

# a wavelet tour of signal processing

The Sparse Way



Third Edition

Stéphane Mallat



Academic Press is an imprint of Elsevier  
30 Corporate Drive, Suite 400  
Burlington, MA 01803

This book is printed on acid-free paper. 

Copyright © 2009 by Elsevier Inc. All rights reserved.

Designations used by companies to distinguish their products are often claimed as trade-marks or registered trademarks. In all instances in which Academic Press is aware of a claim, the product names appear in initial capital or all capital letters. Readers, however, should contact the appropriate companies for more complete information regarding trademarks and registration.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, scanning, or otherwise, without prior written permission of the publisher.

Permissions may be sought directly from Elsevier's Science & Technology Rights Department in Oxford, UK: phone: (+44) 1865 843830, fax: (+44) 1865 853333, e-mail: permissions@elsevier.com. You may also complete your request on-line via the Elsevier homepage (<http://elsevier.com>), by selecting "Support & Contact" then "Copyright and Permission" and then "Obtaining Permissions."

**Library of Congress Cataloging-in-Publication Data**  
Application submitted

ISBN 13: 978-0-12-374370-1

For information on all Academic Press publications,  
visit our Website at [www.books.elsevier.com](http://www.books.elsevier.com)

Printed in the United States  
08 09 10 11 12 10 9 8 7 6 5 4 3 2 1

Working together to grow  
libraries in developing countries

[www.elsevier.com](http://www.elsevier.com) | [www.bookaid.org](http://www.bookaid.org) | [www.sabre.org](http://www.sabre.org)

ELSEVIER BOOK AID International Sabre Foundation

*À la mémoire de mon père, Alexandre.  
Pour ma mère, Francine.*

# Preface to the Sparse Edition

I cannot help but find striking resemblances between scientific communities and schools of fish. We interact in conferences and through articles, and we move together while a global trajectory emerges from individual contributions. Some of us like to be at the center of the school, others prefer to wander around, and a few swim in multiple directions in front. To avoid dying by starvation in a progressively narrower and specialized domain, a scientific community needs also to move on. Computational harmonic analysis is still very much alive because it went beyond wavelets. Writing such a book is about decoding the trajectory of the school and gathering the pearls that have been uncovered on the way. Wavelets are no longer the central topic, despite the previous edition's original title. It is just an important tool, as the Fourier transform is. Sparse representation and processing are now at the core.

In the 1980s, many researchers were focused on building time-frequency decompositions, trying to avoid the uncertainty barrier, and hoping to discover the ultimate representation. Along the way came the construction of wavelet orthogonal bases, which opened new perspectives through collaborations with physicists and mathematicians. Designing orthogonal bases with Xlets became a popular sport with compression and noise-reduction applications. Connections with approximations and sparsity also became more apparent. The search for sparsity has taken over, leading to new grounds where orthonormal bases are replaced by redundant dictionaries of waveforms.

During these last seven years, I also encountered the industrial world. With a lot of naiveness, some bandlets, and more mathematics, I cofounded a start-up with Christophe Bernard, Jérôme Kalifa, and Erwan Le Pennec. It took us some time to learn that in three months good engineering should produce robust algorithms that operate in real time, as opposed to the three years we were used to having for writing new ideas with promising perspectives. Yet, we survived because mathematics is a major source of industrial innovations for signal processing. Semiconductor technology offers amazing computational power and flexibility. However, ad hoc algorithms often do not scale easily and mathematics accelerates the trial-and-error development process. Sparsity decreases computations, memory, and data communications. Although it brings beauty, mathematical understanding is not a luxury. It is required by increasingly sophisticated information-processing devices.

## New Additions

Putting sparsity at the center of the book implied rewriting many parts and adding sections. Chapters 12 and 13 are new. They introduce sparse representations in redundant dictionaries, and inverse problems, super-resolution, and

compressive sensing. Here is a small catalog of new elements in this third edition:

- Radon transform and tomography
- Lifting for wavelets on surfaces, bounded domains, and fast computations
- JPEG-2000 image compression
- Block thresholding for denoising
- Geometric representations with adaptive triangulations, curvelets, and bandlets
- Sparse approximations in redundant dictionaries with pursuit algorithms
- Noise reduction with model selection in redundant dictionaries
- Exact recovery of sparse approximation supports in dictionaries
- Multichannel signal representations and processing
- Dictionary learning
- Inverse problems and super-resolution
- Compressive sensing
- Source separation

## Teaching

This book is intended as a graduate-level textbook. Its evolution is also the result of teaching courses in electrical engineering and applied mathematics. A new website provides software for reproducible experimentations, exercise solutions, together with teaching material such as slides with figures and MATLAB software for numerical classes of <http://wavelet-tour.com>.

More exercises have been added at the end of each chapter, ordered by level of difficulty. Level<sup>1</sup> exercises are direct applications of the course. Level<sup>2</sup> exercises require more thinking. Level<sup>3</sup> includes some technical derivation exercises. Level<sup>4</sup> are projects at the interface of research that are possible topics for a final course project or independent study. More exercises and projects can be found in the website.

## Sparse Course Programs

The Fourier transform and analog-to-digital conversion through linear sampling approximations provide a common ground for all courses (Chapters 2 and 3). It introduces basic signal representations and reviews important mathematical and algorithmic tools needed afterward. Many trajectories are then possible to explore and teach sparse signal processing. The following list notes several topics that can orient a course's structure with elements that can be covered along the way.

Sparse representations with bases and applications:

- Principles of linear and nonlinear approximations in bases (Chapter 9)
- Lipschitz regularity and wavelet coefficients decay (Chapter 6)
- Wavelet bases (Chapter 7)
- Properties of linear and nonlinear wavelet basis approximations (Chapter 9)
- Image wavelet compression (Chapter 10)
- Linear and nonlinear diagonal denoising (Chapter 11)

Sparse time-frequency representations:

- Time-frequency wavelet and windowed Fourier ridges for audio processing (Chapter 4)
- Local cosine bases (Chapter 8)
- Linear and nonlinear approximations in bases (Chapter 9)
- Audio compression (Chapter 10)
- Audio denoising and block thresholding (Chapter 11)
- Compression and denoising in redundant time-frequency dictionaries with best bases or pursuit algorithms (Chapter 12)

Sparse signal estimation:

- Bayes versus minimax and linear versus nonlinear estimations (Chapter 11)
- Wavelet bases (Chapter 7)
- Linear and nonlinear approximations in bases (Chapter 9)
- Thresholding estimation (Chapter 11)
- Minimax optimality (Chapter 11)
- Model selection for denoising in redundant dictionaries (Chapter 12)
- Compressive sensing (Chapter 13)

Sparse compression and information theory:

- Wavelet orthonormal bases (Chapter 7)
- Linear and nonlinear approximations in bases (Chapter 9)
- Compression and sparse transform codes in bases (Chapter 10)
- Compression in redundant dictionaries (Chapter 12)
- Compressive sensing (Chapter 13)
- Source separation (Chapter 13)

Dictionary representations and inverse problems:

- Frames and Riesz bases (Chapter 5)
- Linear and nonlinear approximations in bases (Chapter 9)
- Ideal redundant dictionary approximations (Chapter 12)
- Pursuit algorithms and dictionary incoherence (Chapter 12)
- Linear and thresholding inverse estimators (Chapter 13)
- Super-resolution and source separation (Chapter 13)
- Compressive sensing (Chapter 13)

Geometric sparse processing:

- Time-frequency spectral lines and ridges (Chapter 4)
- Frames and Riesz bases (Chapter 5)
- Multiscale edge representations with wavelet maxima (Chapter 6)
- Sparse approximation supports in bases (Chapter 9)
- Approximations with geometric regularity, curvelets, and bandlets (Chapters 9 and 12)
- Sparse signal compression and geometric bit budget (Chapters 10 and 12)
- Exact recovery of sparse approximation supports (Chapter 12)
- Super-resolution (Chapter 13)

## ACKNOWLEDGMENTS

Some things do not change with new editions, in particular the traces left by the ones who were, and remain, for me important references. As always, I am deeply grateful to Ruzena Bajcsy and Yves Meyer.

I spent the last few years with three brilliant and kind colleagues—Christophe Bernard, Jérôme Kalifa, and Erwan Le Pennec—in a pressure cooker called a “start-up.” Pressure means stress, despite very good moments. The resulting sauce was a blend of what all of us could provide, which brought new flavors to our personalities. I am thankful to them for the ones I got, some of which I am still discovering.

This new edition is the result of a collaboration with Gabriel Peyré, who made these changes not only possible, but also very interesting to do. I thank him for his remarkable work and help.

*Stéphane Mallat*

---

## ACKNOWLEDGMENTS

Some things do not change with new editions, in particular the traces left by the ones who were, and remain, for me important references. As always, I am deeply grateful to Ruzena Bajcsy and Yves Meyer.

I spent the last few years with three brilliant and kind colleagues—Christophe Bernard, Jérôme Kalifa, and Erwan Le Pennec—in a pressure cooker called a “start-up.” Pressure means stress, despite very good moments. The resulting sauce was a blend of what all of us could provide, which brought new flavors to our personalities. I am thankful to them for the ones I got, some of which I am still discovering.

This new edition is the result of a collaboration with Gabriel Peyré, who made these changes not only possible, but also very interesting to do. I thank him for his remarkable work and help.

*Stéphane Mallat*

# Notations

$\langle f, g \rangle$	Inner product (A.6)
$\ f\ $	Euclidean or Hilbert space norm
$\ f\ _1$	$L^1$ or $\ell^1$ norm
$\ f\ _\infty$	$L^\infty$ norm
$f[n] = O(g[n])$	Order of: there exists $K$ such that $f[n] \leq K g[n]$
$f[n] = o(g[n])$	Small order of: $\lim_{n \rightarrow +\infty} \frac{f[n]}{g[n]} = 0$
$f[n] \sim g[n]$	Equivalent to: $f[n] = O(g[n])$ and $g[n] = O(f[n])$
$A < +\infty$	$A$ is finite
$A \gg B$	$A$ is much bigger than $B$
$z^*$	Complex conjugate of $z \in \mathbb{C}$
$[x]$	Largest integer $n \leq x$
$\lceil x \rceil$	Smallest integer $n \geq x$
$(x)_+$	$\max(x, 0)$
$n \bmod N$	Remainder of the integer division of $n$ modulo $N$

## Sets

$\mathbb{N}$	Positive integers including 0
$\mathbb{Z}$	Integers
$\mathbb{R}$	Real numbers
$\mathbb{R}^+$	Positive real numbers
$\mathbb{C}$	Complex numbers
$ \Lambda $	Number of elements in a set $\Lambda$

## Signals

$f(t)$	Continuous time signal
$f[n]$	Discrete signal
$\delta(t)$	Dirac distribution (A.30)
$\delta[n]$	Discrete Dirac (3.32)
$\mathbf{1}_{[a,b]}$	Indicator of a function that is 1 in $[a, b]$ and 0 outside

## Spaces

$C_0$	Uniformly continuous functions (7.207)
$C^p$	$p$ times continuously differentiable functions
$C^\infty$	Infinitely differentiable functions
$\mathbf{W}^s(\mathbb{R})$	Sobolev <sup>s</sup> times differentiable functions (9.8)
$L^2(\mathbb{R})$	Finite energy functions $\int  f(t) ^2 dt < +\infty$
$L^p(\mathbb{R})$	Functions such that $\int  f(t) ^p dt < +\infty$
$\ell^2(\mathbb{Z})$	Finite energy discrete signals $\sum_{n=-\infty}^{+\infty}  f[n] ^2 < +\infty$
$\ell^p(\mathbb{Z})$	Discrete signals such that $\sum_{n=-\infty}^{+\infty}  f[n] ^p < +\infty$
$\mathbb{C}^N$	Complex signals of size $N$
$\mathbf{U} \oplus \mathbf{V}$	Direct sum of two vector spaces

$\mathbf{U} \otimes \mathbf{V}$	Tensor product of two vector spaces (A.19)
$\text{Null}\mathbf{U}$	Null space of an operator $U$
$\text{Im}\mathbf{U}$	Image space of an operator $U$
<b>Operators</b>	
$\text{Id}$	Identity
$f'(t)$	Derivative $\frac{df(t)}{dt}$
$f^{(p)}(t)$	Derivative $\frac{d^p f(t)}{dt^p}$ of order $p$
$\vec{\nabla}f(x, y)$	Gradient vector (6.51)
$f \star g(t)$	Continuous time convolution (2.2)
$f \star g[n]$	Discrete convolution (3.33)
$f \circledast g[n]$	Circular convolution (3.73)
<b>Transforms</b>	
$\hat{f}(\omega)$	Fourier transform (2.6), (3.39)
$\hat{f}[k]$	Discrete Fourier transform (3.49)
$Sf(u, s)$	Short-time windowed Fourier transform (4.11)
$P_S f(u, \xi)$	Spectrogram (4.12)
$Wf(u, s)$	Wavelet transform (4.31)
$P_W f(u, \xi)$	Scalogram (4.55)
$P_V f(u, \xi)$	Wigner-Ville distribution (4.120)
<b>Probability</b>	
$X$	Random variable
$E\{X\}$	Expected value
$\mathcal{H}(X)$	Entropy (10.4)
$\mathcal{H}_d(X)$	Differential entropy (10.20)
$\text{Cov}(X_1, X_2)$	Covariance (A.22)
$F[n]$	Random vector
$R_F[k]$	Autocovariance of a stationary process (A.26)

# Sparse Representations

# 1

Signals carry overwhelming amounts of data in which relevant information is often more difficult to find than a needle in a haystack. Processing is faster and simpler in a sparse representation where few coefficients reveal the information we are looking for. Such representations can be constructed by decomposing signals over elementary waveforms chosen in a family called a *dictionary*. But the search for the Holy Grail of an ideal sparse transform adapted to all signals is a hopeless quest. The discovery of wavelet orthogonal bases and local time-frequency dictionaries has opened the door to a huge jungle of new transforms. Adapting sparse representations to signal properties, and deriving efficient processing operators, is therefore a necessary survival strategy.

An orthogonal basis is a dictionary of minimum size that can yield a sparse representation if designed to concentrate the signal energy over a set of few vectors. This set gives a geometric signal description. Efficient signal compression and noise-reduction algorithms are then implemented with diagonal operators computed with fast algorithms. But this is not always optimal.

In natural languages, a richer dictionary helps to build shorter and more precise sentences. Similarly, dictionaries of vectors that are larger than bases are needed to build sparse representations of complex signals. But choosing is difficult and requires more complex algorithms. Sparse representations in redundant dictionaries can improve pattern recognition, compression, and noise reduction, but also the resolution of new inverse problems. This includes superresolution, source separation, and compressive sensing.

This first chapter is a sparse book representation, providing the story line and the main ideas. It gives a sense of orientation for choosing a path to travel.

---

## 1.1 COMPUTATIONAL HARMONIC ANALYSIS

Fourier and wavelet bases are the journey's starting point. They decompose signals over oscillatory waveforms that reveal many signal properties and provide a path to sparse representations. Discretized signals often have a very large size  $N \geq 10^6$ , and thus can only be processed by fast algorithms, typically implemented with  $O(N \log N)$  operations and memories. Fourier and wavelet transforms

illustrate the strong connection between well-structured mathematical tools and fast algorithms.

### 1.1.1 The Fourier Kingdom

The Fourier transform is everywhere in physics and mathematics because it diagonalizes time-invariant convolution operators. It rules over linear time-invariant signal processing, the building blocks of which are *frequency filtering* operators.

Fourier analysis represents any finite energy function  $f(t)$  as a sum of sinusoidal waves  $e^{i\omega t}$ :

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \hat{f}(\omega) e^{i\omega t} d\omega. \quad (1.1)$$

The amplitude  $\hat{f}(\omega)$  of each sinusoidal wave  $e^{i\omega t}$  is equal to its correlation with  $f$ , also called Fourier transform:

$$\hat{f}(\omega) = \int_{-\infty}^{+\infty} f(t) e^{-i\omega t} dt. \quad (1.2)$$

The more regular  $f(t)$ , the faster the decay of the sinusoidal wave amplitude  $|\hat{f}(\omega)|$  when frequency  $\omega$  increases.

When  $f(t)$  is defined only on an interval, say  $[0, 1]$ , then the Fourier transform becomes a decomposition in a Fourier orthonormal basis  $\{e^{i2\pi m t}\}_{m \in \mathbb{Z}}$  of  $L^2[0, 1]$ . If  $f(t)$  is uniformly regular, then its Fourier transform coefficients also have a fast decay when the frequency  $2\pi m$  increases, so it can be easily approximated with few low-frequency Fourier coefficients. The Fourier transform therefore defines a sparse representation of uniformly regular functions.

Over discrete signals, the Fourier transform is a decomposition in a discrete orthogonal Fourier basis  $\{e^{i2\pi k n/N}\}_{0 \leq k < N}$  of  $\mathbb{C}^N$ , which has properties similar to a Fourier transform on functions. Its embedded structure leads to fast Fourier transform (FFT) algorithms, which compute discrete Fourier coefficients with  $O(N \log N)$  instead of  $N^2$ . This FFT algorithm is a cornerstone of discrete signal processing.

As long as we are satisfied with linear time-invariant operators or uniformly regular signals, the Fourier transform provides simple answers to most questions. Its richness makes it suitable for a wide range of applications such as signal transmissions or stationary signal processing. However, to represent a transient phenomenon—a word pronounced at a particular time, an apple located in the left corner of an image—the Fourier transform becomes a cumbersome tool that requires many coefficients to represent a localized event. Indeed, the support of  $e^{i\omega t}$  covers the whole real line, so  $\hat{f}(\omega)$  depends on the values  $f(t)$  for all times  $t \in \mathbb{R}$ . This global “mix” of information makes it difficult to analyze or represent any local property of  $f(t)$  from  $\hat{f}(\omega)$ .

### 1.1.2 Wavelet Bases

Wavelet bases, like Fourier bases, reveal the signal regularity through the amplitude of coefficients, and their structure leads to a fast computational algorithm.

However, wavelets are well localized and few coefficients are needed to represent local transient structures. As opposed to a Fourier basis, a wavelet basis defines a sparse representation of piecewise regular signals, which may include transients and singularities. In images, large wavelet coefficients are located in the neighborhood of edges and irregular textures.

The story began in 1910, when Haar [291] constructed a piecewise constant function

$$\psi(t) = \begin{cases} 1 & \text{if } 0 \leq t < 1/2 \\ -1 & \text{if } 1/2 \leq t < 1 \\ 0 & \text{otherwise} \end{cases}$$

the dilations and translations of which generate an orthonormal basis

$$\left\{ \psi_{j,n}(t) = \frac{1}{\sqrt{2^j}} \psi\left(\frac{t - 2^j n}{2^j}\right) \right\}_{(j,n) \in \mathbb{Z}^2}$$

of the space  $\mathbf{L}^2(\mathbb{R})$  of signals having a finite energy

$$\|f\|^2 = \int_{-\infty}^{+\infty} |f(t)|^2 dt < +\infty.$$

Let us write  $\langle f, g \rangle = \int_{-\infty}^{+\infty} f(t) g^*(t) dt$ —the inner product in  $\mathbf{L}^2(\mathbb{R})$ . Any finite energy signal  $f$  can thus be represented by its wavelet inner-product coefficients

$$\langle f, \psi_{j,n} \rangle = \int_{-\infty}^{+\infty} f(t) \psi_{j,n}(t) dt$$

and recovered by summing them in this wavelet orthonormal basis:

$$f = \sum_{j=-\infty}^{+\infty} \sum_{n=-\infty}^{+\infty} \langle f, \psi_{j,n} \rangle \psi_{j,n}. \quad (1.3)$$

Each Haar wavelet  $\psi_{j,n}(t)$  has a zero average over its support  $[2^j n, 2^j(n+1)]$ . If  $f$  is locally regular and  $2^j$  is small, then it is nearly constant over this interval and the wavelet coefficient  $\langle f, \psi_{j,n} \rangle$  is nearly zero. This means that large wavelet coefficients are located at sharp signal transitions only.

With a jump in time, the story continues in 1980, when Strömberg [449] found a piecewise linear function  $\psi$  that also generates an orthonormal basis and gives better approximations of smooth functions. Meyer was not aware of this result, and motivated by the work of Morlet and Grossmann over continuous wavelet transform, he tried to prove that there exists no regular wavelet  $\psi$  that generates an orthonormal basis. This attempt was a failure since he ended up constructing a whole family of orthonormal wavelet bases, with functions  $\psi$  that are infinitely continuously differentiable [375]. This was the fundamental impulse that led to a widespread search for new orthonormal wavelet bases, which culminated in the celebrated Daubechies wavelets of compact support [194].

The systematic theory for constructing orthonormal wavelet bases was established by Meyer and Mallat through the elaboration of multiresolution signal approximations [362], as presented in Chapter 7. It was inspired by original ideas developed in computer vision by Burt and Adelson [126] to analyze images at several resolutions. Digging deeper into the properties of orthogonal wavelets and multiresolution approximations brought to light a surprising link with filter banks constructed with conjugate mirror filters, and a fast wavelet transform algorithm decomposing signals of size  $N$  with  $O(N)$  operations [361].

### **Filter Banks**

Motivated by speech compression, in 1976 Croisier, Esteban, and Galand [189] introduced an invertible filter bank, which decomposes a discrete signal  $f[n]$  into two signals of half its size using a filtering and subsampling procedure. They showed that  $f[n]$  can be recovered from these subsampled signals by canceling the aliasing terms with a particular class of filters called *conjugate mirror filters*. This breakthrough led to a 10-year research effort to build a complete filter bank theory. Necessary and sufficient conditions for decomposing a signal in subsampled components with a filtering scheme, and recovering the same signal with an inverse transform, were established by Smith and Barnwell [444], Vaidyanathan [469], and Vetterli [471].

The multiresolution theory of Mallat [362] and Meyer [44] proves that any conjugate mirror filter characterizes a wavelet  $\psi$  that generates an orthonormal basis of  $L^2(\mathbb{R})$ , and that a fast discrete wavelet transform is implemented by cascading these conjugate mirror filters [361]. The equivalence between this continuous time wavelet theory and discrete filter banks led to a new fruitful interface between digital signal processing and harmonic analysis, first creating a culture shock that is now well resolved.

### **Continuous versus Discrete and Finite**

Originally, many signal processing engineers were wondering what is the point of considering wavelets and signals as functions, since all computations are performed over discrete signals with conjugate mirror filters. Why bother with the convergence of infinite convolution cascades if in practice we only compute a finite number of convolutions? Answering these important questions is necessary in order to understand why this book alternates between theorems on continuous time functions and discrete algorithms applied to finite sequences.

A short answer would be “simplicity.” In  $L^2(\mathbb{R})$ , a wavelet basis is constructed by dilating and translating a single function  $\psi$ . Several important theorems relate the amplitude of wavelet coefficients to the local regularity of the signal  $f$ . Dilations are not defined over discrete sequences, and discrete wavelet bases are therefore more complex to describe. The regularity of a discrete sequence is not well defined either, which makes it more difficult to interpret the amplitude of wavelet coefficients. A theory of continuous-time functions gives asymptotic results for discrete

sequences with sampling intervals decreasing to zero. This theory is useful because these asymptotic results are precise enough to understand the behavior of discrete algorithms.

But continuous time or space models are not sufficient for elaborating discrete signal-processing algorithms. The transition between continuous and discrete signals must be done with great care to maintain important properties such as orthogonality. Restricting the constructions to finite discrete signals adds another layer of complexity because of border problems. How these border issues affect numerical implementations is carefully addressed once the properties of the bases are thoroughly understood.

### **Wavelets for Images**

Wavelet orthonormal bases of images can be constructed from wavelet orthonormal bases of one-dimensional signals. Three mother wavelets  $\psi^1(x)$ ,  $\psi^2(x)$ , and  $\psi^3(x)$ , with  $x = (x_1, x_2) \in \mathbb{R}^2$ , are dilated by  $2^j$  and translated by  $2^j n$  with  $n = (n_1, n_2) \in \mathbb{Z}^2$ . This yields an orthonormal basis of the space  $L^2(\mathbb{R}^2)$  of finite energy functions  $f(x) = f(x_1, x_2)$ :

$$\left\{ \psi_{j,n}^k(x) = \frac{1}{2^j} \psi^k \left( \frac{x - 2^j n}{2^j} \right) \right\}_{j \in \mathbb{Z}, n \in \mathbb{Z}^2, 1 \leq k \leq 3}$$

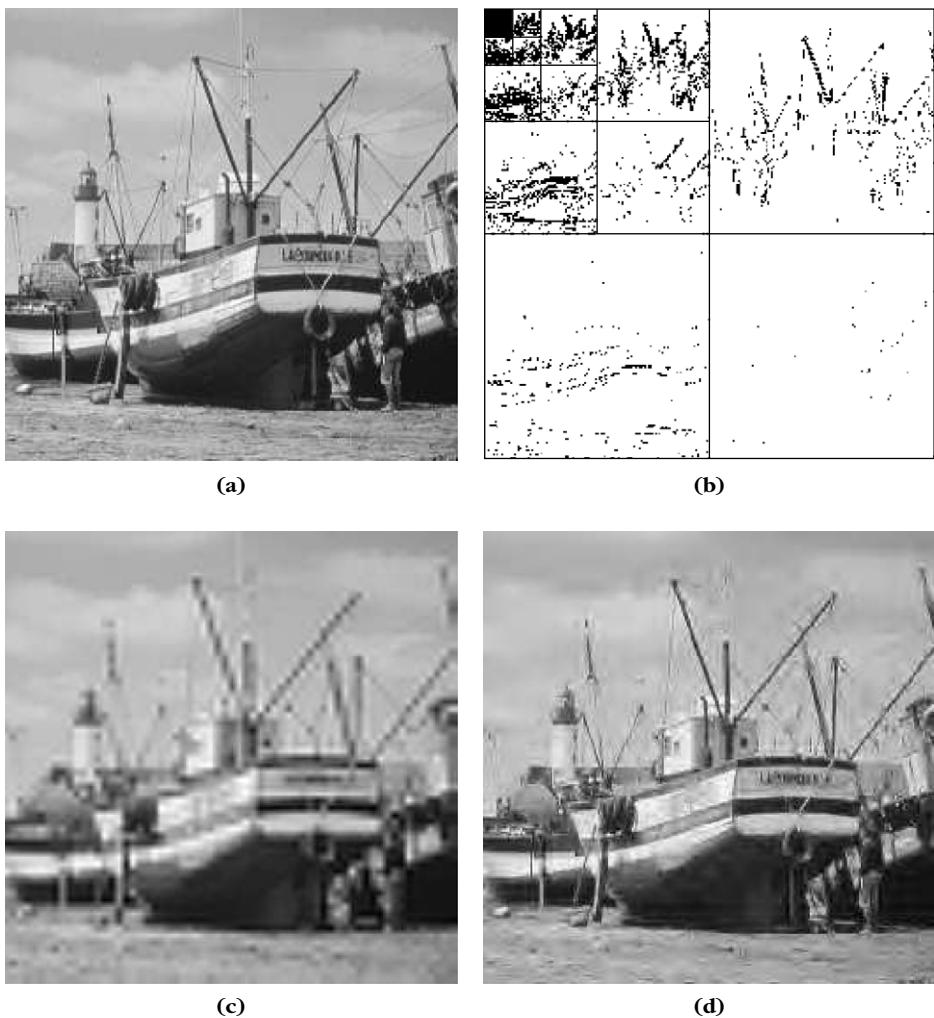
The support of a wavelet  $\psi_{j,n}^k$  is a square of width proportional to the scale  $2^j$ . Two-dimensional wavelet bases are discretized to define orthonormal bases of images including  $N$  pixels. Wavelet coefficients are calculated with the fast  $O(N)$  algorithm described in Chapter 7.

Like in one dimension, a wavelet coefficient  $\langle f, \psi_{j,n}^k \rangle$  has a small amplitude if  $f(x)$  is regular over the support of  $\psi_{j,n}^k$ . It has a large amplitude near sharp transitions such as edges. Figure 1.1(b) is the array of  $N$  wavelet coefficients. Each direction  $k$  and scale  $2^j$  corresponds to a subimage, which shows in black the position of the largest coefficients above a threshold:  $|\langle f, \psi_{j,n}^k \rangle| \geq T$ .

---

## **1.2 APPROXIMATION AND PROCESSING IN BASES**

Analog-to-digital signal conversion is the first step of digital signal processing. Chapter 3 explains that it amounts to projecting the signal over a basis of an approximation space. Most often, the resulting digital representation remains much too large and needs to be further reduced. A digital image typically includes more than  $10^6$  samples and a CD music recording has  $40 \times 10^3$  samples per second. Sparse representations that reduce the number of parameters can be obtained by thresholding coefficients in an appropriate orthogonal basis. Efficient compression and noise-reduction algorithms are then implemented with simple operators in this basis.

**FIGURE 1.1**

**(a)** Discrete image  $f[n]$  of  $N = 256^2$  pixels. **(b)** Array of  $N$  orthogonal wavelet coefficients  $\langle f, \psi_{j,n}^k \rangle$  for  $k = 1, 2, 3$ , and  $4$  scales  $2^j$ ; black points correspond to  $|\langle f, \psi_{j,n}^k \rangle| > T$ . **(c)** Linear approximation from the  $N/16$  wavelet coefficients at the three largest scales. **(d)** Nonlinear approximation from the  $M = N/16$  wavelet coefficients of largest amplitude shown in (b).

### Stochastic versus Deterministic Signal Models

A representation is optimized relative to a signal class, corresponding to all potential signals encountered in an application. This requires building signal models that carry available prior information.

A signal  $f$  can be modeled as a realization of a random process  $F$ , the probability distribution of which is known a priori. A Bayesian approach then tries to minimize

the expected approximation error. Linear approximations are simpler because they only depend on the covariance. Chapter 9 shows that optimal linear approximations are obtained on the basis of principal components that are the eigenvectors of the covariance matrix. However, the expected error of nonlinear approximations depends on the full probability distribution of  $F$ . This distribution is most often not known for complex signals, such as images or sounds, because their transient structures are not adequately modeled as realizations of known processes such as Gaussian ones.

To optimize nonlinear representations, weaker but sufficiently powerful deterministic models can be elaborated. A deterministic model specifies a set  $\Theta$ , where the signal belongs. This set is defined by any prior information—for example, on the time-frequency localization of transients in musical recordings or on the geometric regularity of edges in images. Simple models can also define  $\Theta$  as a ball in a functional space, with a specific regularity norm such as a total variation norm. A stochastic model is richer because it provides the probability distribution in  $\Theta$ . When this distribution is not available, the average error cannot be calculated and is replaced by the maximum error over  $\Theta$ . Optimizing the representation then amounts to minimizing this maximum error, which is called a *minimax* optimization.

### 1.2.1 Sampling with Linear Approximations

Analog-to-digital signal conversion is most often implemented with a linear approximation operator that filters and samples the input analog signal. From these samples, a linear digital-to-analog converter recovers a projection of the original analog signal over an approximation space whose dimension depends on the sampling density. Linear approximations project signals in spaces of lowest possible dimensions to reduce computations and storage cost, while controlling the resulting error.

#### **Sampling Theorems**

Let us consider finite energy signals  $\|\bar{f}\|^2 = \int |\bar{f}(x)|^2 dx$  of finite support, which is normalized to  $[0, 1]$  or  $[0, 1]^2$  for images. A sampling process implements a filtering of  $\bar{f}(x)$  with a low-pass impulse response  $\bar{\phi}_s(x)$  and a uniform sampling to output a discrete signal:

$$f[n] = \bar{f} \star \bar{\phi}_s(ns) \quad \text{for } 0 \leq n < N.$$

In two dimensions,  $n = (n_1, n_2)$  and  $x = (x_1, x_2)$ . These filtered samples can also be written as inner products:

$$\bar{f} \star \bar{\phi}_s(ns) = \int f(u) \bar{\phi}_s(ns - u) du = \langle f(x), \phi_s(x - ns) \rangle$$

with  $\phi_s(x) = \bar{\phi}_s(-x)$ . Chapter 3 explains that  $\phi_s$  is chosen, like in the classic Shannon–Whittaker sampling theorem, so that a family of functions  $\{\phi_s(x - ns)\}_{1 \leq n \leq N}$  is a basis of an appropriate approximation space  $\mathbf{U}_N$ . The best linear approximation of  $\bar{f}$  in  $\mathbf{U}_N$  recovered from these samples is the orthogonal

projection  $\bar{f}_N$  of  $f$  in  $\mathbf{U}_N$ , and if the basis is orthonormal, then

$$\bar{f}_N(x) = \sum_{n=0}^{N-1} f[n] \phi_s(x - ns). \quad (1.4)$$

A sampling theorem states that if  $\bar{f} \in \mathbf{U}_N$  then  $\bar{f} = \bar{f}_N$  so (1.4) recovers  $\bar{f}(x)$  from the measured samples. Most often,  $\bar{f}$  does not belong to this approximation space. It is called *aliasing* in the context of Shannon–Whittaker sampling, where  $\mathbf{U}_N$  is the space of functions having a frequency support restricted to the  $N$  lower frequencies. The approximation error  $\|\bar{f} - \bar{f}_N\|^2$  must then be controlled.

### Linear Approximation Error

The approximation error is computed by finding an orthogonal basis  $\mathcal{B} = \{\bar{g}_m(x)\}_{0 \leq m < +\infty}$  of the whole analog signal space  $L^2[0, 1]^2$ , with the first  $N$  vector  $\{\bar{g}_m(x)\}_{0 \leq m \leq N}$  that defines an orthogonal basis of  $\mathbf{U}_N$ . Thus, the orthogonal projection on  $\mathbf{U}_N$  can be rewritten as

$$\bar{f}_N(x) = \sum_{m=0}^{N-1} \langle \bar{f}, \bar{g}_m \rangle \bar{g}_m(x).$$

Since  $\bar{f} = \sum_{m=0}^{+\infty} \langle \bar{f}, \bar{g}_m \rangle \bar{g}_m$ , the approximation error is the energy of the removed inner products:

$$\varepsilon_l(N, f) = \|\bar{f} - \bar{f}_N\|^2 = \sum_{m=N}^{+\infty} |\langle \bar{f}, \bar{g}_m \rangle|^2.$$

This error decreases quickly when  $N$  increases if the coefficient amplitudes  $|\langle \bar{f}, \bar{g}_m \rangle|$  have a fast decay when the index  $m$  increases. The dimension  $N$  is adjusted to the desired approximation error.

Figure 1.1(a) shows a discrete image  $f[n]$  approximated with  $N = 256^2$  pixels. Figure 1.1(c) displays a lower-resolution image  $f_{N/16}$  projected on a space  $\mathbf{U}_{N/16}$  of dimension  $N/16$ , generated by  $N/16$  large-scale wavelets. It is calculated by setting all the wavelet coefficients to zero at the first two smaller scales. The approximation error is  $\|f - f_{N/16}\|^2 / \|f\|^2 = 14 \times 10^{-3}$ . Reducing the resolution introduces more blur and errors. A linear approximation space  $\mathbf{U}_N$  corresponds to a uniform grid that approximates precisely uniform regular signals. Since images  $\bar{f}$  are often not uniformly regular, it is necessary to measure it at a high-resolution  $N$ . This is why digital cameras have a resolution that increases as technology improves.

### 1.2.2 Sparse Nonlinear Approximations

Linear approximations reduce the space dimensionality but can introduce important errors when reducing the resolution if the signal is not uniformly regular, as shown by Figure 1.1(c). To improve such approximations, more coefficients should be kept where needed—not in regular regions but near sharp transitions and edges.

This requires defining an irregular sampling adapted to the local signal regularity. This optimized irregular sampling has a simple equivalent solution through nonlinear approximations in wavelet bases.

Nonlinear approximations operate in two stages. First, a linear operator approximates the analog signal  $\bar{f}$  with  $N$  samples written  $f[n] = \bar{f} \star \phi_s(ns)$ . Then, a nonlinear approximation of  $f[n]$  is computed to reduce the  $N$  coefficients  $f[n]$  to  $M \ll N$  coefficients in a sparse representation.

The discrete signal  $f$  can be considered as a vector of  $\mathbb{C}^N$ . Inner products and norms in  $\mathbb{C}^N$  are written

$$\langle f, g \rangle = \sum_{n=0}^{N-1} f[n] g^*[n] \quad \text{and} \quad \|f\|^2 = \sum_{n=0}^{N-1} |f[n]|^2.$$

To obtain a sparse representation with a nonlinear approximation, we choose a new orthonormal basis  $\mathcal{B} = \{g_m[n]\}_{m \in \Gamma}$  of  $\mathbb{C}^N$ , which concentrates the signal energy as much as possible over few coefficients. Signal coefficients  $\{\langle f, g_m \rangle\}_{m \in \Gamma}$  are computed from the  $N$  input sample values  $f[n]$  with an orthogonal change of basis that takes  $N^2$  operations in nonstructured bases. In a wavelet or Fourier bases, fast algorithms require, respectively,  $O(N)$  and  $O(N \log_2 N)$  operations.

### Approximation by Thresholding

For  $M < N$ , an approximation  $f_M$  is computed by selecting the “best”  $M < N$  vectors within  $\mathcal{B}$ . The orthogonal projection of  $f$  on the space  $V_\Lambda$  generated by  $M$  vectors  $\{g_m\}_{m \in \Lambda}$  in  $\mathcal{B}$  is

$$f_\Lambda = \sum_{m \in \Lambda} \langle f, g_m \rangle g_m. \quad (1.5)$$

Since  $f = \sum_{m \in \Gamma} \langle f, g_m \rangle g_m$ , the resulting error is

$$\|f - f_\Lambda\|^2 = \sum_{m \notin \Lambda} |\langle f, g_m \rangle|^2. \quad (1.6)$$

We write  $|\Lambda|$  the size of the set  $\Lambda$ . The best  $M = |\Lambda|$  term approximation, which minimizes  $\|f - f_\Lambda\|^2$ , is thus obtained by selecting the  $M$  coefficients of largest amplitude. These coefficients are above a threshold  $T$  that depends on  $M$ :

$$f_M = f_{\Lambda_T} = \sum_{m \in \Lambda_T} \langle f, g_m \rangle g_m \quad \text{with} \quad \Lambda_T = \{m \in \Gamma : |\langle f, g_m \rangle| \geq T\}. \quad (1.7)$$

This approximation is nonlinear because the approximation set  $\Lambda_T$  changes with  $f$ . The resulting approximation error is:

$$\varepsilon_n(M, f) = \|f - f_M\|^2 = \sum_{m \notin \Lambda_T} |\langle f, g_m \rangle|^2. \quad (1.8)$$

Figure 1.1(b) shows that the approximation support  $\Lambda_T$  of an image in a wavelet orthonormal basis depends on the geometry of edges and textures. Keeping large

wavelet coefficients is equivalent to constructing an adaptive approximation grid specified by the scale-space support  $\Lambda_T$ . It increases the approximation resolution where the signal is irregular. The geometry of  $\Lambda_T$  gives the spatial distribution of sharp image transitions and edges, and their propagation across scales. Chapter 6 proves that wavelet coefficients give important information about singularities and local Lipschitz regularity. This example illustrates how approximation support provides “geometric” information on  $f$ , relative to a dictionary, that is a wavelet basis in this example.

Figure 1.1(d) gives the nonlinear wavelet approximation  $f_M$  recovered from the  $M = N/16$  large-amplitude wavelet coefficients, with an error  $\|f - f_M\|^2 / \|f\|^2 = 5 \times 10^{-3}$ . This error is nearly three times smaller than the linear approximation error obtained with the same number of wavelet coefficients, and the image quality is much better.

An analog signal can be recovered from the discrete nonlinear approximation  $f_M$ :

$$\bar{f}_M(x) = \sum_{n=0}^{N-1} f_M[n] \phi_s(x - ns).$$

Since all projections are orthogonal, the overall approximation error on the original analog signal  $\bar{f}(x)$  is the sum of the analog sampling error and the discrete nonlinear error:

$$\|\bar{f} - \bar{f}_M\|^2 = \|\bar{f} - \bar{f}_N\|^2 + \|f - f_M\|^2 = \varepsilon_l(N, f) + \varepsilon_n(M, f).$$

In practice,  $N$  is imposed by the resolution of the signal-acquisition hardware, and  $M$  is typically adjusted so that  $\varepsilon_n(M, f) \geq \varepsilon_l(N, f)$ .

### **Sparsity with Regularity**

Sparse representations are obtained in a basis that takes advantage of some form of regularity of the input signals, creating many small-amplitude coefficients. Since wavelets have localized support, functions with isolated singularities produce few large-amplitude wavelet coefficients in the neighborhood of these singularities. Nonlinear wavelet approximation produces a small error over spaces of functions that do not have “too many” sharp transitions and singularities. Chapter 9 shows that functions having a bounded total variation norm are useful models for images with nonfractal (finite length) edges.

Edges often define regular geometric curves. Wavelets detect the location of edges but their square support cannot take advantage of their potential geometric regularity. More sparse representations are defined in dictionaries of curvelets or bandlets, which have elongated support in multiple directions, that can be adapted to this geometrical regularity. In such dictionaries, the approximation support  $\Lambda_T$  is smaller but provides explicit information about edges’ local geometrical properties such as their orientation. In this context, geometry does not just apply to multidimensional signals. Audio signals, such as musical recordings, also have a complex geometric regularity in time-frequency dictionaries.

### 1.2.3 Compression

Storage limitations and fast transmission through narrow bandwidth channels require compression of signals while minimizing degradation. Transform codes compress signals by coding a sparse representation. Chapter 10 introduces the information theory needed to understand these codes and to optimize their performance.

In a compression framework, the analog signal has already been discretized into a signal  $f[n]$  of size  $N$ . This discrete signal is decomposed in an orthonormal basis  $\mathcal{B} = \{g_m\}_{m \in \Gamma}$  of  $\mathbb{C}^N$ :

$$f = \sum_{m \in \Gamma} \langle f, g_m \rangle g_m.$$

Coefficients  $\langle f, g_m \rangle$  are approximated by quantized values  $Q(\langle f, g_m \rangle)$ . If  $Q$  is a uniform quantizer of step  $\Delta$ , then  $|x - Q(x)| \leq \Delta/2$ ; and if  $|x| < \Delta/2$ , then  $Q(x) = 0$ . The signal  $\tilde{f}$  restored from quantized coefficients is

$$\tilde{f} = \sum_{m \in \Gamma} Q(\langle f, g_m \rangle) g_m.$$

An entropy code records these coefficients with  $R$  bits. The goal is to minimize the signal-distortion rate  $d(R, f) = \|\tilde{f} - f\|^2$ .

The coefficients not quantized to zero correspond to the set  $\Lambda_T = \{m \in \Gamma : |\langle f, g_m \rangle| \geq T\}$  with  $T = \Delta/2$ . For sparse signals, Chapter 10 shows that the bit budget  $R$  is dominated by the number of bits to code  $\Lambda_T$  in  $\Gamma$ , which is nearly proportional to its size  $|\Lambda_T|$ . This means that the “information” about a sparse representation is mostly geometric. Moreover, the distortion is dominated by the nonlinear approximation error  $\|f - f_{\Lambda_T}\|^2$ , for  $f_{\Lambda_T} = \sum_{m \in \Lambda_T} \langle f, g_m \rangle g_m$ . Compression is thus a sparse approximation problem. For a given distortion  $d(R, f)$ , minimizing  $R$  requires reducing  $|\Lambda_T|$  and thus optimizing the sparsity.

The number of bits to code  $\Lambda_T$  can take advantage of any prior information on the geometry. Figure 1.1(b) shows that large wavelet coefficients are not randomly distributed. They have a tendency to be aggregated toward larger scales, and at fine scales they are regrouped along edge curves or in texture regions. Using such prior geometric models is a source of gain in coders such as JPEG-2000.

Chapter 10 describes the implementation of audio transform codes. Image transform codes in block cosine bases and wavelet bases are introduced, together with the JPEG and JPEG-2000 compression standards.

### 1.2.4 Denoising

Signal-acquisition devices add noise that can be reduced by estimators using prior information on signal properties. Signal processing has long remained mostly Bayesian and linear. Nonlinear smoothing algorithms existed in statistics, but these procedures were often ad hoc and complex. Two statisticians, Donoho and Johnstone [221], changed the “game” by proving that simple thresholding in sparse

representations can yield nearly optimal nonlinear estimators. This was the beginning of a considerable refinement of nonlinear estimation algorithms that is still ongoing.

Let us consider digital measurements that add a random noise  $W[n]$  to the original signal  $f[n]$ :

$$X[n] = f[n] + W[n] \quad \text{for } 0 \leq n < N.$$

The signal  $f$  is estimated by transforming the noisy data  $X$  with an operator  $D$ :

$$\tilde{F} = DX.$$

The risk of the estimator  $\tilde{F}$  of  $f$  is the average error, calculated with respect to the probability distribution of noise  $W$ :

$$r(D, f) = E\{\|f - DX\|^2\}.$$

### **Bayes versus Minimax**

To optimize the estimation operator  $D$ , one must take advantage of prior information available about signal  $f$ . In a Bayes framework,  $f$  is considered a realization of a random vector  $F$  and the Bayes risk is the expected risk calculated with respect to the prior probability distribution  $\pi$  of the random signal model  $F$ :

$$r(D, \pi) = E_\pi\{r(D, F)\}.$$

Optimizing  $D$  among all possible operators yields the *minimum Bayes risk*:

$$r_n(\pi) = \inf_{all D} r(D, \pi).$$

In the 1940s, Wald brought in a new perspective on statistics with a decision theory partly imported from the theory of games. This point of view uses deterministic models, where signals are elements of a set  $\Theta$ , without specifying their probability distribution in this set. To control the risk for any  $f \in \Theta$ , we compute the maximum risk:

$$r(D, \Theta) = \sup_{f \in \Theta} r(D, f).$$

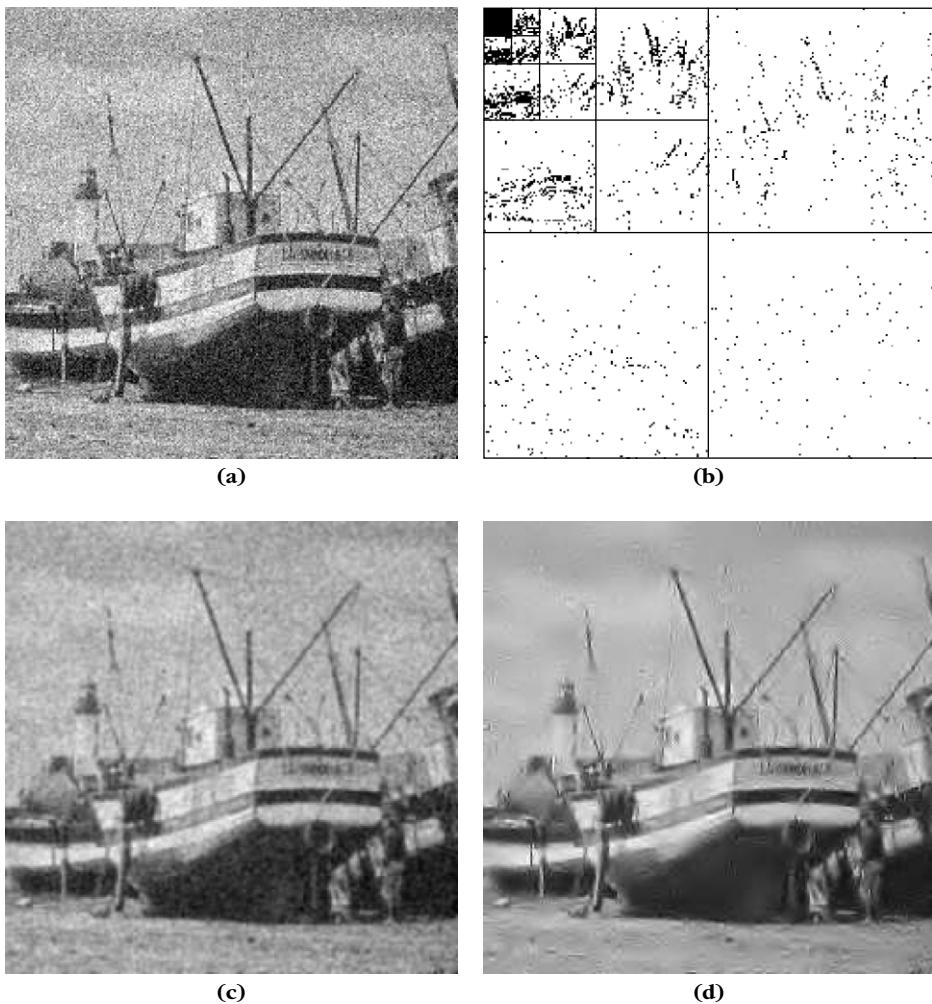
The *minimax risk* is the lower bound computed over all operators  $D$ :

$$r_n(\Theta) = \inf_{all D} r(D, \Theta).$$

In practice, the goal is to find an operator  $D$  that is simple to implement and yields a risk close to the minimax lower bound.

### **Thresholding Estimators**

It is tempting to restrict calculations to linear operators  $D$  because of their simplicity. Optimal linear Wiener estimators are introduced in Chapter 11. Figure 1.2(a) is an image contaminated by Gaussian white noise. Figure 1.2(b) shows an optimized

**FIGURE 1.2**

**(a)** Noisy image  $X$ . **(b)** Noisy wavelet coefficients above threshold,  $|\langle X, \psi_{j,n} \rangle| \geq T$ . **(c)** Linear estimation  $X * h$ . **(d)** Nonlinear estimator recovered from thresholded wavelet coefficients over several translated bases.

linear filtering estimation  $\tilde{F} = X * h[n]$ , which is therefore diagonal in a Fourier basis  $\mathcal{B}$ . This convolution operator averages the noise but also blurs the image and keeps low-frequency noise by retaining the image's low frequencies.

If  $f$  has a sparse representation in a dictionary, then projecting  $X$  on the vectors of this sparse support can considerably improve linear estimators. The difficulty is identifying the sparse support of  $f$  from the noisy data  $X$ . Donoho and

Johnstone [221] proved that, in an orthonormal basis, a simple thresholding of noisy coefficients does the trick. Noisy signal coefficients in an orthonormal basis  $\mathcal{B} = \{g_m\}_{m \in \Gamma}$  are

$$\langle X, g_m \rangle = \langle f, g_m \rangle + \langle W, g_m \rangle \quad \text{for } m \in \Gamma.$$

Thresholding these noisy coefficients yields an orthogonal projection estimator

$$\tilde{F} = X_{\tilde{\Lambda}_T} = \sum_{m \in \tilde{\Lambda}_T} \langle X, g_m \rangle g_m \quad \text{with } \tilde{\Lambda}_T = \{m \in \Gamma : |\langle X, g_m \rangle| \geq T\}. \quad (1.9)$$

The set  $\tilde{\Lambda}_T$  is an estimate of an approximation support of  $f$ . It is hopefully close to the optimal approximation support  $\Lambda_T = \{m \in \Gamma : |f, g_m| \geq T\}$ .

Figure 1.2(b) shows the estimated approximation set  $\tilde{\Lambda}_T$  of noisy-wavelet coefficients,  $|\langle X, \psi_{j,n} \rangle| \geq T$ , that can be compared to the optimal approximation support  $\Lambda_T$  shown in Figure 1.1(b). The estimation in Figure 1.2(d) from wavelet coefficients in  $\tilde{\Lambda}_T$  has considerably reduced the noise in regular regions while keeping the sharpness of edges by preserving large-wavelet coefficients. This estimation is improved with a translation-invariant procedure that averages this estimator over several translated wavelet bases. Thresholding wavelet coefficients implements an adaptive smoothing, which averages the data  $X$  with a kernel that depends on the estimated regularity of the original signal  $f$ .

Donoho and Johnstone proved that for Gaussian white noise of variance  $\sigma^2$ , choosing  $T = \sigma \sqrt{2 \log_e N}$  yields a risk  $E\{\|f - \tilde{F}\|^2\}$  of the order of  $\|f - f_{\Lambda_T}\|^2$ , up to a  $\log_e N$  factor. This spectacular result shows that the estimated support  $\tilde{\Lambda}_T$  does nearly as well as the optimal unknown support  $\Lambda_T$ . The resulting risk is small if the representation is sparse and precise.

The set  $\tilde{\Lambda}_T$  in Figure 1.2(b) “looks” different from the  $\Lambda_T$  in Figure 1.1(b) because it has more isolated points. This indicates that some prior information on the geometry of  $\Lambda_T$  could be used to improve the estimation. For audio noise-reduction, thresholding estimators are applied in sparse representations provided by time-frequency bases. Similar isolated time-frequency coefficients produce a highly annoying “musical noise.” Musical noise is removed with a block thresholding that regularizes the geometry of the estimated support  $\tilde{\Lambda}_T$  and avoids leaving isolated points. Block thresholding also improves wavelet estimators.

If  $W$  is a Gaussian noise and signals in  $\Theta$  have a sparse representation in  $\mathcal{B}$ , then Chapter 11 proves that thresholding estimators can produce a nearly minimax risk. In particular, wavelet thresholding estimators have a nearly minimax risk for large classes of piecewise smooth signals, including bounded variation images.

### 1.3 TIME-FREQUENCY DICTIONARIES

Motivated by quantum mechanics, in 1946 the physicist Gabor [267] proposed decomposing signals over dictionaries of elementary waveforms which he called

time-frequency atoms that have a minimal spread in a time-frequency plane. By showing that such decompositions are closely related to our perception of sounds, and that they exhibit important structures in speech and music recordings, Gabor demonstrated the importance of localized time-frequency signal processing. Beyond sounds, large classes of signals have sparse decompositions as sums of time-frequency atoms selected from appropriate dictionaries. The key issue is to understand how to construct dictionaries with time-frequency atoms adapted to signal properties.

### 1.3.1 Heisenberg Uncertainty

A time-frequency dictionary  $\mathcal{D} = \{\phi_\gamma\}_{\gamma \in \Gamma}$  is composed of waveforms of unit norm  $\|\phi_\gamma\| = 1$ , which have a narrow localization in time and frequency. The time localization  $u$  of  $\phi_\gamma$  and its spread around  $u$ , are defined by

$$u = \int t |\phi_\gamma(t)|^2 dt \quad \text{and} \quad \sigma_{t,\gamma}^2 = \int |t - u|^2 |\phi_\gamma(t)|^2 dt.$$

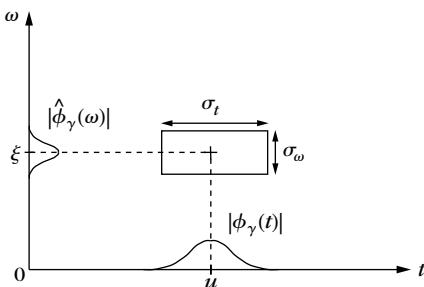
Similarly, the frequency localization and spread of  $\hat{\phi}_\gamma$  are defined by

$$\xi = (2\pi)^{-1} \int \omega |\hat{\phi}_\gamma(\omega)|^2 d\omega \quad \text{and} \quad \sigma_{\omega,\gamma}^2 = (2\pi)^{-1} \int |\omega - \xi|^2 |\hat{\phi}_\gamma(\omega)|^2 d\omega.$$

The Fourier Parseval formula

$$\langle f, \phi_\gamma \rangle = \int_{-\infty}^{+\infty} f(t) \phi_\gamma^*(t) dt = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \hat{f}(\omega) \hat{\phi}_\gamma^*(\omega) d\omega \quad (1.10)$$

shows that  $\langle f, \phi_\gamma \rangle$  depends mostly on the values  $f(t)$  and  $\hat{f}(\omega)$ , where  $\phi_\gamma(t)$  and  $\hat{\phi}_\gamma(\omega)$  are nonnegligible, and hence for  $(t, \omega)$  in a rectangle centered at  $(u, \xi)$ , of size  $\sigma_{t,\gamma} \times \sigma_{\omega,\gamma}$ . This rectangle is illustrated by Figure 1.3 in this time-frequency plane  $(t, \omega)$ . It can be interpreted as a “quantum of information” over an elementary



**FIGURE 1.3**

Heisenberg box representing an atom  $\phi_\gamma$ .

resolution cell. The uncertainty principle theorem proves (see Chapter 2) that this rectangle has a minimum surface that limits the joint time-frequency resolution:

$$\sigma_{t,\gamma} \sigma_{\omega,\gamma} \geq \frac{1}{2}. \quad (1.11)$$

Constructing a dictionary of time-frequency atoms can thus be thought of as covering the time-frequency plane with resolution cells having a time width  $\sigma_{t,\gamma}$  and a frequency width  $\sigma_{\omega,\gamma}$  which may vary but with a surface larger than one-half. Windowed Fourier and wavelet transforms are two important examples.

### 1.3.2 Windowed Fourier Transform

A windowed Fourier dictionary is constructed by translating in time and frequency a time window  $g(t)$ , of unit norm  $\|g\| = 1$ , centered at  $t = 0$ :

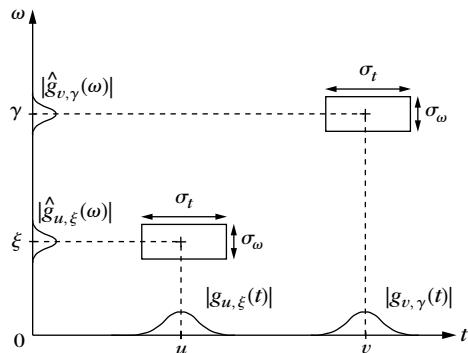
$$\mathcal{D} = \left\{ g_{u,\xi}(t) = g(t-u) e^{i\xi t} \right\}_{(u,\xi) \in \mathbb{R}^2}.$$

The atom  $g_{u,\xi}$  is translated by  $u$  in time and by  $\xi$  in frequency. The time-and-frequency spread of  $g_{u,\xi}$  is independent of  $u$  and  $\xi$ . This means that each atom  $g_{u,\xi}$  corresponds to a Heisenberg rectangle that has a size  $\sigma_t \times \sigma_\omega$  independent of its position  $(u, \xi)$ , as shown by Figure 1.4.

The windowed Fourier transform projects  $f$  on each dictionary atom  $g_{u,\xi}$ :

$$Sf(u, \xi) = \langle f, g_{u,\xi} \rangle = \int_{-\infty}^{+\infty} f(t) g(t-u) e^{-i\xi t} dt. \quad (1.12)$$

It can be interpreted as a Fourier transform of  $f$  at the frequency  $\xi$ , localized by the window  $g(t-u)$  in the neighborhood of  $u$ . This windowed Fourier transform is highly redundant and represents one-dimensional signals by a time-frequency image in  $(u, \xi)$ . It is thus necessary to understand how to select many fewer time-frequency coefficients that represent the signal efficiently.



**FIGURE 1.4**

Time-frequency boxes (“Heisenberg rectangles”) representing the energy spread of two windowed Fourier atoms.

When listening to music, we perceive sounds that have a frequency that varies in time. Chapter 4 shows that a spectral line of  $f$  creates high-amplitude windowed Fourier coefficients  $Sf(u, \xi)$  at frequencies  $\xi(u)$  that depend on time  $u$ . These spectral components are detected and characterized by ridge points, which are local maxima in this time-frequency plane. Ridge points define a time-frequency approximation support  $\Lambda$  of  $f$  with a geometry that depends on the time-frequency evolution of the signal spectral components. Modifying the sound duration or audio transpositions are implemented by modifying the geometry of the ridge support in time frequency.

A windowed Fourier transform decomposes signals over waveforms that have the same time and frequency resolution. It is thus effective as long as the signal does not include structures having different time-frequency resolutions, some being very localized in time and others very localized in frequency. Wavelets address this issue by changing the time and frequency resolution.

### 1.3.3 Continuous Wavelet Transform

In reflection seismology, Morlet knew that the waveforms sent underground have a duration that is too long at high frequencies to separate the returns of fine, closely spaced geophysical layers. Such waveforms are called *wavelets* in geophysics. Instead of emitting pulses of equal duration, he thought of sending shorter waveforms at high frequencies. These waveforms were obtained by scaling the mother wavelet, hence the name of this transform. Although Grossmann was working in theoretical physics, he recognized in Morlet's approach some ideas that were close to his own work on coherent quantum states.

Nearly forty years after Gabor, Morlet and Grossmann reactivated a fundamental collaboration between theoretical physics and signal processing, which led to the formalization of the continuous wavelet transform [288]. These ideas were not totally new to mathematicians working in harmonic analysis, or to computer vision researchers studying multiscale image processing. It was thus only the beginning of a rapid catalysis that brought together scientists with very different backgrounds.

A wavelet dictionary is constructed from a mother wavelet  $\psi$  of zero average

$$\int_{-\infty}^{+\infty} \psi(t) dt = 0,$$

which is dilated with a scale parameter  $s$ , and translated by  $u$ :

$$\mathcal{D} = \left\{ \psi_{u,s}(t) = \frac{1}{\sqrt{s}} \psi\left(\frac{t-u}{s}\right) \right\}_{u \in \mathbb{R}, s > 0}. \quad (1.13)$$

The continuous wavelet transform of  $f$  at any scale  $s$  and position  $u$  is the projection of  $f$  on the corresponding wavelet atom:

$$Wf(u, s) = \langle f, \psi_{u,s} \rangle = \int_{-\infty}^{+\infty} f(t) \frac{1}{\sqrt{s}} \psi^*\left(\frac{t-u}{s}\right) dt. \quad (1.14)$$

It represents one-dimensional signals by highly redundant time-scale images in  $(u, s)$ .

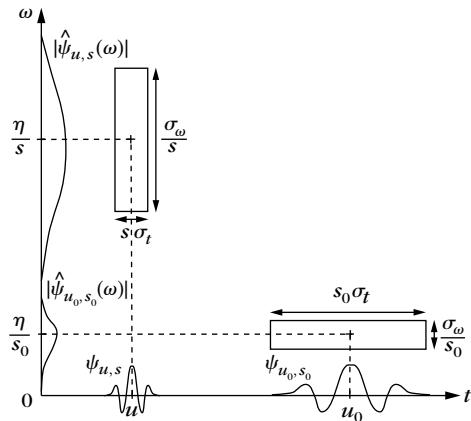
### Varying Time-Frequency Resolution

As opposed to windowed Fourier atoms, wavelets have a time-frequency resolution that changes. The wavelet  $\psi_{u,s}$  has a time support centered at  $u$  and proportional to  $s$ . Let us choose a wavelet  $\psi$  whose Fourier transform  $\hat{\psi}(\omega)$  is nonzero in a positive frequency interval centered at  $\eta$ . The Fourier transform  $\hat{\psi}_{u,s}(\omega)$  is dilated by  $1/s$  and thus is localized in a positive frequency interval centered at  $\xi = \eta/s$ ; its size is scaled by  $1/s$ . In the time-frequency plane, the Heisenberg box of a wavelet atom  $\psi_{u,s}$  is therefore a rectangle centered at  $(u, \eta/s)$ , with time and frequency widths, respectively, proportional to  $s$  and  $1/s$ . When  $s$  varies, the time and frequency width of this time-frequency resolution cell changes, but its area remains constant, as illustrated by Figure 1.5.

Large-amplitude wavelet coefficients can detect and measure short high-frequency variations because they have a narrow time localization at high frequencies. At low frequencies their time resolution is lower, but they have a better frequency resolution. This modification of time and frequency resolution is adapted to represent sounds with sharp attacks, or radar signals having a frequency that may vary quickly at high frequencies.

### Multiscale Zooming

A wavelet dictionary is also adapted to analyze the scaling evolution of transients with zooming procedures across scales. Suppose now that  $\psi$  is real. Since it has a zero average, a wavelet coefficient  $Wf(u, s)$  measures the variation of  $f$  in a neighborhood of  $u$  that has a size proportional to  $s$ . Sharp signal transitions create large-amplitude wavelet coefficients.



**FIGURE 1.5**

Heisenberg time-frequency boxes of two wavelets,  $\psi_{u,s}$  and  $\psi_{u_0,s_0}$ . When the scale  $s$  decreases, the time support is reduced but the frequency spread increases and covers an interval that is shifted toward high frequencies.

Signal singularities have specific scaling invariance characterized by Lipschitz exponents. Chapter 6 relates the pointwise regularity of  $f$  to the asymptotic decay of the wavelet transform amplitude  $|Wf(u, s)|$  when  $s$  goes to zero. Singularities are detected by following the local maxima of the wavelet transform across scales.

In images, wavelet local maxima indicate the position of edges, which are sharp variations of image intensity. It defines scale-space approximation support of  $f$  from which precise image approximations are reconstructed. At different scales, the geometry of this local maxima support provides contours of image structures of varying sizes. This multiscale edge detection is particularly effective for pattern recognition in computer vision [146].

The zooming capability of the wavelet transform not only locates isolated singular events, but can also characterize more complex multifractal signals having nonisolated singularities. Mandelbrot [41] was the first to recognize the existence of multifractals in most corners of nature. Scaling one part of a multifractal produces a signal that is statistically similar to the whole. This self-similarity appears in the continuous wavelet transform, which modifies the analyzing scale. From global measurements of the wavelet transform decay, Chapter 6 measures the singularity distribution of multifractals. This is particularly important in analyzing their properties and testing multifractal models in physics or in financial time series.

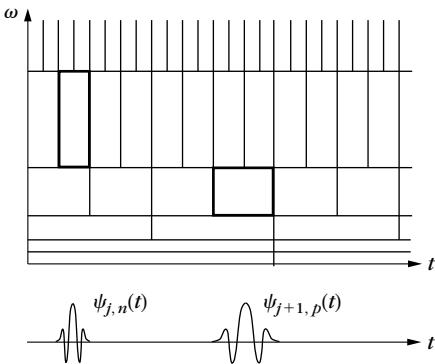
### 1.3.4 Time-Frequency Orthonormal Bases

Orthonormal bases of time-frequency atoms remove all redundancy and define stable representations. A wavelet orthonormal basis is an example of the time-frequency basis obtained by scaling a wavelet  $\psi$  with dyadic scales  $s = 2^j$  and translating it by  $2^j n$ , which is written  $\psi_{j,n}$ . In the time-frequency plane, the Heisenberg resolution box of  $\psi_{j,n}$  is a dilation by  $2^j$  and translation by  $2^j n$  of the Heisenberg box of  $\psi$ . A wavelet orthonormal is thus a subdictionary of the continuous wavelet transform dictionary, which yields a perfect tiling of the time-frequency plane illustrated in Figure 1.6.

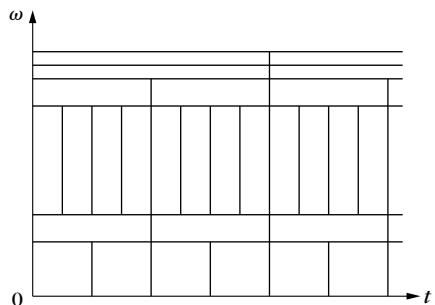
One can construct many other orthonormal bases of time-frequency atoms, corresponding to different tilings of the time-frequency plane. Wavelet packet and local cosine bases are two important examples constructed in Chapter 8, with time-frequency atoms that split the frequency and the time axis, respectively, in intervals of varying sizes.

#### **Wavelet Packet Bases**

Wavelet bases divide the frequency axis into intervals of 1 octave bandwidth. Coifman, Meyer, and Wickerhauser [182] have generalized this construction with bases that split the frequency axis in intervals of bandwidth that may be adjusted. Each frequency interval is covered by the Heisenberg time-frequency boxes of wavelet packet functions translated in time, in order to cover the whole plane, as shown by Figure 1.7.

**FIGURE 1.6**

The time-frequency boxes of a wavelet basis define a tiling of the time-frequency plane.

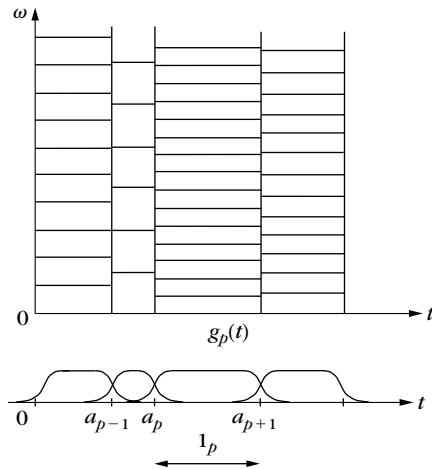
**FIGURE 1.7**

A wavelet packet basis divides the frequency axis in separate intervals of varying sizes. A tiling is obtained by translating in time the wavelet packets covering each frequency interval.

As for wavelets, wavelet-packet coefficients are obtained with a filter bank of conjugate mirror filters that split the frequency axis in several frequency intervals. Different frequency segmentations correspond to different wavelet packet bases. For images, a filter bank divides the image frequency support in squares of dyadic sizes that can be adjusted.

### **Local Cosine Bases**

Local cosine orthonormal bases are constructed by dividing the time axis instead of the frequency axis. The time axis is segmented in successive intervals  $[a_p, a_{p+1}]$ . The local cosine bases of Malvar [368] are obtained by designing smooth windows  $g_p(t)$  that cover each interval  $[a_p, a_{p+1}]$ , and by multiplying them by cosine functions  $\cos(\xi t + \phi)$  of different frequencies. This is yet another idea that has been independently studied in physics, signal processing, and mathematics. Malvar's original construction was for discrete signals. At the same time, the physicist Wilson [486] was designing a local cosine basis, with smooth windows of infinite support,

**FIGURE 1.8**

A local cosine basis divides the time axis with smooth windows  $g_p(t)$  and translates these windows into frequency.

to analyze the properties of quantum coherent states. Malvar bases were also rediscovered and generalized by the harmonic analysts Coifman and Meyer [181]. These different views of the same bases brought to light mathematical and algorithmic properties that opened new applications.

A multiplication by  $\cos(\xi t + \phi)$  translates the Fourier transform  $\hat{g}_p(\omega)$  of  $g_p(t)$  by  $\pm \xi$ . Over positive frequencies, the time-frequency box of the modulated window  $g_p(t) \cos(\xi t + \phi)$  is therefore equal to the time-frequency box of  $g_p$  translated by  $\xi$  along frequencies. Figure 1.8 shows the time-frequency tiling corresponding to such a local cosine basis. For images, a two-dimensional cosine basis is constructed by dividing the image support in squares of varying sizes.

## 1.4 SPARSITY IN REDUNDANT DICTIONARIES

In natural languages, large dictionaries are needed to refine ideas with short sentences, and they evolve with usage. Eskimos have eight different words to describe *snow quality*, whereas a single word is typically sufficient in a Parisian dictionary. Similarly, large signal dictionaries of vectors are needed to construct sparse representations of complex signals. However, computing and optimizing a signal approximation by choosing the best  $M$  dictionary vectors is much more difficult.

### 1.4.1 Frame Analysis and Synthesis

Suppose that a sparse family of vectors  $\{\phi_p\}_{p \in \Lambda}$  has been selected to approximate a signal  $f$ . An approximation can be recovered as an orthogonal projection in

the space  $\mathbf{V}_\Lambda$  generated by these vectors. We then face one of the following two problems.

1. In a *dual-synthesis* problem, the orthogonal projection  $f_\Lambda$  of  $f$  in  $\mathbf{V}_\Lambda$  must be computed from dictionary coefficients,  $\{\langle f, \phi_p \rangle\}_{p \in \Lambda}$ , provided by an analysis operator. This is the case when a signal transform  $\{\langle f, \phi_p \rangle\}_{p \in \Gamma}$  is calculated in some large dictionary and a subset of inner products are selected. Such inner products may correspond to coefficients above a threshold or local maxima values.
2. In a *dual-analysis* problem, the decomposition coefficients of  $f_\Lambda$  must be computed on a family of selected vectors  $\{\phi_p\}_{p \in \Lambda}$ . This problem appears when sparse representation algorithms select vectors as opposed to inner products. This is the case for pursuit algorithms, which compute approximation supports in highly redundant dictionaries.

The frame theory gives energy equivalence conditions to solve both problems with stable operators. A family  $\{\phi_p\}_{p \in \Lambda}$  is a frame of the space  $\mathbf{V}$  it generates if there exists  $B \geq A > 0$  such that

$$\forall h \in \mathbf{V}, \quad A\|h\|^2 \leq \sum_{m \in \Lambda} |\langle h, \phi_p \rangle|^2 \leq B\|h\|^2.$$

The representation is stable since any perturbation of frame coefficients implies a modification of similar magnitude on  $h$ . Chapter 5 proves that the existence of a dual frame  $\{\tilde{\phi}_p\}_{p \in \Lambda}$  that solves both the dual-synthesis and dual-analysis problems:

$$f_\Lambda = \sum_{p \in \Lambda} \langle f, \phi_p \rangle \tilde{\phi}_p = \sum_{p \in \Lambda} \langle f, \tilde{\phi}_p \rangle \phi_p. \quad (1.15)$$

Algorithms are provided to calculate these decompositions. The dual frame is also stable:

$$\forall f \in \mathbf{V}, \quad B^{-1}\|f\|^2 \leq \sum_{m \in \Gamma} |\langle f, \tilde{\phi}_p \rangle|^2 \leq B^{-1}\|f\|^2.$$

The frame bounds  $A$  and  $B$  are redundancy factors. If the vectors  $\{\phi_p\}_{p \in \Gamma}$  are normalized and linearly independent, then  $A \leq 1 \leq B$ . Such a dictionary is called a *Riesz basis* of  $\mathbf{V}$  and the dual frame is biorthogonal:

$$\forall (p, p') \in \Lambda^2, \quad \langle \phi_p, \tilde{\phi}_{p'} \rangle = \delta[p - p'].$$

When the basis is orthonormal, then both bases are equal. Analysis and synthesis problems are then identical.

The frame theory is also used to construct redundant dictionaries that define complete, stable, and redundant signal representations, where  $\mathbf{V}$  is then the whole signal space. The frame bounds measure the redundancy of such dictionaries. Chapter 5 studies the construction of windowed Fourier and wavelet frame dictionaries by

sampling their time, frequency, and scaling parameters, while controlling frame bounds. In two dimensions, directional wavelet frames include wavelets sensitive to directional image structures such as textures or edges.

To improve the sparsity of images having edges along regular geometric curves, Candès and Donoho [134] introduced curvelet frames, with elongated waveforms having different directions, positions, and scales. Images with piecewise regular edges have representations that are asymptotically more sparse by thresholding curvelet coefficients than wavelet coefficients.

### 1.4.2 Ideal Dictionary Approximations

In a redundant dictionary  $\mathcal{D} = \{\phi_p\}_{p \in \Gamma}$ , we would like to find the best approximation support  $\Lambda$  with  $M = |\Lambda|$  vectors, which minimize the error  $\|f - f_\Lambda\|^2$ . Chapter 12 proves that it is equivalent to find  $\Lambda_T$ , which minimizes the corresponding approximation Lagrangian

$$\mathcal{L}_0(T, f, \Lambda) = \|f - f_\Lambda\|^2 + T^2|\Lambda|, \quad (1.16)$$

for some multiplier  $T$ .

Compression and denoising are two applications of redundant dictionary approximations. When compressing signals by quantizing dictionary coefficients, the distortion rate varies, like the Lagrangian (1.16), with a multiplier  $T$  that depends on the quantization step. Optimizing the coder is thus equivalent to minimizing this approximation Lagrangian. For sparse representations, most of the bits are devoted to coding the geometry of the sparse approximation set  $\Lambda_T$  in  $\Gamma$ .

Estimators reducing noise from observations  $X = f + W$  are also optimized by finding a best orthogonal projector over a set of dictionary vectors. The *model selection* theory of Barron, Birgé, and Massart [97] proves that finding  $\hat{\Lambda}_T$ , which minimizes this same Lagrangian  $\mathcal{L}_0(T, X, \Lambda)$ , defines an estimator that has a risk on the same order as the minimum approximation error  $\|f - f_{\Lambda_T}\|^2$  up to a logarithmic factor. This is similar to the optimality result obtained for thresholding estimators in an orthonormal basis.

The bad news is that minimizing the approximation Lagrangian  $\mathcal{L}_0$  is an NP-hard problem and is therefore computationally intractable. It is necessary therefore to find algorithms that are sufficiently fast to compute suboptimal, but “good enough,” solutions.

### Dictionaries of Orthonormal Bases

To reduce the complexity of optimal approximations, the search can be reduced to subfamilies of orthogonal dictionary vectors. In a dictionary of orthonormal bases, any family of orthogonal dictionary vectors can be complemented to form an orthogonal basis  $\mathcal{B}$  included in  $\mathcal{D}$ . As a result, the best approximation of  $f$  from orthogonal vectors in  $\mathcal{B}$  is obtained by thresholding the coefficients of  $f$  in a “best basis” in  $\mathcal{D}$ .

For tree dictionaries of orthonormal bases obtained by a recursive split of orthogonal vector spaces, the fast, dynamic programming algorithm of Coifman and

Wickerhauser [182] finds such a best basis with  $O(P)$  operations, where  $P$  is the dictionary size.

Wavelet packet and local cosine bases are examples of tree dictionaries of time-frequency orthonormal bases of size  $P = N \log_2 N$ . A best basis is a time-frequency tiling that is the best match to the signal time-frequency structures.

To approximate geometrically regular edges, wavelets are not as efficient as curvelets, but wavelets provide more sparse representations of singularities that are not distributed along geometrically regular curves. Bandlet dictionaries, introduced by Le Pennec, Mallat, and Peyré [342, 365], are dictionaries of orthonormal bases that can adapt to the variability of images' geometric regularity. Minimax optimal asymptotic rates are derived for compression and denoising.

### 1.4.3 Pursuit in Dictionaries

Approximating signals only from orthogonal vectors brings rigidity that limits the ability to optimize the representation. Pursuit algorithms remove this constraint with flexible procedures that search for sparse, although not necessarily optimal, dictionary approximations. Such approximations are computed by optimizing the choice of dictionary vectors  $\{\phi_p\}_{p \in \Lambda}$ .

#### **Matching Pursuit**

Matching pursuit algorithms introduced by Mallat and Zhang [366] are greedy algorithms that optimize approximations by selecting dictionary vectors one by one. The vector in  $\phi_{p_0} \in \mathcal{D}$  that best approximates a signal  $f$  is

$$\phi_{p_0} = \operatorname{argmax}_{p \in \Gamma} |\langle f, \phi_p \rangle|$$

and the residual approximation error is

$$Rf = f - \langle f, \phi_{p_0} \rangle \phi_{p_0}.$$

A matching pursuit further approximates the residue  $Rf$  by selecting another best vector  $\phi_{p_1}$  from the dictionary and continues this process over next-order residues  $R^m f$ , which produces a signal decomposition:

$$f = \sum_{m=0}^{M-1} \langle R^m f, \phi_{p_m} \rangle \phi_{p_m} + R^M f.$$

The approximation from the  $M$ -selected vectors  $\{\phi_{p_m}\}_{0 \leq m < M}$  can be refined with an orthogonal back projection on the space generated by these vectors. An orthogonal matching pursuit further improves this decomposition by orthogonalizing progressively the projection directions  $\phi_{p_m}$  during the decompositon. The resulting decompositions are applied to compression, denoising, and pattern recognition of various types of signals, images, and videos.

### Basis Pursuit

Approximating  $f$  with a minimum number of nonzero coefficients  $a[p]$  in a dictionary  $\mathcal{D}$  is equivalent to minimizing the  $\mathbf{1}^0$  norm  $\|a\|_0$ , which gives the number of nonzero coefficients. This  $\mathbf{1}^0$  norm is highly nonconvex, which explains why the resulting minimization is NP-hard. Donoho and Chen [158] thus proposed replacing the  $\mathbf{1}^0$  norm by the  $\mathbf{1}^1$  norm  $\|a\|_1 = \sum_{p \in \Gamma} |a[p]|$ , which is convex. The resulting basis pursuit algorithm computes a synthesis operator

$$f = \sum_{p \in \Gamma} a[p] \phi_p, \text{ which minimizes } \|a\|_1 = \sum_{p \in \Gamma} |a[p]|. \quad (1.17)$$

This optimal solution is calculated with a linear programming algorithm. A basis pursuit is computationally more intense than a matching pursuit, but it is a more global optimization that yields representations that can be more sparse.

In approximation, compression, or denoising applications,  $f$  is recovered with an error bounded by a precision parameter  $\varepsilon$ . The optimization (1.18) is thus relaxed by finding a synthesis such that

$$\|f - \sum_{p \in \Gamma} a[p] \phi_p\| \leq \varepsilon, \text{ which minimizes } \|a\|_1 = \sum_{p \in \Gamma} |a[p]|. \quad (1.18)$$

This is a convex minimization problem, with a solution calculated by minimizing the corresponding  $\mathbf{1}^1$  Lagrangian

$$\mathcal{L}_1(T, f, a) = \|f - \sum_{p \in \Gamma} a[p] \phi_p\|^2 + T \|a\|_1,$$

where  $T$  is a Lagrange multiplier that depends on  $\varepsilon$ . This is called an  $\mathbf{1}^1$  Lagrangian pursuit in this book. A solution  $\tilde{a}[p]$  is computed with iterative algorithms that are guaranteed to converge. The number of nonzero coordinates of  $\tilde{a}$  typically decreases as  $T$  increases.

### Incoherence for Support Recovery

Matching pursuit and  $\mathbf{1}^1$  Lagrangian pursuits are optimal if they recover the approximation support  $\Lambda_T$ , which minimizes the approximation Lagrangian

$$\mathcal{L}_0(T, f, \Lambda) = \|f - f_\Lambda\|^2 + T^2 |\Lambda|,$$

where  $f_\Lambda$  is the orthogonal projection of  $f$  in the space  $\mathbf{V}_\Lambda$  generated by  $\{\phi_p\}_{p \in \Lambda}$ . This is not always true and depends on  $\Lambda_T$ . An *Exact Recovery Criteria* proved by Tropp [464] guarantees that pursuit algorithms do recover the optimal support  $\Lambda_T$  if

$$ERC(\Lambda_T) = \max_{q \notin \Lambda_T} \sum_{p \in \Lambda_T} |\langle \tilde{\phi}_p, \phi_q \rangle| < 1, \quad (1.19)$$

where  $\{\tilde{\phi}_p\}_{p \in \Lambda_T}$  is the biorthogonal basis of  $\{\phi_p\}_{p \in \Lambda_T}$  in  $\mathbf{V}_{\Lambda_T}$ . This criterion implies that dictionary vectors  $\phi_q$  outside  $\Lambda_T$  should have a small inner product with vectors in  $\Lambda_T$ .

This recovery is stable relative to noise perturbations if  $\{\phi_p\}_{p \in \Lambda}$  has Riesz bounds that are not too far from 1. These vectors should be nearly orthogonal and hence have small inner products. These small inner-product conditions are interpreted as a form of incoherence. A stable recovery of  $\Lambda_T$  is possible if vectors in  $\Lambda_T$  are incoherent with respect to other dictionary vectors and are incoherent between themselves. It depends on the geometric configuration of  $\Lambda_T$  in  $\Gamma$ .

## 1.5 INVERSE PROBLEMS

Most digital measurement devices, such as cameras, microphones, or medical imaging systems, can be modeled as a linear transformation of an incoming analog signal, plus noise due to intrinsic measurement fluctuations or to electronic noises. This linear transformation can be decomposed into a stable analog-to-digital linear conversion followed by a discrete operator  $U$  that carries the specific transfer function of the measurement device. The resulting measured data can be written

$$Y[q] = Uf[q] + W[q],$$

where  $f \in \mathbb{C}^N$  is the high-resolution signal we want to recover, and  $W[q]$  is the measurement noise. For a camera with an optic that is out of focus, the operator  $U$  is a low-pass convolution producing a blur. For a magnetic resonance imaging system,  $U$  is a Radon transform integrating the signal along rays and the number  $Q$  of measurements is smaller than  $N$ . In such problems,  $U$  is not invertible and recovering an estimate of  $f$  is an *ill-posed* inverse problem.

Inverse problems are among the most difficult signal-processing problems with considerable applications. When data acquisition is difficult, costly, or dangerous, or when the signal is degraded, super-resolution is important to recover the highest possible resolution information. This applies to satellite observations, seismic exploration, medical imaging, radar, camera phones, or degraded Internet videos displayed on high-resolution screens. Separating mixed information sources from fewer measurements is yet another super-resolution problem in telecommunication or audio recognition.

Incoherence, sparsity, and geometry play a crucial role in the solution of ill-defined inverse problems. With a sensing matrix  $U$  with random coefficients, Candès and Tao [139] and Donoho [217] proved that super-resolution becomes stable for signals having a sufficiently sparse representation in a dictionary. This remarkable result opens the door to new compression sensing devices and algorithms that recover high-resolution signals from a few randomized linear measurements.

### 1.5.1 Diagonal Inverse Estimation

In an ill-posed inverse problem,

$$Y = Uf + W$$

the image space  $\mathbf{Im}U = \{Uh : h \in \mathbb{C}^N\}$  of  $U$  is of dimension  $Q$  smaller than the high-resolution space  $N$  where  $f$  belongs. Inverse problems include two difficulties. In the image space  $\mathbf{Im}U$ , where  $U$  is invertible, its inverse may amplify the noise  $W$ , which then needs to be reduced by an efficient denoising procedure. In the null space  $\mathbf{Null}U$ , all signals  $h$  are set to zero  $Uh = 0$  and thus disappear in the measured data  $Y$ . Recovering the projection of  $f$  in  $\mathbf{Null}U$  requires using some strong prior information. A super-resolution estimator recovers an estimation of  $f$  in a dimension space larger than  $Q$  and hopefully equal to  $N$ , but this is not always possible.

#### Singular Value Decompositions

Let  $f = \sum_{m \in \Gamma} a[m] g_m$  be the representation of  $f$  in an orthonormal basis  $\mathcal{B} = \{g_m\}_{m \in \Gamma}$ . An approximation must be recovered from

$$Y = \sum_{m \in \Gamma} a[m] Ug_m + W.$$

A basis  $\mathcal{B}$  of singular vectors diagonalizes  $U^*U$ . Then  $U$  transforms a subset of  $Q$  vectors  $\{g_m\}_{m \in \Gamma_Q}$  of  $\mathcal{B}$  into an orthogonal basis  $\{Ug_m\}_{m \in \Gamma_Q}$  of  $\mathbf{Im}U$  and sets all other vectors to zero. A singular value decomposition estimates the coefficients  $a[m]$  of  $f$  by projecting  $Y$  on this singular basis and by renormalizing the resulting coefficients

$$\forall m \in \Gamma, \quad \tilde{a}[m] = \frac{\langle Y, Ug_m \rangle}{\|Ug_m\|^2 + h_m^2},$$

where  $h_m^2$  are regularization parameters.

Such estimators recover nonzero coefficients in a space of dimension  $Q$  and thus bring no super-resolution. If  $U$  is a convolution operator, then  $\mathcal{B}$  is the Fourier basis and a singular value estimation implements a regularized inverse convolution.

#### Diagonal Thresholding Estimation

The basis that diagonalizes  $U^*U$  rarely provides a sparse signal representation. For example, a Fourier basis that diagonalizes convolution operators does not efficiently approximate signals including singularities.

Donoho [214] introduced more flexibility by looking for a basis  $\mathcal{B}$  providing a sparse signal representation, where a subset of  $Q$  vectors  $\{g_m\}_{m \in \Gamma_Q}$  are transformed by  $U$  in a Riesz basis  $\{Ug_m\}_{m \in \Gamma_Q}$  of  $\mathbf{Im}U$ , while the others are set to zero. With an appropriate renormalization,  $\{\lambda_m^{-1} Ug_m\}_{m \in \Gamma_Q}$  has a biorthogonal basis  $\{\tilde{\phi}_m\}_{m \in \Gamma_Q}$

that is normalized  $\|\tilde{\phi}_m\| = 1$ . The sparse coefficients of  $f$  in  $\mathcal{B}$  can then be estimated with a thresholding

$$\forall m \in \Gamma_Q, \quad \tilde{a}[m] = \rho_{T_m}(\tilde{\lambda}_m^{-1}\langle Y, \tilde{\phi}_m \rangle) \quad \text{with } \rho_T(x) = x \mathbf{1}_{|x| > T},$$

for thresholds  $T_m$  appropriately defined.

For classes of signals that are sparse in  $\mathcal{B}$ , such thresholding estimators may yield a nearly minimax risk, but they provide no super-resolution since this non-linear projector remains in a space of dimension  $Q < P$ . This result applies to classes of convolution operators  $U$  in wavelet or wavelet packet bases. Diagonal inverse estimators are computationally efficient and potentially optimal in cases where super-resolution is not possible.

### 1.5.2 Super-resolution and Compressive Sensing

Suppose that  $f$  has a sparse representation in some dictionary  $\mathcal{D} = \{g_p\}_{p \in \Gamma}$  of  $P$  normalized vectors. The  $P$  vectors of the transformed dictionary  $\mathcal{D}_U = U\mathcal{D} = \{Ug_p\}_{p \in \Gamma}$  belong to the space  $\mathbf{Im}U$  of dimension  $Q < P$  and thus define a redundant dictionary. Vectors in the approximation support  $\Lambda$  of  $f$  are not restricted a priori to a particular subspace of  $\mathbb{C}^N$ . Super-resolution is possible if the approximation support  $\Lambda$  of  $f$  in  $\mathcal{D}$  can be estimated by decomposing the noisy data  $Y$  over  $\mathcal{D}_U$ . It depends on the properties of the approximation support  $\Lambda$  of  $f$  in  $\Gamma$ .

#### **Geometric Conditions for Super-resolution**

Let  $w_\Lambda = f - f_\Lambda$  be the approximation error of a sparse representation  $f_\Lambda = \sum_{p \in \Lambda} a[p]g_p$  of  $f$ . The observed signal can be written as

$$Y = Uf + W = \sum_{p \in \Lambda} a[p]Ug_p + Uw_\Lambda + W.$$

If the support  $\Lambda$  can be identified by finding a sparse approximation of  $Y$  in  $\mathcal{D}_U$

$$Y_\Lambda = \sum_{p \in \Lambda} \tilde{a}[p]Ug_p,$$

then we can recover a super-resolution estimation of  $f$

$$\tilde{F} = \sum_{p \in \Lambda} \tilde{a}[p]g_p.$$

This shows that super-resolution is possible if the approximation support  $\Lambda$  can be identified by decomposing  $Y$  in the redundant transformed dictionary  $\mathcal{D}_U$ . If the exact recovery criteria is satisfy  $ERC(\Lambda) < 1$  and if  $\{Ug_p\}_{p \in \Lambda}$  is a Riesz basis, then  $\Lambda$  can be recovered using pursuit algorithms with controlled error bounds.

For most operator  $U$ , not all sparse approximation sets can be recovered. It is necessary to impose some further geometric conditions on  $\Lambda$  in  $\Gamma$ , which makes super-resolution difficult and often unstable. Numerical applications to sparse spike deconvolution, tomography, super-resolution zooming, and inpainting illustrate these results.

### Compressive Sensing with Randomness

Candès and Tao [139], and Donoho [217] proved that stable super-resolution is possible for any sufficiently sparse signal  $f$  if  $U$  is an operator with random coefficients. Compressive sensing then becomes possible by recovering a close approximation of  $f \in \mathbb{C}^N$  from  $Q \ll N$  linear measurements [133].

A recovery is stable for a sparse approximation set  $|\Lambda| \leq M$  only if the corresponding dictionary family  $\{Ug_m\}_{m \in \Lambda}$  is a Riesz basis of the space it generates. The *M-restricted isometry conditions* of Candès, Tao, and Donoho [217] imposes uniform Riesz bounds for all sets  $\Lambda \subset \Gamma$  with  $|\Lambda| \leq M$ :

$$\forall c \in \mathbb{C}^{|\Lambda|}, \quad (1 - \delta_M) \|c\|^2 \leq \left\| \sum_{m \in \Lambda} c[m] U g_m \right\|^2 \leq (1 + \delta_M) \|c\|^2. \quad (1.20)$$

This is a strong incoherence condition on the  $P$  vectors of  $\{Ug_m\}_{m \in \Gamma}$ , which supposes that any subset of less than  $M$  vectors is nearly uniformly distributed on the unit sphere of  $\text{Im } U$ .

For an orthogonal basis  $\mathcal{D} = \{g_m\}_{m \in \Gamma}$ , this is possible for  $M \leq C Q (\log N)^{-1}$  if  $U$  is a matrix with independent Gaussian random coefficients. A pursuit algorithm then provides a stable approximation of any  $f \in \mathbb{C}^N$  having a sparse approximation from vectors in  $\mathcal{D}$ .

These results open a new compressive-sensing approach to signal acquisition and representation. Instead of first discretizing linearly the signal at a high-resolution  $N$  and then computing a nonlinear representation over  $M$  coefficients in some dictionary, compressive-sensing measures directly  $M$  randomized linear coefficients. A reconstructed signal is then recovered by a nonlinear algorithm, producing an error that can be of the same order of magnitude as the error obtained by the more classic two-step approximation process, with a more economic acquisition process. These results remain valid for several types of random matrices  $U$ . Examples of applications to single-pixel cameras, video super-resolution, new analog-to-digital converters, and MRI imaging are described.

### Blind Source Separation

Sparsity in redundant dictionaries also provides efficient strategies to separate a family of signals  $\{f_s\}_{0 \leq s < S}$  that are linearly mixed in  $K \leq S$  observed signals with noise:

$$Y_k[n] = \sum_{s=0}^{S-1} u_{k,s} f_s[n] + W_k[n] \quad \text{for } 0 \leq n < N \quad \text{and } 0 \leq k < K.$$

From a stereo recording, separating the sounds of  $S$  musical instruments is an example of source separation with  $k = 2$ . Most often the mixing matrix  $U = \{u_{k,s}\}_{0 \leq k < K, 0 \leq s < S}$  is unknown. Source separation is a super-resolution problem since  $SN$  data values must be recovered from  $QN \leq SN$  measurements. Not knowing the operator  $U$  makes it even more complicated.

If each source  $f_s$  has a sparse approximation support  $\Lambda_s$  in a dictionary  $\mathcal{D}$ , with  $\sum_{s=0}^{S-1} |\Lambda_s| \ll N$ , then it is likely that the sets  $\{\Lambda_s\}_{0 \leq s < S}$  are nearly disjoint. In this

case, the operator  $U$ , the supports  $\Lambda_s$ , and the sources  $f_s$  are approximated by computing sparse approximations of the observed data  $Y_k$  in  $\mathcal{D}$ . The distribution of these coefficients identifies the coefficients of the mixing matrix  $U$  and the nearly disjoint source supports. Time-frequency separation of sounds illustrate these results.

## 1.6 TRAVEL GUIDE

### 1.6.1 Reproducible Computational Science

This book covers the whole spectrum from theorems on functions of continuous variables to fast discrete algorithms and their applications. Section 1.1.2 argues that models based on continuous time functions give useful asymptotic results for understanding the behavior of discrete algorithms. Still, a mathematical analysis alone is often unable to fully predict the behavior and suitability of algorithms for specific signals. Experiments are necessary and such experiments should be reproducible, just like experiments in other fields of science [124].

The reproducibility of experiments requires having complete software and full source code for inspection, modification, and application under varied parameter settings. Following this perspective, computational algorithms presented in this book are available as MATLAB subroutines or in other software packages. Figures can be reproduced and the source code is available. Software demonstrations and selected exercise solutions are available at <http://wavelet-tour.com>. For the instructor, solutions are available at [www.elsevierdirect.com/9780123743701](http://www.elsevierdirect.com/9780123743701).

### 1.6.2 Book Road Map

Some redundancy is introduced between sections to avoid imposing a linear progression through the book. The preface describes several possible programs for a sparse signal-processing course.

All theorems are explained in the text and reading the proofs is not necessary to understand the results. Most of the book's theorems are proved in detail, and important techniques are included. Exercises at the end of each chapter give examples of mathematical, algorithmic, and numeric applications, ordered by level of difficulty from 1 to 4, and selected solutions can be found at <http://wavelet-tour.com>.

The book begins with Chapters 2 and 3, which review the Fourier transform and linear discrete signal processing. They provide the necessary background for readers with no signal-processing background. Important properties of linear operators, projectors, and vector spaces can be found in the Appendix. Local time-frequency transforms and dictionaries are presented in Chapter 4; the wavelet and windowed Fourier transforms are introduced and compared. The measurement of instantaneous frequencies illustrates the limitations of time-frequency resolution. Dictionary stability and redundancy are introduced in Chapter 5 through the frame theory, with examples of windowed Fourier, wavelet, and curvelet frames. Chapter 6

explains the relationship between wavelet coefficient amplitude and local signal regularity. It is applied to the detection of singularities and edges and to the analysis of multifractals.

Wavelet bases and fast filter bank algorithms are important tools presented in Chapter 7. An overdose of orthonormal bases can strike the reader while studying the construction and properties of wavelet packets and local cosine bases in Chapter 8. It is thus important to read Chapter 9, which describes sparse approximations in bases. Signal-compression and denoising applications described in Chapters 10 and 11 give life to most theoretical and algorithmic results in the book. These chapters offer a practical perspective on the relevance of linear and nonlinear signal-processing algorithms. Chapter 12 introduces sparse decompositions in redundant dictionaries and their applications. The resolution of inverse problems is studied in Chapter 13, with super-resolution, compressive sensing, and source separation.

# The Fourier Kingdom

# 2

The story begins in 1807 when Fourier presents a memoir to the Institut de France, where he claims that any periodic function can be represented as a series of harmonically related sinusoids. This idea had a profound impact in mathematical analysis, physics, and engineering, but it took a century and a half to understand the convergence of Fourier series and complete the theory of Fourier integrals.

Fourier was motivated by the study of heat diffusion, which is governed by a linear differential equation. However, the Fourier transform diagonalizes all linear time-invariant operators—the building blocks of signal processing. It therefore is not only the starting point of our exploration but also the basis of all further developments.

## 2.1 LINEAR TIME-INVARIANT FILTERING

Classic signal-processing operations, such as signal transmission, stationary noise removal, or predictive coding, are implemented with linear time-invariant operators. The time invariance of an operator  $L$  means that if the input  $f(t)$  is delayed by  $\tau$ ,  $f_\tau(t) = f(t - \tau)$ , then the output is also delayed by  $\tau$ :

$$g(t) = Lf(t) \Rightarrow g(t - \tau) = Lf_\tau(t). \quad (2.1)$$

For numerical stability, operator  $L$  must have a weak form of continuity, which means that  $Lf$  is modified by a small amount if  $f$  is slightly modified. This weak continuity is formalized by the theory of distributions [61, 64], which guarantees that we are on a safe ground without having to worry further about it.

### 2.1.1 Impulse Response

Linear time-invariant systems are characterized by their response to a Dirac impulse, defined in Section A.7 in the Appendix. If  $f$  is continuous, its value at  $t$  is obtained by an “integration” against a Dirac located at  $t$ . Let  $\delta_u(t) = \delta(t - u)$ :

$$f(t) = \int_{-\infty}^{+\infty} f(u) \delta_u(t) du.$$

The continuity and linearity of  $L$  imply that

$$Lf(t) = \int_{-\infty}^{+\infty} f(u) L\delta_u(t) du.$$

Let  $h$  be the impulse response of  $L$ :

$$h(t) = L\delta(t).$$

The time-invariance proves that  $L\delta_u(t) = h(t-u)$ , therefore

$$Lf(t) = \int_{-\infty}^{+\infty} f(u)h(t-u) du = \int_{-\infty}^{+\infty} h(u)f(t-u) du = h * f(t). \quad (2.2)$$

A time-invariant linear filter thus is equivalent to a convolution with the impulse response  $h$ . The continuity of  $f$  is not necessary. This formula remains valid for any signal  $f$  for which the convolution integral converges.

Let us recall a few useful properties of convolution products:

■ Commutativity

$$f * h(t) = h * f(t). \quad (2.3)$$

■ Differentiation

$$\frac{d}{dt}(f * h)(t) = \frac{df}{dt} * h(t) = f * \frac{dh}{dt}(t). \quad (2.4)$$

■ Dirac convolution

$$f * \delta_\tau(t) = f(t - \tau). \quad (2.5)$$

### **Stability and Causality**

A filter is said to be *causal* if  $Lf(t)$  does not depend on the values  $f(u)$  for  $u > t$ . Since

$$Lf(t) = \int_{-\infty}^{+\infty} h(u)f(t-u) du,$$

this means that  $h(u) = 0$  for  $u < 0$ . Such impulse responses are said to be *causal*.

The *stability* property guarantees that  $Lf(t)$  is bounded if  $f(t)$  is bounded. Since

$$|Lf(t)| \leq \int_{-\infty}^{+\infty} |h(u)| |f(t-u)| du \leq \sup_{u \in \mathbb{R}} |f(u)| \int_{-\infty}^{+\infty} |h(u)| du,$$

it is sufficient that  $\int_{-\infty}^{+\infty} |h(u)| du < +\infty$ . One can verify that this condition is also necessary if  $h$  is a function. We thus say that  $h$  is *stable* if it is integrable.

---

### **EXAMPLE 2.1**

An amplification and delay system is defined by

$$Lf(t) = \lambda f(t - \tau).$$

The impulse response of this filter is  $h(t) = \lambda \delta(t - \tau)$ .

---

---

**EXAMPLE 2.2**

A uniform averaging of  $f$  over intervals of size  $T$  is calculated by

$$Lf(t) = \frac{1}{T} \int_{t-T/2}^{t+T/2} f(u) du.$$

This integral can be rewritten as a convolution of  $f$  with the impulse response  $h = 1/T \mathbf{1}_{[-T/2, T/2]}$ .

---

### 2.1.2 Transfer Functions

Complex exponentials  $e^{i\omega t}$  are eigenvectors of convolution operators. Indeed,

$$Le^{i\omega t} = \int_{-\infty}^{+\infty} h(u) e^{i\omega(t-u)} du,$$

which yields

$$Le^{i\omega t} = e^{it\omega} \int_{-\infty}^{+\infty} h(u) e^{-i\omega u} du = \hat{h}(\omega) e^{i\omega t}.$$

The eigenvalue

$$\hat{h}(\omega) = \int_{-\infty}^{+\infty} h(u) e^{-i\omega u} du$$

is the Fourier transform of  $h$  at the frequency  $\omega$ . Since complex sinusoidal waves  $e^{i\omega t}$  are the eigenvectors of time-invariant linear systems, it is tempting to try to decompose any function  $f$  as a sum of these eigenvectors. We are then able to express  $Lf$  directly from the eigenvalues  $\hat{h}(\omega)$ . The Fourier analysis proves that, under weak conditions on  $f$ , it is indeed possible to write it as a Fourier integral.

---

## 2.2 FOURIER INTEGRALS

To avoid convergence issues, the Fourier integral is first defined over the space  $\mathbf{L}^1(\mathbb{R})$  of integrable functions [54]. It is then extended to the space  $\mathbf{L}^2(\mathbb{R})$  of finite energy functions [23].

### 2.2.1 Fourier Transform in $\mathbf{L}^1(\mathbb{R})$

The Fourier integral

$$\hat{f}(\omega) = \int_{-\infty}^{+\infty} f(t) e^{-i\omega t} dt \tag{2.6}$$

measures “how much” oscillations are at the frequency  $\omega$  there is in  $f$ . If  $f \in \mathbf{L}^1(\mathbb{R})$ , this integral does converge and

$$|\hat{f}(\omega)| \leq \int_{-\infty}^{+\infty} |f(t)| dt < +\infty. \tag{2.7}$$

Thus the Fourier transform is bounded, and one can verify that it is a continuous function of  $\omega$  (Exercise 2.1). If  $\hat{f}$  is also integrable, Theorem 2.1 gives the inverse Fourier transform.

**Theorem 2.1:** *Inverse Fourier Transform.* If  $f \in \mathbf{L}^1(\mathbb{R})$  and  $\hat{f} \in \mathbf{L}^1(\mathbb{R})$  then

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \hat{f}(\omega) e^{i\omega t} d\omega. \quad (2.8)$$

**Proof.** Replacing  $\hat{f}(\omega)$  by its integral expression yields

$$\frac{1}{2\pi} \int_{-\infty}^{+\infty} \hat{f}(\omega) \exp(i\omega t) d\omega = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \left( \int_{-\infty}^{+\infty} f(u) \exp[i\omega(t-u)] du \right) d\omega.$$

We cannot apply the Fubini Theorem directly because  $f(u) \exp[i\omega(t-u)]$  is not integrable in  $\mathbb{R}^2$ . To avoid this technical problem, we multiply by  $\exp(-\varepsilon^2 \omega^2/4)$ , which converges to 1 when  $\varepsilon$  goes to 0. Let us define

$$I_\varepsilon(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \left( \int_{-\infty}^{+\infty} f(u) \exp\left(\frac{-\varepsilon^2 \omega^2}{4}\right) \exp[i\omega(t-u)] du \right) d\omega. \quad (2.9)$$

We compute  $I_\varepsilon$  in two different ways using the Fubini theorem. The integration with respect to  $u$  gives

$$I_\varepsilon(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \hat{f}(\omega) \exp\left(\frac{-\varepsilon^2 \omega^2}{4}\right) \exp(i\omega t) d\omega.$$

Since

$$\left| \hat{f}(\omega) \exp\left(\frac{-\varepsilon^2 \omega^2}{4}\right) \exp[i\omega(t-u)] \right| \leq |\hat{f}(\omega)|$$

and since  $\hat{f}$  is integrable, we can apply the dominated convergence Theorem A.1, which proves that

$$\lim_{\varepsilon \rightarrow 0} I_\varepsilon(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \hat{f}(\omega) \exp(i\omega t) d\omega. \quad (2.10)$$

Let us now compute the integral (2.9) differently by applying the Fubini theorem and integrating with respect to  $\omega$ :

$$I_\varepsilon(t) = \int_{-\infty}^{+\infty} g_\varepsilon(t-u) f(u) du, \quad (2.11)$$

with

$$g_\varepsilon(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \exp(ix\omega) \exp\left(\frac{-\varepsilon^2 \omega^2}{4}\right) d\omega.$$

A change of variable  $\omega' = \varepsilon\omega$  shows that  $g_\varepsilon(x) = \varepsilon^{-1} g_1(\varepsilon^{-1}x)$ , and it is proved in (2.32) that  $g_1(x) = \pi^{-1/2} e^{-x^2}$ . The Gaussian  $g_1$  has an integral equal to 1 and a fast decay. The

squeezed Gaussians  $g_\varepsilon$  have an integral that remains equal to 1, and thus they converge to a Dirac  $\delta$  when  $\varepsilon$  goes to 0. By inserting (2.11), one can thus verify that

$$\lim_{\varepsilon \rightarrow 0} \int_{-\infty}^{+\infty} |I_\varepsilon(t) - f(t)| dt = \lim_{\varepsilon \rightarrow 0} \int \int g_\varepsilon(t-u) |f(u) - f(t)| du dt = 0.$$

Inserting (2.10) proves (2.8). ■

The inversion formula (2.8) decomposes  $f$  as a sum of sinusoidal waves  $e^{i\omega t}$  of amplitude  $\hat{f}(\omega)$ . By using this formula, we can show (Exercise 2.1) that the hypothesis  $\hat{f} \in L^1(\mathbb{R})$  implies that  $f$  must be continuous. Therefore the reconstruction (2.8) is not proved for discontinuous functions. The extension of the Fourier transform to the space  $L^2(\mathbb{R})$  will address this issue.

The most important property of the Fourier transform for signal-processing applications is the convolution theorem 2.2. It is another way to express the fact that sinusoidal waves  $e^{it\omega}$  are eigenvalues of convolution operators.

**Theorem 2.2:** *Convolution.* Let  $f \in L^1(\mathbb{R})$  and  $h \in L^1(\mathbb{R})$ . The function  $g = h \star f$  is in  $L^1(\mathbb{R})$  and

$$\hat{g}(\omega) = \hat{h}(\omega) \hat{f}(\omega). \quad (2.12)$$

**Proof.**

$$\hat{g}(\omega) = \int_{-\infty}^{+\infty} \exp(-it\omega) \left( \int_{-\infty}^{+\infty} f(t-u) h(u) du \right) dt.$$

Since  $|f(t-u)||h(u)|$  is integrable in  $\mathbb{R}^2$ , we can apply the Fubini Theorem A.2, and the change of variable  $(t, u) \rightarrow (v = t-u, u)$  yields

$$\begin{aligned} \hat{g}(\omega) &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \exp[-i(u+v)\omega] f(v) h(u) du dv \\ &= \left( \int_{-\infty}^{+\infty} \exp(-iv\omega) f(v) dv \right) \left( \int_{-\infty}^{+\infty} \exp(-iu\omega) h(u) du \right), \end{aligned}$$

which verifies (2.12). ■

The response  $Lf = g = f \star h$  of a linear time-invariant system can be calculated from its Fourier transform  $\hat{g}(\omega) = \hat{f}(\omega) \hat{h}(\omega)$  with the inverse Fourier formula,

$$g(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \hat{g}(\omega) e^{i\omega t} d\omega, \quad (2.13)$$

which yields

$$Lf(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \hat{h}(\omega) \hat{f}(\omega) e^{i\omega t} d\omega. \quad (2.14)$$

Each frequency component  $e^{it\omega}$  of amplitude  $\hat{f}(\omega)$  is amplified or attenuated by  $\hat{h}(\omega)$ . Such a convolution is thus called a *frequency filtering*, and  $\hat{h}$  is the *transfer function* of the filter.

**Table 2.1** Fourier Transform Properties

Property	Function	Fourier Transform	
	$f(t)$	$\hat{f}(\omega)$	
Inverse	$\hat{f}(t)$	$2\pi f(-\omega)$	(2.15)
Convolution	$f_1 * f_2(t)$	$\hat{f}_1(\omega)\hat{f}_2(\omega)$	(2.16)
Multiplication	$f_1(t)f_2(t)$	$\frac{1}{2\pi}\hat{f}_1 * \hat{f}_2(\omega)$	(2.17)
Translation	$f(t-u)$	$e^{-i u \omega} \hat{f}(\omega)$	(2.18)
Modulation	$e^{i \xi t} f(t)$	$\hat{f}(\omega - \xi)$	(2.19)
Scaling	$f(t/s)$	$ s  \hat{f}(s \omega)$	(2.20)
Time derivatives	$f^{(p)}(t)$	$(i\omega)^p \hat{f}(\omega)$	(2.21)
Frequency derivatives	$(-it)^p f(t)$	$\hat{f}^{(p)}(\omega)$	(2.22)
Complex conjugate	$f^*(t)$	$\hat{f}^*(-\omega)$	(2.23)
Hermitian symmetry	$f(t) \in \mathbb{R}$	$\hat{f}(-\omega) = \hat{f}^*(\omega)$	(2.24)

Table 2.1 summarizes important Fourier transform properties that are often used in calculations. Most of the formulas are proved with a change of variable in the Fourier integral.

### 2.2.2 Fourier Transform in $L^2(\mathbb{R})$

The Fourier transform of the indicator function  $f = \mathbf{1}_{[-1,1]}$  is

$$\hat{f}(\omega) = \int_{-1}^1 e^{-i\omega t} dt = \frac{2 \sin \omega}{\omega}.$$

This function is not integrable because  $f$  is not continuous, but its square is integrable. The inverse Fourier transform, Theorem 2.1, thus does not apply. This motivates the extension of the Fourier transform to the space  $L^2(\mathbb{R})$  of functions  $f$  with a finite energy  $\int_{-\infty}^{+\infty} |f(t)|^2 dt < +\infty$ . By working in the Hilbert space  $L^2(\mathbb{R})$ , we also have access to all the facilities provided by the existence of an inner product. The inner product of  $f \in L^2(\mathbb{R})$  and  $g \in L^2(\mathbb{R})$  is

$$\langle f, g \rangle = \int_{-\infty}^{+\infty} f(t) g^*(t) dt,$$

and the resulting norm in  $L^2(\mathbb{R})$  is

$$\|f\|^2 = \langle f, f \rangle = \int_{-\infty}^{+\infty} |f(t)|^2 dt.$$

Theorem 2.3 proves that inner products and norms in  $\mathbf{L}^2(\mathbb{R})$  are conserved by the Fourier transform up to a factor of  $2\pi$ . Equations (2.25) and (2.26) are called the *Parseval* and *Plancherel* formulas, respectively.

**Theorem 2.3.** If  $f$  and  $h$  are in  $\mathbf{L}^1(\mathbb{R}) \cap \mathbf{L}^2(\mathbb{R})$ , then

$$\int_{-\infty}^{+\infty} f(t) h^*(t) dt = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \hat{f}(\omega) \hat{h}^*(\omega) d\omega. \quad (2.25)$$

For  $h = f$  it follows that

$$\int_{-\infty}^{+\infty} |f(t)|^2 dt = \frac{1}{2\pi} \int_{-\infty}^{+\infty} |\hat{f}(\omega)|^2 d\omega. \quad (2.26)$$

**Proof.** Let  $g = f * \bar{h}$  with  $\bar{h}(t) = h^*(-t)$ . The convolution, Theorem 2.2, and property (2.23) show that  $\hat{g}(\omega) = \hat{f}(\omega) \hat{h}^*(\omega)$ . The reconstruction formula (2.8) applied to  $g(0)$  yields

$$\int_{-\infty}^{+\infty} f(t) h^*(t) dt = g(0) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \hat{g}(\omega) d\omega = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \hat{f}(\omega) \hat{h}^*(\omega) d\omega. \quad \blacksquare$$

### Density Extension in $\mathbf{L}^2(\mathbb{R})$

If  $f \in \mathbf{L}^2(\mathbb{R})$  but  $f \notin \mathbf{L}^1(\mathbb{R})$ , its Fourier transform cannot be calculated with the Fourier integral (2.6) because  $f(t) e^{i\omega t}$  is not integrable. It is defined as a limit using the Fourier transforms of functions in  $\mathbf{L}^1(\mathbb{R}) \cap \mathbf{L}^2(\mathbb{R})$ .

Since  $\mathbf{L}^1(\mathbb{R}) \cap \mathbf{L}^2(\mathbb{R})$  is dense in  $\mathbf{L}^2(\mathbb{R})$ , one can find a family  $\{f_n\}_{n \in \mathbb{Z}}$  of functions in  $\mathbf{L}^1(\mathbb{R}) \cap \mathbf{L}^2(\mathbb{R})$  that converges to  $f$ :

$$\lim_{n \rightarrow +\infty} \|f - f_n\| = 0.$$

Since  $\{f_n\}_{n \in \mathbb{Z}}$  converges, it is a Cauchy sequence, which means that  $\|f_n - f_p\|$  is arbitrarily small if  $n$  and  $p$  are large enough. Moreover,  $f_n \in \mathbf{L}^1(\mathbb{R})$ , so its Fourier transform  $\hat{f}_n$  is well defined.

The Plancherel formula (2.26) proves that  $\{\hat{f}_n\}_{n \in \mathbb{Z}}$  is also a Cauchy sequence because

$$\|\hat{f}_n - \hat{f}_p\| = \sqrt{2\pi} \|f_n - f_p\|$$

is arbitrarily small for large enough  $n$  and  $p$ . A Hilbert space (Appendix A.2) is complete, which means that all Cauchy sequences converge to an element of the space. Thus, there exists  $\hat{f} \in \mathbf{L}^2(\mathbb{R})$  such that

$$\lim_{n \rightarrow +\infty} \|\hat{f} - \hat{f}_n\| = 0.$$

By definition,  $\hat{f}$  is the Fourier transform of  $f$ . This extension of the Fourier transform to  $\mathbf{L}^2(\mathbb{R})$  satisfies the convolution theorem, the Parseval and Plancherel formulas, as well as all properties (2.15 to 2.24).

### Diracs

Diracs are often used in calculations; their properties are summarized in Section A.7 in the Appendix. A Dirac  $\delta$  associates its value to a function at  $t = 0$ . Since  $e^{i\omega t} = 1$  at  $t = 0$ , it seems reasonable to define its Fourier transform by

$$\hat{\delta}(\omega) = \int_{-\infty}^{+\infty} \delta(t) e^{-i\omega t} dt = 1. \quad (2.27)$$

This formula is justified mathematically by the extension of the Fourier transform to tempered distributions [61, 64].

### 2.2.3 Examples

The following examples often appear in Fourier calculations. They also illustrate important Fourier transform properties.

The *indicator function*  $f = \mathbf{1}_{[-T, T]}$  is discontinuous at  $t = \pm T$ . Its Fourier transform is therefore not integrable:

$$\hat{f}(\omega) = \int_{-T}^T e^{-i\omega t} dt = \frac{2 \sin(T\omega)}{\omega}. \quad (2.28)$$

An *ideal low-pass filter* has a transfer function  $\hat{\phi} = \mathbf{1}_{[-\xi, \xi]}$  that selects low frequencies over  $[-\xi, \xi]$ . The impulse response is calculated with the inverse Fourier integral (2.8):

$$\phi(t) = \frac{1}{2\pi} \int_{-\xi}^{\xi} e^{i\omega t} d\omega = \frac{\sin(\xi t)}{\pi t}. \quad (2.29)$$

A *passive electronic circuit* implements analog filters with resistances, capacities, and inductors. The input voltage  $f(t)$  is related to the output voltage  $g(t)$  by a differential equation with constant coefficients:

$$\sum_{k=0}^K a_k f^{(k)}(t) = \sum_{k=0}^M b_k g^{(k)}(t). \quad (2.30)$$

Suppose that the circuit is not charged for  $t < 0$ , which means that  $f(t) = g(t) = 0$ . The output  $g$  is a linear time-invariant function of  $f$  and thus can be written  $g = f \star \phi$ . Computing the Fourier transform of (2.30) and applying (2.22) proves that

$$\hat{\phi}(\omega) = \frac{\hat{g}(\omega)}{\hat{f}(\omega)} = \frac{\sum_{k=0}^K a_k (i\omega)^k}{\sum_{k=0}^M b_k (i\omega)^k}. \quad (2.31)$$

It therefore is a rational function of  $i\omega$ . An ideal low-pass transfer function  $\mathbf{1}_{[-\xi, \xi]}$  thus cannot be implemented by an analog circuit. It must be approximated by a rational function. Chebyshev or Butterworth filters are often used for this purpose [14].

A Gaussian  $f(t) = \exp(-t^2)$  is a  $C^\infty$  function with a fast asymptotic decay. Its Fourier transform is also a Gaussian:

$$\hat{f}(\omega) = \sqrt{\pi} \exp(-\omega^2/4). \quad (2.32)$$

This Fourier transform is computed by showing with an integration by parts that  $\hat{f}'(\omega) = \int_{-\infty}^{+\infty} \exp(-t^2) e^{-i\omega t} dt$  is differentiable and satisfies the differential equation

$$2\hat{f}'(\omega) + \omega\hat{f}(\omega) = 0. \quad (2.33)$$

The solution to this equation is a Gaussian  $\hat{f}(\omega) = K \exp(-\omega^2/4)$ , and since  $\hat{f}(0) = \int_{-\infty}^{+\infty} \exp(-t^2) dt = \sqrt{\pi}$ , we obtain (2.32).

A Gaussian *chirp*  $f(t) = \exp[-(a - ib)t^2]$  has a Fourier transform calculated with a similar differential equation

$$\hat{f}(\omega) = \sqrt{\frac{\pi}{a - ib}} \exp\left(\frac{-(a + ib)\omega^2}{4(a^2 + b^2)}\right). \quad (2.34)$$

A translated *Dirac*  $\delta_\tau(t) = \delta(t - \tau)$  has a Fourier transform calculated by evaluating  $e^{-i\omega t}$  at  $t = \tau$ :

$$\hat{\delta}_\tau(\omega) = \int_{-\infty}^{+\infty} \delta(t - \tau) e^{-i\omega t} dt = e^{-i\omega\tau}. \quad (2.35)$$

The *Dirac comb* is a sum of translated Diracs

$$c(t) = \sum_{n=-\infty}^{+\infty} \delta(t - nT)$$

used to uniformly sample analog signals. Its Fourier transform is derived from (2.35):

$$\hat{c}(\omega) = \sum_{n=-\infty}^{+\infty} e^{-inT\omega}. \quad (2.36)$$

The Poisson formula proves that it is also equal to a Dirac comb with a spacing equal to  $2\pi/T$ .

**Theorem 2.4: Poisson Formula.** In the sense of distribution equalities (A.29),

$$\sum_{n=-\infty}^{+\infty} e^{-inT\omega} = \frac{2\pi}{T} \sum_{k=-\infty}^{+\infty} \delta\left(\omega - \frac{2\pi k}{T}\right). \quad (2.37)$$

**Proof.** The Fourier transform  $\hat{c}$  in (2.36) is periodic with period  $2\pi/T$ . To verify the Poisson formula, it is therefore sufficient to prove that the restriction of  $\hat{c}$  to  $[-\pi/T, \pi/T]$  is equal

to  $2\pi/T \delta$ . The formula (2.37) is proved in the sense of a distribution equality (A.29) by showing that for any test function  $\hat{\theta}(\omega)$  with a support included in  $[-\pi/T, \pi/T]$ ,

$$\langle \hat{c}, \hat{\theta} \rangle = \lim_{N \rightarrow +\infty} \int_{-\infty}^{+\infty} \sum_{n=-N}^N \exp(-inT\omega) \hat{\theta}(\omega) d\omega = \frac{2\pi}{T} \hat{\theta}(0).$$

The sum of the geometric series is

$$\sum_{n=-N}^N \exp(-inT\omega) = \frac{\sin[(N+1/2)T\omega]}{\sin[T\omega/2]}. \quad (2.38)$$

Thus,

$$\langle \hat{c}, \hat{\theta} \rangle = \lim_{N \rightarrow +\infty} \frac{2\pi}{T} \int_{-\pi/T}^{\pi/T} \frac{\sin[(N+1/2)T\omega]}{\pi\omega} \frac{T\omega/2}{\sin[T\omega/2]} \hat{\theta}(\omega) d\omega. \quad (2.39)$$

Let

$$\hat{\psi}(\omega) = \begin{cases} \hat{\theta}(\omega) \frac{T\omega/2}{\sin[T\omega/2]} & \text{if } |\omega| \leq \pi/T \\ 0 & \text{if } |\omega| > \pi/T \end{cases}$$

and  $\psi(t)$  be the inverse Fourier transform of  $\hat{\psi}(\omega)$ . Since  $2\omega^{-1} \sin(a\omega)$  is the Fourier transform of  $\mathbf{1}_{[-a,a]}(t)$ , the Parseval formula (2.25) implies

$$\begin{aligned} \langle \hat{c}, \hat{\theta} \rangle &= \lim_{N \rightarrow +\infty} \frac{2\pi}{T} \int_{-\infty}^{+\infty} \frac{\sin[(N+1/2)T\omega]}{\pi\omega} \hat{\psi}(\omega) d\omega \\ &= \lim_{N \rightarrow +\infty} \frac{2\pi}{T} \int_{-(N+1/2)T}^{(N+1/2)T} \psi(t) dt. \end{aligned} \quad (2.40)$$

When  $N$  goes to  $+\infty$  the integral converges to  $\hat{\psi}(0) = \hat{\theta}(0)$ . ■

## 2.3 PROPERTIES

### 2.3.1 Regularity and Decay

The global regularity of a signal  $f$  depends on the decay of  $|\hat{f}(\omega)|$  when the frequency  $\omega$  increases. The differentiability of  $f$  is studied. If  $\hat{f} \in \mathbf{L}^1(\mathbb{R})$ , then the Fourier inversion formula (2.8) implies that  $f$  is continuous and bounded:

$$|f(t)| \leq \frac{1}{2\pi} \int_{-\infty}^{+\infty} |e^{i\omega t} \hat{f}(\omega)| d\omega = \frac{1}{2\pi} \int_{-\infty}^{+\infty} |\hat{f}(\omega)| d\omega < +\infty. \quad (2.41)$$

Theorem 2.5 applies this property to obtain a sufficient condition that guarantees the differentiability of  $f$  at any order  $p$ .

**Theorem 2.5.** A function  $f$  is bounded and  $p$  times continuously differentiable with bounded derivatives if

$$\int_{-\infty}^{+\infty} |\hat{f}(\omega)| (1 + |\omega|^p) d\omega < +\infty. \quad (2.42)$$

**Proof.** The Fourier transform of the  $k$ th-order derivative  $f^{(k)}(t)$  is  $(i\omega)^k \hat{f}(\omega)$ . Applying (2.41) to this derivative proves that

$$|f^{(k)}(t)| \leq \int_{-\infty}^{+\infty} |\hat{f}(\omega)| |\omega|^k d\omega.$$

Condition (2.42) implies that

$$\int_{-\infty}^{+\infty} |\hat{f}(\omega)| |\omega|^k d\omega < +\infty$$

for any  $k \leq p$ , so  $f^{(k)}(t)$  is continuous and bounded. ■

This result proves that if a constant  $K$  and  $\varepsilon > 0$  exist such that

$$|\hat{f}(\omega)| \leq \frac{K}{1 + |\omega|^{p+1+\varepsilon}}, \quad \text{then } f \in C^p.$$

If  $\hat{f}$  has a compact support, then (2.42) implies that  $f \in C^\infty$ .

The decay of  $|\hat{f}(\omega)|$  depends on the worst singular behavior of  $f$ . For example,  $f = \mathbf{1}_{[-T, T]}$  is discontinuous at  $t = \pm T$ , so  $|\hat{f}(\omega)|$  decays like  $|\omega|^{-1}$ . In this case, it could also be important to know that  $f(t)$  is regular for  $t \neq \pm T$ . This information cannot be derived from the decay of  $|\hat{f}(\omega)|$ . To characterize local regularity of a signal  $f$ , it is necessary to decompose it over waveforms that are sufficiently localized in time, as opposed to sinusoidal waves  $e^{i\omega t}$ . Section 6.1.3 explains that wavelets are particularly appropriate for this purpose.

### 2.3.2 Uncertainty Principle

Can we construct a function  $f$ , with an energy that is highly localized in time and with a Fourier transform  $\hat{f}$  having an energy concentrated in a small-frequency interval? The Dirac  $\delta(t - u)$  has a support restricted to  $t = u$ , but its Fourier transform  $e^{-iu\omega}$  has an energy uniformly spread over all frequencies. We know that  $|\hat{f}(\omega)|$  decays quickly at high frequencies only if  $f$  has regular variations in time. The energy of  $f$  therefore must be spread over a relatively large domain.

To reduce the time spread of  $f$ , we can scale it by  $s < 1$ , while keeping its total energy constant. If

$$f_s(t) = \frac{1}{\sqrt{s}} f\left(\frac{t}{s}\right), \quad \text{then } \|f_s\|^2 = \|f\|^2.$$

The Fourier transform  $\hat{f}_s(\omega) = \sqrt{s} \hat{f}(s\omega)$  is dilated by  $1/s$ , so we lose in frequency localization what we gained in time. Underlying is a trade-off between time and frequency localization.

Time and frequency energy concentrations are restricted by the Heisenberg uncertainty principle. This principle has a particularly important interpretation in quantum mechanics as an uncertainty on the position and momentum of a free particle. The state of a one-dimensional particle is described by a wave function

$f \in L^2(\mathbb{R})$ . The probability density that this particle is located at  $t$  is  $\frac{1}{\|f\|^2} |f(t)|^2$ . The probability density that its momentum is equal to  $\omega$  is  $\frac{1}{2\pi\|f\|^2} |\hat{f}(\omega)|^2$ . The average location of this particle is

$$u = \frac{1}{\|f\|^2} \int_{-\infty}^{+\infty} t |f(t)|^2 dt, \quad (2.43)$$

and the average momentum is

$$\xi = \frac{1}{2\pi\|f\|^2} \int_{-\infty}^{+\infty} \omega |\hat{f}(\omega)|^2 d\omega. \quad (2.44)$$

The variances around these average values are, respectively,

$$\sigma_t^2 = \frac{1}{\|f\|^2} \int_{-\infty}^{+\infty} (t - u)^2 |f(t)|^2 dt \quad (2.45)$$

and

$$\sigma_\omega^2 = \frac{1}{2\pi\|f\|^2} \int_{-\infty}^{+\infty} (\omega - \xi)^2 |\hat{f}(\omega)|^2 d\omega. \quad (2.46)$$

The larger  $\sigma_t$ , the more uncertainty there is concerning the position of the free particle; the larger  $\sigma_\omega$ , the more uncertainty there is concerning its momentum.

**Theorem 2.6:** *Heisenberg Uncertainty.* The temporal variance and the frequency variance of  $f \in L^2(\mathbb{R})$  satisfy

$$\sigma_t^2 \sigma_\omega^2 \geq \frac{1}{4}. \quad (2.47)$$

This inequality is an equality if and only if there exist  $(u, \xi, a, b) \in \mathbb{R}^2 \times \mathbb{C}^2$  such that

$$f(t) = a \exp[i\xi t - b(t - u)^2]. \quad (2.48)$$

**Proof.** The following proof, from Weyl [70], supposes that  $\lim_{|t| \rightarrow +\infty} \sqrt{t} f(t) = 0$ , but the theorem is valid for any  $f \in L^2(\mathbb{R})$ . If the average time and frequency localization of  $f$  is  $u$  and  $\xi$ , then the average time and frequency location of  $\exp(-i\xi t)f(t+u)$  is zero. Thus, it is sufficient to prove the theorem for  $u = \xi = 0$ . Observe that

$$\sigma_t^2 \sigma_\omega^2 = \frac{1}{2\pi\|f\|^4} \int_{-\infty}^{+\infty} |tf(t)|^2 dt \int_{-\infty}^{+\infty} |\omega \hat{f}(\omega)|^2 d\omega. \quad (2.49)$$

Since  $i\omega \hat{f}(\omega)$  is the Fourier transform of  $f'(t)$ , the Plancherel identity (2.26) applied to  $i\omega \hat{f}(\omega)$  yields

$$\sigma_t^2 \sigma_\omega^2 = \frac{1}{\|f\|^4} \int_{-\infty}^{+\infty} |tf(t)|^2 dt \int_{-\infty}^{+\infty} |f'(t)|^2 dt. \quad (2.50)$$

Schwarz's inequality implies

$$\begin{aligned}\sigma_t^2 \sigma_\omega^2 &\geq \frac{1}{\|f\|^4} \left[ \int_{-\infty}^{+\infty} |tf'(t)f^*(t)| dt \right]^2 \\ &\geq \frac{1}{\|f\|^4} \left[ \int_{-\infty}^{+\infty} \frac{t}{2} [f'(t)f^*(t) + f'^*(t)f(t)] dt \right]^2 \\ &\geq \frac{1}{4\|f\|^4} \left[ \int_{-\infty}^{+\infty} t(|f(t)|^2)' dt \right]^2.\end{aligned}$$

Since  $\lim_{|t| \rightarrow +\infty} \sqrt{t}f(t) = 0$ , an integration by parts gives

$$\sigma_t^2 \sigma_\omega^2 \geq \frac{1}{4\|f\|^4} \left[ \int_{-\infty}^{+\infty} |f(t)|^2 dt \right]^2 = \frac{1}{4}. \quad (2.51)$$

To obtain an equality, Schwarz's inequality applied to (2.50) must be an equality. This implies that there exists  $b \in \mathbb{C}$  such that

$$f'(t) = -2b t f(t). \quad (2.52)$$

Thus, there exists  $a \in \mathbb{C}$  such that  $f(t) = a \exp(-bt^2)$ . The other steps of the proof are then equalities so that the lower bound is indeed reached. When  $u \neq 0$  and  $\xi \neq 0$ , the corresponding time and frequency translations yield (2.48). ■

In quantum mechanics, this theorem shows that we cannot arbitrarily reduce the uncertainty as to the position and the momentum of a free particle. In signal processing, the modulated Gaussians (2.48) that have a minimum joint time-frequency localization are called Gabor chirps. As expected, they are smooth functions with fast time asymptotic decay.

### Compact Support

Despite the Heisenberg uncertainty bound, we might still be able to construct a function of compact support whose Fourier transform has a compact support. Such a function would be very useful for constructing a finite impulse response filter with a band-limited transfer function. Unfortunately, Theorem 2.7 proves that it does not exist.

**Theorem 2.7.** If  $f \neq 0$  has a compact support then  $\hat{f}(\omega)$  cannot be zero on a whole interval. Similarly, if  $\hat{f} \neq 0$  has a compact support, then  $f(t)$  cannot be zero on a whole interval.

**Proof.** We prove only the first statement because the second is derived from the first by applying the Fourier transform. If  $\hat{f}$  has a compact support included in  $[-b, b]$ , then

$$f(t) = \frac{1}{2\pi} \int_{-b}^b \hat{f}(\omega) \exp(i\omega t) d\omega. \quad (2.53)$$

If  $f(t) = 0$  for  $t \in [c, d]$ , by differentiating  $n$  times under the integral at  $t_0 = (c + d)/2$ , we obtain

$$f^{(n)}(t_0) = \frac{1}{2\pi} \int_{-b}^b \hat{f}(\omega) (i\omega)^n \exp(i\omega t_0) d\omega = 0. \quad (2.54)$$

Since

$$f(t) = \frac{1}{2\pi} \int_{-b}^b \hat{f}(\omega) \exp[i\omega(t - t_0)] \exp(i\omega t_0) d\omega, \quad (2.55)$$

developing  $\exp[i\omega(t - t_0)]$  as an infinite series yields, for all  $t \in \mathbb{R}$ ,

$$f(t) = \frac{1}{2\pi} \sum_{n=0}^{+\infty} \frac{[i(t - t_0)]^n}{n!} \int_{-b}^b \hat{f}(\omega) \omega^n \exp(i\omega t_0) d\omega = 0. \quad (2.56)$$

This contradicts our assumption that  $f \neq 0$ . ■

### 2.3.3 Total Variation

Total variation measures the total amplitude of signal oscillations. It plays an important role in image processing, where its value depends on the length the image level sets. We show that a low-pass filter can considerably amplify the total variation by creating Gibbs oscillations.

#### *Variations and Oscillations*

If  $f$  is differentiable, its total variation is defined by

$$\|f\|_V = \int_{-\infty}^{+\infty} |f'(t)| dt. \quad (2.57)$$

If  $\{x_p\}_p$  are the abscissa of the local extrema of  $f$ , where  $f'(x_p) = 0$ , then

$$\|f\|_V = \sum_p |f(x_{p+1}) - f(x_p)|.$$

Thus, it measures the total amplitude of the oscillations of  $f$ . For example, if  $f(t) = \exp(-t^2)$ , then  $\|f\|_V = 2$ . If  $f(t) = \sin(\pi t)/(\pi t)$ , then  $f$  has a local extrema at  $x_p \in [p, p+1]$  for any  $p \in \mathbb{Z}$ . Since  $|f(x_{p+1}) - f(x_p)| \sim |p|^{-1}$ , we derive that  $\|f\|_V = +\infty$ .

The total variation of nondifferentiable functions can be calculated by considering the derivative in the general sense of distributions [61, 75]. This is equivalent to approximating the derivative by a finite difference on an interval  $h$  that goes to zero:

$$\|f\|_V = \lim_{h \rightarrow 0} \int_{-\infty}^{+\infty} \frac{|f(t) - f(t - h)|}{|h|} dt. \quad (2.58)$$

The total variation of discontinuous functions is thus well defined. For example, if  $f = \mathbf{1}_{[a,b]}$ , then (2.58) gives  $\|f\|_V = 2$ . We say that  $f$  has a *bounded variation* if  $\|f\|_V < +\infty$ .

Whether  $f'$  is the standard derivative of  $f$  or its generalized derivative in the sense of distributions, its Fourier transform is  $\hat{f}'(\omega) = i\omega \hat{f}(\omega)$ . Therefore

$$|\omega| |\hat{f}(\omega)| \leq \int_{-\infty}^{+\infty} |f'(t)| dt = \|f\|_V,$$

which implies that

$$|\hat{f}(\omega)| \leq \frac{\|f\|_V}{|\omega|}. \quad (2.59)$$

However,  $|\hat{f}(\omega)| = O(|\omega|^{-1})$  is not a sufficient condition to guarantee that  $f$  has bounded variation. For example, if  $f(t) = \sin(\pi t)/(\pi t)$ , then  $\hat{f} = \mathbf{1}_{[-\pi, \pi]}$  satisfies  $|\hat{f}(\omega)| \leq \pi |\omega|^{-1}$  although  $\|f\|_V = +\infty$ . In general, the total variation of  $f$  cannot be evaluated from  $|\hat{f}(\omega)|$ .

### Discrete Signals

Let  $f_N[n] = f \star \phi_N(n/N)$  be a discrete signal obtained with an averaging filter,  $\phi_N(t) = \mathbf{1}_{[0, N^{-1}]}(t)$ , and a uniform sampling at intervals  $N^{-1}$ . The discrete total variation is calculated by approximating the signal derivative by a finite difference over the sampling distance,  $h = N^{-1}$ , and replacing the integral (2.58) by a Riemann sum, which gives

$$\|f_N\|_V = \sum_n |f_N[n] - f_N[n-1]|. \quad (2.60)$$

If  $n_p$  are the abscissa of the local extrema of  $f_N$ , then

$$\|f_N\|_V = \sum_p |f_N[n_{p+1}] - f_N[n_p]|.$$

The total variation thus measures the total amplitude of the oscillations of  $f$ . In accordance with (2.58), we say that the discrete signal has a *bounded variation* if  $\|f_N\|_V$  is bounded by a constant independent of the resolution  $N$ .

### Gibbs Oscillations

Filtering a signal with a low-pass filter can create oscillations that have an infinite total variation. Let  $f_\xi = f \star \phi_\xi$  be the filtered signal obtained with an ideal low-pass filter whose transfer function is  $\hat{\phi}_\xi = \mathbf{1}_{[-\xi, \xi]}$ . If  $f \in L^2(\mathbb{R})$ , then  $f_\xi$  converges to  $f$  in  $L^2(\mathbb{R})$  norm:  $\lim_{\xi \rightarrow +\infty} \|f - f_\xi\| = 0$ . Indeed,  $\hat{f}_\xi = \hat{f} \mathbf{1}_{[-\xi, \xi]}$  and the Plancherel formula (2.26) imply that

$$\|f - f_\xi\|^2 = \frac{1}{2\pi} \int_{-\infty}^{+\infty} |\hat{f}(\omega) - \hat{f}_\xi(\omega)|^2 d\omega = \frac{1}{2\pi} \int_{|\omega| > \xi} |\hat{f}(\omega)|^2 d\omega,$$

which goes to zero as  $\xi$  increases. However, if  $f$  is discontinuous in  $t_0$ , then we show that  $f_\xi$  has Gibbs oscillations in the neighborhood of  $t_0$ , which prevents  $\sup_{t \in \mathbb{R}} |f(t) - f_\xi(t)|$  from converging to zero as  $\xi$  increases.

Let  $f$  be a bounded variation function  $\|f\|_V < +\infty$  that has an isolated discontinuity at  $t_0$ , with a left limit  $f(t_0^-)$  and right limit  $f(t_0^+)$ . It is decomposed as a sum of  $f_c$ , which is continuous in the neighborhood of  $t_0$ , plus a Heaviside step of amplitude  $f(t_0^+) - f(t_0^-)$ :

$$f(t) = f_c(t) + [f(t_0^+) - f(t_0^-)] u(t - t_0),$$

with

$$u(t) = \begin{cases} 1 & \text{if } t \geq 0 \\ 0 & \text{otherwise} \end{cases}. \quad (2.61)$$

Thus,

$$f_\xi(t) = f_c \star \phi_\xi(t) + [f(t_0^+) - f(t_0^-)] u \star \phi_\xi(t - t_0). \quad (2.62)$$

Since  $f_c$  has bounded variation and is uniformly continuous in the neighborhood of  $t_0$ , one can prove (Exercise 2.15) that  $f_c \star \phi_\xi(t)$  converges uniformly to  $f_c(t)$  in a neighborhood of  $t_0$ . The following theorem shows that this is not true for  $u \star \phi_\xi$ , which creates Gibbs oscillations.

**Theorem 2.8:** *Gibbs.* For any  $\xi > 0$ ,

$$u \star \phi_\xi(t) = \int_{-\infty}^{\xi t} \frac{\sin x}{\pi x} dx. \quad (2.63)$$

**Proof.** The impulse response of an ideal low-pass filter, calculated in (2.29), is  $\phi_\xi(t) = \sin(\xi t)/(\pi t)$ . Thus,

$$u \star \phi_\xi(t) = \int_{-\infty}^{+\infty} u(\tau) \frac{\sin \xi(t - \tau)}{\pi(t - \tau)} d\tau = \int_0^{+\infty} \frac{\sin \xi(t - \tau)}{\pi(t - \tau)} d\tau.$$

The change of variable  $x = \xi(t - \tau)$  gives (2.63). ■

The function

$$s(\xi t) = \int_{-\infty}^{\xi t} \frac{\sin x}{\pi x} dx$$

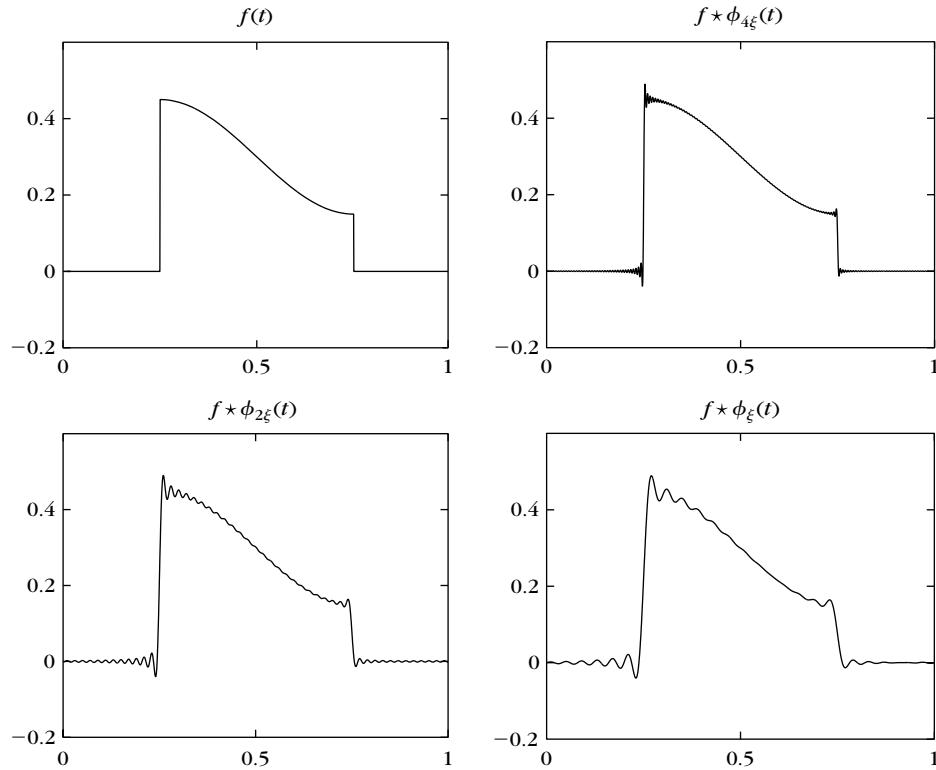
is a sigmoid that increases from 0 at  $t = -\infty$  to 1 at  $t = +\infty$ , with  $s(0) = 1/2$ . It has oscillations of period  $\pi/\xi$ , which are attenuated when the distance to 0 increases; however, their total variation is infinite:  $\|s\|_V = +\infty$ . The maximum amplitude of the Gibbs oscillations occurs at  $t = \pm \pi/\xi$ , with an amplitude independent of  $\xi$ :

$$A = s(\pi) - 1 = \int_{-\infty}^{\pi} \frac{\sin x}{\pi x} dx - 1 \approx 0.045.$$

Inserting (2.63) into (2.62) shows that

$$f(t) - f_\xi(t) = [f(t_0^+) - f(t_0^-)] s(\xi(t - t_0)) + \varepsilon(\xi, t), \quad (2.64)$$

where  $\lim_{\xi \rightarrow +\infty} \sup_{|t - t_0| < \alpha} |\varepsilon(\xi, t)| = 0$  in some neighborhood of size  $\alpha > 0$  around  $t_0$ . The sigmoid  $s(\xi(t - t_0))$  centered at  $t_0$  creates a maximum error of fixed amplitude

**FIGURE 2.1**

Gibbs oscillations created by low-pass filters with cut-off frequencies that decrease from left to right.

for all \$\xi\$. This is seen in Figure 2.1, where the Gibbs oscillations have an amplitude proportional to the jump \$f(t\_0^+) - f(t\_0^-)\$ at all frequencies \$\xi\$.

### ***Image Total Variation***

The total variation of an image \$f(x\_1, x\_2)\$ depends on the amplitude of its variations as well as the length of the contours along which they occur. Suppose that \$f(x\_1, x\_2)\$ is differentiable. The total variation is defined by

$$\|f\|_V = \int \int |\vec{\nabla}f(x_1, x_2)| dx_1 dx_2, \quad (2.65)$$

where the modulus of the gradient vector is

$$|\vec{\nabla}f(x_1, x_2)| = \left( \left| \frac{\partial f(x_1, x_2)}{\partial x_1} \right|^2 + \left| \frac{\partial f(x_1, x_2)}{\partial x_2} \right|^2 \right)^{1/2}.$$

As in one dimension, the total variation is extended to discontinuous functions by taking the derivatives in the general sense of distributions. An equivalent norm is obtained by approximating the partial derivatives by finite differences:

$$\begin{aligned} |\Delta_h f(x_1, x_2)| &= \left( \left| \frac{f(x_1, x_2) - f(x_1 - h, x_2)}{h} \right|^2 \right. \\ &\quad \left. + \left| \frac{f(x_1, x_2) - f(x_1, x_2 - h)}{h} \right|^2 \right)^{1/2}. \end{aligned}$$

One can verify that

$$\|f\|_V \leq \lim_{h \rightarrow 0} \int \int |\Delta_h f(x_1, x_2)| dx_1 dx_2 \leq \sqrt{2} \|f\|_V. \quad (2.66)$$

The finite difference integral gives a larger value when  $f(x_1, x_2)$  is discontinuous along a diagonal line in the  $(x_1, x_2)$  plane.

The total variation of  $f$  is related to the length of its level sets. Let us define

$$\Omega_y = \{(x_1, x_2) \in \mathbb{R}^2 : f(x_1, x_2) > y\}.$$

If  $f$  is continuous, then the boundary  $\partial\Omega_y$  of  $\Omega_y$  is the level set of all  $(x_1, x_2)$  such that  $f(x_1, x_2) = y$ . Let  $H^1(\partial\Omega_y)$  be the length of  $\partial\Omega_y$ . Formally, this length is calculated in the sense of the mono-dimensional Hausdorff measure. Theorem 2.9 relates the total variation of  $f$  to the length of its level sets.

**Theorem 2.9: Co-area Formula.** If  $\|f\|_V < +\infty$ , then

$$\|f\|_V = \int_{-\infty}^{+\infty} H^1(\partial\Omega_y) dy. \quad (2.67)$$

**Proof.** The proof is a highly technical result that is given in [75]. Here we give an intuitive explanation when  $f$  is continuously differentiable. In this case  $\partial\Omega_y$  is a differentiable curve  $x(y, s) \in \mathbb{R}^2$ , which is parameterized by arc-length  $s$ . Let  $\vec{\tau}(x)$  be the vector tangent to this curve in the plane. The gradient  $\vec{\nabla}f(x)$  is orthogonal to  $\vec{\tau}(x)$ . The Frenet coordinate system along  $\partial\Omega_y$  is composed of  $\vec{\tau}(x)$  and of the unit vector  $\vec{n}(x)$  parallel to  $\vec{\nabla}f(x)$ . Let  $ds$  and  $dn$  be the Lebesgue measures in the direction of  $\vec{\tau}$  and  $\vec{n}$ . We then have

$$|\vec{\nabla}f(x)| = \vec{\nabla}f(x) \cdot \vec{n} = \frac{dy}{dn}, \quad (2.68)$$

where  $dy$  is the differential of amplitudes across level sets. The idea of the proof is to decompose the total variation integral over the plane as an integral along the level sets and across level sets, which is written:

$$\|f\|_V = \int \int |\vec{\nabla}f(x_1, x_2)| dx_1 dx_2 = \int \int_{\partial\Omega_y} |\vec{\nabla}f(x(y, s))| ds dn. \quad (2.69)$$

By using (2.68), we can get

$$\|f\|_V = \int \int_{\partial\Omega_y} ds dy.$$

But  $\int_{\partial\Omega_y} ds = H^1(\partial\Omega_y)$  is the length of the level set, which justifies (2.67). ■

The co-area formula gives an important geometrical interpretation of the total image variation. Images have a bounded gray level. Thus, the integral (2.67) is calculated over a finite interval and is proportional to the average length of level sets. It is finite as long as the level sets are not fractal curves. Let  $f = \alpha \mathbf{1}_\Omega$  be proportional to the indicator function of a set  $\Omega \subset \mathbb{R}^2$  that has a boundary  $\partial\Omega$  of length  $L$ . The co-area formula (2.9) implies that  $\|f\|_V = \alpha L$ . In general, bounded variation images must have step edges of finite length.

### ***Discrete Images***

A camera measures light intensity with photoreceptors that perform an averaging and a uniform sampling over a grid that is supposed to be uniform. For a resolution  $N$ , the sampling interval is  $N^{-1}$ . The resulting image can be written  $f_N[n_1, n_2] = f * \phi_N(n_1/N, n_2/N)$ , where  $\phi_N = \mathbf{1}_{[0, N^{-1}]^2}$  and  $f$  is the averaged analog image. Its total variation is defined by approximating derivatives by finite differences and the integral (2.66) by a Riemann sum:

$$\begin{aligned} \|f_N\|_V &= \frac{1}{N} \sum_{n_1} \sum_{n_2} (|f_N[n_1, n_2] - f_N[n_1 - 1, n_2]|^2 \\ &\quad + |f_N[n_1, n_2] - f_N[n_1, n_2 - 1]|^2)^{1/2}. \end{aligned} \quad (2.70)$$

In accordance with (2.66), we say that the image has bounded variation if  $\|f_N\|_V$  is bounded by a constant independent of resolution  $N$ . The co-area formula proves that it depends on the length of the level sets as the image resolution increases. The  $\sqrt{2}$  upper-bound factor in (2.66) comes from the fact that the length of a diagonal line can be increased by  $\sqrt{2}$  if it is approximated by a zig-zag line that remains on the horizontal and vertical segments of the image-sampling grid. Figure 2.2(a) shows a bounded variation image and Figure 2.2(b) displays the level sets obtained by uniformly discretizing amplitude variable  $y$ . The total variation of this image remains nearly constant as resolution varies.

---

## **2.4 TWO-DIMENSIONAL FOURIER TRANSFORM**

The Fourier transform in  $\mathbb{R}^n$  is a straightforward extension of the one-dimensional Fourier transform. The two-dimensional case is briefly reviewed for image-processing applications. The Fourier transform of a two-dimensional integrable

**FIGURE 2.2**

- (a)** The total variation of this image remains nearly constant when resolution  $N$  increases.  
**(b)** Level sets  $\partial\Omega_y$  obtained by uniformly sampling amplitude variable  $y$ .

function,  $f \in \mathbf{L}^1(\mathbb{R}^2)$ , is

$$\hat{f}(\omega_1, \omega_2) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x_1, x_2) \exp[-i(\omega_1 x_1 + \omega_2 x_2)] dx_1 dx_2. \quad (2.71)$$

In polar coordinates,  $\exp[i(\omega_1 x_1 + \omega_2 x_2)]$  can be rewritten

$$\exp[i(\omega_1 x_1 + \omega_2 x_2)] = \exp[i\xi(x_1 \cos \theta + x_2 \sin \theta)],$$

with  $\xi = \sqrt{\omega_1^2 + \omega_2^2}$ . It is a plane wave that propagates in the direction of  $\theta$  and oscillates at frequency  $\xi$ . The properties of a two-dimensional Fourier transform are essentially the same as in one dimension. Let us summarize a few important results. We write  $\omega = (\omega_1, \omega_2)$ ,  $x = (x_1, x_2)$ ,  $\omega \cdot x = \omega_1 x_1 + \omega_2 x_2$ , and  $\iint f(x_1, x_2) dx_1 dx_2 = \iint f(x) dx$ .

- If  $f \in \mathbf{L}^1(\mathbb{R}^2)$  and  $\hat{f} \in \mathbf{L}^1(\mathbb{R}^2)$ , then

$$f(x) = \frac{1}{4\pi^2} \iint \hat{f}(\omega) \exp[i(\omega \cdot x)] d\omega. \quad (2.72)$$

- If  $f \in \mathbf{L}^1(\mathbb{R}^2)$  and  $h \in \mathbf{L}^1(\mathbb{R}^2)$ , then the convolution

$$g(x) = f \star h(x) = \iint f(u) h(x - u) du$$

has a Fourier transform

$$\hat{g}(\omega) = \hat{f}(\omega) \hat{h}(\omega). \quad (2.73)$$

- The Parseval formula proves that

$$\iint f(x) g^*(x) dx = \frac{1}{4\pi^2} \iint \hat{f}(\omega) \hat{g}^*(\omega) d\omega. \quad (2.74)$$

If  $f = g$ , we obtain the Plancherel equality

$$\iint |f(x)|^2 dx = \frac{1}{4\pi^2} \iint |\hat{f}(\omega)|^2 d\omega. \quad (2.75)$$

The Fourier transform of a finite-energy function thus has finite energy. With the same density-based argument as in one dimension, energy equivalence makes it possible to extend the Fourier transform to any function  $f \in L^2(\mathbb{R}^2)$ .

- If  $f \in L^2(\mathbb{R}^2)$  is separable, which means that

$$f(x) = f(x_1, x_2) = g(x_1) h(x_2),$$

then its Fourier transform is

$$\hat{f}(\omega) = \hat{f}(\omega_1, \omega_2) = \hat{g}(\omega_1) \hat{h}(\omega_2),$$

where  $\hat{h}$  and  $\hat{g}$  are the one-dimensional Fourier transforms of  $g$  and  $h$ . For example, the indicator function,

$$f(x_1, x_2) = \begin{cases} 1 & \text{if } |x_1| \leq T, |x_2| \leq T \\ 0 & \text{otherwise} \end{cases} = \mathbf{1}_{[-T, T]}(x_1) \times \mathbf{1}_{[-T, T]}(x_2),$$

is a separable function, the Fourier transform of which is derived from (2.28):

$$\hat{f}(\omega_1, \omega_2) = \frac{4 \sin(T\omega_1) \sin(T\omega_2)}{\omega_1 \omega_2}.$$

- If  $f(x_1, x_2)$  is rotated by  $\theta$ :

$$f_\theta(x_1, x_2) = f(x_1 \cos \theta - x_2 \sin \theta, x_1 \sin \theta + x_2 \cos \theta),$$

then its Fourier transform is rotated by  $\theta$ :

$$\hat{f}_\theta(\omega_1, \omega_2) = \hat{f}(\omega_1 \cos \theta - \omega_2 \sin \theta, \omega_1 \sin \theta + \omega_2 \cos \theta). \quad (2.76)$$

### Radon Transform

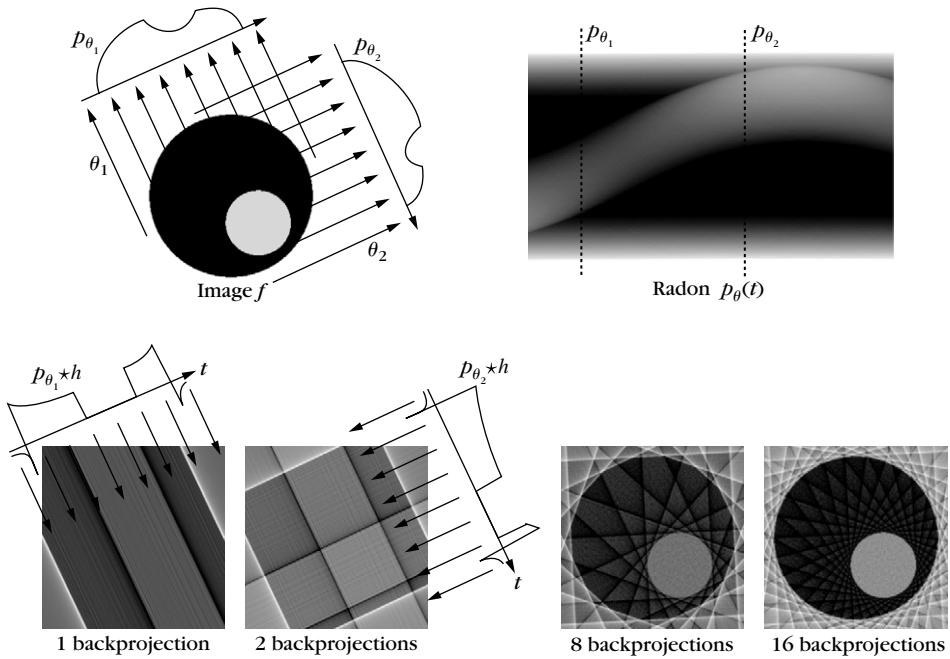
A Radon transform computes integrals of  $f \in L^2(\mathbb{R}^2)$  along rays. It provides a good model for some tomographic systems such as X-ray measurements in medical imaging. It is then necessary to invert the Radon transform to reconstruct the two- or three-dimensional body from these integrals.

Let us write  $\tau_\theta = (\cos \theta, \sin \theta)$ . A ray  $\Delta_{t,\theta}$  is a line defined by its equation

$$x \cdot \tau_\theta = x_1 \cos \theta + x_2 \sin \theta = t.$$

The projection  $p_\theta$  of  $f$  along a parallel line of orientation  $\theta$  is defined by

$$\forall \theta \in [0, \pi), \forall t \in \mathbb{R}, \quad p_\theta(t) = \int_{\Delta_{t,\theta}} f(x) ds = \iint f(x) \delta(x \cdot \tau_\theta - t) dx, \quad (2.77)$$

**FIGURE 2.3**

The Radon transform and its reconstruction with an increasing number of back projections.

where  $\delta$  is the Dirac distribution. The Radon transform maps  $f(x)$  to  $p_\theta(t)$  for  $\theta \in [0, \pi]$ .

In medical imaging applications, a scanner is rotated around an object to compute the projection  $p_\theta$  for many angles  $\theta \in [0, \pi]$ , as illustrated in Figure 2.3. The Fourier slice, Theorem 2.10, relates the Fourier transform of  $p_\theta$  to slices of the Fourier transform of  $f$ .

**Theorem 2.10: Fourier Slice.** The Fourier transform of projections satisfies

$$\forall \theta \in [0, \pi], \forall \xi \in \mathbb{R} \quad \hat{p}_\theta(\xi) = \hat{f}(\xi \cos \theta, \xi \sin \theta).$$

**Proof.** The Fourier transform of the projection is

$$\begin{aligned} \hat{p}_\theta(\xi) &= \int_{-\infty}^{+\infty} \left( \iint f(x) \delta(x \cdot \tau_\theta - t) dx \right) e^{-it\xi} dt \\ &= \iint f(x) \exp(-i(x \cdot \tau_\theta)\xi) dx = \hat{f}(\xi \tau_\theta). \end{aligned}$$

An image  $f$  can be recovered from its projection  $p_\theta$  thanks to the projection slice theorem. Indeed, the Fourier transform  $\hat{f}$ , known along each ray of direction  $\theta$

and  $f$ , is thus obtained with the 2D inverse Fourier transform 2.71. The back-projection theorem (2.11) gives an inversion formula.

**Theorem 2.11:** *Backprojection.* The image  $f$  is recovered using a one-dimensional filter  $h(t)$ :

$$f(x) = \frac{1}{2\pi} \int_0^\pi p_\theta * h(x \cdot \tau_\theta) d\theta, \quad \text{with } \hat{h}(\xi) = |\xi|.$$

**Proof.** The inverse Fourier transform 2.72 in polar coordinates  $(\omega_1, \omega_2) = (\xi \cos \theta, \xi \sin \theta)$ , with  $d\omega_1 d\omega_2 = \xi d\theta d\xi$ , can be written

$$f(x) = \frac{1}{4\pi^2} \int_0^{+\infty} \int_0^{2\pi} \hat{f}(\xi \cos \theta, \xi \sin \theta) \exp(i(x \cdot \tau_\theta) \xi) \xi d\theta d\xi.$$

Using the Fourier slice, Theorem 2.10, with  $p_{\theta+\pi}(t) = p_\theta(-t)$ , this is rewritten as

$$f(x) = \frac{1}{2\pi} \int_0^\pi \left( \frac{1}{2\pi} \int_{-\infty}^{+\infty} |\xi| \hat{p}_\theta(\xi) \exp(i(x \cdot \tau_\theta) \xi) d\xi \right) d\theta. \quad (2.78)$$

The inner integral is the inverse Fourier transform of  $\hat{p}_\theta(\xi) |\xi|$  evaluated at  $x \cdot \tau_\theta \in \mathbb{R}$ . The convolution formula 2.73 shows that it is equal to  $p_\theta * h(x \cdot \tau_\theta)$ . ■

In medical imaging applications, only a limited number of projections is available; thus, the Fourier transform  $\hat{f}$  is partially known. In this case, an approximation of  $f$  can still be recovered by summing the corresponding filtered backprojections  $p_\theta * h(x \cdot \tau_\theta)$ . Figure 2.3 shows this process, and the reconstruction of an image with a geometric object, using an increasing number of evenly spaced projections. Section 13.3 describes a nonlinear super-resolution reconstruction algorithm that recovers a more precise image by using a sparse representation.

## 2.5 EXERCISES

- 2.1 <sup>1</sup> Prove that if  $f \in L^1(\mathbb{R})$ , then  $\hat{f}(\omega)$  is a continuous function of  $\omega$ , and that if  $\hat{f} \in L^1(\mathbb{R})$ , then  $f(t)$  is continuous.
- 2.2 <sup>1</sup> Prove that a filter with impulse response  $h(t)$  is stable only if  $\int |h(t)| dt < \infty$ .
- 2.3 <sup>1</sup> Prove the translation (2.1), scaling (2.1), and time derivative (2.1) properties of the Fourier transform.
- 2.4 <sup>1</sup> Let  $f_r(t) = \operatorname{Re}[f(t)]$  and  $f_i(t) = \operatorname{Im}[f(t)]$  be the real and imaginary parts of  $f(t)$ . Prove that  $\hat{f}_r(\omega) = [\hat{f}(\omega) + \hat{f}^*(-\omega)]/2$  and  $\hat{f}_i(\omega) = [\hat{f}(\omega) - \hat{f}^*(-\omega)]/(2i)$ .

*Note:* Exercises have been ordered by level of difficulty: Level<sup>1</sup> exercises are direct applications of the course. Level<sup>2</sup> require more thinking. Level<sup>3</sup> exercises include some technical derivations. Level<sup>4</sup> are projects at the interface of research; they are possible topics for a final course project or independent study.

**2.5** <sup>1</sup> Prove that if  $\hat{f}(\omega)$  is differentiable and  $\hat{f}(0) = \hat{f}'(0) = 0$ , then

$$\int_{-\infty}^{+\infty} f(t) dt = \int_{-\infty}^{+\infty} t f(t) dt = 0.$$

**2.6** <sup>1</sup> By using the Fourier transform, verify that

$$\int_{-\infty}^{+\infty} \frac{\sin(\pi t)}{(\pi t)} dt = 1 \quad \text{and} \quad \int_{-\infty}^{+\infty} \frac{\sin^3 t}{t^3} dt = \frac{3\pi}{4}.$$

**2.7** <sup>2</sup> Show that the Fourier transform of  $f(t) = \exp(-(a - ib)t^2)$  is

$$\hat{f}(\omega) = \sqrt{\frac{\pi}{a - ib}} \exp\left(-\frac{a + ib}{4(a^2 + b^2)} \omega^2\right).$$

*Hint:* write a differential equation similar to (2.33).

**2.8** <sup>3</sup> *Riemann-Lebesgue.* Prove that if  $f \in \mathbf{L}^1(\mathbb{R})$ , then  $\lim_{\omega \rightarrow \infty} \hat{f}(\omega) = 0$ . *Hint:* Prove it first for  $\mathbf{C}^1$  functions with a compact support and use a density argument.

**2.9** <sup>2</sup> *Stability of passive circuits:*

- (a) Let  $p$  be a complex number with  $\operatorname{Re}[p] < 0$ . Compute the Fourier transforms of  $f(t) = \exp(pt) \mathbf{1}_{[0, +\infty)}(t)$  and of  $f(t) = t^n \exp(pt) \mathbf{1}_{[0, +\infty)}(t)$ .
- (b) A passive circuit relates the input voltage  $f$  to the output voltage  $g$  by a differential equation with constant coefficients:

$$\sum_{k=0}^K a_k f^{(k)}(t) = \sum_{k=0}^M b_k g^{(k)}(t).$$

Prove that this system is stable and causal if and only if the roots of the equation  $\sum_{k=0}^M b_k z^k = 0$  have a strictly negative real part.

(c) A Butterworth filter satisfies

$$|\hat{h}(\omega)|^2 = \frac{1}{1 + (\omega/\omega_0)^{2N}}.$$

For  $N = 3$ , compute  $\hat{h}(\omega)$  and  $h(t)$  so that this filter can be implemented by a stable electronic circuit.

**2.10** <sup>1</sup> For any  $A > 0$ , construct  $f$  such that the time and frequency spread measured, respectively, by  $\sigma_t$  and  $\sigma_\omega$  in (2.46) and (2.45) satisfy  $\sigma_t > A$  and  $\sigma_\omega > A$ .

**2.11** <sup>3</sup> Suppose that  $f(t) \geq 0$  and that its support is in  $[-T, T]$ . Verify that  $|\hat{f}(\omega)| \leq \hat{f}(0)$ . Let  $\omega_c$  be the half-power point defined by  $|\hat{f}(\omega_c)|^2 = |f(0)|^2/2$  and  $|f(\omega)|^2 < |f(0)|^2/2$  for  $\omega < \omega_c$ . Prove that  $\omega_c T \geq \pi/2$ .

**2.12** <sup>2</sup> *Rectification.* A rectifier computes  $g(t) = |f(t)|$  for recovering the envelope of modulated signals [54].

- (a) Show that if  $f(t) = a(t) \sin \omega_0 t$  with  $a(t) \geq 0$  then  $g(t) = |f(t)|$  satisfies

$$\hat{g}(\omega) = -\frac{2}{\pi} \sum_{n=-\infty}^{+\infty} \frac{\hat{a}(\omega - 2n\omega_0)}{4n^2 - 1}.$$

- (b) Suppose that  $\hat{a}(\omega) = 0$  for  $|\omega| > \omega_0$ . Find  $h$  such that  $a(t) = h * g(t)$ .

- 2.13** <sup>2</sup> *Amplitude modulation.* For  $0 \leq n < N$ , we suppose that  $f_n(t)$  is real and that  $\hat{f}_n(\omega) = 0$  for  $|\omega| > \omega_0$ . An amplitude-modulated multiplexed signal is defined by

$$g(t) = \sum_{n=0}^N f_n(t) \cos(2n\omega_0 t).$$

Compute  $\hat{g}(\omega)$  and verify that the width of its support is  $4N\omega_0$ . Find a demodulation algorithm that recovers each  $f_n$  from  $g$ .

- 2.14** <sup>1</sup> Show that  $\|\phi\|_V = +\infty$  if  $\phi(t) = \sin(\pi t)/(\pi t)$ . Show that  $\|\phi\|_V = 2\lambda$  if  $\phi(t) = \lambda \mathbf{1}_{[a,b]}(t)$ .

- 2.15** <sup>3</sup> Let  $f_\xi = f * h_\xi$  with  $\hat{h}_\xi = \mathbf{1}_{[-\xi, \xi]}$ . Suppose that  $f$  has a bounded variation  $\|f\|_V < +\infty$  and that it is continuous in a neighborhood of  $t_0$ . Prove that in a neighborhood of  $t_0$ ,  $f_\xi(t)$  converges uniformly to  $f(t)$  when  $\xi$  goes to  $+\infty$ .

- 2.16** <sup>1</sup> Compute the two-dimensional Fourier transforms of  $f(x) = \mathbf{1}_{[0,1]^2}(x_1, x_2)$  and of  $f(x) = e^{-(x_1^2 + x_2^2)}$ .

- 2.17** <sup>1</sup> Compute the Radon transform of the indicator function of the unit circle:  $f(x_1, x_2) = \mathbf{1}_{x^1 + x^2 \leq 1}$ .

- 2.18** <sup>2</sup> Let  $f(x_1, x_2)$  be an image that has a discontinuity of amplitude  $A$  along a straight line that has an angle  $\theta$  in the plane  $(x_1, x_2)$ . Compute the amplitude of the Gibbs oscillations of  $f * h_\xi(x_1, x_2)$  as a function of  $\xi, \theta$  and  $A$  for  $\hat{h}_\xi(\omega_1, \omega_2) = \mathbf{1}_{[-\xi, \xi]}(\omega_1) \mathbf{1}_{[-\xi, \xi]}(\omega_2)$ .

# Discrete Revolution

# 3

Digital signal processing has taken over. First used in the 1950s at the service of analog signal processing to simulate analog transforms, digital algorithms have invaded most traditional fortresses, including phones, music recording, cameras, televisions, and all information processing. Analog computations performed with electronic circuits are faster than digital algorithms implemented with microprocessors but are less precise and less flexible. Thus, analog circuits are often replaced by digital chips once the computational performance of microprocessors is sufficient to operate in real time for a given application.

Whether sound recordings or images, most discrete signals are obtained by sampling an analog signal. An analog-to-digital conversion is a linear approximation that introduces an error dependent on the sampling rate. Once more, the Fourier transform is unavoidable because the eigenvectors of discrete time-invariant operators are sinusoidal waves. The Fourier transform is discretized for signals of finite size and implemented with a fast Fourier transform (FFT) algorithm.

---

## 3.1 SAMPLING ANALOG SIGNALS

The simplest way to discretize an analog signal  $f$  is to record its sample values  $\{f(ns)\}_{n \in \mathbb{Z}}$  at interval  $s$ . An approximation of  $f(t)$  at any  $t \in \mathbb{R}$  may be recovered by interpolating these samples. The Shannon-Whittaker sampling theorem gives a sufficient condition on the support of the Fourier transform  $\hat{f}$  to recover  $f(t)$  exactly. Aliasing and approximation errors are studied when this condition is not satisfied.

Digital-acquisition devices often do not satisfy the restrictive hypothesis of the Shannon-Whittaker sampling theorem. General linear analog-to-discrete conversion is introduced in Section 3.1.3, showing that a stable uniform discretization is a linear approximation. A digital conversion also approximates discrete coefficients, with a given precision, to store them with a limited number of bits. This quantization aspect is studied in Chapter 10.

### 3.1.1 Shannon-Whittaker Sampling Theorem

Sampling is first studied from the more classic Shannon-Whittaker point of view, which tries to recover  $f(t)$  from its samples  $\{f(ns)\}_{n \in \mathbb{Z}}$ . A discrete signal can

be represented as a sum of Diracs. We associate to any sample  $f(ns)$  a Dirac  $f(ns)\delta(t - ns)$  located at  $t = ns$ . A uniform sampling of  $f$  thus corresponds to the weighted Dirac sum

$$f_d(t) = \sum_{n=-\infty}^{+\infty} f(ns) \delta(t - ns). \quad (3.1)$$

The Fourier transform of  $\delta(t - ns)$  is  $e^{-ins\omega}$ , so the Fourier transform of  $f_d$  is a Fourier series:

$$\hat{f}_d(\omega) = \sum_{n=-\infty}^{+\infty} f(ns) e^{-ins\omega}. \quad (3.2)$$

To understand how to compute  $f(t)$  from the sample values  $f(ns)$  and therefore  $f$  from  $f_d$ , we relate their Fourier transforms  $\hat{f}$  and  $\hat{f}_d$ .

**Theorem 3.1.** The Fourier transform of the discrete signal obtained by sampling  $f$  at interval  $s$  is

$$\hat{f}_d(\omega) = \frac{1}{s} \sum_{k=-\infty}^{+\infty} \hat{f}\left(\omega - \frac{2k\pi}{s}\right). \quad (3.3)$$

**Proof.** Since  $\delta(t - ns)$  is zero outside  $t = ns$ ,

$$f(ns) \delta(t - ns) = f(t) \delta(t - ns),$$

we can rewrite (3.1) as multiplication with a Dirac comb:

$$f_d(t) = f(t) \sum_{n=-\infty}^{+\infty} \delta(t - ns) = f(t) c(t). \quad (3.4)$$

Computing the Fourier transform yields

$$\hat{f}_d(\omega) = \frac{1}{2\pi} \hat{f} \star \hat{c}(\omega). \quad (3.5)$$

The Poisson formula (2.4) proves that

$$\hat{c}(\omega) = \frac{2\pi}{s} \sum_{k=-\infty}^{+\infty} \delta\left(\omega - \frac{2\pi k}{s}\right). \quad (3.6)$$

Since  $\hat{f} \star \delta(\omega - \xi) = \hat{f}(\omega - \xi)$ , inserting (3.6) into (3.5) proves (3.3). ■

Theorem 3.1 proves that sampling  $f$  at interval  $s$  is equivalent to making its Fourier transform  $2\pi/s$  periodic by summing all its translations  $\hat{f}(\omega - 2k\pi/s)$ . The resulting sampling theorem was first proved by Whittaker [482] in 1935 in

a book on interpolation theory. Shannon rediscovered it in 1949 for applications to communication theory [429].

**Theorem 3.2:** *Shannon, Whittaker.* If the support of  $\hat{f}$  is included in  $[-\pi/s, \pi/s]$ , then

$$f(t) = \sum_{n=-\infty}^{+\infty} f(ns) \phi_s(t - ns), \quad (3.7)$$

with

$$\phi_s(t) = \frac{\sin(\pi t/s)}{\pi t/s}. \quad (3.8)$$

**Proof.** If  $n \neq 0$ , the support of  $\hat{f}(\omega - n\pi/s)$  does not intersect the support of  $\hat{f}(\omega)$  because  $\hat{f}(\omega) = 0$  for  $|\omega| > \pi/s$ ; so (3.3) implies

$$\hat{f}_d(\omega) = \frac{\hat{f}(\omega)}{s} \text{ if } |\omega| \leq \frac{\pi}{s}. \quad (3.9)$$

The Fourier transform of  $\phi_s$  is  $\hat{\phi}_s = s \mathbf{1}_{[-\pi/s, \pi/s]}$ . Since the support of  $\hat{f}$  is in  $[-\pi/s, \pi/s]$ , it results from (3.9) that  $\hat{f}(\omega) = \hat{\phi}_s(\omega) \hat{f}_d(\omega)$ . The inverse Fourier transform of this equality gives

$$\begin{aligned} f(t) &= \phi_s \star f_d(t) = \phi_s \star \sum_{n=-\infty}^{+\infty} f(ns) \delta(t - ns) \\ &= \sum_{n=-\infty}^{+\infty} f(ns) \phi_s(t - ns). \end{aligned} \quad \blacksquare$$

The sampling theorem supposes that the support of  $\hat{f}$  is included in  $[-\pi/s, \pi/s]$ , which guarantees that  $f$  has no brutal variations between consecutive samples; thus, it can be recovered with a smooth interpolation. Section 3.1.3 shows that one can impose other smoothness conditions to recover  $f$  from its samples. Figure 3.1 illustrates the different steps for a sampling and reconstruction from samples, in both the time and Fourier domains.

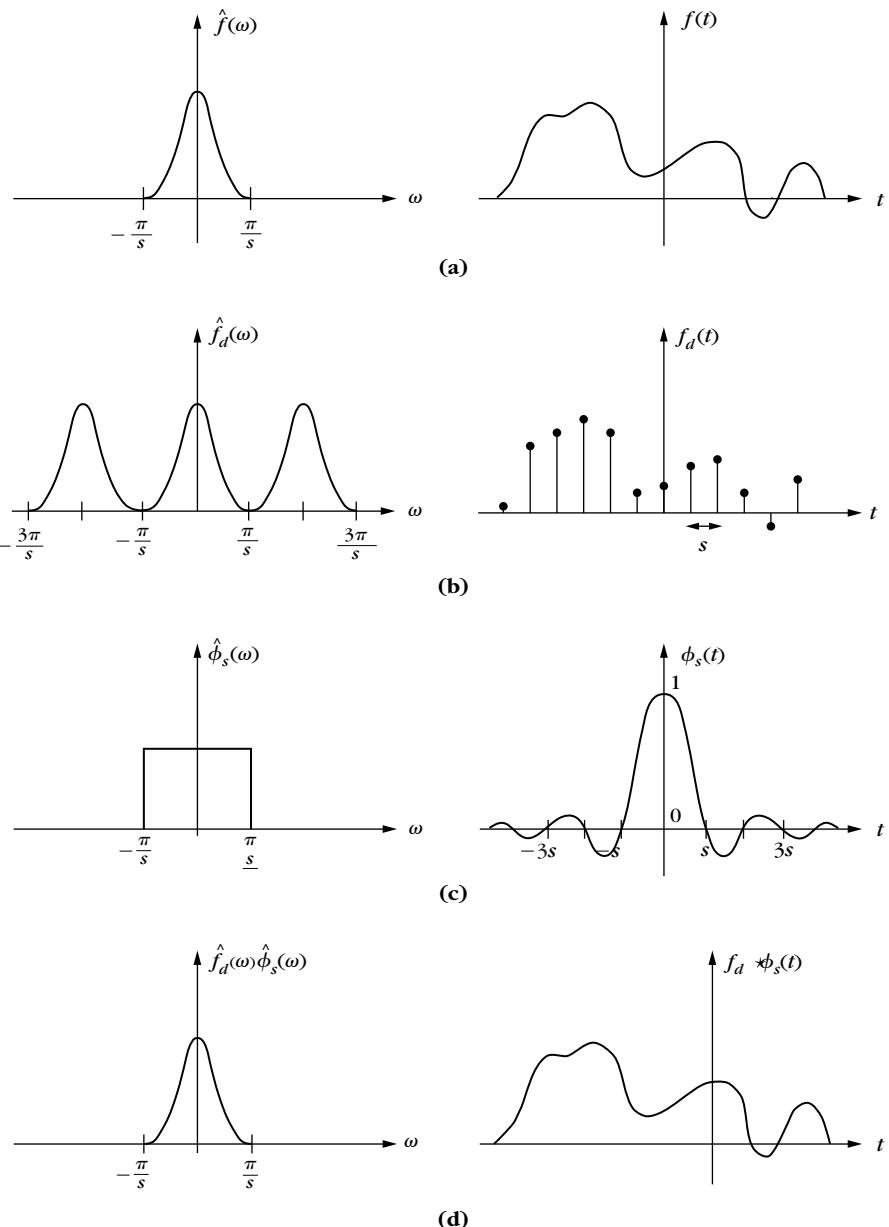
### 3.1.2 Aliasing

The sampling interval  $s$  is often imposed by computation or storage constraints and support of  $\hat{f}$  is generally not included in  $[-\pi/s, \pi/s]$ . In this case the interpolation formula (3.7) does not recover  $f$ . We analyze the resulting error and a filtering procedure to reduce it.

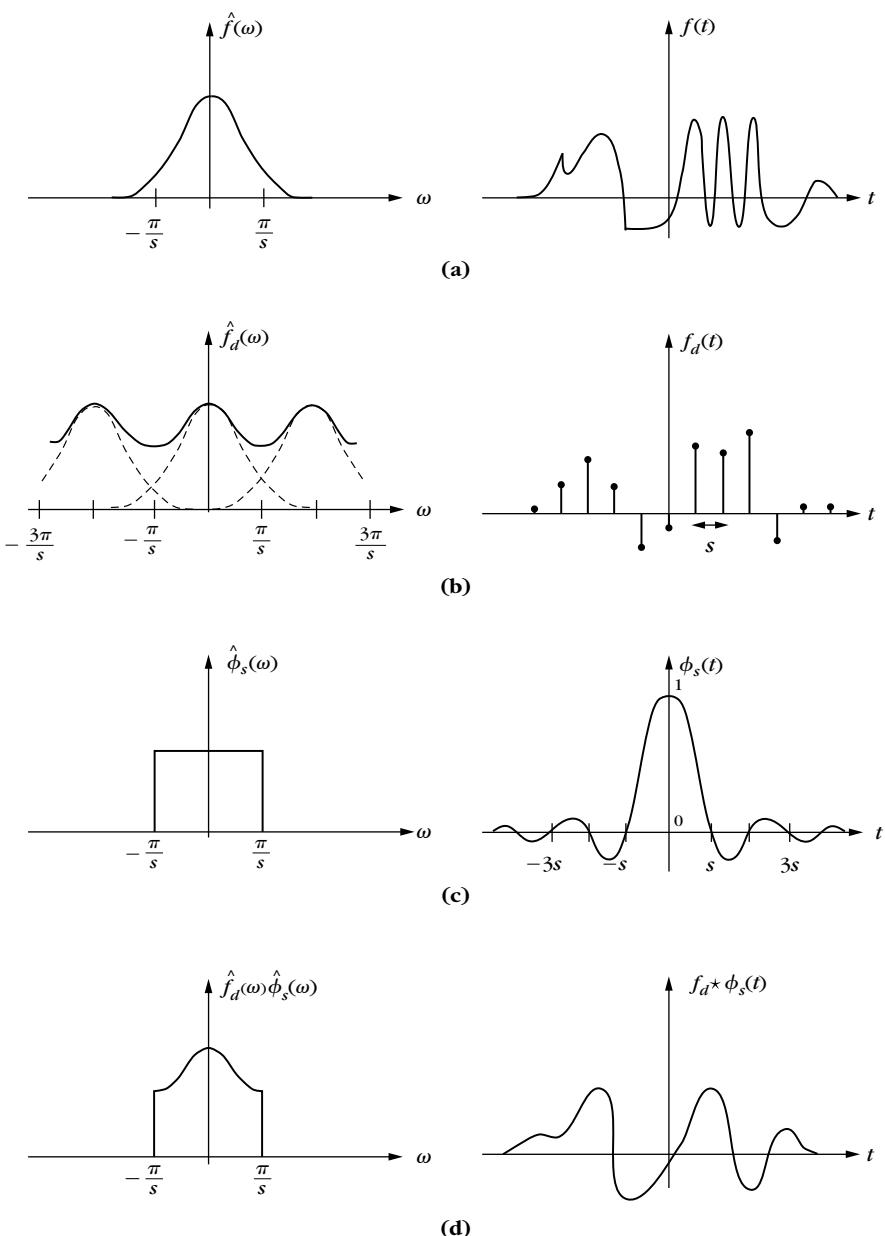
Theorem 3.1 proves that

$$\hat{f}_d(\omega) = \frac{1}{s} \sum_{k=-\infty}^{+\infty} \hat{f}\left(\omega - \frac{2k\pi}{s}\right). \quad (3.10)$$

Suppose that support of  $\hat{f}$  goes beyond  $[-\pi/s, \pi/s]$ . In general, support of  $\hat{f}(\omega - 2k\pi/s)$  intersects  $[-\pi/s, \pi/s]$  for several  $k \neq 0$ , as shown in Figure 3.2. This folding

**FIGURE 3.1**

**(a)** Signal  $f$  and its Fourier transform  $\hat{f}$ . **(b)** A uniform sampling of  $f$  makes its Fourier transform periodic. **(c)** Ideal low-pass filter. **(d)** The filtering of (b) with (c) recovers  $f$ .

**FIGURE 3.2**

**(a)** Signal  $f$  and its Fourier transform  $\hat{f}$ . **(b)** Aliasing produced by an overlapping of  $\hat{f}(\omega - 2k\pi/s)$  for different  $k$ , shown with dashed lines. **(c)** Ideal low-pass filter. **(d)** The filtering of (b) with (c) creates a low-frequency signal that is different from  $f$ .

of high-frequency components over a low-frequency interval is called *aliasing*. In the presence of aliasing, the interpolated signal

$$\phi_s \star f_d(t) = \sum_{n=-\infty}^{+\infty} f(ns) \phi_s(t - ns)$$

has a Fourier transform

$$\hat{f}_d(\omega) \hat{\phi}_s(\omega) = s \hat{f}_d(\omega) \mathbf{1}_{[-\pi/s, \pi/s]}(\omega) = \mathbf{1}_{[-\pi/s, \pi/s]}(\omega) \sum_{k=-\infty}^{+\infty} \hat{f}\left(\omega - \frac{2k\pi}{s}\right), \quad (3.11)$$

which may be completely different from  $\hat{f}(\omega)$  over  $[-\pi/s, \pi/s]$ . The signal  $\phi_s \star f_d$  may not even be a good approximation of  $f$ , as shown by Figure 3.2.

### EXAMPLE 3.1

Let us consider a high-frequency oscillation

$$f(t) = \cos(\omega_0 t) = \frac{e^{i\omega_0 t} + e^{-i\omega_0 t}}{2}.$$

Its Fourier transform is

$$\hat{f}(\omega) = \pi \left( \delta(\omega - \omega_0) + \delta(\omega + \omega_0) \right).$$

If  $2\pi/s > \omega_0 > \pi/s$ , then (3.11) yields

$$\begin{aligned} \hat{f}_d(\omega) \hat{\phi}_s(\omega) \\ = \pi \mathbf{1}_{[-\pi/s, \pi/s]}(\omega) \sum_{k=-\infty}^{+\infty} \left( \delta\left(\omega - \omega_0 - \frac{2k\pi}{s}\right) + \delta\left(\omega + \omega_0 - \frac{2k\pi}{s}\right) \right) \\ = \pi \left( \delta\left(\omega - \frac{2\pi}{s} + \omega_0\right) + \delta\left(\omega + \frac{2\pi}{s} - \omega_0\right) \right), \end{aligned}$$

so

$$f_d \star \phi_s(t) = \cos \left[ \left( \frac{2\pi}{s} - \omega_0 \right) t \right].$$

The aliasing reduces the high-frequency  $\omega_0$  to a lower frequency  $2\pi/s - \omega_0 \in [-\pi/s, \pi/s]$ . The same frequency folding is observed in a film that samples a fast-moving object without enough images per second. A wheel turning rapidly appears as though turning much more slowly in the film.

### Removal of Aliasing

To apply the sampling theorem,  $f$  is approximated by the closest signal  $\tilde{f}$ , the Fourier transform of which has a support in  $[-\pi/s, \pi/s]$ . The Plancherel formula (2.26) proves that

$$\begin{aligned} \|f - \tilde{f}\|^2 &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} |\hat{f}(\omega) - \hat{\tilde{f}}(\omega)|^2 d\omega \\ &= \frac{1}{2\pi} \int_{|\omega| > \pi/s} |\hat{f}(\omega)|^2 d\omega + \frac{1}{2\pi} \int_{|\omega| \leq \pi/s} |\hat{f}(\omega) - \hat{\tilde{f}}(\omega)|^2 d\omega. \end{aligned}$$

This distance is minimum when the second integral is zero and therefore

$$\widehat{\tilde{f}}(\omega) = \widehat{f}(\omega) \mathbf{1}_{[-\pi/s, \pi/s]}(\omega) = \frac{1}{s} \widehat{\phi}_s(\omega) \widehat{f}(\omega). \quad (3.12)$$

It corresponds to  $\tilde{f} = \frac{1}{s} f \star \phi_s$ .

The filtering of  $f$  by  $\phi_s$  avoids aliasing by removing any frequency larger than  $\pi/s$ . Since  $\widehat{\tilde{f}}$  has a support in  $[-\pi/s, \pi/s]$ , the sampling theorem proves that  $\tilde{f}(t)$  can be recovered from the samples  $\tilde{f}(ns)$ . An analog-to-digital converter is therefore composed of a filter that limits the frequency band to  $[-\pi/s, \pi/s]$ , followed by a uniform sampling at interval  $s$ .

### 3.1.3 General Sampling and Linear Analog Conversions

The Shannon-Whittaker theorem is a particular example of linear discrete-to-analog conversion, which does not apply to all digital acquisition devices. This section describes general analog-to-discrete conversion and reverse discrete-to-analog conversion, with general linear filtering and uniform sampling. Analog signals are approximated by linear projections on approximation spaces.

#### Sampling Theorems

We want to recover a stable approximation of  $f \in \mathbf{L}^2(\mathbb{R})$  from a filtering and uniform sampling, which outputs  $\{f \star \bar{\phi}_s(ns)\}_{n \in \mathbb{Z}}$ , for some real filter  $\bar{\phi}_s(t)$ . These samples can be written as inner products in  $\mathbf{L}^2(\mathbb{R})$ :

$$f \star \phi_s(ns) = \int_{-\infty}^{+\infty} f(t) \bar{\phi}_s(ns - t) dt = \langle f(t), \phi_s(t - ns) \rangle, \quad (3.13)$$

with  $\phi_s(t) = \bar{\phi}_s(-t)$ . Let  $\mathbf{U}_s$  be the approximation space generated by linear combination of the  $\{\phi_s(t - ns)\}_{n \in \mathbb{Z}}$ . The approximation  $\tilde{f} \in \mathbf{U}_s$ , which minimizes the maximum possible error  $\|f - \tilde{f}\|$ , is the orthogonal projection of  $f$  on  $\mathbf{U}_s$  (Exercice 3.5). The calculation of this orthogonal projection is stable if  $\{\phi_s(t - ns)\}_{n \in \mathbb{Z}}$  is a Riesz basis of  $\mathbf{U}_s$ , as defined in Section 5.1.1.

Following Definition 5.1, a Riesz basis is a family of linearly independent functions that yields an inner product satisfying an energy equivalence. There exists  $B \geq A > 0$  such that for any  $f \in \mathbf{U}_s$

$$A \|f\|^2 \leq \sum_{n=-\infty}^{+\infty} |\langle f(t), \phi_s(t - ns) \rangle|^2 \leq B \|f\|^2. \quad (3.14)$$

The basis is orthogonal if and only if  $A = B$ . The following generalized sampling theorem computes the orthogonal projection on the approximation space  $\mathbf{U}_s$  [468].

**Theorem 3.3:** *Linear sampling.* Let  $\{\phi_s(t - ns)\}_{n \in \mathbb{Z}}$  be a Riesz basis of  $\mathbf{U}_s$  and  $\bar{\phi}_s(t) = \phi_s(-t)$ . There exists a biorthogonal basis  $\{\tilde{\phi}_s(t - ns)\}_{n \in \mathbb{Z}}$  of  $\mathbf{U}_s$  such that

$$\forall f \in \mathbf{L}^2(\mathbb{R}), \quad P_{\mathbf{U}_s} f(t) = \sum_{n=-\infty}^{+\infty} f \star \bar{\phi}_s(ns) \tilde{\phi}_s(t - ns). \quad (3.15)$$

**Proof.** For any Riesz basis, Section 5.1.2 proves that a biorthogonal basis  $\{\tilde{\phi}_{s,n}(t)\}_{n \in \mathbb{Z}}$  exists that satisfies the biorthogonality relations

$$\forall (n, m) \in \mathbb{Z}^2, \quad \langle \phi_s(t - ns), \tilde{\phi}_{s,m}(t - ms) \rangle = \delta[n - m]. \quad (3.16)$$

Since  $\langle \phi_s(t - (n - m)s), \tilde{\phi}_{s,0}(t) \rangle = \langle \phi_s(t - ns), \tilde{\phi}_{s,0}(t - ms) \rangle = 0$  and since the dual basis is unique, necessarily  $\tilde{\phi}_{s,m}(t) = \tilde{\phi}_{s,0}(t - ms)$ . Section 5.1.2 proves in (5.20) that the orthogonal projection in  $\mathbf{U}_s$  can be written

$$P_{\mathbf{U}_s} f(t) = \sum_{n=-\infty}^{+\infty} \langle f(t), \phi_s(t - ns) \rangle \tilde{\phi}_s(t - ns)$$

which proves (3.15). ■

The orthogonal projection (3.15) can be rewritten as an analog filtering of the discrete signal  $f_d(t) = \sum_{n=-\infty}^{+\infty} f \star \tilde{\phi}_s(ns) \delta(t - ns)$ :

$$P_{\mathbf{U}_s} f(t) = f_d \star \tilde{\phi}_s(t). \quad (3.17)$$

If  $f \in \mathbf{U}_s$ , then  $P_{\mathbf{U}_s} f = f$  so it is exactly reconstructed by filtering the uniformly sampled discrete signal  $\{f \star \tilde{\phi}_s(ns)\}_{n \in \mathbb{Z}}$  with the analog filter  $\tilde{\phi}_s(t)$ . If  $f \notin \mathbf{U}_s$ , then (3.17) recovers the best linear approximation of  $f$  in  $\mathbf{U}_s$ . Section 9.1 shows that the linear approximation error  $\|f - P_{\mathbf{U}_s} f\|$  essentially depends on the uniform regularity of  $f$ . Given some prior information on  $f$ , optimizing the analog discretization filter  $\phi_s$  amounts to optimizing the approximation space  $\mathbf{U}_s$  to minimize this error. The following theorem characterizes filters  $\phi_s$  that generate a Riesz basis and computes the dual filter.

**Theorem 3.4.** A filter  $\phi_s$  generates a Riesz basis  $\{\phi_s(t - ns)\}_{n \in \mathbb{Z}}$  of a space  $\mathbf{U}_s$  if and only if there exists  $B \geq A > 0$  such that

$$\forall \omega \in [0, 2\pi/s], \quad A \leq \frac{1}{s} \sum_{k=-\infty}^{+\infty} |\hat{\phi}_s(\omega - \frac{2k\pi}{s})|^2 \leq B. \quad (3.18)$$

The biorthogonal basis  $\{\tilde{\phi}_s(t - ns)\}_{n \in \mathbb{Z}}$  is defined by the dual filter  $\tilde{\phi}_s$ , which satisfies:

$$\hat{\tilde{\phi}}_s(\omega) = \frac{s \hat{\phi}_s^*(\omega)}{\sum_{k=-\infty}^{+\infty} |\hat{\phi}_s(\omega - 2k\pi/s)|^2}. \quad (3.19)$$

**Proof.** Theorem 5.5 proves that  $\{\phi_s(t - ns)\}_{n \in \mathbb{Z}}$  is a Riesz basis of  $\mathbf{U}_s$  with Riesz bounds  $B \geq A > 0$  if and only if it is linearly independent and

$$\forall a \in \ell^2(\mathbb{Z}), \quad A \|a\|^2 \leq \left\| \sum_{n \in \mathbb{Z}} a[ns] \phi_s(t - ns) \right\|^2 \leq B \|a\|^2, \quad (3.20)$$

with  $\|a\|^2 = \sum_{n \in \mathbb{Z}} |a[ns]|^2$ .

Let us first write these conditions in the Fourier domain. The Fourier transform of  $f(t) = \sum_{n=-\infty}^{+\infty} a[ns] \phi_s(t - ns)$  is

$$\hat{f}(\omega) = \sum_{n=-\infty}^{+\infty} a[ns] e^{-ins\omega} \hat{\phi}_s(\omega) = \hat{a}(\omega) \hat{\phi}_s(\omega), \quad (3.21)$$

where  $\hat{a}(\omega)$  is the Fourier series  $\hat{a}(\omega) = \sum_{n=-\infty}^{+\infty} a[ns] e^{-ins\omega}$ . Let us relate the norm of  $f$  and  $\hat{a}$ . Since  $\hat{a}(\omega)$  is  $2\pi/s$  periodic, inserting (3.21) in the Plancherel formula (2.26) gives

$$\begin{aligned}\|f\|^2 &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} |\hat{f}(\omega)|^2 d\omega \\ &= \frac{1}{2\pi} \int_0^{2\pi/s} \sum_{k=-\infty}^{+\infty} |\hat{a}(\omega + 2k\pi/s)|^2 |\hat{\phi}_s(\omega + 2k\pi/s)|^2 d\omega \\ &= \frac{1}{2\pi} \int_0^{2\pi/s} |\hat{a}(\omega)|^2 \sum_{k=-\infty}^{+\infty} |\hat{\phi}_s(\omega + 2k\pi/s)|^2 d\omega.\end{aligned}\quad (3.22)$$

Section 3.2.2 on Fourier series proves that

$$\|a\|^2 = \sum_{n=-\infty}^{+\infty} |a[ns]|^2 = \frac{s}{2\pi} \int_0^{2\pi/s} |\hat{a}(\omega)|^2 d\omega. \quad (3.23)$$

As a consequence of (3.22) and (3.23), the Riesz bound inequalities (3.20) are equivalent to

$$\forall \hat{a} \in \mathbf{L}^2[0, 2\pi/s], \quad \frac{1}{2\pi} \int_0^{2\pi/s} |\hat{a}(\omega)|^2 \sum_{k=-\infty}^{+\infty} |\hat{\phi}_s(\omega + 2k\pi/s)|^2 d\omega \leq \frac{Bs}{2\pi} \int_0^{2\pi/s} |\hat{a}(\omega)|^2 d\omega \quad (3.24)$$

and

$$\forall \hat{a} \in \mathbf{L}^2[0, 2\pi/s], \quad \frac{1}{2\pi} \int_0^{2\pi/s} |\hat{a}(\omega)|^2 \sum_{k=-\infty}^{+\infty} |\hat{\phi}_s(\omega + 2k\pi/s)|^2 d\omega \geq \frac{As}{2\pi} \int_0^{2\pi/s} |\hat{a}(\omega)|^2 d\omega. \quad (3.25)$$

If  $\hat{\phi}_s$  satisfies (3.18), then clearly (3.24) and (3.25) are valid, which proves (3.22).

Conversely, if  $\{\phi_s(ns - t)\}_{n \in \mathbb{Z}}$  is a Riesz basis. Suppose that either the upper or the lower bound of (3.18) is not satisfied for  $\omega$  in a set of nonzero measures. Let  $\hat{a}$  be the indicator function of this set. Then either (3.24) or (3.25) is not valid for this  $\hat{a}$ . This implies that the Riesz bounds (3.20) are not valid for  $a$  and therefore that it is not a Riesz basis, which contradicts our hypothesis. So (3.18) is indeed valid for almost all  $\omega \in [0, 2\pi/s]$ .

To compute the biorthogonal basis, we are looking for  $\tilde{\phi}_s \in \mathbf{U}_s$  such that  $\{\tilde{\phi}_s(ns - t)\}_{n \in \mathbb{Z}}$  satisfies the biorthogonal relations (3.16). Since  $\tilde{\phi}_s \in \mathbf{U}_s$  we saw in (3.21) that its Fourier transform can be written  $\hat{\tilde{\phi}}_s(\omega) = \hat{a}(\omega) \hat{\phi}_s(\omega)$ , where  $\hat{a}(\omega)$  is  $2\pi/s$  periodic. Let us define  $g(t) = \tilde{\phi}_s \star \tilde{\phi}_s(t)$ . Its Fourier transform is

$$\hat{g}(\omega) = \hat{\phi}_s^*(\omega) \hat{\tilde{\phi}}_s(\omega) = \hat{a}(\omega) |\hat{\phi}_s(\omega)|^2.$$

The biorthogonal relations (3.16) are satisfied if and only if  $g(ns) = 0$  if  $n \neq 0$  and  $g(0) = 1$ . It results that  $g_d(t) = \sum_{n=-\infty}^{+\infty} g(ns) \delta(t - ns) = \delta(t)$ . Theorem 3.1 derives in (3.3) that

$$\hat{g}_d(\omega) = \frac{1}{s} \sum_{k=-\infty}^{+\infty} \hat{g}(\omega - 2k\pi/s) = \frac{\hat{a}(\omega)}{s} \sum_{k=-\infty}^{+\infty} |\hat{\phi}_s(\omega - 2k\pi/s)|^2 = 1.$$

It results that

$$\hat{a}(\omega) = s \left( \sum_{k=-\infty}^{+\infty} |\hat{\phi}_s(\omega - 2k\pi/s)|^2 \right)^{-1},$$

which proves (3.19). ■

This theorem gives a necessary and sufficient condition on the low-pass filter  $\tilde{\phi}_s(t) = \phi_s(-t)$  to recover a stable signal approximation from uniform sampling at interval  $s$ . For various sampling interval  $s$ , the low-pass filter can be obtained by dilating a single filter  $\phi_s(t) = s^{-1/2}\phi(t/s)$  and thus  $\hat{\phi}_s(\omega) = s^{1/2}\hat{\phi}(s\omega)$ . The necessary and sufficient Riesz basis condition (3.18) is then satisfied if and only if

$$\forall \omega \in [-\pi, \pi], \quad A \leq \sum_{k=-\infty}^{+\infty} |\hat{\phi}(\omega - 2k\pi)|^2 \leq B. \quad (3.26)$$

It results from (3.19) that the dual filter satisfies  $\widehat{\tilde{\phi}}_s(\omega) = s^{1/2}\widehat{\tilde{\phi}}(s\omega)$  and therefore  $\tilde{\phi}_s(t) = s^{-1/2}\tilde{\phi}(t/s)$ . When  $A = B = 1$ , the Riesz basis is an orthonormal basis, which proves Corollary 3.1.

**Corollary 3.1.** The family  $\{\phi_s(t - ns)\}_{n \in \mathbb{Z}}$  is an orthonormal basis of the space  $\mathbf{U}_s$  it generates, with  $\phi_s(t) = s^{-1/2}\phi(t/s)$ , if and only if

$$\forall \omega \in [0, 2\pi], \quad \sum_{k=-\infty}^{+\infty} |\hat{\phi}(\omega - 2k\pi)|^2 = 1, \quad (3.27)$$

and the dual filter is  $\tilde{\phi}_s = \phi_s$ .

### Shannon-Whittaker Revisited

Shannon-Whittaker, Theorem 3.2, is defined with a sine-cardinal perfect low-pass filter  $\phi_s$ , which we renormalize here to have a unit norm. The following theorem proves that it samples functions on an orthonormal basis.

**Theorem 3.5.** If  $\phi_s(t) = s^{1/2} \sin(\pi s^{-1}t)/(\pi t)$  then  $\{\phi_s(t - ns)\}_{n \in \mathbb{Z}}$  is an orthonormal basis of the space  $\mathbf{U}_s$  of functions whose Fourier transforms have a support included in  $[-\pi/s, \pi/s]$ . If  $f \in \mathbf{U}_s$ , then

$$f(nT) = s^{-1/2} f \star \phi_s(ns). \quad (3.28)$$

**Proof.** The filter satisfies  $\phi_s(t) = s^{-1/2}\phi(t/s)$  with  $\phi(t) = \sin(\pi t)/(\pi t)$ . The Fourier transform  $\hat{\phi}(\omega) = \mathbf{1}_{[-\pi, \pi]}(\omega)$  satisfies the condition (3.27) of Corollary 3.1, which proves that  $\{\phi_s(t - ns)\}_{n \in \mathbb{Z}}$  is an orthonormal basis of a space  $\mathbf{U}_s$ .

Any  $f(t) = \sum_{n=-\infty}^{+\infty} a[n] \phi_s(t - ns) \in \mathbf{U}_s$  has a Fourier transform that can be written

$$\hat{f}(\omega) = \sum_{n=-\infty}^{+\infty} a[n] e^{-ins\omega} \hat{\phi}_s(\omega) = \hat{a}(\omega) s^{1/2} \mathbf{1}_{[-\pi/s, \pi/s]}, \quad (3.29)$$

which implies that  $f \in \mathbf{U}_s$  if and only if  $f$  has a Fourier transform supported in  $[-\pi/s, \pi/s]$ .

If  $f \in \mathbf{U}_s$ , then decomposing it on the orthonormal basis  $\{\phi_s(t - ns)\}_{n \in \mathbb{Z}}$  gives

$$f(t) = P_{\mathbf{U}_s} f(t) = \sum_{n \in \mathbb{Z}} \langle f(u), \phi_s(u - ns) \rangle \phi_s(t - ns).$$

Since  $\phi_s(ps) = s^{-1/2} \delta[ps]$  and  $\phi_s(-t) = \phi_s(t)$ , the result is that

$$f(ns) = s^{-1/2} \langle f(u), \phi_s(u - ns) \rangle = s^{-1/2} f \star \phi_s(ns). \quad \blacksquare$$

This theorem proves that in the particular case of the Shannon-Whittaker sampling theorem, if  $f \in \mathbf{U}_s$  then the sampled low-pass filtered values  $f \star \phi_s(ns)$  are proportional to the signal samples  $f(ns)$ . This comes from the fact that the sine-cardinal  $\phi(t) = \sin(\pi t/s)/(\pi t/s)$  satisfies the interpolation property  $\phi(ns) = \delta[ns]$ . A generalization of such multiscale interpolations is studied in Section 7.6.

Shannon-Whittaker sampling approximates signals by restricting their Fourier transform to a low-frequency interval. It is particularly effective for smooth signals with a Fourier transform that has energy concentrated at low frequencies. It is also adapted for sound recordings, which are sufficiently approximated by lower-frequency harmonics.

For discontinuous signals, such as images, a low-frequency restriction produces Gibbs oscillations, as described in Section 2.3.3. The image visual quality is degraded by these oscillations, which have a total variation (2.65) that is infinite. A piecewise constant approximation has the advantage of creating no such spurious oscillations.

### Block Sampler

A block sampler approximates signals with piecewise constant functions. The approximation space  $\mathbf{U}_s$  is the set of all functions that are constant on intervals  $[ns, (n+1)s]$ , for any  $n \in \mathbb{Z}$ . Let  $\phi_s(t) = s^{-1/2} \mathbf{1}_{[0,s]}(t)$ . The family  $\{\phi_s(t - ns)\}_{n \in \mathbb{Z}}$  is an orthonormal basis of  $\mathbf{U}_s$  (Exercise 3.1). If  $f \notin \mathbf{U}_s$ , then its orthogonal projection on  $\mathbf{U}_s$  is calculated with a partial decomposition in the block orthonormal basis of  $\mathbf{U}_s$

$$P_{\mathbf{U}_s} f(t) = \sum_{n=-\infty}^{+\infty} \langle f(u), \phi_s(u - ns) \rangle \phi_s(t - ns), \quad (3.30)$$

and each coefficient is proportional to the signal average on  $[ns, (n+1)s]$ :

$$\langle f(u), \phi_s(u - ns) \rangle = f \star \phi_s(ns) = s^{-1/2} \int_{ns}^{(n+1)s} f(u) du.$$

This block analog-to-digital conversion is particularly simple to implement in analog electronics, where integration is performed by a capacity.

In domains where  $f$  is a regular function, a piecewise constant approximation  $P_{\mathbf{U}_s} f$  is not very precise and can be significantly improved. More precise approximations are obtained with approximation spaces  $\mathbf{U}_s$  of higher-order polynomial splines. The resulting approximations can introduce small Gibbs oscillations, but these oscillations have a finite total variation.

### Spline Sampling

Block samplers are generalized by spline sampling with a space  $\mathbf{U}_s$  of spline functions that are  $m - 1$  times continuously differentiable and equal to a polynomial of degree  $m$  on any interval  $[ns, (n + 1)s]$ , for  $n \in \mathbb{Z}$ . When  $m = 1$ , functions in  $\mathbf{U}_s$  are piecewise linear and continuous.

A Riesz basis of polynomial splines is constructed with *box splines*. A box spline  $\phi$  of degree  $m$  is computed by convolving the box window  $\mathbf{1}_{[0,1]}$  with itself  $m + 1$  times and centering it at 0 or  $1/2$ . Its Fourier transform is

$$\hat{\phi}(\omega) = \left( \frac{\sin(\omega/2)}{\omega/2} \right)^{m+1} \exp\left(\frac{-i\varepsilon\omega}{2}\right). \quad (3.31)$$

If  $m$  is even, then  $\varepsilon = 1$  and  $\phi$  have a support centered at  $t = 1/2$ . If  $m$  is odd, then  $\varepsilon = 0$  and  $\phi(t)$  are symmetric about  $t = 0$ . One can verify that  $\hat{\phi}(\omega)$  satisfies the sampling condition (3.26) using a closed-form expression (7.20) of the resulting series. This means that for any  $s > 0$ , a box splines family  $\{\phi_s(t - ns)\}_{n \in \mathbb{Z}}$  defines a Riesz basis of  $\mathbf{U}_s$ , and thus is a stable sampling.

## 3.2 DISCRETE TIME-INVARIANT FILTERS

### 3.2.1 Impulse Response and Transfer Function

Classic discrete signal-processing algorithms most generally are based on time-invariant linear operators [51, 55]. The time invariance is limited to translations on the sampling grid. To simplify notation, the sampling interval is normalized  $s = 1$ , and we denote  $f[n]$  the sample values. A linear discrete operator  $L$  is time-invariant if an input  $f[n]$ , delayed by  $p \in \mathbb{Z}$ ,  $f_p[n] = f[n - p]$ , produces an output also delayed by  $p$ :

$$Lf_p[n] = Lf[n - p].$$

#### Impulse Response

We denote by  $\delta[n]$  the discrete Dirac

$$\delta[n] = \begin{cases} 1 & \text{if } n = 0 \\ 0 & \text{if } n \neq 0 \end{cases}. \quad (3.32)$$

Any signal  $f[n]$  can be decomposed as a sum of shifted Diracs:

$$f[n] = \sum_{p=-\infty}^{+\infty} f[p] \delta[n - p].$$

Let  $L\delta[n] = h[n]$  be the discrete *impulse response*. Linearity and time invariance implies that

$$Lf[n] = \sum_{p=-\infty}^{+\infty} f[p] h[n - p] = f \star h[n]. \quad (3.33)$$

A discrete linear time-invariant operator is thus computed with a discrete convolution. If  $h[n]$  has a finite support, the sum (3.33) is calculated with a finite number of operations. These are called *finite impulse response* (FIR) filters. Convolutions with infinite impulse response filters may also be calculated with a finite number of operations if they can be rewritten with a recursive equation (3.45).

### Causality and Stability

A discrete filter  $L$  is *causal* if  $Lf[p]$  depends only on the values of  $f[n]$  for  $n \leq p$ . The convolution formula (3.33) implies that  $h[n] = 0$  if  $n < 0$ .

The filter is *stable* if any bounded input signal  $f[n]$  produces a bounded output signal  $Lf[n]$ . Since

$$|Lf[n]| \leq \sup_{n \in \mathbb{Z}} |f[n]| \sum_{k=-\infty}^{+\infty} |h[k]|,$$

it is sufficient that  $\sum_{n=-\infty}^{+\infty} |h[n]| < +\infty$ , which means that  $h \in \ell^1(\mathbb{Z})$ . One can verify that this sufficient condition is also necessary. Thus, the filter is stable if and only if  $h \in \ell^1(\mathbb{Z})$  (Exercise 3.6).

### Transfer Function

The Fourier transform plays a fundamental role in analyzing discrete time-invariant operators because discrete sinusoidal waves  $e_{\omega}[n] = e^{i\omega n}$  are eigenvectors:

$$Le_{\omega}[n] = \sum_{p=-\infty}^{+\infty} e^{i\omega(n-p)} h[p] = e^{i\omega n} \sum_{p=-\infty}^{+\infty} h[p] e^{-i\omega p}. \quad (3.34)$$

The eigenvalue is a Fourier series

$$\hat{h}(\omega) = \sum_{p=-\infty}^{+\infty} h[p] e^{-i\omega p}. \quad (3.35)$$

It is the filter *transfer function*.

### EXAMPLE 3.2

The uniform discrete average

$$Lf[n] = \frac{1}{2N+1} \sum_{p=n-N}^{n+N} f[p]$$

is a time-invariant discrete filter that has an impulse response of  $h = (2N+1)^{-1} \mathbf{1}_{[-N,N]}$ . Its transfer function is

$$\hat{h}(\omega) = \frac{1}{2N+1} \sum_{n=-N}^{+N} e^{-in\omega} = \frac{1}{2N+1} \frac{\sin((N+1/2)\omega)}{\sin \omega/2}. \quad (3.36)$$

### 3.2.2 Fourier Series

The properties of Fourier series are essentially the same as the properties of the Fourier transform since Fourier series are particular instances of Fourier transforms for Dirac sums. If  $f(t) = \sum_{n=-\infty}^{+\infty} f[n] \delta(t-n)$ , then  $\hat{f}(\omega) = \sum_{n=-\infty}^{+\infty} f[n] e^{-i\omega n}$ .

For any  $n \in \mathbb{Z}$ ,  $e^{-i\omega n}$  has period  $2\pi$ , so Fourier series have period  $2\pi$ . An important issue to understand is whether all functions with period  $2\pi$  can be written as Fourier series. Such functions are characterized by their restriction to  $[-\pi, \pi]$ . We therefore consider functions  $\hat{a} \in L^2[-\pi, \pi]$  that are square integrable over  $[-\pi, \pi]$ . The space  $L^2[-\pi, \pi]$  is a Hilbert space with the inner product

$$\langle \hat{a}, \hat{b} \rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{a}(\omega) \hat{b}^*(\omega) d\omega \quad (3.37)$$

and the resulting norm

$$\|\hat{a}\|^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |\hat{a}(\omega)|^2 d\omega.$$

Theorem 3.6 proves that any function in  $L^2[-\pi, \pi]$  can be written as a Fourier series.

**Theorem 3.6.** The family of functions  $\{e^{-ik\omega}\}_{k \in \mathbb{Z}}$  is an orthonormal basis of  $L^2[-\pi, \pi]$ .

**Proof.** The orthogonality with respect to the inner product (3.37) is established with a direct integration. To prove that  $\{\exp(-ik\omega)\}_{k \in \mathbb{Z}}$  is a basis, we must show that linear expansions of these vectors are dense in  $L^2[-\pi, \pi]$ .

We first prove that any continuously differentiable function  $\hat{\phi}$  with a support included in  $[-\pi, \pi]$  satisfies

$$\hat{\phi}(\omega) = \sum_{k=-\infty}^{+\infty} \langle \hat{\phi}(\xi), \exp(-ik\xi) \rangle \exp(-ik\omega), \quad (3.38)$$

with a pointwise convergence for any  $\omega \in [-\pi, \pi]$ . Let us compute the partial sum

$$\begin{aligned} S_N(\omega) &= \sum_{k=-N}^N \langle \hat{\phi}(\xi), \exp(-ik\xi) \rangle \exp(-ik\omega) \\ &= \sum_{k=-N}^N \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{\phi}(\xi) \exp(ik\xi) d\xi \exp(-ik\omega) \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{\phi}(\xi) \sum_{k=-N}^N \exp[ik(\xi - \omega)] d\xi. \end{aligned}$$

The Poisson formula (2.37) proves the distribution equality

$$\lim_{N \rightarrow +\infty} \sum_{k=-N}^N \exp[ik(\xi - \omega)] = 2\pi \sum_{k=-\infty}^{+\infty} \delta(\xi - \omega - 2\pi k),$$

and since the support of  $\hat{\phi}$  is in  $[-\pi, \pi]$ , we get

$$\lim_{N \rightarrow +\infty} S_N(\omega) = \hat{\phi}(\omega).$$

Since  $\hat{\phi}$  is continuously differentiable, following the steps (2.38–2.40) in the proof of the Poisson formula shows that  $S_N(\omega)$  converges uniformly to  $\hat{\phi}(\omega)$  on  $[-\pi, \pi]$ .

To prove that linear expansions of sinusoidal waves  $\{\exp(-ik\omega)\}_{k \in \mathbb{Z}}$  are dense in  $\mathbf{L}^2[-\pi, \pi]$ , let us verify that the distance between  $\hat{a} \in \mathbf{L}^2[-\pi, \pi]$  and such a linear expansion is less than  $\varepsilon$ , for any  $\varepsilon > 0$ . Continuously differentiable functions with a support included in  $[-\pi, \pi]$  are dense in  $\mathbf{L}^2[-\pi, \pi]$ ; thus, there exists  $\hat{\phi}$  such that  $\|\hat{a} - \hat{\phi}\| \leq \varepsilon/2$ . The uniform pointwise convergence proves that there exists  $N$  for which

$$\sup_{\omega \in [-\pi, \pi]} |S_N(\omega) - \hat{\phi}(\omega)| \leq \frac{\varepsilon}{2},$$

which implies that

$$\|S_N - \hat{\phi}\|^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |S_N(\omega) - \hat{\phi}(\omega)|^2 d\omega \leq \frac{\varepsilon^2}{4}.$$

It follows that  $\hat{a}$  is approximated by the Fourier series  $S_N$  with an error

$$\|\hat{a} - S_N\| \leq \|\hat{a} - \hat{\phi}\| + \|\hat{\phi} - S_N\| \leq \varepsilon. \quad \blacksquare$$

Theorem 3.6 proves that if  $f \in \ell^2(\mathbb{Z})$ , the Fourier series

$$\hat{f}(\omega) = \sum_{n=-\infty}^{+\infty} f[n] e^{-i\omega n} \quad (3.39)$$

can be interpreted as the decomposition of  $\hat{f}$  in an orthonormal basis of  $\mathbf{L}^2[-\pi, \pi]$ . The Fourier series coefficients can thus be written as inner products in  $\mathbf{L}^2[-\pi, \pi]$ :

$$f[n] = \langle \hat{f}(\omega), e^{-i\omega n} \rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{f}(\omega) e^{i\omega n} d\omega. \quad (3.40)$$

The energy conservation of orthonormal bases (A.10) yields a Plancherel identity:

$$\sum_{n=-\infty}^{+\infty} |f[n]|^2 = \|\hat{f}\|^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |\hat{f}(\omega)|^2 d\omega. \quad (3.41)$$

### Pointwise Convergence

The equality (3.39) is meant in the sense of mean-square convergence

$$\lim_{N \rightarrow +\infty} \left\| \hat{f}(\omega) - \sum_{k=-N}^N f[k] e^{-i\omega k} \right\| = 0.$$

It does not imply a pointwise convergence at all  $\omega \in \mathbb{R}$ .

In 1873, Dubois-Reymond constructed a periodic function  $\hat{f}(\omega)$  that is continuous and has a Fourier series that diverges at some point. On the other hand, if  $\hat{f}(\omega)$  is continuously differentiable, then the proof of Theorem 3.6 shows that its Fourier series converges uniformly to  $\hat{f}(\omega)$  on  $[-\pi, \pi]$ . It was only in 1966 that Carleson [149] was able to prove that if  $\hat{f} \in L^2[-\pi, \pi]$  then its Fourier series converges almost everywhere. The proof is very technical.

### Convolutions

Since  $\{e^{-i\omega k}\}_{k \in \mathbb{Z}}$  are eigenvectors of discrete convolution operators, we also have a discrete convolution theorem.

**Theorem 3.7.** If  $f \in \ell^1(\mathbb{Z})$  and  $h \in \ell^1(\mathbb{Z})$ , then  $g = f \star h \in \ell^1(\mathbb{Z})$  and

$$\hat{g}(\omega) = \hat{f}(\omega) \hat{h}(\omega). \quad (3.42)$$

The proof is identical to the proof of the convolution, Theorem 2.2, if we replace integrals by discrete sums. The reconstruction formula (3.40) shows that a filtered signal can be written

$$f \star h[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{h}(\omega) \hat{f}(\omega) e^{i\omega n} d\omega. \quad (3.43)$$

The transfer function  $\hat{h}(\omega)$  amplifies or attenuates the frequency components  $\hat{f}(\omega)$  of  $f[n]$ .

### EXAMPLE 3.3

An ideal discrete low-pass filter has a  $2\pi$  periodic transfer function that is defined by  $\hat{h}(\omega) = \mathbf{1}_{[-\xi, \xi]}(\omega)$ , for  $\omega \in [-\pi, \pi]$  and  $0 < \xi < \pi$ . Its impulse response is computed with (3.40):

$$h[n] = \frac{1}{2\pi} \int_{-\xi}^{\xi} e^{i\omega n} d\omega = \frac{\sin \xi n}{\pi n}. \quad (3.44)$$

It is a uniform sampling of the ideal analog low-pass filter (2.29).

### EXAMPLE 3.4

A recursive filter computes  $g = Lf$ , which is a solution of a recursive equation

$$\sum_{k=0}^K a_k f[n-k] = \sum_{k=0}^M b_k g[n-k], \quad (3.45)$$

with  $b_0 \neq 0$ . If  $g[n] = 0$  and  $f[n] = 0$  for  $n < 0$ , then  $g$  has a linear and time-invariant dependency on  $f$  and thus can be written  $g = f \star h$ . The transfer function is obtained by computing the Fourier transform of (3.45). The Fourier transform of  $f_k[n] = f[n-k]$  is

$\hat{f}_k(\omega) = \hat{f}(\omega) e^{-ik\omega}$ , so

$$\hat{h}(\omega) = \frac{\hat{g}(\omega)}{\hat{f}(\omega)} = \frac{\sum_{k=0}^K a_k e^{-ik\omega}}{\sum_{k=0}^M b_k e^{-ik\omega}}.$$

It is a rational function of  $e^{-i\omega}$ . If  $b_k \neq 0$  for some  $k > 0$ , then one can verify that the impulse response  $h$  has an infinite support. The stability of such filters is studied in Exercise 3.18. A direct calculation of the convolution sum  $g[n] = f \star h[n]$  would require an infinite number of operations, whereas (3.45) computes  $g[n]$  with  $K + M$  additions and multiplications from its past values.

---

### Window Multiplication

An infinite impulse response filter  $h$ , such as the ideal low-pass filter (3.44), may be approximated by a finite response filter  $\tilde{h}$  by multiplying  $h$  with a window  $g$  of finite support:

$$\tilde{h}[n] = g[n] h[n].$$

One can verify (Exercise 3.6) that a multiplication in time is equivalent to a convolution in the frequency domain:

$$\widehat{\tilde{h}}(\omega) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{h}(\xi) \hat{g}(\omega - \xi) d\xi = \frac{1}{2\pi} \hat{h} \star \hat{g}(\omega). \quad (3.46)$$

Clearly  $\widehat{\tilde{h}} = \hat{h}$  only if  $\hat{g} = 2\pi\delta$ , which would imply that  $g$  has an infinite support and  $g[n] = 1$ . The approximation  $\tilde{h}$  is close to  $\hat{h}$  only if  $\hat{g}$  approximates a Dirac, which means that all its energy is concentrated at low frequencies. In time,  $g$  should therefore have smooth variations.

The rectangular window  $g = \mathbf{1}_{[-N,N]}$  has a Fourier transform  $\hat{g}$  computed in (3.36). It has important side lobes far away from  $\omega = 0$ . The resulting  $\tilde{h}$  is a poor approximation of  $\hat{h}$ . The Hanning window

$$g[n] = \cos^2\left(\frac{\pi n}{2N}\right) \mathbf{1}_{[-N,N]}[n]$$

is smoother and thus has a Fourier transform better concentrated at low frequencies. The spectral properties of other windows are studied in Section 4.2.2.

---

## 3.3 FINITE SIGNALS

Up to now we have considered discrete signals  $f[n]$  defined for all  $n \in \mathbb{Z}$ . In practice,  $f[n]$  is known over a finite domain, say  $0 \leq n < N$ . Convolutions therefore must be modified to take into account the border effects at  $n = 0$  and  $n = N - 1$ . The Fourier transform also must be redefined over finite sequences for numerical computations. The fast Fourier transform algorithm is explained as well as its application to fast convolutions.

### 3.3.1 Circular Convolutions

Let  $\tilde{f}$  and  $\tilde{h}$  be signals of  $N$  samples. To compute the convolution product

$$\tilde{f} \star \tilde{h}[n] = \sum_{p=-\infty}^{+\infty} \tilde{f}[p] \tilde{h}[n-p] \quad \text{for } 0 \leq n < N,$$

we must know  $\tilde{f}[n]$  and  $\tilde{h}[n]$  beyond  $0 \leq n < N$ . One approach is to extend  $\tilde{f}$  and  $\tilde{h}$  with a periodization over  $N$  samples, and to define

$$f[n] = \tilde{f}[n \bmod N], \quad h[n] = \tilde{h}[n \bmod N].$$

The *circular convolution* of two such signals, both with period  $N$ , is defined as a sum over their period:

$$f \circledast h[n] = \sum_{p=0}^{N-1} f[p] h[n-p] = \sum_{p=0}^{N-1} f[n-p] h[p].$$

It is also a signal of period  $N$ .

The eigenvectors of a circular convolution operator

$$Lf[n] = f \circledast h[n]$$

are the discrete complex exponentials  $e_k[n] = \exp(i2\pi kn/N)$ . Indeed,

$$Le_k[n] = \exp\left(\frac{i2\pi kn}{N}\right) \sum_{p=0}^{N-1} h[p] \exp\left(\frac{-i2\pi kp}{N}\right),$$

and the eigenvalue is the discrete Fourier transform of  $h$ :

$$\hat{h}[k] = \sum_{p=0}^{N-1} h[p] \exp\left(\frac{-i2\pi kp}{N}\right).$$

### 3.3.2 Discrete Fourier Transform

The space of signals of period  $N$  is an Euclidean space of dimension  $N$  and the inner product of two such signals  $f$  and  $g$  is

$$\langle f, g \rangle = \sum_{n=0}^{N-1} f[n] g^*[n]. \tag{3.47}$$

Theorem 3.8 proves that any signal with period  $N$  can be decomposed as a sum of discrete sinusoidal waves.

**Theorem 3.8.** The family

$$\left\{ e_k[n] = \exp\left(\frac{i2\pi kn}{N}\right) \right\}_{0 \leq k < N}$$

is an orthogonal basis of the space of signals of period  $N$ .

Since the space is of dimension  $N$ , any orthogonal family of  $N$  vectors is an orthogonal basis. To prove this theorem, it is sufficient to verify that  $\{e_k\}_{0 \leq k < N}$  is orthogonal with respect to the inner product (3.47) (Exercise 3.8). Any signal  $f$  of period  $N$  can be decomposed on this basis:

$$f = \sum_{k=0}^{N-1} \frac{\langle f, e_k \rangle}{\|e_k\|^2} e_k. \quad (3.48)$$

By definition, the *discrete Fourier transform* (DFT) of  $f$  is

$$\hat{f}[k] = \langle f, e_k \rangle = \sum_{n=0}^{N-1} f[n] \exp\left(\frac{-i2\pi kn}{N}\right). \quad (3.49)$$

Since  $\|e_k\|^2 = N$ , (3.48) gives an inverse discrete Fourier formula:

$$f[n] = \frac{1}{N} \sum_{k=0}^{N-1} \hat{f}[k] \exp\left(\frac{i2\pi kn}{N}\right). \quad (3.50)$$

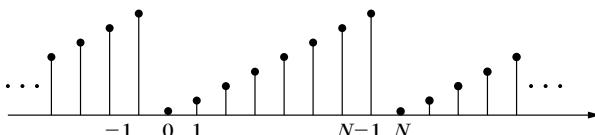
The orthogonality of the basis also implies a Plancherel formula:

$$\|f\|^2 = \sum_{n=0}^{N-1} |f[n]|^2 = \frac{1}{N} \sum_{k=0}^{N-1} |\hat{f}[k]|^2. \quad (3.51)$$

The discrete Fourier transform (DFT) of a signal  $f$  of period  $N$  is computed from its values for  $0 \leq n < N$ . Then why is it important to consider it a periodic signal with period  $N$  rather than a finite signal of  $N$  samples? The answer lies in the interpretation of the Fourier coefficients. The discrete Fourier sum (3.50) defines a signal of period  $N$  for which the samples  $f[0]$  and  $f[N - 1]$  are side by side. If  $f[0]$  and  $f[N - 1]$  are very different, this produces a brutal transition in the periodic signal, creating relatively high-amplitude Fourier coefficients at high frequencies. For example, Figure 3.3 shows that the “smooth” ramp  $f[n] = n$  for  $0 \leq n < N$  has sharp transitions at  $n = 0$  and  $n = N$  once made periodic.

### Circular Convolutions

Since  $\{\exp(i2\pi kn/N)\}_{0 \leq k < N}$  are eigenvectors of circular convolutions, we derive a convolution theorem.



**FIGURE 3.3**

Signal  $f[n] = n$  for  $0 \leq n < N$  made periodic over  $N$  samples.

**Theorem 3.9.** If  $f$  and  $h$  have period  $N$ , then the discrete Fourier transform of  $g = f \circledast h$  is

$$\hat{g}[k] = \hat{f}[k] \hat{h}[k]. \quad (3.52)$$

The proof is similar to the proof of the two previous convolution theorems—2.2 and 3.7. This theorem shows that a circular convolution can be interpreted as a discrete frequency filtering. It also opens the door to fast computations of convolutions using the fast Fourier transform.

### 3.3.3 Fast Fourier Transform

For a signal  $f$  of  $N$  points, a direct calculation of the  $N$  discrete Fourier sums

$$\hat{f}[k] = \sum_{n=0}^{N-1} f[n] \exp\left(\frac{-i2\pi kn}{N}\right), \quad \text{for } 0 \leq k < N, \quad (3.53)$$

requires  $N^2$  complex multiplications and additions. The FFT algorithm reduces the numerical complexity to  $O(N \log_2 N)$  by reorganizing the calculations.

When the frequency index is even, we group the terms  $n$  and  $n + N/2$ :

$$\hat{f}[2k] = \sum_{n=0}^{N/2-1} (f[n] + f[n + N/2]) \exp\left(\frac{-i2\pi kn}{N/2}\right). \quad (3.54)$$

When the frequency index is odd, the same grouping becomes

$$\hat{f}[2k+1] = \sum_{n=0}^{N/2-1} \exp\left(\frac{-i2\pi n}{N}\right) (f[n] - f[n + N/2]) \exp\left(\frac{-i2\pi kn}{N/2}\right). \quad (3.55)$$

Equation (3.54) proves that even frequencies are obtained by calculating the DFT of the  $N/2$  periodic signal

$$f_e[n] = f[n] + f[n + N/2].$$

Odd frequencies are derived from (3.55) by computing the Fourier transform of the  $N/2$  periodic signal:

$$f_o[n] = \exp\left(\frac{-i2\pi n}{N}\right) (f[n] - f[n + N/2]).$$

A DFT of size  $N$  may thus be calculated with two discrete Fourier transforms of size  $N/2$  plus  $O(N)$  operations.

The inverse FFT of  $\hat{f}$  is derived from the forward fast Fourier transform of its complex conjugate  $\hat{f}^*$  by observing that

$$f^*[n] = \frac{1}{N} \sum_{k=0}^{N-1} \hat{f}^*[k] \exp\left(\frac{-i2\pi kn}{N}\right). \quad (3.56)$$

### Complexity

Let  $C(N)$  be the number of elementary operations needed to compute a DFT with the FFT. Since  $f$  is complex, the calculation of  $f_e$  and  $f_o$  requires  $N$  complex additions and  $N/2$  complex multiplications. Let  $KN$  be the corresponding number of elementary operations. We have

$$C(N) = 2 C(N/2) + KN, \quad (3.57)$$

since the Fourier transform of a single point is equal to itself,  $C(1) = 0$ . With the change of variable  $l = \log_2 N$  and the change of function  $T(l) = \frac{C(N)}{N}$ , from (3.57) we derive

$$T(l) = T(l - 1) + K.$$

Since  $T(0) = 0$ , we get  $T(l) = Kl$  and therefore

$$C(N) = KN \log_2(N).$$

Several variations of this fast algorithm exist [49, 237]. The goal is to minimize the constant  $K$ . The most efficient fast DFT to this date is the split-radix FFT algorithm, which is slightly more complicated than the procedure just described; however, it requires only  $N \log_2 N$  real multiplications and  $3N \log_2 N$  additions. When the input signal  $f$  is real, there are half as many parameters to compute, since  $\hat{f}[-k] = \hat{f}^*[k]$ . The number of multiplications and additions is thus reduced by 2.

### 3.3.4 Fast Convolutions

The low computational complexity of a FFT makes it efficient to compute finite discrete convolutions by using the circular convolution, Theorem 3.9. Let  $f$  and  $h$  be two signals with samples that are nonzero only for  $0 \leq n < M$ . The causal signal

$$g[n] = f \star h[n] = \sum_{k=-\infty}^{+\infty} f[k] h[n-k] \quad (3.58)$$

is nonzero only for  $0 \leq n < 2M$ . If  $h$  and  $f$  have  $M$  nonzero samples, calculating this convolution product with the sum (3.58) requires  $M(M+1)$  multiplications and additions. When  $M \geq 32$ , the number of computations is reduced by using the FFT [11, 49].

To use the fast Fourier transform with the circular convolution, Theorem 3.9, the noncircular convolution (3.58) is written as a circular convolution. We define two signals of period  $2M$ :

$$a[n] = \begin{cases} f[n] & \text{if } 0 \leq n < M \\ 0 & \text{if } M \leq n < 2M \end{cases} \quad (3.59)$$

$$b[n] = \begin{cases} h[n] & \text{if } 0 \leq n < M \\ 0 & \text{if } M \leq n < 2M \end{cases} \quad (3.60)$$

By letting  $c = a \circledast b$ , one can verify that  $c[n] = g[n]$  for  $0 \leq n < 2M$ . The  $2M$  nonzero coefficients of  $g$  are thus obtained by computing  $\hat{a}$  and  $\hat{b}$  from  $a$  and  $b$  and then calculating the inverse DFT of  $\hat{c} = \hat{a} \hat{b}$ .

With the fast Fourier transform algorithm, this requires a total of  $O(M \log_2 M)$  additions and multiplications instead of  $M(M + 1)$ . A single FFT or inverse FFT of a real signal of size  $N$  is calculated with  $2^{-1}N \log_2 N$  multiplications, using a split-radix algorithm. The FFT convolution is thus performed with a total of  $3M \log_2 M + 11M$  real multiplications. For  $M \geq 32$ , the FFT algorithm is faster than the direct convolution approach. For  $M \leq 16$ , it is faster to use a direct convolution sum.

### **Fast Overlap–Add Convolutions**

The convolution of a signal  $f$  of  $L$  nonzero samples with a smaller causal signal  $h$  of  $M$  samples is calculated with an overlap–add procedure that is faster than the previous algorithm. The signal  $f$  is decomposed with a sum of  $L/M$  blocks  $f_r$  having  $M$  nonzero samples:

$$f[n] = \sum_{r=0}^{L/M-1} f_r[n - rM], \quad \text{with } f_r[n] = f[n + rM] \mathbf{1}_{[0,M-1]}[n]. \quad (3.61)$$

For each  $0 \leq r < L/M$ , the  $2M$  nonzero samples of  $g_r = f_r \star h$  are computed with the FFT-based convolution algorithm, which requires  $O(M \log_2 M)$  operations. These  $L/M$  convolutions are thus obtained with  $O(L \log_2 M)$  operations. The block decomposition (3.61) implies that

$$f \star h[n] = \sum_{r=0}^{L/M-1} g_r[n - rM]. \quad (3.62)$$

The addition of these  $L/M$  translated signals of size  $2M$  is done with  $2L$  additions. The overall convolution is thus performed with  $O(L \log_2 M)$  operations.

## **3.4 DISCRETE IMAGE PROCESSING**

Two-dimensional signal processing poses many specific geometric and topological problems that do not exist in one dimension [21, 33]. For example, a simple concept, such as causality, is not well defined in two dimensions. We can avoid the complexity introduced by the second dimension by extending one-dimensional algorithms with a separable approach. This not only simplifies the mathematics but also leads to fast numerical algorithms along the rows and columns of images. Section A.5 in the Appendix reviews the properties of tensor products for separable calculations.

### **3.4.1 Two-Dimensional Sampling Theorems**

The light intensity measured by a camera is generally sampled over a rectangular array of picture elements, called *pixels*. One-dimensional sampling theorems are

extended to this two-dimensional sampling array. Other two-dimensional sampling grids such as hexagonal, are also possible, but nonrectangular sampling arrays are not used often.

Let  $s_1$  and  $s_2$  be the sampling intervals along the  $x_1$  and  $x_2$  axes of an infinite rectangular sampling grid. The following renormalizes the axes so that  $s_1 = s_2 = s$ . A discrete image obtained by sampling  $f(x)$  with  $x = (x_1, x_2)$  can be represented as a sum of Diracs located at the grid points:

$$f_d(x) = \sum_{n \in \mathbb{Z}^2} f(sn) \delta(x - ns).$$

The two-dimensional Fourier transform of  $\delta(x - sn)$  is  $e^{-isn \cdot \omega}$  with  $\omega = (\omega_1, \omega_2)$  and  $n \cdot \omega = n_1 \omega_1 + n_2 \omega_2$ . Thus, the Fourier transform of  $f_d$  is a two-dimensional Fourier series:

$$\hat{f}_d(\omega) = \sum_{n \in \mathbb{Z}^2} f(sn) e^{-isn \cdot \omega}. \quad (3.63)$$

It is  $2\pi/s$  periodic along  $\omega_1$  and along  $\omega_2$ . An extension of Theorem 3.1 relates  $\hat{f}_d$  to the two-dimensional Fourier transform  $\hat{f}$  of  $f$ .

**Theorem 3.10.** The Fourier transform of the discrete image  $f_d(x)$  is

$$\hat{f}_d(\omega) = \frac{1}{s^2} \sum_{k \in \mathbb{Z}^2} \hat{f}(\omega - 2k\pi/s), \quad \text{with } k = (k_1, k_2). \quad (3.64)$$

We derive the following two-dimensional sampling theorem, which is analogous to Theorem 3.2.

**Theorem 3.11.** If  $\hat{f}$  has a support included in  $[-\pi/s, \pi/s]^2$ , then

$$f(x) = s \sum_{n \in \mathbb{Z}^2} f(ns) \phi_s(x - ns), \quad (3.65)$$

where

$$\phi_s(x_1, x_2) = \frac{1}{s} \frac{\sin(\pi x_1/s)}{\pi x_1/s} \frac{\sin(\pi x_2/s)}{\pi x_2/s}. \quad (3.66)$$

If the support of  $\hat{f}$  is not included in the low-frequency rectangle  $[-\pi/s, \pi/s]^2$ , the interpolation formula (3.65) introduces aliasing errors. Such aliasing is eliminated by prefiltering  $f$  with the ideal low-pass separable filter  $\phi_s(x)$  having a Fourier transform equal to 1 on  $[-\pi/s, \pi/s]^2$ .

### General Sampling Theorems

As explained in Section 3.1.3, the Shannon-Whittaker sampling theorem is a particular case of more general linear sampling theorems with low-pass filters. The following theorem is a two-dimensional extension of Theorems 3.3 and 3.4; it characterizes these filters to obtain a stable reconstruction.

**Theorem 3.12.** If there exists  $B \geq A > 0$  such that the Fourier transform of  $\phi_s \in \mathbf{L}^2(\mathbb{R}^2)$  satisfies

$$\forall \omega \in [0, 2\pi/s]^2 \quad A \leq \hat{h}(\omega) = \sum_{k \in \mathbb{Z}^2} |\hat{\phi}_s(\omega - 2k\pi/s)|^2 \leq B,$$

then  $\{\phi_s(x - ns)\}_{n \in \mathbb{Z}^2}$  is a Riesz basis of a space  $\mathbf{U}_s$ . The Fourier transform of the dual filter  $\tilde{\phi}_s$  is  $\hat{\tilde{\phi}}_s(\omega) = \hat{\phi}_s^*(\omega)/h(\omega)$ , and the orthogonal projection of  $f \in \mathbf{L}^2(\mathbb{R}^2)$  in  $\mathbf{U}_s$  is

$$P_{\mathbf{U}_s} f(x) = \sum_{n \in \mathbb{Z}^2} f \star \bar{\phi}_s(ns) \tilde{\phi}_s(x - ns), \quad \text{with } \bar{\phi}_s(x) = \phi_s(-x). \quad (3.67)$$

This theorem gives a necessary and sufficient condition to obtain a stable linear reconstruction from samples computed with a linear filter. The proof is a direct extension of the proofs of Theorems 3.3 and 3.4. It recovers a signal approximation as an orthogonal projection by filtering the discrete signal  $f_d(x) = \sum_{n \in \mathbb{Z}^2} f \star \bar{\phi}_s(ns) \delta(x - ns)$ :

$$P_{\mathbf{U}_s} f(x) = f_d \star \tilde{\phi}_s(x).$$

The same as in one dimension, the filter  $\phi_s$  can be obtained by scaling a single filter  $\phi_s(x) = s^{-1}\phi(s^{-1}x)$ . The two-dimensional Shannon-Whittaker theorem is a particular example, where  $\hat{\phi}_s = s\mathbf{1}_{[-\pi/s, \pi/s]^2}$ , which defines an orthonormal basis of the space  $\mathbf{U}_s$  of functions having a Fourier transform supported in  $[-\pi/s, \pi/s]^2$ .

### 3.4.2 Discrete Image Filtering

The properties of two-dimensional space-invariant operators are essentially the same as in one dimension. The sampling interval  $s$  is normalized to 1. A pixel value located at  $n = (n_1, n_2)$  is written  $f[n]$ . A linear operator  $L$  is space-invariant if  $Lf_p[n] = Lf[n-p]$  for any  $f_p[n] = f[n-p]$ , with  $p = (p_1, p_2) \in \mathbb{Z}^2$ . A discrete Dirac is defined by  $\delta[n] = 1$  if  $n = (0, 0)$  and  $\delta[n] = 0$  if  $n \neq (0, 0)$ .

#### Impulse Response

Since  $f[n] = \sum_{p \in \mathbb{Z}^2} f[p] \delta[n-p]$ , linearity and time invariance implies

$$Lf[n] = \sum_{p \in \mathbb{Z}^2} f[p] h[n-p] = f \star h[n], \quad (3.68)$$

where  $h[n]$  is the response of the impulse  $h[n] = L\delta[n]$ . If the impulse response is separable

$$h[n_1, n_2] = h_1[n_1] h_2[n_2], \quad (3.69)$$

then the two-dimensional convolution (3.68) is computed as one-dimensional convolutions along the columns of the image followed by one-dimensional convolutions along the rows (or vice versa):

$$f \star h[n_1, n_2] = \sum_{p_1=-\infty}^{+\infty} h_1[n_1 - p_1] \sum_{p_2=-\infty}^{+\infty} h_2[n_2 - p_2] f[p_1, p_2]. \quad (3.70)$$

This factorization reduces the number of operations. If  $h_1$  and  $h_2$  are finite impulse response filters of size  $M_1$  and  $M_2$ , respectively, then the separable calculation (3.70) requires  $M_1 + M_2$  additions and multiplications per point  $(n_1, n_2)$  as opposed to  $M_1 M_2$  in a nonseparable computation (3.68).

### **Transfer Function**

The Fourier transform of a discrete image  $f$  is defined by the Fourier series:

$$\hat{f}(\omega) = \sum_{n \in \mathbb{Z}^2} f[n] e^{-i\omega \cdot n}, \quad \text{with } \omega \cdot n = n_1 \omega_1 + n_2 \omega_2. \quad (3.71)$$

The two-dimensional extension of the convolution, Theorem 3.7, proves that if  $g[n] = Lf[n] = f \star h[n]$  then its Fourier transform is  $\hat{g}(\omega) = \hat{f}(\omega) \hat{h}(\omega)$ , and  $\hat{h}(\omega)$  is the transfer function of the filter. When a filter is separable  $h[n_1, n_2] = h_1[n_1] h_2[n_2]$ , its transfer function is also separable:

$$\hat{h}(\omega_1, \omega_2) = \hat{h}_1(\omega_1) \hat{h}_2(\omega_2). \quad (3.72)$$

### **3.4.3 Circular Convolutions and Fourier Basis**

The discrete convolution of a finite image  $\tilde{f}$  raises border problems. As in one dimension, these border issues are solved by extending the image, making it periodic along its rows and columns:

$$f[n_1, n_2] = \tilde{f}[n_1 \bmod N_1, n_2 \bmod N_2],$$

where  $N = N_1 N_2$  is the image size. The resulting periodic image  $f[n_1, n_2]$  is defined for all  $(n_1, n_2) \in \mathbb{Z}^2$ , and each of its rows and columns are periodic one-dimensional signals.

A discrete convolution is replaced by a circular convolution over the image period. If  $f$  and  $h$  have a periodicity  $N_1$  and  $N_2$  along  $(n_1, n_2)$ , then

$$f \circledast h[n_1, n_2] = \sum_{p_1=0}^{N_1-1} \sum_{p_2=0}^{N_2-1} f[p_1, p_2] h[n_1 - p_1, n_2 - p_2]. \quad (3.73)$$

### **Discrete Fourier Transform**

The eigenvectors of circular convolutions are two-dimensional discrete sinusoidal waves:

$$e_k[n] = e^{\omega_k \cdot n}, \quad \text{with } \omega_k = (2\pi k_1/N_1, 2\pi k_2/N_2) \quad \text{for } 0 \leq k_1 < N_1, 0 \leq k_2 < N_2.$$

This family of  $N = N_1 N_2$  discrete vectors is the separable product of two one-dimensional discrete Fourier bases  $\{e^{i2\pi k_1 n/N_1}\}_{0 \leq k_1 < N_1}$  and  $\{e^{i2\pi k_2 n/N_2}\}_{0 \leq k_2 < N_2}$ . Thus, Theorem A.3 proves that the family  $\{e_k[n]\}_{0 \leq k_1 < N_1, 0 \leq k_2 < N_2}$  is an orthogonal basis of  $\mathbb{C}^N = \mathbb{C}^{N_1} \otimes \mathbb{C}^{N_2}$  (Exercise 3.23). Any image  $f \in \mathbb{C}^N$  can be decomposed in this orthogonal basis:

$$f[n] = \frac{1}{N} \sum_{k_1=0}^{N_1-1} \sum_{k_2=0}^{N_2-1} \hat{f}[k] e^{i\omega_k \cdot n}, \quad (3.74)$$

where  $\hat{f}$  is the two-dimensional DFT of  $f$

$$\hat{f}[k] = \langle f, e_k \rangle = \sum_{n_1=0}^{N_1-1} \sum_{n_2=0}^{N_2-1} f[n] e^{-i\omega_k \cdot n}. \quad (3.75)$$

### Fast Convolutions

Since  $e^{i\omega_k \cdot n}$  are eigenvectors of two-dimensional circular convolutions, the DFT of  $g = f \star h$  is

$$\hat{g}[k] = \hat{f}[k] \hat{h}[k]. \quad (3.76)$$

A direct computation of  $f \star h$  with the summation (3.73) requires  $O(N^2)$  multiplications. With the two-dimensional FFT described next,  $\hat{f}[k]$  and  $\hat{h}[k]$  as well as the inverse DFT of their product (3.76) are calculated with  $O(N \log N)$  operations. Noncircular convolutions are computed with a fast algorithm by reducing them to circular convolutions, with the same approach as in Section 3.3.4.

### Separable Basis Decomposition

Let  $\mathcal{B}_1 = \{e_{k_1}^1\}_{0 \leq k_1 < N_1}$  and  $\mathcal{B}_2 = \{e_{k_2}^2\}_{0 \leq k_2 < N_2}$  be two orthogonal bases of  $\mathbb{C}^{N_1}$  and  $\mathbb{C}^{N_2}$ . Suppose the calculation of decomposition coefficients of  $f_1 \in \mathbb{C}^{N_1}$  in the basis  $\mathcal{B}_1$  requires  $C_1(N_1)$  operations and of  $f_2 \in \mathbb{C}^{N_2}$  in the basis  $\mathcal{B}_2$  requires  $C_2(N_2)$  operations. One can verify (Exercise 3.23) that the family  $\mathcal{B} = \{e_k[n] = e_{k_1}^1[n_1] e_{k_2}^2[n_2]\}_{0 \leq k_1 < N_1, 0 \leq k_2 < N_2}$  is an orthogonal basis of the space  $\mathbb{C}^N = \mathbb{C}^{N_1} \otimes \mathbb{C}^{N_2}$  of images  $f[n_1, n_2]$  of  $N = N_1 N_2$  pixels. We describe a fast separable algorithm that computes the decomposition coefficients of an image  $f$  in  $\mathcal{B}$  with  $N_2 C_1(N_1) + N_1 C_2(N_2)$  operations as opposed to  $N^2$ . A fast two-dimensional FFT is derived.

Two-dimensional inner products are calculated with

$$\begin{aligned} \langle f, e_{k_1}^1 e_{k_2}^2 \rangle &= \sum_{n_1=0}^{N_1-1} \sum_{n_2=0}^{N_2-1} f[n_1, n_2] e_{k_1}^{1*}[n_1] e_{k_2}^{2*}[n_2] \\ &= \sum_{n_1=0}^{N_1-1} e_{k_1}^{1*}[n_1] \sum_{n_2=0}^{N_2-1} f[n_1, n_2] e_{k_2}^{2*}[n_2]. \end{aligned} \quad (3.77)$$

For  $0 \leq n_1 < N_1$ , we must compute

$$Uf[n_1, k_2] = \sum_{n_2=0}^{N_2-1} f[n_1, n_2] e_{k_2}^{2*}[n_2],$$

which are the decomposition coefficients of the  $N_1$  image rows of size  $N_2$  in the basis  $\mathcal{B}_2$ . The coefficients  $\{\langle f, e_{k_1}^1 e_{k_2}^2 \rangle\}_{0 \leq k_1 < N_1, 0 \leq k_2 < N_2}$  are calculated in (3.77) as the inner products of the columns of the transformed image  $Uf[n_1, k_2]$  in the basis  $\mathcal{B}_1$ . The overall algorithm thus requires performing  $N_1$  one-dimensional transforms

in the basis  $\mathcal{B}_2$  plus  $N_2$  one-dimensional transforms in the basis  $\mathcal{B}_1$ ; it therefore requires  $N_2 C_1(N_1) + N_1 C_2(N_2)$  operations.

The fast Fourier transform algorithm of Section 3.3.3 decomposes signals of size  $N_1$  and  $N_2$  in the discrete Fourier bases  $\mathcal{B}_1 = \{e_{k_1}^1[n_1] = e^{i2\pi k_1 n_1 / N_1}\}_{0 \leq k_1 < N_1}$  and  $\mathcal{B}_2 = \{e_{k_2}^2[n_2] = e^{i2\pi k_2 n_2 / N_2}\}_{0 \leq k_2 < N_2}$ , with  $C_1(N_1) = KN_1 \log_2 N_1$  and  $C_2(N_2) = KN_2 \log_2 N_2$  operations. A separable implementation of a two-dimensional FFT thus requires  $N_2 C_1(N_1) + N_1 C_2(N_2) = KN \log_2 N$  operations, with  $N = N_1 N_2$ . A split-radix FFT corresponds to  $K = 3$ .

## 3.5 EXERCISES

- 3.1** <sup>1</sup> Show that if  $\phi_s(t) = s^{-1/2} \mathbf{1}_{[0,s)}(t)$ , then  $\{\phi_s(t - ns)\}_{n \in \mathbb{Z}}$  is an orthonormal basis of the space of piecewise constant function on intervals  $[ns, (n+1)s)$  for any  $n \in \mathbb{Z}$ .
- 3.2** <sup>2</sup> Prove that if  $f$  has a Fourier transform included in  $[-\pi/s, \pi/s]$ , then

$$\forall u \in \mathbb{R}, \quad f(u) = \frac{1}{s} \langle f(t), \phi_s(t-u) \rangle \quad \text{with} \quad \phi_s(t) = \frac{\sin(\pi t/s)}{\pi t/s}.$$

- 3.3** <sup>2</sup> An interpolation function  $f(t)$  satisfies  $f(n) = \delta[n]$  for any  $n \in \mathbb{Z}$ .
- (a) Prove that  $\sum_{k=-\infty}^{+\infty} \hat{f}(\omega + 2k\pi) = 1$  if and only if  $f$  is an interpolation function.
  - (b) Suppose that  $f(t) = \sum_{n=-\infty}^{+\infty} h[n] \theta(t-n)$  with  $\theta \in \mathbf{L}^2(\mathbb{R})$ . Find  $\hat{h}(\omega)$  as a function of  $\hat{\theta}(\omega)$  so that  $f(t)$  is an interpolation function. Relate  $\hat{f}(\omega)$  to  $\hat{\theta}(\omega)$ , and give a sufficient condition on  $\hat{\theta}$  to guarantee that  $f \in \mathbf{L}^2(\mathbb{R})$ .
- 3.4** <sup>2</sup> Prove that if  $f \in \mathbf{L}^2(\mathbb{R})$  and  $\sum_{n=-\infty}^{+\infty} f(t-n) \in \mathbf{L}^2[0, 1]$ , then

$$\sum_{n=-\infty}^{+\infty} f(t-n) = \sum_{k=-\infty}^{+\infty} \hat{f}(2k\pi) e^{i2\pi kt}.$$

- 3.5** <sup>1</sup> We want to approximate  $f$  by a signal  $\tilde{f}$  in an approximation space  $\mathbf{U}_s$ . Prove that the approximation  $\tilde{f}$  that minimizes  $\|\tilde{f} - f\|$ , is the orthogonal projection of  $f$  in  $\mathbf{U}_s$ .
- 3.6** <sup>2</sup> Prove that the discrete filter  $Lf[n] = f \star h[n]$  is stable if and only if  $h \in \ell^1(\mathbb{Z})$ .
- 3.7** <sup>2</sup> If  $\hat{h}(\omega)$  and  $\hat{g}(\omega)$  are the Fourier transforms of  $h[n]$  and  $g[n]$ , we write

$$\hat{h} \star \hat{g}(\omega) = \int_{-\pi}^{+\pi} \hat{h}(\xi) \hat{g}(\omega - \xi) d\xi.$$

Prove that if  $f[n] = g[n] h[n]$ , then  $\hat{f}(\omega) = (2\pi)^{-1} \hat{h} \star \hat{g}(\omega)$ .

- 3.8** <sup>1</sup> Prove that  $\{e^{i2\pi kn/N}\}_{0 \leq k < N}$  is an orthogonal family and thus an orthogonal basis of  $\mathbb{C}^N$ . What renormalization factor is needed to obtain an orthonormal basis?

- 3.9** <sup>2</sup> Suppose that  $\hat{f}$  has a support in  $[-(n+1)\pi/s, -n\pi/s] \cup [n\pi/s, (n+1)\pi/s]$  and that  $f(t)$  is real. Find an interpolation formula that recovers  $f(t)$  from  $\{f(ns)\}_{n \in \mathbb{Z}}$ .

- 3.10** <sup>3</sup> Suppose that  $\hat{f}$  has a support in  $[-\pi/s, \pi/s]$ .

- (a) Give the filter  $\phi_s(t)$  such that for any  $f$ ,

$$\forall n \in \mathbb{Z}, \quad \tilde{f}(ns) = \int_{(n-1/2)s}^{(n+1/2)s} f(t) dt = f \star \phi_s(ns).$$

- (b) Show that  $\tilde{f}(t) = f \star \phi_s(t)$  can be recovered from  $\{\tilde{f}(ns)\}_{n \in \mathbb{Z}}$  with an interpolation formula.

- (c) Reconstruct  $f$  from  $\tilde{f}$  by inverting  $\phi_s$ .

- (d) Prove that the reconstruction of  $f(t)$  from  $\{\tilde{f}(ns)\}_{n \in \mathbb{Z}}$  is stable.

- 3.11** <sup>2</sup> The linear box spline  $\phi(t)$  is defined in (3.31) for  $m = 1$ .

- (a) Give an analytical formula for  $\phi(t)$  and specify its support.

- (b) Prove with (7.20) that  $\{\phi(t-n)\}_{n \in \mathbb{Z}}$  is a Riesz basis of the space of finite-energy functions that are continuous and linear on intervals  $[ns, (n+1)]$  for  $n \in \mathbb{Z}$ .

- (c) Does the dual filter  $\tilde{\phi}(t)$  have a compact support? Compute its graph numerically.

- 3.12** <sup>1</sup> If  $f[n]$  is defined for  $0 \leq n < N$ , prove that  $|\hat{f}[k]| \leq \sum_{n=0}^{N-1} |f[n]|$  for any  $0 \leq k < N$ .

- 3.13** <sup>2</sup> The discrete and periodic total variation is

$$\|f\|_V = \sum_{n=0}^{N-1} |f[n] - f[n-1]| + |f[N-1] - f[0]|.$$

- (a) Prove that  $\|f\|_V = \sum_{n=0}^{N-1} |f \otimes h[n]|$  where  $h[n]$  is a filter and specify  $\hat{h}[k]$ .

- (b) Derive an upper bound of  $|\hat{f}[k]|$  as a function of  $k^{-1}$ .

- 3.14** <sup>1</sup> Let  $g[n] = (-1)^n h[n]$ . Relate  $\hat{g}(\omega)$  to  $\hat{h}(\omega)$ . If  $h$  is a low-pass filter, what kind of filter is  $g$ ?

- 3.15** <sup>2</sup> Prove the convolution Theorem 3.7.

- 3.16** <sup>2</sup> Let  $h^{-1}$  be the inverse of  $h$  defined by  $h \star h^{-1}[n] = \delta[n]$ .

- (a) Compute  $\hat{h}^{-1}(\omega)$  as a function of  $\hat{h}(\omega)$ .

- (b) Prove that if  $h$  has a finite support, then  $h^{-1}$  has a finite support if and only if  $h[n] = \delta[n-p]$  for some  $p \in \mathbb{Z}$ .

**3.17 1 All pass filters:**

(a) Verify that

$$\hat{h}(\omega) = \prod_{k=1}^K \frac{a_k^* - e^{-i\omega}}{1 + a_k e^{i\omega}}$$

is an all-pass filter; that is,  $|\hat{h}(\omega)| = 1$ .(b) Prove that  $\{h[n-m]\}_{m \in \mathbb{Z}}$  is an orthonormal basis of  $\ell^2(\mathbb{Z})$ .**3.18 2 Recursive filters:**

- (a) Compute the Fourier transform of  $h[n] = a^n \mathbf{1}_{[0,+\infty)}[n]$  for  $|a| < 1$ . Compute the inverse Fourier transform of  $\hat{h}(\omega) = (1 - a e^{-i\omega})^{-p}$ .
- (b) Suppose that  $g = f \star h$  is calculated by a recursive equation with real coefficients

$$\sum_{k=0}^K a_k f[n-k] = \sum_{k=0}^M b_k g[n-k].$$

Write  $\hat{h}(\omega)$  as a function of the parameters  $a_k$  and  $b_k$ .(c) Show that  $h$  is a stable filter if and only if the equation  $\sum_{k=0}^M b_k z^{-k} = 0$  has roots with a modulus strictly smaller than 1.**3.19 1 Discrete interpolation.** Let  $\hat{f}[k]$  be the DFT of a signal  $f[n]$  of size  $N$ . We define a signal  $\tilde{f}[n]$  of size  $2N$  by  $\tilde{\hat{f}}[N/2] = \hat{\tilde{f}}[3N/2] = \hat{f}[N/2]$  and

$$\tilde{\hat{f}}[k] = \begin{cases} 2\hat{f}[k] & \text{if } 0 \leq k < N/2 \\ 0 & \text{if } N/2 < k < 3N/2 \\ 2\hat{f}[k-N] & \text{if } 3N/2 < k < 2N. \end{cases}$$

Prove that  $\tilde{f}$  is an interpolation of  $f$  that satisfies  $\tilde{f}[2n] = f[n]$ .**3.20 2 Decimation.** Let  $x[n] = y[Mn]$  with  $M > 1$ .

- (a) Show that  $\hat{x}(\omega) = M^{-1} \sum_{k=0}^{M-1} \hat{y}(M^{-1}(\omega - 2k\pi))$ .
- (b) Give a sufficient condition on  $\hat{y}(\omega)$  to recover  $y$  from  $x$  and give the interpolation formula that recovers  $y[n]$  from  $x$ .

**3.21 3** We want to compute numerically the Fourier transform of  $f(t)$ . Let  $f_d[n] = f(ns)$  and  $f_p[n] = \sum_{p=-\infty}^{+\infty} f_d[n-pN]$ .

- (a) Prove that the DFT of  $f_p[n]$  is related to the Fourier series of  $f_d[n]$  and to the Fourier transform of  $f(t)$  by

$$\hat{f}_p[k] = \hat{f}_d\left(\frac{2\pi k}{N}\right) = \frac{1}{s} \sum_{l=-\infty}^{+\infty} \hat{f}\left(\frac{2k\pi}{Ns} - \frac{2l\pi}{s}\right).$$

- (b) Suppose that  $|f(t)|$  and  $|\hat{f}(\omega)|$  are negligible when  $t \notin [-t_0, t_0]$  and  $\omega \notin [-\omega_0, \omega_0]$ . Relate  $N$  and  $s$  to  $t_0$  and  $\omega_0$  so that one can compute

an approximate value of  $\hat{f}(\omega)$  for all  $\omega \in \mathbb{R}$  by interpolating the samples  $\hat{f}_p[k]$ . Is it possible to compute exactly  $\hat{f}(\omega)$  with such an interpolation formula?<sup>4</sup>

- (c) Let  $f(t) = (\sin(\pi t)/(\pi t))^4$ . What is the support of  $\hat{f}$ ? Sample  $f$  appropriately and compute  $\hat{f}$  numerically with an FFT algorithm.

- 3.22** <sup>2</sup>The analytic part  $f_a[n]$  of a real discrete signal  $f[n]$  of size  $N$  is defined by

$$\hat{f}_a[k] = \begin{cases} \hat{f}[k] & \text{if } k = 0, N/2 \\ 2\hat{f}[k] & \text{if } 0 < k < N/2 \\ 0 & \text{if } N/2 < k < N. \end{cases}$$

- (a) Compute  $f_a[n]$  for  $f[n] = \cos(2\pi kn/N)$  with  $0 < k < N/2$ .  
 (b) Prove that the real part  $g[n] = \operatorname{Re}(f[n])$  is what satisfies

$$\hat{g}[k] = (\hat{f}[k] + \hat{f}^*[-k])/2.$$

- (c) Prove that  $\operatorname{Re}(f_a) = f$ .

- 3.23** <sup>1</sup>Prove that if  $\{e_{k_1}[n_1]\}_{0 \leq k_1 < N_1}$  is an orthonormal basis of  $\mathbb{C}^{N_1}$  and  $\{e_{k_2}[n_2]\}_{0 \leq k_2 < N_2}$  is an orthonormal basis of  $\mathbb{C}^{N_2}$ , then  $\{e_{k_1}[n_1] e_{k_2}[n_2]\}_{0 \leq k_1 < N_1, 0 \leq k_2 < N_2}$  is an orthogonal basis of the space  $\mathbb{C}^N = \mathbb{C}^{N_1 N_2}$  of images  $f[n_1, n_2]$  of  $N = N_1 N_2$  pixels.

- 3.24** <sup>2</sup>Let  $h[n_1, n_2]$  be a nonseparable filter that is nonzero for  $0 \leq n_1, n_2 < M$ . Let  $f[n_1, n_2]$  be a square image defined for  $0 \leq n_1, n_2 \leq LM$  of  $N = (LM)^2$  pixels. Describe an overlap-add algorithm to compute  $g[n_1, n_2] = f \star h[n_1, n_2]$ . By using an FFT that requires  $K P \log P$  operators to compute the Fourier transform of an image of  $P$  pixels, how many operations does your algorithm require? If  $K = 6$ , for what range of  $M$  is it better to compute the convolution with a direct summation?

- 3.25** <sup>2</sup>Let  $f[n_1, n_2, n_3]$  be a three-dimensional signal of size  $N = N_1 N_2 N_3$ . The discrete Fourier transform is defined as a decomposition in a separable discrete Fourier basis. Give a separable algorithm that decomposes  $f$  in this basis with  $K N \log N$  operations, by using a one-dimensional FFT algorithm that requires  $K P \log P$  operations for a one-dimensional signal of size  $P$ .

# Time Meets Frequency

# 4

When we listen to music, we clearly “hear” the time variation of the sound “frequencies.” These localized frequency events are not “pure” tones but packets of close frequencies. The properties of sounds are revealed by transforms that decompose signals over elementary functions that have a narrow localization in time and frequency. Windowed Fourier transforms and wavelet transforms are two important classes of local time-frequency decompositions. Measuring the time variations of “instantaneous” frequencies illustrates the limitations imposed by the Heisenberg uncertainty. Such frequencies are detected as local maxima in windowed Fourier and wavelet dictionaries and define a signal-approximation support. Audio-processing algorithms are implemented by modifying the geometry of this approximation support.

There is no unique definition of time-frequency energy density; all quadratic distributions are related through the averaging of a single quadratic form called the Wigner-Ville distribution. This framework gives another perspective on windowed Fourier and wavelet transforms.

## 4.1 TIME-FREQUENCY ATOMS

A linear time-frequency transform correlates the signal with a dictionary of waveforms that are concentrated in time and in frequency. The waveforms are called *time-frequency atoms*. Let us consider a general dictionary of atoms  $\mathcal{D} = \{\phi_\gamma\}_{\gamma \in \Gamma}$ , where  $\gamma$  might be a multiindex parameter. We suppose that  $\phi_\gamma \in \mathbf{L}^2(\mathbb{R})$  and that  $\|\phi_\gamma\| = 1$ . The corresponding linear time-frequency transform of  $f \in \mathbf{L}^2(\mathbb{R})$  is defined by

$$\Phi f(\gamma) = \int_{-\infty}^{+\infty} f(t) \phi_\gamma^*(t) dt = \langle f, \phi_\gamma \rangle.$$

The Parseval formula (2.25) proves that

$$\Phi f(\gamma) = \int_{-\infty}^{+\infty} f(t) \phi_\gamma^*(t) dt = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \hat{f}(\omega) \hat{\phi}_\gamma^*(\omega) d\omega. \quad (4.1)$$

If  $\phi_\gamma(t)$  is nearly zero when  $t$  is outside a neighborhood of an abscissa  $u$ , then  $\langle f, \phi_\gamma \rangle$  depends only on the values of  $f$  in this neighborhood. Similarly, if  $\hat{\phi}_\gamma(\omega)$  is negligible for  $\omega$  far from  $\xi$ , then the right integral of (4.1) proves that  $\langle f, \phi_\gamma \rangle$  reveals the properties of  $\hat{f}$  in the neighborhood of  $\xi$ .

### Heisenberg Boxes

The slice of information provided by  $\langle f, \phi_\gamma \rangle$  is represented in a time-frequency plane  $(t, \omega)$  by a rectangle having a position and size that depends on the time-frequency spread of  $\phi_\gamma$ . Since

$$\|\phi_\gamma\|^2 = \int_{-\infty}^{+\infty} |\phi_\gamma(t)|^2 dt = 1,$$

we interpret  $|\phi_\gamma(t)|^2$  as a probability distribution centered at

$$u_\gamma = \int_{-\infty}^{+\infty} t |\phi_\gamma(t)|^2 dt. \quad (4.2)$$

The spread around  $u_\gamma$  is measured by the variance

$$\sigma_t^2(\gamma) = \int_{-\infty}^{+\infty} (t - u_\gamma)^2 |\phi_\gamma(t)|^2 dt. \quad (4.3)$$

The Plancherel formula (2.26) proves that

$$\int_{-\infty}^{+\infty} |\hat{\phi}_\gamma(\omega)|^2 d\omega = 2\pi \|\phi_\gamma\|^2.$$

The center frequency of  $\hat{\phi}_\gamma$  is therefore defined by

$$\xi_\gamma = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \omega |\hat{\phi}_\gamma(\omega)|^2 d\omega, \quad (4.4)$$

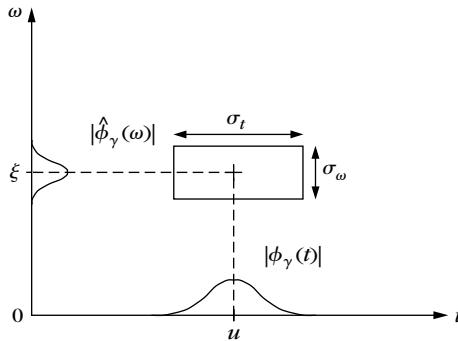
and its spread around  $\xi_\gamma$  is

$$\sigma_\omega^2(\gamma) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} (\omega - \xi_\gamma)^2 |\hat{\phi}_\gamma(\omega)|^2 d\omega. \quad (4.5)$$

The time-frequency resolution of  $\phi_\gamma$  is represented in the time-frequency plane  $(t, \omega)$  by a Heisenberg box centered at  $(u_\gamma, \xi_\gamma)$ , having a time width equal to  $\sigma_t(\gamma)$  and a frequency  $\sigma_\omega(\gamma)$ . Figure 4.1 illustrates this. The Heisenberg uncertainty Theorem 2.6 proves that the area of the rectangle is at least one-half:

$$\sigma_t \sigma_\omega \geq \frac{1}{2}. \quad (4.6)$$

This limits the joint resolution of  $\phi_\gamma$  in time and frequency. The time-frequency plane must be manipulated carefully because a point  $(t_0, \omega_0)$  is ill-defined. There is no function that is concentrated perfectly at a point  $t_0$  and a frequency  $\omega_0$ . Only rectangles with an area of at least one-half may correspond to time-frequency atoms.

**FIGURE 4.1**

Heisenberg box representing an atom  $\phi_\gamma$ .

### Translation-Invariant Dictionaries

For pattern recognition, it can be important to construct signal representations that are translation-invariant. When a pattern is translated, its numerical descriptors are then translated but not modified. Observe that for any  $\phi_\gamma \in \mathcal{D}$  and any shift  $u$ ,

$$\langle f(t-u), \phi_\gamma(t) \rangle = \langle f(t), \phi_\gamma(t+u) \rangle.$$

A translation-invariant representation is thus obtained if  $\phi_\gamma(t+u)$  is in  $\mathcal{D}$  up to a multiplicative constant. Such a dictionary is said to be translation-invariant.

A translation-invariant dictionary is obtained by translating a family of generators  $\{\phi_\gamma\}_{\gamma \in \Gamma}$  and can be written  $\mathcal{D} = \{\phi_{u,\gamma}\}_{\gamma \in \Gamma, u \in \mathbb{R}}$ , with  $\phi_{u,\gamma}(t) = \lambda_{u,\gamma} \phi_\gamma(t-u)$ . The resulting time-frequency transform of  $f$  can then be written as a convolution:

$$\Phi f(u, \gamma) = \langle f, \phi_{u,\gamma} \rangle = \int_{-\infty}^{+\infty} f(t) \lambda_{u,\gamma} \phi_\gamma^*(t-u) dt = \lambda_{u,\gamma} f \star \tilde{\phi}_\gamma(u),$$

with  $\tilde{\phi}_\gamma(t) = \phi_\gamma^*(-t)$ .

### Energy Density

Let us suppose that  $\phi_\gamma(t)$  is centered at  $t=0$  so that  $\phi_{u,\gamma}(t)$  is centered at  $u$ . Let  $\xi_\gamma$  be the center frequency of  $\hat{\phi}_\gamma(\omega)$  defined in (4.4). The time-frequency box of  $\phi_{u,\gamma}$  specifies a neighborhood of  $(u, \xi_\gamma)$ , where the energy of  $f$  is measured by

$$P_\Phi f(u, \xi_\gamma) = |\langle f, \phi_{u,\gamma} \rangle|^2 = \left| \int_{-\infty}^{+\infty} f(t) \phi_{u,\gamma}^*(t) dt \right|^2. \quad (4.7)$$

Section 4.5.1 proves that any such energy density is an averaging of the Wigner-Ville distribution, with a kernel that depends on the atoms  $\phi_{u,\gamma}$ .

---

**EXAMPLE 4.1**

A windowed Fourier atom is constructed with a window  $g$  modulated by the frequency  $\xi$  and translated by  $u$ :

$$\phi_{u,\gamma}(t) = g_{u,\xi}(t) = e^{i\xi t} g(t-u). \quad (4.8)$$

The resulting window Fourier dictionary  $\mathcal{D} = \{g_{u,\xi}(t)\}_{u,\xi \in \mathbb{R}^2}$  is translation-invariant since  $g_{u,\xi} = e^{i\xi u} g_{0,\xi}(t-u)$ . A windowed Fourier dictionary is also frequency-invariant because

$$e^{i\omega t} g_{u,\xi}(t) = g_{u,\xi+\omega}(t) \in \mathcal{D}.$$

This dictionary is thus particularly useful to analyze patterns that are translated in time and frequency.

A wavelet atom is a dilation by  $s$  and a translation by  $u$  of a *mother wavelet*  $\psi$ :

$$\phi_{u,\gamma}(t) = \psi_{u,s}(t) = \frac{1}{\sqrt{s}} \psi\left(\frac{t-u}{s}\right). \quad (4.9)$$

A wavelet dictionary  $\mathcal{D} = \{\psi_{u,s}(t)\}_{u \in \mathbb{R}, s \in \mathbb{R}^+}$  is translation-invariant but also scale-invariant because scaling any wavelet produces a dilated wavelet that remains in the dictionary. A wavelet dictionary can be used to analyze patterns translated and scaled by arbitrary factors.

Wavelets and windowed Fourier atoms have well-localized energy in time while their Fourier transform is mostly concentrated in a limited-frequency band. The properties of the resulting transforms are studied in Sections 4.2 and 4.3.

---



---

## 4.2 WINDOWED FOURIER TRANSFORM

In 1946, Gabor [267] introduced windowed Fourier atoms to measure the “frequency variations” of sounds. A real and symmetric window  $g(t) = g(-t)$  is translated by  $u$  and modulated by the frequency  $\xi$ :

$$g_{u,\xi}(t) = e^{i\xi t} g(t-u). \quad (4.10)$$

It is normalized  $\|g\| = 1$  so that  $\|g_{u,\xi}\| = 1$  for any  $(u, \xi) \in \mathbb{R}^2$ . The resulting windowed Fourier transform of  $f \in \mathbf{L}^2(\mathbb{R})$  is

$$Sf(u, \xi) = \langle f, g_{u,\xi} \rangle = \int_{-\infty}^{+\infty} f(t) g(t-u) e^{-i\xi t} dt. \quad (4.11)$$

This transform is also called the *short time Fourier transform* because the multiplication by  $g(t-u)$  localizes the Fourier integral in the neighborhood of  $t=u$ .

As in (4.7), one can define an energy density called a *spectrogram*, denoted  $P_S$ :

$$P_S f(u, \xi) = |Sf(u, \xi)|^2 = \left| \int_{-\infty}^{+\infty} f(t) g(t-u) e^{-i\xi t} dt \right|^2. \quad (4.12)$$

The spectrogram measures the energy of  $f$  in a time-frequency neighborhood of  $(u, \xi)$  specified by the Heisenberg box of  $g_{u,\xi}$ .

### Heisenberg Boxes

Since  $g$  is even,  $g_{u,\xi}(t) = e^{i\xi t} g(t-u)$  is centered at  $u$ . The time spread around  $u$  is independent of  $u$  and  $\xi$ :

$$\sigma_t^2 = \int_{-\infty}^{+\infty} (t-u)^2 |g_{u,\xi}(t)|^2 dt = \int_{-\infty}^{+\infty} t^2 |g(t)|^2 dt. \quad (4.13)$$

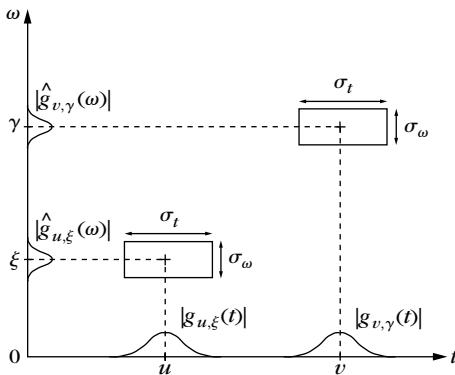
The Fourier transform  $\hat{g}$  of  $g$  is real and symmetric because  $g$  is real and symmetric. The Fourier transform of  $g_{u,\xi}$  is

$$\hat{g}_{u,\xi}(\omega) = \hat{g}(\omega - \xi) \exp[-iu(\omega - \xi)]. \quad (4.14)$$

It is a translation by  $\xi$  of the frequency window  $\hat{g}$ , so its center frequency is  $\xi$ . The frequency spread around  $\xi$  is

$$\sigma_\omega^2 = \frac{1}{2\pi} \int_{-\infty}^{+\infty} (\omega - \xi)^2 |\hat{g}_{u,\xi}(\omega)|^2 d\omega = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \omega^2 |\hat{g}(\omega)|^2 d\omega. \quad (4.15)$$

It is independent of  $u$  and  $\xi$ . Thus,  $g_{u,\xi}$  corresponds to a Heisenberg box of area  $\sigma_t \sigma_\omega$  centered at  $(u, \xi)$ , as illustrated by Figure 4.2. The size of this box is independent of  $(u, \xi)$ , which means that a windowed Fourier transform has the same resolution across the time-frequency plane.



**FIGURE 4.2**

Heisenberg boxes of two windowed Fourier atoms,  $g_{u,\xi}$  and  $g_{v,\gamma}$ .

---

**EXAMPLE 4.2**

A sinusoidal wave  $f(t) = \exp(i\xi_0 t)$ , the Fourier transform of which is a Dirac  $\hat{f}(\omega) = 2\pi\delta(\omega - \xi_0)$ , has a windowed Fourier transform:

$$Sf(u, \xi) = \hat{g}(\xi - \xi_0) \exp[-iu(\xi - \xi_0)].$$

Its energy is spread over the frequency interval  $[\xi_0 - \sigma_\omega/2, \xi_0 + \sigma_\omega/2]$ .

---

**EXAMPLE 4.3**

The windowed Fourier transform of a Dirac  $f(t) = \delta(t - u_0)$  is

$$Sf(u, \xi) = g(u_0 - u) \exp(-i\xi u_0).$$

Its energy is spread in the time interval  $[u_0 - \sigma_t/2, u_0 + \sigma_t/2]$ .

---

**EXAMPLE 4.4**

A linear chirp  $f(t) = \exp(it^2)$  has an “instantaneous” frequency that increases linearly in time. For a Gaussian window  $g(t) = (\pi\sigma^2)^{-1/4} \exp[-t^2/(2\sigma^2)]$ , the windowed Fourier transform of  $f$  is calculated using the Fourier transform (2.34) of Gaussian chirps. One can verify that its spectrogram is

$$P_{Sf}(u, \xi) = |Sf(u, \xi)|^2 = \left( \frac{4\pi\sigma^2}{1 + 4a^2\sigma^4} \right)^{1/2} \exp\left(-\frac{\sigma^2(\xi - 2au)^2}{1 + 4a^2\sigma^4}\right). \quad (4.16)$$

For a fixed time  $u$ ,  $P_{Sf}(u, \xi)$  is a Gaussian that reaches its maximum at the frequency  $\xi(u) = 2au$ . Observe that if we write  $f(t) = \exp[i\phi(t)]$ , then  $\xi(u)$  is equal to the instantaneous frequency, defined as the derivative of the phase:  $\omega(u) = \phi'(u) = 2au$ . Section 4.4.2 explains this results.

---

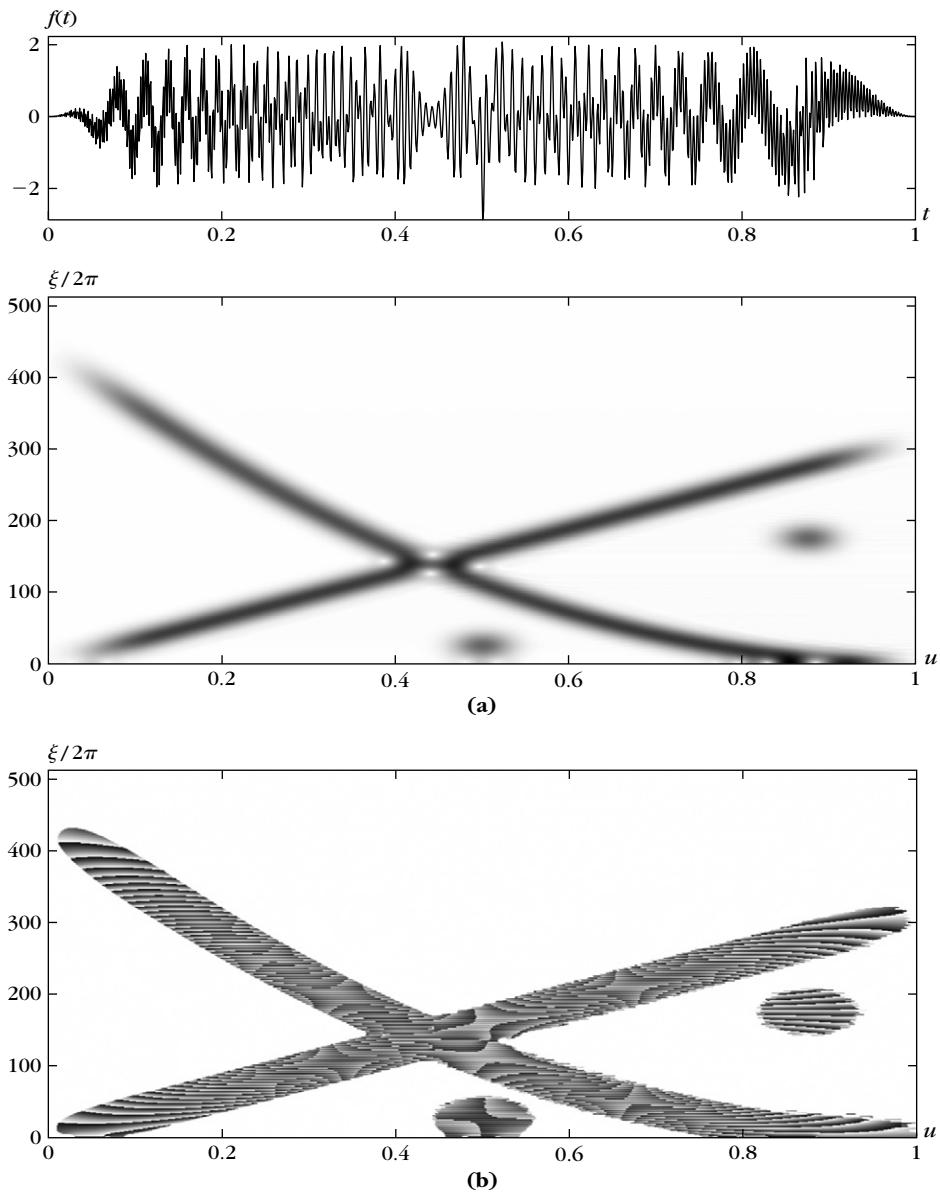
**EXAMPLE 4.5**

Figure 4.3 gives the spectrogram of a signal that includes a linear chirp, a quadratic chirp, and two modulated Gaussians. The spectrogram is computed with a Gaussian window dilated by  $\sigma = 0.05$ . As expected from (4.16), the linear chirp yields large amplitude coefficients along the trajectory of its instantaneous frequency, which is a straight line. The quadratic chirp yields large coefficients along a parabola. The two modulated Gaussians produce low- and high-frequency blobs at  $u = 0.5$  and  $u = 0.87$ .

---

**4.2.1 Completeness and Stability**

When the time-frequency indices  $(u, \xi)$  vary across  $\mathbb{R}^2$ , the Heisenberg boxes of the atoms  $g_{u,\xi}$  cover the whole time-frequency plane. One can expect therefore that  $f$

**FIGURE 4.3**

The signal includes a linear chirp with a frequency that increases, a quadratic chirp with a frequency that decreases, and two modulated Gaussian functions located at  $t = 0.5$  and  $t = 0.87$ . **(a)** Spectrogram  $P_{sf}(u, \xi)$ ; dark points indicate large-amplitude coefficients.

**(b)** Complex phase of  $Sf(u, \xi)$  in regions where the modulus  $P_{sf}(u, \xi)$  is nonzero.

can be recovered from its windowed Fourier transform  $Sf(u, \xi)$ . Theorem 4.1 gives a reconstruction formula and proves that the energy is conserved.

**Theorem 4.1.** If  $f \in L^2(\mathbb{R})$ , then

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} Sf(u, \xi) g(t-u) e^{i\xi t} d\xi du \quad (4.17)$$

and

$$\int_{-\infty}^{+\infty} |f(t)|^2 dt = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} |Sf(u, \xi)|^2 d\xi du. \quad (4.18)$$

**Proof.** The reconstruction formula (4.17) is proved first. Let us apply the Fourier Parseval formula (2.25) to the integral (4.17) with respect to the integration in  $u$ . The Fourier transform of  $f_\xi(u) = Sf(u, \xi)$  with respect to  $u$  is computed by observing that

$$Sf(u, \xi) = \exp(-iu\xi) \int_{-\infty}^{+\infty} f(t) g(t-u) \exp[i\xi(u-t)] dt = \exp(-iu\xi) f \star g_\xi(u),$$

where  $g_\xi(t) = g(t) \exp(i\xi t)$  because  $g(t) = g(-t)$ . Its Fourier transform therefore is

$$\hat{f}_\xi(\omega) = \hat{f}(\omega + \xi) \hat{g}_\xi(\omega + \xi) = \hat{f}(\omega + \xi) \hat{g}(\omega).$$

The Fourier transform of  $g(t-u)$  with respect to  $u$  is  $\hat{g}(\omega) \exp(-it\omega)$ . Thus,

$$\begin{aligned} & \frac{1}{2\pi} \left( \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} Sf(u, \xi) g(t-u) \exp(i\xi t) du \right) d\xi \\ &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} \left( \frac{1}{2\pi} \int_{-\infty}^{+\infty} \hat{f}(\omega + \xi) |\hat{g}(\omega)|^2 \exp(it(\omega + \xi)) d\omega \right) d\xi. \end{aligned}$$

If  $\hat{f} \in L^1(\mathbb{R})$ , we can apply the Fubini theorem (A.2) to reverse the integration order. The inverse Fourier transform proves that

$$\frac{1}{2\pi} \int_{-\infty}^{+\infty} \hat{f}(\omega + \xi) \exp(it(\omega + \xi)) d\xi = f(t).$$

Since  $\frac{1}{2\pi} \int_{-\infty}^{+\infty} |\hat{g}(\omega)|^2 d\omega = 1$ , we derive (4.17). If  $\hat{f} \notin L^1(\mathbb{R})$ , a density argument is used to verify this formula.

Let us now prove the energy conservation (4.18). Since the Fourier transform in  $u$  of  $Sf(u, \xi)$  is  $\hat{f}(\omega + \xi) \hat{g}(\omega)$ , the Plancherel formula (2.26) applied to the right side of (4.18) gives

$$\frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} |Sf(u, \xi)|^2 du d\xi = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \frac{1}{2\pi} \int_{-\infty}^{+\infty} |\hat{f}(\omega + \xi) \hat{g}(\omega)|^2 d\omega d\xi.$$

The Fubini theorem applies and the Plancherel formula proves that

$$\frac{1}{2\pi} \int_{-\infty}^{+\infty} |\hat{f}(\omega + \xi)|^2 d\xi = \|f\|^2,$$

which implies (4.18).  $\blacksquare$

The reconstruction formula (4.17) can be rewritten

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \langle f, g_{u,\xi} \rangle g_{u,\xi}(t) d\xi du. \quad (4.19)$$

It resembles the decomposition of a signal in an orthonormal basis; however, it is not because the functions  $\{g_{u,\xi}\}_{u,\xi \in \mathbb{R}^2}$  are very redundant in  $L^2(\mathbb{R})$ . The second equality (4.18) justifies the interpretation of the spectrogram  $P_S f(u, \xi) = |Sf(u, \xi)|^2$  as an energy density because its time-frequency sum equals the signal energy.

### **Reproducing Kernel**

A windowed Fourier transform represents a one-dimension signal  $f(t)$  by a two-dimensional function  $Sf(u, \xi)$ . Energy conservation proves that  $Sf \in L^2(\mathbb{R}^2)$ . Because  $Sf(u, \xi)$  is redundant, it is not true that any  $\Phi \in L^2(\mathbb{R}^2)$  is the windowed Fourier transform of some  $f \in L^2(\mathbb{R})$ . Theorem 4.2 gives a necessary and sufficient condition for such a function to be a windowed Fourier transform.

**Theorem 4.2.** Let  $\Phi \in L^2(\mathbb{R}^2)$ . There exists  $f \in L^2(\mathbb{R})$  such that  $\Phi(u, \xi) = Sf(u, \xi)$ , if and only if,

$$\Phi(u_0, \xi_0) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \Phi(u, \xi) K(u_0, u, \xi_0, \xi) du d\xi, \quad (4.20)$$

with

$$K(u_0, u, \xi_0, \xi) = \langle g_{u,\xi}, g_{u_0,\xi_0} \rangle. \quad (4.21)$$

**Proof.** Suppose that there exists  $f$  such that  $\Phi(u, \xi) = Sf(u, \xi)$ . Let us replace  $f$  with its reconstruction integral (4.17) in the windowed Fourier transform definition:

$$Sf(u_0, \xi_0) = \int_{-\infty}^{+\infty} \left( \frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} Sf(u, \xi) g_{u,\xi}(t) du d\xi \right) g_{u_0,\xi_0}^*(t) dt. \quad (4.22)$$

Inverting the integral on  $t$  with the integrals on  $u$  and  $\xi$  yields (4.20). To prove that the condition (4.20) is sufficient, we define  $f$  as in the reconstruction formula (4.17):

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \Phi(u, \xi) g(t-u) \exp(i\xi t) d\xi du$$

and show that (4.20) implies that  $\Phi(u, \xi) = Sf(u, \xi)$ .  $\blacksquare$

### Ambiguity Function

The reproducing kernel  $K(u_0, u, \xi_0, \xi)$  measures the time-frequency overlap of the two atoms  $g_{u,\xi}$  and  $g_{u_0,\xi_0}$ . The amplitude of  $K(u_0, u, \xi_0, \xi)$  decays with  $u_0 - u$  and  $\xi_0 - \xi$  at a rate that depends on the energy concentration of  $g$  and  $\hat{g}$ . Replacing  $g_{u,\xi}$  and  $g_{u_0,\xi_0}$  by their expression and making the change of variable  $v = t - (u + u_0)/2$  in the inner product integral (4.21) yields

$$K(u_0, u, \xi_0, \xi) = \exp\left(-\frac{i}{2}(\xi_0 - \xi)(u + u_0)\right) Ag(u_0 - u, \xi_0 - \xi), \quad (4.23)$$

where

$$Ag(\tau, \gamma) = \int_{-\infty}^{+\infty} g\left(v + \frac{\tau}{2}\right) g\left(v - \frac{\tau}{2}\right) e^{-i\gamma v} dv \quad (4.24)$$

is called the *ambiguity function* of  $g$ . Using the Parseval formula to replace this time integral with a Fourier integral gives

$$Ag(\tau, \gamma) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \hat{g}\left(\omega + \frac{\gamma}{2}\right) \hat{g}\left(\omega - \frac{\gamma}{2}\right) e^{i\tau\omega} d\omega. \quad (4.25)$$

The decay of the ambiguity function measures the spread of  $g$  in time and of  $\hat{g}$  in frequency. For example, if  $g$  has a support included in an interval of size  $T$ , then  $Ag(\tau, \omega) = 0$  for  $|\tau| \geq T/2$ . The integral (4.25) shows that the same result applies to the support of  $\hat{g}$ .

### 4.2.2 Choice of Window

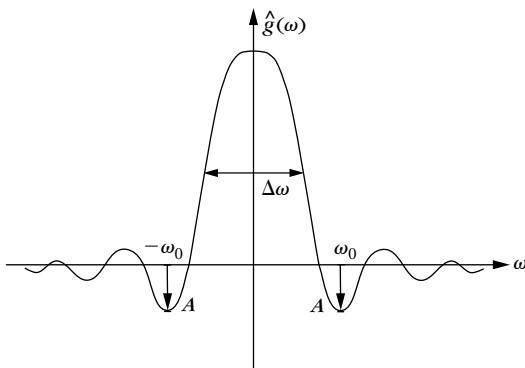
The resolution in time and frequency of the windowed Fourier transform depends on the spread of the window in time and frequency. This can be measured from the decay of the ambiguity function (4.24) or more simply from the area  $\sigma_t \sigma_\omega$  of the Heisenberg box. The uncertainty Theorem 2.6 proves that this area reaches the minimum value  $1/2$ , if, and only if,  $g$  is a Gaussian. The ambiguity function  $Ag(\tau, \gamma)$  is then a two-dimensional Gaussian.

### Window Scale

The time-frequency localization of  $g$  can be modified with a scaling. Suppose that  $g$  has a Heisenberg time and frequency width, respectively, equal to  $\sigma_t$  and  $\sigma_\omega$ . Let  $g_s(t) = s^{-1/2} g(t/s)$  be its dilation by  $s$ . A change of variables in the integrals (4.13) and (4.15) shows that the Heisenberg time and frequency width of  $g_s$  are, respectively,  $s\sigma_t$  and  $\sigma_\omega/s$ . The area of the Heisenberg box is not modified, but it is dilated by  $s$  in time and compressed by  $s$  in frequency. Similarly, a change of variable in the ambiguity integral (4.24) shows that the ambiguity function is dilated in time and frequency, respectively, by  $s$  and  $1/s$ :

$$Ag_s(\tau, \gamma) = Ag\left(\frac{\tau}{s}, s\gamma\right).$$

The choice of a particular scale  $s$  depends on the desired resolution trade-off between time and frequency.

**FIGURE 4.4**

The energy spread of  $\hat{g}$  is measured by its bandwidth  $\Delta\omega$  and the maximum amplitude  $A$  of the first side lobes, located at  $\omega = \pm\omega_0$ .

### **Finite Support**

In numerical applications,  $g$  must have a compact support. Theorem 2.7 proves that its Fourier transform  $\hat{g}$  necessarily has an infinite support. It is a symmetric function with a main lobe centered at  $\omega = 0$ , which decays to zero with oscillations. Figure 4.4 illustrates its behavior. To maximize the frequency resolution of the transform, we must concentrate the energy of  $\hat{g}$  near  $\omega = 0$ . The following three important parameters evaluate the spread of  $\hat{g}$ :

- The root mean-square bandwidth  $\Delta\omega$ , which is defined by

$$\frac{|\hat{g}(\Delta\omega/2)|^2}{|\hat{g}(0)|^2} = \frac{1}{2}.$$

- The maximum amplitude  $A$  of the first side lobes located at  $\omega = \pm\omega_0$  in Figure 4.4. It is measured in decibels:

$$A = 10 \log_{10} \frac{|\hat{g}(\omega_0)|^2}{|\hat{g}(0)|^2}.$$

- The polynomial exponent  $p$ , which gives the asymptotic decay of  $|\hat{g}(\omega)|$  for large frequencies:

$$|\hat{g}(\omega)| = O(\omega^{-p-1}). \quad (4.26)$$

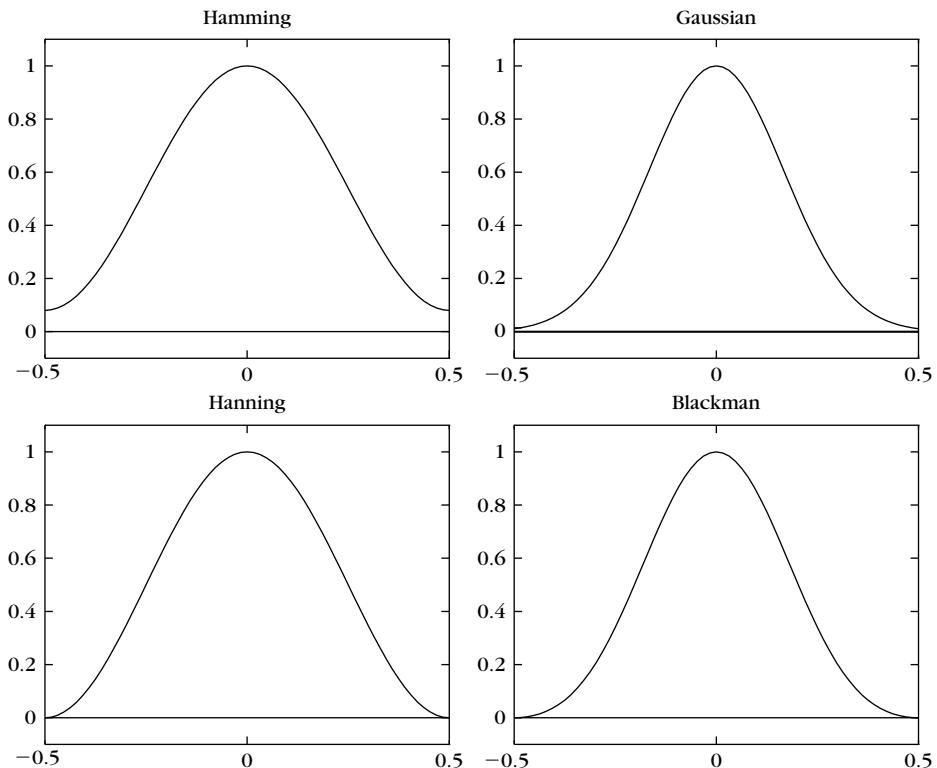
Table 4.1 gives the values of these three parameters for several windows  $g$  having a support restricted to  $[-1/2, 1/2]$  [293]. Figure 4.5 shows the graph of these windows.

