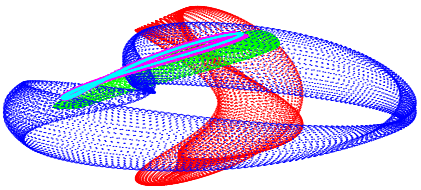
Contribution to the Study of Fourier Methods for Quasi–Periodic Functions and the Vicinity of the Collinear Libration Points

José María Mondelo

Departament de Matemàtica Aplicada i Anàlisi
Universitat de Barcelona



Programa de doctorat de Matemàtica Aplicada i Anàlisi.
Bienni 1996–98.

Memòria presentada per a aspirar al grau de
Doctor en Matemàtiques per la Universitat
de Barcelona

Certifico que la present memòria ha estat realitzada per José María Mondelo González i dirigida per mi.

Barcelona, 23 de maig de 2001

Gerard Gómez i Muntané

A mi familia.

Acknowledgements

I would like to thank my thesis advisor, G. Gómez, for the many hours we have spent together, and for his support and advice, which has never been just academic. I am also indebted with C. Simó, who has always been aware of this work and has given many hints and ideas that have been critical for its developement. I would like to thank also À. Jorba, for several suggestions and for introducing me in the world of parallel computing. J. Masdemont has provided me software to acces the JPL ephemeris and to evaluate the c_i functions of Appendix A. I would like to mention also J. Font and J. Timoneda, who have always provided me one of the most powerful computing enviornments of the Department. They have also shared with me their expertise in hardware and system management, which has allowed me to get my current job. All my colleages of the Department of Applied Mathematics and Analysis of the University of Barcelona have contributed to provide a very friendly environment to work in. I would like to especially mention A. González, F. Naselli and J. Puig, for their friendship and support during my last year there. And finally, I would like to thank my family. Whithout their support, this work would never have been done.

Preface

This work has been organized in three parts. The first two ones contain the main results, and the last one, which has been divided in several appendices, has complementary results.

The first part of the work (Chapters 1 to 5) is dedicated to the development and study of a procedure for the accurate computation of frequencies, as well as the related Fourier coefficients, of a quasi–periodic function, using as only input input a equally–spaced sampling of the function to be analyzed over a finite time interval.

The first technique for the accurate determination of frequencies has been introduced by J. Laskar ([18], [20], [19]). It is based on the maximization of the formula that gives the Fourier coefficients of a function with respect to the harmonic index, but taking it as a real number. This procedure has been applied to the study of the long–term dynamics of the Solar System ([18]), as well as to the study of chemistry and particle accelerator models through the computation of *frequency maps* ([19]). Some methodology for frequency determination has also been introduced in [12],[13],[10],[11]. In these works, the determination of frequencies has been applied to development of semi–analytical models for the motion in the Solar System.

Our procedure takes the methodology developed in [12],[13],[10][11] as a starting point. It is based in asking for equality between the Discrete Fourier Transform (DFT) of the analyzed function and its quasi–periodic approximation. Error estimates are obtained and illustrated with numerical examples. Also, in the line of the previously–mentioned works, we apply our procedure to the development of simplified models for the motion in the Solar System.

The second part of the work (Chapters 6 to 7) is devoted to the study to the dynamics in the vicinity of the collinear equilibrium points of the three–dimensional Restricted Three–Body Problem (RTBP) for the Earth–Moon mass parameter.

The first systematic study of this vicinity has been done in [10] and [16], using as a tool the reduction to the central manifold of the collinear equilibrium points. This is a semi–analytical technique, which limits the region that can be explored by the convergence of the expansions computed. The same methodology has also been applied to the study of the collinear equilibrium points of a model for the Earth–Moon system, called the Quasi–Bicircular Problem ([3]). In this last study, the convergence constraints are still more severe.

In this work, we follow the families of periodic orbits and invariant 2D tori of the center manifolds of the three collinear libration points using purely numerical procedures. With this approach, we can extend the analysis of the phase space done in [10] and [16] to a wider range of energy values, that now include several bifurcations, and also to the L_3 libration point. The methodology used for the continuation of invariant tori is based in [7],

with some modifications in order to account for variable excitations and some additional parameters needed for our exploration. We have followed parallel strategies in order to cope with the large amount of computations required. They have been carried out on HIDRA, one of the Beowulf clusters of the Barcelona Dynamical Systems Group.

The third and last part of this report consists in several appendices, which give some additional results that have been taken apart from the main text in order to improve its readability.

Contents

Part I

Numerical Fourier analysis of quasi–periodic functions

This part is devoted to the development and study of a procedure to compute the frequencies and the related amplitudes of a quasi–periodic function. In Chapter 1 we introduce some notation and methodology related to the Discrete Fourier Transform (DFT), which is the main tool in which our procedure is based. In Chapter 2 we describe the procedure, as well as some aspects of its computer implementation. Chapter 3 is devoted to the obtention of error estimates, which are collected in Theorem 3.4.1 and illustrated with a numerical example in Chapter 4. In Chapter 5, we apply the methodology to the development of simplified models for the motion in the Solar System.

Chapter 1

The Discrete Fourier Transform (DFT)

This chapter gives the basic notation and definitions needed to develop our Fourier analysis procedure in Chapter 2. It is started with the introduction of the DFT and some related notation. After that, we discuss the concept known in the literature as *leakage*, which leads to the introduction of filtering through the use of Hanning functions. We end the Chapter with some comments about how the concept known as *aliasing* arises in our setting.

1.1 Preliminaries and notation

Let D be the space of real valued functions defined on a discrete set of N equally spaced points $t_0, ..., t_{N-1}$ over the interval $[0, T]$, i.e. $t_l = l \cdot \Delta t$, with $\Delta t = T/N$. The equality of the spacing is only a technical requirement, since the DFT could be adapted to a non equally spaced set of samples. Nevertheless, all what follows has been written assuming Δt constant. From now on N is assumed to be even. If f, $g \in D$, we define their discrete scalar product as

$$\langle f, g \rangle = \sum_{l=0}^{N-1} f(t_l) g(t_l).$$

The set of functions $\{\{\varphi_j\}_{j=0}^{N/2}, \{\psi_j\}_{j=1}^{N/2-1}\} \subset D$, being

$$\varphi_j(t) = \cos(\frac{2\pi jt}{T}), \quad \psi_j(t) = \sin(\frac{2\pi jt}{T}),$$

form an orthogonal basis of D. Therefore, every function $f \in D$ can be written as

$$f(t_l) = P_{f,T,N}(t_l) \qquad l = 0, \ldots, N-1,$$

where

$$\begin{aligned} P_{f,T,N}(t) &= \frac{1}{2}\Big(c_{f,T,N}(0) + c_{f,T,N}(\frac{N}{2})\cos(\frac{2\pi\frac{N}{2}t}{T})\Big) \\ &\quad + \sum_{j=1}^{N/2-1}\Big(c_{f,T,N}(j)\cos(\frac{2\pi jt}{T}) + s_{f,T,N}(j)\sin(\frac{2\pi jt}{T})\Big). \end{aligned} \tag{1.1}$$

with

$$c_{f,T,N}(j) = \delta_j \frac{\langle f, \varphi_j \rangle}{\langle \varphi_j, \varphi_j \rangle}, \quad j = 0, \ldots, \frac{N}{2}, \qquad s_{f,T,N}(j) = \frac{\langle f, \psi_j \rangle}{\langle \psi_j, \psi_j \rangle}, \quad j = 1, \ldots, \frac{N}{2} - 1,$$

with

$$\delta_j = \begin{cases} 2, & j = 0, \\ 1, & j = 1, ..., N/2 - 1, \\ 2, & j = N/2, \end{cases}$$

Equation (1.1) defines the *Discrete Fourier Transform in sines and cosines* (DFT) of $f \in D$ as a function of the discrete set of frequencies j/T, $j = 0, \ldots, N/2$. The values $c_{f,T,N}(j)$ and $s_{f,T,N}(j)$ are the coefficients related to the j/T frequency of the trigonometric interpolating polynomial $P_{f,T,N}(t)$ of the function f at the nodes $\{t_l\}_{l=0}^{N-1}$. All the frequencies of $P_{f,T,N}(t)$ are multiples of $1/T$. The DFT coefficients can be explicitly written as

$$\begin{aligned} c_{f,T,N}(j) &= \frac{2}{N} \sum_{l=0}^{N-1} f(t_l) \cos(2\pi \frac{j}{N} l), \quad j = 0, ..., N/2, \\ s_{f,T,N}(j) &= \frac{2}{N} \sum_{l=0}^{N-1} f(t_l) \sin(2\pi \frac{j}{N} l), \quad j = 1, ..., N/2 - 1. \end{aligned}$$

For a general complex–valued function f, using the discrete scalar product $\langle f, g \rangle = \sum_{l=0}^{N-1} f(t_l)\overline{g(t_l)}$ and the orthogonal basis $\{e^{2\pi i \frac{j}{T} t}\}_{j=0}^{N-1}$, we can define the DFT as

$$F_{f,T,N}(j) = \frac{1}{N} \sum_{l=0}^{N-1} f(t_l) e^{-2\pi i \frac{j}{T} t_l}, \quad j = 0, ..., N-1. \tag{1.2}$$

If f takes real values, $F_{N-j}(f) = \overline{F_j(f)}$, the $F_{f,T,N}(j)$, $c_{f,T,N}(j)$, $s_{f,T,N}(j)$ coefficients are related by

$$F_{f,T,N}(j) = \frac{1}{2}(c_{f,T,N}(j) - i s_{f,T,N}(j)) \quad j = 0, \ldots, \frac{N}{2},$$

where we assume $s_{f,T,N}(0) = s_{f,T,N}(N/2) = 0$. This allows to compute efficiently the $c_{f,T,N}(j)$, $s_{f,T,N}(j)$ coefficients using a standard Fast Fourier Transform (FFT) algorithm (see, for instance, [6], [5] or [21]).

The complex–valued function

$$\phi_{f,T}(\alpha) = \frac{1}{T} \int_0^T f(t) e^{-i2\pi\alpha t} dt$$

will be called *Truncated Continuous Fourier Transform* (TCFT) of f. Note that $\phi_{f,T,N}(\frac{j}{T})$, $j \in \mathbb{Z}$, are the coefficients of the Fourier series of f on the interval $[0, T]$. Note also that the DFT can be seen as a Riemann sum of the TCFT, more concretely

$$\phi_{f,T}\Big(\frac{k}{T}\Big) = \frac{1}{T} \int_0^T f(t) e^{-2\pi i \frac{k}{T} t} dt \approx \frac{1}{T} \sum_{l=0}^{N-1} f(t_l) e^{-2\pi i \frac{k}{T} t_l} \frac{T}{N} = F_{f,T,N}(k). \tag{1.3}$$

Consequently, we can obtain the TCFT as the limit when $N \to \infty$ of the DFT. In section 3, Lemma 3.2.4, we will give an explicit bound of the difference between the DFT and the TCFT of a complex exponential term $e^{i2\pi\omega t}$.

1.2 Leakage effect and filtering

For periodic functions, when the length T of the time interval spanned by the samples is not an integer multiple of the period of the function (or equivalently, when the frequency of the function is not an integer multiple of the "basic frequency" $1/T$ associated to the sample interval $[0, T]$), there appear in the DFT spurious frequencies, that is, the DFT is different from zero at frequencies not being multiple of the frequency of the function. This is a phenomenon known as *leakage*, for which we give a graphical example in figure 1.1. Leakage also affects the TCFT.

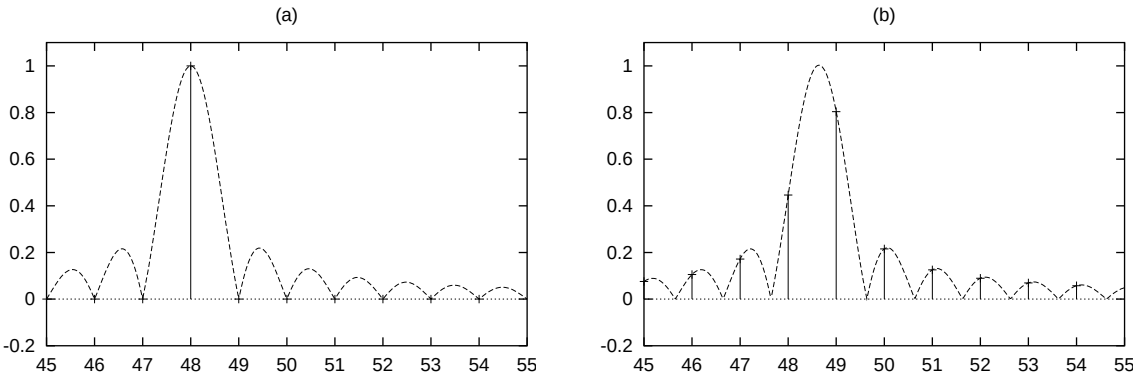


Figure 1.1: Plot of $[(c_{f,T,N}(j))^2 + (s_{f,T,N}(j))^2]^{1/2}$ as a function of j (dashed line) and for $j = 45, \ldots, 55$ (solid vertical lines), with $f(t) = \cos(2\pi\omega t)$, $T = 64$ and $N = 256$. In (a) $\omega = 0.75$, so $48 \cdot (0.75)^{-1} = 64 = T$ and there is no leakage. This is not the case for (b), where $\omega = 0.76$.

For the procedures that will be described later, we are interested in reducing leakage for functions of the form $e^{2\pi i\omega t}$. The way to do this is to use a filter or window function.

Definition 1.2.1 *$H(t)$ is said to be a* filter function of degree $r \geq 0$ for the interval $[0, T]$ *if it is a positive function of class C^{r-1} with $H^{(j)}(0) = H^{(j)}(T) = 0$ for $j = 0, \ldots, r-1$, such that $H^{(r)}$ is continuous except for a finite set of jump discontinuities, and has bounded variation. We also assume that*

$$\frac{1}{T}\int_0^T H(t)dt = 1. \tag{1.4}$$

It is enough to focus on the TCFT, since it is the limit of the DFT when $N \to \infty$ (equation (1.3)). The reduction of leakage for the TCFT of $H(t)e^{2\pi i\omega t}$, which depends directly on the regularity of the filter function, is given by corollary 1.2.1.

Proposition 1.2.1 If g is a filter function of degree r for the interval $[0, 2\pi]$, then

$$\Big|\int_0^{2\pi} g(t)e^{i\alpha t}dt\Big| \leq \frac{2B(g^{(r)}) + V(g^{(r)})}{|\alpha|^{r+1}},$$

where $B(g^{(r)})$ is a bound for $g^{(r)}$ in $[0, 2\pi]$ and $V(g^{(r)})$ is the variation of $g^{(r)}$ in the same interval. Moreover, if we assume $g^{(r)}(0) = g^{(r)}(2\pi) = 0$, then

$$\Big|\int_0^{2\pi} g(t)e^{i\alpha t}dt\Big| \leq \frac{V(g^{(r)})}{|\alpha|^{r+1}}.$$

Proof: Let $t_1, \dots t_{m-1}$ be the jump discontinuities of $g^{(r)}$ in $(0, 2\pi)$, and let $t_0 = 0$, $t_m = 2\pi$ (which may be jump discontinuities or not). Successive integrations by parts yield

$$\begin{aligned}
\int_0^{2\pi} g(t)e^{i\alpha t}dt &= \Big[g(t)\frac{e^{i\alpha t}}{i\alpha}\Big]_{t=0}^{t=2\pi} - \frac{1}{i\alpha}\int_0^{2\pi} g'(t)e^{i\alpha t}dt \\
&= \dots = \Big(\frac{-1}{i\alpha}\Big)^{r-1}\sum_{j=0}^{m-1}\int_{t_j}^{t_{j+1}} g^{(r-1)}(t)e^{i\alpha t}dt \\
&= \Big(\frac{-1}{i\alpha}\Big)^{r-1}\Big(\sum_{j=0}^{m-1}\Big[g^{(r-1)}(t)\frac{e^{i\alpha t}}{i\alpha}\Big]_{t=t_j}^{t=t_{j+1}} + \frac{-1}{i\alpha}\sum_{j=0}^{m-1}\int_{t_j}^{t_{j+1}} g^{(r)}(t)e^{i\alpha t}dt\Big).
\end{aligned}$$

The first sum vanishes because it is a telescopic sum and $g^{(r-1)}(0) = g^{(r-1)}(2\pi) = 0$.

Given $\varepsilon > 0$, let δ_j be such that, for $t^*, t^{**} \in [t_j, t_{j+1}]$, if $|t^* - t^{**}| < \delta_j$ then $|g^{(r)}(t^*) - g^{(r)}(t^{**})| < \varepsilon$ $(j = 0 \div m-1)$. These δ_j exist because $g^{(r)}$ is uniformly continuous in each interval $[t_j, t_{j+1}]$. Reducing δ_j if necessary, we can assume $t_{j+1} - t_j = n_j\delta_j$ for $n_j \in \mathbb{N}$. Define $M = n_0 + \dots + n_{m-1}$ and

$$\begin{aligned}
&s_0 = t_0,\ s_1 = t_0 + \delta_0,\ \dots,\ s_{n_0-1} = t_0 + (n_0-1)\delta_0,\ s_{n_0} = t_0 + n_0\delta_0 = t_1, \\
&s_{n_0+1} = t_1 + \delta_1,\ \dots,\ s_{n_0+\dots+n_{m-1}-1} = t_{m-1} + (n_{m-1}-1)\delta_{m-1},\ s_M = t_m.
\end{aligned}$$

Denoting by $\xi_j = \frac{1}{2}(s_j + s_{j+1})$,

$$\begin{aligned}
\int_0^{2\pi} g(t)e^{i\alpha t}dt &= \Big(\frac{-1}{i\alpha}\Big)^r\sum_{j=0}^{M-1}\int_{s_j}^{s_{j+1}}(g^{(r)}(\xi_j) + g^{(r)}(t) - g^{(r)}(\xi_j))e^{i\alpha t}dt \\
&= \Big(\frac{-1}{i\alpha}\Big)^r\sum_{j=0}^{M-1}\Big(g^{(r)}(\xi_j)\frac{e^{i\alpha s_{j+1}} - e^{i\alpha s_j}}{i\alpha} + \int_{s_j}^{s_{j+1}}(g^{(r)}(t) - g^{(r)}(\xi_j))e^{i\alpha t}dt\Big) \\
&= \Big(\frac{-1}{i\alpha}\Big)^r\Big[\frac{1}{i\alpha}\Big(-g^{(r)}(\xi_0)e^{i\alpha s_0} + g^{(r)}(\xi_{M-1})e^{i\alpha s_M} \\
&\qquad + \sum_{j=1}^{M-1}(g^{(r)}(\xi_{j-1}) - g^{(r)}(\xi_j))e^{i\alpha s_j}\Big) \\
&\qquad + \sum_{j=0}^{M-1}\int_{s_j}^{s_{j+1}}(g^{(r)}(t) - g^{(r)}(\xi_j))e^{i\alpha t}dt\Big],
\end{aligned}$$

and hence

$$\Big|\int_0^{2\pi} g(t)e^{i\alpha t}dt\Big| \leq \frac{1}{|\alpha|^r}\Big[\frac{1}{|\alpha|}(2B(g^{(r)}) + V(g^{(r)})) + 2\pi\varepsilon\Big].$$

Since $\varepsilon > 0$ is arbitrary, doing $\varepsilon \to 0$ we get the first equality of the proposition. The second bound follows immediately from the above computations. □

For $\alpha = j/T$, the above Proposition is a standard result about the decaying of the Fourier coefficients (see, e.g., [5] and references therein). The above proof covers the case $\alpha \neq j/T$.

Corollary 1.2.1 If H is a filter function of degree r for $[0, T]$, then

$$\phi_{H(t)e^{2\pi i\nu t/T},T}(\alpha/T) = O\Big(\frac{1}{|\nu-\alpha|^{r+1}}\Big).$$

Proof: We have

$$\phi_{H(t)e^{2\pi i\nu t/T},T}(\alpha/T) = \frac{1}{T}\int_0^T H(t)e^{2\pi i(\nu-\alpha)t/T}dt = \frac{1}{2\pi}\int_0^{2\pi} H(\frac{T}{2\pi}s)e^{i(\nu-\alpha)s}ds,$$

and hence

$$|\phi_{H(t)e^{2\pi i\nu t/T},T}(\alpha/T)| \leq \frac{2B(H^{(r)}) + V(H^{(r)})}{2\pi|\nu-\alpha|^{r+1}}.$$

□

We will use as a filter function the *Hanning window function*, which is defined as

$$H_T(t) = 1 - \cos\left(2\pi\frac{1}{T}t\right).$$

and has degree 2. To increase the degree of the filter, the Hanning function can be iterated and we can consider *Hanning functions of order* $n_h \in \mathbb{N}$, defined as

$$H_T^{n_h}(t) = q_{n_h}\left(1 - \cos\left(2\pi\frac{1}{T}t\right)\right)^{n_h},$$

where the constants q_{n_h} are computed in order to fulfill (1.4), so

$$q_{n_h} = \left[\frac{1}{T}\int_0^T \left(1 - \cos\left(2\pi\frac{1}{T}t\right)\right)^{n_h} dt\right]^{-1} = \frac{n_h!}{(2n_h-1)!!}.$$

The advantage of the Hanning function with respect to other well-known window functions (see [21]) is its degree of differentiability. For instance, $H^{n_h}(t)$ has degree $2n_h$, whereas a general "triangle window function" $T^{n_t}(t)$ defined as

$$T^{n_t}(t) = \frac{n_t+1}{n_t}\Big(1 - \Big|\frac{2}{T}t - 1\Big|^{n_t}\Big).$$

has degree just n_t. The *Parzen window* and the *Welch window* are the particular cases $n_t = 1$ and $n_t = 2$ of $T^{n_t}(t)$ respectively. The Hanning function has its simplicity as an

additional advantage. The properties of trigonometric functions allow to obtain relations like Lemma 1.2.1.

The DFT coefficients with Hanning order n_h of $f(t)$ are defined as the DFT coefficients of $H_T^{n_h}(t)f(t)$, and will be denoted by

$$\begin{aligned} F^{n_h}_{f,T,N}(j) &= F_{H^{n_h}f,T,N}(j), \\ c^{n_h}_{f,T,N}(j) &= c_{H^{n_h}f,T,N}(j), \\ s^{n_h}_{f,T,N}(j) &= s_{H^{n_h}f,T,N}(j). \end{aligned}$$

Analogously, for the TCFT we will have

$$\phi^{n_h}_{f,T,N}(\alpha) = \frac{1}{T}\int_0^T H^{n_h}(t)f(t)e^{-2\pi i\alpha t}dt.$$

The following lemma relates the coefficients of filtered and non–filtered transforms of a function $f(t)$:

Lemma 1.2.1 The following relations hold:

(a) $$F^{n_h}_{f,T,N}(j) = \frac{q_{n_h}}{2^{n_h}}\sum_{l=-n_h}^{n_h}(-1)^l\binom{2n_h}{n_h+l}F_{f,T,N}(j+l) = \sum_{l=-n_h}^{n_h}\frac{(-1)^l(n_h!)^2F_{f,T,N}(j+l)}{(n_h+l)!(n_h-l)!}.$$

(b) $$\phi^{n_h}_{f,T,N}(\alpha) = \frac{q_{n_h}}{2^{n_h}}\sum_{l=-n_h}^{n_h}(-1)^l\binom{2n_h}{n_h+l}\phi_{f,T,N}(\alpha+\frac{l}{T}) = \sum_{l=-n_h}^{n_h}\frac{(-1)^l(n_h!)^2\phi_{f,T,N}(\alpha+l/T)}{(n_h+l)!(n_h-l)!}.$$

Proof: We only prove (a). Similar calculations are valid for (b). Using that $1-\cos x = 2\sin^2\frac{x}{2}$, we have

$$\begin{aligned} F^{n_h}_{f,T,N}(j) &= \sum_{l=0}^{N-1} q_{n_h}\Big(1-\cos(2\pi\frac{1}{T}t_l)\Big)^{n_h}f(t_l)e^{-2\pi i\frac{j}{T}t_l} \\ &= \sum_{l=0}^{N-1} q_{n_h}2^{n_h}\sin^{2n_h}(\pi\frac{1}{T}t_l)f(t_l)e^{-2\pi i\frac{j}{T}t_l} \\ &= \sum_{l=0}^{N-1} q_{n_h}2^{n_h}\frac{(e^{\pi i\frac{1}{T}t_l}-e^{-\pi i\frac{1}{T}t_l})^{2n_h}}{(2i)^{2n_h}}f(t_l)e^{-2\pi i\frac{j}{T}t_l} \\ &= \sum_{l=0}^{N-1}\frac{q_{n_h}}{(-2)^{n_h}}\sum_{l=0}^{2n_h}\binom{2n_h}{l}e^{\pi i\frac{2n_h-l}{T}t_l}(-1)^l e^{-\pi i\frac{l}{T}t_l}f(t_l)e^{-2\pi i\frac{j}{T}t_l} \\ &= \frac{q_{n_h}}{2^{n_h}}\sum_{l=0}^{2n_h}(-1)^{l-n_h}\binom{2n_h}{l}\sum_{l=0}^{N-1}f(t_l)e^{-2\pi i\frac{j+(l-n_h)}{T}t_l}. \end{aligned}$$

Shifting the index by n_h units, and using that

$$\frac{q_{n_h}}{2^{n_h}}\binom{2n_h}{n_h+l} = \frac{(n_h!)(n_h!)(2n_h)!}{(n_h!)2^{n_h}(2n_h-1)!!(n_h+l)!(n_h-l)!} = \frac{(n_h!)^2(2n_h)!}{(2n_h)!(n_h+l)!(n_h-l)!},$$

we get (a). □

For $c^{n_h}_{f,T,N}(j)$ and $s^{n_h}_{f,T,N}(j)$ relations similar to those of Lemma 1.2.1(a) hold. For instance, for $n_h = 1, 2$ we get

$$\begin{aligned} F^1_{f,T,N}(j) &= -\frac{1}{2}F_{f,T,N}(j-1)+F_{f,T,N}(j)-\frac{1}{2}F_{f,T,N}(j+1), \\ F^2_{f,T,N}(j) &= \frac{1}{6}F_{f,T,N}(j-2)-\frac{2}{3}F_{f,T,N}(j-1)+F_{f,T,N}(j)-\frac{2}{3}F_{f,T,N}(j+1)+\frac{1}{6}F_{f,T,N}(j+2). \end{aligned}$$

As H^{n_h} is a filter function of degree $2n_h$, according to Corollary 1.2.1 we have

$$\phi^{n_h}_{e^{2\pi i\nu t/T},T}(\alpha/T) = O\Big(\frac{1}{(|\nu-\alpha|)^{2n_h+1}}\Big).$$

In fact, it can be explicitly calculated that

$$\phi^{n_h}_{e^{2\pi i\nu t/T},T}(\alpha/T) = \frac{(-1)^{n_h}(n_h!)^2(e^{2\pi i(\nu-\alpha)}-1)}{2\pi i\psi_{n_h}(\nu-\alpha)} \tag{1.5}$$

being

$$\psi_{n_h}(x) = \prod_{l=-n_h}^{n_h}(x+l). \tag{1.6}$$

1.3 Aliasing effect

Apart from leakage, another common effect when performing DFT is *aliasing*. It consists in the fact that any frequency greater than half the frequency associated to the sampling width, i.e. any frequency greater than $\omega_c := \frac{N}{2T}$, is *aliased* in a frequency less than ω_c. This is due to the following fact. Denoting $p^{n_h}_{f,T,N}(j) = ((c^{n_h}_{f,T,N}(j))^2 + (s_{f,T,N})^2(j))^{1/2}$, we have that

$$\begin{aligned} p^{n_h}_{\mathrm{cs}(2\pi(\omega+\frac{N}{T})t,T,N}(j) &= p^{n_h}_{\mathrm{cs}(2\pi\omega t),T,N}(j), \\ p^{n_h}_{\mathrm{cs}(2\pi(-\omega)t),T,N}(j) &= p^{n_h}_{\mathrm{cs}(2\pi\omega t)}(j), \end{aligned}$$

where cs stands for any of the functions cos or sin. If $\omega > \frac{N}{2T}$ and $m \in \mathbb{Z}$ is such that $\widetilde{\omega} := \omega - m\frac{N}{T} \in [-\frac{N}{2T}, \frac{N}{2T}]$, then the frequencies ω and $|\widetilde{\omega}|$ are undistinguisheable from a DFT point of view when using the p function. The frequency ω_c is called the *Nyquist critical frequency* in the literature.

Indeed, when all the frequency components are confined to the interval $[-\omega_c, \omega_c]$, the function is called *band-width limited*. More concretely, in terms of the continuous Fourier Transform, this means

$$\int_{-\infty}^{\infty} f(t)e^{-2\pi i\omega t}dt = 0, \qquad \text{for} \qquad |\omega| > \omega_c.$$

In this case, assuming that we have a sampling of infinite size $\{j \cdot \Delta t\}_{j\in\mathbb{Z}}$, the *Shannon sampling theorem* (see [5], [21]) allows to reconstruct the function $f(t)$ from its samples, namely

$$f(t) = \Delta t \sum_{j=-\infty}^{\infty} f(t_j) \frac{\sin(2\pi\omega_c(t-t_j))}{\pi(t-t_j)},$$

where $\omega_c = \frac{1}{2\Delta t}$. Note that quasi–periodic functions are not band–width limited.

Chapter 2

Procedures for the refined Fourier analysis

This chapter is devoted to the description of our Fourier analysis procedure. It starts with the detailed description of the three steps in which it is carried out, namely: first approximation of frequencies, computation of the related amplitudes and iterative improvement of both frequencies and amplitudes. After that, we discuss some aspects regarding to its practical implementation: the algorithm to follow, the use of trigonometric recurrences and the accurate evaluation of the DFT of sines and cosines.

2.1 Introduction

Given N values, $\{f(t_l)\}_{l=0}^{N-1}$, $t_l \in [0,T]$ of a certain function $f(t)$, which is assumed to be quasi–periodic, we want to find a polynomical trigonometric approximation with a fixed number of frequencies N_f,

$$Q_f(t) = A_0^c + \sum_{l=1}^{N_f} (A_l^c \cos(\frac{2\pi\nu_l t}{T}) + A_l^s \sin(\frac{2\pi\nu_l t}{T})). \tag{2.1}$$

A standard approach to detect the frequencies of a given signal is to look for "peaks" of the modulus of the DFT, $p_{f,T,N}^{n_h}(j)$, which is also known as *power spectral density* in the literature. J. Laskar ([18], [20], [19]) introduced a refinement of this procedure, which consists in looking for maxima of $|\phi_{f,T,N}^{n_h}(j/T)|$, assuming that j takes real values. Additional methodology for frequency determination has been introduced in [13],[11]. It is the starting point for the methodology that will be developed here.

Our procedure is based entirely on the DFT for reasons that will be given below. The basic idea is to ask for the equality between the DFT of the sampled initial function and the DFT of its quasi–periodic approximation. It has three main steps: to get first approximations of the frequencies (either following the standard approach or using the method of Laskar), to compute the related approximated amplitudes and, finally, to perform a simultaneous improvement of frequencies and amplitudes.

2.2 First approximation of frequencies

If f has only one complex exponential term $f(t) = ae^{2\pi i \frac{\nu}{T} t}$, it follows from (1.5) that the modulus of its TCFT is

$$|\phi_{f,T}^{n_h}(\alpha/T)| = \frac{(n_h!)^2 |a| |1 - e^{i2\pi(\nu-\alpha)}|}{2\pi |\psi_{n_h}(\nu - \alpha)|}.$$

This function has a maximum at $\alpha = \nu$ (see Fig. 2.1). So, the problem of finding ν can be reduced to maximize the previous function with respect to α.

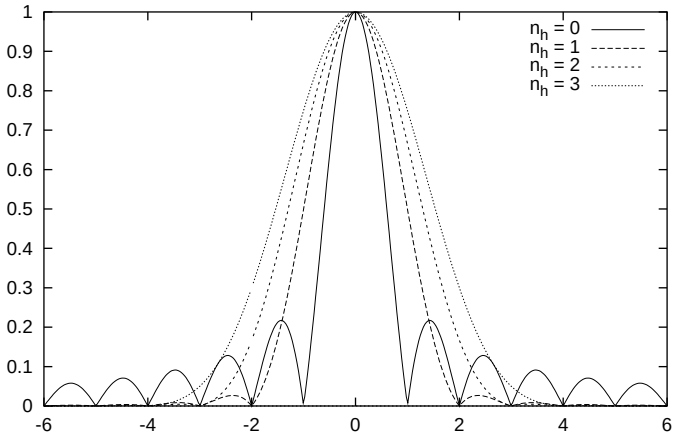


Figure 2.1: Plot of $|\phi_{e^{2\pi i\nu t/T},T}^{n_h}(\alpha/T)|$ as a function of $(\nu - \alpha)$ for $n_h = 0, 1, 2, 3$.

In a more general case, if $f(t)$ has m different frequencies,

$$f(t) = \sum_{l=1}^{m} a_l e^{2\pi i \frac{\nu_l}{T} t},$$

then $|\phi_{f,T}^{n_h}(\alpha/T)|$ does not have its maxima exactly at $\nu_1, \ldots, \nu_m$, but we can write

$$|\phi_{f,T}^{n_h}(\alpha/T) - \phi_{a_j e^{i2\pi\nu_j t/T},T}^{n_h}(\alpha/T)| \leq \sum_{\substack{l=1 \\ l\neq j}}^{m} |\phi_{a_l e^{i2\pi\nu_l t/T},T}^{n_h}(\alpha/T)|.$$

If α is close to ν_j, then $\sum_{l\neq j} |\phi_{a_l e^{i2\pi\nu_l t/T},T}^{n_h}(\alpha/T)|$ (that is, leakage from the other frequencies) will be small, so $|\phi_{f,T}^{n_h}(\alpha/T)|$ will be close to $|\phi_{a_j e^{i2\pi\nu_j t/T},T}^{n_h}(\alpha/T)|$, and therefore will have a maximum near ν_j. In this way, looking to the local maxima of the function $|\phi_{f,T}^{n_h}(\alpha/T)|^2$, we get a first procedure for computing an estimate of the frequencies. This is the method used by Laskar ([18], [20], [19]). In his procedure, used for the computation and analysis of *frequency maps* related to dynamical systems defined by the function $f(t)$, he looks for the local maxima of the function $|\phi_{f,T}^{n_h}(\alpha/T)|^2$ using some numerical quadrature formula for the evaluation of the TCFT at a discrete set of values of the argument α.

Once some values of α near the maxima have been computed, the values of the maxima are refined by interpolation.

Since leakage is responsible for the maxima of $|\phi^{n_h}_{f,T}(\alpha/T)|$ not being the true frequencies, the use of filtering improves this procedure, since it reduces leakage. However, it is not advisable to take n_h too large, since as we increase n_h the "peaks" of the TCFT become wider and may shade nearby frequencies (see Fig. 2.1).

In our procedure, we maximize the modulus of the filtered DFT of the initial function, $|F^{n_h}_{f,T,N}(j)|$ (where j may take real values), instead of approximating the TCFT using a numerical quadrature formula and maximizing this approximation. The reasons for this are:

- Although the DFT suffers from aliasing, whereas the TCFT does not, numerical quadrature formulae suffer aliasing at least as much as the DFT does. For instance, using a Newton-Côtes formula with all the sampling points as nodes, which is assumed to be written as $\int_0^T f(t)dt \approx \sum_{l=0}^{N-1} A_l f(lT/N)$, we have

$$\begin{aligned}
\phi^{n_h}_{f,T}\Big(\frac{\alpha+N}{T}\Big) &= \frac{1}{T}\int_0^T H^{n_h}_T(t) f(t) e^{\frac{-i2\pi(\alpha+N)t}{T}} dt \\
&\approx \frac{1}{T}\sum_{l=0}^{M} A_l H^{n_h}_T(l\tfrac{T}{N}) f(l\tfrac{T}{N}) e^{\frac{-i2\pi(\alpha+N)}{T} l\frac{T}{N}} \\
&= \frac{1}{T}\sum_{l=0}^{M} A_l H^{n_h}_T(l\tfrac{T}{N}) f(l\tfrac{T}{N}) e^{\frac{-i2\pi\alpha}{T} l\frac{T}{N}} \approx \phi^{n_h}_{f,T}(\frac{\alpha}{T})
\end{aligned}$$

- The use of a numerical quadrature formula does not guarantee the accuracy of the theoretical TCFT, since the error formulas include a power of the integration step, which in our case is the sampling width, and it does not need to be small. However, the DFT is close to the TCFT when there is no aliasing, as it will be shown in Lemma 3.2.4.

