step Δt, for different values of ε and on right picture of Figure 5, the error as a function of ε, for different values of the time-step Δt. This is in perfect agreement with the predicted uniform accuracy of the method. Note that the small peaks correspond to the highest values of the remainder $\frac{1}{\varepsilon}|\sqrt{2} - \frac{p}{q}|$.

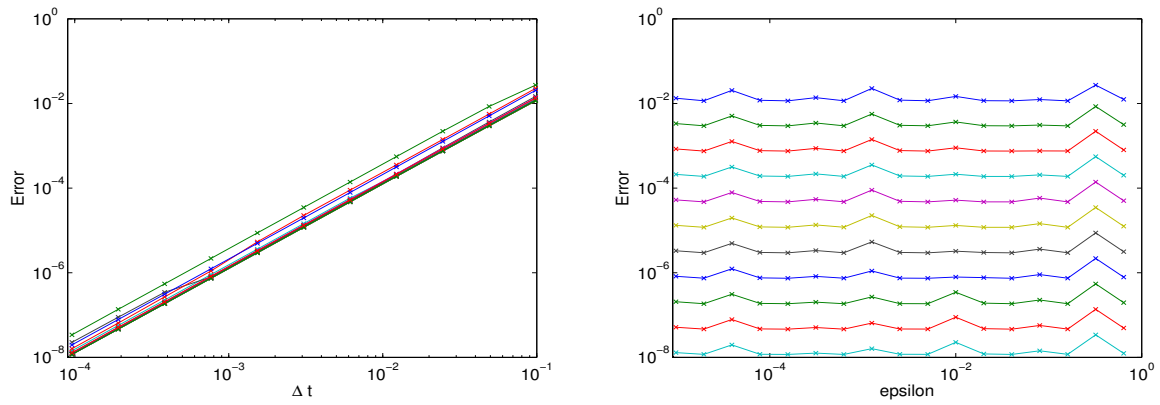

Figure 5: (FPU, $d = 1$) Left: Error as a function of Δt for $\varepsilon = 2^N \times 10^{-2}$, with $N \in \{6, \ldots, 1, 0, -1, \ldots, -10\}$. Right: Error as a function of ε for $\Delta t = 2\pi/2^N$ with $N \in \{6, \ldots, 16\}$

5.2 A Fermi-Pasta-Ulam system with three frequencies

In this subsection, we consider the same FPU system as in the above Subsection 5.1, but with the frequencies

$$\lambda_1 = 1, \qquad \lambda_2 = \lambda_3 = 1, \qquad \lambda_4 = \frac{\pi}{2}, \qquad \lambda_5 = \sqrt{2}.$$

Energy exchanges

We observe the evolution over a long time of the quantities

$$I_1 = \lambda_1 \left(\frac{p_1^2}{2} + \frac{q_1^2}{2} \right), \qquad I_j = \frac{p_1^2}{2} + \frac{\lambda_j^2 q_1^2}{2}, \quad j = 2, \ldots, 5,$$

computed with again the three methods:

- the Strang splitting method (Figure 6) with the time-step $\Delta t = \frac{2\pi}{16}\varepsilon$;
- the MCRM [6] of order 4, with $N = 60$ and with the micro time-step $\Delta t = \frac{2\pi}{16q}$ for the micro-integrator over one period $[0, 2\pi]$, which represents again a computational speed-up of 10 compared to the direct Strang splitting method (Figure 7);
- the TSM [5], implemented with the implicit second-order mid-point scheme (Figure 8), with time-step $\Delta t = \frac{2\pi}{16}$ and 64 discretization points in the variable τ. We observe a computational speed-up of 2.5 compared to the Strang splitting method.