To interpret the three frequency parameters, let us consider the spectrogram of a frequency tone  $f(t) = \exp(i\xi_0 t)$ . If  $\Delta\omega$  is small, then  $|Sf(u, \xi)|^2 = |\hat{g}(\xi - \xi_0)|^2$  has energy concentrated near  $\xi = \xi_0$ . The side lobes of  $\hat{g}$  create “shadows” at  $\xi = \xi_0 \pm \omega_0$ , which can be neglected if  $A$  is also small.

**Table 4.1** Frequency Parameters of Five Windows  $g$ 

Name	$g(t)$	$\Delta\omega$	A	p
Rectangle	1	0.89	-13db	0
Hamming	$0.54 + 0.46 \cos(2\pi t)$	1.36	-43db	0
Gaussian	$\exp(-18t^2)$	1.55	-55db	0
Hanning	$\cos^2(\pi t)$	1.44	-32db	2
Blackman	$0.42 + 0.5 \cos(2\pi t) + 0.08 \cos(4\pi t)$	1.68	-58db	2

Note: Supports are restricted to  $[-1/2, 1/2]$ . The windows are normalized so that  $g(0) = 1$  but  $\|g\| \neq 1$ .

**FIGURE 4.5**

Graphs of four windows  $g$  with supports that are  $[-1/2, 1/2]$ .

If the frequency tone is embedded in a signal that has other components of much higher energy at different frequencies, the tone can still be detected if  $\hat{g}(\omega - \xi)$  attenuates these components rapidly when  $|\omega - \xi|$  increases. This means that  $|\hat{g}(\omega)|$  has a rapid decay, and Theorem 2.5 proves that this decay depends on

the regularity of  $g$ . Property (4.26) is typically satisfied by windows that are  $p$  times differentiable.

### 4.2.3 Discrete Windowed Fourier Transform

The discretization and fast computation of the windowed Fourier transform follow the same ideas as the discretization of the Fourier transform described in Section 3.3. We consider discrete signals of period  $N$ . The window  $g[n]$  is chosen to be a symmetric discrete signal of period  $N$  with unit norm  $\|g\| = 1$ . Discrete windowed Fourier atoms are defined by

$$g_{m,l}[n] = g[n - m] \exp\left(\frac{i2\pi ln}{N}\right).$$

The discrete Fourier transform (DFT) of  $g_{m,l}$  is

$$\hat{g}_{m,l}[k] = \hat{g}[k - l] \exp\left(\frac{-i2\pi m(k - l)}{N}\right).$$

The discrete windowed Fourier transform of a signal  $f$  of period  $N$  is

$$Sf[m, l] = \langle f, g_{m,l} \rangle = \sum_{n=0}^{N-1} f[n] g[n - m] \exp\left(\frac{-i2\pi ln}{N}\right), \quad (4.27)$$

For each  $0 \leq m < N$ ,  $Sf[m, l]$  is calculated for  $0 \leq l < N$  with a DFT of  $f[n]g[n - m]$ . This is performed with  $N$  FFT procedures of size  $N$ , and therefore requires a total of  $O(N^2 \log_2 N)$  operations. Figure 4.3 is computed with this algorithm.

#### Inverse Transform

Theorem 4.3 discretizes the reconstruction formula and the energy conservation of Theorem 4.1.

**Theorem 4.3.** If  $f$  is a signal of period  $N$ , then

$$f[n] = \frac{1}{N} \sum_{m=0}^{N-1} \sum_{l=0}^{N-1} Sf[m, l] g[n - m] \exp\left(\frac{i2\pi ln}{N}\right) \quad (4.28)$$

and

$$\sum_{n=0}^{N-1} |f[n]|^2 = \frac{1}{N} \sum_{l=0}^{N-1} \sum_{m=0}^{N-1} |Sf[m, l]|^2. \quad (4.29)$$

This theorem is proved by applying the Parseval and Plancherel formulas of the discrete Fourier transform, exactly as in the proof of Theorem 4.1 (Exercise 4.1). The energy conservation (4.29) proves that this windowed Fourier transform defines a tight frame, as explained in Chapter 5. The reconstruction formula (4.28) is rewritten

$$f[n] = \frac{1}{N} \sum_{m=0}^{N-1} g[n - m] \sum_{l=0}^{N-1} Sf[m, l] \exp\left(\frac{i2\pi ln}{N}\right).$$

The second sum computes, for each  $0 \leq m < N$ , the inverse DFT of  $Sf[m, l]$  with respect to  $l$ . This is calculated with  $N$  FFT procedures, requiring a total of  $O(N^2 \log_2 N)$  operations.

A discrete windowed Fourier transform is an  $N^2$  image  $Sf[l, m]$  that is very redundant because it is entirely specified by a signal  $f$  of size  $N$ . The redundancy is characterized by a discrete reproducing kernel equation, which is the discrete equivalent of (4.20) (Exercise 4.1).

### 4.3 WAVELET TRANSFORMS

To analyze signal structures of very different sizes, it is necessary to use time-frequency atoms with different time supports. The wavelet transform decomposes signals over dilated and translated wavelets. A wavelet is a function  $\psi \in L^2(\mathbb{R})$  with a zero average:

$$\int_{-\infty}^{+\infty} \psi(t) dt = 0. \quad (4.30)$$

It is normalized  $\|\psi\| = 1$  and centered in the neighborhood of  $t = 0$ . A dictionary of time-frequency atoms is obtained by scaling  $\psi$  by  $s$  and translating it by  $u$ :

$$\mathcal{D} = \left\{ \psi_{u,s}(t) = \frac{1}{\sqrt{s}} \psi\left(\frac{t-u}{s}\right) \right\}_{u \in \mathbb{R}, s \in \mathbb{R}^+}$$

These atoms remain normalized:  $\|\psi_{u,s}\| = 1$ . The wavelet transform of  $f \in L^2(\mathbb{R})$  at time  $u$  and scale  $s$  is

$$Wf(u, s) = \langle f, \psi_{u,s} \rangle = \int_{-\infty}^{+\infty} f(t) \frac{1}{\sqrt{s}} \psi^*\left(\frac{t-u}{s}\right) dt. \quad (4.31)$$

#### *Linear Filtering*

The wavelet transform can be rewritten as a convolution product:

$$Wf(u, s) = \int_{-\infty}^{+\infty} f(t) \frac{1}{\sqrt{s}} \psi^*\left(\frac{t-u}{s}\right) dt = f \star \bar{\psi}_s(u), \quad (4.32)$$

with

$$\bar{\psi}_s(t) = \frac{1}{\sqrt{s}} \psi^*\left(\frac{-t}{s}\right).$$

The Fourier transform of  $\bar{\psi}_s(t)$  is

$$\widehat{\bar{\psi}}_s(\omega) = \sqrt{s} \widehat{\psi}^*(s\omega). \quad (4.33)$$

Since  $\hat{\psi}(0) = \int_{-\infty}^{+\infty} \psi(t) dt = 0$ , it appears that  $\hat{\psi}$  is the transfer function of a band-pass filter. The convolution (4.32) computes the wavelet transform with dilated band-pass filters.

### Analytic Versus Real Wavelets

Like a windowed Fourier transform, a wavelet transform can measure the time evolution of frequency transients. This requires using a complex analytic wavelet, which can separate amplitude and phase components. The properties of this analytic wavelet transform are described in Section 4.3.2, and its application to the measurement of instantaneous frequencies is explained in Section 4.4.3. In contrast, real wavelets are often used to detect sharp signal transitions. Section 4.3.1 introduces elementary properties of real wavelets, which are developed in Chapter 6.

#### 4.3.1 Real Wavelets

Suppose that  $\psi$  is a real wavelet. Since it has a zero average, the wavelet integral

$$Wf(u, s) = \int_{-\infty}^{+\infty} f(t) \frac{1}{\sqrt{s}} \psi^* \left( \frac{t-u}{s} \right) dt$$

measures the variation of  $f$  in a neighborhood of  $u$  proportional to  $s$ . Section 6.1.3 proves that when scale  $s$  goes to zero, the decay of the wavelet coefficient characterizes the regularity of  $f$  in the neighborhood of  $u$ . This has important applications for detecting transients and analyzing fractals. This section concentrates on the completeness and redundancy properties of real wavelet transforms.

#### EXAMPLE 4.6

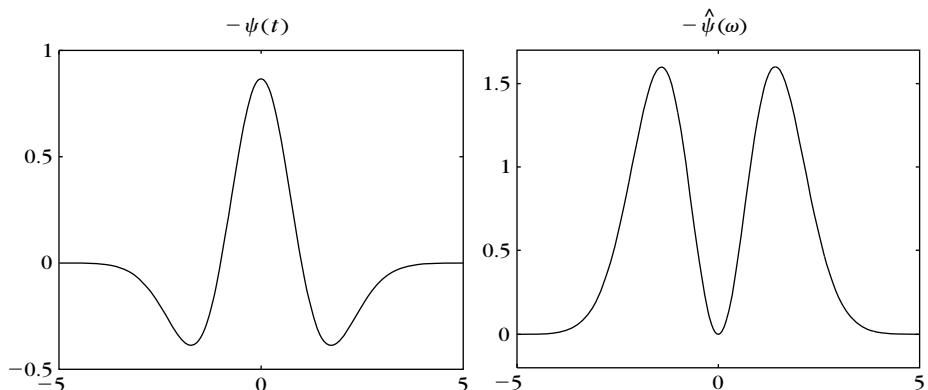
Wavelets equal to the second derivative of a Gaussian are called *Mexican hats*. They were first used in computer vision to detect multiscale edges [487]. The normalized Mexican hat wavelet is

$$\psi(t) = \frac{2}{\pi^{1/4} \sqrt{3\sigma}} \left( \frac{t^2}{\sigma^2} - 1 \right) \exp \left( \frac{-t^2}{2\sigma^2} \right). \quad (4.34)$$

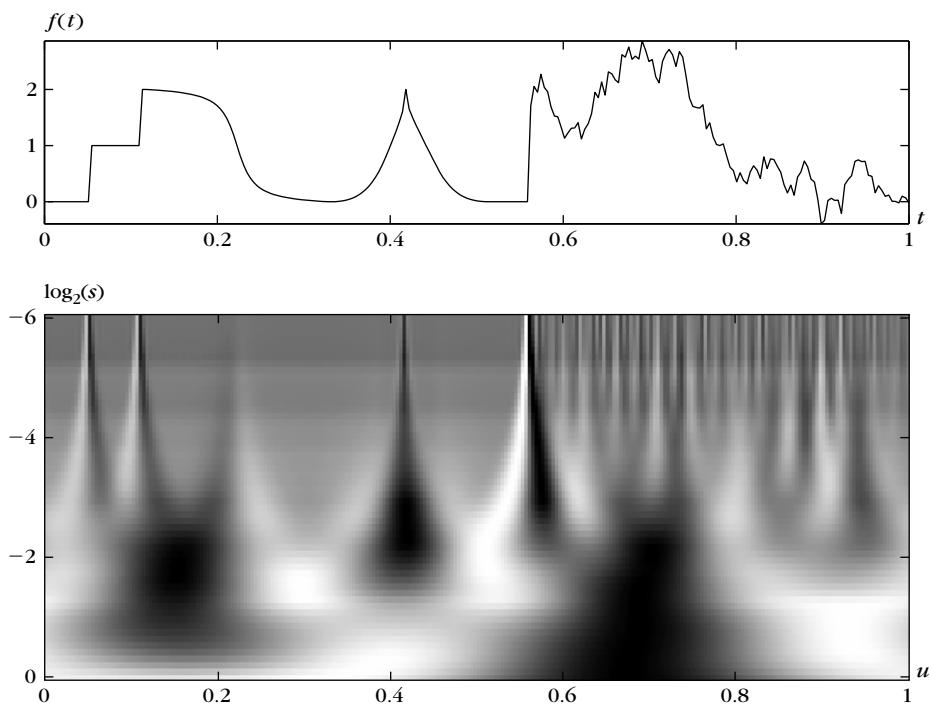
For  $\sigma = 1$ , Figure 4.6 plots  $-\psi$  and its Fourier transform:

$$\hat{\psi}(\omega) = \frac{-\sqrt{8} \sigma^{5/2} \pi^{1/4}}{\sqrt{3}} \omega^2 \exp \left( \frac{-\sigma^2 \omega^2}{2} \right). \quad (4.35)$$

Figure 4.7 shows the wavelet transform of a piecewise regular signal on the left and, almost everywhere, singular on the right. The maximum scale is smaller than 1 because the support of  $f$  is normalized to  $[0, 1]$ . The minimum scale is limited by the sampling interval of the discretized signal used in numerical calculations. When the scale decreases, the wavelet transform has a rapid decay to zero in the regions where the signal is regular. The isolated singularities on the left create cones of large-amplitude wavelet coefficients that converge to the locations of the singularities. This is further explained in Chapter 6.

**FIGURE 4.6**

Mexican-hat wavelet (4.34) for  $\sigma = 1$  and its Fourier transform.

**FIGURE 4.7**

Real wavelet transform  $Wf(u, s)$  computed with a Mexican-hat wavelet (4.34). The vertical axis represents  $\log_2 s$ . Black, gray, and white points correspond, respectively, to positive, zero, and negative wavelet coefficients.

A real wavelet transform is complete and maintains an energy conservation as long as the wavelet satisfies a weak admissibility condition, specified by Theorem 4.4. This theorem was first proved in 1964 by the mathematician Calderón [132] from a different point of view. Wavelets did not appear as such, but Calderón defines a wavelet transform as a convolution operator that decomposes the identity. Grossmann and Morlet [288] were not aware of Calderón's work when they proved the same formula for signal processing.

**Theorem 4.4:** *Calderón, Grossmann and Morlet.* Let  $\psi \in \mathbf{L}^2(\mathbb{R})$  be a real function such that

$$C_\psi = \int_0^{+\infty} \frac{|\hat{\psi}(\omega)|^2}{\omega} d\omega < +\infty. \quad (4.36)$$

Any  $f \in \mathbf{L}^2(\mathbb{R})$  satisfies

$$f(t) = \frac{1}{C_\psi} \int_0^{+\infty} \int_{-\infty}^{+\infty} wf(u, s) \frac{1}{\sqrt{s}} \psi\left(\frac{t-u}{s}\right) du \frac{ds}{s^2}, \quad (4.37)$$

and

$$\int_{-\infty}^{+\infty} |f(t)|^2 dt = \frac{1}{C_\psi} \int_0^{+\infty} \int_{-\infty}^{+\infty} |wf(u, s)|^2 du \frac{ds}{s^2}. \quad (4.38)$$

**Proof.** The proof of (4.38) is almost identical to the proof of (4.18). Let us concentrate on the proof of (4.37). The right integral  $b(t)$  of (4.37) can be rewritten as a sum of convolutions. Inserting  $wf(u, s) = f \star \bar{\psi}_s(u)$  with  $\psi_s(t) = s^{-1/2} \psi(t/s)$  yields

$$\begin{aligned} b(t) &= \frac{1}{C_\psi} \int_0^{+\infty} wf(., s) \star \psi_s(t) \frac{ds}{s^2} \\ &= \frac{1}{C_\psi} \int_0^{+\infty} f \star \bar{\psi}_s \star \psi_s(t) \frac{ds}{s^2}. \end{aligned} \quad (4.39)$$

The “.” indicates the variable over which the convolution is calculated. We prove that  $b = f$  by showing that their Fourier transforms are equal. The Fourier transform of  $b$  is

$$\hat{b}(\omega) = \frac{1}{C_\psi} \int_0^{+\infty} \hat{f}(\omega) \sqrt{s} \hat{\psi}^*(s\omega) \sqrt{s} \hat{\psi}(s\omega) \frac{ds}{s^2} = \frac{\hat{f}(\omega)}{C_\psi} \int_0^{+\infty} |\hat{\psi}(s\omega)|^2 \frac{ds}{s}.$$

Since  $\psi$  is real we know that  $|\hat{\psi}(-\omega)|^2 = |\hat{\psi}(\omega)|^2$ . The change of variable  $\xi = s\omega$  thus proves that

$$\hat{b}(\omega) = \frac{1}{C_\psi} \hat{f}(\omega) \int_0^{+\infty} \frac{|\hat{\psi}(\xi)|^2}{\xi} d\xi = \hat{f}(\omega).$$

■

The theorem hypothesis

$$C_\psi = \int_0^{+\infty} \frac{|\hat{\psi}(\omega)|^2}{\omega} d\omega < +\infty$$

is called the wavelet *admissibility condition*. To guarantee that this integral is finite, we must ensure that  $\hat{\psi}(0) = 0$ , which explains why wavelets must have a zero average. This condition is nearly sufficient. If  $\hat{\psi}(0) = 0$  and  $\hat{\psi}(\omega)$  is continuously differentiable, then the admissibility condition is satisfied. One can verify that  $\hat{\psi}(\omega)$  is continuously differentiable if  $\psi$  has a sufficient time decay:

$$\int_{-\infty}^{+\infty} (1 + |t|) |\psi(t)| dt < +\infty.$$

### Reproducing Kernel

Like a windowed Fourier transform, a wavelet transform is a redundant representation with a redundancy characterized by a reproducing kernel equation. Inserting the reconstruction formula (4.37) into the definition of the wavelet transform yields

$$wf(u_0, s_0) = \int_{-\infty}^{+\infty} \left( \frac{1}{C_\psi} \int_0^{+\infty} \int_{-\infty}^{+\infty} wf(u, s) \psi_{u,s}(t) du \frac{ds}{s^2} \right) \psi_{u_0,s_0}^*(t) dt.$$

Interchanging these integrals gives

$$wf(u_0, s_0) = \frac{1}{C_\psi} \int_{-\infty}^{+\infty} K(u, u_0, s, s_0) wf(u, s) du \frac{ds}{s^2}, \quad (4.40)$$

with

$$K(u_0, u, s_0, s) = \langle \psi_{u,s}, \psi_{u_0,s_0} \rangle. \quad (4.41)$$

The reproducing kernel  $K(u_0, u, s_0, s)$  measures the correlation of two wavelets,  $\psi_{u,s}$  and  $\psi_{u_0,s_0}$ . The reader can verify that any function  $\Phi(u, s)$  is the wavelet transform of some  $f \in L^2(\mathbb{R})$  if and only if it satisfies the reproducing kernel equation (4.40).

### Scaling Function

When  $wf(u, s)$  is known only for  $s < s_0$ , to recover  $f$  we need a complement of information that corresponds to  $wf(u, s)$  for  $s > s_0$ . This is obtained by introducing a *scaling function*  $\phi$  that is an aggregation of wavelets at scales larger than 1. The modulus of its Fourier transform is defined by

$$|\hat{\phi}(\omega)|^2 = \int_1^{+\infty} |\hat{\psi}(s\omega)|^2 \frac{ds}{s} = \int_\omega^{+\infty} \frac{|\hat{\psi}(\xi)|^2}{\xi} d\xi, \quad (4.42)$$

and the complex phase of  $\hat{\phi}(\omega)$  can be arbitrarily chosen. One can verify that  $\|\phi\| = 1$ , and we can derive from the admissibility condition (4.36) that

$$\lim_{\omega \rightarrow 0} |\hat{\phi}(\omega)|^2 = C_\psi. \quad (4.43)$$

The scaling function therefore can be interpreted as the impulse response of a low-pass filter. Let us denote

$$\phi_s(t) = \frac{1}{\sqrt{s}} \phi\left(\frac{t}{s}\right) \text{ and } \bar{\phi}_s(t) = \phi_s^*(-t).$$

The low-frequency approximation of  $f$  at scale  $s$  is

$$If(u, s) = \left\langle f(t), \frac{1}{\sqrt{s}} \phi\left(\frac{t-u}{s}\right) \right\rangle = f \star \bar{\phi}_s(u). \quad (4.44)$$

With a minor modification of Theorem 4.4's proof it can be shown that (Exercise 4.3)

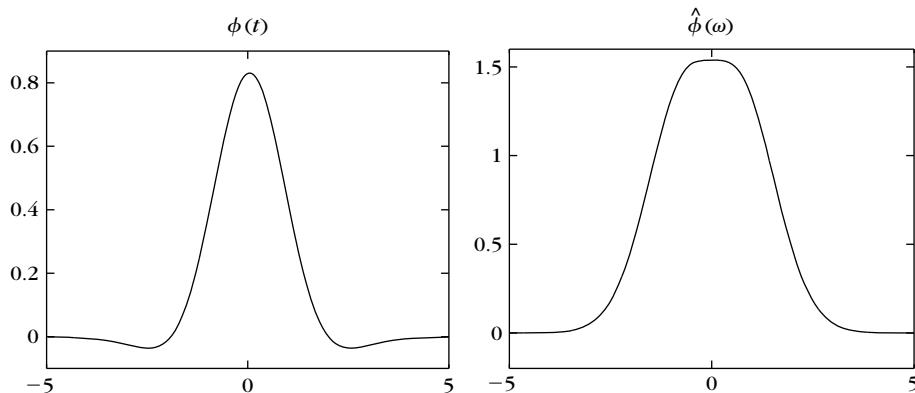
$$f(t) = \frac{1}{C_\psi} \int_0^{s_0} wf(., s) \star \psi_s(t) \frac{ds}{s^2} + \frac{1}{C_\psi s_0} If(., s_0) \star \phi_{s_0}(t). \quad (4.45)$$

### EXAMPLE 4.7

If  $\psi$  is the second-order derivative of a Gaussian with a Fourier transform given by (4.35), then the integration (4.42) yields

$$\hat{\phi}(\omega) = \frac{2\sigma^{3/2}\pi^{1/4}}{\sqrt{3}} \sqrt{\omega^2 + \frac{1}{\sigma^2}} \exp\left(-\frac{\sigma^2\omega^2}{2}\right). \quad (4.46)$$

Figure 4.8 displays  $\phi$  and  $\hat{\phi}$  for  $\sigma = 1$ .



**FIGURE 4.8**

Scaling function associated to a Mexican-hat wavelet and its Fourier transform calculated with (4.46).

### 4.3.2 Analytic Wavelets

To analyze the time evolution of frequency tones, it is necessary to use an analytic wavelet to separate the phase and amplitude information of signals. The properties of the resulting analytic wavelet transform are studied next.

**Analytic Signal**

A function  $f_a \in \mathbf{L}^2(\mathbb{R})$  is said to be *analytic* if its Fourier transform is zero for negative frequencies:

$$\hat{f}_a(\omega) = 0 \quad \text{if } \omega < 0.$$

An analytic function is necessarily complex but is entirely characterized by its real part. Indeed, the Fourier transform of its real part  $f = \operatorname{Re}[f_a]$  is

$$\hat{f}(\omega) = \frac{\hat{f}_a(\omega) + \hat{f}_a^*(-\omega)}{2},$$

and this relation can be inverted:

$$\hat{f}_a(\omega) = \begin{cases} 2\hat{f}(\omega) & \text{if } \omega \geq 0 \\ 0 & \text{if } \omega < 0 \end{cases} \quad (4.47)$$

The analytic part  $f_a(t)$  of a signal  $f(t)$  is the inverse Fourier transform of  $\hat{f}_a(\omega)$  defined by (4.47).

**Discrete Analytic Part**

The analytic part  $f_a[n]$  of a discrete signal  $f[n]$  of size  $N$  is also computed by setting the negative frequency components of its discrete Fourier transform to zero. The Fourier transform values at  $k=0$  and  $k=N/2$  must be carefully adjusted so that  $\operatorname{Re}[f_a] = f$  (Exercise 3.4):

$$\hat{f}_a[k] = \begin{cases} \hat{f}[k] & \text{if } k = 0, N/2 \\ 2\hat{f}[k] & \text{if } 0 < k < N/2 \\ 0 & \text{if } N/2 < k < N \end{cases} \quad (4.48)$$

We obtain  $f_a[n]$  by computing the inverse DFT.

**EXAMPLE 4.8**

The Fourier transform of

$$f(t) = a \cos(\omega_0 t + \phi) = \frac{a}{2} \left( \exp[i(\omega_0 t + \phi)] + \exp[-i(\omega_0 t + \phi)] \right)$$

is

$$\hat{f}(\omega) = \pi a \left( \exp(i\phi) \delta(\omega - \omega_0) + \exp(-i\phi) \delta(\omega + \omega_0) \right).$$

The Fourier transform of the analytic part computed with (4.47) is  $\hat{f}_a(\omega) = 2\pi a \exp(i\phi) \delta(\omega - \omega_0)$  and therefore

$$f_a(t) = a \exp[i(\omega_0 t + \phi)]. \quad (4.49)$$

### Time-Frequency Resolution

An analytic wavelet transform is calculated with an analytic wavelet  $\psi$ :

$$wf(u, s) = \langle f, \psi_{u,s} \rangle = \int_{-\infty}^{+\infty} f(t) \frac{1}{\sqrt{s}} \psi^* \left( \frac{t-u}{s} \right) dt. \quad (4.50)$$

Its time-frequency resolution depends on the time-frequency spread of the wavelet atoms  $\psi_{u,s}$ . We suppose that  $\psi$  is centered at 0, which implies that  $\psi_{u,s}$  is centered at  $t=u$ . With the change of variable  $v = \frac{t-u}{s}$ , we verify that

$$\int_{-\infty}^{+\infty} (t-u)^2 |\psi_{u,s}(t)|^2 dt = s^2 \sigma_t^2, \quad (4.51)$$

with  $\sigma_t^2 = \int_{-\infty}^{+\infty} t^2 |\psi(t)|^2 dt$ . Since  $\hat{\psi}(\omega)$  is zero at negative frequencies, the center frequency  $\eta$  of  $\hat{\psi}$  is

$$\eta = \frac{1}{2\pi} \int_0^{+\infty} \omega |\hat{\psi}(\omega)|^2 d\omega. \quad (4.52)$$

The Fourier transform of  $\psi_{u,s}$  is a dilation of  $\hat{\psi}$  by  $1/s$ :

$$\hat{\psi}_{u,s}(\omega) = \sqrt{s} \hat{\psi}(s\omega) \exp(-i\omega u). \quad (4.53)$$

Its center frequency therefore is  $\eta/s$ . The energy spread of  $\hat{\psi}_{u,s}$  around  $\eta/s$  is

$$\frac{1}{2\pi} \int_0^{+\infty} \left( \omega - \frac{\eta}{s} \right)^2 \left| \hat{\psi}_{u,s}(\omega) \right|^2 d\omega = \frac{\sigma_\omega^2}{s^2}, \quad (4.54)$$

with

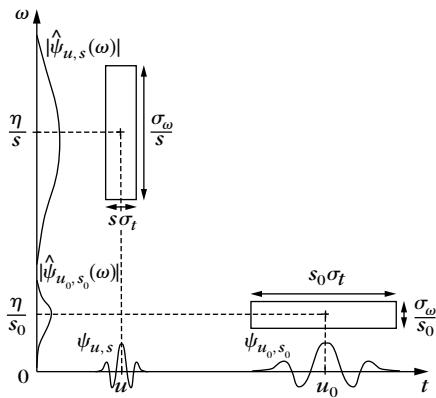
$$\sigma_\omega^2 = \frac{1}{2\pi} \int_0^{+\infty} (\omega - \eta)^2 |\hat{\psi}(\omega)|^2 d\omega.$$

Thus, the energy spread of a wavelet time-frequency atom  $\psi_{u,s}$  corresponds to a Heisenberg box centered at  $(u, \eta/s)$ , of size  $s\sigma_t$  along time and  $\sigma_\omega/s$  along frequency. The area of the rectangle remains equal to  $\sigma_t \sigma_\omega$  at all scales but the resolution in time and frequency depends on  $s$ , as illustrated in Figure 4.9.

An analytic wavelet transform defines a local time-frequency energy density  $P_W f$ , which measures the energy of  $f$  in the Heisenberg box of each wavelet  $\psi_{u,s}$  centered at  $(u, \xi = \eta/s)$ :

$$P_W f(u, \xi) = |W f(u, s)|^2 = \left| W f \left( u, \frac{\eta}{\xi} \right) \right|^2. \quad (4.55)$$

This energy density is called a *scalogram*.

**FIGURE 4.9**

Heisenberg boxes of two wavelets. Smaller scales decrease the time spread but increase the frequency support, which is shifted toward higher frequencies.

### Completeness

An analytic wavelet transform of  $f$  depends only on its analytic part  $f_a$ . Theorem 4.5 derives a reconstruction formula and proves that energy is conserved for real signals.

**Theorem 4.5.** For any  $f \in \mathbf{L}^2(\mathbb{R})$ ,

$$Wf(u, s) = \frac{1}{2} Wf_a(u, s). \quad (4.56)$$

If  $C_\psi = \int_0^{+\infty} \omega^{-1} |\hat{\psi}(\omega)|^2 d\omega < +\infty$  and  $f$  is real, then

$$f(t) = \frac{2}{C_\psi} \operatorname{Re} \left[ \int_0^{+\infty} \int_{-\infty}^{+\infty} Wf(u, s) \psi_s(t-u) du \frac{ds}{s^2} \right], \quad (4.57)$$

and

$$\|f\|^2 = \frac{2}{C_\psi} \int_0^{+\infty} \int_{-\infty}^{+\infty} |Wf(u, s)|^2 du \frac{ds}{s^2}. \quad (4.58)$$

**Proof.** Let us first prove (4.56). The Fourier transform with respect to  $u$  of

$$f_s(u) = wf(u, s) = f \star \bar{\psi}_s(u)$$

is

$$\hat{f}_s(\omega) = \hat{f}(\omega) \sqrt{s} \hat{\psi}^*(s\omega).$$

Since  $\hat{\psi}(\omega) = 0$  at negative frequencies, and  $\hat{f}_a(\omega) = 2\hat{f}(\omega)$  for  $\omega \geq 0$ , we derive that

$$\hat{f}_s(\omega) = \frac{1}{2} \hat{f}_a(\omega) \sqrt{s} \hat{\psi}^*(s\omega),$$

which is the Fourier transform of (4.56).

With the same derivations as in the proof of (4.37), one can verify that the inverse wavelet formula reconstructs the analytic part of  $f$ :

$$f_a(t) = \frac{1}{C_\psi} \int_0^{+\infty} \int_{-\infty}^{+\infty} w f_a(u, s) \psi_s(t-u) \frac{ds}{s^2} du. \quad (4.59)$$

Since  $f = \operatorname{Re}[f_a]$ , inserting (4.56) proves (4.57).

An energy conservation for the analytic part  $f_a$  is proved as in (4.38) by applying the Plancherel formula:

$$\int_{-\infty}^{+\infty} |f_a(t)|^2 dt = \frac{1}{C_\psi} \int_0^{+\infty} \int_{-\infty}^{+\infty} |W f_a(u, s)|^2 du \frac{ds}{s^2}.$$

Since  $W f_a(u, s) = 2 W f(u, s)$  and  $\|f_a\|^2 = 2 \|f\|^2$ , equation (4.58) follows. ■

If  $f$  is real, the change of variable  $\xi = 1/s$  in the energy conservation (4.58) proves that

$$\|f\|^2 = \frac{2}{C_\psi} \int_0^{+\infty} \int_{-\infty}^{+\infty} P_W f(u, \xi) du d\xi.$$

It justifies the interpretation of a scalogram as a time-frequency energy density.

### Wavelet Modulated Windows

An analytic wavelet can be constructed with a frequency modulation of a real and symmetric window  $g$ . The Fourier transform of

$$\psi(t) = g(t) \exp(i\eta t) \quad (4.60)$$

is  $\hat{\psi}(\omega) = \hat{g}(\omega - \eta)$ . If  $\hat{g}(\omega) = 0$  for  $|\omega| > \eta$ , then  $\hat{\psi}(\omega) = 0$  for  $\omega < 0$ . Therefore,  $\psi$  is analytic, as shown in Figure 4.10. Since  $g$  is real and even,  $\hat{g}$  is also real and symmetric. The center frequency of  $\hat{\psi}$  is therefore  $\eta$  and

$$|\hat{\psi}(\eta)| = \sup_{\omega \in \mathbb{R}} |\hat{\psi}(\omega)| = \hat{g}(0). \quad (4.61)$$

A *Gabor wavelet*  $\psi(t) = g(t) e^{i\eta t}$  is obtained with a Gaussian window:

$$g(t) = \frac{1}{(\sigma^2 \pi)^{1/4}} \exp\left(\frac{-t^2}{2\sigma^2}\right). \quad (4.62)$$

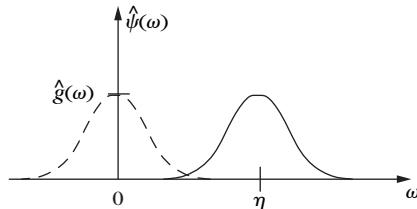


FIGURE 4.10

Fourier transform  $\hat{\psi}(\omega)$  of a wavelet  $\psi(t) = g(t) \exp(i\eta t)$ .

The Fourier transform of this window is  $\hat{g}(\omega) = (4\pi\sigma^2)^{1/4} \exp(-\sigma^2\omega^2/2)$ . If  $\sigma^2\eta^2 \gg 1$  then  $\hat{g}(\omega) \approx 0$  for  $|\omega| > \eta$ . Thus, such Gabor wavelets are considered to be approximately analytic.

---

**EXAMPLE 4.9**

The wavelet transform of  $f(t) = a \exp(i\omega_0 t)$  is

$$Wf(u, s) = a\sqrt{s}\hat{\psi}^*(s\omega_0) \exp(i\omega_0 u) = a\sqrt{s}\hat{g}(s\omega_0 - \eta) \exp(i\omega_0 u).$$

Observe that the normalized scalogram is maximum at  $\xi = \omega_0$ :

$$\frac{\xi}{\eta}P_W f(u, \xi) = \frac{1}{s}|Wf(u, s)|^2 = a^2 \left| \hat{g}\left(\eta\left(\frac{\omega_0}{\xi} - 1\right)\right) \right|^2.$$


---

**EXAMPLE 4.10**

The wavelet transform of a linear chirp  $f(t) = \exp(iat^2) = \exp[i\phi(t)]$  is computed for a Gabor wavelet with a Gaussian window given in (4.62). By using the Fourier transform of Gaussian chirps (2.34), one can verify that

$$\frac{|Wf(u, s)|^2}{s} = \left( \frac{4\pi\sigma^2}{1 + 4s^2a^2\sigma^4} \right)^{1/2} \exp\left( \frac{-\sigma^2}{1 + 4a^2s^4\sigma^4} (\eta - 2asu)^2 \right).$$

As long as  $4a^2s^4\sigma^4 \ll 1$  at a fixed time  $u$ , the renormalized scalogram  $\eta^{-1}\xi P_W f(u, \xi)$  is a Gaussian function of  $s$  that reaches its maximum at

$$\xi(u) = \frac{\eta}{s(u)} = \phi'(u) = 2au. \quad (4.63)$$

Section 4.4.3 explains why the amplitude is maximum at the instantaneous frequency  $\phi'(u)$ .

---

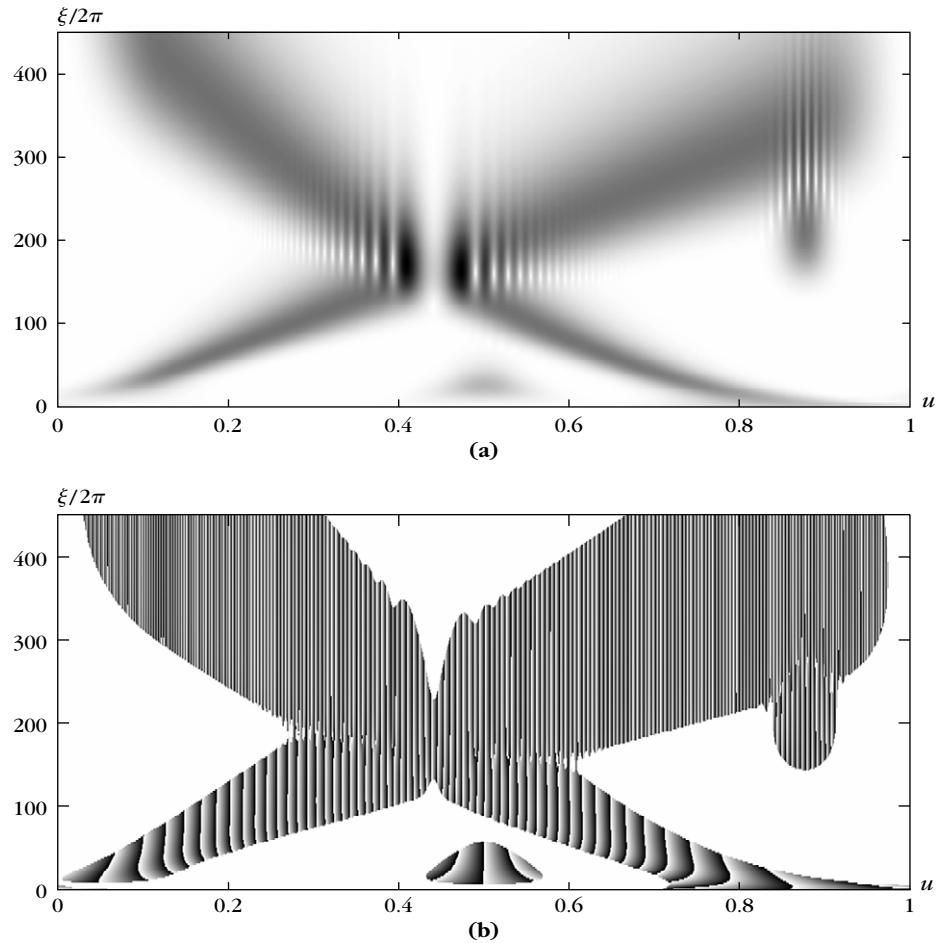
**EXAMPLE 4.11**

Figure 4.11 displays the normalized scalogram  $\eta^{-1}\xi P_W f(u, \xi)$ , and the complex phase  $\Theta_W(u, \xi)$  of  $Wf(u, s)$ , for the signal  $f$  of Figure 4.3. The frequency bandwidth of wavelet atoms is proportional to  $1/s = \xi/\eta$ . The frequency resolution of the scalogram is therefore finer than the spectrogram at low frequencies but coarser than the spectrogram at higher frequencies. This explains why the wavelet transform produces interference patterns between the high-frequency Gabor function at the abscissa  $t = 0.87$  and the quadratic chirp at the same location, whereas the spectrogram in Figure 4.3 separates them well.

---

### 4.3.3 Discrete Wavelets

Let  $\bar{f}(t)$  be a continuous time signal defined over  $[0, 1]$ . Let  $f[n]$  be the discrete signal obtained by a low-pass filtering of  $\bar{f}$  and uniform sampling at intervals  $N^{-1}$ .

**FIGURE 4.11**

**(a)** Normalized scalogram  $\eta^{-1}\xi P_{wf}(u, \xi)$  computed from the signal in Figure 4.3; dark points indicate large-amplitude coefficients. **(b)** Complex phase  $\Theta_W(u, \xi)$  of  $Wf(u, \eta/\xi)$ , where the modulus is nonzero.

Its discrete wavelet transform can only be calculated at scales  $N^{-1} < s < 1$ , as shown in Figure 4.7. It is calculated for  $s = a^j$ , with  $a = 2^{1/v}$ , which provides  $v$  intermediate scales in each octave  $[2^j, 2^{j+1})$ .

Let  $\psi(t)$  be a wavelet with a support included in  $[-K/2, K/2]$ . For  $1 \leq a^j \leq N K^{-1}$ , a discrete wavelet scaled by  $a^j$  is defined by

$$\psi_j[n] = \frac{1}{\sqrt{a^j}} \psi\left(\frac{n}{a^j}\right).$$

This discrete wavelet has  $Ka^j$  nonzero values on  $[-N/2, N/2]$ . The scale  $a^j$  is larger than 1; otherwise, the sampling interval may be larger than the wavelet support.

### **Fast Transform**

To avoid border problems, we treat  $f[n]$  and the wavelets  $\psi_j[n]$  as periodic signals of period  $N$ . The discrete wavelet transform can then be written as a circular convolution with  $\bar{\psi}_j[n] = \psi_j^*[-n]$ :

$$Wf[n, a^j] = \sum_{m=0}^{N-1} f[m] \psi_j^*[m-n] = f \circledast \bar{\psi}_j[n]. \quad (4.64)$$

This circular convolution is calculated with the fast Fourier transform algorithm, which requires  $O(N \log_2 N)$  operations. If  $a = 2^{1/v}$ , there are  $v \log_2(N/(2K))$  scales  $a^j \in [2N^{-1}, K^{-1}]$ . The total number of operations to compute the wavelet transform over all scales therefore is  $O(vN(\log_2 N)^2)$  [408].

To compute the scalogram  $P_W[n, \xi] = |Wf[n, \frac{n}{\xi}]|^2$ , we calculate  $Wf[n, s]$  at any scale  $s$  with a parabola interpolation. Let  $j$  be the closest integer to  $\log_2 s / \log_2 a$ , and  $p(x)$  be the parabola such that

$$p(j-1) = Wf[n, a^{j-1}], \quad p(j) = Wf[n, a^j], \quad p(j+1) = Wf[n, a^{j+1}].$$

A second-order interpolation computes

$$Wf[n, s] = p\left(\frac{\log_2 s}{\log_2 a}\right).$$

Parabolic interpolations are used instead of linear interpolations in order to more precisely locate the ridges defined in Section 4.4.3.

### **Discrete Scaling Filter**

A wavelet transform computed up to a scale  $a^j$  is not a complete signal representation. It is necessary to add low frequencies  $Lf[n, a^j]$  corresponding to scales larger than  $a^j$ . A discrete and periodic scaling filter is computed by sampling the scaling function  $\phi(t)$  defined in (4.42):

$$\phi_J[n] = \frac{1}{\sqrt{a^j}} \phi\left(\frac{n}{a^j}\right) \quad \text{for } n \in [-N/2, N/2].$$

Let  $\bar{\phi}_J[n] = \phi_J^*[-n]$ ; the low frequencies are carried by

$$Lf[n, a^j] = \sum_{m=0}^{N-1} f[m] \phi_J^*[m-n] = f \circledast \bar{\phi}_J[n]. \quad (4.65)$$

### **Reconstruction**

An approximate inverse wavelet transform is implemented by discretizing the integral (4.45). Suppose that  $a^I = 1$  is the finest scale. Since  $ds/s^2 = d \log_e s/s$  and the discrete wavelet transform is computed along an exponential scale sequence  $\{a^j\}_j$  with a logarithmic increment  $d \log_e s = \log_e a$ , we obtain

$$f[n] \approx \frac{\log_e a}{C_\psi} \sum_{j=I}^J \frac{1}{a^j} Wf[., a^j] \circledast \psi_j[n] + \frac{1}{C_\psi a^J} Lf[., a^J] \circledast \phi_J[n]. \quad (4.66)$$

The “.” indicates the variable over which the convolution is calculated. These circular convolutions are calculated using the FFT, with  $O(vN(\log_2 N)^2)$  operations.

Analytic wavelet transforms are often computed over real signals  $f[n]$  that have no energy at low frequencies. The scaling filter component is then negligible. Theorem 4.5 shows that

$$f[n] \approx \frac{2 \log_e a}{C_\psi} \operatorname{Re} \left( \sum_{j=I}^J \frac{1}{a^j} Wf[., a^j] \circledast \psi_j[n] \right). \quad (4.67)$$

The error introduced by the discretization of scales decreases when the number  $v$  of voices per octave increases. However, the approximation of continuous time convolutions with discrete convolutions also creates high-frequency errors. Perfect reconstructions are obtained with a more careful design of the reconstruction filters (Exercise 4.3). Section 5.2.2 describes an exact inverse wavelet transform computed at dyadic scales  $a^j = 2^j$ .

## **4.4 TIME-FREQUENCY GEOMETRY OF INSTANTANEOUS FREQUENCIES**

When listening to music, we perceive several frequencies that change with time. In music, it is associated to the geometric perception of “movements.” This notion of instantaneous frequency remains to be defined. The time variation of several instantaneous frequencies is measured with local maxima of windowed Fourier transforms and wavelet transforms. They define a geometric time-frequency support from which signal approximations are recovered. Audio processing is implemented by modifying this time-frequency support.

### **4.4.1 Analytic Instantaneous Frequency**

The notion of instantaneous frequency is not well defined. It can, however, be uniquely specified with the signal analytic part. A cosine modulation

$$f(t) = a \cos(\omega_0 t + \theta_0) = a \cos \theta(t)$$

has a frequency  $\omega_0$  that is the derivative of the phase  $\theta(t) = \omega_0 t + \theta_0$ . To generalize this notion, real signals  $f$  are written as an amplitude  $a(t)$  modulated with a time-varying phase  $\theta(t)$ :

$$f(t) = a(t) \cos \theta(t), \quad \text{with } a(t) \geq 0. \quad (4.68)$$

The *instantaneous frequency* can be defined as a positive derivative of the phase:

$$\omega(t) = \theta'(t) \geq 0.$$

The derivative is chosen to be positive by adapting the sign of  $\theta(t)$ . However, for a given  $f(t)$ , there are many possible choices of  $a(t)$  and  $\theta(t)$  to satisfy (4.68), so  $\omega(t)$  is not uniquely defined relative to  $f$ .

A particular decomposition (4.68) is obtained from the analytic part  $f_a$  of  $f$ , which has a Fourier transform defined in (4.47) by

$$\hat{f}_a(\omega) = \begin{cases} 2\hat{f}(\omega) & \text{if } \omega \geq 0 \\ 0 & \text{if } \omega < 0 \end{cases} \quad (4.69)$$

This complex signal is represented by separating the modulus and the complex phase:

$$f_a(t) = a(t) \exp[i\theta(t)]. \quad (4.70)$$

Since  $f = \operatorname{Re}[f_a]$ , it follows that

$$f(t) = a(t) \cos \theta(t).$$

We call  $a(t)$  the *analytic amplitude* of  $f(t)$  and  $\theta'(t)$  its *instantaneous frequency*; they are uniquely defined.

### EXAMPLE 4.12

If  $f(t) = a(t) \cos(\omega_0 t + \theta_0)$ , then

$$\hat{f}(\omega) = \frac{1}{2} \left( \exp(i\theta_0) \hat{a}(\omega - \omega_0) + \exp(-i\theta_0) \hat{a}(\omega + \omega_0) \right).$$

If the variations of  $a(t)$  are slow compared to the period  $2\pi/\omega_0$ , which is achieved by requiring that the support of  $\hat{a}$  be included in  $[-\omega_0, \omega_0]$ , then

$$\hat{f}_a(\omega) = \hat{a}(\omega - \omega_0) \exp(i\theta_0),$$

$$\text{so } f_a(t) = a(t) \exp[i(\omega_0 t + \theta_0)].$$


---

If a signal  $f$  is the sum of two sinusoidal waves:

$$f(t) = a \cos(\omega_1 t) + a \cos(\omega_2 t),$$

then

$$f_a(t) = a \exp(i\omega_1 t) + a \exp(i\omega_2 t) = 2a \cos\left(\frac{1}{2}(\omega_1 - \omega_2)t\right) \exp\left(\frac{i}{2}(\omega_1 + \omega_2)t\right).$$

The instantaneous frequency is  $\theta'(t) = (\omega_1 + \omega_2)/2$  and the amplitude is

$$a(t) = 2a \left| \cos\left(\frac{1}{2}(\omega_1 - \omega_2)t\right) \right|.$$

This result is not satisfying because it does not reveal that the signal includes two sinusoidal waves of the same amplitude; it measures an average frequency value. The next sections explain how to measure the instantaneous frequencies of several spectral components by separating them with a windowed Fourier transform or a wavelet transform. We first describe two important applications of instantaneous frequencies.

### **Frequency Modulation**

In signal communications, information can be transmitted through the amplitude  $a(t)$  (amplitude modulation) or the instantaneous frequency  $\theta'(t)$  (frequency modulation) [60]. Frequency modulation is more robust in the presence of additive Gaussian white noise. In addition, it better resists multipath interferences, which destroy the amplitude information. A frequency modulation sends a message  $m(t)$  through a signal

$$f(t) = a \cos \theta(t) \quad \text{with} \quad \theta'(t) = \omega_0 + k m(t).$$

The frequency bandwidth of  $f$  is proportional to  $k$ . This constant is adjusted depending on transmission noise and available bandwidth. At reception, the message  $m(t)$  is restored with a frequency demodulation that computes the instantaneous frequency  $\theta'(t)$  [120].

### **Additive Sound Models**

Musical sounds and voiced speech segments can be modeled with sums of sinusoidal *partials*:

$$f(t) = \sum_{k=1}^K f_k(t) = \sum_{k=1}^K a_k(t) \cos \theta_k(t), \quad (4.71)$$

where  $a_k$  and  $\theta_k'$  vary slowly [413, 414]. Such decompositions are useful for pattern recognition and for modifying sound properties [339]. Sections 4.4.2 and 4.4.3 explain how to compute  $a_k$  and the instantaneous frequency  $\theta_k'$  and reconstruct signals from this information.

Reducing or increasing the duration of a sound  $f$  by a factor  $\alpha$  in time is used for radio broadcasting to adjust recorded sequences to a precise time schedule. A scaling  $f(\alpha t)$  transforms each  $\theta_k(t)$  in  $\theta_k(\alpha t)$  and therefore  $\theta_k'(t)$  in  $\alpha \theta_k'(\alpha t)$ . For sound reduction, with  $\alpha > 1$  all frequencies are thus increased. To avoid modifying the values of  $\theta_k'$  and  $a_k$ , a new sound is synthesized:

$$f_\alpha(t) = \sum_{k=1}^K a_k(\alpha t) \cos\left(\frac{1}{\alpha} \theta_k(\alpha t)\right). \quad (4.72)$$

The partials of  $f_\alpha$  at  $t = t_0/\alpha$  and the partials of  $f$  at  $t = t_0$  have the same amplitudes and the same instantaneous frequencies, therefore the properties of these sounds are perceived as identical.

A frequency transposition with the same duration is calculated by dividing each phase by a constant  $\alpha$  in order to shift the sound harmonics:

$$f_\alpha(t) = \sum_{k=1}^K b_k(t) \cos(\theta_k(t)/\alpha). \quad (4.73)$$

The instantaneous frequency of each partial is now  $\theta_k'(t)/\alpha$ . To maintain the sound properties, the amplitudes  $b_k(t)$  must be adjusted so as not to modify the global frequency envelope  $F(t, \omega)$  of the harmonics:

$$a_k(t) = F(t, \theta_k'(t)) \quad \text{and} \quad b_k(t) = F(t, \theta_k'(t)/\alpha). \quad (4.74)$$

Many types of sounds—musical instruments or speech—are produced by an excitation that propagates across a wave guide. Locally,  $F(t, \omega)$  is the transfer function of the wave guide. In speech processing, it is called a *formant*. This transfer function is often approximated with an autoregressive filter of order  $M$ , in which case:

$$F(t, \omega) = \frac{C}{\sum_{m=0}^{M-1} c_m e^{-im\omega}}. \quad (4.75)$$

The parameters  $c_m$  are identified with (4.74) from the  $a_k$  and the  $b_k$  are then derived with (4.74) and (4.75).

#### 4.4.2 Windowed Fourier Ridges

The spectrogram  $P_S f(u, \xi) = |Sf(u, \xi)|^2$  measures the energy of  $f$  in a time-frequency neighborhood of  $(u, \xi)$ . The ridge algorithm computes the signal instantaneous frequencies and amplitudes from the local maxima of  $P_S f(u, \xi)$ . These local maxima define a geometric support in the time-frequency plane. Modifications of sound durations or frequency transpositions are computed with time or frequency dilations of the ridge support.

Time-frequency ridges were introduced by Delprat, Escudé, Guillemain, Kronland-Martinet, Tchamitchian, and Torrésani [66, 204] to analyze musical sounds and are used to represent time-varying frequency tones for a wide range of signals [66, 289].

The windowed Fourier transform is computed with a symmetric window  $g(t) = g(-t)$  that has a support equal to  $[-1/2, 1/2]$ . The Fourier transform  $\hat{g}$  is a real symmetric function. We suppose that  $|\hat{g}(\omega)| \leq \hat{g}(0)$  for all  $\omega \in \mathbb{R}$ , and that  $\hat{g}(0) = \int_{-1/2}^{1/2} g(t) dt$  is on the order of 1. Table 4.1 listed several examples of such windows. The window  $g$  is normalized so that  $\|g\| = 1$ . For a fixed scale  $s$ ,  $g_s(t) = s^{-1/2}g(t/s)$

has a support of size  $s$  and a unit norm. The corresponding windowed Fourier atoms are

$$g_{s,u,\xi}(t) = g_s(t-u) e^{i\xi t},$$

and the resulting windowed Fourier transform is

$$Sf(u, \xi) = \langle f, g_{s,u,\xi} \rangle = \int_{-\infty}^{+\infty} f(t) g_s(t-u) e^{-i\xi t} dt. \quad (4.76)$$

Theorem 4.6 relates  $Sf(u, \xi)$  to the instantaneous frequency of  $f$ .

**Theorem 4.6.** Let  $f(t) = a(t) \cos \theta(t)$ . If  $\xi \geq 0$ , then

$$\langle f, g_{s,u,\xi} \rangle = \frac{\sqrt{s}}{2} a(u) \exp(i[\theta(u) - \xi u]) \left( \hat{g}(s[\xi - \theta'(u)]) + \varepsilon(u, \xi) \right). \quad (4.77)$$

The corrective term satisfies

$$|\varepsilon(u, \xi)| \leq \varepsilon_{a,1} + \varepsilon_{a,2} + \varepsilon_{\theta,2} + \sup_{|\omega| \geq s\theta'(u)} |\hat{g}(\omega)| \quad (4.78)$$

with

$$\varepsilon_{a,1} \leq \frac{s |a'(u)|}{|a(u)|}, \quad \varepsilon_{a,2} \leq \sup_{|t-u| \leq s/2} \frac{s^2 |a''(t)|}{|a(u)|}, \quad (4.79)$$

and if  $s |a'(u)| |a(u)|^{-1} \leq 1$ , then

$$\varepsilon_{\theta,2} \leq \sup_{|t-u| \leq s/2} s^2 |\theta''(t)|. \quad (4.80)$$

If  $\xi = \theta'(u)$ , then

$$\varepsilon_{a,1} = \frac{s |a'(u)|}{|a(u)|} \left| \hat{g}'(2s\theta'(u)) \right|. \quad (4.81)$$

**Proof.** Observe that

$$\begin{aligned} \langle f, g_{s,u,\xi} \rangle &= \int_{-\infty}^{+\infty} a(t) \cos \theta(t) g_s(t-u) \exp(-i\xi t) dt \\ &= \frac{1}{2} \int_{-\infty}^{+\infty} a(t) (\exp[i\theta(t)] + \exp[-i\theta(t)]) g_s(t-u) \exp[-i\xi t] dt \\ &= I(\theta) + I(-\theta). \end{aligned}$$

We first concentrate on

$$\begin{aligned} I(\theta) &= \frac{1}{2} \int_{-\infty}^{+\infty} a(t) \exp[i\theta(t)] g_s(t-u) \exp(-i\xi t) dt \\ &= \frac{1}{2} \int_{-\infty}^{+\infty} a(t+u) e^{i\theta(t+u)} g_s(t) \exp[-i\xi(t+u)] dt. \end{aligned}$$

This integral is computed by using second-order Taylor expansions:

$$\begin{aligned} a(t+u) &= a(u) + t a'(u) + \frac{t^2}{2} \alpha(t) \text{ with } |\alpha(t)| \leq \sup_{h \in [u, t+u]} |a''(h)| \\ \theta(t+u) &= \theta(u) + t \theta'(u) + \frac{t^2}{2} \beta(t) \text{ with } |\beta(t)| \leq \sup_{h \in [u, t+u]} |\theta''(h)|. \end{aligned}$$

We get

$$\begin{aligned} & 2 \exp(-i(\theta(u) - \xi u)) I(\theta) \\ &= \int_{-\infty}^{+\infty} a(u) g_s(t) \exp(-it(\xi - \theta'(u))) \exp\left(i \frac{t^2}{2} \beta(t)\right) dt \\ &+ \int_{-\infty}^{+\infty} a'(u) t g_s(t) \exp(-it(\xi - \theta'(u))) \exp\left(i \frac{t^2}{2} \beta(t)\right) dt \\ &+ \frac{1}{2} \int_{-\infty}^{+\infty} \alpha(t) t^2 g_s(t) \exp(-i(t\xi + \theta(u) - \theta(t+u))) dt. \end{aligned}$$

A first-order Taylor expansion of  $\exp(ix)$  gives

$$\exp\left(i \frac{t^2}{2} \beta(t)\right) = 1 + \frac{t^2}{2} \beta(t) \gamma(t), \text{ with } |\gamma(t)| \leq 1. \quad (4.82)$$

Since

$$\int_{-\infty}^{+\infty} g_s(t) \exp[-it(\xi - \theta'(u))] dt = \sqrt{s} \hat{g}(s[\xi - \theta'(u)]),$$

inserting (4.82) in the expression of  $I(\theta)$  yields

$$\begin{aligned} & \left| I(\theta) - \frac{\sqrt{s}}{2} a(u) \exp[i(\theta(u) - \xi u)] \hat{g}(\xi - \theta'(u)) \right| \\ & \leq \frac{\sqrt{s} |a(u)|}{4} (\varepsilon_{a,1}^+ + \varepsilon_{a,2} + \varepsilon_{\theta,2}), \end{aligned} \quad (4.83)$$

with

$$\varepsilon_{a,1}^+ = \frac{2|a'(u)|}{|a(u)|} \left| \int_{-\infty}^{+\infty} t \frac{1}{\sqrt{s}} g_s(t) \exp[-it(\xi - \theta'(u))] dt \right|, \quad (4.84)$$

$$\varepsilon_{a,2} = \int_{-\infty}^{+\infty} t^2 |\alpha(t)| \frac{1}{\sqrt{s}} |g_s(t)| dt, \quad (4.85)$$

$$\begin{aligned} \varepsilon_{\theta,2} &= \int_{-\infty}^{+\infty} t^2 |\beta(t)| \frac{1}{\sqrt{s}} |g_s(t)| dt \\ &+ \frac{|a'(u)|}{|a(u)|} \int_{-\infty}^{+\infty} |t^3| |\beta(t)| \frac{1}{\sqrt{s}} |g_s(t)| dt. \end{aligned} \quad (4.86)$$

Applying (4.83) to  $I(-\theta)$  gives

$$|I(-\theta)| \leq \frac{\sqrt{s} |a(u)|}{2} |\hat{g}(\xi + \theta'(u))| + \frac{\sqrt{s} |a(u)|}{4} (\varepsilon_{a,1}^- + \varepsilon_{a,2} + \varepsilon_{\theta,2}),$$

with

$$\varepsilon_{a,1}^- = \frac{2|a'(u)|}{|a(u)|} \left| \int_{-\infty}^{+\infty} t \frac{1}{\sqrt{s}} g_s(t) \exp[-it(\xi + \theta'(u))] dt \right|. \quad (4.87)$$

Since  $\xi \geq 0$  and  $\theta'(u) \geq 0$ , we derive that

$$|\hat{g}(s[\xi + \theta'(u)])| \leq \sup_{|\omega| \geq s\theta'(u)} |\hat{g}(\omega)|;$$

therefore

$$I(\theta) + I(-\theta) = \frac{\sqrt{s}}{2} a(u) \exp[i(\theta(u) - \xi u)] \left( \hat{g}(s[\xi - \theta'(u)]) + \varepsilon(u, \xi) \right),$$

with

$$\varepsilon(u, \xi) = \frac{\varepsilon_{a,1}^+ + \varepsilon_{a,1}^-}{2} + \varepsilon_{a,2} + \varepsilon_{\theta,2} + \sup_{|\omega| \geq s|\theta'(u)|} |\hat{g}(\omega)|.$$

Let us now verify the upper bound (4.79) for  $\varepsilon_{a,1} = (\varepsilon_{a,1}^+ + \varepsilon_{a,1}^-)/2$ . Since  $g_s(t) = s^{-1/2}g(t/s)$ , a simple calculation shows that for  $n \geq 0$

$$\int_{-\infty}^{+\infty} |t|^n \frac{1}{\sqrt{s}} |g_s(t)| dt = s^n \int_{-1/2}^{1/2} |t|^n |g(t)| dt \leq \frac{s^n}{2^n} \|g\|^2 = \frac{s^n}{2^n}. \quad (4.88)$$

Inserting this for  $n = 1$  in (4.84) and (4.87) gives

$$\varepsilon_{a,1} = \frac{\varepsilon_{a,1}^+ + \varepsilon_{a,1}^-}{2} \leq \frac{s|a'(u)|}{|a(u)|}.$$

The upper bounds (4.79) and (4.80) of the second-order terms  $\varepsilon_{a,2}$  and  $\varepsilon_{\theta,2}$  are obtained by observing that the remainder  $\alpha(t)$  and  $\beta(t)$  of the Taylor expansion of  $a(t+u)$  and  $\theta(t+u)$  satisfy

$$\sup_{|t| \leq s/2} |\alpha(t)| \leq \sup_{|t-u| \leq s/2} |a''(t)|, \quad \sup_{|t| \leq s/2} |\beta(t)| \leq \sup_{|t-u| \leq s/2} |\theta''(t)|. \quad (4.89)$$

Inserting this in (4.85) yields

$$\varepsilon_{a,2} \leq \sup_{|t-u| \leq s/2} \frac{s^2 |a''(t)|}{|a(u)|}.$$

When  $s|a'(u)||a(u)|^{-1} \leq 1$ , replacing  $|\beta(t)|$  by its upper bound in (4.86) gives

$$\varepsilon_{\theta,2} \leq \frac{1}{2} \left( 1 + \frac{s|a'(u)|}{|a(u)|} \right) \sup_{|t-u| \leq s/2} s^2 |\theta''(t)| \leq \sup_{|t-u| \leq s/2} s^2 |\theta''(t)|.$$

Let us finally compute  $\varepsilon_a$  when  $\xi = \theta'(u)$ . Since  $g(t) = g(-t)$ , we derive from (4.84) that

$$\varepsilon_{a,1}^+ = \frac{2|a'(u)|}{|a(u)|} \left| \int_{-\infty}^{+\infty} t \frac{1}{\sqrt{s}} g_s(t) dt \right| = 0.$$

We also derive from (2.22) that the Fourier transform of  $t \frac{1}{\sqrt{s}} g_s(t)$  is  $i s \hat{g}'(s\omega)$ , so (4.87) gives

$$\varepsilon_a = \frac{1}{2} \varepsilon_{a,1}^- = \frac{s|a'(u)|}{|a(u)|} |\hat{g}'(2s\theta'(u))|. \quad \blacksquare$$

Delprat et al. [204] give a different proof of a similar result when  $g(t)$  is a Gaussian, using a stationary phase approximation. If we can neglect the corrective term  $\varepsilon(u, \xi)$ , we will see that (4.77) enables us to measure  $a(u)$  and  $\theta'(u)$  from  $Sf(u, \xi)$ . This implies that the decomposition  $f(t) = a(t) \cos \theta(t)$  is uniquely defined. By reviewing the proof of Theorem 4.6, one can verify that  $a$  and  $\theta'$  are the analytic amplitude and instantaneous frequencies of  $f$ .

The expressions (4.79, 4.80) show that the three corrective terms  $\varepsilon_{a,1}$ ,  $\varepsilon_{a,2}$ , and  $\varepsilon_{\theta,2}$  are small if  $a(t)$  and  $\theta'(t)$  have small relative variations over the support of window  $g_s$ . Let  $\Delta\omega$  be the bandwidth of  $\hat{g}$  defined by

$$|\hat{g}(\omega)| \ll 1 \quad \text{for } |\omega| \geq \Delta\omega. \quad (4.90)$$

The term

$$\sup_{|\omega| \geq s|\theta'(u)|} |\hat{g}(\omega)| \text{ of } \varepsilon(u, \xi)$$

is negligible if

$$\theta'(u) \geq \frac{\Delta\omega}{s}.$$

### Ridge Points

Let us suppose that  $a(t)$  and  $\theta'(t)$  have small variations over intervals of size  $s$  and that  $\theta'(t) \geq \Delta\omega/s$  so that the corrective term  $\varepsilon(u, \xi)$  in (4.77) can be neglected. Since  $|\hat{g}(\omega)|$  is maximum at  $\omega = 0$ , (4.77) shows that for each  $u$  the spectrogram  $|Sf(u, \xi)|^2 = |\langle f, g_{s,u,\xi} \rangle|^2$  is maximum at  $\xi(u) = \theta'(u)$ . The corresponding time-frequency points  $(u, \xi(u))$  are called *ridges*. At ridge points, (4.77) becomes

$$Sf(u, \xi) = \frac{\sqrt{s}}{2} a(u) \exp(i[\theta(u) - \xi u]) (\hat{g}(0) + \varepsilon(u, \xi)). \quad (4.91)$$

Theorem 4.6 proves that the  $\varepsilon(u, \xi)$  is smaller at a ridge point because the first-order term  $\varepsilon_{a,1}$  becomes negligible in (4.81). This is shown by verifying that  $|\hat{g}'(2s\theta'(u))|$  is negligible when  $s\theta'(u) \geq \Delta\omega$ . At ridge points, the second-order terms  $\varepsilon_{a,2}$  and  $\varepsilon_{\theta,2}$  are predominant in  $\varepsilon(u, \xi)$ .

The ridge frequency gives the instantaneous frequency  $\xi(u) = \theta'(u)$  and the amplitude is calculated by

$$a(u) = \frac{2 |Sf(u, \xi(u))|}{\sqrt{s} |\hat{g}(0)|}. \quad (4.92)$$

Let  $\Theta_S(u, \xi)$  be the complex phase of  $Sf(u, \xi)$ . If we neglect the corrective term, then (4.91) proves that ridges are also stationary phase points:

$$\frac{\partial \Theta_S(u, \xi)}{\partial u} = \theta'(u) - \xi = 0.$$

Testing the stationarity of the phase locates the ridges more precisely.

### Multiple Frequencies

When the signal contains several spectral lines having frequencies sufficiently apart, the windowed Fourier transform separates each of these components and the ridges detect the evolution in time of each spectral component. Let us consider

$$f(t) = a_1(t) \cos \theta_1(t) + a_2(t) \cos \theta_2(t),$$

where  $a_k(t)$  and  $\theta'_k(t)$  have small variations over intervals of size  $s$  and  $s\theta'_k(t) \geq \Delta\omega$ . Since the windowed Fourier transform is linear, we apply (4.77) to each spectral component and neglect the corrective terms:

$$\begin{aligned} Sf(u, \xi) &= \frac{\sqrt{s}}{2} a_1(u) \hat{g}(s[\xi - \theta'_1(u)]) \exp(i[\theta_1(u) - \xi u]) \\ &\quad + \frac{\sqrt{s}}{2} a_2(u) \hat{g}(s[\xi - \theta'_2(u)]) \exp(i[\theta_2(u) - \xi u]). \end{aligned} \quad (4.93)$$

The two spectral components are discriminated if for all  $u$

$$\hat{g}(s|\theta'_1(u) - \theta'_2(u)|) \ll 1, \quad (4.94)$$

which means that the frequency difference is larger than the bandwidth of  $\hat{g}(s\omega)$ :

$$|\theta'_1(u) - \theta'_2(u)| \geq \frac{\Delta\omega}{s}. \quad (4.95)$$

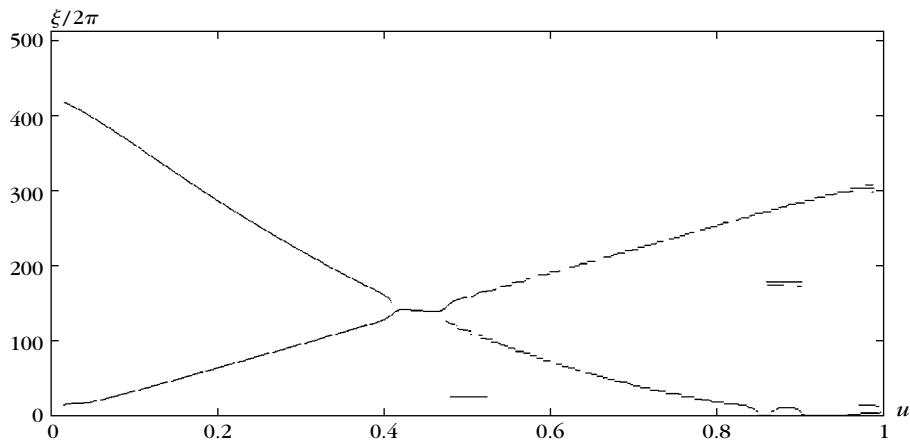
In this case, when  $\xi = \theta'_1(u)$ , the second term of (4.93) can be neglected and the first term generates a ridge point from which we may recover  $\theta'_1(u)$  and  $a_1(u)$  by using (4.92). Similarly, if  $\xi = \theta'_2(u)$ , the first term can be neglected and we have a second ridge point that characterizes  $\theta'_2(u)$  and  $a_2(u)$ . The ridge points are distributed along two time-frequency lines,  $\xi(u) = \theta'_1(u)$  and  $\xi(u) = \theta'_2(u)$ . This result is valid for any number of time-varying spectral components, as long as the distance between any two instantaneous frequencies satisfies (4.95). If two spectral lines are too close, they interfere, thus destroying the ridge pattern.

### Time-Frequency Ridge Support

The number of instantaneous frequencies is typically unknown. The ridge support  $\Lambda$  therefore is defined as the set of all  $(u, \xi)$ —the local maxima of  $|Sf(u, \xi)|^2$  for  $u$  fixed and  $\xi$  varying and points of stationary phase  $\partial\Theta_S(u, \xi)/\partial u \approx 0$ . This support is often reduced by removing small-ridge amplitudes  $|Sf(u, \xi)|$  that are mostly dominated by the noise, or because smaller ridges may be “shadows” of other instantaneous frequencies created by the side lobes of  $\hat{g}(\omega)$ .

Let  $\{g_{s,u,\xi}\}_{(u,\xi)\in\Lambda}$  be the set of ridge atoms. For discrete signals, there is a finite number of ridge points, that define a frame of the space  $\mathbf{V}_\Lambda$  they generate. A ridge signal approximation is computed as an orthogonal projection of  $f$  on  $\mathbf{V}_\Lambda$ . Section 5.1.3 shows that it is obtained with the dual frame  $\{\tilde{g}_{\Lambda,u,\xi}\}_{(u,\xi)\in\Lambda}$  of  $\{g_{s,u,\xi}\}_{(u,\xi)\in\Lambda}$  in  $\mathbf{V}_\Lambda$ :

$$f_\Lambda = \sum_{(u,\xi)\in\Lambda} Sf(u, \xi) \tilde{g}_{\Lambda,u,\xi}. \quad (4.96)$$

**FIGURE 4.12**

Support of larger-amplitude ridges calculated from the spectrogram in Figure 4.3. These ridges give the instantaneous frequencies of the linear and quadratic chirps and of low- and high-frequency transients at  $t = 0.5$  and  $t = 0.87$ .

The dual-synthesis algorithm of Section 5.1.3 computes this orthogonal projection by inverting the symmetric operator

$$Lh = \sum_{(u,\xi) \in \Lambda} \langle h, g_{s,u,\xi} \rangle g_{s,u,\xi}. \quad (4.97)$$

The inversion requires iteration of this operator many times. If there are only a few ridge points, then (4.97) is efficiently computed by evaluating the inner product and the sum for just  $(u, \xi) \in \Lambda$ . If there are many ridge points, it is more efficient to compute the full windowed Fourier transform  $Sh(u, \xi) = \langle h, g_{s,u,\xi} \rangle$  with the FFT algorithm (described in Section 4.2.3); set all coefficients to zero for  $(u, \xi) \notin \Lambda$  and apply the fast inverse windowed Fourier transform over all coefficients. The normalization factor  $N^{-1}$  in (4.28) must be removed (set to 1) to implement (4.97).

Figure 4.12 displays the ridge support computed from the modulus and phase of the windowed Fourier transform shown in Figure 4.3. For  $t \in [0.4, 0.5]$ , the instantaneous frequencies of the linear chirp and the quadratic chirps are close, the frequency resolution of the window is not sufficient to discriminate them. As a result, the ridges detect a single average instantaneous frequency.

### **Time-Scaling and Frequency Transpositions**

A reduction of sound duration by a factor  $\alpha$  is implemented, according to the deformation model (4.72), by dilating the ridge support  $\Lambda$  in time:

$$\Lambda_\alpha = \{(u, \xi) : (\alpha u, \xi) \in \Lambda\}. \quad (4.98)$$

The windowed Fourier coefficients  $c(u, \xi)$  in  $\Lambda_\alpha$  are derived from the modulus and phase of ridge coefficients

$$\forall (v, \xi) \in \Lambda_\alpha, \quad c(v, \xi) = |Sf(\alpha v, \xi)| e^{i\Theta_S(\alpha v, \xi)/\alpha}. \quad (4.99)$$

The scaled signal is reconstructed from these coefficients, with the dual-synthesis algorithm of Section 5.1.3, as in (4.96):

$$f_\alpha = \sum_{(v, \xi) \in \Lambda_\alpha} c(v, \xi) \tilde{g}_{\Lambda_\alpha, v, \xi}.$$

Similarly, a sound transposition is implemented according to the transposition model (4.73) by dilating the ridge support  $\Lambda$  in frequency:

$$\Lambda_\alpha = \{(u, \xi) : (u, \alpha \xi) \in \Lambda\}. \quad (4.100)$$

The transposed coefficient amplitudes  $|c(u, \xi)|$  in  $\Lambda_\alpha$  are calculated with (4.74). At any fixed time  $u_0$ , the ridge amplitudes at all frequencies  $\{a(u_0, \xi) = |Sf(u_0, \xi)|\}_{(u_0, \xi) \in \Lambda}$  are mapped to transposed amplitudes  $\{b(u_0, \eta)\}_{(u_0, \eta) \in \Lambda_\alpha}$  at frequencies  $\eta = \xi/\alpha$  by computing a frequency envelope. The resulting ridge coefficients are

$$\forall (u, \eta) \in \Lambda_\alpha, \quad c(u, \eta) = b(u, \eta) e^{i\Theta_S(u, \alpha \eta)/\alpha}. \quad (4.101)$$

The transposed signal is reconstructed with the dual-synthesis algorithm of Section 5.1.3:

$$f_\alpha = \sum_{(u, \eta) \in \Lambda_\alpha} c(u, \eta) \tilde{g}_{\Lambda_\alpha, u, \eta}.$$

### **Choice of Window**

The measurement of instantaneous frequencies at ridge points is valid only if the size  $s$  of the window  $g_s$  is sufficiently small so that the second-order terms  $\varepsilon_{\alpha,2}$  and  $\varepsilon_{\theta,2}$  in (4.79) and (4.80) are small:

$$\sup_{|t-u| \leq s/2} \frac{s^2 |a_k''(t)|}{|a_k(u)|} \ll 1 \text{ and } \sup_{|t-u| \leq s/2} s^2 |\theta_k''(t)| \ll 1. \quad (4.102)$$

On the other hand, the frequency bandwidth  $\Delta\omega/s$  must also be sufficiently small to discriminate consecutive spectral components in (4.95). The window scale  $s$  therefore must be adjusted as a trade-off between both constraints.

Table 4.1 listed the spectral parameters of several windows of compact support. For instantaneous frequency detection, it is particularly important to ensure that  $\hat{g}$  has negligible side lobes at  $\pm \omega_0$ , as illustrated by Figure 4.4. The reader can verify with (4.77) that these side lobes “react” to an instantaneous frequency  $\theta'(u)$  by creating shadow maxima of  $|Sf(u, \xi)|^2$  at frequencies  $\xi = \theta'(u) \pm \omega_0$ . The ratio of the amplitude of these shadow maxima to the amplitude of the main local maxima at  $\xi = \theta'(u)$  is  $|\hat{g}(\omega_0)|^2 / |\hat{g}(0)|^{-2}$ . They can be removed by thresholding or by testing the stationarity of the phase.

---

**EXAMPLE 4.13**

The sum of two parallel linear chirps

$$f(t) = a_1 \cos(bt^2 + ct) + a_2 \cos(bt^2) \quad (4.103)$$

has two instantaneous frequencies  $\theta_1'(t) = 2bt + c$  and  $\theta_2'(t) = 2bt$ . Figure 4.13 gives a numerical example. The window  $g_s$  has enough frequency resolution to discriminate both chirps if

$$|\theta_1'(t) - \theta_2'(t)| = |c| \geq \frac{\Delta\omega}{s}. \quad (4.104)$$

Its time support is small enough if

$$s^2 |\theta_1''(u)| = s^2 |\theta_2''(u)| = 2b s^2 \ll 1. \quad (4.105)$$

Conditions (4.104) and (4.105) prove that there exists an appropriate window  $g$ , if and only if,

$$\frac{c}{\sqrt{b}} \gg \Delta\omega. \quad (4.106)$$

Since  $g$  is a smooth window with a support  $[-1/2, 1/2]$ , its frequency bandwidth  $\Delta\omega$  is on the order of 1. The linear chirps in Figure 4.13 satisfy (4.106). Ridges are computed with the truncated Gaussian window of Table 4.1, with  $s = 0.5$ .

---



---

**EXAMPLE 4.14**

The hyperbolic chirp

$$f(t) = \cos\left(\frac{\alpha}{\beta - t}\right)$$

for  $0 \leq t < \beta$  has an instantaneous frequency

$$\theta'(t) = \frac{\alpha}{(\beta - t)^2},$$

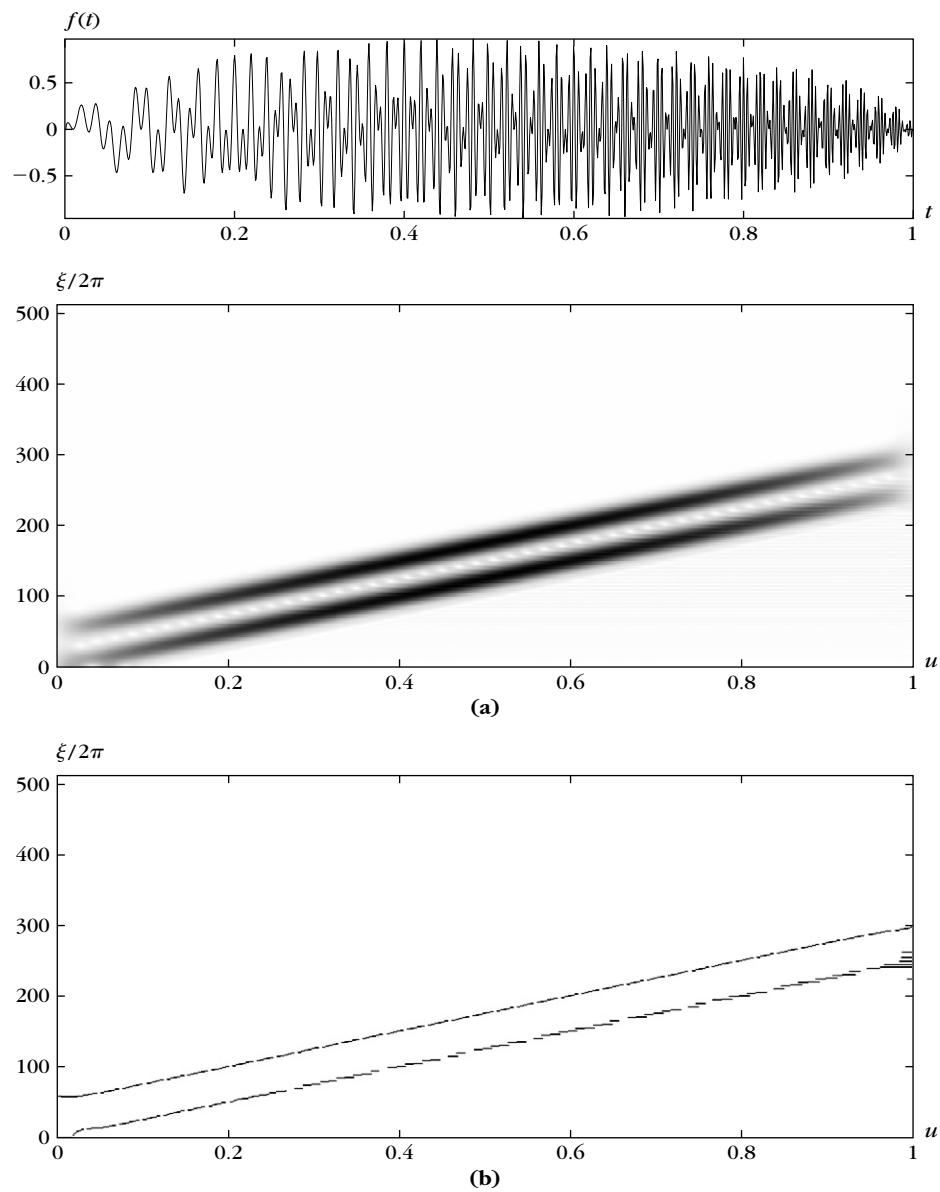
which varies quickly when  $t$  is close to  $\beta$ . The instantaneous frequency of hyperbolic chirps goes from 0 to  $+\infty$  in a finite time interval. This is particularly useful for radars. Such chirps are also emitted by the cruise sonars of bats [204].

The instantaneous frequency of hyperbolic chirps cannot be estimated with a windowed Fourier transform because for any fixed window size instantaneous frequency varies too quickly at high frequencies. When  $u$  is close enough to  $\beta$ , then (4.102) is not satisfied because

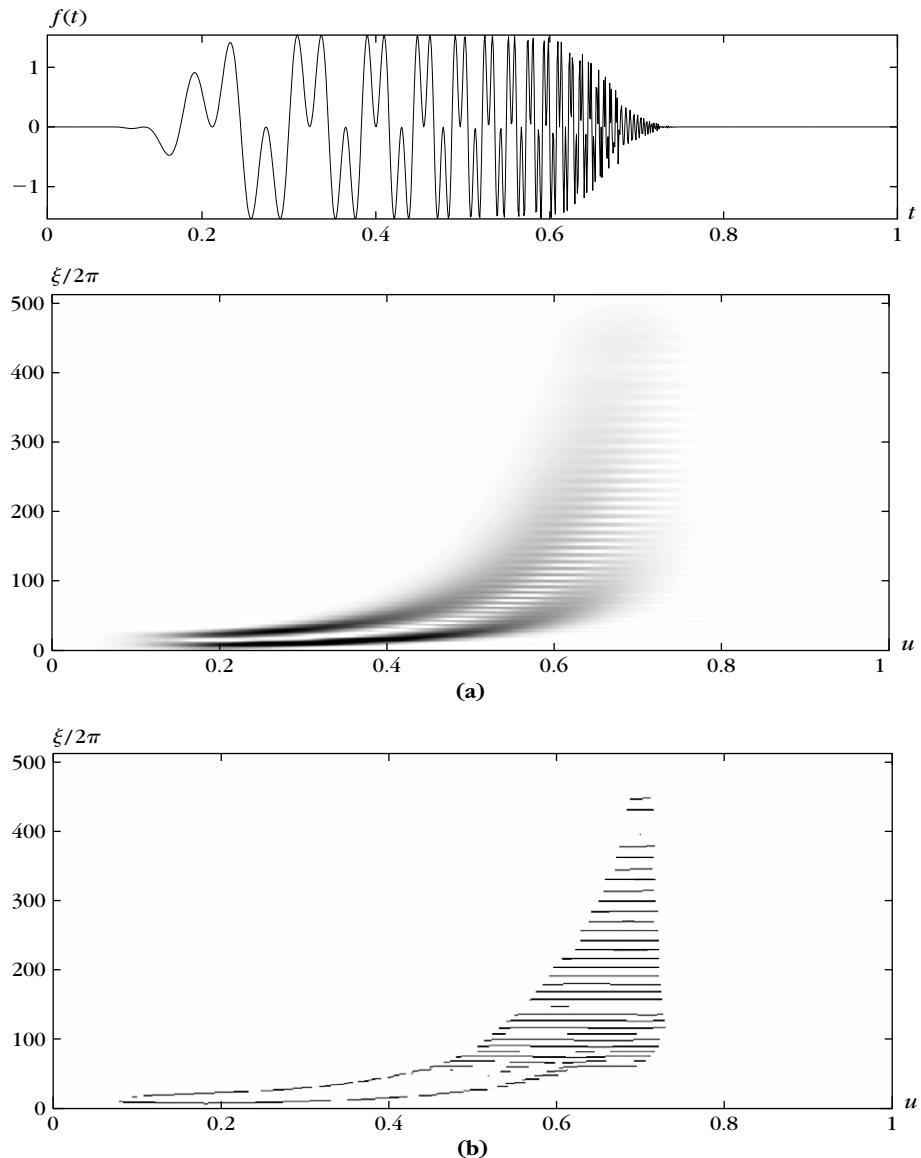
$$s^2 |\theta''(u)| = \frac{s^2 \alpha}{(\beta - u)^3} > 1.$$

Figure 4.14 shows a signal that is a sum of two hyperbolic chirps:

$$f(t) = a_1 \cos\left(\frac{\alpha_1}{\beta_1 - t}\right) + a_2 \cos\left(\frac{\alpha_2}{\beta_2 - t}\right), \quad (4.107)$$

**FIGURE 4.13**

Sum of two parallel linear chirps: **(a)** Spectrogram  $P_S f(u, \xi) = |Sf(u, \xi)|^2$ . **(b)** Ridge support calculated from the spectrogram.

**FIGURE 4.14**

Sum of two hyperbolic chirps: **(a)** Spectrogram  $P_S f(u, \xi)$ . **(b)** Ridge support calculated from the spectrogram.

with  $\beta_1 = 0.68$  and  $\beta_2 = 0.72$ . At the beginning of the signal, the two chirps have close instantaneous frequencies that are discriminated by the windowed Fourier ridge computed with a large window. When getting close to  $\beta_1$  and  $\beta_2$ , the instantaneous frequency varies too quickly relative to window size. The resulting ridges cannot follow the instantaneous frequencies.

### 4.4.3 Wavelet Ridges

Windowed Fourier atoms have a fixed scale and thus cannot follow the instantaneous frequency of rapidly varying events such as hyperbolic chirps. In contrast, an analytic wavelet transform modifies the scale of its time-frequency atoms. The ridge algorithm of Delprat et al. [204] is extended to analytic wavelet transforms to accurately measure frequency tones that are rapidly changing at high frequencies.

An approximately analytic wavelet is constructed in (4.60) by multiplying a window  $g$  with a sinusoidal wave:

$$\psi(t) = g(t) \exp(i\eta t).$$

As in the previous section,  $g$  is a symmetric window, with a support equal to  $[-1/2, 1/2]$ , and a unit norm  $\|g\| = 1$ . Let  $\Delta\omega$  be the bandwidth of  $\hat{g}$  defined in (4.90). If  $\eta > \Delta\omega$ , then

$$\forall \omega < 0, \quad \hat{\psi}(\omega) = \hat{g}(\omega - \eta) \ll 1.$$

The wavelet  $\psi$  is not strictly analytic because its Fourier transform is not exactly equal to zero at negative frequencies.

Dilated and translated wavelets can be rewritten as

$$\psi_{u,s}(t) = \frac{1}{\sqrt{s}} \psi\left(\frac{t-u}{s}\right) = g_{s,u,\xi}(t) \exp(-i\xi u),$$

with  $\xi = \eta/s$  and

$$g_{s,u,\xi}(t) = \sqrt{s} g\left(\frac{t-u}{s}\right) \exp(i\xi t).$$

The resulting wavelet transform uses time-frequency atoms similar to those of a windowed Fourier transform (4.76); however, in this case scale  $s$  varies over  $\mathbb{R}^+$  while  $\xi = \eta/s$ :

$$Wf(u, s) = \langle f, \psi_{u,s} \rangle = \langle f, g_{s,u,\xi} \rangle \exp(i\xi u).$$

Theorem 4.6 computes  $\langle f, g_{s,u,\xi} \rangle$  when  $f(t) = a(t) \cos \theta(t)$ , which gives

$$Wf(u, s) = \frac{\sqrt{s}}{2} a(u) \exp[t\theta(u)] \left( \hat{g}(s[\xi - \theta'(u)]) + \varepsilon(u, \xi) \right). \quad (4.108)$$

The corrective term  $\varepsilon(u, \xi)$  is negligible if  $a(t)$  and  $\theta'(t)$  have small variations over the support of  $\psi_{u,s}$  and if  $\theta'(u) \geq \Delta\omega/s$ .

### Ridge Detection

Instantaneous frequency is measured from ridges defined over the wavelet transform. The normalized scalogram defined by

$$\frac{\xi}{\eta} P_W f(u, \xi) = \frac{|Wf(u, s)|^2}{s} \quad \text{for } \xi = \eta/s$$

is calculated with (4.108):

$$\frac{\xi}{\eta} P_W f(u, \xi) = \frac{1}{4} a^2(u) \left| \hat{g} \left( \eta \left[ 1 - \frac{\theta'(u)}{\xi} \right] \right) + \varepsilon(u, \xi) \right|^2.$$

Since  $|\hat{g}(\omega)|$  is maximum at  $\omega = 0$ , if we neglect  $\varepsilon(u, \xi)$ , this expression shows that the scalogram is maximum at

$$\frac{\eta}{s(u)} = \xi(u) = \theta'(u). \quad (4.109)$$

The corresponding points  $(u, \xi(u))$  are called *wavelet ridges*. The analytic amplitude is given by

$$a(u) = \frac{2 \sqrt{\eta^{-1} \xi P_W f(u, \xi)}}{|\hat{g}(0)|}. \quad (4.110)$$

The complex phase of  $Wf(u, s)$  in (4.108) is  $\Theta_W(u, \xi) = \theta(u)$ ; at ridge points,

$$\frac{\partial \Theta_W(u, \xi)}{\partial u} = \theta'(u) = \xi. \quad (4.111)$$

When  $\xi = \theta'(u)$ , the first-order term  $\varepsilon_{a,1}$  calculated in (4.81) becomes negligible. The corrective term is then dominated by  $\varepsilon_{a,2}$  and  $\varepsilon_{\theta,2}$ . To simplify the expression, we approximate the supremum of  $a''$  and  $\theta''$  in the neighborhood of  $u$  by their value at  $u$ . Since  $s = \eta/\xi = \eta/\theta'(u)$ , (4.79) and (4.80) imply that these second-order terms become negligible if

$$\frac{\eta^2}{|\theta'(u)|^2} \frac{|a''(u)|}{|a(u)|} \ll 1 \quad \text{and} \quad \eta^2 \frac{|\theta''(u)|}{|\theta'(u)|^2} \ll 1. \quad (4.112)$$

The presence of  $\theta'$  in the denominator proves that  $a'$  and  $\theta'$  must have slow variations if  $\theta'$  is small but may vary much more quickly for large instantaneous frequencies.

### Multispectral Estimation

Suppose that  $f$  is a sum of two spectral components:

$$f(t) = a_1(t) \cos \theta_1(t) + a_2(t) \cos \theta_2(t).$$

As in (4.94), we verify that the second instantaneous frequency  $\theta'_2$  does not interfere with the ridge of  $\theta'_1$  if the dilated window has a sufficient spectral resolution at ridge scale  $s = \eta/\xi = \eta/\theta'_1(u)$ :

$$\hat{g}(s|\theta'_1(u) - \theta'_2(u)|) \ll 1. \quad (4.113)$$

Since the bandwidth of  $\hat{g}(\omega)$  is  $\Delta\omega$ , this means that

$$\frac{|\theta_1'(u) - \theta_2'(u)|}{\theta_1'(u)} \geq \frac{\Delta\omega}{\eta}. \quad (4.114)$$

Similarly, the first spectral component does not interfere with the second ridge located at  $s = \eta/\xi = \eta/\theta_2'(u)$  if

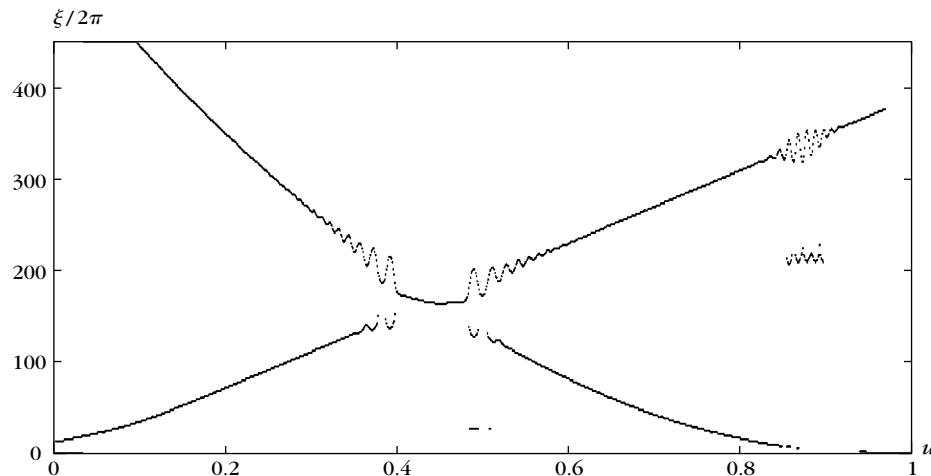
$$\frac{|\theta_1'(u) - \theta_2'(u)|}{\theta_2'(u)} \geq \frac{\Delta\omega}{\eta}. \quad (4.115)$$

To separate spectral lines that have close instantaneous frequencies, these conditions prove that the wavelet must have a small octave bandwidth  $\Delta\omega/\eta$ . The bandwidth  $\Delta\omega$  is a fixed constant on the order of 1. The frequency  $\eta$  is a free parameter that is chosen as a trade-off between the time-resolution condition (4.112) and the frequency bandwidth conditions (4.114) and (4.115).

Figure 4.15 displays the ridges computed from the normalized scalogram and the wavelet phase shown in Figure 4.11. The ridges of the high-frequency transient located at  $t = 0.87$  have oscillations because of interference with the linear chirp above. The frequency-separation condition (4.114) is not satisfied. This is also the case in the time interval  $[0.35, 0.55]$ , where the instantaneous frequencies of the linear and quadratic chirps are too close.

### Ridge Support and Processing

The wavelet ridge support  $\Lambda$  of  $f$  is the set of all ridge points  $(u, s)$  in the time-scale plane or  $(u, \xi = \eta/s)$  in the time-frequency plane, corresponding to local maxima of



**FIGURE 4.15**

Ridge support calculated from the scalogram shown in Figure 4.11; compare with the windowed Fourier ridges in Figure 4.12.

$|Wf(u, s)|/s$  for a fixed  $u$  and  $s$  varying, where the complex phase  $\Theta_W(u, s)$  nearly satisfies (4.111).

As in the windowed Fourier case, an orthogonal projection is computed over the space  $\mathbf{V}_\Lambda$  generated by the ridge wavelets  $\{\psi_{u,s}\}_{(u,s)\in\Lambda}$  by using the dual wavelet frame  $\{\tilde{\psi}_{\Lambda,u,s}\}_{(u,s)\in\Lambda}$ :

$$f_\Lambda = \sum_{(u,s)\in\Lambda} Wf(u, s) \tilde{\psi}_{\Lambda,u,s}. \quad (4.116)$$

It is implemented with the dual-synthesis algorithm of Section 5.1.3 by inverting the symmetric operator

$$Lh = \sum_{(u,s)\in\Lambda} \langle h, \psi_{u,s} \rangle \psi_{u,s}, \quad (4.117)$$

which is performed by computing this operator many times. When there are many ridge points, instead of computing this sum only for  $(u, s) \in \Lambda$ , it may require less operations to compute  $wf(u, s)$  with the fast wavelet transform algorithm of Section 4.3.3. All coefficients  $(u, s) \notin \Lambda$  are set to zero, and the fast inverse wavelet transform algorithm is applied. The inverse wavelet transform formula (4.66) must be modified by removing the renormalization factor  $a^{-j}$  and  $a^{-J}$  in the sum (set them to 1) to implement the operator (4.117).

The same as in the windowed Fourier case, modifications of sound durations or frequency transpositions are computed by modifying ridge support. A reduction of sound duration by a factor  $\alpha$  transforms ridge support  $\Lambda$  into:

$$\Lambda_\alpha = \{(u, s) : (\alpha u, s) \in \Lambda\}. \quad (4.118)$$

A sound transposition is implemented by modifying the scales of the time-scale ridge support  $\Lambda$ , which defines:

$$\Lambda_\alpha = \{(u, s) : (u, s/\alpha) \in \Lambda\}. \quad (4.119)$$

The wavelet coefficients over these supports are derived from the deformation model (4.72) or (4.74), similar to (4.99) and (4.101) for the windowed Fourier transform. Processed signals are recovered from the modified wavelet coefficients and modified supports with the dual-synthesis algorithm of Section 5.1.3.

### EXAMPLE 4.15

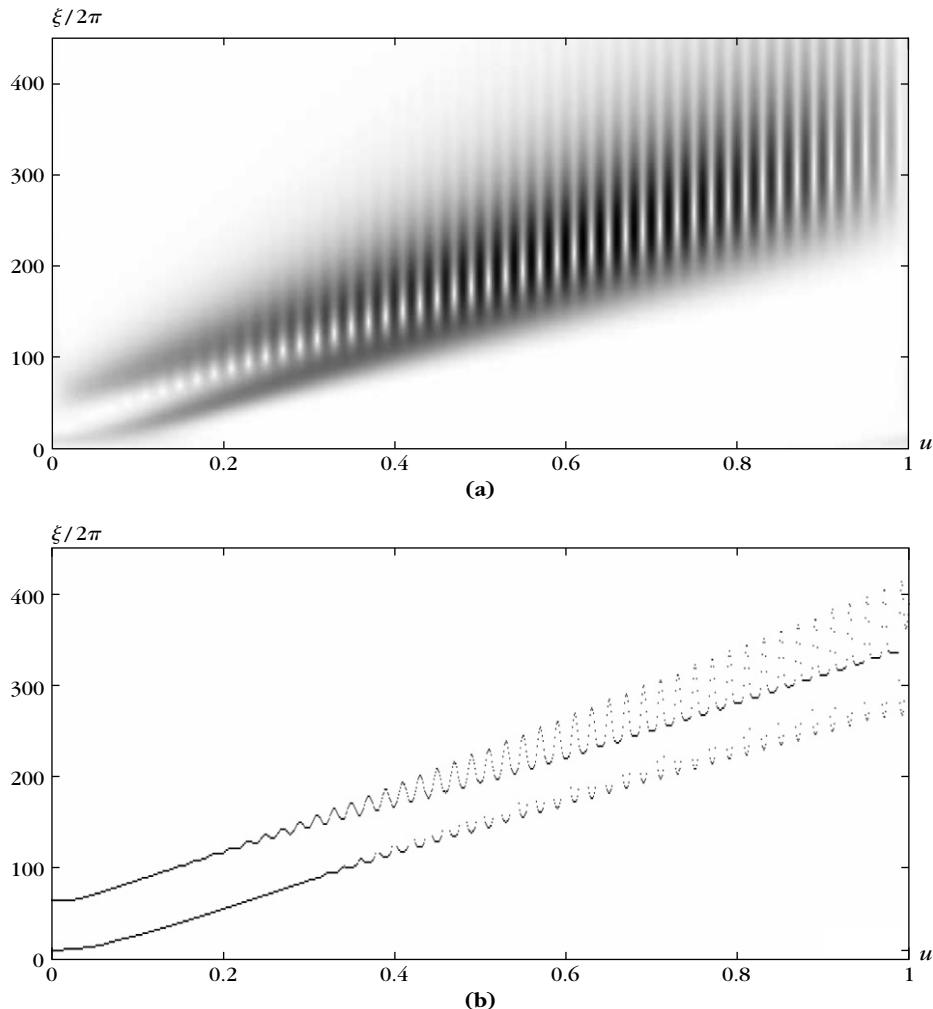
The instantaneous frequencies of two linear chirps

$$f(t) = a_1 \cos(bt^2 + ct) + a_2 \cos(bt^2)$$

are not precisely measured by wavelet ridges. Indeed,

$$\frac{|\theta'_2(u) - \theta'_1(u)|}{\theta'_1(u)} = \frac{c}{bt}$$

converges to zero when  $t$  increases. When it is smaller than  $\Delta\omega/\eta$ , the two chirps interact and create interference patterns like those shown in Figure 4.16. The ridges follow these interferences and do not properly estimate the two instantaneous frequencies, as opposed to the windowed Fourier ridges shown in Figure 4.13.



**FIGURE 4.16**

- (a) Normalized scalogram  $\eta^{-1}\xi P_W f(u, \xi)$  of two parallel linear chirps shown in Figure 4.13.
- (b) Wavelet ridges.

---

**EXAMPLE 4.16**

The instantaneous frequency of a hyperbolic chirp

$$f(t) = \cos\left(\frac{\alpha}{\beta-t}\right)$$

is  $\theta'(t) = \alpha(1-t)^{-2}$ . Wavelet ridges can measure this instantaneous frequency if the time-resolution condition (4.112) is satisfied:

$$\eta^2 \ll \frac{\theta'(t)^2}{|\theta''(t)|} = \frac{\alpha}{|t-\beta|}.$$

This is the case if  $|t-\beta|$  is not too large.

Figure 4.17 displays the scalogram and the ridges of two hyperbolic chirps

$$f(t) = a_1 \cos\left(\frac{\alpha_1}{\beta_1-t}\right) + a_2 \cos\left(\frac{\alpha_2}{\beta_2-t}\right),$$

with  $\beta_1=0.68$  and  $\beta_2=0.72$ . As opposed to the windowed Fourier ridges shown in Figure 4.14, the wavelet ridges follow the rapid time modification of both instantaneous frequencies. This is particularly useful in analyzing the returns of hyperbolic chirps emitted by radar or sonar. Several techniques have been developed to detect chirps with wavelet ridges in the presence of noise [151, 455].

---

### **Better Is More Sparse**

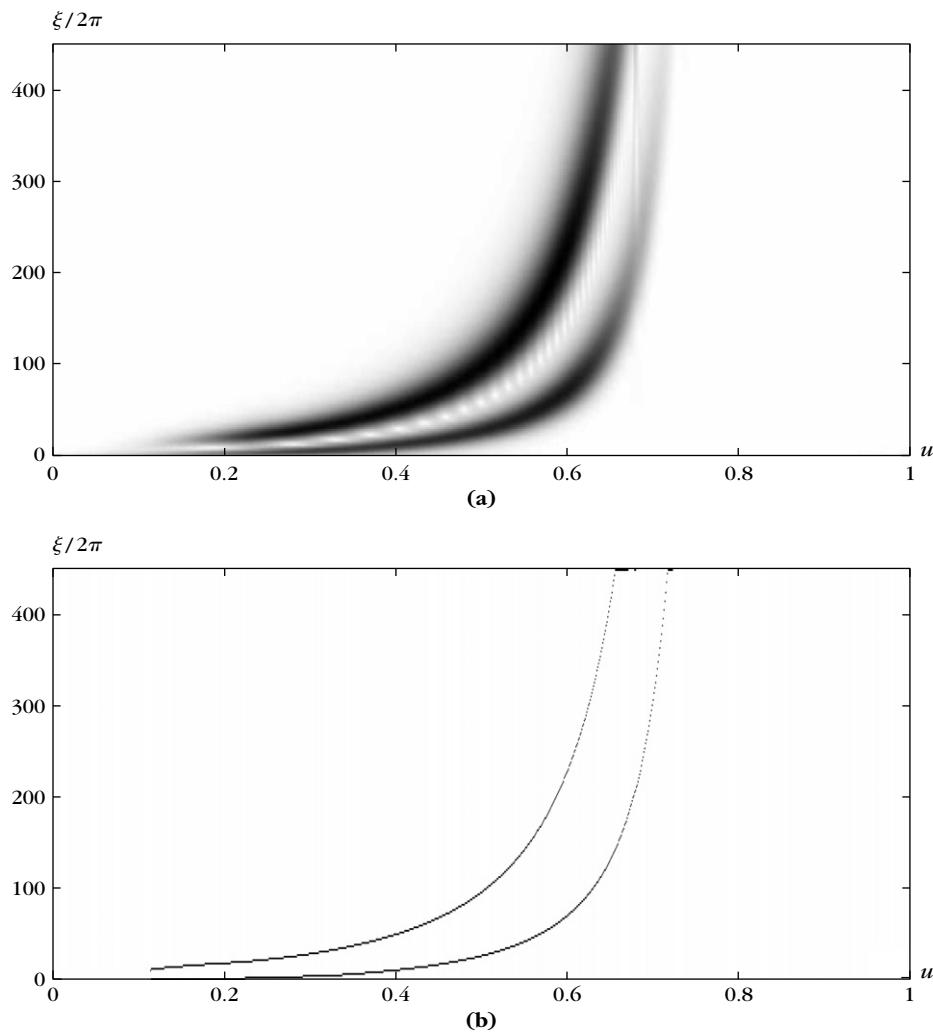
The linear and hyperbolic chirp examples show that the best transform depends on the signal time-frequency property. All examples also show that when the time-frequency transform has a resolution adapted to the signal time-frequency properties, the number of ridge points is reduced. Indeed, if signal structures do not match dictionary time-frequency atoms, then their energy is diffused over many more atoms, which produces more local maxima. Sparsity therefore appears as a natural criterion to adjust the resolution of time-frequency transforms.

Section 12.3.3 studies sparse time-frequency decompositions in very redundant Gabor time-frequency dictionaries, including windowed Fourier and wavelet atoms with a computationally more intense matching pursuit algorithm.

---

## 4.5 QUADRATIC TIME-FREQUENCY ENERGY

Wavelet and windowed-Fourier transforms are computed by correlating the signal with families of time-frequency atoms. The time and frequency resolution of these transforms is limited by the time-frequency resolution of the corresponding atoms. Ideally, one would like to define a density of energy in a time-frequency plane with no loss of resolution.

**FIGURE 4.17**

- (a) Normalized scalogram  $\eta^{-1}\xi P_W f(u, \xi)$  of two hyperbolic chirps shown in Figure 4.14.  
 (b) Wavelet ridges.

The Wigner-Ville distribution is a time-frequency energy density computed by correlating  $f$  with a time and frequency translation of itself. Despite its remarkable properties, the application of the Wigner-Ville distribution is limited by the existence of interference terms. Such interferences can be attenuated by time-frequency averaging but this results in a loss of resolution. It will be proved that the spectrogram, the scalogram, and all squared time-frequency decompositions can be written as a

time-frequency averaging of the Wigner-Ville distribution, which gives a common framework to relate the transforms.

### 4.5.1 Wigner-Ville Distribution

To analyze time-frequency structures, in 1948 Ville [475] introduced in signal processing a quadratic form that had been studied by Wigner [484] in a 1932 article on quantum thermodynamics:

$$P_V f(u, \xi) = \int_{-\infty}^{+\infty} f\left(u + \frac{\tau}{2}\right) f^*\left(u - \frac{\tau}{2}\right) e^{-i\tau\xi} d\tau. \quad (4.120)$$

The Wigner-Ville distribution remains real because it is the Fourier transform of  $f(u + \tau/2)f^*(u - \tau/2)$ , which has a Hermitian symmetry in  $\tau$ ; time and frequency have a symmetric role. This distribution can also be rewritten as a frequency integration by applying the Parseval formula:

$$P_V f(u, \xi) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \hat{f}\left(\xi + \frac{\gamma}{2}\right) \hat{f}^*\left(\xi - \frac{\gamma}{2}\right) e^{i\gamma u} d\gamma. \quad (4.121)$$

#### Time-Frequency Support

The Wigner-Ville transform localizes the time-frequency structures of  $f$ . If the energy of  $f$  is concentrated in time around  $u_0$  and in frequency around  $\xi_0$ , then  $P_V f$  has its energy centered at  $(u_0, \xi_0)$ , with a spread equal to the time and frequency spread of  $f$ . This property is illustrated by Theorem 4.7, which relates the time and frequency support of  $P_V f$  to the support of  $f$  and  $\hat{f}$ .

#### Theorem 4.7.

- If the support of  $f$  is  $[u_0 - T/2, u_0 + T/2]$ , then for all  $\xi$  the support in  $u$  of  $P_V f(u, \xi)$  is included in this interval.
- If the support of  $\hat{f}$  is  $[\xi_0 - \Delta/2, \xi_0 + \Delta/2]$ , then for all  $u$  the support in  $\xi$  of  $P_V f(u, \xi)$  is included in this interval.

**Proof.** Let  $\bar{f}(t) = f(-t)$ ; the Wigner-Ville distribution is rewritten as

$$P_V f(u, \xi) = \int_{-\infty}^{+\infty} f\left(\frac{\tau+2u}{2}\right) \bar{f}^*\left(\frac{\tau-2u}{2}\right) e^{-i\xi\tau} d\tau. \quad (4.122)$$

Suppose that  $f$  has a support equal to  $[u_0 - T/2, u_0 + T/2]$ . The supports of  $f(\tau/2 + u)$  and  $\bar{f}(\tau/2 - u)$  are then, respectively,

$$[2(u_0 - u) - T, 2(u_0 - u) + T] \text{ and } [-2(u_0 + u) - T, -2(u_0 + u) + T].$$

The Wigner-Ville integral (4.122) shows that  $P_V f(u, \xi)$  is nonzero if these two intervals overlap, which is the case only if  $|u_0 - u| < T$ . Support of  $P_V f(u, \xi)$  along  $u$  is therefore included in the support of  $f$ . If the support of  $\hat{f}$  is an interval, then the same derivation based on (4.121) shows that support of  $P_V f(u, \xi)$  along  $\xi$  is included in support of  $\hat{f}$ . ■

**EXAMPLE 4.17**

Theorem 4.7 proves that the Wigner-Ville distribution does not spread the time or frequency support of Diracs or sinusoids, unlike windowed Fourier and wavelet transforms. Direct calculations yield

$$f(t) = \delta(u - u_0) \Rightarrow P_V f(u, \xi) = \delta(u - u_0) \quad (4.123)$$

$$f(t) = \exp(i\xi_0 t) \Rightarrow P_V f(u, \xi) = \frac{1}{2\pi} \delta(\xi - \xi_0) \quad (4.124)$$

**EXAMPLE 4.18**

If  $f$  is a smooth and symmetric window, then its Wigner-Ville distribution  $P_V f(u, \xi)$  is concentrated in a neighborhood of  $u = \xi = 0$ . A Gaussian  $f(t) = (\sigma^2 \pi)^{-1/4} \exp(-t^2/(2\sigma^2))$  is transformed into a two-dimensional Gaussian because its Fourier transform is also Gaussian (2.32), and one can verify that

$$P_V f(u, \xi) = \frac{1}{\pi} \exp\left(\frac{-u^2}{\sigma^2} - \sigma^2 \xi^2\right). \quad (4.125)$$

In this particular case,  $P_V f(u, \xi) = |f(u)|^2 |\hat{f}(\xi)|^2$ .

The Wigner-Ville distribution has important invariance properties. A phase shift does not modify its value:

$$f(t) = e^{i\theta} g(t) \Rightarrow P_V f(u, \xi) = P_V g(u, \xi). \quad (4.126)$$

When  $f$  is translated in time or frequency, its Wigner-Ville transform is also translated:

$$f(t) = g(t - u_0) \Rightarrow P_V f(u, \xi) = P_V g(u - u_0, \xi) \quad (4.127)$$

$$f(t) = \exp(i\xi_0 t) g(t) \Rightarrow P_V f(u, \xi) = P_V g(u, \xi - \xi_0) \quad (4.128)$$

If  $f$  is scaled by  $s$  and thus  $\hat{f}$  is scaled by  $1/s$ , then the time and frequency parameters of  $P_V f$  are also scaled, respectively, by  $s$  and  $1/s$

$$f(t) = \frac{1}{\sqrt{s}} g\left(\frac{t}{s}\right) \Rightarrow P_V f(u, \xi) = P_V f\left(\frac{u}{s}, s\xi\right). \quad (4.129)$$

**EXAMPLE 4.19**

If  $g$  is a smooth and symmetric window, then  $P_V g(u, \xi)$  has its energy concentrated in the neighborhood of  $(0, 0)$ . The time-frequency atom

$$f_0(t) = \frac{a}{\sqrt{s}} \exp(i\theta_0) g\left(\frac{t - u_0}{s}\right) \exp(i\xi_0 t)$$

has a Wigner-Ville distribution that is calculated with (4.126), (4.127), and (4.128):

$$P_V f_0(u, \xi) = |a|^2 P_V g\left(\frac{u - u_0}{s}, s(\xi - \xi_0)\right). \quad (4.130)$$

Its energy therefore is concentrated in the neighborhood of  $(u_0, \xi_0)$ , on an ellipse with axes that are proportional to  $s$  in time and  $1/s$  in frequency.

---

### Instantaneous Frequency

Ville's original motivation for studying time-frequency decompositions was to compute the instantaneous frequency of a signal [475]. Let  $f_a$  be the analytical part of  $f$  obtained in (4.69) by setting  $\hat{f}(\omega)$  to zero for  $\omega < 0$ . We write  $f_a(t) = a(t) \exp[i\theta(t)]$  to define the instantaneous frequency  $\omega(t) = \theta'(t)$ . Theorem 4.8 proves that  $\theta'(t)$  is the “average” frequency computed relative to the Wigner-Ville distribution  $P_V f_a$ .

**Theorem 4.8.** If  $f_a(t) = a(t) \exp[i\theta(t)]$ , then

$$\theta'(u) = \frac{\int_{-\infty}^{+\infty} \xi P_V f_a(u, \xi) d\xi}{\int_{-\infty}^{+\infty} P_V f_a(u, \xi) d\xi}. \quad (4.131)$$

**Proof.** To prove this result, we verify that any function  $g$  satisfies

$$\begin{aligned} & \iint \xi g\left(u + \frac{\tau}{2}\right) g^*\left(u - \frac{\tau}{2}\right) \exp(-i\tau\xi) d\xi d\tau \\ &= -\pi i \left[ g'(u) g^*(u) - g(u) g^{*'}(u) \right]. \end{aligned} \quad (4.132)$$

This identity is obtained by observing that the Fourier transform of  $i\xi$  is the derivative of a Dirac, which gives an equality in the sense of distributions:

$$\int_{-\infty}^{+\infty} \xi \exp(-i\tau\xi) d\xi = -i 2\pi \delta'(\tau).$$

Since  $\int_{-\infty}^{+\infty} \delta'(\tau) h(\tau) d\tau = -h'(0)$ , inserting  $h(\tau) = g(u + \tau/2) g^*(u - \tau/2)$  proves (4.132). If  $g(u) = f_a(u) = a(u) \exp[i\theta(u)]$ , then (4.132) gives

$$\int_{-\infty}^{+\infty} \xi P_V f_a(u, \xi) d\xi = 2\pi a^2(u) \theta'(u).$$

We will see in (4.136) that  $|f_a(u)|^2 = (2\pi)^{-1} \int_{-\infty}^{+\infty} P_V f_a(u, \xi) d\xi$ , and since  $|f_a(u)|^2 = a(u)^2$ , we derive (4.131). ■

This theorem shows that for a fixed  $u$  the mass of  $P_V f_a(u, \xi)$  is typically concentrated in the neighborhood of the instantaneous frequency  $\xi = \theta'(u)$ . For example, a linear chirp  $f(t) = \exp(iat^2)$  is transformed into a Dirac located along the instantaneous frequency  $\xi = \theta'(u) = 2au$ :

$$P_V f(u, \xi) = \delta(\xi - 2au).$$

Similarly, the multiplication of  $f$  by a linear chirp  $\exp(iat^2)$  makes a frequency translation of  $P_V f$  by the instantaneous frequency  $2au$ :

$$f(t) = \exp(iat^2) g(t) \Rightarrow P_V f(u, \xi) = P_V g(u, \xi - 2au). \quad (4.133)$$

### Energy Density

The Moyal [379] formula proves that the Wigner-Ville transform is unitary, which implies energy-conservation properties.

**Theorem 4.9:** *Moyal.* For any  $f$  and  $g$  in  $\mathbf{L}^2(\mathbb{R})$ ,

$$\left| \int_{-\infty}^{+\infty} f(t) g^*(t) dt \right|^2 = \frac{1}{2\pi} \iint P_V f(u, \xi) P_V g(u, \xi) du d\xi. \quad (4.134)$$

**Proof.** Let us compute the integral

$$\begin{aligned} I &= \iint P_V f(u, \xi) P_V g(u, \xi) du d\xi \\ &= \iint \iint \int f\left(u + \frac{\tau}{2}\right) f^*\left(u - \frac{\tau}{2}\right) g\left(u + \frac{\tau'}{2}\right) g^*\left(u - \frac{\tau'}{2}\right) \\ &\quad \exp[-i\xi(\tau + \tau')] d\tau d\tau' du d\xi. \end{aligned}$$

The Fourier transform of  $h(t) = 1$  is  $\hat{h}(\omega) = 2\pi\delta(\omega)$ , which means that we have a distribution equality  $\int \exp[-i\xi(\tau + \tau')] d\xi = 2\pi\delta(\tau + \tau')$ . As a result,

$$\begin{aligned} I &= 2\pi \iint \int f\left(u + \frac{\tau}{2}\right) f^*\left(u - \frac{\tau}{2}\right) g\left(u + \frac{\tau'}{2}\right) g^*\left(u - \frac{\tau'}{2}\right) \delta(\tau + \tau') d\tau d\tau' du \\ &= 2\pi \int \int f\left(u + \frac{\tau}{2}\right) f^*\left(u - \frac{\tau}{2}\right) g\left(u - \frac{\tau}{2}\right) g^*\left(u + \frac{\tau}{2}\right) d\tau du. \end{aligned}$$

The change of variable  $t = u + \tau/2$  and  $t' = u - \tau/2$  yields (4.134). ■

One can consider  $|f(t)|^2$  and  $|\hat{f}(\omega)|^2/(2\pi)$  as energy densities in time and frequency that satisfy a conservation equation:

$$\|f\|^2 = \int_{-\infty}^{+\infty} |f(t)|^2 dt = \frac{1}{2\pi} \int_{-\infty}^{+\infty} |\hat{f}(\omega)|^2 d\omega.$$

Theorem 4.10 shows that these time and frequency densities are recovered with marginal integrals over the Wigner-Ville distribution.

**Theorem 4.10.** For any  $f \in \mathbf{L}^2(\mathbb{R})$ ,

$$\int_{-\infty}^{+\infty} P_V f(u, \xi) du = |\hat{f}(\xi)|^2 \quad (4.135)$$

and

$$\frac{1}{2\pi} \int_{-\infty}^{+\infty} P_V f(u, \xi) d\xi = |f(u)|^2. \quad (4.136)$$

**Proof.** The frequency integral (4.121) proves that the one-dimensional Fourier transform of  $g_\xi(u) = P_V f(u, \xi)$ , with respect to  $u$ , is

$$\hat{g}_\xi(\gamma) = \hat{f}\left(\xi + \frac{\gamma}{2}\right) \hat{f}^*\left(\xi - \frac{\gamma}{2}\right).$$

We derive (4.135) from the fact that it is

$$\hat{g}_\xi(0) = \int_{-\infty}^{+\infty} g_\xi(u) du.$$

Similarly, (4.120) shows that  $P_V f(u, \xi)$  is the one-dimensional Fourier transform of  $f(u + \tau/2)f^*(u - \tau/2)$  with respect to  $\tau$ , where  $\xi$  is the Fourier variable. Its integral in  $\xi$  therefore gives the value for  $\tau = 0$ , which is the identity (4.136). ■

This theorem suggests interpreting the Wigner-Ville distribution as a joint time-frequency energy density. However, the Wigner-Ville distribution misses one fundamental property of an energy density: positivity. Let us compute, for example, the Wigner-Ville distribution of  $f = \mathbf{1}_{[-T, T]}$  with the integral (4.120):

$$P_V f(u, \xi) = \frac{2 \sin(2(T - |u|)\xi)}{\xi} \mathbf{1}_{[-T, T]}(u).$$

It is an oscillating function that takes negative values. In fact, one can prove that translated and frequency-modulated Gaussians are the only functions with positive Wigner-Ville distributions. As we will see in the next section, to obtain positive energy distributions for all signals, it is necessary to average the Wigner-Ville transform, thus losing some time-frequency resolution.

## 4.5.2 Interferences and Positivity

At this point, the Wigner-Ville distribution may seem to be an ideal tool for analyzing the time-frequency structures of a signal. This, however, is not the case because of interferences created by the transform's quadratic properties. Interference can be removed by averaging the Wigner-Ville distribution with appropriate kernels that yield positive time-frequency densities. However, this reduces time-frequency resolution. Spectrograms and scalograms are examples of positive quadratic distributions obtained by smoothing the Wigner-Ville distribution.

### Cross Terms

Let  $f = f_1 + f_2$  be a composite signal. Since the Wigner-Ville distribution is a quadratic form,

$$P_V f = P_V f_1 + P_V f_2 + P_V[f_1, f_2] + P_V[f_2, f_1], \quad (4.137)$$

where  $P_V[h, g]$  is the cross Wigner-Ville distribution of two signals:

$$P_V[h, g](u, \xi) = \int_{-\infty}^{+\infty} h\left(u + \frac{\tau}{2}\right) g^*\left(u - \frac{\tau}{2}\right) e^{-i\tau\xi} d\tau. \quad (4.138)$$

The interference term

$$I[f_1, f_2] = P_V[f_1, f_2] + P_V[f_2, f_1]$$

is a real function that creates nonzero values at unexpected locations of the  $(u, \xi)$  plane.

Let us consider two time-frequency atoms defined by

$$f_1(t) = a_1 e^{i\theta_1} g(t - u_1) e^{i\xi_1 t} \text{ and } f_2(t) = a_2 e^{i\theta_2} g(t - u_2) e^{i\xi_2 t},$$

where  $g$  is a time window centered at  $t = 0$ . Their Wigner-Ville distributions computed in (4.130) are

$$P_V f_1(u, \xi) = a_1^2 P_V g(u - u_1, \xi - \xi_1) \text{ and } P_V f_2(u, \xi) = a_2^2 P_V g(u - u_2, \xi - \xi_2).$$

Since the energy of  $P_V g$  is centered at  $(0, 0)$ , the energy of  $P_V f_1$  and  $P_V f_2$  is concentrated in the neighborhoods of  $(u_1, \xi_1)$  and  $(u_2, \xi_2)$ , respectively. A direct calculation verifies that the interference term is

$$I[f_1, f_2](u, \xi) = 2a_1 a_2 P_V g(u - u_0, \xi - \xi_0) \cos((u - u_0)\Delta\xi - (\xi - \xi_0)\Delta u + \Delta\theta)$$

with

$$u_0 = \frac{u_1 + u_2}{2}, \quad \xi_0 = \frac{\xi_1 + \xi_2}{2}$$

$$\Delta u = u_1 - u_2, \quad \Delta\xi = \xi_1 - \xi_2$$

$$\Delta\theta = \theta_1 - \theta_2 + u_0 \Delta\xi.$$

The interference term is an oscillatory waveform centered at the middle point  $(u_0, \xi_0)$ . This is quite counterintuitive because  $f$  and  $\hat{f}$  have very little energy in the neighborhood of  $u_0$  and  $\xi_0$ . The frequency of the oscillations is proportional to the Euclidean distance  $\sqrt{\Delta\xi^2 + \Delta u^2}$  of  $(u_1, \xi_1)$  and  $(u_2, \xi_2)$ . The direction of the oscillations is perpendicular to the line that joins  $(u_1, \xi_1)$  and  $(u_2, \xi_2)$ . Figure 4.18 on the next page displays the Wigner-Ville distribution of two atoms obtained with a Gaussian window  $g$ . Oscillating interference appears at the middle time-frequency point.

This figure's example shows that interference  $I[f_1, f_2](u, \xi)$  has some energy in regions where  $|f(u)|^2 \approx 0$  and  $|\hat{f}(\xi)|^2 \approx 0$ . Such interferences can have a complicated structure [26, 302], but they are necessarily oscillatory because the marginal integrals (4.135) and (4.136) vanish:

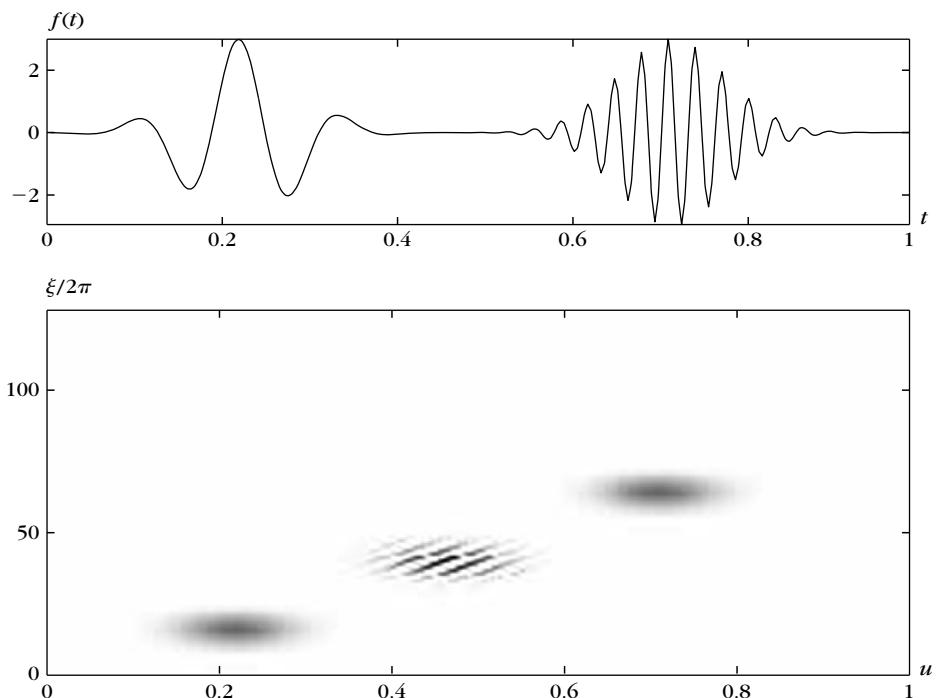
$$\int_{-\infty}^{+\infty} P_V f(u, \xi) d\xi = 2\pi |f(u)|^2, \quad \int_{-\infty}^{+\infty} P_V f(u, \xi) du = |\hat{f}(\xi)|^2.$$

### Analytic Part

Interference terms also exist in a real signal  $f$  with a single instantaneous frequency component. Let  $f_a(t) = a(t) \exp[i\theta(t)]$  be its analytic part:

$$f = \operatorname{Re}[f_a] = \frac{1}{2} (f_a + f_a^*).$$

Theorem 4.8 proves that for fixed  $u$ ,  $P_V f_a(u, \xi)$  and  $P_V f_a^*(u, \xi)$  have an energy concentrated, respectively, in the neighborhood of  $\xi_1 = \theta'(u)$  and  $\xi_2 = -\theta'(u)$ . Both

**FIGURE 4.18**

Wigner-Ville distribution  $P_V f(u, \xi)$  of two Gabor atoms (top); the oscillating interferences are centered at the middle time-frequency location.

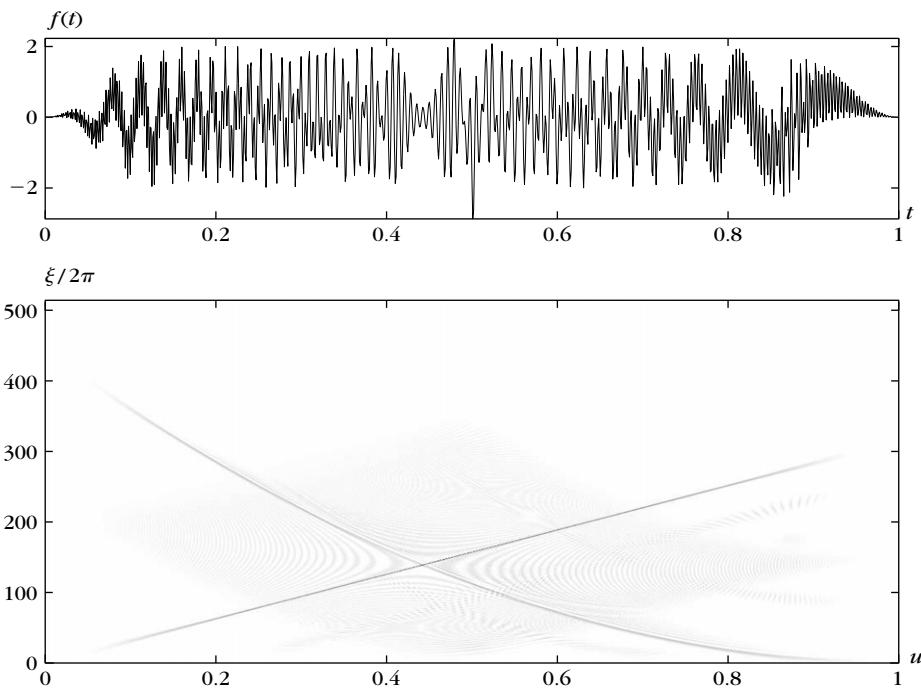
components create an interference term at the intermediate zero frequency  $\xi_0 = (\xi_1 + \xi_2)/2 = 0$ . To avoid this low-frequency interference, we often compute  $P_V f_a$  as opposed to  $P_V f$ .

Figure 4.19 displays  $P_V f_a$  for a real signal  $f$  that includes a linear chirp, a quadratic chirp, and two isolated time-frequency atoms. The linear and quadratic chirps are localized along narrow time-frequency lines; they are spread on wider bands by the spectrogram shown in Figure 4.3 and the scalogram shown in Figure 4.4, respectively. However, interference terms create complex oscillatory patterns that make it difficult to detect the existence of the two time-frequency transients at  $t = 0.5$  and  $t = 0.87$ , which clearly appear in the spectrogram and the scalogram.

### **Positivity**

Since interference terms include positive and negative oscillations, they can be partially removed by smoothing  $P_V f$  with a kernel  $K$ :

$$P_K f(u, \xi) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} P_V f(u', \xi') K(u, u', \xi, \xi') du' d\xi'. \quad (4.139)$$

**FIGURE 4.19**

The Wigner-Ville distribution  $P_V f_a(u, \xi)$  (bottom) of the analytic part of the signal (top).

The time-frequency resolution of this distribution depends on the spread of the kernel  $K$  in the neighborhood of  $(u, \xi)$ . Since interferences take negative values, one can guarantee that all interferences are removed by imposing that this time-frequency distribution remain positive  $P_K f(u, \xi) \geq 0$  for all  $(u, \xi) \in \mathbb{R}^2$ .

The spectrogram (4.12) and scalogram (4.55) are examples of positive time-frequency energy distributions. In general, let us consider a family of time-frequency atoms  $\{\phi_\gamma\}_{\gamma \in \Gamma}$ . Suppose that for any  $(u, \xi)$  there exists a unique atom  $\phi_{\gamma(u, \xi)}$  centered in time and frequency at  $(u, \xi)$ . The resulting time-frequency energy density is

$$Pf(u, \xi) = |\langle f, \phi_{\gamma(u, \xi)} \rangle|^2.$$

The Moyal formula (4.134) proves that this energy density can be written as a time-frequency averaging of the Wigner-Ville distribution:

$$Pf(u, \xi) = \frac{1}{2\pi} \int \int P_V f(u', \xi') P_V \phi_{\gamma(u, \xi)}(u', \xi') du' d\xi'. \quad (4.140)$$

The smoothing kernel is the Wigner-Ville distribution of the atoms:

$$K(u, u', \xi, \xi') = \frac{1}{2\pi} P_V \phi_{\gamma(u, \xi)}(u', \xi').$$

The loss of time-frequency resolution depends on the spread of the distribution  $P_V \phi_{\gamma(u,\xi)}(u', \xi')$  in the neighborhood of  $(u, v)$ .

---

**EXAMPLE 4.20**

A spectrogram is computed with windowed Fourier atoms:

$$\phi_{\gamma(u,\xi)}(t) = g(t - u) e^{i\xi t}.$$

The Wigner-Ville distribution calculated in (4.130) yields

$$K(u, u', \xi, \xi') = \frac{1}{2\pi} P_V \phi_{\gamma(u,\xi)}(u', \xi') = \frac{1}{2\pi} P_V g(u' - u, \xi' - \xi). \quad (4.141)$$

For a spectrogram, the Wigner-Ville averaging (4.140) is therefore a two-dimensional convolution with  $P_V g$ . If  $g$  is a Gaussian window, then  $P_V g$  is a two-dimensional Gaussian. This proves that averaging  $P_V f$  with a sufficiently wide Gaussian defines a positive energy density. The general class of time-frequency distributions obtained by convolving  $P_V f$  with an arbitrary kernel  $K$  is studied in Section 4.5.3.

---

**EXAMPLE 4.21**

Let  $\psi$  be an analytic wavelet with a center frequency that is  $\eta$ . The wavelet atom  $\psi_{u,s}(t) = s^{-1/2} \psi((t - u)/s)$  is centered at  $(u, \xi = \eta/s)$ , and the scalogram is defined by

$$P_W f(u, \xi) = |\langle f, \psi_{u,s} \rangle|^2 \quad \text{for } \xi = \eta/s.$$

Properties (4.127, 4.129) prove that the averaging kernel is

$$K(u, u', \xi, \xi') = \frac{1}{2\pi} P_V \psi \left( \frac{u' - u}{s}, s\xi' \right) = \frac{1}{2\pi} P_V \psi \left( \frac{\xi}{\eta} (u' - u), \frac{\eta}{\xi} \xi' \right).$$


---

Positive time-frequency distributions totally remove the interference terms but produce a loss of resolution. This is emphasized by Theorem 4.11 described by Wigner [485].

**Theorem 4.11:** *Wigner.* There is no positive quadratic energy distribution  $Pf$  that satisfies

$$\int_{-\infty}^{+\infty} Pf(u, \xi) d\xi = 2\pi |f(u)|^2 \quad \text{and} \quad \int_{-\infty}^{+\infty} Pf(u, \xi) du = |\hat{f}(\xi)|^2. \quad (4.142)$$

**Proof.** Suppose that  $Pf$  is a positive quadratic distribution that satisfies these marginals.

Since  $Pf(u, \xi) \geq 0$ , the integrals (4.142) imply that, if the support of  $f$  is included in an interval  $I$ , then  $Pf(u, \xi) = 0$  for  $u \notin I$ . We can associate to the quadratic form  $Pf$  a bilinear distribution defined for any  $f$  and  $g$  by

$$P[f, g] = \frac{1}{4} (P(f+g) - P(f-g)).$$

Let  $f_1$  and  $f_2$  be two nonzero signals having their support in two disjoint intervals  $I_1$  and  $I_2$  so that  $f_1 f_2 = 0$ . Let  $f = af_1 + bf_2$ :

$$Pf = |a|^2 Pf_1 + ab^* P[f_1, f_2] + a^* b P[f_2, f_1] + |b|^2 Pf_2.$$

Since  $I_1$  does not intersect  $I_2$ ,  $Pf_1(u, \xi) = 0$  for  $u \in I_2$ . Remember that  $Pf(u, \xi) \geq 0$  for all  $a$  and  $b$ , so necessarily  $P[f_1, f_2](u, \xi) = P[f_2, f_1](u, \xi) = 0$  for  $u \in I_2$ . Similarly, we can prove that these cross terms are zero for  $u \in I_1$ ; therefore

$$Pf(u, \xi) = |a|^2 Pf_1(u, \xi) + |b|^2 Pf_2(u, \xi).$$

Integrating this equation and inserting (4.142) yields

$$|\hat{f}(\xi)|^2 = |a|^2 |\hat{f}_1(\xi)|^2 + |b|^2 |\hat{f}_2(\xi)|^2.$$

Since  $\hat{f}(\xi) = a\hat{f}_1(\xi) + b\hat{f}_2(\xi)$ , it follows that  $\hat{f}_1(\xi)\hat{f}_2^*(\xi) = 0$ . But this is not possible because  $f_1$  and  $f_2$  have a compact support in time and Theorem 2.7 proves that  $\hat{f}_1$  and  $\hat{f}_2$  are  $C^\infty$  functions that cannot vanish on a whole interval. Thus, we conclude that one cannot construct a positive quadratic distribution  $Pf$  that satisfies the marginals (4.142). ■

### 4.5.3 Cohen's Class

While attenuating the interference terms with a smoothing kernel  $K$ , we may want to retain certain important properties. Cohen [177] introduced a general class of quadratic time-frequency distributions that satisfy the time translation and frequency modulation invariance properties (4.127) and (4.128). If a signal is translated in time or frequency, its energy distribution is translated just by the corresponding amount. This was the beginning of a systematic study of quadratic time-frequency distributions obtained as a weighted average of a Wigner-Ville distribution [8, 26, 178, 301].

Section 2.1 proves that linear translation-invariant operators are convolution products. The translation-invariance properties (4.127, 4.128) therefore are equivalent to having a smoothing kernel in (4.139) be a convolution kernel:

$$K(u, u', \xi, \xi') = K(u - u', \xi - \xi'); \quad (4.143)$$

therefore

$$P_K f(u, \xi) = P_V f \star K(u, \xi) = \int \int K(u - u', \xi - \xi') P_V f(u', \xi') du' d\xi'. \quad (4.144)$$

The spectrogram is an example of Cohen's class distribution with a kernel in (4.141) that is the Wigner-Ville distribution of the window:

$$K(u, \xi) = \frac{1}{2\pi} P_V g(u, \xi) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} g\left(u + \frac{\tau}{2}\right) g\left(u - \frac{\tau}{2}\right) e^{-i\tau\xi} d\tau. \quad (4.145)$$

### Ambiguity Function

The properties of the convolution (4.144) are more easily studied by calculating the two-dimensional Fourier transform of  $P_V f(u, \xi)$  with respect to  $u$  and  $\xi$ . We denote by  $Af(\tau, \gamma)$  this Fourier transform:

$$Af(\tau, \gamma) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} P_V f(u, \xi) \exp[-i(u\gamma + \xi\tau)] du d\xi.$$

Note that the Fourier variables  $\tau$  and  $\gamma$  are inverted with respect to the usual Fourier notation. Since the one-dimensional Fourier transform of  $P_V f(u, \xi)$ , with respect to  $u$ , is  $\hat{f}(\xi + \gamma/2)\hat{f}^*(\xi - \gamma/2)$ , applying the one-dimensional Fourier transform with respect to  $\xi$  gives

$$Af(\tau, \gamma) = \int_{-\infty}^{+\infty} \hat{f}\left(\xi + \frac{\gamma}{2}\right) \hat{f}^*\left(\xi - \frac{\gamma}{2}\right) e^{-i\tau\xi} d\xi. \quad (4.146)$$

The Parseval formula yields

$$Af(\tau, \gamma) = \int_{-\infty}^{+\infty} f\left(u + \frac{\tau}{2}\right) f^*\left(u - \frac{\tau}{2}\right) e^{-i\gamma u} du. \quad (4.147)$$

We recognize the *ambiguity function* encountered in (4.24) when studying the time-frequency resolution of a windowed Fourier transform. It measures the energy concentration of  $f$  in time and in frequency.

### Kernel Properties

The Fourier transform of  $K(u, \xi)$  is

$$\hat{K}(\tau, \gamma) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} K(u, \xi) \exp[-i(u\gamma + \xi\tau)] du d\xi.$$

As in the definition of the ambiguity function (4.146), the Fourier parameters  $\tau$  and  $\gamma$  of  $\hat{K}$  are inverted. Theorem 4.12 gives necessary and sufficient conditions to ensure that  $P_K$  satisfies marginal energy properties such as those of the Wigner-Ville distribution. Wigner's Theorem 4.11 proves that in this case  $P_K f(u, \xi)$  takes negative values.

**Theorem 4.12.** For all  $f \in \mathbf{L}^2(\mathbb{R})$ ,

$$\int_{-\infty}^{+\infty} P_K f(u, \xi) d\xi = 2\pi |f(u)|^2, \quad \int_{-\infty}^{+\infty} P_K f(u, \xi) du = |\hat{f}(\xi)|^2, \quad (4.148)$$

if and only if

$$\forall (\tau, \gamma) \in \mathbb{R}^2, \quad \hat{K}(\tau, 0) = \hat{K}(0, \gamma) = 1. \quad (4.149)$$

**Proof.** Let  $A_K f(\tau, \gamma)$  be the two-dimensional Fourier transform of  $P_K f(u, \xi)$ . The Fourier integral at  $(0, \gamma)$  gives

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} P_K f(u, \xi) e^{-iu\gamma} d\xi du = A_K f(0, \gamma). \quad (4.150)$$

Since the ambiguity function  $Af(\tau, \gamma)$  is the Fourier transform of  $P_V f(u, \xi)$ , the two-dimensional convolution (4.144) gives

$$A_K(\tau, \gamma) = Af(\tau, \gamma) \hat{K}(\tau, \gamma). \quad (4.151)$$

The Fourier transform of  $2\pi|f(u)|^2$  is  $\hat{f} \star \bar{\hat{f}}(\gamma)$ , with  $\bar{\hat{f}}(\gamma) = \hat{f}^*(-\gamma)$ . The relation (4.150) shows that (4.148) is satisfied, if and only if,

$$A_K f(0, \gamma) = Af(0, \gamma) \hat{K}(0, \gamma) = \hat{f} \star \bar{\hat{f}}(\gamma). \quad (4.152)$$

Since  $P_V f$  satisfies the marginal property (4.135), we similarly prove that

$$Af(0, \gamma) = \hat{f} \star \bar{\hat{f}}(\gamma).$$

Requiring (4.152) to be valid for any  $\hat{f}(\gamma)$  is equivalent to requiring that  $\hat{K}(0, \gamma) = 1$  for all  $\gamma \in \mathbb{R}$ . The same derivation applied to other marginal integration yields  $\hat{K}(\xi, 0) = 1$ . ■

In addition to requiring time-frequency translation invariance, it may be useful to guarantee that  $P_K$  satisfies the same scaling property as a Wigner-Ville distribution:

$$g(t) = \frac{1}{\sqrt{s}} f\left(\frac{t}{s}\right) \Rightarrow P_K g(u, \xi) = P_K f\left(\frac{u}{s}, s\xi\right).$$

Such a distribution  $P_K$  is *affine* invariant. One can verify that affine invariance is equivalent to imposing that

$$\forall s \in \mathbb{R}^+, \quad K\left(su, \frac{\xi}{s}\right) = K(u, \xi), \quad (4.153)$$

therefore

$$K(u, \xi) = K(u\xi, 1) = \beta(u\xi).$$

### EXAMPLE 4.22

The Rihaczek distribution is an affine invariant distribution whose convolution kernel is

$$\hat{K}(\tau, \gamma) = \exp\left(\frac{i\tau\gamma}{2}\right). \quad (4.154)$$

A direct calculation shows that

$$P_K f(u, \xi) = f(u) \hat{f}^*(\xi) \exp(-iu\xi). \quad (4.155)$$

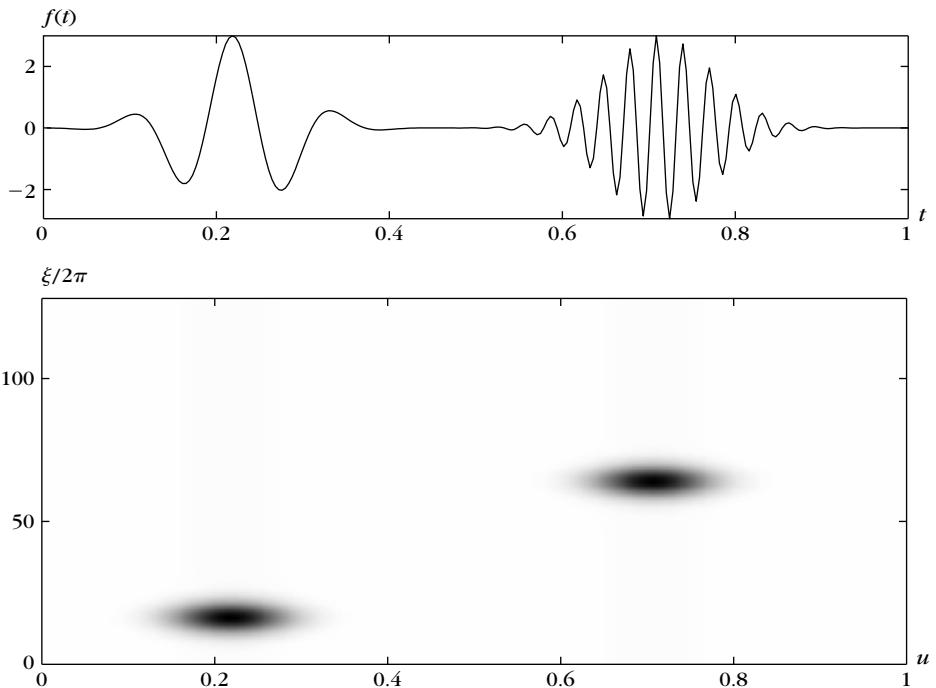
**EXAMPLE 4.23**

The kernel of the *Choi-William* distribution is [161]

$$\hat{K}(\tau, \gamma) = \exp(-\sigma^2 \tau^2 \gamma^2). \quad (4.156)$$

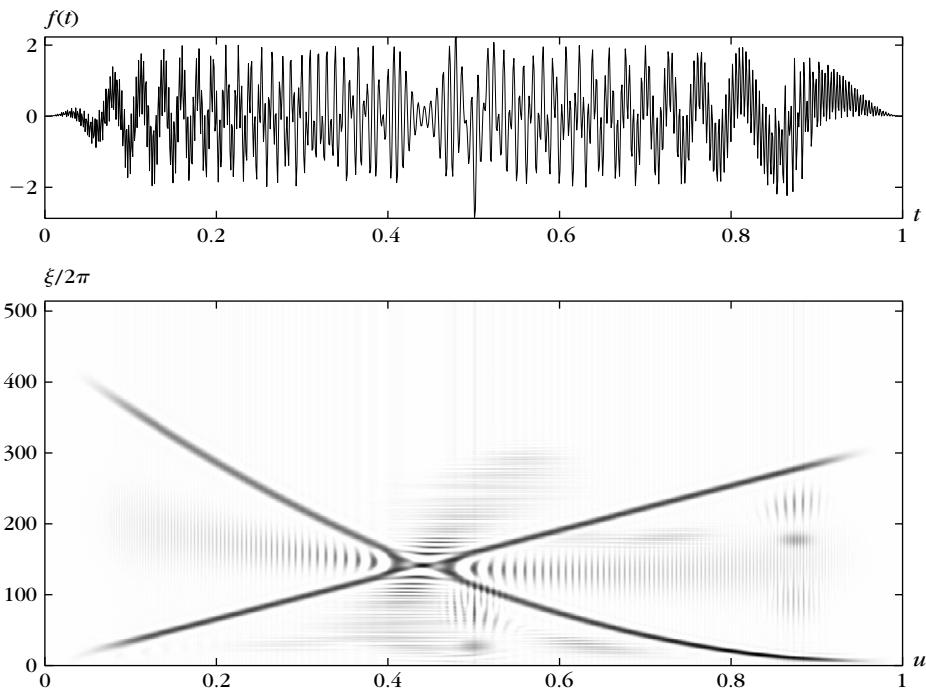
It is symmetric and thus corresponds to a real function  $K(u, \xi)$ . This distribution satisfies the marginal conditions (4.149). Since  $\lim_{\sigma \rightarrow 0} \hat{K}(\tau, \gamma) = 1$ , when  $\sigma$  is small the Choi-William distribution is close to a Wigner-Ville distribution. Increasing  $\sigma$  attenuates the interference terms but spreads  $K(u, \xi)$ , which reduces the time-frequency resolution of the distribution.

Figure 4.20 shows that the interference terms of two modulated Gaussians nearly disappear when the Wigner-Ville distribution of Figure 4.18 is averaged by a Choi-William kernel having a sufficiently large  $\sigma$ . Figure 4.21 gives the Choi-William distribution corresponding to the Wigner-Ville distribution shown in Figure 4.19. The energy of the linear and quadratic chirps are spread over wider time-frequency bands but the interference terms are attenuated, although not totally removed. It remains difficult to isolate the two modulated Gaussians at  $t = 0.5$  and  $t = 0.87$ , which clearly appear in the spectrogram of Figure 4.3.



**FIGURE 4.20**

Choi-William distribution  $P_K f(u, \xi)$  of two Gabor atoms (*top*); the interference term that appears in the Wigner-Ville distribution of Figure 4.18 has nearly disappeared.

**FIGURE 4.21**

Choi-William distribution  $P_K f_a(u, \xi)$  of signal's analytic part (top); the interferences remain visible.

#### 4.5.4 Discrete Wigner-Ville Computations

The Wigner integral (4.120) is the Fourier transform of  $f(u + \tau/2)f^*(u - \tau/2)$ :

$$P_V f(u, \xi) = \int_{-\infty}^{+\infty} f\left(u + \frac{\tau}{2}\right) f^*\left(u - \frac{\tau}{2}\right) e^{-i\tau\xi} d\tau. \quad (4.157)$$

For a discrete signal  $f[n]$ , defined over  $0 \leq n < N$ , the integral is replaced by a discrete sum:

$$P_V f[n, k] = \sum_{p=-N}^{N-1} f\left[n + \frac{p}{2}\right] f^*\left[n - \frac{p}{2}\right] \exp\left(\frac{-i2\pi kp}{N}\right). \quad (4.158)$$

When  $p$  is odd, this calculation requires knowing the value of  $f$  at half integers. Such values are computed by interpolating  $f$  by adding zeroes to its Fourier transform. This is necessary to avoid the aliasing produced by the discretization of the Wigner-Ville integral [165].

The interpolation  $\tilde{f}$  of  $f$  is a signal of size  $2N$  that has a DFT  $\hat{\tilde{f}}$  defined from the discrete Fourier transform  $\hat{f}$  of  $f$  by

$$\hat{\tilde{f}}[k] = \begin{cases} 2\hat{f}[k] & \text{if } 0 \leq k < N/2 \\ 0 & \text{if } N/2 < k < 3N/2 \\ 2\hat{f}[k-N] & \text{if } 3N/2 < k < 2N \\ \hat{f}[N/2] & \text{if } k = N/2, 3N/2 \end{cases}$$

Computing the inverse DFT shows that  $\tilde{f}[2n] = f[n]$  for  $n \in [0, N-1]$ . When  $n \notin [0, 2N-1]$ , we set  $\tilde{f}[n] = 0$ . The Wigner summation (4.158) is calculated from  $\tilde{f}$ :

$$\begin{aligned} P_V f[n, k] &= \sum_{p=-N}^{N-1} \tilde{f}[2n+p] \tilde{f}^*[2n-p] \exp\left(\frac{-i2\pi kp}{N}\right) \\ &= \sum_{p=0}^{2N-1} \tilde{f}[2n+p-N] \tilde{f}^*[2n-p+N] \exp\left(\frac{-i2\pi(2k)p}{2N}\right). \end{aligned}$$

For  $0 \leq n < N$  fixed,  $P_V f[n, k]$  is the discrete Fourier transform of size  $2N$  of  $g[p] = \tilde{f}[2n+p-N] \tilde{f}^*[2n-p+N]$  at the frequency  $2k$ . Thus, the discrete Wigner-Ville distribution is calculated with  $N$  FFT procedures of size  $2N$ , which requires  $O(N^2 \log N)$  operations. To compute the Wigner-Ville distribution of the analytic part  $f_a$  of  $f$ , we use (4.48).

### Cohen's Class

A Cohen's class distribution is calculated with a circular convolution of the discrete Wigner-Ville distribution with a kernel  $K[p, q]$ :

$$P_K[n, k] = P_V \circledast K[n, k]. \quad (4.159)$$

Its two-dimensional discrete Fourier transform is therefore

$$A_K[p, q] = A f[p, q] \hat{K}[p, q]. \quad (4.160)$$

The signal  $A f[p, q]$  is the discrete ambiguity function, calculated with a two-dimensional FFT of the discrete Wigner-Ville distribution  $P_V f[n, k]$ . As in the case of continuous time, we have inverted the index  $p$  and  $q$  of the usual two-dimensional Fourier transform. The Cohen's class distribution (4.159) is obtained by calculating the inverse Fourier transform of (4.160). This also requires a total of  $O(N^2 \log N)$  operations.

---

## 4.6 EXERCISES

**4.1** <sup>2</sup> *Instantaneous frequency.* Let  $f(t) = \exp[i\phi(t)]$ .

- (a) Prove that  $\int_{-\infty}^{+\infty} |Sf(u, \xi)|^2 d\xi = 2\pi$ . Hint:  $Sf(u, \xi)$  is a Fourier transform; use the Parseval formula.
- (b) Similarly, show that

$$\int_{-\infty}^{+\infty} \xi |Sf(u, \xi)|^2 d\xi = 2\pi \int_{-\infty}^{+\infty} \phi'(t) |g(t-u)|^2 dt,$$

and interpret this result.

**4.2** <sup>1</sup> When  $g(t) = (\pi\sigma^2)^{-1/4} \exp(-t^2/(2\sigma^2))$ , compute the ambiguity function  $A_g(\tau, \gamma)$ .

**4.3** <sup>1</sup> Prove that the approximate reconstruction formula (4.66) is exact if and only if

$$\frac{\log_e a}{C_\psi} \sum_{j=I}^J \frac{1}{a^j} \hat{\psi}_j[k] + \frac{1}{C_\psi a^J} \hat{\phi}_J[k] = 1.$$

Compute numerically the left equation value for different  $a$  when  $\psi_j[n]$  is constructed from the Gabor wavelet (4.60) and (4.62) with  $\sigma = 1$  and  $2^J < N/4$ .

**4.4** <sup>1</sup> Let  $g[n]$  be a window with  $L$  nonzero coefficients. For signals of size  $N$ , describe a fast algorithm that computes the discrete windowed Fourier transform (4.27) with  $O(N \log_2 L)$  operations.

**4.5** <sup>3</sup> Let  $K$  be the reproducing kernel (4.21) of a windowed Fourier transform:  $K(u_0, u, \xi_0, \xi) = \langle g_{u,\xi}, g_{u_0,\xi_0} \rangle$ .

- (a) For any  $\Phi \in \mathbf{L}^2(\mathbb{R}^2)$  we define

$$P\Phi(u_0, \xi_0) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \Phi(u, \xi) K(u_0, u, \xi_0, \xi) du d\xi.$$

Prove that  $P$  is an orthogonal projector on the space of functions  $\Phi(u, \xi)$  that are windowed Fourier transforms of functions in  $\mathbf{L}^2(\mathbb{R})$ .

(b) Suppose that for all  $(u, \xi) \in \mathbb{R}^2$  we are given  $\tilde{S}f(u, \xi) = Q(Sf(u, \xi))$ , which is a quantization of the windowed Fourier coefficients. How can we reduce the norm  $\mathbf{L}^2(\mathbb{R}^2)$  of the quantification error  $\varepsilon(u, \xi) = Sf(u, \xi) - Q(Sf(u, \xi))$ ?

**4.6** <sup>3</sup> Prove the wavelet reconstruction formula (4.45).

**4.7** <sup>3</sup> Prove that a scaling function  $\phi$  defined by (4.42) satisfies  $\|\phi\| = 1$ .

- 4.8** <sup>2</sup> Let  $\psi$  be a real and even wavelet such that  $C = \int_0^{+\infty} \omega^{-1} |\hat{\psi}(\omega)| d\omega < +\infty$ . Prove that

$$\forall f \in L^2(\mathbb{R}), \quad f(t) = \frac{1}{C} \int_0^{+\infty} Wf(t, s) \frac{ds}{s^{3/2}}. \quad (4.161)$$

- 4.9** <sup>3</sup> *Analytic continuation.* Let  $f \in L^2(\mathbb{R})$  be a function such that  $\hat{f}(\omega) = 0$  for  $\omega < 0$ . For any complex  $z \in \mathbb{C}$  such that  $\text{Im}(z) \geq 0$ , we define

$$f^{(p)}(z) = \frac{1}{\pi} \int_0^{+\infty} (i\omega)^p \hat{f}(\omega) e^{iz\omega} d\omega.$$

- (a) Verify that if  $f$  is  $C^p$  then  $f^{(p)}(t)$  is the derivative of order  $p$  of  $f(t)$ .
- (b) Prove that if  $\text{Im}(z) > 0$ , then  $f^{(p)}(z)$  is differentiable relative to the complex variable  $z$ . Such a function is said to be *analytic* on the upper-half complex plane.
- (c) Prove that this analytic extension can be written as a wavelet transform

$$f^{(p)}(x + iy) = y^{-p-1/2} Wf(x, y),$$

calculated with an analytic wavelet  $\psi$  that you will specify.

- 4.10** <sup>2</sup> Let  $f(t) = \cos(a \cos bt)$ . We want to compute precisely the instantaneous frequency of  $f$  from the ridges of its windowed Fourier transform. Find a necessary condition on the window support as a function of  $a$  and  $b$ . If  $f(t) = \cos(a \cos bt) + \cos(a \cos bt + ct)$ , find a condition on  $a$ ,  $b$ , and  $c$  in order to measure both instantaneous frequencies with the ridges of a windowed Fourier transform. Verify your calculations with a numerical implementation.
- 4.11** <sup>4</sup> *Noise removal.* We want to suppress noise from audio signals by thresholding ridge coefficients. Implement a dual-synthesis algorithm that reconstructs audio signal approximations from windowed Fourier ridge points (4.96) or wavelet ridge points (4.116), with the conjugate-gradient inverse frame algorithm of Theorem 5.8. Study the SNR of audio denoising by thresholding the ridge coefficients. Try to improve this SNR by averaging the spectrogram values along ridges. Compare the SNR with a linear filtering estimator.
- 4.12** <sup>4</sup> *Sound duration.* Make a program that modifies the sound duration with the formula (4.72) by modifying the ridges of a window Fourier transform with (4.99) or of a wavelet transform with (4.118), and by reconstructing a signal with a dual synthesis.
- 4.13** <sup>4</sup> *Sound transposition.* Implement a sound transposition, with windowed Fourier or wavelet ridges, with the transposition model (4.73). The resulting modifications of the ridge supports are specified by (4.100) and (4.119). The amplitude of the transposed harmonics can be computed with the auto-regressive model (4.75). A signal is restored with a dual-synthesis algorithm.

- 4.14** <sup>4</sup> The sinusoidal model (4.71) is improved for speech signals by adding a nonharmonic component  $B(t)$  to the partials [339]:

$$F(t) = \sum_{k=1}^K a_k(t) \cos \phi_k(t) + B(t). \quad (4.162)$$

Given a signal  $f(t)$  that is considered to be a realization of  $F(t)$ , compute the ridges of a windowed Fourier transform, and find the “main” partials and compute their amplitude  $a_k$  and phase  $\phi_k$ . These partials are subtracted from the signal. Over intervals of fixed size, the residue is modeled as the realization of an autoregressive process  $B(t)$ , of order 10 to 15. Use a standard algorithm to compute the parameters of this autoregressive process [57]. Evaluate the audio quality of the sound restored from the calculated model (4.162). Study an application to audio compression by quantizing and coding the parameters of the model.

- 4.15** <sup>2</sup> Prove that  $Pf(u, \xi) = \|f\|^{-2} |f(u)|^2 |\hat{f}(\xi)|^2$  satisfies the marginal properties (4.135) and (4.136). Why can’t we apply the Wigner Theorem 4.11?
- 4.16** <sup>1</sup> Let  $g_\sigma$  be a Gaussian of variance  $\sigma^2$ . Prove that  $P_\theta f(u, \xi) = P_V f \star \theta(u, \xi)$  is a positive distribution if  $\theta(u, \xi) = g_\sigma(u) g_\beta(\xi)$  with  $\sigma \beta \geq 1/2$ . Hint: Consider a spectrogram calculated with a Gaussian window.
- 4.17** <sup>3</sup> Let  $\{g_n(t)\}_{n \in \mathbb{N}}$  be an orthonormal basis of  $\mathbf{L}^2(\mathbb{R})$ . Prove that

$$\forall (t, \omega) \in \mathbb{R}^2, \quad \sum_{n=0}^{+\infty} P_V g_n(t, \omega) = 1.$$

- 4.18** <sup>2</sup> Let  $f_a(t) = a(t) \exp[i\phi(t)]$  be the analytic part of  $f(t)$ . Prove that

$$\int_{-\infty}^{+\infty} (\xi - \phi'(t))^2 P_V f_a(t, \xi) d\xi = -\pi a^2(t) \frac{d^2 \log a(t)}{dt^2}.$$

- 4.19** <sup>4</sup> To avoid the time-frequency resolution limitations of a windowed Fourier transform, we want to adapt the window size to the signal content. Let  $g(t)$  be a window supported in  $[-\frac{1}{2}, \frac{1}{2}]$ . We denote by  $S_j f(u, \xi)$  the windowed Fourier transform calculated with the dilated window  $g_j(t) = 2^{-j/2} g(2^{-j} t)$ . Find a procedure that computes a single map of ridges by choosing a “best” window size at each  $(u, \xi)$ . One approach is to choose the scale  $2^l$  for each  $(u, \xi)$  such that  $|S_l f(u, \xi)|^2 = \sup_j |S_j f(u, \xi)|^2$ . Test your algorithm on the linear and hyperbolic chirp signals (4.103, 4.107). Test it on the Tweet and Greasy signals in WAVELAB.

## Frames

## 5

A signal representation may provide “analysis” coefficients that are inner products with a family of vectors, or “synthesis” coefficients that compute an approximation by recombining a family of vectors. Frames are families of vectors where “analysis” and “synthesis” representations are stable. Signal reconstructions are computed with a dual frame. Frames are potentially redundant and thus more general than bases, with a redundancy measured by frame bounds. They provide the flexibility needed to build signal representations with unstructured families of vectors.

Complete and stable wavelet and windowed Fourier transforms are constructed with frames of wavelets and windowed Fourier atoms. In two dimensions, frames of directional wavelets and curvelets are introduced to analyze and process multiscale image structures.

## 5.1 FRAMES AND RIESZ BASES

### 5.1.1 Stable Analysis and Synthesis Operators

The frame theory was originally developed by Duffin and Schaeffer [235] to reconstruct band-limited signals from irregularly spaced samples. They established general conditions to recover a vector  $f$  in a Hilbert space  $\mathbf{H}$  from its inner products with a family of vectors  $\{\phi_n\}_{n \in \Gamma}$ . The index set  $\Gamma$  might be finite or infinite. The following frame definition gives an energy equivalence to invert the operator  $\Phi$  defined by

$$\forall n \in \Gamma, \quad \Phi f[n] = \langle f, \phi_n \rangle. \quad (5.1)$$

**Definition 5.1:** *Frame and Riesz Basis.* The sequence  $\{\phi_n\}_{n \in \Gamma}$  is a frame of  $\mathbf{H}$  if there exist two constants  $B \geq A > 0$  such that

$$\forall f \in \mathbf{H}, \quad A \|f\|^2 \leq \sum_{n \in \Gamma} |\langle f, \phi_n \rangle|^2 \leq B \|f\|^2. \quad (5.2)$$

When  $A = B$  the frame is said to be tight. If the  $\{\phi_n\}_{n \in \Gamma}$  are linearly independent then the frame is not redundant and is called a *Riesz basis*.

If the frame condition is satisfied, then  $\Phi$  is called a *frame analysis operator*. Section 5.1.2 proves that (5.2) is a necessary and sufficient condition guaranteeing that  $\Phi$  is invertible on its image space, with a bounded inverse. Thus, a frame defines a complete and stable signal representation, which may also be redundant.

### Frame Synthesis

Let us consider the space of finite energy coefficients

$$\ell^2(\Gamma) = \{a : \|a\|^2 = \sum_{n \in \Gamma} |a[n]|^2 < +\infty\}.$$

The adjoint  $\Phi^*$  of  $\Phi$  is defined over  $\ell^2(\Gamma)$  and satisfies for any  $f \in \mathbf{H}$  and  $a \in \ell^2(\Gamma)$ :

$$\langle \Phi^* a, f \rangle = \langle a, \Phi f \rangle = \sum_{n \in \Gamma} a[n] \langle f, \phi_n \rangle^*.$$

It is therefore the synthesis operator

$$\Phi^* a = \sum_{n \in \Gamma} a[n] \phi_n. \quad (5.3)$$

The frame condition (5.2) can be rewritten as

$$\forall f \in \mathbf{H}, \quad A \|f\|^2 \leq \|\Phi f\|^2 = \langle \Phi^* \Phi f, f \rangle \leq B \|f\|^2, \quad (5.4)$$

with

$$\Phi^* \Phi f = \sum_{m \in \Gamma} \langle f, \phi_m \rangle \phi_m.$$

It results that  $A$  and  $B$  are the infimum and supremum values of the spectrum of the symmetric operator  $\Phi^* \Phi$ , which correspond to the smallest and largest eigenvalues in finite dimension. The eigenvalues are also called *singular values* of  $\Phi$  or *singular spectrum*. Theorem 5.1 derives that the frame synthesis operator is also stable.

**Theorem 5.1.** The family  $\{\phi_n\}_{n \in \Gamma}$  is a frame with bounds  $0 < A \leq B$  if and only if

$$\forall a \in \text{Im } \Phi, \quad A \|a\|^2 \leq \left\| \sum_{n \in \Gamma} a[n] \phi_n \right\|^2 \leq B \|a\|^2. \quad (5.5)$$

**Proof.** Since  $\Phi^* a = \sum_{n \in \Gamma} a[n] \phi_n$ , it results that

$$\left\| \sum_{n \in \Gamma} a[n] \phi_n \right\|^2 = \langle \Phi^* a, a \rangle.$$

The operator  $\Phi$  is a frame if and only if the spectrum of  $\Phi^* \Phi$  is bound by  $A$  and  $B$ . The inequality (5.5) states that the spectrum of  $\Phi^* \Phi$  over  $\text{Im } \Phi$  is also bounded by  $A$  and  $B$ . Both statements are proved to be equivalent by verifying that the supremum and infimum of the spectrum of  $\Phi^* \Phi$  are equal to the supremum and infimum of the spectrum of  $\Phi^* \Phi$ .

In finite dimension, if  $\lambda$  is an eigenvalue of  $\Phi^* \Phi$  with eigenvector  $f$ , then  $\lambda$  is also an eigenvalue of  $\Phi^* \Phi$  with eigenvector  $\Phi f$ . Indeed,  $\Phi^* \Phi (\Phi f) = \lambda f$  so  $\Phi^* \Phi (\Phi f) = \lambda \Phi f$

and  $\Phi f \neq 0$ , because the left frame inequality (5.2) implies that  $\|\Phi f\|^2 \leq A \|f\|^2$ . It results that the maximum and minimum eigenvectors of  $\Phi^* \Phi$  and  $\Phi \Phi^*$  on  $\text{Im} \Phi$  are identical.

In a Hilbert space of infinite dimension, we prove that the supremum and infimum of the spectrum of both operators remain identical by growing the space dimension, and computing the limit of the largest and smallest eigenvalues when the space dimension tends to infinity. ■

This theorem proves that linear combination of frame vectors define a stable signal representation. Section 5.1.2 proves that synthesis coefficients are computed with a dual frame. The operator  $\Phi \Phi^*$  is the *Gram* matrix  $\{\langle \phi_n, \phi_p \rangle\}_{(m,p) \in \ell^2(\Gamma)}$ :

$$\Phi \Phi^* a[p] = \sum_{m \in \Gamma} a[m] \langle \phi_n, \phi_p \rangle. \quad (5.6)$$

One must be careful because (5.5) is only valid for  $a \in \text{Im} \Phi$ . If it is valid for all  $a \in \ell^2(\Gamma)$  with  $A > 0$  then the family is linearly independant and is thus a Riesz basis.

### Redundancy

When the frame vectors are normalized  $\|\phi_n\| = 1$ , Theorem 5.2 shows that the frame redundancy is measured by the frame bounds  $A$  and  $B$ .

**Theorem 5.2.** In a space of finite dimension  $N$ , a frame of  $P \geq N$  normalized vectors has frame bounds  $A$  and  $B$ , which satisfy

$$A \leq \frac{P}{N} \leq B. \quad (5.7)$$

For a tight frame  $A = B = P/N$ .

**Proof.** It results from (5.4) that all eigenvalues of  $\Phi^* \Phi$  are between  $A$  and  $B$ . The trace of  $\Phi^* \Phi$  thus satisfies

$$AN \leq \text{tr}(\Phi^* \Phi) \leq BN.$$

But since the trace is not modified by commuting matrices (Exercise 5.4), and  $\|\phi_n\| = 1$ ,

$$AN \leq \text{tr}(\Phi^* \Phi) = \text{tr}(\Phi \Phi^*) = \sum_{n=1}^P |\langle \phi_n, \phi_n \rangle|^2 = P \leq BN,$$

which implies (5.7). ■

If  $\{\phi_n\}_{n \in \Gamma}$  is a normalized Riesz basis and is therefore linearly independent, then (5.7) proves that  $A \leq 1 \leq B$ . This result remains valid in infinite dimension. Inserting  $f = \phi_n$  in the frame inequality (5.2) proves that the frame is orthonormal if and only if  $B = 1$ , in which case  $A = 1$ .

---

**EXAMPLE 5.1**

Let  $\{g_1, g_2\}$  be an orthonormal basis of an  $N=2$  two-dimensional plane  $\mathbf{H}$ . The  $P=3$  normalized vectors

$$\phi_1 = g_1, \quad \phi_2 = -\frac{g_1}{2} + \frac{\sqrt{3}}{2} g_2, \quad \phi_3 = -\frac{g_1}{2} - \frac{\sqrt{3}}{2} g_2$$

have equal angles of  $2\pi/3$  between each other. For any  $f \in \mathbf{H}$ ,

$$\sum_{n=1}^3 |\langle f, \phi_n \rangle|^2 = \frac{3}{2} \|f\|^2.$$

Thus, these three vectors define a tight frame with  $A=B=3/2$ .

---

**EXAMPLE 5.2**

For any  $0 \leq k < K$ , suppose that  $\{\phi_{k,n}\}_{n \in \Gamma}$  is an orthonormal basis of  $\mathbf{H}$ . The union of these  $K$  orthonormal bases  $\{\phi_{k,n}\}_{n \in \Gamma, 0 \leq k < K}$  is a tight frame with  $A=B=K$ . Indeed, the energy conservation in an orthonormal basis implies that for any  $f \in \mathbf{H}$ ,

$$\sum_{n \in \mathbb{Z}} |\langle f, \phi_{k,n} \rangle|^2 = \|f\|^2,$$

therefore,

$$\sum_{k=0}^{K-1} \sum_{n \in \mathbb{Z}} |\langle f, \phi_{k,n} \rangle|^2 = K \|f\|^2.$$


---

One can verify (Exercise 5.3) that a finite set of  $N$  vectors  $\{\phi_n\}_{1 \leq n \leq N}$  is always a frame of the space  $\mathbf{V}$  generated by linear combinations of these vectors. When  $N$  increases, the frame bounds  $A$  and  $B$  may go respectively to 0 and  $+\infty$ . This illustrates the fact that in infinite dimensional spaces, a family of vectors may be complete and not yield a stable signal representation.

***Irregular Sampling***

Let  $\mathbf{U}_s$  be the space of  $\mathbf{L}^2(\mathbb{R})$  functions having a Fourier transform support included in  $[-\pi/s, \pi/s]$ . For a uniform sampling,  $t_n = ns$ , Theorem 3.5 proves that if  $\phi_s(t) = s^{1/2} \sin(\pi s^{-1} t)/(\pi t)$ , then  $\{\phi_s(t - ns)\}_{n \in \mathbb{Z}}$  is an orthonormal basis of  $\mathbf{U}_s$ . The reconstruction of  $f$  from its samples is then given by the sampling Theorem 3.2.

The irregular sampling conditions of Duffin and Schaeffer [235] for constructing a frame were later refined by several researchers [81, 102, 500]. Gröchenig proved [285] that if  $\lim_{n \rightarrow +\infty} t_n = +\infty$  and  $\lim_{n \rightarrow -\infty} t_n = -\infty$ , and if the maximum sampling distance  $\delta$  satisfies

$$\delta = \sup_{n \in \mathbb{Z}} |t_{n+1} - t_n| < s, \tag{5.8}$$

then

$$\{\lambda_n \phi_s(t - t_n)\}_{n \in \mathbb{Z}} \quad \text{with} \quad \lambda_n = \sqrt{\frac{t_{n+1} - t_{n-1}}{2s}}$$

is a frame with frame bounds  $A \geq (1 - \delta/s)^2$  and  $B \leq (1 + \delta/s)^2$ . The amplitude factor  $\lambda_n$  compensates for the increase of sample density relatively to  $s$ . The reconstruction of  $f$  requires inverting the frame operator  $\Phi f[n] = \langle f(u), \lambda_n, \phi_s(u - t_n) \rangle$ .

### 5.1.2 Dual Frame and Pseudo Inverse

The reconstruction of  $f$  from its frame coefficients  $\Phi f[n]$  is calculated with a pseudo inverse also called Moore-Penrose pseudo inverse. This pseudo inverse is a bounded operator that implements a dual-frame reconstruction. For Riesz bases, this dual frame is a biorthogonal basis.

For any operator  $U$ , we denote by  $\mathbf{Im}U$  the image space of all  $Uf$  and by  $\mathbf{Null}U$  the null space of all  $h$ , such that  $Uh = 0$ .

**Theorem 5.3.** If  $\{\phi_n\}_{n \in \Gamma}$  is a frame but not a Riesz basis, then  $\Phi$  admits an infinite number of left inverses.

**Proof.** We know that  $\mathbf{Null}\Phi^* = (\mathbf{Im}\Phi)^\perp$  is the orthogonal complement of  $\mathbf{Im}\Phi$  in  $\ell^2(\Gamma)$  (Exercise 5.7). If  $\Phi$  is a frame and not a Riesz basis, then  $\{\phi_n\}_{n \in \Gamma}$  is linearly dependent, so there exists  $a \in \mathbf{Null}\Phi^* = (\mathbf{Im}\Phi)^\perp$  with  $a \neq 0$ .

A frame operator  $\Phi$  is injective (one to one). Indeed, the frame inequality (5.2) guarantees that  $\Phi f = 0$  implies  $f = 0$ . Its restriction to  $\mathbf{Im}\Phi$  is thus invertible, which means that  $\Phi$  admits a left inverse. There is an infinite number of left inverses since the restriction of a left inverse to  $(\mathbf{Im}\Phi)^\perp \neq \{0\}$  may be any arbitrary linear operator. ■

The more redundant the frame  $\{\phi_n\}_{n \in \Gamma}$ , the larger the orthogonal complement  $(\mathbf{Im}\Phi)^\perp$  of  $\mathbf{Im}\Phi$  in  $\ell^2(\Gamma)$ . The pseudo inverse, written as  $\Phi^+$ , is defined as the left inverse that is zero on  $(\mathbf{Im}\Phi)^\perp$ :

$$\forall f \in \mathbf{H}, \quad \Phi^+ \Phi f = f \quad \text{and} \quad \forall a \in (\mathbf{Im}\Phi)^\perp, \quad \Phi^+ a = 0. \quad (5.9)$$

Theorem 5.4 computes this pseudo inverse.

**Theorem 5.4: Pseudo Inverse.** If  $\Phi$  is a frame operator, then  $\Phi^* \Phi$  is invertible and the pseudo inverse satisfies

$$\Phi^+ = (\Phi^* \Phi)^{-1} \Phi^*. \quad (5.10)$$

**Proof.** The frame condition in (5.4) is rewritten as

$$\forall f \in \mathbf{H}, \quad A \|f\|^2 \leq \langle \Phi^* \Phi f, f \rangle \leq B \|f\|^2.$$

The result is that  $\Phi^* \Phi$  is an injective self-adjoint operator:  $\Phi^* \Phi f = 0$  if and only if  $f = 0$ . It is therefore invertible. For all  $f \in \mathbf{H}$ ,

$$\Phi^+ \Phi f = (\Phi^* \Phi)^{-1} \Phi^* \Phi f = f,$$

so  $\Phi^+$  is a left inverse. Since  $(\mathbf{Im}\Phi)^\perp = \mathbf{Null}\Phi^*$ , it results that  $\Phi^+ a = 0$  for any  $a \in (\mathbf{Im}\Phi)^\perp = \mathbf{Null}\Phi^*$ . Since this left inverse vanishes on  $(\mathbf{Im}\Phi)^\perp$ , it is the pseudo inverse. ■

**Dual Frame**

The pseudo inverse of a frame operator implements a reconstruction with a dual frame, which is specified by Theorem 5.5.

**Theorem 5.5.** Let  $\{\phi_n\}_{n \in \Gamma}$  be a frame with bounds  $0 < A \leq B$ . The dual operator defined by

$$\forall n \in \Gamma, \quad \tilde{\Phi}f[n] = \langle f, \tilde{\phi}_n \rangle \quad \text{with} \quad \tilde{\phi}_n = (\Phi^* \Phi)^{-1} \phi_n \quad (5.11)$$

satisfies  $\tilde{\Phi}^* = \Phi^+$ , and thus

$$f = \sum_{n \in \Gamma} \langle f, \phi_n \rangle \tilde{\phi}_n = \sum_{n \in \Gamma} \langle f, \tilde{\phi}_n \rangle \phi_n. \quad (5.12)$$

It defines a dual frame as

$$\forall f \in \mathbf{H}, \quad \frac{1}{B} \|f\|^2 \leq \sum_{n \in \Gamma} |\langle f, \tilde{\phi}_n \rangle|^2 \leq \frac{1}{A} \|f\|^2. \quad (5.13)$$

If the frame is tight (i.e.,  $A = B$ ), then  $\tilde{\phi}_n = A^{-1} \phi_n$ .

**Proof.** The dual operator can be written as  $\tilde{\Phi} = \Phi(\Phi^* \Phi)^{-1}$ . Indeed,

$$\tilde{\Phi}f[n] = \langle f, \tilde{\phi}_n \rangle = \langle f, (\Phi^* \Phi)^{-1} \phi_n \rangle = \langle (\Phi^* \Phi)^{-1} f, \phi_n \rangle = \Phi(\Phi^* \Phi)^{-1} f.$$

Thus, we derive from (5.10) that its adjoint is the pseudo inverse of  $\Phi$ :

$$\tilde{\Phi}^* = (\Phi^* \Phi)^{-1} \Phi^* = \Phi^+.$$

It results that  $\Phi^+ \Phi = \tilde{\Phi}^* \Phi = \text{Id}$  and thus that  $\Phi^* \tilde{\Phi} = \text{Id}$ , which proves (5.12).

Let us now prove the frame bounds (5.13). Frame conditions are rewritten in (5.4):

$$\forall f \in \mathbf{H}, \quad A \|f\|^2 \leq \langle \Phi^* \Phi f, f \rangle \leq B \|f\|^2. \quad (5.14)$$

Lemma 5.1 applied to  $L = \Phi^* \Phi$  proves that

$$\forall f \in \mathbf{H}, \quad B^{-1} \|f\|^2 \leq \langle (\Phi^* \Phi)^{-1} f, f \rangle \leq A^{-1} \|f\|^2. \quad (5.15)$$

Since for any  $f \in \mathbf{H}$

$$\|\tilde{\Phi}f\|^2 = \langle \Phi(\Phi^* \Phi)^{-1} f, \Phi(\Phi^* \Phi)^{-1} f \rangle = \langle f, (\Phi^* \Phi)^{-1} f \rangle,$$

the dual-frame bounds (5.13) are derived from (5.15).

If  $A = B$ , then  $\langle \Phi^* \Phi f, f \rangle = A \|f\|^2$ . Thus, the spectrum of  $\Phi^* \Phi$  is reduced to  $A$  and therefore  $\Phi^* \Phi = A \text{Id}$ . As a result,  $\tilde{\phi}_n = (\Phi^* \Phi)^{-1} \phi_n = A^{-1} \phi_n$ .

**Lemma 5.1.** If  $L$  is a self-adjoint operator such that there exist  $B \geq A > 0$  satisfying

$$\forall f \in \mathbf{H}, \quad A \|f\|^2 \leq \langle Lf, f \rangle \leq B \|f\|^2, \quad (5.16)$$

then  $L$  is invertible and

$$\forall f \in \mathbf{H}, \quad \frac{1}{B} \|f\|^2 \leq \langle L^{-1} f, f \rangle \leq \frac{1}{A} \|f\|^2. \quad (5.17)$$

In finite dimensions, since  $L$  is self-adjoint we know that it is diagonalized in an orthonormal basis. The inequality (5.16) proves that its eigenvalues are between  $A$  and  $B$ . It is therefore invertible with eigenvalues between  $B^{-1}$  and  $A^{-1}$ , which proves (5.17). In a Hilbert space of infinite dimension, we prove that same result on the supremum and infimum of the spectrum by growing the space dimension, and computing the limit of the largest and smallest eigenvalues when the space dimension tends to infinity. ■

This theorem proves that  $f$  is reconstructed from frame coefficients  $\Phi f[n] = \langle f, \phi_n \rangle$  with the dual frame  $\{\tilde{\phi}_n\}_{n \in \Gamma}$ . The synthesis coefficients of  $f$  in  $\{\phi_n\}_{n \in \Gamma}$  are the dual-frame coefficients  $\Phi f[n] = \langle f, \phi_n \rangle$ . If the frame is tight, then both decompositions are identical:

$$f = \frac{1}{A} \sum_{n \in \Gamma} \langle f, \phi_n \rangle \phi_n. \quad (5.18)$$

### **Biorthogonal Bases**

A Riesz basis is a frame of vectors that are linearly independent, which implies that  $\text{Im } \Phi = \ell^2(\Gamma)$ , so its dual frame is also linearly independent. Inserting  $f = \phi_p$  in (5.12) yields

$$\phi_p = \sum_{n \in \Gamma} \langle \phi_p, \tilde{\phi}_n \rangle \phi_n,$$

and the linear independence implies that

$$\langle \phi_p, \tilde{\phi}_n \rangle = \delta[p - n].$$

Thus, dual Riesz bases are biorthogonal families of vectors. If the basis is normalized (i.e.,  $\|\phi_n\| = 1$ ), then

$$A \leq 1 \leq B. \quad (5.19)$$

This is proved by inserting  $f = \phi_p$  in the frame inequality (5.13):

$$\frac{1}{B} \|\phi_p\|^2 \leq \sum_{n \in \Gamma} |\langle \phi_p, \tilde{\phi}_n \rangle|^2 = 1 \leq \frac{1}{A} \|\phi_p\|^2.$$

### **5.1.3 Dual-Frame Analysis and Synthesis Computations**

Suppose that  $\{\phi_n\}_{n \in \Gamma}$  is a frame of a subspace  $\mathbf{V}$  of the whole signal space. The best linear approximation of  $f$  in  $\mathbf{V}$  is the orthogonal projection of  $f$  in  $\mathbf{V}$ . Theorem 5.6 shows that this orthogonal projection is computed with the dual frame. Two iterative numerical algorithms are described to implement such computations.

**Theorem 5.6.** Let  $\{\phi_n\}_{n \in \Gamma}$  be a frame of  $\mathbf{V}$ , and  $\{\tilde{\phi}_n\}_{n \in \Gamma}$  its dual frame in  $\mathbf{V}$ . The orthogonal projection of  $f \in \mathbf{H}$  in  $\mathbf{V}$  is

$$P_{\mathbf{V}} f = \sum_{n \in \Gamma} \langle f, \phi_n \rangle \tilde{\phi}_n = \sum_{n \in \Gamma} \langle f, \tilde{\phi}_n \rangle \phi_n. \quad (5.20)$$

**Proof.** Since both frames are dual in  $\mathbf{V}$ , if  $f \in \mathbf{V}$ , then, (5.12) proves that the operator  $P_{\mathbf{V}}$  defined in (5.20) satisfies  $P_{\mathbf{V}}f = f$ . To prove that it is an orthogonal projection it is sufficient to verify that if  $f \in \mathbf{H}$  then  $\langle f - P_{\mathbf{V}}f, \phi_p \rangle = 0$  for all  $p \in \Gamma$ . Indeed,

$$\langle f - P_{\mathbf{V}}f, \phi_p \rangle = \langle f, \phi_p \rangle - \sum_{n \in \Gamma} \langle f, \phi_n \rangle \langle \tilde{\phi}_n, \phi_p \rangle = 0$$

because the dual-frame property implies that  $\sum_{n \in \Gamma} \langle \tilde{\phi}_n, \phi_p \rangle \phi_n = \phi_p$ . ■

If  $\Gamma$  is finite, then  $\{\phi_n\}_{n \in \Gamma}$  is necessarily a frame of the space  $\mathbf{V}$  it generates, and (5.20) reconstructs the best linear approximation of  $f$  in  $\mathbf{V}$ . This result is particularly important for approximating signals from a finite set of vectors.

Since  $\Phi$  is not a frame of the whole signal space  $\mathbf{H}$ , but of a subspace  $\mathbf{V}$  then  $\Phi$  is only invertible on this subspace and the pseudo-inverse definition becomes:

$$\forall f \in \mathbf{V}, \quad \Phi^+ \Phi f = f \quad \text{and} \quad \forall a \in (\text{Im } \Phi)^\perp, \quad \Phi^+ a = 0. \quad (5.21)$$

Let  $\Phi_{\mathbf{V}}$  be the restriction of  $\Phi$  to  $\mathbf{V}$ . The operator  $\Phi^* \Phi_{\mathbf{V}}$  is invertible on  $\mathbf{V}$  and we write  $(\Phi^* \Phi_{\mathbf{V}})^{-1}$  its inverse. Similar to (5.10), we verify that  $\Phi^+ = (\Phi^* \Phi_{\mathbf{V}})^{-1} \Phi^*$ .

### Dual Synthesis

In a dual synthesis problem, the orthogonal projection is computed from the frame coefficients  $\{\Phi f[n] = \langle f, \phi_n \rangle\}_{n \in \Gamma}$  with the dual-frame synthesis operator:

$$P_{\mathbf{V}}f = \tilde{\Phi}^* \Phi f = \sum_{n \in \Gamma} \langle f, \phi_n \rangle \tilde{\phi}_n. \quad (5.22)$$

If the frame  $\{\phi_n\}_{n \in \Gamma}$  does not depend on the signal  $f$ , then the dual-frame vectors are precomputed with (5.11):

$$\forall n \in \Gamma, \quad \tilde{\phi}_n = (\Phi^* \Phi_{\mathbf{V}})^{-1} \phi_n, \quad (5.23)$$

and the dual synthesis is solved directly with (5.22). In many applications, the frame vectors  $\{\phi_n\}_{n \in \Gamma}$  depend on the signal  $f$ , in which case the dual-frame vectors  $\tilde{\phi}_n$  cannot be computed in advance, and it is highly inefficient to compute them. This is the case when coefficients  $\{\langle f, \phi_n \rangle\}_{n \in \Gamma}$  are selected in a redundant transform, to build a sparse signal representation. For example, the time-frequency ridge vectors in Sections 4.4.2 and 4.4.3 are selected from the local maxima of  $f$  in highly redundant windowed Fourier or wavelet transforms.

The transform coefficients  $\Phi f$  are known and we must compute

$$P_{\mathbf{V}}f = \tilde{\Phi}^* \Phi f = (\Phi^* \Phi_{\mathbf{V}})^{-1} \Phi^* \Phi f.$$

A dual-synthesis algorithm computes first

$$y = \Phi^* \Phi f = \sum_{n \in \Gamma} \langle f, \phi_n \rangle \phi_n \in \mathbf{V}$$

and then derives  $P_V f = L^{-1}y = z$  by applying the inverse of the symmetric operator  $L = \Phi^* \Phi_V$  to  $y$ , with

$$\forall h \in V, \quad Lh = \sum_{n \in \Gamma} \langle h, \phi_n \rangle \phi_n. \quad (5.24)$$

The eigenvalues of  $L$  are between  $A$  and  $B$ .

### Dual Analysis

In a dual analysis, the orthogonal projection  $P_V f$  is computed from the frame vectors  $\{\phi_n\}_{n \in \Gamma}$  with the dual-frame analysis operator  $\tilde{\Phi}f[n] = \langle f, \tilde{\phi}_n \rangle$ :

$$P_V f = \Phi^* \tilde{\Phi}f = \sum_{n \in \Gamma} \langle f, \tilde{\phi}_n \rangle \phi_n. \quad (5.25)$$

If  $\{\phi_n\}_{n \in \Gamma}$  does not depend upon  $f$  then  $\{\tilde{\phi}_n\}_{n \in \Gamma}$  is precomputed with (5.23). The  $\{\phi_n\}_{n \in \Gamma}$  may also be selected adaptively from a larger dictionary, to provide a sparse approximation of  $f$ . Computing the orthogonal projection  $P_V f$  is called a *backprojection*. In Section 12.3, matching pursuits implement this back-projection.

When  $\{\phi_n\}_{n \in \Gamma}$  depends on  $f$ , computing the dual frame is inefficient. The dual coefficient  $a[n] = \tilde{\Phi}f[n]$  is calculated directly, as well as

$$P_V f = \Phi^* a = \sum_{n \in \Gamma} a[n] \phi_n. \quad (5.26)$$

Since  $\Phi P_V f = \Phi f$ , we have  $\Phi \Phi^* a = \Phi f$ . Let  $\Phi_{\text{Im}\Phi}^*$  be the restriction of  $\Phi^*$  to  $\text{Im}\Phi$ . Since  $\Phi \Phi_{\text{Im}\Phi}^*$  is invertible on  $\text{Im}\Phi$

$$a = (\Phi \Phi_{\text{Im}\Phi}^*)^{-1} \Phi f.$$

Thus, the dual-analysis algorithm computes  $y = \Phi f = \{\langle f, \phi_n \rangle\}_{n \in \Gamma}$  and derives the dual coefficients  $a = L^{-1}y = z$  by applying the inverse of the Gram operator  $L = \Phi \Phi_{\text{Im}\Phi}^*$  to  $y$ , with

$$Lh[n] = \sum_{p \in \Gamma} h[p] \langle \phi_n, \phi_p \rangle. \quad (5.27)$$

The eigenvalues of  $L$  are also between  $A$  and  $B$ . The orthogonal projection of  $f$  is recovered with (5.26).

### Richardson Inversion of Symmetric Operators

The key computational step of a dual-analysis or a dual-synthesis problem is to compute  $z = L^{-1}y$ , where  $L$  is a symmetric operator with eigenvalues that are between  $A$  and  $B$ . Theorems 5.7 and 5.8 describe two iterative algorithms with exponential convergence. The *Richardson iteration procedure* is simpler but requires knowing the frame bounds  $A$  and  $B$ . *Conjugate gradient* iterations converge more quickly when  $B/A$  is large, and do not require knowing the values of  $A$  and  $B$ .

**Theorem 5.7.** To compute  $z = L^{-1}y$ , let  $z_0$  be an initial value and  $\gamma > 0$  be a relaxation parameter. For any  $k > 0$ , define

$$z_k = z_{k-1} + \gamma (y - Lz_{k-1}). \quad (5.28)$$

If

$$\delta = \max \{ |1 - \gamma A|, |1 - \gamma B| \} < 1, \quad (5.29)$$

then

$$\|z - z_k\| \leq \delta^k \|z - z_0\|, \quad (5.30)$$

and therefore  $\lim_{k \rightarrow +\infty} z_k = z$ .

**Proof.** The induction equation (5.28) can be rewritten as

$$z - z_k = z - z_{k-1} - \gamma L(z - z_{k-1}).$$

Let

$$R = Id - \gamma L,$$

$$z - z_k = R(z - z_{k-1}) = R^k(z - z_0). \quad (5.31)$$

Since the eigenvalues of  $L$  are between  $A$  and  $B$ ,

$$A \|z\|^2 \leq \langle Lz, z \rangle \leq B \|z\|^2.$$

This implies that  $R = I - \gamma L$  satisfies

$$|\langle Rz, z \rangle| \leq \delta \|z\|^2,$$

where  $\delta$  is given by (5.29). Since  $R$  is symmetric, this inequality proves that  $\|R\| \leq \delta$ . Thus, we derive (5.30) from (5.31). The error  $\|z - z_k\|$  clearly converges to zero if  $\delta < 1$ . ■

The convergence is guaranteed for all initial values  $z_0$ . If an estimation  $z_0$  of the solution  $z$  is known, then this estimation can be chosen; otherwise,  $z_0$  is often set to 0. For frame inversion, the Richardson iteration algorithm is sometimes called the *frame algorithm* [19]. The convergence rate is maximized when  $\delta$  is minimum:

$$\delta = \frac{B - A}{B + A} = \frac{1 - A/B}{1 + A/B},$$

which corresponds to the relaxation parameter

$$\gamma = \frac{2}{A + B}. \quad (5.32)$$

The algorithm converges quickly if  $A/B$  is close to 1. If  $A/B$  is small then

$$\delta \approx 1 - 2 \frac{A}{B}. \quad (5.33)$$

The inequality (5.30) proves that we obtain an error smaller than  $\varepsilon$  for a number  $n$  of iterations, which satisfies

$$\frac{\|z - z_k\|}{\|z - z_0\|} \leq \delta^k = \varepsilon.$$

Inserting (5.33) gives

$$k \approx \frac{\log_e \varepsilon}{\log_e(1 - 2A/B)} \approx \frac{-B}{2A} \log_e \varepsilon. \quad (5.34)$$

Therefore, the number of iterations increases proportionally to the frame-bound ratio  $B/A$ .

The exact values of  $A$  and  $B$  are often not known, and  $A$  is generally more difficult to compute. The upper frame bound is  $B = \|\Phi \Phi^*\|_S = \|\Phi^* \Phi\|_S$ . If we choose

$$\gamma < 2 \|\Phi \Phi^*\|_S^{-1}, \quad (5.35)$$

then (5.29) shows that the algorithm is guaranteed to converge, but the convergence rate depends on  $A$ . Since  $0 < A \leq B$ , the optimal relaxation parameter  $\gamma$  in (5.32) is in the range  $\|\Phi \Phi^*\|_S^{-1} \leq \gamma < 2 \|\Phi \Phi^*\|_S^{-1}$ .

### **Conjugate-Gradient Inversion**

The conjugate-gradient algorithm computes  $z = L^{-1}y$  with a gradient descent along orthogonal directions with respect to the norm induced by the symmetric operator  $L$ :

$$\|z\|_L^2 = \|Lz\|^2. \quad (5.36)$$

This  $L$  norm is used to estimate the error. Gröchenig's [287] implementation of the conjugate-gradient algorithm is given by Theorem 5.8.

**Theorem 5.8:** *Conjugate Gradient.* To compute  $z = L^{-1}y$ , we initialize

$$z_0 = 0, \quad r_0 = p_0 = y, \quad p_{-1} = 0. \quad (5.37)$$

For any  $k \geq 0$ , we define by induction:

$$\lambda_k = \frac{\langle r_k, p_k \rangle}{\langle p_k, Lp_k \rangle} \quad (5.38)$$

$$z_{k+1} = z_k + \lambda_k p_k \quad (5.39)$$

$$r_{k+1} = r_k - \lambda_k Lp_k \quad (5.40)$$

$$p_{k+1} = Lp_k - \frac{\langle Lp_k, Lp_k \rangle}{\langle p_k, Lp_k \rangle} p_k - \frac{\langle Lp_k, Lp_{k-1} \rangle}{\langle p_{k-1}, Lp_{k-1} \rangle} p_{k-1}. \quad (5.41)$$

If  $\sigma = \frac{\sqrt{B} - \sqrt{A}}{\sqrt{B} + \sqrt{A}}$ , then

$$\|z - z_k\|_L \leq \frac{2\sigma^k}{1 + \sigma^{2k}} \|z\|_L, \quad (5.42)$$

and therefore  $\lim_{k \rightarrow +\infty} z_k = z$ .

**Proof.** We give the main steps of the proof as outlined by Gröchenig [287].

*Step 1.* Let  $\mathbf{U}_k$  be the subspace generated by  $\{L^j z\}_{1 \leq j \leq k}$ . By induction on  $k$ , we derive from (5.41) that  $p_j \in \mathbf{U}_k$ , for  $j < k$ .

*Step 2.* We prove by induction that  $\{p_j\}_{0 \leq j < k}$  is an orthogonal basis of  $\mathbf{U}_k$  with respect to the inner product  $\langle z, h \rangle_L = \langle z, Lh \rangle$ . Assuming that  $\langle p_k, Lp_j \rangle = 0$ , for  $j \leq k-1$ , it can be shown that  $\langle p_{k+1}, Lp_j \rangle = 0$ , for  $j \leq k$ .

*Step 3.* We verify that  $z_k$  is the orthogonal projection of  $z$  onto  $\mathbf{U}_k$  with respect to  $\langle \cdot, \cdot \rangle_L$ , which means that

$$\forall g \in \mathbf{U}_k, \quad \|z - g\|_L \geq \|z - z_k\|_L.$$

Since  $z_k \in \mathbf{U}_k$ , this requires proving that  $\langle z - z_k, p_j \rangle_L = 0$ , for  $j < k$ .

*Step 4.* We compute the orthogonal projection of  $z$  in embedded spaces  $\mathbf{U}_k$  of dimension  $k$ , and one can verify that  $\lim_{k \rightarrow +\infty} \|z - z_k\|_L = 0$ . The exponential convergence (5.42) is proved in [287]. ■

As opposed to the Richardson algorithm, the initial value  $z_0$  must be set to 0. As in the Richardson iteration algorithm, the convergence is slower when  $A/B$  is small. In this case,

$$\sigma = \frac{1 - \sqrt{A/B}}{1 + \sqrt{A/B}} \approx 1 - 2 \sqrt{\frac{A}{B}}.$$

The upper bound (5.42) proves that we obtain a relative error

$$\frac{\|z - z_k\|_L}{\|z\|_L} \leq \varepsilon$$

for a number of iterations

$$k \approx \frac{\log_e \frac{\varepsilon}{2}}{\log_e \sigma} \approx \frac{-\sqrt{B}}{2\sqrt{A}} \log_e \frac{\varepsilon}{2}.$$

Comparing this result with (5.34) shows that when  $A/B$  is small, the conjugate-gradient algorithm needs much less iterations than the Richardson iteration algorithm to compute  $z = L^{-1}y$  at a fixed precision.

### 5.1.4 Frame Projector and Reproducing Kernel

Frame redundancy is useful in reducing noise added to the frame coefficients. The vector computed with noisy frame coefficients is projected on the image of  $\Phi$  to reduce the amplitude of the noise. This technique is used for high-precision analog-to-digital conversion based on oversampling. The following theorem specifies the orthogonal projector on  $\text{Im}\Phi$ .

**Theorem 5.9:** *Reproducing Kernel.* Let  $\{\phi_n\}_{n \in \Gamma}$  be a frame of  $\mathbf{H}$  or of a subspace  $\mathbf{V}$ . The orthogonal projection from  $\ell^2(\Gamma)$  onto  $\text{Im}\Phi$  is

$$Pa[n] = \Phi \Phi^+ a[n] = \sum_{p \in \Gamma} a[p] \langle \tilde{\phi}_p, \phi_n \rangle. \quad (5.43)$$

Coefficients  $a \in \ell^2(\Gamma)$  are frame coefficients if and only if they satisfy the reproducing kernel equation

$$a[n] = \Phi \Phi^+ a[n] = \sum_{p \in \Gamma} a[p] \langle \tilde{\phi}_p, \phi_n \rangle. \quad (5.44)$$

**Proof.** If  $a \in \text{Im}\Phi$ , then  $a = \Phi f$  and

$$Pa = \Phi \Phi^+ \Phi f = \Phi f = a.$$

If  $a \in (\text{Im}\Phi)^\perp$ , then  $Pa = 0$  because  $\Phi^+ a = 0$ . This proves that  $P$  is an orthogonal projector on  $\text{Im}\Phi$ . Since  $\Phi f[n] = \langle f, \phi_n \rangle$  and  $\Phi^+ a = \sum_{p \in \Gamma} a[p] \tilde{\phi}_p$ , we derive (5.43).

A vector  $a \in \ell^2(\Gamma)$  belongs to  $\text{Im}\Phi$  if and only if  $a = Pa$ , which proves (5.44). ■

The reproducing kernel equation (5.44) expresses the redundancy of frame coefficients. If the frame is not redundant and is a Riesz basis, then  $\langle \tilde{\phi}_p, \phi_n \rangle = 0$ , so this equation vanishes.

### Noise Reduction

Suppose that each frame coefficient  $\Phi f[n]$  is contaminated by an additive noise  $W[n]$ , which is a random variable. Applying the projector  $P$  gives

$$P(\Phi f + W) = \Phi f + PW,$$

with

$$PW[n] = \sum_{p \in \Gamma} W[p] \langle \tilde{\phi}_p, \phi_n \rangle.$$

Since  $P$  is an orthogonal projector,  $\|PW\| \leq \|W\|$ . This projector removes the component of  $W$  that is in  $(\text{Im}\Phi)^\perp$ . Increasing the redundancy of the frame reduces the size of  $\text{Im}\Phi$  and thus increases  $(\text{Im}\Phi)^\perp$ , so a larger portion of the noise is removed. If  $W$  is a white noise, its energy is uniformly distributed in the space  $\ell^2(\Gamma)$ . Theorem 5.10 proves that its energy is reduced by at least  $A$  if the frame vectors are normalized.

**Theorem 5.10.** Suppose that  $\|\phi_n\| = C$ , for all  $n \in \Gamma$ . If  $W$  is a zero-mean white noise of variance  $E\{|W[n]|^2\} = \sigma^2$ , then

$$E\{|PW[n]|^2\} \leq \frac{\sigma^2 C^2}{A}. \quad (5.45)$$

If the frame is tight then this inequality is an equality.

**Proof.** Let us compute

$$E\{|PW[n]|^2\} = E \left\{ \left( \sum_{p \in \Gamma} W[p] \langle \tilde{\phi}_p, \phi_n \rangle \right) \left( \sum_{l \in \Gamma} W[l]^* \langle \tilde{\phi}_l, \phi_n \rangle^* \right) \right\}.$$

Since  $W$  is white,

$$E\{W[p] W^*[l]\} = \sigma^2 \delta[p - l],$$

and therefore,

$$E\{|PW[n]|^2\} = \sigma^2 \sum_{p \in \Gamma} |\langle \tilde{\phi}_p, \phi_n \rangle|^2 \leq \frac{\sigma^2 \|\phi_n\|^2}{A} = \frac{\sigma^2 C^2}{A}.$$

The last inequality is an equality if the frame is tight. ■

### Oversampling

This noise-reduction strategy is used by high-precision analog-to-digital converters. After a low-pass filter, a band-limited analog signal  $f(t)$  is uniformly sampled and quantized. In hardware, it is often easier to increase the sampling rate rather than the quantization precision. Increasing the sampling rate introduces a redundancy between the sample values of the band-limited signal. Thus, these samples can be interpreted as frame coefficients. For a wide range of signals it has been shown that the quantization error is nearly a white noise [277]. Thus, it can be significantly reduced by a frame projector, which in this case is a low-pass convolution operator (Exercise 5.16).

The noise can be further reduced if it is not white and if its energy is better concentrated in  $(\text{Im}\Phi)^\perp$ . This can be done by transforming the quantization noise into a noise that has energy mostly concentrated at high frequencies. Sigma-delta modulators produce such quantization noises by integrating the signal before its quantization [89]. To compensate for the integration, the quantized signal is differentiated. This differentiation increases the energy of the quantized noise at high frequencies and reduces its energy at low frequencies [456].

#### 5.1.5 Translation-Invariant Frames

Section 4.1 introduces translation-invariant dictionaries obtained by translating a family of generators  $\{\phi_n\}_{n \in \Gamma}$ , which are used to construct translation-invariant signal representations. In multiple dimensions for  $\phi_n \in \mathbf{L}^2(\mathbb{R}^d)$ , the resulting dictionary can be written  $\mathcal{D} = \{\phi_{u,n}(x)\}_{n \in \Gamma, u \in \mathbb{R}^d}$ , with  $\phi_{u,n}(x) = \lambda_{u,n} \phi_n(x - u)$ . In a translation-invariant wavelet dictionary, the generators are obtained by dilating a wavelet  $\psi(t)$  with scales  $s_n$ :  $\phi_n(t) = s_n^{-1/2} \psi(x/s_n)$ . In a window Fourier dictionary, the generators are obtained by modulating a window  $g(x)$  at frequencies  $\xi_n$ :  $\phi_n(x) = e^{i\xi_n x} g(x)$ .

The decomposition coefficients of  $f$  in  $\mathcal{D}$  are convolution products

$$\Phi f(u, n) = \langle f, \phi_{u,n} \rangle = \lambda_{u,n} f \star \bar{\phi}_n(u) \quad \text{with} \quad \bar{\phi}_n(x) = \phi_n^*(-x). \quad (5.46)$$

Suppose that  $\Gamma$  is a countable set. The overall index set  $\mathbb{R}^d \times \Gamma$  is not countable, so the dictionary  $\mathcal{D}$  cannot strictly speaking be considered as a frame. However, if we

consider the overall energy of dictionary coefficients, calculated with a sum and a multidimensional integral

$$\sum_{n \in \Gamma} \|\Phi f(u, n)\|^2 = \sum_{n \in \Gamma} \int |\Phi f(u, n)| du,$$

and if there exist two constants  $A > 0$  and  $B > 0$  such that for all  $f \in \mathbf{L}^2(\mathbb{R})$ ,

$$A \|f\|^2 \leq \sum_{n \in \Gamma} \|\Phi f(u, n)\|^2 \leq B \|f\|^2, \quad (5.47)$$

then the frame theory results of the previous section apply. Thus, with an abuse of language, such translation-invariant dictionaries will also be called frames. Theorem 5.11 proves that the frame condition (5.47) is equivalent to a condition on the Fourier transform  $\hat{\phi}_n(\omega)$  of the generators.

**Theorem 5.11.** If there exist two constants  $B \geq A > 0$  such that for almost all  $\omega$  in  $\mathbb{R}^d$

$$A \leq \sum_{n \in \Gamma} |\hat{\phi}_n(\omega)|^2 \leq B, \quad (5.48)$$

then the frame inequality (5.47) is valid for all  $f \in \mathbf{L}^2(\mathbb{R}^d)$ . Any  $\{\tilde{\phi}_n\}_{n \in \Gamma}$  that satisfies for almost all  $\omega$  in  $\mathbb{R}^d$

$$\sum_{n \in \Gamma} \hat{\phi}_n^*(\omega) \hat{\phi}_n(\omega) = 1, \quad (5.49)$$

defines a left inverse

$$f(t) = \sum_{n \in \Gamma} \Phi f(., n) \star \tilde{\phi}_n(t). \quad (5.50)$$

The pseudo inverse (dual frame) is implemented by

$$\widehat{\tilde{\phi}}_n(\omega) = \frac{\hat{\phi}_n(\omega)}{\sum_{n \in \Gamma} |\hat{\phi}_n(\omega)|^2}. \quad (5.51)$$

**Proof.** The frame condition (5.47) means that  $\Phi^* \Phi$  has a spectrum bounded by  $A$  and  $B$ . It results from (5.46) that

$$\Phi^* \Phi f(x) = f \star (\sum_{n \in \Gamma} \phi_n \star \bar{\phi}_n)(x). \quad (5.52)$$

The spectrum of this convolution operator is given by the Fourier transform of  $\sum_{n \in \Gamma} \phi_n \star \bar{\phi}_n(x)$ , which is  $\sum_{n \in \Gamma} |\hat{\phi}_n(\omega)|^2$ . Thus, the frame inequality (5.47) is equivalent to condition (5.48).

Equation (5.50) is proved by taking the Fourier transform on both sides and inserting (5.49).

Theorem 5.5 proves that the dual-frame vectors implementing the pseudo inverse are  $\tilde{\phi}_{n,u} = (\Phi^* \Phi)^{-1} \phi_{n,u}$ . Since  $\Phi^* \Phi$  is the convolution operator (5.52), its inverse is calculated by inverting its transfer function, which yields (5.51). ■

For wavelet or windowed Fourier translation-invariant dictionaries, the theorem condition (5.48) becomes a condition on the Fourier transform of the wavelet  $\hat{\psi}(\omega)$  or on the Fourier transform of the window  $\hat{g}(\omega)$ . As explained in Sections 5.3 and 5.4, more conditions are needed to obtain a frame by discretizing the translation parameter  $u$ .

### **Discrete Translation-Invariant Frames**

For finite-dimensional signals  $f[n] \in \mathbb{C}^N$  a circular translation-invariant frame is obtained with a periodic shift modulo  $N$  of a finite number of generators  $\{\phi_m[n]\}_{0 \leq m < M}$ :

$$\mathcal{D} = \{\phi_{m,p}[n] = \phi_m[(n-p)\text{mod}N]\}_{0 \leq m < M, 0 \leq p < N}.$$

Such translation-invariant frames appear in Section 11.2.3 to define translation-invariant thresholding estimators for noise removal. Similar to Theorem 5.11, Theorem 5.12 gives a necessary and sufficient condition on the discrete Fourier transform  $\hat{\phi}_m[k] = \sum_{n=0}^{N-1} \phi_m[n] e^{-i2\pi k/N}$  of the generators  $\phi_m[n]$  to obtain a frame.

**Theorem 5.12.** A circular translation-invariant dictionary  $\mathcal{D} = \{\phi_{m,p}[n]\}_{0 \leq m < M, 0 \leq n < N}$  is a frame with frame bounds  $0 < A \leq B$  if and only if

$$\forall 0 \leq k < N \quad A \leq \sum_{m=0}^{M-1} |\hat{\phi}_m[k]|^2 \leq B. \quad (5.53)$$

The proof proceeds essentially like the proof of Theorem 5.11, and is left in Exercise 5.8.

## **5.2 TRANSLATION-INVARIANT DYADIC WAVELET TRANSFORM**

The continuous wavelet transform of Section 4.3 decomposes one-dimensional signals  $f \in \mathbf{L}^2(\mathbb{R})$  over a dictionary of translated and dilated wavelets

$$\psi_{u,s}(t) = \frac{1}{\sqrt{s}} \psi\left(\frac{t-u}{s}\right).$$

Translation-invariant wavelet dictionaries are constructed by sampling the scale parameter  $s$  along an exponential sequence  $\{v^j\}_{j \in \mathbb{Z}}$ , while keeping all translation parameters  $u$ . We choose  $v = 2$  to simplify computer implementations:

$$\mathcal{D} = \left\{ \psi_{u,2^j}(t) = \frac{1}{\sqrt{2^j}} \psi\left(\frac{t-u}{2^j}\right) \right\}_{u \in \mathbb{R}, j \in \mathbb{Z}}.$$

The resulting dyadic wavelet transform of  $f \in \mathbf{L}^2(\mathbb{R})$  is defined by

$$Wf(u, 2^j) = \langle f, \psi_{u,2^j} \rangle = \int_{-\infty}^{+\infty} f(t) \frac{1}{\sqrt{2^j}} \psi\left(\frac{t-u}{2^j}\right) dt = f \star \bar{\psi}_{2^j}(u), \quad (5.54)$$

with

$$\tilde{\psi}_{2^j}(t) = \psi_{2^j}(-t) = \frac{1}{2^j} \psi\left(\frac{-t}{2^j}\right).$$

Translation-invariant dyadic wavelet transforms are used in pattern-recognition applications and for denoising with translation-invariant wavelet thresholding estimators, as explained in Section 11.3.1. Fast computations with filter banks are presented in the next two sections.

Theorem 5.11 on translation-invariant dictionaries can be applied to the multiscale wavelet generators  $\phi_n(t) = 2^{-j/2} \psi_{2^j}(t)$ . Since  $\hat{\phi}_n(\omega) = \hat{\psi}(2^j\omega)$ , the Fourier condition (5.48) means that there exist two constants  $A > 0$  and  $B > 0$  such that

$$\forall \omega \in \mathbb{R} - \{0\}, \quad A \leq \sum_{j=-\infty}^{+\infty} |\hat{\psi}(2^j\omega)|^2 \leq B, \quad (5.55)$$

and since  $\Phi f(u, n) = 2^{-j/2} Wf(u, n)$ , Theorem 5.11 proves the frame inequality

$$A \|f\|^2 \leq \sum_{j=-\infty}^{+\infty} \frac{1}{2^j} \|Wf(u, 2^j)\|^2 \leq B \|f\|^2. \quad (5.56)$$

This shows that if the frequency axis is completely covered by dilated dyadic wavelets, as shown in Figure 5.1, then a dyadic wavelet transform defines a complete and stable representation.

Moreover, if  $\tilde{\psi}$  satisfies

$$\forall \omega \in \mathbb{R} - \{0\}, \quad \sum_{j=-\infty}^{+\infty} \hat{\psi}^*(2^j\omega) \hat{\psi}(2^j\omega) = 1, \quad (5.57)$$

then (5.50) applied to  $\tilde{\phi}_n(t) = 2^{-j} \tilde{\psi}(2^{-j}t)$  proves that

$$f(t) = \sum_{j=-\infty}^{+\infty} \frac{1}{2^j} Wf(., 2^j) \star \tilde{\psi}_{2^j}(t). \quad (5.58)$$

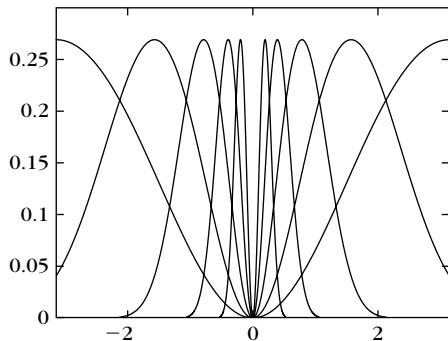
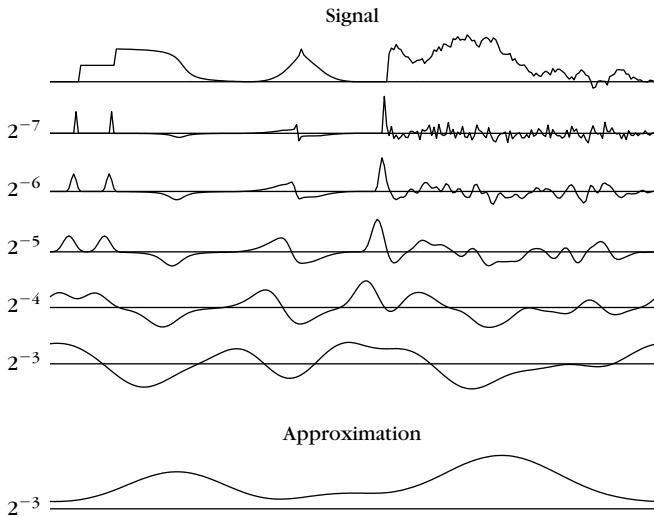


FIGURE 5.1

---

Scaled Fourier transforms  $|\hat{\psi}(2^j\omega)|^2$  computed with (5.69), for  $1 \leq j \leq 5$  and  $\omega \in [-\pi, \pi]$ .

**FIGURE 5.2**

Dyadic wavelet transform  $Wf(u, 2^j)$  computed at scales  $2^{-7} \leq 2^j \leq 2^{-3}$  with filter bank algorithm from Section 5.2.2, for a signal defined over  $[0, 1]$ . The bottom curve carries lower frequencies corresponding to scales larger than  $2^{-3}$ .

Figure 5.2 gives a dyadic wavelet transform computed over five scales with the quadratic spline wavelet shown later in Figure 5.3.

### 5.2.1 Dyadic Wavelet Design

A discrete dyadic wavelet transform can be computed with a fast filter bank algorithm if the wavelet is appropriately designed. The synthesis of these dyadic wavelets is similar to the construction of biorthogonal wavelet bases, explained in Section 7.4. All technical issues related to the convergence of infinite cascades of filters are avoided in this section. Reading Chapter 7 first is necessary for understanding the main results.

Let  $h$  and  $g$  be a pair of finite impulse-response filters. Suppose that  $h$  is a low-pass filter with a transfer function that satisfies  $\hat{h}(0) = \sqrt{2}$ . As in the case of orthogonal and biorthogonal wavelet bases, we construct a scaling function with a Fourier transform:

$$\hat{\phi}(\omega) = \prod_{p=1}^{+\infty} \frac{\hat{h}(2^{-p}\omega)}{\sqrt{2}} = \frac{1}{\sqrt{2}} \hat{h}\left(\frac{\omega}{2}\right) \hat{\phi}\left(\frac{\omega}{2}\right). \quad (5.59)$$

We suppose here that this Fourier transform is a finite-energy function so that  $\phi \in L^2(\mathbb{R})$ . The corresponding wavelet  $\psi$  has a Fourier transform defined by

$$\hat{\psi}(\omega) = \frac{1}{\sqrt{2}} \hat{g}\left(\frac{\omega}{2}\right) \hat{\phi}\left(\frac{\omega}{2}\right). \quad (5.60)$$

Theorem 7.5 proves that both  $\phi$  and  $\psi$  have a compact support because  $h$  and  $g$  have a finite number of nonzero coefficients. The number of vanishing moments of  $\psi$  is equal to the number of zeroes of  $\hat{\psi}(\omega)$  at  $\omega=0$ . Since  $\hat{\phi}(0)=1$ , (5.60) implies that it is also equal to the number of zeros of  $\hat{g}(\omega)$  at  $\omega=0$ .

### Reconstructing Wavelets

Reconstructing wavelets that satisfy (5.49) are calculated with a pair of finite impulse response dual filters  $\tilde{h}$  and  $\tilde{g}$ . We suppose that the following Fourier transform has a finite energy:

$$\hat{\phi}(\omega) = \prod_{p=1}^{+\infty} \frac{\hat{\tilde{h}}(2^{-p}\omega)}{\sqrt{2}} = \frac{1}{\sqrt{2}} \hat{\tilde{h}}\left(\frac{\omega}{2}\right) \hat{\phi}\left(\frac{\omega}{2}\right). \quad (5.61)$$

Let us define

$$\hat{\psi}(\omega) = \frac{1}{\sqrt{2}} \hat{\tilde{g}}\left(\frac{\omega}{2}\right) \hat{\phi}\left(\frac{\omega}{2}\right). \quad (5.62)$$

Theorem 5.13 gives a sufficient condition to guarantee that  $\hat{\psi}$  is the Fourier transform of a reconstruction wavelet.

**Theorem 5.13.** If the filters satisfy

$$\forall \omega \in [-\pi, \pi], \quad \hat{h}(\omega) \hat{h}^*(\omega) + \hat{g}(\omega) \hat{g}^*(\omega) = 2, \quad (5.63)$$

then

$$\forall \omega \in \mathbb{R} - \{0\}, \quad \sum_{j=-\infty}^{+\infty} \hat{\psi}^*(2^j \omega) \hat{\psi}(2^j \omega) = 1. \quad (5.64)$$

**Proof.** The Fourier transform expressions (5.60) and (5.62) prove that

$$\hat{\psi}(\omega) \hat{\psi}^*(\omega) = \frac{1}{2} \hat{\tilde{g}}\left(\frac{\omega}{2}\right) \hat{g}^*\left(\frac{\omega}{2}\right) \hat{\phi}\left(\frac{\omega}{2}\right) \hat{\phi}^*\left(\frac{\omega}{2}\right).$$

Equation (5.63) implies

$$\begin{aligned} \hat{\psi}(\omega) \hat{\psi}^*(\omega) &= \frac{1}{2} \left[ 2 - \hat{\tilde{h}}\left(\frac{\omega}{2}\right) \hat{h}^*\left(\frac{\omega}{2}\right) \right] \hat{\phi}\left(\frac{\omega}{2}\right) \hat{\phi}^*\left(\frac{\omega}{2}\right) \\ &= \hat{\phi}\left(\frac{\omega}{2}\right) \hat{\phi}^*\left(\frac{\omega}{2}\right) - \hat{\tilde{h}}\left(\frac{\omega}{2}\right) \hat{h}^*\left(\frac{\omega}{2}\right). \end{aligned}$$

Therefore,

$$\sum_{j=-l}^k \hat{\psi}(2^j \omega) \hat{\psi}^*(2^j \omega) = \hat{\phi}^*(2^{-l} \omega) \hat{\phi}(2^{-l} \omega) - \hat{\phi}^*(2^k \omega) \hat{\phi}(2^k \omega).$$

Since  $\hat{g}(0)=0$ , (5.63) implies  $\hat{h}(0) \hat{h}^*(0)=2$ . We also impose that  $\hat{h}(0)=\sqrt{2}$ , so one can derive from (5.59) and (5.61) that  $\hat{\phi}(0)=\hat{\phi}^*(0)=1$ . Since  $\phi$  and  $\tilde{\phi}$  belong to  $\mathbf{L}^1(\mathbb{R})$ ,

$\hat{\phi}$  and  $\tilde{\hat{\phi}}$  are continuous, and the Riemann-Lebesgue lemma (Exercise 2.8) proves that  $|\hat{\phi}(\omega)|$  and  $|\tilde{\hat{\phi}}(\omega)|$  decrease to zero when  $\omega$  goes to  $\infty$ . For  $\omega \neq 0$ , letting  $k$  and  $l$  go to  $+\infty$  yields (5.64). ■

Observe that (5.63) is the same as the unit gain condition (7.117) for biorthogonal wavelets. The aliasing cancellation condition (7.116) of biorthogonal wavelets is not required because the wavelet transform is not sampled in time.

### Finite Impulse Response Solution

Let us shift  $h$  and  $g$  to obtain causal filters. The resulting transfer functions  $\hat{h}(\omega)$  and  $\hat{g}(\omega)$  are polynomials in  $e^{-i\omega}$ . We suppose that these polynomials have no common zeros. The Bezout theorem (7.8) on polynomials proves that if  $P(z)$  and  $Q(z)$  are two polynomials of degree  $n$  and  $l$ , with no common zeros, then there exists a unique pair of polynomials  $\tilde{P}(z)$  and  $\tilde{Q}(z)$  of degree  $l - 1$  and  $n - 1$  such that

$$P(z)\tilde{P}(z) + Q(z)\tilde{Q}(z) = 1. \quad (5.65)$$

This guarantees the existence of  $\hat{h}(\omega)$  and  $\hat{g}(\omega)$ , which are polynomials in  $e^{-i\omega}$  and satisfy (5.63). These are the Fourier transforms of the finite impulse response filters  $\tilde{h}$  and  $\tilde{g}$ . However, one must be careful, because the resulting scaling function  $\hat{\phi}$  in (5.61) does not necessarily have a finite energy.

### Spline Dyadic Wavelets

A *box spline* of degree  $m$  is a translation of  $m + 1$  convolutions of  $\mathbf{1}_{[0,1]}$  with itself. It is centered at  $t = 1/2$  if  $m$  is even and at  $t = 0$  if  $m$  is odd. Its Fourier transform is

$$\hat{\phi}(\omega) = \left( \frac{\sin(\omega/2)}{\omega/2} \right)^{m+1} \exp\left( \frac{-i\varepsilon\omega}{2} \right) \quad \text{with } \varepsilon = \begin{cases} 1 & \text{if } m \text{ is even} \\ 0 & \text{if } m \text{ is odd} \end{cases}, \quad (5.66)$$

so

$$\hat{h}(\omega) = \sqrt{2} \frac{\hat{\phi}(2\omega)}{\hat{\phi}(\omega)} = \sqrt{2} \left( \cos \frac{\omega}{2} \right)^{m+1} \exp\left( \frac{-i\varepsilon\omega}{2} \right). \quad (5.67)$$

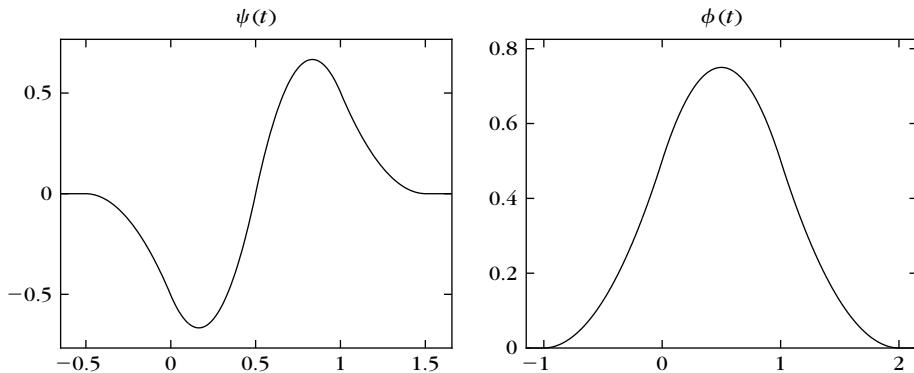
We construct a wavelet that has one vanishing moment by choosing  $\hat{g}(\omega) = O(\omega)$  in the neighborhood of  $\omega = 0$ . For example,

$$\hat{g}(\omega) = -i\sqrt{2} \sin \frac{\omega}{2} \exp\left( \frac{-i\varepsilon\omega}{2} \right). \quad (5.68)$$

The Fourier transform of the resulting wavelet is

$$\hat{\psi}(\omega) = \frac{1}{\sqrt{2}} \hat{g}\left(\frac{\omega}{2}\right) \hat{\phi}\left(\frac{\omega}{2}\right) = \frac{-i\omega}{4} \left( \frac{\sin(\omega/4)}{\omega/4} \right)^{m+2} \exp\left( \frac{-i\omega(1+\varepsilon)}{4} \right). \quad (5.69)$$

It is the first derivative of a box spline of degree  $m + 1$  centered at  $t = (1 + \varepsilon)/4$ . For  $m = 2$ , Figure 5.3 shows the resulting quadratic splines  $\phi$  and  $\psi$ . The

**FIGURE 5.3**

Quadratic spline wavelet and scaling function.

**Table 5.1** Coefficients of Filters Computed from the Transfer Functions (5.67, 5.68, 5.70) for  $m = 2$

$n$	$h[n]/\sqrt{2}$	$\tilde{h}[n]/\sqrt{2}$	$g[n]/\sqrt{2}$	$\tilde{g}[n]/\sqrt{2}$
-2				-0.03125
-1	0.125	0.125		-0.21875
0	0.375	0.375	-0.5	-0.6875
1	0.375	0.375	0.5	0.6875
2	0.125	0.125		0.21875
3				0.03125

*Note:* These filters generate the quadratic spline scaling functions and wavelets shown in Figure 5.3.

dyadic admissibility condition (5.48) is verified numerically for  $A = 0.505$  and  $B = 0.522$ .

To design dual-scaling functions  $\tilde{\phi}$  and wavelets  $\tilde{\psi}$  that are splines, we choose  $\hat{h} = \tilde{h}$ . As a consequence,  $\phi = \tilde{\phi}$  and the reconstruction condition (5.63) imply that

$$\hat{g}(\omega) = \frac{2 - |\hat{h}(\omega)|^2}{\hat{g}^*(\omega)} = -i\sqrt{2} \exp\left(\frac{-i\omega}{2}\right) \sin \frac{\omega}{2} \sum_{n=0}^m \left(\cos \frac{\omega}{2}\right)^{2n}. \quad (5.70)$$

Table 5.1 gives the corresponding filters for  $m = 2$ .

### 5.2.2 Algorithme à Troux

Suppose that the scaling functions and wavelets  $\phi, \psi, \tilde{\phi},$  and  $\tilde{\psi}$  are designed with the filters  $h, g, \tilde{h},$  and  $\tilde{g}$ . A fast dyadic wavelet transform is calculated with a filter bank

algorithm, called *algorithme à trous*, introduced by Holschneider et al. [303]. It is similar to a fast biorthogonal wavelet transform, without subsampling [367, 433].

### Fast Dyadic Transform

The samples  $a_0[n]$  of the input discrete signal are written as a low-pass filtering with  $\phi$  of an analog signal  $f$ , in the neighborhood of  $t = n$ :

$$a_0[n] = f \star \bar{\phi}(n) = \langle f(t), \phi(t - n) \rangle = \int_{-\infty}^{+\infty} f(t) \phi(t - n) dt.$$

This is further justified in Section 7.3.1. For any  $j \geq 0$ , we denote

$$a_j[n] = \langle f(t), \phi_{2^j}(t - n) \rangle \quad \text{with} \quad \phi_{2^j}(t) = \frac{1}{\sqrt{2^j}} \phi\left(\frac{t}{2^j}\right).$$

The dyadic wavelet coefficients are computed for  $j > 0$  over the integer grid

$$d_j[n] = Wf(n, 2^j) = \langle f(t), \psi_{2^j}(t - n) \rangle.$$

For any filter  $x[n]$ , we denote by  $x_j[n]$  the filters obtained by inserting  $2^j - 1$  zeros between each sample of  $x[n]$ . Its Fourier transform is  $\hat{x}(2^j \omega)$ . Inserting zeros in the filters creates holes (*trous* in French). Let  $\tilde{x}_j[n] = x_j[-n]$ . Theorem 5.14 gives convolution formulas that are cascaded to compute a dyadic wavelet transform and its inverse.

**Theorem 5.14.** For any  $j \geq 0$ ,

$$a_{j+1}[n] = a_j \star \tilde{h}_j[n], \quad d_{j+1}[n] = a_j \star \tilde{g}_j[n], \quad (5.71)$$

and

$$a_j[n] = \frac{1}{2} \left( a_{j+1} \star \tilde{h}_j[n] + d_{j+1} \star \tilde{g}_j[n] \right). \quad (5.72)$$

**Proof of (5.71).** Since

$$a_{j+1}[n] = f \star \bar{\phi}_{2^{j+1}}(n) \quad \text{and} \quad d_{j+1}[n] = f \star \bar{\psi}_{2^{j+1}}(n),$$

we verify with (3.3) that their Fourier transforms are, respectively,

$$\hat{a}_{j+1}(\omega) = \sum_{k=-\infty}^{+\infty} \hat{f}(\omega + 2k\pi) \hat{\phi}_{2^{j+1}}^*(\omega + 2k\pi)$$

and

$$\hat{d}_{j+1}(\omega) = \sum_{k=-\infty}^{+\infty} \hat{f}(\omega + 2k\pi) \hat{\psi}_{2^{j+1}}^*(\omega + 2k\pi).$$

The properties (5.61) and (5.62) imply that

$$\begin{aligned} \hat{\phi}_{2^{j+1}}(\omega) &= \sqrt{2^{j+1}} \hat{\phi}(2^{j+1}\omega) = \hat{h}(2^j\omega) \sqrt{2^j} \hat{\phi}(2^j\omega), \\ \hat{\psi}_{2^{j+1}}(\omega) &= \sqrt{2^{j+1}} \hat{\psi}(2^{j+1}\omega) = \hat{g}(2^j\omega) \sqrt{2^j} \hat{\phi}(2^j\omega). \end{aligned}$$

Since  $j \geq 0$ , both  $\hat{h}(2^j\omega)$  and  $\hat{g}(2^j\omega)$  are  $2\pi$  periodic, so

$$\hat{a}_{j+1}(\omega) = \hat{h}^*(2^j\omega) \hat{a}_j(\omega) \quad \text{and} \quad \hat{d}_{j+1}(\omega) = \hat{g}^*(2^j\omega) \hat{a}_j(\omega). \quad (5.73)$$

These two equations are the Fourier transforms of (5.71).

**Proof of (5.72).** Equation (5.73) implies

$$\begin{aligned} \hat{a}_{j+1}(\omega) \hat{h}(2^j\omega) + \hat{d}_{j+1}(\omega) \hat{g}(2^j\omega) = \\ \hat{a}_j(\omega) \hat{h}^*(2^j\omega) \hat{h}(2^j\omega) + \hat{a}_j(\omega) \hat{g}^*(2^j\omega) \hat{g}(2^j\omega). \end{aligned}$$

Inserting the reconstruction condition (5.63) proves that

$$\hat{a}_{j+1}(\omega) \hat{h}(2^j\omega) + \hat{d}_{j+1}(\omega) \hat{g}(2^j\omega) = 2 \hat{a}_j(\omega),$$

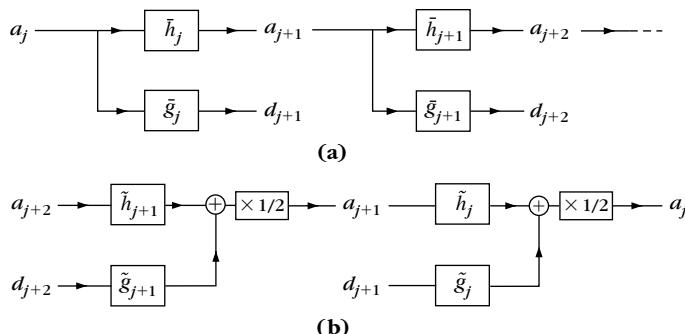
which is the Fourier transform of (5.72). ■

The dyadic wavelet representation of  $a_0$  is defined as the set of wavelet coefficients up to a scale  $2^J$  plus the remaining low-frequency information  $a_J$ :

$$\left[ \{d_j\}_{1 \leq j \leq J}, a_J \right]. \quad (5.74)$$

It is computed from  $a_0$  by cascading the convolutions (5.71) for  $0 \leq j < J$ , as illustrated in Figure 5.4(a). The dyadic wavelet transform of Figure 5.2 is calculated with the filter bank algorithm. The original signal  $a_0$  is recovered from its wavelet representation (5.74) by iterating (5.72) for  $J > j \geq 0$ , as illustrated in Figure 5.4(b).

If the input signal  $a_0[n]$  has a finite size of  $N$  samples, the convolutions (5.71) are replaced by circular convolutions. The maximum scale  $2^J$  is then limited to  $N$ , and for  $J = \log_2 N$ , one can verify that  $a_J[n]$  is constant and equal to  $N^{-1/2} \sum_{n=0}^{N-1} a_0[n]$ . Suppose that  $h$  and  $g$  have, respectively,  $K_h$  and  $K_g$  nonzero samples. The “dilated” filters  $\tilde{h}_j$  and  $\tilde{g}_j$  have the same number of nonzero coefficients. Therefore, the number of multiplications needed to compute  $a_{j+1}$  and  $d_{j+1}$  from  $a_j$  or the reverse



**FIGURE 5.4**

(a) The dyadic wavelet coefficients are computed by cascading convolutions with dilated filters  $\tilde{h}_j$  and  $\tilde{g}_j$ . (b) The original signal is reconstructed through convolutions with  $\tilde{h}_j$  and  $\tilde{g}_j$ . A multiplication by  $1/2$  is necessary to recover the next finer scale signal  $a_j$ .

is equal to  $(K_h + K_g)N$ . Thus, for  $J = \log_2 N$ , the dyadic wavelet representation (5.74) and its inverse are calculated with  $(K_h + K_g)N \log_2 N$  multiplications and additions.

### 5.3 SUBSAMPLED WAVELET FRAMES

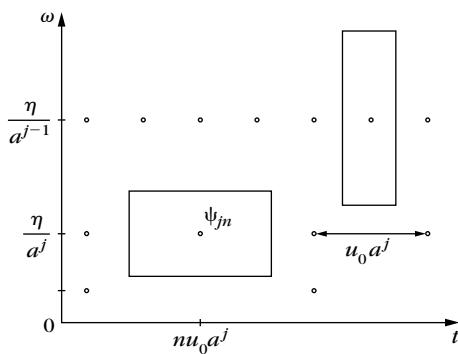
Wavelet frames are constructed by sampling the scale parameter but also the translation parameter of a wavelet dictionary. A real continuous wavelet transform of  $f \in L^2(\mathbb{R})$  is defined in Section 4.3 by

$$Wf(u, s) = \langle f, \psi_{u,s} \rangle \quad \text{with} \quad \psi_{u,s}(t) = \frac{1}{\sqrt{s}} \psi\left(\frac{t-u}{s}\right),$$

where  $\psi$  is a real wavelet. Imposing  $\|\psi\| = 1$  implies that  $\|\psi_{u,s}\| = 1$ .

Intuitively, to construct a frame we need to cover the time-frequency plane with the Heisenberg boxes of the corresponding discrete wavelet family. A wavelet  $\psi_{u,s}$  has an energy in time that is centered at  $u$  over a domain proportional to  $s$ . For positive frequencies, its Fourier transform  $\hat{\psi}_{u,s}$  has a support centered at a frequency  $\eta/s$ , with a spread proportional to  $1/s$ . To obtain a full cover, we sample  $s$  along an exponential sequence  $\{a^j\}_{j \in \mathbb{Z}}$ , with a sufficiently small dilation step  $a > 1$ . The time translation  $u$  is sampled uniformly at intervals proportional to the scale  $a^j$ , as illustrated in Figure 5.5. Let us denote

$$\psi_{j,n}(t) = \frac{1}{\sqrt{a^j}} \psi\left(\frac{t-nu_0a^j}{a^j}\right).$$



**FIGURE 5.5**

The Heisenberg box of a wavelet  $\psi_{j,n}$  scaled by  $s = a^j$  has a time and frequency width proportional to  $a^j$  and  $a^{-j}$ , respectively. The time-frequency plane is covered by these boxes if  $u_0$  and  $a$  are sufficiently small.

In the following, we give without proof, some necessary conditions and sufficient conditions on  $\psi$ ,  $a$  and  $u_0$  so that  $\{\psi_{j,n}\}_{(j,n)\in\mathbb{Z}^2}$  is a frame of  $\mathbf{L}^2(\mathbb{R})$ .

### Necessary Conditions

We suppose that  $\psi$  is real, normalized, and satisfies the admissibility condition of Theorem 4.4:

$$C_\psi = \int_0^{+\infty} \frac{|\hat{\psi}(\omega)|^2}{\omega} d\omega < +\infty. \quad (5.75)$$

**Theorem 5.15: Daubechies.** If  $\{\psi_{j,n}\}_{(j,n)\in\mathbb{Z}^2}$  is a frame of  $\mathbf{L}^2(\mathbb{R})$ , then the frame bounds satisfy

$$A \leq \frac{C_\psi}{u_0 \log_e a} \leq B, \quad (5.76)$$

$$\forall \omega \in \mathbb{R} - \{0\}, \quad A \leq \frac{1}{u_0} \sum_{j=-\infty}^{+\infty} |\hat{\psi}(a^j \omega)|^2 \leq B. \quad (5.77)$$

This theorem is proved in [19, 163]. Condition (5.77) is equivalent to the frame condition (5.55) for a translation-invariant dyadic wavelet transform, for which the parameter  $u$  is not sampled. It requires that the Fourier axis is covered by wavelets dilated by  $\{a^j\}_{j\in\mathbb{Z}}$ . The inequality (5.76), which relates the sampling density  $u_0 \log_e a$  to the frame bounds, is proved in [19]. It shows that the frame is an orthonormal basis if and only if

$$A = B = \frac{C_\psi}{u_0 \log_e a} = 1.$$

Chapter 7 constructs wavelet orthonormal bases of  $\mathbf{L}^2(\mathbb{R})$  with regular wavelets of compact support.

### Sufficient Conditions

Theorem 5.16 proved by Daubechies [19] provides a lower and upper bound for the frame bounds  $A$  and  $B$ , depending on  $\psi$ ,  $u_0$ , and  $a$ .

**Theorem 5.16: Daubechies.** Let us define

$$\theta(\xi) = \sup_{1 \leq |\omega| \leq a} \sum_{j=-\infty}^{+\infty} |\hat{\psi}(a^j \omega)| |\hat{\psi}(a^j \omega + \xi)| \quad (5.78)$$

and

$$\Delta = \sum_{\substack{k=-\infty \\ k \neq 0}}^{+\infty} \left[ \theta\left(\frac{2\pi k}{u_0}\right) \theta\left(\frac{-2\pi k}{u_0}\right) \right]^{1/2}.$$

If  $u_0$  and  $a$  are such that

$$A_0 = \frac{1}{u_0} \left( \inf_{1 \leq |\omega| \leq a} \sum_{j=-\infty}^{+\infty} |\hat{\psi}(a^j \omega)|^2 - \Delta \right) > 0, \quad (5.79)$$

and

$$B_0 = \frac{1}{u_0} \left( \sup_{1 \leq |\omega| \leq a} \sum_{j=-\infty}^{+\infty} |\hat{\psi}(a^j \omega)|^2 + \Delta \right) < +\infty, \quad (5.80)$$

then  $\{\psi_{j,n}\}_{(j,n) \in \mathbb{Z}^2}$  is a frame of  $\mathbf{L}^2(\mathbb{R})$ . The constants  $A_0$  and  $B_0$  are respectively lower and upper bounds of the frame bounds  $A$  and  $B$ .

The sufficient conditions (5.79) and (5.80) are similar to the necessary condition (5.77). If  $\Delta$  is small relative to  $\inf_{1 \leq |\omega| \leq a} \sum_{j=-\infty}^{+\infty} |\hat{\psi}(a^j \omega)|^2$ , then  $A_0$  and  $B_0$  are close to the optimal frame bounds  $A$  and  $B$ . For a fixed dilation step  $a$ , the value of  $\Delta$  decreases when the time-sampling interval  $u_0$  decreases.

### Dual Frame

Theorem 5.5 gives a general formula for computing the dual-wavelet frame vectors

$$\tilde{\psi}_{j,n} = (\Phi^* \Phi)^{-1} \psi_{j,n}. \quad (5.81)$$

One could reasonably hope that the dual functions  $\tilde{\psi}_{j,n}$  would be obtained by scaling and translating a dual wavelet  $\tilde{\psi}$ . The unfortunate reality is that this is generally not true. In general, the operator  $\Phi^* \Phi$  does not commute with dilations by  $a^j$ , so  $(\Phi^* \Phi)^{-1}$  does not commute with these dilations either. On the other hand, one can prove that  $(\Phi^* \Phi)^{-1}$  commutes with translations by  $na^j u_0$ , which means that

$$\tilde{\psi}_{j,n}(t) = \tilde{\psi}_{j,0}(t - na^j u_0). \quad (5.82)$$

Thus, the dual frame  $\{\tilde{\psi}_{j,n}\}_{(j,n) \in \mathbb{Z}^2}$  is obtained by calculating each elementary function  $\tilde{\psi}_{j,0}$  with (5.81), and translating them with (5.82). The situation is much simpler for tight frames, where the dual frame is equal to the original wavelet frame.

### Mexican Hat Wavelet

The normalized second derivative of a Gaussian is

$$\psi(t) = \frac{2}{\sqrt{3}} \pi^{-1/4} (t^2 - 1) \exp\left(\frac{-t^2}{2}\right). \quad (5.83)$$

Its Fourier transform is

$$\hat{\psi}(\omega) = -\frac{\sqrt{8} \pi^{1/4} \omega^2}{\sqrt{3}} \exp\left(\frac{-\omega^2}{2}\right).$$

The graph of these functions is shown in Figure 4.6.

**Table 5.2** Estimated Frame Bounds for the Mexican Hat Wavelet

$a$	$u_0$	$A_0$	$B_0$	$B_0/A_0$
2	0.25	13.091	14.183	1.083
2	0.5	6.546	7.092	1.083
2	1.0	3.223	3.596	1.116
2	1.5	0.325	4.221	12.986
$2^{\frac{1}{2}}$	0.25	27.273	27.278	1.0002
$2^{\frac{1}{2}}$	0.5	13.673	13.639	1.0002
$2^{\frac{1}{2}}$	1.0	6.768	6.870	1.015
$2^{\frac{1}{2}}$	1.75	0.517	7.276	14.061
$2^{\frac{1}{4}}$	0.25	54.552	54.552	1.0000
$2^{\frac{1}{4}}$	0.5	27.276	27.276	1.0000
$2^{\frac{1}{4}}$	1.0	13.586	13.690	1.007
$2^{\frac{1}{4}}$	1.75	2.928	12.659	4.324

Source: Computed with Theorem 5.16 [19].

The dilation step  $\alpha$  is generally set to be  $\alpha = 2^{1/v}$  where  $v$  is the number of intermediate scales (voices) for each octave. Table 5.2 gives the estimated frame bounds  $A_0$  and  $B_0$  computed by Daubechies [19] with the formula of Theorem 5.16. For  $v \geq 2$  voices per octave, the frame is nearly tight when  $u_0 \leq 0.5$ , in which case the dual frame can be approximated by the original wavelet frame. As expected from (5.76), when  $A \approx B$ ,

$$A \approx B \approx \frac{C_\psi}{u_0 \log_e \alpha} = \frac{v}{u_0} C_\psi \log_2 e.$$

The frame bounds increase proportionally to  $v/u_0$ . For  $\alpha = 2$ , we see that  $A_0$  decreases brutally from  $u_0 = 1$  to  $u_0 = 1.5$ . For  $u_0 = 1.75$ , the wavelet family is not a frame anymore. For  $\alpha = 2^{1/2}$ , the same transition appears for a larger  $u_0$ .

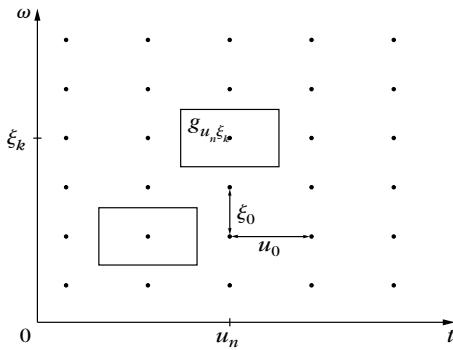
## 5.4 WINDOWED FOURIER FRAMES

Frame theory gives conditions for discretizing the windowed Fourier transform while retaining a complete and stable representation. The windowed Fourier transform of  $f \in L^2(\mathbb{R})$  is defined in Section 4.2 by

$$\mathcal{S}f(u, \xi) = \langle f, g_{u, \xi} \rangle,$$

with

$$g_{u, \xi}(t) = g(t - u) e^{i\xi t}.$$

**FIGURE 5.6**

A windowed Fourier frame is obtained by covering the time-frequency plane with a regular grid of windowed Fourier atoms, translated by  $u_n = n u_0$  in time and by  $\xi_k = k \xi_0$  in frequency.

Setting  $\|g\| = 1$  implies that  $\|g_{u,\xi}\| = 1$ . A discrete windowed Fourier transform representation

$$\{Sf(u_n, \xi_k) = \langle f, g_{u_n, \xi_k} \rangle\}_{(n,k) \in \mathbb{Z}^2}$$

is complete and stable if  $\{g_{u,n, \xi_k}\}_{(n,k) \in \mathbb{Z}^2}$  is a frame of  $\mathbf{L}^2(\mathbb{R})$ .

Intuitively, one can expect that the discrete windowed Fourier transform is complete if the Heisenberg boxes of all atoms  $\{g_{u,n, \xi_k}\}_{(n,k) \in \mathbb{Z}^2}$  fully cover the time-frequency plane. Section 4.2 shows that the Heisenberg box of  $g_{u_n, \xi_k}$  is centered in the time-frequency plane at  $(u_n, \xi_k)$ . Its size is independent of  $u_n$  and  $\xi_k$ . It depends on the time-frequency spread of the window  $g$ . Thus, a complete cover of the plane is obtained by translating these boxes over a uniform rectangular grid, as illustrated in Figure 5.6. The time and frequency parameters  $(u, \xi)$  are discretized over a rectangular grid with time and frequency intervals of size  $u_0$  and  $\xi_0$ . Let us denote

$$g_{n,k}(t) = g(t - nu_0) \exp(ik\xi_0 t).$$

The sampling intervals  $(u_0, \xi_0)$  must be adjusted to the time-frequency spread of  $g$ .

### Window Scaling

Suppose that  $\{g_{n,k}\}_{(n,k) \in \mathbb{Z}^2}$  is a frame of  $\mathbf{L}^2(\mathbb{R})$  with frame bounds  $A$  and  $B$ . Let us dilate the window  $g_s(t) = s^{-1/2}g(t/s)$ . It increases by  $s$  the time width of the Heisenberg box of  $g$  and reduces by  $s$  its frequency width. Thus, we obtain the same cover of the time-frequency plane by increasing  $u_0$  by  $s$  and reducing  $\xi_0$  by  $s$ . Let

$$g_{s,n,k}(t) = g_s(t - nsu_0) \exp\left(ik\frac{\xi_0}{s}t\right). \quad (5.84)$$

We prove that  $\{g_{s,n,k}\}_{(n,k)\in\mathbb{Z}^2}$  satisfies the same frame inequalities as  $\{g_{n,k}\}_{(n,k)\in\mathbb{Z}^2}$ , with the same frame bounds  $A$  and  $B$ , by a change of variable  $t' = ts$  in the inner product integrals.

### 5.4.1 Tight Frames

Tight frames are easier to manipulate numerically since the dual frame is equal to the original frame. Daubechies, Grossmann, and Meyer [197] give sufficient conditions for building a window of compact support that generates a tight frame.

**Theorem 5.17:** *Daubechies, Grossmann, Meyer.* Let  $g$  be a window that has a support included in  $[-\pi/\xi_0, \pi/\xi_0]$ . If

$$\forall t \in \mathbb{R}, \quad \frac{2\pi}{\xi_0} \sum_{n=-\infty}^{+\infty} |g(t - nu_0)|^2 = A > 0, \quad (5.85)$$

then  $\{g_{n,k}(t) = g(t - nu_0) e^{ik\xi_0 t}\}_{(n,k)\in\mathbb{Z}^2}$  is a tight frame  $L^2(\mathbb{R})$  with a frame bound equal to  $A$ .

**Proof.** The function  $g(t - nu_0)f(t)$  has a support in  $[nu_0 - \pi/\xi_0, nu_0 + \pi/\xi_0]$ . Since  $\{e^{ik\xi_0 t}\}_{k\in\mathbb{Z}}$  is an orthogonal basis of this space, we have

$$\begin{aligned} \int_{-\infty}^{+\infty} |g(t - nu_0)|^2 |f(t)|^2 dt &= \int_{nu_0 - \pi/\xi_0}^{nu_0 + \pi/\xi_0} |g(t - nu_0)|^2 |f(t)|^2 dt \\ &= \frac{\xi_0}{2\pi} \sum_{k=-\infty}^{+\infty} |\langle g(u - nu_0)f(u), e^{ik\xi_0 u} \rangle|^2. \end{aligned}$$

Since  $g_{n,k}(t) = g(t - nu_0) e^{ik\xi_0 t}$ , we get

$$\int_{-\infty}^{+\infty} |g(t - nu_0)|^2 |f(t)|^2 dt = \frac{\xi_0}{2\pi} \sum_{k=-\infty}^{+\infty} |\langle f, g_{n,k} \rangle|^2.$$

Summing over  $n$  and inserting (5.85) proves that  $A \|f\|^2 = \sum_{k,n=-\infty}^{+\infty} |\langle f, g_{n,k} \rangle|^2$ , and therefore, that  $\{g_{n,k}\}_{(n,k)\in\mathbb{Z}^2}$  is a tight frame of  $L^2(\mathbb{R})$ . ■

Since  $g$  has a support in  $[-\pi/\xi_0, \pi/\xi_0]$  the condition (5.85) implies that

$$\frac{2\pi}{u_0 \xi_0} \geq 1,$$

so that there is no whole between consecutive windows  $g(t - nu_0)$  and  $g(t - (n+1)u_0)$ . If we impose that  $1 \leq 2\pi/(u_0 \xi_0) \leq 2$ , then only consecutive windows have supports that overlap. The square root of a Hanning window

$$g(t) = \sqrt{\frac{\xi_0}{\pi}} \cos\left(\frac{\xi_0 t}{2}\right) \mathbf{1}_{[-\pi/\xi_0, \pi/\xi_0]}(t)$$

is a positive normalized window that satisfies (5.85) with  $u_0 = \pi/\xi_0$  and a redundancy factor of  $A = 2$ . The design of other windows is studied in Section 8.4.2 for local cosine bases.

### Discrete Window Fourier Tight Frames

To construct a windowed Fourier tight frame of  $\mathbb{C}^N$ , the Fourier basis  $\{e^{ik\xi_0 t}\}_{k \in \mathbb{Z}}$  of  $L^2[-\pi/\xi_0, \pi/\xi_0]$  is replaced by the discrete Fourier basis  $\{e^{i2\pi kn/K}\}_{0 \leq k < K}$  of  $\mathbb{C}^K$ . Theorem 5.18 is a discrete equivalent of Theorem 5.17.

**Theorem 5.18.** Let  $g[n]$  be an  $N$  periodic discrete window with a support and restricted to  $[-N/2, N/2]$  that is included in  $[-K/2, K/2 - 1]$ . If  $M$  divides  $N$  and

$$\forall 0 \leq n < N, \quad K \sum_{m=0}^{N/M-1} |g[n - mM]|^2 = A > 0, \quad (5.86)$$

then  $\{g_{m,k}[n] = g[n - mM] e^{i2\pi kn/K}\}_{0 \leq k < K, 0 \leq m < N/M}$  is a tight frame  $\mathbb{C}^N$  with a frame bound equal to  $A$ .

The proof of this theorem follows the same steps as the proof of Theorem 5.17. It is left in Exercise 5.10. There are  $N/M$  translated windows and thus  $NK/M$  windowed Fourier coefficients. For a fixed window position indexed by  $m$ , the discrete windowed Fourier coefficients are the discrete Fourier coefficients of the windowed signal

$$Sf[m, k] = \langle f, g_{m,k} \rangle = \sum_{n=K/2}^{K/2-1} f[n] g[n - mM] e^{-i2\pi kn/K} \quad \text{for } 0 \leq k < K.$$

They are computed with  $O(K \log_2 K)$  operations with an FFT. Over all windows, this requires a total of  $O(NK/M \log_2 K)$  operations. We generally choose  $1 < K/M \leq 2$  so that only consecutive windows overlap. The square root of a Hanning window  $g[n] = \sqrt{2/K} \cos(\pi n/K)$  satisfies (5.86) for  $M = K/2$  and a redundancy factor  $A = 2$ . Figure 5.7 shows the log spectrogram  $\log |Sf[m, k]|^2$  of the windowed Fourier frame coefficients computed with a square root Hanning window for a musical recording.

#### 5.4.2 General Frames

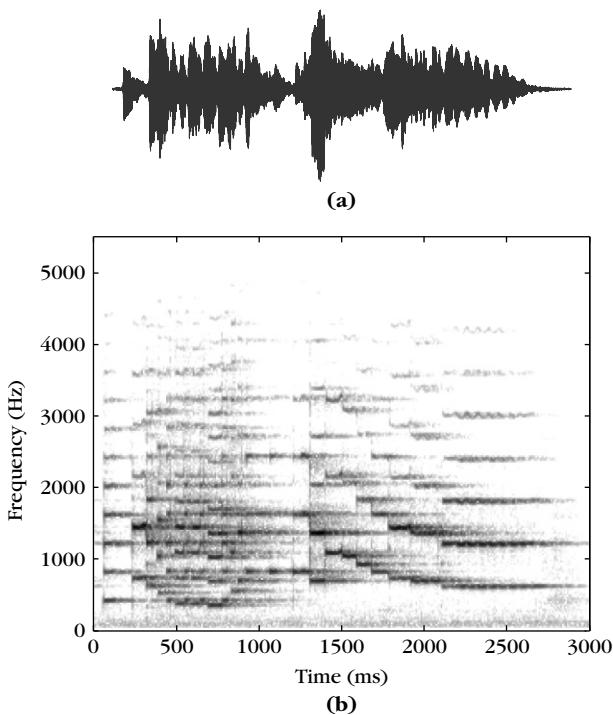
For general windowed Fourier frames of  $L^2(\mathbb{R}^2)$ , Daubechies [19] proved several necessary conditions on  $g$ ,  $u_0$  and  $\xi_0$  to guarantee that  $\{g_{n,k}\}_{(n,k) \in \mathbb{Z}^2}$  is a frame of  $L^2(\mathbb{R})$ . We do not reproduce the proofs, but summarize the main results.

**Theorem 5.19:** *Daubechies.* The windowed Fourier family  $\{g_{n,k}\}_{(n,k) \in \mathbb{Z}^2}$  is a frame only if

$$\frac{2\pi}{u_0 \xi_0} \geq 1. \quad (5.87)$$

The frame bounds  $A$  and  $B$  necessarily satisfy

$$A \leq \frac{2\pi}{u_0 \xi_0} \leq B, \quad (5.88)$$

**FIGURE 5.7**

(a) Musical recording. (b) Log spectrogram  $\log |Sf[m, k]|^2$  computed with a square root Hanning window.

$$\forall t \in \mathbb{R}, \quad A \leq \frac{2\pi}{\xi_0} \sum_{n=-\infty}^{+\infty} |g(t - nu_0)|^2 \leq B, \quad (5.89)$$

$$\forall \omega \in \mathbb{R}, \quad A \leq \frac{1}{u_0} \sum_{k=-\infty}^{+\infty} |\hat{g}(\omega - k\xi_0)|^2 \leq B. \quad (5.90)$$

The ratio  $2\pi/(u_0\xi_0)$  measures the density of windowed Fourier atoms in the time-frequency plane. The first condition (5.87) ensures that this density is greater than 1 because the covering ability of each atom is limited. Inequalities (5.89) and (5.90) are proved in full generality by Chui and Shi [163]. They show that the uniform time translations of  $g$  must completely cover the time axis, and the frequency translations of its Fourier transform  $\hat{g}$  must similarly cover the frequency axis.

Since all windowed Fourier vectors are normalized, the frame is an orthogonal basis only if  $A = B = 1$ . The frame bound condition (5.88) shows that this is possible only at the critical sampling density  $u_0\xi_0 = 2\pi$ . The Balian-Low theorem 5.20 [93] proves that  $g$  is then either nonsmooth or has a slow time decay.

**Theorem 5.20:** *Balian-Low.* If  $\{g_{n,k}\}_{(n,k)\in\mathbb{Z}^2}$  is a windowed Fourier frame with  $u_0\xi_0 = 2\pi$ , then

$$\int_{-\infty}^{+\infty} t^2 |g(t)|^2 dt = +\infty \quad \text{or} \quad \int_{-\infty}^{+\infty} \omega^2 |\hat{g}(\omega)|^2 d\omega = +\infty. \quad (5.91)$$

This theorem proves that we cannot construct an orthogonal windowed Fourier basis with a differentiable window  $g$  of compact support. On the other hand, one can verify that the discontinuous rectangular window

$$g = \frac{1}{\sqrt{u_0}} \mathbf{1}_{[-u_0/2, u_0/2]}$$

yields an orthogonal windowed Fourier basis for  $u_0\xi_0 = 2\pi$ . This basis is rarely used because of the bad frequency localization of  $\hat{g}$ .

### Sufficient Conditions

Theorem 5.21 proved by Daubechies [195] gives sufficient conditions on  $u_0, \xi_0$ , and  $g$  for constructing a windowed Fourier frame.

**Theorem 5.21:** *Daubechies.* Let us define

$$\theta(u) = \sup_{0 \leq t \leq u_0} \sum_{n=-\infty}^{+\infty} |g(t - nu_0)| |g(t - nu_0 + u)| \quad (5.92)$$

and

$$\Delta = \sum_{\substack{k=-\infty \\ k \neq 0}}^{+\infty} \left[ \theta\left(\frac{2\pi k}{\xi_0}\right) \theta\left(\frac{-2\pi k}{\xi_0}\right) \right]^{1/2}. \quad (5.93)$$

If  $u_0$  and  $\xi_0$  satisfy

$$A_0 = \frac{2\pi}{\xi_0} \left( \int_{0 \leq t \leq u_0} \sum_{n=-\infty}^{+\infty} |g(t - nu_0)|^2 - \Delta \right) > 0 \quad (5.94)$$

and

$$B_0 = \frac{2\pi}{\xi_0} \left( \sup_{0 \leq t \leq u_0} \sum_{n=-\infty}^{+\infty} |g(t - nu_0)|^2 + \Delta \right) < +\infty, \quad (5.95)$$

then  $\{g_{n,k}\}_{(n,k)\in\mathbb{Z}^2}$  is a frame. The constants  $A_0$  and  $B_0$  are, respectively, lower bounds and upper bounds of the frame bounds  $A$  and  $B$ .

Observe that the only difference between the sufficient conditions (5.94, 5.95) and the necessary condition (5.89) is the addition and subtraction of  $\Delta$ . If  $\Delta$  is small compared to  $\inf_{0 \leq t \leq u_0} \sum_{n=-\infty}^{+\infty} |g(t - nu_0)|^2$ , then  $A_0$  and  $B_0$  are close to the optimal frame bounds  $A$  and  $B$ .

### Dual Frame

Theorem 5.5 proves that the dual-windowed frame vectors are

$$\tilde{g}_{n,k} = (\Phi^* \Phi)^{-1} g_{n,k}. \quad (5.96)$$

Theorem 5.22 shows that this dual frame is also a windowed Fourier frame, which means that its vectors are time and frequency translations of a new window  $\tilde{g}$ .

**Theorem 5.22.** Dual-windowed Fourier vectors can be rewritten as

$$\tilde{g}_{n,k}(t) = \tilde{g}(t - nu_0) \exp(ik\xi_0 t),$$

where  $\tilde{g}$  is the dual window

$$\tilde{g} = (\Phi^* \Phi)^{-1} g. \quad (5.97)$$

**Proof.** This result is proved by showing first that  $\Phi^* \Phi$  commutes with time and frequency translations proportional to  $u_0$  and  $\xi_0$ . If  $\phi \in L^2(\mathbb{R})$  and  $\phi_{m,l}(t) = \phi(t - mu_0) \exp(il\xi_0 t)$ , we verify that

$$\Phi^* \Phi \phi_{m,l}(t) = \exp(il\xi_0 t) \Phi^* \Phi \phi(t - mu_0).$$

Indeed,

$$\Phi^* \Phi \phi_{m,l} = \sum_{(n,k) \in \mathbb{Z}^2} \langle \phi_{m,l}, g_{n,k} \rangle g_{n,k}$$

and a change of variable yields

$$\langle \phi_{m,l}, g_{n,k} \rangle = \langle \phi, g_{n-m, k-l} \rangle.$$

Consequently,

$$\begin{aligned} \Phi^* \Phi \phi_{m,l}(t) &= \sum_{(n,k) \in \mathbb{Z}^2} \langle \phi, g_{n-m, k-l} \rangle \exp(il\xi_0 t) g_{n-m, k-l}(t - mu_0) \\ &= \exp(il\xi_0 t) \Phi^* \Phi \phi(t - mu_0). \end{aligned}$$

Since  $\Phi^* \Phi$  commutes with these translations and frequency modulations, we verify that  $(\Phi^* \Phi)^{-1}$  necessarily commutes with the same group operations. Thus,

$$\tilde{g}_{n,k}(t) = (\Phi^* \Phi)^{-1} g_{n,k} = \exp(ik\xi_0) (\Phi^* \Phi)^{-1} g_{0,0}(t - nu_0) = \exp(ik\xi_0) \tilde{g}(t - nu_0). \quad \blacksquare$$

### Gaussian Window

The Gaussian window

$$g(t) = \pi^{-1/4} \exp\left(-\frac{t^2}{2}\right) \quad (5.98)$$

has a Fourier transform  $\hat{g}$  that is a Gaussian with the same variance. The time and frequency spreads of this window are identical. Therefore, let us choose equal sampling intervals in time and frequency:  $u_0 = \xi_0$ . For the same product  $u_0 \xi_0$  other

**Table 5.3** Frame Bounds for the Gaussian Window (5.98) and  $u_0 = \xi_0$

$u_0\xi_0$	$A_0$	$B_0$	$B_0/A_0$
$\pi/2$	3.9	4.1	1.05
$3\pi/4$	2.5	2.8	1.1
$\pi$	1.6	2.4	1.5
$4\pi/3$	0.58	2.1	3.6
$1.9\pi$	0.09	2.0	22

choices would degrade the frame bounds. If  $g$  is dilated by  $s$  then the time and frequency sampling intervals must become  $su_0$  and  $\xi_0/s$ .

If the time-frequency sampling density is above the critical value of  $2\pi/(u_0\xi_0) > 1$ , then Daubechies [195] proves that  $\{g_{n,k}\}_{(n,k)\in\mathbb{Z}^2}$  is a frame. When  $u_0\xi_0$  tends to  $2\pi$ , the frame bound  $A$  tends to 0. For  $u_0\xi_0 = 2\pi$ , the family  $\{g_{n,k}\}_{(n,k)\in\mathbb{Z}^2}$  is complete in  $L^2(\mathbb{R})$ , which means that any  $f \in L^2(\mathbb{R})$  is entirely characterized by the inner products  $\{\langle f, g_{n,k} \rangle\}_{(n,k)\in\mathbb{Z}^2}$ . However, the Balian-Low theorem (5.20) proves that it cannot be a frame and one can indeed verify that  $A = 0$  [195]. This means that the reconstruction of  $f$  from these inner products is unstable.

Table 5.3 gives the estimated frame bounds  $A_0$  and  $B_0$  calculated with Theorem 5.21, for different values of  $u_0 = \xi_0$ . For  $u_0\xi_0 = \pi/2$ , which corresponds to time and frequency sampling intervals that are half the critical sampling rate, the frame is nearly tight. As expected,  $A \approx B \approx 4$ , which verifies that the redundancy factor is 4 (2 in time and 2 in frequency). Since the frame is almost tight, the dual frame is approximately equal to the original frame, which means that  $\tilde{g} \approx g$ . When  $u_0\xi_0$  increases we see that  $A$  decreases to zero and  $\tilde{g}$  deviates more and more from a Gaussian. In the limit  $u_0\xi_0 = 2\pi$ , the dual window  $\tilde{g}$  is a discontinuous function that does not belong to  $L^2(\mathbb{R})$ . These results can be extended to discrete windowed Fourier transforms computed with a discretized Gaussian window [501].

## 5.5 MULTISCALE DIRECTIONAL FRAMES FOR IMAGES

To reveal geometric image properties, wavelet frames are constructed with mother wavelets having a direction selectivity, providing information on the direction of sharp transitions such as edges and textures. Directional wavelet frames are described in Section 5.5.1.

Wavelet frames yield high-amplitude coefficients in the neighborhood of edges, and cannot take advantage of their geometric regularity to improve the sparsity of the representation. Curvelet frames, described in Section 5.5.2, are constructed with elongated waveforms that follow directional image structures and improve the representation sparsity.

### 5.5.1 Directional Wavelet Frames

A directional wavelet transform decomposes images over directional wavelets that are translated, rotated, and dilated at dyadic scales. Such transforms appear in many image-processing applications and physiological models. Applications to texture discrimination are also discussed.

A directional wavelet  $\psi^\alpha(x)$  with  $x = (x_1, x_2) \in \mathbb{R}^2$  of angle  $\alpha$  is a wavelet having  $p$  directional vanishing moments along any one-dimensional line of direction  $\alpha + \pi/2$  in the plane

$$\forall \rho \in \mathbb{R}, \quad \int \psi^\alpha(\rho \cos \alpha - u \sin \alpha, \rho \sin \alpha + u \cos \alpha) u^k du = 0 \quad \text{for } 0 \leq k < p, \quad (5.99)$$

but does not have directional vanishing moments along the direction  $\alpha$ . Such a wavelet oscillates in the direction of  $\alpha + \pi/2$  but not in the direction  $\alpha$ . It is orthogonal to any two-dimensional polynomial of degree strictly smaller than  $p$  (Exercise 5.21).

The set  $\Theta$  of chosen directions are typically uniform in  $[0, \pi]$ :  $\Theta = \{\alpha = k\pi/K \text{ for } 0 \leq k < K\}$ . Dilating these directional wavelets by factors  $2^j$  and translating them by any  $u \in \mathbb{R}$  yields a translation-invariant directional wavelet family:

$$\{\psi_{2^j}^\alpha(x - u)\}_{u \in \mathbb{R}^2, j \in \mathbb{Z}, \alpha \in \Theta} \quad \text{with} \quad \psi_{2^j}^\alpha(x) = 2^{-j} \psi^\alpha(2^{-j}x). \quad (5.100)$$

Directional wavelets may be derived by rotating a single mother wavelet  $\psi(x_1, x_2)$  having vanishing moments in the horizontal direction, with a rotation operator  $R_\alpha$  of angle  $\alpha$  in  $\mathbb{R}^2$ .

A dyadic directional wavelet transform of  $f$  computes the inner product with each wavelet:

$$Wf(u, 2^j, \alpha) = \langle f, \psi_{2^j, u}^\alpha \rangle \quad \text{where} \quad \psi_{2^j, u}^\alpha(x) = \psi_{2^j}^\alpha(x - u).$$

This dyadic wavelet transform can also be written as convolutions with directional wavelets:

$$Wf(u, 2^j, \alpha) = f \star \bar{\psi}_{2^j}^\alpha(u) \quad \text{where} \quad \bar{\psi}_{2^j}^\alpha(x) = \psi_{2^j}^\alpha(-x).$$

A wavelet  $\psi_{2^j}^\alpha(x - u)$  has a support dilated by  $2^j$ , located in the neighborhood of  $u$  and oscillates in the direction of  $\alpha + \pi/2$ . If  $f(x)$  is constant over the support of  $\psi_{2^j, u}^\alpha$  along lines of direction  $\alpha + \pi/2$ , then  $\langle f, \psi_{2^j, u}^\alpha \rangle = 0$  because of its directional vanishing moments. In particular, this coefficient vanishes in the neighborhood of an edge having a tangent in the direction  $\alpha + \pi/2$ . If the edge angle deviates from  $\alpha + \pi/2$ , then it produces large amplitude coefficients, with a maximum typically when the edge has a direction  $\alpha$ . Thus, the amplitude of wavelet coefficients depends on the local orientation of the image structures.

Theorem 5.11 proves that the translation-invariant wavelet family is a frame if there exists  $B \geq A > 0$  such that the generators  $\phi_n(x) = 2^{-j}\psi_{2^j}^\alpha(x)$  have Fourier transforms  $\hat{\phi}_n(\omega) = \hat{\psi}^\alpha(\omega)$ , which satisfy

$$\forall \omega = (\omega_1, \omega_2) \in \mathbb{R}^2 - \{(0, 0)\}, \quad A \leq \sum_{\alpha \in \Theta} \sum_{j=-\infty}^{+\infty} |\hat{\psi}^\alpha(2^j \omega)|^2 \leq B. \quad (5.101)$$

It results from Theorem 5.11 that there exists a dual family of reconstructing wavelets  $\{\tilde{\psi}^\alpha\}_{\alpha \in \Theta}$  that have Fourier transforms that satisfy

$$\sum_{j=-\infty}^{+\infty} \sum_{\alpha \in \Theta} \widehat{\tilde{\psi}}^\alpha(2^j \omega) \widehat{\psi}^{\alpha*}(2^j \omega) = 1, \quad (5.102)$$

which yields

$$f(x) = \sum_{j=-\infty}^{+\infty} \frac{1}{2^{2j}} \sum_{\alpha \in \Theta} Wf(\cdot, 2^j, \alpha) \star \tilde{\psi}_{2^j}^\alpha(x). \quad (5.103)$$

Examples of directional wavelets obtained by rotating a single mother wavelet are constructed with Gabor functions and steerable derivatives.

### Gabor Wavelets

In the cat's visual cortex, Hubel and Wiesel [306] discovered a class of cells, called simple cells, having a response that depends on the frequency and direction of the visual stimuli. Numerous physiological experiments [401] have shown that these cells can be modeled as linear filters with impulse responses that have been measured at different locations of the visual cortex. Daugmann [200] showed that these impulse responses can be approximated by *Gabor wavelets*, obtained with a Gaussian window  $g(x_1, x_2) = (2\pi)^{-1} e^{-(x_1^2 + x_2^2)/2}$  multiplied by a sinusoidal wave:

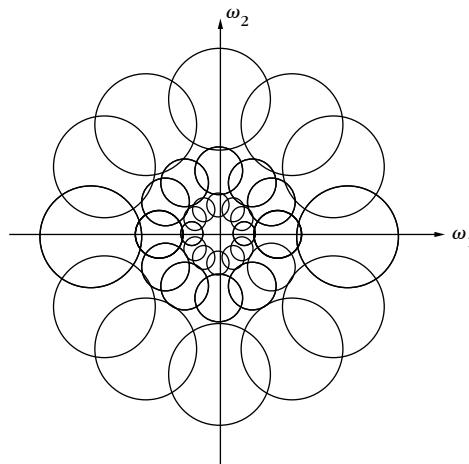
$$\psi^\alpha(x_1, x_2) = g(x_1, x_2) \exp[-i\eta(-x_1 \sin \alpha + x_2 \cos \alpha)]. \quad (5.104)$$

These findings suggest the existence of some sort of wavelet transform in the visual cortex, combined with subsequent nonlinearities [403]. The "physiological" wavelets have a frequency resolution on the order of 1 to 1.5 octaves, and are thus similar to dyadic wavelets.

The Fourier transform of  $g(x_1, x_2)$  is  $\hat{g}(\omega_1, \omega_2) = e^{-(\omega_1^2 + \omega_2^2)/2}$ . It results from (5.104) that

$$\hat{\psi}_{2^j}^\alpha(\omega_1, \omega_2) = \sqrt{2^j} \hat{g}(2^j \omega_1 + \eta \sin \alpha, 2^j \omega_2 - \eta \cos \alpha).$$

In the Fourier plane, the energy of this Gabor wavelet is mostly concentrated around  $(-2^{-j}\eta \sin \alpha, 2^{-j}\eta \cos \alpha)$ , in a neighborhood proportional to  $2^{-j}$ .



**FIGURE 5.8**

Each circle represents the frequency domain in a direction  $\alpha + \pi/2$  where the amplitude of a Gabor wavelet Fourier transform  $|\hat{\psi}_{2^j}^\alpha(\omega)|$  is large. It is proportional to  $2^{-j}$  and its position rotates with  $\alpha$ .

The direction is chosen to be uniform  $\alpha = l\pi/K$  for  $-K < l \leq K$ . Figure 5.8 shows a cover of the frequency plane by dyadic Gabor wavelets 5.105 with  $K = 6$ . If  $K \geq 4$  and  $\eta$  is of the order of 1 then 5.101 is satisfied with stable bounds. Since images  $f(x)$  are real,  $\hat{f}(-\omega) = \hat{f}^*(\omega)$  and  $f$  can be reconstructed by covering only half of the frequency plane, with  $-K < l \leq 0$ . This is a two-dimensional equivalent of the one-dimensional analytic wavelet transform, studied in Section 4.3.2, with wavelets having a Fourier transform support restricted to positive frequencies. For texture analysis, Gabor wavelets provide information on the local image frequencies.

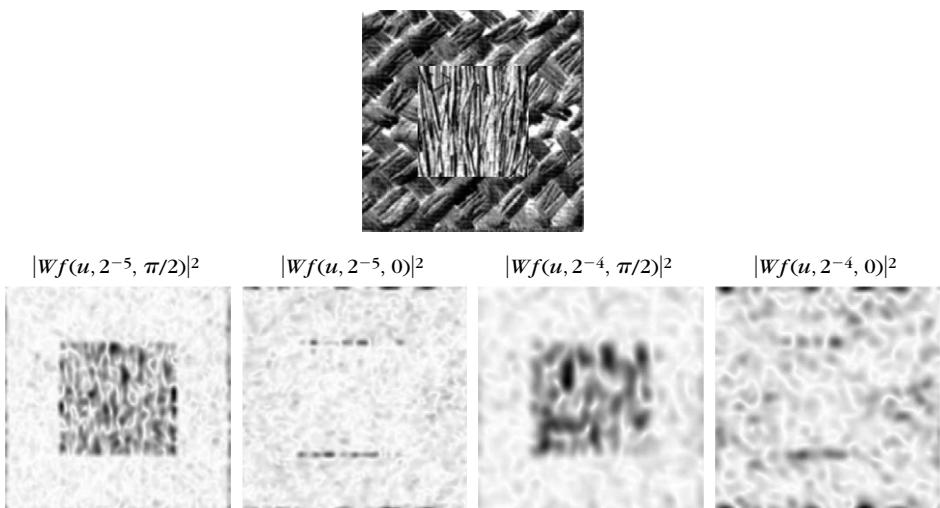
### Texture Discrimination

Despite many attempts, there are no appropriate mathematical models for “homogeneous image textures.” The notion of texture homogeneity is still defined with respect to our visual perception. A texture is said to be homogeneous if it is preattentively perceived as being homogeneous by a human observer.

The texton theory of Julesz [322] was a first important step in understanding the different parameters that influence the perception of textures. The direction of texture elements and their frequency content seem to be important clues for discrimination. This motivated early researchers to study the repartition of texture energy in the Fourier domain [92]. For segmentation purposes it is necessary to localize texture measurements over neighborhoods of varying sizes. Thus, the Fourier transform was replaced by localized energy measurements at the output of filter banks that compute a wavelet transform [315, 338, 402, 467]. Besides the algorithmic efficiency of this approach, this model is partly supported by physiological studies of the visual cortex.

Since  $Wf(u, 2^j, \alpha) = f \star \bar{\psi}_{2^j}^\alpha(u)$ , Gabor wavelet coefficients measure the energy of  $f$  in a spatial neighborhood of  $u$  of size  $2^j$ , and in a frequency neighborhood of  $(-2^{-j}\eta \sin \alpha, 2^{-j}\eta \cos \alpha)$  of size  $2^{-j}$ , where the support of  $\hat{\psi}_{2^j}^\alpha(\omega)$  is located, illustrated in Figure 5.8. Varying the scale  $2^j$  and the angle  $\alpha$  modifies the frequency channel [119]. The wavelet transform energy  $|Wf(u, 2^j, \alpha)|^2$  is large when the angle  $\alpha$  and scale  $2^j$  match the direction and scale of high-energy texture components in the neighborhood of  $u$ . Thus, the amplitude of  $|Wf(u, 2^j, \alpha)|^2$  can be used to discriminate textures. Figure 5.9 shows the dyadic wavelet transform of two textures, computed along horizontal and vertical directions, at the scales  $2^{-4}$  and  $2^{-5}$  (the image support is normalized to  $[0, 1]^2$ ). The central texture is regular vertically and has more energy along horizontal high frequencies than the peripheral texture. These two textures are therefore discriminated by the wavelet of angle  $\alpha = \pi/2$ , whereas the other wavelet with  $\alpha = 0$  produces similar responses for both textures.

For segmentation, one must design an algorithm that aggregates the wavelet responses at all scales and directions in order to find the boundaries of homogeneous textured regions. Both clustering procedures and detection of sharp transitions over wavelet energy measurements have been used to segment the image [315, 402, 467]. These algorithms work well experimentally but rely on ad hoc parameter settings.



**FIGURE 5.9**

Directional Gabor wavelet transform  $|Wf(u, 2^j, \alpha)|^2$  of a texture patch, at the scales  $2^j = 2^{-4}, 2^{-5}$ , along two directions  $\alpha = 0, \pi/2$ . The darker the pixel, the larger the wavelet coefficient amplitude.

### Steerable Wavelets

Steerable directional wavelets along any angle  $\alpha$  can be written as a linear expansion of few mother wavelets [441]. For example, a steerable wavelet in the direction  $\alpha$  can be defined as the partial derivative of order  $p$  of a window  $\mathbf{b}(x)$  in the direction of the vector  $\vec{n} = (-\sin \alpha, \cos \alpha)$ :

$$\psi^\alpha(x) = \frac{\partial^p \theta(x)}{\partial \vec{n}^p} = \left( -\sin \alpha \frac{\partial}{\partial x_1} + \cos \alpha \frac{\partial}{\partial x_2} \right)^p \theta(x). \quad (5.105)$$

Let  $R_\alpha$  be the planar rotation by an angle  $\alpha$ . If the window is invariant under rotations  $\theta(x) = \theta(R_\alpha x)$ , then these wavelets are generated by the rotation of a single mother wavelet:  $\psi^\alpha(x) = \psi(R_\alpha x)$  with  $\psi = \partial^p \theta / \partial x_2^p$ .

Furthermore, the expansion of the derivatives in (5.105) proves that each  $\psi^\alpha$  can be expressed as a linear combination of  $p+1$  partial derivatives

$$\psi^\alpha(x) = \sum_{i=0}^p a_i(\alpha) \rho^i(x), \quad \text{where } a_i(\alpha) = \binom{p}{i} (-\sin \alpha)^i (\cos \alpha)^{p-i}, \quad (5.106)$$

with

$$\forall 0 \leq i \leq p, \quad \rho^i(x) = \frac{\partial^p \theta(x)}{\partial x_1^i \partial x_2^{p-i}}.$$

The waveforms  $\rho^i(x)$  can also be considered as wavelets functions with vanishing moments. It results from 5.106 that the directional wavelet transform at any angle  $\alpha$  can be calculated from  $p+1$  convolutions of  $f$  with the  $\rho^i$  dilated:

$$Wf(u, 2^j, \alpha) = \sum_{i=0}^p a_i(\alpha) (f \star \tilde{\rho}_{2^j}^i)(u) \quad \text{for } \tilde{\rho}_{2^j}^i(x) = 2^{-j} \rho^i(-2^{-j}x).$$

Exercise 5.22 gives conditions on  $\theta$  so that for a set  $\Theta$  of  $p+1$  angles  $\alpha = k\pi/(p+1)$  with  $0 \leq k < p$  the resulting oriented wavelets  $\psi^\alpha$  define a family of dyadic wavelets that satisfy 5.101. Section 6.3 uses such directional wavelets, with  $p=1$ , to detect multiscale edges in images.

### Discretization of the Translation

A translation-invariant wavelet transforms  $Wf(u, 2^j, \alpha)$  for all scales  $2^j$ , and angle  $\alpha$  requires a large amount of memory. To reduce computation and memory storage, the translation parameter is discretized. In the one-dimensional case a frame is obtained by uniformly sampling the translation parameter  $u$  with intervals  $u_0 2^j n$  with  $n = (n_1, n_2) \in \mathbb{Z}^2$ , proportional to the scale  $2^j$ . The discretized wavelet derived from the translation-invariant wavelet family (5.100) is

$$\{\psi_{2^j}^\alpha(x - 2^j u_0 n)\}_{n \in \mathbb{Z}^2, j \in \mathbb{Z}, \alpha \in \Theta} \quad \text{with } \psi_{2^j}^\alpha(x) = 2^{-j} \psi^\alpha(2^{-j}x). \quad (5.107)$$

Necessary and sufficient conditions similar to Theorems 5.15 and 5.16 can be established to guarantee that such a wavelet family defines a frame of  $\mathbf{L}^2(\mathbb{R}^2)$ .

To decompose images in such wavelet frames with a fast filter bank, directional wavelets can be synthesized as a product of discrete filters. The steerable pyramid of Simoncelli et al. [441] decomposes images in such a directional wavelet frame, with a cascade of convolutions with low-pass filters and directional band-pass filters. The filter coefficients are optimized to yield wavelets that satisfy approximately the steerability condition 5.106 and produce a tight frame. The sampling interval is  $u_0 = 1/2$ .

Figure 5.10 shows an example of decomposition on such a steerable pyramid with  $K = 4$  directions. For discrete images of  $N$  pixels, the finest scale is  $2^j = 2N^{-1}$ . Since  $u_0 = 1/2$ , wavelet coefficients at the finest scale define an image of  $N$  pixels for each direction. The wavelet image size then decreases as the scale  $2^j$  increases. The total number of wavelet coefficients is  $4KN/3$  and the tight frame factor is  $4K/3$  [441]. Steerable wavelet frames are used to remove noise with wavelet thresholding estimators [404] and for texture analysis and synthesis [438].

Chapter 9 explains that sparse image representation can be obtained by keeping large-amplitude coefficients above a threshold. Large-amplitude wavelet coefficients appear where the image has a sharp transition, when the wavelet oscillates in a direction approximately perpendicular to the direction of the edge. However, even when directions are not perfectly aligned, wavelet coefficients remain nonnegligible in the neighborhood of edges. Thus, the number of large-amplitude wavelet coefficients is typically proportional to the length of edges in images. Reducing the number of large coefficients requires using waveforms that are more sensitive to direction properties, as shown in the next section.

### 5.5.2 Curvelet Frames

Curvelet frames were introduced by Candès and Donoho [134] to construct sparse representation for images including edges that are geometrically regular. Similar to directional wavelets, curvelet frames are obtained by rotating, dilating, and translating elementary waveforms. However, curvelets have a highly elongated support obtained with a parabolic scaling using different scaling factors along the curvelet width and length. These anisotropic waveforms have a much better direction sensitivity than directional wavelets. Section 9.3.2 studies applications to sparse approximations of geometrically regular images.

#### Dyadic Curvelet Transform

A curvelet is function  $c(x)$  having vanishing moments along the horizontal direction like a wavelet. However, as opposed to wavelets, dilated curvelets are obtained with a parabolic scaling law that produces highly elongated waveforms at fine scales:

$$c_{2^j}(x_1, x_2) \approx 2^{-3j/4} c(2^{-j/2}x_1, 2^{-j}x_2). \quad (5.108)$$

They have a *width* proportional to their *length*<sup>2</sup>. Dilated curvelets are then rotated  $c_{2^j}^\alpha = c_{2^j}(R_\alpha x)$ , where  $R_\alpha$  is the planar rotation of angle  $\alpha$ , and translated like wavelets:

**FIGURE 5.10**

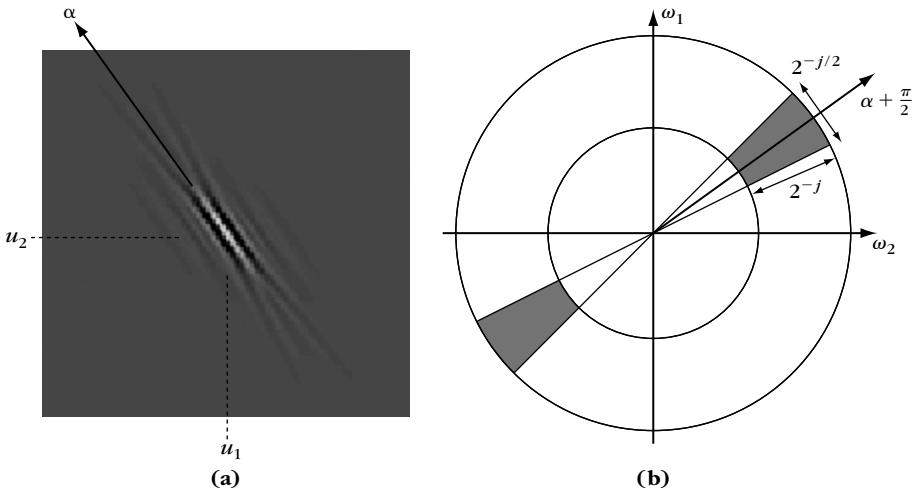
Decomposition of an image in a frame of steerable directional wavelets [441] along four directions:  $\alpha = 0, \pi/4, \pi/2, 3\pi/4$ , at two consecutive scales,  $2^j$  and  $2^j + 1$ . Black, gray, and white pixels correspond respectively to wavelet coefficients of negative, zero, and positive values.

$c_{2^j,u}^\alpha = c_{2^j}^\alpha(x - u)$ . The resulting translation-invariant dyadic curvelet transform of  $f \in \mathbf{L}^2(\mathbb{R}^2)$  is defined by

$$Cf(u, 2^j, \alpha) = \langle f, c_{2^j,u}^\alpha \rangle = f \star \bar{c}_{2^j}^\alpha(u) \quad \text{with} \quad \bar{c}_{2^j}^\alpha(x) = c_{2^j}^\alpha(-x).$$

To obtain a tight frame, the Fourier transform of a curvelet at a scale  $2^j$  is defined by

$$\hat{c}_{2^j}(\omega) = 2^{3j/4} \hat{\psi}(2^j r) \hat{\phi}\left(\frac{2\theta}{2^{|j/2|} \pi}\right), \quad \text{with} \quad \omega = r(\cos \theta, \sin \theta), \quad (5.109)$$

**FIGURE 5.11**

(a) Example of curvelet  $c_{2^j,u}^\alpha(x)$ . (b) The frequency support of  $\hat{c}_{2^j,u}^\alpha(\omega)$  is a wedge obtained as a product of a radial window with an angular window.

where  $\hat{\psi}$  is the Fourier transform of a one-dimensional wavelet and  $\hat{\phi}$  is a one-dimensional angular window that localizes the frequency support of  $\hat{c}_{2^j}$  in a polar parabolic wedge, illustrated in Figure 5.11. The wavelet  $\hat{\psi}$  is chosen to have a compact support in  $[1/2, 2]$  and satisfies the dyadic frequency covering:

$$\forall r \in \mathbb{R}^*, \quad \sum_{j=-\infty}^{+\infty} |\hat{\psi}(2^j r)|^2 = 1. \quad (5.110)$$

One may, for example, choose a Meyer wavelet as defined in (7.82). The angular window  $\hat{\phi}$  is chosen to be supported in  $[-1, 1]$  and satisfies

$$\forall u, \quad \sum_{k=-\infty}^{+\infty} |\hat{\phi}(u-k)|^2 = 1. \quad (5.111)$$

As a result of these two properties, one can verify that for uniformly distributed angles,

$$\Theta_j = \{\alpha = k\pi 2^{\lfloor j/2 \rfloor - 1} \quad \text{for} \quad 0 \leq k < 2^{-\lfloor j/2 \rfloor + 2}\}$$

curvelets cover the frequency plane

$$\forall \omega \in \mathbb{R}^2 - \{0\}, \quad \sum_{j \in \mathbb{Z}} \sum_{\alpha \in \Theta_j} 2^{-3j/2} |\hat{c}_{2^j}^\alpha(\omega)|^2 = 1. \quad (5.112)$$

Real valued curvelets are obtained with a symmetric version of 5.109:  $\hat{c}_{2j}(\omega) + \hat{c}_{2j}(-\omega)$ . Applying Theorem 5.11 proves that a translation-invariant dyadic curvelet dictionary  $\{c_{2j,u}^\alpha\}_{\alpha \in \Theta_j, j \in \mathbb{Z}, u \in \mathbb{R}^2}$  is a dyadic translation-invariant tight frame that defines a complete and stable signal representation [142].

**Theorem 5.23:** *Candès, Donoho.* For any  $f \in \mathbf{L}^2(\mathbb{R}^2)$ ,

$$\|f\|^2 = \sum_{j \in \mathbb{Z}} 2^{-3j/2} \sum_{\alpha \in \Theta_j} \|Cf(\cdot, 2^j, \alpha)\|^2,$$

and

$$f(x) = \sum_{j \in \mathbb{Z}} 2^{-3j/2} \sum_{\alpha \in \Theta_j} Cf(\cdot, 2^j, \alpha) * c_{2j}^\alpha(x).$$

### Curvelet Properties

Since  $\hat{c}_{2j}(\omega)$  is a smooth function with a support included in a rectangle of size proportional to  $2^{-j/2} \times 2^{-j}$ , the spatial curvelet  $c_{2j}(x)$  is a regular function with a fast decay outside a rectangle of size  $2^{j/2} \times 2^j$ . The rotated and translated curvelet  $c_{2j,u}^\alpha$  is supported around the point  $u$  in an elongated rectangle along the direction  $\alpha$ ; its shape has a parabolic ratio  $width = length^2$ , as shown in Figure 5.11.

Since the Fourier transform  $\hat{c}_{2j}(\omega_1, \omega_2)$  is zero in the neighborhood of the vertical axis  $\omega_1 = 0$ ,  $c_{2j}(x_1, x_2)$  has an infinite number of vanishing moments in the horizontal direction

$$\forall \omega_1, \quad \frac{\partial^q \hat{c}_j}{\partial \omega_1^q}(0, \omega_1) = 0 \quad \Rightarrow \quad \forall q \geq 0, \quad \forall x_2, \quad \int c_{2j}(x_1, x_2) x_1^q dx_1 = 0.$$

A rotated curvelet  $c_{2j,u}^\alpha$  has vanishing moments in the direction  $\alpha + \pi/2$ , and therefore oscillates in the direction  $\alpha + \pi/2$ , whereas its support is elongated in the direction  $\alpha$ .

### Discretization of Translation

Curvelet tight frames are constructed by sampling the translation parameter  $u$  [134]. These tight frames provide sparse representations of signals including regular geometric structures.

The curvelet sampling grid depends on the scale  $2^j$  and on the angle  $\alpha$ . Sampling intervals are proportional to the curvelet width  $2^j$  in the direction  $\alpha + \pi/2$  and to its length  $2^{j/2}$  in the direction  $\alpha$ :

$$\forall m = (m_1, m_2) \in \mathbb{Z}^2, \quad u_m^{(j,\alpha)} = R_\alpha(2^{j/2}m_1, 2^j m_2). \quad (5.113)$$

Figure 5.12 illustrates this sampling grid. The resulting dictionary of translated curvelets is

$$\left\{ c_{j,m}^\alpha(x) = c_{2j}^\alpha(x - u_m^{(j,\alpha)}) \right\}_{j \in \mathbb{Z}, \alpha \in \Theta_j, m \in \mathbb{Z}^2}.$$

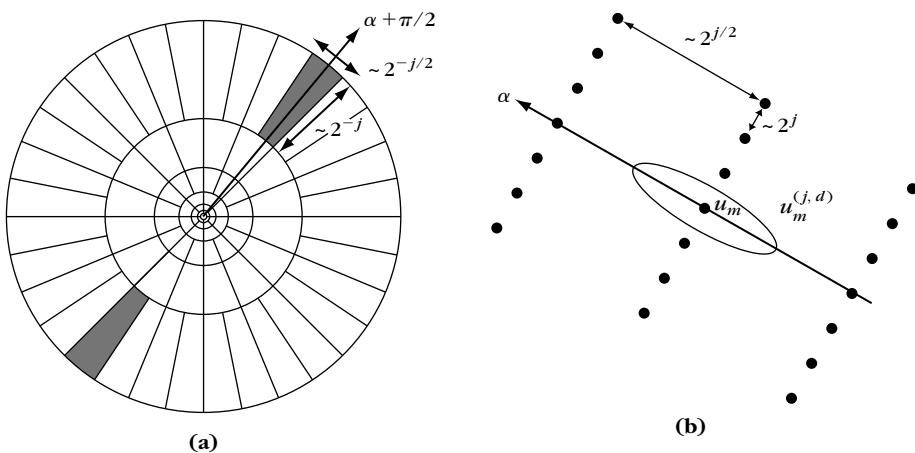


FIGURE 5.12

(a) Curvelet polar tiling of the frequency plane with parabolic wedges. (b) Curvelet spatial sampling grid  $u_m^{(j, \alpha)}$  at a scale  $2^j$  and direction  $\alpha$ .

