Wavelet transform modulus : phase retrieval and scattering

Irène Waldspurger

Thèse de doctorat de l'École normale supérieure

École doctorale 386 : sciences mathématiques de Paris Centre

Discipline : mathématiques appliquées

Wavelet transform modulus: phase retrieval and scattering

Transformée en ondelettes : reconstruction de phase et scattering

---

par

Irène Waldspurger

présentée et soutenue le 10 novembre 2015 devant un jury composé de :

M. Andrés Almansa	Examinateur
M. Habib Ammari	Examinateur
M. Alexandre d'Aspremont	Examinateur
M. Jalal Fadili	Rapporteur
M. Philippe Jaming	Rapporteur
M. Stéphane Mallat	Directeur de thèse

Unité mixte de recherche 8548 : département d'informatique de l'École normale supérieure

Abstract

Automatically understanding the content of a natural signal, like a sound or an image, is in general a difficult task. In their naive representation, signals are indeed complicated objects, belonging to high-dimensional spaces. With a different representation, they can however be easier to interpret.

This thesis considers a representation commonly used in these cases, in particular for the analysis of audio signals: the modulus of the wavelet transform. To better understand the behaviour of this operator, we study, from a theoretical as well as algorithmic point of view, the corresponding inverse problem: the reconstruction of a signal from the modulus of its wavelet transform.

This problem belongs to a wider class of inverse problems: phase retrieval problems. In a first chapter, we describe a new algorithm, *PhaseCut*, which numerically solves a generic phase retrieval problem. Like the similar algorithm *PhaseLift*, *PhaseCut* relies on a convex relaxation of the phase retrieval problem, which happens to be of the same form as relaxations of the widely studied problem MaxCut. We compare the performances of *PhaseCut* and *PhaseLift*, in terms of precision and complexity.

In the next two chapters, we study the specific case of phase retrieval for the wavelet transform. We show that any function with no negative frequencies is uniquely determined (up to a global phase) by the modulus of its wavelet transform, but that the reconstruction from the modulus is not stable to noise, for a strong notion of stability. However, we prove a local stability property. We also present a new non-convex phase retrieval algorithm, which is specific to the case of the wavelet transform, and we numerically study its performances.

Finally, in the last two chapters, we study a more sophisticated representation, built from the modulus of the wavelet transform: the scattering transform. Our goal is to understand which properties of a signal are characterized by its scattering transform. We first prove that the energy of scattering coefficients of a signal, at a given order, is upper bounded by the energy of the signal itself, convolved with a high-pass filter that depends on the order. We then study a generalization of the scattering transform, for stationary processes. We show that, in finite dimension, this generalized transform preserves the norm. In dimension one, we also show that the generalized scattering coefficients of a process characterize the tail of its distribution.

Résumé

Les tâches qui consistent à comprendre automatiquement le contenu d'un signal naturel, comme une image ou un son, sont en général difficiles. En effet, dans leur représentation naïve, les signaux sont des objets compliqués, appartenant à des espaces de grande dimension. Représentés différemment, ils peuvent en revanche être plus faciles à interpréter.

Cette thèse s'intéresse à une représentation fréquemment utilisée dans ce genre de situations, notamment pour analyser des signaux audio : le module de la transformée en ondelettes. Pour mieux comprendre son comportement, nous considérons, d'un point de vue théorique et algorithmique, le problème inverse correspondant : la reconstruction d'un signal à partir du module de sa transformée en ondelettes.

Ce problème appartient à une classe plus générale de problèmes inverses : les problèmes de reconstruction de phase. Dans un premier chapitre, nous décrivons un nouvel algorithme, *PhaseCut*, qui résoud numériquement un problème de reconstruction de phase générique. Comme l'algorithme similaire *PhaseLift*, *PhaseCut* utilise une relaxation convexe, qui se trouve en l'occurence être de la même forme que les relaxations du problème abondamment étudié MaxCut. Nous comparons les performances de *PhaseCut* et *PhaseLift*, en termes de précision et de rapidité.

Dans les deux chapitres suivants, nous étudions le cas particulier de la reconstruction de phase pour la transformée en ondelettes. Nous montrons que toute fonction sans fréquence négative est uniquement déterminée (à une phase globale près) par le module de sa transformée en ondelettes, mais que la reconstruction à partir du module n'est pas stable au bruit, pour une définition forte de la stabilité. On démontre en revanche une propriété de stabilité locale. Nous présentons également un nouvel algorithme de reconstruction de phase, non-convexe, qui est spécifique à la transformée en ondelettes, et étudions numériquement ses performances.

Enfin, dans les deux derniers chapitres, nous étudions une représentation plus sophistiquée, construite à partir du module de transformée en ondelettes : la transformée de scattering. Notre but est de comprendre quelles propriétés d'un signal sont caractérisées par sa transformée de scattering. On commence par démontrer un théorème majorant l'énergie des coefficients de scattering d'un signal, à un ordre donné, en fonction de l'énergie du signal initial, convolé par un filtre passe-haut qui dépend de l'ordre. On étudie ensuite une généralisation de la transformée de scattering, qui s'applique à des processus stationnaires. On montre qu'en dimension finie, cette transformée généralisée préserve la norme. En dimension un, on montre également que les coefficients de scattering généralisés d'un processus caractérisent la queue de distribution du processus.

Remerciements

Pendant les quatre années qu'a duré ma thèse, Stéphane Mallat m'a fait découvrir avec passion un domaine mathématique dont, initialement, je soupçonnais à peine l'existence. J'ai pu bénéficier de sa vision claire et profonde de tous les sujets que nous avons abordés, de ses idées foisonnantes et de sa grande culture scientifique. Je lui suis aussi reconnaissante de sa patience et de sa disponibilité.

Ce sont Albert Cohen et Francis Bach qui, lorsque j'étais en master, m'ont incitée à le contacter ; merci à eux pour ce conseil fructueux.

J'ai beaucoup apprécié les collaborations aussi conviviales qu'instructives que nous avons menées avec d'autres chercheurs de l'ENS. Merci donc à Alexandre d'Aspremont et Fajwel Fogel, ainsi qu'à Habib Ammari et Han Wang. Merci également à Alexandre pour son aide précieuse dans ma recherche de post-doc.

En janvier 2013, j'ai eu la chance de passer un mois à l'université de Yale. Je suis très reconnaissante à Ronald Coifman de m'avoir invitée et de m'avoir consacré autant de temps. Merci également à Amit Singer et Afonso Bandeira pour le séjour à Princeton.

Ce manuscrit a été relu avec beaucoup de soin par Jalal Fadili et Philippe Jaming. Je les remercie sincèrement pour ce travail, ainsi que pour les discussions que nous avons eues, en Savoie et à Bordeaux, sur la reconstruction de phase. Merci aussi à Andrés Almansa, Habib Ammari et Alexandre d'Aspremont d'avoir accepté de participer au jury.

Pendant ma thèse, j'ai été simultanément hébergée par les départements de mathématiques et d'informatique de l'ENS. Tous deux offrent un environnement de travail d'une grande qualité et je suis très heureuse qu'ils m'aient accueillie. Dans les équipes administratives, merci beaucoup à Albane, Bénédicte, Isabelle, Joëlle, Lara, Laurence, Lise-Marie, Michelle, Valérie et Zaïna pour leur efficacité et leur gentillesse. Au service des prestations informatiques, merci à Claudie, Jacques et Ludovic de m'avoir si souvent aidée. Merci également à toutes les personnes de la bibliothèque.

Quant à mon monitorat, merci aux élèves de Paris 6 puis de l'ENS d'en avoir fait une expérience aussi agréable et intéressante.

Dans les deux départements, merci beaucoup à mes collègues pour nos déjeuners et nos discussions. Au DI, ces remerciements s'adressent en particulier aux « glorieux anciens », Joakim, Joan et Laurent, aux « jeunes », Édouard, Grégoire, Mathieu et Vincent, aux « seniors » Carmine, Gilles, Guy, Ivan, Matthew, Sira et Xiu-Yuan, ainsi qu'à tout-e-s les stagiaires ou invité-e-s que je suis heureuse d'avoir rencontré-e-s, Chris, Goël, Maxime, Mia, Michel, Naoufal, Paul, Tomás, Y-Lan et Zhuoran. Au DMA, merci à Cécile, Charles, Clémence, Jaime et Valentine pour l'ambiance chaleureuse qui règne dans le bureau C21.

Au cours de ma scolarité, plusieurs professeur-e-s ont grandement contribué à développer mon intérêt pour les mathématiques. Merci tout particulièrement à Serge Dupont et Yves Duval.

Dans un registre plus personnel, merci à tou-te-s les ami-e-s qui ont rendu ces dernières années si agréables : Athanaric pour la cueillette des mirabelles, Benoît, inoubliable binôme, Catherine, Max et Silvain, qui m'ont fait découvrir Brewberry, Charles-Antoine et Jordi pour nos voyages en Belgique, Esther et sa constante gentillesse, Jacques-Henri et Marie-Karelle pour un voyage au Touquet riche en émotions, Jeremy, compagnon de pizzeria, Jonathan et son chat, les organisateurs du séminaire de l'abbé Mole (Bastien, Julien, Silvain et Vincent), Stefania, reine des bases de données et des tartes vapeur, Thierry, spécialiste ès foie gras et marathons, et Xavier, randonneur véloce et cinéphile distingué.

Sans mes parents, je n'aurais probablement jamais entrepris d'études de mathématiques. Heureusement qu'ils sont là. Je sais aussi que je peux toujours compter sur Héloïse. Merci à elle, et à Jonas. Merci également à Anne, Denis et Noémie.

Merci enfin à Gabriel de préparer le café comme personne.

Chapter 1

Introduction

The goal of data analysis is to develop methods to automatically understand the content of natural signals. Examples of possible tasks are finding which objects are pictured in an image, or transcribing an audio recording of a human voice.

These problems are in general difficult, because, under their raw form, signals tend to be complicated objects, living in high-dimensional spaces. An audio signal of 2 seconds, sampled at 44.1 kHz, is for example an element of $\mathbb{R}^{88200}$. To overcome this difficulty, a common method is to find a better representation of signals, which makes them easier to interpret.

Such representations must satisfy two essential properties. They must be discriminative: signals representing different things must have different representations. On the other hand, signals representing the same thing (two audio recordings of the same voice pronouncing the same words, for example) must have identical representations, even if they are not equal.

This thesis studies the wavelet transform modulus, which is an example of such a representation, used in particular in audio processing.

To understand to what extent it satisfies the previous two properties, we consider the corresponding inverse problem: the reconstruction of a function from its wavelet transform modulus. It is an example of a well-known class of inverse problems: phase retrieval problems. Chapter 2 of this thesis describes an algorithm to solve generic phase retrieval problems. Chapters 3 and 4 are devoted to the specific case of the wavelet transform.

In Chapters 5 and 6, we consider more sophisticated representations that are built from the wavelet transform modulus: scattering transforms.

Section 1.1 of this introduction is about the class of phase retrieval problems. In Section 1.2, we consider the specific case of phase retrieval for the wavelet transform. Section 1.3 is devoted to the scattering transform. In each section, an introduction explains the main definitions and known results, while the latter parts present the contributions of this thesis.

1.1 Phase retrieval problems

Phase retrieval problems are inverse problems of the form:

$$\text{Reconstruct } x \text{ from } \{|L_i(x)|\}_{i\in I}$$

where: x is an unknown element of a complex vector space E
I is an arbitrary set of indexes
for any $i \in I$, $L_i : E \to \mathbb{C}$ is a known linear form
$|.|$ denotes the usual complex modulus

Multiplying x by a unitary complex number has no influence on $\{|L_i(x)|\}_{i\in I}$:

$$\text{if } |\lambda| = 1, \text{ then} \qquad \forall i \in I, \quad |L_i(x)| = |L_i(\lambda x)|$$

so we can not hope to reconstruct exactly x from the set of modulus $\{|L_i(x)|\}_{i\in I}$. We only try to recover x *up to a global phase*, that is up to multiplication by a unitary complex number.

Phase retrieval problems have been studied from the fifties because of their numerous physical applications, for example X-ray imaging [Miao et al., 2008], diffractive imaging [Bunk et al., 2007] or astronomy [Dainty and Fienup, 1987]. In this section, we describe the theoretical and algorithmic issues they raise. We then come to the first contribution of this thesis (chapter 2): a general algorithm to solve phase retrieval problems.

1.1.1 Theoretical issues

Given a specific phase retrieval problem, the main theoretical question that it raises is to know whether the problem is well-posed, in terms of uniqueness and stability: is the reconstruction unique, up to a global phase? If it is unique, is it stable under measurement noise?

We call *uniqueness* in a phase retrieval problem the fact that all vectors $x \in E$ are uniquely determined by $\{|L_i(x)|\}_{i\in I}$, up to a global phase:

$$\begin{array}{rc} \forall x, y \in E \text{ such that} & \forall i \in I, |L_i(x)| = |L_i(y)|, \\ \exists \lambda \in \mathbb{C}, |\lambda| = 1 \text{ such that} & x = \lambda y \end{array} \qquad \text{(Uniqueness)}$$

The definition of *stability* is less canonical. Broadly speaking, we wish that, when the modulus $|L_i(x)|$ are not exactly known, but only up to a small error, it is still possible to reconstruct an approximation of the unknown x:

$$\begin{array}{rc} \forall x, y \in E \text{ such that} & \forall i \in I, |L_i(x)| \approx |L_i(y)|, \\ \exists \lambda \in \mathbb{C}, |\lambda| = 1 \text{ such that} & x \approx \lambda y \end{array} \qquad \text{(Stability)}$$

Various meanings are possible for the symbol "$\approx$". When the spaces E and $(\mathbb{R}^+)^I$ are endowed with metrics, which we respectively denote by d_E and d_I, we can for example decide that the reconstruction is *stable* if there exists $C > 0$ such that, for all $\epsilon > 0$:

$$\begin{array}{rc} \forall x, y \in E \text{ such that} & d_I\left(\{|L_i(x)|\}_{i\in I}, \{|L_i(y)|\}_{i\in I}\right) < \epsilon, \\ \exists \lambda \in \mathbb{C}, |\lambda| = 1 \text{ such that} & d_E(x, \lambda y) < C\epsilon \end{array} \qquad \text{(Strong stability)}$$

This notion of stability can however be too strong. For the wavelet transform, we will have to introduce a weaker notion of *local stability*.

Uniqueness and stability results for phase retrieval problems can be grouped in two main families:

- when the measurements L_i are randomly chosen: at least for particular choices of probability laws, the phase retrieval problem is then well-posed with high probability; (Uniqueness) and (Strong stability) both hold. The drawback of this case is that, in many applications, the measurements are fixed: they can not be chosen at random.

- when the measurements L_i are imposed by concrete applications: although useful, this setting is in general more difficult to theoretically study. The main cases that can be handled are when the L_i have a particular form, which allows to use harmonic analysis properties. Even in these cases, the results are often negative: there is no uniqueness or no stability.

For random measurements, when the ambient space E has finite dimension, there is uniqueness with probability 1 if the probability measure is uniformly continuous with respect to Lebesgue measure [Balan et al., 2006; Conca et al., 2015], and the number of measurements is large enough:

$$\text{Card } I \geq 4 \dim E - 4$$

When $E = \mathbb{C}^n$ and the measurements are independently chosen according to a normal random law, there is uniqueness and (strong) stability with high probability [Candès et al., 2013; Candès and Li, 2014], provided that, for a known constant C:

$$\text{Card } I \geq C \dim E$$

Other probability laws can be considered, as in [Alexeev et al., 2013; Candès et al., 2015; Gross et al., 2015a].

For deterministic measurements, the most well-known case, and also the most important for applications, is the reconstruction of a compactly-supported function from the modulus of its Fourier transform:

$$\text{Reconstruct a compactly-supported } f \in L^2(\mathbb{R}) \text{ from } |\hat{f}|$$

Here, the recovery is only considered up to trivial ambiguities (multiplication by a global phase, translation and complex conjugation). Even up to these ambiguities, there is no uniqueness [Akutowicz, 1956; Walther, 1963]. For uniqueness to hold, it is necessary that the signals f verify additional hypotheses, such as sparsity [Ranieri et al., 2013].

In higher dimensions ($f \in L^2(\mathbb{R}^d)$ with $d \geq 2$), there is uniqueness for "generic" functions f [Barakat and Newsam, 1984] but no stability.

The study of the phase retrieval problem for the Fourier transform relies on tools from harmonic analysis. These tools can be adapted to handle other measurements than the Fourier transform, notably fractional Fourier transforms. In this latter case, uniqueness holds, under simple conditions on the parameters of the fractional Fourier transforms [Jaming, 2014]. However, there is no strong stability [Andreys and Jaming, 2015].

An intermediate way between random measurements and fixed measurements imposed by a concrete situation is to design deterministic measurements so that they satisfy uniqueness or stability properties, as in [Bodmann and Hammen, 2014]. This method is interesting from a theoretical point of view, but it suffers from the same drawback as random measurements: it has in general no physical application.

1.1.2 Algorithms

Even when it is well-posed, numerically solving a phase retrieval problem is also difficult. Two main families of algorithms have been designed for generic phase retrieval problems, each with its advantages and drawbacks:

- iterative methods, which are simple to implement and relatively fast, but tend to return erroneous reconstructions
- methods by convexification, which often achieve exact reconstruction, but have high complexity

Iterative methods are the older family. They consist in building a sequence of approximate solutions, refined step by step until it converges, and returning the limit of this sequence. They notably include the alternate projection algorithm, introduced by Gerchberg and Saxton [1972], and improved by Fienup [1982].

Unfortunately, there is in general no guarantee that the sequence of approximate solutions converges towards the true solution. Phase retrieval problems are indeed highly non-linear, and do not admit a convex formulation. The sequence of approximations can therefore converge only towards a local optimum.

Depending on the considered application, various heuristics have been developed to overcome convergence problems [Millane, 1990]. For particular choices of random measurements, iterative

methods have been proven to converge towards the true solutions when correctly initialized [Netrapalli et al., 2013; Candès et al., 2015]. However, in general, their convergence properties are not well understood. For phase retrieval problems which arise in audio processing (and to which we will come back), they are often disappointing: reconstructed signals tend to present audio artifacts [Sturmel and Daudet, 2011].

Methods by convexification have been introduced by Chai et al. [2011] and by Candès et al. [2013], whose algorithm *PhaseLift* we briefly describe. The principle is to reformulate the phase retrieval problem under a matricial form. This reformulation is still a non-convex problem, but has a convex approximation. At least for precise choices of random measurements, the solution to the approximated problem can be proven to be the same as the solution to the initial problem, with high probability.

To describe the reformulation of the problem into a matricial form, we assume that we are in a finite-dimensional setting: $E = \mathbb{C}^n$, for some $n \in \mathbb{N}^*$. The linear forms $L_i : E \to \mathbb{C}$ are then matrices with one line: $L_i \in \mathcal{M}_{1,n}(\mathbb{C})$. For any i, we set $b_i = |L_i(x)|$. The constraint $|L_i(x)| = b_i$ can be rewritten as:

$$|L_i(x)| = b_i \quad \Longleftrightarrow \quad |L_i(x)|^2 = b_i^2 \quad \Longleftrightarrow \quad x^* L_i^* L_i x = b_i^2 \quad \Longleftrightarrow \quad \mathrm{Tr}\left(L_i^* L_i x x^*\right) = b_i^2$$

The phase retrieval problem then becomes:

$$\text{Find } x \text{ s.t. } \forall i, |L_i(x)| = b_i \quad \Longleftrightarrow \quad \text{Find } x \text{ s.t. } \forall i, \mathrm{Tr}\left(L_i^* L_i x x^*\right) = b_i^2$$

With a change of variable $X = xx^*$, this is equivalent to:

$$\begin{array}{rl} \text{Find} & X \in \mathcal{M}_n(\mathbb{C}) \\ \text{s.t.} & \mathrm{rank}(X) = 1 \\ \text{and} & \forall i, \mathrm{Tr}\left(L_i^* L_i X\right) = b_i^2 \end{array}$$

This problem is not convex, because the set of matrices with rank 1 is not convex. However, it can be approximated by a convex problem, with the same method as used for matrix completion [Candès and Recht, 2009]. This convex problem can be solved in a time which is polynomial in the dimension n and the number of measurements.

Candès et al. [2013]; Candès and Li [2014] have shown that, when the linear forms L_i are chosen according to random normal laws, then the convex approximation and the initial problem have the same solution, provided that:

$$\text{Card } I \geq C \dim E$$

(where C is a known constant).

Even if these correctness guarantees are limited to specific choices of measurements, empirical evaluations show that *PhaseLift* yields exact reconstructions in many cases.

However, the dimension of the reformulated problem is much higher than the dimension of the original problem: X has n^2 entries, while x is an n-dimensional vector. The algorithm thus has a high (although polynomial) complexity, worse than iterative methods. It prevents it from being used for applications where signals are necessarily of high dimension, as in audio processing.

1.1.3 A general phase retrieval algorithm, *PhaseCut* (chapter 2)

Chapter 2 of this thesis describes a new phase retrieval algorithm, called *PhaseCut*, belonging to the class of methods by convexification. It is similar to *PhaseLift*, but uses a different matricial reformulation. It yields an algorithm with a different complexity, more efficient than *PhaseLift* for some instances of phase retrieval problems.

Instead of directly reconstructing $x \in \mathbb{R}^n$, as in *PhaseLift*, we focus on reconstructing $\{L_i(x)\}_{i\in I}$ (which, under mild assumptions, also allows to recover x). By assumption, the modulus $|L_i(x)|$ are known. So it suffices to reconstruct the phases, that is the complex numbers u_i such that:

$$L_i(x) = |L_i(x)|u_i \qquad \text{and} \qquad |u_i| = 1$$

In the same way as, in *PhaseLift*, the constraints $|L_i(x)| = b_i$ could be reformulated by introducing the matrix $X = xx^*$, here, the constraints $|u_i| = 1$ can be reformulated by introducing the matrix $U = uu^*$, where:

$$u = \begin{pmatrix} u_1 \\ \vdots \\ u_m \end{pmatrix}$$

So, as in *PhaseLift*, we obtain a reformulation of the phase retrieval problem in terms of matrices, and this reformulation also has a convex approximation, which we call *PhaseCut*.

Almost the same correctness guarantees hold for *PhaseLift* and *PhaseCut*. We can actually show that *PhaseLift* yields exact reconstruction when and only when a slightly modified version of *PhaseCut* yields exact reconstruction. The same argument allows one to compare the stability to noise of both algorithms: *PhaseCut* is at least as stable to noise as a slightly modified version of *PhaseLift*, but can be more stable, especially when the set of measurements $\{L_i(x)\}_{i\in I}$ is sparse.

The main advantage of *PhaseCut* over *PhaseLift* is its different complexity. At first sight, the complexity of *PhaseCut* seems worse, because the involved matrices are of higher dimension as in *PhaseLift*: they have m^2 entries, with m the number of measurements, while matrices in *PhaseLift* only have n^2 entries, with n the dimension of the unknown x (always smaller than

m). However, it happens that the convex problem obtained with *PhaseCut* has the same form as convex relaxations of *MaxCut*-type problems, which have been widely studied since their introduction in [Delorme and Poljak, 1993; Goemans and Williamson, 1995]. We can thus apply to *PhaseCut* optimization techniques developed for *MaxCut*, which do not apply to *PhaseLift*.

So the complexity of *PhaseCut* can be better or worse than the one of *PhaseLift*, depending on the optimization algorithm used to solve the convex problem. *PhaseLift* will then be more efficient for some problems, while *PhaseCut* will be better suited to others. Our numerical experiments indicate that *PhaseCut* is better for difficult phase retrieval problems, where the reconstruction is not very stable.

1.2 Phase retrieval for the wavelet transform

We now consider a specific phase retrieval problem, with important applications in audio processing: the case of the wavelet transform. We first define the wavelet transform, and give our motivations for studying the corresponding phase retrieval problem. Finally, we describe the theoretical and algorithmic contributions of this thesis.

1.2.1 Wavelet transform modulus

Wavelets have been introduced at the end of the eighties; introductions can be found in [Cohen, 1992; Daubechies, 1992; Mallat, 2009].

In this thesis, we call *wavelet* any function $\psi : \mathbb{R} \to \mathbb{C}$, in $L^1 \cap L^2(\mathbb{R})$ such that:

$$\int_{\mathbb{R}} \psi(t)dt = 0$$

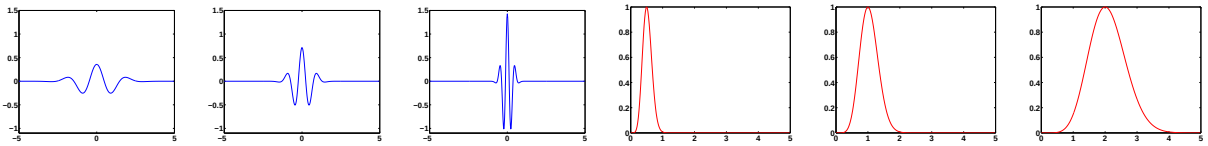
If such a function ψ is fixed, we can define by contraction or dilation of ψ a whole family of wavelets $(\psi_j)_{j\in\mathbb{Z}}$:

$$\begin{array}{rlll} & \forall j \in \mathbb{Z}, t \in \mathbb{R} & \psi_j(t) & = & 2^{-j}\psi(2^{-j}t) \\ \iff & \forall j \in \mathbb{Z}, \omega \in \mathbb{R} & \hat{\psi}_j(\omega) & = & \hat{\psi}(2^j\omega) \end{array}$$

The wavelet transform $W : L^2(\mathbb{R}) \to (L^2(\mathbb{R}))^{\mathbb{Z}}$ is then:

$$\begin{array}{rccc} W: & L^2(\mathbb{R}) & \to & (L^2(\mathbb{R}))^{\mathbb{Z}} \\ & f & \to & \{f \star \psi_j\}_{j\in\mathbb{Z}} \end{array}$$

In general, the mother wavelet ψ is chosen so that it is well-localized around 0, and its Fourier transform $\hat{\psi}$ is well-localized around 1. For any $j \in \mathbb{Z}$, ψ_j can then be interpreted as a band-pass



(a) (b) (c) (d) (e) (f)

Figure 1.1: Example of a wavelet family. (a) (b) (c) Functions $\psi_1, \psi_0, \psi_{-1}$; only the real part is displayed (d) (e) (f) Functions $\hat{\psi}_1, \hat{\psi}_0, \hat{\psi}_{-1}$

filter of characteristic frequency 2^{-j} and bandwidth proportional to 2^{-j}. Informally, the wavelet transform of a function f is thus a decomposition of f in (overlapping) frequency bands.

Figure 1.1 shows an example of a wavelet family, and Figure 1.2a an example of wavelet transform.

Wavelets are generally complex-valued, so too are thus the $f \star \psi_j$. We denote the wavelet transform modulus by:

$$\begin{array}{rccc} |W|: & L^2(\mathbb{R}) & \to & (L^2(\mathbb{R}))^{\mathbb{Z}} \\ & f & \to & \{|f \star \psi_j|\}_{j \in \mathbb{Z}} \end{array}$$

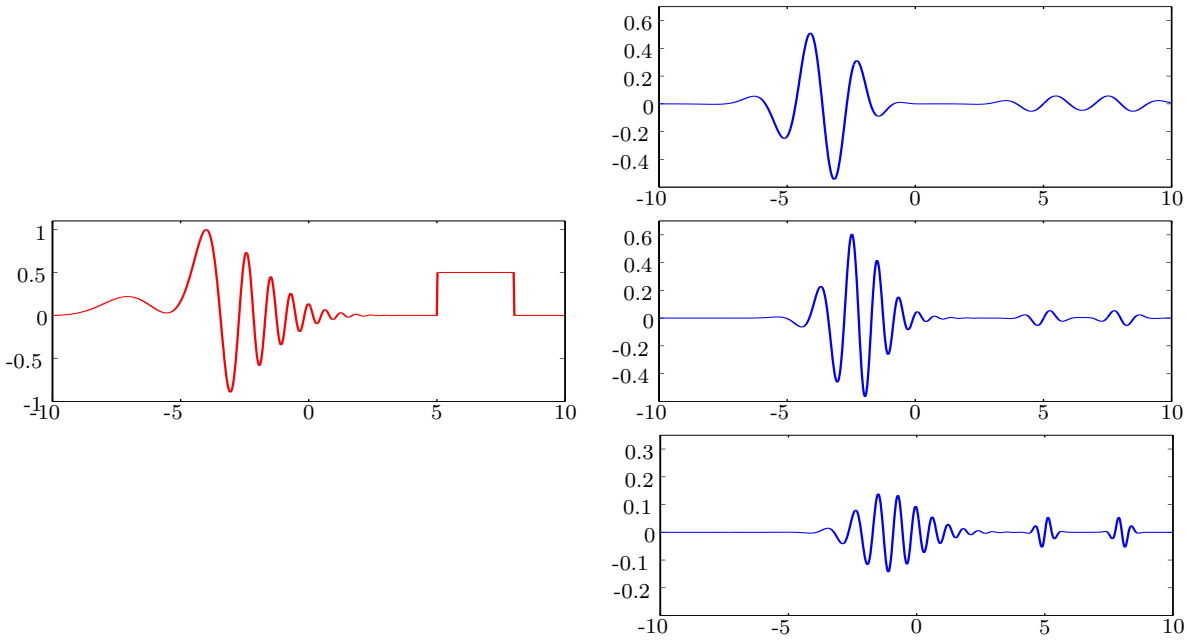
Informally, $|f \star \psi_j(t)|$ represents the energy of the signal f, around time t, in the frequency band centered at 2^{-j}. An example is on Figure 1.2b.

1.2.2 Interest of the phase retrieval problem

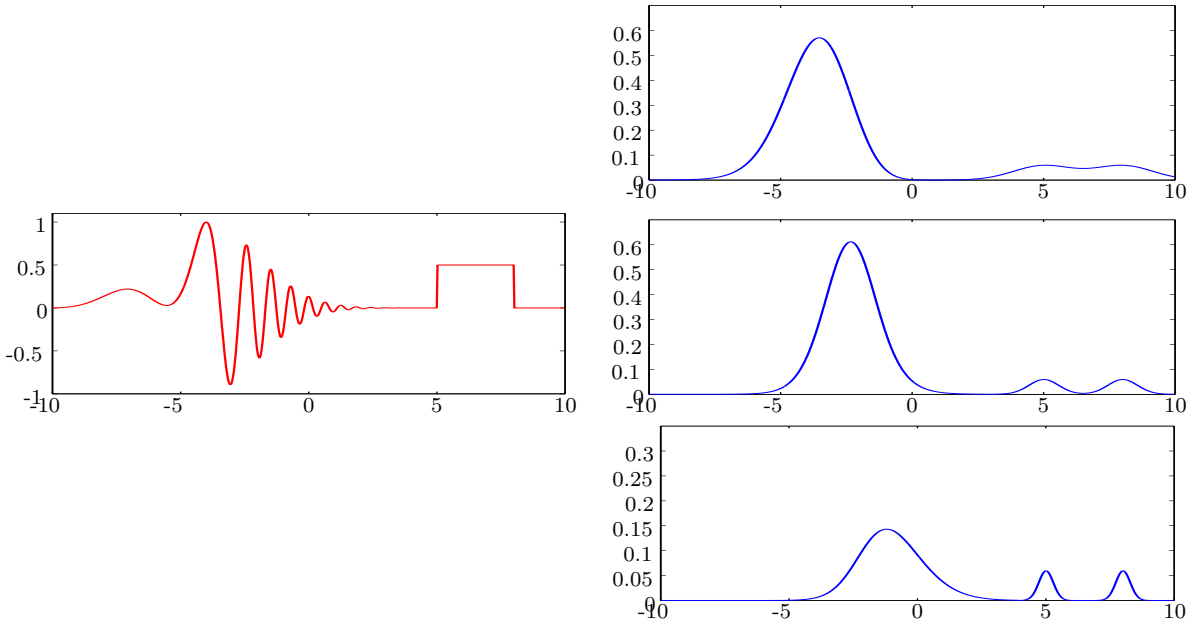
The modulus of the wavelet transform, sometimes called *scalogram*, is a common way to represent and analyze audio signals, relatively similar to the *spectrogram*. It has also been proposed as a model of the auditory cortex [Yang et al., 1992]. Indeed, it possesses the desirable properties mentioned at the beginning of this introduction [Balan et al., 2006; Risset and Wessel, 1999]:

- Audio signals which sound different to the human ear have different scalograms.
- Audio signals which are identical to the ear have almost the same scalograms, although they may not be equal.

To theoretically justify these properties, it is natural to consider the inverse phase retrieval problem: to what extent is it possible to recover an audio signal from the modulus of its wavelet transform? This problem has been investigated from the early eighties [Griffin and Lim, 1984; Nawab et al., 1983], but from an experimental, rather than theoretical, point of view.



(a) Example of wavelet transform. On the left: a function f. On the right, from top to bottom: $f \star \psi_1, f \star \psi_0, f \star \psi_{-1}$ (only real parts are displayed).



(b) Example of modulus of wavelet transform. On the left: the same function f as previously. On the right, from top to bottom: $|f \star \psi_1|, |f \star \psi_0|, |f \star \psi_{-1}|$.

Figure 1.2: Wavelet transform and modulus of the wavelet transform

Besides its theoretical motivations, the phase retrieval problem for the wavelet transform has concrete applications: in audio processing, some tasks require modifying the scalogram of a given signal, then reconstructing a new signal. It then amounts to solve a phase retrieval problem. Examples include blind source separation [Virtanen, 2007] and audio texture synthesis [Bruna and Mallat, 2013b].

For the study of phase retrieval problems, the case of the wavelet transform is also interesting, because it has atypical properties, compared to other known examples:

- It is a non-random case, relevant for applications, where harmonic analysis properties allow to prove a uniqueness result, for a specific choice of the wavelet family.
- We can show that the reconstruction is not stable to noise in a strong sense. However, for a specific choice of the wavelet family, it satisfies a local stability property, which had not been observed until now.

1.2.3 Phase retrieval for the Cauchy wavelet transform (chapter 3)

In chapter 3, we study the phase retrieval problem from a theoretical point of view, for a specific choice of wavelets.

We use Cauchy wavelets (which will be defined in chapter 3). These wavelets satisfy a very particular property: for any function $f \in L^2(\mathbb{R})$,

$$\forall j \in \mathbb{Z}, t \in \mathbb{R}, \qquad f \star \psi_j(t) = c_j F(t + i2^j)$$

where $(c_j)_{j\in\mathbb{Z}}$ is a sequence of explicit constants, and F is the holomorphic extension of (a modification of) f to the complex upper plane. The phase retrieval problem can then be rephrased in terms of holomorphic functions, and analyzed with techniques similar to Akutowicz [1956], by factorizing F into Blaschke products.

We have a uniqueness theorem:

Theorem (3.2)**.** *If $f, g \in L^2(\mathbb{R})$ are two functions such that $\hat{f}(\omega) = \hat{g}(\omega) = 0$ for any $\omega < 0$, and, for at least two distinct values of $j \in \mathbb{Z}$:*

$$|f \star \psi_j| = |g \star \psi_j|$$

then $f = e^{i\phi} g$ for some $\phi \in \mathbb{R}$.

This result is strong in the sense that it does not require the whole wavelet transforms of f and g to be equal in modulus: it suffices that only two components are equal.

However, we show that the reconstruction is not strongly stable, in the sense that, for any $\epsilon > 0$, there exist $f, g \in L^2(\mathbb{R})$ such that:

$$\big|\big| |W|f - |W|g \big|\big|_2 < \epsilon \qquad \text{but} \qquad \forall \phi \in \mathbb{R}, ||f - e^{i\phi} g||_2 \geq 1$$

where $|W|$ still denotes the modulus of the wavelet transform. This holds for all wavelets, not especially Cauchy ones.

We nevertheless have a form of local stability. If $|W|f \approx |W|g$, then our numerical experiments show that:

$$\forall j \in \mathbb{Z}, t \in \mathbb{R}, \qquad f \star \psi_j(t) \approx e^{i\phi_j(t)} g \star \psi_j(t) \tag{1.1}$$

where $(j, t) \to \phi_j(t)$ is a phase whose variation is slow in j and t, except maybe at the points where $f \star \psi_j(t)$ is close to zero. Informally, Equation (1.1) expresses the fact that the wavelet transforms of f and g are approximately equal, up to a global phase, in the neighborhood of each point (j, t) of the time-frequency plane. Hence the term *local stability*.

This numerical result can be formalized and proven for Cauchy wavelets (Theorems 3.17 and 3.18), although the statements are more technical than this simplified exposition.

1.2.4 Phase retrieval for wavelet transforms: a non-convex algorithm (chapter 4)

Results from Chapter 3 indicate that phase retrieval for the wavelet transform is a relatively well-posed problem, as reconstruction is unique and locally stable. However, generic phase retrieval algorithms tend to yield disappointing numerical results:

- Convexification methods like *PhaseLift* [Chai et al., 2011; Candès et al., 2011] and *PhaseCut* (Chapter 2) give very encouraging results for toy problems of small dimension [Sun and Smith [2012], Paragraph 2.4.3]. However, their complexity is too high for them to be used on real-size problems.

- Iterative methods for audio signals have been introduced by [Griffin and Lim, 1984] and subsequently refined [Achan et al., 2004; Bouvrie and Ezzat, 2006]. However, as the problem is non-convex, they tend to get stuck in local optima, and they generally do not yield reconstruction results of high quality.

In Chapter 4, we propose a new iterative method, achieving the same precision as convexification methods on small-size problems, with a complexity roughly linear in the size of the signal, allowing it to be used on real audio signals.

The main tool of this algorithm is a reformulation of the phase retrieval problem. This reformulation uses the notion of holomorphic extension, as in Chapter 3, but applies to any family of (exponentially-decaying wavelets), and not only to Cauchy ones.

The reformulation yields a multiscale algorithm, reconstructing the unknown signal f frequency band by frequency band, starting with the low frequencies and ending with the high ones. At each step, the already recovered phase information is exploited to reconstruct one more frequency band.

Numerical experiments confirm that the algorithm performs well: the reconstruction is of high quality, and degrades proportionally to measurement noise. It is worth noting that, because the phase retrieval problem is locally but not globally stable, the signals that are reconstructed from the wavelet transform modulus may not be very close to the original signals. However, they have almost exactly the same wavelet transform, in modulus. In the case of audio signals, original signals and their reconstructions, although not necessarily equal, are in general indistinguishable to the ear.

We use our algorithm to experimentally validate the theoretical conclusions of Chapter 3. We in particular highlight the fact that reconstruction is more difficult and less stable when the wavelet transform modulus has a lot of values close to zero.

1.3 Scattering transforms

In the last two chapters of the thesis, we consider the integration of the wavelet transform modulus in a more sophisticated representation: the scattering transform.

The scattering operator, introduced in [Mallat, 2012], is a deep representation, obtained by iterative application of the wavelet transform modulus. It has been defined so as to be invariant to translations of the input signal, and stable to small deformations. Since then, it has achieved important success in various data analysis tasks.

After a definition of the scattering (Paragraph 1.3.1), we discuss its main achievements and the theoretical questions it raises (Paragraph 1.3.2). In the frame of this thesis, the main question is to understand which properties of a signal are characterized by its scattering transform, especially which regularity properties. In Paragraphs 1.3.3 and 1.3.4, we describe the results of Chapters 5 and 6 related to this question.

1.3.1 Definition of scattering

The scattering transform is defined as a cascade of wavelet transform modulus, followed by averages.

Starting from a function $f \in L^2(\mathbb{R})$, we compute its local temporal mean by convolving it with a smooth window function ϕ_J, whose support has a characteristic size proportional to 2^J,

for some $J \in \mathbb{Z}$:

$$f \quad \rightarrow \quad f \star \phi_J$$

This temporal mean $f \star \phi_J$ (called *scattering coefficient of order* 0) has the good property of being almost invariant under translations of f:

$$f \star \phi_J \approx f(. - \tau) \star \phi_J \qquad \text{when } |\tau| \ll 2^J$$

However, $f \star \phi_J$ gives a very rough information on f: all high frequencies of f have been lost. Therefore, although it has the advantage of being invariant to translations, $f \star \phi_J$ is not discriminative enough to be an interesting representation of f.

We recover the high frequencies with the wavelet transform (whose very low frequencies can be discarded):

$$f \star \psi_J, \quad f \star \psi_{J-1}, \quad f \star \psi_{J-2}, ...$$

After an application of the complex modulus, we average these components with ϕ_J, to obtain *scattering coefficients of order* 1:

$$|f \star \psi_J| \star \phi_J, \quad |f \star \psi_{J-1}| \star \phi_J, \quad |f \star \psi_{J-2}| \star \phi_J, ...$$

Again, these averaged functions are almost invariant under translations of f. Together with $f \star \phi_J$, they bring more information about f than $f \star \phi_J$ alone, and thus form a more discriminative representation.

However, the high frequencies of the $|f \star \psi_j|$ have been lost during the averaging. We recover them with a new application of the wavelet transform modulus, and so on.

Scattering coefficients of order n are thus defined as the temporal average of the wavelet transform modulus, n times composed with itself. The whole process is schematized on Figure 1.3. Scattering coefficients make up a representation of f that is invariant to translations (when J goes to infinity) and stable to small deformations of f [Mallat, 2012].

This definition can be refined so as to be invariant or stable to other transformations than translations and deformations, for example rotations [Sifre and Mallat, 2013; Oyallon and Mallat, 2015].

1.3.2 Success and open question

Its invariance and stability properties allow the scattering transform to be successfully applied to various data analysis tasks, in particular classification tasks. Applications cover a high range of domains, such as images [Bruna and Mallat, 2013a], audio signals [Andén and Mallat, 2011] or even quantum chemistry [Hirn et al., 2015].

Compared to learned classifiers, the scattering offers competitive performances, with the advantage of being a deterministic transform, that does not need to be trained from data. This

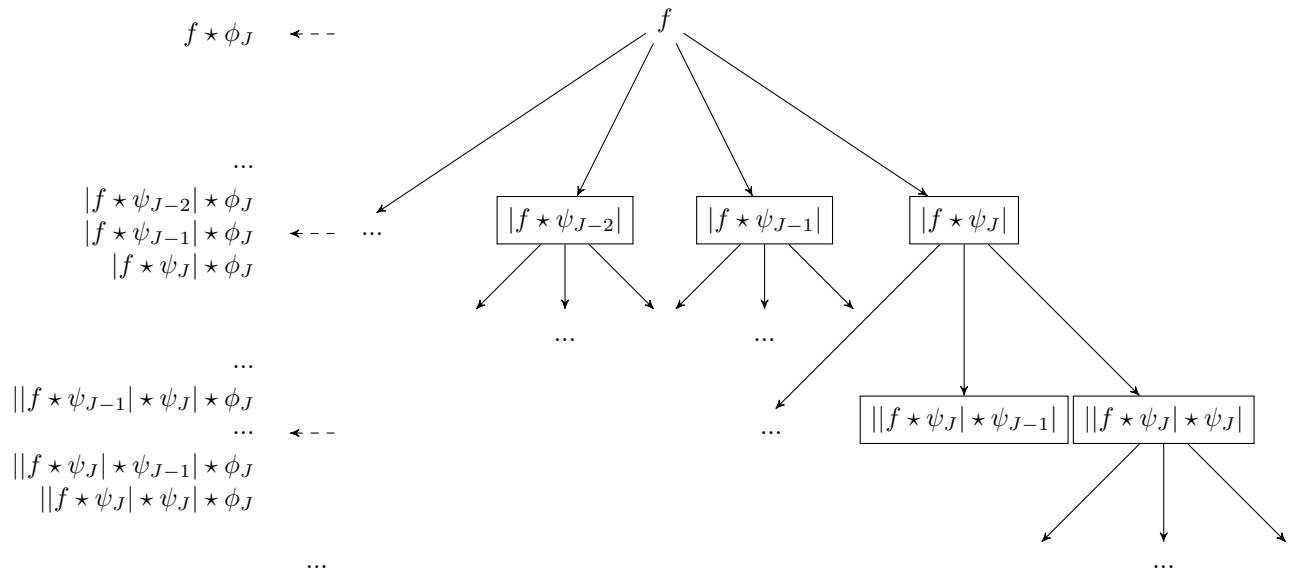


Figure 1.3: Schematic illustration of the scattering transform

gives a simple and generic framework, that can be efficiently implemented [Sifre et al., 2013], and analyzed from a theoretical point of view.

One of the main problems related to scattering is to determine which properties of a signal can be recovered from its scattering transform.

The scattering transform of a signal does not uniquely determine the signal. However, it characterizes some of its properties. For example, from the scattering transform of a sample of an auditive texture, it is not possible to exactly recover the sample. However, a perceptually plausible sample of the same texture can be reconstructed [Bruna and Mallat, 2013b].

In particular, we wonder to what extent the regularity of a signal can be characterized from its scattering transform. It is known that some regularity classes, like Besov spaces, can be characterized by the decay of wavelet coefficients [DeVore et al., 1992]. Are there similar results for scattering coefficients? An indication that this may be possible is that first and second-order scattering coefficients are already known to characterize some structural properties of stationary processes [Bruna et al., 2015].

1.3.3 Exponential decay of scattering coefficients (chapter 5)

In Chapter 5, we prove a result in this optic. We consider the decay of scattering coefficients as a function of their order; we show that it is controlled by the decay of the signal's Fourier transform.

[Mallat, 2012] proves that the scattering transform preserves the L^2-norm, provided that the wavelets satisfy some hypotheses. For any real-valued function $f \in L^2(\mathbb{R})$:

$$\sum_{n \geq 0} ||S_n[f]||_2^2 = ||f||_2^2$$

where $S_n[f]$ is the set of scattering coefficients of f with order n.

However, this result gives no information on how $||S_n(f)||_2$ decays with n. Theorem 5.2 partially solves this problem with an upper bound on $||S_n(f)||_2$. It indeed states that, under relatively general hypotheses on the wavelets, there exist constants $r > 0, a > 1$ such that, for any $n \geq 2$ and any real-valued function $f \in L^2(\mathbb{R})$:

$$\sum_{m \geq n} ||S_m[f]||_2^2 \leq \int_{\mathbb{R}} |\hat{f}(\omega)|^2 \left(1 - e^{-\left(\frac{\omega}{ra^n}\right)^2}\right) d\omega \tag{1.2}$$

The function $\omega \to 1 - e^{-\left(\frac{\omega}{ra^n}\right)^2}$ is a high-pass filter, with a bandwidth proportional to ra^n.

Equation (1.2) shows the existence, in the scattering transform, of a phenomenon of energy propagation towards the low frequencies. The energy of f carried by the frequency band $[-ra^n; ra^n]$ is output in the scattering coefficients with order smaller that n; the energy of scattering coefficients with order at least n comes almost only from higher frequency ranges.

1.3.4 Generalized scattering (chapter 6)

In Chapter 6, we introduce a generalization of the (finite-dimensional) scattering transform for stationary processes (as defined in [Mallat, 2012]). We show that these generalized scattering coefficients characterize some simple properties of the underlying stationary processes.

The generalized scattering starts with a random process X, taking its values in $\mathbb{R}^m$. We iteratively define:

$$\begin{aligned} X_0 &= X \\ \forall n \in \mathbb{N}, \qquad X_{n+1} &= |W_n X_n - \mathbb{E}(W_n X_n)| \end{aligned}$$

where the W_n are linear unitary operators and $|.|$ denotes the coordinate-wise modulus.

The *scattering coefficients* are $\{\mathbb{E}(W_n X_n)\}_{n\in\mathbb{N}}$.

This definition generalizes the scattering, in the sense that the linear operators W_n can represent transformations other than the wavelet transform; the expectation replaces the convolution with the low-pass filter ϕ_J.

Its interest is to offer a mathematical framework for the study of deep learned representations, in particular the ones learned in an unsupervised way. Within this framework, we can indeed consider learning the linear operators W_n, to adapt them to the law of X.

With no assumption on the ambient dimension, we simply show an energy preservation result (Theorem 6.2):

$$\text{if } \mathbb{E}(X^2) < +\infty \qquad\qquad \mathbb{E}(X^2) = \sum_{n=0}^{+\infty} ||\mathbb{E}(W_n X_n)||^2$$

In dimension one (that is, when the processes X_n take their values in $\mathbb{R}$ instead of $\mathbb{R}^m$ with general m), we can analyze the process with more precision. We show that scattering coefficients characterize the distribution tail of X (Theorem 6.4). More precisely, when X is not almost surely bounded and takes only positive values, then:

$$\mathbb{E}(X_n) \sim 2f(S_n) \qquad \text{when } n \to +\infty$$

where: - $f(t) = \mathbb{E}((X-t)1_{X\geq t})$ characterizes the law of X
- $S_n = \sum_{k<n} \mathbb{E}(X_k)$ can be computed from the scattering coefficients

The sequence (S_n) goes to infinity with n but $S_{n+1} - S_n \to 0$, so $(\mathbb{E}(X_n))_{n\in\mathbb{N}}$ precisely describes the asymptotic behavior of f, that is, the distribution tail of X.

Chapter 2

A general phase retrieval algorithm, *PhaseCut*

This chapter has be written in collaboration with Alexandre d'Aspremont and Stéphane Mallat, and published in [Waldspurger et al., 2015].

We consider a generic phase retrieval problem, in finite dimension. So we aim at reconstructing an unknown vector x, from linear complex-valued measurements Ax. The phase of these measurements has been lost; we have only access to their modulus $b = |Ax|$. Such a problem can be ill-posed; the modulus $|Ax|$ may for example not uniquely determine x. However, we are not concerned here with such issues: we assume the phase retrieval problem to be well-posed and focus on the design of a numerical algorithm.

Phase retrieval problems have traditionally been solved with iterative algorithms [Gerchberg and Saxton, 1972; Fienup, 1982]. Despite their success in some cases [Dainty and Fienup, 1987], these algorithms often stall in local minima, whose existence comes from the non-convexity of the problem. To overcome this difficulty, Chai et al. [2011] and Candès et al. [2011] have proposed a convex relaxation for the phase retrieval problem, called *PhaseLift*. It has been proven in [Candès et al., 2013; Candès and Li, 2014] that, with high probability, when the measurement matrix A has independent Gaussian entries, the relaxation is *tight*, meaning that it has the same solution as the original problem. So with high probability, the convex relaxation exactly reconstructs the unknown x, but, as it is convex, it can be solved with polynomial-time algorithms. It is also stable to noise.

The contribution of this chapter is to propose a different convex relaxation, called *PhaseCut*. Compared to *PhaseLift*, which directly reconstructs the unknown vector x, *PhaseCut* focuses on the reconstruction of the phase of Ax, which is an element of the complex unit torus. The

convex relaxation to which we then arrive has the same form as semidefinite programs for solving *MaxCut*-type problems [Delorme and Poljak, 1993; Goemans and Williamson, 1995]. Links had already been established between *MaxCut* and angular synchronization [Singer, 2011] and maximum-likelihood channel detection [Luo et al., 2003; Kisialiou and Luo, 2010; So, 2010], but it is the first time that a connection is drawn with phase retrieval.

We obtain correctness results for *PhaseCut* by showing (with the help of [Voroninski, 2012]) an equivalence with *PhaseLift*: in the noiseless case, the relaxation *PhaseLift* is tight if and only if a slightly modified version of *PhaseCut* is tight (the same equivalence exists between the unmodified *PhaseCut* and a slight modification of *PhaseLift* enjoying the same properties [Demanet and Hand, 2012; Candès and Li, 2014]). We also compare the stability to noise of *PhaseLift* and *PhaseCut*: *PhaseCut* is at least as stable as a variant of *PhaseLift*, but empirically appears to be more stable in some cases, in particular when $b = |Ax|$ is sparse.

Although the convex relaxation *PhaseCut* typically has a larger dimension than*PhaseLift*, it has a simpler structure. We can then apply efficient algorithms designed for *MaxCut*, in particular a provably convergent block coordinate descent algorithm whose complexity per iteration is the same as in the iterative algorithm [Gerchberg and Saxton, 1972]. Depending on the used algorithm, the complexity of *PhaseCut* is thus different from the one of *PhaseLift*, and significantly better in some cases. Depending on the exact considered phase retrieval problem, one or the other algorithm suits better.

From the point of view of combinatorial optimization, showing an equivalence between phase recovery and *MaxCut* allows us to expose a new class of nontrivial problem instances where the semidefinite relaxation for a *MaxCut*-like problem is tight, together with explicit conditions for tightness directly imported from the matrix completion formulation of these problems (these conditions are also hard to check, but hold with high probability for some classes of random experiments).

The chapter is organized as follows. In Section 2.1, we explain how to factorize away the magnitude information in the phase retrieval problem, so as to reformulate it as a non-convex quadratic optimization problem on the phase variables. The convex relaxation is derived in Paragraph 2.1.4. In Section 2.2, we detail several algorithms for solving this problem. In Section 2.3, we prove the equivalence, in terms of tightness, between a variant of *PhaseCut* and *PhaseLift*. We compare the stability to noise and the complexity of both algorithms. In Section 2.4, we perform numerical experiments comparing *PhaseLift*, *PhaseCut* and an iterative baseline, for three different phase retrieval problems. Section 2.5 contains technical lemmas.

Notations

We write $\mathbf{S}_p$ (resp. $\mathbf{H}_p$) the cone of symmetric (resp. Hermitian) matrices of dimension p; $\mathbf{S}_p^+$ (resp. $\mathbf{H}_p^+$) denotes the set of positive symmetric (resp. Hermitian) matrices.

We write $\|\cdot\|_p$ the Schatten p-norm of a matrix, that is the p-norm of the vector of its eigenvalues (in particular, $\|\cdot\|_\infty$ is the spectral norm). The notation $\|A\|_{\ell_1}$ refers to the sum of the modulus of the coefficients of A.

We write $A^\dagger$ the (Moore-Penrose) pseudo-inverse of a matrix A. For $x \in \mathbb{R}^p$, we write $\operatorname{diag}(x)$ the matrix with diagonal x. When $X \in \mathbf{H}_p$ however, $\operatorname{diag}(X)$ is the vector containing the diagonal elements of X. For $X \in \mathbf{H}_p$, X^* is the Hermitian transpose of X, with $X^* = (\bar{X})^T$. Finally, we write b^2 the vector with components b_i^2, $i = 1, \ldots, n$.

2.1 Phase recovery

The phase recovery problem seeks to retrieve a signal $x \in \mathbb{C}^p$ from the amplitude $b = |Ax|$ of n linear measurements, solving

$$\begin{array}{ll} \text{find} & x \\ \text{such that} & |Ax| = b, \end{array} \tag{2.1}$$

in the variable $x \in \mathbb{C}^p$, where $A \in \mathbb{C}^{n\times p}$ and $b \in \mathbb{R}^n$.

2.1.1 Greedy optimization in the signal

Approximate solutions x of the recovery problem in (2.1) are usually computed from $b = |Ax|$ using algorithms inspired from the alternating projection method in [Gerchberg and Saxton, 1972]. These algorithms compute iterates y^k in the set $\mathbf{F}$ of vectors $y \in \mathbb{C}^n$ such that $|y| = b = |Ax|$, which are getting progressively closer to the image of A.

The Gerchberg-Saxton algorithm projects the current iterate y^k on the image of A using the orthogonal projector $AA^\dagger$, then it adjusts to b_i the amplitude of each coordinate, so as to obtain a new element of $\mathbf{F}$. We describe this method explicitly below.

Algorithm 1 Gerchberg-Saxton.

Input: An initial $y^1 \in \mathbf{F}$, i.e. such that $|y^1| = b$.

1: **for** $k = 1, \ldots, N-1$ **do**

2: Set

$$y_i^{k+1} = b_i \frac{(AA^\dagger y^k)_i}{|(AA^\dagger y^k)_i|}, \quad i = 1, \ldots, n. \tag{Gerchberg-Saxton}$$

3: **end for**

Output: $y^N \in \mathbf{F}$.

Because $\mathbf{F}$ is not convex however, this alternating projection method usually converges to a stationary point y^∞ which does not belong to the intersection of $\mathbf{F}$ with the image of A, and hence $|AA^\dagger y^\infty| \neq b$. Several modifications proposed in [Fienup, 1982] do not eliminate the existence of multiple stationary points. To guarantee convergence to a unique solution, which hopefully belongs to the intersection of $\mathbf{F}$ and the image of A, this non-convex optimization problem has recently been relaxed as a semidefinite program [Chai et al., 2011; Candès et al., 2011], where phase recovery is formulated as a matrix completion problem (described in Section 2.3). Although the computational complexity of this relaxation is much higher than that of the Gerchberg-Saxton algorithm, it is able to recover x from $|Ax|$ (up to a multiplicative constant) in a number of cases [Chai et al., 2011; Candès et al., 2011].

2.1.2 Splitting phase and amplitude variables

As opposed to these strategies, we solve the phase recovery problem by explicitly separating the amplitude and phase variables, and by only optimizing the values of the phase variables. In the noiseless case, we can write $Ax = \text{diag}\,(b)u$ where $u \in \mathbb{C}^n$ is a phase vector, satisfying $|u_i| = 1$ for $i = 1, \ldots, n$. Given $b = |Ax|$, the phase recovery problem can thus be written as

$$\min_{\substack{u\in\mathbb{C}^n,\,|u_i|=1,\\ x\in\mathbb{C}^p}} \|Ax - \text{diag}\,(b)u\|_2^2,$$

where we optimize over both variables $u \in \mathbb{C}^n$ and $x \in \mathbb{C}^p$. In this format, the inner minimization problem in x is a standard least squares and can be solved explicitly by setting

$$x = A^\dagger \text{diag}\,(b)u,$$

which means that problem (2.1) is equivalent to the reduced problem

$$\min_{\substack{|u_i|=1\\ u\in\mathbb{C}^n}} \|AA^\dagger \text{diag}\,(b)u - \text{diag}\,(b)u\|_2^2.$$

The objective of this last problem can be rewritten as follows

$$\begin{aligned} \|AA^\dagger \text{diag}\,(b)u - \text{diag}\,(b)u\|_2^2 &= \|(AA^\dagger - \mathbf{I})\text{diag}\,(b)u\|_2^2 \\ &= u^* \text{diag}\,(b^T)\tilde{M}\text{diag}\,(b)u. \end{aligned}$$

where $\tilde{M} = (AA^\dagger - \mathbf{I})^*(AA^\dagger - \mathbf{I}) = \mathbf{I} - AA^\dagger$. Finally, the phase recovery problem (2.1) becomes

$$\begin{array}{ll} \text{minimize} & u^* M u \\ \text{subject to} & |u_i| = 1, \quad i = 1, \ldots n, \end{array} \tag{2.2}$$

in the variable $u \in \mathbb{C}^n$, where the Hermitian matrix

$$M = \operatorname{diag}(b)(\mathbf{I} - AA^\dagger)\operatorname{diag}(b)$$

is positive semidefinite. The intuition behind this last formulation is that $(\mathbf{I} - AA^\dagger)$ is the orthogonal projector on the orthogonal complement of the image of A (the kernel of A^*). So this last problem simply minimizes in the phase vector u the norm of the component of $\operatorname{diag}(b)u$ which is not in the image of A.

2.1.3 Greedy optimization in phase

Having transformed the phase recovery problem (2.1) in the quadratic minimization problem (2.2), we will be able to describe the associated convex relaxation. However, the reformulation (2.2) also yields a natural greedy optimization method; we begin with the latter.

Suppose that we are given an initial vector $u \in \mathbb{C}^n$, and focus on optimizing over a single component u_i for $i = 1, \ldots, n$. The problem is equivalent to solving

$$\begin{array}{ll} \text{minimize} & \bar{u}_i M_{ii} u_i + 2\mathrm{Re}\left(\sum_{j\neq i} \bar{u}_j M_{ji} u_i\right) \\ \text{subject to} & |u_i| = 1, \quad i = 1, \ldots n, \end{array}$$

in the variable $u_i \in \mathbb{C}$ where all the other phase coefficients u_j remain constant. Because $|u_i| = 1$ this then amounts to solving

$$\min_{|u_i|=1} \ \mathrm{Re}\left(u_i \sum_{j\neq i} M_{ji}\bar{u}_j\right)$$

which means

$$u_i = \frac{-\sum_{j\neq i} \bar{M}_{ji} u_j}{\left|\sum_{j\neq i} \bar{M}_{ji} u_j\right|} \tag{2.3}$$

for each $i = 1, \ldots, n$, when u is the optimum solution to problem (2.2). We can use this fact to derive Algorithm 2, a greedy algorithm for optimizing the phase problem.

This greedy algorithm converges to a stationary point u^∞, but it is generally not a global solution of problem (2.2), and hence $|AA^\dagger \operatorname{diag}(u^\infty)b| \neq b$. It has often nearly the same stationary points as the Gerchberg-Saxton algorithm. One can indeed verify that if u^∞ is a stationary point then $y^\infty = \operatorname{diag}(u^\infty)b$ is a stationary point of the Gerchberg-Saxton algorithm. Conversely if b has no zero coordinate and y^∞ is a stable stationary point of the Gerchberg-Saxton algorithm then $u_i^\infty = y_i^\infty/|y_i^\infty|$ defines a stationary point of the greedy algorithm in phase.

If Ax can be computed with a fast algorithm using $O(n \log n)$ operations, which is the case for Fourier or wavelets transform operators for example, then each Gerchberg-Saxton iteration is

Algorithm 2 Greedy algorithm in phase.

Input: An initial $u \in \mathbb{C}^n$ such that $|u_i| = 1$, $i = 1, \ldots, n$. An integer $N > 1$.

1: **for** $k = 1, \ldots, N$ **do**
2: **for** $i = 1, \ldots n$ **do**
3: Set

$$u_i = \frac{-\sum_{j \neq i} \bar{M}_{ji} u_j}{\left|\sum_{j \neq i} \bar{M}_{ji} u_j\right|}$$

4: **end for**
5: **end for**

Output: $u \in \mathbb{C}^n$ such that $|u_i| = 1$, $i = 1, \ldots, n$.

computed with $O(n \log n)$ operations. The greedy phase algorithm above does not take advantage of this fast algorithm and requires $O(n^2)$ operations to update all coordinates u_i for each iteration k. However, we will see in Section 2.2.6 that a small modification of the algorithm allows for $O(n \log n)$ iteration complexity.

2.1.4 Complex *MaxCut*

In this paragraph, we describe the convex relaxation of (2.2). It is an optimization problem over the set of Hermitian matrices. In the second half of the paragraph, we explain how to reformulate it as a real semidefinite program.

Following the classical relaxation argument in [Shor, 1987; Lovász and Schrijver, 1991; Goemans and Williamson, 1995; Nesterov, 1998], we first write $U = uu^* \in \mathbf{H}_n$. Problem (2.2), written

$$QP(M) \stackrel{def}{=} \begin{array}{ll} \text{min.} & u^* M u \\ \text{subject to} & |u_i| = 1, \quad i = 1, \ldots n, \end{array}$$

in the variable $u \in \mathbb{C}^n$, is equivalent to

$$\begin{array}{ll} \text{min.} & \mathrm{Tr}\,(UM) \\ \text{subject to} & \mathrm{diag}\,(U) = 1 \\ & U \succeq 0, \, \mathbf{Rank}(U) = 1, \end{array}$$

in the variable $U \in \mathbf{H}_n$. After dropping the (non-convex) rank constraint, we obtain the following convex relaxation

$$SDP(M) \stackrel{def}{=} \begin{array}{ll} \text{min.} & \mathrm{Tr}\,(UM) \\ \text{subject to} & \mathrm{diag}\,(U) = 1, \, U \succeq 0, \end{array} \qquad (PhaseCut)$$

which is a semidefinite program (SDP) in the matrix $U \in \mathbf{H}_n$ and can be solved efficiently. When the solution of problem *PhaseCut* has rank one, the relaxation is tight and the vector u such that $U = uu^*$ is an optimal solution of the phase recovery problem (2.2). If the solution has rank larger than one, a normalized leading eigenvector v of U is used as an approximate solution, and $\operatorname{diag}(U - vv^T)$ gives a measure of the uncertainty around the coefficients of v.

In practice, semidefinite programming solvers are rarely designed to directly handle problems written over Hermitian matrices and start by reformulating complex programs in $\mathbf{H}_n$ as real semidefinite programs over $\mathbf{S}_{2n}$ based on the simple facts that follow. For $Z, Y \in \mathbf{H}_n$, we define $\mathcal{T}(Z) \in \mathbf{S}_{2n}$ as in [Goemans and Williamson, 2004]

$$\mathcal{T}(Z) = \begin{pmatrix} \operatorname{Re}(Z) & -\operatorname{Im}(Z) \\ \operatorname{Im}(Z) & \operatorname{Re}(Z) \end{pmatrix} \tag{2.4}$$

so that $\operatorname{Tr}(\mathcal{T}(Z)\mathcal{T}(Y)) = 2\operatorname{Tr}(ZY)$. By construction, $Z \in \mathbf{H}_n$ if and only if $\mathcal{T}(Z) \in \mathbf{S}_{2n}$. One can also check that $z = x + iy$ is an eigenvector of Z with eigenvalue λ if and only if

$$\begin{pmatrix} x \\ y \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} -y \\ x \end{pmatrix}$$

are eigenvectors of $\mathcal{T}(Z)$, both with eigenvalue λ (depending on the normalization of z, one corresponds to $(\operatorname{Re}(z), \operatorname{Im}(z))$, the other one to $(\operatorname{Re}(i\,z), \operatorname{Im}(i\,z))$. This means in particular that $Z \succeq 0$ if and only if $\mathcal{T}(Z) \succeq 0$.

We can use these facts to formulate an equivalent semidefinite program over real symmetric matrices, written

$$\begin{array}{ll} \text{minimize} & \operatorname{Tr}(\mathcal{T}(M)X) \\ \text{subject to} & X_{i,i} + X_{n+i,n+i} = 2 \\ & X_{i,j} = X_{n+i,n+j},\ X_{n+i,j} = -X_{i,n+j}, \quad i,j = 1,\ldots,n, \\ & X \succeq 0, \end{array}$$

in the variable X in $\mathbf{S}_{2n}$. This last problem is equivalent to *PhaseCut*. In fact, because of symmetries in $\mathcal{T}(M)$, the equality constraints enforcing symmetry can be dropped, and this problem is equivalent to a *MaxCut*-like problem in dimension $2n$, which reads

$$\begin{array}{ll} \text{minimize} & \operatorname{Tr}(\mathcal{T}(M)X) \\ \text{subject to} & \operatorname{diag}(X) = 1, X \succeq 0, \end{array} \tag{2.5}$$

in the variable X in $\mathbf{S}_{2n}$. As we will see below, formulating a relaxation to the phase recovery problem as a complex *MaxCut*-like semidefinite program has direct computational benefits.

2.2 Algorithms

In the previous section, we have approximated the phase recovery problem (2.2) by a convex relaxation, written

$$\begin{array}{ll} \text{minimize} & \mathrm{Tr}\,(UM) \\ \text{subject to} & \mathrm{diag}\,(U) = 1,\ U \succeq 0, \end{array}$$

which is a semidefinite program in the matrix $U \in \mathbf{H}_n$. The dual, written

$$\max_{w \in \mathbb{R}^n} \ n\lambda_{\min}(M + \mathrm{diag}\,(w)) - 1^T w, \tag{2.6}$$

is a minimum eigenvalue maximization problem in the variable $w \in \mathbb{R}^n$. Both primal and dual can be solved efficiently. When exact phase recovery is possible, the optimum value of the primal problem *PhaseCut* is zero and we must have $\lambda_{\min}(M) = 0$, which means that $w = 0$ is an optimal solution of the dual.

2.2.1 Interior point methods

For small scale problems, with $n \sim 10^2$, generic interior point solvers such as SDPT3 [Toh et al., 1999] solve problem (2.5) with a complexity typically growing as $O\left(n^{4.5}\log(1/\epsilon)\right)$ where $\epsilon > 0$ is the target precision [Ben-Tal and Nemirovski, 2001, §4.6.3]. Exploiting the fact that the $2n$ equality constraints on the diagonal in (2.5) are singletons, Helmberg et al. [1996] derive an interior point method for solving the *MaxCut* problem, with complexity growing as $O\left(n^{3.5}\log(1/\epsilon)\right)$ where the most expensive operation at each iteration is the inversion of a positive definite matrix, which costs $O(n^3)$ flops.

2.2.2 First-order methods

When n becomes large, the cost of running even one iteration of an interior point solver rapidly becomes prohibitive. However, we can exploit the fact that the dual of problem (2.5) can be written (after switching signs) as a maximum eigenvalue minimization problem. Smooth first-order minimization algorithms detailed in [Nesterov, 2007] then produce an ϵ-solution after

$$O\left(\frac{n^3\sqrt{\log n}}{\epsilon}\right)$$

floating point operations. Each iteration requires forming a matrix exponential, which costs $O(n^3)$ flops. This is not strictly smaller than the iteration complexity of specialized interior

point algorithms, but matrix structure often allows significant speedup in this step. Finally, the simplest subgradient methods produce an ϵ-solution in

$$O\left(\frac{n^2 \log n}{\epsilon^2}\right)$$

floating point operations. Each iteration requires computing a leading eigenvector which has complexity roughly $O(n^2 \log n)$.

2.2.3 Block coordinate descent

We can also solve the semidefinite program in *PhaseCut* using a block coordinate descent algorithm. While no explicit complexity bounds are available for this method in our case, the algorithm is particularly simple and has a very low cost per iteration (it only requires computing a matrix vector product). We write i^c the index set $\{1, \ldots, i-1, i+1, \ldots, n\}$ and describe the method as Algorithm 3.

Block coordinate descent is widely used to solve statistical problems where the objective is separable (LASSO is a typical example) and was shown to efficiently solve semidefinite programs arising in covariance estimation [d'Aspremont et al., 2008]. These results were extended by [Wen et al., 2012] to a broader class of semidefinite programs, including *MaxCut*. We briefly recall its simple construction below, applied to a barrier version of the *MaxCut* relaxation *PhaseCut*, written

$$\begin{array}{ll} \text{minimize} & \mathrm{Tr}\,(UM) - \mu \log\det(U) \\ \text{subject to} & \mathrm{diag}\,(U) = 1 \end{array} \tag{2.7}$$

which is a semidefinite program in the matrix $U \in \mathbf{H}_n$, where $\mu > 0$ is the barrier parameter. As in interior point algorithms, the barrier enforces positive semidefiniteness and the value of $\mu > 0$ precisely controls the distance between the optimal solution to (2.7) and the optimal set of *PhaseCut*. We refer the reader to [Boyd and Vandenberghe, 2004] for further details. The key to applying coordinate descent methods to problems penalized by the $\log\det(\cdot)$ barrier is the following block-determinant formula

$$\det(U) = \det(B)\,\det(y - x^T B^{-1} x), \quad \text{when} \quad U = \begin{pmatrix} B & x \\ x^T & y \end{pmatrix}, \quad U \succ 0. \tag{2.8}$$

This means that, all other parameters being fixed, minimizing the function $\det(X)$ in the row and column block of variables x, is equivalent to minimizing the quadratic form $y - x^T Z^{-1} x$, arguably a much simpler problem. Solving the semidefinite program (2.7) row/column by row/column thus amounts to solving the simple problem (2.9) described in the following lemma.