For these simulations, we have taken $\varepsilon = 10^{-3}$, and the rational approximations $\lambda_4 = \frac{\pi}{2} \approx \frac{110}{70}$ and $\lambda_5 = \sqrt{2} \approx \frac{99}{70}$, which give

$$\frac{1}{\varepsilon}\left|\lambda_4 - \frac{p_4}{q}\right| \approx 0.63, \qquad \frac{1}{\varepsilon}\left|\lambda_5 - \frac{p_5}{q}\right| \approx 0.07.$$

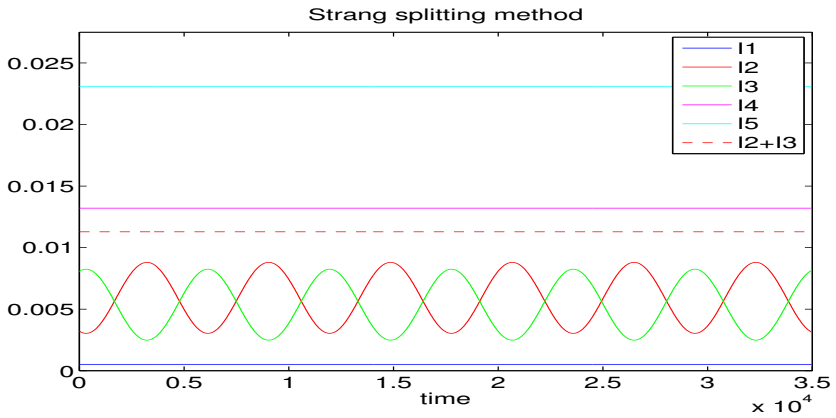

Figure 6: (FPU, $d = 2$) Energy exchanges, Strang splitting method

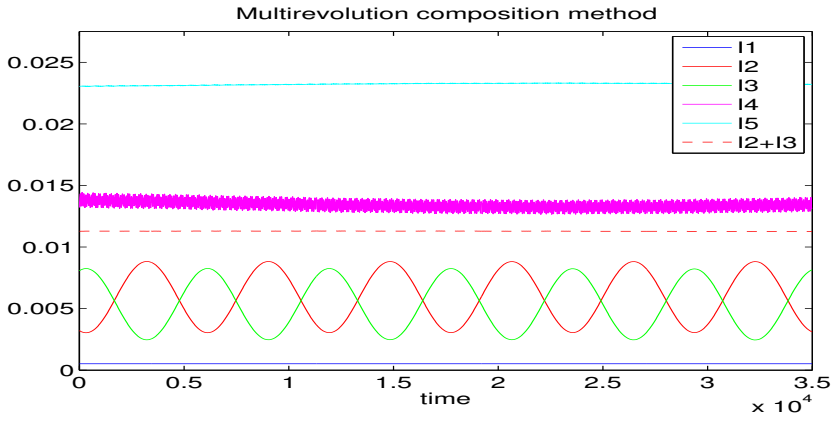

Figure 7: (FPU, $d = 2$) Energy exchanges, multirevolution composition method

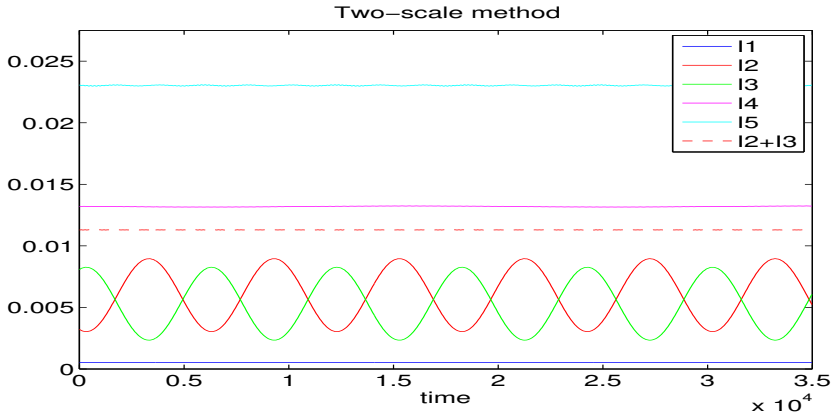

Figure 8: (FPU, $d = 2$) Energy exchanges, two-scale method

Uniform accuracy of the two-scale method

Again, we wish here to observe the uniform accuracy of the two-scale method. The final time is 2π, the number of discretization points for the τ variable is $\max(64, 16q)$. The values of p_4 p_5 and q in the approximations $\frac{p_4}{q}$ and $\frac{p_5}{q}$ of $\lambda_4 = \frac{\pi}{2}$ and $\lambda_5 = \sqrt{2}$, as well as the remainder $\frac{1}{\varepsilon}|\lambda_4 - \frac{p_4}{q}| + \frac{1}{\varepsilon}|\lambda_5 - \frac{p_5}{q}|$ are given in the following table. These values are those which minimize this remainder under the constraint $\varepsilon q^{3/2} \leq 1$.