Theorem 5.24 proves that this curvelet family is a tight frame of  $\mathbf{L}^2(\mathbb{R}^2)$ . The proof is not given here, but can be found in [142].

**Theorem 5.24:** Candès, Donoho. For any  $f \in \mathbf{L}^2(\mathbb{R}^2)$ ,

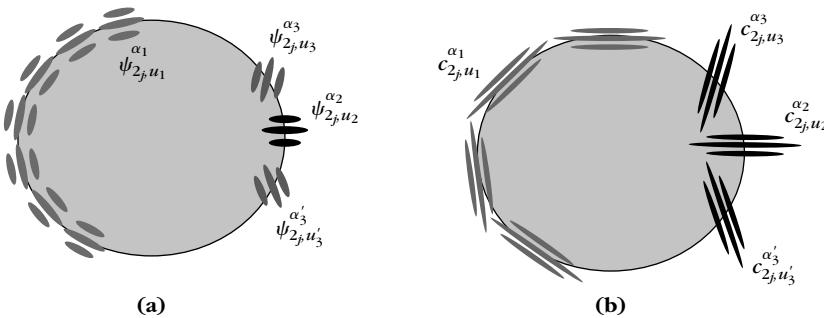
$$\|f\|^2 = \sum_{j \in \mathbb{Z}} \sum_{\alpha \in \Theta_j} \sum_{m \in \mathbb{Z}^2} |\langle f, c_{j,m}^\alpha \rangle|^2 \quad (5.114)$$

and

$$f(x) = \sum_{j \in \mathbb{Z}} \sum_{\alpha \in \Theta_j} \sum_{m \in \mathbb{Z}^2} \langle f, c_{j,m}^\alpha \rangle c_{j,m}^\alpha(x).$$

### Wavelet versus Curvelet Coefficients

In the neighborhood of an edge having a tangent in a direction  $\theta$ , large-amplitude coefficients are created by curvelets and wavelets of direction  $\alpha = \theta$ , which have their vanishing moment in the direction  $\theta + \pi/2$ . These curvelets have a support elongated in the edge direction  $\theta$ , as illustrated in Figure 5.13. In this direction, the sampling grid of a curvelet frame has an interval  $2^{j/2}$ , which is much larger than the sampling interval  $2^j$  of a wavelet frame. Thus, an edge is covered by fewer curvelets than wavelets having a direction equal to the edge direction. If the angle  $\alpha$  of the curvelet deviates from  $\theta$ , then curvelet coefficients decay quickly because of the narrow frequency localizaton of curvelets. This gives a high-directional selectivity to curvelets.

**FIGURE 5.13**

**(a)** Directional wavelets: a regular edge creates more large-amplitude wavelet coefficients than curvelet coefficients. **(b)** Curvelet coefficients have a large amplitude when their support is aligned with the edge direction, but there are few such curvelets.

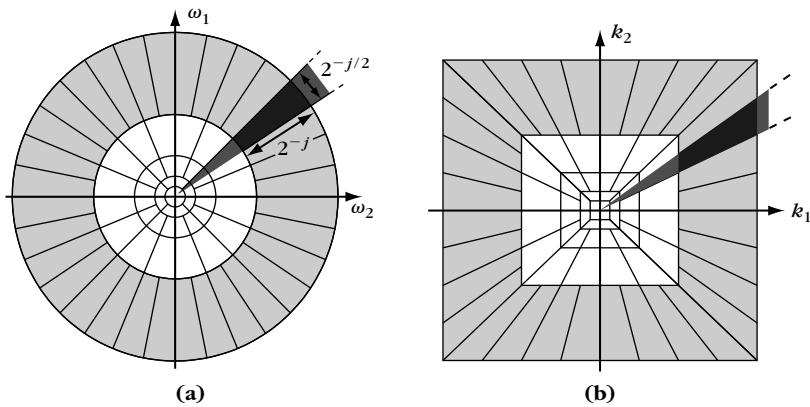
Even though wavelet coefficients vanish when  $\alpha = \theta + \pi/2$ , they have a smaller directional selectivity than curvelets, and wavelet coefficients' amplitudes decay more slowly as  $\alpha$  deviates from  $\theta$ . Indeed, the frequency support of wavelets is much wider and their spatial support is nearly isotropic. As a consequence, an edge produces large-amplitude wavelet coefficients in several directions. This is illustrated by Figure 5.10, where Lena's shoulder edge creates large coefficients in three directions, and small coefficients only when the wavelet and the edge directions are aligned.

As a result, edges and image structures with some directional regularity create fewer large-amplitude curvelet coefficients than wavelet coefficients. The theorem derived from Section 9.3.3 proves that for images having regular edges, curvelet tight frames are asymptotically more efficient than wavelet bases when building sparse representations.

### **Fast Curvelet Decomposition Algorithm**

To compute the curvelet transform of a discrete image  $f[n_1, n_2]$  uniformly sampled over  $N$  pixels, one must take into account the discretization grid, which imposes constraints on curvelet angles. The fast curvelet transform [140] replaces the polar tiling of the Fourier domain by a recto-polar tiling, illustrated in Figure 5.14(b). The directions  $\alpha$  are uniformly discretized so that the slopes of the wedges containing the support of the curvelets are uniformly distributed in each of the north, south, west, and east Fourier quadrants. Each wedge is the support of the two-dimensional DFT  $\hat{c}_j^\alpha[k_1, k_2]$  of a discrete curvelet  $c_j^\alpha[n_1, n_2]$ . The curvelet translation parameters are not chosen according to 5.113, but remain on a subgrid of the original image sampling grid. At a scale  $2^j$ , there is one sampling grid  $(2^{\lfloor j/2 \rfloor} m_1, 2^j m_2)$  for curvelets in the east and west quadrants. In these directions, curvelet coefficients are

$$\langle f[n_1, n_2], c_j^\alpha[n_1 - 2^{\lfloor j/2 \rfloor} m_1, n_2 - 2^j m_2] \rangle = f \star \bar{c}_j^\alpha[2^{\lfloor j/2 \rfloor} m_1, 2^j m_2] \quad (5.115)$$

**FIGURE 5.14**

**(a)** Curvelet frequency plane tiling. The dark gray area is a wedge obtained as the product of a radial window and an angular window. **(b)** Discrete curvelet frequency tiling. The radial and angular windows define trapezoidal wedges as shown in dark gray.

with  $\bar{c}_j^\alpha[n_1, n_2] = c_j^\alpha[-n_1, -n_2]$ . For curvelets in the north and south quadrants the translation grid is  $(2^j m_1, 2^{\lfloor j/2 \rfloor} m_2)$ , which corresponds to curvelet coefficients

$$\langle f[n_1, n_2], c_j^\alpha[n_1 - 2^j m_1, n_2 - 2^{\lfloor j/2 \rfloor} m_2] \rangle = f \star \bar{c}_j^\alpha[2^j m_1, 2^{\lfloor j/2 \rfloor} m_2]. \quad (5.116)$$

The discrete curvelet transform computes the curvelet filtering and sampling with a two-dimensional FFT. The two-dimensional discrete Fourier transforms of  $f[n]$  and  $\bar{c}_j^\alpha[n]$  are  $\hat{f}[k]$  and  $\hat{c}_j^\alpha[k]$ . The algorithm proceeds as:

- Computation of the two-dimensional DFT  $\hat{f}[k]$  of  $f[n]$ .
- For each  $j$  and the corresponding  $2^{-\lfloor j/2 \rfloor + 2}$  angles  $\alpha$ , calculation of  $\hat{f}[k] \hat{c}_j^\alpha[-k]$ .
- Computation of the inverse Fourier transform of  $\hat{f}[k] \hat{c}_j^\alpha[-k]$  on the smallest possible warped frequency rectangle including the wedge support of  $\hat{c}_j^\alpha[-k]$ .

The critical step is the last inverse Fourier transform. A computationally more expensive one would compute  $f \star \bar{c}_j^\alpha[n]$  for all  $n = (n_1, n_2)$  and then subsample this convolution along the grids (5.115) and (5.116).

Instead, the subsampled curvelet coefficients are calculated directly by restricting the FFT to a bounding box that contains the support of  $\hat{c}_j^\alpha[-k]$ . A horizontal or vertical warping maps this bounding box to an elongated rectangular frequency box on which the inverse FFT is calculated. One can verify that the resulting coefficients correspond to the subsampled curvelet coefficients. The overall complexity of the algorithm is then  $O(N \log_2(N))$ , as detailed in [140]. The tight frame redundancy bound obtained with this discrete algorithm is  $A \approx 5$ .

An orthogonal curvelet type transform has been developed by Do and Vetterli [212]. The resulting contourlets are not redundant but do not have the appropriate time and frequency localization needed to obtain asymptotic approximation results similar to curvelets.

## 5.6 EXERCISES

- 5.1** <sup>1</sup> Prove that if  $K \in \mathbb{Z} - \{0\}$ , then  $\{\phi_p[n] = \exp(i2\pi pn/(KN))\}_{0 \leq p < KN}$  is a tight frame of  $\mathbb{C}^N$ . Compute the frame bound.
- 5.2** <sup>1</sup> Prove that if  $K \in \mathbb{R} - \{0\}$ , then  $\{\phi_p(t) = \exp(i2\pi pnt/K)\}_{p \in \mathbb{Z}}$  is a tight frame of  $\mathbf{L}^2[0, 1]$ . Compute the frame bound.
- 5.3** <sup>2</sup> Prove that a finite set of  $N$  vectors  $\{\phi_n\}_{1 \leq n \leq N}$  is always a frame of the space  $\mathbf{V}$  generated by linear combinations of these vectors.
- 5.4** <sup>1</sup> If  $U_1$  and  $U_2$  are two operators from  $\mathbb{C}^N$  to  $\mathbb{C}^N$ , prove that the trace (sum of diagonal values) satisfies  $\text{tr}(U_1 U_2) = \text{tr}(U_2 U_1)$ .
- 5.5** <sup>2</sup> Prove that the translation-invariant frame

$$\{\phi_p[n] = \delta[(n-p)\bmod N] - \delta[(n-p-1)\bmod N]\}_{0 \leq p < N}$$

is a translation-invariant frame of the space  $\mathbf{V} = \{f \in \mathbb{R}^N : \sum_{n=0}^{N-1} f[n] = 0\}$ . Compute the frame bounds. Is it a numerically stable frame when  $N$  is large?

- 5.6** <sup>2</sup> Construct a Riesz basis in  $\mathbb{C}^N$  with a lower frame bound  $A$  that tends to zero and an upper frame bound that tends to  $+\infty$  as  $N$  increases.
- 5.7** <sup>1</sup> If  $U$  is an operator from  $\mathbb{C}^N$  to  $\mathbb{C}^P$ , prove that  $\mathbf{Null}U^*$  is the orthogonal complement of  $\mathbf{Im}U$  in  $\mathbb{C}^P$ .
- 5.8** <sup>3</sup> Prove Theorem 5.12.
- 5.9** <sup>1</sup> Let  $\hat{g} = \mathbf{1}_{[-\omega_0, \omega_0]}$ . Prove that  $\{g(t - p2\pi/\omega_0) \exp(ik\omega_0 t)\}_{(k,p) \in \mathbb{Z}^2}$  is an orthonormal basis of  $\mathbf{L}^2(\mathbb{R})$ .
- 5.10** <sup>2</sup> Let  $g_{m,k}[n] = g[n - mM] \exp(i2\pi kn/K)$ , where  $g[n]$  is a window with a support included in  $[-K/2, K/2 - 1]$ .
  - (a) Prove that  $\sum_{n=mM-M/2}^{mM+M/2-1} |g[n - mM]|^2 |f[n]|^2 = K^{-1} \sum_{k=0}^{K-1} |\langle f, g_{m,k} \rangle|^2$ .
  - (b) Prove Theorem 5.18 with arguments similar to Theorem 5.17.
- 5.11** <sup>2</sup> Compute the trigonometric polynomials  $\widehat{h}(\omega)$  and  $\widehat{\tilde{g}}(\omega)$  of minimum degree that satisfy (5.63) for the spline filters (5.67, 5.68) with  $m = 2$ . Compute numerically the graph of  $\widehat{\phi}$  and  $\tilde{\phi}$ . Are they finite-energy functions?

- 5.12** <sup>1</sup> Compute a cubic spline dyadic wavelet with two vanishing moments using the filter  $h$  defined by (5.67) for  $m = 3$ , with a filter  $g$  having three nonzero coefficients. Compute in WAVELET the dyadic wavelet transform of the Lady signal with this new wavelet. Calculate  $\tilde{g}[n]$  if  $\tilde{h}[n] = h[n]$ .
- 5.13** <sup>2</sup> Prove the tight-frame energy conservation (4.29) of a discrete windowed Fourier transform. Derive (4.28) from general tight-frame properties. Compute the resulting discrete windowed Fourier transform reproducing kernel.
- 5.14** <sup>1</sup> Let  $\{g(t - n\beta) \exp(ik\eta t)\}_{(n,k) \in \mathbb{Z}^2}$  be a windowed Fourier frame defined by  $g(t) = \pi^{-1/4} \exp(-t^2/2)$  with  $\beta = \eta$  and  $\beta \eta < 2\pi$ . With the conjugate gradient algorithm of Theorem 5.8, compute numerically the window  $\tilde{g}(t)$  that generates the dual frame, for the values of  $\beta \eta$  in Table 5.3. Compare  $\tilde{g}$  with  $g$  and explain why they are progressively more different as  $\beta \eta$  tends to  $2\pi$ .
- 5.15** <sup>2</sup> *Sigma-Delta converter:* A signal  $f(t)$  is sampled and quantized. We suppose that  $\hat{f}$  has a support in  $[-\pi/T, \pi/T]$ .
  - (a) Let  $x[n] = f(nT/K)$ . Show that if  $\omega \in [-\pi, \pi]$ , then  $\hat{x}(\omega) \neq 0$  only if  $\omega \in [-\pi/K, \pi/K]$ .
  - (b) Let  $\tilde{x}[n] = Q(x[n])$  be the quantized samples. We now consider  $x[n]$  as a random vector, and we model the error  $x[n] - \tilde{x}[n] = W[n]$  as a white noise process of variance  $\sigma^2$ . Find the filter  $h[n]$  that minimizes

$$\varepsilon = E\{\|\tilde{x} \star h - x\|^2\},$$

and compute this minimum as a function of  $\sigma^2$  and  $K$ .

- (c) Let  $\hat{h}_p(\omega) = (1 - e^{-i\omega})^{-p}$  be the transfer function of a discrete integration of order  $p$ . We quantize  $\tilde{x}[n] = Q(x \star h_p[n])$ . Find the filter  $h[n]$  that minimizes  $\varepsilon = E\{\|\tilde{x} \star h - x\|^2\}$ , and compute this minimum as a function of  $\sigma^2$ ,  $K$ , and  $p$ . For a fixed oversampling factor  $K$ , how can we reduce this error?
- 5.16** <sup>3</sup> *Oversampled analog-to-digital conversion.* Let  $\phi_s(t) = s^{1/2} \sin(\pi t/s)/(\pi t)$ .
  - (a) Prove that the following family is a union of orthogonal bases:
$$\left\{ \phi_s \left( t - k \frac{s}{K} - ns \right) \right\}_{1 \leq k \leq K, n \in \mathbb{Z}}$$

Compute the tight-frame bound.

  - (b) Prove that the frame projector  $P$  defined in Proposition 5.9 is a discrete convolution. Compute its impulse response  $h[n]$ .
  - (c) Characterize the signals  $a[n]$  that belong to the image space  $\text{Im}\Phi$  of this frame.
  - (d) Let  $f(t)$  be a signal with a Fourier transform supported in  $[-\pi/s, \pi/s]$ . Prove that  $f \star \phi_s(ns) = s^{-1/2} f(ns)$ .

- (e) Let  $s_0 = s/K$ . For all  $n \in \mathbb{Z}$ , we measure the oversampled noisy signal  $Y[n] = f \star \phi_s(ns_0) + W[n]$  where  $W[n]$  is a Gaussian white noise of variance  $\sigma^2$ . With the frame projector  $P$ , compute the error  $E\{|PY[Kn] - s^{-1/2}f(ns)|^2\}$  and show that it decreases when  $K$  increases.

- 5.17** <sup>1</sup> Let  $\psi$  be a dyadic wavelet that satisfies (5.48). Let  $\ell^2(\mathbf{L}^2(\mathbb{R}))$  be the space of sequences  $\{g_j(u)\}_{j \in \mathbb{Z}}$  such that  $\sum_{j=-\infty}^{+\infty} \|g_j\|^2 < +\infty$ .

- (a) Verify that if  $f \in \mathbf{L}^2(\mathbb{R})$ , then  $\{Wf(u, 2^j)\}_{j \in \mathbb{Z}} \in \ell^2(\mathbf{L}^2(\mathbb{R}))$ . Next, let  $\tilde{\psi}$  be defined by

$$\widehat{\tilde{\psi}}(\omega) = \frac{\widehat{\psi}(\omega)}{\sum_{j=-\infty}^{+\infty} |\widehat{\psi}(2^j \omega)|^2},$$

and  $W^{-1}$  be the operator defined by

$$W^{-1}\{g_j(u)\}_{j \in \mathbb{Z}} = \sum_{j=-\infty}^{+\infty} \frac{1}{2^j} g_j \star \tilde{\psi}_{2^j}(t).$$

Prove that  $W^{-1}$  is the pseudo inverse of  $W$  in  $\ell^2(\mathbf{L}^2(\mathbb{R}))$ .

- (b) Verify that  $\tilde{\psi}$  has the same number of vanishing moments as  $\psi$ .  
(c) Let  $\mathbf{V}$  be the subspace of  $\ell^2(\mathbf{L}^2(\mathbb{R}))$  that regroups all the dyadic wavelet transforms of functions in  $\mathbf{L}^2(\mathbb{R})$ . Compute the orthogonal projection of  $\{g_j(u)\}_{j \in \mathbb{Z}}$  in  $\mathbf{V}$ .

- 5.18** <sup>1</sup> Prove that if there exist  $A > 0$  and  $B \geq 0$  such that

$$A(2 - |\widehat{h}(\omega)|^2) \leq |\widehat{g}(\omega)|^2 \leq B(2 - |\widehat{h}(\omega)|^2),$$

and if  $\phi$  defined in (5.59) belongs to  $\mathbf{L}^2(\mathbb{R})$ , then the wavelet  $\psi$  given by (5.60) satisfies the dyadic wavelet condition (5.55).

- 5.19** <sup>3</sup> *Zak transform.* The Zak transform associates to any  $f \in \mathbf{L}^2(\mathbb{R})$

$$Zf(u, \xi) = \sum_{l=-\infty}^{+\infty} e^{i2\pi l\xi} f(u-l).$$

- (a) Prove that it is a unitary operator from  $\mathbf{L}^2(\mathbb{R})$  to  $\mathbf{L}^2[0, 1]^2$ :

$$\int_{-\infty}^{+\infty} f(t) g^*(t) dt = \int_0^1 \int_0^1 Zf(u, \xi) Zg^*(u, \xi) du d\xi,$$

by verifying that for  $g = \mathbf{1}_{[0,1]}$ , it transforms the orthogonal basis  $\{g_{n,k}(t) = g(t-n) \exp(i2\pi kt)\}_{(n,k) \in \mathbb{Z}^2}$  of  $\mathbf{L}^2(\mathbb{R})$  into an orthonormal basis of  $\mathbf{L}^2[0, 1]^2$ .

- (b) Prove that the inverse Zak transform is defined by

$$\forall h \in \mathbf{L}^2[0, 1]^2, \quad Z^{-1}h(u) = \int_0^1 h(u, \xi) d\xi.$$

- (c) Prove that if  $g \in L^2(\mathbb{R})$  then  $\{g(t-n) \exp(i2\pi kt)\}_{(n,k) \in \mathbb{Z}^2}$  is a frame of  $L^2(\mathbb{R})$  if and only if there exist  $A > 0$  and  $B$  such that

$$\forall (u, \xi) \in [0, 1]^2, \quad A \leq |Zg(u, \xi)|^2 \leq B, \quad (5.117)$$

where  $A$  and  $B$  are the frame bounds.

- (d) Prove that if (5.117) holds, then the dual window  $\tilde{g}$  of the dual frame is defined by  $Z\tilde{g}(u, \xi) = 1/Zg^*(u, \xi)$ .

- 5.20** <sup>3</sup> Suppose that  $\hat{f}$  has a support in  $[-\pi/T, \pi/T]$ . Let  $\{f(t_n)\}_{n \in \mathbb{Z}}$  be irregular samples that satisfy (5.8). With the conjugate gradient Theorem 5.8, implement numerically a procedure that computes a uniform sampling  $\{f(nT)\}_{n \in \mathbb{Z}}$  (from which  $f$  can be recovered with the sampling Theorem 3.2). Analyze the convergence rate of the conjugate-gradient algorithm as a function of  $\delta$ . What happens if the condition (5.8) is not satisfied?
- 5.21** <sup>2</sup> Prove that if  $\psi(x_1, x_2)$  is a directional wavelet having  $p$  vanishing moments in a direction  $\alpha + \pi/2$  as defined in (5.99), then it is orthogonal to any two-dimensional polynomial of degree  $p - 1$ .
- 5.22** <sup>2</sup> Let  $\vec{n}_k = (\cos(2k\pi/K), \sin(2k\pi/K)) \in \mathbb{R}^2$ .
- (a) Prove that  $\{\vec{n}_k\}_{0 \leq k < K}$  is a tight frame of  $K$  vectors and that for any  $\omega = (\omega_1, \omega_2) \in \mathbb{R}^2$ , it satisfies  $\sum_{k=0}^{K-1} |\omega \cdot \vec{n}_k|^2 = K/2|\omega|^2$  where  $\omega \cdot \vec{n}$  is the inner product in  $\mathbb{R}^2$ .
  - (b) Let  $\psi^k = \partial\theta(x)/\partial\vec{n}_k$  be the derivative of  $\theta(x)$  in the direction of  $\vec{n}_k$  with  $x \in \mathbb{R}^2$ . If  $\theta(x)$  is rotationally invariant (not modified by a rotation of  $x$ ), then prove that the frame condition (5.101) is equivalent to

$$2A/K \leq \sum_{j=-\infty}^{+\infty} 2^{2j} |\omega|^2 |\hat{\theta}(2^j \omega)|^2 \leq 2B/K.$$

- 5.23** <sup>4</sup> Develop a texture classification algorithm with a two-dimensional Gabor wavelet transform using four oriented wavelets. The classification procedure can be based on “feature vectors” that provide local averages of the wavelet transform amplitude at several scales, along these four orientations [315, 338, 402, 467].

# Wavelet Zoom

# 6

A wavelet transform can focus on localized signal structures with a zooming procedure that progressively reduces the scale parameter. Singularities and irregular structures often carry essential information in a signal. For example, discontinuities in images may correspond to occlusion contours of objects in a scene. The wavelet transform amplitude across scales is related to the local signal regularity and Lipschitz exponents. Singularities and edges are detected from wavelet transform local maxima at multiple scales. These maxima define a geometric scale-space support from which signal and image approximations are recovered.

Nonisolated singularities appear in highly irregular signals such as multifractals. The wavelet transform takes advantage of multifractal self-similarities to compute the distribution of their singularities. This singularity spectrum characterizes multifractal properties. Throughout this chapter wavelets are real functions.

---

## 6.1 LIPSCHITZ REGULARITY

To characterize singular structures, it is necessary to precisely quantify the local regularity of a signal  $f(t)$ . Lipschitz exponents provide uniform regularity measurements over time intervals, but also at any point  $v$ . If  $f$  has a singularity at  $v$ , which means that it is not differentiable at  $v$ , then the Lipschitz exponent at  $v$  characterizes this singular behavior.

Section 6.1.1 relates the uniform Lipschitz regularity of  $f$  over  $\mathbb{R}$  to the asymptotic decay of the amplitude of its Fourier transform. This global regularity measurement is useless in analyzing the signal properties at particular locations. Section 6.1.3 studies zooming procedures that measure local Lipschitz exponents from the decay of the wavelet transform amplitude at fine scales.

### 6.1.1 Lipschitz Definition and Fourier Analysis

The Taylor formula relates the differentiability of a signal to local polynomial approximations. Suppose that  $f$  is  $m$  times differentiable in  $[v - h, v + h]$ . Let  $p_v$  be

the Taylor polynomial in the neighborhood of  $v$ :

$$p_v(t) = \sum_{k=0}^{m-1} \frac{f^{(k)}(v)}{k!} (t-v)^k. \quad (6.1)$$

The Taylor formula proves that the approximation error

$$\varepsilon_v(t) = f(t) - p_v(t)$$

satisfies

$$\forall t \in [v-h, v+h], \quad |\varepsilon_v(t)| \leq \frac{|t-v|^m}{m!} \sup_{u \in [v-h, v+h]} |f^m(u)|. \quad (6.2)$$

The  $m$ th-order differentiability of  $f$  in the neighborhood of  $v$  yields an upper bound on the error  $\varepsilon_v(t)$  when  $t$  tends to  $v$ . The Lipschitz regularity refines this upper bound with noninteger exponents. Lipschitz exponents are also called *Hölder* exponents in mathematics literature.

**Definition 6.1:** *Lipschitz.*

- A function  $f$  is pointwise Lipschitz  $\alpha \geq 0$  at  $v$ , if there exists  $K > 0$  and a polynomial  $p_v$  of degree  $m = \lfloor \alpha \rfloor$  such that

$$\forall t \in \mathbb{R}, \quad |f(t) - p_v(t)| \leq K |t - v|^\alpha. \quad (6.3)$$

- A function  $f$  is uniformly Lipschitz  $\alpha$  over  $[a, b]$  if it satisfies (6.3) for all  $v \in [a, b]$  with a constant  $K$  that is independent of  $v$ .
- The Lipschitz regularity of  $f$  at  $v$  or over  $[a, b]$  is the supremum of the  $\alpha$  such that  $f$  is Lipschitz  $\alpha$ .

At each  $v$  the polynomial  $p_v(t)$  is uniquely defined. If  $f$  is  $m = \lfloor \alpha \rfloor$  times continuously differentiable in a neighborhood of  $v$ , then  $p_v$  is the Taylor expansion of  $f$  at  $v$ . Pointwise Lipschitz exponents may vary arbitrarily from abscissa to abscissa. One can construct multifractal functions with nonisolated singularities, where  $f$  has a different Lipschitz regularity at each point. In contrast, uniform Lipschitz exponents provide a more global measurement of regularity, which applies to a whole interval. If  $f$  is uniformly Lipschitz  $\alpha > m$  in the neighborhood of  $v$ , then one can verify that  $f$  is necessarily  $m$  times continuously differentiable in this neighborhood.

If  $0 \leq \alpha < 1$ , then  $p_v(t) = f(v)$  and the Lipschitz condition (6.3) becomes

$$\forall t \in \mathbb{R}, \quad |f(t) - f(v)| \leq K |t - v|^\alpha.$$

A function that is bounded but discontinuous at  $v$  is Lipschitz 0 at  $v$ . If the Lipschitz regularity is  $\alpha < 1$  at  $v$ , then  $f$  is not differentiable at  $v$  and  $\alpha$  characterizes the singularity type.

### Fourier Condition

The uniform Lipschitz regularity of  $f$  over  $\mathbb{R}$  is related to the asymptotic decay of its Fourier transform. Theorem 6.1 can be interpreted as a generalization of Theorem 2.5.

**Theorem 6.1.** A function  $f$  is bounded and uniformly Lipschitz  $\alpha$  over  $\mathbb{R}$  if

$$\int_{-\infty}^{+\infty} |\hat{f}(\omega)| (1 + |\omega|^\alpha) d\omega < +\infty. \quad (6.4)$$

**Proof.** To prove that  $f$  is bounded, we use the inverse Fourier integral (2.8) and (6.4), which shows that

$$|f(t)| \leq \int_{-\infty}^{+\infty} |\hat{f}(\omega)| d\omega < +\infty.$$

Let us now verify the Lipschitz condition (6.3) when  $0 \leq \alpha \leq 1$ . In this case,  $p_v(t) = f(v)$  and the uniform Lipschitz regularity means that there exists  $K > 0$  such that for all  $(t, v) \in \mathbb{R}^2$

$$\frac{|f(t) - f(v)|}{|t - v|^\alpha} \leq K.$$

Since

$$\begin{aligned} f(t) &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} \hat{f}(\omega) \exp(i\omega t) d\omega, \\ \frac{|f(t) - f(v)|}{|t - v|^\alpha} &\leq \frac{1}{2\pi} \int_{-\infty}^{+\infty} |\hat{f}(\omega)| \frac{|\exp(i\omega t) - \exp(i\omega v)|}{|t - v|^\alpha} d\omega. \end{aligned} \quad (6.5)$$

For  $|t - v|^{-1} \leq |\omega|$ ,

$$\frac{|\exp(i\omega t) - \exp(i\omega v)|}{|t - v|^\alpha} \leq \frac{2}{|t - v|^\alpha} \leq 2|\omega|^\alpha.$$

For  $|t - v|^{-1} \geq |\omega|$ ,

$$\frac{|\exp(i\omega t) - \exp(i\omega v)|}{|t - v|^\alpha} \leq \frac{|\omega| |t - v|}{|t - v|^\alpha} \leq |\omega|^\alpha.$$

Cutting the integral (6.5) in two for  $|\omega| < |t - v|^{-1}$  and  $|\omega| \geq |t - v|^{-1}$  yields

$$\frac{|f(t) - f(v)|}{|t - v|^\alpha} \leq \frac{1}{2\pi} \int_{-\infty}^{+\infty} 2|\hat{f}(\omega)| |\omega|^\alpha d\omega = K.$$

If (6.4) is satisfied, then  $K < +\infty$  so  $f$  is uniformly Lipschitz  $\alpha$ .

Let us extend this result for  $m = \lfloor \alpha \rfloor > 0$ . We proved in (2.42) that (6.4) implies that  $f$  is  $m$  times continuously differentiable. One can verify that  $f$  is uniformly Lipschitz  $\alpha$  over  $\mathbb{R}$  if and only if  $f^{(m)}$  is uniformly Lipschitz  $\alpha - m$  over  $\mathbb{R}$ . The Fourier transform of  $f^{(m)}$  is  $(i\omega)^m \hat{f}(\omega)$ . Since  $0 \leq \alpha - m < 1$ , we can use our previous result, which proves that  $f^{(m)}$  is uniformly Lipschitz  $\alpha - m$ , and thus that  $f$  is uniformly Lipschitz  $\alpha$ . ■

The Fourier transform is a powerful tool for measuring the minimum global regularity of functions. However, it is not possible to analyze the regularity of  $f$  at a particular point  $v$  from the decay of  $|\hat{f}(\omega)|$  at high frequencies  $\omega$ . In contrast, since wavelets are well localized in time, the wavelet transform gives Lipschitz regularity over intervals *and* at points.

### 6.1.2 Wavelet Vanishing Moments

To measure the local regularity of a signal, it is not so important to use a wavelet with a narrow frequency support, but vanishing moments are crucial. If the wavelet has  $n$  vanishing moments, then we show that the wavelet transform can be interpreted as a multiscale differential operator of order  $n$ . This yields a first relation between the differentiability of  $f$  and its wavelet transform decay at fine scales.

#### *Polynomial Suppression*

The Lipschitz property (6.3) approximates  $f$  with a polynomial  $p_v$  in the neighborhood of  $v$ :

$$f(t) = p_v(t) + \varepsilon_v(t) \quad \text{with} \quad |\varepsilon_v(t)| \leq K |t - v|^\alpha. \quad (6.6)$$

A wavelet transform estimates the exponent  $\alpha$  by ignoring the polynomial  $p_v$ . For this purpose, we use a wavelet that has  $n > \alpha$  vanishing moments:

$$\int_{-\infty}^{+\infty} t^k \psi(t) dt = 0 \quad \text{for} \quad 0 \leq k < n.$$

A wavelet with  $n$  vanishing moments is orthogonal to polynomials of degree  $n - 1$ . Since  $\alpha < n$ , the polynomial  $p_v$  has degree at most  $n - 1$ . With the change of variable  $t' = (t - u)/s$ , we verify that

$$W p_v(u, s) = \int_{-\infty}^{+\infty} p_v(t) \frac{1}{\sqrt{s}} \psi\left(\frac{t-u}{s}\right) dt = 0. \quad (6.7)$$

Since  $f = p_v + \varepsilon_v$ ,

$$W f(u, s) = W \varepsilon_v(u, s). \quad (6.8)$$

Section 6.1.3 explains how to measure  $\alpha$  from  $|W f(u, s)|$  when  $u$  is in the neighborhood of  $v$ .

#### *Multiscale Differential Operator*

Theorem 6.2 proves that a wavelet with  $n$  vanishing moments can be written as the  $n$ th-order derivative of a function  $\theta$ ; the resulting wavelet transform is a multiscale differential operator. We suppose that  $\psi$  has a fast decay, which means that for any decay exponent  $m \in \mathbb{N}$  there exists  $C_m$  such that

$$\forall t \in \mathbb{R}, \quad |\psi(t)| \leq \frac{C_m}{1 + |t|^m}. \quad (6.9)$$

**Theorem 6.2.** A wavelet  $\psi$  with a fast decay has  $n$  vanishing moments if and only if there exists  $\theta$  with a fast decay such that

$$\psi(t) = (-1)^n \frac{d^n \theta(t)}{dt^n}. \quad (6.10)$$

As a consequence

$$W f(u, s) = s^n \frac{d^n}{du^n} (f \star \bar{\theta}_s)(u), \quad (6.11)$$

with  $\bar{\theta}_s(t) = s^{-1/2}\theta(-t/s)$ . Moreover,  $\psi$  has no more than  $n$  vanishing moments if and only if  $\int_{-\infty}^{+\infty} \theta(t) dt \neq 0$ .

**Proof.** The fast decay of  $\psi$  implies that  $\hat{\psi}$  is  $C^\infty$ . This is proved by setting  $f = \hat{\psi}$  in Theorem 2.5.

The integral of a function is equal to its Fourier transform evaluated at  $\omega = 0$ . The derivative property (2.22) implies that for any  $k < n$ ,

$$\int_{-\infty}^{+\infty} t^k \psi(t) dt = (i)^k \hat{\psi}^{(k)}(0) = 0. \quad (6.12)$$

We can therefore make the factorization

$$\hat{\psi}(\omega) = (-i\omega)^n \hat{\theta}(\omega), \quad (6.13)$$

and  $\hat{\theta}(\omega)$  is bounded. The fast decay of  $\theta$  is proved with an induction on  $n$ . For  $n = 1$ ,

$$\theta(t) = \int_{-\infty}^t \psi(u) du = \int_t^{+\infty} \psi(u) du,$$

and the fast decay of  $\theta$  is derived from (6.9). We then similarly verify that increasing the order of integration by 1 up to  $n$  maintains the fast decay of  $\theta$ .

Conversely,  $|\hat{\theta}(\omega)| \leq \int_{-\infty}^{+\infty} |\theta(t)| dt < +\infty$ , because  $\theta$  has a fast decay. The Fourier transform of (6.10) yields (6.13), which implies that  $\hat{\psi}^{(k)}(0) = 0$  for  $k < n$ . It follows from (6.12) that  $\psi$  has  $n$  vanishing moments.

To test whether  $\psi$  has more than  $n$  vanishing moments, we compute with (6.13)

$$\int_{-\infty}^{+\infty} t^n \psi(t) dt = (i)^n \hat{\psi}^{(n)}(0) = (-i)^n n! \hat{\theta}(0).$$

Clearly,  $\psi$  has no more than  $n$  vanishing moments if and only if  $\hat{\theta}(0) = \int_{-\infty}^{+\infty} \theta(t) dt \neq 0$ .

The wavelet transform (4.32) can be written

$$Wf(u, s) = f \star \bar{\psi}_s(u) \quad \text{with} \quad \bar{\psi}_s(t) = \frac{1}{\sqrt{s}} \psi\left(\frac{-t}{s}\right). \quad (6.14)$$

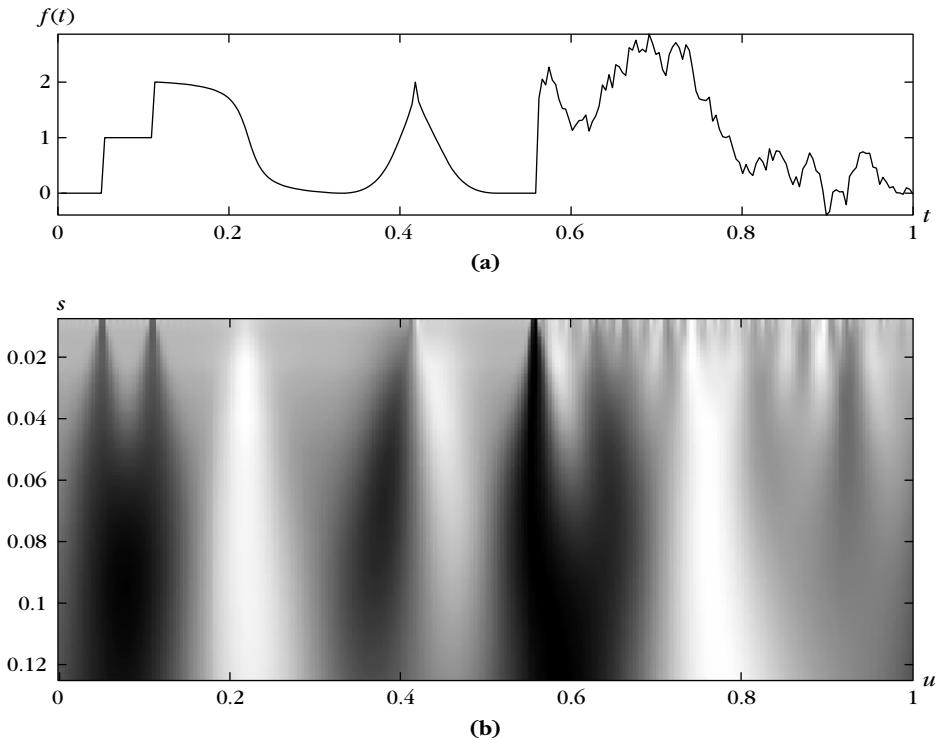
We derive from (6.10) that  $\bar{\psi}_s(t) = s^n \frac{d^n \bar{\theta}_s(t)}{dt^n}$ . Commuting the convolution and differentiation operators yields

$$Wf(u, s) = s^n f \star \frac{d^n \bar{\theta}_s}{dt^n}(u) = s^n \frac{d^n}{du^n} (f \star \bar{\theta}_s)(u). \quad \blacksquare$$

If  $K = \int_{-\infty}^{+\infty} \theta(t) dt \neq 0$ , then the convolution  $f \star \bar{\theta}_s(t)$  can be interpreted as a weighted average of  $f$  with a kernel dilated by  $s$ . So (6.11) proves that  $Wf(u, s)$  is an  $n$ th-order derivative of an averaging of  $f$  over a domain proportional to  $s$ . Figure 6.1 shows a wavelet transform calculated with  $\psi = -\theta'$ , where  $\theta$  is a Gaussian. The resulting  $Wf(u, s)$  is the derivative of  $f$  averaged in the neighborhood of  $u$  with a Gaussian kernel dilated by  $s$ .

Since  $\theta$  has a fast decay, one can verify that

$$\lim_{s \rightarrow 0} \frac{1}{\sqrt{s}} \bar{\theta}_s = K \delta,$$

**FIGURE 6.1**

Wavelet transform  $Wf(u, s)$  calculated with  $\psi = -\theta'$  where  $\theta$  is a Gaussian, for the signal  $f$  shown in (a). Position parameter  $u$  and scale  $s$  vary, respectively, along the horizontal and vertical axes. (b) Black, gray, and white points correspond to positive, zero, and negative wavelet coefficients. Singularities create large-amplitude coefficients in their cone of influence.

in the sense of the weak convergence (A.30). This means that for any  $\phi$  that is continuous at  $u$ ,

$$\lim_{s \rightarrow 0} \phi \star \frac{1}{\sqrt{s}} \bar{\theta}_s(u) = K \phi(u).$$

If  $f$  is  $n$  times continuously differentiable in the neighborhood of  $u$ , then (6.11) implies that

$$\lim_{s \rightarrow 0} \frac{Wf(u, s)}{s^{n+1/2}} = \lim_{s \rightarrow 0} f^{(n)} \star \frac{1}{\sqrt{s}} \bar{\theta}_s(u) = K f^{(n)}(u). \quad (6.15)$$

In particular, if  $f$  is  $C^n$  with a bounded  $n$ th-order derivative, then  $|Wf(u, s)| = O(s^{n+1/2})$ . This is a first relation between the decay of  $|Wf(u, s)|$  when  $s$  decreases and the uniform regularity of  $f$ . Finer relations are studied in the next section.

### 6.1.3 Regularity Measurements with Wavelets

The decay of the wavelet transform amplitude across scales is related to the uniform and pointwise Lipschitz regularity of the signal. Measuring this asymptotic decay is equivalent to zooming into signal structures with a scale that goes to zero. We suppose that the wavelet  $\psi$  has  $n$  vanishing moments and is  $C^n$  with derivatives that have a fast decay. This means that for any  $0 \leq k \leq n$  and  $m \in \mathbb{N}$  there exists  $C_m$  such that

$$\forall t \in \mathbb{R}, \quad |\psi^{(k)}(t)| \leq \frac{C_m}{1 + |t|^m}. \quad (6.16)$$

Theorem 6.3 relates the uniform Lipschitz regularity of  $f$  on an interval to the amplitude of its wavelet transform at fine scales.

**Theorem 6.3.** If  $f \in L^2(\mathbb{R})$  is uniformly Lipschitz  $\alpha \leq n$  over  $[a, b]$ , then there exists  $A > 0$  such that

$$\forall (u, s) \in [a, b] \times \mathbb{R}^+, \quad |Wf(u, s)| \leq A s^{\alpha+1/2}. \quad (6.17)$$

Conversely, suppose that  $f$  is bounded and that  $Wf(u, s)$  satisfies (6.17) for an  $\alpha < n$  that is not an integer. Then  $f$  is uniformly Lipschitz  $\alpha$  on  $[a + \varepsilon, b - \varepsilon]$ , for any  $\varepsilon > 0$ .

**Proof.** This theorem is proved with minor modifications in the proof of Theorem 6.4. Since  $f$  is Lipschitz  $\alpha$  at any  $v \in [a, b]$ , Theorem 6.4 shows in (6.20) that

$$\forall (u, s) \in \mathbb{R} \times \mathbb{R}^+, \quad |Wf(u, s)| \leq A s^{\alpha+1/2} \left( 1 + \left| \frac{u-v}{s} \right|^\alpha \right).$$

For  $u \in [a, b]$ , we can choose  $v = u$ , which implies that  $|Wf(u, s)| \leq A s^{\alpha+1/2}$ . We verify from the proof of (6.20) that the constant  $A$  does not depend on  $v$  because the Lipschitz regularity is uniform over  $[a, b]$ .

To prove that  $f$  is uniformly Lipschitz  $\alpha$  over  $[a + \varepsilon, b - \varepsilon]$ , we must verify that there exists  $K$  such that for all  $v \in [a + \varepsilon, b - \varepsilon]$  we can find a polynomial  $p_v$  of degree  $\lfloor \alpha \rfloor$  such that

$$\forall t \in \mathbb{R}, \quad |f(t) - p_v(t)| \leq K |t - v|^\alpha. \quad (6.18)$$

When  $t \notin [a + \varepsilon/2, b - \varepsilon/2]$ , then  $|t - v| \geq \varepsilon/2$ , and since  $f$  is bounded, (6.18) is verified with a constant  $K$  that depends on  $\varepsilon$ . For  $t \in [a + \varepsilon/2, b - \varepsilon/2]$ , the proof follows the same derivations as the proof of pointwise Lipschitz regularity from (6.21) in Theorem 6.4. The upper bounds (6.26) and (6.27) are replaced by

$$\forall t \in [a + \varepsilon/2, b - \varepsilon/2], \quad |\Delta_j^{(k)}(t)| \leq K 2^{(\alpha-k)j} \quad \text{for } 0 \leq k \leq \lfloor \alpha \rfloor + 1. \quad (6.19)$$

This inequality is verified by computing an upper-bound integral similar to (6.25) but which is divided in two— $u \in [a, b]$  and  $u \notin [a, b]$ . When  $u \in [a, b]$ , the condition (6.21) is replaced by  $|Wf(u, s)| \leq A s^{\alpha+1/2}$  in (6.25). When  $u \notin [a, b]$ , we just use the fact that  $|Wf(u, s)| \leq \|f\| \|\psi\|$  and derive (6.19) from the fast decay of  $|\psi^{(k)}(t)|$ , by observing that  $|t - u| \geq \varepsilon/2$  for  $t \in [a + \varepsilon/2, b - \varepsilon/2]$ . The constant  $K$  depends on  $A$  and  $\varepsilon$  but not on  $v$ . The proof then proceeds like the proof of Theorem 6.4, and since the resulting

constant  $K$  in (6.29) does not depend on  $v$ , the Lipschitz regularity is uniform over  $[a - \varepsilon, b + \varepsilon]$ . ■

The inequality (6.17) is really a condition on the asymptotic decay of  $|Wf(u, s)|$  when  $s$  goes to zero. At large scales it does not introduce any constraint since the Cauchy-Schwarz inequality guarantees that the wavelet transform is bounded:

$$|Wf(u, s)| = |\langle f, \psi_{u,s} \rangle| \leq \|f\| \|\psi\|.$$

When the scale  $s$  decreases,  $Wf(u, s)$  measures fine-scale variations in the neighborhood of  $u$ . Theorem 6.3 proves that  $|Wf(u, s)|$  decays like  $s^{\alpha+1/2}$  over intervals where  $f$  is uniformly Lipschitz  $\alpha$ .

Observe that the upper bound (6.17) is similar to the sufficient Fourier condition of theorem (6.1), which supposes that  $|\hat{f}(\omega)|$  decays faster than  $\omega^{-\alpha}$ . The wavelet scale  $s$  plays the role of a “localized” inverse frequency  $\omega^{-1}$ . As opposed to the Fourier transform theorem (6.1), the wavelet transform gives a Lipschitz regularity condition that is localized over any finite interval and it provides a necessary condition that is nearly sufficient. When  $[a, b] = \mathbb{R}$ , then (6.17) is a necessary and sufficient condition for  $f$  to be uniformly Lipschitz  $\alpha$  on  $\mathbb{R}$ .

If  $\psi$  has exactly  $n$  vanishing moments, then the wavelet transform decay gives no information concerning the Lipschitz regularity of  $f$  for  $\alpha > n$ . If  $f$  is uniformly Lipschitz  $\alpha > n$ , then it is  $C^n$  and (6.15) proves that  $\lim_{s \rightarrow 0} s^{-n-1/2} Wf(u, s) = Kf^{(n)}(u)$  with  $K \neq 0$ . This proves that  $|Wf(u, s)| \sim s^{n+1/2}$  at fine scales despite the higher regularity of  $f$ .

If the Lipschitz exponent  $\alpha$  is an integer, then (6.17) is not sufficient to prove that  $f$  is uniformly Lipschitz  $\alpha$ . When  $[a, b] = \mathbb{R}$ , if  $\alpha = 1$  and  $\psi$  has two vanishing moments, then the class of functions that satisfy (6.17) is called the *Zygmund class* [44]. It is slightly larger than the set of functions that are uniformly Lipschitz 1. For example,  $f(t) = t \log_e t$  belongs to the Zygmund class although it is not Lipschitz 1 at  $t = 0$ .

### **Pointwise Lipschitz Regularity**

The study of pointwise Lipschitz exponents with the wavelet transform is a delicate and beautiful topic that finds its mathematical roots in the characterization of Sobolev spaces by Littlewood and Paley in the 1930s. Characterizing the regularity of  $f$  at a point  $v$  can be difficult because  $f$  may have very different types of singularities that are aggregated in the neighborhood of  $v$ . In 1984, Bony [118] introduced the “two-microlocalization” theory, which refines the Littlewood-Paley approach to provide pointwise characterization of singularities that he used to study the solution of hyperbolic partial differential equations. These technical results became much simpler through the work of Jaffard [312] who proved that the two-microlocalization properties are equivalent to specific decay conditions on the wavelet transform amplitude. Theorem 6.4 gives a necessary and a sufficient condition on the wavelet transform for estimating the Lipschitz regularity of  $f$  at a point  $v$ . Remember that the wavelet  $\psi$  has  $n$  vanishing moments and  $n$  derivatives having a fast decay.

**Theorem 6.4:** Jaffard. If  $f \in L^2(\mathbb{R})$  is Lipschitz  $\alpha \leq n$  at  $v$ , then there exists  $A$  such that

$$\forall (u, s) \in \mathbb{R} \times \mathbb{R}^+, \quad |Wf(u, s)| \leq A s^{\alpha+1/2} \left( 1 + \left| \frac{u-v}{s} \right|^\alpha \right). \quad (6.20)$$

Conversely, if  $\alpha < n$  is not an integer and there exist  $A$  and  $\alpha' < \alpha$  such that

$$\forall (u, s) \in \mathbb{R} \times \mathbb{R}^+, \quad |Wf(u, s)| \leq A s^{\alpha+1/2} \left( 1 + \left| \frac{u-v}{s} \right|^{\alpha'} \right), \quad (6.21)$$

then  $f$  is Lipschitz  $\alpha$  at  $v$ .

**Proof.** The necessary condition is relatively simple to prove but the sufficient condition is much more difficult.

**Proof of (6.20).** Since  $f$  is Lipschitz  $\alpha$  at  $v$ , there exists a polynomial  $p_v$  of degree  $\lfloor \alpha \rfloor < n$  and  $K$  such that  $|f(t) - p_v(t)| \leq K|t - v|^\alpha$ . Since  $\psi$  has  $n$  vanishing moments, we saw in (6.7) that  $Wp_v(u, s) = 0$ , and thus

$$\begin{aligned} |Wf(u, s)| &= \left| \int_{-\infty}^{+\infty} (f(t) - p_v(t)) \frac{1}{\sqrt{s}} \psi\left(\frac{t-u}{s}\right) dt \right| \\ &\leq \int_{-\infty}^{+\infty} K |t-v|^\alpha \frac{1}{\sqrt{s}} \left| \psi\left(\frac{t-u}{s}\right) \right| dt. \end{aligned}$$

The change of variable  $x = (t-u)/s$  gives

$$|Wf(u, s)| \leq \sqrt{s} \int_{-\infty}^{+\infty} K |sx + u - v|^\alpha |\psi(x)| dx.$$

Since  $|a+b|^\alpha \leq 2^\alpha (|a|^\alpha + |b|^\alpha)$ ,

$$|Wf(u, s)| \leq K 2^\alpha \sqrt{s} \left( s^\alpha \int_{-\infty}^{+\infty} |x|^\alpha |\psi(x)| dx + |u-v|^\alpha \int_{-\infty}^{+\infty} |\psi(x)| dx \right),$$

which proves (6.20).

**Proof of (6.21).** The wavelet reconstruction formula (4.37) proves that  $f$  can be decomposed in a Littlewood-Paley-type sum

$$f(t) = \sum_{j=-\infty}^{+\infty} \Delta_j(t) \quad (6.22)$$

with

$$\Delta_j(t) = \frac{1}{C_\psi} \int_{-\infty}^{+\infty} \int_{2^j}^{2^{j+1}} Wf(u, s) \frac{1}{\sqrt{s}} \psi\left(\frac{t-u}{s}\right) \frac{ds}{s^2} du. \quad (6.23)$$

Let  $\Delta_j^{(k)}$  be its  $k$ th-order derivative. To prove that  $f$  is Lipschitz  $\alpha$  at  $v$  we shall approximate  $f$  with a polynomial that generalizes the Taylor polynomial

$$p_v(t) = \sum_{k=0}^{\lfloor \alpha \rfloor} \left( \sum_{j=-\infty}^{+\infty} \Delta_j^{(k)}(v) \right) \frac{(t-v)^k}{k!}. \quad (6.24)$$

If  $f$  is  $n$  times differentiable at  $v$ , then  $p_v$  corresponds to the Taylor polynomial; however, this is not necessarily true. We shall first prove that  $\sum_{j=-\infty}^{+\infty} \Delta_j^{(k)}(v)$  is finite by getting upper bounds on  $|\Delta_j^{(k)}(t)|$ . These sums may be thought of as a generalization of pointwise derivatives.

To simplify the notation, we denote by  $K$  a generic constant that may change value from one line to the next but that does not depend on  $j$  and  $t$ . The hypothesis (6.21) and the asymptotic decay condition (6.16) imply that

$$\begin{aligned} |\Delta_j(t)| &= \frac{1}{C_\psi} \int_{-\infty}^{+\infty} \int_{2^j}^{2^{j+1}} A s^\alpha \left(1 + \left|\frac{u-v}{s}\right|^{\alpha'}\right) \frac{C_m}{1 + |(t-u)/s|^m} \frac{ds}{s^2} du \\ &\leq K \int_{-\infty}^{+\infty} 2^{\alpha j} \left(1 + \left|\frac{u-v}{2^j}\right|^{\alpha'}\right) \frac{1}{1 + |(t-u)/2^j|^m} \frac{du}{2^j}. \end{aligned} \quad (6.25)$$

Since  $|u-v|^{\alpha'} \leq 2^{\alpha'} (|u-t|^{\alpha'} + |t-v|^{\alpha'})$ , the change of variable  $u' = 2^{-j}(u-t)$  yields

$$|\Delta_j(t)| \leq K 2^{\alpha j} \int_{-\infty}^{+\infty} \frac{1 + |u'|^{\alpha'} + |(v-t)/2^j|^{\alpha'}}{1 + |u'|^m} du'.$$

Choosing  $m = \alpha' + 2$  yields

$$|\Delta_j(t)| \leq K 2^{\alpha j} \left(1 + \left|\frac{v-t}{2^j}\right|^{\alpha'}\right). \quad (6.26)$$

The same derivations applied to the derivatives of  $\Delta_j(t)$  yield

$$\forall k \leq \lfloor \alpha \rfloor + 1, \quad |\Delta_j^{(k)}(t)| \leq K 2^{(\alpha-k)j} \left(1 + \left|\frac{v-t}{2^j}\right|^{\alpha'}\right). \quad (6.27)$$

At  $t = v$ , it follows that

$$\forall k \leq \lfloor \alpha \rfloor, \quad |\Delta_j^{(k)}(v)| \leq K 2^{(\alpha-k)j}. \quad (6.28)$$

This guarantees a fast decay of  $|\Delta_j^{(k)}(v)|$  when  $2^j$  goes to zero, because  $\alpha$  is not an integer so  $\alpha > \lfloor \alpha \rfloor$ . At large scales  $2^j$ , since  $|Wf(u, s)| \leq \|f\| \|\psi\|$  with the change of variable  $u' = (t-u)/s$  in (6.23), we have

$$|\Delta_j^{(k)}(v)| \leq \frac{\|f\| \|\psi\|}{C_\psi} \int_{-\infty}^{+\infty} |\psi^{(k)}(u')| du' \int_{2^j}^{2^{j+1}} \frac{ds}{s^{3/2+k}},$$

therefore  $|\Delta_j^{(k)}(v)| \leq K 2^{-(k+1/2)j}$ . Together with (6.28) this proves that the polynomial  $p_v$  defined in (6.24) has coefficients that are finite.

With the Littlewood-Paley decomposition (6.22), we compute

$$|f(t) - p_v(t)| = \left| \sum_{j=-\infty}^{+\infty} \left( \Delta_j(t) - \sum_{k=0}^{\lfloor \alpha \rfloor} \Delta_j^{(k)}(v) \frac{(t-v)^k}{k!} \right) \right|.$$

The sum over scales is divided in two at  $2^j$  such that  $2^j \geq |t - v| \geq 2^{j-1}$ . For  $j \geq J$ , we can use the classical Taylor theorem to bound the Taylor expansion of  $\Delta_j$ :

$$\begin{aligned} I &= \sum_{j=J}^{+\infty} \left| \Delta_j(t) - \sum_{k=0}^{\lfloor \alpha \rfloor} \Delta_j^{(k)}(v) \frac{(t-v)^k}{k!} \right| \\ &\leq \sum_{j=J}^{+\infty} \frac{(t-v)^{\lfloor \alpha \rfloor + 1}}{(\lfloor \alpha \rfloor + 1)!} \sup_{h \in [t, v]} |\Delta_j^{\lfloor \alpha \rfloor + 1}(h)|. \end{aligned}$$

Inserting (6.27) yields

$$I \leq K |t - v|^{\lfloor \alpha \rfloor + 1} \sum_{j=J}^{+\infty} 2^{-j(\lfloor \alpha \rfloor + 1 - \alpha)} \left| \frac{v-t}{2^j} \right|^{\alpha'},$$

and since  $2^j \geq |t - v| \geq 2^{j-1}$ , we get  $I \leq K |v - t|^\alpha$ .

Let us now consider the case  $j < J$ :

$$\begin{aligned} II &= \sum_{j=-\infty}^{j-1} \left| \Delta_j(t) - \sum_{k=0}^{\lfloor \alpha \rfloor} \Delta_j^{(k)}(v) \frac{(t-v)^k}{k!} \right| \\ &\leq K \sum_{j=-\infty}^{j-1} \left( 2^{\alpha j} \left( 1 + \left| \frac{v-t}{2^j} \right|^{\alpha'} \right) + \sum_{k=0}^{\lfloor \alpha \rfloor} \frac{(t-v)^k}{k!} 2^{j(\alpha-k)} \right) \\ &\leq K \left( 2^{\alpha J} + 2^{(\alpha-\alpha')J} |t-v|^{\alpha'} + \sum_{k=0}^{\lfloor \alpha \rfloor} \frac{(t-v)^k}{k!} 2^{j(\alpha-k)} \right), \end{aligned}$$

and since  $2^j \geq |t - v| \geq 2^{j-1}$ , we get  $II \leq K |v - t|^\alpha$ . As a result,

$$|f(t) - p_v(t)| \leq I + II \leq K |v - t|^\alpha, \quad (6.29)$$

which proves that  $f$  is Lipschitz  $\alpha$  at  $v$ . ■

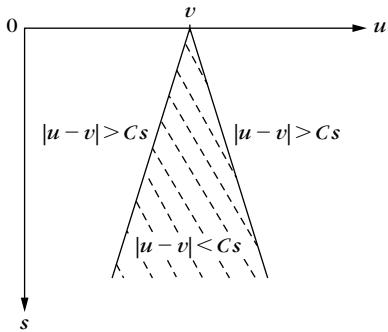
### Cone of Influence

To interpret more easily the necessary condition (6.20) and the sufficient condition (6.21), we shall suppose that  $\psi$  has a compact support equal to  $[-C, C]$ . The *cone of influence* of  $v$  in the scale-space plane is the set of points  $(u, s)$  such that  $v$  is included in the support of  $\psi_{u,s}(t) = s^{-1/2} \psi((t-u)/s)$ . Since the support of  $\psi((t-u)/s)$  is equal to  $[u - Cs, u + Cs]$ , the cone of influence of  $v$  is defined by

$$|u - v| \leq Cs. \quad (6.30)$$

It is illustrated in Figure 6.2. If  $u$  is in the cone of influence of  $v$ , then  $Wf(u, s) = \langle f, \psi_{u,s} \rangle$  depends on the value of  $f$  in the neighborhood of  $v$ . Since  $|u - v|/s \leq C$ , the conditions (6.20, 6.21) can be written as

$$|Wf(u, s)| \leq A' s^{\alpha+1/2},$$

**FIGURE 6.2**

The cone of influence of an abscissa  $v$  consists of the scale-space points  $(u, s)$  for which the support of  $\psi_{u,s}$  intersects  $t = v$ .

which is identical to the uniform Lipschitz condition (6.17) given by Theorem 6.3. In Figure 6.1, the high-amplitude wavelet coefficients are in the cone of influence of each singularity.

### Oscillating Singularities

It may seem surprising that (6.20) and (6.21) also impose a condition on the wavelet transform outside the cone of influence of  $v$ . Indeed, this corresponds to wavelets of which the support does not intersect  $v$ . For  $|u - v| > Cs$ , we get

$$|Wf(u, s)| \leq A' s^{\alpha-\alpha'+1/2} |u - v|^\alpha. \quad (6.31)$$

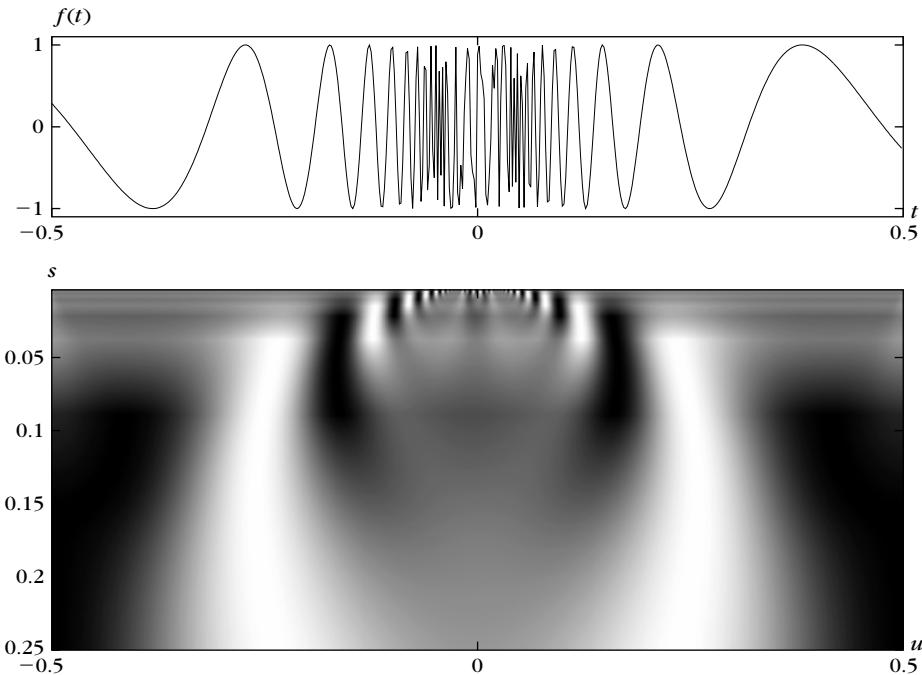
We shall see that it is indeed necessary to impose this decay when  $u$  tends to  $v$  in order to control the oscillations of  $f$  that might generate singularities.

Let us consider the generic example of a highly oscillatory function

$$f(t) = \sin \frac{1}{t},$$

which is discontinuous at  $v = 0$  because of the acceleration of its oscillations. Since  $\psi$  is a smooth  $C^n$  function, if it is centered close to zero, then the rapid oscillations of  $\sin t^{-1}$  produce a correlation integral  $\langle \sin t^{-1}, \psi_{u,s} \rangle$  that is very small. With an integration by parts, one can verify that if  $(u, s)$  is in the cone of influence of  $v = 0$ , then  $|Wf(u, s)| \leq A s^{2+1/2}$ . This looks as if  $f$  is Lipschitz 2 at 0. However, Figure 6.3 shows high-energy wavelet coefficients outside the cone of influence of  $v = 0$ , which are responsible for the discontinuity. To guarantee that  $f$  is Lipschitz  $\alpha$ , the amplitude of such coefficients is controlled by the upper bound (6.31).

To explain why high-frequency oscillations appear outside the cone of influence of  $v$ , we use the results of Section 4.4.3 on the estimation of instantaneous frequencies with wavelet ridges. The instantaneous frequency of  $\sin t^{-1} = \sin \theta(t)$  is  $|\theta'(t)| = t^{-2}$ . Let  $\psi^a$  be the analytic part of  $\psi$ , defined in (4.47). The corresponding

**FIGURE 6.3**

Wavelet transform of  $f(t) = \sin(\alpha t^{-1})$  calculated with  $\psi = -\theta'$ , where  $\theta$  is a Gaussian. High-amplitude coefficients are along a parabola outside the cone of influence of  $t = 0$ .

complex analytic wavelet transform is  $W^a f(u, s) = \langle f, \psi_{u,s}^a \rangle$ . It was proved in (4.109) that for a fixed time  $u$ , the maximum of  $s^{-1/2} |W^a f(u, s)|$  is located at the scale

$$s(u) = \frac{\eta}{\theta'(u)} = \eta u^2,$$

where  $\eta$  is the center frequency of  $\hat{\psi}^a(\omega)$ . When  $u$  varies, the set of points  $(u, s(u))$  defines a *ridge* that is a parabola located outside the cone of influence of  $v = 0$  in the plane  $(u, s)$ . Since  $\psi = \operatorname{Re}[\psi^a]$ , the real wavelet transform is

$$Wf(u, s) = \operatorname{Re}[W^a f(u, s)].$$

The high-amplitude values of  $Wf(u, s)$  are thus located along the same parabola ridge curve in the scale-space plane, which clearly appears in Figure 6.3. Real wavelet coefficients  $Wf(u, s)$  change signs along the ridge because of the variations of the complex phase of  $W^a f(u, s)$ .

The example of  $f(t) = \sin t^{-1}$  can be extended to general oscillating singularities [32]. A function  $f$  has an oscillating singularity at  $v$  if there exist  $\alpha \geq 0$  and  $\beta > 0$  such that for  $t$  in a neighborhood of  $v$ ,

$$f(t) \sim |t - v|^\alpha g\left(\frac{1}{|t - v|^\beta}\right),$$

where  $g(t)$  is a  $C^\infty$  oscillating function that has primitives bounded at any order. The function  $g(t) = \sin t^{-1}$  is a typical example. The oscillations have an instantaneous frequency  $\theta'(t)$  that increases to infinity faster than  $|t|^{-1}$  when  $t$  goes to  $v$ . High-energy wavelet coefficients are located along the ridge  $s(u) = \eta/\theta'(u)$ , and this curve is necessarily outside the cone of influence  $|u - v| \leq Cs$ .

## 6.2 WAVELET TRANSFORM MODULUS MAXIMA

Theorems 6.3 and 6.4 prove that the local Lipschitz regularity of  $f$  at  $v$  depends on the decay at fine scales of  $|Wf(u, s)|$  in the neighborhood of  $v$ . Measuring this decay directly in the time-scale plane  $(u, s)$  is not necessary. The decay of  $|Wf(u, s)|$  can indeed be controlled from its local maxima values. Section 6.2.1 studies the detection and characterization of singularities from wavelet local maxima. Signal approximations are recovered in Section 6.2.2, from the scale-space support of these local maxima at dyadic scales.

### 6.2.1 Detection of Singularities

Singularities are detected by finding the abscissa where the wavelet modulus maxima converge at fine scales. A wavelet *modulus maximum* is defined as a point  $(u_0, s_0)$  such that  $|Wf(u, s_0)|$  is locally maximum at  $u = u_0$ . This implies that

$$\frac{\partial Wf(u_0, s_0)}{\partial u} = 0.$$

This local maximum should be a strict local maximum in either the right or the left neighborhood of  $u_0$  to avoid having any local maxima when  $|Wf(u, s_0)|$  is constant. We call any connected curve  $s(u)$  in the scale-space plane  $(u, s)$  along which all points are modulus maxima a *maxima line*. (See Figure 6.5b on page 218, which shows the wavelet modulus maxima of a signal.)

To better understand the properties of these maxima, the wavelet transform is written as a multiscale differential operator. Theorem 6.2 proves that if  $\psi$  has exactly  $n$  vanishing moments and a compact support, then there exists  $\theta$  of compact support such that  $\psi = (-1)^n \theta^{(n)}$  with  $\int_{-\infty}^{+\infty} \theta(t) dt \neq 0$ . The wavelet transform is rewritten in (6.11) as a multiscale differential operator

$$Wf(u, s) = s^n \frac{d^n}{du^n} (f \star \bar{\theta}_s)(u). \quad (6.32)$$

If the wavelet has only one vanishing moment, wavelet modulus maxima are the maxima of the first-order derivative of  $f$  smoothed by  $\bar{\theta}_s$ , as illustrated by Figure 6.4. These multiscale modulus maxima are used to locate discontinuities and edges in images. If the wavelet has two vanishing moments, the modulus maxima correspond to high curvatures. Theorem 6.5 proves that if  $Wf(u, s)$  has no modulus maxima at fine scales, then  $f$  is locally regular.

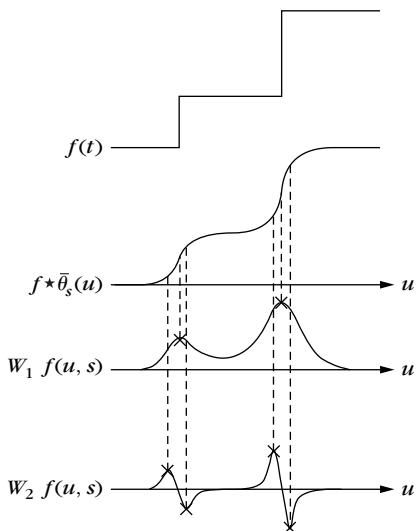


FIGURE 6.4

The convolution  $f \star \bar{\theta}_s(u)$  averages  $f$  over a domain proportional to  $s$ . If  $\psi = -\theta'$ , then  $W_1 f(u, s) = s \frac{d}{du} (f \star \bar{\theta}_s)(u)$  has modulus maxima at sharp variation points of  $f \star \bar{\theta}_s(u)$ . If  $\psi = \theta''$ , then the modulus maxima of  $W_2 f(u, s) = s^2 \frac{d^2}{du^2} (f \star \bar{\theta}_s)(u)$  correspond to locally maximum curvatures.

**Theorem 6.5:** *Hwang, Mallat.* Suppose that  $\psi$  is  $\mathbf{C}^n$  with a compact support, and  $\psi = (-1)^n \theta^{(n)}$  with  $\int_{-\infty}^{+\infty} \theta(t) dt \neq 0$ . Let  $f \in \mathbf{L}^1[a, b]$ . If there exists  $s_0 > 0$  such that  $|Wf(u, s)|$  has no local maximum for  $u \in [a, b]$  and  $s < s_0$ , then  $f$  is uniformly Lipschitz  $n$  on  $[a + \varepsilon, b - \varepsilon]$ , for any  $\varepsilon > 0$ .

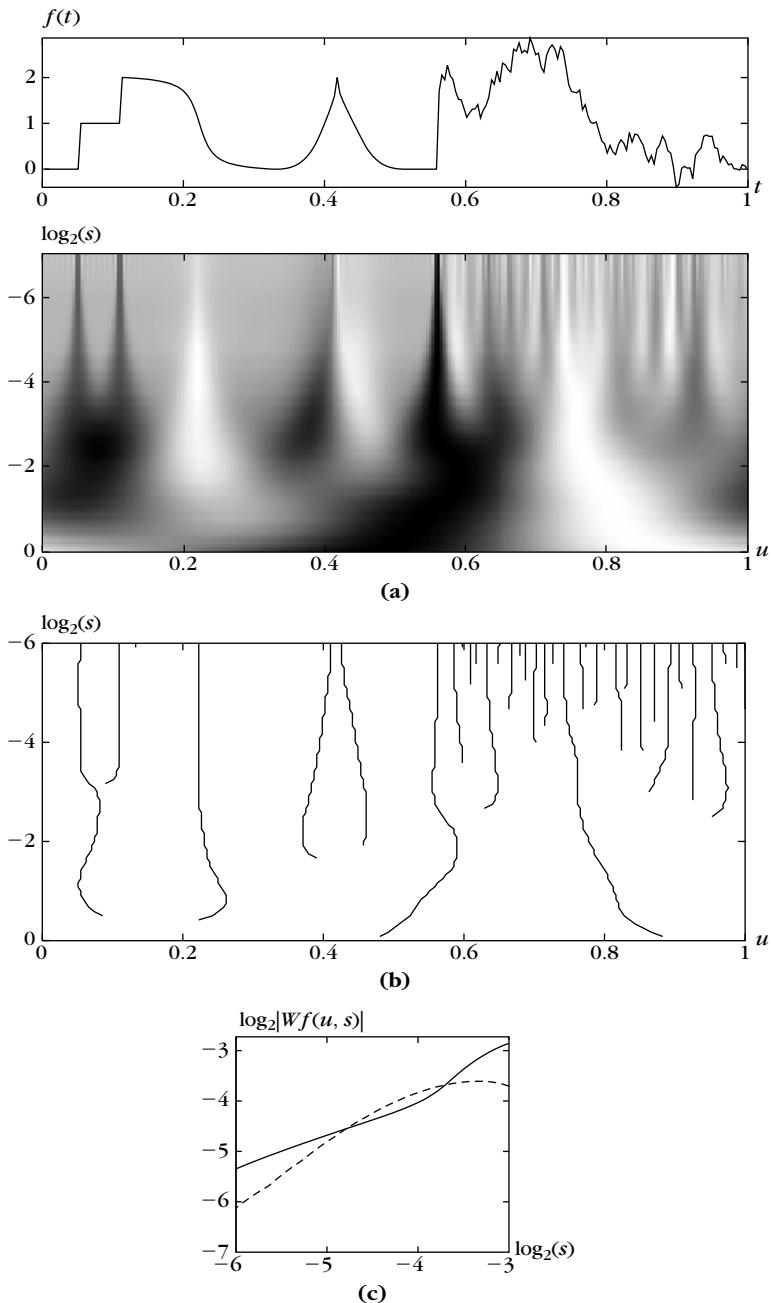
This theorem is proved in [364]. It implies that  $f$  can be singular (not Lipschitz 1) at a point  $v$  only if there is a sequence of wavelet maxima points  $(u_p, s_p)_{p \in \mathbb{N}}$  that converges toward  $v$  at fine scales:

$$\lim_{p \rightarrow +\infty} u_p = v \quad \text{and} \quad \lim_{p \rightarrow +\infty} s_p = 0.$$

These modulus maxima points may or may not be along the same maxima line. This result guarantees that all singularities are detected by following the wavelet transform modulus maxima at fine scales. Figure 6.5 gives an example where all singularities are located by following the maxima lines.

### Maxima Propagation

For all  $\psi = (-1)^n \theta^{(n)}$ , we are not guaranteed that a modulus maxima located at  $(u_0, s_0)$  belongs to a maxima line that propagates toward finer scales. When  $s$  decreases,  $Wf(u, s)$  may have no more maxima in the neighborhood of  $u = u_0$ .



**FIGURE 6.5**

- (a) Wavelet transform  $Wf(u, s)$ ; the horizontal and vertical axes give  $u$  and  $\log_2 s$ , respectively.  
 (b) Modulus maxima of  $Wf(u, s)$ . (c) The full line gives the decay of  $\log_2 |Wf(u, s)|$  as a function of  $\log_2 s$  along the maxima line that converges to the abscissa  $t = 0.05$ . The dashed line gives  $\log_2 |Wf(u, s)|$  along the left maxima line that converges to  $t = 0.42$ .

Theorem 6.6 proves that this is never the case if  $\theta$  is a Gaussian. The wavelet transform  $Wf(u, s)$  can then be written as the solution of the heat-diffusion equation, where  $s$  is proportional to the diffusion time. The maximum principle applied to the heat-diffusion equation proves that maxima may not disappear when  $s$  decreases. Applications of the heat-diffusion equation to the analysis of multiscale averaging have been studied by several computer vision researchers [309, 330, 496].

**Theorem 6.6:** *Hummel, Poggio, Yuille.* Let  $\psi = (-1)^n \theta^{(n)}$ , where  $\theta$  is a Gaussian. For any  $f \in L^2(\mathbb{R})$ , the modulus maxima of  $Wf(u, s)$  belong to connected curves that are never interrupted when the scale decreases.

**Proof.** To simplify the proof, we suppose that  $\theta$  is a normalized Gaussian  $\theta(t) = 2^{-1}\pi^{-1/2} \exp(-t^2/4)$  and that the Fourier transform is  $\hat{\theta}(\omega) = \exp(-\omega^2)$ . Theorem 6.2 proves that

$$Wf(u, s) = s^n f^{(n)} * \theta_s(u), \quad (6.33)$$

where the  $n$ th derivative  $f^{(n)}$  is defined in the sense of distributions. Let  $\tau$  be the diffusion time. The solution of

$$\frac{\partial g(u, \tau)}{\partial \tau} = \frac{\partial^2 g(u, \tau)}{\partial u^2} \quad (6.34)$$

with initial condition  $g(u, 0) = g_0(u)$  is obtained by computing the Fourier transform with respect to  $u$  of (6.34):

$$\frac{\partial \hat{g}(\omega, u)}{\partial \tau} = -\omega^2 \hat{g}(\omega, \tau).$$

It follows that  $\hat{g}(\omega, \tau) = \hat{g}_0(\omega) \exp(-\tau\omega^2)$  and thus,

$$g(u, \tau) = \frac{1}{\sqrt{\tau}} g_0 * \theta_{\sqrt{\tau}}(u).$$

For  $\tau = s^2$ , setting  $g_0 = f^{(n)}$  and inserting (6.33) yields  $Wf(u, s) = s^{n+1} g(u, s^2)$ . Thus, the wavelet transform is proportional to a heat diffusion with initial condition  $f^{(n)}$ .

The maximum principle for the parabolic heat equation [35] proves that a global maximum of  $|g(u, s^2)|$  for  $(u, s) \in [a, b] \times [s_0, s_1]$  is necessarily either on the boundary  $u = a, b$  or at  $s = s_0$ . A modulus maxima of  $Wf(u, s)$  at  $(u_1, s_1)$  is a local maxima of  $|g(u, s^2)|$  for a fixed  $s$  and varying  $u$ . Suppose that a line of modulus maxima is interrupted at  $(u_1, s_1)$ , with  $s_1 > 0$ . One can then verify that there exists  $\varepsilon > 0$  such that a global maximum of  $|g(u, s^2)|$  over  $[u_1 - \varepsilon, u_1 + \varepsilon] \times [s_1 - \varepsilon, s_1]$  is at  $(u_1, s_1)$ . This contradicts the maximum principle, and thus proves that all modulus maxima propagate toward finer scales. ■

Derivatives of Gaussians are most often used to guarantee that all maxima lines propagate up to the finest scales. Chaining together maxima into maxima lines is also a procedure for removing spurious modulus maxima created by numerical errors in regions where the wavelet transform is close to zero.

### Isolated Singularities

A wavelet transform may have a sequence of local maxima that converge to an abscissa  $v$  even though  $f$  is perfectly regular at  $v$ . This is the case of the maxima line of Figure 6.5 that converges to the abscissa  $v = 0.23$ . To detect singularities it is therefore not sufficient to follow the wavelet modulus maxima across scales. The Lipschitz regularity is calculated from the decay of the modulus maxima amplitude.

Let us suppose that for  $s < s_0$  all modulus maxima that converge to  $v$  are included in a cone

$$|u - v| \leq Cs. \quad (6.35)$$

This means that  $f$  does not have oscillations that accelerate in the neighborhood of  $v$ . The potential singularity at  $v$  is necessarily isolated. Indeed, we can derive from Theorem 6.5 that the absence of maxima outside the cone of influence implies that  $f$  is uniformly Lipschitz  $n$  in the neighborhood of any  $t \neq v$  with  $t \in (v - Cs_0, v + Cs_0)$ . The decay of  $|Wf(u, s)|$  in the neighborhood of  $v$  is controlled by the decay of the modulus maxima included in the cone  $|u - v| \leq Cs$ . Theorem 6.3 implies that  $f$  is uniformly Lipschitz  $\alpha$  in the neighborhood of  $v$  if and only if there exists  $A > 0$  such that each modulus maximum  $(u, s)$  in the cone (6.35) satisfies

$$|Wf(u, s)| \leq As^{\alpha+1/2}, \quad (6.36)$$

which is equivalent to

$$\log_2 |Wf(u, s)| \leq \log_2 A + \left(\alpha + \frac{1}{2}\right) \log_2 s. \quad (6.37)$$

Thus, the Lipschitz regularity at  $v$  is the maximum slope of  $\log_2 |Wf(u, s)|$  as a function of  $\log_2 s$  along the maxima lines converging to  $v$ .

In numerical calculations, the finest scale of the wavelet transform is limited by the resolution of the discrete data. From a sampling at intervals  $N^{-1}$ , Section 4.3.3 computes the discrete wavelet transform at scales  $s \geq \lambda N^{-1}$  where  $\lambda$  is large enough to avoid sampling coarsely the wavelets at the finest scale. The Lipschitz regularity  $\alpha$  of a singularity is then estimated by measuring the decay slope of  $\log_2 |Wf(u, s)|$  as a function of  $\log_2 s$  for  $2^j \geq s \geq \lambda N^{-1}$ . The largest scale  $2^j$  should be smaller than the distance between two consecutive singularities to avoid having other singularities influence the value of  $Wf(u, s)$ . The sampling interval  $N^{-1}$  must be small enough to measure  $\alpha$  accurately. The signal in Figure 6.5(a) is defined by  $N = 256$  samples. Figure 6.5(c) shows the decay of  $\log_2 |Wf(u, s)|$  along the maxima line converging to  $t = 0.05$ . It has slope  $\alpha + 1/2 \approx 1/2$  for  $2^{-4} \geq s \geq 2^{-6}$ . As expected,  $\alpha = 0$  because the signal is discontinuous at  $t = 0.05$ . Along the second maxima line converging to  $t = 0.42$  the slope is  $\alpha + 1/2 \approx 1$ , which indicates that the singularity is Lipschitz  $1/2$ .

When  $f$  is a function with singularities that are not isolated, finite resolution measurements are not sufficient to distinguish individual singularities. Section 6.4 describes a global approach that computes the singularity spectrum of multifractals by taking advantage of their self-similarity.

### **Smoothed Singularities**

The signal may have important variations that are infinitely continuously differentiable. For example, at the border of a shadow the gray level of an image varies quickly but is not discontinuous because of the diffraction effect. The smoothness of these transitions is modeled as a diffusion with a Gaussian kernel that has a variance that is measured from the decay of wavelet modulus maxima.

In the neighborhood of a sharp transition at  $v$ , we suppose that

$$f(t) = f_0 \star g_\sigma(t), \quad (6.38)$$

where  $g_\sigma$  is a Gaussian of variance  $\sigma^2$ :

$$g_\sigma(t) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left(\frac{-t^2}{2\sigma^2}\right). \quad (6.39)$$

If  $f_0$  has a Lipschitz  $\alpha$  singularity at  $v$  that is isolated and nonoscillating, it is uniformly Lipschitz  $\alpha$  in the neighborhood of  $v$ . For wavelets that are derivatives of Gaussians, Theorem 6.7 [367] relates the decay of the wavelet transform to  $\sigma$  and  $\alpha$ .

**Theorem 6.7.** Let  $\psi = (-1)^n \theta^{(n)}$  with  $\theta(t) = \lambda \exp(-t^2/(2\beta^2))$ . If  $f = f_0 \star g_\sigma$  and  $f_0$  is uniformly Lipschitz  $\alpha$  on  $[v-h, v+h]$ , then there exists  $A$  such that

$$\forall (u, s) \in [v-h, v+h] \times \mathbb{R}^+, \quad |Wf(u, s)| \leq A s^{\alpha+1/2} \left(1 + \frac{\sigma^2}{\beta^2 s^2}\right)^{-(n-\alpha)/2}. \quad (6.40)$$

**Proof.** The wavelet transform can be written as

$$Wf(u, s) = s^n \frac{d^n}{du^n} (f \star \bar{\theta}_s)(u) = s^n \frac{d^n}{du^n} (f_0 \star g_\sigma \star \bar{\theta}_s)(u). \quad (6.41)$$

Since  $\theta$  is a Gaussian, one can verify with a Fourier transform calculation that

$$\bar{\theta}_s \star g_\sigma(t) = \sqrt{\frac{s}{s_0}} \bar{\theta}_{s_0}(t) \quad \text{with} \quad s_0 = \sqrt{s^2 + \frac{\sigma^2}{\beta^2}}. \quad (6.42)$$

Inserting this result in (6.41) yields

$$Wf(u, s) = s^n \sqrt{\frac{s}{s_0}} \frac{d^n}{du^n} (f_0 \star \bar{\theta}_{s_0})(u) = \left(\frac{s}{s_0}\right)^{n+1/2} Wf_0(u, s_0). \quad (6.43)$$

Since  $f_0$  is uniformly Lipschitz  $\alpha$  on  $[v-h, v+h]$ , Theorem 6.3 proves that there exists  $A > 0$  such that

$$\forall (u, s) \in [v-h, v+h] \times \mathbb{R}^+, \quad |Wf_0(u, s)| \leq A s^{\alpha+1/2}. \quad (6.44)$$

Inserting this in (6.43) gives

$$|Wf(u, s)| \leq A \left(\frac{s}{s_0}\right)^{n+1/2} s_0^{\alpha+1/2}, \quad (6.45)$$

from which we derive (6.40) by inserting the expression (6.42) of  $s_0$ . ■

This theorem explains how the wavelet transform decay relates to the amount of diffusion of a singularity. At large scales  $s \gg \sigma/\beta$ , the Gaussian averaging is not “felt” by the wavelet transform that decays like  $s^{\alpha+1/2}$ . For  $s \leq \sigma/\beta$ , the variation of  $f$  at  $v$  is not sharp relative to  $s$  because of the Gaussian averaging. At these fine scales, the wavelet transform decays like  $s^{n+1/2}$  because  $f$  is  $C^\infty$ .

The parameters  $K$ ,  $\alpha$ , and  $\sigma$  are numerically estimated from the decay of the modulus maxima along the maxima curves that converge toward  $v$ . The variance  $\beta^2$  depends on the choice of wavelet and is known in advance. A regression is performed to approximate

$$\log_2 |Wf(u, s)| \approx \log_2(K) + \left(\alpha + \frac{1}{2}\right) \log_2 s - \frac{n-\alpha}{2} \log_2 \left(1 + \frac{\sigma^2}{\beta^2 s^2}\right).$$

Figure 6.6 gives the wavelet modulus maxima computed with a wavelet that is a second derivative of a Gaussian. The decay of  $\log_2 |Wf(u, s)|$  as a function of  $\log_2 s$  is given along several maxima lines corresponding to smoothed and nonsmoothed singularities. The wavelet is normalized so that  $\beta = 1$  and the diffusion scale is  $\sigma = 2^{-5}$ .

### 6.2.2 Dyadic Maxima Representation

Wavelet transform maxima carry the properties of sharp signal transitions and singularities. By recovering a signal approximation from these maxima, signal singularities can be modified or removed by processing the wavelet modulus maxima.

For fast numerical computations, the detection of wavelet transform maxima is limited to dyadic scales  $\{2^j\}_{j \in \mathbb{Z}}$ . Suppose that  $\psi$  is a dyadic wavelet, which means that there exist  $A > 0$  and  $B$  such that

$$\forall \omega \in \mathbb{R} - \{0\}, \quad A \leq \sum_{j=-\infty}^{+\infty} |\hat{\psi}(2^j \omega)|^2 \leq B. \quad (6.46)$$

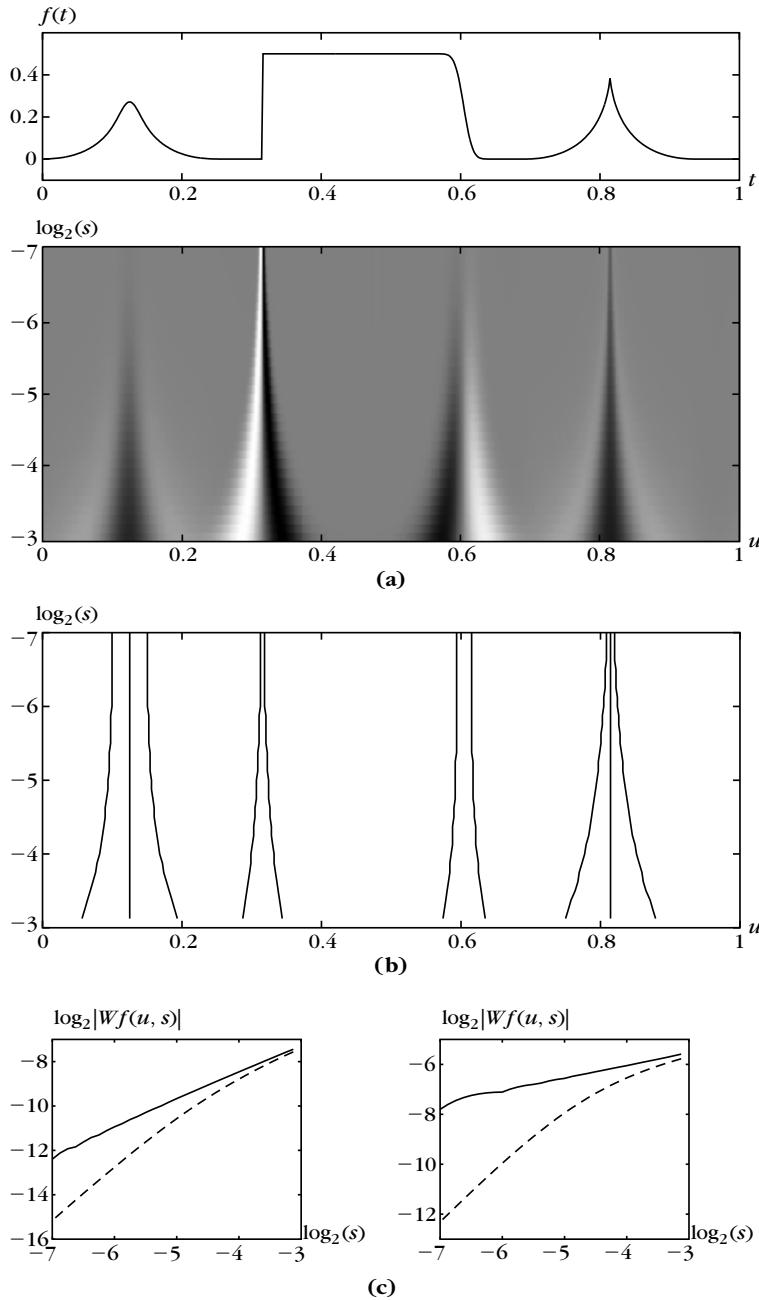
As a consequence of Theorem 5.11 on translation-invariant frames, it is proved in Section 5.2 that the resulting translation-invariant dyadic wavelet transform  $\{Wf(u, 2^j)\}_{j \in \mathbb{Z}}$  is complete and stable. This dyadic wavelet transform has the same properties as a continuous wavelet transform  $Wf(u, s)$ . All theorems of Sections 6.1.3 and 6.2 remain valid if we restrict  $s$  to the dyadic scales  $\{2^j\}_{j \in \mathbb{Z}}$ . Singularities create sequences of maxima that converge toward the corresponding location at fine scales, and the Lipschitz regularity is calculated from the decay of the maxima amplitude.

#### Scale-Space Maxima Support

Mallat and Zhong [367] introduced a dyadic wavelet maxima representation with a scale-space approximation support  $\Lambda$  of modulus maxima  $(u, 2^j)$  of  $Wf$ .

Wavelet maxima can be interpreted as points of 0 or  $\pi$  phase for an appropriate complex wavelet transform. Let  $\psi'$  be the derivative of  $\psi$  and  $\psi'_{u, 2^j}(t) = 2^{-j/2} \psi'(2^{-j}(t - u))$ . If  $Wf$  has a local extremum at  $u_0$ , then

$$\frac{\partial Wf(u_0, 2^j)}{\partial u} = -2^{-j} \langle f, \psi'_{2^j, u_0} \rangle = 0.$$



**FIGURE 6.6**

(a) Wavelet transform  $Wf(u, s)$ . (b) Modulus maxima of a wavelet transform computed  $\psi = \theta''$ , where  $\theta$  is a Gaussian with variance  $\beta = 1$ . (c) Decay of  $\log_2 |Wf(u, s)|$  along maxima curves. The solid and dotted lines (left) correspond to the maxima curves converging to  $t = 0.81$  and  $t = 0.12$ , respectively. They correspond to the curves (right) converging to  $t = 0.38$  and  $t = 0.55$ , respectively. The diffusion at  $t = 0.12$  and  $t = 0.55$  modifies the decay for  $s \leq \sigma = 2^{-5}$ .

Let us introduce a complex wavelet  $\psi^c(t) = \psi(t) + i\psi'(t)$ . If  $(u, s) \in \Lambda$ , then the resulting complex wavelet transform value is

$$W^c f(u, 2^j) = \langle f, \psi_{2^j, u}^c \rangle = \langle f, \psi_{2^j, u} \rangle + i \langle f, \psi'_{2^j, u} \rangle = Wf(u, s), \quad (6.47)$$

because  $\langle f, \psi'_{2^j, u} \rangle = 0$ . The complex wavelet value  $W^c f(u, s)$  has a phase equal to 0 or  $\pi$  depending on the sign of  $Wf(u, s)$ , and a modulus  $|W^c f(u, s)| = |Wf(u, s)|$ .

Figure 6.7(c) gives an example computed with the quadratic spline dyadic wavelet in Figure 5.3. This adaptive sampling of  $u$  produces a translation-invariant representation, which is important for pattern recognition. When  $f$  is translated by  $\tau$  each  $Wf(2^j, u)$  is translated by  $\tau$ , so the maxima support is translated by  $\tau$ , as illustrated by Figure 6.8. This is not the case for wavelet frame coefficients, where the translation parameter  $u$  is sampled with an interval proportional to the scale  $a^j$ , as explained in Section 5.3.

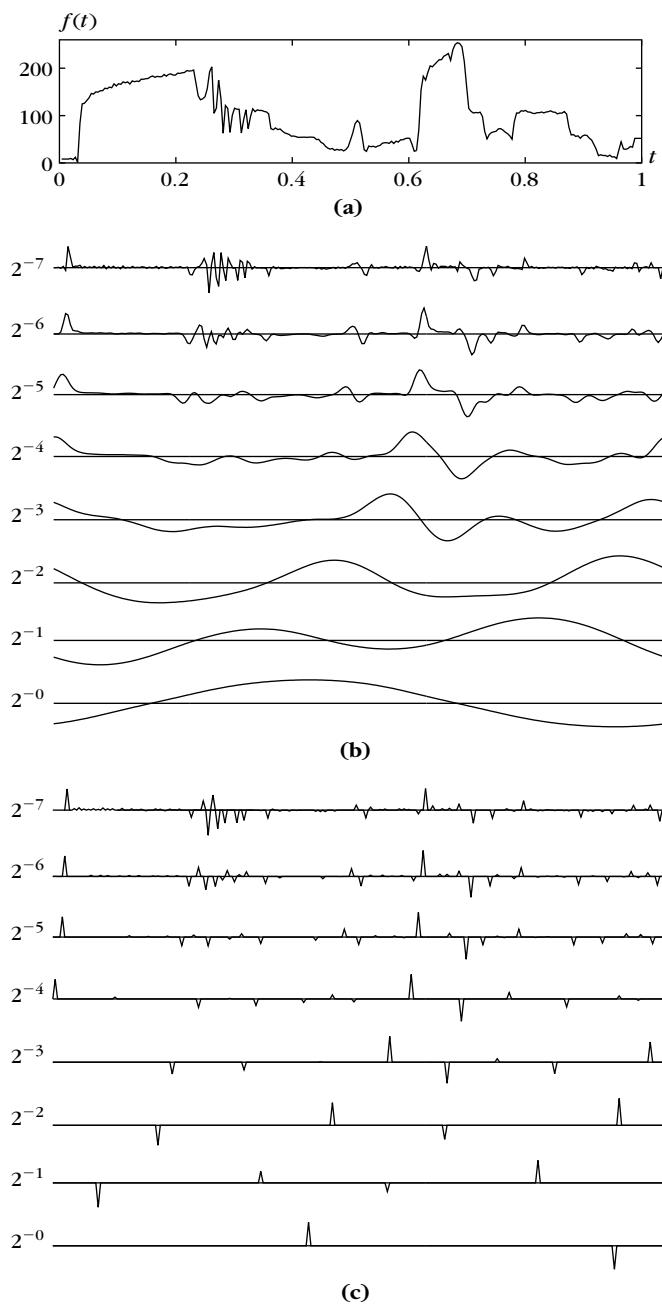
### Approximations from Wavelet Maxima

Mallat and Zhong [367] recover signal approximations from their wavelet maxima with an alternate projection algorithm, but several other algorithms have been proposed [150, 190, 286]. In the following we concentrate on orthogonal projection approximations on the space generated by wavelets in the scale-space maxima support. Numerical experiments show that dyadic wavelets of compact support recover signal approximations with a relative mean-square error that is typically of the order of  $10^{-2}$ .

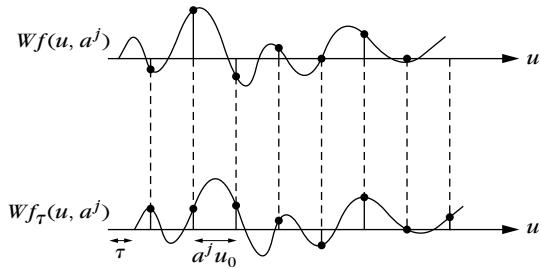
For general dyadic wavelets, Meyer [45] and Berman and Baras [107] proved that exact reconstruction is not possible. They found families of continuous or discrete signals having the same dyadic wavelet transforms and modulus maxima. However, signals with the same wavelet maxima differ from each other by small amplitude errors introducing no oscillation, which explains the success of numerical reconstructions [367]. If the signal has a band-limited Fourier transform and if  $\hat{\psi}$  has a compact support, then Kicey and Lennard [328] proved that wavelet modulus maxima define a complete and stable signal representation.

As a result of (6.47), the wavelet modulus maxima specifies the complex wavelet inner products  $\{\langle f, \psi_{u, 2^j}^c \rangle\}_{(u, 2^j) \in \Lambda}$ . Thus, a modulus maxima approximation can be computed as an orthogonal projection of  $f$  on the space generated by the complex wavelets  $\{\psi_{u, 2^j}^c\}_{(u, 2^j) \in \Lambda}$ . To reduce computations, the explicit extrema condition  $\langle \tilde{f}, \psi'_{u, 2^j} \rangle = 0$  is often removed, because it is indirectly almost obtained by calculating the orthogonal projection over the space  $\mathbf{V}_\Lambda$  generated by the real maxima wavelets  $\{\psi_{u, 2^j}\}_{(u, 2^j) \in \Lambda}$ . Section 5.1.3 shows that this orthogonal projection is obtained from the dual frame  $\{\tilde{\psi}_{u, 2^j}\}_{(u, 2^j) \in \Lambda}$  of  $\{\psi_{u, 2^j}\}_{(u, 2^j) \in \Lambda}$  in  $\mathbf{V}_\Lambda$ :

$$f_\Lambda = \sum_{(u, 2^j) \in \Lambda} \langle f, \psi_{u, 2^j} \rangle \tilde{\psi}_{u, 2^j}. \quad (6.48)$$

**FIGURE 6.7**

(a) Intensity variation along one row of the Lena image. (b) Dyadic wavelet transform computed at all scales  $2N^{-1} \leq 2^j \leq 1$ , with the quadratic spline wavelet  $\psi = -\theta'$  shown in Figure 5.3. (c) Modulus maxima of the dyadic wavelet transform.

**FIGURE 6.8**

If  $f_\tau(t) = f(t - \tau)$ , then  $Wf_\tau(u, a^j) = Wf(u - \tau, a^j)$ . Uniformly sampling  $Wf_\tau(u, a^j)$  and  $Wf(u, a^j)$  at  $u = na^j u_0$  may yield very different values if  $\tau \neq ku_0 a^j$ .

The dual-synthesis algorithm from Section 5.1.3 computes this orthogonal projection by inverting a symmetric operator  $L$  in  $\mathbb{V}_\Lambda$ :

$$Ly = \sum_{(u, 2^j) \in \Lambda} \langle y, \psi_{u, 2^j} \rangle \psi_{u, 2^j}, \quad (6.49)$$

with a conjugate gradient algorithm. Indeed  $f_\Lambda = L^{-1}(Lf)$ .

---

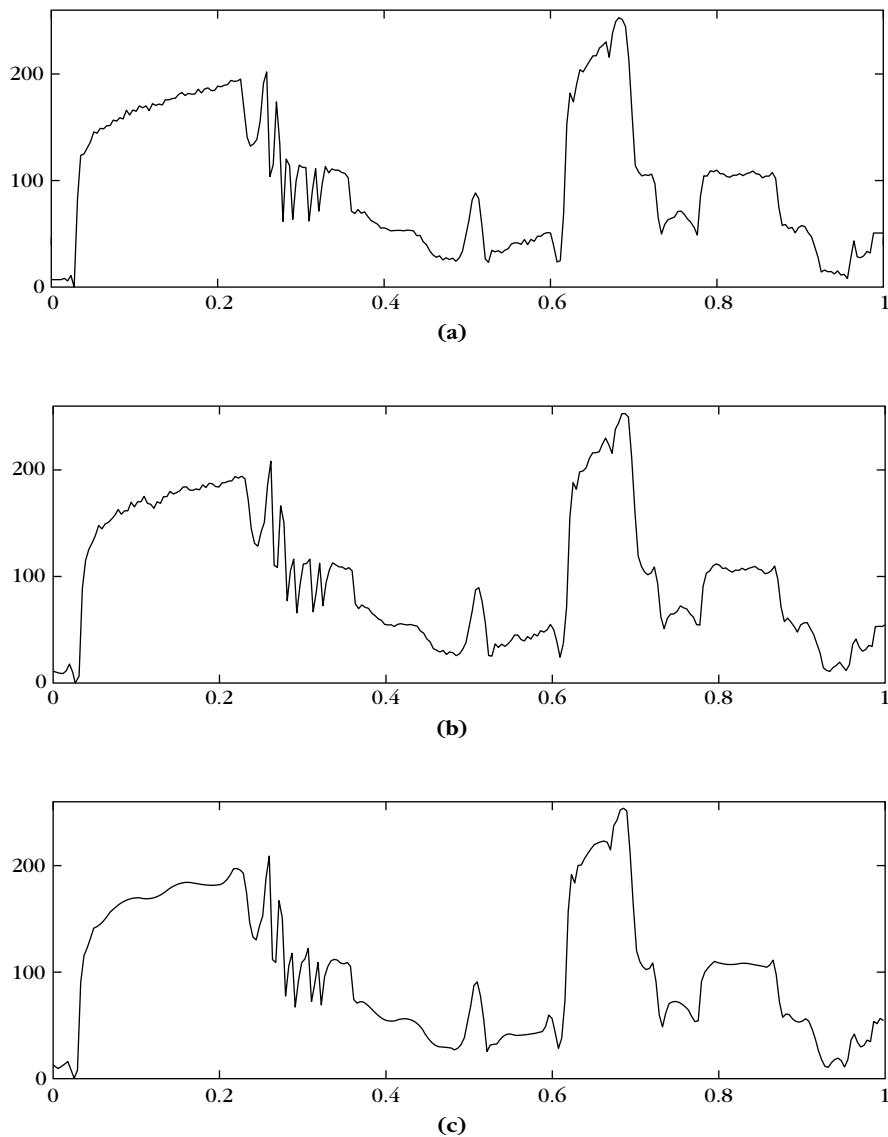
### EXAMPLE 6.1

Figure 6.9(b) shows the approximation  $f_\Lambda$ , recovered with 10 conjugate gradient iterations, from the wavelet maxima in Figure 6.7(c). This reconstruction is calculated with real instead of complex wavelets. After 20 iterations, the reconstruction error is  $\|f - \tilde{f}\| / \|f\| = 2.5 \cdot 10^{-2}$ . Figure 6.9(c) shows the signal reconstructed from 50% of the wavelet maxima that have the largest amplitude. Sharp signal transitions corresponding to large wavelet maxima have not been affected, but small texture variations disappear because the corresponding maxima are removed. The resulting signal is piecewise regular.

---

### *Fast Discrete Calculations*

The conjugate-gradient inversion of the operator (6.49) iterates on this operator many times. If there are many local maxima, it is more efficient to compute  $Wy(u, 2^j) = \langle y, \psi_{u, 2^j} \rangle$  for all  $(u, 2^j)$ , with the “algorithme à trous” (see Section 5.2.2). For a signal of size  $N$ , it cascades convolutions with two filters  $h[n]$  and  $g[n]$ , up to a maximum scale  $J = \log_2 N$ , with  $O(N \log_2 N)$  operations. All nonmaxima coefficients for  $(u, 2^j) \notin \Lambda$  are then set to zero. The reconstruction of  $Ly$  is computed by modifying the filter bank reconstruction given by Theorem 5.14, which also requires  $O(N \log_2 N)$  operations. The decomposition and reconstruction wavelets are the same in (6.49), so the reconstruction filters are  $\tilde{h}[n] = h[n]$  and  $\tilde{g}[n] = g[n]$ . The factor  $1/2$  in (5.72) is also removed because the reconstruction

**FIGURE 6.9**

**(a)** Original signal  $f$ . **(b)** Signal approximation  $f_L$  recovered from the dyadic wavelet maxima shown in Figure 6.7(c). **(c)** Approximation recovered from 50% largest maxima.

wavelets in (6.49) are not attenuated by  $2^{-j}$  as in a nonsampled wavelet reconstruction (5.50). For  $J = \log_2 N$ , we initialize  $\tilde{a}_J[n] = C/\sqrt{N}$  where  $C$  is the average signal value, and for  $\log_2 N > j \geq 0$  we compute

$$\tilde{a}_j[n] = \tilde{a}_{j+1} \star h_j[n] + \tilde{d}_{j+1} \star g_j[n]. \quad (6.50)$$

One can verify that  $Ly[n] = \tilde{a}_0[n]$  with the same derivations as in the proof of Theorem 5.14.

The signal approximations shown in Figure 6.9 are computed with the filters of Table 5.1. About 10 iterations of conjugate gradient are usually sufficient to recover an approximation with  $\|f_\Lambda - f\|/\|f\|$  of the order of  $10^{-2}$ , if all wavelet maxima are kept.

## 6.3 MULTISCALE EDGE DETECTION

Image edges are often important for pattern recognition. This is clearly illustrated by our visual ability to recognize an object from a drawing that gives a rough outline of contours. But, what is an edge? It could be defined as points where the image intensity has sharp transitions. A closer look shows that this definition is often not satisfactory. Image textures do have sharp intensity variations that are often not considered as edges. When looking at a brick wall, we may decide that the edges are the contours of the wall whereas the bricks define a texture. Alternatively, we may include the contours of each brick in the set of edges and consider the irregular surface of each brick as a texture. The discrimination of edges versus textures depends on the scale of analysis.

This has motivated computer vision researchers to detect sharp image variations at different scales [42, 416]. Section 6.3.1 describes the multiscale Canny edge detector [146]. It is equivalent to detecting modulus maxima in a two-dimensional dyadic wavelet transform [367]. Thus, the scale-space support of these modulus maxima correspond to multiscale edges. The Lipschitz regularity of edge points is derived from the decay of wavelet modulus maxima across scales. Image approximations are recovered with an orthogonal projection on the wavelets of the modulus maxima support with no visual degradation. Thus, image-processing algorithms can be implemented on multiscale edges.

### 6.3.1 Wavelet Maxima for Images

#### *Canny Edge Detection*

The Canny algorithm detects points of sharp variation in an image  $f(x_1, x_2)$  by calculating the modulus of its gradient vector

$$\vec{\nabla}f = \left( \frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2} \right). \quad (6.51)$$

The partial derivative of  $f$  in the direction of a unit vector  $\vec{n} = (\cos \alpha, \sin \alpha)$  in the  $x = (x_1, x_2)$  plane is calculated as an inner product with the gradient vector

$$\frac{\partial f}{\partial \vec{n}} = \vec{\nabla}f \cdot \vec{n} = \frac{\partial f}{\partial x_1} \cos \alpha + \frac{\partial f}{\partial x_2} \sin \alpha.$$

The absolute value of this partial derivative is maximum if  $\vec{n}$  is colinear to  $\vec{\nabla}f$ . This shows that  $\vec{\nabla}f(x)$  is parallel to the direction of maximum change of the surface

$f(x)$ . A point  $y \in \mathbb{R}^2$  is defined as an edge if  $|\vec{\nabla}f(y)|$  is locally maximum at  $x = y$  when  $x = y + \lambda \vec{\nabla}f(y)$  and  $|\lambda|$  is small enough. This means that the partial derivatives of  $f$  reach a local maximum at  $x = y$ , when  $x$  varies in a one-dimensional neighborhood of  $y$  along the direction of maximum change of  $f$  at  $y$ . These edge points are inflection points of  $f$ .

### Multiscale Edge Detection

A multiscale version of this edge detector is implemented by smoothing the surface with a convolution kernel  $\theta(x)$  that is dilated. This is computed with two wavelets that are the partial derivatives of  $\theta$ :

$$\psi^1 = -\frac{\partial \theta}{\partial x_1} \quad \text{and} \quad \psi^2 = -\frac{\partial \theta}{\partial x_2}. \quad (6.52)$$

The scale varies along the dyadic sequence  $\{2^j\}_{j \in \mathbb{Z}}$  to limit computations and storage. For  $1 \leq k \leq 2$ , we denote for  $x = (x_1, x_2)$ ,

$$\psi_{2^j}^k(x_1, x_2) = \frac{1}{2^j} \psi^k\left(\frac{x_1}{2^j}, \frac{x_2}{2^j}\right) \quad \text{and} \quad \bar{\psi}_{2^j}^k(x) = \psi_{2^j}^k(-x).$$

In the two directions indexed by  $1 \leq k \leq 2$ , the dyadic wavelet transform of  $f \in L^2(\mathbb{R}^2)$  at  $u = (u_1, u_2)$  is

$$W^k f(u, 2^j) = \langle f(x), \psi_{2^j}^k(x - u) \rangle = f \star \bar{\psi}_{2^j}^k(u). \quad (6.53)$$

Section 5.5 gives necessary and sufficient conditions for obtaining a complete and stable representation.

Let us denote  $\theta_{2^j}(x) = 2^{-j} \theta(2^{-j}x)$  and  $\bar{\theta}_{2^j}(x) = \theta_{2^j}(-x)$ . The two scaled wavelets can be rewritten as

$$\bar{\psi}_{2^j}^1 = 2^j \frac{\partial \bar{\theta}_{2^j}}{\partial x_1} \quad \text{and} \quad \bar{\psi}_{2^j}^2 = 2^j \frac{\partial \bar{\theta}_{2^j}}{\partial x_2}.$$

Thus, let us derive from (6.53) that the wavelet transform components are proportional to the coordinates of the gradient vector of  $f$  smoothed by  $\bar{\theta}_{2^j}$ :

$$\begin{pmatrix} W^1 f(u, 2^j) \\ W^2 f(u, 2^j) \end{pmatrix} = 2^j \begin{pmatrix} \frac{\partial}{\partial u_1} (f \star \bar{\theta}_{2^j})(u) \\ \frac{\partial}{\partial u_2} (f \star \bar{\theta}_{2^j})(u) \end{pmatrix} = 2^j \vec{\nabla} (f \star \bar{\theta}_{2^j})(u). \quad (6.54)$$

The modulus of this gradient vector is proportional to the wavelet transform modulus

$$Mf(u, 2^j) = \sqrt{|W^1 f(u, 2^j)|^2 + |W^2 f(u, 2^j)|^2}. \quad (6.55)$$

Let  $Af(u, 2^j)$  be the angle of the wavelet transform vector (6.54) in the plane  $(x_1, x_2)$ :

$$Af(u, 2^j) = \begin{cases} \alpha(u) & \text{if } W^1 f(u, 2^j) \geq 0 \\ \pi + \alpha(u) & \text{if } W^1 f(u, 2^j) < 0 \end{cases} \quad (6.56)$$

with

$$\alpha(u) = \tan^{-1} \left( \frac{W^2 f(u, 2^j)}{W^1 f(u, 2^j)} \right).$$

The unit vector  $\vec{n}_j(u) = (\cos A f(u, 2^j), \sin A f(u, 2^j))$  is colinear to  $\vec{\nabla}(f \star \bar{\theta}_{2^j})(u)$ . An edge point at the scale  $2^j$  is a point  $v$  such that  $Mf(u, 2^j)$  is locally maximum at  $u=v$  when  $u=v+\lambda \vec{n}_j(v)$  and  $|\lambda|$  is small enough. These points are also called wavelet transform *modulus maxima*. The smoothed image  $f \star \bar{\theta}_{2^j}$  has an inflection point at a modulus maximum location. Figure 6.10 gives an example where the wavelet modulus maxima are located along the contour of a circle.

### Maxima Curves

Edge points are distributed along curves that often correspond to the boundary of important structures. Individual wavelet modulus maxima are chained together to form a maxima curve that follows an edge. At any location, the tangent of the edge curve is approximated by computing the tangent of a level set. This tangent direction is used to chain wavelet maxima that are along the same edge curve.

The level sets of  $g(x)$  are the curves  $x(s)$  in the  $(x_1, x_2)$  plane where  $g(x(s))$  is constant. The parameter  $s$  is the arc-length of the level set. Let  $\vec{\tau} = (\tau_1, \tau_2)$  be the direction of the tangent of  $x(s)$ . Since  $g(x(s))$  is constant when  $s$  varies,

$$\frac{\partial g(x(s))}{\partial s} = \frac{\partial g}{\partial x_1} \tau_1 + \frac{\partial g}{\partial x_2} \tau_2 = \vec{\nabla}g \cdot \vec{\tau} = 0.$$

So,  $\vec{\nabla}g(x)$  is perpendicular to the direction  $\vec{\tau}$  of the tangent of the level set that goes through  $x$ .

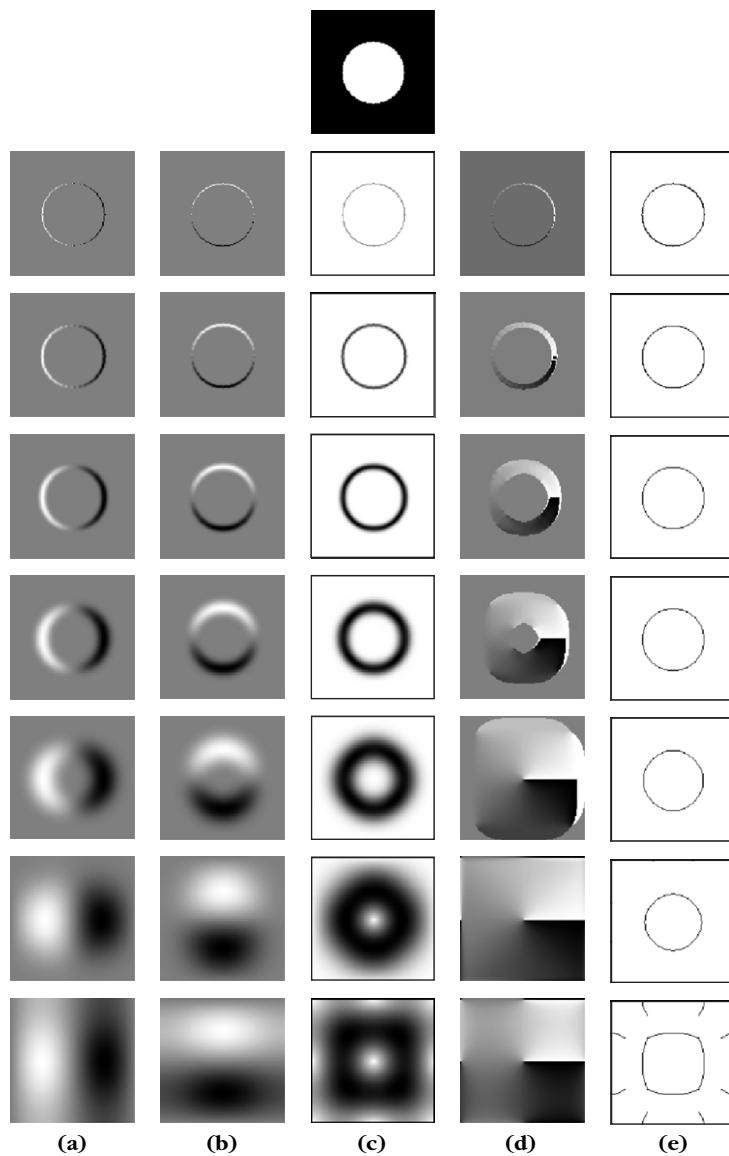
This level set property applied to  $g=f \star \bar{\theta}_{2^j}$  proves that at a maximum point  $v$  the vector  $\vec{n}_j(v)$  of angle  $A f(v, 2^j)$  is perpendicular to the level set of  $f \star \bar{\theta}_{2^j}$  going through  $v$ . If the intensity profile remains constant along an edge, then the inflection points (maxima points) are along a level set. The tangent of the maxima curve is therefore perpendicular to  $\vec{n}_j(v)$ . The intensity profile of an edge may not be constant but its variations are often negligible over a neighborhood of size  $2^j$  for a sufficiently small scale  $2^j$ , unless we are near a corner. The tangent of the maxima curve is then nearly perpendicular to  $\vec{n}_j(v)$ . In discrete calculations, maxima curves are recovered by chaining together any two wavelet maxima at  $v$  and  $v+\vec{n}$ , which are neighbors over the image sampling grid and such that  $\vec{n}$  is nearly perpendicular to  $\vec{n}_j(v)$ .

---

### EXAMPLE 6.2

The dyadic wavelet transform of the image in Figure 6.10 yields modulus images  $Mf(v, 2^j)$  with maxima along the boundary of a disk. This circular edge is also a level set of the image. Thus, the vector  $\vec{n}_j(v)$  of angle  $A f(v, 2^j)$  is perpendicular to the edge at the maxima locations.

---

**FIGURE 6.10**

The very top image has  $N = 128^2$  pixels. (a) Wavelet transform in the horizontal direction with a scale  $2^j$  that increases from top to bottom:  $\{W^1f(\mathbf{u}, 2^j)\}_{-6 \leq j \leq 0}$ ; black, gray, and white pixels correspond to negative, zero, and positive values, respectively. (b) Vertical direction:  $\{W^2f(\mathbf{u}, 2^j)\}_{-6 \leq j \leq 0}$ . (c) Wavelet transform modulus  $\{Mf(\mathbf{u}, 2^j)\}_{-6 \leq j \leq 0}$ ; white and black pixels correspond to zero and large-amplitude coefficients, respectively. (d) Angles  $\{Af(\mathbf{u}, 2^j)\}_{-6 \leq j \leq 0}$  at points where the modulus is nonzero. (e) The wavelet modulus maxima support is shown in black.

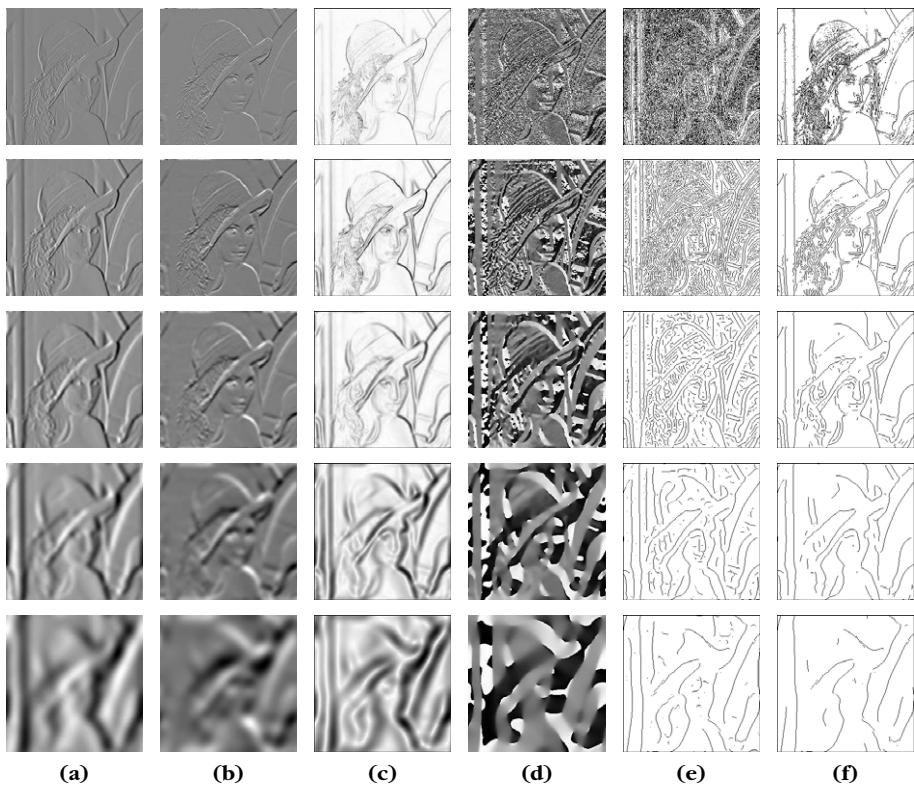


FIGURE 6.11

Multiscale edges of the Lena image shown in Figure 6.12. (a)  $\{W^1 f(u, 2^j)\}_{-7 \leq j \leq -3}$ . (b)  $\{W^2 f(u, 2^j)\}_{-7 \leq j \leq -3}$ . (c)  $\{Mf(u, 2^j)\}_{-7 \leq j \leq -3}$ . (d)  $\{Af(u, 2^j)\}_{-7 \leq j \leq -3}$ . (e) Modulus maxima support. (f) Support of maxima with modulus values above a threshold.

### EXAMPLE 6.3

In the Lena image shown in Figure 6.11, some edges disappear when the scale increases. These correspond to fine-scale intensity variations that are removed by the averaging with  $\bar{\theta}_{2^j}$  when  $2^j$  is large. This averaging also modifies the position of the remaining edges. Figure 6.11(f) displays the wavelet maxima such that  $Mf(v, 2^j) \geq T$  for a given threshold  $T$ . They indicate the location of edges where the image has large amplitude variations.

### Lipschitz Regularity

The decay of the two-dimensional wavelet transform depends on the regularity of  $f$ . We restrict the analysis to Lipschitz exponents  $0 \leq \alpha \leq 1$ . A function  $f$  is said to

be Lipschitz  $\alpha$  at  $v = (v_1, v_2)$  if there exists  $K > 0$  such that for all  $(x_1, x_2) \in \mathbb{R}^2$ ,

$$|f(x_1, x_2) - f(v_1, v_2)| \leq K(|x_1 - v_1|^2 + |x_2 - v_2|^2)^{\alpha/2}. \quad (6.57)$$

If there exists  $K > 0$  such that (6.57) is satisfied for any  $v \in \Omega$ , then  $f$  is uniformly Lipschitz  $\alpha$  over  $\Omega$ . As in one dimension, the Lipschitz regularity of a function  $f$  is related to the asymptotic decay  $|W^1 f(u, 2^j)|$  and  $|W^2 f(u, 2^j)|$  in the corresponding neighborhood. This decay is controlled by  $Mf(u, 2^j)$ . Like in Theorem 6.3, one can prove that  $f$  is uniformly Lipschitz  $\alpha$  inside a bounded domain of  $\mathbb{R}^2$  if and only if there exists  $A > 0$  such that for all  $u$  inside this domain and all scales  $2^j$ ,

$$|Mf(u, 2^j)| \leq A 2^{j(\alpha+1)}. \quad (6.58)$$

Suppose that the image has an isolated edge curve along which  $f$  has Lipschitz regularity  $\alpha$ . The value of  $|Mf(u, 2^j)|$  in a two-dimensional neighborhood of the edge curve can be bounded by the wavelet modulus values along the edge curve. The Lipschitz regularity  $\alpha$  of the edge is estimated with (6.58) by measuring the slope of  $\log_2 |Mf(u, 2^j)|$  as a function of  $j$ . If  $f$  is not singular but has a smooth transition along the edge, the smoothness can be quantified by the variance  $\sigma^2$  of a two-dimensional Gaussian blur. The value of  $\sigma^2$  is estimated by generalizing Theorem 6.7.

### **Reconstruction from Edges**

In his book about vision, Marr [42] conjectured that images can be reconstructed from multiscale edges. For a Canny edge detector, this is equivalent to recovering images from wavelet modulus maxima. Whether dyadic wavelet maxima define a complete and stable representation in two dimensions is still an open mathematical problem. However, the algorithm of Mallat and Zhong [367] recovers an image approximation that is visually identical to the original one. In the following, image approximations are computed by projecting the image on the space generated by wavelets on the modulus maxima support.

Let  $\Lambda$  be the set of all modulus maxima points  $(u, 2^j)$ . Let  $\vec{n}$  be the unit vector in the direction of  $Af(u, 2^j)$  and

$$\psi_{u, 2^j}^3(x) = 2^{2j} \frac{\partial^2 \theta_{2^j}(x - u)}{\partial \vec{n}^2}.$$

Since the wavelet transform modulus  $Mf(u, 2^j)$  has a local extremum at  $u$  in the direction of  $\vec{n}$ , it results that

$$\langle f, \psi_{u, 2^j}^3 \rangle = 0. \quad (6.59)$$

A modulus maxima representation provides the set of inner products  $\{\langle f, \psi_{u, 2^j}^k \rangle\}_{(u, 2^j) \in \Lambda, 1 \leq k \leq 3}$ . A modulus maxima approximation  $f_\Lambda$  can be computed as an orthogonal projection of  $f$  on the space generated by the family of maxima wavelets  $\{\psi_{u, 2^j}^k\}_{(u, 2^j) \in \Lambda, 1 \leq k \leq 3}$ .

To reduce computations, the condition on the third wavelets  $\langle f, \psi_{u,2j}^3 \rangle = 0$  is removed, because it is indirectly almost imposed by the orthogonal projection over the space  $\mathbf{V}_\Lambda$  generated by the other two wavelets for  $k = 1, 2$ . The dual-synthesis algorithm from Section 5.1.3 computes this orthogonal projection  $f_\Lambda$  by inverting a symmetric operator  $L$  in  $\mathbf{V}_\Lambda$ :

$$Ly = \sum_{(u,2j) \in \Lambda} \sum_{k=1}^2 \langle y, \psi_{u,2j}^k \rangle \psi_{u,2j}^k, \quad (6.60)$$

with a conjugate-gradient algorithm. Indeed  $f_\Lambda = L^{-1}(Lf)$ . When keeping all modulus maxima, the resulting modulus maxima approximation  $f_\Lambda$  satisfies  $\|f_\Lambda - f\|/\|f\| \leq 10^{-2}$ . Singularities and edges are nearly perfectly recovered and no spurious oscillations are introduced. The images differ slightly in smooth regions, but visually this is not noticeable.

---

#### EXAMPLE 6.4

The image reconstructed in Figure 6.12(b) is visually identical to the original image. It is recovered with 10 conjugate-gradient iterations. After 20 iterations, the relative mean-square reconstruction error is  $\|\tilde{f} - f\|/\|f\| = 4 \cdot 10^{-3}$ . The thresholding of maxima accounts for the disappearance of image structures from the reconstruction shown in Figure 6.12(c). Sharp image variations are recovered.

---

#### **Denoising by Multiscale Edge Thresholding**

Multiscale edge representations can be used to reduce additive noise. Denoising algorithms by thresholding wavelet coefficients are presented in Section 11.3.1. Block thresholding (see Section 11.4.2) regularizes this coefficient selection by regrouping them in square blocks. Similarly, a noise-removal algorithm can be implemented by thresholding multiscale wavelet maxima, while taking into account their geometric properties.

A simple approach implemented by Hwang and Mallat [364] chains the maxima into curves that are thresholded as a block. In Figure 6.13 noisy modulus maxima are shown on the second row and the third row displays the thresholded modulus maxima chains. At the finest scale shown on the left, the noise is masking the image structures. Maxima chains are selected by using the position of the selected maxima at the previous scale. An image approximation is recovered from the selected wavelet maxima. Edges are well-recovered visually but textures and fine structures are removed. This produces a cartoonlike image.

#### **Illusory Contours**

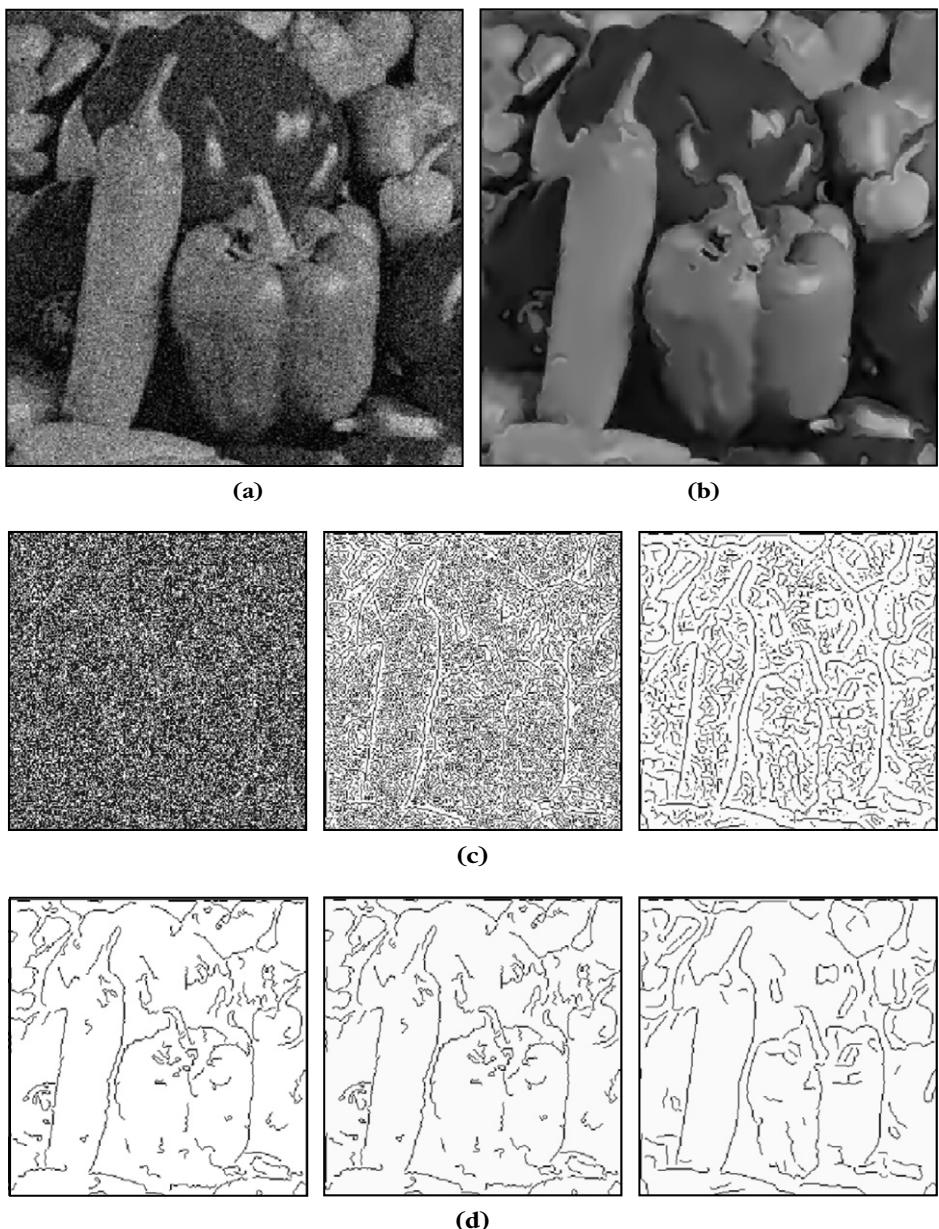
A multiscale wavelet edge detector defines edges as points where the image intensity varies sharply. However, this definition is too restrictive when edges are used to find the contours of objects. For image segmentation, edges must define closed curves

**FIGURE 6.12**

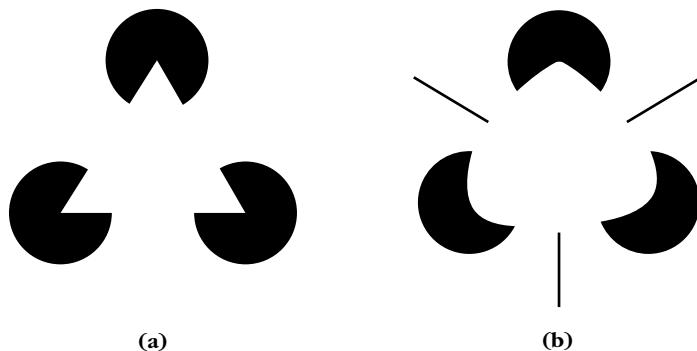
(a) Original Lena image. (b) Image reconstructed from the wavelet maxima displayed in Figure 6.11(e) and larger-scale maxima. (c) Image reconstructed from the thresholded wavelet maxima displayed in Figure 6.11(f) and larger-scale maxima.

that outline the boundaries of each region. Because of noise or light variations, local edge detectors produce contours with holes. Filling these holes requires some prior knowledge about the behavior of edges in the image. The illusion of the Kanizsa triangle [37] shows that such an edge filling is performed by the human visual system.

In Figure 6.14 one can “see” the edges of a straight and a curved triangle although the image gray level remains uniformly white between the black discs. Closing edge curves and understanding illusory contours requires computational models that are

**FIGURE 6.13**

**(a)** Noisy peppers image. **(b)** Peppers image restored from the thresholding maxima chains shown in **(d)**. The images in row **(c)** show the wavelet maxima support of the noisy image—the scale increases from left to right, from  $2^{-7}$  to  $2^{-5}$ . The images in row **(d)** give the maxima support computed with a thresholding selection of the maxima chains.

**FIGURE 6.14**

The illusory edges of a (a) straight and (b) curved triangle are perceived in domains where the images are uniformly white.

not as local as multiscale differential operators. Such contours can be obtained as the solution of a global optimization that incorporates constraints on the regularity of contours and takes into account the existence of occlusions [269].

### 6.3.2 Fast Multiscale Edge Computations

The dyadic wavelet transform of an image of  $N$  pixels is computed with a separable extension of the filter bank algorithm described in Section 5.2.2. A fast multiscale edge detection is derived [367].

#### *Wavelet Design*

Edge-detection wavelets (6.52) are designed as separable products of the one-dimensional dyadic wavelets constructed in Section 5.2.1. Their Fourier transform is

$$\hat{\psi}^1(\omega_1, \omega_2) = \hat{g}\left(\frac{\omega_1}{2}\right)\hat{\phi}\left(\frac{\omega_1}{2}\right)\hat{\phi}\left(\frac{\omega_2}{2}\right), \quad (6.61)$$

and

$$\hat{\psi}^2(\omega_1, \omega_2) = \hat{g}\left(\frac{\omega_2}{2}\right)\hat{\phi}\left(\frac{\omega_1}{2}\right)\hat{\phi}\left(\frac{\omega_2}{2}\right), \quad (6.62)$$

where  $\hat{\phi}(\omega)$  is a scaling function that has energy concentrated at low frequencies and

$$\hat{g}(\omega) = -i\sqrt{2} \sin\left(\frac{\omega}{2}\right) \exp\left(\frac{-i\omega}{2}\right). \quad (6.63)$$

This transfer function is the Fourier transform of a finite difference filter, which is a discrete approximation of a derivative

$$\frac{g[p]}{\sqrt{2}} = \begin{cases} -0.5 & \text{if } p=0 \\ 0.5 & \text{if } p=1 \\ 0 & \text{otherwise.} \end{cases} \quad (6.64)$$

The resulting wavelets  $\psi^1$  and  $\psi^2$  are finite difference approximations of partial derivatives along  $x_1$  and  $x_2$  of  $\theta(x_1, x_2) = 4\phi(2x_1)\phi(2x_2)$ .

To implement the dyadic wavelet transform with a filter bank algorithm, the scaling function  $\hat{\phi}$  is calculated, as in (5.60), with an infinite product:

$$\hat{\phi}(\omega) = \prod_{p=1}^{+\infty} \frac{\hat{h}(2^{-p}\omega)}{\sqrt{2}} = \frac{1}{\sqrt{2}} \hat{h}\left(\frac{\omega}{2}\right) \hat{\phi}\left(\frac{\omega}{2}\right). \quad (6.65)$$

The  $2\pi$  periodic function  $\hat{h}$  is the transfer function of a finite impulse-response low-pass filter  $h[p]$ . We showed in (5.61) that the Fourier transform of a box spline of degree  $m$ ,

$$\hat{\phi}(\omega) = \left( \frac{\sin(\omega/2)}{\omega/2} \right)^{m+1} \exp\left(\frac{-i\varepsilon\omega}{2}\right) \quad \text{with } \varepsilon = \begin{cases} 1 & \text{if } m \text{ is even} \\ 0 & \text{if } m \text{ is odd} \end{cases}$$

is obtained with

$$\hat{h}(\omega) = \sqrt{2} \frac{\hat{\phi}(2\omega)}{\hat{\phi}(\omega)} = \sqrt{2} \left( \cos \frac{\omega}{2} \right)^{m+1} \exp\left(\frac{-i\varepsilon\omega}{2}\right).$$

Table 5.1 gives  $h[p]$  for  $m = 2$ .

### Algorithme à Trou

The one-dimensional *algorithme à trous* (see Section 5.2.2) is extended in two dimensions with convolutions along the image rows and columns.

Each sample  $a_0[n]$  of the normalized discrete image is considered to be an average of the input analog image  $f$  calculated with the kernel  $\phi(x_1)\phi(x_2)$  translated at  $n = (n_1, n_2)$ :

$$a_0[n_1, n_2] = \langle f(x_1, x_2), \phi(x_1 - n_1)\phi(x_2 - n_2) \rangle.$$

This is further justified in Section 7.7.3. For any  $j \geq 0$ , we denote

$$a_j[n_1, n_2] = \langle f(x_1, x_2), \phi_{2^j}(x_1 - n_1)\phi_{2^j}(x_2 - n_2) \rangle.$$

The discrete wavelet coefficients at  $n = (n_1, n_2)$  are

$$d_j^1[n] = W^1 f(n, 2^j) \quad \text{and} \quad d_j^2[n] = W^2 f(n, 2^j).$$

They are calculated with separable convolutions.

For any  $j \geq 0$ , the filter  $h[p]$  “dilated” by  $2^j$  is defined by

$$\bar{h}_j[p] = \begin{cases} h[-p/2^j] & \text{if } p/2^j \in \mathbb{Z} \\ 0 & \text{otherwise;} \end{cases} \quad (6.66)$$

and for  $j > 0$ , a centered finite difference filter is defined by

$$\frac{\bar{g}_j[p]}{\sqrt{2}} = \begin{cases} 0.5 & \text{if } p = -2^{j-1} \\ -0.5 & \text{if } p = 2^{j-1} \\ 0 & \text{otherwise.} \end{cases} \quad (6.67)$$

For  $j=0$ , we define  $\bar{g}_0[0]/\sqrt{2} = -0.5$ ,  $\bar{g}_0[-1]/\sqrt{2} = -0.5$  and  $\bar{g}_0[p] = 0$  for  $p \neq 0, -1$ . A separable two-dimensional filter is written as

$$\alpha\beta[n_1, n_2] = \alpha[n_1]\beta[n_2],$$

and  $\delta[n]$  is a discrete Dirac. Similar to Theorem 5.14, one can prove that for any  $j \geq 0$  and any  $n = (n_1, n_2)$ ,

$$a_{j+1}[n] = a_j \star \bar{h}_j \bar{h}_j[n], \quad (6.68)$$

$$d_{j+1}^1[n] = a_j \star \bar{g}_j \delta[n], \quad (6.69)$$

$$d_{j+1}^2[n] = a_j \star \delta \bar{g}_j[n]. \quad (6.70)$$

Dyadic wavelet coefficients up to the scale  $2^J$  are therefore calculated by cascading the convolutions (6.68–6.70) for  $0 < j \leq J$ . To take into account border problems, all convolutions are replaced by circular convolutions, which means that the input image  $a_0[n]$  is considered to be periodic along its rows and columns. For an image of  $N$  pixels, this algorithm requires  $O(N \log_2 N)$  operations. For a square image with a maximum scale  $J = \log_2 N^{1/2}$ , one can verify that the larger-scale approximation is a constant proportional to the gray-level average  $C$ :

$$a_J[n_1, n_2] = N^{-1/2} \sum_{n_1, n_2=0}^{N^{1/2}-1} a_0[n_1, n_2] = N^{1/2} C.$$

The wavelet transform modulus is  $Mf(n, 2^j) = |d_j^1[n]|^2 + |d_j^2[n]|^2$ , whereas  $Af(n, 2^j)$  is the angle of the vector  $(d_j^1[n], d_j^2[n])$ .

The support  $\Lambda$  of wavelet modulus maxima  $(u, 2^j)$  is the set of points  $Mf(u, 2^j)$ , which is larger than its two neighbors  $Mf(u \pm \vec{\varepsilon}, 2^j)$ , where  $\vec{\varepsilon} = (\varepsilon_1, \varepsilon_2)$  is the vector with coordinates  $\varepsilon_1$  and  $\varepsilon_2$  that are either 0 or 1 and have an angle that is the closest to  $Af(u, 2^j)$ .

### **Reconstruction from Maxima**

The orthogonal projection from wavelet maxima is computed with the dual-synthesis algorithm from Section 5.1.3, which inverts the symmetric operator (6.60) with conjugate-gradient iterations. This requires computing  $Ly$  efficiently for any image  $y[n]$ . For this purpose, the wavelet coefficients of  $y$  are first calculated with the *algorithme à trous*, and all coefficients for  $(u, 2^j) \notin \Lambda$  are set to 0. The signal  $Ly[n]$  is recovered from these nonzero wavelet coefficients. Let  $h_j[n] = \bar{h}_j[-n]$  and  $g_j[n] = \bar{g}_j[-n]$  be the two filters defined with (6.66) and (6.67). The calculation is initialized for  $J = \log_2 N^{1/2}$  by setting  $\tilde{a}_J[n] = CN^{-1/2}$ , where  $C$  is the average image intensity. For  $\log_2 N > j \geq 0$ , we compute

$$\tilde{a}_j[n] = \tilde{a}_{j+1} \star h_j h_j[n] + d_{j+1}^1 \star g_j \delta[n] + d_{j+1}^2 \star \delta g_j[n],$$

and one can verify that  $Ly[n] = \tilde{a}_0[n]$  is recovered with  $O(N \log_2 N)$  operations. The reconstructed images that were shown in Figure 6.12 are obtained with 10 conjugate-gradient iterations implemented with this filter bank algorithm.

## 6.4 MULTIFRACTALS

Signals that are singular at almost every point were originally studied as pathological objects of pure mathematical interest. Mandelbrot [41] was the first to recognize that such phenomena are encountered everywhere. Among the many examples [25] are economic records such as the Dow Jones industrial average, physiological data including heart records, electromagnetic fluctuations in galactic radiation noise, textures in images of natural terrains, variations of traffic flow, and so on.

The singularities of multifractals often vary from point to point, and knowing the distribution of these singularities is important in analyzing their properties. Pointwise measurements of Lipschitz exponents are not possible because of the finite numerical resolution. After discretization, each sample corresponds to a time interval where the signal has an infinite number of singularities that may all be different. The singularity distribution must therefore be estimated from global measurements that take advantage of multifractal self-similarities. Section 6.4.2 computes the fractal dimension of sets of points having the same Lipschitz regularity, with a global partition function calculated from wavelet transform modulus maxima. Applications to fractal noises, such as fractional Brownian motions and hydrodynamic turbulence, are studied in Section 6.4.3.

### 6.4.1 Fractal Sets and Self-Similar Functions

A set  $\mathcal{S} \subset \mathbb{R}^n$  is said to be self-similar if it is the union of disjoint subsets  $\mathcal{S}_1, \dots, \mathcal{S}_k$  that can be obtained from  $\mathcal{S}$  with a scaling, translation, and rotation. This self-similarity often implies an infinite multiplication of details, which creates irregular structures. The triadic Cantor set and the Von Koch curve are simple examples.

---

#### EXAMPLE 6.5

The Von Koch curve is a fractal set obtained by recursively dividing each segment of length  $I$  in four segments of length  $I/3$ , as illustrated in Figure 6.15. Each subdivision multiplies the length by  $4/3$ ; therefore, the limit of these subdivisions is a curve of infinite length.

---

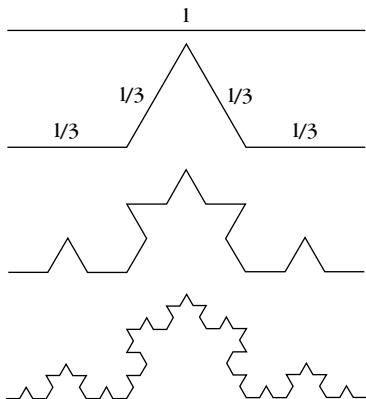
#### EXAMPLE 6.6

The triadic Cantor set is constructed by recursively dividing intervals of size  $I$  in two subintervals of size  $I/3$  and a central hole, illustrated in Figure 6.16. The iteration begins from  $[0, 1]$ . The Cantor set obtained as a limit of these subdivisions is a dust of points in  $[0, 1]$ .

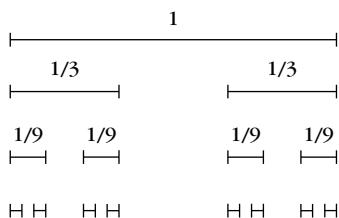
---

#### *Fractal Dimension*

The Von Koch curve has infinite length in a finite square of  $\mathbb{R}^2$ ; therefore, the usual length measurement is not well adapted to characterize the topological properties of such fractal curves. This motivated Hausdorff in 1919 to introduce a new definition

**FIGURE 6.15**

Three iterations of the Von Koch subdivision. The Von Koch curve is the fractal obtained as a limit of an infinite number of subdivisions.

**FIGURE 6.16**

Three iterations of the Cantor subdivision of  $[0, 1]$ . The limit of an infinite number of subdivisions is a closed set in  $[0, 1]$ .

of dimension—the *capacity dimension*—based on the size variations of sets when measured at different scales.

The capacity dimension is a simplification of the Hausdorff dimension that is easier to compute numerically. Let  $\mathcal{S}$  be a bounded set in  $\mathbb{R}^n$ . We count the minimum number  $N(s)$  of balls of radius  $s$  needed to cover  $\mathcal{S}$ . If  $\mathcal{S}$  is a set of dimension  $D$  with a finite length ( $D = 1$ ), surface ( $D = 2$ ), or volume ( $D = 3$ ), then

$$N(s) \sim s^{-D},$$

so

$$D = -\lim_{s \rightarrow 0} \frac{\log N(s)}{\log s}. \quad (6.71)$$

The capacity dimension  $D$  of  $\mathcal{S}$  generalizes this result and is defined by

$$D = -\liminf_{s \rightarrow 0} \frac{\log N(s)}{\log s}. \quad (6.72)$$

The measure of  $\mathcal{S}$  is then

$$M = \limsup_{s \rightarrow 0} N(s) s^D.$$

It may be finite or infinite.

The Hausdorff dimension is a refined fractal measure that considers all covers of  $\mathcal{S}$  with balls of radius smaller than  $s$ . It is most often, but not always, equal to the capacity dimension. In the following examples, the capacity dimension is called *fractal dimension*.

### EXAMPLE 6.7

The Von Koch curve has infinite length because its fractal dimension is  $D > 1$ . We need  $N(s) = 4^n$  balls of size  $s = 3^{-n}$  to cover the whole curve, thus,

$$N(3^{-n}) = (3^{-n})^{-\log 4 / \log 3}.$$

One can verify that at any other scale  $s$ , the minimum number of balls  $N(s)$  to cover this curve satisfies

$$D = -\liminf_{s \rightarrow 0} \frac{\log N(s)}{\log s} = \frac{\log 4}{\log 3}.$$

As expected, it has a fractal dimension between 1 and 2.

### EXAMPLE 6.8

The triadic Cantor set is covered by  $N(s) = 2^n$  intervals of size  $s = 3^{-n}$ , so

$$N(3^{-n}) = (3^{-n})^{-\log 2 / \log 3}.$$

One can also verify that

$$D = -\liminf_{s \rightarrow 0} \frac{\log N(s)}{\log s} = \frac{\log 2}{\log 3}.$$

### *Self-Similar Functions*

Let  $f$  be a continuous function with a compact support  $\mathcal{S}$ . We say that  $f$  is *self-similar* if there exist disjoint subsets  $\mathcal{S}_1, \dots, \mathcal{S}_k$  such that the graph of  $f$  restricted to each  $\mathcal{S}_i$  is an affine transformation of  $f$ . This means that there exist a scale  $l_i > 1$ , a translation  $r_i$ , a weight  $p_i$ , and a constant  $c_i$  such that

$$\forall t \in \mathcal{S}_i, \quad f(t) = c_i + p_i f(l_i(t - r_i)). \quad (6.73)$$

Outside these subsets, we suppose that  $f$  is constant. Generalizations of this definition can also be used [128].

If a function is self-similar then its wavelet transform is also self-similar. Let  $g$  be an affine transformation of  $f$ :

$$g(t) = p f(l(t - r)) + c. \quad (6.74)$$

Its wavelet transform is

$$Wg(u, s) = \int_{-\infty}^{+\infty} g(t) \frac{1}{\sqrt{s}} \psi\left(\frac{t-u}{s}\right) dt.$$

With the change of variable  $t' = l(t - r)$ , since  $\psi$  has a zero average, the affine relation (6.74) implies

$$Wg(u, s) = \frac{p}{\sqrt{l}} Wf(l(u - r), sl).$$

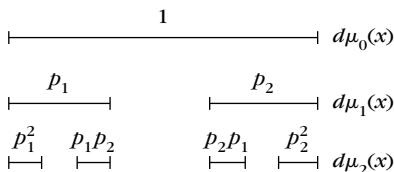
Suppose that  $\psi$  has a compact support included in  $[-K, K]$ . The affine invariance (6.73) of  $f$  over  $S_i = [a_i, b_i]$  produces an affine invariance for all wavelets having a support included in  $S_i$ . For any  $s < (b_i - a_i)/K$  and any  $u \in [a_i + Ks, b_i - Ks]$ ,

$$Wf(u, s) = \frac{p_i}{\sqrt{l_i}} Wf(l_i(u - r_i), sl_i).$$

The wavelet transform's self-similarity implies that the positions and values of its modulus maxima are also self-similar. This can be used to recover unknown affine-invariance properties with a voting procedure based on wavelet modulus maxima [310].

### EXAMPLE 6.9

A Cantor measure is constructed over a Cantor set. Let  $d\mu_0(x) = dx$  be the uniform Lebesgue measure on  $[0, 1]$ . As in the Cantor set construction, this measure is subdivided into three uniform measures over  $[0, 1/3]$ ,  $[1/3, 2/3]$ , and  $[2/3, 1]$  with integrals equal to  $p_1$ , 0, and  $p_2$ , respectively. We impose  $p_1 + p_2 = 1$  to obtain a total measure  $d\mu_1$  on  $[0, 1]$  with an integral equal to 1. This operation is iteratively repeated by dividing each uniform measure of integral  $p$  over  $[a, a+1]$  into three equal parts where the integrals are  $p_1p$ , 0, and  $p_2p$ , respectively, over  $[a, a+1/3]$ ,  $[a+1/3, a+2/3]$ , and  $[a+2/3, a+1]$ . This is illustrated in Figure 6.17. After each subdivision, the resulting measure  $d\mu_n$  has a unit integral. In the limit, we obtain a Cantor measure  $d\mu_\infty$  of unit integral with a support that is the triadic Cantor set.



**FIGURE 6.17**

Two subdivisions of the uniform measure on  $[0, 1]$  with left and right weights  $p_1$  and  $p_2$ . The Cantor measure  $d\mu_\infty$  is the limit of an infinite number of these subdivisions.

---

**EXAMPLE 6.10**

A devil's staircase is the integral of a Cantor measure:

$$f(t) = \int_0^t d\mu_\infty(x). \quad (6.75)$$

It is a continuous function that increases from 0 to 1 on  $[0, 1]$ . The recursive construction of the Cantor measure implies that  $f$  is self-similar:

$$f(t) = \begin{cases} p_1 f(3t) & \text{if } t \in [0, 1/3] \\ p_1 & \text{if } t \in [1/3, 2/3] \\ p_1 + p_2 f(3t - 2) & \text{if } t \in [2/3, 0]. \end{cases}$$

Figure 6.18 displays the devil's staircase obtained with  $p_1 = p_2 = 0.5$ . The wavelet transform in (b) is calculated with a wavelet that is the first derivative of a Gaussian. The self-similarity of  $f$  yields a wavelet transform and modulus maxima that are self-similar. The subdivision of each interval in three parts appears through the multiplication by 2 maxima lines when the scale is multiplied by 3. This Cantor construction is generalized with different interval subdivisions and weight allocations beginning from the same Lebesgue measure  $d\mu_0$  on  $[0, 1]$  [5].

---

### 6.4.2 Singularity Spectrum

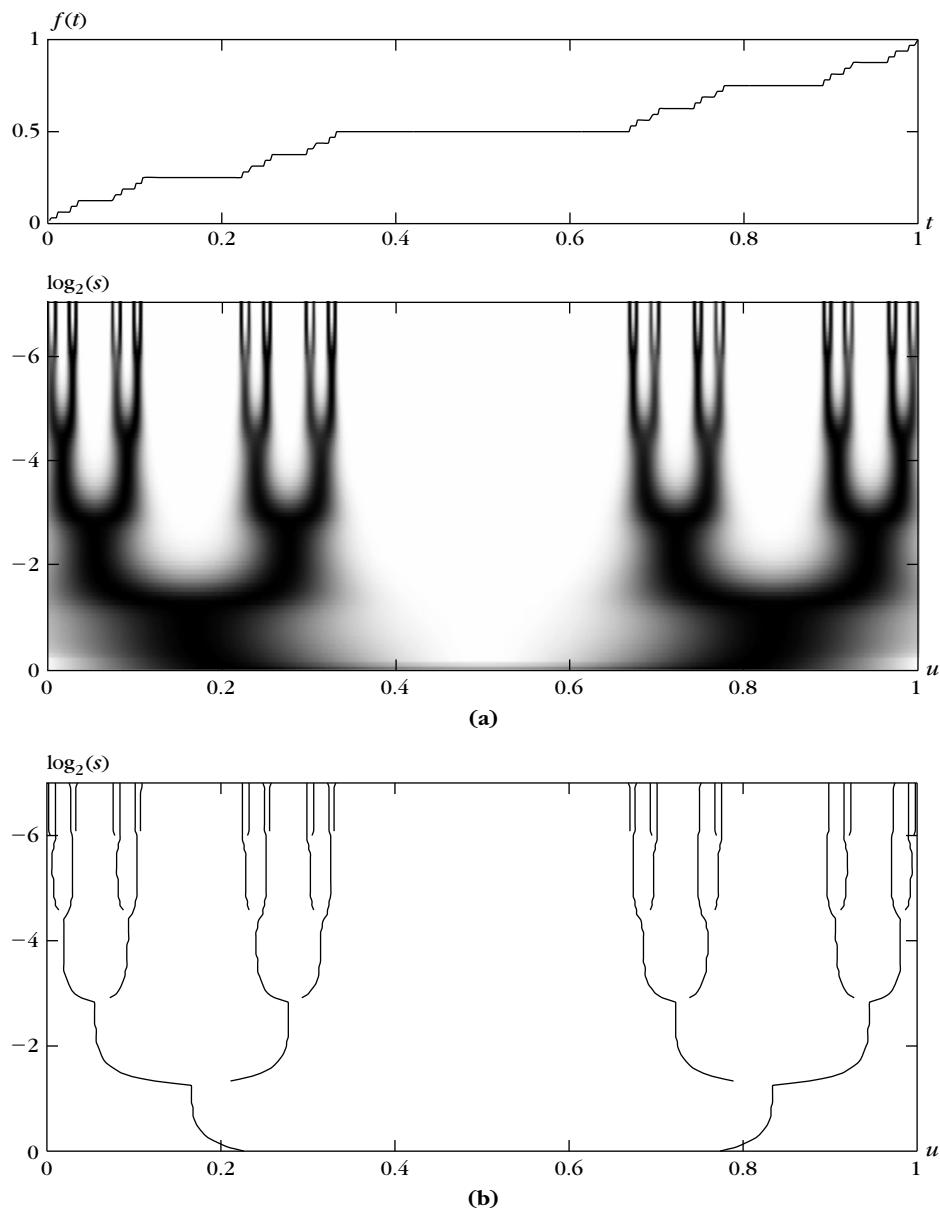
Finding the distribution of singularities in a multifractal signal  $f$  is particularly important for analyzing its properties. The spectrum of singularity measures the global repartition of singularities having different Lipschitz regularity. The pointwise Lipschitz regularity of  $f$  is given by Definition 6.1.

**Definition 6.1:** *Spectrum.* Let  $\mathcal{S}_\alpha$  be the set of all points  $t \in \mathbb{R}$  where the pointwise Lipschitz regularity of  $f$  is equal to  $\alpha$ . The spectrum of singularity  $D(\alpha)$  of  $f$  is the fractal dimension of  $\mathcal{S}_\alpha$ . The support of  $D(\alpha)$  is the set of  $\alpha$  such that  $\mathcal{S}_\alpha$  is not empty.

This spectrum was originally introduced by Frisch and Parisi [264] to analyze the homogeneity of multifractal measures that model the energy dissipation of turbulent fluids. It was then extended by Arneodo, Bacry, and Muzy [381] to multifractal signals. The fractal dimension definition (6.72) shows that if we make a disjoint cover of the support of  $f$  with intervals of size  $s$ , then the number of intervals that intersect  $\mathcal{S}_\alpha$  is

$$N_\alpha(s) \sim s^{-D(\alpha)}. \quad (6.76)$$

The singularity spectrum gives the proportion of Lipschitz  $\alpha$  singularities that appear at any scale  $s$ . A multifractal  $f$  is said to be homogeneous if all singularities have the same Lipschitz exponent  $\alpha_0$ , which means the support of  $D(\alpha)$  is restricted to  $\{\alpha_0\}$ . Fractional Brownian motions are examples of homogeneous multifractals.

**FIGURE 6.18**

Devil's staircase calculated from a Cantor measure with equal weights  $p_1 = p_2 = 0.5$ .

(a) Wavelet transform  $Wf(u, s)$  computed with  $\psi = -\theta'$  where  $\theta$  is Gaussian. (b) Wavelet transform modulus maxima.

### Partition Function

One cannot compute the pointwise Lipschitz regularity of a multifractal because its singularities are not isolated, and the finite numerical resolution is not sufficient to discriminate them. It is possible, however, to measure the singularity spectrum of multifractals from the wavelet transform local maxima using a global partition function introduced by Arneodo, Bacry, and Muzy [381].

Let  $\psi$  be a wavelet with  $n$  vanishing moments. Theorem 6.5 proves that if  $f$  has pointwise Lipschitz regularity  $\alpha_0 < n$  at  $v$ , then the wavelet transform  $Wf(u, s)$  has a sequence of modulus maxima that converges toward  $v$  at fine scales. Thus, the set of maxima at the scale  $s$  can be interpreted as a covering of the singular support of  $f$  with wavelets of scale  $s$ . At these maxima locations,

$$|Wf(u, s)| \sim s^{\alpha_0 + 1/2}.$$

Let  $\{u_p(s)\}_{p \in \mathbb{Z}}$  be the position of all local maxima of  $|Wg(u, s)|$  at a fixed scale  $s$ . The partition function  $\mathcal{Z}$  measures the sum at a power  $q$  of all these wavelet modulus maxima:

$$\mathcal{Z}(q, s) = \sum_p |Wf(u_p, s)|^q. \quad (6.77)$$

At each scale  $s$ , any two consecutive maxima  $u_p$  and  $u_{p+1}$  are supposed to have a distance  $|u_{p+1} - u_p| > \varepsilon s$ , for some  $\varepsilon > 0$ . If not, over intervals of size  $\varepsilon s$ , the sum (6.77) includes only the maxima of largest amplitude. This protects the partition function from the multiplication of very close maxima created by fast oscillations.

For each  $q \in \mathbb{R}$ , the scaling exponent  $\tau(q)$  measures the asymptotic decay of  $\mathcal{Z}(q, s)$  at fine scales  $s$ :

$$\tau(q) = \liminf_{s \rightarrow 0} \frac{\log \mathcal{Z}(q, s)}{\log s}.$$

This typically means that

$$\mathcal{Z}(q, s) \sim s^{\tau(q)}.$$

### Legendre Transform

Theorem 6.8 relates  $\tau(q)$  to the Legendre transform of  $D(\alpha)$  for self-similar signals. This result was established in [91] for a particular class of fractal signals and generalized by Jaffard [313].

**Theorem 6.8:** *Arneodo, Bacry, Jaffard, Muzy.* Let  $\Lambda = [\alpha_{\min}, \alpha_{\max}]$  be the support of  $D(\alpha)$ . Let  $\psi$  be a wavelet with  $n > \alpha_{\max}$  vanishing moments. If  $f$  is a self-similar signal, then

$$\tau(q) = \min_{\alpha \in \Lambda} (q(\alpha + 1/2) - D(\alpha)). \quad (6.78)$$

**Proof.** The detailed proof is long; we only give an intuitive justification. The sum (6.77) over all maxima positions is replaced by an integral over the Lipschitz parameter. At the scale  $s$ ,

(6.76) indicates that the density of modulus maxima that cover a singularity with Lipschitz exponent  $\alpha$  is proportional to  $s^{-D(\alpha)}$ . At locations where  $f$  has Lipschitz regularity  $\alpha$ , the wavelet transform decay is approximated by

$$|Wf(u, s)| \sim s^{\alpha+1/2}.$$

It follows that

$$\mathcal{Z}(q, s) \sim \int_{\Lambda} s^{q(\alpha+1/2)} s^{-D(\alpha)} d\alpha.$$

When  $s$  goes to 0 we derive that  $\mathcal{Z}(q, s) \sim s^{\tau(q)}$  for  $\tau(q) = \min_{\alpha \in \Lambda}(q(\alpha + 1/2) - D(\alpha))$ . ■

This theorem proves that the scaling exponent  $\tau(q)$  is the Legendre transform of  $D(\alpha)$ . It is necessary to use a wavelet with enough vanishing moments to measure all Lipschitz exponents up to  $\alpha_{\max}$ . In numerical calculations  $\tau(q)$  is computed by evaluating the sum  $\mathcal{Z}(q, s)$ . Thus, we need to invert the Legendre transform (6.78) to recover the spectrum of singularity  $D(\alpha)$ .

### Theorem 6.9.

- The scaling exponent  $\tau(q)$  is a concave and increasing function of  $q$ .
- The Legendre transform (6.78) is invertible if and only if  $D(\alpha)$  is concave, in which case

$$D(\alpha) = \min_{q \in \mathbb{R}} \left( q(\alpha + 1/2) - \tau(q) \right). \quad (6.79)$$

- The spectrum  $D(\alpha)$  of self-similar signals is concave.

**Proof.** The proof that  $D(\alpha)$  is concave for self-similar signals can be found in [313]. We concentrate on the properties of the Legendre transform that are important in numerical calculations. To simplify the proof, let us suppose that  $D(q)$  is twice differentiable. The minimum of the Legendre transform (6.78) is reached at a critical point  $q(\alpha)$ . Computing the derivative of  $q(\alpha + 1/2) - D(\alpha)$  with respect to  $\alpha$  gives

$$q(\alpha) = \frac{dD}{d\alpha}, \quad (6.80)$$

with

$$\tau(q) = q \left( \alpha + \frac{1}{2} \right) - D(\alpha). \quad (6.81)$$

Since it is a minimum, the second derivative of  $\tau(q(\alpha))$  with respect to  $\alpha$  is positive, from which we derive that

$$\frac{d^2 D(\alpha(q))}{d\alpha^2} \leq 0.$$

This proves that  $\tau(q)$  depends only on the values where  $D(\alpha)$  has a negative second derivative. Thus, we can recover  $D(\alpha)$  from  $\tau(q)$  only if it is concave.

The derivative of  $\tau(q)$  is

$$\frac{d\tau(q)}{dq} = \alpha + \frac{1}{2} + q \frac{d\alpha}{dq} - \frac{d\alpha}{dq} \frac{dD(\alpha)}{d\alpha} = \alpha + \frac{1}{2} \geq 0. \quad (6.82)$$

Therefore, it is increasing. Its second derivative is

$$\frac{d^2\tau(q)}{dq^2} = \frac{d\alpha}{dq}.$$

Taking the derivative of (6.80) with respect to  $q$  proves that

$$\frac{d\alpha}{dq} \frac{d^2D(\alpha)}{d\alpha^2} = 1.$$

Since  $\frac{d^2D(\alpha)}{d\alpha^2} \leq 0$ , we derive that  $\frac{d^2\tau(q)}{dq^2} \leq 0$ . Thus,  $\tau(q)$  is concave. By using (6.81), (6.82), and the fact that  $\tau(q)$  is concave, we verify that

$$D(\alpha) = \min_{q \in \mathbb{R}} \left( q(\alpha + 1/2) - \tau(q) \right).$$

■

The spectrum  $D(\alpha)$  of self-similar signals is concave and therefore can be calculated from  $\tau(q)$  with the inverse Legendre formula (6.79). This formula is also valid for a much larger class of multifractals; for example, it is verified for statistical self-similar signals such as realizations of fractional Brownian motions. Multifractals having some stochastic self-similarity have a spectrum that can often be calculated as an inverse Legendre transform (6.79). However, let us emphasize that this formula is not exact for any function  $f$  because its spectrum of singularity  $D(\alpha)$  is not necessarily concave. In general, Jaffard proved [313] that the Legendre transform (6.79) gives only an upper bound of  $D(\alpha)$ . These singularity spectrum properties are studied in detail in [46].

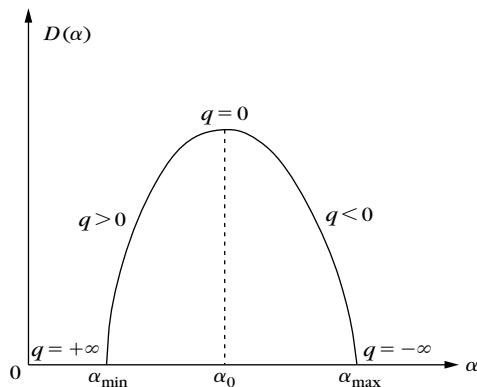
Figure 6.19 illustrates the properties of a concave spectrum  $D(\alpha)$ . The Legendre transform (6.78) proves that its maximum is reached at

$$D(\alpha_0) = \max_{\alpha \in \Lambda} D(\alpha) = -\tau(0).$$

It is the fractal dimension of the Lipschitz exponent  $\alpha_0$  most frequently encountered in  $f$ . Since all other Lipschitz  $\alpha$  singularities appear over sets of lower dimension, if  $\alpha_0 < 1$ , then  $D(\alpha_0)$  is also the fractal dimension of the singular support of  $f$ . The spectrum  $D(\alpha)$  for  $\alpha < \alpha_0$  depends on  $\tau(q)$  for  $q > 0$ , and for  $\alpha > \alpha_0$  it depends on  $\tau(q)$  for  $q < 0$ .

### Numerical Calculations

To compute  $D(\alpha)$ , we assume that the Legendre transform formula (6.79) is valid. We first calculate  $\mathcal{Z}(q, s) = \sum_p |Wf(u_p, s)|^q$ , then derive the decay scaling exponent  $\tau(q)$ , and finally compute  $D(\alpha)$  with a Legendre transform. If  $q < 0$ , then the value of  $\mathcal{Z}(q, s)$  depends mostly on the small-amplitude maxima  $|Wf(u_p, s)|$ . Numerical calculations may then become unstable. To avoid introducing spurious modulus

**FIGURE 6.19**Concave spectrum  $D(\alpha)$ .

maxima created by numerical errors in regions where  $f$  is nearly constant, wavelet maxima are chained to produce maxima curve across scales. If  $\psi = (-1)^p \theta^{(p)}$  where  $\theta$  is a Gaussian, Theorem 6.6 proves that all maxima lines  $u_p(s)$  define curves that propagate up to the limit  $s=0$ . Thus, all maxima lines that do not propagate up to the finest scale are removed in the calculation of  $\mathcal{Z}(q, s)$ . The calculation of the spectrum  $D(\alpha)$  proceeds as follows:

- 1. Maxima.** Compute  $Wf(u, s)$  and the modulus maxima at each scale  $s$ . Chain the wavelet maxima across scales.

- 2. Partition function.** Compute

$$\mathcal{Z}(q, s) = \sum_p |Wf(u_p, s)|^q.$$

- 3. Scaling.** Compute  $\tau(q)$  with a linear regression of  $\log_2 \mathcal{Z}(s, q)$  as a function of  $\log_2 s$ :

$$\log_2 \mathcal{Z}(q, s) \approx \tau(q) \log_2 s + C(q).$$

- 4. Spectrum.** Compute

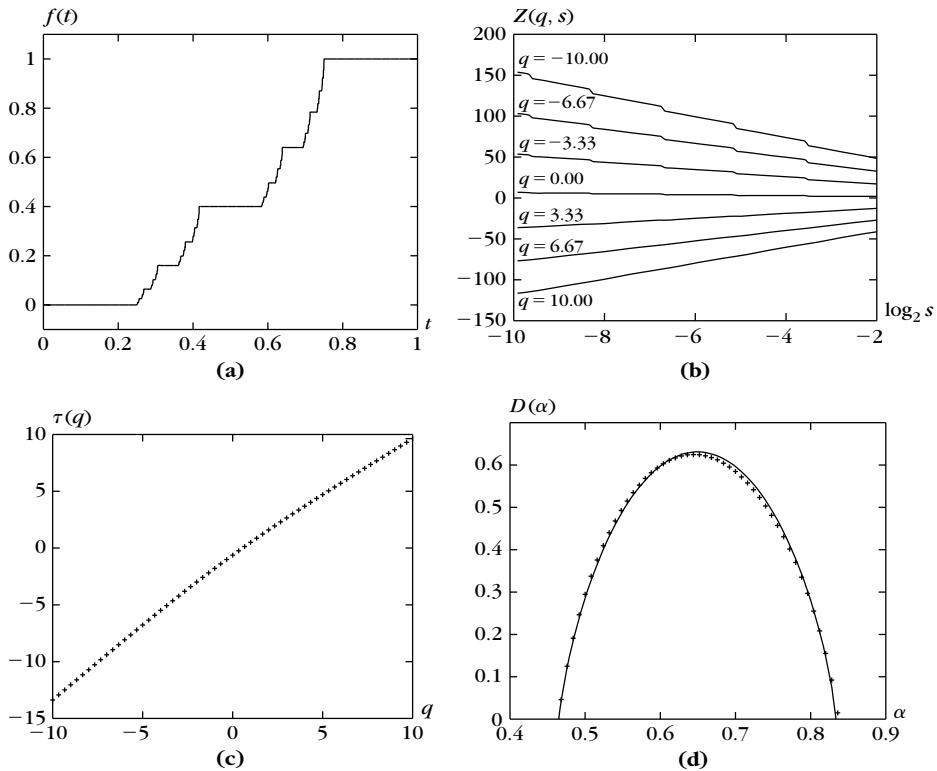
$$D(\alpha) = \min_{q \in \mathbb{R}} (q(\alpha + 1/2) - \tau(q)).$$

---

### EXAMPLE 6.11

The spectrum of singularity  $D(\alpha)$  of the devil's staircase (6.75) is a concave function that can be calculated analytically [292]. Suppose that  $p_1 < p_2$ . The support of  $D(\alpha)$  is  $[\alpha_{\min}, \alpha_{\max}]$  with

$$\alpha_{\min} = \frac{-\log p_2}{\log 3} \quad \text{and} \quad \alpha_{\max} = \frac{-\log p_1}{\log 3}.$$

**FIGURE 6.20**

**(a)** Devil's staircase with  $p_1 = 0.4$  and  $p_2 = 0.6$ . **(b)** Partition function  $Z(q, s)$  for several values of  $q$ . **(c)** Scaling exponent  $\tau(q)$ . **(d)** The theoretical spectrum  $D(\alpha)$  is shown with a solid line. The spectrum values are calculated numerically with a Legendre transform of  $\tau(q)$ .

If  $p_1 = p_2 = 1/2$ , then the support of  $D(\alpha)$  is reduced to a point, which means that all the singularities of  $f$  have the same Lipschitz  $\log 2/\log 3$  regularity. The value  $D(\log 2/\log 3)$  is then the fractal dimension of the triadic Cantor set and is equal to  $\log 2/\log 3$ .

Figure 6.20(a) shows a devil's staircase calculated with  $p_1 = 0.4$  and  $p_2 = 0.6$ . Its wavelet transform is computed with  $\psi = -\theta'$  where  $\theta$  is a Gaussian. The decay of  $\log_2 Z(q, s)$  as a function of  $\log_2 s$  is shown in Figure 6.20(b) for several values of  $q$ . The resulting  $\tau(q)$  and  $D(\alpha)$  are given by Figures 6.20(c, d). There is no numerical instability for  $q < 0$ , because there is no modulus maximum that has an amplitude close to zero. This is not the case if the wavelet transform is calculated with a wavelet that has more vanishing moments.

### Smooth Perturbations

Let  $f$  be a multifractal with a spectrum of singularity  $D(\alpha)$  calculated from  $\tau(q)$ . If a  $C^\infty$  signal  $g$  is added to  $f$  then the singularities are not modified and the singularity

spectrum of  $\tilde{f} = f + g$  remains  $D(\alpha)$ . We study the effect of this smooth perturbation on the spectrum calculation.

The wavelet transform of  $\tilde{f}$  is

$$W\tilde{f}(u, s) = Wf(u, s) + Wg(u, s).$$

Let  $\tau(q)$  and  $\tilde{\tau}(q)$  be the scaling exponent of the partition functions  $\mathcal{Z}(q, s)$  and  $\tilde{\mathcal{Z}}(q, s)$  calculated from the modulus maxima of  $Wf(u, s)$  and  $W\tilde{f}(u, s)$ , respectively. We denote by  $D(\alpha)$  and  $\tilde{D}(\alpha)$  the Legendre transforms of  $\tau(q)$  and  $\tilde{\tau}(q)$ , respectively. Theorem 6.10 relates  $\tau(q)$  and  $\tilde{\tau}(q)$ .

**Theorem 6.10:** *Arneodo, Bacry, Muzy.* Let  $\psi$  be a wavelet with exactly  $n$  vanishing moments. Suppose that  $f$  is a self-similar function.

- If  $g$  is a polynomial of degree  $p < n$ , then  $\tau(q) = \tilde{\tau}(q)$  for all  $q \in \mathbb{R}$ .
- If  $g^{(n)}$  is almost everywhere nonzero, then

$$\tilde{\tau}(q) = \begin{cases} \tau(q) & \text{if } q > q_c \\ (n + 1/2)q & \text{if } q \leq q_c \end{cases} \quad (6.83)$$

where  $q_c$  is defined by  $\tau(q_c) = (n + 1/2)q_c$ .

**Proof.** If  $g$  is a polynomial of degree  $p < n$ , then  $Wg(u, s) = 0$ . The addition of  $g$  does not modify the calculation of the singularity spectrum based on wavelet maxima, so  $\tau(q) = \tilde{\tau}(q)$  for all  $q \in \mathbb{R}$ .

If  $g$  is a  $C^\infty$  function that is not a polynomial then its wavelet transform is generally nonzero. We justify (6.83) with an intuitive argument that is not a proof. A rigorous proof can be found in [91]. Since  $\psi$  has exactly  $n$  vanishing moments, (6.15) proves that

$$|Wg(u, s)| \sim K s^{n+1/2} g^{(n)}(u).$$

We suppose that  $g^{(n)}(u) \neq 0$ . For  $\tau(q) \leq (n + 1/2)q$ , since  $|Wg(u, s)|^q \sim s^{q(n+1/2)}$  has a faster asymptotic decay than  $s^{\tau(q)}$  when  $s$  goes to zero, one can verify that  $\tilde{\mathcal{Z}}(q, s)$  and  $\mathcal{Z}(q, s)$  have the same scaling exponent,  $\tilde{\tau}(q) = \tau(q)$ . If  $\tau(q) > (n + 1/2)q$ , which means that  $q \leq q_c$ , then the decay of  $|W\tilde{f}(u, s)|^q$  is controlled by the decay of  $|Wg(u, s)|^q$ , so  $\tilde{\tau}(q) = (n + 1/2)q$ . ■

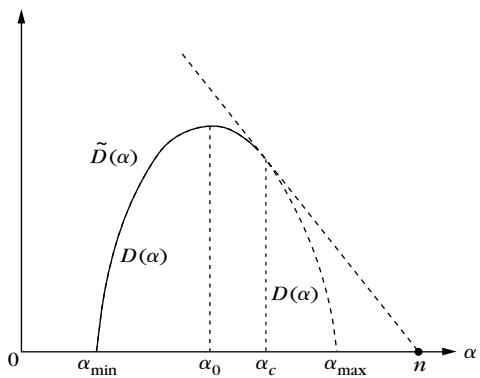
This theorem proves that the addition of a nonpolynomial smooth function introduces a bias in the calculation of the singularity spectrum. Let  $\alpha_c$  be the critical Lipschitz exponent corresponding to  $q_c$ :

$$D(\alpha_c) = q_c (\alpha_c + 1/2) - \tau(q_c).$$

The Legendre transform of  $\tilde{\tau}(q)$  in (6.83) yields

$$\tilde{D}(\alpha) = \begin{cases} D(\alpha) & \text{if } \alpha \leq \alpha_c \\ 0 & \text{if } \alpha = n \\ -\infty & \text{if } \alpha > \alpha_c \text{ and } \alpha \neq n. \end{cases} \quad (6.84)$$

This modification is illustrated by Figure 6.21.

**FIGURE 6.21**

If  $\psi$  has  $n$  vanishing moments, in the presence of a  $C^\infty$  perturbation the computed spectrum  $\tilde{D}(\alpha)$  is identical to the true spectrum  $D(\alpha)$  for  $\alpha \leq \alpha_c$ . Its support is reduced to  $\{n\}$  for  $\alpha > \alpha_c$ .

The bias introduced by the addition of smooth components can be detected experimentally by modifying the number  $n$  of vanishing moments of  $\psi$ . Indeed the value of  $q_c$  depends on  $n$ . If the singularity spectrum varies when changing the number of vanishing moments of the wavelet then it indicates the presence of a bias.

### 6.4.3 Fractal Noises

Fractional Brownian motions are statistically self-similar Gaussian processes that give interesting models for a wide class of natural phenomena [371]. Despite their nonstationarity, one can define a power spectrum that has a power decay. Realizations of fractional Brownian motions are almost everywhere singular, with the same Lipschitz regularity at all points.

We often encounter fractal noise processes that are not Gaussian although their power spectrum has a power decay. Realizations of these processes may include singularities of various types. The spectrum of singularity is then important in analyzing their properties. This is illustrated by an application to hydrodynamic turbulence.

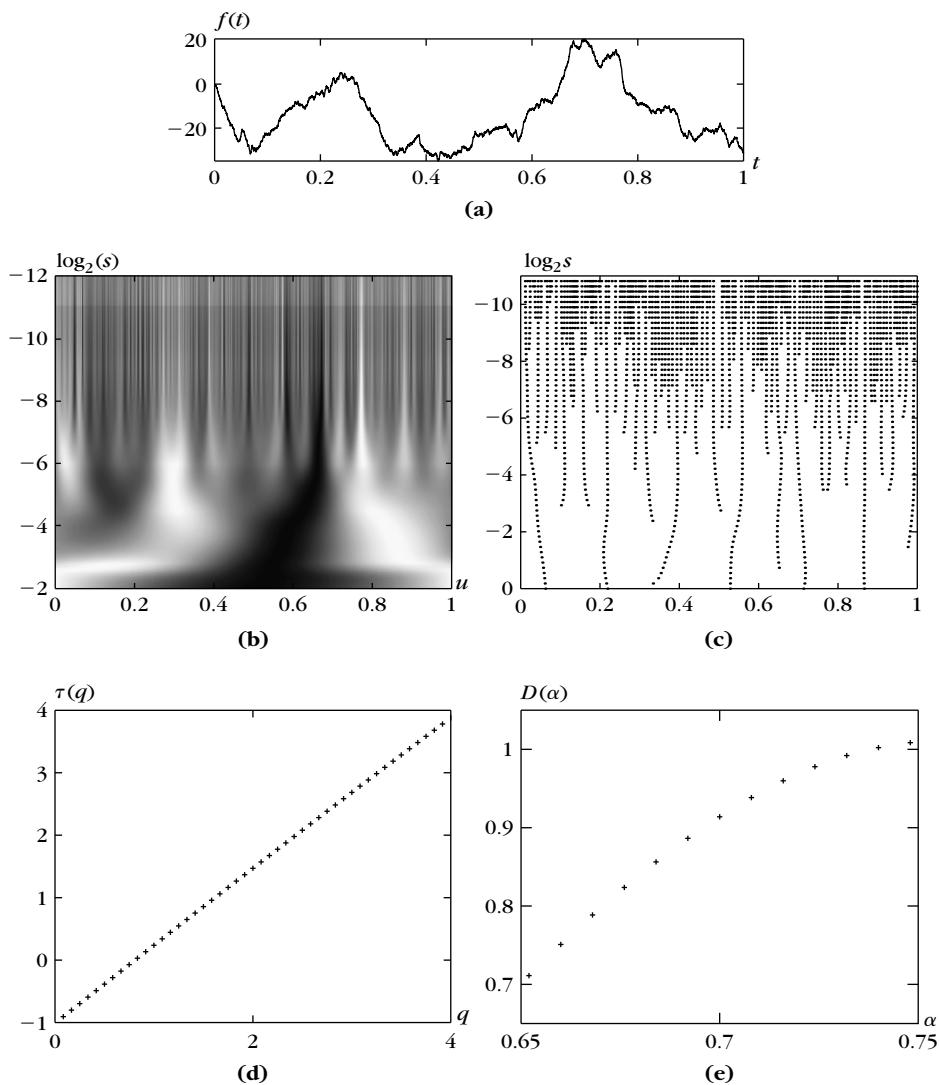
**Definition 6.2:** *Fractional Brownian Motion.* A fractional Brownian motion of Hurst exponent  $0 < H < 1$  is a zero-mean Gaussian process  $B_H$  such that

$$B_H(0) = 0,$$

and

$$E\{|B_H(t) - B_H(t - \Delta)|^2\} = \sigma^2 |\Delta|^{2H}. \quad (6.85)$$

Property (6.85) imposes that the deviation of  $|B_H(t) - B_H(t - \Delta)|$  be proportional to  $|\Delta|^H$ . As a consequence, one can prove that any realization  $f$  of  $B_H$  is almost everywhere singular with a pointwise Lipschitz regularity  $\alpha = H$ . The smaller

**FIGURE 6.22**

(a) One realization of a fractional Brownian motion for a Hurst exponent  $H = 0.7$ . (b) Wavelet transform. (c) Modulus maxima of its wavelet transform. (d) Scaling exponent  $\tau(q)$ . (e) Resulting  $D(\alpha)$  over its support.

$H$  is, the more singular  $f$  is. Figure 6.22(a) shows the graph of one realization for  $H = 0.7$ .

Setting  $\Delta = t$  in (6.85) yields

$$E\{|B_H(t)|^2\} = \sigma^2 |t|^{2H}.$$

Developing (6.85) for  $\Delta = t - u$  also gives

$$E\{B_H(t) B_H(u)\} = \frac{\sigma^2}{2} (|t|^{2H} + |u|^{2H} - |t-u|^{2H}). \quad (6.86)$$

The covariance does not depend only on  $t - u$ , which proves that a fractional Brownian motion is nonstationary.

The statistical self-similarity appears when scaling this process. One can derive from (6.86) that for any  $s > 0$ ,

$$E\{B_H(st) B_H(su)\} = E\{s^H B_H(t) s^H B_H(u)\}.$$

Since  $B_H(st)$  and  $s^H B_H(t)$  are two Gaussian processes with the same mean and covariance, they have the same probability distribution,

$$B_H(st) \equiv s^H B_H(t),$$

where  $\equiv$  denotes an equality of finite-dimensional distributions.

### **Power Spectrum**

Although  $B_H$  is not stationary, one can define a generalized power spectrum. This power spectrum is introduced by proving that the increments of a fractional Brownian motion are stationary and by computing their power spectrum [73].

**Theorem 6.11.** Let  $g_\Delta(t) = \delta(t) - \delta(t - \Delta)$ . The increment

$$I_{H,\Delta}(t) = B_H \star g_\Delta(t) = B_H(t) - B_H(t - \Delta) \quad (6.87)$$

is a stationary process with power spectrum

$$\hat{R}_{I_{H,\Delta}}(\omega) = \frac{\sigma_H^2}{|\omega|^{2H+1}} |\hat{g}_\Delta(\omega)|^2. \quad (6.88)$$

**Proof.** The covariance of  $I_{H,\Delta}$  is computed with (6.86):

$$E\{I_{H,\Delta}(t) I_{H,\Delta}(t - \tau)\} = \frac{\sigma^2}{2} (|\tau - \Delta|^{2H} + |\tau + \Delta|^{2H} - 2|\tau|^{2H}) = R_{I_{H,\Delta}}(\tau). \quad (6.89)$$

The power spectrum  $\hat{R}_{I_{H,\Delta}}(\omega)$  is the Fourier transform of  $R_{I_{H,\Delta}}(\tau)$ . One can verify that the Fourier transform of the distribution  $f(\tau) = |\tau|^{2H}$  is  $\hat{f}(\omega) = -\lambda_H |\omega|^{-(2H+1)}$ , with  $\lambda_H > 0$ . Thus, we derive that the Fourier transform of (6.89) can be written as

$$\hat{R}_{I_{H,\Delta}}(\omega) = 2\sigma^2 \lambda_H |\omega|^{-(2H+1)} \sin^2 \frac{\Delta\omega}{2},$$

which proves (6.88) for  $\sigma_H^2 = \sigma^2 \lambda_H / 2$ . ■

If  $X(t)$  is a stationary process, then we know that  $Y(t) = X \star g(t)$  is also stationary and the power spectrum of both processes is related by

$$\hat{R}_X(\omega) = \frac{\hat{R}_Y(\omega)}{|\hat{g}(\omega)|^2}. \quad (6.90)$$

Although  $B_H(t)$  is not stationary, Theorem 6.11 proves that  $I_{H,\Delta}(t) = B_H \star g_\Delta(t)$  is stationary. As in (6.90), it is tempting to define a “generalized” power spectrum calculated with (6.88):

$$\hat{R}_{B_H}(\omega) = \frac{\hat{R}_{I_{H,\Delta}}(\omega)}{|\hat{g}_\Delta(\omega)|^2} = \frac{\sigma_H^2}{|\omega|^{2H+1}}. \quad (6.91)$$

The nonstationarity of  $B_H(t)$  appears in the energy blow-up at low frequencies. The increments  $I_{H,\Delta}(t)$  are stationary because the multiplication by  $|\hat{g}_\Delta(\omega)|^2 = O(\omega^2)$  removes the explosion of the low-frequency energy. One can generalize this result and verify that if  $g$  is an arbitrary stable filter with a transfer function that satisfies  $|\hat{g}(\omega)| = O(\omega)$ , then  $Y(t) = B_H \star g(t)$  is a stationary Gaussian process with a power spectrum that is

$$\hat{R}_Y(\omega) = \frac{\sigma_H^2}{|\omega|^{2H+1}} |\hat{g}(\omega)|^2. \quad (6.92)$$

### Wavelet Transform

The wavelet transform of a fractional Brownian motion is

$$WB_H(u, s) = B_H \star \bar{\psi}_s(u). \quad (6.93)$$

Since  $\psi$  has at least one vanishing moment, necessarily  $|\hat{\psi}(\omega)| = O(\omega)$  in the neighborhood of  $\omega = 0$ . The wavelet filter  $g = \bar{\psi}_s$  has a Fourier transform  $\hat{g}(\omega) = \sqrt{s} \hat{\psi}^*(s\omega) = O(\omega)$  near  $\omega = 0$ . This proves that for a fixed  $s$  the process  $Y_s(u) = WB_H(u, s)$  is a Gaussian stationary process [258] with a power spectrum that is calculated with (6.92):

$$\hat{R}_{Y_s}(\omega) = s |\hat{\psi}(s\omega)|^2 \frac{\sigma_H^2}{|\omega|^{2H+1}} = s^{2H+2} \hat{R}_{Y_1}(s\omega). \quad (6.94)$$

The self-similarity of the power spectrum and the fact that  $B_H$  is Gaussian are sufficient to prove that  $WB_H(u, s)$  is self-similar across scales:

$$WB_H(u, s) \equiv s^{H+1/2} WB_H\left(\frac{u}{s}, 1\right),$$

where the equivalence means that they have the same finite distributions. Interesting characterizations of fractional Brownian motion properties are also obtained by decomposing these processes in wavelet bases [46, 73, 490].

### EXAMPLE 6.12

Figure 6.22(a) on page 253 displays one realization of a fractional Brownian with  $H = 0.7$ . The wavelet transform and its modulus maxima are shown in Figures 6.22(b) and 6.22(c). The partition function (6.77) is computed from the wavelet modulus maxima. Figure 6.22(d) gives the scaling exponent  $\tau(q)$ , which is nearly a straight line.

Fractional Brownian motions are homogeneous fractals with Lipschitz exponents equal to  $H$ . In this example, the theoretical spectrum  $D(\alpha)$  has a support reduced to  $\{0.7\}$  with  $D(0.7) = 1$ . The estimated spectrum in Figure 6.22(e) is calculated with a Legendre transform of  $\tau(q)$ . Its support is  $[0.65, 0.75]$ . There is an estimation error because the calculations are performed on a signal of finite size.

### **Fractal Noises**

Some physical phenomena produce more general fractal noises  $X(t)$ , which are not Gaussian processes, but they do have stationary increments. As for fractional Brownian motions, one can define a generalized power spectrum with a power decay

$$\hat{R}_X(\omega) = \frac{\sigma_H^2}{|\omega|^{2H+1}}.$$

These processes are transformed into a wide-sense stationary process by a convolution with a stable filter  $g$  that removes the lowest frequencies  $|\hat{g}(\omega)| = O(\omega)$ . Thus, one can determine that the wavelet transform  $Y_s(u) = WX(u, s)$  is a stationary process at any fixed scale  $s$ . Its spectrum is the same as the spectrum (6.94) of fractional Brownian motions. If  $H < 1$ , the asymptotic decay of  $\hat{R}_X(\omega)$  indicates that realizations of  $X(t)$  are singular functions; however, it gives no information about the distribution of these singularities.

As opposed to fractional Brownian motions, general fractal noises have realizations that may include singularities of various types. Such multifractals are differentiated from realizations of fractional Brownian motions by computing their singularity spectrum  $D(\alpha)$ . For example, the velocity fields of fully developed turbulent flows have been modeled by fractal noises, but the calculation of the singularity spectrum clearly shows that these flows differ in important ways from fractional Brownian motions.

### **Hydrodynamic Turbulence**

Fully developed turbulence appears in incompressible flows at high Reynolds numbers. Understanding the properties of hydrodynamic turbulence is a major problem of modern physics, which remains mostly open despite an intense research effort since the first theory of Kolmogorov in 1941 [331]. The number of degrees of liberty of three-dimensional turbulence is considerable, which produces extremely complex spatio-temporal behavior. No formalism is yet able to build a statistical physics framework based on the Navier-Stokes equations that would enable us to understand the global behavior of turbulent flows as it is done in thermodynamics.

In 1941, Kolmogorov [331] formulated a statistical theory of turbulence. The velocity field is modeled as a process  $V(x)$  that has increments with variance

$$E\{|V(x + \Delta) - V(x)|^2\} \sim \varepsilon^{2/3} \Delta^{2/3}.$$

The constant  $\varepsilon$  is a rate of dissipation of energy per unit of mass and time, which is supposed to be independent of the location. This indicates that the velocity field is statistically homogeneous with Lipschitz regularity  $\alpha = H = 1/3$ . The theory predicts that a one-dimensional trace of a three-dimensional velocity field is a fractal noise process with stationary increments that have spectrum decays with a power exponent  $2H + 1 = 5/3$ :

$$\hat{R}_V(\omega) = \frac{\sigma_H^2}{|\omega|^{5/3}}.$$

The success of this theory comes from numerous experimental verifications of this power spectrum decay. However, the theory does not take into account the existence of coherent structures such as vortices. These phenomena contradict the hypothesis of homogeneity, which is at the root of Kolmogorov's 1941 theory.

Kolmogorov [332] modified the homogeneity assumption in 1962 by introducing an energy dissipation rate  $\varepsilon(x)$  that varies with the spatial location  $x$ . This opens the door to “local stochastic self-similar” multifractal models, first developed by Mandelbrot [370] to explain energy exchanges between fine-scale structures and large-scale structures. The spectrum of singularity  $D(\alpha)$  is playing an important role in testing these models [264]. Calculations with wavelet maxima on turbulent velocity fields [5] show that  $D(\alpha)$  is maximum at  $1/3$ , as predicted by the Kolmogorov theory. However,  $D(\alpha)$  does not have a support reduced to  $\{1/3\}$ , which verifies that a turbulent velocity field is not a homogeneous process. Models based on the wavelet transform have been introduced to explain the distribution of vortices in turbulent fluids [13, 251, 252].

## 6.5 EXERCISES

### 6.1 <sup>2</sup> Lipschitz regularity:

- (a) Prove that if  $f$  is uniformly Lipschitz  $\alpha$  on  $[a, b]$ , then it is pointwise Lipschitz  $\alpha$  at all  $t_0 \in [a, b]$ .
- (b) Show that  $f(t) = t \sin t^{-1}$  is Lipschitz 1 at all  $t_0 \in [-1, 1]$  and verify that it is uniformly Lipschitz  $\alpha$  over  $[-1, 1]$  only for  $\alpha \leq 1/2$ . *Hint:* Consider the points  $t_n = (n + 1/2)^{-1} \pi^{-1}$ .

### 6.2 <sup>2</sup> Regularity of derivatives:

- (a) Prove that  $f$  is uniformly Lipschitz  $\alpha > 1$  over  $[a, b]$  if and only if  $f'$  is uniformly Lipschitz  $\alpha - 1$  over  $[a, b]$ .
- (b) Show that  $f$  may be pointwise Lipschitz  $\alpha > 1$  at  $t_0$  while  $f'$  is not pointwise Lipschitz  $\alpha - 1$  at  $t_0$ . Consider  $f(t) = t^2 \cos t^{-1}$  at  $t = 0$ .

### 6.3 <sup>2</sup> Find $f(t)$ that is uniformly Lipschitz 1, but does not satisfy the sufficient Fourier condition (6.1).

### 6.4 <sup>1</sup> Let $f(t) = \cos \omega_0 t$ and $\psi(t)$ be a wavelet that is symmetric about 0.

(a) Verify that

$$Wf(u, s) = \sqrt{s} \hat{\psi}(sw_0) \cos \omega_0 t.$$

(b) Find the equations of the curves of wavelet modulus maxima in the time-scale plane  $(u, s)$ . Relate the decay of  $|Wf(u, s)|$  along these curves to the number  $n$  of vanishing moments of  $\psi$ .

- 6.5** <sup>1</sup> Let  $f(t) = |t|^\alpha$ . Show that  $Wf(u, s) = s^{\alpha+1/2} Wf(u/s, 1)$ . Prove that it is not sufficient to measure the decay of  $|Wf(u, s)|$  when  $s$  goes to zero at  $u=0$  in order to compute the Lipschitz regularity of  $f$  at  $t=0$ .
- 6.6** <sup>3</sup> Let  $f(t) = |t|^\alpha \sin |t|^{-\beta}$  with  $\alpha > 0$  and  $\beta > 0$ . What is the pointwise Lipschitz regularity of  $f$  and  $f'$  at  $t=0$ ? Find the equation of the ridge curve in the  $(u, s)$  plane along which the high-amplitude wavelet coefficients  $|Wf(u, s)|$  converge to  $t=0$  when  $s$  goes to zero. Compute the maximum values of  $\alpha$  and  $\alpha'$  such that  $Wf(u, s)$  satisfies (6.21).
- 6.7** <sup>2</sup> For a complex wavelet, we call *lines of constant phase* the curves in the  $(u, s)$  plane along which the complex phase of  $Wf(u, s)$  remains constant when  $s$  varies.
- (a) If  $f(t) = |t|^\alpha$ , prove that the lines of constant phase converge toward the singularity at  $t=0$  when  $s$  goes to zero. Verify this numerically.
  - (b) Let  $\psi$  be a real wavelet and  $Wf(u, s)$  be the real wavelet transform of  $f$ . Show that the modulus maxima of  $Wf(u, s)$  correspond to lines of constant phase of an analytic wavelet transform, which is calculated with a particular analytic wavelet  $\psi^a$  that you will specify.
- 6.8** <sup>3</sup> Prove that if  $f = \mathbf{1}_{[0, +\infty)}$ , then the number of modulus maxima of  $Wf(u, s)$  at each scale  $s$  is larger than or equal to the number of vanishing moments of  $\psi$ .
- 6.9** <sup>2</sup> The spectrum of singularity of the Riemann function

$$f(t) = \sum_{n=-\infty}^{+\infty} \frac{1}{n^2} \sin n^2 t$$

is defined on its support by  $D(\alpha) = 4\alpha - 2$  if  $\alpha \in [1/2, 3/4]$  and  $D(3/2) = 0$  [304, 313]. Verify this result numerically by computing this spectrum from the partition function of a wavelet transform modulus maxima.

- 6.10** <sup>3</sup> Let  $\psi = -\theta'$  where  $\theta$  is a positive window of compact support. If  $f$  is a Cantor devil's staircase, prove that there exist lines of modulus maxima that converge toward each singularity.
- 6.11** <sup>3</sup> Implement an algorithm that detects oscillating singularities by following the ridges of an analytic wavelet transform when the scale  $s$  decreases. Test your algorithm on  $f(t) = \sin t^{-1}$ .

- 6.12** <sup>2</sup> Implement an algorithm that reconstructs a signal from the local maxima of its dyadic wavelet transform with a dual synthesis (6.48) using a conjugate-gradient algorithm.
- 6.13** <sup>3</sup> Let  $X[n] = f[n] + W[n]$  be a signal of size  $N$ , where  $W$  is a Gaussian white noise of variance  $\sigma^2$ . Implement in WAVELET an estimator of  $f$  that thresholds at  $T = \lambda \sigma$  the maxima of a dyadic wavelet transform of  $X$ . The estimation of  $f$  is reconstructed from the thresholded maxima representation with the dual synthesis (6.48) implemented with a conjugate-gradient algorithm. Compare numerically the risk of this estimator with the risk of a thresholding estimator over the translation-invariant dyadic wavelet transform of  $X$ .
- 6.14** <sup>2</sup> Let  $\theta(t)$  be a Gaussian of variance 1.

(a) Prove that the Laplacian of a two-dimensional Gaussian

$$\psi(x_1, x_2) = \frac{\partial^2 \theta(x_1)}{\partial x^2} \theta(x_2) + \theta(x_1) \frac{\partial^2 \theta(x_2)}{\partial x_2^2}$$

satisfies the dyadic wavelet condition (5.101) (there is only one wavelet).

- (b) Explain why the zero-crossings of this dyadic wavelet transform provide the locations of multiscale edges in images. Compare the position of these zero-crossings with the wavelet modulus maxima obtained with  $\psi^1(x_1, x_2) = -\theta'(x_1) \theta(x_2)$  and  $\psi^2(x_1, x_2) = -\theta(x_1) \theta'(x_2)$ .

- 6.15** <sup>2</sup> The covariance of a fractional Brownian motion  $B_H(t)$  is given by (6.86). Show that the wavelet transform at a scale  $s$  is stationary by verifying that

$$E\left\{ WB_H(u_1, s) WB_H(u_2, s) \right\} = -\frac{\sigma^2}{2} s^{2H+1} \int_{-\infty}^{+\infty} |t|^{2H} \Psi\left(\frac{u_1 - u_2}{s} - t\right) dt,$$

with  $\Psi(t) = \psi \star \bar{\psi}(t)$  and  $\bar{\psi}(t) = \psi(-t)$ .

- 6.16** <sup>2</sup> Let  $X(t)$  be a stationary Gaussian process with a covariance  $R_X(\tau) = E\{X(t)X(t - \tau)\}$  that is twice differentiable. One can prove that the average number of zero-crossings over an interval of size 1 is  $-\pi R_X''(0) (\pi^2 R_X(0))^{-1}$  [53]. Let  $B_H(t)$  be a fractional Brownian motion and  $\psi$  a wavelet that is  $C^2$ . Prove that the average numbers, respectively, of zero-crossings and of modulus maxima of  $WB_H(u, s)$  for  $u \in [0, 1]$  are proportional to  $s$ . Verify this result numerically.
- 6.17** <sup>2</sup> Implement an algorithm that estimates the Lipschitz regularity  $\alpha$  and the smoothing scale  $\sigma$  of sharp variation points in one-dimensional signals by applying the result of Theorem 6.7 on the dyadic wavelet transform maxima.

# Wavelet Bases

# 7

One can construct wavelets  $\psi$  such that the dilated and translated family

$$\left\{ \psi_{j,n}(t) = \frac{1}{\sqrt{2^j}} \psi\left(\frac{t - 2^j n}{2^j}\right) \right\}_{(j,n) \in \mathbb{Z}^2}$$

is an orthonormal basis of  $L^2(\mathbb{R})$ . Behind this simple statement lie very different points of view that open a fruitful exchange between harmonic analysis and discrete signal processing.

Orthogonal wavelets dilated by  $2^j$  carry signal variations at the resolution  $2^{-j}$ . The construction of these bases can be related to multiresolution signal approximations. Following this link leads us to an unexpected equivalence between wavelet bases and conjugate mirror filters used in discrete multirate filter banks. These filter banks implement a fast orthogonal wavelet transform that requires only  $O(N)$  operations for signals of size  $N$ . The design of conjugate mirror filters also gives new classes of wavelet orthogonal bases including regular wavelets of compact support. In several dimensions, wavelet bases of  $L^2(\mathbb{R}^d)$  are constructed with separable products of functions of one variable. Wavelet bases are also adapted to bounded domains and surfaces with lifting algorithms.

## 7.1 ORTHOGONAL WAVELET BASES

Our search for orthogonal wavelets begins with multiresolution approximations. For  $f \in L^2(\mathbb{R})$ , the partial sum of wavelet coefficients  $\sum_{n=-\infty}^{+\infty} \langle f, \psi_{j,n} \rangle \psi_{j,n}$  can indeed be interpreted as the difference between two approximations of  $f$  at the resolutions  $2^{-j+1}$  and  $2^{-j}$ . Multiresolution approximations compute the approximation of signals at various resolutions with orthogonal projections on different spaces  $\{\mathbf{V}_j\}_{j \in \mathbb{Z}}$ . Section 7.1.3 proves that multiresolution approximations are entirely characterized by a particular discrete filter that governs the loss of information across resolutions. These discrete filters provide a simple procedure for designing and synthesizing orthogonal wavelet bases.

### 7.1.1 Multiresolution Approximations

Adapting the signal resolution allows one to process only the relevant details for a particular task. In computer vision, Burt and Adelson [126] introduced a multiresolution pyramid that can be used to process a low-resolution image first and then selectively increase the resolution when necessary. This section formalizes multiresolution approximations, which set the ground for the construction of orthogonal wavelets.

The approximation of a function  $f$  at a resolution  $2^{-j}$  is specified by a discrete grid of samples that provides local averages of  $f$  over neighborhoods of size proportional to  $2^j$ . Thus, a multiresolution approximation is composed of embedded grids of approximation. More formally, the approximation of a function at a resolution  $2^{-j}$  is defined as an orthogonal projection on a space  $\mathbf{V}_j \subset \mathbf{L}^2(\mathbb{R})$ . The space  $\mathbf{V}_j$  regroups all possible approximations at the resolution  $2^{-j}$ . The orthogonal projection of  $f$  is the function  $f_j \in \mathbf{V}_j$  that minimizes  $\|f - f_j\|$ . The following definition, introduced by Mallat [362] and Meyer [44], specifies the mathematical properties of multiresolution spaces. To avoid confusion, let us emphasize that a scale parameter  $2^j$  is the inverse of the resolution  $2^{-j}$ .

**Definition 7.1:** *Multiresolutions.* A sequence  $\{\mathbf{V}_j\}_{j \in \mathbb{Z}}$  of closed subspaces of  $\mathbf{L}^2(\mathbb{R})$  is a multiresolution approximation if the following six properties are satisfied:

$$\forall (j, k) \in \mathbb{Z}^2, \quad f(t) \in \mathbf{V}_j \Leftrightarrow f(t - 2^j k) \in \mathbf{V}_j, \quad (7.1)$$

$$\forall j \in \mathbb{Z}, \quad \mathbf{V}_{j+1} \subset \mathbf{V}_j, \quad (7.2)$$

$$\forall j \in \mathbb{Z}, \quad f(t) \in \mathbf{V}_j \Leftrightarrow f\left(\frac{t}{2}\right) \in \mathbf{V}_{j+1}, \quad (7.3)$$

$$\lim_{j \rightarrow +\infty} \mathbf{V}_j = \bigcap_{j=-\infty}^{+\infty} \mathbf{V}_j = \{0\}, \quad (7.4)$$

$$\lim_{j \rightarrow -\infty} \mathbf{V}_j = \text{Closure} \left( \bigcup_{j=-\infty}^{+\infty} \mathbf{V}_j \right) = \mathbf{L}^2(\mathbb{R}), \quad (7.5)$$

and there exists  $\theta$  such that  $\{\theta(t - n)\}_{n \in \mathbb{Z}}$  is a Riesz basis of  $\mathbf{V}_0$ .

Let us give an intuitive explanation of these mathematical properties. Property (7.1) means that  $\mathbf{V}_j$  is invariant by any translation proportional to the scale  $2^j$ . As we shall see later, this space can be assimilated to a uniform grid with intervals  $2^j$ , which characterizes the signal approximation at the resolution  $2^{-j}$ . The inclusion (7.2) is a causality property that proves that an approximation at a resolution  $2^{-j}$  contains all the necessary information to compute an approximation at a coarser resolution  $2^{-j-1}$ . Dilating functions in  $\mathbf{V}_j$  by 2 enlarges the details by 2 and (7.3) guarantees that it defines an approximation at a coarser resolution  $2^{-j-1}$ . When the resolution  $2^{-j}$  goes to 0 (7.4) implies that we lose all the details of  $f$  and

$$\lim_{j \rightarrow +\infty} \|P_{\mathbf{V}_j} f\| = 0. \quad (7.6)$$

On the other hand, when the resolution  $2^{-j}$  goes  $+\infty$ , property (7.5) imposes that the signal approximation converges to the original signal:

$$\lim_{j \rightarrow -\infty} \|f - P_{V_j} f\| = 0. \quad (7.7)$$

When the resolution  $2^{-j}$  increases, the decay rate of the approximation error  $\|f - P_{V_j} f\|$  depends on the regularity of  $f$ . In Section 9.1.3 we relate this error to the uniform Lipschitz regularity of  $f$ .

The existence of a Riesz basis  $\{\theta(t-n)\}_{n \in \mathbb{Z}}$  of  $V_0$  provides a discretization theorem as explained in Section 3.1.3. The function  $\theta$  can be interpreted as a unit resolution cell; Section 5.1.1 gives the definition of a Riesz basis. It is a family of linearly independent functions such that there exist  $B \geq A > 0$ , which satisfy

$$\forall f \in V_0, \quad A \|f\|^2 \leq \sum_{n=-\infty}^{+\infty} |\langle f(t), \theta(t-n) \rangle|^2 \leq B \|f\|^2. \quad (7.8)$$

This energy equivalence guarantees that signal expansions over  $\{\theta(t-n)\}_{n \in \mathbb{Z}}$  are numerically stable. One may verify that the family  $\{2^{-j/2}\theta(2^{-j}t-n)\}_{n \in \mathbb{Z}}$  is a Riesz basis of  $V_j$  with the same Riesz bounds  $A$  and  $B$  at all scales  $2^j$ . Theorem 3.4 proves that  $\{\theta(t-n)\}_{n \in \mathbb{Z}}$  is a Riesz basis if and only if

$$\forall \omega \in [-\pi, \pi], \quad A \leq \sum_{k=-\infty}^{+\infty} |\hat{\theta}(\omega + 2k\pi)|^2 \leq B. \quad (7.9)$$

### EXAMPLE 7.1: Piecewise Constant Approximations

A simple multiresolution approximation is composed of piecewise constant functions. Space  $V_j$  is the set of all  $g \in L^2(\mathbb{R})$  such that  $g(t)$  is constant for  $t \in [n2^j, (n+1)2^j]$  and  $n \in \mathbb{Z}$ . The approximation at a resolution  $2^{-j}$  of  $f$  is the closest piecewise constant function on intervals of size  $2^j$ . The resolution cell can be chosen to be the box window  $\theta = \mathbf{1}_{[0,1]}$ . Clearly,  $V_j \subset V_{j-1}$ , since functions constant on intervals of size  $2^j$  are also constant on intervals of size  $2^{j-1}$ . The verification of the other multiresolution properties is left to the reader. It is often desirable to construct approximations that are smooth functions, in which case piecewise constant functions are not appropriate.

### EXAMPLE 7.2: Shannon Approximations

Frequency band-limited functions also yield multiresolution approximations. Space  $V_j$  is defined as the set of functions with a Fourier transform support included in  $[-2^{-j}\pi, 2^{-j}\pi]$ . Theorem 3.5 provides an orthonormal basis  $\{\theta(t-n)\}_{n \in \mathbb{Z}}$  of  $V_0$  defined by

$$\theta(t) = \frac{\sin \pi t}{\pi t}. \quad (7.10)$$

All other properties of multiresolution approximation are easily verified.

The approximation at the resolution  $2^{-j}$  of  $f \in \mathbf{L}^2(\mathbb{R})$  is the function  $P_{\mathbf{V}_j}f \in \mathbf{V}_j$  that minimizes  $\|P_{\mathbf{V}_j}f - f\|$ . It is proved in (3.12) that its Fourier transform is obtained with a frequency filtering:

$$\widehat{P_{\mathbf{V}_j}f}(\omega) = \widehat{f}(\omega) \mathbf{1}_{[-2^{-j}\pi, 2^{-j}\pi]}(\omega).$$

This Fourier transform is generally discontinuous at  $\pm 2^{-j}\pi$ , in which case  $|P_{\mathbf{V}_j}f(t)|$  decays like  $|t|^{-1}$  for large  $|t|$ , even though  $f$  might have a compact support.

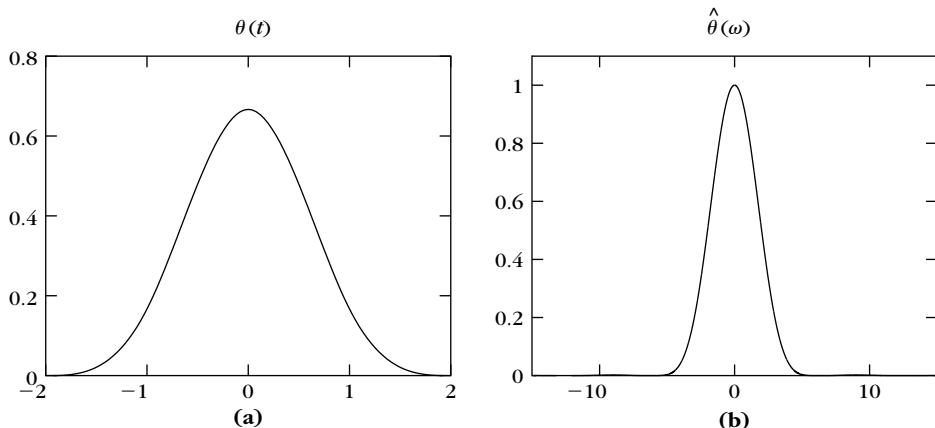
### EXAMPLE 7.3: Spline Approximations

Polynomial spline approximations construct smooth approximations with fast asymptotic decay. The space  $\mathbf{V}_j$  of splines of degree  $m \geq 0$  is the set of functions that are  $m-1$  times continuously differentiable and equal to a polynomial of degree  $m$  on any interval  $[n2^j, (n+1)2^j]$  for  $n \in \mathbb{Z}$ . When  $m=0$ , it is a piecewise constant multiresolution approximation. When  $m=1$ , functions in  $\mathbf{V}_j$  are piecewise linear and continuous.

A Riesz basis of polynomial splines is constructed with *box splines*. A box spline  $\theta$  of degree  $m$  is computed by convolving the box window  $\mathbf{1}_{[0,1]}$  with itself  $m+1$  times and centering at 0 or  $1/2$ . Its Fourier transform is

$$\widehat{\theta}(\omega) = \left( \frac{\sin(\omega/2)}{\omega/2} \right)^{m+1} \exp\left( \frac{-i\varepsilon\omega}{2} \right). \quad (7.11)$$

If  $m$  is even, then  $\varepsilon=1$  and  $\theta$  has a support centered at  $t=1/2$ . If  $m$  is odd, then  $\varepsilon=0$  and  $\theta(t)$  is symmetric about  $t=0$ . Figure 7.1 displays a cubic box spline  $m=3$  and its Fourier transform. For all  $m \geq 0$ , one can prove that  $\{\theta(t-n)\}_{n \in \mathbb{Z}}$  is a Riesz basis of  $\mathbf{V}_0$  by verifying the condition (7.9). This is done with a closed-form expression for the series (7.19).



**FIGURE 7.1**

Cubic box spline  $\theta$  (a) and its Fourier transform  $\widehat{\theta}$  (b).

### 7.1.2 Scaling Function

The approximation of  $f$  at the resolution  $2^{-j}$  is defined as the orthogonal projection  $P_{V_j}f$  on  $V_j$ . To compute this projection, we must find an orthonormal basis of  $V_j$ . Theorem 7.1 orthogonalizes the Riesz basis  $\{\theta(t-n)\}_{n \in \mathbb{Z}}$  and constructs an orthogonal basis of each space  $V_j$  by dilating and translating a single function  $\phi$  called a *scaling function*. To avoid confusing the resolution  $2^{-j}$  and the scale  $2^j$ , in the rest of this chapter the notion of resolution is dropped and  $P_{V_j}f$  is called an approximation at the scale  $2^j$ .

**Theorem 7.1.** Let  $\{V_j\}_{j \in \mathbb{Z}}$  be a multiresolution approximation and  $\phi$  be the scaling function having a Fourier transform

$$\hat{\phi}(\omega) = \frac{\hat{\theta}(\omega)}{\left(\sum_{k=-\infty}^{+\infty} |\hat{\theta}(\omega + 2k\pi)|^2\right)^{1/2}}. \quad (7.12)$$

Let us denote

$$\phi_{j,n}(t) = \frac{1}{\sqrt{2^j}} \phi\left(\frac{t-n}{2^j}\right).$$

The family  $\{\phi_{j,n}\}_{n \in \mathbb{Z}}$  is an orthonormal basis of  $V_j$  for all  $j \in \mathbb{Z}$ .

**Proof.** To construct an orthonormal basis, we look for a function  $\phi \in V_0$ . Thus, it can be expanded in the basis  $\{\theta(t-n)\}_{n \in \mathbb{Z}}$ :

$$\phi(t) = \sum_{n=-\infty}^{+\infty} a[n] \theta(t-n),$$

which implies that

$$\hat{\phi}(\omega) = \hat{a}(\omega) \hat{\theta}(\omega),$$

where  $\hat{a}$  is a  $2\pi$  periodic Fourier series of finite energy. To compute  $\hat{a}$  we express the orthogonality of  $\{\phi(t-n)\}_{n \in \mathbb{Z}}$  in the Fourier domain. Let  $\bar{\phi}(t) = \phi^*(-t)$ . For any  $(n, p) \in \mathbb{Z}^2$ ,

$$\begin{aligned} \langle \phi(t-n), \phi(t-p) \rangle &= \int_{-\infty}^{+\infty} \phi(t-n) \phi^*(t-p) dt \\ &= \phi \star \bar{\phi}(p-n). \end{aligned} \quad (7.13)$$

Thus,  $\{\phi(t-n)\}_{n \in \mathbb{Z}}$  is orthonormal if and only if  $\phi \star \bar{\phi}(n) = \delta[n]$ . Computing the Fourier transform of this equality yields

$$\sum_{k=-\infty}^{+\infty} |\hat{\phi}(\omega + 2k\pi)|^2 = 1. \quad (7.14)$$

Indeed, the Fourier transform of  $\phi \star \bar{\phi}(t)$  is  $|\hat{\phi}(\omega)|^2$ , and we proved in (3.3) that sampling a function periodizes its Fourier transform. The property (7.14) is verified if we choose

$$\hat{a}(\omega) = \left( \sum_{k=-\infty}^{+\infty} |\hat{\theta}(\omega + 2k\pi)|^2 \right)^{-1/2}.$$

We saw in (7.9) that the denominator has a strictly positive lower bound, so  $\hat{\alpha}$  is a  $2\pi$  periodic function of finite energy. ■

### Approximation

The orthogonal projection of  $f$  over  $V_j$  is obtained with an expansion in the scaling orthogonal basis

$$P_{V_j}f = \sum_{n=-\infty}^{+\infty} \langle f, \phi_{j,n} \rangle \phi_{j,n}. \quad (7.15)$$

The inner products

$$a_j[n] = \langle f, \phi_{j,n} \rangle \quad (7.16)$$

provide a discrete approximation at the scale  $2^j$ . We can rewrite them as a convolution product:

$$a_j[n] = \int_{-\infty}^{+\infty} f(t) \frac{1}{\sqrt{2^j}} \phi\left(\frac{t - 2^j n}{2^j}\right) dt = f \star \bar{\phi}_j(2^j n), \quad (7.17)$$

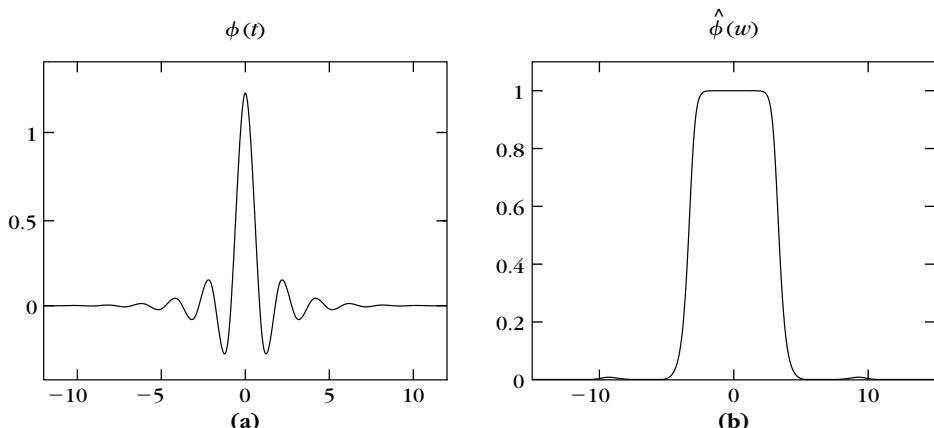
with  $\bar{\phi}_j(t) = \sqrt{2^{-j}} \phi(2^{-j}t)$ . The energy of the Fourier transform  $\hat{\phi}$  is typically concentrated in  $[-\pi, \pi]$ , as illustrated by Figure 7.2. As a consequence, the Fourier transform  $\sqrt{2^j} \hat{\phi}^*(2^j \omega)$  of  $\bar{\phi}_j(t)$  is mostly nonnegligible in  $[-2^{-j}\pi, 2^{-j}\pi]$ . The discrete approximation  $a_j[n]$  is therefore a low-pass filtering of  $f$  sampled at intervals  $2^j$ . Figure 7.3 gives a discrete multiresolution approximation at scales  $2^{-9} \leq 2^j \leq 2^{-4}$ .

---

### EXAMPLE 7.4

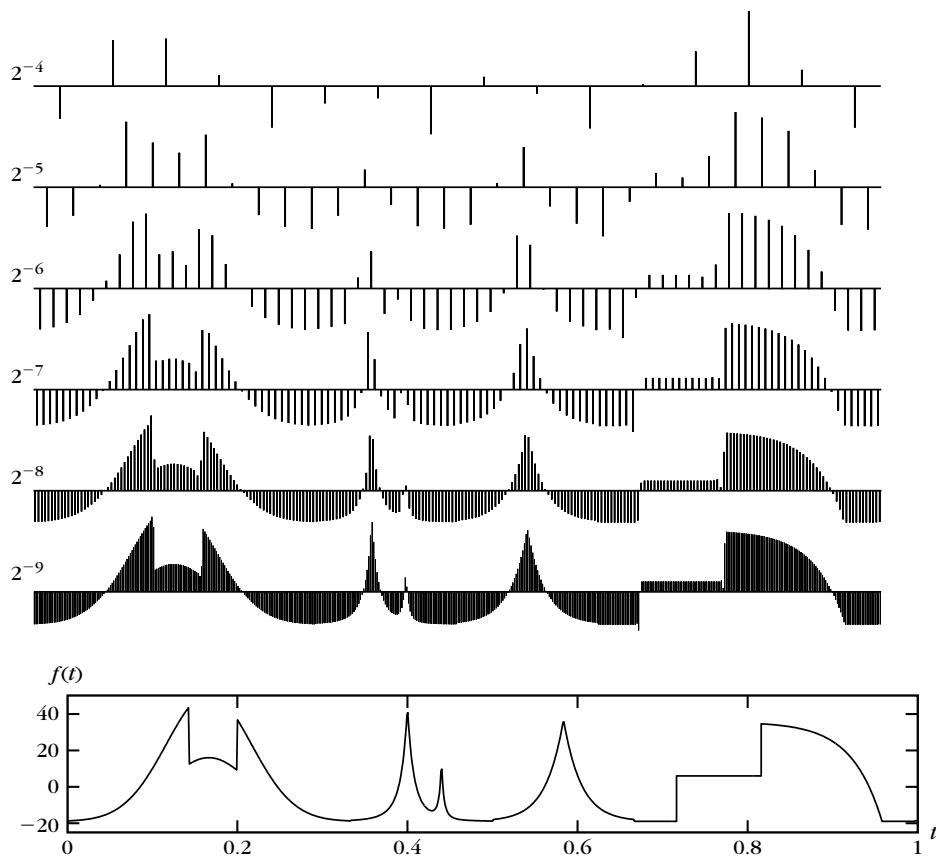
For piecewise constant approximations and Shannon multiresolution approximations we have constructed Riesz bases  $\{\theta(t - n)\}_{n \in \mathbb{Z}}$  that are orthonormal bases, thus  $\phi = \theta$ .

---



**FIGURE 7.2**

Cubic spline-scaling function  $\phi$  (a) and its Fourier transform  $\hat{\phi}$  computed with (7.18) (b).

**FIGURE 7.3**

Discrete multiresolution approximations  $a_j[n]$  at scales  $2^j$ , computed with cubic splines.

**EXAMPLE 7.5**

Spline multiresolution approximations admit a Riesz basis constructed with a box spline  $\theta$  of degree  $m$ , having a Fourier transform given by (7.11). Inserting this expression in (7.12) yields

$$\hat{\phi}(\omega) = \frac{\exp(-i\varepsilon\omega/2)}{\omega^{m+1} \sqrt{S_{2m+2}(\omega)}}, \quad (7.18)$$

with

$$S_n(\omega) = \sum_{k=-\infty}^{+\infty} \frac{1}{(\omega + 2k\pi)^n}, \quad (7.19)$$

and  $\varepsilon = 1$  if  $m$  is even or  $\varepsilon = 0$  if  $m$  is odd. A closed-form expression of  $S_{2m+2}(\omega)$  is obtained by computing the derivative of order  $2m$  of the identity

$$S_2(2\omega) = \sum_{k=-\infty}^{+\infty} \frac{1}{(2\omega + 2k\pi)^2} = \frac{1}{4 \sin^2 \omega}.$$

linear splines,  $m = 1$  and

$$S_4(2\omega) = \frac{1 + 2 \cos^2 \omega}{48 \sin^4 \omega}, \quad (7.20)$$

which yields

$$\hat{\phi}(\omega) = \frac{4 \sqrt{3} \sin^2(\omega/2)}{\omega^2 \sqrt{1 + 2 \cos^2(\omega/2)}}. \quad (7.21)$$

The cubic spline-scaling function corresponds to  $m = 3$ , and  $\hat{\phi}(\omega)$  is calculated with (7.18) by inserting

$$S_8(2\omega) = \frac{5 + 30 \cos^2 \omega + 30 \sin^2 \omega \cos^2 \omega}{105 2^8 \sin^8 \omega} + \frac{70 \cos^4 \omega + 2 \sin^4 \omega \cos^2 \omega + 2/3 \sin^6 \omega}{105 2^8 \sin^8 \omega}. \quad (7.22)$$

This cubic spline-scaling function  $\phi$  and its Fourier transform are displayed in Figure 7.2. It has an infinite support but decays exponentially.

---

### 7.1.3 Conjugate Mirror Filters

A multiresolution approximation is entirely characterized by the scaling function  $\phi$  that generates an orthogonal basis of each space  $\mathbf{V}_j$ . We study the properties of  $\phi$ , which guarantee that the spaces  $\mathbf{V}_j$  satisfy all conditions of a multiresolution approximation. It is proved that any scaling function is specified by a discrete filter called a *conjugate mirror filter*.

#### Scaling Equation

The multiresolution causality property (7.2) imposes that  $\mathbf{V}_j \subset \mathbf{V}_{j-1}$ ; in particular,  $2^{-1/2}\phi(t/2) \in \mathbf{V}_1 \subset \mathbf{V}_0$ . Since  $\{\phi(t-n)\}_{n \in \mathbb{Z}}$  is an orthonormal basis of  $\mathbf{V}_0$ , we can decompose

$$\frac{1}{\sqrt{2}} \phi\left(\frac{t}{2}\right) = \sum_{n=-\infty}^{+\infty} h[n] \phi(t-n), \quad (7.23)$$

with

$$h[n] = \left\langle \frac{1}{\sqrt{2}} \phi\left(\frac{t}{2}\right), \phi(t-n) \right\rangle. \quad (7.24)$$

This scaling equation relates a dilation of  $\phi$  by 2 to its integer translations. The sequence  $h[n]$  will be interpreted as a discrete filter.

The Fourier transform of both sides of (7.23) yields

$$\hat{\phi}(2\omega) = \frac{1}{\sqrt{2}} \hat{h}(\omega) \hat{\phi}(\omega) \quad (7.25)$$

for  $\hat{h}(\omega) = \sum_{n=-\infty}^{+\infty} h[n] e^{-in\omega}$ . Thus, it is tempting to express  $\hat{\phi}(\omega)$  directly as a product of dilations of  $\hat{h}(\omega)$ . For any  $p \geq 0$ , (7.25) implies

$$\hat{\phi}(2^{-p+1}\omega) = \frac{1}{\sqrt{2}} \hat{h}(2^{-p}\omega) \hat{\phi}(2^{-p}\omega). \quad (7.26)$$

By substitution, we obtain

$$\hat{\phi}(\omega) = \left( \prod_{p=1}^P \frac{\hat{h}(2^{-p}\omega)}{\sqrt{2}} \right) \hat{\phi}(2^{-P}\omega). \quad (7.27)$$

If  $\hat{\phi}(\omega)$  is continuous at  $\omega = 0$ , then  $\lim_{P \rightarrow +\infty} \hat{\phi}(2^{-P}\omega) = \hat{\phi}(0)$ , so

$$\hat{\phi}(\omega) = \prod_{p=1}^{+\infty} \frac{\hat{h}(2^{-p}\omega)}{\sqrt{2}} \hat{\phi}(0). \quad (7.28)$$

Theorem 7.2 [44, 362] gives necessary and sufficient conditions on  $\hat{h}(\omega)$  to guarantee that this infinite product is the Fourier transform of a scaling function.

**Theorem 7.2: Mallat, Meyer.** Let  $\phi \in \mathbf{L}^2(\mathbb{R})$  be an integrable scaling function. The Fourier series of  $h[n] = \langle 2^{-1/2}\phi(t/2), \phi(t-n) \rangle$  satisfies

$$\forall \omega \in \mathbb{R}, \quad |\hat{h}(\omega)|^2 + |\hat{h}(\omega + \pi)|^2 = 2, \quad (7.29)$$

and

$$\hat{h}(0) = \sqrt{2}. \quad (7.30)$$

Conversely, if  $\hat{h}(\omega)$  is  $2\pi$  periodic and continuously differentiable in a neighborhood of  $\omega = 0$ , if it satisfies (7.29) and (7.30) and if

$$\inf_{\omega \in [-\pi/2, \pi/2]} |\hat{h}(\omega)| > 0, \quad (7.31)$$

then

$$\hat{\phi}(\omega) = \prod_{p=1}^{+\infty} \frac{\hat{h}(2^{-p}\omega)}{\sqrt{2}} \quad (7.32)$$

is the Fourier transform of a scaling function  $\phi \in \mathbf{L}^2(\mathbb{R})$ .

**Proof.** This theorem is a central result and its proof is long and technical. It is divided in several parts.

**Proof of the necessary condition (7.29).** The necessary condition is proved to be a consequence of the fact that  $\{\phi(t-n)\}_{n \in \mathbb{Z}}$  is orthonormal. In the Fourier domain, (7.14)

gives an equivalent condition:

$$\forall \omega \in \mathbb{R}, \quad \sum_{k=-\infty}^{+\infty} |\hat{\phi}(\omega + 2k\pi)|^2 = 1. \quad (7.33)$$

Inserting  $\hat{\phi}(\omega) = 2^{-1/2}\hat{h}(\omega/2)\hat{\phi}(\omega/2)$  yields

$$\sum_{k=-\infty}^{+\infty} \left| \hat{h}\left(\frac{\omega}{2} + k\pi\right) \right|^2 \left| \hat{\phi}\left(\frac{\omega}{2} + k\pi\right) \right|^2 = 2.$$

Since  $\hat{h}(\omega)$  is  $2\pi$  periodic, separating the even and odd integer terms gives

$$\left| \hat{h}\left(\frac{\omega}{2}\right) \right|^2 \sum_{p=-\infty}^{+\infty} \left| \hat{\phi}\left(\frac{\omega}{2} + 2p\pi\right) \right|^2 + \left| \hat{h}\left(\frac{\omega}{2} + \pi\right) \right|^2 \sum_{p=-\infty}^{+\infty} \left| \hat{\phi}\left(\frac{\omega}{2} + \pi + 2p\pi\right) \right|^2 = 2.$$

Inserting (7.33) for  $\omega' = \omega/2$  and  $\omega' = \omega/2 + \pi$  proves that

$$|\hat{h}(\omega')|^2 + |\hat{h}(\omega' + \pi)|^2 = 2.$$

**Proof of the necessary condition (7.30).** We prove that  $\hat{h}(0) = \sqrt{2}$  by showing that  $\hat{\phi}(0) \neq 0$ . Indeed, we know that  $\hat{\phi}(0) = 2^{-1/2}\hat{h}(0)\hat{\phi}(0)$ . More precisely, we verify that  $|\hat{\phi}(0)| = 1$  is a consequence of the completeness property (7.5) of multiresolution approximations.

The orthogonal projection of  $f \in \mathbf{L}^2(\mathbb{R})$  on  $\mathbf{V}_j$  is

$$P_{\mathbf{V}_j} f = \sum_{n=-\infty}^{+\infty} \langle f, \phi_{j,n} \rangle \phi_{j,n}. \quad (7.34)$$

Property (7.5) expressed in the time and Fourier domains with the Plancherel formula implies that

$$\lim_{j \rightarrow -\infty} \|f - P_{\mathbf{V}_j} f\|^2 = \lim_{j \rightarrow -\infty} 2\pi \|\hat{f} - \widehat{P_{\mathbf{V}_j} f}\|^2 = 0. \quad (7.35)$$

To compute the Fourier transform  $\widehat{P_{\mathbf{V}_j} f}(\omega)$ , we denote  $\phi_j(t) = \sqrt{2^{-j}}\phi(2^{-j}t)$ . Inserting the convolution expression (7.17) in (7.34) yields

$$P_{\mathbf{V}_j} f(t) = \sum_{n=-\infty}^{+\infty} f \star \bar{\phi}_j(2^j n) \phi_j(t - 2^j n) = \phi_j \star \sum_{n=-\infty}^{+\infty} f \star \bar{\phi}_j(2^j n) \delta(t - 2^j n).$$

The Fourier transform of  $f \star \bar{\phi}_j(t)$  is  $\sqrt{2^j} \hat{f}(\omega) \hat{\phi}^*(2^j \omega)$ . A uniform sampling has a periodized Fourier transform calculated in (3.3), and thus,

$$\widehat{P_{\mathbf{V}_j} f}(\omega) = \hat{\phi}(2^j \omega) \sum_{k=-\infty}^{+\infty} \hat{f}\left(\omega - \frac{2k\pi}{2^j}\right) \hat{\phi}^*\left(2^j \left[\omega - \frac{2k\pi}{2^j}\right]\right). \quad (7.36)$$

Let us choose  $\hat{f} = \mathbf{1}_{[-\pi, \pi]}$ . For  $j < 0$  and  $\omega \in [-\pi, \pi]$ , (7.36) gives  $\widehat{P_{\mathbf{V}_j} f}(\omega) = |\hat{\phi}(2^j \omega)|^2$ . The mean-square convergence (7.35) implies that

$$\lim_{j \rightarrow -\infty} \int_{-\pi}^{\pi} \left| 1 - |\hat{\phi}(2^j \omega)|^2 \right|^2 d\omega = 0.$$

Since  $\phi$  is integrable,  $\hat{\phi}(\omega)$  is continuous and thus  $\lim_{j \rightarrow -\infty} |\hat{\phi}(2^j \omega)| = |\hat{\phi}(0)| = 1$ .

We now prove that the function  $\phi$ , having a Fourier transform given by (7.32), is a scaling function. This is divided in two intermediate results.

**Proof that  $\{\phi(t-n)\}_{n \in \mathbb{Z}}$  is orthonormal.** Observe first that the infinite product (7.32) converges and that  $|\hat{\phi}(\omega)| \leq 1$  because (7.29) implies that  $|\hat{h}(\omega)| \leq \sqrt{2}$ . The Parseval formula gives

$$\langle \phi(t), \phi(t-n) \rangle = \int_{-\infty}^{+\infty} \phi(t) \phi^*(t-n) dt = \frac{1}{2\pi} \int_{-\infty}^{+\infty} |\hat{\phi}(\omega)|^2 e^{in\omega} d\omega.$$

Verifying that  $\{\phi(t-n)\}_{n \in \mathbb{Z}}$  is orthonormal is equivalent to showing that

$$\int_{-\infty}^{+\infty} |\hat{\phi}(\omega)|^2 e^{in\omega} d\omega = 2\pi \delta[n].$$

This result is obtained by considering the functions

$$\hat{\phi}_k(\omega) = \prod_{p=1}^k \frac{\hat{h}(2^{-p}\omega)}{\sqrt{2}} \mathbf{1}_{[-2^k\pi, 2^k\pi]}(\omega),$$

and computing the limit, as  $k$  increases to  $+\infty$ , of the integrals

$$I_k[n] = \int_{-\infty}^{+\infty} |\hat{\phi}_k(\omega)|^2 e^{in\omega} d\omega = \int_{-2^k\pi}^{2^k\pi} \prod_{p=1}^k \frac{|\hat{h}(2^{-p}\omega)|^2}{2} e^{in\omega} d\omega.$$

First, let us show that  $I_k[n] = 2\pi\delta[n]$  for all  $k \geq 1$ . To do this, we divide  $I_k[n]$  into two integrals:

$$I_k[n] = \int_{-2^k\pi}^0 \prod_{p=1}^k \frac{|\hat{h}(2^{-p}\omega)|^2}{2} e^{in\omega} d\omega + \int_0^{2^k\pi} \prod_{p=1}^k \frac{|\hat{h}(2^{-p}\omega)|^2}{2} e^{in\omega} d\omega.$$

Let us make the change of variable  $\omega' = \omega + 2^k\pi$  in the first integral. Since  $\hat{h}(\omega)$  is  $2\pi$  periodic, when  $p < k$ , then  $|\hat{h}(2^{-p}[\omega' - 2^k\pi])|^2 = |\hat{h}(2^{-p}\omega')|^2$ . When  $k = p$  the hypothesis (7.29) implies that

$$|\hat{h}(2^{-k}[\omega' - 2^k\pi])|^2 + |\hat{h}(2^{-k}\omega')|^2 = 2.$$

For  $k > 1$ , the two integrals of  $I_k[n]$  become

$$I_k[n] = \int_0^{2^k\pi} \prod_{p=1}^{k-1} \frac{|\hat{h}(2^{-p}\omega)|^2}{2} e^{in\omega} d\omega. \quad (7.37)$$

Since  $\prod_{p=1}^{k-1} |\hat{h}(2^{-p}\omega)|^2 e^{in\omega}$  is  $2^k\pi$  periodic we obtain  $I_k[n] = I_{k-1}[n]$ , and by induction  $I_k[n] = I_1[n]$ . Writing (7.37) for  $k = 1$  gives

$$I_1[n] = \int_0^{2\pi} e^{in\omega} d\omega = 2\pi \delta[n],$$

which verifies that  $I_k[n] = 2\pi\delta[n]$ , for all  $k \geq 1$ .

We shall now prove that  $\hat{\phi} \in \mathbf{L}^2(\mathbb{R})$ . For all  $\omega \in \mathbb{R}$ ,

$$\lim_{k \rightarrow \infty} |\hat{\phi}_k(\omega)|^2 = \prod_{p=1}^{\infty} \frac{|\hat{h}(2^{-p}\omega)|^2}{2} = |\hat{\phi}(\omega)|^2.$$

The Fatou lemma (A.1) on positive functions proves that

$$\int_{-\infty}^{+\infty} |\hat{\phi}(\omega)|^2 d\omega \leq \lim_{k \rightarrow \infty} \int_{-\infty}^{+\infty} |\hat{\phi}_k(\omega)|^2 d\omega = 2\pi, \quad (7.38)$$

because  $I_k[0] = 2\pi$  for all  $k \geq 1$ . Since

$$|\hat{\phi}(\omega)|^2 e^{in\omega} = \lim_{k \rightarrow \infty} |\hat{\phi}_k(\omega)|^2 e^{in\omega},$$

we finally verify that

$$\int_{-\infty}^{+\infty} |\hat{\phi}(\omega)|^2 e^{in\omega} d\omega = \lim_{k \rightarrow \infty} \int_{-\infty}^{+\infty} |\hat{\phi}_k(\omega)|^2 e^{in\omega} d\omega = 2\pi \delta[n] \quad (7.39)$$

by applying the dominated convergence theorem (A.1). This requires verifying the upper-bound condition (A.1). This is done in our case by proving the existence of a constant  $C$  such that

$$|\hat{\phi}_k(\omega)|^2 e^{in\omega} = |\hat{\phi}_k(\omega)|^2 \leq C |\hat{\phi}(\omega)|^2. \quad (7.40)$$

Indeed, we showed in (7.38) that  $|\hat{\phi}(\omega)|^2$  is an integrable function.

The existence of  $C > 0$  satisfying (7.40) is trivial for  $|\omega| > 2^k \pi$  since  $\hat{\phi}_k(\omega) = 0$ . For  $|\omega| \leq 2^k \pi$  since  $\hat{\phi}(\omega) = 2^{-1/k} \hat{h}(\omega/2) \hat{\phi}(\omega/2)$ , it follows that

$$|\hat{\phi}(\omega)|^2 = |\hat{\phi}_k(\omega)|^2 |\hat{\phi}(2^{-k}\omega)|^2.$$

Therefore, to prove (7.40) for  $|\omega| \leq 2^k \pi$ , it is sufficient to show that  $|\hat{\phi}(\omega)|^2 \geq 1/C$  for  $\omega \in [-\pi, \pi]$ .

Let us first study the neighborhood of  $\omega = 0$ . Since  $\hat{h}(\omega)$  is continuously differentiable in this neighborhood and since  $|\hat{h}(\omega)|^2 \leq 2 = |\hat{h}(0)|^2$ , the functions  $|\hat{h}(\omega)|^2$  and  $\log_e |\hat{h}(\omega)|^2$  have derivatives that vanish at  $\omega = 0$ . It follows that there exists  $\varepsilon > 0$  such that

$$\forall |\omega| \leq \varepsilon, \quad 0 \geq \log_e \left( \frac{|\hat{h}(\omega)|^2}{2} \right) \geq -|\omega|.$$

Thus, for  $|\omega| \leq \varepsilon$

$$|\hat{\phi}(\omega)|^2 = \exp \left[ \sum_{p=1}^{+\infty} \log_e \left( \frac{|\hat{h}(2^{-p}\omega)|^2}{2} \right) \right] \geq e^{-|\omega|} \geq e^{-\varepsilon}. \quad (7.41)$$

Now let us analyze the domain  $|\omega| > \varepsilon$ . To do this we take an integer  $l$  such that  $2^{-l} \pi < \varepsilon$ . Condition (7.31) proves that  $K = \inf_{\omega \in [-\pi/2, \pi/2]} |\hat{h}(\omega)| > 0$ , so if  $|\omega| \leq \pi$ ,

$$|\hat{\phi}(\omega)|^2 = \prod_{p=1}^l \frac{|\hat{h}(2^{-p}\omega)|^2}{2} |\hat{\phi}(2^{-l}\omega)|^2 \geq \frac{K^{2l}}{2^l} e^{-\varepsilon} = \frac{1}{C}.$$

This last result finishes the proof of inequality (7.40). Applying the dominated convergence theorem (A.1) proves (7.39) and that  $\{\phi(t-n)\}_{n \in \mathbb{Z}}$  is orthonormal. A simple change of variable shows that  $\{\phi_{j,n}\}_{j \in \mathbb{Z}}$  is orthonormal for all  $j \in \mathbb{Z}$ .

**Proof that  $\{\mathbf{V}_j\}_{j \in \mathbb{Z}}$  is a multiresolution approximation.** To verify that  $\phi$  is a scaling function, we must show that the spaces  $\mathbf{V}_j$  generated by  $\{\phi_{j,n}\}_{j \in \mathbb{Z}}$  define a multiresolution approximation. The multiresolution properties (7.1) and (7.3) are clearly true. The causality  $\mathbf{V}_{j+1} \subset \mathbf{V}_j$  is verified by showing that for any  $p \in \mathbb{Z}$ ,

$$\phi_{j+1,p} = \sum_{n=-\infty}^{+\infty} h[n-2p] \phi_{j,n}.$$

This equality is proved later in (7.107). Since all vectors of a basis of  $\mathbf{V}_{j+1}$  can be decomposed in a basis of  $\mathbf{V}_j$ , it follows that  $\mathbf{V}_{j+1} \subset \mathbf{V}_j$ .

To prove the multiresolution property (7.4) we must show that any  $f \in \mathbf{L}^2(\mathbb{R})$  satisfies

$$\lim_{j \rightarrow +\infty} \|P_{\mathbf{V}_j} f\| = 0. \quad (7.42)$$

Since  $\{\phi_{j,n}\}_{n \in \mathbb{Z}}$  is an orthonormal basis of  $\mathbf{V}_j$ ,

$$\|P_{\mathbf{V}_j} f\|^2 = \sum_{n=-\infty}^{+\infty} |\langle f, \phi_{j,n} \rangle|^2.$$

Suppose first that  $f$  is bounded by  $A$  and has a compact support included in  $[2^J, 2^J]$ . The constants  $A$  and  $J$  may be arbitrarily large. It follows that

$$\begin{aligned} \sum_{n=-\infty}^{+\infty} |\langle f, \phi_{j,n} \rangle|^2 &\leq 2^{-j} \left[ \sum_{n=-\infty}^{+\infty} \int_{-2^J}^{2^J} |f(t)| |\phi(2^{-j}t - n)| dt \right]^2 \\ &\leq 2^{-j} A^2 \left[ \sum_{n=-\infty}^{+\infty} \int_{-2^J}^{2^J} |\phi(2^{-j}t - n)| dt \right]^2. \end{aligned}$$

Applying the Cauchy-Schwarz inequality to  $1 \times |\phi(2^{-j}t - n)|$  yields

$$\begin{aligned} \sum_{n=-\infty}^{+\infty} |\langle f, \phi_{j,n} \rangle|^2 &\leq A^2 2^{J+1} \sum_{n=-\infty}^{+\infty} \int_{-2^J}^{2^J} |\phi(2^{-j}t - n)|^2 2^{-j} dt \\ &\leq A^2 2^{J+1} \int_{S_j} |\phi(t)|^2 dt = A^2 2^{J+1} \int_{-\infty}^{+\infty} |\phi(t)|^2 \mathbf{1}_{S_j}(t) dt, \end{aligned}$$

with  $S_j = \cup_{n \in \mathbb{Z}} [n - 2^{J-j}, n + 2^{J-j}]$  for  $j > J$ . For  $t \notin \mathbb{Z}$ , we obviously have  $\mathbf{1}_{S_j}(t) \rightarrow 0$  for  $j \rightarrow +\infty$ . The dominated convergence theorem (A.1) applied to  $|\phi(t)|^2 \mathbf{1}_{S_j}(t)$  proves that the integral converges to 0 and thus,

$$\lim_{j \rightarrow +\infty} \sum_{n=-\infty}^{+\infty} |\langle f, \phi_{j,n} \rangle|^2 = 0.$$

Property (7.42) is extended to any  $f \in \mathbf{L}^2(\mathbb{R})$  by using the density in  $\mathbf{L}^2(\mathbb{R})$  of bounded function with a compact support, and Theorem A.5.

To prove the last multiresolution property (7.5) we must show that for any  $f \in \mathbf{L}^2(\mathbb{R})$ ,

$$\lim_{j \rightarrow -\infty} \|f - P_{V_j} f\|^2 = \lim_{j \rightarrow -\infty} (\|f\|^2 - \|P_{V_j} f\|^2) = 0. \quad (7.43)$$

We consider functions  $f$  that have a Fourier transform  $\hat{f}$  that has a compact support included in  $[-2^J \pi, 2^J \pi]$  for  $J$  large enough. We proved in (7.36) that the Fourier transform of  $P_{V_j} f$  is

$$\widehat{P_{V_j} f}(\omega) = \hat{\phi}(2^j \omega) \sum_{k=-\infty}^{+\infty} \hat{f}(\omega - 2^{-j} 2k\pi) \hat{\phi}^*(2^j [\omega - 2^{-j} 2k\pi]).$$

If  $j < -J$ , then the supports of  $\hat{f}(\omega - 2^{-j} 2k\pi)$  are disjoint for different  $k$ , so

$$\begin{aligned} \|P_{V_j} f\|^2 &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} |\hat{f}(\omega)|^2 |\hat{\phi}(2^j \omega)|^4 d\omega \\ &\quad + \frac{1}{2\pi} \int_{-\infty}^{+\infty} \sum_{\substack{k=-\infty \\ k \neq 0}}^{+\infty} |\hat{f}(\omega - 2^{-j} 2k\pi)|^2 |\hat{\phi}(2^j \omega)|^2 |\hat{\phi}(2^j [\omega - 2^{-j} 2k\pi])|^2 d\omega. \end{aligned} \quad (7.44)$$

We have already observed that  $|\phi(\omega)| \leq 1$  and (7.41) proves that if  $\omega$  is sufficiently small then  $|\phi(\omega)| \geq e^{-|\omega|}$ , so

$$\lim_{\omega \rightarrow 0} |\hat{\phi}(\omega)| = 1.$$

Since  $|\hat{f}(\omega)|^2 |\hat{\phi}(2^j \omega)|^4 \leq |\hat{f}(\omega)|^2$  and  $\lim_{j \rightarrow -\infty} |\hat{\phi}(2^j \omega)|^4 |\hat{f}(\omega)|^2 = |\hat{f}(\omega)|^2$ , one can apply the dominated convergence theorem (A.1), to prove that

$$\lim_{j \rightarrow -\infty} \int_{-\infty}^{+\infty} |\hat{f}(\omega)|^2 |\hat{\phi}(2^j \omega)|^4 d\omega = \int_{-\infty}^{+\infty} |\hat{f}(\omega)|^2 d\omega = \|f\|^2. \quad (7.45)$$

The operator  $P_{V_j}$  is an orthogonal projector, so  $\|P_{V_j} f\| \leq \|f\|$ . With (7.44) and (7.45), this implies that  $\lim_{j \rightarrow -\infty} (\|f\|^2 - \|P_{V_j} f\|^2) = 0$  and thus verifies (7.43). This property is extended to any  $f \in \mathbf{L}^2(\mathbb{R})$  by using the density in  $\mathbf{L}^2(\mathbb{R})$  of functions having compactly supported Fourier transforms and the result of Theorem A.5. ■

Discrete filters that have transfer functions that satisfy (7.29) are called *conjugate mirror filters*. As we shall see in Section 7.3, they play an important role in discrete signal processing; they make it possible to decompose discrete signals in separate frequency bands with filter banks. One difficulty of the proof is showing that the infinite cascade of convolutions that is represented in the Fourier domain by the product (7.32) does converge to a decent function in  $\mathbf{L}^2(\mathbb{R})$ . The sufficient condition (7.31) is not necessary to construct a scaling function, but it is always satisfied in practical designs of conjugate mirror filters. It cannot just be removed as shown by the example  $\hat{h}(\omega) = \cos(3\omega/2)$ , which satisfies all other conditions. In this case, a simple calculation shows that  $\phi = 1/3 \mathbf{1}_{[-3/2, 3/2]}$ . Clearly  $\{\phi(t-n)\}_{n \in \mathbb{Z}}$  is not orthogonal, so  $\phi$  is not a scaling function. However, the condition (7.31) may be replaced by a weaker but more technical necessary and sufficient condition proved by Cohen [16, 167].

---

**EXAMPLE 7.6**

For a Shannon multiresolution approximation,  $\hat{\phi} = \mathbf{1}_{[-\pi, \pi]}$ . Thus, we derive from (7.32) that

$$\forall \omega \in [-\pi, \pi], \quad \hat{h}(\omega) = \sqrt{2} \mathbf{1}_{[-\pi/2, \pi/2]}(\omega).$$


---

---

**EXAMPLE 7.7**

For piecewise constant approximations,  $\phi = \mathbf{1}_{[0, 1]}$ . Since  $h[n] = \langle 2^{-1/2}\phi(t/2), \phi(t-n) \rangle$ , it follows that

$$h[n] = \begin{cases} 2^{-1/2} & \text{if } n = 0, 1 \\ 0 & \text{otherwise} \end{cases} \quad (7.46)$$


---

---

**EXAMPLE 7.8**

Polynomial splines of degree  $m$  correspond to a conjugate mirror filter  $\hat{h}(\omega)$  that is calculated from  $\hat{\phi}(\omega)$  with (7.25):

$$\hat{h}(\omega) = \sqrt{2} \frac{\hat{\phi}(2\omega)}{\hat{\phi}(\omega)}. \quad (7.47)$$

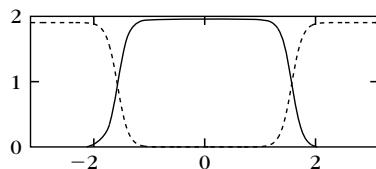
Inserting (7.18) yields

$$\hat{h}(\omega) = \exp\left(\frac{-i\varepsilon\omega}{2}\right) \sqrt{\frac{S_{2m+2}(\omega)}{2^{2m+1} S_{2m+2}(2\omega)}}, \quad (7.48)$$

where  $\varepsilon = 0$  if  $m$  is odd and  $\varepsilon = 1$  if  $m$  is even. For linear splines  $m = 1$ , so (7.20) implies that

$$\hat{h}(\omega) = \sqrt{2} \left[ \frac{1 + 2 \cos^2(\omega/2)}{1 + 2 \cos^2 \omega} \right]^{1/2} \cos^2\left(\frac{\omega}{2}\right). \quad (7.49)$$

For cubic splines, the conjugate mirror filter is calculated by inserting (7.22) in (7.48). Figure 7.4 gives the graph of  $|\hat{h}(\omega)|^2$ . The impulse responses  $h[n]$  of these filters have an infinite support but an exponential decay. For  $m$  odd,  $h[n]$  is symmetric about  $n = 0$ . Table 7.1 gives the coefficients  $h[n]$  above  $10^{-4}$  for  $m = 1, 3$ .



**FIGURE 7.4**

The solid line gives  $|\hat{h}(\omega)|^2$  on  $[-\pi, \pi]$  for a cubic spline multiresolution. The dotted line corresponds to  $|\hat{g}(\omega)|^2$ .

**Table 7.1** Conjugate Mirror Filters  $h[n]$  for Linear Splines  $m=1$  and Cubic Splines  $m=3$

	$n$	$h[n]$		$n$	$h[n]$
$m=1$	0	0.817645956	$m=3$	5, -5	0.042068328
	1, -1	0.397296430		6, -6	-0.017176331
	2, -2	-0.069101020		7, -7	-0.017982291
	3, -3	-0.051945337		8, -8	0.008685294
	4, -4	0.016974805		9, -9	0.008201477
	5, -5	0.009990599		10, -10	-0.004353840
	6, -6	-0.003883261		11, -11	-0.003882426
	7, -7	-0.002201945		12, -12	0.002186714
	8, -8	0.000923371		13, -13	0.001882120
	9, -9	0.000511636		14, -14	-0.001103748
	10, -10	-0.000224296		15, -15	-0.000927187
	11, -11	-0.000122686		16, -16	0.000559952
$m=3$	0	0.766130398		17, -17	0.000462093
	1, -1	0.433923147		18, -18	-0.000285414
	2, -2	-0.050201753		19, -19	-0.000232304
	3, -3	-0.110036987		20, -20	0.000146098
	4, -4	0.032080869			

Note: The coefficients below  $10^{-4}$  are not given.

### 7.1.4 In Which Orthogonal Wavelets Finally Arrive

Orthonormal wavelets carry the details necessary to increase the resolution of a signal approximation. The approximations of  $f$  at the scales  $2^j$  and  $2^{j-1}$  are, respectively, equal to their orthogonal projections on  $\mathbf{V}_j$  and  $\mathbf{V}_{j-1}$ . We know that  $\mathbf{V}_j$  is included in  $\mathbf{V}_{j-1}$ . Let  $\mathbf{W}_j$  be the orthogonal complement of  $\mathbf{V}_j$  in  $\mathbf{V}_{j-1}$ :

$$\mathbf{V}_{j-1} = \mathbf{V}_j \oplus \mathbf{W}_j. \quad (7.50)$$

The orthogonal projection of  $f$  on  $\mathbf{V}_{j-1}$  can be decomposed as the sum of orthogonal projections on  $\mathbf{V}_j$  and  $\mathbf{W}_j$ :

$$P_{\mathbf{V}_{j-1}} f = P_{\mathbf{V}_j} f + P_{\mathbf{W}_j} f. \quad (7.51)$$

The complement  $P_{\mathbf{W}_j} f$  provides the “details” of  $f$  that appear at the scale  $2^{j-1}$  but that disappear at the coarser scale  $2^j$ . Theorem 7.3 [44, 362] proves that one can construct an orthonormal basis of  $\mathbf{W}_j$  by scaling and translating a wavelet  $\psi$ .

**Theorem 7.3:** Mallat, Meyer. Let  $\phi$  be a scaling function and  $h$  the corresponding conjugate mirror filter. Let  $\psi$  be the function having a Fourier transform

$$\hat{\psi}(\omega) = \frac{1}{\sqrt{2}} \hat{g}\left(\frac{\omega}{2}\right) \hat{\phi}\left(\frac{\omega}{2}\right), \quad (7.52)$$

with

$$\hat{g}(\omega) = e^{-i\omega} \hat{h}^*(\omega + \pi). \quad (7.53)$$

Let us denote

$$\psi_{j,n}(t) = \frac{1}{\sqrt{2^j}} \psi\left(\frac{t - 2^j n}{2^j}\right).$$

For any scale  $2^j$ ,  $\{\psi_{j,n}\}_{n \in \mathbb{Z}}$  is an orthonormal basis of  $\mathbf{W}_j$ . For all scales,  $\{\psi_{j,n}\}_{(j,n) \in \mathbb{Z}^2}$  is an orthonormal basis of  $\mathbf{L}^2(\mathbb{R})$ .

**Proof.** Let us prove first that  $\hat{\psi}$  can be written as the product (7.52). Necessarily,  $\psi(t/2) \in \mathbf{W}_1 \subset \mathbf{V}_0$ . Thus, it can be decomposed in  $\{\phi(t-n)\}_{n \in \mathbb{Z}}$ , which is an orthogonal basis of  $\mathbf{V}_0$ :

$$\frac{1}{\sqrt{2}} \psi\left(\frac{t}{2}\right) = \sum_{n=-\infty}^{+\infty} g[n] \phi(t-n), \quad (7.54)$$

with

$$g[n] = \frac{1}{\sqrt{2}} \left\langle \psi\left(\frac{t}{2}\right), \phi(t-n) \right\rangle. \quad (7.55)$$

The Fourier transform of (7.54) yields

$$\hat{\psi}(2\omega) = \frac{1}{\sqrt{2}} \hat{g}(\omega) \hat{\phi}(\omega). \quad (7.56)$$

Lemma 7.1 gives necessary and sufficient conditions on  $\hat{g}$  for designing an orthogonal wavelet.

**Lemma 7.1.** The family  $\{\psi_{j,n}\}_{n \in \mathbb{Z}}$  is an orthonormal basis of  $\mathbf{W}_j$  if and only if

$$|\hat{g}(\omega)|^2 + |\hat{g}(\omega + \pi)|^2 = 2 \quad (7.57)$$

and

$$\hat{g}(\omega) \hat{h}^*(\omega) + \hat{g}(\omega + \pi) \hat{h}^*(\omega + \pi) = 0. \quad (7.58)$$

The lemma is proved for  $j=0$  from which it is easily extended to  $j \neq 0$  with an appropriate scaling. As in (7.14), one can verify that  $\{\psi(t-n)\}_{n \in \mathbb{Z}}$  is orthonormal if and only if

$$\forall \omega \in \mathbb{R}, \quad I(\omega) = \sum_{k=-\infty}^{+\infty} |\hat{\psi}(\omega + 2k\pi)|^2 = 1. \quad (7.59)$$

Since  $\hat{\psi}(\omega) = 2^{-1/2} \hat{g}(\omega/2) \hat{\phi}(\omega/2)$  and  $\hat{g}(\omega)$  is  $2\pi$  periodic,

$$\begin{aligned} I(\omega) &= \sum_{k=-\infty}^{+\infty} |\hat{g}\left(\frac{\omega}{2} + k\pi\right)|^2 |\hat{\phi}\left(\frac{\omega}{2} + k\pi\right)|^2 \\ &= |\hat{g}\left(\frac{\omega}{2}\right)|^2 \sum_{p=-\infty}^{+\infty} |\hat{\phi}\left(\frac{\omega}{2} + 2p\pi\right)|^2 + |\hat{g}\left(\frac{\omega}{2} + \pi\right)|^2 \sum_{p=-\infty}^{+\infty} |\hat{\phi}\left(\frac{\omega}{2} + \pi + 2p\pi\right)|^2. \end{aligned}$$

We know that  $\sum_{p=-\infty}^{+\infty} |\hat{\phi}(\omega + 2p\pi)|^2 = 1$ , so (7.59) is equivalent to (7.57).

The space  $\mathbf{W}_0$  is orthogonal to  $\mathbf{V}_0$  if and only if  $\{\phi(t-n)\}_{n \in \mathbb{Z}}$  and  $\{\psi(t-n)\}_{n \in \mathbb{Z}}$  are orthogonal families of vectors. This means that for any  $n \in \mathbb{Z}$ ,

$$\langle \psi(t), \phi(t-n) \rangle = \psi \star \bar{\phi}(n) = 0.$$

The Fourier transform of  $\psi \star \bar{\phi}(t)$  is  $\hat{\psi}(\omega) \hat{\phi}^*(\omega)$ . The sampled sequence  $\psi \star \bar{\phi}(n)$  is zero if its Fourier series computed with (3.3) satisfies

$$\forall \omega \in \mathbb{R}, \quad \sum_{k=-\infty}^{+\infty} \hat{\psi}(\omega + 2k\pi) \hat{\phi}^*(\omega + 2k\pi) = 0. \quad (7.60)$$

By inserting  $\hat{\psi}(\omega) = 2^{-1/2} \hat{g}(\omega/2) \hat{\phi}(\omega/2)$  and  $\hat{\phi}(\omega) = 2^{-1/2} \hat{h}(\omega/2) \hat{\phi}(\omega/2)$  in this equation, since  $\sum_{k=-\infty}^{+\infty} |\hat{\phi}(\omega + 2k\pi)|^2 = 1$ , we prove as before that (7.60) is equivalent to (7.58).

We must finally verify that  $\mathbf{V}_{-1} = \mathbf{V}_0 \oplus \mathbf{W}_0$ . Knowing that  $\{\sqrt{2}\phi(2t-n)\}_{n \in \mathbb{Z}}$  is an orthogonal basis of  $\mathbf{V}_{-1}$ , it is equivalent to show that for any  $a[n] \in \ell^2(\mathbb{Z})$  there exist  $b[n] \in \ell^2(\mathbb{Z})$  and  $c[n] \in \ell^2(\mathbb{Z})$  such that

$$\sum_{n=-\infty}^{+\infty} a[n] \sqrt{2} \phi(2[t - 2^{-1}n]) = \sum_{n=-\infty}^{+\infty} b[n] \phi(t-n) + \sum_{n=-\infty}^{+\infty} c[n] \psi(t-n). \quad (7.61)$$

This is done by relating  $\hat{b}(\omega)$  and  $\hat{c}(\omega)$  to  $\hat{a}(\omega)$ . The Fourier transform of (7.61) yields

$$\frac{1}{\sqrt{2}} \hat{a}\left(\frac{\omega}{2}\right) \hat{\phi}\left(\frac{\omega}{2}\right) = \hat{b}(\omega) \hat{\phi}(\omega) + \hat{c}(\omega) \hat{\psi}(\omega).$$

Inserting  $\hat{\psi}(\omega) = 2^{-1/2} \hat{g}(\omega/2) \hat{\phi}(\omega/2)$  and  $\hat{\phi}(\omega) = 2^{-1/2} \hat{h}(\omega/2) \hat{\phi}(\omega/2)$  in this equation shows that it is necessarily satisfied if

$$\hat{a}\left(\frac{\omega}{2}\right) = \hat{b}(\omega) \hat{h}\left(\frac{\omega}{2}\right) + \hat{c}(\omega) \hat{g}\left(\frac{\omega}{2}\right). \quad (7.62)$$

Let us define

$$\hat{b}(2\omega) = \frac{1}{2} [\hat{a}(\omega) \hat{h}^*(\omega) + \hat{a}(\omega + \pi) \hat{h}^*(\omega + \pi)]$$

and

$$\hat{c}(2\omega) = \frac{1}{2} [\hat{a}(\omega) \hat{g}^*(\omega) + \hat{a}(\omega + \pi) \hat{g}^*(\omega + \pi)].$$

When calculating the right side of (7.62), we verify that it is equal to the left side by inserting (7.57), (7.58), and using

$$|\hat{h}(\omega)|^2 + |\hat{h}(\omega + \pi)|^2 = 2. \quad (7.63)$$

Since  $\hat{b}(\omega)$  and  $\hat{c}(\omega)$  are  $2\pi$  periodic they are the Fourier series of two sequences  $b[n]$  and  $c[n]$  that satisfy (7.61). This finishes the proof of the lemma.

The formula (7.53)

$$\hat{g}(\omega) = e^{-i\omega} \hat{h}^*(\omega + \pi)$$

satisfies (7.57) and (7.58) because of (7.63). Thus, we derive from Lemma 7.1 that  $\{\psi_{j,n}\}_{(j,n) \in \mathbb{Z}^2}$  is an orthogonal basis of  $\mathbf{W}_j$ .

We complete the proof of the theorem by verifying that  $\{\psi_{j,n}\}_{(j,n) \in \mathbb{Z}^2}$  is an orthogonal basis of  $\mathbf{L}^2(\mathbb{R})$ . Observe first that the detail spaces  $\{\mathbf{W}_j\}_{j \in \mathbb{Z}}$  are orthogonal. Indeed,  $\mathbf{W}_j$  is

orthogonal to  $\mathbf{V}_j$  and  $\mathbf{W}_l \subset \mathbf{V}_{l-1} \subset \mathbf{V}_j$  for  $j < l$ . Thus,  $\mathbf{W}_j$  and  $\mathbf{W}_l$  are orthogonal. We can also decompose

$$\mathbf{L}^2(\mathbb{R}) = \bigoplus_{j=-\infty}^{+\infty} \mathbf{W}_j. \quad (7.64)$$

Indeed,  $\mathbf{V}_{j-1} = \mathbf{W}_j \oplus \mathbf{V}_j$ , and we verify by substitution that for any  $L > J$ ,

$$\mathbf{V}_L = \bigoplus_{j=L-1}^J \mathbf{W}_j \oplus \mathbf{V}_J. \quad (7.65)$$

Since  $\{\mathbf{V}_j\}_{j \in \mathbb{Z}}$  is a multiresolution approximation,  $\mathbf{V}_L$  and  $\mathbf{V}_J$  tend, respectively, to  $\mathbf{L}^2(\mathbb{R})$  and  $\{0\}$  when  $L$  and  $J$  go, respectively, to  $-\infty$  and  $+\infty$ , which implies (7.64). Therefore, a union of orthonormal bases of all  $\mathbf{W}_j$  is an orthonormal basis of  $\mathbf{L}^2(\mathbb{R})$ . ■

The proof of the theorem shows that  $\hat{g}$  is the Fourier series of

$$g[n] = \left\langle \frac{1}{\sqrt{2}} \psi\left(\frac{t}{2}\right), \phi(t-n) \right\rangle, \quad (7.66)$$

which are the decomposition coefficients of

$$\frac{1}{\sqrt{2}} \psi\left(\frac{t}{2}\right) = \sum_{n=-\infty}^{+\infty} g[n] \phi(t-n). \quad (7.67)$$

Calculating the inverse Fourier transform of (7.53) yields

$$g[n] = (-1)^{1-n} h[1-n]. \quad (7.68)$$

This mirror filter plays an important role in the fast wavelet transform algorithm.

### EXAMPLE 7.9

Figure 7.5 displays the cubic spline wavelet  $\psi$  and its Fourier transform  $\hat{\psi}$  calculated by inserting in (7.52) the expressions (7.18) and (7.48) of  $\hat{\phi}(\omega)$  and  $\hat{h}(\omega)$ . The properties of

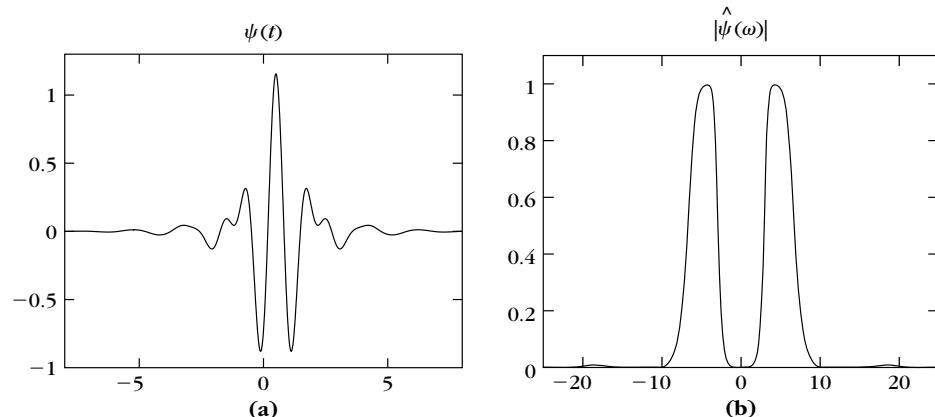
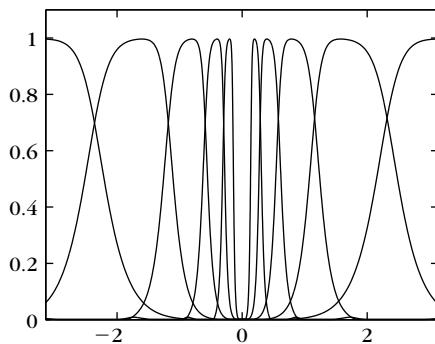


FIGURE 7.5

Battle-Lemarié cubic spline wavelet  $\psi$  (a) and its Fourier transform modulus (b).

**FIGURE 7.6**

Graph of  $|\hat{\psi}(2^j\omega)|^2$  for the cubic spline Battle-Lemarié wavelet, with  $1 \leq j \leq 5$  and  $\omega \in [-\pi, \pi]$ .

this Battle-Lemarié spline wavelet are further studied in Section 7.2.2. Like most orthogonal wavelets, the energy of  $\hat{\psi}$  is essentially concentrated in  $[-2\pi, -\pi] \cup [\pi, 2\pi]$ . For any  $\psi$  that generates an orthogonal basis of  $L^2(\mathbb{R})$ , one can verify that

$$\forall \omega \in \mathbb{R} - \{0\}, \quad \sum_{j=-\infty}^{+\infty} |\hat{\psi}(2^j\omega)|^2 = 1.$$

This is illustrated in Figure 7.6.

The orthogonal projection of a signal  $f$  in a “detail” space  $\mathbf{W}_j$  is obtained with a partial expansion in its wavelet basis:

$$P\mathbf{W}_j f = \sum_{n=-\infty}^{+\infty} \langle f, \psi_{j,n} \rangle \psi_{j,n}.$$

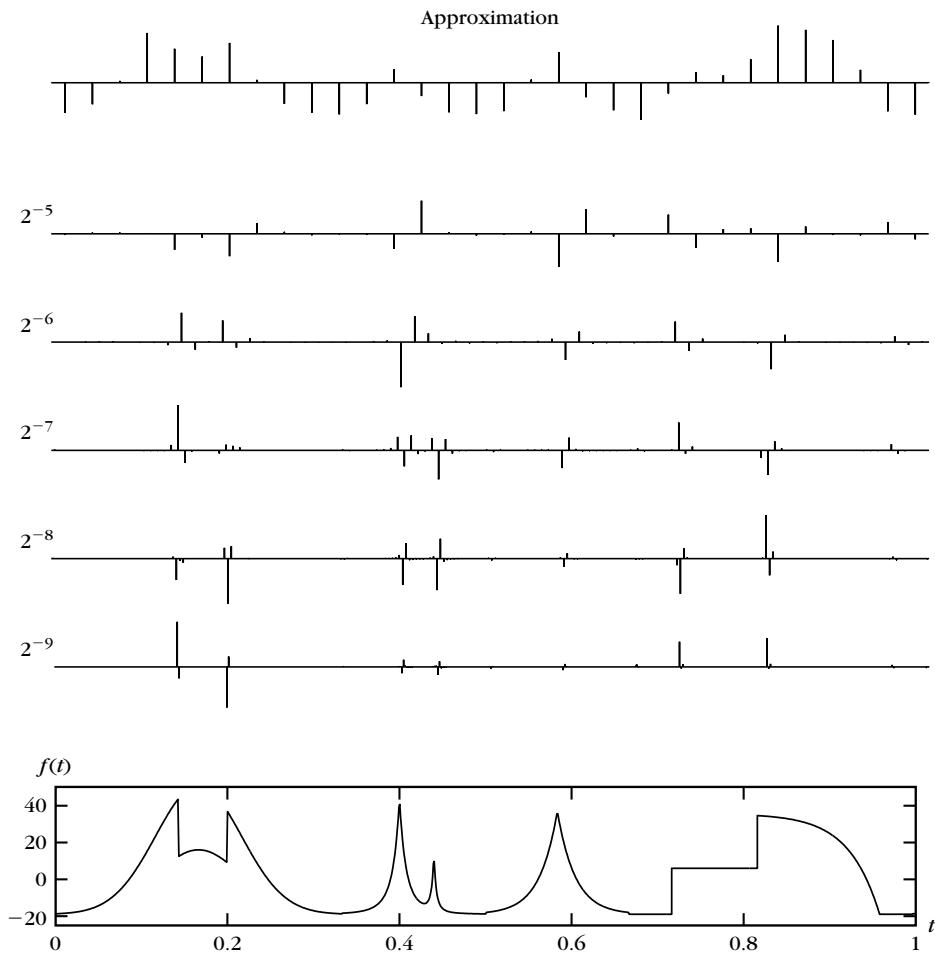
Thus, a signal expansion in a wavelet orthogonal basis can be viewed as an aggregation of details at all scales  $2^j$  that go from 0 to  $+\infty$ :

$$f = \sum_{j=-\infty}^{+\infty} P\mathbf{W}_j f = \sum_{j=-\infty}^{+\infty} \sum_{n=-\infty}^{+\infty} \langle f, \psi_{j,n} \rangle \psi_{j,n}.$$

Figure 7.7 gives the coefficients of a signal decomposed in the cubic spline wavelet orthogonal basis. The calculations are performed with the fast wavelet transform algorithm of Section 7.3. The up or down Diracs give the amplitudes of positive or negative wavelet coefficients at a distance  $2^j n$  at each scale  $2^j$ . Coefficients are nearly zero at fine scales where the signal is locally regular.

### Wavelet Design

Theorem 7.3 constructs a wavelet orthonormal basis from any conjugate mirror filter  $\hat{h}(\omega)$ . This gives a simple procedure for designing and building wavelet orthogonal

**FIGURE 7.7**

Wavelet coefficients  $d_j[n] = \langle f, \psi_{j,n} \rangle$  calculated at scales  $2^j$  with the cubic spline wavelet. Each up or down Dirac gives the amplitude of a positive or negative wavelet coefficient. At the top is the remaining coarse-signal approximation  $a_j[n] = \langle f, \phi_{j,n} \rangle$  for  $J = -5$ .

bases. Conversely, we may wonder whether all wavelet orthonormal bases are associated to a multiresolution approximation and a conjugate mirror filter. If we impose that  $\psi$  has a compact support, then Lemarié [52] proved that  $\psi$  necessarily corresponds to a multiresolution approximation. It is possible, however, to construct pathological wavelets that decay like  $|t|^{-1}$  at infinity, and that cannot be derived from any multiresolution approximation. Section 7.2 describes important classes of wavelet bases and explains how to design  $\hat{h}$  to specify the support, the number of vanishing moments, and the regularity of  $\psi$ .

## 7.2 CLASSES OF WAVELET BASES

### 7.2.1 Choosing a Wavelet

Most applications of wavelet bases exploit their ability to efficiently approximate particular classes of functions with few nonzero wavelet coefficients. This is true not only for data compression but also for noise removal and fast calculations. Therefore, the design of  $\psi$  must be optimized to produce a maximum number of wavelet coefficients  $|\langle f, \psi_{j,n} \rangle|$  that are close to zero. A function  $f$  has few nonnegligible wavelet coefficients if most of the fine-scale (high-resolution) wavelet coefficients are small. This depends mostly on the regularity of  $f$ , the number of vanishing moments of  $\psi$ , and the size of its support. To construct an appropriate wavelet from a conjugate mirror filter  $h[n]$ , we relate these properties to conditions on  $\hat{h}(\omega)$ .

#### *Vanishing Moments*

Let us recall that  $\psi$  has  $p$  vanishing moments if

$$\int_{-\infty}^{+\infty} t^k \psi(t) dt = 0 \quad \text{for } 0 \leq k < p. \quad (7.69)$$

This means that  $\psi$  is orthogonal to any polynomial of degree  $p - 1$ . Section 6.1.3 proves that if  $f$  is regular and  $\psi$  has enough vanishing moments, then the wavelet coefficients  $|\langle f, \psi_{j,n} \rangle|$  are small at fine scales  $2^j$ . Indeed, if  $f$  is locally  $C^k$ , then over a small interval it is well approximated by a Taylor polynomial of degree  $k$ . If  $k < p$ , then wavelets are orthogonal to this Taylor polynomial, and thus produce small-amplitude coefficients at fine scales. Theorem 7.4 relates the number of vanishing moments of  $\psi$  to the vanishing derivatives of  $\hat{\psi}(\omega)$  at  $\omega = 0$  and to the number of zeroes of  $\hat{h}(\omega)$  at  $\omega = \pi$ . It also proves that polynomials of degree  $p - 1$  are then reproduced by the scaling functions.

**Theorem 7.4:** *Vanishing Moments.* Let  $\psi$  and  $\phi$  be a wavelet and a scaling function that generate an orthogonal basis. Suppose that  $|\psi(t)| = O((1+t^2)^{-p/2-1})$  and  $|\phi(t)| = O((1+t^2)^{-p/2-1})$ . The four following statements are equivalent:

1. The wavelet  $\psi$  has  $p$  vanishing moments.
2.  $\hat{\psi}(\omega)$  and its first  $p - 1$  derivatives are zero at  $\omega = 0$ .
3.  $\hat{h}(\omega)$  and its first  $p - 1$  derivatives are zero at  $\omega = \pi$ .
4. For any  $0 \leq k < p$ ,

$$q_k(t) = \sum_{n=-\infty}^{+\infty} n^k \phi(t-n) \text{ is a polynomial of degree } k. \quad (7.70)$$

**Proof.** The decay of  $|\phi(t)|$  and  $|\psi(t)|$  implies that  $\hat{\psi}(\omega)$  and  $\hat{\phi}(\omega)$  are  $p$  times continuously differentiable. The  $k^{\text{th}}$ -order derivative  $\hat{\psi}^{(k)}(\omega)$  is the Fourier transform of  $(-it)^k \psi(t)$ . Thus,

$$\hat{\psi}^{(k)}(0) = \int_{-\infty}^{+\infty} (-it)^k \psi(t) dt.$$

We derive that (1) is equivalent to (2).

Theorem 7.3 proves that

$$\sqrt{2} \hat{\psi}(2\omega) = e^{-i\omega} \hat{h}^*(\omega + \pi) \hat{\phi}(\omega).$$

Since  $\hat{\phi}(0) \neq 0$ , by differentiating this expression we prove that (2) is equivalent to (3).

Let us now prove that (4) implies (1). Since  $\psi$  is orthogonal to  $\{\phi(t - n)\}_{n \in \mathbb{Z}}$ , it is also orthogonal to the polynomials  $q_k$  for  $0 \leq k < p$ . This family of polynomials is a basis of the space of polynomials of degree at most  $p - 1$ . Thus,  $\psi$  is orthogonal to any polynomial of degree  $p - 1$  and in particular to  $t^k$  for  $0 \leq k < p$ . This means that  $\psi$  has  $p$  vanishing moments.

To verify that (1) implies (4) we suppose that  $\psi$  has  $p$  vanishing moments, and for  $k < p$ , we evaluate  $q_k(t)$  defined in (7.70). This is done by computing its Fourier transform:

$$\hat{q}_k(\omega) = \hat{\phi}(\omega) \sum_{n=-\infty}^{+\infty} n^k \exp(-in\omega) = (i)^k \hat{\phi}(\omega) \frac{d^k}{d\omega^k} \sum_{n=-\infty}^{+\infty} \exp(-in\omega).$$

Let  $\delta^{(k)}$  be the distribution that is the  $k^{\text{th}}$ -order derivative of a Dirac, defined in Section A.7 in the Appendix. The Poisson formula (2.4) proves that

$$\hat{q}_k(\omega) = (i)^k \frac{1}{2\pi} \hat{\phi}(\omega) \sum_{l=-\infty}^{+\infty} \delta^{(k)}(\omega - 2l\pi). \quad (7.71)$$

With several integrations by parts, we verify the distribution equality

$$\hat{\phi}(\omega) \delta^{(k)}(\omega - 2l\pi) = \hat{\phi}(2l\pi) \delta^{(k)}(\omega - 2l\pi) + \sum_{m=0}^{k-1} a_{m,l}^k \delta^{(m)}(\omega - 2l\pi), \quad (7.72)$$

where  $a_{m,l}^k$  is a linear combination of the derivatives  $\{\hat{\phi}^{(m)}(2l\pi)\}_{0 \leq m \leq k}$ .

For  $l \neq 0$ , let us prove that  $a_{m,l}^k = 0$  by showing that  $\hat{\phi}^{(m)}(2l\pi) = 0$  if  $0 \leq m < p$ . For any  $P > 0$ , (7.27) implies

$$\hat{\phi}(\omega) = \hat{\phi}(2^{-P}\omega) \prod_{p=1}^P \frac{\hat{h}(2^{-p}\omega)}{\sqrt{2}}. \quad (7.73)$$

Since  $\psi$  has  $p$  vanishing moments, we showed in (3) that  $\hat{h}(\omega)$  has a zero of order  $p$  at  $\omega = \pm\pi$ . But  $\hat{h}(\omega)$  is also  $2\pi$  periodic, so (7.73) implies that  $\hat{\phi}(\omega) = O(|\omega - 2l\pi|^p)$  in the neighborhood of  $\omega = 2l\pi$ , for any  $l \neq 0$ . Thus,  $\hat{\phi}^{(m)}(2l\pi) = 0$  if  $m < p$ .

Since  $a_{m,l}^k = 0$  and  $\phi(2l\pi) = 0$  when  $l \neq 0$ , it follows from (7.72) that

$$\hat{\phi}(\omega) \delta^{(k)}(\omega - 2l\pi) = 0 \quad \text{for } l \neq 0.$$

The only term that remains in the summation (7.71) is  $l = 0$ , and inserting (7.72) yields

$$\hat{q}_k(\omega) = (i)^k \frac{1}{2\pi} \left( \hat{\phi}(0) \delta^{(k)}(\omega) + \sum_{m=0}^{k-1} a_{m,0}^k \delta^{(m)}(\omega) \right).$$

The inverse Fourier transform of  $\delta^{(m)}(\omega)$  is  $(2\pi)^{-1}(-it)^m$ , and Theorem 7.2 proves that  $\hat{\phi}(0) \neq 0$ . Thus, the inverse Fourier transform  $q_k$  of  $\hat{q}_k$  is a polynomial of degree  $k$ . ■

The hypothesis (4) is called the Fix-Strang condition [446]. The polynomials  $\{q_k\}_{0 \leq k < p}$  define a basis of the space of polynomials of degree  $p - 1$ . The Fix-Strang condition proves that  $\psi$  has  $p$  vanishing moments if and only if any polynomial of degree  $p - 1$  can be written as a linear expansion of  $\{\phi(t - n)\}_{n \in \mathbb{Z}}$ . The decomposition coefficients of the polynomials  $q_k$  do not have a finite energy because polynomials do not have a finite energy.

### Size of Support

If  $f$  has an isolated singularity at  $t_0$  and if  $t_0$  is inside the support of  $\psi_{j,n}(t) = 2^{-j/2} \psi(2^{-j}t - n)$ , then  $\langle f, \psi_{j,n} \rangle$  may have a large amplitude. If  $\psi$  has a compact support of size  $K$ , at each scale  $2^j$  there are  $K$  wavelets  $\psi_{j,n}$  with a support including  $t_0$ . To minimize the number of high-amplitude coefficients we must reduce the support size of  $\psi$ . Theorem 7.5 relates the support size of  $h$  to the support of  $\phi$  and  $\psi$ .

**Theorem 7.5: Compact Support.** The scaling function  $\phi$  has a compact support if and only if  $h$  has a compact support and their supports are equal. If the support of  $h$  and  $\phi$  is  $[N_1, N_2]$ , then the support of  $\psi$  is  $[(N_1 - N_2 + 1)/2, (N_2 - N_1 + 1)/2]$ .

**Proof.** If  $\phi$  has a compact support, since

$$h[n] = \frac{1}{\sqrt{2}} \left\langle \phi\left(\frac{t}{2}\right), \phi(t - n) \right\rangle,$$

we derive that  $h$  also has a compact support. Conversely, the scaling function satisfies

$$\frac{1}{\sqrt{2}} \phi\left(\frac{t}{2}\right) = \sum_{n=-\infty}^{+\infty} h[n] \phi(t - n). \quad (7.74)$$

If  $h$  has a compact support then one can prove [194] that  $\phi$  has a compact support. The proof is not reproduced here.

To relate the support of  $\phi$  and  $h$ , we suppose that  $h[n]$  is nonzero for  $N_1 \leq n \leq N_2$  and that  $\phi$  has a compact support  $[K_1, K_2]$ . The support of  $\phi(t/2)$  is  $[2K_1, 2K_2]$ . The sum at the right of (7.74) is a function with a support of  $[N_1 + K_1, N_2 + K_2]$ . The equality proves that the support of  $\phi$  is  $[K_1, K_2] = [N_1, N_2]$ .

Let us recall from (7.68) and (7.67) that

$$\frac{1}{\sqrt{2}} \psi\left(\frac{t}{2}\right) = \sum_{n=-\infty}^{+\infty} g[n] \phi(t - n) = \sum_{n=-\infty}^{+\infty} (-1)^{1-n} h[1-n] \phi(t - n).$$

If the supports of  $\phi$  and  $h$  are equal to  $[N_1, N_2]$ , the sum on the right side has a support equal to  $[N_1 - N_2 + 1, N_2 - N_1 + 1]$ . Thus,  $\psi$  has a support equal to  $[(N_1 - N_2 + 1)/2, (N_2 - N_1 + 1)/2]$ . ■

If  $h$  has a finite impulse response in  $[N_1, N_2]$ , Theorem 7.5 proves that  $\psi$  has a support of size  $N_2 - N_1$  centered at  $1/2$ . To minimize the size of the support, we must synthesize conjugate mirror filters with as few nonzero coefficients as possible.

### Support versus Moments

The support size of a function and the number of vanishing moments are a priori independent. However, we shall see in Theorem 7.7 that the constraints imposed on orthogonal wavelets imply that if  $\psi$  has  $p$  vanishing moments, then its support is at least of size  $2p - 1$ . Daubechies wavelets are optimal in the sense that they have a minimum size support for a given number of vanishing moments. When choosing a particular wavelet, we face a trade-off between the number of vanishing moments and the support size. If  $f$  has few isolated singularities and is very regular between singularities, we must choose a wavelet with many vanishing moments to produce a large number of small wavelet coefficients  $\langle f, \psi_{j,n} \rangle$ . If the density of singularities increases, it might be better to decrease the size of its support at the cost of reducing the number of vanishing moments. Indeed, wavelets that overlap the singularities create high-amplitude coefficients.

The multiwavelet construction of Geronimo, Hardin, and Massupust [271] offers more design flexibility by introducing several scaling functions and wavelets. Exercise 7.16 gives an example. Better trade-off can be obtained between the multiwavelets supports and their vanishing moments [447]. However, multiwavelet decompositions are implemented with a slightly more complicated filter bank algorithm than a standard orthogonal wavelet transform.

### Regularity

The regularity of  $\psi$  has mostly a cosmetic influence on the error introduced by thresholding or quantizing the wavelet coefficients. When reconstructing a signal from its wavelet coefficients

$$f = \sum_{j=-\infty}^{+\infty} \sum_{n=-\infty}^{+\infty} \langle f, \psi_{j,n} \rangle \psi_{j,n},$$

an error  $\varepsilon$  added to a coefficient  $\langle f, \psi_{j,n} \rangle$  will add the wavelet component  $\varepsilon \psi_{j,n}$  to the reconstructed signal. If  $\psi$  is smooth, then  $\varepsilon \psi_{j,n}$  is a smooth error. For image-coding applications, a smooth error is often less visible than an irregular error, even though they have the same energy. Better-quality images are obtained with wavelets that are continuously differentiable than with the discontinuous Haar wavelet. Theorem 7.6 due to Tchamitchian [454] relates the uniform Lipschitz regularity of  $\phi$  and  $\psi$  to the number of zeroes of  $\hat{h}(\omega)$  at  $\omega = \pi$ .

**Theorem 7.6:** *Tchamitchian.* Let  $\hat{h}(\omega)$  be a conjugate mirror filter with  $p$  zeroes at  $\pi$  and that satisfies the sufficient conditions of Theorem 7.2. Let us perform the factorization

$$\hat{h}(\omega) = \sqrt{2} \left( \frac{1 + e^{i\omega}}{2} \right)^p \hat{l}(\omega).$$

If  $\sup_{\omega \in \mathbb{R}} |\hat{l}(\omega)| = B$ , then  $\psi$  and  $\phi$  are uniformly Lipschitz  $\alpha$  for

$$\alpha < \alpha_0 = p - \log_2 B - 1. \quad (7.75)$$

**Proof.** This result is proved by showing that there exist  $C_1 > 0$  and  $C_2 > 0$  such that for all  $\omega \in \mathbb{R}$

$$|\hat{\phi}(\omega)| \leq C_1 (1 + |\omega|)^{-p + \log_2 B} \quad (7.76)$$

$$|\hat{\psi}(\omega)| \leq C_2 (1 + |\omega|)^{-p + \log_2 B}. \quad (7.77)$$

The Lipschitz regularity of  $\phi$  and  $\psi$  is then derived from Theorem 6.1, which shows that if  $\int_{-\infty}^{+\infty} (1 + |\omega|^\alpha) |\hat{f}(\omega)| d\omega < +\infty$ , then  $f$  is uniformly Lipschitz  $\alpha$ .

We proved in (7.32) that  $\hat{\phi}(\omega) = \prod_{j=1}^{+\infty} 2^{-1/2} \hat{h}(2^{-j}\omega)$ . One can verify that

$$\prod_{j=1}^{+\infty} \frac{1 + \exp(i2^{-j}\omega)}{2} = \frac{1 - \exp(i\omega)}{i\omega},$$

thus,

$$|\hat{\phi}(\omega)| = \frac{|1 - \exp(i\omega)|^p}{|\omega|^p} \prod_{j=1}^{+\infty} |\hat{h}(2^{-j}\omega)|. \quad (7.78)$$

Let us now compute an upper bound for  $\prod_{j=1}^{+\infty} |\hat{h}(2^{-j}\omega)|$ . At  $\omega = 0$ , we have  $\hat{h}(0) = \sqrt{2}$  so  $\hat{h}(0) = 1$ . Since  $\hat{h}(\omega)$  is continuously differentiable at  $\omega = 0$ ,  $\hat{h}'(\omega)$  is also continuously differentiable at  $\omega = 0$ . Thus, we derive that there exists  $\varepsilon > 0$  such that if  $|\omega| < \varepsilon$ , then  $|\hat{h}(\omega)| \leq 1 + K|\omega|$ . Consequently,

$$\sup_{|\omega| \leq \varepsilon} \prod_{j=1}^{+\infty} |\hat{h}(2^{-j}\omega)| \leq \sup_{|\omega| \leq \varepsilon} \prod_{j=1}^{+\infty} (1 + K|2^{-j}\omega|) \leq e^{K\varepsilon}. \quad (7.79)$$

If  $|\omega| > \varepsilon$ , there exists  $J \geq 1$  such that  $2^{J-1}\varepsilon \leq |\omega| \leq 2^J\varepsilon$ , and we decompose

$$\prod_{j=1}^{+\infty} \hat{h}(2^{-j}\omega) = \prod_{j=1}^J \hat{h}(2^{-j}\omega) \prod_{j=1}^{+\infty} |\hat{h}(2^{-j-J}\omega)|. \quad (7.80)$$

Since  $\sup_{\omega \in \mathbb{R}} |\hat{h}(\omega)| = B$ , inserting (7.79) yields for  $|\omega| > \varepsilon$

$$\prod_{j=1}^{+\infty} \hat{h}(2^{-j}\omega) \leq B^J e^{K\varepsilon} = e^{K\varepsilon} 2^{J \log_2 B}. \quad (7.81)$$

Since  $2^J \leq \varepsilon^{-1} 2|\omega|$ , this proves that

$$\forall \omega \in \mathbb{R}, \quad \prod_{j=1}^{+\infty} \hat{h}(2^{-j}\omega) \leq e^{K\varepsilon} \left( 1 + \frac{|2\omega|^{\log_2 B}}{\varepsilon^{\log_2 B}} \right).$$

Equation (7.76) is derived from (7.78) and this last inequality. Since  $|\hat{\psi}(2\omega)| = 2^{-1/2} |\hat{h}(\omega + \pi)| |\hat{\phi}(\omega)|$ , (7.77) is obtained from (7.76). ■

This theorem proves that if  $B < 2^{p-1}$ , then  $\alpha_0 > 0$ . It means that  $\phi$  and  $\psi$  are uniformly continuous. For any  $m > 0$ , if  $B < 2^{p-1-m}$ , then  $\alpha_0 > m$ , so  $\psi$  and  $\phi$  are  $m$  times continuously differentiable. Theorem 7.4 shows that the number  $p$  of zeros of  $\hat{h}(\omega)$  at  $\pi$  is equal to the number of vanishing moments of  $\psi$ . A priori, we are

not guaranteed that increasing  $p$  will improve the wavelet regularity since  $B$  might increase as well. However, for important families of conjugate mirror filters such as splines or Daubechies filters,  $B$  increases more slowly than  $p$ , which implies that wavelet regularity increases with the number of vanishing moments. Let us emphasize that the number of vanishing moments and the regularity of orthogonal wavelets are related but it is the number of vanishing moments and not the regularity that affects the amplitude of the wavelet coefficients at fine scales.

### 7.2.2 Shannon, Meyer, Haar, and Battle-Lemarié Wavelets

We study important classes of wavelets with Fourier transforms that are derived from the general formula proved in Theorem 7.3,

$$\hat{\psi}(\omega) = \frac{1}{\sqrt{2}} \hat{g}\left(\frac{\omega}{2}\right) \hat{\phi}\left(\frac{\omega}{2}\right) = \frac{1}{\sqrt{2}} \exp\left(\frac{-i\omega}{2}\right) \hat{h}^*\left(\frac{\omega}{2} + \pi\right) \hat{\phi}\left(\frac{\omega}{2}\right). \quad (7.82)$$

#### Shannon Wavelet

The Shannon wavelet is constructed from the Shannon multiresolution approximation, which approximates functions by their restriction to low-frequency intervals. It corresponds to  $\hat{\phi} = \mathbf{1}_{[-\pi, \pi]}$  and  $\hat{h}(\omega) = \sqrt{2} \mathbf{1}_{[-\pi/2, \pi/2]}(\omega)$  for  $\omega \in [-\pi, \pi]$ . We derive from (7.82) that

$$\hat{\psi}(\omega) = \begin{cases} \exp(-i\omega/2) & \text{if } \omega \in [-2\pi, -\pi] \cup [\pi, 2\pi] \\ 0 & \text{otherwise,} \end{cases} \quad (7.83)$$

and thus,

$$\psi(t) = \frac{\sin 2\pi(t - 1/2)}{2\pi(t - 1/2)} - \frac{\sin \pi(t - 1/2)}{\pi(t - 1/2)}.$$

This wavelet is  $\mathbf{C}^\infty$  but has a slow asymptotic time decay. Since  $\hat{\psi}(\omega)$  is zero in the neighborhood of  $\omega = 0$ , all its derivatives are zero at  $\omega = 0$ . Thus, Theorem 7.4 implies that  $\psi$  has an infinite number of vanishing moments.