Lemma 2.1. *Suppose $\sigma > 0$, $c \in \mathbb{R}^{n-1}$, and $B \in \mathbf{S}_{n-1}$ are such that $b \neq 0$ and $B \succ 0$, then the optimal solution of the block problem*

$$\min_x \; c^T x - \sigma \log(1 - x^T B^{-1} x) \tag{2.9}$$

is given by

$$x = \frac{\sqrt{\sigma^2 + \gamma} - \sigma}{\gamma} Bc$$

where $\gamma = c^T Bc$.

Proof. As in [Wen et al., 2012], a direct consequence of the first order optimality conditions for (2.9). □

Here, we see problem (2.7) as an unconstrained minimization problem over the off-diagonal coefficients of U, and (2.8) shows that each block iteration amounts to solving a minimization subproblem of the form (2.9). Lemma 2.1 then shows that this is equivalent to computing a matrix vector product. Linear convergence of the algorithm is guaranteed by the result in [Boyd and Vandenberghe, 2004, §9.4.3] and the fact that the function $\log\det$ is strongly convex over compact subsets of the positive semidefinite cone. So the complexity of the method is bounded by $O\left(\log \frac{1}{\epsilon}\right)$ but the constant in this bound depends on n here, and the dependence cannot be quantified explicitly.

Algorithm 3 Block Coordinate Descent Algorithm for ***PhaseCut***.

Input: An initial $X^0 = \mathbf{I}_n$ and $\nu > 0$ (typically small). An integer $N > 1$.

1: **for** $k = 1, \ldots, N$ **do**

2: Pick $i \in [1, n]$.

3: Compute

$$x = X^k_{i^c,i^c} M_{i^c,i} \quad \text{and} \quad \gamma = x^* M_{i^c,i}$$

4: If $\gamma > 0$, set

$$X^{k+1}_{i^c,i} = X^{k+1*}_{i,i^c} = -\sqrt{\frac{1-\nu}{\gamma}}\, x$$

else

$$X^{k+1}_{i^c,i} = X^{k+1*}_{i,i^c} = 0.$$

5: **end for**

Output: A matrix $X \succeq 0$ with $\operatorname{diag}(X) = 1$.

2.2.4 Initialization & randomization

Suppose the Hermitian matrix U solves the semidefinite relaxation *PhaseCut*. As in [Goemans and Williamson, 2004; Ben-tal et al., 2003; Zhang and Huang, 2006; So et al., 2007], we generate complex Gaussian vectors $x \in \mathbb{C}^n$ with $x \sim \mathcal{N}_{\mathbb{C}}(0, U)$, and for each sample x, we form $z \in \mathbb{C}^n$ such that

$$z_i = \frac{x_i}{|x_i|}, \quad i = 1, \ldots, n.$$

All the sample points z generated using this procedure satisfy $|z_i| = 1$, hence are feasible points for problem (2.2). This means in particular that $QP(M) \leq \mathbb{E}[z^* M z]$. In fact, this expectation can be computed almost explicitly, using

$$\mathbb{E}[zz^*] = F(U), \quad \text{with} \quad F(w) = \frac{1}{2} e^{i \arg(w)} \int_0^\pi \cos(\theta) \arcsin(|w| \cos(\theta)) d\theta$$

where $F(U)$ is the matrix with coefficients $F(U_{ij})$, $i, j = 1, \ldots, n$. We then get

$$SDP(M) \leq QP(M) \leq \mathrm{Tr}\,(MF(U)) \tag{2.10}$$

In practice, to extract good candidate solutions from the solution U to the SDP relaxation in *PhaseCut*, we sample a few points from $\mathcal{N}_{\mathbb{C}}(0, U)$, normalize their coordinates and simply pick the point which minimizes $z^* M z$.

This sampling procedure also suggests a simple spectral technique for computing rough solutions to problem *PhaseCut*: compute an eigenvector of M corresponding to its lowest eigenvalue and simply normalize its coordinates (this corresponds to the simple bound on *MaxCut* by [Delorme and Poljak, 1993]). The information contained in U can also be used to solve a robust formulation [Ben-Tal et al., 2009] of problem (2.1) given a Gaussian model $u \sim \mathcal{N}_{\mathbb{C}}(0, U)$.

2.2.5 Approximation bounds

The semidefinite program in *PhaseCut* is a *MaxCut*-type graph partitioning relaxation whose performance has been studied extensively. Note however that most approximation results for *MaxCut* study *maximization* problems over positive semidefinite or nonnegative matrices, while we are *minimizing* in *PhaseCut* so, as pointed out in [Kisialiou and Luo, 2010; So and Ye, 2010] for example, we do not inherit the constant approximation ratios that hold in the classical *MaxCut* setting.

2.2.6 Exploiting structure

In some instances, we have additional structural information on the solution of problems (2.1) and (2.2), which usually reduces the complexity of approximating *PhaseCut* and improves the quality of the approximate solutions. We briefly highlight a few examples below.

Alignment

In other instances, we might have prior knowledge that the phases of certain samples are aligned, i.e. that there is an index set I such that $u_i = u_j, \quad$ for all $i, j \in I$, this reduces to the symmetric case discussed above when the phase is arbitrary. W.l.o.g., we can also fix the phase to be one, with $u_i = 1$ for $i \in I$, and solve a constrained version of the relaxation *PhaseCut*

$$\begin{array}{ll} \text{min.} & \operatorname{Tr}(UM) \\ \text{subject to} & U_{ij} = 1, \quad i, j \in I, \\ & \operatorname{diag}(U) = 1,\ U \succeq 0, \end{array}$$

which is a semidefinite program in $U \in \mathbf{H}_n$.

Fast Fourier transform

If the product Mx can be computed with a fast algorithm in $O(n \log n)$ operations, which is the case for Fourier or wavelet transform operators, we significantly speed up the iterations of Algorithm 3 to update all coefficients at once. Each iteration of the modified Algorithm 3 then has cost $O(n \log n)$ instead of $O(n^2)$.

Real valued signal

In some cases, we know that the solution vector x in (2.1) is real valued. Problem (2.1) can be reformulated to explicitly constrain the solution to be real, by writing it

$$\min_{\substack{u \in \mathbb{C}^n,\, |u_i| = 1, \\ x \in \mathbb{R}^p}} \|Ax - \operatorname{diag}(b)u\|_2^2$$

or again, using the operator $\mathcal{T}(\cdot)$ defined in (2.4)

$$\begin{array}{ll} \text{minimize} & \left\| \mathcal{T}(A) \begin{pmatrix} x \\ 0 \end{pmatrix} - \operatorname{diag} \begin{pmatrix} b \\ b \end{pmatrix} \begin{pmatrix} \operatorname{Re}(u) \\ \operatorname{Im}(u) \end{pmatrix} \right\|_2^2 \\ \text{subject to} & u \in \mathbb{C}^n,\ |u_i| = 1 \\ & x \in \mathbb{R}^p. \end{array}$$

The optimal solution of the inner minimization problem in x is given by $x = A_2^\dagger B_2 v$, where

$$A_2 = \begin{pmatrix} \operatorname{Re}(A) \\ \operatorname{Im}(A) \end{pmatrix}, \quad B_2 = \operatorname{diag} \begin{pmatrix} b \\ b \end{pmatrix}, \quad \text{and} \quad v = \begin{pmatrix} \operatorname{Re}(u) \\ \operatorname{Im}(u) \end{pmatrix}$$

hence the problem is finally rewritten

$$\begin{array}{ll} \text{minimize} & \|(A_2 A_2^\dagger B_2 - B_2)v\|_2^2 \\ \text{subject to} & v_i^2 + v_{n+i}^2 = 1, \quad i = 1, \ldots, n, \end{array}$$

in the variable $v \in \mathbb{R}^{2n}$. This can be relaxed as above by the following problem

$$\begin{array}{ll} \text{minimize} & \operatorname{Tr}(VM_2) \\ \text{subject to} & V_{ii} + V_{n+i,n+i} = 1, \quad i = 1, \ldots, n, \\ & V \succeq 0, \end{array}$$

which is a semidefinite program in the variable $V \in \mathbf{S}_{2n}$, where $M_2 = (A_2 A_2^\dagger B_2 - B_2)^T (A_2 A_2^\dagger B_2 - B_2) = B_2^T (\mathbf{I} - A_2 A_2^\dagger) B_2$.

2.3 Matrix completion & exact recovery conditions

In [Chai et al., 2011; Candès et al., 2011], phase recovery (2.1) is cast as a matrix completion problem. In this section, we briefly review this approach and compare it with the semidefinite program in *PhaseCut*.

Given a signal vector $b \in \mathbb{R}^n$ and a sampling matrix $A \in \mathbb{C}^{n \times p}$, we look for a vector $x \in \mathbb{C}^p$ satisfying

$$|a_i^* x| = b_i, \quad i = 1, \ldots, n,$$

where the vector a_i^* is the i^{th} row of A and $x \in \mathbb{C}^p$ is the signal we are trying to reconstruct. The phase recovery problem is then written as

$$\begin{array}{ll} \text{minimize} & \mathbf{Rank}(X) \\ \text{subject to} & \operatorname{Tr}(a_i a_i^* X) = b_i^2, \quad i = 1, \ldots, n \\ & X \succeq 0 \end{array}$$

in the variable $X \in \mathbf{H}_p$, where $X = xx^*$ when exact recovery occurs. This last problem can be relaxed as

$$\begin{array}{ll} \text{minimize} & \operatorname{Tr}(X) \\ \text{subject to} & \operatorname{Tr}(a_i a_i^* X) = b_i^2, \quad i = 1, \ldots, n \\ & X \succeq 0 \end{array} \qquad (\textit{PhaseLift})$$

which is a semidefinite program (called *PhaseLift* by Candès et al. [2011]) in the variable $X \in \mathbf{H}_p$. Recent results in [Candès et al., 2013; Candès and Li, 2014] give explicit (if somewhat stringent) conditions on A and x under which the relaxation is tight (i.e. the optimal X in *PhaseLift* is unique, has rank one, with leading eigenvector x).

In Paragraph 2.3.1, we define a variant of *PhaseLift*, easier to compare with *PhaseCut*. In paragraph 2.3.2, we show that this variant and *PhaseCut* can be seen as projection problems. Paragraph 2.3.3 compares the tightness of *PhaseLift* and a slight modification of *PhaseCut*. Paragraph 2.3.4 discusses the stability to noise and paragraph 2.3.6 compares the complexities of both methods.

2.3.1 Weak formulation

We also introduce a weak version of *PhaseLift*, which is more directly related to *PhaseCut* and is easier to interpret geometrically. It was noted in [Candès et al., 2013] that, when $\mathbf{I} \in \text{span}\{a_i a_i^*\}_{i=1}^n$, the condition $\text{Tr}\,(a_i a_i^* X) = b_i^2, i = 1, ..., n$ determines $\text{Tr}\,(X)$, so in this case the trace minimization objective is redundant and *PhaseLift* is equivalent to

$$\begin{array}{ll} \text{find} & X \\ \text{subject to} & \text{Tr}\,(a_i a_i^* X) = b_i^2, \quad i = 1, \dots, n \\ & X \succeq 0. \end{array} \qquad (\textit{Weak PhaseLift})$$

When $\mathbf{I} \notin \text{span}\{a_i a_i^*\}_{i=1}^n$ on the other hand, *Weak PhaseLift* and *PhaseLift* are not equivalent: solutions of *PhaseLift* solve *Weak PhaseLift* too but the converse is not true. Interior point solvers typically pick a solution at the analytic center of the feasible set of *Weak PhaseLift* which in general can be significantly different from the minimum trace solution.

However, in practice, the removal of trace minimization does not really seem to alter the performances of the algorithm. We will illustrate this affirmation with numerical experiments in Paragraph 2.4.4 and a formal proof is given in [Demanet and Hand, 2012; Candès and Li, 2014] who showed that, in the case of Gaussian random measurements, the relaxation of *Weak PhaseLift* is tight with high probability under the same conditions as *PhaseLift*.

2.3.2 Phase recovery as a projection

We will see in what follows that phase recovery can be interpreted as a projection problem. These results will prove useful later to study stability. The *PhaseCut* reconstruction problem is written

$$\begin{array}{ll} \text{minimize} & \text{Tr}\,(UM) \\ \text{subject to} & \text{diag}\,(U) = 1,\ U \succeq 0, \end{array}$$

with $M = \text{diag}\,(b)(\mathbf{I} - AA^\dagger)\text{diag}\,(b)$. In what follows, we assume $b_i \neq 0$, $i = 1, ..., n$, which means that, after scaling U, solving *PhaseCut* is equivalent to solving

$$\begin{array}{ll} \text{minimize} & \text{Tr}\,(V(\mathbf{I} - AA^\dagger)) \\ \text{subject to} & \text{diag}\,(V) = b^2, V \succeq 0. \end{array} \qquad (2.11)$$

In the following lemma, we show that this last semidefinite program can be understood as a projection problem on a section of the semidefinite cone using the trace (or nuclear) norm. We define

$$\mathcal{F} = \{V \in \mathbf{H}_n : x^* V x = 0, \forall x \in \text{Range}\,(A)^\perp\}$$

which is also $\mathcal{F} = \{V \in \mathbf{H}_n : (\mathbf{I} - AA^\dagger)V(\mathbf{I} - AA^\dagger) = 0\}$, and we now formulate the objective of problem (2.11) as a distance.

Lemma 2.2. *For all $V \in \mathbf{H}_n$ such that $V \succeq 0$,*

$$\mathrm{Tr}\left(V(\mathbf{I} - AA^\dagger)\right) = d_1(V, \mathcal{F}) \tag{2.12}$$

where d_1 is the distance associated to the trace norm.

Proof. Let $\mathcal{B}_1$ (resp. $\mathcal{B}_2$) be an orthonormal basis of Range A (resp. (Range $A)^\perp$). Let T be the transformation matrix from canonical basis to orthonormal basis $\mathcal{B}_1 \cup \mathcal{B}_2$. Then

$$\mathcal{F} = \{V \in \mathbf{H}_n \text{ s.t. } T^{-1}VT = \left(\begin{smallmatrix} S_1 & S_2 \\ S_2^* & 0 \end{smallmatrix}\right), S_1 \in \mathbf{H}_p, S_2 \in \mathcal{M}_{p,n-p}\}$$

As the transformation $X \to T^{-1}XT$ preserves the nuclear norm, for every matrix $V \succeq 0$, if we write

$$T^{-1}VT = \left(\begin{smallmatrix} V_1 & V_2 \\ V_2^* & V_3 \end{smallmatrix}\right)$$

then the orthogonal projection of V onto $\mathcal{F}$ is

$$W = T\left(\begin{smallmatrix} V_1 & V_2 \\ V_2^* & 0 \end{smallmatrix}\right)T^{-1},$$

so $d_1(V, \mathcal{F}) = \|V - W\|_1 = \|\left(\begin{smallmatrix} 0 & 0 \\ 0 & V_3 \end{smallmatrix}\right)\|_1$. As $V \succeq 0$, $\left(\begin{smallmatrix} V_1 & V_2 \\ V_2^* & V_3 \end{smallmatrix}\right) \succeq 0$ hence $\left(\begin{smallmatrix} 0 & 0 \\ 0 & V_3 \end{smallmatrix}\right) \succeq 0$, so $d_1(V, \mathcal{F}) = \mathrm{Tr}\left(\begin{smallmatrix} 0 & 0 \\ 0 & V_3 \end{smallmatrix}\right)$. Because $AA^\dagger$ is the orthogonal projection onto Range (A), we have $T^{-1}(\mathbf{I} - AA^\dagger)T = \left(\begin{smallmatrix} 0 & 0 \\ 0 & \mathbf{I} \end{smallmatrix}\right)$ hence

$$d_1(V, \mathcal{F}) = \mathrm{Tr}\left(\begin{smallmatrix} 0 & 0 \\ 0 & V_3 \end{smallmatrix}\right) = \mathrm{Tr}\left((T^{-1}VT)(T^{-1}(\mathbf{I} - AA^\dagger)T)\right) = \mathrm{Tr}\left(V(\mathbf{I} - AA^\dagger)\right)$$

which is the desired result. □

This means that *PhaseCut* can be written as a projection problem, i.e.

$$\begin{array}{ll} \text{minimize} & d_1(V, \mathcal{F}) \\ \text{subject to} & V \in \mathbf{H}_n^+ \cap \mathcal{H}_b \end{array} \tag{2.13}$$

in the variable $V \in \mathbf{H}_n$, where $\mathcal{H}_b = \{V \in \mathbf{H}_n \text{ s.t. } V_{i,i} = b_i^2, i = 1, ..., n\}$. Moreover, with a_i the i-th row of A, we have for all $X \in \mathbf{H}_p^+$, $\mathrm{Tr}\left(a_i a_i^* X\right) = a_i^* X a_i = \mathrm{diag}\left(AXA^*\right)_i, \quad i = 1, \ldots, n$, so if we call $V = AXA^* \in \mathcal{F}$, when A is injective, $X = A^\dagger V A^{\dagger *}$ and *Weak PhaseLift* is equivalent to

$$\begin{array}{ll} \text{find} & V \in \mathbf{H}_n^+ \cap \mathcal{F} \\ \text{subject to} & \mathrm{diag}\left(V\right) = b^2. \end{array}$$

First order algorithms for *Weak PhaseLift* will typically solve

$$\begin{array}{ll} \text{minimize} & d(\mathrm{diag}\left(V\right), b^2) \\ \text{subject to} & V \in \mathbf{H}_n^+ \cap \mathcal{F} \end{array}$$

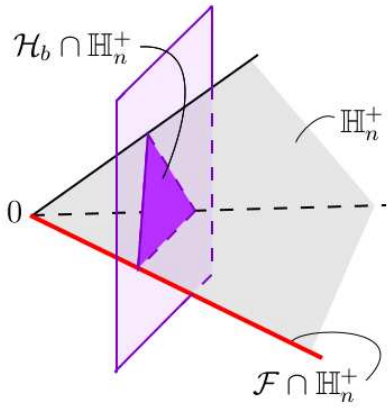


Figure 2.1: Schematic representation of the sets involved in equations (2.13) and (2.14): the cone of positive hermitian matrices $\mathbb{H}_n^+$ (in light grey), its intersection with the affine subspace $\mathcal{H}_b$, and $\mathcal{F} \cap \mathbb{H}_n^+$, which is a face of $\mathbb{H}_n^+$.

for some distance d over $\mathbb{R}^n$. If d is the l^s-norm, for any $s \geq 1$, $d(\mathrm{diag}\,(V), b^2) = d_s(V, \mathcal{H}_b)$, where d_s is the distance generated by the Schatten s-norm, and the algorithm becomes

$$\begin{array}{ll} \text{minimize} & d_s(V, \mathcal{H}_b) \\ \text{subject to} & V \in \mathbf{H}_n^+ \cap \mathcal{F} \end{array} \tag{2.14}$$

which is another projection problem in V.

Thus, *PhaseCut* and *Weak PhaseLift* are comparable, in the sense that both algorithms aim at finding a point of $\mathbb{H}_n^+ \cap \mathcal{F} \cap \mathcal{H}_b$ but *PhaseCut* does so by picking a point of $\mathbb{H}_n^+ \cap \mathcal{H}_b$ and moving towards $\mathcal{F}$ while *Weak PhaseLift* moves a point of $\mathbb{H}_n^+ \cap \mathcal{F}$ towards $\mathcal{H}_b$. We can push the parallel between both relaxations much further. We will show in what follows that, in a very general case, *PhaseLift* and a modified version of *PhaseCut* are simultaneously tight. We will also be able to compare the stability of *Weak PhaseLift* and *PhaseCut* when measurements become noisy.

2.3.3 Tightness of the semidefinite relaxation

We will now formulate a refinement of the semidefinite relaxation in *PhaseCut* and prove that this refinement is equivalent, in terms of tightness, to the relaxation in *PhaseLift* under mild technical assumptions. Suppose u is the optimal phase vector, we know that the optimal solution to (2.1) can then be written $x = A^\dagger \mathrm{diag}\,(b)u$, which corresponds to the matrix $X = A^\dagger \mathrm{diag}\,(b)uu^* \mathrm{diag}\,(b)A^{\dagger *}$ in *PhaseLift*, hence

$$\mathrm{Tr}\,(X) = \mathrm{Tr}\,(\mathrm{diag}\,(b)A^{\dagger *}A^\dagger \mathrm{diag}\,(b)uu^*).$$

Writing $B = \operatorname{diag}(b)A^{\dagger *}A^{\dagger}\operatorname{diag}(b)$, when problem (2.1) is solvable, we look for the "minimum trace" solution among all the optimal points of relaxation *PhaseCut* by solving

$$SDP2(M) \stackrel{def}{=} \begin{array}{ll} \text{min.} & \operatorname{Tr}(BU) \\ \text{subject to} & \operatorname{Tr}(MU) = 0 \\ & \operatorname{diag}(U) = 1,\ U \succeq 0, \end{array} \qquad (PhaseCutMod)$$

which is a semidefinite program in $U \in \mathbf{H}_n$. When problem (2.1) is solvable, then every optimal solution of the semidefinite relaxation *PhaseCut* is a feasible point of relaxation *PhaseCutMod*. In practice, the semidefinite program $SDP(M + \gamma B)$, written

$$\begin{array}{ll} \text{minimize} & \operatorname{Tr}((M + \gamma B)U) \\ \text{subject to} & \operatorname{diag}(U) = 1,\ U \succeq 0, \end{array}$$

obtained by replacing M by $M + \gamma B$ in problem *PhaseCut*, will produce a solution to *PhaseCutMod* whenever $\gamma > 0$ is sufficiently small (this is essentially the exact penalty method detailed in [Bertsekas, 1998, §4.3] for example). This means that all algorithms (greedy or SDP) designed to solve the original *PhaseCut* problem can be recycled to solve *PhaseCutMod* with negligible effect on complexity. We now show that the *PhaseCutMod* and *PhaseLift* relaxations are simultaneously tight when A is injective. An earlier version of this text showed that *PhaseLift* tightness implies *PhaseCutMod* tightness, and the argument was reversed in [Voroninski, 2012] under mild additional assumptions.

Proposition 2.3. *Assume that $b_i \neq 0$ for $i = 1, \ldots, n$, that A is injective and that there is a solution x to (2.1). The function*

$$\begin{array}{rcl} \Phi : \mathbf{H}_p & \to & \mathbf{H}_n \\ X & \mapsto & \Phi(X) = \operatorname{diag}(b)^{-1} A X A^* \operatorname{diag}(b)^{-1} \end{array}$$

is a bijection between the feasible points of PhaseCutMod and those of PhaseLift.

Proof. Note that Φ is injective whenever $b > 0$ and A has full rank. We have to show that U is a feasible point of *PhaseCutMod* if and only if it can be written under the form $\Phi(X)$, where X is feasible for *PhaseLift*. We first show that

$$\operatorname{Tr}(MU) = 0, \quad U \succeq 0, \qquad (2.15)$$

is equivalent to

$$U = \Phi(X) \qquad (2.16)$$

for some $X \succeq 0$. Observe that $\mathrm{Tr}\,(UM) = 0$ means $UM = 0$ because $U, M \succeq 0$, hence $\mathrm{Tr}\,(MU) = 0$ in (2.15) is equivalent to

$$AA^\dagger \mathrm{diag}\,(b) U \mathrm{diag}\,(b) = \mathrm{diag}\,(b) U \mathrm{diag}\,(b)$$

because $b > 0$ and $M = \mathrm{diag}\,(b)(\mathbf{I} - AA^\dagger)\mathrm{diag}\,(b)$. If we set $X = A^\dagger \mathrm{diag}\,(b) U \mathrm{diag}\,(b) A^{\dagger *}$, this last equality implies both

$$AX = AA^\dagger \mathrm{diag}\,(b) U \mathrm{diag}\,(b) A^{\dagger *} = \mathrm{diag}\,(b) U \mathrm{diag}\,(b) A^{\dagger *}$$

and

$$AXA^* = \mathrm{diag}\,(b) U \mathrm{diag}\,(b) A^{\dagger *} A^* = \mathrm{diag}\,(b) U \mathrm{diag}\,(b)$$

which is $U = \Phi(X)$, and shows (2.15) implies (2.16). Conversely, if $U = \Phi(X)$ then:

$$\mathrm{diag}\,(b) U \mathrm{diag}\,(b) = AXA^*$$

and, using $AA^\dagger A = A$, we get $AXA^* = AA^\dagger AXA^* = AA^\dagger \mathrm{diag}\,(b) U \mathrm{diag}\,(b)$ which means $MU = 0$, hence (2.15) is in fact equivalent to (2.16) since $U \succeq 0$ by construction.

Now, if X is feasible for *PhaseLift*, we have shown $\mathrm{Tr}\,(M\Phi(X)) = 0$ and $\phi(X) \succeq 0$, moreover $\mathrm{diag}\,(\Phi(X))_i = \mathrm{Tr}\,(a_i a_i^* X)/b_i^2 = 1$, so $U = \Phi(X)$ is a feasible point of *PhaseCutMod*. Conversely, if U is feasible for *PhaseCutMod*, we have shown that there exists $X \succeq 0$ such that $U = \Phi(X)$ which means $\mathrm{diag}\,(b) U \mathrm{diag}\,(b) = AXA^*$. We also have $\mathrm{Tr}\,(a_i a_i^* X) = b_i^2 U_{ii} = b_i^2$, which means X is feasible for *PhaseLift* and concludes the proof. □

We now have the following central corollary showing the equivalence between *PhaseCutMod* and *PhaseLift* in the noiseless case.

Corollary 2.4. *If A is injective, $b_i \neq 0$ for all $i = 1, ..., n$ and if the reconstruction problem (2.1) admits an exact solution, then PhaseCutMod is tight (i.e. has a unique rank one solution) whenever PhaseLift is.*

Proof. When A is injective, $\mathrm{Tr}\,(X) = \mathrm{Tr}\,(B\Phi(X))$ and $\mathbf{Rank}(X) = \mathbf{Rank}(\Phi(X))$. □

This last result shows that in the noiseless case, the relaxations *PhaseLift* and *PhaseCutMod* are in fact equivalent. In the same way, we could have shown that *Weak PhaseLift* and *PhaseCut* were equivalent. The performances of both algorithms may not match however when the information on b is noisy and perfect recovery is not possible.

Remark 2.5. *Note that Proposition 2.3 and Corollary 2.4 also hold when the initial signal is real and the measurements are complex. In this case, we define the B in PhaseCutMod by $B = B_2 A_2^{\dagger *} A_2^\dagger B_2$ (with the notations of paragraph 2.2.6). We must also replace the definition of Φ by $\Phi(X) = B_2^{-1} A_2 X A_2^* B_2^{-1}$. Furthermore, all steps in the proof of Proposition 2.3 are still valid if we replace M by M_2, A by A_2 and $\mathrm{diag}\,(b)$ by B_2. The only difference is that now $\frac{1}{b_i^2}\mathrm{Tr}\,(a_i a_i^* X) = \mathrm{diag}\,(\Phi(X))_i + \mathrm{diag}\,(\Phi(X))_{n+i}$.*

2.3.4 Stability in the presence of noise

We now consider the case where the vector of measurements b is of the form $b = |Ax_0| + b_{\text{noise}}$. We first introduce a definition of C-stability for *PhaseCut* and *Weak PhaseLift*. The main result of this section is that, when the *Weak PhaseLift* solution in (2.14) is stable at a point, *PhaseCut* is stable too, with a constant of the same order. The converse does not seem to be true when b is sparse.

Definition 2.6. *Let $x_0 \in \mathbb{C}^n, C > 0$. The algorithm PhaseCut (resp. Weak PhaseLift) is said to be C-stable at x_0 iff for all $b_{\text{noise}} \in \mathbb{R}^n$ close enough to zero, every minimizer V of equation (2.13) (resp. (2.14)) with $b = |Ax_0| + b_{\text{noise}}$, satisfies*

$$\|V - (Ax_0)(Ax_0)^*\|_2 \leq C\|Ax_0\|_2\|b_{\text{noise}}\|_2.$$

The following matrix perturbation result motivates this definition, by showing that a C-stable algorithm generates a $O(C\|b_{\text{noise}}\|_2)$-error over the signal it reconstructs.

Proposition 2.7. *Let $C > 0$ be arbitrary. We suppose that $Ax_0 \neq 0$ and $\|V - (Ax_0)(Ax_0)^*\|_2 \leq C\|Ax_0\|_2\|b_{\text{noise}}\|_2 \leq \|Ax_0\|_2^2/2$. Let y be V's main eigenvector, normalized so that $(Ax_0)^*y = \|Ax_0\|_2$. Then*

$$\|y - Ax_0\|_2 = O(C\|b_{\text{noise}}\|_2),$$

and the constant in this last equation does not depend upon A, x_0, C or $\|b\|_2$.

Proof. We use [Karoui and d'Aspremont, 2010, Eq.10] for

$$u = \frac{Ax_0}{\|Ax_0\|_2} \qquad v = \frac{y}{\|Ax_0\|_2} \qquad E = \frac{V - (Ax_0)(Ax_0)^*}{\|Ax_0\|_2^2}$$

This result is based on [Kato, 1995, Eq. 3.29], which gives a precise asymptotic expansion of $u - v$. For our purposes here, we only need the first-order term. See also Bhatia [1997], Stewart and Sun [1990] or Stewart [2001] among others for a complete discussion. We get $\|v - u\| = O(\|E\|_2)$ because if $M = uu^*$, then $\|R\|_\infty = 1$ in [Karoui and d'Aspremont, 2010, Eq.10]. This implies

$$\|y - Ax_0\|_2 = \|Ax_0\|_2\|u - v\| = O\left(\frac{\|V - (Ax_0)(Ax_0)^*\|_2}{\|Ax_0\|_2}\right) = O(C\|b_{\text{noise}}\|)$$

which is the desired result. □

Note that normalizing y differently, we would obtain $\|y - Ax_0\|_2 \leq 4C\|b_{\text{noise}}\|_2$. We now show the main result of this section, according to which *PhaseCut* is "almost as stable as" *Weak PhaseLift*. In numerical applications, the exact values of the stability constants have only a small importance, what matters is that they are of the same order.

Theorem 2.8. *Let $A \in \mathbb{C}^{n\times m}$, for all $x_0 \in \mathbb{C}^n, C > 0$, if Weak PhaseLift is C-stable in x_0, then PhaseCut is $(2C + 2\sqrt{2} + 1)$-stable in x_0.*

Proof. Let $x_0 \in \mathbb{C}^n, C > 0$ be such that *Weak PhaseLift* is C-stable in x_0. Ax_0 is a non-zero vector (because, with our definition, neither *Weak PhaseLift* nor *PhaseCut* may be stable in x_0 if $Ax_0 = 0$ and $A \neq 0$). We set $D = 2C+2\sqrt{2}+1$ and suppose by contradiction that *PhaseCut* is not D-stable in x_0. Let $\epsilon > 0$ be arbitrary. Let $b_{\mathrm{n,PC}} \in \mathbb{R}^n$ be such that $\|b_{\mathrm{n,PC}}\|_2 \leq \max(\|Ax_0\|_2, \epsilon/2)$ and such that, for $b = |Ax_0| + b_{\mathrm{n,PC}}$, the minimizer V_{PC} of (2.13) verifies

$$\|V_{PC} - (Ax_0)(Ax_0)^*\|_2 > D\|Ax_0\|_2\|b_{\mathrm{n,PC}}\|_2$$

Such a V_{PC} must exist or *PhaseCut* would be D-stable in x_0. We call $V_{PC}^{/\!/}$ the restriction of V_{PC} to $\mathrm{Range}\,(A)$ (that is, the matrix such that $x^*(V_{PC}^{/\!/})y = x^*(V_{PC})y$ if $x, y \in \mathrm{Range}\,(A)$ and $x^*(V_{PC}^{/\!/})y = 0$ if $x \in \mathrm{Range}\,(A)^\perp$ or $y \in \mathrm{Range}\,(A)^\perp$) and $V_{PC}^\perp$ the restriction of V_{PC} to $\mathrm{Range}\,(A)^\perp$. Let us set $b_{\mathrm{n,PL}} = \sqrt{V_{PC\,ii}^{/\!/}} - |Ax_0|_{ii}$ for $i = 1, ..., n$. As $V_{PC}^{/\!/} \in \mathbf{H}_n^+ \cap \mathcal{F}$, $V_{PC}^{/\!/}$ minimizes (2.14) for $b = |Ax_0| + b_{\mathrm{n,PL}}$ (because $V_{PC}^{/\!/} \in \mathcal{H}_b$). Lemmas 2.9 and 2.10 (proven in the appendix) imply that $\|V_{PC}^{/\!/} - (Ax_0)(Ax_0)^*\|_2 > C\|Ax_0\|_2\|b_{\mathrm{n,PL}}\|_2$ and $\|b_{\mathrm{n,PL}}\|_2 \leq \epsilon$. As ϵ is arbitrary, *Weak PhaseLift* is not C-stable in x_0, which contradicts our hypotheses. Consequently, *PhaseCut* is $(2C + 2\sqrt{2} + 1)$-stable in x_0. □

The proof of this theorem is based on the fact that, when V_{PC} solves (2.13), one can construct some $V_{PL} = V_{PC}^{/\!/}$ close to V_{PC}, which is an approximate solution of (2.14). It is natural to wonder whether, conversely, from a solution V_{PL} of (2.14), one can construct an approximate solution V_{PC} of (2.13). It does not seem to be the case. One could for example imagine setting $V_{PC} = \mathrm{diag}\,(R)V_{PL}\mathrm{diag}\,(R)$, where $R_i = b_i/\sqrt{V_{PL\,ii}}$. Then V_{PC} would not necessarily minimize (2.13) but at least belong to $\mathcal{H}_b$. But $\|V_{PC} - V_{PL}\|_2$ might be quite large: (2.14) implies that $\|\mathrm{diag}\,(V_{PL}) - b^2\|_s$ is small but, if some coefficients of b are very small, some R_i may still be huge, so $\mathrm{diag}\,(R) \not\approx \mathbf{I}$. This does happen in practice (see Paragraph 2.4.5).

To conclude this section, we relate this definition of stability to the one introduced in [Candès and Li, 2014]. Suppose that A is a matrix of random gaussian independant measurements such that $\mathbb{E}[|A_{i,j}|^2] = 1$ for all i, j. We also suppose that $n \geq c_0 p$ (for some c_0 independent of n and p). In the noisy setting, Candès and Li [2014] showed that the minimizer X of a modified version of *PhaseLift* satisfies with high probability

$$||X - x_0x_0^*||_2 \leq C_0\frac{||\,|Ax_0|^2 - b^2\,||_1}{n} \tag{2.17}$$

for some C_0 independent of all variables. Assuming that the *Weak PhaseLift* solution in (2.14) behaves as *PhaseLift* in a noisy setting and that (2.17) also holds for *Weak PhaseLift*, then

$$||AXA^* - (Ax_0)(Ax_0)^*||_2 \leq ||A||_\infty^2||X - x_0x_0^*||_2$$

$$\leq C_0 \frac{||A||_\infty^2}{n} || \, |Ax_0|^2 - b^2 \, ||_1$$
$$\leq C_0 \frac{||A||_\infty^2}{n} (2||Ax_0||_2 + ||b_{\text{noise}}||_2)||b_{\text{noise}}||_2$$

Consequently, for any $C > 2C_0 \frac{||A||_\infty^2}{n}$, *Weak PhaseLift* is C-stable in all x_0. With high probability, $||A||_\infty^2 \leq (1+1/8)n$ (it is a corollary of [Candès and Li, 2014, Lemma 2.1]) so *Weak PhaseLift* (and thus also *PhaseCut*) is C-stable with high probability for some C independent of all parameters of the problem.

2.3.5 Perturbation results

We recall here sensitivity analysis results for semidefinite programming from Yildirim and Todd [2001]; Yildirim [2003], which produce explicit bounds on the impact of small perturbations in the observation vector b^2 on the solution V of the semidefinite program (2.11). Roughly speaking, these results show that if $b^2 + b_{noise}$ remains in an explicit ellipsoid (called Dikin's ellipsoid), then interior point methods converge back to the solution in one full Newton step, hence the impact on V is linear, equal to the Newton step. These results are more numerical in nature than the stability bounds detailed in the previous section, but they precisely quantify both the size and, perhaps more importantly, the geometry of the stability region.

2.3.6 Complexity comparisons

Both the relaxation in *PhaseLift* and that in *PhaseCut* are semidefinite programs and we highlight below the relative complexity of solving these problems depending on algorithmic choices and precision targets. Note that, in their numerical experiments, [Candès et al., 2011] solve a penalized formulation of problem *PhaseLift*, written

$$\min_{X \succeq 0} \sum_{i=1}^{n} (\text{Tr}\,(a_i a_i^* X) - b_i^2)^2 + \lambda \text{Tr}\,(X) \tag{2.18}$$

in the variable $X \in \mathbf{H}_p$, for various values of the penalty parameter $\lambda > 0$.

The trace norm promotes a low rank solution, and solving a sequence of weighted trace-norm problems has been shown to further reduce the rank in [Fazel et al., 2003; Candès et al., 2011]. This method replaces $\text{Tr}\,(X)$ by $\text{Tr}\,(W_k X)$ where W_0 is initialized to the identity I. Given a solution X_k of the resulting semidefinite program, the weighted matrix is updated to $W_{k+1} = (X_k + \eta I)^{-1}$ (see Fazel et al. [2003] for details). We denote by K the total number of such iterations, typically of the order of 10. Trace minimization is not needed for the semidefinite program (*PhaseCut*), where the trace is fixed because we optimize over a normalized phase vector.

However, weighted trace-norm iterations could potentially improve performance in *PhaseCut* as well.

Recall that p is the size of the signal and n is the number of measured samples with $n = Jp$ in the examples reviewed in Section 2.4. In the numerical experiments in [Candès et al., 2011] as well as in this paper, $J = 3, 4, 5$. The complexity of solving the *PhaseCut* and *PhaseLift* relaxations in *PhaseLift* using generic semidefinite programming solvers such as SDPT3 [Toh et al., 1999], *without exploiting structure*, is given by

$$O\left(J^{4.5}\, p^{4.5} \log \frac{1}{\epsilon}\right) \quad \text{and} \quad O\left(K\, J^2\, p^{4.5} \log \frac{1}{\epsilon}\right)$$

for *PhaseCut* and *PhaseLift* respectively [Ben-Tal and Nemirovski, 2001, § 6.6.3]. The fact that the constraint matrices have only one nonzero coefficient in *PhaseCut* can be exploited (the fact that the constraints $a_i a_i^*$ are rank one in *PhaseLift* helps, but it does not modify the principal complexity term) so we get

$$O\left(J^{3.5}\, p^{3.5} \log \frac{1}{\epsilon}\right) \quad \text{and} \quad O\left(K\, J^2 p^{4.5} \log \frac{1}{\epsilon}\right)$$

for *PhaseCut* and *PhaseLift* respectively using the algorithm in Helmberg et al. [1996] for example. If we use first-order solvers such as TFOCS [Becker et al., 2012], based on the optimal algorithm in [Nesterov, 1983], the dependence on the dimension can be further reduced, to become

$$O\left(\frac{J^3\, p^3}{\epsilon}\right) \quad \text{and} \quad O\left(\frac{K\, J\, p^3}{\epsilon}\right)$$

for a penalized version of the *PhaseCut* relaxation and the penalized formulation of *PhaseLift* in (2.18). While the dependence on the signal dimensions p is somewhat reduced, the dependence on the target precision grows from $\log(1/\epsilon)$ to $1/\epsilon$. Finally, the iteration complexity of the block coordinate descent Algorithm 3 is substantially lower and its convergence is linear, but no fully explicit bounds on the number of iterations are known in our case. The complexity of the method is then bounded by $O\left(\log \frac{1}{\epsilon}\right)$ but the constant in this bound depends on n here, and the dependence cannot be quantified explicitly.

Algorithmic choices are ultimately guided by precision targets. If ϵ is large enough so that a first-order solver or a block coordinate descent can be used, the complexity of *PhaseCut* is not significantly better than that of *PhaseLift*. On the contrary, when ϵ is small, we must use an interior point solver, for which *PhaseCut*'s complexity is an order of magnitude lower than that of *PhaseLift* because its constraint matrices are singletons. In practice, the target value for ϵ strongly depends on the sampling matrix A. For example, when A corresponds to the convolution by 6 Gaussian random filters (Paragraph 2.4.2), to reconstruct a Gaussian white

noise of size 64 with a relative precision of η, we typically need $\epsilon \sim 2.10^{-1}\eta$. For 4 Cauchy wavelets (Paragraph 2.4.3), it is twenty times less, with $\epsilon \sim 10^{-2}\eta$. For other types of signals than Gaussian white noise, we may even need $\epsilon \sim 10^{-3}\eta$.

2.3.7 Greedy refinement

If the *PhaseCut* or *PhaseLift* algorithms do not return a rank one matrix then an approximate solution of the phase recovery problem is obtained by extracting a leading eigenvector v. For *PhaseCut* and *PhaseLift*, $\tilde{x} = A^\dagger \mathrm{diag}\,(b)v$ and $\tilde{x} = v$ are respectively approximate solutions of the phase recovery problem with $|A\tilde{x}| \neq b = |Ax|$. This solution is then refined by applying the Gerchberg-Saxton algorithm initialized with $\tilde{x}$. If $\tilde{x}$ is sufficiently close to x then, according to numerical experiments of Section 2.4, this greedy algorithm converges to $\lambda\, x$ with $|\lambda| = 1$. These greedy iterations require much less operations than *PhaseCut* and *PhaseLift* algorithms, and thus have no significant contribution to the computational complexity.

2.3.8 Sparsity

Minimizing $\mathrm{Tr}\,(X)$ in the *PhaseLift* problem means looking for signals which match the modulus constraints and have minimum ℓ_2 norm. In some cases, we have a priori knowledge that the signal we are trying to reconstruct is sparse, i.e. $\mathbf{Card}(x)$ is small. The effect of imposing sparsity was studied in e.g. [Moravec et al., 2007; Shechtman et al., 2011; Li and Voroninski, 2013].

Assuming $n \leq p$, the set of solutions to $\|Ax - \mathrm{diag}\,(b)u\|_2$ is written $x = A^\dagger \mathrm{diag}\,(b)u + Fv$ where F is a basis for the nullspace of A. In this case, when the rows of A are independent, $AA^\dagger = \mathbf{I}$ and the reconstruction problem with a ℓ_1 penalty promoting sparsity is then written

$$\begin{array}{ll} \text{minimize} & \|A^\dagger \mathrm{diag}\,(b)u + Fv\|_1^2 \\ \text{subject to} & |u_i| = 1, \end{array}$$

in the variables $u \in \mathbb{C}^p$ and $y \in \mathbb{C}^{p-n}$. Using the fact that $\|y\|_1^2 = \|yy^*\|_{\ell_1}$, this can be relaxed as

$$\begin{array}{ll} \text{minimize} & \|VUV^*\|_{\ell_1} \\ \text{subject to} & U \succeq 0,\ |U_{ii}| = 1, \quad i = 1, \ldots, n, \end{array}$$

which is a semidefinite program in the (larger) matrix variable $U \in \mathbf{H}_p$ and $V = (A^\dagger \mathrm{diag}\,(b), F)$.

On the other hand, when $n > p$ and A is injective, the matrix F disappears. We can take sparsity into account by adding an l_1 penalization to *PhaseCut*. As noted in [Voroninski, 2012] however, the effect of an ℓ_1 penalty on least-squares solutions is not completely clear.

2.4 Numerical results

In this section, we compare the numerical performance of the Gerchberg-Saxton, *PhaseCut* and *PhaseLift* algorithms on various phase recovery problems. As in [Candès et al., 2011], the *PhaseLift* problem is solved using the package in [Becker et al., 2012], with reweighting, using $K = 10$ outer iterations and 1000 iterations of the first order algorithm. The *PhaseCut* and Gerchberg-Saxton algorithms described here are implemented in a public software package available at

`http://www.cmap.polytechnique.fr/scattering/code/phaserecovery.zip`

Other numerical experiments about *PhaseCut* have been conducted in the optic of applications to imaging problems, and can be found in [Fogel et al., 2013].

In our experiments, the phase recovery algorithms compute an approximate solution $\tilde{x}$ from $|Ax|$ and the reconstruction error is measured by the relative Euclidean distance up to a complex phase given by

$$\epsilon(x, \tilde{x}) \stackrel{def}{=} \min_{c \in \mathbb{C}, |c|=1} \frac{\|x - c\,\tilde{x}\|}{\|x\|}. \tag{2.19}$$

We also record the error over measured amplitudes, written

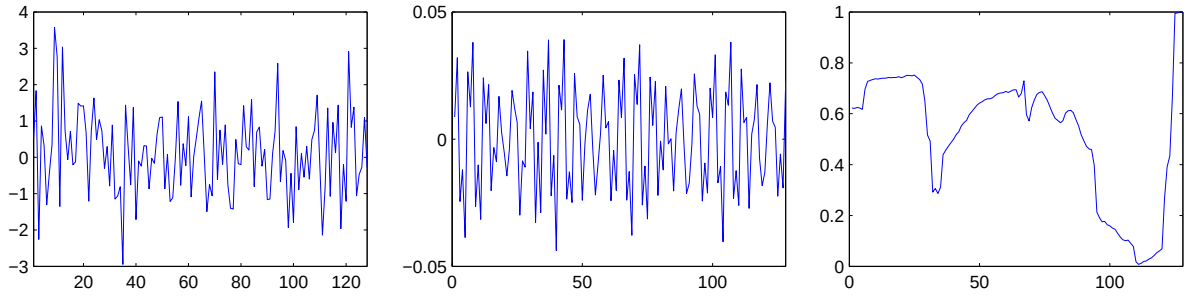
$$\epsilon(|Ax|, |A\tilde{x}|) \stackrel{def}{=} \frac{\||Ax| - |A\tilde{x}|\|}{\|Ax\|}. \tag{2.20}$$

Note that when the phase recovery problem either does not admit a unique solution or is unstable, we usually have $\epsilon(|Ax|, |A\tilde{x}|) \ll \epsilon(x, \tilde{x})$. In the next three subsections, we study these reconstruction errors for three different phase recovery problems, where A is defined as an oversampled Fourier transform, as multiple filterings with random filters, or as a wavelet transform. Numerical results are computed on three different types of test signals x: realizations of a complex Gaussian white noise, sums of complex exponentials $a_\omega\, e^{i\omega m}$ with random frequencies ω and random amplitudes a_ω (the number of exponentials is random, around 6), and signals whose real and imaginary parts are scan-lines of natural images. Each signal has $p = 128$ coefficients. Figure 2.2 shows the real part of sample signals, for each signal type.

2.4.1 Oversampled Fourier transform

The discrete Fourier transform $\hat{y}$ of a signal y of q coefficients is written

$$\hat{y}_k = \sum_{m=0}^{q-1} y_m \exp\left(\frac{-i2\pi km}{q}\right).$$



(a) (b) (c)

Figure 2.2: Real parts of sample test signals. (a) Gaussian white noise. (b) Sum of 6 sinuoids of random frequencies and amplitudes. (c) Scan-line of an image.

In X-ray crystallography or diffraction imaging experiments, compactly supported signals are estimated from the amplitude of Fourier transforms oversampled by a factor $J \geq 2$. The corresponding operator A computes an oversampled discrete Fourier transform evaluated over $n = Jp$ coefficients. The signal x of size p is extended into x^J by adding $(J-1)p$ zeros and

$$(Ax)_k = \hat{x}^J_k = \sum_{m=1}^{p} x_m \exp(-\frac{i2\pi km}{n}).$$

For this oversampled Fourier transform, the phase recovery problem does not have a unique solution [Akutowicz, 1956]. In fact, one can show [Sanz, 1985] that there are as many as 2^{p-1} solutions $\tilde{x} \in \mathbb{C}^p$ such that $|A\tilde{x}| = |Ax|$. Moreover, increasing the oversampling factor J beyond 2 does not reduce the number of solutions.

Because of this intrinsic instability, we will observe that all algorithms perform similarly on this type of reconstruction problems and Table 2.1 shows that the percentage of perfect reconstruction is below 5% for all methods. The only signals which can be perfectly recovered are sums of few sinusoids. Because these test signals are very sparse in the Fourier domain, the number of signals having identical Fourier coefficient amplitudes is considerably smaller than in typical sample signals. As a consequence, there is a small probability (about 5%) of exactly reconstructing the original signal given an arbitrary initialization. None of the Gaussian random noises and image scan lines are exactly recovered. Note that we say that an exact reconstruction is reached when $\epsilon(x, \tilde{x}) < 10^{-2}$ because a few iterations of the Gerchberg-Saxton algorithm from such an approximate solution $\tilde{x}$ will typically converges to x. Numerical results are computed with 100 sample signals in each of the 3 signal classes.

Table 2.2 gives the average relative error $\epsilon(x, \tilde{x})$ over signals that are not perfectly reconstructed, which is of order one here. Despite this large error, Table 2.3 shows that the relative error $\epsilon(|Ax|, |A\tilde{x}|)$ over the Fourier modulus coefficients is below 10^{-3} for all algorithms. This is

	Fourier	Random Filters	Wavelets
Gerchberg-Saxton	5%	49%	0%
PhaseLift with reweighting	3%	100%	62%
PhaseCut	4%	100%	100%

Table 2.1: Percentage of perfect reconstruction from $|Ax|$, over 300 test signals, for the three different operators A (columns) and the three algorithms (rows).

	Fourier	Random Filters	Wavelets
Gerchberg-Saxton	0.9	1.2	1.3
PhaseLift with reweighting	0.8	exact	0.5
PhaseCut	0.8	exact	exact

Table 2.2: Average relative signal reconstruction error $\epsilon(\tilde{x}, x)$ over all test signals that are not perfectly reconstructed, for each operator A and each algorithm.

due to the non-uniqueness of the phase recovery from Fourier modulus coefficients. Recovering a solution $\tilde{x}$ with identical or nearly identical oversampled Fourier modulus coefficients as x does not guarantee that $\tilde{x}$ is proportional to x.

Overall, in this set of ill-posed Fourier experiments, recovery performance is very poor for all methods and the *PhaseCut* and *PhaseLift* relaxations do not improve much on the results of the faster Gerchberg-Saxton algorithm.

2.4.2 Multiple random illumination filters

To guarantee uniqueness of the phase recovery problem, one can add independent measurements by "illuminating" the object through J filters h^j in the context of X-ray imaging or crystallography [Candès et al., 2013]. The resulting operator A is the discrete Fourier transform

	Fourier	Random Filters	Wavelets
Gerchberg-Saxton	9.10^{-4}	0.2	0.3
PhaseLift with reweighting	5.10^{-4}	exact	8.10^{-2}
PhaseCut	6.10^{-4}	exact	exact

Table 2.3: Average relative error $\epsilon(|A\tilde{x}|, |Ax|)$ of coefficient amplitudes, over all test signals that are not perfectly reconstructed, for each operator A and each algorithm.

of x multiplied by each filter h^j of size p

$$(Ax)_{k+pj} = \widehat{(xh^j)}_k = (\widehat{x} \star \hat{h}^j)_k \quad \text{for } 1 \leq j \leq J \text{ and } 0 \leq k < p,$$

where $\hat{x} \star \hat{h}^j$ is the circular convolution between $\hat{x}$ and $\hat{h}^j$.

Candès et al. [2015]; Gross et al. [2015b] prove that, in a close setting to this one, *PhaseLift* is tight with high probability, for a number J of filters logarithmic in p. Candès et al. [2011] empirically observe that, for signals of size $p = 128$, with $J = 4$ filters, perfect recovery is achieved in 100% of their experiments.

Table 2.1 confirms this behavior and shows that the *PhaseCut* algorithm achieves perfect recovery in all our experiments. As predicted by the equivalence results presented in the previous section, we observe that *PhaseCut* and *PhaseLift* have identical performance in these experiments. With 4 filters, the solutions of these two SDP relaxations are not of rank one but are "almost" of rank one, in the sense that their first eigenvector v has an eigenvalue much larger than the others, by a factor of about 5 to 10. Numerically, we observe that the corresponding approximate solutions, $\tilde{x} = \operatorname{diag}(v)b$, yield a relative error $\epsilon(|Ax|, |A\tilde{x}|)$ which, for scan-lines of images and especially for Gaussian signals, is of the order of the ratio between the largest and the second largest eigenvalue of the matrix U. The resulting solutions $\tilde{x}$ are then sufficiently close to x so that a few iterations of the Gerchberg-Saxton algorithm started at $\tilde{x}$ will converge to x.

Table 2.1 shows however that directly applying the Gerchberg-Saxton algorithm starting from a random initialization point yields perfect recovery in only about 50% of our experiments. This percentage decreases as the signal size p increases. The average error $\epsilon(x, \tilde{x})$ on non-recovered signals in Table 2.2 is 1.3 whereas on the average error on the modulus $\epsilon(|Ax|, |A\tilde{x}|)$ is 0.2.

2.4.3 Wavelet transform

Finally, we test the algorithm on the phase retrieval problem that will be discussed in the next two chapters of this thesis: the case of the wavelet transform.

To simplify experiments, we consider wavelets dilated by dyadic factors 2^j, which have a lower frequency resolution than audio wavelets. A discrete wavelet transform is computed by circular convolutions with discrete wavelet filters, i.e.

$$(Ax)_{k+jp} = (x \star \psi^j)_k = \sum_{m=1}^{p} x_m \psi^j_{k-m} \quad \text{for } 1 \leq j \leq J-1 \text{ and } 1 \leq k \leq p$$

where ψ^j_m is a p periodic wavelet filter. It is defined by dilating, sampling and periodizing a

complex wavelet $\psi \in \mathbf{L^2}(\mathbb{C})$, with

$$\psi_m^j = \sum_{k=-\infty}^{\infty} \psi(2^j(m/p - k)) \quad \text{for } 1 \leq m \leq p.$$

Numerical computations are performed with a Cauchy wavelet whose Fourier transform is, up to a scaling factor

$$\hat{\psi}(\omega) = \omega^d \, e^{-\omega} \, \mathbf{1}_{\omega > 0},$$

with $d = 5$. To guarantee that A is an invertible operator, the lowest signal frequencies are carried by a suitable low-pass filter ϕ and

$$(Ax)_{k+Jp} = (x \star \phi)_k \quad \text{for } 1 \leq k \leq p.$$

One can prove that x is always uniquely determined by $|Ax|$, up to a multiplication factor (this will be done in Chapter 3).

We consider the case of real signals x; recall that the results of Paragraph 2.2.6 allow us to explicitly impose the condition that x is real in the *PhaseCut* recovery algorithm. For *PhaseLift* in Candès et al. [2011], this condition is enforced by imposing that $X = xx^*$ is real. For the Gerchberg-Saxton algorithm, when x is real, we simply project at each iteration on the image of $\mathbb{R}^p$ by A, instead of projecting on the image of $\mathbb{C}^p$ by A.

Numerical experiments are performed on the real part of the complex test signals. Table 2.1 shows that Gerchberg-Saxton does not reconstruct exactly any test signal from the modulus of its wavelet coefficients. The average relative error $\epsilon(\tilde{x}, x)$ in Table 2.2 is 1.2 where the coefficient amplitudes have an average error $\epsilon(|A\tilde{x}|, |Ax|)$ of 0.3 in Table 2.3.

PhaseLift reconstructs 62% of test signals, but the reconstruction rate varies with the signal type. The proportions of exactly reconstructed signals among random noises, sums of sinusoids and image scan-lines are 27%, 60% and 99% respectively. Indeed, image scan-lines have a large proportion of wavelet coefficients whose amplitudes are negligible. The proportion of phase coefficients having a strong impact on the reconstruction of x is thus much smaller for scan-line images than for random noises, which reduces the number of significant variables to recover. Sums of sinuoids of random frequency have wavelet coefficients whose sparsity is intermediate between image scan-lines and Gaussian white noises, which explains the intermediate performance of *PhaseLift* on these signals. The overall average error $\epsilon(\tilde{x}, x)$ on non-reconstructed signals is 0.5. Despite this relatively important error, $\tilde{x}$ and x are usually almost equal on most of their support, up to a sign switch, and the importance of the error is precisely due to the number of sign switches.

The *PhaseCut* algorithm exactly reconstructs all test signals. Moreover, the recovered matrix U is always of rank one and it is therefore not necessary to refine the solution with Gerchberg-Saxton iterations. At first sight, this difference in performance between *PhaseCut* and *PhaseLift*

may seem to contradict the equivalence results of Paragraph 2.3.3 (which are valid both when x is real and when x is complex). It can be explained however by the fact that 10 steps of reweighting and 1000 inner iterations per step are not enough to let *PhaseLift* fully converge. In these experiments, the precision required to get perfect reconstruction is very high and, consequently, the number of first-order iterations required to achieve it is too large (see Paragraph 2.3.6). With an interior-point-solver, this number would be much smaller but the time required per iteration would become prohibitively large. The much simpler structure of the *PhaseCut* relaxation allows us to solve these larger problems more efficiently.

2.4.4 Impact of trace minimization

We saw in Paragraph 2.3.1 that, in the absence of noise, *PhaseCut* was very similar to a simplified version of *PhaseLift*, *Weak PhaseLift*, in which no trace minimization is performed. Here, we confirm empirically that *Weak PhaseLift* and *PhaseLift* are essentially equivalent. Minimizing the trace is usually used as rank minimization heuristic, with recovery guarantees in certain settings [Fazel et al., 2003; Candès and Recht, 2009; Chandrasekaran et al., 2012] but it does not seem to make much difference here. In fact, Demanet and Hand [2012]; Candès and Li [2014] showed that in the setting where measurements are randomly chosen according to independant Gaussian laws, *Weak PhaseLift* has a unique (rank one) solution with high probability, i.e. the feasible set of *PhaseLift* is a singleton and trace minimization has no impact. Of course, from a numerical point of view, solving the feasibility problem *Weak PhaseLift* is about as hard as solving the trace minimization problem *PhaseLift*, so this result simplifies analysis but does not really affect numerical performance.

Figure 2.3 compares the performances of *PhaseLift* and *Weak PhaseLift* as a function of n (the number of measurements). We plot the percentage of successful reconstructions (*left*) and the percentage of cases where the relaxation was exact, i.e. the reconstructed matrix X was rank one (*right*). The plot shows clear phase transitions when the number of measurements increases. For *PhaseLift*, these transitions happen respectively at $n = 155 \approx 2.5p$ and $n = 285 \approx 4.5p$, while for *Weak PhaseLift*, the values become $n = 170 \approx 2.7p$ and $n = 295 \approx 4.6p$, so the transition thresholds are very similar. Note that, in the absence of noise, *Weak PhaseLift* and *PhaseCut* have the same solutions, up to a linear transformation (see Paragraph 2.3.2), so we can expect the same results when comparing *PhaseCut* with *PhaseCutMod*.

2.4.5 Reconstruction in the presence of noise

Numerical stability is crucial for practical applications. In this last subsection, we suppose that the vector b of measurements is of the form

$$b = |Ax| + b_{\text{noise}}$$

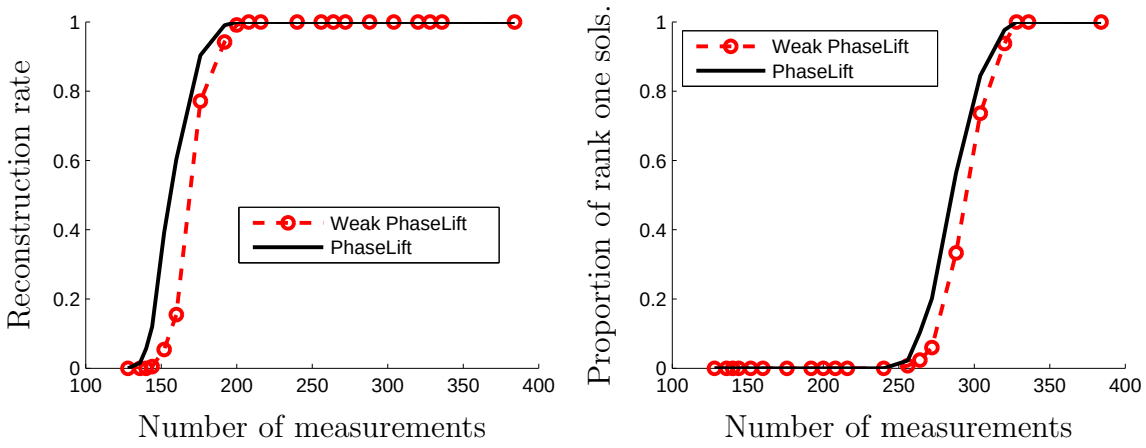


Figure 2.3: Comparison of *PhaseLift* and *Weak PhaseLift* performance, for 64-sized signals, as a function of the number of measurements. Reconstruction rate, after Gerchberg-Saxton iterations *(left)* and proportion of rank one solutions *(right)*.

with $\|b_{\text{noise}}\|_2 = o(\|Ax\|_2)$. In our experiments, b_{noise} is always a Gaussian white noise.

Two reasons can explain numerical instabilities in the solution $\tilde{x}$. First, the reconstruction problem itself can be unstable, with $\|\tilde{x} - cx\| \gg \||A\tilde{x}| - |Ax|\|$ for all $c \in \mathbb{C}$. Second, the algorithm may fail to reconstruct $\tilde{x}$ such that $\||A\tilde{x}| - b\| \approx \|b_{\text{noise}}\|$. No algorithm can overcome the first cause but good reconstruction methods will overcome the second one. In the following paragraphs, to complement the results in Paragraph 2.3.4, we will demonstrate empirically that *PhaseCut* is stable, and compare its performances with *PhaseLift*. We will observe in particular that *PhaseCut* appears to be more stable than *PhaseLift* when b is sparse.

Wavelet transform

Figure 2.4 displays the performance of *PhaseCut* in the wavelet transform case. It shows that *PhaseCut* is stable up to around $5 - 10\%$ of noise. Indeed, the reconstructed $\tilde{x}$ usually satisfies $\epsilon(|Ax|, |A\tilde{x}|) = \| \, |Ax| - |A\tilde{x}| \, \|_2 \leq \|b_{\text{noise}}\|_2$, which is the best we can hope for. Wavelet transform is a case where the underlying phase retrieval problem may present instabilities, therefore the reconstruction error $\epsilon(x, \tilde{x})$ is sometimes much larger than $\epsilon(|Ax|, |A\tilde{x}|)$. This remark applies especially to sums of sinusoids, which represent the most unstable case.

When all coefficients of Ax have approximately the same amplitude, *PhaseLift* and *PhaseCut* produce similar results, but when Ax is sparse, *PhaseLift* appears less stable. We gave a qualitative explanation of this behavior at the end of Paragraph 2.3.4 which seems to be confirmed by the results in Figure 2.4. Indeed, the performance of *PhaseLift* and *PhaseCut* are equivalent in the case of Gaussian random filters (where measurements are never sparse), they are a bit worse

in the case of sinusoids (where measurements are sometimes sparse) and quite unsatisfactory for scan-lines of images (where measurements are always sparse).

Multiple random illumination filters

Candès and Li [2014] prove that, if A is a Gaussian matrix, the reconstruction problem is stable with high probability, and *PhaseLift* reconstructs a $\tilde{x}$ such that

$$\epsilon(\tilde{x}, x) \leq O\left(\frac{\|b_{\text{noise}}\|_2}{\|Ax\|_2}\right).$$

The same result seems to hold for A corresponding to Gaussian random illumination filters (cf. Paragraph 2.4.2). Moreover, *PhaseCut* is as stable as *PhaseLift*. Actually, up to 20% of noise, when followed by some Gerchberg-Saxton iterations, *PhaseCut* and *PhaseLift* almost always reconstruct the same function. Figure 2.5 displays the corresponding empirical performance, confirming that both algorithms are stable. The relative reconstruction errors are approximately linear in the amount of noise, with

$$\epsilon(|A\tilde{x}|, |Ax|) \approx 0.8 \times \frac{\|b_{\text{noise}}\|_2}{\|Ax\|_2} \quad \text{and} \quad \epsilon(\tilde{x}, x) \approx 2 \times \frac{\|b_{\text{noise}}\|_2}{\|Ax\|_2}$$

in our experiments.

The impact of the sparsity of b discussed in the last paragraph may seem irrelevant here: if A and x are independently chosen, Ax is never sparse. However, if we do not choose A and x independently, we may achieve partial sparsity. We performed tests for the case of five Gaussian random filters, where we chose $x \in \mathbb{C}^{64}$ such that $(Ax)_k = 0$ for $k \leq 60$. This choice has no particular physical interpretation but it allows us to check that the influence of sparsity in $|Ax|$ over *PhaseLift* is not specific to the wavelet transform. Figure 2.5 displays the relative error over the reconstructed matrix in the sparse and non-sparse cases. If we denote by $X_{\text{pl}} \in \mathbb{C}^{p \times p}$ (resp. $X_{\text{pc}} \in \mathbb{C}^{n \times n}$) the matrix reconstructed by *PhaseLift* (resp. *PhaseCut*), this relative error is defined by

$$\epsilon = \frac{\|AX_{\text{pl}}A^* - (Ax)(Ax)^*\|_2}{\|(Ax)(Ax)^*\|_2} \qquad (\text{for } \textit{PhaseLift})$$

$$\epsilon = \frac{\|\text{diag}\,(b) X_{\text{pc}} \text{diag}\,(b) - (Ax)(Ax)^*\|_2}{\|(Ax)(Ax)^*\|_2} \qquad (\text{for } \textit{PhaseCut})$$

In the non-sparse case, both algorithms yield very similar error $\epsilon \approx 7\|b_{\text{noise}}\|_2/\|Ax\|_2$ (the difference for a relative noise of 10^{-4} may come from a computational artifact). In the sparse case, there are less phases to reconstruct, because we do not need to reconstruct the phase of

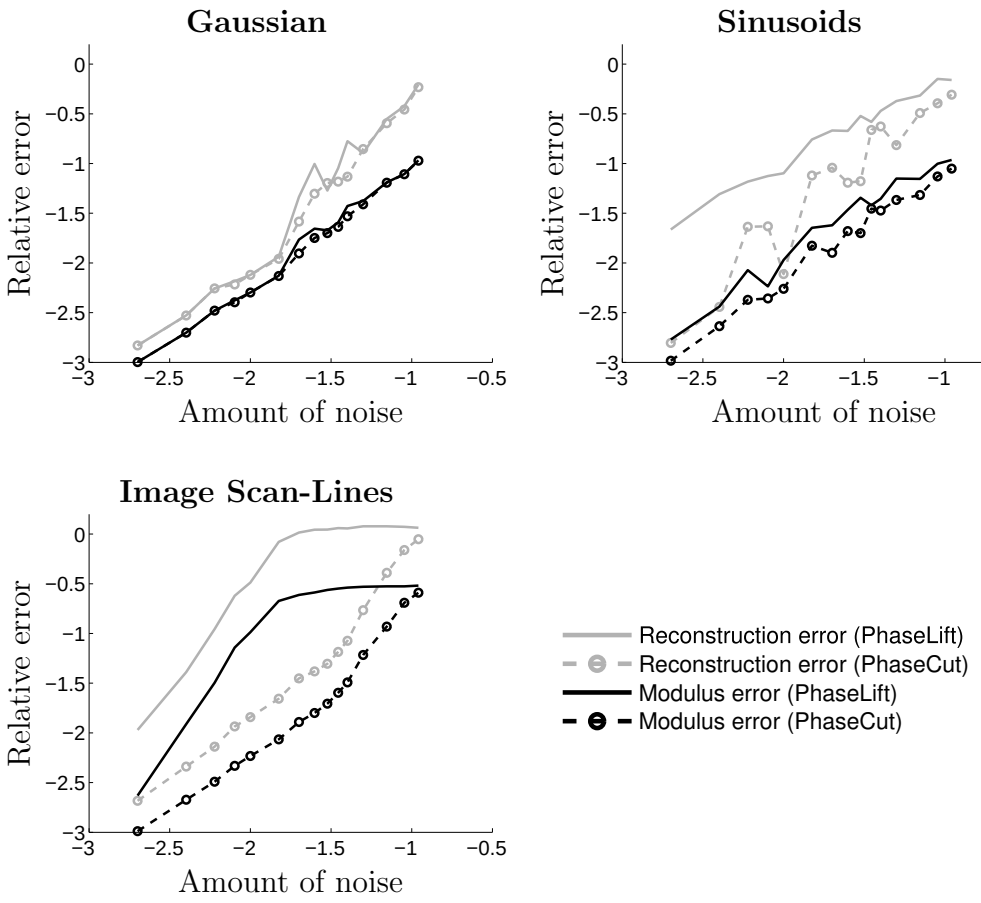


Figure 2.4: Mean reconstruction errors versus amount of noise for *PhaseLift* and *PhaseCut*, both in decimal logarithmic scale, for three types of signals: Gaussian white noises, sums of sinusoids and scan-lines of images. Both algorithms were followed by a few hundred Gerchberg-Saxton iterations.

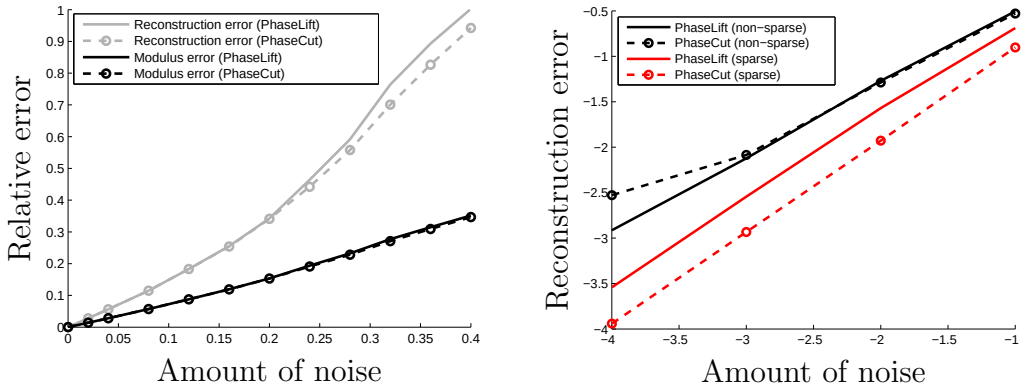


Figure 2.5: *Left:* Mean performances of *PhaseLift* and *PhaseCut*, followed by Gerchberg-Saxton iterations, for four Gaussian random illumination filters. The x-axis represents the relative noise level, $\|b_{\text{noise}}\|_2/\|Ax\|_2$ and the y-axis the relative error on the result, which is either $\epsilon(\tilde{x}, x)$ or $\epsilon(|A\tilde{x}|, |Ax|)$. *Right:* Loglog plot of the relative error over the matrix reconstructed by *PhaseLift* (resp. *PhaseCut*) when A represents the convolution by five Gaussian filters. Black curves correspond to Ax non-sparse, red ones to sparse Ax.

null measurements. Consequently, the problem is better constrained and we expect the algorithms to be more stable. Indeed, the relative errors over the reconstructed matrices are smaller. However, in this case, the performance of *PhaseLift* and *PhaseCut* do not match anymore: $\epsilon \approx 3\|b_{\text{noise}}\|_2/\|Ax\|_2$ for *PhaseLift* and $\epsilon \approx 1.2\|b_{\text{noise}}\|_2/\|Ax\|_2$ for *PhaseCut*. This remark has no practical impact in our particular example here because taking a few Gerchberg-Saxton iterations would likely make both methods converge towards the same solution, but it confirms the importance of accounting for the sparsity of $|Ax|$.