ε	0.64	0.32	0.16	0.08	0.04	0.02	0.01	5×10^{-3}	2.5×10^{-3}	1.25×10^{-3}
p_4	2	3	3	8	11	11	11	53	80	110
p_5	1	3	3	7	10	10	10	48	72	99
q	1	2	2	5	7	7	7	34	51	70
remainder	1.32	0.49	0.98	0.54	0.37	0.75	1.5	2.88	1.85	0.56

ε	6.25×10^{-4}	3.12×10^{-4}	1.56×10^{-4}	7.81×10^{-5}	3.91×10^{-5}
p_4	201	311	421	732	732
p_5	181	280	379	659	659
q	128	198	268	466	466
remainder	1.02	0.52	0.86	0.89	1.78

On the left of Figure 9, we plot the error as a function of the time-step Δt, for different values of ε and on the right of Figure 9, the error as a function of ε, for different values of the time-step Δt.

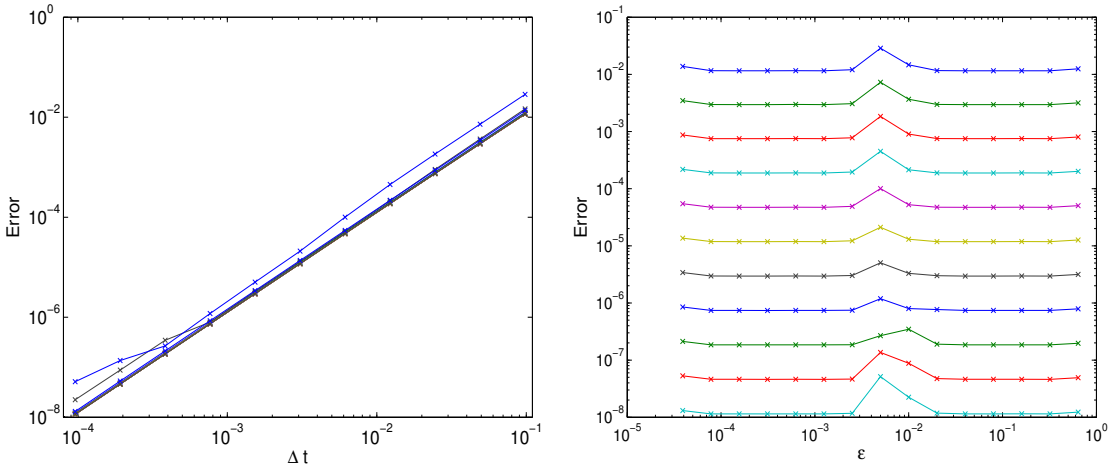

Figure 9: (FPU, $d = 1$) Left: error as a function of Δt for $\varepsilon = 2^N \times 10^{-2}$, with $N \in \{6, \ldots, 1, 0, -1, \ldots, -10\}$. Right: error as a function of ε for $\Delta t = 2\pi/2^N$ with $N \in \{6, \ldots, 16\}$

5.3 Three coupled nonlinear Schrödinger equations

In this section, we consider a multi-component non-linear Schrödinger system posed in infinite dimension, which models multi-component Bose-Einstein condensates. Roughly speaking, harmonic oscillators are here replaced by Laplacian operators with periodic boundary conditions. The interested reader may find more details on the physical aspects of the model under consideration in references [2] and [13]. The important point therein is that different components may have different "trapping" potentials and thus oscillate with different frequencies. In this section, we shall use as test problem the following coupled system of three non-linear Schrödinger equations, where the components u_1, u_2 and u_3 are discretized in x by trigonometric polynomials (accordingly Fast Fourier Transform (FFT) is used in our numerical experiments):