We use Newton's method to maximize $|F^{n_h}_{f,T,N}(j)|$. For that, we need to evaluate $\frac{\partial}{\partial j}|F^{n_h}_{f,T,N}(j)|$, and $\frac{\partial^2}{\partial j^2}|F^{n_h}_{f,T,N}(j/T)|$. Expressions for these functions can be obtained from (1.2). We take the "peaks" of the DFT as initial approximations for Newton's method. That is, given j_0 such that $p_{j_0-1} < p_{j_0} > p_{j_0+1}$, we use $\alpha = j_0$ as initial approximation.

2.3 Computation of the amplitudes assuming known frequencies

Once we know the frequencies $\{\nu_l\}_{l=1}^{N_f}$ in (2.1), we can compute the related amplitudes $\{A^c_l\}_{l=0}^{N_f}$, $\{A^s_l\}_{l=1}^{N_f}$ by asking the DFT of the current quasi–periodic approximation Q_f of f to be equal to the DFT of the sampled data $\{f(t_l)\}_{l=0}^{N-1}$. That is,

$$F^{n_h}_{Q_f,T,N}(j) = F^{n_h}_{f,T,N}(j), \tag{2.2}$$

for suitable values of j. Since we are interested in real functions, we will use the sines–cosines form of the DFT instead of the complex one. In order to get a square system for the $1+2N_f$ unknowns, we select values of j in (2.2) in such a manner that we get

$$\begin{array}{rcl} A_0^c + \sum_{l=1}^{N_f} \left(A_l^c \overline{c}_{\nu_l,N}^{n_h}(0) + A_l^s \widetilde{c}_{\nu_l,N}^{n_h}(0) \right) & = & c_{f,T,N}^{n_h}(0), \\ A_0^c c_1^{n_h}(j_i) + \sum_{l=1}^{N_f} \left(A_l^c \overline{c}_{\nu_l,N}^{n_h}(j_i) + A_l^s \widetilde{c}_{\nu_l,N}^{n_h}(j_i) \right) & = & c_{f,T,N}^{n_h}(j_i), \\ \sum_{l=1}^{N_f} \left(A_l^c \overline{s}_{\nu_l,N}^{n_h}(j_i) + A_l^s \widetilde{s}_{\nu_l,N}^{n_h}(j_i) \right) & = & s_{f,T,N}^{n_h}(j_i), \end{array} \tag{2.3}$$

where

$$\begin{array}{rclcrcl} c_1^{n_h}(j) & = & c_{1,T,N}^{n_h}(j), & & & & \\ \overline{c}_{\nu,N}^{n_h}(j) & = & c_{\cos(\frac{2\pi\nu}{T}),T,N}^{n_h}(j), & & \overline{s}_{\nu,N}^{n_h}(j) & = & s_{\cos(\frac{2\pi\nu}{T}),T,N}^{n_h}(j), \\ \widetilde{c}_{\nu,N}^{n_h}(j) & = & c_{\sin(\frac{2\pi\nu}{T}),T,N}^{n_h}(j), & & \widetilde{s}_{\nu,N}^{n_h}(j) & = & s_{\sin(\frac{2\pi\nu}{T}),T,N}^{n_h}(j), \end{array} \tag{2.4}$$

and the j_i are chosen as the closest integers to ν_i, that is, such that $|j_i - \nu_i| \leq 1/2$ for $i = 1 \div N_f$. Note that $c_1^{n_h}(j)$ is independent of T and N. The fact that $\overline{c}_{\nu,N}^{n_h}(j)$, $\overline{s}_{\nu,N}^{n_h}(j)$, $\widetilde{c}_{\nu,N}^{n_h}(j)$ and $\widetilde{s}_{\nu,N}^{n_h}(j)$ do not depend on T will be shown in Section 2.5.3.

In this way we get a $(1+2N_f) \times (1+2N_f)$ linear system, which, assuming that $j_i \geq 1 + n_h$ for $i = 1 \div N_f$, can be written in compact block form as

$$\begin{pmatrix} 2 & u_1 & \dots & u_{N_f} \\ 0 & B_1^1 & \dots & B_{N_f}^1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & B_1^{N_f} & \dots & B_{N_f}^{N_f} \end{pmatrix} \begin{pmatrix} A_0^c \\ v_1 \\ \vdots \\ v_{N_f} \end{pmatrix} = \begin{pmatrix} c_{f,T,N}^{n_h}(0) \\ w_1 \\ \vdots \\ w_{N_f} \end{pmatrix}, \tag{2.5}$$

where

$$\begin{array}{rclcrcl} u_l & = & \begin{pmatrix} \overline{c}_{\nu_l,N}^{n_h}(0) & \widetilde{c}_{\nu_l,N}^{n_h}(0) \end{pmatrix}, & & B_{i,l} & = & \begin{pmatrix} \overline{c}_{\nu_l,N}^{n_h}(j_i) & \widetilde{c}_{\nu_l,N}^{n_h}(j_i) \\ \overline{s}_{\nu_l,N}^{n_h}(j_i) & \widetilde{s}_{\nu_l,N}^{n_h}(j_i) \end{pmatrix}, \\ v_i & = & \begin{pmatrix} A_i^c \\ A_i^s \end{pmatrix}, & & w_i & = & \begin{pmatrix} c_{f,T,N}^{n_h}(j_i) \\ s_{f,T,N}^{n_h}(j_i) \end{pmatrix}. \end{array} \tag{2.6}$$

Since the DFT decreases as $|\nu - j|$ goes away from zero, this system is near to block–diagonal and therefore is well conditioned. In theorem 3.4.1 we give a bound of the inverse of its coefficient matrix. Moreover, because of its structure, it is very well suited for a 2×2 block Jacobi method, which can be written, if we remove the first equation of (2.5), as

$$v_i^{(n+1)} = B_{i,i}^{-1} \Big(- \sum_{\substack{j=1 \\ j \neq i}}^{N_f} B_{i,j} v_j^{(n)} + w_i \Big), \qquad i = 1 \div N_f. \tag{2.7}$$

Once we have values for $\{A_l^c, A_l^s\}_{l=1}^{N_f}$, we can compute A_0^c from the first equation of (2.5). In corollary 3.4.1 we give a result about the convergence of this Jacobi procedure. In practice,

the results have shown that the convergence is very fast when starting from the values given directly by the DFT. Usually 3 or 4 iterates of the block Jacobi method are enough for Hanning level $n_h = 2$ and a tolerance of 10^{-15} for the maximum difference between two consecutive iterates (these values correspond to the analysis of a trigonometric polynomial with three frequencies).

2.4 Simultaneous improvement of frequencies and amplitudes

Given approximations of frequencies and amplitudes, we can improve them simultaneously by solving a system of equations similar to the one used in the previous section. With respect to that system, we need now an additional equation for each frequency, since frequencies are now unknown. We therefore solve iteratively the system

$$\begin{array}{rcl}
A_0^c + \sum_{l=1}^{N_f} \left(A_l^c \overline{c}_{\nu_l,N}^{n_h}(0) + A_l^s \widetilde{c}_{\nu_l,N}^{n_h}(0) \right) & = & c_{f,T,N}^{n_h}(0), \\
A_0^c c_1^{n_h}(j_i) + \sum_{l=1}^{N_f} \left(A_l^c \overline{c}_{\nu_l,N}^{n_h}(j_i) + A_l^s \widetilde{c}_{\nu_l,N}^{n_h}(j_i) \right) & = & c_{f,T,N}^{n_h}(j_i), \\
\sum_{l=1}^{N_f} \left(A_l^c \overline{s}_{\nu_l,N}^{n_h}(j_i) + A_l^s \widetilde{s}_{\nu_l,N}^{n_h}(j_i) \right) & = & s_{f,T,N}^{n_h}(j_i), \\
A_0^c cs_1^{n_h}(j_i^+) + \sum_{l=1}^{N_f} \left(A_l^c \overline{cs}_{\nu_l,N}^{n_h}(j_i^+) + A_l^s \widetilde{cs}_{\nu_l,N}^{n_h}(j_i^+) \right) & = & cs_{f,T,N}^{n_h}(j_i^+),
\end{array} \tag{2.8}$$

for $\{\nu_l\}_{l=1}^{N_f}$, $\{A_l^c\}_{l=0}^{N_f}$, $\{A_l^s\}_{l=1}^{N_f}$, where j_i and j_i^+ are defined as

$$\begin{array}{l}
j_i = [\nu_i],\ j_i^+ = [\nu_i] + 1 \quad \text{if } \nu_i - [\nu_i] \le 1/2, \\
j_i = [\nu_i] + 1,\ j_i^+ = [\nu_i] \quad \text{otherwise},
\end{array}$$

for $i = 1 \div N_f$. In the last equation of (2.8), cs denotes either c or s; the criterium to choose one or the other is given bellow.

If $j_i \ge 1 + n_h$ for $i = 1 \div N_f$, the differential of (2.8) with respect to the unknowns

$$\left(\begin{array}{ccccccc} A_0^c & \nu_1 & A_1^c & A_1^s & \dots & A_{N_f}^c & A_{N_f}^s \end{array} \right),$$

which is needed in order to apply Newton's method, can be written as

$$M = \begin{pmatrix} 2 & v_1 & \dots & v_{N_f} \\ 0 & B_1^1 & \dots & B_{N_f}^1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & B_1^{N_f} & \dots & B_{N_f}^{N_f} \end{pmatrix},$$

being

$$
\begin{aligned}
v_l &= \left(A_l^c \partial\overline{c}_{\nu_l,N}^{n_h}(0) + A_l^s \partial\widetilde{c}_{\nu_l,N}^{n_h}(0) \quad \overline{c}_{\nu_l,N}^{n_h}(0) \quad \widetilde{c}_{\nu_l,N}^{n_h}(0) \right), \\
B_{i,l} &= \begin{pmatrix} A_l^c \partial\overline{c}_{\nu_l,N}^{n_h}(j_i) + A_l^s \partial\widetilde{c}_{\nu_l,N}^{n_h}(j_i) & \overline{c}_{\nu_l,N}^{n_h}(j_i) & \widetilde{c}_{\nu_l,N}^{n_h}(j_i) \\ A_l^c \partial\overline{s}_{\nu_l,N}^{n_h}(j_i) + A_l^s \partial\widetilde{s}_{\nu_l,N}^{n_h}(j_i) & \overline{s}_{\nu_l,N}^{n_h}(j_i) & \widetilde{s}_{\nu_l,N}^{n_h}(j_i) \\ A_l^c \partial\overline{cs}_{\nu_l,N}^{n_h}(j_i^+) + A_l^s \partial\widetilde{cs}_{\nu_l,N}^{n_h}(j_i^+) & \overline{cs}_{\nu_l,N}^{n_h}(j_i^+) & \widetilde{cs}_{\nu_l,N}^{n_h}(j_i^+) \end{pmatrix},
\end{aligned}
\tag{2.9}
$$

where ∂ denotes derivative with respect to ν. As in the preceding section, this matrix is close to block–diagonal and, therefore, the system to be solved at each Newton iteration is well conditioned.

For each block $B_{i,l}$, the criterium to choose cs from c and s is to set it equal to the one that minimizes $\|B_{i,i}^{-1}\|_\infty$. This further improves the well–conditioning of the system, and is theoretically justified in Section 3 (Lemma 3.2.9).

With the exception of rounding errors, the only source of error in this procedure is the leakage from frequencies that we are skipping, as will be shown in Section 3. In particular, this method is exact for trigonometric polynomials (the combination of the procedures of Sections 2.2 and 2.3 is not).

As a final remark note that, because of the use of the DFT, both this procedure and the one of the previous section suffer from the aliasing effect introduced in Section 1.3.

2.5 Implementation details

In this section we give some details for the implementation of the procedures described in the previous section.

2.5.1 Algorithm for the procedure

Starting from the sampling $\{f(t_l)\}_{l=0}^{N-1}$ of a function which is known to have a quasi–periodic behavior, we carry out its Fourier analysis by finding initial approximations for the frequencies using the procedure of Section 2.2, obtaining the related amplitudes using Section 2.3 and iteratively refining the approximations of frequencies and amplitudes through Section 2.4.

In order to prevent some frequencies to "hide" nearby frequencies of lower amplitude, it is advisable to proceed iteratively, in such a way that at each iteration we only consider those frequencies whose amplitude is greater than a given tolerance.

Concretely, the algorithm used for the numerical examples of the last section is the following.

Algorithm 2.5.1 *Provided a minimum amplitude b_{min} for the frequencies to be computed, and a number of iterations n for the procedure, first define*

$$
p_{max} = \max_{j=1 \div \frac{N}{2}} p_{f,T,N}^{n_h}(j), \qquad db = (b_{min}/p_{max})^{1/n},
$$

where $p_{f,T,N}^{n_h}(j) = ((c_{f,T,N}^{n_h}(j))^2 + (s_{f,T,N}^{n_h}(j))^2)^{1/2}$, and set

$$
Q_f(t) = 0, \qquad b = p_{max}, \qquad N_f = 0.
$$

$Q_f(t)$ will be the current quasi–periodic approximation of f, b the minimum amplitude of the frequencies to be detected in the current iteration, and N_f the number of frequencies computed. Then, while $b > b_{min}$, proceed as:

1. *Set $b \leftarrow b \cdot db$. Let $k_{N_f+1}, \ldots, k_{N_f+m}$ be the peaks of the modulus of the DFT of $f - Q_f$ with minimum amplitude b, that is, $\{k_{N_f+1}, \ldots k_{N_f+m}\} = \{j \in \mathbb{Z} : n_h + 2 \leq j \leq \frac{N}{2} - n_h - 2,\ p^{n_h}_{f-Q_f,T,N}(j) \geq b,\ p^{n_h}_{f-Q_f,T,N}(j-1) \leq p^{n_h}_{f-Q_f,T,N}(j) \geq p^{n_h}_{f-Q_f,T,N}(j+1)\}$. For each k_l, apply the procedure of Section 2.2 to obtain ν_l.*

2. *Solve (2.3), according to Section 2.3, to get $\{A^c_l\}_{l=0}^{N_f+m}$ and $\{A^s_l\}_{l=1}^{N_f+m}$ from $\{\nu_l\}_{l=1}^{N_f+m}$.*

3. *Solve (2.8), according to Section 2.4 to iteratively refine $\{\nu_l\}_{l=1}^{N_f+m}$, $\{A^c_l\}_{l=0}^{N_f+m}$ and $\{A^s_l\}_{l=1}^{N_f+m}$.*

4. *Update the number of frequencies and the current quasi–periodic approximation,*

$$N_f \leftarrow N_f + m, \qquad Q_f(t) \leftarrow A^c_0 + \sum_{l=1}^{N_f} (A^c_l \cos(\frac{2\pi\nu_l t}{T}) + A^s_l \sin(\frac{2\pi\nu_l t}{T}))$$

   and go to step 1.

We stop the algorithm if

- *N_f reaches a given maximum number of frequencies, or if*
- *$\max_{l=0\div N-1} |f(t_l) - Q_f(t_l)|$ is under a given tolerance, or if*
- *$\max_{j=0\div j/2} p^{n_h}_{f-Q_f,T,N}(j)$ is under a given tolerance, or if*
- *there appear two frequencies too close. We usually consider ν_{l_1}, ν_{l_2} to be too close if $|\nu_{l_1} - \nu_{l_2}| < 2 + n_h$.*

In practice, the DFT approximation is good enough for Newton's method of Section 2.4 to converge. That is, we can skip the preliminary determination of the frequencies of Section 2.2 by setting $\nu_l = k_l$ in step 1, and then compute the amplitudes related to these frequencies following step 2. It may be useful to use the procedure of Section 2.2 anyway when N_f is very large and we want to save some Newton iterates in step 3, since we have to solve a $(1 + 3N_f) \times (1 + 3N_f)$ linear system at each Newton iterate.

2.5.2 Use of trigonometric recurrences

Large amounts of computing time can be saved if we avoid the evaluation of the sin and cos functions when we have to evaluate the DFT using its definition in the procedure of Section 2.2, or when we have to compute $\{f(t_l) - Q_f(t_l)\}_{l=0}^{N-1}$ in step 1 of the algorithm given above. This can be accomplished through the use of trigonometric recurrences for the evaluation of $\cos(lx)$ and $\sin(lx)$ for $l \in \mathbb{N}$ and $x \in \mathbb{R}$. One has to be careful in choosing

such recurrences, in order to avoid numerical instability (see [29] for a discussion). The recurrence we have used is given in [29], p. 24: first set

$$dc_1 := -2\sin^2\frac{x}{2}, \quad t := 2dc_1, \quad ds_1 := \sqrt{-dc_1(2+dc_1)}, \quad s_0 := 0, \quad c_0 := 1,$$

and then compute, for $m := 1, 2, \ldots$,

$$\begin{array}{rcll} c_m & := & c_{m-1} + dc_m, & dc_{m+1} := t \cdot c_m + dc_m, \\ s_m & := & s_{m-1} + ds_m, & ds_{m+1} := t \cdot s_m + ds_m. \end{array}$$

Just for illustrating purposes, we compare in Fig. 2.2 the errors produced by the trigonometric recurrence previously given with the following one: first set

$$cc_0 = 1, \quad cc_1 = \cos x, \quad ss_0 = 0, \quad ss_1 = \sin x,$$

and then compute, for $m := 2, 3, \ldots$,

$$\begin{array}{rcl} cc_{m+1} & = & (2\cos x)cc_m - cc_{m-1}, \\ ss_{m+1} & = & (2\cos x)ss_m - ss_{m-1}. \end{array}$$

2.5.3 Evaluation of the DFT of sines and cosines

Special care must be taken in the evaluation of the DFT of sines and cosines, in order to avoid cancellations and singularities. In this section we describe some of the strategies followed in our implementation, especially those related to small values of $\nu - j$.

The DFT of sines and cosines can be evaluated from the complex DFT of a complex exponential term through the following formulae:

$$\begin{array}{rcl} \overline{c}^{n_h}_{\nu,N}(j) & = & \mathrm{Re}\, F^{n_h}_{e^{i2\pi\nu t/T},T,N}(j) + \mathrm{Re}\, F^{n_h}_{e^{i2\pi(-\nu)t/T},T,N}(j), \\ \overline{s}^{n_h}_{\nu,N}(j) & = & -\mathrm{Im}\, F^{n_h}_{e^{i2\pi\nu t/T},T,N}(j) - \mathrm{Im}\, F^{n_h}_{e^{i2\pi(-\nu)t/T},T,N}(j), \\ \widetilde{c}^{n_h}_{\nu,N}(j) & = & \mathrm{Im}\, F^{n_h}_{e^{i2\pi\nu t/T},T,N}(j) - \mathrm{Im}\, F^{n_h}_{e^{i2\pi(-\nu)t/T},T,N}(j), \\ \widetilde{s}^{n_h}_{\nu,N}(j) & = & \mathrm{Re}\, F^{n_h}_{e^{i2\pi\nu t/T},T,N}(j) - \mathrm{Re}\, F^{n_h}_{e^{i2\pi(-\nu)t/T},T,N}(j). \end{array}$$

Derivating with respect to ν, we get similar relations that allow to obtain $\partial\overline{c}^{n_h}_{\nu,T,N}(j)$, $\partial\widetilde{c}^{n_h}_{\nu,T,N}(j)$, $\partial\overline{s}^{n_h}_{\nu,T,N}(j)$, $\partial\widetilde{s}^{n_h}_{\nu,T,N}(j)$ from $\partial F^{n_h}_{e^{i2\pi(-\nu)t/T},T,N}(j)$. As before, ∂ denotes derivative with respect to ν.

The non-filtered complex DFT of a complex exponential term is a geometric progression,

$$F_{e^{i2\pi\nu t/T},T,N}(j) = \frac{1}{N}\sum_{l=0}^{N-1} e^{i2\pi(\nu-j)l/N} = \frac{1-e^{i2\pi(\nu-j)}}{N(1-e^{i2\pi(\nu-j)/N})},$$

and its derivative with respect to ν is

$$\partial F^{n_h}_{e^{i2\pi\nu t/T},T,N}(j) = \frac{i2\pi}{N}\cdot\frac{\frac{1}{N}e^{i2\pi(\nu-j)} - e^{i2\pi(\nu-j)} + \frac{N-1}{N}e^{i2\pi(N+1)(\nu-j)/N}}{(1-e^{i2\pi(\nu-j)/N})^2}.$$

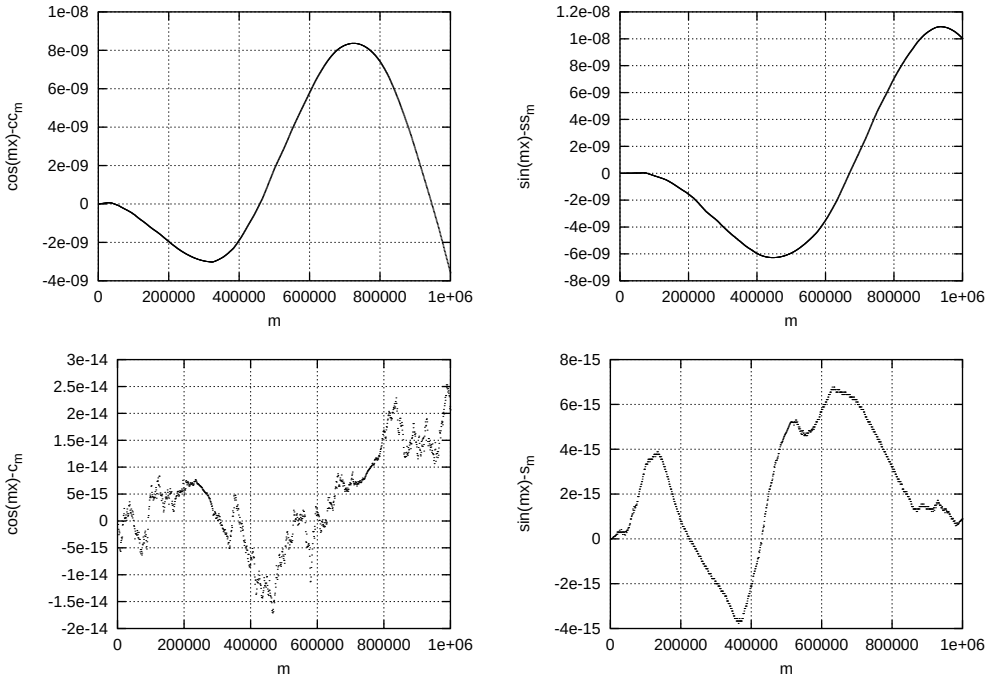


Figure 2.2: Illustration of the numerical instability of trigonometric recurrences. For $x = 2\pi \times 10^{-6}$ and $n = 10^6$, we have evaluated $\{\cos(mx), \sin(mx)\}_{m=0}^{n}$ and the values $\{c_m, s_m\}_{m=0}^{n}$ and $\{cc_m, ss_m\}_{m=0}^{n}$ using the recurrences detailed in the text. In the left–hand plots we show the differences $\cos(mx) - cc_m$ (top) and $\cos(mx) - c_m$ (bottom). In the right–hand ones, we show the differences $\sin(mx) - ss_m$ (top) and $\sin(mx) - s_m$. These values are machine and compiler–dependent. The program that has computed these plots has been compiled with GNU gcc 2.95.2 with the optimization option '–O3' on an Intel Pentium II processor.

Using the relations of Lemma 1.2.1 and its derivatives, we can get the filtered complex DFT of $e^{i2\pi\nu t/T}$, as well as its derivatives with respect to ν, from the non–filtered values.

The computation of $F_{e^{i2\pi\nu t/T},T,N}(j)$ is organized as follows:

$$F_{e^{i2\pi\nu t/T},T,N}(j) = \frac{1}{N}\Big(\frac{ac+bd}{c^2+d^2} + i\frac{bc-ad}{c^2+d^2}\Big),$$

being

$$\begin{aligned} a &= 1-\cos(2\pi(\nu-j)) = 2\sin^2(\pi(\nu-j)), & b &= -\sin(2\pi(\nu-j)),\\ c &= 1-\cos(2\pi(\nu-j)/N) = 2\sin^2(\pi(\nu-j)/N), & d &= -\sin(2\pi(\nu-j)/N). \end{aligned}$$

For a and c, we use the second expressions in order to avoid cancellations.

Concerning to $\partial F_{e^{i2\pi\nu t/T},T,N}(j)$, we compute it as

$$\partial F_{e^{i2\pi\nu t/T},T,N}(j) = \frac{2\pi}{N}\Big(\frac{ad-bc}{c^2+d^2} + i\frac{ac+bd}{c^2+d^2}\Big),$$

being

$$\begin{aligned} a &= \frac{1}{N}\cos\Big(\frac{2\pi(\nu-j)}{N}\Big) - \cos(2\pi(\nu-j)) + \frac{N-1}{N}\cos\Big(\frac{2\pi(N+1)(\nu-j)}{N}\Big)\\ &= \frac{2}{N}\sin\Big(\frac{\pi(2+N)(\nu-j)}{N}\Big)\sin(\pi(\nu-j)) - 2\sin\Big(\frac{\pi(2N+1)(\nu-j)}{N}\Big)\sin\Big(\frac{\pi(\nu-j)}{N}\Big),\\ b &= \frac{1}{N}\sin\Big(\frac{2\pi(\nu-j)}{N}\Big) - \sin(2\pi(\nu-j)) - 2\sin\Big(\frac{\pi(2N+1)(\nu-j)}{N}\Big)\sin\Big(\frac{\pi(\nu-j)}{N}\Big)\\ &= -\frac{2}{N}\cos\Big(\frac{\pi(2+N)(\nu-j)}{N}\Big)\sin(\pi(\nu-j)) + 2\cos\Big(\frac{2\pi(N+1)(\nu-j)}{N}\Big)\sin\Big(\frac{\pi(\nu-j)}{N}\Big),\\ e &= 1-\cos\Big(\frac{2\pi(\nu-j)}{N}\Big) = 2\sin^2\Big(\frac{\pi(\nu-j)}{N}\Big),\\ f &= -\sin\Big(\frac{2\pi(\nu-j)}{N}\Big),\\ c &= e^2-f^2,\\ d &= 2ef. \end{aligned}$$

As before, for a, b, and e we use the second form in order to avoid cancellations, although the second expression of b does not remove cancellations completely.

Since $F_{e^{i2\pi\nu t/T},T,N}(j) = h(\nu-j)$, being

$$h(\alpha) = \frac{1}{N}\sum_{l=0}^{N-1} e^{i2\pi\alpha l/N} = \frac{1-e^{i2\pi\alpha}}{N(1-e^{i2\pi\alpha/N})},$$

we can use the Taylor expansion of h to evaluate $F_{e^{i2\pi\nu t/T},T,N}(j)$ and $\partial F_{e^{i2\pi\nu t/T},T,N}(j)$ for $|\nu-j|$ small. Indeed, we could use this Taylor expansion for any $|\nu-j|$, because h is an entire function (it is a finite sum of exponentials), but the convergence of the expansion is slow for large $|\nu-j|$. We have set a threshold δ in such a way that for $|\nu-j|\geq\delta$ we use the previous formulation and for $|\nu-j|<\delta$ we use the Taylor expansion. The value

of δ is chosen in order to have fast convergence of the Taylor expansion and a small error due to the cancellation in the second expression of b. We have taken $\delta = 0.1$, since

$$\lim_{N\to\infty} \frac{-\beta+\gamma}{\beta} = 0.1563,$$

being

$$\beta = \frac{2}{N}\cos\Big(\frac{\pi(2+N)\delta}{N}\Big)\sin(\pi\delta), \qquad \gamma = 2\cos\Big(\frac{2\pi(N+1)\delta}{N}\Big)\sin\Big(\frac{\pi\delta}{N}\Big),$$

and, in this way, the maximum loss of precision due to the cancellation of the second expression of b is one order of magnitude. Moreover, since

$$\Big|\frac{h^{(k)}(0)}{k!}\delta^k\Big| \leq \frac{(2\pi)^k}{(k+1)!}|\delta|^k,$$

the convergence of the Taylor expansion is fast for $|\nu - j| < \delta$.

For the evaluation of the Taylor expansion of h, we have used that

$$h^{(k)}(0) = \frac{(i2\pi)^k}{N^{k+1}}\sum_{l=0}^{N-1} l^k,$$

and, for $k \geq 1$,

$$\sum_{l=1}^{N-1} l^k = \frac{(N-1)^{k+1}}{k+1} + \frac{(N-1)^k}{2} + \frac{1}{2}\binom{k}{1}B_2(N-1)^{k-1} + \frac{1}{4}\binom{k}{3}B_4(N-1)^{k-3} + \ldots,$$

where the sum ends at either $N-1$ or $(N-1)^2$, and B_i are the Bernoulli numbers.

Chapter 3

Error estimates

In this Chapter we develop error estimates for the numerical procedure described in the previous Chapter. We cover both the case of computation of the amplitudes from known frequencies and the iterative improvement of frequencies and amplitudes. The Chapter is ended with Theorem 3.4.1, which gives bounds for the error of our Fourier analysis procedure in terms of the parameters used for the analysis and the properties of the analyzed function.

3.1 Introduction and notation

In order to derive error bounds, we will assume through this section that the function f to be analyzed is real analytic and quasi–periodic, that is

$$f(t) = \sum_{k\in\mathbb{Z}^m} a_k e^{i2\pi k\omega t} = A_0^c + \sum_{\substack{k\in\mathbb{Z}^m \\ k\omega>0}} (A_k^c \cos(2\pi k\omega t) + A_k^s \sin(2\pi k\omega t)) \tag{3.1}$$

where $k\omega = k_1\omega_1 + \ldots + k_m\omega_m$, $A_k = 2\,\mathrm{Re}\, a_k$, $B_k = -2\,\mathrm{Im}\, a_k$, the frequency vector $\omega = (\omega_1, \ldots, \omega_m)$ is assumed to satisfy a Diophantine condition of the form

$$|k\omega| \geq \frac{D}{|k|^\tau}, \tag{3.2}$$

with $D, \tau > 0$, and the Fourier coefficients of f satisfy the *Cauchy estimates*,

$$|a_k| \leq Ce^{-\delta|k|}, \quad \forall k \in \mathbb{Z}^m. \tag{3.3}$$

We will also assume that we want to compute the frequencies of f up to order $|k| \leq r_0 - 1$ as well as its related amplitudes. That is, we want to approximate f by a trigonometric polynomial

$$p(t) = A_0^c + \sum_{l=1}^{N_f} (A_l^c \cos(\frac{2\pi\nu_l}{T}t) + A_l^s \sin(\frac{2\pi\nu_l}{T}t)),$$

being

$$\{\nu_1, \ldots, \nu_{N_f}\} = \{Tk\omega : k \in \mathbb{Z}^m, 1 \leq |k| \leq r_0 - 1, Tk\omega > 0\}. \tag{3.4}$$

We will give error bounds for two cases: the case in which we want to compute the amplitudes from known frequencies (Section 2.3) and the case in which both frequencies and amplitudes are unknown (Section 2.4). In order to perform error analysis for the second case, we split the right–hand side of (2.8) and rewrite it as

$$\begin{aligned}
A_0^c + \sum_{l=1}^{N_f} (A_l^c \overline{c}_{\nu_l,N}^{n_h}(0) + A_l^s \widetilde{c}_{\nu_l,N}^{n_h}(0)) &= c_{p,T,N}^{n_h}(0) + c_{f-p,T,N}^{n_h}(0) \\
A_0^c c_1^{n_h}(j_i) + \sum_{l=1}^{N_f} (A_l^c \overline{c}_{\nu_l,N}^{n_h}(j_i) + A_l^s \widetilde{c}_{\nu_l,N}^{n_h}(j_i)) &= c_{p,T,N}^{n_h}(j_i) + c_{f-p,T,N}^{n_h}(j_i) \\
\sum_{l=1}^{N_f} (A_l^c \overline{s}_{\nu_l,N}^{n_h}(j_i) + A_l^s \widetilde{s}_{\nu_l,N}^{n_h}(j_i)) &= s_{p,T,N}^{n_h}(j_i) + s_{f-p,T,N}^{n_h}(j_i) \\
\underbrace{A_0^c cs_1^{n_h}(j_i^+) + \sum_{l=1}^{N_f} (A_l^c \overline{cs}_{\nu_l,N}^{n_h}(j_i^+) + A_l^s \widetilde{cs}_{\nu_l,N}^{n_h}(j_i^+))}_{g(y+\Delta y)} &= \underbrace{cs_{p,T,N}^{n_h}(j_i^+)}_{b} + \underbrace{cs_{f-p,T,N}^{n_h}(j_i^+)}_{\Delta b}.
\end{aligned} \tag{3.5}$$

We would get the exact frequencies and amplitudes, which we denote as y for short, if we solved $g(y) = b$. But the system to be solved is $g(y+\Delta y) = b+\Delta b$, and therefore the error we have (assuming no rounding errors) can be bounded (in the first order approximation) by

$$\|\Delta y\|_\infty \lesssim \|Dg(y)^{-1}\|_\infty \|\Delta b\|_\infty$$

A similar argument is applied to the case in which the frequencies are known and we want to compute the amplitudes. In this case, g, y, Δy, b and Δb are defined as

$$\begin{aligned}
A_0^c + \sum_{l=1}^{N_f} (A_l^c \overline{c}_{\nu_l,N}^{n_h}(0) + A_l^s \widetilde{c}_{\nu_l,N}^{n_h}(0)) &= c_{p,T,N}^{n_h}(0) + c_{f-p,T,N}^{n_h}(0) \\
A_0^c c_1^{n_h}(j_i) + \sum_{l=1}^{N_f} (A_l^c \overline{c}_{\nu_l,N}^{n_h}(j_i) + A_l^s \widetilde{c}_{\nu_l,N}^{n_h}(j_i)) &= c_{p,T,N}^{n_h}(j_i) + c_{f-p,T,N}^{n_h}(j_i) \\
\underbrace{\sum_{l=1}^{N_f} (A_l^c \overline{s}_{\nu_l,N}^{n_h}(j_i) + A_l^s \widetilde{s}_{\nu_l,N}^{n_h}(j_i))}_{g(y+\Delta y)} &= \underbrace{s_{p,T,N}^{n_h}(j_i)}_{b} + \underbrace{s_{f-p,T,N}^{n_h}(j_i)}_{\Delta b}.
\end{aligned} \tag{3.6}$$

This section is devoted to the computation of bounds for $\|Dg(y)^{-1}\|_\infty$ and $\|\Delta b\|_\infty$ in terms of T, N, n_h and the properties of the analyzed function f. From now on, and unless otherwise stated, we will use the supremum norm.