Since  $\hat{\psi}(\omega)$  has a compact support we know that  $\psi(t)$  is  $\mathbf{C}^\infty$ . However,  $|\psi(t)|$  decays only like  $|t|^{-1}$  at infinity because  $\hat{\psi}(\omega)$  is discontinuous at  $\pm\pi$  and  $\pm 2\pi$ .

#### Meyer Wavelets

A Meyer wavelet [375] is a frequency band-limited function that has a Fourier transform that is smooth, unlike the Fourier transform of the Shannon wavelet. This smoothness provides a much faster asymptotic decay in time. These wavelets are constructed with conjugate mirror filters  $\hat{h}(\omega)$  that are  $\mathbf{C}^n$  and satisfy

$$\hat{h}(\omega) = \begin{cases} \sqrt{2} & \text{if } \omega \in [-\pi/3, \pi/3] \\ 0 & \text{if } \omega \in [-\pi, -2\pi/3] \cup [2\pi/3, \pi]. \end{cases} \quad (7.84)$$

The only degree of freedom is the behavior of  $\hat{h}(\omega)$  in the transition bands  $[-2\pi/3, -\pi/3] \cup [\pi/3, 2\pi/3]$ . It must satisfy the quadrature condition

$$|\hat{h}(\omega)|^2 + |\hat{h}(\omega + \pi)|^2 = 2, \quad (7.85)$$

and to obtain  $\mathbf{C}^n$  junctions at  $|\omega| = \pi/3$  and  $|\omega| = 2\pi/3$ , the  $n$  first derivatives must vanish at these abscissa. One can construct such functions that are  $\mathbf{C}^\infty$ .

The scaling function  $\hat{\phi}(\omega) = \prod_{p=1}^{+\infty} 2^{-1/2} \hat{h}(2^{-p}\omega)$  has a compact support and one can verify that

$$\hat{\phi}(\omega) = \begin{cases} 2^{-1/2} \hat{h}(\omega/2) & \text{if } |\omega| \leq 4\pi/3 \\ 0 & \text{if } |\omega| > 4\pi/3. \end{cases} \quad (7.86)$$

The resulting wavelet (7.82) is

$$\hat{\psi}(\omega) = \begin{cases} 0 & \text{if } |\omega| \leq 2\pi/3 \\ 2^{-1/2} \hat{g}(\omega/2) & \text{if } 2\pi/3 \leq |\omega| \leq 4\pi/3 \\ 2^{-1/2} \exp(-i\omega/2) \hat{h}(\omega/4) & \text{if } 4\pi/3 \leq |\omega| \leq 8\pi/3 \\ 0 & \text{if } |\omega| > 8\pi/3. \end{cases} \quad (7.87)$$

The functions  $\phi$  and  $\psi$  are  $\mathbf{C}^\infty$  because their Fourier transforms have a compact support. Since  $\hat{\psi}(\omega) = 0$  in the neighborhood of  $\omega = 0$ , all its derivatives are zero at  $\omega = 0$ , which proves that  $\psi$  has an infinite number of vanishing moments.

If  $\hat{h}$  is  $\mathbf{C}^n$ , then  $\hat{\psi}$  and  $\hat{\phi}$  are also  $\mathbf{C}^n$ . The discontinuities of the  $(n+1)^{\text{th}}$  derivative of  $\hat{h}$  are generally at the junction of the transition band  $|\omega| = \pi/3, 2\pi/3$ , in which case one can show that there exists  $A$  such that

$$|\phi(t)| \leq A (1 + |t|)^{-n-1} \quad \text{and} \quad |\psi(t)| \leq A (1 + |t|)^{-n-1}.$$

Although the asymptotic decay of  $\psi$  is fast when  $n$  is large, its effective numerical decay may be relatively slow, which is reflected by the fact that  $A$  is quite large. As a consequence, a Meyer wavelet transform is generally implemented in the Fourier domain. Section 8.4.2 relates these wavelet bases to lapped orthogonal transforms applied in the Fourier domain. One can prove [19] that there exists no orthogonal wavelet that is  $\mathbf{C}^\infty$  and has an exponential decay.

### EXAMPLE 7.10

To satisfy the quadrature condition (7.85), one can verify that  $\hat{h}$  in (7.84) may be defined on the transition bands by

$$\hat{h}(\omega) = \sqrt{2} \cos \left[ \frac{\pi}{2} \beta \left( \frac{3|\omega|}{\pi} - 1 \right) \right] \quad \text{for } |\omega| \in [\pi/3, 2\pi/3],$$

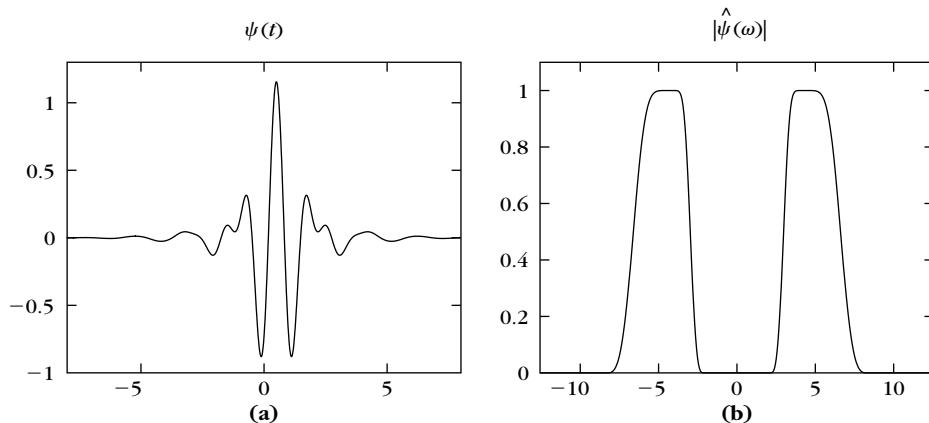
where  $\beta(x)$  is a function that goes from 0 to 1 on the interval  $[0, 1]$  and satisfies

$$\forall x \in [0, 1], \quad \beta(x) + \beta(1-x) = 1. \quad (7.88)$$

An example due to Daubechies [19] is

$$\beta(x) = x^4 (35 - 84x + 70x^2 - 20x^3). \quad (7.89)$$

The resulting  $\hat{h}(\omega)$  has  $n = 3$  vanishing derivatives at  $|\omega| = \pi/3, 2\pi/3$ . Figure 7.8 displays the corresponding wavelet  $\psi$ .

**FIGURE 7.8**

Meyer wavelet  $\psi$  (a) and its Fourier transform modulus computed with (7.89) (b).

### Haar Wavelets

The Haar basis is obtained with a multiresolution of piecewise constant functions. The scaling function is  $\phi = \mathbf{1}_{[0,1]}$ . The filter  $h[n]$  given in (7.46) has two nonzero coefficients equal to  $2^{-1/2}$  at  $n = 0$  and  $n = 1$ . Thus,

$$\frac{1}{\sqrt{2}} \psi\left(\frac{t}{2}\right) = \sum_{n=-\infty}^{+\infty} (-1)^{1-n} h[1-n] \phi(t-n) = \frac{1}{\sqrt{2}} (\phi(t-1) - \phi(t)),$$

so

$$\psi(t) = \begin{cases} -1 & \text{if } 0 \leq t < 1/2 \\ 1 & \text{if } 1/2 \leq t < 1 \\ 0 & \text{otherwise.} \end{cases} \quad (7.90)$$

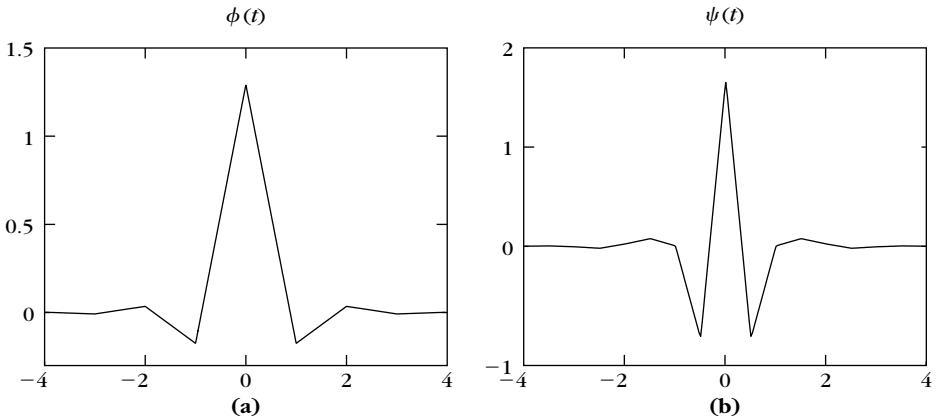
The Haar wavelet has the shortest support among all orthogonal wavelets. It is not well adapted to approximating smooth functions because it has only one vanishing moment.

### Battle-Lemarié Wavelets

Polynomial spline wavelets introduced by Battle [99] and Lemarié [345] are computed from spline multiresolution approximations. The expressions of  $\hat{\phi}(\omega)$  and  $\hat{h}(\omega)$  are given, respectively, by (7.18) and (7.48). For splines of degree  $m$ ,  $\hat{h}(\omega)$  and its first  $m$  derivatives are zero at  $\omega = \pi$ . Theorem 7.4 derives that  $\psi$  has  $m + 1$  vanishing moments. It follows from (7.82) that

$$\hat{\psi}(\omega) = \frac{\exp(-i\omega/2)}{\omega^{m+1}} \sqrt{\frac{S_{2m+2}(\omega/2 + \pi)}{S_{2m+2}(\omega) S_{2m+2}(\omega/2)}}.$$

This wavelet  $\psi$  has an exponential decay. Since it is a polynomial spline of degree  $m$ , it is  $m - 1$  times continuously differentiable. Polynomial spline wavelets are less

**FIGURE 7.9**

Linear spline Battle-Lemarié scaling function  $\phi$  (a) and wavelet  $\psi$  (b).

regular than Meyer wavelets but have faster time asymptotic decay. For  $m$  odd,  $\psi$  is symmetric about  $1/2$ . For  $m$  even, it is antisymmetric about  $1/2$ . Figure 7.5 gives the graph of the cubic spline wavelet  $\psi$  corresponding to  $m = 3$ . For  $m = 1$ , Figure 7.9 displays linear splines  $\phi$  and  $\psi$ . The properties of these wavelets are further studied in [15, 106, 164].

### 7.2.3 Daubechies Compactly Supported Wavelets

Daubechies wavelets have a support of minimum size for any given number  $p$  of vanishing moments. Theorem 7.5 proves that wavelets of compact support are computed with finite impulse-response conjugate mirror filters  $h$ . We consider real causal filters  $h[n]$ , which implies that  $\hat{h}$  is a trigonometric polynomial:

$$\hat{h}(\omega) = \sum_{n=0}^{N-1} h[n] e^{-in\omega}.$$

To ensure that  $\psi$  has  $p$  vanishing moments, Theorem 7.4 shows that  $\hat{h}$  must have a zero of order  $p$  at  $\omega = \pi$ . To construct a trigonometric polynomial of minimal size, we factor  $(1 + e^{-i\omega})^p$ , which is a minimum-size polynomial having  $p$  zeros at  $\omega = \pi$ :

$$\hat{h}(\omega) = \sqrt{2} \left( \frac{1 + e^{-i\omega}}{2} \right)^p R(e^{-i\omega}). \quad (7.91)$$

The difficulty is to design a polynomial  $R(e^{-i\omega})$  of minimum degree  $m$  such that  $\hat{h}$  satisfies

$$|\hat{h}(\omega)|^2 + |\hat{h}(\omega + \pi)|^2 = 2. \quad (7.92)$$

As a result,  $h$  has  $N = m + p + 1$  nonzero coefficients. Theorem 7.7 by Daubechies [194] proves that the minimum degree of  $R$  is  $m = p - 1$ .

**Theorem 7.7:** Daubechies. A real conjugate mirror filter  $h$ , such that  $\hat{h}(\omega)$  has  $p$  zeroes at  $\omega = \pi$ , has at least  $2p$  nonzero coefficients. Daubechies filters have  $2p$  nonzero coefficients.

**Proof.** The proof is constructive and computes Daubechies filters. Since  $h[n]$  is real,  $|\hat{h}(\omega)|^2$  is an even function and can be written as a polynomial in  $\cos \omega$ . Thus,  $|R(e^{-i\omega})|^2$  defined in (7.91) is a polynomial in  $\cos \omega$  that we can also write as a polynomial  $P(\sin^2(\omega/2))$ ,

$$|\hat{h}(\omega)|^2 = 2 \left( \cos \frac{\omega}{2} \right)^{2p} P \left( \sin^2 \frac{\omega}{2} \right). \quad (7.93)$$

The quadrature condition (7.92) is equivalent to

$$(1-y)^p P(y) + y^p P(1-y) = 1, \quad (7.94)$$

for any  $y = \sin^2(\omega/2) \in [0, 1]$ . To minimize the number of nonzero terms of the finite Fourier series  $\hat{h}(\omega)$ , we must find the solution  $P(y) \geq 0$  of minimum degree, which is obtained with the Bezout theorem on polynomials. ■

**Theorem 7.8:** Bezout. Let  $Q_1(y)$  and  $Q_2(y)$  be two polynomials of degrees  $n_1$  and  $n_2$  with no common zeroes. There exist two unique polynomials  $P_1(y)$  and  $P_2(y)$  of degrees  $n_2 - 1$  and  $n_1 - 1$  such that

$$P_1(y) Q_1(y) + P_2(y) Q_2(y) = 1. \quad (7.95)$$

The proof of this classical result is in [19]. Since  $Q_1(y) = (1-y)^p$  and  $Q_2(y) = y^p$  are two polynomials of degree  $p$  with no common zeros, the Bezout theorem proves that there exist two unique polynomials  $P_1(y)$  and  $P_2(y)$  such that

$$(1-y)^p P_1(y) + y^p P_2(y) = 1.$$

The reader can verify that  $P_2(y) = P_1(1-y) = P(1-y)$  with

$$P(y) = \sum_{k=0}^{p-1} \binom{p-1+k}{k} y^k. \quad (7.96)$$

Clearly,  $P(y) \geq 0$  for  $y \in [0, 1]$ . Thus,  $P(y)$  is the polynomial of minimum degree satisfying (7.94) with  $P(y) \geq 0$ .

Now we need to construct a minimum-degree polynomial

$$R(e^{-i\omega}) = \sum_{k=0}^m r_k e^{-ik\omega} = r_0 \prod_{k=0}^m (1 - a_k e^{-i\omega})$$

such that  $|R(e^{-i\omega})|^2 = P(\sin^2(\omega/2))$ . Since its coefficients are real,  $R^*(e^{-i\omega}) = R(e^{i\omega})$ , and thus,

$$|R(e^{-i\omega})|^2 = R(e^{-i\omega}) R(e^{i\omega}) = P \left( \frac{2 - e^{i\omega} - e^{-i\omega}}{4} \right) = Q(e^{-i\omega}). \quad (7.97)$$

This factorization is solved by extending it to the whole complex plane with the variable  $z = e^{-i\omega}$ :

$$R(z) R(z^{-1}) = r_0^2 \prod_{k=0}^m (1 - a_k z) (1 - a_k z^{-1}) = Q(z) = P \left( \frac{2 - z - z^{-1}}{4} \right). \quad (7.98)$$

Let us compute the roots of  $Q(z)$ . Since  $Q(z)$  has real coefficients if  $c_k$  is a root, then  $c_k^*$  is also a root, and since it is a function of  $z+z^{-1}$  if  $c_k$  is a root, then  $1/c_k$  and thus  $1/c_k^*$  are also roots. To design  $R(z)$  that satisfies (7.98), we choose each root  $a_k$  of  $R(z)$  among a pair  $(c_k, 1/c_k)$  and include  $a_k^*$  as a root to obtain real coefficients. This procedure yields a polynomial of minimum degree  $m=p-1$ , with  $r_0^2=Q(0)=P(1/2)=2^{p-1}$ . The resulting filter  $h$  of minimum size has  $N=p+m+1=2p$  nonzero coefficients.

Among all possible factorizations, the minimum-phase solution  $R(e^{i\omega})$  is obtained by choosing  $a_k$  among  $(c_k, 1/c_k)$  to be inside the unit circle  $|a_k| \leq 1$  [51]. The resulting causal filter  $h$  has an energy maximally concentrated at small abscissa  $n \geq 0$ . It is a Daubechies filter of order  $p$ .

The constructive proof of this theorem synthesizes causal conjugate mirror filters of size  $2p$ . Table 7.2 gives the coefficients of these Daubechies filters for  $2 \leq p \leq 10$ . Theorem 7.9 derives that Daubechies wavelets calculated with these conjugate mirror filters have a support of minimum size.

**Theorem 7.9:** *Daubechies.* If  $\psi$  is a wavelet with  $p$  vanishing moments that generates an orthonormal basis of  $L^2(\mathbb{R})$ , then it has a support of size larger than or equal to  $2p-1$ . A Daubechies wavelet has a minimum-size support equal to  $[-p+1, p]$ . The support of the corresponding scaling function  $\phi$  is  $[0, 2p-1]$ .

This theorem is a direct consequence of Theorem 7.7. The support of the wavelet, and that of the scaling function, are calculated with Theorem 7.5. When  $p=1$  we get the Haar wavelet. Figure 7.10 displays the graphs of  $\phi$  and  $\psi$  for  $p=2, 3, 4$ .

The regularity of  $\phi$  and  $\psi$  is the same since  $\psi(t)$  is a finite linear combination of  $\phi(2t-n)$ . However, this regularity is difficult to estimate precisely. Let  $B = \sup_{\omega \in \mathbb{R}} |R(e^{-i\omega})|$  where  $R(e^{-i\omega})$  is the trigonometric polynomial defined in (7.91). Theorem 7.6 proves that  $\psi$  is at least uniformly Lipschitz  $\alpha$  for  $\alpha < p - \log_2 B - 1$ . For Daubechies wavelets,  $B$  increases more slowly than  $p$ , and Figure 7.10 shows indeed that the regularity of these wavelets increases with  $p$ . Daubechies and Lagarias [198] have established a more precise technique that computes the exact Lipschitz regularity of  $\psi$ . For  $p=2$  the wavelet  $\psi$  is only Lipschitz 0.55, but for  $p=3$  it is Lipschitz 1.08, which means that it is already continuously differentiable. For  $p$  large,  $\phi$  and  $\psi$  are uniformly Lipschitz  $\alpha$ , for  $\alpha$  of the order of  $0.2p$  [168].

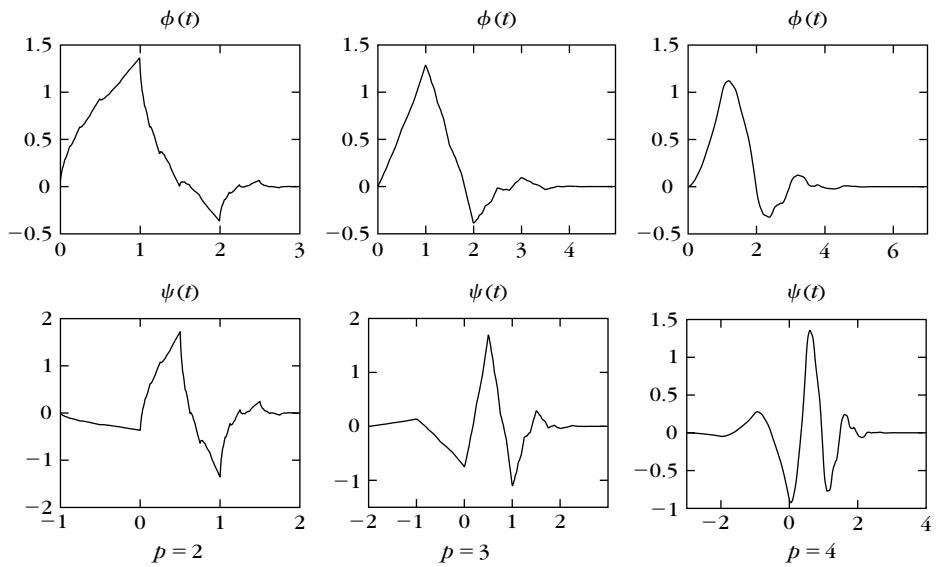
### Symmlets

Daubechies wavelets are very asymmetric because they are constructed by selecting the minimum-phase square root of  $Q(e^{-i\omega})$  in (7.97). One can show [51] that filters corresponding to a minimum-phase square root have their energy optimally concentrated near the starting point of their support. Thus, they are highly nonsymmetric, which yields very asymmetric wavelets.

To obtain a symmetric or antisymmetric wavelet, the filter  $h$  must be symmetric or antisymmetric with respect to the center of its support, which means that  $\hat{h}(\omega)$  has a linear complex phase. Daubechies proved [194] that the Haar filter is the

**Table 7.2** Daubechies Filters for Wavelets with  $p$  Vanishing Moments

	$n$	$h_p[n]$		$n$	$h_p[n]$		$n$	$h_p[n]$
$p = 2$	0	0.482962913145		8	-0.031582039317		2	0.604823123690
	1	0.836516303738		9	0.000553842201		3	0.657288078051
	2	0.224143868042		10	0.004777257511		4	0.133197385825
	3	-0.129409522551		11	-0.001077301085		5	-0.293273783279
$p = 3$	0	0.332670552950	$p = 7$	0	0.077852054085		6	-0.096840783223
	1	0.806891509311		1	0.396539319482		7	0.148540749338
	2	0.459877502118		2	0.729132090846		8	0.030725681479
	3	-0.135011020010		3	0.469782287405		9	-0.067632829061
	4	-0.085441273882		4	-0.143906003929		10	0.000250947115
	5	0.035226291882		5	-0.224036184994		11	0.022361662124
$p = 4$	0	0.230377813309		6	0.071309219267		12	-0.004723204758
	1	0.714846570553		7	0.080612609151		13	-0.004281503682
	2	0.630880767930		8	-0.038029936935		14	0.001847646883
	3	-0.027983769417		9	-0.016574541631		15	0.000230385764
	4	-0.187034811719		10	0.012550998556		16	-0.000251963189
	5	0.030841381836		11	0.000429577973		17	0.000039347320
	6	0.032883011667		12	-0.001801640704			
	7	-0.010597401785		13	0.000353713800			
$p = 5$	0	0.160102397974	$p = 8$	0	0.054415842243		2	0.527201188932
	1	0.603829269797		1	0.312871590914		3	0.688459039454
	2	0.724308528438		2	0.675630736297		4	0.281172343661
	3	0.138428145901		3	0.585354683654		5	-0.249846424327
	4	-0.242294887066		4	-0.015829105256		6	-0.195946274377
	5	-0.032244869585		5	-0.284015542962		7	0.127369340336
	6	0.077571493840		6	0.000472484574		8	0.093057364604
	7	-0.006241490213		7	0.128747426620		9	-0.071394147166
	8	-0.012580751999		8	-0.017369301002		10	-0.029457536822
	9	0.003335725285		9	-0.04408825393		11	0.033212674059
$p = 6$	0	0.111540743350		10	0.013981027917		12	0.003606553567
	1	0.494623890398		11	0.008746094047		13	-0.010733175483
	2	0.751133908021		12	-0.004870352993		14	0.001395351747
	3	0.315250351709		13	-0.000391740373		15	0.001992405295
	4	-0.226264693965		14	0.000675449406		16	-0.000685856695
	5	-0.129766867567		15	-0.000117476784		17	-0.000116466855
	6	0.097501605587	$p = 9$	0	0.038077947364		18	0.000093588670
	7	0.027522865530		1	0.243834674613		19	-0.000013264203

**FIGURE 7.10**

Daubechies scaling function  $\phi$  and wavelet  $\psi$  with  $p$  vanishing moments.

only real compactly supported conjugate mirror filter that has a linear phase. The Daubechies *symmlet* filters are obtained by optimizing the choice of the square root  $R(e^{-i\omega})$  of  $Q(e^{-i\omega})$  to obtain an almost linear phase. The resulting wavelets still have a minimum support  $[-p+1, p]$  with  $p$  vanishing moments, but they are more symmetric, as illustrated by Figure 7.11 for  $p = 8$ . The coefficients of the symmlet filters are in WAVELET. Complex conjugate mirror filters with a compact support and a linear phase can be constructed [352], but they produce complex wavelet coefficients that have real and imaginary parts that are redundant when the signal is real.

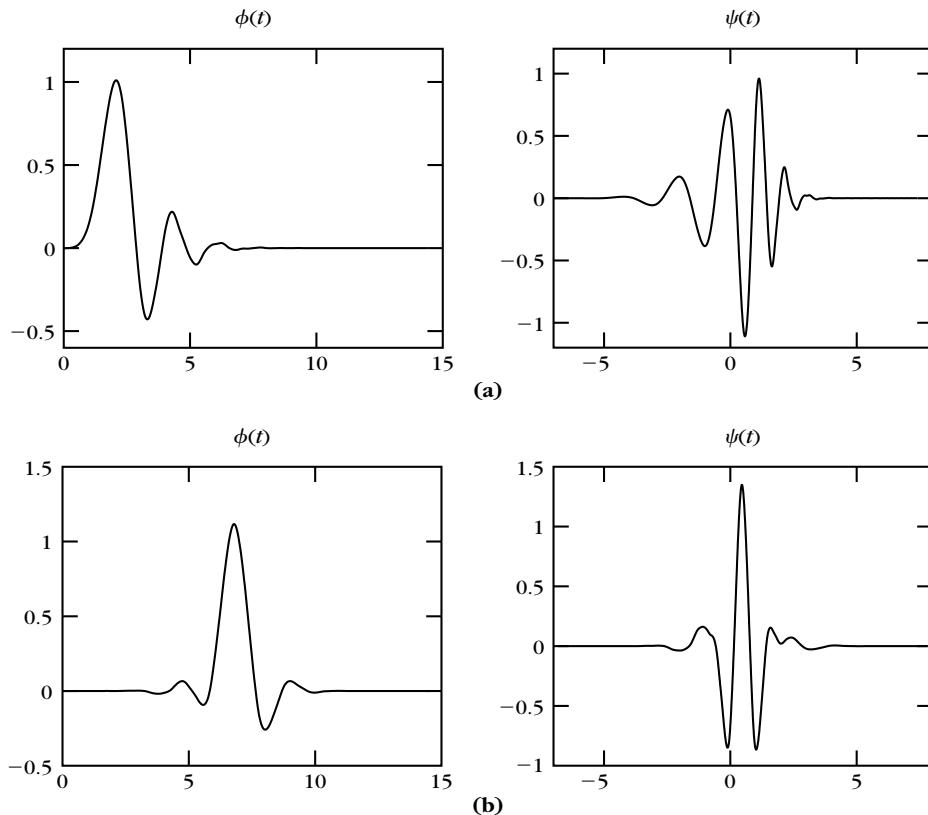
### Coiflets

For an application in numerical analysis, Coifman asked Daubechies [194] to construct a family of wavelets  $\psi$  that have  $p$  vanishing moments and a minimum-size support, with scaling functions that also satisfy

$$\int_{-\infty}^{+\infty} \phi(t) dt = 1 \quad \text{and} \quad \int_{-\infty}^{+\infty} t^k \phi(t) dt = 0 \quad \text{for } 1 \leq k < p. \quad (7.99)$$

Such scaling functions are useful in establishing precise quadrature formulas. If  $f$  is  $C^k$  in the neighborhood of  $2^J n$  with  $k < p$ , then a Taylor expansion of  $f$  up to order  $k$  shows that

$$2^{-J/2} \langle f, \phi_{J,n} \rangle \approx f(2^J n) + O(2^{(k+1)J}). \quad (7.100)$$

**FIGURE 7.11**

Daubechies (a) and symmlet (b) scaling functions and wavelets with  $p = 8$  vanishing moments.

Thus, at a fine scale  $2^J$ , the scaling coefficients are closely approximated by the signal samples. The order of approximation increases with  $p$ . The supplementary condition (7.99) requires increasing the support of  $\psi$ ; the resulting coiflet has a support of size  $3p - 1$  instead of  $2p - 1$  for a Daubechies wavelet. The corresponding conjugate mirror filters are tabulated in WAVELET.

### **Audio Filters**

The first conjugate mirror filters with finite impulse response were constructed in 1986 by Smith and Barnwell [443] in the context of perfect filter bank reconstruction, explained in Section 7.3.2. These filters satisfy the quadrature condition  $|\hat{h}(\omega)|^2 + |\hat{h}(\omega + \pi)|^2 = 2$ , which is necessary and sufficient for filter bank reconstruction. However,  $\hat{h}(0) \neq \sqrt{2}$ , so the infinite product of such filters does not yield a wavelet basis of  $L^2(\mathbb{R})$ . Instead of imposing any vanishing moments, Smith and Barnwell [443], and later Vaidyanathan and Hoang [470], designed their filters to

reduce the size of the transition band, where  $|\hat{h}(\omega)|$  decays from nearly  $\sqrt{2}$  to nearly 0 in the neighborhood of  $\pm \pi/2$ . This constraint is important in optimizing the transform code of audio signals (see Section 10.3.3). However, many cascades of these filters exhibit wild behavior. The Vaidyanathan-Hoang filters are tabulated in WAVELET. Many other classes of conjugate mirror filters with finite impulse response have been constructed [69, 79]. Recursive conjugate mirror filters may also be designed [300] to minimize the size of the transition band for a given number of zeroes at  $\omega = \pi$ . These filters have a fast but noncausal recursive implementation for signals of finite size.

## 7.3 WAVELETS AND FILTER BANKS

Decomposition coefficients in a wavelet orthogonal basis are computed with a fast algorithm that cascades discrete convolutions with  $h$  and  $g$ , and subsamples the output. Section 7.3.1 derives this result from the embedded structure of multiresolution approximations. A direct filter bank analysis is performed in Section 7.3.2, which gives more general perfect reconstruction conditions on the filters. Section 7.3.3 shows that perfect reconstruction filter banks decompose signals in a basis of  $\ell^2(\mathbb{Z})$ . This basis is orthogonal for conjugate mirror filters.

### 7.3.1 Fast Orthogonal Wavelet Transform

We describe a fast filter bank algorithm that computes the orthogonal wavelet coefficients of a signal measured at a finite resolution. A fast wavelet transform decomposes successively each approximation  $P_{\mathbf{V}_j}f$  into a coarser approximation  $P_{\mathbf{V}_{j+1}}f$ , plus the wavelet coefficients carried by  $P_{\mathbf{W}_{j+1}}f$ . In the other direction, the reconstruction from wavelet coefficients recovers each  $P_{\mathbf{V}_j}f$  from  $P_{\mathbf{V}_{j+1}}f$  and  $P_{\mathbf{W}_{j+1}}f$ .

Since  $\{\phi_{j,n}\}_{n \in \mathbb{Z}}$  and  $\{\psi_{j,n}\}_{n \in \mathbb{Z}}$  are orthonormal bases of  $\mathbf{V}_j$  and  $\mathbf{W}_j$ , the projection in these spaces is characterized by

$$a_j[n] = \langle f, \phi_{j,n} \rangle \quad \text{and} \quad d_j[n] = \langle f, \psi_{j,n} \rangle.$$

Theorem 7.10 [360, 361] shows that these coefficients are calculated with a cascade of discrete convolutions and subsamplings. We denote  $\check{x}[n] = x[-n]$  and

$$\check{x}[n] = \begin{cases} x[p] & \text{if } n = 2p \\ 0 & \text{if } n = 2p + 1. \end{cases} \quad (7.101)$$

**Theorem 7.10:** *Mallat.* At the decomposition,

$$a_{j+1}[p] = \sum_{n=-\infty}^{+\infty} h[n - 2p] a_j[n] = a_j \star \bar{h}[2p], \quad (7.102)$$

$$d_{j+1}[p] = \sum_{n=-\infty}^{+\infty} g[n - 2p] a_j[n] = a_j \star \bar{g}[2p]. \quad (7.103)$$

At the reconstruction,

$$\begin{aligned} a_j[p] &= \sum_{n=-\infty}^{+\infty} h[p - 2n] a_{j+1}[n] + \sum_{n=-\infty}^{+\infty} g[p - 2n] d_{j+1}[n] \\ &= \check{a}_{j+1} \star h[p] + \check{d}_{j+1} \star g[p]. \end{aligned} \quad (7.104)$$

**Proof of (7.102).** Any  $\phi_{j+1,p} \in \mathbf{V}_{j+1} \subset \mathbf{V}_j$  can be decomposed in the orthonormal basis  $\{\phi_{j,n}\}_{n \in \mathbb{Z}}$  of  $\mathbf{V}_j$ :

$$\phi_{j+1,p} = \sum_{n=-\infty}^{+\infty} \langle \phi_{j+1,p}, \phi_{j,n} \rangle \phi_{j,n}. \quad (7.105)$$

With the change of variable  $t' = 2^{-j}t - 2p$ , we obtain

$$\begin{aligned} \langle \phi_{j+1,p}, \phi_{j,n} \rangle &= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2^{j+1}}} \phi\left(\frac{t - 2^{j+1}p}{2^{j+1}}\right) \frac{1}{\sqrt{2^j}} \phi^*\left(\frac{t - 2^j n}{2^j}\right) dt \\ &= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2}} \phi\left(\frac{t}{2}\right) \phi^*(t - n + 2p) dt \\ &= \left\langle \frac{1}{\sqrt{2}} \phi\left(\frac{t}{2}\right), \phi(t - n + 2p) \right\rangle = h[n - 2p]. \end{aligned} \quad (7.106)$$

Thus, (7.105) implies that

$$\phi_{j+1,p} = \sum_{n=-\infty}^{+\infty} h[n - 2p] \phi_{j,n}. \quad (7.107)$$

Computing the inner product of  $f$  with the vectors on each side of this equality yields (7.102).

**Proof of (7.103).** Since  $\psi_{j+1,p} \in \mathbf{W}_{j+1} \subset \mathbf{V}_j$ , it can be decomposed as

$$\psi_{j+1,p} = \sum_{n=-\infty}^{+\infty} \langle \psi_{j+1,p}, \phi_{j,n} \rangle \phi_{j,n}.$$

As in (7.106), the change of variable  $t' = 2^{-j}t - 2p$  proves that

$$\langle \psi_{j+1,p}, \phi_{j,n} \rangle = \left\langle \frac{1}{\sqrt{2}} \psi\left(\frac{t}{2}\right), \phi(t - n + 2p) \right\rangle = g[n - 2p], \quad (7.108)$$

and thus,

$$\psi_{j+1,p} = \sum_{n=-\infty}^{+\infty} g[n - 2p] \phi_{j,n}. \quad (7.109)$$

Taking the inner product with  $f$  on each side gives (7.103).

**Proof of (7.104).** Since  $\mathbf{W}_{j+1}$  is the orthogonal complement of  $\mathbf{V}_{j+1}$  in  $\mathbf{V}_j$ , the union of the two bases  $\{\psi_{j+1,n}\}_{n \in \mathbb{Z}}$  and  $\{\phi_{j+1,n}\}_{n \in \mathbb{Z}}$  is an orthonormal basis of  $\mathbf{V}_j$ . Thus, any  $\phi_{j,p}$  can be decomposed in this basis:

$$\begin{aligned}\phi_{j,p} &= \sum_{n=-\infty}^{+\infty} \langle \phi_{j,p}, \phi_{j+1,n} \rangle \phi_{j+1,n} \\ &\quad + \sum_{n=-\infty}^{+\infty} \langle \phi_{j,p}, \psi_{j+1,n} \rangle \psi_{j+1,n}.\end{aligned}$$

Inserting (7.106) and (7.108) yields

$$\phi_{j,p} = \sum_{n=-\infty}^{+\infty} h[p - 2n] \phi_{j+1,n} + \sum_{n=-\infty}^{+\infty} g[p - 2n] \psi_{j+1,n}.$$

Taking the inner product with  $f$  on both sides of this equality gives (7.104). ■

Theorem 7.10 proves that  $a_{j+1}$  and  $d_{j+1}$  are computed by taking every other sample of the convolution of  $a_j$  with  $\bar{h}$  and  $\bar{g}$ , respectively, as illustrated by Figure 7.12. The filter  $\bar{h}$  removes the higher frequencies of the inner product sequence  $a_j$ , whereas  $\bar{g}$  is a high-pass filter that collects the remaining highest frequencies. The reconstruction (7.104) is an interpolation that inserts zeroes to expand  $a_{j+1}$  and  $d_{j+1}$  and filters these signals, as shown in Figure 7.12.

An *orthogonal wavelet representation* of  $a_L = \langle f, \phi_{L,n} \rangle$  is composed of wavelet coefficients of  $f$  at scales  $2^L < 2^j \leq 2^J$ , plus the remaining approximation at the largest scale  $2^J$ :

$$[\{d_j\}_{L \leq j \leq J}, a_J]. \quad (7.110)$$

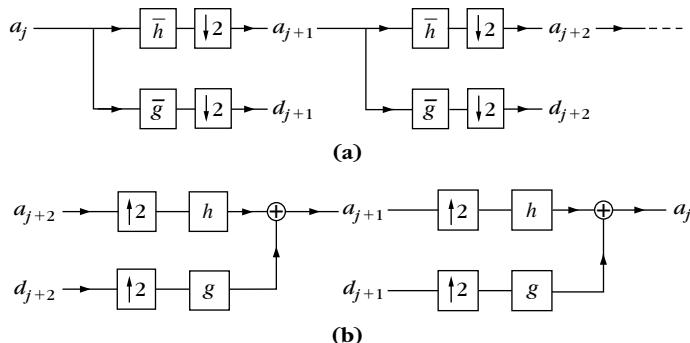


FIGURE 7.12

(a) A fast wavelet transform is computed with a cascade of filterings with  $\bar{h}$  and  $\bar{g}$  followed by a factor 2 subsampling. (b) A fast inverse wavelet transform reconstructs progressively each  $a_j$  by inserting zeroes between samples of  $a_{j+1}$  and  $d_{j+1}$ , filtering and adding the output.

It is computed from  $a_L$  by iterating (7.102) and (7.103) for  $L \leq j < J$ . Figure 7.7 gives a numerical example computed with the cubic spline filter of Table 7.1. The original signal  $a_L$  is recovered from this wavelet representation by iterating the reconstruction (7.104) for  $J > j \geq L$ .

### Initialization

Most often the discrete input signal  $b[n]$  is obtained by a finite-resolution device that averages and samples an analog input signal. For example, a CCD camera filters the light intensity by the optics and each photoreceptor averages the input light over its support. Thus, a pixel value measures average light intensity. If the sampling distance is  $N^{-1}$ , to define and compute the wavelet coefficients, we need to associate to  $b[n]$  a function  $f(t) \in \mathbf{V}_L$  approximated at the scale  $2^L = N^{-1}$ , and compute  $a_L[n] = \langle f, \phi_{L,n} \rangle$ . Exercise 7.6 explains how to compute  $a_L[n] = \langle f, \phi_{L,n} \rangle$  so that  $b[n] = f(N^{-1}n)$ .

A simpler and faster approach considers

$$f(t) = \sum_{n=-\infty}^{+\infty} b[n] \phi\left(\frac{t - 2^L n}{2^L}\right) \in \mathbf{V}_L.$$

Since  $\{\phi_{L,n}(t) = 2^{-L/2} \phi(2^{-L}t - n)\}_{n \in \mathbb{Z}}$  is orthonormal and  $2^L = N^{-1}$ ,

$$b[n] = N^{1/2} \langle f, \phi_{L,n} \rangle = N^{1/2} a_L[n].$$

But  $\hat{\phi}(0) = \int_{-\infty}^{\infty} \phi(t) dt = 1$ , so

$$N^{1/2} a_L[n] = \int_{-\infty}^{+\infty} f(t) \frac{1}{N^{-1}} \phi\left(\frac{t - N^{-1}n}{N^{-1}}\right) dt$$

is a weighted average of  $f$  in the neighborhood of  $N^{-1}n$  over a domain proportional to  $N^{-1}$ . Thus, if  $f$  is regular,

$$b[n] = N^{1/2} a_L[n] \approx f(N^{-1}n). \quad (7.111)$$

If  $\psi$  is a coiflet and  $f(t)$  is regular in the neighborhood of  $N^{-1}n$ , then (7.100) shows that  $N^{-1/2} a_L[n]$  is a high-order approximation of  $f(N^{-1}n)$ .

### Finite Signals

Let us consider a signal  $f$  with a support in  $[0, 1]$  and that is approximated with a uniform sampling at intervals  $N^{-1}$ . The resulting approximation  $a_L$  has  $N = 2^{-L}$  samples. This is the case in Figure 7.7 with  $N = 1024$ . Computing the convolutions with  $\bar{h}$  and  $\bar{g}$  at abscissa close to 0 or close to  $N$  requires knowing the values of  $a_L[n]$  beyond the boundaries  $n = 0$  and  $n = N - 1$ . These boundary problems may be solved with one of the three approaches described in Section 7.5.

Section 7.5.1 explains the simplest algorithm, which periodizes  $a_L$ . The convolutions in Theorem 7.10 are replaced by circular convolutions. This is equivalent to decomposing  $f$  in a periodic wavelet basis of  $\mathbf{L}^2[0, 1]$ . This algorithm has the disadvantage of creating large wavelet coefficients at the borders.

If  $\psi$  is symmetric or antisymmetric, we can use a folding procedure described in Section 7.5.2, which creates smaller wavelet coefficients at the border. It decomposes  $f$  in a folded wavelet basis of  $L^2[0, 1]$ . However, we mentioned in Section 7.2.3 that Haar is the only symmetric wavelet with a compact support. Higher-order spline wavelets have a symmetry, but  $h$  must be truncated in numerical calculations.

The most efficient boundary treatment is described in Section 7.5.3, but the implementation is more complicated. Boundary wavelets that keep their vanishing moments are designed to avoid creating large-amplitude coefficients when  $f$  is regular. The fast algorithm is implemented with special boundary filters and requires the same number of calculations as the two other methods.

### Complexity

Suppose that  $h$  and  $g$  have  $K$  nonzero coefficients. Let  $a_L$  be a signal of size  $N = 2^{-L}$ . With appropriate boundary calculations, each  $a_j$  and  $d_j$  has  $2^{-j}$  samples. Equations (7.102) and (7.103) compute  $a_{j+1}$  and  $d_{j+1}$  from  $a_j$  with  $2^{-j}K$  additions and multiplications. Therefore, the wavelet representation (7.110) is calculated with at most  $2KN$  additions and multiplications. The reconstruction (7.104) of  $a_j$  from  $a_{j+1}$  and  $d_{j+1}$  is also obtained with  $2^{-j}K$  additions and multiplications. The original signal  $a_L$  is also recovered from the wavelet representation with at most  $2KN$  additions and multiplications.

### Wavelet Graphs

The graphs of  $\phi$  and  $\psi$  are computed numerically with the inverse wavelet transform. If  $f = \phi$ , then  $a_0[n] = \delta[n]$  and  $d_j[n] = 0$  for all  $L < j \leq 0$ . The inverse wavelet transform computes  $a_L$  and (7.111) shows that

$$N^{1/2} a_L[n] \approx \phi(N^{-1}n).$$

If  $\phi$  is regular and  $N$  is large enough, we recover a precise approximation of the graph of  $\phi$  from  $a_L$ .

Similarly, if  $f = \psi$ , then  $a_0[n] = 0$ ,  $d_0[n] = \delta[n]$ , and  $d_j[n] = 0$  for  $L < j < 0$ . Then  $a_L[n]$  is calculated with the inverse wavelet transform and  $N^{1/2} a_L[n] \approx \psi(N^{-1}n)$ . The Daubechies wavelets and scaling functions in Figure 7.10 are calculated with this procedure.

### 7.3.2 Perfect Reconstruction Filter Banks

The fast discrete wavelet transform decomposes signals into low-pass and high-pass components subsampled by 2; the inverse transform performs the reconstruction. The study of such classical multirate filter banks became a major signal-processing topic in 1976, when Croisier, Esteban, and Galand [189] discovered that it is possible to perform such decompositions and reconstructions with *quadrature mirror filters* (Exercise 7.7). However, besides the simple Haar filter, a quadrature mirror filter cannot have a finite impulse response. In 1984, Smith and Barnwell [444] and Mintzer [376] found necessary and sufficient conditions for obtaining perfect

reconstruction orthogonal filters with a finite impulse response that they called *conjugate mirror filters*. The theory was completed by the biorthogonal equations of Vetterli [471, 472] and the general paraunitary matrix theory of Vaidyanathan [469]. We follow this digital signal-processing approach, which gives a simple understanding of conjugate mirror filter conditions. More complete presentations of filter bank properties can be found in [1, 2, 63, 68, 69].

### Filter Bank

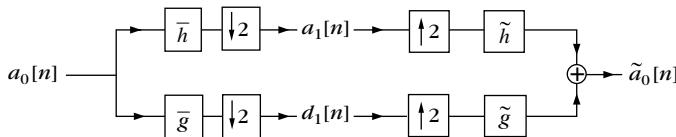
A two-channel multirate filter bank convolves a signal  $a_0$  with a low-pass filter  $\bar{h}[n] = h[-n]$  and a high-pass filter  $\bar{g}[n] = g[-n]$  and subsamples by 2 the output:

$$a_1[n] = a_0 \star \bar{h}[2n] \quad \text{and} \quad d_1[n] = a_0 \star \bar{g}[2n]. \quad (7.112)$$

A reconstructed signal  $\tilde{a}_0$  is obtained by filtering the zero expanded signals with a dual low-pass filter  $\tilde{h}$  and a dual high-pass filter  $\tilde{g}$ , as shown in Figure 7.13. With the zero insertion notation (7.101) it yields

$$\tilde{a}_0[n] = \check{a}_1 \star \tilde{h}[n] + \check{d}_1 \star \tilde{g}[n]. \quad (7.113)$$

We study necessary and sufficient conditions on  $h, g, \bar{h}$ , and  $\bar{g}$  to guarantee a perfect reconstruction  $\tilde{a}_0 = a_0$ .



**FIGURE 7.13**

The input signal is filtered by a low-pass and a high-pass filter and subsampled. The reconstruction is performed by inserting zeroes and filtering with dual filters  $\tilde{h}$  and  $\tilde{g}$ .

### Subsampling and Zero Interpolation

Subsamplings and expansions with zero insertions have simple expressions in the Fourier domain. Since  $\hat{x}(\omega) = \sum_{n=-\infty}^{+\infty} x[n] e^{-in\omega}$ , the Fourier series of the subsampled signal  $y[n] = x[2n]$  can be written as

$$\hat{y}(2\omega) = \sum_{n=-\infty}^{+\infty} x[2n] e^{-i2n\omega} = \frac{1}{2} (\hat{x}(\omega) + \hat{x}(\omega + \pi)). \quad (7.114)$$

The component  $\hat{x}(\omega + \pi)$  creates a frequency folding. This *aliasing* must be canceled at the reconstruction.

The insertion of zeros defines

$$y[n] = \check{x}[n] = \begin{cases} x[p] & \text{if } n = 2p \\ 0 & \text{if } n = 2p + 1, \end{cases}$$

that has a Fourier transform

$$\hat{y}(\omega) = \sum_{n=-\infty}^{+\infty} x[n] e^{-i2n\omega} = \hat{x}(2\omega). \quad (7.115)$$

Theorem 7.11 gives Vetterli's [471] biorthogonal conditions, which guarantee that  $\tilde{a}_0 = a_0$ .

**Theorem 7.11:** *Vetterli.* The filter bank performs an exact reconstruction for any input signal if and only if

$$\hat{h}^*(\omega + \pi) \hat{\tilde{h}}(\omega) + \hat{g}^*(\omega + \pi) \hat{\tilde{g}}(\omega) = 0, \quad (7.116)$$

and

$$\hat{h}^*(\omega) \hat{\tilde{h}}(\omega) + \hat{g}^*(\omega) \hat{\tilde{g}}(\omega) = 2. \quad (7.117)$$

**Proof.** We first relate the Fourier transform of  $a_1$  and  $d_1$  to the Fourier transform of  $a_0$ . Since  $h$  and  $g$  are real, the transfer functions of  $\tilde{h}$  and  $\tilde{g}$  are, respectively,  $\hat{h}(-\omega) = \hat{h}^*(\omega)$  and  $\hat{g}(-\omega) = \hat{g}^*(\omega)$ . By using (7.114), we derive from the definition (7.112) of  $a_1$  and  $d_1$  that

$$\hat{a}_1(2\omega) = \frac{1}{2} (\hat{a}_0(\omega) \hat{h}^*(\omega) + \hat{a}_0(\omega + \pi) \hat{h}^*(\omega + \pi)), \quad (7.118)$$

$$\hat{d}_1(2\omega) = \frac{1}{2} (\hat{a}_0(\omega) \hat{g}^*(\omega) + \hat{a}_0(\omega + \pi) \hat{g}^*(\omega + \pi)). \quad (7.119)$$

The expression (7.113) of  $\tilde{a}_0$  and the zero insertion property (7.115) also imply

$$\hat{\tilde{a}}_0(\omega) = \hat{a}_1(2\omega) \hat{\tilde{h}}(\omega) + \hat{d}_1(2\omega) \hat{\tilde{g}}(\omega). \quad (7.120)$$

Thus,

$$\begin{aligned} \hat{\tilde{a}}_0(\omega) &= \frac{1}{2} (\hat{h}^*(\omega) \hat{\tilde{h}}(\omega) + \hat{g}^*(\omega) \hat{\tilde{g}}(\omega)) \hat{a}_0(\omega) \\ &\quad + \frac{1}{2} (\hat{h}^*(\omega + \pi) \hat{\tilde{h}}(\omega) + \hat{g}^*(\omega + \pi) \hat{\tilde{g}}(\omega)) \hat{a}_0(\omega + \pi). \end{aligned}$$

To obtain  $a_0 = \tilde{a}_0$  for all  $a_0$ , the filters must cancel the aliasing term  $\hat{a}_0(\omega + \pi)$  and guarantee a unit gain for  $\hat{a}_0(\omega)$ , which proves equations (7.116) and (7.117). ■

Theorem 7.11 proves that the reconstruction filters  $\tilde{h}$  and  $\tilde{g}$  are entirely specified by the decomposition filters  $h$  and  $g$ . In matrix form, it can be rewritten

$$\begin{pmatrix} \hat{h}(\omega) & \hat{g}(\omega) \\ \hat{h}(\omega + \pi) & \hat{g}(\omega + \pi) \end{pmatrix} \times \begin{pmatrix} \hat{\tilde{h}}^*(\omega) \\ \hat{\tilde{g}}^*(\omega) \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \end{pmatrix}. \quad (7.121)$$

The inversion of this  $2 \times 2$  matrix yields

$$\begin{pmatrix} \hat{\tilde{h}}^*(\omega) \\ \hat{\tilde{g}}^*(\omega) \end{pmatrix} = \frac{2}{\Delta(\omega)} \begin{pmatrix} \hat{g}(\omega + \pi) \\ -\hat{h}(\omega + \pi) \end{pmatrix}, \quad (7.122)$$

where  $\Delta(\omega)$  is the determinant

$$\Delta(\omega) = \hat{h}(\omega)\hat{g}(\omega + \pi) - \hat{h}(\omega + \pi)\hat{g}(\omega). \quad (7.123)$$

The reconstruction filters are stable only if the determinant does not vanish for all  $\omega \in [-\pi, \pi]$ . Vaidyanathan [469] has extended this result to multirate filter banks with an arbitrary number  $M$  of channels by showing that the resulting matrices of filters satisfy paraunitary properties [68].

### **Finite Impulse Response**

When all filters have a finite impulse response, the determinant  $\Delta(\omega)$  can be evaluated. This yields simpler relations between the decomposition and reconstruction filters.

**Theorem 7.12.** Perfect reconstruction filters satisfy

$$\hat{h}^*(\omega)\hat{\tilde{h}}(\omega) + \hat{h}^*(\omega + \pi)\hat{\tilde{h}}(\omega + \pi) = 2. \quad (7.124)$$

For finite impulse-response filters, there exist  $a \in \mathbb{R}$  and  $l \in \mathbb{Z}$  such that

$$\hat{g}(\omega) = a e^{-i(2l+1)\omega} \hat{\tilde{h}}^*(\omega + \pi) \quad \text{and} \quad \hat{\tilde{g}}(\omega) = a^{-1} e^{-i(2l+1)\omega} \hat{h}^*(\omega + \pi). \quad (7.125)$$

**Proof.** Equation (7.122) proves that

$$\hat{\tilde{h}}^*(\omega) = \frac{2}{\Delta(\omega)} \hat{g}(\omega + \pi) \quad \text{and} \quad \hat{\tilde{g}}^*(\omega) = \frac{-2}{\Delta(\omega)} \hat{h}(\omega + \pi). \quad (7.126)$$

Thus,

$$\hat{g}(\omega)\hat{\tilde{g}}^*(\omega) = -\frac{\Delta(\omega + \pi)}{\Delta(\omega)} \hat{\tilde{h}}^*(\omega + \pi)\hat{h}(\omega + \pi). \quad (7.127)$$

The definition (7.123) implies that  $\Delta(\omega + \pi) = -\Delta(\omega)$ . Inserting (7.127) in (7.117) yields (7.124).

The Fourier transform of finite impulse-response filters is a finite series in  $\exp(\pm i\omega)$ . Therefore, the determinant  $\Delta(\omega)$  defined by (7.123) is a finite series. Moreover, (7.126) proves that  $\Delta^{-1}(\omega)$  must also be a finite series. A finite series in  $\exp(\pm i\omega)$  that has an inverse that is also a finite series must have a single term. Since  $\Delta(\omega) = -\Delta(\omega + \pi)$  the exponent  $n$  must be odd. This proves that there exist  $l \in \mathbb{Z}$  and  $a \in \mathbb{R}$  such that

$$\Delta(\omega) = -2a \exp[i(2l+1)\omega]. \quad (7.128)$$

Inserting this expression in (7.126) yields (7.125). ■

The factor  $a$  is a gain that is inverse for the decomposition and reconstruction filters and  $l$  is a reverse shift. We generally set  $a = 1$  and  $l = 0$ . In the time domain (7.125) can then be rewritten as

$$g[n] = (-1)^{1-n} \tilde{h}[1-n] \quad \text{and} \quad \tilde{g}[n] = (-1)^{1-n} h[1-n]. \quad (7.129)$$

The two pairs of filters  $(h, g)$  and  $(\tilde{h}, \tilde{g})$  play a symmetric role and can be inverted.

### Conjugate Mirror Filters

If we impose that the decomposition filter  $h$  is equal to the reconstruction filter  $\tilde{h}$ , then (7.124) is the condition of Smith and Barnwell [444] and Mintzer [376] that defines conjugate mirror filters:

$$|\hat{h}(\omega)|^2 + |\hat{h}(\omega + \pi)|^2 = 2. \quad (7.130)$$

It is identical to the filter condition (7.29) that is required in order to synthesize orthogonal wavelets. Section 7.3.3 proves that it is also equivalent to discrete orthogonality properties.

### 7.3.3 Biorthogonal Bases of $\ell^2(\mathbb{Z})$

The decomposition of a discrete signal in a multirate filter bank is interpreted as an expansion in a basis of  $\ell^2(\mathbb{Z})$ . Observe first that the low-pass and high-pass signals of a filter bank computed with (7.112) can be rewritten as inner products in  $\ell^2(\mathbb{Z})$ :

$$a_1[l] = \sum_{n=-\infty}^{+\infty} a_0[n] h[n - 2l] = \langle a_0[n], h[n - 2l] \rangle, \quad (7.131)$$

$$d_1[l] = \sum_{n=-\infty}^{+\infty} a_0[n] g[n - 2l] = \langle a_0[n], g[n - 2l] \rangle. \quad (7.132)$$

The signal recovered by the reconstructing filters is

$$a_0[n] = \sum_{l=-\infty}^{+\infty} a_1[l] \tilde{h}[n - 2l] + \sum_{l=-\infty}^{+\infty} d_1[l] \tilde{g}[n - 2l]. \quad (7.133)$$

Inserting (7.131) and (7.132) yields

$$a_0[n] = \sum_{l=-\infty}^{+\infty} \langle f[k], h[k - 2l] \rangle \tilde{h}[n - 2l] + \sum_{l=-\infty}^{+\infty} \langle f[k], g[k - 2l] \rangle \tilde{g}[n - 2l]. \quad (7.134)$$

We recognize the decomposition of  $a_0$  over dual families of vectors  $\{\tilde{h}[n - 2l], \tilde{g}[n - 2l]\}_{l \in \mathbb{Z}}$  and  $\{h[n - 2l], g[n - 2l]\}_{l \in \mathbb{Z}}$ . Theorem 7.13 proves that these two families are biorthogonal.

**Theorem 7.13.** If  $h$ ,  $g$ ,  $\tilde{h}$ , and  $\tilde{g}$  are perfect reconstruction filters, and their Fourier transforms are bounded, then  $\{\tilde{h}[n - 2l], \tilde{g}[n - 2l]\}_{l \in \mathbb{Z}}$  and  $\{h[n - 2l], g[n - 2l]\}_{l \in \mathbb{Z}}$  are biorthogonal Riesz bases of  $\ell^2(\mathbb{Z})$ .

**Proof.** To prove that these families are biorthogonal we must show that for all  $n \in \mathbb{Z}$

$$\langle \tilde{h}[n], h[n - 2l] \rangle = \delta[l] \quad (7.135)$$

$$\langle \tilde{g}[n], g[n - 2l] \rangle = \delta[l] \quad (7.136)$$

and

$$\langle \tilde{h}[n], g[n - 2l] \rangle = \langle \tilde{g}[n], h[n - 2l] \rangle = 0. \quad (7.137)$$

For perfect reconstruction filters, (7.124) proves that

$$\frac{1}{2} \left( \hat{h}^*(\omega) \hat{\tilde{h}}(\omega) + \hat{h}^*(\omega + \pi) \hat{\tilde{h}}(\omega + \pi) \right) = 1.$$

In the time domain, this equation becomes

$$\bar{h} \star \tilde{h}[2l] = \sum_{k=-\infty}^{+\infty} \tilde{h}[n] \bar{h}[n - 2l] = \delta[l], \quad (7.138)$$

which verifies (7.135). The same proof as for (7.124) shows that

$$\frac{1}{2} \left( \hat{g}^*(\omega) \hat{\tilde{g}}(\omega) + \hat{g}^*(\omega + \pi) \hat{\tilde{g}}(\omega + \pi) \right) = 1.$$

In the time domain, this equation yields (7.136). It also follows from (7.122) that

$$\frac{1}{2} \left( \hat{g}^*(\omega) \hat{\tilde{h}}(\omega) + \hat{g}^*(\omega + \pi) \hat{\tilde{h}}(\omega + \pi) \right) = 0,$$

and

$$\frac{1}{2} \left( \hat{h}^*(\omega) \hat{\tilde{g}}(\omega) + \hat{h}^*(\omega + \pi) \hat{\tilde{g}}(\omega + \pi) \right) = 0.$$

The inverse Fourier transforms of these two equations yield (7.137).

To finish the proof, one must show the existence of Riesz bounds. The reader can verify that this is a consequence of the fact that the Fourier transform of each filter is bounded. ■

### Orthogonal Bases

A Riesz basis is orthonormal if the dual basis is the same as the original basis. For filter banks this means that  $h = \tilde{h}$  and  $g = \tilde{g}$ . The filter  $h$  is then a conjugate mirror filter

$$|\hat{h}(\omega)|^2 + |\hat{h}(\omega + \pi)|^2 = 2. \quad (7.139)$$

The resulting family  $\{h[n - 2l], g[n - 2l]\}_{l \in \mathbb{Z}}$  is an orthogonal basis of  $\ell^2(\mathbb{Z})$ .

### Discrete Wavelet Bases

The construction of conjugate mirror filters is simpler than the construction of orthogonal wavelet bases of  $L^2(\mathbb{R})$ . Why then should we bother with continuous time models of wavelets, since in any case, all computations are discrete and rely on conjugate mirror filters? The reason is that conjugate mirror filters are most often used in filter banks that cascade several levels of filterings and subsamplings. Thus, it is necessary to understand the behavior of such a cascade [407]. In a wavelet filter bank tree, the output of the low-pass filter  $\bar{h}$  is subdecomposed, whereas the output of the high-pass filter  $\bar{g}$  is not; this is illustrated in Figure 7.12. Suppose that the sampling distance of the original discrete signal is  $N^{-1}$ . We denote  $a_L[n]$  for this discrete signal, with  $2^L = N^{-1}$ . At the depth  $j - L \geq 0$  of this filter bank tree, the low-pass signal  $a_j$  and high-pass signal  $d_j$  can be written as

$$a_j[l] = a_L \star \bar{\phi}_j[2^{j-L}l] = \langle a_L[n], \phi_j[n - 2^{j-L}l] \rangle$$

and

$$d_j[l] = a_L \star \bar{\psi}_j[2^{j-L}l] = \langle a_L[n], \psi_j[n - 2^{j-L}l] \rangle.$$

The Fourier transforms of these equivalent filters are

$$\hat{\phi}_j(\omega) = \prod_{p=0}^{j-L-1} \hat{h}(2^p \omega) \quad \text{and} \quad \hat{\psi}_j(\omega) = \hat{g}(2^{j-L-1} \omega) \prod_{p=0}^{j-L-2} \hat{h}(2^p \omega). \quad (7.140)$$

A filter bank tree of depth  $J - L \geq 0$  decomposes  $a_L$  over the family of vectors

$$\left[ \left\{ \phi_j[n - 2^{j-L}l] \right\}_{l \in \mathbb{Z}}, \quad \left\{ \psi_j[n - 2^{j-L}l] \right\}_{l < j \leq J, l \in \mathbb{Z}} \right]. \quad (7.141)$$

For conjugate mirror filters, one can verify that this family is an orthonormal basis of  $\ell^2(\mathbb{Z})$ . These discrete vectors are close to a uniform sampling of the continuous time-scaling functions  $\phi_j(t) = 2^{-j/2} \phi(2^{-j}t)$  and wavelets  $\psi_j(t) = 2^{-j/2} \phi(2^{-j}t)$ . When the number  $L - j$  of successive convolutions increases, one can verify that  $\phi_j[n]$  and  $\psi_j[n]$  converge, respectively, to  $N^{-1/2} \phi(N^{-1}n)$  and  $N^{-1/2} \psi(N^{-1}n)$ .

The factor  $N^{-1/2}$  normalizes the  $\ell^2(\mathbb{Z})$  norm of these sampled functions. If  $L - j = 4$ , then  $\phi_j[n]$  and  $\psi_j[n]$  are already very close to these limit values. Thus, the impulse responses  $\phi_j[n]$  and  $\psi_j[n]$  of the filter bank are much closer to continuous time-scaling functions and wavelets than they are to the original conjugate mirror filters  $h$  and  $g$ . This explains why wavelets provide appropriate models for understanding the applications of these filter banks. Chapter 8 relates more general filter banks to wavelet packet bases.

If the decomposition and reconstruction filters of the filter bank are different, the resulting basis (7.141) is nonorthogonal. The stability of this discrete wavelet basis does not degrade when the depth  $J - L$  of the filter bank increases. The next section shows that the corresponding continuous time wavelet  $\psi(t)$  generates a Riesz basis of  $\mathbf{L}^2(\mathbb{R})$ .

## 7.4 BIORTHOGONAL WAVELET BASES

The stability and completeness properties of biorthogonal wavelet bases are described for perfect reconstruction filters  $h$  and  $\tilde{h}$  having a finite impulse response. The design of linear phase wavelets with compact support is explained in Section 7.4.2.

### 7.4.1 Construction of Biorthogonal Wavelet Bases

An infinite cascade of perfect reconstruction filters  $(h, g)$  and  $(\tilde{h}, \tilde{g})$  yields two scaling functions and wavelets having a Fourier transform that satisfies

$$\hat{\phi}(2\omega) = \frac{1}{\sqrt{2}} \hat{h}(\omega) \hat{\phi}(\omega), \quad \hat{\tilde{\phi}}(2\omega) = \frac{1}{\sqrt{2}} \hat{\tilde{h}}(\omega) \hat{\tilde{\phi}}(\omega), \quad (7.142)$$

$$\hat{\psi}(2\omega) = \frac{1}{\sqrt{2}} \hat{g}(\omega) \hat{\phi}(\omega), \quad \hat{\tilde{\psi}}(2\omega) = \frac{1}{\sqrt{2}} \hat{\tilde{g}}(\omega) \hat{\tilde{\phi}}(\omega). \quad (7.143)$$

In the time domain, these relations become

$$\phi(t) = \sqrt{2} \sum_{n=-\infty}^{+\infty} h[n] \phi(2t-n), \quad \tilde{\phi}(t) = \sqrt{2} \sum_{n=-\infty}^{+\infty} \tilde{h}[n] \tilde{\phi}(2t-n) \quad (7.144)$$

$$\psi(t) = \sqrt{2} \sum_{n=-\infty}^{+\infty} g[n] \phi(2t-n), \quad \tilde{\psi}(t) = \sqrt{2} \sum_{n=-\infty}^{+\infty} \tilde{g}[n] \tilde{\phi}(2t-n). \quad (7.145)$$

The perfect reconstruction conditions are given by Theorem 7.12. If we normalize the gain and shift to  $a = 1$  and  $l = 0$ , the filters must satisfy

$$\hat{h}^*(\omega) \hat{h}(\omega) + \hat{h}^*(\omega + \pi) \hat{h}(\omega + \pi) = 2, \quad (7.146)$$

and

$$\hat{g}(\omega) = e^{-i\omega} \hat{h}^*(\omega + \pi), \quad \hat{\tilde{g}}(\omega) = e^{-i\omega} \hat{h}^*(\omega + \pi). \quad (7.147)$$

Wavelets should have a zero average, which means that  $\hat{\psi}(0) = \tilde{\hat{\psi}}(0) = 0$ . This is obtained by setting  $\hat{g}(0) = \tilde{\hat{g}}(0) = 0$  and thus  $\hat{h}(\pi) = \tilde{\hat{h}}(\pi) = 0$ . The perfect reconstruction condition (7.146) implies that  $\hat{h}^*(0) \hat{h}(0) = 2$ . Since both filters are defined up to multiplicative constants equal to  $\lambda$  and  $\lambda^{-1}$ , respectively, we adjust  $\lambda$  so that  $\hat{h}(0) = \tilde{\hat{h}}(0) = \sqrt{2}$ .

In the following, we also suppose that  $h$  and  $\tilde{h}$  are finite impulse-response filters. One can then prove [19] that

$$\hat{\phi}(\omega) = \prod_{p=1}^{+\infty} \frac{\hat{h}(2^{-p}\omega)}{\sqrt{2}} \quad \text{and} \quad \hat{\tilde{\phi}}(\omega) = \prod_{p=1}^{+\infty} \frac{\hat{h}(2^{-p}\omega)}{\sqrt{2}} \quad (7.148)$$

are the Fourier transforms of distributions of compact support. However, these distributions may exhibit wild behavior and have infinite energy. Some further conditions must be imposed to guarantee that  $\hat{\phi}$  and  $\hat{\tilde{\phi}}$  are the Fourier transforms of finite energy functions. Theorem 7.14 gives sufficient conditions on the perfect reconstruction filters for synthesizing biorthogonal wavelet bases of  $\mathbf{L}^2(\mathbb{R})$ .

**Theorem 7.14:** Cohen, Daubechies, Feauveau. Suppose that there exist strictly positive trigonometric polynomials  $P(e^{i\omega})$  and  $\tilde{P}(e^{i\omega})$  such that

$$\left| \hat{h}\left(\frac{\omega}{2}\right) \right|^2 P(e^{i\omega/2}) + \left| \hat{h}\left(\frac{\omega}{2} + \pi\right) \right|^2 P(e^{i(\omega/2 + \pi)}) = 2P(e^{i\omega}), \quad (7.149)$$

$$\left| \hat{\tilde{h}}\left(\frac{\omega}{2}\right) \right|^2 \tilde{P}(e^{i\omega/2}) + \left| \hat{\tilde{h}}\left(\frac{\omega}{2} + \pi\right) \right|^2 \tilde{P}(e^{i(\omega/2 + \pi)}) = 2\tilde{P}(e^{i\omega}), \quad (7.150)$$

and that  $P$  and  $\tilde{P}$  are unique (up to normalization). Suppose that

$$\inf_{\omega \in [-\pi/2, \pi/2]} |\hat{h}(\omega)| > 0, \quad \inf_{\omega \in [-\pi/2, \pi/2]} |\hat{\tilde{h}}(\omega)| > 0. \quad (7.151)$$

Then, the functions  $\hat{\phi}$  and  $\hat{\tilde{\phi}}$  defined in (7.148) belong to  $\mathbf{L}^2(\mathbb{R})$ , and  $\phi$ ,  $\tilde{\phi}$  satisfy biorthogonal relations

$$\langle \phi(t), \tilde{\phi}(t-n) \rangle = \delta[n]. \quad (7.152)$$

The two wavelet families  $\{\psi_{j,n}\}_{(j,n)\in\mathbb{Z}^2}$  and  $\{\tilde{\psi}_{j,n}\}_{(j,n)\in\mathbb{Z}^2}$  are biorthogonal Riesz bases of  $\mathbf{L}^2(\mathbb{R})$ .

The proof of this theorem is in [172] and [19]. The hypothesis (7.151) is also imposed by Theorem 7.2, which constructs orthogonal bases of scaling functions. The conditions (7.149) and (7.150) do not appear in the construction of wavelet orthogonal bases because they are always satisfied with  $P(e^{i\omega}) = \tilde{P}(e^{i\omega}) = 1$ , and one can prove that constants are the only invariant trigonometric polynomials [341].

Biorthogonality means that for any  $(j, j', n, n') \in \mathbb{Z}^4$ ,

$$\langle \psi_{j,n}, \tilde{\psi}_{j',n'} \rangle = \delta[n - n'] \delta[j - j']. \quad (7.153)$$

Any  $f \in \mathbf{L}^2(\mathbb{R})$  has two possible decompositions in these bases:

$$f = \sum_{n,j=-\infty}^{+\infty} \langle f, \psi_{j,n} \rangle \psi_{j,n} = \sum_{n,j=-\infty}^{+\infty} \langle f, \tilde{\psi}_{j,n} \rangle \tilde{\psi}_{j,n}. \quad (7.154)$$

The Riesz stability implies that there exist  $A > 0$  and  $B > 0$  such that

$$A \|f\|^2 \leq \sum_{n,j=-\infty}^{+\infty} |\langle f, \psi_{j,n} \rangle|^2 \leq B \|f\|^2, \quad (7.155)$$

$$\frac{1}{B} \|f\|^2 \leq \sum_{n,j=-\infty}^{+\infty} |\langle f, \tilde{\psi}_{j,n} \rangle|^2 \leq \frac{1}{A} \|f\|^2. \quad (7.156)$$

### Multiresolutions

Biorthogonal wavelet bases are related to multiresolution approximations. The family  $\{\phi(t - n)\}_{n \in \mathbb{Z}}$  is a Riesz basis of the space  $\mathbf{V}_0$  it generates, whereas  $\{\tilde{\phi}(t - n)\}_{n \in \mathbb{Z}}$  is a Riesz basis of another space  $\tilde{\mathbf{V}}_0$ . Let  $\mathbf{V}_j$  and  $\tilde{\mathbf{V}}_j$  be the spaces defined by

$$\begin{aligned} f(t) \in \mathbf{V}_j &\Leftrightarrow f(2^j t) \in \mathbf{V}_0, \\ f(t) \in \tilde{\mathbf{V}}_j &\Leftrightarrow f(2^j t) \in \tilde{\mathbf{V}}_0. \end{aligned}$$

One can verify that  $\{\mathbf{V}_j\}_{j \in \mathbb{Z}}$  and  $\{\tilde{\mathbf{V}}_j\}_{j \in \mathbb{Z}}$  are two multiresolution approximations of  $\mathbf{L}^2(\mathbb{R})$ . For any  $j \in \mathbb{Z}$ ,  $\{\phi_{j,n}\}_{n \in \mathbb{Z}}$  and  $\{\tilde{\phi}_{j,n}\}_{n \in \mathbb{Z}}$  are Riesz bases of  $\mathbf{V}_j$  and  $\tilde{\mathbf{V}}_j$ . The dilated wavelets  $\{\psi_{j,n}\}_{n \in \mathbb{Z}}$  and  $\{\tilde{\psi}_{j,n}\}_{n \in \mathbb{Z}}$  are bases of two detail spaces  $\mathbf{W}_j$  and  $\tilde{\mathbf{W}}_j$  such that

$$\mathbf{V}_j \oplus \mathbf{W}_j = \mathbf{V}_{j-1} \quad \text{and} \quad \tilde{\mathbf{V}}_j \oplus \tilde{\mathbf{W}}_j = \tilde{\mathbf{V}}_{j-1}.$$

The biorthogonality of the decomposition and reconstruction wavelets implies that  $\mathbf{W}_j$  is not orthogonal to  $\mathbf{V}_j$  but is to  $\tilde{\mathbf{V}}_j$ , whereas  $\tilde{\mathbf{W}}_j$  is not orthogonal to  $\tilde{\mathbf{V}}_j$  but is to  $\mathbf{V}_j$ .

### Fast Biorthogonal Wavelet Transform

The perfect reconstruction filter bank discussed in Section 7.3.2 implements a fast biorthogonal wavelet transform. For any discrete signal input  $b[n]$  sampled

at intervals  $N^{-1} = 2^L$ , there exists  $f \in \mathbf{V}_L$  such that  $a_L[n] = \langle f, \phi_{L,n} \rangle = N^{-1/2} b[n]$ . The wavelet coefficients are computed by successive convolutions with  $\tilde{h}$  and  $\tilde{g}$ . Let  $a_j[n] = \langle f, \phi_{j,n} \rangle$  and  $d_j[n] = \langle f, \psi_{j,n} \rangle$ . As in Theorem 7.10, one can prove that

$$a_{j+1}[n] = a_j \star \tilde{h}[2n], \quad d_{j+1}[n] = a_j \star \tilde{g}[2n]. \quad (7.157)$$

The reconstruction is performed with the dual filters  $\tilde{h}$  and  $\tilde{g}$ :

$$a_j[n] = \check{d}_{j+1} \star \tilde{h}[n] + \check{d}_{j+1} \star \tilde{g}[n]. \quad (7.158)$$

If  $a_L$  includes  $N$  nonzero samples, the biorthogonal wavelet representation  $\{d_j\}_{L \leq j \leq J}, a_J\}$  is calculated with  $O(N)$  operations by iterating (7.157) for  $L \leq j < J$ . The reconstruction of  $a_L$  by applying (7.158) for  $J > j \geq L$  requires the same number of operations.

## 7.4.2 Biorthogonal Wavelet Design

The support size, the number of vanishing moments, the regularity, wavelet ordering, and the symmetry of biorthogonal wavelets is controlled with an appropriate design of  $h$  and  $\tilde{h}$ .

### Support

If the perfect reconstruction filters  $h$  and  $\tilde{h}$  have a finite impulse response, then the corresponding scaling functions and wavelets also have a compact support. As in Section 7.2.1, one can show that if  $h[n]$  and  $\tilde{h}[n]$  are nonzero, respectively, for  $N_1 \leq n \leq N_2$  and  $\tilde{N}_1 \leq n \leq \tilde{N}_2$ , then  $\phi$  and  $\tilde{\phi}$  have a support equal to  $[N_1, N_2]$  and  $[\tilde{N}_1, \tilde{N}_2]$ , respectively. Since

$$g[n] = (-1)^{1-n} h[1-n] \quad \text{and} \quad \tilde{g}[n] = (-1)^{1-n} \tilde{h}[1-n],$$

the supports of  $\psi$  and  $\tilde{\psi}$  defined in (7.145) are, respectively,

$$\left[ \frac{N_1 - \tilde{N}_2 + 1}{2}, \frac{N_2 - \tilde{N}_1 + 1}{2} \right] \quad \text{and} \quad \left[ \frac{\tilde{N}_1 - N_2 + 1}{2}, \frac{\tilde{N}_2 - N_1 + 1}{2} \right]. \quad (7.159)$$

Thus, both wavelets have a support of the same size and equal to

$$l = \frac{N_2 - N_1 + \tilde{N}_2 - \tilde{N}_1}{2}. \quad (7.160)$$

### Vanishing Moments

The number of vanishing moments of  $\psi$  and  $\tilde{\psi}$  depends on the number of zeroes at  $\omega = \pi$  of  $\hat{h}(\omega)$  and  $\hat{\tilde{h}}(\omega)$ . Theorem 7.4 proves that  $\psi$  has  $\tilde{p}$  vanishing moments if the derivatives of its Fourier transform satisfy  $\hat{\psi}^{(k)}(0) = 0$  for  $k \leq \tilde{p}$ . Since  $\hat{\phi}(0) = 1$ , (7.4.1) implies that it is equivalent to impose that  $\hat{g}(\omega)$  has a zero of order  $\tilde{p}$  at  $\omega = 0$ . Since  $\hat{g}(\omega) = e^{-i\omega} \hat{h}^*(\omega + \pi)$ , this means that  $\hat{h}(\omega)$  has a zero of order  $\tilde{p}$  at  $\omega = \pi$ . Similarly, the number of vanishing moments of  $\tilde{\psi}$  is equal to the number  $p$  of zeroes of  $\hat{h}(\omega)$  at  $\pi$ .

### Regularity

Although the regularity of a function is a priori independent of the number of vanishing moments, the smoothness of biorthogonal wavelets is related to their vanishing moments. The regularity of  $\phi$  and  $\psi$  is the same because (7.145) shows that  $\psi$  is a finite linear expansion of  $\phi$  translated. Tchamitchian's theorem (7.6) gives a sufficient condition for estimating this regularity. If  $\hat{h}(\omega)$  has a zero of order  $p$  at  $\pi$ , we can perform the factorization

$$\hat{h}(\omega) = \left( \frac{1 + e^{-i\omega}}{2} \right)^p \hat{l}(\omega). \quad (7.161)$$

Let  $B = \sup_{\omega \in [-\pi, \pi]} |\hat{l}(\omega)|$ . Theorem 7.6 proves that  $\phi$  is uniformly Lipschitz  $\alpha$  for

$$\alpha < \alpha_0 = p - \log_2 B - 1.$$

Generally,  $\log_2 B$  increases more slowly than  $p$ . This implies that the regularity of  $\phi$  and  $\psi$  increases with  $p$ , which is equal to the number of vanishing moments of  $\psi$ . Similarly, one can show that the regularity of  $\tilde{\psi}$  and  $\tilde{\phi}$  increases with  $\tilde{p}$ , which is the number of vanishing moments of  $\psi$ . If  $\hat{h}$  and  $\tilde{h}$  have different numbers of zeroes at  $\pi$ , the properties of  $\psi$  and  $\tilde{\psi}$  can be very different.

### Ordering of Wavelets

Since  $\psi$  and  $\tilde{\psi}$  might not have the same regularity and number of vanishing moments, the two reconstruction formulas

$$f = \sum_{n,j=-\infty}^{+\infty} \langle f, \psi_{j,n} \rangle \tilde{\psi}_{j,n}, \quad (7.162)$$

$$f = \sum_{n,j=-\infty}^{+\infty} \langle f, \tilde{\psi}_{j,n} \rangle \psi_{j,n} \quad (7.163)$$

are not equivalent. The decomposition (7.162) is obtained with the filters  $(h, g)$ , and the reconstruction with  $(\tilde{h}, \tilde{g})$ . The inverse formula (7.163) corresponds to  $(\tilde{h}, \tilde{g})$  at the decomposition and  $(h, g)$  at the reconstruction.

To produce small wavelet coefficients in regular regions we must compute the inner products using the wavelet with the maximum number of vanishing moments. The reconstruction is then performed with the other wavelet, which is generally the smoothest one. If errors are added to the wavelet coefficients, for example with a quantization, a smooth wavelet at the reconstruction introduces a smooth error. The number of vanishing moments of  $\psi$  is equal to the number  $\tilde{p}$  of zeroes at  $\pi$  of  $\hat{h}$ . Increasing  $\tilde{p}$  also increases the regularity of  $\tilde{\psi}$ . Thus, it is better to use  $h$  at the decomposition and  $\tilde{h}$  at the reconstruction if  $\hat{h}$  has fewer zeroes at  $\pi$  than  $\hat{\tilde{h}}$ .

### Symmetry

It is possible to construct smooth biorthogonal wavelets of compact support that are either symmetric or antisymmetric. This is impossible for orthogonal wavelets,

besides the particular case of the Haar basis. Symmetric or antisymmetric wavelets are synthesized with perfect reconstruction filters having a linear phase. If  $h$  and  $\tilde{h}$  have an odd number of nonzero samples and are symmetric about  $n = 0$ , the reader can verify that  $\phi$  and  $\tilde{\phi}$  are symmetric about  $t = 0$ , while  $\psi$  and  $\tilde{\psi}$  are symmetric with respect to a shifted center. If  $h$  and  $\tilde{h}$  have an even number of nonzero samples and are symmetric about  $n = 1/2$ , then  $\phi(t)$  and  $\tilde{\phi}(t)$  are symmetric about  $t = 1/2$ , while  $\psi$  and  $\tilde{\psi}$  are antisymmetric with respect to a shifted center. When the wavelets are symmetric or antisymmetric, wavelet bases over finite intervals are constructed with the folding procedure of Section 7.5.2.

### 7.4.3 Compactly Supported Biorthogonal Wavelets

We study the design of biorthogonal wavelets with a minimum-size support for a specified number of vanishing moments. Symmetric or antisymmetric compactly supported spline biorthogonal wavelet bases are constructed with a technique introduced in [172].

**Theorem 7.15:** Cohen, Daubechies, Feauveau. Biorthogonal wavelets  $\psi$  and  $\tilde{\psi}$  with, respectively,  $\tilde{p}$  and  $p$  vanishing moments have a support size of at least  $p + \tilde{p} - 1$ . CDF biorthogonal wavelets have a minimum support size  $p + \tilde{p} - 1$ .

**Proof.** The proof follows the same approach as the proof of Daubechies' theorem (7.7). One can verify that  $p$  and  $\tilde{p}$  must necessarily have the same parity. We concentrate on filters  $h[n]$  and  $\tilde{h}[n]$  that have a symmetry with respect to  $n = 0$  or  $n = 1/2$ . The general case proceeds similarly. We can then factor

$$\hat{h}(\omega) = \sqrt{2} \exp\left(\frac{-i\varepsilon\omega}{2}\right) \left(\cos \frac{\omega}{2}\right)^p L(\cos \omega), \quad (7.164)$$

$$\hat{\tilde{h}}(\omega) = \sqrt{2} \exp\left(\frac{-i\varepsilon\omega}{2}\right) \left(\cos \frac{\omega}{2}\right)^{\tilde{p}} \tilde{L}(\cos \omega), \quad (7.165)$$

with  $\varepsilon = 0$  for  $p$  and  $\tilde{p}$  for even values and  $\varepsilon = 1$  for odd values. Let  $q = (p + \tilde{p})/2$ . The perfect reconstruction condition

$$\hat{h}^*(\omega) \hat{h}(\omega) + \hat{h}^*(\omega + \pi) \hat{\tilde{h}}(\omega + \pi) = 2$$

is imposed by writing

$$L(\cos \omega) \tilde{L}(\cos \omega) = P\left(\sin^2 \frac{\omega}{2}\right), \quad (7.166)$$

where the polynomial  $P(y)$  must satisfy for all  $y \in [0, 1]$

$$(1 - y)^q P(y) + y^q P(1 - y) = 1. \quad (7.167)$$

We saw in (7.96) that the polynomial of minimum degree satisfying this equation is

$$P(y) = \sum_{k=0}^{q-1} \binom{q-1+k}{k} y^k. \quad (7.168)$$

The spectral factorization (7.166) is solved with a root attribution similar to (7.98). The resulting minimum support of  $\psi$  and  $\tilde{\psi}$  specified by (7.160) is then  $p + \tilde{p} - 1$ . ■

### Spline Biorthogonal Wavelets

Let us choose

$$\hat{h}(\omega) = \sqrt{2} \exp\left(\frac{-i\varepsilon\omega}{2}\right) \left(\cos\frac{\omega}{2}\right)^p \quad (7.169)$$

with  $\varepsilon = 0$  for  $p$  even and  $\varepsilon = 1$  for  $p$  odd. The scaling function computed with (7.148) is then a box spline of degree  $p - 1$ :

$$\hat{\phi}(\omega) = \exp\left(\frac{-i\varepsilon\omega}{2}\right) \left(\frac{\sin(\omega/2)}{\omega/2}\right)^p.$$

Since  $\psi$  is a linear combination of box splines  $\phi(2t - n)$ , it is a compactly supported polynomial spline of the same degree.

The number of vanishing moments  $\tilde{p}$  of  $\psi$  is a free parameter, which must have the same parity as  $p$ . Let  $q = (p + \tilde{p})/2$ . The biorthogonal filter  $\tilde{h}$  of minimum length is obtained by observing that  $L(\cos \omega) = 1$  in (7.164). Thus, the factorization (7.166) and (7.168) imply that

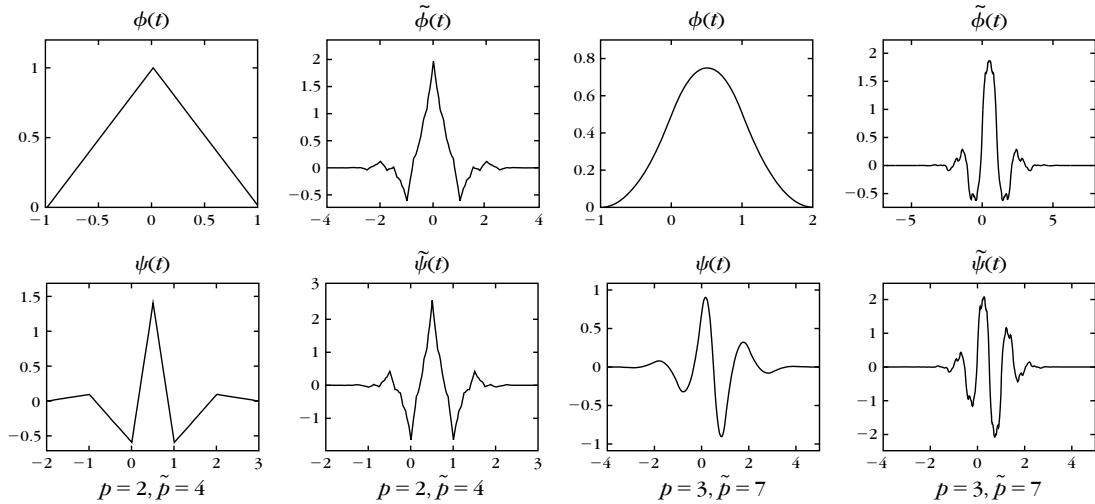
$$\hat{\tilde{h}}(\omega) = \sqrt{2} \exp\left(\frac{-i\varepsilon\omega}{2}\right) \left(\cos\frac{\omega}{2}\right)^{\tilde{p}} \sum_{k=0}^{q-1} \binom{q-1+k}{k} \left(\sin\frac{\omega}{2}\right)^{2k}. \quad (7.170)$$

These filters satisfy the conditions of Theorem 7.14 and therefore generate biorthogonal wavelet bases. Table 7.3 gives the filter coefficients for  $(p=2, \tilde{p}=4)$  and  $(p=3, \tilde{p}=7)$ ; see Figure 7.14 for the resulting dual wavelet and scaling functions.

**Table 7.3** Perfect Reconstruction Filters  $h$  and  $\tilde{h}$  for Compactly Supported Spline Wavelets

$n$	$p, \tilde{p}$	$h[n]$	$\tilde{h}[n]$
0		0.70710678118655	0.99436891104358
1, -1	$p=2$	0.35355339059327	0.41984465132951
2, -2	$\tilde{p}=4$		-0.17677669529664
3, -3			-0.06629126073624
4, -4			0.03314563036812
0, 1		0.53033008588991	0.95164212189718
-1, 2	$p=3$	0.17677669529664	-0.02649924094535
-2, 3	$\tilde{p}=7$		-0.30115912592284
-3, 4			0.03133297870736
-4, 5			0.07466398507402
-5, 6			-0.01683176542131
-6, 7			-0.00906325830378
-7, 8			0.00302108610126

Note:  $\hat{h}$  and  $\hat{\tilde{h}}$  have, respectively,  $\tilde{p}$  and  $p$  zeros at  $\omega = \pi$ .

**FIGURE 7.14**

Spline biorthogonal wavelets and scaling functions of compact support corresponding to Table 7.3 filters.

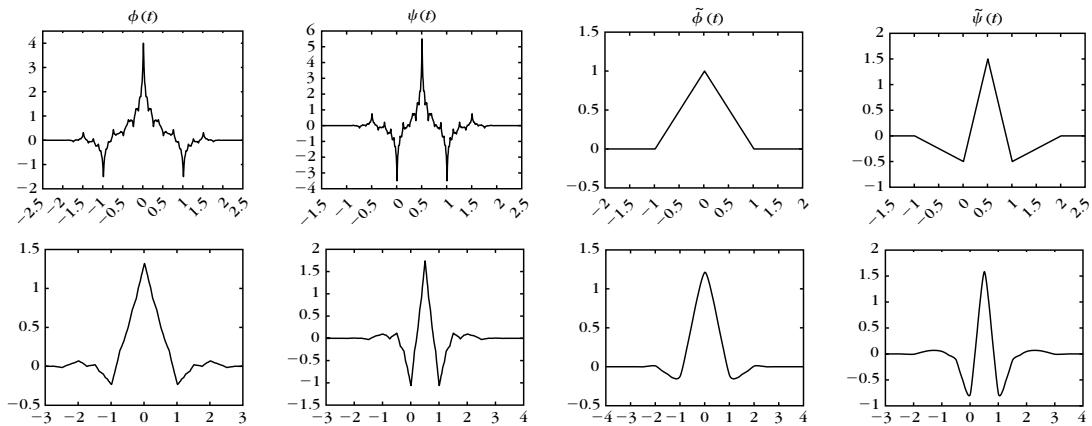
### Closer Filter Length

Biorthogonal filters  $h$  and  $\tilde{h}$  of more similar length are obtained by factoring the polynomial  $P(\sin^2 \frac{\omega}{2})$  in (7.166) with two polynomial  $L(\cos \omega)$  and  $\tilde{L}(\cos \omega)$  of similar degree. There is a limited number of possible factorizations. For  $q = (p + \tilde{p})/2 < 4$ , the only solution is  $L(\cos \omega) = 1$ . For  $q = 4$  there is one nontrivial factorization, and for  $q = 5$  there are two. Table 7.4 gives the resulting coefficients of the filters  $h$  and  $\tilde{h}$  of most similar length, computed by Cohen, Daubechies, and Feauveau [172]. These filters also satisfy the conditions of Theorem 7.14 and therefore define biorthogonal wavelet bases.

Figure 7.15 gives the scaling functions and wavelets for  $p = \tilde{p} = 2$  and  $p = \tilde{p} = 4$ , which correspond to filter sizes  $5/3$  and  $9/7$ , respectively. For  $p = \tilde{p} = 4$ ,  $\phi, \psi$  are similar to  $\tilde{\phi}, \tilde{\psi}$ , which indicates that this basis is nearly orthogonal. This particular set of filters is often used in image compression and recommended for JPEG-2000. The quasi-orthogonality guarantees a good numerical stability and the symmetry allows one to use the folding procedure of Section 7.5.2 at the boundaries. There are also enough vanishing moments to create small wavelet coefficients in regular image domains. Section 7.8.5 describes their lifting implementation, which is simple and efficient. Filter sizes  $5/3$  are also recommended for lossless compression with JPEG-2000, because they use integer operations with a lifting algorithm. The design of other compactly supported biorthogonal filters is discussed extensively in [172, 473].

**Table 7.4** Perfect Reconstruction Filters of Most Similar Length

$p, \tilde{p}$	$n$	$h[n]$	$\tilde{h}[n]$
$p = 2$ $\tilde{p} = 2$	0	1.06066017177982	0.70710678118655
	-1, 1	0.35355339059327	0.35355339059327
	-2, 2	-0.17677669529664	0
$p = 4$ $\tilde{p} = 4$	0	0.85269867900889	0.78848561640637
	-1, 1	0.37740285561283	0.41809227322204
	-2, 2	-0.11062440441844	-0.04068941760920
	-3, 3	-0.02384946501956	-0.06453888262876
	-4, 4	0.03782845554969	0
$p = 5$ $\tilde{p} = 5$	0	0.89950610974865	0.73666018142821
	-1, 1	0.47680326579848	0.34560528195603
	-2, 2	-0.09350469740094	-0.05446378846824
	-3, 3	-0.13670658466433	0.00794810863724
	-4, 4	-0.00269496688011	0.03968708834741
	-5, 5	0.01345670945912	0
$p = 5$ $\tilde{p} = 5$	0	0.54113273169141	1.32702528570780
	-1, 1	0.34335173921766	0.47198693379091
	-2, 2	0.06115645341349	-0.36378609009851
	-3, 3	0.00027989343090	-0.11843354319764
	-4, 4	0.02183057133337	0.05382683783789
	-5, 5	0.00992177208685	0

**FIGURE 7.15**

Biorthonormal wavelets and scaling functions calculated with the filters of Table 7.4, with  $p = 2$  and  $\tilde{p} = 2$  (top row) and  $p = 4$  and  $\tilde{p} = 4$  (bottom row).

---

## 7.5 WAVELET BASES ON AN INTERVAL

To decompose signals  $f$  defined over an interval  $[0, 1]$ , it is necessary to construct wavelet bases of  $L^2[0, 1]$ . Such bases are synthesized by modifying the wavelets  $\psi_{j,n}(t) = 2^{-j/2}\psi(2^{-j}t - n)$  of a basis  $\{\psi_{j,n}\}_{(j,n)\in\mathbb{Z}^2}$  of  $L^2(\mathbb{R})$ . *Inside* wavelets  $\psi_{j,n}$ , have a support included in  $[0, 1]$ , and are not modified. *Boundary* wavelets  $\psi_{j,n}$ , have a support that overlaps  $t = 0$  or  $t = 1$ , and are transformed into functions having a support in  $[0, 1]$ , which are designed in order to provide the necessary complement to generate a basis of  $L^2[0, 1]$ . If  $\psi$  has a compact support, then there is a constant number of boundary wavelets at each scale.

The main difficulty is to construct boundary wavelets that keep their vanishing moments. The next three sections describe different approaches to constructing boundary wavelets. Periodic wavelets have no vanishing moments at the boundary, whereas folded wavelets have one vanishing moment. The custom-designed boundary wavelets of Section 7.5.3 have as many vanishing moments as the inside wavelets but are more complicated to construct. Scaling functions  $\phi_{j,n}$  are also restricted to  $[0, 1]$  by modifying the scaling functions  $\phi_{j,n}(t) = 2^{-j/2}\phi(2^{-j}t - n)$  associated with the wavelets  $\psi_{j,n}$ . The resulting wavelet basis of  $L^2[0, 1]$  is composed of  $2^{-J}$  scaling functions at a coarse scale  $2^J < 1$ , plus  $2^{-j}$  wavelets at each scale  $2^j \leq 2^J$ :

$$\left[ \{\phi_{J,n}^{\text{int}}\}_{0 \leq n < 2^{-J}}, \{\psi_{j,n}^{\text{int}}\}_{-\infty < j \leq J, 0 \leq n < 2^{-j}} \right]. \quad (7.171)$$

On any interval  $[a, b]$ , a wavelet orthonormal basis of  $L^2[a, b]$  is constructed with a dilation by  $b - a$  and a translation by  $a$  of the wavelets in (7.171).

### *Discrete Basis of $\mathbb{C}^N$*

The decomposition of a signal in a wavelet basis over an interval is computed by modifying the fast wavelet transform algorithm of Section 7.3.1. A discrete signal  $b[n]$  of  $N$  samples is associated to the approximation of a signal  $f \in L^2[0, 1]$  at a scale  $N^{-1} = 2^L$  with (7.111):

$$N^{-1/2} b[n] = a_L[n] = \langle f, \phi_{L,n}^{\text{int}} \rangle \quad \text{for } 0 \leq n < 2^{-L}.$$

Its wavelet coefficients can be calculated at scales  $1 \geq 2^j > 2^L$ . We set

$$a_j[n] = \langle f, \phi_{j,n}^{\text{int}} \rangle \quad \text{and} \quad d_j[n] = \langle f, \psi_{j,n}^{\text{int}} \rangle \quad \text{for } 0 \leq n < 2^{-j}. \quad (7.172)$$

The wavelets and scaling functions with support inside  $[0, 1]$  are identical to the wavelets and scaling functions of a basis of  $L^2(\mathbb{R})$ . Thus, the corresponding coefficients  $a_j[n]$  and  $d_j[n]$  can be calculated with the decomposition and reconstruction equations given by Theorem 7.10. However, these convolution formulas must be modified near the boundary where the wavelets and scaling functions are modified. Boundary calculations depend on the specific design of the boundary wavelets, as explained in the next three sections. The resulting filter bank algorithm still computes the  $N$  coefficients of the wavelet representation  $[a_J, \{d_j\}_{L \leq j \leq J}]$  of  $a_L$  with  $O(N)$  operations.

Wavelet coefficients can also be written as discrete inner products of  $a_L$  with discrete wavelets:

$$a_j[n] = \langle a_L[m], \phi_{j,n}^{\text{int}}[m] \rangle \quad \text{and} \quad d_j[n] = \langle a_L[m], \psi_{j,n}^{\text{int}}[m] \rangle. \quad (7.173)$$

As in Section 7.3.3, we verify that

$$\left[ \{\phi_{j,n}^{\text{int}}[m]\}_{0 \leq n < 2^{-j}}, \{\psi_{j,n}^{\text{int}}[m]\}_{L \leq j \leq J, 0 \leq n < 2^{-j}} \right]$$

is an orthonormal basis of  $\mathbb{C}^N$ .

### 7.5.1 Periodic Wavelets

A wavelet basis  $\{\psi_{j,n}\}_{(j,n) \in \mathbb{Z}^2}$  of  $\mathbf{L}^2(\mathbb{R})$  is transformed into a wavelet basis of  $\mathbf{L}^2[0, 1]$  by periodizing each  $\psi_{j,n}$ . The periodization of  $f \in \mathbf{L}^2(\mathbb{R})$  over  $[0, 1]$  is defined by

$$f^{\text{pér}}(t) = \sum_{k=-\infty}^{+\infty} f(t+k). \quad (7.174)$$

The resulting periodic wavelets are

$$\psi_{j,n}^{\text{pér}}(t) = \frac{1}{\sqrt{2^j}} \sum_{k=-\infty}^{+\infty} \psi\left(\frac{t - 2^j n + k}{2^j}\right).$$

For  $j \leq 0$ , there are  $2^{-j}$  different  $\psi_{j,n}^{\text{pér}}$  indexed by  $0 \leq n < 2^{-j}$ . If the support of  $\psi_{j,n}$  is included in  $[0, 1]$ , then  $\psi_{j,n}^{\text{pér}}(t) = \psi_{j,n}(t)$  for  $t \in [0, 1]$ . Thus, the restriction to  $[0, 1]$  of this periodization modifies only the boundary wavelets with a support that overlaps  $t = 0$  or  $t = 1$ .

As indicated in Figure 7.16, such wavelets are transformed into boundary wavelets that have two disjoint components near  $t = 0$  and  $t = 1$ . Taken separately, the components near  $t = 0$  and  $t = 1$  of these boundary wavelets have no vanishing moments, and thus create large signal coefficients, as we shall see later. Theorem 7.16 proves that periodic wavelets together with periodized scaling functions  $\phi_{j,n}^{\text{pér}}$  generate an orthogonal basis of  $\mathbf{L}^2[0, 1]$ .



FIGURE 7.16

The restriction to  $[0, 1]$  of a periodic wavelet  $\psi_{j,n}^{\text{pér}}$  has two disjoint components near  $t = 0$  and  $t = 1$ .

**Theorem 7.16.** For any  $J \leq 0$ ,

$$\left[ \{\psi_{j,n}^{\text{pér}}\}_{-\infty < j \leq J, 0 \leq n < 2^{-j}}, \{\phi_{j,n}^{\text{pér}}\}_{0 \leq n < 2^{-J}} \right] \quad (7.175)$$

is an orthogonal basis of  $\mathbf{L}^2[0, 1]$ .

**Proof.** The orthogonality of this family is proved with Lemma 7.2.

**Lemma 7.2.** Let  $\alpha(t), \beta(t) \in L^2(\mathbb{R})$ . If  $\langle \alpha(t), \beta(t+k) \rangle = 0$  for all  $k \in \mathbb{Z}$ , then

$$\int_0^1 \alpha^{\text{pér}}(t) \beta^{\text{pér}}(t) dt = 0. \quad (7.176)$$

To verify (7.176) we insert the definition (7.174) of periodized functions:

$$\begin{aligned} \int_0^1 \alpha^{\text{pér}}(t) \beta^{\text{pér}}(t) dt &= \int_{-\infty}^{+\infty} \alpha(t) \beta^{\text{pér}}(t) dt \\ &= \sum_{k=-\infty}^{+\infty} \int_{-\infty}^{+\infty} \alpha(t) \beta(t+k) dt = 0. \end{aligned}$$

Since  $[\{\psi_{j,n}\}_{-\infty < j \leq J, n \in \mathbb{Z}}, \{\phi_{J,n}\}_{n \in \mathbb{Z}}]$  is orthogonal in  $L^2(\mathbb{R})$ , we can verify that any two different wavelets or scaling functions  $\alpha^{\text{pér}}$  and  $\beta^{\text{pér}}$  in (7.175) have necessarily a nonperiodized version that satisfies  $\langle \alpha(t), \beta(t+k) \rangle = 0$  for all  $k \in \mathbb{Z}$ . Thus, this lemma proves that (7.175) is orthogonal in  $L^2[0, 1]$ .

To prove that this family generates  $L^2[0, 1]$ , we extend  $f \in L^2[0, 1]$  with zeros outside  $[0, 1]$  and decompose it in the wavelet basis of  $L^2(\mathbb{R})$ :

$$f = \sum_{j=-\infty}^J \sum_{n=-\infty}^{+\infty} \langle f, \psi_{j,n} \rangle \psi_{j,n} + \sum_{n=-\infty}^{+\infty} \langle f, \phi_{J,n} \rangle \phi_{J,n}. \quad (7.177)$$

This zero extension is periodized with the sum (7.174), which defines  $f^{\text{pér}}(t) = f(t)$  for  $t \in [0, 1]$ . Periodizing (7.177) proves that  $f$  can be decomposed over the periodized wavelet family (7.175) in  $L^2[0, 1]$ . ■

Theorem 7.16 shows that periodizing a wavelet orthogonal basis of  $L^2(\mathbb{R})$  defines a wavelet orthogonal basis of  $L^2[0, 1]$ . If  $J = 0$ , then there is a single scaling function, and one can verify that  $\phi_{0,0}(t) = 1$ . The resulting scaling coefficient  $\langle f, \phi_{0,0} \rangle$  is the average of  $f$  over  $[0, 1]$ .

Periodic wavelet bases have the disadvantage of creating high-amplitude wavelet coefficients in the neighborhood of  $t = 0$  and  $t = 1$ , because the boundary wavelets have separate components with no vanishing moments. If  $f(0) \neq f(1)$ , the wavelet coefficients behave as if the signal were discontinuous at the boundaries. This can also be verified by extending  $f \in L^2[0, 1]$  into an infinite 1 periodic signal  $f^{\text{pér}}$  and by showing that

$$\int_0^1 f(t) \psi_{j,n}^{\text{pér}}(t) dt = \int_{-\infty}^{+\infty} f^{\text{pér}}(t) \psi_{j,n}(t) dt. \quad (7.178)$$

If  $f(0) \neq f(1)$ , then  $f^{\text{pér}}(t)$  is discontinuous at  $t = 0$  and  $t = 1$ , which creates high-amplitude wavelet coefficients when  $\psi_{j,n}$  overlaps the interval boundaries.

### Periodic Discrete Transform

For  $f \in L^2[0, 1]$  let us consider

$$a_j[n] = \langle f, \phi_{j,n}^{\text{pér}} \rangle \quad \text{and} \quad d_j[n] = \langle f, \psi_{j,n}^{\text{pér}} \rangle.$$

We verify as in (7.178) that these inner products are equal to the coefficients of a periodic signal decomposed in a nonperiodic wavelet basis:

$$a_j[n] = \langle f^{\text{pér}}, \phi_{j,n} \rangle \quad \text{and} \quad d_j[n] = \langle f^{\text{pér}}, \psi_{j,n} \rangle.$$

Thus, the convolution formulas of Theorem 7.10 apply if we take into account the periodicity of  $f^{\text{pér}}$ . This means that  $a_j[n]$  and  $d_j[n]$  are considered as discrete signals of period  $2^{-j}$ , and all convolutions in (7.102–7.104) must therefore be replaced by circular convolutions. Despite the poor behavior of periodic wavelets near the boundaries, they are often used because the numerical implementation is particularly simple.

### 7.5.2 Folded Wavelets

Decomposing  $f \in L^2[0, 1]$  in a periodic wavelet basis was shown in (7.178) to be equivalent to a decomposition of  $f^{\text{pér}}$  in a regular basis of  $L^2(\mathbb{R})$ . Let us extend  $f$  with zeros outside  $[0, 1]$ . To avoid creating discontinuities with such a periodization, the signal is folded with respect to  $t = 0$ :  $f_0(t) = f(t) + f(-t)$ . The support of  $f_0$  is  $[-1, 1]$  and it is transformed into a 2 periodic signal, as illustrated in Figure 7.17:

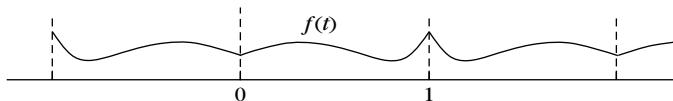
$$f^{\text{repl}}(t) = \sum_{k=-\infty}^{+\infty} f_0(t - 2k) = \sum_{k=-\infty}^{+\infty} f(t - 2k) + \sum_{k=-\infty}^{+\infty} f(2k - t). \quad (7.179)$$

Clearly  $f^{\text{repl}}(t) = f(t)$  if  $t \in [0, 1]$ , and it is symmetric with respect to  $t = 0$  and  $t = 1$ . If  $f$  is continuously differentiable, then  $f^{\text{repl}}$  is continuous at  $t = 0$  and  $t = 1$ , but its derivative is discontinuous at  $t = 0$  and  $t = 1$  if  $f'(0) \neq 0$  and  $f'(1) \neq 0$ .

Decomposing  $f^{\text{repl}}$  in a wavelet basis  $\{\psi_{j,n}\}_{(j,n) \in \mathbb{Z}^2}$  is equivalent to decomposing  $f$  on a folded wavelet basis. Let  $\psi_{j,n}^{\text{repl}}$  be the folding of  $\psi_{j,n}$  with the summation (7.179). One can verify that

$$\int_0^1 f(t) \psi_{j,n}^{\text{repl}}(t) dt = \int_{-\infty}^{+\infty} f^{\text{repl}}(t) \psi_{j,n}(t) dt. \quad (7.180)$$

Suppose that  $f$  is regular over  $[0, 1]$ . Then  $f^{\text{repl}}$  is continuous at  $t = 0, 1$  and produces smaller boundary wavelet coefficients than  $f^{\text{pér}}$ . However, it is not continuously differentiable at  $t = 0, 1$ , which creates bigger wavelet coefficients at the boundary than inside.



**FIGURE 7.17**

The folded signal  $f^{\text{repl}}(t)$  is 2 periodic, symmetric about  $t = 0$  and  $t = 1$ , and equal to  $f(t)$  on  $[0, 1]$ .

To construct a basis of  $L^2[0, 1]$  with the folded wavelets  $\psi_{j,n}^{\text{repl}}$ , it is sufficient for  $\psi(t)$  to be either symmetric or antisymmetric with respect to  $t = 1/2$ . The Haar wavelet is the only real compactly supported wavelet that is symmetric or antisymmetric and that generates an orthogonal basis of  $L^2(\mathbb{R})$ . On the other hand, if we loosen up the orthogonality constraint, Section 7.4 proves that there exist biorthogonal bases constructed with compactly supported wavelets that are either symmetric or antisymmetric. Let  $\{\psi_{j,n}\}_{(j,n) \in \mathbb{Z}^2}$  and  $\{\tilde{\psi}_{j,n}\}_{(j,n) \in \mathbb{Z}^2}$  be such biorthogonal wavelet bases. If we fold the wavelets as well as the scaling functions, then for  $J \leq 0$ ,

$$\left[ \{\psi_{j,n}^{\text{repl}}\}_{-\infty < j \leq J, 0 \leq n < 2^{-j}}, \{\phi_{j,n}^{\text{repl}}\}_{0 \leq n < 2^{-J}} \right] \quad (7.181)$$

is a Riesz basis of  $L^2[0, 1]$  [174]. The biorthogonal basis is obtained by folding the dual wavelets  $\tilde{\psi}_{j,n}$  and is given by

$$\left[ \{\tilde{\psi}_{j,n}^{\text{repl}}\}_{-\infty < j \leq J, 0 \leq n < 2^{-j}}, \{\tilde{\phi}_{j,n}^{\text{repl}}\}_{0 \leq n < 2^{-J}} \right]. \quad (7.182)$$

If  $J = 0$ , then  $\phi_{0,0}^{\text{repl}} = \tilde{\phi}_{0,0}^{\text{repl}} = 1$ .

Biorthogonal wavelets of compact support are characterized by a pair of finite perfect reconstruction filters  $(h, \tilde{h})$ . The symmetry of these wavelets depends on the symmetry and size of the filters, as explained in Section 7.4.2. A fast folded wavelet transform is implemented with a modified filter bank algorithm, where the treatment of boundaries is slightly more complicated than for periodic wavelets. The symmetric and antisymmetric cases are considered separately.

### Folded Discrete Transform

For  $f \in L^2[0, 1]$ , we consider

$$a_j[n] = \langle f, \phi_{j,n}^{\text{repl}} \rangle \quad \text{and} \quad d_j[n] = \langle f, \psi_{j,n}^{\text{repl}} \rangle.$$

We verify as in (7.180) that these inner products are equal to the coefficients of a folded signal decomposed in a nonfolded wavelet basis:

$$a_j[n] = \langle f^{\text{repl}}, \phi_{j,n} \rangle \quad \text{and} \quad d_j[n] = \langle f^{\text{repl}}, \psi_{j,n} \rangle.$$

The convolution formulas of Theorem 7.10 apply if we take into account the symmetry and periodicity of  $f^{\text{repl}}$ . The symmetry properties of  $\phi$  and  $\psi$  imply that  $a_j[n]$  and  $d_j[n]$  also have symmetry and periodicity properties, which must be taken into account in the calculations of (7.102–7.104).

Symmetric biorthogonal wavelets are constructed with perfect reconstruction filters  $h$  and  $\tilde{h}$  of odd size that are symmetric about  $n = 0$ . Then  $\phi$  is symmetric about 0, whereas  $\psi$  is symmetric about  $1/2$ . As a result, one can verify that  $a_j[n]$  is  $2^{-j+1}$  periodic and symmetric about  $n = 0$  and  $n = 2^{-j}$ . Thus, it is characterized by  $2^{-j+1}$  samples for  $0 \leq n < 2^{-j}$ . The situation is different for  $d_j[n]$ , which is  $2^{-j+1}$  periodic but symmetric with respect to  $-1/2$  and  $2^{-j} - 1/2$ . It is characterized by  $2^{-j}$  samples for  $0 \leq n < 2^{-j}$ .

To initialize this algorithm, the original signal  $a_L[n]$  defined over  $0 \leq n < N - 1$  must be extended by one sample at  $n = N$ , and considered to be symmetric with respect to  $n = 0$  and  $n = N$ . The extension is done by setting  $a_L[N] = a_L[N - 1]$ . For any  $J < L$ , the resulting discrete wavelet representation  $\{d_j\}_{L < j \leq J}, a_J$  is characterized by  $N + 1$  coefficients. To avoid adding one more coefficient, one can modify symmetry at the right boundary of  $a_L$  by considering that it is symmetric with respect to  $N - 1/2$  instead of  $N$ . The symmetry of the resulting  $a_j$  and  $d_j$  at the right boundary is modified accordingly by studying the properties of the convolution formula (7.157). As a result, these signals are characterized by  $2^{-j}$  samples and the wavelet representation has  $N$  coefficients. A simpler implementation of this folding technique is given with a lifting in Section 7.8.5. This folding approach is used in most applications because it leads to simpler data structures that keep the number of coefficients constant. However, the discrete coefficients near the right boundary cannot be written as inner products of some function  $f(t)$  with dilated boundary wavelets.

Antisymmetric biorthogonal wavelets are obtained with perfect reconstruction filters  $h$  and  $\hat{h}$  of even size that are symmetric about  $n = 1/2$ . In this case,  $\phi$  is symmetric about  $1/2$  and  $\psi$  is antisymmetric about  $1/2$ . As a result,  $a_j$  and  $d_j$  are  $2^{-j+1}$  periodic and, respectively, symmetric and antisymmetric about  $-1/2$  and  $2^{-j} - 1/2$ . They are both characterized by  $2^{-j}$  samples for  $0 \leq n < 2^{-j}$ . The algorithm is initialized by considering that  $a_L[n]$  is symmetric with respect to  $-1/2$  and  $N - 1/2$ . There is no need to add another sample. The resulting discrete wavelet representation  $\{d_j\}_{L < j \leq J}, a_J$  is characterized by  $N$  coefficients.

### 7.5.3 Boundary Wavelets

Wavelet coefficients are small in regions where the signal is regular only if the wavelets have enough vanishing moments. The restriction of periodic and folded “boundary” wavelets to the neighborhood of  $t = 0$  and  $t = 1$  have, respectively, 0 and 1 vanishing moments. Therefore, these boundary wavelets cannot fully take advantage of the signal regularity. They produce large inner products, as if the signal were discontinuous or had a discontinuous derivative. To avoid creating large-amplitude wavelet coefficients at the boundaries, one must synthesize boundary wavelets that have as many vanishing moments as the original wavelet  $\psi$ . Initially introduced by Meyer, this approach has been refined by Cohen, Daubechies, and Vial [174]. The main results are given without proofs.

#### **Multiresolution of $L^2[0, 1]$**

A wavelet basis of  $L^2[0, 1]$  is constructed with a multiresolution approximation  $\{\mathbf{V}_j^{\text{int}}\}_{-\infty < j \leq 0}$ . A wavelet has  $p$  vanishing moments if it is orthogonal to all polynomials of degree  $p - 1$  or smaller. Since wavelets at a scale  $2^j$  are orthogonal to functions in  $\mathbf{V}_j^{\text{int}}$ , to guarantee that they have  $p$  vanishing moments we make sure that polynomials of degree  $p - 1$  are inside  $\mathbf{V}_j^{\text{int}}$ .

We define an approximation space  $\mathbf{V}_j^{\text{int}} \subset \mathbf{L}^2[0, 1]$  with a compactly supported Daubechies scaling function  $\phi$  associated to a wavelet with  $p$  vanishing moments. Theorem 7.7 proves that the support of  $\phi$  has size  $2p - 1$ . We translate  $\phi$  so that its support is  $[-p + 1, p]$ . At a scale  $2^j \leq (2p)^{-1}$ , there are  $2^{-j} - 2p$  scaling functions with a support inside  $[0, 1]$ :

$$\phi_{j,n}^{\text{int}}(t) = \phi_{j,n}(t) = \frac{1}{\sqrt{2^j}} \phi\left(\frac{t - 2^j n}{2^j}\right) \quad \text{for } p \leq n < 2^{-j} - p.$$

To construct an approximation space  $\mathbf{V}_j^{\text{int}}$  of dimension  $2^{-j}$  we add  $p$  scaling functions with a support on the left boundary near  $t = 0$ :

$$\phi_{j,n}^{\text{int}}(t) = \frac{1}{\sqrt{2^j}} \phi_n^{\text{left}}\left(\frac{t}{2^j}\right) \quad \text{for } 0 \leq n < p,$$

and  $p$  scaling functions on the right boundary near  $t = 1$ :

$$\phi_{j,n}^{\text{int}}(t) = \frac{1}{\sqrt{2^j}} \phi_{2^{-j}-1-n}^{\text{right}}\left(\frac{t-1}{2^j}\right) \quad \text{for } 2^{-j} - p \leq n < 2^{-j}.$$

Theorem 7.17 constructs appropriate boundary scaling functions  $\{\phi_n^{\text{left}}\}_{0 \leq n < p}$  and  $\{\phi_n^{\text{right}}\}_{0 \leq n < p}$ .

**Theorem 7.17:** *Cohen, Daubechies, Vial.* One can construct boundary scaling functions  $\phi_n^{\text{left}}$  and  $\phi_n^{\text{right}}$  so that if  $2^{-j} \geq 2p$ , then  $\{\phi_{j,n}^{\text{int}}\}_{0 \leq n < 2^{-j}}$  is an orthonormal basis of a space  $\mathbf{V}_j^{\text{int}}$  satisfying

$$\begin{aligned} \mathbf{V}_j^{\text{int}} &\subset \mathbf{V}_{j-1}^{\text{int}} \\ \lim_{j \rightarrow -\infty} \mathbf{V}_j^{\text{int}} &= \text{Closure} \left( \bigcup_{j=-\infty}^{-\log_2(2p)} \mathbf{V}_j^{\text{int}} \right) = \mathbf{L}^2[0, 1], \end{aligned}$$

and the restrictions to  $[0, 1]$  of polynomials of degree  $p - 1$  are in  $\mathbf{V}_j^{\text{int}}$ .

**Proof.** A sketch of the proof is given. All details are in [174]. Since the wavelet  $\psi$  corresponding to  $\phi$  has  $p$  vanishing moments, the Fix-Strang condition (7.70) implies that

$$q_k(t) = \sum_{n=-\infty}^{+\infty} n^k \phi(t-n) \tag{7.183}$$

is a polynomial of degree  $k$ . At any scale  $2^j$ ,  $q_k(2^{-j}t)$  is still a polynomial of degree  $k$ , and for  $0 \leq k < p$  this family defines a basis of polynomials of degree  $p - 1$ . To guarantee that polynomials of degree  $p - 1$  are in  $\mathbf{V}_j^{\text{int}}$  we impose that the restriction of  $q_k(2^{-j}t)$  to  $[0, 1]$  can be decomposed in the basis of  $\mathbf{V}_j^{\text{int}}$ :

$$\begin{aligned} q_k(2^{-j}t) \mathbf{1}_{[0,1]}(t) &= \sum_{n=0}^{p-1} a[n] \phi_n^{\text{left}}(2^{-j}t) + \sum_{n=p}^{2^{-j}p-1} n^k \phi(2^{-j}t - n) \\ &\quad + \sum_{n=0}^{p-1} b[n] \phi_n^{\text{right}}(2^{-j}t - 2^{-j}). \end{aligned} \tag{7.184}$$

Since the support of  $\phi$  is  $[-p+1, p]$ , the condition (7.184) together with (7.183) can be separated into two nonoverlapping left and right conditions. With a change of variable, we verify that (7.184) is equivalent to

$$\sum_{n=-p+1}^p n^k \phi(t-n) \mathbf{1}_{[0,+\infty)}(t) = \sum_{n=0}^{p-1} a[n] \phi_n^{\text{left}}(t), \quad (7.185)$$

and

$$\sum_{n=-p}^{p-1} n^k \phi(t-n) \mathbf{1}_{(-\infty,0]}(t) = \sum_{n=0}^{p-1} b[n] \phi_n^{\text{right}}(t). \quad (7.186)$$

The embedding property  $\mathbf{V}_j^{\text{int}} \subset \mathbf{V}_{j-1}^{\text{int}}$  is obtained by imposing that the boundary scaling functions satisfy scaling equations. We suppose that  $\phi_n^{\text{left}}$  has a support  $[0, p+n]$  and satisfies a scaling equation of the form

$$2^{-1/2} \phi_n^{\text{left}}(2^{-1}t) = \sum_{l=0}^{p-1} H_{n,l}^{\text{left}} \phi_l^{\text{left}}(t) + \sum_{m=p}^{p+2n} h_{n,m}^{\text{left}} \phi(t-m), \quad (7.187)$$

whereas  $\phi_n^{\text{right}}$  has a support  $[-p-n, 0]$  and satisfies a similar scaling equation on the right. The constants  $H_{n,l}^{\text{left}}$ ,  $h_{n,m}^{\text{left}}$ ,  $H_{n,l}^{\text{right}}$ , and  $h_{n,m}^{\text{right}}$  are adjusted to verify the polynomial reproduction equations (7.185) and (7.186), while producing orthogonal scaling functions. The resulting family  $\{\phi_{j,n}^{\text{int}}\}_{n \leqslant n < 2^{-j}}$  is an orthonormal basis of a space  $\mathbf{V}_j^{\text{int}}$ .

The convergence of the spaces  $\mathbf{V}_j^{\text{int}}$  to  $\mathbf{L}^2[0, 1]$  when  $2^j$  goes to 0 is a consequence of the fact that the multiresolution spaces  $\mathbf{V}_j$  generated by the Daubechies scaling function  $\{\phi_{j,n}\}_{n \in \mathbb{Z}}$  converge to  $\mathbf{L}^2(\mathbb{R})$ . ■

The proof constructs the scaling functions through scaling equations specified by discrete filters. At the boundaries, the filter coefficients are adjusted to construct orthogonal scaling functions with a support in  $[0, 1]$ , and to guarantee that polynomials of degree  $p-1$  are reproduced by these scaling functions. Table 7.5 gives the filter coefficients for  $p=2$ .

### Wavelet Basis of $\mathbf{L}^2[0, 1]$

Let  $\mathbf{W}_j^{\text{int}}$  be the orthogonal complement of  $\mathbf{V}_j^{\text{int}}$  in  $\mathbf{V}_{j-1}^{\text{int}}$ . The support of the Daubechies wavelet  $\psi$  with  $p$  vanishing moments is  $[-p+1, p]$ . Since  $\psi_{j,n}$  is orthogonal to any  $\phi_{j,l}$ , we verify that an orthogonal basis of  $\mathbf{W}_j^{\text{int}}$  can be constructed with the  $2^{-j} - 2p$  inside wavelets with support in  $[0, 1]$ :

$$\psi_{j,n}^{\text{int}}(t) = \psi_{j,n}(t) = \frac{1}{\sqrt{2^j}} \psi\left(\frac{t-2^j n}{2^j}\right) \quad \text{for } p \leqslant n < 2^{-j} - p,$$

to which are added  $2p$  left and right boundary wavelets

$$\psi_{j,n}^{\text{int}}(t) = \frac{1}{\sqrt{2^j}} \psi_n^{\text{left}}\left(\frac{t}{2^j}\right) \quad \text{for } 0 \leqslant n < p,$$

**Table 7.5** Left and Right Border Coefficients for a Daubechies Wavelet with  $p = 2$  Vanishing Moments

$k$	$l$	$H_{k,l}^{\text{left}}$	$G_{k,l}^{\text{left}}$	$k$	$m$	$h_{k,m}^{\text{left}}$	$g_{k,m}^{\text{left}}$
0	0	0.6033325119	-0.7965436169	0	2	-0.398312997	-0.2587922483
0	1	0.690895531	0.5463927140	1	2	0.8500881025	0.227428117
1	0	0.03751746045	0.01003722456	1	3	0.2238203570	-0.8366028212
1	1	0.4573276599	0.1223510431	1	4	-0.129227434	0.4830129218
$k$	$l$	$H_{k,l}^{\text{right}}$	$G_{k,l}^{\text{right}}$	$k$	$m$	$h_{k,m}^{\text{right}}$	$g_{k,m}^{\text{left}}$
-2	-2	0.1901514184	-0.3639069596	-2	-5	0.4431490496	0.235575950
-2	-1	-0.1942334074	0.3717189665	-2	-4	0.7675566693	0.4010695194
-1	-2	0.434896998	0.8014229620	-2	-3	0.3749553316	-0.7175799994
-2	-1	0.8705087534	-0.2575129195	-1	-3	0.2303890438	-0.5398225007
		$h[-1]$	$h[0]$			$h[1]$	$h[2]$
		0.482962913145	0.836516303738			0.224143868042	-0.129409522551

Note: The inside filter coefficients are at the bottom of the table. A table of coefficients for  $p \geq 2$  vanishing moments can be retrieved over the Internet at the FTP site <ftp://math.princeton.edu/pub/user/ingrid/interval-tables>.

$$\psi_{j,n}^{\text{int}}(t) = \frac{1}{\sqrt{2^j}} \psi_{2^{-j}-1-n}^{\text{right}}\left(\frac{t-1}{2^j}\right) \quad \text{for } 2^{-j}-p \leq n < 2^{-j}.$$

Since  $\mathbf{W}_j^{\text{int}} \subset \mathbf{V}_{j-1}^{\text{int}}$ , the left and right boundary wavelets at any scale  $2^j$  can be expanded into scaling functions at the scale  $2^{j-1}$ . For  $j=1$  we impose that the left boundary wavelets satisfy equations of the form

$$\frac{1}{\sqrt{2}} \psi_n^{\text{left}}\left(\frac{t}{2}\right) = \sum_{l=0}^{p-1} G_{n,l}^{\text{left}} \phi_l^{\text{left}}(t) + \sum_{m=p}^{p+2n} g_{n,m}^{\text{left}} \phi(t-m). \quad (7.188)$$

The right boundary wavelets satisfy similar equations. The coefficients  $G_{n,l}^{\text{left}}, g_{n,m}^{\text{left}}$ ,  $G_{n,l}^{\text{right}}$ , and  $g_{n,m}^{\text{right}}$  are computed so that  $\{\psi_{j,n}^{\text{int}}\}_{0 \leq n < 2^{-j}}$  is an orthonormal basis of  $\mathbf{W}_j^{\text{int}}$ . Table 7.5 gives the values of these coefficients for  $p=2$ .

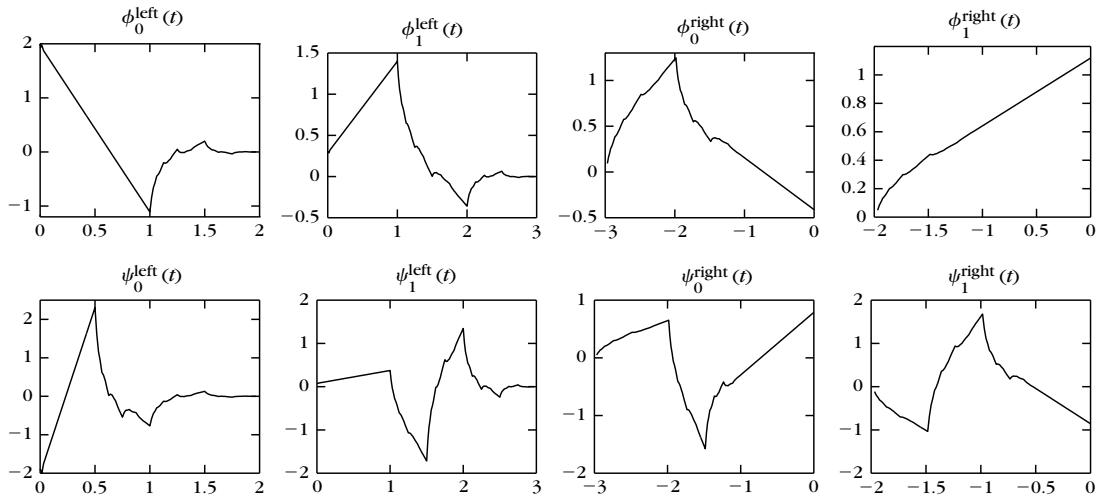
For any  $2^J \leq (2p)^{-1}$  the multiresolution properties prove that

$$\mathbf{L}^2[0, 1] = \mathbf{V}_J^{\text{int}} \oplus_{J=-\infty}^J \mathbf{W}_j^{\text{int}},$$

which implies that

$$\left[ \{\phi_{J,n}^{\text{int}}\}_{0 \leq n < 2^{-J}}, \{\psi_{j,n}^{\text{int}}\}_{-\infty < j \leq J, 0 \leq n < 2^{-j}} \right] \quad (7.189)$$

is an orthonormal wavelet basis of  $\mathbf{L}^2[0, 1]$ . The boundary wavelets, like the inside wavelets, have  $p$  vanishing moments because polynomials of degree  $p-1$  are included in the space  $\mathbf{V}_j^{\text{int}}$ . Figure 7.18 displays the  $2p=4$  boundary scaling functions and wavelets.

**FIGURE 7.18**

Boundary scaling functions and wavelets with  $p = 2$  vanishing moments.

### **Fast Discrete Algorithm**

For any  $f \in \mathbf{L}^2[0, 1]$  we denote

$$a_j[n] = \langle f, \phi_{j,n}^{\text{int}} \rangle \quad \text{and} \quad d_j[n] = \langle f, \psi_{j,n}^{\text{int}} \rangle \quad \text{for } 0 \leq n \leq 2^{-j}.$$

Wavelet coefficients are computed with a cascade of convolutions identical to Theorem 7.10 as long as filters do not overlap signal boundaries. A Daubechies filter  $h$  is considered here to have a support located at  $[-p+1, p]$ . At the boundary, the usual Daubechies filters are replaced by boundary filters that relate boundary wavelets and scaling functions to the finer-scale scaling functions in (7.187) and (7.188).

**Theorem 7.18:** Cohen, Daubechies, Vial. If  $0 \leq k < p$ ,

$$\begin{aligned} a_j[k] &= \sum_{l=0}^{p-1} H_{k,l}^{\text{left}} a_{j-1}[l] + \sum_{m=p}^{p+2k} h_{k,m}^{\text{left}} a_{j-1}[m], \\ d_j[k] &= \sum_{l=0}^{p-1} G_{k,l}^{\text{left}} a_{j-1}[l] + \sum_{m=p}^{p+2k} g_{k,m}^{\text{left}} a_{j-1}[m]. \end{aligned}$$

If  $p \leq k < 2^{-j} - p$ ,

$$\begin{aligned} a_j[k] &= \sum_{l=-\infty}^{+\infty} h[l-2k] a_{j-1}[l], \\ d_j[k] &= \sum_{l=-\infty}^{+\infty} g[l-2k] a_{j-1}[l]. \end{aligned}$$

If  $-p \leq k < 0$ ,

$$a_j[2^{-j} + k] = \sum_{l=-p}^{-1} H_{k,l}^{\text{right}} a_{j-1}[2^{-j+1} + l] + \sum_{m=-p+2k+1}^{-p-1} h_{k,m}^{\text{right}} a_{j-1}[2^{-j+1} + m],$$

$$d_j[2^{-j} + k] = \sum_{l=-p}^{-1} G_{k,l}^{\text{right}} a_{j-1}[2^{-j+1} + l] + \sum_{m=-p+2k+1}^{-p-1} g_{k,m}^{\text{right}} a_{j-1}[2^{-j+1} + m].$$

This cascade algorithm decomposes  $a_L$  into a discrete wavelet transform  $[a_j, \{d_j\}_{L \leq j \leq J}]$  with  $O(N)$  operations. The maximum scale must satisfy  $2^J \leq (2p)^{-1}$ , because the number of boundary coefficients remains equal to  $2p$  at all scales. The implementation is more complicated than the folding and periodic algorithms described in Sections 7.5.1 and 7.5.2, but does not require more computations. The signal  $a_L$  is reconstructed from its wavelet coefficients, by inverting the decomposition formula in Theorem 7.18.

**Theorem 7.19:** Cohen, Daubechies, Vial. If  $0 \leq l \leq p-1$ ,

$$a_{j-1}[l] = \sum_{k=0}^{p-1} H_{k,l}^{\text{left}} a_j[k] + \sum_{k=0}^{p-1} G_{k,l}^{\text{left}} d_j[k].$$

If  $p \leq l \leq 3p-2$ ,

$$a_{j-1}[l] = \sum_{k=(l-p)/2}^{p-1} h_{k,l}^{\text{left}} a_j[k] + \sum_{k=-\infty}^{+\infty} h[l-2k] a_j[k]$$

$$+ \sum_{k=(l-p)/2}^{p-1} g_{k,l}^{\text{left}} d_j[k] + \sum_{k=-\infty}^{+\infty} g[l-2k] d_j[k].$$

If  $3p-1 \leq l \leq 2^{-j+1} - 3p$ ,

$$a_{j-1}[l] = \sum_{k=-\infty}^{+\infty} h[l-2k] a_j[k] + \sum_{k=-\infty}^{+\infty} g[l-2k] d_j[k].$$

If  $-p-1 \geq l \geq -3p+1$ ,

$$a_{j-1}[2^{-j+1} + l] = \sum_{k=-p}^{(l+p-1)/2} h_{k,l}^{\text{right}} a_j[2^{-j} + k] + \sum_{k=-\infty}^{+\infty} h[l-2k] a_j[2^{-j} + k]$$

$$+ \sum_{k=-p}^{(l+p-1)/2} g_{k,l}^{\text{right}} d_j[2^{-j} + k] + \sum_{k=-\infty}^{+\infty} g[l-2k] d_j[2^{-j} + k].$$

If  $-1 \geq l \geq -p$ ,

$$a_{j-1}[2^{-j+1} + l] = \sum_{k=-p}^{-1} H_{k,l}^{\text{right}} a_j[2^{-j} + k] + \sum_{k=-p}^{-1} G_{k,l}^{\text{right}} d_j[2^{-j} + k].$$

The original signal  $a_L$  is reconstructed from the orthogonal wavelet representation  $[a_J, \{d_j\}_{L < j \leq J}]$  by iterating these equations for  $L < j \leq J$ . This reconstruction is performed with  $O(N)$  operations.

## 7.6 MULTISCALE INTERPOLATIONS

Multiresolution approximations are closely connected to the generalized interpolations and sampling theorems studied in Section 3.1.3. Section 7.6.1 constructs general classes of interpolation functions from orthogonal scaling functions and derives new sampling theorems. Interpolation bases have the advantage of easily computing the decomposition coefficients from the sample values of the signal. Section 7.6.2 constructs interpolation wavelet bases.

### 7.6.1 Interpolation and Sampling Theorems

Section 3.1.3 explains that a sampling scheme approximates a signal by its orthogonal projection onto a space  $\mathbf{U}_s$  and samples this projection at intervals  $s$ . The space  $\mathbf{U}_s$  is constructed so that any function in  $\mathbf{U}_s$  can be recovered by interpolating a uniform sampling at intervals  $s$ . We relate the construction of interpolation functions to orthogonal scaling functions and compute the orthogonal projector on  $\mathbf{U}_s$ .

An *interpolation function* any  $\phi$  such that  $\{\phi(t-n)\}_{n \in \mathbb{Z}}$  is a Riesz basis of the space  $\mathbf{U}_1$  it generates, and that satisfies

$$\phi(n) = \begin{cases} 1 & \text{if } n = 0 \\ 0 & \text{if } n \neq 0. \end{cases} \quad (7.190)$$

Any  $f \in \mathbf{U}_1$  is recovered by interpolating its samples  $f(n)$ :

$$f(t) = \sum_{n=-\infty}^{+\infty} f(n) \phi(t-n). \quad (7.191)$$

Indeed, we know that  $f$  is a linear combination of the basis vector  $\{\phi(t-n)\}_{n \in \mathbb{Z}}$  and the interpolation property (7.190) yields (7.191). The Whittaker sampling Theorem 3.2 is based on the interpolation function

$$\phi(t) = \frac{\sin \pi t}{\pi t}.$$

In this case, space  $\mathbf{U}_1$  is the set of functions having a Fourier transform support included in  $[-\pi, \pi]$ .

Scaling an interpolation function yields a new interpolation for a different sampling interval. Let us define  $\phi_s(t) = \phi(t/s)$  and

$$\mathbf{U}_s = \left\{ f \in \mathbf{L}^2(\mathbb{R}) \quad \text{with} \quad f(st) \in \mathbf{U}_1 \right\}.$$

One can verify that any  $f \in \mathbf{U}_s$  can be written as

$$f(t) = \sum_{n=-\infty}^{+\infty} f(ns) \phi_s(t - ns). \quad (7.192)$$

### Scaling Autocorrelation

We denote by  $\phi_o$  an orthogonal scaling function, defined by the fact that  $\{\phi_o(t - n)\}_{n \in \mathbb{Z}}$  is an orthonormal basis of a space  $\mathbf{V}_0$  of a multiresolution approximation. Theorem 7.2 proves that this scaling function is characterized by a conjugate mirror filter  $h_o$ . Theorem 7.20 defines an interpolation function from the autocorrelation of  $\phi_o$  [423].

**Theorem 7.20.** Let  $\bar{\phi}_o(t) = \phi_o(-t)$  and  $\bar{h}_o[n] = h_o[-n]$ . If  $|\hat{\phi}_o(\omega)| = O((1 + |\omega|)^{-1})$ , then

$$\phi(t) = \int_{-\infty}^{+\infty} \phi_o(u) \phi_o(u - t) du = \phi_o \star \bar{\phi}_o(t) \quad (7.193)$$

is an interpolation function. Moreover,

$$\phi\left(\frac{t}{2}\right) = \sum_{n=-\infty}^{+\infty} h[n] \phi(t - n) \quad (7.194)$$

with

$$h[n] = \sum_{m=-\infty}^{+\infty} h_o[m] h_o[m - n] = h_o \star \bar{h}_o[n]. \quad (7.195)$$

**Proof.** Observe first that

$$\phi(n) = \langle \phi_o(t), \phi_o(t - n) \rangle = \delta[n],$$

which proves the interpolation property (7.190). To prove that  $\{\phi(t - n)\}_{n \in \mathbb{Z}}$  is a Riesz basis of the space  $\mathbf{U}_1$  it generates, we verify the condition (7.9). The autocorrelation  $\phi(t) = \phi_o \star \bar{\phi}_o(t)$  has a Fourier transform  $\hat{\phi}(\omega) = |\hat{\phi}_o(\omega)|^2$ . Thus, condition (7.9) means that there exist  $B \geq A > 0$  such that

$$\forall \omega \in [-\pi, \pi], \quad A \leq \sum_{k=-\infty}^{+\infty} |\hat{\phi}_o(\omega - 2k\pi)|^4 \leq B. \quad (7.196)$$

We proved in (7.14) that the orthogonality of a family  $\{\phi_o(t - n)\}_{n \in \mathbb{Z}}$  is equivalent to

$$\forall \omega \in [-\pi, \pi], \quad \sum_{k=-\infty}^{+\infty} |\hat{\phi}_o(\omega + 2k\pi)|^2 = 1. \quad (7.197)$$

Therefore, the right inequality of (7.196) is valid for  $A = 1$ . Let us prove the left inequality. Since  $|\hat{\phi}_o(\omega)| = O((1 + |\omega|)^{-1})$ , one can verify that there exists  $K > 0$  such that for all  $\omega \in [-\pi, \pi]$ ,  $\sum_{|k|>K} |\hat{\phi}_o(\omega + 2k\pi)|^2 < 1/2$ , so (7.197) implies that  $\sum_{k=-K}^K |\hat{\phi}_o(\omega + 2k\pi)|^2 \geq 1/2$ . It follows that

$$\sum_{k=-K}^K |\hat{\phi}_o(\omega + 2k\pi)|^4 \geq \frac{1}{4(2K + 1)},$$

which proves (7.196) for  $A^{-1} = 4(2K + 1)$ .

Since  $\phi_o$  is a scaling function, (7.23) proves that there exists a conjugate mirror filter  $h_o$  such that

$$\frac{1}{\sqrt{2}} \phi_o\left(\frac{t}{2}\right) = \sum_{n=-\infty}^{+\infty} h_o[n] \phi_o(t-n).$$

Computing  $\phi(t) = \phi_o \star \bar{\phi}_o(t)$  yields (7.194) with  $h[n] = h_o \star \bar{h}_o[n]$ . ■

Theorem 7.20 proves that the autocorrelation of an orthogonal scaling function  $\phi_o$  is an interpolation function  $\phi$  that also satisfies a scaling equation. One can design  $\phi$  to approximate regular signals efficiently by their orthogonal projection in  $\mathbf{U}_s$ . Definition 6.1 measures the regularity of  $f$  with a Lipschitz exponent, which depends on the difference between  $f$  and its Taylor polynomial expansion. Theorem 7.21 gives a condition for recovering polynomials by interpolating their samples with  $\phi$ . It derives an upper bound for the error when approximating  $f$  by its orthogonal projection in  $\mathbf{U}_s$ .

**Theorem 7.21:** Fix, Strang. Any polynomial  $q(t)$  of degree smaller or equal to  $p-1$  is decomposed into

$$q(t) = \sum_{n=-\infty}^{+\infty} q(n) \phi(t-n) \quad (7.198)$$

if and only if  $\hat{h}(\omega)$  has a zero of order  $p$  at  $\omega = \pi$ .

Suppose that this property is satisfied. If  $f$  has a compact support and is uniformly Lipschitz  $\alpha \leq p$ , then there exists  $C > 0$  such that

$$\forall s > 0, \quad \|f - P_{\mathbf{U}_s} f\| \leq C s^\alpha. \quad (7.199)$$

**Proof.** The main steps of the proof are given without technical detail. Let us set  $s = 2^j$ .

One can verify that the spaces  $\{\mathbf{V}_j = \mathbf{U}_{2^j}\}_{j \in \mathbb{Z}}$  define a multiresolution approximation of  $L^2(\mathbb{R})$ . The Riesz basis of  $\mathbf{V}_0$  required by Definition 7.1 is obtained with  $\theta = \phi$ . This basis is orthogonalized by Theorem 7.1 to obtain an orthogonal basis of scaling functions. Theorem 7.3 derives a wavelet orthonormal basis  $\{\psi_{j,n}\}_{(j,n) \in \mathbb{Z}^2}$  of  $L^2(\mathbb{R})$ .

Using Theorem 7.4, one can verify that  $\psi$  has  $p$  vanishing moments if and only if  $\hat{h}(\omega)$  has  $p$  zeros at  $\pi$ . Although  $\phi$  is not the orthogonal scaling function, the Fix-Strang condition (7.70) remains valid. It is also equivalent that for  $k < p$ ,

$$q_k(t) = \sum_{n=-\infty}^{+\infty} n^k \phi(t-n)$$

is a polynomial of degree  $k$ . The interpolation property (7.191) implies that  $q_k(n) = n^k$  for all  $n \in \mathbb{Z}$ , so  $q_k(t) = t^k$ . Since  $\{t^k\}_{0 \leq k < p}$  is a basis for polynomials of degree  $p-1$ , any polynomial  $q(t)$  of degree  $p-1$  can be decomposed over  $\{\phi(t-n)\}_{n \in \mathbb{Z}}$  if and only if  $\hat{h}(\omega)$  has  $p$  zeros at  $\pi$ .

We indicate how to prove (7.199) for  $s = 2^j$ . The truncated family of wavelets  $\{\psi_{l,n}\}_{l \leq j, n \in \mathbb{Z}}$  is an orthogonal basis of the orthogonal complement of  $\mathbf{U}_{2^j} = \mathbf{V}_j$  in  $L^2(\mathbb{R})$ . Thus,

$$\|f - P_{\mathbf{U}_{2^j}} f\|^2 = \sum_{l=-\infty}^j \sum_{n=-\infty}^{+\infty} |\langle f, \psi_{l,n} \rangle|^2.$$

If  $f$  is uniformly Lipschitz  $\alpha$ , since  $\psi$  has  $p$  vanishing moments, Theorem 6.3 proves that there exists  $A > 0$  such that

$$|Wf(2^l n, 2^l)| = |\langle f, \psi_{l,n} \rangle| \leq A 2^{(\alpha+1/2)l}.$$

To simplify the argument we suppose that  $\psi$  has a compact support, although this is not required. Since  $f$  also has a compact support, one can verify that the number of nonzero  $\langle f, \psi_{l,n} \rangle$  is bounded by  $K 2^{-l}$  for some  $K > 0$ . Thus,

$$\|f - P_{U_{2^j}} f\|^2 \leq \sum_{l=-\infty}^j K 2^{-l} A^2 2^{(2\alpha+1)l} \leq \frac{KA^2}{1-2^{-\alpha}} 2^{2\alpha j},$$

which proves (7.199) for  $s = 2^j$ . ■

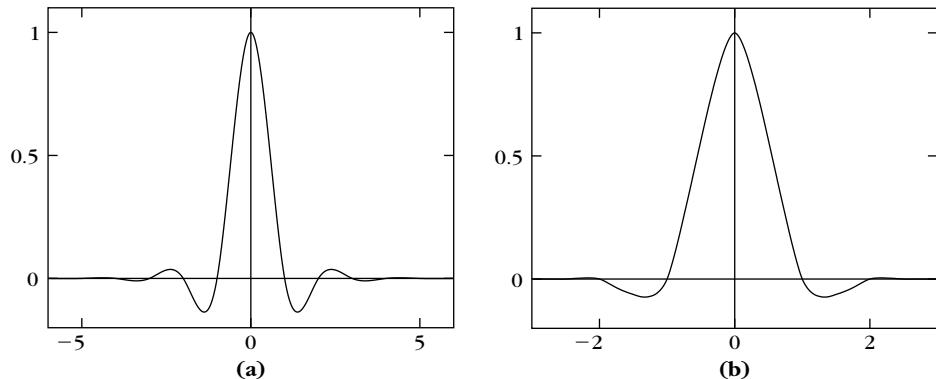
As long as  $\alpha \leq p$ , the larger the Lipschitz exponent  $\alpha$ , the faster the error  $\|f - P_{U_s} f\|$  decays to zero when the sampling interval  $s$  decreases. If a signal  $f$  is  $C^k$  with a compact support, then it is uniformly Lipschitz  $k$ , so Theorem 7.21 proves that  $\|f - P_{U_s} f\| \leq C s^k$ .

### EXAMPLE 7.11

A cubic spline-interpolation function is obtained from the linear spline-scaling function  $\phi_o$ . The Fourier transform expression (7.5) yields

$$\hat{\phi}(\omega) = |\hat{\phi}_o(\omega)|^2 = \frac{48 \sin^4(\omega/2)}{\omega^4 (1 + 2 \cos^2(\omega/2))}. \quad (7.200)$$

Figure 7.19(a) gives the graph of  $\phi$ , which has an infinite support but exponential decay. With Theorem 7.21, one can verify that this interpolation function recovers polynomials of degree



**FIGURE 7.19**

(a) Cubic spline-interpolation function. (b) Deslauriers-Dubuc interpolation function of degree 3.

3 from a uniform sampling. The performance of spline interpolation functions for generalized sampling theorems is studied in [162, 468].

---

### EXAMPLE 7.12

Deslauriers-Dubuc [206] interpolation functions of degree  $2p - 1$  are compactly supported interpolation functions of minimal size that decompose polynomials of degree  $2p - 1$ . One can verify that such an interpolation function is the autocorrelation of a scaling function  $\phi_o$ . To reproduce polynomials of degree  $2p - 1$ , Theorem 7.21 proves that  $\hat{h}(\omega)$  must have a zero of order  $2p$  at  $\pi$ . Since  $h[n] = h_o \star h_o[n]$ , it follows that  $\hat{h}(\omega) = |\hat{h}_o(\omega)|^2$ , and thus  $\hat{h}_o(\omega)$  has a zero of order  $p$  at  $\pi$ . The Daubechies theorem (7.7) designs minimum-size conjugate mirror filters  $h_o$  that satisfy this condition. Daubechies filters  $h_o$  have  $2p$  nonzero coefficients and the resulting scaling function  $\phi_o$  has a support of size  $2p - 1$ . The autocorrelation  $\phi$  is the Deslauriers-Dubuc interpolation function, which support  $[-2p + 1, 2p - 1]$ .

For  $p = 1$ ,  $\phi_o = \mathbf{1}_{[0,1]}$  and  $\phi$  are the piecewise linear tent functions with a support that  $[-1, 1]$ . For  $p = 2$ , the Deslauriers-Dubuc interpolation function  $\phi$  is the autocorrelation of the Daubechies 2 scaling function, shown in Figure 7.10. The graph of this interpolation function is in Figure 7.19(b). Polynomials of degree  $2p - 1 = 3$  are interpolated by this function.

The scaling equation (7.194) implies that any autocorrelation filter verifies  $h[2n] = 0$  for  $n \neq 0$ . For any  $p \geq 0$ , the nonzero values of the resulting filter are calculated from the coefficients of the polynomial (7.168) that is factored to synthesize Daubechies filters. The support of  $h$  is  $[-2p + 1, 2p - 1]$  and

$$h[2n+1] = (-1)^{p-n} \frac{\prod_{k=0}^{2p-1} (k - p + 1/2)}{(n + 1/2)(p - n - 1)!(p + n)!} \quad \text{for } -p \leq n < p. \quad (7.201)$$


---

### Dual Basis

If  $f \notin \mathbf{U}_s$ , then it is approximated by its orthogonal projection  $P_{\mathbf{U}_s} f$  on  $\mathbf{U}_s$  before the samples at intervals  $s$  are recorded. This orthogonal projection is computed with a biorthogonal basis  $\{\tilde{\phi}_s(t - ns)\}_{n \in \mathbb{Z}}$  [82]. Theorem 3.4 proves that  $\tilde{\phi}_s(t) = s^{-1} \tilde{\phi}(s^{-1}t)$  where the Fourier transform of  $\tilde{\phi}$  is

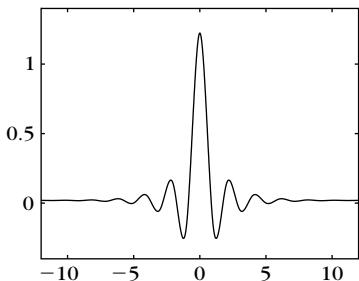
$$\tilde{\phi}(\omega) = \frac{\hat{\phi}^*(\omega)}{\sum_{k=-\infty}^{+\infty} |\hat{\phi}(\omega + 2k\pi)|^2}. \quad (7.202)$$

Figure 7.20 gives the graph of the cubic spline  $\tilde{\phi}$  associated to the cubic spline-interpolation function. The orthogonal projection of  $f$  over  $\mathbf{U}_s$  is computed by decomposing  $f$  in the biorthogonal bases

$$P_{\mathbf{U}_s} f(t) = \sum_{n=-\infty}^{+\infty} \langle f(u), \tilde{\phi}_s(u - ns) \rangle \phi_s(t - ns). \quad (7.203)$$

Let  $\tilde{\phi}_s(t) = \tilde{\phi}_s(-t)$ . The interpolation property (7.190) implies that

$$P_{\mathbf{U}_s} f(ns) = \langle f(u), \tilde{\phi}_s(u - ns) \rangle = f \star \tilde{\phi}_s(ns). \quad (7.204)$$

**FIGURE 7.20**

The dual cubic spline  $\tilde{\phi}(t)$  associated to the cubic spline–interpolation function  $\phi(t)$  shown in Figure 7.19(a).

Therefore, this discretization of  $f$  through a projection onto  $\mathbf{U}_s$  is obtained by a filtering with  $\tilde{\phi}_s$  followed by a uniform sampling at intervals  $s$ . The best linear approximation of  $f$  is recovered with the interpolation formula (7.203).

## 7.6.2 Interpolation Wavelet Basis

An interpolation function  $\phi$  can recover a signal  $f$  from a uniform sampling  $\{f(ns)\}_{n \in \mathbb{Z}}$  if  $f$  belongs to an appropriate subspace  $\mathbf{U}_s$  of  $\mathbf{L}^2(\mathbb{R})$ . Donoho [213] has extended this approach by constructing interpolation wavelet bases of the whole space of uniformly continuous signals with the supremum norm. The decomposition coefficients are calculated from sample values instead of inner product integrals.

### **Subdivision Scheme**

Let  $\phi$  be an interpolation function that is the autocorrelation of an orthogonal scaling function  $\phi_0$ . Let  $\phi_{j,n}(t) = \phi(2^{-j}t - n)$ . The constant  $2^{-j/2}$  that normalizes the energy of  $\phi_{j,n}$  is not added because we shall use a supremum norm  $\|f\|_\infty = \sup_{t \in \mathbb{R}} |f(t)|$  instead of the  $\mathbf{L}^2(\mathbb{R})$  norm, and

$$\|\phi_{j,n}\|_\infty = \|\phi\|_\infty = |\phi(0)| = 1.$$

We define the interpolation space  $\mathbf{V}_j$  of functions

$$g = \sum_{n=-\infty}^{+\infty} a[n] \phi_{j,n},$$

where  $a[n]$  has at most a polynomial growth in  $n$ . Since  $\phi$  is an interpolation function,  $a[n] = g(2^j n)$ . This space  $\mathbf{V}_j$  is not included in  $\mathbf{L}^2(\mathbb{R})$  since  $a[n]$  may not have a finite energy. The scaling equation (7.194) implies that  $\mathbf{V}_{j+1} \subset \mathbf{V}_j$  for any  $j \in \mathbb{Z}$ . If the autocorrelation filter  $h$  has a Fourier transform  $\hat{h}(\omega)$  that has a zero of order  $p$  at  $\omega = \pi$ , then Theorem 7.21 proves that polynomials of a degree smaller than  $p - 1$  are included in  $\mathbf{V}_j$ .

For  $f \notin \mathbf{V}_j$ , we define a simple projector on  $\mathbf{V}_j$  that interpolates the dyadic samples  $f(2^j n)$ :

$$P_{\mathbf{V}_j} f(t) = \sum_{n=-\infty}^{+\infty} f(2^j n) \phi_j(t - 2^j n). \quad (7.205)$$

This projector has no orthogonality property but satisfies  $P_{\mathbf{V}_j} f(2^j n) = f(2^j n)$ . Let  $\mathbf{C}_0$  be the space of functions that are uniformly continuous over  $\mathbb{R}$ . Theorem 7.22 proves that any  $f \in \mathbf{C}_0$  can be approximated with an arbitrary precision by  $P_{\mathbf{V}_j} f$  when  $2^j$  goes to zero.

**Theorem 7.22:** *Donoho.* Suppose that  $\phi$  has an exponential decay. If  $f \in \mathbf{C}_0$ , then

$$\lim_{j \rightarrow -\infty} \|f - P_{\mathbf{V}_j} f\|_\infty = \lim_{j \rightarrow -\infty} \sup_{t \in \mathbb{R}} |f(t) - P_{\mathbf{V}_j} f(t)| = 0. \quad (7.206)$$

**Proof.** Let  $\omega(\delta, f)$  denote the modulus of continuity

$$\omega(\delta, f) = \sup_{|h| \leq \delta} \sup_{t \in \mathbb{R}} |f(t+h) - f(t)|. \quad (7.207)$$

By definition,  $f \in \mathbf{C}_0$  if  $\lim_{\delta \rightarrow 0} \omega(\delta, f) = 0$ .

Any  $t \in \mathbb{R}$  can be written as  $t = 2^j(n + h)$  with  $n \in \mathbb{Z}$  and  $|h| \leq 1$ . Since  $P_{\mathbf{V}_j} f(2^j n) = f(2^j n)$ ,

$$\begin{aligned} |f(2^j(n+h)) - P_{\mathbf{V}_j} f(2^j(n+h))| &\leq |f(2^j(n+h)) - f(2^j n)| \\ &\quad + |P_{\mathbf{V}_j} f(2^j(n+h)) - P_{\mathbf{V}_j} f(2^j n)| \\ &\leq \omega(2^j, f) + \omega(2^j, P_{\mathbf{V}_j} f). \end{aligned}$$

Lemma 7.3 proves that  $\omega(2^j, P_{\mathbf{V}_j} f) \leq C_\phi \omega(2^j, f)$  where  $C_\phi$  is a constant independent of  $j$  and  $f$ . Taking a supremum over  $t = 2^j(n + h)$  implies the final result:

$$\sup_{t \in \mathbb{R}} |f(t) - P_{\mathbf{V}_j} f(t)| \leq (1 + C_\phi) \omega(2^j, f) \rightarrow 0 \quad \text{when } j \rightarrow -\infty.$$

**Lemma 7.3.** There exists  $C_\phi > 0$  such that for all  $j \in \mathbb{Z}$  and  $f \in \mathbf{C}_0$ ,

$$\omega(2^j, P_{\mathbf{V}_j} f) \leq C_\phi \omega(2^j, f). \quad (7.208)$$

Let us set  $j = 0$ . For  $|h| \leq 1$ , a summation by parts gives

$$P_{\mathbf{V}_0} f(t+h) - P_{\mathbf{V}_0} f(t) = \sum_{n=-\infty}^{+\infty} (f(n+1) - f(n)) \theta_h(t-n),$$

where

$$\theta_h(t) = \sum_{k=1}^{+\infty} (\phi(t+h-k) - \phi(t-k)).$$

Thus,

$$|P_{V_0}f(t+h) - P_{V_0}f(t)| \leq \sup_{n \in \mathbb{Z}} |f(n+1) - f(n)| \sum_{n=-\infty}^{+\infty} |\theta_h(t-n)|. \quad (7.209)$$

Since  $\phi$  has an exponential decay, there exists a constant  $C_\phi$  such that if  $|h| \leq 1$  and  $t \in \mathbb{R}$ , then  $\sum_{n=-\infty}^{+\infty} |\theta_h(t-n)| \leq C_\phi$ . Taking a supremum over  $t$  in (7.209) proves that

$$\omega(1, P_{V_0}f) \leq C_\phi \sup_{n \in \mathbb{Z}} |f(n+1) - f(n)| \leq C_\phi \omega(1, f).$$

Scaling this result by  $2^j$  yields (7.208). ■

### Interpolation Wavelets

The projection  $P_{V_j}f(t)$  interpolates the values  $f(2^j n)$ . When reducing the scale by 2, we obtain a finer interpolation  $P_{V_{j-1}}f(t)$  that also goes through the intermediate samples  $f(2^j(n+1/2))$ . This refinement can be obtained by adding “details” that compensate for the difference between  $P_{V_j}f(2^j(n+1/2))$  and  $f(2^j(n+1/2))$ . To do this, we create a “detail” space  $\mathbf{W}_j$  that provides the values  $f(t)$  at intermediate dyadic points  $t = 2^j(n+1/2)$ . This space is constructed from interpolation functions centered at these locations, namely  $\phi_{j-1,2n+1}$ . We call *interpolation wavelets*

$$\psi_{j,n} = \phi_{j-1,2n+1}.$$

Observe that  $\psi_{j,n}(t) = \psi(2^{-j}t - n)$  with

$$\psi(t) = \phi(2t - 1).$$

The function  $\psi$  is not truly a wavelet since it has no vanishing moment. However, we shall see that it plays the same role as a wavelet in this decomposition. We define  $\mathbf{W}_j$  to be the space of all sums  $\sum_{n=-\infty}^{+\infty} a[n] \psi_{j,n}$ . Theorem 7.23 proves that it is a (nonorthogonal) complement of  $\mathbf{V}_j$  in  $\mathbf{V}_{j-1}$ .

**Theorem 7.23.** For any  $j \in \mathbb{Z}$ ,

$$\mathbf{V}_{j-1} = \mathbf{V}_j \oplus \mathbf{W}_j.$$

If  $f \in \mathbf{V}_{j-1}$ , then

$$f = \sum_{n=-\infty}^{+\infty} f(2^j n) \phi_{j,n} + \sum_{n=-\infty}^{+\infty} d_j[n] \psi_{j,n},$$

with

$$d_j[n] = f\left(2^j(n+1/2)\right) - P_{V_j}f\left(2^j(n+1/2)\right). \quad (7.210)$$

**Proof.** Any  $f \in \mathbf{V}_{j-1}$  can be written as

$$f = \sum_{n=-\infty}^{+\infty} f(2^{j-1}n) \phi_{j-1,n}.$$

The function  $f - P_{V_j}f$  belongs to  $V_{j-1}$  and vanishes at  $\{2^j n\}_{n \in \mathbb{Z}}$ . Thus, it can be decomposed over the intermediate interpolation functions  $\phi_{j-1,2n+1} = \psi_{j,n}$ :

$$f(t) - P_{V_j}f(t) = \sum_{n=-\infty}^{+\infty} d_j[n] \psi_{j,n}(t) \in W_j.$$

This proves that  $V_{j-1} \subset V_j \oplus W_j$ . By construction we know that  $W_j \subset V_{j-1}$ , so  $V_{j-1} = V_j \oplus W_j$ . Setting  $t = 2^{j-1}(2n + 1)$  in this formula also verifies (7.210). ■

Theorem 7.23 refines an interpolation from a coarse grid  $2^j n$  to a finer grid  $2^{j-1}n$  by adding “details” with coefficients  $d_j[n]$  that are the interpolation errors  $f(2^j(n + 1/2)) - P_{V_j}f(2^j(n + 1/2))$ . The following Theorem 7.24 defines an interpolation wavelet basis of  $C_0$  in the sense of uniform convergence.

**Theorem 7.24.** If  $f \in C_0$ , then

$$\lim_{\substack{m \rightarrow +\infty \\ l \rightarrow -\infty}} \|f - \sum_{n=-m}^m f(2^J n) \phi_{J,n} - \sum_{j=l}^J \sum_{n=-m}^m d_j[n] \psi_{j,n}\|_\infty = 0. \quad (7.211)$$

The formula (7.211) decomposes  $f$  into a coarse interpolation at intervals  $2^J$  plus layers of details that give the interpolation errors on successively finer dyadic grids. The proof is done by choosing  $f$  to be a continuous function with a compact support, in which case (7.211) is derived from Theorem 7.23 and (7.206). The density of such functions in  $C_0$  (for the supremum norm) allows us to extend this result to any  $f$  in  $C_0$ . We shall write

$$f = \sum_{n=-\infty}^{+\infty} f(2^J n) \phi_{J,n} + \sum_{j=-\infty}^J \sum_{n=-\infty}^{+\infty} d_j[n] \psi_{j,n},$$

which means that  $[\{\phi_{J,n}\}_{n \in \mathbb{Z}}, \{\psi_{j,n}\}_{n \in \mathbb{Z}, j \leq J}]$  is a basis of  $C_0$ . In  $L^2(\mathbb{R})$ , “biorthogonal” scaling functions and wavelets are formally defined by

$$f(2^J n) = \langle f, \tilde{\phi}_{J,n} \rangle = \int_{-\infty}^{+\infty} f(t) \tilde{\phi}_{J,n}(t) dt,$$

$$d_j[n] = \langle f, \tilde{\psi}_{j,n} \rangle = \int_{-\infty}^{+\infty} f(t) \tilde{\psi}_{j,n}(t) dt. \quad (7.212)$$

Clearly,  $\tilde{\phi}_{J,n}(t) = \delta(t - 2^J n)$ . Similarly, (7.210) and (7.205) implies that  $\tilde{\psi}_{j,n}$  is a finite sum of Diracs. These dual-scaling functions and wavelets do not have a finite energy, which illustrates the fact that  $[\{\phi_{J,n}\}_{n \in \mathbb{Z}}, \{\psi_{j,n}\}_{n \in \mathbb{Z}, j \leq J}]$  is not a Riesz basis of  $L^2(\mathbb{R})$ .

If  $\hat{h}(\omega)$  has  $p$  zeros at  $\pi$ , then one can verify that  $\tilde{\psi}_{j,n}$  has  $p$  vanishing moments. With similar derivations as in the proof of (6.20) in Theorem 6.4, one can show that if  $f$  is uniformly Lipschitz  $\alpha \leq p$ , then there exists  $A > 0$  such that

$$|\langle f, \tilde{\psi}_{j,n} \rangle| = |d_j[n]| \leq A 2^{aj}.$$

A regular signal yields small-amplitude wavelet coefficients at fine scales. Thus, we can neglect these coefficients and still reconstruct a precise approximation of  $f$ .

### Fast Calculations

The interpolating wavelet transform of  $f$  is calculated at scale  $1 \geq 2^j > N^{-1} = 2^L$  from its sample values  $\{f(N^{-1}n)\}_{n \in \mathbb{Z}}$ . At each scale  $2^j$ , the values of  $f$  in between samples  $\{2^j n\}_{n \in \mathbb{Z}}$  are calculated with the interpolation (7.205):

$$\begin{aligned} P_{V_j} f(2^j(n + 1/2)) &= \sum_{k=-\infty}^{+\infty} f(2^j k) \phi(n - k + 1/2) \\ &= \sum_{k=-\infty}^{+\infty} f(2^j k) h_i[n - k], \end{aligned} \quad (7.213)$$

where the interpolation filter  $h_i$  is a subsampling of the autocorrelation filter  $h$  in (7.195):

$$h_i[n] = \phi(n + 1/2) = h[2n + 1]. \quad (7.214)$$

The wavelet coefficients are computed with (7.210):

$$d_j[n] = f(2^j(n + 1/2)) - P_{V_j} f(2^j(n + 1/2)).$$

The reconstruction of  $f(N^{-1}n)$  from the wavelet coefficients is performed recursively by recovering the samples  $f(2^{j-1}n)$  from the coarser sampling  $f(2^j n)$  with the interpolation (7.213) to which is added  $d_j[n]$ . If  $h_i[n]$  is a finite filter of size  $K$  and if  $f$  has a support in  $[0, 1]$ , then the decomposition and reconstruction algorithms require  $KN$  multiplications and additions.

A Deslauriers-Dubuc interpolation function  $\phi$  has the shortest support while including polynomials of degree  $2p - 1$  in the spaces  $V_j$ . The corresponding interpolation filter  $h_i[n]$  defined by (7.214) has  $2p$  nonzero coefficients for  $-p \leq n < p$ , which are calculated in (7.201). If  $p = 2$ , then  $h_i[1] = h_i[-2] = -1/16$  and  $h_i[0] = h_i[-1] = 9/16$ . Suppose that  $q(t)$  is a polynomial of degree smaller or equal to  $2p - 1$ . Since  $q = P_{V_j} q$ , (7.213) implies a Lagrange interpolation formula

$$q(2^j(n + 1/2)) = \sum_{k=-\infty}^{+\infty} q(2^j k) h_i[n - k].$$

The Lagrange filter  $h_i$  of size  $2p$  is the shortest filter that recovers intermediate values of polynomials of degree  $2p - 1$  from a uniform sampling.

To restrict the wavelet interpolation bases to a finite interval  $[0, 1]$  while reproducing polynomials of degree  $2p - 1$ , the filter  $h_i$  is modified at the boundaries. Suppose that  $f(N^{-1}n)$  is defined for  $0 \leq n < N$ . When computing the interpolation

$$P_{V_j} f(2^j(n + 1/2)) = \sum_{k=-\infty}^{+\infty} f(2^j k) h_i[n - k],$$

if  $n$  is too close to 0 or to  $2^{-j} - 1$ , then  $h_i$  must be modified to ensure that the support of  $h_i[n - k]$  remains inside  $[0, 2^{-j} - 1]$ . The interpolation  $P_{V_j} f(2^j(n + 1/2))$

is then calculated from the closest  $2p$  samples  $f(2^j k)$  for  $2^j k \in [0, 1]$ . The new interpolation coefficients are computed in order to recover exactly all polynomials of degree  $2p - 1$  [450]. For  $p = 2$ , the problem occurs only at  $n = 0$  and the appropriate boundary coefficients are

$$h_i[0] = \frac{5}{16}, \quad h_i[-1] = \frac{15}{16}, \quad h_i[-2] = \frac{-5}{16}, \quad h_i[-3] = \frac{1}{16}.$$

The symmetric boundary filter  $h_i[-n]$  is used on the other side at  $n = 2^{-j} - 1$ .

## 7.7 SEPARABLE WAVELET BASES

To any wavelet orthonormal basis  $\{\psi_{j,n}\}_{(j,n) \in \mathbb{Z}^2}$  of  $\mathbf{L}^2(\mathbb{R})$ , one can associate a separable wavelet orthonormal basis of  $\mathbf{L}^2(\mathbb{R}^2)$ :

$$\left\{ \psi_{j_1,n_1}(x_1) \psi_{j_2,n_2}(x_2) \right\}_{(j_1,j_2,n_1,n_2) \in \mathbb{Z}^4}. \quad (7.215)$$

The functions  $\psi_{j_1,n_1}(x_1) \psi_{j_2,n_2}(x_2)$  mix information at two different scales  $2^{j_1}$  and  $2^{j_2}$  along  $x_1$  and  $x_2$ , which we often want to avoid. Separable multiresolutions lead to another construction of separable wavelet bases with wavelets that are products of functions dilated at the same scale. These multiresolution approximations also have important applications in computer vision, where they are used to process images at different levels of details. Lower-resolution images are indeed represented by fewer pixels and might still carry enough information to perform a recognition task.

Signal decompositions in separable wavelet bases are computed with a separable extension of the filter bank algorithm described in Section 7.7.3. Section 7.7.4 constructs separable wavelet bases in any dimension, and explains the corresponding fast wavelet transform algorithm. Nonseparable wavelet bases can also be constructed [85, 334] but they are used less often in image processing. Section 7.8.3 gives examples of nonseparable quincunx biorthogonal wavelet bases, which have a single quasi-istropic wavelet at each scale.

### 7.7.1 Separable Multiresolutions

As in one dimension, the notion of resolution is formalized with orthogonal projections in spaces of various sizes. The approximation of an image  $f(x_1, x_2)$  at the resolution  $2^{-j}$  is defined as the orthogonal projection of  $f$  on a space  $\mathbf{V}_j^2$  that is included in  $\mathbf{L}^2(\mathbb{R}^2)$ . The space  $\mathbf{V}_j^2$  is the set of all approximations at the resolution  $2^{-j}$ . When the resolution decreases, the size of  $\mathbf{V}_j^2$  decreases as well. The formal definition of a multiresolution approximation  $\{\mathbf{V}_j^2\}_{j \in \mathbb{Z}}$  of  $\mathbf{L}^2(\mathbb{R}^2)$  is a straightforward extension of Definition 7.1 that specifies multiresolutions of  $\mathbf{L}^2(\mathbb{R})$ . The same causality, completeness, and scaling properties must be satisfied.

We consider the particular case of separable multiresolutions. Let  $\{\mathbf{V}_j\}_{j \in \mathbb{Z}}$  be a multiresolution of  $\mathbf{L}^2(\mathbb{R})$ . A separable two-dimensional multiresolution is composed of the tensor product spaces

$$\mathbf{V}_j^2 = \mathbf{V}_j \otimes \mathbf{V}_j. \quad (7.216)$$

The space  $\mathbf{V}_j^2$  is the set of finite energy functions  $f(x_1, x_2)$  that are linear expansions of separable functions:

$$f(x_1, x_2) = \sum_{m=-\infty}^{+\infty} a[m] f_m(x_1) g_m(x_2) \quad \text{with } f_m \in \mathbf{V}_j, \quad g_m \in \mathbf{V}_j.$$

Section A.5 reviews the properties of tensor products. If  $\{\mathbf{V}_j\}_{j \in \mathbb{Z}}$  is a multiresolution approximation of  $\mathbf{L}^2(\mathbb{R})$ , then  $\{\mathbf{V}_j^2\}_{j \in \mathbb{Z}}$  is a multiresolution approximation of  $\mathbf{L}^2(\mathbb{R}^2)$ .

Theorem 7.1 demonstrates the existence of a scaling function  $\phi$  such that  $\{\phi_{j,m}\}_{m \in \mathbb{Z}}$  is an orthonormal basis of  $\mathbf{V}_j$ . Since  $\mathbf{V}_j^2 = \mathbf{V}_j \otimes \mathbf{V}_j$ , Theorem A.6 proves that for  $x = (x_1, x_2)$  and  $n = (n_1, n_2)$ ,

$$\left\{ \phi_{j,n}^2(x) = \phi_{j,n_1}(x_1) \phi_{j,n_2}(x_2) = \frac{1}{2^j} \phi\left(\frac{x_1 - 2^j n_1}{2^j}\right) \phi\left(\frac{x_2 - 2^j n_2}{2^j}\right) \right\}_{n \in \mathbb{Z}^2}$$

is an orthonormal basis of  $\mathbf{V}_j^2$ . It is obtained by scaling by  $2^j$  the two-dimensional separable scaling function  $\phi^2(x) = \phi(x_1) \phi(x_2)$  and translating it on a two-dimensional square grid with intervals  $2^j$ .

---

### EXAMPLE 7.13: Piecewise Constant Approximation

Let  $\mathbf{V}_j$  be the approximation space of functions that are constant on  $[2^j m, 2^j(m+1)]$  for any  $m \in \mathbb{Z}$ . The tensor product defines a two-dimensional piecewise constant approximation. The space  $\mathbf{V}_j^2$  is the set of functions that are constant on any square  $[2^j n_1, 2^j(n_1+1)] \times [2^j n_2, 2^j(n_2+1)]$ , for  $(n_1, n_2) \in \mathbb{Z}^2$ . The two-dimensional scaling function is

$$\phi^2(x) = \phi(x_1) \phi(x_2) = \begin{cases} 1 & \text{if } 0 \leq x_1 \leq 1 \text{ and } 0 \leq x_2 \leq 1 \\ 0 & \text{otherwise.} \end{cases}.$$


---

---

### EXAMPLE 7.14: Shannon Approximation

Let  $\mathbf{V}_j$  be the space of functions with Fourier transforms that have a support included in  $[-2^{-j}\pi, 2^{-j}\pi]$ . Space  $\mathbf{V}_j^2$  is the set of functions the two-dimensional Fourier transforms of which have a support included in the low-frequency square  $[-2^{-j}\pi, 2^{-j}\pi] \times [-2^{-j}\pi, 2^{-j}\pi]$ . The two-dimensional scaling function is a perfect two-dimensional low-pass filter the Fourier transform of which is

$$\hat{\phi}(\omega_1) \hat{\phi}(\omega_2) = \begin{cases} 1 & \text{if } |\omega_1| \leq 2^{-j}\pi \text{ and } |\omega_2| \leq 2^{-j}\pi \\ 0 & \text{otherwise.} \end{cases}.$$


---

---

**EXAMPLE 7.15: Spline Approximation**

Let  $\mathbf{V}_j$  be the space of polynomial spline functions of degree  $p$  that are  $\mathbf{C}^{p-1}$  with nodes located at  $2^{-j}m$  for  $m \in \mathbb{Z}$ . The space  $\mathbf{V}_j^2$  is composed of two-dimensional polynomial spline functions that are  $p - 1$  times continuously differentiable. The restriction of  $f(x_1, x_2) \in \mathbf{V}_j^2$  to any square  $[2^j n_1, 2^j(n_1 + 1)] \times [2^j n_2, 2^j(n_2 + 1)]$  is a separable product  $q_1(x_1)q_2(x_2)$  of two polynomials of degree at most  $p$ .

---

### **Multiresolution Vision**

An image of  $512 \times 512$  pixels often includes too much information for real-time vision processing. Multiresolution algorithms process less image data by selecting the relevant details that are necessary to perform a particular recognition task [58]. The human visual system uses a similar strategy. The distribution of photoreceptors on the retina is not uniform. The visual acuity is greatest at the center of the retina where the density of receptors is maximum. When moving apart from the center, the resolution decreases proportionally to the distance from the retina center [428].

The high-resolution visual center is called the *fovea*. It is responsible for high-acuity tasks such as reading or recognition. A retina with a uniform resolution equal to the highest fovea resolution would require about 10,000 times more photoreceptors. Such a uniform resolution retina would increase considerably the size of the optic nerve that transmits the retina information to the visual cortex and the size of the visual cortex that processes these data.

Active vision strategies [83] compensate the nonuniformity of visual resolution with eye saccades, which move successively the fovea over regions of a scene with a high information content. These saccades are partly guided by the lower-resolution information gathered at the periphery of the retina. This multiresolution sensor has the advantage of providing high-resolution information at selected locations and a large field of view with relatively little data.

Multiresolution algorithms implement in software [125] the search for important high-resolution data. A uniform high-resolution image is measured by a camera but only a small part of this information is processed. Figure 7.21 displays a pyramid of progressively lower-resolution images calculated with a filter bank presented in Section 7.7.3. Coarse to fine algorithms analyze first the lower-resolution image and selectively increase the resolution in regions where more details are needed. Such algorithms have been developed for object recognition and stereo calculations [284].

### **7.7.2 Two-Dimensional Wavelet Bases**

A separable wavelet orthonormal basis of  $\mathbf{L}^2(\mathbb{R}^2)$  is constructed with separable products of a scaling function  $\phi$  and a wavelet  $\psi$ . The scaling function  $\phi$  is associated to a one-dimensional multiresolution approximation  $\{\mathbf{V}_j\}_{j \in \mathbb{Z}}$ . Let  $\{\mathbf{V}_j^2\}_{j \in \mathbb{Z}}$  be the separable two-dimensional multiresolution defined by  $\mathbf{V}_j^2 = \mathbf{V}_j \otimes \mathbf{V}_j$ . Let  $\mathbf{W}_j^2$  be the detail

**FIGURE 7.21**

Multiresolution approximations  $a_j[n_1, n_2]$  of an image at scales  $2^j$  for  $-5 \geq j \geq -8$ .

space equal to the orthogonal complement of the lower-resolution approximation space  $\mathbf{V}_j^2$  in  $\mathbf{V}_{j-1}^2$ :

$$\mathbf{V}_{j-1}^2 = \mathbf{V}_j^2 \oplus \mathbf{W}_j^2. \quad (7.217)$$

To construct a wavelet orthonormal basis of  $\mathbf{L}^2(\mathbb{R}^2)$ , Theorem 7.25 builds a wavelet basis of each detail space  $\mathbf{W}_j^2$ .

**Theorem 7.25.** Let  $\phi$  be a scaling function and  $\psi$  be the corresponding wavelet generating a wavelet orthonormal basis of  $\mathbf{L}^2(\mathbb{R})$ . We define three wavelets:

$$\psi^1(x) = \phi(x_1)\psi(x_2), \quad \psi^2(x) = \psi(x_1)\phi(x_2), \quad \psi^3(x) = \psi(x_1)\psi(x_2), \quad (7.218)$$

and denote for  $1 \leq k \leq 3$ ,

$$\psi_{j,n}^k(x) = \frac{1}{2^j} \psi^k\left(\frac{x_1 - 2^j n_1}{2^j}, \frac{x_2 - 2^j n_2}{2^j}\right).$$

The wavelet family

$$\left\{\psi_{j,n}^1, \psi_{j,n}^2, \psi_{j,n}^3\right\}_{n \in \mathbb{Z}^2} \quad (7.219)$$

is an orthonormal basis of  $\mathbf{W}_j^2$ , and

$$\left\{\psi_{j,n}^1, \psi_{j,n}^2, \psi_{j,n}^3\right\}_{(j,n) \in \mathbb{Z}^3} \quad (7.220)$$

is an orthonormal basis of  $\mathbf{L}^2(\mathbb{R}^2)$ .

**Proof.** Equation (7.217) is rewritten as

$$\mathbf{V}_{j-1} \otimes \mathbf{V}_{j-1} = (\mathbf{V}_j \otimes \mathbf{V}_j) \oplus \mathbf{W}_j^2. \quad (7.221)$$

The one-dimensional multiresolution space  $\mathbf{V}_{j-1}$  can also be decomposed into  $\mathbf{V}_{j-1} = \mathbf{V}_j \oplus \mathbf{W}_j$ . By inserting this in (7.221), the distributivity of  $\oplus$  with respect to  $\otimes$  proves that

$$\mathbf{W}_j^2 = (\mathbf{V}_j \otimes \mathbf{W}_j) \oplus (\mathbf{W}_j \otimes \mathbf{V}_j) \oplus (\mathbf{W}_j \otimes \mathbf{W}_j). \quad (7.222)$$

Since  $\{\phi_{j,m}\}_{m \in \mathbb{Z}}$  and  $\{\psi_{j,m}\}_{m \in \mathbb{Z}}$  are orthonormal bases of  $\mathbf{V}_j$  and  $\mathbf{W}_j$ , we derive that

$$\{\phi_{j,n_1}(x_1) \psi_{j,n_2}(x_2), \psi_{j,n_1}(x_1) \phi_{j,n_2}(x_2), \psi_{j,n_1}(x_1) \psi_{j,n_2}(x_2)\}_{(n_1, n_2) \in \mathbb{Z}^2}$$

is an orthonormal basis of  $\mathbf{W}_j^2$ . As in the one-dimensional case, the overall space  $\mathbf{L}^2(\mathbb{R}^2)$  can be decomposed as an orthogonal sum of the detail spaces at all resolutions:

$$\mathbf{L}^2(\mathbb{R}^2) = \bigoplus_{j=-\infty}^{+\infty} \mathbf{W}_j^2. \quad (7.223)$$

Thus,

$$\{\phi_{j,n_1}(x_1) \psi_{j,n_2}(x_2), \psi_{j,n_1}(x_1) \phi_{j,n_2}(x_2), \psi_{j,n_1}(x_1) \psi_{j,n_2}(x_2)\}_{(j, n_1, n_2) \in \mathbb{Z}^3}$$

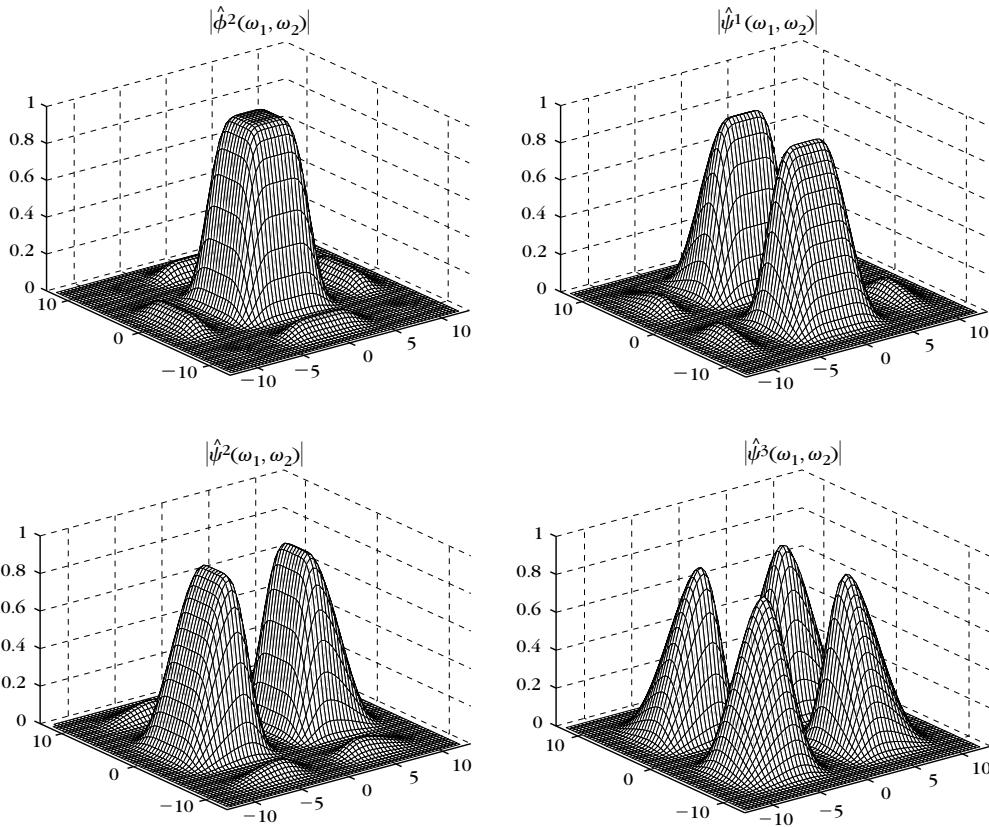
is an orthonormal basis of  $\mathbf{L}^2(\mathbb{R}^2)$ . ■

The three wavelets extract image details at different scales and in different directions. Overpositive frequencies,  $\hat{\phi}$  and  $\hat{\psi}$  have an energy mainly concentrated, respectively, on  $[0, \pi]$  and  $[\pi, 2\pi]$ . The separable wavelet expressions (7.218) imply that

$$\hat{\psi}^1(\omega_1, \omega_2) = \hat{\phi}(\omega_1) \hat{\psi}(\omega_2), \quad \hat{\psi}^2(\omega_1, \omega_2) = \hat{\psi}(\omega_1) \hat{\phi}(\omega_2),$$

and  $\hat{\psi}^3(\omega_1, \omega_2) = \hat{\psi}(\omega_1) \hat{\psi}(\omega_2)$ . Thus,  $|\hat{\psi}^1(\omega_1, \omega_2)|$  is large at low horizontal frequencies  $\omega_1$  and high vertical frequencies  $\omega_2$ , whereas  $|\hat{\psi}^2(\omega_1, \omega_2)|$  is large at high horizontal frequencies and low vertical frequencies, and  $|\hat{\psi}^3(\omega_1, \omega_2)|$  is large at high horizontal and vertical frequencies. Figure 7.22 displays the Fourier transform of separable wavelets and scaling functions calculated from a one-dimensional Daubechies 4 wavelet.

Suppose that  $\psi(t)$  has  $p$  vanishing moments and is orthogonal to one-dimensional polynomials of degree  $p - 1$ . The wavelet  $\psi^1$  has  $p$  one-dimensional directional vanishing moments along  $x_2$  in the sense that it is orthogonal to any function  $f(x_1, x_2)$  that is a polynomial of degree  $p - 1$  along  $x_2$  for  $x_1$  fixed. It is a horizontal directional wavelet that yields large coefficients for horizontal edges, as explained in Section 5.5.1. Similarly,  $\psi^2$  has  $p - 1$  directional vanishing moments along  $x_1$  and is a vertical directional wavelet. This is illustrated by the decomposition of a square later in Figure 7.24. The wavelet  $\psi^3$  has directional vanishing moments along both  $x_1$  and  $x_2$  and is therefore not a directional wavelet. It produces large coefficients

**FIGURE 7.22**

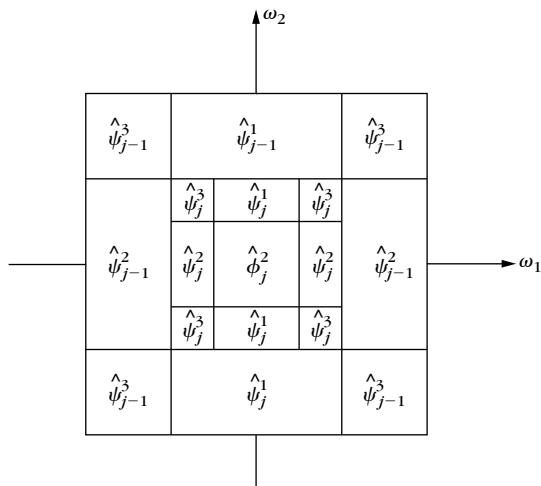
Fourier transforms of a separable scaling function and of three separable wavelets calculated from a one-dimensional Daubechies 4 wavelet.

at corners or junctions. The three wavelets  $\psi^k$  for  $k = 1, 2, 3$  are orthogonal to two-dimensional polynomials of degree  $p - 1$ .

### EXAMPLE 7.16

For a Shannon multiresolution approximation, the resulting two-dimensional wavelet basis paves the two-dimensional Fourier plane  $(\omega_1, \omega_2)$  with dilated rectangles. The Fourier transforms  $\hat{\phi}$  and  $\hat{\psi}$  are the indicator functions of  $[-\pi, \pi]$  and  $[-2\pi, -\pi] \cup [\pi, 2\pi]$ , respectively. The separable space  $\mathbf{V}_j^2$  contains functions with a two-dimensional Fourier transform support included in the low-frequency square  $[-2^{-j}\pi, 2^{-j}\pi] \times [-2^{-j}\pi, 2^{-j}\pi]$ . This corresponds to the support of  $\hat{\phi}_{j,n}^2$  indicated in Figure 7.23.

The detail space  $\mathbf{W}_j^2$  is the orthogonal complement of  $\mathbf{V}_j^2$  in  $\mathbf{V}_{j-1}^2$  and thus includes functions with Fourier transforms supported in the frequency annulus between the two squares

**FIGURE 7.23**

These dyadic rectangles indicate the regions where the energy of  $\hat{\psi}_{j,n}^k$  is mostly concentrated for  $1 \leq k \leq 3$ . Image approximations at the scale  $2^j$  are restricted to the lower-frequency square.

$[-2^{-j}\pi, 2^{-j}\pi] \times [-2^{-j}\pi, 2^{-j}\pi]$  and  $[-2^{-j+1}\pi, 2^{-j+1}\pi] \times [-2^{-j+1}\pi, 2^{-j+1}\pi]$ . As shown in Figure 7.23, this annulus is decomposed in three separable frequency regions, which are the Fourier supports of  $\hat{\psi}_{j,n}^k$  for  $1 \leq k \leq 3$ . Dilating these supports at all scales  $2^j$  yields an exact cover of the frequency plane  $(\omega_1, \omega_2)$ .

For general separable wavelet bases, Figure 7.23 gives only an indication of the domains where the energy of the different wavelets is concentrated. When the wavelets are constructed with a one-dimensional wavelet of compact support, the resulting Fourier transforms have side lobes that appear in Figure 7.22.

---

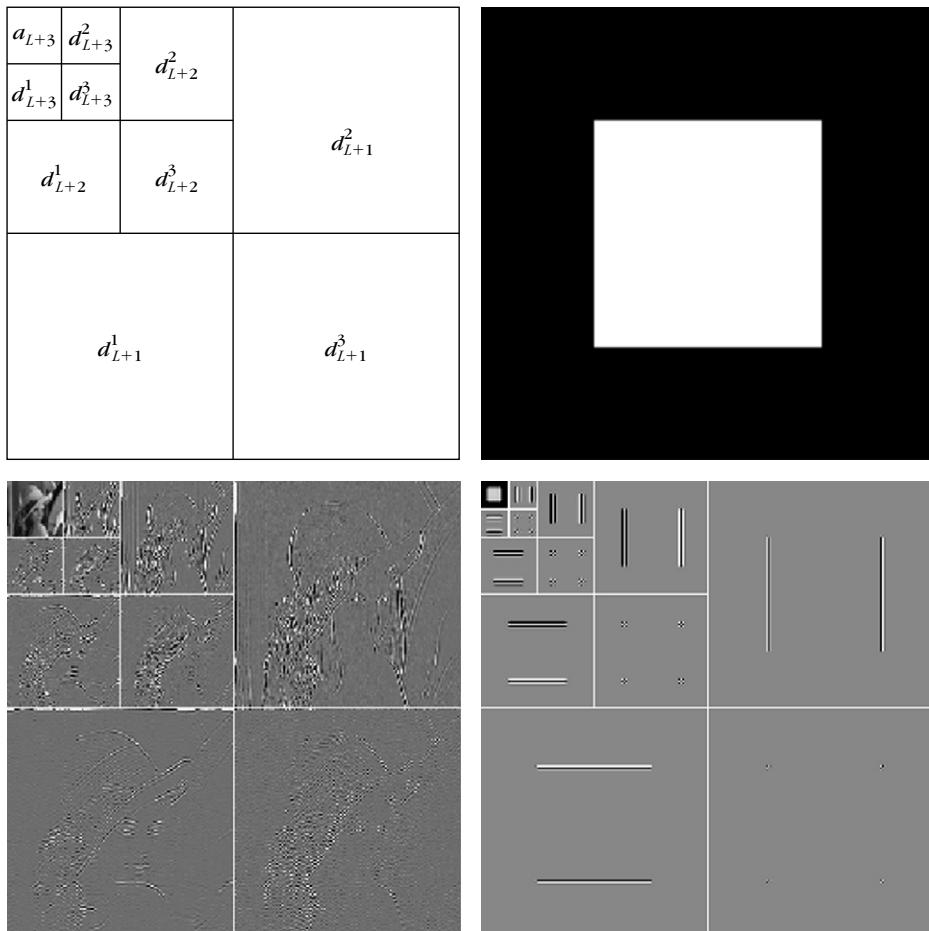
### EXAMPLE 7.17

Figure 7.24 gives two examples of wavelet transforms computed using separable Daubechies wavelets with  $p = 4$  vanishing moments. They are calculated with the filter bank algorithm from Section 7.7.3. Coefficients of large amplitude in  $d_j^1$ ,  $d_j^2$ , and  $d_j^3$  correspond, respectively, to vertical high frequencies (horizontal edges), horizontal high frequencies (vertical edges), and high frequencies in both directions (corners). Regions where the image intensity varies smoothly yield nearly zero coefficients, shown in gray in the figure. The large number of nearly zero coefficients makes it particularly attractive for compact image coding.

---

### Separable Biorthogonal Bases

One-dimensional biorthogonal wavelet bases are extended to separable biorthogonal bases of  $\mathbf{L}^2(\mathbb{R}^2)$  with the same approach as in Theorem 7.25. Let  $\phi, \psi$  and  $\tilde{\phi}$ ,

**FIGURE 7.24**

Separable wavelet transforms of the Lena image and of a white square in a black background, decomposed on 3 and 4 octaves, respectively. Black, gray, and white pixels correspond, respectively, to positive, zero, and negative wavelet coefficients. The disposition of wavelet image coefficients  $d_j^k[n, m] = \langle f, \psi_{j,n}^k \rangle$  is illustrated on the top left.

$\tilde{\psi}$  be two dual pairs of scaling functions and wavelets that generate biorthogonal wavelet bases of  $L^2(\mathbb{R})$ . The dual wavelets of  $\psi^1$ ,  $\psi^2$ , and  $\psi^3$  defined by (7.218) are

$$\tilde{\psi}^1(x) = \tilde{\phi}(x_1) \tilde{\psi}(x_2), \quad \tilde{\psi}^2(x) = \tilde{\psi}(x_1) \tilde{\phi}(x_2), \quad \tilde{\psi}^3(x) = \tilde{\psi}(x_1) \tilde{\psi}(x_2). \quad (7.224)$$

One can verify that

$$\left\{ \psi_{j,n}^1, \psi_{j,n}^2, \psi_{j,n}^3 \right\}_{(j,n) \in \mathbb{Z}^3} \quad (7.225)$$

and

$$\left\{ \tilde{\psi}_{j,n}^1, \tilde{\psi}_{j,n}^2, \tilde{\psi}_{j,n}^3 \right\}_{(j,n) \in \mathbb{Z}^3} \quad (7.226)$$

are biorthogonal Riesz bases of  $\mathbf{L}^2(\mathbb{R}^2)$ .

### 7.7.3 Fast Two-Dimensional Wavelet Transform

The fast wavelet transform algorithm presented in Section 7.3.1 is extended in two dimensions. At all scales  $2^j$  and for any  $n = (n_1, n_2)$ , we denote

$$a_j[n] = \langle f, \phi_{j,n}^2 \rangle \quad \text{and} \quad d_j^k[n] = \langle f, \psi_{j,n}^k \rangle \quad \text{for } 1 \leq k \leq 3.$$

For any pair of one-dimensional filters  $y[m]$  and  $z[m]$  we write the product filter  $yz[n] = y[n_1]z[n_2]$  and  $\bar{y}[m] = y[-m]$ . Let  $h[m]$  and  $g[m]$  be the conjugate mirror filters associated to the wavelet  $\psi$ .

The wavelet coefficients at the scale  $2^{j+1}$  are calculated from  $a_j$  with two-dimensional separable convolutions and subsamplings. The decomposition formulas are obtained by applying the one-dimensional convolution formulas (7.102) and (7.103) of Theorem 7.10 to the separable two-dimensional wavelets and scaling functions for  $n = (n_1, n_2)$ :

$$a_{j+1}[n] = a_j \star \bar{h}\bar{h}[2n], \quad (7.227)$$

$$d_{j+1}^1[n] = a_j \star \bar{h}\bar{g}[2n], \quad (7.228)$$

$$d_{j+1}^2[n] = a_j \star \bar{g}\bar{h}[2n], \quad (7.229)$$

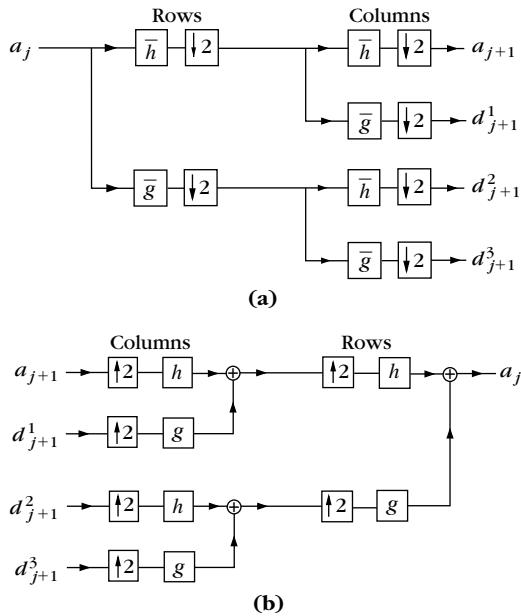
$$d_{j+1}^3[n] = a_j \star \bar{g}\bar{g}[2n]. \quad (7.230)$$

We showed in (3.70) that a separable two-dimensional convolution can be factored into one-dimensional convolutions along the rows and columns of the image. With the factorization illustrated in Figure 7.25(a), these four convolutions equations are computed with only six groups of one-dimensional convolutions. The rows of  $a_j$  are first convolved with  $\bar{h}$  and  $\bar{g}$  and subsampled by 2. The columns of these two output images are then convolved, respectively, with  $\bar{h}$  and  $\bar{g}$  and subsampled, which gives the four subsampled images  $a_{j+1}$ ,  $d_{j+1}^1$ ,  $d_{j+1}^2$ , and  $d_{j+1}^3$ .

We denote by  $\check{y}[n] = \check{y}[n_1, n_2]$  the image twice the size of  $y[n]$ , obtained by inserting a row of zeros and a column of zeros between pairs of consecutive rows and columns. The approximation  $a_j$  is recovered from the coarser-scale approximation  $a_{j+1}$  and the wavelet coefficients  $d_{j+1}^k$  with two-dimensional separable convolutions derived from the one-dimensional reconstruction formula (7.104)

$$a_j[n] = \check{a}_{j+1} \star hh[n] + \check{d}_{j+1}^1 \star hg[n] + \check{d}_{j+1}^2 \star gh[n] + \check{d}_{j+1}^3 \star gg[n]. \quad (7.231)$$

These four separable convolutions can also be factored into six groups of one-dimensional convolutions along rows and columns, illustrated in Figure 7.25(b).

**FIGURE 7.25**

**(a)** Decomposition of  $a_j$  with six groups of one-dimensional convolutions and subsamplings along the image rows and columns. **(b)** Reconstruction of  $a_j$  by inserting zeros between the rows and columns of  $a_{j+1}$  and  $d_{j+1}^k$ , and filtering the output.

Let  $b[n]$  be an input image with pixels at a distance  $2^L$ . We associate to  $b[n]$  a function  $f(x) \in \mathbf{V}_L^2$  approximated at the scale  $2^L$ . Its coefficients  $a_L[n] = \langle f, \phi_{L,n}^2 \rangle$  are defined like in (7.111) by

$$b[n] = 2^{-L} a_L[n] \approx f(2^L n). \quad (7.232)$$

The wavelet image representation of  $a_L$  is computed by iterating (7.227–7.230) for  $L \leq j < J$ :

$$\left[ a_J, \{d_j^1, d_j^2, d_j^3\}_{L \leq j \leq J} \right]. \quad (7.233)$$

The image  $a_L$  is recovered from this wavelet representation by computing (7.231) for  $J > j \geq L$ .

### Finite Image and Complexity

When  $a_L$  is a finite image of  $N = N_1 N_2$  pixels, we face boundary problems when computing the convolutions (7.227–7.231). Since the decomposition algorithm is separable along rows and columns, we use one of the three one-dimensional boundary techniques described in Section 7.5. The resulting values are decomposition

coefficients in a wavelet basis of  $\mathbf{L}^2[0, 1]^2$ . Depending on the boundary treatment, this wavelet basis is a periodic basis, a folded basis, or a boundary adapted basis.

For square images with  $N_1 = N_2$ , the resulting images  $a_j$  and  $d_j^k$  have  $2^{-2j}$  samples. Thus, the images of the wavelet representation (7.233) include a total of  $N$  samples. If  $h$  and  $g$  have size  $K$ , the reader can verify that  $2K2^{-2(j-1)}$  multiplications and additions are needed to compute the four convolutions (7.227–7.230) with the factorization of Figure 7.25(a). Thus, the wavelet representation (7.233) is calculated with fewer than  $8/3 KN$  operations. The reconstruction of  $a_L$  by factoring the reconstruction equation (7.231) requires the same number of operations.

### **Fast Biorthogonal Wavelet Transform**

The decomposition of an image in a biorthogonal wavelet basis is performed with the same fast wavelet transform algorithm. Let  $(\tilde{h}, \tilde{g})$  be the perfect reconstruction filters associated to  $(h, g)$ . The inverse wavelet transform is computed by replacing the filters  $(h, g)$  that appear in (7.231) by  $(\tilde{h}, \tilde{g})$ .

#### **7.7.4 Wavelet Bases in Higher Dimensions**

Separable wavelet orthonormal bases of  $\mathbf{L}^2(\mathbb{R}^p)$  are constructed for any  $p \geq 2$  with a procedure similar to the two-dimensional extension. Let  $\phi$  be a scaling function and  $\psi$  a wavelet that yields an orthogonal basis of  $\mathbf{L}^2(\mathbb{R})$ . We denote  $\theta^0 = \phi$  and  $\theta^1 = \psi$ . To any integer  $0 \leq \varepsilon < 2^p$  written in binary form  $\varepsilon = \varepsilon_1, \dots, \varepsilon_p$ , we associate the  $p$ -dimensional functions defined in  $x = (x_1, \dots, x_p)$  by

$$\psi^\varepsilon(x) = \theta^{\varepsilon_1}(x_1) \dots \theta^{\varepsilon_p}(x_p).$$

For  $\varepsilon = 0$ , we obtain a  $p$ -dimensional scaling function

$$\psi^0(x) = \phi(x_1) \dots \phi(x_p).$$

Nonzero indexes  $\varepsilon$  correspond to  $2^p - 1$  wavelets. At any scale  $2^j$  and for  $n = (n_1, \dots, n_p)$ , we denote

$$\psi_{j,n}^\varepsilon(x) = 2^{-pj/2} \psi^\varepsilon\left(\frac{x_1 - 2^j n_1}{2^j}, \dots, \frac{x_p - 2^j n_p}{2^j}\right).$$

**Theorem 7.26.** The family obtained by dilating and translating the  $2^p - 1$  wavelets for  $\varepsilon \neq 0$ ,

$$\left\{ \psi_{j,n}^\varepsilon \right\}_{1 \leq \varepsilon < 2^p, (j,n) \in \mathbb{Z}^{p+1}}, \quad (7.234)$$

is an orthonormal basis of  $\mathbf{L}^2(\mathbb{R}^p)$ .

The proof is done by induction on  $p$ . It follows the same steps as the proof of Theorem 7.25, which associates to a wavelet basis of  $\mathbf{L}^2(\mathbb{R})$  a separable wavelet basis of  $\mathbf{L}^2(\mathbb{R}^2)$ . For  $p = 2$ , we verify that the basis (7.234) includes three elementary wavelets. For  $p = 3$ , there are seven different wavelets.

### Fast Wavelet Transform

Let  $b[n]$  be an input  $p$ -dimensional discrete signal sampled at intervals  $2^L$ . We associate  $b[n]$  to an approximation  $f$  at the scale  $2^L$  with scaling coefficients  $a_L[n] = \langle f, \psi_{L,n}^0 \rangle$  that satisfy

$$b[n] = 2^{-Lp/2} a_L[n] \approx f(2^L n).$$

The wavelet coefficients of  $f$  at scales  $2^j > 2^L$  are computed with separable convolutions and subsamplings along the  $p$  signal dimensions. We denote

$$a_j[n] = \langle f, \psi_{j,n}^0 \rangle \quad \text{and} \quad d_j^\varepsilon[n] = \langle f, \psi_{j,n}^\varepsilon \rangle \quad \text{for } 0 < \varepsilon < 2^p.$$

The fast wavelet transform is computed with filters that are separable products of the one-dimensional filters  $h$  and  $g$ . The separable  $p$ -dimensional low-pass filter is

$$h^0[n] = h[n_1] \dots h[n_p].$$

Let us denote  $u^0[m] = h[m]$  and  $u^1[m] = g[m]$ . To any integer  $\varepsilon = \varepsilon_1, \dots, \varepsilon_p$  written in a binary form, we associate a separable  $p$ -dimensional band-pass filter

$$g^\varepsilon[n] = u^{\varepsilon_1}[n_1] \dots u^{\varepsilon_p}[n_p].$$

Let  $\bar{g}^\varepsilon[n] = g^\varepsilon[-n]$ . One can verify that

$$a_{j+1}[n] = a_j \star \bar{h}^0[2n], \tag{7.235}$$

$$d_{j+1}^\varepsilon[n] = a_j \star \bar{g}^\varepsilon[2n]. \tag{7.236}$$

We denote by  $\check{y}[n]$  the signal obtained by adding a zero between any two samples of  $y[n]$  that are adjacent in the  $p$ -dimensional lattice  $n = (n_1, \dots, n_p)$ . It doubles the size of  $y[n]$  along each direction. If  $y[n]$  has  $M^p$  samples, then  $\check{y}[n]$  has  $(2M)^p$  samples. The reconstruction is performed with

$$a_j[n] = \check{a}_{j+1} \star h^0[n] + \sum_{\varepsilon=1}^{2^p-1} \check{d}_{j+1}^\varepsilon \star g^\varepsilon[n]. \tag{7.237}$$

The  $2^p$  separable convolutions needed to compute  $a_j$  and  $\{d_j^\varepsilon\}_{1 \leq \varepsilon \leq 2^p}$  as well as the reconstruction (7.237) can be factored in  $2^{p+1} - 2$  groups of one-dimensional convolutions along the rows of  $p$ -dimensional signals. This is a generalization of the two-dimensional case, illustrated in Figure 7.25. The wavelet representation of  $a_L$  is

$$\left[ \{d_j^\varepsilon\}_{1 \leq \varepsilon \leq 2^p, L \leq j \leq J}, a_J \right]. \tag{7.238}$$

It is computed by iterating (7.235) and (7.236) for  $L \leq j < J$ . The reconstruction of  $a_L$  is performed with the partial reconstruction (7.237) for  $J > j \geq L$ .

If  $a_L$  is a finite signal of size  $N_1, \dots, N_p$ , the one-dimensional convolutions are modified with one of the three boundary techniques described in Section 7.5. The resulting algorithm computes decomposition coefficients in a separable wavelet

basis of  $\mathbf{L}^2[0, 1]^p$ . If  $N_1 = \dots = N_p$ , the signals  $a_j$  and  $d_j^e$  have  $2^{-pj}$  samples. Like  $a_L$ , the wavelet representation (7.238) is composed of  $N$  samples. If the filter  $h$  has  $K$  nonzero samples, then the separable factorization of (7.235) and (7.236) requires  $pK2^{-p(j-1)}$  multiplications and additions. Thus, the wavelet representation (7.238) is computed with fewer than  $p(1 - 2^{-p})^{-1}KN$  multiplications and additions. The reconstruction is performed with the same number of operations.

## 7.8 LIFTING WAVELETS

The lifting scheme, introduced by Sweldens [451, 452], factorizes orthogonal and biorthogonal wavelet transforms into elementary spatial operators called *liftings*. It has two main applications. The first one is an acceleration of the fast wavelet transform algorithm. The filter bank convolution and subsampling operations are factorized into elementary filterings on even and odd samples, which reduces the number of operations by nearly 2. Border treatments are also simplified. This is also called a *paraunitary filter bank implementation*. Readers mainly interested in this fast lifting implementation can skip directly to Section 7.8.5, which can be read independently.

The second application is the design of wavelets adapted to multidimensional bounded domains and surfaces, which is not possible with a Fourier transform approach. Section 7.8.1 introduces biorthogonal multiresolution and wavelet bases over nonstationary grids for arbitrary domains. The lifting construction of biorthogonal wavelet bases is explained in Section 7.8.2, with the resulting fast lifting wavelet transform. Section 7.8.3 describes an application to nonseparable quincunx wavelet bases for images. The construction of wavelet bases over bounded domains and surfaces is explained in Section 7.8.4, with computer graphics examples.

### 7.8.1 Biorthogonal Bases over Nonstationary Grids

The lifting scheme constructs wavelet bases over an arbitrary domain  $\Omega$  to represent functions of finite energy defined over  $\Omega$ . This section defines biorthogonal filters and wavelets that may be modified both in space and in scale to be adapted to the domain geometry. Section 7.8.2 explains the calculation of these filters and wavelets with a lifting scheme.

#### *Embedded Grids*

Biorthogonal multiresolutions, defined in Section 7.4, are generalized by considering nested spaces

$$\{0\} \subset \dots \subset \mathbf{V}_{j+1} \subset \mathbf{V}_j \subset \mathbf{V}_{j-1} \subset \dots \subset \mathbf{L}^2(\Omega),$$

which are defined over embedded approximation grids  $\mathcal{G}_j \subset \mathcal{G}_{j-1}$ . Each index  $n \in \mathcal{G}_j$  is associated to a sampling point  $x_n \in \Omega$ . Since the sampling grids are embedded,

this position does not change when the index is moved to finer grids  $n \in \mathcal{G}_k$  for  $k \leq j$ . Each space  $\mathbf{V}_j$  is equipped with a Riesz basis  $\{\phi_{j,n}\}_{n \in \mathcal{G}_j}$  parameterized by the approximation grid  $\mathcal{G}_j$ .

Embedded grids are decomposed with complementary grids  $\mathcal{C}_j$ :

$$\mathcal{G}_{j-1} = \mathcal{G}_j \cup \mathcal{C}_j.$$

For example, over the interval  $[0, 1]$ , the grid  $\mathcal{G}_{j-1}$  is the set of  $2^{j-1}n \in [0, 1]$  for  $0 \leq n < 2^{-j}$ . It is decomposed into the even grid points of  $2^{j-1}2n$  of  $\mathcal{G}_j$  and the odd grid points  $2^{j-1}(2n+1)$  of  $\mathcal{C}_j$ . A corresponding vector space decomposition is defined  $\mathbf{V}_{j-1} = \mathbf{V}_j \oplus \mathbf{W}_j$ , where the detail space  $\mathbf{W}_j$  has a Riesz basis of wavelets  $\{\psi_{j,m}\}_{m \in \mathcal{C}_j}$  indexed on the complementary grid  $\mathcal{C}_j$ .

A dual-biorthogonal wavelet family  $\{\tilde{\psi}_{j,m}\}_{m \in \mathcal{G}_j}$  and a dual basis of scaling functions  $\{\tilde{\phi}_{j,n}\}_{n \in \mathcal{G}_j}$  satisfy the biorthogonality conditions

$$\langle \phi_{j,n}, \tilde{\phi}_{j,n'} \rangle = \delta[n - n'] \quad \text{and} \quad \langle \psi_{j,m}, \tilde{\psi}_{j',m'} \rangle = \delta[j - j'] \delta[m - m']. \quad (7.239)$$

The resulting wavelet families  $\{\psi_{j,m}\}_{m \in \mathcal{G}_j, j \in \mathbb{Z}}$  and  $\{\tilde{\psi}_{j,m}\}_{m \in \mathcal{G}_j, j \in \mathbb{Z}}$  are biorthogonal wavelet bases of  $\mathbf{L}^2(\Omega)$ , which implies that for any  $f \in \mathbf{L}^2(\Omega)$ ,

$$f = \sum_{n \in \mathcal{G}_j} \langle f, \phi_{j,n} \rangle \tilde{\phi}_{j,n} + \sum_{j \geq J} \sum_{m \in \mathcal{C}_j} \langle f, \psi_{j,m} \rangle \tilde{\psi}_{j,m},$$

where  $2^J$  is an arbitrary coarsest scale. The difference with the biorthogonal wavelets of Section 7.4 is that scaling functions and wavelets are typically not translations and dilations of mother scaling functions and wavelets to be adapted to  $\Omega$ . As a result, the decomposition filters are not convolution filters.

### Spatially Varying Filters

The spatially varying filters associated to this biorthogonal multiresolution satisfy

$$\phi_{j,n} = \sum_{l \in \mathcal{G}_{j-1}} h_j[n, l] \phi_{j-1,l} \quad \text{and} \quad \psi_{j,m} = \sum_{l \in \mathcal{G}_{j-1}} g_j[m, l] \phi_{j-1,l}. \quad (7.240)$$

Over a translation-invariant domain,  $h_j[n, l] = h[n - 2l]$  are the perfect reconstruction filters of Section 7.3.2. Dual filters  $\tilde{h}$  and  $\tilde{g}$  are defined similarly by

$$\tilde{\phi}_{j,n} = \sum_{l \in \mathcal{G}_{j-1}} \tilde{h}_j[n, l] \tilde{\phi}_{j-1,l} \quad \text{and} \quad \tilde{\psi}_{j,m} = \sum_{l \in \mathcal{G}_{j-1}} \tilde{g}_j[m, l] \tilde{\phi}_{j-1,l}. \quad (7.241)$$

The biorthogonality relations (7.239) between wavelets and scaling functions imply equivalent biorthogonality filter relations for all  $n, n' \in \mathcal{G}_j$  and  $m, m' \in \mathcal{C}_j$ :

$$\sum_{l \in \mathcal{G}_{j-1}} g_j[m, l] \tilde{g}_j[m', l] = \delta[m - m'] \quad \sum_{l \in \mathcal{G}_{j-1}} h_j[n, l] \tilde{h}_j[n', l] = \delta[n - n'], \quad (7.242)$$

$$\sum_{l \in \mathcal{G}_{j-1}} g_j[m, l] \tilde{h}_j[n, l] = 0 \quad \sum_{l \in \mathcal{G}_{j-1}} h_j[n, l] \tilde{g}_j[m, l] = 0. \quad (7.243)$$

Wavelets and scaling functions can be written as an infinite product of these filters. If these products converge in  $L^2(\Omega)$  and the filters that satisfy (7.242) and (7.243), then the resulting wavelets and scaling functions define biorthogonal bases of  $L^2(\Omega)$ , which satisfy (7.239) [452].

To simplify notations, the filters  $h_j$  and  $g_j$  are also written as matrices that transform discrete vector

$$\forall a \in \mathbb{C}^{|\mathcal{G}_{j-1}|}, \quad \forall n \in \mathcal{G}_j, (H_j a)[n] = \sum_{l \in \mathcal{G}_{j-1}} h_j[n, l] a[l]$$

$$\forall a \in \mathbb{C}^{|\mathcal{G}_{j-1}|}, \quad \forall m \in \mathcal{C}_j, (G_j a)[m] = \sum_{l \in \mathcal{G}_{j-1}} g_j[m, l] a[l]$$

and similarly for the dual matrices  $\tilde{H}_j$  and  $\tilde{G}_j$ .

The biorthogonality conditions (7.242) are rewritten as

$$\begin{bmatrix} \tilde{H}_j \\ \tilde{G}_j \end{bmatrix} \begin{bmatrix} H_j^* & G_j^* \end{bmatrix} = \begin{bmatrix} \text{Id}_{\mathcal{G}_j} & 0 \\ 0 & \text{Id}_{\mathcal{C}_j} \end{bmatrix}, \quad (7.244)$$

where  $A^*$  is the complex transpose of a matrix  $A$ .

### ***Vanishing Moments***

Wavelets with  $p_1$  vanishing moments are orthogonal to polynomials of a degree strictly smaller than  $p_1$ . Let  $d_1$  be the dimension of the space of polynomials of degree  $q - 1$  in dimension  $d$ . If  $d = 1$ , then  $d_1 = p_1$ , and if  $d = 2$ , then  $d_1 = p_1(p_1 + 1)/2$ . Such wavelets are orthogonal to a basis  $\{q^{(k)}\}_{0 \leq k < d_1}$  of this polynomial space, defined over  $\Omega \subset \mathbb{R}^d$ :

$$\forall j, \forall m \in \mathcal{C}_j, \forall k < d_1, \quad \int_{\Omega} \psi_{j,m}(x) q^{(k)}(x) dx = 0, \quad (7.245)$$

and similarly for dual wavelets with  $p_2$  vanishing moments,

$$\forall j, \forall m \in \mathcal{C}_j, \forall k < d_2, \quad \int_{\Omega} \tilde{\psi}_{j,m}(x) q^{(k)}(x) dx = 0.$$

Lifting steps are used to increase the number of vanishing moments.

### **7.8.2 Lifting Scheme**

The lifting scheme builds filters over arbitrary domains  $\Omega$  as a succession of elementary lifting steps applied to lazy wavelets that are Diracs. Each lifting step transforms a family of biorthogonal filters into new biorthogonal filters that define wavelets with more vanishing moments.

### ***Lazy Wavelets***

A lifting begins from lazy wavelets, which are Diracs on grid points. The lazy discrete orthogonal wavelet transform just splits the coefficients of a grid  $\mathcal{G}_{j-1}$  into the two

subgrids  $\mathcal{G}_j$  and  $\mathcal{C}_j$ . It corresponds to filters that are Diracs on these grids:

$$\forall l \in \mathcal{G}_{j-1}, \forall n \in \mathcal{G}_j, \forall m \in \mathcal{C}_j, \quad \begin{cases} h_j^o[n, l] = \tilde{h}_j^o[n, l] = \delta[n - l], \\ g_j^o[m, l] = \tilde{g}_j^o[m, l] = \delta[m - l]. \end{cases}$$

For  $a \in \mathbb{C}^{|\mathcal{G}_{j-1}|}$ , the vector  $H_j^o a \in \mathbb{C}^{|\mathcal{G}_j|}$  is the restriction of  $a$  to  $\mathcal{G}_j$ , and  $G_j^o a \in \mathbb{C}^{|\mathcal{C}_j|}$  is the restriction of  $a$  to  $\mathcal{C}_j$ .

Since each index  $n \in \mathcal{G}_j$  is associated to a sampling point  $x_n \in \Omega$  that does not depend on the scale index  $j$ , one can verify that  $\tilde{\psi}_{j,m}^o(x) = \psi_{j,m}^o = \delta(x - x_m)$  and  $\tilde{\phi}_{j,n}^o(x) = \phi_{j,n}^o = \delta(x - x_n)$ , meaning that for any continuous function  $f(x)$ :

$$\int_{\Omega} f(x) \psi_{j,m}^o(x) dx = \int_{\Omega} f(x) \tilde{\psi}_{j,m}^o(x) dx = f(x_m).$$

This lazy wavelet basis is improved with a succession of liftings.

### ***Increasing Vanishing Moments, Stability, and Regularity***

A lifting modifies biorthogonal filters in order to increase the number of vanishing moments of the resulting biorthogonal wavelets, and hopefully their stability and regularity.

Increasing the number of vanishing moments decreases the amplitude of wavelet coefficients in regions where the signal is regular, which produces a more sparse representation. However, increasing the number of vanishing moments with a lifting also increases the wavelet support, which is an adverse effect that increases the number of large coefficients produced by isolated singularities.

Each lifting step maintains the filter biorthogonality but provides no control on the Riesz bounds and thus on the stability of the resulting wavelet biorthogonal basis. When a basis is orthogonal then the dual basis is equal to the original basis. Having a dual basis that is similar to the original basis is therefore an indication of stability. As a result, stability is generally improved when dual wavelets have as much vanishing moments as original wavelets and a support of similar size. This is why a lifting procedure also increases the number of vanishing moments of dual wavelets. It can also improve the regularity of the dual wavelet.

A lifting design is computed by adjusting the number of vanishing moments. The stability and regularity of the resulting biorthogonal wavelets are measured a posteriori, hoping for the best. This is the main weakness of this wavelet design procedure.

### ***Prediction***

Starting from an initial set of biorthogonal filters  $\{h_j^o, g_j^o, \tilde{h}_j^o, \tilde{g}_j^o\}_j$ , a prediction step modifies each filter  $g_j^o$  to increase the number of vanishing moments of  $\psi_{j,n}$ . This is done with an operator  $P_j$  that predicts the values in the grid  $\mathcal{C}_j$  from samples in the grid  $\mathcal{G}_j$ :

$$\forall a \in \mathbb{C}^{|\mathcal{G}_j|}, \forall m \in \mathcal{C}_j, (P_j a)[m] = \sum_{n \in \mathcal{G}_j} p_j[m, n] a[n].$$

The number of vanishing moments of  $\psi_{j,n}$  is increased by modifying the filter  $g_j^o$  with this predictor

$$h_j = h_j^o \quad \text{and} \quad g_j[m, l] = g_j^o[m, l] - \sum_{n \in \mathcal{G}_j} p_j[m, n] h_j^o[n, l]. \quad (7.246)$$

Biorthogonality is maintained by also modifying the dual filter  $\tilde{h}_j^o$ :

$$\tilde{g}_j = \tilde{g}_j^o \quad \text{and} \quad \tilde{h}_j[n, l] = \tilde{h}_j^o[n, l] + \sum_{m \in \mathcal{C}_j} p_j[m, n] \tilde{g}_j^o[m, l].$$

The filter lifting (7.246) implies a retransformation of the scaling and wavelet coefficients computed with the original filters  $h_j^o$  and  $g_j^o$ . The lifted scaling coefficients  $\{a_j[n] = \langle f, \phi_{j,n} \rangle\}_{n \in \mathcal{G}_j}$  and detail coefficients  $\{d_j[m] = \langle f, \psi_{j,m} \rangle\}_{m \in \mathcal{C}_j}$  are computed from the coefficients  $\{d_j^o[m], a_j^o[n]\}_{n \in \mathcal{G}_j, m \in \mathcal{C}_j}$  corresponding to  $h_j^o$  and  $g_j^o$  by applying the predict operator

$$\forall m \in \mathcal{C}_j, \quad d_j[m] = d_j^o[m] - \sum_{n \in \mathcal{G}_j} p_j[m, n] a_j^o[n],$$

while the scaling coefficients are not modified:  $a_j[n] = a_j^o[n]$ . If  $P_j$  is a good predictor of  $d_j^o[m]$  from  $a_j^o[n]$  on the coarse grid  $\mathcal{G}_j$ , then the resulting coefficients  $d_j[m]$  are smaller, which is an indication that the wavelet has more vanishing moments.

The prediction (7.246) of the filters is rewritten with matrix notations

$$\begin{cases} H_j = H_j^o \\ G_j = G_j^o - P_j H_j^o \end{cases} \quad \text{and} \quad \begin{cases} \tilde{H}_j = \tilde{H}_j^o + P_j^* \tilde{G}_j^o \\ \tilde{G}_j = \tilde{G}_j^o. \end{cases} \quad (7.247)$$

Since  $H_j = H_j^o$  is not modified, the scaling functions  $\phi_{j,n} = \phi_{j,n}^o$  are not modified. In contrast, since  $\tilde{H}_j^o$  is modified, both the dual-scaling and wavelet functions are modified:

$$\phi_{j,n} = \phi_{j,n}^o, \quad \psi_{j,m} = \psi_{j,m}^o - \sum_{n \in \mathcal{G}_j} p_j[m, n] \phi_{j,n}^o, \quad (7.248)$$

$$\tilde{\phi}_{j,n} = \sum_{l \in \mathcal{G}_{j-1}} h_j^o[n, l] \tilde{\phi}_{j-1,l} + \sum_{m \in \mathcal{C}_j} p_j[m, n] \tilde{\psi}_{j,m}, \quad \tilde{\psi}_{j,m} = \sum_{l \in \mathcal{G}_{j-1}} g_j^o[m, l] \tilde{\phi}_{j-1,l}. \quad (7.249)$$

Theorem 7.27 proves that this lifting step maintains the biorthogonality conditions [451].

**Theorem 7.27:** *Sweldens.* The prediction (7.247) transforms the biorthogonal filters  $\{H_j^o, G_j^o, \tilde{H}_j^o, \tilde{G}_j^o\}$  into a new set of biorthogonal filters  $\{H_j, G_j, \tilde{H}_j, \tilde{G}_j\}$ .

**Proof.** The lifting step (7.247) is written in matrix notation as

$$\begin{bmatrix} H_j \\ G_j \end{bmatrix} = \begin{bmatrix} \text{Id}_{\mathcal{G}_j} & 0 \\ -P_j & \text{Id}_{\mathcal{C}_j} \end{bmatrix} \begin{bmatrix} H_j^o \\ G_j^o \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \tilde{H}_j \\ \tilde{G}_j \end{bmatrix} = \begin{bmatrix} \text{Id}_{\mathcal{G}_j} & P_j^* \\ 0 & \text{Id}_{\mathcal{C}_j} \end{bmatrix} \begin{bmatrix} \tilde{H}_j^o \\ \tilde{G}_j^o \end{bmatrix}.$$

The proof of the biorthogonality relation (7.244) follows from

$$\begin{bmatrix} \text{Id}_{\mathcal{G}_j} & P_j \\ 0 & \text{Id}_{\mathcal{C}_j} \end{bmatrix} \begin{bmatrix} \text{Id}_{\mathcal{G}_j} & -P_j \\ 0 & \text{Id}_{\mathcal{C}_j} \end{bmatrix} = \begin{bmatrix} \text{Id}_{\mathcal{G}_j} & 0 \\ 0 & \text{Id}_{\mathcal{C}_j} \end{bmatrix}. \quad \blacksquare$$

To increase the number of vanishing moments of  $\psi_{j,m}$ , and get  $\int_{\Omega} \psi_{j,m}(x) q^{(k)}(x) dx = 0$  for a basis of  $d_1$  polynomial of degree  $p_1$ , (7.248) shows that predict coefficients must satisfy

$$\int_{\Omega} \psi_{j,m}^o(x) q^{(k)}(x) dx = \sum_{n \in \mathcal{G}_j} p_j[m, n] \int_{\Omega} \phi_{j,n}^o(x) q^{(k)}(x) dx \quad \text{for } 0 \leq k < d_1. \quad (7.250)$$

The predictor  $\{p_j[m, n]\}_n$  can be chosen to have  $d_1$  nonzero coefficients that solve this  $d_1 \times d_1$  linear system for each  $m$ .

### **Update**

The prediction (7.248) modifies  $\psi_{j,n}$  but does not change  $\tilde{\psi}_{j,n}$ . The roles of  $\psi_{j,m}$  and  $\tilde{\psi}_{j,m}$  are reversed by applying a lifting step to increase the number of vanishing moments of  $\tilde{\psi}_{j,m}$  as well. The goal is to obtain dual wavelets  $\tilde{\psi}_{j,n}$  that are as similar as possible to  $\psi_{j,n}$  in order to improve the stability of the basis. It requires the use of an *update* operator  $U_j$  defined by

$$\forall b \in \mathbb{C}^{|\mathcal{C}_j|}, \forall n \in \mathcal{G}_j, (U_j b)[n] = \sum_{m \in \mathcal{C}_j} u_j[n, m] b[m].$$

The update step is then

$$\begin{cases} H_j = H_j^o + U_j G_j^o \\ G_j = G_j^o \end{cases} \quad \text{and} \quad \begin{cases} \tilde{H}_j = \tilde{H}_j^o \\ \tilde{G}_j = \tilde{G}_j^o - U_j^* \tilde{H}_j^o. \end{cases} \quad (7.251)$$

Since predict and update steps are equivalent operations on dual filters, Theorem 7.27 shows that this update operation defines new filters that remain biorthogonal.

Let  $\{d_j^o[m], a_j^o[n]\}_{n \in \mathcal{G}_j, m \in \mathcal{C}_j}$  be the wavelet and scaling coefficients corresponding to the filters  $h_j^o$  and  $g_j^o$ . The wavelet coefficients are not modified  $d_j[n] = d_j^o[n]$ , and  $\{a_j[n]\}_{n \in \mathcal{G}_j}$  is computed by applying the update operator

$$\forall n \in \mathcal{G}_j, a_j[n] = a_j^o[n] + \sum_{m \in \mathcal{C}_j} u_j[n, m] d_j^o[m].$$

The updated scaling functions and wavelets are:

$$\phi_{j,n} = \sum_{l \in \mathcal{G}_{j-1}} h_j^o[n, l] \phi_{j-1,l} + \sum_{m \in \mathcal{C}_j} u_j[n, m] \psi_{j,m}, \quad \psi_{j,m} = \sum_{l \in \mathcal{G}_{j-1}} g_j^o[m, l] \phi_{j-1,l}, \quad (7.252)$$

$$\tilde{\phi}_{j,n} = \tilde{\phi}_{j,n}^o, \quad \tilde{\psi}_{j,m} = \tilde{\psi}_{j,m}^o - \sum_{n \in \mathcal{G}_j} u_j[n, m] \tilde{\phi}_{j,n}^o. \quad (7.253)$$

Theorem 7.28 proves that this update does not modify the number of vanishing moments of the analyzing wavelet.

**Theorem 7.28.** If the wavelets  $\{\psi_{j,m}^o\}_{j,m}$  have  $p_1$  vanishing moments, then the wavelets  $\{\psi_{j,m}\}_{j,m}$  obtained with the update (7.252) have  $p_1$  vanishing moments.

**Proof.** Equation (7.253) shows that  $\tilde{\phi}_{j,n}^o = \tilde{\phi}_{j,n}$ . Since the original multiresolution has  $p_1$  vanishing moments, it is orthogonal to a basis of  $d_1$  polynomials  $\{q^{(k)}\}_{0 \leq k < d_1}$  of degree  $q - 1$  as defined in (7.245):

$$\forall k < d_1, \quad q^{(k)} = \sum_{n \in \mathcal{G}_j} \langle q^{(k)}, \phi_{j,n}^o \rangle \tilde{\phi}_{j,n}^o = \sum_{n \in \mathcal{G}_j} \langle q^{(k)}, \phi_{j,n}^o \rangle \tilde{\phi}_{j,n}.$$

Taking the inner product of this relation with each  $\phi_{j,n'}$  leads to

$$\forall k < d_1, \quad \langle q^{(k)}, \phi_{j,n'} \rangle = \sum_{n \in \mathcal{G}_j} \langle q^{(k)}, \phi_{j,n}^o \rangle \langle \tilde{\phi}_{j,n}, \phi_{j,n'} \rangle = \langle q^{(k)}, \phi_{j,n'}^o \rangle. \quad (7.254)$$

Using the refinement equation (7.252) gives

$$\begin{aligned} \forall k < d_1, \quad \langle \psi_{j,m}, q^{(k)} \rangle &= \sum_{l \in \mathcal{G}_{j-1}} g_j^o[m, l] \langle \phi_{j-1,l}, q^{(k)} \rangle = \sum_{l \in \mathcal{G}_{j-1}} g_j^o[m, l] \langle \phi_{j-1,l}^o, q^{(k)} \rangle \\ &= \langle \sum_{l \in \mathcal{G}_{j-1}} g_j^o[m, l] \phi_{j-1,l}^o, q^{(k)} \rangle = \langle \psi_{j,m}^o, q^{(k)} \rangle = 0, \end{aligned}$$

where we used  $\langle \phi_{j-1,l}, q^{(k)} \rangle = \langle \phi_{j-1,l}^o, q^{(k)} \rangle$ , which follows from (7.254). ■

To increase the number of vanishing moments of  $\tilde{\psi}_{j,m}$  and to get  $\int_{\Omega} \tilde{\psi}_{j,m}(x) q^{(k)}(x) dx = 0$  for a basis of  $d_2$  polynomials of degree  $p_2$  in  $\Omega$ , (7.252) shows that update coefficients must satisfy

$$\forall k < d_2, \quad \int_{\Omega} \tilde{\psi}_{j,m}^o(x) q^{(k)}(x) dx = \sum_{n \in \mathcal{G}_j} u_j[n, m] \int_{\Omega} \tilde{\phi}_{j,n}^o(x) q^{(k)}(x) dx. \quad (7.255)$$

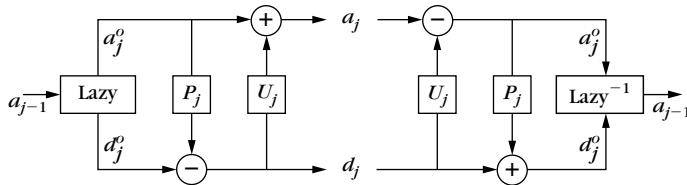
Thus, the update coefficients  $\{u_j[m, n]\}_n$  can be chosen to have  $d_2$  nonzero coefficients, which solves this  $d_2 \times d_2$  linear system for each  $m$ .

### Predict plus Update Design and Algorithm

Wavelets synthesized with a lifting are constructed with one predict step followed by one update step, because there is no technique that controls the wavelet stability and regularity over more lifting steps. Beginning from Dirac lazy wavelets  $\{\psi_{j,m}^o, \tilde{\psi}_{j,m}^o\}_{j,m}$  with no vanishing moment, a prediction (7.247) obtains biorthogonal wavelets  $\{\psi_{j,m}^1, \tilde{\psi}_{j,m}^1\}_{j,m}$  with, respectively,  $p_1$  and zero vanishing moments. An update (7.251) then yields biorthogonal wavelets  $\{\psi_{j,m}, \tilde{\psi}_{j,m}\}_{j,m}$  having, respectively,  $p_1$  and  $p_2$  vanishing moments.

A fast wavelet transform computes the coefficients

$$\forall n \in \mathcal{G}_j, \quad a_j[n] = \langle f, \phi_{j,n} \rangle \quad \text{and} \quad \forall m \in \mathcal{C}_j, \quad d_j[m] = \langle f, \psi_{j,m} \rangle$$

**FIGURE 7.26**

Predict and update decomposition of  $a_{j-1}$  into  $a_j$  and  $d_j$ , followed by the reconstruction of  $a_{j-1}$  with the same update and predict operators.

by replacing convolution with conjugate mirror filters by a succession of lifting and update steps. The algorithm takes in input a discrete signal  $a_L \in \mathbb{C}^{|\mathcal{G}_L|}$  of length  $N = |\mathcal{G}_L|$ , and applies the lazy decomposition, a predict, and an update operator, as illustrated in Figure 7.26. For  $j = L + 1, \dots, J$ , it computes

1. *Split:*  $\forall m \in \mathcal{C}_j, d_j^o[m] = a_{j-1}[m], \quad \forall n \in \mathcal{G}_j, a_j^o[n] = a_{j-1}[n]$ .
2. *Forward predict:*  $\forall m \in \mathcal{C}_j, \quad d_j[m] = d_j^o[m] - \sum_{n \in \mathcal{G}_j} p_j[m, n] a_j^o[n]$ .
3. *Forward update:*  $\forall n \in \mathcal{G}_j, \quad a_j[n] = a_j^o[n] + \sum_{m \in \mathcal{C}_j} u_j[n, m] d_j[m]$ .

This fast biorthogonal wavelet transform (7.157) requires  $O(N)$  operations.

The reconstruction of  $a_L$  from  $\{d_j\}_{J \leq j < L}$  and  $a_J$  inverts these predict and update steps. For  $j = J, \dots, L + 1$ , it computes

1. *Backward update:*  $\forall n \in \mathcal{G}_j, \quad a_j^o[n] = a_j[n] - \sum_{m \in \mathcal{C}_j} u_j[n, m] d_j[m]$ .
2. *Backward predict:*  $\forall m \in \mathcal{C}_j, \quad d_j^o[m] = d_j[m] + \sum_{n \in \mathcal{G}_j} p_j[m, n] a_j^o[n]$ .
3. *Merge:*  $\forall m \in \mathcal{C}_j, \quad a_{j-1}[m] = d_j^o[m], \quad \forall n \in \mathcal{G}_j, \quad a_{j-1}[n] = a_j^o[n]$ .

### **Vanishing Moments**

The predict and update filters are computed to create vanishing moments on the resulting wavelets. After the prediction applied to the lazy Diracs  $\psi_{j,m}^0(x) = \delta(x - x_m), \phi_{j,n}^0(x) = \delta(x - x_n)$ , the resulting wavelets and scaling functions derived from (7.248) and (7.249) are:

$$\phi_{j,n}^1(x) = \delta(x - x_n), \quad \psi_{j,m}^1(x) = \delta(x - x_m) - \sum_{n \in \mathcal{G}_j} p_j[m, n] \delta(x - x_n) \quad (7.256)$$

$$\tilde{\phi}_{j,n}^1(x) = \delta(x - x_n) + \sum_{m \in \mathcal{C}_j} p_j[m, n] \delta(x - x_m), \quad \tilde{\psi}_{j,m}^1(x) = \delta(x - x_m). \quad (7.257)$$

According to (7.250), the wavelet  $\psi_{j,m}^1$  has  $p_1$  vanishing moments if for each  $m$ , the  $p_1$  coefficients of  $\{p_j[m, n]\}_n$  solve the  $d_1 \times d_1$  linear system:

$$\forall m \in \mathcal{C}_j, \forall k < d_1, \quad q^{(k)}(x_m) = \sum_{n \in \mathcal{G}_j} p_j[m, n] q^{(k)}(x_n). \quad (7.258)$$

Following (7.253), after the update of the dual wavelet and the scaling functions, are

$$\tilde{\phi}_{j,n}^2 = \tilde{\phi}_{j,n}^1 = \tilde{\phi}_{j-1,n}^2 + \sum_{m \in \mathcal{C}_j} p_j[m, n] \tilde{\phi}_{j-1,m}^2 \quad (7.259)$$

$$\tilde{\psi}_{j,m}^2 = \tilde{\psi}_{j,m}^1 - \sum_{n \in \mathcal{G}_j} u_j[n, m] \tilde{\phi}_{j,n}^1 = \tilde{\phi}_{j-1,m}^2 - \sum_{n \in \mathcal{G}_j} u_j[n, m] \tilde{\phi}_{j,m}^2. \quad (7.260)$$

Theorem 7.28 proves that the vanishing moments of  $\psi_{j,m}^1$  are transferred to  $\psi_{j,m}^2$  after the dual lifting step. According to (7.258), the dual wavelet  $\tilde{\psi}_{j,m}^2$  has  $p_2$  vanishing moments if for each  $m$  the  $d_2$  coefficients of  $\{u_j[n, m]\}_n$  solve the  $d_2 \times d_2$  linear system:

$$\forall k < d_2, \quad \sum_{n \in \mathcal{G}_j} I_j^k(n) u_j[n, m] = I_{j-1}^k(m) \quad \text{with} \quad I_j^k(n) = \langle \tilde{\phi}_{j,n}, q^{(k)} \rangle. \quad (7.261)$$

The inner products  $I_j^k(n)$  are computed iteratively with (7.259):

$$I_j^k(n) = I_{j-1}^k(n) + \sum_{m \in \mathcal{C}_j} p_j[m, n] I_{j-1}^k(m). \quad (7.262)$$

The recurrence is initialized at the finest scale  $j = L$  by setting  $I_L^k(n) = q^{(k)}(x_n)$ , where  $x_n \in \Omega$  is the point associated to the index  $n \in \mathcal{G}_L$ . More elaborated initializations using quadrature formula can also be used [427].

### **Linear Splines 5/3 Biorthogonal Wavelets**

Linear spline wavelets are obtained with a two-step lifting construction beginning from lazy wavelets. The one-dimensional grid  $\mathcal{G}_{j-1}$  is a uniform sampling at intervals  $2^{j-1}$  and the two subgrids  $\mathcal{G}_j$  and  $\mathcal{C}_j$  correspond to even and odd subsampling. Figure 7.27 illustrates these embedded one-dimensional grids.

The lazy wavelet transform splits the coefficients  $a_{j-1}$  into two groups

$$\forall n \in \mathcal{G}_j, \quad a_j^o[n] = a_{j-1}[n], \quad \text{and} \quad \forall m \in \mathcal{C}_j, \quad d_j^o[m] = a_{j-1}[m].$$

The value of  $d_j^o$  in  $\mathcal{C}_j$  is predicted with a linear interpolation of neighbor values in  $\mathcal{G}_j$

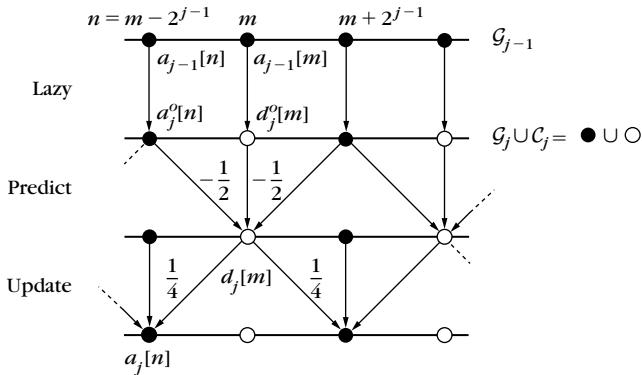
$$\forall m \in \mathcal{C}_j, \quad d_j^o[m] = d_j^o[m] - \frac{a_j^o[n + 2^{j-1}] + a_j^o[n - 2^{j-1}]}{2}. \quad (7.263)$$

This lifting step creates wavelets with two vanishing moments because this linear interpolation predicts exactly the values of polynomials of degree 0 and 1.

A symmetric update step computes

$$\begin{aligned} \forall n \in \mathcal{G}_j, \quad a_j[n] = & a_j^o[n] + (u_j[n, m - 2^{j-1}] d_j[m - 2^{j-1}] \\ & + u_j[n, m + 2^{j-1}] d_j[m + 2^{j-1}]). \end{aligned} \quad (7.264)$$

To obtain two vanishing moments, the inner products  $I_j^k(n) = \langle \tilde{\phi}_{j,n}(t), t^k \rangle$  are computed iteratively with (7.262), using two nonzero update coefficients  $u_j[n, m -$

**FIGURE 7.27**

Predict and update steps for the construction of linear spline wavelets.

$2^{j-1}]$  and  $u_j[n, m + 2^{j-1}]$  for each  $m$ . For  $k=0$  we get  $I_j^0(n) = \langle \tilde{\phi}_{j,n}, 1 \rangle = 2^{j-L}$ . Since  $t$  is antisymmetric, if this equation is valid for  $k=0$  and if  $u_j[n, m - 2^{j-1}] = u_j[n, m + 2^{j-1}]$ , then it is valid for  $k=1$ . Solving (7.261) for  $k=0$  gives  $u_j[n, m - 2^{j-1}] = u_j[n, m + 2^{j-1}] = 1/4$  and thus,

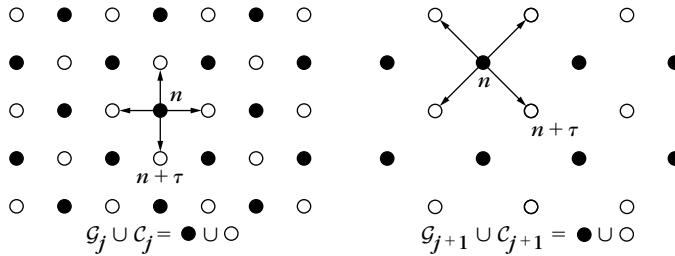
$$a_j[n] = a_j^o[n] + \frac{1}{4}(d_j[m - 2^{j-1}] + d_j[m + 2^{j-1}]). \quad (7.265)$$

Figure 7.27 illustrates the succession of predict and update. One can verify (Exercise 7.20) that the resulting biorthogonal wavelets correspond to the spline biorthogonal wavelets computed with  $p_1 = p_2 = 2$  vanishing moments (shown in Figure 7.15). The dual-scaling functions and wavelets are compactly supported linear splines. Higher-order biorthogonal spline wavelets are constructed with a prediction (7.263) and an update (7.264) providing more vanishing moments.

### 7.8.3 Quincunx Wavelet Bases

Separable two-dimensional wavelet bases are constructed in Section 7.7.2 from one-dimensional wavelet bases. They are implemented with separable filter banks that increase the scale by 2, by dividing the image grid in a coarse grid that keeps one point out of four, plus three detail grids of the same size and that correspond to three different wavelets. Other regular subsamplings of the image array lead to nonseparable wavelet bases. A quincunx subsampling divides the image grid into a coarse grid that keeps one point out of two and a detail grid of the same size that corresponds to a quincunx wavelet. Thus, the scale increases only by a factor  $2^{1/2}$ . A quincunx wavelet is more isotropic than separable wavelets.

Biorthogonal or orthogonal quincunx wavelets are constructed with perfect reconstruction or conjugate mirror filters, defined with a quincunx subsampling, which yields conditions on their transfer functions [170, 253, 254, 431]. Kovačević

**FIGURE 7.28**

Two successive quincunx subsampling for  $j$  and  $j + 1$ , where  $j$  is odd.

and Sweldens [333] give a simple construction of biorthogonal quincunx wavelets from lazy wavelets, with a predict followed by an update lifting.

We denote by  $(a, b)^*$  a transposed two-dimensional vector column. Embedded quincunx grids are defined by

$$\forall j, \quad \begin{cases} \mathcal{G}_j = L^j \mathbb{Z}^2 = \{L^j(n_1, n_2)^* : n \in \mathbb{Z}^2\} \\ \mathcal{C}_j = L^j \mathbb{Z}^2 + L^{j-1} e_1 = \mathcal{G}_j + L^{j-1} e_1, \end{cases}$$

where  $L = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$ ,  $e_1 = (1, 0)^*$  and  $e_2 = (0, 1)^*$ .

Figure 7.28 shows two quincunx subsampled grids, depending on the parity of  $j$ .

Each point  $n \in \mathcal{G}_j$  and  $m \in \mathcal{C}_j$  have four neighbors

$$\{n + \tau\}_{\Delta_j} \subset \mathcal{C}_j \quad \text{and} \quad \{m + \tau\}_{\Delta_j} \subset \mathcal{G}_j,$$

where the set of shifts has the following four elements:

$$\Delta_j = \{\pm L^{j-1} e_1, \pm L^{j-1} e_2\}.$$

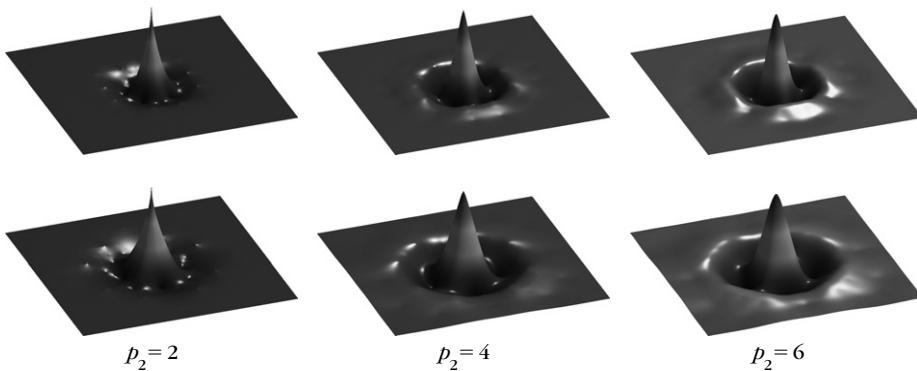
The simplest symmetric prediction operator on these grids is the symmetric averaging on the four neighbors, which performs a linear interpolation:

$$\forall a \in \mathbb{C}^{|\mathcal{G}_j|}, \quad \forall m \in \mathcal{C}_j, \quad (P_j a)[m] = \frac{1}{4} \sum_{\tau \in \Delta_j} a[m + \tau]. \quad (7.266)$$

This prediction operator applied to lazy wavelets yields a wavelet that is orthogonal to constant and linear polynomial in  $\mathbb{R}^2$ , which gives  $p_1 = 2$  vanishing moments. The update operator is defined with the same symmetry on the four neighbors

$$\forall b \in \mathbb{C}^{|\mathcal{C}_j|}, \quad \forall n \in \mathcal{G}_j, \quad (U_j b)[n] = \lambda \sum_{\tau \in \Delta_j} b[n + \tau]. \quad (7.267)$$

Choosing  $\lambda = 1/4$  satisfies the vanishing moments conditions (7.261) for constant and linear polynomials [333].

**FIGURE 7.29**

Quincunx dual wavelets  $\tilde{\psi}_{j,m}$  and  $\tilde{\psi}_{j+1,m}$  at two consecutive scales  $2^j$  and  $2^{j+1/2}$  (first and second row) with progressively more vanishing moments.

Figure 7.29 shows dual wavelets  $\tilde{\psi}_{j,m}$  and  $\tilde{\psi}_{j+1,m}$  at two consecutive scales  $2^j$  and  $2^{j+1/2}$ , corresponding to the predict and update operators (7.266) and (7.267). They have  $p_2 = 2$  vanishing moments and are nearly isotropic. The analyzing wavelets  $\psi_{j,m}$  also have  $p_1 = 2$  vanishing moments but are more irregular. The irregularity of analyzing wavelets is not a problem since the reconstruction is performed by dual wavelets. Wavelets with more vanishing moments are obtained by replacing the four-neighborhood linear interpolation (7.266) by higher-order polynomial interpolations [333]. Figure 7.29 shows the resulting dual wavelets  $\tilde{\psi}_{j,m}$  for  $p_2 = 4$  and  $p_2 = 6$ , which correspond to wavelets  $\psi_{j,m}$  with as much vanishing moments  $p_1 = p_2$ , but which are more irregular. Increasing the number of vanishing moments improves the regularity of dual wavelets, but it reduces the stability of these biorthogonal bases, which limits their application.

Figure 7.30 shows an example of quincunx wavelet image transform with  $p_1 = 2$  vanishing moments. It is computed with the fast lifting algorithm from Section 7.8.2 with the predict and update operators (7.266) and (7.267). The quasi-isotropic quincunx wavelet detects sharp transitions in all directions.

#### 7.8.4 Wavelets on Bounded Domains and Surfaces

Processing three-dimensional surfaces and signals defined on surfaces is important in multimedia and computer graphics applications. Biorthogonal wavelets on triangulated meshes were introduced by Lounsbery et al. [354], and Schröder and Sweldens [427] improved these techniques with lifting schemes. Lifted wavelets are used to compress functions defined on a surface  $\Omega \subset \mathbb{R}^3$ , and in particular on a sphere to process geographical data on Earth. The sphere is represented with a recursively subdivided three-dimensional mesh, and the signal is processed using lifted wavelets on this embedded mesh.

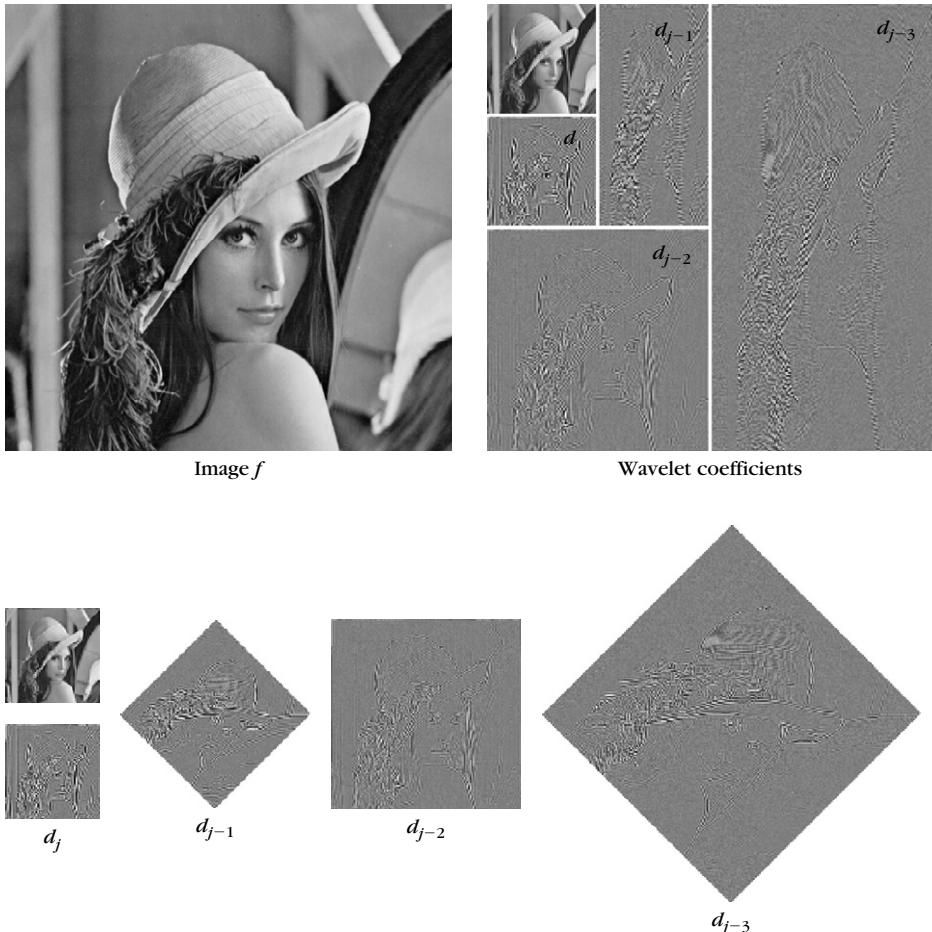
**FIGURE 7.30**

Image decomposition in a biorthogonal quincunx wavelet basis with  $p_1 = 2$  vanishing moments. The top right image shows wavelet coefficients packed over the image-sampling array. These coefficients are displayed as square quincunx grids (*below*) with a rotation for odd scales.

Lifted wavelets are also defined on a two-dimensional parametric domain  $\Omega$  with an arbitrary topology, to compress a three-dimensional surface  $S \subset \mathbb{R}^3$ , viewed as a mapping from  $\Omega$  to  $\mathbb{R}^3$ . The surface is represented as a three-dimensional mesh, and the lifted wavelet transform computes three coefficients for each vertex of the mesh—one per coordinate. Denoising and compression of surfaces are then implemented by thresholding and quantizing these wavelet coefficients. Such multiresolution processes have applications in video games, where a large amount of three-dimensional surface data must be displayed in real time. Lifting wavelets also

finds applications in computer-aided design, where surfaces are densely sampled to represent manufactured objects, and should be compressed to reduce the storage requirements.

### Semiregular Triangulation

A triangulation is a data structure frequently used to index points  $x_n \in \Omega$  on a surface. Embedded indexes  $n \in \mathcal{G}_j$  with a triangulation topology are defined by recursively subdividing a coarse triangulated mesh.

For each scale  $j$ , a triangulation  $(E_j, T_j)$  is composed of edges  $E_j \subset \mathcal{G}_j \times \mathcal{G}_j$  that link pairs of points on the grid, and triangles  $T_j \subset \mathcal{G}_j \times \mathcal{G}_j \times \mathcal{G}_j$ . Each triangle of  $T_j$  is composed of three edges in  $E_j$ . These triangulations are supposed to be embedded using the 1:4 subdivision rule of each triangle, illustrated in Figure 7.31, as follows.

- For each edge  $e \in E_j$ , a midpoint  $\gamma(e) \in \mathcal{G}_{j-1}$  is added to the vertices

$$\mathcal{G}_{j-1} = \mathcal{G}_j \cup \{\gamma(e) : e \in E_j\}.$$

- Each edge is subdivided into two finer edges

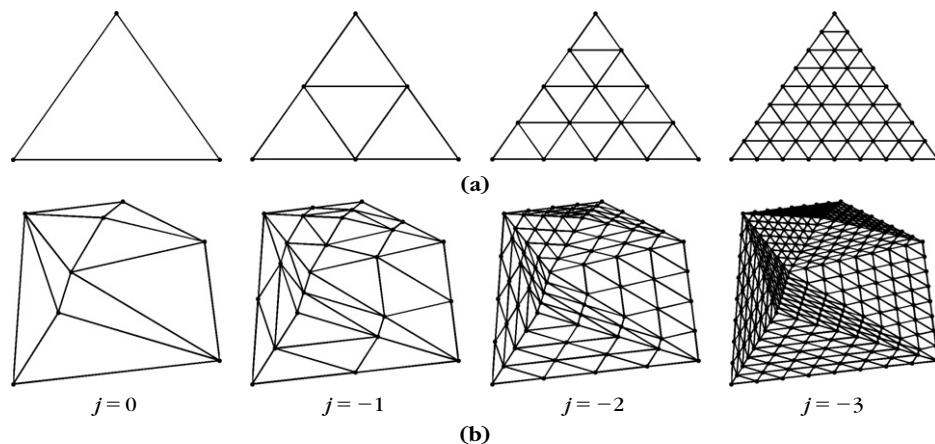
$$\forall e = (n_0, n_1) \in E_j, \quad \sigma_1(e) = (n_0, \gamma(e)) \quad \text{and} \quad \sigma_2(e) = (\gamma(e), n_1).$$

The subdivided set of edges is then

$$E_{j-1} = \{\sigma_i(e) : i = 1, 2 \text{ and } e \in E_j\}.$$

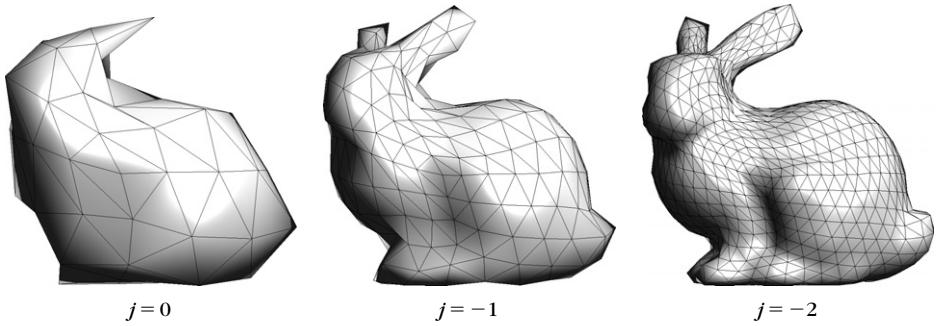
- Each triangle face  $f = (n_0, n_1, n_2) \in F_j$  is subdivided into four faces

$$\begin{aligned} \mu_1(f) &= (n_0, \gamma(n_0, n_1), \gamma(n_0, n_2)), & \mu_2(f) &= (n_1, \gamma(n_1, n_0), \gamma(n_1, n_2)), \\ \mu_3(f) &= (n_2, \gamma(n_2, n_0), \gamma(n_2, n_1)), & \mu_4(f) &= (\gamma(n_0, n_1), \gamma(n_1, n_2), \gamma(n_2, n_0)). \end{aligned}$$



**FIGURE 7.31**

(a) Iterated regular subdivision 1:4 of one triangle in four equal subtriangles. (b) Planar triangulation  $(\mathcal{G}_0, E_0, T_0)$  of a domain  $\Omega$  in  $\mathbb{R}^2$ , successively refined with a 1:4 subdivision.

**FIGURE 7.32**

Examples of semiregular mesh  $\{G_j, E_j, T_j\}_j$  of a domain  $\Omega$  that is a surface  $S$  in  $\mathbb{R}^3$ .

The subdivided set of faces is then

$$F_{j-1} = \{\mu_i(f) : i = 1, 2, 3, 4 \text{ and } f \in F_j\}.$$

Figure 7.31 shows an example of coarse planar triangulation with points  $x_n$  in a domain  $\Omega \in \mathbb{R}^2$ .

The semiregular mesh  $\{E_j, T_j\}_j$  and the corresponding sample locations  $x_n$  are usually computed from an input surface  $S$ , represented either with a parametric continuous function or with an unstructured set of polygons. This requires using a hierarchical remeshing process to compute the embedded triangulation [354]. Figure 7.32 shows an example of semiregular triangulation of a three-dimensional surface  $S$ , and in this case the points  $x_n$  belong to a domain  $\Omega$  that is the surface  $S$  in  $\mathbb{R}^3$ .

### Wavelets to Process Functions on Surfaces

Let us consider a domain  $\Omega \in \mathbb{R}^3$  that is a three-dimensional surface, and each  $n \in G_j$  indexes a point  $x_n \in \Omega$  of the surface. Figure 7.33 shows the local labeling convention for the neighboring vertices in  $G_j$  of a given index  $m \in C_j$  in a butterfly neighborhood.

A predictor  $P_j$  is computed in this local neighborhood

$$\forall m \in C_j, (P_j a)[m] = \sum_{i=1,2} \lambda_i a[v_m^i] + \sum_{i=1,2} \mu_i a[w_m^i] + \sum_{i,j=1}^2 \nu_{i,j} a[z_m^{i,j}], \quad (7.268)$$

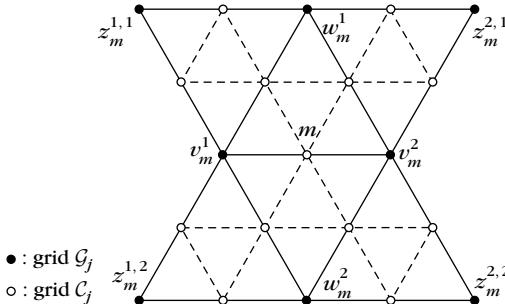
where the parameters  $\lambda_i, \mu_i, \nu_{i,j}$  are calculated by solving (7.258) to obtain vanishing moments for a collection of low-degree polynomials  $\{q^{(k)}\}_{0 \leq k < d_1}$  [427].

The update operator ensures that the dual wavelets have one vanishing moment. It is calculated on the direct neighbors in  $C_j$  of each point  $n \in G_j$ :

$$\forall n \in G_j, \mathcal{V}_n = \{m = \gamma(n, n') \in C_j : (n, n') \in E_j\}.$$

The update operator is parameterized as follows:

$$\forall b \in \mathbb{C}^{|C_j|}, \forall n \in G_j, (U_j b)[n] = \lambda_n \sum_{m \in \mathcal{V}_n} b[m], \quad (7.269)$$

**FIGURE 7.33**

Labeling of points in the butterfly neighborhood of a vertex  $m \in \mathcal{C}_j$ .

where each  $\lambda_n$  is computed to solve the system (7.261) to obtain vanishing moments. This requires the iterative computation of the moments  $I_j^k(n)$ , as explained in (7.262), with an initialization  $I_L^k(n) = 1$  at the finest scale  $L$  of the mesh. In a perfectly regular triangulation, where all the points have six neighbors, a constant weight  $\lambda_n = \lambda$  can be used, and one can check that  $\lambda_n = 1/24$  guarantees one vanishing moment.

Processing signals on a sphere is an important application of these wavelets on surfaces [427]. The triangulated mesh is obtained by a 1:4 subdivision of a regular polyhedron, for instance a tetrahedron. The positions of the points  $x_n$  for  $n \in \mathcal{G}_j$  are defined recursively by projecting midpoints of the edges on the sphere:

$$\forall (n_0, n_1) \in E_j, \quad x_{\gamma(n_0, n_1)} = \frac{x_{n_0} + x_{n_1}}{\|x_{n_0} + x_{n_1}\|}.$$

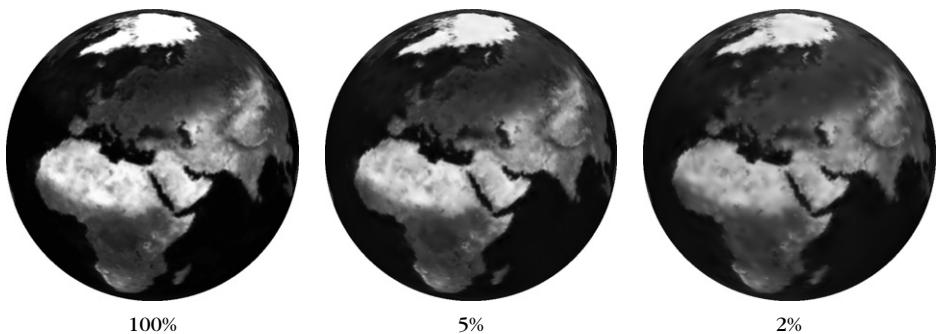
The signal  $f \in \mathbb{C}^{|\mathcal{G}_L|}$  is defined on the finest mesh  $\mathcal{G}_L$ .

A nonlinear approximation is obtained by setting to zero wavelet coefficients that satisfy  $|\langle f, \psi_{j,m} \rangle| \leq T \|\psi_{j,m}\|$  where  $T$  is a given threshold. The value of  $\|\psi_{j,m}\|$  can be approximated from the size of its support  $\|\psi_{j,m}\| \sim \sqrt{\text{supp}(\psi_{j,n})}$  [427]. Figure 7.34 shows an example of such a nonlinear approximation with an image of Earth defined as a function on the sphere.

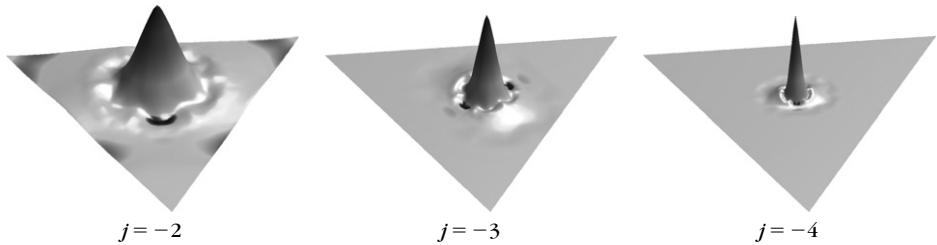
### Wavelets to Process Surfaces

A three-dimensional surface  $\mathcal{S} \subset \mathbb{R}^3$  is represented as a mapping from a two-dimensional parameter domain  $\Omega$  to  $\mathbb{R}^3$ . This surface is discretized with a semiregular mesh  $\{\mathcal{G}_j, E_j, T_j\}_{L \leq j \leq 0}$ , and thus  $\Omega$  can be chosen as the finest grid  $\mathcal{G}_L$  viewed as an abstract domain. The surface is a discrete mapping from  $\Omega = \mathcal{G}_L$  to  $\mathbb{R}^3$  that assigns to each  $n \in \Omega$  three values  $(f_1[n], f_2[n], f_3[n]) \in \mathcal{S}$ , which is a position in the three-dimensional space.

Processing the discrete surface is equivalent to processing the three signals  $(f_1, f_2, f_3)$  where each  $f_i \in \mathbb{C}^{|\mathcal{G}_L|}$  is defined on the finest grid  $\mathcal{G}_L$ . Since points in

**FIGURE 7.34**

Nonlinear approximation of a function defined on a sphere using a decreasing number of large wavelet coefficients.

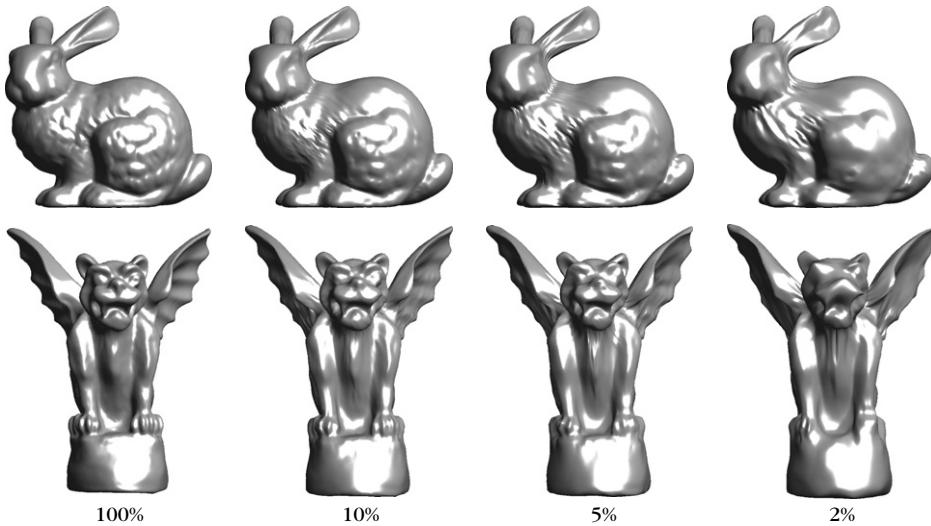
**FIGURE 7.35**

Example of dual wavelets  $\psi_{j,k}(x)$  for  $x$  on a subdivided equilateral triangle. The height over the triangle indicates the value of the wavelet.

the parametric domain  $\Omega$  do not have positions in Euclidean space, the notion of vanishing moments is not well defined, and the predict operator is computed using weights that are calculated as if the faces of the mesh were equilateral triangles. One can verify that the resulting parameters of the predict operator (7.268) are  $\lambda_i = 1/2$ ,  $\mu_i = 1/4$ , and  $\nu_{i,j} = -1/16$  [427]. Figure 7.35 shows the corresponding wavelets on a subdivided equilateral triangle.

These wavelets are used to compress each of the three signals  $f_i \in \mathbb{C}^{|\mathcal{G}_L|}$  by uniformly quantizing the normalized coefficients  $\langle f_i, \psi_{j,m} \rangle / \|\psi_{j,m}\|$ . The resulting set of quantized coefficients are within strings with an entropy coder algorithm, described in Section 10.2.1. The quantization and coding of sparse signal representation is described in Section 10.4.

Figure 7.36 shows an example of three-dimensional surface approximation using biorthogonal wavelets on triangulated meshes. Wavelet coders based on lifting offer state-of-the-art results in surface compression [327]. There is no control on the Riesz bounds of the lifted wavelet basis, because the lifted basis depends on the surface properties, but good approximation results are obtained.

**FIGURE 7.36**

Nonlinear surface approximation using a decreasing proportion of large wavelet coefficients.

### 7.8.5 Faster Wavelet Transform with Lifting

Lifting is used to reduce the number of operations of a one-dimensional fast wavelet transform by factorizing the filter bank convolutions. It also reduces the memory requirements by implementing all operations, in place, within the input signal array.

Before the introduction of liftings for wavelet design, the factorization of filter bank convolutions was studied as paraunitary filter banks by Vaidyanathan [68] and other signal-processing researchers. Instead of implementing a filter bank with convolutions and subsamplings, specific filters are designed for even and odd signal coefficients. These filters are shown to be factorized as a succession of elementary operators that correspond to the predict and update operators of a lifting. This filtering architecture reduces by up to 2 the number of additions and multiplications, and simplifies folding border treatments for symmetric wavelets.

The fast wavelet transform algorithm in Sections 7.3.1 and 7.3.2 decomposes iteratively the scaling coefficients  $a_j[n]$  at a scale  $2^j$  into larger-scale coefficients  $a_{j+1}[n]$  and detail wavelet coefficients  $d_{j+1}[n]$ , with convolutions and subsampling using two filters  $h$  and  $g$ . According to (7.157),

$$a_{j+1}[n] = a_j \star \bar{h}[2n], \quad d_{j+1}[n] = a_j \star \bar{g}[2n], \quad (7.270)$$

with  $\bar{h}[n] = h[-n]$  and  $\bar{g}[n] = g[-n]$ . The reconstruction is performed with the dual filters  $\tilde{h}$  and  $\tilde{g}$  according to (7.158),

$$a_j[n] = \check{a}_{j+1} \star \tilde{h}[n] + \check{d}_{j+1} \star \tilde{g}[n]. \quad (7.271)$$

Sweldens and Daubechies [199] proved that the convolutions and subsamplings (7.270) can be implemented with a succession of lifting steps and the reconstruction (7.271) by inverting these liftings.

Each uniform one-dimensional signal sampling grid  $\mathcal{G}_j$  of  $a_j[n]$  is divided into even samples in  $\mathcal{G}_{j+1}$  and odd samples in  $\mathcal{C}_j$  where  $a_{j+1}$  and  $d_{j+1}$  are, respectively, defined. Starting from the lazy wavelet transform that splits even and odd samples of  $a_j[n]$ ,  $K$  couples of predict and update convolution operators  $\{P^{(k)}, U^{(k)}\}_{1 \leq k \leq K}$  are sequentially applied. Each predict operator  $P^{(k)}$  corresponds to a filter  $p^{(k)}[n]$  and each update operator to a filter  $u^{(k)}[m]$ . The filters have a small support of typically two coefficients. They are computed from the biorthogonal filters  $(h, g, \tilde{h}, \tilde{g})$  with a Euclidean division algorithm [199]. A final scaling step renormalizes the filter coefficients.

Let  $a_L[n]$  be the input finest-scale signal for  $n \in \mathcal{G}_L$ . The lifting implementation of a fast wavelet transform proceeds as follow. For  $j = L+1, \dots, J$ :

1. *Even-odd split:*  $\forall m \in \mathcal{C}_j, d_j^{(0)}[m] = a_{j-1}[m]$ , and  $\forall n \in \mathcal{G}_j, a_j^{(0)}[n] = a_{j-1}[n]$ . The following steps 2 and 3 are performed for  $k = 1, \dots, K$ .
2. *Forward predict:*  $\forall m \in \mathcal{C}_j, d_j^{(k)}[m] = d_j^{(k-1)}[m] - \sum_{n \in \mathcal{G}_j} p^{(k)}[m-n] a_j^{(k-1)}[n]$ .
3. *Forward update:*  $\forall n \in \mathcal{G}_j, a_j^{(k)}[n] = a_j^{(k-1)}[n] + \sum_{m \in \mathcal{C}_j} u^{(k)}[n-m] d_j^{(k)}[m]$ .
4. *Scaling:* The output coefficients are  $d_j = d_j^{(K)} / \zeta$  and  $a_j = \zeta a_j^{(K)}$ .

The fast inverse wavelet transform recovers  $a_L$  from the set of wavelet coefficients  $\{d_j\}_{0 \leq j < L}$  and  $a_J$  by inverting these predict and update operators. For  $j = J, \dots, L+1$ :

1. *Inverse scaling:* Initialize  $d_j^{(K)} = \zeta d_j$  and  $a_j^{(K)} = a_j / \zeta$ . The following steps, 2 and 3, are performed for  $k = K, \dots, 1$ .
2. *Backward update:*  $\forall n \in \mathcal{G}_j, a_j^{(k-1)}[n] = a_j^{(k)}[n] - \sum_{m \in \mathcal{C}_j} u^{(k)}[n-m] d_j^{(k)}[m]$ .
3. *Backward predict:*  $\forall m \in \mathcal{C}_j, d_j^{(k-1)}[m] = d_j^{(k)}[m] + \sum_{n \in \mathcal{G}_j} p^{(k)}[m-n] a_j^{(k-1)}[n]$ .
4. *Merge even-odd samples:*  $\forall m \in \mathcal{C}_j, a_{j-1}[m] = d_j^{(0)}[m]$  and  $\forall n \in \mathcal{G}_j, a_{j-1}[n] = a_j^{(0)}[n]$ .

One can verify (Exercise 7.21) that this implementation divides the number of operations by up to a factor of 2, compared to direct convolutions and subsamplings calculated in (7.270) and (7.271). Moreover, this algorithm proceeds “in place,” which means that it only uses the memory of the original array  $\mathcal{G}_L$  with

coefficients that are progressively modified by the lifting operations; it then stores the resulting wavelet coefficients. Similarly, the reconstruction operates in place by reconstructing progressively the coefficients in the array  $\mathcal{G}_L$ .

### **Symmetric Lifting on the Interval**

The lifting implementation is described in more detail for symmetric wavelets on an interval, corresponding to symmetric biorthogonal filters  $(h, g, \tilde{h}, \tilde{g})$ . Predict and update convolutions are modified at the boundaries to implement folding boundary conditions.

The input sampling grid  $\mathcal{G}_L$  has  $N = 2^{-L}$  samples at integer positions  $0 \leq n < N$ . The embedded subgrids at scales  $0 \geq 2^j > 2^L$  are

$$\mathcal{G}_j = \{kN2^j : 0 \leq k < 2^{-j}\} \quad \text{and} \quad \mathcal{C}_j = \{kN2^j + N2^{j-1} : 0 \leq k < 2^{-j}\}. \quad (7.272)$$

The resulting lifting operators  $\{P^{(k)}, U^{(k)}\}_{1 \leq k \leq K}$  are two-tap symmetric convolution operators. A prediction of parameter  $\lambda$  is defined by

$$\forall m \in \mathcal{C}_j, \quad m < N - N2^{j-1}, \quad (P_\lambda a)[m] = \lambda (a[m + N2^{j-1}] + a[m - N2^{j-1}]). \quad (7.273)$$

An update of parameter  $\mu$  is defined as

$$\forall n \in \mathcal{G}_j, \quad n > 0, \quad (U_\mu b)[n] = \mu (b[n + N2^{j-1}] + b[n - N2^{j-1}]). \quad (7.274)$$

At the interval boundaries, the predict and update operators must be modified for  $m = N - N2^{j-1}$  in (7.273) and  $n = 0$  in (7.274), because one of the two neighbors falls outside the grids  $\mathcal{G}_j$  and  $\mathcal{C}_j$ .

Symmetric boundary conditions, described in Section 7.5.2, are implemented with a folding, which leads to the following definition of boundary predict and update operators:

$$(P_\lambda a)[N - N2^{j-1}] = 2\lambda a[N - N2^j] \quad \text{and} \quad (U_\mu b)[0] = 2\mu b[N2^{j-1}].$$

This ensures that the lifting operators have one vanishing moment, and that the resulting boundary wavelets also have one vanishing moment.

### **Factorization of 5/3 and 9/7 Biorthogonal Wavelets**

The 9/7 biorthogonal and 5/3 wavelets in Figure 7.15 are recommended for image compression in JPEG-2000 and are often used in wavelet image-processing applications. The 5/3 biorthogonal wavelet has  $p_1 = p_2 = 2$  vanishing moments, while the 9/7 wavelet has  $p_1 = p_2 = 4$  vanishing moments.

The 5/3 wavelet transform is implemented with  $K = 1$  pair of predict and update steps, presented in Section 7.8.2. They correspond to

$$P^{(1)} = P_{1/2}, \quad U^{(1)} = U_{1/4}, \quad \text{and} \quad \zeta = \sqrt{2}.$$

The 9/7 wavelet transform is implemented with  $K = 2$  pairs of predict and update steps defined as

$$\begin{cases} P^{(1)} = P_\alpha, & U^{(1)} = U_\beta, \\ P^{(2)} = P_\gamma, & U^{(2)} = U_\delta, \end{cases} \quad \text{where} \quad \begin{cases} \alpha = 1.58613434342059, & \beta = -0.0529801185729, \\ \gamma = -0.8829110755309, & \delta = 0.4435068520439, \end{cases}$$

and the scaling constant is  $\zeta = 1.1496043988602$ .

## 7.9 EXERCISES

- 7.1** <sup>2</sup> Let  $h$  be a conjugate mirror filter associated to a scaling function  $\phi$ .
- Prove that if  $\hat{h}(\omega)$  is  $C^p$  and has a zero of order  $p$  at  $\pi$ , then the  $l$ th-order derivative  $\hat{\phi}^{(l)}(2k\pi) = 0$  for any  $k \in \mathbb{Z} - \{0\}$  and  $l < p$ .
  - Derive that if  $q < p$ , then  $\sum_{n=-\infty}^{+\infty} n^q \phi(n) = \int_{-\infty}^{+\infty} t^q \phi(t) dt$ .
- 7.2** <sup>2</sup> Prove that  $\sum_{n=-\infty}^{+\infty} \phi(t-n) = 1$  if  $\phi$  is an orthogonal scaling function.
- 7.3** <sup>2</sup> Let  $\phi_m$  be the Battle-Lemarié scaling function of degree  $m$  defined in (7.18). Let  $\phi$  be the Shannon scaling function defined by  $\hat{\phi} = \mathbf{1}_{[-\pi, \pi]}$ . Prove that  $\lim_{m \rightarrow +\infty} \|\phi_m - \phi\| = 0$ .
- 7.4** <sup>3</sup> Suppose that  $h[n]$  is nonzero only for  $0 \leq n < K$ . We denote  $m[n] = \sqrt{2} h[n]$ . The scaling equation is  $\phi(t) = \sum_{n=0}^{K-1} m[n] \phi(2t-n)$ .
- Suppose that  $K = 2$ . Prove that if  $t$  is a dyadic number that can be written in binary form with  $i$  digits,  $t = 0.\varepsilon_1\varepsilon_2 \cdots \varepsilon_i$  with  $\varepsilon_k \in \{0, 1\}$ , then  $\phi(t)$  is the product

$$\phi(t) = m[\varepsilon_0] \times m[\varepsilon_1] \times \cdots \times m[\varepsilon_i] \times \phi(0).$$

- For  $K = 2$ , show that if  $m[0] = 4/3$  and  $m[1] = 2/3$ , then  $\phi(t)$  is singular at all dyadic points. Verify numerically with WAVELET that the resulting scaling equation does not define a finite-energy function  $\phi$ .
- Show that one can find two matrices  $M[0]$  and  $M[1]$  such that the  $K$ -dimensional vector  $\Phi(t) = [\phi(t), \phi(t+1), \dots, \phi(t+K-1)]^T$  satisfies

$$\Phi(t) = M[0] \Phi(2t) + M[1] \Phi(2t-1).$$

- Show that one can compute  $\Phi(t)$  at any dyadic number  $t = 0.\varepsilon_1\varepsilon_2 \cdots \varepsilon_i$  with a product of matrices:

$$\Phi(t) = M[\varepsilon_0] \times M[\varepsilon_1] \times \cdots \times M[\varepsilon_i] \times \Phi(0).$$

- 7.5** <sup>2</sup> Let us define

$$\phi_{k+1}(t) = \sqrt{2} \sum_{n=-\infty}^{+\infty} h[n] \phi_k(2t-n), \quad (7.275)$$

with  $\phi_0 = \mathbf{1}_{[0, 1]}$ , and  $a_k[n] = \langle \phi_k(t), \phi_k(t-n) \rangle$ .

(a) Let

$$P\hat{f}(\omega) = \frac{1}{2} \left( |\hat{h}\left(\frac{\omega}{2}\right)|^2 \hat{f}\left(\frac{\omega}{2}\right) + |\hat{h}\left(\frac{\omega}{2} + \pi\right)|^2 \hat{f}\left(\frac{\omega}{2} + \pi\right) \right).$$

Prove that  $\hat{a}_{k+1}(\omega) = P\hat{a}_k(\omega)$ .

- (b) Prove that if there exists  $\phi$  such that  $\lim_{k \rightarrow +\infty} \|\phi_k - \phi\| = 0$ , then 1 is an eigenvalue of  $P$  and  $\hat{\phi}(\omega) = \prod_{p=1}^{+\infty} 2^{-1/2} \hat{h}(2^{-p}\omega)$ . What is the degree of freedom on  $\phi_0$  in order to still converge to the same limit  $\phi$ ?
- (c) Implement numerically the computations of  $\phi_k(t)$  for the Daubechies conjugate mirror filter with  $p=6$  zeros at  $\pi$ . How many iterations are needed to obtain  $\|\phi_k - \phi\| < 10^{-4}$ ? Try to improve the rate of convergence by modifying  $\phi_0$ .

**7.6** <sup>3</sup> Let  $b[n] = f(N^{-1}n)$  with  $2^L = N^{-1}$  and  $f \in \mathbf{V}_L$ . We want to recover  $a_L[n] = \langle f, \phi_{L,n} \rangle$  from  $b[n]$  to compute the wavelet coefficients of  $f$  with Theorem 7.10.

- (a) Let  $\phi_L[n] = 2^{-L/2} \phi(2^{-L}n)$ . Prove that  $b[n] = a_L \star \phi_L[n]$ .
- (b) Prove that if there exists  $C > 0$  such that for all  $\omega \in [-\pi, \pi]$ ,

$$\hat{\phi}_d(\omega) = \sum_{k=-\infty}^{+\infty} \hat{\phi}(\omega + 2k\pi) \geq C,$$

then  $a_L$  can be calculated from  $b$  with a stable filter  $\phi_L^{-1}[n]$ .

- (c) If  $\phi$  is a cubic spline-scaling function, compute numerically  $\phi_L^{-1}[n]$ . For a given numerical precision, compare the number of operations needed to compute  $a_L$  from  $b$  with the number of operations needed to compute the fast wavelet transform of  $a_L$ .
- (d) Show that calculating  $a_L$  from  $b$  is equivalent to performing a change of basis in  $\mathbf{V}_L$  from a Riesz interpolation basis to an orthonormal basis.

**7.7** <sup>2</sup> *Quadrature mirror filters.* We define a multirate filter bank with four filters  $h, g, \tilde{h}$ , and  $\tilde{g}$ , which decomposes a signal  $a_0[n]$

$$a_1[n] = a_0 \star h[2n], \quad d_1[n] = a_0 \star g[2n].$$

Using the notation (7.101), we reconstruct

$$\tilde{a}_0[n] = \check{a}_1 \star \tilde{h}[n] + \check{d}_1 \star \tilde{g}[n].$$

- (a) Prove that  $\tilde{a}_0[n] = a_0[n-l]$  if

$$\hat{g}(\omega) = \hat{h}(\omega + \pi), \quad \widehat{\tilde{h}}(\omega) = \hat{h}(\omega), \quad \widehat{\tilde{g}}(\omega) = -\widehat{\tilde{h}}(\omega + \pi),$$

and if  $h$  satisfies the quadrature mirror condition

$$\hat{h}^2(\omega) - \hat{h}^2(\omega + \pi) = 2 e^{-il\omega}.$$

- (b) Show that  $\ell$  is necessarily odd.  
 (c) Verify that the Haar filter (7.46) is a quadrature mirror filter (it is the only finite impulse response solution).

- 7.8** <sup>1</sup> Let  $f$  be a function of support  $[0, 1]$  that is equal to different polynomials of degree  $q$  on the intervals  $\{[\tau_k, \tau_{k+1}]\}_{0 \leq k < K}$  with  $\tau_0 = 0$  and  $\tau_K = 1$ . Let  $\psi$  be a Daubechies wavelet with  $p$  vanishing moments. If  $q < p$ , compute the number of nonzero wavelet coefficients  $\langle f, \psi_{j,n} \rangle$  at a fixed scale  $2^j$ . How should we choose  $p$  to minimize this number? If  $q > p$ , what is the maximum number of nonzero wavelet coefficients  $\langle f, \psi_{j,n} \rangle$  at a fixed scale  $2^j$ ?  
**7.9** <sup>2</sup> Let  $\theta$  be a box spline of degree  $m$  obtained by  $m + 1$  convolutions of  $\mathbf{1}_{[0,1]}$  with itself.

- (a) Prove that

$$\theta(t) = \frac{1}{m!} \sum_{k=0}^{m+1} (-1)^k \binom{m+1}{k} ([t-k]_+)^m,$$

where  $[x]_+ = \max(x, 0)$ . Hint: Write  $\mathbf{1}_{[0,1]} = \mathbf{1}_{[0,+\infty)} - \mathbf{1}_{(1,+\infty)}$ .

- (b) Let  $A_m$  and  $B_m$  be the Riesz bounds of  $\{\theta(t-n)\}_{n \in \mathbb{Z}}$ . With (7.9) prove that  $\lim_{m \rightarrow +\infty} B_m = +\infty$ . Compute numerically  $A_m$  and  $B_m$  for  $m \in \{0, \dots, 5\}$  with MATLAB.

- 7.10** <sup>1</sup> Prove that if  $\{\psi_{j,n}\}_{(j,n) \in \mathbb{Z}^2}$  is an orthonormal basis of  $\mathbf{L}^2(\mathbb{R})$ , then for all  $\omega \in \mathbb{R} - \{0\}$ ,  $\sum_{j=-\infty}^{+\infty} |\hat{\psi}(2^j \omega)|^2 = 1$ . Find an example showing that the converse is not true.

- 7.11** <sup>3</sup> Let us define

$$\hat{\psi}(\omega) = \begin{cases} 1 & \text{if } 4\pi/7 \leq |\omega| \leq \pi \text{ or } 4\pi \leq |\omega| \leq 4\pi + 4\pi/7 \\ 0 & \text{otherwise.} \end{cases}$$

Prove that  $\{\psi_{j,n}\}_{(j,n) \in \mathbb{Z}^2}$  is an orthonormal basis of  $\mathbf{L}^2(\mathbb{R})$ . Prove that  $\psi$  is not associated to a scaling function  $\phi$  that generates a multiresolution approximation.

- 7.12** <sup>2</sup> Express the Coiflet property (7.99) as an equivalent condition on the conjugate mirror filter  $\hat{h}(e^{i\omega})$ .

- 7.13** <sup>1</sup> Prove that  $\psi(t)$  has  $p$  vanishing moments if and only if, for all  $j \geq 0$ , the discrete wavelets  $\psi_j[n]$  defined in (7.140) have  $p$  discrete vanishing moments

$$\sum_{n=-\infty}^{+\infty} n^k \psi_j[n] = 0 \quad \text{for } 0 \leq k < p.$$

- 7.14** <sup>2</sup> Let  $\psi(t)$  be a compactly supported wavelet calculated with Daubechies conjugate mirror filters  $(h, g)$ . Let  $\psi'_{j,n}(t) = 2^{-j/2} \psi'(2^{-j}t - n)$  be the derivative wavelets.

- (a) Verify that  $h_1$  and  $g_1$  defined by

$$\hat{h}_1(\omega) = 2 \hat{h}(\omega) (e^{i\omega} - 1)^{-1}, \quad \hat{g}_1(\omega) = 2 (e^{i\omega} - 1) \hat{g}(\omega)$$

are finite impulse response filters.

- (b) Prove that the Fourier transform of  $\psi'(t)$  can be written as

$$\widehat{\psi'}(\omega) = \frac{\widehat{g}_1(2^{-1}\omega)}{\sqrt{2}} \prod_{p=2}^{+\infty} \frac{\widehat{h}_1(2^{-p}\omega)}{\sqrt{2}}.$$

- (c) Describe a fast filter bank algorithm to compute the derivative wavelet coefficients  $\langle f, \psi'_{j,n} \rangle$  [108].

- 7.15** <sup>3</sup> Let  $\psi(t)$  be a compactly supported wavelet calculated with Daubechies conjugate mirror filters  $(h, g)$ . Let  $\hat{h}^a(\omega) = |\hat{h}(\omega)|^2$ . We verify that  $\hat{\psi}^a(\omega) = \hat{\psi}(\omega) \hat{h}^a(\omega/4 - \pi/2)$  is an almost analytic wavelet.

- (a) Prove that  $\psi^a$  is a complex wavelet such that  $\operatorname{Re}[\psi^a] = \psi$ .  
(b) Compute numerically  $\psi^a(\omega)$  for a Daubechies wavelet with four vanishing moments. Explain why  $\psi^a(\omega) \approx 0$  for  $\omega < 0$ .  
(c) Let  $\psi_{j,n}^a(t) = 2^{-j/2} \psi^a(2^{-j}t - n)$ . Using the fact that

$$\hat{\psi}^a(\omega) = \frac{\hat{g}(2^{-1}\omega)}{\sqrt{2}} \frac{\hat{h}(2^{-2}\omega)}{\sqrt{2}} \frac{|\hat{h}(2^{-2}\omega - 2^{-1}\pi)|^2}{\sqrt{2}} \prod_{k=3}^{+\infty} \frac{\hat{h}(2^{-k}\omega)}{\sqrt{2}},$$

show that we can modify the fast wavelet transform algorithm to compute the “analytic” wavelet coefficients  $\langle f, \psi_{j,n}^a \rangle$  by inserting a new filter.

- (d) Let  $\phi$  be the scaling function associated to  $\psi$ . We define separable two-dimensional “analytic” wavelets by:

$$\begin{aligned} \psi^1(x) &= \psi^a(x_1) \phi(x_2), & \psi^2(x) &= \phi(x_1) \psi^a(x_2), \\ \psi^3(x) &= \psi^a(x_1) \psi^a(x_2), & \psi^4(x) &= \psi^a(x_1) \psi^a(-x_2). \end{aligned}$$

Let  $\psi_{j,n}^k(x) = 2^{-j} \psi^k(2^{-j}x - n)$  for  $n \in \mathbb{Z}^2$ . Modify the separable wavelet filter bank algorithm from Section 7.7.3 to compute the “analytic” wavelet coefficients  $\langle f, \psi_{j,n}^k \rangle$ .

- (e) Prove that  $\{\psi_{j,n}^k\}_{1 \leq k \leq 4, j \in \mathbb{Z}, n \in \mathbb{Z}^2}$  is a frame of the space of *real* functions  $f \in \mathbf{L}^2(\mathbb{R}^2)$  [108].

- 7.16** <sup>2</sup> *Mutiwavelets.* We define the following two scaling functions:

$$\phi_1(t) = \phi_1(2t) + \phi_1(2t - 1)$$

$$\phi_2(t) = \frac{1}{2} (\phi_2(2t) + \phi_2(2t - 1) - \phi_1(2t) + \phi_1(2t - 1)).$$

- (a) Compute the functions  $\phi_1$  and  $\phi_2$ . Prove that  $\{\phi_1(t - n), \phi_2(t - n)\}_{n \in \mathbb{Z}}$  is an orthonormal basis of a space  $\mathbf{V}_0$  that will be specified.

- (b) Find  $\psi_1$  and  $\psi_2$  with a support on  $[0, 1]$  that are orthogonal to each other and to  $\phi_1$  and  $\phi_2$ . Plot these wavelets. Verify that they have two vanishing moments and that they generate an orthonormal basis of  $L^2(\mathbb{R})$ .

- 7.17** <sup>3</sup> Let  $f^{\text{repl}}$  be the folded function defined in (7.179).
- (a) Let  $\alpha(t), \beta(t) \in L^2(\mathbb{R})$  be two functions that are either symmetric or anti-symmetric about  $t = 0$ . If  $\langle \alpha(t), \beta(t + 2k) \rangle = 0$  and  $\langle \alpha(t), \beta(2k - t) \rangle = 0$  for all  $k \in \mathbb{Z}$ , then prove that

$$\int_0^1 \alpha^{\text{repl}}(t) \beta^{\text{repl}}(t) dt = 0.$$

- (b) Prove that if  $\psi, \tilde{\psi}, \phi, \tilde{\phi}$  are either symmetric or antisymmetric with respect to  $t = 1/2$  or  $t = 0$ , and generate biorthogonal bases of  $L^2(\mathbb{R})$ , then the folded bases (7.181) and (7.182) are biorthogonal bases of  $L^2[0, 1]$ . Hint: Use the same approach as in Theorem 7.16.

- 7.18** <sup>3</sup> A recursive filter has a Fourier transform that is a ratio of trigonometric polynomials as in (2.31).

- (a) Let  $p[n] = h \star \bar{h}[n]$  with  $\bar{h}[n] = h[-n]$ . Verify that if  $h$  is a recursive conjugate mirror filter, then  $\hat{p}(\omega) + \hat{p}(\omega + \pi) = 2$  and there exists  $\hat{r}(\omega) = \sum_{k=0}^{K-1} r[k] e^{-ik\omega}$  such that

$$\hat{p}(\omega) = \frac{2|\hat{r}(\omega)|^2}{|\hat{r}(\omega)|^2 + |\hat{r}(\omega + \pi)|^2}. \quad (7.276)$$

- (b) Suppose that  $K$  is even and that  $r[K/2 - 1 - k] = r[K/2 + k]$ . Verify that

$$\hat{p}(\omega) = \frac{|\hat{r}(\omega)|^2}{2|\hat{r}(\omega) + \hat{r}(\omega + \pi)|^2}. \quad (7.277)$$

- (c) If  $\hat{r}(\omega) = (1 + e^{-i\omega})^{K-1}$  with  $K = 6$ , compute  $\hat{h}(\omega)$  with the factorization (7.277), and verify that it is a stable filter (Exercise 3.18). Compute numerically this filter and plot the graph of the corresponding wavelet  $\psi(t)$ .

- 7.19** <sup>2</sup> *Balancing.* Suppose that  $h, \tilde{h}$  define a pair of perfect reconstruction filters satisfying (7.124).

- (a) Prove that

$$h_{\text{new}}[n] = \frac{1}{2}(h[n] + h[n-1]), \quad \tilde{h}_{\text{new}}[n] = \frac{1}{2}(\tilde{h}[n] + \tilde{h}[n-1])$$

defines a new pair of perfect reconstruction filters. Verify that  $\hat{h}_{\text{new}}(\omega)$  and  $\hat{\tilde{h}}_{\text{new}}(\omega)$  have, respectively, one more and one less zero at  $\pi$  than  $\hat{h}(\omega)$  and  $\hat{\tilde{h}}(\omega)$  [63].