$$i\partial_t u_1(t,x) = -\frac{\omega_1}{\varepsilon}\Delta u_1(t,x) + \left(\alpha_{11}(x)|u_1(t,x)|^2 + \alpha_{12}(x)|u_2(t,x)|^2 + \alpha_{13}(x)|u_3(t,x)|^2\right) u_1(t,x)$$
$$i\partial_t u_2(t,x) = -\frac{\omega_2}{\varepsilon}\Delta u_2(t,x) + \left(\alpha_{12}(x)|u_1(t,x)|^2 + \alpha_{22}(x)|u_2(t,x)|^2 + \alpha_{23}(x)|u_3(t,x)|^2\right) u_2(t,x)$$
$$i\partial_t u_3(t,x) = -\frac{\omega_3}{\varepsilon}\Delta u_3(t,x) + \left(\alpha_{13}(x)|u_1(t,x)|^2 + \alpha_{23}(x)|u_2(t,x)|^2 + \alpha_{33}(x)|u_3(t,x)|^2\right) u_3(t,x)$$

on the interval $[0, 2\pi]$, with periodic boundary conditions and the following set of coefficients:

$$\omega_1 = \omega_2 = 1, \quad \omega_3 = \sqrt{2}, \quad \alpha_{11}(x) = 2\cos(2x), \quad \alpha_{12} = \alpha_{13} = \alpha_{22} = \alpha_{23} \equiv 1,$$

and with initial data

$$u_1(0,x) = \frac{1}{2} + \frac{4}{10}e^{-ix}, \quad u_2(0,x) = \frac{1}{4} + \frac{4}{10}e^{ix}, \quad u_3(0,x) = \frac{1}{4} + \frac{6}{10}e^{ix}.$$

Energy exchanges

We observe on Figures 10 and 11 the evolution of the total energy and of

$$I_j = \omega_j \int_0^{2\pi} |\nabla u_j|^2 dx,$$

computed by three methods, for $\varepsilon = 10^{-4}$:

- the Strang splitting method with the time-step $\Delta t = \frac{2\pi}{32}\varepsilon$;
- the MCRM [6] of order 4, with $N = 60$ and with the micro time-step $\Delta t = \frac{2\pi}{32q}$ for the micro-integrator over one period $[0, 2\pi/q]$, which represents a computational gain of a factor 10 compared to the direct Strang splitting method;
- the TSM [5], with micro time step $T/128$ and 2048 discretization points in τ. We observe a computational gain of a factor 5 compared to the Strang splitting method.

For the three methods, we take 32 discretization points for the x variable.

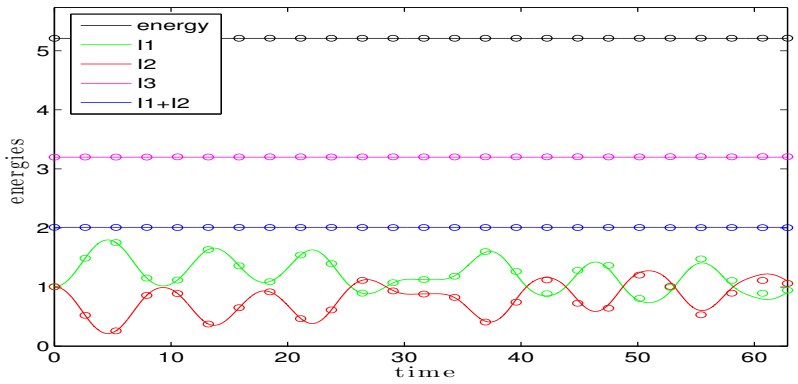

Figure 10: Energy exchanges for $\varepsilon = 0.0001$, computed with the Strang splitting method (plain lines) and the MRCM (circles) with $N = 60$

Uniform accuracy the two-scale method

Finally, we check here the uniform accuracy of the two-scale method. The final time is 0.2, the number of discretization points is 32 in x and 4096 in τ. The values of p and q in the approximation $\frac{p}{q}$ of $\sqrt{2}$ are the same as given in table of Subsection 5.1.