3.2 Error bounds for $\|Dg(y)^{-1}\|_\infty$

In order to simplify the expressions to be manipulated, we will bound the TCFT instead of the DFT. That is, we will use $\mathcal{C}_{f,T}^{n_h}$ and $\mathcal{S}_{f,T}^{n_h}$ defined by

$$\phi_{f,T}^{n_h}(\frac{j}{T}) = \frac{1}{2}(\mathcal{C}_{f,T}^{n_h}(j) - i\mathcal{S}_{f,T}^{n_h}(j)), \tag{3.7}$$

(here $i = \sqrt{-1}$). As in the discrete case (2.4), we will note

$$\begin{array}{rclcrcl}
\mathcal{C}_1^{n_h}(j) & = & \mathcal{C}_{1,T,N}^{n_h}(j), & & & & \\
\overline{\mathcal{C}}_\nu^{n_h}(j) & = & \mathcal{C}_{\cos(\frac{2\pi\nu t}{T}),T}^{n_h}(j), & & \overline{\mathcal{S}}_\nu^{n_h}(j) & = & \mathcal{S}_{\cos(\frac{2\pi\nu t}{T}),T}^{n_h}(j), \\
\widetilde{\mathcal{C}}_\nu^{n_h}(j) & = & \mathcal{C}_{\sin(\frac{2\pi\nu t}{T}),T}^{n_h}(j), & & \widetilde{\mathcal{S}}_\nu^{n_h}(j) & = & \mathcal{S}_{\sin(\frac{2\pi\nu t}{T}),T}^{n_h}(j),
\end{array}$$

and the derivatives of $\overline{\mathcal{C}}_\nu^{n_h}(j)$, $\overline{\mathcal{S}}_\nu^{n_h}(j)$, $\widetilde{\mathcal{C}}_\nu^{n_h}(j)$ and $\widetilde{\mathcal{S}}_\nu^{n_h}(j)$ with respect to ν will be denoted as $\partial\overline{\mathcal{C}}_\nu^{n_h}(j)$, $\partial\overline{\mathcal{S}}_\nu^{n_h}(j)$, $\partial\widetilde{\mathcal{C}}_\nu^{n_h}(j)$ and $\partial\widetilde{\mathcal{S}}_\nu^{n_h}(j)$, respectively. We give expressions for these transforms in the following

Lemma 3.2.1 *Denote $\psi_{n_h}(x) = \prod_{l=-n_h}^{n_h}(x+l)$. We have*

$$\begin{aligned}
\overline{\mathcal{C}}_\nu^{n_h}(j) &= \frac{(-1)^{n_h}(n_h!)^2}{2\pi}\Big(\frac{\sin\left(2\pi(\nu-j)\right)}{\psi_{n_h}(\nu-j)} + \frac{\sin\left(2\pi(-\nu-j)\right)}{\psi_{n_h}(-\nu-j)}\Big), \\
\widetilde{\mathcal{C}}_\nu^{n_h}(j) &= \frac{(-1)^{n_h}(n_h!)^2}{2\pi}\Big(\frac{1-\cos\left(2\pi(\nu-j)\right)}{\psi_{n_h}(\nu-j)} - \frac{1-\cos\left(2\pi(-\nu-j)\right)}{\psi_{n_h}(-\nu-j)}\Big), \\
\overline{\mathcal{S}}_\nu^{n_h}(j) &= \frac{(-1)^{n_h}(n_h!)^2}{2\pi}\Big(-\frac{1-\cos\left(2\pi(\nu-j)\right)}{\psi_{n_h}(\nu-j)} - \frac{1-\cos\left(2\pi(-\nu-j)\right)}{\psi_{n_h}(-\nu-j)}\Big), \\
\widetilde{\mathcal{S}}_\nu^{n_h}(j) &= \frac{(-1)^{n_h}(n_h!)^2}{2\pi}\Big(\frac{\sin\left(2\pi(\nu-j)\right)}{\psi_{n_h}(\nu-j)} - \frac{\sin\left(2\pi(-\nu-j)\right)}{\psi_{n_h}(-\nu-j)}\Big), \\
\partial\overline{\mathcal{C}}_\nu^{n_h}(j) &= \frac{(-1)^{n_h}(n_h!)^2}{2\pi}\Big(\frac{h_r(\nu-j)}{\psi_{n_h}(\nu-j)} - \frac{h_r(-\nu-j)}{\psi_{n_h}(-\nu-j)}\Big), \\
\partial\widetilde{\mathcal{C}}_\nu^{n_h}(j) &= \frac{(-1)^{n_h}(n_h!)^2}{2\pi}\Big(\frac{h_i(\nu-j)}{\psi_{n_h}(\nu-j)} + \frac{h_i(-\nu-j)}{\psi_{n_h}(-\nu-j)}\Big), \\
\partial\overline{\mathcal{S}}_\nu^{n_h}(j) &= \frac{(-1)^{n_h}(n_h!)^2}{2\pi}\Big(-\frac{h_i(\nu-j)}{\psi_{n_h}(\nu-j)} + \frac{h_i(-\nu-j)}{\psi_{n_h}(-\nu-j)}\Big), \\
\partial\widetilde{\mathcal{S}}_\nu^{n_h}(j) &= \frac{(-1)^{n_h}(n_h!)^2}{2\pi}\Big(\frac{h_r(\nu-j)}{\psi_{n_h}(\nu-j)} + \frac{h_r(-\nu-j)}{\psi_{n_h}(-\nu-j)}\Big),
\end{aligned}$$

where

$$\begin{aligned}
h_r(x) &= 2\pi\cos(2\pi x) - r_{n_h}(x)\sin(2\pi x), \\
h_i(x) &= 2\pi\sin(2\pi x) - r_{n_h}(x)(1-\cos(2\pi x)), \\
r_{n_h}(x) &= \sum_{l=-n_h}^{n_h}\frac{1}{x+l} = \frac{\psi'_{n_h}(x)}{\psi_{n_h}(x)}.
\end{aligned}$$

Proof: We have

$$\begin{aligned}
\overline{\mathcal{C}}_\nu^{n_h}(j) &= \mathcal{C}_{\cos(2\pi\nu t/T),T}^{n_h}(j) \overset{(3.7)}{=} 2\,\mathrm{Re}\,\phi_{\cos(2\pi\nu t/T),T}^{n_h}(j/T) \\
&= 2\,\mathrm{Re}\,\phi_{(e^{i2\pi\nu t/T}+e^{i2\pi(-\nu)t/T})/2,T}^{n_h}(j/T) \\
&= \mathrm{Re}\,\phi_{e^{i2\pi\nu t/T},T}^{n_h}(j/T) + \mathrm{Re}\,\phi_{e^{i2\pi(-\nu)t/T},T}^{n_h}(j/T),
\end{aligned}$$

$$\begin{aligned}
\widetilde{\mathcal{C}}_{\nu}^{n_h}(j) &= \operatorname{Im}\phi^{n_h}_{e^{i2\pi\nu t/T},T}(j/T) - \operatorname{Im}\phi^{n_h}_{e^{i2\pi(-\nu)t/T},T}(j/T),\\
\overline{\mathcal{S}}_{\nu}^{n_h}(j) &= -\operatorname{Im}\phi^{n_h}_{e^{i2\pi\nu t/T},T}(j/T) - \operatorname{Im}\phi^{n_h}_{e^{i2\pi(-\nu)t/T},T}(j/T),\\
\widetilde{\mathcal{S}}_{\nu}^{n_h}(j) &= \operatorname{Re}\phi^{n_h}_{e^{i2\pi\nu t/T},T}(j/T) - \operatorname{Re}\phi^{n_h}_{e^{i2\pi(-\nu)t/T},T}(j/T).
\end{aligned}$$

Then, the lemma follows from (1.5) and

$$\frac{d}{d\nu}\phi^{n_h}_{e^{i2\pi\nu t/T},T}(j/T) = \frac{(-1)^{n_h}(n_h!)^2}{2\pi\psi_{n_h}(\nu-j)}\Big(2\pi e^{i2\pi(\nu-j)} - i(1-e^{i2\pi(\nu-j)})\sum_{l=-n_h}^{n_h}\frac{1}{\nu-j+l}\Big).$$

□

In order to bound the error due to the approximation of the DFT by the TCFT, we need the following lemmas.

Lemma 3.2.2 *(Discrete Poisson summation formula) If $n_h \geq 1$, we have*

$$F^{n_h}_{f,T,N}(j) = \sum_{l=-\infty}^{\infty}\phi^{n_h}_{f,T}\Big(\frac{j+lN}{T}\Big).$$

In particular, $\overline{c}^{n_h}_{\nu,N}(j) = \sum_{l=-\infty}^{\infty}\overline{c}^{n_h}_{\nu}(j+lN)$, and analogous identities hold for $\widetilde{c}^{n_h}_{\nu,N}(j)$, $\overline{s}^{n_h}_{\nu,N}(j)$, $\widetilde{s}^{n_h}_{\nu,N}(j)$, and their derivatives with respect to ν.

Proof: This is a known result (see, for instance, [5]). We give a proof here for completeness, and also to clarify the need for the hypothesis $n_h \geq 1$.

We first note that, by definition of the TCFT, the Fourier expansion of $H^{n_h}_T(t)f(t)$ with respect to the interval $[0,T]$ is

$$\sum_{k=-\infty}^{\infty}\phi^{n_h}_{f,T}(\frac{k}{T})e^{\frac{i2\pi kt}{T}}.$$

The function $H^{n_h}_T(t)f(t)$ coincides with its Fourier expansion for all $t \in [0,T]$ because, since $n_h \geq 1$, we have $H^{n_h}_T(0)f(0) = H^{n_h}_T(T)f(T) = 0$.

Then, using the definition (1.2) of the complex DFT and the above Fourier expansion,

$$\begin{aligned}
F^{n_h}_{f,T,N}(j) &= \frac{1}{N}\sum_{l=0}^{N-1}\Big(\sum_{k=-\infty}^{\infty}\phi^{n_h}_{f,T}(\frac{k}{T})e^{\frac{i2\pi k}{T}l\frac{T}{N}}\Big)e^{-\frac{i2\pi j}{T}l\frac{T}{N}}\\
&= \frac{1}{N}\sum_{k=-\infty}^{\infty}\phi^{n_h}_{f,T}(\frac{k}{T})\sum_{l=0}^{N-1}e^{\frac{i2\pi(k-j)l}{N}},
\end{aligned}$$

and the lemma follows from the fact that the inner sum above is equal to N if $k-j$ is an integer multiple of N and zero otherwise. □

Lemma 3.2.3 *For $|x| \geq n_h + 2$ we have*

$$\begin{aligned} |r_{n_h}(x)| &\leq \ln(|x| + n_h) - \ln(|x| - n_h - 1), \\ |h_r(x)|, |h_i(x)| &\leq 2\pi + 2(\ln(|x| + n_h) - \ln(|x| - n_h - 1)). \end{aligned}$$

Proof: For the first inequality, we have

$$|r_{n_h}(x)| = \sum_{l=-n_h}^{n_h} \frac{1}{|x| + l} \leq \int_{|x|-n_h}^{|x|+n_h+1} \frac{1}{z-1} dz = \ln(|x| + n_h) - \ln(|x| - n_h - 1).$$

The bounds for $|h_r(x)|, |h_i(x)|$ follow from this one. □

Lemma 3.2.4 *For $j \geq 0$, $N - j - |\nu| - n_h > 0$, we have*

$$|F^{n_h}_{e^{i2\pi\nu t/T},T,N}(j) - \phi^{n_h}_{e^{i2\pi\nu t/T},T,N}(\frac{j}{T})| \leq \frac{2(n_h!)^2(1 + \frac{1}{2n_h})}{\pi(N - j - |\nu| - n_h)^{1+2n_h}},$$

for $j, \nu \geq 0$, $N - j - \nu - n_h > 0$,

$$|\widetilde{\overline{cs}}^{n_h}_{\nu,N}(j) - \widetilde{\overline{\mathcal{CS}}}^{n_h}_{\nu}(j)| \leq \frac{4(n_h!)^2(1 + \frac{1}{2n_h})}{\pi(N - j - \nu - n_h)^{1+2n_h}},$$

and for $j, \nu \geq 0$, $N - j - \nu - n_h \geq 2$,

$$|\partial\widetilde{\overline{cs}}^{n_h}_{\nu,N}(j) - \partial\widetilde{\overline{\mathcal{CS}}}^{n_h}_{\nu}(j)| \leq \frac{4(n_h!)^2(1 + \frac{1}{2n_h})(\pi + \ln(N-j-\nu+n_h) - \ln(N-j-\nu-n_h-1))}{\pi(N-j-\nu-n_h)^{1+2n_h}}.$$

In the previous expressions, $\widetilde{\overline{cs}}$ denotes one of $\overline{c}$, $\widetilde{c}$, $\overline{s}$, $\widetilde{s}$, and $\widetilde{\overline{\mathcal{CS}}}$ denotes one of $\overline{\mathcal{C}}$, $\widetilde{\mathcal{C}}$, $\overline{\mathcal{S}}$, $\widetilde{\mathcal{S}}$.

Proof: For the first inequality, using the Discrete Poisson summation formula (Lemma 3.2.2),

$$\begin{aligned} |F^{n_h}_{e^{i2\pi\nu t/T},T,N}(j) - \phi^{n_h}_{e^{i2\pi\frac{\nu}{T}t},T}(\frac{j}{T})| &\leq \sum_{l=1}^{\infty} \Big(|\phi^{n_h}_{e^{i2\pi\frac{\nu}{T}t},T}(\frac{j + lN}{T})| + |\phi^{n_h}_{e^{i2\pi\frac{\nu}{T}t},T}(\frac{j - lN}{T})| \Big) \\ &\overset{(1.5)}{\leq} \sum_{l=1}^{\infty} \Big(\frac{(n_h!)^2}{\pi\psi_{n_h}(|\nu - j - lN|)} + \frac{(n_h!)^2}{\pi\psi_{n_h}(|\nu - j + lN|)} \Big). \end{aligned}$$

Now, if $N - j - |\nu| > n_h$, we have $|\nu - j \pm lN| \geq |j \pm lN| - |\nu| \geq lN - j - |\nu|$, and we can bound the previous series by

$$\begin{aligned} &\sum_{l=1}^{\infty} \frac{2(n_h!)^2}{\pi(lN - j - |\nu| - n_h)^{1+2n_h}} \\ &\qquad \leq \frac{2(n_h)^2}{N\pi} \Big(\frac{N}{(N - j - |\nu| - n_h)^{1+2n_h}} + \int_{2N-j-|\nu|-n_h}^{\infty} \frac{1}{(y - N)^{1+2n_h}} dy \Big) \\ &\qquad \leq \frac{2(n_h)^2}{N\pi} \Big(\frac{N}{(N - j - |\nu| - n_h)^{1+2n_h}} + \frac{1/(2n_h)}{(N - j - |\nu| - n_h)^{2n_h}} \Big) \\ &\qquad \leq \frac{2(n_h!)^2(1 + \frac{1}{2n_h})}{(N - j - |\nu| - n_h)^{1+2n_h}}. \end{aligned}$$

Regarding to the other two inequalities, if we take into account that

$$\begin{aligned} |\widetilde{\overline{\mathcal{CS}}}_{\nu}^{n_h}(j)| &\leq \frac{2(n_h!)^2}{\pi\psi_{n_h}(|\nu-j|)}, \\ |\partial\widetilde{\overline{\mathcal{CS}}}_{\nu}^{n_h}(j)| &\leq \frac{2(n_h!)^2(\pi+\ln(|\nu-j|+n_h)-\ln(|\nu-j|-n_h-1))}{\pi\psi_{n_h}(|\nu-j|)}, \end{aligned}$$

we can easily adapt the previous sequence of inequalities to both cases. Note that, for the second inequality above, we need the hypothesis $N-j-\nu-n_h \geq 2$ in order to apply Lemma 3.2.3. □

In order to bound $\|Dg(y)^{-1}\|$, we will further simplify $\overline{\mathcal{C}}_{\nu}^{n_h}(j)$, $\widetilde{\mathcal{C}}_{\nu}^{n_h}(j)$, etc., by eliminating the second term in the sums given by Lemma 3.2.1. For this we introduce $\overline{\mathfrak{c}}$, $\widetilde{\mathfrak{c}}$, ..., according to the following

Definition 3.2.1 *We define*

$$\begin{aligned} \overline{\mathfrak{c}}_{\nu}^{n_h}(j) &= \frac{(-1)^{n_h}(n_h!)^2}{2\pi}\cdot\frac{\sin(2\pi(\nu-j))}{\psi_{n_h}(\nu-j)}, & \overline{\mathfrak{s}}_{\nu}^{n_h}(j) &= -\widetilde{\mathfrak{c}}_{\nu}^{n_h}(j), \\ \widetilde{\mathfrak{c}}_{\nu}^{n_h}(j) &= \frac{(-1)^{n_h}(n_h!)^2}{2\pi}\cdot\frac{1-\cos(2\pi(\nu-j))}{\psi_{n_h}(\nu-j)}, & \widetilde{\mathfrak{s}}_{\nu}^{n_h}(j) &= \overline{\mathfrak{c}}_{\nu}^{n_h}(j), \\ \partial\overline{\mathfrak{c}}_{\nu}^{n_h}(j) &= \frac{(-1)^{n_h}(n_h!)^2}{2\pi}\cdot\frac{h_r(\nu-j)}{\psi_{n_h}(\nu-j)}, & \partial\overline{\mathfrak{s}}_{\nu}^{n_h}(j) &= -\partial\widetilde{\mathfrak{c}}_{\nu}^{n_h}(j), \\ \partial\widetilde{\mathfrak{c}}_{\nu}^{n_h}(j) &= \frac{(-1)^{n_h}(n_h!)^2}{2\pi}\cdot\frac{h_i(\nu-j)}{\psi_{n_h}(\nu-j)}, & \partial\widetilde{\mathfrak{s}}_{\nu}^{n_h}(j) &= \partial\overline{\mathfrak{c}}_{\nu}^{n_h}(j). \end{aligned}$$

In the following lemma, we bound the error after this simplification.

Lemma 3.2.5 *If $|\nu+j| > n_h+2$, we have*

$$|\partial\overline{\mathfrak{c}}_{\nu}^{n_h}(j)-\partial\overline{\mathcal{C}}_{\nu}^{n_h}(j)| \leq \frac{(n_h!)^2(\pi+\ln(|-\nu-j|+n_h)-\ln(|-\nu-j|-n_h-1))}{\pi(|-\nu-j|-n_h)^{1+2n_h}},$$

and the same bound holds for $|\partial\widetilde{\mathfrak{c}}_{\nu}^{n_h}(j)-\partial\widetilde{\mathcal{C}}_{\nu}^{n_h}(j)|$, $|\partial\overline{\mathfrak{s}}_{\nu}^{n_h}(j)-\partial\overline{\mathcal{S}}_{\nu}^{n_h}(j)|$, $|\partial\widetilde{\mathfrak{s}}_{\nu}^{n_h}(j)-\partial\widetilde{\mathcal{S}}_{\nu}^{n_h}(j)|$. We also have

$$|\overline{\mathfrak{c}}_{\nu}^{n_h}(j)-\overline{\mathcal{C}}_{\nu}^{n_h}(j)| \leq \frac{(n_h!)^2}{\pi(|-\nu-j|-n_h)^{1+2n_h}},$$

and the same bound holds for $|\widetilde{\mathfrak{c}}_{\nu}^{n_h}(j)-\widetilde{\mathcal{C}}_{\nu}^{n_h}(j)|$, $|\overline{\mathfrak{s}}_{\nu}^{n_h}(j)-\overline{\mathcal{S}}_{\nu}^{n_h}(j)|$, $|\widetilde{\mathfrak{s}}_{\nu}^{n_h}(j)-\widetilde{\mathcal{S}}_{\nu}^{n_h}(j)|$.

Proof: It is a direct application of Definition 3.2.1 and Lemma 3.2.3. □

3.2.1 Error estimation for known frequencies

If we assume in (3.6) that $j_i \geq n_h + 1$ for $i = 1 \div N_f$, then $c_1^{n_h}(j_i) = 0$ for $i = 1 \div N_f$ and the first equation of system (3.6) is uncoupled with the other ones. Therefore, we can write $Dg(y)$ as

$$M = \begin{pmatrix} 2 & B_{0,1} & \dots & B_{0,N_f} \\ 0 & B_{1,1} & \dots & B_{1,N_f} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & B_{N_f,1} & \dots & B_{N_f,N_f} \end{pmatrix},$$

where $B_{0,l} = u_l$ are 1×2 blocks, being v_l as defined in (2.6), and $B_{i,l}$, $i, l = 1 \div N_f$, are 2×2 blocks defined as in (2.6). Let us split M in its block–diagonal and block–off–diagonal parts, that is $M = M_D + M_O$, being

$$M_D = \begin{pmatrix} 2 & 0 & \dots & 0 \\ 0 & B_{1,1} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & B_{N_f,N_f} \end{pmatrix}, \quad M_O = \begin{pmatrix} 0 & B_{0,1} & \dots & B_{0,N_f} \\ 0 & 0 & \dots & B_{1,N_f} \\ 0 & \vdots & \ddots & \vdots \\ 0 & B_{N_f,1} & \dots & 0 \end{pmatrix}.$$

From Definition 3.2.1 and lemmas 3.2.5 and 3.2.4, under suitable hypothesis, $\overline{cs}_{\nu,N}^{n_h}(j)$ and $\widetilde{cs}_{\nu,N}^{n_h}(j)$ decrease as ν goes away from j, and therefore M is close to its diagonal part. What we will do is to obtain bounds for $\|M_D^{-1}\|$ and $\|M_O\|$ and then use them to bound $\|M^{-1}\|$ using the following

Lemma 3.2.6 *If M and ΔM are $n \times n$ matrices satisfying that M is invertible and $\|M^{-1}\|\|\Delta M\| < 1$, then $M + \Delta M$ is inversible and satisfies*

$$\|(M + \Delta M)^{-1}\| \leq \frac{\|M^{-1}\|}{1 - \|M^{-1}\|\|\Delta M\|}$$

Proof: See [29], p. 188. □

To simplify notation, we introduce the following

Definition 3.2.2 *We will denote by $\mathcal{M}$, $\mathcal{M}_D$, $\mathcal{M}_O$ and $\mathcal{B}_{i,l}$ the equivalents of M, M_D, M_O and $B_{i,l}$, respectively, but replacing $\overline{c}_\nu^{n_h}$, $\widetilde{c}_\nu^{n_h}$, etc. by $\overline{\mathcal{C}}_\nu^{n_h}$, $\widetilde{\mathcal{C}}_\nu^{n_h}$, etc. That is, by replacing the DFT by the TCFT. We will also denote by $\mathfrak{M}$, $\mathfrak{M}_D$, $\mathfrak{M}_O$ and $\mathfrak{B}_{i,l}$ the equivalents of M, M_D, M_O and $B_{i,l}$, respectively, but replacing $\overline{c}_\nu^{n_h}$, $\widetilde{c}_\nu^{n_h}$, etc. by $\overline{\mathfrak{c}}_\nu^{n_h}$, $\widetilde{\mathfrak{c}}_\nu^{n_h}$, etc. In this way, for instance,*

$$\mathfrak{B}_{i,l} = \begin{pmatrix} \overline{\mathfrak{c}}_{\nu_l}^{n_h}(j_i) & \widetilde{\mathfrak{c}}_{\nu_l}^{n_h}(j_i) \\ \overline{\mathfrak{s}}_{\nu_l}^{n_h}(j_i) & \widetilde{\mathfrak{s}}_{\nu_l}^{n_h}(j_i) \end{pmatrix}$$

In the following proposition, we give bounds for $\|\mathfrak{M}_D^{-1}\|$ and $\|\mathfrak{M}_O\|$.

Proposition 3.2.1 *If* $\frac{TD}{(2r_0-2)^\tau} > \frac{1}{2} + n_h$, *where* D, τ *are given by (3.2) and* r_0 *is given by (3.4), then*

$$\|\mathfrak{M}_D^{-1}\| \leq \frac{5}{3}, \qquad \|\mathfrak{M}_O\| \leq \frac{\sqrt{2} N_f (n_h!)^2}{\pi (\frac{TD}{(2r_0-1)^\tau} - \frac{1}{2} - n_h)^{1+2n_h}}.$$

Proof: From the definitions of $\mathfrak{M}_D$ and $\mathfrak{M}_O$, we have

$$\|\mathfrak{M}_D^{-1}\| \leq \max(\max_{i=1,\ldots,N_f} \|(\mathfrak{B}_{i,i})^{-1}\|, \frac{1}{2}), \qquad \|\mathfrak{M}_O\| \leq \max_{i=0,\ldots,N_f} \sum_{l=1}^{N_f} \|\mathfrak{B}_{i,l}\| \tag{3.8}$$

For the second bound, we have used the fact that the bounds that will be found for $\|\mathfrak{M}_{i,l}\|$ are valid for $\|v_l\|$. So, in order to take into account the first row of $\|\mathfrak{M}_O\|$, we have allowed the sum to run for $l = i$.

Denoting $\rho_{i,l} = \nu_l - j_i$ and using the trigonometric identities $\sin(2\varepsilon)/2 = \sin(\varepsilon)\cos(\varepsilon)$, $(1 - \cos(2\varepsilon))/2 = \sin^2(\varepsilon)$, we can write

$$\mathfrak{B}_{i,l} = \frac{(-1)^{n_h}(n_h!)^2 \sin(\pi\rho_{i,l})}{\pi\psi_{n_h}(\rho_{i,l})} \begin{pmatrix} \cos(\pi\rho_{i,l}) & \sin(\pi\rho_{i,l}) \\ -\sin(\pi\rho_{i,l}) & \cos(\pi\rho_{i,l}) \end{pmatrix}.$$

Therefore

$$(\mathfrak{B}_{i,l})^{-1} = \frac{\pi\psi_{n_h}(\rho_{i,l})}{(-1)^{n_h}(n_h!)^2 \sin(\pi\rho_{i,l})} \begin{pmatrix} \cos(\pi\rho_{i,l}) & -\sin(\pi\rho_{i,l}) \\ \sin(\pi\rho_{i,l}) & \cos(\pi\rho_{i,l}) \end{pmatrix}$$

and

$$\begin{aligned} \|\mathfrak{B}_{i,l}\| &= \Big|\frac{(n_h!)^2 \sin(\pi\rho_{i,l})}{\pi\psi_{n_h}(\rho_{i,l})}\Big| (|\cos(\pi\rho_{i,l})| + |\sin(\pi\rho_{i,l})|), \\ \|(\mathfrak{B}_{i,i})^{-1}\| &= \Big|\frac{\pi\psi_{n_h}(\rho_{i,i})}{(n_h!)^2 \sin(\pi\rho_{i,i})}\Big| (|\cos(\pi\rho_{i,i})| + |\sin(\pi\rho_{i,i})|). \end{aligned} \tag{3.9}$$

Let us define

$$F_1(\rho_{i,i}) = \Big|\frac{\pi\psi_{n_h}(\rho_{i,i})}{(n_h!)^2 \sin(\pi\rho_{i,i})}\Big|$$

Recalling that the j_i, $i = 1 \div N_f$, were chosen such that $|\nu_i - j_i| < 0.5$, we only have to bound (3.9) for $-0.5 \leq \rho_{i,i} \leq 0.5$. From the definition (1.6) of ψ_{n_h}, we can write $F_1(\rho_{i,i})$ as

$$F_1(\rho_{i,i}) = \Big|\frac{\pi\rho_{i,i}}{\sin(\pi\rho_{i,i})}\Big| \prod_{l=1}^{n_h} \frac{l^2 - \rho_{i,i}^2}{l^2},$$

and it is readily checked that if $-0.5 \leq \rho_{i,i} \leq 0.5$ then $F_1(\rho_{i,i})$ decreases as $n_h \to \infty$. The limit is a positive value because of the Weierstrass factorization formula for the sine (see, for instance, [31]):

$$\sin(\pi z) = \pi z \prod_{n=1}^{\infty} \Big(1 - \frac{z^2}{n^2}\Big) \qquad \text{for } z \in \mathbb{C}. \tag{3.10}$$

Therefore, $\mathfrak{B}_{i,i}$ is invertible and we can bound $\|(\mathfrak{B}_{i,i})^{-1}\|$ for $n_h \in \mathbb{N}$ by the bound for $n_h = 0$, which is

$$\|(\mathfrak{B}_{i,i})^{-1}\| \leq \left|\frac{\pi\rho_{i,i}}{\sin(\pi\rho_{i,i})}\right|(|\cos(\pi\rho_{i,i})| + |\sin(\pi\rho_{i,i})|) \leq \frac{5}{3} \qquad \text{for } -0.5 \leq \rho_{i,i} \leq 0.5. \quad (3.11)$$

For the actual behavior of $\|\mathfrak{B}_{i,i}^{-1}\|$ in terms of $\rho_{i,i}$, see Fig. 3.1.

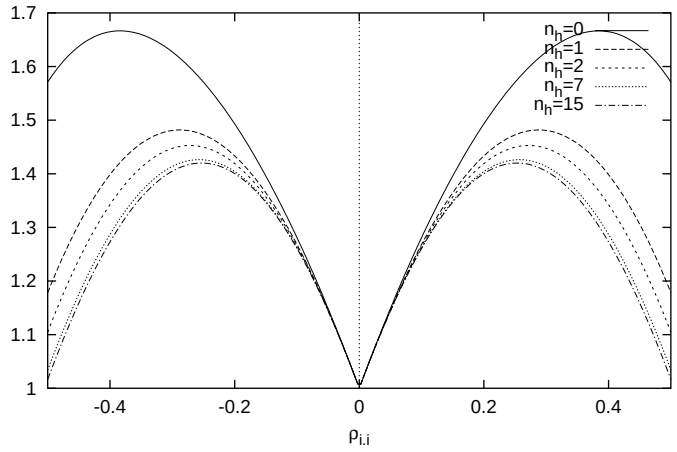


Figure 3.1: Graph of $\|\mathfrak{B}_{i,i}^{-1}\|_\infty$ for $-0.5 \leq \rho_{i,i} \leq 0.5$ and $n_h = 0, 1, 3, 5, 15$.

Concerning $\|\mathfrak{B}_{i,l}\|$, for $\rho_{i,l} \in \mathbb{R}$ and $n_h \in \mathbb{N}$, we have

$$\|\mathfrak{B}_{i,l}\| \leq \frac{(n_h!)^2}{\pi|\psi_{n_h}(\rho_{i,l})|}\max_{x\in\mathbb{R}}(|\cos(\pi x)| + |\sin(\pi x)|) = \frac{\sqrt{2}(n_h!)^2}{\pi|\psi_{n_h}(\rho_{i,l})|}. \quad (3.12)$$

If we define $j_0 = 0$, the previous bound is also valid for $\|\mathfrak{B}_{0,l}\|$. For the actual behavior of $\|\mathfrak{B}_{i,l}\|$ in terms of $\rho_{i,l}$, see Fig. 3.2.

Now from (3.8), (3.11) and (3.12),

$$\|\mathfrak{M}_D^{-1}\| \leq 5/3, \qquad \|\mathfrak{M}_O\| \leq \max_{i=0\div N_f}\sum_{l=1}^{N_f}\frac{\sqrt{2}(n_h!)^2}{\pi|\psi_{n_h}(\nu_l - j_i)|}.$$

Since by definition $|\nu_i - j_i| \leq 1/2$, we have $|\nu_l - j_i| \geq |\nu_l - \nu_i| - 1/2 = T|(k_l - k_i)\omega| - 1/2$. Using the Diophantine condition (3.2),

$$T|(k_l - k_i)\omega| - \frac{1}{2} \;\geq\; \frac{TD}{|k_l - k_i|^\tau} - \frac{1}{2} \;\geq\; \frac{TD}{(2r_0 - 2)^\tau} - \frac{1}{2}.$$

Now since by hypothesis $\frac{TD}{(2r_0-2)^\tau} > \frac{1}{2} + n_h$, we get

$$\|\mathfrak{M}_O\| \leq \frac{\sqrt{2}N_f(n_h!)^2}{\pi(\frac{TD}{(2r_0-2)^\tau} - \frac{1}{2} - n_h)^{1+2n_h}},$$

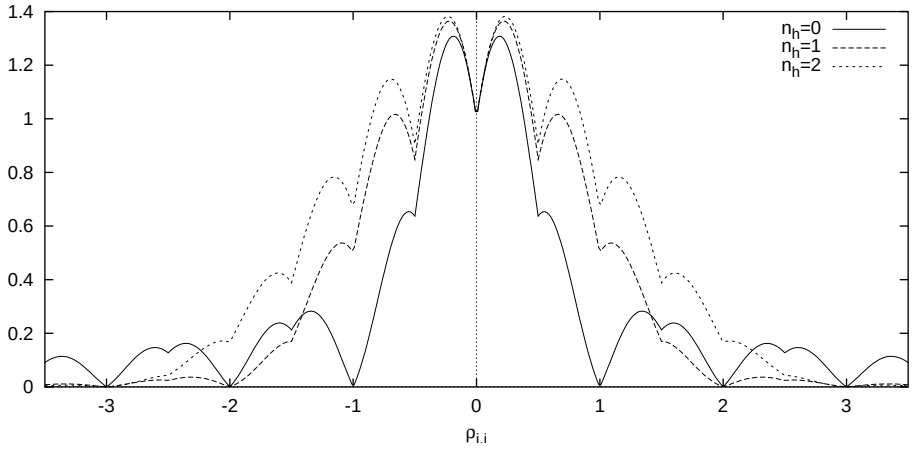


Figure 3.2: Graph of $\|\mathfrak{B}_{i,l}(\rho_{i,l})\|_\infty$ for $n_h = 0, 1, 2$.

and this ends the proposition. □

From the bounds of $\|\mathfrak{M}_D^{-1}\|$ and $\|\mathfrak{M}_O\|$ and Lemma 3.2.6, we can get a bound for $\|M^{-1}\|$. For that, we need bounds of $\|\mathcal{M}_D - \mathfrak{M}_D\|$, $\|M_D - \mathcal{M}_D\|$, $\|\mathcal{M}_O - \mathfrak{M}_O\|$ and $\|M_O - \mathcal{M}_O\|$, which are given in the following lemmas.

Lemma 3.2.7 *If $[\nu_{min}] > n_h$, where $\nu_{min} = \min\{\nu_1, \ldots, \nu_{N_f}\}$ and $[\,]$ denotes integer part, we have*

$$\|\mathcal{M}_D - \mathfrak{M}_D\| \leq \frac{2(n_h!)^2}{\pi(2[\nu_{min}] - n_h)^{1+2n_h}}, \qquad \|\mathcal{M}_O - \mathfrak{M}_O\| \leq \frac{2N_f(n_h!)^2}{\pi([\nu_{min}] - n_h)^{1+2n_h}}$$

Proof: We have

$$\begin{aligned} \|\mathcal{M}_D - \mathfrak{M}_D\| &\leq \max_{i=1\div N_f} \|\mathcal{B}_{i,i} - \mathfrak{B}_{i,i}\| \\ &= \max_{i=1\div N_f} (|\overline{\mathcal{CS}}_{\nu_i}^{n_h}(j_i) - \overline{\mathfrak{cs}}_{\nu_i}^{n_h}(j_i)| + |\widetilde{\mathcal{CS}}_{\nu_i}^{n_h}(j_i) - \widetilde{\mathfrak{cs}}_{\nu_i}^{n_h}(j_i)|), \end{aligned}$$

where either $\mathcal{CS} = \mathcal{C}$ and $\mathfrak{cs} = \mathfrak{c}$ or $\mathcal{CS} = \mathcal{S}$ and $\mathfrak{cs} = \mathfrak{s}$. Now, using Lemma 3.2.5 and the hypothesis, we get the first inequality:

$$\|\mathcal{M}_D - \mathfrak{M}_D\| \leq \max_{i=1\div N_f} \frac{2(n_h!)^2}{\pi\psi_{n_h}(-\nu_i - j_i)} \leq \frac{2(n_h!)^2}{\pi(2[\nu_{min}] - n_h)^{1+2n_h}}.$$

As for the second inequality,

$$\|\mathcal{M}_O - \mathfrak{M}_O\| \leq \max_{i=0\div N_f} \sum_{\substack{l=1 \\ l\neq i}}^{N_f} (|\overline{\mathcal{CS}}_{\nu_l}^{n_h}(j_i) - \overline{\mathfrak{cs}}_{\nu_l}^{n_h}(j_i)| + |\widetilde{\mathcal{CS}}_{\nu_l}^{n_h}(j_i) + \widetilde{\mathfrak{cs}}_{\nu_l}^{n_h}(j_i)|),$$

where we denote $j_0 = 0$. Using Lemma 3.2.5 again,

$$\|\mathcal{M}_O - \mathfrak{M}_O\| \leq \max_{i=0\div N_f} \sum_{l=1}^{N_f} \frac{2(n_h!)^2}{\pi\psi_{n_h}(-\nu_l - j_i)} \leq \frac{2N_f(n_h!)^2}{\pi([\nu_{min}] - n_h)^{1+2n_h}}.$$

We lose the factor 2 in front of $[\nu_{min}]$ with respect to the bound of $\|\mathcal{M}_D - \mathfrak{M}_D\|$ because we have to consider the first row ($i = 0$). □

Lemma 3.2.8 *Assume that $N - T(2r_0 - 2)\|\omega\|_\infty - \frac{1}{2} - n_h > 0$, where r_0 is given by (3.4). Then*

$$\begin{aligned} \|M_D - \mathcal{M}_D\| &\leq \frac{8(n_h!)^2(1+\frac{1}{2n_h})}{\pi(N - T(2r_0-2)\|\omega\|_\infty - \frac{1}{2} - n_h)^{1+2n_h}}, \\ \|M_O - \mathcal{M}_O\| &\leq \frac{8(n_h!)^2 N_f(1+\frac{1}{2n_h})}{\pi(N - T(2r_0-2)\|\omega\|_\infty - \frac{1}{2} - n_h)^{1+2n_h}}. \end{aligned}$$

Proof: Using Lemma 3.2.4,

$$\|M_D - \mathcal{M}_D\| \leq \max_{i=1\div N_f} \|B_{i,i} - \mathcal{B}_{i,i}\| \leq \max_{i=1\div N_f} \frac{8(n_h!)^2(1+\frac{1}{2n_h})}{\pi(N - j_i - \nu_i - n_h)^{1+2n_h}},$$

and, since by definition $|j_i - \nu_i| < 1/2$ and as $\nu_i = Tk_i\omega$ with $|k_i| \leq r_0 - 1$, we have that $j_i + \nu_i \leq 2\nu_i + 1/2 \leq T(2r_0 - 2)\|\omega\|_\infty + 1/2$, and the first inequality follows immediately. A similar argument proves the second inequality. □

The bound for $\|M^{-1}\|$ that follows from the previous results will be given in Theorem 3.4.1.