2.5 Technical lemmas

We now prove two technical lemmas used in the proof of Theorem 2.8.

Lemma 2.9. *Under the assumptions and notations of Theorem 2.8, we have*

$$\|V_{PC}^{/\!/} - (Ax_0)(Ax_0)^*\|_2 > 2C\|Ax_0\|_2\|b_{\text{n,PC}}\|_2$$

Proof. We first give an upper bound of $\|V_{PC} - V_{PC}^{/\!/}\|_2$. We use the Cauchy-Schwarz inequality: for every positive matrix X and all x, y, $|x^*Xy| \leq \sqrt{x^*Xx}\sqrt{y^*Xy}$. Let $\{f_i\}$ be an hermitian base of Range (A) diagonalizing $V_{PC}^{/\!/}$ and $\{g_i\}$ an hermitian base of Range $(A)^{\perp}$ diagonalizing

$V_{PC}^{\perp}$. As $\{f_i\} \cap \{g_i\}$ is an hermitian base of $\mathbb{C}^n$, we have

$$\begin{aligned}
\|V_{PC} - V_{PC}^{/\!/}\|_2^2 &= \sum_{i,i'} |f_i^*(V_{PC} - V_{PC}^{/\!/}) f_{i'}|^2 + \sum_{i,j} |f_i^*(V_{PC} - V_{PC}^{/\!/}) g_j|^2 \\
&\quad + \sum_{i,j} |g_j^*(V_{PC} - V_{PC}^{/\!/}) f_i|^2 + \sum_{j,j'} |g_j^*(V_{PC} - V_{PC}^{/\!/}) g_{j'}|^2 \\
&= 2\sum_{i,j} |f_i^*(V_{PC}) g_j|^2 + \sum_i |g_i^*(V_{PC}^{\perp}) g_i|^2 \\
&\leq 2\sum_{i,j} |f_i^*(V_{PC}) f_i \| g_j^*(V_{PC}) g_j| + (\sum_i g_i^*(V_{PC}^{\perp}) g_i)^2 \\
&= 2\mathrm{Tr}\, V_{PC}^{/\!/} \mathrm{Tr}\, V_{PC}^{\perp} + (\mathrm{Tr}\, V_{PC}^{\perp})^2 \\
&\leq \left(\sqrt{2} \sqrt{\mathrm{Tr}\, V_{PC}^{/\!/}} \sqrt{\mathrm{Tr}\, V_{PC}^{\perp}} + \mathrm{Tr}\, V_{PC}^{\perp} \right)^2 \qquad (2.21)
\end{aligned}$$

Let us now bound $\mathrm{Tr}\, V_{PC}^{\perp}$. We first note that $\mathrm{Tr}\, V_{PC}^{\perp} = \mathrm{Tr}\,((\mathbf{I} - AA^\dagger) V_{PC} (\mathbf{I} - AA^\dagger)) = \mathrm{Tr}\,(V_{PC}(\mathbf{I} - AA^\dagger)) = d_1(V_{PC}, \mathcal{F})$ (according to Lemma 2.2). Let $u \in \mathbb{C}^n$ be such that, for all i, $|u_i| = 1$ and $(Ax_0)_i = u_i |Ax_0|_i$. We set $b = |Ax_0| + b_{\mathrm{n,PC}}$ and $V = (b \times u)(b \times u)^*$. As $V \in \mathbf{H}_n^+ \cap \mathcal{H}_b$ and V_{PC} minimizes (2.13),

$$\begin{aligned}
\mathrm{Tr}\, V_{PC}^{\perp} &= d_1(V_{PC}, \mathcal{F}) \leq d_1(V, \mathcal{F}) = d_1((Ax_0 + b_{\mathrm{n,PC}} u)(Ax_0 + b_{\mathrm{n,PC}} u)^*, \mathcal{F}) \\
&= d_1((b_{\mathrm{n,PC}} u)(b_{\mathrm{n,PC}} u)^*, \mathcal{F}) \\
&\leq \|(b_{\mathrm{n,PC}} u)(b_{\mathrm{n,PC}} u)^*\|_1 = \mathrm{Tr}\,(b_{\mathrm{n,PC}} u)(b_{\mathrm{n,PC}} u)^* = \|b_{\mathrm{n,PC}}\|_2^2
\end{aligned}$$

We also have $\mathrm{Tr}\, V_{PC}^{/\!/} = \mathrm{Tr}\, V_{PC} - \mathrm{Tr}\, V_{PC}^{\perp}$. This equality comes from the fact that, if $\{f_i\}$ is an hermitian base of $\mathrm{Range}\,(A)$ and $\{g_i\}$ an hermitian base of $\mathrm{Range}\,(A)^{\perp}$, then

$$\mathrm{Tr}\, V_{PC} = \sum_i f_i V_{PC} f_i^* + \sum_i g_i V_{PC} g_i^* = \sum_i f_i V_{PC}^{/\!/} f_i^* + \sum_i g_i V_{PC}^{\perp} g_i^* = \mathrm{Tr}\, V_{PC}^{/\!/} + \mathrm{Tr}\, V_{PC}^{\perp}$$

As $V_{PC}^{\perp} \succeq 0$, $\mathrm{Tr}\, V_{PC}^{/\!/} \leq \mathrm{Tr}\, V_{PC} = \||Ax_0| + b_{\mathrm{n,PC}}\|_2^2$ and, by combining this with relations (2.21) and (2.22), we get

$$\begin{aligned}
\|V_{PC} - V_{PC}^{/\!/}\|_2 &\leq \sqrt{2} \||Ax_0| + b_{\mathrm{n,PC}}\|_2 \|b_{\mathrm{n,PC}}\|_2 + \|b_{\mathrm{n,PC}}\|_2^2 \\
&\leq \sqrt{2} \|Ax_0\|_2 \|b_{\mathrm{n,PC}}\|_2 + (1 + \sqrt{2}) \|b_{\mathrm{n,PC}}\|_2^2
\end{aligned}$$

And, by reminding that we assumed $\|b_{\mathrm{n,PC}}\|_2 \leq \|Ax_0\|_2$,

$$\|V_{PC}^{/\!/} - (Ax_0)(Ax_0)^*\|_2 \geq \|V_{PC} - (Ax_0)(Ax_0)^*\|_2 - \|V_{PC}^{/\!/} - V_{PC}\|_2$$

$$> D\|Ax_0\|_2\|b_{\mathrm{n,PC}}\|_2 - \sqrt{2}\|Ax_0\|_2\|b_{\mathrm{n,PC}}\|_2 - (1+\sqrt{2})\|b_{\mathrm{n,PC}}\|_2^2$$
$$\geq (D - 2\sqrt{2} - 1)\|Ax_0\|_2\|b_{\mathrm{n,PC}}\|_2 = 2C\|Ax_0\|_2\|b_{\mathrm{n,PC}}\|_2$$

which concludes the proof. □

Lemma 2.10. *Under the assumptions and notations of Theorem 2.8, we have* $\|b_{\mathrm{n,PL}}\|_2 \leq 2\|b_{\mathrm{n,PC}}\|$.

Proof. Let e_i be the i-th vector of $\mathbb{C}^n$'s canonical base. We set $e_i = f_i + g_i$ where $f_i \in \mathrm{Range}\,(A)$ and $g_i \in \mathrm{Range}\,(A)^\perp$.

$$\begin{aligned} V_{PC\,ii} &= e_i^* V_{PC} e_i \\ &= f_i^* V_{PC}^{/\!/} f_i + 2\mathrm{Re}\,(f_i^* V_{PC} g_i) + g_i^* V_{PC}^{\perp} g_i \\ &= V_{PC\,ii}^{/\!/} + 2\mathrm{Re}\,(f_i^* V_{PC} g_i) + V_{PC\,ii}^{\perp} \end{aligned}$$

Because $|f_i^* V_{PC} g_i| \leq \sqrt{f_i^* V_{PC} f_i}\sqrt{g_i^* V_{PC} g_i} = \sqrt{V_{PC\,ii}^{/\!/}}\sqrt{V_{PC\,ii}^{\perp}}$,

$$\begin{aligned} &(\sqrt{V_{PC\,ii}^{/\!/}} - \sqrt{V_{PC\,ii}^{\perp}})^2 \leq V_{PC\,ii} \leq (\sqrt{V_{PC\,ii}^{/\!/}} + \sqrt{V_{PC\,ii}^{\perp}})^2 \\ &\Rightarrow \sqrt{V_{PC\,ii}^{/\!/}} - \sqrt{V_{PC\,ii}^{\perp}} \leq \sqrt{V_{PC\,ii}} \leq \sqrt{V_{PC\,ii}^{/\!/}} + \sqrt{V_{PC\,ii}^{\perp}} \end{aligned}$$

So

$$\begin{aligned} |b_{\mathrm{n,PL},i}| &= |\sqrt{V_{PC\,ii}^{/\!/}} - |Ax_0|_i| \\ &\leq |\sqrt{V_{PC\,ii}^{/\!/}} - \sqrt{V_{PC\,ii}}| + |\sqrt{V_{PC\,ii}} - |Ax_0|_i| \\ &\leq \sqrt{V_{PC\,ii}^{\perp}} + b_{\mathrm{n,PC},i} \end{aligned}$$

and, by (2.22),

$$\begin{aligned} \|b_{\mathrm{n,PL}}\|_2 &\leq \| \left\{ \sqrt{V_{PC\,ii}^{\perp}} \right\}_i \|_2 + \|b_{\mathrm{n,PC}}\|_2 \\ &= \sqrt{\mathrm{Tr}\, V_{PC}^{\perp}} + \|b_{\mathrm{n,PC}}\|_2 \leq 2\|b_{\mathrm{n,PC}}\|_2 \end{aligned}$$

which concludes the proof. □

Acknowledgments

The authors are grateful to Richard Baraniuk, Emmanuel Candès, Rodolphe Jenatton, Amit Singer and Vlad Voroninski for very constructive comments. In particular, Vlad Voroninski showed in [Voroninski, 2012] that the argument in the first version of this text, proving that *PhaseCutMod* is tight when *PhaseLift* is, could be reversed under mild technical conditions and pointed out an error in our handling of sparsity constraints. AA would like to acknowledge support from a starting grant from the European Research Council (project SIPA), and SM acknowledges support from ANR grant BLAN 012601.

Chapter 3

Phase retrieval for the Cauchy wavelet transform

After presenting an algorithm for solving generic phase retrieval problems in the previous chapter, we now consider a specific phase retrieval problem: the case of the wavelet transform. A wavelet family $(\psi_j)_{j\in\mathbb{Z}}$ being fixed, the problem is:

$$\text{reconstruct } f \in L^2(\mathbb{R}) \text{ from } \{|f \star \psi_j|\}_{j\in\mathbb{Z}}$$

In this chapter, we study the well-posedness of this problem, in terms of uniqueness and stability: is any function f uniquely determined by its wavelet transform modulus? Is the reconstruction stable to noise?

We restrict ourselves to a specific class of wavelets, namely Cauchy wavelets, whose link with harmonic analysis makes it easier to study.

Besides its applications in audio processing, this problem has a theoretical interest in the general context of phase retrieval.

The wavelet transform is one of the only known natural[1] operators for which the phase retrieval problem is now known to be well-posed (uniqueness of the reconstruction, and a form of stability). It in particular contrasts with the Fourier transform [Walther, 1963; Akutowicz, 1956] and the fractional Fourier transform [Jaming, 2014], which can be studied with the same tools: in the first case, there is no uniqueness; in the second one, there is uniqueness, but there is no stability in a strong sense, and no result for a weaker form of stability is known.

Moreover, the notion of *local stability* which appears in the case of the wavelet transform is new to phase retrieval.

[1] "natural" in the sense that it appears in other fields than phase recovery

The uniqueness result is Corollary 3.2. It shows that two analytical signals whose wavelet transforms are identical in modulus are equal up to a global phase. A signal f is said to be analytical if $\hat{f}(\omega) = 0$ when $\omega < 0$. This condition may seem restrictive, but it is actually not, because almost the same result holds for real-valued signals (Corollary 3.3), which covers all conceivable applications.

In Theorem 3.11, we prove that the reconstruction operator is continuous. This is a weak notion of stability to noise, and we show that no strong stability holds, in the sense that, for any $\epsilon > 0$, there exist functions f, g such that:

$$\left|\left|\{|f \star \psi_j|\}_j - \{|g \star \psi_j|\}_j\right|\right|_2 < \epsilon \qquad \text{but} \qquad ||f - g||_2 \geq 1$$

So two functions can have almost the same wavelet transform modulus without being close in the L^2-norm sense.

However, we have a form of local stability. Theorems 3.17 and 3.18 can be approximately summarized by the following informal assertion: if the wavelet transform modulus of two functions are close, then the wavelet transforms are close, up to a global phase, in the neighborhood of each point (j, t) of the time-frequency plane, except maybe at points where the wavelet transform is close to zero.

Our proof techniques naturally yield a reconstruction algorithm. We have implemented it; it yields precise reconstruction results and is relatively stable to noise. Because it only works for Cauchy wavelets, it has limited application. However, it allows us to empirically confirm our theoretical results.

Section 3.1 is concerned with uniqueness results. Section 3.2 proves the weak stability theorem. Section 3.3 explains why there is no strong stability. The local stability results are proved in Section 3.4. Section 3.5 describes the algorithm and our numerical experiments. Finally, Section 3.6 proves useful technical lemmas.

The results of this chapter have been published in [Mallat and Waldspurger, 2014].

Notations

For any $f \in L^1(\mathbb{R})$, we denote by $\hat{f}$ or $\mathcal{F}(f)$ the Fourier transform of f:

$$\hat{f}(\omega) = \int_{\mathbb{R}} f(x) e^{-i\omega x} dx \quad \forall \omega \in \mathbb{R}$$

We extend this definition to L^2 by continuity.

We denote by $\mathcal{F}^{-1} : L^2(\mathbb{R}) \to L^2(\mathbb{R})$ the inverse Fourier transform and recall that, for any $f \in L^1 \cap L^2(\mathbb{R})$:

$$\mathcal{F}^{-1}(f)(x) = \frac{1}{2\pi} \int_{\mathbb{R}} \hat{f}(\omega) e^{i\omega x} d\omega$$

We denote by $\mathbb{H}$ the Poincaré half-plane: $\mathbb{H} = \{z \in \mathbb{C} \text{ s.t. } \operatorname{Im} z > 0\}$.

3.1 Uniqueness of the reconstruction for Cauchy wavelets

3.1.1 Definition of the wavelet transform; comparison with Fourier

The most important phase retrieval problem, which naturally arises in several physical settings, is the case of the Fourier transform:

$$\text{reconstruct } f \in L^2(\mathbb{R}) \text{ from } |\hat{f}|$$

Without additional assumptions over f, the reconstruction is clearly impossible: any choice of phase $\phi : \mathbb{R} \to \mathbb{R}$ yield a signal $g = \mathcal{F}^{-1}(|\hat{f}|e^{i\phi}) \in L^2(\mathbb{R})$ such that $|\hat{g}| = |\hat{f}|$.

To avoid this problem, one may for example require that f is compactly supported. However, Akutowicz [1957]; Walther [1963] showed that, even with this constraint, the reconstruction is still not possible.

More precisely, their result is the following one. If $f \in L^2(\mathbb{R})$ is a compactly supported function, then its Fourier transform $\hat{f}$ admits a holomorphic extension F over all $\mathbb{C}$: $F(z) = \int_{\mathbb{R}} f(x) e^{-izx} dx$. If $g \in L^2(\mathbb{R})$ is another compactly supported function and G is this holomorphic extension of its Fourier transform, the equality $|\hat{f}| = |\hat{g}|$ happens to be equivalent to:

$$\forall z \in \mathbb{C}, \quad F(z)\overline{F(\overline{z})} = G(z)\overline{G(\overline{z})}$$

This in turn is essentially equivalent to:

$$\{z_n\} \cup \{\overline{z}_n\} = \{z'_n\} \cup \{\overline{z}'_n\} \tag{3.1}$$

where the (z_n) and (z'_n) are the respective zeros of F and G over $\mathbb{C}$, counted with multiplicity. This means that F and G must have the same zeros, up to symmetry with respect to the real axis.

Conversely, for every choice of $\{z'_n\}$ satisfying (3.1), it is possible to find a compactly supported g such that the zeroes of G are the z'_n, which implies $|\hat{f}| = |\hat{g}|$.

A similar result can be established in the case where the function $f \in L^2(\mathbb{R})$ is assumed to be identically zero on the negative real line [Akutowicz, 1956] instead of compactly supported.

Let us now define the wavelet transform and compare it with the Fourier transform.

Let $\psi \in L^1 \cap L^2(\mathbb{R})$ be a wavelet, that is a function such that $\int_{\mathbb{R}} \psi(x)dx = 0$. Let $a > 1$ be fixed; we call a the *dilation factor*. We define a family of wavelets by:

$$\forall x \in \mathbb{R} \quad \psi_j(x) = a^{-j}\psi(a^{-j}x) \qquad \Leftrightarrow \qquad \forall \omega \in \mathbb{R} \quad \hat{\psi}_j(\omega) = \hat{\psi}(a^j\omega)$$

The wavelet transform operator is:

$$f \in L^2(\mathbb{R}) \rightarrow \{f \star \psi_j\}_{j\in\mathbb{Z}} \in (L^2(\mathbb{R}))^{\mathbb{Z}}$$

This operator is unitary if the so-called Littlewood-Paley condition is satisfied:

$$\left(\sum_j |\hat{\psi}_j(\omega)|^2 = 1, \forall \omega \in \mathbb{R}\right) \qquad \Rightarrow \qquad \left(||f||_2^2 = \sum_j ||f \star \psi_j||_2^2 \quad \forall f \in L^2(\mathbb{R})\right) \tag{3.2}$$

The phase retrieval problem associated with this operator is:

$$\text{reconstruct } f \in L^2(\mathbb{R}) \text{ from } \{|f \star \psi_j|\}_{j\in\mathbb{Z}}$$

This problem may or may not be well-posed, depending on which wavelet family we use.

The simplest case is the one where the wavelets are Shannon wavelets:

$$\hat{\psi} = 1_{[1;a]} \qquad \Rightarrow \qquad \forall j \in \mathbb{Z}, \quad \hat{\psi}_j(\omega) = 1_{[a^{-j};a^{-j+1}]}$$

Reconstructing f amounts to reconstruct $\hat{f}1_{[a^{-j};a^{-j+1}]} = \hat{f}\hat{\psi}_j$ for all j. For each j, we have only two pieces of information about $\hat{f}\hat{\psi}_j$: its support is included in $[a^{-j}; a^{-j+1}]$ and the modulus of its inverse Fourier transform is $|f \star \psi_j|$. From the results of the Fourier transform case, it is not enough to determine uniquely $\hat{f}\hat{\psi}_j$. Thus, for Shannon wavelets, the phase retrieval problem is as ill-posed as for the Fourier transform.

In this example, the problem comes from the fact that the $\hat{\psi}_j$ have non-overlapping supports. Thus, reconstructing f is equivalent to reconstructing independently each $f \star \psi_j$, and that is not possible.

However, in general, the $\hat{\psi}_j$ have overlapping supports and the $f \star \psi_j$ are not independent for different values of j. They satisfy the following relation:

$$(f \star \psi_j) \star \psi_k = (f \star \psi_k) \star \psi_j \quad \forall j, k \in \mathbb{Z} \tag{3.3}$$

Thus, there is "redundancy" in the wavelet decomposition of f. We can hope that this redundancy compensates the loss of phase of $|f \star \psi_j|$. In the following, we show that, at least for specific wavelets, it is the case.

3.1.2 Uniqueness theorem for Cauchy wavelets

In this paragraph, we consider wavelets of the following form:

$$\begin{aligned} \hat{\psi}(\omega) &= \rho(\omega)\omega^p e^{-\omega} 1_{\omega>0} \\ \hat{\psi}_j(\omega) &= \hat{\psi}(a^j\omega) \quad \forall \omega \in \mathbb{R} \end{aligned} \tag{3.4}$$

where $p > 0$ and $\rho \in L^\infty(\mathbb{R})$ is such that $\rho(a\omega) = \rho(\omega)$ for almost every $\omega \in \mathbb{R}$ and $\rho(\omega) \neq 0, \forall \omega$.

The presence of ρ allows some flexibility in the choice of the family. In particular, if it is properly chosen, the Littlewood-Paley condition (3.2) may be satisfied. However, the proofs are the same with or without ρ.

When $\rho = 1$, the wavelets of the form (3.4) are called *Cauchy wavelets of order p*. The figure 3.1 displays an example of such wavelets. For these wavelets, the wavelet transform has the property to be a set of sections of a holomorphic function along horizontal lines.

If $f \in L^2(\mathbb{R})$, its analytic part f_+ is defined by:

$$\hat{f}_+(\omega) = 2\hat{f}(\omega)1_{\omega>0} \tag{3.5}$$

We define:

$$F(z) = \frac{1}{2\pi}\int_{\mathbb{R}} \omega^p \hat{f}_+(\omega) e^{i\omega z} d\omega \qquad \forall z \text{ s.t. } \operatorname{Im} z > 0 \tag{3.6}$$

When f_+ is sufficiently regular, F is the holomorphic extension of its p-th derivative.

For each $y > 0$, if we denote by $F(. + iy)$ the function $x \in \mathbb{R} \to F(x + iy)$:

$$F(. + iy) = \mathcal{F}^{-1}\left(2\omega^p \hat{f}(\omega) 1_{\omega>0} e^{-y\omega}\right)$$

Consequently, for each $j \in \mathbb{Z}$:

$$\frac{a^{pj}}{2} F(. + ia^j) = f \star \psi_j \quad \forall j \in \mathbb{Z} \tag{3.7}$$

So $f \star \psi_j$ is the restriction of F to the horizontal line $\mathbb{R} + ia^j$. In this case, the relation (3.3) is equivalent to the fact that, for all j, k, $f \star \psi_j$ and $f \star \psi_k$ are the restrictions of the *same* holomorphic function to the lines $\mathbb{R} + ia^j$ and $\mathbb{R} + ia^k$.

Reconstructing f_+ from $\{|f \star \psi_j|\}_{j\in\mathbb{Z}}$ now amounts to reconstruct the holomorphic function $F : \mathbb{H} = \{z \in \mathbb{C}, \operatorname{Im} z > 0\} \to \mathbb{C}$ from its modulus on an infinite set of horizontal lines. The figure 3.2 shows these lines for $a = 2$. Our phase retrieval problem thus reduces to a harmonic analysis problem. Actually, knowing $|F|$ on only two lines is already enough to recover F and one of the two lines may even be $\mathbb{R}$, the boundary of $\mathbb{H}$.

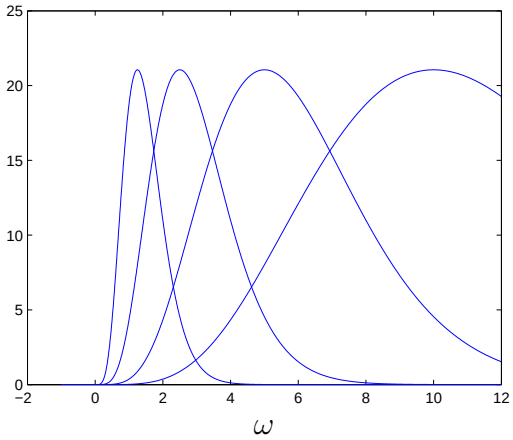


Figure 3.1: Cauchy wavelets of order $p = 5$ for $j = 2, 1, 0, -1$, $a = 2$

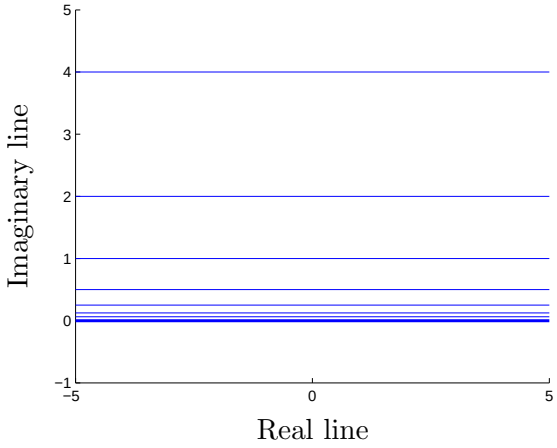


Figure 3.2: Lines in $\mathbb{C}$ over which $|F|$ is known (for $a = 2$)

Theorem 3.1. *Let $\alpha > 0$ be fixed. Let $F, G : \mathbb{H} \to \mathbb{C}$ be holomorphic functions such that, for some $M > 0$:*

$$\int_{\mathbb{R}} |F(x+iy)|^2 dx < M \quad \text{and} \quad \int_{\mathbb{R}} |G(x+iy)|^2 dx < M \qquad \forall y > 0 \tag{3.8}$$

We suppose that:

$$|F(x+i\alpha)| = |G(x+i\alpha)| \text{ for a.e. } x \in \mathbb{R}$$
$$\lim_{y \to 0^+} |F(x+iy)| = \lim_{y \to 0^+} |G(x+iy)| \text{ for a.e. } x \in \mathbb{R}$$

Then, for some $\phi \in \mathbb{R}$:

$$F = e^{i\phi} G \tag{3.9}$$

The proof is given in section 3.1.4.

Corollary 3.2. *We consider wavelets $(\psi_j)_{j \in \mathbb{Z}}$ of the form (3.4). Let $f, g \in L^2(\mathbb{R})$ be such that, for some $j, k \in \mathbb{Z}$ with $j \neq k$:*

$$|f \star \psi_j| = |g \star \psi_j| \quad \text{and} \quad |f \star \psi_k| = |g \star \psi_k| \tag{3.10}$$

We denote by f_+ and g_+ the analytic parts of f and g (as defined in (3.5))
There exists $\phi \in \mathbb{R}$ such that:

$$f_+ = e^{i\phi} g_+ \tag{3.11}$$

Another uniqueness result for the wavelet transform can be found in [Jaming, 2014, Thm 4.4]. However, it is of different nature, because it concerns wavelet transforms with continuous frequency parameter.

Proof. We may assume that $j < k$. We define F and G as in (3.6), with the additional ρ:

$$F(z) = \frac{1}{2\pi}\int_{\mathbb{R}} \omega^p \rho(\omega)\hat{f}_+(\omega)e^{i\omega z}d\omega \qquad G(z) = \frac{1}{2\pi}\int_{\mathbb{R}} \omega^p \rho(\omega)\hat{g}_+(\omega)e^{i\omega z}d\omega \qquad \forall z \in \mathbb{H}$$

For each $y > 0$, $F(. + iy) = \mathcal{F}^{-1}(2\omega^p\rho(\omega)e^{-y\omega}1_{\omega>0}\hat{f}(\omega))$. For $y = a^j$ and $y = a^k$, it implies $F(. + ia^j) = \frac{2}{a^{jp}} f \star \psi_j$ and $F(. + ia^k) = \frac{2}{a^{kp}} f \star \psi_k$. From (3.10):

$$|F(. + ia^j)| = \frac{2}{a^{pj}}|f \star \psi_j| = \frac{2}{a^{pj}}|g \star \psi_j| = |G(. + ia^j)|$$
$$|F(. + ia^k)| = \frac{2}{a^{pk}}|f \star \psi_k| = \frac{2}{a^{pk}}|g \star \psi_k| = |G(. + ia^k)|$$

So the functions $F(. + ia^j)$ and $G(. + ia^j)$ coincide in modulus on two horizontal lines: $\mathbb{R}$ and $\mathbb{R} + i(a^k - a^j)$. From Theorem 3.1, they are equal up to a global phase. As ρ does not vanish, it implies that f_+ and g_+ are equal up to this global phase.

In order to be able to apply Theorem 3.1, we must verify that the condition (3.8) holds for $F(. + ia^j)$ and $G(. + ia^j)$. For any $y > a^j$:

$$\begin{aligned} & F(. + iy) = \mathcal{F}^{-1}\left(2\omega^p\rho(\omega)\hat{f}(\omega)e^{-y\omega}\right) \\ \Rightarrow \quad & ||F(. + iy)||_2^2 = \frac{1}{2\pi}||2\omega^p\rho(\omega)\hat{f}(\omega)e^{-y\omega}1_{\omega\geq 0}||_2^2 \\ & \leq \frac{1}{2\pi}||2\omega^p\rho(\omega)\hat{f}(\omega)e^{-a^j\omega}1_{\omega\geq 0}||_2^2 \\ & = \left(\frac{2}{a^{jp}}\right)^2 ||f \star \psi_j||_2^2 \end{aligned}$$

The same inequality holds for G: the condition (3.8) is true for $M = \left(\frac{2}{a^{jp}}\right)^2 ||f \star \psi_j||_2^2$. □

We have just proven that the modulus of the wavelet transform uniquely determines, up to a global phase, the analytic part of a function, that is its positive frequencies. On the contrary, as wavelets are analytic ($\hat{\psi}_j(\omega) = 0$ if $\omega < 0$), the wavelet transform contains no information about the negative frequencies. In practice, signals are often real so negative frequencies are determined by positive ones and this latter limitation is not really important.

Corollary 3.3. *Let $f, g \in L^2(\mathbb{R})$ be real-valued functions; f_+ and g_+ are their analytic parts. We assume that, for some $j, k \in \mathbb{Z}$ such that $j \neq k$:*

$$|f \star \psi_j| = |g \star \psi_j| \quad \text{and} \quad |f \star \psi_k| = |g \star \psi_k|$$

Then, for some $\phi \in \mathbb{R}$:

$$f_+ = e^{i\phi} g_+ \quad \Leftrightarrow \quad f = \mathrm{Re}\,(e^{i\phi} g_+)$$

Remark 3.4. *Although Corollary 3.2 holds for only two wavelets and does not require $|f \star \psi_s| = |g \star \psi_s|$ for each $s \in \mathbb{Z}$, the reconstruction of f from only two components, $|f \star \psi_j|$ and $|f \star \psi_k|$, is very unstable in practice. Indeed, $\hat{\psi}_j$ and $\hat{\psi}_k$ are concentrated around characteristic frequencies of order 2^{-j} and 2^{-k}. Thus, from $f \star \psi_j$ and $f \star \psi_k$ (and even more so from $|f \star \psi_j|$ and $|f \star \psi_k|$), reconstructing the frequencies of f which are not close to 2^{-j} or 2^{-k} is numerically impossible. It is an ill-conditioned deconvolution problem.*

Before ending this section, let us note that, with a proof similar to the one of Corollary 3.2, Theorem 3.1 also implies the following result.

Corollary 3.5. *Let $\alpha > 0$ be fixed. Let $f, g \in L^2(\mathbb{R})$ be such that $\hat{f}(\omega) = \hat{g}(\omega) = 0$ for every $\omega < 0$.*

If $|\hat{f}| = |\hat{g}|$ and $|\widehat{f(t)e^{-\alpha t}}| = |\widehat{g(t)e^{-\alpha t}}|$, then, for some $\phi \in \mathbb{R}$:

$$f = e^{i\phi} g$$

This says that there is uniqueness in the phase retrieval problem associated to the masked Fourier transform, in the case where there are two masks, $t \to 1$ and $t \to e^{-\alpha t}$.

3.1.3 Discrete case

Naturally, the functions we have to deal with in practice are generally not in $L^2(\mathbb{R})$. They are instead discrete finite signals. In this section, we explain how to switch from the continuous to the discrete finite setting. As we will see, all results derived in the continuous case have a discrete equivalent but proofs become simpler because they use polynomials instead of holomorphic functions.

Let $f \in \mathbb{C}^n$ be a discrete function. We assume n is even. The discrete Fourier transform of f is:

$$\hat{f}[k] = \sum_{s=0}^{n-1} f[s] e^{-\frac{2\pi i s k}{n}} \qquad \text{for } k = -\frac{n}{2} + 1, ..., \frac{n}{2}$$

The analytic part of f is $f_+ \in \mathbb{C}^n$ such that:

$$\hat{f}_+[k] = 0 \text{ if } -\frac{n}{2} + 1 \leq k < 0$$

$$\hat{f}_+[k] = \hat{f}[k] \text{ if } k = 0 \text{ or } k = \frac{n}{2}$$
$$\hat{f}_+[k] = 2\hat{f}[k] \text{ if } 0 < k < \frac{n}{2}$$

When f is real, $f = \mathrm{Re}\,(f_+)$.

We consider wavelets of the following form, for $p > 0$ and $a > 1$:

$$\hat{\psi}_j[k] = \rho(a^j k)(a^j k)^p e^{-a^j k} 1_{k \geq 0} \quad \text{for all } j \in \mathbb{Z}, k = -\frac{n}{2} + 1, ..., \frac{n}{2} \tag{3.12}$$

where $\rho : \mathbb{R}^+ \to \mathbb{C}$ is such that $\rho(ax) = \rho(x)$ for every x and ρ does not vanish.

As in the continuous case, the set $\{|f \star \psi_j|\}_{j \in \mathbb{Z}}$ almost uniquely determines f_+. Naturally, the global phase still cannot be determined. The mean value of f_+ can also not be determined, because $\hat{\psi}_j[0] = 0$ for all j. To determine the mean value and the global phase, we would need some additional information, for example the value of $f \star \phi$ for some low frequency signal ϕ.

Theorem 3.6 (Discrete version of 3.2)**.** *Let $f, g \in \mathbb{C}^n$ be discrete signals and $(\psi_j)_{j \in \mathbb{Z}}$ a family of wavelets of the form (3.12). Let $j, l \in \mathbb{Z}$ be two distinct integers. Then:*

$$|f \star \psi_j| = |g \star \psi_j| \quad \textit{and} \quad |f \star \psi_l| = |g \star \psi_l| \tag{3.13}$$

if and only if, for some $\phi \in \mathbb{R}, c \in \mathbb{C}$:

$$f_+ = e^{i\phi} g_+ + c$$

Proof. We first assume $f_+ = e^{i\phi} g_+ + c$. Taking the Fourier transform of this equality yields:

$$\hat{f}[k] = e^{i\phi} \hat{g}[k] \qquad \text{for all } k = 1, ..., \frac{n}{2}$$

As $\hat{\psi}_j[k] = 0$ for $k = -\frac{n}{2} + 1, ..., 0$:

$$\hat{f}[k]\hat{\psi}_j[k] = e^{i\phi}\hat{g}[k]\hat{\psi}_j[k] \qquad \text{for all } k = -\frac{n}{2} + 1, ..., \frac{n}{2}$$
$$\Rightarrow \quad (f \star \psi_j = e^{i\phi}(g \star \psi_j))$$

So $|f \star \psi_j| = |g \star \psi_j|$ and, similarly, $|f \star \psi_l| = |g \star \psi_l|$.

We now suppose conversely that $|f \star \psi_j| = |g \star \psi_j|$ and $|f \star \psi_l| = |g \star \psi_l|$. We define:

$$F(z) = \frac{1}{n} \sum_{k=1}^{n/2} \hat{f}[k] \rho(k) k^p z^k \qquad G(z) = \frac{1}{n} \sum_{k=1}^{n/2} \hat{g}[k] \rho(k) k^p z^k \qquad \forall z \in \mathbb{C}$$

These polynomials are the discrete equivalents of functions F and G used in the proof of 3.2. For all $s = -\frac{n}{2} + 1, ..., \frac{n}{2}$:

$$\begin{aligned} F(e^{-a^j} e^{\frac{2\pi i s}{n}}) &= \frac{1}{n} \sum_{k=1}^{n/2} \hat{f}[k] \rho(k) k^p e^{-a^j k} e^{\frac{2\pi i k s}{n}} \\ &= a^{-jp} \frac{1}{n} \sum_{k=-n/2+1}^{n/2} \hat{f}[k] \hat{\psi}_j[k] e^{\frac{2\pi i k s}{n}} \\ &= a^{-jp} \left(f \star \psi_j[s] \right) \end{aligned}$$

Similarly, $G(e^{-a^j} e^{\frac{2\pi i s}{n}}) = a^{-jp}(g \star \psi_j[s])$ for all $s = -\frac{n}{2} + 1, ..., \frac{n}{2}$.

Thus, $f \star \psi_j$ and $g \star \psi_j$ can be seen as the restrictions of F and G to the circle of radius e^{-a^j}. This is similar to the continuous case, where $f \star \psi_j$ and $g \star \psi_j$ were the restrictions of functions F, G to horizontal lines.

The equality (3.13) implies:

$$\begin{aligned} &\left| F(e^{-a^j} e^{\frac{2\pi i s}{n}}) \right|^2 = \left| G(e^{-a^j} e^{\frac{2\pi i s}{n}}) \right|^2 \qquad \text{for all } s = -\frac{n}{2} + 1, ..., \frac{n}{2} \\ \Leftrightarrow\ & F(e^{-a^j} e^{\frac{2\pi i s}{n}}) \overline{F}(e^{-a^j} e^{-\frac{2\pi i s}{n}}) = G(e^{-a^j} e^{\frac{2\pi i s}{n}}) \overline{G}(e^{-a^j} e^{-\frac{2\pi i s}{n}}) \qquad \text{for all } s = -\frac{n}{2} + 1, ..., \frac{n}{2} \end{aligned}$$

The functions $z \to F(e^{-a^j} z)\overline{F}(e^{-a^j} \frac{1}{z})$ and $z \to G(e^{-a^j} z)\overline{G}(e^{-a^j} \frac{1}{z})$ are polynomials of degree $n-2$ (up to multiplication by $z^{n/2-1}$). They share n common values so they are equal. The same is true for l instead of j so:

$$F(e^{-a^j} z)\overline{F}\left(e^{-a^j} \frac{1}{z}\right) = G(e^{-a^j} z)\overline{G}\left(e^{-a^j} \frac{1}{z}\right) \qquad \forall z \in \mathbb{C} \tag{3.14}$$

$$F(e^{-a^l} z)\overline{F}\left(e^{-a^l} \frac{1}{z}\right) = G(e^{-a^l} z)\overline{G}\left(e^{-a^l} \frac{1}{z}\right) \qquad \forall z \in \mathbb{C} \tag{3.15}$$

If we show that these equalities imply $F = e^{i\phi} G$ for some $\phi \in \mathbb{R}$, the proof will be finished. Indeed, from the definition of F and G, we will then have $\hat{f}[k] = e^{i\phi} \hat{g}[k]$ for all $k = 1, ..., \frac{n}{2}$ so $\hat{f}_+[k] = e^{i\phi} \hat{g}_+[k]$ for all $k \neq 0$. It implies $f_+ = e^{i\phi} g_+ + c$ for $c = \frac{1}{n} \left(\hat{f}_+[0] - e^{i\phi} \hat{g}_+[0] \right)$.

It suffices to show that F and G have the same roots (with multiplicity) because then, they will be proportional and, from (3.14), (3.15), the proportionality constant must be of modulus 1.

For each $z \in \mathbb{C}$, let $\mu_F(z)$ (resp. $\mu_G(z)$) be the multiplicity of z as a root of F (resp. G). The polynomials of (3.14) are of respective degree $n - 2\mu_F(0)$ and $n - 2\mu_G(0)$ so $\mu_F(0) = \mu_G(0)$.

For all $z \neq 0$, the multiplicity of $e^{a^j} z$ as a zero of (3.14) is:

$$\mu_F(z) + \mu_F\left(\frac{e^{-2a^j}}{\overline{z}}\right) = \mu_G(z) + \mu_G\left(\frac{e^{-2a^j}}{\overline{z}}\right)$$

and the multiplicity of $e^{2a^j - a^l} z$ as a zero of (3.15) is:

$$\mu_F(e^{2(a^j-a^l)}z) + \mu_F\left(\frac{e^{-2a^j}}{\overline{z}}\right) = \mu_G(e^{2(a^j-a^l)}z) + \mu_G\left(\frac{e^{-2a^j}}{\overline{z}}\right)$$

Subtracting this last equality to the previous one implies that, for all z:

$$\mu_F(z) - \mu_G(z) = \mu_F(e^{2(a^j-a^l)}z) - \mu_G(e^{2(a^j-a^l)}z)$$

By applying this equality several times, we get, for all $n \in \mathbb{N}$:

$$\begin{aligned}
\mu_F(z) - \mu_G(z) &= \mu_F(e^{2(a^j-a^l)}z) - \mu_G(e^{2(a^j-a^l)}z) \\
&= \mu_F(e^{4(a^j-a^l)}z) - \mu_G(e^{4(a^j-a^l)}z) \\
&= ... \\
&= \mu_F(e^{2n(a^j-a^l)}z) - \mu_G(e^{2n(a^j-a^l)}z)
\end{aligned}$$

As F and G have a finite number of roots, $\mu_F(e^{2n(a^j-a^l)}z) - \mu_G(e^{2n(a^j-a^l)}z) = 0$ if n is large enough. So $\mu_F(z) = \mu_G(z)$ for all $z \in \mathbb{C}$. □

As in the section 3.1.2, a very similar proof gives a uniqueness result for the case of the Fourier transform with masks, if the masks are well-chosen.

Theorem 3.7 (Discrete version of 3.5)**.** *Let $\alpha > 0$ be fixed. Let $f, g \in \mathbb{C}^{2n-1}$ be two discrete signals with support in $\{0, ..., n-1\}$:*

$$f[s] = g[s] = 0 \text{ for } s = n, ..., 2n-2$$

If $|\hat{f}| = |\hat{g}|$ and $|\widehat{f[s]e^{-s\alpha}}| = |\widehat{g[s]e^{-s\alpha}}|$, then, for some $\phi \in \mathbb{R}$:

$$f = e^{i\phi} g$$

Remark that this theorem describes systems of $4n - 2$ linear measurements whose moduli are enough to recover each complex signal of dimension n. It is known that $4n - 4$ generic measurements always achieve this property ([Balan et al., 2006; Conca et al., 2015]). However, it is in general difficult to find deterministic systems for which it can be proven.

3.1.4 Proof of Theorem 3.1

Theorem (3.1). *Let $\alpha > 0$ be fixed. Let $F, G : \mathbb{H} \to \mathbb{C}$ be holomorphic functions such that, for some $M > 0$:*

$$\int_{\mathbb{R}} |F(x+iy)|^2 dx < M \quad \text{and} \quad \int_{\mathbb{R}} |G(x+iy)|^2 dx < M \qquad \forall y > 0 \tag{3.8}$$

We suppose that:

$$|F(x+i\alpha)| = |G(x+i\alpha)| \text{ for a.e. } x \in \mathbb{R}$$
$$\lim_{y\to 0^+} |F(x+iy)| = \lim_{y\to 0^+} |G(x+iy)| \text{ for a.e. } x \in \mathbb{R}$$

Then, for some $\phi \in \mathbb{R}$:

$$F = e^{i\phi} G \tag{3.16}$$

Proof of Theorem 3.1. This proof relies on the ideas used by Akutowicz [1956].

If $F = 0$, the theorem is true: G is null over a whole line and, as G is holomorphic, $G = 0$. The same reasoning holds if $G = 0$. We now assume $F \neq 0, G \neq 0$.

The central point of the proof is to factorize the functions $F, F(.+i\alpha), G, G(.+i\alpha)$ as in the following lemma.

Lemma 3.8. [Kryloff, 1939][2] *The function F admits the following factorization:*

$$F(z) = e^{ic+i\beta z} B(z) D(z) S(z)$$

Here, c and β are real numbers. The function B is a Blaschke product. It is formed with the zeros of F in the upper half-plane $\mathbb{H}$. We call (z_k) these zeros, counted with multiplicity, with the exception of i. We call m the multiplicity of i as zero.

$$B(z) = \left(\frac{z-i}{z+i}\right)^m \prod_k \frac{|z_k - i|}{z_k - i} \frac{|z_k + i|}{z_k + i} \frac{z - z_k}{z - \overline{z}_k} \tag{3.17}$$

This product converges over $\mathbb{H}$, which is equivalent to:

$$\sum_k \frac{\operatorname{Im} z_k}{1 + |z_k|^2} < +\infty \tag{3.18}$$

[2] Non Russian speaking readers may also deduce this theorem from Rudin [1987, Thm 17.17]: functions over $\mathbb{H}$ may be turned into functions over $D(0,1)$ by composing them with the conformal application $z \in D(0,1) \to \frac{1-z}{1+z} i \in \mathbb{H}$. The main difficulty is to show that if $H : \mathbb{H} \to \mathbb{C}$ satisfies (3.8), then $\tilde{H} : z \in D(0,1) \to H\left(\frac{1-z}{1+z} i\right) \in \mathbb{C}$ is of class H^2 and Rudin's theorem can be applied.

The functions D and S are defined by:

$$D(z) = \exp\left(\frac{1}{\pi i}\int_{\mathbb{R}} \frac{1+tz}{t-z}\frac{\log|F(t)|}{1+t^2}dt\right) \tag{3.19}$$

$$S(z) = \exp\left(\frac{i}{\pi}\int_{\mathbb{R}} \frac{1+tz}{t-z}dE(t)\right) \tag{3.20}$$

In the first equation, $|F(t)|$ is the limit of $|F|$ on $\mathbb{R}$. In the second one, dE is a positive bounded measure, singular with respect to Lebesgue measure.

Both integrals converge absolutely for any $z \in \mathbb{H}$.

The same factorization can be applied to $F(.+i\alpha), G$ and $G(.+i\alpha)$:

$$\begin{array}{cc} F(z) = e^{ic_F+i\beta_F z}B_F(z)D_F(z)S_F(z) & G(z) = e^{ic_G+i\beta_G z}B_G(z)D_G(z)S_G(z) \\ F(z+i\alpha) = e^{i\tilde{c}_F+i\tilde{\beta}_F z}\tilde{B}_F(z)\tilde{D}_F(z)\tilde{S}_F(z) & G(z+i\alpha) = e^{i\tilde{c}_G+i\tilde{\beta}_G z}\tilde{B}_G(z)\tilde{D}_G(z)\tilde{S}_G(z) \end{array}$$

As $F(.+i\alpha)$ and $G(.+i\alpha)$ are analytic on the real line, they actually have no singular part S. The proof may be found in [Garnett, 1981, Thm 6.3]; it is done for functions on the unit disk but also holds for functions on $\mathbb{H}$.

$$\tilde{S}_F = \tilde{S}_G = 1 \tag{3.21}$$

Because $\lim\limits_{y\to 0^+}|F(.+iy)| = \lim\limits_{y\to 0^+}|G(.+iy)|$ and $|F(.+i\alpha)| = |G(.+i\alpha)|$, we have $D_F = D_G$ and $\tilde{D}_F = \tilde{D}_G$. We show that it implies a relation between the B's, that is, a relation between the zeros of F and G. From this relation, we will be able to prove that F and G have the same zeros and that, up to a global phase, they are equal.

For all $z \in \mathbb{H}$:

$$\begin{aligned} &\frac{e^{ic_F+i\beta_F(z+i\alpha)}B_F(z+i\alpha)D_F(z+i\alpha)S_F(z+i\alpha)}{e^{i\tilde{c}_F+i\tilde{\beta}_F z}\tilde{B}_F(z)\tilde{D}_F(z)} \\ &\qquad = \frac{F(z+i\alpha)}{F(z+i\alpha)} = 1 \\ &\qquad = \frac{G(z+i\alpha)}{G(z+i\alpha)} \\ &\qquad = \frac{e^{ic_G+i\beta_G(z+i\alpha)}B_G(z+i\alpha)D_G(z+i\alpha)S_G(z+i\alpha)}{e^{i\tilde{c}_G+i\tilde{\beta}_G z}\tilde{B}_G(z)\tilde{D}_G(z)} \end{aligned}$$

$$\Rightarrow \quad \frac{B_F(z+i\alpha)\tilde{B}_G(z)}{B_G(z+i\alpha)\tilde{B}_F(z)} = e^{iC+iBz}\frac{S_G(z+i\alpha)}{S_F(z+i\alpha)} \tag{3.22}$$

for some $C, B \in \mathbb{R}$

Equality (3.22) holds only for $z \in \mathbb{H}$. It is a priori not even defined for $z \in \mathbb{C} - \mathbb{H}$. Before going on, we must show that (3.22) is meaningful and still valid over all $\mathbb{C}$. This is the purpose of the two following lemmas, whose proofs may be found in Paragraph 3.6.1.

For $z \in \mathbb{H}$, we denote by $\mu_F(z)$ (resp. $\mu_G(z)$) the multiplicity of z as a zero of F (resp. G).

Lemma 3.9. *There exists a meromorphic function $B_w : \mathbb{C} \to \mathbb{C}$ such that:*

$$B_w(z) = \frac{B_F(z+i\alpha)\tilde{B}_G(z)}{B_G(z+i\alpha)\tilde{B}_F(z)} \qquad \forall z \in \mathbb{H}$$

Moreover, for every $z \in \mathbb{H}$, the multiplicity of $\overline{z} - i\alpha$ as a pole of B_w is:

$$(\mu_F(z) - \mu_G(z)) - (\mu_F(z+2i\alpha) - \mu_G(z+2i\alpha)) \tag{3.23}$$

Lemma 3.10. *For every $z \in \mathbb{H}$, $\frac{S_G(z+i\alpha)}{S_F(z+i\alpha)} = 1$.*

Equation (3.22) and Lemmas 3.9 and 3.10 give, for all $z \in \mathbb{H}$ and thus all $z \in \mathbb{C}$ (because functions are meromorphic):

$$B_w(z) = e^{iC+iBz} \qquad \forall z \in \mathbb{C}$$

The function e^{iC+iBz} has no zero nor pole so, from (3.23), for all $z \in \mathbb{H}$:

$$(\mu_F(z) - \mu_G(z)) - (\mu_F(z+2i\alpha) - \mu_G(z+2i\alpha)) = 0$$

So if $\mu_F(z) \neq \mu_G(z)$ for some z, we may by symmetry assume that $\mu_F(z) > \mu_G(z)$ and, in this case, for all $n \in \mathbb{N}^*$:

$$\begin{aligned}\mu_F(z+2ni\alpha) - \mu_G(z+2ni\alpha) &= \dots \\ &= \mu_F(z+2i\alpha) - \mu_G(z+2i\alpha) \\ &= \mu_F(z) - \mu_G(z) > 0\end{aligned}$$

In particular, $z + 2ni\alpha$ is a zero of F for all $n \in \mathbb{N}^*$. But this is impossible because, if it is the case, $\frac{\mathrm{Im}(z+2ni\alpha)}{1+|z+2ni\alpha|^2} \sim \frac{1}{2n\alpha}$ and:

$$\sum_k \frac{\mathrm{Im}\, z_k}{1+|z_k|^2} = +\infty$$

where the (z_k) are the zeros of F over $\mathbb{H}$. It is in contradiction with (3.18).

So for all $z \in \mathbb{H}$, $\mu_F(z) = \mu_G(z)$. This implies that $B_F = B_G$ and $\tilde{B}_F = \tilde{B}_G$. So, for all $z \in \mathbb{H}$:

$$F(z+i\alpha) = e^{i\tilde{c}_F + i\tilde{\beta}_F z}\tilde{B}_F(z)\tilde{D}_F(z) = e^{i\tilde{c}_F + i\tilde{\beta}_F z}\tilde{B}_G(z)\tilde{D}_G(z) = e^{i\gamma+i\delta z}G(z+i\alpha)$$

$$\text{with } \gamma = \tilde{c}_F - \tilde{c}_G \text{ and } \delta = \tilde{\beta}_F - \tilde{\beta}_G$$

The functions F and G are meromorphic over $\mathbb{H}$ so the last equality actually holds over all $\{z \in \mathbb{C} \text{ s.t. } \operatorname{Im} z > -\alpha\}$.

$$\begin{aligned} |\lim_{y\to 0^+} F(x+iy)| &= |\lim_{y\to 0^+} e^{i\gamma + i\delta(x+iy-i\alpha)} G(x+iy)| \\ &= e^{\delta\alpha} |\lim_{y\to 0^+} G(x+iy)| \end{aligned}$$

Consequently, because δ is real and $\alpha \neq 0$, $\delta = 0$. So:

$$F(z) = e^{i\gamma} G(z) \qquad \forall z \in \mathbb{H}$$

□

3.2 Weak stability of the reconstruction

In the previous section, we proved that the operator $U : f \to \{|f \star \psi_j|\}$ was injective, up to a global phase, for Cauchy wavelets. So we can theoretically reconstruct any function f from $U(f)$. However, if we want the reconstruction to be possible in practice, we also need it to be stable to a small amount of noise:

$$(U(f_1) \approx U(f_2)) \quad \Rightarrow \quad (f_1 \approx f_2)$$

In this section, we show that it is, in some sense, the case: U^{-1} is continuous.

Contrarily to the ones of the previous section, this result is not specific to Cauchy wavelets: it holds for all reasonable wavelets, as soon as U is injective.

3.2.1 Definitions

As in the previous section, we consider only functions without negative frequencies:

$$L^2_+(\mathbb{R}) = \{f \in L^2(\mathbb{R}) \text{ s.t. } \hat{f}(\omega) = 0 \text{ for a.e. } \omega < 0\}$$

As the reconstruction is always up to a global phase, we need to define the quotient $L^2_+(\mathbb{R})/S^1$:

$$f = g \text{ in } L^2_+(\mathbb{R})/S^1 \quad \Leftrightarrow \quad f = e^{i\phi} g \text{ for some } \phi \in \mathbb{R}$$

The set $L^2_+(\mathbb{R})/S^1$ is equipped with a natural metric:

$$D_2(f,g) = \inf_{\phi\in\mathbb{R}} ||f - e^{i\phi} g||_2$$

Remark that $D_2(f,0) = ||f||_2$.
We also define:

$$l^2(\mathbb{Z}, L^2(\mathbb{R})) = \left\{ (h_j)_{j\in\mathbb{Z}} \in L^2(\mathbb{R})^{\mathbb{Z}} \text{ s.t. } \sum_j ||h_j||_2^2 < +\infty \right\}$$

$$\left|\left|(h_j) - (h'_j)\right|\right|_2 = \sqrt{\sum_{j\in\mathbb{Z}} ||h_j - h'_j||_2^2} \quad \text{for any } (h_j), (h'_j) \in l^2(\mathbb{Z}, L^2(\mathbb{R}))$$

We are interested in the operator U:

$$\begin{array}{rccc} U: & L^2_+(\mathbb{R})/S^1 & \to & l^2(\mathbb{Z}, L^2(\mathbb{R})) \\ & f & \to & (|f \star \psi_j|)_{j\in\mathbb{Z}} \end{array} \tag{3.24}$$

We require two conditions over the wavelets. They must be analytic:

$$\hat{\psi}_j(\omega) = 0 \text{ for a.e. } \omega < 0, j \in \mathbb{Z} \tag{3.25}$$

and satisfy an approximate Littlewood-Paley inequality:

$$A \leq \sum_{j\in\mathbb{Z}} |\hat{\psi}_j(\omega)|^2 \leq B \qquad \text{for a.e. } \omega > 0, \text{for some } A, B > 0 \tag{3.26}$$

This last inequality and the fact that $D_2(f,0) = ||f||_2$ imply:

$$\forall f \in L^2_+(\mathbb{R})/S^1, \quad \sqrt{A} D_2(f,0) \leq ||U(f)||_2 \leq \sqrt{B} D_2(f,0) \tag{3.27}$$

In particular, it ensures the continuity of U.

3.2.2 Weak stability theorem

Theorem 3.11. *We suppose that, for all $j \in \mathbb{Z}$, $\psi_j \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ and that (3.25) and (3.26) hold. We also suppose that U is injective. Then:*
(i) The image of U, $I_U = \{U(f) \text{ s.t. } f \in L^2_+(\mathbb{R})/S^1\}$ is closed in $l^2(\mathbb{Z}, L^2(\mathbb{R}))$.
(ii) The application $U^{-1} : I_U \to L^2_+(\mathbb{R})/S^1$ is continuous.

Proof. What we have to prove is the following: if $(U(f_n))_{n\in\mathbb{N}}$ converges towards a limit $v \in l^2(\mathbb{Z}, L^2(\mathbb{R}))$, then $v = U(g)$ for some $g \in L^2_+(\mathbb{R})/S^1$ and $f_n \to g$ in $L^2_+(\mathbb{R})/S^1$.

So let $(U(f_n))_{n\in\mathbb{N}}$ be a sequence of elements in I_U, which converges in $l^2(\mathbb{Z}, L^2(\mathbb{R}))$. Let $v = (h_j)_{j\in\mathbb{Z}} \in L^2_{\mathbb{Z}}(\mathbb{R})$ be the limit. We show that $v \in I_U$.

Lemma 3.12. *For all $j \in \mathbb{Z}$, $\{f_n \star \psi_j\}_{n\in\mathbb{N}}$ is relatively compact in $L^2(\mathbb{R})$ (that is, the closure of this set in $L^2(\mathbb{R})$ is compact).*

The proof of this lemma is given in Paragraph 3.6.2. It uses the Riesz-Fréchet-Kolmogorov theorem, which gives an explicit characterization of the relatively compact subsets of $L^2(\mathbb{R})$.

For every $j \in \mathbb{Z}$, $\{f_n \star \psi_j\}_{n\in\mathbb{N}}$ is thus included in a compact subset of $L^2(\mathbb{R})$. In a compact set, every sequence admits a convergent subsequence: there exists $\phi : \mathbb{N} \to \mathbb{N}$ injective such that $(f_{\phi(n)} \star \psi_j)_{n\in\mathbb{N}}$ converges in $L^2(\mathbb{R})$. Actually, we can choose ϕ such that $(f_{\phi(n)} \star \psi_j)_n$ converges for any j (and not only for a single one). We denote by l_j the limits.

Lemma 3.13 (Proof in Paragraph 3.6.2)**.** *There exists $g \in L^2_+(\mathbb{R})$ such that $l_j = g \star \psi_j$ for every j. Moreover, $f_{\phi(n)} \to g$ in $L^2(\mathbb{R})$.*

As U is continuous, $U(g) = \lim_n U(f_{\phi(n)}) = v$. So v belongs to I_U.

The g such that $U(g) = v$ is uniquely defined in $L^2_+(\mathbb{R})/S^1$ because U is injective (it does not depend on the choice of ϕ). We must now show that $f_n \to g$.

From Lemma 3.13, $(f_n)_n$ admits a subsequence $(f_{\phi(n)})$ which converges to g. By the same reasoning, every subsequence $(f_{\psi(n)})_n$ of $(f_n)_n$ admits a subsequence which converges to g. This implies that $(f_n)_n$ globally converges to g.

□

Remark 3.14. *The same proof gives a similar result for wavelets on $\mathbb{R}^d$, of the form $(\psi_{j,\gamma})_{j\in\mathbb{Z},\gamma\in\Gamma}$, for Γ a finite set of parameters.*

3.3 The reconstruction is not uniformly continuous

Theorem 3.11 states that the operator $U : f \to \{|f \star \psi_j|\}_{j\in\mathbb{Z}}$ has a continuous inverse U^{-1}, when it is invertible. However, U^{-1} is not uniformly continuous. Indeed, for any $\epsilon > 0$, there exist $g_1, g_2 \in L^2_+(\mathbb{R})/S^1$ such that:

$$||U(g_1) - U(g_2)|| < \epsilon \quad \text{but} \quad ||g_1 - g_2|| \geq 1 \tag{3.28}$$

In this section, we describe a way to construct such "unstable" pairs (g_1, g_2): we start from any g_1 and modulate each $g_1 \star \psi_j$ by a low-frequency phase. We then (approximately) invert this modified wavelet transform and obtain g_2.

This construction seems to be "generic" in the sense that it includes all the instabilities that we have been able to observe in practice.

3.3.1 A simple example

To begin with, we give a simple example of instabilities and relate it to known results about the stability in general phase retrieval problems.

In phase retrieval problems with (a finite number of) real measurements, the stability of the reconstruction operator is characterized by the following theorem ([Bandeira et al., 2014], [Balan and Wang, 2015]).

Theorem 3.15. *Let $A \in \mathbb{R}^{m\times n}$ be a measurement matrix. For any $S \subset \{1, ..., m\}$, we denote by A_S the matrix obtained by discarding the rows of A whose indexes are not in S. We call λ_S^2 the lower frame bound of A_S, that is, the largest real number such that:*

$$||A_S x||_2^2 \geq \lambda_S^2 ||x||_2^2 \qquad \forall x \in \mathbb{R}^n$$

Then, for any $x, y \in \mathbb{R}^n$:

$$|| \, |Ax| - |Ay| \, ||_2 \geq \left(\min_S \sqrt{\lambda_S^2 + \lambda_{S^c}^2} \right) . \min(||x - y||_2, ||x + y||_2)$$

Moreover, $\min_S \sqrt{\lambda_S^2 + \lambda_{S^c}^2}$ is the optimal constant.

This theorem implies that, in the real case, the reconstruction operator has a Lipschitz constant exactly equal to $1/\left(\min_S \sqrt{\lambda_S^2 + \lambda_{S^c}^2}\right)$. In the complex case, it is only possible to prove that the Lipschitz constant is at least $1/\left(\min_S \sqrt{\lambda_S^2 + \lambda_{S^c}^2}\right)$.

Theorem 3.16. *Let $A \in \mathbb{C}^{m\times n}$ be a measurement matrix. There exist $x, y \in \mathbb{C}^n$ such that:*

$$|| \, |Ax| - |Ay| \, ||_2 \leq \left(\min_S \sqrt{\lambda_S^2 + \lambda_{S^c}^2} \right) . \min_{|\eta|=1} (||x - \eta y||_2)$$

Consequently, if the set of measurements can be divided in two parts S and S^c such that λ_S^2 and $\lambda_{S^c}^2$ are very small, then the reconstruction is not stable.

Such a phenomenon occurs in the case of the wavelet transform. We define:

$$S = \{\psi_j \text{ s.t. } j \geq 0\} \text{ and } S^c = \{\psi_j \text{ s.t. } j < 0\}$$

Let us fix a small $\epsilon > 0$. We choose $f_1, f_2 \in L^2(\mathbb{R})$ such that:

$$\hat{f}_1(x) = 0 \text{ if } |x| < 1/\epsilon \quad \text{ and } \quad \hat{f}_2(x) = 0 \text{ if } x \notin [-\epsilon; \epsilon]$$

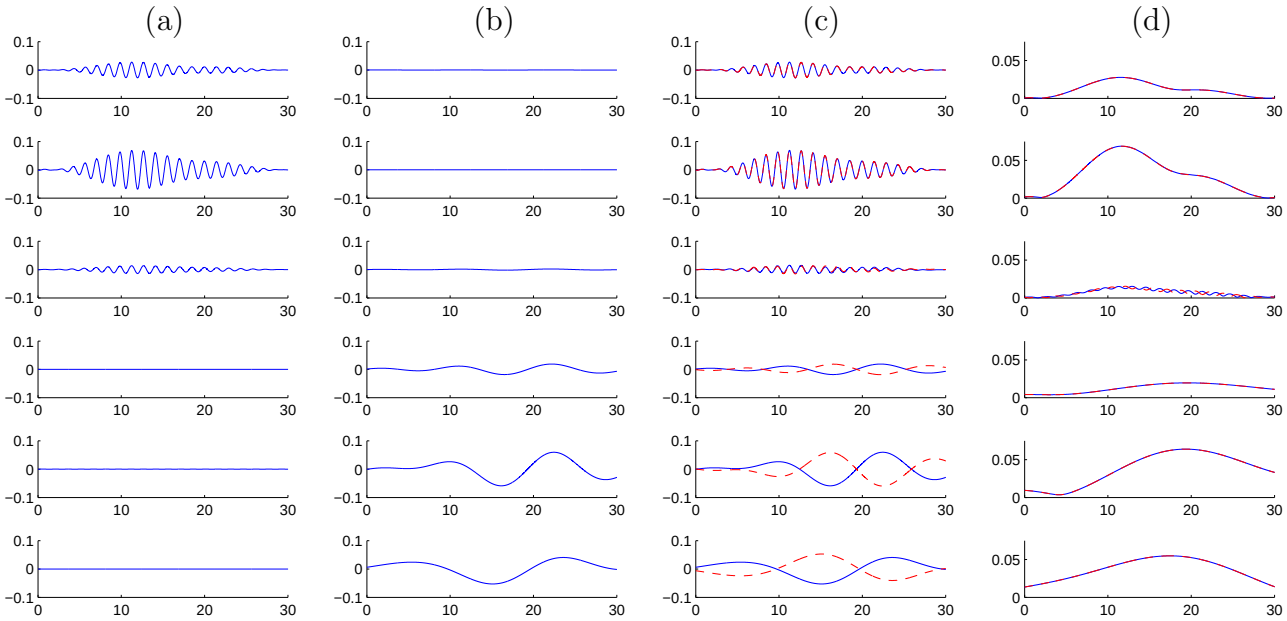


Figure 3.3: (a) Wavelet transform of f_1 (b) Wavelet transform of f_2 (c) Wavelet transform of $f_1 + f_2$ (solid blue) and $f_1 - f_2$ (dashed red) (d) Modulus of the wavelet transforms of $f_1 + f_2$ and $f_1 - f_2$; the two modulus are almost equal
In each column, each graph corresponds to a specific frequency; the highest frequency is on top and the lowest one at bottom. For complex functions, only the real part is displayed.

For every $\psi_j \in S$, $f_1 \star \psi_j \approx 0$ because the characteristic frequency of ψ_j is smaller than 1 and f_1 is a very high frequency function. So:

$$|(f_1 + f_2) \star \psi_j| \approx |f_2 \star \psi_j| = |-f_2 \star \psi_j| \approx |(f_1 - f_2) \star \psi_j|$$

And similarly, for $\psi_j \in S^c$, $f_2 \star \psi_j \approx 0$ and:

$$|(f_1 + f_2) \star \psi_j| \approx |f_1 \star \psi_j| \approx |(f_1 - f_2) \star \psi_j|$$

As a consequence:

$$\{|(f_1 + f_2) \star \psi_j|\}_{j \in \mathbb{Z}} \approx \{|(f_1 - f_2) \star \psi_j|\}_{j \in \mathbb{Z}}$$

Nevertheless, $f_1 + f_2$ and $f_1 - f_2$ may not be close in $L^2(\mathbb{R})/S^1$: $g_1 = f_1 + f_2$ and $g_2 = f_1 - f_2$ satisfy (3.28).

The figure 3.3 displays an example of this kind.

3.3.2 A wider class of instabilities

We now describe the construction of more general "unstable" pairs (g_1, g_2).
Let $g_1 \in L^2(\mathbb{R})$ be any function. We aim at finding $g_2 \in L^2(\mathbb{R})$ such that, for all $j \in \mathbb{Z}$:

$$(g_1 \star \psi_j)e^{i\phi_j} \approx g_2 \star \psi_j \tag{3.29}$$

for some real functions ϕ_j.

In other words, we must find phases ϕ_j such that $(g_1 \star \psi_j)e^{i\phi_j}$ is approximately equal to the wavelet transform of some $g_2 \in L^2(\mathbb{R})$. Any phases $\phi_j(t)$ which vary slowly both in t and in j satisfy this property.

Indeed, if the $\phi_j(t)$ vary "slowly enough", we set:

$$g_2 = \sum_{j\in\mathbb{Z}} \left((g_1 \star \psi_j)e^{i\phi_j}\right) \star \tilde{\psi}_j$$

where $\{\tilde{\psi}_j\}_{j\in\mathbb{Z}}$ are the dual wavelets associated to $\{\psi_j\}$.

Then, for all $k \in \mathbb{Z}, t \in \mathbb{R}$:

$$\begin{aligned} g_2 \star \psi_k(t) &= \sum_{j\in\mathbb{Z}} \left((g_1 \star \psi_j)e^{i\phi_j}\right) \star \tilde{\psi}_j \star \psi_k(t) \\ &= \sum_{j\in\mathbb{Z}} \int_{\mathbb{R}} e^{i\phi_j(t-u)}(g_1 \star \psi_j)(t-u)(\tilde{\psi}_j \star \psi_k)(u)\, du \end{aligned}$$

$$\begin{aligned} (g_1 \star \psi_k(t))e^{i\phi_k(t)} &= e^{i\phi_k(t)} \sum_{j\in\mathbb{Z}} (g_1 \star \psi_j) \star (\tilde{\psi}_j \star \psi_k)(t) \\ &= \sum_{j\in\mathbb{Z}} \int_{\mathbb{R}} e^{i\phi_k(t)}(g_1 \star \psi_j)(t-u)(\tilde{\psi}_j \star \psi_k)(u)\, du \end{aligned}$$

So:

$$g_2 \star \psi_k(t) - (g_1 \star \psi_k(t))e^{i\phi_k(t)} = \sum_{j\in\mathbb{Z}} \int_{\mathbb{R}} (e^{i\phi_j(t-u)} - e^{i\phi_k(t)})(g_1 \star \psi_j)(t-u)(\tilde{\psi}_j \star \psi_k)(u)\, du \tag{3.30}$$

The function $\tilde{\psi}_j \star \psi_k(u)$ is negligible if j is not of the same order as k or if u is too far away from 0. It means that, for some $C \in \mathbb{N}, U \in \mathbb{R}$ (which may depend on k):

$$g_2 \star \psi_k(t) - (g_1 \star \psi_k(t))e^{i\phi_k(t)} \approx \sum_{|j-k|\leq C} \int_{[-U;U]} (e^{i\phi_j(t-u)} - e^{i\phi_k(t)})(g_1 \star \psi_j)(t-u)(\tilde{\psi}_j \star \psi_k)(u)\, du$$

If $\phi_j(t-u)$ does not vary much over $[k-C;k+C]\times[-U;U]$, it gives the desired relation:

$$g_2 \star \psi_k(t) - (g_1 \star \psi_k(t))e^{i\phi_k(t)} \approx 0$$

which is (3.29).

To summarize, we have described a way to construct $g_1, g_2 \in L^2(\mathbb{R})$ such that $|g_1 \star \psi_j| \approx |g_2 \star \psi_j|$ for all j. The principle is to multiply the wavelet transform of g_1 by any set of phases $\{e^{i\phi_j(t)}\}_{j\in\mathbb{Z}}$ whose variations are slow enough in j and t.

How slow the variations must be depends on g_1. Indeed, at the points (j,t) where $g_1 \star \psi_j(t)$ is small, the phase may vary more rapidly because, then, the presence of $g_1 \star \psi_j(t-u)$ in (3.30) compensates for a bigger $(e^{i\phi_j(t-u)} - e^{i\phi_k(t)})$.

All instabilities g_1, g_2 that we were able to observe in practice were of the form we described: each time, the wavelet transforms of g_1 and g_2 were equal up to a phase whose variation was slow in j and t, except at the points where $g_1 \star \psi_j$ was small.