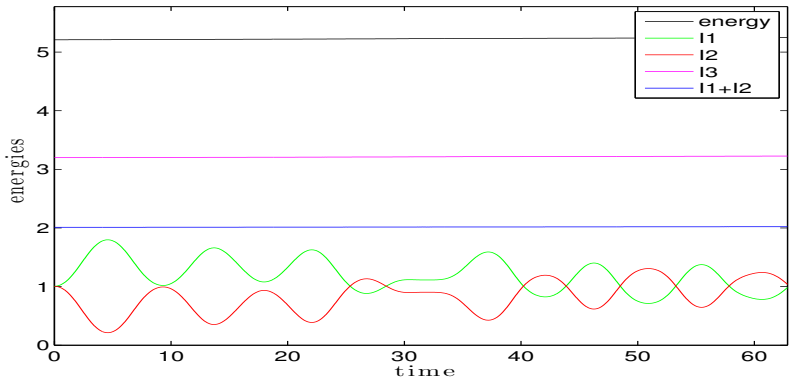

Figure 11: Energy exchanges for $\varepsilon = 0.0001$, computed with the two-scale method

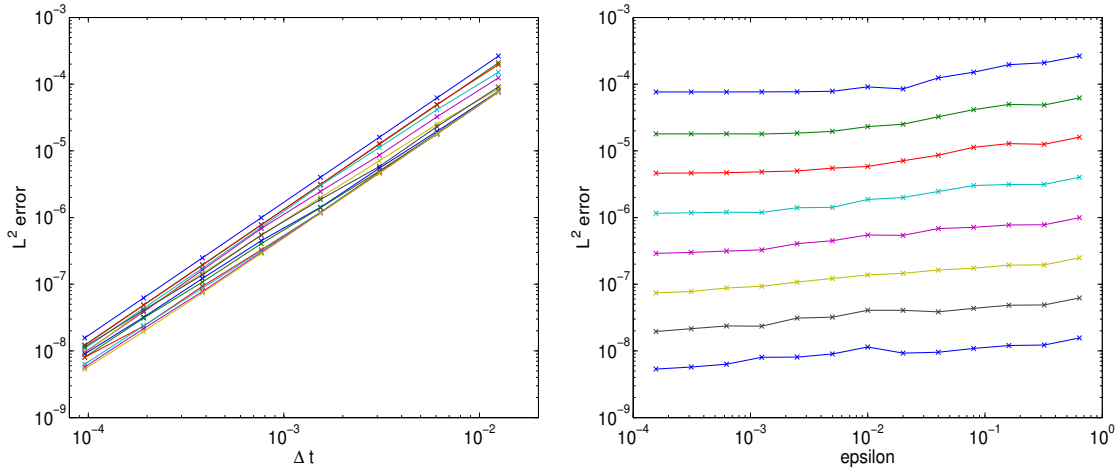

Figure 12: Left: L^2 error versus Δt for $\varepsilon = 2^N \times 10^{-2}$, with $N \in \{6, \ldots, 1, 0, -1, \ldots, -6\}$. Right: L^2 error versus ε for $\Delta t = 2\pi/2^N$ with $N \in \{9, \ldots, 16\}$

Appendix

In order to keep the paper as self-contained as possible, we recall in this section the main results of [8, 11] as used in the proof of the averaging results of Section 3. For the sake of simplicity, we assume here that the norm on $\mathbb{C}^n$ is the Euclidean norm in accordance with [11] and with Section 3.1.

Averaging of periodically forced problems

Consider a periodic highly-oscillatory differential equation of the form

$$\dot{v}^{[\varepsilon,\mu]}(t) = f^{\varepsilon}_{t/\mu}\Big(v^{[\varepsilon,\mu]}(t)\Big), \quad v^{[\varepsilon,\mu]}(0) = v_0 \in \mathbb{R}^n, \quad t \in [0,T], \tag{23}$$

where the function $(\tau, v) \mapsto f^{\varepsilon}_{\tau}(v)$ is assumed to be 2π-periodic in τ and where ε is a small parameter with values in the interval $\mathcal{J} :=]-\varepsilon_0, \varepsilon_0[$. We emphasize right away that **no**

regularity of the function f^ε in terms of ε is required, though all later **boundedness assumptions need to be uniform with respect to** ε. The main assumption of the averaging result derived in [8] requires the definition of the following $\mathbb{C}^n$-extension of the domain $\mathcal{K} \subset \mathbb{R}^n$ in which we wish to study the differential equation[4](23): for all $\rho \geq 0$,

$$\mathcal{K}_\rho = \{v + w \in \mathbb{C}^n; v \in \bar{\mathcal{K}}, \|w\| \leq \rho\}.$$