- (b) Relate this balancing operation to a lifting.

- (c) The Deslauriers-Dubuc filters are  $\hat{h}(\omega) = 1$  and

$$\widehat{\tilde{h}}(\omega) = \frac{1}{16} (-e^{-3i\omega} + 9e^{-i\omega} + 16 + 9e^{i\omega} - e^{3i\omega}).$$

Compute  $h_{new}$  and  $\tilde{h}_{new}$  as well as the corresponding biorthogonal wavelets  $\psi_{new}$ ,  $\tilde{\psi}_{new}$ , after one balancing and again after a second balancing.

- 7.20** <sup>1</sup> Compute numerically the wavelets and scaling functions associated to the predict and update lifting steps (7.264) and (7.265). Verify that you obtain the 5/3 wavelets displayed in Figure 7.15.
- 7.21** <sup>1</sup> Give the reduction of the number of operations when implementing a fast wavelet transform with 5/3 and 7/9 biorthogonal wavelets with the lifting algorithm described in Section 7.8.5, compared with a direct implementation with (7.270) and (7.271) by using the coefficients in Table 7.4.
- 7.22** <sup>1</sup> For a Deslauriers-Dubuc interpolation wavelet of degree 3, compute the dual wavelet  $\tilde{\psi}$  in (7.212), which is a sum of Diracs. Verify that it has four vanishing moments.
- 7.23** <sup>2</sup> Prove that a Deslauriers-Dubuc interpolation function of degree  $2p - 1$  converges to a sinc function when  $p$  goes to  $+\infty$ .
- 7.24** <sup>3</sup> Let  $\phi$  be an autocorrelation scaling function that reproduces polynomials of degree  $p - 1$  as in (7.198). Prove that if  $f$  is uniformly Lipschitz  $\alpha$ , then under the same hypotheses as in Theorem 7.22, there exists  $K > 0$  such that

$$\|f - P_{V_j} f\|_\infty \leq K 2^{\alpha j}.$$

- 7.25** <sup>2</sup> Let  $\phi(t)$  be an interpolation function that generates an interpolation wavelet basis of  $C_0(\mathbb{R})$ . Construct a separable interpolation wavelet basis of the space  $C_0(\mathbb{R}^p)$  of uniformly continuous  $p$ -dimensional signals  $f(x_1, \dots, x_p)$ . Hint: Construct  $2^p - 1$  interpolation wavelets by appropriately translating  $\phi(x_1) \cdots \phi(x_p)$ .
- 7.26** <sup>3</sup> *Fractional Brownian.* Let  $\psi(t)$  be a compactly supported wavelet with  $p$  vanishing moments that generates an orthonormal basis of  $L^2(\mathbb{R})$ . The covariance of a fractional Brownian motion  $B_H(t)$  is given by (6.86).
- (a) Prove that  $E\{|\langle B_H, \psi_{j,n} \rangle|^2\}$  is proportional to  $2^{j(2H+1)}$ . Hint: Use Exercise 6.15.
  - (b) Prove that the decorrelation between the same scale wavelet coefficients increases when the number  $p$  of vanishing moments of  $\psi$  increases:

$$E\{\langle B_H, \psi_{j,n} \rangle \langle B_H, \psi_{l,m} \rangle\} = O\left(2^{j(2H+1)} |n-m|^{2(H-p)}\right).$$

- (c) In two dimensions, synthesize “approximate” fractional Brownian motion images  $\tilde{B}_H$  with wavelet coefficients  $\langle B_H, \psi_{j,n}^k \rangle$  that are independent Gaussian random variables, with variances proportional to  $2^{j(2H+2)}$ . Adjust  $H$  in order to produce textures that look like clouds in the sky.
- 7.27** <sup>2</sup> *Image mosaic.* Let  $f_0[n_1, n_2]$  and  $f_1[n_1, n_2]$  be two square images of  $N$  pixels. We want to merge the center of  $f_0[n_1, n_2]$  for  $N^{1/2}/4 \leq n_1, n_2 < 3N^{1/2}/4$  in the center of  $f_1$ . Compute numerically the wavelet coefficients of  $f_0$  and  $f_1$ . At each scale  $2^j$  and orientation  $1 \leq k \leq 3$ , replace the  $2^{-2j}/4$  wavelet coefficients corresponding to the center of  $f_1$  by the wavelet coefficients of  $f_0$ . Reconstruct an image from this manipulated wavelet representation. Explain why the image  $f_0$  seems to be merged in  $f_1$ , without the strong boundary effects that are obtained when directly replacing the pixels of  $f_1$  by the pixels of  $f_0$ .
- 7.28** <sup>3</sup> *Foveal vision.* A foveal image has a maximum resolution at the center, with a resolution that decreases linearly as a function of the distance to the center. Show that one can construct an approximate foveal image by keeping a constant number of nonzero wavelet coefficients at each scale  $2^j$ . Implement this algorithm numerically.
- 7.29** <sup>2</sup> *High contrast.* We consider a color image specified by three color channels: red  $r[n]$ , green  $g[n]$ , and blue  $b[n]$ . The intensity image  $(r + g + b)/3$  averages the variations of the three color channels. To create a high-contrast image  $f$ , for each wavelet  $\psi_{j,n}^k$  we set  $\langle f, \psi_{j,n}^k \rangle$  to be the coefficient among  $\langle r, \psi_{j,n}^k \rangle$ ,  $\langle g, \psi_{j,n}^k \rangle$ , and  $\langle b, \psi_{j,n}^k \rangle$ , which has the maximum amplitude. Implement this algorithm numerically and evaluate its performance for different types of multispectral images. How does the choice of  $\psi$  affect the results?

# Wavelet Packet and Local Cosine Bases

# 8

Different types of time-frequency structures are encountered in complex signals such as speech recordings. This motivates the design of bases with time-frequency properties that may be adapted. Wavelet bases are one particular family of bases that represent piecewise smooth signals effectively. Other time-frequency bases are constructed to approximate different types of signals such as audio recordings.

Orthonormal wavelet packet bases are computed with conjugate mirror filters that divide the frequency axis in separate intervals of various sizes. Different conjugate mirror filter banks correspond to different wavelet packet bases. If the signal properties change over time, it is preferable to isolate different time intervals with translated windows. Local cosine bases are constructed by multiplying these windows with cosine functions. Wavelet packets segment the frequency axis and are uniformly translated in time, whereas local cosine bases divide the time axis and are uniformly translated in frequency. Both types of bases are extended in two dimensions for image processing.

## 8.1 WAVELET PACKETS

### 8.1.1 Wavelet Packet Tree

Wavelet packets were introduced by Coifman, Meyer, and Wickerhauser [182] by generalizing the link between multiresolution approximations and wavelets. A space  $\mathbf{V}_j$  of a multiresolution approximation is decomposed in a lower-resolution space  $\mathbf{V}_{j+1}$  plus a detail space  $\mathbf{W}_{j+1}$ . This is done by dividing the orthogonal basis  $\{\phi_j(t - 2^j n)\}_{n \in \mathbb{Z}}$  of  $\mathbf{V}_j$  into two new orthogonal bases

$$\{\phi_{j+1}(t - 2^{j+1} n)\}_{n \in \mathbb{Z}} \text{ of } \mathbf{V}_{j+1} \quad \text{and} \quad \{\psi_{j+1}(t - 2^{j+1} n)\}_{n \in \mathbb{Z}} \text{ of } \mathbf{W}_{j+1}.$$

The decompositions (7.107) and (7.109) of  $\phi_{j+1}$  and  $\psi_{j+1}$  in the basis  $\{\phi_j(t - 2^j n)\}_{n \in \mathbb{Z}}$  are specified by a pair of conjugate mirror filters  $h[n]$  and

$$g[n] = (-1)^{1-n} h[1-n].$$

Theorem 8.1 generalizes this result to any space  $\mathbf{U}_j$  that admits an orthogonal basis of functions translated by  $n2^j$  for  $n \in \mathbb{Z}$ .

**Theorem 8.1:** *Coifman, Meyer, Wickerhauser.* Let  $\{\theta_j(t - 2^j n)\}_{n \in \mathbb{Z}}$  be an orthonormal basis of a space  $\mathbf{U}_j$ . Let  $h$  and  $g$  be a pair of conjugate mirror filters. Define

$$\theta_{j+1}^0(t) = \sum_{n=-\infty}^{+\infty} h[n] \theta_j(t - 2^j n) \quad \text{and} \quad \theta_{j+1}^1(t) = \sum_{n=-\infty}^{+\infty} g[n] \theta_j(t - 2^j n). \quad (8.1)$$

The family

$$\{\theta_{j+1}^0(t - 2^{j+1} n), \theta_{j+1}^1(t - 2^{j+1} n)\}_{n \in \mathbb{Z}}$$

is an orthonormal basis of  $\mathbf{U}_j$ .

**Proof.** This proof is very similar to the proof of Theorem 7.3. The main steps are outlined. Theorem 3.4 shows that  $\{\theta_j(t - 2^j n)\}_{n \in \mathbb{Z}}$  is orthogonal if and only if

$$\frac{1}{2^j} \sum_{k=-\infty}^{+\infty} \left| \hat{\theta}_j \left( \omega + \frac{2k\pi}{2^j} \right) \right|^2 = 1. \quad (8.2)$$

We derive from (8.1) that the Fourier transform of  $\theta_{j+1}^0$  is

$$\hat{\theta}_{j+1}^0(\omega) = \hat{\theta}_j(\omega) \sum_{n=-\infty}^{+\infty} h[n] \exp(-i2^j n\omega) = \hat{h}(2^j \omega) \hat{\theta}_j(\omega). \quad (8.3)$$

Similarly, the Fourier transform of  $\theta_{j+1}^1$  is

$$\hat{\theta}_{j+1}^1(\omega) = \hat{g}(2^j \omega) \hat{\theta}_j(\omega). \quad (8.4)$$

Proving that  $\{\theta_{j+1}^0(t - 2^{j+1} n)\}$  and  $\{\theta_{j+1}^1(t - 2^{j+1} n)\}_{n \in \mathbb{Z}}$  are two families of orthogonal vectors is equivalent to showing that for  $l = 0$  or  $l = 1$

$$\frac{1}{2^{j+1}} \sum_{k=-\infty}^{+\infty} \left| \hat{\theta}_{j+1}^l \left( \omega + \frac{2k\pi}{2^{j+1}} \right) \right|^2 = 1. \quad (8.5)$$

These two families of vectors yield orthogonal spaces if and only if

$$\frac{1}{2^{j+1}} \sum_{k=-\infty}^{+\infty} \hat{\theta}_{j+1}^0 \left( \omega + \frac{2k\pi}{2^{j+1}} \right) \hat{\theta}_{j+1}^{1*} \left( \omega + \frac{2k\pi}{2^{j+1}} \right) = 0. \quad (8.6)$$

The relations (8.5) and (8.6) are verified by replacing  $\hat{\theta}_{j+1}^0$  and  $\hat{\theta}_{j+1}^1$  by (8.3) and (8.4), respectively, and by using the orthogonality of the basis (8.2) and the conjugate mirror filter properties

$$|\hat{h}(\omega)|^2 + |\hat{h}(\omega + \pi)|^2 = 2,$$

$$|\hat{g}(\omega)|^2 + |\hat{g}(\omega + \pi)|^2 = 2,$$

$$\hat{g}(\omega) \hat{h}^*(\omega) + \hat{g}(\omega + \pi) \hat{h}^*(\omega + \pi) = 0.$$

To prove that the family  $\{\theta_{j+1}^0(t - 2^{j+1}n), \theta_{j+1}^1(t - 2^{j+1}n)\}_{n \in \mathbb{Z}}$  generates the same space as  $\{\theta_j(t - 2^j n)\}_{n \in \mathbb{Z}}$ , we must prove that for any  $a[n] \in \ell^2(\mathbb{Z})$  there exist  $b[n] \in \ell^2(\mathbb{Z})$  and  $c[n] \in \ell^2(\mathbb{Z})$  such that

$$\sum_{n=-\infty}^{+\infty} a[n] \theta_j(t - 2^j n) = \sum_{n=-\infty}^{+\infty} b[n] \theta_{j+1}^0(t - 2^{j+1}n) + \sum_{n=-\infty}^{+\infty} c[n] \theta_{j+1}^1(t - 2^{j+1}n). \quad (8.7)$$

To do this, we relate  $\hat{b}(\omega)$  and  $\hat{c}(\omega)$  to  $\hat{a}(\omega)$ . The Fourier transform of (8.7) yields

$$\hat{a}(2^j \omega) \hat{\theta}_j(\omega) = \hat{b}(2^{j+1} \omega) \hat{\theta}_{j+1}^0(\omega) + \hat{c}(2^{j+1} \omega) \hat{\theta}_{j+1}^1(\omega). \quad (8.8)$$

One can verify that

$$\hat{b}(2^{j+1} \omega) = \frac{1}{2} \left( \hat{a}(2^j \omega) \hat{h}^*(2^j \omega) + \hat{a}(2^j \omega + \pi) \hat{h}^*(2^j \omega + \pi) \right)$$

and

$$\hat{c}(2^{j+1} \omega) = \frac{1}{2} \left( \hat{a}(2^j \omega) \hat{g}^*(2^j \omega) + \hat{a}(2^j \omega + \pi) \hat{g}^*(2^j \omega + \pi) \right)$$

satisfy (8.8). ■

Theorem 8.1 proves that conjugate mirror filters transform an orthogonal basis  $\{\theta_j(t - 2^j n)\}_{n \in \mathbb{Z}}$  in two orthogonal families  $\{\theta_{j+1}^0(t - 2^{j+1}n)\}_{n \in \mathbb{Z}}$  and  $\{\theta_{j+1}^1(t - 2^{j+1}n)\}_{n \in \mathbb{Z}}$ . Let  $\mathbf{U}_{j+1}^0$  and  $\mathbf{U}_{j+1}^1$  be the spaces generated by each of these families. Clearly  $\mathbf{U}_{j+1}^0$  and  $\mathbf{U}_{j+1}^1$  are orthogonal and

$$\mathbf{U}_{j+1}^0 \oplus \mathbf{U}_{j+1}^1 = \mathbf{U}_j.$$

Computing the Fourier transform of (8.1) relates the Fourier transforms of  $\theta_{j+1}^0$  and  $\theta_{j+1}^1$  to the Fourier transform of  $\theta_j$ :

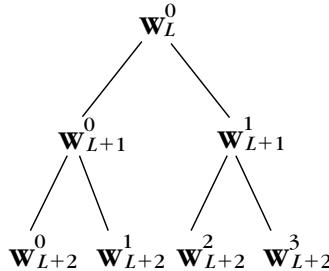
$$\hat{\theta}_{j+1}^0(\omega) = \hat{h}(2^j \omega) \hat{\theta}_j(\omega), \quad \hat{\theta}_{j+1}^1(\omega) = \hat{g}(2^j \omega) \hat{\theta}_j(\omega). \quad (8.9)$$

Since the transfer functions  $\hat{h}(2^j \omega)$  and  $\hat{g}(2^j \omega)$  have their energy concentrated in different frequency intervals, this transformation can be interpreted as a division of the frequency support of  $\hat{\theta}_j$ .

### **Binary Wavelet Packet Tree**

Instead of dividing only the approximation spaces  $\mathbf{V}_j$  to construct detail spaces  $\mathbf{W}_j$  and wavelet bases, Theorem 8.1 proves that we can set  $\mathbf{U}_j = \mathbf{W}_j$  and divide these detail spaces to derive new bases. The recursive splitting of vector spaces is represented in a binary tree. If the signals are approximated at the scale  $2^L$ , to the root of the tree we associate the approximation space  $\mathbf{V}_L$ . This space admits an orthogonal basis of scaling functions  $\{\phi_L(t - 2^L n)\}_{n \in \mathbb{Z}}$  with  $\phi_L(t) = 2^{-L/2} \phi(2^{-L} t)$ .

Any node of the binary tree is labeled by  $(j, p)$ , where  $j - L \geq 0$  is the depth of the node in the tree, and  $p$  is the number of nodes that are on its left at the same

**FIGURE 8.1**

Binary tree of wavelet packet spaces.

depth  $j - L$ . Such a tree is illustrated in Figure 8.1. To each node  $(j, p)$  we associate a space  $\mathbf{W}_j^p$ , which admits an orthonormal basis  $\{\psi_j^p(t - 2^j n)\}_{n \in \mathbb{Z}}$  by going down the tree. At the root we have  $\mathbf{W}_L^0 = \mathbf{V}_L$  and  $\psi_L^0 = \phi_L$ . Suppose now that we have already constructed  $\mathbf{W}_j^p$  and its orthonormal basis  $\mathcal{B}_j^p = \{\psi_j^p(t - 2^j n)\}_{n \in \mathbb{Z}}$  at the node  $(j, p)$ . The two wavelet packet orthogonal bases at the children nodes are defined by the splitting relations (8.1):

$$\psi_{j+1}^{2p}(t) = \sum_{n=-\infty}^{+\infty} h[n] \psi_j^p(t - 2^j n) \quad (8.10)$$

and

$$\psi_{j+1}^{2p+1}(t) = \sum_{n=-\infty}^{+\infty} g[n] \psi_j^p(t - 2^j n). \quad (8.11)$$

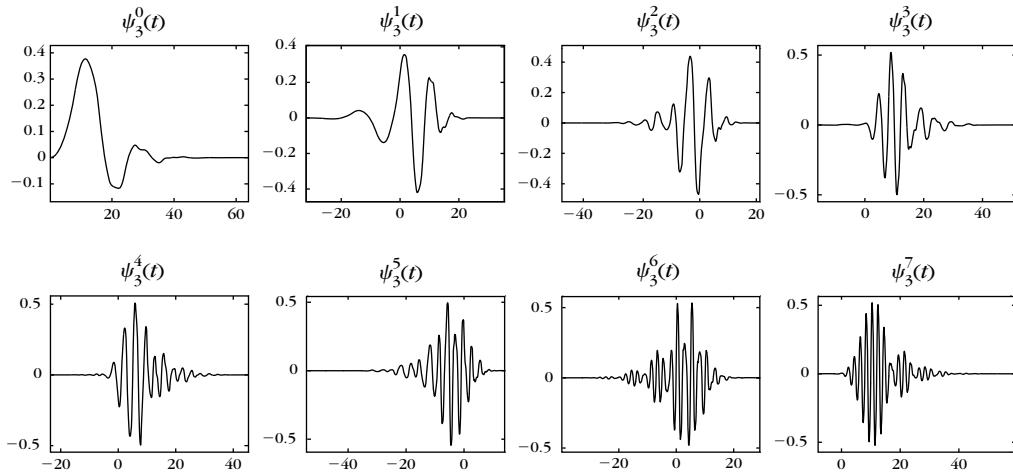
Since  $\{\psi_j^p(t - 2^j n)\}_{n \in \mathbb{Z}}$  is orthonormal,

$$h[n] = \langle \psi_{j+1}^{2p}(u), \psi_j^p(u - 2^j n) \rangle, \quad g[n] = \langle \psi_{j+1}^{2p+1}(u), \psi_j^p(u - 2^j n) \rangle. \quad (8.12)$$

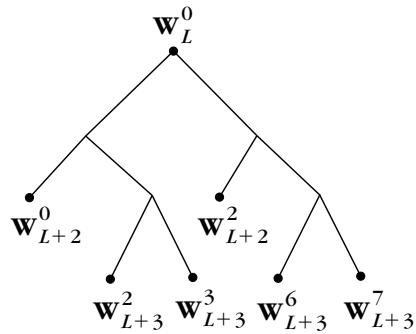
Theorem 8.1 proves that  $\mathcal{B}_{j+1}^{2p} = \{\psi_{j+1}^{2p}(t - 2^{j+1} n)\}_{n \in \mathbb{Z}}$  and  $\mathcal{B}_{j+1}^{2p+1} = \{\psi_{j+1}^{2p+1}(t - 2^{j+1} n)\}_{n \in \mathbb{Z}}$  are orthonormal bases of two orthogonal spaces  $\mathbf{W}_{j+1}^{2p}$  and  $\mathbf{W}_{j+1}^{2p+1}$  such that

$$\mathbf{W}_{j+1}^{2p} \oplus \mathbf{W}_{j+1}^{2p+1} = \mathbf{W}_j^p. \quad (8.13)$$

This recursive splitting defines a binary tree of wavelet packet spaces where each parent node is divided in two orthogonal subspaces. Figure 8.2 displays the eight wavelet packets  $\psi_j^p$  at the depth  $j - L = 3$ , calculated with a Daubechies 5 filter. These wavelet packets are frequency ordered from left to right, as explained in Section 8.1.2.

**FIGURE 8.2**

Wavelet packets computed with a Daubechies 5 filter at the depth  $j - L = 3$  of the wavelet packet tree, with  $L = 0$ . They are ordered from low to high frequencies.

**FIGURE 8.3**

Example of an admissible wavelet packet binary tree.

### **Admissible Tree**

We call any binary tree where each node has either zero or two children an *admissible tree*, as shown in Figure 8.3. Let  $\{j_i, p_i\}_{1 \leq i \leq I}$  be the leaves of an admissible binary tree. By applying the recursive splitting (8.13) along the branches of an admissible tree, we verify that the spaces  $\{\mathbf{W}_{j_i}^{p_i}\}_{1 \leq i \leq I}$  are mutually orthogonal and add up to  $\mathbf{W}_L^0$ :

$$\mathbf{W}_L^0 = \bigoplus_{i=1}^I \mathbf{W}_{j_i}^{p_i}. \quad (8.14)$$

The union of the corresponding wavelet packet bases

$$\{\psi_{j_i}^{p_i}(t - 2^{j_i}n)\}_{n \in \mathbb{Z}, 1 \leq i \leq I}$$

thus defines an orthogonal basis of  $\mathbf{W}_L^0 = \mathbf{V}_L$ .

### **Number of Wavelet Packet Bases**

The number of different wavelet packet orthogonal bases of  $\mathbf{V}_L$  is equal to the number of different admissible binary trees. Theorem 8.2 proves that there are more than  $2^{2^{J-1}}$  different wavelet packet orthonormal bases included in a full wavelet packet binary tree of depth  $J$ .

**Theorem 8.2.** The number  $B_J$  of wavelet packet bases in a full wavelet packet binary tree of depth  $J$  satisfies

$$2^{2^{J-1}} \leq B_J \leq 2^{\frac{5}{4}2^{J-1}}. \quad (8.15)$$

**Proof.** This result is proved by induction on the depth  $J$  of the wavelet packet tree. The number  $B_J$  of different orthonormal bases is equal to the number of different admissible binary trees of depth of at most  $J$ , which have nodes with either zero or two children. For  $J = 0$ , the tree is reduced to its root, so  $B_0 = 1$ .

Observe that the set of trees of depth of at most  $J + 1$  is composed of trees of depth of at least 1 and at most  $J + 1$  plus one tree of depth 0 that is reduced to the root. A tree of depth of at least 1 has a left and a right subtree that are admissible trees of depth of at most  $J$ . The configuration of these trees is a priori independent and there are  $B_J$  admissible trees of depth  $J$ , so

$$B_{J+1} = B_J^2 + 1. \quad (8.16)$$

Since  $B_1 = 2$  and  $B_{J+1} \geq B_J^2$ , we prove by induction that  $B_J \geq 2^{2^{J-1}}$ . Moreover,

$$\log_2 B_{J+1} = 2 \log_2 B_J + \log_2(1 + B_J^{-2}).$$

If  $J \geq 1$ , then  $B_J \geq 2$ , so

$$\log_2 B_{J+1} \leq 2 \log_2 B_J + \frac{1}{4}. \quad (8.17)$$

Since  $B_1 = 2$ ,

$$\log_2 B_{J+1} \leq 2^J + \frac{1}{4} \sum_{j=0}^{J-1} 2^j \leq 2^J + \frac{2^J}{4},$$

$$\text{so } B_J \leq 2^{\frac{5}{4}2^{J-1}}. \quad \blacksquare$$

For discrete signals of size  $N$ , we shall see that the wavelet packet tree is at most of depth  $J = \log_2 N$ . This theorem proves that the number of wavelet packet bases satisfies  $2^{N/2} \leq B_{\log_2 N} \leq 2^{5N/8}$ .

### Wavelet Packets on Intervals

To construct wavelet packet bases of  $L^2[0, 1]$ , we use the border techniques developed in Section 7.5 to design wavelet bases of  $L^2[0, 1]$ . The simplest approach constructs periodic bases. As in the wavelet case, the coefficients of  $f \in L^2[0, 1]$  in a periodic wavelet packet basis are the same as the decomposition coefficients of  $f^{\text{Per}}(t) = \sum_{k=-\infty}^{+\infty} f(t+k)$  in the original wavelet packet basis of  $L^2(\mathbb{R})$ . The periodization of  $f$  often creates discontinuities at the borders  $t = 0$  and  $t = 1$ , which generate large-amplitude wavelet packet coefficients.

Section 7.5.3 describes a more sophisticated technique that modifies the filters  $h$  and  $g$  in order to construct boundary wavelets that keep their vanishing moments. A generalization to wavelet packets is obtained by using these modified filters in Theorem 8.1. This avoids creating the large-amplitude coefficients at the boundary, which is typical of the periodic case.

### Biorthogonal Wavelet Packets

Nonorthogonal wavelet bases are constructed in Section 7.4 with two pairs of perfect reconstruction filters  $(h, g)$  and  $(\tilde{h}, \tilde{g})$  instead of a single pair of conjugate mirror filters. The orthogonal splitting Theorem 8.1 is extended into a biorthogonal splitting by replacing the conjugate mirror filters with these perfect reconstruction filters. A Riesz basis  $\{\theta_j(t - 2^j n)\}_{n \in \mathbb{Z}}$  of  $\mathbf{U}_j$  is transformed into two Riesz bases  $\{\theta_{j+1}^0(t - 2^{j+1} n)\}_{n \in \mathbb{Z}}$  and  $\{\theta_{j+1}^1(t - 2^{j+1} n)\}_{n \in \mathbb{Z}}$  of two nonorthogonal spaces  $\mathbf{U}_{j+1}^0$  and  $\mathbf{U}_{j+1}^1$  such that

$$\mathbf{U}_{j+1}^0 \oplus \mathbf{U}_{j+1}^1 = \mathbf{U}_j.$$

A binary tree of nonorthogonal wavelet packet Riesz bases can be derived by induction using this vector space division. As in the orthogonal case, the wavelet packets at the leaves of an admissible binary tree define a basis of  $\mathbf{W}_L^0$ , but this basis is not orthogonal.

The lack of orthogonality is not a problem by itself as long as the basis remains stable. Cohen and Daubechies proved [171] that when the depth  $j - L$  increases, the angle between the spaces  $\mathbf{W}_j^p$  located at the same depth can become progressively smaller. This indicates that some of the wavelet packet bases constructed from an admissible binary tree become unstable. Thus, we concentrate on orthogonal wavelet packets constructed with conjugate mirror filters.

#### 8.1.2 Time-Frequency Localization

##### Time Support

If the conjugate mirror filters  $h$  and  $g$  have a finite impulse response of size  $K$ , Theorem 7.5 proves that  $\phi$  has a support size of  $K - 1$ , so  $\psi_L^0 = \phi_L$  has a support size of  $(K - 1)2^L$ . Since

$$\psi_{j+1}^{2p}(t) = \sum_{n=-\infty}^{+\infty} h[n] \psi_j^p(t - 2^j n), \quad \psi_{j+1}^{2p+1}(t) = \sum_{n=-\infty}^{+\infty} g[n] \psi_j^p(t - 2^j n), \quad (8.18)$$

an induction on  $j$  shows that the support size of  $\psi_j^p$  is  $(K - 1)2^j$ . Thus, the parameter  $j$  specifies the scale  $2^j$  of the support. The wavelet packets in Figure 8.2 are constructed with a Daubechies filter of  $K = 10$  coefficients with  $j = 3$  and thus have a support size of  $2^3(10 - 1) = 72$ .

### Frequency Localization

The frequency localization of wavelet packets is more complicated to analyze. The Fourier transform of (8.18) proves that the Fourier transforms of wavelet packet children are related to their parent by

$$\hat{\psi}_{j+1}^{2p}(\omega) = \hat{h}(2^j\omega) \hat{\psi}_j^p(\omega), \quad \hat{\psi}_{j+1}^{2p+1}(\omega) = \hat{g}(2^j\omega) \hat{\psi}_j^p(\omega). \quad (8.19)$$

The energy of  $\hat{\psi}_j^p$  is mostly concentrated over a frequency band and the two filters  $\hat{h}(2^j\omega)$  and  $\hat{g}(2^j\omega)$  select the lower- or higher-frequency components within this band. To relate the size and position of this frequency band to the indexes  $(p, j)$ , we consider a simple example.

### Shannon Wavelet Packets

Shannon wavelet packets are computed with perfect discrete low-pass and high-pass filters

$$|\hat{h}(\omega)| = \begin{cases} \sqrt{2} & \text{if } \omega \in [-\pi/2 + 2k\pi, \pi/2 + 2k\pi] \text{ with } k \in \mathbb{Z} \\ 0 & \text{otherwise} \end{cases} \quad (8.20)$$

and

$$|\hat{g}(\omega)| = \begin{cases} \sqrt{2} & \text{if } \omega \in [\pi/2 + 2k\pi, 3\pi/2 + 2k\pi] \text{ with } k \in \mathbb{Z} \\ 0 & \text{otherwise.} \end{cases} \quad (8.21)$$

In this case, it is relatively simple to calculate the frequency support of the wavelet packets. The Fourier transform of the scaling function is

$$\hat{\phi}_L^0 = \hat{\phi}_L = \mathbf{1}_{[-2^{-L}\pi, 2^{-L}\pi]}. \quad (8.22)$$

Each multiplication with  $\hat{h}(2^j\omega)$  or  $\hat{g}(2^j\omega)$  divides the frequency support of the wavelet packets in two. The delicate point is to realize that  $\hat{h}(2^j\omega)$  does not always play the role of a low-pass filter because of the side lobes that are brought into the interval  $[-2^{-L}\pi, 2^{-L}\pi]$  by the dilation. At depth  $j - L$ , Theorem 8.3 proves that  $\hat{\psi}_j^p$  is proportional to the indicator function of a pair of frequency intervals, which are labeled  $I_j^k$ . The permutation that relates  $p$  and  $k$  is characterized recursively [71].

**Theorem 8.3:** Coifman, Wickerhauser. For any  $j - L > 0$  and  $0 \leq p < 2^{j-L}$ , there exists  $0 \leq k < 2^{j-L}$  such that

$$|\hat{\psi}_j^p(\omega)| = 2^{j/2} \mathbf{1}_{I_j^k}(\omega), \quad (8.23)$$

where  $I_j^k$  is a symmetric pair of intervals

$$I_j^k = [-(k+1)\pi 2^{-j}, -k\pi 2^{-j}] \cup [k\pi 2^{-j}, (k+1)\pi 2^{-j}]. \quad (8.24)$$

The permutation  $k = G[p]$  satisfies for any  $0 \leq p < 2^{j-L}$

$$G[2p] = \begin{cases} 2G[p] & \text{if } G[p] \text{ is even} \\ 2G[p] + 1 & \text{if } G[p] \text{ is odd} \end{cases} \quad (8.25)$$

$$G[2p+1] = \begin{cases} 2G[p] + 1 & \text{if } G[p] \text{ is even} \\ 2G[p] & \text{if } G[p] \text{ is odd.} \end{cases} \quad (8.26)$$

**Proof.** Equations (8.23), (8.25), and (8.26) are proved by induction on depth  $j-L$ . For  $j-L=0$ , (8.22) shows that (8.23) is valid. Suppose that (8.23) is valid for  $j=l \geq L$  and any  $0 \leq p < 2^{l-L}$ . We first prove that (8.25) and (8.26) are verified for  $j=l$ . From these two equations we then easily carry the induction hypothesis to prove that (8.23) is true for  $j=l+1$  and for any  $0 \leq p < 2^{l+1-L}$ .

Equations (8.20) and (8.21) imply that

$$|\hat{h}(2^l \omega)| = \begin{cases} \sqrt{2} & \text{if } \omega \in [-2^{-l-1}(4m-1)\pi, 2^{-l-1}(4m+1)\pi] \text{ with } m \in \mathbb{Z} \\ 0 & \text{otherwise} \end{cases} \quad (8.27)$$

$$|\hat{g}(2^l \omega)| = \begin{cases} \sqrt{2} & \text{if } \omega \in [-2^{-l-1}(4m+1)\pi, 2^{-l-1}(4m+3)\pi] \text{ with } m \in \mathbb{Z} \\ 0 & \text{otherwise.} \end{cases} \quad (8.28)$$

Since (8.23) is valid for  $l$ , the support of  $\hat{\psi}_l^p$  is

$$I_l^k = [-(2k+2)\pi 2^{-l-1}, -2k\pi 2^{-l-1}] \cup [2k\pi 2^{-l-1}, (2k+2)\pi 2^{-l-1}].$$

The two children are defined by

$$\hat{\psi}_{l+1}^{2p}(\omega) = \hat{h}(2^l \omega) \hat{\psi}_l^p(\omega), \quad \hat{\psi}_{l+1}^{2p+1}(\omega) = \hat{g}(2^l \omega) \hat{\psi}_l^p(\omega).$$

Thus, we derive (8.25) and (8.26) by checking the intersection of  $I_l^k$  with the supports of  $\hat{h}(2^l \omega)$  and  $\hat{g}(2^l \omega)$  specified by (8.27) and (8.28). ■

For Shannon wavelet packets, Theorem 8.3 proves that  $\hat{\psi}_j^p$  has a frequency support located over two intervals of size  $2^{-j}\pi$ , centered at  $\pm(k+1/2)\pi 2^{-j}$ . The Fourier transform expression (8.23) implies that these Shannon wavelet packets can be written as cosine-modulated windows

$$\psi_j^p(t) = 2^{-j/2+1} \theta(2^{-j}t) \cos[2^{-j}\pi(k+1/2)(t-\tau_{j,p})], \quad (8.29)$$

with

$$\theta(t) = \frac{\sin(\pi t/2)}{\pi t},$$

and thus,

$$\hat{\theta}(\omega) = \mathbf{1}_{[-\pi/2, \pi/2]}(\omega).$$

The translation parameter  $\tau_{j,p}$  can be calculated from the complex phase of  $\hat{\psi}_j^p$ .

### Frequency Ordering

It is often easier to label  $\psi_j^k$  a wavelet packet  $\psi_j^p$  that has a Fourier transform centered at  $\pm(k + 1/2)\pi 2^{-j}$  with  $k = G[p]$ . This means changing its position in the wavelet packet tree from node  $p$  to node  $k$ . The resulting wavelet packet tree is frequency ordered. The left child always corresponds to a lower-frequency wavelet packet and the right child to a higher-frequency one.

The permutation  $k = G[p]$  is characterized by the recursive equations (8.25) and (8.26). The inverse permutation  $p = G^{-1}[k]$  is called a *Gray code* in coding theory. This permutation is implemented on binary strings by deriving the following relations from (8.25) and (8.26). If  $p_i$  is the  $i^{\text{th}}$  binary digit of the integer  $p$  and  $k_i$  the  $i^{\text{th}}$  digit of  $k = G[p]$ , then

$$k_i = \left( \sum_{l=i}^{+\infty} p_l \right) \bmod 2, \quad (8.30)$$

$$p_i = (k_i + k_{i+1}) \bmod 2. \quad (8.31)$$

### Compactly Supported Wavelet Packets

Wavelet packets of compact support have a more complicated frequency behavior than Shannon wavelet packets, but the previous analysis provides important insights. If  $h$  is a finite impulse response filter,  $\hat{h}$  does not have a support restricted to  $[-\pi/2, \pi/2]$  over the interval  $[-\pi, \pi]$ . It is true, however, that the energy of  $\hat{h}$  is mostly concentrated in  $[-\pi/2, \pi/2]$ . Similarly, the energy of  $\hat{g}$  is mostly concentrated in  $[-\pi, -\pi/2] \cup [\pi/2, \pi]$ , for  $\omega \in [-\pi, \pi]$ . As a consequence, the localization properties of Shannon wavelet packets remain qualitatively valid. The energy of  $\hat{\psi}_j^p$  is mostly concentrated over

$$I_j^k = [-(k + 1)\pi 2^{-j}, -k\pi 2^{-j}] \cup [k\pi 2^{-j}, (k + 1)\pi 2^{-j}],$$

with  $k = G[p]$ . The larger the proportion of energy of  $\hat{h}$  in  $[-\pi/2, \pi/2]$ , the more concentrated the energy of  $\hat{\psi}_j^p$  in  $I_j^k$ . The energy concentration of  $\hat{h}$  in  $[-\pi/2, \pi/2]$  is increased by having more zeroes at  $\pi$ , so that  $\hat{h}(\omega)$  remains close to zero in  $[-\pi, -\pi/2] \cup [\pi/2, \pi]$ . Theorem 7.4 proves that this is equivalent to imposing that the wavelets constructed in the wavelet packet tree have many vanishing moments.

These qualitative statements must be interpreted carefully. The side lobes of  $\hat{\psi}_j^p$  beyond the intervals  $I_j^k$  are not completely negligible. For example, wavelet packets created with a Haar filter are discontinuous functions. Thus,  $|\hat{\psi}_j^p(\omega)|$  decays like  $|\omega|^{-1}$  at high frequencies, which indicates the existence of large side lobes outside  $I_j^k$ . It is also important to note that contrary to Shannon wavelet packets, compactly supported wavelet packets cannot be written as dilated windows modulated by cosine functions of varying frequency. When the scale increases, wavelet packets generally do not converge to cosine functions. They may have a wild behavior with localized oscillations of considerable amplitude.

### Walsh Wavelet Packets

Walsh wavelet packets are generated by the Haar conjugate mirror filter

$$h[n] = \begin{cases} \frac{1}{\sqrt{2}} & \text{if } n = 0, 1 \\ 0 & \text{otherwise.} \end{cases}$$

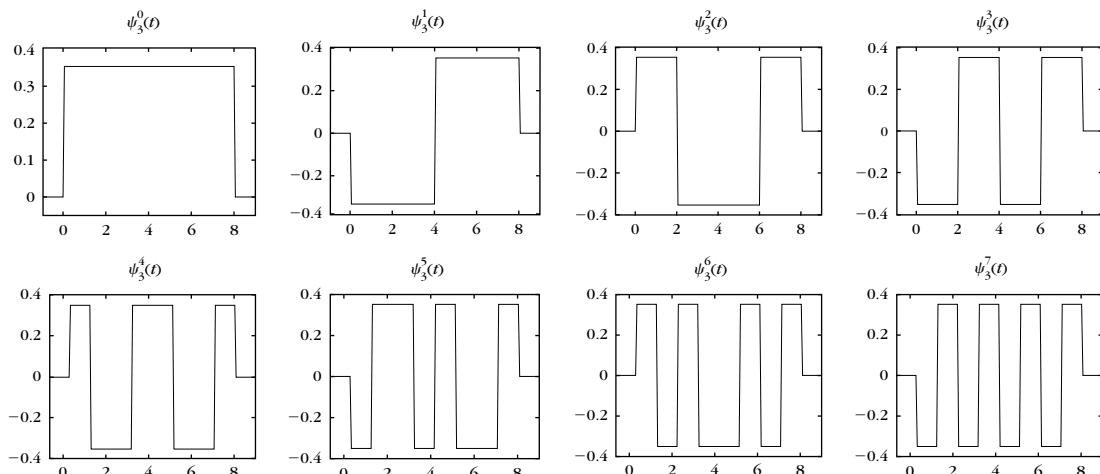
They have very different properties from Shannon wavelet packets since the filter  $h$  is well localized in time but not in frequency. The corresponding scaling function is  $\phi = \mathbf{1}_{[0, 1]}$  and the approximation space  $\mathbf{V}_L = \mathbf{W}_L^0$  is composed of functions that are constant over the intervals  $[2^L n, 2^L(n+1))$  for  $n \in \mathbb{Z}$ . Since all wavelet packets created with this filter belong to  $\mathbf{V}_L$ , they are piecewise constant functions. The support size of  $h$  is  $K = 2$ , so Walsh functions  $\psi_j^p$  have a support size of  $2^j$ . The wavelet packet recursive relations (8.18) become

$$\psi_{j+1}^{2p}(t) = \frac{1}{\sqrt{2}} \psi_j^p(t) + \frac{1}{\sqrt{2}} \psi_j^p(t - 2^j), \quad (8.32)$$

and

$$\psi_{j+1}^{2p+1}(t) = \frac{1}{\sqrt{2}} \psi_j^p(t) - \frac{1}{\sqrt{2}} \psi_j^p(t - 2^j). \quad (8.33)$$

Since  $\psi_j^p$  has a support size of  $2^j$ , it does not intersect the support of  $\psi_j^p(t - 2^j)$ . Thus, these wavelet packets are constructed by juxtaposing  $\psi_j^p$  with its translated version that has a sign that might be changed. Figure 8.4 shows the Walsh functions at depth  $j - L = 3$  of the wavelet packet tree. Theorem 8.4 computes the number of oscillations of  $\psi_j^p$ .



**FIGURE 8.4**

Frequency-ordered Walsh wavelet packets computed with a Haar filter at depth  $j - L = 3$  of the wavelet packet tree, with  $L = 0$ .

**Theorem 8.4.** The support of a Walsh wavelet packet  $\psi_j^P$  is  $[0, 2^j]$ . Over its support,  $\psi_j^P(t) = \pm 2^{-j/2}$ . It changes sign  $k = G[p]$  times, where  $G[p]$  is the permutation defined by (8.25) and (8.26).

**Proof.** By induction on  $j$ , we derive from (8.32) and (8.33) that the support is  $[0, 2^j]$  and that  $\psi_j^P(t) = \pm 2^{-j/2}$  over its support. Let  $k$  be the number of times that  $\psi_j^P$  changes sign.

The number of times that  $\psi_{j+1}^{2p}$  and  $\psi_{j+1}^{2p+1}$  change sign is either  $2k$  or  $2k + 1$  depending on the sign of the first and last nonzero values of  $\psi_j^P$ . If  $k$  is even, then the sign of the first and last nonzero values of  $\psi_j^P$  are the same. Thus, the number of times  $\psi_{j+1}^{2p}$  and  $\psi_{j+1}^{2p+1}$  change sign is, respectively,  $2k$  and  $2k + 1$ . If  $k$  is odd, then the sign of the first and last nonzero values of  $\psi_j^P$  are different. The number of times  $\psi_{j+1}^{2p}$  and  $\psi_{j+1}^{2p+1}$  change sign is then  $2k + 1$  and  $2k$ . These recursive properties are identical to (8.25) and (8.26). ■

Therefore, a Walsh wavelet packet  $\psi_j^P$  is a square wave with  $k = G[p]$  oscillations over a support size of  $2^j$ . This result is similar to (8.29), which proves that a Shannon wavelet packet  $\psi_j^P$  is a window modulated by a cosine of frequency  $2^{-j}k\pi$ . In both cases, the oscillation frequency of wavelet packets is proportional to  $2^{-j}k$ .

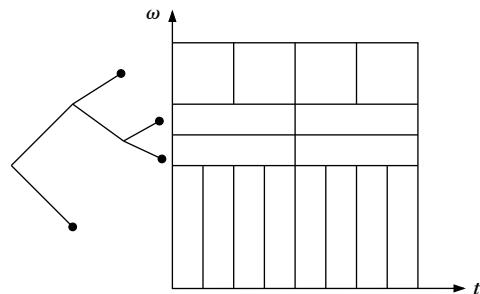
### Heisenberg Boxes

For display purposes, we associate to any wavelet packet  $\psi_j^P(t - 2^j n)$  a Heisenberg rectangle, which indicates the time and frequency domains where the energy of this wavelet packet is mostly concentrated. The time support of the rectangle is set to be the same as the time support of a Walsh wavelet packet  $\psi_j^P(t - 2^j n)$ , which is equal to  $[2^j n, 2^j(n + 1)]$ . The frequency support of the rectangle is defined as the positive-frequency support  $[k\pi 2^{-j}, (k + 1)\pi 2^{-j}]$  of Shannon wavelet packets with  $k = G[p]$ . The scale  $2^j$  modifies the time and frequency elongation of this time-frequency rectangle, but its surface remains constant. The indices  $n$  and  $k$  give its localization in time and frequency. General wavelet packets—for example, computed with Daubechies filters—have a time and frequency spread that is much wider than this Heisenberg rectangle. However, this convention has the advantage of associating a wavelet packet basis to an exact paving of the time-frequency plane. Figure 8.5 shows an example of such a paving and the corresponding wavelet packet tree.

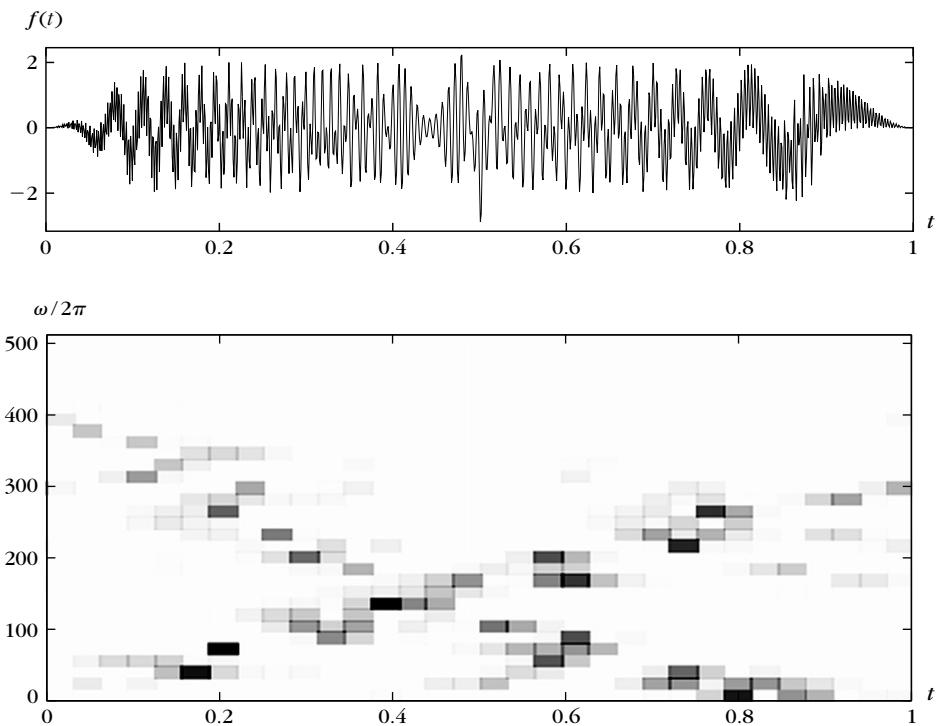
Figure 8.6 displays the decomposition of a multichirp signal having a spectrogram shown in Figure 4.3. The wavelet packet basis is computed with a Daubechies 10 filter. As expected, the large-amplitude coefficients are along the trajectory of linear and quadratic chirps that appear in Figure 4.3. We also see the trace of the two modulated Gaussian functions located at  $t = 0.5$  and  $t = 0.87$ .

### 8.1.3 Particular Wavelet Packet Bases

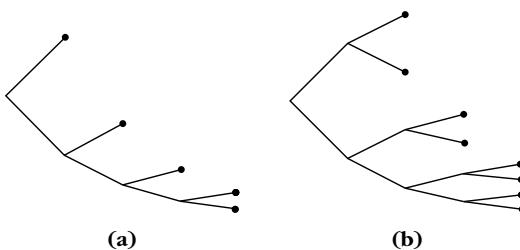
Among the many wavelet packet bases, we describe here the properties of  $M$ -band wavelet bases, “local cosine type” bases, and “best” bases. The wavelet packet

**FIGURE 8.5**

The wavelet packet tree (*left*) divides the frequency axis in several intervals. The Heisenberg boxes (*right*) of the corresponding wavelet packet basis.

**FIGURE 8.6**

Wavelet packet decomposition of the multichirp signal the spectrogram of which is shown in Figure 4.3. The darker the gray level of each Heisenberg box, the larger the amplitude  $|\langle f, \psi_j^B \rangle|$  of the corresponding wavelet packet coefficient.

**FIGURE 8.7**

(a) Wavelet packet tree of a dyadic wavelet basis. (b) Wavelet packet tree of an  $M$ -band wavelet basis with  $M = 2$ .

tree is frequency ordered, which means that  $\psi_j^k$  has a Fourier transform with an energy essentially concentrated in the interval  $[k\pi 2^{-j}, (k+1)\pi 2^{-j}]$  for positive frequencies.

### **M-Band Wavelets**

The standard dyadic wavelet basis is an example of a wavelet packet basis of  $\mathbf{V}_L$ , obtained by choosing the admissible binary tree shown in Figure 8.7(a). Its leaves are the nodes  $k = 1$  at all depth  $j - L$  and thus correspond to the wavelet packet basis

$$\{\psi_j^1(t - 2^j n)\}_{n \in \mathbb{Z}, j > L}$$

constructed by dilating a single wavelet  $\psi^1$ :

$$\psi_j^1(t) = \frac{1}{\sqrt{2^j}} \psi^1\left(\frac{t}{2^j}\right).$$

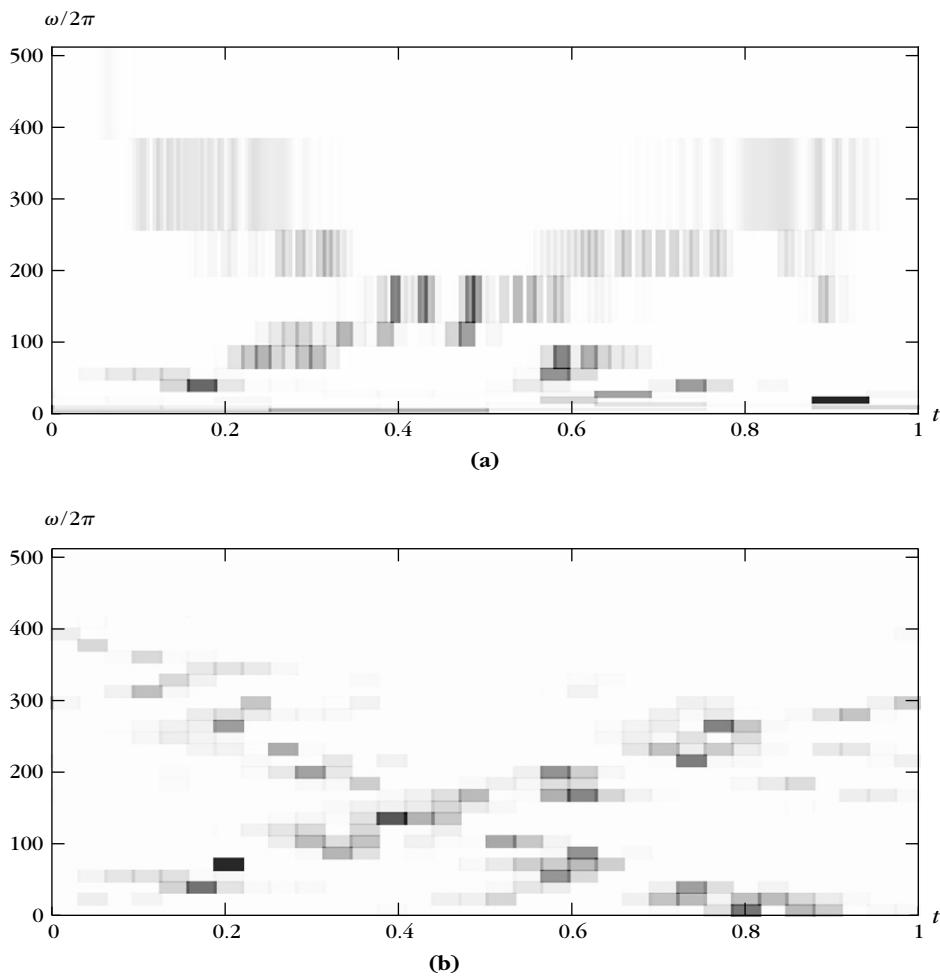
The energy of  $\hat{\psi}^1$  is mostly concentrated in the interval  $[-2\pi, -\pi] \cup [\pi, 2\pi]$ . The octave bandwidth for positive frequencies is the ratio between the bandwidth of the pass band and its distance to the zero frequency. It is equal to 1 octave. This quantity remains constant by dilation and specifies the frequency resolution of the wavelet transform.

Wavelet packets include other wavelet bases constructed with several wavelets having a better frequency resolution. Let us consider the admissible binary trees of Figure 8.7(b), which have leaves that are indexed by  $k = 2$  and  $k = 3$  at all depth  $j - L$ . The resulting wavelet packet basis of  $\mathbf{V}_L$  is

$$\{\psi_j^2(t - 2^j n), \psi_j^3(t - 2^j n)\}_{n \in \mathbb{Z}, j > L+1}.$$

These wavelet packets can be rewritten as dilations of two elementary wavelets  $\psi^2$  and  $\psi^3$ :

$$\psi_j^2(t) = \frac{1}{\sqrt{2^{j-1}}} \psi^2\left(\frac{t}{2^{j-1}}\right), \quad \psi_j^3(t) = \frac{1}{\sqrt{2^{j-1}}} \psi^3\left(\frac{t}{2^{j-1}}\right).$$

**FIGURE 8.8**

**(a)** Heisenberg boxes of a two-band wavelet decomposition of the multichirp signal shown in Figure 8.6. **(b)** Decomposition of the same signal in a pseudo-local cosine wavelet packet basis.

Over positive frequencies, the energy of  $\hat{\psi}^2$  and  $\hat{\psi}^3$  is mostly concentrated, respectively, in  $[\pi, 3\pi/2]$  and  $[3\pi/2, 2\pi]$ . Thus, the octave bandwidths of  $\hat{\psi}^2$  and  $\hat{\psi}^3$  are equal to  $1/2$  and  $1/3$ , respectively. Wavelets  $\hat{\psi}^2$  and  $\hat{\psi}^3$  have a higher-frequency resolution than  $\hat{\psi}^1$ , but their time support is twice as large. Figure 8.8(a) gives a two-band wavelet decomposition of the multichirp signal shown in Figure 8.6, calculated with a Daubechies 10 filter.

Higher-resolution wavelet bases can be constructed with an arbitrary number of  $M = 2^l$  wavelets. In a frequency-ordered wavelet packet tree, we define an admissible

binary tree with leaves indexed by  $2^l \leq k < 2^{l+1}$  at the depth  $j - L > l$ . The resulting wavelet packet basis

$$\{\psi_j^k(t - 2^j n)\}_{M \leq k < 2M, j > L+l}$$

can be written as dilations and translations of  $M$  elementary wavelets

$$\psi_j^k(t) = \frac{1}{\sqrt{2^{j-l}}} \psi^k\left(\frac{t}{2^{j-l}}\right).$$

The support size of  $\psi^k$  is proportional to  $M = 2^l$ . For positive frequencies, the energy of  $\hat{\psi}^k$  is mostly concentrated in  $[k\pi 2^{-l}, (k+1)\pi 2^{-l}]$ . The octave bandwidth is therefore  $\pi 2^{-l}/(k\pi 2^{-l}) = k^{-1}$  for  $M \leq k < 2M$ . The  $M$  wavelets  $\{\psi^k\}_{M \leq k < 2M}$  have an octave bandwidth smaller than  $M^{-1}$  but a time support  $M$  times larger than the support of  $\psi^1$ . Such wavelet bases are called *M-band wavelets*. More general families of M-band wavelets can also be constructed with other *M-band* filter banks studied in [68].

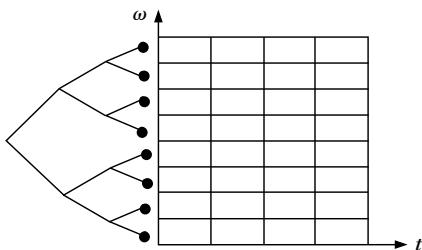
### Pseudo-Local Cosine Bases

Pseudo-local cosine bases are constructed with an admissible binary tree that is a full tree of depth  $J - L \geq 0$ . The leaves are the nodes indexed by  $0 \leq k < 2^{J-L}$  and the resulting wavelet packet basis is

$$\{\psi_J^k(t - 2^J n)\}_{n \in \mathbb{Z}, 0 \leq k < 2^{J-L}}. \quad (8.34)$$

If these wavelet packets are constructed with a conjugate mirror filter of size  $K$ , they have a support of size  $(K-1)2^J$ . For positive frequencies, the energy of  $\hat{\psi}_J^k$  is concentrated in  $[k\pi 2^{-J}, (k+1)\pi 2^{-J}]$ . Therefore, the bandwidth of all these wavelet packets is approximately constant and equal to  $\pi 2^{-J}$ . The Heisenberg boxes of these wavelet packets have the same size and divide the time-frequency plane in the rectangular grid illustrated in Figure 8.9.

Shannon wavelet packets  $\psi_J^k$  are written in (8.29) as a dilated window  $\theta$  modulated by cosine functions of frequency  $2^{-J}(k+1/2)\pi$ . In this case, the uniform



**FIGURE 8.9**

Admissible tree (*left*) and Heisenberg boxes (*right*) of a wavelet packet pseudo-local cosine basis.

wavelet packet basis (8.34) is a local cosine basis with windows of constant size. This result is not valid for wavelet packets constructed with different conjugate mirror filters. Nevertheless, the time and frequency resolution of uniform wavelet packet bases (8.34) remains constant, like that of local cosine bases constructed with windows of constant size. Figure 8.8(b) gives the decomposition coefficients of a signal in such a uniform wavelet packet basis.

### **Best Basis**

Applications of orthogonal bases often rely on their ability to efficiently approximate signals with only a few nonzero vectors. Choosing a wavelet packet basis that concentrates the signal energy over a few coefficients also reveals its time-frequency structures. Section 12.2.2 describes a fast algorithm that searches for a “best” basis that minimizes a Schur concave cost function among all wavelet packet bases. The wavelet packet basis of Figure 8.6 is calculated with this best basis search.

### **8.1.4 Wavelet Packet Filter Banks**

Wavelet packet coefficients are computed with a filter bank algorithm that generalizes the fast discrete wavelet transform. This algorithm is a straightforward iteration of the two-channel filter bank decomposition presented in Section 7.3.2. Therefore, it was used in signal processing by Crosnier, Esteban, and Galand [189] when they introduced the first family of perfect reconstruction filters. The algorithm is presented here from a wavelet packet point of view.

To any discrete signal input  $b[n]$  sampled at intervals  $N^{-1} = 2^L$ , as in (7.111) we associate  $f \in \mathbf{V}_L$  with decomposition coefficients  $a_L[n] = \langle f, \phi_{L,n} \rangle$  that satisfy

$$b[n] = N^{1/2} a_L[n] \approx f(N^{-1}n). \quad (8.35)$$

For any node  $(j, p)$  of the wavelet packet tree, we denote the wavelet packet coefficients

$$d_j^p[n] = \langle f(t), \psi_j^p(t - 2^j n) \rangle.$$

At the root of the tree  $d_L^0[n] = a_L[n]$  is computed from  $b[n]$  with (8.35).

### **Wavelet Packet Decomposition**

We denote  $\check{x}[n] = x[-n]$  and by  $\check{x}$  the signal obtained by inserting a zero between each sample of  $x$ . Theorem 8.5 generalizes the fast wavelet transform/Theorem 7.10.

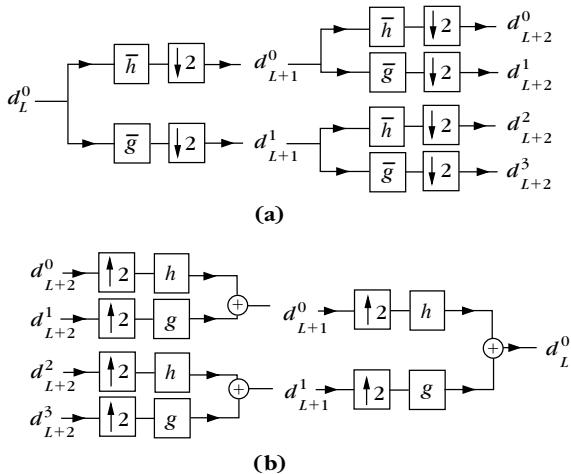
**Theorem 8.5.** At the decomposition

$$d_{j+1}^{2p}[k] = d_j^p \star \bar{h}[2k] \quad \text{and} \quad d_{j+1}^{2p+1}[k] = d_j^p \star \bar{g}[2k]. \quad (8.36)$$

At the reconstruction

$$d_j^p[k] = \check{d}_{j+1}^{2p} \star h[k] + \check{d}_{j+1}^{2p+1} \star g[k]. \quad (8.37)$$

The proof of these equations is identical to the proof of Theorem 7.10. The coefficients of wavelet packet children  $d_{j+1}^{2p}$  and  $d_{j+1}^{2p+1}$  are obtained by subsampling

**FIGURE 8.10**

- (a) Wavelet packet filter bank decomposition with successive filterings and subsamplings.  
 (b) Reconstruction by inserting zeros and filtering the outputs.

the convolutions of  $d_j^p$  with  $\bar{h}$  and  $\bar{g}$ . Iterating these equations along the branches of a wavelet packet tree computes all wavelet packet coefficients, as illustrated by Figure 8.10(a). From the wavelet packet coefficients at the leaves  $\{j_i, p_i\}_{1 \leq i \leq I}$  of an admissible subtree, we recover  $d_L^0$  at the top of the tree by computing (8.37) for each node inside the tree, as illustrated by Figure 8.10(b). ■

### **Finite Signals**

If  $a_L$  is a finite signal of size  $2^{-L} = N$ , we are facing the same border convolution problems as in a fast discrete wavelet transform. One approach explained in Section 7.5.1 is to periodize the wavelet packet basis. The convolutions (8.36) are then replaced by circular convolutions. To avoid introducing sharp transitions with the periodization, one can also use the border filters described in Section 7.5.3. In either case,  $d_j^p$  has  $2^{-j}$  samples. At any depth  $j - L$  of the tree, the wavelet packet signals  $\{d_j^p\}_{0 \leq p < 2^{j-L}}$  include a total of  $N$  coefficients. Since the maximum depth is  $\log_2 N$ , there are at most  $N \log_2 N$  coefficients in a full wavelet packet tree.

In a full wavelet packet tree of depth  $\log_2 N$ , all coefficients are computed by iterating (8.36) for  $L \leq j < 0$ . If  $h$  and  $g$  have  $K$  nonzero coefficients, this requires  $KN \log_2 N$  additions and multiplications. This is quite spectacular since there are more than  $2^{N/2}$  different wavelet packet bases included in this wavelet packet tree.

The computational complexity to recover  $a_L = d_L^0$  from the wavelet packet coefficients of an admissible tree increases with the number of inside nodes of the admissible tree. When the admissible tree is the full binary tree of depth  $\log_2 N$ , the number of operations is maximum and equal to  $KN \log_2 N$  multiplications and

additions. If the admissible subtree is a wavelet tree, we need fewer than  $2KN$  multiplications and additions.

### **Discrete Wavelet Packet Bases of $\ell^2(\mathbb{Z})$**

The signal decomposition in a conjugate mirror filter bank can also be interpreted as an expansion in discrete wavelet packet bases of  $\ell^2(\mathbb{Z})$ . This is proved with a result similar to Theorem 8.1.

**Theorem 8.6.** Let  $\{\theta_j[m - 2^{j-L}n]\}_{n \in \mathbb{Z}}$  be an orthonormal basis of a space  $\mathbf{U}_j$  with  $j - L \in \mathbb{N}$ . Define

$$\theta_{j+1}^0[m] = \sum_{n=-\infty}^{+\infty} h[n] \theta_j[m - 2^{j-L}n], \quad \theta_{j+1}^1[m] = \sum_{n=-\infty}^{+\infty} g[n] \theta_j[m - 2^{j-L}n]. \quad (8.38)$$

The family

$$\left\{ \theta_{j+1}^0[m - 2^{j+1-L}n], \theta_{j+1}^1[m - 2^{j+1-L}n] \right\}_{n \in \mathbb{Z}}$$

is an orthonormal basis of  $\mathbf{U}_j$ .

The proof is similar to the proof of Theorem 8.1. As in the continuous-time case, we derive from this theorem a binary tree of discrete wavelet packets. At the root of the discrete wavelet packet tree is the space  $\mathbf{W}_L^0 = \ell^2(\mathbb{Z})$  of discrete signals obtained with a sampling interval  $N^{-1} = 2^L$ . It admits a canonical basis of Diracs  $\{\psi_L^0[m - n] = \delta[m - n]\}_{n \in \mathbb{Z}}$ . The signal  $a_L[m]$  is specified by its sample values in this basis. One can verify that the convolutions and subsamplings (8.36) compute

$$d_j^p[n] = \langle a_L[m], \psi_j^p[m - 2^{j-L}n] \rangle,$$

where  $\{\psi_j^p[m - 2^{j-L}n]\}_{n \in \mathbb{Z}}$  is an orthogonal basis of a space  $\mathbf{W}_j^p$ . These discrete wavelet packets are recursively defined for any  $j \geq L$  and  $0 \leq p < 2^{j-L}$  by

$$\psi_{j+1}^{2p}[m] = \sum_{n=-\infty}^{+\infty} h[n] \psi_j^p[m - 2^{j-L}n], \quad \psi_{j+1}^{2p+1}[m] = \sum_{n=-\infty}^{+\infty} g[n] \psi_j^p[m - 2^{j-L}n]. \quad (8.39)$$

## **8.2 IMAGE WAVELET PACKETS**

### **8.2.1 Wavelet Packet Quad-Tree**

We construct wavelet packet bases of  $L^2(\mathbb{R}^2)$  with separable products of wavelet packets  $\psi_j^p(x_1 - 2^j n_1) \psi_j^q(x_2 - 2^j n_2)$  having the same scale along  $x_1$  and  $x_2$ . These separable wavelet packet bases are associated to quad-trees, and divide the two-dimensional Fourier plane  $(\omega_1, \omega_2)$  into square regions of varying sizes. Separable wavelet packet bases are extensions of separable wavelet bases.

If images are approximated at the scale  $2^L$ , to the root of the quad-tree we associate the approximation space  $\mathbf{V}_L^2 = \mathbf{V}_L \otimes \mathbf{V}_L \subset \mathbf{L}^2(\mathbb{R}^2)$  defined in Section 7.7.1. In Section 8.1.1 we explain how to decompose  $\mathbf{V}_L$  with a binary tree of wavelet packet spaces  $\mathbf{W}_j^p \subset \mathbf{V}_L$  that admit an orthogonal basis  $\{\psi_j^p(t - 2^j n)\}_{n \in \mathbb{Z}}$ . The two-dimensional wavelet packet quad-tree is composed of separable wavelet packet spaces. Each node of this quad-tree is labeled by a scale  $2^j$  and two integers  $0 \leq p < 2^{j-L}$  and  $0 \leq q < 2^{j-L}$ , and corresponds to a separable space

$$\mathbf{W}_j^{p,q} = \mathbf{W}_j^p \otimes \mathbf{W}_j^q. \quad (8.40)$$

The resulting separable wavelet packet for  $x = (x_1, x_2)$  is

$$\psi_j^{p,q}(x) = \psi_j^p(x_1) \psi_j^q(x_2).$$

Theorem 7.25 proves that an orthogonal basis of  $\mathbf{W}_j^{p,q}$  is obtained with a separable product of the wavelet packet bases of  $\mathbf{W}_j^p$  and  $\mathbf{W}_j^q$ , which can be written as

$$\left\{ \psi_j^{p,q}(x - 2^j n) \right\}_{n \in \mathbb{Z}^2}.$$

At the root  $\mathbf{W}_L^{0,0} = \mathbf{V}_L^2$ , the wavelet packet is a two-dimensional scaling function

$$\psi_L^{0,0}(x) = \phi_L^2(x) = \phi_L(x_1) \phi_L(x_2).$$

One-dimensional wavelet packet spaces satisfy

$$\mathbf{W}_j^p = \mathbf{W}_{j+1}^{2p} \oplus \mathbf{W}_{j+1}^{2p+1} \quad \text{and} \quad \mathbf{W}_j^q = \mathbf{W}_{j+1}^{2q} \oplus \mathbf{W}_{j+1}^{2q+1}.$$

Inserting these equations in (8.40) proves that  $\mathbf{W}_j^{p,q}$  is the direct sum of the four orthogonal subspaces

$$\mathbf{W}_j^{p,q} = \mathbf{W}_{j+1}^{2p,2q} \oplus \mathbf{W}_{j+1}^{2p+1,2q} \oplus \mathbf{W}_{j+1}^{2p,2q+1} \oplus \mathbf{W}_{j+1}^{2p+1,2q+1}. \quad (8.41)$$

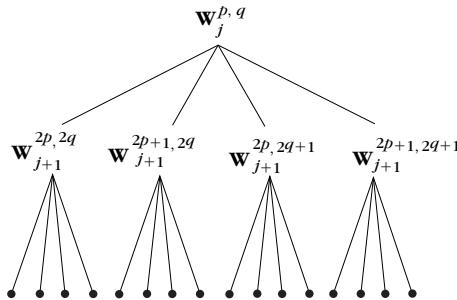
These subspaces are located at the four children nodes in the quad-tree, as shown by Figure 8.11. We call any quad-tree that has nodes with either zero or four children an *admissible quad-tree*. Let  $\{j_i, p_i, q_i\}_{0 \leq i \leq I}$  be the indices of the nodes at the leaves of an admissible quad-tree. Applying recursively the reconstruction sum (8.41) along the branches of this quad-tree gives an orthogonal decomposition of  $\mathbf{W}_L^{0,0}$ :

$$\mathbf{W}_L^{0,0} = \bigoplus_{i=1}^I \mathbf{W}_{j_i}^{p_i, q_i}.$$

The union of the corresponding wavelet packet bases

$$\left\{ \psi_{j_i}^{p_i, q_i}(x - 2^{j_i} n) \right\}_{(n_1, n_2) \in \mathbb{Z}^2, 1 \leq i \leq I}$$

is therefore an orthonormal basis of  $\mathbf{V}_L^2 = \mathbf{W}_L^{0,0}$ .

**FIGURE 8.11**

A wavelet packet quad-tree for images is constructed recursively by decomposing each separable space  $\mathbf{W}_j^{p,q}$  in four subspaces.

### Number of Wavelet Packet Bases

The number of different bases in a full wavelet packet quad-tree of depth  $J$  is equal to the number of admissible subtrees. Theorem 8.7 proves that there are more than  $2^{4^{J-1}}$  such bases.

**Theorem 8.7.** The number  $B_J$  of wavelet packet bases in a full wavelet packet quad-tree of depth  $J$  satisfies

$$2^{4^{J-1}} \leq B_J \leq 2^{\frac{49}{48}4^{J-1}}.$$

**Proof.** This result is proved with induction, as in the proof of Theorem 8.7. The reader can verify that  $B_J$  satisfies an induction relation similar to (8.16):

$$B_{J+1} = B_J^4 + 1. \quad (8.42)$$

Since  $B_0 = 1$ ,  $B_1 = 2$ , and  $B_{J+1} \geq B_J^4$ , we derive that  $B_J \geq 2^{4^{J-1}}$ . Moreover, for  $J \geq 1$

$$\log_2 B_{J+1} = 4 \log_2 B_J + \log_2(1 + B_J^{-4}) \leq 4 \log_2 B_J + \frac{1}{16} \leq 4^J + \frac{1}{16} \sum_{j=0}^{J-1} 4^j,$$

which implies that  $B_J \geq 2^{\frac{49}{48}4^{J-1}}$ . ■

For an image of  $N = N_1 N_2$  pixels, if  $N_1 = N_2$ , then the wavelet packet quad-tree has a depth at most  $\log_2 N_1 = \log_2 N^{1/2}$ . Thus, the number of wavelet packet bases satisfies

$$2^{\frac{N}{4}} \leq B_{\log_2 N} \leq 2^{\frac{49}{48}\frac{N}{4}}. \quad (8.43)$$

### Spatial and Frequency Localization

The spatial and frequency localization of two-dimensional wavelet packets is derived from the time-frequency analysis performed in Section 8.1.2. If the conjugate mirror

filter  $h$  has  $K$  nonzero coefficients, we proved that  $\psi_j^p$  has a support size of  $2^j(K - 1)$ , thus  $\psi_j^p(x_1)\psi_j^q(x_2)$  has a square support of width  $2^j(K - 1)$ .

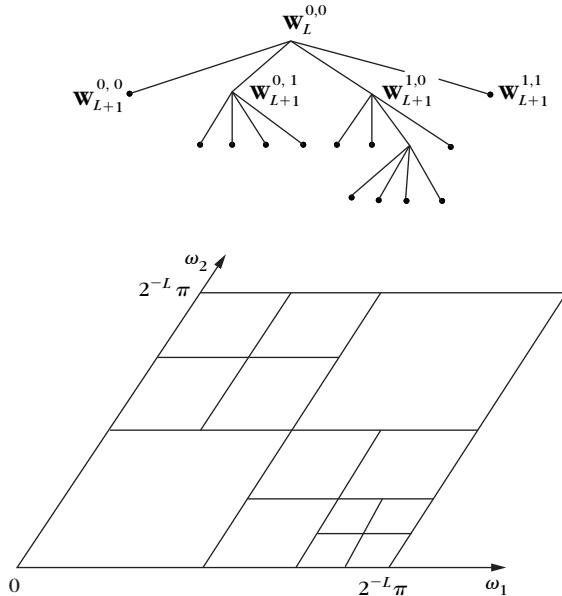
We showed that the Fourier transform of  $\psi_j^p$  has its energy mostly concentrated in

$$[-(k+1)2^{-j}\pi, -k2^{-j}\pi] \cup [k2^{-j}\pi, (k+1)2^{-j}\pi],$$

where  $k = G[p]$  is specified by Theorem 8.3. Therefore, the Fourier transform of a two-dimensional wavelet packet  $\psi_j^{p,q}$  has its energy mostly concentrated in

$$[k_12^{-j}\pi, (k_1+1)2^{-j}\pi] \times [k_22^{-j}\pi, (k_2+1)2^{-j}\pi], \quad (8.44)$$

with  $k_1 = G[p]$  and  $k_2 = G[q]$ , and in the three squares that are symmetric with respect to the two axes  $\omega_1 = 0$  and  $\omega_2 = 0$ . An admissible wavelet packet quad-tree decomposes the positive-frequency quadrant into squares of dyadic sizes, as illustrated in Figure 8.12. For example, the leaves of a full wavelet packet quad-tree of depth  $j - L$  define a wavelet packet basis that decomposes the positive-frequency quadrant into squares of constant width equal to  $2^{-j}\pi$ . This wavelet packet basis is similar to a two-dimensional local cosine basis with windows of constant size.



**FIGURE 8.12**

A wavelet packet quad-tree decomposes the positive-frequency quadrant into squares of progressively smaller sizes as we go down the tree.

### 8.2.2 Separable Filter Banks

The decomposition coefficients of an image in a separable wavelet packet basis are computed with a separable extension of the filter bank algorithm described in Section 8.1.4. Let  $b[n]$  be an input image with pixels at a distance  $2^L$ . We associate  $b[n]$  to a function  $f \in \mathbf{V}_L^2$  approximated at the scale  $2^L$ , with decomposition coefficients  $a_L[n] = \langle f(x), \phi_L^2(x - 2^L n) \rangle$  that are defined like in (7.232):

$$b[n] = 2^{-L} a_L[n] \approx f(2^L n).$$

The wavelet packet coefficients

$$d_j^{p,q}[n] = \langle f, \psi_j^{p,q}(x - 2^j n) \rangle$$

characterize the orthogonal projection of  $f$  in  $\mathbf{W}_j^{p,q}$ . At the root,  $d_L^{0,0} = a_L$ .

#### **Separable Filter Bank**

From the separability of wavelet packet bases and the one-dimensional convolution formula of Theorem (8.5), we derive that for any  $n = (n_1, n_2)$ ,

$$d_j^{2p,2q}[n] = d_j^{p,q} \star \bar{h}\bar{h}[2n], \quad d_j^{2p+1,2q}[n] = d_j^{p,q} \star \bar{g}\bar{h}[2n], \quad (8.45)$$

$$d_{j+1}^{2p,2q+1}[n] = d_j^{p,q} \star \bar{h}\bar{g}[2n], \quad d_{j+1}^{2p+1,2q+1}[n] = d_j^{p,q} \star \bar{g}\bar{g}[2n]. \quad (8.46)$$

Thus, the coefficients of a wavelet packet quad-tree are computed by iterating these equations along the quad-tree's branches. The calculations are performed with separable convolutions along the rows and columns of the image, as illustrated in Figure 8.13.

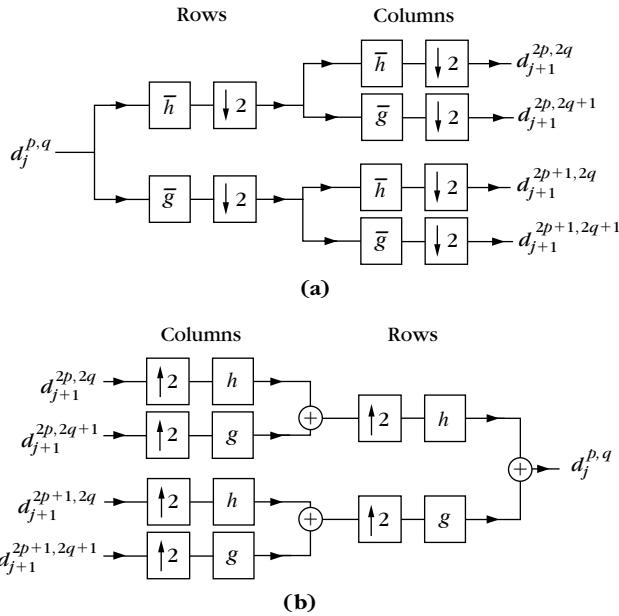
At the reconstruction,

$$\begin{aligned} d_j^{p,q}[n] &= \check{d}_{j+1}^{2p,2q} \star hh[n] + \check{d}_{j+1}^{2p+1,2q} \star gh[n] \\ &\quad + \check{d}_{j+1}^{2p,2q+1} \star hg[n] + \check{d}_{j+1}^{2p+1,2q+1} \star gg[n]. \end{aligned} \quad (8.47)$$

The image  $a_L = d_L^{0,0}$  is reconstructed from wavelet packet coefficients stored at the leaves of any admissible quad-tree by repeating the partial reconstruction (8.47) in the inside nodes of this quad-tree.

#### **Finite Images**

If the image  $a_L$  has  $N = 2^{-2L}$  pixels, the one-dimensional convolution border problems are solved with one of the two approaches described in Sections 7.5.1 and 7.5.3. Each wavelet packet image  $d_j^{p,q}$  includes  $2^{-2j}$  pixels. At depth  $j - L$ , there are  $N$  wavelet packet coefficients in  $\{d_j^{p,q}\}_{0 \leq p,q < 2^{j-L}}$ . Thus, a quad-tree of maximum depth  $\log_2 N^{1/2}$  includes  $N \log_2 N^{1/2}$  coefficients. If  $h$  and  $g$  have  $K$  nonzero coefficients, the one-dimensional convolutions that implement (8.45) and (8.46) require  $2K2^{-2j}$  multiplications and additions. Thus, all wavelet packet coefficients at depth  $j + 1 - L$  are computed from wavelet packet coefficients located at depth  $j - L$  with

**FIGURE 8.13**

**(a)** Wavelet packet decomposition implementing (8.45) and (8.46) with one-dimensional convolutions along the rows and columns of  $d_1^{p,q}$ . **(b)** Wavelet packet reconstruction implementing (8.47).

$2KN$  calculations. The  $N \log_2 N^{1/2}$  wavelet packet coefficients of a full tree of depth  $\log_2 N^{1/2}$  are therefore obtained with  $KN \log_2 N$  multiplications and additions. The numerical complexity of reconstructing  $a_L$  from a wavelet packet basis depends on the number of inside nodes of the corresponding quad-tree. The worst case is a reconstruction from the leaves of a full quad-tree of depth  $\log_2 N^{1/2}$ , which requires  $KN \log_2 N$  multiplications and additions.

### 8.3 BLOCK TRANSFORMS

Wavelet packet bases are designed by dividing the frequency axis in intervals of varying sizes. These bases are particularly well adapted to decomposing signals that have different behavior in different frequency intervals. If  $f$  has properties that vary in time, it is then more appropriate to decompose  $f$  in a *block basis* that segments the time axis in intervals with sizes that are adapted to the signal structures. Section 8.3.1 explains how to generate a block basis of  $L^2(\mathbb{R})$  from any basis of  $L^2[0, 1]$ . The block cosine bases described in Sections 8.3.2 and 8.3.3 are used for signal and image compression.

### 8.3.1 Block Bases

Block orthonormal bases are obtained by dividing the time axis in consecutive intervals  $[a_p, a_{p+1}]$  with

$$\lim_{p \rightarrow -\infty} a_p = -\infty \quad \text{and} \quad \lim_{p \rightarrow +\infty} a_p = +\infty.$$

The size  $l_p = a_{p+1} - a_p$  of each interval is arbitrary. Let  $g = \mathbf{1}_{[0,1]}$ . An interval is covered by the dilated rectangular window

$$g_p(t) = \mathbf{1}_{[a_p, a_{p+1}]}(t) = g\left(\frac{t - a_p}{l_p}\right). \quad (8.48)$$

Theorem 8.8 constructs a block orthogonal basis of  $\mathbf{L}^2(\mathbb{R})$  from a single orthonormal basis of  $\mathbf{L}^2[0, 1]$ .

**Theorem 8.8.** If  $\{e_k\}_{k \in \mathbb{Z}}$  is an orthonormal basis of  $\mathbf{L}^2[0, 1]$ , then

$$\left\{ g_{p,k}(t) = g_p(t) \frac{1}{\sqrt{l_p}} e_k\left(\frac{t - a_p}{l_p}\right) \right\}_{(p,k) \in \mathbb{Z}} \quad (8.49)$$

is a block orthonormal basis of  $\mathbf{L}^2(\mathbb{R})$ .

**Proof.** One can verify that the dilated and translated family

$$\left\{ \frac{1}{\sqrt{l_p}} e_k\left(\frac{t - a_p}{l_p}\right) \right\}_{k \in \mathbb{Z}} \quad (8.50)$$

is an orthonormal basis of  $\mathbf{L}^2[a_p, a_{p+1}]$ . If  $p \neq q$ , then  $\langle g_{p,k}, g_{q,k} \rangle = 0$  since their supports do not overlap. Thus, the family (8.49) is orthonormal. To expand a signal  $f$  in this family, it is decomposed as a sum of separate blocks

$$f(t) = \sum_{p=-\infty}^{+\infty} f(t) g_p(t),$$

and each block  $f(t)g_p(t)$  is decomposed in the basis (8.50). ■

#### Block Fourier Basis

A block basis is constructed with the Fourier basis of  $\mathbf{L}^2[0, 1]$ :

$$\left\{ e_k(t) = \exp(i2k\pi t) \right\}_{k \in \mathbb{Z}}.$$

The time support of each block Fourier vector  $g_{p,k}$  is  $[a_p, a_{p+1}]$  of size  $l_p$ . The Fourier transform of  $g = \mathbf{1}_{[0,1]}$  is

$$\hat{g}(\omega) = \frac{\sin(\omega/2)}{\omega/2} \exp\left(\frac{i\omega}{2}\right)$$

and

$$\hat{g}_{p,k}(\omega) = \sqrt{l_p} \hat{g}(l_p\omega - 2k\pi) \exp\left(\frac{-i2\pi ka_p}{l_p}\right).$$

It is centered at  $2k\pi l_p^{-1}$  and has a slow asymptotic decay proportional to  $l_p^{-1}|\omega|^{-1}$ . Because of this poor frequency localization, even though a signal  $f$  is smooth, its decomposition in a block Fourier basis may include large high-frequency coefficients. This can also be interpreted as an effect of periodization.

### Discrete Block Bases

For all  $p \in \mathbb{Z}$ , we suppose that  $a_p \in \mathbb{Z}$ . Discrete block bases are built with discrete rectangular windows having supports on intervals  $[a_p, a_{p+1} - 1]$ :

$$g_p[n] = \mathbf{1}_{[a_p, a_{p+1}-1]}(n).$$

Since dilations are not defined in a discrete framework, we generally cannot derive bases of intervals of varying sizes from a single basis. Thus, Theorem 8.9 supposes that we can construct an orthonormal basis of  $\mathbb{C}^l$  for any  $l > 0$ . The proof is straightforward.

**Theorem 8.9.** Suppose that  $\{e_{k,l}\}_{0 \leq k < l}$  is an orthogonal basis of  $\mathbb{C}^l$  for any  $l > 0$ . The family

$$\left\{ g_{p,k}[n] = g_p[n] e_{k,l_p}[n - a_p] \right\}_{0 \leq k < l_p, p \in \mathbb{Z}} \quad (8.51)$$

is a block orthonormal basis of  $\ell^2(\mathbb{Z})$ .

A discrete block basis is constructed with discrete Fourier bases

$$\left\{ e_{k,l}[n] = \frac{1}{\sqrt{l}} \exp\left(\frac{i2\pi kn}{l}\right) \right\}_{0 \leq k < l}.$$

The resulting block Fourier vectors  $g_{p,k}$  have sharp transitions at the window border, and thus are not well localized in frequency. As in the continuous case, the decomposition of smooth signals  $f$  may produce large-amplitude, high-frequency coefficients because of border effects.

### Block Bases of Images

General block bases of images are constructed by partitioning the plane  $\mathbb{R}^2$  into rectangles  $\{[a_p, b_p] \times [c_p, d_p]\}_{p \in \mathbb{Z}}$  of arbitrary length  $l_p = b_p - a_p$  and width  $w_p = d_p - c_p$ . Let  $\{e_k\}_{k \in \mathbb{Z}}$  be an orthonormal basis of  $\mathbf{L}^2[0, 1]$  and  $g = \mathbf{1}_{[0, 1]}$ . We denote

$$g_{p,k,j}(x, y) = g\left(\frac{x - a_p}{l_p}\right) g\left(\frac{y - c_p}{w_p}\right) \frac{1}{\sqrt{l_p w_p}} e_k\left(\frac{x - a_p}{l_p}\right) e_j\left(\frac{y - c_p}{w_p}\right).$$

The family  $\{g_{p,k,j}\}_{(k,j) \in \mathbb{Z}^2}$  is an orthonormal basis of  $\mathbf{L}^2([a_p, b_p] \times [c_p, d_p])$ , and thus  $\{g_{p,k,j}\}_{(p,k,j) \in \mathbb{Z}^3}$  is an orthonormal basis of  $\mathbf{L}^2(\mathbb{R}^2)$ .

For discrete images, we define discrete windows that cover each rectangle

$$g_p = \mathbf{1}_{[a_p, b_p - 1] \times [c_p, d_p - 1]}.$$

If  $\{e_{k,l}\}_{0 \leq k < l}$  is an orthogonal basis of  $\mathbb{C}^l$  for any  $l > 0$ , then

$$\left\{ g_{p,k,j}[n_1, n_2] = g_p[n_1, n_2] e_{k,l_p}[n_1 - a_p] e_{j,w_p}[n_2 - c_p] \right\}_{(k,j,p) \in \mathbb{Z}^3}$$

is a block basis of  $\ell^2(\mathbb{Z}^2)$ .

### 8.3.2 Cosine Bases

If  $f \in L^2[0, 1]$  and  $f(0) \neq f(1)$ , even though  $f$  might be a smooth function, the Fourier coefficients

$$\langle f(u), e^{i2k\pi u} \rangle = \int_0^1 f(u) e^{-i2k\pi u} du$$

have a relatively large amplitude at high frequencies  $2k\pi$ . Indeed, the Fourier series expansion

$$f(t) = \sum_{k=-\infty}^{+\infty} \langle f(u), e^{i2k\pi u} \rangle e^{i2k\pi t}$$

is a function of period 1, equal to  $f$  over  $[0, 1]$ , and that is therefore discontinuous if  $f(0) \neq f(1)$ . This shows that the restriction of a smooth function to an interval generates large Fourier coefficients. As a consequence, block Fourier bases are rarely used. A cosine I basis reduces this border effect by restoring a periodic extension  $\tilde{f}$  of  $f$ , which is continuous if  $f$  is continuous. Thus, high-frequency cosine I coefficients have a smaller amplitude than Fourier coefficients.

#### Cosine I Basis

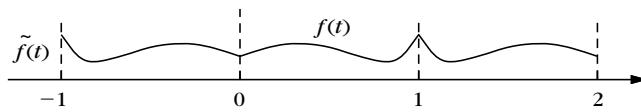
We define  $\tilde{f}$  to be the function of period 2 that is symmetric at about 0 and equal to  $f$  over  $[0, 1]$ :

$$\tilde{f}(t) = \begin{cases} f(t) & \text{for } t \in [0, 1] \\ f(-t) & \text{for } t \in (-1, 0). \end{cases} \quad (8.52)$$

If  $f$  is continuous over  $[0, 1]$ , then  $\tilde{f}$  is continuous over  $\mathbb{R}$ , as shown in Figure 8.14. However, if  $f$  has a nonzero right derivative at 0 or left derivative at 1, then  $\tilde{f}$  is nondifferentiable at integer points.

The Fourier expansion of  $\tilde{f}$  over  $[0, 2]$  can be written as a sum of sine and cosine terms:

$$\tilde{f}(t) = \sum_{k=0}^{+\infty} a[k] \cos\left(\frac{2\pi kt}{2}\right) + \sum_{k=1}^{+\infty} b[k] \sin\left(\frac{2\pi kt}{2}\right).$$

**FIGURE 8.14**

The function  $\tilde{f}(t)$  is an extension of  $f(t)$ ; it is symmetric at about 0 and of period 2.

The sine coefficients  $b[k]$  are zero because  $\tilde{f}$  is even. Since  $f(t) = \tilde{f}(t)$  over  $[0, 1]$ , this proves that any  $f \in L^2[0, 1]$  can be written as a linear combination of the cosines  $\{\cos(k\pi t)\}_{k \in \mathbb{N}}$ . One can verify that this family is orthogonal over  $[0, 1]$ . Therefore, it is an orthogonal basis of  $L^2[0, 1]$ , as stated by Theorem 8.10.

**Theorem 8.10:** *Cosine I.* The family

$$\left\{ \lambda_k \sqrt{2} \cos(\pi k t) \right\}_{k \in \mathbb{N}} \quad \text{with} \quad \lambda_k = \begin{cases} 2^{-1/2} & \text{if } k = 0 \\ 1 & \text{if } k \neq 0 \end{cases}$$

is an orthonormal basis of  $L^2[0, 1]$ . ■

### Block Cosine Basis

Let us divide the real line with square windows  $g_p = \mathbf{1}_{[a_p, a_{p+1}]}$ . Theorem 8.8 proves that

$$\left\{ g_{p,k}(t) = g_p(t) \sqrt{\frac{2}{l_p}} \lambda_k \cos\left(\pi k \frac{t - a_p}{l_p}\right) \right\}_{k \in \mathbb{N}, p \in \mathbb{Z}}$$

is a block basis of  $L^2(\mathbb{R})$ . The decomposition coefficients of a smooth function have a faster decay at high frequencies in a block cosine basis than in a block Fourier basis, because cosine bases correspond to a smoother signal extension beyond the intervals  $[a_p, a_{p+1}]$ .

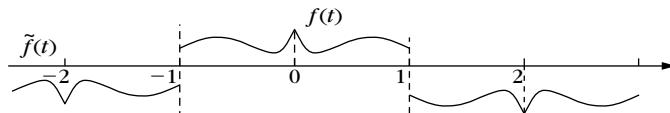
### Cosine IV Basis

Other cosine bases are constructed from the Fourier series, with different extensions of  $f$  beyond  $[0, 1]$ . The cosine IV basis appears in fast numerical computations of cosine I coefficients. It is also used to construct local cosine bases with smooth windows in Section 8.4.2.

Any  $f \in L^2[0, 1]$  is extended into a function  $\tilde{f}$  of period 4, which is symmetric about 0 and antisymmetric about 1 and  $-1$ :

$$\tilde{f}(t) = \begin{cases} f(t) & \text{if } t \in [0, 1] \\ f(-t) & \text{if } t \in (-1, 0) \\ -f(2-t) & \text{if } t \in [1, 2] \\ -f(2+t) & \text{if } t \in [-1, -2] \end{cases}$$

If  $f(1) \neq 0$ , the antisymmetry at 1 creates a function  $\tilde{f}$  that is discontinuous at  $f(2n+1)$  for any  $n \in \mathbb{Z}$ , as shown in Figure 8.15. Therefore, this extension is less regular than the cosine I extension (8.52).

**FIGURE 8.15**

A cosine IV extends  $f(t)$  into a signal  $\tilde{f}(t)$  of period 4, which is symmetric with respect to 0 and antisymmetric with respect to 1.

Since  $\tilde{f}$  is 4 periodic, it can be decomposed as a sum of sines and cosines of period 4:

$$\tilde{f}(t) = \sum_{k=0}^{+\infty} a[k] \cos\left(\frac{2\pi kt}{4}\right) + \sum_{k=1}^{+\infty} b[k] \sin\left(\frac{2\pi kt}{4}\right).$$

The symmetry about 0 implies that

$$b[k] = \frac{1}{2} \int_{-2}^2 \tilde{f}(t) \sin\left(\frac{2\pi kt}{4}\right) dt = 0.$$

For even frequencies, the antisymmetry about 1 and  $-1$  yields

$$a[2k] = \frac{1}{2} \int_{-2}^2 \tilde{f}(t) \cos\left(\frac{2\pi(2k)t}{4}\right) dt = 0.$$

Thus, the only nonzero components are cosines of odd frequencies:

$$\tilde{f}(t) = \sum_{k=0}^{+\infty} a[2k+1] \cos\left(\frac{(2k+1)2\pi t}{4}\right). \quad (8.53)$$

Since  $f(t) = \tilde{f}(t)$  over  $[0, 1]$ , this proves that any  $f \in L^2[0, 1]$  is decomposed as a sum of such cosine functions. One can verify that the restriction of these cosine functions to  $[0, 1]$  is orthogonal in  $L^2[0, 1]$ , which implies Theorem 8.11.

**Theorem 8.11: Cosine IV.** The family

$$\left\{ \sqrt{2} \cos\left[\left(k + \frac{1}{2}\right)\pi t\right] \right\}_{k \in \mathbb{N}}$$

is an orthonormal basis of  $L^2[0, 1]$ . ■

The cosine transform IV is not used in block transforms because it has the same drawbacks as a block Fourier basis. Block cosine IV coefficients of a smooth  $f$  have a slow decay at high frequencies, because such a decomposition corresponds to a discontinuous extension of  $f$  beyond each block. Section 8.4.2 explains how to avoid this issue with smooth windows.

### 8.3.3 Discrete Cosine Bases

Discrete cosine bases are derived from the discrete Fourier basis with the same approach as in the continuous time case. To simplify notations, the sampling distance is normalized to 1. If the sampling distance was originally  $N^{-1}$ , then the frequency indexes that appear in this section must be multiplied by  $N$ .

#### *Discrete Cosine I*

A signal  $f[n]$  defined for  $0 \leq n < N$  is extended by symmetry with respect to  $-1/2$  into a signal  $\tilde{f}[n]$  of size  $2N$ :

$$\tilde{f}[n] = \begin{cases} f[n] & \text{for } 0 \leq n < N \\ f[-n-1] & \text{for } -N \leq n \leq -1. \end{cases} \quad (8.54)$$

The  $2N$  discrete Fourier transform of  $\tilde{f}$  can be written as a sum of sine and cosine terms:

$$\tilde{f}[n] = \sum_{k=0}^{N-1} a[k] \cos \left[ \frac{k\pi}{N} \left( n + \frac{1}{2} \right) \right] + \sum_{k=0}^{N-1} b[k] \sin \left[ \frac{k\pi}{N} \left( n + \frac{1}{2} \right) \right].$$

Since  $\tilde{f}$  is symmetric about  $-1/2$ , then necessarily  $b[k] = 0$  for  $0 \leq k < N$ . Moreover,  $f[n] = \tilde{f}[n]$  for  $0 \leq n < N$ , so any signal  $f \in \mathbb{C}^N$  can be written as a sum of these cosine functions. The reader can also verify that these discrete cosine signals are orthogonal in  $\mathbb{C}^N$ . Thus, we obtain Theorem 8.12.

**Theorem 8.12: Cosine I.** The family

$$\left\{ \lambda_k \sqrt{\frac{2}{N}} \cos \left[ \frac{k\pi}{N} \left( n + \frac{1}{2} \right) \right] \right\}_{0 \leq k < N} \quad \text{with} \quad \lambda_k = \begin{cases} 2^{-1/2} & \text{if } k = 0 \\ 1 & \text{otherwise} \end{cases}$$

is an orthonormal basis of  $\mathbb{C}^N$ . ■

This theorem proves that any  $f \in \mathbb{C}^N$  can be decomposed into

$$f[n] = \frac{2}{N} \sum_{k=0}^{N-1} \hat{f}_I[k] \lambda_k \cos \left[ \frac{k\pi}{N} \left( n + \frac{1}{2} \right) \right], \quad (8.55)$$

where

$$\hat{f}_I[k] = \left\langle f[n], \lambda_k \cos \left[ \frac{k\pi}{N} \left( n + \frac{1}{2} \right) \right] \right\rangle = \lambda_k \sum_{n=0}^{N-1} f[n] \cos \left[ \frac{k\pi}{N} \left( n + \frac{1}{2} \right) \right] \quad (8.56)$$

is the discrete cosine transform I (DCT-I) of  $f$ . The next section describes a fast discrete cosine transform that computes  $\hat{f}_I$  with  $O(N \log_2 N)$  operations.

### Discrete Block Cosine Transform

Let us divide the integer set  $\mathbb{Z}$  with discrete windows  $g_p[n] = \mathbf{1}_{[a_p, a_p+1]}(n)$  with  $a_p \in \mathbb{Z}$ . Theorem 8.9 proves that the corresponding block basis

$$\left\{ g_{p,k}[n] = g_p[n] \lambda_k \sqrt{\frac{2}{l_p}} \cos \left[ \frac{k\pi}{l_p} \left( n + \frac{1}{2} - a_p \right) \right] \right\}_{0 \leq k < N, p \in \mathbb{Z}}$$

is an orthonormal basis of  $\ell^2(\mathbb{Z})$ . Over each block of size  $l_p = a_{p+1} - a_p$ , the fast DCT-I algorithm computes all coefficients with  $O(l_p \log_2 l_p)$  operations. Section 10.5.1 describes the JPEG image compression standard, which decomposes images in a separable block cosine basis. A block cosine basis is used as opposed to a block Fourier basis, because it yields smaller-amplitude, high-frequency coefficients, which improves the coding performance.

### Discrete Cosine IV

To construct a discrete cosine IV basis, a signal  $f$  of  $N$  samples is extended into a signal  $\tilde{f}$  of period  $4N$ , which is symmetric with respect to  $-1/2$  and antisymmetric with respect to  $N-1/2$  and  $-N+1/2$ . As in (8.53), the decomposition of  $\tilde{f}$  over a family of sines and cosines of period  $4N$  has no sine terms and no cosine terms of even frequency. Since  $\tilde{f}[N] = f[n]$  for  $0 \leq n < N$ , we derive that  $f$  can also be written as a linear expansion of these odd-frequency cosines, which are orthogonal in  $\mathbb{C}^N$ . Thus, we obtain Theorem 8.13.

**Theorem 8.13:** *Cosine IV.* The family

$$\left\{ \sqrt{\frac{2}{N}} \cos \left[ \frac{\pi}{N} \left( k + \frac{1}{2} \right) \left( n + \frac{1}{2} \right) \right] \right\}_{0 \leq k < N}$$

is an orthonormal basis of  $\mathbb{C}^N$ . ■

This theorem proves that any  $f \in \mathbb{C}^N$  can be decomposed into

$$f[n] = \frac{2}{N} \sum_{k=0}^{N-1} \hat{f}_{IV}[k] \cos \left[ \frac{\pi}{N} \left( k + \frac{1}{2} \right) \left( n + \frac{1}{2} \right) \right], \quad (8.57)$$

where

$$\hat{f}_{IV}[k] = \sum_{n=0}^{N-1} f[n] \cos \left[ \frac{\pi}{N} \left( k + \frac{1}{2} \right) \left( n + \frac{1}{2} \right) \right] \quad (8.58)$$

is the discrete cosine transform IV (DCT-IV) of  $f$ .

#### 8.3.4 Fast Discrete Cosine Transforms

The DCT-IV of a signal of size  $N$  is related to the discrete Fourier transform (DFT) of a complex signal of size  $N/2$  with a formula introduced by Duhamel, Mahieux, and Petit [40, 236]. By computing this DFT with the fast Fourier transform (FFT)

described in Section 3.3.3, we need  $O(N \log_2 N)$  operations to compute the DCT-IV. The DCT-I coefficients are then calculated through an induction relation with the DCT-IV, due to Wang [479].

### Fast DCT-IV

To clarify the relation between a DCT-IV and a DFT, we split  $f[n]$  in two half-size signals of odd and even indices:

$$\begin{aligned} b[n] &= f[2n], \\ c[n] &= f[N - 1 - 2n]. \end{aligned}$$

The DCT-IV (8.58) is rewritten as

$$\begin{aligned} \hat{f}_{IV}[k] &= \sum_{n=0}^{N/2-1} b[n] \cos \left[ \left( 2n + \frac{1}{2} \right) \left( k + \frac{1}{2} \right) \frac{\pi}{N} \right] + \\ &\quad \sum_{n=0}^{N/2-1} c[n] \cos \left[ \left( N - 1 - 2n + \frac{1}{2} \right) \left( k + \frac{1}{2} \right) \frac{\pi}{N} \right] \\ &= \sum_{n=0}^{N/2-1} b[n] \cos \left[ \left( n + \frac{1}{4} \right) \left( k + \frac{1}{2} \right) \frac{2\pi}{N} \right] + \\ &\quad (-1)^k \sum_{n=0}^{N/2-1} c[n] \sin \left[ \left( n + \frac{1}{4} \right) \left( k + \frac{1}{2} \right) \frac{2\pi}{N} \right]. \end{aligned}$$

Thus, the even-frequency indices can be expressed as a real part,

$$\begin{aligned} \hat{f}_{IV}[2k] &= \\ &\text{Re} \left\{ \exp \left[ \frac{-i\pi k}{N} \right] \sum_{n=0}^{N/2-1} (b[n] + ic[n]) \exp \left[ -i(n + \frac{1}{4}) \frac{\pi}{N} \right] \exp \left[ \frac{-i2\pi kn}{N/2} \right] \right\}, \end{aligned} \tag{8.59}$$

whereas the odd coefficients correspond to an imaginary part,

$$\begin{aligned} \hat{f}_{IV}[N - 2k - 1] &= \\ &- \text{Im} \left\{ \exp \left[ \frac{-i\pi k}{N} \right] \sum_{n=0}^{N/2-1} (b[n] + ic[n]) \exp \left[ -i(n + \frac{1}{4}) \frac{\pi}{N} \right] \exp \left[ \frac{-i2\pi kn}{N/2} \right] \right\}. \end{aligned} \tag{8.60}$$

For  $0 \leq n < N/2$ , we denote

$$g[n] = (b[n] + i c[n]) \exp \left[ -i \left( n + \frac{1}{4} \right) \frac{\pi}{N} \right].$$

The DFT  $\hat{g}[k]$  of  $g[n]$  is computed with an FFT of size  $N/2$ . Equations (8.59) and (8.60) prove that

$$\hat{f}_{IV}[2k] = \text{Re} \left\{ \exp \left[ \frac{-i\pi k}{N} \right] \hat{g}[k] \right\},$$

and

$$\hat{f}_{IV}[N - 2k - 1] = -\text{Im} \left\{ \exp \left[ \frac{-i\pi k}{N} \right] \hat{g}[k] \right\}.$$

The DCT-IV coefficients  $\hat{f}_{IV}[k]$  are obtained with one FFT of size  $N/2$  plus  $O(N)$  operations, which makes a total of  $O(N \log_2 N)$  operations. To normalize the DCT-IV, the resulting coefficients must be multiplied by  $\sqrt{2/N}$ . An efficient implementation of the DCT-IV with a split-radix FFT requires [40]

$$\mu_{DCT-IV}(N) = \frac{N}{2} \log_2 N + N \quad (8.61)$$

real multiplications and

$$\alpha_{DCT-IV}(N) = \frac{3N}{2} \log_2 N \quad (8.62)$$

additions.

The inverse DCT-IV of  $\hat{f}_{IV}$  is given by (8.57). Up to the proportionality constant  $2/N$ , this sum is the same as (8.58), where  $\hat{f}_{IV}$  and  $f$  are interchanged. This proves that the inverse DCT-IV is computed with the same fast algorithm as the forward DCT-IV.

### **Fast DCT-I**

A DCT-I is calculated with an induction relation that involves the DCT-IV. Regrouping the terms  $f[n]$  and  $f[N - 1 - n]$  of a DCT-I (8.56) yields

$$\hat{f}_I[2k] = \lambda_k \sum_{n=0}^{N/2-1} (f[n] + f[N - 1 - n]) \cos \left[ \frac{\pi k}{N/2} \left( n + \frac{1}{2} \right) \right], \quad (8.63)$$

$$\hat{f}_I[2k + 1] = \sum_{n=0}^{N/2-1} (f[n] - f[N - 1 - n]) \cos \left[ \frac{\pi (k + 1/2)}{N/2} \left( n + \frac{1}{2} \right) \right]. \quad (8.64)$$

The even index coefficients of the DCT-I are equal to the DCT-I of the signal  $f[n] + f[N - 1 - n]$  of length  $N/2$ . The odd coefficients are equal to the DCT-IV of the signal  $f[n] - f[N - 1 - n]$  of length  $N/2$ . Thus, the number of multiplications of a DCT-I is related to the number of multiplications of a DCT-IV by the induction relation

$$\mu_{DCT-I}(N) = \mu_{DCT-I}(N/2) + \mu_{DCT-IV}(N/2), \quad (8.65)$$

while the number of additions is

$$\alpha_{DCT-I}(N) = \alpha_{DCT-I}(N/2) + \alpha_{DCT-IV}(N/2) + N. \quad (8.66)$$

Since the number of multiplications and additions of a DCT-IV is  $O(N \log_2 N)$ , this induction relation proves that the number of multiplications and additions of this algorithm is also  $O(N \log_2 N)$ .

If the DCT-IV is implemented with a split-radix FFT, by inserting (8.61) and (8.62) in the recurrence equations (8.65) and (8.66), we derive that the number of multiplications and additions to compute a DCT-I of size  $N$  is

$$\mu_{DCT-I}(N) = \frac{N}{2} \log_2 N + 1, \quad (8.67)$$

and

$$\alpha_{DCT-I}(N) = \frac{3N}{2} \log_2 N - N + 1. \quad (8.68)$$

The inverse DCT-I is computed with a similar recursive algorithm. Applied to  $\hat{f}_I$ , it is obtained by computing the inverse DCT-IV of the odd index coefficients  $\hat{f}_I[2k+1]$  with (8.64) and an inverse DCT-I of size  $N/2$  applied to the even coefficients  $\hat{f}_I[2k]$  with (8.63). From the values  $f[n] + f[N-1-n]$  and  $f[n] - f[N-1-n]$ , we recover  $f[n]$  and  $f[N-1-n]$ . The inverse DCT-IV is identical to the forward DCT-IV up to a multiplicative constant. Thus, the inverse DCT-I requires the same number of operations as the forward DCT-I.

## 8.4 LAPPED ORTHOGONAL TRANSFORMS

Cosine and Fourier block bases are computed with discontinuous rectangular windows that divide the real line in disjoint intervals. Multiplying a signal with a rectangular window creates discontinuities that produce large-amplitude coefficients at high frequencies. To avoid these discontinuity artifacts, it is necessary to use smooth windows.

The Balian-Low theorem (5.20) proves that for any  $u_0$  and  $\xi_0$ , there exists no differentiable window  $g$  of compact support such that

$$\left\{ g(t - nu_0) \exp(ik\xi_0 t) \right\}_{(n,k) \in \mathbb{Z}^2}$$

is an orthonormal basis of  $L^2(\mathbb{R})$ . This negative result discouraged any research in this direction, until Malvar discovered in discrete signal processing that one could create orthogonal bases with smooth windows modulated by a cosine IV basis [368, 369]. This result was independently rediscovered for continuous-time functions by Coifman and Meyer [181] with a different approach that we shall follow here. The roots of these new orthogonal bases are lapped projectors, which split signals in orthogonal components with overlapping supports [43]. Section 8.4.1 introduces these lapped projectors, the construction of continuous time and discrete lapped orthogonal bases are explained in the following sections. The particular case of local cosine bases is studied in more detail.

### 8.4.1 Lapped Projectors

Block transforms compute the restriction of  $f$  to consecutive intervals  $[a_p, a_{p+1}]$  and decompose this restriction in an orthogonal basis of  $[a_p, a_{p+1}]$ . Formally, the

restriction of  $f$  to  $[a_p, a_{p+1}]$  is an orthogonal projection on space  $\mathbf{W}^p$  of functions with a support included in  $[a_p, a_{p+1}]$ . To avoid the discontinuities introduced by this projection, we introduce new orthogonal projectors that perform a smooth deformation of  $f$ .

### Projectors on Half Lines

Let us first construct two orthogonal projectors that decompose any  $f \in \mathbf{L}^2(\mathbb{R})$  in two orthogonal components  $P^+f$  and  $P^-f$  with supports that are, respectively,  $[-1, +\infty)$  and  $(-\infty, 1]$ . For this purpose we consider a monotone increasing profile function  $\beta$  such that

$$\beta(t) = \begin{cases} 0 & \text{if } t < -1 \\ 1 & \text{if } t > 1 \end{cases} \quad (8.69)$$

and

$$\forall t \in [-1, 1], \quad \beta^2(t) + \beta^2(-t) = 1. \quad (8.70)$$

A naive definition

$$P^+f(t) = \beta^2(t)f(t) \quad \text{and} \quad P^-f(t) = \beta^2(-t)f(t)$$

satisfies the support conditions but does not define orthogonal functions. Since the supports of  $P^+f(t)$  and  $P^-f(t)$  overlap only on  $[-1, 1]$ , the orthogonality is obtained by creating functions having a different symmetry with respect to 0 on  $[-1, 1]$ :

$$P^+f(t) = \beta(t)[\beta(t)f(t) + \beta(-t)f(-t)] = \beta(t)p(t), \quad (8.71)$$

and

$$P^-f(t) = \beta(-t)[\beta(-t)f(t) - \beta(t)f(-t)] = \beta(-t)q(t). \quad (8.72)$$

The functions  $p(t)$  and  $q(t)$  are, respectively, even and odd, and since  $\beta(t)\beta(-t)$  is even, it follows that

$$\langle P^+f, P^-f \rangle = \int_{-1}^1 \beta(t)\beta(-t)p(t)q^*(t)dt = 0. \quad (8.73)$$

Clearly,  $P^+f$  belongs to space  $\mathbf{W}^+$  of functions  $f \in \mathbf{L}^2(\mathbb{R})$  such that there exists  $p(t) = p(-t)$  with

$$f(t) = \begin{cases} 0 & \text{if } t < -1 \\ \beta(t)p(t) & \text{if } t \in [-1, 1]. \end{cases}$$

Similarly,  $P^-f$  is in space  $\mathbf{W}^-$  composed of  $f \in \mathbf{L}^2(\mathbb{R})$  such that there exists  $q(t) = -q(-t)$  with

$$f(t) = \begin{cases} 0 & \text{if } t > 1 \\ \beta(-t)q(t) & \text{if } t \in [-1, 1]. \end{cases}$$

Functions in  $\mathbf{W}^+$  and  $\mathbf{W}^-$  may have an arbitrary behavior on  $[1, +\infty)$  and  $(-\infty, -1]$ , respectively. Theorem 8.14 characterizes  $P^+$  and  $P^-$ . We denote the identity operator by  $\text{Id}$ .

**Theorem 8.14:** *Coifman, Meyer.* Operators  $P^+$  and  $P^-$  are orthogonal projectors on  $\mathbf{W}^+$  and  $\mathbf{W}^-$ , respectively. Spaces  $\mathbf{W}^+$  and  $\mathbf{W}^-$  are orthogonal and

$$P^+ + P^- = \text{Id}. \quad (8.74)$$

**Proof.** To verify that  $P^+$  is a projector we show that any  $f \in \mathbf{W}^+$  satisfies  $P^+f = f$ . If  $t < -1$ , then  $P^+f(t) = f(t) = 0$ , and if  $t > 1$ , then  $P^+f(t) = f(t) = 1$ . If  $t \in [-1, 1]$ , then  $f(t) = \beta(t)p_0(t)$  and inserting (8.71) yields

$$P^+f(t) = \beta(t)[\beta^2(t)p_0(t) + \beta^2(-t)p_0(-t)] = \beta(t)p_0(t),$$

because  $p_0(t)$  is even and  $\beta(t)$  satisfies (8.70). Projector  $P^+$  is proved to be orthogonal by showing that it is self-adjoint:

$$\begin{aligned} \langle P^+f, g \rangle &= \int_{-1}^1 \beta^2(t)f(t)g^*(t)dt + \int_{-1}^1 \beta(t)\beta(-t)f(-t)g^*(t)dt + \\ &\quad \int_1^{+\infty} f(t)g^*(t)dt. \end{aligned}$$

A change of variable  $t' = -t$  in the second integral verifies that this formula is symmetric in  $f$  and  $g$  and thus  $\langle P^+f, g \rangle = \langle f, P^+g \rangle$ . Identical derivations prove that  $P^-$  is an orthogonal projector on  $\mathbf{W}^-$ .

The orthogonality of  $\mathbf{W}^-$  and  $\mathbf{W}^+$  is proved in (8.73). To verify (8.74), for  $f \in \mathbf{L}^2(\mathbb{R})$  we compute

$$P^+f(t) + P^-f(t) = f(t)[\beta^2(t) + \beta^2(-t)] = f(t). \quad \blacksquare$$

These half-line projectors are generalized by decomposing signals in two orthogonal components with supports that are, respectively,  $[a - \eta, +\infty)$  and  $(-\infty, a + \eta]$ . For this purpose, we scale and translate the profile function  $\beta(\frac{t-a}{\eta})$  so that it increases from 0 to 1 on  $[a - \eta, a + \eta]$ , as illustrated in Figure 8.16. The symmetry with respect to 0, which transforms  $f(t)$  in  $f(-t)$ , becomes a symmetry with respect to  $a$ , which transforms  $f(t)$  in  $f(2a - t)$ . The resulting projectors are

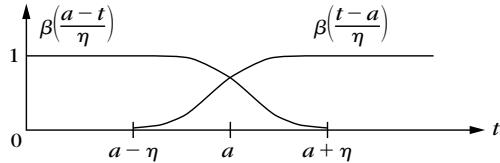
$$P_{a,\eta}^+f(t) = \beta\left(\frac{t-a}{\eta}\right) \left[ \beta\left(\frac{t-a}{\eta}\right)f(t) + \beta\left(\frac{a-t}{\eta}\right)f(2a-t) \right], \quad (8.75)$$

and

$$P_{a,\eta}^-f(t) = \beta\left(\frac{a-t}{\eta}\right) \left[ \beta\left(\frac{a-t}{\eta}\right)f(t) - \beta\left(\frac{t-a}{\eta}\right)f(2a-t) \right]. \quad (8.76)$$

A straightforward extension of Theorem 8.14 proves that  $P_{a,\eta}^+$  is an orthogonal projector on space  $\mathbf{W}_{a,\eta}^+$  of functions  $f \in \mathbf{L}^2(\mathbb{R})$  such that there exists  $p(t) = p(2a - t)$  with

$$f(t) = \begin{cases} 0 & \text{if } t < a - \eta \\ \beta(\eta^{-1}(t-a))p(t) & \text{if } t \in [a - \eta, a + \eta]. \end{cases} \quad (8.77)$$

**FIGURE 8.16**

A multiplication with  $\beta\left(\frac{t-a}{\eta}\right)$  and  $\beta\left(\frac{a-t}{\eta}\right)$  restricts the support of functions to  $[a-\eta, +\infty)$  and  $(-\infty, a+\eta]$ .

Similarly,  $P_{a,\eta}^-$  is an orthogonal projector on space  $\mathbf{W}_{a,\eta}^-$  composed of  $f \in \mathbf{L}^2(\mathbb{R})$  such that there exists  $q(t) = -q(2a-t)$  with

$$f(t) = \begin{cases} 0 & \text{if } t < -1 \\ \beta(\eta^{-1}(a-t)) q(t) & \text{if } t \in [a-\eta, a+\eta]. \end{cases} \quad (8.78)$$

Spaces  $\mathbf{W}_{a,\eta}^+$  and  $\mathbf{W}_{a,\eta}^-$  are orthogonal and

$$P_{a,\eta}^+ + P_{a,\eta}^- = \text{Id.} \quad (8.79)$$

### **Projectors on Intervals**

A lapped projector splits a signal in two orthogonal components that overlap on  $[a-\eta, a+\eta]$ . Repeating such projections at different locations performs a signal decomposition into orthogonal pieces with supports that overlap. Let us divide the time axis in overlapping intervals:

$$I_p = [a_p - \eta_p, a_{p+1} + \eta_{p+1}]$$

with

$$\lim_{p \rightarrow -\infty} a_p = -\infty \quad \text{and} \quad \lim_{p \rightarrow +\infty} a_p = +\infty. \quad (8.80)$$

To ensure that  $I_{p-1}$  and  $I_{p+1}$  do not intersect for any  $p \in \mathbb{Z}$ , we impose that

$$a_{p+1} - \eta_{p+1} \geq a_p + \eta_p,$$

and thus,

$$l_p = a_{p+1} - a_p \geq \eta_{p+1} + \eta_p. \quad (8.81)$$

The support of  $f$  is restricted to  $I_p$  by the operator

$$P_p = P_{a_p, \eta_p}^+ P_{a_{p+1}, \eta_{p+1}}^-. \quad (8.82)$$

Since  $P_{a_p, \eta_p}^+$  and  $P_{a_{p+1}, \eta_{p+1}}^-$  are orthogonal projections on  $\mathbf{W}_{a_p, \eta_p}^+$  and  $\mathbf{W}_{a_{p+1}, \eta_{p+1}}^-$ , it follows that  $P_p$  is an orthogonal projector on

$$\mathbf{W}^p = \mathbf{W}_{a_p, \eta_p}^+ \cap \mathbf{W}_{a_{p+1}, \eta_{p+1}}^-. \quad (8.83)$$

Let us divide  $I_p$  in two overlapping intervals  $O_p$  and  $O_{p+1}$ , and a central interval  $C_p$ :

$$I_p = [a_p - \eta_p, a_{p+1} + \eta_{p+1}] = O_p \cup C_p \cup O_{p+1} \quad (8.84)$$

with

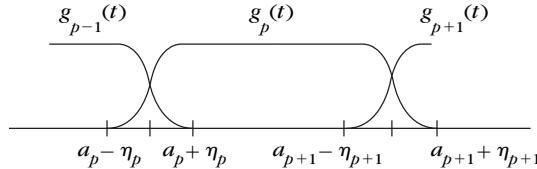
$$O_p = [a_p - \eta_p, a_p + \eta_p] \quad \text{and} \quad C_p = [a_p + \eta_p, a_{p+1} - \eta_{p+1}].$$

Space  $\mathbf{W}^p$  is characterized by introducing a window  $g_p$  support in  $I_p$ , and that has a raising profile on  $O_p$  and a decaying profile on  $O_{p+1}$ :

$$g_p(t) = \begin{cases} 0 & \text{if } t \notin I_p \\ \beta(\eta_p^{-1}(t - a_p)) & \text{if } t \in O_p \\ 1 & \text{if } t \in C_p \\ \beta(\eta_{p+1}^{-1}(a_{p+1} - t)) & \text{if } t \in O_{p+1}. \end{cases} \quad (8.85)$$

This window is illustrated in Figure 8.17. It follows from (8.77), (8.78), and (8.83) that  $\mathbf{W}^p$  is the space of functions  $f \in \mathbf{L}^2(\mathbb{R})$  that can be written as

$$f(t) = g_p(t) h(t) \quad \text{with} \quad h(t) = \begin{cases} h(2a_p - t) & \text{if } t \in O_p \\ -h(2a_{p+1} - t) & \text{if } t \in O_{p+1}. \end{cases} \quad (8.86)$$



**FIGURE 8.17**

Each window  $g_p$  has a support  $[a_p - \eta_p, a_{p+1} + \eta_{p+1}]$  with an increasing profile and a decreasing profile over  $[a_p - \eta_p, a_p + \eta_p]$  and  $[a_{p+1} - \eta_{p+1}, a_{p+1} + \eta_{p+1}]$ .

The function  $h$  is symmetric with respect to  $a_p$  and antisymmetric with respect to  $a_{p+1}$ , with an arbitrary behavior in  $C_p$ . Projector  $P_p$  on  $\mathbf{W}^p$  defined in (8.82) can be rewritten as

$$P_p f(t) = \begin{cases} P_{a_p, \eta_p}^- f(t) & \text{if } t \in O_p \\ f(t) & \text{if } t \in C_p \\ P_{a_{p+1}, \eta_{p+1}}^+ f(t) & \text{if } t \in O_{p+1} = g_p(t) h_p(t), \end{cases} \quad (8.87)$$

where  $h_p(t)$  is calculated by inserting (8.75) and (8.76):

$$h_p(t) = \begin{cases} g_p(t)f(t) + g_p(2a_p - t)f(2a_p - t) & \text{if } t \in O_p \\ f(t) & \text{if } t \in C_p \\ g_p(t)f(t) - g_p(2a_{p+1} - t)f(2a_{p+1} - t) & \text{if } t \in O_{p+1}. \end{cases} \quad (8.88)$$

Theorem 8.15 derives a decomposition of the identity.

**Theorem 8.15.** Operator  $P_p$  is an orthogonal projector on  $\mathbf{W}^p$ . If  $p \neq q$ , then  $\mathbf{W}^p$  is orthogonal to  $\mathbf{W}^q$  and

$$\sum_{p=-\infty}^{+\infty} P_p = \text{Id}. \quad (8.89)$$

**Proof.** If  $p \neq q$  and  $|p - q| > 1$ , then functions in  $\mathbf{W}^p$  and  $\mathbf{W}^q$  have supports that do not overlap, so these spaces are orthogonal. If  $q = p + 1$ , then

$$\mathbf{W}^p = \mathbf{W}_{a_p, \eta_p}^+ \cap \mathbf{W}_{a_{p+1}, \eta_{p+1}}^- \quad \text{and} \quad \mathbf{W}^{p+1} = \mathbf{W}_{a_{p+1}, \eta_{p+1}}^+ \cap \mathbf{W}_{a_{p+2}, \eta_{p+2}}^-.$$

Since  $\mathbf{W}_{a_{p+1}, \eta_{p+1}}^-$  is orthogonal to  $\mathbf{W}_{a_{p+1}, \eta_{p+1}}^+$ , it follows that  $\mathbf{W}^p$  is orthogonal to  $\mathbf{W}^{p+1}$ .

To prove (8.89), we first verify that

$$P_p + P_{p+1} = P_{a_p, \eta_p}^+ P_{a_{p+2}, \eta_{p+2}}^-. \quad (8.90)$$

This is shown by decomposing  $P_p$  and  $P_{p+1}$  with (8.87) and inserting

$$P_{a_{p+1}, \eta_{p+1}}^+ + P_{a_{p+1}, \eta_{p+1}}^- = \text{Id}.$$

As a consequence,

$$\sum_{p=n}^m P_p = P_{a_n, \eta_n}^+ P_{a_m, \eta_m}^- \quad (8.91)$$

For any  $f \in \mathbf{L}^2(\mathbb{R})$ ,

$$\|f - P_{a_n, \eta_n}^+ P_{a_m, \eta_m}^- f\|^2 \geq \int_{-\infty}^{a_n + \eta_n} |f(t)|^2 dt + \int_{a_m - \eta_m}^{+\infty} |f(t)|^2 dt$$

and inserting (8.80) proves that

$$\lim_{\substack{n \rightarrow -\infty \\ m \rightarrow +\infty}} \|f - P_{a_n, \eta_n}^+ P_{a_m, \eta_m}^- f\|^2 = 0.$$

The summation (8.91) implies (8.89). ■

### Discretized Projectors

Projectors  $P_p$  that restrict the signal support to  $[a_p - \eta_p, a_{p+1} + \eta_{p+1}]$  are easily extended for discrete signals. Suppose that  $\{a_p\}_{p \in \mathbb{Z}}$  are half integers, which means that  $a_p + 1/2 \in \mathbb{Z}$ . The windows  $g_p(t)$  defined in (8.85) are uniformly sampled  $g_p[n] = g_p(n)$ . As in (8.86) we define the space  $\mathbf{W}^p \subset \ell^2(\mathbb{Z})$  of discrete signals

$$f[n] = g_p[n] h[n] \quad \text{with} \quad h[n] = \begin{cases} h[2a_p - n] & \text{if } n \in O_p \\ -h[2a_{p+1} - n] & \text{if } n \in O_{p+1}. \end{cases} \quad (8.92)$$

The orthogonal projector  $P_p$  on  $\mathbf{W}^p$  is defined by an expression identical to (8.87) and (8.88):

$$P_p f[n] = g_p[n] h_p[n] \quad (8.93)$$

with

$$h_p[n] = \begin{cases} g_p[n]f[n] + g_p[2a_p - n]f[2a_p - n] & \text{if } n \in O_p \\ f[n] & \text{if } n \in C_p \\ g_p[n]f[n] - g_p[2a_{p+1} - n]f[2a_{p+1} - n] & \text{if } n \in O_{p+1}. \end{cases} \quad (8.94)$$

Finally, we prove as in Theorem 8.15 that if  $p \neq q$ , then  $\mathbf{W}^p$  is orthogonal to  $\mathbf{W}^q$  and

$$\sum_{p=-\infty}^{+\infty} P_p = \text{Id.} \quad (8.95)$$

### 8.4.2 Lapped Orthogonal Bases

An orthogonal basis of  $\mathbf{L}^2(\mathbb{R})$  is defined from a basis  $\{e_k\}_{k \in \mathbb{N}}$  of  $\mathbf{L}^2[0, 1]$  by multiplying a translation and dilation of each vector with a smooth window  $g_p$  defined in (8.85). A local cosine basis of  $\mathbf{L}^2(\mathbb{R})$  is derived from a cosine IV basis of  $\mathbf{L}^2[0, 1]$ .

The support of  $g_p$  is  $[a_p - \eta_p, a_{p+1} + \eta_{p+1}]$  with  $l_p = a_{p+1} - a_p$ , as illustrated in Figure 8.17. The design of these windows also implies symmetry and quadrature properties on overlapping intervals:

$$g_p(t) = g_{p+1}(2a_{p+1} - t) \quad \text{for } t \in [a_{p+1} - \eta_{p+1}, a_{p+1} + \eta_{p+1}], \quad (8.96)$$

and

$$g_p^2(t) + g_{p+1}^2(t) = 1 \quad \text{for } t \in [a_{p+1} - \eta_{p+1}, a_{p+1} + \eta_{p+1}].$$

Each  $e_k \in \mathbf{L}^2[0, 1]$  is extended over  $\mathbb{R}$  into a function  $\tilde{e}_k$  that is symmetric with respect to 0 and antisymmetric with respect to 1. The resulting  $\tilde{e}_k$  has period 4 and is defined over  $[-2, 2]$  by

$$\tilde{e}_k(t) = \begin{cases} e_k(t) & \text{if } t \in [0, 1] \\ e_k(-t) & \text{if } t \in (-1, 0) \\ -e_k(2-t) & \text{if } t \in [1, 2) \\ -e_k(2+t) & \text{if } t \in [-1, -2). \end{cases}$$

Theorem 8.16 derives an orthonormal basis of  $\mathbf{L}^2(\mathbb{R})$ .

**Theorem 8.16:** *Coifman, Malvar, Meyer.* Let  $\{e_k\}_{k \in \mathbb{N}}$  be an orthonormal basis of  $\mathbf{L}^2[0, 1]$ . The family

$$\left\{ g_{p,k}(t) = g_p(t) \frac{1}{\sqrt{l_p}} \tilde{e}_k\left(\frac{t-a_p}{l_p}\right) \right\}_{k \in \mathbb{N}, p \in \mathbb{Z}} \quad (8.97)$$

is an orthonormal basis of  $\mathbf{L}^2(\mathbb{R})$ .

**Proof.** Since  $\tilde{e}_k(l_p^{-1}(t - a_p))$  is symmetric with respect to  $a_p$  and antisymmetric with respect to  $a_{p+1}$ , it follows from (8.86) that  $g_{p,k} \in \mathbf{W}^p$  for all  $k \in \mathbb{N}$ . Theorem 8.15 proves that

spaces  $\mathbf{W}^p$  and  $\mathbf{W}^q$  are orthogonal for  $p \neq q$  and that  $\mathbf{L}^2(\mathbb{R}) = \bigoplus_{p=-\infty}^{+\infty} \mathbf{W}^p$ . To prove that (8.97) is an orthonormal basis of  $\mathbf{L}^2(\mathbb{R})$ , we thus need to show that

$$\left\{ g_{p,k}(t) = g_p(t) \frac{1}{\sqrt{l_p}} \tilde{e}_k \left( \frac{t - a_p}{l_p} \right) \right\}_{k \in \mathbb{N}, p \in \mathbb{Z}} \quad (8.98)$$

is an orthonormal basis of  $\mathbf{W}^p$ .

Let us prove first that any  $f \in \mathbf{W}^p$  can be decomposed over this family. Such a function can be as written as  $f(t) = g_p(t) h(t)$ , where the restriction of  $h$  to  $[a_p, a_{p+1}]$  is arbitrary, and  $h$  is, respectively, symmetric and antisymmetric with respect to  $a_p$  and  $a_{p+1}$ . Since  $\{\tilde{e}_k\}_{k \in \mathbb{N}}$  is an orthonormal basis of  $\mathbf{L}^2[0, 1]$ , clearly

$$\left\{ \frac{1}{\sqrt{l_p}} \tilde{e}_k \left( \frac{t - a_p}{l_p} \right) \right\}_{k \in \mathbb{N}} \quad (8.99)$$

is an orthonormal basis of  $\mathbf{L}^2[a_p, a_{p+1}]$ . The restriction of  $h$  to  $[a_p, a_{p+1}]$  can therefore be decomposed in this basis. This decomposition remains valid for all  $t \in [a_p - \eta_p, a_{p+1} + \eta_{p+1}]$  since  $h(t)$  and the  $l_p^{-1/2} \tilde{e}_k(l_p^{-1}(t - a_p))$  have the same symmetry with respect to  $a_p$  and  $a_{p+1}$ . Therefore,  $f(t) = h(t)g_p(t)$  can be decomposed over the family (8.98). Lemma 8.1 finishes the proof by showing that the orthogonality of functions in (8.98) is a consequence of the orthogonality of (8.99) in  $\mathbf{L}^2[a_p, a_{p+1}]$ .

**Lemma 8.1.** If  $f_b(t) = h_b(t)g_p(t) \in \mathbf{W}^p$  and  $f_c(t) = h_c(t)g_p(t) \in \mathbf{W}^p$ , then

$$\langle f_b, f_c \rangle = \int_{a_p - \eta_p}^{a_{p+1} + \eta_{p+1}} f_b(t) f_c^*(t) dt = \int_{a_p}^{a_{p+1}} h_b(t) h_c^*(t) dt. \quad (8.100)$$

Let us evaluate

$$\langle f_b, f_c \rangle = \int_{a_p - \eta_p}^{a_{p+1} + \eta_{p+1}} h_b(t) h_c^*(t) g_p^2(t) dt. \quad (8.101)$$

We know that  $h_b(t)$  and  $h_c(t)$  are symmetric with respect to  $a_p$ , so

$$\int_{a_p - \eta_p}^{a_p + \eta_p} h_b(t) h_c^*(t) g_p^2(t) dt = \int_{a_p}^{a_p + \eta_p} h_b(t) h_c^*(t) [g_p^2(t) + g_p^2(2a_p - t)] dt.$$

Since  $g_p^2(t) + g_p^2(2a_p - t) = 1$  over this interval, we obtain

$$\int_{a_p - \eta_p}^{a_p + \eta_p} h_b(t) h_c^*(t) g_p^2(t) dt = \int_{a_p}^{a_p + \eta_p} h_b(t) h_c(t) dt. \quad (8.102)$$

The functions  $h_b(t)$  and  $h_c(t)$  are antisymmetric with respect to  $a_{p+1}$ , so  $h_b(t)h_c^*(t)$  is symmetric about  $a_{p+1}$ . Thus we prove similarly that

$$\int_{a_{p+1} - \eta_{p+1}}^{a_{p+1} + \eta_{p+1}} h_b(t) h_c^*(t) g_{p+1}^2(t) dt = \int_{a_{p+1} - \eta_{p+1}}^{a_{p+1}} h_b(t) h_c^*(t) dt. \quad (8.103)$$

Since  $g_p(t) = 1$  for  $t \in [a_p + \eta_p, a_{p+1} - \eta_{p+1}]$ , inserting (8.102) and (8.103) in (8.101) proves the lemma property (8.100). ■

Theorem 8.16 is similar to the block basis theorem (8.8) but it has the advantage of using smooth windows  $g_p$  as opposed to the rectangular windows that are indicator functions of  $[a_p, a_{p+1}]$ . It yields smooth functions  $g_{p,k}$  only if the extension  $\tilde{e}_k$  of  $e_k$  is a smooth function. This is the case for the cosine IV basis  $\{e_k(t) = \sqrt{2} \cos[(k + 1/2)\pi t]\}_{k \in \mathbb{N}}$  of  $L^2[0, 1]$  defined in Theorem 8.11. Indeed  $\cos[(k + 1/2)\pi t]$  has a natural symmetric and antisymmetric extension with respect to 0 and 1 over  $\mathbb{R}$ . Corollary 8.1 derives a local cosine basis.

**Corollary 8.1.** The family of local cosine functions

$$\left\{ g_{p,k}(t) = g_p(t) \sqrt{\frac{2}{l_p}} \cos \left[ \pi \left( k + \frac{1}{2} \right) \frac{t - a_p}{l_p} \right] \right\}_{k \in \mathbb{N}, p \in \mathbb{Z}} \quad (8.104)$$

is an orthonormal basis of  $L^2(\mathbb{R})$ .

### Cosine–Sine I Basis

Other bases can be constructed with functions having a different symmetry. To maintain the orthogonality of the windowed basis, we must ensure that consecutive windows  $g_p$  and  $g_{p+1}$  are multiplied by functions that have an opposite symmetry with respect to  $a_{p+1}$ . For example, we can multiply  $g_{2p}$  with functions that are symmetric with respect to both ends  $a_{2p}$  and  $a_{2p+1}$ , and multiply  $g_{2p+1}$  with functions that are antisymmetric with respect to  $a_{2p+1}$  and  $a_{2p+2}$ . Such bases can be constructed with the cosine I basis  $\{\sqrt{2}\lambda_k \cos(\pi kt)\}_{k \in \mathbb{Z}}$  defined in Theorem 8.10, with  $\lambda_0 = 2^{-1/2}$  and  $\lambda_k = 1$  for  $k \neq 0$ , and with the sine I family  $\{\sqrt{2} \sin(\pi kt)\}_{k \in \mathbb{N}^*}$ , which is also an orthonormal basis of  $L^2[0, 1]$ . The reader can verify that if

$$g_{2p,k}(t) = g_{2p}(t) \sqrt{\frac{2}{l_{2p}}} \lambda_k \cos \left[ \pi k \frac{t - a_{2p}}{l_{2p}} \right]$$

$$g_{2p+1,k}(t) = g_{2p+1}(t) \sqrt{\frac{2}{l_{2p+1}}} \sin \left[ \pi k \frac{t - a_{2p+1}}{l_{2p+1}} \right],$$

then  $\{g_{p,k}\}_{k \in \mathbb{N}, p \in \mathbb{Z}}$  is an orthonormal basis of  $L^2(\mathbb{R})$ .

### Lapped Transforms in Frequency

Lapped orthogonal projectors can also divide the frequency axis in separate overlapping intervals. This is done by decomposing the Fourier transform  $\hat{f}(\omega)$  of  $f(t)$  over a local cosine basis defined on the frequency axis  $\{g_{p,k}(\omega)\}_{p \in \mathbb{Z}, k \in \mathbb{N}}$ . This is also equivalent to decomposing  $f(t)$  on its inverse Fourier transform  $\{\frac{1}{2\pi} \hat{g}_{p,k}(-t)\}_{p \in \mathbb{Z}, k \in \mathbb{N}}$ . As opposed to wavelet packets, which decompose signals in dyadic-frequency bands, this approach offers complete flexibility on the size of the frequency intervals  $[a_p - \eta_p, a_{p+1} + \eta_{p+1}]$ .

A signal decomposition in a Meyer wavelet or wavelet packet basis can be calculated with a lapped orthogonal transform applied in the Fourier domain. Indeed, the Fourier transform (7.87) of a Meyer wavelet has a compact support and

$\{|\hat{\psi}(2^j\omega)|\}_{j \in \mathbb{Z}}$  can be considered as a family asymmetric window, with support that only overlaps with adjacent windows with appropriate symmetry properties. These windows cover the whole frequency axis:  $\sum_{j=-\infty}^{+\infty} |\hat{\psi}(2^j\omega)|^2 = 1$ . As a result, the Meyer wavelet transform can be viewed as a lapped orthogonal transform applied in the Fourier domain. Thus, it can be efficiently implemented with the folding algorithm from Section 8.4.4.

### 8.4.3 Local Cosine Bases

The local cosine basis defined in (8.104) is composed of functions

$$g_{p,k}(t) = g_p(t) \sqrt{\frac{2}{l_p}} \cos \left[ \pi \left( k + \frac{1}{2} \right) \frac{t - a_p}{l_p} \right]$$

with a compact support  $[a_p - \eta_p, a_{p+1} + \eta_{p+1}]$ . The energy of their Fourier transforms is also well concentrated. Let  $\hat{g}_p$  be the Fourier transform of  $g_p$ ,

$$\hat{g}_{p,k}(\omega) = \frac{\exp(-ia_p \xi_{p,k})}{2} \sqrt{\frac{2}{l_p}} (\hat{g}_p(\omega - \xi_{p,k}) + \hat{g}_p(\omega + \xi_{p,k})),$$

where

$$\xi_{p,k} = \frac{\pi(k + 1/2)}{l_p}.$$

The bandwidth of  $\hat{g}_{p,k}$  around  $\xi_{p,k}$  and  $-\xi_{p,k}$  is equal to the bandwidth of  $\hat{g}_p$ . If sizes  $\eta_p$  and  $\eta_{p+1}$  of the variation intervals of  $g_p$  are proportional to  $l_p$ , then this bandwidth is proportional to  $l_p^{-1}$ .

For smooth functions  $f$ , we want to guarantee that the inner products  $\langle f, g_{p,k} \rangle$  have a fast decay when the center frequency  $\xi_{p,k}$  increases. The Parseval formula proves that

$$\langle f, g_{p,k} \rangle = \frac{\exp(ia_p \xi_{p,k})}{2\pi} \sqrt{\frac{2}{l_p}} \int_{-\infty}^{+\infty} \hat{f}(\omega) (\hat{g}_p^*(\omega - \xi_{p,k}) + \hat{g}_p^*(\omega + \xi_{p,k})) d\omega.$$

The smoothness of  $f$  implies that  $|\hat{f}(\omega)|$  has a fast decay at large frequencies  $\omega$ . Therefore, this integral will become small when  $\xi_{p,k}$  increases if  $g_p$  is a smooth window, because  $|\hat{g}_p(\omega)|$  has a fast decay.

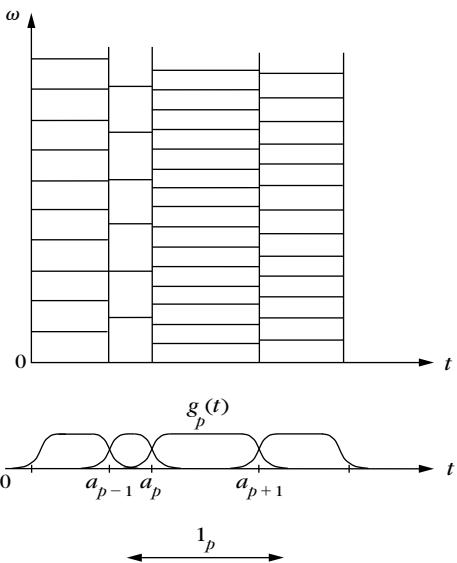
#### Window Design

The regularity of  $g_p$  depends on the regularity of profile  $\beta$ , as shown by (8.85). This profile must satisfy

$$\beta^2(t) + \beta^2(-t) = 1 \quad \text{for } t \in [-1, 1], \tag{8.105}$$

plus  $\beta(t) = 0$  if  $t < -1$  and  $\beta(t) = 1$  if  $t > 1$ . One example is

$$\beta_0(t) = \sin\left(\frac{\pi}{4}(1+t)\right) \quad \text{for } t \in [-1, 1],$$

**FIGURE 8.18**

Heisenberg boxes of local cosine vectors define a regular grid over the time-frequency plane.

but its derivative at  $t = \pm 1$  is nonzero, so  $\beta$  is not differentiable at  $\pm 1$ . Windows of higher regularity are constructed with a profile  $\beta_k$  defined by induction for  $k \geq 0$  by

$$\beta_{k+1}(t) = \beta_k \left( \sin \frac{\pi t}{2} \right) \quad \text{for } t \in [-1, 1].$$

For any  $k \geq 0$ , one can verify that  $\beta_k$  satisfies (8.105) and has  $2^k - 1$  vanishing derivatives at  $t = \pm 1$ . The resulting  $\beta$  and  $g_p$  are therefore  $2^k - 1$  times continuously differentiable.

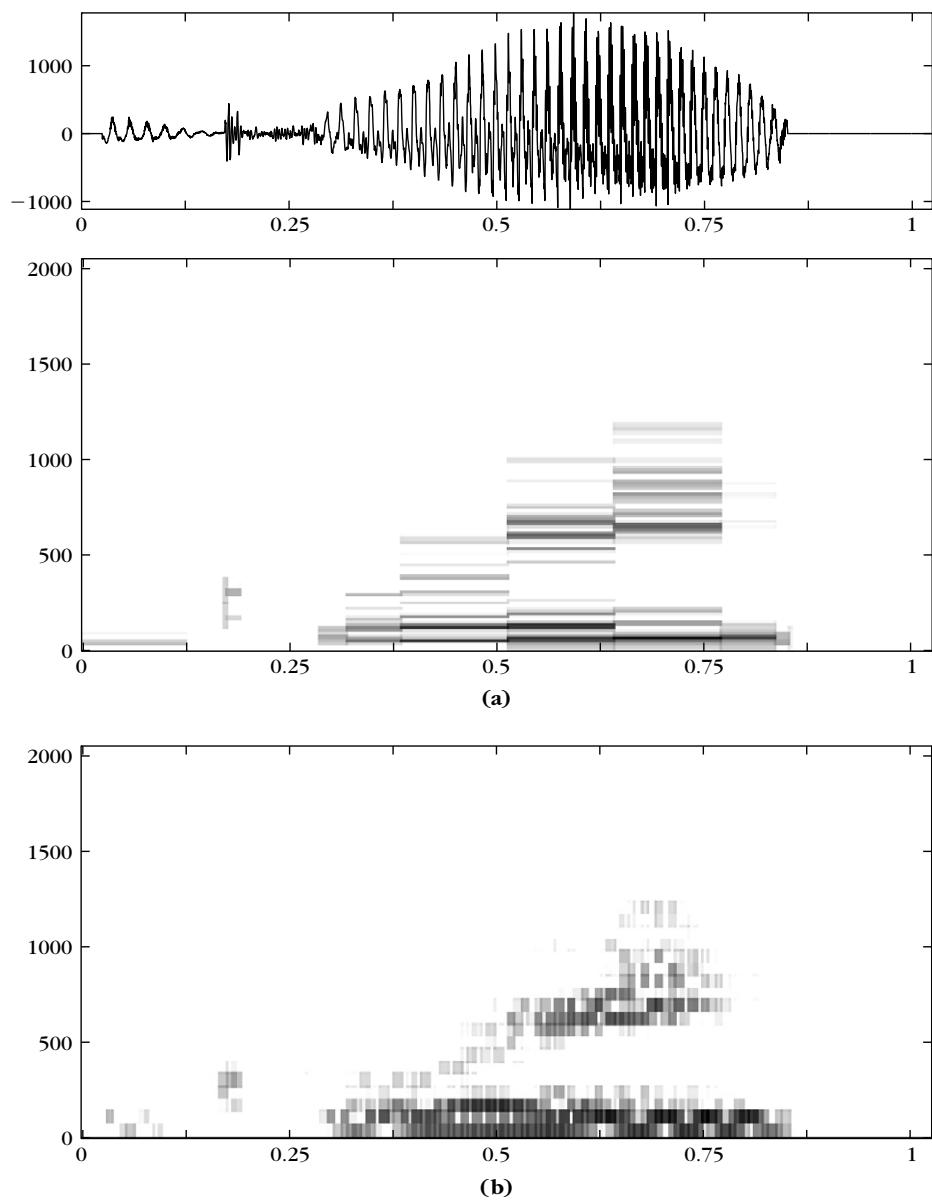
### **Heisenberg Box**

A local cosine basis can be symbolically represented as an exact paving of the time-frequency plane. The time and frequency region of high-energy concentration for each local cosine vector  $g_{p,k}$  is approximated by a Heisenberg rectangle

$$[a_p, a_{p+1}] \times \left[ \xi_{p,k} - \frac{\pi}{2l_p}, \xi_{p,k} + \frac{\pi}{2l_p} \right],$$

as illustrated in Figure 8.18. A local cosine basis  $\{g_{p,k}\}_{k \in \mathbb{N}, p \in \mathbb{Z}}$  corresponds to a time-frequency grid which varies in time.

Figure 8.19(a) shows the decomposition of a digital recording of the sound “grea” coming from the word “greasy.” The window sizes are adapted to the signal structures with the best basis algorithm described in Section 12.2.2. High-amplitude

**FIGURE 8.19**

(a) The signal at the top is a recording of the sound “grea” in the word “greasy.” This signal is decomposed in a local cosine basis with windows of varying sizes. The larger the amplitude of  $|\langle f, g_{p,k} \rangle|$ , the darker the gray level of the Heisenberg box. (b) Decomposition in a local cosine basis with small windows of constant size.

coefficients are along spectral lines in the time-frequency plane. Most Heisenberg boxes appear in white, which indicates that the corresponding inner product is nearly zero. Thus, this signal can be approximated with a few nonzero local cosine vectors. Figure 8.19(b) decomposes the same signal in a local cosine basis composed of small windows of constant size. The signal time-frequency structures do not appear as well as in Figure 8.19(a).

### **Translation and Phase**

Cosine modulations as opposed to complex exponentials do not provide easy access to phase information. The translation of a signal can induce important modifications of its decomposition coefficients in a cosine basis. Consider, for example,

$$f(t) = g_{p,k}(t) = g_p(t) \sqrt{\frac{2}{l_p}} \cos \left[ \pi \left( k + \frac{1}{2} \right) \frac{t - a_p}{l_p} \right].$$

Since the basis is orthogonal,  $\langle f, g_{p,k} \rangle = 1$  and all other inner products are zero. After a translation by  $\tau = l_p/(2k+1)$ ,

$$f_\tau(t) = f \left( t - \frac{l_p}{2k+1} \right) = g_p(t) \sqrt{\frac{2}{l_p}} \sin \left[ \pi \left( k + \frac{1}{2} \right) \frac{t - a_p}{l_p} \right].$$

The opposite parity of sine and cosine implies that  $\langle f_\tau, g_{p,k} \rangle \approx 0$ . In contrast,  $\langle f_\tau, g_{p,k-1} \rangle$  and  $\langle f_\tau, g_{p,k+1} \rangle$  become nonzero. After translation, a signal component initially represented by a cosine of frequency  $\pi(k+1/2)/l_p$  is therefore spread over cosine vectors of different frequencies.

This example shows that the local cosine coefficients of a pattern are severely modified by any translation. We are facing the same translation distortions as observed in Section 5.1.5 for wavelets and time-frequency frames. This lack of translation invariance makes it difficult to use these bases for pattern recognition.

### **8.4.4 Discrete Lapped Transforms**

Lapped orthogonal bases are discretized by replacing the orthogonal basis of  $L^2[0, 1]$  with a discrete basis of  $\mathbb{C}^N$ , and uniformly sampling the windows  $g_p$ . Discrete local cosine bases are derived with discrete cosine IV bases.

Let  $\{a_p\}_{p \in \mathbb{Z}}$  be a sequence of half integers  $a_p + 1/2 \in \mathbb{Z}$  with

$$\lim_{p \rightarrow -\infty} a_p = -\infty \quad \text{and} \quad \lim_{p \rightarrow +\infty} a_p = +\infty.$$

A discrete lapped orthogonal basis is constructed with the discrete projectors  $P_p$  defined in (8.93). These operators are implemented with the sampled windows  $g_p[n] = g_p(n)$ . Suppose that  $\{e_{k,l}[n]\}_{0 \leq k < l}$  is an orthogonal basis of signals defined for  $0 \leq n < l$ . These vectors are extended over  $\mathbb{Z}$  with a symmetry with respect to  $-1/2$  and an antisymmetry with respect to  $l - 1/2$ . The resulting extensions have a

period  $4l$  and are defined over  $[-2l, 2l - 1]$  by

$$\tilde{e}_{l,k}[n] = \begin{cases} e_{l,k}[n] & \text{if } n \in [0, l-1] \\ e_{l,k}[-1-n] & \text{if } n \in [-l, -1] \\ -e_k[2l-1-n] & \text{if } n \in [l, 2l-1] \\ -e_k[2l+n] & \text{if } n \in [-2l, -l-1]. \end{cases}$$

Theorem 8.17 proves that multiplying these vectors with the discrete windows  $g_p[n]$  yields an orthonormal basis of  $\ell^2(\mathbb{Z})$ .

**Theorem 8.17:** *Coifman, Malvar, Meyer.* Suppose that  $\{e_{k,l}\}_{0 \leq k < l}$  is an orthogonal basis of  $\mathbb{C}^l$  for any  $l > 0$ . The family

$$\left\{ g_{p,k}[n] = g_p[n] \tilde{e}_{k,l_p}[n - a_p] \right\}_{0 \leq k < l_p, p \in \mathbb{Z}} \quad (8.106)$$

is a lapped orthonormal basis of  $\ell^2(\mathbb{Z})$ .

The proof of this theorem is identical to the proof of Theorem 8.16 since we have a discrete equivalent of the spaces  $\mathbf{W}^p$  and their projectors. It is also based on a discrete equivalent of Lemma 8.1, which is verified with the same derivations. Beyond the proof of Theorem 8.17, we shall see that Lemma 8.2 is important for quickly computing the decomposition coefficients  $\langle f, g_{p,k} \rangle$ . ■

**Lemma 8.2.** Any  $f_b[n] = g_p[n] h_b[n] \in \mathbf{W}^p$  and  $f_c[n] = g_p[n] h_c[n] \in \mathbf{W}^p$  satisfy

$$\langle f_b, f_c \rangle = \sum_{a_p - \eta_p < n < a_{p+1} + \eta_{p+1}} f_b[n] f_c^*[n] = \sum_{a_p < n < a_{p+1}} h_b[n] h_c^*[n]. \quad (8.107)$$

Theorem 8.17 is similar to the discrete block basis theorem (8.9) but constructs an orthogonal basis with smooth discrete windows  $g_p[n]$ . The discrete cosine IV bases

$$\left\{ e_{l,k}[n] = \sqrt{\frac{2}{l}} \cos \left[ \frac{\pi}{l} \left( k + \frac{1}{2} \right) \left( n + \frac{1}{2} \right) \right] \right\}_{0 \leq k < l}$$

have the advantage of including vectors that have a natural symmetric and antisymmetric extension with respect to  $-1/2$  and  $l - 1/2$ . This produces a discrete local cosine basis of  $\ell^2(\mathbb{Z})$ .

**Corollary 8.2.** The family

$$\left\{ g_{p,k}[n] = g_p[n] \sqrt{\frac{2}{l_p}} \cos \left[ \pi \left( k + \frac{1}{2} \right) \frac{n - a_p}{l_p} \right] \right\}_{0 \leq k < l_p, p \in \mathbb{Z}} \quad (8.108)$$

is an orthonormal basis of  $\ell^2(\mathbb{Z})$ .

### Fast Lapped Orthogonal Transform

A fast algorithm introduced by Malvar [40] replaces the calculations of  $\langle f, g_{p,k} \rangle$  by a computation of inner products in the original bases  $\{e_{l,k}\}_{0 \leq k < l}$  with a folding procedure. In a discrete local cosine basis, these inner products are calculated with the fast DCT-IV algorithm.

To simplify notations, as in Section 8.4.1 we decompose  $I_p = [a_p - \eta_p, a_{p+1} + \eta_{p+1}]$  into  $I_p = O_p \cup C_p \cup O_{p+1}$  with

$$O_p = [a_p - \eta_p, a_p + \eta_p] \quad \text{and} \quad C_p = [a_p + \eta_p, a_{p+1} - \eta_{p+1}].$$

The orthogonal projector  $P_p$  on space  $\mathbf{W}^p$  generated by  $\{g_{p,k}\}_{0 \leq k < l_p}$  was calculated in (8.93):

$$P_p f[n] = g_p[n] h_p[n],$$

where  $h_p$  is a folded version of  $f$ :

$$h_p[n] = \begin{cases} g_p[n]f[n] + g_p[2a_p - n]f[2a_p - n] & \text{if } n \in O_p \\ f[n] & \text{if } n \in C_p \\ g_p[n]f[n] - g_p[2a_{p+1} - n]f[2a_{p+1} - n] & \text{if } n \in O_{p+1}. \end{cases} \quad (8.109)$$

Since  $g_{p,k} \in \mathbf{W}^p$ ,

$$\langle f, g_{p,k} \rangle = \langle P_p f, g_{p,k} \rangle = \langle g_p h_p, g_p \tilde{e}_{l_p,k} \rangle.$$

Since  $\tilde{e}_{l_p,k}[n] = e_{l_p,k}[n]$  for  $n \in [a_p, a_{p+1}]$ , Lemma 8.2 derives that

$$\langle f, g_{p,k} \rangle = \sum_{a_p < n < a_{p+1}} h_p[n] e_{l_p,k}[n] = \langle h_p, e_{l_p,k} \rangle_{[a_p, a_{p+1}]} \quad (8.110)$$

This proves that the decomposition coefficients  $\langle f, g_{p,k} \rangle$  can be calculated by folding  $f$  into  $h_p$  and computing the inner product with the orthogonal basis  $\{e_{l_p,k}\}_{0 \leq k < l_p}$  defined over  $[a_p, a_{p+1}]$ .

For a discrete cosine basis, the DCT-IV coefficients

$$\langle h_p, e_{l_p,k} \rangle_{[a_p, a_{p+1}]} = \sum_{a_p < n < a_{p+1}} h_p[n] \sqrt{\frac{2}{l_p}} \cos \left[ \pi \left( k + \frac{1}{2} \right) \frac{n - a_p}{l_p} \right] \quad (8.111)$$

are computed with the fast DCT-IV algorithm from Section 8.3.4, which requires  $O(l_p \log_2 l_p)$  operations. The inverse lapped transform recovers  $h_p[n]$  over  $[a_p, a_{p+1}]$  from the  $l_p$  inner products  $\{\langle h_p, e_{l_p,k} \rangle_{[a_p, a_{p+1}]} \}_{0 \leq k < l_p}$ . In a local cosine IV basis, this is done with the fast inverse DCT-IV, which is identical to the forward DCT-IV and requires  $O(l_p \log_2 l_p)$  operations. The reconstruction of  $f$  is done by applying (8.95), which proves that

$$f[n] = \sum_{p=-\infty}^{+\infty} P_p f[n] = \sum_{p=-\infty}^{+\infty} g_p[n] h_p[n]. \quad (8.112)$$

Let us denote  $O_p^- = [a_p - \eta_p, a_p]$  and  $O_p^+ = [a_p, a_p + \eta_p]$ . The restriction of (8.112) to  $[a_p, a_{p+1}]$  gives

$$f[n] = \begin{cases} g_p[n] h_p[n] + g_{p-1}[n] h_{p-1}[n] & \text{if } n \in O_p^+ \\ h_p[n] & \text{if } n \in C_p \\ g_p[n] h_p[n] + g_{p+1}[n] h_{p+1}[n] & \text{if } n \in O_{p+1}^-. \end{cases}$$

The symmetry of the windows guarantees that  $g_{p-1}[n] = g_p[2a_p - n]$  and  $g_{p+1}[n] = g_p[2a_{p+1} - n]$ . Since  $h_{p-1}[n]$  is antisymmetric with respect to  $a_p$  and  $h_{p+1}[n]$  is symmetric with respect to  $a_{p+1}$ , we can recover  $f[n]$  on  $[a_p, a_{p+1}]$  from the values of  $h_{p-1}[n]$ ,  $h_p[n]$ , and  $h_{p+1}[n]$  computed, respectively, on  $[a_{p-1}, a_p]$ ,  $[a_p, a_{p+1}]$ , and  $[a_{p+1}, a_{p+2}]$ :

$$f[n] = \begin{cases} g_p[n] h_p[n] - g_p[2a_p - n] h_{p-1}[2a_p - n] & \text{if } n \in O_p^+ \\ h_p[n] & \text{if } n \in C_p \\ g_p[n] h_p[n] + g_p[2a_{p+1} - n] h_{p+1}[2a_{p+1} - n] & \text{if } n \in O_{p+1}^-. \end{cases} \quad (8.113)$$

This unfolding formula is implemented with  $O(l_p)$  calculations. Thus, the inverse local cosine transform requires  $O(l_p \log_2 l_p)$  operations to recover  $f[n]$  on each interval  $[a_p, a_{p+1}]$  of length  $l_p$ .

### Finite Signals

If  $f[n]$  is defined for  $0 \leq n < N$ , the extremities of the first and last interval must be  $a_0 = -1/2$  and  $a_q = N - 1/2$ . A fast local cosine algorithm needs  $O(l_p \log_2 l_p)$  additions and multiplications to decompose or reconstruct the signal on each interval of length  $l_p$ . On the whole signal of length  $N$ , it thus needs a total of  $O(N \log_2 L)$  operations, where  $L = \sup_{0 \leq p < q} l_p$ .

Since we do not know the values of  $f[n]$  for  $n < 0$ , at the left border we set  $\eta_0 = 0$ . This means that  $g_0[n]$  jumps from 0 to 1 at  $n = 0$ . The resulting transform on the left boundary is equivalent to a straight DCT-IV. Section 8.3.2 shows that since cosine IV vectors are even on the left boundary, the DCT-IV is equivalent to a symmetric signal extension followed by a discrete Fourier transform. This avoids creating discontinuity artifacts at the left border.

At the right border, we also set  $\eta_q = 0$  to limit the support of  $g_{q-1}$  to  $[0, N - 1]$ . Section 8.4.4 explains that since cosine IV vectors are odd on the right boundary, the DCT-IV is equivalent to an antisymmetric signal extension. If  $f[N - 1] \neq 0$ , this extension introduces a sharp signal transition that creates artificial high frequencies. To reduce this border effect, we replace the cosine IV modulation

$$g_{q-1,k}[n] = g_{q-1}[n] \sqrt{\frac{2}{l_{q-1}}} \cos \left[ \pi \left( k + \frac{1}{2} \right) \frac{n - a_{q-1}}{l_{q-1}} \right]$$

with a cosine I modulation,

$$g_{q-1,k}[n] = g_{q-1}[n] \sqrt{\frac{2}{l_{q-1}}} \lambda_k \cos \left[ \pi k \frac{n - a_{q-1}}{l_{q-1}} \right].$$

The orthogonality with the other elements of the basis is maintained because these cosine I vectors, like cosine IV vectors, are even with respect to  $a_{q-1}$ . Since  $\cos[\pi k(n - a_{q-1})/l_{q-1}]$  is also symmetric with respect to  $a_q = N - 1/2$ , computing a DCT-I is equivalent to performing a symmetric signal extension at the right boundary, which avoids discontinuities. In the fast local cosine transform, we thus compute a DCT-I of the last folded signal  $h_{q-1}$  instead of a DCT-IV. The reconstruction algorithm uses an inverse DCT-I to recover  $h_{q-1}$  from these coefficients.

## 8.5 LOCAL COSINE TREES

Corollary 8.1 constructs local cosine bases for any segmentation of the time axis into intervals  $[a_p, a_{p+1}]$  of arbitrary lengths. This result is more general than the construction of wavelet packet bases that can only divide the frequency axis into dyadic intervals with a length proportional to a power of 2. However, Coifman and Meyer [181] showed that restricting the intervals to dyadic sizes has the advantage of creating a tree structure similar to a wavelet packet tree. “Best” local cosine bases can then be adaptively chosen with the fast dynamical programming algorithm, described in Section 12.2.2.

### 8.5.1 Binary Tree of Cosine Bases

A local cosine tree includes orthogonal bases that segment the time axis in dyadic intervals. For any  $j \geq 0$ , the interval  $[0, 1]$  is divided in  $2^j$  intervals of length  $2^{-j}$  by setting

$$a_{p,j} = p 2^{-j} \quad \text{for } 0 \leq p \leq 2^j.$$

These intervals are covered by windows  $g_{p,j}$  defined by (8.85) with a support  $[a_{p,j} - \eta, a_{p+1,j} + \eta]$ :

$$g_{p,j}(t) = \begin{cases} \beta(\eta^{-1}(t - a_{p,j})) & \text{if } t \in [a_{p,j} - \eta, a_{p,j} + \eta] \\ 1 & \text{if } t \in [a_{p,j} + \eta, a_{p+1,j} - \eta] \\ \beta(\eta^{-1}(a_{p+1,j} - t)) & \text{if } t \in [a_{p+1,j} - \eta, a_{p+1,j} + \eta] \\ 0 & \text{otherwise.} \end{cases} \quad (8.114)$$

To ensure that the support of  $g_{p,j}$  is in  $[0, 1]$  for  $p = 0$  and  $p = 2^j - 1$ , we modify, respectively, the left and right sides of these windows by setting  $g_{0,j}(t) = 1$  if  $t \in [0, \eta]$ , and  $g_{2^j-1,j}(t) = 1$  if  $t \in [1 - \eta, 1]$ . It follows that  $g_{0,0} = \mathbf{1}_{[0,1]}$ . The size  $\eta$  of the raising and decaying profiles of  $g_{p,j}$  is independent of  $j$ . To guarantee that windows overlap only with their two neighbors, length  $a_{p+1,j} - a_{p,j} = 2^{-j}$  must be larger than size  $2\eta$  of the overlapping intervals, and thus

$$\eta \leq 2^{-j-1}. \quad (8.115)$$

Similar to wavelet packet trees, a local cosine tree is constructed by recursively dividing spaces built with local cosine bases. A tree node at depth  $j$  and position  $p$

is associated to space  $\mathbf{W}_j^p$  generated by the local cosine family

$$\mathcal{B}_j^p = \left\{ g_{p,j}(t) \sqrt{\frac{2}{2^{-j}}} \cos \left[ \pi \left( k + \frac{1}{2} \right) \frac{t - a_{p,j}}{2^{-j}} \right] \right\}_{k \in \mathbb{Z}}. \quad (8.116)$$

Any  $f \in \mathbf{W}_j^p$  has a support in  $[a_{p,j} - \eta, a_{p+1,j} + \eta]$  and can be written as  $f(t) = g_{p,j}(t) h(t)$ , where  $h(t)$  is symmetric and antisymmetric, respectively, to  $a_{p,j}$  and  $a_{p+1,j}$ . Theorem 8.18 shows that  $\mathbf{W}_j^p$  is divided in two orthogonal spaces  $\mathbf{W}_{j+1}^{2p}$  and  $\mathbf{W}_{j+1}^{2p+1}$  that are built over the two half intervals.

**Theorem 8.18:** Coifman, Meyer. For any  $j \geq 0$  and  $p < 2^j$ , spaces  $\mathbf{W}_{j+1}^{2p}$  and  $\mathbf{W}_{j+1}^{2p+1}$  are orthogonal and

$$\mathbf{W}_j^p = \mathbf{W}_{j+1}^{2p} \oplus \mathbf{W}_{j+1}^{2p+1}. \quad (8.117)$$

**Proof.** The orthogonality of  $\mathbf{W}_{j+1}^{2p}$  and  $\mathbf{W}_{j+1}^{2p+1}$  is proved by Theorem 8.15. We denote  $P_{p,j}$  as the orthogonal projector on  $\mathbf{W}_j^p$ . With the notation of Section 8.4.1, this projector is decomposed into two splitting projectors at  $a_{p,j}$  and  $a_{p+1,j}$ :

$$P_{p,j} = P_{a_{p,j}, \eta}^+ P_{a_{p+1,j}, \eta}^-.$$

Equation (8.90) proves that

$$P_{2p,j+1} + P_{2p+1,j+1} = P_{a_{2p,j+1}, \eta}^+ P_{a_{2p+2,j+1}, \eta}^- = P_{a_{p,j}, \eta}^+ P_{a_{p+1,j}, \eta}^- = P_{p,j}.$$

This equality on orthogonal projectors implies (8.117). ■

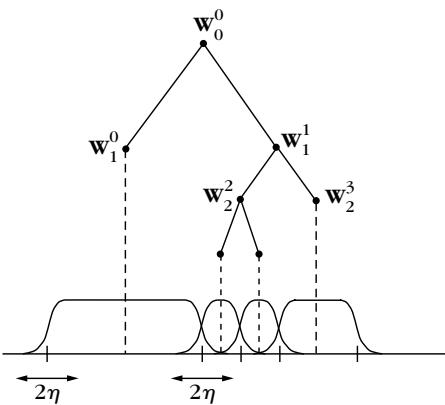
Space  $\mathbf{W}_j^p$  located at node  $(j, p)$  of a local cosine tree is therefore the sum of the two spaces  $\mathbf{W}_{j+1}^{2p}$  and  $\mathbf{W}_{j+1}^{2p+1}$  located at the children nodes. Since  $g_{0,0} = \mathbf{1}_{[0,1]}$  it follows that  $\mathbf{W}_0^0 = \mathbf{L}^2[0, 1]$ . The maximum depth  $J$  of the binary tree is limited by the support condition  $\eta \leq 2^{-J-1}$ , and thus

$$J \leq -\log_2(2\eta). \quad (8.118)$$

### Admissible Local Cosine Bases

As in a wavelet packet binary tree, many local cosine orthogonal bases are constructed from this local cosine tree. We call any subtree of the local cosine tree with nodes that have either zero or two children *an admissible binary tree*. Let  $\{j_i, p_i\}_{1 \leq i \leq I}$  be the indices at the leaves of a particular admissible binary tree. Applying the splitting property (8.117) along the branches of this subtree proves that

$$\mathbf{L}^2[0, 1] = \mathbf{W}_0^0 = \bigoplus_{i=1}^I \mathbf{W}_{j_i}^{p_i}.$$

**FIGURE 8.20**

An admissible binary tree of local cosine spaces divides the time axis in windows of dyadic lengths.

Thus, the union of local cosine bases  $\bigcup_{i=1}^I \mathcal{B}_{j_i}^{p_i}$  is an orthogonal basis of  $\mathbf{L}^2[0, 1]$ . This can also be interpreted as a division of the time axis into windows of various length, as illustrated by Figure 8.20.

The number  $B_J$  of different dyadic local cosine bases is equal to the number of different admissible subtrees of depth of at most  $J$ . For  $J = -\log_2(2\eta)$ , Theorem 8.2 proves that

$$2^{1/(4\eta)} \leq B_J \leq 2^{3/(8\eta)}.$$

Figure 8.19 on page 421 shows the decomposition of a sound recording in two dyadic local cosine bases selected from the binary tree. The basis in (a) is calculated with the best basis algorithm of Section 12.2.2.

### **Choice of $\eta$**

At all scales  $2^j$ , windows  $g_{p,j}$  of a local cosine tree have raising and decaying profiles of the same size  $\eta$ . Thus, these windows can be recombined independently from their scale. If  $\eta$  is small compared to the interval size  $2^{-j}$ , then  $g_{p,j}$  has a relatively sharp variation at its borders compared to the size of its support. Since  $\eta$  is not proportional to  $2^{-j}$ , the energy concentration of  $\hat{g}_{p,j}$  is not improved when the window size  $2^{-j}$  increases. Even though  $f$  may be very smooth over  $[a_{p,j}, a_{p+1,j}]$ , the border variations of the window create relatively large coefficients up to a frequency of the order of  $\pi/\eta$ .

To reduce the number of large coefficients we must increase  $\eta$ , but this also increases the minimum window size in the tree, which is  $2^{-J} = 2\eta$ . The choice of  $\eta$  is therefore the result of a trade-off between window regularity and the maximum resolution of the time subdivision. There is no equivalent limitation in the construction of wavelet packet bases.

### 8.5.2 Tree of Discrete Bases

For discrete signals of size  $N$ , a binary tree of discrete cosine bases is constructed like a binary tree of continuous-time cosine bases. To simplify notations, the sampling distance is normalized to 1. If it is equal to  $N^{-1}$ , then frequency parameters must be multiplied by  $N$ .

The subdivision points are located at half integers:

$$a_{p,j} = p N 2^{-j} - 1/2 \quad \text{for } 0 \leq p \leq 2^j.$$

The discrete windows are obtained by sampling the windows  $g_p(t)$  defined in (8.114),  $g_{p,j}[n] = g_{p,j}(n)$ . The same border modification is used to ensure that the support of all  $g_{p,j}[n]$  is in  $[0, N-1]$ .

A node at depth  $j$  and position  $p$  in the binary tree corresponds to space  $\mathbf{W}_j^p$  generated by the discrete local cosine family

$$\mathcal{B}_j^p = \left\{ g_{p,j}[n] \sqrt{\frac{2}{2^{-j}N}} \cos \left[ \pi \left( k + \frac{1}{2} \right) \frac{n - a_{p,j}}{2^{-j}N} \right] \right\}_{0 \leq k < N 2^{-j}}.$$

Since  $g_{0,0} = \mathbf{1}_{[0,N-1]}$ , the space  $\mathbf{W}_0^0$  at the root of the tree includes any signal defined over  $0 \leq n < N$ , so  $\mathbf{W}_0^0 = \mathbb{C}^N$ . As in Theorem 8.18 we verify that  $\mathbf{W}_j^p$  is orthogonal to  $\mathbf{W}_j^q$  for  $p \neq q$  and that

$$\mathbf{W}_j^p = \mathbf{W}_{j+1}^{2p} \oplus \mathbf{W}_{j+1}^{2p+1}. \quad (8.119)$$

The splitting property (8.119) implies that the union of local cosine families  $\mathcal{B}_j^p$  located at the leaves of an admissible subtree is an orthogonal basis of  $\mathbf{W}_0^0 = \mathbb{C}^N$ . The minimum window size is limited by  $2\eta \leq 2^{-j}N$ , so the maximum depth of this binary tree is  $J = \log_2 \frac{N}{2\eta}$ . Thus, one can construct more than  $2^{2^{J-1}} = 2^{N/(4\eta)}$  different discrete local cosine bases within this binary tree.

#### Fast Calculations

The fast local cosine transform algorithm described in Section 8.4.4 requires  $O(2^{-j}N \log_2(2^{-j}N))$  operations to compute the inner products of  $f$  with the  $2^{-j}N$  vectors in the local cosine family  $\mathcal{B}_j^p$ . The total number of operations to perform these computations at all nodes  $(j, p)$  of the tree, for  $0 \leq p < 2^j$  and  $0 \leq j \leq J$ , is therefore  $O(NJ \log_2 N)$ . The local cosine decompositions in Figure 8.19 are calculated with this fast algorithm. To improve the right border treatment, Section 8.4.4 explains that the last DCT-IV should be replaced by a DCTI at each scale  $2^j$ . The signal  $f$  is recovered from the local cosine coefficients at the leaves of any admissible binary tree with the fast local cosine reconstruction algorithm, which needs  $O(N \log_2 N)$  operations.

### 8.5.3 Image Cosine Quad-Tree

A local cosine binary tree is extended in two dimensions into a quad-tree, which recursively divides square image windows into four smaller windows. This separable

approach is similar to the extension of wavelet packet bases in two dimensions, described in Section 8.2.

Let us consider square images of  $N$  pixels. A node of the quad-tree is labeled by its depth  $j$  and two indices  $p$  and  $q$ . Let  $g_{p,j}[n]$  be the discrete one-dimensional window defined in Section 8.5.2. At depth  $j$ , a node  $(p, q)$  corresponds to a separable space

$$\mathbf{W}_j^{p,q} = \mathbf{W}_j^p \otimes \mathbf{W}_j^q, \quad (8.120)$$

which is generated by a separable local cosine basis of  $2^{-2j}N$  vectors

$$\mathcal{B}_j^{p,q} = \left\{ g_{p,j}[n_1] g_{q,j}[n_2] \frac{2}{2^{-j}N^{1/2}} \cos \left[ \pi \left( k_1 + \frac{1}{2} \right) \frac{n_1 - a_{p,j}}{2^{-j}N^{1/2}} \right] \cos \left[ \pi \left( k_2 + \frac{1}{2} \right) \frac{n_2 - a_{q,j}}{2^{-j}N^{1/2}} \right] \right\}_{0 \leq k_1, k_2 < 2^{-j}N^{1/2}}$$

We know from (8.119) that

$$\mathbf{W}_j^p = \mathbf{W}_{j+1}^{2p} \oplus \mathbf{W}_{j+1}^{2p+1} \quad \text{and} \quad \mathbf{W}_j^q = \mathbf{W}_{j+1}^{2q} \oplus \mathbf{W}_{j+1}^{2q+1}.$$

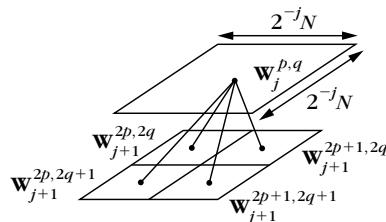
Inserting these equations in (8.120) proves that  $\mathbf{W}_j^{p,q}$  is the direct sum of four orthogonal subspaces:

$$\mathbf{W}_j^{p,q} = \mathbf{W}_{j+1}^{2p,2q} \oplus \mathbf{W}_{j+1}^{2p+1,2q} \oplus \mathbf{W}_{j+1}^{2p,2q+1} \oplus \mathbf{W}_{j+1}^{2p+1,2q+1}. \quad (8.121)$$

Space  $\mathbf{W}_j^{p,q}$  at node  $(j, p, q)$  is therefore decomposed in the four subspaces located at the four children nodes of the quad-tree. This decomposition can also be interpreted as a division of the square window  $g_{p,j}[n_1] g_{q,j}[n_2]$  into four subwindows of equal sizes, as illustrated in Figure 8.21. The space located at the root of the tree is

$$\mathbf{W}_0^{0,0} = \mathbf{W}_0^0 \otimes \mathbf{W}_0^0. \quad (8.122)$$

It includes all images of  $N$  pixels. Size  $\eta$  of the raising and decaying profiles of the one-dimensional windows defines the maximum depth  $J = \log_2 \frac{N^{1/2}}{2\eta}$  of the quad-tree.



**FIGURE 8.21**

Functions in  $\mathbf{W}_j^{p,q}$  have a support located in a square region of the image. It is divided into four subspaces that cover smaller squares in the image.

**FIGURE 8.22**

The grid shows the support of the windows  $g_{j,p}[n_1]g_{j,q}[n_2]$  of a “best” local cosine basis selected in the local cosine quad-tree.

### **Admissible Quad-Trees**

An admissible subtree of this local cosine quad-tree has nodes that have either zero or four children. Applying the decomposition property (8.121) along the branches of an admissible quad-tree proves that spaces  $\mathbf{W}_{j_i}^{p_i, q_i}$  located at the leaves decompose  $\mathbf{W}_0^{0,0}$  in orthogonal subspaces. The union of the corresponding two-dimensional local cosine bases  $\mathcal{B}_{j_i}^{p_i, q_i}$  is therefore an orthogonal basis of  $\mathbf{W}_0^{0,0}$ . We proved in (8.42) that there are more than  $2^{4^{J-1}} = 2^{N/16\eta^2}$  different admissible trees of maximum depth  $J = \log_2 \frac{N^{1/2}}{2\eta}$ . These bases divide the image plane into squares of varying sizes. Figure 8.22 gives an example of image decomposition in a local cosine basis corresponding to an admissible quad-tree. This local cosine basis is selected with the best basis algorithm of Section 12.2.2.

### **Fast Calculations**

The decomposition of an image  $f[n]$  over a separable local cosine family  $\mathcal{B}_j^{p,q}$  requires  $O(2^{-2j}N \log_2(2^{-j}N))$  operations with a separable implementation of the fast one-dimensional local cosine transform. For a full local cosine quad-tree of depth  $J$ , these calculations are performed for  $0 \leq p, q < 2^j$  and  $0 \leq j \leq J$ , which requires  $O(NJ \log_2 N)$  multiplications and additions. The original image is recovered from the local cosine coefficients at the leaves of any admissible subtree with  $O(N \log_2 N)$  computations.

---

## 8.6 EXERCISES

- 8.1** <sup>2</sup> Prove the discrete splitting Theorem 8.6.
- 8.2** <sup>2</sup> Meyer wavelet packets are calculated with a Meyer conjugate mirror filter (7.84). Compute the size of the frequency support of  $\hat{\psi}_j^p$  as a function of  $2^j$ . Study the convergence of  $\psi_{j,n}(t)$  when the scale  $2^j$  goes to  $+\infty$ .
- 8.3** <sup>1</sup> Extend the separable wavelet packet tree of Section 8.2.2 for discrete  $p$ -dimensional signals. Verify that the wavelet packet tree of a  $p$ -dimensional discrete signal of  $N$  samples includes  $O(N \log_2 N)$  wavelet packet coefficients that are calculated with  $O(K N \log_2 N)$  operations if the conjugate mirror filter  $h$  has  $K$  nonzero coefficients.
- 8.4** <sup>2</sup> Anisotropic wavelet packets  $\psi_j^p[a - 2^{L-j}n_1] \psi_l^q[b - 2^{L-l}n_2]$  may have different scales  $2^j$  and  $2^l$  along the rows and columns. A decomposition over such wavelet packets is calculated with a filter bank that filters and subsamples the image rows  $j-L$  times, whereas the columns are filtered and subsampled  $l-L$  times. For an image  $f[n]$  of  $N$  pixels, show that a dictionary of anisotropic wavelet packets includes  $O(N(\log_2 N)^2)$  different vectors. Compute the number of operations needed to decompose  $f$  in this dictionary.
- 8.5** <sup>2</sup> *Hartley transform.* Let  $\text{cas}(t) = \cos(t) + \sin(t)$ . We define

$$\mathcal{B} = \left\{ g_k[n] = \frac{1}{\sqrt{N}} \text{cas}\left(\frac{2\pi nk}{N}\right) \right\}_{0 \leq k < N}.$$

- (a) Prove that  $\mathcal{B}$  is an orthonormal basis of  $\mathbb{C}^N$ .
- (b) For any signal  $f[n]$  of size  $N$ , find a fast Hartley transform algorithm based on the FFT, which computes  $\{\langle f, g_k \rangle\}_{0 \leq k < N}$  with  $O(N \log_2 N)$  operations.
- 8.6** <sup>2</sup> Prove that  $\{\sqrt{2} \sin[(k+1/2)\pi t]\}_{k \in \mathbb{Z}}$  is an orthonormal basis of  $L^2[0, 1]$ . Find a corresponding discrete orthonormal basis of  $\mathbb{C}^N$ .
- 8.7** <sup>2</sup> Prove that  $\{\sqrt{2} \sin(k\pi t)\}_{k \in \mathbb{Z}}$  is an orthonormal basis of  $L^2[0, 1]$ . Find a corresponding discrete orthonormal basis of  $\mathbb{C}^N$ .
- 8.8** <sup>3</sup> *Lapped Fourier basis:*
- (a) Construct a lapped orthogonal basis  $\{\tilde{g}_{p,k}\}_{(p,k) \in \mathbb{Z}}$  of  $L^2(\mathbb{R})$  from the Fourier basis  $\{\exp(i2\pi kt)\}_{k \in \mathbb{Z}}$  of  $L^2[0, 1]$ .
- (b) Explain why this local Fourier basis does not contradict the Balian-Low Theorem 5.20.

- (c) Let  $f \in L^2(\mathbb{R})$  be such that  $|\hat{f}(\omega)| = O((1 + |\omega|^p)^{-1})$  for some  $p > 0$ . Compute the rate of decay of  $|\langle f, \tilde{g}_{p,k} \rangle|$  when the frequency index  $|k|$  increases. Compare it with the rate of decay of  $|\langle f, g_{p,k} \rangle|$ , where  $g_{p,k}$  is a local cosine vector (8.104). How do the two bases compare for signal-processing applications?
- 8.9** <sup>3</sup> Describe a fast algorithm to compute the Meyer orthogonal wavelet transform with a lapped transform applied in the Fourier domain. Calculate the numerical complexity of this algorithm for periodic signals of size  $N$ . Compare this result with the numerical complexity of the standard fast wavelet transform algorithm, where the convolutions with Meyer conjugate mirror filters are calculated with an FFT.
- 8.10** <sup>3</sup> *Mirror wavelets.* Let  $(h, g)$  be a pair of conjugate mirror filters. A mirror wavelet packet first decomposes the input signal  $a_0$  into  $a_1$  and  $d_1$  by filtering it with  $\bar{h}[n] = h[-n]$  and  $\bar{g}[n] = g[-n]$ , respectively, and subsampling the output. The signals  $a_1$  and  $d_1$  are subdecomposed with an orthogonal wavelet transform filter bank tree. Decomposing  $a_1$  and  $d_1$  with a cascade of  $j - 2$  filtering with  $\bar{h}$  and a filtering with  $\bar{g}$  yields wavelet coefficients  $\langle a_0, \psi_{j,m} \rangle$  and wavelet packet coefficients that we write as  $\langle a_0, \tilde{\psi}_{j,m} \rangle$ . Prove that the discrete Fourier transform of these *mirror wavelets* satisfies  $|\tilde{\psi}_{j,m}[k]| = |\widehat{\psi}_{j,m}[N/2 - k]|$ . Show that if the filters  $\bar{h}$  and  $\bar{g}$  that decompose  $d_1$  are replaced by  $h$  and  $g$ , then the resulting mirror wavelets satisfy  $\tilde{\psi}_{j,m}[n] = (-1)^{n-1} \psi_{j,m}[1-n]$ .
- 8.11** <sup>2</sup> *Arbitrary Walsh tilings:*
- Prove that two Walsh wavelet packets  $\psi_{j,n}^p$  and  $\psi_{j',n'}^{p'}$  are orthogonal if their Heisenberg boxes, defined in Section 8.1.2, do not intersect in the time-frequency plane [71].
  - A dyadic tiling of the time-frequency plane is an exact cover  $\{[2^j n, 2^j(n+1)] \times [k\pi 2^{-j}, (k+1)\pi 2^{-j}]\}_{(j,n,k) \in I}$ , where the index set  $I$  is adjusted to guarantee that the time-frequency boxes do not intersect and that they leave no hole. Prove that any such tiling corresponds to a Walsh orthonormal basis of  $L^2(\mathbb{R})$   $\{\psi_{j,n}^p\}_{(p,j,n) \in I}$ .
- 8.12** <sup>3</sup> *Double tree.* We want to construct a dictionary of block wavelet packet bases, which has the freedom to segment both the time and frequency axes. For this purpose, as in a local cosine basis dictionary, we construct a binary tree, which divides  $[0, 1]$  in  $2^j$  intervals  $[p2^{-j}, (p+1)2^{-j}]$  that correspond to nodes indexed by  $p$  at the depth  $j$  of the tree. At each of these nodes, we construct another tree of wavelet packet orthonormal bases of  $L^2[p2^{-j}, (p+1)2^{-j}]$  [299].
- Define admissible subtrees in this double tree, with leaves corresponding to orthonormal bases of  $L^2[0, 1]$ . Give an example of an admissible tree and draw the resulting tiling of the time-frequency plane.

- (b) Give a recursive equation that relates the number of admissible subtrees of depth  $J + 1$  and of depth  $J$ . Give an upper bound and a lower bound for the total number of orthogonal bases in this double-tree dictionary.
- (c) Can one find a basis in a double tree that is well adapted to implement an efficient transform code for audio signals? Justify your answer.
- 8.13** <sup>3</sup> An anisotropic local cosine basis for images is constructed with rectangular windows that have a width  $2^j$  that may be different from their height  $2^l$ . Similar to a local cosine tree, such bases are calculated by progressively dividing windows, but the horizontal and vertical divisions of these windows are done independently. Show that a dictionary of anisotropic local cosine bases can be represented as a graph. Implement numerically an algorithm that decomposes images in a graph of anisotropic local cosine bases.

# Approximations in Bases

# 9

It is time to wonder why we are constructing so many different orthonormal bases. In signal processing, orthogonal bases are of interest because they can provide sparse representations of certain types of signals with few vectors. Compression and denoising are applications studied in Chapters 10 and 11.

Approximation theory studies the error produced by different approximation schemes. Classic sampling theorems are linear approximations that project the analog signal over low-frequency vectors chosen a priori in a basis. The discrete signal representation may be further reduced with a linear projection over the first few vectors of an orthonormal basis. However, better nonlinear approximations are obtained by choosing the approximation vectors depending on the signal. In a wavelet basis, these nonlinear approximations locally adjust the approximation resolution to the signal regularity.

Approximation errors depend on the signal regularity. For uniformly regular signals, linear and nonlinear approximations perform similarly, whether in a wavelet or in a Fourier basis. When the signal regularity is not uniform, nonlinear approximations in a wavelet basis can considerably reduce the error of linear approximations. This is the case for piecewise regular signals or bounded variation signals and images. Geometric approximations of piecewise regular images with regular edge curves are studied with adaptive triangulations and curvelets.

---

## 9.1 LINEAR APPROXIMATIONS

Analog signals are discretized in Section 3.1.3 with inner products in a basis. In the following sections we compute the resulting linear approximation error in wavelet and Fourier bases, which depends on the uniform signal regularity. For signals modeled as realizations of a random vector, in Section 9.1.4 we prove that the optimal basis is the Karhunen-Loëve basis (principal components), which diagonalizes the covariance matrix.

### 9.1.1 Sampling and Approximation Error

Approximation errors of linear sampling processes are related to the error of linear approximations in an orthogonal basis. These errors are computed from the decay of signal coefficients in this basis.

An analog signal  $f(t)$  is discretized with a low-pass filter  $\bar{\phi}_s(t)$  and a uniform sampling interval  $s$ :

$$f \star \bar{\phi}_s(ns) = \int_{-\infty}^{+\infty} f(u) \bar{\phi}_s(ns - u) du = \langle f(t), \phi_s(t - ns) \rangle, \quad (9.1)$$

with  $\bar{\phi}_s(t) = \phi_s(-t)$ . Let us consider an analog signal of compact support, normalized to  $[0, 1]$ . At a resolution  $N$  corresponding to  $s = N^{-1}$ , the discretization is performed over  $N$  functions  $\{\phi_n(t) = \phi_s(t - ns)\}_{0 \leq n < N}$  that are modified at the boundaries to maintain their support in  $[0, 1]$ . They define a Riesz basis of an approximation space  $\mathbf{U}_N \subset \mathbf{L}^2[0, 1]$ . The best linear approximation of  $f$  in  $\mathbf{U}_N$  is the orthogonal projection  $f_N$  of  $f$  in  $\mathbf{U}_N$ , recovered with the biorthogonal basis  $\{\tilde{\phi}_n(t)\}_{1 \leq n \leq N}$ :

$$f_N(t) = \sum_{n=0}^{N-1} \langle f, \phi_n \rangle \tilde{\phi}_n(t). \quad (9.2)$$

To compute the approximation error  $\|f - f_N\|$ , we introduce an orthonormal basis  $\mathcal{B} = \{g_m\}_{m \in \mathbb{N}}$  of  $\mathbf{L}^2[0, 1]$ , with  $N$  vectors  $\{g_m\}_{0 \leq m < N}$  defining an orthogonal basis of the same approximation space  $\mathbf{U}_N$ . Fourier and wavelet bases provide such bases for many classic approximation spaces. The orthogonal projection  $f_N$  of  $f$  in  $\mathbf{U}_N$  can be decomposed on the first  $N$  vectors of this basis:

$$f_N = \sum_{m=0}^{N-1} \langle f, g_m \rangle g_m.$$

Since  $\mathcal{B}$  is an orthonormal basis of  $\mathbf{L}^2[0, 1]$ ,  $f = \sum_{m=0}^{+\infty} \langle f, g_m \rangle g_m$ , so

$$f - f_N = \sum_{m=N}^{+\infty} \langle f, g_m \rangle g_m,$$

and the resulting approximation error is

$$\varepsilon_l(N, f) = \|f - f_N\|^2 = \sum_{m=N}^{+\infty} |\langle f, g_m \rangle|^2. \quad (9.3)$$

The fact that  $\|f\|^2 = \sum_{m=0}^{+\infty} |\langle f, g_m \rangle|^2 < +\infty$  implies that the error decays to zero:

$$\lim_{N \rightarrow +\infty} \varepsilon_l(N, f) = 0.$$

However, the decay rate of  $\varepsilon_l(N, f)$  as  $N$  increases depends on the decay of  $|\langle f, g_m \rangle|$  as  $m$  increases. Theorem 9.1 gives equivalent conditions on the decay of  $\varepsilon_l(N, f)$  and  $|\langle f, g_m \rangle|$ .

**Theorem 9.1.** For any  $s > 1/2$ , there exists  $A, B > 0$  such that if  $\sum_{m=0}^{+\infty} |m|^{2s} |\langle f, g_m \rangle|^2 < +\infty$ , then

$$A \sum_{m=0}^{+\infty} m^{2s} |\langle f, g_m \rangle|^2 \leq \sum_{N=0}^{+\infty} N^{2s-1} \varepsilon_l(N, f) \leq B \sum_{m=0}^{+\infty} m^{2s} |\langle f, g_m \rangle|^2, \quad (9.4)$$

and thus  $\varepsilon_l(N, f) = o(N^{-2s})$ .

**Proof.** By inserting (9.3), we compute

$$\sum_{N=0}^{+\infty} N^{2s-1} \varepsilon_l(N, f) = \sum_{N=0}^{+\infty} \sum_{m=N}^{+\infty} N^{2s-1} |\langle f, g_m \rangle|^2 = \sum_{m=0}^{+\infty} |\langle f, g_m \rangle|^2 \sum_{N=0}^m N^{2s-1}.$$

For any  $s > 1/2$ ,

$$\int_0^m x^{2s-1} dx \leq \sum_{N=0}^m N^{2s-1} \leq \int_1^{m+1} x^{2s-1} dx,$$

which implies that  $\sum_{N=0}^m N^{2s-1} \sim m^{2s}$  and thus proves (9.4).

To verify that  $\varepsilon_l(N, f) = o(N^{-2s})$ , observe that  $\varepsilon_l(m, f) \geq \varepsilon_l(N, f)$  for  $m \leq N$ , so

$$\varepsilon_l(N, f) \sum_{m=N/2}^{N-1} m^{2s-1} \leq \sum_{m=N/2}^{N-1} m^{2s-1} \varepsilon_l(m, f) \leq \sum_{m=N/2}^{+\infty} m^{2s-1} \varepsilon_l(m, f). \quad (9.5)$$

Since  $\sum_{m=1}^{+\infty} m^{2s-1} \varepsilon_l(m, f) < +\infty$ , it follows that

$$\lim_{N \rightarrow +\infty} \sum_{m=N/2}^{+\infty} m^{2s-1} \varepsilon_l(m, f) = 0.$$

Moreover, there exists  $C > 0$  such that  $\sum_{m=N/2}^{N-1} m^{2s-1} \geq C N^{2s}$ , so (9.5) implies that  $\lim_{N \rightarrow +\infty} \varepsilon_l(N, f) N^{2s} = 0$ . ■

This theorem proves that the linear approximation error of  $f$  in basis  $\mathcal{B}$  decays faster than  $N^{-2s}$  if  $f$  belongs to the space

$$\mathbf{W}_{\mathcal{B},s} = \left\{ f \in \mathbf{H} : \sum_{m=0}^{+\infty} m^{2s} |\langle f, g_m \rangle|^2 < +\infty \right\}.$$

One can also prove that this linear approximation is asymptotically optimal over this space [20]. Indeed, there exists nonlinear or nonlinear approximation scheme with error decays that are at least like  $N^{-\alpha}$ , with  $\alpha > 2s$ , for all  $f \in \mathbf{W}_{\mathcal{B},s}$ .

In the next sections we prove that if  $\mathcal{B}$  is a Fourier or wavelet basis, then  $\mathbf{W}_{\mathcal{B},s}$  is a Sobolev space, and therefore that linear approximations of Sobolev functions are optimal in Fourier and wavelet bases. However, we shall also see that for more complex functions, the linear approximation of  $f$  from the first  $N$  vectors of  $\mathcal{B}$  is not always precise because these vectors are not necessarily the best ones to approximate  $f$ . Nonlinear approximations calculated with vectors chosen adaptively depending on  $f$  are studied in Section 9.2.

### 9.1.2 Linear Fourier Approximations

The Shannon-Whittaker sampling theorem performs a perfect low-pass filter that keeps the signal low frequencies. Thus, it is equivalent to a linear approximation over the lower frequencies of a Fourier basis. Linear Fourier approximations are asymptotically optimal for uniformly regular signals. The approximation error is related to the Sobolev differentiability. It is also calculated for nonuniform regular signals, such as discontinuous signals having a bounded total variation.

Theorem 3.6 proves (modulo a change of variable) that  $\{e^{i2\pi mt}\}_{m \in \mathbb{Z}}$  is an orthonormal basis of  $L^2[0, 1]$ . Thus, we can decompose  $f \in L^2[0, 1]$  in the Fourier series

$$f(t) = \sum_{m=-\infty}^{+\infty} \langle f(u), e^{i2\pi mu} \rangle e^{i2\pi mt}, \quad (9.6)$$

with

$$\langle f(u), e^{i2\pi mu} \rangle = \int_0^1 f(u) e^{-i2\pi mu} du.$$

The decomposition (9.6) defines a periodic extension of  $f$  for all  $t \in \mathbb{R}$ . The decay of the Fourier coefficients  $|\langle f(u), e^{i2\pi mu} \rangle|$  as  $m$  increases depends on the regularity of this periodic extension.

The linear approximation of  $f \in L^2[0, 1]$  by the  $N$  sinusoids of lower frequencies is obtained by a linear filtering that sets all higher frequencies to zero:

$$f_N(t) = \sum_{|m| \leq N/2} \langle f(u), e^{i2\pi mu} \rangle e^{i2\pi mt}.$$

It projects  $f$  in space  $U_N$  of functions having Fourier coefficients that are zero above the frequency  $N\pi$ .

#### Error Decay versus Sobolev Differentiability

The decay of the linear Fourier approximation error depends on the Sobolev differentiability. The regularity of  $f$  can be measured by the number of times it is differentiable. Sobolev differentiability extends derivatives to nonintegers with a Fourier decay condition. To avoid boundary issues, we first consider functions  $f(t)$  defined for all  $t \in \mathbb{R}$ .

Recall that the Fourier transform of the derivative  $f'(t)$  is  $i\omega \hat{f}(\omega)$ . The Plancherel formula proves that  $f' \in L^2(\mathbb{R})$  if

$$\int_{-\infty}^{+\infty} |\omega|^2 |\hat{f}(\omega)|^2 d\omega = 2\pi \int_{-\infty}^{+\infty} |f'(t)|^2 dt < +\infty.$$

This suggests replacing the usual pointwise definition of the derivative by a definition based on the Fourier transform. We say that  $f \in L^2(\mathbb{R})$  is differentiable in the *sense of Sobolev* if

$$\int_{-\infty}^{+\infty} |\omega|^2 |\hat{f}(\omega)|^2 d\omega < +\infty. \quad (9.7)$$

This integral imposes that  $|\hat{f}(\omega)|$  has a sufficiently fast decay when the frequency  $\omega$  goes to  $+\infty$ . As in Section 2.3.1, the regularity of  $f$  is measured from the asymptotic decay of its Fourier transform.

This definition is generalized for any  $s > 0$ . Space  $\mathbf{W}^s(\mathbb{R})$  of  $s$  times differentiable Sobolev functions is the space of functions  $f \in \mathbf{L}^2(\mathbb{R})$  having a Fourier transform that satisfies [67]

$$\int_{-\infty}^{+\infty} |\omega|^{2s} |\hat{f}(\omega)|^2 d\omega < +\infty. \quad (9.8)$$

If  $s > n + 1/2$ , then one can verify (Exercise 9.7) that  $f$  is  $n$  times continuously differentiable.

Let  $\mathbf{W}^s[0, 1]$  be the space of functions in  $\mathbf{L}^2[0, 1]$  that can be extended outside  $[0, 1]$  into a function  $f \in \mathbf{W}^s(\mathbb{R})$ . To avoid border problems at  $t = 0$  or at  $t = 1$ , let us consider functions  $f$  that have supports that are strictly included in  $(0, 1)$ . A simple regular extension on  $\mathbb{R}$  is obtained by setting its value to 0 outside  $[0, 1]$ , and  $f \in \mathbf{W}^s[0, 1]$  if this extension is in  $\mathbf{W}^s(\mathbb{R})$ . In this case, one can prove (not trivial) that the Sobolev integral condition (9.8) reduces to a discrete sum, meaning that  $f \in \mathbf{W}^s[0, 1]$  if and only if

$$\sum_{m=-\infty}^{+\infty} |m|^{2s} |\langle f(u), e^{i2\pi mu} \rangle|^2 < +\infty. \quad (9.9)$$

For such differentiable functions in the sense of Sobolev, Theorem 9.2 computes the approximation error

$$\varepsilon_l(N, f) = \|f - f_N\|^2 = \int_0^1 |f(t) - f_N(t)|^2 dt = \sum_{|m|>N/2} |\langle f(u), e^{i2\pi mu} \rangle|^2. \quad (9.10)$$

**Theorem 9.2.** Let  $f \in \mathbf{L}^2[0, 1]$  be a function with support strictly included in  $(0, 1)$ . Then  $f \in \mathbf{W}^s[0, 1]$  if and only if

$$\sum_{N=1}^{+\infty} N^{2s} \frac{\varepsilon_l(N, f)}{N} < +\infty, \quad (9.11)$$

which implies  $\varepsilon_l(N, f) = o(N^{-2s})$ . ■

The proof relies on the fact that functions in  $\mathbf{W}^s[0, 1]$  with a support in  $(0, 1)$  are characterized by (9.9). Therefore, this theorem is a consequence of Theorem 9.1. Thus, the linear Fourier approximation decays quickly if and only if  $f$  has a large regularity exponent  $s$  in the sense of Sobolev.

### Discontinuities and Bounded Variation

If  $f$  is discontinuous, then  $f \notin \mathbf{W}^s[0, 1]$  for any  $s > 1/2$ . Thus, Theorem 9.2 proves that  $\varepsilon_l(N, f)$  can decay like  $N^{-\alpha}$  only if  $\alpha \leq 1$ . For bounded variation functions,

which are introduced in Section 2.3.3, Theorem 9.3 proves that  $\varepsilon_l(N, f) = O(N^{-1})$ . A function has a bounded variation if

$$\|f\|_V = \int_0^1 |f'(t)| dt < +\infty.$$

The derivative must be taken in the sense of distributions because  $f$  may be discontinuous. If  $f = \mathbf{1}_{[0,1/2]}$ , then  $\|f\|_V = 2$ . Recall that  $a[N] \sim b[N]$  if  $a[N] = O(b[N])$  and  $b[N] = O(a[N])$ .

**Theorem 9.3.**

- If  $\|f\|_V < +\infty$ , then  $\varepsilon_l(N, f) = O(\|f\|_V^2 N^{-1})$ .
- If  $f = C \mathbf{1}_{[0,1/2]}$ , then  $\varepsilon_l(N, f) \sim \|f\|_V^2 N^{-1}$ .

**Proof.** If  $\|f\|_V < +\infty$ , then

$$\begin{aligned} |\langle f(u), \exp(i2m\pi u) \rangle| &= \left| \int_0^1 f(u) \exp(-i2m\pi u) du \right| \\ &= \left| \int_0^1 f'(u) \frac{\exp(-i2m\pi u)}{-i2m\pi} dt \right| \leq \frac{\|f\|_V}{2|m|\pi}. \end{aligned}$$

Thus,

$$\varepsilon_l(N, f) = \sum_{|m|>N/2} |\langle f(u), \exp(i2m\pi u) \rangle|^2 \leq \frac{\|f\|_V^2}{4\pi^2} \sum_{|m|>N/2} \frac{1}{m^2} = O(\|f\|_V^2 N^{-1}).$$

If  $f = C \mathbf{1}_{[0,1/2]}$ , then  $\|f\|_V = 2C$  and

$$|\langle f(u), \exp(i2m\pi u) \rangle| = \begin{cases} 0 & \text{if } m \neq 0 \text{ is even} \\ C/(\pi|m|) & \text{if } m \text{ is odd,} \end{cases}$$

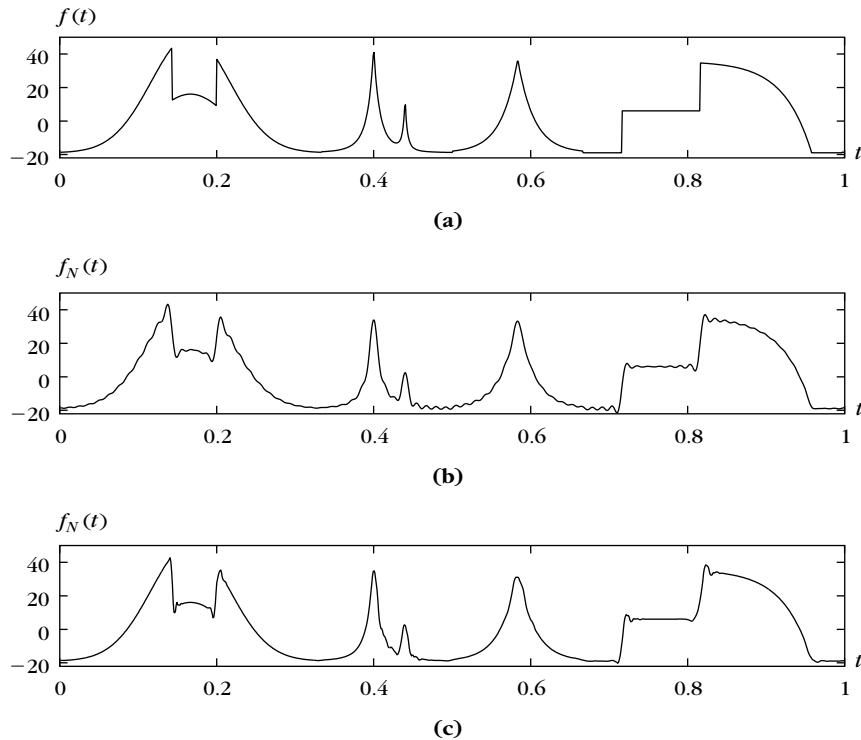
so  $\varepsilon_l(N, f) \sim C^2 N^{-1}$ . ■

This theorem shows that when  $f$  is discontinuous with bounded variations, then  $\varepsilon_l(N, f)$  decays typically like  $N^{-1}$ . Figure 9.1(b) shows a bounded variation signal approximated by Fourier coefficients of lower frequencies. The approximation error is concentrated in the neighborhood of discontinuities where the removal of high frequencies creates Gibbs oscillations (see Section 2.3.1).

### Localized Approximations

To localize Fourier series approximations over intervals, we multiply  $f$  by smooth windows that cover each of these intervals. The Balian-Low theorem (5.20) proves that one cannot build local Fourier bases with smooth windows of compact support. However, in Section 8.4.2 we construct orthonormal bases by replacing complex exponentials by cosine functions. For appropriate windows  $g_p$  of compact support  $[a_p - \eta_p, a_{p+1} + \eta_{p+1}]$ , Corollary 8.1 constructs an orthonormal basis of  $\mathbf{L}^2(\mathbb{R})$ :

$$\left\{ g_{p,k}(t) = g_p(t) \sqrt{\frac{2}{l_p}} \cos \left[ \pi \left( k + \frac{1}{2} \right) \frac{t - a_p}{l_p} \right] \right\}_{k \in \mathbb{N}, p \in \mathbb{Z}}.$$

**FIGURE 9.1**

**(a)** Original signal  $f$ . **(b)** Signal  $f_N$  approximated from  $N = 128$  lower-frequency Fourier coefficients with  $\|f - f_N\|/\|f\| = 8.63 \cdot 10^{-2}$ . **(c)** Signal  $f_N$  approximated from larger-scale Daubechies 4 wavelet coefficients with  $N = 128$  and  $\|f - f_N\|/\|f\| = 8.58 \cdot 10^{-2}$ .

Writing  $f$  in this local cosine basis is equivalent to segmenting it into several windowed components  $f_p(t) = f(t)g_p(t)$ , which are decomposed in a cosine IV basis. If  $g_p$  is  $C^\infty$ , the regularity of  $g_p(t)f(t)$  is the same as the regularity of  $f$  over  $[a_p - \eta_p, a_{p+1} + \eta_{p+1}]$ . Section 8.3.2 relates cosine IV coefficients to Fourier series coefficients. It follows from Theorem 9.2 that if  $f_p \in W^s(\mathbb{R})$ , then the approximation

$$f_{p,N} = \sum_{k=0}^{N-1} \langle f, g_{p,k} \rangle g_{p,k}$$

yields an error of

$$\varepsilon_l(N, f_p) = \|f_p - f_{p,N}\|^2 = o(N^{-2s}).$$

Thus, the approximation error in a local cosine basis depends on the local regularity of  $f$  over each window support.

### 9.1.3 Multiresolution Approximation Errors with Wavelets

Wavelets are constructed as bases of orthogonal complements of multiresolution approximation spaces. Thus, the projection error on multiresolution approximation spaces depends on the decay of wavelet coefficients. Linear approximations with wavelets behave essentially like Fourier approximations but with a better treatment of boundaries. They are also asymptotically optimal for uniformly regular signals. The linear error decay is computed for Sobolev differentiable functions and for uniformly Lipschitz  $\alpha$  functions.

#### Uniform Approximation Grid

Section 7.5 constructs multiresolution approximation spaces  $\mathbf{U}_N = \mathbf{V}_L$  of  $\mathbf{L}^2[0, 1]$  with their orthonormal basis of  $N = 2^{-L}$  scaling functions  $\{\phi_{L,n}(t)\}_{0 \leq n < 2^{-L}}$ . These scaling functions  $\phi_{L,n}(t) = \phi_L(t - 2^L n)$  with  $\phi_L(t) = 2^{-L/2}\phi(2^{-L}t)$  are finite elements translated over a uniform grid, modified near 0 and 1 so that their support remains in  $[0, 1]$ . The resulting projection of  $f$  in such a space is

$$f_N = P_{\mathbf{V}_L} f = \sum_{n=0}^{2^{-L}-1} \langle f, \phi_{L,n} \rangle \phi_{L,n}, \quad (9.12)$$

and

$$\langle f, \phi_{L,n} \rangle = \int f(t) \phi_L(t - 2^L n) dt = f \star \bar{\phi}_L(ns) \quad \text{with} \quad \bar{\phi}_L(t) = \phi_L(-t).$$

A different orthogonal basis of  $\mathbf{V}_L$  is obtained from wavelets at scales  $2^j > 2^L$  and scaling functions at a large scale  $2^J$ :

$$\left[ \{\phi_{J,n}\}_{0 \leq n < 2^{-J}}, \{\psi_{J,n}\}_{-J < j \leq J, 0 \leq n < 2^{-j}} \right]. \quad (9.13)$$

The approximation (9.12) can also be written as a wavelet approximation:

$$f_N = P_{\mathbf{V}_L} f = \sum_{j=L+1}^J \sum_{n=0}^{2^{-j}-1} \langle f, \psi_{j,n} \rangle \psi_{j,n} + \sum_{n=0}^{2^{-J}-1} \langle f, \phi_{J,n} \rangle \phi_{J,n}. \quad (9.14)$$

Since wavelets define an orthonormal basis of  $\mathbf{L}^2[0, 1]$ ,

$$\left[ \{\phi_{J,n}\}_{0 \leq n < 2^{-J}}, \{\psi_{J,n}\}_{-\infty < j \leq J, 0 \leq n < 2^{-j}} \right], \quad (9.15)$$

the approximation error is the energy of wavelet coefficients at scales smaller than  $2^L$ :

$$\varepsilon_l(N, f) = \|f - f_N\|^2 = \sum_{j=-\infty}^L \sum_{n=0}^{2^{-j}-1} |\langle f, \psi_{j,n} \rangle|^2. \quad (9.16)$$

In the following we suppose that wavelets  $\psi_{j,n}$  are  $\mathbf{C}^q$  ( $q$  times differentiable) and have  $q$  vanishing moments. The treatment of boundaries is the key difference

with Fourier approximations. Fourier approximations consider that the signal is periodic and if  $f(0) \neq f(1)$ , then the approximation of  $f$  behaves as if  $f$  was discontinuous. The periodic orthogonal wavelet bases in Section 7.5.1 do essentially the same. To improve this result, wavelets at the boundaries must keep their  $q$  vanishing moments, which requires us to modify them near the boundaries so that their supports remain in  $[0, 1]$ . Section 7.5.3 constructs such wavelet bases. They take advantage of the regularity of  $f$  over  $[0, 1]$  with no condition on  $f(0)$  and  $f(1)$ . For mathematical analysis, we only use these wavelets, without explicitly writing their shape modifications at the boundaries, to simplify notations. In numerical experiments, the folding boundary solution of Section 7.5.2 is more often used because it has a simpler algorithmic implementation. Folded wavelets have one vanishing moment at the boundary, which is often sufficient in applications.

### **Approximation Error versus Sobolev Regularity**

Like a Fourier basis, a wavelet basis provides an efficient approximation of uniformly regular signals. The decay of wavelet linear approximation errors is first related to the differentiability in the sense of Sobolev. Let  $\mathbf{W}^s[0, 1]$  be the Sobolev space of functions that are restrictions over  $[0, 1]$  of  $s$  times differentiable Sobolev function  $\mathbf{W}^s(\mathbb{R})$  defined over  $\mathbb{R}$ . If  $\psi$  has  $q$  vanishing moments, then (6.11) proves that the wavelet transform is a multiscale differential operator of order  $q$  at least. To test the differentiability of  $f$  up to order  $s$ , we need  $q > s$ . Theorem 9.4 gives a necessary and sufficient condition on wavelet coefficients so that  $f \in \mathbf{W}^s[0, 1]$ .

**Theorem 9.4.** Let  $0 < s < q$  be a Sobolev exponent. A function  $f \in \mathbf{L}^2[0, 1]$  is in  $\mathbf{W}^s[0, 1]$  if and only if

$$\sum_{j=-\infty}^J \sum_{n=0}^{2^{-j}-1} 2^{-2sj} |\langle f, \psi_{j,n} \rangle|^2 < +\infty. \quad (9.17)$$

**Proof.** We give an intuitive justification but not a proof of this result. To simplify, we suppose that the support of  $f$  is included in  $(0, 1)$ . If we extend  $f$  by zeros outside  $[0, 1]$ , then  $f \in \mathbf{W}^s(\mathbb{R})$ , which means that

$$\int_{-\infty}^{+\infty} |\omega|^{2s} |\hat{f}(\omega)|^2 d\omega < +\infty. \quad (9.18)$$

The low-frequency part of this integral always remains finite because  $f \in \mathbf{L}^2(\mathbb{R})$ :

$$\int_{|\omega| \leqslant 2^{-J}\pi} |\omega|^{2s} |\hat{f}(\omega)|^2 d\omega \leqslant 2^{-2sJ} \pi^{2s} \int_{|\omega| \leqslant \pi} |\hat{f}(\omega)|^2 d\omega \leqslant 2^{-2sJ} \pi^{2s} \|f\|^2.$$

The energy of  $\hat{\psi}_{j,n}$  is essentially concentrated in the intervals,  $[-2^{-j}2\pi, -2^{-j}\pi] \cup [2^{-j}\pi, 2^{-j}2\pi]$ . As a consequence,

$$\sum_{n=0}^{2^{-j}-1} |\langle f, \psi_{j,n} \rangle|^2 \sim \int_{2^{-j}\pi \leqslant |\omega| \leqslant 2^{-j+1}\pi} |\hat{f}(\omega)|^2 d\omega.$$

Over this interval  $|\omega| \sim 2^{-j}$ , so

$$\sum_{n=0}^{2^{-j}-1} 2^{-2sj} |\langle f, \psi_{j,n} \rangle|^2 \sim \int_{2^{-j}\pi \leq |\omega| \leq 2^{-j+1}\pi} |\omega|^{2s} |\hat{f}(\omega)|^2 d\omega.$$

It follows that

$$\sum_{j=-\infty}^J \sum_{n=0}^{2^{-j}-1} 2^{-2sj} |\langle f, \psi_{j,n} \rangle|^2 \sim \int_{|\omega| \geq 2^{-J}\pi} |\omega|^{2s} |\hat{f}(\omega)|^2 d\omega,$$

which explains why (9.18) is equivalent to (9.17). ■

This theorem proves that the Sobolev regularity of  $f$  is equivalent to a fast decay of the wavelet coefficients  $|\langle f, \psi_{j,n} \rangle|$  when scale  $2^j$  decreases. If  $\psi$  has  $q$  vanishing moments but is not  $q$  times continuously differentiable, then  $f \in \mathbf{W}^s[0, 1]$  implies (9.17), but the opposite implication is not true. Theorem 9.5 uses the decay condition (9.17) to characterize the linear approximation error with  $N$  wavelets.

**Theorem 9.5.** Let  $0 < s < q$  be a Sobolev exponent. A function  $f \in \mathbf{L}^2[0, 1]$  is in  $\mathbf{W}^s[0, 1]$  if and only if

$$\sum_{N=1}^{+\infty} N^{2s} \frac{\varepsilon_l(N, f)}{N} < +\infty, \quad (9.19)$$

which implies  $\varepsilon_l(N, f) = o(N^{-2s})$ .

**Proof.** Let us write the wavelets  $\psi_{j,n} = g_m$  with  $m = 2^{-j} + n$ . One can verify that the Sobolev condition (9.17) is equivalent to

$$\sum_{m=0}^{+\infty} |m|^{2s} |\langle f, g_m \rangle|^2 < +\infty.$$

The proof ends by applying Theorem 9.1. ■

If the wavelet has  $q$  vanishing moments but is not  $q$  times continuously differentiable, then  $f \in \mathbf{W}^s[0, 1]$  implies (9.19) but the opposite implication is false. Theorem 9.5 proves that  $f \in \mathbf{W}^s[0, 1]$  if and only if the approximation error  $\varepsilon_l(N, f)$  decays slightly faster than  $N^{-2s}$ . The wavelet approximation error is of the same order as the Fourier approximation error calculated in (9.11). However, Fourier approximations impose that the support of  $f$  is strictly included in  $[0, 1]$ , whereas wavelet approximation does not impose this condition or any other boundary condition because of the finer wavelet treatment of boundaries previously explained.

### Lipschitz Regularity

A different measure of uniform regularity is provided by Lipschitz exponents, which compute the error of a local polynomial approximation. A function  $f$  is uniformly

Lipschitz  $\alpha$  over  $[0, 1]$  if there exists  $K > 0$ , such that for any  $v \in [0, 1]$ , one can find a polynomial  $p_v$  of degree  $\lfloor \alpha \rfloor$  such that

$$\forall t \in [0, 1], \quad |f(t) - p_v(t)| \leq K |t - v|^\alpha. \quad (9.20)$$

The infimum of the  $K$  that satisfy (9.20) is the *homogeneous* Hölder  $\alpha$  norm  $\|f\|_{\tilde{C}^\alpha}$ . The Hölder  $\alpha$  norm of  $f$  also imposes that  $f$  is bounded:

$$\|f\|_{C^\alpha} = \|f\|_{\tilde{C}^\alpha} + \|f\|_\infty. \quad (9.21)$$

Space  $C^\alpha[0, 1]$  of functions  $f$  such that  $\|f\|_{C^\alpha} < +\infty$  is called a Hölder space. Theorem 9.6 characterizes the decay of wavelet coefficients.

**Theorem 9.6.** There exists  $B \geq A > 0$  such that

$$A \|f\|_{\tilde{C}^\alpha} \leq \sup_{j \geq J, 0 \leq n < 2^{-j}} 2^{-j(\alpha+1/2)} |\langle f, \psi_{j,n} \rangle| \leq B \|f\|_{\tilde{C}^\alpha}. \quad (9.22)$$

**Proof.** The proof of the equivalence between uniform Lipschitz regularity and the coefficient decay of a continuous wavelet transform is given in Theorem 6.3. This theorem gives a nearly equivalent result in the context of orthonormal wavelet coefficients that correspond to a sampling of a continuous wavelet transform computed with the same mother wavelet. Thus, the theorem proof is an adaptation of the proof of Theorem 6.3. This is illustrated by proving the right inequality of (9.22).

If  $f$  is uniformly Lipschitz  $\alpha$  on the support of  $\psi_{j,n}$ , since  $\psi_{j,n}$  is orthogonal the polynomial  $p_{2^j n}$ , approximating  $f$  at  $v = 2^j n$  yields

$$|\langle f, \psi_{j,n} \rangle| = |\langle f - p_{2^j n}, \psi_{j,n} \rangle| \leq \|f\|_{\tilde{C}^\alpha} \int 2^{-j/2} |\psi(2^{-j}(t - 2^j n))| |t - 2^j n|^\alpha dt. \quad (9.23)$$

With a change of variable, we get

$$|\langle f, \psi_{j,n} \rangle| \leq \|f\|_{\tilde{C}^\alpha} 2^{j(\alpha+1/2)} \int |\psi(t)| |t|^\alpha dt,$$

which proves the right inequality of (9.22). Observe that we do not use the wavelet regularity in this proof.

The left inequality of (9.22) is proved by following the continuous wavelet transform theorem (6.3) steps and replacing integrals by discrete sums over the position and scale of orthogonal wavelets. The regularity of wavelets plays an important role as shown by the proof of Theorem 6.3. In this case, there is no boundary issue because wavelets are adapted to the interval  $[0, 1]$  and keep their vanishing moments at the boundaries. ■

Similar to Theorem 9.4 for Sobolev differentiability, this theorem proves that uniform Lipschitz regularity is characterized by the decay of orthogonal wavelet coefficients when the scale  $2^j$  decreases. Hölder and Sobolev spaces belong to the larger class of Besov spaces, defined in Section 9.2.3.

If  $\psi$  has  $q$  vanishing moments but is not  $q$  times continuously differentiable, then the proof of Theorem 9.6 shows that the right inequality of (9.22) is valid.

Theorem 9.7 derives the decay of linear approximation errors for wavelets having  $q$  vanishing moments but that are not necessarily  $\mathbf{C}^q$ .

**Theorem 9.7.** If  $f$  is uniformly Lipschitz  $0 < \alpha \leq q$  over  $[0, 1]$ , then  $\varepsilon_l(N, f) = O(\|f\|_{\tilde{\mathcal{C}}^\alpha}^2 N^{-2\alpha})$ .

**Proof.** Theorem 9.6 proves that

$$|\langle f, \psi_{j,n} \rangle| \leq B \|f\|_{\tilde{\mathcal{C}}^\alpha} 2^{j(\alpha+1/2)}. \quad (9.24)$$

There are  $2^{-j}$  wavelet coefficients at scale  $2^j$ , so there are  $2^{-k}$  wavelet coefficients at scales  $2^j > 2^k$ . The right inequality (9.22) implies that

$$\varepsilon_l(2^{-k}, f) = \sum_{j=-\infty}^k \sum_{n=0}^{2^{-j}-1} |\langle f, \psi_{j,n} \rangle|^2 \leq B^2 \sum_{j=-\infty}^k 2^{-j} 2^{j(2\alpha+1)} = \frac{B^2 \|f\|_{\tilde{\mathcal{C}}^\alpha}^2 2^{2\alpha k}}{1 - 2^{-2\alpha}}.$$

For  $k = -\lfloor \log_2 N \rfloor$ , we derive that  $\varepsilon_l(N, f) = O(\|f\|_{\tilde{\mathcal{C}}^\alpha}^2 2^{2\alpha k}) = O(\|f\|_{\tilde{\mathcal{C}}^\alpha}^2 N^{-2\alpha})$ . ■

### Discontinuity and Bounded Variation

If  $f$  is not uniformly regular, then linear wavelet approximations perform poorly. If  $f$  has a discontinuity in  $(0, 1)$ , then  $f \notin \mathbf{W}^s[0, 1]$  for  $s > 1/2$ , so Theorem 9.5 proves that we cannot have  $\varepsilon_l(N, f) = O(N^{-\alpha})$  for  $\alpha > 1$ .

If  $f$  has a bounded total variation norm  $\|f\|_V$ , then Theorem 9.14 will prove that wavelet approximation errors satisfy  $\varepsilon_l(N, f) = O(\|f\|_V^2 N^{-1})$ . The same Fourier approximation result was obtained in Theorem 9.3. If  $f = C \mathbf{1}_{[0,1/2]}$ , then one can verify that wavelet approximation gives  $\varepsilon_l(N, f) \sim \|f\|_V^2 N^{-1}$  (Exercise 9.10).

Figure 9.1 gives an example of a discontinuous signal with bounded variation that is approximated by its larger-scale wavelet coefficients. The largest-amplitude errors are in the neighborhood of singularities, where the scale should be refined. The relative approximation error  $\|f - f_N\|/\|f\| = 8.56 \cdot 10^{-2}$  is almost the same as in a Fourier basis.

#### 9.1.4 Karhunen-Loève Approximations

Suppose that signals are modeled as realizations of a random process  $F$ . We prove that the basis that minimizes the average linear approximation error is the Karhunen-Loève basis, which diagonalizes the covariance operator of  $F$ . To avoid the subtleties of diagonalizing an infinite dimensional operator, we consider signals of finite dimension  $P$ , which means that  $F[n]$  is a random vector of size  $P$ .

Section A.6 in the Appendix reviews the covariance properties of random vectors. If  $F[n]$  does not have a zero mean, we subtract the expected value  $E\{F[n]\}$  from  $F[n]$  to get a zero mean. The random vector  $F$  can be decomposed in an orthogonal basis  $\{g_m\}_{0 \leq m < P}$ :

$$F = \sum_{m=0}^{P-1} \langle F, g_m \rangle g_m.$$

Each coefficient

$$\langle F, g_m \rangle = \sum_{n=0}^{P-1} F[n] g_m^*[n]$$

is a random variable (see Section A.6 in the Appendix). The approximation from the first  $N$  vectors of the basis is the orthogonal projection on the space  $\mathbf{U}_N$  generated by these vectors:

$$F_N = \sum_{m=0}^{N-1} \langle F, g_m \rangle g_m.$$

The resulting mean-square error is

$$E\{\varepsilon_l(N, F)\} = E\left\{\|F - F_N\|^2\right\} = \sum_{m=N}^{P-1} E\{|\langle F, g_m \rangle|^2\}.$$

The error is related to the covariance of  $F$  defined by

$$R_F[n, m] = E\{F[n] F^*[m]\}.$$

Let  $K_F$  be the covariance operator represented by this matrix. It is symmetric and positive and is thus diagonalized in an orthogonal basis called a *Karhunen-Loève basis*. This basis is not unique if several eigenvalues are equal. Theorem 9.8 proves that a Karhunen-Loève basis is optimal for linear approximations.

**Theorem 9.8.** For all  $N \geq 1$ , the expected approximation error

$$E\{\varepsilon_l(N, F)\} = E\{\|F - F_N\|^2\} = \sum_{m=N}^{P-1} E\{|\langle F, g_m \rangle|^2\}$$

is minimum if and only if  $\{g_m\}_{0 \leq m < P}$  is a Karhunen-Loève basis that diagonalizes the covariance  $K_F$  of  $F$  with vectors indexed in decreasing eigenvalue order:

$$\langle K_F g_m, g_m \rangle \geq \langle K_F g_{m+1}, g_{m+1} \rangle \quad \text{for } 0 \leq m < P-1.$$

**Proof.** Let us first observe that

$$E\{\varepsilon_l(F, N)\} = \sum_{m=N}^{P-1} \langle K_F g_m, g_m \rangle, \tag{9.25}$$

because for any vector  $z[n]$ ,

$$\begin{aligned} E\{|\langle F, z \rangle|^2\} &= E\left\{\sum_{n=0}^{P-1} \sum_{m=0}^{P-1} F[n] F[m] z[n] z^*[m]\right\} \\ &= \sum_{n=0}^{P-1} \sum_{m=0}^{P-1} R_F[n, m] z[n] z^*[m] \\ &= \langle K_F z, z \rangle. \end{aligned}$$

We now prove that (9.25) is minimum if the basis diagonalizes  $K_F$ . Let us consider an arbitrary orthonormal basis  $\{h_m\}_{0 \leq m < P}$ . The trace  $\text{tr}(K_F)$  of  $K_F$  is independent of the basis:

$$\text{tr}(K_F) = \sum_{m=0}^{P-1} \langle K_F h_m, h_m \rangle.$$

Thus, the basis that minimizes  $\sum_{m=N}^{P-1} \langle K_F h_m, h_m \rangle$ , maximizes  $\sum_{m=0}^{N-1} \langle K_F h_m, h_m \rangle$ .

Let  $\{g_m\}_{0 \leq m < P}$  be a basis that diagonalizes  $K_F$ :

$$K_F g_m = \sigma_m^2 g_m \quad \text{with} \quad \sigma_m^2 \geq \sigma_{m+1}^2 \quad \text{for} \quad 0 \leq m < P-1.$$

The theorem is proved by verifying that for all  $N \geq 0$ ,

$$\sum_{m=0}^{N-1} \langle K_F h_m, h_m \rangle \leq \sum_{m=0}^{N-1} \langle K_F g_m, g_m \rangle = \sum_{m=0}^{N-1} \sigma_m^2.$$

To relate  $\langle K_F h_m, h_m \rangle$  to the eigenvalues  $\{\sigma_i^2\}_{0 \leq i < P}$ , we expand  $h_m$  in the basis  $\{g_i\}_{0 \leq i < P}$ :

$$\langle K_F h_m, h_m \rangle = \sum_{i=0}^{P-1} |\langle h_m, g_i \rangle|^2 \sigma_i^2. \quad (9.26)$$

Thus,

$$\sum_{m=0}^{N-1} \langle K_F h_m, h_m \rangle = \sum_{m=0}^{N-1} \sum_{i=0}^{P-1} |\langle h_m, g_i \rangle|^2 \sigma_i^2 = \sum_{i=0}^{P-1} q_i \sigma_i^2$$

with

$$0 \leq q_i = \sum_{m=0}^{N-1} |\langle h_m, g_i \rangle|^2 \leq 1 \quad \text{and} \quad \sum_{i=0}^{P-1} q_i = N.$$

We evaluate

$$\begin{aligned} \sum_{m=0}^{N-1} \langle K_F h_m, h_m \rangle - \sum_{i=0}^{N-1} \sigma_i^2 &= \sum_{i=0}^{P-1} q_i \sigma_i^2 - \sum_{i=0}^{N-1} \sigma_i^2 \\ &= \sum_{i=0}^{P-1} q_i \sigma_i^2 - \sum_{i=0}^{N-1} \sigma_i^2 + \sigma_{N-1}^2 \left( N - \sum_{i=0}^{P-1} q_i \right) \\ &= \sum_{i=0}^{N-1} (\sigma_i^2 - \sigma_{N-1}^2) (q_i - 1) + \sum_{i=N}^{P-1} q_i (\sigma_i^2 - \sigma_{N-1}^2). \end{aligned}$$

Since the eigenvalues are listed in order of decreasing amplitude, it follows that

$$\sum_{m=0}^{N-1} \langle K_F h_m, h_m \rangle - \sum_{m=0}^{N-1} \sigma_m^2 \leq 0.$$

Suppose that this last inequality is an equality. We finish the proof by showing that  $\{h_m\}_{0 \leq m < P}$  must be a Karhunen-Loëve basis. If  $i < N$ , then  $\sigma_i^2 \neq \sigma_{N-1}^2$  implies  $q_i = 1$ . If  $i \geq N$ , then  $\sigma_i^2 \neq \sigma_{N-1}^2$  implies  $q_i = 0$ . This is valid for all  $N \geq 0$  if  $\langle h_m, g_i \rangle \neq 0$  only when  $\sigma_i^2 = \sigma_m^2$ . This means that the change of basis is performed inside each eigenspace of  $K_F$ , so  $\{h_m\}_{0 \leq m < P}$  also diagonalizes  $K_F$ . ■

The eigenvectors  $g_m$  of the covariance matrix are called *principal components*. Theorem 9.8 proves that a Karhunen-Loëve basis yields the smallest expected linear error when approximating a class of signals by their projection on  $N$  orthogonal vectors.

Theorem 9.8 has a simple geometrical interpretation. The realizations of  $F$  define a cloud of points in  $\mathbb{C}^P$ . The density of this cloud specifies the probability distribution of  $F$ . The vectors  $g_m$  of the Karhunen-Loëve basis give the directions of the principal axes of the cloud. Large eigenvalues  $\sigma_m^2$  correspond to directions  $g_m$  along which the cloud is highly elongated. Theorem 9.8 proves that projecting the realizations of  $F$  on these principal components yields the smallest average error. If  $F$  is a Gaussian random vector, the probability density is uniform along ellipsoids with axes proportional to  $\sigma_m$  in the direction of  $g_m$ . Thus, these principal directions are truly the preferred directions of the process.

### Random-Shift Processes

If the process is not Gaussian, its probability distribution can have a complex geometry, and a linear approximation along the principal axes may not be efficient. As an example, we consider a random vector  $F[n]$  of size  $P$  that is a random-shift modulo  $P$  of a deterministic signal  $f[n]$  of zero mean,  $\sum_{n=0}^{P-1} f[n] = 0$ :

$$F[n] = f[(n - Q) \bmod P]. \quad (9.27)$$

Shift  $Q$  is an integer random variable with a probability distribution that is uniform on  $[0, P - 1]$ :

$$\Pr(Q = p) = \frac{1}{P} \quad \text{for } 0 \leq p < P.$$

This process has a zero mean:

$$E\{F[n]\} = \frac{1}{P} \sum_{p=0}^{P-1} f[(n - p) \bmod P] = 0,$$

and its covariance is

$$\begin{aligned} R_F[n, k] &= E\{F[n]F[k]\} = \frac{1}{P} \sum_{p=0}^{P-1} f[(n - p) \bmod P]f[(k - p) \bmod P] \\ &= \frac{1}{P} f \circledast \bar{f}[n - k] \quad \text{with} \quad \bar{f}[n] = f[-n]. \end{aligned} \quad (9.28)$$

Thus,  $R_F[n, k] = R_F[n - k]$  with

$$R_F[k] = \frac{1}{P} f \circledast \bar{f}[k].$$

Since  $R_F$  is  $P$  periodic,  $F$  is a circular stationary random vector, as defined in Section A.6 in the Appendix. The covariance operator  $K_F$  is a circular convolution with  $R_F$ , and is therefore diagonalized in the discrete Fourier Karhunen-Loève basis  $\{\frac{1}{\sqrt{P}} \exp\left(\frac{i2\pi mn}{P}\right)\}_{0 \leq m < P}$ . The eigenvalues are given by the Fourier transform of  $R_F$ :

$$\sigma_m^2 = \hat{R}_F[m] = \frac{1}{P} |\hat{f}[m]|^2. \quad (9.29)$$

Theorem 9.8 proves that a linear approximation yields a minimum error in this Fourier basis. To better understand this result, let us consider an extreme case where  $f[n] = \delta[n] - \delta[n - 1]$ . Theorem 9.8 guarantees that the Fourier Karhunen-Loève basis produces a smaller expected approximation error than does a canonical basis of Diracs  $\{g_m[n] = \delta[n - m]\}_{0 \leq m < P}$ . Indeed, we do not know a priori the abscissa of the nonzero coefficients of  $F$ , so there is no particular Dirac that is better adapted to perform the approximation. Since the Fourier vectors cover the whole support of  $F$ , they always absorb part of the signal energy:

$$E \left\{ \left| \left\langle F[n], \frac{1}{\sqrt{P}} \exp\left(\frac{i2\pi mn}{P}\right) \right\rangle \right|^2 \right\} = \hat{R}_F[m] = \frac{4}{P} \sin^2\left(\frac{\pi k}{P}\right).$$

Therefore, selecting  $N$  higher-frequency Fourier coefficients yields a better mean-square approximation than choosing a priori  $N$  Dirac vectors to perform the approximation.

The linear approximation of  $F$  in a Fourier basis is not efficient because all the eigenvalues  $\hat{R}_F[m]$  have the same order of magnitude. A simple nonlinear algorithm can improve this approximation. In a Dirac basis,  $F$  is exactly reproduced by selecting the two Diracs corresponding to the largest-amplitude coefficients having positions  $Q$  and  $Q - 1$  that depend on each realization of  $F$ . A nonlinear algorithm that selects the largest-amplitude coefficient for each realization of  $F$  is not efficient in a Fourier basis. Indeed, the realizations of  $F$  do not have their energy concentrated over a few large-amplitude Fourier coefficients. This example shows that when  $F$  is not a Gaussian process, a nonlinear approximation may be much more precise than a linear approximation, and the Karhunen-Loève basis is no longer optimal.

## 9.2 NONLINEAR APPROXIMATIONS

Digital images or sounds are signals discretized over spaces of a large dimension  $N$ , because linear approximation error has a slow decay. Digital camera images have  $N \geq 10^6$  pixels, whereas one second of a CD recording has  $N = 40 \cdot 10^3$  samples. Sparse signal representations are obtained by projecting such signals over less vectors selected adaptively in an orthonormal basis of discrete signals in  $\mathbb{C}^N$ . This is equivalent to performing a nonlinear approximation of the input analog signal in a basis of  $L^2[0, 1]$ .

Section 9.2.1 analyzes the properties of the resulting nonlinear approximation error. Sections 9.2.2 and 9.2.3 prove that nonlinear wavelet approximations are equivalent to adaptive grids, and can provide sparse representations of signals including singularities. Approximations of functions in Besov spaces and with bounded variations are studied in Section 9.2.3.

### 9.2.1 Nonlinear Approximation Error

The discretization of an input analog signal  $f$  computes  $N$  sample values  $\{\langle f, \phi_n \rangle\}_{0 \leq n < N}$  that specify the projection  $f_N$  of  $f$  over an approximation space  $\mathbf{U}_N$  of dimension  $N$ . A nonlinear approximation further approximates this projection over a basis providing a sparse representation.

Let  $\mathcal{B} = \{g_m\}_{m \in \mathbb{N}}$  be an orthonormal basis of  $\mathbf{L}^2[0, 1]$  or  $\mathbf{L}^2[0, 1]^2$ , with the first  $N$  vectors defining a basis of  $\mathbf{U}_N$ . The orthogonal projection in  $\mathbf{U}_N$  can be written as

$$f_N(x) = \sum_{m=0}^{N-1} \langle f, g_m \rangle g_m(x),$$

and the linear approximation error is

$$\|f - f_N\|^2 = \sum_{n=N}^{+\infty} |\langle f, g_m \rangle|^2.$$

Let us reproject  $f_N$  over a subset of  $M < N$  vectors  $\{g_m\}_{m \in \Lambda}$  with  $\Lambda \subset [0, N - 1]$ :

$$f_\Lambda(x) = \sum_{m \in \Lambda} \langle f, g_m \rangle g_m(x).$$

The approximation error is the sum of the remaining coefficients:

$$\|f_N - f_\Lambda\|^2 = \sum_{m \notin \Lambda} |\langle f, g_m \rangle|^2. \quad (9.30)$$

The approximation set that minimizes this error is the set  $\Lambda_T$  of  $M$  vectors corresponding to the largest inner-product amplitude  $|\langle f, g_m \rangle|$ , and thus above a threshold  $T$  that depends on  $M$ :

$$\Lambda_T = \{m : 0 \leq m < N, |\langle f, g_m \rangle| \geq T\} \quad \text{with} \quad |\Lambda_T| = M. \quad (9.31)$$

The minimum approximation error is the energy of coefficients below  $T$ :

$$\varepsilon_n(M, f) = \|f_N - f_{\Lambda_T}\|^2 = \sum_{m \notin \Lambda_T} |\langle f, g_m \rangle|^2.$$

In the following, we often write that  $f_M = f_{\Lambda_T}$  is the best  $M$ -term approximation.

The overall error is the sum of the linear error when projecting  $f$  on  $\mathbf{U}_N$  and the nonlinear approximation error:

$$\varepsilon_n(M, f) = \|f - f_M\|^2 = \|f - f_N\|^2 + \|f_N - f_M\|^2. \quad (9.32)$$

If  $N$  is large enough so that all coefficients above  $T$  are in the first  $N$ ,

$$T \geq \max_{|m| \geq N} |\langle f, g_m \rangle| \quad \text{and hence} \quad N > \arg \max_m \{|\langle f, g_m \rangle| \geq T\}, \quad (9.33)$$

then the  $M$  largest signal coefficients are among the first  $N$  and the nonlinear error (9.32) is the minimum error obtained from  $M$  coefficients chosen anywhere in the infinite basis  $\mathcal{B} = \{g_m\}_{m \in \mathbb{N}}$ . In this case, the linear approximation space  $\mathbf{U}_N$  and  $f_N$  do not play any explicit role in the error  $\varepsilon_n(M, f)$ . If  $|\langle f, g_m \rangle| \leq C m^{-\beta}$  for some  $\beta > 0$ , then we can choose  $N \geq C^\beta T^\beta$ . In the following, this condition is supposed to be satisfied.

### Discrete Numerical Computations

The linear approximation space  $\mathbf{U}_N$  is important for discrete computations. A nonlinear approximation  $f_{\Lambda_T}$  is computed by calculating the nonlinear approximation of the discretized signal  $a[n] = \langle f, \phi_n \rangle$  for  $0 \leq n < N$ , and performing a discrete-to-analog conversion.

Since both the discretization family  $\{\phi_n\}_{0 \leq n < N}$  and the approximation basis  $\{g_m\}_{0 \leq m < N}$  are orthonormal bases of  $\mathbf{U}_N$ ,  $\{h_m[n] = \langle g_m, \phi_n \rangle\}_{0 \leq m < N}$  is an orthonormal basis of  $\mathbb{C}^N$ . Analog signal inner products in  $L^2[0, 1]$  and their discretization in  $\mathbb{C}^N$  are then equal:

$$\langle a[n], h_m[n] \rangle = \langle f(x), g_m(x) \rangle \quad \text{for } 0 \leq m < N. \quad (9.34)$$

The nonlinear approximation of the signal  $a[n]$  in the basis  $\{h_m\}_{0 \leq m < N}$  of  $\mathbb{C}^N$  is

$$a_{\Lambda_T}[n] = \sum_{m \in \Lambda_T} \langle a, h_m \rangle h_m[n] \quad \text{with} \quad \Lambda_T = \{m : |\langle a, h_m \rangle| \geq T\}.$$

It results from (9.34) that the analog conversion of this discrete signal is the nonlinear analog approximation:

$$f_{\Lambda_T}(x) = \sum_{n=0}^{N-1} a_{\Lambda_T}[n] \phi_n(x) = \sum_{n=0}^{N-1} a_{\Lambda_T}[n] \phi_n(x) = \sum_{m \in \Lambda_T} \langle f, g_m \rangle g_m(x).$$

The number of operations to compute  $a_{\Lambda_T}$  is dominated by the number of operations to compute the  $N$  signal coefficients  $\{\langle a, h_m \rangle\}_{0 \leq m < N}$ , which takes  $O(N \log_2 N)$  operations in a discrete Fourier basis, and  $O(N)$  in a discrete wavelet basis. Thus, reducing  $N$  decreases the number of operations and does not affect the nonlinear approximation error, as long as (9.33) is satisfied. Given this equivalence between discrete and analog nonlinear approximations, we now concentrate on analog functions to relate this error to their regularity.

### Approximation Error

To evaluate the nonlinear approximation error  $\varepsilon_n(M, f)$ , the coefficients  $\{|\langle f, g_m \rangle|\}_{m \in \mathbb{N}}$  are sorted in decreasing order. Let  $f_{\mathcal{B}}^r[k] = \langle f, g_{m_k} \rangle$  be the coefficient of rank  $k$ :

$$|f_{\mathcal{B}}^r[k]| \geq |f_{\mathcal{B}}^r[k+1]| \quad \text{with} \quad k > 0.$$

The best  $M$ -term nonlinear approximation computed from the  $M$  largest coefficients is:

$$f_M = \sum_{k=1}^M f_B^r[k] g_{m_k}. \quad (9.35)$$

The resulting error is

$$\varepsilon_n(M, f) = \|f - f_M\|^2 = \sum_{k=M+1}^{+\infty} |f_B^r[k]|^2.$$

Theorem 9.9 relates the decay of this approximation error as  $M$  increases to the decay of  $|f_B^r[k]|$  as  $k$  increases.

**Theorem 9.9.** Let  $s > 1/2$ . If there exists  $C > 0$  such that  $|f_B^r[k]| \leq C k^{-s}$ , then

$$\varepsilon_n(M, f) \leq \frac{C^2}{2s-1} M^{1-2s}. \quad (9.36)$$

Conversely, if  $\varepsilon_n(M, f)$  satisfies (9.36), then

$$|f_B^r[k]| \leq \left(1 - \frac{1}{2s}\right)^{-s} C k^{-s}. \quad (9.37)$$

**Proof.** Since

$$\varepsilon_n(M, f) = \sum_{k=M+1}^{+\infty} |f_B^r[k]|^2 \leq C^2 \sum_{k=M+1}^{+\infty} k^{-2s},$$

and

$$\sum_{k=M+1}^{+\infty} k^{-2s} \leq \int_M^{+\infty} x^{-2s} dx = \frac{M^{1-2s}}{2s-1}, \quad (9.38)$$

we derive (9.36).

Conversely, let  $\alpha < 1$ ,

$$\varepsilon_n(\alpha M, f) \geq \sum_{k=\alpha M+1}^M |f_B^r[k]|^2 \geq (1-\alpha) M |f_B^r[M]|^2.$$

So if (9.36) is satisfied,

$$|f_B^r[M]|^2 \leq \frac{\varepsilon_n(\alpha M, f)}{1-\alpha} M^{-1} \leq \frac{C^2}{2s-1} \frac{\alpha^{1-2s}}{1-\alpha} M^{-2s}.$$

For  $\alpha = 1 - 1/(2s)$ , we get (9.37) for  $k = M$ . ■

### **L<sup>p</sup> Spaces**

Theorem 9.10 relates the decay of sorted inner products to their  $\ell^p$  norm

$$\|f\|_{\mathcal{B},p} = \left( \sum_{m=0}^{+\infty} |\langle f, g_m \rangle|^p \right)^{1/p}.$$

It derives a decay upper bound of the error  $\varepsilon_n(M, f)$ .

**Theorem 9.10.** Let  $p < 2$ . If  $\|f\|_{\mathcal{B},p} < +\infty$ , then

$$|f_{\mathcal{B}}^r[k]| \leq \|f\|_{\mathcal{B},p} k^{-1/p} \quad (9.39)$$

and  $\varepsilon_n(M, f) = o(M^{1-2/p})$ .

**Proof.** We prove (9.39) by observing that

$$\|f\|_{\mathcal{B},p}^p = \sum_{n=1}^{+\infty} |f_{\mathcal{B}}^r[n]|^p \geq \sum_{n=1}^k |f_{\mathcal{B}}^r[n]|^p \geq k |f_{\mathcal{B}}^r[k]|^p.$$

To show that  $\varepsilon_n(M, f) = o(M^{1-2/p})$ , we set

$$S[k] = \sum_{n=k}^{2k-1} |f_{\mathcal{B}}^r[n]|^p \geq k |f_{\mathcal{B}}^r[2k]|^p.$$

Thus,

$$\begin{aligned} \varepsilon_n(M, f) &= \sum_{k=M+1}^{+\infty} |f_{\mathcal{B}}^r[k]|^2 \leq \sum_{k=M+1}^{+\infty} S[k/2]^{2/p} (k/2)^{-2/p} \\ &\leq \sup_{k>M/2} |S[k]|^{2/p} \sum_{k=M+1}^{+\infty} (k/2)^{-2/p}. \end{aligned}$$

Since  $\|f\|_{\mathcal{B},p}^p = \sum_{n=1}^{+\infty} |f_{\mathcal{B}}^r[n]|^p < +\infty$ , it follows that  $\lim_{k \rightarrow +\infty} \sup_{k>M/2} |S[k]| = 0$ . Thus, we derive from (9.38) that  $\varepsilon_n(M, f) = o(M^{1-2/p})$ . ■

This theorem specifies spaces of functions that are well approximated by a few vectors of an orthogonal basis  $\mathcal{B}$ . We denote

$$\mathbf{B}_{\mathcal{B},p} = \left\{ f \in \mathbf{H} : \|f\|_{\mathcal{B},p} < +\infty \right\}. \quad (9.40)$$

If  $f \in \mathbf{B}_{\mathcal{B},p}$ , then Theorem 9.10 proves that  $\varepsilon_n(M, f) = o(M^{1-2/p})$ . This is called a *Jackson inequality* [20]. Conversely, if  $\varepsilon_n(M, f) = O(M^{1-2/p})$ , then the *Bernstein inequality* (9.37) for  $s = 1/p$  shows that  $f \in \mathbf{B}_{\mathcal{B},q}$  for any  $q > p$ . Section 9.2.3 studies the properties of spaces  $\mathbf{B}_{\mathcal{B},p}$  for wavelet bases.

### 9.2.2 Wavelet Adaptive Grids

A nonlinear approximation in a wavelet orthonormal basis keeps the largest-amplitude coefficients. We saw in Section 6.1.3 that these coefficients occur near singularities. Thus, wavelet nonlinear approximation defines an adaptive grid that refines the approximation scale in the neighborhood of the signal sharp transitions. Such approximations are particularly well adapted to piecewise regular signals. The precision of nonlinear wavelet approximation is also studied for bounded variation functions and more general Besov space functions [209].

We consider a wavelet basis adapted to  $\mathbf{L}^2[0, 1]$ , constructed in Section 7.5.3 with compactly supported wavelets that are  $\mathbf{C}^q$  with  $q$  vanishing moments:

$$\mathcal{B} = \left[ \{\phi_{j,n}\}_{0 \leq n < 2^{-j}}, \{\psi_{j,n}\}_{-\infty < j \leq L, 0 \leq n < 2^{-j}} \right].$$

To simplify notation, we write  $\phi_{j,n} = \psi_{j+1,n}$ .

If the analog signal  $f \in \mathbf{L}^2[0, 1]$  is approximated at the scale  $2^L$  with  $N = 2^{-L}$  samples  $\{\langle f, \phi_{L,n} \rangle\}_{0 \leq n < 2^{-L}}$ , then the corresponding  $N$  wavelet coefficients  $\{\langle f, \psi_{j,n} \rangle\}_{n,j>L}$  are computed with  $O(N)$  operations with the fast wavelet transform algorithm of Section 7.3.1. The best nonlinear approximation of  $f \in \mathbf{L}^2[0, 1]$  from  $M$  wavelet coefficients above  $T$  at scales  $2^j > 2^L$  is

$$f_M = \sum_{(j,n) \in \Lambda_T} \langle f, \psi_{j,n} \rangle \psi_{j,n} \quad \text{with} \quad \Lambda_T = \{(j, n) : j > L, |\langle f, \psi_{j,n} \rangle| \geq T\}.$$

The approximation error is  $\varepsilon_n(M, f) = \sum_{(j,n) \notin \Lambda_T} |\langle f, \psi_{j,n} \rangle|^2$ . Theorem 9.11 proves that if  $f$  is bounded, then for a sufficiently large  $N$  the approximation support  $\Lambda_T$  corresponds to all possible wavelet coefficients above  $T$ .

**Theorem 9.11.** If  $f$  is bounded, then all wavelets producing coefficients above  $T$  are in an approximation space  $\mathbf{V}_L$  of dimension  $N = 2^{-L} = O(\|f\|_\infty^2 T^{-2})$ .

**Proof.** If  $f$  is bounded, then

$$\begin{aligned} |\langle f, \psi_{j,n} \rangle| &= \left| \int_0^1 f(t) 2^{-j/2} \psi(2^{-j}t - n) dt \right| \\ &\leq 2^{j/2} \sup_t |f(t)| \int_0^1 |\psi(t)| dt = 2^{j/2} \|f\|_\infty \|\psi\|_1. \end{aligned} \tag{9.41}$$

So,  $|\langle f, \psi_{j,n} \rangle| \geq T$  implies that  $2^j \geq T^2 \|f\|_\infty^{-2} \|\psi\|_1^{-2}$ , which proves the theorem for  $2^L = T^2 \|f\|_\infty^{-2} \|\psi\|_1^{-2}$ . ■

This theorem shows that for bounded signals, if the discretization  $2^L$  is sufficiently small, then the nonlinear approximation error computed from the first  $N = 2^{-L}$  wavelet coefficients is equal to the approximation error obtained by selecting the  $M$  largest wavelet coefficients in the infinite-dimensional wavelet basis. In the following, we suppose that this condition is satisfied, and thus do not have to worry about the discretization scale.

### Piecewise Regular Signals

Piecewise regular signals define a first simple model where nonlinear wavelet approximations considerably outperform linear approximations. We consider signals with a finite number of singularities and that are uniformly regular between singularities. Theorem 9.12 characterizes the linear and nonlinear wavelet approximation error decay for such signals.

**Theorem 9.12.** If  $f$  has  $K$  discontinuities on  $[0, 1]$  and is uniformly Lipschitz  $\alpha$  between these discontinuities, with  $1/2 < \alpha < q$ , then

$$\varepsilon_l(M, f) = O(K \|f\|_{C^\alpha}^2 M^{-1}) \quad \text{and} \quad \varepsilon_n(M, f) = O(\|f\|_{C^\alpha}^2 M^{-2\alpha}). \quad (9.42)$$

**Proof.** We distinguish type I wavelets  $\psi_{j,n}$  for  $n \in I_j$ , with a support including a point where  $f$  is discontinuous, from type II wavelets for  $n \in II_j$ , with a support that is included in a domain where  $f$  is uniformly Lipschitz  $\alpha$ .

Let  $C$  be the support size of  $\psi$ . At a scale  $2^j$ , each wavelet  $\psi_{j,n}$  has a support of size  $C2^j$ , translated by  $2^j n$ . Thus, there are at most  $|I_j| \leq C K$  type I wavelets  $\psi_{j,n}$  with supports that include at least one of the  $K$  discontinuities of  $f$ . Since  $\|f\|_\infty \leq \|f\|_{C^\alpha}$ , (9.41) shows for  $\alpha = 0$  that there exists  $B_0$  such that  $|\langle f, \psi_{j,n} \rangle| \leq B_0 \|f\|_{C^\alpha} 2^{j/2}$ .

At fine scales  $2^j$ , there are much more type II wavelets  $n \in II_j$ , but this number  $|II_j|$  is smaller than the total number  $2^{-j}$  of wavelets at this scale. Since  $f$  is uniformly Lipschitz  $\alpha$  on the support of  $\psi_{j,n}$ , the right inequality of (9.22) proves that there exists  $B$  such that

$$|\langle f, \psi_{j,n} \rangle| \leq B \|f\|_{C^\alpha} 2^{j(\alpha+1/2)}. \quad (9.43)$$

This linear approximation error from  $M = 2^{-k}$  wavelets satisfies

$$\begin{aligned} \varepsilon_l(M, f) &= \sum_{j \leq k} \left( \sum_{n \in I_j} |\langle f, \psi_{j,n} \rangle|^2 + \sum_{n \in II_j} |\langle f, \psi_{j,n} \rangle|^2 \right) \\ &\leq \sum_{j \leq k} \left( C K B_0^2 \|f\|_{C^\alpha}^2 2^j + 2^{-j} B^2 \|f\|_{C^\alpha}^2 2^{(2\alpha+1)j} \right) \\ &\leq \|f\|_{C^\alpha}^2 2 C K B_0^2 2^k + \|f\|_{C^\alpha}^2 (1 - 2^{-2\alpha})^{-1} B^2 2^{2\alpha k}. \end{aligned}$$

This inequality proves that for  $\alpha > 1/2$  the error term of type I wavelets dominates, and  $\varepsilon_l(M, f) = O(\|f\|_{C^\alpha}^2 K M^{-1})$ .

To compute the nonlinear approximaton error  $\varepsilon(M, f)$ , we evaluate the decay of ordered wavelet coefficients. Let  $f_B^r[k] = \langle f, \psi_{jk, n_k} \rangle$  be the coefficient of rank  $k$ :  $|f_B^r[k]| \geq |f_B^r[k+1]|$  for  $k \geq 1$ . Let  $f_{B,I}^r[k]$  and  $f_{B,II}^r[k]$  be the values of the wavelet coefficient of rank  $k$  in the type I and type II wavelets.

For  $l > 0$ , there are at most  $l K C$  type I coefficients at scales  $2^j > 2^{-l}$  and type I wavelet coefficients satisfy  $|\langle f, \psi_{j,n} \rangle| \leq B_0 \|f\|_{C^\alpha} 2^{-l/2}$  at scales  $2^j \leq 2^{-l}$ . It results that

$$f_{B,I}^r[IKC] \leq B_0 \|f\|_{C^\alpha} 2^{-l/2},$$

so  $f_{B,I}^r[k] = O(\|f\|_{C^\alpha} 2^{-k/(2KC)})$  has an exponential decay.

For  $l \geq 0$ , there are at most  $2^l$  type II wavelet coefficients at scales  $2^j > 2^{-l}$ , and type II wavelet coefficients satisfy  $|\langle f, \psi_{j,n} \rangle| \leq B \|f\|_{C^\alpha} 2^{-l(\alpha+1/2)}$  at scales  $2^j \leq 2^{-l}$ . It results that

$$f_{B,II}^r[2^{-l}] \leq B \|f\|_{C^\alpha} 2^{-l(\alpha+1/2)}.$$

It follows that  $f_{B,II}^r[k] = O(\|f\|_{C^\alpha} k^{-\alpha-1/2})$  for all  $k > 0$ .

Since type I coefficients have a much faster decay than type II coefficients, putting them together gives  $f_B^r[k] = O(\|f\|_{C^\alpha} k^{-\alpha-1/2})$ . From the inequality (9.36) of Theorem 9.9, it results that  $\varepsilon_n(M, f) = O(\|f\|_{C^\alpha}^2 M^{-2\alpha})$ . ■

Although there are few large wavelet coefficients created by the potential  $K$  discontinuities, the theorem proof shows that the linear approximation error is dominated by these discontinuities. On the contrary, these few wavelet coefficients have a negligible impact on the nonlinear approximation error. Thus, it decays as if there were no such discontinuities and  $f$  was uniformly Lipschitz  $\alpha$  over its whole support.

### Adaptive Grids

The approximation  $f_M$  calculated from the  $M$  largest-amplitude wavelet coefficients can be interpreted as an adaptive grid approximation where the approximation scale is refined in the neighborhood of singularities.

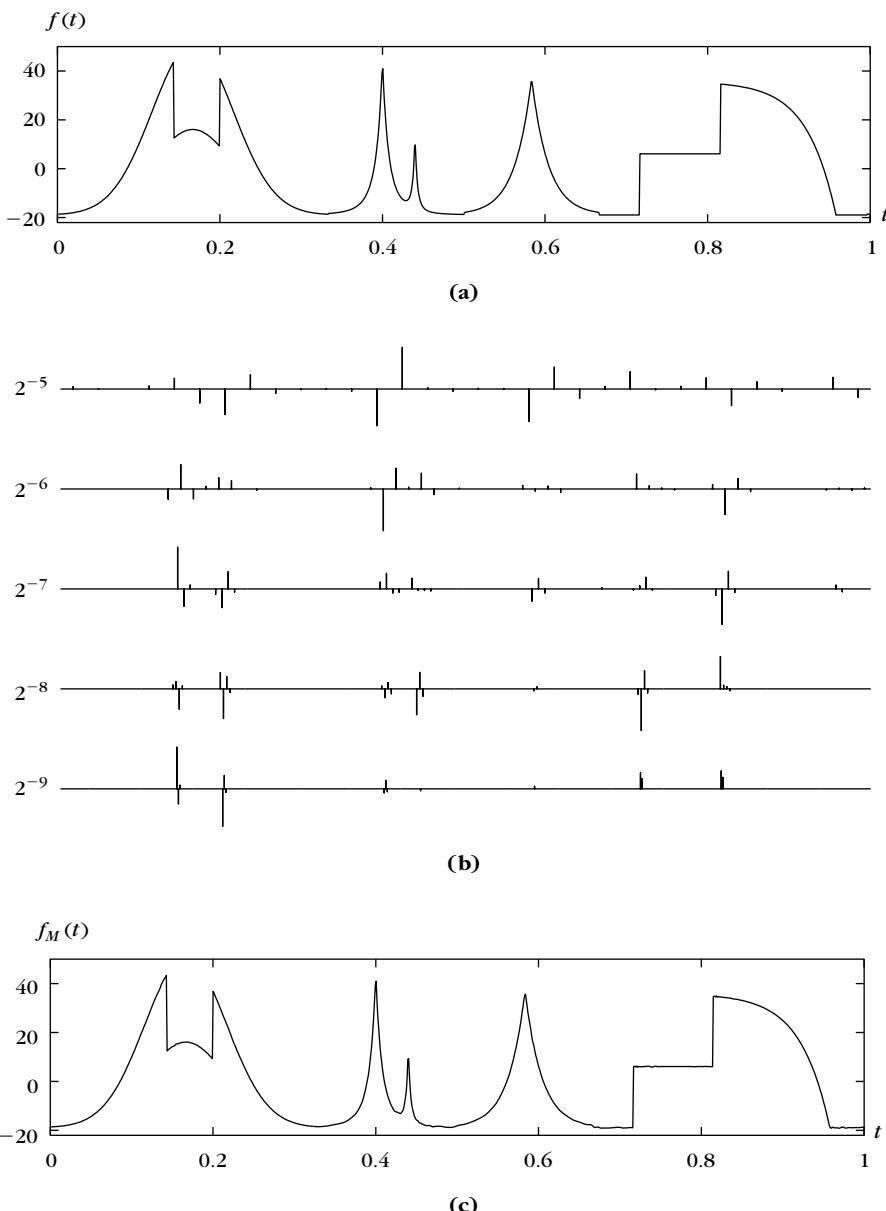
A nonlinear approximation keeps all coefficients above a threshold  $|\langle f, \psi_{j,n} \rangle| \geq T$ . In a region where  $f$  is uniformly Lipschitz  $\alpha$ , since  $|\langle f, \psi_{j,n} \rangle| \sim A 2^{j(\alpha+1/2)}$ , the coefficients above  $T$  are typically at scales

$$2^j > 2^l = \left(\frac{T}{A}\right)^{2/(2\alpha+1)}.$$

Setting all wavelet coefficients below the scale  $2^l$  to zero is equivalent to computing a local approximation of  $f$  at the scale  $2^l$ . The smaller the local Lipschitz regularity  $\alpha$ , the finer the approximation scale  $2^l$ .

Figure 9.2 shows the nonlinear wavelet approximation of a piecewise regular signal. Up and down Diracs correspond to positive and negative wavelet coefficients with an amplitude above  $T$ . The largest-amplitude wavelet coefficients are in the cone of influence of each singularity. The scale-space approximation support  $\Lambda_T$  specifies the geometry of the signal-sharp transitions. Since the approximation scale is refined in the neighborhood of each singularity, they are much better restored than in the fixed-scale linear approximation shown in Figure 9.1. The nonlinear approximation error in this case is 17 times smaller than the linear approximation error.

Nonlinear wavelet approximations are nearly optimal compared to adaptive spline approximations. A spline approximation  $\tilde{f}_M$  is calculated by choosing  $K$  nodes  $t_1 < t_2 < \dots < t_K$  inside  $[0, 1]$ . Over each interval  $[t_k, t_{k+1}]$ ,  $f$  is approximated by the closest polynomial of degree  $r$ . This polynomial spline  $\tilde{f}_M$  is specified by  $M = K(r+2)$  parameters, which are the node locations  $\{t_k\}_{1 \leq k \leq K}$  plus the  $K(r+1)$  parameters of the  $K$  polynomials of degree  $r$ . To reduce  $\|f - \tilde{f}_M\|$ , the nodes must be closely spaced when  $f$  is irregular and farther apart when  $f$  is smooth. However, finding the  $M$  parameters that minimize  $\|f - \tilde{f}_M\|$  is a difficult nonlinear optimization.

**FIGURE 9.2**

**(a)** Original signal  $f$ . **(b)** Each Dirac corresponds to one of the largest  $M = 0.15 N$  wavelet coefficients, calculated with a symmlet 4. **(c)** Nonlinear approximation  $f_M$  recovered from the  $M$  largest wavelet coefficients shown in (b),  $\|f - f_M\|/\|f\| = 5.1 \cdot 10^{-3}$ .

A spline wavelet basis of Battle-Lemarié gives nonlinear approximations that are also spline functions, but nodes  $t_k$  are restricted to dyadic locations  $2^j n$  with a scale  $2^j$  that is locally adapted to the signal regularity. It is computed with  $O(N)$  operations by projecting the signal in an approximation space of dimension  $N = O(T^2)$ . For large classes of signals, including balls of Besov spaces, the maximum approximation errors with wavelets or with optimized splines have the same decay rate when  $M$  increases [210]. Therefore, the computational overhead of an optimized spline approximation is not worth it.

### 9.2.3 Approximations in Besov and Bounded Variation Spaces

Studying the performance of nonlinear wavelet approximations more precisely requires introducing new spaces. As previously, we write the coarse-scale scaling functions  $\phi_{J,n} = \psi_{J+1,n}$ . The *Besov* space  $\mathbf{B}_{\beta,\gamma}^s[0, 1]$  is the set of functions  $f \in \mathbf{L}^2[0, 1]$  such that

$$\|f\|_{s,\beta,\gamma} = \left( \sum_{j=-\infty}^{J+1} \left[ 2^{-j(s+1/2-1/\beta)} \left( \sum_{n=0}^{2^{-j}-1} |\langle f, \psi_{j,n} \rangle|^{\beta} \right)^{1/\beta} \right]^{\gamma} \right)^{1/\gamma} < +\infty. \quad (9.44)$$

Frazier, Jawerth [260], and Meyer [375] proved that  $\mathbf{B}_{\beta,\gamma}^s[0, 1]$  does not depend on the particular choice of wavelet basis, as long as the wavelets in the basis have  $q > s$  vanishing moments and are in  $\mathbf{C}^q$ . Space  $\mathbf{B}_{\beta,\gamma}^s[0, 1]$  corresponds typically to functions that have a “derivative of order  $s$ ” that is in  $\mathbf{L}^\beta[0, 1]$ . The index  $\gamma$  is a fine-tuning parameter, which is less important. We need  $q > s$ , because a wavelet with  $q$  vanishing moments can test the differentiability of a signal only up to the order  $q$ . When removing the coarsest-scale scaling functions  $\phi_{J,n} = \psi_{J+1,n}$  from (9.44), the norm  $\|f\|_{s,\beta,\gamma}$  is called an homogeneous Besov norm that we shall write as  $\|f\|_{s,\beta,\gamma}^*$  and the corresponding homogeneous Besov space is  $\tilde{\mathbf{B}}_{\beta,\gamma}^s[0, 1]$ .

If  $\beta \geq 2$ , then functions in  $\mathbf{B}_{\beta,\gamma}^s[0, 1]$  have a uniform regularity of order  $s$ . For  $\beta = \gamma = 2$ , Theorem 9.4 proves that  $\mathbf{B}_{2,2}^s[0, 1] = \mathbf{W}^s[0, 1]$  is the space of  $s$  times differentiable functions in the sense of Sobolev. Theorem 9.5 proves that this space is characterized by the decay of the linear approximation error  $\varepsilon_l(N, f)$  and that  $\varepsilon_l(N, f) = o(N^{-2s})$ . Since  $\varepsilon_n(M, f) \leq \varepsilon_l(M, f)$ , clearly  $\varepsilon_n(M, f) = o(M^{-2s})$ . One can verify (Exercise 9.11) that nonlinear approximations do not improve linear approximations over Sobolev spaces. For  $\beta = \gamma = \infty$ , Theorem 9.6 proves that the homogeneous Besov norm is an homogeneous Hölder norm

$$\|f\|_{s,\infty,\infty}^* = \sup_{j \geq J,n} 2^{-j(\alpha+1/2)} |\langle f, \psi_{j,n} \rangle| \sim \|f\|_{\tilde{\mathbf{C}}^s}, \quad (9.45)$$

and the corresponding space  $\tilde{\mathbf{B}}_{\infty,\infty}^s[0, 1]$  is the homogeneous Hölder space of functions that are uniformly Lipschitz  $s$  on  $[0, 1]$ .

For  $\beta < 2$ , functions in  $\mathbf{B}_{\beta,\gamma}^s[0, 1]$  are not necessarily uniformly regular. Therefore, the adaptivity of nonlinear approximations significantly improves the decay rate of

the error. In particular, if  $p = \beta = \gamma$  and  $s = 1/2 + 1/p$ , then the Besov norm is a simple  $\ell^p$  norm:

$$\|f\|_{s,\beta,\gamma} = \left( \sum_{j=-\infty}^{J+1} \sum_{n=0}^{2^{-j}-1} |\langle f, \psi_{j,n} \rangle|^p \right)^{1/p}.$$

Theorem 9.10 proves that if  $f \in \mathbf{B}_{\beta,\gamma}^s[0, 1]$ , then  $\varepsilon_n(M, f) = o(M^{1-2/p})$ . The smaller  $p$  is, the faster the error decay. The proof of Theorem 9.12 shows that although  $f$  may be discontinuous, if the number of discontinuities is finite and if  $f$  is uniformly Lipschitz  $\alpha$  between these discontinuities, then its sorted wavelet coefficients satisfy  $|f'_B[k]| = O(k^{-\alpha-1/2})$ , so  $f \in \mathbf{B}_{\beta,\gamma}^s[0, 1]$  for  $1/p < \alpha + 1/2$ . This shows that these spaces include functions that are not  $s$  times differentiable at all points. The linear approximation error  $\varepsilon_l(M, f)$  for  $f \in \mathbf{B}_{\beta,\gamma}^s[0, 1]$  can decrease arbitrarily slowly because the  $M$  wavelet coefficients at the largest scales may be arbitrarily small. A nonlinear approximation is much more efficient in these spaces.

### Bounded Variation

Bounded variation functions are important examples of signals for which a nonlinear approximation yields a much smaller error than a linear approximation. The total variation norm is defined in (2.57) by

$$\|f\|_V = \int_0^1 |f'(t)| dt.$$

The derivative  $f'$  must be understood in the sense of distributions in order to include discontinuous functions. To compute the linear and nonlinear wavelet approximation error for bounded variation signals, Theorem 9.13 computes an upper and a lower bound of  $\|f\|_V$  from the modulus of wavelet coefficients.

**Theorem 9.13.** Consider a wavelet basis constructed with  $\psi$  such that  $\|\psi\|_V < +\infty$ . There exist  $A, B > 0$  such that for all  $f \in \mathbf{L}^2[0, 1]$ ,

$$\|f\|_V \leq B \sum_{j=-\infty}^{J+1} \sum_{n=0}^{2^{-j}-1} 2^{-j/2} |\langle f, \psi_{j,n} \rangle| = B \|f\|_{1,1,1}, \quad (9.46)$$

and

$$\|f\|_V \geq A \sup_{j \leq J} \left( \sum_{n=0}^{2^{-j}-1} 2^{-j/2} |\langle f, \psi_{j,n} \rangle| \right) = A \|f\|_{1,1,\infty}^*. \quad (9.47)$$

**Proof.** By decomposing  $f$  in the wavelet basis

$$f = \sum_{j=-\infty}^J \sum_{n=0}^{2^{-j}-1} \langle f, \psi_{j,n} \rangle \psi_{j,n} + \sum_{n=0}^{2^{-J}-1} \langle f, \phi_{J,n} \rangle \phi_{J,n}$$

we get

$$\|f\|_V \leq \sum_{j=-\infty}^J \sum_{n=0}^{2^{-j}-1} |\langle f, \psi_{j,n} \rangle| \|\psi_{j,n}\|_V + \sum_{n=0}^{2^{-J}-1} |\langle f, \phi_{J,n} \rangle| \|\phi_{J,n}\|_V. \quad (9.48)$$

The wavelet basis includes wavelets, with supports that are inside  $(0, 1)$ , and border wavelets, which are obtained by dilating and translating a finite number of mother wavelets. To simplify notations we write the basis as if there were a single mother wavelet:  $\psi_{j,n}(t) = 2^{-J/2}\psi(2^{-j}t - n)$ . Thus, we verify with a change of variable that

$$\|\psi_{j,n}\|_V = \int_0^1 2^{-j/2} 2^{-j} |\psi'(2^{-j}t - n)| dt = 2^{-j/2} \|\psi\|_V.$$

Since  $\phi_{J,n}(t) = 2^{-J/2}\phi(2^{-J}t - n)$ , we also prove that  $\|\phi_{J,n}\|_V = 2^{-J/2} \|\phi\|_V$ . Thus, the inequality (9.46) is derived from (9.48).

Since  $\psi$  has at least one vanishing moment, its primitive  $\theta$  is a function with the same support, which we suppose is included in  $[-K/2, K/2]$ . To prove (9.47) for  $j \leq J$ , we make an integration by parts:

$$\begin{aligned} \sum_{n=0}^{2^{-j}-1} |\langle f, \psi_{j,n} \rangle| &= \sum_{n=0}^{2^{-j}-1} \left| \int_0^1 f(t) 2^{-j/2} \psi(2^{-j}t - n) dt \right| \\ &= \sum_{n=0}^{2^{-j}-1} \left| \int_0^1 f'(t) 2^{j/2} \theta(2^{-j}t - n) dt \right| \\ &\leq 2^{j/2} \sum_{n=0}^{2^{-j}-1} \int_0^1 |f'(t)| |\theta(2^{-j}t - n)| dt. \end{aligned}$$

Since  $\theta$  has a support in  $[-K/2, K/2]$ ,

$$\sum_{n=0}^{2^{-j}-1} |\langle f, \psi_{j,n} \rangle| \leq 2^{j/2} K \sup_{t \in \mathbb{R}} |\theta(t)| \int_0^1 |f'(t)| dt \leq A^{-1} 2^{j/2} \|f\|_V. \quad (9.49)$$

This inequality proves (9.47). ■

This theorem shows that the total variation norm is bounded by two Besov norms:

$$A \|f\|_{1,1,\infty}^* \leq \|f\|_V \leq B \|f\|_{1,1,1}.$$

The lower bound is an homogeneous norm because the addition of a constant to  $f$  does not modify  $\|f\|_V$ . Space  $\mathbf{BV}[0, 1]$  of bounded variation functions is therefore embedded in the corresponding Besov spaces:

$$\mathbf{B}_{1,1}^1[0, 1] \subset \mathbf{BV}[0, 1] \subset B\mathbf{B}_{1,\infty}^1[0, 1].$$

Theorem 9.14 derives linear and nonlinear wavelet approximation errors for bounded variation signals.

**Theorem 9.14.** For all  $f \in \mathbf{BV}[0, 1]$  and  $M > 2q$ ,

$$\varepsilon_l(M, f) = O(\|f\|_V^2 M^{-1}), \quad (9.50)$$

and

$$\varepsilon_n(M, f) = O(\|f\|_V^2 M^{-2}). \quad (9.51)$$

**Proof.** Section 7.5.3 shows a wavelet basis of  $\mathbf{L}^2[0, 1]$  with boundary wavelets keeping their  $q$  vanishing moments that has  $2q$  different boundary wavelets and scaling functions, so the largest wavelet scale can be  $2^J = (2q)^{-1}$  but not smaller.

There are  $2^{-j}$  wavelet coefficients at a scale  $2^j$ , so for any  $L \leq J$ , there are  $2^{-L}$  wavelet and scaling coefficients at scales  $2^j > 2^L$ . The resulting linear wavelet approximation error is

$$\varepsilon_l(2^{-L}, f) = \sum_{j=-\infty}^L \sum_{n=0}^{2^{-j}-1} |\langle f, \psi_{j,n} \rangle|^2. \quad (9.52)$$

We showed in (9.47) that

$$\sum_{n=0}^{2^{-j}-1} |\langle f, \psi_{j,n} \rangle| \leq A^{-1} 2^{j/2} \|f\|_V,$$

and thus that

$$\sum_{n=0}^{2^{-j}-1} |\langle f, \psi_{j,n} \rangle|^2 \leq A^{-2} 2^j \|f\|_V^2.$$

It results from (9.52) that

$$\varepsilon_l(2^{-L}, f) \leq 2A^{-2} 2^L \|f\|_V^2.$$

Setting  $M = 2^{-L}$ , we derive (9.50).

Let us now prove the nonlinear approximation error bound (9.51). Let  $f_B^r[k]$  be the wavelet coefficient of rank  $k$ , excluding all the scaling coefficients  $\langle f, \phi_{J,n} \rangle$ , since we cannot control their value with  $\|f\|_V$ . We first show that there exists  $B_0$  such that for all  $f \in \mathbf{BV}[0, 1]$ ,

$$|f_B^r[k]| \leq B_0 \|f\|_V k^{-3/2}. \quad (9.53)$$

To take into account the fact that (9.53) does not apply to the  $2^J$  scaling coefficients  $\langle f, \phi_{J,n} \rangle$ , an upper bound of  $\varepsilon_n(M, f)$  is obtained by selecting the  $2^J$  scaling coefficients plus the  $M - 2^J$  biggest wavelet coefficients; thus,

$$\varepsilon_n(M, f) \leq \sum_{k=M-2^J+1}^{+\infty} |f_B^r[k]|^2. \quad (9.54)$$

For  $M > 2q = 2^{-J}$ , inserting (9.53) in (9.54) proves (9.51).

The upper bound (9.53) is proved by computing an upper bound of the number of coefficients larger than an arbitrary threshold  $T$ . At scale  $2^J$ , we denote by  $f_B^r[j, k]$

the coefficient of rank  $k$  among  $\{\langle f, \psi_{j,n} \rangle\}_{0 \leq n \leq 2^{-j}}$ . The inequality (9.49) proves that for all  $j \leq J$ ,

$$\sum_{n=0}^{2^{-j}-1} |\langle f, \psi_{j,n} \rangle| \leq A^{-1} 2^{j/2} \|f\|_V.$$

Thus, it follows from (9.39) that

$$|f_B^r[j, k]| \leq A^{-1} 2^{j/2} \|f\|_V k^{-1} = C 2^{j/2} k^{-1}.$$

Thus, at scale  $2^j$ , the number  $k_j$  of coefficients larger than  $T$  satisfies

$$k_j \leq \min(2^{-j}, 2^{j/2} C T^{-1}).$$

The total number  $k$  of coefficients larger than  $T$  is

$$\begin{aligned} k &= \sum_{j=-\infty}^J k_j \leq \sum_{2^j \geq (C^{-1}T)^{2/3}} 2^{-j} + \sum_{2^j > (C^{-1}T)^{2/3}} 2^{j/2} C T^{-1} \\ &\leq 6(C T^{-1})^{2/3}. \end{aligned}$$

By choosing  $T = |f_B^r[k]|$ , since  $C = A^{-1} \|f\|_V$ , we get

$$|f_B^r[k]| \leq 6^{3/2} A^{-1} \|f\|_V k^{-3/2},$$

which proves (9.53). ■

The asymptotic decay rate of linear and nonlinear approximation errors in Theorem 9.14 cannot be improved. If  $f \in \mathbf{BV}[0, 1]$  has discontinuities, then  $\varepsilon_l(M, f)$  decays like  $M^{-1}$  and  $\varepsilon_n(M, f)$  decays like  $M^{-2}$ . One can also prove [207] that this error-decay rate for all bounded variation functions cannot be improved by any type of nonlinear approximation scheme. In this sense, wavelets are optimal for approximating bounded variation functions.

### 9.3 SPARSE IMAGE REPRESENTATIONS

Approximation of images is more complex than one-dimensional signals, because singularities often belong to geometrical structures such as edges or textures. Nonlinear wavelet approximation defines adaptive approximation grids that are numerically highly effective. These approximations are optimal for bounded variation images, but not for images having edges that are geometrically regular. Section 9.3.2 introduces a piecewise regular image model with regular edges, and study adaptive triangulation approximations. Section 9.3.3 proves that curvelet frames yield asymptotically optimal approximation errors for such piecewise  $C^2$  regular images.

### 9.3.1 Wavelet Image Approximations

Linear and nonlinear approximations of functions in  $L^2[0, 1]^d$  can be calculated in separable wavelet bases. We concentrate on the two-dimensional case for image processing, and compute approximation errors for bounded variation images.

Section 7.7.4 constructs a separable wavelet basis of  $L^2[0, 1]^2$  from a wavelet basis of  $L^2[0, 1]$ , with separable products of wavelets and scaling functions. We suppose that all wavelets of the basis of  $L^2[0, 1]$  are  $C^q$  with  $q$  vanishing moments. The wavelet basis of  $L^2[0, 1]^2$  includes three mother wavelets  $\{\psi^l\}_{1 \leq l \leq 3}$  that are dilated by  $2^j$  and translated over a square grid of interval  $2^j$  in  $[0, 1]^2$ . As modulo modifications near the borders, these wavelets can be written as

$$\psi_{j,n}^l(x) = \frac{1}{2^j} \psi^l \left( \frac{x_1 - 2^j n_1}{2^j}, \frac{x_2 - 2^j n_2}{2^j} \right). \quad (9.55)$$

They have  $q$  vanishing moments in the sense that they are orthogonal to two-dimensional polynomials of a degree strictly smaller than  $q$ . If we limit the scales to  $2^j \leq 2^J$ , we must complete the wavelet family with two-dimensional scaling functions

$$\phi_{J,n}^2(x) = \frac{1}{2^J} \phi^2 \left( \frac{x_1 - 2^J n_1}{2^J}, \frac{x_2 - 2^J n_2}{2^J} \right)$$

to obtain the orthonormal basis of  $L^2[0, 1]^2$ :

$$\mathcal{B} = \left( \{\phi_{J,n}^2\}_{2^J n \in [0,1)^2} \cup \{\psi_{j,n}^l\}_{j \leq J, 2^j n \in [0,1)^2, 1 \leq l \leq 3} \right).$$

#### Linear Image Approximation

The linear discretization of an analog image in  $L^2[0, 1]^2$  can be defined by  $N = 2^{-2L}$  samples  $\{\langle f, \phi_{L,n}^2 \rangle\}_{2^L n \in [0,1)^2}$ , which characterize the orthogonal projection of  $f$  in the approximation space  $V_L$ . The precision of such linear approximations depends on the uniform image regularity.

Local image regularity can be measured with Lipschitz exponents. A function  $f$  is uniformly Lipschitz  $\alpha$  over a domain  $\Omega \subset \mathbb{R}^2$  if there exists  $K > 0$ , such that for any  $v \in \Omega$  one can find a polynomial  $p_v$  of degree  $\lfloor \alpha \rfloor$  such that

$$\forall x \in \Omega, \quad |f(x) - p_v(x)| \leq K |x - v|^\alpha. \quad (9.56)$$

The infimum of  $K$ , which satisfies (9.56), is the *homogeneous Hölder  $\alpha$  norm*  $\|f\|_{\tilde{C}^\alpha}$ . The Hölder  $\alpha$  norm of  $f$  also imposes that  $f$  is bounded:  $\|f\|_{C^\alpha} = \|f\|_{\tilde{C}^\alpha} + \|f\|_\infty$ . We write  $C^\alpha[0, 1]^2$  as the Hölder space of functions for which  $\|f\|_{C^\alpha} < +\infty$ . Similar to Theorem 9.15 in one dimension, Theorem 9.6 computes the linear approximation error decay of such functions, with  $C^q$  wavelets having  $q$  vanishing moments.

**Theorem 9.15.** There exists  $B \geq A > 0$  such that

$$A \|f\|_{\tilde{C}^\alpha} \leq \sup_{1 \leq l \leq 3, j \geq J, 0 \leq n < 2^{-j}} 2^{-j(\alpha+1)} |\langle f, \psi_{j,n}^l \rangle| \leq B \|f\|_{\tilde{C}^\alpha}. \quad (9.57)$$

**Proof.** The proof is essentially the same as the proof of Theorem 9.6 in one dimension. We shall only prove the right inequality. If  $f$  is uniformly Lipschitz  $\alpha$  on the support of  $\psi_{j,n}^l$ , since  $\psi_{j,n}^l$  is orthogonal to the polynomial  $p_{2^j n}$  approximating  $f$  at  $v = 2^j n$ , we get

$$\begin{aligned} |\langle f, \psi_{j,n}^l \rangle| &= |\langle f - p_{2^j n}, \psi_{j,n}^l \rangle| \leq \|f\|_{\tilde{\mathbf{C}}^\alpha} \int \int 2^{-j} |\psi^l(2^{-j}(x - 2^j n))| |x - 2^j n|^\alpha dx \\ &\leq \|f\|_{\tilde{\mathbf{C}}^\alpha} 2^{(\alpha+1)j} \int \int |\psi^l(x)| |x|^\alpha dx, \end{aligned}$$

which proves the right inequality of (9.57). The wavelet regularity is not used to prove this inequality. The left inequality requires that the wavelets are  $\mathbf{C}^q$ . ■

Theorem 9.16 computes the linear wavelet approximation error decay for images that are uniformly Lipschitz  $\alpha$ . It requires that wavelets have  $q$  vanishing moments, but no regularity condition is needed.

**Theorem 9.16.** If  $f$  is uniformly Lipschitz  $0 < \alpha \leq q$  over  $[0, 1]^2$ , then  $\varepsilon_l(N, f) = O(\|f\|_{\tilde{\mathbf{C}}^\alpha}^2 N^{-\alpha})$ .

**Proof.** There are  $3 2^{-2j}$  wavelet coefficients at scale  $2^j$  and  $2^{-2k}$  wavelet coefficients and scaling coefficients at scales  $2^j > 2^k$ . The right inequality of (9.57) proves that

$$|\langle f, \psi_{j,n}^l \rangle| \leq B \|f\|_{\tilde{\mathbf{C}}^\alpha} 2^{(\alpha+1)j}.$$

As a result,

$$\begin{aligned} \varepsilon_l(2^{-2k}, f) &= \sum_{j=-\infty}^k \sum_{l=1}^3 \sum_{2^j n \in [0, 1]^2} |\langle f, \psi_{j,n}^l \rangle|^2 \leq 3B^2 \|f\|_{\tilde{\mathbf{C}}^\alpha}^2 \sum_{j=-\infty}^k 2^{-2j} 2^{j(2\alpha+2)} \\ &= \frac{3B^2 \|f\|_{\tilde{\mathbf{C}}^\alpha}^2 2^{2\alpha k}}{1 - 2^{-2\alpha}}. \end{aligned}$$

For  $2k = -\lfloor \log_2 N \rfloor$ , we derive that  $\varepsilon_l(N, f) = O(\|f\|_{\tilde{\mathbf{C}}^\alpha}^2 2^{2\alpha k}) = O(\|f\|_{\tilde{\mathbf{C}}^\alpha}^2 N^{-\alpha})$ . ■

One can prove that this decay rate is optimal in the sense that no approximation scheme can improve decay rate  $N^{-\alpha}$  over all uniformly Lipschitz  $\alpha$  functions [20].

### NonLinear Approximation of Piecewise Regular Images

If an image has singularities, then linear wavelet approximations introduce large errors. In one dimension, an isolated discontinuity creates a constant number of large wavelet coefficients at each scale. As a result, nonlinear wavelet approximations are marginally influenced by a finite number of isolated singularities. Theorem 9.12 proves that if  $f$  is uniformly Lipschitz  $\alpha$  between these singularities, then the asymptotic error decay behaves as if there was no singularity. In two dimensions, if  $f$  is uniformly Lipschitz  $\alpha$ , then Theorem 9.16 proves that the linear approximation error from  $M$  wavelets satisfies  $\varepsilon_l(M, f) = O(M^{-\alpha})$ . Thus, we can

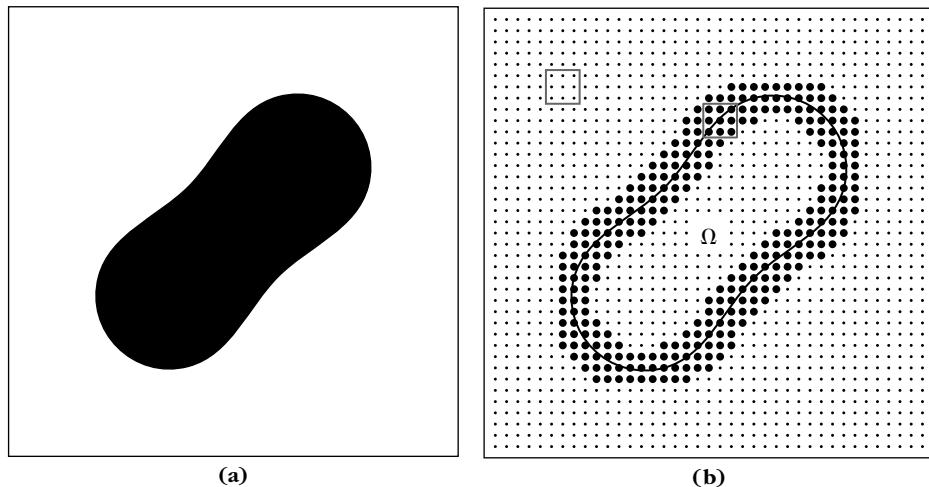
hope that piecewise regular images yield a nonlinear approximation error having the same asymptotic decay. Regretfully, this is wrong.

A piecewise regular image has discontinuities along curves of dimension 1, which create a nonnegligible number of high-amplitude wavelet coefficients. As a result, even though the function may be infinitely differentiable between discontinuities, the nonlinear approximation error  $\varepsilon_n(M, f)$  decays only like  $M^{-1}$ .

As an example, suppose that  $f = C \mathbf{1}_\Omega$  is the indicator function of a set  $\Omega$  the border  $\partial\Omega$  of which has a finite length, as shown in Figure 9.3. If the support of  $\psi_{j,n}^l$  does not intersect the border  $\partial\Omega$ , then  $\langle f, \psi_{j,n}^l \rangle = 0$  because  $f$  is constant over the support of  $\psi_{j,n}^l$ . The wavelets  $\psi_{j,n}^l$  have a square support of size proportional to  $2^j$ , which is translated on a grid of interval  $2^j$ . Since  $\partial\Omega$  has a finite length  $L$ , there are on the order of  $L 2^{-j}$  wavelets with supports that intersect  $\partial\Omega$ . Figure 9.3(b) illustrates the position of these coefficients.

Since  $f$  is bounded, the result from (9.57) for  $\alpha = 0$  is that  $|\langle f, \psi_{j,n}^l \rangle| = O(C 2^j)$ . Along the border, wavelet coefficients typically have an amplitude of  $|\langle f, \psi_{j,n}^l \rangle| \sim C 2^j$ . Thus, the  $M$  largest coefficients are typically at scales  $2^j \geq L/M$ . Selecting these  $M$  largest coefficients yields an error of

$$\varepsilon_n(M, f) \sim \sum_{j=-\infty}^{\log_2(L/M)-1} L 2^{-j} C^2 2^{2j} = (C L)^2 M^{-1}. \quad (9.58)$$



**FIGURE 9.3**

(a) Image  $f = \mathbf{1}_\Omega$ . (b) At the scale  $2^j$ , the wavelets  $\psi_{j,n}^l$  have a square support of width proportional to  $2^j$ . This support is translated on a grid of interval  $2^j$ , which is indicated by the smaller dots. The darker dots correspond to wavelets with support that intersect the frontier of  $\Omega$ , for which  $\langle f, \psi_{j,n}^l \rangle \neq 0$ .

Thus, the large number of wavelet coefficients produced by the edges of  $f$  limit the error decay to  $M^{-1}$ .

Two questions then arise. Is the class of images that have a wavelet approximation error that decays like  $M^{-1}$  sufficiently large enough to incorporate interesting image models? The following section proves that this class includes all bounded variation images. The second question is: Is it possible to find sparse signal representations that are better than wavelets to approximate images having regular edges  $z$ . We address this question in Section 9.3.2.

### Bounded Variation Images

Bounded variation functions provide good models for large classes of images that do not have irregular textures. The total variation of  $f$  is defined in Section 2.3.3 by

$$\|f\|_V = \int_0^1 \int_0^1 |\vec{\nabla} f(x)| dx. \quad (9.59)$$

The partial derivatives of  $\vec{\nabla} f$  must be taken in the general sense of distributions in order to include discontinuous functions. Let  $\partial\Omega_t$  be the level set defined as the boundary of

$$\Omega_t = \{x \in \mathbb{R}^2 : f(x) > t\}.$$

Theorem 2.9 proves that the total variation depends on length  $H^1(\partial\Omega_t)$  of level sets:

$$\int_0^1 \int_0^1 |\vec{\nabla} f(x)| dx = \int_{-\infty}^{+\infty} H^1(\partial\Omega_t) dt. \quad (9.60)$$

It results that if  $f = C \mathbf{1}_\Omega$ , then  $\|f\|_V = C L$  where  $L$  is the length of the boundary of  $\Omega$ . Thus, indicator functions of sets have a bounded total variation that is proportional to the length of their “edges.”

Linear and nonlinear approximation errors of bounded variation images are computed by evaluating the decay of their wavelet coefficients across scales. We denote with  $f_B^r[k]$  the rank  $k$  wavelet coefficient of  $f$ , without including the  $2^{2J}$  scaling coefficients  $\langle f, \phi_{J,n}^2 \rangle$ . Theorem 9.17 gives upper and lower bounds on  $\|f\|_V$  from wavelet coefficients. Wavelets are supposed to have a compact support and need only one vanishing moment.

**Theorem 9.17:** Cohen, DeVore, Petrushev, Xu. There exist  $A, B_1, B_2 > 0$  such that if  $\|f\|_V < +\infty$ , then

$$\sum_{j=-\infty}^J \sum_{l=1}^3 \sum_{2^j n \in [0,1]^2} |\langle f, \psi_{j,n}^l \rangle| + \sum_{2^J n \in [0,1]^2} |\langle f, \phi_{J,n}^2 \rangle| \geq A \|f\|_V, \quad (9.61)$$

$$\sup_{\substack{-\infty < j \leq J \\ 1 \leq l \leq 3}} \left( \sum_{2^j n \in [0,1]^2} |\langle f, \psi_{j,n}^l \rangle| \right) \leq B_1 \|f\|_V, \quad (9.62)$$

and

$$|f_B^r[k]| \leq B_2 \|f\|_V k^{-1}. \quad (9.63)$$

**Proof.** In two dimensions, a wavelet total variation does not depend on scale and position.

Indeed, with a change of variable  $x' = 2^{-j}x - n$ , we get

$$\|\psi_{j,n}^l\|_V = \iint |\vec{\nabla} \psi_{j,n}^l(x)| dx = \iint |\vec{\nabla} \psi^l(x')| dx' = \|\psi^l\|_V.$$

Similarly,  $\|\phi_{j,n}^2\|_V = \|\phi^2\|_V$ . The inequalities (9.61) and (9.62) are proved with the same derivation steps as in Theorem 9.13 for one-dimensional bounded variation functions.

The proof of (9.63) is technical and can be found in [175]. The inequality (9.62) proves that wavelet coefficients have a bounded  $L^1$  norm at each scale  $2^j$ . It results from (9.39) in Theorem 9.10 that ranked wavelet coefficients at each scale  $2^j$  have a decay bounded by  $B_1 \|f\|_V k^{-1}$ . The inequality (9.63) is finer since it applies to the ranking of wavelet coefficients at all scales. ■

The Lena image is an example of finite-resolution approximation of a bounded variation image. Figure 9.4 shows that its sorted wavelet coefficients  $\log_2 |f_B^r[k]|$  decays with a slope that reaches  $-1$  as  $\log_2 k$  increases, which verifies that  $|f_B^r[k]| = O(k^{-1})$ . In contrast, the Mandrill image shown in Figure 10.7 does not have a bounded total variation because of the fur texture. As a consequence,  $\log_2 |f_B^r[k]|$  decays more slowly, in this case with a slope that reaches  $-0.65$ .

A function with finite total variation does not necessarily have a bounded amplitude, but images do have a bounded amplitude. Theorem 9.18 incorporates this hypothesis to compute linear approximation errors.

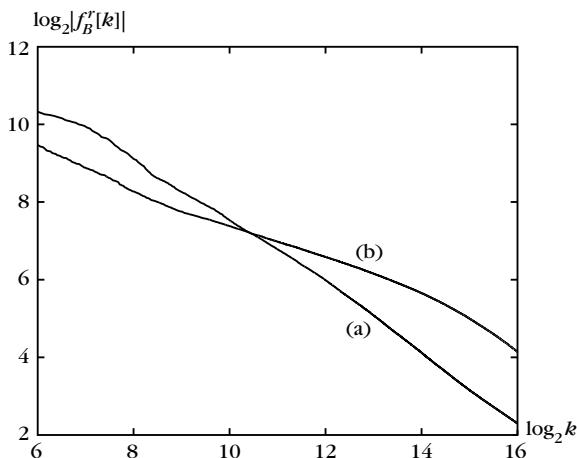


FIGURE 9.4

Sorted wavelet coefficients  $\log_2 |f_B^r[k]|$  as a function of  $\log_2 k$  for two images. (a) Lena image shown in Figure 9.5(a). (b) Mandrill image shown in Figure 10.7.

**Theorem 9.18.** If  $\|f\|_V < +\infty$  and  $\|f\|_\infty < +\infty$ , then

$$\varepsilon_l(M, f) = O(\|f\|_V \|f\|_\infty M^{-1/2}) \quad (9.64)$$

and

$$\varepsilon_n(M, f) = O(\|f\|_V^2 M^{-1}). \quad (9.65)$$

**Proof.** The linear approximation error from  $M = 2^{-2m}$  wavelets is

$$\varepsilon_l(2^{-2m}, f) = \sum_{j=-\infty}^m \sum_{l=1}^3 \sum_{2^j n \in [0, 1]^2} |\langle f, \psi_{j,n}^l \rangle|^2. \quad (9.66)$$

We shall verify that there exists  $B > 0$  such that for all  $j$  and  $l$ ,

$$\sum_{2^j n \in [0, 1]^2} |\langle f, \psi_{j,n}^l \rangle|^2 \leq B \|f\|_V \|f\|_\infty 2^j. \quad (9.67)$$

Applying this upper bound to the sum (9.66) proves that

$$\varepsilon_l(2^{-2m}, f) \leq 6B1 \|f\|_V \|f\|_\infty 2^m,$$

from which (9.64) is derived.