3.4 Local stability result

The goal of this section is to give a partial formal justification to the fact that has been non-rigorously discussed in section 3.3.2: when two functions g_1, g_2 satisfy $|g_1 \star \psi_j| \approx |g_2 \star \psi_j|$ for all j, then the wavelet transforms $\{g_1 \star \psi_j(t)\}_j$ and $\{g_2 \star \psi_j(t)\}_j$ are equal up to a phase whose variation is slow in t and j, except eventually at the points where $|g_1 \star \psi_j(t)|$ is small.

In the whole section, we consider $f^{(1)}, f^{(2)}$ two non-zero functions. We denote by $F^{(1)}, F^{(2)}$ the holomorphic extensions defined in (3.6). We recall that, for all $j \in \mathbb{Z}$:

$$f \star \psi_j(x) = \frac{a^{pj}}{2} F(x + ia^j) \qquad \forall x \in \mathbb{R} \tag{3.31}$$

We define:

$$N_j = \sup_{x\in\mathbb{R}, s=1,2} |f^{(s)} \star \psi_j(x)|$$

3.4.1 Main principle

From $|f\star\psi_j|$, one can calculate $|f\star\psi_j|^2$ and thus, from (3.31), $|F(x+ia^j)|^2$, for all $x \in \mathbb{R}$. But this last function coincides with $G_j(z) = F(z+ia^j)\overline{F(\overline{z}+ia^j)}$ on the horizontal line $\operatorname{Im} z = 0$. As G_j is holomorphic, it is uniquely determined by its values on one line. Consequently, G_j is uniquely determined from $|f \star \psi_j|$.

Combining the functions G_j for different values of j allows to write explicit reconstruction formulas. The stability of these formulas can be studied, to obtain relations of the following

form, for $K > 0$:

$$\Big(|f^{(1)} \star \psi_k| \approx |f^{(2)} \star \psi_k| \quad \forall k \in \mathbb{Z}\Big)$$
$$\Rightarrow \Big((f^{(1)} \star \psi_j)(\overline{f^{(1)} \star \psi_{j+K}}) \approx (f^{(2)} \star \psi_j)(\overline{f^{(2)} \star \psi_{j+K}}) \quad \forall j \in \mathbb{Z}\Big)$$

These relations imply that, for each j, the phases of $f^{(1)} \star \psi_j$ and $f^{(2)} \star \psi_j$ are approximately equal up to multiplication by the phase of $\frac{\overline{f^{(1)} \star \psi_{j+K}}}{\overline{f^{(2)} \star \psi_{j+K}}}$. If K is not too small, this last phase is low-frequency, compared to the phase of $f^{(1)} \star \psi_j$ and $f^{(2)} \star \psi_j$.

The results we obtain are local, in the sense that if the approximate equality $|f^{(1)} \star \psi_k| \approx |f^{(2)} \star \psi_k|$ only holds on a (large enough) interval of $\mathbb{R}$, the equality $(f^{(1)} \star \psi_j)(\overline{f^{(1)} \star \psi_{j+K}}) \approx (f^{(2)} \star \psi_j)(\overline{f^{(2)} \star \psi_{j+K}})$ still holds (also on an interval of $\mathbb{R}$).

Our main technical difficulty was to handle properly the fact that the G_j's may have zeros (which is a problem because we need to divide by G_j in order to get reconstruction formulas). We know that, when the wavelet transform has a lot of zeros, the reconstruction becomes unstable. On the other hand, if they are only a few isolated zeros, the reconstruction is stable and this must appear in our theorems.

They are several ways to write reconstruction formulas, which give different stability results. In the dyadic case ($a = 2$), there is a relatively simple method. We present it first. Then we handle the case where $a < 2$. We do not consider the case where $a > 2$. Indeed, it has less practical interest for us. Moreover, when the value of a increases, the reconstruction becomes much less stable.

3.4.2 Case $a = 2$

In the dyadic case, we only assume that two consecutive moduli are approximately known, on an interval of $\mathbb{R}$: $|f \star \psi_j|$ and $|f \star \psi_{j+1}|$. We also assume that, on this interval, the moduli are never too close to 0. Then we show these moduli determine in a stable way:

$$\frac{f \star \psi_{j+2}}{f \star \psi_{j+1}}$$

Theorem 3.17. *Let $\epsilon, c, \lambda \in]0;1[$, $M > 0$ be fixed, with $c \geq \epsilon$.*

We assume that, for all $x \in [-M2^j; M2^j]$:

$$\big||f^{(1)} \star \psi_j(x)|^2 - |f^{(2)} \star \psi_j(x)|^2\big| \leq \epsilon N_j^2$$
$$\big||f^{(1)} \star \psi_{j+1}(x)|^2 - |f^{(2)} \star \psi_{j+1}(x)|^2\big| \leq \epsilon N_{j+1}^2$$

and:

$$|f^{(1)} \star \psi_j(x)|^2, |f^{(2)} \star \psi_j(x)|^2 \geq cN_j^2$$
$$|f^{(1)} \star \psi_{j+1}(x)|^2, |f^{(2)} \star \psi_{j+1}(x)|^2 \geq cN_{j+1}^2$$

Then, for all $x \in [-\lambda^2 M 2^j; \lambda^2 M 2^j]$*:*

$$\left| \frac{f^{(1)} \star \psi_{j+2}}{f^{(1)} \star \psi_{j+1}}(x) - \frac{f^{(2)} \star \psi_{j+2}}{f^{(2)} \star \psi_{j+1}}(x) \right| \leq \frac{A}{c} \left(\frac{N_{j-1}}{N_{j+1}} \right)^{4/3} \epsilon^{(1/3-\alpha_M)(4/5-\alpha'_M)}$$

if $1/3 - \alpha_M > 0$ *and* $4/5 - \alpha'_M > 0$*, where:*

- A *is a constant which depends only on* p*.*
- $\alpha_M, \alpha'_M \to 0$ *exponentially when* $M \to +\infty$*.*

Principle of the proof. Here, we only give a broad outline of the proof. A rigorous one is given in Paragraph 3.6.3, with all the necessary technical details.

As explained in the paragraph 3.4.1, $|f^{(1)} \star \psi_{j+1}|$ uniquely determines the values of $z \to F^{(1)}(z + i2^{j+1})\overline{F^{(1)}(\overline{z} + i2^{j+1})}$ on the line $\operatorname{Im} z = 0$. Thus, it uniquely determines all the values (because the function is holomorphic) and in particular (for $z = x + i2^j$):

$$F^{(1)}(x + i3.2^j)\overline{F^{(1)}(x + i2^j)} \quad \forall x \in \mathbb{R}$$

Moreover, this determination is a stable operation:

$$\left(|f^{(1)} \star \psi_{j+1}(x)|^2 \approx |f^{(2)} \star \psi_{j+1}(x)|^2 \quad \forall x \in \mathbb{R} \right)$$
$$\Rightarrow \left(F^{(1)}(x + i3.2^j)\overline{F^{(1)}(x + i2^j)} \approx F^{(2)}(x + i3.2^j)\overline{F^{(2)}(x + i2^j)} \quad \forall x \in \mathbb{R} \right)$$

If we divide this last expression by $|F^{(1)}(x+i2^j)|^2 \approx |F^{(2)}(x+i2^j)|^2$ (whose values we know from $|f \star \psi_j|^2$):

$$\frac{F^{(1)}(x + i3.2^j)}{F^{(1)}(x + i2^j)} \approx \frac{F^{(2)}(x + i3.2^j)}{F^{(2)}(x + i2^j)} \quad \text{for } x \in \mathbb{R}$$

As previously, using the holomorphy of F allows to replace, in the last expression, the real number x by $x + i2^j$:

$$\frac{F^{(1)}(x + i2^{j+2})}{F^{(1)}(x + i2^{j+1})} \approx \frac{F^{(2)}(x + i2^{j+2})}{F^{(2)}(x + i2^{j+1})} \quad \text{for } x \in \mathbb{R}$$

By (3.31), this is the same as:

$$\frac{f^{(1)} \star \psi_{j+2}}{f^{(1)} \star \psi_{j+1}} \approx \frac{f^{(2)} \star \psi_{j+2}}{f^{(2)} \star \psi_{j+1}}$$

□

From this theorem, if $f^{(s)} \star \psi_{j+2}$ has no small values either on $[-\lambda^2 M 2^j; \lambda^2 M 2^j]$, then:

$$\text{phase}(f^{(1)} \star \psi_{j+1}) - \text{phase}(f^{(2)} \star \psi_{j+1}) \approx \text{phase}(f^{(1)} \star \psi_{j+2}) - \text{phase}(f^{(2)} \star \psi_{j+2})$$

If more than two consecutive components of the wavelet transform have almost the same modulus (and all these components do not come close to 0), one can iterate this approximate equality. It gives:

$$\text{phase}(f^{(1)} \star \psi_{j+1}) - \text{phase}(f^{(2)} \star \psi_{j+1}) \approx \text{phase}(f^{(1)} \star \psi_{j+K}) - \text{phase}(f^{(2)} \star \psi_{j+K})$$

This holds for any $K \in \mathbb{N}^*$ but with an approximation error that becomes larger and larger as K increases.

When K is large enough, this means that $f^{(1)} \star \psi_{j+1}$ and $f^{(2)} \star \psi_{j+1}$ are equal up to a low-frequency phase.

3.4.3 Case $a < 2$

For this section, we fix:

- $j \in \mathbb{Z}$: the frequency of the component whose phase we want to estimate
- $K \in \mathbb{N}^*$ even: the number of components of the wavelet transform whose modulus are approximately equal
- $\epsilon, \kappa \in]0;1[$: they will control the difference between $|f^{(1)} \star \psi_j|$ and $|f^{(2)} \star \psi_j|$, as well as the minimal value of those functions.
- $M > 0$: we will assume that the approximate equality between the modulus holds on $[-Ma^{j+K}; Ma^{j+K}]$.
- $k \in \mathbb{N}^*$ such that $a^{-k} < 2 - a$: this number will control the stability with which one can derive information about $f \star \psi_{l-1}$ from $|f \star \psi_l|$. Typically, for $a \leq 1.5$, we may take $k = 3$.

We define:

- $J \in [j + K - 1; j + K]$ such that $a^J = \frac{2}{a+1} a^{j+K} + \frac{a-1}{a+1} a^j$: we will prove that $f^{(1)} \star \psi_j$ and $f^{(2)} \star \psi_j$ are equal up to a phase which is concentrated around a^J in frequencies (that is, a much lower-frequency phase than the phase of $f \star \psi_j$).
- $c = 1 - \frac{a-1}{1-a^{-k}} \in]0;1[$ and $d_M = c - 4\frac{e^{-\pi M/(K+2)}}{1-e^{-\pi M/(K+2)}}$, which converges exponentially to c when $\frac{M}{K}$ goes to ∞.

Theorem 3.18. *We assume that* $\kappa \geq \epsilon^{2(1-c)}$.

We assume that, for $x \in [-Ma^{j+K}; Ma^{j+K}]$ *and* $l = j+1, ..., j+K$*:*

$$\left| |f^{(1)} \star \psi_l(x)|^2 - |f^{(2)} \star \psi_l(x)|^2 \right| \leq \epsilon N_l^2 \tag{3.32}$$

$$|f^{(1)} \star \psi_l(x)|^2, |f^{(2)} \star \psi_l(x)|^2 \geq \kappa N_l^2 \tag{3.33}$$

Then, for any $x \in \left[-\frac{Ma^{j+K}}{2}; \frac{Ma^{j+K}}{2}\right]$*, as soon as* $d_M < 1$*:*

$$\frac{1}{N_J N_j} \left| \left(\overline{f^{(1)} \star \psi_J(x)}\right) \left(f^{(1)} \star \psi_j(x)\right) - \left(\overline{f^{(2)} \star \psi_J(x)}\right) \left(f^{(2)} \star \psi_j(x)\right) \right| \leq \frac{C_K}{\kappa^{K/4}} \epsilon^{d_M} \tag{3.34}$$

where $C_K = \frac{6}{1-\sqrt{\kappa}} \prod_{s=0}^{K/2-1} \left(a^{p(k-1)} \frac{N_{n_s-1-k}}{N_{n_s-2}}\right)$

As in the dyadic case $a = 2$, this theorem shows that, if two functions $f^{(1)}$ and $f^{(2)}$ have their wavelet transforms almost equal in moduli, then, for each j, $f^{(1)} \star \psi_j \approx f^{(2)} \star \psi_j$ up to multiplication by a low-frequency function.

In contrast to the dyadic case, we are not able to show directly that:

$$\frac{f^{(1)} \star \psi_j}{f^{(2)} \star \psi_j} \approx \frac{f^{(1)} \star \psi_{j+1}}{f^{(2)} \star \psi_{j+1}}$$

Because of that, the inequality we get is less good than in the dyadic case: the bound in (3.34) is exponential in K instead of being proportional to K.

With a slightly different method, we could have obtained a better bound, proportional to K. This better bound would have been valid for any $a > 1$, but under the condition that $f \star \psi_l$ does not come close to 0 for some explicit non-integer values of l, which would have been rather unsatisfactory because, in practice, these values of l do not seem to play a particular role.

Principle of the proof. The full proof may be found in Paragraph 3.6.4. Its principle is to show, by induction over $s = 0, ..., K/2$, that:

$$(\overline{f^{(1)} \star \psi_{J_s}})(f^{(1)} \star \psi_{j+K-2s}) \approx (\overline{f^{(2)} \star \psi_{J_s}})(f^{(2)} \star \psi_{j+K-2s}) \tag{3.35}$$

where J_s is an explicit number in the interval $[j+K-1; j+K]$.

For $s = 0$, we set $J_s = j+K$ and (3.35) just says:

$$\left|f^{(1)} \star \psi_{j+K}\right|^2 \approx \left|f^{(2)} \star \psi_{j+K}\right|^2$$

which is true by hypothesis.

Then, to go from s to $s+1$, we use the fact that:

$$(\overline{f^{(1)} \star \psi_{j+K-2s}})(f^{(1)} \star \psi_l) \approx (\overline{f^{(2)} \star \psi_{j+K-2s}})(f^{(2)} \star \psi_l) \tag{3.36}$$

if we choose l such that $a^l = 2a^{j+K-2s-1} - a^{j+K-2s}$: we can check that, up to multiplication by a constant, $(\overline{f^{(r)} \star \psi_{j+K-2s}})(f^{(r)} \star \psi_l)$ is the evaluation on the line $a^{j+K-2s} - a^{j+K-2s-1}$ of the holomorphic extension of $|f^{(r)} \star \psi_{j+K-2s-1}|^2$. The holomorphic extension is a stable transformation (in a sense that has to be made precise). As $|f^{(1)} \star \psi_{j+K-2s-1}|^2 \approx |f^{(2)} \star \psi_{j+K-2s-1}|^2$, this implies (3.36).

Multiplying (3.35) and (3.36) and dividing by $|f^{(1)} \star \psi_{j+K-2s}|^2 \approx |f^{(2)} \star \psi_{j+K-2s}|^2$ yields:

$$(\overline{f^{(1)} \star \psi_{J_s}})(f^{(1)} \star \psi_l) \approx (\overline{f^{(2)} \star \psi_{J_s}})(f^{(2)} \star \psi_l) \tag{3.37}$$

If J_{s+1} is suitably chosen, $(f^{(r)} \star \psi_{J_{s+1}})(f^{(r)} \star \psi_{j+K-2(s+1)})$ may be seen as the restriction to a line of the holomorphic extension of $(\overline{f^{(r)} \star \psi_{J_s}})(f^{(r)} \star \psi_l)$. Because, again, taking the holomorphic extension is relatively stable, the relation (3.37) implies the recurrence hypothesis (3.35) at order $s+1$.

For $s = K/2$, the recurrence hypothesis is equivalent to the stated result. □

3.5 Numerical experiments

In the previous section, we proved a form of stability for the phase retrieval problem associated to the Cauchy wavelet transform. The proof implicitly relied on the existence of an explicit reconstruction algorithm. In this section, we describe a practical implementation of this algorithm and its performances.

The main goal of our numerical experiments is to investigate the issue of stability. Theorems 3.17 and 3.18 prove that the reconstruction is, in some sense, stable, at least when the wavelet transform does not have small values. Are these results confirmed by the implementation? To what extent does the presence of small values make the reconstruction unstable?

As we will see, our algorithm can fail when large parts of the wavelet transform are close to zero. In all other cases, it seems to succeed and to be stable to noise, even when the amount of noise over the wavelet transform is relatively high ($\sim 10\%$). The presence of a small number of zeroes in the wavelet transform is not a problem.

In practical applications, the wavelet transforms of the signals of interest (mostly audio signals) always have a lot of small values. The algorithm that we present is thus mostly a theoretical tool. Without modifications, it is not intended for real applications. Nevertheless, the results it gives for audio signals are better than expected so, with some more work, it could be suited to practical applications in audio processing. This will be the subject of future work.

The code is available at http://www.di.ens.fr/~waldspurger/cauchy_phase_retrieval.html, along with examples of reconstruction for audio signals. It only handles the dyadic case $a = 2$ but could easily be extended to other values of a.

3.5.1 Description of the algorithm

In practice, we must restrict our wavelet transform to a finite number of components. So we only consider the $|f \star \psi_j|$ for $j \in \{J_{\min}, ..., J_{\max}\}$. To compensate for the loss of the $|f \star \psi_j|$ with $j > J_{\max}$, we give to our algorithm an additional information about the low-frequency, under the form of $f \star \phi_{J_{\max}}$, where $\hat{\phi}_{J_{\max}}$ is negligible outside an neighborhood of 0 of size $\sim a^{-J_{\max}}$.

The algorithm takes as input the functions $|f \star \psi_{J_{\min}}|, |f \star \psi_{J_{\min}+1}|, ..., |f \star \psi_{J_{\max}}|, f \star \phi_{J_{\max}}$, for some unknown f, and tries to reconstruct f. The input functions may be contaminated by some noise. To simplify the implementation, we have assumed that the probability distribution of the noise was known.

For any real numbers j, k_1, k_2 such that $j \in \mathbb{Z}$ and $2.a^j = a^{k_1} + a^{k_2}$, it comes from the reasoning of the previous section that $|f \star \psi_j|$ uniquely determines $(f \star \psi_{k_1}).(\overline{f \star \psi_{k_2}})$. More precisely, we have, for all $\omega \in \mathbb{R}$:

$$\widehat{(f \star \psi_{k_1}).(\overline{f \star \psi_{k_2}})}(\omega) = |\widehat{f \star \psi_j}|^2(\omega) e^{(a^{k_2} - a^j)\omega} \frac{a^{k_1+k_2}}{a^{2j}} \tag{3.38}$$

The algorithm begins by fixing real numbers $k_{J_{\min}-1}, k_{J_{\min}}, ..., k_{J_{\max}}$ such that:

$$\begin{gathered} k_{J_{\min}-1} < J_{\min} < k_{J_{\min}} < J_{\min} + 1 < ... < J_{\max} < k_{J_{\max}} \\ \forall j, \qquad 2.a^j = a^{k_{j-1}} + a^{k_j} \end{gathered} \tag{3.39}$$

Then, for all j, it applies (3.38) to determine $g_j \overset{\text{def}}{=} (f \star \psi_{k_{j-1}}).(\overline{f \star \psi_{k_j}})$. Because of the exponential function present in (3.38), the g_j may take arbitrarily high values in the frequency band $\{(a^{k_2} - a^j)\omega \gg 1\}$. To avoid this, we truncate the high frequencies of g_j.

The function $f \star \psi_{k_{J_{\max}}}$ may be approximately determined from $f \star \phi_{J_{\max}}$. From this function and the g_j, the algorithm estimates all the $f \star \psi_{k_j}$. As this estimation involves divisions by functions which may be close to zero at some points, it is usually not very accurate. In particular, the estimated set $\{f \star \psi_{k_j}\}_j$ do not generally satisfy the constraint that it must belong to the range of the function $f \in L^2(\mathbb{R}) \to \{f \star \psi_{k_j}\}_{J_{\min}-1 \leq j \leq J_{\max}}$.

Thus, in a second step, the algorithm refines the estimation. To do this, it attempts to minimize an error function which takes into account both the fact that $(f \star \psi_{k_{j-1}}).(\overline{f \star \psi_{k_j}})$ is known for every j and the fact that $\{f \star \psi_{k_{j-1}}\}_{J_{\min}-1 \leq j \leq J_{\max}}$ must belong to the range of $f \in L^2(\mathbb{R}) \to \{f \star \psi_{k_j}\}_{J_{\min}-1 \leq j \leq J_{\max}}$. The minimization is performed by gradient descent, using the previously found estimations as initialization.

Finally, we deduce f from the $f \star \psi_{k_{j-1}}$ and refine this estimation one more time by a few steps of the classical Gerchberg-Saxton algorithm ([Gerchberg and Saxton, 1972]). This final refinement step is useful, because the Gerchberg-Saxton algorithm converges much faster than the gradient descent. According to our tests, the performances of the algorithm would be approximately the same with more gradient descent iterations and no final refinement. However, the execution time would be much longer.

The principle of the algorithm is summarized by the pseudo-code 4.

Algorithm 4 Reconstruction algorithm

Input: $\{|f \star \psi_j|\}_{J_{\min} \leq j \leq J_{\max}}$ and $f \star \phi_{J_{\max}}$
1: Choose $k_{J_{\min}-1}, ..., k_{J_{\max}}$ as in (3.39).
2: **for all** j **do**
3: Determine $g_j = (f \star \psi_{k_{j-1}}).(\overline{f \star \psi_{k_j}})$ from $|f \star \psi_j|^2$.
4: **end for**
5: Determine $f \star \psi_{k_{J_{\max}}}$ from $f \star \phi_{J_{\max}}$.
6: **for all** j **do**
7: Estimate $h_j \approx f \star \psi_{k_j}$.
8: **end for**
9: Refine the estimation with a gradient descent.
10: Deduce f from $\{f \star \psi_{k_j}\}_{J_{\min}-1 \leq j \leq J_{\max}}$.
11: Refine the estimation of f with the Gerchberg-Saxton algorithm.
Output: f

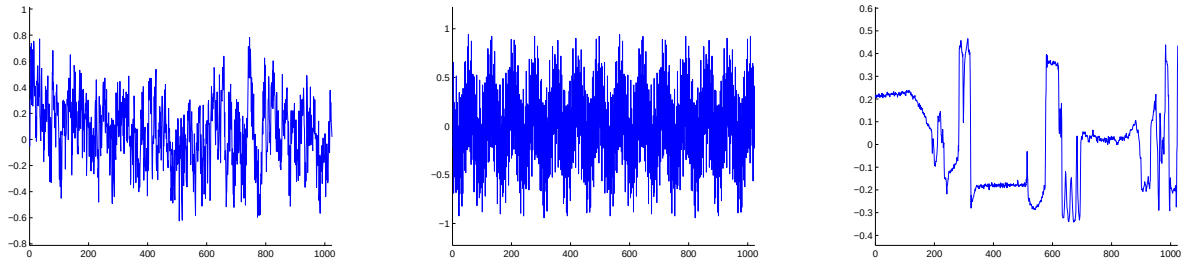
3.5.2 Input signals

We study the performances of this algorithm on three classes of input signals with finite size n. The figure 3.4 shows an example for each of these three classes.

The first class contains realizations of Gaussian processes with renormalized frequencies. More precisely, the signals f of this class satisfy:

$$\hat{f}[n] = \frac{X_n}{\sqrt{n+2}}$$

where the X_n are independent realizations of a Gaussian random variable $X \sim \mathcal{N}(0,1)$. The normalization $\frac{1}{\sqrt{n+2}}$ ensures that all dyadic frequency bands contain approximately the same amount of energy.



(a) (b) (c)

Figure 3.4: examples of signals: (a) realization of a Gaussian process (b) sum of sinusoids (c) piecewise regular

The second class consists in sums of a few sinusoids. The amplitudes, phases and frequencies of the sinusoids are randomly chosen. In each dyadic frequency band, there is approximately the same mean number of sinusoids (slightly smaller than 1).

The signals of the third class are random lines extracted from real images. They usually are structured signals, with smooth regular parts and large discontinuities at a small number of points.

To study the influence of the size of the signals on the reconstruction, we perform tests for signals of size $N = 128, N = 1024$ and $N = 8192$. For each N, we used $\log_2(N) - 1$ Cauchy wavelets of order $p = 3$. Our low-pass filter is a Gaussian function of the form $\hat{\phi}[k] = \exp(-\alpha k^2/2)$, with α independent of N.

3.5.3 Noise

The inputs that are provided to the algorithm are not exactly $\{|f \star \psi_j|\}, f \star \phi_{J_{\max}}$ but $\{|f \star \psi_j| + n_{\psi,j}\}, f \star \phi_{J_{\max}} + n_\phi$. The $n_{\psi,j}$ and the n_ϕ represent an additive noise. In all our experiments, this noise is white and Gaussian.

We measure the amplitude of the noise in relative l^2-norm:

$$\text{relative noise} = \frac{\sqrt{||n_\phi||_2^2 + \sum_j ||n_{\psi,j}||_2^2}}{\sqrt{||f \star \phi_{J_{\max}}||_2^2 + \sum_j ||f \star \psi_j||_2^2}}$$

3.5.4 Results

The results are displayed on the figure 3.5.

The x-axis displays the relative error induced by the noise over the input and the y-axis represents the reconstruction error, both over the reconstructed function and over the modulus of the wavelet transform of the reconstructed function.

For an input signal f and output f_{rec}, we define the relative error between f and f_{rec} by:

$$\text{function error } = \frac{||f - f_{rec}||_2}{||f||_2}$$

and the relative error over the modulus of the wavelet transform by:

$$\text{modulus error } = \frac{\sqrt{||f \star \phi_{J_{\max}} - f_{rec} \star \phi_{J_{\max}}||_2^2 + \sum_j || \, |f \star \psi_j| - |f_{rec} \star \psi_j| \, ||_2^2}}{\sqrt{||f \star \phi_{J_{\max}}||_2^2 + \sum_j ||f \star \psi_j||_2^2}}$$

The modulus error describes the capacity of the algorithm to reconstruct a signal whose wavelet transform is close, in modulus, to the one which has been provided as input. The function error, on the other hand, quantifies the intrinsic stability of the phase retrieval problem. If the modulus error is small but the function error is large, it means that there are several functions whose wavelet transforms are almost equal in moduli and the reconstruction problem is ill-posed.

An ideal reconstruction algorithm would yield a small modulus error (that is, proportional to the noise over the input). Nevertheless, the function error could be large or small, depending on the well-posedness of the phase retrieval problem.

We expect that our algorithm may fail when the input modulus contain very small values (because the algorithm performs divisions, which become very unstable in presence of zeroes).

For almost each of the signals that we consider, there exist x's such that $f \star \psi_{k_j}(x) \approx 0$ but the number of such points vary greatly, depending on which class the signal belongs. As an example, the wavelet transforms of the three signals of the figure 3.4 are displayed in 3.6.

For Gaussian signals, there are generally not many points at which the wavelet transform vanishes. The positions of these points do not seem to be correlated in either space or frequency.

For piecewise regular signals, there are more of this points but they are usually distributed in such a way that if $f \star \psi_j(x) \approx 0$, then $f \star \psi_k(x) \approx 0$ for all wavelets ψ_k of higher frequencies than ψ_j. This distribution makes the reconstruction easier.

When the signals are sums of sinusoids, it often happens that some components of the wavelet transform are totally negligible: for some j, $f \star \psi_j(x) \approx 0$ for any x. The negligible frequencies may be either high, low or intermediate.

From the results shown in 3.5, it is clear that the number of zeros influences the reconstruction, but also that isolated zeroes do not prevent reconstruction. The algorithm performs well on

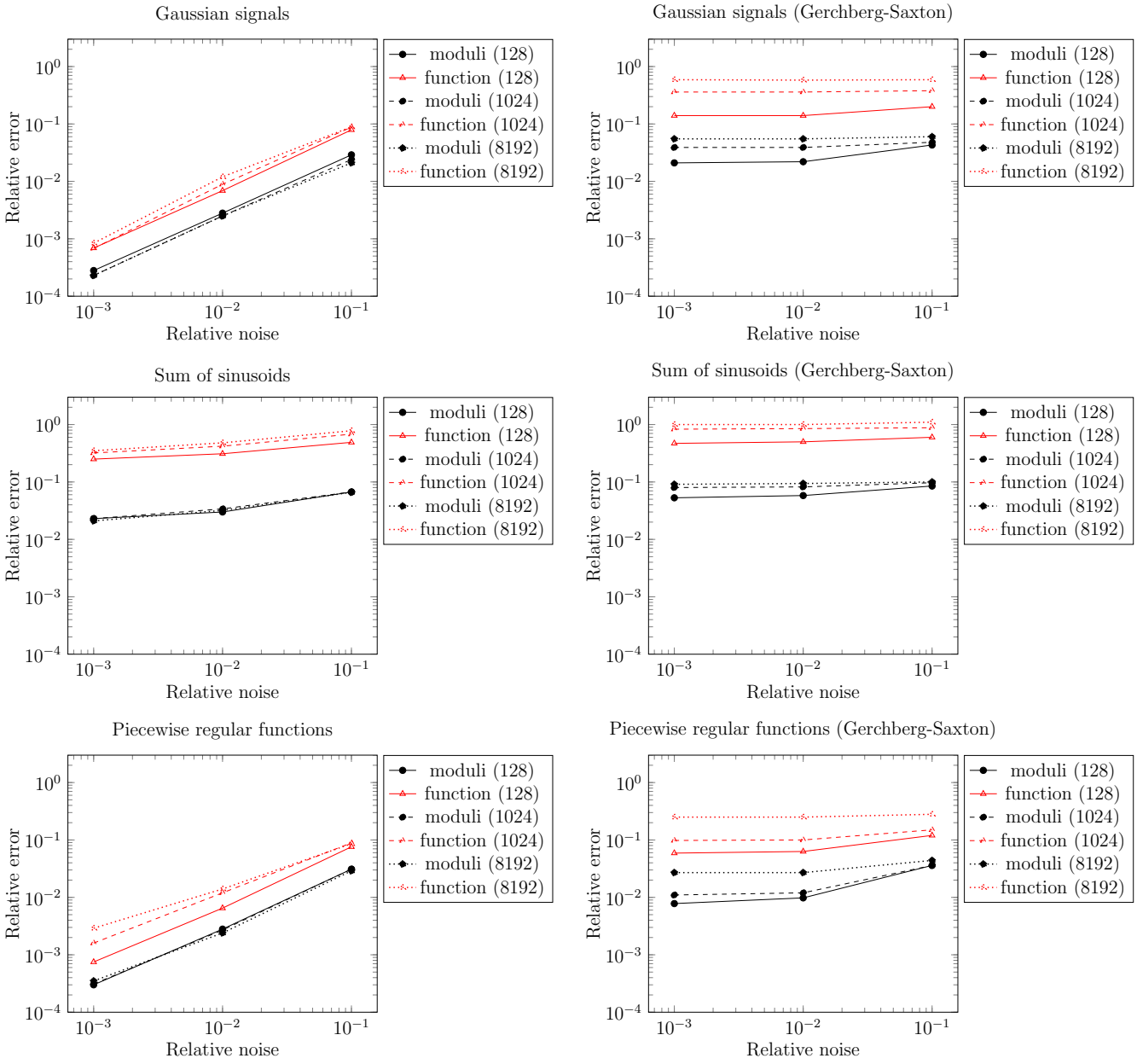
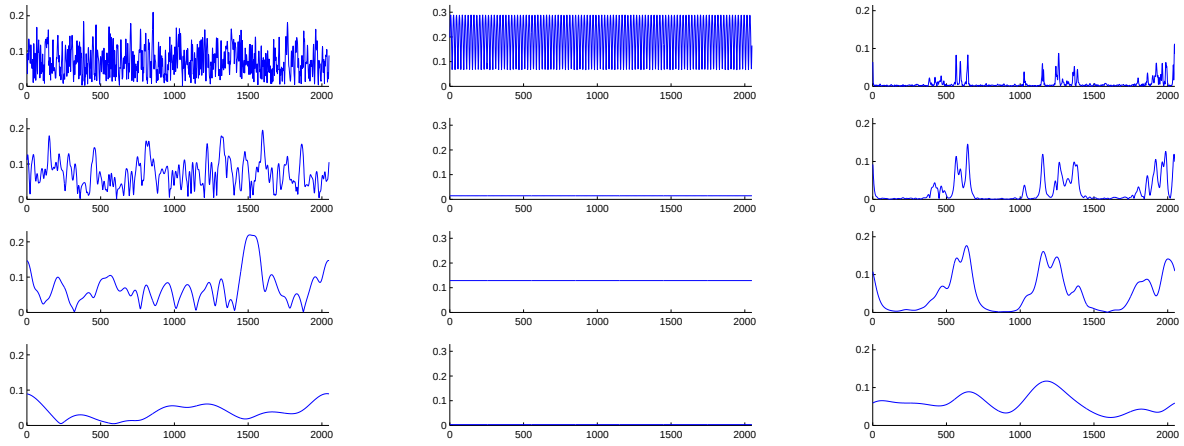


Figure 3.5: Reconstruction results for the three considered classes of signals. Left column: our algorithm. Right column: alternate projections (Gerchberg-Saxton)



(a) (b) (c)

Figure 3.6: wavelet transforms, in modulus, of the signals of the figure 3.4: (a) realization of a Gaussian process (b) sum of sinusoids (c) piecewise regular
Each column represents the wavelet transform of one signal. Each graph corresponds to one frequency component of the wavelet transform. For sake of visibility, only 4 components are shown, although nine were used in the calculation.

Gaussian or piecewise regular signals. The distance in modulus between the wavelet transform of the reconstructed signal and of the original one is proportional to the amount of noise (and generally significantly smaller). This holds up to large levels of noise (10%). By comparison, the classical Gerchberg-Saxton algorithm is much less efficient.

However, the algorithm often fails when the input signal is a sum of sinusoids. Not surprisingly, the most difficult signals in this class are the ones for which the sinusoids are not equally distributed among frequency bands and the wavelet transform has a lot of zeroes. The relative error over the modulus of the wavelet transform is then often of several percent, even when the relative error induced by the noise is of the order of 0.1%.

In the section 3.3, we explained why, for any function f, it is generally possible to construct g such that f and g are not close but their wavelet transform have almost the same modulus. This construction holds provided that the time and frequency support of f is large enough.

Increasing the time and frequency support of f amounts here to increase the size N of the signals. Thus, we expect the function error to increase with N. It is indeed the case but this effect is very weakly perceptible on Gaussian signals. It is stronger on piecewise regular functions, probably because the wavelet transforms of these signals have more zeroes; their reconstruction is thus less stable.

In the case of the sums of sinusoids, because of the failure of the algorithm, we can not draw

firm conclusions regarding the stability of the reconstruction. We nevertheless suspect that this class of signals is the least stable of all and that these instabilities are the cause of the incorrect behavior of our algorithm.

3.6 Technical lemmas

3.6.1 Lemmas of the proof of Theorem 3.1

Proof of Lemma 3.9. We recall equation (3.22):

$$\frac{B_F(z+i\alpha)\tilde{B}_G(z)}{B_G(z+i\alpha)\tilde{B}_F(z)} = e^{iC+iBz}\frac{S_G(z+i\alpha)}{S_F(z+i\alpha)} \tag{3.22}$$

We want to show that the left part of this equality admits a meromorphic extension to $\mathbb{C}$. We also want this meromorphic extension to have the same poles (with multiplicity) than it would if all four functions $B_F, B_G, \tilde{B}_F$ and $\tilde{B}_G$ were meromorphically defined over all $\mathbb{C}$.

We first remark that $\tilde{B}_F$ and $\tilde{B}_G$ admit meromorphic extensions to $\mathbb{C}$. Indeed, if the $(z_k)_k$ are the zeros of $F(.+i\alpha)$ in $\mathbb{H}$, this set has no accumulation point in $\overline{H}$: if z_∞ was an accumulation point, $z_\infty + i\alpha \in \mathbb{H}$ would be an accumulation point of the zeros of F and, as F is holomorphic, it would be the null function. From the classical properties of Blaschke products, $\tilde{B}_F$ converge over $\mathbb{C}$ and so does $\tilde{B}_G$.

On the contrary, B_F and B_G may not admit meromorphic extensions over $\mathbb{C}$. But their quotient B_F/B_G does.

We define:

$$B'_F(z) = \left(\frac{z-i}{z+i}\right)^{m_F} \prod_k \frac{|z_k^F - i|}{z_k^F - i}\frac{|z_k^F + i|}{z_k^F + i}\frac{z - z_k^F}{z - \overline{z}_k^F}$$

where the (z_k^F)'s are the zeros of F, each z_k^F being counted, not with multiplicity $\mu_F(z_k^F)$, but with multiplicity $\max(0, \mu_F(z_k^F) - \mu_G(z_k^F))$ (and m_F is still the multiplicity of i as a zero of F).

Similarly:

$$B'_G(z) = \left(\frac{z-i}{z+i}\right)^{m_G} \prod_k \frac{|z_k^G - i|}{z_k^G - i}\frac{|z_k^G + i|}{z_k^G + i}\frac{z - z_k^G}{z - \overline{z}_k^G}$$

where the (z_k^G)'s are the zeros of G counted with multiplicity $\max(0, \mu_G(z_k^G) - \mu_F(z_k^G))$.

We define:

$$B_{F,G}(z) = \prod_k \frac{|z_k^{F,G} - i|}{z_k^{F,G} - i}\frac{|z_k^{F,G} + i|}{z_k^{F,G} + i}\frac{z - z_k^{F,G}}{z - \overline{z}_k^{F,G}}$$

where the $z_k^{F,G}$ are the zeros of F or G, counted with multiplicity $\min(\mu_F(z_k^{F,G}), \mu_G(z_k^{F,G}))$. The function $B_{F,G}$ corresponds to the "common part" of B_F and B_G, which we may factorize in the quotient B_F/B_G.

The products $B'_F, B'_G, B_{F,G}$ converge over $\mathbb{H}$ and, for all $z \in \mathbb{H}$:

$$B_F(z) = B'_F(z)B_{F,G}(z) \qquad B_G(z) = B'_G(z)B_{F,G}(z)$$

So for all $z \in \mathbb{H}$:

$$\frac{B_F(z+i\alpha)\tilde{B}_G(z)}{B_G(z+i\alpha)\tilde{B}_F(z)} = \frac{B'_F(z+i\alpha)\tilde{B}_G(z)}{B'_G(z+i\alpha)\tilde{B}_F(z)}$$

If we show that B'_F and B'_G converge over $\mathbb{C}$, we can take $B_w(z) = \frac{B'_F(z+i\alpha)\tilde{B}_G(z)}{B'_G(z+i\alpha)\tilde{B}_F(z)}$. It will be meromorphic over $\mathbb{C}$.

To prove this, we first establish a relation between the zeros of F and G.

Let z be such that $0 < \operatorname{Im} z \leq \alpha$. The zeros of B_F are the zeros of F in $\mathbb{H}$, counted with multiplicity. Thus, $z - i\alpha$ is a zero of $B_F(. + i\alpha)$ with multiplicity $\mu_F(z)$. It is a zero of $B_G(. + i\alpha)$ with multiplicity $\mu_G(z)$.

Because $\operatorname{Im}(z - i\alpha) \leq 0$, it is not a zero of $\tilde{B}_F$ (resp. $\tilde{B}_G$) but may be a pole. As a pole, its multiplicity is the multiplicity of $\overline{z - i\alpha} = \overline{z} + i\alpha$ as a zero of $F(. + i\alpha)$ (resp. $G(. + i\alpha)$): it is $\mu_F(\overline{z} + 2i\alpha)$ (resp. $\mu_G(\overline{z} + 2i\alpha)$).

The right part of (3.22), $e^{iC+iBz}\frac{S_G(z+i\alpha)}{S_F(z+i\alpha)}$ has no zero neither pole over $\{z \in \mathbb{C} \text{ s.t. } \operatorname{Im} z > -\alpha\}$ (from the definition of S_G and S_F given in (3.20)). So neither does the left part. In particular, $z - i\alpha$ is not a zero and is not a pole:

$$\mu_F(z) - \mu_G(z) - \mu_G(\overline{z} + 2i\alpha) + \mu_F(\overline{z} + 2i\alpha) = 0 \tag{3.40}$$

We now explain why B'_F converges over $\mathbb{C}$. The same result will hold for B'_G. From the properties of Blaschke products, B'_F converges over $\mathbb{C}$ if (z_k^F) has no accumulation point in $\mathbb{R}$.

By contradiction, we assume that some subsequence of (z_k^F), denoted by $(z_{\phi(k)}^F)$, converges to $\lambda \in \mathbb{R}$. Because the z_k^F's appear in B'_F with multiplicity $\max(0, \mu_F(z_k^F) - \mu_G(z_k^F))$, we must have:

$$\mu_F(z_{\phi(k)}^F) - \mu_G(z_{\phi(k)}^F) > 0 \qquad \forall k \in \mathbb{N}$$

We can assume that, for all k, $0 < \operatorname{Im} z_{\phi(k)}^F \leq \alpha$. From (3.40):

$$\mu_G(\overline{z}_{\phi(k)}^F + 2i\alpha) - \mu_F(\overline{z}_{\phi(k)}^F + 2i\alpha) = \mu_F(z_{\phi(k)}^F) - \mu_G(z_{\phi(k)}^F) > 0$$

Consequently, $\overline{z}_{\phi(k)}^F + 2i\alpha$ is a zero of G for all k. As $z_{\phi(k)}^F \to \lambda \in \mathbb{R}$, $\lambda + 2i\alpha \in \mathbb{H}$ is an accumulation point of the zeros of G. This is impossible because G is holomorphic over $\mathbb{H}$ and we have assumed that it was not the null function.

To conclude, we have to prove Equation (3.23).

For any $z \in \mathbb{H}$, the multiplicity of $\overline{z} - i\alpha$ as a pole of $B'_F(. + i\alpha)$ is the multiplicity of z as a zero of B'_F, that is $\max(0, \mu_F(z) - \mu_G(z))$. Its multiplicity as a pole of $B'_G(. + i\alpha)$ is $\max(0, \mu_G(z) - \mu_F(z))$. As a pole of $\tilde{B}_F$ (resp. $\tilde{B}_G$), it is $\mu_F(z + 2i\alpha)$ (resp. $\mu_G(z + 2i\alpha)$).

The multiplicity of $\overline{z} - i\alpha$ as a pole of B_w is then, as required:

$$\begin{aligned}\max(0, \mu_F(z) - \mu_G(z)) - \max(0, \mu_G(z) - \mu_F(z)) - \mu_F(z + 2i\alpha) + \mu_G(z + 2i\alpha)\\ = (\mu_F(z) - \mu_G(z)) - (\mu_F(z + 2i\alpha) - \mu_G(z + 2i\alpha))\end{aligned}$$

□

Proof of Lemma 3.10. We call dE_F and dE_G the singular measures appearing in the definitions of S_F and S_G (see (3.20)).

From equation (3.22) and Lemma 3.9, for any $z \in \mathbb{H}$:

$$\exp\left(\frac{i}{\pi}\int_{\mathbb{R}} \frac{1+tz}{t-z}(dE_G - dE_F)(t)\right) = \frac{S_G(z)}{S_F(z)} = B_w(z - i\alpha)e^{-iC-iB(z-i\alpha)}$$

The function $z \to B_w(z - i\alpha)e^{-iC-iB(z-i\alpha)}$ is meromorphic over $\mathbb{C}$. From the following lemma, $dE_G - dE_F$ must then be the null measure, so $S_G = S_F$ over $\mathbb{H}$.

Lemma 3.19. *Let dE be a real bounded measure, singular with respect to Lebesgue measure. We define:*

$$S(z) = \exp\left(\frac{i}{\pi}\int_{\mathbb{R}} \frac{1+tz}{t-z}dE(t)\right) \qquad \forall z \in \mathbb{H}$$

If S admits a meromorphic extension in the neighborhood of each point of $\mathbb{R}$, then $dE = 0$.

Proof. Let $s(z) = -\log|S(z)|$ for all $z \in \mathbb{H}$. This is well-defined and:

$$s(x + iy) = \frac{1}{\pi}\int_{\mathbb{R}} \frac{y}{(t-x)^2 + y^2}(1 + t^2)dE(t) \qquad \forall x, y \in \mathbb{R} \text{ s.t. } y > 0$$

This is the Poisson integral of $(1 + t^2)dE(t)$. So, as dE is bounded, $(1 + t^2)dE(t)$ is the limit, in the sense of distributions, of $s(t + iy)dt$ when $y \to 0^+$. The principle of the proof will then be to show that $s(. + iy)$ also converges to $-\log|S_{|\mathbb{R}}|$, where $S_{|\mathbb{R}}$ is the extension of S to $\mathbb{R}$, so $dE = -\frac{\log|S(t)|dt}{1+t^2}$. The singularity of dE will imply $\log|S_{|\mathbb{R}}| = 0$ and $dE = 0$.

We still denote by $S(t)$ the meromorphic extension of S to a neighborhood of $\overline{H}$. Let $\{r_k\}$ be the zeros or poles of S.

When $y \to 0^+$, $s(. + iy)$ tends to $-\log|S|$ almost everywhere. On every compact of $\mathbb{R} - \{r_k\}$, the convergence is uniform, and thus in L^1.

Let r_k be any zero or pole and $\epsilon > 0$ be such that S admits a meromorphic extension over a neighborhood of $[r_k - \epsilon; r_k + \epsilon] \times [-\epsilon; \epsilon]$ and $r_j \notin [r_k - \epsilon; r_k + \epsilon]$ for all $j \neq k$. There exist $h : [r_k - \epsilon; r_k + \epsilon] \times [-\epsilon; \epsilon] \to \mathbb{C}$ holomorphic and $m \in \mathbb{Z}$ such that:

$$S(z) = (z - r_k)^m h(z) \quad \forall z \in [r_k - \epsilon; r_k + \epsilon] \times [-\epsilon; \epsilon] \qquad \text{and} \quad h(r_k) \neq 0$$

For all $y \in]0; \epsilon[$:

$$\begin{aligned}\int_{r_k-\epsilon}^{r_k+\epsilon} |s(t+iy) + \log|S(t)||dt &= \int_{r_k-\epsilon}^{r_k+\epsilon} \big|m \log|t - r_k + iy| + \log|h(t+iy)| \\ &\qquad - m \log|t - r_k| - \log|h(t)|\big| dt \\ &\leq m \int_{r_k-\epsilon}^{r_k+\epsilon} \big|\log|t - r_k + iy| - \log|t - r_k|\big| dt \\ &\quad + \int_{r_k-\epsilon}^{r_k+\epsilon} \big|\log|h(t+iy)| - \log|h(t)|\big| dt \end{aligned} \tag{3.41}$$

As $\log|h|$ is continuous, $\log|h(. + iy)|$ converges uniformly to $\log|h_{|\mathbb{R}}|$ over $[r_k - \epsilon; r_k + \epsilon]$:

$$\int_{r_k-\epsilon}^{r_k+\epsilon} \big|\log|h(t+iy)| - \log|h(t)|\big| dt \to 0 \quad \text{when } y \to 0^+$$

As $\log|. - r_k + iy|$ converges to $\log|. - r_k|$ in $L^1([r_k - \epsilon; r_k + \epsilon])$:

$$\int_{r_k-\epsilon}^{r_k+\epsilon} \big|\log|t - r_k + iy| - \log|t - r_k|\big| dt \to 0$$

So, by (3.41), $s(. + iy)$ converges in L^1 to $t \in \mathbb{R} \to -\log|S(t)|$, over $[r_k - \epsilon; r_k + \epsilon]$. As the sequence (r_k) has no accumulation point in $\mathbb{R}$, $s(. + iy) \to -\log|S_{\mathbb{R}}|$ (in L^1) over each compact set of $\mathbb{R}$.

For every $f \in \mathcal{C}_c^0(\mathbb{R})$:

$$\int_{\mathbb{R}} f(t)(1+t^2)dE(t) = \lim_{y \to 0^+} \int_{\mathbb{R}} s(t+iy)f(t)dt = -\int_{\mathbb{R}} \log|S(t)|f(t)dt$$

We deduce that $dE(t) = -\frac{\log|S(t)|dt}{1+t^2}$. As dE is singular with respect to Lebesgue measure, we must have $\log|S(t)| = 0$ for all $t \in \mathbb{R}$ and $dE = 0$. □

□

3.6.2 Lemmas of the proof of Theorem 3.11

Proof of Lemma 3.12. We first recall the Riesz-Fréchet-Kolmogorov theorem.

Theorem (Riesz-Fréchet-Kolomogorov)**.** *Let $p \in [1; +\infty[$. Let $\mathcal{F}$ be a subset of $L^p(\mathbb{R})$. The set $\mathcal{F}$ is relatively compact if and only if:*

(i) $\mathcal{F}$ is bounded.

(ii) For every $\epsilon > 0$, there exists some compact $K \subset \mathbb{R}$ such that:

$$\sup_{f \in \mathcal{F}} ||f||_{L^p(\mathbb{R}-K)} \leq \epsilon$$

(iii) For every $\epsilon > 0$, there exists $\delta > 0$ such that:

$$\sup_{f \in \mathcal{F}} ||f(. + h) - f||_p \leq \epsilon \qquad \forall h \in [-\delta; \delta]$$

We want to apply this theorem to $p = 2$ and $\mathcal{F} = \{f_n \star \psi_j\}_{n \in \mathbb{N}}$.

First of all, $\mathcal{F}$ is bounded: actually, from (3.27), $(f_n)_{n \in \mathbb{N}}$ itself is bounded (because $(U(f_n))_n$ converges and thus is bounded). It implies that $\{f_n \star \psi_j\}_n$ is bounded because $||f_n \star \psi_j||_2 \leq ||f_n||_2 ||\psi_j||_1$ (by Young's inequality).

Let us now prove (ii). Let any $\epsilon > 0$ be fixed.

The sequence $(|f_n \star \psi_j|)_n$ converges in $L^2(\mathbb{R})$ (to h_j, because $U(f_n) \to (h_j)_{j \in \mathbb{Z}}$ in $l^2(\mathbb{Z}, L^2(\mathbb{R}))$). So $\{|f_n \star \psi_j|\}_n$ is relatively compact in $L^2(\mathbb{R})$. By the Riesz-Fréchet-Kolmogorov theorem, there exists $K \subset \mathbb{R}$ a compact set such that:

$$\sup_{n \in \mathbb{N}} || \, |f_n \star \psi_j| \, ||_{L^2(\mathbb{R}-K)} \leq \epsilon$$

But, for all n, $|| \, |f_n \star \psi_j| \, ||_{L^2(\mathbb{R}-K)} = ||f_n \star \psi_j||_{L^2(\mathbb{R}-K)}$ so (ii) holds:

$$\sup_{n \in \mathbb{N}} ||f_n \star \psi_j||_{L^2(\mathbb{R}-K)} \leq \epsilon$$

We finally check (iii). Let $\epsilon > 0$ be fixed. For any $h \in \mathbb{R}$:

$$||(f_n \star \psi_j)(. + h) - (f_n \star \psi_j)||_2 = ||f_n \star (\psi_j(. - h) - \psi_j)||_2 \leq ||f_n||_2 ||\psi_j(. - h) - \psi_j||_1$$

As $\sup_n ||f_n||_2 < +\infty$ and $\lim_{h \to 0} ||\psi_j(. - h) - \psi_j||_1 = 0$ (this property holds for any L^1 function), we have, for $\delta > 0$ small enough:

$$\sup_n ||(f_n \star \psi_j)(. + h) - (f_n \star \psi_j)||_2 \leq \epsilon \qquad \forall h \in [-\delta; \delta]$$

□

Proof of Lemma 3.13. We want to find $g \in L^2_+(\mathbb{R})$ such that $\hat{l}_j = \hat{g}\hat{\psi}_j$ for every $j \in \mathbb{Z}$.

If $\omega \leq 0$, we set $\hat{g}(\omega) = 0$. Then, for each j, we set $\hat{g} = \hat{l}_j/\hat{\psi}_j$ on the support of $\hat{\psi}_j$, which we denote by Supp $\hat{\psi}_j$. This definition is correct in the sense that:

$$\text{if } j_1 \neq j_2, \quad \frac{\hat{l}_{j_1}}{\hat{\psi}_{j_1}} = \frac{\hat{l}_{j_2}}{\hat{\psi}_{j_2}} \text{ a.e. on Supp } \hat{\psi}_{j_1} \cap \text{Supp } \hat{\psi}_{j_2}$$

Indeed, for all n, $(f_{\phi(n)} \star \psi_{j_1}) \star \psi_{j_2} = (f_{\phi(n)} \star \psi_{j_2}) \star \psi_{j_1}$ so, by taking the limit in n, $l_{j_1} \star \psi_{j_2} = l_{j_2} \star \psi_{j_1}$ and $\hat{l}_{j_1}\hat{\psi}_{j_2} = \hat{l}_{j_2}\hat{\psi}_{j_1}$.

We can note that, for all j, $\hat{g}\hat{\psi}_j = \hat{l}_j$. It is true on Supp $\hat{\psi}_j$, by definition. And, on $\mathbb{R}-$Supp $\hat{\psi}_j$, $\hat{l}_j = 0 = \hat{g}\hat{\psi}_j$ because $\hat{l}_j$ is the L^2-limit of $\hat{f}_{\phi(n)}\hat{\psi}_j$ and $\hat{f}_{\phi(n)}\hat{\psi}_j = 0$ on $\mathbb{R} - \text{Supp } \hat{\psi}_j$.

The $\hat{g}$ we just defined belongs to $L^2(\mathbb{R})$. Indeed, by (3.26):

$$||\hat{g}||_2^2 \leq \frac{1}{B}\int_{\mathbb{R}^+} |\hat{g}|^2 \sum_j |\hat{\psi}_j|^2 = \frac{1}{B}\int_{\mathbb{R}^+} \sum_j |\hat{l}_j|^2 = \frac{1}{B}\sum_j ||l_j||_2^2$$

As $f_{\phi(n)} \star \psi_j$ goes to l_j when n goes to ∞ and $U(f_{\phi(n)}) = \{|f_{\phi(n)} \star \psi_j|\}_j$ goes to $(h_j)_{j\in\mathbb{Z}} \in l^2(\mathbb{Z}, L^2(\mathbb{R}))$, we must have $|l_j| = h_j$ for each j. So $\frac{1}{B}\sum\limits_j ||l_j||_2^2 = \frac{1}{B}\sum\limits_j ||h_j||_2^2 = \frac{1}{B}||(h_j)_{j\in\mathbb{Z}}||_2^2 < +\infty$ and $\hat{g}$ belongs to $L^2(\mathbb{R})$.

As $\hat{g} \in L^2(\mathbb{R})$, it is the Fourier transform of some $g \in L^2(\mathbb{R})$. For all $j \in \mathbb{Z}$, as $\hat{g}\hat{\psi}_j = \hat{l}_j$, we have $g \star \psi_j = l_j$.

We now show that $f_{\phi(n)} \to g$ when $n \to \infty$.

For every $J, n \in \mathbb{N}$:

$$\begin{aligned}
\sqrt{\sum_{|j|>J} ||f_{\phi(n)} \star \psi_j||_2^2} &= \sqrt{\sum_{|j|>J} ||\, U(f_{\phi(n)})_j \,||_2^2} \\
&\leq \sqrt{\sum_{|j|>J} ||U(f_{\phi(n)})_j - h_j||_2^2} + \sqrt{\sum_{|j|>J} ||h_j||_2^2} \\
&\leq ||U(f_{\phi(n)}) - (h_j)||_2 + \sqrt{\sum_{|j|>J} ||h_j||_2^2}
\end{aligned}$$

So $\limsup\limits_n \left(\sum\limits_{|j|>J} ||f_{\phi(n)} \star \psi_j||_2^2 \right) \leq \sum\limits_{|j|>J} ||h_j||_2^2$ and:

$$\limsup_n \left(\sum_{j\in\mathbb{Z}} ||f_{\phi(n)} \star \psi_j - g \star \psi_j||_2^2 \right) \leq \limsup_n \left(\sum_{|j|\leq J} ||f_{\phi(n)} \star \psi_j - g \star \psi_j||_2^2 \right)$$

$$
\begin{aligned}
&+ \limsup_n \left(\sum_{|j|>J} ||f_{\phi(n)} \star \psi_j - g \star \psi_j||_2^2 \right) \\
&= \limsup_n \left(\sum_{|j|>J} ||f_{\phi(n)} \star \psi_j - g \star \psi_j||_2^2 \right) \\
&\leq \sum_{|j|>J} ||h_j||_2^2
\end{aligned}
$$

This last quantity may be as small as desired, for J large enough, so $\sum_{j\in\mathbb{Z}} ||f_{\phi(n)} \star \psi_j - g \star \psi_j||_2^2 \to 0$. By (3.26):

$$
\begin{aligned}
\sum_{j\in\mathbb{Z}} ||f_{\phi(n)} \star \psi_j - g \star \psi_j||_2^2 &= \int_{\mathbb{R}} \left|\hat{f}_{\phi(n)} - \hat{g}\right|^2 (\sum_j |\hat{\psi}_j|^2) \\
&\geq A \int_{\mathbb{R}} \left|\hat{f}_{\phi(n)} - \hat{g}\right|^2 \\
&= \frac{A}{2\pi} ||f_{\phi(n)} - g||_2^2
\end{aligned}
$$

so $||f_{\phi(n)} - g||_2 \to 0$. □

3.6.3 Proof of Theorem 3.17

In this section, we prove Theorem 3.17, which gives a stability result for the case of dyadic wavelets.

For all $y > 0$, we define:

$$\mathcal{N}(y) = \sup_{x\in\mathbb{R}, s=1,2} |F^{(s)}(x + iy)|$$

The following lemma is not necessary to our proof but we will use it to progressively simplify our inequalities.

Lemma 3.20. *For all $y_1, y_2 \in \mathbb{R}_+^*$, if $y_1 < y_2$:*

$$\mathcal{N}(y_1) \geq \mathcal{N}(y_2) \tag{3.42}$$

and for all $y_3 \in [y_1; y_2]$:

$$\mathcal{N}(y_3) \leq \mathcal{N}(y_1)^{\frac{y_2-y_3}{y_2-y_1}} \mathcal{N}(y_2)^{\frac{y_3-y_1}{y_2-y_1}} \tag{3.43}$$

Proof. The second inequality comes directly from Theorem 3.24, applied to functions $F^{(1)}$ and $F^{(2)}$ on the band $\{z \in \mathbb{C} \text{ s.t. } y_1 < \operatorname{Im} z < y_2\}$.

The first inequality may be derived from (3.43). The function $\mathcal{N}(y)$ is bounded when $y \to +\infty$.

Keeping y_1 and y_3 fixed in (3.43) and letting y_2 go to $+\infty$ then gives:

$$\mathcal{N}(y_3) \leq \mathcal{N}(y_1)$$

□

Remark 3.21. *When $y \to +\infty$, then $\mathcal{N}(y) \to 0$ because, from (3.6) and the Hölder inequality:*

$$\begin{aligned} |F^{(s)}(x+iy)| &= \left| \frac{1}{2\pi} \int_{\mathbb{R}} \omega^p \hat{f}_+(\omega) e^{i\omega(x+iy)} d\omega \right| \\ &\leq \frac{1}{2\pi} ||\hat{f}_+||_2 ||\omega \to \omega^p e^{i\omega(x+iy)}||_2 \\ &\leq \frac{1}{2\pi} ||\hat{f}_+||_2 ||\omega \to \omega^p e^{-\omega y}||_2 \end{aligned}$$

which decreases geometrically to zero when $y \to +\infty$.

We can now prove the theorem.

Proof of Theorem 3.17. From the relation (3.31) between $F^{(s)}$ and the $f^{(s)} \star \psi_j$ and from the hypotheses, the following inequalities hold for all $x \in [-M2^j; M2^j]$:

$$\begin{gathered} \left| |F^{(1)}(x+i2^j)|^2 - |F^{(2)}(x+i2^j)|^2 \right| \leq \epsilon \mathcal{N}(2^j)^2 \\ \left| |F^{(1)}(x+i2^{j+1})|^2 - |F^{(2)}(x+i2^{j+1})|^2 \right| \leq \epsilon \mathcal{N}(2^{j+1})^2 \\ |F^{(1)}(x+i2^j)|^2, |F^{(2)}(x+i2^j)|^2 \geq c\mathcal{N}(2^j)^2 \\ |F^{(1)}(x+i2^{j+1})|^2, |F^{(2)}(x+i2^{j+1})|^2 \geq c\mathcal{N}(2^{j+1})^2 \end{gathered}$$

Let us set, for all z such that $-2^{j+1} < \operatorname{Im} z < 2^{j+1}$:

$$G(z) = F^{(1)}(z+i2^{j+1})\overline{F^{(1)}(\overline{z}+i2^{j+1})} - F^{(2)}(z+i2^{j+1})\overline{F^{(2)}(\overline{z}+i2^{j+1})}$$

For all z such that $\operatorname{Im} z = 0$:

$$\begin{aligned} |G(z)| = \left| |F^{(1)}(z+i2^{j+1})|^2 - |F^{(2)}(z+i2^{j+1})|^2 \right| &\leq \epsilon \mathcal{N}(2^{j+1})^2 && \text{if } |\operatorname{Re} z| \leq M2^j \\ &\leq \mathcal{N}(2^{j+1})^2 && \text{if } |\operatorname{Re} z| > M2^j \end{aligned}$$

and for all z such that $\operatorname{Im} z = 3.2^{j-1}$:

$$\begin{aligned}|G(z)| &= |F^{(1)}(\operatorname{Re} z + 7.2^{j-1}i)\overline{F^{(1)}(\operatorname{Re} z + 2^{j-1}i)} - F^{(2)}(\operatorname{Re} z + 7.2^{j-1}i)\overline{F^{(2)}(\operatorname{Re} z + 2^{j-1}i)}| \\ &\leq 2\mathcal{N}(7.2^{j-1})\mathcal{N}(2^{j-1})\end{aligned}$$

We apply Lemma 3.25 for $a = 0, b = 3.2^{j-1}, t = 2/3, A = \mathcal{N}(2^{j+1})^2, B = 2\mathcal{N}(7.2^{j-1})\mathcal{N}(2^{j-1})$. It implies that, for all $x \in [-\lambda M 2^j; \lambda M 2^j]$:

$$|G(x + i2^j)| \leq 2^{2/3}\epsilon^{1/3-\alpha_M}\mathcal{N}(2^{j+1})^{2/3}\mathcal{N}(2^{j-1})^{2/3}\mathcal{N}(7.2^{j-1})^{2/3}$$

where $\alpha_M = \frac{4}{3}\frac{\exp\left(-\frac{2\pi}{3}(1-\lambda)M\right)}{1-\exp\left(-\frac{2\pi}{3}(1-\lambda)M\right)}$.

Replacing G by its definition gives, for all $x \in [-\lambda M 2^j; \lambda M 2^j]$:

$$\begin{aligned}&|F^{(1)}(x + i3.2^j)\overline{F^{(1)}(x + i2^j)} - F^{(2)}(x + i3.2^j)\overline{F^{(2)}(x + i2^j)}| \\ &\qquad \leq 2^{2/3}\epsilon^{1/3-\alpha_M}\mathcal{N}(2^{j+1})^{2/3}\mathcal{N}(2^{j-1})^{2/3}\mathcal{N}(7.2^{j-1})^{2/3} \\ &\qquad \leq 2\epsilon^{1/3-\alpha_M}\mathcal{N}(2^{j+1})^{4/3}\mathcal{N}(2^{j-1})^{2/3}\end{aligned}$$

We used Equation (3.42) to obtain the last inequality.

So, for all $x \in [-\lambda M 2^j; \lambda M 2^j]$:

$$\begin{aligned}&\Big|F^{(1)}(x + i3.2^j)\overline{F^{(1)}(x + i2^j)}F^{(2)}(x + i2^j)\overline{F^{(2)}(x + i2^j)} \\ &\quad -F^{(2)}(x + i3.2^j)\overline{F^{(2)}(x + i2^j)}F^{(1)}(x + i2^j)\overline{F^{(1)}(x + i2^j)}\Big| \\ &\quad \leq |F^{(1)}(x + i3.2^j)\overline{F^{(1)}(x + i2^j)} - F^{(2)}(x + i3.2^j)\overline{F^{(2)}(x + i2^j)}|.|F^{(2)}(x + i2^j)\overline{F^{(2)}(x + i2^j)}| \\ &\quad + |F^{(2)}(x + i3.2^j)\overline{F^{(2)}(x + i2^j)}||F^{(2)}(x + i2^j)\overline{F^{(2)}(x + i2^j)} - F^{(1)}(x + i2^j)\overline{F^{(1)}(x + i2^j)}| \\ &\quad \leq 2\epsilon^{1/3-\alpha_M}\mathcal{N}(2^{j+1})^{4/3}\mathcal{N}(2^{j-1})^{2/3}|F^{(2)}(x + i2^j)|^2 + \epsilon\mathcal{N}(2^j)^2|F^{(2)}(x + i3.2^j)\overline{F^{(2)}(x + i2^j)}|\end{aligned}$$

Dividing by $|\overline{F^{(1)}(x + i2^j)F^{(2)}(x + i2^j)}|$ gives:

$$\begin{aligned}&|F^{(1)}(x + i3.2^j)F^{(2)}(x + i2^j) - F^{(2)}(x + i3.2^j)F^{(1)}(x + i2^j)| \\ &\qquad \leq 2\epsilon^{1/3-\alpha_M}\mathcal{N}(2^{j+1})^{4/3}\mathcal{N}(2^{j-1})^{2/3}\frac{|F^{(2)}(x + i2^j)|}{|F^{(1)}(x + i2^j)|} \\ &\qquad + \epsilon\mathcal{N}(2^j)^2\frac{|F^{(2)}(x + i3.2^j)|}{|F^{(1)}(x + i2^j)|}\end{aligned}$$

For each $x \in [-\lambda M 2^j; \lambda M 2^j]$, this relation also holds if we switch the roles of $F^{(1)}$ and $F^{(2)}$. Thus, we can assume that $|F^{(2)}(x+i2^j)| \leq |F^{(1)}(x+i2^j)|$. Using also the fact that $|F^{(1)}(x+i2^j)| \geq \sqrt{c}\mathcal{N}(2^j)$ yields (always for $x \in [-\lambda M 2^j; \lambda M 2^j]$):

$$|F^{(1)}(x + i3.2^j)F^{(2)}(x + i2^j) - F^{(2)}(x + i3.2^j)F^{(1)}(x + i2^j)|$$

$$
\begin{aligned}
&\leq 2\epsilon^{1/3-\alpha_M}\mathcal{N}(2^{j+1})^{4/3}\mathcal{N}(2^{j-1})^{2/3} + \frac{\epsilon}{\sqrt{c}}\mathcal{N}(2^j)\mathcal{N}(3.2^j)\\
&= 2\mathcal{N}(2^j)\mathcal{N}(3.2^j)\left(\frac{\mathcal{N}(2^{j+1})^{4/3}\mathcal{N}(2^{j-1})^{2/3}}{\mathcal{N}(2^j)\mathcal{N}(3.2^j)}\epsilon^{1/3-\alpha_M} + \frac{\epsilon}{2\sqrt{c}}\right)\\
&\leq 2\mathcal{N}(2^j)\mathcal{N}(3.2^j)\left(\left(\frac{\mathcal{N}(2^{j-1})}{\mathcal{N}(2^{j+1})}\right)^{2/3}\epsilon^{1/3-\alpha_M} + \frac{\epsilon}{2\sqrt{c}}\right)\\
&\leq 3\mathcal{N}(2^j)\mathcal{N}(3.2^j)\left(\frac{\mathcal{N}(2^{j-1})}{\mathcal{N}(2^{j+1})}\right)^{2/3}\epsilon^{1/3-\alpha_M} \qquad (3.44)
\end{aligned}
$$

In the middle, we used Equation (3.43): $\mathcal{N}(2^{j+1}) \leq \mathcal{N}(2^j)^{1/2}\mathcal{N}(3.2^j)^{1/2}$. For the last inequality, we used the fact that $c \geq \epsilon$ so $\frac{\epsilon}{2\sqrt{c}} \leq \frac{\sqrt{\epsilon}}{2} \leq \frac{\epsilon^{1/3-\alpha_M}}{2} \leq \left(\frac{\mathcal{N}(2^{j-1})}{\mathcal{N}(2^{j+1})}\right)^{2/3}\frac{\epsilon^{1/3-\alpha_M}}{2}$.

For all z such that $\operatorname{Im} z > -2^j$, we set:

$$H(z) = F^{(1)}(z+i3.2^j)F^{(2)}(z+i2^j) - F^{(2)}(z+i3.2^j)F^{(1)}(z+i2^j)$$

From (3.44):

$$
\begin{aligned}
|H(z)| &\leq 2\mathcal{N}(2^j)\mathcal{N}(3.2^j) && \text{if } \operatorname{Im} z = 0 \text{ and } |\operatorname{Re} z| > \lambda M 2^j\\
&\leq 2\mathcal{N}(2^j)\mathcal{N}(3.2^j)\min\left(1, \frac{3}{2}\left(\frac{\mathcal{N}(2^{j-1})}{\mathcal{N}(2^{j+1})}\right)^{2/3}\epsilon^{1/3-\alpha_M}\right) && \text{if } \operatorname{Im} z = 0 \text{ and } |\operatorname{Re} z| \leq \lambda M 2^j\\
&\leq 2\mathcal{N}(2^{j+3})\mathcal{N}(6.2^j) && \text{if } \operatorname{Im} z = 5.2^j
\end{aligned}
$$

We may apply Lemma 3.25 again. For all $x \in [-\lambda^2 M 2^j; \lambda^2 M 2^j]$:

$$
\begin{aligned}
|H(x+i2^j)| &\leq 2\min\left(1, \frac{3}{2}\left(\frac{\mathcal{N}(2^{j-1})}{\mathcal{N}(2^{j+1})}\right)^{2/3}\epsilon^{1/3-\alpha_M}\right)^{4/5-\alpha'_M}\\
&\quad \times \mathcal{N}(2^j)^{4/5}\mathcal{N}(3.2^j)^{4/5}\mathcal{N}(2^{j+3})^{1/5}\mathcal{N}(6.2^j)^{1/5}\\
&\leq 2\min\left(1, \frac{3}{2}\left(\frac{\mathcal{N}(2^{j-1})}{\mathcal{N}(2^{j+1})}\right)^{2/3}\epsilon^{1/3-\alpha_M}\right)^{4/5-\alpha'_M}\mathcal{N}(2^j)^{4/5}\mathcal{N}(2^{j+1})^{6/5}
\end{aligned}
$$

where $\alpha'_M = \frac{2}{5}\frac{\exp\left(-\frac{\pi}{5}\lambda(1-\lambda)M)\right)}{1-\exp\left(-\frac{\pi}{5}\lambda(1-\lambda)M)\right)}$.