Assumption 4. *There exist $R > 0$ and an open set $\mathcal{U}$ containing $\mathcal{K}_R$, such that for any $\varepsilon \in \mathcal{J}$ and any $\tau \in \mathbb{T}$, $f^\varepsilon_\tau(\cdot)$ can be extended to a map from $\mathcal{U}$ to $\mathbb{C}^n$ that is analytic on $\mathcal{K}_R$. In addition, the Fourier coefficients $\hat{f}^\varepsilon_k$, $k \in \mathbb{Z}$, of f^ε, satisfy the uniform (in ε) bound*

$$\forall \varepsilon \in \mathcal{J}, \quad \sum_{k\in\mathbb{Z}} \|\hat{f}^\varepsilon_k\|_R \leq M$$

for some $M < +\infty$ independent of ε.

As already noticed in [8], this assumption does not imply that f is differentiable with respect to τ, only that f^ε is jointly continuous in $\mathbb{T}\times\mathcal{K}_R$ (and again not necessarily continuous w.r.t. ε), and that

$$\forall \varepsilon \in \mathcal{J}, \quad \forall \tau \in \mathbb{T}, \quad \|f^\varepsilon_\tau\|_R \leq M.$$

We are now in position to formulate Theorem 3.4 of [8].

Theorem 5.1. *Suppose that f^ε satisfies Assumption 4. Then for any $\varepsilon \in \mathcal{J}$ and for any $(\mu, N) \in \mathbb{C} \times \mathbb{N}^*$ such that $|\mu|N \leq c := \frac{R}{8M}$, there exists a near-identity change of variables $v = \Phi^{[\varepsilon,\mu,N]}_{t/\mu}(V)$ with $\Phi^{[\varepsilon,\mu,N]} : \mathbb{T} \times \mathcal{K}_{R/2} \to \mathcal{K}_R$ transforming equation (23) into the equation*

$$\dot{V} = F^{[\varepsilon,\mu,N]}(V) + R^{[\varepsilon,\mu,N]}_{t/\mu}(V), \quad V(0) = v_0,$$

with averaged vector field $F^{[\varepsilon,\mu,N]} : \mathcal{K}_{R/2} \to \mathbb{C}^n$ and remainder $R^{[\varepsilon,\mu,N]} : \mathbb{T} \times \mathcal{K}_{R/2} \to \mathbb{C}^n$ satisfying the following bounds

$$\|F^{[\varepsilon,\mu,N]} - \hat{f}^\varepsilon_0\|_{R/2} \leq \frac{M}{2}|\mu| \quad \text{and} \quad \forall \tau \in \mathbb{T}, \quad \|R^{[\varepsilon,\mu,N]}_\tau\|_{R/2} \leq \frac{5\left(\frac{|\mu|N}{c}\right)^N}{1 - \frac{|\mu|N}{c}}\, M.$$

Besides, if $|\mu| \leq c/e$ and $N = N^{[\mu]}$ is chosen as the integer part of $c/(e|\mu|) \geq 1$, then

$$\forall \tau \in \mathbb{T}, \quad \|R^{[\varepsilon,\mu,N^{[\mu]}]}_\tau\|_{R/2} \leq \frac{5e^2}{e-1} M \exp\Big(-\frac{c}{e|\mu|}\Big).$$

Conserved quantities in autonomous Hamiltonian problems

We consider here the more specific situation of an autonomous Hamiltonian problem

$$\dot{u}^{[\varepsilon,\mu]} = J^{-1}\nabla_u \mathcal{H}^{[\varepsilon,\mu]}(u^{[\varepsilon,\mu]}), \quad u^\varepsilon(0) = u_0 \in X, \tag{24}$$

[4]The domain $\mathcal{K}$ is usually defined as an open subset of $\mathbb{R}^n$ containing all solutions of (23) for all values of $t \in [0, T]$, all sufficiently small values of μ and all values of ε.