3.2.2 General case

As in the case of known frequencies, we assume in (3.5) that $j_i > n_h$ for $i = 1 \div N_f$ so the first equation of system (3.5) is uncoupled with the other ones and $M = Dg(y)$ can be written as

$$M = \begin{pmatrix} 2 & B_{0,1} & \dots & B_{0,N_f} \\ 0 & B_{1,1} & \dots & B_{1,N_f} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & B_{N_f,1} & \dots & B_{N_f,N_f} \end{pmatrix}$$

where $B_{0,l} = v_l$ are 1×3 blocks, being v_l as defined in (2.9), and $B_{i,l}$, $i, l = 1 \div N_f$, are 3×3 blocks defined as in (2.9). We split M in its block–diagonal and block–off–diagonal parts,

$$M_D = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & B_{1,1} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & B_{N_f,N_f} \end{pmatrix}, \quad M_O = \begin{pmatrix} 0 & B_{0,1} & \dots & B_{0,N_f} \\ 0 & 0 & \dots & B_{1,N_f} \\ 0 & \vdots & \ddots & \vdots \\ 0 & B_{N_f,1} & \dots & 0 \end{pmatrix}.$$

As before, we will obtain bounds for $\|M_D^{-1}\|$ and $\|M_O\|$ and then use them to bound $\|M\|$ through Lemma 3.2.6.

In order to obtain bounds for $\|M_D\|$, we first state the following

Definition 3.2.3 *We will denote by $\mathcal{M}$, $\mathcal{M}_D$, $\mathcal{M}_O$ and $\mathcal{B}_{i,l}$ the equivalents of M, M_D, M_O and $B_{i,l}$, respectively, but replacing $\overline{c}_\nu^{n_h}$, $\widetilde{c}_\nu^{n_h}$, etc. by $\overline{\mathcal{C}}_\nu^{n_h}$, $\widetilde{\mathcal{C}}_\nu^{n_h}$, etc. That is, by replacing the DFT by the TCFT. We will also denote by $\mathfrak{M}$, $\mathfrak{M}_D$, $\mathfrak{M}_O$ and $\mathfrak{B}_{i,l}$ the equivalents of M, M_D, M_O and $B_{i,l}$, respectively, but replacing $\overline{c}_\nu^{n_h}$, $\widetilde{c}_\nu^{n_h}$, etc. by $\overline{\mathfrak{c}}_\nu^{n_h}$, $\widetilde{\mathfrak{c}}_\nu^{n_h}$, etc. For instance,*

$$\mathfrak{B}_{i,i} = \begin{pmatrix} A_i^c \partial\overline{\mathfrak{c}}_{\nu_i,N}^{n_h}(j_i) + A_i^s \partial\widetilde{\mathfrak{c}}_{\nu_i,N}^{n_h}(j_i) & \overline{\mathfrak{c}}_{\nu_i,N}^{n_h}(j_i) & \widetilde{\mathfrak{c}}_{\nu_i,N}^{n_h}(j_i) \\ A_i^c \partial\overline{\mathfrak{s}}_{\nu_i,N}^{n_h}(j_i) + A_i^s \partial\widetilde{\mathfrak{s}}_{\nu_i,N}^{n_h}(j_i) & \overline{\mathfrak{s}}_{\nu_i,N}^{n_h}(j_i) & \widetilde{\mathfrak{s}}_{\nu_i,N}^{n_h}(j_i) \\ A_i^c \partial\overline{\mathfrak{cs}}_{\nu_i,N}^{n_h}(j_i^+) + A_i^s \partial\widetilde{\mathfrak{cs}}_{\nu_i,N}^{n_h}(j_i^+) & \overline{\mathfrak{cs}}_{\nu_i,N}^{n_h}(j_i^+) & \widetilde{\mathfrak{cs}}_{\nu_i,N}^{n_h}(j_i^+) \end{pmatrix},$$

where $\mathfrak{cs}$ denotes either $\mathfrak{c}$ or $\mathfrak{s}$.

In order to invert $\mathfrak{M}_D$, we only have to invert a block $\mathfrak{B}_{i,i}$. The possibility to do that is established by the following

Lemma 3.2.9 *If $(A_i^s, A_i^c) \neq (0,0)$, $\mathfrak{B}_{i,i}$ is invertible either setting $\mathfrak{cs} = \mathfrak{c}$ or $\mathfrak{cs} = \mathfrak{s}$.*

Proof: Consider the matrix

$$\mathfrak{A} = \begin{pmatrix} \partial\overline{\mathfrak{c}}_{\nu_i}^{n_h}(j_i) & \partial\widetilde{\mathfrak{c}}_{\nu_i}^{n_h}(j_i) & \overline{\mathfrak{c}}_{\nu_i}^{n_h}(j_i) & \widetilde{\mathfrak{c}}_{\nu_i}^{n_h}(j_i) \\ \partial\overline{\mathfrak{s}}_{\nu_i}^{n_h}(j_i) & \partial\widetilde{\mathfrak{s}}_{\nu_i}^{n_h}(j_i) & \overline{\mathfrak{s}}_{\nu_i}^{n_h}(j_i) & \widetilde{\mathfrak{s}}_{\nu_i}^{n_h}(j_i) \\ \partial\overline{\mathfrak{c}}_{\nu_i}^{n_h}(j_i^+) & \partial\widetilde{\mathfrak{c}}_{\nu_i}^{n_h}(j_i^+) & \overline{\mathfrak{c}}_{\nu_i}^{n_h}(j_i^+) & \widetilde{\mathfrak{c}}_{\nu_i}^{n_h}(j_i^+) \\ \partial\overline{\mathfrak{s}}_{\nu_i}^{n_h}(j_i^+) & \partial\widetilde{\mathfrak{s}}_{\nu_i}^{n_h}(j_i^+) & \overline{\mathfrak{s}}_{\nu_i}^{n_h}(j_i^+) & \widetilde{\mathfrak{s}}_{\nu_i}^{n_h}(j_i^+) \end{pmatrix},$$

and denote by $\mathfrak{A}_{l_1,l_2,l_3}^{i_1,i_2,i_3}$ the submatrix of $\mathfrak{A}$ obtained by selecting the rows i_1, i_2, i_3 and the columns l_1, l_2, l_3. Then, the determinant of a block $\mathfrak{B}_{i,i}$ is

$$\det \mathfrak{B}_{i,i} = \begin{cases} A_i^c \det \mathfrak{A}_{1,3,4}^{1,2,3} + A_i^s \det \mathfrak{A}_{2,3,4}^{1,2,3} & \text{if we set } \mathfrak{cs} = \mathfrak{c}, \\ A_i^c \det \mathfrak{A}_{1,3,4}^{1,2,4} + A_i^s \det \mathfrak{A}_{2,3,4}^{1,2,4} & \text{if we set } \mathfrak{cs} = \mathfrak{s}. \end{cases}$$

To see that there exists a choice of $\mathfrak{cs}$ that makes $\det \mathfrak{B}_{i,i} \neq 0$ is equivalent to see that the system

$$\begin{cases} A_i^c \det \mathfrak{A}_{1,3,4}^{1,2,3} + A_i^s \det \mathfrak{A}_{2,3,4}^{1,2,3} &= 0 \\ A_i^c \det \mathfrak{A}_{1,3,4}^{1,2,4} + A_i^s \det \mathfrak{A}_{2,3,4}^{1,2,4} &= 0 \end{cases},$$

with unknowns A_i^c, A_i^s, has unique solution. That is, that the determinant

$$\det \mathfrak{D} = \begin{vmatrix} \det \mathfrak{A}_{1,3,4}^{1,2,3} & \det \mathfrak{A}_{2,3,4}^{1,2,3} \\ \det \mathfrak{A}_{1,3,4}^{1,2,4} & \det \mathfrak{A}_{2,3,4}^{1,2,4} \end{vmatrix} \tag{3.13}$$

is different from zero. From Definition 3.2.1, and since $\nu_i - j_i^+ = \nu_i - j_i - \operatorname{sign}(\nu_i - j_i)$, this determinant only depends on the difference $\varepsilon = \nu_i - j_i$ which, by definition, ranges from $-1/2$ to $1/2$. In order to prove the lemma, we only need to see that the previous determinant is different from zero in this range (see Fig. 3.3 for the numerical evidence).

In order to simplify the notation, we denote

$$\mathfrak{A} = \begin{pmatrix} a_1 & a_2 & a_3 & a_4 \\ -a_2 & a_1 & -a_4 & a_3 \\ a_5 & a_6 & a_7 & a_8 \\ -a_6 & a_5 & -a_8 & a_7 \end{pmatrix}$$

and

$$A = \mathfrak{A}^{1,2,3}_{1,3,4}, \quad B = \mathfrak{A}^{1,2,4}_{1,3,4}, \quad C = \mathfrak{A}^{1,2,3}_{2,3,4}, \quad D = \mathfrak{A}^{1,2,4}_{2,3,4},$$

so that $\det \mathfrak{D} = \det A \det D - \det B \det C$.

First note that, using $-(\varepsilon - \mathrm{sign}(\varepsilon)) = -\varepsilon - \mathrm{sign}(-\varepsilon)$ and the fact that $\overline{\mathfrak{c}}^{n_h}_\nu(j)$ and $\widetilde{\mathfrak{c}}^{n_h}_\nu(j)$ are even and odd in ε respectively, we can check that $\det \mathfrak{D}$ is even in ε and therefore we can restrict to $[0, 1/2]$ the range of ε to be considered.

Let $0 < \varepsilon < 1/2$ and assume $\det \mathfrak{D} = 0$. We note that $\det A = \det D$ and $\det B = -\det C$, so that $\det \mathfrak{D} = (\det A)^2 + (\det B)^2$ and we have $\det A = \det B = 0$. Expanding through the first column, we get

$$\begin{aligned} \det A &= -a_1(a_3a_7 + a_4a_8) + a_2(a_3a_8 - a_4a_7) + a_5(a_3^2 + a_4^2) \\ \det B &= a_2(a_3a_7 + a_4a_8) + a_1(a_3a_8 - a_4a_7) - a_6(a_3^2 + a_4^2) \end{aligned}$$

The $a_3a_8 - a_4a_7$ term is readily checked to be zero. We denote $\psi = \psi_{n_h}(\varepsilon)$ and $\psi_m = \psi_{n_h}(\varepsilon - 1)$. We check that $a_3a_7 + a_4a_8$ has the same numerator as $a_3^2 + a_4^2$, due to the 1–periodicity in ε of the numerators of $a_1, \ldots, a_8$. The denominators are different: $\psi\psi_m$ for $a_3a_7 + a_4a_8$ and ψ^2 for $a_3^2 + a_4^2$. Setting $\det A = \det B = 0$ and simplifying numerators, we get

$$a_1\psi = a_5\psi_m, \qquad a_2\psi = a_6\psi_m.$$

Now, using the expressions for a_1 and a_5 from Definition 3.2.1, as well as $a_1\psi = a_5\psi_m$, we obtain $r_{n_h}(\varepsilon) = r_{n_h}(\varepsilon - 1)$, that is $\psi'/\psi = \psi'_m\psi_m$ (here $'$ denotes derivative), and therefore

$$\frac{d}{d\varepsilon}\frac{\psi_m}{\psi} = \frac{\psi'_m\psi - \psi_m\psi'}{\psi_m^2} = 0. \tag{3.14}$$

The condition $a_2\psi = a_6\psi_m$ leads to the same conclusion.

But

$$\frac{\psi_m}{\psi} = \frac{\varepsilon - n_h - 1}{\varepsilon + n_h} = 1 - \frac{2n_h + 1}{\varepsilon + n_h},$$

and its derivative with respect to ε is different from zero for $0 < \varepsilon \leq \frac{1}{2}$, which is in contradiction with (3.14).

For $\varepsilon = 0$, $\det \mathfrak{D}$ is checked to be different from zero using the expressions of Definition 3.2.1 (it is necessary to compute the limits when $\varepsilon \to 0$). □

Now that we know that a block $\mathfrak{B}_{i,i}$ is invertible, in order to actually invert it we state the following

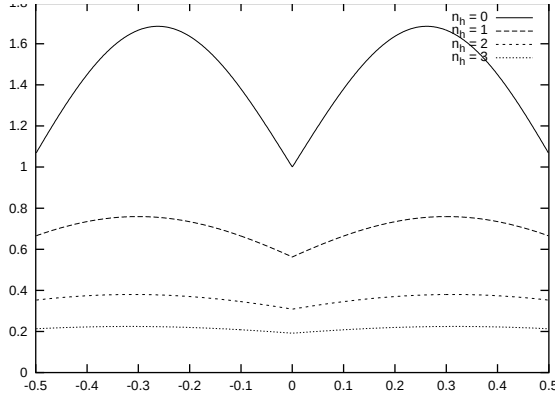


Figure 3.3: Plot of the determinant (3.13) for $-1/2 \leq \nu_i - j_i \leq 1/2$ and $n_h = 0, 1, 2, 3$.

n_h	0	1	2	3
G_{n_h}	4.84	8.83	13.3	17.7

Table 3.1: Some values of the G_{n_h} constants.

Definition 3.2.4 *For $n_h \in \mathbb{N}$, we define G_{n_h} to be an upper bound of*

$$\max_{\substack{\theta\in[0,2\pi]\\ |\nu-j|\leq\frac{1}{2}}} \min_{\mathfrak{cs}\in\{\mathfrak{c},\mathfrak{s}\}} \left\| \begin{pmatrix} (\cos\theta)\partial\overline{\mathfrak{c}}^{n_h}_{\nu,N}(j) + (\sin\theta)\partial\widetilde{\mathfrak{c}}^{n_h}_{\nu,N}(j) & \overline{\mathfrak{c}}^{n_h}_{\nu,N}(j) & \widetilde{\mathfrak{c}}^{n_h}_{\nu,N}(j) \\ (\cos\theta)\partial\overline{\mathfrak{s}}^{n_h}_{\nu,N}(j) + (\sin\theta)\partial\widetilde{\mathfrak{s}}^{n_h}_{\nu,N}(j) & \overline{\mathfrak{s}}^{n_h}_{\nu,N}(j) & \widetilde{\mathfrak{s}}^{n_h}_{\nu,N}(j) \\ (\cos\theta)\partial\overline{\mathfrak{cs}}^{n_h}_{\nu,N}(j^+) + (\sin\theta)\partial\widetilde{\mathfrak{cs}}^{n_h}_{\nu,N}(j^+) & \overline{\mathfrak{cs}}^{n_h}_{\nu,N}(j^+) & \widetilde{\mathfrak{cs}}^{n_h}_{\nu,N}(j^+) \end{pmatrix}^{-1} \right\|.$$

In table 3.1 we give some values of the G_{n_h} constants found numerically. Just for illustration purposes, in figure 3.4 we display the behavior of

$$\min_{\mathfrak{cs}\in\{\mathfrak{c},\mathfrak{s}\}} \left\| \begin{pmatrix} (\cos\theta)\partial\overline{\mathfrak{c}}^{n_h}_{\nu,N}(j) + (\sin\theta)\partial\widetilde{\mathfrak{c}}^{n_h}_{\nu,N}(j) & \overline{\mathfrak{c}}^{n_h}_{\nu,N}(j) & \widetilde{\mathfrak{c}}^{n_h}_{\nu,N}(j) \\ (\cos\theta)\partial\overline{\mathfrak{s}}^{n_h}_{\nu,N}(j) + (\sin\theta)\partial\widetilde{\mathfrak{s}}^{n_h}_{\nu,N}(j) & \overline{\mathfrak{s}}^{n_h}_{\nu,N}(j) & \widetilde{\mathfrak{s}}^{n_h}_{\nu,N}(j) \\ (\cos\theta)\partial\overline{\mathfrak{cs}}^{n_h}_{\nu,N}(j^+) + (\sin\theta)\partial\widetilde{\mathfrak{cs}}^{n_h}_{\nu,N}(j^+) & \overline{\mathfrak{cs}}^{n_h}_{\nu,N}(j^+) & \widetilde{\mathfrak{cs}}^{n_h}_{\nu,N}(j^+) \end{pmatrix}^{-1} \right\| \qquad (3.15)$$

in terms of θ and $\nu - j$.

In order to relate the bound of the previous definition to the bound of an actual block $\mathfrak{B}_{i,i}$, we will use the following

Lemma 3.2.10 *Let $\lambda \neq 0$ be a real number and v_1, v_2, v_3 3–dimensional row vectors. Then*

$$\left\| \begin{pmatrix} \lambda v_1 & v_2 & v_3 \end{pmatrix}^{-1} \right\|_\infty \leq \max(\frac{1}{\lambda}, 1) \left\| \begin{pmatrix} v_1 & v_2 & v_3 \end{pmatrix}^{-1} \right\|_\infty.$$

Proof: Define w_1, w_2, w_3 according to

$$\begin{pmatrix} v_1 & v_2 & v_3 \end{pmatrix}^{-1} = \frac{1}{\det(v_1, v_2, v_3)} \begin{pmatrix} w_1^\top \\ w_2^\top \\ w_3^\top \end{pmatrix}.$$

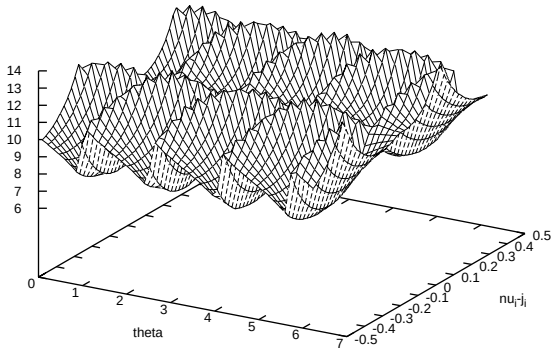


Figure 3.4: Plot of (3.15) for $0 \leq \theta \leq 2\pi$ and $-1/2 \leq \nu_i - j_i \leq 1/2$.

Then,

$$\begin{pmatrix} \lambda v_1 & v_2 & v_3 \end{pmatrix}^{-1} = \frac{1}{\det(\lambda v_1, v_2, v_3)} \begin{pmatrix} w_1^\top \\ \lambda w_2^\top \\ \lambda w_3^\top \end{pmatrix} = \frac{1}{\det(v_1, v_2, v_3)} \begin{pmatrix} w_1^\top/\lambda \\ w_2^\top \\ w_3^\top \end{pmatrix},$$

and therefore,

$$\begin{aligned} \left\| \begin{pmatrix} \lambda v_1 & v_2 & v_3 \end{pmatrix}^{-1} \right\|_\infty &= \frac{1}{|\det(v_1, v_2, v_3)|} \left\| \begin{pmatrix} w_1^\top/\lambda \\ w_2^\top \\ w_3^\top \end{pmatrix} \right\|_\infty \\ &= \frac{\max(\frac{1}{\lambda}\|w_1\|_1, \|w_2\|_1, \|w_3\|_1)}{|\det(v_1, v_2, v_3)|} \\ &\leq \max(\frac{1}{\lambda}, 1) \frac{\max(\|w_1\|_1, \|w_2\|_1, \|w_3\|_1)}{|\det(v_1, v_2, v_3)|} \\ &= \max(\frac{1}{\lambda}, 1) \left\| \begin{pmatrix} v_1 & v_2 & v_3 \end{pmatrix}^{-1} \right\|_\infty. \end{aligned}$$

□

Let us denote $A_i = ((A_i^c)^2 + (A_i^s)^2)^{1/2}$. Using Definition 3.2.4 and Lemma 3.2.10 we have

$$\|(\mathfrak{B}_{i,i})^{-1}\| \leq \max(A_i^{-1}, 1) G_{n_h},$$

and therefore,

$$\|\mathfrak{M}_D^{-1}\| \leq \max(A_{min}^{-1}, 1) G_{n_h},$$

where $A_{min} = \min\{A_1, \ldots, A_{N_f}\}$.

Now we bound the simplified off–diagonal part of M.

Lemma 3.2.11 *Assume* $\frac{TD}{(2r_0-2)^\tau} > 3 + n_h$. *Then,*

$$\|\mathfrak{M}_O\| \leq \frac{(n_h!)^2\Big[\sqrt{2}(\sum_{l=1}^{N_f} A_l)\Big(\pi + \ln\big(\frac{TD}{(2r_0-2)^\tau} - 1 + n_h\big) - \ln\big(\frac{TD}{(2r_0-2)^\tau} - 2 - n_h\big)\Big) + 2N_f\Big]}{\pi\big(\frac{TD}{(2r_0-2)^\tau} - 1 - n_h\big)^{1+2n_h}}$$

Proof: We first note that, from Lemma 3.2.3 and Definition 3.2.1,

$$\begin{aligned}
|\overline{\mathfrak{c}}_\nu^{n_h}(j)|, |\widetilde{\mathfrak{c}}_\nu^{n_h}(j)|, |\overline{\mathfrak{s}}_\nu^{n_h}(j)|, |\widetilde{\mathfrak{s}}_\nu^{n_h}(j)| &\leq \frac{(n_h!)^2}{\pi(|\nu-j| - n_h)^{1+2n_h}}, \\
|\partial\overline{\mathfrak{c}}_\nu^{n_h}(j)|, |\partial\widetilde{\mathfrak{c}}_\nu^{n_h}(j)|, |\partial\overline{\mathfrak{s}}_\nu^{n_h}(j)|, |\partial\widetilde{\mathfrak{s}}_\nu^{n_h}(j)| &\leq \\
&\frac{(n_h!)^2(\pi + \ln(|\nu-j| + n_h) - \ln(|\nu-j| - n_h - 1))}{\pi(|\nu-j| - n_h)^{1+2n_h}}.
\end{aligned}$$

Therefore, using $|A_l^c| + |A_l^s| \leq \sqrt{2}((A_l^c)^2 + (A_l^s)^2)^{1/2} = \sqrt{2}A_l$,

$$\begin{aligned}
\|\mathfrak{B}_{i,l}(j)\| &\leq \max_{\substack{j=j_i,j_i^+ \\ \mathfrak{cs}=\mathfrak{c},\mathfrak{s}}} \Big(|A_i^c||\partial\widetilde{\mathfrak{cs}}_{\nu_i}^{n_h}(j)| + |A_i^s||\partial\widetilde{\mathfrak{cs}}_{\nu_i}^{n_h}(j)| + |\widetilde{\mathfrak{cs}}_{\nu_i}^{n_h}(j)| + |\widetilde{\mathfrak{cs}}_{\nu_i}^{n_h}(j)|\Big) \\
&\leq \max_{j=j_i,j_i^+} \frac{(n_h!)^2(\sqrt{2}A_l(\pi + \ln(|\nu_l-j| + n_h) - \ln(|\nu_l-j| - n_h - 1)) + 2)}{\pi(|\nu_l-j| - n_h)^{1+2n_h}}.
\end{aligned}$$

Now, for $i, l = 1 \div N_f$ and $j = j_i, j_i^+$ there exists i_j such that $j \in \{[\nu_{i_j}], [\nu_{i_j}] + 1\}$, so $|\nu_{i_j} - j| \leq 1$. As stated at the beginning of this section, we also have that there exists k_{i_j}, with $|k_{i_j}| \leq r_0 - 1$, such that $\nu_{i_j} = Tk_{i_j}\omega$. Then, using the Diophantine condition (3.2), we obtain

$$\begin{aligned}
|\nu_l - j| &\geq |\nu_l - \nu_{i_j}| - |j - \nu_{i_j}| \geq T|(k_l - k_{i_j})\omega| - 1 \geq \frac{TD}{|k_l - k_{i_j}|^\tau} - 1 \\
&\geq \frac{TD}{(2r_0-2)^\tau} - 1,
\end{aligned}$$

and the lemma follows from

$$\|\mathfrak{M}_O\| \leq \max_{i=0\div N_f} \sum_{l=1}^{N_f} \|\mathfrak{B}_{i,l}\|,$$

where we denote $j_0 = 0$. □

In order to bound $\|M^{-1}\|$ from $\|M_D^{-1}\|$ and $\|M_O^{-1}\|$ by applying Lemma 3.2.6, we need the bounds of $\|\mathcal{M}_D - \mathfrak{M}_D\|$, $\|M_D - \mathcal{M}_D\|$, $\|\mathcal{M}_O - \mathfrak{M}_O\|$ and $\|M_O - \mathcal{M}_O\|$. We calculate them in the following lemmas.

Lemma 3.2.12 *Denote* $\nu_{min} = \min\{\nu_1, \ldots, \nu_{N_f}\}$, $A_{max} = \max\{A_0, \ldots, A_{N_f}\}$. *If* $[\nu_{min}] \geq 2 + n_h$, *we have*

$$\|\mathfrak{M}_D - \mathcal{M}_D\| \leq \frac{(n_h!)^2\Big(\sqrt{2}A_{max}[\pi + \ln(2[\nu_{min}] + n_h) - \ln(2[\nu_{min}] - n_h - 1)] + 2\Big)}{\pi(2[\nu_{min}] - n_h)^{1+2n_h}},$$

$$\|\mathfrak{M}_O - \mathcal{M}_O\| \leq \frac{(n_h!)^2\Big(\sqrt{2}(\sum_{l=1}^{N_f} A_l)(\pi + \ln([\nu_{min}]+n_h) - \ln([\nu_{min}]-n_h-1)) + 2N_f\Big)}{\pi([\nu_{min}] - n_h)^{1+2n_h}},$$

Proof: From Definition 3.2.3 and Lemma 3.2.5,

$$\begin{aligned}\|\mathfrak{B}_{i,l} - \mathcal{B}_{i,l}\| &\leq \max_{j=j_i,j_i^+} \left(\begin{array}{c} |A_l^c||\partial\overline{\mathfrak{cs}}_{\nu_l}^{n_h}(j) - \partial\overline{\mathcal{CS}}_{\nu_l}^{n_h}(j)| + |A_l^s||\partial\widetilde{\mathfrak{cs}}_{\nu_l}^{n_h}(j) - \partial\widetilde{\mathcal{CS}}_{\nu_l}^{n_h}(j)|+ \\ |\overline{\mathfrak{cs}}_{\nu_l}^{n_h}(j) - \overline{\mathcal{CS}}_{\nu_l}^{n_h}(j)| + |\widetilde{\mathfrak{cs}}_{\nu_l}^{n_h}(j) - \widetilde{\mathcal{CS}}_{\nu_l}^{n_h}(j)| \end{array} \right) \\ &\leq \frac{(n_h!)^2\Big(\sqrt{2}A_l(\ln(|-\nu_l-j|+n_h) - \ln(|-\nu_l-j|-n_h-1)) + 2\Big)}{\pi(|-\nu_l-j|-n_h)^{1+2n_h}},\end{aligned}$$

where either $\mathfrak{cs} = \mathfrak{c}$ and $\mathcal{CS} = \mathcal{C}$ or $\mathfrak{cs} = \mathfrak{s}$ and $\mathcal{CS} = \mathcal{S}$. For the second inequality we have used that

$$|A_l^c| + |A_l^s| \leq \sqrt{2}((A_l^c)^2 + (A_l^s)^2)^{1/2} = \sqrt{2}A_l.$$

Now, the first inequality of the lemma follows from

$$\|\mathfrak{M}_D - \mathcal{M}_D\| \leq \max_{i=1\div N_f} \|\mathfrak{B}_{i,i} - \mathcal{B}_{i,i}\|$$

and the fact that $|-\nu_i - j_i|, |-\nu_i - j_i^+| \geq 2[\nu_{min}]$ for $i = 1 \div N_f$. The second inequality follows from

$$\|\mathfrak{M}_O - \mathcal{M}_O\| \leq \max_{i=0\div N_f} \sum_{\substack{l=1 \\ l\neq i}}^{N_f} \|\mathfrak{B}_{i,l} - \mathcal{B}_{i,l}\|$$

and the fact that $|-\nu_l - j_i|, |-\nu_l - j_i^+| \geq [\nu_{min}]$ for $i = 0 \div N_f, l = 1 \div N_f$ (we denote $j_0 = j_0^+ = 0$). □

Lemma 3.2.13 *Assume $N - T(2r_0 - 2)\|\omega\|_\infty > 3 + n_h$. Then,*

$$\begin{aligned}&\|\mathcal{M}_D - M_D\| \\ &\qquad \leq \frac{4(n_h!)^2\Big(\sqrt{2}A_{max}(\pi + \ln(N-\Omega+n_h) - \ln(N-\Omega-1-n_h)) + 2\Big)(1+\frac{1}{2n_h})}{\pi(N-\Omega-n_h)^{1+2n_h}}, \\ &\|\mathcal{M}_O - M_O\| \\ &\qquad \leq \frac{4(n_h!)^2\Big(\sqrt{2}(\sum_{l=1}^{N_f} A_l)(\pi + \ln(N-\Omega+n_h) - \ln(N-\Omega-1-n_h)) + 2N_f\Big)(1+\frac{1}{2n_h})}{\pi(N-\Omega-n_h)^{1+2n_h}}.\end{aligned}$$

being $\Omega = T(2r_0 - 2)\|\omega\|_\infty + 1$.

Proof: For a 3×3 block, we apply Lemma 3.2.4 and obtain

$$\begin{aligned}
&\|\mathfrak{B}_{i,l} - \mathcal{B}_{i,l}\| \\
&\quad \leq \max_{j=j_i,j_i^+} \left(\begin{array}{c} |A_i^c||\partial\overline{\mathcal{CS}}_{\nu_l}^{n_h}(j) - \partial\overline{cs}_{\nu_l}^{n_h}(j)| + |A_i^s||\partial\widetilde{\mathcal{CS}}_{\nu_l}^{n_h}(j) - \partial\widetilde{cs}_{\nu_l}^{n_h}(j)| + \\ |\overline{\mathcal{CS}}_{\nu_l}^{n_h}(j) - \overline{cs}_{\nu_l}^{n_h}(j)| + |\widetilde{\mathcal{CS}}_{\nu_l}^{n_h}(j) - \widetilde{cs}_{\nu_l}^{n_h}(j)|, \end{array} \right) \\
&\quad \leq \frac{4(n_h!)^2\Big(\sqrt{2}A_i(\pi + \ln(N-j-\nu_l+n_h) - \ln(N-j-\nu_l-n_h-1)) + 2\Big)(1+\frac{1}{2n_h})}{\pi(N-j-\nu_l-n_h)^{1+2n_h}},
\end{aligned}$$

where either $\mathfrak{cs} = \mathfrak{c}$ and $\mathcal{CS} = \mathcal{C}$ or $\mathfrak{cs} = \mathfrak{s}$ and $\mathcal{CS} = \mathcal{S}$. As $j \in \{j_i, j_i^+\}$ and $\nu_l = Tk_l\omega$ with $1 \leq |k_l| \leq r_0 - 1$, we have that $j + \nu_l \leq Tk_i\omega + 1 + Tk_l\omega \leq T(2r_0-2)\|\omega\|_\infty + 1 = \Omega$ for some $|k_i| \leq r_0 - 1$. Using this, the lemma follows from

$$\begin{aligned}
\|M_D - \mathcal{M}_D\| &\leq \max_{i=1\div N_f} \|B_{i,i} - \mathcal{B}_{i,i}\|, \\
\|M_O - \mathcal{M}_O\| &\leq \max_{i=0\div N_f} \sum_{l=1}^{N_f} \|B_{i,l} - \mathcal{B}_{i,l}\|,
\end{aligned}$$

where we denote $j_0 = j_0^+ = 0$. □

From these lemmas follows a bound for $\|Dg(y)^{-1}\|$, as will be stated in theorem 3.4.1.

3.3 Error bounds for $\|\Delta b\|_\infty$

We give first three definitions and one lemma in order to be able to bound finite sums of the type $\sum_{j=r_0}^{r_1} j^\alpha e^{-\delta j}$, with r_1 either finite or infinite.

Definition 3.3.1 *Given $z \in \mathbb{R}$, we define*

$$\forall x \in \mathbb{R}, \quad [x]_z = \max\{z + n : n \in \mathbb{Z}, z + n \leq x\} = z + [x - z],$$

where [ ] denotes integer part.

Note that, for all $m \in \mathbb{Z}$, we have $[x]_z = [x]_{z+m}$.

In what follows, we will use the incomplete Gamma functions $\gamma(\alpha, x)$ and $\Gamma(\alpha, x)$, which are defined as (see, for instance, [1])

$$\gamma(\alpha, x) = \int_0^x e^{-t}t^{\alpha-1}dt, \qquad \Gamma(\alpha, x) = \int_x^\infty e^{-t}t^{\alpha-1}dt.$$

In order to be able to bound sums by integrals taking into account the intervals of monotonicity is convenient to introduce the following

Definition 3.3.2 *For $j_1, j_2, \alpha, \delta > 0$ we define the functions*

$$\begin{aligned} G_f(j_1,j_2,\alpha,\delta) &= \frac{1}{\delta^{\alpha+1}}\chi_{\{j_1\le\frac{\alpha}{\delta}-1\}}\Big(\gamma(\alpha+1,\delta\min([\frac{\alpha}{\delta}]_{j_1},j_2+1))-\gamma(\alpha+1,\delta j_1)\Big)+ \\ &\quad \chi_{\{j_1\le\frac{\alpha}{\delta},j_2>\frac{\alpha}{\delta}-1\}}([\frac{\alpha}{\delta}]_{j_1})^\alpha e^{-\delta[\frac{\alpha}{\delta}]_{j_1}}+ \\ &\quad \chi_{\{j_1<\frac{\alpha}{\delta}+1,j_2>\frac{\alpha}{\delta}\}}([\frac{\alpha}{\delta}]_{j_1}+1)^\alpha e^{-\delta([\frac{\alpha}{\delta}]_{j_1}+1)}+ \\ &\quad \frac{1}{\delta^{\alpha+1}}\chi_{\{j_2>\frac{\alpha}{\delta}+1\}}\Big(\gamma(\alpha+1,\delta j_2)-\gamma(\alpha+1,\max([\frac{\alpha}{\delta}]_{j_1}+1,j_1-1))\Big), \end{aligned}$$

and

$$\begin{aligned} G_\infty(j_1,\alpha,\delta) &= \frac{1}{\delta^{\alpha+1}}\chi_{\{j_1\le\frac{\alpha}{\delta}-1\}}\Big(\gamma(\alpha+1,\delta[\frac{\alpha}{\delta}]_{j_1})-\gamma(\alpha+1,\delta j_1)\Big)+ \\ &\quad \chi_{\{j_1\le\frac{\alpha}{\delta}\}}([\frac{\alpha}{\delta}]_{j_1})^\alpha e^{-\delta[\frac{\alpha}{\delta}]_{j_1}}+ \\ &\quad \chi_{\{j_1<\frac{\alpha}{\delta}+1\}}([\frac{\alpha}{\delta}]_{j_1}+1)^\alpha e^{-\delta([\frac{\alpha}{\delta}]_{j_1}+1)}+ \\ &\quad \frac{1}{\delta^{\alpha+1}}\Gamma(\alpha+1,\delta\max([\frac{\alpha}{\delta}]_{j_1}+1,j_1-1)). \end{aligned}$$

In the above formulas, $\chi_{\{condition\}}$ equals 1 if condition is true and 0 otherwise.