Replacing H by its definition and dividing by $|F^{(1)}(x+i2^{j+1})F^{(2)}(x+i2^{j+1})|$ (which is greater that $c\mathcal{N}(2^{j+1})^2$) gives:

$$\left|\frac{F^{(1)}(x+i2^{j+2})}{F^{(1)}(x+i2^{j+1})} - \frac{F^{(2)}(x+i2^{j+2})}{F^{(2)}(x+i2^{j+1})}\right|$$

$$\leq \frac{2}{c}\min\left(1,\left(\frac{3}{2}\frac{\mathcal{N}(2^{j-1})}{\mathcal{N}(2^{j+1})}\right)^{2/3}\epsilon^{1/3-\alpha_M}\right)^{4/5-\alpha'_M}\left(\frac{\mathcal{N}(2^j)}{\mathcal{N}(2^{j+1})}\right)^{4/5}$$

As soon as $4/5-\alpha'_M>0$ and $1/3-\alpha_M>0$:

$$\begin{aligned}\left|\frac{F^{(1)}(x+i2^{j+2})}{F^{(1)}(x+i2^{j+1})}-\frac{F^{(2)}(x+i2^{j+2})}{F^{(2)}(x+i2^{j+1})}\right| &\leq \frac{3}{c}\left(\frac{\mathcal{N}(2^{j-1})}{\mathcal{N}(2^{j+1})}\right)^{8/15}\left(\frac{\mathcal{N}(2^j)}{\mathcal{N}(2^{j+1})}\right)^{4/5}\epsilon^{(1/3-\alpha_M)(4/5-\alpha'_M)}\\ &\leq \frac{3}{c}\left(\frac{\mathcal{N}(2^{j-1})}{\mathcal{N}(2^{j+1})}\right)^{4/3}\epsilon^{(1/3-\alpha_M)(4/5-\alpha'_M)}\\ &= \frac{3}{c}\left(\frac{N_{j-1}}{N_{j+1}}2^{2p}\right)^{4/3}\epsilon^{(1/3-\alpha_M)(4/5-\alpha'_M)}\end{aligned}$$

So:

$$\left|\frac{f^{(1)}\star\psi_{j+2}(x)}{f^{(1)}\star\psi_{j+1}(x)}-\frac{f^{(2)}\star\psi_{j+2}(x)}{f^{(2)}\star\psi_{j+1}(x)}\right|\leq\frac{3}{c}2^{\frac{11p}{3}}\left(\frac{N_{j-1}}{N_{j+1}}\right)^{4/3}\epsilon^{(1/3-\alpha_M)(4/5-\alpha'_M)}$$

which is the desired result with $A=3.2^{\frac{11p}{3}}$. □

3.6.4 Proof of Theorem 3.18

In this whole section, as in the paragraph 3.4.3, k is assumed to be a fixed integer such that:

$$a^{-k}<2-a$$

and we define:

$$c=1-\frac{a-1}{1-a^{-k}}$$

Lemma 3.22. *Let the following numbers be fixed:*

$$\epsilon\in]0;1[\qquad M>0\qquad \mu\in[0;M[\qquad j\in\mathbb{Z}$$

We assume that, for all $x\in[-Ma^j;Ma^j]$:

$$\left||F^{(1)}(x+ia^j)|^2-|F^{(2)}(x+ia^j)|^2\right|\leq\epsilon\mathcal{N}(a^j)^2$$

Then, for all $x\in[-(M-\mu)a^j;(M-\mu)a^j]$:

$$\begin{aligned}&\left|\overline{F^{(1)}(x+i(2a^j-a^{j+1}))}F^{(1)}(x+ia^{j+1})-\overline{F^{(2)}(x+i(2a^j-a^{j+1}))}F^{(2)}(x+ia^{j+1})\right|\\ &\qquad\leq\mathcal{N}(a^j)^{2c}\left(2\mathcal{N}(a^{j+1})\mathcal{N}(a^{j-k})\right)^{1-c}\epsilon^{c-\alpha}\end{aligned}$$

where:

$$\alpha=2\frac{e^{-\pi\mu}}{1-e^{-\pi\mu}}$$

Proof. We set:

$$H(z) = \overline{F^{(1)}(\overline{z} + ia^j)}F^{(1)}(z + ia^j) - \overline{F^{(2)}(\overline{z} + ia^j)}F^{(2)}(z + ia^j)$$

When $y = 0$, $|H(x + iy)| = \left||F^{(1)}(x + ia^j)|^2 - |F^{(2)}(x + ia^j)|^2\right|$. So:

$$\begin{aligned} |H(x + iy)| &\leq \epsilon\mathcal{N}(a^j)^2 \text{ if } x \in [-Ma^j; Ma^j] \\ &\leq \mathcal{N}(a^j)^2 \text{ if } x \notin [-Ma^j; Ma^j] \end{aligned}$$

When $y = a^j - a^{j-k}$:

$$\begin{aligned} |H(x + iy)| &= |\overline{F^{(1)}(x + ia^{j-k})}F^{(1)}(x + i(2a^j - a^{j-k})) - \overline{F^{(2)}(x + ia^{j-k})}F^{(2)}(x + i(2a^j - a^{j-k}))| \\ &\leq 2\mathcal{N}(2a^j - a^{j-k})\mathcal{N}(a^{j-k}) \qquad (\forall x \in \mathbb{R}) \end{aligned}$$

We apply Lemma 3.25 to H, restricted to the band $\{z \in \mathbb{C} \text{ s.t. } \operatorname{Im} z \in [0; a^j - a^{j-k}]\}$. From this lemma, when $y = a^{j+1} - a^j$ and $x \in [-\mu Ma^j; \mu Ma^j]$:

$$|H(x + iy)| \leq \epsilon^{f(x+iy)}\mathcal{N}(a^j)^{2c}\left(2\mathcal{N}(2a^j - a^{j-k})\mathcal{N}(a^{j-k})\right)^{1-c}$$

where $c = 1 - \frac{a-1}{1-a^{-k}}$ and:

$$f(x + iy) \geq c - 2\frac{a-1}{1-a^{-k}}\frac{e^{-\pi\frac{Ma^j-|x|}{a^j-a^{j-k}}}}{1 - e^{-\pi\frac{Ma^j-|x|}{a^j-a^{j-k}}}}$$

Because of the definition of k, $\frac{a-1}{1-a^{-k}} \leq 1$. Moreover, $\frac{Ma^j-|x|}{a^j-a^{j-k}} \geq \frac{\mu}{1-a^{-k}} \geq \mu$, so:

$$f(x + iy) \geq c - 2\frac{e^{-\pi\mu}}{1 - e^{-\pi\mu}} = c - \alpha$$

Replacing H by its definition yields:

$$\begin{aligned} \left|\overline{F^{(1)}(x + i(2a^j - a^{j+1}))}F^{(1)}(x + ia^{j+1}) - \right. & \left. \overline{F^{(2)}(x + i(2a^j - a^{j+1}))}F^{(2)}(x + ia^{j+1})\right| \\ &= |H(x + i(a^{j+1} - a^j))| \\ &\leq \epsilon^{c-\alpha}\mathcal{N}(a^j)^{2c}\left(2\mathcal{N}(2a^j - a^{j-k})\mathcal{N}(a^{j-k})\right)^{1-c} \end{aligned}$$

To conclude, it suffices to note that, because of the way we chose k, $2a^j - a^{j-k} \geq a^{j+1}$ so, from 3.20, $\mathcal{N}(2a^j - a^{j-k}) \leq \mathcal{N}(a^{j+1})$. □

Theorem 3.23. *Let the following numbers be fixed:*

$$\epsilon, \kappa \in]0;1[\text{ with } \kappa \geq \epsilon^{2(1-c)} \qquad M > 0 \qquad \mu \in [0; M[\qquad j \in \mathbb{Z} \qquad K \in \mathbb{N}$$

We assume that, for any $n \in \{j+1, ..., j+K\}$ and $x \in [-Ma^{j+K}; Ma^{j+K}]$:

$$\left| |F^{(1)}(x+ia^n)|^2 - |F^{(2)}(x+ia^n)|^2 \right| \leq \epsilon \mathcal{N}(a^n)^2 \tag{3.45}$$

$$|F^{(1)}(x+ia^n)|^2, |F^{(2)}(x+ia^n)|^2 \geq \kappa \mathcal{N}(a^n)^2 \tag{3.46}$$

We define recursively:

$$\begin{array}{cccc} & n_0 = j+K & & w_0 = a^{j+K} \\ \forall l \in \mathbb{N} & n_{l+1} = n_l - 2 & & w_{l+1} = w_l - (a-1)^2 a^{n_{l+1}} \end{array}$$

We define:

$$D_l = \prod_{s=0}^{l-1} \left(\frac{\mathcal{N}(a^{n_s-1-k})}{\mathcal{N}(a^{n_s-2})} \right) \qquad \text{and} \qquad c_l = c - 2\left(1 + \frac{2}{a}\frac{a^2-1}{a+2}\left(\sum_{s=0}^{l-1} a^{-2s}\right)\right)\left(\frac{e^{-\pi\mu}}{1-e^{-\pi\mu}}\right)$$

For any $l \geq 0$ such that $n_l \geq j$ and $M - (l+1)\mu > 0$, we have, provided that $c_l < 1$:

$$\begin{aligned} \frac{1}{\mathcal{N}(w_l)\mathcal{N}(a^{n_l})} \left| \overline{F^{(1)}(x+iw_l)} F^{(1)}(x+ia^{n_l}) - \overline{F^{(2)}(x+iw_l)} F^{(2)}(x+ia^{n_l}) \right| & \\ \leq 3D_l \left(\frac{2\kappa^{-l/2} - \kappa^{-(l-1)/2} - 1}{1-\sqrt{\kappa}} \right) \epsilon^{c_l} & \\ \left(\forall x \in [-(M-(l+1)\mu)a^{j+K}; (M-(l+1)\mu)a^{j+K}]\right) & \end{aligned} \tag{3.47}$$

Proof. We proceed by induction over l.

For $l = 0$, (3.47) is a direct consequence of (3.45). Indeed, $w_0 = a^{n_0}, D_0 = 1, c_0 < 1$ so, for $x \in [-Ma^{j+K}; Ma^{j+K}]$:

$$\frac{1}{\mathcal{N}(a^{n_0})^2} \left| |F^{(1)}(x+ia^{n_0})|^2 - |F^{(2)}(x+ia^{n_0})|^2 \right| \leq \epsilon \leq 3D_0\epsilon^{c_0}$$

We now suppose that (3.47) holds for l and prove it for $l+1$.

We proceed in two parts. First, we use the induction hypothesis to bound the function $\left| \overline{F^{(1)}(x+iw_l)} F^{(1)}(x+i(2a^{n_l-1} - a^{n_l})) - \overline{F^{(2)}(x+iw_l)} F^{(2)}(x+i(2a^{n_l-1} - a^{n_l})) \right|$. We then use this bound to obtain the desired result.

First part: by triangular inequality,

$$\begin{aligned}
&\left|\overline{F^{(1)}(x+iw_l)}\ F^{(1)}(x+i(2a^{n_l-1}-a^{n_l}))-\overline{F^{(2)}(x+iw_l)}F^{(2)}(x+i(2a^{n_l-1}-a^{n_l}))\right| \\
&\qquad\le \left|\overline{F^{(1)}(x+iw_l)}F^{(1)}(x+ia^{n_l})-\overline{F^{(2)}(x+iw_l)}F^{(2)}(x+ia^{n_l})\right| \qquad (3.48)\\
&\qquad\qquad\times\left|\frac{F^{(1)}(x+i(2a^{n_l-1}-a^{n_l}))}{F^{(1)}(x+ia^{n_l})}\right| \\
&\qquad+\left|\frac{F^{(2)}(x+ia^{n_l})}{F^{(1)}(x+ia^{n_l})}-\frac{\overline{F^{(1)}(x+ia^{n_l})}}{\overline{F^{(2)}(x+ia^{n_l})}}\right| \qquad (3.49)\\
&\qquad\qquad\times\left|\overline{F^{(2)}(x+iw_l)}F^{(1)}(x+i(2a^{n_l-1}-a^{n_l}))\right| \\
&\qquad+\left|\overline{F^{(1)}(x+ia^{n_l})}F^{(1)}(x+i(2a^{n_l-1}-a^{n_l}))\right. \\
&\qquad\qquad\left.-\overline{F^{(2)}(x+ia^{n_l})}F^{(2)}(x+i(2a^{n_l-1}-a^{n_l}))\right| \qquad (3.50)\\
&\qquad\qquad\times\left|\frac{\overline{F^{(2)}(x+iw_l)}}{\overline{F^{(2)}(x+ia^{n_l})}}\right|
\end{aligned}$$

By the induction hypothesis, for $x\in[-(M-2l\mu)a^{j+K};(M-2l\mu)a^{j+K}]$, (3.48) is bounded by:

$$\begin{aligned}
&\left|\overline{F^{(1)}(x+iw_l)}F^{(1)}(x+ia^{n_l})\ -\overline{F^{(2)}(x+iw_l)}F^{(2)}(x+ia^{n_l})\right| \\
&\qquad\le 3D_l\left(\frac{2\kappa^{-l/2}-\kappa^{-(l-1)/2}-1}{1-\sqrt{\kappa}}\right)\mathcal{N}(w_l)\mathcal{N}(a^{n_l})\epsilon^{c_l}
\end{aligned}$$

Because of (3.45) and (3.46) (for $n=n_l$), (3.49) is bounded by:

$$\left|\frac{F^{(2)}(x+ia^{n_l})}{F^{(1)}(x+ia^{n_l})}-\frac{\overline{F^{(1)}(x+ia^{n_l})}}{\overline{F^{(2)}(x+ia^{n_l})}}\right|=\left|\frac{|F^{(2)}(x+ia^{n_l})|^2-|F^{(1)}(x+ia^{n_l})|^2}{F^{(1)}(x+ia^{n_l})F^{(2)}(x+ia^{n_l})}\right|\le\frac{\epsilon}{\kappa}$$

Finally, from Lemma 3.22 applied to $j=n_l-1$, (3.50) is bounded by:

$$\begin{aligned}
&\left|\overline{F^{(1)}(x+ia^{n_l})}F^{(1)}(x+i(2a^{n_l-1}-a^{n_l}))\ -\ \overline{F^{(2)}(x+ia^{n_l})}F^{(2)}(x+i(2a^{n_l-1}-a^{n_l}))\right| \\
&\qquad\le\mathcal{N}(a^{n_l-1})^{2c}(2\mathcal{N}(a^{n_l})\mathcal{N}(a^{n_l-1-k}))^{1-c}\epsilon^{c-\alpha}
\end{aligned}$$

for all $x\in[-Ma^{j+K}+\mu a^j;Ma^{j+K}-\mu a^j]\supset[-(M-(l+1)\mu)a^{j+K};(M-(2l+1)\mu)a^{j+K}]$.

We insert these bounds into the triangular inequality. We also use the fact that $|F^{(1)}(x+ia^{n_l})|,|F^{(2)}(x+ia^{n_l})|\ge\sqrt{\kappa}\mathcal{N}(a^{n_l})$. We get, for any $x\in[-(M-(l+1)\mu)a^{j+K};(M-(l+1)\mu)a^{j+K}]$:

$$\left|\overline{F^{(1)}(x+iw_l)}\ F^{(1)}(x+i(2a^{n_l-1}-a^{n_l}))-\overline{F^{(2)}(x+iw_l)}F^{(2)}(x+i(2a^{n_l-1}-a^{n_l}))\right|$$

$$\leq \frac{1}{\sqrt{\kappa}} 3D_l \left(\frac{2\kappa^{-l/2} - \kappa^{-(l-1)/2} - 1}{1 - \sqrt{\kappa}} \right) \mathcal{N}(w_l)\mathcal{N}(2a^{n_l-1} - a^{n_l})\epsilon^{c_l}$$
$$+ \frac{\epsilon}{\kappa}\mathcal{N}(w_l)\mathcal{N}(2a^{n_l-1} - a^{n_l})$$
$$+ \frac{2^{1-c}}{\sqrt{\kappa}} \frac{\mathcal{N}(w_l)}{\mathcal{N}(a^{n_l})^c}\mathcal{N}(a^{n_l-1})^{2c}\mathcal{N}(a^{n_l-1-k})^{1-c}\epsilon^{c-\alpha}$$

We must now simplify this inequality.

First, $2a^{n_l-1} - a^{n_l} = ca^{n_l-1} + (1-c)a^{n_l-1-k}$ so, from Lemma 3.20, $\mathcal{N}(2a^{n_l-1} - a^{n_l}) \leq \mathcal{N}(a^{n_l-1})^c\mathcal{N}(a^{n_l-1-k})^{1-c}$. So:

$$\left|\overline{F^{(1)}(x+iw_l)}\ F^{(1)}(x + i(2a^{n_l-1} - a^{n_l})) - \overline{F^{(2)}(x+iw_l)}F^{(2)}(x + i(2a^{n_l-1} - a^{n_l}))\right|$$
$$\leq \mathcal{N}(w_l)\mathcal{N}(a^{n_l-1})^c\mathcal{N}(a^{n_l-1-k})^{1-c}$$
$$\times \left(\frac{1}{\sqrt{\kappa}} 3D_l \left(\frac{2\kappa^{-l/2} - \kappa^{-(l-1)/2} - 1}{1 - \sqrt{\kappa}} \right) \epsilon^{c_l} + \frac{\epsilon}{\kappa} + \frac{2^{1-c}}{\sqrt{\kappa}} \frac{\mathcal{N}(a^{n_l-1})^c}{\mathcal{N}(a^{n_l})^c}\epsilon^{c-\alpha} \right)$$

Now we note that $1 \leq \frac{\mathcal{N}(a^{n_l-1})^c}{\mathcal{N}(a^{n_l})^c}$ (from Lemma 3.20 again, because $a^{n_l-1} \leq a^{n_l}$). Because $\kappa \geq \epsilon^{2(1-c)}$, we also have $\frac{\epsilon}{\kappa} \leq \frac{\epsilon^c}{\sqrt{\kappa}} \leq \frac{\epsilon^{c-\alpha}}{\sqrt{\kappa}}$. And as $c - \alpha \geq c_l$, $\epsilon^{c-\alpha} \leq \epsilon^{c_l}$. This gives:

$$\left|\overline{F^{(1)}(x+iw_l)}\ F^{(1)}(x + i(2a^{n_l-1} - a^{n_l})) - \overline{F^{(2)}(x+iw_l)}F^{(2)}(x + i(2a^{n_l-1} - a^{n_l}))\right|$$
$$\leq \frac{\epsilon^{c_l}}{\sqrt{\kappa}}\mathcal{N}(w_l)\frac{\mathcal{N}(a^{n_l-1})^{2c}\mathcal{N}(a^{n_l-1-k})^{1-c}}{\mathcal{N}(a^{n_l})^c} \left(3D_l \left(\frac{2\kappa^{-l/2} - \kappa^{-(l-1)/2} - 1}{1 - \sqrt{\kappa}} \right) + 1 + 2^{1-c} \right)$$

If we bound 2^{1-c} by 2 and notice that $D_l \geq 1$ (because, from 3.20, it is a product of terms bigger that 1), we have:

$$\left|\overline{F^{(1)}(x+iw_l)}\ F^{(1)}(x + i(2a^{n_l-1} - a^{n_l})) - \overline{F^{(2)}(x+iw_l)}F^{(2)}(x + i(2a^{n_l-1} - a^{n_l}))\right|$$
$$\leq \frac{\epsilon^{c_l}}{\sqrt{\kappa}} 3D_l\mathcal{N}(w_l)\frac{\mathcal{N}(a^{n_l-1})^{2c}\mathcal{N}(a^{n_l-1-k})^{1-c}}{\mathcal{N}(a^{n_l})^c} \left(\left(\frac{2\kappa^{-l/2} - \kappa^{-(l-1)/2} - 1}{1 - \sqrt{\kappa}} \right) + 1 \right)$$
$$\leq 3\epsilon^{c_l} D_l\mathcal{N}(w_l)\frac{\mathcal{N}(a^{n_l-1})^{2c}\mathcal{N}(a^{n_l-1-k})^{1-c}}{\mathcal{N}(a^{n_l})^c} \left(\frac{2\kappa^{-(l+1)/2} - \kappa^{-l/2} - 1}{1 - \sqrt{\kappa}} \right)$$

Finally, from 3.20, we have $\mathcal{N}(a^{n_l-1}) \leq \mathcal{N}(a^{n_l})^{1/2}\mathcal{N}(2a^{n_l-1} - a^{n_l})^{1/2}$ so:

$$\left|\overline{F^{(1)}(x+iw_l)}\ F^{(1)}(x + i(2a^{n_l-1} - a^{n_l})) - \overline{F^{(2)}(x+iw_l)}F^{(2)}(x + i(2a^{n_l-1} - a^{n_l}))\right|$$
$$\leq 3\epsilon^{c_l} D_l\mathcal{N}(w_l)\mathcal{N}(2a^{n_l-1} - a^{n_l}) \left(\frac{\mathcal{N}(a^{n_l-1-k})}{\mathcal{N}(2a^{n_l-1} - a^{n_l})} \right)^{1-c} \left(\frac{2\kappa^{-(l+1)/2} - \kappa^{-l/2} - 1}{1 - \sqrt{\kappa}} \right)$$

Second part: we define, for any $z \in \mathbb{C}$ such that $-2a^{n_l-1} + a^{n_l} < \operatorname{Im} z < w_l$:

$$H(z) = \overline{F^{(1)}(\overline{z} + iw_l)}F^{(1)}(z + i(2a^{n_l-1} - a^{n_l})) - \overline{F^{(2)}(\overline{z} + iw_l)}F^{(2)}(z + i(2a^{n_l-1} - a^{n_l}))$$

We write:

$$B = \frac{3}{2}D_l \left(\frac{\mathcal{N}(a^{n_l-1-k})}{\mathcal{N}(2a^{n_l-1} - a^{n_l})}\right)^{1-c} \left(\frac{2\kappa^{-(l+1)/2} - \kappa^{-l/2} - 1}{1 - \sqrt{\kappa}}\right)$$

From the first part:

$$\begin{aligned}|H(x+iy)| &\leq 2\mathcal{N}(w_l)\mathcal{N}(2a^{n_l-1} - a^{n_l})B\epsilon^{c_l} \\ &\qquad \text{if } y = 0, x \in [-(M-(l+1)\mu)a^{j+K}; (M-(l+1)\mu)a^{j+K}] \\ &\leq 2\mathcal{N}(w_l)\mathcal{N}(2a^{n_l-1} - a^{n_l}) \\ &\qquad \text{if } y = 0, x \notin [-(M-(l+1)\mu)a^{j+K}; (M-(l+1)\mu)a^{j+K}]\end{aligned}$$

Moreover, if we set $y_l = w_l - 2a^{n_l-1} + a^{n_l}$:

$$\begin{aligned}H(x+iy_l) &= \overline{F^{(1)}(x + i(2a^{n_l-1} - a^{n_l}))}F^{(1)}(x + iw_l) - \overline{F^{(2)}(x + i(2a^{n_l-1} - a^{n_l}))}F^{(2)}(x + iw_l) \\ &= \overline{H(x)}\end{aligned}$$

Thus, we also have:

$$\begin{aligned}|H(x+iy)| &\leq 2\mathcal{N}(w_l)\mathcal{N}(2a^{n_l-1} - a^{n_l})B\epsilon^{c_l} \\ &\qquad \text{if } y = y_l, x \in [-(M-(l+1)\mu)a^{j+K}; (M-(l+1)\mu)a^{j+K}] \\ &\leq 2\mathcal{N}(w_l)\mathcal{N}(2a^{n_l-1} - a^{n_l}) \\ &\qquad \text{if } y = y_l, x \notin [-(M-(l+1)\mu)a^{j+K}; (M-(l+1)\mu)a^{j+K}]\end{aligned}$$

We apply Lemma 3.26 with $a = 0, b = y_l$. For $\operatorname{Im} z = (a-1)^2 a^{n_l-2}$ and $|\operatorname{Re} z| \leq (M-(l+1)\mu)a^{j+K}$:

$$|H(z)| \leq 2\mathcal{N}(w_l)\mathcal{N}(2a^{n_l-1} - a^{n_l})(B\epsilon^{c_l})^{f(z)} \tag{3.51}$$

with $f(z) \geq 1 - 4\frac{(a-1)^2 a^{n_l-2}}{y_l}\left(\frac{e^{-\pi\frac{(M-(l+1)\mu)a^{j+K}-|\operatorname{Re} z|}{y_l}}}{1-e^{-\pi\frac{(M-(l+1)\mu)a^{j+K}-|\operatorname{Re} z|}{y_l}}}\right)$.

From the definition of w_l, one may check that (w_l) is a decreasing sequence which converges to $\frac{2a^{j+K}}{a+1}$ when l goes to ∞. So, for any $l \geq 0$:

$$y_l \leq w_l \leq w_0 = a^{j+K}$$

$$y_l \geq \frac{2a^{j+K}}{a+1} - 2a^{n_l-1} + a^{n_l} \geq \frac{2a^{j+K}}{a+1} - 2a^{j+K-1} + a^{j+K} = \frac{(a-1)(a+2)}{(a+1)} a^{j+K-1}$$

From this we deduce:

$$f(z) \geq 1 - 4\,\frac{a^2-1}{a+2} a^{-2l-1} \left(\frac{e^{-\pi \frac{(M-(l+1)\mu)a^{j+K}-|\mathrm{Re}\ z|}{a^{j+K}}}}{1-e^{-\pi \frac{(M-(l+1)\mu)a^{j+K}-|\mathrm{Re}\ z|}{a^{j+K}}}} \right)$$

So, when $|\mathrm{Re}\, z| \leq (M-(l+2)\mu)a^{j+K}$, $f(z) \geq 1 - 4\,\frac{a^2-1}{a+2} a^{-2l-1} \left(\frac{e^{-\pi\mu}}{1-e^{-\pi\mu}}\right)$.

As $B \geq 1$ and $f(z) \leq 1$, $B^{f(z)} \leq B$. Moreover, $c_l \leq 1$ so $c_l f(z) \geq c_l - (1-f(z))$ if $1-f(z) \geq 0$. Equation (3.51) thus gives:

$$\begin{aligned}
|H(z)| &\leq 2\mathcal{N}(w_l)\mathcal{N}(2a^{n_l-1} - a^{n_l}) B \epsilon^{c_l - 4\,\frac{a^2-1}{a+2} a^{-2l-1}\left(\frac{e^{-\pi\mu}}{1-e^{-\pi\mu}}\right)} \\
&= 2\mathcal{N}(w_l)\mathcal{N}(2a^{n_l-1} - a^{n_l}) B \epsilon^{c_{l+1}} \\
&= 3D_l \mathcal{N}(w_l)\mathcal{N}(2a^{n_l-1} - a^{n_l})^c \mathcal{N}(a^{n_l-1-k})^{1-c} \left(\frac{2\kappa^{-(l+1)/2} - \kappa^{-l/2} - 1}{1-\sqrt{\kappa}} \right) \epsilon^{c_{l+1}}
\end{aligned}$$

Because $w_l \geq w_{l+1}$ and $2a^{n_l-1} - a^{n_l} \geq a^{n_l-1-k}$, we have $\mathcal{N}(w_l) \leq \mathcal{N}(w_{l+1})$ and $\mathcal{N}(2a^{n_l-1} - a^{n_l}) \leq \mathcal{N}(a^{n_l-1-k})$. Thus:

$$\begin{aligned}
|H(z)| &\leq 3D_l \mathcal{N}(w_{l+1})\mathcal{N}(a^{n_l-2}) \frac{\mathcal{N}(a^{n_l-1-k})}{\mathcal{N}(a^{n_l-2})} \left(\frac{2\kappa^{-(l+1)/2} - \kappa^{-l/2} - 1}{1-\sqrt{\kappa}} \right) \epsilon^{c_{l+1}} \\
&= 3D_{l+1} \mathcal{N}(w_{l+1})\mathcal{N}(a^{n_l-2}) \left(\frac{2\kappa^{-(l+1)/2} - \kappa^{-l/2} - 1}{1-\sqrt{\kappa}} \right) \epsilon^{c_{l+1}}
\end{aligned}$$

So, for any $x \in [-(M-(l+2)\mu)a^{j+K}; (M-(l+2)\mu)a^{j+K}]$:

$$\begin{aligned}
&\frac{1}{\mathcal{N}(w_{l+1})\mathcal{N}(a^{n_{l+1}})} \left| \overline{F^{(1)}(x+iw_{l+1})} F^{(1)}(x+ia^{n_{l+1}}) - \overline{F^{(2)}(x+iw_{l+1})} F^{(2)}(x+ia^{n_{l+1}}) \right| \\
&\qquad = |H(x+i(a-1)^2 a^{n_l-2})| \\
&\qquad \leq 3D_{l+1} \left(\frac{2\kappa^{-(l+1)/2} - \kappa^{-l/2} - 1}{1-\sqrt{\kappa}} \right) \epsilon^{c_{l+1}}
\end{aligned}$$

This is exactly the induction hypothesis at the order $l+1$. □

Proof of Theorem 3.18. We will obtain the desired theorem as a corollary of the previous one (3.23).

The conditions (3.45) and (3.46) in the statement of Theorem 3.23 are equivalent to (3.32) and (3.33), required in Theorem 3.18.

Thus, if we fix $\mu \in [0; M[$, we have that, for any $l \geq 0$ such that $n_l \geq j$ and $M-(l+1)\mu > 0$, under the condition that $c_l < 1$:

$$\frac{1}{\mathcal{N}(w_l)\mathcal{N}(a^{n_l})}\left|\overline{F^{(1)}(x+iw_l)}F^{(1)}(x+ia^{n_l}) - \overline{F^{(2)}(x+iw_l)}F^{(2)}(x+ia^{n_l})\right| \leq 3D_l\left(\frac{2\kappa^{-l/2}-\kappa^{-(l-1)/2}-1}{1-\sqrt{\kappa}}\right)\epsilon^{c_l}$$
$$\left(\forall x \in [-(M-(l+1)\mu)a^{j+K}; (M-(l+1)\mu)a^{j+K}]\right)$$

where the constants are defined as in 3.23.

We can check that, for any l, $w_l = \frac{a^{j+K}}{a+1}\left(2+(a-1)a^{-2l}\right)$.

We take $l = K/2$. We then have $w_l = \frac{2}{a+1}a^{j+K} + \frac{a-1}{a+1}a^j = a^J$ and $n_l = j$. For this l, the previous inequality is equivalent to:

$$\frac{1}{N_J N_j}\left|\overline{f^{(1)}\star\psi_J(x)}f^{(1)}\star\psi_j(x) - \overline{f^{(2)}\star\psi_J(x)}\ f^{(2)}\star\psi_j(x)\right| \leq 3D_l\left(\frac{2\kappa^{-K/4}-\kappa^{-(K-2)/4}-1}{1-\sqrt{\kappa}}\right)\epsilon^{c_l}$$

We observe that $c_l \geq \lim_{l\to\infty} c_l = c - 2\left(1+2\frac{a}{a+2}\right)\left(\frac{e^{-\pi\mu}}{1-e^{-\pi\mu}}\right) \geq c - 4\left(\frac{e^{-\pi\mu}}{1-e^{-\pi\mu}}\right)$ and $\frac{2\kappa^{-K/4}-\kappa^{-(K-2)/4}-1}{1-\sqrt{\kappa}} \leq \frac{2\kappa^{-K/4}}{1-\sqrt{\kappa}}$.

So, for any $x \in [-(M-\mu(1+K/2))a^{j+K}; (M-\mu(1+K/2))a^{j+K}]$:

$$\frac{1}{N_J N_j}\left|\overline{f^{(1)}\star\psi_J(x)}f^{(1)}\star\psi_j(x) - \overline{f^{(2)}\star\psi_J(x)}\ f^{(2)}\star\psi_j(x)\right| \leq 6D_l\frac{\kappa^{-K/4}}{1-\sqrt{\kappa}}\epsilon^{c-4\left(\frac{e^{-\pi\mu}}{1-e^{-\pi\mu}}\right)}$$

From Equation (3.31):

$$D_l = \prod_{s=0}^{K/2-1}\left(\frac{\mathcal{N}(a^{n_s-1-k})}{\mathcal{N}(a^{n_s-2})}\right) = \prod_{s=0}^{K/2-1}\left(a^{p(k-1)}\frac{N_{n_s-1-k}}{N_{n_s-2}}\right)$$

For $\mu = \frac{M}{K+2}$, our last inequality is exactly the desired result. □

3.6.5 Bounds for holomorphic functions

In the proofs of the section 3.4, we often have to consider holomorphic functions defined on a band of the complex plane. We want to obtain information about their values inside the band from their values on the boundary of the band. This is the purpose of the three theorems contained in this section.

In the whole section, a, b are fixed real numbers such that $a < b$. We write $B_{a,b} = \{z \in \mathbb{C} \text{ s.t. } a < \operatorname{Im} z < b\}$. We consider a holomorphic function $W : B_{a,b} \to \mathbb{C}$ which satisfies the following properties:

(i) W is bounded on $B_{a,b}$.

(ii) W admits a continuous extension over $\overline{B}_{a,b}$, which we still denote by W.

The first theorem we need is a well-known fact. We recall its proof because it is very short and relies on the same idea that will also be used in the other proofs.

Theorem 3.24. *We suppose that, for some $A, B > 0$:*

$$|W(z)| \leq A \text{ if } \operatorname{Im} z = a$$
$$|W(z)| \leq B \text{ if } \operatorname{Im} z = b$$

Then, for all $t \in]0;1[$ and all $z \in \mathbb{C}$ such that $\operatorname{Im} z = (1-t)a + tb$:

$$|W(z)| \leq A^{1-t} B^t$$

Proof. For every $\epsilon > 0$ and $z \in \overline{B}_{a,b}$:

$$L(z) = \log(|W(z)|) - \frac{(b - \operatorname{Im} z)\log(A) + (\operatorname{Im} z - a)\log(B)}{b-a} - \epsilon \log|z + i(1-a)|$$

is subharmonic on $B_{a,b}$ and continuous on $\overline{B}_{a,b}$. It is upper-bounded and takes negative values on $\partial B_{a,b}$. Moreover, $L(z) \to -\infty$ when $\operatorname{Re}(z) \to \pm\infty$. From the maximum principle, this function must be negative on $B_{a,b}$.

Letting ϵ go to 0 implies:

$$\log(|W(z)|) \leq \frac{(b - \operatorname{Im} z)\log(A) + (\operatorname{Im} z - a)\log(B)}{b-a} \qquad \forall z \in \overline{B}_{a,b}$$
$$\Rightarrow \quad |W(z)| \leq A^{\frac{b - \operatorname{Im} z}{b-a}} B^{\frac{\operatorname{Im} z - a}{b-a}}$$

□

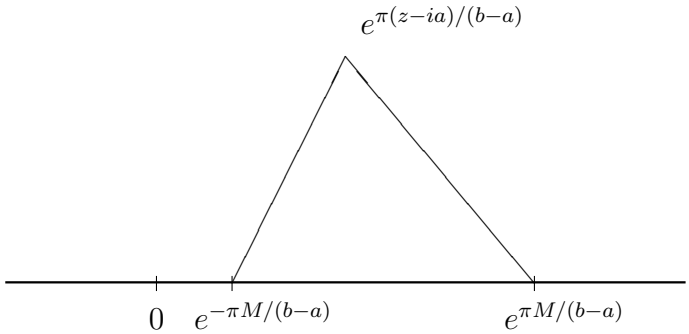


Figure 3.7: Positions of the points used in the definition of f

Lemma 3.25. *Let $A, B, \epsilon > 0$ be fixed real numbers, with $\epsilon \leq 1$. We assume that:*

$$|W(z)| \leq B \text{ if } \operatorname{Im} z = b$$
$$|W(z)| \leq A \text{ if } \operatorname{Im} z = a \text{ and } \operatorname{Re} z \notin [-M; M]$$
$$|W(z)| \leq \epsilon A \text{ if } \operatorname{Im} z = a \text{ and } \operatorname{Re} z \in [-M; M]$$

Then, for all z such that $a < \operatorname{Im} z < b$, if $t \in [0; 1]$ is such that $\operatorname{Im} z = (1-t)a + tb$:

$$|W(z)| \leq \epsilon^{f(z)} A^{1-t} B^t$$

where:

$$f(z) = \frac{1}{\pi} \arg\left(\frac{e^{\pi M/(b-a)} - e^{\pi(z-ia)/(b-a)}}{e^{-\pi M/(b-a)} - e^{\pi(z-ia)/(b-a)}} \right)$$

and this function satisfies, when $|\operatorname{Re} z| \leq M$: $f(z) \geq (1-t) - 2t \frac{e^{-\pi \frac{M - |\operatorname{Re}\ z|}{b-a}}}{1 - e^{-\pi \frac{M - |\operatorname{Re}\ z|}{b-a}}}$.

Proof. The function f may be continuously extended to $\overline{B}_{a,b} - \{-M + ia; M + ia\}$. By looking at the figure 3.7, one sees that:

$$\begin{aligned} f(x + ia) &= 0 \text{ for all } x \in \mathbb{R} - [-M; M] \\ &= 1 \text{ for all } x \in]-M; M[\\ f(x + ib) &= 0 \text{ for all } x \in \mathbb{R} \end{aligned}$$

We set:

$$f(-M + ia) = f(M + ia) = 1$$

This definition makes the extension of f upper semi-continuous on $\overline{B}_{a,b}$ (because $f \leq 1$ on all $B_{a,b}$).

For any $\eta > 0$, the following function is subharmonic on $B_{a,b}$:

$$L(z) = \log(|W(z)|) - \log(\epsilon) f(z) - \frac{(b - \operatorname{Im} z)\log(A) + (\operatorname{Im} z - a)\log(B)}{b-a} - \eta \log|z + i(1-a)|$$

It is upper semi-continuous on $\overline{B}_{a,b}$ and tends to $-\infty$ when $\operatorname{Re} z \to \pm\infty$. Thus, this function admits a local maximum over $\overline{B}_{a,b}$. This maximum is attained on $\partial B_{a,b}$, because L is subharmonic.

From the hypotheses, one can check that $L(z) \leq 0$ for all $z \in \partial B_{a,b}$. The function L is thus negative on the whole band $\overline{B}_{a,b}$. Letting η go to zero gives, for all $z \in B_{a,b}$ such that $\operatorname{Im} z = (1-t)a + tb$:

$$|W(z)| \leq \epsilon^{f(z)} A^{1-t} B^t$$

We are only left to show that $f(z) \geq (1-t) - 2t\frac{e^{-\pi\frac{M-|\operatorname{Re} z|}{b-a}}}{1-e^{-\pi\frac{M-|\operatorname{Re} z|}{b-a}}}$ when $\operatorname{Im} z = (1-t)a + tb$.

If we write $x = \operatorname{Re}(z)$, we have:

$$\begin{aligned} f(z) &= \frac{1}{\pi}\arg\left(-e^{-i\pi t}\frac{1 - e^{\pi(x-M)/(b-a)}e^{\pi i t})}{1 - e^{-\pi(M+x)/(b-a)}e^{-\pi i t}}\right) \\ &= (1-t) + \frac{1}{\pi}\arg\left(\frac{1 - e^{\pi(x-M)/(b-a)}e^{\pi i t})}{1 - e^{-\pi(M+x)/(b-a)}e^{-\pi i t}}\right) \end{aligned}$$

We note that:

$$\begin{aligned} \left|\arg\left(1 - e^{\pi\frac{x-M}{b-a}}e^{i\pi t}\right)\right| &\leq \left|\tan\left(1 - e^{\pi\frac{x-M}{b-a}}e^{i\pi t}\right)\right| \\ &= |\sin(\pi t)|\,\frac{e^{\pi\frac{x-M}{b-a}}}{1 - e^{\pi\frac{x-M}{b-a}}\cos(\pi t)} \\ &\leq |\sin(\pi t)|\,\frac{e^{\pi\frac{x-M}{b-a}}}{1 - e^{\pi\frac{x-M}{b-a}}} \\ &\leq \pi t\frac{e^{\pi\frac{x-M}{b-a}}}{1 - e^{\pi\frac{x-M}{b-a}}} \leq \pi t\frac{e^{-\pi\frac{M-|\operatorname{Re} z|}{b-a}}}{1 - e^{-\pi\frac{M-|\operatorname{Re} z|}{b-a}}} \end{aligned}$$

And the same inequality holds for $\left|\arg\left(1 - e^{-\pi\frac{M+x}{b-a}}e^{-i\pi t}\right)\right|$. This implies the result.

□

The proof of the third result is similar to the proof of the second one. We do not reproduce it.

Lemma 3.26. *Let $M, A, \epsilon > 0$ be fixed real numbers, with $\epsilon \leq 1$. We assume that:*

$$\begin{array}{lll} |W(x+ia)| \leq A & |W(x+ib)| \leq A & \forall x \in \mathbb{R} - [-M; M] \\ |W(x+ia)| \leq \epsilon A & |W(x+ib)| \leq \epsilon A & \forall x \in [-M; M] \end{array}$$

Then, for all z such that $a < \operatorname{Im} z < b$:

$$|W(z)| \leq \epsilon^{f(z)} A$$

where:

$$f(z) = \frac{1}{\pi} \arg \left(\frac{e^{\pi M/(b-a)} - e^{\pi(z-ia)/(b-a)}}{e^{-\pi M/(b-a)} - e^{\pi(z-ia)/(b-a)}} . \frac{-e^{-\pi M/(b-a)} - e^{\pi(z-ia)/(b-a)}}{-e^{\pi M/(b-a)} - e^{\pi(z-ia)/(b-a)}} \right)$$

and this function satisfies, when $|\operatorname{Re} z| \leq M$: $f(z) \geq 1 - 4t \left(\frac{e^{-\pi \frac{M - |\operatorname{Re}\ z|}{b-a}}}{1 - e^{-\pi \frac{M - |\operatorname{Re}\ z|}{b-a}}} \right)$, for $t = \frac{\operatorname{Im}\ z-a}{b-a}$.

Chapter 4

Phase retrieval for wavelet transforms: a non-convex algorithm

In Chapter 3, we proved that, at least for a specific choice of wavelets, the phase retrieval problem for the wavelet transform is well-posed: any function is uniquely determined by the modulus of its wavelet transform (sometimes called *scalogram* in the context of audio signals), and the reconstruction has a form of stability to noise. In this chapter, we propose an algorithm to numerically solve the problem.

Algorithms used to reconstruct audio signals from their scalogram (or spectrogram, which is similar) are divided into the same two classes as generic phase retrieval problems: iterative and convexified methods.

Iterative algorithms have been introduced for the spectrogram in [Griffin and Lim, 1984]. While simple and relatively fast, they tend to produce distinct auditive artifacts. Improvements have been achieved in particular in [Bouvrie and Ezzat, 2006] (by applying the algorithm to small temporal windows, and not to the whole signal at once) and in [Achan et al., 2004; Eldar et al., 2015] (in the case where additional information is available about the signal). The Douglas-Rachford method [Fienup, 1982; Bauschke et al., 2002] can also yield significant improvements; however, from our experiments, it does not remove all artifacts[1].

Methods by convexification seem to perform very well on small signals [Sun and Smith [2012] or Paragraph 2.4.3 of this thesis]. Nevertheless, their high complexity prevents them to be used on real audio signals.

In this chapter, we propose a new iterative algorithm, combining the advantages of both families. Its complexity is roughly linear in the signal size, up to logarithmic factors, so that it can be used on real-size problems. However, the quality of reconstructed signals is as good as

[1] A few examples are available at `http://www.di.ens.fr/~waldspurger/wavelets_phase_retrieval.html`.

with a convexified method.

This new algorithm is multiscale: it reconstructs the signal frequency band by frequency band, starting with the low frequencies. Its main tool is a reformulation of the phase retrieval problem, which extends to more general wavelets the reconstruction algorithm presented in the previous chapter (Section 3.5) for the case of Cauchy wavelets. This reformulation has two advantages. First, it gives a simple method to propagate the phase information reconstructed in low frequency bands towards higher frequency bands. Second, it naturally yields a local optimization algorithm to refine approximate solutions; this local optimization method, although non-convex, seems to be robust to the problem of local minima.

In Section 4.1, we describe our reformulation and prove its equivalence with the original problem. We explain its advantages. In Section 4.2, we describe the resulting algorithm, including a new multigrid error correction step. In Section 4.3, we discuss the superiority of multiscale algorithms over non-multiscale ones. Finally, Section 4.4 is devoted to numerical results. It shows that our algorithm is both precise and stable to noise. Moreover, it uses the algorithm to numerically investigate the intrinsic stability of the phase retrieval problem, in line with the theoretical results of Chapter 3.

Definitions and assumptions

All signals $f[n]$ are of finite length N. Their discrete Fourier transform is defined by:

$$\hat{f}[k] = \sum_{s=0}^{N-1} f[n] e^{-2\pi i \frac{kn}{N}} \qquad k = 0, ..., N-1$$

and the convolution always refers to the circular convolution.

We define a family of wavelets $(\psi_j)_{0 \leq j \leq J}$ by:

$$\hat{\psi}_j[k] = \hat{\psi}(a^j k) \qquad k = 0, ..., N-1$$

where the *dilation factor* a can be any number in $(1; +\infty)$ and $\psi : \mathbb{R} \to \mathbb{C}$ is a fixed *mother wavelet*. We assume that J is sufficiently large so that $\hat{\psi}_J$ is negligible outside a small set of points. An example is shown in Figure 4.1.

The *wavelet transform* is defined by:

$$\forall f \in \mathbb{R}^N, \quad Wf = \{f \star \psi_j\}_{0 \leq j \leq J}$$

The problem we consider here consists in reconstructing functions from the modulus of their wavelet transform:

$$\text{Reconstruct } f \text{ from } \{|f \star \psi_j|\}_{0 \leq j \leq J}$$

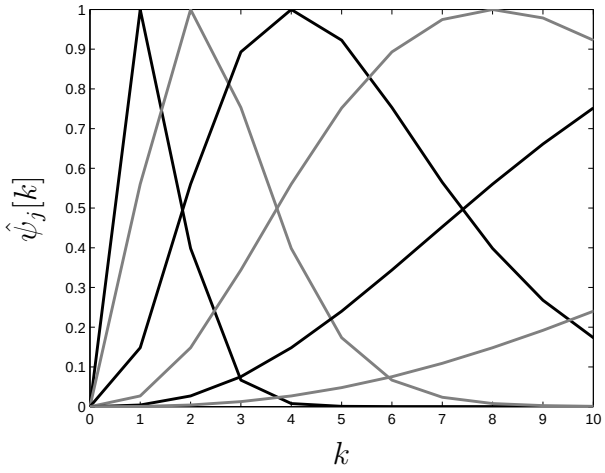


Figure 4.1: Example of wavelets; the figure shows ψ_j for $J-5 \leq j \leq J$ in the Fourier domain. Only the real part is displayed.

Multiplying a function by a unitary complex does not change the modulus of its wavelet transform, so we only aim at reconstructing functions up to multiplication by a unitary complex, that is *up to a global phase*.

All signals are assumed to be analytic:

$$\hat{f}[k] = 0 \quad \text{when } N/2 < k \leq N-1 \tag{4.1}$$

Equivalently, we could assume the signals to be real but set the $\hat{\psi}_j[k]$ to zero for $N/2 < k \leq N-1$.

4.1 Reformulation of the phase retrieval problem

In the first part of this section, we reformulate the phase retrieval problem for the wavelet transform, by introducing two auxiliary wavelet families.

We then describe the two main advantages of this reformulation. First, it allows to propagate the phase information from the low-frequencies to the high ones, and so enables us to perform the reconstruction scale by scale. Second, from this reformulation, we can define a natural objective function to locally optimize approximate solutions. Although non-convex, this function has few local minima; hence, the local optimization algorithm is efficient.

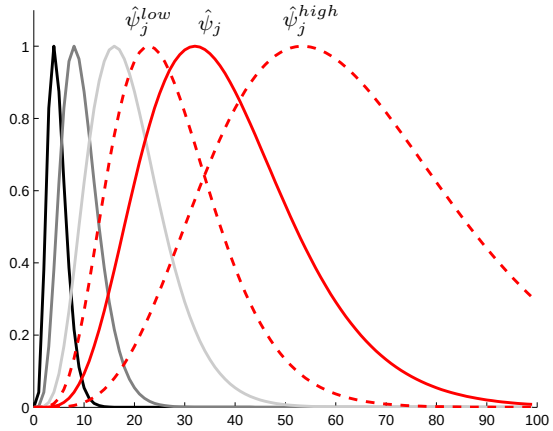


Figure 4.2: $\psi_J, ..., \psi_{j+1}, \psi_j$ (in the Fourier domain), along with ψ_j^{low} and ψ_j^{high} (dashed lines)

4.1.1 Introduction of auxiliary wavelets and reformulation

Let us fix $r \in]0;1[$ and define:

$$\forall k = 0, ..., N-1, \qquad \hat{\psi}_j^{low}[k] = \hat{\psi}_j[k] r^k$$
$$\hat{\psi}_j^{high}[k] = \hat{\psi}_j[k] r^{-k}$$

This definition is illustrated by Figure 4.2. The wavelet $\hat{\psi}_j^{low}$ has a lower characteristic frequency than $\hat{\psi}_j$ and $\hat{\psi}_j^{high}$ a higher one. The following theorem explains how to rewrite a condition on the modulus of $f \star \psi_j$ as a condition on $f \star \psi_j^{low}$ and $f \star \psi_j^{high}$.

Theorem 4.1. *Let $j \in \{0, ..., J\}$ and $g_j \in (\mathbb{R}^+)^N$ be fixed. Let Q_j be the function whose Fourier transform is:*

$$\hat{Q}_j[k] = r^k \widehat{g_j^2}[k] \tag{4.2}$$
$$\forall k = \left\lfloor \frac{N}{2} \right\rfloor - N + 1, ..., \left\lfloor \frac{N}{2} \right\rfloor$$

For any $f \in \mathbb{C}^N$ satisfying the analycity condition (4.1), the following two properties are equivalent:

1. $|f \star \psi_j| = g_j$
2. $(f \star \psi_j^{low})\overline{(f \star \psi_j^{high})} = Q_j$

Proof. The proof consists in showing that the second inequality is the analytic extension of the first one, in a sense that will be precisely defined.

For any function $h : \{0, ..., N-1\} \to \mathbb{C}$, let $P(h)$ be:

$$\forall z \in \mathbb{C} \qquad P(h)(z) = \sum_{k=\lfloor \frac{N}{2} \rfloor - N + 1}^{\lfloor \frac{N}{2} \rfloor} \hat{h}[k] z^k$$

Up to a change of coordinates, $P(|f \star \psi_j|^2)$ and $P(g_j^2)$ are equal to $P((f \star \psi_j^{low})\overline{(f \star \psi_j^{high})})$ and $P(Q_j)$:

Lemma 4.2. *For any f satisfying the analycity condition (4.1):*

$$\begin{aligned} \forall z \in \mathbb{C} \qquad & P(|f \star \psi_j|^2)(rz) = P((f \star \psi_j^{low})\overline{(f \star \psi_j^{high})})(z) \\ \textit{and} \quad & P(g_j^2)(rz) = P(Q_j)(z) \end{aligned}$$

This lemma is proved in the appendix 4.5.2. It implies the result because then:

$$\begin{aligned} |f \star \psi_j| = g_j &\iff |f \star \psi_j|^2 = g_j^2 \\ &\iff \forall z, P(|f \star \psi_j|^2)(z) = P(g_j^2)(z) \\ &\iff \forall z, P(|f \star \psi_j|^2)(rz) = P(g_j^2)(rz) \\ &\iff \forall z, P((f \star \psi_j^{low})\overline{(f \star \psi_j^{high})})(z) = P(Q_j)(z) \\ &\iff (f \star \psi_j^{low})\overline{(f \star \psi_j^{high})} = Q_j \end{aligned}$$

□

By applying simultaneously Theorem 4.1 to all indexes j, we can reformulate the phase retrieval problem $|f \star \psi_j| = g_j, \forall j$ in terms of the $f \star \psi_j^{low}$'s and $f \star \psi_j^{high}$'s.

Corollary 4.3 (Reformulation of the phase retrieval problem)**.** *Let $(g_j)_{0 \le j \le J}$ be a family of signals in $(\mathbb{R}^+)^N$. For each j, let Q_j be defined as in (4.2). Then the following two problems are equivalent:*

Find f satisfying (4.1) such that:

$$\forall j, \quad |f \star \psi_j| = g_j$$

$$\iff$$

Find f satisfying (4.1) such that:

$$\forall j, \quad (f \star \psi_j^{low})\overline{(f \star \psi_j^{high})} = Q_j \quad (4.3)$$

4.1.2 Phase propagation across scales

This new formulation yields a natural multiscale reconstruction algorithm, in which one reconstructs f frequency band by frequency band, starting from the low frequencies.

Indeed, once $f \star \psi_J, ..., f \star \psi_{j+1}$ have been reconstructed, it is possible to estimate $f \star \psi_j^{low}$ by deconvolution. This deconvolution is stable to noise because, if r is sufficiently small, then the frequency band covered by ψ_j^{low} is almost included in the frequency range covered by $\psi_J, ..., \psi_{j+1}$ (see figure 4.2). From $f \star \psi_j^{low}$, one can reconstruct $f \star \psi_j^{high}$, using (4.3):

$$f \star \psi_j^{high} = \frac{\overline{Q}_j}{\overline{f \star \psi_j^{low}}} \tag{4.4}$$

Finally, one reconstructs $f \star \psi_j$ from $f \star \psi_j^{high}$ and $f \star \psi_j^{low}$.

The classical formulation of the phase retrieval problem does not allow the conception of such a multiscale algorithm. Indeed, from $f \star \psi_J, ..., f \star \psi_{j+1}$, it is not possible to directly estimate $f \star \psi_j$: it would require performing a highly unstable deconvolution. The introduction of the two auxiliary wavelet families is essential.

4.1.3 Local optimization of approximate solutions

From the reformulation (4.3), we can define a natural objective function for the local optimization of approximate solutions to the phase retrieval problem. This is also possible from the classical formulation but the objective function then has numerous local minima, which make it difficult to globally minimize. Empirically, the objective function associated to the reformulation suffers dramatically less from this drawback.

The objective function has $2J + 3$ variables: $(h_j^{low})_{0 \leq j \leq J}$, $(h_j^{high})_{0 \leq j \leq J}$ and f. The intuition is that f is the signal we aim at reconstructing and the h_j^{low}, h_j^{high} correspond to the $f \star \psi_j^{low}$'s and $f \star \psi_j^{high}$'s. The objective function is:

$$\begin{aligned} &\text{obj}(h_J^{low}, ..., h_0^{low}, h_J^{high}, ..., h_0^{high}, f) \\ = &\sum_{j=0}^{J} ||h_j^{low} \overline{h_j^{high}} - Q_j||_2^2 \\ &+ \lambda \sum_{j=0}^{J} \left(||f \star \psi_j^{low} - h_j^{low}||_2^2 + ||f \star \psi_j^{high} - h_j^{high}||_2^2 \right) \end{aligned} \tag{4.5}$$

We additionally constrain the variables $(h_j^{low})_{0 \leq j \leq J}$ and $(h_j^{high})_{0 \leq j \leq J}$ to satisfy:

$$\forall j = 0, ..., J-1 \qquad h_j^{low} \star \psi_{j+1}^{high} = h_{j+1}^{high} \star \psi_j^{low} \tag{4.6}$$

The first term of the objective ensures that Equalities (4.3) are satisfied, while the second term and the additional constraint (4.6) enforce the fact that the h_j^{low}'s and h_j^{high}'s must be the wavelet transforms of the *same* function f.

The parameter λ is a positive real number. In our implementation, we choose a small λ, so that the first term dominates over the second one.

A similar objective function can also be derived directly from the classical formulation. However, empirically, it appears to have much more local minima than the function (4.5); hence, it is more difficult to efficiently minimize. A possible explanation is that the set of zeroes of the first term of (4.5) (which dominates the second one) has a smaller dimension when the reformulation is used, thus reducing the number of local minima it contains.

4.2 Description of the algorithm

In this section, we describe our implementation of the multiscale reconstruction algorithm introduced in Section 4.1. We explain the general organization in Paragraph 4.2.1. We then describe our exhaustive search method for solving phase retrieval problems of very small size (paragraph 4.2.2), which our algorithm uses to initialize the multiscale reconstruction. In Paragraph 4.2.3, we describe an additional multigrid correction step.

4.2.1 Organization of the algorithm

We start by reconstructing $f \star \psi_J$ from $|f \star \psi_J|$ and $|f \star \psi_{J-1}|$. We use an exhaustive search method, described in the next paragraph 4.2.2, which takes advantage of the fact that $\hat{\psi}_J$ and $\hat{\psi}_{J-1}$ have very small supports.

We then reconstruct the components of the wavelet transform scale by scale, as described in Section 4.1.

At each scale, we reconstruct $f \star \psi_j^{low}$ by propagating the phase information coming from $f \star \psi_J, ..., f \star \psi_{j+1}$ (as explained in Paragraph 4.1.2). This estimation can be imprecise, so we refine it by local optimization, using the objective function defined in Paragraph 4.1.3, from which we drop all the terms with higher scales than j. The local optimization algorithm we use in the implementation is L-BFGS ([Nocedal, 1980]), a low-memory approximation of a second order method.

We then reconstruct $f \star \psi_j^{high}$ by Equation (4.4).

At the end of the reconstruction, we run a few steps of the classical Gerchberg-Saxton algorithm to further refine the estimation.

The pseudo-code 5 summarizes the structure of the implementation.

Algorithm 5 overview of the algorithm

Input: $\{|f \star \psi_j|\}_{0 \leq j \leq J}$

Initialization: reconstruct $f \star \psi_J$ by exhaustive search
for all $j = J : (-1) : 0$ **do**
    Estimate $f \star \psi_j^{low}$ by phase propagation
    Refine the values of $f \star \psi_J^{low}, ..., f \star \psi_j^{low}, f \star \psi_J^{high}, ..., f \star \psi_{j+1}^{high}$ by local optimization
    Do an error correction step
    Refine again
    Compute $f \star \psi_j^{high}$ by $f \star \psi_h^{high} = \overline{Q}_j / \overline{f \star \psi_j^{low}}$
end for
Compute f
Refine f with Gerchberg-Saxton

Output: f

4.2.2 Reconstruction by exhaustive search for small problems

In this paragraph, we explain how to reconstruct $f \star \psi_j$ from $|f \star \psi_j|$ and $|f \star \psi_{j-1}|$ by exhaustive search, in the case where the support of $\hat{\psi}_j$ and $\hat{\psi}_{j-1}$ is small.

This is the method we use to initialize our multiscale algorithm. It is also useful for the multigrid error correction step described in the next paragraph 4.2.3.

Lemma 4.4. *Let $m \in \mathbb{R}^N$ and $K \in \mathbb{N}^*$ be fixed. We consider the problem:*

$$\begin{aligned} &\text{Find } g \in \mathbb{C}^N \text{ s.t. } |g| = m \\ &\qquad\qquad \text{and } Supp(\hat{g}) \subset \{1, ..., K\} \end{aligned}$$

This problem has at most 2^{K-1} solutions, up to a global phase, and there exist a simple algorithm which, from m and N, returns the list of all possible solutions.

Proof. This lemma is a consequence of classical results about the phase retrieval problem for the Fourier transform. It can for example be derived from [Hayes, 1982]. We give a proof in the appendix 4.5.1. □

We apply this lemma to $m = |f \star \psi_j|$ and $|f \star \psi_{j-1}|$, and construct the lists of all possible $f \star \psi_j$'s and of all possible $f \star \psi_{j-1}$'s. The true $f \star \psi_j$ and $f \star \psi_{j-1}$ are the only pair in these two lists which satisfy the equality:

$$(f \star \psi_j) \star \psi_{j-1} = (f \star \psi_{j-1}) \star \psi_j$$

This solves the problem.

The number of elements in the lists is exponential in the size of the supports of $\hat{\psi}_j$ and $\hat{\psi}_{j-1}$, so this algorithm has a prohibitive complexity when the supports become large. Otherwise, our numerical experiments show that it works well.

4.2.3 Error correction

When the modulus are noisy, there can be errors during the phase propagation step. The local optimization generally corrects them, if run for a sufficient amount of time, but, for the case where some errors are left, we add, at each scale, a multigrid error correction step. This step does not totally remove the errors but greatly reduces their amplitude.

Principle

First, we determine the values of n for which $f \star \psi_j^{low}[n]$ and $f \star \psi_{j+1}^{high}[n]$ seems to have been incorrectly reconstructed. We use the fact that $f \star \psi_j^{low}$ and $f \star \psi_{j+1}^{high}$ must satisfy:

$$(f \star \psi_j^{low}) \star \psi_{j+1}^{high} = (f \star \psi_{j+1}^{high}) \star \psi_j^{low}$$

The points where this equality does not hold provide a good estimation of the places where the values of $f \star \psi_j^{low}$ and $f \star \psi_{j+1}^{high}$ are erroneous.

We then construct a set of smooth "windows" $w_1, ..., w_S$, whose supports cover the interval on which errors have been found (see figure 4.3), such that each window has a small support. For each s, we reconstruct $(f \star \psi_j^{low}).w_s$ and $(f \star \psi_{j+1}^{high}).w_s$, by expressing these functions as the solutions to phase retrieval problems of small size, which we can solve by the exhaustive search method described in Paragraph 4.2.2.

As w_s is smooth, the multiplication by w_s approximately commutes with the convolution by ψ_j, ψ_{j+1}:

$$|(f.w_s) \star \psi_j| \approx |(f \star \psi_j).w_s| = w_s|f \star \psi_j|$$
$$|(f.w_s) \star \psi_{j+1}| \approx |(f \star \psi_{j+1}).w_s| = w_s|f \star \psi_{j+1}|$$

The wavelets ψ_j and ψ_{j+1} have a small support in the Fourier domain, if we truncate them to the support of w_s, so we can solve this problem by exhaustive search, and reconstruct $(f.w_s) \star \psi_j$ and $(f.w_s) \star \psi_{j+1}$.

From $(f.w_s) \star \psi_j$ and $(f.w_s) \star \psi_{j+1}$, we reconstruct $(f \star \psi_j^{low}).w_s \approx (f.w_s) \star \psi_j^{low}$ and $(f \star \psi_{j+1}^{high}).w_s \approx (f.w_s) \star \psi_{j+1}^{high}$ by deconvolution.

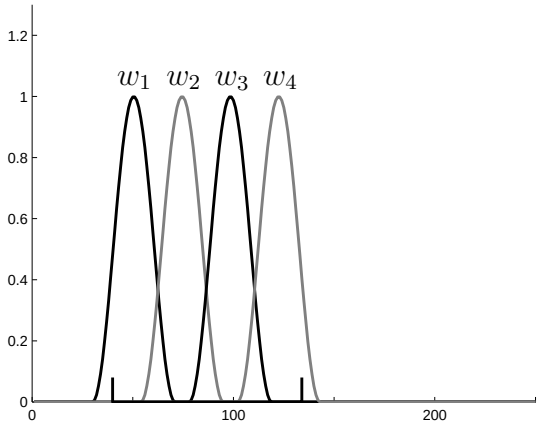


Figure 4.3: Four window signals, whose supports cover the interval in which errors have been detected

Usefulness of the error correction step

The error correction step does not perfectly correct the errors, but greatly reduces the amplitude of large ones.

Figure 4.4 shows an example of this phenomenon. It deals with the reconstruction of a difficult audio signal, representing a human voice saying "I'm sorry". Figure 4.4a shows $f \star \psi_7^{low}$ after the multiscale reconstruction at scale 7, but before the error correction step. The reconstruction presents large errors. Figure 4.4b shows the value after the error correction step. It is still not perfect but much closer to the ground truth.

So the error correction step must be used when large errors are susceptible to occur, and turned off otherwise: it makes the algorithm faster without reducing its precision.

Figure 4.5 illustrates this affirmation by showing the mean reconstruction error for the same audio signal as previously. When 200 iterations only are allowed at each local optimization step, there are large errors in the multiscale reconstruction; the error correction step significantly reduces the reconstruction error. When 2000 iterations are allowed, all the large errors can be corrected during the local optimization steps and the error correction step is not useful.

4.3 Multiscale versus non-multiscale

Our reconstruction algorithm has very good reconstruction performances, mainly because it uses the reformulation of the phase retrieval problem introduced in Section 4.1. However, the

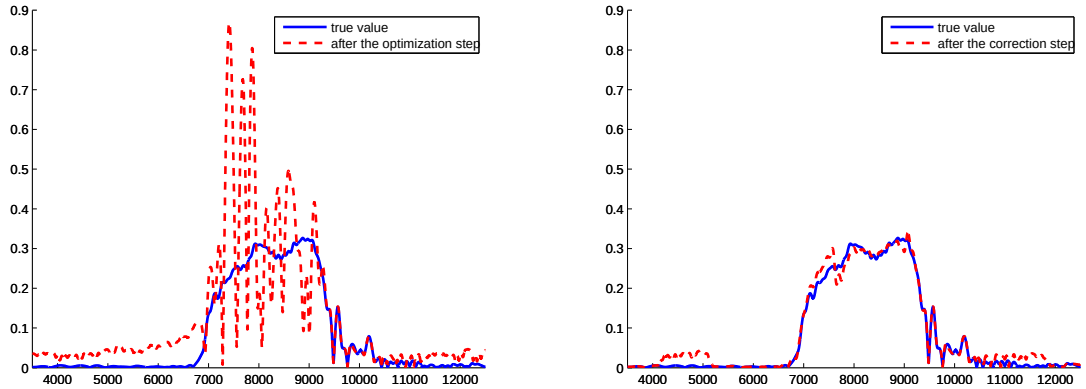


Figure 4.4: For an audio signal, the reconstructed value of $f \star \psi_7^{low}$ at the scale 7 of the multiscale algorithm, in modulus (dashed line); the solid line represents the ground true. (a) Before the error correction step (b) After the error correction step

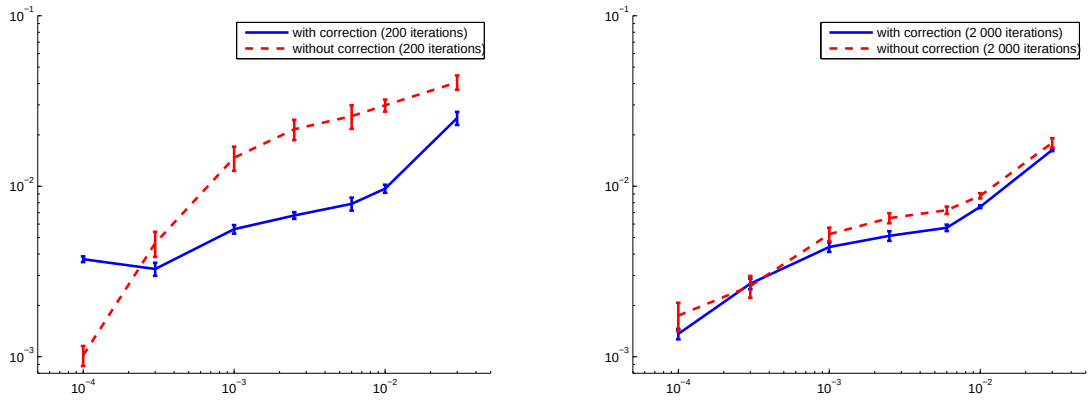


Figure 4.5: Mean reconstruction error (4.9) as a function of the noise, for an audio signal representing a human voice. (a) Maximal number of iterations per local optimization step equal to 200 (b) Maximal number equal to 2000.

quality of its results is also due to its multiscale structure. It is indeed known that, for the reconstruction of functions from their spectrogram or scalogram, multiscale algorithms perform better than non-multiscale ones [Bouvrie and Ezzat, 2006; Bruna, 2013].

In this section, we propose two justifications for this phenomenon (paragraph 4.3.1). We then introduce a multiscale version of the classical Gerchberg-Saxton algorithm, and numerically verify that it yields better reconstruction results than the usual non-multiscale version (paragraph 4.3.2).

4.3.1 Advantages of the multiscale reconstruction

At least two factors can explain the superiority of multiscale methods, where the $f \star \psi_j$'s are reconstructed one by one, and not all at the same time.

First, they can partially remedy the possible ill-conditioning of the problem. In particular, if the $f \star \psi_j$'s have very different norms, then a non-multiscale algorithm will be more sensitive to the components with a high norm. It may neglect the information given by $|f \star \psi_j|$, for the values of j such that this function has a small norm. With an multiscale algorithm where all the $|f \star \psi_j|$'s are successively considered, this happens less frequently.

Second, iterative algorithms, like Gerchberg-Saxton, are very sensitive to the choice of their starting point (hence the care given to their initialization in the literature [Netrapalli et al., 2013; Candès et al., 2015]). If all the components are reconstructed at the same time and the starting point is randomly chosen, the algorithm almost never converges towards the correct solution: it gets stuck in a local minima. In a multiscale algorithm, the starting point at each scale can be chosen so as to be consistent with the values reconstructed at lower scales; it yields much better results.

4.3.2 Multiscale Gerchberg-Saxton

To justify the efficiency of the multiscale approach, we introduce a multiscale version of the classical Gerchberg-Saxton algorithm [Gerchberg and Saxton, 1972] (by alternate projections) and compare its performances with the non-multiscale algorithm.

The multiscale algorithm reconstructs $f \star \psi_J$ by exhaustive search (paragraph 4.2.2).

Then, for each j, once $f \star \psi_J, ..., f \star \psi_{j+1}$ are reconstructed, an initial guess for $f \star \psi_j$ is computed by deconvolution. The frequencies of $f \star \psi_j$ for which the deconvolution is too unstable are set to zero. The regular Gerchberg-Saxton algorithm is then simultaneously applied to $f \star \psi_J, ..., f \star \psi_j$.

We test this algorithm on realizations of Gaussian random processes (see Section 4.4.2 for details), of various lengths. On Figure 4.6, we plot the mean reconstruction error obtained with

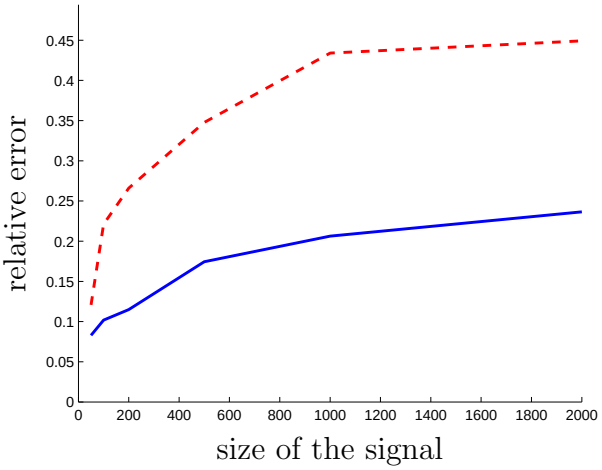


Figure 4.6: Mean reconstruction error, as a function of the size of the signal; the solid blue line corresponds to the multiscale algorithm and the dashed red one to the non-multiscale one.

the regular Gerchberg-Saxton algorithm and the error obtained with the multiscale version (see Paragraph 4.4.1 for the definition of the reconstruction error).