The upper bound (9.67) is calculated with (9.62), which shows that there exists  $B_2 > 0$  such that for all  $j$  and  $l$ ,

$$\sum_{2^j n \in [0, 1]^2} |\langle f, \psi_{j,n}^l \rangle| \leq B_2 \|f\|_V. \quad (9.68)$$

The amplitude of a wavelet coefficient can also be bounded:

$$|\langle f, \psi_{j,n}^l \rangle| \leq \|f\|_\infty \|\psi_{j,n}^l\|_1 = \|f\|_\infty 2^j \|\psi^l\|_1,$$

where  $\|\psi^l\|_1$  is the  $\mathbf{L}^1[0, 1]^2$  norm of  $\psi^l$ . If  $B_3 = \max_{1 \leq l \leq 3} \|\psi^l\|_1$ , this yields

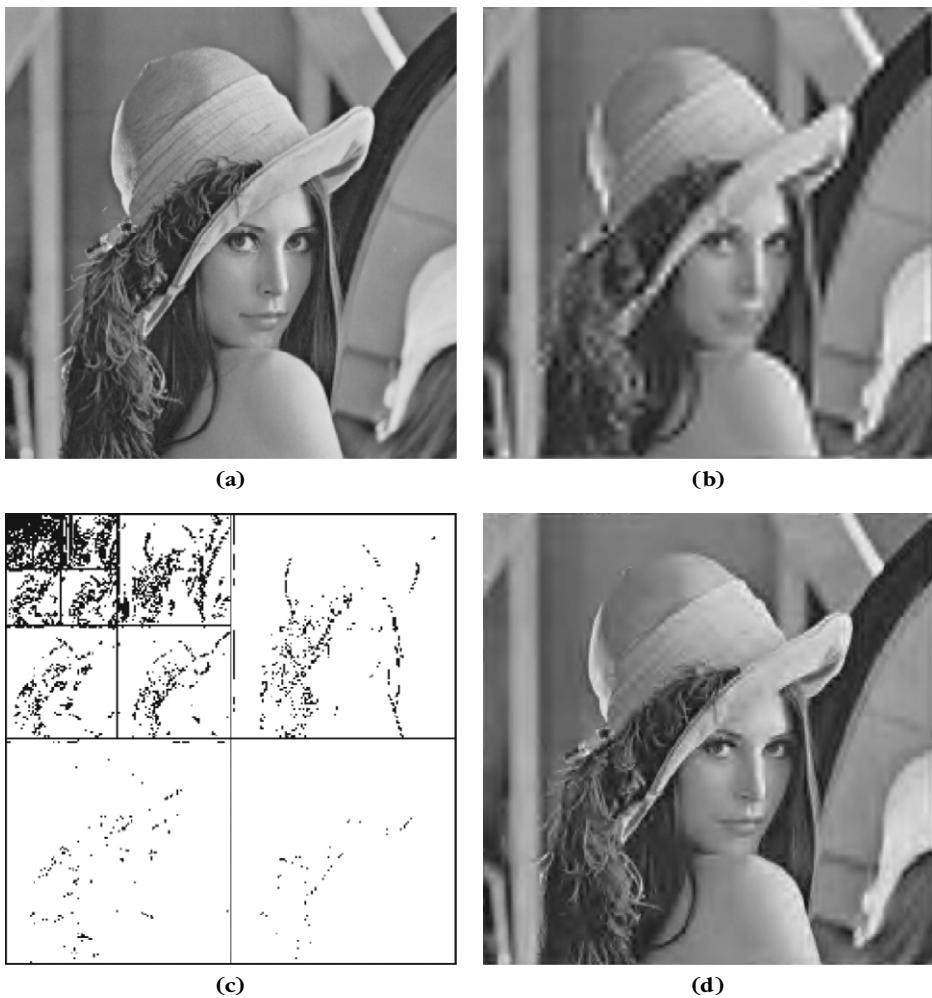
$$|\langle f, \psi_{j,n}^l \rangle| \leq B_3 2^j \|f\|_\infty. \quad (9.69)$$

Since  $\sum_n |a_n|^2 \leq \sup_n |a_n| \sum_n |a_n|$ , we get (9.67) from (9.68) and (9.69).

The nonlinear approximation error is a direct consequence of the sorted coefficient decay (9.63). It results from (9.63) and (9.36) that  $\varepsilon_n(M, f) = O(\|f\|_V^2 M^{-1})$ . ■

This theorem shows that nonlinear wavelet approximations of bounded variation images can yield much smaller errors than linear approximations. The decay bounds of this theorem are tight in the sense that one can find functions, for which  $\varepsilon_l(M, f)$  and  $\varepsilon_n(M, f)$  decay, respectively, like  $M^{-1/2}$  and  $M^{-1}$ . This is typically the case for bounded variation images including discontinuities along edges, for example,  $f = C \mathbf{1}_\Omega$ . The Lena image in Figure 9.5(a) is another example.

Figure 9.5(b) is a linear approximation calculated with  $M = N/16$  largest-scale wavelet coefficients. This approximation produces a uniform blur and creates Gibbs oscillations in the neighborhood of contours. Figure 9.5(c) gives the nonlinear

**FIGURE 9.5**

- (a) Lena image  $f$  of  $N = 256^2$  pixels. (b) Linear approximations  $f_M$  calculated from the  $M = N/16$  symmlet 4 wavelet coefficients at the largest scales:  $\|f - f_M\|/\|f\| = 0.036$ .  
 (c) The support of the  $M = N/16$  largest-amplitude wavelet coefficients are shown in black.  
 (d) Nonlinear approximation  $f_M$  calculated from the  $M$  largest-amplitude wavelet coefficients:  $\|f - f_M\|/\|f\| = 0.011$ .

approximation support  $\Lambda_T$  of  $M = N/16$  largest-scale wavelet coefficients. Large-amplitude coefficients are located where the image intensity varies sharply, in particular along the edges. The resulting nonlinear approximation is shown in Figure 9.5(d). The nonlinear approximation error is much smaller than the linear

approximation error— $\varepsilon_n(M, f) \leq \varepsilon_l(M, f)/10$ —and the image quality is indeed better. As in one dimension, this nonlinear wavelet approximation can be interpreted as an adaptive grid approximation, which refines the approximation resolution near edges and textures by keeping wavelet coefficients at smaller scales.

### 9.3.2 Geometric Image Models and Adaptive Triangulations

Bounded variation image models correspond to images that have level sets with a finite average length, but they do not imply any geometrical regularity of these level sets. The level sets and “edges” of many images such as Lena are often piecewise regular curves. This geometric regularity can be used to improve the sparsity of image representations.

When an image is uniformly Lipschitz  $\alpha$ , Theorem 9.16 proves that wavelet nonlinear approximations have an error  $\varepsilon_l(M, f) = O(M^{-\alpha})$ , which is optimal. However, as soon as the image is discontinuous along an edge, then the error decay rate drops to  $\varepsilon_n(M, f) = O(M^{-1})$ , because edges create a number of large wavelet coefficients that is proportional to their length. This decay rate is improved by representations taking advantage of edge geometric regularities. We introduce a piecewise regular image model that incorporates the geometric regularity of edges. Approximations of piecewise regular images are studied with adaptive triangulations.

#### Piecewise $C^\alpha$ Image Models

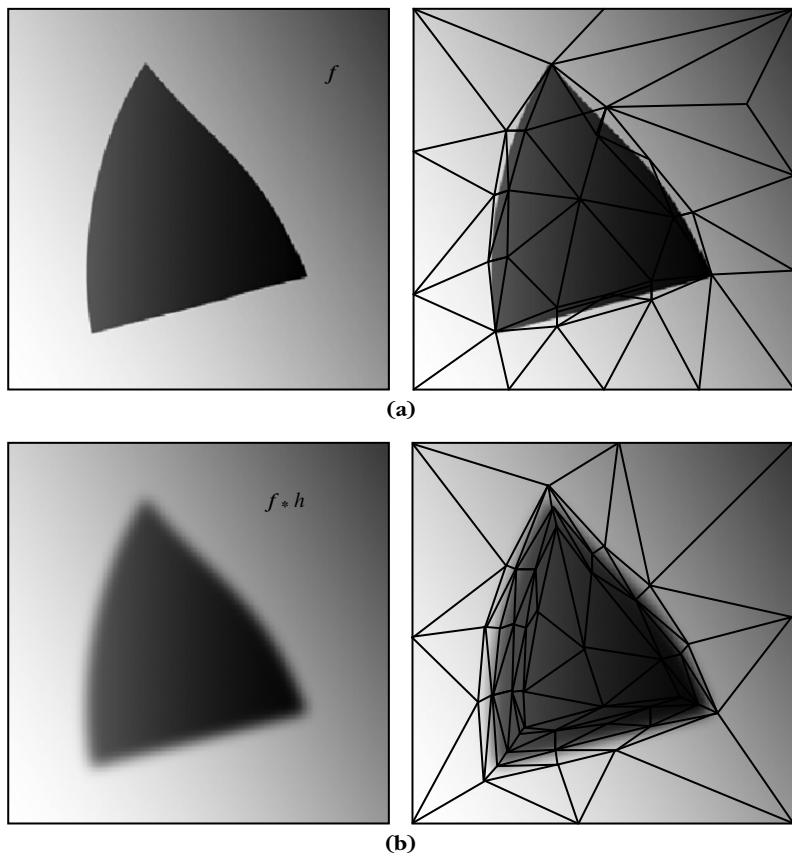
Piecewise regular image models include edges that are also piecewise regular. These edges are typically occlusion contours of objects in images. The regularity is measured in the sense of uniform Lipschitz regularity with Hölder norms  $\|f\|_{C^\alpha}$ , defined in (9.20) and (9.56) for one- and two-dimensional functions, respectively. Edges are supposed to be a finite union of curves  $e_k$  that are uniformly Lipschitz  $\alpha$  in  $[0, 1]^2$ ; between edges, the image is supposed to be uniformly Lipschitz  $\alpha$ . To model the blur introduced by the optics or by diffraction phenomena, the image model incorporates a convolution by an unknown regular kernel  $h_s$ , with scale  $s$  that is a parameter.

**Definition 9.1.** A function  $f \in L^2[0, 1]^2$  is said to be a piecewise  $C^\alpha$  with a blurring scale  $s \geq 0$ , if  $f = \bar{f} \star h_s$  where  $\bar{f}$  is uniformly Lipschitz  $\alpha$  on  $\Omega = [0, 1]^2 - \{e_k\}_{1 \leq k < K}$ . If  $s > 0$ , then  $h_s(x) = s^{-2}h(s^{-1}x)$  where  $h$  is a uniformly Lipschitz  $\alpha$  kernel with a support in  $[-1, 1]$ , and if  $s = 0$ , then  $h_0 = \delta$ . Curves  $e_k$  are uniformly Lipschitz  $\alpha$  and do not intersect tangentially.

When the blurring scale  $s = 0$ , then  $f = \bar{f} \star h_0 = \bar{f}$  is typically discontinuous along the edges. When  $s > 0$ , then  $\bar{f} \star h_s$  is blurred and discontinuities along edges are diffused on a neighborhood of size  $s$ , as shown in Figure 9.6.

#### Approximations with Adapted Triangulations

Wavelet approximations are inefficient to recover regular edges because it takes many wavelets to cover the edge at each scale, as shown by Figure 9.3. To improve

**FIGURE 9.6**

Adaptive triangulations for piecewise linear approximations of a piecewise  $\mathbf{C}^2$  image, without (a) and with (b) a blurring kernel.

this approximation, it is necessary to use elongated approximation elements. For  $\mathbf{C}^2$  piecewise regular images, it is proved that piecewise linear approximations over  $M$  adapted triangles can reach the optimal  $O(M^{-2})$  error decay, as if the image had no singularities.

For the solution of partial differential equations, the local optimization of anisotropic triangles was introduced by Babuška and Aziz [90]. Adaptive triangulations are indeed useful in numerical analysis, where shocks or boundary layers require anisotropic refinements of finite elements [77, 410, 434]. A planar triangulation  $(\mathcal{V}, \mathcal{T})$  of  $[0, 1]^2$  is composed of vertices  $\mathcal{V} = \{x_i\}_{0 \leq i < p}$  and disjoint triangular faces  $\mathcal{T} = \{T_k\}_{0 \leq k < M}$  that cover the image domain

$$\bigcup_{k=0}^{M-1} T_k = [0, 1]^2.$$

Let  $M$  be the number of triangles. We consider an image approximation  $\tilde{f}_M$  obtained with linear interpolations of the image values at the vertices. For each  $x_i \in \mathcal{V}$ ,  $\tilde{f}_M(x_i) = f(x_i)$ , and  $\tilde{f}_M(x)$  is linear on each triangular face  $T_k$ . A function  $f$  is well approximated with  $M$  triangles if the shapes of these triangles are optimized to capture the regularity of  $f$ .

Theorem 9.19 proves that adapted triangulations for piecewise  $C^2$  images leads to the optimal decay rate  $O(M^{-2})$  by sketching the construction of a well-adapted triangulation.

**Theorem 9.19.** If  $f$  is a piecewise  $C^2$  image, then there exists  $C$  such that for any  $M$  one can construct a triangulation  $(\mathcal{V}, \mathcal{T})$  with  $M$  triangles over which the piecewise linear interpolation  $\tilde{f}_M$  satisfies

$$\|f - \tilde{f}_M\|^2 \leq CM^{-2}. \quad (9.70)$$

**Proof.** A sketch of the proof is given. Let us first consider the case where the blurring scale is  $s = 0$ . Each edge curve  $e_k$  is  $C^2$  and can be covered by a band  $\beta_k$  of width  $\varepsilon^2$ . This band is a union of straight tubes of width  $\varepsilon^2$  and length  $\varepsilon/C$  where  $C$  is the maximum curvature of the edge curves. Each tube is decomposed in two elongated triangles as illustrated in Figure 9.6. The number of such triangles is  $2CL_k\varepsilon^{-1}$  where  $L_k$  is the length of  $e_k$ . Triangle vertices should also be appropriately adjusted at junctions or corners. Let  $L = \sum_k L_k$  be the total length of edges and  $\beta = \cup_k \beta_k$  be the band covering all edges. This band is divided in  $2CL\varepsilon^{-1}$  triangles. Since  $f$  is bounded,  $\|f - \tilde{f}_M\|_{L^2(\beta)}^2 \leq \text{area}(\beta) \|f\|_\infty^2 \leq L\varepsilon^2 \|f\|_\infty^2$ .

The complementary  $\beta^c = [0, 1]^2 - \beta$  is covered at the center by nearly equilateral triangles of surface of the order of  $\varepsilon$ , as shown in Figure 9.6. There are  $O(\varepsilon^{-1})$  such triangles. Performing this packing requires us to use a boundary layer of triangles that connect the large nearly isotropic triangles of width  $\varepsilon^{1/2}$  to the anisotropic triangles of size  $\sim \varepsilon^2 \times \varepsilon$ , along edges. One can verify that such a boundary layer can be constructed with  $O(LC\varepsilon^{-1})$  triangles. Since  $f$  is  $C^2$  on  $\Omega - \beta$ , the approximation error of a linear interpolation  $\tilde{f}$  over this triangulation is  $\|f - \tilde{f}_M\|_{L^\infty(\beta^c)} = O(\|f\|_{C^2(\beta^c)} \varepsilon^2)$ . Thus, the total error satisfies

$$\|f - \tilde{f}_M\|^2 = O(L\|f\|_\infty^2 \varepsilon^2 + \|f\|_{C^2}^2 \varepsilon^2)$$

with a number of triangles  $M = O((CL + 1)\varepsilon^{-1})$ , which verifies (9.70) for  $s = 0$ .

Suppose now that  $s > 0$ . According to Definition 9.1,  $f = \tilde{f} \star h_s$ , so edges are diffused into sharp transitions along a tube of width  $s$ . Since  $h$  is  $C^2$ , within this tube  $f$  is  $C^2$ , but it has large-amplitude derivatives if  $s$  is small. This defines an overall band  $\beta$  of surface  $Ls$  where the derivatives of  $f$  are potentially large. The triangulation of domain  $[0, 1]^2 - \beta$  can be treated similarly to the case  $s = 0$ . Thus, we concentrate on the triangulation of band  $\beta$  and show that one can find a triangulation with  $O(\varepsilon^{-1})$  triangles that yields an error in  $O(\varepsilon^2)$  over the band.

Band  $\beta$  has a surface  $Ls$  and can thus be covered by  $L\varepsilon^{-1}$  triangles of surface  $\varepsilon s$ . Let us compute the aspect ratio of these triangles to minimize the resulting error. A blurred piecewise  $C^2$  image  $f = \tilde{f} \star h_s$  has an anisotropic regularity at a point  $x$  close to an edge curve  $e_k$  where  $\tilde{f}$  is discontinuous. Let  $\tau_1(x)$  be the unit vector that is tangent to  $e_k$  at point  $x$ , and  $\tau_2(x)$  be the perpendicular vector. In the system of coordinates  $(\tau_1, \tau_2)$ , for

any  $u = u_1, \tau_1 + u_2 \tau_2$  in the neighborhood of  $x$ , one can prove that [342, 365]

$$\left| \frac{\partial^{i_1+i_2} f}{\partial u_1^{i_1} \partial u_2^{i_2}}(u) \right| = O(s^{-i_1/2-i_2}). \quad (9.71)$$

Thus, for  $s$  small the derivatives are much larger along  $\tau_2$  than along  $\tau_1$ . The error of local linear approximation can be computed with a Taylor decomposition

$$f(x + \Delta) = f(x) + \langle \nabla_x f, h \rangle + \frac{1}{2} \langle H_x(f) \Delta, \Delta \rangle + O(\|\Delta\|^2),$$

where  $H_x(f) \in \mathbb{R}^{2 \times 2}$  is the symmetric Hessian tensor of second derivatives. Let us decompose  $\Delta = \Delta_1 \tau_1 + \Delta_2 \tau_2$ :

$$|f(x + \Delta) - (f(x) + \langle \nabla_x f, \Delta \rangle)| = O(s^{-1} |\Delta_1|^2 + s^{-2} |\Delta_2|^2 + s^{-3/2} |\Delta_1| |\Delta_2|). \quad (9.72)$$

For a triangle of width and length  $\Delta_1$  and  $\Delta_2$ , we want to minimize the error for a given surface  $s\varepsilon \sim \Delta_1 \Delta_2$ , which is obtained with  $\Delta_2/\Delta_1 = \sqrt{s}$ . It results that  $\Delta_1 \sim s^{1/3} \varepsilon^{1/2}$  and  $\Delta_2 \sim s^{3/4} \varepsilon^{1/2}$ , which yields a linear approximation error (9.72) on this triangle:

$$|f(x + \Delta) - (f(x) + \langle \nabla_x f, \Delta \rangle)| = O(s^{-1/2} \varepsilon).$$

This gives

$$\|f - \tilde{f}_M\|_{L^2(\beta)}^2 \leq \text{area}(\beta) \|f - \tilde{f}_M\|_\infty^2 = O(L\varepsilon^2).$$

It proves that the  $O(\varepsilon^{-1})$  triangles produce an error of  $O(\varepsilon^2)$  on band  $\beta$ . Since the same result is valid on  $[0, 1]^2 - \beta$ , it yields (9.70). ■

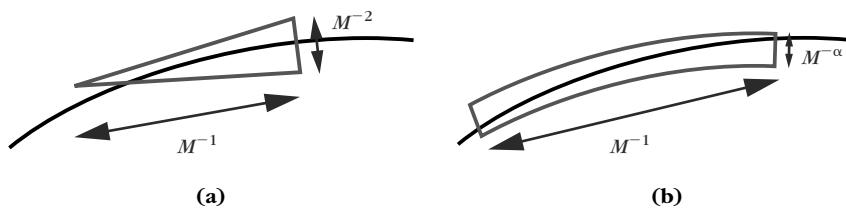
This theorem proves that an adaptive triangulation can yield an optimal decay rate  $O(M^{-2})$  for a piecewise  $C^2$  image. The decay rate of the error is independent from the blurring scale  $s$ , which may be zero or not. However, the triangulation depends on  $s$  and on the edge geometry.

Wherever  $f$  is uniformly  $C^2$ , it is approximated over large, nearly isotropic triangles of size  $O(M^{1/2})$ . In the neighborhood of edges, to introduce an error  $O(M^{-2})$ , triangles must be narrow in the direction of the discontinuity and as long as possible to cover the discontinuity with a minimum number of triangles. Since edges are  $C^2$ , they can be covered with triangles of width and length proportional to  $M^{-2}$  and  $M^{-1}$ , as illustrated in Figure 9.7.

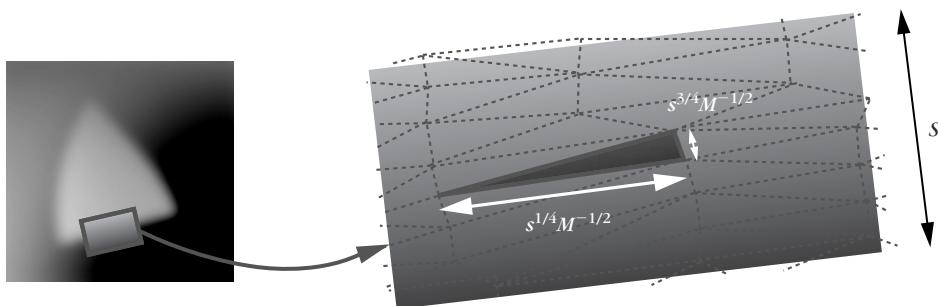
If the image is blurred at a scale  $s$ , then the discontinuities are a diffused neighborhood of size  $s$  where the image has sharp transitions. Theorem 9.19 shows that the tube of width  $s$  around each edge should be covered with triangles of a width and length proportional to  $s^{3/4} M^{-1/2}$  and  $s^{1/4} M^{-1/2}$ , as illustrated in Figure 9.8.

### Algorithms to Design Adapted Triangulations

It is difficult to adapt the triangulation according to Theorem 9.19 because the geometry of edges and the blurring scale  $s$  are unknown. There is currently no adaptive triangulation algorithm that guarantees finding an approximation with an error decay of  $O(M^{-2})$  for all piecewise  $C^2$  images. Most algorithms use greedy strategies

**FIGURE 9.7**

**(a)** Adapted triangle and **(b)** finite element to approximate a discontinuous function around a singularity curve.

**FIGURE 9.8**

Aspect ratio of triangles for the approximation of a blurred contour.

that iteratively construct a triangulation by progressively reducing the approximation error. Some algorithms progressively increase the number of triangles, while others decimate a dense triangulation.

Delaunay refinement algorithms, introduced by Ruppert [421] and Chew [160], proceed by iteratively inserting points to improve the triangulation precision. Each point is added at a circumcenter of one triangle, and is chosen to reduce the approximation error as much as possible. These algorithms control the shape of triangles and are used to compute isotropic triangulations of images where the size of the triangles varies to match the local image regularity. Extensions of these vertex insertion methods capture anisotropic regularity by locally modifying the notion of circumcenter [337] or with a local optimization of the vertex location [77].

Triangulation-thinning algorithms start with a dense triangulation of the domain and progressively remove either a vertex, an edge, or a face to increase the approximation error as slowly as possible until  $M$  triangles remain [238, 268, 305]. These methods do not control the shape of the resulting triangles, but can create effective anisotropic approximation meshes. They have been used for image compression [205, 239].

These adaptive triangulations are most often used to mesh the interior of a two- or three-dimensional domain having a complex boundary, or to mesh a two-dimensional surface embedded in three-dimensional space.

### Approximation of Piecewise $C^\alpha$ Images

Adaptive triangulations could be generalized to higher-order approximations of piecewise  $C^\alpha$  images to obtain an error in  $O(M^{-\alpha})$  with  $M$  finite elements for  $\alpha > 2$ . This would require computing polynomial approximations of order  $p = \lceil \alpha \rceil - 1$  on each finite element, and the support of these finite elements should also approximate the geometry of edges at an order  $p$ . To produce an error  $\|f - f_M\|^2 = O(M^{-\alpha})$ , it is indeed necessary to cover edge curves with  $O(M)$  finite elements of width  $O(M^{-\alpha})$ , as illustrated in Figure 9.7. However, this gets extremely complicated and has never been implemented numerically. Section 12.2.4 introduces bandlet approximations, which reach the  $O(M^{-\alpha})$  error decay for piecewise  $C^\alpha$  images, by choosing a best basis in a dictionary of orthonormal bases.

### 9.3.3 Curvelet Approximations

Candès and Donoho [135] showed that a simple thresholding of curvelet coefficients yields nearly optimal approximations of piecewise  $C^2$  images, as opposed to a complex triangulation optimization.

Section 5.5.2 introduced curvelets, defined with a translation, rotation, and anisotropic stretching

$$c_{2j,u}^\theta(x_1, x_2) = c_{2j}(R_\theta(x - u)) \quad \text{where} \quad c_{2j}(x_1, x_2) \approx 2^{-3j/4} c(2^{-j/2}x_1, 2^{-j}x_2), \quad (9.73)$$

where  $R_\theta$  is the planar rotation of angle  $\theta$ . A curvelet  $c_{j,m}^\theta$  is elongated in the direction  $\theta$ , with a *width* proportional to its *length*<sup>2</sup>. This parabolic scaling corresponds to the scaling of triangles used by Theorem 9.19 to approximate a discontinuous image along  $C^2$  edges.

A tight frame  $\mathcal{D} = \{c_{j,m}^\theta\}_{j,m,\theta}$  of curvelets  $c_{j,m}^\theta(x) = c_{2j}^\theta(x - u_m^{(j,\theta)})$  is obtained with  $2^{-\lfloor j/2 \rfloor + 2}$  equispaced angles  $\theta$  at each scale  $2^j$ , and a translation grid defined by

$$\forall m = (m_1, m_2) \in \mathbb{Z}^2, \quad u_m^{(j,\theta)} = R_\theta(2^{j/2}m_1, 2^j m_2). \quad (9.74)$$

This construction yields a tight frame of  $\mathbf{L}^2[0, 1]^2$  with periodic boundary conditions [135].

An  $M$ -term thresholding curvelet approximation is defined by

$$f_M = \sum_{(\theta,j,m) \in \Lambda_T} \langle f, c_{j,m}^\theta \rangle c_{j,m}^\theta \quad \text{with} \quad \Lambda_T = \left\{ (j, \theta, m) : |\langle f, c_{j,m}^\theta \rangle| > T \right\},$$

where  $M = |\Lambda_T|$  is the number of approximation curvelets. Since curvelets define a tight frame with a frame bound of  $A > 0$ ,

$$\|f - f_M\|^2 \leq A^{-1} \sum_{(\theta,j,m) \notin \Lambda_T} |\langle f, c_{j,m}^\theta \rangle|^2. \quad (9.75)$$

Although the frame is tight, this is not an equality because the thresholded curvelet coefficients of  $f$  are not the curvelet coefficients of  $f_M$  due to the frame redundancy. Theorem 9.20 shows that a thresholding curvelets approximation of a piecewise  $\mathbf{C}^2$  image has an error with a nearly optimal asymptotic decay.

**Theorem 9.20:** *Candès, Donoho.* Let  $f$  be a piecewise  $\mathbf{C}^2$  image. An  $M$ -term curvelet approximation satisfies

$$\|f - f_M\|^2 = O(M^{-2}(\log M)^3). \quad (9.76)$$

**Proof.** The detailed proof can be found in [135]. We give the main ideas by analyzing how curvelet atoms interact with regular parts and with edges in a piecewise regular image. The blurring  $f = \bar{f} * h_s$  is nearly equivalent to translating curvelet coefficients from a scale  $2^j$  to a scale  $2^j + s$ , and thus does not introduce any difficulty in the approximation. In the following, we suppose that  $s = 0$ .

The approximation error (9.76) is computed with an upper bound on the energy of curvelet coefficients (9.75). This sum is divided in three sets of curvelet coefficients. Type I curvelets have a support mostly concentrated where the image is uniformly  $\mathbf{C}^2$ ; these coefficients are small. Type II curvelets are in a neighborhood of an edge, and are elongated along the direction of the tangent to the edge; these coefficients are large. Type III curvelets are in a neighborhood of an edge with an angle different from the tangent angle; these coefficients get quickly small as the difference of angle increases. An upper bound of the error is computed by selecting the largest  $M/3$  curvelet coefficients for each type of curvelet coefficient and by computing the energy of the leftover curvelet coefficients for each type of coefficient.

A type I curvelet is located at a position  $u_m^{(j,\theta)}$  at a distance larger than  $K 2^{j/2}$  from edges. Since curvelets have vanishing moments, keeping type I curvelets at scales larger than  $2^j$  is equivalent to implementing a linear wavelet approximation at a scale  $2^j$ . Theorem 9.16 shows that for linear wavelet approximations of  $\mathbf{C}^2$  images, keeping  $M/3$  larger-scale coefficients yields an error that decreases like  $O(M^{-2})$ .

For type II curvelets in the neighborhood of an edge, with an angle  $\theta$  aligned with the local direction of the tangent to the edge, the sampling interval in this direction is also  $2^{j/2}$ . Thus, there are  $O(L 2^{-j/2})$  type II curvelets along the edge curves of length  $L$  in the image. These curvelet coefficients are bounded:

$$|\langle f, c_{j,m}^\theta \rangle| \leq \|f\|_\infty \|c_{2^j}\|_1 = \|f\|_\infty O(2^{3j/4})$$

because  $c_{2^j}(x_1, x_2) \approx 2^{-3j/4} c(2^{-j}x_1, 2^{-j/2}x_2)$ . If we keep the  $M/3$  larger-scale type II curvelets, since there are  $O(L 2^{-j/2})$  such curvelets at each scale, it amounts to keeping them at scales above  $2^l = O(L^2 M^{-2})$ . The leftover type II curvelets have an energy  $O(L 2^{-l/2}) \|f\|_\infty^2 O(2^{3l/2}) = O(M^{-2})$ .

Type III curvelet coefficients  $|\langle f, c_{j,m}^\theta \rangle|$  are located in the neighborhood of an edge with an angle  $\theta$  that deviates from the local orientation  $\theta_0$  of the edge tangent. The coefficients have a fast decay as  $|\theta - \theta_0|$  increases. This is shown in the Fourier domain by verifying that the Fourier domain where the energy of the localized edge patch is concentrated becomes increasingly disjoint from the Fourier support of  $\hat{c}_{j,m}^\theta$  as  $|\theta - \theta_0|$  increases. The error analysis of these curvelet coefficients is the most technical aspect of the proof. One can prove that selecting the  $M/3$  type III largest curvelets yields an error

among type III curvelets that is  $O(M^{-2}(\log_2 M)^3)$  [135]. This error dominates the overall approximation error and yields (9.76). ■

The curvelet approximation rate is optimal up to a  $(\log_2 M)^3$  factor. The beauty of this result comes from its simplicity. Unlike an optimal triangulation that must adapt the geometry of each element, curvelets define a fixed frame with coefficients that are selected by a simple thresholding. However, this simplicity has a downside. Curvelet approximations are optimal for piecewise  $C^\alpha$  images for  $\alpha = 2$ , but they are not optimal if  $\alpha > 2$  or  $\alpha < 1$ . In particular, curvelets are not optimal for bounded variation functions and their nonlinear approximation error does not have the  $M^{-1}$  decay of wavelets. Irregular geometric structures and isolated singularities are approximated with more curvelets than wavelets. Section 12.2.4 studies approximations in bandlet dictionaries, which are adapted to the unknown geometric image regularity.

In most images, curvelet frame approximations are not as effective as wavelet orthonormal bases because they have a redundancy factor  $A$  that is at least 5, and because most images include some structures that are more irregular than just piecewise  $C^2$  elements. Section 11.3.2 describes curvelet applications to noise removal.

---

## 9.4 EXERCISES

- 9.1** <sup>2</sup> Suppose that  $\{g_m\}_{m \in \mathbb{Z}}$  and  $\{\tilde{g}_m\}_{m \in \mathbb{Z}}$  are two dual frames of  $L^2[0, 1]$ .
- Let  $f_N = \sum_{m=0}^{N-1} \langle f, g_m \rangle \tilde{g}_m$ . Prove that the result (9.4) of Theorem 9.1 remains valid.
  - Let  $\langle f, g_{m_k} \rangle$  be the coefficient of rank  $k$ :  $|\langle f, g_{m_k} \rangle| \geq |\langle f, g_{m_{k+1}} \rangle|$ . Prove that if  $|\langle f, g_{m_k} \rangle| = O(k^{-s})$  with  $s > 1/2$ , then the best approximation  $f_M$  of  $f$  with  $M$  frame vectors satisfies  $\|f - f_M\|^2 = O(M^{1-2s})$ .
- 9.2** <sup>1</sup> *Color images.* A color pixel is represented by red, green, and blue components  $(r, g, b)$ , which are considered as orthogonal coordinates in a three-dimensional color space. The red  $r[n_1, n_2]$ , green  $g[n_1, n_2]$ , and blue  $b[n_1, n_2]$  image pixels are modeled as values taken by, respectively, three random variables  $R$ ,  $G$ , and  $B$ , which are the three coordinates of a color vector. Estimate numerically the  $3 \times 3$  covariance matrix of this color random vector from several images and compute the Karhunen-Loëve basis that diagonalizes it. Compare the color images reconstructed from the two Karhunen-Loëve color channels of highest variance with a reconstruction from the red and green channels.
- 9.3** <sup>2</sup> Suppose that  $\mathcal{B} = \{g_m\}_{m \in \mathbb{Z}}$  and  $\tilde{\mathcal{B}} = \{\tilde{g}_m\}_{m \in \mathbb{Z}}$  are two dual frames of  $L^2[0, 1]$ . Let  $f_N = \sum_{m=0}^{N-1} \langle f, g_m \rangle \tilde{g}_m$ . Prove that the result (9.4) of Theorem 9.1 remains valid even though  $\mathcal{B}$  is not an orthonormal basis.

- 9.4** <sup>2</sup> Let  $\vec{f} = (f_k)_{0 \leq k < K}$  be a multichannel signal where each  $f_k$  is a signal of size  $N$ . We write  $\|\vec{f}\|_F^2 = \sum_{k=0}^{K-1} \|f_k\|^2$ . Let  $\vec{f}_M = (f_{k,M})_{0 \leq k < K}$  be the multichannel signal obtained by projecting all  $f_k$  on the same  $M$  vectors of an orthonormal basis  $\mathcal{B} = \{g_m\}_{0 \leq m < N}$ . Prove that the best  $M$ -term approximation  $\vec{f}_M$  that minimizes  $\|\vec{f} - \vec{f}_M\|_F^2$  is obtained by selecting the  $M$  vectors  $g_m \in \mathcal{B}$  that maximize  $\sum_{k=0}^{K-1} |\langle f_k, g_m \rangle|^2$ .
- 9.5** <sup>1</sup> Verify that for  $f = C \mathbf{1}_{[0,1/2]}$ , a linear approximation with the  $N$  largest scale wavelets over  $[0, 1]$  produces an error that satisfies  $\varepsilon_l(N, f) \sim \|f\|_V^2 N^{-1}$ .
- 9.6** <sup>2</sup> Prove that for any  $f \in \mathbf{L}^2[0, 1]$ , if  $\|f\|_V < +\infty$ , then  $\|f\|_\infty < +\infty$ . Verify that one can find an image  $f \in \mathbf{L}^2[0, 1]^2$  such that  $\|f\|_V < +\infty$  and  $\|f\|_\infty = +\infty$ .
- 9.7** <sup>2</sup> Prove that if  $f \in \mathbf{W}^s(\mathbb{R})$  with  $s > p + 1/2$ , then  $f \in \mathbf{C}^p$ .
- 9.8** <sup>2</sup> The family of discrete polynomials  $\{p_k[n] = n^k\}_{0 \leq k < N}$  is a basis of  $\mathbb{C}^N$ .
  - Implement numerically a Gram-Schmidt algorithm that orthogonalizes  $\{p_k\}_{0 \leq k < N}$ .
  - Let  $f$  be a signal of size  $N$ . Compute the polynomial  $f_k$  of degree  $k$  that minimizes  $\|f - f_k\|$ . Perform numerical experiments on signals  $f$  that are uniformly smooth and piecewise smooth. Compare the approximation error with the error obtained by approximating  $f$  with the  $k$  lower-frequency Fourier coefficients.
- 9.9** <sup>3</sup> Let  $f$  be a function with a finite total variation  $\|f\|_V$  on  $[0, 1]$ . For a quantization step  $\Delta$ ,  $[0, 1]$  is divided into consecutive intervals  $[t_k, t_{k+1}]$  with  $m\Delta \leq f(t) \leq (m+1)\Delta$  for  $t \in [t_k, t_{k+1}]$ . In each  $[t_k, t_{k+1}]$ ,  $f(t)$  is approximated by its average  $a_k = (t_{k+1} - t_k)^{-1} \int_{t_k}^{t_{k+1}} f(t) dt$ . Let  $M$  be the total number of such intervals and  $f_M(t) = a_k$  for  $t \in [t_k, t_{k+1}]$  be a piecewise constant approximation of  $f$ . Prove that there exists a constant  $C$  such that
- $$\|f - f_M\|^2 \leq C \|f\|_V^2 M^{-2}.$$
- 9.10** <sup>3</sup> Let  $\alpha[M]$  be a decreasing sequence such that  $\lim_{M \rightarrow +\infty} \alpha[M] = 0$ . By using (9.61) prove that there exists a bounded variation function  $f \in \mathbf{L}^2[0, 1]^2$  such that  $\varepsilon_l(f, M) \geq \alpha[M]$  (the amplitude of  $f$  is not bounded).
- 9.11** <sup>1</sup> Consider a wavelet basis of  $\mathbf{L}^2[0, 1]$  constructed with wavelets having  $q > s$  vanishing moments and that are  $\mathbf{C}^q$ . Construct functions  $f \in \mathbf{W}^s[0, 1]$  for which the linear and nonlinear approximation errors in this basis are identical:  $\varepsilon_l(f, M) = \varepsilon_n(f, M)$  for any  $M \geq 0$ .
- 9.12** <sup>2</sup> Let  $f(t)$  be a piecewise polynomial signal of degree 3 defined on  $[0, 1]$ , with  $K$  discontinuities. We denote by  $f_{l,M}$  and  $f_{n,M}$ , respectively, the linear and nonlinear approximations of  $f$  from  $M$  vectors chosen from a Daubechies wavelet basis of  $\mathbf{L}^2[0, 1]$ , with four vanishing moments.
  - Give upper bounds as a function of  $K$  and  $M$  of  $\|f - f_{l,M}\|^2$  and  $\|f - f_{n,M}\|^2$ .

- (b) The Piece-polynomial signal  $f$  in WAVELAB is piecewise polynomial with degree 3. Decompose it in a Daubechies wavelet basis with four vanishing moments, and compute  $\|f - f_K\|$  and  $\|f - \tilde{f}_K\|$  as a function of  $K$ . Verify your analytic formula.

**9.13** <sup>2</sup> Let  $f[n]$  be defined over  $[0, N]$ . We denote by  $f_{p,k}[n]$  the signal that is piecewise constant on  $[0, k]$ , takes at most  $p$  different values, and minimizes

$$\varepsilon_{p,k} = \|f - f_{p,k}\|_{[0,k]}^2 = \sum_{n=0}^k |f[n] - f_{p,k}[n]|^2.$$

- (a) Compute as a function of  $f[n]$  the value  $a_{l,k}$  that minimizes  $c_{l,k} = \sum_{n=l}^k |f[n] - a_{l,k}|^2$ .
- (b) Prove that

$$\varepsilon_{p,k} = \min_{l \in [0, k-1]} \{\varepsilon_{p-1,l} + c_{l,k}\}.$$

Derive a bottom-up algorithm that computes progressively  $f_{p,k}$  for  $0 \leq k \leq N$  and  $1 \leq p \leq K$ , and obtains  $f_{K,N}$  with  $O(KN^2)$  operations. Implement this algorithm in WAVELAB.

- (c) Compute the nonlinear approximation of  $f$  with the  $K$  largest-amplitude Haar wavelet coefficients and the resulting approximation error. Compare this error with  $\|f - f_{K,N}\|$  as a function of  $K$  for the lady and the Piece-polynomial signals in WAVELAB. Explain your results.

**9.14** <sup>2</sup> *Approximation of oscillatory functions:*

- (a) Let  $f(t) = a(t) \exp[i\phi(t)]$ . If  $a(t)$  and  $\phi'(t)$  remain nearly constant on the support of  $\psi_{j,n}$ , then show with an approximate calculation that

$$\langle f, \psi_{j,n} \rangle \approx a(2^j n) \sqrt{2^j} \hat{\psi}(2^j \phi'(2^j n)). \quad (9.77)$$

- (b) Let  $f(t) = \sin t^{-1} \mathbf{1}_{[-1/\pi, 1/\pi]}(t)$ . Show that the  $\ell^p$  norm of the wavelet coefficients of  $f$  is finite if and only if  $p < 1$ . Use the approximate formula.
- (c) Compute an upper bound of the nonlinear approximation error  $\varepsilon_n(f, M)$  of  $\sin t^{-1}$  from  $M$  wavelet coefficients. Verify your theoretical estimate with a numerical calculation in WAVELAB.

# Compression

# 10

Reducing a liter of orange juice to a few grams of concentrated powder is what lossy compression is about. The taste of the restored beverage is similar to the taste of orange juice but has often lost some subtlety. In this book we are more interested in sounds and images, but we face the same trade-off between quality and compression. Saving money for data storage and improving transmissions through channels with limited bandwidth are major compression applications.

A transform coder decomposes a signal in an orthogonal basis and quantizes the decomposition coefficients. The distortion of the restored signal is minimized by optimizing the quantization, the basis, and the bit allocation. The basic information theory necessary for understanding quantization properties is introduced. Distortion rate theory is first studied at high bit rates, in a Bayes framework, where signals are realizations of a random vector that has a probability distribution that is known *a priori*. This applies to audio coding, where signals are often modeled with Gaussian processes.

High signal-compression factors are obtained in sparse representations, where few nonzero coefficients are kept. Most of the bits are devoted to code the geometry of these nonzero coefficients, and the distortion is dominated by the resulting nonlinear approximation term. Wavelet image transform codes illustrate these properties. JPEG and JPEG-2000 image-compression standards are described.

---

## 10.1 TRANSFORM CODING

When production models are available, as in speech signals, compression algorithms can code parameters that produce a signal approximation. When no such model is available, as in general audio signals or images, then transform codes provide efficient compression algorithms with sparse representations in orthonormal bases. Section 10.1.1 reviews different types of compression algorithms and Section 10.1.2 concentrates on transform codes.

### 10.1.1 Compression State of the Art

#### *Speech*

Speech coding is used for telephony, where it may be of limited quality but good intelligibility, and for higher-quality teleconferencing. Telephone speech is limited to the frequency band of 200–3400 Hz and is sampled at 8 kHz. A pulse code modulation (PCM), which quantizes each sample on 8 bits, produces a code with 64 kb/s ( $64 \cdot 10^3$  bits per second). This can be considerably reduced by removing some of the speech redundancy.

The production of speech signals is well understood. Model-based analysis-synthesis codes give intelligible speech at 2 kb/s. This is widely used for defense telecommunications [316, 460]. Digital cellular telephony uses 8 kb/s or less to reproduce more natural voices. Linear predictive codes (LPCs) restore speech signals by filtering white noise or a pulse train with linear filters defined by parameters that are estimated and coded. For higher bit rates, the quality of LPC speech production is enhanced by exciting the linear filters with waveforms chosen from a larger family. These code-excited linear prediction (CELP) codes provide nearly perfect telephone quality at 16 kb/s.

#### *Audio*

Audio signals include speech but also music and all types of sounds. On a compact disc, the audio signal is limited to a maximum frequency of 20 kHz. It is sampled at 44.1 kHz and each sample is coded on 16 bits. The bit rate of the resulting PCM code is 706 kb/s. For compact discs and digital audio tapes, signals must be coded with hardly any noticeable distortion. This is also true for multimedia CD-ROM and digital television sounds.

No models are available for general audio signals. At present, the best compression is achieved by transform coders that decompose the signal in a local time-frequency basis. To reduce perceived distortion, perceptual coders [317] adapt the quantization of time-frequency coefficients to our hearing sensitivity. Compact disc-quality sounds are restored with 128 kb/s; nearly perfect audio signals are obtained with 64 kb/s.

#### *Images*

A gray-level image typically has  $512 \times 512$  pixels, each coded with 8 bits. Like audio signals, images include many types of structures that are difficult to model. Currently, the best image-compression algorithms are the JPEG and JPEG-2000 compression standards, which are transform codes in cosine bases and wavelet bases.

The efficiency of these bases comes from their ability to construct precise nonlinear image approximations with few nonzero vectors. With fewer than 1 bit/pixel, visually perfect images are reconstructed. At 0.25 bit/pixel, the image remains of good quality.

### Video

Applications of digital video range from low-quality videophones, teleconferencing, and Internet video browsing, to high-resolution television. The most effective compression algorithms remove time redundancy with a motion compensation. Local image displacements are measured from one frame to the next, and are coded as motion vectors. Each frame is predicted from a previous one by compensating for the motion. An error image is calculated and compressed with a transform code. The MPEG video-compression standards are based on such motion compensation [344] with a JPEG-type compression of prediction error images.

Standard-definition television (SDTV) format has interlaced images of  $720 \times 426$  (NTSC) or  $720 \times 576$  pixels (PAL) with, respectively, 50 and 60 images per second. MPEG-2 codes these images with typically 5 Mb/s. Internet videos are often smaller images of  $320 \times 240$  pixels with typically 20 or 30 images per second, and are often coded with 200 to 300 kb/s for real-time browsing. The full high-definition television (HDTV) format corresponds to images of  $1920 \times 1080$  pixels. MPEG-2 codes these images with 12 to 24 Mb/s. With MPEG-4, the bit rate goes down to 12 Mb/s or less.

#### 10.1.2 Compression in Orthonormal Bases

A transform coder decomposes signals in an orthonormal basis  $\mathcal{B} = \{g_m\}_{0 \leq m < N}$  and optimizes the compression of the decomposition coefficients. The performance of such a transform code is first studied from a Bayes point of view, by supposing that the signal is the realization of a random process  $F[n]$  of size  $N$ , which has a probability distribution that is known a priori.

Let us decompose  $F$  over  $\mathcal{B}$ :

$$F = \sum_{m=0}^{N-1} F_{\mathcal{B}}[m] g_m.$$

Each coefficient  $F_{\mathcal{B}}[m]$  is a random variable defined by

$$F_{\mathcal{B}}[m] = \langle F, g_m \rangle = \sum_{n=0}^{N-1} F[n] g_m^*[n].$$

To center the variations of  $F_{\mathcal{B}}[m]$  at zero, we code  $F_{\mathcal{B}}[m] - E\{F_{\mathcal{B}}[m]\}$  and store the mean value  $E\{F_{\mathcal{B}}[m]\}$ . This is equivalent to supposing that  $F_{\mathcal{B}}[m]$  has a zero mean.

### Quantization

To construct a finite code, each coefficient  $F_{\mathcal{B}}[m]$  is approximated by a quantized variable  $\tilde{F}_{\mathcal{B}}[m]$ , which takes its values over a finite set of real numbers. A scalar quantization approximates each  $F_{\mathcal{B}}[m]$  independently. If the coefficients  $F_{\mathcal{B}}[m]$  are highly dependent, quantizer performance is improved by vector quantizers that approximate the vector of  $N$  coefficients  $\{F_{\mathcal{B}}[m]\}_{0 \leq m < N}$  together [27]. Scalar quantizers require fewer computations and are thus more often used. If the basis

$\{g_m\}_{0 \leq m < N}$  can be chosen so that the coefficients  $F_{\mathcal{B}}[m]$  are nearly independent, the improvement of a vector quantizer becomes marginal. After quantization, the reconstructed signal is

$$\tilde{F} = \sum_{m=0}^{N-1} \tilde{F}_{\mathcal{B}}[m] g_m.$$

### Distortion Rate

Let us evaluate the distortion introduced by this quantization. Ultimately, we want to restore a signal that is perceived as nearly identical to the original signal. Perceptual transform codes are optimized with respect to our sensitivity to degradations in audio signals and images [317]. However, distances that evaluate perceptual errors are highly nonlinear and thus difficult to manipulate mathematically. A mean-square norm often does not properly quantify the perceived distortion, but reducing a mean-square distortion generally enhances the coder performance. Weighted mean-square distances can provide better measurements of perceived errors and are optimized like a standard mean-square norm.

In the following, we try to minimize the average coding distortion, evaluated with a mean-square norm. Since the basis is orthogonal, this distortion can be written as

$$d = E\{\|F - \tilde{F}\|^2\} = \sum_{m=0}^{N-1} E\{|F_{\mathcal{B}}[m] - \tilde{F}_{\mathcal{B}}[m]|^2\}.$$

The average number of bits allocated to encode a quantized coefficient  $\tilde{F}_{\mathcal{B}}[m]$  is denoted as  $R_m$ . For a given  $R_m$ , a scalar quantizer is designed to minimize  $E\{|F_{\mathcal{B}}[m] - \tilde{F}_{\mathcal{B}}[m]|^2\}$ . The total mean-square distortion  $d$  depends on the average total bit budget

$$R = \sum_{m=0}^{N-1} R_m.$$

The function  $d(R)$  is called the *distortion rate*. For a given  $R$ , the bit allocation  $\{R_m\}_{0 \leq m < N}$  must be adjusted in order to minimize  $d(R)$ .

### Choice of Basis

The distortion rate of an optimized transform code depends on the orthonormal basis  $\mathcal{B}$ . We see in Section 10.3.2 that the Karhunen-Loëve basis minimizes  $d(R)$  for high-resolution quantizations of signals that are realizations of a Gaussian process. However, this is not true when the process is non-Gaussian.

To achieve a high compression rate, the transform code must produce many zero-quantized coefficients, and thus define a sparse signal representation. Section 10.4 shows that  $d(R)$  then depends on the precision of nonlinear approximations in the basis  $\mathcal{B}$ .

---

## 10.2 DISTORTION RATE OF QUANTIZATION

Quantized coefficients take their values over a finite set and can thus be coded with a finite number of bits. Section 10.2.1 reviews entropy codes of random sources. Section 10.2.2 studies the optimization of scalar quantizers, in order to reduce the mean-square error for a given bit allocation.

### 10.2.1 Entropy Coding

Let  $X$  be a random source that takes its values among a finite alphabet of  $K$  symbols  $\mathcal{A} = \{x_k\}_{1 \leq k \leq K}$ . The goal is to minimize the average bit rate needed to store the values of  $X$ . We consider codes that associate to each symbol  $x_k$  a binary word  $w_k$  of length  $l_k$ . A sequence of values produced by the source  $X$  is coded by aggregating the corresponding binary words.

All symbols  $x_k$  can be coded with binary words of the same size  $l_k = \lceil \log_2 K \rceil$  bits. However, the average code length may be reduced with a *variable-length code* using smaller binary words for symbols that occur frequently. Let us denote with  $p_k$  the probability of occurrence of a symbol  $x_k$ :

$$p_k = \Pr\{X = x_k\}.$$

The average bit rate to code each symbol emitted by the source  $X$  is

$$R_X = \sum_{k=1}^K l_k p_k. \quad (10.1)$$

We want to optimize the code words  $\{w_k\}_{1 \leq k \leq K}$  in order to minimize  $R_X$ .

#### **Prefix Code**

Codes with words of varying lengths are not always uniquely decodable. Let us consider the code that associates to  $\{x_k\}_{1 \leq k \leq 4}$  the code words

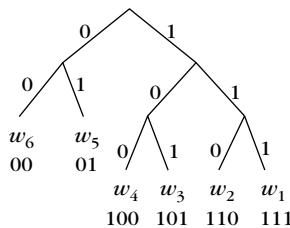
$$\{w_1 = 0, w_2 = 10, w_3 = 110, w_4 = 101\}. \quad (10.2)$$

The sequence 1010 can either correspond to  $w_2 w_2$  or to  $w_4 w_1$ . To guarantee that any aggregation of code words is uniquely decodable, the *prefix* condition imposes that no code word may be the prefix (beginning) of another one. The code (10.2) does not satisfy this condition since  $w_2$  is the prefix of  $w_4$ . The code

$$\{w_1 = 0, w_2 = 10, w_3 = 110, w_4 = 111\}$$

satisfies this prefix condition. Any code that satisfies the prefix condition is clearly uniquely decodable.

A prefix code is characterized by a binary tree that has  $K$  leaves corresponding to the symbols  $\{x_k\}_{1 \leq k \leq K}$ . Figure 10.1 shows an example for a prefix code of  $K = 6$  symbols. The left and right branches of the binary tree are, respectively, coded by 0 and 1. The binary code word  $w_k$  associated to  $x_k$  is the succession of 0 and 1

**FIGURE 10.1**

Prefix tree corresponding to a code with six symbols. The code word  $w_k$  of each leaf is indicated below it.

corresponding, respectively, to the left and right branches along the path from the root to the leaf  $x_k$ . The binary code produced by such a binary tree is always a prefix code. Indeed,  $w_m$  is a prefix of  $w_k$  if and only if  $x_m$  is an ancestor of  $x_k$  in the binary tree. This is not possible since both symbols correspond to a leaf of the tree. Conversely, we can verify that any prefix code can be represented by such a binary tree.

The length  $l_k$  of the code word  $w_k$  is the depth in the binary tree of the corresponding leaf. Thus, the optimization of a prefix code is equivalent to the construction of an optimal binary tree that distributes the depth of the leaves in order to minimize

$$R_X = \sum_{k=1}^K l_k p_k. \quad (10.3)$$

Therefore, higher-probability symbols should correspond to leaves higher in the tree.

### **Shannon Entropy**

The Shannon theorem [429] proves that entropy is a lower bound for the average bit rate  $R_X$  of any prefix code.

**Theorem 10.1:** *Shannon.* Let  $X$  be a source with symbols  $\{x_k\}_{1 \leq k \leq K}$  that occur with probabilities  $\{p_k\}_{1 \leq k \leq K}$ . The average bit rate  $R_X$  of a prefix code satisfies

$$R_X \geq \mathcal{H}(X) = - \sum_{k=1}^K p_k \log_2 p_k. \quad (10.4)$$

Moreover, there exists a prefix code such that

$$R_X \leq \mathcal{H}(X) + 1 \quad (10.5)$$

and  $\mathcal{H}(X)$  is called the *entropy* of  $X$ .

**Proof.** This theorem is based on the Kraft inequality given by Lemma 10.1.

**Lemma 10.1: Kraft.** Any prefix code satisfies

$$\sum_{k=1}^K 2^{-l_k} \leq 1. \quad (10.6)$$

Conversely, if  $\{l_k\}_{1 \leq k \leq K}$  is a positive sequence that satisfies (10.6), then a sequence of binary words  $\{w_k\}_{1 \leq k \leq K}$  of length  $\{l_k\}_{1 \leq k \leq K}$  exists that satisfies the prefix condition.

To prove (10.6), we construct a full binary tree  $T$  the leaves of which are at the depth  $m = \max\{l_1, l_2, \dots, l_K\}$ . Inside this tree, we can locate node  $n_k$  at depth  $l_k$  that codes the binary word  $w_k$ . We denote  $T_k$  to the subtree with the root of  $n_k$ , as illustrated in Figure 10.2. This subtree has a depth  $m - l_k$  and thus contains  $2^{m-l_k}$  nodes at the level  $m$  of  $T$ . There are  $2^m$  nodes at the depth  $m$  of  $T$  and the prefix condition implies that the subtrees  $T_1, \dots, T_K$  have no node in common, so

$$\sum_{k=1}^K 2^{m-l_k} \leq 2^m,$$

which proves (10.6).

Conversely, we consider  $\{l_k\}_{1 \leq k \leq K}$  that satisfies (10.6), with  $l_1 \leq l_2 \leq \dots \leq l_K$  and  $m = \max\{l_1, l_2, \dots, l_K\}$ . Again, we construct a full binary tree  $T$  with leaves at depth  $m$ . Let  $S_1$  be the  $2^{m-l_1}$  first nodes at level  $m$ ,  $S_2$  be the next  $2^{m-l_2}$  nodes, and so on, as illustrated in Figure 10.2. Since  $\sum_{k=1}^K 2^{m-l_k} \leq 2^m$ , the sets  $\{S_k\}_{1 \leq k \leq K}$  have fewer than  $2^m$  elements and can thus be constructed at level  $m$  of the tree. The nodes of a set  $S_k$  are the leaves of a subtree  $T_k$  of  $T$ . The root  $n_k$  of  $T_k$  is at depth  $l_k$  and corresponds to a binary word  $w_k$ . By construction, all these subtrees  $T_k$  are distinct, so  $\{w_k\}_{1 \leq k \leq K}$  is a prefix code where each code word  $w_k$  has a length  $l_k$ . This finishes the lemma proof.

To prove the two inequalities (10.4) and (10.5) of the theorem, we consider the minimization of

$$R_X = \sum_{k=1}^K p_k l_k$$

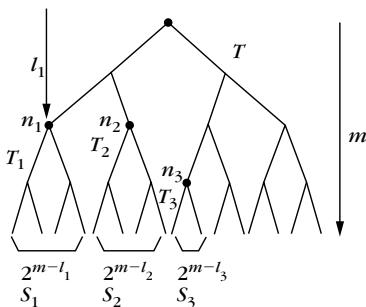


FIGURE 10.2

The leaves at depth  $m$  of tree  $T$  are regrouped as sets  $S_k$  of  $2^{m-l_k}$  nodes that are the leaves of tree  $T_k$ , having its root  $n_k$  at depth  $l_k$ . Here,  $m = 4$  and  $l_1 = 2$ , so  $S_1$  has  $2^2$  nodes.

under the Kraft inequality constraint

$$\sum_{k=1}^K 2^{-l_k} \leq 1.$$

If we admit noninteger values for  $l_k$ , we can verify with Lagrange multipliers that the minimum is reached for  $l_k = -\log_2 p_k$ . The value of this minimum is the entropy lower bound:

$$R_X = \sum_{k=1}^K p_k l_k = - \sum_{k=1}^K p_k \log_2 p_k = \mathcal{H}(X),$$

which proves (10.4).

To guarantee that  $l_k$  is an integer, the Shannon code is defined by

$$l_k = \lceil -\log_2 p_k \rceil, \quad (10.7)$$

where  $\lceil x \rceil$  is the smallest integer larger than  $x$ . Since  $l_k \geq -\log_2 p_k$ , the Kraft inequality is satisfied:

$$\sum_{k=1}^K 2^{-l_k} \leq \sum_{k=1}^K 2^{\log_2 p_k} = 1.$$

Lemma 10.1 proves that there exists a prefix code with binary words  $w_k$  that have length  $w_k$ . For this code,

$$R_X = \sum_{k=1}^K p_k l_k \leq \sum_{k=1}^K p_k (-\log_2 p_k + 1) = \mathcal{H}(X) + 1,$$

which proves (10.5). ■

The entropy  $\mathcal{H}(X)$  measures the uncertainty as to the outcome of the random variable  $X$ , and

$$0 \leq \mathcal{H}(X) \leq \log_2 K.$$

The maximum value  $\log_2 K$  corresponds to a sequence with a uniform probability distribution  $p_k = 1/K$  for  $1 \leq k \leq K$ . Since no value is more probable than any other, the uncertainty as to the outcome of  $X$  is maximum. The minimum entropy value  $\mathcal{H}(X) = 0$  corresponds to a source where one symbol  $x_k$  occurs with probability 1. There is no uncertainty as to the outcome of  $X$  because we know in advance that it will be equal to  $x_k$ .

### Huffman Code

The entropy lower bound  $\mathcal{H}(X)$  is nearly reachable with an optimized prefix code. The *Huffman algorithm* is a dynamical programming algorithm that constructs a binary tree that minimizes the average bit rate  $R_X = \sum_{k=1}^K p_k l_k$ . This tree is called

an *optimal prefix-code tree*. Theorem 10.2 gives an induction rule that constructs the tree from bottom up by aggregating lower-probability symbols.

**Theorem 10.2:** *Huffman.* Let us consider  $K$  symbols with their probability of occurrence sorted in increasing order  $p_k \leq p_{k+1}$ :

$$\{(x_1, p_1), (x_2, p_2), (x_3, p_3), \dots, (x_K, p_K)\}. \quad (10.8)$$

We aggregate the two lower-probability symbols  $x_1$  and  $x_2$  in a single symbol  $x_{1,2}$  of probability

$$p_{1,2} = p_1 + p_2.$$

An optimal prefix–code tree for the  $K$  symbols (10.8) is obtained by constructing an optimal prefix–code tree for the  $K - 1$  symbols,

$$\{(x_{1,2}, p_{1,2}), (x_3, p_3), \dots, (x_K, p_K)\}, \quad (10.9)$$

and by dividing the leaf  $x_{1,2}$  into two children nodes corresponding to  $x_1$  and  $x_2$ .

The proof of this theorem [27, 307] is left to the reader. The Huffman rule reduces the construction of an optimal prefix–code tree of  $K$  symbols (10.8) to the construction of an optimal code of  $K - 1$  symbols (10.9) plus an elementary operation. The Huffman algorithm iterates this regrouping  $K - 1$  times to grow an optimal prefix–code tree progressively from bottom to top. The Shannon theorem (10.1) proves that the average bit rate of the Huffman optimal prefix code satisfies

$$\mathcal{H}(X) \leq R_X \leq \mathcal{H}(X) + 1. \quad (10.10)$$

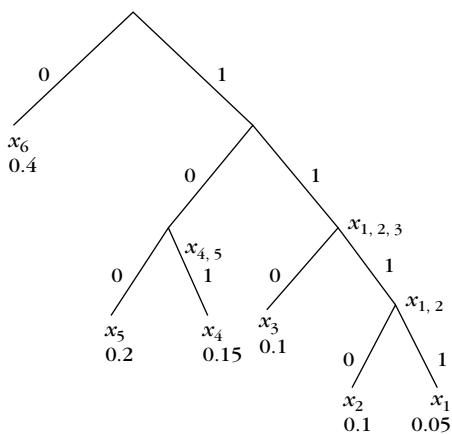
As explained in the proof of Theorem 10.1, the bit rate may be up to 1 bit more than the entropy lower bound because this lower bound is obtained with  $I_k = -\log_2 p_k$ , which is generally not possible since  $I_k$  must be an integer. In particular, lower bit rates are achieved when one symbol has a probability close to 1.

### EXAMPLE 10.1

We construct the Huffman code with six symbols  $\{x_k\}_{1 \leq k \leq 6}$  of probabilities

$$\{p_1 = 0.05, p_2 = 0.1, p_3 = 0.1, p_4 = 0.15, p_5 = 0.2, p_6 = 0.4\}.$$

The symbols  $x_1$  and  $x_2$  are the lower-probability symbols, which are regrouped in a symbol  $x_{1,2}$  with a probability of  $p_{1,2} = p_1 + p_2 = 0.15$ . At the next iteration, the lower probabilities are  $p_3 = 0.1$  and  $p_{1,2} = 0.15$ , so we regroup  $x_{1,2}$  and  $x_3$  in a symbol  $x_{1,2,3}$  with a probability of 0.25. The next two lower-probability symbols are  $x_4$  and  $x_5$ , which are regrouped in a symbol  $x_{4,5}$  with a probability of 0.35. We then group  $x_{4,5}$  and  $x_{1,2,3}$ , which yields  $x_{1,2,3,4,5}$  with a probability of 0.6, which is finally aggregated with  $x_6$ . This finishes the tree, as illustrated in Figure 10.3. The resulting average bit rate (10.3) is  $R_X = 2.35$ , whereas the entropy is

**FIGURE 10.3**

Prefix tree grown with the Huffman algorithm for a set of  $K = 6$  symbols  $x_k$ , with probabilities  $p_k$  indicated at the leaves of the tree.

$\mathcal{H}(X) = 2.28$ . This Huffman code is better than the prefix code in Figure 10.1, which has an average bit rate of  $R_X = 2.4$ .

### Block Coding

As mentioned earlier, the inequality (10.10) shows that a Huffman code may require 1 bit above the entropy because the length  $l_k$  of each binary word must be an integer, whereas the optimal value  $-\log_2 p_k$  is generally a real number. To reduce this overhead the symbols are coded together in blocks of size  $n$ .

Let us consider the block of  $n$  independent random variables  $\vec{X} = X_1, \dots, X_n$ , where each  $X_k$  takes its values in the alphabet  $\mathcal{A} = \{x_k\}_{1 \leq k \leq K}$  with the same probability distribution as  $X$ . The vector  $\vec{X}$  can be considered as a random variable taking its values in the alphabet  $\mathcal{A}^n$  of size  $K^n$ . To each block of symbols  $\vec{s} \in \mathcal{A}^n$ , we associate a binary word of length  $l(\vec{s})$ . The average number of bits per symbol for such a code is

$$R_X = \frac{1}{n} \sum_{\vec{s} \in \mathcal{A}^n} p(\vec{s}) l(\vec{s}).$$

Theorem 10.3 proves that the resulting Huffman code has a bit rate that converges to the entropy of  $X$  as  $n$  increases.

**Theorem 10.3.** The Huffman code for a block of size  $n$  requires an average number of bits per symbol that satisfies

$$\mathcal{H}(X) \leq R_X \leq \mathcal{H}(X) + \frac{1}{n}. \quad (10.11)$$

**Proof.** The entropy of  $\vec{X}$  considered as a random variable is

$$\mathcal{H}(\vec{X}) = \sum_{\vec{s} \in \mathcal{A}^n} p(\vec{s}) \log_2 p(\vec{s}).$$

Denote by  $R_{\vec{X}}$  the average number of bits to code each block  $\vec{X}$ . Applying (10.10) shows that with a Huffman code,  $R_{\vec{X}}$  satisfies

$$\mathcal{H}(\vec{X}) \leq R_{\vec{X}} \leq \mathcal{H}(\vec{X}) + 1. \quad (10.12)$$

Since the random variables  $X_i$  that compose  $\vec{X}$  are independent,

$$p(\vec{s}) = p(s_1, \dots, s_n) = \prod_{i=1}^n p(s_i).$$

Thus, we derive that  $\mathcal{H}(\vec{X}) = n \mathcal{H}(X)$ , and since  $R = \vec{R}/n$ , we obtain (10.11) from (10.12). ■

Coding together the symbols in blocks is equivalent to coding each symbol  $x_k$  with an average number of bits  $l_k$  that is not an integer. This explains why block coding can nearly reach the entropy lower bound. The Huffman code can also be adaptively modified for long sequences in which the probability of occurrence of the symbols may vary [18]. The probability distribution is computed from the histogram (cumulative distribution) of the  $N$  most recent symbols that were decoded. The next  $N$  symbols are coded with a new Huffman code calculated from the updated probability distribution. However, recomputing the Huffman code after updating the probability distribution is computationally expensive. Arithmetic codes have a causality structure that makes it easier to adapt the code to a varying probability distribution.

### **Arithmetic Code**

Like a block Huffman code, an arithmetic code [411] records the symbols  $\{x_k\}_{1 \leq k \leq K}$  in blocks to be coded. However, an arithmetic code is more structured; it progressively constructs the code of a whole block as each symbol is taken into account. When the probability  $p_k$  of each symbol  $x_k$  is not known, an adaptive arithmetic code progressively learns the probability distribution of the source and adapts the encoding.

We consider a block of symbols  $\vec{s} = s_1, s_2, \dots, s_n$  produced by a random vector  $\vec{X} = X_1, \dots, X_n$  of  $n$  independent random variables. Each  $X_k$  has the same probability distribution  $p(x)$  as the source  $X$ , with  $p(x_j) = p_j$ . An arithmetic code represents each  $\vec{s}$  by an interval  $[a_n, a_n + b_n]$  included in  $[0, 1]$ , with a length equal to the probability of occurrence of this sequence:

$$b_n = \prod_{k=1}^n p(s_k).$$

This interval is defined by induction as follows. We initialize  $a_0 = 0$  and  $b_0 = 1$ . Let  $[a_i, a_i + b_i]$  be the interval corresponding to the first  $i$  symbols  $s_1, \dots, s_i$ . Suppose that the next symbol  $s_{i+1}$  is equal to  $x_j$  so that  $p(s_{i+1}) = p_j$ . The new interval  $[a_{i+1}, a_{i+1} + b_{i+1}]$  is a subinterval of  $[a_i, a_i + b_i]$  with a size reduced by  $p_j$ :

$$a_{i+1} = a_i + b_i \sum_{k=1}^{j-1} p_k \quad \text{and} \quad b_{i+1} = b_i p_j.$$

The final interval  $[a_n, a_n + b_n]$  characterizes the sequence  $s_1, \dots, s_n$  unambiguously because the  $K^n$  different blocks of symbols  $\vec{s}$  correspond to  $K^n$  different intervals that make a partition of  $[0, 1]$ . Since these intervals are nonoverlapping,  $[a_n, a_n + b_n]$  is characterized by coding a number  $c_n \in [a_n, a_n + b_n]$  in binary form. The binary expression of the chosen numbers  $c_n$  for each of the  $K^n$  intervals defines a prefix code so that a sequence of such numbers is uniquely decodable. The value of  $c_n$  is progressively calculated by adding refinement bits when  $[a_i, a_i + b_i]$  is reduced in the next subinterval  $[a_{i+1}, a_{i+1} + b_{i+1}]$  until  $[a_n, a_n + b_n]$ .

There are efficient implementations that avoid numerical errors caused by the finite precision of arithmetic calculations when calculating  $c_n$  [488]. The resulting binary number  $c_n$  has  $d_n$  digits with

$$-\lceil \log_2 b_n \rceil \leq d_n \leq -\lceil \log_2 b_n \rceil + 2.$$

Since  $\log_2 b_n = \sum_{i=1}^n \log_2 p(s_i)$  and  $\mathcal{H}(X) = E\{\log_2 X\}$ , one can verify that the average number of bits per symbol of this arithmetic code satisfies

$$\mathcal{H}(X) \leq R_X \leq \mathcal{H}(X) + \frac{2}{n}. \quad (10.13)$$

When the successive values  $X_k$  of the blocks are not independent, the upper and lower bounds (10.13) remain valid because the successive symbols are encoded as if they were independent.

An arithmetic code has a causal structure in the sense that the first  $i$  symbols of a sequence  $s_1, \dots, s_i, s_{i+1}, \dots, s_n$  are specified by an interval  $[a_i, a_i + b_i]$  that does not depend on the value of the last  $n - i$  symbols. Since the sequence is progressively coded and decoded, one can implement an adaptive version that progressively learns the probability distribution  $p(x)$  [377, 412]. When coding  $s_{i+1}$ , this probability distribution can be approximated by the histogram (cumulative distribution)  $p_i(x)$  of the first  $i$  symbols. The subinterval of  $[a_i, a_i + b_i]$  associated to  $s_{i+1}$  is calculated with this estimated probability distribution. Suppose that  $s_{i+1} = x_j$ ; we denote  $p_i(x_j) = p_{i,j}$ . The new interval is defined by

$$a_{i+1} = a_i + b_i \sum_{k=1}^{j-1} p_{i,k} \quad \text{and} \quad b_{i+1} = b_i p_{i,j}. \quad (10.14)$$

The decoder is able to recover  $s_{i+1}$  by recovering the first  $i$  symbols of the sequence and computing the cumulative probability distribution  $p_i(x)$  of these symbols. The interval  $[a_{i+1}, a_{i+1} + b_{i+1}]$  is then calculated from  $[a_i, a_i + b_i]$  with (10.14). The initial distribution  $p_0(x)$  can be set to be uniform.

If the symbols of the block are produced by independent random variables, then as  $i$  increases, the estimated probability distribution  $p_i(x)$  converges to the probability distribution  $p(x)$  of the source. As the total block size  $n$  increases to  $+\infty$ , one can prove that the average bit rate of this adaptive arithmetic code converges to the entropy of the source. Under weaker Markov random-chain hypotheses this result also remains valid [412].

### Noise Sensitivity

Huffman and arithmetic codes are more compact than a simple fixed-length code of size  $\log_2 K$ , but they are also more sensitive to errors. For a constant-length code, a single bit error modifies the value of only one symbol. In contrast, a single bit error in a variable-length code may modify the whole symbol sequence. In noisy transmissions where such errors might occur, it is necessary to use an error correction code that introduces a slight redundancy in order to suppress the transmission errors [18].

### 10.2.2 Scalar Quantization

If the source  $X$  has arbitrary real values, it cannot be coded with a finite number of bits. A scalar quantizer  $Q$  approximates  $X$  by  $\tilde{X} = Q(X)$ , which takes its values over a finite set. We study the optimization of such a quantizer in order to minimize the number of bits needed to code  $\tilde{X}$  for a given mean-square error

$$d = E\{(X - \tilde{X})^2\}.$$

Suppose that  $X$  takes its values in  $[a, b]$ , which may correspond to the whole real axis. We decompose  $[a, b]$  in  $K$  intervals  $\{(y_{k-1}, y_k]\}_{1 \leq k \leq K}$  of variable length, with  $y_0 = a$  and  $y_K = b$ . A scalar quantizer approximates all  $x \in (y_{k-1}, y_k]$  by  $x_k$ :

$$\forall x \in (y_{k-1}, y_k], \quad Q(x) = x_k.$$

The intervals  $(y_{k-1}, y_k]$  are called *quantization bins*. Rounding off integers is a simple example where the quantization bins  $(y_{k-1}, y_k] = (k - \frac{1}{2}, k + \frac{1}{2}]$  have size 1 and  $x_k = k$  for any  $k \in \mathbb{Z}$ .

#### High-Resolution Quantizer

Let  $p(x)$  be the probability density of the random source  $X$ . The mean-square quantization error is

$$d = E\{(X - \tilde{X})^2\} = \int_{-\infty}^{+\infty} (x - Q(x))^2 p(x) dx. \quad (10.15)$$

A quantizer is said to have a *high resolution* if  $p(x)$  is approximately constant on each quantization bin  $(y_{k-1}, y_k]$  of size  $\Delta_k = y_k - y_{k-1}$ . This is the case if the sizes  $\Delta_k$  are sufficiently small relative to the rate of variation of  $p(x)$ , so that one can neglect these variations in each quantization bin. We then have

$$p(x) = \frac{p_k}{\Delta_k} \quad \text{for } x \in (y_{k-1}, y_k], \quad (10.16)$$

where

$$p_k = \Pr\{X \in (y_{k-1}, y_k]\}.$$

Theorem 10.4 computes the mean-square error under this high-resolution hypothesis.

**Theorem 10.4.** For a high-resolution quantizer, the mean-square error  $d$  is minimized when  $x_k = (y_k + y_{k-1})/2$ , which yields

$$d = \frac{1}{12} \sum_{k=1}^K p_k \Delta_k^2. \quad (10.17)$$

**Proof.** The quantization error (10.15) can be rewritten as

$$d = \sum_{k=1}^K \int_{y_{k-1}}^{y_k} (x - x_k)^2 p(x) dx.$$

Replacing  $p(x)$  by its expression (10.16) gives

$$d = \sum_{k=1}^K \frac{p_k}{\Delta_k} \int_{y_{k-1}}^{y_k} (x - x_k)^2 dx. \quad (10.18)$$

One can verify that each integral is minimum for  $x_k = (y_k + y_{k-1})/2$ , which yields (10.17). ■

### Uniform Quantizer

The uniform quantizer is an important special case where all quantization bins have the same size

$$y_k - y_{k-1} = \Delta \quad \text{for } 1 \leq k \leq K.$$

For a high-resolution uniform quantizer, the average quadratic distortion (10.17) becomes

$$d = \frac{\Delta^2}{12} \sum_{k=1}^K p_k = \frac{\Delta^2}{12}. \quad (10.19)$$

It is independent of the probability density  $p(x)$  of the source.

### Entropy Constrained Quantizer

We want to minimize the number of bits required to code the quantized values  $\tilde{X} = Q(X)$  for a fixed distortion  $d = E\{(X - \tilde{X})^2\}$ . The Shannon theorem (10.1) proves that the minimum average number of bits to code  $\tilde{X}$  is the entropy  $H(\tilde{X})$ . Huffman and arithmetic codes produce bit rates close to this entropy lower bound. Thus, we design a quantizer that minimizes  $H(\tilde{X})$ .

The quantized source  $\tilde{X}$  takes  $K$  possible values  $\{x_k\}_{1 \leq k \leq K}$  with probabilities

$$p_k = \Pr(\tilde{X} = x_k) = \Pr(X \in (y_{k-1}, y_k]) = \int_{y_{k-1}}^{y_k} p(x) dx.$$

Its entropy is

$$H(\tilde{X}) = - \sum_{k=1}^K p_k \log_2 p_k.$$

For a high-resolution quantizer, Theorem 10.5 by Gish and Pierce [273] relates  $\mathcal{H}(\tilde{X})$  to the *differential entropy* of  $X$  defined by

$$\mathcal{H}_d(X) = - \int_{-\infty}^{+\infty} p(x) \log_2 p(x) dx. \quad (10.20)$$

**Theorem 10.5:** *Gish, Pierce.* If  $Q$  is a high-resolution quantizer with respect to  $p(x)$ , then

$$\mathcal{H}(\tilde{X}) \geq \mathcal{H}_d(X) - \frac{1}{2} \log_2(12d). \quad (10.21)$$

This inequality is an equality if and only if  $Q$  is a uniform quantizer.

**Proof.** By definition, a high-resolution quantizer satisfies (10.16), so  $p_k = p(x)\Delta_k$  for  $x \in (y_{k-1}, y_k]$ . Thus,

$$\begin{aligned} \mathcal{H}(\tilde{X}) &= - \sum_{k=1}^K p_k \log_2 p_k \\ &= - \sum_{k=1}^K \int_{y_{k-1}}^{y_k} p(x) \log_2 p(x) dx - \sum_{k=1}^K p_k \log_2 \Delta_k \\ &= \mathcal{H}_d(X) - \frac{1}{2} \sum_{k=1}^K p_k \log_2 \Delta_k^2. \end{aligned} \quad (10.22)$$

The Jensen inequality for a concave function  $\phi(x)$  proves that if  $p_k \geq 0$  with  $\sum_{k=1}^K p_k = 1$ , then for any  $\{a_k\}_{1 \leq k \leq K}$ ,

$$\sum_{k=1}^K p_k \phi(a_k) \leq \phi\left(\sum_{k=1}^K p_k a_k\right). \quad (10.23)$$

If  $\phi(x)$  is strictly concave, the inequality is an equality if and only if all  $a_k$  are equal when  $p_k \neq 0$ . Since  $\log_2(x)$  is strictly concave, we derive from (10.17) and (10.23) that

$$\frac{1}{2} \sum_{k=1}^K p_k \log_2(\Delta_k^2) \leq \frac{1}{2} \log_2 \left( \sum_{k=1}^K p_k \Delta_k^2 \right) = \frac{1}{2} \log_2(12d).$$

Inserting this in (10.22) proves that

$$\mathcal{H}(\tilde{X}) \geq \mathcal{H}_d(X) - \frac{1}{2} \log_2(12d).$$

This inequality is an equality if and only if all  $\Delta_k$  are equal, which corresponds to a uniform quantizer. ■

This theorem proves that for a high-resolution quantizer, the minimum average bit rate  $R_X = \mathcal{H}(\tilde{X})$  is achieved by a uniform quantizer and

$$R_X = \mathcal{H}_d(X) - \frac{1}{2} \log_2(12d). \quad (10.24)$$

In this case,  $d = \Delta^2/12$ , so

$$R_X = \mathcal{H}_d(X) - \log_2 \Delta. \quad (10.25)$$

The distortion rate is obtained by taking the inverse of (10.24):

$$d(R_X) = \frac{1}{12} 2^{2\mathcal{H}_d(X)} 2^{-2R_X}. \quad (10.26)$$

---

## 10.3 HIGH BIT RATE COMPRESSION

Section 10.3.1 studies the distortion rate performance of a transform coding computed with high-resolution quantizers. For Gaussian processes, Section 10.3.2 proves that the optimal basis is the Karhunen-Loëve basis. An application to audio compression is studied in Section 10.3.3.

### 10.3.1 Bit Allocation

Let us optimize the transform code of a random vector  $F[n]$  decomposed in an orthonormal basis  $\{g_m\}_{0 \leq m < N}$ :

$$F = \sum_{m=0}^{N-1} F_B[m] g_m.$$

Each  $F_B[m]$  is a zero-mean source that is quantized into  $\tilde{F}_B[m]$  with an average bit budget  $R_m$ . For a high-resolution quantization, Theorem 10.5 proves that the error  $d_m = E\{|F_B[m] - \tilde{F}_B[m]|^2\}$  is minimized with a uniform scalar quantization, and  $R_m = \mathcal{H}_d(X) - \log_2 \Delta_m$  where  $\Delta_m$  is the bin size.

In many applications, the overall bit budget  $R$  is fixed by some memory or transmission bandwidth constraints. Thus, we need to optimize the choice of the quantization steps  $\{\Delta_m\}_{0 \leq m < N}$  to minimize the total distortion

$$d = \sum_{m=0}^{N-1} d_m$$

for a fixed-bit budget

$$R = \sum_{m=0}^{N-1} R_m.$$

The following bit allocation theorem (10.6) proves that the transform code is optimized when all  $\Delta_m$  are equal, by minimizing the distortion rate Lagrangian

$$\mathcal{L}(R, d) = d + \lambda R = \sum_{m=0}^{N-1} (d_m + \lambda R_m). \quad (10.27)$$

**Theorem 10.6.** For a fixed-bit budget  $R$  with a high-resolution quantization, the total distortion  $d$  is minimum for

$$\Delta_m^2 = 2^{2\bar{\mathcal{H}}_d} 2^{-2\bar{R}} \quad \text{for } 0 \leq m < N \quad (10.28)$$

with

$$\bar{R} = \frac{R}{N} \quad \text{and} \quad \bar{\mathcal{H}}_d = \frac{1}{N} \sum_{m=0}^{N-1} \mathcal{H}_d(F_B[m]).$$

The resulting distortion rate is

$$d(\bar{R}) = \frac{N}{12} 2^{2\bar{\mathcal{H}}_d} 2^{-2\bar{R}}. \quad (10.29)$$

**Proof.** For uniform high-resolution quantizations, (10.26) proves that

$$d_m(R_m) = \frac{1}{12} 2^{2\mathcal{H}_d(F_B[m])} 2^{-2R_m} \quad (10.30)$$

is a convex function of  $R_m$ , and the bit budget condition can be written as

$$R = \sum_{m=0}^{N-1} R_m = \sum_{m=0}^{N-1} \mathcal{H}_d(F_B[m]) - \sum_{m=0}^{N-1} \frac{1}{2} \log_2(12 d_m). \quad (10.31)$$

The minimization of  $d = \sum_{m=0}^{N-1} d_m$  under the equality constraint  $\sum_{m=0}^{N-1} R_m = R$  is thus a convex minimization obtained by finding the multiplier  $\lambda$ , which minimizes the distortion rate Lagrangian  $\mathcal{L}(R, d) = \sum_{m=0}^{N-1} (d_m(R_m) + \lambda R_m)$ . At the minimum, the relative variation of the distortion rate for each coefficient is constant and equal to the Lagrange multiplier:

$$\lambda = -\frac{\partial d_m(R_m)}{\partial R_m} = 2d_m \log_e 2 \quad \text{for } 0 \leq m < N.$$

Thus,

$$\Delta_m^2 / 12 = d_m = \frac{d}{N} = \frac{\lambda}{2 \log_e 2}. \quad (10.32)$$

Since  $d_m = d/N$ , the bit budget condition (10.31) becomes

$$R = \sum_{m=0}^{N-1} \mathcal{H}_d(F_B[m]) - \frac{N}{2} \log_2 \left( \frac{12d}{N} \right).$$

Inverting this equation gives the expression (10.29) of  $d(R)$  and inserting this result in (10.32) yields (10.28). ■

This theorem shows that the transform code is optimized if it introduces the same expected error  $d_m = \Delta_m^2 / 12 = d/N$  along each direction  $g_m$  of the basis  $\mathcal{B}$ . Then, the number of bits  $R_m$  used to encode  $F_B[m]$  depends only on its differential entropy:

$$R_m = \mathcal{H}_d(F_B[m]) - \frac{1}{2} \log_2 \left( \frac{12d}{N} \right). \quad (10.33)$$

Let  $\sigma_m^2$  be the variance of  $F_B[m]$ , and let  $F_B[m]/\sigma_m$  be the normalized random variable of variance 1. A simple calculation shows that

$$\mathcal{H}_d(F_B[m]) = \mathcal{H}_d(F_B[m]/\sigma_m) + \log_2 \sigma_m.$$

The “optimal bit allocation”  $R_m$  in (10.33) may become negative if the variance  $\sigma_m$  is too small, which is clearly not an admissible solution. In practice,  $R_m$  must be a positive integer, but with this condition the resulting optimal solution has no simple analytic expression (Exercise 10.8). If we neglect the integer bit constraint, (10.33) gives the optimal bit allocation as long as  $R_m \geq 0$ .

### **Weighted Mean-Square Error**

We mentioned that a mean-square error often does not measure the perceived distortion of images or audio signals very well. When the vectors  $g_m$  are localized in time and frequency, a mean-square norm sums the errors at all times and frequencies with equal weights. Thus, it hides the temporal and frequency properties of the error  $F - \tilde{F}$ . Better norms can be constructed by emphasizing certain frequencies more than others in order to match our audio or visual sensitivity, which varies with the signal frequency. A weighted mean-square norm is defined by

$$d = \sum_{m=0}^{N-1} \frac{d_m}{w_m^2}, \quad (10.34)$$

where  $\{w_m^2\}_{0 \leq m < N}$  are constant weights.

Theorem 10.6 applies to weighted mean-square errors by observing that

$$d = \sum_{m=0}^{N-1} d_m^w,$$

where  $d_m^w = d_m / w_m^2$  is the quantization error of  $F_B^w[m] = F_B[m] / w_m$ . Theorem 10.6 proves that bit allocation is optimized by uniformly quantizing all  $F_B^w[m]$  with the same bin size  $\Delta$ . This implies that coefficients  $F_B[m]$  are uniformly quantized with a bin size  $\Delta_m = \Delta w_m$ ; it follows that  $d_m = w_m^2 d / N$ . As expected, larger weights increase the error in the corresponding direction. The uniform quantization  $Q_{\Delta_m}$  with bins of size  $\Delta_m$  can be computed from a quantizer  $Q$  that associates to any real number its closest integer:

$$Q_{\Delta_m}(F_B[m]) = \Delta_m Q\left(\frac{F_B[m]}{\Delta_m}\right) = \Delta w_m Q\left(\frac{F_B[m]}{\Delta w_m}\right). \quad (10.35)$$

### **10.3.2 Optimal Basis and Karhunen-Loèv**

Transform code performance depends on the choice of an orthonormal basis  $\mathcal{B}$ . For high-resolution quantizations, (10.29) proves that the distortion rate  $d(\bar{R})$  is optimized by choosing a basis  $\mathcal{B}$  that minimizes the average differential entropy

$$\bar{\mathcal{H}}_d = \frac{1}{N} \sum_{m=0}^{N-1} \mathcal{H}_d(F_B[m]).$$

In general, we do not know how to compute this optimal basis because the probability density of  $F_B[m] = \langle F, g_m \rangle$  may depend on  $g_m$  in a complicated way.

### Gaussian Process

If  $F$  is a Gaussian random vector, then the coefficients  $F_{\mathcal{B}}[m]$  are Gaussian random variables in any basis. In this case, the probability density of  $F_{\mathcal{B}}[m]$  depends only on the variance  $\sigma_m^2$ :

$$p_m(x) = \frac{1}{\sigma_m \sqrt{2\pi}} \exp\left(\frac{-x^2}{2\sigma_m^2}\right).$$

With a direct integration, we verify that

$$\mathcal{H}_d(F_{\mathcal{B}}[m]) = - \int_{-\infty}^{+\infty} p_m(x) \log_2 p_m(x) dx = \log_2 \sigma_m + \log_2 \sqrt{2\pi e}.$$

Inserting this expression in (10.29) yields

$$d(\bar{R}) = N \frac{\pi e}{6} \rho^2 2^{-2\bar{R}}, \quad (10.36)$$

where  $\rho^2$  is the geometrical mean of all variances:

$$\rho^2 = \left( \prod_{m=0}^{N-1} \sigma_m^2 \right)^{1/N}.$$

Therefore, the basis must be chosen to minimize  $\rho^2$ .

**Theorem 10.7.** The geometrical mean variance  $\rho^2$  is minimized in a Karhunen-Loëve basis of  $F$ .

**Proof.** Let  $K$  be the covariance operator of  $F$ ,

$$\sigma_m^2 = \langle Kg_m, g_m \rangle.$$

Observe that

$$\log_2 \rho^2 = \frac{1}{N} \sum_{m=0}^{N-1} \log_2 (\langle Kg_m, g_m \rangle). \quad (10.37)$$

Lemma 10.2 proves that since  $C(x) = \log_2(x)$  is strictly concave,  $\sum_{m=0}^{N-1} \log_2 (\langle Kg_m, g_m \rangle)$  is minimum if and only if  $\{g_m\}_{0 \leq m < N}$  diagonalizes  $K$ , and thus if it is a Karhunen-Loëve basis.

**Lemma 10.2.** Let  $K$  be a covariance operator and  $\{g_m\}_{0 \leq m < N}$  be an orthonormal basis. If  $C(x)$  is strictly concave, then  $\sum_{m=0}^{N-1} C(\langle Kg_m, g_m \rangle)$  is minimum if and only if  $K$  is diagonal in this basis.

To prove this lemma, let us consider a Karhunen-Loëve basis  $\{h_m\}_{0 \leq m < N}$  that diagonalizes  $K$ . As in (9.26), by decomposing  $g_m$  in the basis  $\{h_i\}_{0 \leq i < N}$ , we obtain

$$\langle Kg_m, g_m \rangle = \sum_{i=0}^{N-1} |\langle g_m, h_i \rangle|^2 \langle Kh_i, h_i \rangle. \quad (10.38)$$

Since  $\sum_{i=0}^{N-1} |\langle g_m, h_i \rangle|^2 = 1$ , applying the Jensen inequality (A.2) to the concave function  $C(x)$  proves that

$$C(\langle Kg_m, g_m \rangle) \geq \sum_{i=0}^{N-1} |\langle g_m, h_i \rangle|^2 C(\langle Kh_i, h_i \rangle). \quad (10.39)$$

Thus,

$$\sum_{m=0}^{N-1} C(\langle Kg_m, g_m \rangle) \geq \sum_{m=0}^{N-1} \sum_{i=0}^{N-1} |\langle g_m, h_i \rangle|^2 C(\langle Kh_i, h_i \rangle).$$

Since  $\sum_{m=0}^{N-1} |\langle g_m, h_i \rangle|^2 = 1$ , we derive that

$$\sum_{m=0}^{N-1} C(\langle Kg_m, g_m \rangle) \geq \sum_{i=0}^{N-1} C(\langle Kh_i, h_i \rangle).$$

This inequality is an equality if and only if for all  $m$  (10.39) is an equality. Since  $C(x)$  is strictly concave, this is possible only if all values  $\langle Kh_i, h_i \rangle$  are equal as long as  $\langle g_m, h_i \rangle \neq 0$ . Thus, we derive that  $g_m$  belongs to an eigenspace of  $K$  and is also an eigenvector of  $K$ . Thus,  $\{g_m\}_{0 \leq m < N}$  diagonalizes  $K$  as well. ■

Together with the distortion rate (10.36), this result proves that a high bit rate transform code of a Gaussian process is optimized in a Karhunen-Loëve basis. The Karhunen-Loëve basis diagonalizes the covariance matrix, which means that the decomposition coefficients  $F_B[m] = \langle F, g_m \rangle$  are uncorrelated. If  $F$  is a Gaussian random vector, then the coefficients  $F_B[m]$  are jointly Gaussian. In this case, being uncorrelated implies that they are independent. The optimality of a Karhunen-Loëve basis is therefore quite intuitive since it produces coefficients  $F_B[m]$  that are independent. The independence of the coefficients justifies using a scalar quantization rather than a vector quantization.

### Coding Gain

The Karhunen-Loëve basis  $\{g_m\}_{0 \leq m < N}$  of  $F$  is a priori not well structured. The decomposition coefficients  $\{\langle f, g_m \rangle\}_{0 \leq m < N}$  of a signal  $f$  are thus computed with  $N^2$  multiplications and additions, which is often too expensive in real-time coding applications. Transform codes often approximate this Karhunen-Loëve basis by a more structured basis that admits a faster decomposition algorithm. The performance of a basis is evaluated by the coding gain [34]

$$G = \frac{E\{\|F\|^2\}}{N \rho^2} = \frac{\sum_{m=0}^{N-1} \sigma_m^2}{N \left( \prod_{m=0}^{N-1} \sigma_m^2 \right)^{1/N}}. \quad (10.40)$$

Theorem 10.7 proves that  $G$  is maximum in a Karhunen-Loëve basis.

### Non-Gaussian Processes

When  $F$  is not Gaussian, the coding gain  $G$  no longer measures the coding performance of the basis. Indeed, the distortion rate (10.29) depends on the average differential entropy factor  $2^{2\bar{H}_d}$ , which is not proportional to  $\rho^2$ . Therefore, the Karhunen-Loëve basis is not optimal.

Circular stationary processes with piecewise smooth realizations are examples of non-Gaussian processes that are not well compressed in their Karhunen-Loëve basis, which is the discrete Fourier basis. In Section 10.4 we show that wavelet bases yield better distortion rates because they can approximate these signals with few nonzero coefficients.

#### 10.3.3 Transparent Audio Code

The compact disc standard samples high-quality audio signals at 44.1 kHz. Samples are quantized with 16 bits, producing a PCM of 706 kb/s. Audio codes must be “transparent,” which means that they should not introduce errors that can be heard by an “average” listener.

Sounds are often modeled as realizations of Gaussian processes. This justifies the use of a Karhunen-Loëve basis to minimize the distortion rate of transform codes. To approximate the Karhunen-Loëve basis, we observe that many audio signals are locally stationary over a sufficiently small time interval. This means that over this time interval, the signal can be approximated by a realization of a stationary process. One can show [192] that the Karhunen-Loëve basis of locally stationary processes is well approximated by a local cosine basis with appropriate window sizes. The local stationarity hypothesis is not always valid, especially for attacks of musical instruments, but bases of local time-frequency atoms remain efficient for most audio segments.

Bases of time-frequency atoms are also well adapted to matching the quantization errors with our hearing sensitivity. Instead of optimizing a mean-square error as in Theorem 10.6, perceptual coders [317] adapt the quantization so that errors fall below an auditory threshold, which depends on each time-frequency atom  $g_m$ .

### Audio Masking

A large-amplitude stimulus often makes us less sensitive to smaller stimuli of a similar nature. This is called a *masking effect*. In a sound, a small-amplitude quantization error may not be heard if it is added to a strong signal component in the same frequency neighborhood. Audio masking takes place in critical frequency bands  $[\omega_c - \Delta\omega/2, \omega_c + \Delta\omega/2]$  that have been measured with psychophysical experiments [425]. A strong narrow-band signal having a frequency energy in the interval  $[\omega_c - \Delta\omega/2, \omega_c + \Delta\omega/2]$  decreases the hearing sensitivity within this frequency interval. However, it does not influence the sensitivity outside this frequency range. In the frequency interval  $[0, 20 \text{ kHz}]$ , there are approximately 25 critical bands. Below 700 Hz, the bandwidths of critical bands are on the order of 100 Hz. Above 700 Hz

the bandwidths increase proportionally to the center frequency  $\omega_c$ :

$$\Delta\omega \approx \begin{cases} 100 & \text{for } \omega_c \leq 700 \\ 0.15 \omega_c & \text{for } 700 \leq \omega_c \leq 20,000. \end{cases} \quad (10.41)$$

The masking effect also depends on the nature of the sound, particularly its tonality. A tone is a signal with a narrow-frequency support as opposed to a noiselike signal with a frequency spectrum that is spread out. A tone has a different masking influence than a noise-type signal; this difference must be taken into account [442].

### ***Adaptive Quantization***

To take advantage of audio masking, transform codes are implemented in orthogonal bases of local time-frequency atoms  $\{g_m\}_{0 \leq m < N}$ , with frequency supports inside critical bands. To measure the effect of audio masking at different times, the signal energy is computed in each critical band. This is done with an FFT over short time intervals, on the order of 10 ms, where signals are considered to be approximately stationary. The signal tonality is estimated by measuring the spread of its Fourier transform. The maximum admissible quantization error in each critical band is estimated depending on both the total signal energy in the band and the signal tonality. This estimation is done with approximate formulas that are established with psychophysical experiments [335]. For each vector  $g_m$  having a Fourier transform inside a given critical band, the inner product  $\langle f, g_m \rangle$  is uniformly quantized according to the maximum admissible error. Quantized coefficients are then entropy coded.

Although the SNR may be as low as 13 db, such an algorithm produces a nearly transparent audio code because the quantization error is below the perceptual threshold in each critical band. The most important degradations introduced by such transform codes are *pre-echoes*. During a silence, the signal remains zero, but it can suddenly reach a large amplitude due to a beginning speech or a musical attack. In a short time interval containing this attack, the signal energy may be quite large in each critical band. By quantizing the coefficients  $\langle f, g_m \rangle$  we introduce an error both in the silent part and in the attack. The error is not masked during the silence and will clearly be heard. It is perceived as a “pre-echo” of the attack. This pre-echo is due to the temporal variation of the signal, which does not respect the local stationarity hypothesis. However, it can be detected and removed with postprocessings.

### ***Choice of Basis***

The MUSICAM (Masking-pattern Universal Subband Integrated Coding and Multiplexing) coder [203] used in the MPEG-I standard [121] is the simplest perceptual subband coder. It decomposes the signal in 32 equal frequency bands of 750-Hz bandwidth, with a filter bank constructed with frequency-modulated windows of 512 samples. This decomposition is similar to a signal expansion in a local cosine basis, but the modulated windows used in MUSICAM are not orthogonal. The quantization levels are adapted in each frequency band in order to take into account the masking properties of the signal. Quantized coefficients are not entropy coded. This

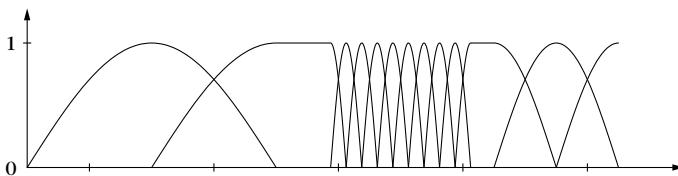
system compresses audio signals up to 128 kb/s without audible impairment. It is often used for digital radio transmissions where small defects are admissible.

MP3 is the standard MPEG-1 layer 3, which operates on a large range of audio quality: Telephone quality is obtained at 8 kb/s on signals sampled at 2.5 kHz; FM radio quality is obtained at 60 kb/s for a stereo signal sampled at 11 kHz; and CD quality is obtained for bit rates going from 112 to 128 kb/s for stereo signals sampled at 22.1 kHz. To maintain a compatibility with the previous standards, the signal is decomposed both with a filter bank and a local cosine basis. The size of the windows can be adapted as in the MPEG-2 AAC (Advanced Audio Coding) standard described next. The size of the quantization bins is computed with a perceptual masking model. The parameters of the models are not specified in the standard and are thus transmitted with the signal. A Huffman code stores the quantized coefficients. For a pair of stereo signals  $f_1$  and  $f_2$ , the compression is improved by coding an average signal  $a = (f_1 + f_2)/2$  and a difference signal  $d = (f_1 - f_2)/2$ , and by adapting the perceptual model for the quantization of each of these signals.

MPEG-2 AAC is a standard that offers a better audio quality for a given compression ratio. The essential difference with MP3 is that it directly uses a decomposition in a local cosine basis. In order to reduce pre-echo distortions, and to adapt the basis to the stationarity intervals of the signal, the size  $2^j$  of the windows can vary from 256 to 2048. However, on each interval of 2048 samples, the size of the window must remain constant, as illustrated in Figure 10.4. Each window has a raising and a decaying profile that is as large as possible, while overlapping only the two adjacent windows. The profile used by the standard is

$$\beta(t) = \sin\left(\frac{\pi}{4}(1+t)\right).$$

Section 8.4.1 explains in (8.85) how to construct the windows  $g_p(t)$  on each interval from this profile. The discrete windows of the local cosine basis are obtained with a uniform sampling:  $g_p[n] = g_p(n)$ . The choice of windows can also be interpreted as a best-basis choice, further studied in Section 12.2.3. However, as opposed to the bases of the local cosine trees from Section 8.5, the windows have raising and decaying profiles of varying sizes, which are best adapted to the segmentation. The strategy to choose the window sizes is not imposed by the standard, and the



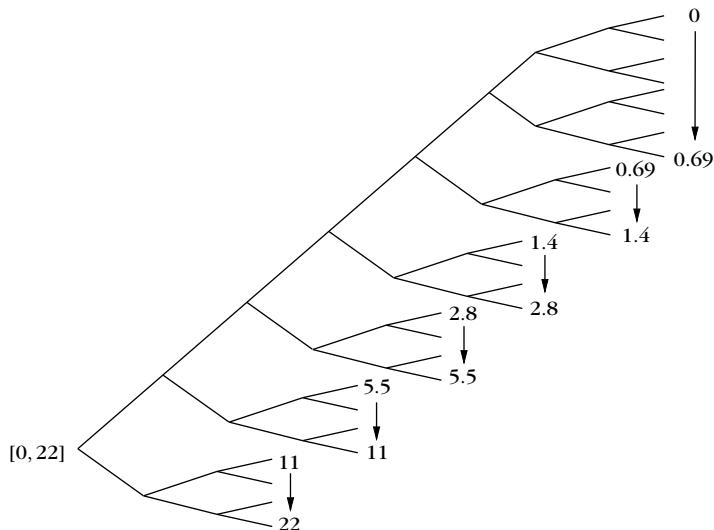
**FIGURE 10.4**

Succession of windows of various sizes on intervals of 2048 samples, which satisfy the orthogonality constraints of local cosine bases.

code transmits the selected window size. A best-basis algorithm must measure the coding efficiency for windows of different sizes for a given audio segment. In the neighborhood of an attack, it is necessary to choose small windows in order to reduce the pre-echo. Like in MP3, the local cosine coefficients are quantized with a perceptual model that depends on the signal energy in each critical frequency band, and an entropy code is applied.

The AC systems produced by Dolby are similar to MPEG-2 AAC. The signal is decomposed in a local cosine basis, and the window size can be adapted to the local signal content. After a perceptual quantization, a Huffman entropy code is used. These coders operate on a variety of bit rates from 64 kb/s to 192 kb/s.

To best match human perception, transform code algorithms have also been developed in wavelet packet bases, with a frequency decomposition that matches the critical frequency bands [483]. Sinha and Tewfik [442] propose the wavelet packet basis shown in Figure 10.5, which is an  $M = 4$  wavelet basis. The properties of M-band wavelet bases are explained in Section 8.1.3. These four wavelets have a bandwidth of  $1/4$ ,  $1/5$ ,  $1/6$ , and  $1/7$  octaves, respectively. The lower-frequency interval  $[0, 700]$  is decomposed with eight wavelet packets of the same bandwidth in order to match the critical frequency bands (10.41). These wavelet packet coefficients are quantized with perceptual models and are entropy coded. Nearly transparent audio codes are obtained at 70 kb/s.



**FIGURE 10.5**

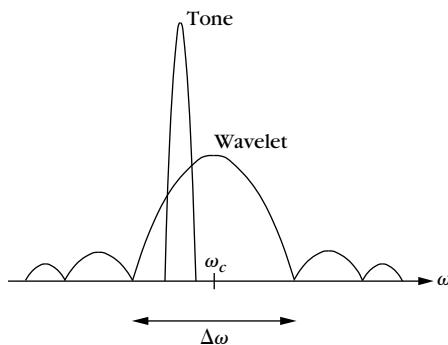
Wavelet packet tree that decomposes the frequency interval  $[0, 22 \text{ kHz}]$  in 24 frequency bands covered by  $M = 4$  wavelets dilated over six octaves, plus 8 low-frequency bands of the same bandwidth. The frequency bands are indicated at the leaves in kHz.

Wavelets produce smaller pre-echo distortions compared to local cosine bases. At the sound attack, the largest wavelet coefficients appear at fine scales. Because fine-scale wavelets have a short support, a quantization error creates a distortion that is concentrated near the attack. However, these bases have the disadvantage of introducing a bigger coding delay than local cosine bases. The coding delay is approximately equal to the maximum time support of the vector used in the basis. It is typically larger for wavelets and wavelet packets than for local cosine vectors.

### **Choice of Filter**

Wavelet and wavelet packet bases are constructed with a filter bank of conjugate mirror filters. For perceptual audio coding, the Fourier transform of each wavelet or wavelet packet must have its energy well concentrated in a single critical band. Second-order lobes that may appear in other frequency bands should have a negligible amplitude. Indeed, a narrow-frequency tone creates large-amplitude coefficients for all wavelets with a frequency support covering this tone, as shown in Figure 10.6. Quantizing the wavelet coefficients is equivalent to adding small wavelets with amplitude equal to the quantization error. If the wavelets excited by the tone have important second-order lobes in other frequency intervals, the quantization errors introduce some energy in these frequency intervals that is not masked by the energy of the tone, thereby introducing audible distortion.

To create wavelets and wavelet packets with small second-order frequency lobes, the transfer function of the corresponding conjugate mirror filter  $\hat{h}(\omega)$  must have a zero of high order at  $\omega = \pi$ . Theorem 7.7 proves that conjugate mirror filters with  $p$  zeros at  $\omega = \pi$  have at least  $2p$  nonzero coefficients, and correspond to wavelets of size  $2p - 1$ . Thus, increasing  $p$  produces a longer coding delay. Numerical experiments [442] show that increasing  $p$  up to 30 can enhance the perceptual quality of the audio code, but the resulting filters have at least 60 nonzero coefficients.



**FIGURE 10.6**

A high-energy, narrow-frequency tone can excite a wavelet having a Fourier transform with second-order lobes outside the critical band of width  $\Delta\omega$ . The quantization then creates audible distortion.

## 10.4 SPARSE SIGNAL COMPRESSION

Sparse representations provide high signal-compression factors by only coding a few nonzero coefficients. The high-resolution quantization hypothesis is not valid anymore. Section 10.4.1 shows that the distortion is dominated by the nonlinear approximation term, and most of the bits are devoted to coding the position of nonzero coefficients. These results are illustrated by a wavelet image transform code. Section 10.4.2 refines such transform codes with an embedding strategy with progressive coding capabilities.

The performance of transform codes was studied from a Bayes point of view, by considering signals as realizations of a random vector that has a known probability distribution. However, there is no known stochastic model that incorporates the diversity of complex signals such as nonstationary textures and edges in images. Classic processes and, in particular, Gaussian processes or homogeneous Markov random fields are not appropriate. This section introduces a different framework where the distortion rate is computed with deterministic signal models.

### 10.4.1 Distortion Rate and Wavelet Image Coding

The signal is considered as a deterministic vector  $f \in \mathbb{C}^N$  that is decomposed in an orthonormal basis  $\mathcal{B} = \{g_m\}_{0 \leq m < N}$ :

$$f = \sum_{m=0}^{N-1} f_{\mathcal{B}}[m] g_m \quad \text{with} \quad f_{\mathcal{B}}[m] = \langle f, g_m \rangle.$$

A transform code quantizes all coefficients and reconstructs

$$\tilde{f} = \sum_{m=0}^{N-1} Q(f_{\mathcal{B}}[m]) g_m. \quad (10.42)$$

Let  $R$  be the number of bits used to code the  $N$  quantized coefficients  $Q(f_{\mathcal{B}}[m])$ . The coding distortion is

$$d(R, f) = \|f - \tilde{f}\|^2 = \sum_{m=0}^{N-1} |f_{\mathcal{B}}[m] - Q(f_{\mathcal{B}}[m])|^2. \quad (10.43)$$

We denote by  $p(x)$  the histogram of the  $N$  coefficients  $f_{\mathcal{B}}[m]$ , normalized so that  $\int p(x) dx = 1$ . The quantizer approximates each  $x \in (y_{k-1}, y_k]$  by  $Q(x) = x_k$ . The proportion of quantized coefficients equal to  $x_k$  is

$$p_k = \int_{y_{k-1}}^{y_k} p(x) dx. \quad (10.44)$$

Suppose that the quantized coefficients of  $f$  can take at most  $K$  different quantized values  $x_k$ . A variable-length code represents the quantized values equal to  $x_k$  with an average of  $l_k$  bits, where the lengths  $l_k$  are specified independently from  $f$ . It is implemented with a prefix code or an arithmetic code over blocks of quantized

values that are large enough so that the  $l_k$  can be assumed to take any real values that satisfy the Kraft inequality (10.6)  $\sum_{k=1}^K 2^{-l_k} \leq 1$ . Encoding a signal with  $K$  symbols requires a total number of bits

$$R = N \sum_{k=1}^K p_k l_k. \quad (10.45)$$

The bit budget  $R$  reaches its minimum for  $l_k = -\log_2 p_k$  and thus,

$$R \geq \mathcal{H}(f) = -N \sum_{k=1}^K p_k \log_2 p_k. \quad (10.46)$$

In practice, we do not know in advance the values of  $p_k$ , which depend on the signal  $f$ . An *adaptive variable-length code*, as explained in Section 10.2.1, computes an estimate  $\tilde{p}_k$  of  $p_k$  with an empirical histogram of the already coded coefficients. It sets  $l_k = -\log_2 \tilde{p}_k$  and updates the estimate  $\tilde{p}_k$  and thus  $l_k$  as the coding progresses. Under appropriate ergodicity assumptions, the estimated  $\tilde{p}_k$  converge to  $p_k$  and  $R$  to the entropy  $\mathcal{H}$ . The  $N$  quantized signal coefficients can be modeled as values taken by a random variable  $X$  with a probability distribution equal to the histogram  $p(x)$ . The distortion (10.43) can then be rewritten as

$$d(R, f) = N E\{|X - Q(X)|^2\},$$

and the bit budget of an adaptive variable-length code converges to entropy  $R = \mathcal{H}(Q(X)) = \mathcal{H}(f)$ .

If the high-resolution quantization assumption is valid, which means that  $p(x)$  is nearly constant over quantization intervals, then Theorem 10.5 proves that  $d(R, f)$  is minimum if and only if  $Q$  is a uniform quantizer. The resulting distortion rate computed in (10.26) is

$$d(R, f) = \frac{N}{12} 2^{2\mathcal{H}_d(f)} 2^{-2R/N}, \quad (10.47)$$

with  $\mathcal{H}_d(f) = -\int p(x) \log_2 p(x) dx$ . It predicts an exponential distortion rate decay. The high-resolution quantization assumption is valid if the quantization bins are small enough, which means that  $R/N$  is sufficiently large.

Coded sequences of quantized coefficients are often not homogeneous and ergodic sources. Adaptive variable-length codes can then produce a bit budget that is below the entropy (10.46). For example, the wavelet coefficients of an image often have a larger amplitude at large scales. When coding coefficients from large to fine scales, an adaptive arithmetic code progressively adapts the estimated probability distribution. Thus, it produces a total bit budget that is often smaller than the entropy  $\mathcal{H}(f)$  obtained with a fixed code globally optimized for the  $N$  wavelet coefficients.

### Wavelet Image Code

A simple wavelet image code is introduced to illustrate the properties of low bit rate transform coding in sparse representations. The image is decomposed in a



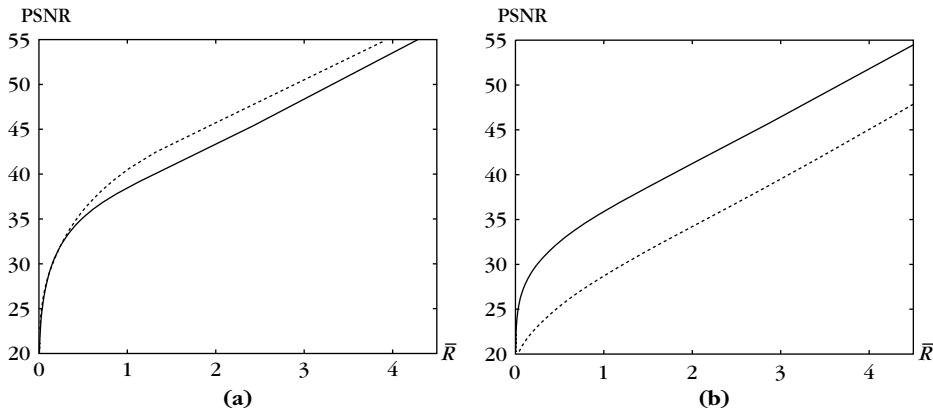
**FIGURE 10.7**

Result of a wavelet transform code with an adaptive arithmetic coding using  $\bar{R} = 0.5$  bit/pixel for images of  $N = 512^2$  pixels: **(a)** Lena, **(b)** GoldHill, **(c)** boats, and **(d)** mandrill.

separable wavelet basis. All wavelet coefficients are uniformly quantized and coded with an adaptive arithmetic code. Figure 10.7 shows examples of coded images with  $R/N = 0.5$  bit/pixel.

The peak signal-to-noise ratio (PSNR) is defined by

$$PSNR(R, f) = 10 \log_{10} \frac{N 255^2}{d(R, f)}.$$

**FIGURE 10.8**

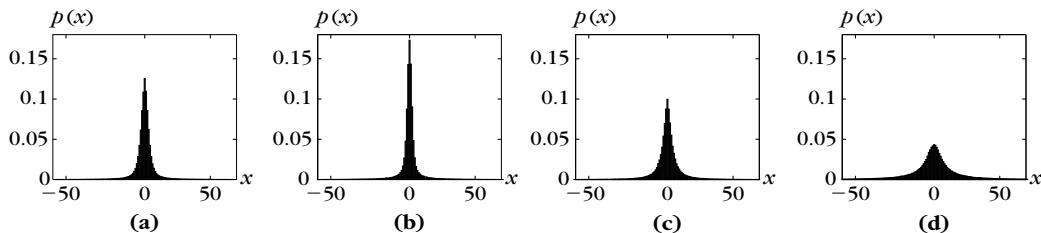
PSNR as a function of  $R/N$ : (a) Lena image (solid line) and boats image (dotted line);  
(b) GoldHill image (solid line) and mandrill image (dotted line).

The high-resolution distortion rate formula (10.47) predicts that there exists a constant  $K$  such that

$$PSNR(R, f) = (20 \log_{10} 2) \bar{R} + K \quad \text{with} \quad \bar{R} = R/N.$$

Figure 10.8 shows that  $PSNR(R, f)$  has indeed a linear growth for  $\bar{R} \geq 1$ , but not for  $\bar{R} < 1$ .

At low bit rates  $\bar{R} \leq 1$ , the quantization interval  $\Delta$  is relatively large. The normalized histogram  $p(x)$  of wavelet coefficients in Figure 10.9 has a narrow peak in the neighborhood of  $x = 0$ . Thus,  $p(x)$  is poorly approximated by a constant in the zero bin  $[-\Delta/2, \Delta/2]$  where  $Q(x) = 0$ . The high-resolution quantization hypothesis is not valid in this zero bin, which explains why the distortion rate formula (10.47) is incorrect. For the mandrill image, the high-resolution hypothesis remains

**FIGURE 10.9**

Normalized histograms of orthogonal wavelet coefficients for each image: (a) Lena, (b) boats, (c) GoldHill, and (d) mandrill.

valid up to  $\bar{R} = 0.5$  because the histogram of its wavelet coefficients is wider in the neighborhood of  $x = 0$ .

### Geometric Bit Budget

If the basis  $\mathcal{B}$  is chosen so that many coefficients  $f_{\mathcal{B}}[m] = \langle f, g_m \rangle$  are close to zero, then the histogram  $p(x)$  has a sharp high-amplitude peak at  $x = 0$ , as in the wavelet histograms shown in Figure 10.9. The distortion rate is calculated at low bit rates, where the high-resolution quantization does not apply.

The bit budget  $R$  is computed by considering separately the set of *significant coefficients*

$$\Lambda_{\Delta/2} = \{m : Q(|f_{\mathcal{B}}[m]|) \neq 0\} = \{m : |f_{\mathcal{B}}[m]| \geq \Delta/2\}.$$

This set is the approximation support that specifies the geometry of a sparse transform coding. Figure 10.10 shows the approximation support of the quantized wavelet coefficients that code the four images in Figure 10.7. The total bit budget  $R$  to code all quantized coefficients is divided into the number of bits  $R_0$  needed to code the coefficients quantized to zero, plus the number of bits  $R_1$  to code significant coefficients:

$$R = R_0 + R_1.$$

The bit budget  $R_0$  can also be interpreted as a geometric bit budget that codes the position of significant coefficients and thus  $\Lambda_{\Delta/2}$ . Let  $M = |\Lambda_{\Delta/2}| \leq N$  be the number of significant coefficients that is coded with  $\log_2 N$  bits. There are  $\binom{N}{M}$  different sets of  $M$  coefficients chosen among  $N$ . To code an approximation support  $\Lambda_{\Delta/2}$  without any other prior information requires a number of bits

$$R_0 = \log_2 N + \log_2 \binom{N}{M} \sim M \left( 1 + \log_2 \frac{N}{M} \right).$$

Theorem 10.8 shows that the number of bits  $R_1$  to code  $M$  significant is typically proportional to  $M$  with a variable-length code. If  $M \ll N$ , then the overall bit budget  $R = R_0 + R_1$  is dominated by the geometric bit budget  $R_0$ .

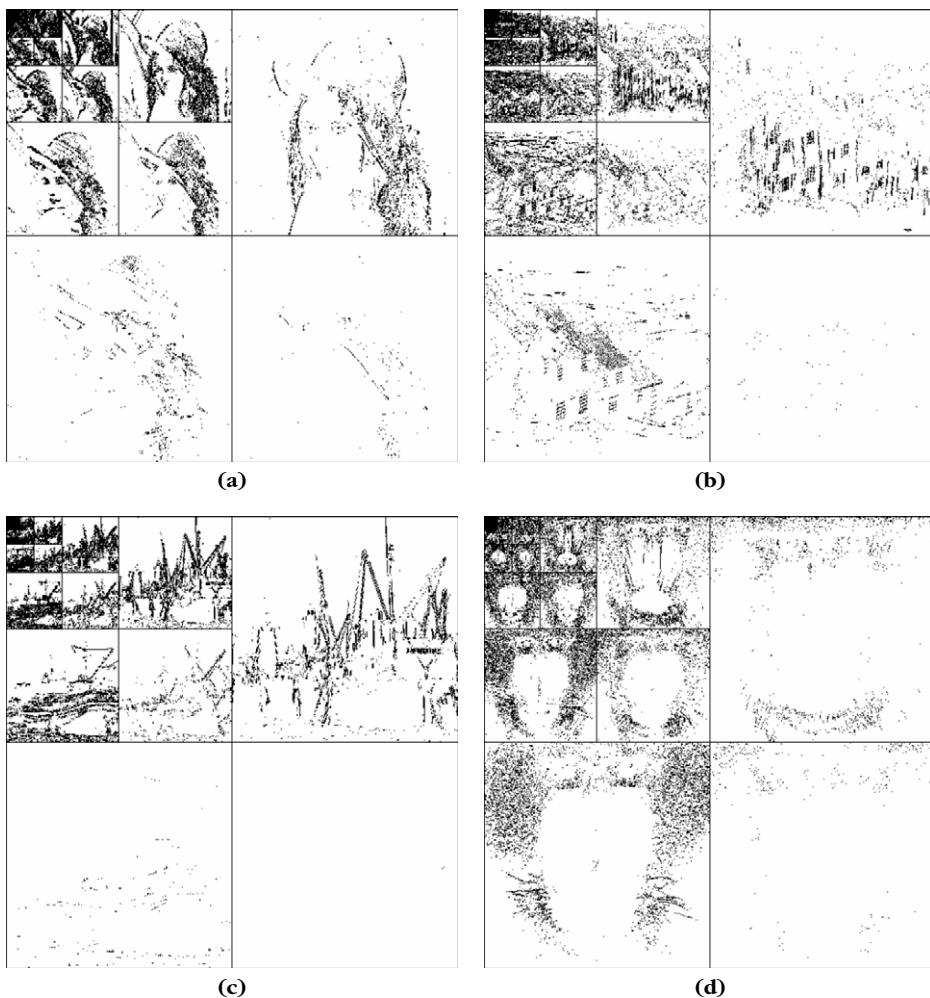
This approximation support  $\Lambda_{\Delta/2}$  can be coded with an entropy coding of the binary *significance map*

$$b[m] = \begin{cases} 0 & \text{if } Q(f_{\mathcal{B}}[m]) = 0 \\ 1 & \text{if } Q(f_{\mathcal{B}}[m]) \neq 0. \end{cases} \quad (10.48)$$

The proportions of 0 and 1 in the significance map are, respectively,  $p_0 = (N - M)/N$  and  $p_1 = M/N$ . An arithmetic code of this significance map yields a bit budget of

$$R_0 \geq H_0 = -N \left( p_0 \log_2 p_0 + p_1 \log_2 p_1 \right), \quad (10.49)$$

which is of the same order as  $\log_2 \binom{N}{M}$  (Exercise 10.5).



**FIGURE 10.10**

Significance maps of quantized wavelet coefficients for images coded with  $\bar{R} = 0.5$  bit/pixel:  
**(a)** Lena, **(b)** GoldHill, **(c)** boats, and **(d)** mandrill.

***Distortion and Nonlinear Approximation***

The distortion  $d(R, f)$  is calculated by separating the significant coefficients in  $\Lambda_{\Delta/2}$  from other coefficients for which  $O(f_B[m]) = 0$ :

$$d(R, f) = \sum_{m=0}^{N-1} |f_B[m] - Q(f_B[m])|^2 \quad (10.50)$$

$$= \sum_{m \notin \Lambda_{\Delta/2}} |f_B[m]|^2 + \sum_{m \in \Lambda_{\Delta/2}} |f_B[m] - Q(f_B[m])|^2. \quad (10.51)$$

Let  $f_M = \sum_{m \in \Lambda_{\Delta/2}} f_B[m] g_m$  be the best  $M$ -term approximation of  $f$  from the  $M$  significant coefficients above  $\Delta/2$ . The first sum of  $d(R, f)$  can be rewritten as a nonlinear approximation error:

$$\|f - f_M\|^2 = \sum_{m \notin \Lambda_{\Delta/2}} |f_B[m]|^2.$$

Since  $0 \leq |x - Q(x)| \leq \Delta/2$ , (10.50) implies that

$$\|f - f_M\|^2 \leq d(R, f) \leq \|f - f_M\|^2 + M \frac{\Delta^2}{4}. \quad (10.52)$$

Theorem 10.8 shows that the nonlinear approximation error  $\|f - f_M\|^2$  dominates the distortion rate behavior [208].

In Section 9.2.1 we prove that nonlinear approximation errors depend on the decay of the sorted coefficients of  $f$  in  $B$ . We denote by  $f_B^r[k] = f_B[m_k]$  the coefficient of rank  $k$ , defined by  $|f_B^r[k]| \geq |f_B^r[k+1]|$  for  $1 \leq k \leq N$ . We write  $|f_B^r[k]| \sim C k^{-s}$  if there exist two constants  $A, B > 0$  independent of  $C, k$ , and  $N$  such that  $A C k^{-s} \leq |f_B^r[k]| \leq B C k^{-s}$ . Theorem 10.8 computes the resulting distortion rate [363].

**Theorem 10.8:** *Falzon, Mallat.* Let  $Q$  be a uniform quantizer. There exists a variable-length code such that for all  $s > 1/2$  and  $C > 0$ , if  $|f_B^r[k]| \sim C k^{-s}$ , then

$$d(R, f) \sim C^2 R^{1-2s} \left(1 + \log_2 \frac{N}{R}\right)^{2s-1} \quad \text{for } R \leq N. \quad (10.53)$$

**Proof.** Since the sorted coefficients satisfy  $|f_B^r[k]| \sim C k^{-s}$  and  $|f_B^r[M]| \sim \Delta$ , we derive that

$$M \sim C^{1/s} \Delta^{-1/s}. \quad (10.54)$$

Since  $s > 1/2$ , the approximation error is

$$\|f - f_M\|^2 = \sum_{k=M+1}^N |f_B^r[k]|^2 \sim \sum_{k=M+1}^N C^2 k^{-2s} \sim C^2 M^{1-2s}. \quad (10.55)$$

But (10.54) shows that  $M \Delta^2 \sim C^2 M^{1-2s}$ , so (10.52) yields

$$d(R, f) \sim C^2 M^{1-2s}. \quad (10.56)$$

We now relate the bit budget  $R = R_0 + R_1$  to  $M$ . The number of bits  $R_0$  to code  $M$  and the significance set  $\Lambda_{\Delta/2}$  of size  $M$  is

$$R_0 = \log_2 N + \log_2 \binom{N}{M} \sim M \left(1 + \log_2 \frac{N}{M}\right).$$

Let us decompose  $R_1 = R_a + R_s$ , where  $R_a$  is the number of bits that code the amplitude of the  $M$  significant coefficients of  $f$ , and  $R_s$  is the number of bits that code their sign. Clearly,  $0 \leq R_s \leq M$ . The amplitude of coefficients is coded with a logarithmic variable-length code, which does not depend on the distribution of these coefficients.

Let  $p_j$  be the fraction of  $M$  significant coefficients such that  $|\mathcal{Q}(f_B^0[k])| = j\Delta$ , and thus  $|f_B^r[k]| \in [(j - 1/2)\Delta, (j + 1/2)\Delta]$ . Since  $|f_B^r[k]| \sim C k^{-s}$ ,

$$p_j M \sim C^{1/s} \Delta^{-1/s} (j - 1/2)^{-1/s} - C^{1/s} \Delta^{-1/s} (j + 1/2)^{-1/s} \sim s^{-1} C^{1/s} \Delta^{-1/s} j^{-1/s-1}.$$

But  $M \sim \sim C^{1/s} \Delta^{-1/s}$  so

$$p_j \sim s^{-1} j^{-1/s-1}. \quad (10.57)$$

Let us consider a logarithmic variable-length code

$$l_j = \log_2(\pi^2/6) + 2 \log_2 j.$$

which satisfies the Kraft inequality (10.6) because

$$\sum_{j=1}^{+\infty} 2^{-l_j} = \frac{6}{\pi^2} \sum_{j=1}^{+\infty} j^{-2} = 1.$$

This variable-length code produces a bit budget

$$R_a = -M \sum_{j=1}^{+\infty} p_j l_j \sim M s^{-1} \sum_{j=1}^{+\infty} j^{-1-1/s} (\log_2(\pi^2/6) + 2 \log_2 j) \sim M. \quad (10.58)$$

As a result,  $R_1 = R_s + R_a \sim M$ , and thus

$$R = R_0 + R_1 \sim M \left( 1 + \log_2 \frac{N}{M} \right). \quad (10.59)$$

Inverting this equation gives

$$M \sim R \left( 1 + \log_2 \frac{N}{R} \right)^{-1},$$

and since  $d(R, f) \sim C^2 M^{1-2s}$  in (10.56), it implies (10.53). ■

The equivalence sign  $\sim$  means that lower and upper bounds of  $d(R, f)$  are obtained by multiplying the right expression of (10.53) by two constants  $A, B > 0$  that are independent of  $C, R$ , and  $N$ . Thus, it specifies the increase of  $d(R, f)$  as  $R$  decreases. Theorem 10.8 proves that at low bit rates, the distortion is proportional to  $R^{1-2s}$ , as opposed to  $2^{-2R/N}$  in the high bit rate distortion formula (10.47). The bit budget is dominated by the geometric bit budget  $R_0$ , which codes the significance map. At low bit rates, to minimize the distortion one must find a basis  $\mathcal{B}$  that yields the smallest  $M$ -term approximation error.

### Twice-Larger Zero Bin

Since the high-resolution quantization hypothesis does not hold, a uniform quantizer does not minimize the distortion rate. The wavelet coefficient histograms in Figure 10.9 are highly peaked and can be modeled in a first approximation by

Laplacian distributions having an exponential decay  $p(x) = \mu/2 e^{-\mu|x|}$ . For Laplacian distributions, one can prove [394] that optimal quantizers that minimize the distortion rate with an entropy coder have a zero bin  $[-\Delta, \Delta]$  that is twice larger than other quantization bins, which must have a constant size  $\Delta$ .

Doubling the size of the zero bin often improves the distortion at low bit rates. This is valid for wavelet transform codes but also for image transform codes in block cosine bases [363]. It reduces the proportion of significant coefficients and improves the bit budget by a factor that is not offset by the increase of the quantization error. A larger zero bin increases the quantization error too much, degrading the overall distortion rate. Thus, the quantizer becomes

$$Q(x) = \begin{cases} 0 & \text{if } |x| < \Delta \\ \text{sign}(x) (\lfloor x/\Delta \rfloor + 1/2) \Delta & \text{if } |x| \geq \Delta \end{cases} \quad (10.60)$$

and the significance map becomes  $\Lambda_\Delta = \{m : |f_B[m]| \leq \Delta\}$ . Theorem 10.8 remains valid for this quantizer with a twice-larger zero bin. Modifying the size of other quantization bins has a marginal effect. One can verify that the distortion rate equivalence (10.53) also holds for a nonuniform quantizer that is adjusted to minimize the distortion rate.

### Bounded Variation Images

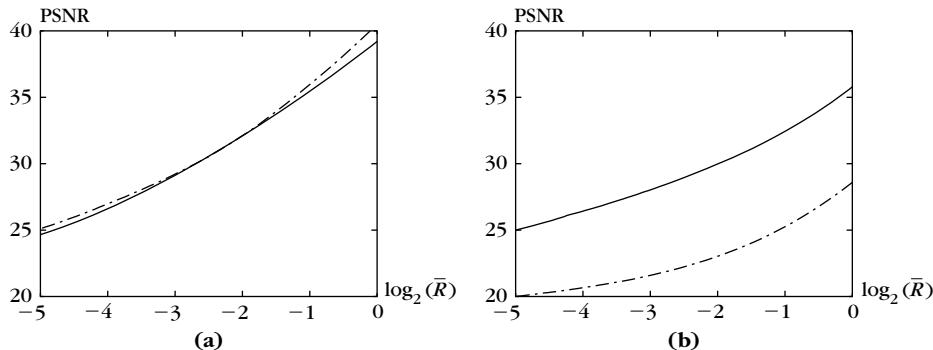
Section 2.3.3 explains that large classes of images have a bounded total variation that is proportional to the length of their contours. Theorem 9.17 proves that the sorted wavelet coefficients  $f_B^r[k]$  of bounded variation images satisfy  $|f_B^r[k]| = O(\|f\|_V k^{-1})$ . If the image is discontinuous along an edge curve, then it creates large-amplitude wavelet coefficients and  $|f_B^r[k]| \sim \|f\|_V k^{-1}$ . This decay property is verified by the wavelet coefficients of the Lena and boat images, which can be considered as discretizations of bounded variation functions. Theorem 10.8 derives for  $s = 1$  that

$$d(R, f) \sim \|f\|_V^2 R^{-1} \left( 1 + \log_2 \frac{N}{R} \right). \quad (10.61)$$

Figure 10.11(a) shows the PSNR computed numerically from a wavelet transform code of the Lena and boats images. Since  $PSNR = 10 \log_{10} d(R, f) + K$ , it results from (10.61) that it increases almost linearly as a function of  $\log_2 \bar{R}$ , with a slope of  $10 \log_{10} 2 \approx 3$  db/bit for  $\bar{R} = R/N$ .

### More Irregular Images

The mandrill and GoldHill are examples of images that do not have a bounded variation. This appears in the fact that their sorted wavelet coefficients satisfy  $|f_B^r[k]| \sim C k^{-s}$  for  $s < 1$ . Since  $PSNR = 10 \log_{10} d(R, f) + K$ , it results from (10.53) that it increases with a slope of  $(2s - 1) 10 \log_{10} 2$  as a function of  $\log_2 \bar{R}$ . For the GoldHill image,  $s \approx 0.8$ , so the PSNR increases by 1.8 db/bit. Mandrill is even more irregular, with  $s \approx 2/3$ , so at low bit rates  $\bar{R} < 1/4$  the PSNR increases by only 1 db/bit. Such images can be modeled as the discretization of functions in Besov spaces with a regularity index  $s/2 + 1/2$  smaller than 1.

**FIGURE 10.11**

PSNR as a function of  $\log_2(\bar{R})$ . (a) Lena image (solid line) and boats image (dotted line)  
 (b) GoldHill image (solid line) and mandrill image (dotted line).

### **Distortion Rate for Analog Signals**

The input signal  $f[n]$  is most often the discretization of an analog signal  $\bar{f}(x)$ , and can be written as  $f[n] = \langle \bar{f}, \phi_n \rangle$  where  $\{\phi_n\}_{0 \leq n < N}$  is a Riesz basis of an approximation space  $\mathbf{U}_N$ . To simplify explanations, we suppose that this basis is orthonormal. Let  $\tilde{f}_N$  be the orthogonal projection of  $f$  in  $\mathbf{U}_N$  that can be recovered from  $f[n]$ . At the end of the processing chain, the coded signal  $\tilde{f}[n]$  is converted into an analog signal in  $\mathbf{U}_N$ :

$$\tilde{f}_N(x) = \sum_{n=0}^{N-1} \tilde{f}[n] \phi_n(x).$$

Since  $\{\phi_n\}_{0 \leq n < N}$  is orthonormal, the norms over analog and discrete signals are equal:  $\|\tilde{f}_N - \tilde{f}\| = \|f - \tilde{f}\|$ .

The overall analog distortion rate  $d(R, \bar{f}) = \|\bar{f} - \tilde{f}_N\|^2$  satisfies

$$d(R, \bar{f}) = \|\bar{f}_N - \bar{f}\|^2 + \|\bar{f}_N - \tilde{f}_N\|^2 = \|\bar{f}_N - \bar{f}\|^2 + \|f - \tilde{f}\|^2.$$

The sampling resolution  $N$  can be chosen so that the linear approximation error  $\|\bar{f}_N - \bar{f}\|^2$  is smaller or of the same order as the compression error  $\|f - \tilde{f}\|^2$ . For most functions, such as bounded variation functions or Besov space functions, the linear approximation error satisfies  $\|\bar{f}_N - \bar{f}\|^2 = O(N^{-\beta})$  for some  $\beta > 0$ . If discrete distortion rate  $\|f - \tilde{f}\|^2$  satisfies the decay (10.53) of Theorem 10.8, then for  $N = R^{(2s-1)/\beta}$ , we get

$$d(R, \bar{f}) = O\left(R^{1-2s} |\log_2 R|^{2s-1}\right). \quad (10.62)$$

This result applies to bounded variation images coded in wavelet bases. Donoho [215] proved that for such functions the decay exponent  $R^{-1}$  cannot be improved

by any other signal coder. In that sense, a wavelet transform coding is optimal for bounded variation images. Functions in Besov spaces coded in wavelet bases also have a distortion rate that satisfies (10.62) for an exponent  $s$  that depends on the space. The optimality of wavelet transform codes in Besov spaces is also studied in [173].

### 10.4.2 Embedded Transform Coding

For rapid transmission or fast browsing from a database, a coarse signal approximation should be quickly provided, and progressively enhanced as more bits are transmitted. Embedded coders offer this flexibility by grouping the bits in order of significance. The decomposition coefficients are sorted and the first bits of the largest coefficients are sent first. A signal approximation can be reconstructed at any time from the bits already transmitted.

Embedded coders store geometric bit planes and can thus take advantage of any prior information about the location of large versus small coefficients. Such prior information is available for natural images decomposed on wavelet bases. Section 10.5.2 explains how JPEG-2000 uses this prior information to optimize wavelet coefficients coding.

#### *Embedding*

The decomposition coefficients  $f_{\mathcal{B}}[m] = \langle f, g_m \rangle$  are partially ordered by grouping them in index sets  $\Theta_k$  defined for any  $k \in \mathbb{Z}$  by

$$\Theta_k = \{m : 2^k \leq |f_{\mathcal{B}}[m]| < 2^{k+1}\} = \Lambda_{2^k} - \Lambda_{2^{k+1}},$$

which are the difference between two significance maps for twice-larger quantization steps. The set  $\Theta_k$  is coded with a binary significance map  $b_k[m]$ :

$$b_k[m] = \begin{cases} 0 & \text{if } m \notin \Theta_k \\ 1 & \text{if } m \in \Theta_k. \end{cases} \quad (10.63)$$

An embedded algorithm quantizes  $f_{\mathcal{B}}[m]$  uniformly with a quantization step  $\Delta = 2^n$  that is progressively reduced. Let  $m \in \Theta_k$  with  $k \geq n$ . The amplitude  $|Q(f_{\mathcal{B}}[m])|$  of the quantized number is represented in base 2 by a binary string with nonzero digits between bit  $k$  and bit  $n$ . Bit  $k$  is necessarily 1 because  $2^k \leq |Q(f_{\mathcal{B}}[m])| < 2^{k+1}$ . Thus,  $k - n$  bits are sufficient to specify this amplitude, to which is added 1 bit for the sign.

The embedded coding is initiated with the largest quantization step that produces at least one nonzero quantized coefficient. Each coding iteration with a reduced quantization step  $2^n$  is called a *bit plane coding pass*. In the loop, to reduce the quantization step from  $2^{n+1}$  to  $2^n$ , the algorithm first codes the significance map  $b_n[m]$ . It then codes the sign of  $f_{\mathcal{B}}[m]$  for  $m \in \Theta_n$ . Afterwards, the code stores the  $n$ th bit of all amplitudes  $|Q(f_{\mathcal{B}}[m])|$  for  $m \in \Theta_k$  with  $k > n$ . If necessary, the coding

precision is improved by decreasing  $n$  and continuing the encoding. The different steps of the algorithm can be summarized as follows [422]:

1. *Initialization:* Store the index  $n$  of the first nonempty set  $\Theta_n$ :

$$n = \left\lceil \sup_m \log_2 |f_B[m]| \right\rceil. \quad (10.64)$$

2. *Significance coding:* Store the significance map  $b_n[m]$  for  $m \notin \Theta_{n+1}$ .
3. *Sign coding:* Code the sign of  $f_B[m]$  for  $m \in \Theta_n$ .
4. *Quantization refinement:* Store the  $n$ th bit of all coefficients  $|f_B[m]| > 2^{n+1}$ . These are coefficients that belong to some set  $\Theta_k$  for  $k > n$ , having a position already stored. Their  $n$ th bit is stored in the order in which its position was recorded in the previous passes.
5. *Precision refinement:* Decrease  $n$  by 1 and go to step 2.

This algorithm may be stopped at any time in the loop, providing a code for any specified number of bits. The decoder then restores the significant coefficients up to a precision  $\Delta = 2^n$ . In general, only part of the coefficients are coded with a precision  $2^n$ . Valid truncation points of a bit stream correspond to the end of the quantization refinement step for a given pass, so that all coefficients are coded with the same precision.

### **Distortion Rate**

The distortion rate is analyzed when the algorithm is stopped at step 4. All coefficients above  $\Delta = 2^n$  are uniformly quantized with a bin size  $\Delta = 2^n$ . The zero-quantization bin  $[-\Delta, \Delta]$  is therefore twice as big as the other quantization bins, which improves coder efficiency as previously explained.

Once the algorithm stops, we denote by  $M$  the number of significant coefficients above  $\Delta = 2^n$ . The total number of bits of the embedded code is

$$R = R_0^e + R_1^e,$$

where  $R_0^e$  is the number of bits needed to code all significance maps  $b_k[m]$  for  $k \geq n$ , and  $R_1^e$  the number of bits used to code the amplitude of the quantized significant coefficients  $Q(f_B[m])$ , knowing that  $m \in \Theta_k$  for  $k > n$ .

To appreciate the efficiency of this embedding strategy, let us compare the bit budget  $R_0^e + R_1^e$  to the number of bits  $R_0 + R_1$  used by the direct transform code from Section 10.4.1. The value  $R_0$  is the number of bits that code the overall significance map

$$b[m] = \begin{cases} 0 & \text{if } |f_B[m]| \leq \Delta \\ 1 & \text{if } |f_B[m]| > \Delta \end{cases} \quad (10.65)$$

and  $R_1$  is the number of bits that code the quantized significant coefficients.

An embedded strategy codes  $Q(f_B[m])$  knowing that  $m \in \Theta_k$  and thus that  $2^k \leq |Q(f_B[m])| < 2^{k+1}$ , whereas a direct transform code knows only that  $|Q(f_B[m])| > \Delta = 2^n$ . Thus, fewer bits are needed for embedded codes:

$$R_1^e \leq R_1. \quad (10.66)$$

However, this improvement may be offset by the supplement of bits needed to code the significance maps  $\{b_k[m]\}_{k>n}$  of the sets  $\{\Theta_k\}_{k>n}$ . A direct transform code records a single significance map  $b[m]$ , which specifies  $\Lambda_{2^n} = \cup_{k \geq n} \Theta_k$ . It provides less information and is therefore coded with fewer bits:

$$R_0^e \geq R_0. \quad (10.67)$$

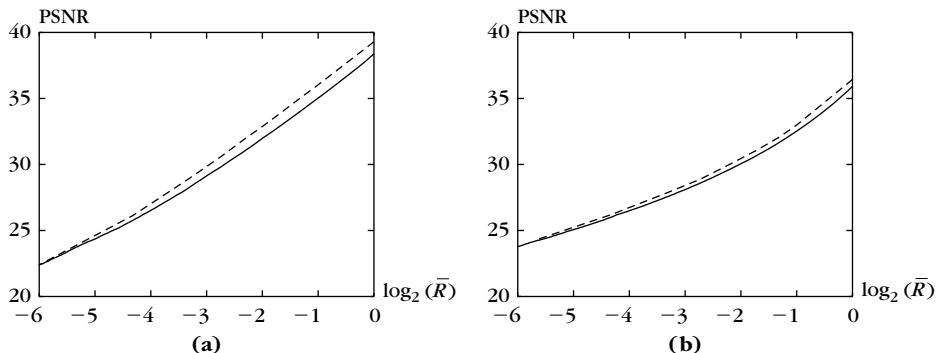
An embedded code brings an improvement over a direct transform code if

$$R_0^e + R_1^e \leq R_0 + R_1.$$

This happens if there is some prior information about the position of large coefficients  $|f_B[m]|$  versus smaller ones. An appropriate coder can then reduce the number of bits needed to encode the partial sorting of all coefficients provided by the significance maps  $\{b_k[m]\}_{k>n}$ . The use of such prior information produces an overhead of  $R_0^e$  relative to  $R_0$  that is smaller than the gain of  $R_1^e$  relative to  $R_1$ . This is the case for most images coded with embedded transform codes implemented in wavelet bases [422] and for the block cosine I basis [492].

Figure 10.12 compares the PSNR of the SPIHT wavelet-embedded code by Said and Pearlman [422] with the PSNR of the direct wavelet transform code that performs an entropy coding of the significance map and of the significance coefficients, described in Section 10.4.1. For any quantization step, both transform codes yield the same distortion but the embedded code reduces the bit budget:

$$R_0^e + R_1^e \leq R_0 + R_1.$$



**FIGURE 10.12**

Comparison of the PSNR obtained with an embedded wavelet transform code (dotted line) and a direct wavelet transform code (solid line): (a) Lena image, and (b) GoldHill image.

As a consequence, the PSNR curve of the embedded code is a translation to the left of the PSNR of the direct transform code. For a fixed bit budget per pixel  $1 \geq R/N \geq 2^{-8}$ , the embedded SPIHT coder gains about 1 db on the Lena image and 1/2 db on the GoldHill image. The GoldHill is an image with more texture. Its wavelet representation is not as sparse as the Lena image, and therefore its distortion rate has a slower decay.

## 10.5 IMAGE-COMPRESSION STANDARDS

The JPEG image-compression standard is a transform code in an orthogonal block cosine basis described in Section 10.5.1. It is still the most common image standard used by digital cameras and for transmission of photographic images on the Internet and cellular phones. JPEG-2000 is the most recent compression standard performing a transform coding in a wavelet basis, summarized in Section 10.5.2. It is mostly used for professional imaging applications.

### 10.5.1 JPEG Block Cosine Coding

The JPEG image-compression standard [478] is a transform coding in a block cosine I basis. Its implementation is relatively simple, which makes it particularly attractive for consumer products.

Theorem 8.12 proves that the following cosine I family is an orthogonal basis of an image block of  $L \times L$  pixels:

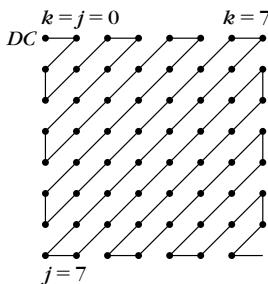
$$\left\{ g_{k,j}[n, m] = \lambda_k \lambda_j \frac{2}{L} \cos \left[ \frac{k\pi}{L} \left( n + \frac{1}{2} \right) \right] \cos \left[ \frac{j\pi}{L} \left( m + \frac{1}{2} \right) \right] \right\}_{0 \leq k, j \leq L} \quad (10.68)$$

with

$$\lambda_p = \begin{cases} 1/\sqrt{2} & \text{if } p = 0 \\ 1 & \text{otherwise.} \end{cases} \quad (10.69)$$

In the JPEG standard, images of  $N$  pixels are divided in  $N/64$  blocks of  $8 \times 8$  pixels. Each image block is expanded in this separable cosine basis with a fast separable DCT-I transform.

JPEG quantizes the block cosine coefficients uniformly. In each block of 64 pixels, a significance map gives the position of zero versus nonzero quantized coefficients. Lower-frequency coefficients are located in the upper right of each block, whereas high-frequency coefficients are in the lower right, as illustrated in Figure 10.13. Many image blocks have significant coefficients only at low frequencies and thus in the upper left of each block. To take advantage of this prior knowledge, JPEG codes the significance map with a run-length code. Each block of 64 coefficients is scanned in zig-zag order as indicated in Figure 10.13. In this scanning order, JPEG registers the size of the successive runs of coefficients quantized to zero, which are efficiently coded together with the values of the following nonzero quantized coefficients.

**FIGURE 10.13**

A block of 64 cosine coefficients has the zero-frequency (*DC*) coefficient at the upper left. The run-length makes a zig-zag scan from low to high frequencies.

Insignificant high-frequency coefficients often produce a long sequence of zeros at the end of the block, which is coded with an end-of-block (EOB) symbol.

In each block  $i$ , there is one cosine vector  $g_{0,0}^i[n, m]$  of frequency zero, which is equal to 1/8 over the block and 0 outside. The inner product  $\langle f, g_{0,0}^i \rangle$  is proportional to the average of the image over the block. Let  $DC^i = Q(\langle f, g_{0,0}^i \rangle)$  be the quantized zero-frequency coefficient. Since the blocks are small, these averages are often close for adjacent blocks, and JPEG codes the differences  $DC^i - DC^{i-1}$ .

### **Weighted Quantization**

Our visual sensitivity depends on the frequency of the image content. We are typically less sensitive to high-frequency oscillatory patterns than to low-frequency variations. To minimize the visual degradation of the coded images, JPEG performs a quantization with intervals that are proportional to weights specified in a table, which is not imposed by the standard. This is equivalent to optimizing a weighted mean-square error (10.34). Table 10.1 is an example of an  $8 \times 8$  weight matrix that is used in JPEG [478]. The weights at the lowest frequencies, corresponding to the upper left portion of Table 10.1, are roughly 10 times smaller than at the highest frequencies, corresponding to the bottom right portion.

### **Distortion Rate**

At 0.25 to 0.5 bit/pixel, the quality of JPEG images is moderate. At 0.2 bit/pixel, Figure 10.14 shows that there are blocking effects due to the discontinuities of the square windows. At 0.75 to 1 bit/pixel, images coded with the JPEG standard are of excellent quality. Above 1 bit/pixel, the visual image quality is perfect. The JPEG standard is often used for  $\bar{R} \in [0.5, 1]$ .

At low bit rates, the artifacts at the block borders are reduced by replacing the block cosine basis by a local cosine basis [40, 87], designed in Section 8.4.4. If the image is smooth over a block, a local cosine basis creates lower-amplitude, high-frequency coefficients, which slightly improves the coder performance. The quantization errors for smoothly overlapping windows also produce more regular

<b>Table 10.1</b> Matrix of Weights $w_{k,j}$							
16	11	10	16	24	40	51	61
12	12	14	19	26	58	60	55
14	13	16	24	40	57	69	56
14	17	22	29	51	87	80	62
18	22	37	56	68	108	103	77
24	35	55	64	81	194	113	92
49	64	78	87	103	121	120	101
72	92	95	98	121	100	103	99

*Note: These weights are used to quantize the block cosine coefficient corresponding to each cosine vector  $g_{k,j}$  [69]. The order is the same as in Figure 10.13.*

gray-level image fluctuations at the block borders. However, the improvement has not been significant enough to motivate replacing the block cosine basis by a local cosine basis in the JPEG standard.

### Implementation of JPEG

The baseline JPEG standard [478] uses an intermediate representation that combines run-length and amplitude values. In each block, the 63 (nonzero frequency) quantized coefficients indicated in Figure 10.13 are integers that are scanned in zig-zag order. A JPEG code is a succession of symbols  $S_1 = (L, B)$  of 8 bits followed by symbols  $S_2$ . The  $L$  variable is the length of a consecutive run of zeros, coded on 4 bits. Thus, its value is limited to the interval  $[0, 15]$ . Actual zero-runs can have a length greater than 15. The symbol  $S_1 = (15, 0)$  is interpreted as a run length of size 16 followed by another run length. When the run of zeros includes the last 63rd coefficient of the block, a special EOB symbol  $S_1 = (0, 0)$  is used, which terminates the coding of the block. For high compression rates, the last run of zeros may be very long. The EOB symbol stops the coding at the beginning of this last run of zeros.

The  $B$  variable of  $S_1$  is coded on 4 bits and gives the number of bits used to code the value of the next nonzero coefficient. Since the image gray-level values are in the interval  $[0, 2^8]$ , one can verify that the amplitude of the block cosine coefficients remains in  $[-2^{10}, 2^{10} - 1]$ . For any integers in this interval, Table 10.2 gives the number of bits used by the code. For example, 70 is coded on  $B = 7$  bits. There are  $2^7$  different numbers that are coded with 7 bits. If  $B$  is nonzero, after the symbol  $S_1$ , the symbol  $S_2$  of length  $B$  specifies the amplitude of the following nonzero coefficient. This variable-length code is a simplified entropy code. High-amplitude coefficients appear less often and are thus coded with more bits.

For  $DC$  coefficients (zero frequency), the differential values  $DC^i - DC^{i-1}$  remain in the interval  $[-2^{11}, 2^{11} - 1]$ . They are also coded with a succession of two symbols.



**FIGURE 10.14**

Image compression with JPEG: left column, 0.5 bit/pixel; right column, 0.2 bit/pixel.

**Table 10.2** The Value of Coefficients Coded on  $B$  Bits

<b><math>B</math></b>	<b>Range of Values</b>
1	-1, 1
2	-3, -2, 2, 3
3	-7 ... -4, 4 ... 7
4	-15 ... -8, 8 ... 15
5	-31 ... -16, 16 ... 31
6	-63 ... -32, 32 ... 63
7	-127 ... -64, 64 ... 127
8	-255 ... -128, 128 ... 255
9	-511 ... -256, 256 ... 511
10	-1023 ... -512, 512 ... 1023

*Note: The values belong to sets of  $2^B$  values that are indicated in the second column.*

In this case,  $S_1$  is reduced to the variable  $B$  that gives the number of bits of the next symbol  $S_2$ , which codes  $DC^i - DC^{i-1}$ .

For both  $DC$  and the other coefficients, the  $S_1$  symbols are encoded with a Huffman entropy code. JPEG does not impose the Huffman tables, which may vary depending on the type of image. An arithmetic entropy code can also be used. For coefficients that are not zero frequency, the  $L$  and  $B$  variables are lumped together because their values are often correlated, and the entropy code of  $S_1$  takes advantage of this correlation.

### 10.5.2 JPEG-2000 Wavelet Coding

The JPEG-2000 image-compression standard is a transform code in a wavelet basis. It introduces typically less distortions than JPEG but this improvement is moderate above 1 bit/pixel. At low bit rates, JPEG-2000 degrades more progressively than JPEG. JPEG-2000 is implemented with an embedded algorithm that provides scalable codes for progressive transmission. Region of interest can also be defined to improve the resolution on specific image parts. Yet, the algorithmic complexity overhead of the JPEG-2000 algorithm has mostly limited its applications to professional image processing, such as medical imaging, professional photography, or digital cinema.

Although mostly inspired by the EBCOT algorithm by Taubman and Marcellin [65], JPEG-2000 is the result of several years of research to optimize wavelet image codes. It gives elegant solutions to key issues of wavelet image coding that will be reviewed together with the description of the standard. Taubman and Marcellin's

book [65] explains all the details. JPEG-2000 brings an improvement of typically more than 1 db relatively to an adaptive arithmetic entropy coding of wavelet coefficients, described in Section 10.4.1. Good visual-quality images are obtained in Figure 10.15 with 0.2 bit/pixel, which considerably improves the JPEG compression results shown in Figure 10.14. At 0.05 bit/pixel the JPEG-2000 recovers a decent approximation, which is not possible with JPEG.

### ***Choice of Wavelet***

To optimize the transform code one must choose a wavelet basis that produces as many zero-quantized coefficients as possible. A two-dimensional separable wavelet basis is constructed from a one-dimensional wavelet basis generated by a mother wavelet  $\psi$ . The wavelet choice does not modify the asymptotic behavior of the distortion rate (10.61) but it influences the multiplicative constant. Three criteria may influence the choice of  $\psi$ : number of vanishing moments, support size, and regularity.

High-amplitude coefficients occur when the supports of the wavelets overlap a brutal transition like an edge. The number of high-amplitude wavelet coefficients created by an edge is proportional to the width of the wavelet support, which should thus be as small as possible. For smooth regions, wavelet coefficients are small at fine scales if the wavelet has enough vanishing moments to take advantage of the image regularity. However, Theorem 7.9 shows that the support size of  $\psi$  increases proportionally to the number of vanishing moments. The choice of an optimal wavelet is therefore a trade-off between the number of vanishing moments and support size.

The wavelet regularity is important for reducing the visibility of artifacts. A quantization error adds a wavelet multiplied by the amplitude of the quantized error to the image. If the wavelet is irregular, the artifact is more visible because it looks like an edge or a texture patch [88]. This is the case for Haar wavelets. Continuously differentiable wavelets produce errors that are less visible, but more regularity often does not improve visual quality.

To avoid creating large-amplitude coefficients at the image border, it is best to use the folding technique from Section 7.5.2, which is much more efficient than the periodic extension from Section 7.5.1. However, it requires using wavelets that are symmetric or antisymmetric. Besides Haar, there is no symmetric or antisymmetric wavelet of compact support that generates an orthonormal basis. Biorthogonal wavelet bases that are nearly orthogonal can be constructed with symmetric or antisymmetric wavelets. Therefore, they are used more often for image compression.

Overall, many numerical studies have shown that the symmetric 9/7 biorthogonal wavelets in Figure 7.15 give the best distortion rate performance for wavelet image-transform codes. They provide an appropriate trade-off between the vanishing moments, support, and regularity requirements. This biorthogonal wavelet basis is nearly orthogonal and thus introduces no numerical instability. They have an efficient lifting implementation, described in Section 7.8.5. JPEG-2000 offers the choice to use the symmetric 5/3 biorthogonal wavelets, which can be implemented

**FIGURE 10.15**

JPEG-2000 transform coding: left column, 0.2 bit/pixel; right column, 0.05 bit/pixel.

with fewer lifting operations and exact integer operations. The 5/3 wavelet coefficients can be coded as scaled integers and thus can be restored exactly if the quantization step is sufficiently small. This provides a lossless coding mode to JPEG-2000, which means a coding with no error. Lossless coding yields much smaller compression ratios, typically of about 1.7 for most images [65].

### ***Intra- and Cross-Scale Correlation***

The significance maps in Figure 10.10 show that significant coefficients tend to be aggregated along contours or in textured regions. Indeed, wavelet coefficients have a large amplitude where the signal has sharp transitions. At each scale and for each direction, a wavelet image coder can take advantage of the correlation between neighbor wavelet coefficient amplitude, induced by the geometric image regularity. This was not done by the wavelet coder from Section 10.4.1, which makes a binary encoding of each coefficient independently from its neighbors. Taking advantage of this intrascale amplitude correlation is an important source of improvement for JPEG-2000.

Figure 10.10 also shows that wavelet coefficient amplitudes are often correlated across scales. If a wavelet coefficient is large and thus significant, the coarser scale coefficient located at the same position is also often significant. Indeed, the wavelet coefficient amplitude often increases when the scale increases. If an image  $f$  is uniformly Lipschitz  $\alpha$  in the neighborhood of  $(x_0, y_0)$ , then (6.58) proves that for wavelets  $\psi_{j,p,q}^l$  located in this neighborhood, there exists  $A \geq 0$  such that

$$|\langle f, \psi_{j,p,q}^l \rangle| \leq A 2^{j(\alpha+1)}.$$

The worst singularities are often discontinuities, so  $\alpha \geq 0$ . This means that in the neighborhood of singularities without oscillations, the amplitude of wavelet coefficients decreases when the scale  $2^j$  decreases. This property is not always valid, in particular for oscillatory patterns. High-frequency oscillations create coefficients at large scales  $2^j$  that are typically smaller than at the fine scale that matches the period of oscillation.

To take advantage of such correlations across scales, wavelet zero-trees have been introduced by Lewis and Knowles [348]. Shapiro [432] used this zero-tree structure to code the embedded significance maps of wavelet coefficients by relating these coefficients across scales with quad-trees. This was further improved by Said and Pearlman [422] with a set partitioning technique. Yet, for general natural images, the coding improvement obtained by algorithms using cross-scale correlation of wavelet coefficient amplitude seems to be marginal compared to approaches that concentrate on intrascale correlation due to geometric structures. This approach was, therefore, not retained by the JPEG-2000 expert group.

### ***Weighted Quantization and Regions of Interest***

Visual distortions introduced by quantization errors of wavelet coefficients depend on the scale  $2^j$ . Errors at large scales are more visible than at fine scales [481]. This can be taken into account by quantizing the wavelet coefficients with intervals

$\Delta_j = \Delta w_j$  that depend on the scale  $2^j$ . For  $\bar{R} \leq 1$  bit/pixel,  $w_j = 2^{-j}$  is appropriate for the three finest scales. The distortion in (10.34) shows that choosing such weights is equivalent to minimizing a weighted mean-square error.

Such a weighted quantization is implemented like in (10.35) by quantizing weighted wavelet coefficients  $f_B[m]/w_j$  with a uniform quantizer. The weights are inverted during the decoding process. JPEG-2000 supports a general weighting scheme that codes weighted coefficients  $w[m]f_B[m]$  where  $w[m]$  can be designed to emphasize some region of interest  $\Omega \subset [0, 1]^2$  in the image. The weights are set to  $w[m] = w > 1$  for the wavelet coefficients  $f_B[m] = \langle f, \psi_{j,p,q}^l \rangle$  where the support of  $\psi_{j,p,q}^l$  intersects  $\Omega$ . As a result, the wavelet coefficients inside  $\Omega$  are given a higher priority during the coding stage, and the region  $\Omega$  is coded first within the compressed stream. This provides a mechanism to more precisely code regions of interest in images—for example, a face in a crowd.

### Overview of the JPEG-2000 Coder

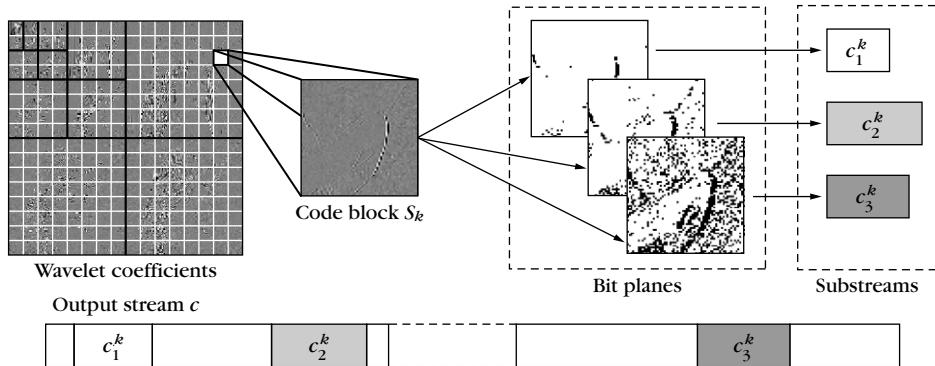
The JPEG-2000 compression standard [65] implements a generic embedded coder, described in Section 10.4.2, and takes advantage of the intrascale dependencies between wavelet coefficients of natural images. The three primitive operations of the algorithm, which code the significance, the sign, and the amplitude refinement, are implemented with a binary adaptive arithmetic coder. This coder exploits the aggregation of large-amplitude wavelet coefficients by creating a small number of context tokens that depend on the neighbors of each coefficient. The coding of a binary symbol uses a conditional probability that depends on the context value of this coefficient.

Wavelet coefficients are subdivided in squares (code blocks) that are coded independently. Each code block  $S_k$  is a square of  $L \times L$  coefficients  $\{f_B[m]\}_{m \in S_k}$ , with typical sizes  $L = 32$  or  $L = 64$ . The coding of each code block  $S_k$  follows the generic embedded coder detailed in Section 10.4.2 and generates a binary stream  $c^k$ . This stream is composed of substreams  $c^k = (c_1^k, c_2^k, \dots)$  where each  $c_n^k$  corresponds to the bit plane of the  $n$ th bit. The whole set of substreams  $\{c_n^k\}_{k,n}$  is reorganized in a global embedded binary stream  $c$  that minimizes the rate distortion for any bit budget. The square segmentation improves the scalability of the bit stream, since the generated code can be truncated optimally at a very large number of points in the stream. It also provides more efficient, parallel, and memory friendly implementation of the wavelet coder. Figure 10.16 illustrates the successive coding steps of JPEG-2000.

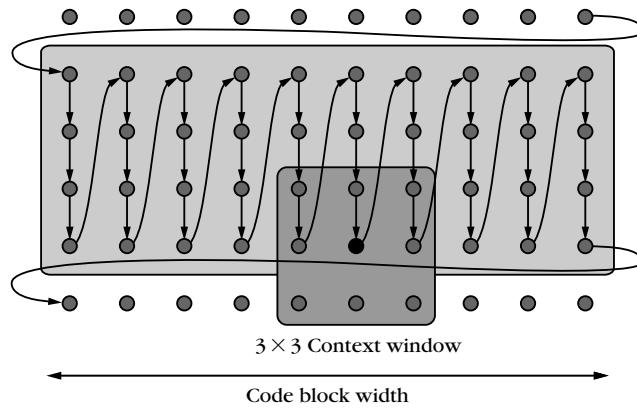
### Conditional Coding of Bit Planes

The coefficients  $f_B[m]$  in each code block are processed sequentially using an ordering of the positions  $m \in S_k$ . The specific choice of ordering used in the JPEG-2000 standard is depicted in Figure 10.17. It scans each square in bands of size  $4 \times L$  coefficients.

The conditional coding of coefficients uses an instantaneous significance  $\sigma[m]$  for each bit plane  $n$ . If the coefficient has not yet been processed, then  $\sigma[m]$  carries

**FIGURE 10.16**

Overview of JPEG-2000 compression process.

**FIGURE 10.17**

JPEG-2000 scans square blocks of  $L \times L$  wavelet coefficients, as bands of size  $4 \times L$ .

the value of the previous pass, and  $\sigma[m] = 0$  if  $|f_B[m]| \geq 2^{n+1}$ , and  $\sigma[m] = 1$  otherwise. If the coefficient has already been processed, then  $\sigma[m] = 0$  if  $|f_B[m]| \geq 2^n$ , and  $\sigma[m] = 1$  otherwise. Observe that  $\sigma[m]$  can be computed by both the coder and the decoder from the information that has already been processed.

In the following, we explain how the three primitive operations—significance, sign, and amplitude refinement coding—of the generic embedded coder are implemented using a conditional arithmetic coder.

### **Significance Coding**

If the coefficient of index  $m$  was insignificant at the previous bit planes, meaning  $|f_B[m]| \leq 2^{n+1}$ , JPEG-2000 encodes the significance bit  $b_n[m] \in \{0, 1\}$ . JPEG-2000

takes advantage of redundancies between neighboring coefficients in the square  $S_k$ . A conditional arithmetic coder stores the value of  $b_n[m]$  by using a conditional probability distribution,

$$P_{\varepsilon, \eta} = P(b_n[m] = \varepsilon | \kappa^0[m] = \eta) \quad \text{for } \varepsilon \in \{0, 1\} \quad \text{and} \quad \eta \in \{0, \dots, \kappa^0_{\max} - 1\},$$

where the context  $\kappa^0[m]$  for a coefficient  $f_B[m] = \langle f, \psi_{j,p,q}^I \rangle$  at position  $(p, q)$  is evaluated by both the compressor and decompressor from a set of fixed rules using the significance  $\sigma[(p, q) + \varepsilon]$  with  $\varepsilon = (\pm 1, \pm 1)$  in a  $3 \times 3$  context window around the position  $(p, q)$ . These rules can be found in [65].

### **Sign Coding**

If the coefficient of index  $m$  has been coded as significant, which means that  $b_n[m] = 1$ , then JPEG-2000 encodes its sign  $s[m] = \text{sign}(f_B[m]) \in \{+1, -1\}$  with an adaptive arithmetic coder that also uses a conditional probability:

$$P(s[m] = \varepsilon | \kappa^s[m] = \eta) \quad \text{for } \varepsilon \in \{+1, -1\} \quad \text{and} \quad \eta \in \{0, \dots, \kappa^s_{\max} - 1\}.$$

The context depends on the  $3 \times 3$  neighboring coefficients, which can be significant and positive, significant and negative, or insignificant, thus allowing 81 unique configurations. JPEG-2000 uses a reduced set of only  $\kappa^s_{\max} = 5$  context values for  $\kappa^s[m]$  that are calculated from horizontal and vertical sign agreement indicators [65].

### **Amplitude Refinement Coding**

If the coefficient is already significant from the previous bit plane, which means that  $|f_B[m]| > 2^{n+1}$ , then JPEG-2000 codes  $\gamma_n[m]$ , which is the  $n$ th bit of  $|f_B[m]|$ , to refine the amplitude of the quantized coefficients, with an adaptive arithmetic coder using a conditional probability:

$$P(\gamma_n[m] = \varepsilon | \kappa^a[m] = \eta) \quad \text{for } \varepsilon \in \{0, 1\} \quad \text{and} \quad \eta \in \{0, \dots, \kappa^a_{\max} - 1\}.$$

Figure 10.9 shows that the histograms of wavelet coefficients of natural images are usually highly peaked around zero. For a small quantized value  $Q(|f_B[m]|)$ , the event  $\gamma_n[m] = 0$  is thus more likely than the event  $\gamma_n[m] = 1$ . In contrast, for a large quantized value  $Q(|f_B[m]|)$ , both events have approximately the same probability. JPEG-2000 uses a context  $\kappa^a[m]$  that discriminates these two cases by checking whether  $|f_B[m]| > 2^{n+2}$  (one already knows that  $|f_B[m]| > 2^{n+1}$ ). This context also takes into account the significance of the neighboring coefficients by using the significance context  $\kappa^0[m]$ .

### **Optimal Truncation Points and Substream Packing**

The bit plane coding algorithm generates an embedded bit stream  $c^k$  for each code block  $\{f_B[m]\}_{m \in S_k}$  of wavelet coefficients. Each bit stream  $c^k$  has a set of valid truncation points  $\{R_n^k, R_{n_{\max}-1}^k, \dots\}$  that corresponds to the end of the coding pass for each bit plane. The quantity  $R_n^k - R_{n+1}^k$  is the number of bits of the substream  $c_n^k$  generated by the coding of the bit plane of index  $n$  of the code block  $S_k$ . The goal

of the rate distortion optimization is to pack together all the substreams  $\{c_n^k\}_{k,n}$  in an optimized order so as to minimize the resulting global distortion when the code is interrupted at a truncation point.

JPEG-2000 offers a mechanism of fractional bit plane coding that further increases the number of these valid truncation points. These fractional coding passes do not code the whole significance map  $b_n$  at once, but rather begin by coding only those coefficient positions  $(p, q)$  that have at least one significant neighbor  $(p, q) + \varepsilon$  such that  $\sigma[(p, q) + \varepsilon] = 1$ . The sign and magnitude refinement pass are then applied, and afterward the coefficients that have not been processed in the current bit plane are encoded. In the following, we assume a set of valid truncation points  $\{R_n^k\}_n$  for each code block  $S_k$ .

Assuming an orthogonal wavelet transform, the distortion at the truncation point  $R_n^k$ , after the coding of the bit plane of index  $n$ , is

$$d_n^k = \sum_{m \in S_k} |Q(f_B[m]) - f_B[m]|^2, \quad (10.70)$$

where  $Q$  is the quantizer of bin size  $2^n$ . This distortion can be computed by the encoder from the already coded bit planes of indexes greater than  $n$ . JPEG-2000 uses the biorthogonal 9/7 wavelet transform, which is close to being orthogonal, so the distortion computation 10.70, although not exact, is accurate enough in practice.

For a given number of bits  $R$ , we must find for each code block  $S_k$  an optimized truncation point  $R_{n_k}^k$  of index  $n_k$  that solves

$$\min_{\{n_k\}_k} \sum_k d_{n_k}^k \quad \text{subject to} \quad \sum_k R_{n_k}^k \leq R. \quad (10.71)$$

This optimal set  $\{n_k\}_k$  depends on the desired total number of bits  $R$ . To obtain an embedded stream  $c$ , the rate distortion optimization (10.70) must be computed for an increasing set of  $\ell_{\max}$  bit budgets  $\{R^{(\ell)}\}_{0 \leq \ell < \ell_{\max}}$ . For each number of bits  $R^{(\ell)}$ , (10.70) is solved with  $R = R^{(\ell)}$ , defining a set  $n_k^{(\ell)}$  of optimal truncation points. The final stream  $c$  is obtained by successively appending the substreams  $c_n^k$  for  $n_k^{(\ell-1)} < n \leq n_k^{(\ell)}$ .

### Rate Distortion Optimization

To build the final stream  $c$ , the optimization (10.70) is solved for a large number of bit budgets  $R = R^{(\ell)}$ . As in Section 10.3.1, following a classic distortion rate approach [392, 435], the constraint minimization is replaced by a Lagrangian optimization. The resulting rate distortion Lagrangian over all square blocks  $S_k$  is

$$\mathcal{L}(\{d_{n_k}^k\}_k, \{R_{n_k}^k\}_k) = \sum_k (d_{n_k}^k + \lambda R_{n_k}^k).$$

A set of indexes  $\{n_k\}_k$  that minimizes  $\mathcal{L}$  is necessarily a minimizer of (10.70) for a bit budget  $R = \sum_k R_{n_k}^k$ . The optimization (10.70) is performed by minimizing  $\mathcal{L}$  for many values of  $\lambda$  and by retaining the smallest value of  $\lambda$  that guarantees  $\sum_k R_{n_k}^k \leq R$ .

For a fixed  $\lambda$ , to minimize  $\mathcal{L}$ , an independent minimization is first done over each block  $S_k$  of the partial Lagrangian  $d_n^k + \lambda R_n^k$ . In each block  $S_k$ , it gives truncation points that are associated to a  $\lambda$ . These truncation points are then globally recombined across all blocks  $S_k$  by ordering them according to their Lagrange variable  $\lambda$ , which provides the global sequence of all truncation points and defines the final embedded JPEG-2000 stream.

## 10.6 EXERCISES

- 10.1** <sup>1</sup> Let  $X$  be a random variable that takes its values in  $\{x_k\}_{1 \leq k \leq 7}$  with probabilities  $\{0.49, 0.26, 0.12, 0.04, 0.04, 0.03, 0.02\}$ .
- Compute the entropy  $\mathcal{H}(X)$ . Construct a binary Huffman code and calculate the average bit rate  $R_X$ .
  - Suppose that the symbols are coded with digits that may take three values  $(-1, 0, 1)$  instead of two as in a bit representation. Variable-length ternary prefix codes can be represented with ternary trees. Extend the Huffman algorithm to compute a ternary prefix code for  $X$  that has a minimal average length.
- 10.2** <sup>1</sup> Let  $x_1$  be the symbol of highest probability of a random variable  $X$ , and  $l_1$  the length of its binary word in a Huffman code. Show that if  $p_1 > 2/5$ , then  $l_1 = 1$ . Verify that if  $p_1 < 1/3$ , then  $l_1 \geq 2$ .
- 10.3** <sup>1</sup> Let  $X$  be a random variable equal to  $x_1$  or  $x_2$  with probabilities  $p_1 = 1 - \varepsilon$  and  $p_2 = \varepsilon$ . Verify that  $\mathcal{H}(X)$  converges to 0 when  $\varepsilon$  goes to 0. Show that the Huffman code has an average number of bits that converges to 1 when  $\varepsilon$  goes to 0.
- 10.4** <sup>2</sup> Prove the Huffman code Theorem 10.2.
- 10.5** <sup>2</sup> Let  $\mathcal{H}_0 = (N - M) \log_2(N/N - M) + M \log_2(N/M)$  be the entropy of a binary coding of the position of  $M$  significant coefficients among  $N$  as in (10.49). Show that  $\mathcal{H}_0 \leq \log_2 \binom{N}{M}$  and compute the difference between the two, by using the Stirling formula  $\lim_{n \rightarrow \infty} (2\pi n)^{-1/2} (n/e)^{-n} n! = 1$ .
- 10.6** <sup>2</sup> Let  $X$  be a random variable with a probability density  $p(x)$ . Let  $Q$  be a quantizer with quantization bins that are  $\{(y_{k-1}, y_k]\}_{1 \leq k \leq K}$ .
- Prove that  $E\{|X - Q(X)|^2\}$  is minimum if and only if

$$Q(x) = x_k = \frac{\int_{y_{k-1}}^{y_k} x p(x) dx}{\int_{y_{k-1}}^{y_k} p(x) dx} \quad \text{for } x \in (y_{k-1}, y_k].$$

- Suppose that  $p(x)$  is a Gaussian with variance  $\sigma^2$ . Find  $x_0$  and  $x_1$  for a “1 bit” quantizer defined by  $y_0 = -\infty$ ,  $y_1 = 0$ , and  $y_2 = +\infty$ .

- 10.7** <sup>1</sup> Consider a pulse code modulation that quantizes each sample of a Gaussian random vector  $F[n]$  and codes it with an entropy code that uses the same number of bits for each  $n$ . If the high-resolution quantization hypothesis is satisfied, prove that the distortion rate is

$$d(\bar{R}) = \frac{\pi e}{6} E\{\|F\|^2\} 2^{-2\bar{R}}.$$

- 10.8** <sup>3</sup> Let  $d = \sum_{m=0}^{N-1} d_m$  be the total distortion of a transform code. We suppose that the distortion rate  $d_m(r)$  for coding the  $m$ th coefficient is convex. Let  $R = \sum_{m=0}^{N-1} R_m$  be the total number of bits.

- (a) Prove with the distortion rate Lagrangian that there exists a unique bit allocation that minimizes  $d(R)$  for  $R$  fixed, and that it satisfies  $\frac{\partial d_m(R_m)}{\partial r} = -\lambda$  where  $\lambda$  is a constant that depends on  $R$ .
- (b) To impose that each  $R_m$  is a positive integer, we use a greedy iterative algorithm that allocates the bits one by one. Let  $\{R_{m,p}\}_{0 \leq m < N}$  be the bit allocation after  $p$  iterations, which means that a total of  $p$  bits have been allocated. The next bit is added to  $R_{k,p}$  such that

$$\left| \frac{\partial d_k(R_{k,p})}{\partial r} \right| = \max_{0 \leq m < N} \left| \frac{\partial d_m(R_{m,p})}{\partial r} \right|.$$

Justify this strategy. Prove that this algorithm gives an optimal solution if all curves  $d_m(r)$  are convex and if  $d_m(n+1) - d_m(n) \approx \frac{\partial d_m(n)}{\partial r}$  for all  $n \in \mathbb{N}$ .

- 10.9** <sup>2</sup> Let  $X[m]$  be a binary first-order Markov chain, which is specified by the transition probabilities  $p_{01} = \Pr\{X[m] = 1 | X[m-1] = 0\}$ ,  $p_{00} = 1 - p_{01}$ ,  $p_{10} = \Pr\{X[m] = 0 | X[m-1] = 1\}$ , and  $p_{11} = 1 - p_{10}$ .

- (a) Prove that  $p_0 = \Pr\{X[m] = 0\} = p_{10}/(p_{10} + p_{01})$  and that  $p_1 = \Pr\{X[m] = 1\} = p_{01}/(p_{10} + p_{01})$ .
- (b) A run-length code records the length  $Z$  of successive runs of 0 values of  $X[m]$  and the length  $I$  of successive runs of 1. Show that if  $Z$  and  $I$  are entropy coded, the average number of bits per sample of the run-length code, denoted  $\bar{R}$ , satisfies

$$\bar{R} \geq \bar{R}_{\min} = p_0 \frac{\mathcal{H}(Z)}{E\{Z\}} + p_1 \frac{\mathcal{H}(I)}{E\{I\}}.$$

- (c) Let  $\mathcal{H}_0 = -p_{01} \log_2 p_{01} - (1 - p_{01}) \log_2 (1 - p_{01})$  and  $\mathcal{H}_1 = -p_{10} \log_2 p_{10} - (1 - p_{10}) \log_2 (1 - p_{10})$ . Prove that

$$\bar{R}_{\min} = \mathcal{H}(X) = p_0 \mathcal{H}_0 + p_1 \mathcal{H}_1,$$

which is the average information gained by moving one step ahead in the Markov chain.

- (d) Suppose that the binary significance map of the transform code of a signal of size  $N$  is a realization of a first-order Markov chain. We denote

$\alpha = 1/E\{Z\} + 1/E\{I\}$ . Let  $M$  be the number of significant coefficients (equal to 1). If  $M \ll N$ , then show that

$$\bar{R}_{\min} \approx \frac{M}{N} \left( \alpha \log_2 \frac{N}{M} + \beta \right) \quad (10.72)$$

with  $\beta = \alpha \log_2 e - 2\alpha \log_2 \alpha - (1-\alpha) \log_2 (1-\alpha)$ .

- (e) Implement a run-length code for the binary significance maps of wavelet image coefficients  $d_j^l[n, m] = \langle f, \psi_{j,n,m}^l \rangle$  for  $j$  and  $l$  fixed. See whether (10.72) approximates the bit rate  $\bar{R}$  calculated numerically as a function of  $N/M$  for the Lena and Barbara images. How does  $\alpha$  vary depending on the scale  $2^j$  and the orientation  $l = 1, 2, 3$ ?
- 10.10** <sup>4</sup> Implement a transform code in a block cosine basis with an arithmetic code and with a local cosine transform over blocks of the same size. Compare the compression rates in DCT-I and local cosine bases, as well as the visual image quality for  $\bar{R} \in [0.2, 1]$ .
- 10.11** <sup>4</sup> Implement a wavelet transform code for color images. Transform the red, green, and blue channels in the color Karhunen-Loëve basis calculated in Exercise 9.2 or with any standard color-coordinate system such as  $Y, U, V$ . Perform a transform code in a wavelet basis with the multichannel decomposition (12.157), which uses the same approximation support for all color channels. Use an arithmetic coder to binary encode together the three color coordinates of wavelet coefficients. Compare numerically the distortion rate of the resulting algorithm with the distortion rate of a wavelet transform code applied to each color channel independently.
- 10.12** <sup>4</sup> Develop a video compression algorithm in a three-dimensional wavelet basis [474]. In the time direction, choose a Haar wavelet in order to minimize the coding delay. This yields zero coefficients at locations where there is no movement in the image sequence. Implement a separable three-dimensional wavelet transform and an arithmetic coding of quantized coefficients. Compare the compression result with an MPEG-2 motion-compensated compression in a DCT basis.

## Denoising

## 11

Removing noise from signals is possible only if some prior information is available. This information is encapsulated in an operator designed to reduce the noise while preserving the signal. Ideally, the joint probability distribution of the signal and the noise is known. Bayesian calculations then derive optimal operators that minimize the average estimation error. However, such probabilistic models are often not available for complex signals such as natural images.

Simpler signal models can be incorporated in the design of a basis or a frame, which takes advantage of known signal properties to build a sparse representation. Efficient nonlinear estimators are then computed by thresholding the resulting coefficients. For one-dimensional signals and images, thresholding estimators are studied in wavelet bases, time-frequency representations, and curvelet frames. Block thresholdings are introduced to regularize these operators, which improves the estimation of audio recordings and images.

The optimality of estimators is analyzed in a minimax framework, where the maximum estimation error is minimized over a predefined set of signals. When signals are not uniformly regular, nonlinear thresholding estimators in wavelet bases are shown to be much more efficient than linear estimators; they nearly reach the minimax risk over different signal classes, such as bounded variation signals and images.

---

### 11.1 ESTIMATION WITH ADDITIVE NOISE

Digital acquisition devices, such as cameras or microphones, output noisy measurements of an incoming analog signal  $\bar{f}(x)$ . These measurements can be modeled by a filtering of  $\bar{f}(x)$  with the sensor responses  $\phi_n(x)$ , to which is added a noise  $W[n]$ :

$$X[n] = \langle \bar{f}, \phi_n \rangle + W[n] \quad \text{for } 0 \leq n < N. \quad (11.1)$$

The noise  $W$  incorporates intrinsic physical fluctuations of the incoming signal. For example, an image intensity with low illumination has a random variation depending on the number of photons captured by each sensor. It also includes noises introduced by the measurement device, such as electronic noises or transmission errors. The aggregated noise  $W$  is modeled by a random vector that has a probability distribution that is supposed to be known a priori and is often Gaussian.

Let us denote the discretized signal by  $f[n] = \langle \bar{f}, \phi_n \rangle$ . This analog-to-digital acquisition is supposed to be stable, so that, according to Section 3.1.3, an analog signal approximation of  $\bar{f}(x)$  can be recovered from  $f[n]$  with a linear projector. One must then optimize an estimation  $\tilde{F}[n] = DX[n]$  of  $f[n]$  calculated from the noisy measurements (11.1) that are rewritten as

$$X[n] = f[n] + W[n] \quad \text{for } 0 \leq n < N.$$

The *decision operator*  $D$  is designed to minimize the estimation error  $f - \tilde{F}$ , measured by a *loss function*.

For audio signals or images, the loss function should measure the perceived audio or visual degradation. A mean-square distance is certainly not a perfect model of perceptual degradation, but it is mathematically simple and sufficiently precise in most applications. Throughout this chapter, the loss function is thus chosen to be a square Euclidean norm. The risk of the estimator  $\tilde{F}$  of  $f$  is the average loss, calculated with respect to the probability distribution of the noise  $W$ :

$$r(D, f) = E\{\|f - DX\|^2\}. \quad (11.2)$$

The decision operator  $D$  is optimized with the prior information available on the signal. The Bayes framework supposes that signals are realizations of a random vector that has a known probability distribution, and a Bayes estimator minimizes the expected risk. A major difficulty is to acquire enough information to model this prior probability distribution. The minimax framework uses simpler deterministic models, which define signals as elements of a predefined set  $\Theta$ . The expected risk cannot be computed, but the maximum risk can be minimized over  $\Theta$ . Section 11.1.2 relates minimax and Bayes estimators through the minimax theorem.

### 11.1.1 Bayes Estimation

A Bayesian model considers signals  $f$  are realizations of a random vector  $F$  with a probability distribution  $\pi$  known a priori. This probability distribution is called the *prior distribution*. The noisy data are thus rewritten as

$$X[n] = F[n] + W[n] \quad \text{for } 0 \leq n < N.$$

We suppose that noise and signal values  $W[k]$  and  $F[n]$  are independent for any  $0 \leq k, n < N$ . The joint distribution of  $F$  and  $W$  is thus the product of the distributions of  $F$  and  $W$ . It specifies the conditional probability distribution of  $F$  given the observed data  $X$ , also called the *posterior distribution*. This posterior distribution is used to optimize the decision operator  $D$  that computes an estimation  $\tilde{F} = DX$  of  $F$  from the data  $X$ .

The *Bayes risk* is the expected risk calculated with respect to the prior probability distribution  $\pi$  of the signal:

$$r(D, \pi) = E_{\pi}\{r(D, F)\}.$$

By inserting (11.2), it can be rewritten as an expected value for the joint probability distribution of the signal and the noise:

$$r(D, \pi) = E\{\|F - \tilde{F}\|^2\} = \sum_{n=0}^{N-1} E\{|F[n] - \tilde{F}[n]|^2\}.$$

Let  $\mathcal{O}_n$  be the set of all operators (linear and nonlinear) from  $\mathbb{C}^N$  to  $\mathbb{C}^N$ . Optimizing  $D$  yields the *minimum Bayes risk*:

$$r_n(\pi) = \inf_{D \in \mathcal{O}_n} r(D, \pi).$$

Theorem 11.1 proves that there exist a *Bayes decision operator*  $D$  and a corresponding *Bayes estimator*  $\tilde{F}$  that achieve this minimum risk.

**Theorem 11.1.** The Bayes estimator  $\tilde{F}$  that yields the minimum Bayes risk  $r_n(\pi)$  is the conditional expectation

$$\tilde{F}[n] = E\{F[n] \mid X[0], X[1], \dots, X[N-1]\}. \quad (11.3)$$

**Proof.** Let  $\pi_n(y)$  be the probability distribution of the value  $y$  of  $F[n]$ . The minimum risk is obtained by finding  $\tilde{F}[n] = D_n(x)$  that minimizes  $r(D_n, \pi_n) = E\{|F[n] - \tilde{F}[n]|^2\}$  for each  $0 \leq n < N$ . This risk depends on the conditional distribution  $P_n(x|y)$  of the data  $X = x$  given  $F[n] = y$ :

$$r(D_n, \pi_n) = \int \int (D_n(x) - y)^2 dP_n(x|y) d\pi_n(y).$$

Let  $P(x) = \int P_n(x|y) d\pi_n(y)$  be the marginal distribution of  $X$  and  $\pi_n(y|x)$  be the posterior distribution of  $F[n]$  given  $X$ . The Bayes formula gives

$$r(D_n, \pi_n) = \int \left[ \int (D_n(x) - y)^2 d\pi_n(y|x) \right] dP(x).$$

The double integral is minimized by minimizing the inside integral for each  $x$ . This quadratic form is minimum when its derivative vanishes:

$$\frac{\partial}{\partial D_n(x)} \int (D_n(x) - y)^2 d\pi_n(y|x) = 2 \int (D_n(x) - y) d\pi_n(y|x) = 0,$$

which implies that

$$D_n(x) = \int y d\pi_n(y|x) = E\{F[n] \mid X = x\},$$

so  $D_n(X) = E\{F[n] \mid X\}$ . ■

### Linear Estimation

The conditional expectation (11.3) is generally a complicated nonlinear function of the data  $\{X[k]\}_{0 \leq k < N}$ , and is difficult to evaluate. To simplify this problem, we

restrict the decision operator  $D$  to be linear. Let  $\mathcal{O}_l$  be the set of all linear operators from  $\mathbb{C}^N$  to  $\mathbb{C}^N$ . The *linear minimum Bayes risk* is:

$$r_l(\pi) = \inf_{D \in \mathcal{O}_l} r(D, \pi).$$

The linear estimator  $\tilde{F} = DX$  that achieves this minimum risk is called the *Wiener estimator*. Theorem 11.2 gives a necessary and sufficient condition that specifies this estimator. We suppose that  $E\{F[n]\} = 0$ , which can be enforced by subtracting  $E\{F[n]\}$  from  $X[n]$  to obtain a zero-mean signal.

**Theorem 11.2.** A linear estimator  $\tilde{F}$  is a Wiener estimator if and only if

$$E\{(F[n] - \tilde{F}[n])X[k]\} = 0 \quad \text{for } 0 \leq k, n < N. \quad (11.4)$$

**Proof.** For each  $0 \leq n < N$ , we must find a linear estimation

$$\tilde{F}[n] = D_n X = \sum_{k=0}^{N-1} h[n, k] X[k],$$

which minimizes

$$r(D_n, \pi_n) = E \left\{ \left( F[n] - \sum_{k=0}^{N-1} h[n, k] X[k] \right) \left( F[n] - \sum_{k=0}^{N-1} h[n, k] X[k] \right)^T \right\}. \quad (11.5)$$

The minimum of this quadratic form is reached if and only if for each  $0 \leq k < N$ ,

$$\frac{\partial r(D_n, \pi_n)}{\partial h[n, k]} = -2 E \left\{ \left( F[n] - \sum_{l=0}^{N-1} h[n, l] X[l] \right) X[k] \right\} = 0,$$

which verifies (11.4). ■

If  $F$  and  $W$  are independent Gaussian random vectors, then the linear optimal estimator is also optimal among nonlinear estimators. Indeed, two jointly Gaussian random vectors are independent if they are noncorrelated [53]. Since  $F[n] - \tilde{F}[n]$  is jointly Gaussian with  $X[k]$ , the noncorrelation (11.4) implies that  $F[n] - \tilde{F}[n]$  and  $X[k]$  are independent for any  $0 \leq k, n < N$ . In this case, we can verify that  $\tilde{F}$  is the Bayes estimator (11.3):  $\tilde{F}[n] = E\{F[n] | X\}$ . Theorem 11.3 computes the Wiener estimator from the covariance  $R_F$  and  $R_W$  of the signal  $F$  and of the noise  $W$ . The properties of covariance operators are described in Section A.6 of the Appendix.

**Theorem 11.3: Wiener.** If the signal  $F$  and the noise  $W$  are independent random vectors of covariance  $R_F$  and  $R_W$ , then the linear Wiener estimator  $\tilde{F} = DF$  that minimizes  $E\{\|\tilde{F} - F\|^2\}$  is

$$\tilde{F} = R_F (R_F + R_W)^{-1} X. \quad (11.6)$$

**Proof.** Let  $\tilde{F}[n]$  be a linear estimator of  $F[n]$ :

$$\tilde{F}[n] = \sum_{l=0}^{N-1} h[n, l] X[l]. \quad (11.7)$$

This equation can be rewritten as a matrix multiplication by introducing the  $N \times N$  matrix  $H = (h[n, l])_{0 \leq n, l < N}$ :

$$\tilde{F} = H X. \quad (11.8)$$

Theorem 11.2 proves that an optimal linear estimator satisfies the noncorrelation condition (11.4), which implies that for  $0 \leq n, k < N$ ,

$$E\{F[n] X[k]\} = E\{\tilde{F}[n] X[k]\} = \sum_{l=0}^{N-1} h[n, l] E\{X[l] X[k]\}.$$

Since  $X[k] = F[k] + W[k]$  and  $E\{F[n] W[k]\} = 0$ , it results that

$$E\{F[n] F[k]\} = \sum_{l=0}^{N-1} h[n, l] (E\{F[l] F[k]\} + E\{W[l] W[k]\}). \quad (11.9)$$

Let  $R_F$  and  $R_W$  be the covariance matrices of  $F$  and  $W$ , defined by  $E\{F[n] F[k]\}$  and  $E\{W[n] W[k]\}$ , respectively. Equation (11.9) can be rewritten as a matrix equation:

$$R_F = H (R_F + R_W).$$

Inverting this equation gives

$$H = R_F (R_F + R_W)^{-1}. \quad \blacksquare$$

The optimal linear estimator (11.6) is simple to compute since it only depends on second-order covariance moments of the signal and of the noise.

### **Estimation in a Karhunen-Loève Basis**

Since a covariance operator is symmetric, it is diagonalized in an orthonormal basis that is called a *Karhunen-Loève basis* or a basis of *principal components*. If the covariance operators  $R_F$  and  $R_W$  are diagonal in the same Karhunen-Loève basis  $\mathcal{B} = \{g_m\}_{0 \leq m < N}$ , then Corollary 11.1 derives from Theorem 11.3 that the Wiener estimator is diagonal in this basis. We write

$$X_{\mathcal{B}}[m] = \langle X, g_m \rangle, \quad F_{\mathcal{B}}[m] = \langle F, g_m \rangle, \quad \tilde{F}_{\mathcal{B}}[m] = \langle \tilde{F}, g_m \rangle,$$

$$W_{\mathcal{B}}[m] = \langle W, g_m \rangle \quad \text{and} \quad \sigma_{\mathcal{B}}[m]^2 = E\{|\langle W, g_m \rangle|^2\}.$$

**Corollary 11.1.** If there exists a Karhunen-Loève basis  $\mathcal{B} = \{g_m\}_{0 \leq m < N}$  that diagonalizes the covariance matrices  $R_F$  and  $R_W$  of  $F$  and  $W$ , then the Wiener estimator that minimizes  $E\{\|\tilde{F} - F\|^2\}$  is

$$\tilde{F} = \sum_{m=0}^{N-1} \frac{E\{|F_{\mathcal{B}}[m]|^2\}}{E\{|F_{\mathcal{B}}[m]|^2\} + \sigma_{\mathcal{B}}[m]^2} X_{\mathcal{B}}[m] g_m. \quad (11.10)$$

The resulting minimum linear Bayes risk is

$$r_l(\pi) = \sum_{m=0}^{N-1} \frac{E\{|F_B[m]|^2\} \sigma_B[m]^2}{E\{|F_B[m]|^2\} + \sigma_B[m]^2}. \quad (11.11)$$

**Proof.** The diagonal values of  $R_F$  and  $R_W$  are  $\langle R_F g_m, g_m \rangle = E\{|F_B[m]|^2\}$  and  $\langle R_W g_m, g_m \rangle = E\{|W_B[m]|^2\} = \sigma_B^2[m]$ . Since  $R_F$  and  $R_W$  are diagonal in  $B$ , the linear operator  $R_F (R_F + R_W)^{-1}$  in (11.6) is also diagonal in  $B$ , with diagonal values equal to  $E\{|F_B[m]|^2\} (E\{|F_B[m]|^2\} + \sigma_B[m]^2)^{-1}$ . So (11.6) proves that the Wiener estimator is

$$\tilde{F}_B = R_F (R_F + R_W)^{-1} X = \sum_{m=0}^{N-1} \frac{E\{|F_B[m]|^2\}}{E\{|F_B[m]|^2\} + \sigma_B[m]^2} X_B[m] g_m, \quad (11.12)$$

which proves (11.10).

The resulting risk is

$$E\{\|F - \tilde{F}\|^2\} = \sum_{m=0}^{N-1} E\left\{|F_B[m] - \tilde{F}_B[m]|^2\right\}. \quad (11.13)$$

Inserting (11.12) in (11.13) with  $X_B[m] = F_B[m] + W_B[m]$ , where  $F_B[m]$  and  $W_B[m]$  are independent, yields (11.11). ■

This corollary proves that the Wiener estimator is implemented with a diagonal attenuation of each data coefficient  $X_B[m]$  by a factor that depends on the signal-to-noise ratio  $E\{|F_B[m]|^2\}/\sigma_B[m]^2$  in the direction of  $g_m$ . The smaller the signal-to-noise ratio (SNR), the more attenuation is required. If  $F$  and  $W$  are Gaussian processes, then the Wiener estimator is optimal among linear and nonlinear estimators of  $F$ .

If  $W$  is a white noise, then its coefficients are uncorrelated with the same variance:

$$E\{W[n] W[k]\} = \sigma^2 \delta[n - k].$$

Its covariance matrix is therefore  $R_W = \sigma^2 \text{Id}$ . It is diagonal in all orthonormal bases and, in particular, in a Karhunen-Loève basis of  $F$ . Thus, Theorem 11.1 can be applied with  $\sigma_B[m] = \sigma$  for  $0 \leq m < N$ .

### Frequency Filtering

Suppose that  $F$  and  $W$  are zero-mean, wide-sense circular stationary random vectors. The properties of such processes are reviewed in Section A.6 of the Appendix. Their covariance satisfies

$$E\{F[n] F[k]\} = R_F[n - k], \quad E\{W[n] W[k]\} = R_W[n - k],$$

where  $R_F[n]$  and  $R_W[n]$  are  $N$  periodic. These matrices correspond to circular convolution operators and are therefore diagonal in the discrete Fourier basis

$$\left\{ g_m[n] = \frac{1}{\sqrt{N}} \exp\left(\frac{i2m\pi n}{N}\right) \right\}_{0 \leq m < N}.$$

The eigenvalues  $E\{|F_B[m]|^2\}$  and  $\sigma_B[m]^2$  are the discrete Fourier transforms of  $R_F[n]$  and  $R_W[n]$ , also called *power spectra*:

$$E\{|F_B[m]|^2\} = \sum_{n=0}^{N-1} R_F[n] \exp\left(-\frac{i2m\pi n}{N}\right) = \hat{R}_F[m],$$

$$\sigma_B[m]^2 = \sum_{n=0}^{N-1} R_W[n] \exp\left(-\frac{i2m\pi n}{N}\right) = \hat{R}_W[m].$$

The Wiener estimator (11.10) is then a diagonal operator in the discrete Fourier basis, computed with the frequency filter

$$\hat{h}[m] = \frac{\hat{R}_F[m]}{\hat{R}_F[m] + \hat{R}_W[m]}. \quad (11.14)$$

It is therefore a circular convolution:

$$\tilde{F}[n] = DX = X \star h[n].$$

The resulting risk is calculated with (11.11):

$$r_I(\pi) = E\{\|F - \tilde{F}\|^2\} = \sum_{m=0}^{N-1} \frac{\hat{R}_F[m] \hat{R}_W[m]}{\hat{R}_F[m] + \hat{R}_W[m]}. \quad (11.15)$$

The numerical value of the risk is often specified by the signal-to-noise ratio, which is measured in decibels:

$$\text{SNR}_{\text{db}} = 10 \log_{10} \left( \frac{E\{\|F\|^2\}}{E\{\|F - \tilde{F}\|^2\}} \right). \quad (11.16)$$

### EXAMPLE 11.1

Figure 11.1(a) shows a realization of a Gaussian process  $F$  obtained as a convolution of a Gaussian white noise  $B$  of variance  $\beta^2$  with a low-pass filter  $g$ :

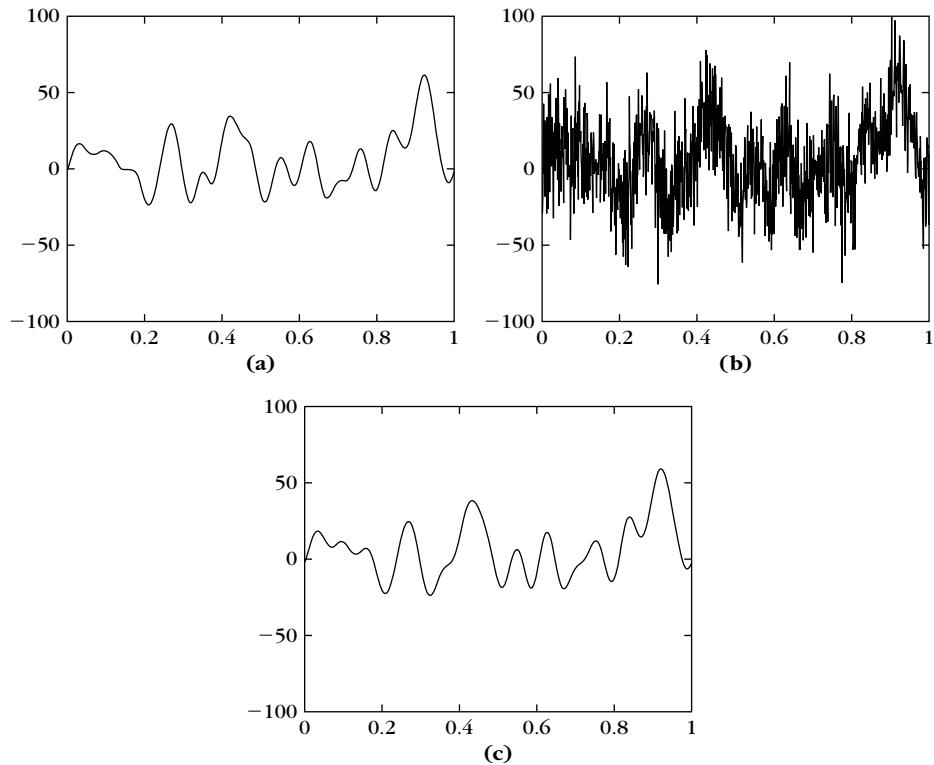
$$F[n] = B \star g[n],$$

with

$$g[n] = C \cos^2\left(\frac{\pi n}{2K}\right) \mathbf{1}_{[-K, K]}[n].$$

Theorem A.7 proves that

$$\hat{R}_F[m] = \hat{R}_B[m] |\hat{g}[m]|^2 = \beta^2 |\hat{g}[m]|^2.$$

**FIGURE 11.1**

**(a)** Realization of a Gaussian process  $F$ . **(b)** Noisy signal obtained by adding a Gaussian white noise ( $\text{SNR} = -0.48 \text{ db}$ ). **(c)** Wiener estimation  $\tilde{F}$  ( $\text{SNR} = 15.2 \text{ db}$ ).

The noisy signal  $X$  shown in Figure 11.1(b) is contaminated by a Gaussian white noise  $W$  of variance  $\sigma^2$ , so  $\hat{R}_W[m] = \sigma^2$ . The Wiener estimation  $\tilde{F}$  is calculated with the frequency filter (11.14)

$$\hat{h}[m] = \frac{\beta^2 |\hat{g}[m]|^2}{\beta^2 |\hat{g}[m]|^2 + \sigma^2}.$$

This linear estimator is also an optimal nonlinear estimator because  $F$  and  $W$  are jointly Gaussian random vectors.

### Piecewise Regular

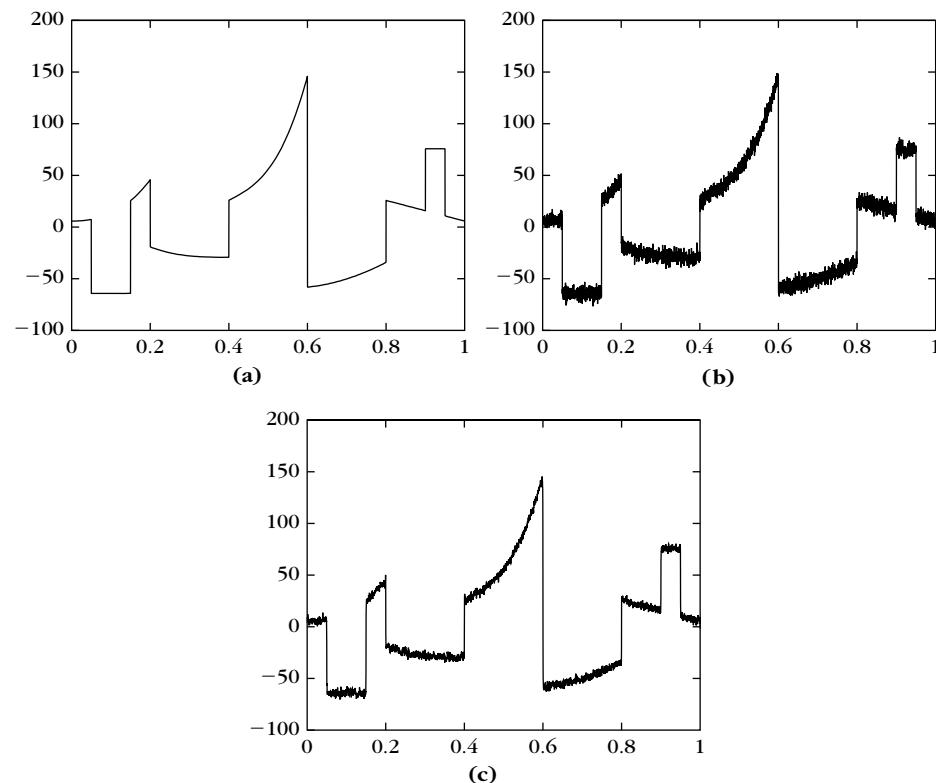
The limitations of linear estimators appear clearly for processes with realizations that are piecewise regular signals. A simple example is a random-shift process  $F$  constructed by translating randomly a piecewise regular signal  $f[n]$  of zero-mean,  $\sum_{n=0}^{N-1} f[n] = 0$ :

$$F[n] = f[(n - Q) \bmod N]. \quad (11.17)$$

The translation variable  $Q$  is an integer random variable with a probability distribution on  $[0, N - 1]$ . It is proved in (9.28) that  $F$  is a circular wide-sense stationary process with a power spectrum calculated in (9.29):

$$\hat{R}_F[m] = \frac{1}{N} |\hat{f}[m]|^2. \quad (11.18)$$

Figure 11.2 shows an example of a piecewise polynomial signal  $f$  of degree  $d = 3$  contaminated by a Gaussian white noise  $W$  of variance  $\sigma^2$ . Assuming that we know  $|\hat{f}[m]|^2$ , the Wiener estimator  $\tilde{f}$  is calculated as a circular convolution with the filter in (11.14). This Wiener filter is a low-pass filter that averages the noisy data to attenuate the noise in regions where the realization of  $F$  is regular, but this averaging is limited to avoid degrading the discontinuities too much. As a result, some noise is left in the smooth regions and the discontinuities are averaged a little. The risk



**FIGURE 11.2**

**(a)** Piecewise polynomial of degree 3. **(b)** Noisy signal degraded by Gaussian white noise (SNR = 21.9 dB). **(c)** Wiener estimation (SNR = 25.9 dB).

calculated in (11.15) is normalized by the total noise energy  $E\{\|W\|^2\} = N\sigma^2$ :

$$\frac{r_l(\pi)}{N\sigma^2} = \sum_{m=0}^{N-1} \frac{N^{-1} |\hat{f}[m]|^2}{|\hat{f}[m]|^2 + N\sigma^2}. \quad (11.19)$$

Suppose that  $f$  has discontinuities of amplitude on the order of  $C \geq \sigma$  and that the noise energy is not negligible:  $N\sigma^2 \geq C^2$ . Using the fact that  $|\hat{f}[m]|$  decays typically like  $CN m^{-1}$ , a direct calculation of the risk (11.19) gives

$$\frac{r_l(\pi)}{N\sigma^2} \sim \frac{C}{\sigma N^{1/2}}. \quad (11.20)$$

The equivalence  $\sim$  means that upper and lower bounds of the left side are obtained by multiplying the right side by two constants  $A, B > 0$  that are independent of  $C$ ,  $\sigma$ , and  $N$ .

The estimation of  $F$  can be improved by nonlinear operators, which average the data  $X$  over large domains where  $F$  is regular, but do not make any averaging where  $F$  is discontinuous. Many estimators have been studied [262, 380], to recover the position of the discontinuities of  $f$  in order to adapt the data averaging. These algorithms have long remained *ad hoc* implementations of intuitively appealing ideas. Wavelet thresholding estimators perform such an adaptive smoothing and Section 11.5.3 proves that the normalized risk decays like  $N^{-1}(\log N)^2$  as opposed to  $N^{-1/2}$  in (11.20).

### 11.1.2 Minimax Estimation

Although we may have some prior information, it is rare that we know the probability distribution of complex signals. Presently, there exists no stochastic model that takes into account the diversity of natural images. However, many images, such as the one in Figure 2.2, have some form of piecewise regularity, with a bounded total variation. Models are often defined over the original analog signal  $\tilde{f}$  that is measured with sensors having a response  $\phi_n$ . The resulting discrete signal  $f[n] = \langle \tilde{f}, \phi_n \rangle$  then belongs to a particular set  $\Theta$  in  $\mathbb{C}^N$  derived from the analog model. This prior information defines a signal set  $\Theta$ , but it does not specify the probability distribution of signals in  $\Theta$ . The more prior information, the smaller the set  $\Theta$ .

Knowing that  $f \in \Theta$ , we want to estimate this signal from the noisy data

$$X[n] = f[n] + W[n].$$

The risk of an estimation  $\tilde{F} = DX$  is  $r(D, f) = E\{\|DX - f\|^2\}$ . The expected risk over  $\Theta$  cannot be computed because the probability distribution of signals in  $\Theta$  is unknown. To control the risk for any  $f \in \Theta$ , we thus try to minimize the maximum risk:

$$r(D, \Theta) = \sup_{f \in \Theta} E\{\|DX - f\|^2\}.$$

The *minimax risk* is the lower bound computed over all linear and nonlinear operators  $D$ :

$$r_n(\Theta) = \inf_{D \in \mathcal{O}_n} r(D, \Theta).$$

In practice, we must find a decision operator  $D$  that is simple to implement and such that  $r(D, \Theta)$  is close to the minimax risk  $r_n(\Theta)$ .

As a first step, as for Wiener estimators in the Bayes framework, the problem is simplified by restricting  $D$  to be a linear operator. The *linear minimax risk* over  $\Theta$  is the lower bound:

$$r_l(\Theta) = \inf_{D \in \mathcal{O}_l} r(D, \Theta).$$

This strategy is efficient only if  $r_l(\Theta)$  is of the same order as  $r_n(\Theta)$ .

### Bayes Priors

A Bayes estimator supposes that we know the prior probability distribution  $\pi$  of signals in  $\Theta$ . If available, this supplement of information can only improve the signal estimation. The central result of game and decision theory shows that minimax estimations are Bayes estimations for a “least-favorable” prior distribution.

Let  $F$  be the signal random vector with a probability distribution that is given by the prior  $\pi$ . For a decision operator  $D$ , the expected risk is  $r(D, \pi) = E_\pi\{r(D, F)\}$ . The minimum Bayes risks for linear and nonlinear operators are defined by:

$$r_l(\pi) = \inf_{D \in \mathcal{O}_l} r(D, \pi) \quad \text{and} \quad r_n(\pi) = \inf_{D \in \mathcal{O}_n} r(D, \pi).$$

Let  $\Theta^*$  be the set of all probability distributions of random vectors with realizations in  $\Theta$ . The minimax theorem (11.4) relates a minimax risk and the maximum Bayes risk calculated for priors in  $\Theta^*$ .

**Theorem 11.4: Minimax.** For any subset  $\Theta$  of  $\mathbb{C}^N$ ,

$$r_l(\Theta) = \sup_{\pi \in \Theta^*} r_l(\pi) \quad \text{and} \quad r_n(\Theta) = \sup_{\pi \in \Theta^*} r_n(\pi). \quad (11.21)$$

**Proof.** For any  $\pi \in \Theta^*$ ,

$$r(D, \pi) \leq r(D, \Theta) \quad (11.22)$$

because  $r(D, \pi)$  is an average risk over realizations of  $F$  that are in  $\Theta$ , whereas  $r(D, \Theta)$  is the maximum risk over  $\Theta$ . Let  $\mathcal{O}$  be a convex set of operators (either  $\mathcal{O}_l$  or  $\mathcal{O}_n$ ). The inequality (11.22) implies that

$$\sup_{\pi \in \Theta^*} r(\pi) = \sup_{\pi \in \Theta^*} \inf_{D \in \mathcal{O}} r(D, \pi) \leq \inf_{D \in \mathcal{O}} r(D, \Theta) = r(\Theta). \quad (11.23)$$

The main difficulty is to prove the reverse inequality:  $r(\Theta) \leq \sup_{\pi \in \Theta^*} r(\pi)$ . When  $\Theta$  is a finite set, the proof gives a geometrical interpretation of the minimum Bayes risk and the minimax risk. The extension to an infinite set  $\Theta$  is sketched.

Suppose that  $\Theta = \{f_i\}_{1 \leq i \leq p}$  is a finite set of signals. We define a risk set:

$$R = \{(y_1, \dots, y_p) \in \mathbb{C}^p : \exists D \in \mathcal{O} \quad \text{with} \quad y_i = r(D, f_i) \text{ for } 1 \leq i \leq p\}.$$

This set is convex in  $\mathbb{C}^p$  because  $\mathcal{O}$  is convex. We begin by giving geometrical interpretations to the Bayes risk and the minimax risk.

A prior  $\pi \in \Theta^*$  is a vector of discrete probabilities  $(\pi_1, \dots, \pi_p)$  and

$$r(\pi, D) = \sum_{i=1}^p \pi_i r(D, f_i). \quad (11.24)$$

The equation  $\sum_{i=1}^p \pi_i y_i = b$  defines a hyperplane  $P_b$  in  $\mathbb{C}^p$ . Computing  $r(\pi) = \inf_{D \in \mathcal{O}} r(D, \pi)$  is equivalent to finding the infimum  $b_0 = r(\pi)$  of all  $b$  for which  $P_b$  intersects  $R$ . The plane  $P_{b_0}$  is tangent to  $R$  as shown in Figure 11.3.

The minimax risk  $r(\Theta)$  has a different geometrical interpretation. Let  $Q_c = \{(y_1, \dots, y_p) \in \mathbb{C}^p : y_i \leq c\}$ . One can verify that  $r(\Theta) = \inf_{D \in \mathcal{O}} \sup_{f_i \in \Theta} r(D, f_i)$  is the infimum  $c_0 = r(\Theta)$  of all  $c$  such that  $Q_c$  intersects  $R$ .

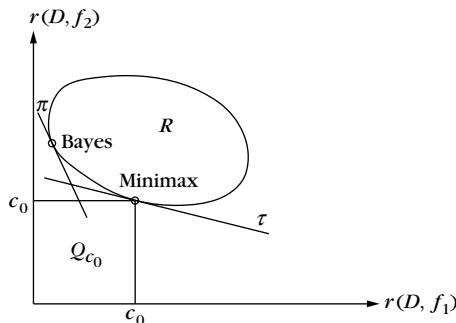
To prove that  $r(\Theta) \leq \sup_{\pi \in \Theta^*} r(\pi)$ , we look for a prior distribution  $\tau \in \Theta^*$  such that  $r(\tau) = r(\Theta)$ . Let  $\tilde{Q}_{c_0}$  be the interior of  $Q_{c_0}$ . Since  $\tilde{Q}_{c_0} \cap R = \emptyset$  and both  $\tilde{Q}_{c_0}$  and  $R$  are convex sets, the hyperplane separation theorem says that there exists a hyperplane of equation

$$\sum_{i=1}^p \tau_i y_i = \tau \cdot y = b, \quad (11.25)$$

with  $\tau \cdot y \leq b$  for  $y \in \tilde{Q}_{c_0}$  and  $\tau \cdot y \geq b$  for  $y \in R$ . Each  $\tau_i \geq 0$ , for if  $\tau_j < 0$ , then for  $y \in \tilde{Q}_{c_0}$ , we obtain a contradiction by taking  $y_j$  to  $-\infty$  with the other coordinates being fixed. Indeed,  $\tau \cdot y$  goes to  $+\infty$  and since  $y$  remains in  $\tilde{Q}_{c_0}$ , it contradicts the fact that  $\tau \cdot y \leq b$ . We can normalize  $\sum_{i=1}^p \tau_i = 1$  by dividing each side of (11.25) by  $\sum_{i=1}^p \tau_i > 0$ . So  $\tau$  corresponds to a probability distribution. By letting  $y \in \tilde{Q}_{c_0}$  converge to the corner point  $(c_0, \dots, c_0)$ , since  $y \cdot \tau \leq b$ , we derive that  $c_0 \leq b$ . Moreover, since  $\tau \cdot y \geq b$  for all  $y \in R$ ,

$$r(\tau) = \inf_{D \in \mathcal{O}} \sum_{i=1}^p \tau_i r(D, f_i) \geq c \geq c_0 = r(\Theta).$$

So,  $r(\Theta) \leq \sup_{\pi \in \Theta^*} r(\pi)$ , which, together with (11.23), proves that  $r(\Theta) = \sup_{\pi \in \Theta^*} r(\pi)$ .



**FIGURE 11.3**

At the Bayes point, a hyperplane defined by the prior  $\pi$  is tangent to the risk set  $R$ . The least-favorable prior  $\tau$  defines a hyperplane that is tangential to  $R$  at the minimax point.

The extension of this result to an infinite set of signals  $\Theta$  is done with a compactness argument. When  $\mathcal{O} = \mathcal{O}_l$  or  $\mathcal{O} = \mathcal{O}_n$ , for any prior  $\pi \in \Theta^*$ , we know from Theorems 11.1 and 11.2 that  $\inf_{D \in \mathcal{O}} r(D, \pi)$  is reached by some Bayes decision operator  $D \in \mathcal{O}$ . One can verify that there exists a subset of operators  $\mathcal{C}$  that includes the Bayes operator for any prior  $\pi \in \Theta^*$ , and such that  $\mathcal{C}$  is compact for an appropriate topology. When  $\mathcal{O} = \mathcal{O}_l$ , one can choose  $\mathcal{C}$  to be the set of linear operators of norm smaller than 1, which is compact because it belongs to a finite-dimensional space of linear operators. Moreover, the risk  $r(f, D)$  can be shown to be continuous in this topology with respect to  $D \in \mathcal{C}$ .

Let  $c < r(\Theta)$ . For any  $f \in \Theta$ , we consider the set of operators  $\mathcal{S}_f = \{D \in \mathcal{C} : r(D, f) > c\}$ . The continuity of  $r$  implies that  $\mathcal{S}_f$  is an open set. For each  $D \in \mathcal{C}$  there exists  $f \in \Theta$  such that  $D \in \mathcal{S}_f$ , so  $\mathcal{C} = \cup_{f \in \Theta} \mathcal{S}_f$ . Since  $\mathcal{C}$  is compact, there exists a finite covering  $\mathcal{C} = \cup_{1 \leq i \leq p} \mathcal{S}_{f_i}$ . The minimax risk over  $\Theta_c = \{f_i\}_{1 \leq i \leq p}$  satisfies

$$r(\Theta_c) = \inf_{D \in \mathcal{O}} \sup_{1 \leq i \leq p} r(D, f_i) \geq c.$$

Since  $\Theta_c$  is a finite set, we proved that there exists  $\tau_c \in \Theta_c^* \subset \Theta^*$  such that  $r(\tau_c) = r(\Theta_c)$ . But  $r(\Theta_c) \geq c$ , so letting  $c$  go to  $r(\Theta)$  implies that  $\sup_{\pi \in \Theta^*} r(\pi) \geq r(\Theta)$ . Together with (11.23) this shows that  $\inf_{\tau \in \Theta^*} r(\tau) = r(\Theta)$ . ■

A distribution  $\tau \in \Theta^*$  such that  $r(\tau) = \inf_{\pi \in \Theta^*} r(\pi)$  is called a *least-favorable* prior distribution. The minimax theorem proves that the minimax risk is the minimum Bayes risk for a least-favorable prior.

In signal processing, minimax calculations are often hidden behind apparently orthodox Bayes estimations. Let us consider an example involving images. It has been observed that histograms of the wavelet coefficients of “natural” images can be modeled with generalized Gaussian distributions [361, 440]. This means that natural images belong to a certain set  $\Theta$ , but it does not specify a prior distribution over this set. To compensate for the lack of knowledge about the dependency of wavelet coefficients spatially and across scales, one may be tempted to create a “simple probabilistic model” where all wavelet coefficients are considered to be independent. This model is clearly simplistic since images have geometrical structures that create strong dependencies both spatially and across scales (see Figure 7.24). However, calculating a Bayes estimator with this inaccurate prior model may give valuable results when estimating images. Why? Because this “simple” prior is often close to a least-favorable prior. The resulting estimator and risk are thus good approximations of the minimax optimum. If not chosen carefully, a “simple” prior may yield an optimistic risk evaluation that is not valid for real signals.

On the other hand, the minimax approach may seem very pessimistic since we always consider the maximum risk over  $\Theta$ ; this is sometimes the case. However, when the set  $\Theta$  is large, one can often verify that a “typical” signal of  $\Theta$  has a risk of the order of the maximum risk over  $\Theta$ . A minimax calculation attempts to isolate the class of signals that are the most difficult to estimate, and one can check whether these signals are indeed typically encountered in an application. If this is not the case, then it indicates that the model specified by  $\Theta$  is not well adapted to this application.

## 11.2 DIAGONAL ESTIMATION IN A BASIS

It is generally not possible to compute the optimal Bayes or minimax estimator that minimizes the risk among all possible operators. To manage this complexity, the most classical strategy limits the choice of operators among linear operators. This comes at a cost, because the minimum risk among linear estimators may be well above the minimum risk obtained with nonlinear estimators. Figure 11.2 is an example where the linear Wiener estimation can be considerably improved with a nonlinear averaging. This section studies a particular class of nonlinear estimators that are diagonal in a basis  $\mathcal{B}$ . If the basis  $\mathcal{B}$  defines a sparse signal representation, then such diagonal estimators are nearly optimal among all nonlinear estimators.

Section 11.2.1 computes a lower bound for the risk when estimating an arbitrary signal  $f$  with a diagonal operator. Donoho and Johnstone [221] made a fundamental breakthrough by showing that thresholding estimators have a risk that is close to this lower bound. The general properties of thresholding estimators are introduced in Sections 11.2.2 and 11.2.3.

### 11.2.1 Diagonal Estimation with Oracles

We consider estimators computed with a diagonal operator in an orthonormal basis  $\mathcal{B} = \{g_m\}_{0 \leq m < N}$ . Lower bounds for the risk are computed with “oracles,” which simplify the estimation by providing information about the signal that is normally not available. These lower bounds are closely related to errors when approximating signals from a few vectors selected in  $\mathcal{B}$ .

The noisy data

$$X[n] = f[n] + W[n] \quad \text{for } 0 \leq n < N \quad (11.26)$$

are decomposed in  $\mathcal{B}$ . We write

$$X_{\mathcal{B}}[m] = \langle X, g_m \rangle, \quad f_{\mathcal{B}}[m] = \langle f, g_m \rangle \quad \text{and} \quad W_{\mathcal{B}}[m] = \langle W, g_m \rangle.$$

The inner product of (11.26) with  $g_m$  gives

$$X_{\mathcal{B}}[m] = f_{\mathcal{B}}[m] + W_{\mathcal{B}}[m].$$

We suppose that  $W$  is a zero-mean *white noise* of variance  $\sigma^2$ , which means

$$E\{W[n] W[k]\} = \sigma^2 \delta[n - k].$$

The noise coefficients

$$W_{\mathcal{B}}[m] = \sum_{n=0}^{N-1} W[n] g_m^*[n]$$

also define a white noise of variance  $\sigma^2$ . Indeed,

$$\begin{aligned} E\{W_{\mathcal{B}}[m] W_{\mathcal{B}}[p]\} &= \sum_{n=0}^{N-1} \sum_{k=0}^{N-1} g_m[n] g_p[k] E\{W[n] W[k]\} \\ &= \sigma^2 \langle g_p, g_m \rangle = \sigma^2 \delta[p - m]. \end{aligned}$$

Since the noise remains white in all bases, it does not influence the choice of basis.

A *diagonal operator* independently estimates each  $f_{\mathcal{B}}[m]$  by multiplying  $X_{\mathcal{B}}[m]$  by a factor  $a_m(X_{\mathcal{B}}[m])$ . The resulting estimator is

$$\tilde{F} = D X = \sum_{m=0}^{N-1} a_m(X_{\mathcal{B}}[m]) X_{\mathcal{B}}[m] g_m. \quad (11.27)$$

The operator  $D$  is linear when  $a_m(X_{\mathcal{B}}[m])$  is a constant independent of  $X_{\mathcal{B}}[m]$ .

### Oracle Attenuation

For a given signal  $f$  let us find the constant  $a_m = a_m(X_{\mathcal{B}}[m])$  that minimizes the risk  $r(D, f)$  of the estimator (11.27):

$$r(D, f) = E\left\{\|f - \tilde{F}\|^2\right\} = \sum_{m=0}^{N-1} E\{|f_{\mathcal{B}}[m] - a_m X_{\mathcal{B}}[m]|^2\}. \quad (11.28)$$

We shall see that  $|a_m| \leq 1$ , which means that the diagonal operator  $D$  should attenuate the noisy coefficients.

Since  $X_{\mathcal{B}} = f_{\mathcal{B}} + W_{\mathcal{B}}$  and  $E\{|W_{\mathcal{B}}[m]|^2\} = \sigma^2$ , and since  $a_m$  is a constant that does not depend on the noise, it follows that

$$E\{|f_{\mathcal{B}}[m] - X_{\mathcal{B}}[m] a_m|^2\} = |f_{\mathcal{B}}[m]|^2 (1 - a_m)^2 + \sigma^2 a_m^2. \quad (11.29)$$

This risk is minimum for

$$a_m = \frac{|f_{\mathcal{B}}[m]|^2}{|f_{\mathcal{B}}[m]|^2 + \sigma^2}, \quad (11.30)$$

in which case

$$r_{\inf}(f) = E\{\|f - \tilde{F}\|^2\} = \sum_{m=0}^{N-1} \frac{|f_{\mathcal{B}}[m]|^2 \sigma^2}{|f_{\mathcal{B}}[m]|^2 + \sigma^2}. \quad (11.31)$$

Observe that the attenuation factor  $a_m$  and the resulting risk have the same structure as the Wiener filter (11.10). However, the Wiener filter is a linear operator that depends on the expected SNRs that are constant values. This attenuation factor depends on the unknown original signal-to-noise ratio  $|f_{\mathcal{B}}[m]|^2/\sigma^2$ . Since  $|f_{\mathcal{B}}[m]|$  is unknown, the attenuation factor  $a_m$  cannot be computed. It is considered as an *oracle* information. The resulting oracle risk  $r_{\inf}(f)$  is a lower bound that is normally not reachable. However, Section 11.2.2 shows that one can get close to  $r_{\inf}(f)$  with a simple thresholding.

### Oracle Projection

The analysis of diagonal estimators can be simplified by restricting  $a_m \in \{0, 1\}$ . When  $a_m = 1$ , the estimator  $\tilde{F} = DX$  selects the coefficient  $X_B[m]$ , and it removes it if  $a_m = 0$ . The operator  $D$  is then an orthogonal projector on a selected subset of vectors of the basis  $\mathcal{B}$ .

The nonlinear projector that minimizes the risk (11.29) is defined by

$$a_m = \begin{cases} 1 & \text{if } |f_{\mathcal{B}}[m]| \geq \sigma \\ 0 & \text{if } |f_{\mathcal{B}}[m]| < \sigma. \end{cases} \quad (11.32)$$

The resulting oracle projector is

$$DX = \sum_{m \in \Lambda_{\sigma}} X_B[m] g_m \quad \text{with} \quad \Lambda_{\sigma} = \{0 \leq m < N : |f_{\mathcal{B}}[m]| \geq \sigma\}. \quad (11.33)$$

It is an orthogonal projection on the set  $\{g_m\}_{m \in \Lambda_{\sigma}}$  of basis vectors that best approximate  $f$ . This “oracle” projector cannot either be implemented because  $a_m$  and  $\Lambda_{\sigma}$  depend on  $f_{\mathcal{B}}[m]$  instead of  $X_B[m]$ . The resulting risk is computed with (11.29):

$$r_{\text{pr}}(f) = E\{\|f - \tilde{F}\|^2\} = \sum_{m=0}^{N-1} \min(|f_{\mathcal{B}}[m]|^2, \sigma^2). \quad (11.34)$$

Since for any  $(x, y) \in \mathbb{R}^2$ ,

$$\min(x, y) \geq \frac{xy}{x+y} \geq \frac{1}{2} \min(x, y),$$

the risk of the oracle projector (11.34) is of the same order as the risk of an oracle attenuation (11.31):

$$r_{\text{pr}}(f) \geq r_{\text{inf}}(f) \geq \frac{1}{2} r_{\text{pr}}(f). \quad (11.35)$$

The risk of an oracle projector can also be related to the approximation error of  $f$  in the basis  $\mathcal{B}$ . Let  $M = |\Lambda_{\sigma}|$  be the number of coefficients such that  $|f_{\mathcal{B}}[m]| \geq \sigma$ . The best  $M$ -term approximation of  $f$ , defined in Section 9.2.1, is the orthogonal projection on the  $M$  vectors  $\{g_m\}_{m \in \Lambda_{\sigma}}$  that yield the largest-amplitude coefficients:

$$f_M = \sum_{m \in \Lambda_{\sigma}} f_{\mathcal{B}}[m] g_m.$$

The nonlinear oracle projector risk can be rewritten as

$$r_{\text{pr}}(f) = \sum_{m=0}^{N-1} \min(|f_{\mathcal{B}}[m]|^2, \sigma^2) = \sum_{m \notin \Lambda_{\sigma}} |f_{\mathcal{B}}[m]|^2 + M \sigma^2 \quad (11.36)$$

$$= \varepsilon_n(M, f) + M \sigma^2, \quad (11.37)$$

where

$$\varepsilon_n(M, f) = \|f - f_M\|^2 = \sum_{m \notin \Lambda_\sigma} |f_B[m]|^2$$

is the best  $M$ -term approximation error. It is the bias produced by setting signal coefficients to zero, and  $M\sigma^2$  is the variance of the remaining noise on the  $M$  coefficients that are kept. Theorem 11.5 proves that when  $\sigma$  decreases, the decay of this risk is characterized by the decay of the nonlinear approximation error  $\varepsilon_n(M, f)$  as  $M$  increases.

**Theorem 11.5.** If  $\varepsilon_n(M, f) \leq C^2 M^{1-2s}$  with  $1 \leq C/\sigma \leq N^s$ , then

$$r_{\text{pr}}(f) \leq 3C^{1/s} \sigma^{2-1/s}. \quad (11.38)$$

**Proof.** Observe that

$$r_{\text{pr}}(f) = \min_{0 \leq M \leq N} (\varepsilon_n(M, f) + M\sigma^2). \quad (11.39)$$

Let  $M_0$  be defined by  $2M_0\sigma^2 > C^2 M_0^{1-2s} \geq M_0\sigma^2$ . Since  $1 \leq C/\sigma \leq N^s$ , necessarily  $1 \leq M_0 \leq N$ . Inserting  $\varepsilon_n(M_0, f) \leq C^2 M_0^{1-2s}$  with  $s > 1/2$  and  $M_0 \leq C^{1/s} \sigma^{-1/s}$  yields

$$r_{\text{pr}}(f) \leq \varepsilon_n(M_0, f) + M_0\sigma^2 \leq 3M_0\sigma^2 \leq 3C^{1/s} \sigma^{2-1/s}, \quad (11.40)$$

which proves (11.38). ■

### Linear Projection

Oracle estimators cannot be implemented because  $a_m$  is a constant that depends on the unknown signal  $f$ . Let us consider linear projectors obtained by setting  $a_m$  to be equal to 1 on the first  $M$  vectors and 0 otherwise:

$$\tilde{F} = D_M X = \sum_{m=0}^{M-1} X_B[m] g_m. \quad (11.41)$$

The risk (11.28) becomes

$$r(D_M, f) = \sum_{m=M}^{N-1} |f_B[m]|^2 + M\sigma^2 = \varepsilon_l(M, f) + M\sigma^2, \quad (11.42)$$

where  $\varepsilon_l(M, f)$  is the linear approximation error computed in (9.3). The two terms  $\varepsilon_l(M, f)$  and  $M\sigma^2$  are, respectively, the bias and the variance components of the estimator. To minimize  $r(D_M, f)$ , the parameter  $M$  is adjusted so that the bias is of the same order as the variance. When the noise variance  $\sigma^2$  decreases, Theorem 11.6 proves that resulting risk depends on the decay of  $\varepsilon_l(M, f)$  as  $M$  increases.

**Theorem 11.6.** If  $\varepsilon_l(M, f) \leq C^2 M^{1-2s}$  with  $1 \leq C/\sigma \leq N^s$ , then

$$r(D_{M_0}, f) \leq 3C^{1/s} \sigma^{2-1/s} \quad \text{for } (C/(2\sigma))^{1/s} < M_0 \leq (C/\sigma)^{1/s}. \quad (11.43)$$

**Proof.** As in (11.40), we verify that

$$r(D_{M_0}, f) = \varepsilon_l(M_0, f) + M_0 \sigma^2 \leq 3C^{1/s} \sigma^{2-1/s},$$

for  $(C/(2\sigma))^{1/s} < M_0 \leq (C/\sigma)^{1/s}$ , which proves (11.43). ■

Theorems 11.5 and 11.6 prove that the performances of oracle projection estimators and optimized linear projectors depend, respectively, on the precision of nonlinear and linear approximations in the basis  $\mathcal{B}$ . Having an approximation error that decreases quickly means that one can then construct a sparse and precise signal representation with only a few vectors in  $\mathcal{B}$ . Section 9.2 shows that nonlinear approximations can be much more precise, in which case the risk of a nonlinear oracle projection is much smaller than the risk of a linear projection. The next section shows that thresholding estimators are nonlinear projection estimators that have a risk close to the oracle projection risk.

### 11.2.2 Thresholding Estimation

In a basis  $\mathcal{B} = \{g_m\}_{0 \leq m < N}$ , a diagonal estimator of  $f$  from  $X = f + W$  can be written as

$$\tilde{F} = DX = \sum_{m=0}^{N-1} a_m(X_{\mathcal{B}}[m]) X_{\mathcal{B}}[m] g_m. \quad (11.44)$$

We suppose that  $W$  is a Gaussian white noise of variance  $\sigma^2$ . When  $a_m$  are thresholding functions, the risk of this estimator is shown to be close to the lower bounds obtained with oracle estimators.

#### Hard Thresholding

A hard-thresholding estimator is implemented with

$$a_m(x) = \begin{cases} 1 & \text{if } |x| \geq T \\ 0 & \text{if } |x| < T, \end{cases} \quad (11.45)$$

and can thus be rewritten as

$$\tilde{F} = DX = \sum_{m \in \tilde{\Lambda}_T} X_{\mathcal{B}}[m] g_m \quad \text{with} \quad \tilde{\Lambda}_T = \{0 \leq m < N : |X_{\mathcal{B}}[m]| \geq T\}. \quad (11.46)$$

It is an orthogonal projection of  $X$  on the set of basis vectors  $\{g_m\}_{m \in \tilde{\Lambda}_T}$ . This estimator can also be rewritten with a hard-thresholding function

$$\tilde{F} = \sum_{m=0}^{N-1} \rho_T(X_{\mathcal{B}}[m]) g_m \quad \text{with} \quad \rho_T(x) = \begin{cases} x & \text{if } |x| > T \\ 0 & \text{if } |x| \leq T. \end{cases} \quad (11.47)$$

The risk of this thresholding is

$$r_{\text{th}}(f) = r(D, f) = \sum_{m=0}^{N-1} E\{|f_{\mathcal{B}}[m] - \rho_T(X_{\mathcal{B}}[m])|^2\},$$

with  $X_{\mathcal{B}}[m] = f_{\mathcal{B}}[m] + W_{\mathcal{B}}[m]$ , and thus

$$|f_{\mathcal{B}}[m] - \rho_T(X_{\mathcal{B}}[m])|^2 = \begin{cases} |W_{\mathcal{B}}[m]|^2 & \text{if } |X_{\mathcal{B}}[m]| > T \\ |f_{\mathcal{B}}[m]|^2 & \text{if } |X_{\mathcal{B}}[m]| \leq T. \end{cases}$$

Since a hard thresholding is a nonlinear projector in the basis  $\mathcal{B}$ , the thresholding risk is larger than the risk (11.34) of an oracle projector:

$$r_{\text{th}}(f) \geq r_{\text{pr}}(f) = \sum_{m=0}^{N-1} \min(|f_{\mathcal{B}}[m]|^2, \sigma^2).$$

### Soft Thresholding

An oracle attenuation (11.30) yields a risk  $r_{\text{inf}}(f)$  that is smaller than the risk  $r_{\text{pr}}(f)$  of an oracle projection, by slightly decreasing the amplitude of all coefficients in order to reduce the added noise. A soft attenuation, although nonoptimal, is implemented by

$$0 \leq a_m(x) = \max \left( 1 - \frac{T}{|x|}, 0 \right) \leq 1. \quad (11.48)$$

The resulting diagonal estimator  $\tilde{F}$  in (11.44) can be written as in (11.47) with a soft-thresholding function, which decreases the amplitude of all noisy coefficients by  $T$ :

$$\rho_T(x) = \begin{cases} x - T & \text{if } x \geq T \\ x + T & \text{if } x \leq -T \\ 0 & \text{if } |x| \leq T. \end{cases} \quad (11.49)$$

The threshold  $T$  is generally chosen so that there is a high probability that it is just above the maximum level of the noise coefficients  $|W_{\mathcal{B}}[m]|$ . Reducing the amplitude of all noisy coefficients by  $T$  thus ensures that the amplitude of an estimated coefficient is smaller than the amplitude of the original one:

$$|\rho_T(X_{\mathcal{B}}[m])| \leq |f_{\mathcal{B}}[m]|. \quad (11.50)$$

In a wavelet basis where large-amplitude coefficients are created by sharp signal variations, this estimation restores a signal that is at least as regular as the original signal  $f$ , without adding sharp transitions due to the noise.

### Thresholding Risk

Theorem 11.7 [221] proves that for an appropriate choice of  $T$ , the risk of a thresholding is close to the risk of an oracle projector  $r_{\text{pr}}(f) = \sum_{m=0}^{N-1} \min(|f_{\mathcal{B}}[m]|^2, \sigma^2)$ . We denote by  $\mathcal{O}_d$  the set of all linear or nonlinear operators that are diagonal in  $\mathcal{B}$ .

**Theorem 11.7:** *Donoho, Johnstone.* Let  $T = \sigma \sqrt{2 \log_e N}$ . The risk  $r_{\text{th}}(f)$  of a hard- or soft-thresholding estimator satisfies for all  $N \geq 4$ ,

$$r_{\text{th}}(f) \leq (2 \log_e N + 1) (\sigma^2 + r_{\text{pr}}(f)). \quad (11.51)$$

The factor  $2 \log_e N$  is optimal among diagonal estimators in  $\mathcal{B}$ :

$$\lim_{N \rightarrow +\infty} \inf_{D \in \mathcal{O}_d} \sup_{f \in \mathbb{C}^N} \frac{E\{\|f - \tilde{F}\|^2\}}{\sigma^2 + r_{\text{pr}}(f)} \frac{1}{2 \log_e N} = 1. \quad (11.52)$$

**Proof.** The proof of (11.51) is given for a soft thresholding. For a hard thresholding, the proof is similar although slightly more complicated. For a threshold  $\lambda$ , a soft thresholding is computed with

$$\rho_\lambda(x) = (x - \lambda \operatorname{sign}(x)) \mathbf{1}_{|x|>\lambda}.$$

Let  $X$  be a Gaussian random variable of mean  $\mu$  and variance 1. The risk when estimating  $\mu$  with a soft thresholding of  $X$  is

$$r(\lambda, \mu) = E\{|\rho_\lambda(X) - \mu|^2\} = E\{|(X - \lambda \operatorname{sign}(X)) \mathbf{1}_{|X|>\lambda} - \mu|^2\}. \quad (11.53)$$

If  $X$  has a variance  $\sigma^2$  and a mean  $\mu$ , then by considering  $\tilde{X} = X/\sigma$ , we verify that

$$E\{|\rho_\lambda(X) - \mu|^2\} = \sigma^2 r\left(\frac{\lambda}{\sigma}, \frac{\mu}{\sigma}\right).$$

Since  $f_{\mathcal{B}}[m]$  is a constant,  $X_{\mathcal{B}}[m] = f_{\mathcal{B}}[m] + W_{\mathcal{B}}[m]$  is a Gaussian random variable of mean  $f_{\mathcal{B}}[m]$  and variance  $\sigma^2$ . The risk of the soft-thresholding estimator  $\tilde{F}$  with a threshold  $T$  is thus

$$r_{\text{th}}(f) = \sigma^2 \sum_{m=0}^{N-1} r\left(\frac{T}{\sigma}, \frac{f_{\mathcal{B}}[m]}{\sigma}\right). \quad (11.54)$$

An upper bound of this risk is calculated with Lemma 11.1.

**Lemma 11.1.** If  $\mu \geq 0$ , then

$$r(\lambda, \mu) \leq r(\lambda, 0) + \min(1 + \lambda^2, \mu^2). \quad (11.55)$$

To prove (11.55), we first verify that if  $\mu \geq 0$ , then

$$0 \leq \frac{\partial r(\lambda, \mu)}{\partial \mu} = 2\mu \int_{-\lambda+\mu}^{\lambda-\mu} \phi(x) dx \leq 2\mu, \quad (11.56)$$

where  $\phi(x)$  is the normalized Gaussian probability density

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right).$$

Indeed (11.53) shows that

$$r(\lambda, \mu) = \mu^2 \int_{-\lambda+\mu}^{\lambda-\mu} \phi(x) dx + \int_{\lambda-\mu}^{+\infty} (x - \lambda)^2 \phi(x) dx + \int_{-\infty}^{-\lambda+\mu} (x + \lambda)^2 \phi(x) dx. \quad (11.57)$$

We obtain (11.56) by differentiating with respect to  $\mu$ .

Since  $\int_{-\infty}^{+\infty} \phi(x) dx = \int_{-\infty}^{+\infty} x^2 \phi(x) dx = 1$  and  $\frac{\partial r(\lambda, \mu)}{\partial \mu} \geq 0$ , necessarily

$$r(\lambda, \mu) \leq \lim_{\mu \rightarrow +\infty} r(\lambda, \mu) = 1 + \lambda^2. \quad (11.58)$$

Moreover, since  $\frac{\partial r(\lambda, s)}{\partial s} \leq 2s$ ,

$$r(\lambda, \mu) - r(\lambda, 0) = \int_0^\mu \frac{\partial r(\lambda, s)}{\partial s} ds \leq \mu^2. \quad (11.59)$$

The inequality (11.55) of the lemma is finally derived from (11.58) and (11.59):

$$r(\lambda, \mu) \leq \min(r(\lambda, 0) + \mu^2, 1 + \lambda^2) \leq r(\lambda, 0) + \min(1 + \lambda^2, \mu^2).$$

By inserting the inequality (11.55) of the lemma in (11.54), we get

$$r_{\text{th}}(f) \leq N\sigma^2 r\left(\frac{T}{\sigma}, 0\right) + \sigma^2 \sum_{m=0}^{N-1} \min\left(\frac{T^2 + \sigma^2}{\sigma^2}, \frac{|f_B[m]|^2}{\sigma^2}\right). \quad (11.60)$$

The expression (11.57) shows that  $r(\lambda, 0) = 2 \int_0^{+\infty} x^2 \phi(x + \lambda) dx$ . For  $T = \sigma \sqrt{2 \log_e N}$  and  $N \geq 4$ , one can verify that

$$N r\left(\frac{T}{\sigma}, 0\right) \leq 2 \log_e N + 1. \quad (11.61)$$

Moreover,

$$\begin{aligned} \sigma^2 \min\left(\frac{T^2 + \sigma^2}{\sigma^2}, \frac{|f_B[m]|^2}{\sigma^2}\right) &= \min(2\sigma^2 \log_e N + \sigma^2, |f_B[m]|^2) \\ &\leq (2 \log_e N + 1) \min(\sigma^2, |f_B[m]|^2). \end{aligned} \quad (11.62)$$

Inserting (11.61) and (11.62) in (11.60) proves (11.51).

Since the soft- and hard-thresholding estimators are particular instances of diagonal estimators, the inequality (11.51) implies that

$$\lim_{N \rightarrow +\infty} \inf_{D \in \mathcal{O}_d} \sup_{f \in \mathbb{C}^N} \frac{E\{\|f - \tilde{F}\|^2\}}{\sigma^2 + r_{\text{pr}}(f)} \frac{1}{2 \log_e N} \leq 1. \quad (11.63)$$

To prove that the limit is equal to 1, for  $N$  fixed, we compute a lower bound by replacing the supremum over all signals  $f$  by an expected value over the distribution of a particular signal process  $F$ . The coefficients  $F_B[m]$  are chosen to define a very sparse sequence. They are independent random variables having a high probability  $1 - \alpha_N$  to be equal to 0 and a low probability  $\alpha_N$  to be equal to a value  $\mu_N$  that is on the order of  $\sigma \sqrt{2 \log_e N}$ , but smaller. By adjusting  $\mu_N$  and  $\alpha_N$ , Donoho and Johnstone [221] prove that the Bayes estimator  $\tilde{F}$  of  $F$  tends to zero as  $N$  increases and they derive a lower bound of the left side of (11.63) that tends to 1. ■

The upper bound (11.51) proves that the risk  $r_{\text{th}}(f)$  of a thresholding estimator is at most  $2 \log_e N$  times larger than the risk  $r_{\text{pr}}(f)$  of an oracle projector. Moreover, (11.52) proves that the  $2 \log_e N$  factor cannot be improved by any other diagonal estimator. For  $r_{\text{pr}}(f)$  to be small, (11.37) shows that  $f$  must be well approximated

by a few vectors in  $\mathcal{B}$ . One can verify [221] that the theorem remains valid if  $r_{\text{pr}}(f)$  is replaced by the risk  $r_{\text{inf}}(f)$  of an oracle attenuation, which is smaller.

### Choice of Threshold

The threshold  $T$  must be chosen just above the maximum level of the noise. Indeed, if  $f = 0$  and thus  $X_{\mathcal{B}} = W_{\mathcal{B}}$ , then to ensure that  $\tilde{F} \approx 0$ , the noise coefficients  $|W_{\mathcal{B}}[m]|$  must have a high probability of being below  $T$ . However, if  $f \neq 0$ , then  $T$  must not be too large, so that we do not set to zero too many coefficients such that  $|f_{\mathcal{B}}[m]| \geq \sigma$ . Since  $W_{\mathcal{B}}$  is a vector of  $N$  independent Gaussian random variables of variance  $\sigma^2$ , one can prove [7] that the maximum amplitude of the noise has a very high probability of being just below  $T = \sigma\sqrt{2\log_e N}$ :

$$\lim_{N \rightarrow +\infty} \text{pr} \left( T - \frac{\sigma \log_e \log_e N}{\log_e N} \leq \max_{0 \leq m \leq N} |W_{\mathcal{B}}[m]| \leq T \right) = 1. \quad (11.64)$$

This explains why the theorem chooses this value. That the threshold  $T$  increases with  $N$  may seem counterintuitive. This is due to the tail of the Gaussian distribution, which creates larger-and-larger-amplitude noise coefficients when the sample size increases. The threshold  $T = \sigma\sqrt{2\log_e N}$  is not optimal and, in general, a lower threshold reduces the risk.

A soft thresholding computed for  $T = \sigma\sqrt{2\log_e N}$  often produces a risk that is larger than with a hard thresholding. A soft thresholding reduces to nearly 0 the amplitude of coefficients just above  $T$  or just below  $-T$ , whereas a hard thresholding leaves them as is. To obtain nearly the same risk for a hard thresholding and a soft thresholding, it is often necessary to reduce by two the threshold of the soft thresholding. Section 11.2.3 explains how to adapt the threshold  $T$  to the data  $X$ .

### Upper-Bound Interpretation

Despite the technicality of the proof, the factor  $2\log_e N$  of the upper bound (11.51) can be easily explained. The ideal coefficient selection (11.32) sets  $X_{\mathcal{B}}[m]$  to zero if and only if  $|f_{\mathcal{B}}[m]| \leq \sigma$ , whereas a hard thresholding sets  $X_{\mathcal{B}}[m]$  to zero when  $|X_{\mathcal{B}}[m]| \leq T$ . If  $|f_{\mathcal{B}}[m]| \leq \sigma$ , then it is very likely that  $|X_{\mathcal{B}}[m]| \leq T$ , because  $T$  is above the noise level. In this case, the hard thresholding sets  $X_{\mathcal{B}}[m]$  to zero as the oracle projector (11.32) does. If  $|f_{\mathcal{B}}[m]| \geq 2T$ , then it is likely that  $|X_{\mathcal{B}}[m]| \geq T$  because  $|W_{\mathcal{B}}[m]| \leq T$ . In this case, the hard thresholding and the oracle projector retain  $X_{\mathcal{B}}[m]$ .

The hard thresholding may behave differently from the ideal coefficient selection when  $|f_{\mathcal{B}}[m]|$  is on the order of  $T$ . The ideal selection yields a risk:  $\min(\sigma^2, |f_{\mathcal{B}}[m]|^2) = \sigma^2$ . If we are unlucky and  $|X_{\mathcal{B}}[m]| \leq T$ , then the thresholding sets  $X_{\mathcal{B}}[m]$  to zero, which produces a risk

$$|f_{\mathcal{B}}[m]|^2 \sim T^2 = 2\log_e N \sigma^2.$$

In this worst case, the thresholding risk is  $2\log_e N$  times larger than the ideal selection risk. Since the proportion of coefficients  $|f_{\mathcal{B}}[m]|$  on the order of  $T$  is often

small, the ratio between the hard-thresholding risk and the oracle projection risk is generally significantly smaller than  $2 \log_e N$ .

### Colored Noise

Thresholding estimators can be adapted when the noise  $W$  is not white. We suppose that  $E\{W[n]\} = 0$ . Since  $W$  is not white,  $\sigma_B[m]^2 = E\{|W_B[m]|^2\}$  depends on each vector  $g_m$  of the basis. As in (11.32) and (11.34), we verify that an oracle projector that keeps all coefficients such that  $|f_B[m]| \geq \sigma_B[m]$  and sets to zero all others has a risk

$$r_{\text{pr}}(f) = \sum_{m=0}^{N-1} \min(|f_B[m]|^2, \sigma_B^2[m]). \quad (11.65)$$

Any linear or nonlinear projector in the basis  $B$  has a risk larger than  $r_{\text{pr}}(f)$ .

Since the noise variance depends on  $m$ , a thresholding estimator must vary the threshold  $T_m$  as a function of  $m$ . Such a hard- or soft-thresholding estimator can be written as

$$\tilde{F} = DX = \sum_{m=0}^{N-1} \rho_{T_m}(X_B[m]) g_m. \quad (11.66)$$

Theorem 11.8 generalizes Theorem 11.7 to compute the thresholding risk  $r_{\text{th}}(f) = E\{\|f - \tilde{F}\|^2\}$ .

**Theorem 11.8:** *Donoho, Johnstone.* Let  $\tilde{F}$  be a hard- or soft-thresholding estimator with

$$T_m = \sigma_B[m] \sqrt{2 \log_e N} \quad \text{for } 0 \leq m < N.$$

Let  $\bar{\sigma}^2 = N^{-1} \sum_{m=0}^{N-1} \sigma_B[m]^2$ . For any  $N \geq 4$ ,

$$r_{\text{th}}(f) \leq (2 \log_e N + 1) (\bar{\sigma}^2 + r_{\text{pr}}(f)). \quad (11.67)$$

The proof of (11.67) is identical to the proof of (11.51). The thresholds  $T_m$  are chosen to be just above the amplitude of each noisy coefficient  $W_B[m]$ .

### Frame Thresholding Estimators

The properties of thresholding estimators remain valid for nonorthogonal Riesz bases and frames. The redundancy of frames often produces a smaller risk than with an orthogonal basis, thanks to their redundancy. They are thus most often used in numerical applications.

Let us recall that  $\{\phi_p\}_{0 \leq p < P}$  with  $P \geq N$  is a frame of  $\mathbb{C}^N$  if there exists  $0 < A \leq B$ , such that for any  $f \in \mathbb{C}^N$

$$A \|f\|^2 \leq \sum_{p=0}^{P-1} |\langle f, \phi_p \rangle|^2 \leq B \|f\|^2.$$

When  $P = N$ , the frame is a Riesz basis, otherwise it is redundant. In the following, we consider that all frame vectors are normalized  $\|\phi_p\| = 1$ . Theorem 5.2 then proves that  $A \leq P/N \leq B$ .

Theorem 5.5 proves that there exists a dual frame  $\{\tilde{\phi}_p\}_{0 \leq p < P}$  such that

$$f = \sum_{p=0}^{P-1} \langle f, \phi_p \rangle \tilde{\phi}_p.$$

A signal  $f$  can be estimated from noisy coefficients  $X = f + W$  by thresholding its frame coefficients

$$\tilde{F} = \sum_{p=0}^{P-1} \rho_T(\langle X, \phi_p \rangle) \tilde{\phi}_p. \quad (11.68)$$

The resulting risk is  $r_{\text{th}}(f) = E\{\|\tilde{F} - f\|^2\}$ . Let us write

$$r_{\text{pr}}(f) = \sum_{p=0}^{P-1} \min(|\langle f, \phi_p \rangle|^2, \sigma^2).$$

Using an oracle, we verify in Exercise 11.12 that  $r_{\text{th}}(f) \geq r_{\text{pr}}(f)/B$ . Moreover, for a threshold  $T = \sigma \sqrt{2 \log_e P}$ , with the same derivation steps as in the proof of Theorem 11.8, one can prove that for any  $P \geq 4$ ,

$$r_{\text{th}}(f) \leq \frac{2 \log_e P + 1}{A} (\sigma^2 + r_{\text{pr}}(f)). \quad (11.69)$$

This proves that thresholding estimators in frames behave like thresholding estimators in orthogonal bases. The threshold  $T = \sigma \sqrt{2 \log_e P}$  is a conservative upper bound that is too large in most numerical experiments. For a tight frame,  $A = B = P/N$ . The thresholding estimator then behaves as the average of  $A$  estimators in  $A$  orthogonal bases. This averaging often reduces the resulting risk.

### 11.2.3 Thresholding Improvements

The thresholding risk is often reduced by choosing a threshold smaller than  $\sigma \sqrt{2 \log_e N}$ . A threshold adapted to the data is calculated by minimizing an estimation of the risk. Different thresholding functions are also considered, and an important improvement is introduced with a translation-invariant thresholding algorithm.

#### Sure Thresholds

To study the impact of the threshold on the risk, we denote by  $r_{\text{th}}(f, T)$  the risk of a soft-thresholding estimator calculated with a threshold  $T$ . An estimate of  $r_{\text{th}}(f, T)$  is calculated from the noisy data  $X$ , and  $T$  is optimized by minimizing the estimated risk.

To estimate the risk  $r_{\text{th}}(f, T)$ , observe that if  $|X_B[m]| < T$ , then the soft thresholding sets this coefficient to zero, which produces a risk equal to  $|f_B[m]|^2$ . Since

$$E\{|X_B[m]|^2\} = |f_B[m]|^2 + \sigma^2,$$

one can estimate  $|f_B[m]|^2$  with  $|X_B[m]|^2 - \sigma^2$ . If  $|X_B[m]| \geq T$ , the soft thresholding subtracts  $T$  from the amplitude of  $X_B[m]$ . The expected risk is the sum of the noise energy plus the bias introduced by the reduction of the amplitude of  $X_B[m]$  by  $T$ . It is estimated by  $\sigma^2 + T^2$ . The resulting estimator of  $r_{\text{th}}(f, T)$  is

$$\text{Sure}(X, T) = \sum_{m=0}^{N-1} C(X_B[m]), \quad (11.70)$$

with

$$C(u) = \begin{cases} u^2 - \sigma^2 & \text{if } u \leq T \\ \sigma^2 + T^2 & \text{if } u > T. \end{cases} \quad (11.71)$$

Theorem 11.9 [222] proves that  $\text{Sure}(X, T)$  is an unbiased risk estimator. It is called a *Stein unbiased risk estimator* (Sure) [445].

**Theorem 11.9:** *Donoho, Johnstone.* For a soft thresholding, the risk estimator  $\text{Sure}(X, T)$  is unbiased:

$$E\{\text{Sure}(X, T)\} = r_{\text{th}}(f, T). \quad (11.72)$$

**Proof.** A soft-thresholding estimator performs a soft thresholding of each noisy coordinate.

As in (11.54), we thus derive that the resulting risk is the sum of the soft-thresholding risk for each coordinate

$$r_{\text{th}}(f, T) = E\{\|f - \tilde{F}\|^2\} = \sigma^2 \sum_{m=0}^{N-1} r(T, f_B[m], \sigma), \quad (11.73)$$

where  $r(\lambda, \mu, \sigma)$  is the risk when estimating  $\mu$  by soft thresholding a Gaussian random variable  $X$  of mean  $\mu$  and variance  $\sigma^2$ :

$$r(\lambda, \mu, \sigma) = E\{|\rho_\lambda(X) - \mu|^2\} = E\{|(X - \lambda \text{sign}(X)) \mathbf{1}_{|X|>\lambda} - \mu|^2\}. \quad (11.74)$$

Let us rewrite

$$r(T, \mu, \sigma) = E\{(X - g(X) - \mu)^2\}, \quad (11.75)$$

where  $g(x) = T \text{sign}(x) + (x - T \text{sign}(x)) \mathbf{1}_{|x|<T}$  is a weakly differentiable function (in the sense of distributions). This risk is calculated by the following Stein formula [445].

**Lemma 11.1:** *Stein.* Let  $g(x)$  be a weakly differentiable function. If  $X$  is a Gaussian random vector of mean  $\mu$  and variance  $\sigma^2$ , then

$$E\{|X + g(X) - \mu|^2\} = \sigma^2 + E\{|g(X)|^2\} + 2\sigma^2 E\{g'(X)\}. \quad (11.76)$$

To prove this lemma, let us develop (11.76)

$$E\{|X + g(X) - \mu|^2\} = E\{(X - \mu)^2\} + E\{|g(X)|^2\} - 2E\{(X - \mu)g(X)\}. \quad (11.77)$$

The probability density of  $X$  is the Gaussian  $\phi_\sigma(y - \mu)$ . The change of variable  $x = y - \mu$  shows that

$$E\{(X - \mu)g(X)\} = \int_{-\infty}^{+\infty} x g(x + \mu) \phi_\sigma(x) dx.$$

Since  $x \phi_\sigma(x) = -\sigma^2 \phi'_\sigma(x)$ , an integration by parts gives

$$\begin{aligned} E\{(X - \mu)g(X)\} &= -\sigma^2 \int_{-\infty}^{+\infty} g(x + \mu) \phi'_\sigma(x) dx \\ &= \sigma^2 \int_{-\infty}^{+\infty} g'(x) \phi_\sigma(x - \mu) dx = E\{g'(X)\}. \end{aligned}$$

Inserting this result in (11.77) proves (11.76).

For the soft-thresholding risk,  $g(x) = T \text{sign}(x) + (x - T \text{sign}(x)) \mathbf{1}_{|x|<T}$ , and thus  $g'(x) = \mathbf{1}_{|x|\leq T}$ . Using the fact that  $E\{\mathbf{1}_{|X|\geq T}\} + E\{\mathbf{1}_{|X|<T}\} = 1$ , the Stein unbiased risk formula (11.76) implies that

$$r(T, \mu, \sigma) = (\sigma^2 + T^2) E\{\mathbf{1}_{|X|\geq T}\} + E\left\{(|X|^2 - \sigma^2) \mathbf{1}_{|X|\leq T}\right\} = E\{C(|X|^2)\}, \quad (11.78)$$

where  $C(x)$  is defined in (11.71). Inserting this expression in (11.73) proves (11.72). ■

These results suggest choosing the threshold that minimizes the Sure estimator

$$\tilde{T} = \operatorname{argmin}_T \text{Sure}(X, T).$$

Although the estimator  $\text{Sure}(X, T)$  of  $r_{\text{th}}(f, T)$  is unbiased, its variance may induce errors leading to a threshold  $\tilde{T}$  that is too small. This happens if the signal energy is small relative to the noise energy:  $\|f\|^2 \ll E\{\|W\|^2\} = N\sigma^2$ . In this case, one must impose  $T = \sigma\sqrt{2 \log_e N}$  in order to remove all the noise. Since  $E\{\|X\|^2\} = \|f\|^2 + N\sigma^2$ , we estimate  $\|f\|^2$  with  $\|X\|^2 - N\sigma^2$  and compare this value with a minimum energy level  $\varepsilon_N = \sigma^2 N^{1/2} (\log_e N)^{3/2}$ . The resulting Sure threshold is

$$T = \begin{cases} \sigma\sqrt{2 \log_e N} & \text{if } \|X\|^2 - N\sigma^2 \leq \varepsilon_N \\ \tilde{T} & \text{if } \|X\|^2 - N\sigma^2 > \varepsilon_N. \end{cases} \quad (11.79)$$

Let  $\Theta$  be a signal set and  $\min_T r_{\text{th}}(\Theta)$  be the minimax risk of a soft thresholding obtained by optimizing the choice of  $T$  depending on  $\Theta$ . Donoho and Johnstone [222] prove that the threshold empirically computed with (11.79) yields a risk  $r_{\text{th}}(\Theta)$  equal to  $\min_T r_{\text{th}}(\Theta)$ , plus a corrective term that decreases rapidly when  $N$  increases if  $\varepsilon_N = \sigma^2 N^{1/2} (\log_e N)^{3/2}$ .

Exercise 11.13 studies a similar risk estimator for hard thresholding. However, this risk estimator is biased. We cannot guarantee that the threshold that minimizes this estimated risk is nearly optimal for hard-thresholding estimations.

### Other Thresholdings and Masking Noise

Besides hard and soft thresholdings, other diagonal attenuation functions in (11.66) can improve a diagonal signal estimation. A whole family of diagonal attenuations is defined by

$$a_m(x) = \max\left(1 - \frac{T^\beta}{|x|^\beta}, 0\right) \quad \text{with } \beta > 0.$$

For  $\beta = 1$ , it corresponds to the soft thresholding (11.48). When  $\beta$  tends to  $+\infty$ , it yields the hard thresholding (11.45).

If  $\beta = 2$ , then  $a_m(x)$  is a James-Stein shrinkage [445]. It is intermediate between a hard and a soft attenuation. Since  $E\{|X_B[m]|^2\} = |f_B[m]|^2 + \sigma^2$ , for  $T = \sigma$  the attenuation

$$a_m(X_B[m]) = \max\left(\frac{|X_B[m]|^2 - \sigma^2}{|X_B[m]|^2}, 0\right)$$

can be interpreted as an empirical estimation of the oracle attenuation factor  $a_m = |f_B[m]|^2 / (|f_B[m]|^2 + \sigma^2)$ . Is also called an *empirical Wiener attenuation*.

Thresholding signal coefficients can introduce perceptual artifacts that reduce the perceived quality of the estimation. The next section shows that in wavelet bases, thresholding noisy image coefficients removes fine texture and can produce cartoonlike images with no textures. Other artifacts can be created by thresholding estimators. Leaving some noise reduces our perceptual sensitivity to these artifacts and can thus improve the perceived signal quality, although it may increase the mean-square norm of the error. It is implemented with attenuation factors that remain strictly positive:

$$a_m(x) = \max\left(1 - \frac{T^\beta}{|x|^\beta}, \varepsilon\right) \quad \text{with } \varepsilon > 0. \quad (11.80)$$

If  $|X_B[m]| \leq T$ , then  $a_m(X_B[m]) = \varepsilon$ , so this thresholding leaves a reduced noise of variance  $\varepsilon^2 \sigma^2$ , which masks potential artifacts.

### Translation-Invariant Thresholding

In many applications, signal models are translation invariant. This is often the case for audio signals, where the recording beginning may be arbitrarily shifted, or for images that are translated by changing the camera position. For stochastic signal models with random processes, translation invariance means that the process is stationary. For a deterministic model that specifies a set  $\Theta$  where the signal belongs, it means that any  $f \in \Theta$  remains in  $\Theta$  after a translation. For signals embedded in additive noise, if the noise is stationary and  $\Theta$  is translation invariant, then the minimization of the maximum risk over  $\Theta$  is achieved with translation-invariant estimators. Theorem 11.12 proves this result for linear minimax estimators, which is also valid for nonlinear estimators.

Coifman and Donoho [179] have introduced translation-invariant thresholding estimators that reduce the risk for translation-invariant sets  $\Theta$ . For signals of finite

length  $N$ , to avoid boundary issues, we consider circular translations modulo  $N$ :  $f_p[n] = f[(n-p)\text{mod}N]$ . Observe that if  $\mathcal{B} = \{g_m\}_{0 \leq m < N}$  is an orthonormal basis of  $\mathbb{C}^N$ , then the translated basis  $\mathcal{B}_p = \{g_{p,m}[n] = g_m[(n-p)\text{mod}N]\}_{0 \leq m < N}$  is also an orthogonal basis of  $\mathbb{C}^N$  for any  $0 \leq p < N$ . If there is no translation information on the signal, all the bases  $\{\mathcal{B}_p\}_{0 \leq p < N}$  are a priori equivalent for thresholding estimations. Coifman and Donoho [179] thus proposed to average the thresholding estimations obtained in these  $N$  bases. This is equivalent to decompose the signal in a translation-invariant dictionary that is a union of these  $N$  translated orthonormal bases:

$$\mathcal{D} = \bigcup_{p=0}^{N-1} \mathcal{B}_p = \{g_{p,m}\}_{0 \leq m, p < N}. \quad (11.81)$$

The energy conservation in each orthogonal basis  $\mathcal{B}_p$  implies a global energy conservation over the  $N^2$  dictionary vectors

$$\|f\|^2 = \frac{1}{N} \sum_{m=0}^{N-1} \sum_{p=0}^{N-1} |\langle f, g_{p,m} \rangle|^2,$$

which proves that this dictionary is a tight frame.

The resulting translation-invariant estimator of  $f$  from noisy data  $X = f + W$  is obtained by thresholding the translation-invariant tight frame coefficients of  $X$ :

$$\tilde{F}[n] = \frac{1}{N} \sum_{p=0}^{N-1} \sum_{m=0}^{N-1} \rho_T(\langle X, g_{p,m} \rangle) g_{p,m}[n], \quad (11.82)$$

where  $\rho_T$  is a hard- or a soft-thresholding operator. Since this estimator is obtained by averaging  $N$  thresholding estimators in orthogonal bases, the resulting thresholding risk satisfies the same upper bound as in Theorem 11.7.

A priori, this translation-invariant thresholding requires  $N$  times more operations than a thresholding estimation in the original basis  $\mathcal{B}$ . However, this is not the case when the original basis  $\mathcal{B}$  includes translated vectors. In this case, the translation-invariant dictionary  $\mathcal{D}$  has less than  $N^2$  different vectors. For example, a wavelet orthogonal basis yields a translation-invariant dyadic wavelet dictionary that has only  $N \log_2 N$  different wavelets.

Translation-invariant tight frames are not necessarily derived from an orthogonal basis. Theorem 5.12 proves that a translation-invariant dictionary obtained by translating  $Q$  generators  $\{g_q\}_{0 \leq q < Q}$  is a tight frame if and only if their discrete Fourier transforms satisfy  $\sum_{q=0}^{Q-1} |\hat{g}_q[k]|^2 = Q$  for all  $0 \leq k < N$ . The simplicity of this condition offers more flexibility to build translation-invariant thresholding estimators than from orthogonal bases.

### 11.3 THRESHOLDING SPARSE REPRESENTATIONS

Thresholding estimators are particularly efficient in a basis that can precisely approximate signals with few nonzero coefficients. The basis must therefore be chosen from prior information on signal properties in order to obtain sparse representations.

Wavelet bases are particularly efficient to estimate piecewise regular signals. Noise removal from images is studied in Section 11.3.2, with wavelet bases and curvelet frames. For audio signals, sparse representations are obtained with localized time-frequency transforms. An important limitation of diagonal-thresholding operators is illustrated in Section 11.3.3, with the creation of a “musical noise” when thresholding windowed Fourier coefficients for audio noise removal.

### 11.3.1 Wavelet Thresholding

Thresholding wavelet coefficients implements an adaptive signal averaging with a kernel that is locally adapted to the signal regularity [4]. Numerical examples illustrate the properties of these estimators for piecewise regular one-dimensional signals. The minimax optimality of wavelet thresholding estimators is studied in Section 11.5.3.

A filter bank of conjugate mirror filters decomposes a discrete signal in a discrete orthogonal wavelet basis, defined in Section 7.3.3. The discrete wavelets  $\psi_{j,m}[n] = \psi_j[n - N2^j m]$  are translated modulo modifications near the boundaries, which are explained in Section 7.5. The support of the signal is normalized to  $[0, 1]$  with  $N$  samples spaced by  $N^{-1}$ . The scale parameter  $2^j$  thus varies from  $2^L = N^{-1}$  up to  $2^J < 1$ :

$$\mathcal{B} = \left[ \{\psi_{j,m}[n]\}_{L \leq j \leq J, 0 \leq m < 2^{-j}}, \{\phi_{J,m}[n]\}_{0 \leq m < 2^{-J}} \right]. \quad (11.83)$$

A thresholding estimator in this wavelet basis can be written as

$$\tilde{f} = \sum_{j=L+1}^J \sum_{m=0}^{2^{-j}} \rho_T(\langle X, \psi_{j,m} \rangle) \psi_{j,m} + \sum_{m=0}^{2^{-J}} \rho_T(\langle X, \phi_{J,m} \rangle) \phi_{J,m}, \quad (11.84)$$

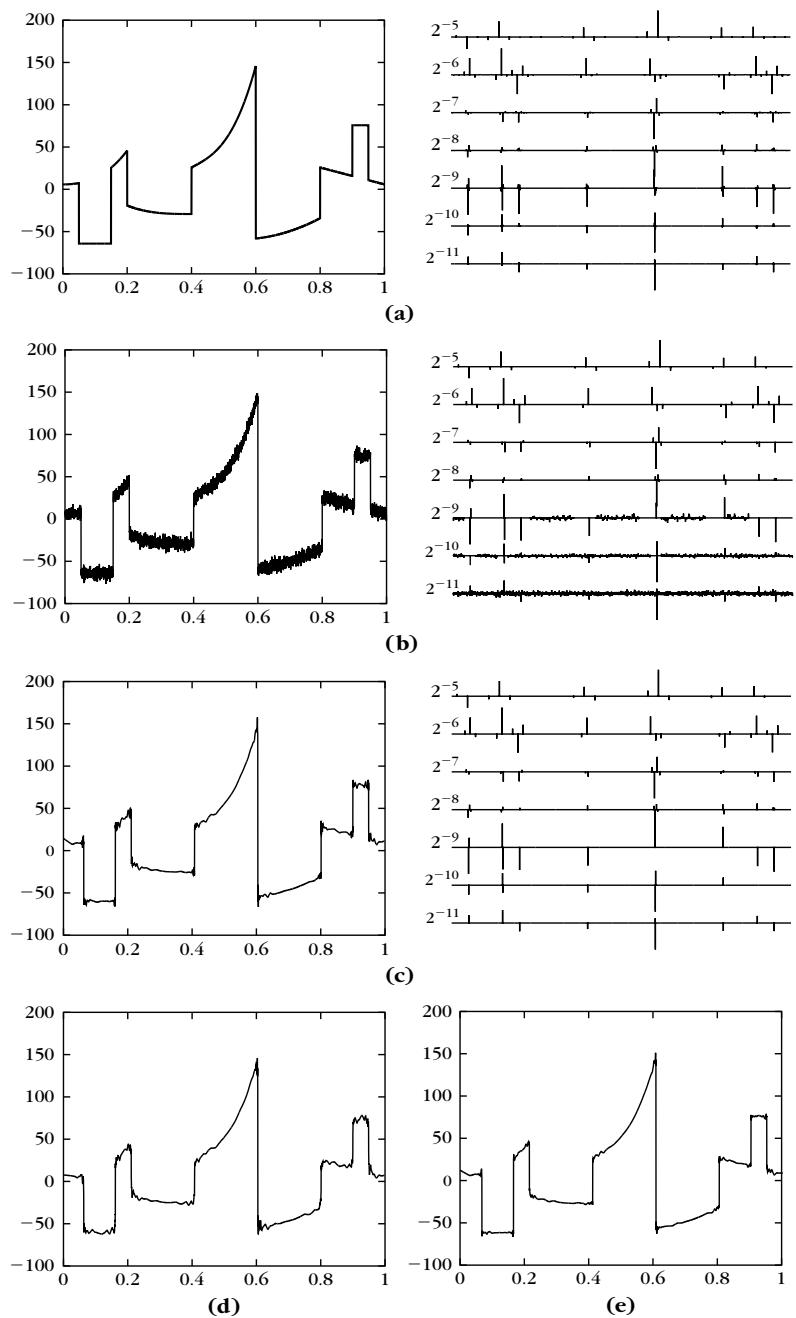
where  $\rho_T$  is a hard thresholding (11.47) or a soft thresholding (11.49). The upper bound (11.51) proves that the estimation risk is small if the energy of  $f$  is absorbed by a few wavelet coefficients.

#### **Adaptive Smoothing**

The thresholding sets all coefficients  $|\langle X, \psi_{j,m} \rangle| \leq T$  to zero. This performs an adaptive smoothing that depends on the regularity of the signal  $f$ . Since  $T$  is above the maximum amplitude of the noise coefficients  $|\langle W, \psi_{j,m} \rangle|$ , if

$$|\langle X, \psi_{j,m} \rangle| = |\langle f, \psi_{j,m} \rangle + \langle W, \psi_{j,m} \rangle| \geq T,$$

then  $|\langle f, \psi_{j,m} \rangle|$  has a high probability of being at least of the order  $T$ . At fine scales  $2^j$ , these coefficients are in the neighborhood of sharp signal transitions, as shown in Figure 11.4(b). By keeping them, we avoid smoothing these sharp variations. In the regions where  $|\langle X, \psi_{j,m} \rangle| < T$ , the coefficients  $\langle f, \psi_{j,m} \rangle$  are likely to be small, which means that  $f$  is locally regular. Setting wavelet coefficients to zero is equivalent to locally averaging the noisy data  $X$ , which is done only if the underlying signal  $f$  appears to be regular.



**FIGURE 11.4**

**(a)** Piecewise polynomial signal (*left*) and its wavelet transform (*right*). **(b)** Noisy signal (SNR = 21.9 db) (*left*) and its wavelet transform (*right*). **(c)** Estimation reconstructed from wavelet coefficients above the threshold (*right*) (SNR = 30.8 db). **(d)** Estimation with wavelet soft thresholding (SNR = 23.8 db). **(e)** Estimation with translation-invariant hard thresholding (SNR = 33.7 db).

### Noise Variance Estimation

To estimate the variance  $\sigma^2$  of the noise  $W[n]$  from the data  $X[n] = W[n] + f[n]$ , we need to suppress the influence of  $f[n]$ . When  $f$  is piecewise smooth, a robust estimator is calculated from the median of the finest-scale wavelet coefficients [221].

The signal  $X$  of size  $N$  has  $N/2$  wavelet coefficients  $\{\langle X, \psi_{l,m} \rangle\}_{0 \leq m < N/2}$  at the finest scale  $2^l = 2N^{-1}$ . The coefficient  $|\langle f, \psi_{l,m} \rangle|$  is small if  $f$  is smooth over the support of  $\psi_{l,m}$ , in which case  $\langle X, \psi_{l,m} \rangle \approx \langle W, \psi_{l,m} \rangle$ . In contrast,  $|\langle f, \psi_{l,m} \rangle|$  is large if  $f$  has a sharp transition in the support of  $\psi_{l,m}$ . A piecewise regular signal has few sharp transitions, and thus produces a number of large coefficients that is small compared to  $N/2$ . At the finest scale, the signal  $f$  thus influences the value of a small portion of large-amplitude coefficients  $\langle X, \psi_{l,m} \rangle$  that are considered to be “outliers.” All others are approximately equal to  $\langle W, \psi_{l,m} \rangle$ , which are independent Gaussian random variables of variance  $\sigma^2$ .

A robust estimator of  $\sigma^2$  is calculated from the median of  $\{\langle X, \psi_{l,m} \rangle\}_{0 \leq m < N/2}$ . The median of  $P$  coefficients  $\text{Med}(\alpha_p)_{0 \leq p < P}$  is the value of the middle coefficient  $\alpha_{n_0}$  of rank  $P/2$ . As opposed to an average, it does not depend on the specific values of coefficients  $\alpha_p > \alpha_{n_0}$ . If  $M$  is the median of the absolute value of  $P$  independent Gaussian random variables of zero mean and variance  $\sigma_0^2$ , then one can show that

$$E\{M\} \approx 0.6745 \sigma_0.$$

The variance  $\sigma^2$  of the noise  $W$  is estimated from the median  $M_X$  of  $\{|\langle X, \psi_{l,m} \rangle|\}_{0 \leq m < N/2}$  by neglecting the influence of  $f$ :

$$\tilde{\sigma} = \frac{M_X}{0.6745}. \quad (11.85)$$

Indeed,  $f$  is responsible for few large-amplitude outliers, and these have little impact on  $M_X$ .

### Hard or Soft Thresholding

If  $T = \sigma \sqrt{2 \log_e N}$ , then (11.50) shows that a soft thresholding guarantees with a high probability that

$$|\langle \tilde{F}, \psi_{j,m} \rangle| = |\rho_T(\langle X, \psi_{j,m} \rangle)| \leq |\langle f, \psi_{j,m} \rangle|.$$

The estimator  $\tilde{F}$  is at least as regular as  $f$  because its wavelet coefficients have a smaller amplitude. This is not true for the hard-thresholding estimator, which leaves the coefficients above  $T$  unchanged, and which can therefore be larger than those of  $f$  because of the additive noise component.

Figure 11.4(a) shows a piecewise polynomial signal of degree at most 3, and its wavelet coefficients calculated with a symmlet 4. Figure 11.4(c) gives an estimation computed with a hard thresholding of the noisy wavelet coefficients in Figure 11.4(b). An estimator  $\tilde{\sigma}^2$  of the noise variance  $\sigma^2$  is calculated with the median (11.85) and the threshold is set to  $T = \tilde{\sigma} \sqrt{2 \log_e N}$ . Thresholding wavelet

coefficients remove the noise in the domain where  $f$  is regular but some traces of the noise remain in the neighborhood of singularities. The resulting SNR is 30.8 db. The soft-thresholding estimation of Figure 11.4(d) attenuates the noise effect at the discontinuities but the reduction by  $T$  of the coefficient amplitude is much too strong, which reduces the SNR to 23.8 db. As already explained, to obtain comparable SNR values, the threshold of the soft thresholding must be about half the size of the hard-thresholding one. In this example, reducing the threshold by 2 increases the SNR of the soft thresholding to 28.6 db.

### Multiscale Sure Thresholds

Piecewise regular signals have a proportion of large coefficients  $|\langle f, \psi_{j,m} \rangle|$  that increases when the scale  $2^j$  increases. Indeed, a singularity creates the same number of large coefficients at each scale, whereas the total number of wavelet coefficients increases when the scale decreases. To use this prior information, one can adapt the threshold choice to the scale  $2^j$ . At large scale  $2^j$  the threshold  $T_j$  should be smaller in order to avoid setting too many large-amplitude signal coefficients to zero, which would increase the risk.

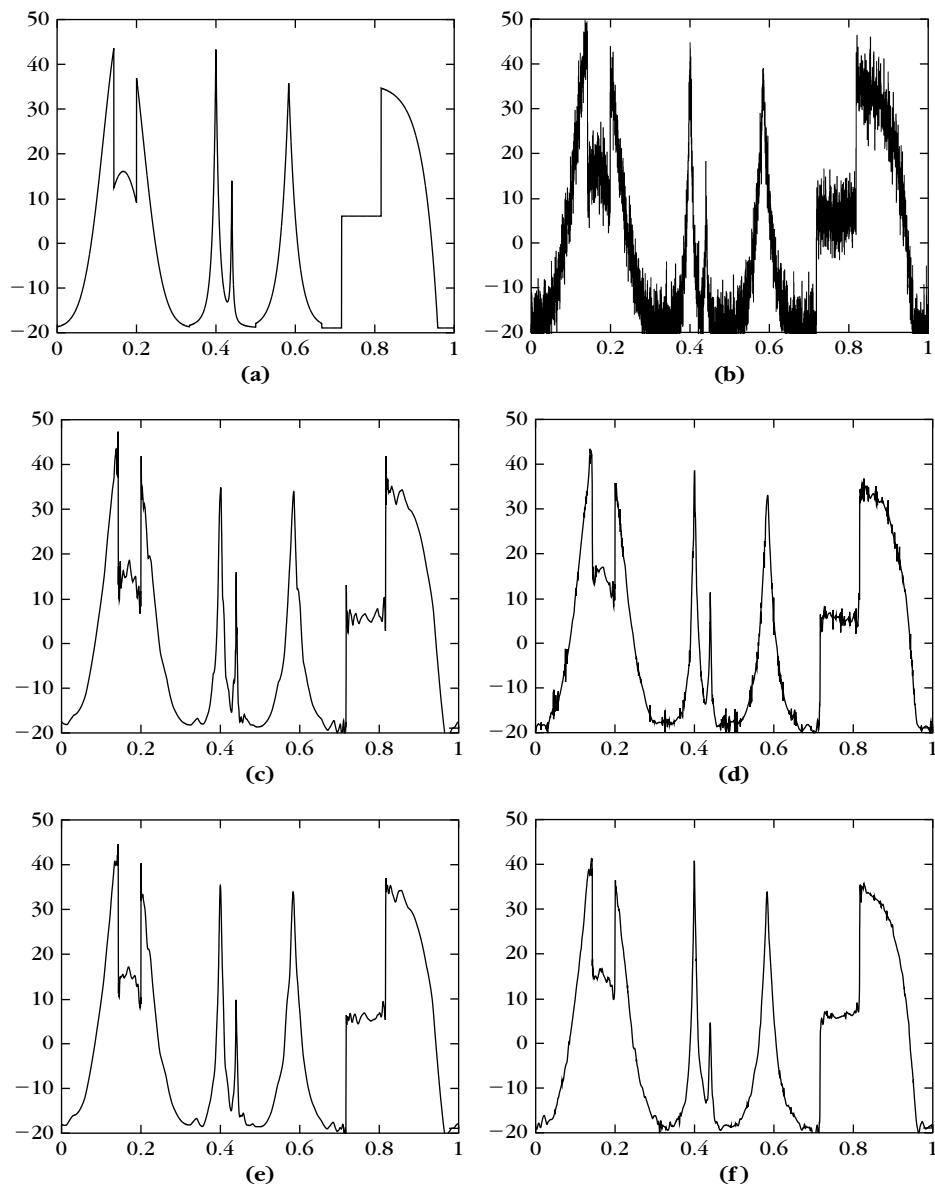
Section 11.2.3 explains how to compute the threshold value for a soft thresholding from the coefficients of the noisy data. We first compute an estimate  $\tilde{\sigma}^2$  of the noise variance  $\sigma^2$  with the median formula (11.85) at the finest scale. At each scale  $2^j$ , a different threshold is calculated from the  $2^{-j}$  noisy coefficients  $\{X, \psi_{j,m}\}_{0 \leq m < 2^{-j}}$  with the algorithm of Section 11.2.3. A Sure threshold  $T_j$  is calculated with (11.79) at each scale  $2^j$ . A soft thresholding is then applied at each scale  $2^j$ , with a threshold  $T_j$ . For a hard thresholding, we have no reliable formula to estimate the risk and thus compute an adapted threshold by minimizing the estimated risk. However, ad hoc hard thresholds may be computed by multiplying by 2 the Sure threshold calculated for a soft thresholding.

Figure 11.5(c) is a hard-thresholding estimation calculated with the same threshold  $T = \tilde{\sigma} \sqrt{2 \log_e N}$  at all scales  $2^j$ . The SNR is 23.3 db. Figure 11.5(d) is obtained by a soft thresholding with Sure thresholds  $T_j$  adapted at each scale  $2^j$ . The SNR is 24.1 db. A soft thresholding with the threshold  $T = \tilde{\sigma} / 2 \sqrt{2 \log_e N}$  at all scales gives a smaller SNR equal to 21.7 db. The adaptive calculation of thresholds clearly improves the estimation.

### Translation Invariance

Thresholding noisy wavelet coefficients creates small ripples near discontinuities, as seen in Figures 11.4(c,d) and 11.5(c,d). Indeed, setting a coefficient  $\langle f, \psi_{j,m} \rangle$  to zero subtracts  $\langle f, \psi_{j,m} \rangle \psi_{j,m}$  from  $f$ , which introduces oscillations whenever  $\langle f, \psi_{j,m} \rangle$  is nonnegligible. Figures 11.4(e) and 11.5(e,f) show that averaging the signal estimation over translated wavelet bases reduces these oscillations, significantly improving the SNR.

A translation-invariant wavelet thresholding estimator decomposes the noisy data  $X$  over a dictionary obtained by translating each orthogonal wavelet  $\psi_{j,m}[n] = \psi_j[n - N2^j m]$  by any factor  $0 \leq p < N$  modulo  $N$ . Suppose that each of the  $J-L$

**FIGURE 11.5**

**(a)** Original signal. **(b)** Noisy signal ( $\text{SNR} = 13.1 \text{ db}$ ). **(c)** Estimation by a hard thresholding in a wavelet basis (symmlet 4) with  $T = \tilde{\sigma}\sqrt{2\log_e N}$  ( $\text{SNR} = 23.3 \text{ db}$ ). **(d)** Soft thresholding calculated with Sure thresholds  $T_j$  adapted to each scale  $2^j$  ( $\text{SNR} = 24.5 \text{ db}$ ). **(e)** Translation-invariant hard thresholding with  $T = \tilde{\sigma}\sqrt{2\log_e N}$  ( $\text{SNR} = 25.7 \text{ db}$ ). **(f)** Translation-invariant soft thresholding with Sure thresholds ( $\text{SNR} = 25.6 \text{ db}$ ).

wavelet  $\psi_j[n]$  is  $N$  periodic. This yields a translation-invariant dyadic wavelet tight frame including  $(J-L)N$  wavelets:

$$\mathcal{D} = \{\psi_j[n-p], \phi_j[n-p]\}_{L < j \leq J, 0 \leq p < N},$$

and the resulting translation-invariant thresholding estimator can be written as

$$\tilde{F}[n] = \sum_{j=L+1}^J \sum_{p=0}^{N-1} \rho_T(\langle X[q], \psi_j[q-p] \rangle) \psi_j[n-p] + \sum_{p=0}^{N-1} \rho_T(\langle X[q], \phi_J[q-p] \rangle) \phi_J[n-p].$$

The decomposition coefficients of  $X$  in this dictionary are provided by the dyadic wavelet transform defined in Section 5.2:

$$WX[2^j, p] = \langle X[n], \psi_j[n-p] \rangle \quad \text{for } 0 \leq p < N.$$

The *algorithme à trous* from Section 5.2.2 computes these  $(J-L)N$  coefficients for  $L < j \leq J$  with  $O(N(J-L))$  operations and reconstructs a signal from the thresholded coefficients with the same number of operations. Since  $(J-L) \leq \log_2 N$ , the total number of operations is bounded by  $O(N \log_2 N)$ .

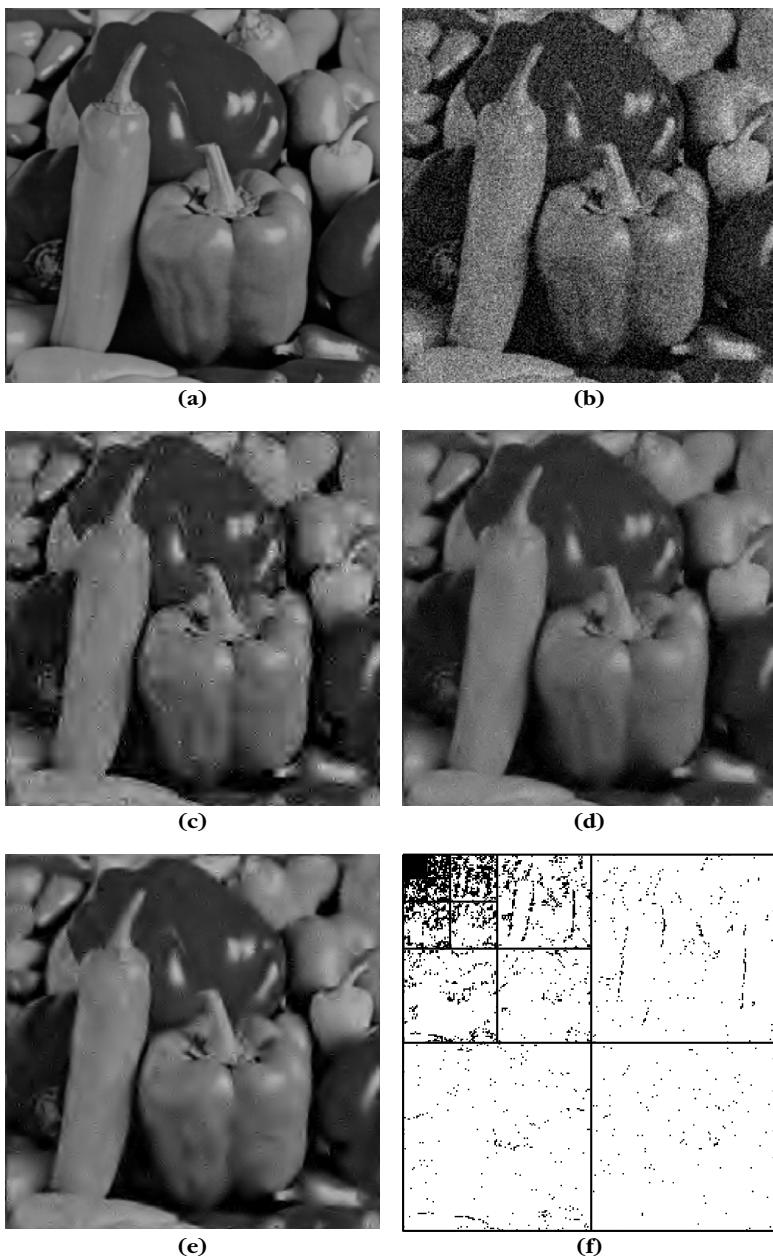
### 11.3.2 Wavelet and Curvelet Image Denoising

Reducing noise by thresholding wavelet coefficients is particularly effective for piecewise regular images, which have sparse wavelet representations. When images include edges or textures that have a regular geometry, then curvelet frames can improve wavelet thresholding estimators.

#### Wavelet Bases

Figure 11.6(b) shows an example of an image contaminated by an additive Gaussian white noise of variance  $\sigma^2$ . This image is decomposed in a separable two-dimensional biorthogonal wavelet basis, generated by a 7/9 mother wavelet. As in one dimension, an estimator  $\tilde{\sigma}$  of  $\sigma$  is computed from the median  $M_X$  of the finest-scale noisy wavelet coefficient amplitudes with (11.85). For images of  $N = 512^2$  pixels, the universal threshold of Theorem 11.7 is  $T = \tilde{\sigma}\sqrt{2\log_e N} \approx 5\tilde{\sigma}$ . Wavelet hard-thresholding estimators are improved by choosing  $T = 3\tilde{\sigma}$ , which significantly increases the SNR and the visual quality of the image. Figure 11.6(c) gives an example. Figure 11.6(d) shows a soft-thresholding estimation with  $T = 3\tilde{\sigma}/2$  from the same wavelet coefficients. For hard- and soft-thresholdings estimations, low-frequency scaling coefficients are not thresholded. A hard thresholding at  $T$  and a soft thresholding at  $T/2$  set the same wavelet coefficients to zero, which are shown in white in Figure 11.6(f). A hard thresholding does not modify the other coefficients shown in black, whereas a soft thresholding reduces their amplitude by  $T/2$ . Coefficients are mostly kept near edges, but some isolated noise coefficients above  $3\tilde{\sigma}$  remain in regular regions.

These isolated noise wavelet coefficients above the threshold produce small wavelet oscillation artifacts that are more visible with a hard thresholding. The visual quality of edges is also affected by small Gibbs-like oscillations, which also

**FIGURE 11.6**

(a) Original image. (b) Noisy image ( $\text{SNR} = 18 \text{ db}$ ). (c) Hard thresholding in a 7/9 separable wavelet basis ( $\text{SNR} = 21.6 \text{ db}$ ). (d) Soft thresholding ( $\text{SNR} = 22.6 \text{ db}$ ). (e) Translation-invariant hard thresholding ( $\text{SNR} = 24.7 \text{ db}$ ). (f) Wavelet coefficients above  $T = 3\hat{\sigma}$  are shown in black. All other coefficients are set to zero by the hard and soft thresholding.

appear in the one-dimensional estimations in Figure 11.4(c) and Figure 11.5(c). The soft thresholding improves the SNR by 1 db relatively to the hard thresholding, and for most images an improvement between 0.5 db and 1 db is observed, with or without finer optimizations of thresholds. A Sure optimization of thresholds at each scale with (11.79) further increases the soft thresholding SNR by over 0.5 db.

A translation-invariant wavelet thresholding estimator is computed by decomposing the image in a two-dimensional translation-invariant dyadic wavelet tight frame. A fast dyadic wavelet transform is implemented with a separable filter bank, similar to the two-dimensional fast orthogonal transform described in Section 7.7.3. The one-dimension filterings and subsamplings along the image rows and columns are replaced by the filterings of the *algorithme à trous* in Section 5.2.2. It requires  $O(N \log_2 N)$  operations. Figure 11.6(e) is calculated with translation-invariant hard thresholding, which gives a much higher SNR of 24.7 db and a better visual quality. A translation-invariant soft thresholding gives an SNR of 23.6 db. Although a soft thresholding is typically better than a hard thresholding in an orthogonal or biorthogonal basis, a hard thresholding improves a soft-thresholding SNR in a translation-invariant wavelet frame and yields the best results. A translation-invariant hard thresholding often removes fine textures that affect the visual image quality. By maintaining a small masking noise with  $\rho_T(x) = |x|$  if  $|x| \geq T$  and  $\rho_T(x) = \varepsilon |x|$  if  $|x| > T$ , the restored image can look more natural.

In Section 11.5.3 we prove that a thresholding in a wavelet basis has a nearly minimax risk for bounded variation images. When the noise variance  $\sigma^2$  decreases to zero, the wavelet thresholding risk is bounded by  $O(\sigma \log \sigma)$ . Irregular or oscillatory textures are not as well estimated because they do not have a sparse wavelet representation and create many nonnegligible wavelet coefficients. Block thresholding algorithms, presented in Section 11.4.2, can improve texture estimation with wavelets.

### Curvelet Frames

Section 9.3.3 shows that images including structures that are geometrically regular, such as  $C^2$  piecewise regular images, have a representation that is asymptotically more sparse with curvelets than with wavelets. Thresholding curvelet coefficients can then improve wavelet thresholding estimators. This is also valid for textures including geometrically regular structures.