where $n = 2m$ is now assumed to be even, J is the matrix

$$J = \begin{pmatrix} 0 & \mathrm{Id} \\ -\mathrm{Id} & 0 \end{pmatrix}, \quad \mathrm{Id} \in \mathcal{M}(\mathbb{R}^m),$$

and

$$\mathcal{H}^{[\varepsilon,\mu]}(u) = \frac{1}{\mu} I^\varepsilon(u) + K^\varepsilon(u)$$

where the flow $\chi_t^{[\varepsilon]}$ of the Hamiltonian system

$$\dot{u} = J^{-1} \nabla_u I^\varepsilon(u)$$

is assumed to be 2π-periodic, *independently of* ε. Problem (24) can be reformulated by performing the change of variables $u = \chi^\varepsilon_{t/\mu}(v)$ so that v satisfies the differential equation

$$\dot{v}^{[\varepsilon,\mu]} = f^\varepsilon_{t/\mu}(v^{[\varepsilon,\mu]}) = J^{-1} \nabla_v K^\varepsilon_{t/\mu}(v^{[\varepsilon,\mu]})$$

where $K^\varepsilon_\tau = K^\varepsilon \circ \chi^\varepsilon_\tau$ for all $\tau \in \mathbb{T}$ and all $\varepsilon \in \mathcal{J}$.

Assumption 5. *There exist $R > 0$ and an open set $\mathcal{U}$ containing $\mathcal{K}_R$, such that for any $\varepsilon \in \mathcal{J}$ and any $\tau \in \mathbb{T}$, $K^\varepsilon_\tau(\cdot)$ can be extended to a map from $\mathcal{U}$ to $\mathbb{C}^n$ that is analytic on $\mathcal{K}_R$. In addition, the Fourier coefficients $\hat{H}^\varepsilon_k$, $k \in \mathbb{Z}$, of K^ε_τ, satisfy the following uniform (in ε) bound*

$$\forall \varepsilon \in \mathcal{J}, \quad \sum_{k \in \mathbb{Z}} \|\hat{H}^\varepsilon_k\|_R \leq M$$

for some $M < +\infty$ independent *of ε.*

Theorem 5.2. *Suppose that K^ε satisfies Assumption 5. Then for any $\varepsilon \in \mathcal{J}$ and for any $(\mu, N) \in \mathbb{C} \times \mathbb{N}^*$ such that $|\mu|(N+1) \leq \frac{1}{L} := \frac{R^2}{8eM}$, the vector field $F^{[\varepsilon,\mu,N]}$ of Theorem 5.1 is Hamiltonian with Hamiltonian $\tilde{K}^{[\varepsilon,\mu,N]}$ and there exists a modified invariant $\tilde{I}^{[\varepsilon,\mu,N]}$ such*

$$\mathcal{H}^{[\varepsilon,\mu]} = \frac{1}{\mu} \tilde{I}^{[\varepsilon,\mu,N]} + \tilde{K}^{[\varepsilon,\mu,N]}$$

where the three terms are "almost in involution" in the sense that

1. *For all $\varepsilon \in \mathcal{J}$ and all $u \in \mathcal{K}$,*

$$|\{\mathcal{H}^{[\varepsilon,\mu]}(u), \tilde{I}^{[\varepsilon,\mu,N]}(u)\}| \leq \left(\frac{R}{8e}\right)^2 \Big(L|\mu|(N+1)\Big)^{(N+1)}. \tag{25}$$

2. *Assume that $L|\mu| \leq 1/(2e)$ and choose $N = N^{[\mu]}$ as the integer part of $L^{-1}|\mu|^{-1}e^{-1} - 1$. Then for all $\varepsilon \in \mathcal{J}$ and all $u \in \mathcal{K}$,*

$$|\{\mathcal{H}^{[\varepsilon,\mu]}(u), \tilde{I}^{[\varepsilon,\mu,N^{[\mu]}]}(u)\}| \leq \frac{M}{8L} \exp\left(-\frac{1}{eL|\mu|}\right). \tag{26}$$

Acknowledgments

The authors have been supported by projects Lodiquas and Moonrise from the ANR (The French National Research Agency). Besides, they wish to thank Y. Bugeaud, X. Caruso, G. Hanrot and G. Wanner for very enlightening discussions on the arithmetic parts of this paper.

References

[1] M. Avellaneda, Th. Y. Hou, and G. Papanicolaou, *Finite difference approximations for partial differential equations with rapidly oscillating coefficients*, RAIRO Modél. Math. Anal. Numér., 25, pp. 693-710,1991.

[2] W. Bao, *Ground states and dynamics of multi-component Bose-Einstein condensates*, Multiscale Modeling and Simulation: a SIAM Interdisciplinary Journal, Vol. 2, pp. 210-236, 2004.