None of the algorithms is able to perfectly reconstruct the signals, in particular when their size increases. However, the multiscale algorithm clearly yields better results, with a mean error approximately twice smaller.

4.4 Numerical results

In this section, we describe the behavior of our algorithm. We compare it with Gerchberg-Saxton and with *PhaseLift*. We show that it is much more precise than Gerchberg-Saxton. It is comparable with *PhaseLift* in terms of precision, but significantly faster, so it allows to reconstruct larger signals.

The performances strongly depend on the type of signals we consider. The main source of difficulty for our algorithm is the presence of small values in the wavelet transform, especially in the low frequencies.

Indeed, the reconstruction of $f \star \psi_j^{high}$ by Equation (4.4) involves a division by $f \star \psi_j^{low}$. When $f \star \psi_j^{low}$ has small values, this operation is unstable and induces errors.

As we will see in Section 4.4.3, the signals whose wavelet transform has many small values are also the signals for which the phase retrieval problem is the least stable (in the sense that two functions can have wavelet transforms almost equal in modulus without being close in l^2-norm). This suggests that this class of functions is intrinsically the most difficult to reconstruct; it is

not an artifact of our algorithm.

We describe our experimental setting in Paragraph 4.4.1. In Paragraph 4.4.2, we give detailed numerical results for various types of signals. In Paragraph 4.4.3, we use our algorithm to investigate the stability to noise of the underlying phase retrieval problem. Finally, in Paragraph 4.4.4, we study the influence of various parameters on the quality of the reconstruction.

Code and audio examples are available at:

http://www.di.ens.fr/~waldspurger/wavelets_phase_retrieval.html

4.4.1 Experimental setting

At each reconstruction trial, we choose a signal f and compute its wavelet transform $\{|f \star \psi_j|\}_{0 \leq j \leq J}$. We corrupt it with a random noise n_j:

$$h_j = |f \star \psi_j| + n_j \tag{4.7}$$

We measure the amplitude of the noise in l^2-norm, relatively to the l^2-norm of the wavelet transform:

$$\text{amount of noise} = \frac{\sqrt{\sum_j ||n_j||_2^2}}{\sqrt{\sum_j ||f \star \psi_j||_2^2}} \tag{4.8}$$

We run the algorithm on the noisy wavelet transform $\{h_j\}_{0 \leq j \leq J}$. It returns a reconstructed signal f_{rec}. We quantify the reconstruction error by the difference, in relative l^2-norm, between the modulus of the wavelet transform of the original signal f and the modulus of the wavelet transform of the reconstructed signal f_{rec}:

$$\text{reconstruction error} = \frac{\sqrt{\sum_j || \, |f \star \psi_j| - |f_{rec} \star \psi_j| \, ||_2^2}}{\sqrt{\sum_j ||f \star \psi_j||_2^2}} \tag{4.9}$$

Alternatively, we could measure the difference between f and f_{rec}:

$$\text{error on the signal} = \frac{||f - f_{rec}||_2}{||f||_2} \tag{4.10}$$

But we know that the reconstruction of a function from the modulus of its wavelet transform is not stable to noise (Section 3.3 of this document). So we do not hope the difference between

f and f_{rec} to be small. We just want the algorithm to reconstruct a signal f_{rec} whose wavelet transform is close to the wavelet transform of f, in modulus. Thus, the reconstruction error (4.9) is more relevant to measure the performances of the algorithm.

In all the experiments, unless otherwise specified, we use dyadic Morlet wavelets, to which we subtract Gaussian functions of small amplitude so that they have zero mean:

$$\hat{\psi}(\omega) = \exp(-p(\omega - 1)^2) - \beta \exp(-p\omega^2)$$

where $\beta > 0$ is chosen so that $\hat{\psi}(0) = 0$ and the parameter p is arbitrary (it controls the frequency bandwidth of the wavelets). For $N = 256$, our family of wavelets contains eight elements, which are plotted on Figure 4.18a. The performances of the algorithm strongly depend on the choice of the wavelet family; this is discussed in Paragraph 4.4.4.

The maximal number of iterations per local optimization step is set to 10000 (with an additional stopping criterion, so that the 10000-th iteration is not always reached). We study the influence of this parameter in Paragraph 4.4.4.

The noises are realizations of Gaussian white noises.

The error correction step described in Paragraph 4.2.3 is always turned on.

Gerchberg-Saxton is applied in a multiscale fashion, as described in Paragraph 4.3.2, which yields better results than the regular implementation.

We use *PhaseLift* [Candès et al., 2011] with ten steps of reweighting, followed by 2000 iterations of the Gerchberg-Saxton algorithm. In our experiments with *PhaseLift*, we only consider signals of size $N = 256$. Handling larger signals is difficult with a straightforward Matlab implementation.

4.4.2 Results

We describe four classes of signals, whose wavelet transforms have more or less small values. For each class, we plot the reconstruction error of our algorithm, Gerchberg-Saxton and *PhaseLift* as a function of the noise error.

Realizations of Gaussian random processes

We first consider realizations of Gaussian random processes. A signal f in this class is defined by:

$$\begin{aligned} \hat{f}[k] &= \frac{X_k}{\sqrt{k+1}} && \text{if } k \in \{1, ..., N/2\} \\ &= 0 && \text{if not} \end{aligned}$$

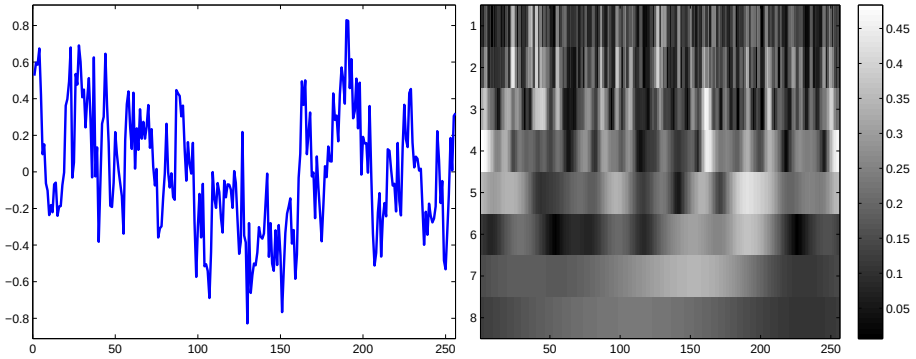


Figure 4.7: Realization of a Gaussian process (left) and modulus of its wavelet transform (right)

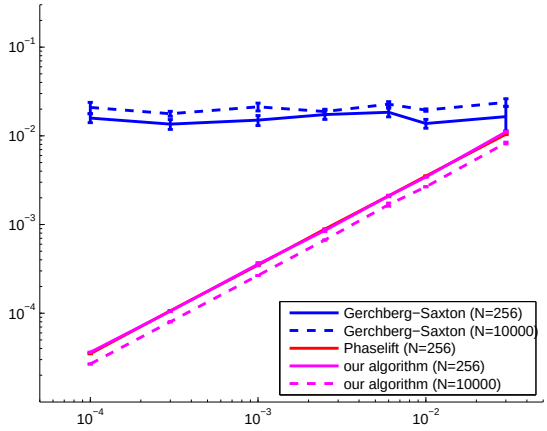


Figure 4.8: Mean reconstruction error as a function of the noise, for Gaussian signals of size $N = 256$ or 10000

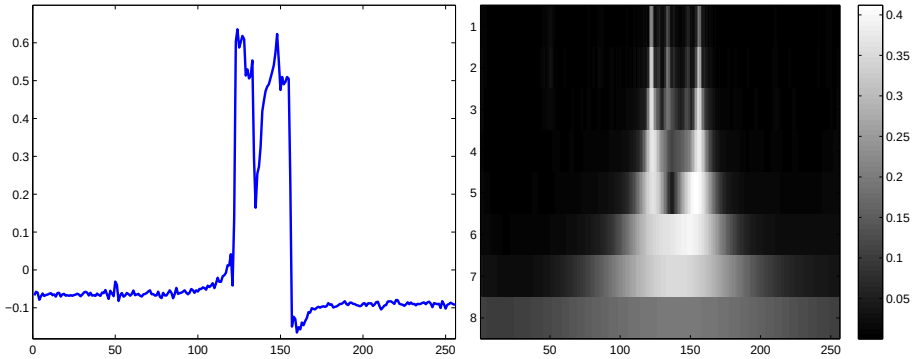


Figure 4.9: Line from an image (left) and modulus of its wavelet transform (right)

where $X_1, ..., X_{N/2}$ are independent realizations of complex Gaussian centered variables. The role of the $\sqrt{k+1}$ is to ensure that all components of the wavelet transform approximately have the same l^2-norm (in expectation). An example is displayed on Figure 4.7, along with the modulus of its wavelet transform.

The wavelet transforms of these signals have few small values, disposed in a seemingly random pattern. This is the most favorable class for our algorithm.

The reconstruction results are shown in Figure 4.8. Even for large signals ($N = 10000$), the mean reconstruction error is proportional to the input noise (generally 2 or 3 times smaller); this is the best possible result. The performances of *PhaseLift* are exactly the same, but Gerchberg-Saxton often fails.

Lines from images

The second class consists in lines randomly extracted from photographs. These signals have oscillating parts (corresponding to the texture zones of the initial image) and smooth parts, with large discontinuities in between. Their wavelet transforms generally contain a lot a small values, but, as can be seen in Figure 4.9, the distribution of these small values is particular. They are more numerous at high frequencies and the non-small values tend to concentrate on vertical lines of the time-frequency plane.

This distribution is favorable to our algorithm: small values in the wavelet transform are mostly a problem when they are in the low frequencies and prevent the correct initialization of the reconstruction at medium or high frequencies. Small values at high frequencies are not a problem.

Indeed, as in the case of Gaussian signals, the reconstruction error is proportional to the input noise (figure 4.10). This is also the case for *PhaseLift* but not for Gerchberg-Saxton.

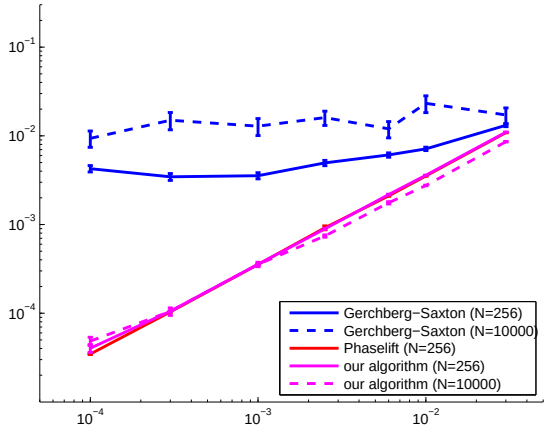


Figure 4.10: Mean reconstruction error as a function of the noise, for lines extracted from images, of size $N = 256$ or 10000

Sums of a few sinusoids

The next class of signals contains sums of a few numbers of sinusoids, multiplied by a window function w to avoid boundary effects. Formally, a signal in this class is of the form:

$$f[n] = \left[\sum_{k=1}^{N/2} \alpha_k \exp\left(i \frac{2\pi k n}{N} \right) \right] \times w[n]$$

where the α_k are zero with high probability and realizations of complex Gaussian centered variables with small probability.

The wavelet transforms of these signals often have components of very small amplitude, which may be located at any frequential scale (figure 4.11). This can prevent the reconstruction.

The results are on Figure 4.12. Our algorithm performs much better than Gerchberg-Saxton but the results are not as good as for the two previous classes of signals.

In most reconstruction trials, the signal is correctly reconstructed, up to an error proportional to the noise. But, with a small probability, the reconstruction fails. The same phenomenon occurs for *PhaseLift*.

The probability of failure seems a bit higher for *PhaseLift* than for our algorithm. For example, when the signals are of size 256 and the noise has a relative norm of 0.01%, the reconstruction error is larger than the noise error 20% of the time for *PhaseLift* and only 10% of the time for our algorithm. However, *PhaseLift* has a smaller mean reconstruction error because, in these failure cases, the result it returns, although not perfect, is more often close to

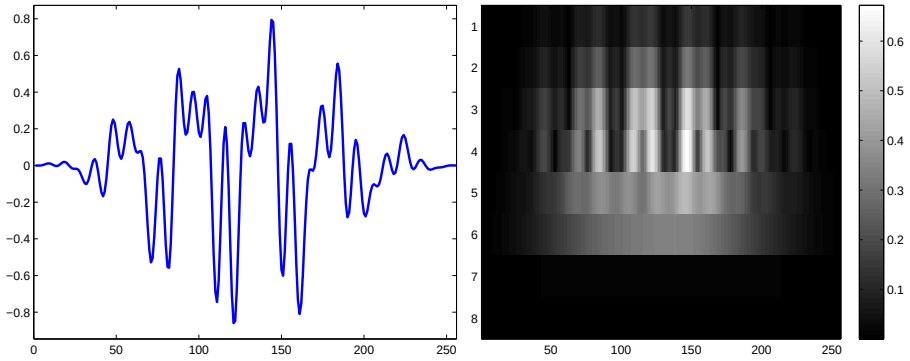


Figure 4.11: Random sum of sinusoids, multiplied by a window function (left) and modulus of its wavelet transform (right)

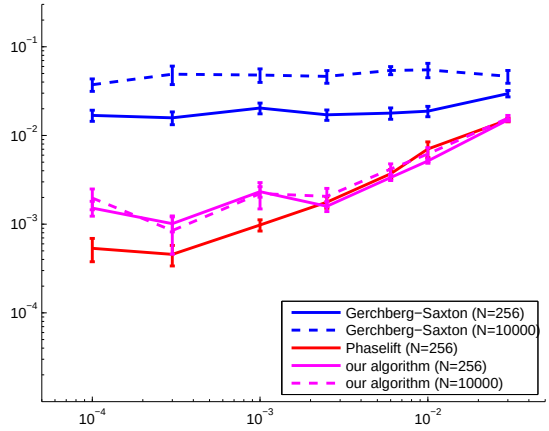
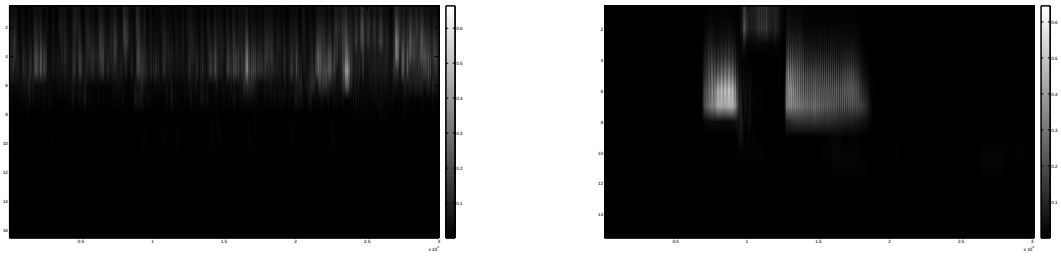


Figure 4.12: Mean reconstruction error as a function of the noise, for random sums of sinusoids multiplied by a window function, of size $N = 256$ or 10000



(a) (b)

Figure 4.13: Wavelet transforms of the audio signals (a) Rimsky-Korsakov (b) "I'm sorry"

the truth: the mean reconstruction error in the failure cases is 0.2% for *PhaseLift* versus 1.7% for our algorithm.

Audio signals

Finally, we test our algorithm on real audio signals. These signals are difficult to reconstruct because they do not contain very low frequencies (as the human ear cannot hear them, these frequencies are not included in the recordings), so the first components of their wavelet transforms are very small.

The reconstruction results may vary from one audio signal to the other. We focus here on two representative examples.

The first signal is an extract of five seconds of a musical piece played by an orchestra (the *Flight of the Bumblebee*, by Rimsky-Korsakov). Figure 4.13a shows the modulus of its wavelet transform. It has 16 components and 9 of them (the ones with lower characteristic frequencies) seem negligible, compared to the other ones. However, its non-negligible components have a moderate number of small values.

The second signal is a recording of a human voice saying "I'm sorry" (figure 4.13b). The low-frequency components of its wavelet transform are also negligible, but even the high-frequency components tend to have small values, which makes the reconstruction even more difficult.

The results are presented in Figures 4.14 and 4.15. For relatively high levels of noise (0.5% or higher), the results, in the sense of the l^2-norm, are satisfying: the reconstruction error is smaller or equal to the amount of noise.

In the high precision regime (that is, for 0.1% of noise or less), the lack of low frequencies does not allow a perfect reconstruction. Nevertheless, the results are still good: the reconstruction error is of the order of 0.1% or 0.2% when the noise error is below 0.1%. More iterations in the optimization steps can further reduce this error. By comparison, the reconstruction error with Gerchberg-Saxton is always several percent, even when the noise is small.

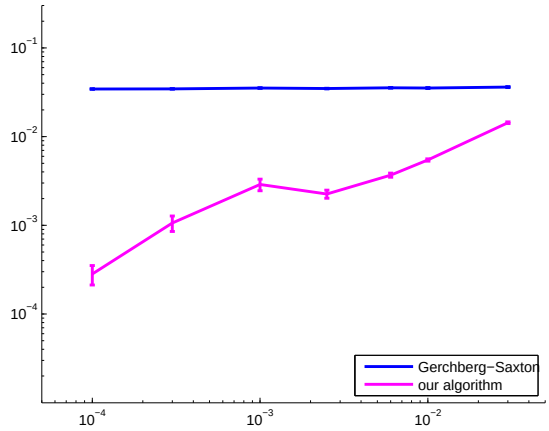


Figure 4.14: mean reconstruction error as a function of the noise, for the audio signal "Rimsky-Korsakov"

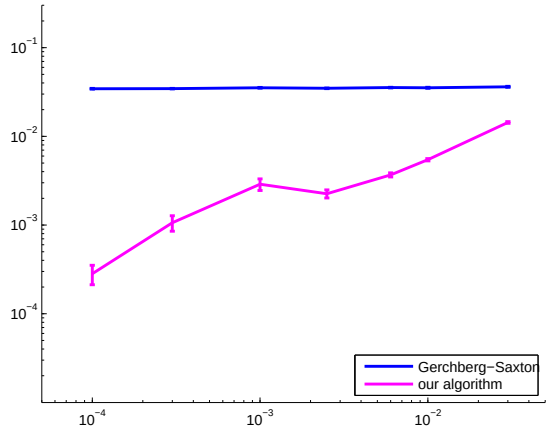


Figure 4.15: mean reconstruction error as a function of the noise, for the audio signal "I'm sorry"

4.4.3 Stability of the reconstruction

In this section, we use our reconstruction algorithm to investigate the stability of the reconstruction. From Section 3.3, we know that the reconstruction is not globally stable to noise: the reconstruction error (4.9) can be small (the modulus of the wavelet transform is almost exactly reconstructed), even if the error on the signal (4.10) is not small (the difference between the initial signal and its reconstruction is large).

We show that this phenomenon can occur for all classes of signals, but is all the more frequent when the wavelet transform has a lot of small values, especially in the low frequency components.

We also experimentally show that, when this phenomenon happens, the original and reconstructed signals have their wavelet transforms $\{f \star \psi_j(t)\}_{j\in\mathbb{Z},t\in\mathbb{R}}$ equal up to multiplication by a phase $\{e^{i\phi_j(t)}\}_{j\in\mathbb{Z},t\in\mathbb{R}}$, which varies slowly in both j and t, except maybe at the points where $f\star\psi_j(t)$ is close to zero. For Cauchy wavelets, we have proven an approximation of this assertion in Section 3.4; we conjecture this result to be valid for more general wavelets.

We perform a large number of reconstruction trials, with various reconstruction parameters. This gives us a large number of pairs (f, f_{rec}), such that $\forall j, t, |f \star \psi_j(t)| \approx |f_{rec} \star \psi_j(t)|$. For each one of these pairs, we compute:

$$\text{error on the modulus} = \frac{\sqrt{\sum_j || \, |f \star \psi_j| - |f_{rec} \star \psi_j| \, ||_2^2}}{\sqrt{\sum_j ||f \star \psi_j||_2^2}} \tag{4.9}$$

$$\text{error on the signal} = \frac{||f - f_{rec}||_2}{||f||_2} \tag{4.10}$$

The results are plotted on Figure 4.16, where each point corresponds to one reconstruction trial. The x-coordinate represents the error on the modulus and the y-coordinate the error on the signal.

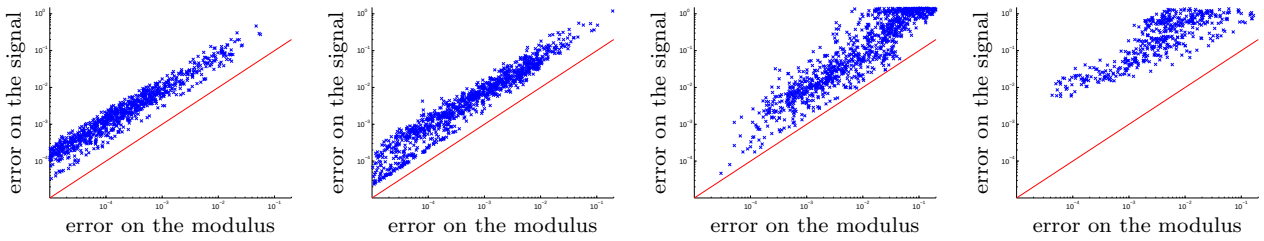
We always have:

$$\text{error on the modulus} \leq C \times (\text{error on the function})$$

with C a constant of the order of 1. This is not surprising because the modulus of the wavelet transform is a Lipschitz operator, with a constant close to 1.

As expected, the converse inequality is not true: the error on the function can be significantly larger than the error on the modulus. For each class, an important number of reconstruction trials yield errors such that:

$$\text{error on the signal} \approx 30 \times \text{error on the modulus}$$



(a) (b) (c) (d)

Figure 4.16: error on the signal (4.10) as a function of the error on the modulus of the wavelet transform (4.9), for several reconstruction trials; the red line $y = x$ is here to serve as a reference (a) Gaussian signals (b) lines from images (c) sums of sinusoids (d) audio signal "I'm sorry"

For realizations of Gaussian random processes or for lines extracted from images (figures 4.16a and 4.16b), the ratio between the two errors never exceeds 30 (except for one outlier). But for sums of a few sinusoids (4.16c) or audio signals (4.16d), we may even have:

$$\text{error on the signal} \geq 100 \times \text{error on the modulus}$$

So instabilities appear in the reconstruction of all kinds of signals, but are stronger for sums of sinusoids and audio signals, that is for the signals whose wavelet transforms have a lot of small values, especially in the low frequencies.

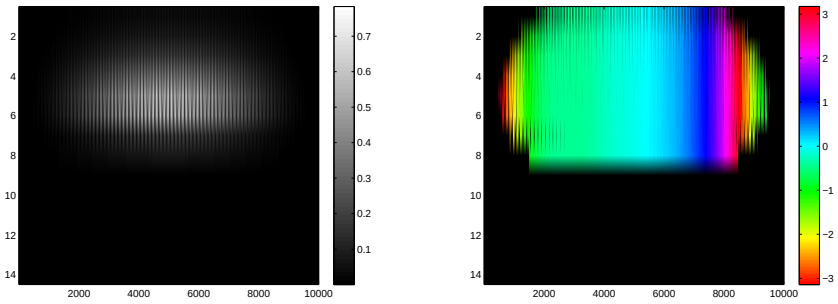
These results have a theoretical justification. Let us recall that, in Section 3.3, we have explained how, from any signal f, it is possible to construct g such that $|f \star \psi_j| \approx |g \star \psi_j|$ for all j but $f \not\approx g$ in the l^2-norm sense.

The principle of the construction is to multiply each $f \star \psi_j(t)$ by a phase $e^{i\phi_j(t)}$. The function $(j,t) \to e^{i\phi_j(t)}$ must be chosen so that it varies slowly in both j and t, except maybe at the points (j,t) where $f \star \psi_j(t)$ is small. Then there exist a signal g such that $(f \star \psi_j(t))e^{i\phi_j(t)} \approx g \star \psi_j(t)$ for any j, t. Taking the modulus of this approximate equality yields:

$$\forall j,t \qquad |f \star \psi_j(t)| \approx |g \star \psi_j(t)|$$

However, we may not have $f \approx g$.

This construction works for any signal f (unless the wavelet transform is very localized in the time frequency domain), but the number of possible $\{e^{i\phi_j(t)}\}_{j,t}$ is larger when the wavelet transform of f has a lot of small values, because the constraint of slow variation is relaxed at the points where the wavelet transform is small (especially when the small values are in the low frequencies). This is probably why instabilities occur for all kinds of signals, but more frequently when the wavelet transforms have a lot of zeroes.



(a) (b)

Figure 4.17: (a) modulus of the wavelet transform of a signal (b) phase difference between the wavelet transform of this signal and of its reconstruction (black points correspond to places where the modulus is too small for the phase to be meaningful)

From our experiments, it seems that the previous construction describes all the instabilities: when the wavelet transforms of f and f_{rec} have almost the same modulus and f is not close to f_{rec}, then the wavelet transforms of f and f_{rec} are equal up to slow-varying phases $\{e^{i\phi_j(t)}\}_{j,t}$.

Figure 4.17 shows an example. The signal is a sum of sinusoids. The relative difference between the modulus is 0.3%, but the difference between the initial and reconstructed signals is more than a hundred times larger; it is 46%. The right subfigure shows the difference between the phases of the two wavelet transforms. It indeed varies slowly, in both time and frequency (actually, it is almost constant along the frequency axis), and a bit faster at the extremities, where the wavelet transform is closer to zero.

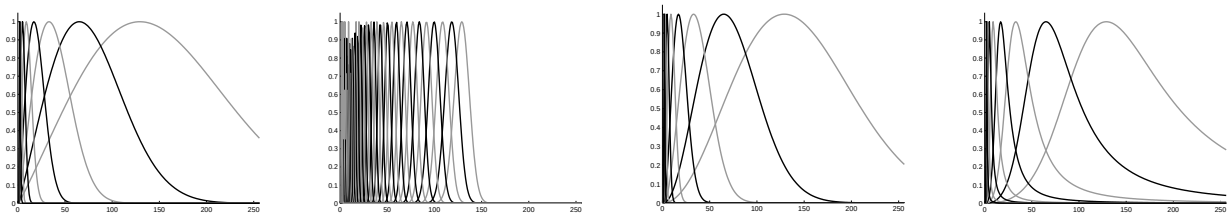
4.4.4 Influence of the parameters

In this paragraph, we analyze the importance of the two main parameters of the algorithm: the choice of the wavelets (paragraph 4.4.4) and the number of iterations allowed per local optimization step (paragraph 4.4.4).

Choice of the wavelets

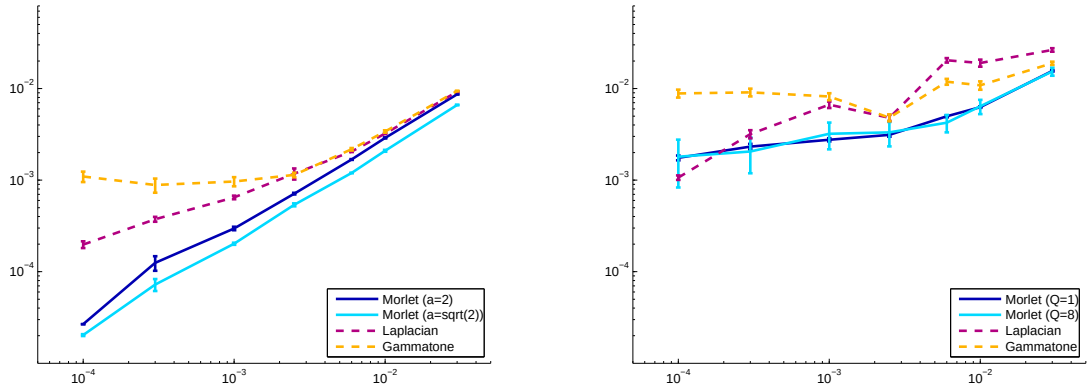
Two properties of the wavelets are especially important: the exponential decay of the wavelets in the Fourier domain (so that the Q_j's (4.2) are correctly computed) and the amount of overlap between two neighboring wavelets (if the overlap is too small, then $f \star \psi_J, ..., f \star \psi_{j+1}$ contain not much information about $f \star \psi_j$ and the multiscale approach is less efficient).

We compare the reconstruction results for four families of wavelets.



(a) (b) (c) (d)

Figure 4.18: Four wavelet families. (a) Morlet (b) Morlet with dilation factor $2^{1/8}$ (c) Laplacian (d) Gammatone



(a) (b)

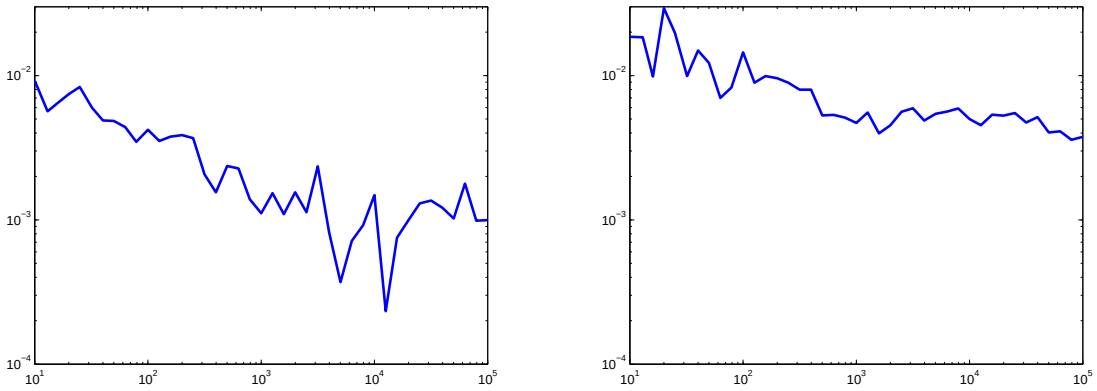
Figure 4.19: Mean reconstruction error as a function of the noise for the four wavelet families displayed in 4.18. (a) Lines from images (b) Audio signal "I'm sorry"

The first family (figure 4.18a) is the one we used in all the previous experiments. It contains dyadic Morlet wavelets. The second family (figure 4.18b) also contains Morlet wavelets, with a smaller bandwidth (Q-factor ≈ 8) and a dilation factor of $2^{1/8}$ instead of 2. This is the kind of wavelets used in audio processing. The third family (figure 4.18c) consists in dyadic Laplacian wavelets $\hat{\psi}(\omega) = \omega^2 e^{1-\omega^2}$. Finally, the wavelets of the fourth family (figure 4.18d) are (derivatives of) Gammatone wavelets.

Figure 4.19 displays the mean reconstruction error as a function of the noise, for two classes of signals: lines randomly extracted from natural images and audio signals.

Morlet wavelets have a fast decay and consecutive wavelets overlap well. This does not depend upon the dilation factor so the reconstruction performances are similar for the two Morlet families (figures 4.19a and 4.19b).

Laplacian wavelets are similar, but the overlap between consecutive wavelets is not as good.



(a) (b)

Figure 4.20: for the audio signal "I'm sorry", reconstruction error as a function of the maximal number of iterations (a) with 0.01% of noise (b) with 0.6% of noise

So Laplacian wavelets globally have the same behavior as Morlet wavelets but require significantly more computational effort to reach the same precision. Figures 4.19a and 4.19b have been generated with a maximal number of iterations per optimization step equal to 30000 (instead of 10000) and the reconstruction error is still larger.

The decay of Gammatone wavelets is polynomial instead of exponential. The products Q_j cannot be efficiently estimated and our method performs significantly worse. In the case of lines extracted from images (4.19a), the reconstruction error stagnates at 0.1%, even when the noise is of the order of 0.01%. For audio signals (4.19b), it is around 1% for any amount of noise.

Number of iterations in the optimization step

The maximal number of iterations allowed per local optimization step (paragraph 4.1.3) can have a huge impact on the quality of the reconstruction.

Figure 4.20 represents, for an audio signal, the reconstruction error as a function of this number of iterations. As the objective functions are not convex, there are no guarantees on the speed of the decay when the number of iterations increases. It can be slow and even non-monotonic. Nevertheless, it clearly globally decays.

The execution time is roughly proportional to the number of iterations. It is thus important to adapt this number to the desired application, so as to reach the necessary precision level without making the algorithm excessively slow.

4.5 Proof of Lemmas 4.4 and 4.2

4.5.1 Proof of Lemma 4.4

Lemma. *(4.4) Let $m \in \mathbb{R}^N$ and $K \in \mathbb{N}^*$ be fixed. We consider the problem:*

$$\begin{aligned} &\text{Find } g \in \mathbb{C}^N \text{ s.t. } |g| = m \\ &\qquad\qquad \text{and } Supp(\hat{g}) \subset \{1, ..., K\} \end{aligned}$$

This problem has at most 2^{K-1} solutions, up to a global phase, and there exist a simple algorithm which, from m and N, returns the list of all possible solutions.

Proof. We define:

$$P(g)(X) = \hat{g}[1] + \hat{g}[2]X + ... + \hat{g}[K]X^{K-1}$$

We show that the constraint $|g| = m$ amounts to knowing $P(g)(X)\overline{P(g)}(1/X)$. This is in turn equivalent to knowing the roots of $P(g)$ (and thus knowing g) up to inversion with respect to the unit circle. There are in general $K-1$ roots, and each one can be inverted. This gives 2^{K-1} solutions.

We set:

$$\begin{aligned} Q(g)(X) &= \overline{P(g)}(1/X) \\ &= \overline{\hat{g}[K]}X^{-(K-1)} + \overline{\hat{g}[K-1]}X^{-(k-2)} + ... + \overline{\hat{g}[1]} \end{aligned}$$

Equation $|g|^2 = m^2$ is equivalent to $\widehat{|g|^2} = \widehat{m^2}$, that is $\frac{1}{N}\hat{g} \star \hat{\bar{g}} = \widehat{m^2}$. For each $k \in \{-(K-1), ..., K-1\}$:

$$\hat{g} \star \hat{\bar{g}}[k] = \sum_s \hat{g}[k-s]\overline{\hat{g}[-s]}$$

This number is the coefficient of order k of $P(g)(X)Q(g)(X)$, so $|g| = m$ if and only if:

$$P(g)(X)Q(g)(X) = N \sum_{k=-(K-1)}^{K-1} \widehat{m^2}[k]X^k \tag{4.11}$$

Let us denote by $r_1, ..., r_{K-1}$ the roots of $P(g)(X)$, so that:

$$\begin{aligned} P(g)(X) &= \hat{g}[K](X - r_1)...(X - r_{K-1}) \\ Q(g)(X) &= \overline{\hat{g}[K]}(1/X - \overline{r}_1)...(1/X - \overline{r}_{K-1}) \end{aligned}$$

From (4.11), the equality $|g| = m$ holds if and only if $\hat{g}[K], r_1, ..., r_{K-1}$ satisfy:

$$|\hat{g}[K]|^2 \prod_{j=1}^{K-1} (X - r_j)(1/X - \overline{r}_j)$$
$$= N \sum_{k=-(K-1)}^{K-1} \widehat{m^2}[k] X^k \tag{4.12}$$

If we denote by $s_1, 1/\overline{s}_1, ..., s_{K-1}, 1/\overline{s}_{K-1}$ the roots of the polynomial function $\sum_{k=-(K-1)}^{K-1} \widehat{m^2}[k] X^k$, then the only possible choices for $r_1, ..., r_{K-1}$ are, up to permutation:

$$r_1 = s_1 \text{ or } 1/\overline{s}_1 \qquad r_2 = s_2 \text{ or } 1/\overline{s}_2 \qquad \dots$$

So there are 2^{K-1} possibilities. Once the r_j have been chosen, $\hat{g}[K]$ is uniquely determined by (4.12), up to multiplication by a unitary complex.

From $r_1, ..., r_{K-1}, \hat{g}[K]$, $P(g)$ is uniquely determined and so is g. The algorithm is summarized in 6. □

Algorithm 6 reconstruction by exhaustive search for a small problem

Input: K,m

1: Compute the roots of $\sum_{k=-(K-1)}^{K-1} \widehat{m^2}[k] X^k$
2: Group them by pairs $(s_1, 1/\overline{s}_1), ..., (s_{K-1}, 1/\overline{s}_{K-1})$
3: List the 2^{K-1} elements $(r_1, ..., r_{K-1})$ of $\{s_1, 1/\overline{s}_1\} \times ... \times \{s_{K-1}, 1/\overline{s}_{K-1}\}$
4: **for all** the elements **do**
5: Compute the corresponding $\hat{g}[K]$ by (4.12)
6: Compute the coefficients of $P(g)(X) = \hat{g}[K](X - r_1)...(X - r_{K-1})$
7: Apply an IFFT to the coefficients to obtain g
8: **end for**

Output: the list of 2^{K-1} possible values for g

4.5.2 Proof of Lemma 4.2

Lemma (4.2). *For any f satisfying the analycity condition (4.1):*

$$\forall z \in \mathbb{C} \qquad P(|f \star \psi_j|^2)(rz) = P((f \star \psi_j^{low})\overline{(f \star \psi_j^{high})})(z)$$
$$\text{and} \quad P(g_j^2)(rz) = P(Q_j)(z)$$

Proof. Recall that, by definition, for any $h \in \mathbb{C}^N$:

$$\forall z \in \mathbb{C} \qquad P(h)(z) = \sum_{k=\left\lfloor \frac{N}{2} \right\rfloor - N+1}^{\left\lfloor \frac{N}{2} \right\rfloor} \hat{h}[k] z^k$$

So for any two signals h, H, the condition $P(h)(rz) = P(H)(z), \forall z \in \mathbb{C}$ is equivalent to:

$$\forall k = \left\lfloor \frac{N}{2} \right\rfloor - N + 1, ..., \left\lfloor \frac{N}{2} \right\rfloor \qquad \hat{h}[k] r^k = \hat{H}[k] \tag{4.13}$$

Applied to g_j^2 and Q_j, this property yields the equality $P(g_j^2)(rz) = P(Q_j)(z), \forall z \in \mathbb{C}$: by the definition of Q_j in (4.2), Equation (4.13) is clearly satisfied.

Let us now show that:

$$P(|f \star \psi_j|^2)(rz) = P((f \star \psi_j^{low})\overline{(f \star \psi_j^{high})})(z), \forall z \in \mathbb{C}$$

It suffices to prove that (4.13) holds, that is:

$$\forall k \in \left\lfloor \frac{N}{2} \right\rfloor - N + 1, ..., \left\lfloor \frac{N}{2} \right\rfloor,$$
$$\widehat{|f \star \psi_j|^2}[k] r^k = \widehat{(f \star \psi_j^{low})\overline{(f \star \psi_j^{high})}}[k]$$

Indeed, because the analycity condition (4.1) holds, we have for all k:

$$\begin{aligned}
\widehat{|f \star \psi_j|^2}[k] &= \frac{1}{N} \left(\widehat{f \star \psi_j} \right) \star \left(\widehat{\overline{f \star \psi_j}} \right)[k] \\
&= \frac{1}{N} \sum_{l=1}^{\lfloor N/2 \rfloor} \hat{f}[l] \hat{\psi}_j[l] \overline{\hat{f}[l-k] \hat{\psi}_j[l-k]} \\
&= \frac{r^{-k}}{N} \sum_{l=1}^{\lfloor N/2 \rfloor} \hat{f}[l] \hat{\psi}_j^{low}[l] \overline{\hat{f}[l-k] \hat{\psi}_j^{high}[l-k]} \\
&= \frac{r^{-k}}{N} \left(\widehat{f \star \psi_j^{low}} \right) \star \left(\widehat{\overline{f \star \psi_j^{high}}} \right)[k] \\
&= r^{-k} \left(\widehat{(f \star \psi_j^{low})\overline{(f \star \psi_j^{high})}} \right)[k]
\end{aligned}$$

□

Chapter 5

Exponential decay of scattering coefficients

In Chapters 3 and 4, we have studied, from the point of view of phase retrieval, the wavelet transform modulus. Our motivation was to understand the behavior of this operator, commonly used in data analysis as a convenient representation of signals, especially audio signals. However, if the wavelet transform modulus is an ubiquitous tool in signal processing, it can be insufficient for some tasks. It is indeed not invariant enough to perceptually negligible modifications of the input signal. It can then be used as a building block to construct more sophisticated representations, with more invariance properties.

In this chapter, we consider such a representation, the scattering transform. Introduced in [Mallat, 2012], the scattering transform has the property to be stable to translations and small deformations of the input signal. It gives excellent results in tasks which are naturally invariant to this transformations, like classification of audio signals and images [Bruna and Mallat, 2013a; Andén and Mallat, 2011].

The scattering is defined as a cascade of wavelet transform modulus. At each step of the cascade, the components of the wavelet transform are averaged, to produce so-called *scattering coefficients*. Each scattering coefficient has an order, which is the number of wavelet transform modulus that have been necessary to compute it.

Our long-term goal is to understand which properties of input signals are characterized by their scattering transform, in particular which regularity properties.

In this chapter, we focus on the relation between the order of scattering coefficients and the frequency band that they describe in the input signal. Theorem 5.2 indeed states that scattering coefficients of order n do almost not contain information on the frequential band centered at 0, of bandwidth proportional ra^n, for some constants $r > 0, a > 1$.

This indicates that there exists in scattering a phenomenon of propagation from high fre-

quencies towards low frequencies. Intuitively, the information initially contained in a frequency band of width A is moved, after one application of the wavelet transform modulus, to a frequency band of width Aa^{-1}, for some $a > 1$. After a second application of the wavelet transform modulus, it is in a band of width Aa^{-2}. And so on, until it is in the frequency band covered by the averaging operator, and disappears, absorbed in a scattering coefficient.

In Section 5.1, we precisely define the scattering transform and some known results about it. We also describe in more detail the propagation phenomenon. In Section 5.2, we state the theorem. In Section 5.3, we give the principle of the proof. Section 5.4 adapts the theorem to the scattering transform on stationary processes. Finally, Section 5.5 proves the lemmas necessary for the proof of the main theorem.

5.1 The scattering transform

As in the previous chapters, once a wavelet $\psi \in L^1 \cap L^2(\mathbb{R})$ (such that $\int_{\mathbb{R}} \psi = 0$) has been chosen, we define a family of wavelets $(\psi_j)_{j\in\mathbb{Z}}$ by:

$$\begin{aligned} & \forall j \in \mathbb{Z}, t \in \mathbb{R} \qquad \psi_j(t) = 2^{-j}\psi(2^{-j}t) \\ \iff \quad & \forall j \in \mathbb{Z}, \omega \in \mathbb{R} \qquad \hat{\psi}_j(\omega) = \hat{\psi}(2^j\omega) \end{aligned}$$

5.1.1 Definition

We follow the definition of [Mallat, 2012].

The scattering transform consists in a cascade of modulus of wavelet transforms. After each application of the modulus, the resulting functions are locally averaged. The set of averages constitutes the scattering transform.

The averaging is performed with a real-valued positive function $\phi \in L^1 \cap L^2(\mathbb{R})$ such that $\hat{\phi}(0) = 1$. We define:

$$\forall J \in \mathbb{Z}, t \in \mathbb{R} \qquad \phi_J(t) = 2^{-J}\phi(2^{-J}t)$$

The convolution with ϕ_J represents an average on an interval of characteristic size 2^J.

We now formally define the cascade of modulus of wavelet transforms. For any function $f \in L^2(\mathbb{R})$, we set:

$$U[\emptyset]f = f$$

and iteratively define, for any n-uplet $(j_1, ..., j_n) \in \mathbb{Z}^n$, with $n \geq 1$:

$$U[(j_1, ..., j_n)]f = \left|U[(j_1, ..., j_{n-1})]f \star \psi_{j_n}\right|$$

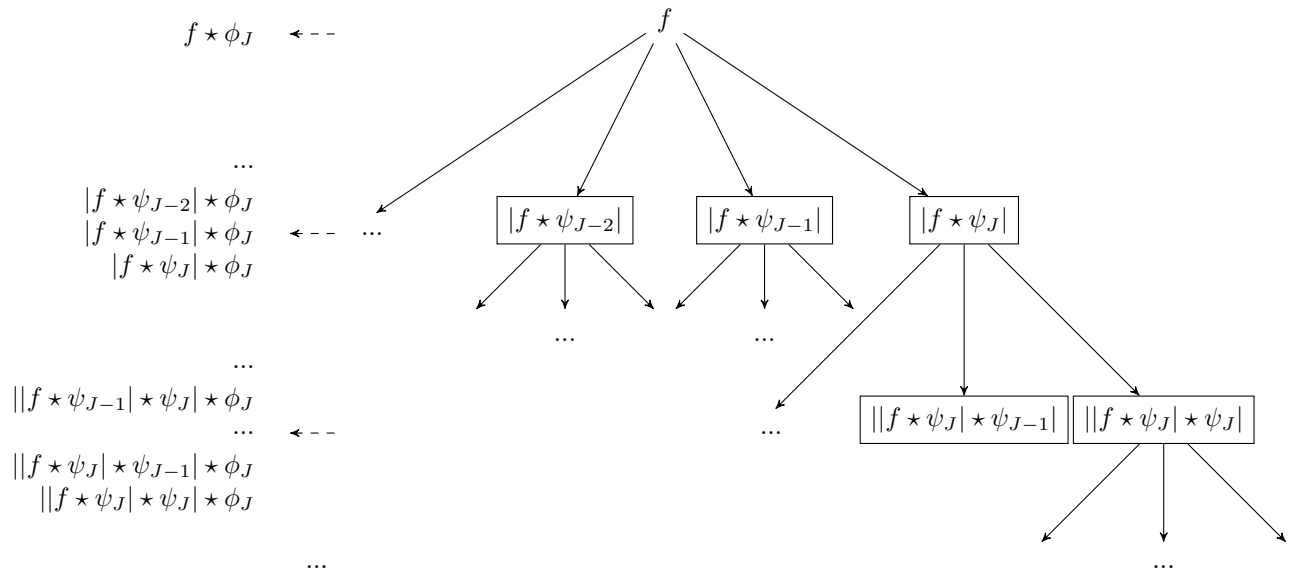


Figure 5.1: Schematic illustration of the scattering transform: the tree on the right represents the cascade of modulus of wavelet transforms; the output scattering coefficients are on the left.

For any $J \in \mathbb{Z}$, we set $\mathcal{P}_J = \{(j_1, ..., j_n), n \in \mathbb{N}, j_1, ..., j_n \in \{-\infty, ..., J\}\}$; we refer to the elements of $\mathcal{P}_J$ as *paths*. We denote the length (that is, the number of elements) of a path p by $|p|$.

For any $p \in \mathcal{P}_J$, we define:

$$S_J[p]f = U[p]f \star \phi_J$$

The *scattering coefficients* associated to f at scale J are the set $\{S_J[p]f\}_{p \in \mathcal{P}_J}$.

The computation of the scattering coefficients is schematized in Figure 5.1.

5.1.2 Norm preservation and energy propagation

When the wavelets are suitably chosen, the scattering transform preserves the norm [Mallat, 2012, Theorem 2.6]:

$$\sum_{p \in \mathcal{P}_J} ||S_J[p]f||_2^2 = ||f||_2^2 \tag{5.1}$$

In this paragraph, we briefly describe this result. To explain it, we introduce the informal idea that the repeated application of the modulus of the wavelet transform moves the energy contained in the high frequencies of the initial signal towards the low frequencies.

For the moment, we assume the wavelets to be analytical:

$$\hat{\psi}(\omega) = 0 \text{ if } \omega < 0$$

and, together with the low-pass filter ϕ, to satisfy a Littlewood-Paley condition:

$$\forall \omega \geq 0 \qquad |\hat{\phi}(\omega)|^2 + \frac{1}{2}\sum_{j \leq 0} |\hat{\psi}_j(\omega)|^2 = 1$$

This condition ensures that, for any scale $J \in \mathbb{Z}$, the wavelet transform with low-pass filter $f \to \{f \star \phi_J, \{f \star \psi_j\}_{j \leq J}\}$ is unitary over the set of real-valued functions:

$$\forall f \in L^2(\mathbb{R}, \mathbb{R}) \qquad ||f \star \phi_J||_2^2 + \sum_{j \leq J} ||f \star \psi_j||_2^2 = ||f||_2^2$$

Consequently, the application $W_J : f \to \{f \star \phi_J, \{|f \star \psi_j|\}_{j \leq J}\}$ preserves the norm. As the scattering transform is computed by recursively applying this operator to the input function, we can prove by iteration over n that, for any length $n \geq 0$:

$$\forall f \in L^2(\mathbb{R}, \mathbb{R}) \qquad \sum_{p \in \mathcal{P}_J, |p| < n} ||S_J[p]f||_2^2 + \sum_{p \in \mathcal{P}_J, |p| = n} ||U[p]f||_2^2 = ||f||_2^2$$

If we can prove that $\sum_{p \in \mathcal{P}_J, |p|=n} ||U[p]f||_2^2$ goes to zero when n goes to infinity, then we can prove Equation (5.1). Under an additional condition on the wavelets, this is done in [Mallat, 2012, Theorem 2.6]:

Theorem 5.1. *A family $(\phi, \{\psi_j\}_{j \in \mathbb{Z}})$ is said to be admissible if there exists $\eta \in \mathbb{R}$ and a positive real-valued function $\rho \in L^2(\mathbb{R})$ such that:*

$$\forall \omega \in \mathbb{R}, \quad |\hat{\rho}(\omega)| \leq |\hat{\phi}(2\omega)| \qquad \textit{and} \qquad \hat{\rho}(0) = 1$$

and that the function:

$$\hat{\Psi}(\omega) = |\hat{\rho}(\omega - \eta)|^2 - \sum_{k=1}^{+\infty} k(1 - |\hat{\rho}(2^{-k}(\omega - \eta))|^2)$$

satisfies:

$$\inf_{1 \leq \omega \leq 2} \sum_{j=-\infty}^{+\infty} \hat{\Psi}(2^j \omega) |\hat{\psi}_j(\omega)|^2 > 0 \tag{5.2}$$

If $(\phi, \{\psi_j\}_{j\in\mathbb{Z}})$ *is admissible, then, for any real-valued function* $f \in L^2(\mathbb{R})$ *and any scale* $J \in \mathbb{Z}$*:*

$$\sum_{p\in\mathcal{P}_J, |p|=n} ||U[p]f||_2^2 \to 0 \qquad \text{when } n \to +\infty \tag{5.3}$$

which implies the norm preservation:

$$\sum_{p\in\mathcal{P}_J} ||S_J[p]f||_2^2 = ||f||_2^2$$

Intuitively, the property (5.3) holds because the iterative application of the modulus of the wavelet transform moves the energy carried by the high frequencies of the signal towards the low frequencies. At each step of the scattering transform, the energy in the lowest frequency bands is output by convolution with ϕ_J. The remaining part is shifted towards the lower frequencies by a new application of the modulus of the wavelet transform and so on.

For the displacement of the energy towards the low frequencies, the modulus is essential: for each signal f, $f \star \psi_j$ is a function whose energy is concentrated in a frequency band of characteristic size 2^{-j}, with a mean frequency also of the order of 2^{-j}. After application of the modulus, $|f \star \psi_j|$ tends to have its energy concentrated in a frequency band of characteristic size still equal to 2^{-j}, but now centered around zero. So the frequencies that it contains are globally lower than the frequencies of $f \star \psi_j$. An example of this phenomenon is displayed on Figure 5.2.

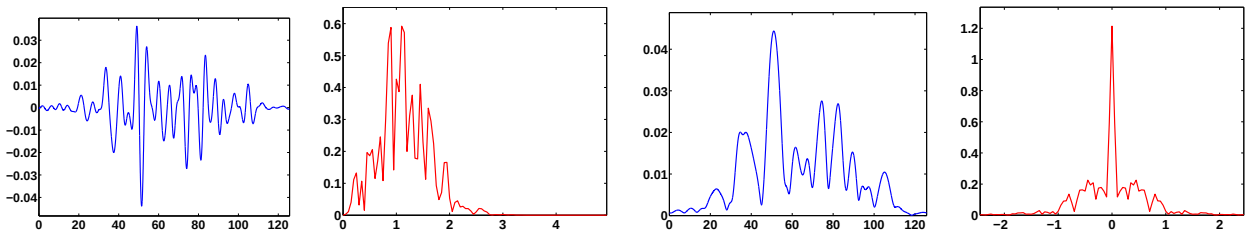
According to this simplistic reasoning, the modulus of the wavelet transform approximately moves the energy contained in a frequency band around 2^j to the frequency band $[-2^{j-1}, 2^{j-1}]$. By iteratively applying this argument, we expect the energy to arrive in the frequency band $[-2^J, 2^J]$ (and thus disappear) after a number of scattering steps proportional to j. The theorem of the next section will formalize this idea.

5.2 Theorem statement

The theorem of this section relies on similar ideas to the ones of Theorem 5.1 but improves it in two aspects:

- It gives a precise bound on $\sum_{p\in\mathcal{P}_J, |p|=n} ||U[p]f||_2^2$, instead of simply ensuring that this quantity goes to zero. This bound formalizes the idea described in the end of the previous paragraph, that the energy of the signal carried by the frequencies around 2^j disappears after a number of scattering step proportional to j.

- It holds for a much more general class of wavelets that Theorem 5.1. In particular, it does not require the wavelets to be analytical. It notably applies to Morlet wavelets, and also



(a) (b) (c) (d)

Figure 5.2: Illustration of the shift towards low frequencies due to the modulus. (a) $f \star \psi_j$ (real part) (b) $\widehat{f \star \psi_j}$ (in modulus) (c) $|f \star \psi_j|$ (d) $\widehat{|f \star \psi_j|}$. Observe that $\widehat{f \star \psi_j}$ and $\widehat{|f \star \psi_j|}$ are localized on frequency bands of the same width, but that, for $\widehat{|f \star \psi_j|}$, the band is centered around 0, and thus globally lower in frequency.

to compactly-supported wavelets like Selesnick wavelets [Selesnick, 2001], two important choices for practical applications. The wavelets have to satisfy three simple conditions, which will be commented after the statement of the theorem.

For any $a > 0$, we denote by χ_a the Gaussian function:

$$\forall t \in \mathbb{R} \qquad \chi_a(t) = \sqrt{\pi} a \exp\left(-(\pi a)^2 t^2\right)$$

whose Fourier transform satisfies:

$$\forall \omega \in \mathbb{R} \qquad \hat{\chi}_a(\omega) = \exp(-(\omega/a)^2)$$

Theorem 5.2. *Let $(\psi_j)_{j\in\mathbb{Z}}$ be a family of wavelets. We assume the following Littlewood-Paley inequality to be satisfied:*

$$\forall \omega \in \mathbb{R} \qquad \frac{1}{2}\sum_{j\in\mathbb{Z}} \left(|\hat{\psi}_j(\omega)|^2 + |\hat{\psi}_j(-\omega)|^2\right) \leq 1 \tag{5.4}$$

We also assume that:

$$\forall j \in \mathbb{Z}, \omega > 0 \qquad |\hat{\psi}_j(-\omega)| \leq |\hat{\psi}_j(\omega)| \tag{5.5}$$

and, for any ω, the inequality is strict for at least one value of j.

Finally, we assume that, for some $\epsilon > 0$:

$$\hat{\psi}(\omega) = O(|\omega|^{1+\epsilon}) \tag{5.6}$$

when $\omega \to 0$.

Then, for any $J \in \mathbb{Z}$, there exist $r > 0, a > 1$ such that, for all $n \geq 2$ and $f \in L^2(\mathbb{R}, \mathbb{R})$:

$$\sum_{p \in \mathcal{P}_J, |p|=n} ||U[p]f||_2^2 \leq ||f||_2^2 - ||f \star \chi_{ra^n}||_2^2 = \int_{\mathbb{R}} |\hat{f}(\omega)|^2 \left(1 - |\hat{\chi}_{ra^n}(\omega)|^2\right) d\omega \tag{5.7}$$

Before turning to the proof of this theorem, let us briefly comment the conditions (5.4), (5.5) and (5.6).

The Littlewood-Paley inequality (5.4) is equivalent to the fact that the wavelet transform is contractive over $L^2(\mathbb{R}, \mathbb{R})$. Without a condition of this kind, the wavelet transform can amplify some frequencies of the initial signal, and the energy contained in the long paths of length n will not necessarily decrease when n increases.

The condition (5.5) describes the fact that, although they do not need to be analytical, the wavelets must give more weight to the positive frequencies than to the negative ones. If this condition is not met, the phenomenon of energy shift towards the low frequencies may not happen. This is in particular the case if the wavelets are real.

The condition (5.6) states that the wavelets have a bit more than one zero momentum. It is necessary for our proof, but it is not clear whether the theorem stays true without it or not.

5.3 Principle of the proof

The proof proceeds by iteration over n.

In this paragraph, we show that, if a is close enough to 1, the fact that the property holds for n implies that it also holds for $n+1$. The initialization, for $n = 2$, relies on the same principle, but is more technical; it is done in the paragraph 5.5.3.

Lemma 5.3. *For any $j \in \mathbb{Z}, x \in \mathbb{R}_+^*, \delta_j \in \mathbb{R}$:*

$$|| \, |f \star \psi_j| \star \chi_x||_2^2 \geq ||f \star \psi_j \star (\chi_x e^{2\pi i \delta_j \cdot})||_2^2 = \int_{\mathbb{R}} |\hat{f}(\omega)|^2 |\hat{\psi}_j(\omega)|^2 |\hat{\chi}_x(\omega - \delta_j)|^2 d\omega$$

This lemma is already present in [Mallat, 2012]. For sake of completeness, we prove it again in the paragraph 5.5.1.

If the property (5.7) holds for $n \geq 2$, then:

$$\sum_{p \in \mathcal{P}_J, |p|=n+1} ||U[p]f||_2^2 = \sum_{j \leq J} \left(\sum_{p \in \mathcal{P}_J, |p|=n} ||U[j,p]f||_2^2 \right)$$

$$= \sum_{j\leq J}\left(\sum_{p\in\mathcal{P}_J,|p|=n}||U[p]\big(|f\star\psi_j|\big)||_2^2\right)$$
$$\leq \sum_{j\leq J}\Big(||\,|f\star\psi_j|\,||_2^2 - ||\,|f\star\psi_j|\star\chi_{ra^n}||_2^2\Big) \tag{5.8}$$

By Lemma 5.3, for any choice of $(\delta_j)_{j\leq J}$:

$$\sum_{p\in\mathcal{P}_J,|p|=n+1}||U[p]f||_2^2 \leq \sum_{j\leq J}\Big(||f\star\psi_j||_2^2 - ||f\star\psi_j\star(\chi_{ra^n}e^{2\pi i\delta_j})||_2^2\Big)$$
$$= \int_{\mathbb{R}}|\hat{f}(\omega)|^2\left(\sum_{j\leq J}|\hat{\psi}_j(\omega)|^2(1-|\hat{\chi}_{ra^n}(\omega-\delta_j)|^2)\right)d\omega$$

We symmetrize the expression, by taking into account the fact that f is real, so $|\hat{f}(\omega)| = |\hat{f}(-\omega)|$:

$$\sum_{p\in\mathcal{P}_J,|p|=n+1}||U[p]f||_2^2 \leq \int_{\mathbb{R}}|\hat{f}(\omega)|^2\times\frac{1}{2}\Big(\sum_{j\leq J}|\hat{\psi}_j(\omega)|^2(1-|\hat{\chi}_{ra^n}(\omega-\delta_j)|^2)$$
$$+|\hat{\psi}_j(-\omega)|^2(1-|\hat{\chi}_{ra^n}(-\omega-\delta_j)|^2)\Big)d\omega$$

To conclude, it is thus sufficient to show that, if the δ_j's are well-chosen, then:

$$\frac{1}{2}\Big(\sum_{j\leq J}|\hat{\psi}_j(\omega)|^2(1-|\hat{\chi}_{ra^n}(\omega-\delta_j)|^2)+|\hat{\psi}_j(-\omega)|^2(1-|\hat{\chi}_{ra^n}(-\omega-\delta_j)|^2)\Big)$$
$$\leq 1-|\hat{\chi}_{ar^{n+1}}(\omega)|^2 \qquad\qquad (\forall\omega\in\mathbb{R})$$

which is a direct consequence of the next lemma, proven in the paragraph 5.5.2.

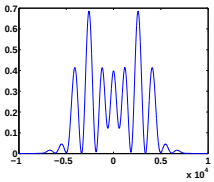
Lemma 5.4. *For any $x>0$, if $a>1$ is close enough to 1, then there exist $(\delta_j)_{j\leq J}$ such that:*

$$\forall\omega\in\mathbb{R}\qquad \frac{1}{2}\left(\sum_{j\in\mathbb{Z}}|\hat{\psi}_j(\omega)|^2(1-|\hat{\chi}_x(\omega-\delta_j)|^2)\right.$$
$$\left.+\sum_{j\in\mathbb{Z}}|\hat{\psi}_j(-\omega)|^2(1-|\hat{\chi}_x(-\omega-\delta_j)|^2)\right)\leq 1-|\hat{\chi}_{ax}(\omega)|^2$$

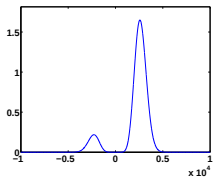
The proof is summarized in Figure 5.3.

Inductive hypothesis: the energy contained in the paths of length $n+1$ beginning by j is at most the energy of $f \star \psi_j$, multiplied by a high-pass filter (Equation (5.8)).

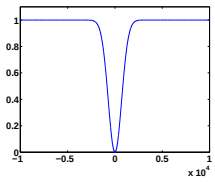
$$\sum_{|p|=n} ||U[j,p]f||_2^2 \quad \leq \int$$



$$\times$$



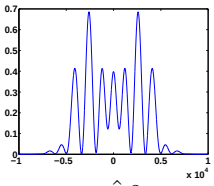
$$\times$$



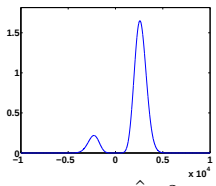
$$|\hat{f}|^2 \qquad |\hat{\psi}_j|^2 \qquad 1 - |\hat{\chi}_{ar^n}|^2$$

Because of the modulus around $f \star \psi_j$, the high-pass filter can be arbitrarily shifted in frequency.

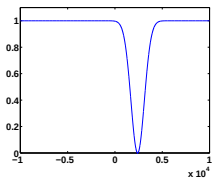
$$\sum_{|p|=n} ||U[j,p]f||_2^2 \quad \leq \int$$



$$\times$$

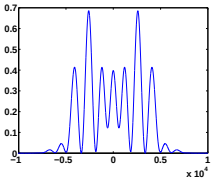


$$\times$$

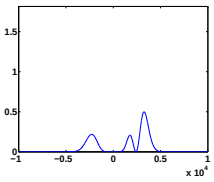


$$|\hat{f}|^2 \qquad |\hat{\psi}_j|^2 \qquad 1 - |\hat{\chi}_{ar^n}(. - \delta_j)|^2$$

$$= \int$$



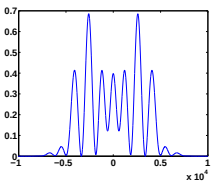
$$\times$$



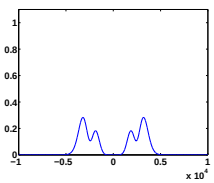
$$|\hat{f}|^2 \qquad |\hat{\psi}_j|^2(1 - |\hat{\chi}_{ar^n}(. - \delta_j)|^2)$$

We consider real signals so we can symmetrize the filter.

$$= \int$$



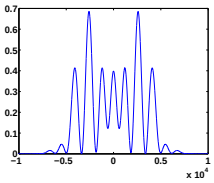
$$\times$$



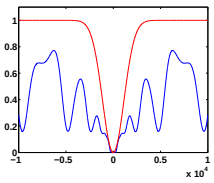
$$|\hat{f}|^2 \qquad \tfrac{1}{2}\left(|\hat{\psi}_j|^2(1 - |\hat{\chi}_{ar^n}(. - \delta_j)|^2) + |\hat{\psi}_j(-.)|^2(1 - |\hat{\chi}_{ar^n}(-. - \delta_j)|^2)\right)$$

Summing over j gives:

$$\sum_{|p|=n+1} ||U[p]f||_2^2 \quad \leq \int$$



$$\times$$



$$|\hat{f}|^2 \qquad \sum_{j \leq J} \tfrac{1}{2}\left(|\hat{\psi}_j|^2(1 - |\hat{\chi}_{ar^n}(. - \delta_j)|^2) + |\hat{\psi}_j(-.)|^2(1 - |\hat{\chi}_{ar^n}(-. - \delta_j)|^2)\right)$$

$$\leq \int \qquad |\hat{f}|^2 \qquad \times \qquad 1 - |\hat{\chi}_{ar^{n+1}}|^2 \qquad \text{(shown in red)}$$

Figure 5.3: Schematic summary of the proof of Theorem 5.2

5.4 Adaptation of the theorem to stationary processes

In what follows, X is a real-valued stationary stochastic process, with continuous and integrable autocovariance R_X.

The scattering transform of X is defined in the same way as for elements of $L^2(\mathbb{R})$:

$$\begin{gathered} U[\emptyset]X = X \\ \forall n \geq 1, (j_1, ..., j_n) \in \mathbb{Z}^n \qquad U[(j_1, ..., j_n)]X = |U[(j_1, ..., j_{n-1})]X \star \psi_{j_n}| \\ \forall p \in \mathcal{P}_J \qquad S_J[p]f = U[p]f \star \phi_J \end{gathered}$$

The results proven for the deterministic wavelet transform tend to be also valid for stationary processes, if one replaces the squared L^2-norms by the expectations of the squared modulus. In particular, Theorem 5.2 can be adapted to the case of stationary processes.

Theorem 5.5. *Let $(\psi_j)_{j\in\mathbb{Z}}$ be a family of wavelets, satisfying the same conditions (5.4), (5.5) and (5.6) as in Theorem 5.2.*

Then, for any $J \in \mathbb{Z}$, there exist $r > 0, a > 1$ such that, for all $n \geq 2$ and $f \in L^2(\mathbb{R}, \mathbb{R})$:

$$\sum_{p\in\mathcal{P}_J, |p|=n} \mathbb{E}(|U[p]X|_2^2) \leq \mathbb{E}(|X|^2) - \mathbb{E}(|X \star \chi_{ra^n}|^2) = \int_{\mathbb{R}} \hat{R}_X(\omega)(1 - |\hat{\chi}_{ra^n}(\omega)|^2)d\omega$$

The proof of this theorem is exactly the same as the one of Theorem 5.2. The only lines to be modified are the ones where we use the fact that the Fourier transform is an isometry of $L^2(\mathbb{R})$. We use instead the property according to which, for any stationary process with continuous and integrable autocovariance R_Y, and for any $h \in L^1 \cap L^2(\mathbb{R})$:

$$\begin{aligned} \mathbb{E}(|Y \star h|^2) &= |\mathbb{E}(Y \star h)|^2 + \int_{\mathbb{R}} |\hat{h}(\omega)|^2 \hat{R}_Y(\omega)d\omega \\ &= \left|\mathbb{E}(Y) \times \int_{\mathbb{R}} h(s)ds\right|^2 + \int_{\mathbb{R}} |\hat{h}(\omega)|^2 \hat{R}_Y(\omega)d\omega \end{aligned}$$

5.5 Proof of Lemmas

5.5.1 Proof of Lemma 5.3

Lemma (5.3). *For any $j \in \mathbb{Z}, x \in \mathbb{R}_+^*, \delta_j \in \mathbb{R}$:*

$$|| \, |f \star \psi_j| \star \chi_x||_2^2 \geq ||f \star \psi_j \star (\chi_x e^{2\pi i \delta_j \cdot})||_2^2 = \int_{\mathbb{R}} |\hat{f}(\omega)|^2 |\hat{\psi}_j(\omega)|^2 |\hat{\chi}_x(\omega - \delta_j)|^2 d\omega$$

Proof. For any t, because χ_x is a positive function:

$$\begin{aligned}
|f \star \psi_j| \star \chi_x(t) &= |(f \star \psi_j)e^{-2\pi i \delta_j \cdot}| \star \chi_x(t) \\
&= |(f \star \psi_j)e^{-2\pi i \delta_j \cdot}| \star |\chi_x|(t) \\
&\geq |((f \star \psi_j)e^{-2\pi i \delta_j \cdot}) \star \chi_x(t)| \\
&= |e^{-2\pi i \delta_j t}\big(f \star \psi_j \star (\chi_x \star e^{2\pi i \delta_j \cdot})\big)(t)| \\
&= |f \star \psi_j \star (\chi_x \star e^{2\pi i \delta_j \cdot})(t)|
\end{aligned}$$

This implies the inequality. The equality is a consequence of the unitarity of the Fourier transform. □

5.5.2 Proof of Lemma 5.4

Lemma (5.4)**.** *For any $x > 0$, if $a > 1$ is close enough to 1, then there exist $(\delta_j)_{j \leq J}$ such that:*

$$\begin{aligned}
\forall \omega \in \mathbb{R} \qquad & \frac{1}{2}\left(\sum_{j\in\mathbb{Z}} |\hat{\psi}_j(\omega)|^2(1 - |\hat{\chi}_x(\omega - \delta_j)|^2)\right. \\
& \left. + \sum_{j\in\mathbb{Z}} |\hat{\psi}_j(-\omega)|^2(1 - |\hat{\chi}_x(-\omega - \delta_j)|^2)\right) \leq 1 - |\hat{\chi}_{ax}(\omega)|^2
\end{aligned}$$

Proof. We are going to look for δ_j's of the form $\delta_j = \delta 2^{-j}$.

$$\begin{aligned}
&\frac{1}{2}\left(\sum_{j\in\mathbb{Z}} |\hat{\psi}_j(\omega)|^2(1 - |\hat{\chi}_x(\omega - \delta 2^{-j})|^2) + \sum_{j\in\mathbb{Z}} |\hat{\psi}_j(-\omega)|^2(1 - |\hat{\chi}_x(-\omega - \delta 2^{-j})|^2)\right) \\
=&\frac{1}{2}\left(\sum_{j\in\mathbb{Z}} |\hat{\psi}_j(\omega)|^2(1 - e^{2\left(-\frac{\omega^2}{x^2} + \frac{2\delta\omega 2^{-j}}{x^2} - \frac{\delta^2 2^{-2j}}{x^2}\right)}) + \sum_{j\in\mathbb{Z}} |\hat{\psi}_j(-\omega)|^2(1 - e^{2\left(-\frac{\omega^2}{x^2} - \frac{2\delta\omega 2^{-j}}{x^2} - \frac{\delta^2 2^{-2j}}{x^2}\right)})\right)
\end{aligned}$$

Let us define:

$$\forall \omega \in \mathbb{R}^* \qquad S(\omega) = \frac{1}{2}\left(\sum_{j\in\mathbb{Z}} |\hat{\psi}_j(\omega)|^2 + \sum_{j\in\mathbb{Z}} |\hat{\psi}_j(-\omega)|^2\right)$$

This function is never 0 if $\omega \neq 0$: if $\omega > 0$, there exist j such that $|\hat{\psi}_j(\omega)| > |\hat{\psi}_j(-\omega)| \geq 0$ so $S(\omega) > 0$; as S is even, we also have $S(\omega) > 0$ if $\omega < 0$.