Lemma 3.3.1 *The functions G_f and G_∞ satisfy*

$$\sum_{j=j_1}^{j_2} j^\alpha e^{-\delta j} \le G_f(j_1,j_2,\alpha,\delta), \qquad \sum_{j=j_1}^{\infty} j^\alpha e^{-\delta j} \le G_\infty(j_1,\alpha,\delta).$$

Proof: To obtain the expressions for G_f, G_∞ in Definition 3.3.2 we have bounded the previous sums by integrals. This can be done easily for the subintervals of j of length 1, starting at j_0, for which the function $j^\alpha e^{-\delta j}$ is monotone. Some care must be taken for the intervals around the maximum of the function, which is attained at $j = \frac{\alpha}{\delta}$. This is the reason for the definition 3.3.1. Both inequalities follow after a careful examination of all the possibilities for the relative position between $[j_1, j_2]$ and the maximum $\frac{\alpha}{\delta}$. □

We recall from (3.5) that Δb is defined as,

$$\Delta b = \begin{pmatrix} c^{n_h}_{f-p,T,N}(0) \\ c^{n_h}_{f-p,T,N}(j_i) \\ s^{n_h}_{f-p,T,N}(j_i) \\ cs^{n_h}_{f-p,T,N}(j_i^+) \end{pmatrix},$$

where i ranges from 1 to N_f and cs denotes either c or s. We want to determine the trigonometric approximation $p(t)$ of $f(t)$ using frequencies up to order $r_0 - 1$, that is, $\{k\omega : |k| \le r_0 - 1\}$, so

$$f(t) - p(t) = \sum_{|k|=r_0}^{\infty} a_k e^{i2\pi k\omega t}.$$

Therefore, denoting by J the set of indices $\{0, j_i, j_i^+ : i = 1 \div N_f\}$, we have

$$\begin{aligned} \|\Delta b\| &\leq 2 \max_{j\in J} |F^{n_h}_{f-p,T,N}(j)| \\ &\leq 2 \max_{j\in J} \sum_{|k|\geq r_0} |a_k||F^{n_h}_{e^{i2\pi k\omega t},T,N}(j)| \\ &\leq 2 \max_{j\in J} \sum_{|k|=r_0}^{r_*} |a_k||F^{n_h}_{e^{i2\pi k\omega t},T,N}(j)| + 2 \sum_{|k|=r_*+1}^{\infty} |a_k|. \end{aligned} \tag{3.16}$$

We will keep r_* as an unknown quantity for the moment, and bound the first term of the above sum but replacing the DFT by the TCFT.

Lemma 3.3.2 *The following inequality is fulfilled:*

$$\#\{k : |k| = j\} \leq \frac{2^m}{(m-1)!}\Big(j + \frac{m}{2}\Big)^{m-1}.$$

Proof: See [17], p. 114. □

Lemma 3.3.3 *If* $\frac{TD}{(r_*+r_0-2)^\tau} > 1 + n_h$, *we have*

$$\sum_{|k|=r_0}^{r_*-1} |a_k||\phi^{n_h}_{e^{i2\pi k\omega t},T}(\frac{j}{T})| \leq$$

$$\frac{2^m C(n_h!)^2 e^{\delta(r_0-1)} \sum_{l=0}^{m-1} \binom{m-1}{l} (\frac{m}{2}-r_0+1)^{m-1-l} G_f(2r_0-1, r_*+r_0-2, l+\tau(1+2n_h), \delta)}{E_*(m-1)!\pi(TD)^{1+2n_h}}$$

where

$$E_* = \frac{(z_* - 1 - n_h)^{1+2n_h}}{z_*^{1+2n_h}}, \quad z_* = \frac{TD}{(r_* + r_0 - 1)^\tau}.$$

Proof: Using the Cauchy estimates and (1.5),

$$\begin{aligned} \sum_{|k|=r_0}^{r_*-1} |a_k||\phi^{n_h}_{e^{i2\pi k\omega t},T}(\frac{j}{T})| &\leq C \sum_{|k|=r_0}^{r_*-1} e^{-\delta|k|} \frac{(n_h!)^2}{\pi\psi_{n_h}(|Tk\omega - j|)} \\ &\leq \frac{C(n_h!)^2}{\pi} \sum_{|k|=r_0}^{r_*-1} \frac{e^{-\delta|k|}}{(|Tk\omega - j| - n_h)^{1+2n_h}} \qquad (3.17) \\ &\leq \frac{C(n_h!)^2}{\pi} \sum_{|k|=r_0}^{r_*-1} \frac{e^{-\delta|k|}}{(\frac{TD}{(|k|+r_0-1)^\tau} - 1 - n_h)^{1+2n_h}} \qquad (3.18) \end{aligned}$$

For the last step we have used that, since $j \in \{j_i, j_i^+\}$ for some $i = 1 \div N_f$, there exists $k_j \in \mathbb{Z}^m$ such that $|j - Tk_j\omega| \leq 1$, so

$$\begin{aligned} |Tk\omega - j| &\geq |Tk\omega - Tk_j\omega| - |j - Tk_j\omega| \geq T|(k-k_j)\omega| - 1 \geq \frac{TD}{|k-k_j|^\tau} - 1 \\ &\geq \frac{TD}{(|k| + r_0 - 1)^\tau} - 1. \end{aligned}$$

In order to be able to sum the above series with the aid of the incomplete Gamma functions, we choose E_* such that $(x - 1 - n_h)^{1+2n_h} \geq E_* x^{1+2n_h}$ for $x \in \{\frac{TD}{(|k|+r_0-1)^\tau}\}_{|k|=1}^{r_*-1}$. This is accomplished setting $E_* = \frac{(z_*-1-n_h)^{1+2n_h}}{z_*^{1+2n_h}}$ with $z_* = \frac{TD}{(r_*+r_0-2)^\tau}$. Therefore,

$$\begin{aligned} \Phi := \sum_{|k|=r_0}^{r_*-1} |a_k||\phi^{n_h}_{e^{i2\pi k\omega t},T}(\frac{j}{T})| &\leq \frac{C(n_h!)^2}{E_*\pi} \sum_{|k|=r_0}^{r_*-1} \frac{e^{-\delta|k|}}{(\frac{TD}{(|k|+r_0-1)^\tau})^{1+2n_h}} \\ &= \frac{C(n_h!)^2}{E_*\pi(TD)^{1+2n_h}} \sum_{|k|=r_0}^{r_*-1} e^{-\delta|k|}(|k| + r_0 - 1)^{\tau(1+2n_h)}. \end{aligned}$$

Now we apply Lemma 3.3.2,

$$\Phi \leq \frac{C(n_h!)^2}{E_*\pi(TD)^{1+2n_h}} \sum_{j=r_0}^{r_*-1} \frac{2^m}{(m-1)!}(j + \frac{m}{2})^{m-1} e^{-\delta j}(j + r_0 - 1)^{\tau(1+2n_h)},$$

shift the index j,

$$\Phi \leq \frac{2^m C(n_h!)^2}{E_*(m-1)!\pi(TD)^{1+2n_h}} \sum_{j=2r_0-1}^{r_*+r_0-2} (j + \frac{m}{2} - r_0 + 1)^{m-1} e^{-\delta(j-r_0+1)} j^{\tau(1+2n_h)},$$

and expand by Netwon's binomial,

$$\begin{aligned} \Phi &\leq \frac{2^m C(n_h!)^2}{E_*(m-1)!\pi(TD)^{1+2n_h}} \sum_{l=0}^{m-1} \binom{m-1}{l} (\frac{m}{2} - r_0 + 1)^{m-1-l} \sum_{j=2r_0-1}^{r_*+r_0-2} j^{l+\tau(1+2n_h)} e^{-\delta(j-r_0+1)} \\ &= \frac{2^m C(n_h!)^2 e^{\delta(r_0-1)}}{E_*(m-1)!\pi(TD)^{1+2n_h}} \sum_{l=0}^{m-1} \binom{m-1}{l} (\frac{m}{2} - r_0 + 1)^{m-1-l} \sum_{j=2r_0-1}^{r_*+r_0-2} j^{l+\tau(1+2n_h)} e^{-\delta j}. \end{aligned}$$

Now, to show the lemma, we only have to apply Lemma 3.3.1. □

In the proof of the previous lemma, we bounded the continuous Fourier transform of a complex exponential term as

$$|\phi^{n_h}_{e^{i2\pi k\omega t},T}(\frac{j}{T})| \leq \frac{(n_h!)^2}{\pi(\frac{TD}{(|k|+r_0-2)^\tau} - 1 - n_h)^{1+2n_h}}.$$

Therefore, an intrinsic way to choose r_* is to take it equal to the last value of $|k|$ for which the previous bound is < 1. In addition to that, and in order to avoid an excessive

amplification of the bound due to the introduction of the E_* constant in the proof of the previous lemma, we will restrict r_* so that $z_* \geq 2(1+n_h)$ and

$$E_* \geq \frac{1}{2^{1+2n_h}}. \tag{3.19}$$

Therefore,

$$r_* = \left[\Big(\frac{TD}{\max((\frac{(n_h!)^2}{\pi})^{\frac{1}{1+2n_h}} + 1 + n_h, 2(1+n_h))} \Big)^{\frac{1}{\tau}} - r_0 + 2 \right].$$

Now that we have chosen r_*, we need to bound the error due to the approximation of the discrete Fourier transform by the continuous one. This is done in the two following lemmas.

Lemma 3.3.4 *If $N - T(r_* + r_0 - 2)\|\omega\|_\infty > 1 + n_h$, then*

$$\Big| \sum_{|k|=r_0}^{r_*-1} a_k F^{n_h}_{e^{i2\pi k\omega t},T,N}(\frac{j}{T}) - \sum_{|k|=r_0}^{r_*-1} a_k \phi^{n_h}_{e^{i2\pi k\omega t},T}(\frac{j}{T}) \Big| \leq$$
$$\frac{2^{m+1} C (n_h!)^2 (1+\frac{1}{2n_h}) e^{\delta \frac{m}{2}} G_f(r_0 + \frac{m}{2}, r_* - 1 + \frac{m}{2}, m-1, \delta)}{\pi (m-1)! (N - T(r_* + r_0 - 2)\|\omega\|_\infty - 1 - n_h)^{1+2n_h}}$$

Proof: Using Lemma 3.2.4 and the Cauchy estimates (3.3),

$$\Big| \sum_{|k|=r_0}^{r_*-1} a_k F^{n_h}_{e^{i2\pi \frac{Tk\omega}{T} t},T,N}(\frac{j}{T}) - \sum_{|k|=r_0}^{r_*-1} a_k \phi^{n_h}_{e^{i2\pi \frac{Tk\omega}{T} t},T}(\frac{j}{T}) \Big| \leq$$
$$\leq \; C \sum_{|k|=r_0}^{r_*-1} e^{-\delta|k|} \frac{2(n_h!)^2(1+\frac{1}{2n_h})}{\pi (N - T(r_* + r_0 - 2)\|\omega\|_\infty - 1 - n_h)^{1+2n_h}},$$

since, for $|k| = r_0 \div r_* - 1$ and $j \in \{0, j_i, j_i^+ : i = 1 \div N_f\}$ there exists k_j with $|k_j| \leq r_0 - 1$ and $|Tk_j\omega - j| \leq 1$, and therefore

$$j + |Tk\omega| \leq T|k_j\omega| + 1 + T|k\omega| \leq T(|k_j| + |k|)\|\omega\|_\infty + 1 \leq T(r_0 + r_* - 2)\|\omega\|_\infty + 1.$$

Using Lemmas 3.3.2 and 3.3.1 and shifting the summation index by $\frac{m}{2}$ units, we get

$$\begin{aligned} \sum_{|k|=r_0}^{r_*-1} e^{-\delta|k|} &\leq \frac{2^m e^{\delta\frac{m}{2}}}{(m-1)!} \sum_{j=r_0+\frac{m}{2}}^{r_*-1+\frac{m}{2}} j^{m-1} e^{-\delta j} \\ &\leq \frac{2^m e^{\delta\frac{m}{2}} G_f(r_0 + \frac{m}{2}, r_* - 1 + \frac{m}{2}, m-1, \delta)}{(m-1)!}, \end{aligned}$$

from which the lemma follows. □

Note that the hypothesis of the previous lemma gives a new constraint for r_*, which is fulfilled if we take

$$r_* = \min\left(\left[\Big(\frac{TD}{\max((\frac{(n_h!)^2}{\pi})^{\frac{1}{1+2n_h}} + 1 + n_h, 2(1+n_h))} \Big)^{\frac{1}{\tau}} - r_0 + 2 \right], \left[\frac{N - 1 - n_h}{T\|\omega\|_\infty} - r_0 + 1 \right] \right)$$

Now, in order to complete the bound for $\|\Delta b\|$ we only have to bound the remainder.

Lemma 3.3.5 *The following inequality holds:*

$$\sum_{|k|=r_*}^{\infty} |a_k| \leq \frac{2^m C e^{\delta \frac{m}{2}} G_\infty(r_* + \frac{m}{2}, m-1, \delta)}{(m-1)!}$$

Proof: It follows from the Cauchy estimates (3.3), Lemma 3.3.2 and Definition 3.3.2. □

From lemmas 3.3.3, 3.3.4 and 3.3.5 follows a bound for $\|\Delta b\|_\infty$ that is stated in Theorem 3.4.1.

A more explicit description of the behavior of $\sum_{|k|=r_*+1}^{\infty} |a_k|$ is given in proposition 3.3.1. First we need two lemmas.

Lemma 3.3.6 *Define $P_l(j) = \#\{k \in \mathbb{Z}^l : |k| = j\}$. Then, for $j \geq 1$ the following recurrence is satisfied:*

$$P_l(j) = 2 + 2\sum_{s=1}^{j-1} P_{l-1}(s) + P_{l-1}(j), \tag{3.20}$$

with $P_1(j) = 2$. Moreover, $P_l(j)$ is a polynomial in j of degree $l-1$.

Proof: It is obvious that $P_1(j) = 2$. Assume $l \geq 2$. Then every $k \in \mathbb{Z}^l$ can be splitted as $k = (k_1, k_2)$ with $k_1 \in \mathbb{Z}^{l-1}$ and $k_2 \in \mathbb{Z}$. In this way

$$\{k \in \mathbb{Z}^l : |k| = j\} = \{(0, \pm j)\} \cup \Big(\bigcup_{s=1}^{j-1} \{(k_1, \pm(j-s)) : |k_1| = s\}\Big) \cup \{(k_1, 0) : |k_1| = j\},$$

and (3.20) follows from the fact that $\#\{(0, \pm j)\} = 2$, $\#\{(k_1, \pm(j-s)) : |k_1| = s\} = 2P_{l-1}(s)$ and $\#\{(k_1, 0) : |k_1| = j\} = P_{l-1}(j)$.

We see that $P_l(j)$ has degree $l-1$ in j by induction on l. For $l = 0$ it is true by definition. Assume it true for $l-1$, that is

$$P_{l-1}(j) = \sum_{r=0}^{l-2} c_r j^r.$$

Then

$$P_l(j) = 2 + 2\sum_{s=1}^{j-1}\sum_{r=0}^{l-2} c_r s^r + P_{l-1}(j) = 2 + 2\sum_{r=0}^{l-2} c_r \sum_{s=1}^{j-1} s^r + P_{l-1}(j),$$

and the property follows from the fact that

$$\sum_{s=1}^{j-1} s^r = \frac{1}{r+1}\sum_{s=0}^{r} \binom{r+1}{s} B_s j^{r-s+1}$$

(B_s are the Bernoulli numbers, see e.g. [22]) is a polynomial in j of degree $r+1$. □

Lemma 3.3.7 *(a) For $l, r \in \mathbb{N}$, $x \in \mathbb{C}$, $|x| < 1$ we have*

$$\sum_{j \geq r} j^l x^j = \frac{Q_l(x)}{(1-x)^{l+1}}, \tag{3.21}$$

where $Q_0(x) = x^r$ and $Q_l(x) = (x - x^2)Q'_{l-1}(x) + lxQ_{l-1}(x)$ for $l \geq 1$.

(b) $Q_l(x)$ is a polynomial in x with minimum degree r and maximum degree $r + l$. Moreover, every coefficient in x is a polynomial in r with maximum degree l.

Proof: For $l = 0$, (3.21) is the sum of a geometric series. Assume (3.21) true for $l - 1$. Then

$$\sum_{j \geq r} j^l x^j = x \sum_{j \geq r} j^l x^{j-1} = x \frac{d}{dx} \sum_{j \geq r} j^{l-1} x^j,$$

and using the induction hypothesis,

$$\begin{aligned} x \frac{d}{dx} \sum_{j \geq r} j^{l-1} x^j &= x \frac{d}{dx} \frac{Q_{l-1}(x)}{(1-x)^l} \\ &= \frac{(x - x^2)Q'_{l-1}(x) + lxQ_{l-1}(x)}{(1-x)^{l+1}}. \end{aligned}$$

As for (b), $Q_0(x)$ verifies (b) trivially and, assuming that (b) is true for $Q_{l-1}(x)$, it is readily checked that $Q_l(x) = (x - x^2)Q'_{l-1}(x) + lxQ_{l-1}(x)$ also verifies (b). □

Proposition 3.3.1 *We have*

$$\sum_{|k|=r}^{\infty} |a_k| = O(r^{m-1} e^{-r\delta}) \qquad \text{as } r \to +\infty.$$

Proof: Using the Cauchy estimates (3.3) and Lemma 3.3.6,

$$\sum_{|k|=r}^{\infty} |a_k| \leq C \sum_{|k|=r}^{\infty} e^{-\delta|k|} \leq C \sum_{j=r}^{\infty} P_m(j) e^{-\delta j}, \tag{3.22}$$

where $P_m(j)$ has degree $m - 1$ in j. Assume $P_m(j) = \sum_{s=0}^{m-1} c_{m,s} j^s$ and define $x = e^{-\delta}$. Then

$$\begin{aligned} C \sum_{j=r}^{\infty} P_m(j) e^{-\delta j} &= C \sum_{s=0}^{m-1} c_{m,s} \sum_{j=r}^{\infty} j^s x^j = C \sum_{s=0}^{m-1} c_{m,s} \frac{Q_s(x)}{(1-x)^{s+1}} \\ &= C \sum_{s=0}^{m-1} \Big(\frac{c_{m,s}}{(1-x)^{s+1}} \sum_{l=0}^{s} p_{s,l}(r) x^{r+l} \Big), \end{aligned}$$

where, following Lemma 3.3.7(b), we have expanded $Q_s(x)$ as $\sum_{l=0}^{s} p_{s,l}(r)x^{r+l}$, with $p_{s,l}(r)$ of maximum degree s in r.

To show that this expression is $O(r^{m-1}x^r)$ when $r \to \infty$, it is enough to see that

$$\lim_{r\to\infty} \frac{1}{r^{m-1}x^r} C \sum_{s=0}^{m-1} \Big(\frac{c_{m,s}}{(1-x)^{s+1}} \sum_{l=0}^{s} p_{s,l}(r)x^{r+l} \Big) = C \sum_{s=0}^{m-1} \Big(\frac{c_{m,s}}{(1-x)^{s+1}} \sum_{l=0}^{s} \Big(\lim_{r\to\infty} \frac{p_{s,l}(r)}{r^{m-1}} \Big) x^l \Big)$$

does not depend on r. This is true, since from Lemma 3.3.7(b) the $p_{s,l}$ polynomials are of degree $\leq s \leq m-1$ and therefore the limit in the right-hand side of the above equation does not depend on r. □

Lemmas 3.3.6 and 3.3.7 also allow to improve the bound of Lemma 3.3.5 for concrete values of m. For instance, if $m = 2$ we have

$$\sum_{\substack{|k|\geq r \\ |k|\in\mathbb{Z}^2}} |a_k| \leq 4C \frac{re^{-\delta r} + (1-r)e^{-\delta(r+1)}}{(1-e^{-\delta})^2}.$$

3.4 Final results

We end this section by gathering all the previous results in a single theorem that gives the bound for the error in frequencies and amplitudes. We consider both the case of known and the case of unknown frequencies in a single theorem.

Theorem 3.4.1 *Assume that we perform Fourier analysis of an analytic quasi–periodic function*

$$f(t) = \sum_{k\in\mathbb{Z}^m} a_k e^{i2\pi\omega t},$$

that satisfies the Cauchy estimates with constants $C, \delta > 0$,

$$|a_k| \leq Ce^{-\delta|k|},$$

and whose frequency vector $\omega = (\omega_1, \ldots, \omega_m)$ satisfies a Diophantine condition of the form

$$|k\omega| > \frac{D}{|k|^\tau},$$

with $D, \tau > 0$. Assume we sample f in N points equally spaced over the interval $[0, T]$, and that we want to determine the frequencies $Tk\omega$ with $1 \leq |k| \leq r_0 - 1$, $Tk\omega > 0$, and the related amplitudes, from which we have approximations close enough to the actual ones. Assume that we carry out the procedure of section 2.4 with $n_h \geq 1$ and get an approximation of f of the form

$$p(t) = A_0^c + \sum_{l=1}^{N_f} \Big(A_l^c \cos(\frac{2\pi\nu_l t}{T}) + A_l^s \sin(\frac{2\pi\nu_l t}{T}) \Big).$$

Assume also that N is such that $N - T(2r_0 - 2)\|\omega\|_\infty > 3 + n_h$, and T is such that $\frac{TD}{(2r_0-2)^\tau} > 3 + n_h$ and $[\nu_{min}] > 2 + n_h$. Then, the error in frequencies and amplitudes, which we denote as Δy, can be bounded, in the first order approximation, as

$$\|\Delta y\| \lesssim \|M^{-1}\|\|\Delta b\|, \tag{3.23}$$

where

$$\|M^{-1}\| \leq \frac{\|M_D^{-1}\|}{1 - \|M_D^{-1}\|\|M_O\|}$$

and

$$\begin{aligned}
\|M_O\| \;\leq\; & \frac{(n_h!)^2}{\pi}\Bigg(\frac{\sqrt{2}(\sum_{l=1}^{N_f} A_l)(\pi + \ln(\frac{TD}{(2r_0-2)^\tau} - 1 + n_h) - \ln(\frac{TD}{(2r_0-2)^\tau} - 2 - n_h)) + 2N_f}{(\frac{TD}{(2r_0-2)^\tau} - 1 - n_h)^{1+2n_h}} \\
& + \frac{\sqrt{2}(\sum_{l=1}^{N_f} A_l)(\pi + \ln([\nu_{min}] + n_h) - \ln([\nu_{min}] - 1 - n_h)) + 2N_f}{([\nu_{min}] - n_h)^{1+2n_h}} \\
& + \frac{4\Big(\sqrt{2}(\sum_{l=1}^{N_f} A_l)(\pi + \ln(N - \Omega_0 + n_h) - \ln(N - \Omega_0 - 1 - n_h)) + 2N_f\Big)(1 + \frac{1}{2n_h})}{(N - \Omega_0 - n_h)^{1+2n_h}}\Bigg)
\end{aligned}$$

and

$$\|M_D^{-1}\| \leq \frac{\|\mathcal{M}_D^{-1}\|}{1 - \|\mathcal{M}_D^{-1}\|\varepsilon_1}, \qquad \|\mathcal{M}_D^{-1}\| \leq \frac{\|\mathfrak{M}_D^{-1}\|}{1 - \|\mathfrak{M}_D^{-1}\|\varepsilon_2}, \qquad \|\mathfrak{M}_D^{-1}\| \leq \frac{G_{n_h}}{\min(1, A_{min})},$$

with G_{n_h} as in Definition 3.2.4, being

$$\begin{aligned}
\varepsilon_1 \;&=\; \frac{4(n_h!)^2\Big(\sqrt{2}A_{max}(\pi + \ln(N - \Omega_0 + n_h) - \ln(N - \Omega_0 - 1 - n_h)) + 2\Big)(1 + \frac{1}{2n_h})}{\pi(N - \Omega_0 - n_h)^{1+2n_h}}, \\
\varepsilon_2 \;&=\; \frac{(n_h!)^2\Big(\sqrt{2}A_{max}(\pi + \ln(2[\nu_{min}] + n_h) - \ln(2[\nu_{min}] - n_h - 1)) + 2\Big)}{\pi(2[\nu_{min}] - n_h)^{1+2n_h}}, \\
\Omega_0 \;&=\; T(2r_0 - 2)\|\omega\|_\infty + 1,
\end{aligned}$$

As for $\|\Delta b\|$,

$$\begin{aligned}
\|\Delta b\| \;\leq\; & \frac{2^{m+1}C}{(m-1)!}\Bigg(\\
& \chi_{\{r_* > r_0\}} \frac{(n_h!)^2 e^{\delta(r_0-1)} \sum_{l=0}^{m-1} \binom{m-1}{l} (\frac{m}{2} - r_0 + 1)^{m-1-l} G_f(2r_0 - 1, r_0 + r_* - 2, l + \tau(1 + 2n_h), \delta)}{E_* \pi (TD)^{1+2n_h}} \\
& + \chi_{\{r_* > r_0\}} \frac{2(n_h!)^2 e^{\delta\frac{m}{2}}(1 + \frac{1}{2n_h}) G_f(r_0 + \frac{m}{2}, r_* - 1 + \frac{m}{2}, m - 1, \delta)}{\pi(N - \Omega - n_h)^{1+2n_h}} \\
& + e^{\delta\frac{m}{2}} G_\infty(r_* + \frac{m}{2}, m - 1, \delta)\Bigg),
\end{aligned} \tag{3.24}$$

where

$$\begin{aligned}
\Omega &= T(r_* + r_0 - 2)\|\omega\|_\infty + 1 \\
r_* &= \max\Big(r_0, \min\Big(\Big[\Big(\frac{TD}{\max((\frac{(n_h!)^2}{\pi})^{\frac{1}{1+2n_h}} + 1 + n_h, 2(1+n_h))}\Big)^{\frac{1}{\tau}} - r_0 + 2\Big], \\
&\qquad \Big[\frac{N - 1 - n_h}{T\|\omega\|_\infty} - r_0 + 1\Big]\Big)\Big) \qquad (3.25) \\
E_* &= \frac{(z_* - 1 - n_h)^{1+2n_h}}{z_*^{1+2n_h}}, \\
z_* &= \frac{TD}{(r_* + r_0 - 2)^\tau},
\end{aligned}$$

and the G_f, G_∞ functions are those of Definition 3.3.2.

If we assume that the frequencies $\{Tk\omega\}_{|k|\le r_0-1}$ are known and want to compute the amplitudes using the procedure described in Section 2.3, formula (3.23) is still valid, where the bounds for $\|\Delta b\|$ are the same as before and the bounds for $\|M^{-1}\|$ are given by

$$\|M^{-1}\| \le \frac{\|M_D^{-1}\|}{1 - \|M_D^{-1}\|\|M_O\|}$$

being

$$\begin{aligned}
\|M_O\| \;\le\; & \frac{N_f(n_h!)^2}{\pi}\Big(\frac{\sqrt{2}}{\pi(\frac{TD}{(2r_0-1)^\tau} - \frac{1}{2} - n_h)^{1+2n_h}} + \frac{2}{\pi([\nu_{min}] - n_h)^{1+2n_h}} \\
& + \frac{8(1 + \frac{1}{2n_h})}{\pi(N - T(2r_0-2)\|\omega\|_\infty - \frac{1}{2} - n_h)^{1+2n_h}}\Big)
\end{aligned}$$

and

$$\|M_D^{-1}\| \le \frac{\|\mathcal{M}_D^{-1}\|}{1 - \|\mathcal{M}_D^{-1}\|\varepsilon_1}, \qquad \|\mathcal{M}_D^{-1}\| \le \frac{\|\mathfrak{M}_D^{-1}\|}{1 - \|\mathfrak{M}_D^{-1}\|\varepsilon_2}, \qquad \|\mathfrak{M}_D^{-1}\| \le \frac{5}{3},$$

being

$$\varepsilon_1 = \frac{8(n_h!)^2(1 + \frac{1}{2n_h})}{\pi(N - T(2r_0-2)\|\omega\|_\infty - \frac{1}{2} - n_h)^{1+2n_h}}, \qquad \varepsilon_2 = \frac{2(n_h!)^2}{\pi(2[\nu_{min}] - n_h)^{1+2n_h}}.$$

Remark 3.4.1 *The bound for $\|\Delta y\|$ given by the previous theorem can be improved by replacing the first term in the bound for $\|\Delta b\|$ by any of the intermediate inequalities of the proof of Lemma 3.3.3. In this case, it may be necessary to modify the definition of r_*. We will give examples in the following section.*

Corollary 3.4.1 *The block Jacobi method as stated in (2.7), used to obtain the amplitudes from known frequencies, is convergent provided that*

$$\|M_D^{-1}\|\|M_O\| < 1,$$

where for $\|M_D^{-1}\|$ and $\|M_O\|$ we use the bounds given in the previous theorem in the case of known frequencies, but replacing N_f by $N_f - 1$.

Proof: The norm of the iteration matrix of the block Jacobi method (2.7) is

$$\max_{i=1\div N_f}\left(\|B_{i,i}^{-1}\|\sum_{\substack{l=1\\ l\neq i}}^{N_f}\|B_{i,l}\|\right) < \|M_D^{-1}\|\|M_O\|.$$

The reason for replacing N_f by $N_f - 1$ in the bounds of the previous theorem is that we apply the block Jacobi method to system (2.3) without its first equation. □

Chapter 4

A numerical example

In this Chapter we apply the procedure developed in Chapter 2 to a family of quasi–periodic functions for which explicit expressions for its frequencies and amplitudes can be computed. These expressions are used to test the error estimates developed in Chapter 3.

4.1 The family of functions analyzed

In order to illustrate the procedures described and to test the error bounds obtained, we have analyzed a family of quasiperiodic functions from which the Fourier coefficients can be explicitly calculated. The functions are

$$f_\mu(t) = \frac{\sin(2\pi\omega_1 t + \varphi_1)}{1 - \mu\cos(2\pi\omega_1 t + \varphi_1)} \cdot \frac{\sin(2\pi\omega_2 t + \varphi_2)}{1 - \mu\cos(2\pi\omega_2 t + \varphi_2)}, \qquad \mu \in [0, 1).$$

They verify $f_\mu(t) = \sum_{k\in\mathbb{Z}^2} a_k^{\mu,\varphi} e^{2\pi i(\omega,k)t}$ with

$$a_k^{\mu,\varphi} = \begin{cases} \frac{-\operatorname{sign}(k_1 k_2)}{\mu^2} c_2^{|k|} e^{i(k,\varphi)} & \text{if } k_1, k_2 \neq 0 \\ 0 & \text{if } k_1 = 0 \text{ or } k_2 = 0 \end{cases} \tag{4.1}$$

being

$$\omega = (\omega_1, \omega_2), \quad \varphi = (\varphi_1, \varphi_2) \quad \text{and} \quad c = \frac{1 - \sqrt{1 - \mu^2}}{\mu}$$

The parameter μ is directly related to the parameter δ in the Cauchy estimates (3.3), namely

$$\delta = \operatorname{Im} \arccos \frac{1}{\mu} = -\log c.$$

4.2 Numerical results

In this section we will show the results corresponding to apply the algorithm described in 2.5.1 to the f_μ functions for $\omega = (1, \sqrt{2})$, $\varphi = (\sqrt{0.2}, \sqrt{0.3})$, $n_h = 2$ and several values of μ, T and N. For the chosen value of ω, the parameters D and τ of the Diophantine

condition (3.2) are 0.85355 and 1, respectively. We have stopped the procedure when all the frequencies (with nonzero amplitudes) of order $|k| \leq 5$ have been refined. The error of the Fourier approximation as well as the corresponding bound, as given by theorem 3.4.1, are shown in Fig. 4.1.

It must be noted that the error in frequencies and amplitudes is much smaller than the difference between the analyzed function f and its computed quasi–periodic approximation Q_f. For instance, in the case of $\mu = 0.9$, from (4.1) the maximum amplitude of the frequencies not determined is $c_2{}^6/\mu^2 = 0.6268$, whereas we reach errors as small as 10^{-14} for some values of T and N. This is due to the fact that the truncation error of our procedure is not introduced by the difference $f - Q_f$ but by its DFT, as is seen in Section 3.

We observe that, for every value of T, as N increases the error decreases and becomes constant after a value of N. We also note that the minimum error for each value of T decreases as we increase T. This behavior of the error in terms of the parameters T, N, can be explained in terms of the bound (3.16).

Let r_1 be such that $2C\sum_{|k|=r_1}^{\infty} e^{-\delta|k|}|F^{n_h}_{e^{i2\pi k\omega t},T,N}(Tk\omega - j)|$ is small (this might be different from the order r_* of Section 3, which is "the order up to which the TCFT helps"). Then the frequencies of order greater than r_1 can be considered irrelevant and we can focus in frequencies of orders from r_0 to $r_1 - 1$. If N is large enough, we can replace the DFT by the TCFT in (3.16), that is

$$\|\Delta b\| \lesssim \max_{j\in\{0,j_i,j_i^+\}_{i=1}^{N_f}} 2C \sum_{|k|=r_0}^{r_1-1} e^{-\delta|k|}|\phi^{n_h}_{e^{i2\pi k\omega t},T}(\frac{Tk\omega - j}{T})|.$$

In order to normalize, we note that $|\phi^{n_h}_{e^{i2\pi k\omega t},T}(\frac{Tk\omega-j}{T})| \leq |g^{n_h}(Tk\omega - j)| = |\widetilde{g}^{n_h}(Tk\omega - j)|$, being

$$g^{n_h}(\alpha) = \frac{(-1)^{n_h}(n_h!)^2(e^{i2\pi\alpha} - 1)}{2\pi i\psi_{n_h}(\alpha)}, \qquad \widetilde{g}^{n_h}(\alpha) = \frac{(-1)^{n_h}(n_h!)^2}{\pi i\psi_{n_h}(\alpha)}.$$

The moduli of these functions are plotted in Fig. 4.2. As T increases, the differences $|Tk\omega - j|$ become larger and, since $|\widetilde{g}^{n_h}(\alpha)|$ decreases with $|\alpha|$, this explains why, for sufficiently large N, the error decreases as T increases.