[3] Y. Bugeaud, *Approximation by algebraic numbers*, Cambridge Tracts in Mathematics Vol. 160, Cambridge University Press, 2004.

[4] M.P. Calvo, P. Chartier, A. Murua and J.M. Sanz-Serna, *A stroboscopic numerical method for highly oscillatory problems*, in Numerical Analysis and Multiscale Computations, B. Engquist, O. Runborg and R. Tsai, editors, Lect. Notes Comput. Sci. Eng., Vol. 82, Springer 2011, 73-87.

[5] P. Chartier, N. Crouseilles, M. Lemou and F. Méhats, *Uniformly accurate numerical schemes for highly-oscillatory Klein-Gordon and nonlinear Schrödinger equation*, Numer. Math., Vol. 129, Issue 2, pp 211-250, February 2015.

[6] P. Chartier, J. Makazaga, A. Murua, and G. Vilmart, *Multirevolution composition methods for highly-oscillatory differential equations*, Numer. Math., Vol. 128, No. 1, pp 167-192, 2014.

[7] M.P. Calvo, P. Chartier, A. Murua, and J.-M. Sanz-Serna, *Numerical stroboscopic averaging for ODEs and DAEs*, Appl. Numer. Math., Vol. 61, No. 10, pp. 1077-1095, 2011.

[8] P. Chartier, A. Murua, J.M. Sanz-Serna, *A formal series approach to averaging: exponentially small error estimates*, Discrete and Continuous Dynamical Systems (DCDS-A), Vol. 32, no. 9, 2012.

[9] P. Chartier, A. Murua and J.M. Sanz-Serna, *Higher-order averaging, formal series and numerical integration I: B-series*, FOCM, Vol. 10, No. 6, 2010.

[10] P. Chartier, A. Murua and J.M. Sanz-Serna, *Higher-order averaging, formal series and numerical integration II: the quasi-periodic case*, FOCM, 2012.

[11] P. Chartier, A. Murua and J.M. Sanz-Serna, *Higher-order averaging, formal series and numerical integration III: error bounds*, FOCM, 2013.

[12] E. Hairer, C. Lubich and G. Wanner, *Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations*, Springer Series in Computational Mathematics 31, 2006.

[13] D.S. Hall, M.R. Matthews, J.R. Ensher, C.E. Wieman, and E.A. Cornell, *Dynamics of component separation in a binary mixture of Bose-Einstein condensates*, Phys. Rev. Lett., Vol. 81, pp. 1539?1542, 1998.

[14] R. Dáger and E. Zuazua, *Controllability of star-shaped networks of strings*, C. R. Acad. Sci. Paris, Vol. 332, No. 7, pp 621-626, 2001.

[15] G.H. Hardy and E.M. Wright, *An introduction to the theory of numbers*, edited and revised by D. R. Heath-Brown and J. H. Silverman. With a foreword by Andrew Wiles. 6th ed., Oxford University Press, 2008.

[16] B. Grébert and C. Villegas-Blas, *On the energy exchange between resonant modes in nonlinear Schrödinger equations*, Ann. I. H. Poincaré, Vol. 28, 2011.

[17] J. C. Lagarias, *Best simultaneous diophantine approximations I, Growth Rates of Best Approximations denominators*, Trans. Amer. Math. Soc., vol. 72 No. 2, 1982.

[18] J. C. Lagarias, *Best simultaneous diophantine approximations II, Behavior of consecutive best approximations*, Pacific Journal of Mathematics, Vol. 102, No. 1, 1982.

[19] R. Orive and E. Zuazua, *Finite Difference Approximation of Homogenization Problems for Elliptic Equations*, Multiscale Modeling & Simulation, Vol. 4, No. 1, pp. 36-87, 2005.

[20] J.A. Sanders, JF. Verhulst and J. Murdock, *Averaging methods in nonlinear dynamical systems*, Applied Mathematical Sciences 59, Springer, 2007.

[21] C. Simo, *Averaging under fast quasi-periodic forcing*, in Hamiltonian Mechanics, Integrability and Chaotic Behavior, Edited by J. Seimenis, NATO Asi Series, Vol. 331.