Curvelet tight frames  $\{c_{j,m}^\alpha\}_{j,m,\alpha}$ , presented in Section 5.5.2, are composed of anisotropic waveforms with different scales and directions. Curvelets have an elongated support proportional to  $2^{j/2}$  in a direction  $\alpha \in [0, \pi]$  and a width proportional to  $2^j$  in the perpendicular direction. They are translated along a grid with intervals that are, respectively,  $2^{j/2}$  in the direction  $\alpha$  and  $2^j$  in the direction  $\alpha + \pi/2$ . In numerical implementations, normalized curvelets have a frame bound  $A = B \geq 5$ , which corresponds to a minimum redundancy factor of 5. A thresholding curvelet estimation of  $f$  from a noisy observation  $X = f + W$  can be written as

$$\tilde{F} = \sum_{j,m,\alpha} \rho_T((X, c_{j,m}^\alpha)) c_{j,m}^\alpha.$$

Suppose that  $f$  is obtained by discretizing a  $C^2$  piecewise regular image with  $C^2$  edge curves, as specified by Definition 9.1. Theorem 9.20 proves that a nonlinear curvelet approximation error has a decay bounded by  $O(M^{-2}(\log M)^3)$ , which improves the asymptotic error decay of wavelet approximations. Theorem 11.5 for  $s = 3/2$  proves that a nonlinear approximation error  $\varepsilon_n(M, f) = O(M^{-2})$  yields an oracle projection risk that satisfies  $r_{\text{pr}}(f) = O(\sigma^{4/3})$ . A thresholding estimator has the same decay up to a  $\log_e P$  factor where  $P = AN$  is the total number of curvelets.

Taking into account the  $(\log M)^3$  factor, Candès and Donoho [141] derive that the risk of a curvelet thresholding of a  $C^2$  piecewise regular image satisfies

$$E\{\|\tilde{F} - f\|^2\} = O(|\log \sigma|^2 \sigma^{4/3}),$$

when the noise variance  $\sigma$  decreases to zero. This improves the risk decay  $O(|\log \sigma| \sigma)$  of a wavelet thresholding estimator. We later prove in (11.152), with  $\alpha = 2$ , that the nonlinear minimax risk over uniformly  $C^2$  Lipschitz images decays like  $\sigma^{4/3}$ . A curvelet thresholding estimator nearly achieves this decay despite the presence of edges, and is therefore asymptotically minimax for the class of  $C^2$  piecewise regular images, up to the  $|\log \sigma|^2$  factor.

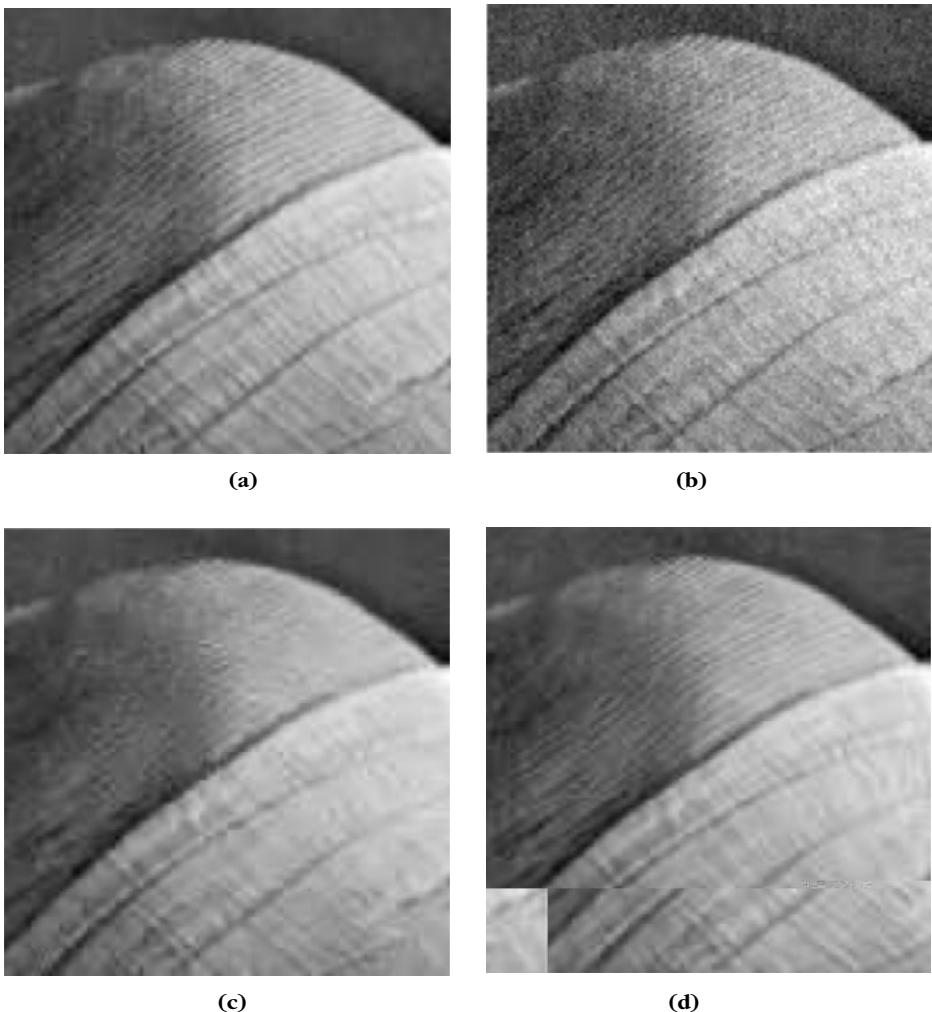
The threshold  $T = \sigma \sqrt{\log_e P}$  is conservative and Figure 11.7 gives an example with  $T = 3\sigma$ , which improves the SNR. When the image has textures with regular stripes as in Lena's hat, a curvelet thresholding gives a better SNR than a translation-invariant wavelet thresholding. However, despite the asymptotic improvements of curvelets on  $C^2$  piecewise regular images, the pepper image in Figure 11.6 is better estimated with wavelets than with curvelets. As opposed to wavelets, curvelets do not have a compact spatial support and their decay is not exponential because their Fourier transform has a compact support. This increases the number of high-amplitude curvelet coefficients created by edges, which impacts the image estimation.

Irregular textures or pointwise singularities have a representation that is more sparse with wavelets than with curvelets, and are thus better estimated by a wavelet thresholding. Other image representations may also be used. A thresholding in a best bandlet basis, presented in Section 12.2.4, adapts the basis to the geometric image regularity, and chooses wavelets when there is no such regularity.

### 11.3.3 Audio Denoising by Time-Frequency Thresholding

Audio signals, whether music or speech, often have a sparse time-frequency representation. Such signals are well approximated by relatively few coefficients in appropriate time-frequency bases or frames. Thus, one may expect that thresholding these time-frequency representations yields effective noise-removal algorithms. Although this is true from a SNR point of view, diagonal thresholding algorithms degrade the audio signal quality by introducing a “musical noise.” This musical noise is produced by isolated noisy time-frequency coefficients above the threshold.

Sparse audio representations are obtained in wavelet packet and local cosine orthogonal bases that have the time-frequency localization that can be adapted to



**FIGURE 11.7**

(a) Original image. (b) Noisy image ( $\text{SNR} = 22 \text{ dB}$ ). (c) Translation-invariant wavelet hard thresholding ( $\text{SNR} = 25.3 \text{ dB}$ ). (d) Curvelet tight frame hard thresholding ( $\text{SNR} = 26 \text{ dB}$ ).

the signal properties. Window Fourier frames also have a time-frequency localization that can be adjusted by choosing an appropriate window size. Thresholding the complex modulus of windowed Fourier frame coefficients seems to better preserve perceptual sound quality than thresholding real wavelet packet or local cosine coefficients. This could be explained by a better restoration of the phase, which is perceptually important for sounds. We shall thus concentrate on windowed Fourier frame thresholding.

A discrete windowed Fourier tight frame of  $\mathbb{C}^N$  is constructed in Section 5.4 by translating and modulating a window  $g[n]$ , which has a support included in  $[-K/2, K/2 - 1]$ . If  $M$  divides  $N$  and

$$\sum_{m=0}^{N/M-1} |g[n - mM]|^2 = \frac{A}{K} \quad \text{for } 0 \leq n < N$$

then Theorem 5.18 proves that

$$\{g_{m,k}[n] = g[n - mM] e^{i2\pi kn/K}\}_{0 \leq k < K, 0 \leq m < N/M}$$

is a tight frame of  $\mathbb{C}^N$ , with a frame bound equal to  $A$ . Numerical experiments in Figure 11.8 are performed using a square root Hanning window  $g[n] = \sqrt{2/K} \cos(\pi n/K)$  with  $M = K/2$  and thus  $A = 2$ . The resulting windowed Fourier frame coefficients for  $0 \leq k < K, 0 \leq m < N/M$  are

$$Sf[m, k] = \langle f, g_{m,k} \rangle = \sum_{n=-K/2}^{K/2-1} f[n] g[n - mM] e^{-i2\pi kn/K}.$$

Audio noises are often stationary but not necessarily white. The time-frequency noise variance thus only depends on the frequency and depends on the noise power spectrum  $\sigma_B^2[m, k] = \sigma_B^2[k]$ . A windowed Fourier thresholding estimator can then be written as

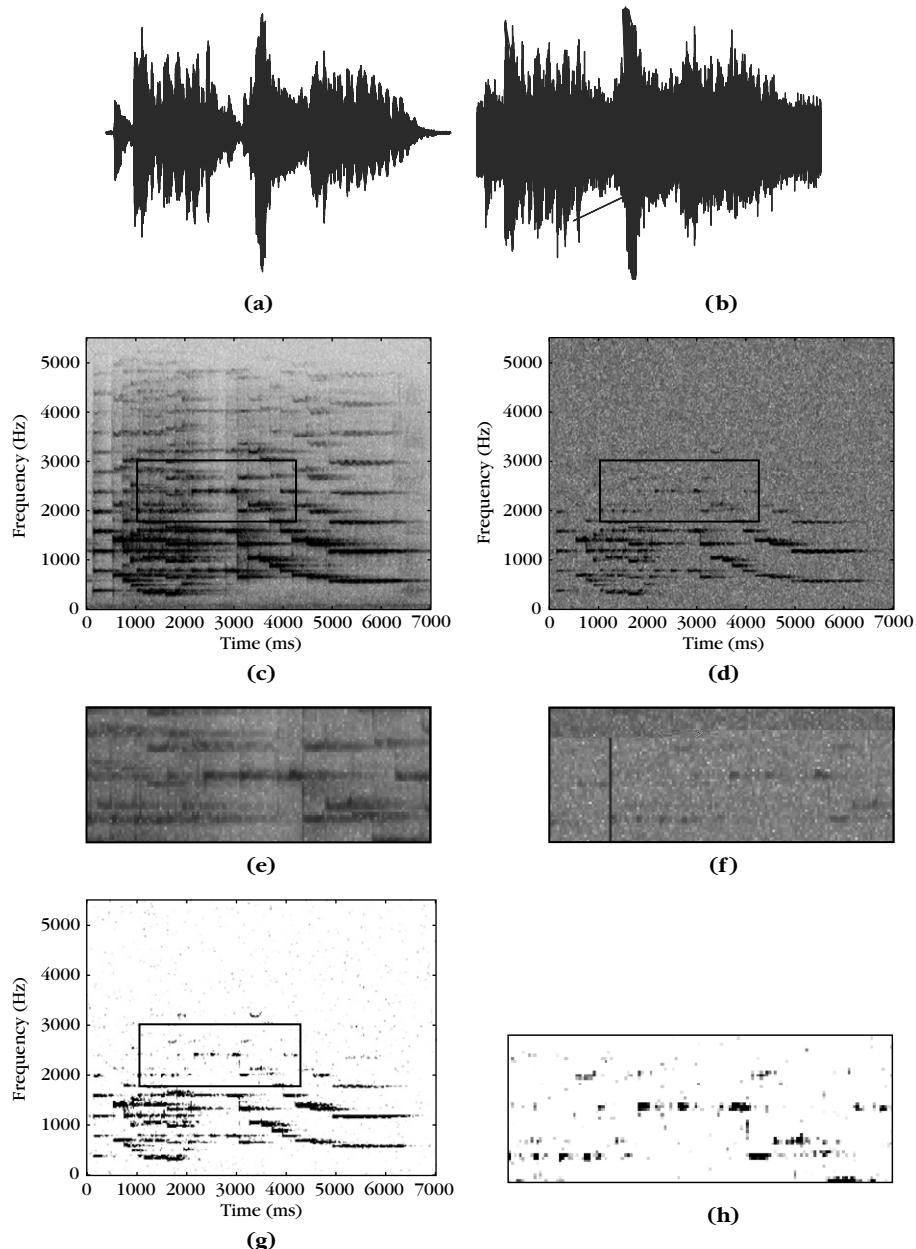
$$\tilde{F} = \sum_{m=0}^{N/M-1} \sum_{k=0}^{K-1} \rho_{T_k}(\langle X, g_{m,k} \rangle) g_{m,k} \quad (11.86)$$

with a threshold  $T_k^2 = \lambda \sigma_B^2[k]$ . Since the early work on time-frequency audio denoising [109], many types of thresholding functions have been studied for time-frequency audio noise removal [430]. The James-Stein estimator, called empirical Wiener estimator or “power subtraction” in audio noise removal, is often used,

$$a_{k,m}(\langle X, g_{m,k} \rangle) = \frac{\rho_{T_k}(\langle X, g_{m,k} \rangle)}{\langle X, g_{m,k} \rangle} = \max \left( 1 - \frac{T_k^2}{|\langle X, g_{m,k} \rangle|^2}, \varepsilon \right), \quad (11.87)$$

with a masking noise factor  $\varepsilon$  that is often nonzero.

To illustrate the musical noise produced by a spectrogram thresholding, Figure 11.8 shows the denoising of a short recording of a Mozart oboe concerto with a white Gaussian noise. Figures 11.8(c, d) give, respectively, the log spectrograms  $\log |Sf[m, k]|$  and  $\log |SX[m, k]|$  of the original signal  $f$  and of the noisy sound  $X$ . Figure 11.8(g) displays the attenuation factors  $a_{k,m}$  in (11.87) with  $\varepsilon = 0$ . Black points correspond to  $a_{k,m} = 1$  and white points to  $a_{k,m} = 0$ . For this Mozart recording, when the noisy signal has a SNR that ranges between -2 db up to 15 db, the SNR improvement of this time-frequency soft-thresholding estimator is between 8 db and 10 db, which is important. However, as it can be observed in the zoom in Figure 11.8(h), there are isolated attenuation coefficients  $a_{k,m} \approx 1$  corresponding to black points, which retain noise coefficients in time-frequency regions where the signal has no energy. Similar isolated points appear in the estimation support

**FIGURE 11.8**

**(a, b)** Original and noisy “Mozart” recording (0 dB). **(c, d)** Log spectrograms of the original and noisy signals. **(e, f)** Zoom on the spectrograms in (c, d). **(g)** Attenuation factors (11.87) computed from noisy coefficients in (d). Black and white pixels correspond, respectively, to 1 and 0. **(h)** Zoom in on the attenuation factors in (g).

$\tilde{\Lambda}_T$  of Figure 11.6(f) for a wavelet image estimation. Because of these isolated attenuation coefficients  $a_{k,m} \approx 1$ , the estimator (11.86) restores windowed Fourier vectors  $g_{m,k}[n]$  that are perceived as a “musical noise.” Despite its small energy, this musical noise is clearly perceived because it is not masked by a sound component at a close frequency and time. Audio masking properties are explained in Section 10.3.3. Despite the SNR improvement, this “musical noise” can be more annoying than the original white noise. Translation-invariant spectrogram thresholding barely improves the musical noise problem. It can be reduced by increasing thresholds, but this attenuates too much audio signal information, and thus also degrades the sound quality. A nonzero masking noise factor  $\varepsilon$ , which maintains a background noise, can be used to reduce the perception of musical noises.

Section 11.4 shows that effective musical noise reduction requires using nondiagonal time-frequency estimators, which regularize the time-frequency estimation by processing coefficients in groups.

## 11.4 NONDIAGONAL BLOCK THRESHOLDING

A diagonal estimator in a basis processes each coefficient independently and thus does not take advantage of potential dependencies between neighbor coefficients. Ideally, an optimized representation takes advantage of all structural signal correlations to improve the signal sparsity. In practice, this is not the case. When a coefficient has a large amplitude, it is likely that some other neighborhood coefficients are also nonnegligible, because of signal dependencies that are not fully taken into account by the representation. For example, wavelet image transforms do not capture the geometric regularity of edges, which produce large wavelet coefficients along curves.

Block thresholding estimators introduced by Cai [129] take advantage of such properties by grouping coefficients in blocks and by taking a decision over these blocks. This grouping regularizes thresholding estimators, which improves the resulting risk. It also avoids leaving isolated noise coefficients above the threshold, perceived as “musical noises” in audio signals and that appear as isolated oscillations in images.

Block thresholding estimators are introduced in Section 11.4.1 together with their mathematical properties. Section 11.4.2 studies the improvements of block thresholding estimations in wavelet bases for piecewise regular signals and images. For audio noise, we show in Section 11.4.3 that time-frequency block thresholdings are effective estimators that avoid introducing musical noises.

### 11.4.1 Block Thresholding in Bases and Frames

A block thresholding estimator implements thresholding decisions over groups of coefficients. The input noisy signal  $X = f + W$  is decomposed in an orthonormal basis  $\mathcal{B} = \{g_m\}_{0 \leq m < M}$  of  $\mathbb{C}^N$ , and we write

$$X_{\mathcal{B}}[m] = \langle X, g_m \rangle, \quad f_{\mathcal{B}}[m] = \langle f, g_m \rangle, \quad W_{\mathcal{B}}[m] = \langle W, g_m \rangle, \quad \text{and} \quad \sigma_{\mathcal{B}}^2[m] = E\{|W_{\mathcal{B}}[m]|^2\}.$$

The interval  $[0, N - 1]$  of all indexes  $m$  is partitioned in  $Q$  disjoint blocks  $\{B_q\}_{1 \leq q \leq Q}$  of indexes that are grouped together. A block diagonal estimator multiplies all coefficients in  $B_q$  with the same attenuation factor  $a_q$

$$\tilde{F} = DX = \sum_{q=1}^Q a_q \sum_{m \in B_q} X_B[m] g_m, \quad (11.88)$$

where each  $a_q$  depends on all coefficients  $X_B[m]$  for  $m \in B_q$ . If all blocks are reduced to a single coefficient, then a block thresholding is a diagonal estimator, otherwise this estimator is not diagonal. Lower bounds of the risk are first computed with “oracles.”

### Oracle Block Attenuations

If  $a_q$  is a constant in the block  $B_q$ , then the risk  $r(D, f)$  of the block estimator (11.88) is

$$r(D, f) = E \left\{ \|f - \tilde{F}\|^2 \right\} = \sum_{q=1}^Q \sum_{m \in B_q} E\{|f_B[m] - a_q X_B[m]|^2\}. \quad (11.89)$$

Since  $X_B = f_B + W_B$  and  $E\{|W_B[m]|^2\} = \sigma_B^2[m]$ , it follows that

$$\sum_{m \in B_q} E\{|f_B[m] - a_q X_B[m]|^2\} = (1 - a_q)^2 \|f_B\|_{B_q}^2 + a_q^2 \|\sigma_B\|_{B_q}^2, \quad (11.90)$$

with

$$\|f_B\|_{B_q}^2 = \sum_{m \in B_q} |f_B[m]|^2 \quad \text{and} \quad \|\sigma_B\|_{B_q}^2 = \sum_{m \in B_q} \sigma_B^2[m].$$

This error is minimized by an oracle attenuation factor,

$$a_q = \frac{\|f_B\|_{B_q}^2}{\|f_B\|_{B_q}^2 + \|\sigma_B\|_{B_q}^2}. \quad (11.91)$$

In a Bayesian framework, estimating this coefficient amounts to estimating an a priori SNR  $\|f_B\|_{B_q}^2 / \|\sigma_B\|_{B_q}^2$  regularized over a block.

An oracle block projection estimator simplifies the estimation by imposing that  $a_q \in \{0, 1\}$ . The minimization of the risk (11.90) gives

$$a_q = \begin{cases} 1 & \text{if } \|f_B\|_{B_q} \geq \|\sigma_B\|_{B_q} \\ 0 & \text{if } \|f_B\|_{B_q} < \|\sigma_B\|_{B_q}. \end{cases} \quad (11.92)$$

This oracle estimator is thus an orthogonal projection of  $X$

$$DX = \sum_{m \in \bar{\Lambda}_\sigma} X_B[m] g_m,$$

where  $\bar{\Lambda}_\sigma$  is the union of all blocks  $B_q$  such that  $\|f_B\|_{B_q} \geq \|\sigma_B\|_{B_q}$ . It is a block approximation of the set of coefficients  $\Lambda_\sigma = \{0 \leq m < N : |f_B[m]| \geq \sigma_B[m]\}$ , which defines the diagonal oracle projector (11.33). The resulting minimum projection risk is

$$\bar{r}_{\text{pr}}(f) = E\{\|f - \tilde{F}\|^2\} = \sum_{q=1}^Q \min(\|f_B\|_{B_q}^2, \|\sigma_B\|_{B_q}^2). \quad (11.93)$$

The risk  $\bar{r}_{\text{pr}}(f)$  of an oracle block projector is always larger than the risk  $r_{\text{pr}}(f)$  of a diagonal oracle projector calculated in (11.65),

$$\bar{r}_{\text{pr}}(f) \geq r_{\text{pr}}(f) = \sum_{m=0}^{N-1} \min(|f_B[m]|^2, \sigma_B[m]^2), \quad (11.94)$$

because a single thresholding decision is taken over a whole block  $B_q$  and not for each coefficient. Both risks are equal if the thresholding decisions are the same, which means that within each block  $B_q$ , all coefficients are below the noise or all coefficients are above the noise:

$$\forall m \in B_q \ \sigma_B[m] \leq |f_B[m]| \quad \text{or} \quad \forall m \in B_q \ \sigma_B[m] > |f_B[m]|.$$

It is nearly the case if large coefficients are aggregated together, and if the blocks are not too large.

### **Block Thresholding**

To approximate oracle block projections, block thresholding decisions are computed from the empirical noisy signal energy on each block

$$\|X_B\|_{B_q}^2 = \sum_{m \in B_q} |X_B[m]|^2.$$

The resulting block thresholding estimator is

$$\tilde{F} = DX = \sum_{q=1}^Q a_q(\|X_B\|_{B_q}) \sum_{m \in B_q} X_B[n] g_m. \quad (11.95)$$

The soft block thresholding of Cai [129, 130] is implemented with the James-Stein thresholding rule:

$$0 \leq a_q(x) = a_{T_q}(x) = \max\left(1 - \frac{T_q^2}{x^2}, 0\right) \leq 1 \quad (11.96)$$

for a threshold  $T_q^2 = \lambda \|\sigma_B\|_{B_q}^2$  that is proportional to the noise energy. A hard block thresholding is implemented with a hard-thresholding decision:

$$a_q(x) = a_{T_q}(x) = \begin{cases} 1 & \text{if } |x| > T_q \\ 0 & \text{if } |x| \leq T_q, \end{cases} \quad (11.97)$$

with  $T_q^2 = \lambda \|\sigma_B\|_{B_q}^2$ , but it is usually not used because its mathematical and numerical properties are not as effective as a soft James-Stein block thresholding.

**Thresholding Risk**

Let us denote  $\bar{r}_{\text{th}}(f) = E\{\|\tilde{F} - f\|^2\}$  as the risk of a soft block thresholding estimator. Suppose that  $W$  is a Gaussian white noise of variance  $\sigma^2$  and that all blocks have the same size  $L$ . The noise energy in each block is then  $\|\sigma_B\|_{B_q}^2 = L\sigma^2$ . Theorem 11.10 [129] computes an upper bound of the block thresholding risk, which is related to the risk  $\bar{r}_{\text{pr}}(f)$  of an oracle block projector.

**Theorem 11.10:** *Cai.* Let  $T^2 = \lambda L\sigma^2$ . The risk  $\bar{r}_{\text{th}}(f)$  of a soft block thresholding estimator satisfies

$$\bar{r}_{\text{th}}(f) \leq \lambda \bar{r}_{\text{pr}}(f) + 4N\sigma^2 P(\chi_L^2 > \lambda L). \quad (11.98)$$

If  $L = \log_e N$  and  $T = \sigma\sqrt{\lambda_* \log_e N}$  with  $\lambda_* = 4.50524$ , then

$$\bar{r}_{\text{th}}(f) \leq \lambda_* \bar{r}_{\text{pr}}(f) + 2\sigma^2. \quad (11.99)$$

**Proof.** The noisy coefficient  $X_B[m] = f_B[m] + W_B[m]$  is a Gaussian random variable of mean  $f_B[m]$  and variance  $\sigma^2$ . Over a block  $B$  of size  $L$  the soft block thresholding estimator can be written as

$$\tilde{F}_B = X_B + g(X_B),$$

where  $g$  is defined over any vector  $x[m]$  for  $m \in B$  by

$$g(x) = (1 - \frac{\lambda L \sigma^2}{\|x\|_B^2})_+ x - x.$$

The resulting block thresholding risk can thus be computed with a sure estimation, using the following Stein lemma 11.2 which generalizes the one-dimensional lemma (11.1).

**Lemma 11.2:** *Stein.* Let  $g(x) = (g_1(x), \dots, g_L(x))$  be a weakly differentiable function from  $\mathbb{R}^L$  to  $\mathbb{R}^L$ . Let us write  $\nabla \cdot g(x) = \sum_{l=1}^L \frac{\partial g_l(x)}{\partial x[l]}$ . If  $X$  is a Gaussian random vector of mean  $\mu \in \mathbb{R}^L$  and covariance matrix  $\sigma^2 \text{Id}$ , then

$$E\{\|X + g(X) - \mu\|_B^2\} = E\{L\sigma^2 + \|g(X)\|_B^2 + 2\sigma^2 \nabla \cdot g(X)\}. \quad (11.100)$$

The proof is a multidimensional extension of the one-dimensional proof given for Lemma 11.1 and can be found in [445]. For  $g(x) = (1 - \frac{\lambda L \sigma^2}{\|x\|_B^2})_+ x - x$ , Lemma 11.2 implies that

$$E\{\|\tilde{F}_B - f_B\|_B^2\} = E\{\text{Sure}(X_B, T, L, \sigma)\}, \quad (11.101)$$

and applying (11.100) with an algebraic calculation gives

$$\begin{aligned} \text{Sure}(X_B, \lambda, L, \sigma) &= \sigma^2 \left( L + \frac{\lambda^2 L^2 - 2\lambda L(L-2)}{\|X_B\|_B^2 / \sigma^2} \mathbf{1}_{(\|X_B\|_B^2 > \lambda L \sigma^2)} \right. \\ &\quad \left. + (\|X_B\|_B^2 / \sigma^2 - 2L) \mathbf{1}_{(\|X_B\|_B^2 \leq \lambda L \sigma^2)} \right). \end{aligned} \quad (11.102)$$

Since  $\|\sigma^{-1} X_B\|^2$  is a sum of  $L$  independent squared normal random variables, it has a  $\chi_L^2$  distribution with  $L$  degrees of freedom.

The overall thresholding risk is

$$\bar{r}_{\text{th}}(f) = \sum_{q=1}^Q E\{\|\tilde{f}_B - f_B\|_{B_q}^2\}.$$

By inserting (11.101) and (11.102), through several technical lemma that are not reproduced here, Cai [129] derives that

$$\bar{r}_{\text{th}}(f) \leq \sum_{q=1}^Q \min(\|f_B\|_{B_q}^2, T^2) + 4N\sigma^2 P(\chi_L^2 > \lambda L). \quad (11.103)$$

Since each block has a size  $L, T^2 = \lambda L \sigma^2 = \lambda \|f_B\|_{B_q}^2$ , so (11.103) implies (11.98). Cai also proves in [129] that  $P(\chi_L^2 > \lambda L) \leq 1/(2N)$  if  $L = \log_e N$  and  $\lambda = \lambda_*$  with  $\lambda_* - \log_e \lambda_* = 3$ . Inserting this result in (11.98) proves (11.99). ■

The upper bound (11.98) of the block thresholding risk  $\bar{r}_{\text{th}}(f)$  has two terms that balance the bias and variance of this estimator. The second term  $N\sigma^2 P(\chi_L^2 > \lambda L)$  is the average error produced by blocks of noisy coefficients above the threshold when the signal is zero. For a diagonal thresholding ( $L = 1$ ), this residual noise corresponds to the “musical noise” that appears when thresholding time-frequency representations of noisy audio recordings. When the block size  $L$  increases, this residual noise energy decreases. However, the bias of the oracle risk  $\bar{r}_{\text{pr}}(f)$ , and thus of  $\bar{r}_{\text{th}}(f)$ , increases with  $L$ . Indeed, larger blocks reduce the flexibility of block thresholding, which computes a single attenuation factor over each block. If large signal coefficients have a tendency to be aggregated, then increasing  $L$  up to a maximum value reduces the residual noise energy more than it increases the bias of  $\bar{r}_{\text{pr}}(f)$ . This is why a block thresholding can reduce the risk of a diagonal thresholding. Setting  $L = \log_e N$  and  $\lambda = \lambda_*$  gives in (11.99) a block thresholding risk that is of the same order as the oracle risk  $\bar{r}_{\text{pr}}(f)$ , but it may not minimize the thresholding risk  $\bar{r}_{\text{th}}(f)$  because it may increase the oracle risk too much.

For a given block size  $L$ , the threshold  $T$  and thus  $\lambda$  are adjusted to balance the error produced by signal coefficients set to zero and the remaining noise energy of coefficients above the threshold. It can be computed by maximizing a Sure estimation of the SNR, as described in the next subsection. One can also set a priori the residual noise probability

$$P(\chi_L^2 > \lambda L) = \delta. \quad (11.104)$$

This strategy is used when this residual noise affects the perceived signal quality more than the SNR, as is the case for time-frequency audio denoising. For  $\delta = 0.1\%$ , Table 11.1 gives the values of  $\lambda$  depending on  $L$ .

### **Sure Block Size Estimation**

To optimize the choice of thresholds and block sizes, as in Section 11.2.3, the Sure approach by Cai and Zhou [131] minimizes the Stein unbiased risk estimator. The

**Table 11.1** Thresholding Parameter  $\lambda$  Calculated for Different Block Size  $L$ , with a Residual Noise Probability  $\delta = 0.1\%$

$L$	4	8	16	32	64	128
$\lambda$	4.7	3.5	2.5	2.0	1.8	1.5

Sure risk estimator of a soft block thresholding over a block  $B$  of  $L$  coefficient is calculated in (11.102) by supposing that the noise is Gaussian and has a constant variance  $\sigma^2$  in the directions of all the vectors of  $B$ . The block noise energy is then  $\|\sigma_B\|_B^2 = L\sigma^2$ . A global risk estimator is obtained by inserting this expression in (11.102) and by summing over the  $Q = N/L$  blocks of a signal of size  $N$ :

$$\begin{aligned} \text{Sure}(X_B, \lambda, L, \sigma_B) &= \|\sigma_B\|^2 \\ &+ \sum_{q=1}^Q \left( \frac{\lambda^2 \|\sigma_B\|_{B_q}^2 - 2\lambda \|\sigma_B\|_{B_q}^2 (L-2)}{\|X_B\|_{B_q}^2 / \|\sigma_B\|_{B_q}^2} \mathbf{1}_{(\|X_B\|_{B_q}^2 > \lambda \|\sigma_B\|_{B_q}^2)} \right. \\ &\quad \left. + (\|X_B\|_{B_q}^2 - 2\|\sigma_B\|_{B_q}^2) \mathbf{1}_{(\|X_B\|_{B_q}^2 \leq \lambda \|\sigma_B\|_{B_q}^2)} \right). \end{aligned} \quad (11.105)$$

If the noise is Gaussian white with a variance  $\sigma^2$ , then  $\|\sigma_B\|_{B_q}^2 = L\sigma^2$  for all blocks  $1 \leq q \leq Q$ . If the noise is not white but its covariance is nearly diagonalized in  $B$ , then this formula remains approximately valid as long as the noise coefficient variances remain nearly constant over each block  $B_q$ . Indeed, over each block the noise behaves as a white noise.

The Sure minimization approach by Cai and Zhou [131] computes the block size and threshold parameters that minimize the Sure estimated risk:

$$(\tilde{L}, \tilde{\lambda}) = \arg \min_{L, \lambda} \text{Sure}(X_B, \lambda, L, \sigma_B). \quad (11.106)$$

In applications, the minimization is performed over a limited set of possible values for the block size  $L$  and for  $\lambda$ . A different  $\lambda$  may also be chosen with (11.104) by adjusting the probability  $\delta$  of the residual noise.

In Section 11.2.3 we explain that when the signal is small relative to the noise energy, the variance of the Sure risk estimator is large and the resulting computed thresholds may be too small. As in (11.79), this case is avoided by estimating the signal energy  $\|f\|^2$  with  $\|X\|^2 - N\sigma^2$  and comparing it with  $\varepsilon_N = \sigma^2 N^{1/2} (\log_e N)^{3/2}$ . When the noise energy is too small, the block size is set to 1 and we use the universal threshold  $\sigma\sqrt{2\log_e N}$ . The resulting threshold and block sizes are derived from the Sure minimization parameters (11.106) with:

$$(T, L) = \begin{cases} (\sigma\sqrt{2\log_e N}, 1) & \text{if } \|X\|^2 - N\sigma^2 \leq \varepsilon_N \\ (\sigma\sqrt{\tilde{\lambda}\tilde{L}}, \tilde{L}) & \text{if } \|X\|^2 - N\sigma^2 > \varepsilon_N. \end{cases} \quad (11.107)$$

### **Frame Block Thresholding and Translation Invariance**

Block thresholding results extend from orthogonal bases to general frames, by thresholding blocks of frame coefficients with the same thresholding formula. Since Theorem 11.10 is proved with a calculation on a single block, this theorem can be extended from orthogonal bases to frames with minor modifications. Optimal block sizes may also be estimated with the Sure procedure previously described.

Similar to diagonal thresholding, block thresholding can be improved with a translation-invariant procedure that decomposes the signal over a translation-invariant tight frame (11.81)

$$\mathcal{D} = \{g_{m,p}[n] = g_m[(n-p)\bmod N]\}_{0 \leq m, p < N}. \quad (11.108)$$

A fully translation-invariant block thresholding requires using overlapping blocks of equal sizes that are translated. A block  $B_{m,p}$  obtained by translating by  $p$  the block  $B_{m,0}$  is associated to each coefficient  $X_B[m, p] = \langle X, g_{m,p} \rangle$ . The resulting block thresholding estimator is

$$\tilde{F}[n] = \sum_{m=0}^{N-1} \sum_{p=0}^{N-1} a_T(\|X_B\|_{B_{m,p}}) X_B[m, p] g_{m,p}[n], \quad (11.109)$$

where  $a_T$  is the soft attenuation (11.96). In this case, a thresholding decision is performed for each coefficient, but this thresholding decision is regularized by the block energy averaging.

### **11.4.2 Wavelet Block Thresholding**

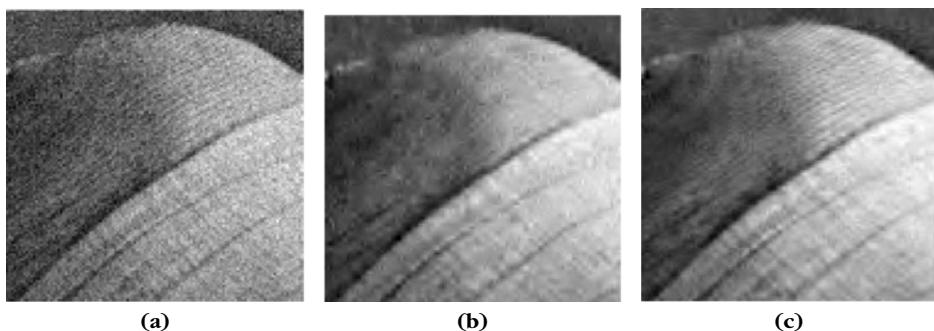
Block thresholding improves wavelet thresholding estimations when large-amplitude wavelet coefficients are often aggregated together. With appropriate block sizes, block thresholds are smaller than diagonal thresholds, which preserves more signal information with less residual noise. The Sure block thresholding chooses a threshold and block size that is fixed at each scale  $2^j$ , but that varies with  $2^j$ .

#### **One-Dimensional Signals**

In one dimension, a pointwise signal singularity produces at each scale  $2^j$  about three large-amplitude orthogonal wavelet coefficients and even more of smaller relative amplitude. This suggests using blocks of size  $L_j \approx 3$  for piecewise regular signals that have isolated singularities. This block size adjustment is performed automatically by the Sure block size and threshold optimizations (11.106) and (11.107). For the piecewise regular signals in Figures 11.4 and 11.5, the block sizes are indeed between 2 and 4. Since the blocks are small, the improvement of this block thresholding over a diagonal thresholding with  $L_j = 1$  is marginal, of about 0.3 db for the signals in Figures 11.4 and 11.5.

#### **Images**

In images, sharp transitions are often distributed along edge curves or in texture regions. Large-amplitude wavelet coefficients are thus not only aggregated because

**FIGURE 11.9**

**(a)** Noisy image zoom (SNR = 28.1 db on the whole image). **(b)** Denoising with a Sure diagonal soft thresholding of orthogonal wavelet coefficients (SNR = 33.4 db). **(c)** Denoising with a Sure soft block thresholding estimation (SNR = 34.2 db).

of the wavelet support but also because of the geometric image regularity. Block thresholding algorithms take advantage of this geometric property to reduce the estimation risk.

For two-dimensional wavelet bases, blocks have a fixed size of  $L_j \times L_j$  pixels that are optimized at each scale  $2^j$  with the Sure optimization together with the thresholds. The resulting soft block thresholding improves the SNR by about 1 db over images such as Lena. Figure 11.9(b) shows a diagonal soft-thresholding estimation ( $L_j = 1$ ) over orthogonal wavelet coefficients with a Sure estimation that adapts thresholds at each scale. The Sure threshold optimization improves the SNR by 0.7 db relative to a soft thresholding with  $T = 3\sigma/2$  at all scales. Figure 11.9(c) gives a Sure block thresholding estimation over orthogonal wavelet coefficients. Block width  $L_j$  is typically equal to 3 or 4 and  $\lambda$  remains nearly equal to 1. Thus, blocks include between 9 and 16 coefficients. Sure block thresholding restores the texture on Lena's hat much better because of block averaging. On the whole Lena image, the SNR of the noisy image is 28.1 db. A Sure diagonal wavelet soft thresholding gives an SNR of 33.4 db, and a Sure soft block thresholding an SNR of 34.2 db.

With a translation-invariant wavelet transform, a diagonal hard thresholding with  $T = 3\sigma$  yields an SNR of 34.5 db, which is better than with a soft thresholding, and a soft Sure block thresholding gives an SNR of 35 db. A Sure block thresholding chooses block sizes  $L_j$  between  $3 \times 2^j$  and  $4 \times 2^j$  with  $\lambda$  remaining around 1. The improvement is then only 0.5 db, but the image quality is visually improved because textures are better restored.

### 11.4.3 Time-Frequency Audio Block Thresholding

Most effective audio denoising algorithms are implemented with nondiagonal adaptive attenuations of time-frequency signal coefficients. Section 11.3.3 explains

that diagonal thresholding algorithms introduce a “musical noise” corresponding to the residual noise above the threshold in time-frequency regions where there is no signal energy. To regularize this estimation, Ephraim and Malah [247, 248] have introduced a time-recursive filtering of time-frequency coefficients that considerably reduces the musical noise. This has led to a large body of complex time-frequency estimators regularized with a time-frequency averaging [147, 176].

Time-frequency block thresholdings also avoid introducing musical noise due to the block averaging that regularizes the thresholding decision. Yu, Mallat, and Bacry [495] showed that automatic parameter adjustments by minimizing the Sure estimated risk produce a low-risk and high-audio perceptual quality.

In Section 11.3.3, on audio time-frequency thresholding, the noisy signal  $X$  is decomposed in a windowed Fourier tight frame of  $\mathbb{C}^N$

$$SX[m, k] = \langle X, g_{m,k} \rangle = \sum_{n=-K/2}^{K/2-1} X[n] g[n - mM] e^{-i2\pi kn/K},$$

for  $0 \leq k < K$  and  $0 \leq m < N/M$ . A block thresholding computes the noisy spectrogram signal energy over blocks  $B_q$  that define a partition of the time-frequency index plane:

$$\|X_B\|_{B_q}^2 = \sum_{(m,k) \in B_q} |SX[m, k]|^2.$$

The resulting soft block thresholding estimator is

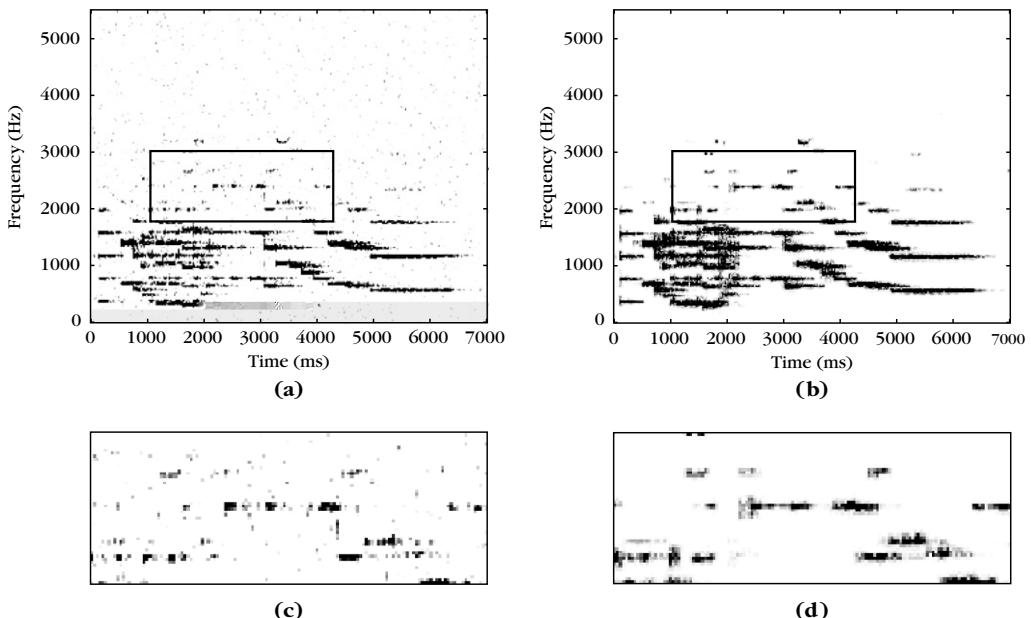
$$\tilde{F} = \sum_{q=1}^Q a_{T_k}(\|X_B\|_{B_q}) \sum_{m \in B_q} SX[m, k] g_m, \quad (11.110)$$

where  $a_{T_k}$  implements the soft James-Stein attenuation (11.96). The audio noise is often stationary, in which case its variance only depends on the frequency index  $\sigma_B^2[m, k] = \sigma_B^2[k]$ .

To nearly remove all musical noise, the residual noise probability  $P(\chi_L^2 > \lambda L) = \delta$  is set to a low value, for example,  $\delta = 0.1\%$  [495]. The block size  $L$  is adjusted by minimizing the resulting Sure estimated risk. Section 11.4.1 explains that ideally a block includes either signal coefficients that are all above the threshold or all below the threshold in order to minimize the bias error. Thus, block estimation can be improved by adjusting their shape. In regions where the signal includes attacks, it is preferable to use blocks that are narrow in time with a larger frequency width, because the signal energy is delocalized in frequency but concentrated in time. This eliminates “pre-echo” artifacts on signal onsets and results in less distortion on signal transients. For a musical signal, including precise harmonics that have a narrow frequency width, blocks should be narrow in frequency and more elongated in time in order to match the signal time-frequency resolution.

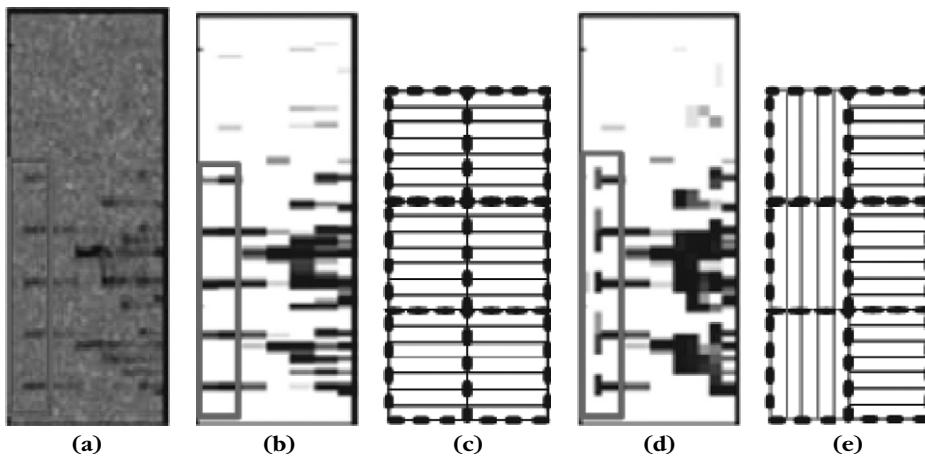
If  $(L_m, L_k)$  are the time and frequency widths of a block, its total size is  $L = L_m \times L_k$ . Sure parameter estimation independently adjusts the time and frequency widths of time-frequency blocks  $B_q$ . Long audio signals are divided into segments of  $N$  coefficients, and over each subpiece of size  $N$  the block sizes and thresholds are computed by minimizing the Sure risk estimator (11.105) with  $L = L_m \times L_k$ . This Sure risk is calculated over a set of possible time and frequency widths  $(L_m, L_k)$ , and for each of them  $\lambda$  is computed by adjusting the residual noise probability  $P(\chi_L^2 > \lambda L) = \delta$ .

Numerical experiments are performed with 15 possible block sizes  $L_m \times L_k$  with  $L_m = 8, 4, 2$  and  $L_k = 16, 8, 4, 2, 1$ . Figure 11.10 compares the attenuation coefficients  $a_{m,k}$  of a diagonal thresholding in (a) and of a block thresholding in (b). The zoom in Figures 11.10(c, d) shows that nondiagonal block thresholding attenuation factors are more regular and do not include isolated points corresponding to a residual noise above the threshold, perceived as a musical noise [495]. When the SNR of the noisy Mozart recording ranges between -2 db and 15 db, the SNR improvement of a block thresholding relative to a soft diagonal thresholding is between 1 db and 1.5 db [495]. More signal components are recovered because the thresholding factors  $\lambda$  of a block thresholding are at least twice smaller than



**FIGURE 11.10**

**(a, b)** Attenuation coefficients calculated, respectively, with a soft thresholding and a block thresholding on the spectrogram of the noisy “Mozart” signal in Figure 11.8. Black pixels correspond to 1 and white to 0. **(c, d)** Zooms over rectangular regions indicated in (a, b).

**FIGURE 11.11**

Zoom on the onset of “Mozart.” **(a)** Log spectrogram. **(b)** Attenuation coefficients of a fixed block thresholding. **(c)** Block sizes ( $L_m = 8, L_k = 1$ ) at the signal onset indicated by a vertical rectangle in (b). **(d)** Attenuation coefficients of an adaptive block thresholding. **(e)** Adapted block sizes ( $L_m = 2, L_k = 9$ ) at the signal onset and ( $L_m = 16, L_k = 1$ ) afterwards.

with a diagonal thresholding. Besides this SNR improvement, the audio quality is considerably improved because of the musical noise removal. Better SNR is also obtained with a block thresholding than with classical Ephraim and Malah-type algorithms [247, 248].

Figure 11.11(a) zooms on the onset of the “Mozart” signal that has a log spectrogram shown in Figure 11.8(b). The attenuation factors of block thresholding with a fixed block size  $L_m = 8$  and  $L_k = 1$  are displayed in Figure 11.11(b). At the beginning of the harmonics, blocks of large attenuation factors spread before the signal’s onset. Figure 11.11(c) illustrates the horizontal blocks used to compute the block attenuation factors in Figure 11.11(b). In the time interval where the blocks exceed the signal onset, a moderate attenuation is performed, and since noise is not eliminated, a transient noise component is heard before the signal begins, which is perceived as a “pre-echo” artifact. In Figures 11.11(c,d), the minimization of the Sure estimation risk chooses blocks of shorter length  $L$  just before and after the onset, which nearly eliminates the “pre-echo” artifact. After onset, more narrow horizontal blocks are selected ( $L_m = 16, L_k = 1$ ) to better capture narrow harmonic signal structures.

## 11.5 DENOISING MINIMAX OPTIMALITY

Section 11.2.2 proves that a thresholding estimator in an orthogonal basis is nearly optimal compared to any diagonal estimator in this basis. It remains to be understood how these estimators compare to all possible linear and nonlinear estimators.

There is often no appropriate stochastic model for complex signals. Thus, we use deterministic models where any prior information is used to specify the smallest possible set  $\Theta$  of potential signals. In this minimax framework, our goal is to minimize the maximum risk over  $\Theta$  and understand under which conditions linear or nonlinear diagonal estimators in a basis  $\mathcal{B}$  can nearly reach this minimax risk.

### **Discrete Minimax Risk**

Given noisy data  $X[n] = f[n] + W[n]$  for  $0 \leq n < N$ , where  $W$  is a Gaussian white noise of variance  $\sigma^2$ , we study the property of estimators  $\tilde{F} = DX$  for any  $f \in \Theta$ . The maximum risk over  $\Theta$  is

$$r(D, \Theta) = \sup_{f \in \Theta} r(D, f) \quad \text{with} \quad r(D, f) = E\{\|\tilde{F} - f\|^2\}.$$

The *linear minimax risk* and *nonlinear minimax risk* are the minimum achievable risk, respectively, over the class  $\mathcal{O}_l$  of all linear operators and  $\mathcal{O}_n$  of all operators (linear and nonlinear) from  $\mathbb{C}^N$  to  $\mathbb{C}^N$ :

$$r_l(\Theta) = \inf_{D \in \mathcal{O}_l} r(D, \Theta) \quad \text{and} \quad r_n(\Theta) = \inf_{D \in \mathcal{O}_n} r(D, \Theta).$$

Sections 11.5.1 and 11.5.2 provide tools to compute linear and nonlinear minimax risks depending on the geometric properties of  $\Theta$ , and explain how they compare to the risk of diagonal estimators in a basis  $\mathcal{B}$ . Block thresholding estimators are not specifically studied because they have almost the same asymptotic properties as diagonal thresholding estimators when the noise variance  $\sigma$  tends to zero [130].

### **Analog Minimax Risk**

Discrete signals most often result from the discretization of analog signals. Signal models are defined over analog signals  $\tilde{f}$ , by specifying a prior set of functions  $\Theta$  where the analog signal belongs. It may, for example, be derived from information on the signal regularity. The discretization of these analog signals defines a set  $\Theta$  of discrete signals in  $\mathbb{C}^N$  where the estimation is computed. Since analog signals are often restored at the end of the processing chain, we must compute the resulting risk and relate it to the risk computed over discrete signals.

Section 11.5.3 computes the risk of linear and thresholding wavelet estimators for different types of signal and image models, by applying the results of Sections 11.5.1 and 11.5.2. For uniformly regular signals and images, linear wavelet estimators are proved to be asymptotically optimal among all linear and nonlinear estimators. However, when the signal regularity is not uniform, thresholding estimators considerably outperform linear estimators. The risk is computed for piecewise regular signals as well as for bounded variation signals and images, where it is proved that wavelet thresholding estimators are nearly minimax optimal. Readers more interested in algorithms and numerical applications may skip the following sections, which are mathematically more involved.

### 11.5.1 Linear Diagonal Minimax Estimation

To estimate  $f[n] \in \Theta$  from noisy measurements  $X[n]$ , we first study the maximum risk of a linear diagonal operator in an orthogonal basis  $\mathcal{B} = \{g_m\}_{0 \leq m < N}$ . Such an estimator can be written as

$$\tilde{F} = DX = \sum_{m=0}^{N-1} a_m X_{\mathcal{B}}[m] g_m, \quad (11.111)$$

where each  $a_m$  is a constant. Let  $\mathcal{O}_{l,d}$  be the set of all linear diagonal operators  $D$ . Since  $\mathcal{O}_{l,d} \subset \mathcal{O}_l$ , the *linear diagonal minimax risk* is larger than the linear minimax risk

$$r_{l,d}(\Theta) = \inf_{D \in \mathcal{O}_{l,d}} r(D, \Theta) \geq r_l(\Theta) = \inf_{D \in \mathcal{O}_l} r(D, \Theta).$$

We characterize diagonal estimators that achieve the linear diagonal minimax risk. If  $\Theta$  is translation invariant, we prove that diagonal operators in a discrete Fourier basis reach the global linear minimax risk:  $r_{l,d}(\Theta) = r_l(\Theta)$ .

#### Quadratic Convex Hull

The “square” of a set  $\Theta$  in the basis  $\mathcal{B}$  is defined by

$$(\Theta)_{\mathcal{B}}^2 = \{\tilde{f} : \tilde{f} = \sum_{m=0}^{N-1} |f_{\mathcal{B}}[m]|^2 g_m \text{ with } f \in \Theta\}. \quad (11.112)$$

We say that  $\Theta$  is *quadratically convex* in  $\mathcal{B}$  if  $(\Theta)_{\mathcal{B}}^2$  is a convex set. A hyperrectangle  $\mathcal{R}_h$  in  $\mathcal{B}$  of vertex  $h \in \mathbb{C}^N$  is a simple example of a quadratically convex set defined by

$$\mathcal{R}_h = \left\{ f : |f_{\mathcal{B}}[m]| \leq |h_{\mathcal{B}}[m]| \text{ for } 0 \leq m < N \right\}.$$

The *quadratic convex hull*  $\text{QH}[\Theta]$  of  $\Theta$  in the basis  $\mathcal{B}$  is defined by

$$\text{QH}[\Theta] = \left\{ f : \sum_{m=0}^{N-1} |f_{\mathcal{B}}[m]|^2 g_m \text{ is in the convex hull of } (\Theta)_{\mathcal{B}}^2 \right\}. \quad (11.113)$$

It is the largest set with a square  $(\text{QH}[\Theta])_{\mathcal{B}}^2$  equal to the convex hull of  $(\Theta)_{\mathcal{B}}^2$ .

The risk of a diagonal estimator is larger than the risk of an oracle attenuation (11.30). As a result, the oracle risk (11.31) gives a lower bound of the minimax linear diagonal risk  $r_{l,d}(\Theta)$ :

$$r_{l,d}(\Theta) \geq r_{\text{inf}}(\Theta) = \sup_{f \in \Theta} \sum_{m=0}^{N-1} \frac{\sigma^2 |f_{\mathcal{B}}[m]|^2}{\sigma^2 + |f_{\mathcal{B}}[m]|^2}. \quad (11.114)$$

Theorem 11.11 proves that this inequality is an equality if  $\Theta$  is quadratically convex.

**Theorem 11.11.** If  $\Theta$  is a bounded and closed set, then there exists a worst signal  $h \in QH[\Theta]$  such that  $r_{\inf}(h) = r_{\inf}(QH[\Theta])$  in the basis  $\mathcal{B}$ . Moreover, the linear diagonal operator  $D$  defined by

$$a_m = \frac{|h_{\mathcal{B}}[m]|^2}{\sigma^2 + |h_{\mathcal{B}}[m]|^2} \quad (11.115)$$

achieves the linear diagonal minimax risk

$$r(D, \Theta) = r_{l,d}(\Theta) = r_{\inf}(QH[\Theta]). \quad (11.116)$$

**Proof.** The risk  $r(D, f)$  of the diagonal operator (11.111) is

$$r(D, f) = \sum_{m=0}^{N-1} \left( \sigma^2 |a_m|^2 + |1 - a_m|^2 |f_{\mathcal{B}}[m]|^2 \right). \quad (11.117)$$

Since it is a linear function of  $|f_{\mathcal{B}}[m]|^2$ , it reaches the same maximum in  $\Theta$  and in  $QH[\Theta]$ . This proves that  $r(D, \Theta) = r(D, QH[\Theta])$  and thus that  $r_{l,d}(\Theta) = r_{l,d}(QH[\Theta])$ .

To verify that  $r_{l,d}(\Theta) = r_{\inf}(QH[\Theta])$ , we prove that  $r_{l,d}(QH[\Theta]) = r_{\inf}(QH[\Theta])$ . Since (11.114) shows that  $r_{\inf}(QH[\Theta]) \leq r_{l,d}(QH[\Theta])$  to get the reverse inequality, it is sufficient to prove that the linear estimator defined by (11.115) satisfies  $r(D, QH[\Theta]) \leq r_{\inf}(QH[\Theta])$ . Since  $\Theta$  is bounded and closed,  $QH[\Theta]$  is also bounded and closed and thus compact, which guarantees the existence of  $h \in QH[\Theta]$  such that  $r_{\inf}(h) = r_{\inf}(QH[\Theta])$ . The risk of this estimator is calculated with (11.117):

$$\begin{aligned} r(D, f) &= \sum_{m=0}^{N-1} \frac{|f_{\mathcal{B}}[m]|^2 \sigma^4 + \sigma^2 |h_{\mathcal{B}}[m]|^4}{(\sigma^2 + |h_{\mathcal{B}}[m]|^2)^2} \\ &= \sum_{m=0}^{N-1} \frac{\sigma^2 |h_{\mathcal{B}}[m]|^2}{\sigma^2 + |h_{\mathcal{B}}[m]|^2} + \sigma^4 \sum_{m=0}^{N-1} \frac{|f_{\mathcal{B}}[m]|^2 - |h_{\mathcal{B}}[m]|^2}{(\sigma^2 + |h_{\mathcal{B}}[m]|^2)^2}. \end{aligned}$$

To show that  $r(D, f) \leq r_{\inf}(QH[\Theta])$ , we verify that the second summation is negative. Let  $0 \leq \eta \leq 1$  and  $y$  be a vector with decomposition coefficients in  $\mathcal{B}$  satisfying

$$|y_{\mathcal{B}}[m]|^2 = (1 - \eta) |h_{\mathcal{B}}[m]|^2 + \eta |f_{\mathcal{B}}[m]|^2.$$

Since  $QH[\Theta]$  is quadratically convex, necessarily  $y \in QH[\Theta]$ , so

$$J(\eta) = \sum_{m=0}^{N-1} \frac{\sigma^2 |y_{\mathcal{B}}[m]|^2}{\sigma^2 + |y_{\mathcal{B}}[m]|^2} \leq \sum_{m=0}^{N-1} \frac{\sigma^2 |h_{\mathcal{B}}[m]|^2}{\sigma^2 + |h_{\mathcal{B}}[m]|^2} = J(0).$$

Since the maximum of  $J(\eta)$  is at  $\eta = 0$ ,

$$J'(0) = \sum_{m=0}^{N-1} \frac{|f_{\mathcal{B}}[m]|^2 - |h_{\mathcal{B}}[m]|^2}{(\sigma^2 + |h_{\mathcal{B}}[m]|^2)^2} \leq 0,$$

which finishes the proof. ■

The worse signal  $h$ , which maximizes  $r_{\inf}(h)$  in  $QH[\Theta]$ , is a signal with coefficients  $|h_{\mathcal{B}}[m]|$  that have a slow decay. It yields conservative amplification factors  $a_m$  that keep many coefficients. This theorem implies that  $r_{l,d}(\Theta) = r_{l,d}(QH[\Theta])$ . To take advantage of the fact that  $\Theta$  may be much smaller than its quadratic convex hull, it is thus necessary to use nonlinear diagonal estimators.

### Translation-Invariant Set

Signals such as sounds or images are often arbitrarily translated in time or in space, depending on the beginning of the recording or the position of the camera. To simplify border effects, we consider signals of period  $N$ . We say that  $\Theta$  is *circular translation invariant* if for any  $f[n] \in \Theta$ , then  $f[(n-p)\text{mod}N] \in \Theta$  for all  $0 \leq p < N$ .

If the set is translation invariant and the noise is stationary, then we show that the best linear estimator is also translation invariant, which means that it is a convolution. Such an operator is diagonal in the discrete Fourier basis  $\mathcal{B} = \{g_m[n] = N^{-1/2} \exp(i2\pi mn/N)\}_{0 \leq m < N}$ . The decomposition coefficients of  $f$  in this basis are proportional to its discrete Fourier transform:

$$f_{\mathcal{B}}[m] = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} f[n] \exp\left(\frac{-i2\pi mn}{N}\right) = \frac{\hat{f}[m]}{\sqrt{N}}.$$

For a set  $\Theta$ , the lower bound  $r_{\inf}(\Theta)$  in (11.114) becomes

$$r_{\inf}(\Theta) = \sup_{f \in \Theta} \sum_{m=0}^{N-1} \frac{\sigma^2 N^{-1} |\hat{f}[m]|^2}{\sigma^2 + N^{-1} |\hat{f}[m]|^2}. \quad (11.118)$$

Theorem 11.12 proves that diagonal operators in the discrete Fourier basis achieve the linear minimax risk.

**Theorem 11.12.** Let  $\Theta$  be a closed and bounded set. Let  $h \in QH[\Theta]$  be such that  $r_{\inf}(h) = r_{\inf}(QH[\Theta])$  and

$$\hat{h}_0[m] = \frac{|\hat{h}[m]|^2}{N \sigma^2 + |\hat{h}[m]|^2}. \quad (11.119)$$

If  $\Theta$  is circular translation invariant, then  $\tilde{F} = DX = X \otimes h_0$  achieves the linear minimax risk

$$r_l(\Theta) = r(D, \Theta) = r_{\inf}(QH[\Theta]). \quad (11.120)$$

**Proof.** Since  $r_l(\Theta) \leq r_{l,d}(\Theta)$ , Theorem 11.11 proves in (11.116) that

$$r_l(\Theta) \leq r_{\inf}(QH[\Theta]).$$

Moreover, the risk  $r_{\inf}(QH[\Theta])$  is achieved by the diagonal estimator (11.115). In the discrete Fourier basis it corresponds to a circular convolution with a transfer function given by (11.119).

We show that  $r_l(\Theta) \geq r_{\inf}(QH[\Theta])$  by using particular Bayes priors. If  $f \in QH[\Theta]$ , then there exists a family  $\{f_i\}_i$  of elements in  $\Theta$  such that for any  $0 \leq m < N$ ,

$$|\hat{f}[m]|^2 = \sum_i p_i |\hat{f}_i[m]|^2 \quad \text{with} \quad \sum_i p_i = 1.$$

To each  $f_i \in \Theta$  we associate a random shift vector  $F_i[n] = f_i[n - Q_i]$  as in (9.27). Each  $F_i[n]$  is circular stationary, and its power spectrum is computed in (9.29):  $\hat{R}_{F_i}[m] = N^{-1} |\hat{f}_i[m]|^2$ . Let  $F$  be a random vector that has a probability  $p_i$  to be equal to  $F_i$ . It

is circular stationary and its power spectrum is  $\hat{R}_F[m] = N^{-1}|\hat{f}[m]|^2$ . We denote by  $\pi_f$  the probability distribution of  $F$ . The risk  $r_l(\pi_f)$  of the Wiener filter is calculated in (11.15):

$$r_l(\pi_f) = \sum_{m=0}^{N-1} \frac{\hat{R}_F[m] \hat{R}_W[m]}{\hat{R}_F[m] + \hat{R}_W[m]} = \sum_{m=0}^{N-1} \frac{N^{-1} |\hat{f}[m]|^2 \sigma^2}{N^{-1} |\hat{f}[m]|^2 + \sigma^2}. \quad (11.121)$$

Since  $\Theta$  is translation invariant, the realizations of  $F$  are in  $\Theta$ , so  $\pi_f \in \Theta^*$ . The minimax theorem (11.4) proves in (11.21) that  $r_l(\pi_f) \leq r_l(\Theta)$ . Since this is true for any  $f \in QH[\Theta]$ , taking a supremum with respect to  $f$  in (11.121) proves that  $r_l(QH[\Theta]) \leq r_l(\Theta)$ , which finishes the proof. ■

### 11.5.2 Thresholding Optimality over Orthosymmetric Sets

We study geometrical conditions on  $\Theta$  to nearly reach the linear minimax risk  $r_l(\Theta)$  and the nonlinear minimax risk  $r_n(\Theta)$  with diagonal estimators in a basis  $\mathcal{B} = \{g_m\}_{0 \leq m < N}$ . Since the oracle attenuation (11.30) yields a smaller risk than any linear or nonlinear diagonal estimator, the maximum risk on  $\Theta$  of any diagonal estimator has a lower bound calculated with the oracle risk (11.31):

$$r_{\inf}(\Theta) = \sup_{f \in \Theta} \sum_{m=0}^{N-1} \frac{\sigma^2 |f_{\mathcal{B}}[m]|^2}{\sigma^2 + |f_{\mathcal{B}}[m]|^2}.$$

Theorem 11.7 proves that thresholding estimators have a risk that is close to this oracle lower bound. Thus, we need to understand under what conditions  $r_n(\Theta)$  is on the order of  $r_{\inf}(\Theta)$ , and compare it with  $r_l(\Theta)$ .

#### *Hyperrectangle*

We first consider hyperrectangles, which are building blocks for computing the minimax risk over any set  $\Theta$ . A hyperrectangle in  $\mathcal{B}$

$$\mathcal{R}_h = \{f : |f_{\mathcal{B}}[m]| \leq |h_{\mathcal{B}}[m]| \quad \text{for } 0 \leq m < N\} \quad (11.122)$$

is a separable set along the basis directions  $g_m$ . The risk lower bound for diagonal estimators is

$$r_{\inf}(\mathcal{R}_h) = \sum_{m=0}^{N-1} \frac{\sigma^2 |h_{\mathcal{B}}[m]|^2}{\sigma^2 + |h_{\mathcal{B}}[m]|^2}.$$

Theorem 11.13 proves that for a hyperrectangle, the nonlinear minimax risk is very close to the linear minimax risk.

**Theorem 11.13.** On a hyperrectangle  $\mathcal{R}_h$  the linear and nonlinear minimax risks are reached by diagonal estimators. They satisfy

$$r_l(\mathcal{R}_h) = r_{\inf}(\mathcal{R}_h), \quad (11.123)$$

and

$$\mu r_{\inf}(\mathcal{R}_h) \leq r_n(\mathcal{R}_h) \leq r_{\inf}(\mathcal{R}_h) \quad \text{with } \mu \leq 1/1.25. \quad (11.124)$$

**Proof.** We first show that a linear minimax estimator is necessarily diagonal in  $\mathcal{B}$ . Let  $\tilde{F} = DX$  be the estimator obtained with a linear operator  $D$  represented by the matrix  $A = (a_{m,n})_{0 \leq n, m \leq N}$  in  $\mathcal{B}$ :

$$\tilde{F}_{\mathcal{B}} = A X_{\mathcal{B}}.$$

Let  $\text{tr}A$  be the trace of  $A$ , and  $A^*$  be its complex transpose. Since  $X = f + W$  where  $W$  is a white noise of variance  $\sigma^2$ , a direct calculation shows that

$$r(D, f) = E\{\|\tilde{F} - f\|^2\} = \sigma^2 \text{tr} AA^* + (Af_{\mathcal{B}} - f_{\mathcal{B}})^* (Af_{\mathcal{B}} - f_{\mathcal{B}}). \quad (11.125)$$

If  $D_d$  is the diagonal operator with coefficients that are  $a_m = a_{m,m}$ , the risk is then

$$r(D_d, f) = \sum_{m=0}^{N-1} \left( \sigma^2 |a_{m,m}|^2 + |1 - a_{m,m}|^2 |f_{\mathcal{B}}[m]|^2 \right). \quad (11.126)$$

To prove that the maximum risk over  $\mathcal{R}_h$  is minimized when  $A$  is diagonal, we show that  $r(D_d, \mathcal{R}_h) \leq r(D, \mathcal{R}_h)$ . For this purpose, we use a prior probability distribution  $\pi \in \mathcal{R}_h^*$  corresponding to a random vector  $F$  with realizations that are in  $\mathcal{R}_h$ :

$$F_{\mathcal{B}}[m] = S[m] h_{\mathcal{B}}[m]. \quad (11.127)$$

The random variables  $S[m]$  are independent and equal to 1 or  $-1$  with probability  $1/2$ . The expected risk  $r(D, \pi) = E\{\|F - \tilde{F}\|^2\}$  is derived from (11.125) by replacing  $f$  by  $F$  and taking the expected value with respect to the probability distribution  $\pi$  of  $F$ . If  $m \neq p$ , then  $E\{F_{\mathcal{B}}[m] F_{\mathcal{B}}[p]\} = 0$ , so we get

$$\begin{aligned} r(D, \pi) &= \sigma^2 \sum_{m=0}^{N-1} |a_{m,m}|^2 + \sum_{m=0}^{N-1} |h_{\mathcal{B}}[m]|^2 \left[ |a_{m,m} - 1|^2 + \sum_{\substack{p=0 \\ p \neq m}}^{N-1} |a_{m,p}|^2 \right] \\ &\geq \sigma^2 \sum_{m=0}^{N-1} |a_{m,m}|^2 + \sum_{m=0}^{N-1} |1 - a_{m,m}|^2 |h_{\mathcal{B}}[m]|^2 = r(D_d, h). \end{aligned} \quad (11.128)$$

Since the realizations of  $F$  are in  $\mathcal{R}_h$ , (11.122) implies that  $r(D, \mathcal{R}_h) \geq r(D, \pi)$ , so  $r(D, \mathcal{R}_h) \geq r(D_d, h)$ . To prove that  $r(D, \mathcal{R}_h) \geq r(D_d, \mathcal{R}_h)$ , it is now sufficient to verify that  $r(D_d, \mathcal{R}_h) = r(D_d, h)$ . To minimize  $r(D_d, f)$ , (11.126) proves necessarily that  $a_{m,m} \in [0, 1]$ . In this case, (11.126) implies

$$r(D_d, \mathcal{R}_h) = \sup_{f \in \mathcal{R}_h} r(D_d, f) = r(D_d, h).$$

Now that we know that the minimax risk is achieved by a diagonal operator, we apply Theorem 11.11, which proves in (11.116) that the minimax risk among linear diagonal operators is  $r_{\inf}(\mathcal{R}_h)$  because  $\mathcal{R}_h$  is quadratically convex. So,  $r_l(\mathcal{R}_h) = r_{\inf}(\mathcal{R}_h)$ .

To prove that the nonlinear minimax risk is also obtained with a diagonal operator, we use the minimax Theorem 11.4, which proves that

$$r_n(\mathcal{R}_h) = \sup_{\pi \in \mathcal{R}_h^*} \inf_{D \in \mathcal{O}_n} r(D, \pi). \quad (11.129)$$

The set  $\mathcal{R}_h$  can be written as a product of intervals along each direction  $g_m$ . As a consequence, to any prior  $\pi \in \mathcal{R}_h^*$  corresponding to a random vector  $F$ , we associate a prior  $\pi' \in \mathcal{R}_h^*$  corresponding to  $F'$  such that  $F'_B[m]$  has the same distribution as  $F_B[m]$  but with  $F'_B[m]$  independent from  $F'_B[p]$  for  $p \neq m$ . We then verify that for any operator  $D$ ,  $r(D, \pi) \leq r(D, \pi')$ . The supremum over  $\mathcal{R}_h^*$  in (11.129) can thus be restricted to processes that have independent coordinates. This independence also implies that the Bayes estimator that minimizes  $r(D, \pi)$  is diagonal in  $\mathcal{B}$ . The minimax theorem (11.14) proves that the minimax risk is reached by diagonal estimators.

Since  $r_n(\mathcal{R}_h) \leq r_l(\mathcal{R}_h)$ , we derive the upper bound in (11.124) from the fact that  $r_l(\mathcal{R}_h) = \mathcal{R}_{\inf}(\mathcal{R}_h)$ . The lower bound (11.124) is obtained by computing the Bayes risk  $r_n(\pi) = \inf_{D \in \mathcal{O}_n} r(D, \pi)$  for the prior  $\pi$  corresponding to  $F$  defined in (11.127), and verifying that  $r_n(\pi) \geq \mu r_{\inf}(\mathcal{R}_h)$ . We see from (11.129) that  $r_n(\mathcal{R}_h) \geq r_n(\pi)$ , which implies (11.124). ■

The bound  $\mu > 0$  was proved by Ibragimov and Khas'minskii [311], but the essentially sharp bound  $1/1.25$  was obtained by Donoho, Liu, and MacGibbon [193]. They showed that  $\mu$  depends on the variance  $\sigma^2$  of the noise, and that if  $\sigma^2$  tends to 0 or to  $+\infty$ , then  $\mu$  tends to 1. For hyperrectangles, linear estimators are thus asymptotically optimal compared to nonlinear estimators.

### Orthosymmetric Sets

To differentiate the properties of linear and nonlinear estimators, we consider more complex sets that can be written as unions of hyperrectangles. We say that  $\Theta$  is *orthosymmetric* in  $\mathcal{B}$  if for any  $f \in \Theta$  and for any  $a_m$  with  $|a_m| \leq 1$ , then

$$\sum_{m=0}^{N-1} a_m f_B[m] g_m \in \Theta.$$

Such a set can be written as a union of hyperrectangles:

$$\Theta = \bigcup_{f \in \Theta} \mathcal{R}_f. \quad (11.130)$$

An upper bound of  $r_n(\Theta)$  is obtained with the maximum risk  $r_{\text{th}}(\Theta) = \sup_{f \in \Theta} r_{\text{th}}(f)$  of a hard- or soft-thresholding estimator in the basis  $\mathcal{B}$ , with a threshold  $T = \sigma \sqrt{2 \log_e N}$ .

**Theorem 11.14.** If  $\Theta$  is orthosymmetric in  $\mathcal{B}$ , then the linear minimax estimator is reached by linear diagonal estimators and

$$r_{l,d}(\Theta) = r_l(\Theta) = r_{\inf}(\text{QH}[\Theta]). \quad (11.131)$$

The nonlinear minimax risk satisfies

$$\frac{1}{1.25} r_{\inf}(\Theta) \leq r_n(\Theta) \leq r_{\text{th}}(\Theta) \leq (2 \log_e N + 1) (\sigma^2 + r_{\inf}(\Theta)). \quad (11.132)$$

**Proof.** Since  $\Theta$  is orthosymmetric,  $\Theta = \bigcup_{f \in \Theta} \mathcal{R}_f$ . On each hyperrectangle  $\mathcal{R}_f$ , we showed in (11.128) that the maximum risk of a linear estimator is reduced by letting it be

diagonal in  $\mathcal{B}$ . The minimax linear estimation in  $\Theta$  is therefore diagonal:  $r_l(\Theta) = r_{l,d}(\Theta)$ . Theorem 11.11 proves in (11.116) that  $r_{l,d}(\Theta) = r_{\inf}(\text{QH}[\Theta])$ , which implies (11.131).

Since  $\Theta = \bigcup_{f \in \Theta} \mathcal{R}_f$ , we also derive that  $r_n(\Theta) \geq \sup_{f \in \Theta} r_n(\mathcal{R}_f)$ . So (11.124) implies that

$$r_n(\Theta) \geq \frac{1}{1.25} r_{\inf}(\Theta).$$

Theorem 11.7 proves in (11.51) that the thresholding risk satisfies

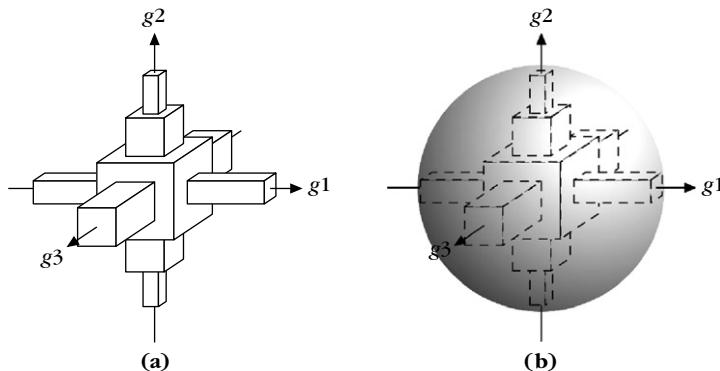
$$r_{\text{th}}(f) \leq (2 \log_e N + 1) (\sigma^2 + r_{\text{pr}}(f)).$$

A modification of the proof shows that this upper bound remains valid if  $r_{\text{pr}}(f)$  is replaced by  $r_{\inf}(f)$  [221]. Taking a supremum over all  $f \in \Theta$  proves the upper bound (11.132), given that  $r_n(\Theta) \leq r_{\text{th}}(\Theta)$ . ■

This theorem shows that  $r_n(\Theta)$  always remains within a factor  $2 \log_e N$  of the lower bound  $r_{\inf}(\Theta)$  and that the thresholding risk  $r_{\text{th}}(\Theta)$  is at most  $2 \log_e N$  times larger than  $r_n(\Theta)$ . In some cases, the factor  $2 \log_e N$  can even be reduced to a constant independent of  $N$ .

Unlike the nonlinear risk  $r_n(\Theta)$ , the linear minimax risk  $r_l(\Theta)$  may be much larger than  $r_{\inf}(\Theta)$ . This depends on the convexity of  $\Theta$ . If  $\Theta$  is quadratically convex, then  $\Theta = \text{QH}[\Theta]$ , so (11.131) implies that  $r_l(\Theta) = r_{\inf}(\Theta)$ . Since  $r_n(\Theta) \geq r_{\inf}(\Theta)/1.25$ , the risk of linear and nonlinear minimax estimators are of the same order. In this case, there is no reason for working with nonlinear as opposed to linear estimators. When  $\Theta$  is an orthosymmetric ellipsoid, Exercise 11.14 computes the minimax linear estimator of Pinsker [400] and the resulting risk.

If  $\Theta$  is not quadratically convex, then its hull  $\text{QH}[\Theta]$  may be much bigger than  $\Theta$ . This is the case when  $\Theta$  has a star shape that is elongated in the directions of the basis vectors  $g_m$ , as illustrated in Figure 11.12. The linear risk  $r_l(\Theta) = r_{\inf}(\text{QH}[\Theta])$



**FIGURE 11.12**

(a) Example of orthosymmetric set  $\Theta$  in three dimensions. (b) The quadratically convex hull  $\text{QH}[\Theta]$  is a larger ellipsoid including  $\Theta$ .

may then be much larger than  $r_{\inf}(\Theta)$ . Since  $r_n(\Theta)$  and  $r_{\text{th}}(\Theta)$  are on the order of  $r_{\inf}(\Theta)$ , they are then much smaller than  $r_l(\Theta)$ . A thresholding estimator thus brings an important improvement over any linear estimator.

---

**EXAMPLE 11.2**

Let  $\Theta$  be an  $\ell^p$  ball defined by

$$\Theta = \{f : \sum_{m=0}^{N-1} \beta_m^p |f_{\mathcal{B}}[m]|^p \leq C^p\}. \quad (11.133)$$

It is an orthosymmetric set. Its square is

$$(\Theta)_{\mathcal{B}}^2 = \{f : \sum_{m=0}^{N-1} \beta_m^p |f_{\mathcal{B}}[m]|^{p/2} \leq C^p\}.$$

If  $p \geq 2$ , then  $(\Theta)_{\mathcal{B}}^2$  is convex, so  $\Theta$  is quadratically convex. If  $p < 2$ , the convex hull of  $(\Theta)_{\mathcal{B}}^2$  is  $\{f : \sum_{m=0}^{N-1} \beta_m^2 |f_{\mathcal{B}}[m]|^2 \leq C^2\}$ , so the quadratic convex hull of  $\Theta$  is

$$\text{QH}[\Theta] = \{f : \sum_{m=0}^{N-1} \beta_m^2 |f_{\mathcal{B}}[m]|^2 \leq C^2\}. \quad (11.134)$$

The smaller  $p$ , the larger the difference between  $\Theta$  and  $\text{QH}[\Theta]$ .

---

### Risk Calculation

Let us consider the maximum linear and nonlinear approximation errors on  $\Theta$ , with  $M$  vectors selected from the basis  $\mathcal{B}$ :

$$\varepsilon_l(M, \Theta) = \sup_{f \in \Theta} \varepsilon_l(M, f) \quad \text{and} \quad \varepsilon_N(M, \Theta) = \sup_{f \in \Theta} \varepsilon_N(M, f).$$

Let  $r(D_M, \Theta)$  be the risk associated to a linear diagonal projector

$$\tilde{F} = D_M X = \sum_{m=0}^{M-1} X_{\mathcal{B}}[m] g_m.$$

Theorem 11.15 proves that  $r(D_M, \Theta)$  depends on the linear approximation error  $\varepsilon_l(M, \Theta)$ . Similarly, the thresholding risk, which is of the order of  $r_{\inf}(\Theta)$ , depends on the nonlinear approximation error  $\varepsilon_N(M, \Theta)$ .

**Theorem 11.15.** Let  $s > 1/2$  and  $C$  be such that  $1 \leq C/\sigma \leq N^s$ . If  $\varepsilon_l(M, \Theta) \leq C^2 M^{1-2s}$ , then

$$r(D_{M_0}, \Theta) \leq 3 C^{1/s} \sigma^{2-1/s} \quad \text{with} \quad (C/(2\sigma))^{1/s} \leq M_0 \leq (C/\sigma)^{1/s}. \quad (11.135)$$

$$\text{If } \varepsilon_N(M, \Theta) \leq C^2 M^{1-2s} \quad \text{then} \quad r_{\inf}(\Theta) \leq 3 C^{1/s} \sigma^{2-1/s}. \quad (11.136)$$

In particular, if

$$\Theta_{C,s} = \left\{ f : \left( \sum_{m=0}^{N-1} |f_B[m]|^{1/s} \right)^s \leq C \right\}, \quad (11.137)$$

then  $r_{\inf}(\Theta_{C,s}) \sim C^{1/s} \sigma^{2-1/s}$ .

**Proof.** The property (11.135) is a consequence of (11.43) in Theorem 11.6. Since  $r_{\inf}(f) \leq r_{\text{pr}}(f)$ , we derive (11.136) from (11.38) in Theorem 11.5.

Theorem 9.9 together with Theorem 9.10 prove that any  $f \in \Theta_{C,s}$  satisfies  $\varepsilon_n(M, f) \leq C^2 M^{1-2s}/(2s-1)$ , which implies that  $r_{\inf}(\Theta_{C,s}) = O(C^{1/s} \sigma^{2-1/s})$ . To get a reverse inequality, we consider  $f \in \Theta_{C,s}$  such that  $|f_B[m]| = \sigma$  for  $0 \leq m < \lfloor (C/\sigma)^{1/s} \rfloor$  and  $f_B[m] = 0$  for  $m \geq \lfloor (C/\sigma)^{1/s} \rfloor$ . In this case,

$$r_p(f) = \lfloor (C/\sigma)^{1/s} \rfloor \sigma^2 \sim C^{1/s} \sigma^{2-1/s}.$$

Since  $r_{\inf}(\Theta_{C,s}) \geq r_{\text{pr}}(f)/2$ , it follows that  $r_{\inf}(\Theta_{C,s}) \sim \sigma^{2-1/s} C^{1/s}$ . ■

The hypothesis  $C/\sigma \geq 1$  guarantees that the largest signal coefficient is not dominated by the noise, whereas  $C/\sigma \leq N^s$  indicates that the smallest coefficient has an amplitude smaller than the noise. This is typically the domain of application for noise-removal algorithms. If  $s$  is large, then  $r_{\inf}(\Theta)$  is almost on the order of  $\sigma^2$ . This risk is much smaller than the noise energy  $E\{\|W\|^2\} = N\sigma^2$ , which means that the estimation removes most of the noise.

### 11.5.3 Nearly Minimax with Wavelet Estimation

Analog signal models are defined over functions by characterizing their regularity. Discrete signal models are derived through the discretization process. We consider uniformly regular and piecewise regular signals as well as bounded variation signals and images. The risk obtained by linear and thresholding wavelet estimators is computed with the tools proved in Sections 11.5.1 and 11.5.2, and their optimality is demonstrated by comparing them to the linear and nonlinear minimax risks.

All asymptotic calculations are performed for a Gaussian white noise, with a variance  $\sigma^2$  that decreases to zero. We write  $r_0(\Theta) \sim r_1(\Theta)$  if there exists two constant  $B \geq A > 0$  that do not depend on the parameters of the set  $\Theta$  or on  $\sigma$  such that  $A r_0(\Theta) \leq r_1(\Theta) \leq B r_0(\Theta)$ .

#### Estimation of Discrete Signals and Functions

Let  $\bar{\Theta}$  be a set of functions in  $L^2[0, 1]$ , which defines an analog signal model. A Gaussian white noise model can characterize the random fluctuations of many sensor outputs. The observed noise process is written as

$$dX(dx) = \bar{f}(x) dx + \sigma dW(dx), \quad (11.138)$$

where the noise  $W(dx)$  is a standard Wiener process, and  $\bar{f}(x)$  is the analog signal of interest, which belongs to the functional set  $\bar{\Theta}$ . The acquisition device outputs noisy measurements  $X[n] = \langle X, \bar{\phi}_n \rangle$  where the  $\bar{\phi}_n(x)$  are the sensor responses. We

suppose here that the acquisition device performs a low-pass filtering and uniform sampling, and thus that  $\phi_n(x) = \bar{\phi}_s(ns - x)$  where  $s$  is the sampling interval. The resulting  $N$  noisy measurements can be written as

$$X[n] = f[n] + W[n],$$

where  $f[n]$  is the discrete signal and  $W[n]$  is the discrete noise:

$$f[n] = \bar{f} \star \bar{\phi}_s(ns) = \int \bar{f}(x) \bar{\phi}_s(ns - x) dx \quad \text{and} \quad W[n] = \sigma \int \bar{\phi}_s(ns - x) dW(dx). \quad (11.139)$$

The resulting set of possible discrete signals is

$$\Theta = \{f \in \mathbb{C}^N : f[n] = \bar{f} \star \bar{\phi}_s(ns) \quad \text{with} \quad \bar{f} \in \bar{\Theta}\}.$$

The properties of the Wiener process imply that if  $\{\bar{\phi}_n(x) = \bar{\phi}_s(ns - x)\}_n$  is an orthonormal family of a space  $\mathbf{U}_N$  in  $\mathbf{L}^2[0, 1]$ , then  $W[n]$  is a Gaussian white noise of variance  $\sigma^2$ .

Let us consider a wavelet orthonormal basis of  $\mathbf{L}^2[0, 1]$ . To simplify explanations, we suppose that  $\bar{\phi}_s(ns - x) = \phi_L(x - 2^L n)$  where  $\phi_L$  is the scaling function at a scale  $s = 2^L = N^{-1}$  associated to this wavelet orthonormal basis. As a result,  $\{\bar{\phi}_s(ns - x) = \phi_{L,n}(x)\}_{0 \leq n < N}$  is a scaling orthonormal basis. The discretized signal then corresponds to scaling coefficients  $f[n] = \langle \bar{f}, \phi_{L,n} \rangle$ , and the sampling approximation space  $\mathbf{U}_N$  is a multiresolution approximation space  $\mathbf{V}_L$ . This hypothesis can be relaxed without modifying the theorems, as long as  $\bar{\phi}_s$  has a fast decay and defines a Riesz basis.

Since  $f[n]$  are the decomposition coefficients of  $\bar{f}$  in the orthonormal basis  $\{\phi_{L,n}\}_{0 \leq n < N}$ , from any discrete estimator  $\tilde{F}[n] = DX[n]$  of  $f[n]$ , an analog estimator of  $\bar{f}(x)$  is derived:

$$\tilde{F}(x) = \sum_{n=0}^{2^{-L}-1} \tilde{F}[n] \phi_{L,n}(x). \quad (11.140)$$

The resulting risk is

$$E\{\|\bar{f} - \tilde{F}\|^2\} = E\{\|P_{\mathbf{V}_L} \bar{f} - \tilde{F}\|^2\} + \|\bar{f} - P_{\mathbf{V}_L} \bar{f}\|^2.$$

Since  $f[n]$  and  $\tilde{F}[n]$  are the coefficients of  $P_{\mathbf{V}_L} \bar{f}$  and  $\tilde{F}$  in an orthonormal basis, the  $\mathbf{L}^2$  error norm is equal to the coefficient error norm in  $\mathbb{C}^N$ :  $\|P_{\mathbf{V}_L} \bar{f} - \tilde{F}\|^2 = \|f - \tilde{F}\|^2$ , and thus

$$E\{\|\bar{f} - \tilde{F}\|^2\} = E\{\|f - \tilde{F}\|^2\} + \|\bar{f} - P_{\mathbf{V}_L} \bar{f}\|^2. \quad (11.141)$$

Taking a supremum over  $\bar{f} \in \bar{\Theta}$ , we get

$$r(D, \Theta) \leq r(D, \bar{\Theta}) \leq r(D, \Theta) + \varepsilon_l(N, \bar{\Theta}), \quad (11.142)$$

where  $\varepsilon_l(N, \bar{\Theta})$  is the maximum linear approximation when projecting functions in  $\bar{\Theta}$  over the space  $\mathbf{U}_N = \mathbf{V}_L$  of dimension  $N$ . We typically have  $\varepsilon_l(N, \bar{\Theta}) = O(N^{-\beta})$  for

some  $\beta > 0$ , and since  $r(D, \Theta)$  typically decays like  $\sigma^\gamma$  for some  $\gamma > 0$ , by choosing  $N \sim \sigma^{-\gamma/\beta}$  the linear approximation error is of the same order as the estimation risk, and we get

$$r(D, \Theta) \sim r(D, \bar{\Theta}). \quad (11.143)$$

For a sufficiently large resolution  $N$ , the estimation errors over discrete and analog signals are of the same order.

### **Uniformly Regular Signals**

Models of uniformly regular signals are defined with a bound over their Hölder norm. Section 9.1.3 proves that such functions have optimal linear approximations in a wavelet basis. We derive that linear wavelet estimators have a maximum risk that is nearly minimax among all linear and nonlinear estimators.

The homogeneous Hölder norm of a uniformly Lipschitz  $\alpha$  function  $\bar{f}$  is the infimum  $\|\bar{f}\|_{\tilde{C}^\alpha}$  of all  $K$  that satisfy

$$\forall (t, v) \in [0, 1]^2, \quad |\bar{f}(t) - p_v(t)| \leq K |t - v|^\alpha,$$

where  $p_v(t)$  is a polynomial of degree  $\lfloor \alpha \rfloor$ . A set of uniformly Lipschitz  $\alpha$  functions provides a good model for uniformly regular signals:

$$\bar{\Theta}_\alpha = \{\bar{f} \in \mathbf{L}^2[0, 1] : \|\bar{f}\|_{\tilde{C}^\alpha} \leq C_\alpha\}, \quad (11.144)$$

where  $\alpha$  and  $C_\alpha$  measure this uniform regularity.

Since  $f[n] = \langle \bar{f}, \phi_{L,n} \rangle$ , we derive in Section 7.3.1 that the orthogonal wavelet coefficients and scaling coefficients of  $\bar{f}(x)$  in  $\mathbf{L}^2[0, 1]$  at a scale  $2^j > 2^L$  are the discrete wavelet coefficients of  $f$  in  $\mathbb{C}^N$ :

$$\langle \bar{f}(x), \psi_{j,m}(x) \rangle = \langle f[n], \psi_{j,m}[n] \rangle \quad \text{and} \quad \langle \bar{f}(x), \phi_{j,m}(x) \rangle = \langle f[n], \phi_{j,m}[n] \rangle.$$

Estimating wavelet or scaling coefficients of  $\bar{f}$  at scales  $2^j > 2^L$  is thus equivalent to estimating the coefficients of  $f$  from the noisy observation  $X$ .

A linear wavelet projector over a family of  $2^{-k} < 2^{-L} = N$  scaling functions is defined by

$$\tilde{F}[n] = D_k X[n] = \sum_{m=0}^{2^{-k}-1} \langle X, \phi_{k,m} \rangle \phi_{k,m}[n]. \quad (11.145)$$

It can also be rewritten as a projection over wavelets at scales  $2^j > 2^k$ :

$$\tilde{F}[n] = D_k X[n] = \sum_{j=k+1}^J \sum_{m=0}^{2^{-j}-1} \langle X, \psi_{j,m} \rangle \psi_{j,m} + \sum_{m=0}^{2^{-J}-1} \langle X, \phi_{J,m} \rangle \phi_{J,m}. \quad (11.146)$$

This amounts to setting all wavelet coefficients  $\langle X, \psi_{j,n} \rangle$  at scales  $2^j \geq 2^k$  to zero. An analog estimator  $\bar{F}(x)$  is associated to  $\tilde{F}[n] = D_k X[n]$  with (11.140). Theorem 11.16 proves that if the projection scale  $2^k$  is appropriately adjusted, then

this linear estimator produces a maximum risk over uniformly Lipschitz  $\alpha$  functions, which is nearly minimax among all linear and nonlinear operators. The proof follows a typical approach to compute minimax rates for such estimators. The set  $\Theta_\alpha$  is embedded in hyperrectangles over which calculations can be carried. Asymptotic risk decays are computed when the noise variance  $\sigma^2$  decreases to zero.

**Theorem 11.16.** Over uniformly Lipschitz  $\alpha$  functions, a linear wavelet projector with cut-off scale  $2^k \sim (\sigma/C_\alpha)^{1/(\alpha+1/2)}$  satisfies

$$r(D_k, \bar{\Theta}_\alpha) \sim r_l(\bar{\Theta}_\alpha) \sim r_n(\bar{\Theta}_\alpha) \sim C_\alpha^{1/(\alpha+1/2)} \sigma^{2-1/(\alpha+1/2)}. \quad (11.147)$$

**Proof.** We shall prove that

$$r(D_k, \Theta_\alpha) \sim r_l(\Theta_\alpha) \sim r_n(\Theta_\alpha) \sim C_\alpha^{1/(\alpha+1/2)} \sigma^{2-1/(\alpha+1/2)} \quad (11.148)$$

by showing that  $\Theta_\alpha$  is nearly a hyperrectangle in a wavelet basis. Theorem 9.7 proves that  $\varepsilon_l(N, \bar{\Theta}_\alpha) = O(N^{-2\alpha})$ . For  $N$  sufficiently large, the theorem result (11.147) is derived from (11.148), by verifying that the maximum risk over  $\bar{\Theta}_\alpha$  is of the same order as the maximum risk over  $\Theta_\alpha$ , with the same argument as in (11.143).

Theorem 9.6 proves in (9.22) that there exists  $B \geq A > 0$  such that

$$A \|\tilde{f}\|_{\tilde{C}^\alpha} \leq \sup_{j \geq J, 0 \leq n < 2^{-j}} 2^{-j(\alpha+1/2)} |\langle f, \psi_{j,n} \rangle| \leq B \|\tilde{f}\|_{\tilde{C}^\alpha}. \quad (11.149)$$

Let us define

$$\mathcal{R}_\lambda = \{f \in \mathbb{C}^N : \sup_{j \geq J, 0 \leq n < 2^{-j}} |\langle f, \psi_{j,n} \rangle| \leq \lambda C_\alpha 2^{j(\alpha+1/2)}\}.$$

This set is hyperrectangles in a wavelet basis, as defined in (11.122). We know that  $\langle f, \psi_{j,n} \rangle = \langle \tilde{f}, \psi_{j,n} \rangle$  for  $j > L$ , and that  $f \in \Theta_\alpha$  if and only if  $\tilde{f} \in \bar{\Theta}_\alpha$ , thus  $\|\tilde{f}\|_{\tilde{C}^\alpha} \leq C_\alpha$ . It results from (11.149) that

$$\mathcal{R}_A \subset \Theta_\alpha \subset \mathcal{R}_B.$$

Theorem 11.13 proves in (11.123) that  $r_{\inf}(\mathcal{R}_A)/1.125 \leq r_n(\mathcal{R}_A)$ , so as a consequence of this embedding, the maximum risk over  $\Theta_\alpha$  satisfies

$$r_{\inf}(\mathcal{R}_A)/1.125 \leq r_n(\mathcal{R}_A) \leq r_n(\Theta_\alpha) \leq r_l(\Theta_\alpha) \leq r(D_k, \mathcal{R}_B). \quad (11.150)$$

Let us compute

$$r_{\inf}(\mathcal{R}_A) \leq A^2 \sum_{j=L+1}^J \sum_{m=0}^{2^j} \min(\sigma^2, C_\alpha^2 2^{j(2\alpha+1)}) \sim A^2 \sigma^2 2^{-k} \quad \text{with } 2^k = (\sigma/C_\alpha)^{1/(\alpha+1/2)}. \quad (11.151)$$

Theorem 9.7 proves that the linear approximation error over  $\Theta_\alpha$  satisfies  $\varepsilon_l(M, \Theta_\alpha) = O(C_\alpha^2 M^{-2\alpha})$ . Theorem 11.15 derives in (11.135) for  $s = \alpha + 1/2$  that the linear projector corresponding to  $2^{-k} \sim M_0 \sim (C/\sigma)^{1/(\alpha+1/2)}$  yields a maximum risk that satisfies  $r(D_k, \Theta_\alpha) \sim \sigma^2 (\sigma/C_\alpha)^{-1/(\alpha+1/2)}$ . Inserting this and (11.151) in (11.150) proves (11.148). ■

This theorem proves that for uniformly regular images, it is not worth using sophisticated nonlinear estimators. Linear estimators are optimal and a simple projector in a wavelet basis is nearly optimal. This theorem remains valid in two dimensions for images. For images, Theorem 9.16 proves that over a set  $\overline{\Theta}_\alpha^2$  of uniformly Lipschitz  $\alpha$  images with a homogeneous Hölder norm bounded by  $C_\alpha$ , the linear approximation error has a different decay than in one dimension, and satisfies

$$\varepsilon_l(N, \overline{\Theta}_\alpha^2) = O(C_\alpha^2 N^{-\alpha}).$$

The same proof as in Theorem 11.16 shows that linear estimators remain optimal with a risk that satisfies

$$r(D_k, \overline{\Theta}_\alpha^2) \sim r_l(\overline{\Theta}_\alpha^2) \sim r_n(\overline{\Theta}_\alpha^2) \sim C_\alpha^{2/(\alpha+1)} \sigma^{2-2/(\alpha+1)}, \quad (11.152)$$

with a linear wavelet projector  $D_k$  the cut-off scale of which satisfies  $2^k \sim (\sigma/C_\alpha)^{2/(\alpha+1)}$ .

### Piecewise Regular Signals

When signals are not uniformly regular, linear estimators are not optimal anymore. For piecewise regular signals, optimal nonlinear estimators average the noisy data  $X = f + W$  over domains where  $f$  is regular, but avoid averaging  $X$  across the discontinuities of  $f$ . These adaptive smoothing algorithms require estimating the positions of the discontinuities of  $f$  from  $X$ . A wavelet thresholding algorithm implements a similar adaptive averaging and produces a nearly minimax risk.

A piecewise uniformly Lipschitz  $\alpha$  function is defined as a function with a Hölder norm bounded on consecutive intervals  $[t_k, t_{k+1}]$ , where the  $t_k$  are the locations of at most  $K$  discontinuities:

$$\overline{\Theta}_{\alpha,K} = \{\bar{f} \in \mathbf{L}^2[0, 1] : \exists \{t_k\}_{0 \leq k < K} \in [0, 1]^K \text{ with } \|\bar{f}\|_{\tilde{C}^\alpha([t_k, t_{k+1}])} \leq C_\alpha \text{ for } 0 \leq k < K\}. \quad (11.153)$$

This model uses a standard Hölder norm  $\|f\|_{C^\alpha} = \|f\|_{\tilde{C}^\alpha} + \|f\|_\infty$ , which imposes that  $f$  is uniformly Lipschitz  $\alpha$  and uniformly bounded, so that the amplitudes of all discontinuities are bounded. The discretization of signals  $\bar{f}(x) \in \overline{\Theta}_{\alpha,K}$  defines a discrete set of signals  $f[n] \in \Theta_{\alpha,K}$ . Figure 11.2 shows a piecewise regular signal with  $K = 9$ .

A wavelet thresholding estimator of  $f[n]$  is defined by

$$\tilde{F}[n] = \sum_{j>L,m} \rho_T(\langle X, \psi_{j,m} \rangle) \psi_{j,m}[n]. \quad (11.154)$$

An analog thresholding estimator  $\tilde{F}(x)$  is associated to  $\tilde{F}[n]$  with (11.140). Theorem 11.17 proves that such thresholding estimators yield a nearly minimax risk, and that this risk is almost the same as the minimax risk in (11.147) for functions having no discontinuities. Linear estimators blur singularities and thus produce a much larger risk.

**Theorem 11.17.** Over piecewise Lipschitz  $\alpha$  functions, a linear wavelet projector with cut-off scale  $2^k \sim \sigma/C_\alpha$  satisfies

$$r(D_k, \bar{\Theta}_{\alpha,K}) \sim r_l(\bar{\Theta}_{\alpha,K}) \sim C_\alpha K \sigma. \quad (11.155)$$

For any  $\sigma > 0$ , a thresholding wavelet estimator with  $N \sim (C_\alpha/\sigma)^{2-1/(\alpha+1/2)}$  and  $T = \sigma \sqrt{2 \log_e N}$  satisfies

$$r_{\text{th}}(\bar{\Theta}_\alpha) = O\left(C_\alpha^{1/(\alpha+1/2)} \sigma^{2-1/(\alpha+1/2)} |\log(\sigma/C_\alpha)|\right). \quad (11.156)$$

**Proof.** We shall first show that

$$r(D_k, \Theta_{\alpha,K}) \sim r_l(\Theta_{\alpha,K}) \sim C_\alpha K \sigma. \quad (11.157)$$

Theorem 9.12 proves that  $\varepsilon_l(N, \bar{\Theta}_\alpha) = O(N^{-1})$ . For  $N$  sufficiently large, (11.155) is derived from (11.157) by proving that the maximum risk over  $\bar{\Theta}_{\alpha,K}$  is equivalent to the maximum risk over  $\Theta_{\alpha,K}$ , with the same argument as in (11.143).

One can verify that  $h[n] = C_\alpha N^{-1/2} \mathbf{1}_{[0,N/2]}[n] \in \Theta_{\alpha,K}$  for any  $K \geq 1$  and  $\alpha > 0$ . An upper bound of the linear minimax risk is computed over a smaller translation-invariant set obtained by translating  $h$  modulo  $N$ :

$$\Theta_0 = \left\{ f \in \mathbb{C}^N : \exists p \in [0, N-1] \quad \text{with} \quad f[n] = h[(n-p) \bmod N] \right\} \subset \Theta_{\alpha,K}.$$

It results that  $r_l(\Theta_{\alpha,K}) \geq r_l(\Theta_0)$ . Theorem 11.12 proves that the linear minimax risk over this translation-invariant set is reached by a diagonal operator in the discrete Fourier basis. Any  $f \in \Theta_0$  satisfies  $|\hat{f}[2m]|^2 = |\hat{h}[2m]| = C_\alpha N^{-1/2} |\sin(2\pi m/N)|^{-1}$  and  $|\hat{f}[2m+1]|^2 = |\hat{h}[2m+1]| = 0$ . Since  $\Theta_0$  is included in a hyperrectangle defined by  $\hat{h}$ , we derive from (11.120) and (11.118) that

$$r_l(\Theta_0) = r_{\inf}(QH[\Theta_0]) = \sum_{m=0}^{N/2-1} \frac{\sigma^2 N^{-2} C_\alpha^2 |\sin(2\pi m/N)|^{-2}}{\sigma^2 + N^{-2} C_\alpha^2 |\sin(2\pi m/N)|^{-2}} \sim C_\alpha \sigma.$$

A similar calculation shows that if  $\Theta_0$  is generated by a signal  $h$  having  $K$  discontinuities of amplitude  $C_\alpha$  instead of a single one, then  $r_l(\Theta_0) \sim K C_\alpha \sigma$ .

Theorem 9.12 proves that the linear approximation error over  $\Theta_{\alpha,K}$  satisfies  $\varepsilon_l(M, \Theta_{\alpha,K}) = O(K C_\alpha^2 M^{-1})$ . Theorem 11.15 derives in (11.135) for  $s=1$  that the linear projector corresponding to  $2^{-k} \sim M_0 \sim C/\sigma$  yields a maximum risk that satisfies  $r(D_k, \Theta_{\alpha,K}) \sim C_\alpha \sigma$ . Since  $r_l(\Theta_{\alpha,K}) \geq r_l(\Theta_0) \sim K C_\alpha \sigma$  and  $r(D_k, \Theta_{\alpha,K}) \leq r_l(\Theta_{\alpha,K})$ , we derive (11.157).

Let us now prove the nonlinear minimax risk result (11.156). For  $T = \sigma \sqrt{2 \log_e N}$ , the thresholding risk satisfies

$$r_{\text{th}}(\Theta) \leq (2 \log_e N + 1) (\sigma^2 + r_{\inf}(\Theta)).$$

Moreover, Theorem 9.12 proves that  $\varepsilon_n(M, \Theta_{\alpha,K}) = O(C_\alpha^2 M^{-2\alpha})$ . Thus, we derive from (11.136) in Theorem 11.15 for  $s=\alpha+1/2$  that  $r_{\inf}(\Theta_{\alpha,K}) = O(C_\alpha^{1/(\alpha+1/2)} \sigma^{2\alpha/(\alpha+1/2)})$ , so

$$r_{\text{th}}(\Theta_\alpha) = O\left(\log_e N C_\alpha^{1/(\alpha+1/2)} \sigma^{2-1/(\alpha+1/2)}\right).$$

Theorem 9.12 proves that  $\varepsilon_l(N, \bar{\Theta}_\alpha) = O(C_\alpha^2 N^{-1})$ . We derive (11.156) with the same argument as in (11.143), by having  $\varepsilon_l(N, \bar{\Theta}_\alpha) = O(C_\alpha^{1/(\alpha+1/2)} \sigma^{2-1/(\alpha+1/2)})$ , which is achieved with  $N \sim (C_\alpha/\sigma)^{2-1/(\alpha+1/2)}$  and thus  $|\log N| \sim |\log(\sigma/C_\alpha)|$ . ■

### Bounded Variation Signals

Bounded variation signals may have discontinuities and include piecewise regular signals. However, it defines a more general model that does not impose any uniform regularity between singularities. The total variation of a signal over  $[0, 1]$  is defined by

$$\|\bar{f}\|_V = \int_0^1 |\bar{f}'(x)| dx,$$

for which we derive a set of bounded variation signals

$$\bar{\Theta}_V = \left\{ \bar{f} \in \mathbf{L}^2[0, 1] : \|\bar{f}\|_V \leq C_V \right\}.$$

Theorem 11.18 proves that nonlinear estimators can have much lower risk than linear estimators for bounded variation functions. It also shows that wavelet thresholding estimations nearly reach the nonlinear minimax rate.

**Theorem 11.18:** *Donoho, Johnstone.* Over bounded variation functions, a linear wavelet projector with cut-off scale  $2^k \sim \sigma/C_\alpha$  satisfies

$$r(D_k, \bar{\Theta}_V) \sim r_l(\bar{\Theta}_V) \sim C_V \sigma. \quad (11.158)$$

There exists  $B \geq A > 0$  such that for any  $\sigma > 0$ , a thresholding wavelet estimator for  $N \sim (C_V/\sigma)^{4/3}$  and  $T = \sigma \sqrt{2 \log_e N}$  satisfies

$$A C_V^{2/3} \sigma^{4/3} \leq r_n(\bar{\Theta}_V) \leq r_{\text{th}}(\bar{\Theta}_V) \leq B C_V^{2/3} \sigma^{4/3} |\log(\sigma/C_V)|, \quad (11.159)$$

**Proof.** We first prove that

$$r(D_k, \Theta_V) \sim r_l(\Theta_V) \sim C_V \sigma \quad \text{with} \quad 2^{-k} \sim \sigma/C_\alpha. \quad (11.160)$$

Theorem 9.14 proves that  $r_l(\bar{\Theta}_V) = O(C_V^2 N^{-1})$ . For  $N$  sufficiently large, (11.158) is derived from (11.160) by verifying that the maximum risk over  $\bar{\Theta}_V$  and  $\Theta_V$  is equivalent, with the same argument as in (11.143).

To compute the risk over  $\Theta_V$ , this set is embedded in two sets that are orthosymmetric in the wavelet basis. This embedding is derived from an upper bound and a lower bound of the wavelet coefficients of  $\bar{f}$ . Theorem 9.13 proves that there exists  $A_2 > 0$  and  $B_2 > 0$  such that

$$\|\bar{f}\|_V \leq B_2 \sum_{j=-\infty}^{J+1} \sum_{n=0}^{2^{-j}-1} 2^{-j/2} |\langle \bar{f}, \psi_{j,n} \rangle| \quad (11.161)$$

and

$$\|\bar{f}\|_V \geq A_2 \sup_{j \leq J} \left( \sum_{n=0}^{2^{-j}-1} 2^{-j/2} |\langle \bar{f}, \psi_{j,n} \rangle| \right). \quad (11.162)$$

We know that  $\langle f, \psi_{j,n} \rangle = \langle \bar{f}, \psi_{j,n} \rangle$  for  $j > L$ , and  $f \in \Theta_V$  if and only if  $\bar{f} \in \bar{\Theta}_V$ . Thus, it results from (11.161) and (11.162) that for any  $q \geq L$ ,

$$\Theta_q \subset \Theta_V \subset \Theta_\infty, \quad (11.163)$$

with

$$\Theta_q = \left\{ f \in \mathbb{C}^N : \sum_{m=0}^{2^{-q}-1} 2^{-q/2} |\langle f, \psi_{q,m} \rangle| \leq B_2^{-1} C_V \quad \text{and} \quad \langle f, \psi_{j,m} \rangle = 0 \text{ for } j \neq q \right\}$$

and

$$\Theta_\infty = \left\{ f \in \mathbb{C}^N : \sup_{j \leq J} \left( \sum_{m=0}^{2^{-j}-1} 2^{-j/2} |\langle f, \psi_{j,m} \rangle| \right) \leq A_2^{-1} C_V \right\}.$$

These two sets are orthosymmetric in the wavelet basis because they only depend on the modulus of wavelet coefficients. If  $f$  is in one of these sets, it remains in these sets when reducing the amplitude of its wavelet coefficients.

It results from (11.131) in Theorem 11.14 that

$$r_{\inf}(\text{QH}[\Theta_q]) = r_l(\Theta_q) \leq r_l(\Theta_V) \leq r_l(\Theta_\infty) = r(D_k, \Theta_\infty). \quad (11.164)$$

The proof of (9.50) in Theorem 9.14 proceeds by showing that there exists  $B_3$  such that for all  $f \in \Theta_\infty$ , the linear approximation error satisfies

$$\varepsilon_l(M, f) \leq B_3 \|f\|_V^2 M^{-1}.$$

Theorem 11.15 derives in (11.6) for  $s = 1$  that

$$r(D_k \Theta_\infty) \leq 3B_3 C_V \sigma \quad \text{with} \quad 2^{-k} \sim \sigma / C_\alpha. \quad (11.165)$$

Property (11.134) implies that

$$\text{QH}[\Theta_q] = \left\{ f \in \mathbb{C}^N : \sum_{m=0}^{2^{-q}-1} 2^{-q} |\langle f, \psi_{q,m} \rangle|^2 \leq B_2^{-2} C_V^2 \quad \text{and} \quad \langle f, \psi_{j,m} \rangle = 0 \text{ for } j \neq q \right\}.$$

For  $2^q = B_2 \sigma / C_V$ , if  $\langle f, \psi_{q,m} \rangle = \sigma$  and  $\langle f, \psi_{j,m} \rangle = 0$  for  $j \neq q$ , then  $f \in \text{QH}[\Theta_q]$ , and thus

$$r_{\inf}(\text{QH}[\Theta_q]) \geq r_{\inf}(f) = 2^{-q-1} \sigma^2 = 2^{-1} B_2^{-1} C_V \sigma.$$

Together with (11.164) and (11.165) it implies (11.160).

To prove the nonlinear minimax risk result (11.156), we first show that

$$A C_V^{2/3} \sigma^{4/3} \leq r_n(\Theta_V) \leq r_{\text{th}}(\Theta_V) \leq \log N B C_V^{2/3} \sigma^{4/3}. \quad (11.166)$$

Theorem 11.14 implies that

$$\frac{1}{1.25} r_{\inf}(\Theta_q) \leq r_n(\Theta_q)$$

and

$$r_{\text{th}}(\Theta_\infty) \leq (2 \log_e N + 1) (\sigma^2 + r_{\inf}(\Theta_\infty)).$$

Since  $\Theta_q \subset \Theta_V \subset \Theta_\infty$ , it results that

$$\frac{1}{1.25} r_{\inf}(\Theta_q) \leq r_n(\Theta_V) \leq r_{\text{th}}(\Theta_V) \leq (2 \log_e N + 1) (\sigma^2 + r_{\inf}(\Theta_\infty)). \quad (11.167)$$

Theorem 9.14 proves that there exists  $B_4$  such that for each  $f \in \Theta_\infty$ , the nonlinear approximation error satisfies

$$\varepsilon_n(M, f) \leq B_4 C_V^2 M^{-2}.$$

Applying (11.136) in Theorem 11.15 for  $s = 3/2$  shows that

$$r_{\inf}(\Theta_\infty) \leq 3B_4^{2/3} C_V^{2/3} \sigma^{2-2/3}. \quad (11.168)$$

Inserting this in (11.167) proves the right upper bound of (11.166).

To prove the left lower bound of (11.166), we choose  $2^q = (B_2 \sigma / C_V)^{2/3}$ . If  $\langle f, \psi_{q,m} \rangle = \sigma$  and  $\langle f, \psi_{j,m} \rangle = 0$  for  $j \neq q$ , then  $f \in \Theta_q$ , so

$$r_{\inf}(\Theta_q) \geq r_{\inf}(f) = 2^{-q-1} \sigma^2 = 2^{-1} B_2^{-2/3} C_V^{2/3} \sigma^{4/3}.$$

Inserting this inequality in (11.167) proves the left lower bound of (11.166).

Theorem 9.14 proves that  $\varepsilon_l(\Theta_V) = O(N^{-1})$ . Thus, we derive (11.159) from (11.166) by verifying that the maximum risk over  $\Theta_V$  and  $\bar{\Theta}_V$  is equivalent, with the same argument as in (11.143). Indeed, the linear approximation risk is sufficiently small for  $N \sim (C_V/\sigma)^{4/3}$ , and thus  $|\log N| \sim |\log(\sigma/C_V)|$ . ■

This theorem proves that when the noise variance  $\sigma$  decreases, the nonlinear minimax risk has a faster asymptotic decay than the linear minimax risk and the thresholding risk is asymptotically equivalent to the nonlinear minimax risk up to a  $|\log \sigma|$  factor. The proof shows that the set of bounded variation functions  $\Theta_V$  can be embedded in two sets that are close enough and that are orthosymmetric in a wavelet basis. It computes the linear and nonlinear risk from the linear and nonlinear approximation errors in these orthosymmetric sets. Similar minimax and thresholding risks can also be calculated in balls of any Besov space, introduced in Section 9.2.3, leading to similar near-optimality results [223].

For  $\alpha > 1$ , a piecewise regular signal with  $K$  discontinuities has a total variation that satisfies  $\|f\|_V \leq B K \|f\|_{C^\alpha}$  for some constant  $B > 0$ . The linear minimax rate is the same for piecewise regular signals in (11.155) and for the much larger class of bounded variation signals in (11.158), because the risk is dominated by the error in the neighborhood of singularities. However, the nonlinear minimax rate for a piecewise regular signal in (11.156) decays faster than for bounded variation signals in (11.159), because nonlinear estimators take advantage of the signal regularity between singularities.

### Bounded Variation Images

Images with edges of finite length and no highly irregular textures have level sets of finite average length. Theorem 2.9 proves that this average length is equal to the total image variation defined by

$$\|f\|_V = \int_0^1 \int_0^1 |\vec{\nabla} f(x_1, x_2)| dx_1 dx_2. \quad (11.169)$$

The partial derivatives of  $\bar{\nabla}f$  are understood in the general sense of distributions to include discontinuous functions. Images also have bounded intensity values. A simple and yet quite powerful image model is thus obtained with functions having a bounded total variation and a bounded amplitude:

$$\bar{\Theta}_V = \left\{ \bar{f} \in \mathbf{L}^2[0, 1]^2 : \|f\|_V \leq C_V, \|f\|_\infty \leq C_\infty \right\}.$$

A camera outputs discrete images obtained by a local averaging of the incoming light intensity. According to (11.138), with a white noise model, the noisy image can be written as  $X[n] = f[n] + W[n]$  for  $n = (n_1, n_2)$ , with  $f[n] = \bar{f} \star \phi_s(ns)$  and  $W[n] = \int \bar{\phi}_s(ns - x) dW(dx)$  for  $x = (x_1, x_2)$ .

Let us consider a separable wavelet orthonormal basis of  $\mathbf{L}^2[0, 1]^2$ . Like in one dimension, we suppose that the low-pass filter is a two-dimensional scaling function  $\bar{\phi}_s(x) = \phi_L^2(-x)$  associated to this wavelet orthonormal basis at a scale  $s = 2^L = N^{-1/2}$ . The discrete image obtained at the camera output can thus be considered as scaling coefficients  $f[n] = \langle \bar{f}, \phi_{L,n}^2 \rangle$ , and the resulting wavelet coefficients of  $\bar{f}(x)$  are the discrete wavelet coefficients of  $f[n]$  in  $\mathbb{C}^N$ :

$$\langle \bar{f}(x), \psi_{j,m}^l(x) \rangle = \langle f[n], \psi_{j,m}^l[n] \rangle \quad \text{for } j > L, 2^j m \in [0, 1]^2 \quad \text{and } l = 1, 2, 3.$$

We suppose that  $\phi_L^2(x)$  has a compact support and a finite total variation. Let  $\Theta_V$  be the set of images  $f[n]$  obtained by discretizing analog functions  $\bar{f} \in \bar{\Theta}_V$ . One can verify that  $\Theta_V$  is a set of discrete images having a bounded amplitude and a bounded discrete total variation as defined by (2.70).

Similar to (11.154), a wavelet thresholding estimator  $\tilde{F}[n]$  is computed from the wavelet coefficients of a noisy image  $X$  at scales  $2^j > 2^L$ , from which an analog estimator  $\tilde{F}(x) = \sum_n \tilde{F}[n] \phi_{L,n}^2(x)$  is recovered like in (11.140). Theorem 11.19 proves that this wavelet thresholding estimation of bounded variation images is more efficient than any linear estimation, and yields a risk that is nearly minimax.

**Theorem 11.19:** *Donoho, Johnstone.* Let  $C = C_V + C_\infty$ . Over bounded variation images, a linear wavelet projector with cut-off scale  $2^k \sim (\sigma/C)^{4/3}$  satisfies

$$r(D_k, \bar{\Theta}_V) \sim r_l(\bar{\Theta}_V) \sim C^{4/3} \sigma^{2/3}. \quad (11.170)$$

There exists  $B \geq A > 0$  such that for any  $\sigma > 0$ , a thresholding wavelet estimator for  $N \sim (C/\sigma)^2$  and  $T = \sigma \sqrt{2 \log_e N}$  satisfies

$$A C \sigma \leq r_n(\bar{\Theta}_V) \leq r_{\text{th}}(\bar{\Theta}_V) \leq B C \sigma |\log(\sigma/C)|. \quad (11.171)$$

**Proof.** We first prove

$$r(D_k, \Theta_V) \sim r_l(\Theta_V) \sim C^{4/3} \sigma^{2/3} \quad \text{with } 2^k \sim (\sigma/C)^{4/3}. \quad (11.172)$$

Theorem 9.18 proves in (9.64) that

$$\varepsilon_l(N, \bar{\Theta}_V) = O(C^2 N^{-1/2}) \quad \text{with } C = C_V + C_\infty.$$

For  $N$  sufficiently large, the theorem result (11.170) is derived from (11.172), with the same argument as in (11.143).

The risk over  $\Theta_V$  is calculated by embedding this set in two orthosymmetries in the wavelet basis. Let us define

$$\begin{aligned} \Theta_q = & \left\{ f \in \mathbb{C}^N : \sum_{\substack{2^q m \in [0,1]^2 \\ 1 \leq l \leq 3}} |\langle f, \psi_{q,m}^l \rangle| \leq B_2^{-1} C \quad \text{and} \quad |\langle f, \psi_{q,m}^l \rangle| \leq B_2^{-1} C 2^q \right. \\ & \left. \text{and} \quad \langle f, \psi_{j,m}^l \rangle = 0 \quad \text{for} \quad j \neq q \right\} \end{aligned}$$

and

$$\Theta_\infty = \left\{ f \in \mathbb{C}^N : \sup_{\substack{j \leq J \\ 1 \leq l \leq 3}} \left( \sum_{2^j m \in [0,1]^2} |\langle f, \psi_{j,m}^l \rangle|^2 \right) \leq A_2^{-2} C^2 2^J \right\}.$$

The upper bound (9.61) of Theorem 9.17 implies that there exists  $B_2 > 0$  such that  $\Theta_q \subset \Theta_V$  for any  $q \geq L$  with  $N = 2^{-2L}$ . One can also derive from (9.67) that  $\Theta_V \subset \Theta_\infty$  for some  $0 < A_2 < B_2$ . This embedding implies that

$$r_{\inf}(\text{QH}[\Theta_q]) \leq r_l(\Theta_V) \leq r(D_k, \Theta_\infty). \quad (11.173)$$

The proof of (9.64) in Theorem 9.18 proceeds by showing that there exists  $B_3$  such that the linear approximation error of any  $f \in \Theta_\infty$  satisfies

$$\varepsilon_l(M, f) \leq B_3 \|f\|_V \|f\|_\infty M^{-1/2} \leq B_3 C^2 M^{-1/2}.$$

Theorem 11.15 derives in (11.135) for  $s = 3/4$  that

$$r(D_k, \Theta_\infty) \leq 3B_3^{2/3} C^{4/3} \sigma^{2/3} \quad \text{with} \quad 2^{-k} \sim (C/\sigma)^{4/3}. \quad (11.174)$$

Using the inequality  $\sum_n |c_n|^2 \leq \sup_n |c_n| \sum_n |c_n|$ , one can also verify that

$$\text{QH}[\Theta_q] = \left\{ f \in \mathbb{C}^N : \sum_{2^q m \in [0,1]^2} 2^{-q} |\langle f, \psi_{q,m}^l \rangle|^2 \leq B_2^{-2} C^2 \right\}.$$

For  $2^q = (B_2 \sigma / C)^{2/3}$ , if  $\langle f, \psi_{q,m}^l \rangle = \sigma$  and  $\langle f, \psi_{j,m}^l \rangle = 0$  for  $j \neq q$ , then  $f \in \text{QH}[\Theta_q]$ , so

$$r_{\inf}(\text{QH}[\Theta_q]) \geq r_{\inf}(f) = 2^{-2q-1} \sigma^2 = 2^{-1} B_2^{-4/3} C^{4/3} \sigma^{2/3}.$$

Inserting this result and (11.174) in (11.173) proves (11.172).

To prove the nonlinear minimax risk result (11.171), we first show that

$$A C \sigma \leq r_n(\Theta_V) \leq r_{\text{th}}(\Theta_V) \leq B C \log N \sigma. \quad (11.175)$$

This requires a different embedding of  $\Theta_V$ . Let  $f_B^r[k]$  be the sorted wavelet coefficients in decreasing amplitude order. Theorem 9.17 proves that there exists  $A_4 > 0$  such that

$$\Theta_V \subset \Theta_1^* = \left\{ f \in \mathbb{C}^N : |f_B^r[k]| \leq A_4^{-1} C_V k^{-1} \right\},$$

where  $\Theta_1^*$  is also an orthosymmetric set. Theorem 11.14 implies that

$$\frac{1}{1.25} r_{\inf}(\Theta_q) \leq r_n(\Theta_q)$$

and

$$r_{\text{th}}(\Theta_1^*) \leq (2 \log_e N + 1) (\sigma^2 + r_{\inf}(\Theta_1^*)).$$

Since  $\Theta_q \subset \Theta_V \subset \Theta_1^*$ , it results that

$$\frac{1}{1.25} r_{\inf}(\Theta_q) \leq r_n(\Theta_V) \leq r_{\text{th}}(\Theta_V) \leq (2 \log_e N + 1) (\sigma^2 + r_{\inf}(\Theta_1^*)). \quad (11.176)$$

Theorem 9.17 proves that there exists  $B_4$  such that for each  $f \in \Theta_1^*$ , the nonlinear approximation error satisfies

$$\varepsilon_n(M, f) \leq B_4 C_V^2 M^{-1}.$$

Applying (11.136) in Theorem 11.15 for  $s = 1$  shows that

$$r_{\inf}(\Theta_1^*) \leq 3B_4^{1/2} C_V \sigma. \quad (11.177)$$

Inserting this in (11.176) proves the right upper bound of (11.175).

To prove the left lower bound of (11.175), we choose  $2^q = B_2 \sigma / C$ . If  $\langle f, \psi_{q,m} \rangle = \sigma$  for  $2^{-q}$  indexes  $m$ , and if  $\langle f, \psi_{q,m} \rangle = 0$  for the  $2^{-2q} - 2^{-q}$  others, and if  $\langle f, \psi_{j,m} \rangle = 0$  for  $j \neq q$ , then we verify that  $f \in \Theta_q$  and thus that

$$r_{\inf}(\Theta_q) \geq r_{\inf}(f) = 2^{-q-1} \sigma^2 = 2^{-1} B_2^{-1} C \sigma.$$

Inserting this inequality in (11.176) proves the left lower bound of (11.175).

Theorem 9.18 proves that  $\varepsilon_l(\bar{\Theta}_V) = O(C^2 N^{-1/2})$ . We derive (11.171) from (11.175) with the same argument as in (11.143), by setting  $N \sim (C/\sigma)^2$  and thus  $|\log N| \sim |\log(\sigma/C_V)|$ , so that  $\varepsilon_l(\Theta_V) = O(C\sigma)$ . ■

This theorem proves that wavelet thresholding estimators are nearly minimax over bounded variation images and yield a risk with a decay that is faster than any linear estimator. In two dimensions, the hypothesis that images have a bounded amplitude is important to control linear approximation errors that play a role both for linear and nonlinear estimators.

For images having some geometric regularity, such as the  $C^2$  piecewise regular images in Section 11.3.2, a thresholding estimator in a curvelet frame has a risk with an asymptotic decay that is faster for small  $\sigma$ . Indeed, curvelet frames yield nonlinear approximations with a smaller asymptotic error for such images. This is also valid for bandlet estimations presented in Section 12.2.4.

## 11.6 EXERCISES

- 11.1** <sup>2</sup> *Linear prediction.* Let  $F[n]$  be a zero-mean, wide-sense stationary random vector with covariance  $R_F[k]$ . We predict the future  $F[n+l]$  from past values  $\{F[n-k]\}_{0 \leq k < N}$  with  $\tilde{F}[n+l] = \sum_{k=0}^{N-1} a_k F[n-k]$ .

- (a) Prove that  $r = E\{|F[n+I] - \tilde{F}[n+I]|^2\}$  is minimum if and only if

$$\sum_{k=0}^{N-1} a_k R_F[q-k] = R_F[q+I] \quad \text{for } 0 \leq q < N.$$

Verify that  $r = R_F[0] - \sum_{k=0}^{N-1} a_k R_F[k+I]$  is the resulting minimum risk.

*Hint:* Use Proposition 11.2.

- (b) Suppose that  $R_F[n] = \rho^{|n|}$  with  $|\rho| < 1$ . Compute  $\tilde{F}[n+I]$  and  $r$ .

- 11.2** <sup>1</sup> Let  $X = F + W$  where the signal  $F$  and the noise  $W$  are zero-mean, wide-sense circular stationary random vectors. Let  $\tilde{F}[n] = X \circledast h[n]$  and  $r(D, \pi) = E\{\|F - \tilde{F}\|^2\}$ . The minimum risk  $r_l(\pi)$  is obtained with the Wiener filter (11.14). A frequency selective filter  $h$  has a discrete Fourier transform  $\hat{h}[m]$  that can only take the values 0 or 1. Find the frequency selective filter that minimizes  $r(D, \pi)$ . Prove that  $r_l(\pi) \leq r(D, \pi) \leq 2r_l(\pi)$ .

- 11.3** <sup>2</sup> Let  $\{g_m\}_{0 \leq m < N}$  be an orthonormal basis. We consider the space  $\mathbf{V}_p$  of signals generated by the first  $p$  vectors  $\{g_m\}_{0 \leq m < p}$ . We want to estimate  $f \in \Theta = \mathbf{V}_p$  from  $X = f + W$ , where  $W$  is a white Gaussian noise of variance  $\sigma^2$ .

- (a) Let  $\tilde{F} = DX$  be the orthogonal projection of  $X$  in  $\mathbf{V}_p$ . Prove that the resulting risk is minimax among linear operators:

$$r(D, \Theta) = r_n(\Theta) = p\sigma^2.$$

- (b) Find the linear minimax estimator over the space of discrete polynomial signals of size  $N$  and degree  $d$ . Compute the linear minimax risk.

- 11.4** <sup>1</sup> Let  $|\langle f, g_{m_k} \rangle| \geq |\langle f, g_{m_{k+1}} \rangle|$  for  $k \geq 1$  be the sorted decomposition coefficients of  $f$  in  $\mathcal{B} = \{g_m\}_{0 \leq m < N}$ . We want to estimate  $f$  from  $X = f + W$  where  $W$  is a Gaussian white noise of variance  $\sigma^2$ . If  $|\langle f, g_{m_k} \rangle| = 2^{-k/2}$ , compute the oracle projection risk  $r_p$  in (11.34) as a function of  $\sigma^2$  and  $N$ . Give an upper bound on the risk  $r$  if we threshold at  $T = \sigma\sqrt{2\log_e N}$  the decomposition coefficients of  $X$ . The same question if  $|\langle f, g_{m_k} \rangle| = k^{-1}$ . Explain why the estimation is more precise in one case than in the other.

- 11.5** <sup>3</sup> Let  $\Theta_{d,K}$  be a set of signals that are piecewise polynomial of degree  $q$ , with at most  $K$  discontinuities with  $N$  samples. Let  $X = f + W$  where  $W$  is a Gaussian white noise of variance  $\sigma^2$ .

- (a) Prove that the minimax risk satisfies  $r_n(\Theta_{d,K}) \geq K(d+1)\sigma^2$ .  
 (b) Prove that a thresholding risk in a Daubechies wavelet basis with  $d+1$  vanishing moments satisfies  $r_{\text{th}}(\Theta_{d,K}) = O(K(d+1)(\log_e N)^2\sigma^2)$ .

- 11.6** <sup>2</sup> Let  $F = f[(n-P) \bmod N]$  be the random-shift process (11.17) obtained with a Dirac doublet  $f[n] = \delta[n] - \delta[n-1]$ . We want to estimate  $F$  from  $X = F + W$  where  $W$  is a Gaussian white noise of variance  $\sigma^2 = 4N^{-1}$ .

- (a) Specify the Wiener filter  $\tilde{F}$  and prove that the resulting risk satisfies  $r_l(\pi) = E\{\|F - \tilde{F}\|^2\} \geq 1$ .

- (b) Compare your numerical results with a translation-invariant hard wavelet thresholding. Analyze the similarities between your algorithm that computes  $s(l)$  and the strategy used by the wavelet thresholding to smooth or not to smooth certain parts of the noisy signal.

- 11.12** <sup>3</sup> *Risk of frame thresholding.* Let  $\{\phi_p\}_{0 \leq p < P}$  with  $P \geq N$  be a frame of  $\mathbb{C}^N$  with frame bounds  $B \geq A > 0$ . For  $X = f + W$  where  $W$  is a Gaussian white noise of variance  $\sigma^2$ , prove that the risk of a thresholding estimator (11.68) satisfies  $r_{\text{th}}(f) \geq B^{-1} \sum_{p=0}^{P-1} \min(|\langle f, \phi_p \rangle|^2, \sigma^2)$ .
- 11.13** <sup>3</sup> Let  $r_{\text{th}}(f, T)$  be the risk of an estimator of  $f$  obtained by hard thresholding at  $T$  the decomposition coefficient of  $X = f + W$  in a basis  $\mathcal{B}$ . The noise  $W$  is Gaussian white with a variance  $\sigma^2$ . This risk is estimated by

$$\tilde{r}(X, T) = \sum_{m=0}^{N-1} C(X_{\mathcal{B}}[m]),$$

with

$$C(u) = \begin{cases} u^2 - \sigma^2 & \text{if } u \leq T \\ \sigma^2 & \text{if } u > T \end{cases}.$$

- (a) Justify qualitatively the definition of this estimator as it is done for (11.71) in the case of a soft-thresholding estimator.  
(b) Let  $\phi_\sigma(x) = (2\pi\sigma^2)^{-1/2} \exp(-x^2/(2\sigma^2))$ . With calculations similar to the proof of Theorem 11.9, show that

$$r_{\text{th}}(T) - E\{\tilde{r}(X, T)\} = 2T\sigma^2 \sum_{m=0}^{N-1} [\phi_\sigma(T - f_{\mathcal{B}}[m]) + \phi_\sigma(T + f_{\mathcal{B}}[m])].$$

- (c) Implement an algorithm that finds  $\tilde{T}$  that minimizes  $\tilde{r}(X, T)$ . Study numerically the performance of  $\tilde{T}$  to estimate noisy signals with a hard thresholding in a wavelet basis.

- 11.14** <sup>2</sup> We want to estimate a signal  $f$  that belongs to an ellipsoid

$$\Theta = \left\{ f : \sum_{m=0}^{N-1} \beta_m^2 |f_{\mathcal{B}}[m]|^2 \leq C^2 \right\}$$

from  $X = f + W$ , where  $W$  is a Gaussian white noise of variance  $\sigma^2$ . We denote  $x_+ = \max(x, 0)$ .

- (a) Using Proposition 11.14, prove that the minimax linear risk on  $\Theta$  satisfies

$$r_l(\Theta) = \sigma^2 \sum_{m=0}^{N-1} a[m], \quad (11.179)$$

with  $a[m] = (\frac{\lambda}{\beta_m} - 1)_+$  where  $\lambda$  is a Lagrange multiplier calculated with

$$\sum_{m=0}^{N-1} \beta_m \left( \frac{\lambda}{\beta_m} - 1 \right)_+ = \frac{C^2}{\sigma^2}. \quad (11.180)$$

- (b) By analogy to Sobolev spaces, the set  $\Theta$  of signals having a discrete derivative of order  $s$  with an energy bounded by  $C^2$  is defined from the discrete Fourier transform:

$$\Theta = \{f : \sum_{m=-N/2+1}^{N/2} |m|^{2s} N^{-1} |\hat{f}[m]|^2 \leq C^2\}. \quad (11.181)$$

Show that the minimax linear estimator  $D$  in  $\Theta$  is a circular convolution  $DX = X \circledast h$ . Explain how to compute the transfer function  $\hat{h}[m]$ .

- (c) Show that the minimax linear risk satisfies

$$r_l(\Theta) \sim C^{2/(2s+1)} \sigma^{2-2/(2s+1)}.$$

- 11.15** <sup>3</sup> Let  $h \in \mathbb{C}^N$ , then consider the set of shift signals  $\Theta_h = \{h_p[n] = h[(n-p) \bmod N] \text{ for } 0 \leq p < N\}$ . Let  $X = f + W$  with  $f \in \Theta_h$ .

- (a) Find a linear estimator that is diagonal in the Fourier basis and that yields a minimax risk over  $\Theta_h$ .  
 (b) For  $h[n] = 1_{[0,N/2]}[n]$ , prove that  $r_l(\Theta_h)/(N\sigma^2) \sim N^{-1/2}$  for all  $N > 0$ .

# Sparsity in Redundant Dictionaries

# 12

Complex signals such as audio recordings or images often include structures that are not well represented by few vectors in any single basis. Indeed, small dictionaries such as bases have a limited capability of sparse expression. Natural languages build sparsity from large redundant dictionaries of words, which evolve in time. Biological perception systems also seem to incorporate robust and redundant representations that generate sparse encodings at later stages. Larger dictionaries incorporating more patterns can increase sparsity and thus improve applications to compression, denoising, inverse problems, and pattern recognition.

Finding the set of  $M$  dictionary vectors that approximate a signal with a minimum error is *NP-hard* in redundant dictionaries. Thus, it is necessary to rely on “good” but nonoptimal approximations, obtained with computational algorithms. Several strategies and algorithms are investigated. Best-basis algorithms restrict the approximations to families of orthogonal vectors selected in dictionaries of orthonormal bases. They lead to fast algorithms, illustrated with wavelet packets, local cosine, and bandlet orthonormal bases. To avoid the rigidity of orthogonality, matching pursuits find freedom in greediness. One by one they select the best approximation vectors in the dictionary. But greediness has its own pitfalls. A basis pursuit implements more global optimizations, which enforce sparsity by minimizing the  $\mathbf{1}^1$  norm of decomposition coefficients.

Sparse signal decompositions in redundant dictionaries are applied to noise removal, signal compression, and pattern recognition, and multichannel signals such as color images are studied. Pursuit algorithms can nearly reach optimal  $M$ -term approximations in *incoherent* dictionaries that include vectors that are sufficiently different. Learning and updating dictionaries are studied by optimizing the approximation of signal examples.

---

## 12.1 IDEAL SPARSE PROCESSING IN DICTIONARIES

Computing an optimal  $M$ -term approximation in redundant dictionaries is computationally intractable, but it sets a goal that will guide most of the following sections

and algorithms. The resulting compression algorithms and denoising estimators are described in Sections 12.1.2 and 12.1.3.

### 12.1.1 Best $M$ -Term Approximations

Let  $\mathcal{D} = \{\phi_p\}_{p \in \Gamma}$  be a dictionary of  $P$  unit norm vectors  $\|\phi_p\| = 1$  in a signal space  $\mathbb{C}^N$ . We study sparse approximations of  $f \in \mathbb{C}^N$  with vectors selected in  $\mathcal{D}$ . Let  $\{\phi_p\}_{p \in \Lambda}$  be a subset of vectors in  $\mathcal{D}$ . We denote by  $|\Lambda|$  the cardinal of the index set  $\Lambda$ . The orthogonal projection of  $f$  on the space  $\mathbf{V}_\Lambda$  generated by these vectors is

$$f_\Lambda = \sum_{p \in \Gamma} a[p] \phi_p \quad \text{with } a[p] \neq 0 \quad \text{only for } p \in \Lambda. \quad (12.1)$$

The set  $\Lambda \subset \Gamma$  is called the support of the approximation coefficients  $a[p]$ . Its cardinal  $|\Lambda| = \|a\|_0$  is the  $\mathbf{1}^0$  pseudo-norm giving the number of nonzero coefficients of  $a$ . This support carries geometrical information about  $f$  relative to  $\mathcal{D}$ . In a wavelet basis, it gives the multiscale location of singularities and edges. In a time-frequency dictionary, it provides the location of transients and time-frequency evolution of harmonics.

The best  $M$ -term approximation  $f_\Lambda$  minimizes the approximation error  $\|f - f_\Lambda\|$  with  $|\Lambda| = M$  dictionary vectors. If  $\mathcal{D}$  is an orthonormal basis, then Section 9.2.1 proves that the best approximation vectors are obtained by thresholding the orthogonal signal coefficients at some level  $T$ . This is not valid if  $\mathcal{D}$  is redundant, but Theorem 12.1 proves that a best approximation is still obtained by minimizing an  $\mathbf{1}^0$  Lagrangian where  $T$  appears as a Lagrange multiplier:

$$\mathcal{L}_0(T, f, \Lambda) = \|f - f_\Lambda\|^2 + T^2 |\Lambda| = \|f - \sum_{p \in \Gamma} a[p] \phi_p\|^2 + T^2 \|a\|_0. \quad (12.2)$$

This Lagrangian penalizes the approximation error  $\|f - f_\Lambda\|^2$  by the number of approximation vectors.

**Theorem 12.1.** In a dictionary  $\mathcal{D} = \{\phi_p\}_{p \in \Gamma}$ ,

$$\Lambda_T = \operatorname{argmin}_{\Lambda \subset \Gamma} \mathcal{L}_0(T, f, \Lambda) = \operatorname{argmin}_{\Lambda \subset \Gamma} \|f - f_\Lambda\|^2 + |\Lambda| T^2 \quad (12.3)$$

is a best approximation support, which satisfies for all  $\Lambda \subset \Gamma$ ,

$$\|f - f_{\Lambda_T}\| \leq \|f - f_\Lambda\|^2 \quad \text{if } |\Lambda| \leq |\Lambda_T|. \quad (12.4)$$

$$\text{If } \mathcal{L}_0(T, f, \Lambda_T) \leq C T^{2-1/s} \text{ with } s \geq 1/2, \text{ then } \|f - f_{\Lambda_T}\|^2 \leq C^{2s} |\Lambda_T|^{1-2s}. \quad (12.5)$$

**Proof.** The minimization (12.3) implies that any  $\Lambda \subset \Gamma$  satisfies

$$\|f - f_\Lambda\|^2 + |\Lambda| T^2 \geq \|f - f_{\Lambda_T}\|^2 + |\Lambda_T| T^2.$$

Therefore, if  $|\Lambda| \leq |\Lambda_T|$ , then  $\|f - f_\Lambda\| \geq \|f - f_{\Lambda_T}\|$ , which proves (12.4).

If  $\|f - f_{\Lambda_T}\|^2 + |\Lambda_T|T^2 \leq C T^{2-1/s}$ , then  $|\Lambda_T| \leq C T^{-1/s}$ , so if  $s \geq 1/2$ ,

$$\|f - f_{\Lambda_T}\|^2 \leq C T^{2-1/s} \leq C^{2s} |\Lambda_T|^{1-2s}. \quad \blacksquare$$

This theorem proves in (12.4) that minimizing the  $\mathbf{1}^0$  Lagrangian yields a best approximation  $f_M = f_{\Lambda_T}$  of  $f$  with  $M = |\Lambda_T|$  terms in  $\mathcal{D}$ . The decay of the approximation error is controlled in (12.5) by the Lagrangian decay as a function of  $T$ . If  $\mathcal{D}$  is an orthonormal basis, Theorem 12.2 derives that the resulting approximation is a thresholding at  $T$ .

**Theorem 12.2.** If  $\mathcal{D}$  is an orthonormal basis, then the best approximation support is

$$\Lambda_T = \operatorname{argmin}_{\Lambda \subset \Gamma} \|f - f_{\Lambda}\|^2 + |\Lambda| T^2 = \{p \in \Gamma : |\langle f, \phi_p \rangle| \geq T\} \quad (12.6)$$

and

$$\mathcal{L}_0(T, f, \Lambda_T) = \sum_{p \in \Gamma} \min(|\langle f, \phi_p \rangle|^2, T^2). \quad (12.7)$$

**Proof.** If  $\mathcal{D}$  is an orthonormal basis, then  $f_{\Lambda} = \sum_{p \in \Lambda} \langle f, \phi_p \rangle \phi_p$ , so

$$\begin{aligned} \|f - f_{\Lambda}\|^2 + |\Lambda| T^2 &= \sum_{p \notin \Lambda} |\langle f, \phi_p \rangle|^2 + |\Lambda| T^2 \\ &\geq \sum_{p \in \Gamma} \min(|\langle f, \phi_p \rangle|^2, T^2) = \|f - f_{\Lambda_T}\|^2 + |\Lambda_T| T^2. \end{aligned} \quad \blacksquare$$

### NP-Hard Support Covering

In general, computing the approximation support (12.3) which minimizes the  $\mathbf{1}^0$  Lagrangian is proved by Davis, Mallat, and Avellaneda [201] to be an NP-hard problem. This means that there exists dictionaries where finding this solution belongs to a class of NP-complete problems, for which it has been conjectured for the last 40 years that the solution cannot be found with algorithms of polynomial complexity.

The proof [201] shows that for particular dictionaries, finding a best approximation is equivalent to a set-covering problem, which is known to be NP-hard. Let us consider a simple dictionary  $\mathcal{D} = \{\phi_p\}_{p \in \Lambda}$  with vectors having exactly three nonzero coordinates in an orthonormal basis  $\mathcal{B} = \{g_m\}_{0 \leq m < N}$ ,

$$\phi_p = \sum_{m \in \Omega_p} g_m \quad \text{with} \quad |\Omega_p| = 3.$$

If the sets  $\{\Omega_p\}_{p \in \Lambda}$  define an exact partition of a subset  $\Omega$  of  $\{0, \dots, N-1\}$ , then  $f_{\Omega} = \sum_{m \in \Omega} g_m$  has an exact and optimal dictionary decomposition:

$$f = \sum_{p \in \Lambda} \phi_p \quad \text{with} \quad |\Lambda| = |\Omega|/3.$$

Finding such an exact decomposition for any  $\Omega$ , if it exists, is an NP-hard *three-sets covering problem*. Indeed, the choice of one element in the solution influences the choice of all others, which essentially requires us to try all possibilities. This argument shows that redundancy makes the approximation problem much more complex. In some dictionaries such as orthonormal bases, it is possible to find optimal  $M$ -term approximations with fast algorithms, but these are particular cases.

Since optimal solutions cannot be calculated exactly, it is necessary to find algorithms of reasonable complexity that find “good” if not optimal solutions. Section 12.2 describes the search for optimal solutions restricted to sets of orthogonal vectors in well-structured *tree dictionaries*. Sections 12.3 and 12.4 study pursuit algorithms that search for more flexible and thus nonorthogonal sets of vectors, but that are not always optimal. Pursuit algorithms may yield optimal solutions, if the optimal support  $\Lambda$  satisfies exact recovery properties (studied in Section 12.5).

### 12.1.2 Compression by Support Coding

Chapter 10 describes transform code algorithms that quantize and code signal coefficients in an orthonormal basis. Increasing the dictionary size can reduce the approximation error by offering more choices. However, it also increases the number of bits needed to code which approximation vectors compress a signal. Optimizing the distortion rate is a trade-off between both effects.

We consider a transform code that approximates  $f$  by its orthogonal projection  $f_\Lambda$  on the space  $\mathbf{V}_\Lambda$  generated by the dictionary vectors  $\{\phi_p\}_{p \in \Lambda}$ , and that quantizes the resulting coefficients. The quantization error is reduced by orthogonalizing the family  $\{\phi_p\}_{p \in \Lambda}$ , for example, with a Gram-Schmidt algorithm, which yields an orthonormal basis  $\{g_p\}_{p \in \Lambda}$  of  $\mathbf{V}_\Lambda$ . The orthogonal projection on  $\mathbf{V}_\Lambda$  can then be written as

$$f_\Lambda = \sum_{p \in \Lambda} \langle f, g_p \rangle g_p. \quad (12.8)$$

These coefficients are uniformly quantized with

$$Q(x) = \begin{cases} 0 & \text{if } |x| < \Delta/2, \\ \text{sign}(x) k \Delta & \text{if } (k - 1/2) \Delta \leq |x| < (k + 1/2) \Delta, \end{cases} \quad (12.9)$$

and the signal recovered from quantized coefficients is

$$\tilde{f} = \sum_{p \in \Lambda} Q(\langle f, g_p \rangle) g_p. \quad (12.10)$$

The set  $\Lambda$  is further restricted to coefficients  $|\langle f, g_p \rangle| \geq \Delta/2$ , and thus  $Q(\langle f, g_p \rangle) \neq 0$ , which has no impact on  $\tilde{f}$ .

### Distortion Rate

Let us compute the distortion rate as a function of the dictionary size  $P$ . The compression distortion is decomposed in an approximation error plus a quantization error:

$$\|f - \tilde{f}\|^2 = \|f - f_\Lambda\|^2 + \|f_\Lambda - \tilde{f}\|^2. \quad (12.11)$$

Since  $|x - Q(x)| \leq \Delta/2$ ,

$$\|f_\Lambda - \tilde{f}\|^2 \leq \sum_{p \in \Lambda} |\langle f, g_p \rangle - Q(\langle f, g_p \rangle)|^2 \leq |\Lambda| \frac{\Delta^2}{4}. \quad (12.12)$$

With (12.11), we derive that the coding distortion is smaller than the  $\mathbf{I}^0$  Lagrangian (12.2):

$$d = \|f - \tilde{f}\|^2 \leq \|f - f_\Lambda\|^2 + |\Lambda| T^2 = \mathcal{L}_0(T, f, \Lambda) \quad \text{for } T = \Delta/2. \quad (12.13)$$

This result shows that minimizing the  $\mathbf{I}^0$  Lagrangian reduces the compression distortion.

Having a larger dictionary offers more possibilities to choose  $\Lambda$  and further reduce the Lagrangian. Suppose that some optimization process finds an approximation support  $\Lambda_T$  such that

$$\mathcal{L}_0(T, f, \Lambda_T) \leq C T^{2-1/s}, \quad (12.14)$$

where  $C$  and  $s$  depend on the dictionary design and size. The number  $M$  of nonzero quantized coefficients is  $M = |\Lambda_T| \leq C T^{-1/s}$ . Thus, the distortion rate satisfies

$$d(R, f) = \|f - \tilde{f}\|^2 \leq C^{2s} M^{1-2s}, \quad (12.15)$$

where  $R$  is the total number of bits required to code the quantized coefficients of  $\tilde{f}$  with a variable-length code.

As in Section 10.4.1, the bit budget  $R$  is decomposed into  $R_0$  bits that code the support set  $\Lambda_T \subset \Gamma$ , plus  $R_1$  bits to code the  $M$  nonzero quantized values  $Q(\langle f, g_p \rangle)$  for  $p \in \Lambda$ . Let  $P$  be the dictionary size. We first code  $M = |\Lambda_T| \leq P$  with  $\log_2 P$  bits. There are  $\binom{P}{M}$  subsets of size  $M$  in a set of size  $P$ . Coding  $\Lambda_T$  without any other prior geometric information thus requires  $R_0 = \log_2 \binom{P}{M} \sim M \log_2(P/M)$  bits. As in (10.48), this can be implemented with an entropy coding of the binary significance map

$$\forall p \in \Gamma, b[p] = \begin{cases} 1 & \text{if } p \in \Lambda_T \\ 0 & \text{if } p \notin \Lambda_T. \end{cases} \quad (12.16)$$

The proportion  $p_k$  of quantized coefficients of amplitude  $|Q_\Delta(\langle f, g_p \rangle)| = k\Delta$  typically has a decay of  $p_k = (k^{-1+\varepsilon})$  for  $\varepsilon > 0$ , as in (10.57). We saw in (10.58) that coding the amplitude of the  $M$  nonzero coefficients with a logarithmic variable length  $l_k = \log_2(\pi^2/6) + 2 \log_2 k$ , and coding their sign, requires a total number of

bits  $R_1 \sim M$  bits. For  $M \ll P$ , it results that the total bit budget is dominated by the number of bits  $R_0$  to code the approximation support  $\Lambda_T$ ,

$$R = R_0 + R_1 \sim R_0 \sim M \log_2(P/M),$$

and hence that

$$M \sim R |\log_2(P/R)|^{-1}.$$

For a distortion satisfying (12.15), we get

$$d(R, f) = O\left(C^{2s} R^{1-2s} |\log_2(P/R)|^{2s-1}\right). \quad (12.17)$$

When coding the approximation support  $\Lambda_T$  in a large dictionary of size  $P$  as opposed to an orthonormal basis of size  $N$ , it introduces a factor  $\log_2 P$  in the distortion rate (12.17) instead of the  $\log_2 N$  factor in (10.8). This is worth it only if it is compensated by a reduction of the approximatton constant  $C$  or an increase of the decay exponent  $s$ .

### **Distortion Rate for Analog Signals**

A discrete signal  $f[n]$  is most often obtained with a linear discretization that projects an analog signal  $\tilde{f}(x)$  on an approximation space  $\mathbf{U}_N$  of size  $N$ . This linear approximation error typically decays like  $O(N^{-\beta})$ . From the discrete compressed signal  $\tilde{f}[n]$ , a discrete-to-analog conversion restores an analog approximation  $\tilde{\tilde{f}}_N(x) \in \mathbf{U}_N$  of  $\tilde{f}(x)$ .

Let us choose a discrete resolution  $N \sim R^{(2s-1)/\beta}$ . If the dictionary has a polynomial size  $P = O(N^\gamma)$ , then similar to (10.62), we derive from (12.17) that

$$d(R, \tilde{f}) = \|\tilde{f} - \tilde{\tilde{f}}_N\|^2 = O\left(R^{1-2s} |\log_2 R|^{2s-1}\right). \quad (12.18)$$

Thus, the distortion rate in a dictionary of polynomial size essentially depends on the constant  $C$  and the exponent  $s$  of the  $\mathbf{1}^0$  Lagrangian decay  $\mathcal{L}_0(T, f, \Lambda_T) \leq C T^{2-1/s}$  in (12.14). To optimize the asymptotic distortion rate decay, one must find dictionaries of polynomial sizes that maximize  $s$ . Section 12.2.4 gives an example of a bandlet dictionary providing such optimal approximations for piecewise regular images.

### **12.1.3 Denoising by Support Selection in a Dictionary**

A hard thresholding in an orthonormal basis is an efficient nonlinear projection estimator, if the basis defines a sparse signal approximation. Such estimators can be improved by increasing the dictionary size. A denoising estimator in a redundant dictionary also projects the observed data on a space generated by an optimized set  $\Lambda$  of vectors. Selecting this support is more difficult than for signal approximation or compression because the noise impacts the choice of  $\Lambda$ . The *model selection theory* proves that a nearly optimal set is estimated by minimizing the  $\mathbf{1}^0$  Lagrangian, with an appropriate multiplier  $T$ .

Noisy signal observations are written as

$$X[n] = f[n] + W[n] \quad \text{for } 0 \leq n < N,$$

where  $W[n]$  is a Gaussian white noise of variance  $\sigma^2$ . Let  $\mathcal{D} = \{\phi_p\}_{p \in \Gamma}$  be a dictionary of  $P$ -unit norm vectors. To any subfamily of vectors,  $\{\phi_p\}_{p \in \Lambda}$  corresponds an orthogonal projection estimator on the space  $\mathbf{V}_\Lambda$  generated by these vectors:

$$X_\Lambda = \sum_{p \in \Lambda} a[p] \phi_p.$$

The orthogonal projection in  $\mathbf{V}_\Lambda$  satisfies  $X_\Lambda = f_\Lambda + W_\Lambda$ , so

$$\|f - X_\Lambda\|^2 = \|f - f_\Lambda\|^2 + \|W_\Lambda\|^2. \quad (12.19)$$

The bias term  $\|f - f_\Lambda\|^2$  is the signal approximation error, which decreases when  $|\Lambda|$  increases. On the contrary, the noise energy  $\|W_\Lambda\|^2$  in  $\mathbf{V}_\Lambda$  increases when  $|\Lambda|$  increases. Reducing the risk amounts to finding a projection support  $\Lambda$  that balances these two terms to minimize their sum.

Since  $\mathbf{V}_\Lambda$  is a space of dimension  $|\Lambda|$ , the projection of a white noise of variance  $\sigma^2$  satisfies  $E\{\|W_\Lambda\|^2\} = |\Lambda| \sigma^2$ . However, there are  $2^P$  possible subsets  $\Lambda$  in  $\Gamma$ , and  $\|W_\Lambda\|^2$  may potentially take much larger values than  $|\Lambda| \sigma^2$  for some particular sets  $\Lambda$ . A concentration inequality proved in Lemma 12.1 of Theorem 12.3 shows that for any subset  $\Lambda$  of  $\Gamma$ ,

$$\|W_\Lambda\|^2 \leq (\lambda^2 \sigma^2 \log_e P) |\Lambda| = T^2 |\Lambda| \quad \text{for } T = \lambda \sigma \sqrt{\log_e P},$$

with a probability that tends to 1 as  $P$  increases for  $\lambda$  large enough. It results that the estimation error is bounded by the approximation Lagrangian:

$$\|f - X_\Lambda\|^2 \leq \mathcal{L}_0(T, f, \Lambda) = \|f - f_\Lambda\|^2 + T^2 |\Lambda|. \quad (12.20)$$

However, the set  $\Lambda_T$  that minimizes this Lagrangian,

$$\Lambda_T = \operatorname{argmin}_{\Lambda \subset \Gamma} \left( \|f - f_\Lambda\|^2 + T^2 |\Lambda| \right), \quad (12.21)$$

can only be found by an oracle because it depends on  $f$ , which is unknown. Thus, we need to find an estimator that is nearly as efficient as the oracle projector on this subset  $\Lambda_T$  of vectors.

### **Penalized Empirical Error**

Estimating the oracle set  $\Lambda_T$  in (12.21) requires us to estimate  $\|f - f_\Lambda\|^2$  for any  $\Lambda \subset \Gamma$ . A crude estimator is given by the empirical norm

$$\|X - X_\Lambda\|^2 = \|X\|^2 - \|X_\Lambda\|^2.$$

This may seem naive because it yields a large error,

$$\|X - X_\Lambda\|^2 - \|f - f_\Lambda\|^2 = (\|X\|^2 - \|f\|^2) - (\|X_\Lambda\|^2 - \|f_\Lambda\|^2).$$

Since  $X = f + W$  and  $X_\Lambda = f_\Lambda + W_\Lambda$ , the expected error is

$$E\{\|X\|^2 - \|f\|^2\} - E\{\|X_\Lambda\|^2 - \|f_\Lambda\|^2\} = (N - |\Lambda|)\sigma^2,$$

which is of the order of  $N\sigma^2$  if  $|\Lambda| \ll N$ . However, the first large term does not influence the choice of  $\Lambda$ . The component that depends on  $\Lambda$  is the smaller term  $\|X_\Lambda\|^2 - \|f_\Lambda\|^2$ , which is only of the order  $|\Lambda|\sigma^2$ .

Thus, we estimate  $\|f - f_\Lambda\|^2$  with  $\|X - X_\Lambda\|^2$  in the oracle formula (12.21), and define the best empirical estimation  $X_{\tilde{\Lambda}_T}$  as the orthogonal projection on a space  $\mathbf{V}_{\tilde{\Lambda}_T}$ , where  $\tilde{\Lambda}_T$  minimizes the penalized empirical risk:

$$\tilde{\Lambda}_T = \operatorname{argmin}_{\Lambda \in \Gamma} (\|X - X_\Lambda\|^2 + T^2 |\Lambda|). \quad (12.22)$$

Theorem 12.3 proves that this estimated set  $\tilde{\Lambda}_T$  yields a risk that is within a factor of 4 of the risk obtained by the oracle set  $\Lambda_T$  in (12.21). This theorem is a consequence of the more general model selection theory of Barron, Birgé, and Massart [97], where the optimization of  $\Lambda$  is interpreted as a model selection. Theorem 12.3 was also proved by Donoho and Johnstone [220] for estimating a best basis in a dictionary of orthonormal bases.

**Theorem 12.3:** Barron, Birgé, Massart, Donoho, Johnstone. Let  $\sigma^2$  be the noise variance and  $T = \lambda \sigma \sqrt{\log_e P}$  with  $\lambda \geq \sqrt{32 + \frac{8}{\log_e P}}$ . For any  $f \in \mathbb{C}^N$ , the best empirical set

$$\tilde{\Lambda}_T = \operatorname{argmin}_{\Lambda \subset \Gamma} (\|X - X_\Lambda\|^2 + T^2 |\Lambda|) \quad (12.23)$$

yields a projection estimator  $\tilde{F} = X_{\tilde{\Lambda}_T}$  of  $f$ , which satisfies

$$E[\|\tilde{F} - f\|^2] \leq 4 \min_{\Lambda \subset \Gamma} (\|f - f_\Lambda\|^2 + T^2 |\Lambda|) + \frac{32\sigma^2}{P}. \quad (12.24)$$

**Proof.** Concentration inequalities are at the core of this result. Indeed, the penalty  $T^2 |\Lambda|$  must dominate the random fluctuations of the projected noise. We give a simplified proof provided in [233]. Lemma 12.1 uses a concentration inequality for Gaussian variables to ensure with high probability that the noise energy is simultaneously small in all the subspaces  $\mathbf{V}_\Lambda$  spanned by subsets of vectors in  $\mathcal{D}$ .

**Lemma 12.1.** For any  $u \geq 0$  and any  $\Lambda \subset \Gamma$ ,

$$\sigma^{-1} \|W_\Lambda\| \leq \sqrt{|\Lambda|} + \sqrt{4|\Lambda| \log_e P + 2u}, \quad (12.25)$$

with a probability greater than  $1 - 2e^{-u}/P$ .

This lemma is based on Tsirelson's lemma which proves that for any function  $L$  from  $\mathbb{C}^N$  to  $\mathbb{C}$  that is 1-Lipschitz ( $|L(f) - L(g)| \leq \|f - g\|$ ), and for any normalized Gaussian white noise vector  $W'$  of variance  $\sigma^2 = 1$ ,

$$\operatorname{Proba}\left(L(W') \geq E\{L(W')\} + t\right) \leq e^{-t^2/2}.$$

The orthogonal projection's norm  $L(W) = \|W_\Lambda\|$  is 1-Lipschitz. Applying Tsirelson's lemma to  $W' = \sigma^{-1}W$  for  $t = \sqrt{4|\Lambda| \log_e P + 2u}$  yields

$$\text{Proba} \left\{ \|W'_\Lambda\| \geq E\{\|W'_\Lambda\|\} + \sqrt{4|\Lambda| \log_e P + 2u} \right\} \leq P^{-2|\Lambda|} e^{-u}.$$

Since  $E\{\|W'_\Lambda\|\} \leq (E\{\|W'_\Lambda\|^2\})^{1/2} = \sqrt{|\Lambda|}$ , it results that

$$\text{Proba} \left\{ \|W'_\Lambda\| \geq \sqrt{|\Lambda|} + \sqrt{4|\Lambda| \log_e P + 2u} \right\} \leq P^{-2|\Lambda|} e^{-u}.$$

Let us now compute the probability of the existence of a set  $\tilde{\Lambda}$  that does not satisfy the lemma condition (12.25), by considering each subset of  $\Gamma$ :

$$\begin{aligned} & \text{Proba} \left\{ \exists \tilde{\Lambda} \subset \Gamma, \|W'_{\tilde{\Lambda}}\| \geq \sqrt{|\tilde{\Lambda}|} + \sqrt{4|\tilde{\Lambda}| \log_e P + 2u} \right\} \\ & \leq \sum_{\tilde{\Lambda} \subset \Gamma} \text{Proba} \left\{ \|W'_{\tilde{\Lambda}}\| \geq \sqrt{|\tilde{\Lambda}|} + \sqrt{4|\tilde{\Lambda}| \log_e P + 2u} \right\} \\ & \leq \sum_{\tilde{\Lambda} \subset \Gamma} P^{-2|\tilde{\Lambda}|} e^{-u} \leq \sum_{n=1}^P \binom{P}{n} P^{-2n} e^{-u} \\ & \leq \sum_{n=1}^P P^{-n} e^{-u} \leq \frac{P^{-1}}{1-P^{-1}} e^{-u}. \end{aligned}$$

It results that for  $P \geq 2$ ,

$$\text{Proba} \left\{ \exists \tilde{\Lambda} \subset \Gamma, \|W'_{\tilde{\Lambda}}\| \geq \sqrt{|\tilde{\Lambda}|} + \sqrt{4|\tilde{\Lambda}| \log_e P + 2u} \right\} \leq \frac{2}{P} e^{-u},$$

from which we get (12.25) by observing that  $W'_\Lambda = \sigma^{-1}W_\Lambda$  because  $W' = \sigma^{-1}W$ . This finishes the proof of Lemma 12.1.

By construction, the best empirical set  $\tilde{\Lambda}_T$  compared to the oracle set  $\Lambda_T$  in (12.21) satisfies

$$\|X - X_{\tilde{\Lambda}_T}\|^2 + T^2 |\tilde{\Lambda}_T| \leq \|X - X_{\Lambda_T}\|^2 + T^2 |\Lambda_T|.$$

By using  $\|X - X_{\tilde{\Lambda}_T}\|^2 = \|X - f\|^2 + \|f - X_{\tilde{\Lambda}_T}\|^2 + 2\langle X - f, f - X_{\tilde{\Lambda}_T} \rangle$  and a similar equality for  $\|X - X_{\Lambda_T}\|^2$  together with the equalities  $\|f - X_{\Lambda_T}\|^2 = \|f - f_{\Lambda_T}\|^2 + \|W_{\Lambda_T}\|^2$  and  $\langle X - f, X_{\tilde{\Lambda}_T} - X_{\Lambda_T} \rangle = \langle X - f, X_{\tilde{\Lambda}_T} - f_{\Lambda_T} \rangle = \|W_{\Lambda_T}\|^2$ , we derive that

$$\|f - X_{\tilde{\Lambda}_T}\|^2 + T^2 |\tilde{\Lambda}_T| \leq \|f - f_{\Lambda_T}\|^2 + T^2 |\Lambda_T| + 2|\langle X - f, X_{\tilde{\Lambda}_T} - f_{\Lambda_T} \rangle|. \quad (12.26)$$

The vectors  $\{\phi_p\}_{p \in \tilde{\Lambda}_T \cup \Lambda_T}$  generate a space  $\mathbf{V}_{\tilde{\Lambda}_T} + \mathbf{V}_{\Lambda_T}$  of dimension smaller or equal to  $|\tilde{\Lambda}_T| + |\Lambda_T|$ . We denote by  $W_{\tilde{\Lambda}_T \cup \Lambda_T}$  the orthogonal projection of the noise  $W$  on this space. The inner product is bounded by writing

$$\begin{aligned} |2\langle X - f, X_{\tilde{\Lambda}_T} - f_{\Lambda_T} \rangle| &= |2\langle W_{\tilde{\Lambda}_T \cup \Lambda_T}, X_{\tilde{\Lambda}_T} - f_{\Lambda_T} \rangle| \\ &\leq 2\|W_{\tilde{\Lambda}_T \cup \Lambda_T}\| (\|X_{\tilde{\Lambda}_T} - f\| + \|f - f_{\Lambda_T}\|). \end{aligned}$$

Lemma 12.1 implies

$$\begin{aligned} |2\langle X - f, X_{\tilde{\Lambda}_T} - f_{\Lambda_T} \rangle| &\leq 2\sigma \left( \sqrt{|\tilde{\Lambda}_T| + |\Lambda_T|} + \sqrt{4(|\tilde{\Lambda}_T| + |\Lambda_T|) \log_e P + 2u} \right) \\ &\quad \left( \|X_{\tilde{\Lambda}_T} - f\| + \|f - f_{\Lambda_T}\| \right), \end{aligned}$$

with a probability greater than  $1 - \frac{2}{P}e^{-u}$ . Applying  $2xy \leq \beta^{-2}x^2 + \beta^2y^2$  successively with  $\beta = 1/2$  and  $\beta = 1$  gives

$$\begin{aligned} |2\langle X - f, X_{\tilde{\Lambda}_T} - f_{\Lambda_T} \rangle| &\leq (1/2)^{-2} 2\sigma^2 (|\tilde{\Lambda}_T| + |\Lambda_T| + 4(|\tilde{\Lambda}_T| + |\Lambda_T|) \log_e P + 2u) \\ &\quad + (1/2)^2 2(\|X_{\tilde{\Lambda}_T} - f\|^2 + \|f - f_{\Lambda_T}\|^2). \end{aligned}$$

Inserting this bound in (12.26) yields

$$\begin{aligned} \frac{1}{2} \|f - X_{\tilde{\Lambda}_T}\|^2 &\leq \frac{3}{2} \|f - f_{\Lambda_T}\|^2 + \sigma^2(\lambda^2 \log_e P + 8(1 + 4 \log_e P))|\Lambda_T| \\ &\quad + \sigma^2(8(1 + 4 \log_e P) - \lambda^2 \log_e P)|\tilde{\Lambda}_T| + 16\sigma^2 u, \end{aligned}$$

So that if  $\lambda^2 \geq 32 + \frac{8}{\log_e P}$ ,

$$\|f - X_{\tilde{\Lambda}_T}\|^2 \leq 3\|f - f_{\Lambda_T}\|^2 + 4\sigma^2\lambda^2 \log_e P|\Lambda_T| + 32\sigma^2 u,$$

which implies for  $T = \lambda \sigma \sqrt{\log_e P}$  that

$$\|f - X_{\tilde{\Lambda}_T}\|^2 \leq 4(\|f - f_{\Lambda_T}\|^2 + T^2|\Lambda_T|) + 32\sigma^2 u,$$

where this result holds with probability greater than  $1 - \frac{2}{P}e^{-u}$ .

Since this is valid for all  $u \geq 0$ , one has

$$\text{Proba} \left\{ \|f - X_{\tilde{\Lambda}_T}\|^2 - 4(\|f - f_{\Lambda_T}\|^2 + T^2|\Lambda_T|) \geq 32\sigma^2 u \right\} \leq \frac{2}{P}e^{-u},$$

which implies by integration over  $u$  that

$$E \left[ \|f - X_{\tilde{\Lambda}_T}\|^2 - 4(\|f - f_{\Lambda_T}\|^2 + T^2|\Lambda_T|) \right] \leq 32\sigma^2 \frac{2}{P},$$

which proves the theorem result (12.24). ■

This theorem proves that the selection of a best-penalized empirical projection produces a risk that is within a factor of 4 of the minimal oracle risk obtained by selecting the best dictionary vectors that approximate  $f$ . Birgé and Massart [114] obtain a better lower bound for  $\lambda$  (roughly  $\lambda > \sqrt{2}$  and thus  $T > \sigma \sqrt{2 \log_e P}$ ) and a multiplicative factor smaller than 4 with a more complex proof using Talgrand's concentration inequalities.

If  $\mathcal{D}$  is an orthonormal basis, then Theorem 12.2 proves that the optimal estimator  $\tilde{F} = X_{\tilde{\Lambda}_T}$  is a hard-thresholding estimator at  $T$ . Thus, this theorem generalizes the thresholding estimation theorem (11.7) of Donoho and Johnstone that computes an upper bound of the thresholding risk in an orthonormal basis with  $P = N$ .

The minimum Lagrangian value  $\mathcal{L}_0(T, f, \Lambda_T)$  is reduced by increasing the size of the dictionary  $\mathcal{D}$ . However, this is paid by also increasing  $T$  proportionally to  $\log_e P$ , so that the penalization term  $T^2|\Lambda|$  is big enough to dominate the impact of the noise on the selection of dictionary vectors. Increasing  $\mathcal{D}$  is thus worth it only if the decay of  $\mathcal{L}_0(T, f, \Lambda_T)$  compensates the increase of  $T$ , as in the compression application of Section 12.1.2.

### **Estimation Risk for Analog Signals**

Discrete signals are most often obtained by discretizing analog signals, and the estimation risk can also be computed on the input analog signal, as in Section 11.5.3. Let  $f[n]$  be the discrete signal obtained by approximating an analog signal  $\bar{f}(x)$  in an approximation space  $\mathbf{U}_N$  of size  $N$ . An estimator  $\tilde{F}[n]$  of  $f[n]$  is converted into an analog estimation  $\bar{F}(x) \in \mathbf{U}_N$  of  $\bar{f}(x)$ , with a discrete-to-analog conversion. We verify as in (11.141) that the total risk is the sum of the discrete estimation risk plus a linear approximation error:

$$E\{\|\bar{f} - \bar{F}\|^2\} = E\{\|f - \tilde{F}\|^2\} + \|\bar{f} - P_{\mathbf{U}_N}\bar{f}\|^2.$$

Suppose that the dictionary has a polynomial size  $P = O(N^\gamma)$  and that the  $\mathbf{l}^0$  Lagrangian decay satisfies

$$\mathcal{L}_0(T, f, \Lambda_T) = \min_{\Lambda \subset \Gamma} \|f - f_\Lambda\|^2 + |\Lambda| T^2 \leq C T^{2-1/s}.$$

If the linear approximation error satisfies  $\|\bar{f} - P_{\mathbf{U}_N}\bar{f}\| = O(N^{-\beta})$ , then by choosing  $N \sim \sigma^{(2s-1)/\beta}$ , we derive from (12.24) in Theorem 12.3 that

$$E\{\|\bar{F} - \bar{f}\|^2\} = O(\sigma^{2-1/s} |\log \sigma|^{2-1/s}). \quad (12.27)$$

When the noise variance  $\sigma^2$  decreases, the risk decay depends on the decay exponent  $s$  of the  $\mathbf{l}^0$  Lagrangian. Optimized dictionaries should thus increase  $s$  as much as possible for any given class  $\Theta$  of signals.

## **12.2 DICTIONARIES OF ORTHONORMAL BASES**

To reduce the complexity of sparse approximations selected in a redundant dictionary, this section restricts such approximations to families of orthogonal vectors. Eliminating approximations from nonorthogonal vectors reduces the number of possible approximation sets  $\Lambda$  in  $\mathcal{D}$ , which simplifies the optimization. In an orthonormal basis, an optimal nonlinear approximation selects the largest-amplitude coefficients. Dictionaries of orthonormal bases take advantage of this property by regrouping orthogonal dictionary vectors in a multitude of orthonormal bases.

**Definition 12.1.** A dictionary  $\mathcal{D}$  is said to be a dictionary of orthonormal bases of  $\mathbb{C}^N$  if any family of orthogonal vectors in  $\mathcal{D}$  also belongs to an orthonormal basis  $\mathcal{B}$  of  $\mathbb{C}^N$  included in  $\mathcal{D}$ .

A dictionary of orthonormal bases  $\mathcal{D} = \{\phi_p\}_{p \in \Gamma}$  is thus a family of  $P > N$  vectors that can also be viewed as a union of orthonormal bases, many of which share common vectors. Wavelet packets and local cosine bases in Chapter 8 define dictionaries of orthonormal bases. In Section 12.2.1 we prove that finding a best signal approximation with orthogonal dictionary vectors can be casted as a search for a best orthonormal basis in which orthogonal vectors are selected by a thresholding. Compression and denoising algorithms are implemented in such a best basis. Tree-structured dictionaries are introduced in Section 12.2.2, in order to compute best bases with a fast dynamic programming algorithm.

### 12.2.1 Approximation, Compression, and Denoising in a Best Basis

Sparse approximations of signals  $f \in \mathbb{C}^N$  are constructed with orthogonal vectors selected from a dictionary  $\mathcal{D} = \{\phi_p\}_{p \in \Gamma}$  of orthonormal bases with compression and denoising applications.

#### **Best Basis**

We denote by  $\Lambda_o \subset \Gamma$  a collection of *orthonormal* vectors in  $\mathcal{D}$ . Sets of nonorthogonal vectors are not considered. The orthogonal projection of  $f$  on the space generated by these vectors is then  $f_{\Lambda_o} = \sum_{p \in \Lambda_o} \langle f, \phi_p \rangle \phi_p$ .

In an orthonormal basis  $\mathcal{B} = \{\phi_p\}_{p \in \Gamma_B}$ , Theorem 12.2 proves that the Lagrangian  $\mathcal{L}_o(T, f, \Lambda_o) = \|f - f_{\Lambda_o}\|^2 + T^2 |\Lambda_o|$  is minimized by selecting coefficients above  $T$ . The resulting minimum is

$$\mathcal{L}_o(T, f, \mathcal{B}) = \underset{\Lambda_o \subset \Gamma_B}{\operatorname{argmin}} \|f - f_{\Lambda_o}\|^2 + T^2 |\Lambda_o| = \sum_{p \in \Gamma_B} \min(|\langle f, \phi_p \rangle|^2, T^2). \quad (12.28)$$

Theorem 12.4 derives that a best approximation from orthogonal vectors in  $\mathcal{D}$  is obtained by thresholding coefficients in a best basis that minimizes this  $\mathbf{l}^0$  Lagrangian.

**Theorem 12.4.** In the best basis

$$\mathcal{B}_T = \underset{\mathcal{B} \subset \mathcal{D}}{\operatorname{argmin}} \mathcal{L}_o(T, f, \mathcal{B}) = \underset{\mathcal{B} \subset \mathcal{D}}{\operatorname{argmin}} \sum_{p \in \Gamma_B} \min(|\langle f, \phi_p \rangle|^2, T^2), \quad (12.29)$$

the thresholded set  $\Lambda_T = \{p \in \Gamma_{\mathcal{B}_T} : |\langle f, \phi_p \rangle| \geq T\}$  satisfies

$$\mathcal{L}_o(T, f, \Lambda_T) = \mathcal{L}_o(T, f, \mathcal{B}_T) = \min_{\Lambda_o \subset \Gamma} \|f - f_{\Lambda_o}\|^2 + |\Lambda_o| T^2, \quad (12.30)$$

and for all  $\Lambda_o \subset \Gamma$ ,

$$\|f - f_{\Lambda_T}\| \leq \|f - f_{\Lambda_o}\| \text{ if } |\Lambda_o| \leq |\Lambda_T|. \quad (12.31)$$

**Proof.** Since any vector in  $\mathcal{D}$  belongs to an orthonormal basis  $\mathcal{B} \subset \mathcal{D}$ , we can write

$\Gamma = \cup_{\mathcal{B} \subset \mathcal{D}} \Gamma_{\mathcal{B}}$ , so (12.28) with (12.29) implies (12.30). The optimal approximation result (12.31), like (12.4), comes from the fact that  $\|f - f_{\Lambda_T}\|^2 + |\Lambda_T| T^2 \leq \|f - f_{\Lambda_o}\|^2 + |\Lambda_o| T^2$ . ■

This theorem proves in (12.31) that the thresholding approximation  $f_M = f_{\Lambda_T}$  of  $f$  in the best orthonormal basis  $\mathcal{B}_T$  is the best approximation of  $f$  from  $M = |\Lambda_T|$  orthogonal vectors in  $\mathcal{D}$ .

### **Compression in a Best Orthonormal Basis**

Section 12.1.2 shows that quantizing signal coefficients over orthogonal dictionary vectors  $\{\phi_p\}_{p \in \Lambda_o}$  yields a distortion

$$d(R, f) \leq \mathcal{L}_0(T, f, \Lambda_o) = \|f - f_{\Lambda_o}\|^2 + T^2 |\Lambda_o|,$$

where  $2T = \Delta$  is the quantization step. A best transform code that minimizes this Lagrangian upper bound is thus implemented in the best orthonormal basis  $\mathcal{B}_T$  in (12.29).

With an entropy coding of the significance map (12.16), the number of bits  $R_0$  to code the indexes of the  $M$  nonzero quantized coefficients among  $P$  dictionary elements is  $R_0 = \log_2 \binom{M}{P} \sim M \log_2(P/M)$ . However, the number of sets  $\Lambda_o$  of orthogonal vectors in  $\mathcal{D}$  is typically much smaller than the number  $2^P$  of subsets  $\Lambda$  in  $\Gamma$  and the resulting number  $R_0$  of bits is thus smaller.

A best-basis search improves the distortion rate  $d(R, f)$  if the Lagrangian approximation reduction is not compensated by the increase of  $R_0$  due to the increase of the number of orthogonal vector sets. For example, if the original signal is piecewise smooth, then a best wavelet packet basis does not concentrate the signal energy much more efficiently than a wavelet basis. Despite the fact that a wavelet packet dictionary includes a wavelet basis, the distortion rate in a best wavelet packet basis is then larger than in a single wavelet basis. For geometrically regular images, Section 12.2.4 shows that a dictionary of bandlet orthonormal bases reduces the distortion rate of a wavelet basis.

### **Denoising in a Best Orthonormal Basis**

To estimate a signal  $f$  from noisy signal observations

$$X[n] = f[n] + W[n] \quad \text{for } 0 \leq n < N,$$

where  $W[n]$  is a Gaussian white noise, Theorem 12.3 proves that a nearly optimal estimator is obtained by minimizing a penalized empirical Lagrangian,

$$\mathcal{L}_0(T, X, \Lambda) = \|X - X_\Lambda\|^2 + T^2 |\Lambda|. \quad (12.32)$$

Restricting  $\Lambda$  to be a set  $\Lambda_o$  of orthogonal vectors in  $\mathcal{D}$  reduces the set of possible *signal models*. As a consequence, Theorem 12.3 remains valid for this subfamily of models. Theorem 12.4 proves in (12.30) that the Lagrangian (12.32) is minimized by thresholding coefficients in a best basis. A best-basis thresholding thus yields a risk that is within a factor of 4 of the best estimation obtained by an oracle. This best basis can be calculated with a fast algorithm described in Section 12.2.2.

#### **12.2.2 Fast Best-Basis Search in Tree Dictionaries**

Tree dictionaries of orthonormal bases are constructed with a recursive split of orthogonal vector spaces and by defining specific orthonormal bases in each subspace. For any additive cost function such as the  $\mathbf{l}^0$  Lagrangian (12.2), a fast dynamic

programming algorithm finds a best basis with a number of operations proportional to the size  $P$  of the dictionary.

### Recursive Split of Vector Spaces

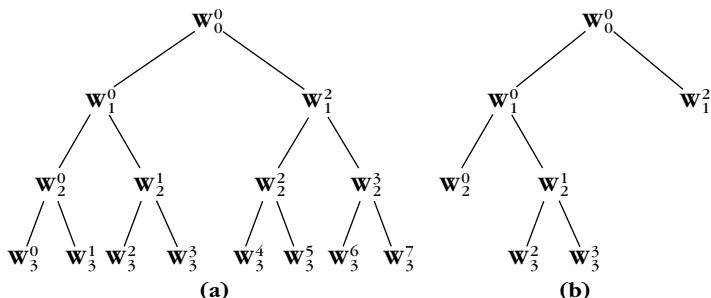
A tree dictionary  $\mathcal{D} = \{\phi_p\}_{p \in \Gamma}$  is obtained by recursively dividing vector spaces into  $q$  orthogonal subspaces, up to a maximum recursive depth. This recursive split is represented by a tree. A vector space  $\mathbf{W}_d^l$  is associated to each tree node at a depth  $d$  and position  $l$ . The  $q$  children of this node correspond to an orthogonal partition of  $\mathbf{W}_d^l$  into  $q$  orthogonal subspaces  $\mathbf{W}_{d+1}^{ql+i}$  at depth  $d+1$ , located at the positions  $ql+i$  for  $0 \leq i < q$ :

$$\mathbf{W}_d^l = \bigoplus_{i=0}^{q-1} \mathbf{W}_{d+1}^{ql+i}. \quad (12.33)$$

Space  $\mathbf{W}_0^0$  at the root of the tree is the full signal space  $\mathbb{C}^N$ . One or several specific orthonormal bases are constructed for each space  $\mathbf{W}_d^l$ . The dictionary  $\mathcal{D}$  is the union of all these specific orthonormal bases for all the spaces  $\mathbf{W}_d^l$  of the tree.

Chapter 8 defines dictionaries of wavelet packet and local cosine bases along binary trees ( $q=2$ ) for one-dimensional signals and along quad-trees ( $q=4$ ) for images. These dictionaries are constructed with a single basis for each space  $\mathbf{W}_d^l$ . For signals of size  $N$ , they have  $P=N \log_2 N$  vectors. The bandlet dictionary in Section 12.2.4 is also defined along a quad-tree, but each space  $\mathbf{W}_d^l$  has several specific orthonormal bases corresponding to different image geometries.

An admissible subtree of a full dictionary tree is a subtree where each node is either a leaf or has its  $q$  children. Figure 12.1(b) gives an example of a binary admissible tree. We verify by induction that the vector spaces at the leaves of an admissible tree define an orthogonal partition of  $\mathbf{W}_0^0 = \mathbb{C}^N$  into orthogonal subspaces. The union of orthonormal bases of these spaces is therefore an orthonormal basis of  $\mathbb{C}^N$ .



**FIGURE 12.1**

(a) Full binary tree of depth 3, indexing all possible spaces  $\mathbf{W}_d^l$ . (b) Example of admissible subarea.

### Additive Costs

A best basis can be defined with a cost function that is not necessarily the  $\mathbf{1}^0$  Lagrangian (12.2). An additive cost function of a signal  $f$  in a basis  $\mathcal{B} = \{\phi_p\}_{p \in \Gamma_{\mathcal{B}}}$  is defined as a sum of independent contributions from each coefficient in  $\mathcal{B}$ :

$$\mathcal{C}(f, \mathcal{B}) = \sum_{p \in \Gamma_{\mathcal{B}}} C(|\langle f, \phi_p \rangle|). \quad (12.34)$$

A best basis of  $\mathbb{C}^N$  in  $\mathcal{D}$  minimizes the resulting cost,

$$\widehat{\mathcal{B}} = \underset{\mathcal{B} \subset \mathcal{D}}{\operatorname{argmin}} \mathcal{C}(f, \mathcal{B}). \quad (12.35)$$

The  $\mathbf{1}^0$  Lagrangian (12.2) is an example of additive cost function,

$$\mathcal{C}(f, \mathcal{B}) = \mathcal{L}_0(T, f, \mathcal{B}) = \sum_{p \in \Gamma_{\mathcal{B}}} C(|\langle f, \phi_p \rangle|) \quad \text{with} \quad C(x) = \min(T^2, x^2). \quad (12.36)$$

The minimization of an  $\mathbf{1}^1$  norm is obtained with

$$\mathcal{C}(f, \mathcal{B}) = \|f_{\mathcal{B}}\|_1 = \sum_{p \in \Gamma_{\mathcal{B}}} C(|\langle f, \phi_p \rangle|) \quad \text{for} \quad C(x) = |x|. \quad (12.37)$$

In Section 12.4.1 we introduce a basis pursuit algorithm that also minimizes the  $\mathbf{1}^1$  norm of signal coefficients in a redundant dictionary. A basis pursuit selects a best basis but without imposing any orthogonal constraint.

### Fast Best-Basis Selection

The fast best-basis search algorithm, introduced by Coifman and Wickerhauser [180], relies on the dictionary tree structure and on the cost additivity. This algorithm is a particular instance of the Classification and Regression Tree (CART) algorithm by Breiman et al. [9]. It explores all tree nodes, from bottom to top, and at each node it computes the best basis  $\widehat{\mathcal{B}}_d^l$  of the corresponding space  $\mathbf{W}_d^l$ .

The cost additivity property (12.34) implies that an orthonormal basis  $\mathcal{B} = \cup_{i=0}^{q-1} \mathcal{B}_i$ , which is a union of  $q$  orthonormal families  $\mathcal{B}_i$ , has a cost equal to the sum of their cost:

$$\mathcal{C}(f, \mathcal{B}) = \sum_{i=0}^{q-1} \mathcal{C}(f, \mathcal{B}_i).$$

As a result, the best basis  $\widehat{\mathcal{B}}_d^l$ , which minimizes this cost among all bases of  $\mathbf{W}_d^l$ , is either one of the specific bases of  $\mathbf{W}_d^l$  or a union of the best bases  $\widehat{\mathcal{B}}_{d+1}^{ql+i}$  that were previously calculated for each of its subspace  $\mathbf{W}_{d+1}^{ql+i}$  for  $0 \leq i < q$ . The decision is thus performed by minimizing the resulting cost, as described in Algorithm 12.1.

**ALGORITHM 12.1***Initialization*

- Compute all dictionary coefficients  $\{\langle f, \phi_p \rangle\}_{p \in \Gamma}$ .
- Initialize the cost of each tree space  $\mathbf{W}_d^l$  by finding the basis  $\mathcal{B}_d^l$  of minimum cost among all specific bases  $\mathcal{B}$  of  $\mathbf{W}_d^l$ :

$$\mathcal{B}_d^l = \underset{\mathcal{B} \text{ of } \mathbf{W}_d^l}{\operatorname{argmin}} \mathcal{C}(f, \mathcal{B}) = \underset{\mathcal{B} \text{ of } \mathbf{W}_d^l}{\operatorname{argmin}} \sum_{p \in \Gamma_B} C(|\langle f, \phi_p \rangle|). \quad (12.38)$$

*Cost Update*

- For each tree node  $(d, l)$ , visited from the bottom to the top ( $d$  decreasing), if we are not at the bottom and if

$$\mathcal{C}(f, \mathcal{B}_d^l) > \sum_{i=0}^{q-1} \mathcal{C}(f, \widehat{\mathcal{B}}_{d+1}^{ql+i}),$$

then set  $\widehat{\mathcal{B}}_d^l = \cup_{i=0}^{q-1} \widehat{\mathcal{B}}_{d+1}^{ql+i}$ ; otherwise set  $\widehat{\mathcal{B}}_d^l = \mathcal{B}_d^l$ . ■

This algorithm outputs the best basis  $\widehat{\mathcal{B}} = \widehat{\mathcal{B}}_0^0$  of  $\mathbb{C}^N = \mathbf{W}_0^0$  that has a minimum cost among all bases of the dictionary. For wavelet packet and local cosine dictionaries, there is a single specific basis per space  $\mathbf{W}_d^l$ , so (12.38) is reduced to computing the cost in this basis. In a bandlet dictionary there are many specific bases for each  $\mathbf{W}_d^l$  corresponding to different geometric image models.

For a dictionary of size  $P$ , the number of comparisons and additions to construct this best basis is  $O(P)$ . The algorithmic complexity is thus dominated by the computation of the  $P$  dictionary coefficients  $\{\langle f, \phi_p \rangle\}_{p \in \Gamma}$ . If implemented with  $O(P)$  operations with a fast transform, then the overall computational algorithmic complexity is  $O(P)$ . This is the case for wavelet packet, local cosine, and bandlet dictionaries.

### 12.2.3 Wavelet Packet and Local Cosine Best Bases

A best wavelet packet or local cosine basis selects time-frequency atoms that match the time-frequency resolution of signal structures. Therefore, it adapts the time-frequency geometry of the approximation support  $\Lambda_T$ . Wavelet packet and local cosine dictionaries are constructed in Chapter 8. We evaluate these approximations through examples that also reveal their limitations.

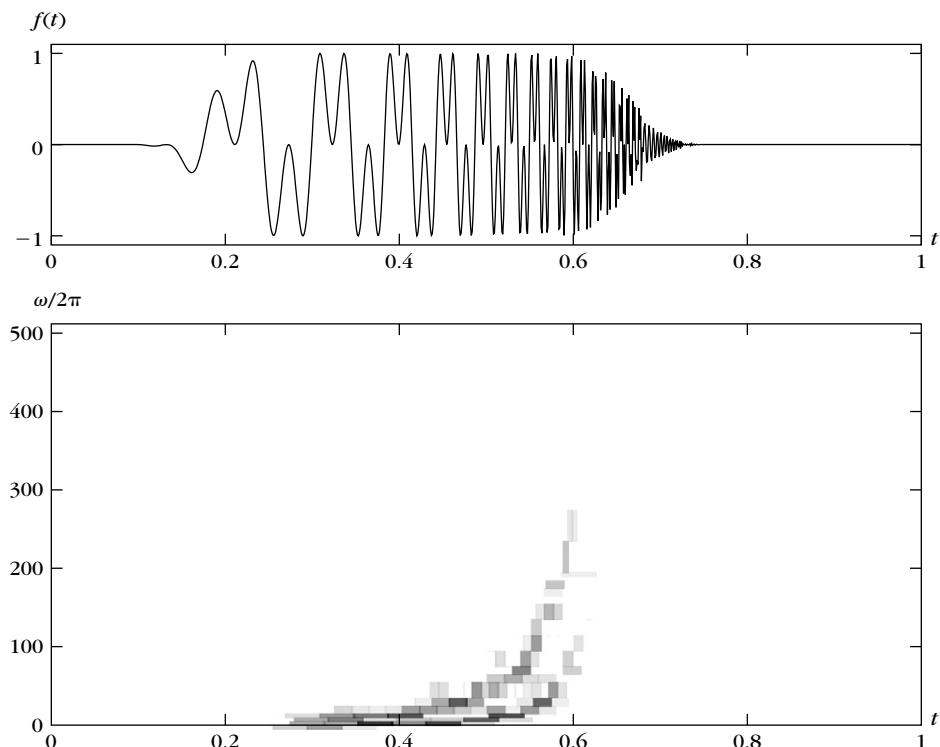
#### ***Best Orthogonal Wavelet Packet Approximations***

A wavelet packet orthogonal basis divides the frequency axis into intervals of varying dyadic sizes  $2^j$ . Each frequency interval is covered by a wavelet packet function that is uniformly translated in time. A best wavelet packet basis can thus be interpreted as a “best” segmentation of the frequency axis in dyadic sizes intervals.

A signal is well approximated by a best wavelet packet basis, if in any frequency interval, the high-energy structures have a similar time-frequency spread. The time translation of the wavelet packet that covers this frequency interval is then well

adapted to approximating all the signal structures in this frequency range that appear at different times. In the best basis computed by minimizing the  $\mathbf{L}^0$  Lagrangian in (12.36), Theorem 12.4 proves that the set of  $\Lambda_T$  of wavelet packet coefficients above  $T$  correspond to the orthogonal wavelet packet vectors that best approximate  $f$  in the whole wavelet packet dictionary. These wavelet packets are represented by Heisenberg boxes, as explained in Section 8.1.2.

Figure 12.2 gives the best wavelet packet approximation set  $\Lambda_T$  of a signal composed of two hyperbolic chirps. The proportion of wavelet packet coefficients that are retained is  $M/N = |\Lambda_T|/N = 8\%$ . The resulting best  $M$ -term orthogonal approximation  $f_{\Lambda_T}$  has a relative error  $\|f - f_{\Lambda_T}\|/\|f\| = 0.11$ . The wavelet packet tree was calculated with the symmlet 8 conjugate mirror filter. The time support of chosen wavelet packets is reduced at high frequencies to adapt itself to the chirps' rapid modification of frequency content. The energy distribution revealed by the wavelet packet Heisenberg boxes in  $\Lambda_T$  is similar to the scalogram calculated in Figure 4.17.



**FIGURE 12.2**

The top signal includes two hyperbolic chirps. The Heisenberg boxes of the best orthogonal wavelet packets in  $\Lambda_T$  are shown in the bottom image. The darkness of each rectangle is proportional to the amplitude of the corresponding wavelet packet coefficient.

Figure 8.6 gives another example of a best wavelet packet basis for a different multichirp signal, calculated with the entropy cost  $C(x) = |x| \log_e |x|$  in (12.34).

Let us mention that the application of best wavelet packet bases to pattern recognition is difficult because these dictionaries are not translation invariant. If the signal is translated, its wavelet packet coefficients are severely modified and the Lagrangian minimization may yield a different basis. This remark applies to local cosine bases as well.

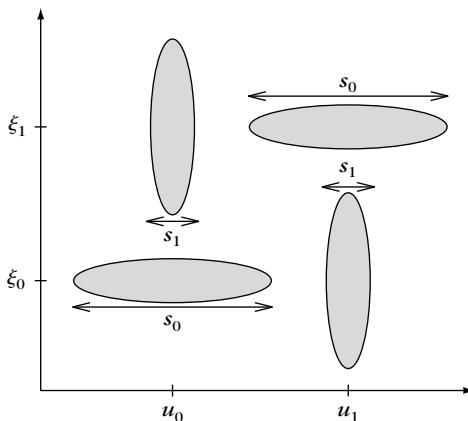
If the signal includes *different types* of high-energy structures, located at different times but in the same frequency interval, there is no wavelet packet basis that is well adapted to all of them. Consider, for example, a sum of four transients centered, respectively, at  $u_0$  and  $u_1$  at two different frequencies  $\xi_0$  and  $\xi_1$ :

$$\begin{aligned} f(t) = & \frac{K_0}{\sqrt{s_0}} g\left(\frac{t-u_0}{s_0}\right) e^{i\xi_0 t} + \frac{K_1}{\sqrt{s_1}} g\left(\frac{t-u_1}{s_1}\right) e^{i\xi_0 t} \\ & + \frac{K_2}{\sqrt{s_1}} g\left(\frac{t-u_0}{s_1}\right) e^{i\xi_1 t} + \frac{K_3}{\sqrt{s_0}} g\left(\frac{t-u_1}{s_0}\right) e^{i\xi_1 t}. \end{aligned} \quad (12.39)$$

The smooth window  $g$  has a Fourier transform  $\hat{g}$  whose energy is concentrated at low frequencies. The Fourier transform of the four transients have their energy concentrated in frequency bands centered, respectively, at  $\xi_0$  and  $\xi_1$ :

$$\begin{aligned} \hat{f}(\omega) = & K_0 \sqrt{s_0} \hat{g}(s_0(\omega - \xi_0)) e^{-iu_0(\omega - \xi_0)} + K_1 \sqrt{s_1} \hat{g}(s_1(\omega - \xi_0)) e^{-iu_1(\omega - \xi_0)} \\ & + K_2 \sqrt{s_1} \hat{g}(s_1(\omega - \xi_1)) e^{-iu_0(\omega - \xi_1)} + K_3 \sqrt{s_0} \hat{g}(s_0(\omega - \xi_1)) e^{-iu_1(\omega - \xi_1)}. \end{aligned}$$

If  $s_0$  and  $s_1$  have different values, the time and frequency spread of these transients is different, which is illustrated in Figure 12.3. In the best wavelet packet basis



**FIGURE 12.3**

Time-frequency energy distribution of the four elementary atoms in (12.39).

selection, the first transient  $K_0 s_0^{-1/2} g(s_0^{-1}(t - u_0)) \exp(i\xi_0 t)$  “votes” for a wavelet packet basis with a scale  $2^j$  that is of the order  $s_0$  at the frequency  $\xi_0$  whereas  $K_1 s_1^{-1/2} g(s_1^{-1}(t - u_1)) \exp(i\xi_0 t)$  “votes” for a wavelet packet basis with a scale  $2^j$  that is close to  $s_1$  at the same frequency. The “best” wavelet packet is adapted to the transient of highest energy. The energy of the smaller transient is then spread across many “best” wavelet packets. The same thing happens for the second pair of transients located in the frequency neighborhood of  $\xi_1$ .

Speech recordings are examples of signals that have properties that rapidly change in time. At two different instants in the same frequency neighborhood, the signal may have totally different energy distributions. A best orthogonal wavelet packet basis is not adapted to this time variation and gives poor nonlinear approximations. Sections 12.3 and 12.4 show that a more flexible nonorthogonal approximation with wavelet packets, computed with a pursuit algorithm, can have the required flexibility.

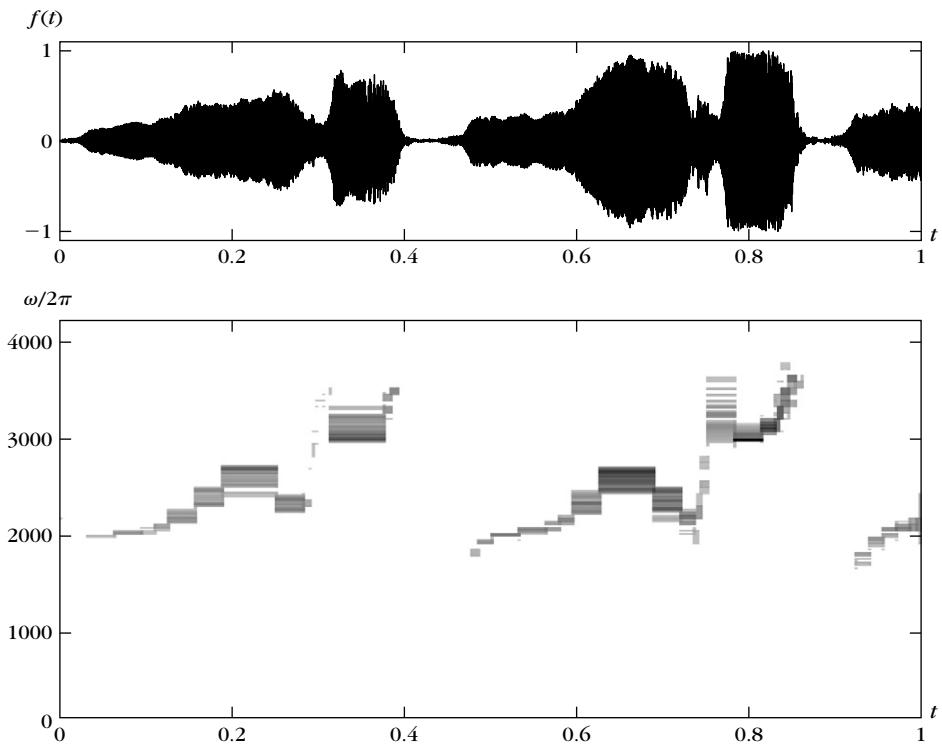
As in one dimension, an image is well approximated in a best wavelet packet basis if its structures within a given frequency band have similar properties across the whole image. For natural scene images, a best wavelet packet often does not provide much better nonlinear approximations than the wavelet basis included in this wavelet packet dictionary. However, for specific classes of images such as fingerprints, one may find wavelet packet bases that significantly outperform the wavelet basis [122].

### ***Best Orthogonal Local Cosine Representations***

Tree dictionaries of local cosine bases are constructed in Section 8.5 with  $P = N \log_2 N$  local cosine vectors. They divide the time axis into intervals of varying dyadic sizes. A best local cosine basis adapts the time segmentation to the variations of the signal time-frequency structures. It is computed with  $O(N \log_2 N)$  operations with the best-basis search algorithm from Section 12.2.2.

In comparison with wavelet packets, we gain time adaptation but we lose frequency flexibility. A best local cosine basis is therefore well adapted to approximating signals with properties that may vary in time, but that do not include structures of very different time and frequency spreads at any given time. Figure 12.4 shows the Heisenberg boxes of the set  $\Lambda_T$  of orthogonal local cosine vectors that best approximate the recording of a bird song, computed by minimizing the  $1^0$  Lagrangian (12.36). The chosen threshold  $T$  yields a relative approximation error  $\|f - f_{\Lambda_T}\| / \|f\| = 5 \times 10^{-2}$  with  $|\Lambda_T|/N = 11\%$  coefficients. The selected local cosine vectors have a time and frequency resolution adapted to the transients and harmonic structures of the signal. Figure 8.19 shows a best local cosine basis that is calculated with an entropy cost function for a speech recording.

The sum of four transients (12.39) is not efficiently represented in a wavelet packet basis but neither is it well approximated in a best local cosine basis. Indeed, if the scales  $s_0$  and  $s_1$  are very different, at  $u_0$  and  $u_1$  this signal includes two transients at the frequency  $\xi_0$  and  $\xi_1$ , respectively, that have a very different time-frequency spread. In each time neighborhood, the size of the window is adapted to

**FIGURE 12.4**

Recording of a bird song (*top*). The Heisenberg boxes of the best orthogonal local cosine vectors in  $\Lambda_T$  are shown in the bottom image. The darkness of each rectangle is proportional to the amplitude of the local cosine coefficient.

the transient of highest energy. The energy of the second transient is spread across many local cosine vectors. Efficient approximations of such signals require more flexibility, which is provided by the pursuit algorithms from Sections 12.3 and 12.4.

Figure 12.5 gives a denoising example with a best local cosine estimator. The signal in Figure 12.5(b) is the bird song contaminated by an additive Gaussian white noise of variance  $\sigma^2$  with an SNR of 12 db. According to Theorem 12.3, a best orthogonal projection estimator is computed by selecting a set  $\tilde{\Lambda}_T$  of best orthogonal local cosine dictionary vectors, which minimizes an empirical penalized risk. This penalized risk corresponds to the empirical  $\mathbf{1}^0$  Lagrangian (12.23), which is minimized by the best-basis algorithm. The chosen threshold of  $T = 3.5\sigma$  is well below the theoretical universal threshold of  $T = \sigma\sqrt{2 \log_e P}$ , which improves the SNR. The Heisenberg boxes of local cosine vectors indexed by  $\tilde{\Lambda}_T$  are shown in boxes of the remaining coefficients in Figure 12.5(c). The orthogonal projection  $\tilde{F} = X_{\tilde{\Lambda}}$  is shown in Figure 12.5(d).

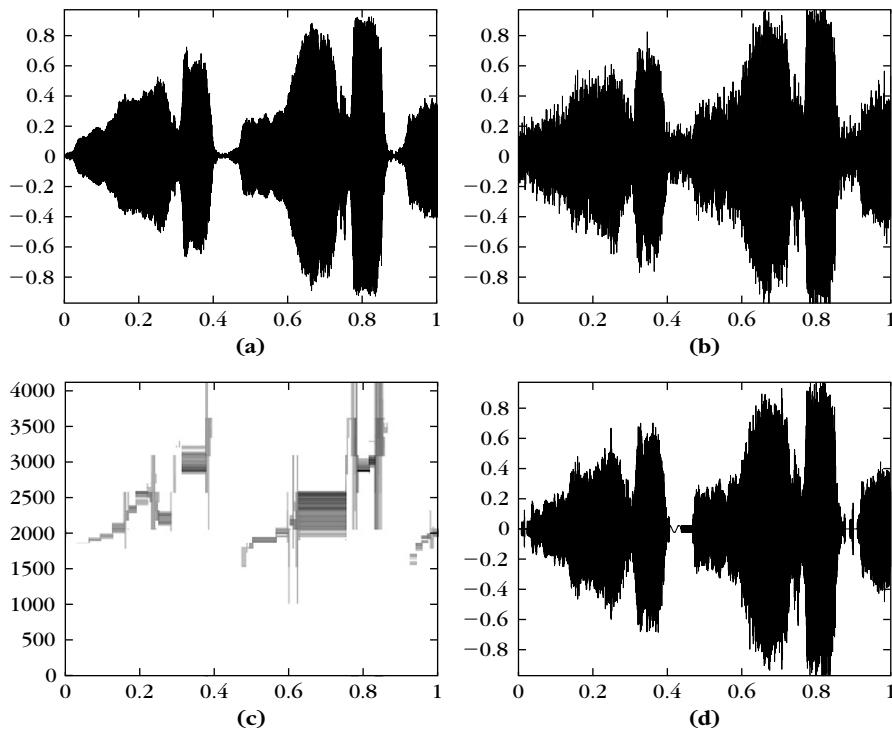


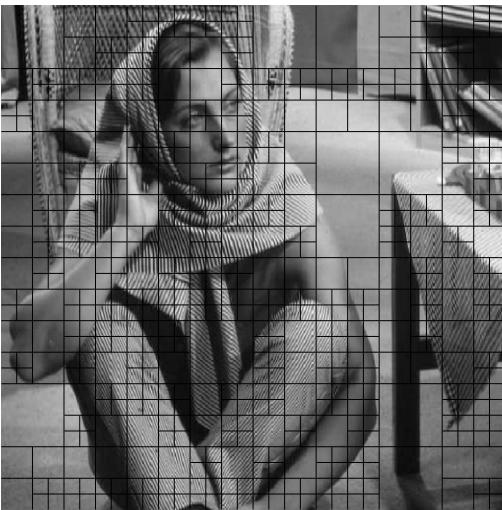
FIGURE 12.5

(a) Original bird song. (b) Noisy signal ( $\text{SNR} = 12 \text{ db}$ ). (c) Heisenberg boxes of the set  $\tilde{\Lambda}_T$  of estimated best orthogonal local cosine vectors. (d) Estimation reconstructed from noisy local cosine coefficients in  $\tilde{\Lambda}_T$  ( $\text{SNR} = 15.5 \text{ db}$ ).

In two dimensions, a best local cosine basis divides an image into square windows that have a size adapted to the spatial variations of local image structures. Figure 12.6 shows the best-basis segmentation of the Barbara image, computed by minimizing the  $\mathbf{l}^1$  norm of its coefficients, with the  $\mathbf{l}^1$  cost function (12.37). The squares are bigger in regions where the image structures remain nearly the same. Figure 8.22 shows another example of image segmentation with a best local cosine basis, also computed with an  $\mathbf{l}^1$  norm.

#### 12.2.4 Bandlets for Geometric Image Regularity

Bandlet dictionaries are constructed to improve image representations by taking advantage of their geometric regularity. Wavelet coefficients are not optimally sparse but inherit geometric image regularity. A bandlet transform applies a directional wavelet transform over wavelet coefficients to reduce the number of large coefficients. This directional transformation depends on a geometric approximation model calculated from the image. Le Pennec, Mallat, and Peyré [342, 365, 396]

**FIGURE 12.6**

The grid shows the approximate support of square overlapping windows in the best local cosine basis, computed with an  $\mathbf{1}^1$  cost.

introduced dictionaries of orthogonal bandlet bases, where the best-basis selection optimizes the geometric approximation model. For piecewise  $C^\alpha$  images, the resulting  $M$ -term bandlet approximations have an optimal asymptotic decay in  $O(M^{-\alpha})$ .

### **Approximation of Piecewise $C^\alpha$ Images**

Definition 9.1 defines a piecewise  $C^\alpha$  image  $f$  as a function that is uniformly Lipschitz  $\alpha$  everywhere outside a set of edge curves, which are also uniformly Lipschitz  $\alpha$ . This image may also be blurred by an unknown convolution kernel. If  $f$  is uniformly Lipschitz  $\alpha$  without edges, then Theorem 9.16 proves that a linear wavelet approximation has an optimal error decay  $\varepsilon_l(M, f) = \|f - f_M\|^2 = O(M^{-\alpha})$ . Edges produce a larger linear approximation error  $\varepsilon_l(M, f) = O(M^{-1/2})$ , which is improved by a nonlinear wavelet approximation  $\varepsilon_n(M, f) = O(M^{-1})$ , but without recovering the  $O(M^{-\alpha})$  decay. For  $\alpha = 2$ , Section 9.3 shows that a piecewise linear approximation over an optimized adaptive triangulation with  $M$  triangles reaches the error decay  $O(M^{-2})$ . Thresholding curvelet frame coefficients also yields a nonlinear approximation error  $\varepsilon_n(M, f) = O(M^{-2}(\log M)^3)$  that is nearly optimal. However, curvelet approximations are not as efficient as wavelets for less regular functions such as bounded variation images. If  $f$  is piecewise  $C^\alpha$  with  $\alpha > 2$ , curvelets cannot improve the  $M^{-2}$  decay either.

The beauty of wavelet and curvelet approximation comes from their simplicity. A simple thresholding directly selects the signal approximation support. However, for images with geometric structures of various regularity, these approximations do not remain optimal when the regularity exponent  $\alpha$  changes. It does not seem

possible to achieve this result without using a redundant dictionary, which requires a more sophisticated approximation scheme.

Elegant adaptive approximation schemes in redundant dictionaries have been developed for images having some geometric regularity. Several algorithms are based on the lifting technique described in Section 7.8, with lifting coefficients that depend on the estimated image regularity [155, 234, 296, 373, 477]. The image can also be segmented adaptively in dyadic squares of various sizes, and approximated on each square by a finite element such as a wedget, which is a step edge along a straight line with an orientation that is adjusted [216]. Refinements with polynomial edges have also been studied [436], but these algorithms do not provide  $M$ -term approximation errors that decay like  $O(M^{-\alpha})$  for all piecewise regular  $C^\alpha$  images.

### **Bandletization of Wavelet Coefficients**

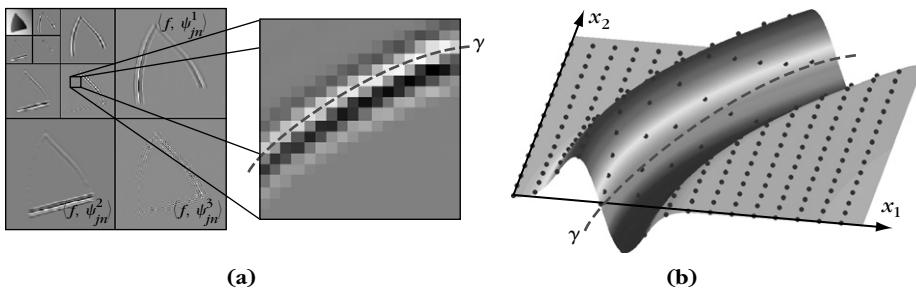
A bandlet transform takes advantage of the geometric regularity captured by a wavelet transform. The decomposition coefficients of  $f$  in an orthogonal wavelet basis can be written as

$$\langle f, \psi_{j,n}^k \rangle = f \star \bar{\psi}_j^k(2^j n) \quad \text{with} \quad \bar{\psi}_j^k(x) = 2^{-j} \psi_j^k(-2^{-j}x), \quad (12.40)$$

for  $x = (x_1, x_2)$  and  $n = (n_1, n_2)$ . The function  $f \star \bar{\psi}_j^k(x)$  has the directional regularity of  $f$ , for example along an edge, and it is regularized by the convolution with  $\bar{\psi}_j^k(x)$ . Figure 12.7 shows a zoom on wavelet coefficients near an edge.

Bandlets retransform wavelet coefficients to take advantage of their directional regularity. This is implemented with a directional wavelet transform applied over wavelet coefficients, which creates new vanishing moments in appropriate directions. The resulting bandlets are written as

$$\phi_p(x) = \sum_n \tilde{\psi}_{i,l,m}[n] \psi_{j,n}^k(x) \quad \text{with} \quad p = (k, j, l, i, m), \quad (12.41)$$



**FIGURE 12.7**

Orthogonal wavelet coefficients at a scale  $2^j$  are samples of a function  $f \star \bar{\psi}_j^k(x)$ , shown in (a). The filtered image  $f \star \bar{\psi}_j^k(x)$  varies regularly when moving along an edge  $\gamma$  (b).

where  $\tilde{\psi}_{i,l,m}[n]$  is a directional wavelet of length  $2^i$ , of width  $2^l$ , and that has a position indexed by  $m$  in the wavelet coefficient array. The bandlet function  $\phi_p(x)$  is a finite linear combination of wavelets  $\psi_{j,n}^k(x)$  and thus has the same regularity as these wavelets. Since  $\psi_{j,n}^k(x)$  has a square support of width proportional to  $2^j$ , the bandlet  $\phi_p$  has a support length proportional to  $2^{j+i}$  and a support width proportional to  $2^{j+l}$ .

If the regularity exponent  $\alpha$  is known then in a neighborhood of an edge, one would like to have elongated bandlets with an aspect ratio defined by  $2^{l+j} = (2^{i+j})^\alpha$ , and thus  $l = \alpha i + (\alpha - 1)j$ . Curvelets satisfy this property for  $\alpha = 2$ . However, when  $\alpha$  is not known in advance and may change, the scale parameters  $i$  and  $l$  must be adjusted adaptively.

As a result of (12.41), the bandlet coefficients of a signal  $\bar{f}$  can be written as

$$\langle \bar{f}, \phi_p \rangle = \sum_n \tilde{\psi}_{i,l,m}[n] \langle \bar{f}, \psi_{j,n}^k \rangle.$$

They are computed by applying a discrete directional wavelet transform on the signal wavelet coefficients  $\langle \bar{f}, \psi_{j,n}^k \rangle$  for each  $k$  and  $2^j$ . This is also called a *bandletization* of wavelet coefficients.

### Geometric Approximation Model

The discrete directional wavelets  $\{\tilde{\psi}_{i,l,m}[n]\}_{i,l,m}$  are defined with a geometric approximation model providing information about the directional image regularity. Many constructions are possible [359]. We describe here a geometric approximation model that is piecewise parallel and yields orthogonal bandlet bases.

For each scale  $2^j$  and direction  $k$ , the array of wavelet transform coefficients  $\{\langle f, \psi_{j,n}^k \rangle\}_n$  is divided into squares of various sizes, as shown in Figure 12.8(b). In regular image regions, wavelet coefficients are small and do not need to be retransformed. Near junctions, the image is irregular in all directions and these few wavelet coefficients are not retransformed either. It is in the neighborhood of edges and directional image structures that an appropriate retransformation can improve the wavelet sparsity.

A *geometric flow* is defined over each *edge square*. It provides the direction along which the discrete bandlets  $\tilde{\psi}_{i,l,m}[n]$  are constructed. It is a vector field, which is parallel horizontally or vertically and points in local directions in which  $f \star \bar{\psi}_j^k(x)$  is the most regular. Figure 12.9(a) gives an example. The segmentation of wavelet coefficients in squares and the specification of a geometric flow in each square defines a geometric approximation model that is used to construct a bandlet basis.

### Bandlets with Alpert Wavelets

Let us consider a square of wavelet coefficients where a geometric flow is defined. We suppose that the flow is parallel vertically. Its vectors can thus be written  $\vec{\gamma}(x) = (1, \tilde{\gamma}'(x_1))$ . Let  $\tilde{\gamma}(x_1)$  be a primitive of  $\tilde{\gamma}'(x_1)$ . Wavelet coefficients are translated vertically with a warping operator  $w(x_1, x_2) = (x_1, x_2 - \tilde{\gamma}(x_1))$  so that the resulting

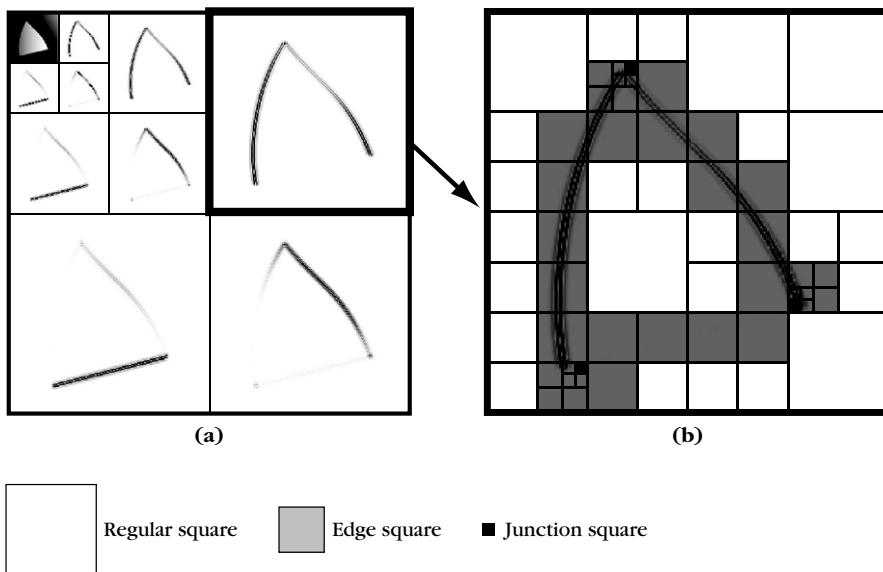


FIGURE 12.8

(a) Wavelet coefficients of the image. (b) Example of segmentation of an array of wavelet coefficients  $\langle f, \psi_{j,n}^k \rangle$  for a particular direction  $k$  and scale  $2^j$ .

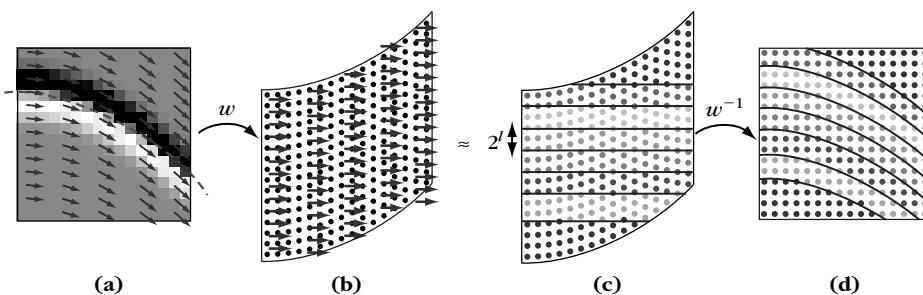


FIGURE 12.9

(a) Square of wavelet coefficients including an edge. A geometric flow nearly parallel to the edge is shown with arrows. (b) A vertical warping  $w$  maps the flow onto a horizontal flow. (c) Support of directional wavelets  $\tilde{\psi}_{i,l,m}[w(n)]$  of length  $2^l$  and width  $2^l$  in the warped domain. (d) Directional wavelets  $\tilde{\psi}_{i,l,m}[n]$  in the square of wavelet coefficients.

geometric flow becomes horizontal, as shown in Figure 12.9(b). In the warped domain, the regularity of  $f \star \tilde{\psi}_j^k(w(x))$  is now horizontal.

Warped directional wavelets  $\tilde{\psi}_{i,l,m}[w(n)]$  are defined to take advantage of this horizontal regularity over the translated orthogonal wavelet coefficients, which are

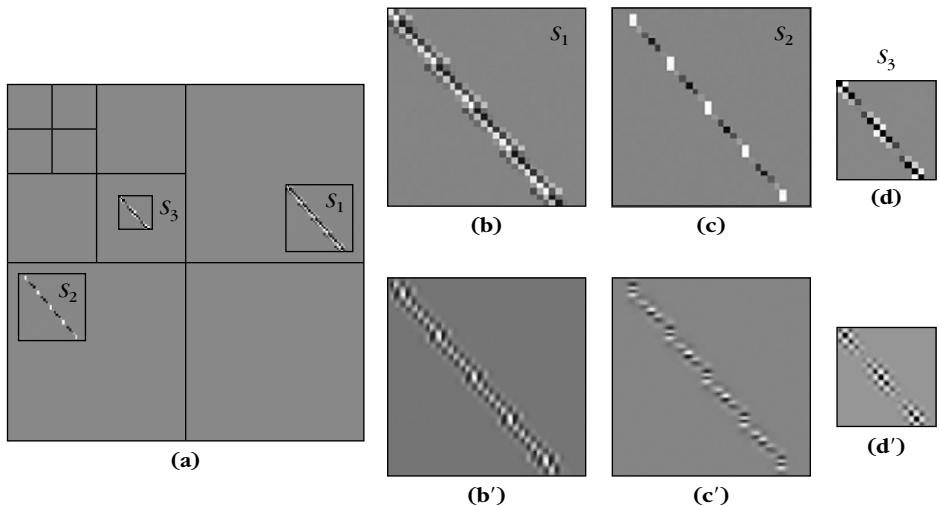


FIGURE 12.10

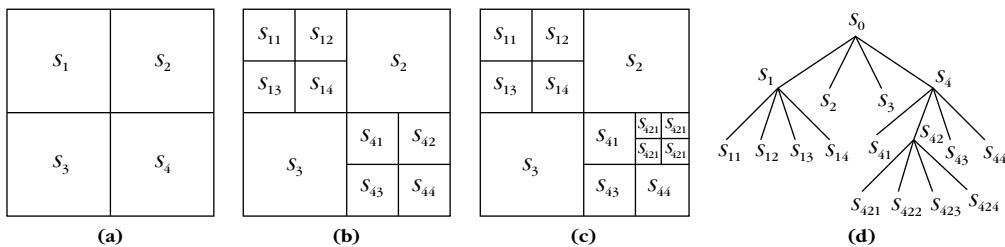
(a) Squares of wavelet coefficients on which bandlets are computed. (b–d) Directional Alpert wavelets  $\tilde{\psi}_{i,l,m}[n]$  of different length  $2^i$  and width  $2^l$ . (b'–d') Bandlet functions  $\phi_p(x)$  computed from the directional wavelets in (b–d), and the wavelets  $\psi_j^k(x)$  corresponding to the squares in (a).

not located on a square grid anymore. Directional wavelets can be constructed with piecewise polynomial Alpert wavelets [84], which are adapted to nonuniform sampling grids [365] and have  $q$  vanishing moments. Over a square of width  $2^i$ , a discrete Alpert wavelet  $\tilde{\psi}_{i,l,m}[w(n)]$  has a length  $2^i$ , a total of  $2^i \times 2^l$  coefficients on its support, and thus a width of the order of  $2^l$  and a position  $m2^l$ . These directional wavelets are horizontal in the warped domain, as shown in Figure 12.9(c). After inverse warping,  $\tilde{\psi}_{i,l,m}[n]$  is parallel to the geometric flow in the original wavelet square, and  $\{\tilde{\psi}_{i,l,m}[n]\}_{i,l,m}$  is an orthonormal basis over the square of  $2^{2i}$  wavelet coefficients. The fast Alpert wavelet transform computes  $2^{2i}$  bandlet coefficients in a square of  $2^{2i}$  coefficients with  $O(2^{2i})$  operations.

Figure 12.10 shows in (b), (c), and (d) several directional Alpert wavelets  $\tilde{\psi}_{i,l,m}[n]$  on squares of different lengths  $2^i$ , and for different width  $2^l$ . The corresponding bandlet functions  $\phi_p(x)$  are computed in (b'), (c'), and (d'), with the wavelets  $\psi_j^k(x)$  corresponding to the squares shown in Figure 12.10(a).

### Dictionary of Bandlet Orthonormal Bases

A bandlet orthonormal basis is defined by segmenting each array of wavelet coefficients  $\langle f, \psi_{j,n}^k \rangle$  in squares of various sizes, and by applying an Alpert wavelet transform along the geometric flow defined in each square. A dictionary of bandlet

**FIGURE 12.11**

**(a–c)** Construction of a dyadic segmentation by successive subdivisions of squares. **(d)** Quad-tree representation of the segmentation. Each leaf of the tree corresponds to a square in the final segmentation.

orthonormal bases is associated to a family of geometric approximation models corresponding to different segmentations and different geometric flows. Choosing a best basis is equivalent to finding an image's best geometric approximation model.

To compute a best basis with the fast algorithm in Section 12.2.2, a tree-structured dictionary is constructed. Each array of wavelet coefficients is divided in squares obtained with a dyadic segmentation. Figure 12.11 illustrates such a segmentation. Each square is recursively subdivided into four squares of the same size until the appropriate size is reached. This subdivision is represented by a quad-tree, where the division of a square appears as the subdivision of a node in four children nodes. The leafs of the quad-tree correspond to the squares defining the dyadic segmentation, as shown in Figure 12.11. At a scale \$2^j\$, the size \$2^i\$ of a square defines the length \$2^{j+i}\$ of its bandlets. Optimizing this segmentation is equivalent to locally adjusting this length. The resulting bandlet dictionary has a tree structure. Each node of this tree corresponds to a space \$\mathbf{W}\_d^l\$ generated by a square of \$2^{2i}\$ orthogonal wavelets at a given wavelet scale \$2^j\$ and orientation \$k\$. A bandlet orthonormal basis of \$\mathbf{W}\_d^l\$ is associated to each geometric flow.

The number of different geometric flows depends on the geometry's required precision. Suppose that the edge curve in the square is parametrized horizontally and defined by \$(x\_1, \gamma(x\_1))\$. For a piecewise \$C^\alpha\$ image, \$\gamma(x\_1)\$ is uniformly Lipschitz \$\alpha\$. Tangent vectors to the edge are \$(1, \gamma'(x\_1))\$ and \$\gamma'(x\_1)\$ is uniformly Lipschitz \$\alpha - 1\$. If \$\alpha \leq q\$, then it can be approximated by a polynomial \$\tilde{\gamma}'(x\_1)\$ of degree \$q - 2\$ with

$$\forall (x_1, x_2) \in S, \quad \|\tilde{\gamma}'(x_1) - \gamma'(x_1)\|_\infty = O(2^{i(\alpha-1)}). \quad (12.42)$$

The polynomial \$\tilde{\gamma}'(x\_1)\$ is specified by \$q - 1\$ parameters that must be quantized to limit the number of possible flows. To satisfy (12.42), these parameters are quantized with a precision \$2^{-i}\$. The total number of possible flows in the square of width \$2^i\$ is thus \$O(2^{i(q-1)})\$. A bandlet dictionary \$\mathcal{D} = \{\phi\_p\}\_{p \in \Gamma}\$ of order \$q\$ is constructed

with  $\mathbf{C}^q$  wavelets having  $q$  vanishing moments, and with polynomial flows of degree  $q - 2$ .

### Bandlet Approximation

A best  $M$ -term bandlet signal approximation is computed by finding a best basis  $\mathcal{B}_T$  and the corresponding best approximation support  $\Lambda_T$ , which minimize the  $\mathbf{l}^0$  Lagrangian

$$\mathcal{L}_0(T, f, \Lambda_T) = \mathcal{L}_0(T, f, \mathcal{B}_T) = \operatorname{argmin}_{\mathcal{B} \subset \mathcal{D}} \sum_{p \in \Gamma_{\mathcal{B}}} \min(|\langle f, \phi_p \rangle|^2, T^2). \quad (12.43)$$

This minimization chooses a best dyadic square segmentation of each wavelet coefficient array, and a best geometric flow in each square. It is implemented with the best-basis algorithm from Section 12.2.2.

An image  $\bar{f} \in \mathbf{L}^2[0, 1]^2$  is first approximated by its orthogonal projection in an approximation space  $\mathbf{V}_L$  of dimension  $N = 2^{-2L}$ . The resulting discrete signal  $f[n] = \langle \bar{f}, \phi_{L,n}^2 \rangle$  has the same wavelet coefficients as  $\bar{f}$  at scales  $2^j > 2^L$ , and thus the same bandlet coefficients at these scales. A best approximation support  $\Lambda_T$  calculated from  $f$  yields an  $M = |\Lambda_T|$  term approximation of  $\bar{f}$ :

$$\bar{f}_M(x) = \bar{f}_{\Lambda_T} = \sum_{p \in \Lambda_T} \langle \bar{f}, \phi_p \rangle \phi_p(x).$$

Theorem 12.5, proved in [342, 365], computes the nonlinear approximation error  $\|\bar{f} - \bar{f}_M\|$  for piecewise regular images.

**Theorem 12.5:** Le Pennec, Mallat, Peyré. Let  $\bar{f} \in \mathbf{L}^2[0, 1]^2$  be a piecewise  $\mathbf{C}^\alpha$  regular image. In a bandlet dictionary of order  $q \geq \alpha$ , for  $T > 0$  and  $2^L = N^{-1/2} \sim T^2$ ,

$$\mathcal{L}_0(T, f, \Lambda_T) = O(T^{2-2/(\alpha+1)}). \quad (12.44)$$

For  $M = |\Lambda_T|$ , the resulting best bandlet approximation  $\bar{f}_M$  has an error

$$\|\bar{f} - \bar{f}_M\| = O(M^{-\alpha}). \quad (12.45)$$

**Proof.** The proof finds a bandlet orthogonal basis  $\mathcal{B} = \{\phi_p\}_{p \in \Gamma_{\mathcal{B}}}$  such that

$$\mathcal{L}_0(T, f, \mathcal{B}) = \sum_{p \in \Gamma_{\mathcal{B}}} \min(|\langle f, \phi_p \rangle|^2, T^2) = O(T^{2-2/(\alpha+1)}). \quad (12.46)$$

Since  $\mathcal{L}_0(T, f, \Lambda_T) = \mathcal{L}_0(T, f, \mathcal{B}_T) \leq \mathcal{L}_0(T, f, \mathcal{B})$ , it implies (12.44). Theorem 12.1 derives in (12.5) that  $\|f - f_{\Lambda_T}\|^2 = O(M^{-\alpha})$  with  $M = O(T^{-2/(\alpha+1)})$ . A piecewise regular image has a bounded total variation, so Theorem 9.18 proves that a linear approximation error with  $N$  larger-scale wavelets has an error  $\|\bar{f} - \bar{f}_N\|^2 = O(N^{-1/2})$ . Since  $N^{-1/2} \sim T^2 = O(M^{-\alpha})$ , it results that

$$\|\bar{f} - \bar{f}_M\|^2 = \|\bar{f} - \bar{f}_N\|^2 + \|f - f_{\Lambda_T}\|^2 = O(M^{-\alpha}),$$

which proves (12.45).

We give the main ideas for constructing a bandlet basis  $\mathcal{B}$  that satisfies (12.46). Detailed derivations can be found in [365]. Following Definition 9.1, a function  $\bar{f}$  is piecewise  $\mathbf{C}^\alpha$

with a blurring scale  $s \geq 0$  if  $\tilde{f} = \tilde{f} \star h_s$  where  $\tilde{f}$  is uniformly Lipschitz  $\alpha$  on  $\Omega = [0, 1]^2 - \{e_r\}_{1 \leq r < K}$ , where the edge curves  $e_r$  are uniformly Lipschitz  $\alpha$  and do not intersect tangentially. Since  $h_s$  is a regular kernel of size  $s$ , the wavelet coefficients of  $\tilde{f}$  at a scale  $2^j$  behave as the wavelet coefficients of  $\tilde{f}$  at a scale  $2^{j'} \sim 2^j + s$  multiplied by  $s^{-1}$ . Thus, it has a marginal impact on the proof. We suppose that  $s = 0$  and consider a signal  $\tilde{f}$  that is not blurred.

Wavelet coefficients  $\langle \tilde{f}, \psi_{j,n}^k \rangle = \langle f, \psi_{j,n}^k \rangle$  are computed at scales  $2^j > 2^L = T^2$ . A dyadic segmentation of each wavelet coefficient array  $\{\langle \tilde{f}, \psi_{j,n}^k \rangle\}_n$  is computed according to Figure 12.8, at each scale  $2^j > 2^L$  and orientation  $k = 1, 2, 3$ . Wavelet arrays are divided into three types of squares. In each type of square a geometric flow is specified, so that the resulting bandlet basis  $\mathcal{B}$  has a Lagrangian that satisfies  $\mathcal{L}_o(T, f, \mathcal{B}) = O(T^{2-2/(\alpha+1)})$ . This is proved by verifying that the number of coefficients above  $T$  is  $O(T^{-2/(\alpha+1)})$  and that the energy of coefficients below  $T$  is  $O(T^{2-2/(\alpha+1)})$ .

- *Regular squares* correspond to coefficients  $\langle \tilde{f}, \psi_{j,n}^k \rangle$ , such that  $f$  is uniformly Lipschitz  $\alpha$  over the support of all  $\psi_{j,n}^k$ .
- *Edge squares* include coefficients corresponding to wavelets with support that intersects a single edge curve. This edge curve can be parametrized horizontally or vertically in each square.
- *Junction squares* include coefficients corresponding to wavelets with support that intersects at least two different edge curves.

Over regular squares, since  $\tilde{f}$  is uniformly Lipschitz  $\alpha$ , Theorem 9.15 proves in (9.15) that  $|\langle \tilde{f}, \psi_{j,n}^k \rangle| = O(2^{-j(\alpha+1)})$ . These small wavelet coefficients do not need to be retransformed and no geometric flow is defined over these squares. The number of coefficients above  $T$  in such squares is indeed  $O(T^{-2/(\alpha+1)})$  and the energy of coefficients below  $T$  is  $O(T^{2-2/(\alpha+1)})$ .

Since edges do not intersect tangentially, one can construct junction squares of width  $2^i \leq C$  where  $C$  does not depend on  $2^j$ . As a result, over the  $|\log_2 T^2|$  scales  $2^j \geq T^2$ , there are only  $O(|\log_2 T|)$  wavelet coefficients in these junction squares, which thus have a marginal impact on the approximation.

At a scale  $2^j$ , an edge square  $S$  of width  $2^i$  has  $O(2^i 2^{-j})$  large coefficients having an amplitude  $O(2^j)$  along the edge. Bandlets are created to reduce the number of these large coefficients that dominate the approximation error. Suppose that the edge curve in  $S$  is parametrized horizontally and defined by  $(x_1, \gamma(x_1))$ . Following (12.42), a geometric flow of vectors  $(1, \tilde{\gamma}'(x_1))$  is defined over the square, where  $\tilde{\gamma}'(x_1)$  is a polynomial of degree  $q-2$ , which satisfies

$$\|\tilde{\gamma}'(x_1) - \gamma'(x_1)\|_\infty = O(2^{i(\alpha-1)}).$$

Let  $w(x_1, x_2) = (x_1, x_2 - \tilde{\gamma}(x_1))$  be the warping that maps this flow to a horizontal flow, as illustrated in Figure 12.9. One can prove [365] that the warped wavelet transform satisfies

$$\left| \frac{\partial^{p_1+p_2} f \star \tilde{\psi}_j^k(w(x))}{\partial^{p_1} x_1 \partial^{p_2} x_2} \right| = O(2^j 2^{-j(p_1/\alpha+p_2)}) \text{ for any } 0 \leq p_1, p_2 \leq q.$$

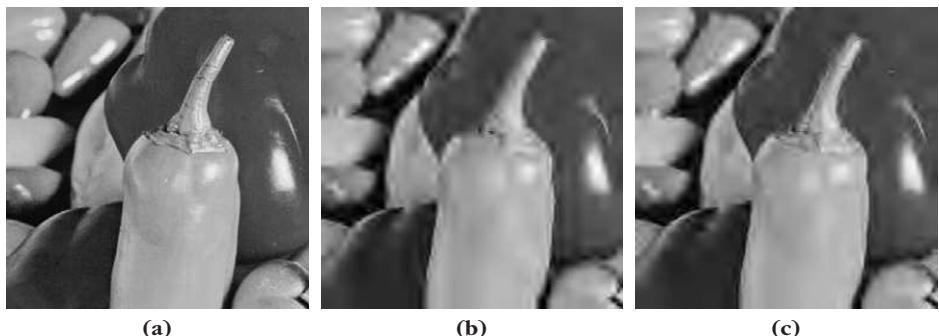
The bandlet transform takes advantage of the regularity of  $f \star \bar{\psi}_j^k(w(x))$  along  $x_1$  with Alpert directional wavelets having  $q$  vanishing moments along  $x_1$ . Computing the amplitude of the resulting bandlet coefficients shows that there are  $O(2^i T^{-2/(\alpha+1/2)})$  bandlet coefficients of amplitude larger than  $T$  and the error of all coefficients below  $T$  is  $O(2^i T^{-2/(\alpha+1)})$ . The total length of edge squares is proportional to the total length of edges in the image, which is  $O(1)$ . Summing the errors over all squares gives a total number of bandlet coefficients, which is  $O(T^{-1/(\alpha+1/2)})$ , and a total error, which is  $(O(T^{2-2/(\alpha+1)}))$ .

As a result, the bandlet basis  $\mathcal{B}$  defined over the three types of squares satisfies  $\mathcal{L}_0(T, f, \mathcal{B}) = O(T^{2-2/(\alpha+1)})$ , which finishes the proof. ■

The best-basis algorithm finds a best geometry to approximate each image. This theorem proves that the resulting approximation error decays as quickly as if the image was uniformly Lipschitz  $\alpha$  over its whole support  $[0, 1]^2$ . Moreover, this result is adaptive in the sense that it is valid for all  $\alpha \leq q$ .

The downside of bandlet approximations is the dictionary size. In a square of width  $2^i$ , we need  $O(2^{i(q-1)})$  polynomial flows, each of which defines a new bandlet family. As a result, a bandlet dictionary of order  $q$  includes  $P = O(N^{1+(q-1)^3/q})$  different bandlets. The total number of operations to compute a best bandlet approximation is  $O(P)$ , which becomes very large for  $q > 2$ . A fast implementation is described in [398] for  $q = 2$  where  $P = O(N^{3/2})$ . Theorem 12.5 is then reduced to  $\alpha \leq 2$ . It still recovers the  $O(M^{-2})$  decay for  $C^2$  images obtained in Theorem 9.19 with piecewise linear approximations over an adaptive triangulation.

Figure 12.12 shows a comparison of the approximation of a piecewise regular image with  $M$  largest orthogonal wavelet coefficients and the best  $M$  orthogonal bandlet coefficients. Wavelet approximations exhibit more ringing artifacts along edges because they do not capture the anisotropic regularity of edges.



**FIGURE 12.12**

**(a)** Original image. **(b)** Approximation with  $M/N = 1\%$  largest-amplitude wavelet coefficients (SNR = 21.8 dB). **(c)** Approximation with  $M/N = 1\%$  best bandlet vectors computed in a best bandlet basis (SNR = 23.2 dB).

### Bandlet Compression

Following the results of Section 12.1.2, a bandlet compression algorithm is implemented by quantizing the best bandlet coefficients of  $f$  with a bin size  $\Delta = 2T$ . According to Section 12.1.2, the approximation support  $\Lambda_T$  is coded on  $R_0 = M \log_2(P/N)$  bits and the amplitude of nonzero quantized coefficients with  $R_1 \sim M$  bits [365]. If  $f$  is the discretization of a piecewise  $C^\alpha$  image  $\bar{f}$ , since  $\mathcal{L}_0(T, f, \Lambda_T) = O(T^{2-2/(\alpha+1)})$ , we derive from (12.17) with  $s = (\alpha + 1)/2$  that the distortion rate satisfies

$$d(R, f) = O(R^{-\alpha} |\log_2(P/R)|^{-\alpha}).$$

Analog piecewise  $C^\alpha$  images are linearly approximated in a multiresolution space of dimension  $N$  with an error  $\|\bar{f} - \bar{f}_N\|^2 = O(N^{-1/2})$ . Taking this into account, we verify that the analog distortion rate satisfies the asymptotic decay rate (12.18)

$$d(R, \bar{f}) = O\left(R^{-\alpha} |\log_2 R|^{-\alpha}\right).$$

Although bandlet compression improves the asymptotic decay of wavelet compression, such coders are not competitive with a JPEG-2000 wavelet image coder, which requires less computations. Moreover, when images have no geometric regularity, despite the fact that the decay rate is the same as with wavelets, bandlets introduce an overhead because of the large dictionary size.

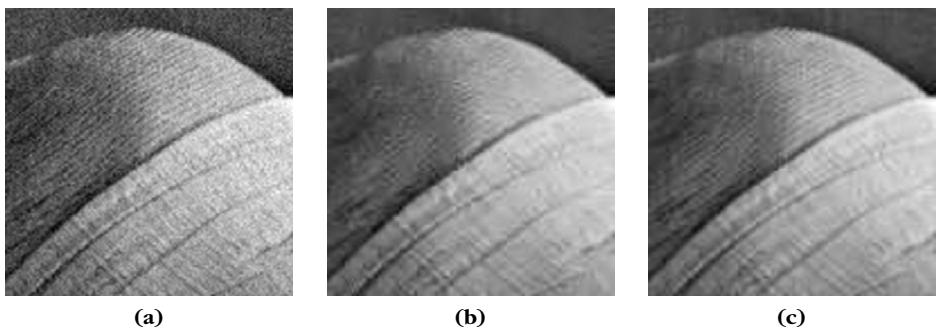
### Bandlet Denoising

Let  $W$  be a Gaussian white noise of variance  $\sigma^2$ . To estimate  $f$  from  $X = f + W$ , a best bandlet estimator  $\tilde{F} = X_{\tilde{\Lambda}_T}$  is computed according to Section 12.2.1 by projecting  $X$  on an optimized family of orthogonal bandlets indexed by  $\tilde{\Lambda}_T$ . It is obtained by thresholding at  $T$  the bandlet coefficients of  $X$  in the best bandlet basis  $\tilde{\mathcal{B}}_T$ , which minimizes  $\mathcal{L}_0(T, X, \mathcal{B})$ , for  $T = \sigma \sqrt{2 \log_e P}$ .

An analog estimator  $\bar{F}$  of  $\bar{f}$  is reconstructed from the noisy signal coefficients in  $\tilde{\Lambda}_T$  with the analog bandlets  $\{\phi_p(x)\}_{p \in \tilde{\Lambda}_T}$ . If  $\bar{f}$  is a piecewise  $C^\alpha$  image  $\bar{f}$ , then Theorem 12.5 proves that  $\mathcal{L}_0(T, f, \Lambda_T) = O(T^{2-2/(\alpha+1)})$ . The computed risk decay (12.27) thus applies for  $s = (\alpha + 1)/2$ :

$$E\{\|\tilde{F} - \bar{f}\|^2\} = O(\sigma^{2-2/(\alpha+1)} |\log \sigma|^{2-2/(\alpha+1)}). \quad (12.47)$$

This decay rate [233] shows that a bandlet estimation over piecewise  $C^\alpha$  images nearly reaches the minimax risk  $r_n \sim \sigma^{2-2/(\alpha+1)}$  calculated in (11.152) for uniformly  $C^\alpha$  images. Figure 12.13 gives a numerical example comparing a best bandlet estimation and a translation-invariant wavelet thresholding estimator for an image including regular geometric structures. The threshold is  $T = 3\sigma$  instead of  $T = \sigma \sqrt{2 \log_e P}$ , because it improves the SNR.

**FIGURE 12.13**

**(a)** Noisy image (SNR = 22 db). **(b)** Translation-invariant wavelet hard thresholding (SNR = 25.3 db). **(c)** Best bandlet thresholding estimation (SNR = 26.4 db).

## 12.3 GREEDY MATCHING PURSUITS

Computing an optimal  $M$ -term approximation  $f_M$  of a signal  $f$  with  $M$  vectors selected in a redundant dictionary  $\mathcal{D}$  is NP-hard. Pursuit strategies construct nonoptimal yet efficient approximations with computational algorithms. Matching pursuits are greedy algorithms that select the dictionary vectors one by one, with applications to compression, denoising, and pattern recognition.

### 12.3.1 Matching Pursuit

Matching pursuit introduced by Mallat and Zhang [366] computes signal approximations from a redundant dictionary, by iteratively selecting one vector at a time. It is related to projection pursuit algorithms used in statistics [263] and to shape-gain vector quantizations [27].

Let  $\mathcal{D} = \{\phi_p\}_{p \in \Gamma}$  be a dictionary of  $P > N$  vectors having a unit norm. This dictionary is supposed to be complete, which means that it includes  $N$  linearly independent vectors that define a basis of the signal space  $\mathbb{C}^N$ . A matching pursuit begins by projecting  $f$  on a vector  $\phi_{p_0} \in \mathcal{D}$  and by computing the residue  $Rf$ :

$$f = \langle f, \phi_{p_0} \rangle \phi_{p_0} + Rf. \quad (12.48)$$

Since  $Rf$  is orthogonal to  $\phi_{p_0}$ ,

$$\|f\|^2 = |\langle f, \phi_{p_0} \rangle|^2 + \|Rf\|^2. \quad (12.49)$$

To minimize  $\|Rf\|$ , we must choose  $\phi_{p_0} \in \mathcal{D}$  such that  $|\langle f, \phi_{p_0} \rangle|$  is maximum. In some cases it is computationally more efficient to find a vector  $\phi_{p_0}$  that is almost optimal:

$$|\langle f, \phi_{p_0} \rangle| \geq \alpha \sup_{p \in \Gamma} |\langle f, \phi_p \rangle|, \quad (12.50)$$

where  $\alpha \in (0, 1]$  is a relaxation factor. The pursuit iterates this procedure by sub-decomposing the residue. Let  $R^0f = f$ . Suppose that the  $m$ th-order residue  $R^m f$  is already computed for  $m \geq 0$ . The next iteration chooses  $\phi_{p_m} \in \mathcal{D}$  such that

$$|\langle R^m f, \phi_{p_m} \rangle| \geq \alpha \sup_{p \in \Gamma} |\langle R^m f, \phi_p \rangle|, \quad (12.51)$$

and projects  $R^m f$  on  $\phi_{p_m}$ :

$$R^m f = \langle R^m f, \phi_{p_m} \rangle \phi_{p_m} + R^{m+1} f. \quad (12.52)$$

The orthogonality of  $R^{m+1} f$  and  $\phi_{p_m}$  implies

$$\|R^m f\|^2 = |\langle R^m f, \phi_{p_m} \rangle|^2 + \|R^{m+1} f\|^2. \quad (12.53)$$

Summing (12.52) from  $m$  between 0 and  $M - 1$  yields

$$f = \sum_{m=0}^{M-1} \langle R^m f, \phi_{p_m} \rangle \phi_{p_m} + R^M f. \quad (12.54)$$

Similarly, summing (12.53) from  $m$  between 0 and  $M - 1$  gives

$$\|f\|^2 = \sum_{m=0}^{M-1} |\langle R^m f, \phi_{p_m} \rangle|^2 + \|R^M f\|^2. \quad (12.55)$$

### **Convergence of Matching Pursuit**

A matching pursuit has an exponential decay if the residual  $\|R^m f\|$  has a minimum rate of decay. The conservation of energy (12.53) implies

$$\frac{\|R^{m+1} f\|^2}{\|R^m f\|^2} = 1 - \left| \left\langle \frac{R^m f}{\|R^m f\|}, \phi_{p_m} \right\rangle \right|^2 \leq 1 - \mu^2(R^m f, \mathcal{D}), \quad (12.56)$$

where  $\mu(r, \mathcal{D})$  is the coherence of a vector relative to the dictionary, defined by

$$\mu(r, \mathcal{D}) = \max_{p \in \Gamma} \left| \left\langle \frac{r}{\|r\|}, \phi_p \right\rangle \right| \leq 1.$$

Theorem 12.6 proves that

$$\mu_{\min}(\mathcal{D}) = \inf_{r \in \mathbb{C}^N, r \neq 0} \mu(r, \mathcal{D}) > 0,$$

and thus that matching pursuits converge exponentially.

**Theorem 12.6.** The residual  $R^m f$  computed by a matching pursuit with relaxation parameter  $\alpha \in (0, 1]$  satisfies

$$\|R^m f\|^2 \leq (1 - \alpha^2 \mu_{\min}(\mathcal{D})^2)^m \|f\|^2 \quad \text{with} \quad 1 \geq \mu_{\min}(\mathcal{D}) > 0. \quad (12.57)$$

As a consequence,

$$f = \sum_{m=0}^{+\infty} \langle R^m f, \phi_{p_m} \rangle \phi_{p_m}, \quad \text{and} \quad \|f\|^2 = \sum_{m=0}^{+\infty} |\langle R^m f, \phi_{p_m} \rangle|^2. \quad (12.58)$$

**Proof.** The atom  $\phi_{p_m}$  selected by a matching pursuit satisfies  $|\langle R^m f, \phi_{p_m} \rangle| \geq \alpha \sup_{p \in \Gamma} |\langle R^m f, \phi_p \rangle|$ . It results from (12.56) that

$$\frac{\|R^{m+1} f\|^2}{\|R^m f\|^2} \leq 1 - \alpha^2 \mu_{\min}^2(\mathcal{D}).$$

Iterating on this equation proves that

$$\|R^m f\|^2 \leq (1 - \alpha^2 \mu_{\min}^2(\mathcal{D}))^m \|f\|^2. \quad (12.59)$$

To verify that  $\mu_{\min}(\mathcal{D}) > 0$ , a contrario lets us suppose that  $\mu_{\min}(\mathcal{D}) = 0$ . There exist  $\{f_m\}_{m \in \mathbb{N}}$  with  $\|f_m\| = 1$  such that

$$\lim_{m \rightarrow +\infty} \sup_{p \in \Gamma} |\langle f_m, \phi_p \rangle| = 0. \quad (12.60)$$

Since the unit sphere of  $\mathbb{C}^N$  is compact, there exists a subsequence  $\{f_{m_k}\}_{k \in \mathbb{N}}$  that converges to a unit vector  $f \in \mathbb{C}^N$ . It follows that

$$\sup_{p \in \Gamma} |\langle f, \phi_p \rangle| = \lim_{k \rightarrow +\infty} \sup_{p \in \Gamma} |\langle f_{m_k}, \phi_p \rangle| = 0, \quad (12.61)$$

so  $\langle f, \phi_p \rangle = 0$  for all  $\phi_p \in \mathcal{D}$ . Since  $\mathcal{D}$  contains a basis of  $\mathbb{C}^N$ , necessarily  $f = 0$ , which is not possible because  $\|f\| = 1$ . It results that, necessarily,  $\mu_{\min}(\mathcal{D}) > 0$ .

This proves that  $1 - \alpha^2 \mu_{\min}^2(\mathcal{D}) < 1$  and thus that  $\lim_{m \rightarrow +\infty} \|R^m f\| = 0$ . Inserting this in (12.54) and (12.55) proves (12.58). ■

Matching pursuits often converge more slowly when the size  $N$  of the signal space increases because  $\mu_{\min}(\mathcal{D})$  can become close to 0. In the limit of infinite-dimensional spaces, Jones' theorem proves that the matching pursuit still converges but the convergence is not exponential [319, 366]. Section 12.3.2 describes an orthogonalized matching pursuit that converges in fewer than  $N$  iterations.

### Backprojection

A matching pursuit computes an approximation  $\tilde{f}_M = \sum_{m=0}^{M-1} \langle R^m f, \phi_{p_m} \rangle \phi_{p_m}$  that belongs to space  $\mathbf{V}_M$  generated by  $M$  vectors  $\{\phi_{p_m}\}_{0 \leq m < M}$ . However, in general  $\tilde{f}_M$  is not equal to the orthogonal projection  $f_M$  on  $f$  in  $\mathbf{V}_M$ , and thus  $\|f - \tilde{f}_M\| \geq \|f - f_M\|$ . In finite dimension, an infinite number of matching pursuit iterations is typically necessary to completely remove the error  $\|f - \tilde{f}_M\|$ , although in most applications this approximation error becomes sufficiently small for  $M \ll N$ . To reduce the matching pursuit error, Mallat and Zhang [366] introduced a backprojection that computes the coefficients  $\tilde{a}[m]$  of the orthogonal projection

$$f_M = \sum_{m=0}^{M-1} \tilde{a}[m] \phi_{p_m}.$$

Let  $y[m] = \langle f, \phi_{p_m} \rangle$ . Section 5.1.3 shows that the decomposition coefficients of this dual-analysis problem are obtained by inverting the Gram operator

$$\tilde{a} = L^{-1}y \quad \text{with} \quad La[m] = \sum_{n=0}^{M-1} a[n] \langle \phi_{p_n}, \phi_{p_m} \rangle.$$

This inversion can be computed with a conjugate-gradient algorithm or with a Richardson gradient descent from the initial coefficients  $a_0[m] = \langle R^m f, \phi_{p_m} \rangle$  provided by the matching pursuit. Let  $\gamma$  be a relaxation parameter that satisfies

$$\delta = \max \{ |1 - \gamma A_M|, |1 - \gamma B_M| \} < 1,$$

where  $B_M \geq A_M > 0$  are the frame bounds of  $\{\phi_{p_m}\}_{0 \leq m < N}$  in  $\mathbf{V}_M$ . Theorem 5.7 proves that

$$a_k = a_{k-1} + \gamma (y - La_{k-1})$$

converges to the solution:  $\lim_{k \rightarrow +\infty} a_k = \tilde{a}$ . A safe choice is  $\gamma = 2/B$  where  $B \geq B_M$  is the upper frame bound of the overall dictionary  $\mathcal{D}$ .

### **Fast Network Calculations**

A matching pursuit is implemented with a fast algorithm that computes  $\langle R^{m+1} f, \phi_p \rangle$  from  $\langle R^m f, \phi_p \rangle$  with an *updating* formula. Taking an inner product with  $\phi_p$  on each side of (12.52) yields

$$\langle R^{m+1} f, \phi_p \rangle = \langle R^m f, \phi_p \rangle - \langle R^m f, \phi_{p_m} \rangle \langle \phi_{p_m}, \phi_p \rangle. \quad (12.62)$$

In neural network language, this is an inhibition of  $\langle R^m f, \phi_p \rangle$  by the selected pattern  $\phi_{p_m}$  with a weight  $\langle \phi_{p_m}, \phi_p \rangle$  that measures its correlation with  $\phi_p$ . To reduce the computational load, it is necessary to construct dictionaries with vectors having a sparse interaction. This means that each  $\phi_p \in \mathcal{D}$  has nonzero inner products with only a small fraction of all other dictionary vectors. It can also be viewed as a network that is not fully connected. Dictionaries can be designed so that nonzero weights  $\langle \phi_\alpha, \phi_p \rangle$  are retrieved from memory or computed with  $O(1)$  operations. A matching pursuit with a relative precision  $\varepsilon$  is implemented with the following steps:

**1. Initialization.** Set  $m = 0$  and compute  $\{\langle f, \phi_p \rangle\}_{p \in \Gamma}$  in  $\mathcal{D}$ .

**2. Best match.** Find  $\phi_{p_m} \in \mathcal{D}$  such that

$$|\langle R^m f, \phi_{p_m} \rangle| = \max_{p \in \Gamma} |\langle R^m f, \phi_p \rangle|. \quad (12.63)$$

**3. Update.** For all  $\phi_p \in \mathcal{D}$  with  $\langle \phi_{p_m}, \phi_p \rangle \neq 0$ ,

$$\langle R^{m+1} f, \phi_p \rangle = \langle R^m f, \phi_p \rangle - \langle R^m f, \phi_{p_m} \rangle \langle \phi_{p_m}, \phi_p \rangle. \quad (12.64)$$

**4. Stopping rule.** If

$$\|R^{m+1}f\|^2 = \|R^m f\|^2 - |\langle R^m f, \phi_{p_m} \rangle|^2 \leq \varepsilon^2 \|f\|^2,$$

then stop. Otherwise,  $m = m + 1$  and go to 2.

If  $\mathcal{D}$  is very redundant, computations at steps 1, 2, and 3 are reduced by performing the calculations in a subdictionary  $\mathcal{D}_\Delta = \{\phi_p\}_{p \in \Gamma_\Delta} \subset \mathcal{D}$ . The subdictionary  $\mathcal{D}_\Delta$  is constructed so that if  $\tilde{\phi}_{p_m} = \operatorname{argmax}_{\phi_p \in \mathcal{D}_\Delta} |\langle R^m f, \phi_p \rangle|$ , then

$$|\langle R^m f, \tilde{\phi}_{p_m} \rangle| \geq \alpha \max_{p \in \Gamma} |\langle R^m f, \phi_p \rangle|. \quad (12.65)$$

The selected atom  $\tilde{\phi}_{p_m} \in \mathcal{D}_\Delta$  is improved with a local search in the larger dictionary  $\mathcal{D}$ , among all atoms  $\phi_p$  “close” to  $\tilde{\phi}_{p_m}$ , in the sense that  $|\langle \phi_p, \tilde{\phi}_{p_m} \rangle| > C$  for a pre-defined constant  $C$ . This local search finds  $\phi_{p_m}$ , which locally maximizes the residue correlation

$$|\langle R^m f, \phi_{p_m} \rangle| = \max_{p \in \Gamma, |\langle \phi_p, \phi_{p_m} \rangle| > C} |\langle R^m f, \phi_p \rangle|.$$

The updating (12.64) is restricted to vectors  $\phi_p \in \mathcal{D}_\Delta$ . The construction of hierarchical dictionaries can also reduce the calculations needed to compute inner products in  $\mathcal{D}$  from inner products in  $\mathcal{D}_\Delta$  [387].

The dictionary must incorporate important signal features, which depend on the signal class. Section 12.3.3 studies dictionaries of Gabor atoms. Section 12.3.4 describes applications to noise reduction. Specific dictionaries for inverse electromagnetic problems, face recognition, and data compression are constructed in [80, 374, 399]. Dictionary learning is studied in Section 12.7.

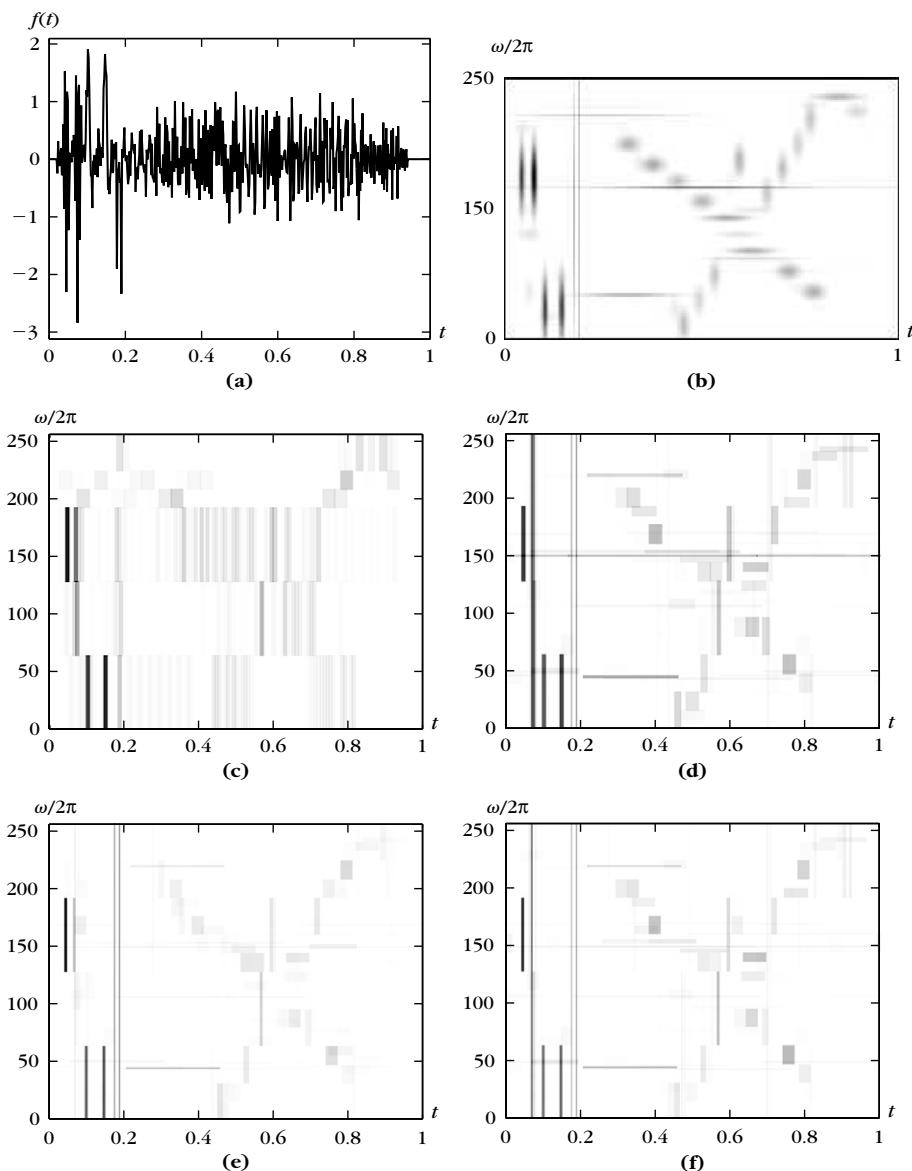
### Wavelet Packets and Local Cosines Dictionaries

Wavelet packet and local cosine trees constructed in Sections 8.2.1 and 8.5.3 are dictionaries with  $P = N \log_2 N$  vectors. For each dictionary vector, there are few other dictionary vectors having nonzero inner products that can be stored in tables to compute the updating formula (12.64). Each matching pursuit iteration then requires  $O(N \log_2 N)$  operations.

In a dictionary of wavelet packet bases calculated with a Daubechies 8 filter, the best basis shown in Figure 12.14(c) optimizes the division of the frequency axis, but it has no flexibility in time. It is, therefore, not adapted to the time evolution of the signal components. The matching pursuit flexibility adapts the wavelet packet choice to local signal structures; Figure 12.14(d) shows that it better reveals its time-frequency properties than the best wavelet packet basis.

### Translation Invariance

Representing a signal structure independently from its location is a form of translation invariance that is important for pattern recognition. Decompositions in

**FIGURE 12.14**

**(a)** Signal synthesized with a sum of chirps, truncated sinusoids, short time transients, and Diracs. The time-frequency images display the atoms selected by different adaptive time-frequency transforms. The darkness is proportional to the coefficient amplitude. **(b)** Gabor matching pursuit. Each dark blob is the Wigner-Ville distribution of a selected Gabor atom. **(c)** Heisenberg boxes of a best wavelet packet basis calculated with a Daubechies 8 filter. **(d)** Wavelet packets selected by a matching pursuit. **(e)** Wavelet packets of a basis pursuit. **(f)** Wavelet packets of an orthogonal matching pursuit.

orthonormal bases lack this translation invariance. Matching pursuits are translation invariant if calculated in translation-invariant dictionaries. A dictionary  $\mathcal{D}$  is *translation invariant* if for any  $\phi_p \in \mathcal{D}$ , then  $\phi_p[n-p] \in \mathcal{D}$  for  $0 \leq p < N$ . Suppose that the matching decomposition of  $f$  in  $\mathcal{D}$  is [201]

$$f[n] = \sum_{m=0}^{M-1} \langle R^m f, \phi_{p_m} \rangle \phi_{p_m}[n] + R^M f[n]. \quad (12.66)$$

Then the matching pursuit of  $f_p[n] = f[n-p]$  selects a translation by  $p$  of the same vectors  $\phi_{p_m}$  with the same decomposition coefficients

$$f_p[n] = \sum_{m=0}^{M-1} \langle R^m f, \phi_{p_m} \rangle \phi_{p_m}[n-p] + R^M f_p[n].$$

Thus, patterns can be characterized independently of their position.

Translation invariance is generalized as an invariance with respect to any group action [201]. A frequency translation is another example of a group operation. If the dictionary is invariant under the action of a group, then the pursuit remains invariant under the action of the same group. Section 12.3.3 gives an example of a Gabor dictionary, which is translation invariant in time and frequency.

### 12.3.2 Orthogonal Matching Pursuit

Matching pursuit approximations are improved by orthogonalizing the directions of projection with a Gram-Schmidt procedure. The resulting orthogonal pursuit converges with a finite number of iterations. This orthogonalization was introduced by Mallat and Zhang together with the nonorthogonal pursuit algorithm in Zhang thesis [74]. The higher computational cost of the Gram-Schmidt algorithm may seem discouraging (reviewers suppressed it from the first publication in [366]), but the improved precision of this orthogonalization becomes important for the inverse problems studied in Chapter 13. It appeared in [202] and was proposed independently by Pati, Rezaifar, and Krishnaprasad [395].

The vector  $\phi_{p_m}$  selected by the matching algorithm is a priori not orthogonal to the previously selected vectors  $\{\phi_{p_l}\}_{0 \leq l < m}$ . When subtracting the projection of  $R^m f$  over  $\phi_{p_m}$ , the algorithm reintroduces new components in the directions of  $\{\phi_{p_l}\}_{0 \leq l < m}$ . This is avoided by projecting the residues on an orthogonal family  $\{u_l\}_{0 \leq l < m}$  computed from  $\{\phi_{p_l}\}_{0 \leq l < m}$ .

Let us initialize  $u_0 = \phi_{p_0}$ . For  $m \geq 0$ , an orthogonal matching pursuit selects  $\phi_{p_m}$  that satisfies

$$|\langle R^m f, \phi_{p_m} \rangle| \geq \alpha \sup_{p \in \Gamma} |\langle R^m f, \phi_p \rangle|. \quad (12.67)$$

The Gram-Schmidt algorithm orthogonalizes  $\phi_{p_m}$  with respect to  $\{\phi_{p_l}\}_{0 \leq l < m}$  and defines

$$u_m = \phi_{p_m} - \sum_{l=0}^{m-1} \frac{\langle \phi_{p_m}, u_l \rangle}{\|u_l\|^2} u_l. \quad (12.68)$$

The residue  $R^m f$  is projected on  $u_m$  instead of  $\phi_{p_m}$ :

$$R^m f = \frac{\langle R^m f, u_m \rangle}{\|u_m\|^2} u_m + R^{m+1} f. \quad (12.69)$$

Summing this equation for  $0 \leq m < k$  yields

$$\begin{aligned} f &= \sum_{m=0}^{k-1} \frac{\langle R^m f, u_m \rangle}{\|u_m\|^2} u_m + R^k f \\ &= P_{V_k} f + R^k f, \end{aligned} \quad (12.70)$$

where  $P_{V_k}$  is the orthogonal projector on the space  $V_k$  generated by  $\{u_m\}_{0 \leq m < k}$ . The Gram-Schmidt algorithm ensures that  $\{\phi_{p_m}\}_{0 \leq m < k}$  is also a basis of  $V_k$ . For any  $k \geq 0$  the residue  $R^k f$  is the component of  $f$  that is orthogonal to  $V_k$ . For  $m = k$ , (12.68) implies that

$$\langle R^m f, u_m \rangle = \langle R^m f, \phi_{p_m} \rangle. \quad (12.71)$$

Since  $V_k$  has dimension  $k$  there exists  $M \leq N$ , but most often  $M = N$ , such that  $f \in V_M$ , so  $R^M f = 0$  and inserting (12.71) in (12.70) for  $k = M$  yields

$$f = \sum_{m=0}^{M-1} \frac{\langle R^m f, \phi_{p_m} \rangle}{\|u_m\|^2} u_m. \quad (12.72)$$

The algorithm stops after  $M \leq N$  iterations. The energy conservation resulting from this decomposition in a family of orthogonal vectors is

$$\|f\|^2 = \sum_{m=0}^{M-1} \frac{|\langle R^m f, \phi_{p_m} \rangle|^2}{\|u_m\|^2}. \quad (12.73)$$

The exponential convergence rate of the matching pursuit in Theorem 12.6 remains valid for an orthogonal matching pursuit, but it also converges in less than  $N$  iterations.

To expand  $f$  over the original dictionary vectors  $\{\phi_{p_m}\}_{0 \leq m < M}$ , we must perform a change of basis. The triangular Gram-Schmidt relations (12.68) are inverted to expand  $u_m$  in  $\{\phi_{p_k}\}_{0 \leq k \leq m}$ :

$$u_m = \sum_{k=0}^m b[k, m] \phi_{p_k}. \quad (12.74)$$

Inserting this expression into (12.72) gives

$$f = \sum_{k=0}^{M-1} a[p_k] \phi_{p_k}, \quad (12.75)$$

with

$$a[p_k] = \sum_{m=k}^{M-1} b[k, m] \frac{\langle R^m f, \phi_{p_m} \rangle}{\|u_m\|^2}.$$

The Gram-Schmidt summation (12.68) must be carefully implemented to avoid numerical instabilities [29]. A Gram-Schmidt orthogonalization of  $M$  vectors requires  $O(NM^2)$  operations. In wavelet packet, local cosine, and Gabor dictionaries,  $M$  matching pursuit iterations are calculated with  $O(MN \log_2 N)$  operations.

For  $M$  large, the Gram-Schmidt orthogonalization very significantly increases the computational complexity of a matching pursuit. A final matching pursuit orthogonal backprojection requires at most  $O(M^3)$  operators, but both algorithms may not give the same results because they do not necessarily select the same vectors. An orthogonal pursuit can improve the approximation precision as shown in Sections 13.3 and 13.4 for the resolution of inverse problems.

Figure 12.14(f) displays the wavelet packets selected by an orthogonal matching pursuit. A comparison with Figure 12.14(d) shows that the orthogonal and nonorthogonal pursuits select nearly the same wavelet packets having a high-amplitude inner product. These vectors are called *coherent structures*. They are selected during the first few iterations. A mathematical interpretation of these coherent structures is given in Section 12.5.2. Most often, during the first few iterations, a matching pursuit selects nearly orthogonal vectors, so orthogonal and nonorthogonal pursuits are nearly identical. When the number of iterations increases and gets close to  $N$ , the residues of an orthogonal pursuit have norms that decrease faster than for a nonorthogonal pursuit. For large-size signals, where the number of iterations is a small fraction of  $N$ , the nonorthogonal pursuit is more often used, but the orthogonalization or a backprojection becomes important if a high-approximation precision is needed.

### 12.3.3 Gabor Dictionaries

Gabor dictionaries are constructed with Gaussian windows, providing optimal time and frequency energy concentration. For images, directional Gabor dictionaries lead to efficient representations, particularly for video compression.

#### *Time-Frequency Gabor Dictionary*

A time and frequency translation-invariant Gabor dictionary is constructed by Qian and Chen [405] as well as Mallat and Zhang [366], by scaling, modulating, and translating a Gaussian window on the signal-sampling grid. For each scale  $2^j$ , a discrete Gaussian window is defined by

$$g_j[n] = K_j 2^{-j/2+1/4} \exp(-\pi(2^{-j}n)^2), \quad (12.76)$$

where the constant  $K_j \approx 1$  is adjusted so that  $\|g_j\| = 1$ . A Gabor time-frequency frame is derived with time intervals  $u_j = 2^j \Delta^{-1}$  and frequency intervals  $\xi_j = 2\pi\Delta^{-1}2^{-j}$ :

$$\mathcal{D}_{j,\Delta} = \left\{ \phi_p[n] = g_j[n - qu_j] \exp(i\xi_j kn) \right\}_{0 \leq q < \Delta N 2^{-j}, 0 \leq k < \Delta 2^j}. \quad (12.77)$$

It includes  $P = \Delta^2 N$  vectors. Asymptotically for  $N$  large, this family of Gabor signals has the same properties as the frames of the Gabor functions studied in Section 5.4.

Theorem 5.19 proves that a necessary condition to obtain a frame is that  $u_j \xi_j = 2\pi\Delta^{-2} < 2\pi$ , and thus  $\Delta > 1$ . Table 5.3 shows that for  $\Delta \geq 2$ , this Gabor dictionary is nearly a tight frame with  $A \approx B \approx \Delta^2$ .

A multiscale Gabor dictionary is a union of such tight frames

$$\mathcal{D}_\Delta = \bigcup_{j=k}^{\log_2 N - k} \mathcal{D}_{j,\Delta}, \quad (12.78)$$

with typically  $k \geq 2$  to avoid having too-small or too-large windows. Its size is thus  $P \leq \Delta^2 N \log_2 N$ , and for  $\Delta \geq 2$  it is nearly a tight frame with frame bounds  $\Delta^2 \log_2 N$ . A translation-invariant dictionary is a much larger dictionary obtained by setting  $u_j = 1$  and  $\xi_j = 2\pi/N$  in (12.77), and it thus includes  $P \approx N^2 \log_2 N$  vectors.

A matching pursuit decomposes real signals in the multiscale Gabor dictionary (12.78) by grouping atoms  $\phi_{p^+}$  and  $\phi_{p^-}$  with  $p^\pm = (qu_j, \pm k\xi_j, 2^j)$ . At each iteration, instead of projecting  $R^m f$  over an atom  $\phi_p$ , the matching pursuit computes its projection on the plane generated by  $(\phi_{p^+}, \phi_{p^-})$ . Since  $R^m f[n]$  is real, one can verify that this is equivalent to projecting  $R^m f$  on a real vector that can be written as

$$\phi_p^\gamma[n] = K_{j,\gamma} g_j[n - qu_j] \cos(k\xi_j n + \gamma).$$

The constant  $K_{j,\gamma}$  sets the norm of this vector to 1 and the phase  $\gamma$  is optimized to maximize the inner product with  $R^m f$ . Matching pursuit iterations yield

$$f = \sum_{m=0}^{+\infty} \langle R^m f, \phi_{p_m}^{\gamma_m} \rangle \phi_{p_m}^{\gamma_m}. \quad (12.79)$$

The time-frequency signal geometry is characterized by the time-frequency and scale support  $\Lambda_M = \{p_m = (q_m u_{j_m}, k \xi_{j_m}, 2^{j_m})\}_{0 \leq m \leq M}$  of the  $M$  selected Gabor atoms. It is more easily visualized with a time-frequency energy distribution obtained by summing the Wigner-Ville distribution  $P_V \phi_{p_m}[n, k]$  of the complex atoms  $\phi_{p_m}$ :

$$P_M f[n, k] = \sum_{m=0}^{+\infty} |\langle R^m f, \phi_{p_m}^{\gamma_m} \rangle|^2 P_V \phi_{p_m}[n, k]. \quad (12.80)$$

Since the window is Gaussian,  $P_V \phi_{p_m}$  is a two-dimensional Gaussian blob centered at  $(q_m u_{j_m}, k_m \xi_{j_m})$  in the time-frequency plane. It is scaled by  $2^{j_m}$  in time and by  $N 2^{-j_m}$  in frequency.

### Computations

A matching pursuit in a translation-invariant Gabor dictionary of size  $P = N^2 \log_2 N$  is implemented by restricting most computations in the multiscale dictionary  $\mathcal{D}_\Delta$  of smaller size  $P \leq 4N \log_2 N$  for  $\Delta = 2$ . At each iteration, a Gabor atom  $\phi_{\tilde{p}_m}$  that best matches  $R^m f$  is selected in  $\mathcal{D}_\Delta$ . The position and frequency of this atom are

then refined with (12.65). It finds a close time-frequency atom  $\phi_{p_m}$  in the larger translation-invariant dictionary, which has a better correlation with  $R^m f$ . The inner product update (12.64) is then computed only for atoms in  $\mathcal{D}_\Delta$ , with an analytical formula. Two Gabor atoms that have positions, frequencies, and scales  $p_1 = (u_1, \xi_1, s_1)$  and  $p_2 = (u_2, \xi_2, s_2)$  have an inner product

$$\begin{aligned} \langle g_{p_1}, g_{p_2} \rangle &= \frac{\sqrt{2s_1 s_2}}{\sqrt{s_1^2 + s_2^2}} \exp\left(-\frac{i(s_1^2 u_2 + s_2^2 u_1)(\xi_2 - \xi_1) + \pi(u_2 - u_1)^2}{s_1^2 + s_2^2} - \frac{(\xi_2 - \xi_1)^2}{4\pi(s_1^{-2} + s_2^{-2})}\right) \\ &\quad + O\left(\exp\left(\frac{-\pi N^2}{s_1^2 + s_2^2}\right) + \exp\left(\frac{-\pi}{s_1^{-2} + s_2^{-2}}\right)\right). \end{aligned} \quad (12.81)$$

The error terms can be neglected if the scales  $s_1$  and  $s_2$  are not too small or too close to  $N$ . The resulting matching pursuit in a translation-invariant Gabor dictionary requires marginally more computations than a matching pursuit in  $\mathcal{D}_\Delta$ .

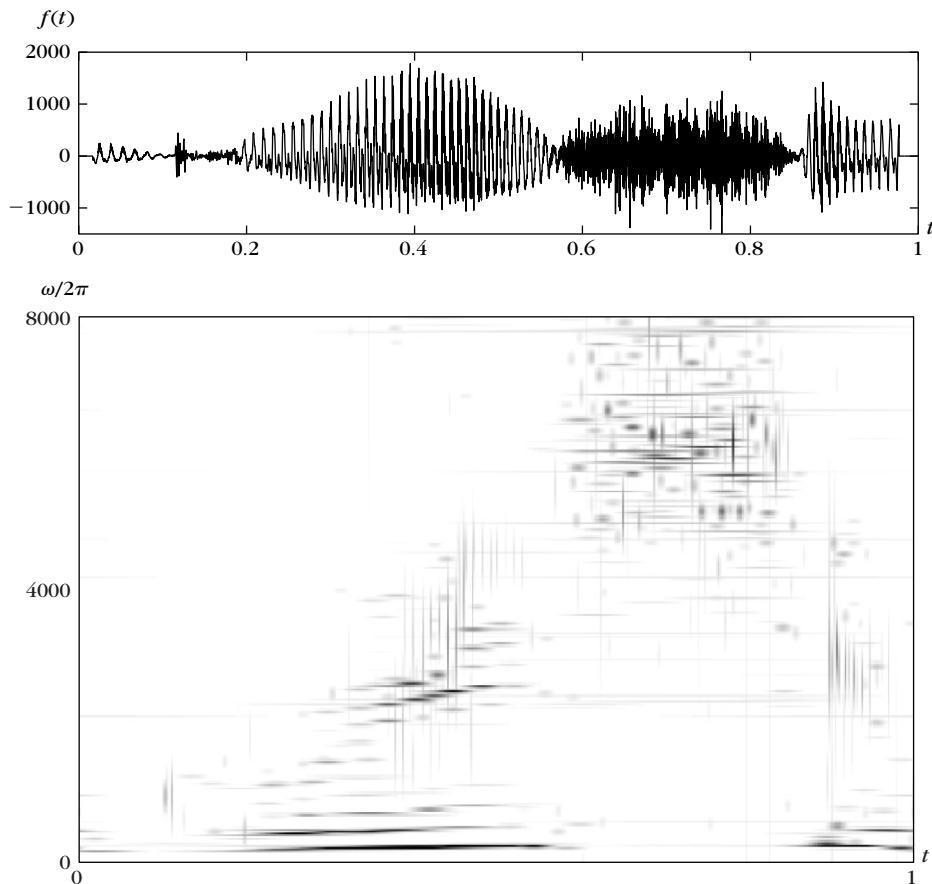
Figure 12.14(b) shows the matching pursuit decomposition of a signal having localized time-frequency structures. This representation is more sparse than the matching pursuit decomposition in the wavelet packet dictionary shown in Figure 12.14(d). Indeed, Gabor dictionary atoms are translated on a finer time-frequency grid than wavelet packets, and they have a better time-frequency localization. As a result, the matching pursuits find Gabor atoms that better match the signal structures.

Figure 12.15 gives the Gabor matching pursuit decomposition of the word “greasy,” sampled at 16 kHz. The time-frequency energy distribution shows the low-frequency component of the “g” and the quick-burst transition to the “ea.” The “ea” has many harmonics that are lined up. The “s” is a noise with a time-frequency energy spread over a high-frequency interval. Most of the signal energy is characterized by a few time-frequency atoms. For  $m = 250$  atoms,  $\|R^m f\|/\|f\| = 0.169$ , even though the signal has 5782 samples, and the sound recovered from these atoms is of good audio quality.

Matching pursuits in Gabor dictionaries provide sparse representation of oscillatory signals, with frequency and scale parameters that are used to characterize the signal structures. For example, studies have been carried in cognitive neurophysiology for the analysis of gamma and high-gamma oscillations in electroencephalogram (EEG) signals [406], which are highly nonstationary. Matching pursuit decompositions are also used to predict epilepsy patterns [22], allowing physicians to identify periods of seizure initiation by analyzing the selected atom properties [259, 320].

In Figure 12.14(b), the two chirps with frequencies that increase and decrease linearly are decomposed in many Gabor atoms. To improve the representation of signals having time-varying spectral lines, the dictionary can include Gabor chirps having an instantaneous frequency that varies linearly in time:

$$\phi_p[n] = g_j[n - qu_j] \exp(i\xi_j(k + cn)n).$$

**FIGURE 12.15**

Speech recording of the word “greasy” sampled at 16 kHz. In the time-frequency image, the dark blobs of various sizes are the Wigner-Ville distributions of Gabor functions selected by the matching pursuit.

Their Wigner-Ville distribution  $P_V \phi_P[n, k]$  is localized around an oriented segment in the time-frequency plane. Such atoms can more efficiently represent progressive frequency variations of the signal spectral components. However, increasing the dictionary size also increases intermediate memory storage and computational complexity. To incorporate Gabor chirps, Gribonval [278] reduces the matching pursuit complexity by first optimizing the scale-time-frequency parameters  $(2^j, q, k)$  for  $c = 0$ , and then adjusting  $c$  instead of jointly optimizing all parameters.

#### ***Directional Image Gabor Dictionaries***

A sparse representation of image structures such as edges, corners, and textures requires using a large dictionary of vectors. Section 5.5.1 describes redundant

dictionaries of directional wavelets and curvelets. Matching pursuit decompositions over two-dimensional directional Gabor wavelets are introduced in [105]. They are constructed with a separable product of Gaussian windows  $g_j[n]$  in (12.76), with angle directions  $\theta = k\pi/C$  where  $C$  is typically 4 or 8:

$$\mathcal{D}_\Delta = \left\{ g_j[n_1 - q_1 2^j \Delta^{-1}] g_j[n_2 - q_2 2^j \Delta^{-1}] \exp(-i 2^{-j} \eta(n_1 \cos \theta + n_2 \sin \theta)) \right\}_{q_1, q_2, j, \theta},$$

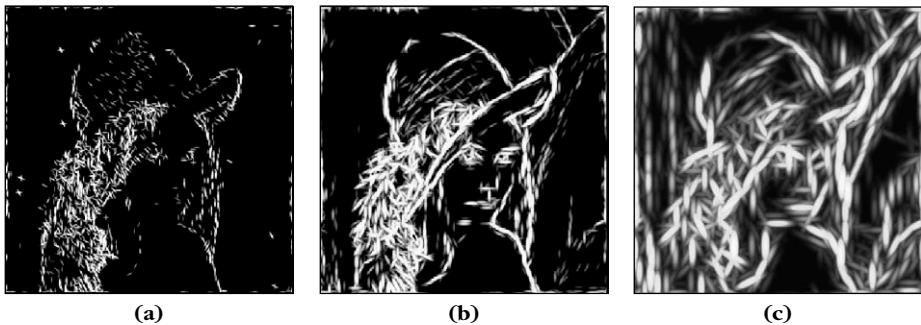
with  $\Delta \geq 2$  and  $\eta < 2\pi$ . This dictionary is a redundant directional wavelet frame. As opposed to the frame decompositions in Section 5.5.1, a matching pursuit yields a sparse image representation by selecting a few Gabor atoms best adapted to the image.

Figure 12.16 shows the atoms selected by a matching pursuit on the Lena image. Each selected atom is displayed as an ellipse at a position  $2^j \Delta^{-1}(q_1, q_2)$ , of width proportional to  $2^j$  and oriented in direction  $\theta$ , with a grayscale amplitude proportional to the matching pursuit coefficient.

To better capture the anisotropic regularity of edges, more Gabor atoms are incorporated in the dictionary, with an anisotropic stretching of their support. This redundant dictionary, which includes directional wavelets and curvelets, can be applied to low-bit rate image compression [257].

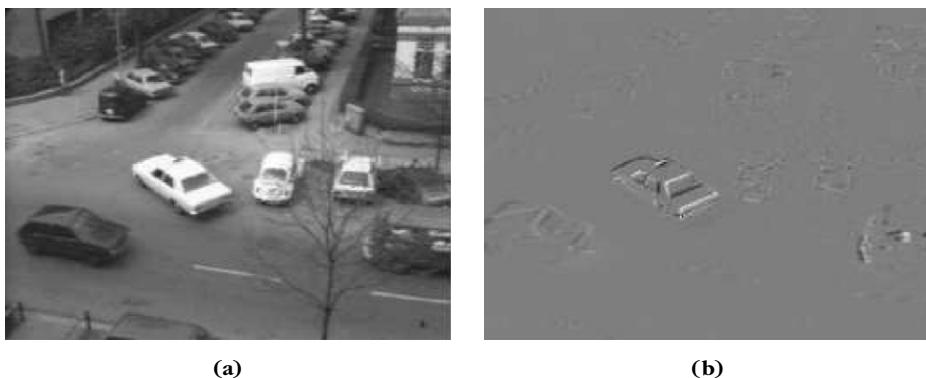
### Video Compression

In MPEG-1, -2, -4 video compression standards, motion vectors are coded to predict an image from a previous one, with a motion compensation [156, 437]. Figure 12.17(b) shows a prediction error image. It is the difference between the image in Figure 12.17(a) and a prediction obtained by moving pixel blocks of a previous image by using computed motion vectors. When the motion is not accurate, it yields errors along edges and sharp structures. These errors typically define oriented oscillatory structures. In MPEG compression standards, prediction error images are



**FIGURE 12.16**

Directional Gabor wavelets selected by a matching pursuit at several scales  $2^j$ : (a)  $2^1$ , (b)  $2^2$ , and (c)  $2^3$ . Each ellipse gives the direction, scale, and position of a selected Gabor atom.

**FIGURE 12.17**

(a) Image of video sequences with three cars moving on a street. (b) Motion compensation error.

compressed with the discrete cosine transform (DCT) introduced in Section 8.3.3. The most recent MPEG-4 H.264 standard adapts the size and shape of the DCT blocks to optimize the distortion rate.

Neff and Zakhor [386] introduced a video matching pursuit compression in two-dimensional Gabor dictionaries that efficiently compresses prediction error images. Section 12.1.2 explains that an orthogonalization reduces the quantization error, but the computational complexity of the orthogonalization is too important for real-time video calculations. Compression is thus implemented with a nonorthogonal matching pursuit iteration (12.52), modified to quantize the selected inner product with  $Q(x)$ :

$$R^{m+1}f = R^m f - Q(\langle R^m f, \phi_{p_m} \rangle) \phi_{p_m}.$$

Initially implemented in a separable Gabor dictionary [386], this procedure is refined in hierarchical dictionaries providing fast algorithms for larger directional Gabor dictionaries, which improves the compression efficiency [387]. Other dictionaries reducing computations have been proposed [80, 191, 318, 351, 426], with distortion rate models to adjust the quantizer to the required bit budget [388]. This lead to a video coder, recognized in 2002 by the MPEG-4 standard expert group as having the best distortion rate with a realistic implementation among all existing solutions. However, industrial priorities have maintained a DCT solution for the new MPEG-4 standard.

### 12.3.4 Coherent Matching Pursuit Denoising

If we cannot interpret the information carried by a signal component, it is often misconstrued as noise. In a crowd speaking a foreign language, we perceive surrounding conversations as background noise. In contrast, our attention is easily attracted by

a remote conversation spoken in a known language. What is important here is not the information content but whether this information is in a coherent format with respect to our system of interpretation. The decomposition of a signal in a dictionary of vectors can similarly be considered as a signal interpretation. Noises are then defined as signal components that do not have a strong correlation with any vector of the dictionary. In the absence of any knowledge concerning the noise, Mallat and Zhang [366] introduced a coherent matching pursuit denoising that selects coherent structures having a high correlation with vectors in the dictionary. These coherent structures typically correspond to the approximation support of  $f$  that can be identified in  $\mathcal{D}$ .

### **Denoising by Thresholding**

Let  $X[n] = f[n] + W[n]$  be noisy measurements. The dictionary estimator of Theorem 12.3 in Section 12.1.3 projects  $X$  on a best set of dictionary vectors  $\Lambda \subset \Gamma$ , which minimizes the  $\mathbf{1}^0$  Lagrangian  $\|X - X_\Lambda\|^2 + T^2 |\Lambda|$ .

For an orthogonal matching pursuit approximation that selects one by one the vectors that are orthogonalized, this is equivalent to thresholding at  $T$  the resulting decomposition (12.72):

$$\tilde{F} = \sum_{m=0}^{N-1} \rho_T \left( \frac{\langle R^m X, \phi_{p_m} \rangle}{\|u_m\|^2} \right) u_m.$$

Since the amplitude of residual coefficients  $|\langle R^m X, \phi_{p_m} \rangle| / \|u_m\|^2$  almost systematically decreases as  $m$  increases, it is nearly equivalent to stop the matching pursuit decomposition at the first iteration  $M$  such that  $|\langle R^M X, \phi_{p_M} \rangle| / \|u_M\|^2 < T$ . Thus, the threshold becomes a stopping criteria.

For a nonorthogonal matching pursuit, despite the nonorthogonality of selected coefficients, a denoising algorithm can also be implemented by stopping the decomposition (12.58) as soon as  $|\langle R^m, \phi_{p_m} \rangle| < T$ . The resulting estimator is

$$\tilde{F} = \sum_{m=0}^{M-1} \langle R^m X, \phi_{p_m} \rangle \phi_{p_m} \quad \text{with} \quad |\langle R^M, \phi_{p_M} \rangle| < T,$$

and can be optimized with a back projection computing the orthogonal projection of  $X$  in  $\{\phi_{p_m}\}_{0 \leq m < M}$ . Coherent denoising provides a different approach that does not rely on a particular noise model and does not set in advance the threshold  $T$ .

### **Coherent Denoising**

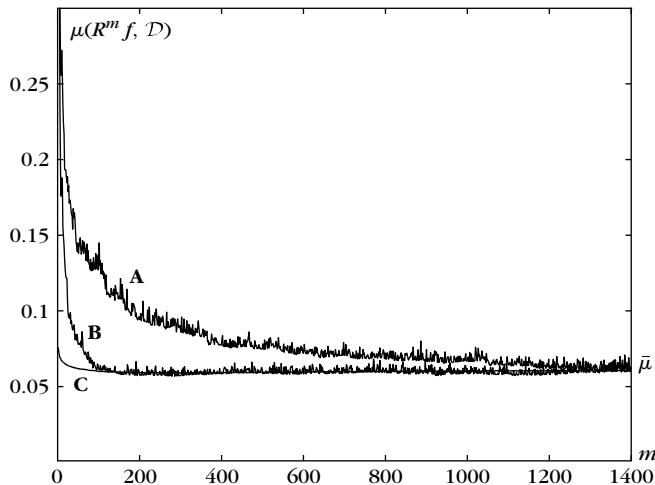
A coherent matching pursuit denoising selects signal structures having a correlation with dictionary vectors that is above an average defined over a matching pursuit attractor. A matching pursuit behaves like a nonlinear chaotic map, and it has been proved by Davis, Mallat, and Avellaneda [201] that for particular dictionaries, the normalized residues  $R^m f / \|R^m f\|$  converge to an attractor. This attractor

is a set of normalized signals  $h$  that do not correlate well with any  $\phi_p \in \mathcal{D}$  because all coherent structures of  $f$  producing maximum inner products with vectors in  $\mathcal{D}$  are progressively removed by the pursuit. Signals on the attractor do not correlate well with any dictionary vector and are thus considered as an incoherent noise with respect to  $\mathcal{D}$ . The coherence of  $f$  relative to  $\mathcal{D}$  is defined in (12.3.1) by  $\mu(f, \mathcal{D}) = \max_{p \in \Gamma} |\langle f / \|f\|, \phi_p \rangle|$ . For signals in the attractor, this coherence has a small amplitude, and we denote the average coherence of this attractor as  $\bar{\mu}$ , which depends on  $\mathcal{D}$  [201]. This average coherence is defined by

$$\bar{\mu} = \lim_{m \rightarrow +\infty} E\{\mu(R^m W', \mathcal{D})\},$$

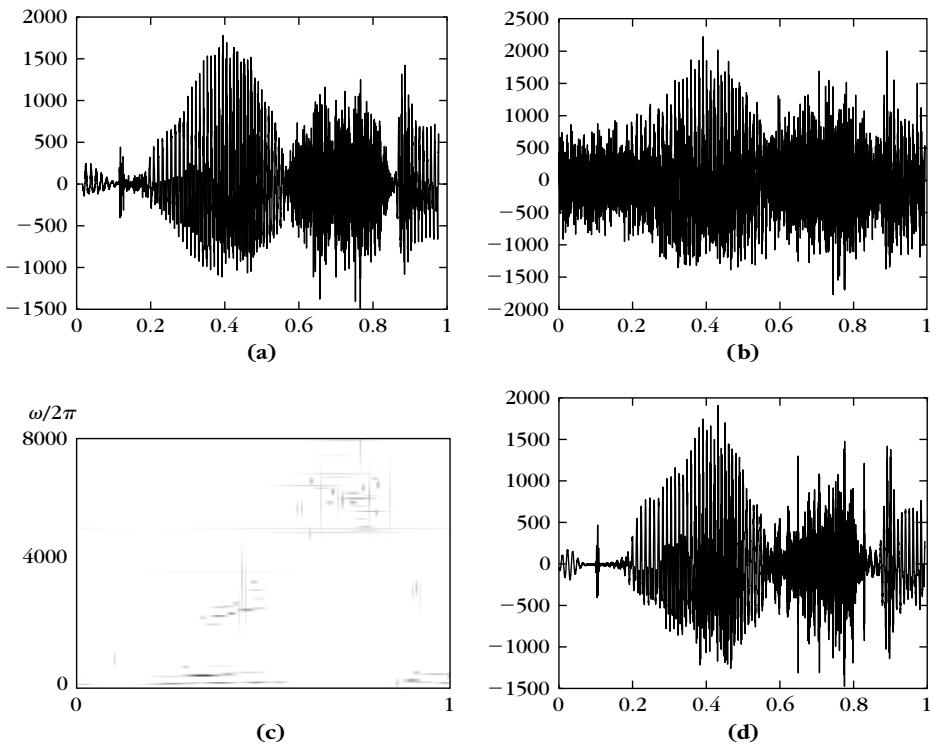
where  $W'$  is a Gaussian white noise of variance  $\sigma^2 = 1$ . The bottom regular curve C in Figure 12.18 gives the value of  $E\{\mu(R^m W', \mathcal{D})\}$  that is nearly equal to  $\bar{\mu} = 0.06$  for  $m \geq 40$  in a Gabor dictionary.