In order to consider the case in which N is not large, we note $|F^{n_h}_{e^{i2\pi k\omega t,T,N},T,N}(j)| = |h^{n_h}_N(Tk\omega - j)| \leq |\widetilde{h}^{n_h}_N(Tk\omega - j)|$, being

$$\begin{aligned} h^0_N(\alpha) &= \frac{1 - e^{i2\pi\alpha}}{N(1 - e^{i2\pi\alpha/N})}, \quad h^{n_h}_N(\alpha) = \frac{q_{n_h}}{2^{n_h}} \sum_{l=-n_h}^{n_h} (-1)^l \binom{2n_h}{n_h + l} h^0_N(\alpha + l), \\ \widetilde{h}^0_N(\alpha) &= \frac{2}{N(1 - e^{i2\pi\alpha/N})}, \quad \widetilde{h}^{n_h}_N(\alpha) = \frac{q_{n_h}}{2^{n_h}} \sum_{l=-n_h}^{n_h} (-1)^l \binom{2n_h}{n_h + l} \widetilde{h}^0_N(\alpha + l). \end{aligned}$$

The moduli of these functions are plotted in Fig. 4.2. Now, if N is not large enough, it may happen that one of the $|Tk\omega - j|$ approaches N and raises the bound (3.16). This explains the fact that, for a fixed value of T, as we decrease N the error ends up increasing.

The qualitative behaviour of the bound given by theorem 3.4.1 is not the same as the one of the real error. For each value of T, as N increases, the bound given by theorem

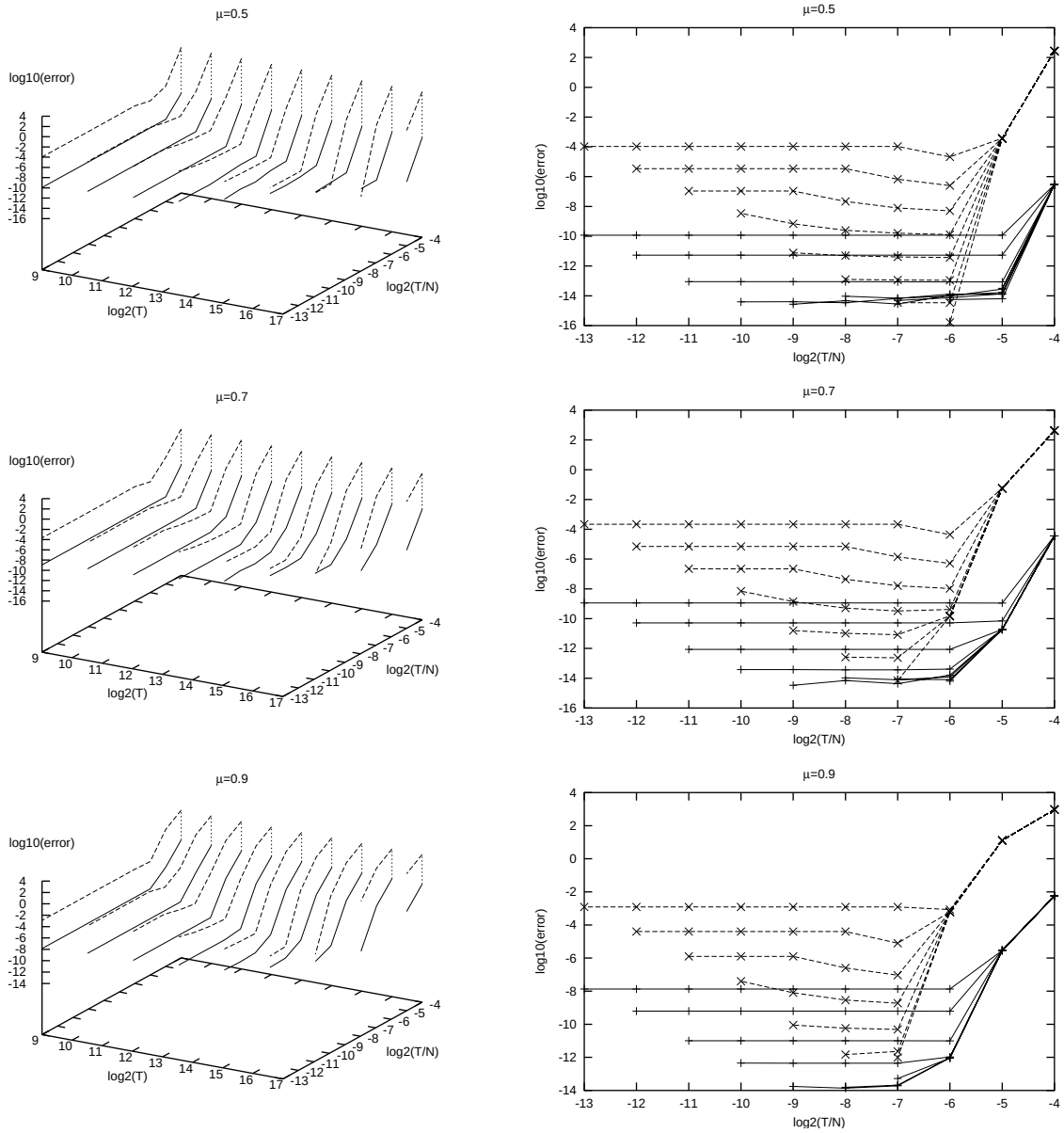


Figure 4.1: Fourier analysis of the f_μ functions for $\mu = 0.5, 0.7, 0.9$ and several values of T and N. Points corresponding to analysis with the same value of T have been joined by lines. The solid lines represent the error in frequencies and amplitudes of the corresponding Fourier analysis. This means that we have represented the maximum value between the error in the frequencies in the error in the amplitudes. The points joined by dashed lines correspond to the bound given by theorem 3.4.1. The right–hand figures are the (y, z) projection of the left–hand side ones.

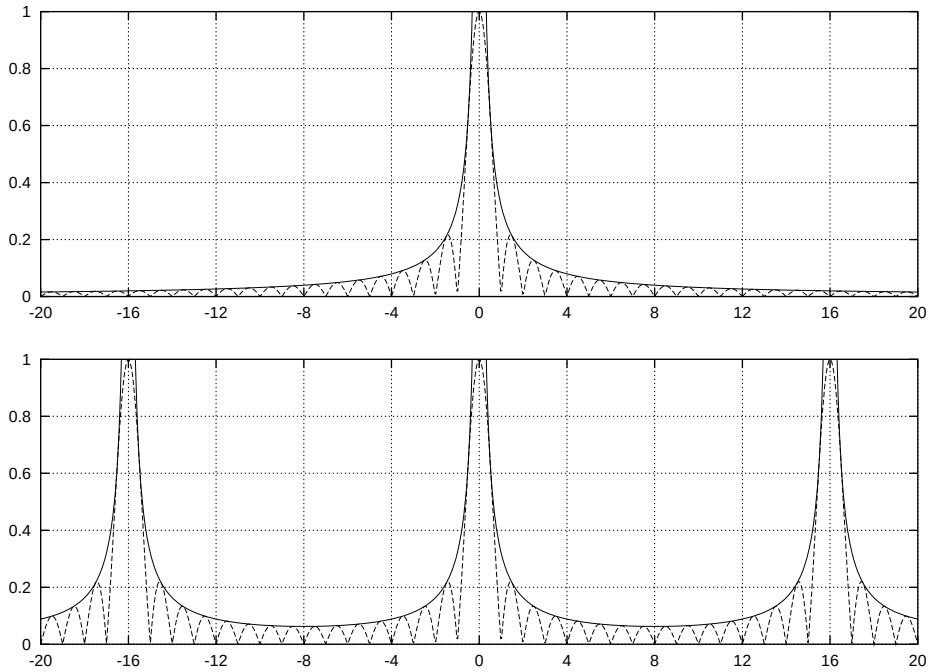


Figure 4.2: Top: graph of the functions $|g^{n_h}(\alpha)|$ (dashed line) and $|\widetilde{g}^{n_h}(\alpha)|$ (solid line) for $n_h = 0$. Bottom: graph of the functions $|h_N^{n_h}(\alpha)|$ (dashed line) and $|\widetilde{h}_N^{n_h}(\alpha)|$ (solid line) for $n_h = 0$ and $N = 16$.

3.4.1 decreases up to a minimum value, then increases slightly and becomes constant. This increasing is due to the introduction of the E_* constant in the proof of the Lemma 3.3.3, which can enlarge the bound by a factor $1/E_*$ (at most 32 for $n_h = 2$, see (3.19)). In Fig. 4.3, we evaluate the bound for $\|\Delta y\|$ replacing the first term in (3.24) by (3.18), which is the last bound in the proof of 3.3.3 before the introduction of E_*. We see how the increasing of the bound after the minimum of Fig 4.1 disappears.

The drawback of this approach is that the sum in (3.18) runs over multiindices $|k| = r_0 \div r_* - 1$ instead of their orders $j = r_0 \div r_* - 1$, and its evaluation can be prohibitive in terms of computing time, especially if the number of basic frequencies m is large. An alternative could be to lower r_* in (3.25) in order to raise the minimum value of E_*. For instance, if we set r_* equal to

$$r_* = \max\Big(r_0, \min\Big(\Big[\Big(\frac{TD}{\max\big(\big(\frac{(n_h!)^2}{\pi}\big)^{\frac{1}{1+2n_h}} + 1 + n_h, \frac{1+n_h}{1-(1/2)^{1/(1+2n_h)}}\big)}\Big)^{1/\tau} - r_0 + 2\Big], \Big[\frac{N-1-n_h}{T\|\omega\|_\infty} - r_0 + 1\Big]\Big)\Big),$$

then the minimum allowed value of E_* is $1/2$. But this can lead to a worse global bound if the Fourier coefficients $|a_k|$ decrease slowly, as we illustrate in table 4.1. A different alternative is to take as value of r_* the one that minimizes the bound given by theorem 3.4.1. The results in this case are given in Fig. 4.5. Of course, they are worse that the ones of Fig. 4.3.

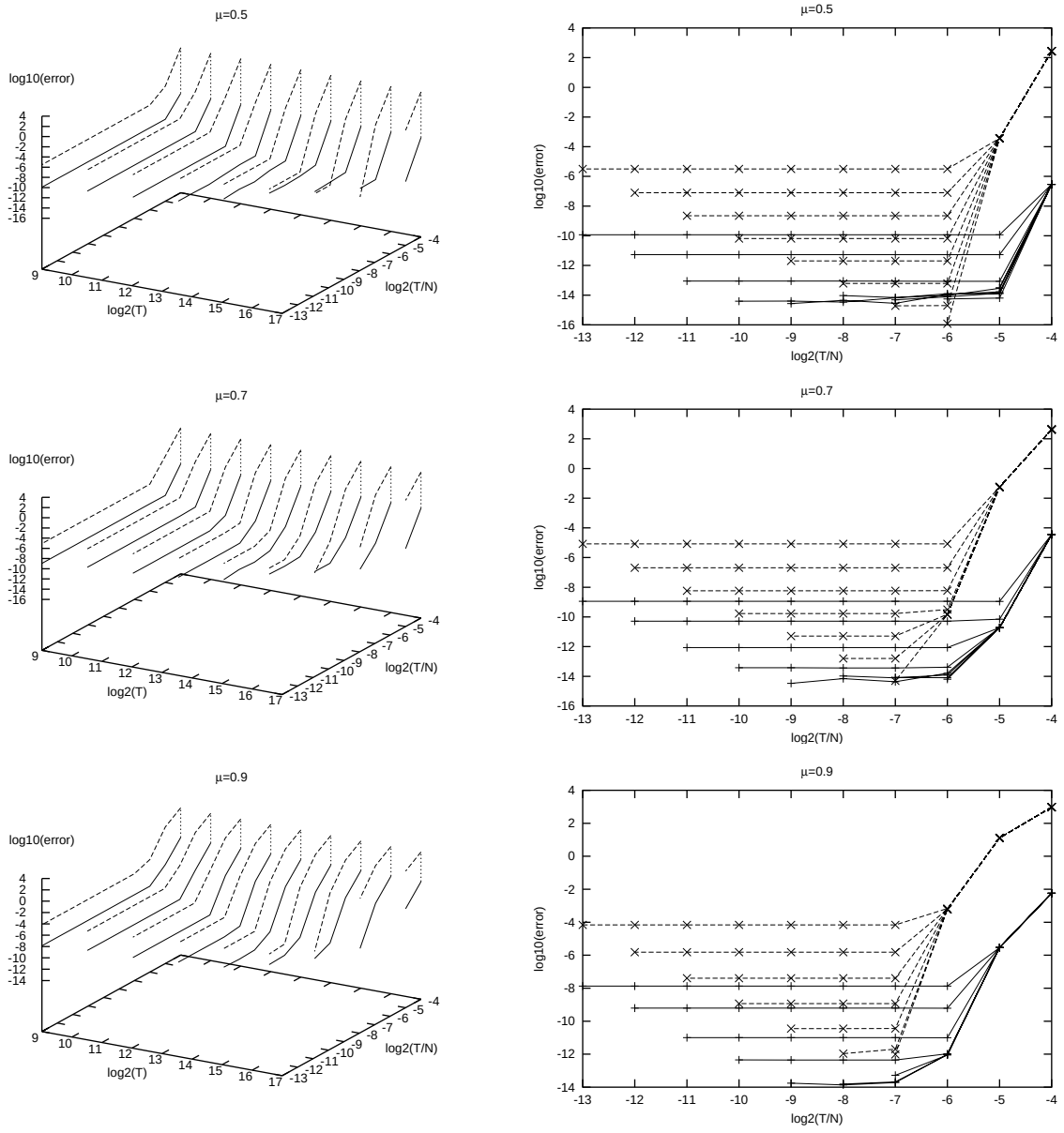


Figure 4.3: This is the same exploration as the one of Fig. 4.1, except that the dashed lines represent the bound obtained replacing the first term in (3.24) by (3.18).

$\max 1/E_*$	μ	T	N	r_*	actual $1/E_*$	bound
32	0.9	1024	262144	61	3.53332	4.53023E–06
32	0.9	1024	524288	121	16.4812	2.08997E–05
32	0.9	1024	1048576	141	31.2714	3.96552E–05
2	0.9	1024	262144	33	1.97207	1.32447E–02
2	0.9	1024	524288	33	1.97207	1.32447E–02
2	0.9	1024	1048576	33	1.97207	1.32447E–02

Table 4.1: Computation of the bound given by theorem 3.4.1 using two different maximum allowed values for $1/E_*$. We see how, by lowering the maximum value of $1/E_*$, the bound can increase drastically.

In Fig. 4.3, the bound is still several orders of magnitude larger than the actual error. This is due to the Diophantine condition, which give only a lower bound for the difference between frequencies. This difference reaches the Diophantine condition in very few cases, as shown in Fig. 4.4. In Fig. 4.6 we evaluate the bound of theorem 3.4.1 by replacing the first term of the bound of $\|\Delta b\|$ by (3.17). We see that in this case there is a very good agreement between the error predicted and the actual error.

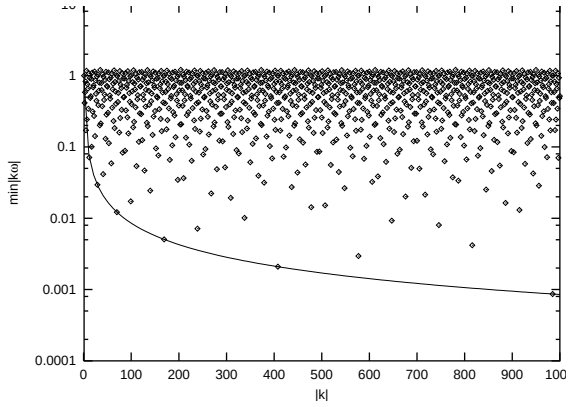


Figure 4.4: Illustration of the non–optimality of the Diophantine condition. The points represent the values of $\min_{|k|=\text{const.}} |k\omega|$ for $|k| = 1 \div 1000$. The curve represents the values of the Diophantine condition $0.85355/|k|$. The only points that are approximately on the curve $0.85355/|k|$ correspond to the values $|k|$=1, 2, 5, 12, 29, 70, 169, 408, 985.

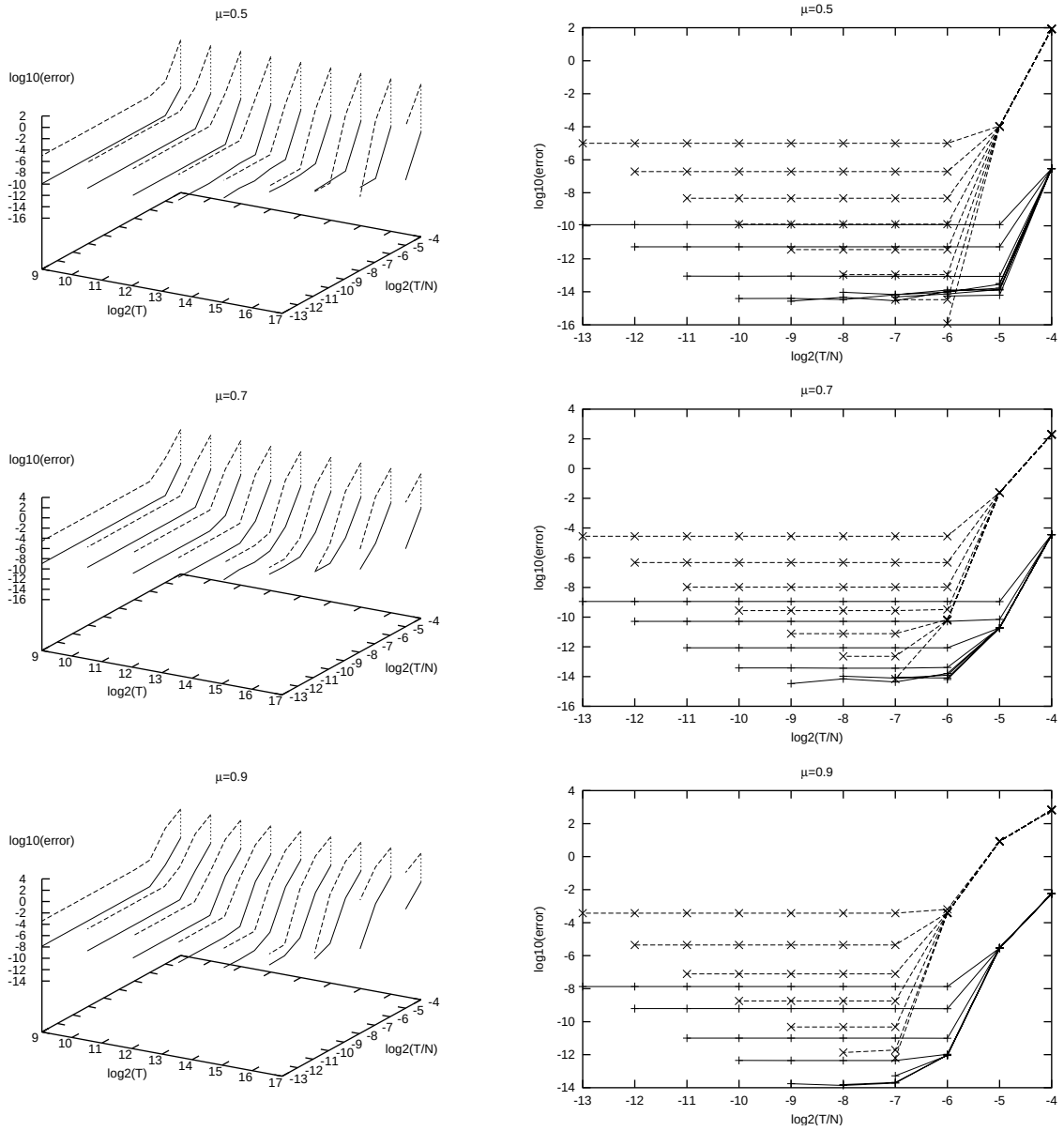


Figure 4.5: This is the same exploration as the one of Fig. 4.1, except that the error bound represented by the dashed lines is obtained by minimizing $\|\Delta b\|$ with respect to r_* in theorem 3.4.1, instead of using (3.25).

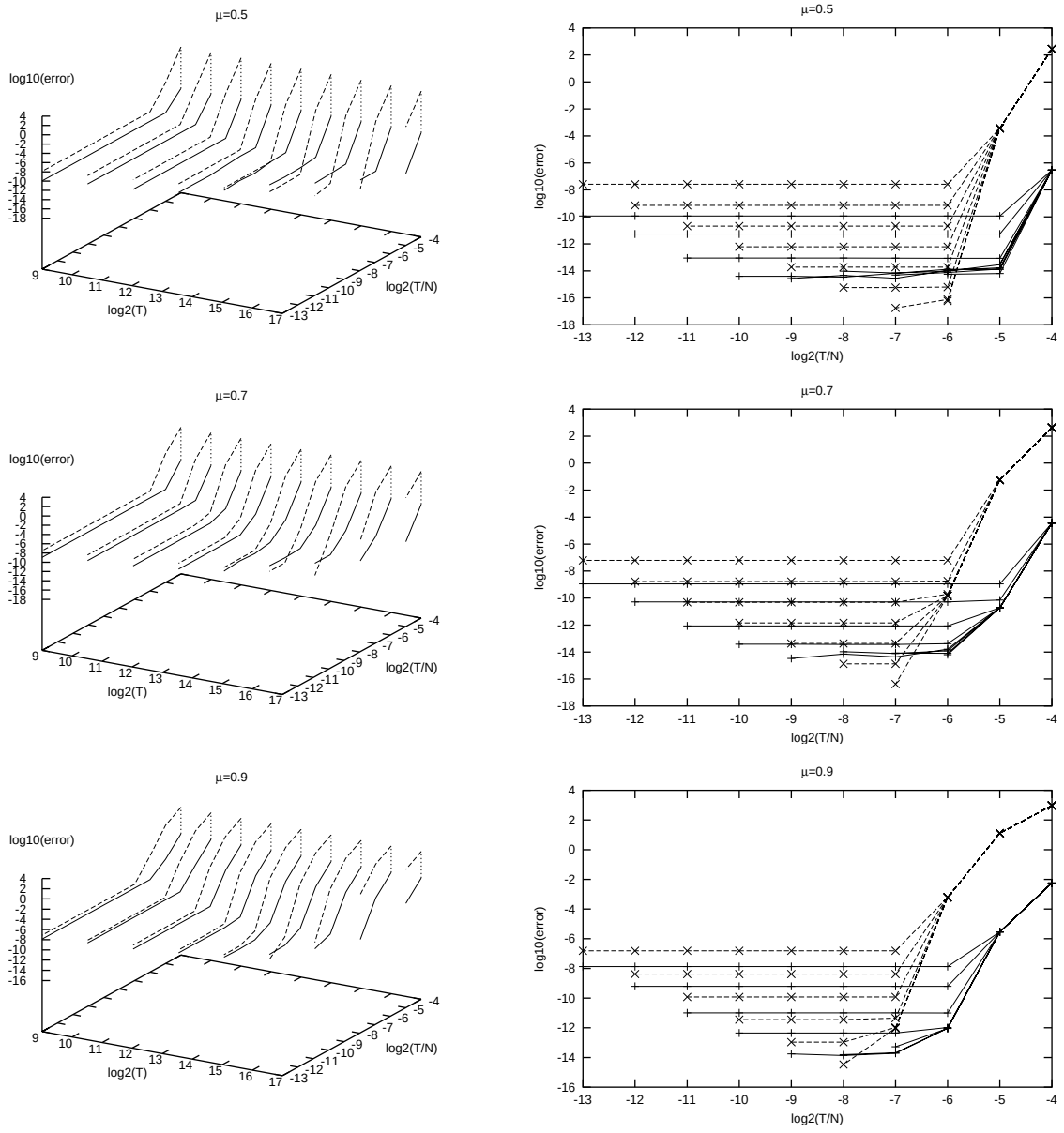


Figure 4.6: This is the same exploration as the one of Fig. 4.1, except that the error bound represented by the dashed lines is obtained by replacing the first term of (3.24) by (3.17).

Chapter 5

Application to the development of Solar System models

In this chapter we apply the procedures of Chapter 2 to the development of simplified models for the motion in the Solar System. They are based on Fourier analysis of the time–dependent part of the real Solar System equations of motion written as a perturbation of the RTBP (see Appendix A). We develop models for the Earth–Moon and Sun–Earth+Moon systems by selecting frequencies from the computed Fourier expansions in a suitable manner. These models are tested against other well–known models through the computation of residual accelerations along selected orbits.

5.1 Introduction

Through this chapter, we will denote the bodies of the Solar System as

$$\mathcal{S} = \{P_1, \ldots, P_9, P_{10}, P_{11}\} \tag{5.1}$$

where P_1,...,P_{11} denote Mercury, Venus, Earth, Mars, Jupiter, Saturn, Uranus, Neptune, Pluto, the Moon and the Sun, respectively. We will also denote the Earth, the Moon and the Sun as E, M and S, respectively. The mass of $P_I \in \mathcal{S}$ will be denoted as m_{P_I}.

Sometimes we will be interested in considering the Earth and the Moon as a single body, located at the Earth–Moon barycentre. We will denote this "virtual" body as P_{12}. In this case, we will consider a modified Solar System

$$\mathcal{S} = \{P_1, P_2, P_4, \ldots, P_9, P_{11}, P_{12}\}, \tag{5.2}$$

which is denoted as before in order to reduce notation.

Let us consider two bodies $I, J \in \mathcal{S}$ (either the "true" Solar System or the modified one) with $m_I > m_J$, which we will call *primaries*. We can choose coordinates $(x, y, z)^\top$ and time units t such that

- the bodies I, J remain fixed at the positions $(\mu_{I,J}, 0, 0)^\top$ and $(\mu_{I,J} - 1, 0, 0)^\top$, respectively, being

$$\mu_{I,J} = \frac{m_J}{m_I + m_J},$$

- the body J completes a revolution around I in 2π time units.

Such coordinates will be called *adimensional* and are introduced in Appendix A (Section A.4). In these coordinates, the equations of motion of a particle under the Newtonian attraction of the bodies of the Solar System can be written as

$$\left\{\begin{array}{rcl} \ddot{x} &=& c_1 + c_4\dot{x} + c_5\dot{y} + c_7 x + c_8 y + c_9 z + c_{13}\dfrac{\partial\Omega}{\partial x}, \\ \ddot{y} &=& c_2 - c_5\dot{x} + c_4\dot{y} + c_6\dot{z} - c_8 x + c_{10} y + c_{11} z + c_{13}\dfrac{\partial\Omega}{\partial y}, \\ \ddot{z} &=& c_3 - c_6\dot{y} + c_4\dot{z} + c_9 x - c_{11} y + c_{12} z + c_{13}\dfrac{\partial\Omega}{\partial z}, \end{array}\right. \tag{5.3}$$

being

$$\begin{array}{rcl} \Omega &=& \dfrac{1-\mu_{I,J}}{\sqrt{(x-\mu_{I,J})^2+y^2+z^2}} + \dfrac{\mu_{I,J}}{\sqrt{(x-\mu_{I,J}+1)^2+y^2+z^2}} \\ && + \displaystyle\sum_{\substack{j\in\mathcal{S} \\ j\neq I,J}} \dfrac{\mu_{I,J,j}}{\sqrt{(x-x_j)^2+(y-y_j)^2+(z-z_j)^2}} \end{array} \tag{5.4}$$

where

$$\mu_{I,J,j} = \frac{m_j}{m_I + m_J},$$

and $(x_j, y_j, z_j)^\top$ are the adimensional coordinates of the body $j \in \mathcal{S}$. In system (5.3), $\{c_i\}_{i=1\div 13}$ are time–dependent functions which can be computed in terms of the positions, velocities, accelerations and over–accelerations of the two primaries I, J. The actual formulae are given in Appendix A. If we set $c_5 = 2$, $c_7 = c_{10} = c_{13} = 1$ and the remaining c_i equal to zero, and we skip the sum in (5.4), then (5.3) become the RTBP equations (A.1) with mass parameter $\mu_{I,J}$. Therefore, we can see (5.3) as a perturbation of the RTBP equations. We can get an idea of the order of this perturbation by looking at the coefficient A_1 of the Fourier expansions of the c_i functions in Appendix A, tables C.1 to C.13 and C.41 to C.53.

In order to evaluate the previous system of equations, we need the positions of the bodies of the Solar System, as well as its derivatives with respect to time up to order three. They can be computed from any analytical or numerical planetary ephemeris. In the computations we have used the JPL ephemeris DE406, because of its high precision over a 6000–year time span. It has the drawback of introducing discontinuities in accelerations and over–accelerations, which are discussed in Appendix B. Higher precision can be obtained by using DE405, but the time span is reduced to 600 years in this case.

In the following sections, we will develop intermediate models between the RTBP and the "real" Solar System (5.3). The strategy followed is to "add basic frequencies" to the RTBP, these frequencies being computed by applying the techniques described in Chapter 2 to the $\{c_i\}_{i=1\div 13}$ and $\{(x_j, y_j, z_j)^\top\}_{j\in\mathcal{S}}$ functions. The models will be developed for

- the Earth–Moon case, which means to consider $I = P_3$ and $J = P_{10}$, being $\mathcal{S}$ as in (5.1), and

- the Sun–Earth+Moon case, which means to consider $I = P_{11}$ and $J = P_{12}$, being $\mathcal{S}$ as in (5.2).

Although we will only cover these two cases, the methodology used can be applied to any pair of primaries.

5.2 Fourier analysis of the time-dependent part of the real Solar System in adimensional coordinates

We give in this section the results of the Fourier analysis of the time–dependent part of system (5.3).

5.2.1 Fourier analysis of the c_i functions

Applying the algorithm described in section 2.5.1, we have performed Fourier analysis of the $\{c_i\}_{i=1\div 13}$ functions, both for the Earth–Moon case and the Sun–Earth+Moon case. The parameters used have been: number of total iterates in the Fourier procedure $n = 10$, minimum value of the frequency threshold $b_{min} =$ 1E–10, Hanning level $n_h = 2$ and several values of the length of the time interval T and the number of points N whose choice will be discussed bellow. The frequency threshold b of the algorithm of 2.5.1 has never reached the value b_{min}, because all the analysis have finished due to the detection of close frequencies. In each analysis, we have computed the maximum difference between the analyzed c_i function and its quasi–periodic approximation, that is,

$$d_{max} = \max_{l=0\div N-1} |c_i(l\tfrac{T}{N}) - Q_{c_i}(l\tfrac{T}{N})|, \tag{5.5}$$

where c_i is the analyzed function and Q_{c_i} its quasi–periodic approximation. In figures 5.1 (Earth–Moon case) and 5.2 (Sun–Earth+Moon case), we have represented the minimum of d_{max} with respect to N for each value of T.

Since we have no a priori information of the behavior of the c_i functions, we have based our choice of the T, N parameters according to the following criteria:

- We have chosen time intervals starting at Jan 1st 2001 and of length at least 95 years.
- We have followed strategies to avoid aliasing.
- We have considered that a set of frequencies is "better" than another if the value of d_{max} related to the first set is smaller than the one related to the second one.

Due to our implementation of the Fourier analysis procedures, the N parameter must range over powers of two. For consistency, the T parameter has also been chosen to range over a geometric progression. The time interval of all analysis starts in January 1st, 2001. The smallest time interval length, T_{min}, has been taken of 95 years (34698.75

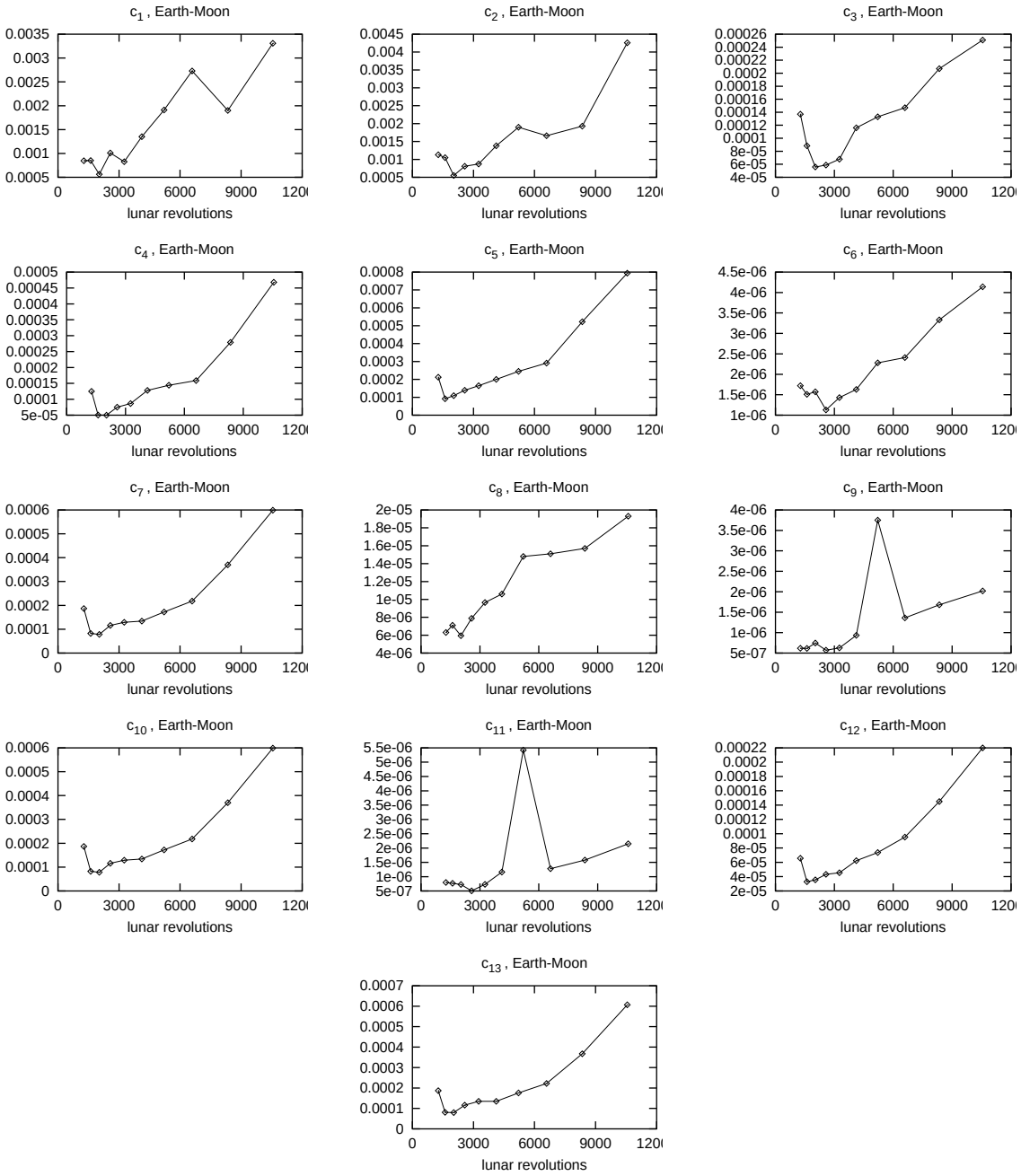


Figure 5.1: Error results of the Fourier analysis of the c_i functions in the Earth–Moon case. For each value of T explored, we have represented the minimum value of d_{max} with respect to N.