The function $y \to 1 - e^{2\left(-\frac{\omega^2}{x^2} + y\right)}$ is concave so, for any $\omega \neq 0$:

$$\frac{1}{2}\left(\sum_{j\in\mathbb{Z}} |\hat{\psi}_j(\omega)|^2(1 - e^{2\left(-\frac{\omega^2}{x^2} + \frac{2\delta\omega 2^{-j}}{x^2} - \frac{\delta^2 2^{-2j}}{x^2}\right)}) + \sum_{j\in\mathbb{Z}} |\hat{\psi}_j(-\omega)|^2(1 - e^{2\left(-\frac{\omega^2}{x^2} - \frac{2\delta\omega 2^{-j}}{x^2} - \frac{\delta^2 2^{-2j}}{x^2}\right)})\right)$$

$$= S(\omega)\sum_{j\in\mathbb{Z}}\frac{\frac{1}{2}|\hat{\psi}_j(\omega)|^2}{S(\omega)}(1-e^{2\left(-\frac{\omega^2}{x^2}+\frac{2\delta\omega 2^{-j}}{x^2}-\frac{\delta^2 2^{-2j}}{x^2}\right)}) + \sum_{j\in\mathbb{Z}}\frac{\frac{1}{2}|\hat{\psi}_j(-\omega)|^2}{S(\omega)}(1-e^{2\left(-\frac{\omega^2}{x^2}-\frac{2\delta\omega 2^{-j}}{x^2}-\frac{\delta^2 2^{-2j}}{x^2}\right)})$$

$$\leq S(\omega)\left(1-\exp\left(2\left(-\frac{\omega^2}{x^2}+\sum_{j\in\mathbb{Z}}\frac{\frac{1}{2}|\hat{\psi}_j(\omega)|^2}{S(\omega)}\left(\frac{2\delta\omega 2^{-j}}{x^2}-\frac{\delta^2 2^{-2j}}{x^2}\right)\right.\right.\right.$$
$$\left.\left.\left.+\sum_{j\in\mathbb{Z}}\frac{\frac{1}{2}|\hat{\psi}_j(-\omega)|^2}{S(\omega)}\left(-\frac{2\delta\omega 2^{-j}}{x^2}-\frac{\delta^2 2^{-2j}}{x^2}\right)\right)\right)\right)$$

$$= S(\omega)\left(1-\exp\left(2\left(-\frac{\omega^2}{x^2}+\frac{2\delta\omega}{x^2}F_1(\omega)-\frac{\delta^2}{x^2}F_2(\omega)\right)\right)\right) \tag{5.9}$$

if we set:

$$F_1(\omega) = \sum_{j\in\mathbb{Z}}\left(\frac{1}{2}\frac{|\hat{\psi}_j(\omega)|^2-|\hat{\psi}_j(-\omega)|^2}{S(\omega)}\right)2^{-j}$$
$$F_2(\omega) = \sum_{j\in\mathbb{Z}}\left(\frac{1}{2}\frac{|\hat{\psi}_j(\omega)|^2+|\hat{\psi}_j(-\omega)|^2}{S(\omega)}\right)2^{-2j}$$

The functions F_1 and F_2 are both continuous. Indeed, for any j, $\hat{\psi}_j$ is continuous, because $\psi \in L^1(\mathbb{R})$. The function S is also continuous, and lower-bounded by a strictly positive constant. Finally, the sums converge uniformly on every compact subset of $\mathbb{R}$, because of the assumption (5.6). This implies the continuity.

Because of the hypothesis (5.5), $F_1(\omega) > 0$ when $\omega > 0$. Moreover, for any $\omega > 0$, $F_1(2\omega) = 2F_1(\omega)$. By a compacity argument, there exist $c > 0$ such that:

$$\forall\omega > 0 \qquad F_1(\omega) \geq c\omega$$

which, as F_1 is odd, implies:

$$\forall\omega \in \mathbb{R} \qquad \omega F_1(\omega) \geq c\omega^2 \tag{5.10}$$

Similarly, there exist $C > 0$ such that:

$$\forall\omega \in \mathbb{R} \qquad F_2(\omega) \leq C\omega^2 \tag{5.11}$$

By combining (5.10) and (5.11) with (5.9), we get that, for all $\omega \neq 0$:

$$\frac{1}{2}\left(\sum_{j\in\mathbb{Z}}|\hat{\psi}_j(\omega)|^2(1-|\hat{\chi}_x(\omega-\delta 2^{-j})|^2)+\sum_{j\in\mathbb{Z}}|\hat{\psi}_j(-\omega)|^2(1-|\hat{\chi}_x(-\omega-\delta 2^{-j})|^2)\right)$$

$$\leq S(\omega)\left(1-\exp\left(-2\frac{\omega^2}{x^2}(1-2c\delta+C\delta^2)\right)\right)$$

If we take $\delta = c/C$, this yields, for any a such that $1 < a \leq \frac{1}{\sqrt{1-c^2/C}}$:

$$\begin{aligned}
&\frac{1}{2}\left(\sum_{j\in\mathbb{Z}}|\hat{\psi}_j(\omega)|^2(1-|\hat{\chi}_x(\omega-\delta 2^{-j})|^2)+\sum_{j\in\mathbb{Z}}|\hat{\psi}_j(-\omega)|^2(1-|\hat{\chi}_x(-\omega-\delta 2^{-j})|^2)\right)\\
&\leq S(\omega)\left(1-\exp\left(-2\frac{\omega^2}{x^2}\left(1-\frac{c^2}{C}\right)\right)\right)\\
&\leq S(\omega)\left(1-\exp\left(-2\frac{\omega^2}{(ax)^2}\right)\right)\\
&\leq 1-\exp\left(-2\frac{\omega^2}{(ax)^2}\right)\\
&= 1-|\hat{\chi}_{ax}(\omega)|^2
\end{aligned}$$

The bound $S(\omega) \leq 1$ comes from the Littlewood-Paley inequality (5.4).

The inequalities are also true for $\omega = 0$ because all terms are then equal to zero. □

5.5.3 Initialization

In this paragraph, we prove that Theorem 5.2 holds for $n = 2$. More precisely, we prove that, for any $a > 1$, there exist $r > 0$ such that Equation (5.7) is valid for $n = 2$.

Proof. For any real-valued function $g \in L^2(\mathbb{R})$:

$$\begin{aligned}
\sum_{j\leq J}||g\star\psi_j||_2^2 &= \int_{\mathbb{R}}|\hat{g}(\omega)|^2\left(\sum_{j\leq J}|\hat{\psi}_j(\omega)|^2\right)d\omega\\
&= \int_{\mathbb{R}}|\hat{g}(\omega)|^2\left(\frac{1}{2}\sum_{j\leq J}|\hat{\psi}_j(\omega)|^2+|\hat{\psi}_j(-\omega)|^2\right)d\omega
\end{aligned} \tag{5.12}$$

If the following inequality held, for some $r > 0, a > 1$:

$$\frac{1}{2}\sum_{j\leq J}\left(|\hat{\psi}_j(\omega)|^2+|\hat{\psi}_j(-\omega)|^2\right)\leq 1-|\hat{\chi}_{ra}(\omega)|^2$$

then Theorem 5.2 would be valid for $n = 1$ and it could be proven for $n = 2$ in the exact same way as in Section 5.3. Unfortunately, this inequality is not necessarily valid (in particular, it is never the case if the wavelet transform is unitary and the wavelets are band-limited).

The next lemma (proven at the end of the paragraph) nevertheless shows that the inequality is satisfied, if one allows the Gaussian function to be replaced by a more general function.

Lemma 5.6. *There exist a real-valued positive function $\phi \in L^1 \cap L^2(\mathbb{R})$ such that:*

$$|\hat{\phi}(\omega)|^2 = 1 - O(\omega^2) \qquad \text{when } \omega \to 0 \tag{5.13}$$

and:

$$\forall \omega \in \mathbb{R} \qquad \frac{1}{2}\sum_{j \leq J} \left(|\hat{\psi}_j(\omega)|^2 + |\hat{\psi}_j(-\omega)|^2 \right) \leq 1 - |\hat{\phi}(\omega)|^2 \tag{5.14}$$

The rest of the proof consists in showing how to adapt Section 5.3, when the Gaussian function has been replaced by the ϕ of the lemma.

Because of (5.12), with ϕ defined as in Lemma 5.6, we have, for any real-valued function $g \in L^2(\mathbb{R})$:

$$\begin{aligned} \sum_{j \leq J} ||g \star \psi_j||_2^2 &\leq \int_{\mathbb{R}} |\hat{g}(\omega)|^2 (1 - |\hat{\phi}(\omega)|^2) \\ &= ||g||_2^2 - ||g \star \phi||_2^2 \end{aligned}$$

So:

$$\begin{aligned} \sum_{p \in \mathcal{P}_J, |p|=2} ||U[p]f||_2^2 &= \sum_{j_1 \leq J} \sum_{j_2 \leq J} || \, ||f \star \psi_{j_1}| \star \psi_{j_2}| \, ||_2^2 \\ &\leq \sum_{j \leq J} ||f \star \psi_j||_2^2 - || \, |f \star \psi_j| \star \phi||_2^2 \end{aligned} \tag{5.15}$$

Lemma 5.3 is still true when χ_x is replaced by ϕ, because the only property of χ_x which is really needed is its positivity. So for any $\delta \in \mathbb{R}$:

$$\forall j \in \mathbb{Z} \qquad || \, |f \star \psi_j| \star \phi||_2^2 \geq \int_{\mathbb{R}} |\hat{f}(\omega)|^2 |\hat{\psi}_j(\omega)|^2 |\hat{\phi}(\omega - \delta)|^2 d\omega$$

In Section 5.3, we used this same inequality, with a different δ for each value of j. Here, we do not need δ to vary as a function of j; however, we will need to consider different values of δ and to average the inequalities over these different values.

If we combine the last inequality with (5.15), we obtain:

$$\forall \delta \in \mathbb{R} \qquad \sum_{p \in \mathcal{P}_J, |p|=2} ||U[p]f||_2^2 \leq \int_{\mathbb{R}} |\hat{f}(\omega)|^2 (1 - |\hat{\phi}(\omega - \delta)|^2) \left(\sum_{j \leq J} |\hat{\psi}_j(\omega)|^2 \right) d\omega$$

As it holds for any $\delta \in \mathbb{R}$, we have, for any positive $c \in L^1(\mathbb{R})$ whose integral over $\mathbb{R}$ is 1:

$$\sum_{p\in\mathcal{P}_J,|p|=2} ||U[p]f||_2^2 \leq \int_{\mathbb{R}} c(\delta) \int_{\mathbb{R}} |\hat{f}(\omega)|^2(1-|\hat{\phi}(\omega-\delta)|^2)\left(\sum_{j\leq J}|\hat{\psi}_j(\omega)|^2\right) d\omega d\delta$$
$$\leq \int_{\mathbb{R}} |\hat{f}(\omega)|^2(1-|\hat{\phi}|^2 \star c(\omega))\left(\sum_{j\leq J}|\hat{\psi}_j(\omega)|^2\right) d\omega$$

We limit ourselves to the case where c is even. As $|\hat{\phi}|^2$ is also even, the last inequality yields, by using the fact that f is real:

$$\sum_{p\in\mathcal{P}_J,|p|=2} ||U[p]f||_2^2 \leq \int_{\mathbb{R}} |\hat{f}(\omega)|^2(1-|\hat{\phi}|^2 \star c(\omega)) \times \frac{1}{2}\left(\sum_{j\leq J}|\hat{\psi}_j(\omega)|^2 + |\hat{\psi}_j(-\omega)|^2\right) d\omega$$

The conclusion comes from a last lemma, proven at the end of the paragraph.

Lemma 5.7. *If we take $c(\delta) = \frac{1}{\sqrt{\pi}}\exp(-\delta^2)$, then there exists $x > 0$ such that:*

$$\forall \omega \in \mathbb{R} \qquad (1-|\hat{\phi}|^2 \star c(\omega)) \times \frac{1}{2}\left(\sum_{j\leq J}|\hat{\psi}_j(\omega)|^2 + |\hat{\psi}_j(-\omega)|^2\right) \leq 1 - |\hat{\chi}_x(\omega)|^2$$

□

We now give the proofs of Lemmas 5.6 and 5.7.

Proof of Lemma 5.6. Let $\gamma : \mathbb{R} \to \mathbb{R}$ be any even, rapidly decreasing and band-limited function such that $\int_{\mathbb{R}} \gamma^2 = 1$. We set $\phi_0 = \gamma^2$. This is also an even, rapidly decreasing and band-limited function.

Because $\int_{\mathbb{R}} \gamma^2 = 1$, we have $\hat{\phi}_0(0) = 1$. As ϕ_0 is even, $\hat{\phi}_0'(0) = 0$. But the second derivative is strictly negative: $\hat{\phi}_0''(0) = -(2\pi)^2 \int_{\mathbb{R}} t^2\phi_0(t)dt < 0$. We deduce from these relations that there exists $\alpha > 0$ such that:

$$|\hat{\phi}_0(\omega)|^2 = 1 - \alpha\omega^2 + o(\omega^2) \qquad \text{when } \omega \to 0 \tag{5.16}$$

Let us show that, for M large enough, $\phi : t \to M^{-1}\phi_0(M^{-1}t)$ satisfies the desired properties. By construction, it is a real-valued positive function. It is rapidly decreasing, so $\phi \in L^1 \cap L^2(\mathbb{R})$. The property (5.13) holds; only the inequality (5.14) is left to prove.

Because of Equation (5.16) and because $\hat{\phi}_0$ is compactly-supported, there exists $\tilde{\alpha} > 0$ such that for all $\omega \in \text{Supp}(\hat{\phi}_0)$:

$$|\hat{\phi}_0(\omega)|^2 \leq 1 - \tilde{\alpha}\omega^2$$

By the assumption (5.6), $|\hat{\psi}(\omega)| = O(|\omega|^{1+\epsilon})$ when $\omega \to 0$, for some $\epsilon > 0$. This implies that:

$$\frac{1}{2}\sum_{j\leq J}\left(|\hat{\psi}_j(\omega)|^2 + |\hat{\psi}_j(-\omega)|^2\right) = o(\omega^2) \qquad \text{when } \omega \to 0$$

Because this sum is moreover bounded by 1 on all $\mathbb{R}$, there exists $A > 0$ such that:

$$\forall \omega \in \mathbb{R} \qquad \frac{1}{2}\sum_{j\leq J}\left(|\hat{\psi}_j(\omega)|^2 + |\hat{\psi}_j(-\omega)|^2\right) \leq A\omega^2$$

If $M \geq \sqrt{A/\tilde{\alpha}}$, then, on the support of $\hat{\phi}$:

$$\begin{aligned}
&|\hat{\phi}(\omega)|^2 + \frac{1}{2}\sum_{j\leq J}\left(|\hat{\psi}_j(\omega)|^2 + |\hat{\psi}_j(-\omega)|^2\right) \\
&= |\hat{\phi}_0(M\omega)|^2 + \frac{1}{2}\sum_{j\leq J}\left(|\hat{\psi}_j(\omega)|^2 + |\hat{\psi}_j(-\omega)|^2\right) \\
&\leq 1 - \tilde{\alpha}M^2\omega^2 + A\omega^2 \\
&\leq 1
\end{aligned}$$

Outside the support of $\hat{\phi}$, the inequality is also true, because of the Littlewood-Paley condition (5.4). So Equation (5.14) holds. □

Proof of Lemma 5.7. Let us define:

$$F(\omega) = (1 - |\hat{\phi}|^2 \star c(\omega)) \times \frac{1}{2}\left(\sum_{j\leq J}|\hat{\psi}_j(\omega)|^2 + |\hat{\psi}_j(-\omega)|^2\right)$$

We are going to prove that F has the following three properties:

1. $\forall \omega \in \mathbb{R}, F(\omega) < 1$
2. $F(\omega) = O(\omega^2)$ when $\omega \to 0$
3. There exists $x > 0$ such that $F(\omega) \leq 1 - |\hat{\chi}_x(\omega)|^2$ if $|\omega|$ is large enough.

These three properties imply that F is bounded by $1 - |\hat{\chi}_x|^2$ on all $\mathbb{R}$, for x small enough. This assertion relies on a compacity argument; as it is relatively straightforward, we do not prove it.

The first property is an immediate consequence of the Littlewood-Paley inequality (5.4) and of the fact that $|\hat{\phi}|^2 \star c > 0$ over $\mathbb{R}$.

The second one is a consequence of a fact that has been explained in the proof of Lemma 5.6:

$$\frac{1}{2}\left(\sum_{j\leq J}|\hat{\psi}_j(\omega)|^2+|\hat{\psi}_j(-\omega)|^2\right)=o(\omega^2)\qquad\text{when }\omega\to 0$$

For the last one, we remark that, for any $\omega\in\mathbb{R}$ such that $\omega\geq 1$:

$$\begin{aligned}|\hat{\phi}|^2\star c(\omega)&=\int_{\mathbb{R}}|\hat{\phi}(\delta)|^2c(\omega-\delta)d\delta\\&\geq\int_0^1|\hat{\phi}(\delta)|^2c(\omega-\delta)d\delta\\&\geq c(\omega)\int_0^1|\hat{\phi}(\delta)|^2d\delta\end{aligned}$$

As all the functions are even, this is also true for $\omega\leq -1$. So when $|\omega|$ is large enough:

$$|\hat{\phi}|^2\star c(\omega)\geq\frac{1}{\sqrt{\pi}}\exp(-\omega^2)\left(\int_0^1|\hat{\phi}(\delta)|^2d\delta\right)\geq\exp(-2\omega^2)=|\hat{\chi}_1(\omega)|^2$$

This proves the third property and concludes. □

Chapter 6

Generalized scattering

In this last chapter, we define a generalization of the scattering transform for stationary processes, by keeping the structure in cascade described in Section 5.1 but replacing the wavelet transform by arbitrary linear functions.

The resulting operator takes as input a random process X_0 and, from it, defines a sequence of processes by the iterative application of the following operation:

$$X_n \to X_{n+1} = |W_n X_n - \mathbb{E}(W_n X_n)| \tag{6.1}$$

where the W_n are arbitrary linear operators and $|.|$ denotes the pointwise modulus. The scattering coefficients are, by definition, the $\mathbb{E}(W_n X)$ for $n \in \mathbb{N}$.

The interest of this generalized scattering is to offer a mathematical framework for the theoretical analysis of deep neural networks. Indeed, it models the architecture of a neural network, with the modulus used as non-linearity.

The first questions that arise are again: for a given choice of W_n, what information does the generalized scattering preserve about the input process? What do scattering coefficients say about the regularity of the input process? Which processes have the same scattering coefficients? For neural networks, these questions have been empirically studied in [Simonyan et al., 2014; Mahendran and Vedaldi, 2015; Dosovitskiy and Brox, 2015].

In the future, we would also be interested in studying the learning of the linear operators W_n. Indeed, while, in the regular scattering, the W_n are fixed (they represent the wavelet transform), they can here be learned from the data. We are especially interested in unsupervised learning: how to adapt the W_n to the law of the input random process, so that the resulting scattering operator yields a relevant representation of realizations of this process? The question of unsupervised learning has attracted a great deal of attention in the last ten years [Bengio et al., 2013]. On large-scale datasets, it tends to be supplanted by purely supervised learning [Russakovsky et al., 2015], but it is competitive when labeled data is less abundant or does not

have a clear spatial or temporal structure [Sermanet et al., 2013; Chen et al., 2014]. It could possibly be further improved.

In this chapter, we focus on the first body of questions, and describe a few properties of the initial process that can be retrieved from the scattering coefficients. In finite dimension, we prove that the scattering operator preserves the norm when the W_n are unitary (Theorem 6.2). We study in more detail the one-dimensional case, where all the X_n take their values in $\mathbb{R}$. We show that the scattering coefficients characterize the tail of the distribution of X_0 (Theorem 6.4), but not necessarily the distribution in the low values.

In Section 6.1, we define the generalized scattering, and explain its link with the regular scattering. In Section 6.2, we prove the energy preservation theorem. Section 6.3 is devoted to the one-dimensional case. We conclude with numerical experiments in Section 6.4.

The first two sections of this chapter come from [Mallat and Waldspurger, 2013].

6.1 Definition of the generalized scattering

In this paragraph, we explain the generalization of the regular scattering as defined in [Mallat, 2012] which leads to the definition that we are going to study.

We consider a finite-dimensional version of the regular scattering, where signals are of finite dimension N and we use only a finite number J of wavelets. This is the setting of all concrete applications.

The scattering of a stationary process X_0 is written as the following sequence of computations:

$$X_0 \in \mathbb{R}^N$$
$$\forall n \geq 0 \qquad X_{n+1} = |W_n X_n| \in (\mathbb{R}^N)^{J^{n+1}}$$

where W_n is the wavelet transform operator:

$$W_n : (g_s)_{1 \leq s \leq J^n} \in (\mathbb{R}^N)^{J^n} \quad \rightarrow \quad (g_s \star \psi_j)_{1 \leq s \leq J^n, 1 \leq j \leq J} \in (\mathbb{C}^N)^{J^{n+1}}$$

and the scattering coefficients are the $A_n X_n, n \in \mathbb{N}$, where A_n is the averaging operator:

$$A_n : (g_s)_{1 \leq s \leq J^n} \in (\mathbb{R}^N)^{J^n} \quad \rightarrow \quad (g_s \star \phi_J)_{1 \leq s \leq J^n} \in (\mathbb{R}^N)^{J^n}$$

When J is large enough, $g_s \star \phi_J$ is approximately the spatial mean of g_s:

$$\forall s, \forall k = 1, ..., N \qquad (g_s \star \phi_J)_k \approx \frac{1}{N} \sum_{l=1}^{N} (g_s)_l \overset{def}{=} \text{mean}(g_s)$$

If X is a sufficiently ergodic stationary process, then the spatial mean is an approximation of the expectation:

$$\forall k = 1, ..., N \qquad \text{mean}(g_s) \approx \mathbb{E}((g_s)_k)$$

So the scattering coefficients are approximately the set of $\mathbb{E}(X_n), n \in \mathbb{N}$.

If the wavelets have been well-chosen, the wavelet transform has the set of constant signals as kernel, and it is unitary on the set of signals with mean zero. So:

$$X_{n+1} = |W_n X_n| = |W_n(X_n - \text{mean}(X_n))| \approx |W_n(X_n - \mathbb{E}(X_n))|$$

And we arrive at Definition (6.1).

We summarize this definition. Let an increasing sequence $(N_n)_{n\in\mathbb{N}}$ of integers be fixed. For each n, let $W_n : \mathbb{R}^{N_n} \to \mathbb{R}^{N_{n+1}}$ be a linear unitary operator. Starting from a random process X_0, taking its values in $\mathbb{R}^{N_0}$, we iteratively define:

$$X_{n+1} = |W_n(X_n - \mathbb{E}(X_n))| \tag{6.2}$$

The scattering coefficients of X_0 are defined as $\{\mathbb{E}(X_n)\}_{n\in\mathbb{N}}$.

6.2 Energy preservation

6.2.1 One-dimensional case

We shall prove an energy preservation result. To simplify the exposition of the proof, we begin with the one-dimensional case: X_0 is a real-valued random variable, with finite second-order moment.

In dimension 1, there are only two unitary operators: Id or $-Id$. The modulus removes the sign, so the generalized scattering reduces to:

$$\forall n \geq 0 \qquad X_{n+1} = |X_n - \mathbb{E}(X_n)|$$

From the definition, we immediately see that:

$$\forall n \geq 0 \qquad \mathbb{E}(X_{n+1}^2) = \mathbb{E}(X_n^2) - \mathbb{E}(X_n)^2$$

So, for any $N \geq 0$:

$$\mathbb{E}(X_0^2) - \mathbb{E}(X_N^2) = \sum_{n=0}^{N-1} \mathbb{E}(X_n)^2 \tag{6.3}$$

Our first result is that this equality goes to the limit $N \to +\infty$: the l^2-norm of the sequence of scattering coefficients $(\mathbb{E}(X_n))_{n\in\mathbb{N}}$ is the norm of X_0.

Theorem 6.1. *Let X_0 be such that $\mathbb{E}(||X_0||^2) < +\infty$. Then:*

$$\mathbb{E}(X_0^2) = \sum_{n=0}^{+\infty} \mathbb{E}(X_n)^2$$

Proof. From Equation (6.3), it is enough to show that $\mathbb{E}(X_n^2)$ goes to zero when n goes to infinity. This equation also implies that $\mathbb{E}(X_n) \to 0$ when $n \to +\infty$.

For any $n \geq 1$, X_n is positive. So if the X_n's were uniformly bounded by a constant $M > 0$, we would have, when $n \to +\infty$:

$$\big(\mathbb{E}(X_n) \to 0\big) \quad \Rightarrow \quad \big(\mathbb{E}(X_n^2) \leq M\mathbb{E}(X_n) \to 0\big)$$

Here, the processes X_n may not be uniformly bounded. However, they are uniformly bounded on an event of arbitrarily small probability.

Indeed, let us fix any $M > \mathbb{E}(X_0^2)$.

For any $n \geq 1$, $X_n \geq 0$ so:

$$X_{n+1}^2 = |X_n - \mathbb{E}(X_n)|^2 \leq X_n^2 + \mathbb{E}(X_n)^2$$

which implies (using Equation (6.3) for the second inequality):

$$\forall n \geq 1, \qquad X_n^2 \leq X_1^2 + \sum_{s=1}^{n-1} \mathbb{E}(X_s)^2 \leq X_1^2 + \mathbb{E}(X_0)^2 \leq X_1^2 + M^2$$

We thus have the inclusion, for any $n \geq 1$:

$$\{X_1 \leq M\} \subset \{X_n \leq \sqrt{2}M\}$$

and:

$$\begin{aligned}
\mathbb{E}(X_n^2) &= \mathbb{E}(X_n^2 1_{X_n \leq \sqrt{2}M}) + \mathbb{E}(X_n^2 1_{X_n > \sqrt{2}M}) \\
&\leq \sqrt{2}M\mathbb{E}(X_n 1_{X_n \leq \sqrt{2}M}) + \mathbb{E}((X_1^2 + M^2) 1_{X_n > \sqrt{2}M}) \\
&\leq \sqrt{2}M\mathbb{E}(X_n) + \mathbb{E}((X_1^2 + M^2) 1_{X_1 > M}) \\
&\leq \sqrt{2}M\mathbb{E}(X_n) + 2\mathbb{E}(X_1^2 1_{X_1 > M})
\end{aligned}$$

Because $\mathbb{E}(X_n) \to 0$ when $n \to +\infty$:

$$\limsup_{n \to \infty} \mathbb{E}(X_n^2) \leq 2\mathbb{E}(X_1^2 1_{X_1 > M})$$

Letting M go to infinity proves:

$$\mathbb{E}(X_n^2) \to 0 \qquad \text{when } n \to +\infty$$

□

6.2.2 Generalization to higher dimensions

We now remove the assumption of dimension one and return to the generalized scattering as defined in Equation (6.2):

$$X_{n+1} = |W_n(X_n - \mathbb{E}(X_n))|$$

Theorem 6.2. *Let X_0 be such that $\mathbb{E}(||X_0||^2) < +\infty$. Then:*

$$\mathbb{E}(||X_0||^2) = \sum_{n=0}^{+\infty} ||\mathbb{E}(X_n)||^2$$

Proof. The proof follows the one of Theorem 6.1.

We first remark that, for any $n \in \mathbb{N}$:

$$\begin{aligned}\mathbb{E}(||X_{n+1}||^2) &= \mathbb{E}(||W_n(X_n - \mathbb{E}(X_n))||^2) \\ &= \mathbb{E}(||X_n - \mathbb{E}(X_n)||^2) \\ &= \mathbb{E}(||X_n||^2) - ||\mathbb{E}(X_n)||^2\end{aligned}$$

which implies a formula analog to Equation (6.3):

$$\forall N \geq 0 \qquad \mathbb{E}(||X_0||^2) - \mathbb{E}(||X_N||^2) = \sum_{n=0}^{N-1} ||\mathbb{E}(X_n)||^2$$

To prove the theorem, it suffices to show that $\mathbb{E}(||X_n||^2) \to 0$ when $n \to +\infty$. We temporarily admit a lemma:

Lemma 6.3. $\mathbb{E}(||X_n||) \to 0$ *when* $n \to +\infty$

From this lemma, we can deduce the result. As the reasoning is the same as in the one-dimensional case, we only review its main steps.

We start by fixing $M > \mathbb{E}(||X_0||^2)$. For any $n \geq 1$:

$$||X_n||^2 \leq ||X_1||^2 + M^2$$

from which we deduce:

$$\mathbb{E}(||X_n||^2) \leq \sqrt{2}M\mathbb{E}(||X_n||) + 2\mathbb{E}(||X_1^2||1_{||X_1||>M})$$

Letting n then M go to infinity yields:

$$\mathbb{E}(||X_n||^2) \to 0 \qquad \text{when } n \to +\infty$$

This is what was needed; only Lemma 6.3 is left to show. □

Proof of Lemma 6.3. We deduce the convergence $\mathbb{E}(||X_n||) \to 0$ from the weaker convergence $||\mathbb{E}(X_n)|| \to 0$, which comes from Equation (6.2.2). The proof relies on a compacity argument, and strongly uses the fact that X_0 takes its values in a finite-dimensional space.

For any $z \in \mathbb{R}^{N_0}$ and $a > 0$, we denote by $B(z, a)$ the set $\{z' \in \mathbb{R}^{N_0} \text{ s.t. } ||z' - z|| < a\}$.

Let us fix $M > \mathbb{E}(||X_0||^2)$, and $\epsilon > 0$. Let $z_1, ..., z_S$ be a finite number of elements of $\mathbb{R}^{N_0}$ such that:

$$B(0, M) \subset \bigcup_{s=1}^{S} B(z_s, \epsilon)$$

We iteratively define:

$$\phi_0 : \text{Id} : \mathbb{R}^{N_0} \to \mathbb{R}^{N_0}$$
$$\forall n \geq 0, \qquad \phi_{n+1} : x \in \mathbb{R}^{N_0} \to |W_n(\phi_n(x) - \mathbb{E}(X_n))| \in \mathbb{R}^{N_{n+1}}$$

so that, for any $n \in \mathbb{N}$, $X_n = \phi_n(X_0)$.

These applications are all 1-Lipschitz: they are compositions of 1-Lipschitz operators. So if we denote by $\alpha : \mathbb{R}^{N_0} \to \{z_1, ..., z_s\}$ any measurable function such that $||z - \alpha(z)|| < \epsilon$ for any $z \in B(0, M)$, we have:

$$\begin{aligned}
\mathbb{E}(||X_n||) &= \mathbb{E}(||X_n||1_{||X_0||<M}) + \mathbb{E}(||X_n||1_{||X_0||\geq M}) \\
&= \mathbb{E}(||\phi_n(X_0)||1_{||X_0||<M}) + \mathbb{E}(||X_n||1_{||X_0||\geq M}) \\
&\leq \mathbb{E}\left(\left(||\phi_n(\alpha(X_0))|| + ||\phi_n(\alpha(X_0)) - \phi_n(X_0)||\right)1_{||X_0||<M}\right) + \mathbb{E}(||X_n||1_{||X_0||\geq M}) \\
&\leq \mathbb{E}\left(\left(||\phi_n(\alpha(X_0))|| + ||\alpha(X_0) - X_0||\right)1_{||X_0||<M}\right) + \mathbb{E}(||X_n||1_{||X_0||\geq M}) \\
&\leq \mathbb{E}(||\phi_n(\alpha(X_0))||1_{||X_0||<M}) + \mathbb{E}(||X_n||1_{||X_0||\geq M}) + \epsilon \\
&= \sum_{s=1}^{S} ||\phi_n(z_s)||\mathbb{P}\left(||X_0|| < M \text{ and } \alpha(X_0) = z_s\right) + \mathbb{E}(||X_n||1_{||X_0||\geq M}) + \epsilon \\
&= \sum_{s=1}^{S} ||\mathbb{E}(\phi_n(z_s)1_{||X_0||<M \text{ and } \alpha(X_0)=z_s})|| + \mathbb{E}(||X_n||1_{||X_0||\geq M}) + \epsilon \\
&\leq \sum_{s=1}^{S} ||\mathbb{E}(\phi_n(X_0)1_{||X_0||<M \text{ and } \alpha(X_0)=z_s})|| + \mathbb{E}(||X_n||1_{||X_0||\geq M}) + 2\epsilon \\
&\leq \sum_{s=1}^{S} ||\mathbb{E}(\phi_n(X_0)1_{||X_0||<M \text{ and } \alpha(X_0)=z_s})|| + \sqrt{5}\mathbb{E}(||X_0||1_{||X_0||\geq M}) + 2\epsilon
\end{aligned}$$

where the last inequality comes from the fact that $||X_1||^2 = ||X_0 - \mathbb{E}(X_0)||^2 \leq 2||X_0||^2 + 2||\mathbb{E}(X_0)||^2 \leq 2||X_0||^2 + 2M$, so $||X_n||^2 \leq ||X_1||^2 + M \leq 2||X_0||^2 + 3M$, and the one before from the fact that ϕ_n is 1-Lipschitz.

Moreover, for any $n \geq 1$, using the positivity of the coordinates of X_n:

$$\begin{aligned}
||\mathbb{E}(X_n)|| &\geq ||\mathbb{E}(X_n 1_{||X_0||<M})|| \\
&= ||\mathbb{E}(\phi_n(X_0) 1_{||X_0||<M})|| \\
&= \left|\left| \sum_{s=1}^{S} \mathbb{E}(\phi_n(X_0) 1_{||X_0||<M \text{ and } \alpha(X_0)=z_s}) \right|\right| \\
&\geq \frac{1}{\sqrt{S}} \sum_{s=1}^{S} ||\mathbb{E}(\phi_n(X_0) 1_{||X_0||<M \text{ and } \alpha(X_0)=z_s})||
\end{aligned}$$

As $||\mathbb{E}(X_n)|| \to 0$ when $n \to +\infty$, this result implies:

$$\begin{aligned}
&\sum_{s=1}^{S} ||\mathbb{E}(\phi_n(X_0) 1_{||X_0||<M \text{ and } \alpha(X_0)=z_s})|| \to 0 \qquad \text{when } n \to +\infty \\
\Rightarrow \qquad &\limsup_n \mathbb{E}(||X_n||) \leq \sqrt{5}\mathbb{E}(||X_0|| 1_{||X_0|| \geq M}) + 2\epsilon
\end{aligned}$$

Letting M and ϵ go respectively to $+\infty$ and 0 implies:

$$\mathbb{E}(||X_n||) \to 0 \qquad \text{when } n \to +\infty$$

□

6.3 Characterization of the distribution tail

In this section, we focus on the one-dimensional case:

$$\forall n \geq 0 \qquad X_{n+1} = |X_n - \mathbb{E}(X_n)|$$

We assume that the process X_0 takes only positive values and is not almost surely bounded. We show that the sequence of generalized scattering coefficients $(\mathbb{E}(X_n))_{n \in \mathbb{N}}$ precisely describes the distribution tail of X_0.

We will not tackle the subject under this point of view, but it is somewhat related to the problem of estimating a probability density function from samples of the corresponding distribution [Scott, 2015; Antoniadis, 1997].

We define:

$$\begin{aligned}
f : \mathbb{R}^+ &\to \mathbb{R}^+ \\
y &\to \mathbb{E}((X_0 - y) 1_{X_0 \geq y})
\end{aligned} \tag{6.4}$$

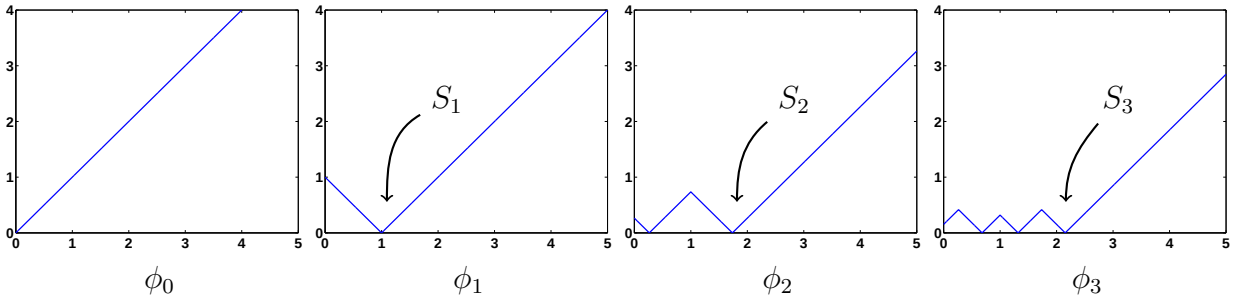


Figure 6.1: The functions $\phi_0, \phi_1, \phi_2, \phi_3$, for a Laplacian probability distribution $p(x) = e^{-x}1_{x\geq 0}$.

This decreasing function fully characterizes the distribution of X_0. The next theorem shows that the knowledge of $(\mathbb{E}(X_n))_{n\in\mathbb{N}}$ gives an equivalent of this function in $+\infty$. So the scattering coefficients precisely describe the distribution of X_0 in the high values, although they do not uniquely determine the law of X_0.

Theorem 6.4. *When $n \to +\infty$, if X_0 takes positive values and is not almost surely bounded:*

$$\mathbb{E}(X_n) \sim 2f\left(\sum_{s=0}^{n-1}\mathbb{E}(X_s)\right)$$

The sequence $(\mathbb{E}(X_n))_{n\in\mathbb{N}}$ thus provides an equivalent of f at infinity, because $\sum_{s=0}^{n-1}\mathbb{E}(X_s)$ goes to infinity with n (it is proven at the beginning of Paragraph 6.3.4).

6.3.1 Proof

To simplify the notations, we set:

$$\forall n \in \mathbb{N} \qquad S_n = \sum_{s=0}^{n-1}\mathbb{E}(X_s)$$

We iteratively define:

$$\phi_0 = \mathrm{Id} : \mathbb{R}^+ \to \mathbb{R}^+$$
$$\forall n \geq 0 \qquad \phi_{n+1} : x \in \mathbb{R}^+ \to |\phi_n(x) - \mathbb{E}(X_n)| \in \mathbb{R}^+$$

By definition, we have $X_n = \phi_n(X_0)$, and in particular $\mathbb{E}(X_n) = \mathbb{E}(\phi_n(X_0))$.

Figure 6.1 illustrates this definition. For n large enough, we see that ϕ_n has two distinct parts: on $[0; S_n]$, it is an oscillating function with small values and, on $[S_n; +\infty[$, we have $\phi_n(x) = x - S_n$

(this is proven by iteration over n). The value of $\mathbb{E}(X_n)$ is the sum of the expectations of these two parts:

$$\mathbb{E}(X_n) = \mathbb{E}(\phi_n(X_0)) = \mathbb{E}(\phi_n(X_0)1_{X_0 \leq S_n}) + \mathbb{E}(\phi_n(X_0)1_{X_0 > S_n})$$

Informally, at infinity, these two terms balance each other, so when n goes to $+\infty$:

$$\mathbb{E}(X_n) \sim 2\mathbb{E}(\phi_n(X_0)1_{X_0 > S_n}) = 2\mathbb{E}((X_0 - S_n)1_{X_0 > S_n}) = 2f(S_n)$$

To rigorously prove this equivalence, we begin with upper and lower bounds for $\mathbb{E}(X_n)$. We define:

$$\forall n \geq 0 \qquad m_n = \max_{x \in [0; S_n]} \phi_n(x)$$

The next lemma is proven in Paragraph 6.3.2.

Lemma 6.5. *For any $n \geq 1$:*

$$2f(S_n) \leq \mathbb{E}(X_n) \leq 2f(S_n) + 2\max(0, m_{n-1} - \mathbb{E}(X_{n-1})) \tag{6.5}$$

To establish Theorem 6.4, we have to bound the error term $2\max(0, m_{n-1} - \mathbb{E}(X_{n-1}))$.

We first list a few useful properties, which can be directly deduced from the definitions. Their detailed proof is in Paragraph 6.3.3.

Lemma 6.6. *For all $n \geq 0$:*

$$m_{n+1} = \max(m_n - \mathbb{E}(X_n), \mathbb{E}(X_n)) \tag{6.6}$$

and:

$$\mathbb{E}(X_{n+1}) \leq 2\mathbb{E}(X_n) \tag{6.7}$$

Moreover, if $\mathbb{E}(X_n) \geq \mathbb{E}(X_{n+1})$, then:

$$\mathbb{E}(X_{n+2}) \leq \mathbb{E}(X_{n+1}) \left(2 - \frac{\mathbb{E}(X_{n+1})}{\mathbb{E}(X_n)}\right) \tag{6.8}$$

If $\mathbb{E}(X_n) \leq \mathbb{E}(X_{n+1})$, then:

$$\mathbb{E}(X_{n+1}) \geq \mathbb{E}(X_{n+2}) \geq \mathbb{E}(X_{n+1}) \left(\frac{\mathbb{E}(X_{n+1})}{\mathbb{E}(X_n)} - 1\right) \tag{6.9}$$

and:

$$\mathbb{E}(X_{n+3}) \leq \mathbb{E}(X_{n+1}) - (\mathbb{E}(X_{n+1}) - \mathbb{E}(X_{n+2})) \left(\frac{\mathbb{E}(X_{n+1})}{\mathbb{E}(X_n)} - \frac{\mathbb{E}(X_{n+2})}{\mathbb{E}(X_{n+1})}\right) \tag{6.10}$$

In any case, the above properties imply:

$$\mathbb{E}(X_{n+2}) \leq \max\left(\mathbb{E}(X_n), \mathbb{E}(X_{n+1})\right) \tag{6.11}$$

These properties do not directly show that the error term $2\max(0, m_{n-1} - \mathbb{E}(X_{n-1}))$ of Lemma 6.5 is small. However, they allow to show that there are infinitely large values for which it is small.

Lemma 6.7. *Let $c \in]0; 1/2[$ be fixed. There exists an extraction $\phi : \mathbb{N} \to \mathbb{N}$ such that:*

$$\forall n \geq 1 \qquad \max(0, m_{\phi(n)-1} - \mathbb{E}(X_{\phi(n)-1})) < c\mathbb{E}(X_{\phi(n)}) \quad \textit{and} \quad m_{\phi(n)+1} = \mathbb{E}(X_{\phi(n)})$$

The proof of this lemma is in Paragraph 6.3.4. We immediately rewrite it under a more practical form:

Corollary 6.8. *Let $\alpha > 2$ be fixed. There exists an extraction $\phi : \mathbb{N} \to \mathbb{N}$ such that:*

$$\forall n \geq 1 \qquad \mathbb{E}(X_{\phi(n)}) \leq \alpha f(S_{\phi(n)}) \quad \textit{and} \quad m_{\phi(n)+1} = \mathbb{E}(X_{\phi(n)})$$

Proof of Corollary 6.8. If we use the extraction of Lemma 6.7 for a value of c such that $\frac{2}{1-c} \leq \alpha$, then, for any $n \geq 1$, from Lemma 6.5:

$$\begin{aligned} \mathbb{E}(X_{\phi(n)}) &\leq 2f(S_{\phi(n)}) + \max(0, m_{\phi(n)-1} - \mathbb{E}(X_{\phi(n)-1})) \\ &\leq 2f(S_{\phi(n)}) + c\mathbb{E}(X_{\phi(n)}) \\ \Rightarrow \qquad \mathbb{E}(X_{\phi(n)}) &\leq \frac{2}{1-c} f(S_{\phi(n)}) \leq \alpha f(S_{\phi(n)}) \end{aligned}$$

□

Using again the properties of Lemma 6.6, we can show that, for n large enough, if n satisfies the equation of Corollary 6.8, then so does $n + 2$. It is proven in Paragraph 6.3.5.

Lemma 6.9. *Let $\alpha \in]2; 5/2[$ be fixed. For any n large enough, if:*

$$\mathbb{E}(X_n) \leq \alpha f(S_n) \quad \textit{and} \quad m_{n+1} = \mathbb{E}(X_n)$$

then:

$$\mathbb{E}(X_{n+2}) \leq \alpha f(S_{n+2}) \quad \textit{and} \quad m_{n+3} = \mathbb{E}(X_{n+2})$$

Combining this lemma and Corollary 6.8, we obtain that, for all $\alpha > 2$:

$$\begin{array}{lllll} & \mathbb{E}(X_{2n}) \leq \alpha f(S_{2n}) & \text{and} \quad m_{2n+1} = \mathbb{E}(X_{2n}) & \text{for all } n \text{ large enough} \\ \text{or} & \mathbb{E}(X_{2n+1}) \leq \alpha f(S_{2n+1}) & \text{and} \quad m_{2n+2} = \mathbb{E}(X_{2n+1}) & \text{for all } n \text{ large enough} \end{array}$$

As α can take any value larger than 2 and, by Lemma 6.5, $\mathbb{E}(X_n) \geq 2f(S_n)$ for all n:

$$\begin{array}{lllll} & \mathbb{E}(X_{2n}) \sim 2f(S_{2n}) & \text{and} \quad m_{2n+1} = \mathbb{E}(X_{2n}) & \text{when } n \text{ goes to infinity} \\ \text{or} & \mathbb{E}(X_{2n+1}) \sim 2f(S_{2n+1}) & \text{and} \quad m_{2n+2} = \mathbb{E}(X_{2n+1}) & \text{when } n \text{ goes to infinity} \end{array}$$

Let us for example assume that we are in the first case: $\mathbb{E}(X_{2n}) \sim 2f(S_{2n})$. A final lemma, proven in Paragraph 6.3.6, concludes.

Lemma 6.10.

$$\begin{array}{lllll} \textit{If} & \mathbb{E}(X_{2n}) \sim 2f(S_{2n}) & \textit{and} & m_{2n+1} = \mathbb{E}(X_{2n}) & \textit{when } n \textit{ goes to infinity,} \\ \textit{then} & \mathbb{E}(X_{2n+1}) \sim 2f(S_{2n+1}) & & & \textit{when } n \textit{ goes to infinity} \end{array}$$

6.3.2 Proof of Lemma 6.5

Proof of Lemma 6.5. For any $n \geq 1$, when $x \geq 0$:

$$\begin{aligned} \phi_{n-1}(x) + \phi_n(x) &= \phi_{n-1}(x) + |\phi_{n-1}(x) - \mathbb{E}(X_{n-1})| \\ &= \mathbb{E}(X_{n-1}) + 2(\phi_{n-1}(x) - \mathbb{E}(X_{n-1}))1_{\phi_{n-1}(x) > \mathbb{E}(X_{n-1})} \\ &= \mathbb{E}(X_{n-1}) + 2(x - S_n)1_{x \geq S_n} + 2(\phi_{n-1}(x) - \mathbb{E}(X_{n-1}))1_{x < S_n \text{ and } \phi_{n-1}(x) > \mathbb{E}(X_{n-1})} \end{aligned}$$

When $x \in [S_{n-1}; S_n]$:

$$\phi_{n-1}(x) = x - S_{n-1} \leq S_n - S_{n-1} = \mathbb{E}(X_{n-1})$$

So $x < S_n$ and $\phi_{n-1}(x) > \mathbb{E}(X_{n-1})$ actually imply $x < S_{n-1}$. Consequently, because, by definition of m_{n-1}, $\phi_{n-1}(x) \leq m_{n-1}$ on $[0; S_{n-1}]$:

$$0 \leq 2(\phi_{n-1}(x) - \mathbb{E}(X_{n-1}))1_{x < S_n \text{ and } \phi_{n-1}(x) > \mathbb{E}(X_{n-1})} \leq 2\max(0, m_{n-1} - \mathbb{E}(X_{n-1}))$$

From the first equality of the proof, this means:

$$\begin{aligned} \mathbb{E}(X_{n-1}) + 2(x - S_n)1_{x \geq S_n} &\leq \phi_{n-1}(x) + \phi_n(x) \\ &\leq \mathbb{E}(X_{n-1}) + 2(x - S_n)1_{x \geq S_n} + 2\max(0, m_{n-1} - \mathbb{E}(X_{n-1})) \end{aligned}$$

Replacing x by X_0 and taking the expectation gives:

$$2f(S_n) \leq \mathbb{E}(X_n) \leq 2f(S_n) + 2\max(0, m_{n-1} - \mathbb{E}(X_{n-1})) \tag{6.5}$$

□

6.3.3 Proof of Lemma 6.6

Proof of Lemma 6.6.
Equation (6.6): for any $x \in [0; S_n]$:

$$\begin{aligned} \phi_{n+1}(x) &= |\phi_n(x) - \mathbb{E}(X_n)| \\ &= \max(\phi_n(x) - \mathbb{E}(X_n), \mathbb{E}(X_n) - \phi_n(x)) \\ &\leq \max(m_n - \mathbb{E}(X_n), \mathbb{E}(X_n)) \end{aligned}$$

For any $x \in [S_n; S_{n+1}]$:

$$\phi_{n+1}(x) = |x - S_n - \mathbb{E}(X_n)| = |X - S_{n+1}| = S_{n+1} - x \leq S_{n+1} - S_n = \mathbb{E}(X_n)$$

So:

$$m_{n+1} = \max_{x \in [0; S_{n+1}]} \phi_{n+1}(x) \leq \max(m_n - \mathbb{E}(X_n), \mathbb{E}(X_n))$$

Moreover, if $x \in [0; S_n]$ is such that $\phi_n(x) = m_n$, then $\phi_{n+1}(x) = |m_n - \mathbb{E}(X_n)| \geq m_n - \mathbb{E}(X_n)$, so $m_{n+1} \geq m_n - \mathbb{E}(X_n)$.

As $\phi_{n+1}(S_n) = S_{n+1} - S_n = \mathbb{E}(X_n)$, we also have $m_{n+1} \geq \mathbb{E}(X_n)$. So:

$$m_{n+1} = \max(m_n - \mathbb{E}(X_n), \mathbb{E}(X_n)) \tag{6.6}$$

Equation (6.7): for any n, X_n only takes positive values.

$$\mathbb{E}(X_{n+1}) = \mathbb{E}(|X_n - \mathbb{E}(X_n)|) \leq \mathbb{E}(|X_n| + |\mathbb{E}(X_n)|) = \mathbb{E}(X_n) + \mathbb{E}(X_n) = 2\mathbb{E}(X_n) \tag{6.7}$$

Equation (6.8): we remark that, if $\mathbb{E}(X_n) \geq \mathbb{E}(X_{n+1})$, then:

$$\forall x \in \mathbb{R}^+ \qquad ||x - \mathbb{E}(X_n)| - \mathbb{E}(X_{n+1})| \leq \frac{\mathbb{E}(X_{n+1})}{\mathbb{E}(X_n)} x + \left(1 - \frac{\mathbb{E}(X_{n+1})}{\mathbb{E}(X_n)}\right) |x - \mathbb{E}(X_n)| \tag{6.12}$$

Indeed, the functions on both sides of the inequality are affine on the following intervals:

$$[0; \mathbb{E}(X_n) - \mathbb{E}(X_{n+1})] \qquad [\mathbb{E}(X_n) - \mathbb{E}(X_{n+1}); \mathbb{E}(X_n)]$$
$$[\mathbb{E}(X_n); \mathbb{E}(X_n) + \mathbb{E}(X_{n+1})] \qquad [\mathbb{E}(X_n) + \mathbb{E}(X_{n+1}); +\infty[$$

On the last interval, both functions have a slope equal to 1. So to prove the inequality, it suffices to verify that it holds at the junction points:

$$0 \qquad \mathbb{E}(X_n) - \mathbb{E}(X_{n+1}) \qquad \mathbb{E}(X_n) \qquad \mathbb{E}(X_n) + \mathbb{E}(X_{n+1})$$

which can be done by a simple computation.

Replacing x by X_n in (6.12), then taking the expectation yields:

$$\begin{aligned}
\mathbb{E}(X_{n+2}) &= \mathbb{E}(||X_n - \mathbb{E}(X_n)| - \mathbb{E}(X_{n+1})|) \\
&\leq \frac{\mathbb{E}(X_{n+1})}{\mathbb{E}(X_n)} \mathbb{E}(X_n) + \left(1 - \frac{\mathbb{E}(X_{n+1})}{\mathbb{E}(X_n)}\right) \mathbb{E}(|X_n - \mathbb{E}(X_n)|) \\
&= \mathbb{E}(X_{n+1}) \left(2 - \frac{\mathbb{E}(X_{n+1})}{\mathbb{E}(X_n)}\right)
\end{aligned} \tag{6.8}$$

Equation (6.9): this proof is similar to the previous one.

If $\mathbb{E}(X_n) \leq \mathbb{E}(X_{n+1})$, then:

$$\forall x \in \mathbb{R}^+ \qquad ||x - \mathbb{E}(X_n)| - \mathbb{E}(X_{n+1})| \geq \alpha + \beta x + \gamma |x - \mathbb{E}(X_n)|$$

where we have set:

$$\alpha = -\mathbb{E}(X_n) \qquad \beta = 1 - \frac{\mathbb{E}(X_{n+1})}{\mathbb{E}(X_n)} \qquad \gamma = \frac{\mathbb{E}(X_{n+1})}{\mathbb{E}(X_n)}$$

By the same reasoning as in the proof of Equation (6.8), this inequality is true because the two functions involved have a slope equal to 1 at infinity, and because the inequality holds in $0, \mathbb{E}(X_n)$ and $\mathbb{E}(X_n) + \mathbb{E}(X_{n+1})$.

If we replace x by X_n and take the expectation:

$$\mathbb{E}(X_{n+2}) \geq \alpha + \beta \mathbb{E}(X_n) + \gamma \mathbb{E}(X_{n+1}) = \mathbb{E}(X_{n+1}) \left(\frac{\mathbb{E}(X_{n+1})}{\mathbb{E}(X_n)} - 1 \right) \tag{6.13}$$

The proof of the other inequality is easier. By the triangular inequality:

$$\begin{aligned} \forall x \in \mathbb{R}^+ \qquad ||x - \mathbb{E}(X_n)| - \mathbb{E}(X_{n+1})| &= ||x - \mathbb{E}(X_n)| - | - \mathbb{E}(X_{n+1})|| \\ &\leq |x - \mathbb{E}(X_n) + \mathbb{E}(X_{n+1})| \\ &= x + \mathbb{E}(X_{n+1}) - \mathbb{E}(X_n) \end{aligned}$$

If we replace x by X_n and take the expectation:

$$\mathbb{E}(X_{n+2}) \leq \mathbb{E}(X_{n+1})$$

So:

$$\mathbb{E}(X_{n+1}) \geq \mathbb{E}(X_{n+2}) \geq \mathbb{E}(X_{n+1}) \left(\frac{\mathbb{E}(X_{n+1})}{\mathbb{E}(X_n)} - 1 \right) \tag{6.9}$$

Equation (6.10): if $\mathbb{E}(X_n) \leq \mathbb{E}(X_{n+1})$, then, by the previous equation, $\mathbb{E}(X_{n+2}) \leq \mathbb{E}(X_{n+1})$. Moreover:

$$\begin{aligned} \mathbb{E}(X_{n+2}) &\geq \mathbb{E}(X_{n+1}) \left(\frac{\mathbb{E}(X_{n+1})}{\mathbb{E}(X_n)} - 1 \right) \\ &= \mathbb{E}(X_{n+1}) - \mathbb{E}(X_n) + \frac{1}{\mathbb{E}(X_n)} (\mathbb{E}(X_{n+1}) - \mathbb{E}(X_n))^2 \\ &\geq \mathbb{E}(X_{n+1}) - \mathbb{E}(X_n) \end{aligned}$$

It allows us to show the following inequality:

$$\forall x \in \mathbb{R}^+ \quad |||x-\mathbb{E}(X_n)|-\mathbb{E}(X_{n+1})|-\mathbb{E}(X_{n+2})| \leq \alpha+\beta x+\gamma|x-\mathbb{E}(X_n)|+\delta||x-\mathbb{E}(X_n)|-\mathbb{E}(X_{n+1})|$$

where:

$$\alpha = (\mathbb{E}(X_{n+1}) - \mathbb{E}(X_{n+2}))\left(\frac{\mathbb{E}(X_n)}{\mathbb{E}(X_{n+1})} - 1\right) \quad \beta = (\mathbb{E}(X_{n+1}) - \mathbb{E}(X_{n+2}))\left(\frac{1}{\mathbb{E}(X_n)} - \frac{1}{\mathbb{E}(X_{n+1})}\right)$$
$$\gamma = 1 - \left(\frac{\mathbb{E}(X_{n+1}) - \mathbb{E}(X_{n+2})}{\mathbb{E}(X_n)}\right) \qquad \delta = \frac{\mathbb{E}(X_{n+1}) - \mathbb{E}(X_{n+2})}{\mathbb{E}(X_{n+1})}$$

By the same argument as in the proofs of Equations (6.8) and (6.9), it suffices to remark that the functions on both sides of the inequality have a slope equal to 1 at infinity and to check the inequality at points:

$$0 \qquad \mathbb{E}(X_n) - \mathbb{E}(X_{n+1}) + \mathbb{E}(X_{n+2}) \qquad \mathbb{E}(X_n) \qquad \mathbb{E}(X_n) + \mathbb{E}(X_{n+1}) - \mathbb{E}(X_{n+2})$$
$$\mathbb{E}(X_n) + \mathbb{E}(X_{n+1}) \qquad \mathbb{E}(X_n) + \mathbb{E}(X_{n+1}) + \mathbb{E}(X_{n+2})$$

which can be done by a simple computation.

Replacing x by X_n and computing the expectation gives:

$$\begin{aligned}\mathbb{E}(X_{n+3}) &\leq \alpha + \beta\mathbb{E}(X_n) + \gamma\mathbb{E}(X_{n+1}) + \delta\mathbb{E}(X_{n+2}) \\ &= \mathbb{E}(X_{n+1}) - (\mathbb{E}(X_{n+1}) - \mathbb{E}(X_{n+2}))\left(\frac{\mathbb{E}(X_{n+1})}{\mathbb{E}(X_n)} - \frac{\mathbb{E}(X_{n+2})}{\mathbb{E}(X_{n+1})}\right)\end{aligned} \tag{6.10}$$

Equation (6.11): if $\mathbb{E}(X_n) \leq \mathbb{E}(X_{n+1})$, then, by Equation (6.9):

$$\mathbb{E}(X_{n+2}) \leq \mathbb{E}(X_{n+1}) \leq \max(\mathbb{E}(X_n), \mathbb{E}(X_{n+1}))$$

If $\mathbb{E}(X_n) > \mathbb{E}(X_{n+1})$, then, by Equation (6.8):

$$\begin{aligned}\mathbb{E}(X_{n+2}) &\leq \mathbb{E}(X_{n+1})\left(2 - \frac{\mathbb{E}(X_{n+1})}{\mathbb{E}(X_n)}\right) \\ &= \mathbb{E}(X_n)\left(1 - \left(1 - \frac{\mathbb{E}(X_{n+1})}{\mathbb{E}(X_n)}\right)^2\right) \\ &\leq \mathbb{E}(X_n) \\ &\leq \max(\mathbb{E}(X_n), \mathbb{E}(X_{n+1}))\end{aligned}$$

So in any case:

$$\mathbb{E}(X_{n+2}) \leq \max(\mathbb{E}(X_n), \mathbb{E}(X_{n+1})) \tag{6.11}$$

□

6.3.4 Proof of Lemma 6.7

Proof of Lemma 6.7. We need to show that there exist arbitrarily large values of n for which the following equation holds:

$$\max(0, m_{n-1} - \mathbb{E}(X_{n-1})) < c\mathbb{E}(X_n) \quad \text{and} \quad m_{n+1} = \mathbb{E}(X_n)$$

By contradiction, we assume that there exists $N \in \mathbb{N}$ such that the equation is never satisfied when $n \geq N$. It means that, for all $n \geq N$:

$$\max(0, m_{n-1} - \mathbb{E}(X_{n-1})) \geq c\mathbb{E}(X_n) \quad \text{or} \quad m_{n+1} \neq \mathbb{E}(X_n) \tag{6.14}$$

We will show that:

$$\sum_{n \geq N} \mathbb{E}(X_n) < +\infty \tag{6.15}$$

This is absurd because then, $S_n = \sum_{k=0}^{n-1} \mathbb{E}(X_k)$ converges to a finite limit L. By continuity, $f(S_n) \to f(L)$ so, from Lemma 6.5, $\mathbb{E}(X_n) \geq 2f(S_n) \geq 2f(L)$ for all $n \geq 1$. So $\mathbb{E}(X_n)$ is lower bounded by a strictly positive constant and Equation (6.15) does not hold.

We distinguish the values of n, depending on which of the following two properties they satisfy:

$$\begin{aligned} P_1(n): &\quad m_n = \mathbb{E}(X_{n-1}) \\ P_2(n): &\quad m_n = m_{n-1} - \mathbb{E}(X_{n-1}) \end{aligned}$$

The following assertions are true; they are proven at the end of the proof.

(1) If $n \geq N$ and $P_1(n), P_1(n+1)$, then:

$$\mathbb{E}(X_{n+1}) \leq (1 - c^2)\mathbb{E}(X_{n-1}) \qquad \text{(Property (1))}$$

(2) If $n \leq n'$ and $P_1(n), P_2(n+1), ..., P_2(n')$, then:

$$\mathbb{E}(X_{n-1}) \geq \mathbb{E}(X_n) + \mathbb{E}(X_{n+1}) + ... + \mathbb{E}(X_{n'-1}) \qquad \text{(Property (2))}$$

These properties imply that there exists $C > 0$ such that, for any n, n', n'' with $N \leq n < n' < n''$:

$$\text{If } P_1(n), ..., P_1(n'-1), P_2(n'), ..., P_2(n''-1) \text{ then } \sum_{k=n-1}^{n''-2} \mathbb{E}(X_k) \leq C\mathbb{E}(X_{n-1}) \tag{6.16}$$

Indeed, because of Property (1), for any $k \geq 0$:

$$\begin{array}{rcrl} \text{If } n+2k-1 < n' & \text{then} & \mathbb{E}(X_{n+2k-1}) & \leq (1-c^2)^k \mathbb{E}(X_{n-1}) \\ \text{If } n+2k < n' & \text{then} & \mathbb{E}(X_{n+2k}) & \leq (1-c^2)^k \mathbb{E}(X_n) \end{array} \tag{6.17}$$

So:

$$\begin{aligned} \sum_{k=n-1}^{n'-1} \mathbb{E}(X_k) &\leq (\mathbb{E}(X_{n-1}) + \mathbb{E}(X_n)) \times \left(1 + (1-c^2) + (1-c^2)^2 + ...\right) \\ &\leq \frac{1}{c^2} \left(\mathbb{E}(X_{n-1}) + \mathbb{E}(X_n)\right) \\ &\leq \frac{3}{c^2} \mathbb{E}(X_{n-1}) \end{aligned}$$

For the last inequality, we have used Equation (6.7) of Lemma 6.6.

Because of Property (2):

$$\sum_{k=n'}^{n''-2} \mathbb{E}(X_k) \leq \mathbb{E}(X_{n'-2})$$

From relations (6.17), $\mathbb{E}(X_{n'-2}) \leq \mathbb{E}(X_{n-1})$ or $\mathbb{E}(X_{n'-2}) \leq \mathbb{E}(X_n)$ (depending on the parity of $n'-n$). So in any case, from Equation (6.7) of Lemma 6.6:

$$\mathbb{E}(X_{n'-2}) \leq 2\mathbb{E}(X_{n-1}) \qquad \Rightarrow \qquad \sum_{k=n'}^{n''-2} \mathbb{E}(X_k) \leq 2\mathbb{E}(X_{n-1})$$

and:

$$\sum_{k=n-1}^{n''-2} \mathbb{E}(X_k) \leq \left(2 + \frac{3}{c^2}\right) \mathbb{E}(X_{n-1})$$

which proves Equation (6.16).

If $P_1(n)$ is satisfied for all n large enough, then, by Property (1), $(\mathbb{E}(X_n))_{n\in\mathbb{N}}$ decays geometrically and $\sum_{n\geq N} \mathbb{E}(X_n) < +\infty$. On the other hand, if $P_2(n)$ is satisfied for all n large enough, then, by Property (2), we also have $\sum_{n\geq N} \mathbb{E}(X_n) < +\infty$. We can thus assume that there exist arbitrarily large values of n such that $P_1(n)$ holds, and arbitrarily large values such that $P_2(n)$ holds.

Even if we have to replace N by a larger integer, we can assume that $P_1(N)$ holds.

We define the sequence of integers $(n_s)_{s\in\mathbb{N}^*}$ such that $n_1 = N$ and:

$$P_1(n_1) \quad P_1(n_1+1) \quad \ldots \quad P_1(n_2-1) \quad P_2(n_2) \quad \ldots \quad P_2(n_3-1) \quad P_1(n_3) \quad \ldots$$

From Equation (6.16):

$$\sum_{n\geq N-1}\mathbb{E}(X_n)=\sum_{s\geq 0}\sum_{k=n_{2s+1}-1}^{n_{2s+3}-2}\mathbb{E}(X_k)\leq C\sum_{s\geq 0}\mathbb{E}(X_{n_{2s+1}-1})$$

So the proof is over if we show that $\sum_{s\geq 0}\mathbb{E}(X_{n_{2s+1}})<+\infty$. This is a consequence of the following equation:

$$\mathbb{E}(X_{n_{2s+3}-1})\leq\frac{25}{32}\mathbb{E}(X_{n_{2s+1}-1}) \tag{6.18}$$

The proof of this equation is done at the end of the paragraph. □

Proof of the properties (1) and (2).
Property (1): we begin with the case where $P_1(n+2)$ holds.

In this case, by definition of $P_1(n+2)$, $m_{n+2}=\mathbb{E}(X_{n+1})$ so, by Equation (6.14):

$$\max(0,m_n-\mathbb{E}(X_n))\geq c\mathbb{E}(X_{n+1})$$

As $P_1(n)$ holds, $m_n=\mathbb{E}(X_{n-1})$, so:

$$\mathbb{E}(X_{n-1})-\mathbb{E}(X_n)\geq c\mathbb{E}(X_{n+1})$$

As $\mathbb{E}(X_{n+1})$ is positive, this equation in particular implies $\mathbb{E}(X_{n-1})\geq\mathbb{E}(X_n)$ so, by Equation (6.8) Lemma 6.6:

$$\mathbb{E}(X_{n+1})\leq\mathbb{E}(X_n)\left(2-\frac{\mathbb{E}(X_n)}{\mathbb{E}(X_{n-1})}\right)$$

Combining the last two equations:

$$\mathbb{E}(X_{n+1})\leq\min\left(\frac{1}{c}(\mathbb{E}(X_{n-1})-\mathbb{E}(X_n)),\mathbb{E}(X_n)\left(2-\frac{\mathbb{E}(X_n)}{\mathbb{E}(X_{n-1})}\right)\right)$$

Computing the maximum of the right side when $\mathbb{E}(X_n)$ varies in $[0;\mathbb{E}(X_{n-1})]$ yields:

$$\mathbb{E}(X_{n+1})\leq\mathbb{E}(X_{n-1})\left(1+\frac{1}{2c}-\frac{1}{2c}\sqrt{1+4c^2}\right)$$

Using the inequality $c<1/2$ and the fact that, for any positive x, $\sqrt{1+x}\geq 1+x/2-x^2/8$, we obtain:

$$\begin{aligned}\mathbb{E}(X_{n+1})&\leq\mathbb{E}(X_{n-1})(1-c+c^3)\\&\leq\mathbb{E}(X_{n-1})(1-c^2)\end{aligned}$$

Let us now handle the case where $P_1(n+2)$ does not hold. From Equation (6.6) of Lemma 6.6, it means that $\mathbb{E}(X_{n+1}) < m_{n+1} - \mathbb{E}(X_{n+1})$, so $\mathbb{E}(X_{n+1}) < m_{n+1}/2 = \mathbb{E}(X_n)/2$.

If $\mathbb{E}(X_n) \leq \mathbb{E}(X_{n-1})$, then:

$$\mathbb{E}(X_{n+1}) \leq \frac{\mathbb{E}(X_n)}{2} \leq \frac{\mathbb{E}(X_{n-1})}{2} \leq (1-c^2)\mathbb{E}(X_{n-1})$$

If $\mathbb{E}(X_n) > \mathbb{E}(X_{n-1})$, then, from Equation (6.9) of Lemma 6.6:

$$\mathbb{E}(X_{n+1}) \geq \mathbb{E}(X_n)\left(\frac{\mathbb{E}(X_n)}{\mathbb{E}(X_{n-1})} - 1\right)$$

As we have seen that $\mathbb{E}(X_{n+1}) < \mathbb{E}(X_n)/2$:

$$\begin{aligned} \frac{\mathbb{E}(X_n)}{2} &\geq \mathbb{E}(X_n)\left(\frac{\mathbb{E}(X_n)}{\mathbb{E}(X_{n-1})} - 1\right) \\ \iff \qquad \mathbb{E}(X_n) &\leq \frac{3}{2}\mathbb{E}(X_{n-1}) \end{aligned}$$

So, as $c < 1/2$:

$$\mathbb{E}(X_{n+1}) < \frac{\mathbb{E}(X_n)}{2} \leq \frac{3}{4}\mathbb{E}(X_{n-1}) \leq (1-c^2)\mathbb{E}(X_{n-1})$$

Property (2): because $P_1(n), P_2(n+1), ..., P_2(n')$,

$$\begin{aligned} m_n &= \mathbb{E}(X_{n-1}) \\ m_{n+1} &= m_n - \mathbb{E}(X_n) = \mathbb{E}(X_{n-1}) - \mathbb{E}(X_n) \\ &... \\ m_{n'} &= \mathbb{E}(X_{n-1}) - \mathbb{E}(X_n) - ... - \mathbb{E}(X_{n'-1}) \end{aligned}$$

The result follows from the fact that $m_{n'} = \max(\mathbb{E}(X_{n'-1}), m_{n'-1} - \mathbb{E}(X_{n'-1}))$ is always positive. □

Proof of Equation (6.18). We divide the proof in three cases:

1. $n_{2s+2} - n_{2s+1}$ odd or $\mathbb{E}(X_{n_{2s+2}-2}) \leq \mathbb{E}(X_{n_{2s+2}-3})$
2. $n_{2s+3} = n_{2s+2} + 1$, $n_{2s+2} - n_{2s+1}$ even and $\mathbb{E}(X_{n_{2s+2}-2}) > \mathbb{E}(X_{n_{2s+2}-3})$
3. $n_{2s+3} \geq n_{2s+2} + 2$, $n_{2s+2} - n_{2s+1}$ even and $\mathbb{E}(X_{n_{2s+2}-2}) > \mathbb{E}(X_{n_{2s+2}-3})$

Case 1: $n_{2s+2} - n_{2s+1}$ odd or $\mathbb{E}(X_{n_{2s+2}-2}) \leq \mathbb{E}(X_{n_{2s+2}-3})$
We begin with showing the inequality $\mathbb{E}(X_{n_{2s+3}-1}) \leq \frac{3}{4}\mathbb{E}(X_{n_{2s+2}-2})$.