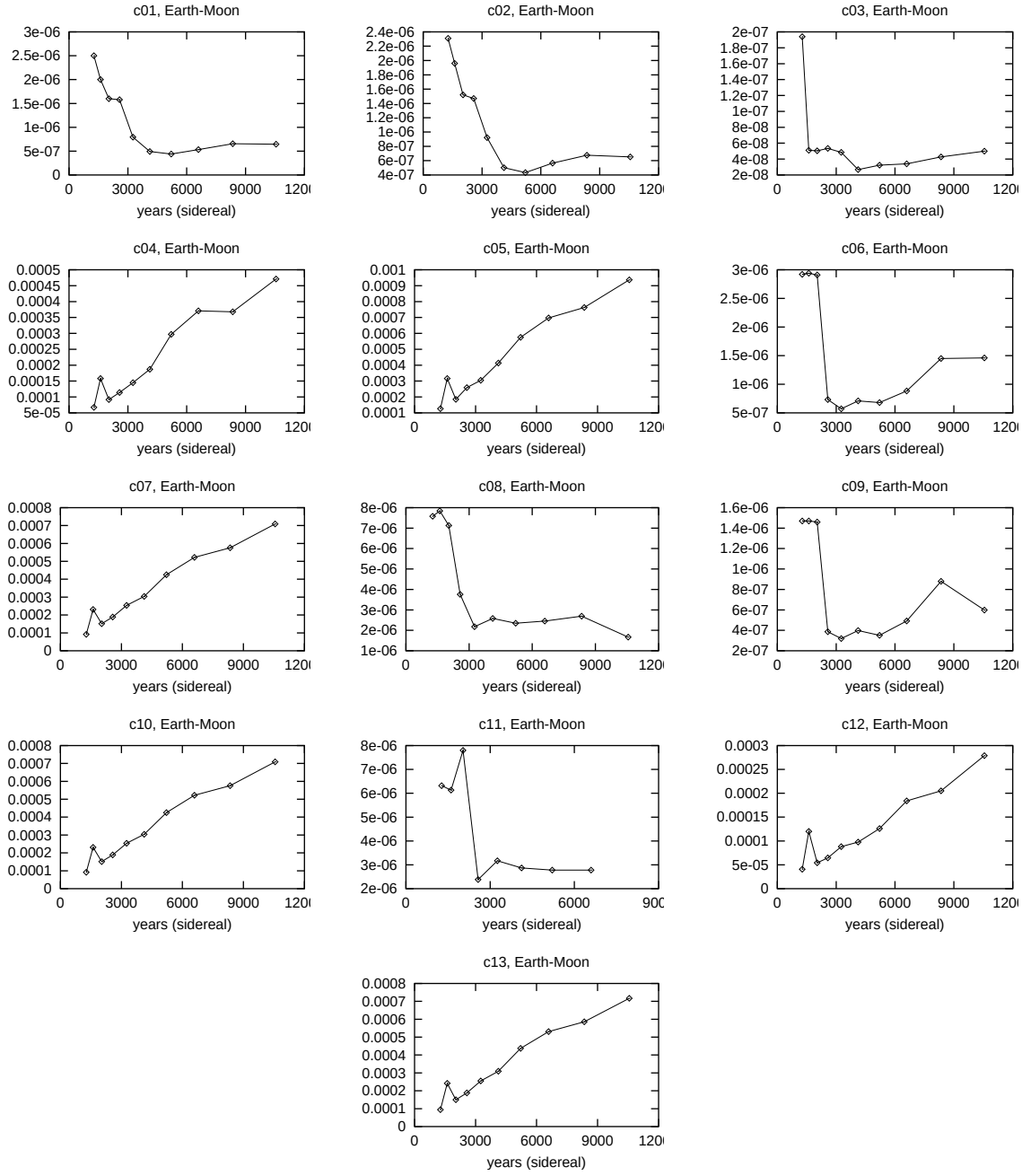


Figure 5.2: Same as Fig. 5.1 but for the Sun–Earth+Moon case.

Julian days). The greatest time-interval length, T_{max}, has been chosen as the maximum time interval given by the JPL DE406 ephemerides after Jan 1st 2001, which is 364938 Julian days (999.15 years). Therefore, we have let T range over the set $\{\delta^n T_{min}\}_{n=0}^{10}$ where $\delta = (T_{max}/T_{min})^{1/10}$. The time units used for these Fourier analysis are revolutions of the secondary (J) around the primary (I), or equivalently, adimensional time divided by 2π. The reason for this is that, in this way, the frequency 1.0 corresponds to one revolution of J around I, which has a more intuitive meaning (one lunar month in the Earth–Moon case, one sidereal year in the Sun–Earth+Moon case). Moreover, in order to evaluate the trigonometric approximations of the c_i functions, we only have to multiply the frequencies found by adimensional time, without the need of an additional 2π factor. For instance, T_{min} is equal to 7979.72 in the Earth–Moon case, which means that during this time span the Moon has given 7979.72 revolutions around the Earth. For the Sun–Earth+Moon case, $T_{min} = 596.891$ (see Appendix A.4 for the details).

The maximum number of samples N_{max} has been chosen to be 2^{20}, in order to allow for "comfortable" runs on machines with 64MB of memory (or, equivalently, bi–processor machines with 128MB). For each value of T, the minimum number of samples has been chosen such that $\frac{T}{2N} \geq 1.5$, in order to make the maximum detectable frequency to be at least 1.5.

In order to control aliasing, two different strategies have been followed. The first one is based on time–domain, and consists in computing the difference between the initial function and its quasi-periodic approximation over a refinement of the grid used for the Fourier analysis. This difference will be denoted as α_1. If it increases significantly with respect to the difference over the Fourier samples, then aliasing is very likely to occur. We usually take $16N$ points equally spaced on $[0, T]$ for this test.

The second anti–aliasing strategy is based on frequency–domain. It consists in computing the number of rightmost consecutive harmonics of the residual DFT that have modulus less than a fraction of the maximum modulus of the residual DFT. Then, we divide this number by $N/2$, the total number of harmonics. That is, we compute

$$\alpha_2 = \frac{\max\{j : p^{n_h}_{c_i-q,T,N}(i) \leq p_{max}/25 \text{ for } i = j \div N/2\}}{N/2}$$

being $p_{max} := \max_{j=0\div N/2} p^{n_h}_{c_i-q,T,N}(j)$, where $p^{n_h}_{c_i-q,T,N}(j)$ is defined as in Section 1.3, c_i is the analyzed function and q its quasi–periodic approximation. Then, for instance, a value of 0.2 for α_1 means that there are no frequencies greater than $0.8\omega_{max}$, being $\omega_{max} = \frac{N}{2T}$, with amplitude greater than 1/25 times the modulus of the residual DFT, so we do not expect aliasing in the corresponding Fourier analysis. We are assuming here that amplitudes decrease as frequencies increase, which is ensured by the Cauchy estimates (3.3) for an analytic quasi–periodic function.

As an example of aliasing and how the two previously–described strategies detect it, we have represented in Fig. 5.3 the residual DFT of some of the Fourier analysis of the c_1 function in the Earth-Moon case. We give some details about these analysis in Table 5.1. In the left plot, we see that for $N = 16384$ there are frequencies of high amplitude near $\omega_{max} = \frac{T}{2N} = \frac{16384}{2\times 2033.24} = 4.02903$. As we increase N, the amplitude of the frequencies near ω_{max} decrease and the values of d_{max} and the first anti–aliasing strategy of Table 5.1 become closer.

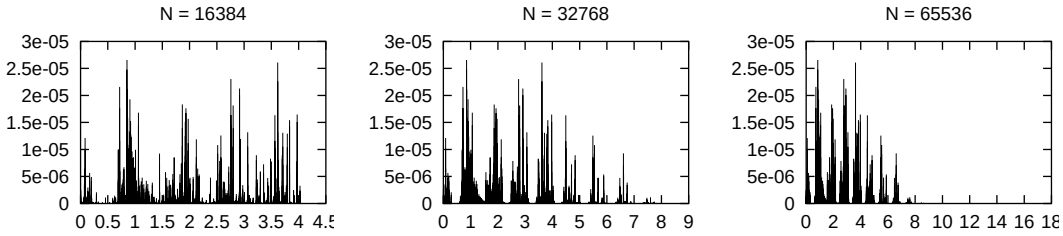


Figure 5.3: Modulus of the residual DFT some of the Fourier analysis of the c_1 function in the Earth–Moon case. From left to right, the values ω_{max} of the right–end of the DFT window are: 4.02903, 8.05806 and 16.1161.

day_0	day_f	T	N	p_{max}	d_{max}	α_1	α_2
366	55917.4	2033.24	16384	2.66E–05	4.90E–04	2.29E–03	0.0007
366	55917.4	2033.24	32768	2.66E–05	5.30E–04	5.67E–04	0.1633
366	55917.4	2033.24	65536	2.66E–05	5.63E–04	5.67E–04	0.5816

Table 5.1: Parameters associated to the Fourier analysis of Fig. 5.3. From left to right: day_0 and day_f are the starting and ending Julian days of the time interval used for each Fourier analysis, taking Jan 1st, 2001 as origin, T is the length of the Fourier interval, in J–revolutions, N is the number of points used, p_{max} is the maximum modulus of the residual DFT, d_{max} is the maximum difference between c_1 and its quasi–periodic approximation over the Fourier analysis samples, and α_1, α_2 are the values of the two anti–aliasing strategies described in the text.

According to this, for the results displayed in figures 5.1 and 5.2 only those analysis with $\alpha \geq 0.2$ have been taken into account.

For the generation of simplified models for the Solar System, among all the analysis performed we have selected the best ones in terms of minimum p_{max}. They are given in tables 5.2 (Earth–Moon) and 5.3 (Sun–Earth+Moon).

5.2.2 Fourier analysis of the positions of the planets

In order to complete the quasi–periodic approximation of all the time-dependent part in the vector–field (5.3), we give in this section the results of the Fourier analysis of the positions of the Solar System bodies in adimensional coordinates. For each coordinate x_{P_i}, y_{P_i}, z_{P_i}, we have performed Fourier analysis using the same parameters as for the analysis of the c_i functions. The minimum value of p_{max} with respect to N for fixed values of T is plotted in figures 5.4 (Earth–Moon) and 5.6 (Sun–Earth+Moon). The best analysis are given in tables 5.4 and 5.5.

function	T (days)	T (years)	T (J–rev.)	N	p_{max}	d_{max}
c_1	55551.4	152.091	2033.24	65536	2.66E–05	5.63E–04
c_2	55551.4	152.091	2033.24	65536	2.67E–05	5.49E–04
c_3	55551.4	152.091	2033.24	32768	3.30E–06	5.58E–05
c_4	55551.4	152.091	2033.24	65536	2.31E–06	5.01E–05
c_5	43904.0	120.203	1606.94	32768	4.85E–06	9.16E–05
c_6	70288.7	192.440	2572.64	32768	3.92E–08	1.13E–06
c_7	55551.4	152.091	2033.24	65536	3.51E–06	7.81E–05
c_8	55551.4	152.091	2033.24	524288	1.96E–07	5.94E–06
c_9	70288.7	192.440	2572.64	65536	1.97E–08	5.69E–07
c_{10}	55551.4	152.091	2033.24	65536	3.51E–06	7.83E–05
c_{11}	70288.7	192.440	2572.64	65536	1.67E–08	5.05E–07
c_{12}	43904.0	120.203	1606.94	32768	1.58E–06	3.29E–05
c_{13}	55551.4	152.091	2033.24	65536	3.51E–06	7.99E–05

Table 5.2: Values of the parameters for the best Fourier analyses of the c_i functions for the Earth–Moon case.

function	T (days)	T (J–rev)	N	p_{max}	d_{max}
c_1	142382.6	389.815	65536	4.95E–08	4.40E–07
c_2	142382.6	389.815	65536	4.95E–08	4.33E–07
c_3	112529.5	308.083	131072	2.28E–09	2.68E–08
c_4	34698.8	94.998	4096	8.34E–06	6.74E–05
c_5	34698.8	94.998	4096	1.75E–05	1.26E–04
c_6	88935.7	243.488	262144	1.76E–08	5.71E–07
c_7	34698.8	94.998	4096	1.36E–05	9.17E–05
c_8	288422.1	789.642	524288	9.65E–08	1.67E–06
c_9	88935.7	243.488	131072	9.71E–09	3.19E–07
c_{10}	34698.8	94.998	4096	1.36E–05	9.17E–05
c_{11}	70288.7	192.436	524288	2.35E–08	2.38E–06
c_{12}	34698.8	94.998	4096	3.92E–06	4.06E–05
c_{13}	34698.8	94.998	4096	1.34E–05	9.47E–05

Table 5.3: Values of the parameters for the best Fourier analyses of the c_i functions for the Sun–Earth+Moon case. Note that, in this case, J–revolutions are sidereal years.

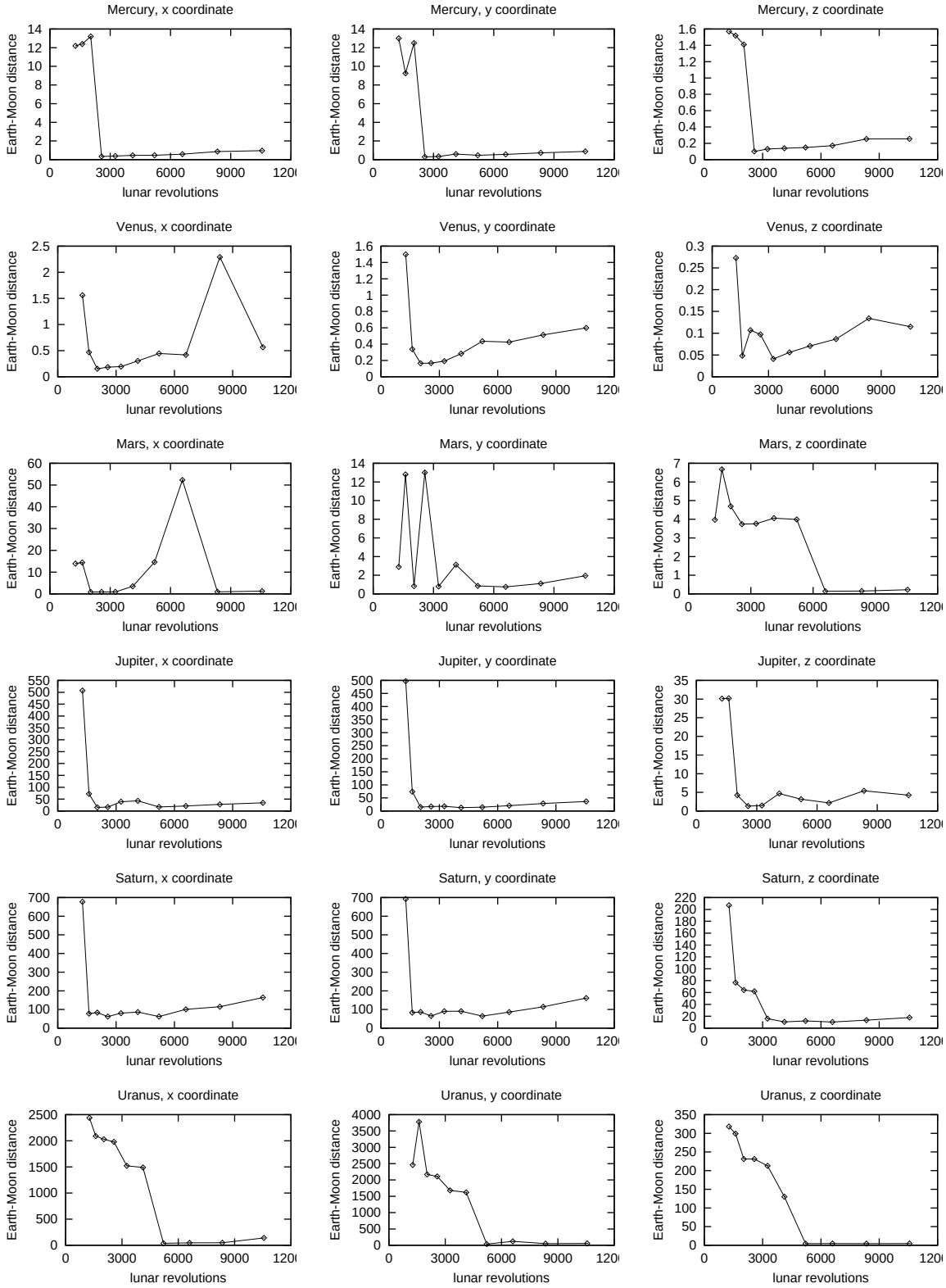


Figure 5.4: Error results of the Fourier analysis of the coordinates of the Solar System bodies (in adimensional coordinates) for the Earth–Moon case. For each value of T explored, we have represented the minimum value of d_{max} with respect to N. They are continued in Fig. 5.5.

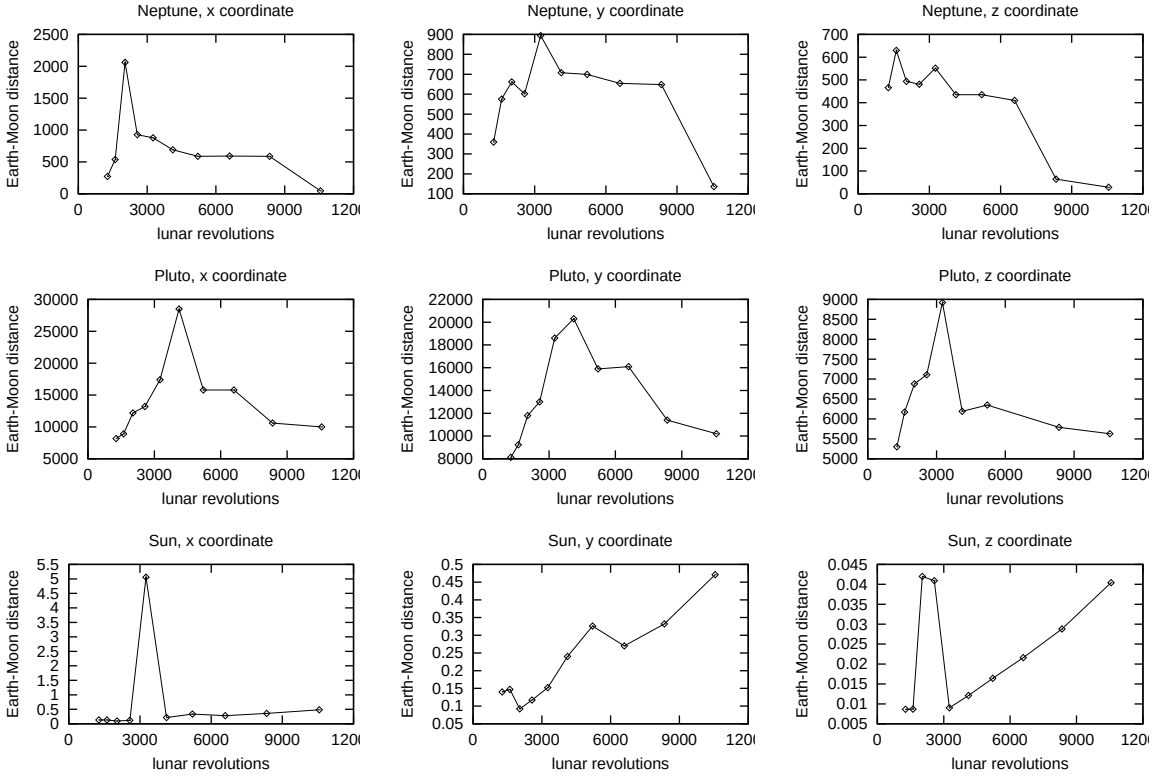


Figure 5.5: Continuation of Fig. 5.4.

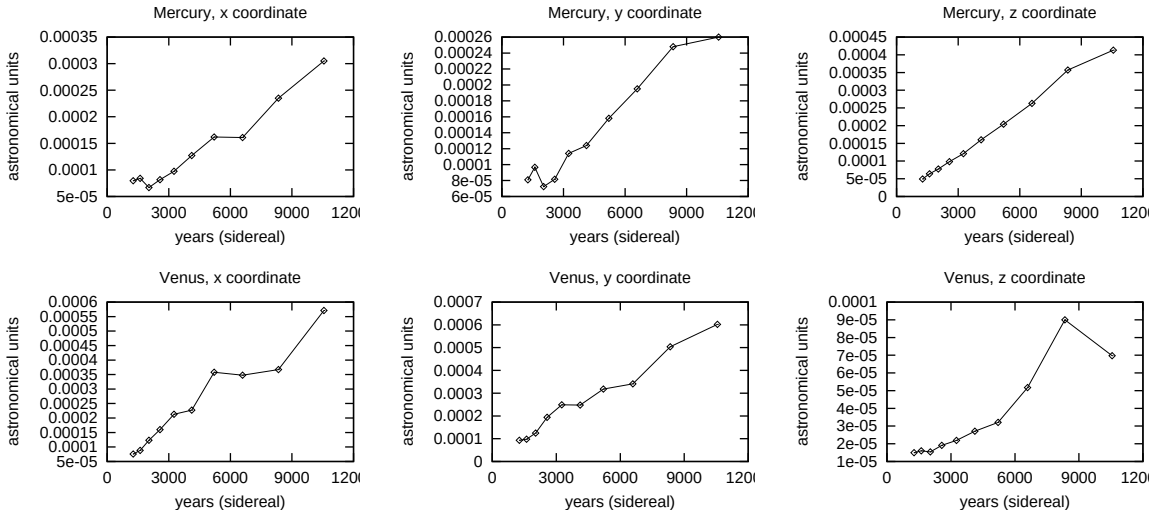


Figure 5.6: Same as Fig. 5.4 but for the Sun–Earth+Moon case (continued in Fig. 5.7).

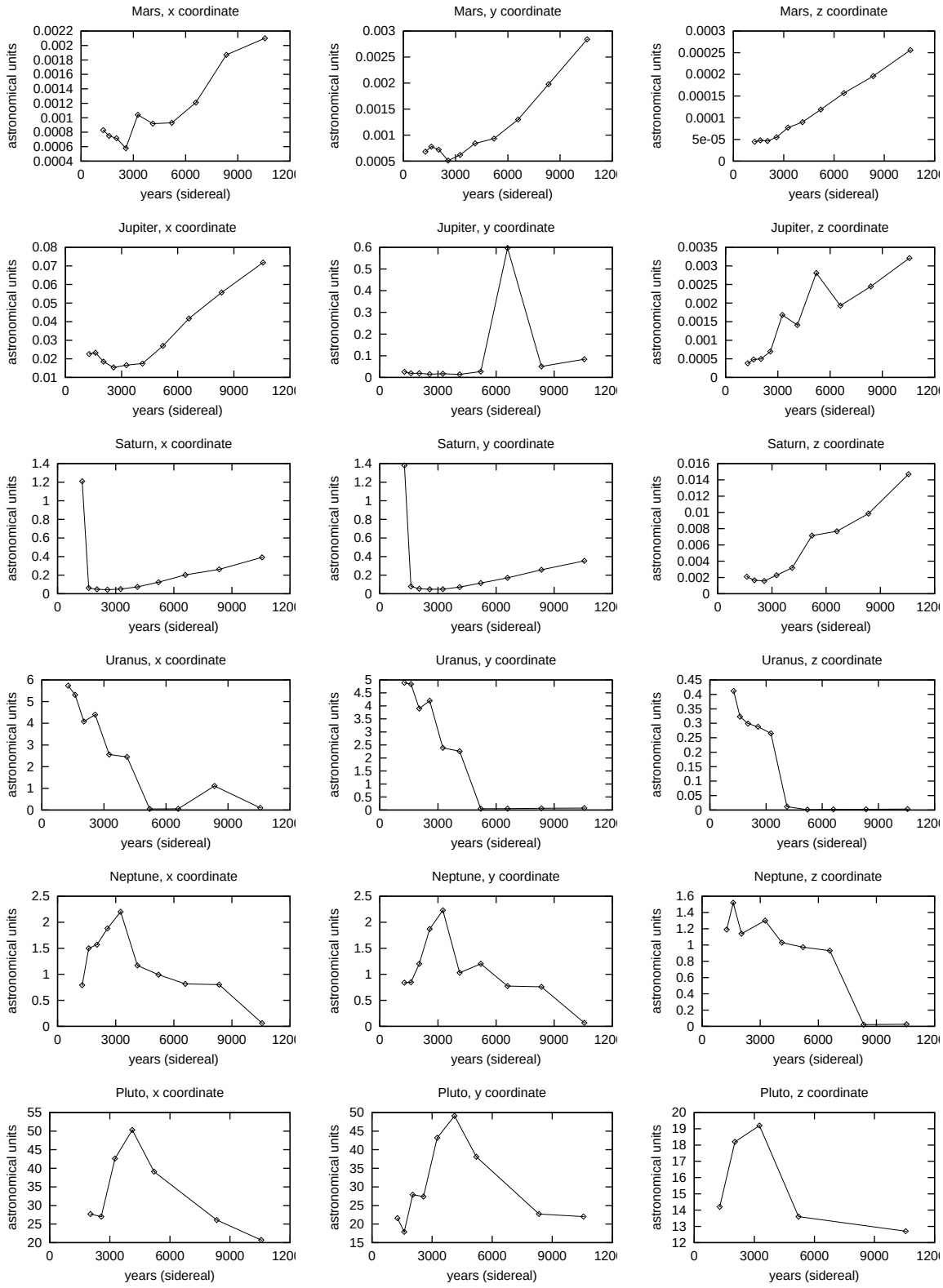


Figure 5.7: Continuation of Fig. 5.6.

body	coord.	T (days)	T (years)	T (J–rev)	N	p_{max}	d_{max}
Mercury	x	70288.7	192.440	2572.64	65536	1.37E–02	3.41E–01
Mercury	y	70288.7	192.440	2572.64	65536	1.08E–02	2.89E–01
Mercury	z	70288.7	192.440	2572.64	32768	3.18E–03	9.99E–02
Venus	x	55551.4	152.091	2033.24	65536	5.13E–03	1.53E–01
Venus	y	55551.4	152.091	2033.24	65536	5.60E–03	1.65E–01
Venus	z	88935.7	243.493	3255.14	65536	1.25E–03	4.10E–02
Mars	x	55551.4	152.091	2033.24	65536	3.61E–02	8.43E–01
Mars	y	180155.5	493.239	6593.89	131072	3.21E–02	7.53E–01
Mars	z	180155.5	493.239	6593.89	131072	3.26E–03	1.38E–01
Jupiter	x	55551.4	152.091	2033.24	32768	1.40E+00	1.53E+01
Jupiter	y	112529.5	308.089	4118.71	65536	5.39E–01	1.31E+01
Jupiter	z	70288.7	192.440	2572.64	32768	1.37E–01	1.31E+00
Saturn	x	70288.7	192.440	2572.64	32768	6.07E+00	6.19E+01
Saturn	y	142382.6	389.822	5211.36	65536	2.53E+00	6.46E+01
Saturn	z	180155.5	493.239	6593.89	65536	3.87E–01	1.04E+01
Uranus	x	142382.6	389.822	5211.36	131072	2.33E+00	3.75E+01
Uranus	y	142382.6	389.822	5211.36	131072	2.33E+00	3.76E+01
Uranus	z	364938.0	999.146	13357.14	131072	2.42E–01	4.14E+00
Neptune	x	288422.1	789.657	10556.57	262144	3.12E+00	4.52E+01
Neptune	y	364938.0	999.146	13357.14	262144	2.37E+00	4.51E+01
Neptune	z	364938.0	999.146	13357.14	131072	1.80E+00	2.72E+01
Pluto	x	364938.0	999.146	13357.14	262144	4.15E+00	1.69E+02
Pluto	y	364938.0	999.146	13357.14	262144	2.08E+01	2.93E+02
Pluto	z	364938.0	999.146	13357.14	131072	2.42E+00	5.16E+01
Sun	x	55551.4	152.091	2033.24	65536	4.41E–03	9.73E–02
Sun	y	55551.4	152.091	2033.24	65536	4.41E–03	9.21E–02
Sun	z	34698.8	95.000	1270.01	16384	8.49E–04	8.65E–03

Table 5.4: Best Fourier analysis parameters for the positions of the Solar System bodies in adimensional coordinates in the Earth–Moon case.

body	coord.	T (days)	T (J–rev)	N	p_{max}	d_{max}
Mercury	x	55551.4	152.089	16384	6.56E–06	6.68E–05
Mercury	y	55551.4	152.089	16384	6.56E–06	7.24E–05
Mercury	z	34698.8	94.998	8192	1.64E–06	4.93E–05
Venus	x	34698.8	94.998	4096	7.61E–06	7.57E–05
Venus	y	34698.8	94.998	4096	7.61E–06	9.32E–05
Venus	z	34698.8	94.998	4096	1.93E–06	1.50E–05
Mars	x	70288.7	192.436	8192	4.87E–05	5.80E–04
Mars	y	70288.7	192.436	8192	4.87E–05	5.11E–04
Mars	z	34698.8	94.998	4096	3.00E–06	4.48E–05
Jupiter	x	70288.7	192.436	8192	3.56E–03	1.54E–02
Jupiter	y	112529.5	308.083	16384	9.35E–04	1.41E–02
Jupiter	z	34698.8	94.998	4096	7.68E–05	3.82E–04
Saturn	x	70288.7	192.436	8192	1.29E–02	4.29E–02
Saturn	y	70288.7	192.436	8192	1.29E–02	4.82E–02
Saturn	z	70288.7	192.436	8192	5.39E–04	1.57E–03
Uranus	x	142382.6	389.815	16384	5.57E–03	4.82E–02
Uranus	y	142382.6	389.815	16384	5.57E–03	5.10E–02
Uranus	z	142382.6	389.815	16384	2.45E–04	1.25E–03
Neptune	x	288422.1	789.642	32768	5.40E–03	6.15E–02
Neptune	y	288422.1	789.642	32768	5.41E–03	6.71E–02
Neptune	z	227949.2	624.079	32768	4.10E–03	2.10E–02
Pluto	x	364938.0	999.127	65536	1.27E–02	2.92E–01
Pluto	y	364938.0	999.127	65536	1.41E–02	3.72E–01
Pluto	z	364938.0	999.127	32768	6.19E–03	7.64E–02

Table 5.5: Best Fourier analysis parameters for the positions of the Solar System bodies in adimensional coordinates in the Sun–Earth+Moon case.

5.3 Generation of simplified Solar System models

In this section we will develop simplified Solar System models based on the Fourier expansions computed in the previous section. The models obtained will be compared with other models though the computation of residual accelerations along selected orbits.

5.3.1 Adjustment by linear combinations of basic frequencies

In order to turn the output of our Fourier analysis procedures into the usual form of a quasi–periodic function (3.1), we need to adjust frequencies as linear combinations, with integer coefficients, of basic ones. We will distinguish two cases:

- the case in which we do not know the basic frequencies, which need to be extracted from the list of frequencies to be adjusted, and
- the case in which the basic frequencies are known.

A simple approach for the first case would be: choose a maximum order of the linear combinations to be found, and a tolerance for the adjustment of frequencies as linear combination of the basic ones. Then, for each frequency, try out all the linear combinations of the current set of basic frequencies up the chosen maximum order. If one of these linear combinations fulfills the requirements, take it, otherwise add the current frequency to the set of basic frequencies.

This procedure may add extra basic frequencies (and thus end up with a rationally dependent set) in some cases, for instance, if the current frequency is an integer divisor of one of the basic frequencies. In order to avoid this, when the current frequency cannot be adjusted as a linear combination of the current basis, for each frequency in the current basis we can try to substitute it by the non–adjusted one and see if all the pre–processed frequencies adjust to this modified basis. If this is not the case, the new frequency is added to the basic set.

These considerations lead to the following

Algorithm 5.3.1 *Given* $\{f_1, \ldots, f_{N_f}\}$ *the set of frequencies to be adjusted as linear combination of basic ones to be selected in the set, a tolerance* tol *for the adjustments and a maximum order* maxor *for the linear combinations to be found, compute the basis* $\{\omega_1, \ldots, \omega_{n_b}\}$ *and the linear combinations* $\{(k_1^i, \ldots, k_{n_b}^i)\}_{i=1 \div N_f}$ *as*

$\omega_1 \leftarrow f_1$, $k_1^1 \leftarrow 1$, $n_b \leftarrow 1$
for $i = 2 \div N_f$
  if $f_i \in \mathrm{lc}(\{\omega_l\}_l, \mathrm{tol}, \mathrm{maxor})$
    $(k_1^i, \ldots, k_{n_b}^i) = \mathrm{adjust}(f_i, \{\omega_l\}_l, \mathrm{tol}, \mathrm{maxor})$
  else
    if $\exists j \in \{1, \ldots, n_b\} : f_1, \ldots, f_i \in \mathrm{lc}(\{\omega_1, \ldots, \overset{(j}{f_i}, \ldots, \omega_{N_f}\}, \mathrm{tol}, \mathrm{maxor})$
      $\omega_j \leftarrow f_i$
      for $l = 1 \div i$
        $(k_1^l, \ldots, k_{n_b}^l) = \mathrm{adjust}(f_i, \{\omega_m\}_m, \mathrm{tol}, \mathrm{maxor})$

else

$n_b \leftarrow n_b + 1$
$(k_1^i, \ldots k_{n_b}^i) = (0, \ldots, 0, 1)$
for $l = 1 \div i - 1$
$k_{n_b}^l = 0$

In this formulation, we have introduced two functions lc *and* adjust*, defined as follows:*

- $\mathrm{lc}(\{\omega_i\}_{i=1\div n_b}, \mathrm{tol}, \mathrm{maxor})$ *is defined as the set of real numers* f *such that there exists* $(k_1, \ldots, k_{n_b})$ *with* k_i *integer,* $|k_1| + \ldots + |k_{n_b}| \leq \mathrm{maxor}$ *and* $|f - k_1\omega_1 - \ldots - k_{n_b}\omega_{n_b}| \leq \mathrm{tol}$*,*

- *for* f *real,* $\mathrm{adjust}(f, \{\omega_i\}_{i=1\div n_b}, \mathrm{tol}, \mathrm{maxor})$ *returns the first* $(k_1, \ldots, k_{n_b})$*, in increasing order and increasing lexicographical order within each order, with order* $\leq \mathrm{maxor}$*, such that* $|f - k_1\omega_1 - \ldots - k_{n_b}\omega_{n_b}| \leq \mathrm{tol}$*. In the case that there is no* $(k_1, \ldots, k_{n_b})$ *of order less than* maxor *with* $|f - k_1\omega_1 - \ldots - k_{n_b}\omega_{n_b}| \leq \mathrm{tol}$*, the one with minimum* $|f - k_1\omega_1 - \ldots - k_{n_b}\omega_{n_b}|$ *is returned.*

Of course, in an actual implementation the role of these functions is accomplished by the same code.

In the second case, in which the basic frequencies $\{\omega_1, \ldots, \omega_{n_b}\}$ are know, we can just take the best linear combination for each frequency. This can be stated as

Algorithm 5.3.2 *Given* $\{f_1, \ldots, f_{N_f}\}$ *the set of frequencies to be adjusted as linear combination of the frequency basis* $\{\omega_1, \ldots, \omega_{n_b}\}$*, a tolerance* tol *for the adjustments and a maximum order* maxor *for the linear combinations to be found, compute the linear combinations* $\{(k_1^i, \ldots, k_{n_b}^i)\}_{i=1\div N_f}$ *as*

for $i = 1 \div N_f$
$(k_1^i, \ldots, k_{n_b}^i) = \mathrm{adjust}(f_i, \{\omega_l\}_l, \mathrm{tol}, \mathrm{maxor})$

5.3.2 Simplified models for the Earth–Moon case

In a rather accurate theory for the lunar motion, as the simplified Brown theory given in [8], the fundamental parameters can be expressed in terms of five basic frequencies:

- The mean longitude of the Moon, which is equal to 1.0.

- The mean elongation of the Moon from the Sun, 0.925195997455093. This is the frequency of the time-dependent part in the Bicircular Problem (BCP) and the Quasi–Bicircular Problem, (QBCP, see Appendix A).

- The mean longitude of the lunar perigee, which is equal to $8.45477852931292 \times 10^{-3}$.

- The longitude of the mean ascending node of the lunar orbit on the ecliptic, $4.01883841204748 \times 10^{-3}$.

- the Sun's mean longitude of perigee, $3.57408131981537 \times 10^{-6}$.

The units used for these frequencies are cycles per lunar revolution. In what follows, these frequencies will be denoted $\{\omega_1, \ldots, \omega_5\}$.

The value of the last frequency in the above set is close to the lower amplitudes of our Fourier expansions, which is close to the precision we can expect in the determination of frequencies, since all Fourier analysis have stopped due to the detection of too close frequencies. In order to avoid the difficulties due this fact, and in order to have a set of basic frequencies with astronomical meaning, we have adopted these frequencies as the basic set, instead of the ones provided by Algorithm 5.3.1.

For the simplified models to be developed in this section, we will only take into account the coordinates of the Sun in (5.4). This avoids the introduction of additional basic frequencies, and is also enough for our purposes, as it will become clear later. In this way, we will only use the Fourier expansions of $c_1, \ldots, c_{13}$ and x_S, y_S, z_S.

Starting from the frequency basis $\{\omega_i\}_{i=1\div 5}$, we will look for a new basis $\{\nu_i\}_{i=1\div 5}$. In terms of this new basis, we will generate 5 models SSSM_i, $i = 1 \div 5$, in such a way that the equations of motion of SSSM_i are

$$\left\{\begin{array}{rcl} \ddot{x} &=& c_1^i + c_4\dot{x} + c_5^i\dot{y} + c_7^i x + c_8^i y + c_9^i z + c_{13}^i \dfrac{\partial \Omega^i}{\partial x} \\ \ddot{y} &=& c_2^i - c_5^i\dot{x} + c_4^i\dot{y} + c_6^i\dot{z} - c_8^i x + c_{10}^i y + c_{11}^i z + c_{13} \dfrac{\partial \Omega^i}{\partial y} \\ \ddot{z} &=& c_3^i - c_6\dot{y} + c_4^i\dot{z} + c_9^i x - c_{11}^i y + c_{12}^i z + c_{13}^i \dfrac{\partial \Omega^i}{\partial z} \end{array}\right.$$

being

$$\begin{array}{rcl} \Omega &=& \dfrac{1-\mu_{E,M}}{\sqrt{(x-\mu_{E,M})^2+y^2+z^2}} + \dfrac{\mu_{E,M}}{\sqrt{(x-\mu_{E,M}+1)^2+y^2+z^2}} \\ && + \dfrac{\mu_{E,M,S}}{\sqrt{(x-x_S^i)^2+(y-y_S^i)^2+(z-z_S^i)^2}}. \end{array}$$

Here c_j^i, $j = 1 \div 13$ and x_S^i, y_S^i, z_S^i stand for their Fourier expansions, computed in the previous section, but keeping only the frequencies that are expressed as linear combinations (with integer coefficients) of the frequencies $\nu_1, \ldots, \nu_i$.

We have used Algorithm 5.3.2 of the previous section with $tol = 10^{-6}$ and $maxor = 20$ to adjust the frequencies found in the analysis of table 5.2 as linear combinations of the $\{\omega_i\}_{i=1\div 5}$. The results for the first 15 frequencies detected in each c_i and x_S, y_S, z_S are shown in tables 5.6 to 5.16. The full expansions are given in Appendix C.

We will take $\nu_1 = \omega_2$ as the first frequency of our new basis. The reason for that is that it is the main frequency of c_1, c_2, x_S and y_S, and in this way it can be considered the main "planar frequency". This is coherent with the fact that ω_2 is also the frequency of the BCP and QBCP models (see Appendix A).

We observe that, except for c_3, c_6, c_9, c_{11} and z_S, the main frequencies of the remaining functions can be expressed as linear combinations of ω_2 and $\omega_1 - \omega_3$. Thus, we will take $\nu_2 = \omega_1 - \omega_3$. Note that, in this way, c_i for $i = 3, 6, 9, 11$ and z_S will be poorly approximated in SSSM_2, but this will not give a poor global approximation because c_i for $i = 3, 6, 9, 11$ are smaller than the remaining c_i, and z_S is also smaller than x_S, y_S.

freq	ampl	err	k_1	k_2	k_3	k_4	k_5	order
0.00000000000	3.49728E–04	0.00000E+00	0	0	0	0	0	0
0.92519578630	2.16240E+00	–2.11120E–07	0	1	0	0	0	1
1.91674083000	1.77450E–01	–3.88880E–07	1	1	–1	0	0	3
0.85039537680	7.53250E–02	–1.92240E–07	–1	2	0	0	1	4
0.06634926290	7.39730E–02	3.88600E–08	1	–1	–1	0	0	3
1.78404231460	3.41170E–02	–4.56330E–07	–1	3	1	0	0	5
2.77558735940	2.39690E–02	–6.33010E–07	0	3	0	0	0	3
2.90828587990	1.36950E–02	–5.60520E–07	2	1	–2	0	0	5
1.84194060340	7.31080E–03	–1.87100E–07	0	2	–1	0	1	4
1.08284144950	3.91270E–03	–2.29900E–07	2	–1	0	2	0	5

Table 5.6: First 10 frequencies of the Fourier analysis of c_1. The frequencies have been adjusted as linear combinations of $\{\omega_i\}_{i=1\div5}$. From left to right the columns are: frequency, in cycles per lunar revolution, amplitude, error ($\mathrm{freq} - k_1\omega_1 - \ldots - k_5\omega_5$), coefficients of the linear combination that approximates freq, and order of the linear combination ($|k_1| + \ldots |k_5|$).

freq.	ampl.	err.	k_1	k_2	k_3	k_4	k_5	order
0.00000000000	–6.70000E–09	0.00000E+00	0	0	0	0	0	0
0.92519578630	2.16960E+00	–2.11120E–07	0	1	0	0	0	1
1.91674083000	1.77820E–01	–3.88890E–07	1	1	–1	0	0	3
0.85039537680	7.58320E–02	–1.92220E–07	–1	2	0	0	1	4
0.06634926260	4.64680E–02	3.85950E–08	1	–1	–1	0	0	3
1.78404231460	3.41920E–02	–4.56320E–07	–1	3	1	0	0	5
2.77558735930	2.39940E–02	–6.33040E–07	0	3	0	0	0	3
2.90828587990	1.37170E–02	–5.60520E–07	2	1	–2	0	0	5
1.84194060340	7.33860E–03	–1.87100E–07	0	2	–1	0	1	4
1.08284144950	3.95090E–03	–2.29860E–07	2	–1	0	2	0	5

Table 5.7: Same as table 5.6 but for the c_2 function.

freq.	ampl.	err.	k_1	k_2	k_3	k_4	k_5	order
0.00000000000	–1.41400E–07	0.00000E+00	0	0	0	0	0	0
0.07882283210	1.90520E–01	–8.87040E–09	1	–1	0	1	0	3
0.15362345870	6.56920E–03	1.89270E–07	2	–2	0	1	–1	6
0.91272221270	5.20890E–03	–1.67780E–07	0	1	–1	–1	0	3
1.07036787670	5.21170E–03	–1.85760E–07	2	–1	–1	1	0	5
0.78002369690	1.09620E–03	–2.35620E–07	–2	3	1	–1	0	7
0.93766935940	1.08610E–03	–2.54980E–07	0	1	1	1	0	3
1.92921440450	3.96610E–04	–4.31370E–07	1	1	0	1	0	3
1.77156873960	3.60950E–04	–4.14380E–07	–1	3	0	–1	0	5
0.00402218340	3.31660E–04	–2.29130E–07	0	0	0	1	1	2

Table 5.8: Same as table 5.6 but for the c_3 function.

freq.	ampl.	err.	k_1	k_2	k_3	k_4	k_5	order
0.00000000000	0.00000E+00	0.00000E+00	0	0	0	0	0	0
0.99154505160	1.07920E–01	–1.69890E–07	1	0	–1	0	0	2
1.85039157300	2.94710E–02	–4.21940E–07	0	2	0	0	0	2
0.85884652970	1.68610E–02	–2.43690E–07	–1	2	1	0	0	4
1.98309009370	8.82140E–03	–3.49210E–07	2	0	–2	0	0	4
2.84193661720	3.80000E–03	–5.99130E–07	1	2	–1	0	0	4
1.77559111020	1.93150E–03	–4.56240E–07	–1	3	0	0	1	5
2.97463513490	6.78820E–04	–5.29470E–07	3	0	–3	0	0	6
2.70923810000	7.14080E–04	–6.68310E–07	–1	4	1	0	0	6
0.78404586970	6.38340E–04	–4.75270E–07	–2	3	1	0	1	7

Table 5.9: Same as table 5.6 but for the c_4 function.

freq.	ampl.	err.	k_1	k_2	k_3	k_4	k_5	order
0.00000000000	2.00003E+00	0.00000E+00	0	0	0	0	0	0
0.99154503470	2.17650E–01	–1.86770E–07	1	0	–1	0	0	2
1.85039156830	4.29420E–02	–4.26650E–07	0	2	0	0	0	2
0.85884653190	3.81670E–02	–2.41550E–07	–1	2	1	0	0	4
1.98309007300	1.48070E–02	–3.69960E–07	2	0	–2	0	0	4
2.84193660360	5.36300E–03	–6.12800E–07	1	2	–1	0	0	4
1.77559105180	2.84910E–03	–5.14630E–07	–1	3	0	0	1	5
0.78404613980	1.55830E–03	–2.05220E–07	–2	3	1	0	1	7
0.91674466300	1.30720E–03	–1.30020E–07	0	1	–1	0	1	3
0.92519587730	1.12100E–03	–1.20200E–07	0	1	0	0	0	1

Table 5.10: Same as table 5.6 but for the c_5 function.

freq.	ampl.	err.	k_1	k_2	k_3	k_4	k_5	order
0.00000000000	0.00000E+00	0.00000E+00	0	0	0	0	0	0
0.84637295300	1.44550E–03	–2.03520E–07	–1	2	0	–1	0	4
1.00401861550	1.44530E–03	–2.22890E–07	1	0	0	1	0	2
0.01247357960	1.89340E–04	–3.72940E–08	0	0	1	1	0	2
0.14517208260	1.88980E–04	1.76520E–08	2	–2	–1	1	0	6
0.77157269570	8.78480E–05	–3.23250E–08	–2	3	0	–1	1	7
0.92921810820	3.54680E–05	–3.01790E–07	0	1	0	1	1	3
1.07881913160	3.51310E–05	–1.35230E–07	2	–1	0	1	–1	5
0.92117316420	1.72620E–05	–4.20780E–07	0	1	0	–1	–1	3
0.21997272480	1.09360E–05	2.31350E–07	3	–3	–1	1	–1	9

Table 5.11: Same as table 5.6 but for the c_6 function.

freq.	ampl.	err.	k_1	k_2	k_3	k_4	k_5	order
0.00000000000	1.00478E+00	0.00000E+00	0	0	0	0	0	0
0.99154504270	1.65040E–01	–1.78730E–07	1	0	–1	0	0	2
0.85884652970	3.24780E–02	–2.43700E–07	–1	2	1	0	0	4
1.85039157280	1.84070E–02	–4.22070E–07	0	2	0	0	0	2
1.98309009370	1.35090E–02	–3.49200E–07	2	0	–2	0	0	4
2.84193661730	3.29470E–03	–5.99110E–07	1	2	–1	0	0	4
0.13269851610	1.45030E–03	6.80760E–08	2	–2	–2	0	0	6
0.78404586980	1.39870E–03	–4.75200E–07	–2	3	1	0	1	7
1.77559111010	1.25920E–03	–4.56320E–07	–1	3	0	0	1	5
2.97463513460	1.08280E–03	–5.29830E–07	3	0	–3	0	0	6

Table 5.12: Same as table 5.6 but for the c_7 function.

freq.	ampl.	err.	k_1	k_2	k_3	k_4	k_5	order
0.00000000000	–7.00000E–10	0.00000E+00	0	0	0	0	0	0
1.85039159880	8.24730E–03	–3.96070E–07	0	2	0	0	0	2
2.84193667480	9.04550E–04	–5.41620E–07	1	2	–1	0	0	4
0.85884652020	9.17510E–04	–2.53210E–07	–1	2	1	0	0	4
1.77559103310	5.07100E–04	–5.33340E–07	–1	3	0	0	1	5
0.99154507800	1.95970E–04	–1.43510E–07	1	0	–1	0	0	2
2.70923811510	1.73420E–04	–6.53300E–07	–1	4	1	0	0	6
3.70078319630	1.14930E–04	–7.93520E–07	0	4	0	0	0	4
1.92519194710	9.57130E–05	–4.76300E–07	1	1	0	0	–1	3
3.83348174850	8.00900E–05	–6.89320E–07	2	2	–2	0	0	6

Table 5.13: Same as table 5.6 but for the c_8 function.

freq.	ampl.	err.	k_1	k_2	k_3	k_4	k_5	order
0.00000000000	–0.00000E+00	0.00000E+00	0	0	0	0	0	0
0.84637295300	7.24530E–04	–2.03520E–07	–1	2	0	–1	0	4
1.00401861550	7.24450E–04	–2.22890E–07	1	0	0	1	0	2
0.01247357980	4.82170E–05	–3.71330E–08	0	0	1	1	0	2
0.14517208260	4.80940E–05	1.76380E–08	2	–2	–1	1	0	6
0.77157269560	4.41510E–05	–3.24260E–08	–2	3	0	–1	1	7
1.99556365120	4.00140E–05	–4.08710E–07	2	0	–1	1	0	4
1.83791798820	3.99960E–05	–3.89770E–07	0	2	–1	–1	0	4
0.92921810800	1.78950E–05	–3.01960E–07	0	1	0	1	1	3
1.07881913140	1.76710E–05	–1.35470E–07	2	–1	0	1	–1	5

Table 5.14: Same as table 5.6 but for the c_9 function.

freq.	ampl.	err.	k_1	k_2	k_3	k_4	k_5	order
0.00000000000	1.00478E+00	0.00000E+00	0	0	0	0	0	0
0.99154504270	1.65030E–01	–1.78730E–07	1	0	–1	0	0	2
0.85884652970	3.24780E–02	–2.43700E–07	–1	2	1	0	0	4
1.85039157280	1.84070E–02	–4.22070E–07	0	2	0	0	0	2
1.98309009370	1.35090E–02	–3.49200E–07	2	0	–2	0	0	4
2.84193661730	3.29470E–03	–5.99110E–07	1	2	–1	0	0	4
0.13269851610	1.45030E–03	6.80760E–08	2	–2	–2	0	0	6
0.78404586980	1.39870E–03	–4.75200E–07	–2	3	1	0	1	7
1.77559111010	1.25920E–03	–4.56320E–07	–1	3	0	0	1	5
2.97463513460	1.08280E–03	–5.29830E–07	3	0	–3	0	0	6

Table 5.15: Same as table 5.6 but for the c_{10} function.

freq.	ampl.	err.	k_1	k_2	k_3	k_4	k_5	order
0.00000000000	–0.00000E+00	0.00000E+00	0	0	0	0	0	0
1.00401861560	7.20820E–04	–2.22850E–07	1	0	0	1	0	2
0.84637295300	6.06950E–04	–2.03500E–07	–1	2	0	–1	0	4
0.14517208280	4.66020E–05	1.78760E–08	2	–2	–1	1	0	6
1.99556364910	3.64300E–05	–4.10800E–07	2	0	–1	1	0	4
1.83791798780	3.65390E–05	–3.90160E–07	0	2	–1	–1	0	4
0.77157269620	3.33090E–05	–3.18310E–08	–2	3	0	–1	1	7
0.01247358100	3.15620E–05	–3.59500E–08	0	0	1	1	0	2
2.85441018260	2.40050E–05	–6.50710E–07	1	2	0	1	0	4
2.69676451900	2.36110E–05	–6.32380E–07	–1	4	0	–1	0	6

Table 5.16: Same as table 5.6 but for the c_{11} function.

freq.	ampl.	err.	k_1	k_2	k_3	k_4	k_5	order
0.00000000000	–1.61183E–03	0.00000E+00	0	0	0	0	0	0
0.99154502640	5.38970E–02	–1.95110E–07	1	0	–1	0	0	2
1.85039157030	2.69200E–02	–4.24600E–07	0	2	0	0	0	2
0.85884654110	8.04870E–03	–2.32340E–07	–1	2	1	0	0	4
1.98309004860	7.32970E–03	–3.94350E–07	2	0	–2	0	0	4
2.84193659510	4.58070E–03	–6.21280E–07	1	2	–1	0	0	4
1.77559129770	1.70370E–03	–2.68790E–07	–1	3	0	0	1	5
2.70923811570	8.46260E–04	–6.52680E–07	–1	4	1	0	0	6
2.97463506060	7.75820E–04	–6.03790E–07	3	0	–3	0	0	6
3.70078314150	5.70720E–04	–8.48320E–07	0	4	0	0	0	4

Table 5.17: Same as table 5.6 but for the c_{12} function.

freq.	ampl.	err.	k_1	k_2	k_3	k_4	k_5	order
0.00000000000	1.00747E+00	0.00000E+00	0	0	0	0	0	0
0.99154504270	1.64840E–01	–1.78730E–07	1	0	–1	0	0	2
0.85884652970	3.15620E–02	–2.43700E–07	–1	2	1	0	0	4
1.85039157290	2.66550E–02	–4.22010E–07	0	2	0	0	0	2
1.98309009370	1.34800E–02	–3.49210E–07	2	0	–2	0	0	4
2.84193661730	4.19930E–03	–5.99110E–07	1	2	–1	0	0	4
1.77559111020	1.76690E–03	–4.56280E–07	–1	3	0	0	1	5
0.13269851610	1.47040E–03	6.80410E–08	2	–2	–2	0	0	6
0.78404586980	1.34660E–03	–4.75200E–07	–2	3	1	0	1	7
2.97463513490	1.07950E–03	–5.29490E–07	3	0	–3	0	0	6

Table 5.18: Same as table 5.6 but for the c_{13} function.

freq.	ampl.	err.	k_1	k_2	k_3	k_4	k_5	order
0.00000000000	–6.27023E–02	0.00000E+00	0	0	0	0	0	0
0.92519578630	3.86480E+02	–2.11130E–07	0	1	0	0	0	1
1.91674083000	3.17140E+01	–3.88890E–07	1	1	–1	0	0	3
0.06634926280	1.32180E+01	3.87440E–08	1	–1	–1	0	0	3
0.99999608230	1.03360E+01	–3.43580E–07	1	0	0	0	–1	2
1.78404231420	6.09200E+00	–4.56710E–07	–1	3	1	0	0	5
2.77558735980	4.27790E+00	–6.32540E–07	0	3	0	0	0	3
0.85039537680	3.79560E+00	–1.92230E–07	–1	2	0	0	1	4
2.90828587090	2.44750E+00	–5.69500E–07	2	1	–2	0	0	5
1.99154129500	1.00420E+00	–3.52350E–07	2	0	–1	0	–1	4

Table 5.19: Same as table 5.6 but for the x_S function.

freq.	ampl.	err.	k_1	k_2	k_3	k_4	k_5	order
0.00000000000	1.60785E–05	0.00000E+00	0	0	0	0	0	0
0.92519578630	3.87760E+02	–2.11130E–07	0	1	0	0	0	1
1.91674083000	3.17800E+01	–3.88890E–07	1	1	–1	0	0	3
0.99999608230	1.03360E+01	–3.43590E–07	1	0	0	0	–1	2
0.06634926280	8.30700E+00	3.87360E–08	1	–1	–1	0	0	3
1.78404231280	6.10530E+00	–4.58110E–07	–1	3	1	0	0	5
2.77558735590	4.28220E+00	–6.36440E–07	0	3	0	0	0	3
0.85039537680	3.85420E+00	–1.92230E–07	–1	2	0	0	1	4
2.90828587990	2.45150E+00	–5.60490E–07	2	1	–2	0	0	5
1.99154129500	1.00460E+00	–3.52360E–07	2	0	–1	0	–1	4

Table 5.20: Same as table 5.6 but for the y_S function.

freq.	ampl.	err.	k_1	k_2	k_3	k_4	k_5	order
0.00000000000	4.24394E–04	0.00000E+00	0	0	0	0	0	0
0.07882283000	3.40520E+01	–1.09480E–08	1	–1	0	1	0	3
0.91272219540	9.30940E–01	–1.85070E–07	0	1	–1	–1	0	3
0.00402231670	9.11850E–01	–9.57650E–08	0	0	0	1	1	2
1.07036785680	9.31450E–01	–2.05600E–07	2	–1	–1	1	0	5
0.15362321080	3.22470E–01	–5.86630E–08	2	–2	0	1	–1	6
0.93766936950	1.93940E–01	–2.44860E–07	0	1	1	1	0	3
0.78002371740	1.95730E–01	–2.15110E–07	–2	3	1	–1	0	7
1.92921439960	7.07670E–02	–4.36290E–07	1	1	0	1	0	3
1.77156873300	6.44060E–02	–4.20950E–07	–1	3	0	–1	0	5

Table 5.21: Same as table 5.6 but for the z_S function.

The remaining ν_i have been taken in order to make the sequence of models SSSM_3, SSSM_4, SSSM_5 decreasing in error in the residual accelerations test that will be discussed bellow. After some trials, we have set

- $\nu_3 = \omega_1 - \omega_2 + \omega_4$, which is the main frequency of c_3,
- $\nu_4 = \omega_1 - \omega_5$, which is the first frequency of x_S which cannot be expressed in terms of ν_1, ν_2, and
- $\nu_5 = \omega_5 - \omega_2$, which is the first frequency of c_3 that cannot be expressed in terms of $\nu_1, \nu_2, \nu_3, \nu_4$.

In this way, we have

$$\begin{pmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \\ \nu_4 \\ \nu_5 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 \\ 1 & -1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & -1 \\ 0 & -1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \\ \omega_4 \\ \omega_5 \end{pmatrix}.$$

Since the above matrix is unimodular, $\{\nu_i\}_{i=1\div 5}$ is a valid basic set of frequencies.

Using residual accelerations, the SSSM_i models, as well as the RTBP, the Bicircular Problem (BCP) and the Quasi–Bicircular problem (QBCP, see Appendix A) have been compared with the real Solar System, as given by (5.3) and (5.4) with the c_i and x_i, y_i, z_i functions evaluated from the JPL DE406 ephemeris files. We have proceed as follows. Given two models to be compared, with differential equations $\ddot{\mathbf{r}} = f(\mathbf{r}, t)$ and $\ddot{\mathbf{r}} = g(r, t)$, respectively, and given a trajectory (positions and velocities) $\gamma : \mathbb{R} \to \mathbb{R}^6$, which does not need to be a trajectory of any of the models, we compute the "mean relative residual acceleration over γ" as

$$\frac{1}{L}\int_0^T \frac{\|f(\gamma(s),t) - g(\gamma(s),t)\|}{\|g(\gamma(s),t)\|}\|\gamma'(s)\|ds, \tag{5.6}$$

where t is a fixed epoch (in adimensional units) and

$$L = \int_0^T \|\gamma'(s)\| ds$$

is the length of the trajectory.

It must be noted that, the BCP and the QBCP as stated in appendix A assume that, for $t = 0$, the vector from the Earth to the Moon and the one from the Earth–Moon barycenter to the Sun form an angle of 180 degrees. Therefore, we must set the origin of adimensional time, both in the SSSM$_i$ models and the real Solar System, such that Earth, Moon and Sun are in a configuration close to the one of the BCP and the QBCP for $t = 0$. For the test of Table 5.22, we have chosen as $t = 0$ the first epoch after Jan 1st, 2001 in which the projection of the vector from the Earth–Moon barycenter to the Sun over the Earth–Moon instantaneous plane of motion forms an angle of 180 degrees with the vector from the Earth to the Moon. This is the Julian day 2451919.3489 (Jan 9th, 2001).

The results of the residual accelerations test are given in Table 5.22. From this table, it becomes clear that the best one–frequency models that we can use, using the residual acceleration criterium, are the BCP and the QBCP. But, when we allow two or more frequencies, the models we get fit the JPL one much better. As it has been said, only the Sun has been taken into account in all the intermediate models. By adding additional Solar System bodies, the residual accelerations are of the same order of magnitude than the ones obtained just using the Sun.

5.3.3 Simplified models for the Sun–Earth+Moon case

In this case, we will extract the basic frequencies from the Fourier analysis of Section 5.2 using Algorithm 5.3.1 for its determination.

From the numerical data obtained (see Appendix C), we first observe that the maximum modulus of the highest Fourier coefficient of c_1, c_2, c_3, c_6, c_8, c_9, c_{11} is 3.521E–05, whereas the minimum modulus of the highest Fourier coefficient of the remaining c_i is 1.669E–02. Therefore, in order to detect basic frequencies, we will only take into consideration the c_4, c_5, c_7, c_{10}, c_{12} and c_{13} functions. In addition to this simplification, we will not consider any Solar System body in (5.4), since, just using the c_i, we are already taking the Sun into account.

Applying Algorithm 5.3.1 to the c_{13} function, setting tol = 1E–5, $maxor = 20$, we get the following 4 basic frequencies:

$$\nu_1 = 0.9999926164,\ \nu_2 = 0.6255242728,\ \nu_3 = 0.9147445983,\ \nu_4 = 1.8313395538.$$

These 4 basic frequencies allow to adjust the frequencies of the best analysis of the c_4, c_5, c_7, c_{10} and c_{12} functions. For that, we have applied the second algorithm of section 5.3.1 with tol = 1E–5 amd $maxor = 20$. The results are given in tables 5.23 to 5.27. With these frequencies, we construct the SSSM$_1$, ..., SSSM$_4$ as we did in the Earth–Moon case.

In Table 5.29, we compare the models RTBP, SSSM$_1$ and SSSM$_4$ with the real Solar System using the same residual acceleration test that we used in the Earth–Moon case. We

z-a.	RTBP	BCP	QBCP	$SSSM_1$	$SSSM_2$	$SSSM_3$	$SSSM_4$	$SSSM_5$
0.020	0.140126	0.146459	0.138580	0.365299	0.095769	0.010674	0.001374	0.000727
0.022	0.138397	0.144693	0.136908	0.359442	0.094562	0.010534	0.001360	0.000724
0.025	0.136603	0.142856	0.135174	0.353302	0.093293	0.010388	0.001346	0.000720
0.028	0.134760	0.140962	0.133392	0.346913	0.091967	0.010235	0.001331	0.000716
0.031	0.132882	0.139025	0.131578	0.340305	0.090590	0.010076	0.001315	0.000711
0.034	0.130985	0.137059	0.129747	0.333509	0.089166	0.009913	0.001299	0.000707
0.038	0.129087	0.135080	0.127914	0.326550	0.087699	0.009744	0.001282	0.000702
0.043	0.127204	0.133103	0.126097	0.319452	0.086191	0.009570	0.001265	0.000696
0.048	0.125352	0.131141	0.124312	0.312235	0.084643	0.009393	0.001247	0.000691
0.053	0.123549	0.129209	0.122576	0.304915	0.083056	0.009211	0.001229	0.000685
0.059	0.121813	0.127324	0.120905	0.297505	0.081429	0.009024	0.001210	0.000678
0.066	0.120162	0.125502	0.119319	0.290018	0.079760	0.008833	0.001191	0.000671
0.073	0.118614	0.123757	0.117835	0.282462	0.078045	0.008637	0.001171	0.000664
0.082	0.117189	0.122108	0.116473	0.274845	0.076280	0.008436	0.001150	0.000655
0.091	0.115905	0.120571	0.115249	0.267173	0.074461	0.008229	0.001128	0.000646
0.102	0.114778	0.119161	0.114181	0.259453	0.072581	0.008016	0.001105	0.000636
0.113	0.113823	0.117895	0.113283	0.251690	0.070634	0.007796	0.001081	0.000625
0.126	0.113052	0.116784	0.112566	0.243889	0.068612	0.007568	0.001056	0.000612
0.141	0.112471	0.115836	0.112037	0.236056	0.066510	0.007331	0.001030	0.000598
0.157	0.112080	0.115057	0.111695	0.228199	0.064322	0.007085	0.001002	0.000583
0.175	0.111872	0.114443	0.111533	0.220325	0.062042	0.006831	0.000973	0.000566
0.195	0.111829	0.113984	0.111535	0.212440	0.059667	0.006566	0.000942	0.000547
0.217	0.111928	0.113663	0.111672	0.204551	0.057196	0.006292	0.000910	0.000526
0.242	0.112133	0.113450	0.111909	0.196665	0.054632	0.006008	0.000875	0.000504
0.269	0.112400	0.113311	0.112201	0.188782	0.051978	0.005716	0.000840	0.000481
0.300	0.112678	0.113200	0.112492	0.180899	0.049240	0.005417	0.000802	0.000456

Table 5.22: Mean residual accelerations between several models and the real Solar System over selected halo orbits of the RTBP around L_2 in the Earth–Moon case. The first column displays the z–amplitude of the halo orbit used as test orbit. The remaining columns show the mean residual acceleration between the corresponding model and the real Solar System over the test orbit.

note that the $SSSM_4$ model gives worse results than $SSSM_1$. This is not a contradiction. Examining table 5.23 to 5.27 we can see that the maximum amplitude of the frequencies of c_4, c_5, c_7, c_{10} and c_{12} that are not multiple of ν_1 is 6.695E–05. Because of that, adding frequencies does not improve significantly the approximation of the c_i functions, and in this way the structure of the equations 5.3 "takes over" the fact that the c_i terms of $SSSM_4$ are closer to the ones of the real Solar System than the corresponding terms of $SSSM_1$.

Therefore, for the Sun–Earth+Moon case, we will give $SSSM_1$ as simplified Solar System model. Note that this is a model with very few frequencies that significantly improves the RTBP.

freq.	ampl.	err.	k_1	k_2	k_3	k_4	order
0.00000000000	1.30000E–09	0.00000E+00	0	0	0	0	0
0.99999261980	3.33720E–02	3.38800E–09	1	0	0	0	1
1.99998564390	8.35280E–04	4.11070E–07	2	0	0	0	2
1.25103997640	3.93800E–05	–8.56920E–06	0	2	0	0	2
1.83134352170	3.40050E–05	3.96790E–06	0	0	0	1	1
0.91473091670	2.84920E–05	–1.36820E–05	0	0	1	0	1
2.99997409570	1.97160E–05	–3.75350E–06	3	0	0	0	3
1.87659754110	9.29780E–06	2.47230E–05	0	3	0	0	3

Table 5.23: Frequencies of the best analysis of c_4 adjusted as linear combinations of $\{\nu_i\}_{i=1\div4}$. From left to right the columns are: frequency, in cycles per lunar revolution, amplitude, error (freq. $- k_1\nu_1 - \ldots k_4\nu_4$), coefficients of the linear combination that approximates freq., and order of the linear combination ($|k_1| + \ldots + |k_4|$).

freq.	ampl.	err.	k_1	k_2	k_3	k_4	order
0.00000000000	2.00000E+00	0.00000E+00	0	0	0	0	0
0.99999261700	6.67490E–02	5.51530E–10	1	0	0	0	1
1.99998563790	1.39230E–03	4.05090E–07	2	0	0	0	2
1.25103998380	6.69550E–05	–8.56180E–06	0	2	0	0	2
0.91475203530	6.12480E–05	7.43700E–06	0	0	1	0	1
1.83134663800	4.85690E–05	7.08420E–06	0	0	0	1	1
2.99997541480	3.01690E–05	–2.43440E–06	3	0	0	0	3
0.62552353770	2.92970E–05	–7.35060E–07	0	1	0	0	1

Table 5.24: Same as Table 5.23 but for the c_5 function.

freq.	ampl.	err.	k_1	k_2	k_3	k_4	order
0.00000000000	1.00042E+00	0.00000E+00	0	0	0	0	0
0.99999261500	5.00800E–02	–1.41660E–09	1	0	0	0	1
1.99998562010	1.25350E–03	3.87270E–07	2	0	0	0	2
0.91475953220	4.82370E–05	1.49340E–05	0	0	1	0	1
1.25103999430	4.22440E–05	–8.55130E–06	0	2	0	0	2
2.99998010500	3.08040E–05	2.25580E–06	3	0	0	0	3
0.62552269280	2.71900E–05	–1.58000E–06	0	1	0	0	1
1.83133006690	1.76890E–05	–9.48690E–06	0	0	0	1	1

Table 5.25: Same as Table 5.23 but for the c_7 function.

freq.	ampl.	err.	k_1	k_2	k_3	k_4	order
0.00000000000	1.00042E+00	0.00000E+00	0	0	0	0	0
0.99999261500	5.00800E–02	–1.41650E–09	1	0	0	0	1
1.99998562010	1.25350E–03	3.87270E–07	2	0	0	0	2
0.91475953220	4.82370E–05	1.49340E–05	0	0	1	0	1
1.25103999430	4.22440E–05	–8.55130E–06	0	2	0	0	2
2.99998010500	3.08040E–05	2.25580E–06	3	0	0	0	3
0.62552269280	2.71900E–05	–1.58000E–06	0	1	0	0	1
1.83133006690	1.76890E–05	–9.48690E–06	0	0	0	1	1

Table 5.26: Same as Table 5.23 but for the c_{10} function.

freq.	ampl.	err.	k_1	k_2	k_3	k_4	order
0.00000000000	–1.39300E–04	0.00000E+00	0	0	0	0	0
0.99999262330	1.66930E–02	6.87550E–09	1	0	0	0	1
1.99998564990	6.96230E–04	4.17110E–07	2	0	0	0	2
1.83134558880	3.11050E–05	6.03500E–06	0	0	0	1	1
1.25103987210	2.46550E–05	–8.67350E–06	0	2	0	0	2
2.99997235010	2.26070E–05	–5.49910E–06	3	0	0	0	3
0.91470513360	1.30450E–05	–3.94650E–05	0	0	1	0	1
1.87659675410	8.75900E–06	2.39360E–05	0	3	0	0	3
2.50211836990	5.41250E–06	2.12790E–05	0	4	0	0	4

Table 5.27: Same as Table 5.23 but for the c_{12} function.

freq.	ampl.	err.	k_1	k_2	k_3	k_4	order
0.00000000000	1.00042E+00	0.00000E+00	0	0	0	0	0
0.99999261640	5.00800E–02	5.35290E–12	1	0	0	0	1
1.99998562580	1.25340E–03	3.93030E–07	2	0	0	0	2
1.25104010020	4.71180E–05	–8.44540E–06	0	2	0	0	2
0.91474459830	4.67440E–05	–4.82540E–11	0	0	1	0	1
2.99997729050	3.07760E–05	–5.58700E–07	3	0	0	0	3
1.83133955380	2.81230E–05	–9.85990E–12	0	0	0	1	1
0.62552427280	1.62760E–05	1.35640E–11	0	1	0	0	1

Table 5.28: Same as Table 5.23 but for the c_{13} function.

z-a.	RTBP	$SSSM_1$	$SSSM_4$
0.020000	3.446497E–02	9.901526E–05	8.905454E–04
0.022288	3.429997E–02	9.844882E–05	8.842048E–04
0.024838	3.411184E–02	9.779360E–05	8.768670E–04
0.027680	3.390024E–02	9.701858E–05	8.684772E–04
0.030846	3.366579E–02	9.616913E–05	8.589500E–04
0.034375	3.341007E–02	9.521763E–05	8.482675E–04
0.038308	3.313580E–02	9.416327E–05	8.364166E–04
0.042691	3.284681E–02	9.300703E–05	8.234040E–04
0.047575	3.254789E–02	9.175134E–05	8.092527E–04
0.053018	3.224472E–02	9.039967E–05	7.939978E–04
0.059084	3.194355E–02	8.895610E–05	7.776813E–04
0.065843	3.165101E–02	8.742482E–05	7.603471E–04
0.073376	3.137381E–02	8.582841E–05	7.420444E–04
0.081771	3.111844E–02	8.413352E–05	7.227963E–04
0.091126	3.089082E–02	8.236183E–05	7.026421E–04
0.101551	3.069597E–02	8.051628E–05	6.816096E–04
0.113169	3.053770E–02	7.859979E–05	6.597243E–04
0.126117	3.041819E–02	7.661569E–05	6.370130E–04
0.140545	3.033772E–02	7.450252E–05	6.135638E–04
0.156624	3.029470E–02	7.240496E–05	5.893022E–04
0.174543	3.028516E–02	7.020714E–05	5.643885E–04
0.194512	3.030323E–02	6.801648E–05	5.388121E–04
0.216766	3.034115E–02	6.579492E–05	5.127031E–04
0.241565	3.038961E–02	6.350846E–05	4.862056E–04
0.269202	3.043825E–02	6.123496E–05	4.593820E–04
0.300000	3.047577E–02	5.898080E–05	4.323859E–04

Table 5.29: Mean relative residual accelerations between several models and the real Solar System over selected halo orbits of the RTBP around L_2 in the Sun–Earth+Moon case.

Part II

The neighborhood of the collinear equilibrium points in the RTBP