If $n_{2s+3} \geq n_{2s+2} + 2$, we actually have:

$$\mathbb{E}(X_{n_{2s+3}-1}) \leq \frac{2}{3}\mathbb{E}(X_{n_{2s+2}-2})$$

Indeed, we use Property (2) for $n = n_{2s+2} - 1$ and $n' = n_{2s+3} - 1$:

$$\mathbb{E}(X_{n_{2s+2}-2}) \geq \mathbb{E}(X_{n_{2s+2}-1}) + ... + \mathbb{E}(X_{n_{2s+3}-2})$$

In particular:

$$\mathbb{E}(X_{n_{2s+2}-2}) \geq \mathbb{E}(X_{n_{2s+3}-3}) + \mathbb{E}(X_{n_{2s+3}-2}) \tag{6.19}$$

If $\mathbb{E}(X_{n_{2s+3}-3}) \leq \mathbb{E}(X_{n_{2s+3}-2})$ then, from Equation (6.9), $\mathbb{E}(X_{n_{2s+3}-1}) \leq \mathbb{E}(X_{n_{2s+3}-2})$. As, moreover, $2\mathbb{E}(X_{n_{2s+3}-3}) \geq \mathbb{E}(X_{n_{2s+3}-2})$ (Equation (6.7)):

$$\begin{aligned} \mathbb{E}(X_{n_{2s+2}-2}) &\geq \frac{3}{2}\mathbb{E}(X_{n_{2s+3}-2}) \geq \frac{3}{2}\mathbb{E}(X_{n_{2s+3}-1}) \\ \Rightarrow \qquad \mathbb{E}(X_{n_{2s+3}-1}) &\leq \frac{2}{3}\mathbb{E}(X_{n_{2s+2}-2}) \end{aligned}$$

If, on the other hand, $\mathbb{E}(X_{n_{2s+3}-3}) > \mathbb{E}(X_{n_{2s+3}-2})$, then, by Equation (6.11):

$$\mathbb{E}(X_{n_{2s+3}-1}) \leq \mathbb{E}(X_{n_{2s+3}-3})$$

Then Equation (6.19), combined with the fact that $\mathbb{E}(X_{n_{2s+3}-1}) \leq 2\mathbb{E}(X_{n_{2s+3}-2})$ yields:

$$\begin{aligned} \mathbb{E}(X_{n_{2s+2}-2}) &\geq \mathbb{E}(X_{n_{2s+3}-1})\left(1 + \frac{1}{2}\right) \\ \Rightarrow \qquad \mathbb{E}(X_{n_{2s+3}-1}) &\leq \frac{2}{3}\mathbb{E}(X_{n_{2s+2}-2}) \end{aligned}$$

If $n_{2s+3} = n_{2s+2} + 1$, then $\mathbb{E}(X_{n_{2s+2}-1}) \leq m_{n_{2s+2}-1}/2 = \mathbb{E}(X_{n_{2s+2}-2})/2$, because $P_2(n_{2s+2})$ holds and $m_{n_{2s+2}} = \max(\mathbb{E}(X_{n_{2s+2}-1}), m_{n_{2s+2}-1} - \mathbb{E}(X_{n_{2s+2}-1}))$. From Equation (6.8):

$$\mathbb{E}(X_{n_{2s+3}-1}) = \mathbb{E}(X_{n_{2s+2}}) \leq \mathbb{E}(X_{n_{2s+2}-1})\left(2 - \frac{\mathbb{E}(X_{n_{2s+2}-1})}{\mathbb{E}(X_{n_{2s+2}-2})}\right)$$

This upper bound is increasing in $\mathbb{E}(X_{n_{2s+2}-1})$ over $[0; \mathbb{E}(X_{n_{2s+2}-2})/2]$, so:

$$\mathbb{E}(X_{n_{2s+3}-1}) \leq \frac{1}{2}\left(2 - \frac{1}{2}\right)\mathbb{E}(X_{n_{2s+2}-2}) = \frac{3}{4}\mathbb{E}(X_{n_{2s+2}-2})$$

In either case, we have, as announced:

$$\mathbb{E}(X_{n_{2s+3}-1}) \leq \frac{3}{4}\mathbb{E}(X_{n_{2s+2}-2})$$

If $n_{2s+2} - n_{2s+1}$ is odd, Property (1) implies:

$$\mathbb{E}(X_{n_{2s+2}-2}) \leq (1-c^2)^{(n_{2s+2}-n_{2s+1}-1)/2}\mathbb{E}(X_{n_{2s+1}-1}) \leq \mathbb{E}(X_{n_{2s+1}-1})$$

Combining the previous two equations:

$$\mathbb{E}(X_{n_{2s+3}-1}) \leq \frac{3}{4}\mathbb{E}(X_{n_{2s+1}-1}) \leq \frac{25}{32}\mathbb{E}(X_{n_{2s+1}-1})$$

If $n_{2s+2} - n_{2s+1}$ is even but $\mathbb{E}(X_{n_{2s+2}-2}) \leq \mathbb{E}(X_{n_{2s+2}-3})$, then, again by Property (1):

$$\mathbb{E}(X_{n_{2s+2}-2}) \leq \mathbb{E}(X_{n_{2s+2}-3}) \leq (1-c^2)^{(n_{2s+2}-n_{2s+1}-2)/2}\mathbb{E}(X_{n_{2s+1}-1}) \leq \mathbb{E}(X_{n_{2s+1}-1})$$

so we reach again the same conclusion:

$$\mathbb{E}(X_{n_{2s+3}-1}) \leq \frac{3}{4}\mathbb{E}(X_{n_{2s+1}-1}) \leq \frac{25}{32}\mathbb{E}(X_{n_{2s+1}-1})$$

Case 2: $n_{2s+3} = n_{2s+2} + 1$, $n_{2s+2} - n_{2s+1}$ even and $\mathbb{E}(X_{n_{2s+2}-2}) > \mathbb{E}(X_{n_{2s+2}-3})$

Because $P_1(n_{2s+2}-1)$ and $P_2(n_{2s+2})$ hold, and because $m_{n_{2s+2}} = \max(\mathbb{E}(X_{n_{2s+2}-1}), m_{n_{2s+2}-1} - \mathbb{E}(X_{n_{2s+2}-1}))$:

$$\mathbb{E}(X_{n_{2s+2}-1}) \leq \frac{1}{2}m_{n_{2s+2}-1} = \frac{1}{2}\mathbb{E}(X_{n_{2s+2}-2})$$

As $\mathbb{E}(X_{n_{2s+2}-2}) > \mathbb{E}(X_{n_{s+2}-3})$, Equation (6.10) of Lemma 6.6 implies that:

$$\begin{aligned}\mathbb{E}(X_{n_{2s+3}-1}) &= \mathbb{E}(X_{n_{2s+2}}) \\ &\leq \mathbb{E}(X_{n_{2s+2}-2}) - (\mathbb{E}(X_{n_{2s+2}-2}) - \mathbb{E}(X_{n_{2s+2}-1}))\left(\frac{\mathbb{E}(X_{n_{2s+2}-2})}{\mathbb{E}(X_{n_{s+2}-3})} - \frac{\mathbb{E}(X_{n_{2s+2}-1})}{\mathbb{E}(X_{n_{2s+2}-2})}\right)\end{aligned}$$

This upper bound is increasing in $\mathbb{E}(X_{n_{2s+2}-1})$ on $[0; \mathbb{E}(X_{n_{2s+2}-2})/2]$: it can be seen by computing the derivative and using the fact that:

$$\mathbb{E}(X_{n_{2s+2}-2}) - \mathbb{E}(X_{n_{2s+2}-1}) \geq \frac{1}{2}\mathbb{E}(X_{n_{2s+2}-2}) > 0$$

$$\text{and} \quad \frac{\mathbb{E}(X_{n_{2s+2}-2})}{\mathbb{E}(X_{n_{s+2}-3})} - \frac{\mathbb{E}(X_{n_{2s+2}-1})}{\mathbb{E}(X_{n_{2s+2}-2})} > 1 - \frac{1}{2} = \frac{1}{2} > 0$$

So, in the upper bound, we can replace $\mathbb{E}(X_{n_{2s+2}-1})$ by $\mathbb{E}(X_{n_{2s+2}-2})/2$ and get:

$$\begin{aligned}\mathbb{E}(X_{n_{2s+3}-1}) &\leq \mathbb{E}(X_{n_{2s+2}-2}) - \frac{1}{2}\mathbb{E}(X_{n_{2s+2}-2})\left(\frac{\mathbb{E}(X_{n_{2s+2}-2})}{\mathbb{E}(X_{n_{s+2}-3})} - \frac{1}{2}\right) \\ &= \frac{\mathbb{E}(X_{n_{2s+2}-2})}{2}\left(\frac{5}{2} - \frac{\mathbb{E}(X_{n_{2s+2}-2})}{\mathbb{E}(X_{n_{s+2}-3})}\right)\end{aligned}$$

When $\mathbb{E}(X_{n_{2s+2}-2})$ varies, the upper bound attains its maximum in $\mathbb{E}(X_{n_{2s+2}-2}) = \frac{5}{4}\mathbb{E}(X_{n_{2s+2}-3})$, which yields:

$$\mathbb{E}(X_{n_{2s+3}-1}) \leq \frac{25}{32}\mathbb{E}(X_{n_{2s+2}-3})$$

By Property (1), as $n_{2s+2} - n_{2s+1}$ is even:

$$\mathbb{E}(X_{n_{2s+3}-1}) \leq \frac{25}{32}\mathbb{E}(X_{n_{2s+2}-3}) \leq \frac{25}{32}(1-c^2)^{(n_{2s+2}-n_{2s+1}-2)/2}\mathbb{E}(X_{n_{2s+1}-1}) \leq \frac{25}{32}\mathbb{E}(X_{n_{2s+1}-1})$$

Case 3: $n_{2s+3} \geq n_{2s+2}+2$, $n_{2s+2} - n_{2s+1}$ even and $\mathbb{E}(X_{n_{2s+2}-2}) > \mathbb{E}(X_{n_{2s+2}-3})$
As $P_1(n_{2s+2}-1)$ holds, $m_{n_{2s+2}-1} = \mathbb{E}(X_{n_{2s+2}-2})$. For any $n = n_{2s+2}, ..., n_{2s+3}-1$, $P_2(n)$ holds so, as $m_n = \max(\mathbb{E}(X_{n-1}), m_n - \mathbb{E}(X_{n-1}))$:

$$m_n = m_{n-1} - \mathbb{E}(X_{n-1}) \qquad \text{and} \qquad \mathbb{E}(X_{n-1}) \leq \frac{1}{2}m_{n-1}$$

By iteration:

$$\begin{aligned}m_n &= \mathbb{E}(X_{n_{2s+2}-2}) - \mathbb{E}(X_{n_{2s+2}-1}) - ... - \mathbb{E}(X_{n-1}) \\ \mathbb{E}(X_{n-1}) &\leq \frac{1}{2}\big(\mathbb{E}(X_{n_{2s+2}-2}) - \mathbb{E}(X_{n_{2s+2}-1}) - ... - \mathbb{E}(X_{n-2})\big)\end{aligned}$$

In particular:

$$\begin{aligned}\mathbb{E}(X_{n_{2s+3}-3}) &\leq \frac{1}{2}\mathbb{E}(X_{n_{2s+2}-2}) \\ \mathbb{E}(X_{n_{2s+3}-2}) &\leq \frac{1}{2}\mathbb{E}(X_{n_{2s+2}-2})\end{aligned}$$

From Equation (6.11) of Lemma 6.6:

$$\mathbb{E}(X_{n_{2s+3}-1}) \leq \max\big(\mathbb{E}(X_{n_{2s+3}-3}), \mathbb{E}(X_{n_{2s+3}-2})\big) \leq \frac{1}{2}\mathbb{E}(X_{n_{2s+2}-2}) \tag{6.20}$$

Now we show that:

$$\mathbb{E}(X_{n_{2s+2}-2}) \leq \frac{3}{2}\mathbb{E}(X_{n_{2s+2}-3})$$

Indeed, by the same reasoning as above, as $P_1(n_{2s+2}-1)$ and $P_2(n_{2s+2})$ hold:

$$\mathbb{E}(X_{n_{2s+2}-1}) \leq \frac{1}{2} m_{n_{2s+2}-1} = \frac{1}{2}\mathbb{E}(X_{n_{2s+2}-2})$$

But, as $\mathbb{E}(X_{n_{2s+2}-2}) > \mathbb{E}(X_{n_{2s+2}-3})$, from this equation and Equation (6.9) of Lemma 6.6:

$$\frac{1}{2}\mathbb{E}(X_{n_{2s+2}-2}) \geq \mathbb{E}(X_{n_{2s+2}-1}) \geq \mathbb{E}(X_{n_{2s+2}-2}) \left(\frac{\mathbb{E}(X_{n_{2s+2}-2})}{\mathbb{E}(X_{n_{2s+2}-3})} - 1 \right)$$

$$\Rightarrow \qquad \mathbb{E}(X_{n_{2s+2}-2}) \leq \frac{3}{2}\mathbb{E}(X_{n_{2s+2}-3})$$

Combining this with (6.20) yields:

$$\mathbb{E}(X_{n_{2s+3}-1}) \leq \frac{3}{4}\mathbb{E}(X_{n_{2s+2}-3})$$

From Property (1), as $n_{2s+2} - n_{2s+1}$ is even:

$$\begin{aligned}
\mathbb{E}(X_{n_{2s+3}-1}) &\leq \frac{3}{4}\mathbb{E}(X_{n_{2s+2}-3}) \\
&\leq \frac{3}{4}(1-c^2)^{(n_{2s+2}-n_{2s+1}-2)/2}\mathbb{E}(X_{n_{2s+1}-1}) \\
&\leq \frac{3}{4}\mathbb{E}(X_{n_{2s+1}-1}) \\
&\leq \frac{25}{32}\mathbb{E}(X_{n_{2s+1}-1})
\end{aligned}$$

□

6.3.5 Proof of Lemma 6.9

Proof of Lemma 6.9. We first remark that, if $\mathbb{E}(X_n) \leq \alpha f(S_n)$, then:

$$\frac{f(S_{n+2})}{f(S_n)} \text{ is arbitrarily close to 1 when } n \to +\infty \tag{6.21}$$

Indeed, $S_{n+2} - S_n = \mathbb{E}(X_n) + \mathbb{E}(X_{n+1}) \leq 3\mathbb{E}(X_n)$ by Equation (6.7) of Lemma 6.6. If $\mathbb{E}(X_n) \leq \alpha f(S_n)$, then:

$$S_{n+2} - S_n \leq 3\alpha f(S_n)$$

The derivative of f goes to zero at infinity, so:

$$f(S_{n+2}) - f(S_n) = o(S_{n+2} - S_n) = o(f(S_n))$$

which implies Equation (6.21).

We fix $x \in]0; 1/2 - 1/\alpha[$ and divide the proof in three cases.

First case: $\mathbb{E}(X_{n+1}) \leq \mathbb{E}(X_n)(1-x)$
Then, from Equation (6.8) of Lemma 6.6:

$$\mathbb{E}(X_{n+2}) \leq \mathbb{E}(X_{n+1}) \left(2 - \frac{\mathbb{E}(X_{n+1})}{\mathbb{E}(X_n)}\right)$$

This bound is increasing in $\mathbb{E}(X_{n+1})$ over $]-\infty; \mathbb{E}(X_n)[$ so:

$$\begin{aligned}
\mathbb{E}(X_{n+2}) &\leq (1-x)\mathbb{E}(X_n) \left(2 - \frac{(1-x)\mathbb{E}(X_n)}{\mathbb{E}(X_n)}\right) \\
&= (1-x^2)\mathbb{E}(X_n) \\
&\leq \alpha(1-x^2) f(S_n)
\end{aligned}$$

We get:

$$\frac{\mathbb{E}(X_{n+2})}{f(S_{n+2})} \leq \alpha(1-x^2)\frac{f(S_n)}{f(S_{n+2})}$$

so, by Equation (6.21), if n is large enough:

$$\frac{\mathbb{E}(X_{n+2})}{f(S_{n+2})} \leq \alpha$$

We must also prove that $m_{n+3} = \mathbb{E}(X_{n+2})$. By Lemma 6.5:

$$\begin{aligned}
\mathbb{E}(X_{n+2}) &\geq 2f(S_{n+2}) \\
&= 2f(S_n)\frac{f(S_{n+2})}{f(S_n)} \\
&\geq \mathbb{E}(X_n)\frac{2}{\alpha}\frac{f(S_{n+2})}{f(S_n)}
\end{aligned}$$

For n large enough, from Equation (6.21) and the fact that $\alpha < 4$, we conclude that:

$$\mathbb{E}(X_{n+2}) \geq \frac{1}{2}\mathbb{E}(X_n) \geq \frac{1}{2}m_{n+2}$$

Indeed, by assumption, $m_{n+1} = \mathbb{E}(X_n)$. As $\mathbb{E}(X_{n+1}) \leq \mathbb{E}(X_n)$, it implies that $m_{n+2} = \max(\mathbb{E}(X_{n+1}), m_{n-1} - \mathbb{E}(X_{n+1})) \leq \mathbb{E}(X_n)$.

As $m_{n+3} = \max(\mathbb{E}(X_{n+2}), m_{n+2} - \mathbb{E}(X_{n+2}))$, we deduce from the last equation that:

$$m_{n+3} = \mathbb{E}(X_{n+2})$$

Second case: $\mathbb{E}(X_n)(1-x) < \mathbb{E}(X_{n+1}) \leq \mathbb{E}(X_n)$.
From Lemma 6.5:

$$\begin{aligned}\mathbb{E}(X_{n+2}) &\leq 2f(S_{n+2}) + 2\max(0, m_{n+1} - \mathbb{E}(X_{n+1})) \\ &= 2f(S_{n+2}) + 2(\mathbb{E}(X_n) - \mathbb{E}(X_{n+1})) \\ &\leq 2f(S_{n+2}) + 2x\mathbb{E}(X_n) \\ &\leq 2f(S_n) + 2x\alpha f(S_n)\end{aligned}$$

So:

$$\frac{\mathbb{E}(X_{n+2})}{f(S_{n+2})} \leq 2(1+\alpha x)\frac{f(S_n)}{f(S_{n+2})}$$

We have chosen $x < \frac{1}{2} - \frac{1}{\alpha}$, so that $2(1+\alpha x) < \alpha$. From Equation (6.21), this implies, for n large enough:

$$\frac{\mathbb{E}(X_{n+2})}{f(S_{n+2})} \leq \alpha$$

The proof that $m_{n+3} = \mathbb{E}(X_{n+2})$ is identical to the one done in the first case.

Third case: $\mathbb{E}(X_{n+1}) > \mathbb{E}(X_n)$
From Lemma 6.5:

$$\mathbb{E}(X_{n+2}) = 2f(S_{n+2}) \leq \alpha f(S_{n+2})$$

because $\max(0, m_{n+1} - \mathbb{E}(X_{n+1})) = \max(0, \mathbb{E}(X_n) - \mathbb{E}(X_{n+1})) = 0$.

Let us show that $m_{n+3} = \mathbb{E}(X_{n+2})$.

As $m_{n+1} = \mathbb{E}(X_n)$ and $\mathbb{E}(X_{n+1}) > \mathbb{E}(X_n)$, we have $m_{n+2} = \max(\mathbb{E}(X_{n+1}), m_{n+1} - \mathbb{E}(X_{n+1})) = \mathbb{E}(X_{n+1})$.

From Equation (6.9) of Lemma 6.6:

$$\mathbb{E}(X_{n+2}) \geq \mathbb{E}(X_{n+1})\left(\frac{\mathbb{E}(X_{n+1})}{\mathbb{E}(X_n)} - 1\right)$$

By solving a polynomial equation of degree 2, we see that it implies:

$$\mathbb{E}(X_{n+1}) \leq \frac{\mathbb{E}(X_n)}{2} + \sqrt{\mathbb{E}(X_n)\left(\mathbb{E}(X_{n+2}) + \frac{\mathbb{E}(X_n)}{4}\right)}$$

Using $\mathbb{E}(X_n) \leq \alpha f(S_n)$ and $\mathbb{E}(X_{n+2}) = 2f(S_{n+2}) \leq 2f(S_n)$ yields:

$$m_{n+2} = \mathbb{E}(X_{n+1}) \leq \left(\frac{\alpha}{2} + \sqrt{\alpha \left(2 + \frac{\alpha}{4} \right)} \right) f(S_n)$$

If $\alpha < 5/2$, then $\frac{\alpha}{2} + \sqrt{\alpha \left(2 + \frac{\alpha}{4} \right)} < 4$. Combining this with Equation (6.21) and the fact that $\mathbb{E}(X_{n+2}) = 2f(S_{n+2})$ gives, for all n large enough:

$$\frac{m_{n+2}}{\mathbb{E}(X_{n+2})} < 2$$

As $m_{n+3} = \max(\mathbb{E}(X_{n+2}), m_{n+2} - \mathbb{E}(X_{n+2}))$, it implies that $m_{n+3} = \mathbb{E}(X_{n+2})$. □

6.3.6 Proof of Lemma 6.10

Proof of Lemma 6.10. We assume that $\mathbb{E}(X_{2n}) \sim 2f(S_{2n})$ when n goes to $+\infty$ and that, for all n large enough:

$$m_{2n+1} = \mathbb{E}(X_{2n})$$

We must show that $\mathbb{E}(X_{2n+1}) \sim 2f(S_{2n+1})$.

As in the proof of Lemma 6.9, we see that $\mathbb{E}(X_{2n}) \sim 2f(S_{2n})$ implies:

$$f(S_n) \sim f(S_{n+1}) \quad \text{when } n \to +\infty$$

From Equation (6.7) of Lemma 6.6, $\mathbb{E}(X_{2n+1}) \leq 2\mathbb{E}(X_{2n})$ for any $n \geq 1$. Combining this with $\mathbb{E}(X_{2n}) \sim 2f(S_{2n}) \sim 2f(S_{2n+1})$ yields:

$$\limsup_{n \to +\infty} \frac{\mathbb{E}(X_{2n+1})}{f(S_{2n+1})} \leq 4$$

We define:

$$M = \limsup_{n \to +\infty} \frac{\mathbb{E}(X_{2n+1})}{f(S_{2n+1})} \leq 4$$

and show that $M = 2$. As, from Lemma 6.5, $\mathbb{E}(X_{2n+1}) \geq 2f(S_{2n+1})$, it implies $\mathbb{E}(X_{2n+1}) \sim 2f(S_{2n+1})$.

We first find an upper bound for $\mathbb{E}(X_{2n+1})/f(S_{2n+1})$ as a function of $\mathbb{E}(X_{2n-1})/f(S_{2n-1})$. We distinguish two cases.

First case: $\mathbb{E}(X_{2n-1}) > \mathbb{E}(X_{2n})$

Then, from Equation (6.8) of Lemma 6.6:

$$\mathbb{E}(X_{2n+1}) \leq \mathbb{E}(X_{2n}) \left(2 - \frac{\mathbb{E}(X_{2n})}{\mathbb{E}(X_{2n-1})}\right)$$

The upper bound is increasing in $\mathbb{E}(X_{2n-1})$. By definition of M, for all $M' > M$, $\mathbb{E}(X_{2n-1}) < M' f(S_{2n-1})$ when n is large enough. So, for n large enough:

$$\mathbb{E}(X_{2n+1}) \leq \mathbb{E}(X_{2n}) \left(2 - \frac{\mathbb{E}(X_{2n})}{M' f(S_{2n-1})}\right) \sim 4 f(S_{2n+1}) \left(1 - \frac{1}{M'}\right)$$

We used the fact that $\mathbb{E}(X_{2n}) \sim 2f(S_{2n}) \sim 2f(S_{2n-1}) \sim 2f(S_{2n+1})$.

As this inequality holds for any $M' > M$, we obtain:

$$\limsup_{n \to +\infty,\ \mathbb{E}(X_{2n-1}) > \mathbb{E}(X_{2n})} \frac{\mathbb{E}(X_{2n+1})}{f(S_{2n+1})} \leq 4 \left(1 - \frac{1}{M}\right) \tag{6.22}$$

Second case: $\mathbb{E}(X_{2n-1}) \leq \mathbb{E}(X_{2n})$.
By assumption, for n large enough, $m_{2n-1} = \mathbb{E}(X_{2(n-1)})$. As $\mathbb{E}(X_{2(n-1)}) \sim 2f(S_{2(n-1)}) \sim 2f(S_{2n-1})$, the following inequality holds for all n large enough:

$$\mathbb{E}(X_{2n-1}) \geq 2f(S_{2n-1}) \geq \frac{m_{2n-1}}{2}$$

As $m_{2n} = \max(\mathbb{E}(X_{2n-1}), m_{2n-1} - \mathbb{E}(X_{2n-1}))$, we must have $m_{2n} = \mathbb{E}(X_{2n-1})$. So $m_{2n} - \mathbb{E}(X_{2n}) = \mathbb{E}(X_{2n-1}) - \mathbb{E}(X_{2n}) \leq 0$ and, by Lemma 6.5:

$$\frac{\mathbb{E}(X_{2n+1})}{f(S_{2n+1})} = 2 \tag{6.23}$$

We can now conclude. If we combine Equations (6.22) and (6.23), we obtain:

$$\limsup_{n \to +\infty} \frac{\mathbb{E}(X_{2n+1})}{f(S_{2n+1})} \leq \max\left(2, 4\left(1 - \frac{1}{M}\right)\right)$$

A simple computation shows that, if $M > 2$, then this upper bound is strictly smaller than M, which is a contradiction with the definition of M. So $M = 2$. □

6.4 Numerical illustrations

6.4.1 Characterization of the distribution tail

We perform numerical experiments for two different choices of X_0. The first process X_0 follows a Laplace law, with a probability density function equal to:

$$p(x) = e^{-x} 1_{x \geq 0}$$

The second process follows a variant of a Pareto law:

$$p(x) = \frac{8}{7} \left(\frac{1}{1 + |x - 1|} \right)^3$$

For each of these two choices, we plot the density function p on Figure 6.2 (graphs (a) and (c)). We also plot the first pairs $(\mathbb{E}(X_n), S_n)$, with S_n defined as in the last section:

$$S_n = \sum_{k=0}^{n-1} \mathbb{E}(X_k)$$

These pairs correspond to the red crosses of graphs (b) and (d). They are displayed along with the function $2f$, where is f defined as in Equation (6.4):

$$\begin{array}{rcl} f : \mathbb{R}^+ \to & & \mathbb{R}^+ \\ y \to & & \mathbb{E}((X_0 - y) 1_{X_0 \geq y}) \end{array} \tag{6.4}$$

As stated by Theorem 6.4:

$$\mathbb{E}(X_n) \sim 2f(S_n)$$

In our two examples, the convergence of $\mathbb{E}(X_n)/(2f(S_n))$ towards 1 seems fast. From $n = 5$, it is difficult to distinguish $\mathbb{E}(X_n)$ from $2f(S_n)$. In both cases, the sequence $(\mathbb{E}(X_n))_{n \in \mathbb{N}}$ provides a very precise estimation of f in the high values.

We remark that, although $(\mathbb{E}(X_n))_{n \in \mathbb{N}}$ precisely characterizes the distribution tail of X_0, it conveys a much less precise information about the distribution at small values.

In particular, the sequence $(\mathbb{E}(X_n))_{n \in \mathbb{N}}$ does not uniquely determine the law of X_0. An example of two different processes associated to the same sequence $(\mathbb{E}(X_n))_{n \in \mathbb{N}}$ is shown on Figure 6.3.

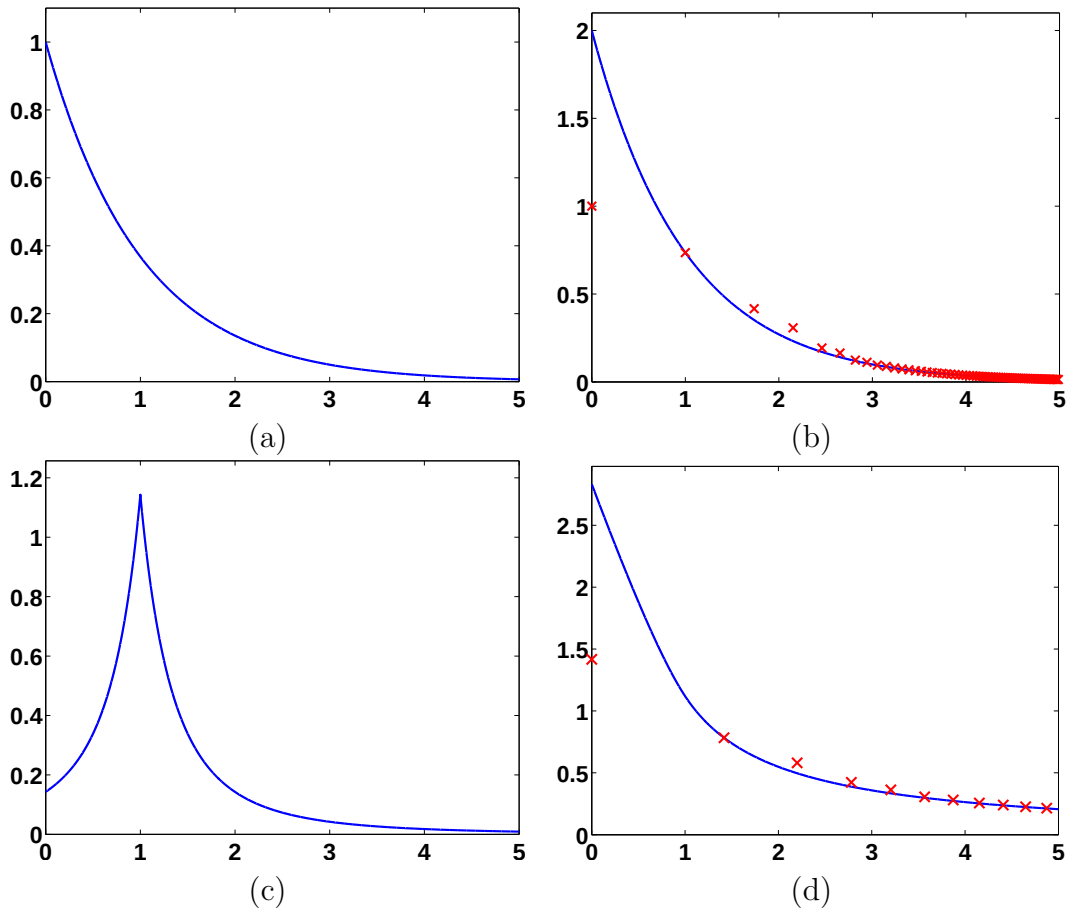


Figure 6.2:
(a) Laplace probability density function $p(x) = e^{-x}1_{x\geq 0}$
(b) Function $2f$, with f defined as in Equation (6.4), and the pairs $(S_n, \mathbb{E}(X_n))$ for $n = 0, ..., 69$
(c) Pareto-like probability density function $p(x) = (8/7)(1 + |x - 1|)^{-3}1_{x\geq 0}$
(d) Function $2f$, and the pairs $(S_n, \mathbb{E}(X_n))$ for $n = 0, ..., 10$

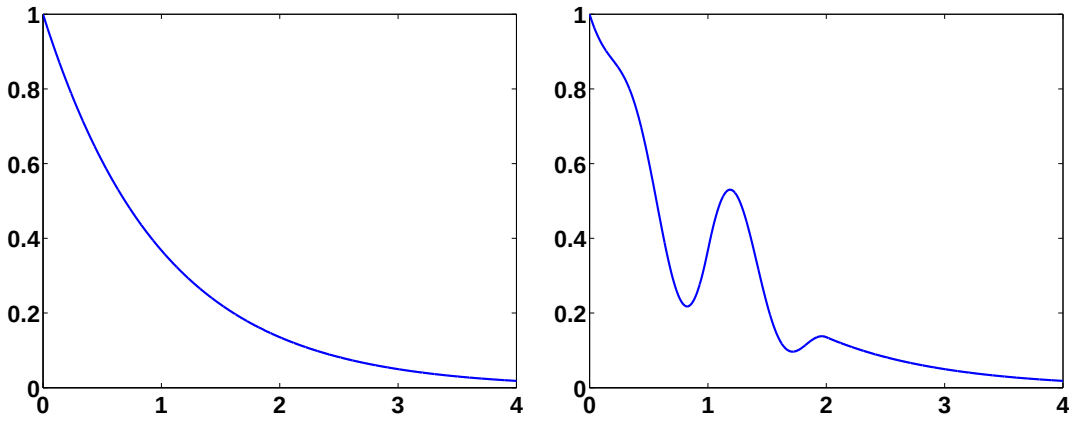


Figure 6.3: Two probability density functions associated to the same sequence $(\mathbb{E}(X_n))_{n\in\mathbb{N}}$

6.4.2 Probability distribution of X_n

When the initial process X_0 admits a probability density function, so do the X_n; we denote by p_n their density function. In Figure 6.4 is displayed p_{10}, the probability density function of X_{10}, for the Pareto-like distribution.

It is discontinuous, although the initial probability density function is continuous over $\mathbb{R}^+$. Indeed, the iterative definition $X_{n+1} = |X_n - \mathbb{E}(X_n)|$ does not preserve the continuity of the density function.

In general, we observe that the density function p_n presents two peaks, one around zero and the other one around $\mathbb{E}(X_n)$. The mass of each peak converges to $1/2$ when n goes to infinity. The tail of p_n, on the other hand, is equal to:

$$p_n(x) = p_0(x + S_n) \qquad \text{when } x \gtrsim \mathbb{E}(X_n)$$

For the Pareto-like distribution, the two parts of p_{10} are respectively shown on graphs (a) and (b) of Figure 6.4.

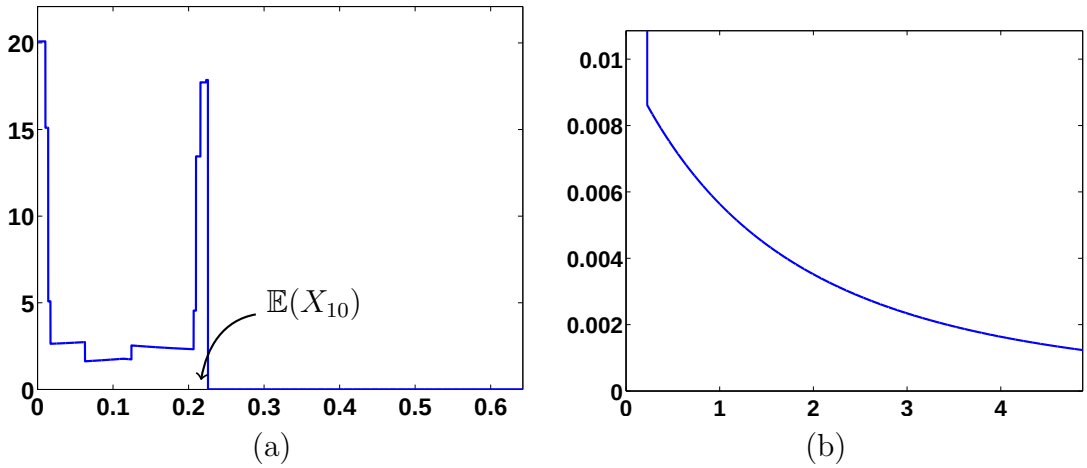


Figure 6.4: For the Pareto-like distribution, probability density function of X_{10}
(a) around zero (b) for large values (Note the change of scale between the two graphs.)

Conclusion: relation with learned deep representations and future work

In this thesis, we have studied the modulus of the wavelet transform. We have seen that this operator produces a representation of audio signals which is both discriminative (two signals that are different for a human person have different representations) and stable to transformations which a person can not hear (multiplication by a slow-varying phase in the time-frequency domain). In chapters 5 and 6, we have moreover seen how it could be used to construct deep representations.

Thanks to these properties, the modulus of the wavelet transform is a crucial tool for the analysis of audio signals. In audio processing, almost all representations used to analyze or classify sounds rely on the modulus of a time-frequency transform (spectrogram or scalogram, often under the form of mel-frequency cepstral coefficients). In recent works, the modulus of the wavelet transform is sometimes followed by sophisticated additional transformations [Hinton et al., 2012]. However, it is always the starting point of the representation.

The goal of this conclusion is to explain why the situation is somewhat similar in image processing, although it may be less obvious, and to describe future directions of research.

In image processing, time-frequency representations have also been the cause of important successes, with in particular descriptors like SIFT [Lowe, 1999] or HOG [Dalal and Triggs, 2005]. However, a recent and very successful trend in image processing consists in learning deep representations from end to end, using raw pixels as inputs [Krizhevsky et al., 2012; Simonyan and Zisserman, 2015; Szegedy et al., 2015]. In this case, no modulus of wavelet transform is explicitly computed.

Nevertheless, it is known that part of filters in the first layer of a neural network spontaneously tend to resemble Gabor filters, responding to particular orientations and scales [Zeiler and Fergus, 2014]. So they are the same filters as in a wavelet transform, and they are also followed by a non-linearity, even if it is not exactly a modulus. Indeed, we are going to see that, at least in AlexNet [Krizhevsky et al., 2012], the first half of the first convolutional layer is essentially the modulus of a wavelet transform.

This observation is coherent with the results of [Yosinski et al., 2014], suggesting that the first layers of a neural network are relatively universal: they do not depend much on the dataset on which the representation has been learned. It is also in line with [Perronnin and Larlus, 2015], who show that replacing the first layers of a neural network by more traditional descriptors (Fisher vectors) simplifies the learning, while retaining a high classification accuracy.

In the first paragraph of this conclusion, we compare the filters of the first layer of AlexNet with the ones of a wavelet transform. In the second paragraph, we discuss the non-linearity. We conclude with preliminary remarks on the second layer.

The images are generated with *Caffe* [Jia et al., 2014]. We use the version of AlexNet implemented in *Caffe*, which slightly differs from the original one.

First convolutional layer

AlexNet contains eight layers. The first five are convolutional, and the last three fully-connected. At the first layer, there are 96 filters, of size $3 \times 11 \times 11$. The first dimension corresponds to the three color channels of the input image, while the second and third dimensions are spatial. The 96 filters are displayed in color on Figure 7.1.

In the implementation of AlexNet, due to memory constraints, the filters of each convolutional layer are split in two halves, stored on different GPUs. There is no communication between the two GPUs, except in the third convolutional layer and in the fully-connected ones. Even if this is not imposed by the architecture, the first GPU tends to learn black-and-white filters, while the second one learns colored filters.

Here, we focus our analysis on the black-and-white filters, that is, the first 48 ones. They are displayed on Figure 7.2 (a). Except for numbers $11, 25, 43, 46$ and 47, these filters have zero-mean and present oscillating patterns, with distinct orientations. They are similar to filters of a two-dimensional wavelet transform.

As shown in Figure 7.2 (b), their Fourier transforms are very well-localized in frequency. The distribution of characteristic frequencies in the plane is shown on Figure 7.3. It is relatively uniform across the different orientations. However, it is logarithmic in scale: the three circular rings of figure 7.3 (obtained one from each other by dilations of factor 2) contain approximately the same number of frequencies. This is the same distribution as in the case of a wavelet transform.

Figure 7.2 (c) shows, for each of the 48 filters, its closest approximation in a frame of real Gabor wavelets. Except for filters $11, 25, 43, 46$ and 47, the approximation is good.

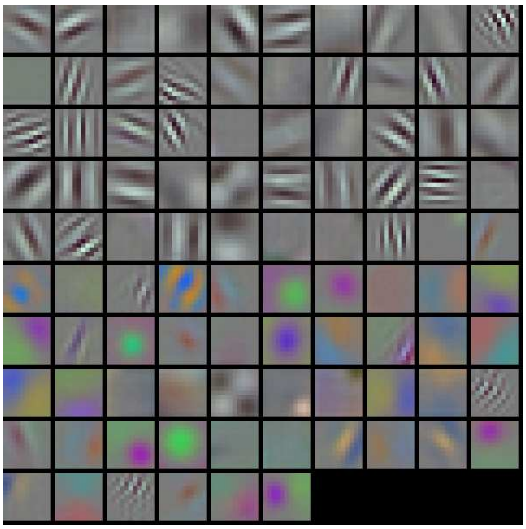
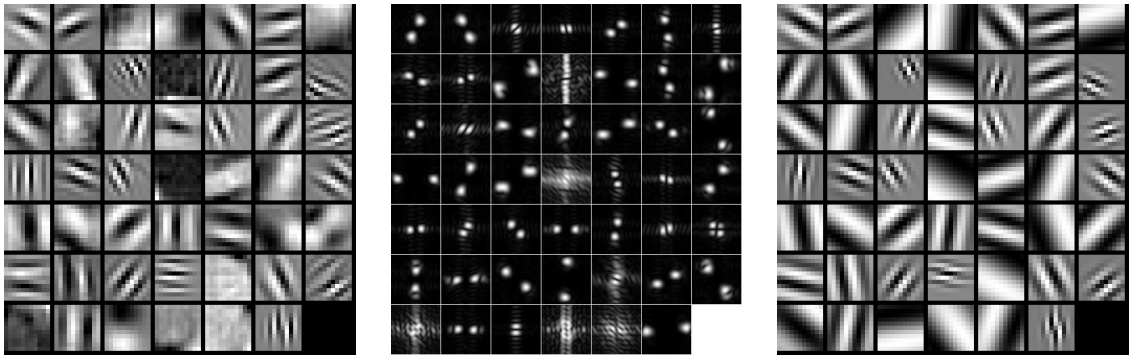


Figure 7.1: The 92 filters of the first layer of AlexNet. Each filter has size $3 \times 11 \times 11$.



(a) (b) (c)

Figure 7.2: (a) the 48 first filters of the first layer of AlexNet (normalized for a better visibility) (b) their Fourier transform (amplitude only) (c) their approximation in a frame of Gabor wavelets

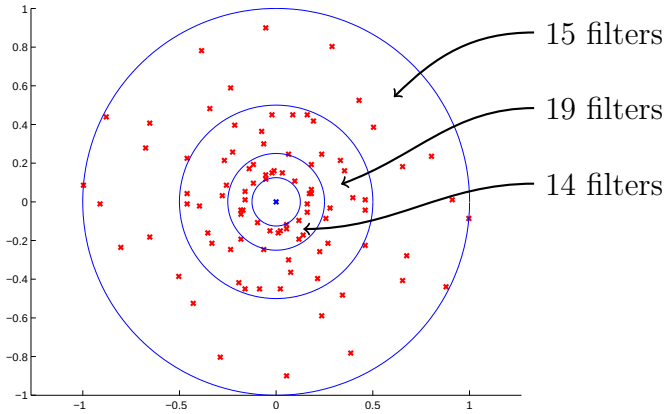


Figure 7.3: Distribution of characteristic frequencies, for the filters of Figure 7.2. The frequencies have been normalized so as to fit in the square $[-1;1]\times[-1;1]$. The blue circles divide the space in three dyadic frequential rings; the number of frequencies in each ring is indicated on the right.

Non-linearity

In AlexNet (as in most deep networks), the filters are real-valued. This is a priori an important difference with the complex wavelet transforms used throughout this thesis. Moreover, the non-linearity used in AlexNet is a Rectified Linear Unit (ReLU: $x \to \max(0,x)$), followed by a max-pooling; it is not a modulus.

However, depending on the choice some parameters, the combination "real wavelet transform + ReLU + max-pooling" can produce coefficients very similar to the ones obtained with a complex wavelet transform followed by a modulus.

Indeed, if ψ is a complex-valued wavelet, sufficiently localized in frequency, with characteristic frequency $\nu \in \mathbb{R}^2$, then, for each real-valued function f, the convolved function $f \star \psi$ is of the form:

$$f \star \psi(x) \approx f_0(x) e^{i\langle \nu, x\rangle} \qquad (x \in \mathbb{R}^2)$$

where f_0 is low-frequency.
So:

$$|f \star \psi(x)| = |f_0(x)| \qquad (x \in \mathbb{R}^2) \tag{7.1}$$

If one replaces the complex wavelet ψ by its real part $\mathrm{Re}\,(\psi) = (\psi + \overline{\psi})/2$, then, because f is real-valued:

$$\begin{aligned} f \star \mathrm{Re}\,(\psi)(x) &= \mathrm{Re}\,(f \star \psi)(x) \\ &= \mathrm{Re}\,(f_0(x) e^{i\langle \nu, x\rangle}) \end{aligned}$$

As f_0 is a low-frequency function, it can be considered approximately constant in the neighborhood of each point x_0: $f_0(x) \approx |f_0(x_0)|e^{i\phi_0}$, where ϕ_0 is a phase depending on x_0. So:

$$f \star \mathrm{Re}\,(\psi)(x) \approx \mathrm{Re}\,(|f_0(x_0)|e^{i(\phi_0+\langle\nu,x\rangle)}) = |f_0(x_0)|\cos(\phi_0 + \langle\nu, x\rangle)$$

Combining with ReLU and max-pooling yields:

$$\begin{aligned}\text{max-pool}(\text{ReLU}(f \star \mathrm{Re}\,(\psi)))(x) &\approx \max_{||y-x||_1 \leq M} \Big(\max\Big(0, |f_0(x_0)|\cos(\phi_0 + \langle\nu, y\rangle)\Big)\Big) \\ &= |f_0(x_0)| \max_{||y-x||_1 \leq M} \Big(\max\Big(0, \cos(\phi_0 + \langle\nu, y\rangle)\Big)\Big)\end{aligned}$$

where M is the spatial extent of the max-pooling.
If $2M$ is larger than the period of the cosine function, $2\pi/||\nu||_2$, then the cosine always reaches 1 on the set $\{||y - x||_1 \leq M\}$, whatever the value of x is. So:

$$\text{max-pool}(\text{ReLU}(f \star \mathrm{Re}\,(\psi)))(x) \approx |f_0(x_0)| \tag{7.2}$$

which is exactly Equation (7.1).

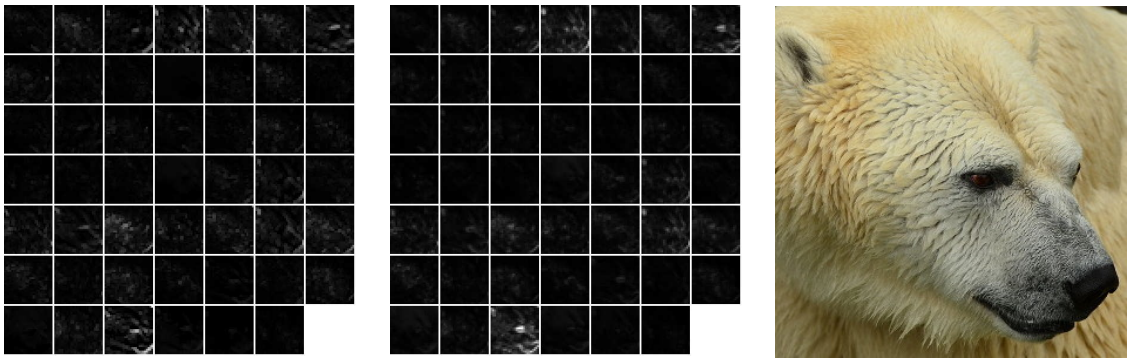
This reasoning seems to apply for AlexNet. Figure 7.4 (a) displays, for the input image shown in (c), the coefficients returned by AlexNet after the first max-pooling layer (for the first 48 channels). In subfigure (b) are displayed the same coefficients, where the real filters have been replaced by complex ones, and the combination of ReLU and max-pooling has been replaced by a modulus. The coefficients in (a) and (b) are not exactly identical, but all the same very similar. In particular, in each case, the non-negligible coefficients are located in the same channels, at the same spatial position.

Second layer

At the second layer, there are 256 filters. Again, we focus on the first half. Each filter has dimensions $48 \times 5 \times 5$. The first dimension corresponds to the indexes of the 48 first filters of the first layer. As first-layer filters are localized in the frequency domain, it can be seen a frequential index. The second and third dimensions are spatial variables.

As for the first layer, we would like to describe the transformation performed by these filters in terms of a simple operator like the wavelet transform. This question is still unsolved for us; in this last paragraph, we limit ourselves to preliminary considerations. The key point is to understand the interaction between frequential and spatial variables.

To partially overcome the difficulty raised by the three-dimensionality of filters, a possible approach is to use the fact that most filters at this layer can be written as a linear combination



(a) (b) (c)

Figure 7.4: (a) Coefficients after the first max-pooling step (b) Same coefficients, where real filters have been converted to analytic ones and the max-pooling has been replaced by a modulus (c) Input image (source: Antoine Letarte, http://commons.wikimedia.org/wiki/File:Polar_bear_headshot_2011.jpg)

of a small number of filters that are separable in space and frequency variables:

$$\psi_{\nu,k,l} = \sum_{s=1}^{S} \psi_{\nu}^{s,freq} \psi_{k,l}^{s,spat} \qquad \nu = 1, ..., 48 \quad k, l = 1, ..., 5$$

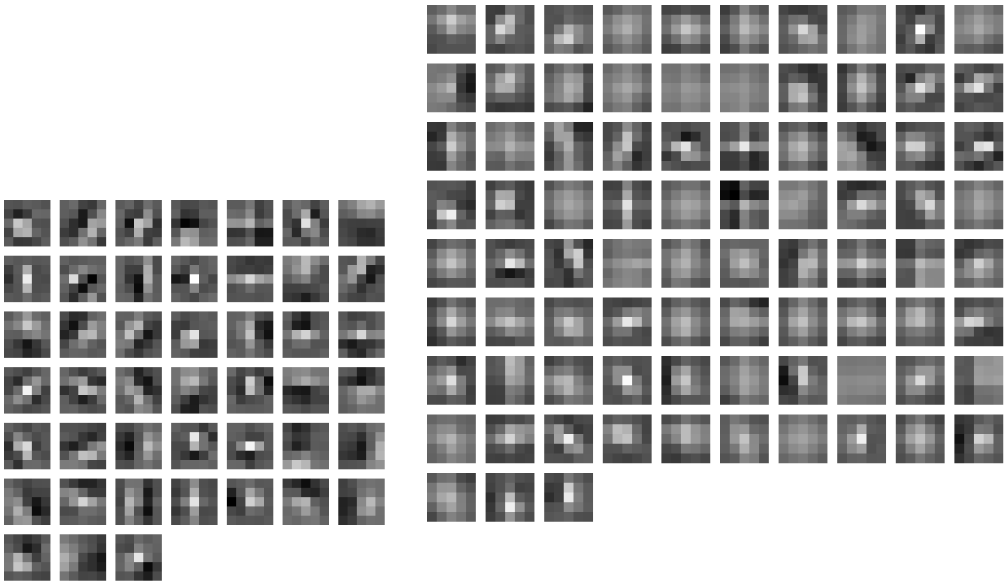
The integer S is small, typically of the order 5. This observation is not new; it has in particular been used to reduce the computational time needed to solve classification tasks with deep networks [Jaderberg et al., 2014].

Although S is in general not equal to 1, the first term of the decomposition, $\psi_{\nu}^{1,freq} \psi_{k,l}^{1,spat}$, already gives an idea of the operations performed by second-layer filters.

The spatial components $\psi^{1,spat}$ are displayed on Figure 7.5. They fall into two categories: zero-mean filters which again tend to look like wavelets (subfigure (a)) and filters with non-zero mean, the majority of which seem to perform local averages (subfigure(b)).

The frequential components $\psi_{\nu}^{1,freq}$ are shown on Figure 7.6. Each component is seen as a function from $\mathbb{R}^2$ to $\mathbb{R}$: to a point $(x, y) \in \mathbb{R}^2$ we associate the value $\psi_{\nu}^{1,freq}$, for ν the index of the first-layer filter whose characteristic frequency is the closest to (x, y).

Some filters are almost constant; they seem to perform averages in the frequential domain (last filter of the fifth row or second filter of the third row). Others, in polar coordinates (ρ, θ), do not vary much in r but present oscillations in θ (sixth filter of the first row or third filter of the seventh row). These filters are similar to the ones of the second-layer of scattering networks used for image analysis: in these networks, the second layer is a joint wavelet transform in both spatial and frequential variables (which allows to achieve invariance or stability to the action



(a) (b)

Figure 7.5: Spatial components for the filters of the second layer of AlexNet. (a) with zero mean (b) with non-zero mean

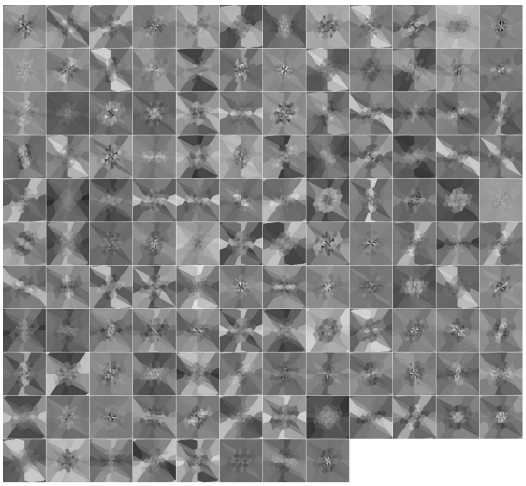


Figure 7.6: Frequential components for the filters of the second layer of AlexNet. Each component is pictured as a function from the frequency plane $\mathbb{R}^2$ to $\mathbb{R}$.

of the rotation group) [Sifre and Mallat, 2013; Oyallon and Mallat, 2015]. There are also more complicated shapes.

Future work will consist in precisely analyzing these second-layer filters, and see to what extent they can be replaced by operators with simple analytic expressions. We can then determine what information it brings on the geometric properties of images that deep networks implicitly use, and see if it suggests possible ways to improve the first layers of deep representations.

Contents

Bibliography

K. Achan, S. T. Roweis, and B. J. Frey. Probabilistic inference of speech signals from phaseless spectrograms. In *Advances in Neural Information Processing Systems 16*, pages 1393–1400, 2004.

E. J. Akutowicz. On the determination of the phase of a Fourier integral, I. *Transactions of the American Mathematical Society*, 83(1):179–192, 1956.

E. J. Akutowicz. On the determination of the phase of a Fourier integral, II. *Proceedings of the American Mathematical Society*, 8(2):234–238, 1957.

B. Alexeev, A. S. Bandeira, M. Fickus, and D. G. Mixon. Phase retrieval with polarization. *SIAM Journal on Imaging Sciences*, 7:35–66, 2013.

J. Andén and S. Mallat. Multiscale scattering for audio classification. In *Proceedings of the International Society of Music Information Retrieval 2011 Conference*, 2011.

S. Andreys and P. Jaming. Zak transform and non-uniqueness in an extension of pauli's phase retrieval problem. *Preprint*, 2015. https://hal.archives-ouvertes.fr/hal-01103583v1.

A. Antoniadis. Wavelets in statistics: a review. *Journal of the Italian Statistical Society*, 6(2): 97–144, 1997.

R. Balan and Y. Wang. Invertibility and robustness of phaseless reconstruction. *Applied and Computational Harmonic Analysis*, 38(3):469–488, 2015.

R. Balan, P. Casazza, and D. Edidin. On signal reconstruction without noisy phase. *Applied and Computational Harmonic Analysis*, 20:345–356, 2006.

A. S. Bandeira, J. Cahill, D. G. Mixon, and A. A. Nelson. Saving phase: Injectivity and stability for phase retrieval. *Applied and Computational Harmonic Analysis*, 37(1):106–125, 2014.

R. Barakat and G. Newsam. Necessary conditions for a unique solution to two-dimensional phase recovery. *Journal of Mathematical Physics*, 25(11):3190–3193, 1984.

H. H. Bauschke, P. L. Combettes, and D. R. Luke. Phase retrieval, error reduction algorithm, and fienup variants: a view from convex optimization. *Journal of the Optical Society of America*, 19:1334–1345, 2002.

S. Becker, E. J. Candès, and M. Grant. Tfocs v1.1. User guide, 2012.

A. Ben-Tal and A. S. Nemirovski. *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*. Society for Industrial and Applied Mathematics, 2001.

A. Ben-tal, A. Nemirovski, and C. Roos. Extended matrix cube theorems with applications to mu-theory in control. *Mathematics of Operations Research*, 28(3):497–523, 2003.

A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust Optimization*. Princeton Series in Applied Mathematics. Princeton University Press, 2009.

Y. Bengio, A. Courville, and P. Vincent. Representation learning: a review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:1798–1828, 2013.

D. Bertsekas. *Nonlinear programming*. Athena scientific, 1998.

R. Bhatia. *Matrix analysis*, volume 169. Springer Verlag, 1997.

B. G. Bodmann and N. Hammen. Stable phase retrieval with low-redundancy frames. *Advances in Computational Mathematics*, 41(2):317–331, 2014.

J. Bouvrie and T. Ezzat. An incremental algorihtm for signal reconstruction from stft magnitude. In *International conference on spoken language processing*, 2006.

S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

J. Bruna. *Scattering representations for recognition*. PhD thesis, École Polytechnique, Palaiseau, 2013.

J. Bruna and S. Mallat. Invariant scattering convolution networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1872–1886, 2013a.

J. Bruna and S. Mallat. Audio texture synthesis with scattering moments. *Preprint*, 2013b. http://arxiv.org/abs/1311.0407.

J. Bruna, S. Mallat, E. Bacry, and J.-F. Muzy. Intermittent process analysis with scattering moments. *Annals of Statistics*, 43(1):323–351, 2015.

O. Bunk, A. Diaz, F. Pfeiffer, C. David, B. Schmitt, D. K. Satapathy, and J. F. van der Veen. Diffractive imaging for periodic samples: retrieving one-dimensional concentration profiles across microfluidic channels. *Acta Crystallographica Section A*, 63(4):306–314, 2007.

E. J. Candès and X. Li. Solving quadratic equations via phaselift when there are about as many equations as unknowns. *Foundations of Computational Mathematics*, 14(5):1017–1026, 2014.

E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

E. J. Candès, Y. C. Eldar, T. Strohmer, and V. Voroninski. Phase retrieval via matrix completion. *SIAM Journal on Imaging Sciences*, 6(1):199–225, 2011.

E. J. Candès, T. Strohmer, and V. Voroninski. Phaselift: exact and stable signal recovery from magnitude measurements via convex programming. *Communications in Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

E. J. Candès, X. Li, and M. Soltanolkotabi. Phase retrieval from coded diffraction patterns. *Applied and Computational Harmonic Analysis*, 39(2):277–299, 2015.

E. J. Candès, X. Li, and M. Soltanolkotabi. Phase retrieval via wirtinger flow: theory and algorithms. *IEEE Transactions of Information Theory*, 61(4):1985–2007, 2015.

A. Chai, M. Moscoso, and G. Papanicolaou. Array imaging using intensity-only measurements. *Inverse Problems*, 27(1), 2011.

V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

X. Chen, X. Cheng, and S. Mallat. Unsupervised deep haar scattering on graphs. In *Avances in Neural Information Processing Systems 25*, pages 1709–1717, 2014.

A. Cohen. *Ondelettes et traitement numérique du signal*. Recherches en mathématiques appliquées. Masson, 1992.

A. Conca, D. Edidin, M. Hering, and C. Vinzant. Algebraic characterization of injectivity in phase retrieval. *Applied and Computational Harmonic Analysis*, 32(2):346–356, 2015.

J. C. Dainty and J. R. Fienup. Phase retrieval and image reconstruction for astronomy. In H. Stark, editor, *Image recovery: theory and application*, pages 231–275. Academic Press, 1987.

N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 886–893, 2005.

A. d'Aspremont, O. Banerjee, and L. El Ghaoui. First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and Applications*, 30(1):56–66, 2008.

I. Daubechies. *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, 1992.

C. Delorme and S. Poljak. Laplacian eigenvalues and the maximum cut problem. *Mathematical Programming*, 62(3):557–574, 1993.

L. Demanet and P. Hand. Stable optimizationless recovery from phaseless linear measurements. *Journal of Fourier Analysis and Applications*, 20(1):199–221, 2012.

R. A. DeVore, B. Jawerth, and V. Popov. Compression of wavelet decompositions. *American Journal of Mathematics*, 114(4):737–785, 1992.

A. Dosovitskiy and T. Brox. Inverting convolutional networks with convolutional networks. Technical report, , 2015. http://arxiv.org/abs/1506.02753.

Y. C. Eldar, P. Sidorenko, D. G. Mixon, S. Barel, and O. Cohen. Sparse phase retrieval from short-time fourier measurements. *IEEE Signal Processing Letters*, 22(5):638–642, 2015.

M. Fazel, H. Hindi, and S. P. Boyd. Log-det heuristic for matrix rank minimization with applications to hankel and euclidean distance matrices. *American Control Conference*, 3: 2156–2162, 2003.

J. R. Fienup. Phase retrieval algorithms: a comparison. *Applied Optics*, 21(15):2758–2769, 1982.

F. Fogel, I. Waldspurger, and A. d'Aspremont. Phase retrieval for imaging problems. *to appear in Mathematical programming computation*, 2013.

John B. Garnett. *Bounded analytic functions*. Academic Press, 1981.

R. Gerchberg and W. Saxton. A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik*, 35:237–246, 1972.

M. X. Goemans and D. P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM*, 42(6):1115–1145, 1995.

M. X. Goemans and D. P. Williamson. Approximation algorithms for max-3-cut and other problems via complex semidefinite programming. *Journal of Computer and System Sciences*, 68(2):442 – 470, 2004.

D. Griffin and J. S. Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 32:236–243, 1984.

D. Gross, F. Krahmer, and R. Kueng. A partial derandomization of PhaseLift using spherical designs. *Journal of Fourier Analysis and Applications*, 21(2):229–266, 2015a.

D. Gross, F. Krahmer, and R. Kueng. Improved recovery guarantees for phase retrieval from coded diffraction patterns. *To appear in Applied and Computational Harmonic Analysis*, 2015b.

M. H. Hayes. The reconstruction of a multidimensional sequence from the phase or magnitude of its fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 30 (2):140–154, 1982.

C. Helmberg, F. Rendl, R. J. Vanderbei, and H. Wolkowicz. An interior-point method for semidefinite programming. *SIAM Journal on Optimization*, 6(2):342–361, 1996.

G. Hinton, L. Deng, D. Yu, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. S. G. Dahl, and B. Kingsbury. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

M. Hirn, N. Poilvert, and S. Mallat. Quantum energy regression using scattering transforms. *Submitted to the International Conference on Machine Learning*, 2015. http://arxiv.org/abs/1502.02077.

M. Jaderberg, A. Vedaldi, and A. Zisserman. Speeding up convolutional neural networks with low rank expansions. In *British Machine Vision Conference*, 2014.

P. Jaming. Uniqueness results in an extension of pauli's phase retrieval problem. *Applied and Computational Harmonic Analysis*, 37:413–441, 2014.

Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678, 2014.

N. El Karoui and A. d'Aspremont. Second-order accurate distributed eigenvector computation for extremely large matrices. *Electronical Journal of Statistics*, 4, 2010.

T. Kato. *Perturbation theory for linear operators*. Springer, 1995.

M. Kisialiou and Z.-Q. Luo. Probabilistic analysis of semidefinite relaxation for binary quadratic minimization. *SIAM Journal on Optimization*, 20(4):1906–1922, 2010.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105, 2012.

W. Kryloff. On functions which are regular in a half-plane. *Matematicheskii Sbornik*, 6, 1939.

X. Li and V. Voroninski. Sparse signal recovery from quadratic measurements via convex programming. *SIAM Journal on Mathematical Analysis*, 45(5):3019–3033, 2013.

L. Lovász and A. Schrijver. Cones of matrices and set-functions and 0-1 optimization. *SIAM Journal on Optimization*, 1(2):166–190, 1991.

D. G. Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Vision*, pages 1150–1157, 1999.

Z.-Q. Luo, X. Luo, and M. Kisialiou. An efficient quasi-maximum likelihood decoder for PSK signals. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 561–564, 2003.

A. Mahendran and A. Vedaldi. Understanding deep image representations by inverting them. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 5188–5196, 2015.

S. Mallat. *A wavelet tour of signal processing. The Sparse Way. Third Edition.* Elsevier, 2009.

S. Mallat. Group invariant scattering. *Communications in Pure and Applied Mathematics*, 65 (10):1331–1398, 2012.

S. Mallat and I. Waldspurger. Deep learning by scattering. *Preprint*, 2013. http://arxiv.org/abs/1306.5532.

S. Mallat and I. Waldspurger. Phase retrieval for the cauchy wavelet transform. *to appear in the Journal of Fourier Analysis and Applications*, 2014.

J. Miao, T. Ishikawa, Q. Shen, and T. Earnest. Extending x-ray crystallography to allow the imaging of noncrystalline materials, cells, and single protein complexes. *Annual Review of Physical Chemistry*, 59:387–410, 2008.

R. P. Millane. Phase retrieval in cristallography and optics. *Journal of Optical Society of America A*, 7(3):394–411, 1990.

M. L. Moravec, J. K. Romberg, and R. G. Baraniuk. Compressive phase retrieval. In *Wavelets XII in SPIE International Symposium on Optical Science and Technology*, 2007.

S. H. Nawab, T. F. Quatieri, and J. S. Lim. Signal reconstruction from short-time fourier transform magnitude. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, pages 986–998, 1983.

Y. Nesterov. A method of solving a convex programming problem with convergence rate $o(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

Y. Nesterov. Semidefinite relaxation and nonconvex quadratic optimization. *Optimization Methods and Software*, 9(1-3):141–160, 1998.

Y. Nesterov. Smoothing technique and its applications in semidefinite optimization. *Mathematical Programming*, 110(2):245–259, 2007.

P. Netrapalli, P. Jain, and S. Sanghavi. Phase retrieval using alternating minimization. In *Advances in Neural Information Processing Systems 26*, pages 1796–2804, 2013.

J. Nocedal. Updating quasi-newton matrices with limited storage. *Mathematics of Computation*, pages 773–782, 1980.

E. Oyallon and S. Mallat. Deep roto-translation scattering for object classification. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 2865–2873, 2015.

F. Perronnin and D. Larlus. Fisher vectors meet neural networks: a hybrid classification architecture. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 3743–3752, 2015.

J. Ranieri, A. Chebira, Y. M. Lu, and M. Vetterli. Phase retrieval for sparse signals: uniqueness conditions. *Preprint*, 2013. http://arxiv.org/abs/1308.3058.

J.-C. Risset and D. L. Wessel. Exploration of timbre by analysis and synthesis. In D. Deutsch, editor, *The psychology of music*, pages 113–169. Academic Press, 1999.

W. Rudin. *Real and complex analysis, third edition*. McGraw-Hill International Editions, 1987.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, pages 1–42, 2015.

J. L. C. Sanz. Mathematical considerations for the problem of fourier transform phase retrieval from magnitude. *SIAM Journal on Applied Mathematics*, 45(4):651–664, 1985.

D. W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization, 2nd Edition*. Wiley, 2015.

I. W. Selesnick. Hilbert transform pairs of wavelet bases. *IEEE Signal Processing Letters*, 8, 2001.

P. Sermanet, K. Kavukcuogly, S. Chintala, and Y. LeCun. Pedestrian detection with unsupervised multi-stage feature learning. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 3626–3633, 2013.

Y. Shechtman, A. Szameit, E. Osherovic, E. Bullkich, H. Dana, S. Gazit, S. Shoham, M. Zibulevsky, I. Yavneh, E. B. Kley, Y. C. Eldar, O. Cohen, and M. Segev. Sparsity-based single-shot sub-wavelength coherent diffractive imaging. In *Frontiers in Optics 2011/Laser Science XXVII*, 2011.

N. Z. Shor. Quadratic optimization problems. *Soviet Journal of Computer and System Sciences*, 25, 1987.

L. Sifre and S. Mallat. Rotation, scaling and deformation invariant scattering for texture discrimination. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 1233–1240, 2013.

L. Sifre, J. Andén, M. Kapoko, E. Oyallon, and V. Lostanlen. Scatnet v0.2. Software, 2013. http://www.di.ens.fr/data/software/scatnet/.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutionnal networks: visualising image classification models and saliency maps. In *International Conference on Learning Representations Workshop*, 2014.

A. Singer. Angular synchronization by eigenvectors and semidefinite programming. *Applied and Computational Harmonic Analysis*, 30(1):20–36, 2011.

A. M.-C. So. Probabilistic analysis of the semidefinite relaxation detector in digital communications. In *Proceedings of the 21st annual ACM-SIAM Symposium on Discrete Algorithms*, pages 698–711, 2010.

A. M.-C. So and Y. Ye. Probabilistic analysis of semidefinite relaxation detectors for multiple-input, multiple-output. In D. P. Palomar and Y. C. Eldar, editors, *Convex optimization in signal processing and applications*. Cambridge University Press, 2010.

A. M.-C. So, J. Zhang, and Y. Ye. On approximating complex quadratic optimization problems via semidefinite programming relaxations. *Mathematical Programming*, 110(1):93–110, 2007.

G. W. Stewart. *Matrix algorithms, Volume II: eigensystems*. SIAM, 2001.

G. W. Stewart and J. Sun. *Matrix perturbation theory*. Academic Press, 1990.

N. Sturmel and L. Daudet. Signal reconstruction from STFT magnitude: a state of the art. In *Proceedings of the 14th international conference on digital audio effects*, pages 375–386, 2011.

D. L. Sun and J. O. Smith. Estimating a signal from a magnitude spectrogram via convex optimization. In *Audio Engineering Society 133rd Convention*, 2012.

C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

K. C. Toh, M. J. Todd, and T. H. Tutuncu. Sdpt3 - a MATLAB software package for semidefinite programming. *Optimization Methods and Software*, 11:545–581, 1999.

T. Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):1066–1074, 2007.

V. Voroninski. A comparison between the phaselift and phasecut algorithms. Working paper, 2012.

I. Waldspurger, A. d'Aspremont, and S. Mallat. Phase recovery, maxcut and complex semidefinite programming. *Mathematical Programming*, 149(1-2):47–81, 2015.

A. Walther. The question of phase retrieval in optics. *Optica Acta*, 10(1):41–49, 1963.

Z. Wen, D. Goldfarb, and K. Scheinberg. Block coordinate descent methods for semidefinite programming. In *Handbook on semidefinite, conic and polynomial optimization*, pages 533–564. Springer, 2012.

X. Yang, K. Wang, and S. A. Shamma. Auditory representations of acoustic signals. *IEEE Transactions on Information Theory*, 38(2):824–839, 1992.

E. A. Yildirim. An interior-point perspective on sensitivity analysis in semidefinite programming. *Mathematics of Operations Research*, 28(4):649–676, 2003.

E. A. Yildirim and M. Todd. Sensitivity analysis in linear programming and semidefinite programming using interior-point methods. *Mathematical Programming*, 90(2):229–261, 2001.

J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems 27*, pages 3320–3328, 2014.

M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Proceedings of the European Conference on Computer Vision*, pages 818–833, 2014.

S. Zhang and Y. Huang. Complex quadratic optimization and semidefinite programming. *SIAM Journal on Optimization*, 16(3):871–890, 2006.