The convergence of the pursuit to the attractor implies that for  $m \geq M$  iterations, the residue  $R^m f$  has a normalized correlation  $\mu(R^m f, \mathcal{D})$  that is nearly equal to  $\bar{\mu}$ . Curve A in Figure 12.18 gives the decay of  $\mu(R^m f, \mathcal{D})$  as a function of  $m$  for the “greasy” signal  $f$  in Figure 12.19(a). After about  $M = 1400$  iterations, it reaches the average coherence level of the attractor. The corresponding 1400 time-frequency atoms are shown in Figure 12.15. Curve B in Figure 12.18 shows the decay of  $\mu(R^m X, \mathcal{D})$  for the noisy signal  $X = f + W$  in Figure 12.19(b), which has an SNR of



**FIGURE 12.18**

Decay of the correlation  $\mu(R^m f, \mathcal{D})$  as a function of the number of iterations  $m$  for two signals decomposed in a Gabor dictionary. **A:**  $f$  is the recording of “greasy” shown in Figure 12.19(a). **B:**  $f$  is the noisy “greasy” signal shown in Figure 12.19(b). **C:**  $E\{\mu(R^m W', \mathcal{D})\}$  that is for a normalized Gaussian white noise  $W'$ .

**FIGURE 12.19**

- (a) Speech recording of “greasy.” (b) Recording of “greasy” plus a Gaussian white noise (SNR = 1.5 dB). (c) Time-frequency distribution of the  $M = 76$  coherent Gabor structures. (d) Estimation  $\tilde{F}$  reconstructed from the 76 coherent structures (SNR = 6.8 dB).

1.5 db. The high-amplitude noise destroys most coherent structures and the attractor is reached after  $M = 76$  iterations.

A coherent matching pursuit denoising with a relaxation parameter  $\alpha = 1$  decomposes a signal as long as the coherence of the residue is above  $\bar{\mu}$  and stops after:

$$\tilde{F} = \sum_{m=0}^{M-1} \langle R^m X, \phi_{p_m} \rangle \phi_{p_m} \quad \text{with} \quad \mu(R^M X, \mathcal{D}) = \frac{|\langle R^M X, \phi_{p_M} \rangle|}{\|R^M X\|} < \bar{\mu}.$$

This estimator can thus also be interpreted as a thresholding of the matching pursuit of  $X$  with a threshold that is adaptively adjusted to

$$T = \bar{\mu} \|R^M X\|.$$

The time-frequency energy distribution of the  $M = 76$  coherent Gabor atoms of the noisy signal is shown in Figure 12.19(c). The estimation  $\tilde{F}$  calculated from the

76 coherent structures is shown in Figure 12.19(d). The SNR of this estimation is 6.8 db. The white noise has been removed with no estimation of the variance, and the restored speech signal has a good intelligibility because its main time-frequency components are retained.

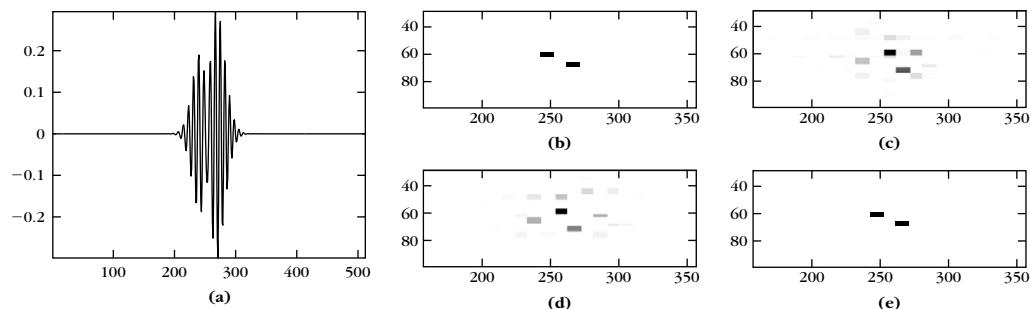
## 12.4 $\mathbf{l}^1$ PURSUITS

To reduce inefficiencies produced by the greediness of matching pursuits,  $\mathbf{l}^1$  pursuits perform a more global optimization, which replaces the  $\mathbf{l}^0$  norm minimization of a best  $M$ -term approximation by an  $\mathbf{l}^1$  norm. The  $\mathbf{l}^0$  Lagrangian studied from Section 12.1.1 is thus replaced by the  $\mathbf{l}^1$  Lagrangian from Section 12.4.2. Although they are not optimal in general, Section 12.5 proves that matching pursuits and basis pursuits can compute nearly optimal  $M$ -term approximations, depending on the signal approximation support and the dictionary.

### 12.4.1 Basis Pursuit

Each step of a matching pursuit performs a local optimization that can be fooled. A basis pursuit minimizes a global criterion that avoids some mistakes made by greedy pursuits. A simple but typical example of failure happens when a linear combination of two vectors  $f = \phi_m + \phi_q$  happens to be highly correlated with a third vector  $\phi_r \in \mathcal{D}$ . A matching pursuit may choose  $\phi_r$  instead of  $\phi_m$  or  $\phi_q$ , and many other vectors are then needed to correct this wrong initial choice, which produces a nonoptimal representation.

Figure 12.20 illustrates this phenomenon with a dictionary  $\mathcal{D}_{j,\Delta} = \{\phi_p\}_{p \in \Gamma}$  of one-dimensional Gabor atoms specified in (12.77). Each Gabor function is a Gaussian translated in time and frequency with an oversampled time-frequency grid



**FIGURE 12.20**

(a) Signal  $f = \phi_m + \phi_q$ . (b) Reduced Heisenberg boxes of the two Gabor atoms  $\phi_m$  and  $\phi_q$ , shown in the time-frequency plane. (c) Atoms selected by a matching pursuit. The darkness of each box is proportional to selected coefficients' amplitude. (d) Atoms selected by an orthogonal matching pursuit. (e) A basis pursuit recovers the two original atoms.

calculated with  $\Delta = 1/4$ . It has  $P = 16N$  vectors of size  $N$ . Figure 12.20 shows a signal  $f = \phi_m + \phi_q$  where  $\phi_m$  and  $\phi_q$  are two Gabor functions having nearly the same position and frequency. Let  $\sigma_t$  and  $\sigma_\omega$  be the time and frequency variance of these Gabor atoms. Figure 12.20(b) represents these atoms in the time-frequency plane, with two reduced Heisenberg rectangles, of time width  $\sigma_t/\Delta$  and frequency width  $\sigma_\omega/\Delta$ , so that all dictionary coefficients can be visualized. The full-size Heisenberg boxes  $\sigma_t \times \sigma_\omega$  of  $\phi_m$  and  $\phi_q$  overlap widely, which makes it difficult to discriminate them in  $f$ . Figures 12.20(c, d) show that a matching pursuit and an orthogonal matching pursuit select a first time-frequency atom with a time and frequency location intermediate between these two atoms, and then other subsequent vectors to compensate for this initial mistake. Such nonoptimal greedy choices are observed on real signals decomposed in redundant dictionaries. High-resolution greedy pursuits can reduce the loss of resolution in time with nonlinear correlation measures [279, 314], but the greediness can still have adverse effects.

### $\mathbf{I}^1$ Minimization

Avoiding this greediness suboptimality requires using a global criterion that enforces the sparsity of the decomposition coefficients of  $f$  in  $\mathcal{D} = \{\phi_p\}_{p \in \Gamma}$ . Let  $\Phi f[p] = \langle f, \phi_p \rangle$  be the decomposition operator in  $\mathcal{D}$ . The reconstruction from dictionary vectors is implemented by the adjoint (5.3)

$$f[n] = \Phi^* a[n] = \sum_{p=0}^{P-1} a[p] \phi_p[n]. \quad (12.82)$$

Since the dictionary is redundant, there are many possible reconstructions. The basis pursuit introduced by Chen and Donoho [159] finds the vector  $\tilde{a}$  of coefficients having a minimum  $\mathbf{I}^1$  norm

$$\tilde{a} = \underset{a \in \mathbb{R}^P}{\operatorname{argmin}} \|a\|_1 \quad \text{subject to} \quad \Phi^* a = f. \quad (12.83)$$

This is a convex minimization that can be written as a linear programming and is thus calculated with efficient algorithms, although computationally more intensive than a matching pursuit. If the solution of (12.83) is not unique, then any valid solution may be used.

The signal in Figure 12.20 can be written exactly as a sum of two dictionary vectors and the basis pursuit minimization recovers this best representation by selecting the appropriate dictionary vectors, as opposed to matching pursuit algorithms. Minimizing the  $\mathbf{I}^1$  norm of the decomposition coefficients  $a[p]$  avoids cancellation effects when selecting an inappropriate vector in the representation, which is then canceled by other redundant dictionary vectors. Indeed, these cancellations increase the  $\mathbf{I}^1$  norm of the resulting coefficients. As a result, the global optimization of a basis pursuit can provide a more accurate representation of sparse signals than matching pursuits for highly correlated and redundant dictionaries. This is further studied in Section 12.5.

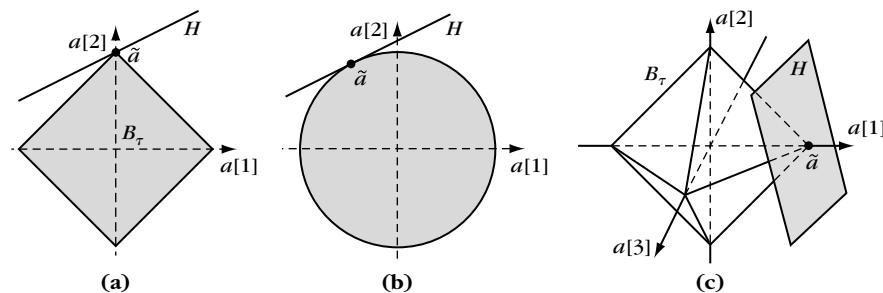


FIGURE 12.21

(a, b) Comparison of the minimum  $\mathbf{l}^1$  and  $\mathbf{l}^2$  solutions of  $\Phi^* \mathbf{a} = f$  for  $P = 2$ . (c) Geometry of  $\mathbf{l}^1$  minimization for  $P = 3$ . Solution  $\tilde{\mathbf{a}}$  typically has fewer nonzero coefficients for  $\mathbf{l}^1$  than for  $\mathbf{l}^2$ .

### Sparsity

A geometric interpretation of basis pursuit also explains why it can recover sparse solutions. For a dictionary of size  $P$  the decomposition coefficients  $a[p]$  define a vector in  $\mathbf{a} \in \mathbb{R}^P$ . Let  $H$  be the affine subspace of  $\mathbb{R}^P$  of coordinate vectors that recover  $f \in \mathbb{R}^N$ ,

$$H = \{\mathbf{a} \in \mathbb{R}^P : \Phi^* \mathbf{a} = f\} = \mathbf{a}_0 + \mathbf{Null}(\Phi^*) \subset \mathbb{R}^P, \quad \text{where } \Phi^* \mathbf{a}_0 = f. \quad (12.84)$$

The dimension of  $H$  is  $P - N$ . A basis pursuit (12.83) finds in  $H$  an element  $\tilde{\mathbf{a}}$  of minimum  $\mathbf{l}^1$  norm. It can be found by inflating the  $\mathbf{l}^1$  ball

$$B_\tau = \{\mathbf{a} \in \mathbb{R}^P : \|\mathbf{a}\|_1 \leq \tau\} \subset \mathbb{R}^P, \quad (12.85)$$

by increasing  $\tau$  until it intersects  $H$ . This geometric configuration is depicted for  $P = 2$  and  $P = 3$  in Figure 12.21.

The  $\mathbf{l}^1$  ball remains closer to the coordinate axes of  $\mathbb{R}^P$  than the  $\mathbf{l}^2$  ball. When the dimension  $P$  increases, the volume of the  $\mathbf{l}^1$  ball becomes much smaller than the volume of the  $\mathbf{l}^2$  ball. Thus, the optimal solution  $\tilde{\mathbf{a}}$  is likely to have more zeros or coefficients close to zero when it is computed by minimizing an  $\mathbf{l}^1$  norm rather than an  $\mathbf{l}^2$  norm. This is illustrated by Figure 12.21.

### Basis Pursuit and Best-Basis Selection

Theorem 12.7 proves that a basis pursuit selects vectors that are independent, unless it is a degenerated case where the solution is not unique, which happens rarely. We denote by  $\tilde{\Lambda} = \{p : \tilde{\mathbf{a}}[p] \neq 0\}$  the support of  $\tilde{\mathbf{a}}$ .

**Theorem 12.7.** A basis pursuit (12.83) admits a solution  $\tilde{\mathbf{a}}$  with support  $\tilde{\Lambda}$  that corresponds to a family  $\{\phi_p\}_{p \in \tilde{\Lambda}}$  of linearly independent dictionary vectors.

**Proof.** If  $\{\phi_p\}_{p \in \tilde{\Lambda}}$  is linearly dependent, then there exists  $h \in \mathbf{Null}(\Phi^*)$  with  $h \neq 0$  and  $h[p] = 0$  for  $p \notin \tilde{\Lambda}$ . For  $\lambda$  small enough such that  $\text{sign}(\tilde{\mathbf{a}} + \lambda h) = \text{sign}(\tilde{\mathbf{a}})$ , the mapping  $\lambda \mapsto \|\tilde{\mathbf{a}} + \lambda h\|_1$

is locally affine until at least one of the components of  $\tilde{a} + \lambda h$  vanishes. Since  $\|\tilde{a}\|_1$  is minimum,  $\|\tilde{a} + \lambda h\|_1$  is constant for  $\lambda$  small enough, and thus  $\|\tilde{a} + \lambda h\|_1 = \|\tilde{a}\|_1$  for all such  $\lambda$ . The minimization of  $\|\alpha\|_1$  with  $\Phi^* \alpha = f$  is therefore nonunique.

Furthermore, for a critical value of  $\lambda$ , one of the components of  $\tilde{a} + \lambda h$  vanishes. The support of  $\tilde{a} + \lambda h$  is strictly included in  $\Lambda$  and  $\|\tilde{a} + \lambda h\|_1$  is minimum. Setting  $\tilde{a}_1 = \tilde{a} + \lambda h$  and iterating this argument shows that there exists a solution supported inside  $\tilde{\Lambda}$  that indexes vectors that are linearly independent. ■

Signals of size  $N$  can rarely be written exactly as a sum of less than  $N$  dictionary vectors, and the  $N$  independent vectors selected by a basis pursuit thus define a basis of  $\mathbb{C}^N$ . A basis pursuit can therefore be interpreted as a *best-basis* algorithm. Among all possible bases of  $\mathcal{D}$ , it selects a basis  $\mathcal{B} = \{\phi_{p_m}\}_{0 \leq m < N}$ , which yields decomposition coefficients  $\{a[p_m]\}_{0 \leq m < N}$  of minimum  $\ell^1$  norm. Unlike the best-basis selection algorithm in Section 12.2.2, it does not restrict the search to orthonormal bases, which provides much more flexibility.

Signal denoising or compression applications can be implemented by thresholding or quantizing the decomposition coefficients of a basis pursuit. However, there is no control on the stability of the selected basis  $\mathcal{B}$ . The potential instabilities of the basis do not provide a good control on the resulting error, but results are typically slightly better than with a matching pursuit.

### **Linear Programming for the Resolution of Basis Pursuit**

The basis pursuit minimization of (12.83) is a convex optimization problem that can be reformulated as a linear programming problem. A standard-form linear programming problem [28] is a constrained optimization over positive vectors  $d[p]$  of size  $L$ . Let  $b[n]$  be a vector of size  $N < L$ ,  $c[p]$  a nonzero vector of size  $L$ , and  $A[n, p]$  an  $L \times N$  matrix. A linear programming problem finds  $d[p] \in \mathbb{R}^L$  such that  $d[p] \geq 0$ , which is the solution of the minimization problem

$$\tilde{d} = \operatorname{argmin}_{d \in (\mathbb{R}^+)^L} \sum_{p=0}^{L-1} d[p] c[p] \quad \text{subject to} \quad Ad = b. \quad (12.86)$$

The basis pursuit optimization

$$\tilde{a} = \operatorname{argmin}_{a \in \mathbb{R}^P} \|a\|_1 \quad \text{subject to} \quad \Phi^* a = f \quad (12.87)$$

is recast as linear programming by introducing slack variables  $u[p] \geq 0$  and  $v[p] \geq 0$  such that  $a = u - v$ . One can then define

$$A = (\Phi^*, -\Phi^*) \in \mathbb{R}^{N \times 2P} \quad c = 1, \quad d = (u, v) \in \mathbb{R}^{2P}, \quad \text{and} \quad b = f.$$

Since

$$\|a\|_1 = \sum_{p=0}^{L-1} d[p] c[p] \quad \text{and} \quad Ad = \Phi^* u - \Phi^* v = f,$$

this shows that (12.87) is written as a linear programming problem (12.86) of size  $L = 2P$ . The matrix  $A$  of size  $N \times L$  has rank  $N$  because the dictionary  $\mathcal{D}$  includes  $N$  linearly independent vectors.

The collection of feasible points  $\{d : Ad = b, d \geq 0\}$  is a convex polyhedron in  $\mathbb{R}^L$ . Theorem 12.7 proves there exists a solution of the linear programming problem with at most  $N$  nonzero coefficients. The linear cost (12.86) can thus be minimized over the subpolyhedron of vectors having  $N$  nonzero coefficients. These  $N$  nonzero coefficients correspond to  $N$  column vectors  $\mathcal{B} = \{\phi_{p_m}\}_{0 \leq m < N}$  that form a basis.

One can also prove [28] that if the cost is not minimum at a given vertex, then there exists an adjacent vertex with a cost that is smaller. The simplex algorithm takes advantage of this property by jumping from one vertex to an adjacent vertex while reducing the cost (12.86). Going to an adjacent vertex means that one of the zero coefficients of  $d[p]$  becomes nonzero while one nonzero coefficient is set to zero. This is equivalent to modifying the basis  $\mathcal{B}$  by replacing one vector by another vector of  $\mathcal{D}$ . The simplex algorithm thus progressively improves the basis by appropriate modifications of its vectors, one at a time. In the worst case, all vertices of the polyhedron will be visited before finding the solution, but the average case is much more favorable.

Since the 1980s, more effective interior point procedures have been developed. Karmarkar's interior point algorithm [325] begins in the middle of the polyhedron and converges by iterative steps toward the vertex solution, while remaining inside the convex polyhedron. For finite precision calculations, when the algorithm has converged close enough to a vertex, it jumps directly to the corresponding vertex, which is guaranteed to be the solution. The middle of the polyhedron corresponds to a decomposition of  $f$  over all vectors of  $\mathcal{D}$ , typically with  $P > N$  nonzero coefficients. When moving toward a vertex some coefficients progressively decrease while others increase to improve the cost (12.86). If only  $N$  decomposition coefficients are significant, jumping to the vertex is equivalent to setting all other coefficients to zero. Each step requires computing the solution of a linear system. If  $A$  is an  $N \times L$  matrix, then Karmarkar's algorithm terminates with  $O(L^{3.5})$  operations. Mathematical work on interior point methods has led to a large variety of approaches that are summarized in [355].

Besides linear programming, let us also mention that simple iterative algorithms can also be implemented to compute the basis pursuit solution [184].

### ***Application to Wavelet Packet and Local Cosine Dictionaries***

Dictionaries of wavelet packets and local cosines have  $P = N \log_2 N$  time-frequency atoms. A straightforward implementation of interior point algorithms thus requires  $O(N^{3.5} \log_2^3 N)$  operations. By using the fast wavelet packet and local cosine transforms together with heuristic computational rules, the number of operations is considerably reduced [158]. The algorithm still remains computationally intense.

Figure 12.14 is an example of a synthetic signal with localized time-frequency structures, decomposed on a wavelet packet dictionary. The flexibility of a basis pursuit decomposition in Figure 12.14(e) gives a much more sparse representation

than a best orthogonal wavelet packet basis in Figure 12.14(c). In this case, the resulting representation is very similar to the matching pursuit decompositions in Figures 12.14(d, f). Figure 12.20 shows that a basis pursuit can improve matching pursuit decompositions when the signal includes close time-frequency structures that are not distinguished by a matching pursuit that selects an intermediate time-frequency atom [158].

### 12.4.2 $\ell^1$ Lagrangian Pursuit

Compression and denoising applications of basis pursuit decompositions create sparse representations by quantizing or thresholding the resulting coefficients. This is not optimal because the selected basis  $\mathcal{B} \subset \mathcal{D}$  is not orthogonal and may even be unstable. It is then more efficient to directly solve the sparse denoising or approximation problem with a Lagrangian approach.

#### *Basis Pursuit Approximation and Denoising*

To suppress an additive noise or to approximate a signal with an error  $\varepsilon$ , instead of thresholding the coefficients obtained with a basis pursuit, Chen, Donoho, and Saunders [159] compute the solution

$$\tilde{f} = \sum_{p=0}^{P-1} \tilde{a}[p] \phi_p = \Phi^* \tilde{a}$$

with decomposition coefficients  $\tilde{a}$  that are a solution of a minimization problem that incorporates the precision parameter  $\varepsilon$ :

$$\tilde{a} = \underset{\alpha \in \mathbb{R}^P}{\operatorname{argmin}} \|\alpha\|_1 \quad \text{subject to} \quad \|\Phi^* \alpha - f\|^2 \leq \varepsilon. \quad (12.88)$$

In a denoising problem,  $f$  is replaced by the noisy signal  $X = f + W$  where  $W$  is the additive noise. It is then called a *basis pursuit denoising* algorithm.

Computing the solution of this quadratically constrained problem is more complicated than the linear programming problem corresponding to a basis pursuit. However, it can be recast as a second-order cone program, which is solved using interior point methods and, in particular, log-barrier methods [10] that extend the interior point algorithms for linear programming problems. These general-purpose algorithms can also be optimized to take into account the separability of the  $\ell^1$  norm. Iterative algorithms converging to the solution have also been developed [183].

The minimization problem (12.88) is convex and thus can also be solved through its Lagrangian formulation:

$$\tilde{a} = \underset{\alpha \in \mathbb{C}^P}{\operatorname{argmin}} \mathcal{L}_1(T, f, \alpha) = \underset{\alpha \in \mathbb{C}^P}{\operatorname{argmin}} \frac{1}{2} \|f - \Phi^* \alpha\|^2 + T \|\alpha\|_1. \quad (12.89)$$

In the following, this Lagrangian minimization will be called a *Lagrangian basis pursuit* or  $\ell^1$  *pursuit*. For each  $\varepsilon > 0$ , there exists a Lagrangian multiplier  $T$  such that convex minimization (12.88) and the Lagrangian minimization (12.89) have

a common solution [266, 463]. In a denoising problem, where  $f$  is replaced by  $X = f + W$ , it is easier to adjust  $T$  than  $\varepsilon$  to the noise level. Indeed, we shall see that a Lagrangian pursuit performs a generalized soft thresholding by  $T$ . For a Gaussian white noise  $\sigma$ , one can typically set  $T = \lambda \sigma$  with  $\lambda \leq \sqrt{2 \log_e P}$ , where  $P$  is the dictionary size. Section 12.4.3 describes algorithms for computing a solution of the Lagrangian minimization (12.89).

Figure 12.22 shows an example of basis pursuit denoising of an image contaminated by a Gaussian white noise. The dictionary includes a translation-invariant dyadic wavelet frame and a tight frame of local cosine vectors with a redundancy factor of 16. The resulting estimation is better than in a translation-invariant wavelet dictionary. Indeed, local cosine vectors provide more sparse representations of the image oscillatory textures.

### **Convexification of $\mathbf{l}^0$ with $\mathbf{l}^1$**

Theorem 12.1 proves that a best  $M$ -term approximation  $f_{\Lambda_T}$  in a dictionary  $\mathcal{D}$  is obtained by minimizing the  $\mathbf{l}^0$  Lagrangian (12.30)

$$\mathcal{L}_0(T, f, \Lambda) = \|f - f_\Lambda\|^2 + T^2 |\Lambda| = \|f - \sum_{p \in \Lambda} a[p] \phi_p\|^2 + T^2 \|a\|_0. \quad (12.90)$$

Since the  $\mathbf{l}^0$  pseudo norm is not convex, Section 12.1.1 explains that the minimization of  $\mathcal{L}_0$  is intractable. An  $\mathbf{l}^0$  norm can be approximated by an  $\mathbf{l}^q$  pseudo-norm

$$\|a\|_q = \left( \sum_{p \in \Lambda} |a[p]|^q \right)^{1/q},$$

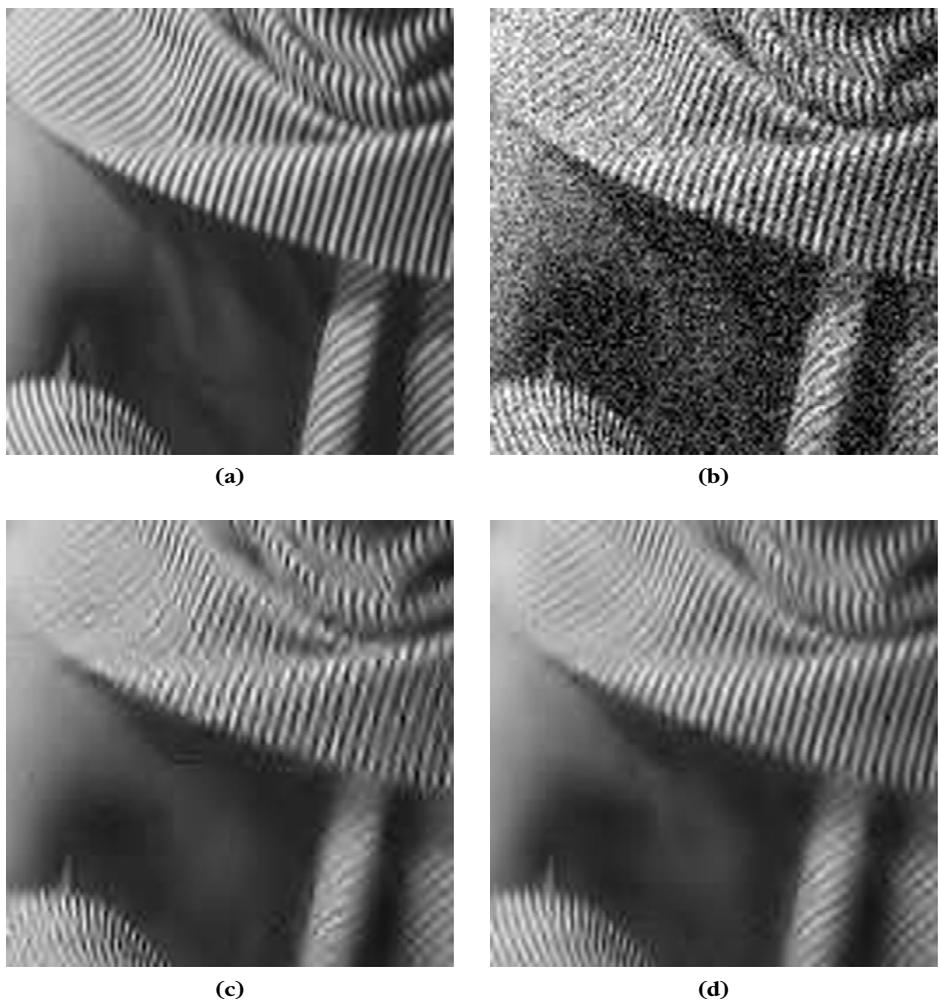
which is nonconvex for  $0 < q < 1$ , and convex and thus a norm for  $q \geq 1$ . As  $q$  decreases, Figure 12.23 shows in  $P=2$  dimensions that the unit ball of vectors  $\|a\|_q \leq 1$  approaches the  $\mathbf{l}^0$  unit “ball,” which corresponds to the two axes. The  $\mathbf{l}^1$  Lagrangian minimization (12.89) can thus be interpreted as a convexification of the  $\mathbf{l}^0$  Lagrangian minimization.

Let  $\tilde{\Lambda} = \{p \in \Gamma : \tilde{a}[p] \neq 0\}$  be the support of the  $\mathbf{l}^1$  Lagrangian pursuit solution  $\tilde{a}$ . For  $|\tilde{\Lambda}| = M$ , the support  $\tilde{\Lambda}$  is typically not equal to the best  $M$ -term approximation support  $\Lambda_T$  obtained by minimizing the  $\mathbf{l}^0$  Lagrangian and  $\|f - \tilde{f}\| \geq \|f - f_{\Lambda_T}\|$ . However, Section 12.5.3 proves that  $\tilde{\Lambda}$  may include the main approximation vectors of  $\Lambda_T$  and  $\|f - \tilde{f}\|$  can be of the same order as  $\|f - f_{\Lambda_T}\|$  if  $\Lambda_T$  satisfies an exact recovery condition that is specified.

### **Generalized Soft Thresholding**

A Lagrangian pursuit computes a sparse approximation  $\tilde{f} = \Phi^* \tilde{a}$  of  $f$  by minimizing

$$\tilde{a} = \underset{a \in \mathbb{C}^P}{\operatorname{argmin}} \mathcal{L}_1(T, f, a) \quad \text{where} \quad \mathcal{L}_1(T, f, a) = \frac{1}{2} \|f - \Phi^* a\|^2 + T \|a\|_1. \quad (12.91)$$

**FIGURE 12.22**

**(a)** Original image  $f$ . **(b)** Noisy image  $X = f + W$  (SNR = 12.5 dB). **(c)** Translation-invariant wavelet denoising, (SNR = 18.6 dB). **(d)** Basis pursuit denoising in a dictionary that is a union of a translation-invariant wavelet frame and a frame of redundant local cosine vectors (SNR = 19.8 dB).

Increasing  $T$  often reduces the support  $\tilde{\Lambda}$  of  $\tilde{a}$  but increases the approximation error  $\|f - \Phi^*\tilde{a}\|$ . The restriction of the dictionary operator and its adjoint to  $\tilde{\Lambda}$  is written as

$$\Phi_{\tilde{\Lambda}} = \{ \langle f, \phi_p \rangle \}_{p \in \tilde{\Lambda}} \quad \text{and} \quad \Phi_{\tilde{\Lambda}}^* a = \sum_{p \in \tilde{\Lambda}} a[p] \phi_p.$$

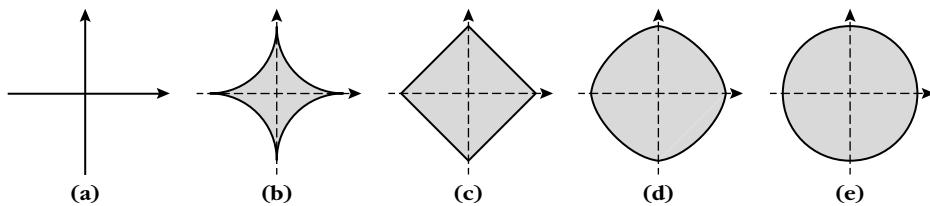


FIGURE 12.23

Unit balls of  $\mathbf{I}^q$  functionals: (a)  $q = 0$ , (b)  $q = 0.5$ , (c)  $q = 1$ , (d)  $q = 1.5$ , and (e)  $q = 2$ .

Theorem 12.8, proved by Fuchs [266], gives necessary and sufficient conditions on  $\tilde{\alpha}$  and its support  $\tilde{\Lambda}$  to be a minimizer of  $\mathcal{L}_1(T, f, \alpha)$ .

**Theorem 12.8: Fuchs.** A vector  $\tilde{\alpha}$  is a minimum of  $\mathcal{L}_1(T, f, \alpha)$  if and only if there exists  $h \in \mathbb{R}^P$  such that

$$\Phi(\Phi^* \tilde{\alpha} - f) + Th = 0 \quad \text{where} \quad \begin{cases} h[p] = \text{sign}(\tilde{\alpha}[p]) & \text{if } \tilde{\alpha}[p] \neq 0 \\ |h[p]| \leq 1 & \text{if } \tilde{\alpha}[p] = 0. \end{cases} \quad (12.92)$$

There exists a solution  $\tilde{\alpha}$  with support  $\tilde{\Lambda}$  that corresponds to a family  $\{\phi_p\}_{p \in \tilde{\Lambda}}$  of linearly independent dictionary vectors. The restriction  $\tilde{\alpha}_{\tilde{\Lambda}}$  over its support satisfies

$$\tilde{\alpha}_{\tilde{\Lambda}} + T(\Phi_{\tilde{\Lambda}} \Phi_{\tilde{\Lambda}}^*)^{-1} \text{sign}(\tilde{\alpha}_{\tilde{\Lambda}}) = \Phi_{\tilde{\Lambda}}^* f, \quad (12.93)$$

where  $\Phi_{\tilde{\Lambda}}^* f = (\Phi_{\tilde{\Lambda}} \Phi_{\tilde{\Lambda}}^*)^{-1} \Phi_{\tilde{\Lambda}}$  is the pseudo inverse of  $\Phi_{\tilde{\Lambda}}^*$ .

**Proof.** If  $\tilde{\alpha}[p] \neq 0$ , then  $\mathcal{L}_1(T, f, \alpha)$  is differentiable along the coordinate  $\alpha[p]$  at the point  $\tilde{\alpha}[p]$ . Setting this derivative to 0 shows that  $\tilde{\alpha}$  is minimum if and only if

$$\langle \phi_p, \Phi^* \tilde{\alpha} - f \rangle + T \text{sign}(\tilde{\alpha}[p]) = 0, \quad \text{so} \quad h[p] = \text{sign}(\tilde{\alpha}[p]). \quad (12.94)$$

If  $\tilde{\alpha}[p] = 0$ , let us consider the vector  $\alpha[q] = \tilde{\alpha}[q] + \tau \delta[q - p]$  for  $\tau \in \mathbb{R}$ . The corresponding Lagrangian value is

$$\mathcal{L}_1(T, f, \alpha) = \mathcal{L}_1(T, f, \tilde{\alpha}) + \tau \langle \phi_p, \Phi^* \tilde{\alpha} - f \rangle + \frac{\tau^2}{2} + T|\tau| \geq \mathcal{L}_1(\tilde{\alpha}).$$

Since  $\tilde{\alpha}$  is a minimizer

$$\forall \tau, \quad \tau \langle \phi_p, \Phi^* \tilde{\alpha} - f \rangle + T|\tau| + \frac{\tau^2}{2} \geq 0.$$

By separately considering the cases  $\tau > 0$  and  $\tau < 0$ , we verify that when  $\tau$  goes to zero it implies that

$$|\langle \phi_p, \Phi^* \tilde{\alpha} - f \rangle| \leq T. \quad (12.95)$$

Conditions (12.94) and (12.95) are equivalent to (12.92).

Conversely, if  $h$  satisfies (12.92), then for any  $\alpha$  we verify that

$$|\alpha[p]| \geq |\tilde{\alpha}[p]| + h[p](\alpha[p] - \tilde{\alpha}[p]),$$

and thus

$$\|a\|_1 \geq \| \tilde{a} \|_1 + \sum_p h[p](a[p] - \tilde{a}[p]) = \| \tilde{a} \|_1 + \langle h, a - \tilde{a} \rangle.$$

Since  $\|\Phi^* \tilde{a} - f\|^2$  is differentiable and convex, this leads to

$$\begin{aligned} \mathcal{L}_1(T, f, a) &\geq \frac{1}{2} \|\Phi^* \tilde{a} - f\|^2 + \langle a - \tilde{a}, \Phi(\Phi^* \tilde{a} - f) \rangle + T (\| \tilde{a} \|_1 + \langle h, a - \tilde{a} \rangle) \\ &\geq \mathcal{L}_1(T, f, \tilde{a}) + \langle a - \tilde{a}, \Phi(\Phi^* \tilde{a} - f) + Th \rangle = \mathcal{L}_1(T, f, \tilde{a}) \end{aligned}$$

because of (12.92), and thus  $\tilde{a}$  minimizes  $\mathcal{L}_1(T, f, a)$ . The conditions (12.92) therefore are necessary and sufficient.

The proof of the existence of a solution  $\tilde{a}$  corresponding to linear independent vectors  $\{\phi_p\}_{p \in \Lambda(\tilde{a})}$  is identical to the proof of Theorem 12.7 in the basis pursuit context. Over the support  $\tilde{\Lambda}$ , the necessary and sufficient condition (12.92) can be rewritten as

$$\Phi_{\tilde{\Lambda}} (\Phi_{\tilde{\Lambda}}^* \tilde{a}_{\tilde{\Lambda}} - f) + T \text{sign}(\tilde{a}_{\tilde{\Lambda}}) = 0,$$

which implies (12.93). ■

Let  $f_{\tilde{\Lambda}} = \Phi_{\tilde{\Lambda}}^* a_{\tilde{\Lambda}}$  be the orthogonal projection of  $f$  over the space  $V_{\tilde{\Lambda}}$  generated by  $\{\phi_p\}_{\tilde{\Lambda}}$ . Its coefficients are  $a_{\tilde{\Lambda}} = \Phi_{\tilde{\Lambda}}^{*+} f_{\tilde{\Lambda}} = \Phi_{\tilde{\Lambda}}^{*+} f$ . Theorem 12.8 proves in (12.93) that

$$\tilde{a}_{\tilde{\Lambda}} = a_{\tilde{\Lambda}} - T(\Phi_{\tilde{\Lambda}} \Phi_{\tilde{\Lambda}}^*)^{-1} \text{sign}(\tilde{a}_{\tilde{\Lambda}}), \quad (12.96)$$

which shows that the  $\mathbf{l}^1$  minimization attenuates by an amount proportional to  $T$  the amplitude of the coefficients  $a_{\tilde{\Lambda}}$  of the orthogonal projection  $f_{\tilde{\Lambda}}$ . This can be interpreted as a generalized soft thresholding. If  $\mathcal{D}$  is an orthonormal basis, then  $\Phi \Phi^* = \text{Id}$  and  $\tilde{a}$  is a classical soft thresholding of the coefficient  $\Phi f$  of  $f$  in this orthonormal basis.

Since  $\tilde{f} = \Phi^* \tilde{a} \in V_{\tilde{\Lambda}}$ , we know that  $\|f - \tilde{f}\| \geq \|f - f_{\tilde{\Lambda}}\|$ . Once the optimal Lagrangian support  $\tilde{\Lambda}$  is recovered, similar to the matching pursuit backprojection, the sparse approximation  $\tilde{f} = \Phi_{\tilde{\Lambda}}^* \tilde{a}_{\tilde{\Lambda}}$  of  $f$  can be improved by computing the orthogonal projection  $f_{\tilde{\Lambda}}$  and its coefficients  $a_{\tilde{\Lambda}}$ .

### 12.4.3 Computations of $\mathbf{l}^1$ Minimizations

The relaxed formulation of an  $\mathbf{l}^1$  Lagrangian pursuit

$$\tilde{a} = \underset{a \in \mathbb{C}^P}{\operatorname{argmin}} \mathcal{L}_1(T, f, a) = \underset{a \in \mathbb{C}^P}{\operatorname{argmin}} \frac{1}{2} \|f - \Phi^* a\|^2 + T \|a\|_1 \quad (12.97)$$

cannot be formulated as a linear program, as opposed to the basis pursuit minimization (12.83). Several approaches have been developed to compute this minimization.

#### *Iterative Thresholding*

Several authors [100, 185, 196, 241, 255, 466] have proposed an iterative algorithm that solves (12.89) with a soft thresholding to decrease the  $\mathbf{l}^1$  norm of the

coefficients  $a$ , and a gradient descent to decrease the value of  $\|f - \Phi^*a\|$ . The coefficient vector  $a$  may be complex, and  $|a[p]|$  is then the complex modulus.

**1. Initialization.** Choose  $a_0$  (e.g., 0), set  $k = 0$ , and compute  $b = \Phi f$ .

**2. Gradient step.** Update

$$\bar{a}_k = a_k + \gamma (b - \Phi \Phi^* a_k), \quad (12.98)$$

where  $\gamma < 2 \|\Phi \Phi^*\|_S^{-1}$ .

**3. Soft thresholding.** Compute

$$a_{k+1}[p] = \rho_{\gamma T}(\bar{a}_k[p]) \quad \text{where} \quad \rho_{\gamma T}(x) = x \max\left(1 - \frac{\gamma T}{|x|}, 0\right). \quad (12.99)$$

**4. Stop.** If  $\|a_k - a_{k+1}\|$  is smaller than a fixed-tolerance criterion, stop the iterations, otherwise set  $k \leftarrow k + 1$  and go back to 2.

This algorithm includes the same gradient step as the Richardson iteration algorithm in Theorem 5.7, which inverts the symmetric operator  $L = \Phi \Phi^*$ . The convergence condition  $\gamma < 2 \|\Phi \Phi^*\|_S^{-1}$  is identical to the convergence condition (5.35) of the Richardson algorithm. Theorem 12.9 proves that the convergence is guaranteed for any  $a_0$ .

**Theorem 12.9.** The sequence  $a_k$  defined by

$$a_{k+1}[p] = \rho_{\gamma T}(a_k[p] + \gamma \Phi(f - \Phi^* a_k)[p]) \quad \text{with} \quad \gamma < 2 \|\Phi \Phi^*\|_S^{-1} \quad (12.100)$$

converges to a solution of (12.89) for any  $a_0 \in \mathbb{C}^P$ .

**Proof.** The following proof is due to Daubechies, Defries, and DeMol [196], showing that  $a_k$  converges to a minimum of the energy

$$\mathcal{L}_1(T, f, a) = \frac{1}{2} \|\Phi^* a - f\|^2 + T \|a\|_1 \quad (12.101)$$

for  $\gamma < \|\Phi \Phi^*\|_S^{-1}$ . A proof of convergence for  $\gamma < 2 \|\Phi \Phi^*\|_S^{-1}$  can be found in [185].

To simplify notations, the dependencies of the Lagrangian on  $T$  and  $f$  are dropped, and it is written  $\mathcal{L}_1(a)$ . Lemma 12.2 proves that  $a_{k+1}$  is the minimum of a modified surrogate Lagrangian  $\tilde{\mathcal{L}}_1(a, a_k)$  that approximates  $\mathcal{L}_1(a)$  and depends on the previous iterate.

**Lemma 12.2.** Let  $\xi$  be the operator that associates to a vector  $b$  the vector

$$\xi(b) = \{\rho_{\gamma T}(\bar{b}[p])\}_{p \in \Gamma} \quad \text{with} \quad \bar{b} = b + \gamma \Phi(f - \Phi^* b). \quad (12.102)$$

For any  $b \in \mathbb{C}^N$ ,  $\xi(b)$  is the unique minimizer of  $\tilde{\mathcal{L}}_1(a, b)$  over all  $a$ , where

$$\tilde{\mathcal{L}}_1(a, b) = \mathcal{L}_1(a) + \frac{1}{2\gamma} \|a - b\|^2 - \frac{1}{2} \|\Phi^* a - \Phi^* b\|^2 \quad \text{and} \quad \mathcal{L}_1(a) = \frac{1}{2} \|f - \Phi^* a\|^2 + T \|a\|_1.$$

Furthermore,

$$\forall h \in \mathbb{C}^N, \quad \tilde{\mathcal{L}}_1(\xi(b) + h, b) \geq \tilde{\mathcal{L}}_1(\xi(b), b) + \frac{1}{2\gamma} \|h\|^2. \quad (12.103)$$

**Proof.** The modified Lagrangian  $\tilde{\mathcal{L}}_1(a, b)$  is expanded as

$$\begin{aligned} \tilde{\mathcal{L}}_1(a, b) &= \frac{1}{2\gamma} \|a\|^2 - \frac{1}{\gamma} \langle a, b \rangle + \langle \Phi^* a, \Phi^* b - f \rangle + T \|a\|_1 + C_1 \\ \gamma \tilde{\mathcal{L}}_1(a, b) &= \frac{1}{2} \|a - \bar{b}\|^2 + T\gamma \|a\|_1 + C_2, \end{aligned}$$

where  $C_1$  and  $C_2$  are two constants that do not depend on  $a$ . This proves that  $\tilde{\mathcal{L}}_1(a, b)$  is a strictly convex function with respect to  $a$ . Since

$$\frac{1}{2} \|a - \bar{b}\|^2 + T\gamma \|a\|_1 = \sum_{p \in \Gamma} (|a[p] - \bar{b}[p]|^2 + T\gamma |a[p]|),$$

each term of this sum can be minimized independently, and Lemma 12.3 proves that the minimum is reached by  $\rho_{\gamma T}(\bar{b}[p])$ . Moreover, at the minimum, the result (12.104) of Lemma 12.3 implies (12.103). Lemma 12.3 is proved by verifying (12.104) with a direct algebraic calculation.

**Lemma 12.3.** The scalar energy  $e(\alpha) = |\beta - \alpha|^2/2 + T|\alpha|$  is minimized by  $\alpha = \rho_T(\beta)$  and

$$e(h + \rho_T(\beta)) - e(\rho_T(\beta)) \geq \frac{|h|^2}{2}. \quad (12.104)$$

■

We now prove that the  $a_k$  defined in equation (12.100) converge to a fixed point of  $\xi$ , and that this fixed point is a minimizer of  $\mathcal{L}_1(a)$ . The difficulty is that  $\xi$  is not strictly contracting. Without loss of generality, we assume that  $\gamma = 1$  and that  $\|\Phi^* b\| < \|b\|$ . Otherwise,  $\Phi^*$  is replaced by  $\sqrt{\gamma}\Phi^*$ ,  $f$  by  $\sqrt{\gamma}f$ , and  $T$  by  $\gamma T$ . The thresholding operator  $\rho_T$  satisfies

$$\forall \alpha, \alpha' \in \mathbb{C}, \quad |\rho_T(\alpha) - \rho_T(\alpha')| \leq |\alpha - \alpha'|.$$

This implies the contractions of the mapping  $\xi$ :

$$\|\xi(a) - \xi(a')\| \leq \|(\text{Id} - \Phi\Phi^*)(a - a')\| \leq \|\text{Id} - \Phi\Phi^*\| \|a - a'\| \leq \|a - a'\|,$$

because  $\|\Phi^*\|_S < 1$ .

In the following, we write  $L = \sqrt{\text{Id} - \Phi\Phi^*}$ , which satisfies  $\|La\|^2 = \|a\|^2 - \|\Phi^* a\|^2$ . Lemma 12.2 shows that  $a_{k+1}$  is the minimizer of  $a \mapsto \tilde{\mathcal{L}}_1(a, a_k)$ , and thus

$$\mathcal{L}_1(a_{k+1}) \leq \mathcal{L}_1(a_{k+1}) + \frac{1}{2} \|L(a_{k+1} - a_k)\|^2 = \tilde{\mathcal{L}}_1(a_{k+1}, a_k) \leq \tilde{\mathcal{L}}_1(a_k, a_k) = \mathcal{L}_1(a_k), \quad (12.105)$$

so  $\{\mathcal{L}_1(a_k)\}_k$  is a nonincreasing sequence. Similarly,

$$\tilde{\mathcal{L}}_1(a_{k+2}, a_{k+1}) \leq \mathcal{L}_1(a_{k+1}) \leq \mathcal{L}_1(a_{k+1}) + \frac{1}{2} \|L(a_{k+1} - a_k)\|^2 = \tilde{\mathcal{L}}_1(a_{k+1}, a_k),$$

so  $\{\tilde{\mathcal{L}}_1(a_{k+1}, a_k)\}_k$  is also a nonincreasing sequence.

Since  $\|\Phi^*\|_S < 1$ , the operator  $L$  is positive definite, and if one denotes  $A > 0$  the smallest eigenvalue of  $L$ ,

$$\sum_{k=0}^K \|a_{k+1} - a_k\|^2 \leq \frac{1}{A} \sum_{k=0}^K \|L(a_{k+1} - a_k)\|^2.$$

It results from (12.105) that  $1/2\|L(a_{k+1} - a_k)\|^2 \leq \mathcal{L}_1(a_k) - \mathcal{L}_1(a_{k+1})$ , and thus

$$\begin{aligned} \sum_{k=0}^K \|a_{k+1} - a_k\|^2 &\leq \frac{1}{2A} \sum_{k=0}^K (\mathcal{L}_1(a_k) - \mathcal{L}_1(a_{k+1})) \\ &= \frac{1}{2A} (\mathcal{L}_1(a_0) - \mathcal{L}_1(a_K)) < \frac{1}{2A} \mathcal{L}_1(a_0). \end{aligned}$$

It follows that the series  $\sum_k \|a_{k+1} - a_k\|^2$  converges, and thus

$$\|\xi(a_k) - a_k\| \rightarrow 0 \quad \text{when } k \rightarrow +\infty. \quad (12.106)$$

Since

$$\|a_k\|_1 \leq \frac{1}{T} \mathcal{L}_1(a_k) \leq \frac{1}{T} \mathcal{L}_1(a_0),$$

the sequence  $\{a_k\}_k$  is bounded. As the vectors are in the finite-dimensional space  $\mathbb{C}^N$ , there exists a subsequence  $\{a_{\gamma(k)}\}_k$  that converges to some  $\tilde{a}$ . Equation (12.106) proves that  $\xi(a_{\gamma(k)})$  also converges to  $\tilde{a}$ , which is thus a fixed point of  $\xi$ .

Since  $\xi$  is contracting,

$$\|a_{k+1} - \tilde{a}\| = \|\xi(a_k - \tilde{a})\| \leq \|a_k - \tilde{a}\|.$$

The sequence  $\{\|a_k - \tilde{a}\|\}_k$  is decreasing and thus convergent. But  $\{\|a_{\gamma(k)} - \tilde{a}\|\}_k$  converges to 0 so the whole sequence  $a_k$  is converging to  $\tilde{a}$ .

Given that  $\xi(\tilde{a}) = \tilde{a}$ , Lemma 12.2 with  $\gamma = 1$  proves that

$$\mathcal{L}_1(\tilde{a} + h) + \frac{1}{2} \|h\|^2 - \frac{1}{2} \|\Phi^* h\|^2 = \tilde{\mathcal{L}}_1(\tilde{a} + h, \tilde{a}) \geq \tilde{\mathcal{L}}_1(\tilde{a}, \tilde{a}) + \frac{1}{2} \|h\|^2 = \mathcal{L}_1(\tilde{a}) + \frac{1}{2} \|h\|^2,$$

which proves that  $\mathcal{L}_1(\tilde{a} + h) \geq \mathcal{L}_1(\tilde{a}) + \frac{1}{2} \|\Phi^* h\|^2$ , and thus  $\tilde{a}$  is a minimizer of  $\mathcal{L}_1$ . ■

This theorem proves the convergence of the algorithm but provides no bound on the decay rate. It incorporates in the loop the Richardson gradient descent so its convergence is slower than the convergence of a Richardson inversion. Theorem 5.7 proves that this convergence depends on frame bounds. In this iterative thresholding algorithm, it depends on the frame bounds  $B_\Lambda \geq A_\Lambda > 0$  of the family of vectors  $\{\phi_p\}_{p \in \Lambda}$  over the support  $\Lambda$  of the solution as it evolves during the convergence. For a gradient descent alone, Theorem 5.7 proves that error decreases by a factor  $\delta = \max\{|1 - \gamma A_\Lambda|, |1 - \gamma B_\Lambda|\}$ . This factor is small if  $A_\Lambda$  is small, which indicates that this family of vectors defines a nearly unstable frame. Once the final support  $\tilde{\Lambda}$  is recovered, the convergence is fast only if  $A_{\tilde{\Lambda}}$  is also not too small. This property guarantees the numerical stability of the solution.

### **Backprojection**

The amplitudes of the coefficients  $\tilde{a}$  on the support  $\tilde{\Lambda}$  are reduced relative to the coefficients  $a_{\tilde{\Lambda}}$  of the orthogonal projection  $f_{\tilde{\Lambda}}$  of  $f$  on the space generated by  $\{\phi_p\}_{p \in \tilde{\Lambda}}$ . The approximation error is reduced by a backprojection that recovers  $a_{\tilde{\Lambda}}$  from  $\tilde{a}$  as in a matching pursuit.

Implementing this backprojection with the Richardson algorithm is equivalent to continuing the iterations with the same gradient step (12.98) and by replacing the soft thresholding (12.99) by an orthogonal projector on  $\tilde{\Lambda}$ :

$$a_{k+1}[p] = \begin{cases} 0 & \text{if } p \notin \tilde{\Lambda} \\ \bar{a}_k[p] & \text{if } p \in \tilde{\Lambda}. \end{cases} \quad (12.107)$$

The convergence is then guaranteed by the Richardson theorem (5.7) and depends on  $A_{\tilde{\Lambda}}$ .

### **Automatic Threshold Updating**

To solve the minimization under an error constraint of  $\varepsilon$ ,

$$\tilde{a} = \underset{a \in \mathbb{R}^P}{\operatorname{argmin}} \|a\|_1 \quad \text{subject to} \quad \|\Phi^* a - f\| \leq \varepsilon, \quad (12.108)$$

the Lagrange multiplier  $T$  must be adjusted to  $\varepsilon$ . A sequence of  $\tilde{a}_l = \underset{a \in \mathbb{C}^P}{\operatorname{argmin}} \frac{1}{2} \|f - \Phi^* a\|^2 + T_l \|a\|_1$  can be calculated so that  $\|\Phi^* \tilde{a}_l - f\|$  converges to  $\varepsilon$ . The error  $\|f - \Phi^* \tilde{a}_l\|$  is an increasing function of  $T_l$  but not strictly. A possible threshold adjustment proposed by Chambolle [152] is

$$T_{l+1} = T_l \frac{\varepsilon}{\|f - \Phi^* \tilde{a}_l\|}. \quad (12.109)$$

One can also update a threshold  $T_k$  with (12.109) at each step  $k$  of the soft-thresholding iteration, which works numerically well, although there is no proof of convergence.

### **Other Algorithms**

Several types of algorithms can solve the  $1^1$  Lagrangian minimization (12.97). It includes primal-dual schemes [497], specialized interior points with preconditioned conjugate gradient [329], Bregman iterations [494], split Bregman iterations [275], two-step iterative thresholding [211], SGPL1 [104], gradient pursuit [115], gradient projection [256], fixed-point continuation [240], gradient methods [389, 390], coordinate-wise descent [261], and sequential subspace optimization [382].

Continuation methods like homotopy [228, 393] keep track of the solution  $\tilde{a}_l$  of (12.97) for a decreasing sequence of thresholds  $T_l$ . For a given  $T_l$ , one can compute the next smallest  $T_{l+1}$  where the support of the optimal solution includes a new component (or more rarely where a new component disappears) and the position of this component. At each iteration, the solution is then computed with the implicit equation (12.93) by calculating the pseudo inverse  $\Phi_{\Lambda}^{*+} f$ . These algorithms are thus

quite fast if the final solution  $\tilde{a}$  is highly sparse, because only a few iterations are then necessary and the size of all matrices remains small.

It is difficult to compare all these algorithms because their speed of convergence depends on the sparsity of the solution and on the frame bounds of the dictionary vectors over the support solution. The iterative thresholding algorithm of Theorem 12.9 has the advantage of simplicity.

#### 12.4.4 Sparse Synthesis versus Analysis and Total Variation Regularization

Matching pursuit and basis pursuit algorithms assume that the signal has a *sparse synthesis* in a dictionary  $\mathcal{D} = \{\phi_p\}_{p \in \Gamma}$  and compute this representation with as few vectors as possible. The sparse synthesis hypothesis should not be confused with a *sparse analysis* hypothesis, which assumes that a linear signal transform  $\Phi f = \{\langle f, \phi_p \rangle\}_{p \in \Gamma}$  is sparse. A sparse analysis can often be related to some form of regularity of  $f$ . For example, piecewise regular signals have a sparse wavelet transform. Similarly, total variation regularizations make a sparse analysis assumption on the sparsity of the gradient vector field. Sparse analysis and synthesis assumptions are equivalent in an orthonormal basis, but are very different when the dictionary is redundant. These aspects are clarified and algorithms are provided to solve sparse analysis problems, including total variation regularizations.

##### **Sparse Synthesis and Analysis with $\mathbf{l}^1$ Norms**

A basis pursuit incorporates the sparse synthesis assumption by minimizing the  $\mathbf{l}^1$  norm of the synthesis coefficients. To approximate  $f$ , the Lagrangian formulation computes

$$\tilde{f}_s = \Phi^* \tilde{a} \quad \text{with} \quad \tilde{a} = \operatorname{argmin}_{a \in \mathbb{C}^P} \frac{1}{2} \|f - \Phi^* a\|^2 + T \|a\|_1. \quad (12.110)$$

In a denoising problem,  $f$  is replaced by the input noisy signal  $X = f + W$ , and if  $W$  is a Gaussian white noise of variance  $\sigma^2$ , then  $T$  is proportional to  $\sigma$ .

A sparse analysis approximation of  $f$  computes

$$\tilde{f}_a = \operatorname{argmin}_{f \in \mathbb{R}^N} \|\Phi f\|_1 \quad \text{with} \quad \|f - y\| \leq \varepsilon, \quad (12.111)$$

where  $\varepsilon$  is the approximation precision. This problem is convex. Similarly, in a denoising problem,  $f$  is replaced by the input noisy signal  $X = f + W$ . A solution  $\tilde{f}_a$  of (12.111) can be computed as a solution of the Lagrangian formulation

$$\tilde{f}_a = \operatorname{argmin}_{h \in \mathbb{R}^N} \frac{1}{2} \|f - h\|^2 + T \|\Phi h\|_1, \quad (12.112)$$

where  $T$  depends on  $\varepsilon$ . In a denoising problem,  $f$  is also replaced by the noisy signal  $X = f + W$ , and  $T$  is proportional to the noise standard deviation  $\sigma$ .

If  $\mathcal{D}$  is an orthonormal basis, then  $\Phi^* = \Phi^{-1}$  so  $\tilde{f}_s = \tilde{f}_a$ . The analysis and synthesis prior assumptions are then equivalent. This is, however, not the case when the dictionaries are redundant [243]. In a redundant dictionary, a sparse synthesis assumes that a signal is well approximated by few well-chosen dictionary vectors. However, it often has a nonsparse representation with badly chosen dictionary vectors. For example, a high-frequency oscillatory texture has a sparse synthesis in a dictionary of wavelet and local cosine because it is well represented by local cosine vectors, but it does not have a sparse analysis representation in this dictionary because the wavelet coefficients of these textures are not sparse.

Thresholding algorithms in redundant translation-invariant dictionaries such as translation-invariant wavelet frames rely on a sparse analysis assumption. They select all large frame coefficients, and the estimation is precise if there are relatively few of them. Sparse analysis constraints can often be interpreted as imposing some form of signal regularity condition—for example, with a total variation norm.

In Section 12.4.3 we describe an iterative thresholding that solves the sparse synthesis Lagrangian minimization (12.110). When the dictionary is redundant, the sparse analysis minimization (12.112) must integrate that the  $L^1$  norm is carried over a redundant set of coefficients  $a[p] = \Phi h[p]$  that cannot be adjusted independently. If  $\Phi$  is a frame, then Theorem 5.9 proves that  $a$  satisfies a reproducing kernel equation  $a = \Phi\Phi^+a$  where  $\Phi\Phi^+$  is an orthogonal projector on the space of frame coefficients. One could think of solving (12.112) with the iterative soft-thresholding algorithm that minimizes (12.110), and project with the projector  $\Phi\Phi^+a$  at each iteration on the constraints, but this algorithm is not guaranteed to converge.

### Total Variation Denoising

Rudin, Osher, and Fatemi's [420] total variation regularization algorithm assumes that the image gradient  $\vec{\nabla}f$  is sparse, which is enforced by minimizing its  $L^1$  norm, and thus the total image variation  $\iint |\vec{\nabla}f(x)| dx$ . Over discrete images, the gradient vector is computed with horizontal and vertical finite differences. Let  $\tau_1 = (1, 0)$  and  $\tau_2 = (0, 1)$ :

$$D_k f[p] = f[p] - f[p - \tau_k] = \langle f, \phi_p^k \rangle \quad \text{with} \quad \phi_p^k = \delta[n - p] - \delta[n - p - \tau_k] \quad \text{for } k = 1, 2.$$

The discrete total variation norm is the complex  $L^1$  norm

$$\|f\|_V = \sum_p \sqrt{|D_1 f[p]|^2 + |D_2 f[p]|^2} = \|\Phi f\|_1,$$

where  $\Phi$  is a complex valued analysis operator

$$\Phi f = D_1 f + i D_2 f = \{\langle f, \phi_p \rangle\}_p \quad \text{for} \quad \phi_p = \phi_p^1 + i \phi_p^2.$$

The discrete image gradient  $\Phi f = D_1 f + i D_2 f$  has the same size  $N$  as the image but has twice the scalar values because of the two partial derivative coordinates. Thus, it defines a redundant representation in  $\mathbb{R}^N$ . The finite difference  $D_1$  and  $D_2$  are convolution operators with transfer functions that vanish at the 0 frequency. As

**FIGURE 12.24**

(a) Noisy image  $X = f + w$  (SNR = 16 dB). (b) Translation-invariant wavelet thresholding estimation (SNR = 22.9 dB). (c) Estimation by total variation regularization (SNR = 21.9 dB).

a result,  $\Phi$  is not a frame of  $\mathbb{R}^N$ . However, it is invertible on the subspace  $\mathbf{V} = \{f \in \mathbb{R}^N : \sum_n f[n] = 0\}$  of dimension  $N - 1$ . Since we are in finite dimension, it defines a frame of  $\mathbf{V}$ , but the pseudo inverse  $\Phi^+$  in (5.21) becomes numerically unstable when  $N$  increases because the frame bound ratio  $A/B$  tends to zero. Theorem 5.9 proves that  $\Phi\Phi^+$  remains an orthogonal projector on the space of gradient coefficients.

Figure 12.24 shows an example of image denoising with a total variation regularization. The total variation minimization recovers an image with a gradient vector that is as sparse as possible, which has a tendency to produce piecewise constant regions when the thresholding increases. As a result, the total variation image denoising most often yields a smaller SNR than wavelet thresholding estimators, unless the image is piecewise constant.

### **Computation of Sparse Analysis Approximations and Denoising with $\mathbb{L}^1$ Norms**

A sparse analysis approximation is defined by

$$\tilde{f}_a = S_T(f) = \underset{h \in \mathbb{R}^N}{\operatorname{argmin}} \frac{1}{2} \|f - h\|^2 + T \|\Phi h\|_1, \quad (12.113)$$

and  $f$  is replaced by the noisy data  $X = f + W$  in a denoising problem. Chambolle [152] sets a dual problem by verifying that the regularization term  $\|\Phi h\|_1$  can be replaced by a set of dual variables:

$$\|\Phi h\|_1 = \max_{r \in \mathcal{K}} \langle r, h \rangle \quad \text{where} \quad \mathcal{K} = \{\Phi^* a : \|a\|_\infty \leq 1\}, \quad (12.114)$$

with  $\|a\|_\infty = \max_p |a[p]|$ , and where  $|\cdot|$  is the modulus of complex numbers. Because of (12.114), the regularization  $J(h) = \|\Phi h\|_1$  is “linearized” by taking a maximum over inner products with vectors in a convex set  $r \in \mathcal{K}$ . This decomposition remains valid for any positively homogeneous functional  $J$ , which satisfies  $J(\lambda h) = |\lambda| J(h)$ .

The minimization (12.113) is thus rewritten as a minimization and maximization:

$$\tilde{f}_a = \operatorname{argmin}_{h \in \mathbb{R}^N} \frac{1}{2} \|f - h\|^2 + T\|\Phi h\|_1 = \operatorname{argmin}_{h \in \mathbb{R}^N} \max_{r \in \mathcal{K}} \frac{1}{2} \|f - h\|^2 + T\langle r, h \rangle.$$

Inverting the min and max and computing the solution of the min problem shows that  $\tilde{f}_a$  is written as

$$\tilde{f}_a = f - T\tilde{r} \quad \text{where} \quad \tilde{r} = \operatorname{argmin}_{r \in \mathcal{K}} \|f - Tr\|,$$

which means that  $\tilde{r}$  is the projection of  $f/T$  on the convex set  $\mathcal{K}$ .

The solution of (12.113) is thus also the solution of the following convex problem:

$$\tilde{f}_a = f - T\Phi^*\tilde{a} \quad \text{where} \quad \tilde{a} = \operatorname{argmin}_{\|a\|_\infty \leq 1, a \in \mathbb{C}^P} \|\Phi^*a - f/T\|. \quad (12.115)$$

The solution  $\tilde{a}$  is not necessarily unique, but both  $\tilde{r} = \Phi^*\tilde{a}$  and  $\tilde{f}_a$  are unique. With this dual formulation, the redundancy of  $a = \Phi f$  in the  $\mathbf{l}^1$  norm of (12.113) does not appear anymore.

Starting from a choice of  $a_0$ , for example  $a_0 = 0$ , Chambolle [152] computes  $\tilde{a}$  by iterating between a gradient step to minimize  $\|\Phi^*a - f/T\|^2$ ,

$$b^{(k+1)} = \Phi(\Phi^*a^{(k)} - f/T),$$

and a “projection” step to ensure that  $|a[p]| \leq 1$  for all  $p$ . An orthogonal projection at each iteration on the constraint gives

$$\tilde{a}^{(k+1)} = a^{(k)}[p] + \gamma b^{(k+1)}[p] \quad \text{and} \quad a^{(k+1)}[p] = \frac{\tilde{a}^{(k+1)}[p]}{\max(|\tilde{a}^{(k+1)}[p]|, 1)}.$$

One can verify [152] that the convergence is guaranteed if  $\gamma < 2 \|\Phi\Phi^*\|_S^{-1}$ . To satisfy the constraints, another possibility is to set

$$a^{(k+1)}[p] = \frac{a^{(k)}[p] + \gamma b^{(k+1)}[p]}{1 + \gamma |b^{(k+1)}[p]|}. \quad (12.116)$$

For the gradient operator  $\Phi^* = \nabla$  discretized with finite differences in 2D,  $\|\Phi\Phi^*\|_S = 1/8$  so one can choose  $\gamma = 1/4$ . The convergence is often fast during the first few iterations, leading to a visually satisfying result, but the remaining iterations tend to converge slowly to the final solution.

Other iterative algorithms have been proposed to solve (12.113), for instance, fixed-point iterations [476], second-order cone programming [274], splitting [480], splitting with Bregman iterations [275], and primal-dual methods [153, 497]. Primal-dual methods tend to have a faster convergence than the Chambolle algorithm.

Chambolle [152] proves that the sparse analysis minimization (12.111) with a precision  $\varepsilon$  is obtained by iteratively computing the solution  $\tilde{a}_l$  of Lagrangian problems (12.113) for thresholds  $T_l$  that are adjusted with

$$T_{l+1} = T_l \frac{\varepsilon}{\|f - \Phi^*\tilde{a}_l\|}. \quad (12.117)$$

The convergence proof relies on the fact that  $\|f - S_T(f)\|$  is a strictly increasing function of  $T$  for the analysis problem, which is not true for the synthesis problem.

### **Computation of Sparse Analysis Inverse Problems with $L^1$ Norms**

In inverse problems studied in Chapter 13,  $f$  is estimated from  $Y = Uh + W$ , where  $U$  is a linear operator and  $W$  an additive noise. A sparse analysis estimation of  $f$  computes

$$\tilde{F}_a = \operatorname{argmin}_{h \in \mathbb{R}^N} \frac{1}{2} \|Y - Uh\|^2 + T \|\Phi h\|_1. \quad (12.118)$$

The sparse Lagrangian analysis approximation operator  $S_T$  in (12.113) can be used to replace the wavelet soft thresholding in the iterative thresholding algorithm (12.100), which leads to the iteration

$$\tilde{F}_{k+1} = S_{\gamma T} (\tilde{F}_k + \mu U^* (Y - U \tilde{F}_k)).$$

One can prove with the general machinery of proximal iterations that  $\tilde{F}_k$  converges to a solution  $\tilde{F}$  of (12.118) if  $\mu < \|U^* U\|_S^{-1}$  [186]. The algorithm is implemented with two embedded loops. The outer loop on  $k$  computes  $\tilde{F}_k + \mu U^* (Y - U \tilde{F}_k)$ , followed by the inner loop, which computes the Lagrangian approximation  $S_T$ , for example, with Chambolle algorithm.

## **12.5 PURSUIT RECOVERY**

Matching pursuits and basis pursuits are nonoptimal sparse approximation algorithms in a redundant dictionary  $\mathcal{D}$ , but are computationally efficient. However, pursuit approximations can be nearly as precise as optimal  $M$ -term approximations, depending on the properties of the approximation supports in  $\mathcal{D}$ .

Beyond approximation, this section studies the ability of pursuit algorithms to recover a specific set  $\Lambda$  of vectors providing a sparse signal approximation in a redundant dictionary. The stability of this recovery is important for pattern recognition when selected dictionary vectors are used to analyze the signal information. It is also at the core of the sparse super-resolution algorithms introduced in Section 13.3.

The stability of sparse signal approximations in redundant dictionaries is related to the dictionary coherence in Section 12.5.1. The exact recovery of signals with matching pursuits and basis pursuits are studied in Sections 12.5.2 and 12.5.3, together with the precision of  $M$ -term approximations.

### **12.5.1 Stability and Incoherence**

Let  $\mathcal{D} = \{\phi_p\}_{p \in \Gamma}$  be a redundant dictionary of normalized vectors. Given a family of linearly independent approximation vectors  $\{\phi_p\}_{p \in \Lambda}$  selected by some algorithm,

the best approximation of  $f$  is its orthogonal projection  $f_\Lambda$  in the space  $\mathbf{V}_\Lambda$  generated by these vectors

$$f_\Lambda = \sum_{p \in \Lambda} a[p] \phi_p.$$

The calculation of the coefficients  $a[p]$  is stable if  $\{\phi_p\}_{p \in \Lambda}$  is a Riesz basis of  $\mathbf{V}_\Lambda$ , with Riesz bounds  $B_\Lambda \geq A_\Lambda > 0$ , which satisfy:

$$\forall a \in \mathbb{C}^{|\Lambda|}, \quad A_\Lambda \|a\|^2 \leq \left\| \sum_{p \in \Lambda} a[p] \phi_p \right\|^2 \leq B_\Lambda \|a\|^2. \quad (12.119)$$

The closer  $A_\Lambda/B_\Lambda$  to 1, the more stable the basis. Gradient descent algorithms compute the coefficients  $a[p]$  with a convergence rate that also depends on  $A_\Lambda/B_\Lambda$ . Theorem 12.10 relates the Riesz bounds to the dictionary *mutual coherence*, introduced by Donoho and Huo [230],

$$\mu(\mathcal{D}) = \sup_{(q,q) \in \Gamma^2, p \neq q} |\langle \phi_p, \phi_q \rangle|.$$

**Theorem 12.10.** The Riesz bounds of  $\{\phi_p\}_{p \in \Lambda}$  satisfy

$$\delta_\Lambda = \max(1 - A_\Lambda, B_\Lambda - 1) \leq \max_{p \in \Lambda} \sum_{q \in \Lambda, p \neq q} |\langle \phi_p, \phi_q \rangle| \leq (|\Lambda| - 1) \mu(\mathcal{D}). \quad (12.120)$$

**Proof.** Theorem 5.1 proves that the constants  $A_\Lambda$  and  $B_\Lambda$  are lower and upper bounds of the eigenvalues of the Gram matrix  $G_\Lambda = \{\langle \phi_p, \phi_q \rangle\}_{(p,q) \in \Lambda^2}$ . Let  $v \neq 0$  be an eigenvector satisfying  $G_\Lambda v = \lambda v$ . Let  $|v[p]| = \max_{q \in \Lambda} |v[q]|$ . Since  $\|\phi_p\|^2 = 1$ ,

$$v[p] + \sum_{q \in \Lambda, q \neq p} \langle \phi_p, \phi_q \rangle v[q] = \lambda v[p] \implies |1 - \lambda| \leq \sum_{q \in \Lambda, q \neq p} |\langle \phi_p, \phi_q \rangle| \frac{|v[q]|}{|v[p]|}.$$

It results that

$$\delta_\Lambda = \max_\lambda |1 - \lambda| \leq \sum_{q \in \Lambda, q \neq p} |\langle \phi_p, \phi_q \rangle| \leq (|\Lambda| - 1) \mu(\mathcal{D}),$$

which proves (12.120). ■

The Riesz bound ratio  $A_\Lambda/B_\Lambda$  is close to 1 if  $\delta_\Lambda$  is small and thus if the vectors in  $\Lambda$  have a small correlation. The upper bound (12.120) proves that sufficiently small sets  $|\Lambda| < \mu(\mathcal{D})$  are Riesz bases with  $A_\Lambda > 0$ . To increase the maximum size of these sets, one should construct dictionaries that are as *incoherent* as possible. The upper bounds (12.120) are simple but relatively crude, because they only depend on inner products of pairs of vectors, whereas the Riesz stability of  $\{\phi_p\}_{p \in \Lambda}$  depends on the distribution of this whole group of vectors on the unit sphere of  $\mathbb{C}^N$ . Section 13.4 proves that much better bounds can be calculated over dictionaries of random vectors.

The mutual coherence of an orthonormal basis  $\mathcal{D}$  is 0, and one can verify (Exercise 12.5) that if we add a vector  $g$ , then  $\mu(\mathcal{D} \cup \{g\}) \geq \frac{1}{\sqrt{N}}$ . However, this level of incoherence can be reached with larger dictionaries. Let us consider the dictionary of  $P = 2N$  vectors, which is a union of a Dirac basis and a discrete Fourier basis:

$$\mathcal{D} = \{\delta[n-p]\}_{0 \leq p < N} \cup \{N^{-1/2} e^{-i2\pi pn/N}\}_{0 \leq p < N}. \quad (12.121)$$

Its mutual coherence is  $\mu(\mathcal{D}) = N^{-1/2}$ . The right upper bound of (12.120) thus proves that any family of  $|\Lambda| \leq \sqrt{N}/2$  Dirac and Fourier vectors defines a basis with a Riesz bound ratio  $A_\Lambda/B_\Lambda \geq 1/3$ . One can construct larger frames  $\mathcal{D}$  of  $P = N^2$  vectors, called Grassmannian frames, that have a coherence in  $O(1/\sqrt{N})$  [448].

Dictionaries often do not have a very small coherence, so the inequality  $\delta_\Lambda \leq (|\Lambda| - 1)\mu(\mathcal{D})$  applies to relatively small sets  $\Lambda$ . For example, in the Gabor dictionary  $\mathcal{D}_{j,\Delta}$  defined in (12.77), the mutual coherence  $\mu(\mathcal{D})$  is maximized by two neighboring Gabor atoms, and the inner product formula (12.81) proves that  $\mu(\mathcal{D}) = e^{-\pi\Delta^{-2}/2} = 0.67$  for  $\Delta = 2$ . The mutual coherence upper bound is therefore useless in this case, but the first upper bound of (12.120) can be used with (12.81) to verify that Gabor atoms that are sufficiently far in time and frequency generate a Riesz basis.

## 12.5.2 Support Recovery with Matching Pursuit

We first study the reconstruction of signals  $f$  that have an exact sparse representation in  $\mathcal{D}$ ,

$$f = \sum_{p \in \Lambda} a[p] \phi_p.$$

An exact recovery condition on  $\Lambda$  is established to guarantee that a matching pursuit selects only approximation vectors in  $\Lambda$ . The optimality of matching pursuit approximations is then analyzed for more general signals. We suppose that the matching pursuit relaxation factor is  $\alpha = 1$ .

An orthogonal matching pursuit, or a matching pursuit with backprojection, computes the orthogonal projection  $f_{\tilde{\Lambda}}$  of  $f$  on a family of vectors  $\{\phi_p\}_{p \in \tilde{\Lambda}}$  selected one by one. At a step  $m$ , a matching pursuit selects an atom in  $\Lambda$  if and only if the correlation of the residual  $R^m f$  with vectors in the complement  $\Lambda^c$  of  $\Lambda$  is smaller than the correlation with vectors in  $\Lambda$ :

$$C(R^m f, \Lambda^c) = \frac{\max_{q \in \Lambda^c} |\langle R^m f, \phi_q \rangle|}{\max_{p \in \Lambda} |\langle R^m f, \phi_p \rangle|} < 1. \quad (12.122)$$

The relative correlation of a vector  $h$  with vectors in  $\Lambda^c$  relative to  $\Lambda$  is defined by

$$C(h, \Lambda^c) = \frac{\max_{q \in \Lambda^c} |\langle h, \phi_q \rangle|}{\max_{p \in \Lambda} |\langle h, \phi_p \rangle|}. \quad (12.123)$$

Theorem 12.11, proved by Tropp [461], gives a condition that guarantees the recovery of  $\Lambda$  with a matching pursuit.

**Theorem 12.11:** *Tropp.* If  $\{\tilde{\phi}_{p,\Lambda}\}_{p \in \Lambda}$  is the dual basis of  $\{\phi_p\}_{p \in \Lambda}$  in the space  $\mathbf{V}_\Lambda$ , then

$$\text{ERC}(\Lambda) = \max_{q \in \Lambda^c} \sum_{p \in \Lambda} |\langle \tilde{\phi}_{p,\Lambda}, \phi_q \rangle| = \sup_{h \in \mathbf{V}_\Lambda} C(h, \Lambda^c). \quad (12.124)$$

If  $f \in \mathbf{V}_\Lambda$  and the following exact recovery condition (ERC) is satisfied,

$$\text{ERC}(\Lambda) < 1, \quad (12.125)$$

then a matching pursuit of  $f$  selects only vectors in  $\{\phi_p\}_{p \in \Lambda}$  and an orthogonal matching pursuit recovers  $f$  with at most  $|\Lambda|$  iterations.

**Proof.** Let us first prove that  $\sup_{h \in \mathbf{V}_\Lambda} C(h, \Lambda^c) \leq \text{ERC}(\Lambda)$ . Let  $\Phi_\Lambda^+$  be the pseudo inverse of  $\Phi_\Lambda$  with  $\Phi_\Lambda f[p] = \langle f, \phi_p \rangle$  for  $p \in \Lambda$ . Theorem 5.6 proves that  $\Phi_\Lambda^+ \Phi_\Lambda$  is an orthogonal projector in  $\mathbf{V}_\Lambda$ , so if  $h \in \mathbf{V}_\Lambda$  and  $q \in \Lambda^c$ ,

$$\begin{aligned} |\langle h, \phi_q \rangle| &= |\langle \Phi_\Lambda^+ \Phi_\Lambda h, \phi_q \rangle| = |\langle \Phi_\Lambda h, \\ (\Phi_\Lambda^+)^* \phi_q \rangle| &\leq \|\Phi_\Lambda h\|_\infty \max_{q \in \Lambda^c} \|(\Phi_\Lambda^+)^* \phi_q\|_1. \end{aligned} \quad (12.126)$$

Theorem 5.5 proves that the dual-frame operator satisfies  $\tilde{\Phi}_\Lambda^* = \Phi_\Lambda^+$  and thus that  $(\Phi_\Lambda^+)^* f[p] = \tilde{\Phi}_\Lambda[p] = \langle f, \tilde{\phi}_{p,\Lambda} \rangle$  for  $p \in \Lambda$ . It results that

$$\text{ERC}(\Lambda) = \max_{q \in \Lambda^c} \|(\Phi_\Lambda^+)^* \phi_q\|_1 = \max_{q \in \Lambda^c} \sum_{p \in \Lambda} |\langle \tilde{\phi}_{p,\Lambda}, \phi_q \rangle|, \quad (12.127)$$

and (12.126) implies that

$$\forall h \in \mathbf{V}_\Lambda, \quad \max_{q \in \Lambda^c} |\langle h, \phi_q \rangle| \leq \max_{p \in \Lambda} |\langle h, \phi_p \rangle| \text{ERC}(\Lambda),$$

which proves that  $\text{ERC}(\Lambda) \geq \sup_{h \in \mathbf{V}_\Lambda} C(h, \Lambda^c)$ .

We now prove the reverse inequality. Let  $q_0 \in \Lambda^c$  be the index such that

$$\sum_{p \in \Lambda} |\langle \tilde{\phi}_{p,\Lambda}, \phi_{q_0} \rangle| = \max_{q \in \Lambda^c} \sum_{p \in \Lambda} |\langle \tilde{\phi}_{p,\Lambda}, \phi_q \rangle|.$$

Introducing

$$h = \sum_{p \in \Lambda} \text{sign}(\langle \tilde{\phi}_{p,\Lambda}, \phi_{q_0} \rangle) \tilde{\phi}_{p,\Lambda} \in \mathbf{V}_\Lambda \quad (12.128)$$

leads to

$$\begin{aligned} \text{ERC}(\Lambda) &= \max_{q \in \Lambda^c} \sum_{p \in \Lambda} |\langle \tilde{\phi}_{p,\Lambda}, \phi_q \rangle| = |\langle h, \phi_{q_0} \rangle| \\ &\leq \max_{q \in \Lambda^c} |\langle h, \phi_q \rangle| \leq C(h, \Lambda^c) \max_{p \in \Lambda} |\langle h, \phi_p \rangle|. \end{aligned}$$

Since  $|\langle h, \phi_p \rangle| = |\text{sign}(\langle \tilde{\phi}_{p,\Lambda}, \phi_{q_0} \rangle)| = 1$ , it results that  $C(h, \Lambda^c) \geq \text{ERC}(\Lambda)$  and thus  $\sup_{h \in \mathbf{V}_\Lambda} C(h, \Lambda^c) \geq \text{ERC}(\Lambda)$ , which finishes the proof of (12.124).

Suppose now that  $f = R^0 f \in \mathbf{V}_\Lambda$  and  $\text{ERC}(\Lambda) < 1$ . We prove by induction that a matching pursuit selects only vectors in  $\{\phi_p\}_{p \in \Lambda}$ . Suppose that the first  $m < M$  matching pursuit

vectors are in  $\{\phi_p\}_{p \in \Lambda}$  and thus that  $R^m f \in \mathbf{V}_\Lambda$ . If  $R^m f \neq 0$ , then (12.125) implies that  $C(R^m f, \Lambda^c) < 1$  and thus that the next vector is selected in  $\Lambda$ . Since  $\dim(\mathbf{V}_\Lambda) \leq |\Lambda|$ , an orthogonal pursuit converges in less than  $|\Lambda|$  iterations. ■

This theorem proves that if a signal can be exactly decomposed over  $\{\phi_p\}_{p \in \Lambda}$ , then  $\text{ERC}(\Lambda) < 1$  guarantees that a matching pursuit reconstructs  $f$  with vectors in  $\{\phi_p\}_{p \in \Lambda}$ . A nonorthogonal pursuit may, however, require more than  $|\Lambda|$  iterations to select all vectors in this family.

If  $\text{ERC}(\Lambda) > 1$ , then there exists  $f \in \mathbf{V}_\Lambda$  such that  $C(f, \Lambda^c) > 1$ . As a result, there exists a vector  $\phi_q$  with  $q \in \Lambda^c$ , which correlates  $f$  better than any other vectors in  $\Lambda$ , and that will be selected by the first iteration of a matching pursuit. This vector may be removed, however, from the approximation support at the end of the decomposition. Indeed, if the remaining iterations select all vectors  $\{\phi_p\}_{p \in \Lambda}$ , then an orthogonal matching pursuit decomposition or a backprojection will associate a coefficient 0 to  $\phi_q$  because  $f = P_{\mathbf{V}_\Lambda} f$ . In particular, we may have  $\text{ERC}(\Lambda) > 1$  but  $\text{ERC}(\tilde{\Lambda}) < 1$  with  $\Lambda \subset \tilde{\Lambda}$ , in which case an orthogonal pursuit recovers exactly the support of any  $f \in \mathbf{V}_\Lambda$  with  $|\tilde{\Lambda}| \geq |\Lambda|$  iterations.

Theorem 12.12 gives an upper bound proved by Tropp [461], which relates  $\text{ERC}(\Lambda)$  to the support size  $|\Lambda|$ . A tighter bound proved by Gribonval and Nielsen [281] and Dossal [231] depends on inner products of dictionary vectors in  $\Lambda$  relative to vectors in the complement  $\Lambda^c$ .

**Theorem 12.12:** *Tropp, Gribonval, Nielsen, Dossal.* For any  $\{\phi_p\}_{p \in \Lambda} \subset \mathcal{D}$ ,

$$\text{ERC}(\Lambda) \leq \frac{\max_{q \in \Lambda^c} \sum_{p \in \Lambda} |\langle \phi_p, \phi_q \rangle|}{1 - \max_{q \in \Lambda} \sum_{p \in \Lambda, p \neq q} |\langle \phi_p, \phi_q \rangle|} \leq \frac{|\Lambda| \mu(\mathcal{D})}{1 - (|\Lambda| - 1) \mu(\mathcal{D})}. \quad (12.129)$$

**Proof.** It is shown in (12.127) that  $\text{ERC}(\Lambda) = \max_{q \in \Lambda^c} \|(\Phi_\Lambda^*)^* \phi_q\|_1$ . We verify that

$$(\Phi_\Lambda^*)^* h = \Phi_\Lambda (\Phi_\Lambda^* \Phi_\Lambda)^{-1} h = (\Phi_\Lambda^*)^+ h = (\Phi_\Lambda \Phi_\Lambda^*)^{-1} \Phi_\Lambda h, \quad (12.130)$$

and thus that we can also write  $\text{ERC}(\Lambda) = \max_{q \in \Lambda^c} \|(\Phi_\Lambda \Phi_\Lambda^*)^{-1} \Phi_\Lambda \phi_q\|_1$ . Introducing the operator norm associated to the  $\mathbf{l}^1$  norm

$$\|A\|_{1,1} = \max_{b \neq 0} \frac{\|Ab\|_1}{\|b\|_1} = \max_j \sum_i |a_{i,j}|$$

for a matrix  $A = (a_{i,j})_{i,j}$ , leads to

$$\text{ERC}(\Lambda) = \max_{q \in \Lambda^c} \|(\Phi_\Lambda \Phi_\Lambda^*)^{-1} \Phi_\Lambda \phi_q\|_1 \leq \max_{q \in \Lambda^c} \|(\Phi_\Lambda \Phi_\Lambda^*)^{-1}\|_{1,1} \max_{q \in \Lambda^c} \|\Phi_\Lambda \phi_q\|_1. \quad (12.131)$$

The second term is the numerator of (12.129),

$$\max_{q \in \Lambda^c} \|\Phi_\Lambda \phi_q\|_1 = \max_{q \in \Lambda^c} \sum_{p \in \Lambda} |\langle \phi_p, \phi_q \rangle|. \quad (12.132)$$

The Gram matrix is rewritten as  $G = \Phi_\Lambda \Phi_\Lambda^* = \text{Id} + H$ , and a Neumann expansion of  $G^{-1}$  gives

$$\|(\Phi_\Lambda \Phi_\Lambda^*)^{-1}\|_{1,1} \leq \sum_{k \geq 0} (\|H\|_{1,1})^k \leq \frac{1}{1 - \|H\|_{1,1}}, \quad (12.133)$$

with

$$\|H\|_{1,1} = \max_{q \in \Lambda} \sum_{p \in \Lambda, p \neq q} |\langle \phi_p, \phi_q \rangle|. \quad (12.134)$$

Inserting this result in (12.133) and inserting (12.132) and (12.133) in (12.131) prove the first inequality of (12.129).

The second inequality is derived from the fact that  $|\langle \phi_p, \phi_q \rangle| \leq \mu(\mathcal{D})$  for any  $p \neq q$ . ■

This theorem gives upper bounds of  $\text{ERC}(\Lambda)$  that can easily be computed. It proves that  $\text{ERC}(\Lambda) < 1$  if the vectors  $\{\phi_p\}_{p \in \Lambda}$  are not too correlated between themselves and with the vectors in the complement  $\Lambda^c$ . In a Gabor dictionary  $\mathcal{D}_{j,\Delta}$  defined in (12.77), the upper bound (12.129) with the Gabor inner product formula (12.81) proves that  $\text{ERC}(\Lambda) < 1$  for families of sufficiently separated time-frequency Gabor atoms indexed by  $\Lambda$ . Theorem 12.11 proves that any combination of such Gabor atoms is recovered by a matching pursuit. Suppose that the Gabor atoms are defined with a Gaussian window that has a variance in time and frequency that is  $\sigma_t^2$  and  $\sigma_\omega^2$ , respectively. Their Heisenberg box has a size  $\sigma_t \times \sigma_\omega$ . If  $|\Lambda| = 2$ , then one can verify that  $\text{ERC}(\Lambda) > 1$  if the Heisenberg boxes of these two atoms intersect. If the time distance of the two Gabor atoms is larger than  $1.5 \sigma_t$ , or if the frequency distance is larger than  $1.5 \sigma_\omega$ , then  $\text{ERC}(\Lambda) < 1$ .

The second upper bound in (12.129) proves that  $\text{ERC}(\Lambda) < 1$  for any sufficiently small set  $\Lambda$

$$|\Lambda| < \frac{1}{2} \left( 1 + \frac{1}{\mu(\mathcal{D})} \right). \quad (12.135)$$

Very sparse approximation sets are thus more easily recovered. The Dirac and Fourier dictionary (12.121) has a low mutual coherence  $\mu(\mathcal{D}) = N^{-1/2}$ . Condition (12.135) together with Theorem 12.11 proves that any combination of  $|\Lambda| \leq N^{1/2}/2$  Fourier and Dirac vectors are recovered by a matching pursuit. The upper bound (12.135) is, however, quite brutal, and in a Gabor dictionary where  $\mu(\mathcal{D}_{j,\Delta}) = e^{-\pi\Delta^{-2}/2}$  and  $\Delta \geq 2$ , it applies only to  $|\Lambda| = 1$ , which is useless.

### Nearly Optimal Approximations with Matching Pursuits

Let  $f_\Lambda$  be the best approximation of  $f$  from  $M = |\Lambda|$  vectors in  $\mathcal{D}$ . If  $f_\Lambda = f$  and  $\text{ERC}(\Lambda) < 1$ , then Theorem 12.11 proves that  $f_\Lambda$  is recovered by a matching pursuit, but it is rare that a signal is exactly a combination of few dictionary vectors. Theorem 12.13, proved by Tropp [461], shows that if  $\text{ERC}(\Lambda) < 1$ , then  $|\Lambda|$  orthogonal

matching pursuit iterations recover the “main approximation vectors” in  $\Lambda$ , and thus produce an error comparable to  $\|f - f_\Lambda\|$ .

**Theorem 12.13:** *Tropp.* Let  $A_\Lambda > 0$  be the lower Riesz bound of  $\{\phi_p\}_{p \in \Lambda}$ . If  $\text{ERC}(\Lambda) < 1$ , then an orthogonal matching pursuit approximation  $\tilde{f}_M$  on  $f$  computed with  $M = |\Lambda|$  iterations satisfies

$$\|f - \tilde{f}_M\|^2 \leq \left(1 + \frac{|\Lambda|}{A_\Lambda (1 - \text{ERC}(\Lambda))^2}\right) \|f - f_\Lambda\|^2. \quad (12.136)$$

**Proof.** The residual at step  $m < |\Lambda|$  is denoted  $R^m f = f - \tilde{f}_m$  where  $\tilde{f}_m$  is the orthogonal projection of  $f$  on the space generated by the dictionary vectors selected by the first  $m$  iterations. The theorem is proved by induction by showing that either  $\tilde{f}_m$  satisfies (12.136) or  $\tilde{f}_m \in \mathbf{V}_\Lambda$ , which means that the first  $m$  vectors selected by the orthogonal pursuit are indexed in  $\Lambda$ . If (12.136) is satisfied for some  $m \leq M = |\Lambda|$ , since  $\|R^M f\| \leq \|R^m f\|$ , it results that  $\tilde{f}_M$  satisfies (12.136). If this is not the case, then the induction proof will show that the  $M = |\Lambda|$  selected vectors are in  $\{\phi_p\}_{p \in \Lambda}$ . Since an orthogonal pursuit selects linearly independent vectors, it implies that  $\tilde{f}_M = \tilde{f}_\Lambda$  satisfies (12.136).

The induction step is proved by supposing that  $\tilde{f}_m \in \mathbf{V}_\Lambda$ , and verifying that either  $C(R^m f, \Lambda^c) < 1$ , in which case the next selected vector is indexed in  $\Lambda$ , or that (12.136) is satisfied. We write  $\Phi_\Lambda^* a = \sum_{p \in \Lambda} a[p] \phi_p$ . Since  $f - f_\Lambda$  is orthogonal to  $\mathbf{V}_\Lambda$  and  $f = R^m f + \tilde{f}_m$  for  $p \in \Lambda$ , we have  $\langle R^m f, \phi_p \rangle = \langle f_\Lambda - \tilde{f}_m, \phi_p \rangle$ , so

$$C(R^m f, \Lambda^c) = \frac{\max_{q \in \Lambda^c} |\langle R^m f, \phi_q \rangle|}{\max_{p \in \Lambda} |\langle R^m f, \phi_p \rangle|} \leq C_1 + C(f_\Lambda - \tilde{f}_m, \Lambda^c), \quad (12.137)$$

with

$$C_1 = \frac{\max_{q \in \Lambda^c} |\langle f - f_\Lambda, \phi_q \rangle|}{\max_{p \in \Lambda} |\langle f_\Lambda - \tilde{f}_m, \phi_p \rangle|} \quad \text{and} \quad C(f_\Lambda - \tilde{f}_m, \Lambda^c) = \frac{\max_{q \in \Lambda^c} |\langle f_\Lambda - \tilde{f}_m, \phi_q \rangle|}{\max_{p \in \Lambda} |\langle f_\Lambda - \tilde{f}_m, \phi_p \rangle|}.$$

Since  $f_\Lambda - \tilde{f}_m \in \mathbf{V}_\Lambda$ , Theorem 12.11 proves that  $C(f_\Lambda - \tilde{f}_m, \Lambda^c) \leq \text{ERC}(\Lambda)$ . Since  $\Phi_\Lambda(f_\Lambda - \tilde{f}_m)[p] = \langle f_\Lambda - \tilde{f}_m, \phi_p \rangle$  and  $\|\Phi_\Lambda h\|^2 \geq A_\Lambda \|h\|^2$ , we get

$$\max_{p \in \Lambda} |\langle f_\Lambda - \tilde{f}_m, \phi_p \rangle| \geq \frac{1}{\sqrt{|\Lambda|}} \|\Phi_\Lambda(f_\Lambda - \tilde{f}_m)\| \geq \frac{\sqrt{A_\Lambda}}{\sqrt{|\Lambda|}} \|f_\Lambda - \tilde{f}_m\|. \quad (12.138)$$

Since  $\max_{q \in \Lambda^c} |\langle f - f_\Lambda, \phi_q \rangle| \leq \|f - f_\Lambda\|$ , inserting these inequalities in (12.137) gives

$$C(R^m f, \Lambda^c) \leq \frac{\sqrt{|\Lambda|}}{\sqrt{A_\Lambda}} \frac{\|f - f_\Lambda\|}{\|f_\Lambda - \tilde{f}_m\|} + \text{ERC}(\Lambda).$$

If  $C(R^m f, \Lambda^c) \geq 1$  then

$$\|f_\Lambda - \tilde{f}_m\|^2 \leq \frac{|\Lambda|}{A_\Lambda (1 - \text{ERC}(\Lambda))^2} \|f - f_\Lambda\|^2.$$

Since  $f - f_\Lambda$  is orthogonal to  $f_\Lambda - \tilde{f}_m$  this condition is equivalent to

$$\|f - \tilde{f}_m\|^2 = \|f - f_\Lambda\|^2 + \|f_\Lambda - \tilde{f}_m\|^2 \leq \left(1 + \frac{|\Lambda|}{A_\Lambda(1 - \text{ERC}(\Lambda))^2}\right) \|f - f_\Lambda\|^2,$$

which proves that (12.136) is satisfied, and thus finishes the induction proof. ■

The proof shows that an orthogonal matching pursuit selects the few first vectors in  $\Lambda$ . These are the “coherent signal structures” having a strong correlation with the dictionary vectors, observed in the numerical experiments in Section 12.3. At some point, the remaining vectors in  $\Lambda$  may not be sufficiently well correlated with  $f$  relative to other dictionary vectors; Theorem 12.13 computes the approximation error at this stage. This result is thus conservative since it does not take into account the approximation improvement obtained by the other vectors. Gribonval and Vandergheynst proved [283] that a nonorthogonal matching pursuit satisfies a similar theorem, but with more than  $M = |\Lambda|$  iterations. Orthogonal and nonorthogonal matching pursuits select the same “coherent structures.” Theorem 12.14 derives an approximation result that depends on only the number of  $M$ -term approximations and on the dictionary mutual coherence.

**Theorem 12.14.** Let  $f_M$  be the best  $M$ -term approximation of  $f$  from  $\mathcal{D}$  dictionary vectors. If  $M \leq \frac{1}{3\mu(\mathcal{D})}$ , then an orthogonal matching pursuit approximation  $\tilde{f}_M$  with  $M$  iterations satisfies

$$\|f - \tilde{f}_M\|^2 \leq (1 + 6M) \|f - f_M\|^2. \quad (12.139)$$

**Proof.** Let  $\Lambda$  be the approximation support of the best  $M$ -term dictionary approximation  $f_M$  with  $|\Lambda| = M$ . If  $|\Lambda| \leq \frac{1}{3\mu(\mathcal{D})}$ , then (12.129) proves that  $(1 - \text{ERC}(\Lambda))^{-1} < 2$ . Theorem 12.10 shows that if  $M \leq \frac{1}{3\mu(\mathcal{D})}$ , then  $A_\Lambda \geq 2/3$ . Theorem 12.13 derives in (12.136) that  $\tilde{f}_M$  satisfies (12.139). ■

In the dictionary (12.121) of Fourier and Dirac vectors where  $\mu(\mathcal{D}) = N^{-1/2}$ , this theorem proves that  $M$  orthogonal matching pursuit iterations are nearly optimal if  $M \leq N^{1/2}/3$ . This result is attractive because it is simple, but in practice the condition  $M \leq \frac{1}{3\mu(\mathcal{D})}$  is very restrictive because the dictionary coherence is often not so small. As previously explained, the mutual coherence of a Gabor dictionary is typically above  $1/2$ , and this theorem thus does not apply.

### 12.5.3 Support Recovery with $\ell^1$ Pursuits

An  $\ell^1$  Lagrangian pursuit has properties similar to an orthogonal matching pursuit, with some improvements that are explained. If the best  $M$ -term approximation  $f_\Lambda$  of  $f$  satisfies  $\text{ERC}(\Lambda) < 1$ , then an  $\ell^1$  Lagrangian pursuit also computes a signal approximation with an error comparable to the minimum  $M$ -term error.

An  $\ell^1$  Lagrangian pursuit (12.89) computes a sparse approximation  $\tilde{f} = \Phi^* \tilde{\alpha}$  of  $f$ , which satisfies

$$\tilde{\alpha} = \underset{\alpha \in \mathbb{C}^P}{\operatorname{argmin}} \frac{1}{2} \|f - \Phi^* \alpha\|^2 + T \|\alpha\|_1. \quad (12.140)$$

Let  $\tilde{\Lambda}$  be the support of  $\tilde{a}$ . Theorem 12.8 proves that  $\tilde{a}$  is the solution of (12.140) if and only if there exists  $h \in \mathbb{R}^P$  such that

$$\Phi(\Phi^*\tilde{a} - f) + Th = 0 \quad \text{where} \quad \begin{cases} h[p] = \text{sign}(\tilde{a}[p]) & \text{if } p \in \tilde{\Lambda} \\ |h[p]| \leq 1 & \text{if } p \notin \tilde{\Lambda}. \end{cases} \quad (12.141)$$

Theorem 12.15, proved by Tropp [463] and Fuchs [266], shows that the resulting approximation satisfies nearly the same error upper bound as an orthogonal matching pursuit in Theorem 12.13.

**Theorem 12.15:** *Fuchs, Tropp.* Let  $A_\Lambda > 0$  be the lower Riesz bound of  $\{\phi_p\}_{p \in \Lambda}$ . If  $\text{ERC}(\Lambda) < 1$  and

$$T = \lambda \frac{\|f - f_\Lambda\|}{1 - \text{ERC}(\Lambda)} \quad \text{with } \lambda > 1, \quad (12.142)$$

then there exists a unique solution  $\tilde{a}$  with support that satisfies  $\tilde{\Lambda} \subset \Lambda$ , and  $\tilde{f} = \Phi^*\tilde{a}$  satisfies

$$\|f - \tilde{f}\|^2 \leq \left(1 + \frac{\lambda^2 |\Lambda|}{A_\Lambda (1 - \text{ERC}(\Lambda))^2}\right) \|f - f_\Lambda\|^2. \quad (12.143)$$

**Proof.** The proof begins by computing a solution  $\tilde{a}$  with support that is in  $\Lambda$ , and then proves that it is unique. We denote by  $a_\Lambda$  a vector defined over the index set  $\Lambda$ . To compute a solution with a support in  $\Lambda$ , we consider a solution  $\tilde{a}_\Lambda$  of the following problem:

$$\tilde{a}_\Lambda = \underset{a_\Lambda \in \mathbb{C}^{|\Lambda|}}{\operatorname{argmin}} \frac{1}{2} \|f - \Phi_\Lambda^* a_\Lambda\|^2 + T \|a_\Lambda\|_1. \quad (12.144)$$

Let  $\tilde{a}$  be defined by  $\tilde{a}[p] = \tilde{a}_\Lambda[p]$  for  $p \in \Lambda$  and  $\tilde{a}[p] = 0$  for  $p \in \Lambda^c$ . It has a support  $\tilde{\Lambda} \subset \Lambda$ . We prove that  $\tilde{a}$  is also the solution of the  $\mathbf{I}^1$  Lagrangian minimization (12.140) if (12.142) is satisfied.

Let  $h$  be defined by

$$Th = \Phi(f - \Phi^*\tilde{a}_\Lambda) = \Phi(f - \Phi_\Lambda^*\tilde{a}_\Lambda). \quad (12.145)$$

The optimality condition (12.141) applied to the minimization (12.144) implies that

$$\forall p \in \Lambda, \quad h[p] = \text{sign}(\tilde{a}_\Lambda[p]).$$

To prove that  $\tilde{a}_\Lambda$  is the solution of (12.140), we must verify that  $|h[q]| \leq 1$  for  $q \in \Lambda^c$ . Equation (12.145) shows that the coefficients  $h_\Lambda$  of  $h$  inside  $\Lambda$  satisfy

$$Th_\Lambda = \Phi_\Lambda(f - \Phi_\Lambda^*\tilde{a}_\Lambda).$$

Since  $A_\Lambda > 0$ , the vectors indexed by  $\Lambda$  are linearly independent and

$$\tilde{a}_\Lambda = \Phi_\Lambda^+ f - T(\Phi_\Lambda \Phi_\Lambda^*)^{-1} h_\Lambda = a_\Lambda - T(\Phi_\Lambda \Phi_\Lambda^*)^{-1} h_\Lambda, \quad (12.146)$$

and (12.145) implies that

$$h_{\Lambda^c} = \frac{1}{T} \Phi_{\Lambda^c} (f - f_\Lambda + T \Phi_\Lambda^* (\Phi_\Lambda \Phi_\Lambda^*)^{-1} h_\Lambda). \quad (12.147)$$

The expression (12.147) of  $h_{\Lambda^c}$  shows that

$$\|h_{\Lambda^c}\|_\infty = T^{-1} \max_{q \in \Lambda^c} |\langle \phi_q, f - f_\Lambda + T\Phi_\Lambda^*(\Phi_\Lambda\Phi_\Lambda^*)^{-1}h_\Lambda \rangle| \quad (12.148)$$

$$\leq T^{-1} \max_{q \in \Lambda^c} |\langle \phi_q, f - f_\Lambda \rangle| + \max_{q \in \Lambda^c} |\langle (\Phi_\Lambda\Phi_\Lambda^*)^{-1}\Phi_\Lambda\phi_q, h_\Lambda \rangle|. \quad (12.149)$$

We saw in (12.130) that  $(\Phi_\Lambda\Phi_\Lambda^*)^{-1}\Phi_\Lambda = \Phi_\Lambda(\Phi_\Lambda^*\Phi_\Lambda)^{-1} = (\Phi_\Lambda^+)^*$ . Since  $\text{ERC}(\Lambda) = \max_{q \in \Lambda^c} \|(\Phi^+)^*\phi_q\|_1$  and  $\|h_\Lambda\|_\infty \leq 1$ , we get

$$\|h_{\Lambda^c}\|_\infty \leq T^{-1} \max_{q \in \Lambda^c} |\langle \phi_q, f - f_\Lambda \rangle| + \text{ERC}(\Lambda) \leq T^{-1} \|f - f_\Lambda\| + \text{ERC}(\Lambda). \quad (12.150)$$

But  $T > \|f - f_\Lambda\|(1 - \text{ERC}(\Lambda))^{-1}$ , so  $\|h_{\Lambda^c}\|_\infty < 1$ , which proves that  $\tilde{a}$  is indeed a solution with  $\tilde{\Lambda} \subset \Lambda$ .

To prove that  $\tilde{a}$  is the unique solution of (12.140), suppose that  $\tilde{a}_1$  is another solution. Then, necessarily,  $\Phi^*\tilde{a} = \Phi^*\tilde{a}_1$  because otherwise the coefficients  $(\tilde{a} + \tilde{a}_1)/2$  would have a strictly smaller Lagrangian. This proves that

$$\forall p \notin \Lambda, \quad |\langle f - \Phi^*\tilde{a}, \phi_p \rangle| = |\langle f - \Phi^*\tilde{a}_1, \phi_p \rangle| < T,$$

and thus that  $\tilde{a}_1$  is also supported inside  $\Lambda$ . Since  $\Phi_\Lambda^*\tilde{a}_\Lambda = \Phi_\Lambda^*\tilde{a}_{1,\Lambda}$  and  $\Phi_\Lambda^*$  is invertible,  $\tilde{a} = \tilde{a}_1$ , which proves that the solution of (12.140) is unique.

Let us now prove the approximation result (12.143). The optimality conditions (12.141) prove that

$$\|\Phi_\Lambda(\Phi^*\tilde{a} - f)\|_\infty = \max_{p \in \Lambda} |\langle \phi_p, \tilde{f} - f \rangle| \leq T.$$

For any  $p \in \Lambda$   $\langle \phi_p, f \rangle = \langle \phi_p, f_\Lambda \rangle$ , so

$$\max_{p \in \Lambda} |\langle \phi_p, \tilde{f} - f_\Lambda \rangle| \leq T.$$

Since  $\tilde{\Lambda} \subset \Lambda$ , it results that  $\tilde{f} - f_\Lambda \in \mathbf{V}_\Lambda$ , and since  $\|\Phi_\Lambda h\|^2 \geq A_\Lambda \|h\|^2$ , we get

$$T \geq \max_{p \in \Lambda} |\langle f_\Lambda - \tilde{f}, \phi_p \rangle| \geq \frac{1}{\sqrt{|\Lambda|}} \|\Phi_\Lambda(f_\Lambda - \tilde{f})\| \geq \frac{\sqrt{A_\Lambda}}{\sqrt{|\Lambda|}} \|f_\Lambda - \tilde{f}\|. \quad (12.151)$$

Since  $T = \lambda \|f - f_\Lambda\|(1 - \text{ERC}(\Lambda))^{-1}$ , it results that

$$\|f_\Lambda - \tilde{f}\| \leq \frac{\lambda \sqrt{|\Lambda|}}{\sqrt{A_\Lambda}(1 - \text{ERC}(\Lambda))} \|f - f_\Lambda\|. \quad (12.152)$$

Moreover,  $f - f_\Lambda$  is orthogonal to  $f_\Lambda - \tilde{f} \in \mathbf{V}_\Lambda$ , so

$$\|f - \tilde{f}\|^2 = \|f - f_\Lambda\|^2 + \|f_\Lambda - \tilde{f}\|^2.$$

Inserting (12.152) proves (12.143). ■

This theorem proves that if  $T$  is sufficiently large, then the  $\ell^1$  Lagrangian pursuit selects only vectors in  $\Lambda$ . At one point it stops selecting vectors in  $\Lambda$  but the error has already reached the upper bound (12.143). If  $f = f_\Lambda$  and  $\text{ERC}(\Lambda) < 1$ , then

(12.143) proves that a basis pursuit recovers all the atoms in  $\Lambda$  and thus reconstructs  $f$ . It is, therefore, an Exact Recovery Criterion for a basis pursuit.

An exact recovery is obtained by letting  $T$  go to zero in a Lagrangian pursuit, which is equivalent to solve a basis pursuit

$$\tilde{a} = \underset{a \in \mathbb{C}^P}{\operatorname{argmin}} \|a\|_1 \quad \text{subject to} \quad \Phi^* a = f. \quad (12.153)$$

Suppose that  $f_M = f_\Lambda$  is the best  $M$ -term approximation of  $f$  from  $M \leq \frac{1}{3\mu(\mathcal{D})}$  dictionary vectors. Similar to Theorem 12.14, Theorem 12.15 implies that the  $\mathbf{l}^1$  pursuit approximation computed with  $T = \|f - f_M\|/2$  satisfies

$$\|f - \tilde{f}\|^2 \leq (1 + 6M) \|f - f_M\|^2. \quad (12.154)$$

Indeed, as in the proof of Theorem 12.14, we verify that  $(\operatorname{ERC}(\Lambda) - 1)^{-1} < 2$  and  $A_\Lambda \geq 2/3$ .

Although  $\operatorname{ERC}(\Lambda) > 1$ , the support  $\Lambda$  of  $f \in \mathbf{V}_\Lambda$  may still be recovered by a basis pursuit. Figure 12.20 gives an example with two Gabor atoms with Heisenberg boxes, which overlap and that are recovered by a basis pursuit despite the fact that  $\operatorname{ERC}(\Lambda) > 1$ . For an  $\mathbf{l}^1$  Lagrangian pursuit, the condition  $\operatorname{ERC}(\Lambda) < 1$  can be refined with a more precise sufficient criterion introduced by Fuchs [266]. It depends on the sign of the coefficients  $a[p]$  supported in  $\Lambda$ , which recover  $f = \Phi^* a$ . In (12.150) as well as in all subsequent derivations and thus in the statement of Theorem 12.15,  $\operatorname{ERC}(\Lambda) = \max_{q \in \Lambda^c} \sum_{p \in \Lambda} |\langle \tilde{\phi}_{p,\Lambda}, \phi_q \rangle|$  can be replaced by

$$F(f, \Lambda) = \max_{q \in \Lambda^c} \sum_{p \in \Lambda} \langle \tilde{\phi}_{p,\Lambda}, \phi_q \rangle \operatorname{sign}(a[p]). \quad (12.155)$$

In particular, if  $F(f, \Lambda) < 1$ , then the support  $\Lambda$  of  $f$  is recovered by a basis pursuit. Moreover, Dossal [231] showed that if there exists  $\tilde{\Lambda}$  with  $\Lambda \subset \tilde{\Lambda}$  such that  $F(f, \tilde{\Lambda}) < 1$ , with arbitrary signs for  $a[p]$  when  $p \in \tilde{\Lambda} - \Lambda$ , then  $f$  is also recovered by a basis pursuit. By using this criteria, one can verify that if  $|\Lambda|$  is small, even though  $\Lambda$  may include very correlated vectors such as Gabor atoms of close time and frequency, we are more likely to recover  $\Lambda$  with a basis pursuit than with an orthogonal matching pursuit, but this recovery can be unstable.

### Image Source Separation

The ability to recover sparse approximation supports in redundant dictionaries has applications to source separation with a single measurement, as proposed by Elad et al. [244]. Let  $f = f_0 + f_1$  be a mixture of two signals  $f_0$  and  $f_1$  that have a sparse representation over different dictionaries  $\mathcal{D}_0 = \{\phi_p\}_{p \in \Gamma_0}$  and  $\mathcal{D}_1 = \{\phi_p\}_{p \in \Gamma_1}$ . If a sparse representation  $\tilde{a}$  of  $f$  in  $\mathcal{D} = \mathcal{D}_0 \cup \mathcal{D}_1$  nearly recovers the approximation support of  $f_0$  in  $\mathcal{D}_0$  and of  $f_1$  in  $\mathcal{D}_1$ , then both signals are separately approximated with

$$\tilde{f}_0 = \sum_{p \in \Gamma_0} \tilde{a}[p] \phi_p \quad \text{and} \quad \tilde{f}_1 = \sum_{p \in \Gamma_1} \tilde{a}[p] \phi_p. \quad (12.156)$$

An application is given to separate edges from oscillatory textures in images.



FIGURE 12.25

Image separation  $f = f_0 + f_1$  in a dictionary that is a union of a wavelet and a local cosine dictionary: (a) image  $f$ , (b) piecewise regular component  $f_0$ , and (c) oscillatory texture  $f_1$ .

To take into account the differences between edges and textures, Meyer [47] introduced an image model  $f = f_0 + f_1$ , where  $f_0$  is a bounded variation function including edges, and  $f_1$  is an oscillatory texture function that belongs to a different functional space. Theorem 9.17 proves that bounded variation images  $f_0$  are sparse in a translation-invariant dictionary  $\mathcal{D}_0$  of wavelets. Dictionaries of curvelets in Section 5.5.2 or bandlets in Section 12.2.4 can also improve the approximations of geometrically regular edges in  $f_0$ .

The oscillatory image  $f_1$  has well-defined local frequencies and is therefore sparse in a dictionary  $\mathcal{D}_1$  of two-dimensional local cosine bases, defined in Section 8.5.3. A dictionary  $\mathcal{D} = \mathcal{D}_0 \cup \mathcal{D}_1$  is defined as a union of a wavelet dictionary and a local cosine dictionary [244]. A sparse representation  $\tilde{\alpha}$  of  $f$  in  $\mathcal{D}$  is computed with an  $\ell^1$  basis pursuit, and approximations  $\tilde{f}_0$  and  $\tilde{f}_1$  of  $f_0$  and  $f_1$  are computed with (12.156). Figure 12.25 shows that this algorithm can indeed separate oscillating textures from piecewise regular image variations in such a dictionary.

## 12.6 MULTICHANNEL SIGNALS

Multiple channel measurements often have strong dependencies that a representation should take into account. For color images, the green, blue, and red (RGB) channels are highly correlated. Indeed, edges and sharp variations typically occur at the same location in each color channel. Stereo audio recordings or multiple point recordings of EEGs also output dependent measurement vectors. Taking into account the structural dependancies of these channels improves compression or denoising applications, but also provides solutions to the source separation problems studied in Section 13.5.

A signal with  $K$  channels is considered as a signal vector  $\vec{f}[n] = (f_k[n])_{0 \leq k < K}$ . The Euclidean norm of a vector  $\vec{a} = (a_k)_{0 \leq k < K} \in \mathbb{C}^K$  is written as  $\|\vec{a}\|^2 = \sum_{k=0}^{K-1} |a_k|^2$ .

The Frobenius norm of a signal vector is

$$\|\vec{f}\|_F^2 = \sum_{k=0}^{K-1} \|f_k\|^2.$$

### **Whitening with Linear Channel Decorrelation**

A linear decorrelation and renormalization of the signal channels can be implemented with an operator  $O$ , which often improves further multichannel processing. The empirical covariance of the  $K$  channels is computed from  $L$  signal vector examples  $\vec{x}_l = (x_{l,k})_{0 \leq k < K}$ :

$$\mu_k = \frac{1}{LN} \sum_{l=0}^{L-1} \sum_{n=0}^{N-1} x_{l,k}[n]$$

and

$$c_{k,k'} = \frac{1}{LN} \sum_{l=0}^{L-1} \sum_{n=0}^{N-1} (x_{l,k}[n] - \mu_k) (x_{l,k'}[n] - \mu_{k'}).$$

Let  $C = (c_{k,k'})_{0 \leq k, k' \leq K}$  be the empirical covariance matrix. The whitening operator  $O = C^{-1/2}$  performs a decorrelation and a renormalization of all channels.

For color images, the change of coordinates from  $(R, G, B)$  to  $(Y, U, V)$  typically implements such a decorrelation. In noise-removal applications, the noise can be decorrelated across channels by computing  $C$  from recordings  $\vec{x}_l$  of the noise.

#### **12.6.1 Approximation and Denoising by Thresholding in Bases**

We consider multichannel signals over which a whitening operator may already have been applied. Approximation and denoising operators are defined by simultaneously thresholding all the channel coefficients in a dictionary. Let  $\mathcal{D} = \{\phi_p\}_{p \in \Gamma}$  be a basis of unit vectors. For any  $\vec{f} = (f_k)_{0 \leq k < K}$ , we write an inner product vector:

$$\langle \vec{f}, \phi_p \rangle = (\langle f_k, \phi_p \rangle)_{0 \leq k < K} \in \mathbb{C}^K.$$

If  $\mathcal{D}$  is an orthonormal basis, then one can verify (Exercise 9.4) that a best  $M$ -term approximation  $\vec{f}_M$  that minimizes  $\|\vec{f} - \vec{f}_M\|_F^2$  is obtained by selecting the  $M$  inner product vectors having the largest norm  $\|\langle \vec{f}, \phi_p \rangle\|$ . Such nonlinear approximations are thus calculated by thresholding the norm of these multichannel inner product vectors:

$$\vec{f}_{\Lambda_T} = \sum_{p \in \Lambda_T} \langle \vec{f}, \phi_p \rangle \phi_p \quad \text{with} \quad \Lambda_T = \{p \in \Gamma : \|\langle \vec{f}, \phi_p \rangle\| \geq T\}. \quad (12.157)$$

Let  $\vec{W}$  be a random noise vector. We suppose that a whitening operator has been applied so that  $\vec{W}[n] = (W_k[n])_{0 \leq k < K}$  is decorrelated across channels, and that each  $W_k[n]$  is a Gaussian white noise. A multichannel estimation of  $\vec{f}$  from

noisy measurements  $\vec{X} = \vec{f} + \vec{W}$  is implemented with a block thresholding, as defined in Section 11.4.1. A hard-block thresholding estimation is an orthogonal projector

$$\tilde{F} = \sum_{p \in \tilde{\Lambda}_T} \langle \vec{X}, \phi_p \rangle \phi_p \quad \text{with} \quad \tilde{\Lambda}_T = \{p \in \Gamma : \|\langle \vec{X}, \phi_p \rangle\| \geq T\}.$$

A James-Stein soft-block thresholding attenuates the amplitude of each inner product vector:

$$\tilde{F} = \sum_{p \in \Gamma} \max \left( 1 - \frac{T^2}{\|\langle \vec{X}, \phi_p \rangle\|^2}, 0 \right) \langle \vec{X}, \phi_p \rangle \phi_p.$$

The risk properties of such block thresholding estimators are studied in Section 11.4.1. Vector thresholding of color images improves the SNR and better preserves colors by attenuating all color channels with the same factors.

### 12.6.2 Multichannel Pursuits

Multichannel signals decomposition in redundant dictionaries are implemented with pursuit algorithms that simultaneously approximate all the channels. Several studies describe the properties of multichannel pursuits and their generalizations and applications [157, 282, 346, 464].

#### **Matching Pursuit**

The matching pursuit algorithm in Section 12.3.1 is extended by searching for dictionary elements that maximize the norm of the multichannel inner product vector. We set the relaxation parameter  $\alpha = 1$ . Let  $\mathcal{D} = \{\phi_p\}_{p \in \Gamma}$  be a dictionary of unit vectors. The matching pursuit algorithm is initialized with  $R^0 \vec{f} = \vec{f}$ . At each iteration  $m$ , it selects a best vector  $\phi_{p_m} \in \mathcal{D}$  such that

$$\|\langle R^m \vec{f}, \phi_{p_m} \rangle\| = \operatorname{argmax}_{p \in \Gamma} \|\langle R^m \vec{f}, \phi_p \rangle\|. \quad (12.158)$$

The orthogonal projection on this vector defines a new residue

$$R^{m+1} \vec{f} = R^m \vec{f} - \langle R^m \vec{f}, \phi_{p_m} \rangle \phi_{p_m}$$

with an energy conservation

$$\|R^{m+1} \vec{f}\|_F^2 = \|R^m \vec{f}\|_F^2 + \|\langle R^m \vec{f}, \phi_{p_m} \rangle\|^2.$$

Theorem 12.6 thus remains valid by using the Frobenius norm over signal vectors, which proves the exponential convergence of the algorithm:

$$\vec{f} = \sum_{m=0}^{+\infty} \langle \vec{f}, \phi_{p_m} \rangle \phi_{p_m}$$

After  $M$  matching pursuit iterations, a backprojection algorithm recovers the orthogonal projection of  $\vec{f}$  on the selected atoms  $\{\phi_{p_m}\}_{0 \leq m < M}$ . It can be computed with a Richardson gradient descent described afterwards for backprojecting an  $\mathbf{l}^1$  pursuit, which is initialized with  $\vec{a}_0[p_m] = \langle R^m \vec{f}, \phi_{p_m} \rangle$ .

The orthogonal matching pursuit of Section 12.3.2 is similarly extended. At a step  $m$ , a vector  $\phi_{p_m}$  is selected as in (12.158). A Gram-Schmidt orthogonalization decomposes  $\phi_{p_m}$  into its projection over the previously selected vectors  $\{\phi_{p_k}\}_{0 \leq k < m}$  plus an orthogonal complement  $u_m$ . The orthogonalized residue is then

$$R^{m+1} \vec{f} = R^m \vec{f} - \frac{\langle R^m \vec{f}, u_m \rangle}{\|u_m\|^2} u_m.$$

It decomposes  $\vec{f}$  over an orthogonal family  $\{u_m\}_m$  and for signals of size  $N$ :

$$\vec{f} = \sum_{m=0}^{N-1} \frac{\langle R^m \vec{f}, u_m \rangle}{\|u_m\|^2} u_m.$$

The orthogonal matching pursuit properties thus remain essentially the same.

### **Multichannel $\mathbf{l}^1$ Pursuits**

Sparse multichannel signal representations in redundant dictionaries can also be computed with  $\mathbf{l}^1$  pursuits, which minimize an  $\mathbf{l}^1$  norm of the coefficients. An  $\mathbf{l}^1$  norm over multichannel vectors of coefficients is defined by

$$\|\vec{a}\|_1 = \sum_{p \in \Gamma} \|\vec{a}[p]\|.$$

We denote

$$\Phi^* \vec{a} = \sum_{p \in \Gamma} \vec{a}[p] \phi_p.$$

A sparse  $\mathbf{l}^1$  pursuit approximation of a vector  $\vec{f}$  at a precision  $\varepsilon$  is defined by  $\tilde{\vec{f}} = \Phi^* \tilde{\vec{a}}[p]$  with

$$\tilde{\vec{a}} = \underset{\vec{a} \in \mathbb{C}^{PK}}{\operatorname{argmin}} \|\vec{a}\|_1 \quad \text{subject to} \quad \|\vec{f} - \Phi^* \vec{a}\|_F \leq \varepsilon.$$

A solution of this convex optimization is computed with an  $\mathbf{l}^1$  Lagrangian minimization

$$\tilde{\vec{a}} = \underset{\vec{a} \in \mathbb{C}^{PK}}{\operatorname{argmin}} \frac{1}{2} \|\vec{f} - \Phi^* \vec{a}\|_F^2 + T \|\vec{a}\|_1, \quad (12.159)$$

where  $T$  depends on  $\varepsilon$ . Several authors have studied the approximation properties of  $\mathbf{l}^1$  vector pursuits and their generalizations [157, 188, 462].

The  $\mathbf{l}^1$  Lagrangian minimization (12.159) is numerically solved with the iterative thresholding algorithm of Section 12.4.3, which is adjusted as follows. We write  $\Phi\vec{f}[p] = \langle \vec{f}, \phi_p \rangle$ .

**1. Initialization.** Choose  $\vec{a}_0$ , set  $k = 0$ , and compute  $\vec{b} = \Phi\vec{f}$ .

**2. Gradient step.** Update

$$\tilde{\vec{a}}_k = \vec{a}_k + \gamma (\vec{b} - \Phi \Phi^* \vec{a}_k), \quad (12.160)$$

where  $\gamma < 2 \|\Phi\Phi^*\|_S^{-1}$ .

**3. Soft thresholding.** Compute

$$\vec{a}_{k+1}[p] = \rho_{\gamma T}(\tilde{\vec{a}}_k[p]), \quad (12.161)$$

where  $\rho_{\gamma T}(\vec{x}) = \vec{x} \max\left(1 - \frac{\gamma T}{\|\vec{x}\|}, 0\right)$ .

**4. Stop.** If  $\|\vec{a}_k - \vec{a}_{k+1}\|_F$  is smaller than a fixed-tolerance criterion, stop the iterations; otherwise, set  $k \leftarrow k + 1$  and go back to 2.

If  $\tilde{\Lambda}$  is the support of the computed solution after convergence, then like for a matching pursuit, a backprojection algorithm computes the orthogonal projection of  $\vec{f}$  on the selected atoms  $\{\phi_p\}_{p \in \tilde{\Lambda}}$ . It can be implemented with a Richardson gradient descent. It continues the gradient descent iterations of step 2, and replaces the soft thresholding in step 3 by a projector defined by  $\vec{a}_{k+1}[p] = \tilde{\vec{a}}_k[p]$  if  $p \in \tilde{\Lambda}$  and  $\vec{a}_{k+1}[p] = 0$  if  $p \notin \tilde{\Lambda}$ .

### Multichannel Dictionaries

Multichannel signals  $\vec{f}$  have been decomposed over dictionaries of scalar signals  $\{\phi_p\}_{p \in \Gamma}$ , thus implying that the same dictionary elements  $\{\phi_p\}_{p \in \Lambda}$  are appropriate to approximate all signal channels  $(f_k)_{0 \leq k < K}$ . More flexibility can be provided by dictionaries of multiple channel signals  $\mathcal{D} = \{\phi_p\}_{p \in \Gamma}$  where each  $\vec{\phi}_p = (\phi_{p,k})_{0 \leq k < K}$  includes  $K$  channels. In the context of color images, this means constructing a dictionary of color vectors having three color channels. Applying color dictionaries to color images can indeed improve the restitution of colors in noise-removal applications [357].

Inner product vectors and projectors over dictionary vectors are written as

$$\langle \vec{f}, \vec{\phi}_p \rangle = \left( \langle f_k, \phi_{p,k} \rangle \right)_{0 \leq k < K} \quad \text{and} \quad \langle \vec{f}, \vec{\phi}_p \rangle \vec{\phi}_p = \left( \langle f_k, \phi_{p,k} \rangle \phi_{p,k} \right)_{0 \leq k < K}.$$

The thresholds and pursuit algorithms of Sections 12.6.1 and 12.6.2 decompose  $\vec{f}$  in a dictionary of scalar signals  $\{\phi_p\}_{p \in \Gamma}$ . They are directly extended to decompose  $\vec{f}$  in a dictionary of signal vectors  $\mathcal{D} = \{\vec{\phi}_p\}_{p \in \Gamma}$ . In all formula and algorithms

$$\langle \vec{f}, \phi_p \rangle \quad \text{is replaced by} \quad \langle \vec{f}, \vec{\phi}_p \rangle$$

and

$$\langle \vec{f}, \phi_p \rangle \phi_p \text{ is replaced by } \langle \vec{f}, \vec{\phi}_p \rangle \vec{\phi}_p.$$

For a fixed index  $p$ , instead of decomposing all signal channels  $f_k$  on the same  $\phi_p$ , the resulting algorithms decompose each  $f_k$  on a potentially different  $\phi_{p,k}$ , but all channels make a simultaneous choice of a dictionary vector  $\vec{\phi}_p$ . Computations and mathematical properties are otherwise the same. The difficulty introduced by this flexibility is to construct dictionaries specifically adapted to each signal channel. Dictionary learning provides an approach to solve this issue.

## 12.7 LEARNING DICTIONARIES

For a given dictionary size  $P$ , the dictionary should be optimized to best approximate signals in a given set  $\Theta$ . Prior information on signals can lead to appropriate dictionary design, for example, with Gabor functions, wavelets, or local cosine vectors. These dictionaries, however, can be optimized by better taking into account the signal properties derived from examples. Olshausen and Field [391] argue that such a learning process is part of biological evolution, and could explain how the visual pathway has been optimized to extract relevant information from visual scenes. Many open questions remain on dictionary learning, but numerical algorithms show that learning is possible and can improve applications.

Let us consider a family of  $K$  signal examples  $\{f_k\}_{0 \leq k < K}$ . We want to find a dictionary  $\mathcal{D} = \{\phi_p\}_{p \in \Gamma}$  of size  $|\Gamma| = P$  in which each  $f_k$  has an “optimally” sparse approximation

$$\tilde{f}_k = \sum_{p \in \Gamma} a[k, p] \phi_p, \quad (12.162)$$

given a precision  $\|f_k - \tilde{f}_k\|^2 \leq \varepsilon$ . This sparse decomposition may be computed with a matching pursuit, an orthogonal matching pursuit, or an  $\mathbf{l}^1$  pursuit. Let  $\Phi f_k = \{\langle f_k, \phi_p \rangle\}_{p \in \Gamma}$  be the dictionary operator with rows equal to the dictionary vectors  $\phi_p$ . The learning process iteratively adjusts  $\mathcal{D}$  to optimize the sparse representation of all examples.

### **Dictionary Update**

Following the work of Olshausen and Field [391], several algorithms have been proposed to optimize  $\mathcal{D}$ , and thus  $\Phi$ , from a family of examples [78, 246, 336, 347]. It is a highly non-convex optimization problem that therefore can be trapped in local minima.

The approach of Engan, Aase, and Husoy [246] performs alternate optimizations, similar to the Lloyd-Max algorithm for learning code books in vector

quantization [27]. Let us write  $\vec{f} = \{f_k\}_{0 \leq k < K}$ . Its Frobenius norm is

$$\|\vec{f}\|_F^2 = \sum_{k=0}^{K-1} \|f_k\|^2.$$

According to (12.162), the approximation vector  $\tilde{\vec{f}} = \{\tilde{f}_k\}_{0 \leq k < K}$  can be written as  $\tilde{\vec{f}} = A\Phi$ . The algorithm alternates between the calculation of the matrix of sparse signal coefficients  $A = (a[k, p])_{0 \leq k < K, p \in \Gamma}$  and a modification of the dictionary vectors  $\phi_p$  to minimize the Frobenius norm of the residual error

$$\|\vec{f} - A\Phi\|_F^2 = \sum_{k=0}^{K-1} \|f_k - \sum_{p \in \Gamma} a[k, p] \phi_p\|^2.$$

The matrix  $\Phi$  can be considered as a vector transformed by the operator  $A$ . As explained in Section 5.1.3, the error  $\|\vec{f} - A\Phi\|_F^2$  is minimum if  $A\Phi$  is the orthogonal projection of  $\vec{f}$  in the image space of  $A$ . It results that  $\Phi$  is computed with the pseudo inverse  $A^+$  of  $A$ :

$$\Phi = A^+ \vec{f} = (A^* A)^{-1} A^* \vec{f}.$$

The inversion of the operator  $L = A^* A$  can be implemented with the conjugate-gradient algorithm in Theorem 5.8. The resulting learning algorithm proceeds as follows:

1. *Initialization.* Each vector  $\phi_p$  is initialized as a white Gaussian noise with a norm scaled to 1.
2. *Sparse approximation.* Calculation with a pursuit of the matrix  $A = (a[k, p])_{0 \leq k < K, p \in \Gamma}$  of sparse approximation coefficients

$$\|f_k - \sum_{p \in \Gamma} a[k, p] \phi_p\| \leq \varepsilon \quad \text{for } 0 \leq k < K. \quad (12.163)$$

3. *Dictionary update.* Minimization of the residual error (12.163) with

$$\Phi = A^+ \vec{f} = (A^* A)^{-1} A^* \vec{f}. \quad (12.164)$$

4. *Dictionary normalization.* Each resulting row  $\phi_p$  of  $\Phi$  is normalized:  $\|\phi_p\| = 1$ .
5. *Stopping criterion.* After a fixed number of iterations, or if  $\Phi$  is marginally modified, then stop; otherwise, go back to 2.

This algorithm is computationally very intensive since it includes a sparse approximation at each iteration and requires the inversion of the  $P \times P$  matrix  $A^* A$ . The sparse approximations are often calculated with an orthogonal pursuit that provides a good precision versus calculation trade-off. When the sparse coefficients  $a[k, p]$  are computed with an  $\ell^1$  pursuit, then one can prove that the algorithm converges to a stationary point [466].

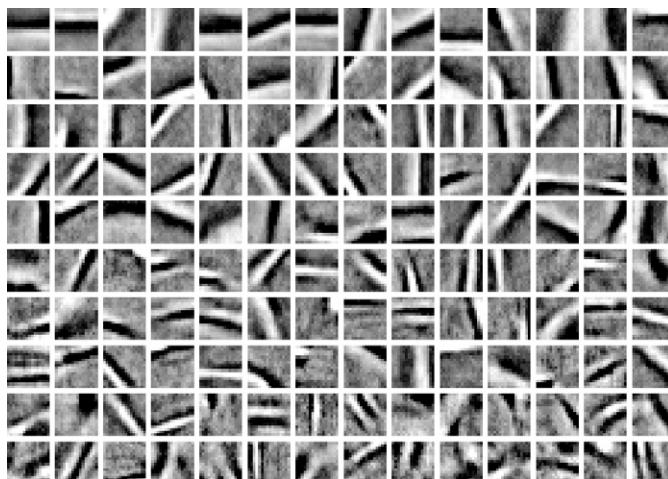
To compress structured images such as identity photographs, Bryt and Elad showed that such learning algorithms are able to construct highly efficient dictionaries [123]. It is also used for video compression with matching pursuit by optimizing a predefined dictionary, which improves the distortion rate [426].

### ***Translation-Invariant Dictionaries***

For noise removal or inverse problems, the estimation of stationary signals is improved with translation-invariant dictionaries. As a result, the algorithm must only learn the  $P$  generators of this translation-invariant dictionary. A maximum support size for these dictionary vectors is set to a relatively small value of typically  $N = 16 \times 16$  pixels. The examples  $\{f_k\}_{0 \leq k < K}$  are then chosen to be small patches of  $N$  pixels, extracted from images.

Optimized translation-invariant dictionaries lead to high-quality, state-of-the-art noise-removal results [242]. For color images, it can also incorporate intrinsic redundancy between the different color channels by learning color vectors [357]. Section 12.6 explains how to extend pursuit algorithms to multiple channel signals such as color images. Besides denoising, these dictionaries are used in super-resolution inverse problems—for example, to recover missing color pixels in high-resolution color demosaicing [357].

Figure 12.26 shows an example of a dictionary learned with this algorithm with  $N = 16^2$  and  $P = 2N$ . The calculated dictionary vectors  $\phi_p$  look similar to the directional wavelets of Section 5.5. Olshausen and Field [391] observed that these dictionary vectors learned from natural images also look similar to the impulse response of *simple cell neurons* in the visual cortical area V1. It supports the idea of a biological adaptation to process visual scenes.



**FIGURE 12.26**

Dictionary  $\{\phi_p\}_{p \in \Gamma}$  learned from examples extracted from natural images.

---

## 12.8 EXERCISES

**12.1** <sup>2</sup> Best wavelet packet and local cosine approximations:

- (a) Synthesize a discrete signal that is well approximated by a few vectors in a best wavelet packet basis, but that requires many more vectors to obtain an equivalent approximation in a best local cosine basis. Test your signal numerically.
- (b) Design a signal that is well approximated in a best local cosine basis but requires many more vectors to approximate it efficiently in a best wavelet packet basis. Verify your result numerically.

**12.2** <sup>3</sup> Describe a coding algorithm that codes the position of  $M$  nonzero orthogonal coefficients in a best wavelet packet or local cosine dictionary tree of size  $P = N \log_2 N$ , and that requires less than  $R_0 = \log_2 \binom{M}{P}$  bits. How many bits does your algorithm require?

**12.3** <sup>2</sup> A double tree of block wavelet packet bases is defined in Exercise 8.12. Describe a fast best-basis algorithm that requires  $O(N(\log_2 N)^2)$  operations to find the block wavelet packet basis that minimizes an additive cost (12.34) [299].

**12.4** <sup>2</sup> In a dictionary  $\mathcal{D} = \{\phi_p\}_{p \in \Gamma}$  of orthonormal bases, we write  $a_o[p]$  a vector of orthogonal coefficients with a support  $\Lambda_0$  corresponding to an orthonormal family  $\{\phi_p\}_{p \in \Lambda_0}$ . We want to minimize the  $\mathbf{l}^1$  Lagrangian among orthogonal coefficients:

$$\mathcal{L}_1(T, f, \tilde{a}) = \min_{a_o} \frac{1}{2} \|f - \sum_{p \in \Gamma} a_o[p] \phi_p\|^2 + T \sum_{p \in \Gamma} |a_o[p]|.$$

- (a) Verify that  $\mathcal{L}_1(T, f, a_0) = 1/2 \sum_{p \in \Lambda_0} |\langle f, \phi_p \rangle - a_o[p]|^2 + T \sum_{p \in \Lambda_0} |a_o[p]|$ .
- (b) Prove that  $e(\alpha) = (\beta - \alpha)^2 / 2 + T|\alpha|$  is minimized by  $\alpha = \rho_T(\beta)$  where  $\rho_T(x) = x \max(1 - T/|x|, 0)$  is a soft thresholding.
- (c) Prove that if the dictionary is reduced to a single orthonormal basis  $\mathcal{B} = \{\phi_p\}_{p \in \Gamma_B}$ , then the minimum  $\mathbf{l}^1$  Lagrangian is

$$\mathcal{L}_1(T, f, \mathcal{B}) = \mathcal{L}_1(T, f, \tilde{a}) = \sum_{p \in \Gamma_B} C(|\langle f, \phi_p \rangle|)$$

$$\text{with } C(x) = \frac{1}{2} \min(T^2, x^2) + T|x|.$$

- (d) In a dictionary of orthonormal bases, describe a best-basis algorithm that finds the minimizer  $\tilde{a}$  of  $\mathcal{L}_1(T, f, a_0)$  among all vectors of orthogonal coefficients selected in the dictionary.

**12.5** <sup>2</sup> Prove that if we add a vector  $h$  to an orthonormal basis  $\mathcal{B}$ , we obtain a redundant dictionary with a mutual coherence that satisfies  $\mu(\mathcal{B} \cup \{h\}) \geq N^{-1/2}$ .

**12.6** <sup>2</sup> Let  $\mathcal{D}_{j,\Delta}$  be a Gabor dictionary as defined in (12.77). Let  $\Lambda$  be an index set of two Gabor atoms in  $\mathcal{D}_{j,\Delta}$ .

- (a) Compute  $\text{ERC}(\Lambda)$  as a function of the distance in time and frequency of both Gabor atoms with (12.81).
- (b) Compute the first upper bound in (12.12) of  $\text{ERC}(\Lambda)$  and compare its value with  $\text{ERC}(\Lambda)$ .
- (c) Compute the minimum time distance (for atoms at the same frequency) and the minimum frequency distance (for atoms at the same time) so that  $\text{ERC}(\Lambda) < 1$ .

**12.7** <sup>2</sup> Let  $\mathcal{D} = \{\delta[n - k], e^{i2\pi kn/N}\}_{0 \leq k < N}$  be a Dirac-Fourier dictionary.

- (a) Prove that a matching pursuit residue calculated with a relaxation factor  $\alpha = 1$  satisfies  $\|R^m f\| \leq \|f\| \exp(-m/(2N))$ .
- (b) Prove that if  $f$  is a combination of  $M \leq N^{1/2}/2$  Fourier and Dirac vectors, then the matching pursuit reconstructs  $f$  exactly as a combination of these  $M$  Fourier and Dirac vectors.

**12.8** <sup>3</sup> *Uncertainty principle* [227]. We consider two orthogonal bases  $\mathcal{B}_0, \mathcal{B}_1$  of  $\mathbb{R}^N$ .

- (a) Prove that for any  $f \neq 0$ ,

$$\|f_{\mathcal{B}_0}\|_0 \|f_{\mathcal{B}_1}\|_0 \geq \frac{1}{\mu(\mathcal{B}_0 \cup \mathcal{B}_1)^2}. \quad (12.165)$$

*Hint:* Show that  $\|f_{\mathcal{B}_0}\|_\infty \leq \mu(\mathcal{B}_0 \cup \mathcal{B}_1) \|f_{\mathcal{B}_1}\|_1$  and use the Cauchy-Schwartz inequality.

- (b) Prove that for any  $f \neq 0$ ,

$$\|f_{\mathcal{B}_0}\|_0 + \|f_{\mathcal{B}_1}\|_0 \geq \frac{2}{\mu(\mathcal{B}_0 \cup \mathcal{B}_1)}. \quad (12.166)$$

**12.9** <sup>3</sup> *Cumulated coherence.* Let  $\mathcal{D} = \{\phi_p\}_{p \in \Gamma}$  be a dictionary with  $\|\phi_p\| = 1$ . The cumulated mutual coherence of  $\mathcal{D}$  is

$$\mu_M(\mathcal{D}) = \max_{|\Lambda| \leq M} \max_{q \in \Lambda^c} \sum_{p \in \Lambda} |\langle \phi_p, \phi_q \rangle|.$$

- (a) Prove that  $\mu_M(\mathcal{D}) \leq M \mu(\mathcal{D})$ .
- (b) Prove that  $\mu_{M-1}(\mathcal{D}) + \mu_M(\mathcal{D}) < 1$  implies that  $\text{ERC}(\Lambda) < 1$  for any  $\Lambda$  such that  $|\Lambda| \leq M$ . *Hint:* Use Theorem 12.12.
- (c) For  $\beta < 1$  and for any  $p \geq 0$ , let

$$\forall n \in \mathbb{Z}, \quad \phi_p[n] = \begin{cases} 0 & \text{if } n < p \\ \beta^{n-p} \sqrt{1-\beta^2} & \text{if } n \geq p \end{cases}$$

Show that  $\mathcal{D}$  spans  $\ell^2(\mathbb{Z})$  and that  $|\langle \phi_p, \phi_{p'} \rangle| = \beta^{|p-p'|}$ . *Hint:* Consider the case  $p > p'$

- (d) Show that  $\mu_M(\mathcal{D}) < 2\beta/(1-\beta)$  while  $\mu(\mathcal{D})M$  grows unbounded with  $M$ .

**12.10** <sup>2</sup> *Spark.* Let  $\mathcal{D} = \{\phi_p\}_{p \in \Gamma}$  be a dictionary with  $\|\phi_p\| = 1$ . The spark of  $\mathcal{D}$ , introduced in [229], is

$$\text{spark}(\mathcal{D}) = \min_{b \in \mathbf{Null}(\Phi^*), b \neq 0} \|b\|_0.$$

- (a) Show that if  $\|\alpha\|_0 < \text{spark}(\mathcal{D})/2$ , then  $\alpha$  is the unique solution of

$$\min_b \|b\|_0 \quad \text{subject to} \quad \Phi^* b = \Phi^* \alpha. \quad (12.167)$$

*Hint:* If  $b$  is a solution of this problem,  $\alpha - b \in \mathbf{Null}(\Phi^*)$ .

- (b) Show that  $\text{spark}(\mathcal{D}) \geq 1 + 1/\mu(\mathcal{D})$ . Next, deduce that if  $\|\alpha\|_0 < (1 + 1/\mu(\mathcal{D}))/2$ , then  $\alpha$  is the unique solution of (12.167). *Hint:* Use Theorem 12.10.

## Inverse Problems

## 13

Recovering high-resolution and high-quality signals from partial and noisy measurements is the dream behind inverse problems. It is present in most signal processing, from medical imaging to analog and digital conversions, from seismic exploration to high-definition video display. Measurements are modeled with a linear operator applied to the input signal, but this operator is typically not invertible. Computing a precise signal estimation is thus not possible without strong a priori information on the signal.

The input data belong to a space of limited dimension that defines the measurement resolution. Estimating the signal at this resolution is already challenging because some signal components are attenuated and can thus barely be discriminated from the noise. Partially inverting the operator can considerably amplify the noise and do more harm than good.

Linear estimators implement a partial and regularized linear inversion that is related to a singular value decomposition. Nonlinear estimations improve these estimations by capturing more prior information on the signal. Sparse estimation algorithms incorporate this prior information in the design of a dictionary in which the estimated signal has a sparse representation. The estimation procedure and the resulting estimation risk depend on the dictionary property.

If there exists a basis providing a sparse signal representation and with vectors that nearly diagonalize the measurement operator, then thresholding estimators can have a nearly minimax risk. When diagonal thresholding estimators fail, super-resolution may have its chance. Super-resolution is more ambitious and computes a signal estimation at a resolution that is higher than the data resolution. Pursuit decompositions can compute sparse super-resolution estimations in redundant and incoherent dictionaries, but super-resolution is not always possible.

Compressive sensing gives a new perspective on inverse problems by stabilizing super-resolution with random measurements. Randomness is a powerful tool to build incoherent dictionaries. Compressive sensing suggests designing new signal-acquisition devices that recover high-resolution signals from lower-resolution randomized measurements.

Ending the book with a cocktail party leads us to blind source separation. Recovering simultaneously several conversations or signals from few mixed

measurements is another super-resolution problem where sparsity again plays a central role.

## 13.1 LINEAR INVERSE ESTIMATION

Let us consider measurements obtained with a linear operator applied to an incoming analog signal  $\bar{f}(x)$  to which noise is added:

$$Y[q] = \bar{U}\bar{f}[q] + W[q], \quad (13.1)$$

where  $\bar{U}\bar{f}[q] = \langle \bar{f}, \bar{u}_q \rangle$  is the measurement output of a sensor. The operator  $\bar{U}$  and the noise variance are supposed to be known or measured with a calibration procedure.

To invert the degradation numerically,  $\bar{U}$  is factorized into a stable sampling operator  $\bar{\Phi}_s$  followed by a discrete operator  $U$ , which carries the degradation and may provide less than  $N$  measurements:

$$Y[q] = U\bar{\Phi}_s\bar{f}[q] + W[q]. \quad (13.2)$$

As explained in Section 3.1.3, the sampling operator  $\bar{\Phi}_s\bar{f}$  projects  $\bar{f}$  over a Riesz basis of an approximation space  $\mathbf{U}_N$ . It is partially inverted by a discrete-to-analog converter that recovers the orthogonal projection of  $\bar{f}$  over  $\mathbf{U}_N$ .

The goal is to recover the best possible estimate of the high-resolution signal  $f[n] = \bar{\Phi}_s\bar{f}[n]$  from

$$Y[q] = Uf[q] + W[q]. \quad (13.3)$$

Let  $Q$  be the dimension of the image space  $\mathbf{Im}U$  of the operator  $U$ . Linear estimators recover the projection of  $f$  in a space of dimension at most  $Q$  and thus do not provide any super-resolution. They are introduced by imposing a regularity condition on the solution through a quadratic variational problem. This will lead us to regularized singular value decompositions.

### 13.1.1 Quadratic and Tikhonov Regularizations

Suppose that  $f$  has some form of regularity, expressed by a regularization operator  $\Phi$  that yields small energy coefficients. A linear estimation  $\tilde{F}$  of  $f$  is computed from  $Y = Uf + W \in \mathbb{R}^N$  as a solution of a quadratic optimization

$$\tilde{F} = \underset{h \in \mathbb{R}^N}{\operatorname{argmin}} \|\Phi h\|^2 \quad \text{subject to} \quad \|Uh - Y\|^2 \leq \varepsilon, \quad (13.4)$$

where  $\varepsilon$  is of the order of the noise energy  $\|W\|^2$ . If  $W$  is a Gaussian white noise of variance  $\sigma^2$ , then  $E\{\|W\|^2\} = N\sigma^2$ . If the noise  $W$  is not white but has an invertible covariance  $K_W$ , then  $\|Uh - Y\|^2$  is replaced by  $\|K_W^{-1/2}(Uh - Y)\|^2$  where  $K_W^{-1/2}$  performs a “whitening” of the error  $Uh - Y$  before applying the Euclidean norm.

Minimizing  $\|\Phi \tilde{F}\|^2$  yields coefficients  $\Phi \tilde{F}[n]$  of small amplitude that are rather uniformly spread. A Tikhonov regularization corresponds to a finite different approximation of a first-order derivative or of a gradient operator  $\Phi = \vec{\nabla}$  in multiple dimensions. In this case, the solution  $\tilde{F}$  has a finite-energy derivative and is thus differentiable in the sense of Sobolev. Section 13.3 gives examples of linear Tikhonov regularizations to estimate missing image pixels.

Since (13.4) is a strictly convex minimization, its solution can be computed as a solution of a Lagrangian minimization

$$\tilde{F} = \underset{h \in \mathbb{R}^N}{\operatorname{argmin}} \frac{1}{2} \|Uh - Y\|^2 + T^2 \|\Phi h\|^2, \quad (13.5)$$

where  $T$  is adjusted as a function of  $\varepsilon$ . The theorem 13.1 computes the resulting linear estimator.

**Theorem 13.1.** The solution of the quadratic minimization

$$\tilde{F} = \underset{h \in \mathbb{R}^N}{\operatorname{argmin}} \|Y - Uh\|^2 + T^2 \|\Phi h\|^2 \quad (13.6)$$

is the linear estimator

$$\tilde{F} = (U^* U + T^2 \Phi^* \Phi)^{-1} U^* Y. \quad (13.7)$$

**Proof.** Since the Lagrangian (13.5) is quadratic relative to the signal coordinates, its minimum is obtained by setting its partial derivatives to zero. The partial derivative of  $\|Y - Uh\|^2 + T^2 \|\Phi h\|^2$  with respect to  $h[n]$  is  $U^*(Uh - Y)[n] + T^2 \Phi^* \Phi h[n]$ . Setting these derivatives to zero leads to the optimal solution (13.7). ■

The linear estimator (13.7) applies  $U^*$  to the data  $Y$ , which projects these data in  $\text{Im } U^* = (\text{Null } U)^\perp$ , which is a space of dimension  $Q$ . It results that  $\tilde{F}$  remains in a space of dimension  $Q$  and thus does not provide any super-resolution. This optimal linear estimator can be interpreted as a pseudo inversion of  $U$  followed by a linear denoising estimator.

Let  $U^+$  be the pseudo inverse, defined as the operator that inverts the restriction of  $U$  to  $\text{Im } U$  and that is equal to 0 on  $(\text{Im } U)^\perp$ . The range of  $U^+$  is  $(\text{Null } U)^\perp$  and  $U^+ U f$  is the orthogonal projection of  $f$  in  $\text{Im } U^* = (\text{Null } U)^\perp$ . Applying this pseudo inverse on the data  $Y$  gives

$$X = U^+ Y = U^+ U f + U^+ W \in (\text{Null } U)^\perp. \quad (13.8)$$

The resulting inverted noise  $Z = U^+ W$  is typically amplified. The estimator (13.7) applies a linear denoising operator  $D$  that reduces this amplified noise

$$\tilde{F} = DX \quad \text{with} \quad D = (U^* U + T^2 \Phi^* \Phi)^{-1} U^* U. \quad (13.9)$$

This linear inverse estimator is thus a linear pseudo inverse followed by a linear denoising estimator.

### 13.1.2 Singular Value Decompositions

Linear estimators that are solutions of a quadratic regularization problem may be written as a diagonal singular value decomposition. For deconvolution problems, it defines diagonal estimators in a Fourier basis. The matrix  $U^*U$  is symmetric and can thus be diagonalized in an orthonormal basis  $\mathcal{B}_S = \{e_k\}_{0 \leq k < N}$ , which is called a basis of *singular vectors*

$$U^*Ue_k = \lambda_k^2 e_k \quad \text{for } 0 \leq k < N.$$

The eigenvalues  $\lambda_k = \|Ue_k\|$  of  $U^*U$  are called *singular values* and define the *singular spectrum* of  $U$ . Let  $\{e_k\}_{k \in \Gamma_Q}$  be the set of  $Q$  basis vectors such that  $Ue_k \neq 0$ . It is an orthonormal basis of  $(\text{Null } U)^\perp$  and one can verify that  $\{Ue_k\}_{k \in \Gamma_Q}$  is an orthogonal basis of  $\text{Im } U$ .

If  $W$  is a white noise of variance  $\sigma^2$ , the covariance of the inverted noise  $Z = U^+W$  is  $K_Z = \sigma^2 U^+ U^{+*} = \sigma^2 (U^*U)^+$ . The covariance  $K_Z$  is thus also diagonalized by  $\mathcal{B}_S$ . This basis is therefore a Karhunen-Loëve basis of the inverted noise  $Z$ . The variance of the inverted noise in this basis is

$$E\{|\langle Z, e_k \rangle|^2\} = \langle K_Z e_k, e_k \rangle = \sigma^2 (U^*U)^+ e_k = \sigma^2 \lambda_k^{-2} \quad \text{for } k \in \Gamma_Q. \quad (13.10)$$

Suppose that  $\mathcal{B}_S$  also diagonalizes  $\Phi^*\Phi$ . The eigenvalues of  $\Phi^*\Phi$  are then  $\|\Phi e_k\|^2$ . Since  $\langle U^*Y, e_k \rangle = \langle Y, Ue_k \rangle$ , it results from (13.7) that the solution of the quadratic regularization (13.6) is diagonal in this basis and can be written as

$$\tilde{F} = \sum_{k \in \Gamma_Q} \frac{\langle Y, Ue_k \rangle}{\lambda_k^2 + \sigma^2 \|\Phi e_k\|^2} e_k. \quad (13.11)$$

This is called a *singular value decomposition* (SVD). The coefficients  $\|\Phi e_k\|^2$  regularize this estimation when the singular spectrum  $\lambda_k^2$  becomes too small. Appropriate choices for  $\|\Phi e_k\|^2$  lead to efficient operators for a variety of applications [111], the simplest one being  $\Phi = \text{Id}$  so that  $\|\Phi e_k\| = 1$ .

#### Oracle SVD Risk

To understand how to adjust  $\|\Phi e_k\|$  in order to minimize the risk, let us consider a particular signal  $f$ . An *oracle SVD* operator chooses the regularization operator  $\Phi$  depending on  $f$  to minimize the risk  $E\{\|\tilde{F} - f\|^2\}$ . Inserting (13.11) in  $E\{\|\tilde{F} - f\|^2\}$  and setting to zero partial derivatives relative to  $\|\Phi e_k\|$  proves that the risk is minimized by  $\|\Phi e_k\| = |\langle f, e_k \rangle|^{-1}$ , and the resulting minimum oral risk is

$$r_{\inf}(f) = \sum_{k \in \Gamma_Q} \frac{|\langle f, e_k \rangle|^2 \sigma^2}{|\langle f, e_k \rangle|^2 \lambda_k^2 + \sigma^2} + \sum_{k \notin \Gamma_Q} |\langle f, e_k \rangle|^2. \quad (13.12)$$

For a linear operator,  $\|\Phi e_k\|$  must remain constant for all  $f$  in a signal set  $\Theta$ . The oracle choice shows that one can find  $\|\Phi e_k\|$ , which produces a small maximum error over  $\Theta$  if the energy of all  $f \in \Theta$  is concentrated over a small number of fixed-basis vectors, and thus if  $\mathcal{B}_S$  provides efficient linear approximations of vectors in  $\Theta$ .

### Deconvolutions

Many inverse problems involve a convolution operator  $Uf[n] = f \circledast u[n]$  that we suppose to be circular to simplify border problems. In this case,  $U^*Uf = f \circledast u \circledast \tilde{u}[n]$  with  $\tilde{u}[n] = u[-n]$ . The singular basis that diagonalizes  $UU^*$  is therefore the discrete Fourier basis

$$\mathcal{B}_S = \{e_k[n] = N^{-1/2} e^{j2\pi kn/N}\}_{0 \leq k < N}.$$

The singular spectrum is  $\lambda_k^2 = |\hat{u}[k]|^2$ , and **NullU** is the space of signals  $f$  with a Fourier transform  $\hat{f}[k]$  that is nonzero only when  $\hat{u}[k] = 0$ .

An SVD deconvolution is obtained with a regularization operator  $\Phi$  that is also a convolution  $\Phi f[n] = f \circledast \phi[n]$ , so that  $\Phi^*\Phi$  is also diagonalized in the discrete Fourier basis with eigenvalues  $\|\Phi e_k\|^2 = |\hat{\phi}[k]|^2$ . The resulting diagonal SVD operator (13.11) is a convolution  $\tilde{f} = Y \circledast d$  with a transfer function that is

$$\hat{d}[k] = \frac{\hat{u}^*[k]}{|\hat{u}[k]|^2 + \sigma^2 |\hat{\phi}[k]|^2}. \quad (13.13)$$

When  $u$  is a low-pass filter, it typically restores the lower frequencies and sets to zero larger frequencies where  $|\hat{u}[k]| \ll \sigma |\hat{\phi}[k]|$ . There is no super-resolution since no signal component is restored when  $\hat{u}[k] = 0$ . A super-resolution estimator would also recover frequencies that have been totally removed by  $U$ , which is the case of the sparse spike deconvolutions in Section 13.3.2.

In a Tikhonov regularization,  $\Phi$  is a finite-difference approximation of a first-order derivative  $\phi[n] = \delta[n] - \delta[n-1]$ , so  $|\hat{\phi}[k]| = 2|\sin(\pi k/N)|$ . The resulting estimator attenuates the signal high frequencies and thus restores a regular estimation. If  $f$  is not uniformly regular, then this regularization produces a large error. Section 13.2 shows that such estimators can be improved by nonlinear estimators that take advantage of a sparse representation in a different basis, such as a wavelet basis.

## 13.2 THRESHOLDING ESTIMATORS FOR INVERSE PROBLEMS

Linear inverse estimators can be factorized into a pseudo inverse followed by a linear denoising operator that attenuates the amplified noise. Replacing the linear denoising operator by a nonlinear thresholding estimator in an appropriate orthogonal basis can improve the estimation. Following the work of Donoho [214], conditions are given to obtain a nearly minimax risk. Section 13.2.2 studies applications to signal and image deconvolutions with wavelet and wavelet packet bases.

### 13.2.1 Thresholding in Bases of Almost Singular Vectors

Suppose that there exists an orthonormal basis  $\mathcal{B} = \{g_m\}_{0 \leq m < N}$  in which  $f$  has a sparse representation. Its decomposition coefficients in  $\mathcal{B}$ ,

$$f = \sum_{m=0}^{N-1} a[m] g_m,$$

are the decomposition coefficients of

$$Uf = \sum_{m=0}^{N-1} a[m] Ug_m$$

in  $\{Ug_m\}_{0 \leq m < N}$ . Let  $U\mathcal{B} = \{Ug_m\}_{m \in \Gamma_Q}$  be the transformed basis of  $Q \leq N$  vectors such that  $Ug_m \neq 0$ . If  $U\mathcal{B}$  is a basis of  $\text{Im } U$ , then  $a[m]$  for  $m \in \Gamma_Q$  is calculated from the inner products of  $Uf$  with the vectors of a dual basis. An estimation of  $f$  from  $Y = Uf + W$  is derived with a thresholding that reduces the noise. Since  $a[m]$  is not estimated and thus set to zero if  $Ug_m = 0$ , the resulting estimator  $\tilde{F}$  of  $f$  belongs to the space  $(\text{Null } U)^\perp$  of dimension  $Q$ . It does not perform any super-resolution. Conditions are established on the basis  $\mathcal{B}$  relative to the signal class and the operator  $U$  to obtain a nearly minimax estimator.

### Thresholding Biorthogonal Bases

The transformed basis  $U\mathcal{B} = \{Ug_m\}_{m \in \Gamma_Q}$  is supposed to be a basis of the finite-dimensional space  $\text{Im } U$ . Its dual basis is characterized by biorthogonality relations. Let us renormalize the transformed basis  $\{\tilde{\lambda}_m^{-1} Ug_m\}_{m \in \Gamma_Q}$  so that the biorthogonal basis  $\{\tilde{\phi}_m\}_{m \in \Gamma_Q}$  in  $\text{Im } U$  is normalized. This basis is characterized by the biorthogonality relations:

$$\forall (m, p) \in \Gamma_Q^2, \quad \langle \tilde{\phi}_p, \tilde{\lambda}_m^{-1} Ug_m \rangle = \delta[p - m], \quad (13.14)$$

and  $\tilde{\lambda}_m$  is adjusted so that  $\|\tilde{\phi}_m\| = 1$ .

If  $f = \sum_{m=0}^{N-1} a[m] g_m$ , then  $Uf = \sum_{m \in \Gamma_Q} \tilde{\lambda}_m a[m] (\tilde{\lambda}_m^{-1} Ug_m)$ , so the coefficients  $a[m]$  are obtained by decomposing  $Uf$  in the dual basis:

$$a[m] = \tilde{\lambda}_m^{-1} \langle Uf, \tilde{\phi}_m \rangle \quad \text{for } m \in \Gamma_Q.$$

From  $Y = Uf + W$ , we get

$$\langle Y, \tilde{\lambda}_m^{-1} \tilde{\phi}_m \rangle = a[m] + \langle W, \tilde{\lambda}_m^{-1} \tilde{\phi}_m \rangle.$$

Since  $\|\tilde{\phi}_m\| = 1$ , if  $W$  is a Gaussian white noise of variance  $\sigma^2$ , then  $\langle W, \tilde{\lambda}_m^{-1} \tilde{\phi}_m \rangle$  is a Gaussian random variable of variance  $\sigma^2 \tilde{\lambda}_m^{-2}$ . Donoho [214] proposed to estimate  $a[m]$  and thus the projection of  $f$  in  $(\text{Null } U)^\perp$  with a thresholding

$$\tilde{F} = \sum_{m \in \Gamma_Q} \rho_{T_m} \left( \langle Y, \tilde{\lambda}_m^{-1} \tilde{\phi}_m \rangle \right) g_m \in (\text{Null } U)^\perp, \quad (13.15)$$

where the thresholds are  $T_m = \tilde{\lambda}_m^{-1} \sigma \sqrt{2 \log_e Q}$ .

### Diagonal Estimation with Amplified Noise

To better understand and compute the thresholding estimator (13.15), it is decomposed into a linear pseudo inverse followed by a thresholding denoising estimator. Let us write  $U_{\Gamma_Q}$  as the restriction of  $U$  to the space  $(\text{Null } U)^\perp$  generated by the  $Q$

vectors  $\{g_m\}_{m \in \Gamma_Q}$ . Like in (13.8), we apply the pseudo inverse  $U^+$ , which inverts the restriction  $U_{\Gamma_Q}$  of  $U$  to  $(\text{Null } U)^\perp$ ,

$$X = U^+ Y = U^+ U f + U^+ W \in (\text{Null } U)^\perp,$$

where  $U^+ U$  is the orthogonal projector on  $(\text{Null } U)^\perp$ . The inverted noise  $Z = U^+ W$  has a covariance on  $(\text{Null } U)^\perp$  that is  $K_Z = \sigma^2 U^+ U^{+\perp}$ . Since  $U^+ U f$  has a sparse decomposition in the basis  $\{g_m\}_{m \in \Gamma_Q}$  of  $(\text{Null } U)^\perp$ , it can be estimated by thresholding its coefficients in this basis, according to (11.66). Thresholds are proportional to the noise variance  $\sigma_B[m]^2 = E\{|\langle Z, g_m \rangle|^2\} = \langle g_m, K_Z g_m \rangle$ .

Theorem 13.2 proves that the noise amplification of  $\sigma_B[m]$  relative to  $\sigma$  is specified by the normalization factors  $\tilde{\lambda}_m$ , and it derives a thresholding estimator of  $X$ .

**Theorem 13.2.** The renormalization factors satisfy

$$\tilde{\lambda}_m^{-2} = \langle (U_{\Gamma_Q}^* U_{\Gamma_Q})^{-1} g_m, g_m \rangle = \frac{\sigma_B[m]^2}{\sigma^2}, \quad (13.16)$$

and the thresholding estimator

$$\tilde{F} = \sum_{m=0}^{Q-1} \rho_{T_m}(\langle X, g_m \rangle) g_m \in (\text{Null } U)^\perp \quad (13.17)$$

with  $T_m = \sqrt{2 \log_e Q} \sigma_B[m]$  is equal to the thresholding inverse estimator (13.15).

**Proof.** The symmetric operator  $U_{\Gamma_Q}^* U_{\Gamma_Q}$  is invertible over  $(\text{Null } U)^\perp$ . We derive from the biorthogonality relations that  $U_{\Gamma_Q}^* \tilde{\phi}_m = \tilde{\lambda}_m g_m$  for  $m \in \Gamma_Q$  and thus that  $\tilde{\phi}_m = \tilde{\lambda}_m U_{\Gamma_Q} (U_{\Gamma_Q}^* U_{\Gamma_Q})^{-1} g_m$ . Since  $\|\tilde{\phi}_m\| = 1$ , it results that  $\tilde{\lambda}_m^{-1} = \|U_{\Gamma_Q} (U_{\Gamma_Q}^* U_{\Gamma_Q})^{-1} g_m\|$ , so

$$\tilde{\lambda}_m^{-2} = \langle U_{\Gamma_Q} (U_{\Gamma_Q}^* U_{\Gamma_Q})^{-1} g_m, U_{\Gamma_Q} (U_{\Gamma_Q}^* U_{\Gamma_Q})^{-1} g_m \rangle = \langle (U_{\Gamma_Q}^* U_{\Gamma_Q})^{-1} g_m, g_m \rangle.$$

If  $g_m \in (\text{Null } U)^\perp$ , then  $K_Z g_m = \sigma^2 (U_{\Gamma_Q}^* U_{\Gamma_Q})^{-1} g_m$ , so  $\sigma^2 \tilde{\lambda}_m^{-2} = \langle K_Z g_m, g_m \rangle = \sigma_B[m]^2$ , which finishes the proof of (13.16).

As a consequence,

$$\langle Y, \tilde{\lambda}_m^{-1} \tilde{\phi}_m \rangle = \langle (U_{\Gamma_Q}^* U_{\Gamma_Q})^{-1} U_{\Gamma_Q}^* Y, g_m \rangle = \langle X, g_m \rangle,$$

so (13.15) and (13.17) threshold the same coefficients. Since  $\sigma \tilde{\lambda}_m^{-1} = \sigma_B[m]$ , the thresholds in (13.17) and (13.15) are also identical. Thus, both estimators are equal. ■

This theorem proves that the thresholding estimator (13.15) can be interpreted as a linear pseudo inverse of  $U$  followed by a thresholding estimator that reduces the amplified noise. We know from Chapter 11 that such estimators are highly efficient if  $f$  is sparse in  $\mathcal{B}$  and if the amplified noise  $Z$  has nearly independent coefficients in the basis  $\mathcal{B}$ . It implies that  $\mathcal{B}$  is a basis of “almost singular vectors” that “nearly diagonalizes” the covariance  $K_Z$  and thus  $U^* U$ .

Implementing the thresholding estimator with (13.17) rather than (13.15) may require less operations if signal coefficients in  $\mathcal{B}$  are computed with a fast algorithm and if the pseudo inverse  $U^+$  is also implemented with a fast algorithm. This is the case for deconvolution estimators in wavelet packet bases.

It is proved in Theorem 11.8 that the risk produced by a thresholding estimator is of the same order as the oracle risk (11.65) obtained by setting to zero all coefficients  $\langle X, g_m \rangle$  for which  $|\langle f, g_m \rangle| \leq \sigma_{\mathcal{B}}[m] = \tilde{\lambda}_m^{-1} \sigma$ . If over a signal class  $\Theta$  we have  $\sigma \tilde{\lambda}_m^{-1} \geq \sup_{f \in \Theta} |\langle f, g_m \rangle|$ , then the oracle systematically sets  $\langle X, g_m \rangle$  to zero because the amplified noise is too large relative to the signal. The estimation risk is thus reduced by doing the same. This is equivalent to reducing the set  $\{g_m\}_{m \in \Gamma_Q}$  to a subset  $\{g_m\}_{m \in \Gamma_{Q_0}}$  for which

$$\sigma \tilde{\lambda}_m^{-1} < \sup_{f \in \Theta} |\langle f, g_m \rangle|. \quad (13.18)$$

Suppressing the directions  $g_m$  corresponding to singular values  $\tilde{\lambda}_m$  that are too small is important in numerical applications.

### Nearly Minimax

It now remains to be understood under which conditions such a thresholding estimator is nearly optimal among all possible nonlinear estimators over a signal set  $\Theta$ . Let  $r_{\text{th}}(\Theta) = \sup_{f \in \Theta} E\{\|f - \tilde{f}\|^2\}$  be the maximum risk over  $\Theta$  of thresholding estimators (13.15) and (13.17). Theorem 13.3 proves that  $r_{\text{th}}(\Theta)$  is close to the nonlinear minimax risk  $r_n(\Theta)$  if  $\Theta$  is orthosymmetric in  $\mathcal{B}$ , and if the transformed basis  $U\mathcal{B}$  satisfies a Riesz stability property, which implies that  $\mathcal{B}$  “nearly diagonalizes”  $U^*U$ . Section 11.5.2 explains that  $\Theta$  is orthosymmetric in  $\mathcal{B}$  if any  $f \in \Theta$  remains in  $\Theta$  when reducing the amplitude of any of its decomposition coefficients in  $\mathcal{B}$ . Such a set is aligned with the vectors’ directions in  $\mathcal{B}$ , which provides sparse signal approximations.

**Theorem 13.3: Donoho.** If  $\Theta$  is orthosymmetric in  $\mathcal{B}$  and there exists  $B > 0$  such that

$$\forall a \in \mathbb{C}^Q, \quad \left\| \sum_{m \in \Gamma_Q} a[m] \tilde{\lambda}_m^{-1} U g_m \right\|^2 \leq B \|a\|^2$$

with (13.19)

$$\tilde{\lambda}_m^{-2} = \langle (U_{\Gamma_Q}^* U_{\Gamma_Q})^{-1} g_m, g_m \rangle,$$

then for thresholds  $T_m = \sigma \tilde{\lambda}_m^{-1} \sqrt{2 \log_e Q}$ , the maximum thresholding risk  $r_{\text{th}}(\Theta)$  satisfies for  $N \geq 4$

$$r_n(\Theta) \leq r_{\text{th}}(\Theta) \leq (2 \log_e Q + 1) \left( \bar{\sigma}^2 + 1.25 B r_n(\Theta) \right) \quad (13.20)$$

with  $\bar{\sigma}^2 = Q^{-1} \sigma^2 \sum_{m \in \Gamma_Q} \tilde{\lambda}_m^{-2}$ .

**Proof.** The main steps of the proof are given without detail. The thresholding risk  $r_{\text{th}}(f) = E\{\|\tilde{f} - f\|^2\}$  is first compared to the minimum oracle risk  $r_{\text{inf}}(f)$  of diagonal estimators.

Computing this oracle risk as in (11.31) gives

$$r_{\inf}(f) = \sum_{m=0}^{Q-1} \frac{\sigma_m^2 |f_{\mathcal{B}}[m]|^2}{\sigma_m^2 + |f_{\mathcal{B}}[m]|^2} + \sum_{m=Q}^{N-1} |f_{\mathcal{B}}[m]|^2. \quad (13.21)$$

Since a thresholding estimator is a diagonal estimator,  $r_{\text{th}}(f) \geq r_{\inf}(f)$ . The result (11.67) of Theorem 11.7 can be refined by replacing  $r_{\text{pr}}(f)$  by  $r_{\inf}(f)$ :

$$r_{\inf}(f) \leq r_{\text{th}}(f) \leq (2 \log_e Q + 1) \left( \tilde{\sigma}^2 + r_{\inf}(f) \right) \quad \text{with} \quad \tilde{\sigma}^2 = Q^{-1} \sum_{m \in \Gamma_Q} \sigma_{\mathcal{B}}^2[m], \quad (13.22)$$

and  $\sigma_{\mathcal{B}}^2[m] = \langle K_Z g_m, g_m \rangle = \sigma^2 \tilde{\lambda}_m^{-2}$ . If  $\Theta$  is orthosymmetric in  $\mathcal{B}$ , then we prove that

$$r_{\inf}(\Theta) \geq 1.25 B r_n(\Theta). \quad (13.23)$$

The proof of this result considers first the particular case where  $\text{NullU} = \{0\}$  and  $U^* U$  transforms  $\mathcal{B}$  in an orthogonal basis. It implies that  $U^* U$  is diagonal in  $\mathcal{B}$  and thus that the covariance matrix  $K_Z$  of the inverted noise  $Z = U^+ W$  is also diagonal in  $\mathcal{B}$ . As a result, the noise coefficients in  $\mathcal{B}$  are independent. Since  $\Theta$  is orthosymmetric in  $\mathcal{B}$ , renormalizing the noise gives a white noise, and Theorem 11.14 implies that diagonal estimators in  $\mathcal{B}$  are nearly minimax, with

$$r_n(\Theta) \geq \frac{1}{1.25} r_{\inf}(\Theta). \quad (13.24)$$

This result is then extended for a nondiagonal covariance  $K_Z$ . Let  $K_d$  be the diagonal matrix in  $\mathcal{B}$  with a diagonal equal to the diagonal of  $K_Z$ . One can verify that

$$K_Z \geq B^{-1} K_d \iff \forall f \in \mathbb{C}^N, \quad \langle K_Z f, f \rangle \geq B^{-1} \langle K_d f, f \rangle. \quad (13.25)$$

As a consequence of this inequality, a noise augmentation lemma proves that the minimax risk with a noise of covariance  $K_Z$  is necessarily larger than the minimax risk when the noise covariance is  $B^{-1} K_d$ . Using this result, (13.23) is derived from (13.24), which applies to  $B^{-1} K_d$ . When  $\text{NullU}$  is not empty,  $U^* U$  is written as a limit of operators  $U_k$  with a null space  $\text{NullU}_k$  that is not empty and verifies (13.23) according to this proof. The result is then proved for  $U$  by taking the limit on  $k$ .

Inequalities (13.22) and (13.23) imply

$$\frac{B}{1.25} r_{\inf}(\Theta) \leq r_n(\Theta) \leq r_{\text{th}}(\Theta) \leq (2 \log_e Q + 1) \left( \tilde{\sigma}^2 + r_{\inf}(\Theta) \right), \quad (13.26)$$

which proves (13.20). ■

The constant  $B$  in (13.19) is the upper Riesz bound of the normalized basis  $\{\tilde{\lambda}_m^{-1} U g_m\}_{m \in \Gamma_Q}$ . Having a stable normalized basis is a requirement to stabilize the thresholding estimation. If  $B$  is not too large and if  $\Theta$  is orthosymmetric, then this theorem proves that the maximum risk  $r_{\text{th}}(\Theta)$  of a thresholding estimator is of the same order as the nonlinear minimax risk  $r_n(\Theta)$ . This result is further improved by reducing  $\{g_m\}_{m \in \Gamma_Q}$  according to (13.18) into a subset  $\{g_m\}_{m \in \Gamma_{Q_0}}$ , which satisfies  $\sigma \tilde{\lambda}_m^{-1} < \sup_{f \in \Theta} |\langle f, g_m \rangle|$ . Indeed, one can verify [324] that it yields a tighter

inequality (13.20) where  $Q$  is replaced by  $Q_0$  and  $\bar{\sigma}^2$  by  $\bar{\sigma}_0^2 = Q_0^{-1} \sigma^2 \sum_{m \in \Gamma_{Q_0}} \tilde{\lambda}_m^{-2}$ . This sum does not include the smallest  $\tilde{\lambda}_m$  and is thus potentially much smaller.

When  $\Theta$  is not orthosymmetric but can be embedded in two close orthosymmetric sets, then applying this theorem to each orthosymmetric set gives a similar result. In this case, since the thresholding estimator performs no super-resolution and is nearly minimax, it also implies that there is no super-resolution estimator that provides a significant improvement for all signals in  $\Theta$ . Indeed, the basis  $\mathcal{B}$  is optimal to represent  $f$ , and it includes  $N - Q$  vectors that are completely cancelled by the operator  $U$ , along which the signal coefficients cannot be recovered.

### Almost Singular Vectors with Narrow Spectrum

Section 5.1.2 proves that for a normalized Riesz basis the upper Riesz bound satisfies  $B \geq 1$ , and if  $B = 1$ , then the basis is orthonormal. In this case,  $g_m$  is an eigenvector of  $U^* U$  with an eigenvalue  $\tilde{\lambda}_m^2 = \|U g_m\|^2$ . This is generally not the case, but to get a small constant  $B$ , the basis  $\mathcal{B}$  must nearly diagonalize  $U^* U$ , and the normalization constants  $\tilde{\lambda}_m$  are then approximately equal to singular values.

Let  $\mathcal{B}_s = \{e_k\}_{k \in \Gamma}$  be a singular basis that diagonalizes  $U^* U$  with singular values  $\{\lambda_k^2\}_{k \in \Gamma}$ . Each  $g_m \in \mathcal{B}$  is a mix of singular values  $\lambda_k^2$ , which define its singular spectrum support. Theorem 13.4 relates  $B$  to the relative variations of the singular spectrum for vectors in  $\mathcal{B}$ .

**Theorem 13.4.** Let  $\mathbf{C}^N = \bigoplus_{l=1}^L \mathbf{U}_l$  be a partition in orthogonal spaces  $\mathbf{U}_l$  generated by families of singular vectors  $\{e_k\}_{k \in \Gamma_l}$  with singular values  $\{\lambda_k^2\}_{k \in \Gamma_l}$ . Let  $\mathcal{B} = \{g_m\}_{m \in \Gamma}$  be an orthonormal basis of  $\mathbb{C}^N$  obtained as a union of orthonormal bases of each  $\mathbf{U}_l$ . If the constant  $B$  satisfies

$$B \left( \min_{k \in \Gamma_l} \lambda_k^2 \right) \geq \max_{k \in \Gamma_l} \lambda_k^2 \quad \text{for } 0 \leq l < L, \quad (13.27)$$

then

$$\forall a \in \mathbb{C}^Q, \quad \left\| \sum_{m \in \Gamma_Q} a[p] \tilde{\lambda}_m^{-1} U g_m \right\|^2 \leq B \|a\|^2 \quad (13.28)$$

with

$$\tilde{\lambda}_m^{-2} = ((U_{\Gamma_Q}^* U_{\Gamma_Q})^{-1} g_m, g_m).$$

**Proof.** Let  $\lambda_{l,\min} = \min_{k \in \Gamma_l} \lambda_k$  and  $\lambda_{l,\max} = \max_{k \in \Gamma_l} \lambda_k$ . We write  $U_{\Gamma_l}$  as the restriction of  $U$  to the space  $\mathbf{U}_l$ . Since each space  $\mathbf{U}_l$  is generated by  $\{e_k\}_{k \in \Gamma_l}$ , it results that

$$\lambda_{l,\min} \text{Id} \leq U_{\Gamma_l}^* U_{\Gamma_l} \leq \lambda_{l,\max} \text{Id}.$$

For  $g_m \in \mathbf{U}_l$ , we thus have  $\lambda_{l,\min} \leq \tilde{\lambda}_m \leq \lambda_{l,\max}$ . Moreover,

$$\begin{aligned} \left\| \sum_{m \in \Gamma_Q} a[p] \tilde{\lambda}_m^{-1} U g_m \right\|^2 &= \sum_{l=1}^L \left\| \sum_{g_m \in \mathbf{U}_l} a[p] \tilde{\lambda}_m^{-1} U g_m \right\|^2 \leq \sum_{l=1}^L \lambda_{l,\min}^{-2} \left\| \sum_{g_m \in \mathbf{U}_l} a[p] U g_m \right\|^2 \\ &\leq \sum_{l=1}^L \lambda_{l,\min}^{-2} \lambda_{l,\max}^2 \left\| \sum_{g_m \in \mathbf{U}_l} a[p] g_m \right\|^2 \leq B \|a\|^2, \end{aligned}$$

which proves (13.28). ■

The constant  $B$  in (13.27) is the maximum relative variation of singular values in the spaces  $\mathbf{U}_l$  where the bases vectors  $g_m$  belong. In this case, each  $g_m$  has a decomposition over singular vectors with singular values that have relative variations bounded by  $B$ . If the vectors  $g_m$  have a *narrow singular spectrum*, then  $B$  gets close to 1.

The condition (13.27) is not strictly necessary and can be relaxed by just imposing that most of the energy of  $g_m$  is concentrated over singular values that have relative variation bounded by  $B$ . For deconvolutions, wavelets and wavelet packets are examples of bases providing sparse signal representations with a narrow singular spectrum in that sense. Other inverse problems have been more recently investigated with this approach [326]. When signals in  $\Theta$  have a sparse representation in a basis with vectors that have a *spread spectrum*, then thresholding estimators are not optimal, but Section 13.3 shows that there may be an opportunity for a super-resolution estimation of  $f$ .

### 13.2.2 Thresholding Deconvolutions

The deconvolution estimation of  $f$  from  $Y = Uf + W$  with  $Uf[n] = f \circledast u[n]$  is studied in Section 13.1 with linear operators that are diagonal in the discrete Fourier basis

$$\mathcal{B}_S = \{g_m[n] = N^{-1/2} e^{i2\pi mn/N}\}_{0 \leq m < N}.$$

The operator  $U^* U$  is a convolution diagonalized in this Fourier basis, and its transfer function is equal to the singular values  $\lambda_k^2 = |\hat{u}[k]|^2$ . Signals including singularities are not well approximated in a Fourier basis, and the resulting linear estimators produce a large risk. To reduce this risk with a thresholding estimator, one must find a basis  $\mathcal{B}$  providing a sparse signal representation with vectors having a narrow spectrum. Theorem 13.4 shows that each vector  $g_m \in \mathcal{B}$  must have a Fourier transform  $\hat{g}_m[k]$  that as energy is concentrated over frequencies  $k$  for which  $\lambda_k^2 = |\hat{u}[k]|^2$  has small relative variations. We consider two types of deconvolution problems where such bases can be constructed with wavelets or wavelet packets.

#### *Homogeneous Deconvolutions with Wavelets*

Derivative and integral operators are examples of convolution operators with transfer functions that vanish at the zero frequency or at infinity, with a homogeneous decay. After discretization, a homogeneous convolution operator  $Uf[n] = f \circledast u[n]$  has by definition a transfer function that satisfies  $|\hat{u}[k]| \sim |k|^p$ . A first-order derivative  $u[n] = \delta[n+1] - \delta[n-1]$  is homogeneous with  $p=1$ :  $|\hat{u}[k]| \sim |k|$ . A derivative of order  $p$  yields  $|\hat{u}[k]| \sim |k|^p$ . Their inverse is singular at  $k=0$  and  $\text{Null } U$  is thus reduced to constant signals. Integrations are homogeneous convolutions with  $p < 0$  and their inverse becomes singular at high frequencies.

Wavelet bases provide sparse representations of piecewise regular signals. Harmonic analysis results [44] also prove that singular homogeneous operators are “nearly diagonalized” in a wavelet basis. Indeed, a wavelet  $\psi_{j,m}$  has a Fourier transform  $\hat{\psi}_{j,m}[k]$  mostly concentrated on a dyadic frequency interval  $k \in [2^{-j-1}, 2^{-j}]$ .

Over such an interval,  $|\hat{u}[k]| \sim |k|^p$  varies by a factor of the order of  $2^p$  that does not depend on the scale. Suppose that the discrete wavelets have  $q > p$  vanishing moments and correspond to the discretization of a regular wavelet  $\psi(t)$  that is  $\mathbf{C}^q$ . For  $L = -\log_2 N$ , one can then verify that a periodic orthonormal wavelet family

$$\{\psi_{j,m}\}_{L < j \leq 0, 0 \leq m < 2^{-j}} \quad (13.29)$$

is transformed into a Riesz basis by a homogeneous convolution operator  $U$  with an upper Riesz bound  $B \sim 2^{2p}$  after renormalization. The constant scaling signal  $\phi_0[n] = N^{-1/2}$  is not included because it is in **NullU**.

The transformed wavelets  $U\psi_{j,m}$  are similar to wavelets and are called *vaguelettes* by Donoho [214]. The asymptotic minimax optimality of wavelet thresholding estimators is proved by Theorem 13.3 for homogeneous deconvolutions of signals that have a sparse signal representation in a wavelet basis. This includes bounded variation signals and images.

### Mirror Wavelets Deconvolution

Analog-acquisition devices often remove high frequencies with a low-pass filter that vanishes at some maximum frequency. The sampling rate discretization is adjusted to this maximum frequency to avoid aliasing. If the low-pass transfer function has a smooth decay in the neighborhood of the maximum frequency, then the discretized signal is blurred. Optical systems often produce such a blur.

The maximum analog frequency is mapped by the discretization to the highest discrete frequency  $2\pi k/N = \pm\pi/2$ . The discrete signal blurring can thus be written as a discrete low-pass filter  $Uf[n] = f \circledast u[n]$  with a transfer function  $\hat{u}[k]$  that has a zero of order  $p \geq 1$  at the maximum frequency index  $k = N/2$ :

$$|\hat{u}[k]| \sim |k - N/2|^p. \quad (13.30)$$

It results that **NullU** corresponds to signals  $h[n]$  such that  $\hat{h}[k] \neq 0$  only for  $k = \pm N/2$ , and thus  $h[n] = c(-1)^n$ . Since wavelet bases provide sparse representation of piecewise regular signals, they could be a good candidate to implement a thresholding deconvolution estimator. This requires to nearly diagonalize  $U^*U$ , and thus that the singular spectrum  $|\hat{u}[k]|^2$  and its inverse  $|\hat{u}[k]|^{-2}$  have small relative variations over the support of the Fourier transform of each basis vector.

At scales  $2^j > 2N^{-1}$ ,  $\hat{\psi}_{j,m}[k]$  has a frequency support nearly included in an interval  $[2^{-j-1}, 2^{-j}]$  where  $|\hat{u}[k]|^{-2}$  remains nearly constant. However, at the finest scale  $2^{L+1} = 2N^{-1}$ , wavelets  $|\hat{\psi}_{L+1,m}[k]|$  have a spread spectrum because their energy is mainly concentrated in the higher-frequency band  $[N/4, N/2]$ , where  $|\hat{u}[k]|$  varies by a huge factor on the order of  $N^{2p}$ . These fine-scale wavelets must therefore be replaced by wavelet packet vectors having a smaller-frequency support adjusted to the rapid relative variation of  $|\hat{u}[k]|$ .

To efficiently approximate piecewise regular signals, these wavelet packets must have the smallest possible spatial support, and thus the largest possible

frequency support. The optimal trade-off is obtained with wavelet packets that we denote  $\tilde{\psi}_{j,m}$  that have a discrete Fourier transform  $\hat{\tilde{\psi}}_{j,m}[k]$  mostly concentrated in  $[N/2 - 2^{-j}, N/2 - 2^{-j-1}]$ , as illustrated by Figure 13.1. Over such intervals,  $|\hat{u}[k]|^{-2}$  varies by a relative factor of  $2^{2p}$  that does not depend on the scale  $2^j$ . These particular wavelet packets, introduced by Kalifa and Mallat [323, 324], are called *mirror wavelets* because they are related to discrete wavelets by

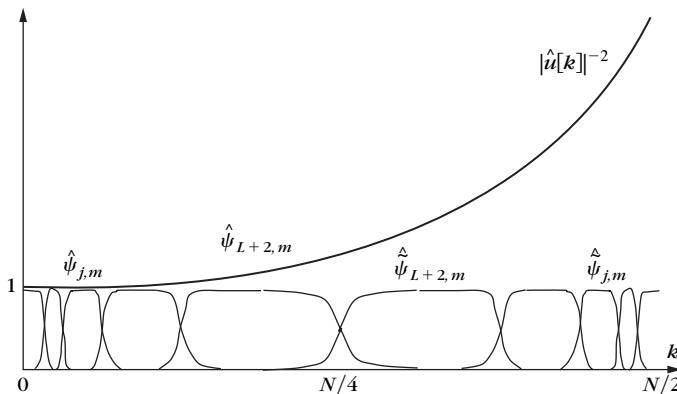
$$|\hat{\tilde{\psi}}_{j,m}[k]| = |\hat{\psi}_{j,m}[N/2 - k]| \quad \text{and} \quad \tilde{\psi}_{j,m}[n] = (-1)^{n-1} \psi_{j,m}[1-n].$$

A mirror wavelet basis is a wavelet packet basis composed of wavelets  $\psi_{j,m}$  at scales  $2^j > 2^{L+1} = 2N^{-1}$  and of mirror wavelets to replace wavelets at the finest scale  $2^{L+1}$ :

$$\left\{ \psi_{j,m}, \tilde{\psi}_{j,m} \right\}_{0 \leq m < 2^{-j}, L+1 \leq j \leq 0}.$$

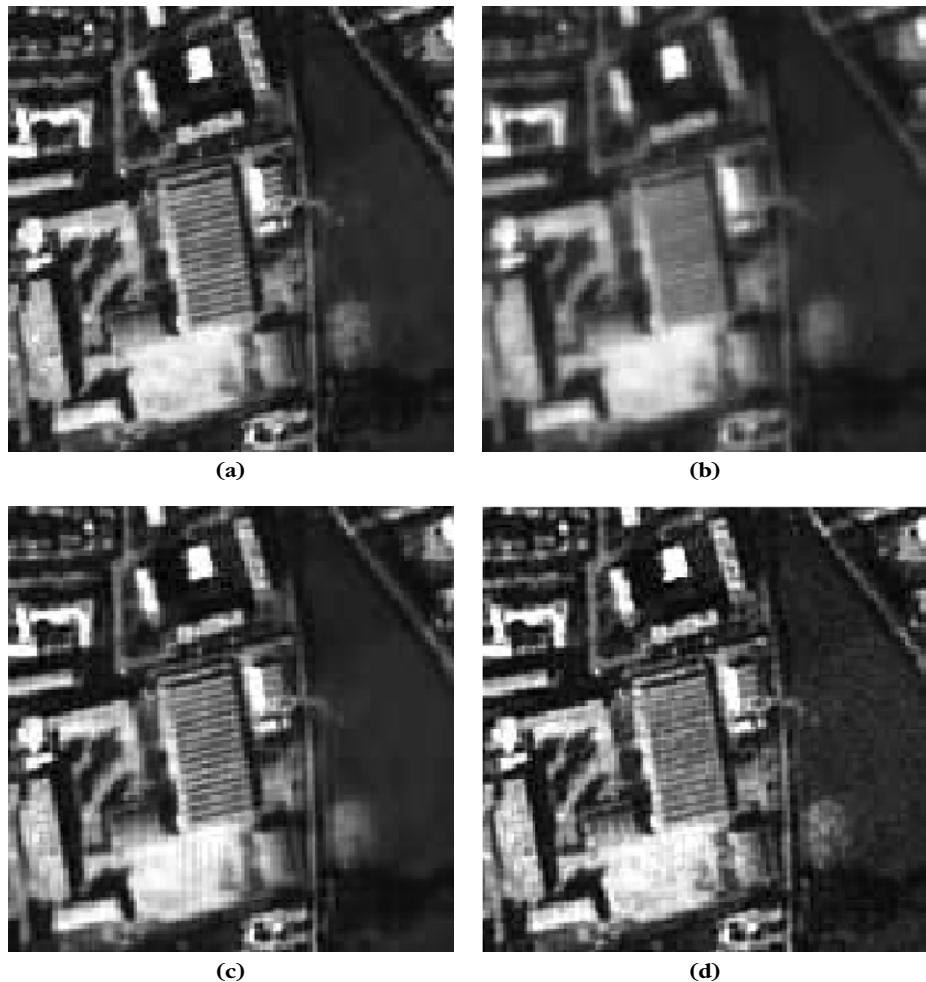
The highest-frequency mirror wavelet  $\tilde{\psi}_{0,0}[n] = N^{-1/2}(-1)^n$  belongs to **NullU** and is therefore not included in the estimation. The fast mirror wavelet transform studied in Exercise 8.10 is implemented with a wavelet packet filter bank, as described in Section 8.1.4.

If these wavelets and wavelet packets are constructed with conjugate mirror filters that define a continuous time wavelet  $\psi(t)$  that is  $C^q$  with  $q > p$  vanishing moments, then Kalifa and Mallat [323] prove that a thresholding estimator in a mirror wavelet basis yields a quasi-minimax deconvolution estimator for bounded variation signals. The resulting risk is then much smaller than the risk obtained by a linear singular value decomposition estimator.



**FIGURE 13.1**

The singular spectrum  $|\hat{u}[k]|^2$  of a low-pass filter decreases to zero at high frequencies. The support of mirror wavelets is reduced at high frequencies so that the relative variation of  $|\hat{u}[k]|^{-2}$  remains uniformly bounded over their support.

**FIGURE 13.2**

(a) Original airplane image. (b) Simulation of a satellite image provided by CNES (SNR = 31.1 db). (c) Deconvolution with a translation-invariant thresholding in a mirror wavelet basis (SNR = 34.1 db). (d) Deconvolution calculated with a circular convolution, which yields a nearly minimax risk for bounded variation images (SNR = 32.7 db).

### ***Deconvolution of Images***

For separable low-pass filters that vanish at the highest frequencies, nearly optimal deconvolution of bounded variation images is calculated with a separable extension of the deconvolution estimator in a mirror wavelet basis. Such restoration algorithms are used in wavelet packet and mirror wavelet bases [324, 417, 418] for deblurring satellite images. The exposition time of the satellite photoreceptors cannot be

reduced too much because the light intensity reaching the satellite is small and must not be dominated by electronic noises. The satellite movement thus produces a blur, which is aggravated by the imperfection of the optics. The electronics of the photoreceptors add a Gaussian white noise. The Figure 13.2(b), provided by the French spatial agency CNES, is a simulated satellite image calculated from the airplane image shown in Figure 13.2(a).

Figure 13.2(c) shows an example of deconvolution calculated in the mirror wavelet basis. The thresholding is performed with a translation-invariant algorithm. This can be compared with the linear estimation in Figure 13.2(d), calculated with a circular convolution estimator that has a maximum risk over bounded variation images close to the minimax linear risk. The linear deconvolution sharpens the image but leaves a visible noise in the regular parts of the image. The thresholding algorithm mostly removes the noise in these regions while improving the restoration of edges and oscillatory parts. Algorithms alternating between linear estimations and thresholding estimations in a wavelet basis can also provide efficient deconvolutions of such images [385].

### 13.3 SUPER-RESOLUTION

Numerically increasing the resolution of measured data has major industrial applications when data acquisition is difficult or costly. In geophysics, the highest possible resolution must be recovered from relatively few measurements obtained by sending waves underground and measuring reflections with sensors distributed on the sea or on the ground. In X-ray imaging, the radiation time of a patient and thus the data acquisition are also limited. For Earth observation, improving resolution usually means sending a new satellite, which is not a light project. On the consumer front, the resolution of videos in standard-definition television formats (PAL or NTSC) must be numerically increased to match the larger resolution of high-definition flat panel televisions. Many more examples can be found.

The linear and thresholding estimators in Sections 13.1 and 13.2 estimate the projection of  $f \in \mathbb{C}^N$  in a space of dimension  $Q_0 \leq Q = \dim(\mathbf{Im}\mathbf{U}) \leq N$ , which provides no super-resolution. Given  $Q < N$  independent measurements, super-resolution aims at estimating the projection of  $f$  in a space of dimension larger than  $Q$ , and if possible of dimension  $N$ .

#### 13.3.1 Sparse Super-resolution Estimation

Similar to Section 13.2, we suppose that  $f$  has a sparse approximation in a dictionary  $\mathcal{D} = \{g_p\}_{p \in \Gamma}$  of size  $|\Gamma| = P > Q$ , but with  $Ug_p \neq 0$  for all  $p \in \Gamma$ . This dictionary may not generate whole space  $\mathbb{C}^N$ . The projection of  $f$  is estimated from  $Y = Uf + W$  over the space generated by this dictionary, which has a dimension larger than  $Q$  and is adjusted depending on super-resolution capabilities.

Sparsity means that  $f$  is precisely approximated by its orthogonal projection  $f_\Lambda$  over a subspace generated by a small number  $|\Lambda|$  of vectors  $\{g_p\}_{p \in \Lambda}$  chosen in  $\mathcal{D}$ :

$$f_\Lambda = \sum_{p \in \Lambda} a[p] g_p. \quad (13.31)$$

The error  $w_\Lambda = f - f_\Lambda$  thus has a small norm. The approximation vectors  $\{g_p\}_{p \in \Lambda}$  are not restricted a priori to a space of dimension  $Q$ . Since  $Y = Uf + W$ , it results from (13.31) that

$$Y = \sum_{p \in \Lambda} a[p] U g_p + W' \quad \text{with} \quad W' = U w_\Lambda + W. \quad (13.32)$$

The coefficients  $a[p]$  can be estimated with a sparse denoising estimation of  $Y$  in the transformed and normalized dictionary

$$\mathcal{D}_U = \left\{ \frac{U g_p}{\|U g_p\|} \right\}_{p \in \Gamma}. \quad (13.33)$$

A major difficulty is that  $\mathcal{D}_U$  is a redundant dictionary with  $P > Q$  vectors that are in the operator image space  $\mathbf{Im}U$  of dimension  $Q$ .

Let  $\tilde{Y}$  be a sparse approximation of  $Y$  computed with an algorithm that selects a subset of dictionary vectors  $\{U g_p / \|U g_p\|\}_{p \in \tilde{\Lambda}}$ :

$$\tilde{Y} = \sum_{p \in \tilde{\Lambda}} \tilde{a}[p] \frac{U g_p}{\|U g_p\|}. \quad (13.34)$$

An  $\mathbf{l}^1$  Lagrangian pursuit (12.89) computes such an approximation with:

$$\tilde{a} = \underset{a \in \mathbb{C}^P}{\operatorname{argmin}} \frac{1}{2} \|Y - \sum_{p \in \Gamma} a[p] \frac{U g_p}{\|U g_p\|}\|^2 + T \|a\|_1. \quad (13.35)$$

This minimization can be solved by the iterative thresholding algorithm in Theorem 12.9. A matching pursuit or an orthogonal matching pursuit can also compute the coefficients  $\tilde{a}$  of a sparse approximation  $\tilde{Y}$  in  $\mathcal{D}_U$ . An estimation  $\tilde{F}$  of  $f$  is derived by inverting  $U$  on the decomposition (13.34) of  $\tilde{Y}$ :

$$\tilde{F} = \sum_{p \in \tilde{\Lambda}} \frac{\tilde{a}[p]}{\|U g_p\|} g_p. \quad (13.36)$$

If  $\mathcal{D}$  is an orthonormal basis that diagonalizes the operator  $U^*U$ , then one can verify that an  $\mathbf{l}^1$  Lagrangian pursuit computes an estimator  $\tilde{F}$  that is identical to the soft-thresholding inverse estimator (13.15), but it provides no super-resolution.

Several conditions are necessary to recover a super-resolution estimation of  $f$ :

- *Stability.*  $\{U g_p / \|U g_p\|\}_{p \in \Lambda}$  must be a Riesz basis.
- *Support recovery.* Decomposing  $Y$  in  $\mathcal{D}_U$  must recover a support  $\tilde{\Lambda}$  that closely approximates the approximation support  $\Lambda$  of  $f$ .
- *Spread singular spectrum.* Vectors in  $\mathcal{D}$  must mix singular values of different amplitudes so that each  $\|U g_p\|$  is not too small.

### Stability with Incoherence and Sparsity

Suppose that the approximation support  $\Lambda$  is given by some oracle. When there is no noise  $W=0$ , the decomposition coefficients  $a[p]$  of  $f_\Lambda$  in  $\{g_p\}_{p \in \Lambda}$  can be recovered from the decomposition of  $Y = Uf$  in  $\{Ug_p/\|Ug_p\|\}_{p \in \Lambda}$  only if this family is linearly independent. If this is the case, in presence of noise, a stable computation also requires a nonzero lower Riesz bound  $A_\Lambda > 0$ :

$$\forall a \in \mathbb{C}^{|\Lambda|}, \quad A_\Lambda \|a\|^2 \leq \left\| \sum_{p \in \Lambda} a[p] \frac{Ug_p}{\|Ug_p\|} \right\|^2. \quad (13.37)$$

This is a nontrivial condition because  $\mathcal{D}_U$  is a redundant dictionary of  $P > Q$  vectors in a space of dimension  $Q$ , and the support  $\Lambda$  may a priori recombine any subset of vectors in  $\mathcal{D}_U$ . A vector  $\phi_p$  for  $p \in \Lambda$  should not be closely approximated by a linear combination of few other vectors in  $\Lambda$ , which is an incoherence property. Section 12.5.1 shows that the condition  $A_\Lambda > 0$  is less difficult to obtain if the number of vectors  $|\Lambda|$  is small relative to  $Q$ , and thus if the signal approximation is sparse.

### Support Recovery

The computed support  $\tilde{\Lambda}$  must provide a good estimation of  $\Lambda$  in the sense that the projection  $f_{\tilde{\Lambda}}$  of  $f$  in the space generated by  $\{g_p\}_{p \in \tilde{\Lambda}}$  should have an error  $\|f - f_{\tilde{\Lambda}}\|$  comparable to  $\|f - f_\Lambda\|$ . If  $\mathcal{D}$  is an orthonormal basis, it implies recovering a subset of the support of  $\Lambda$ , which carries the coefficients  $\langle f, g_p \rangle$  of large amplitude. Decomposing  $Y$  in  $\mathcal{D}_U$  may not recover this support because  $Y$  does not have a unique decomposition in this redundant dictionary. If there is no noise and  $f = f_\Lambda$  then  $Y = Uf_\Lambda$  belongs to the space  $\mathbf{V}_\Lambda$  generated by  $\{Ug_p\}_{p \in \Lambda}$ . Let  $\Lambda^c$  be the complement of  $\Lambda$ . The *exact recovery criteria* (ERC) by Tropp [461] imposes that

$$\text{ERC}(\Lambda) = \sup_{h \in \mathbf{V}_\Lambda} \frac{\max_{q \in \Lambda^c} |\langle h, Ug_q \rangle| / \|Ug_q\|}{\max_{p \in \Lambda} |\langle h, Ug_p \rangle| / \|Ug_p\|} < 1.$$

Theorems 12.11 and 12.15 then prove that the support  $\Lambda$  of  $f$  is recovered by decomposing  $Y$  with a matching pursuit or an  $\mathbf{l}^1$  basis pursuit in  $\mathcal{D}$ . Theorem 12.12 also proves that the condition  $\text{ERC}(\Lambda) < 1$  requires that vectors in  $\Lambda$  have a small mutual correlation and a small correlation with any vector in the complement  $\Lambda^c$ . This is again an incoherence property over dictionary vectors. In presence of noise, the Riesz stability (13.37) is crucial to partially recover this support with an orthogonal matching pursuit or an  $\mathbf{l}^1$  Lagrangian pursuit, as proved by Theorems 12.13 and 12.15.

### Spread Singular Spectrum

The coefficients  $\tilde{a}$  calculated from  $Y = Uf + W$  carry the projection of the noise over the space generated by  $\{Ug_p/\|Ug_p\|\}_{p \in \tilde{\Lambda}}$ . This noise is amplified by the normalization factors  $1/\|Ug_p\|$  in  $\tilde{F}$ , which should not be too large. Recovering a signal coordinate in the direction of  $g_p \in \mathcal{D}$  must not amplify the noise above the maximum

signal coefficient that can be recovered. If  $W$  is a white noise of variance  $\sigma^2$ , it implies that

$$\forall p \in \Gamma, \quad \sigma \|Ug_p\|^{-1} < \sup_{f \in \Theta} |\langle f, g_p \rangle|, \quad (13.38)$$

where  $\Theta$  is the set of all possible signals. The same condition appears in (13.18) for thresholding inverse estimators.

Let  $\mathcal{B}_S = \{e_k\}_{0 \leq k \leq N}$  be a singular vector basis that diagonalizes  $U^*U$  with singular values  $\{\lambda_k^2\}_{0 \leq k \leq N}$ . Since

$$\|Ug_p\|^2 = \langle U^*Ug_p, g_p \rangle = \sum_{k=0}^{N-1} \lambda_k^2 |\langle g_p, e_k \rangle|^2,$$

to guarantee that  $\|Ug_p\|$  is not too small,  $g_p$  must have part of its energy spread over singular vectors  $\{e_k\}_k$  having relatively large singular values  $\{\lambda_k^2\}_k$ . However, to recover a super-resolution estimation of  $f$  in **NullU** or in directions  $e_k$  where  $\lambda_k^2$  is small, the vectors  $g_p$  must also be spread over these directions. Each  $g_p$  should thus have a spread spectrum that mixes small and large singular spectrum values. Ideally,  $|\langle g_p, e_k \rangle| = N^{-1/2}$  and  $\|Ug_p\|^2 = N^{-1} \sum_{k=0}^{N-1} \lambda_k^2$ . This condition is opposite to the narrow spectrum condition in Theorem 13.4 for thresholding estimators, which cannot perform any super-resolution.

If all vectors  $g_p$  have a fully spread spectrum, then  $\|Ug_p\|$  is approximately constant for all  $p \in \Gamma$ . The normalization then has a more marginal impact and  $Y$  can be decomposed in a nonnormalized transformed dictionary  $\mathcal{D}_U = \{Ug_p\}_{p \in \Gamma}$ . The normalized  $\mathbf{l}^1$  Lagrangian minimization (13.35) and (13.36) are then replaced by

$$\tilde{F} = \sum_{p \in \tilde{\Lambda}} \tilde{a}[p] g_p \quad \text{with} \quad \tilde{a} = \underset{a \in \mathbb{C}^p}{\operatorname{argmin}} \frac{1}{2} \|Y - \sum_{p \in \Lambda} a[p] U g_p\|^2 + T \|a\|_1, \quad (13.39)$$

which can simplify computations.

Suppose that some vectors  $g_p \in \mathcal{D}$  do not have a sufficiently spread spectrum and do not even satisfy the maximum noise amplification condition (13.38). If these vectors are not removed from the dictionary, the transformed dictionary should not be normalized to avoid numerical instabilities. This is equivalent to solve the nonnormalized  $\mathbf{l}^1$  Lagrangian minimization (13.39). Directions  $g_p$  for which  $\|Ug_p\|$  is small are then barely recovered, which is indirectly equivalent to removing these vectors from the dictionary, but the lack of normalization penalizes the recovery of other directions.

### **Super-Resolution Recovery**

Let us consider a normalized sparse super-resolution estimation calculated with an  $\mathbf{l}^1$  Lagrangian pursuit:

$$\tilde{F} = \sum_{p \in \tilde{\Lambda}} \tilde{a}[p] \frac{g_p}{\|Ug_p\|} \quad \text{with} \quad \tilde{a} = \underset{a \in \mathbb{C}^p}{\operatorname{argmin}} \frac{1}{2} \|Y - \Phi_U^* a\|^2 + T \|a\|_1, \quad (13.40)$$

where  $\Phi_U^* \alpha = \sum_{p \in \Gamma} \alpha[p] U g_p / \|U g_p\|$ . Theorem 13.5 computes a conservative upper bound of the estimation error by setting a threshold  $T$  large enough so that the support  $\tilde{\Lambda}$  of  $\tilde{\alpha}$  satisfies  $\tilde{\Lambda} \subset \Lambda$  with a high probability. The theorem assumes that the approximation family  $\{g_p\}_{p \in \Lambda}$  of  $f$  is a Riesz basis of the space it generates, and we write  $\bar{B}_\Lambda$ , the upper Riesz bound. The theorem also assumes that the transformed vectors  $\{U g_p / \|U g_p\|\}_{p \in \Lambda}$  define a Riesz basis with lower Riesz bounds  $A_\Lambda > 0$ , and that the exact recovery condition  $\text{ERC}(\Lambda) < 1$  is satisfied.

**Theorem 13.5.** If  $\text{ERC}(\Lambda) < 1$  and

$$T = \lambda \frac{\|U\|_s \|f - f_\Lambda\| + \sigma \sqrt{2 \log_e P}}{1 - \text{ERC}(\Lambda)} \quad \text{with } \lambda > 1, \quad (13.41)$$

then there exists a unique  $\mathbf{l}^1$  pursuit solution  $\tilde{\alpha}$  with a support that satisfies  $\tilde{\Lambda} \subset \Lambda$  and the estimator  $\tilde{F} = \sum_{p \in \tilde{\Lambda}} \tilde{\alpha}[p] g_p / \|U g_p\|$  has an error

$$\|\tilde{F} - f\|^2 \leq \|f - f_\Lambda\|^2 + \frac{\bar{B}_\Lambda (\lambda + 2)^2 |\Lambda| \left( \|U\|_s \|f - f_\Lambda\| + \sigma \sqrt{2 \log_e P} \right)^2}{(\min_{p \in \Lambda} \|U g_p\|^2) A_\Lambda^2 (1 - \text{ERC}(\Lambda))^2}, \quad (13.42)$$

with a probability that tends to 1 as  $P$  increases.

**Proof.** The proof is derived from the proof of Theorem 12.15. To compute a solution with a support in  $\Lambda$ , we also consider a solution  $\tilde{\alpha}_\Lambda$  of the  $\mathbf{l}^1$  Lagrangian minimization over  $\Lambda$ :

$$\tilde{\alpha}_\Lambda = \underset{\alpha_\Lambda \in \mathbb{C}^{|\Lambda|}}{\operatorname{argmin}} \frac{1}{2} \|Y - \Phi_{U\Lambda}^* \alpha_\Lambda\|^2 + T \|\alpha_\Lambda\|_1,$$

and  $\tilde{\alpha}$  is defined by  $\tilde{\alpha}[p] = \tilde{\alpha}_\Lambda[p]$  for  $p \in \Lambda$  and  $\tilde{\alpha}[p] = 0$  for  $p \in \Lambda^c$ . Let  $h$  be defined by

$$Th = \Phi_U(Y - \Phi_U^* \tilde{\alpha}_\Lambda) = \Phi_U(Y - \Phi_{U\Lambda}^* \tilde{\alpha}_\Lambda). \quad (13.43)$$

To prove that  $\tilde{\alpha}$  is a solution of the  $\mathbf{l}^1$  Lagrangian pursuit (13.40), according to Theorem 12.8, we must verify that  $\|h_{\Lambda^c}\|_\infty \leq 1$ . Like in (12.150), we prove that

$$\|h_{\Lambda^c}\|_\infty \leq T^{-1} \max_{q \in \Lambda^c} |\langle \phi_q, Y - Y_\Lambda \rangle| + \text{ERC}(\Lambda), \quad (13.44)$$

where  $Y_\Lambda = P_{V_\Lambda} Y$  is the orthogonal projection of  $Y$  on the space  $V_\Lambda$  generated by  $\{U g_p / \|U g_p\|\}_{p \in \Lambda}$ . Since  $Y = U f + W$ ,

$$\langle Y - Y_\Lambda, \phi_q \rangle = \langle U f - P_{V_\Lambda} U f, \phi_q \rangle + \langle W - P_{V_\Lambda} W, \phi_q \rangle.$$

Since there are  $P$  vectors in the dictionary, and  $W$  is a white noise of variance  $\sigma^2$ ,

$$\max_{q \in \Gamma} |\langle W - P_{V_\Lambda} W, \phi_q \rangle| \leq \sigma \sqrt{2 \log_e P},$$

with a probability that tends to 1 as  $P$  increases. It results that

$$\max_{q \in \Gamma} |\langle Y - Y_\Lambda, \phi_q \rangle| \leq \|U f - P_{V_\Lambda} U f\| + \sigma \sqrt{2 \log_e P}. \quad (13.45)$$

Since  $Uf_\Lambda \in \mathbf{V}_\Lambda$ ,

$$\|Uf - P_{\mathbf{V}_\Lambda} Uf\| \leq \|Uf - Uf_\Lambda\| \leq \|U\|_s \|f - f_\Lambda\|.$$

Equation (13.44) together with (13.45) implies that

$$\|h_{\Lambda^c}\|_\infty \leq \frac{\|U\|_s \|f - f_\Lambda\| + \sigma \sqrt{2 \log_e P}}{T} + \text{ERC}(\Lambda). \quad (13.46)$$

If

$$T = \frac{\lambda (\|U\|_s \|f - f_\Lambda\| + \sigma \sqrt{2 \log_e P})}{1 - \text{ERC}(\Lambda)} \text{ with } \lambda > 1, \quad (13.47)$$

then  $\|h_{\Lambda^c}\|_\infty < 1$  and Theorem 12.8 prove that  $\tilde{a}$  is indeed an  $\mathbf{l}^1$  Lagrangian pursuit solution of (13.40). The same argument as in the proof of Theorem 13.5 shows that this solution is unique.

Since  $\tilde{\Lambda} \subset \Lambda$ , the error bound (13.42) can be computed from the coefficients restricted to  $\Lambda$ . Similar to (12.146), we prove that the coefficients  $\tilde{a}_\Lambda$  of  $\tilde{Y}$  satisfy

$$\tilde{a}_\Lambda = \Phi_{U\Lambda}^{*+} Y - T(\Phi_{U\Lambda} \Phi_{U\Lambda}^*)^{-1} h_\Lambda. \quad (13.48)$$

Writing  $f_\Lambda = \sum_{p \in \Lambda} a_\Lambda[p] g_p / \|Ug_p\|$ , we get

$$Y = \sum_{p \in \Lambda} a_\Lambda[p] \frac{Ug_p}{\|Ug_p\|} + U(f - f_\Lambda) + W.$$

It results that

$$\Phi_{U\Lambda}^{*+} Y = a_\Lambda + \Phi_{U\Lambda}^{*+} U(f - f_\Lambda) + \Phi_{U\Lambda}^{*+} W.$$

We derive from (13.48) that

$$\begin{aligned} \|\tilde{a}_\Lambda - a_\Lambda\| &\leq \|\Phi_{U\Lambda}^{*+} U(f - f_\Lambda)\| + \|\Phi_{U\Lambda}^{*+} W\| + \|T(\Phi_{U\Lambda} \Phi_{U\Lambda}^*)^{-1} h_\Lambda\| \\ &\leq \frac{1}{\sqrt{A_\Lambda}} (\|P_{\mathbf{V}_\Lambda} W\| + \|U\|_s \|f - f_\Lambda\|) + \frac{T \sqrt{|\Lambda|}}{A_\Lambda}, \end{aligned}$$

where we used that  $\|h_\Lambda\| \leq \sqrt{|\Lambda|}$ . In a dictionary of size  $P$ , Lemma 12.1 proves that the energy of the noise projected in any space generated by  $|\Lambda|$  dictionary vectors satisfies

$$\|W_\Lambda\| \leq 2\sigma \sqrt{2|\Lambda| \log_e P} \quad (13.49)$$

with a probability that tends to 1 as  $P$  increases. Inserting the value of  $T$  in (13.47) gives

$$\begin{aligned} \|\tilde{a}_\Lambda - a_\Lambda\| &\leq \frac{2\sigma \sqrt{|\Lambda| 2 \log_e P} + \|U\|_s \|f - f_\Lambda\|}{\sqrt{A_\Lambda}} \\ &\quad + \frac{\sqrt{|\Lambda|} \lambda (\|U\|_s \|f - f_\Lambda\| + \sigma \sqrt{2 \log_e P})}{A_\Lambda (1 - \text{ERC}(\Lambda))} \\ \|\tilde{a}_\Lambda - a_\Lambda\| &\leq \frac{(\lambda + 2) \sqrt{|\Lambda|} (\|U\|_s \|f - f_\Lambda\| + \sigma \sqrt{2 \log_e P})}{A_\Lambda (1 - \text{ERC}(\Lambda))}. \end{aligned} \quad (13.50)$$

Since

$$\|\tilde{F} - f_\Lambda\| = \left\| \sum_{p \in \Lambda} (\tilde{a}_\Lambda[p] - a_\Lambda[p]) \frac{g_p}{\|Ug_p\|} \right\| \leq \frac{\sqrt{\bar{B}_\Lambda} \|\tilde{a}_\Lambda - a_\Lambda\|}{\min_{p \in \Lambda} \|Ug_p\|},$$

it results that

$$\|\tilde{F} - f_\Lambda\| \leq \frac{\sqrt{\bar{B}_\Lambda} (\lambda + 2) \sqrt{|\Lambda|} (\|U\|_S \|f - f_\Lambda\| + \sigma \sqrt{2 \log_e P})}{\min_{p \in \Lambda} \|Ug_p\| A_\Lambda (1 - \text{ERC}(\Lambda))}. \quad (13.51)$$

Since  $\tilde{F} - f_\Lambda \in \mathbf{V}_\Lambda$ , we have  $\|\tilde{F} - f\|^2 = \|\tilde{F} - f_\Lambda\|^2 + \|f - f_\Lambda\|^2$ . Inserting (13.51) proves (13.42). ■

The result of this theorem is conservative but shows the main sources of instabilities of sparse super-resolution algorithms. It proves that part of the approximation support of  $\Lambda$  can be recovered by approximating the noisy data  $Y$  if  $\text{ERC}(\Lambda) < 1$ . The multiplier  $T$  behaves as a soft threshold and must be above the noise, thus the term  $\sigma \sqrt{2 \log_e P}$ . The normalization factors  $\|Ug_p\|$  cannot be too small to avoid amplifying the noise.

### 13.3.2 Sparse Spike Deconvolution

Seismic sparse spike deconvolution is probably the first super-resolution algorithm used in industry for seismic exploration. This problem perfectly illustrates the main super-resolution ideas and difficulties. Mineral and oil seismic explorations measure underground reflectivity by sending pressure waves. The reflected pressure waves are recorded at the surface as a function of time and spatial position. Seismic inversion includes different steps such as migration and stacking to invert the wave propagation equation. After these inversions, at a given position of the surface, the resulting seismic data  $Y$  are approximately related to the underground reflectivity  $f$  through a convolution equation  $Y[n] = u * f[n] + W[n]$  where  $n$  is a time variable that is related to depth.

The convolution kernel  $u$  is called a *seismic wavelet* in geophysics, which is the origin of this name chosen by the geophysicist Morlet [276]. It depends on the pressure wave sent underground but also on the subsequent inversion operations. These seismic wavelets are calibrated from reflectivity and seismic data measured along wells that are drilled in the ground. The singular basis  $\mathcal{B}_S$  that diagonalizes  $U^* U$  is the Fourier basis, and the singular values are given by the transfer function  $|\hat{u}[k]|^2$ , which is a band-pass filter. The noise  $W$  includes not only random measurement noise but also model errors, for example, neglecting multiple reflections in the wave propagation equation.

In a simple model, the underground impedance is approximated by piecewise constant functions corresponding to layers of homogeneous rocks. The reflectivity  $f$  is then a set of Diracs corresponding to the difference of impedance at the interfaces between different geophysical layers:

$$f[n] = \sum_{p \in \Lambda} a[p] \delta[n - p \Delta].$$

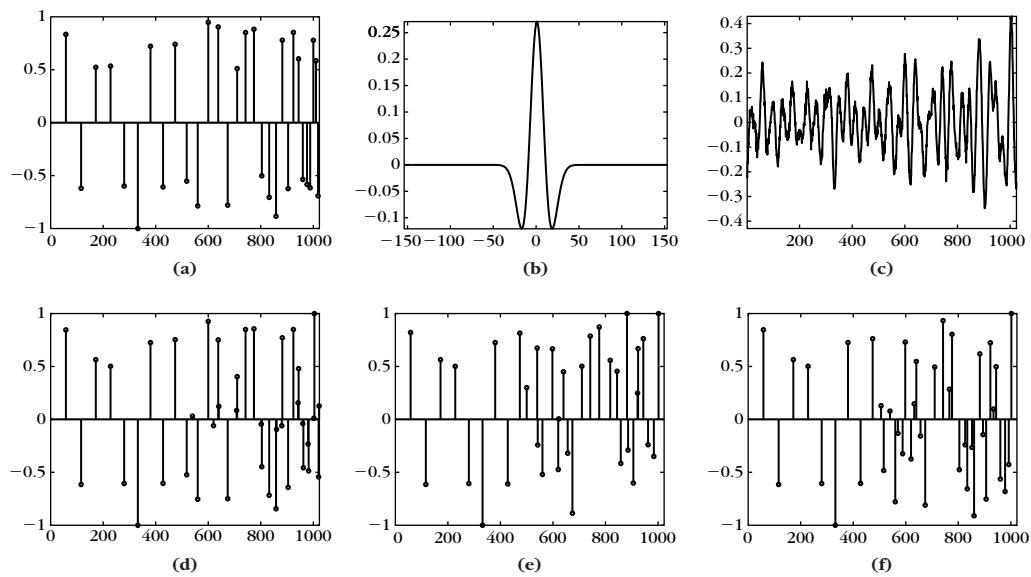


FIGURE 13.3

- (a) Sparse spikes signal  $f$ ; the distance between spikes decreases from left to right, from  $30\Delta$  to  $5\Delta$ .  
 (b) Seismic wavelet  $u$  (zoom). (c) Measured seismic data  $y$ . (d) Sparse spikes deconvolution with an  $\ell^1$  pursuit, (e) a matching pursuit, and (f) an orthogonal matching pursuit.

Thus, it has a sparse representation in a Dirac orthonormal dictionary

$$\mathcal{D} = \{g_p[n] = \delta[n - p\Delta]\}_{0 \leq p < P},$$

translated on a grid of interval  $\Delta = N/P$  that defines the resolution of the sparse spike deconvolution. This dictionary fully satisfies the spread spectrum hypothesis since each Dirac has a flat spectrum in a Fourier basis.

Figure 13.3 shows a synthetic example of a sparse spike signal  $f$  of size  $N = 1024$  and the resulting noisy observation  $Y[n]$ . The seismic wavelet  $u[n]$  is the second derivative of a Gaussian. Without noise, the dimension  $Q = \dim(\text{Im } U)$  of the observation space is the number of frequencies such that  $|\hat{u}[k]| \neq 0$ . In presence of noise, a linear or thresholding estimator can recover an estimation of  $f$  in a space of lower dimension  $Q_0$  for which the amplified noise  $|\hat{u}[k]|^{-1}\sigma$  is not above the maximum amplitude of signal coefficients. In this case,  $Q_0 \approx 100$ .

If  $f$  only includes Diracs that are far away so that  $u[n - p\Delta]$  barely overlaps with  $u[n - q\Delta]$  for  $(p, q) \in \Lambda^2$ , then the locations  $p\Delta$  can be detected with a “match filtering,” which is equivalent to a matching pursuit in the dictionary  $\mathcal{D}_U$ . An accurate estimation  $\tilde{a}[p]$  of each  $a[p]$  is then derived, which yields an estimation  $\tilde{F}$  of  $f$ . This is a super-resolution estimation since the low and high frequencies of  $f$  are restored although they were fully removed by the band-pass filter  $U$ . The main difficulty of sparse spike estimation is to recover thin geophysical layers corresponding to closely located Diracs, producing overlapping seismic wavelets  $u[n - p\Delta]$ .

To identify these closely located *spikes*, in 1973 Clearbout and Muir [166] proposed to use an  $\mathbf{1}^1$  minimization in the Dirac basis. Since  $Ug_p[n] = u \star \delta[n - p\Delta] = u[n - p\Delta]$ , the transformed dictionary is a family of translated seismic wavelets:

$$\mathcal{D}_U = \left\{ \frac{u[n - p\Delta]}{\|u\|} \right\}_{0 \leq p < P}. \quad (13.52)$$

In 1986, Santosa and Symes [424] implemented this idea with an  $\mathbf{1}^1$  relaxed minimization, which is a Lagrangian basis pursuit (13.35) of the seismic signal  $Y[n]$  in the transformed dictionary of translated wavelets (13.52). It yields a sparse set of coefficients  $\tilde{a}[p]$  from which a sparse spike estimation of  $f$  is derived according to (13.36):

$$\tilde{F} = \sum_{p \in \Gamma} \frac{\tilde{a}[p]}{\|u\|} \delta[n - p\Delta]. \quad (13.53)$$

The resolution  $\Delta$  is set relative to the scale  $s$  of the wavelet. Increasing  $\Delta$  reduces the maximum resolution of the sparse spike, but Section 12.5 shows that when  $\Delta$  is too small, computations become unstable. Close wavelets  $u[n - p\Delta]$  and  $u[n - (p + 1)\Delta]$  become too similar to choose between them, and the Lagrangian pursuit algorithm converges more slowly.

If  $\text{ERC}(\Lambda) < 1$  and the noise is small, Theorems 12.13 and 12.15 prove that an orthogonal matching pursuit as well as an  $\mathbf{1}^1$  Lagrangian pursuit recover the spikes in the support  $\Lambda$ . Theorem 12.12 shows that  $\text{ERC}(\Lambda)$  decreases as the distance between spikes in  $\Lambda$  increases [232]. If  $u$  is the second-order derivative of a Gaussian

$$u[n] = \lambda (1 - s^{-2} n^2) e^{-s^{-2} n^2/2},$$

then a numerical calculation shows that  $\text{ERC}(\Lambda) < 1$  if the distance between any two consecutive spikes in  $p$  and  $q$  satisfies  $|p\Delta - q\Delta| \geq 5s$ .

The sparse spike deconvolutions in Figure 13.3 are calculated with a wavelet scaled by  $s = 10$  that corresponds to  $Q_0 \approx 100$  frequency measurements over a signal of size  $N = 1024$ . The transformed dictionary includes  $P = 512$  waveforms separated by  $\Delta = 2$ . If all spikes have a distance of  $25\Delta$ , then  $\text{ERC}(\Lambda) < 1$ , and they can thus be recovered by an  $\mathbf{1}^1$  Lagrangian pursuit or a matching pursuit. Figure 13.3 shows a sparse spike signal  $f$  and the measurement  $Y$  with noise. The  $\mathbf{1}^1$  pursuit and matching pursuits are computed with a backprojection to restore the amplitude of spikes. The spikes have a spacing that decreases nonlinearly from  $30\Delta$  to  $5\Delta$  from left to right. The three algorithms recover all spikes up to a spacing of  $22\Delta$  (middle of the figure), whereas  $\text{ERC}(\Lambda) < 1$  is only for a spacing of  $25\Delta$  or larger. The three algorithms begin to fail below  $22\Delta$  but the  $\mathbf{1}^1$  pursuit yields a higher SNR. In this example,  $\|\tilde{F} - f\|/\|f\|$  is 0.45 for an  $\mathbf{1}^1$  pursuit, and 0.9 for a matching pursuit and for an orthogonal matching pursuit. A matching pursuit and an orthogonal matching pursuit selects first the same “coherent structures” corresponding to the spikes on the left that have a distance larger than  $22\Delta$ .

Although slightly pessimistic, the  $\text{ERC}(\Lambda) < 1$  gives a good prediction for the recovery of signal components, but the  $\mathbf{l}^1$  pursuit can still recover information below this limit. It improves the result of matching pursuit algorithms at a computational cost. Donoho et al. [225–227] prove that the  $\mathbf{l}^1$  Lagrangian pursuit has the ability to recover closer spikes if they are not too numerous, but computations become unstable when their distance is reduced.

### 13.3.3 Recovery of Missing Data

Applications of super-resolution are studied for image zooming, Radon transform inversion in medical imaging, and image restoration with missing pixels. In missing data problems, partial observations are specified by a set of noisy measurements

$$Y[q] = Uf[q] + W[q] = \langle f, u_q \rangle + W[q] \quad \text{with } q \in \Omega \quad \text{and } |\Omega| = Q < N.$$

The family  $\{u_q\}_{q \in \Omega}$  is a basis of a subspace  $\mathbf{V}$  of dimension  $Q$ . A linear estimation of the orthogonal projection  $P_{\mathbf{V}}f$  of  $f$  in  $\mathbf{V}$  can be computed with a dual basis:

$$\tilde{F}_l[n] = \sum_{q \in \Omega} Y[q] \tilde{u}_q[n] = P_{\mathbf{V}}f[n] + \sum_{q \in \Omega} W[q] \tilde{u}_q[n] \in \mathbf{V}. \quad (13.54)$$

For super-resolution, the dictionary  $\mathcal{D}$  must include vectors  $g_p$  with a spread spectrum. If  $\{u_q\}_{q \in \Omega}$  is an orthonormal family, then  $U^*U = P_{\mathbf{V}}$ . So  $\mathbf{V}$  is an eigenspace with singular value 1 and  $\mathbf{V}^\perp$  is the other eigenspace with singular value 0. Each  $g_p$  should thus have orthogonal projections in  $\mathbf{V}$  and in  $\mathbf{V}^\perp$  that are relatively large.

#### *Image Inpainting*

Inpainting is an example of missing data recovery for damaged images where pixel values are available in a known region  $\Omega$ , and missing in its complement  $\Omega^c$ :

$$Uf[q] = f[q] \quad \text{for } q \in \Omega \quad \text{with } |\Omega| = Q < N. \quad (13.55)$$

Elad et al. [244] as well as Fadili, Starck, and Murtagh [250] studied inpainting solutions with the sparse Lagrangian  $\mathbf{l}^1$  pursuit minimization. This dictionary must include vectors with restrictions to  $\Omega$  and  $\Omega^c$  that have sufficiently large energy, while providing a sparse image representation. Figure 13.4 gives an example similar to [250], where the grid of the bird cage is removed from  $\Omega$ . The interpolation  $\tilde{F}$  in Figure 13.4(c) is computed with a translation-invariant dyadic wavelet dictionary  $\mathcal{D}$ . Fine-scale wavelets with a support nearly inside  $\Omega^c$  must be removed from the dictionary to compute a normalized Lagrangian estimation (13.35). To simplify computations, they are kept in the dictionary and the super-resolution estimation  $\tilde{F}$  is calculated with a nonnormalized  $\mathbf{l}^1$  pursuit (13.39).

Figure 13.5(c) shows a second inpainting example in a dictionary  $\mathcal{D}$  that is the union of a translation-invariant wavelet dictionary and a tight frame of local cosine vectors with a redundancy factor of 4. The restriction of the original image to  $\Omega$  is shown in Figure 13.5(a). Local cosine vectors have a support larger than the size of the holes and satisfy the noise-amplification conditions, but this is not the

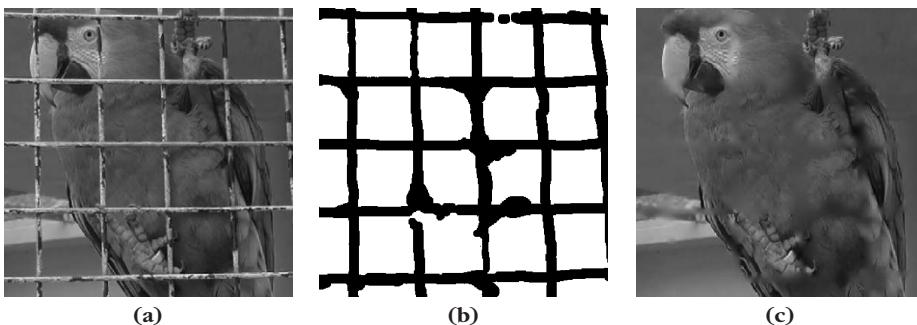


FIGURE 13.4

(a) Original image  $f$ . (b) Available pixels in  $\Omega$  are shown in white. (c) Estimation  $\tilde{F}$  computed with an  $L^1$  Lagrangian pursuit in a translation-invariant wavelet dictionary.

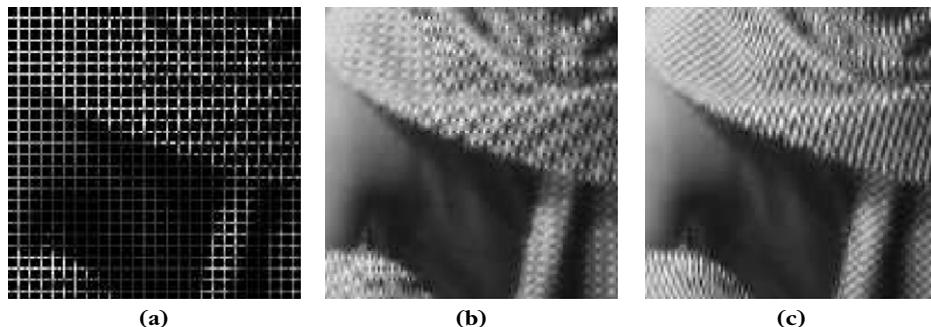


FIGURE 13.5

(a) Observed image restricted to  $\Omega$ . (b) Linear Tikhonov interpolation (SNR = 16.2db).

(c) Interpolation with an  $L^1$  Lagrangian pursuit in a wavelet and local cosine dictionary (SNR = 18.2 db).

case of fine-scale wavelets inside  $\Omega_c$ . The estimation  $\tilde{F}$  is thus also calculated with a nonnormalized  $L^1$  pursuit. The resulting SNR calculated relative to the original image without a hole is 18.2 db. It improves by 2 db the SNR obtained with a linear Tikhonov regularization in Figure 13.4(b).

### **Linear Tikhonov Regularization**

As any inverse problem, missing data can be computed with a linear inverse estimator, studied in Section 13.1, which recovers

$$\tilde{F} = \underset{h \in \mathbb{R}^N}{\operatorname{argmin}} \| \Phi h \|^2 \quad \text{with} \quad \| Y - U \tilde{F} \|^2 \leq \varepsilon, \quad (13.56)$$

where  $\Phi$  is a regularization operator. A Tikhonov minimization regularizes the solution with a gradient operator  $\Phi h = \vec{\nabla} h$ , and  $\|\vec{\nabla} h\|^2$  is then a Sobolev norm that tends

to recover a uniformly regular image. The Lagrangian formula of this convex problem gives

$$\tilde{F} = \underset{h \in \mathbb{R}^N}{\operatorname{argmin}} \frac{1}{2} \|Uh - y\|^2 + T^2 \|\Phi h\|^2 \quad \text{with} \quad \Phi h = \tilde{\nabla} \Phi, \quad (13.57)$$

where  $T$  is adjusted as a function of  $\varepsilon$ . The solution computed in (13.7) is

$$\tilde{F} = (U^* U + T^2 \Phi^* \Phi)^{-1} U^* Y.$$

It is calculated by inverting the symmetric operator  $L = U^* U + T^2 \Phi^* \Phi$  with the conjugate-gradient algorithm or the Richardson gradient descent (see Section 5.1.3). As explained in Section 13.1, these linear estimators compute the solution in a space of dimension  $Q$  and thus do not perform any super-resolution.

For an inpainting problem, where image values are known in  $\Omega$ , one can verify that the resulting solution satisfies  $\Delta \tilde{F}[n] = 0$  for  $n \in \Omega^c$  with boundary conditions specified by image values in  $\Omega$ . The Tikhonov regularization thus diffuses the image values in  $\Omega^c$  with an isotropic heat-diffusion equation. If the noise is neglected and thus  $\varepsilon = 0$ , then image values in  $\Omega$  are not regularized and the boundary values of  $\Omega^c$  are the values of  $f$  at the boundary of  $\Omega$ . Figure 13.4(b) is an example of inpainting computed with a Tikhonov regularization. The SNR computed relative to the original image (without holes) is 16.2 db.

Total variation regularizations often do not outperform a linear Tikhonov regularization for image inpainting. Masnou and Morel [372] improved total variation regularization algorithms by also minimizing the  $\mathbf{1}^1$  norm of the curvature of level sets. The solution is obtained with a nonlinear partial differential equation, which performs an anisotropic diffusion of the image values in the holes. Other partial differential equations that impose more geometric regularity have also been studied [94, 110, 154, 249, 465]. The algorithms give good results but can have instabilities when the domain  $\Omega^c$  is nonconvex and complex, as in Figure 13.4.

### ***Image Scaling and Deinterlacing***

Image and video screens often have more pixels than the images that are displayed. To fit the whole screen, images must be scaled, while restoring as many details as possible and minimizing artifacts. This is a major challenge for videos, in particular for high-definition television (HDTV). Indeed, most current television images are in an interlaced standard-definition television (SDTV) format (PAL or NTSC). Interlacing means that one image out of two carries only the even rows and the next one carries only the odd rows, which is adapted to CRT television displays. Flat HDTV screens simultaneously display the even and odd rows of each image. The number of rows and columns of high-definition images is also at least twice as large as SDTV formats [358]. Thus, to display SDTV interlaced images on HDTV screens requires us to increase the number of pixels by at least 8 for each image. Moreover, recent screens display 120 or 100 images per second as opposed to 60 or 50, which also requires us to double the number of images in time. In such scaling applications, the image is known over a coarse regular spatial or space-time

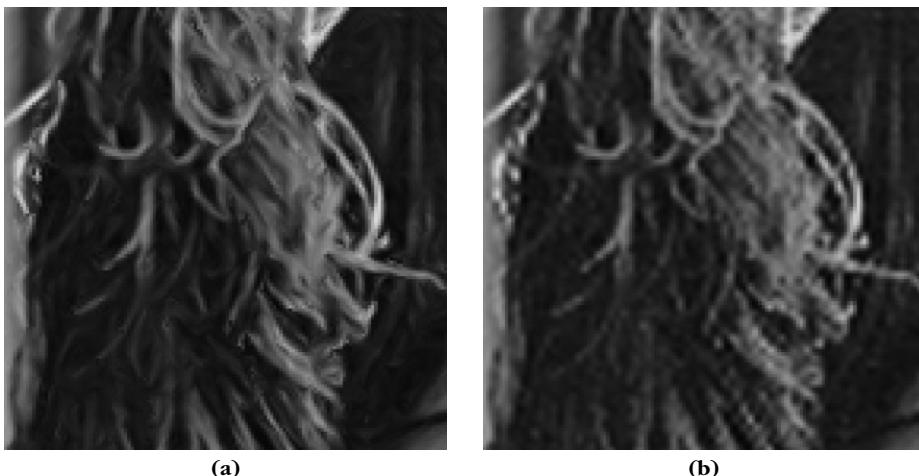
grid  $\Omega$  having  $Q$  pixels, and image values must be interpolated on a finer grid with  $N$  pixels. For SDTV to HDTV conversion,  $N \geq 16Q$ . The noise  $W[n]$  is often complex because it is dominated by compression artifacts as opposed to camera noise.

If the noise can be neglected, then known image values are preserved and a linear interpolation computes

$$\tilde{F}[n] = \sum_{m \in \Omega} Y[m] \theta[n - m].$$

The interpolation kernel satisfies  $\theta[n - m] = \delta[n - m]$  for any  $(n, m) \in \Omega^2$ , so that  $\tilde{F}[m] = Y[m]$  for  $m \in \Omega$ . A quadratic minimization (13.56) with  $\varepsilon = 0$  implements such an interpolation. For a Tikhonov regularization with  $\Phi = \nabla$ , the kernel  $\theta$  computes a linear interpolation. Cubic spline interpolation kernels  $\theta$  are most often used in image processing and correspond to a third-order differential operator  $\Phi$  [458]. Figure 13.6(a) shows an example of a linear scaling by four along the image rows and columns with a cubic spline interpolation. The image is blurred and oscillations appear along directional structures such as contours.

Instead of interpolating all pixels with a predefined linear kernel, adaptive directional interpolations adapt the interpolation kernel for each missing pixel, depending on the observed image regularity. If the image regularity is not isotropic, an elongated kernel is used to perform the interpolation along a direction where the image is locally the most regular. Along edges, the interpolation kernel is typically elongated in the direction of the edge tangent. Such techniques are used in industry



**FIGURE 13.6**

- (a) Separable linear interpolation by a factor of four along rows and columns with cubic splines.  
 (b) Nonlinear directional interpolation with the same factors.

for video deinterlacing and scaling. Finding locally the best interpolation directions and optimizing the shapes of the kernels are difficult problems, most often solved with ad hoc algorithms. Yet, good results are obtained even on complex images, as shown in Figure 13.6(b).

Nonlinear scaling can be computed with a sparse super-resolution estimator in a dictionary  $\mathcal{D}$  with vectors  $g_p$  that intersect both  $\Omega$  and  $\Omega^c$ , while providing a sparse image representation. Such dictionaries must include elongated directional waveforms of large support such as curvelets or bandlets in order to take advantage of the directional image regularity [290]. A Lagrangian pursuit estimator can then be interpreted as an adaptive image interpolation. The interpolation directions and the size of the scale of the interpolation kernels correspond to the direction and size of the reconstructing dictionary waveforms  $Ug_p$  for  $p \in \tilde{\Lambda}$ , computed to recover a sparse representation of the observed image  $Y$  on  $\Omega$ .

The ability to achieve some super-resolution depends on the geometry of  $f$  relative to  $\Omega$ . If  $\Omega$  is a square subsampled grid, then one can verify that no super-resolution is possible along a strictly horizontal or vertical edge. When edge angles are very close to horizontal and vertical, some super-resolution is possible but limited by instabilities. Similarly, when video images do not move, constant values in time provide no information to increase the spatial resolution.

### ***Tomography Inversion***

A two-dimensional X-ray tomographic imaging system measures the Radon transform of body slices  $\bar{f}(x)$  along a limited number of angles  $\{\theta_1, \dots, \theta_L\}$  in order to reduce the exposition time of patients. The Radon transform of  $\bar{f}(x)$  along a ray parameterized by  $x_1 \cos \theta + x_2 \sin \theta = \tau$  is

$$\forall \tau \in \mathbb{R}, \quad \overline{U}\bar{f}(\theta, \tau) = p_\theta(\tau) = \int \int \bar{f}(x) \delta(x_1 \cos \theta + x_2 \sin \theta - \tau) dx.$$

The Fourier slice theorem (2.10) proves that

$$\hat{p}_\theta(\omega) = \widehat{\bar{f}}(\omega \cos \theta, \omega \sin \theta). \quad (13.58)$$

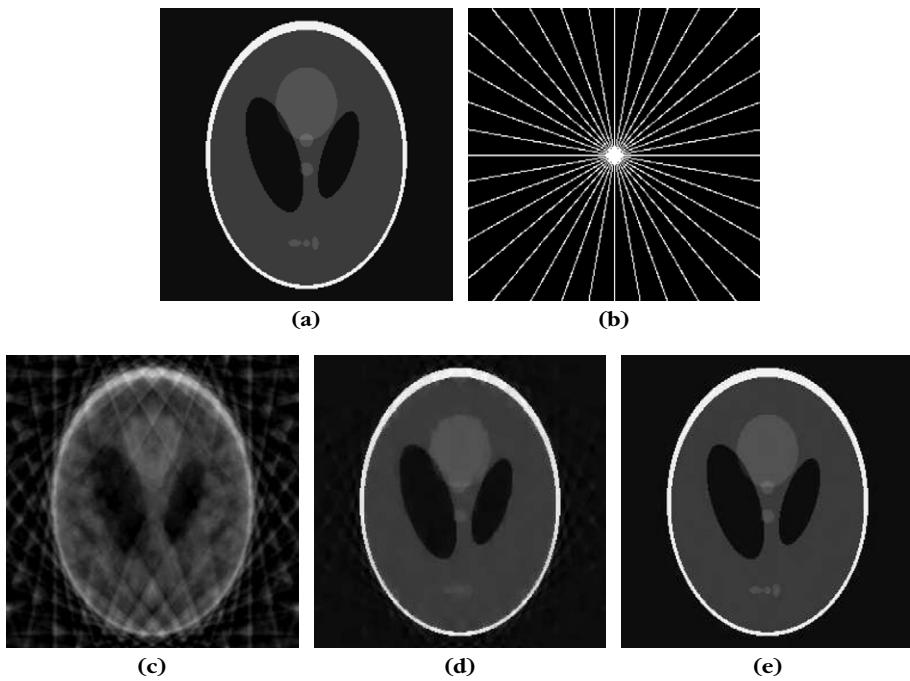
A linear orthogonal projection of  $f$  can be computed from these observations over the space  $\mathbf{V}$  generated by Fourier vectors at the available frequencies. It is obtained with the backprojection theorem (2.11) as a partial sum:

$$P_{\mathbf{V}}\bar{f}(x) = \frac{1}{L} \sum_{\ell=1}^L p_{\theta_\ell} * h(x_1 \cos \theta_\ell + x_2 \sin \theta_\ell) \quad \text{with} \quad \hat{h}(\xi) = |\xi|. \quad (13.59)$$

Tomographic measurement systems provide discrete measurements with noise

$$Y[l, q] = Uf[l, q] + W[l, q] = p_{\theta_l}[q] + W[l, q] \quad \text{for } 1 \leq l \leq L,$$

from which we want to recover an estimation  $\tilde{F}$  of a high-resolution image  $f \in \mathbb{C}^N$ . According to the Fourier slice theorem, the Fourier transform of  $Y[l, q]$  along  $q$  gives

**FIGURE 13.7**

- (a) Original phantom image.
- (b) Frequency plane showing in white the frequency rays in  $\Omega$ .
- (c) Reconstruction with a linear orthogonal projection computed with a backprojection.
- (d) Lagrangian pursuit estimation in a Haar translation-invariant wavelet dictionary.
- (e) Inversion with a total variation regularization.

noisy measurements of  $\hat{f}[m]$  along rays  $m \in \Omega$ , as illustrated in Figure 13.7. Inverse tomography is thus a missing Fourier data recovery problem. The Fourier basis is a basis of singular vectors that diagonalize  $U^*U$  with singular values equal to 1 on  $\Omega$ . A super-resolution recovery requires using a dictionary of spread spectrum vectors, which have a Fourier transform as delocalized as possible.

Medical image models are often piecewise regular, and thus have sparse wavelet approximations. Simple piecewise constant phantom images, as shown in Figure 13.7, are often used to evaluate inversion algorithms. In this case, the most sparse wavelet representation is obtained with Haar wavelets, which are discontinuous and thus have a Fourier transform that is quite spread out. Figure 13.7(c) shows the image reconstructed with a linear backprojection. The resulting image is highly oscillatory because of the missing frequencies. Figure 13.7(d) shows a nonnormalized Lagrangian  $\ell^1$  pursuit estimation (13.39) with a Haar translation-invariant wavelet dictionary. It recovers a much more precise and sharp piecewise constant estimation.

### Total Variation Regularization

Instead of supposing that the solution has a sparse synthesis in a dictionary as we did in this section, a sparse analysis assumes that a particular linear image transform  $\Phi f$  is sparse. As explained in Section 12.4.4, an estimator  $\tilde{F}$  of  $f$  can be defined as

$$\tilde{F} = \underset{h \in \mathbb{R}^N}{\operatorname{argmin}} \|\Phi h\|_1 \quad \text{with} \quad \|Y - U \tilde{F}\|^2 \leq \varepsilon. \quad (13.60)$$

For images, Rudin, Osher, and Fatemi [420] introduced this approach with  $\Phi f = \vec{\nabla} f$ , in which case  $\|\Phi f\|_1 = \|\vec{\nabla} f\|_1 = \|f\|_V$  is the total image variation. The Lagrangian formulation then computes

$$\tilde{F} = \underset{h \in \mathbb{R}^N}{\operatorname{argmin}} \frac{1}{2} \|y - Uh\|^2 + T \|\Phi h\|_1 \quad \text{with} \quad \Phi h = \vec{\nabla} h. \quad (13.61)$$

Section 12.4.4 describes an iterative algorithm solving this minimization.

This minimization (13.60) looks similar to the Tikhonov regularization (13.56), where the  $\mathbf{l}^2$  norm  $\|\Phi h\|$  is replaced by a  $\mathbf{l}^1$  norm  $\|\Phi h\|_1$ , but the estimator properties are completely different. A  $\mathbf{l}^2$  norm is minimized by maintaining small-amplitude coefficients distributed uniformly, which yields a uniformly regular signal with a Tikhonov regularization computed with  $\Phi = \vec{\nabla}$ . As explained in Section 12.4.1, the minimization of a  $\mathbf{l}^1$  norm tends to produce many zero- or small-amplitude coefficients and few large-amplitude ones. For  $\Phi = \vec{\nabla}$ , the coarea theorem (2.9) proves that the total image variation  $\|\vec{\nabla} f\|_1 = \|f\|_V$  is the average length of the level sets of  $f$ .

The phantom image of Figure 13.7(a) is ideal for total variation estimation. Indeed, the gradient is zero everywhere outside the edges of the image objects, which have a length that is not too large. Figure 13.7(e) is obtained by minimizing the Lagrangian formulation (13.61) of the total variation minimization with the Radon transform operator  $U$ . Without noise, this total variation regularization performs an almost exact recovery of the input image  $f$ , which is not the case of the Lagrangian pursuit with Haar wavelets. Indeed, the gradient field is more sparse than with a multiscale Haar wavelet transform. Haar wavelets do not restore the boundaries of the image phantoms as precisely as a total variation regularization, which minimizes the length of restored contours.

Real medical images are not piecewise constant and include much more complex structures. Total variation estimations are therefore not as spectacular on real images. They have a tendency to remove textures and oscillatory structures by producing flat image areas, which can reduce the SNR.

---

## 13.4 COMPRESSIVE SENSING

Super-resolution is not always possible and often unstable for usual measurement operators  $U$ . Candès and Tao [138, 139] as well as Donoho [217, 218] observed that sparse super-resolution becomes stable for all sufficiently sparse signals when

$U$  is a random measurement operator that computes random linear combinations of all signal values. This remarkable result opens the door to compressive sensing strategies, where randomized linear measurements can recover a higher-resolution approximation of signals that have a sparse approximation in some dictionary. It also gives a conceptual probabilistic framework where random mixing appears as an efficient information acquisition strategy for structured information.

### 13.4.1 Incoherence with Random Measurements

As opposed to previously studied inverse problems, in compressive sensing, we have the luxury to design the measurement operator that is not imposed. Random sensing operators create highly incoherent transformed dictionaries where sufficiently sparse signals have a stable recovery.

#### **Compressive Sensing Acquisition and Recovery**

The compressive sensing acquisition of an analog signal  $\bar{f}(x)$  is implemented with a continuous sensing operator  $\bar{U}$  that provides  $Q$  measurements

$$Y[q] = \bar{U}\bar{f}[q] + W[q] = \langle \bar{u}_q, \bar{f} \rangle + W[q]. \quad (13.62)$$

In analog compressive sensing, the hardware device outputs randomized analog measurements with transfer functions  $\bar{u}_q(x)$  that are realizations of a random process. This acquisition is modeled with a stable high-resolution analog-to-digital converter  $\bar{\Phi}_s$  followed by a discrete operator  $U$ , which outputs  $Q$  random combinations of these high-resolution measurements:  $\bar{U} = U\bar{\Phi}_s$ . Measurements can thus be rewritten as

$$Y[q] = Uf[q] + W[q] \quad \text{for } 0 \leq q < Q,$$

where  $f = \bar{\Phi}_s \bar{f} \in \mathbb{C}^N$  is a high-resolution discretization of  $\bar{f}(x)$  with  $N \gg Q$ . A super-resolution operator computes an estimation  $\tilde{F} \in \mathbb{C}^N$  of  $f$  from the vector  $Y$  of  $Q$  measurements.

Suppose that  $f$  has a sparse approximation in a dictionary  $\mathcal{D} = \{g_p\}_{p \in \Gamma}$ . According to the sparse super-resolution algorithm described in Section 13.3, an estimation of  $f \in \mathbb{C}^N$  is computed by finding a sparse approximation of  $Y$  in the transformed dictionary  $\mathcal{D}_U = \{Ug_p\}_{p \in \Gamma}$ . Suppose that  $U$  is a random matrix with coefficients that are obtained with independent random variables of the same distribution and variance  $Q^{-1}$ . For large  $N$ , the law of large numbers guarantees with high probability that  $\|Ug_p\|^2$  is close to 1. It implies with a high probability that any  $g_p \in \mathcal{D}$  has a spread singular spectrum relative to the singular vectors and singular values of  $U^*U$ . Each  $Ug_p$  is thus already nearly normalized.

Let  $\Phi f[p] = \langle f, g_p \rangle$ . The super-resolution estimator is

$$\tilde{F} = \sum_{p \in \tilde{\Lambda}} \tilde{a}[p] g_p = \Phi^* \tilde{a},$$

where  $\tilde{a}$  are sparse approximation coefficients of  $Y$  in the transformed dictionary  $\mathcal{D}_U$ . They can be computed with an  $\mathbf{l}^1$  pursuit minimization

$$\tilde{a} = \operatorname{argmin}_{a \in \mathbb{C}^P} \|a\|_1 \text{ subject to } \left\| \sum_{p \in \Gamma} a[p] U g_p - Y \right\| \leq \varepsilon, \quad (13.63)$$

or as a solution of an  $\mathbf{l}^1$  Lagrangian pursuit

$$\tilde{a} = \operatorname{argmin}_{a \in \mathbb{C}^P} \frac{1}{2} \left\| \sum_{p \in \Gamma} a[p] U g_p - Y \right\|^2 + T \|a\|_1. \quad (13.64)$$

The sparse approximation  $\tilde{a}$  can also be computed with a matching pursuit decomposition of  $Y$  in  $\mathcal{D}_U$ .

### **Restricted Isometry and Incoherence**

Section 12.5 explains that an estimation of a sparse approximation  $f_\Lambda$  is possible by decomposing  $Y$  in the transformed dictionary  $\mathcal{D}_U = \{U g_p\}_{p \in \Gamma}$  only if  $\{U g_p\}_{p \in \Gamma}$  is a frame with a frame bound ratio  $A_\Lambda/B_\Lambda$  that is not close to 0. If the vectors are normalized vectors, then  $A_\Lambda \leq 1 \leq B_\Lambda$ , and it is equivalent to impose that  $\delta_\Lambda = \max(1 - A_\Lambda, B_\Lambda - 1)$  is not too small. To get a stable recovery of all sparse signals, compressive sensing imposes a uniform bound on all sufficiently sparse sets  $\Lambda$ :

$$\delta_\Lambda \geq \delta_M(\mathcal{D}_U) > 0 \text{ if } |\Lambda| \leq M,$$

where  $\delta_M(\mathcal{D}_U)$  is called an *M-restricted isometry bound*. It results that for all  $\Lambda$  with  $|\Lambda| \leq M$ ,

$$\forall a \in \mathbb{C}^{|\Lambda|}, \quad (1 - \delta_M(\mathcal{D}_U)) \sum_{p \in \Lambda} |a[p]|^2 \leq \left\| \sum_{p \in \Lambda} a[p] U g_p \right\|^2 \leq (1 + \delta_M(\mathcal{D}_U)) \sum_{p \in \Lambda} |a[p]|^2. \quad (13.65)$$

Theorem 12.10 relates  $\delta_M(\mathcal{D}_U)$  to the dictionary mutual coherence

$$\delta_M(\mathcal{D}_U) \leq (M - 1) \mu(\mathcal{D}_U) \quad \text{with} \quad \mu(\mathcal{D}_U) = \max_{(p, q) \in \Gamma^2, p \neq q} \langle U g_p, U g_q \rangle. \quad (13.66)$$

However, the mutual coherence  $\mu(\mathcal{D}_U)$  does not provide a tight upper bound of  $\delta_M(\mathcal{D}_U)$ . Indeed, it depends on the correlation of pairs of dictionary vectors, whereas  $\delta_M(\mathcal{D}_U) > 0$  measures the stability of potentially much larger groups of  $M$  dictionary vectors. Restricted isometry bounds are stronger measures of the dictionary incoherence. A simple geometric interpretation explains why random measurement operators define incoherent dictionaries with  $\delta_M(\mathcal{D}_U) > 0$  for relatively large  $M$ .

We know that all dictionary vectors  $U g_p$  belong to the space **ImU** of dimension  $Q$ . An orthonormal basis is a stable family of  $Q$  vectors that are perfectly spread on the unit sphere of **ImU**. A family  $\{U g_p\}_{p \in \Gamma}$  is a stable Riesz basis of a subspace if these points remain well distributed on this sphere. For this result to be valid for any collection of less than  $M$  vectors in a dictionary of size  $P > Q$ , we need to distribute as uniformly as possible these  $P$  vectors on the unit sphere of **ImU**. A natural

idea is to define such vectors as  $P$  realizations of a Gaussian white noise. Indeed, a Gaussian white noise vector of dimension  $Q$  and variance  $Q^{-1}$  has a probability density that is constant on all spheres of  $\mathbb{R}^Q$ , and each realization has a norm that is close to 1 when  $Q$  is large. It is thus highly likely that these  $P$  realizations are well spread on the unit sphere. If  $\mathcal{D}$  is an orthonormal basis and  $U$  is a Gaussian random matrix with coefficients that are independent Gaussian random variables, then  $\{Ug_p\}_{p \in \Lambda}$  are also  $Q$  independent Gaussian random variables of variance  $Q^{-1}$ . Random matrix operators are thus good candidates to build a transformed dictionary that satisfies the  $M$ -restricted isometry condition (13.65) for a relatively large  $M$ . For the mathematical analysis, we shall suppose in the following that the dictionary  $\mathcal{D}$  is an orthonormal basis.

### Gaussian and Bernoulli Random Sensing Matrices

Up to now, random matrices are the only universal large-size matrices that ensure that the vectors  $\{Ug_p\}_p$  are nearly uniformly spread around the unit sphere of  $\mathbb{C}^Q$  with a high probability for any fixed orthogonal basis  $\mathcal{D}$ . This is necessary to guarantee that any collection of less than  $M$  vectors defines a Riesz basis for  $M$  relatively large. All known deterministic sensing matrices  $U$  have some regularity that prevents the set of vectors  $\{Ug_p\}_{p \in \Gamma}$  to be sufficiently well distributed.

A Gaussian random matrix  $U$  has coefficients that are realizations of independent Gaussian random variables of mean 0 and variance  $Q^{-1}$ . Its rows and columns are thus realizations of Gaussian white noise random vectors. The mutual coherence  $\mu(\mathcal{D}_U)$  of  $\mathcal{D}_U = \{Ug_p\}_p$  can be shown to be  $O(\sqrt{(\log N)/Q})$  with a high probability [230]. The inequality (13.66) derives that  $\delta_M(\mathcal{D}_U) < 1$  for  $M = O(\sqrt{Q}/(\log N))$ . Theorems 12.14 and 12.15 also prove that the approximation support of any signal with  $M = O(\sqrt{Q}/(\log N))$  nonzero coefficients is recovered by an orthogonal matching pursuit or  $\ell^1$  Lagrangian pursuit. Candès and Tao [139] as well as Donoho [217] proved that this result can be considerably improved.

**Theorem 13.6:** *Candès, Tao, Donoho.* Let  $U$  be a Gaussian matrix and  $\mathcal{D}$  be an orthonormal basis. For any  $\delta < 1$ , there exists a constant  $\beta$  such that for

$$M \leq \frac{\beta Q}{\log(N/Q)}, \quad (13.67)$$

the dictionary  $\mathcal{D}_U = U\mathcal{D}$  satisfies  $\delta_M(\mathcal{D}_U) \leq \delta$  with a probability that increases toward 1 exponentially fast with  $N$ .

**Proof.** Let  $\Phi$  be the analysis matrix associated to  $\mathcal{D}$ :  $\Phi f = \langle f, g_p \rangle$ . The matrix  $\Phi_U$  is associated to the transformed dictionary  $\mathcal{D}_U = U\mathcal{D}$  is  $\Phi_U = \Phi U^*$ . Indeed,  $\Phi_U f = \langle f, Ug_p \rangle$ . If  $U$  is a Gaussian random matrix, then its columns  $\{u_q[p]\}_{p \in \Gamma, 0 \leq q < Q}$  are realizations of  $Q$  independent Gaussian white noise. It results that  $\Phi_U = \{(g_p, u_q)\}_{p \in \Gamma, 0 \leq q < Q}$  are the decomposition coefficients of these  $Q$  white noises in an orthonormal basis that remain independent Gaussian random variables. The matrix  $\Phi_U$  is thus also a Gaussian random matrix with  $Q$  columns that are realizations of independent Gaussian white noises. Its restriction  $\Phi_\Lambda$  to  $|\Lambda|$  rows indexed in  $\Lambda$  is also a Gaussian random matrix with  $Q$  columns.

Since  $\Phi_{U\Lambda}^* \mathbf{a} = \sum_{p \in \Lambda} a[p] U g_p$ ,

$$\langle \Phi_{U\Lambda} \Phi_{U\Lambda}^* \mathbf{a}, \mathbf{a} \rangle = \left\| \sum_{p \in \Lambda} a[p] U g_p \right\|^2.$$

Proving the  $M$ -restricted isometry bound (13.65) satisfies  $\delta_M(\mathcal{D}_U) \leq \delta$  is thus equivalent to proving that the largest and smallest eigenvalues of  $\Phi_{U\Lambda} \Phi_{U\Lambda}^*$  are between  $1 - \delta$  and  $1 + \delta$  for any set of  $|\Lambda| \leq M$  rows chosen among  $N$ . This means finding an upper and lower bound of the maximum and minimum singular value of a Gaussian random matrix having  $|\Lambda|$  rows and  $Q$  columns with a probability that tends to 1 when  $N$  increases.

The matrix  $\Phi_{U\Lambda} \Phi_{U\Lambda}^*$  is an  $|\Lambda|$  by  $|\Lambda|$  symmetric matrix with coefficients that are inner products between  $\Lambda$  realizations of independent Gaussian white noises of size  $Q$  chosen among  $N$ . Upper bounds of Gaussian matrix singular values are relatively classic results [245] and recent concentration inequalities have been proved on the smallest singular value [353]. Candès and Tao [139] as well Donoho [217] prove that these concentration inequalities imply that for any  $\delta < 1$ , there exists a constant  $\beta$  such that for all  $M \leq \beta Q / (\log N/Q)$ , the maximum and minimum eigenvalues are bounded by  $1 - \delta$  and  $1 + \delta$ . ■

This theorem proves that up to a logarithmic factor,  $\delta_M < 1$  for  $M$  proportional to  $Q$ , as opposed to  $\sqrt{Q}$  as in the result obtained with the dictionary mutual coherence. However, the constant  $\beta$  derived from upper and lower bounds of Gaussian matrix singular values is very small. Gaussian random matrices are universal in the sense that this result does not depend on the orthonormal basis  $\mathcal{D}$ .

Implementing in hardware an operator  $\tilde{U}\tilde{f}$  that projects signals over independent Gaussian white noise processes can be difficult. The numerical super-resolution estimation  $\tilde{F}$  of  $f \in \mathbb{C}^N$  also requires us to store the transformed dictionary  $\{Ug_p\}_{p \in \mathbb{Z}}$  in a memory of size  $O(QN)$ , which is huge when  $Q$  and  $N$  are large. The estimation with an  $\mathbf{l}^1$  Lagrangian pursuit or with an orthogonal matching pursuit then iteratively decomposes signals in this unstructured dictionary. It requires  $QN$  additions and multiplications each time, which is again too much when  $Q$  and  $N$  are large. This suggests finding other random matrices providing  $M$ -restricted isometry bounds  $\delta_M(\mathcal{D}_U)$  similar to Gaussian random matrices, but with less memory and computational requirements.

*Bernoulli random matrices* have random entries that are independent Bernoulli random variables, thus taking values  $\pm 1$  with a probability  $1/2$ . These matrices are renormalized by  $Q^{-1/2}$  so that  $\|Ug_p\| = 1$ . Candès and Tao [139] prove that Theorem 13.6 still holds for Bernoulli matrices. The constant  $\beta$  is smaller, however, because realizations of Bernoulli processes do not have a uniform probability density over spheres of  $\mathbb{R}^Q$  as Gaussian white noises do. The vectors  $Ug_p$  of the transformed dictionary are therefore not as uniformly distributed as with a Gaussian process. Bernoulli random matrices replace all multiplications by additions and subtractions, which requires less operations and storage, but the computational complexity and memory storage remain large when  $N$  is large.

### **Random Projectors**

More structured random operators are constructed with a random subsampling of the decomposition coefficients of a signal in an orthonormal basis  $\mathcal{B} = \{u_m\}_{0 \leq m < N}$ . The operator  $U$  is then an orthogonal projector on a family of  $Q$  vectors  $\{u_q\}_{q \in \Gamma_Q}$  where the set  $\Gamma_Q$  is randomly chosen with a uniform probability distribution among all subsets of size  $Q$  in an index set of size  $N$ . Since  $U$  is an orthogonal projection, decomposing coefficients in the transformed dictionary can be written as  $\langle f, U g_p \rangle = \langle U f, g_p \rangle$ . They are computed with a fast algorithm if the projector  $U$  is implemented with a fast algorithm as well as signal decompositions in the orthonormal basis  $\mathcal{D} = \{g_p\}_{p \in \Gamma}$ . If  $U$  is a projection over Fourier basis vectors of random frequencies,  $Uf$  is computed with  $O(N \log N)$  operations with an FFT.

Theorem 13.7, proved by Candès, Romberg, and Tao [139] and Rudelson and Vershynin [419], shows that randomized subsampled orthogonal transforms have low restricted isometry constants if the mutual coherence  $\mu(\mathcal{B} \cup \mathcal{D})$  of the union of bases is small.

**Theorem 13.7:** *Candès, Romberg, Tao, Rudelson, Vershynin.* Let  $\mathcal{D}$  be an orthonormal basis and  $U$  be a projector on a randomly chosen subset of vectors in an orthonormal basis  $\mathcal{B}$ . For any  $\delta < 1$ , there exists  $\beta$  such that for all  $M$  satisfying

$$M \leq \frac{\beta Q}{N \mu(\mathcal{B} \cup \mathcal{D})^2 (\log N)^5}, \quad (13.68)$$

the dictionary  $\mathcal{D}_U = U\mathcal{D}$  satisfies  $\delta_M(\mathcal{D}_U) \leq \delta$  with a probability that increases toward 1 like  $1 - N^{-c}$  where  $c$  is a constant.

A randomized subsampled orthogonal transform is not universal as opposed to Gaussian or Bernouilli random matrices. The upper bound depends on the mutual coherence of the random sampling orthonormal basis  $\mathcal{B}$  with the basis  $\mathcal{D}$  that provides a sparse signal representation. If  $\mathcal{B}$  is a discrete Fourier basis and  $\mathcal{D}$  is a Dirac basis then  $\mu(\mathcal{B} \cup \mathcal{D}) = N^{-1/2}$ , and (13.68) becomes  $M \leq CQ / (\log N)^5$ . In applications, random subsampling projectors are often used because they are computationally more efficient than Gaussian or Bernouilli random matrices [95].

### **Randomized Sparse Spike Deconvolutions**

A random Fourier sampling for sparse spike signals can be interpreted as a randomized sparse spike deconvolution. A comparison with standard sparse spike deconvolutions shows the importance of randomization to improve the signal recovery.

In a sparse spike deconvolution problem,  $Y = f \star u + W$  and  $f = \sum_{p \in \Lambda} a[p] \delta[n - p\Delta]$  is sparse in the Dirac basis  $\mathcal{D}$ . As explained in Section 13.3.2, sparse spike deconvolutions estimate the coefficients  $a[p]$  by decomposing  $Y$  in the transformed dictionary  $\mathcal{D}_U = \{u[n - \Delta p]\}_p$  with  $u$  normalized  $\|u\| = 1$ . Since  $\langle u[n - p\Delta], u[n - q\Delta] \rangle = u \star \bar{u}[\Delta(p - q)]$  with  $\bar{u}[n] = u[-n]$ , the dictionary mutual coherence is

$$\mu(\mathcal{D}_U) = \max_{p \neq 0} |\bar{u} \star u[p\Delta]|. \quad (13.69)$$

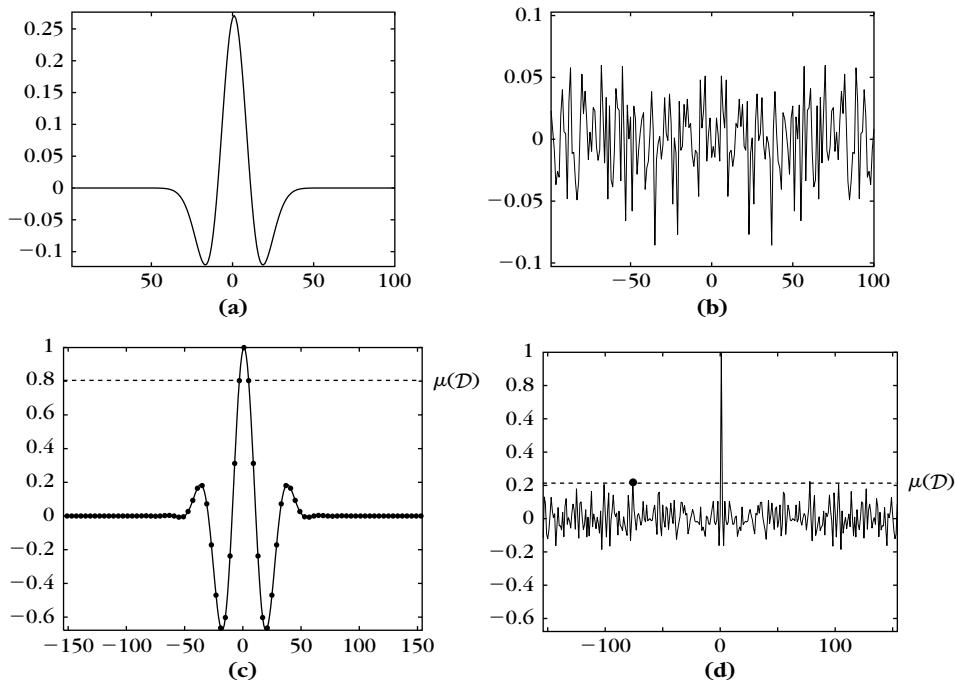


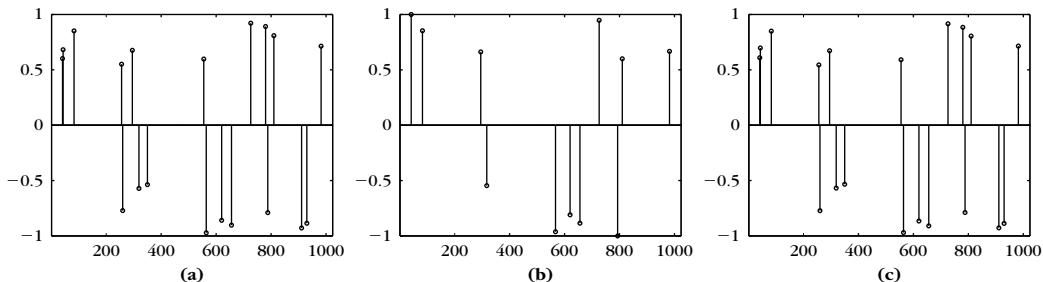
FIGURE 13.8

(a) Gaussian second-derivative wavelet  $u_1[n]$ . (b) Filter  $u_2[n]$  with  $\hat{u}_2[k]$  equal to 1 over  $Q$  random symmetric frequencies. (c) Value of  $u_1 \star \bar{u}_1[n]$  with  $\mu(\mathcal{D}_U) = 0.8$ . The dots have a spacing of  $\Delta$ . (d) Value of  $u_2 \star \bar{u}_2[n]$  with  $\mu(\mathcal{D}_U) = 0.2$ .

A seismic wavelet is a band-pass filter. Figure 13.8(a) gives an example of filter  $u_1$ , which is the second derivative of a Gaussian scaled by  $s$ . Since it is regular, the resulting dictionary coherence is close to 1 if  $\Delta/s$  is small, as illustrated in Figure 13.8(c) where  $\mu(\mathcal{D}_U) = 0.8$ .

A random Fourier sampling with  $Q/2$  random positive frequencies and  $Q/2$  symmetric negative frequencies is a convolution with a real filter  $u_2$  that has a Fourier transform  $\hat{u}_2[k]$  that is 1 over  $Q$  frequencies. Figure 13.8(b) shows that  $u_2[n]$  is highly irregular with a uniformly spread energy. As a result,  $\mathcal{D}_U$  has a low mutual coherence  $\mu(\mathcal{D}_U) = \max_{p>0} |u_2 \star \bar{u}_2[p\Delta]| = 0.2$ , shown in Figure 13.8(d).

The Gaussian derivative filter  $u_1$  has a Fourier transform that is nonnegligible over  $Q = 100$  frequencies like the Fourier random sampling filter  $u_2$ . Both filters thus provide  $Q$  frequency measurements. Figure 13.9(a) shows an example of sparse spike signal. Figures 13.9(b, c) give the estimated sparse spike signal  $\tilde{F}_1$  and  $\tilde{F}_2$  recovered from  $Y_1 = f \star u_1 + W$  and  $Y_2 = f \star u_2 + W$  with an  $L^1$  pursuit. As expected, close spikes are not recovered with the Gaussian derivative  $u_1$ , but they are recovered with  $u_2$ . This random Fourier sampling filter is able to recover the location and sign of 19 Diracs, including very close ones, from  $Q = 100$  Fourier frequencies.

**FIGURE 13.9**

(a) Sparse spike  $f$ . (b) Estimation  $\tilde{F}_1$  from  $Y_1 = u_1 \star f + W$ . (c) Estimation  $\tilde{F}_2$  from  $Y_2 = u_2 \star f + W$ .

The dictionary incoherence not only improves the stability and precision of the signal recovery but also the convergence of the iterative thresholding algorithm detailed in Theorem 12.9. This could suggest sending random waves for seismic exploration as opposed to usual “wavelets,” but regrettably seismic wavelet design must also take into account geophysical constraints and the attenuation of underground wave propagation.

### 13.4.2 Approximations with Compressive Sensing

Compressive sensing provides an alternative to linear and nonlinear approximation strategies. Previously described approximation schemes first compute a linear approximation in a high-resolution space  $\mathbf{U}_N$  with an error  $\varepsilon$ . A sparse support of size  $M \ll N$  is then calculated in some dictionary with an error that remains of the order of  $\varepsilon$ . Compressive sensing suggests to directly perform a sparse signal measurement with  $Q$  measurements while restoring a signal with the same approximation error. It is proved that compressive sensing recovery can indeed have an approximation error that has the same asymptotic decay as a nonlinear approximation error, but the devil is in the constants.

#### **Approximation Error**

If  $f$  is measured at a high resolution  $N$ , a nonlinear approximation is computed in an orthonormal basis  $\mathcal{D} = \{g_p\}_{p \in \Gamma}$  of  $\mathbb{C}^N$  by finding the  $M$ -largest coefficients. Let  $\{f_{\mathcal{D}}^r[k]\}_{1 \leq k \leq N}$  be the ranked coefficients  $\langle f, g_p \rangle$  in  $\mathcal{D}$ :  $|f_{\mathcal{D}}^r[k]| \geq |f_{\mathcal{D}}^r[k+1]|$ . The  $M$ -term approximation error is the energy of leftover smaller coefficients

$$\|f - f_M\|^2 = \sum_{k > M} |f_{\mathcal{D}}^r[k]|^2.$$

Instead of measuring  $f$  at a high resolution  $N$ , computing  $N$  coefficients  $\langle f, g_p \rangle$ , and throwing away the  $N - M$  smallest ones, we want to acquire a smaller number of  $Q$  measurements with the same error. Since we do not know where the  $M$ -largest coefficients of  $f$  in  $\mathcal{D}$  are, compressed sensing performs  $Q \geq M$

spread random measurements. Recovering an estimation nearly as precise as this  $M$ -term approximation strongly relies on  $M$ -restricted isometry conditions in the transformed dictionary

$$\begin{aligned} \forall \mathbf{a} \in \mathbb{C}^{|\Lambda|}, \quad (1 - \delta_M(\mathcal{D}_U)) \sum_{p \in \Lambda} |\mathbf{a}[p]|^2 &\leq \|\sum_{p \in \Lambda} \mathbf{a}[p] U g_p\|^2 \\ &\leq (1 + \delta_M(\mathcal{D}_U)) \sum_{p \in \Lambda} |\mathbf{a}[p]|^2 \end{aligned} \tag{13.70}$$

for all  $\Lambda \subset \Gamma$  with  $|\Lambda| \leq M$ .

Theorem 13.8 by Candès, Romberg, and Tao [137] gives an error bound to compare the nonlinear approximation error with the compressive sensing estimation  $\tilde{F}$  computed from  $Y = Uf + W$ . In this context, the noise  $W$  has a small energy that only limits the computational precision. A similar theorem is given by Donoho [219]. We write  $\Phi f = \langle f, g_p \rangle$  as the decomposition operator of  $\mathcal{D}$ .

**Theorem 13.8:** *Candès, Romberg, Tao.* Let  $U$  be a sensing matrix and  $\mathcal{D} = \{g_p\}_{p \in \Gamma}$  be an orthonormal basis. Suppose that  $M$  satisfies  $\delta_{3M}(\mathcal{D}_U) \leq \delta < 1/3$  with  $\mathcal{D}_U = \{Ug_p\}_{p \in \Gamma}$ . Let  $\tilde{F} = \sum_{p \in \Gamma} \tilde{\mathbf{a}}[p] g_p$  be defined by

$$\tilde{\mathbf{a}} = \underset{b \in \mathbb{C}^N}{\operatorname{argmin}} \|b\|_1 \quad \text{subject to} \quad \|\sum_{p \in \Gamma} b[p] U g_p - Y\| \leq \|W\|.$$

There exists a constant  $C$  that only depends on  $\delta$  such that

$$\|f - \tilde{F}\| \leq \frac{C}{\sqrt{M}} \sum_{k > M} |f_{\mathcal{D}}^r[k]| + C \|W\|, \tag{13.71}$$

where  $f_{\mathcal{D}}^r[k]$  are the ranked coefficients  $f_{\mathcal{D}}$  ordered by decaying magnitude.

**Proof.** The following proof uses arguments provided in [169]. We write  $a[p] = \langle f, g_p \rangle$  and  $\tilde{a} = \langle \tilde{f}, g_p \rangle$  as the decomposition coefficients in the orthonormal basis  $\mathcal{D}$ . Let  $\Lambda \subset \Gamma$  be the indexes  $p$  of the  $M$ -largest coefficients  $|a[p]|$ . The coefficient error  $b = a - \tilde{a}$  is evaluated on the complement  $\Lambda^c$  of  $\Lambda$  in  $\Gamma$ . This complement is subdivided into  $\Lambda^c = \Lambda_1 \cup \dots \cup \Lambda_k$ , where  $\Lambda_1$  are the indexes  $p$  of the  $2M$  largest coefficients  $|b_{\Lambda^c}[p]|$ ,  $\Lambda_2$  are the indexes of the next  $2M$  largest coefficients, and so on until  $\Lambda_k$ , which may contain fewer than  $2M$  elements. We also write  $\Lambda_{01} = \Lambda \cup \Lambda_1$  and  $\Lambda_{01}^c = \Lambda_2 \cup \dots \cup \Lambda_k$ .

Since for any  $p \in \Lambda_{j+1}$  and  $p' \in \Lambda_j$ ,  $|b[p]| \leq |b[p']|$ , so  $|b[p]| \leq (2M)^{-1} \|b_{\Lambda_j}\|_1$  and thus

$$\|b_{\Lambda_{j+1}}\| \leq (2M)^{-1/2} \|b_{\Lambda_j}\|_1.$$

Since  $\|b_{\Lambda_{01}^c}\| \leq \sum_{j \geq 2} \|b_{\Lambda_j}\|$ , it implies

$$\|b_{\Lambda_{01}^c}\| \leq (2M)^{-1/2} \sum_{j \geq 1} \|b_{\Lambda_j}\|_1 = (2M)^{-1/2} \|b_{\Lambda^c}\|_1. \tag{13.72}$$

Since  $\|\Phi^* a - Y\| = \|W\|$ , the definition of  $\tilde{a}$  implies  $\|\tilde{a}\|_1 \leq \|a\|_1$ . Triangular inequalities on the restriction of  $b = a - \tilde{a}$  to  $\Lambda$  and  $\Lambda^c$  give

$$\begin{cases} \|\tilde{b}_{\Lambda^c}\|_1 \leq \|\tilde{a}_{\Lambda^c}\|_1 + \|a_{\Lambda^c}\|_1, \\ \|\tilde{b}_{\Lambda}\|_1 \geq \|a_{\Lambda}\|_1 - \|\tilde{a}_{\Lambda}\|_1. \end{cases}$$

Together with  $\|\tilde{a}\|_1 \leq \|a\|_1$ , it implies that

$$\|b_{\Lambda^c}\|_1 \leq \|b_\Lambda\|_1 + 2\|a_{\Lambda^c}\|_1 = \|b_\Lambda\|_1 + 2\eta \quad \text{where} \quad \eta = \sum_{k>M} |f_D^r[k]|. \quad (13.73)$$

Equations (13.72) and (13.73) lead to

$$\|b_{\Lambda_{01}^c}\| \leq \frac{1}{\sqrt{2M}} \|b_\Lambda\|_1 + \frac{2}{\sqrt{2M}} \eta \leq \frac{1}{\sqrt{2}} \|b_\Lambda\| + \frac{2}{\sqrt{2M}} \eta \leq \frac{1}{\sqrt{2}} \|b_{\Lambda_{01}}\| + \frac{2}{\sqrt{2M}} \eta.$$

The recovery error is thus written as a function of  $\|b_{\Lambda_{01}}\|$  as follows:

$$\|a - \tilde{a}\| = \|b\| \leq \|b_{\Lambda_{01}}\| + \|b_{\Lambda_{01}^c}\| \leq \left(1 + \frac{1}{\sqrt{2}}\right) \|b_{\Lambda_{01}}\| + \frac{2}{\sqrt{2M}} \eta. \quad (13.74)$$

Since  $\|\Phi^* \tilde{a} - Y\| \leq \|W\|$ , the vector  $b$  satisfies

$$\|\Phi^* b\| \leq \|\Phi^* a - Y\| + \|\Phi^* \tilde{a} - Y\| \leq 2\|W\|. \quad (13.75)$$

The reversed triangular inequality gives

$$2\|W\| \geq \|\Phi^* b\| \geq \|\Phi^* b_{\Lambda_{01}}\| - \|\Phi^* b_{\Lambda_{01}^c}\|. \quad (13.76)$$

The restricted isometry inequality (13.70) applies to  $b_{\Lambda_j}$  which is a vector with less than  $2M$  coefficients, and  $\|\Phi^* b_{\Lambda_{01}^c}\|$  is bounded with the same argument as in (13.72),

$$\|\Phi^* b_{\Lambda_{01}^c}\| \leq \sqrt{1 + \delta_{2M}(\mathcal{D}_U)} \sum_{j \geq 2} \|b_{\Lambda_j}\| \leq \frac{\sqrt{1 + \delta_{2M}(\mathcal{D}_U)}}{\sqrt{2M}} \|b_{\Lambda^c}\|_1,$$

which, together with (13.73), leads to

$$\begin{aligned} \|\Phi^* b_{\Lambda_{01}^c}\| &\leq \frac{\sqrt{1 + \delta_{2M}(\mathcal{D}_U)}}{\sqrt{2M}} (\|b_\Lambda\|_1 + 2\eta) \\ &\leq \frac{\sqrt{1 + \delta_{2M}(\mathcal{D}_U)}}{\sqrt{2}} \|b_\Lambda\| + 2 \frac{\sqrt{1 + \delta_{2M}(\mathcal{D}_U)}}{\sqrt{2M}} \eta. \end{aligned} \quad (13.77)$$

The restricted isometry inequality (13.70) applied to  $b_{\Lambda_{01}}$ , which has at most  $3M$  nonzero coefficients, gives  $\|\Phi^* b_{\Lambda_{01}}\| \geq \sqrt{1 - \delta_{3M}(\mathcal{D}_U)} \|b_{\Lambda_{01}}\|$ . Inserting this inequality and the upper bound (13.77) on  $\|\Phi^* b_{\Lambda_{01}^c}\|$  in (13.76) with  $\|b_\Lambda\| \leq \|b_{\Lambda_{01}}\|$  gives

$$2\|W\| \geq A \|b_{\Lambda_{01}}\| - 2 \frac{\sqrt{1 + \delta_{2M}(\mathcal{D}_U)}}{\sqrt{2M}} \eta \quad \text{where} \quad A = \sqrt{1 - \delta_{3M}(\mathcal{D}_U)} - \frac{1}{\sqrt{2}} \sqrt{1 + \delta_{2M}(\mathcal{D}_U)}.$$

Since  $\delta_{2M}(\mathcal{D}_U) \leq \delta_{3M}(\mathcal{D}_U)$ , the condition  $\delta_{3M}(\mathcal{D}_U) < 1/3$  implies that  $A > 0$ , so that  $\|b_{\Lambda_{01}}\|$  is bounded as follows:

$$\|b_{\Lambda_{01}}\| \leq 2 \frac{\sqrt{1 + \delta_{2M}(\mathcal{D}_U)}}{A\sqrt{2M}} \eta + \frac{2}{A} \|W\|.$$

Using this bound on  $\|b_{\Lambda_{01}}\|$  in the recovery error  $\|a - \tilde{a}\|$  of (13.74) gives the result of the theorem. ■

The theorem implies in particular that if  $\delta_{3M}(\mathcal{D}_U) < 1/3$  and if there is no noise  $W = 0$ , then a signal  $f = \sum_{p \in \Lambda} a[p]g_p$  that has  $|\Lambda| \leq M$  nonzero coefficients is exactly recovered by a basis pursuit. However, the theorem is much more powerful since it proves stability of compressed sensing approximations by relating the error to the decay of smaller-amplitude signal coefficients. The proof does not rely on the orthogonality or the independence of vectors in  $\mathcal{D}$  and thus can be extended to redundant frames. Similar results hold for modified orthogonal matching pursuit algorithms [383, 384], and other algorithmic approaches have also been developed to recover sparse Fourier expansions from few randomized point-wise evaluations [272].

When  $W$  is a Gaussian white noise of variance  $\sigma^2$ , following the work of Candès and Tao [143], Bickel, Ritov, and Tsybakov [113] proved that the noise term  $\|W\|$  that appears in (13.71) can be reduced to its projection over the space generated by the vectors in the approximation support of  $\tilde{F}$ . For a dictionary of size  $P$ , they proved that under a similar  $M$ -restricted isometry condition, solving an  $\mathbf{l}^1$  Lagrangian pursuit with a threshold  $T = \lambda \sigma \sqrt{2 \log_e P}$  yields almost the same result as the model selection theorem (12.3), which minimizes an  $\mathbf{l}^0$  Lagrangian. If  $\tilde{\Lambda}$  is the support of the solution  $\tilde{a}$ , then the noise term  $\|W\|$  is replaced by  $|\tilde{\Lambda}|T$  up to a multiplicative factor.

### Compressive Sensing versus Nonlinear Approximations

The nonlinear approximation error  $\|f - f_M\|^2$  in a basis  $\mathcal{D}$  is typically computed by assuming that the sorted coefficients have a decay  $|f_{\mathcal{D}}[k]| = O(k^{-s})$  in which case Theorem 9.9 proves that  $\|f - f_M\| = O(M^{-s+1/2})$ . Theorem 13.9 proves that for  $s > 1$ , the compressive sensing error  $\|\tilde{F} - f\|$  has nearly the same asymptotic decay rate.

**Theorem 13.9.** Suppose that  $|f_{\mathcal{D}}^r[k]| = O(k^{-s})$  with  $s > 1$ . Let  $U$  be a Gaussian or a Bernoulli random matrix. Any vector  $Y$  of  $Q$  measurements yields an approximation

$$\tilde{F} = \sum_{p \in \Gamma} \tilde{a}[p]g_p \quad \text{with} \quad \tilde{a} = \underset{b \in \mathbb{C}^N}{\operatorname{argmin}} \|b\|_1 \quad \text{subject to} \quad \left\| \sum_{p \in \Gamma} \tilde{b}[p]Ug_p - Y \right\| \leq \|W\|,$$

which satisfies

$$\|f - \tilde{F}\| = O\left((\beta Q)^{1/2-s} |\log(N/Q)|^{s-1/2} + \|W\|\right), \quad (13.78)$$

with a probability that increases exponentially to 1 with  $N$ .

**Proof.** Theorem 13.6 proves in (13.67) that the condition  $\delta_{3M} < 1$  of Theorem 13.8 is achieved for  $M \leq \beta Q / (\log N/Q)$  if  $U$  is a Gaussian random matrix. This result is also valid for Bernoulli matrices. Since  $|f_{\mathcal{D}}^r[k]| = O(k^{-s})$ , there exists  $C > 0$  such that

$$\frac{1}{\sqrt{M}} \sum_{k > M} |f_{\mathcal{D}}^r[k]| \leq \frac{1}{\sqrt{M}} \sum_{k > M} C k^{-s} \leq \frac{C}{M^{1/2}} \cdot \frac{M^{1-s}}{1-s},$$

and inserting this in (13.71) implies (13.78). ■

If  $U$  computes a random sampling in a Fourier basis and if  $\mathcal{D}$  is a Dirac basis, then Theorem 13.7 implies that the upper bound (13.78) is satisfied if  $|\log(N/Q)|^{1/2-s}$  is replaced by  $|\log N|^{5(1/2-s)}$ . This theorem proves that with  $Q$  measurements a compressed sensing recovers an approximation  $\tilde{F}$  with an error having the same asymptotic decays  $Q^{1/2-s}$  as a nonlinear approximation, up to a logarithmic factor  $\log(N/Q)$ . This is a spectacular result because a standard linear approximation can have a much slower decay depending on the distribution of large coefficients. However, the applicability of this result depends on the constants that are involved, and evaluating them is a delicate topic.

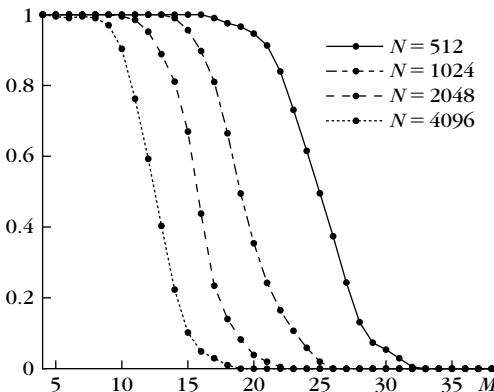
### Perfect Recovery Constants

Theorem 13.8 proves that an  $M$  sparse signal having only  $M$  nonzero coefficients is exactly reconstructed if  $\delta_{3M}(\mathcal{D}_U) < 1/3$ . Theorem 13.6 implies that this is valid if  $M = \beta Q / (\log N/Q)$ , but lower bounds of  $\beta$  computed so far are very small, which gives very pessimistic upper bounds of  $Q/M$ .

A lower bound of  $Q/M$  can be computed by looking for “bad” sparse signals  $f$  with  $M$  coefficients and that are not recovered exactly with  $Q = CM$  measurements  $Y = Uf$  without noise. For  $N = 1024$  and  $Q = 100$ , one can find such signals with  $M = 6$ , which proves that  $Q/M \geq 100/6$ . This is still a large constant, which a priori should increase with  $N$  because of the  $\log(N/Q)$  factor.

More encouraging evaluations of  $Q/M$  are computed “on average” with Monte-Carlo simulations. Results are slightly improved by renormalizing the vectors  $Ug_m$  of a Gaussian random matrix  $U$  in order to adjust their norm exactly to 1. A random  $M$  sparse signal  $F$  is defined with  $M$  nonzero coefficient positions that are randomly distributed in  $\{0, \dots, N-1\}$  with amplitudes that are independent Gaussian random variables of unit variance. For realizations  $f$  of  $F$ , a basis pursuit approximation  $\tilde{f}$  is computed from the  $Q$  measurements  $Y = Uf$  with no noise. The ratio of perfect recovery is the probability that an  $M$  sparse signal  $f$  drawn from this distribution is exactly recovered by the basis pursuit. It is evaluated numerically by Monte-Carlo sampling.

Figure 13.10 shows the recovery performance of basis pursuit for  $Q = 100$  and for several values of  $N$ . For  $Q = 100$  and  $N = 1024$ , the average perfect recovery ratio reaches 1 for  $M = 13$ , which is much better than the worse case  $M = 6$  previously mentioned. The recovery performance deteriorates when  $N$  increases, which is consistent with Theorem 13.6. The perfect recovery ratio reaches 1 for  $Q/M = 6.2$  when  $N = 512$ , for  $Q/M = 7.7$  when  $N = 1024$ , for  $Q/M = 11$  when  $N = 2048$ , and for  $Q/M = 16$  when  $N = 4096$ . Bernoulli matrices yield nearly the same recovery ratio as Gaussian matrices. If the orthogonal basis  $\mathcal{D}$  is a Dirac basis and  $U$  is a random Fourier sampling matrix, then the ratio of perfect recovery remains nearly the same. These results do not prevent us from obtaining better results in examples with large-amplitude coefficients, such as in Figure 13.9 where  $M = 19$  elements are recovered from  $Q = 100$  measures for  $N = 1024$ .

**FIGURE 13.10**

Perfect recovery ratio calculated with a basis pursuit, for  $Q = 100$  random Gaussian measurements for several signal size values  $N$ .

### Approximation Constants

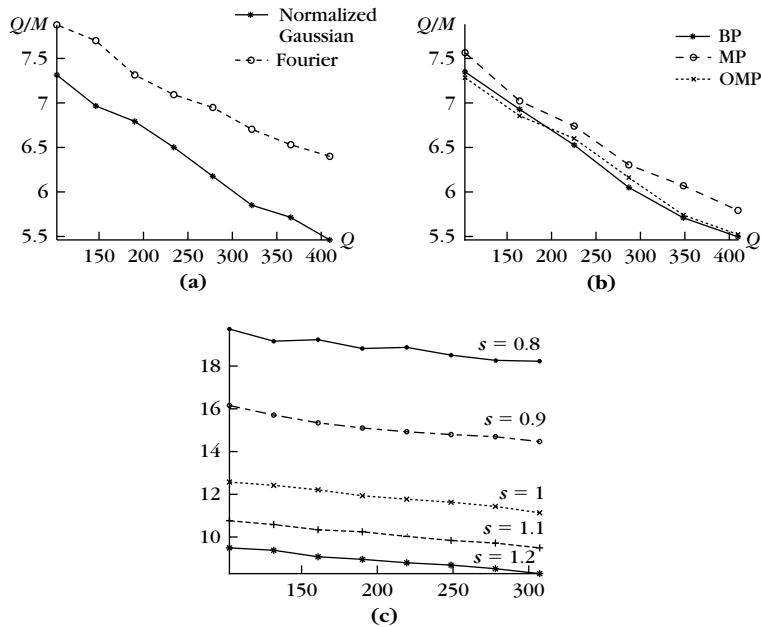
Perfect recovery experiments give very partial information because most signals are not  $M$  sparse. To better understand the range of applicability of compressed sensing, we give a numerical indication of the ratio  $Q/M$  for which the compressed sensing approximation produces an error equal to an  $M$ -term nonlinear approximation. This is again performed with Monte-Carlo simulations that compute an average case over a particular signal model. It is therefore not computed from worst cases.

Large signal classes can be modeled by the decay of their sorted coefficients in an orthonormal basis  $\mathcal{D}$ . Bounded variation images have sorted wavelet coefficients that have a decay that is  $O(k^{-s})$  for  $s = 1$ . For the boat, Lena, and Barbara images previously shown, the exponent is indeed close to 1. For the mandrill image in Figure 10.7,  $s = 0.7$  because of the irregular textures. Let  $F[n]$  be a random vector with coefficients  $\{\langle F, g_p \rangle\}_{0 \leq p < N}$  in  $\mathcal{D}$  that are a random permutation of the values  $\{(-1)^k k^{-\alpha}\}_{0 \leq k < N}$ . The ranked coefficients of  $F$  always satisfy  $|F_{\mathcal{D}}^r[k]| = k^{-s}$ , but there is no prior information on the location of large versus small coefficients. For all realizations, the  $M$ -term approximation error is therefore

$$\varepsilon_n[M] = \sum_{|k| > M} k^{-2s} \approx (2s - 1)^{-1/2} M^{-2s+1}. \quad (13.79)$$

We study the average value of  $Q/M$  as a function of the number  $Q$  of measurements needed to have a compressed sensing error that is equal to the  $M$  term approximation error  $\varepsilon_n[M]$ . It compares the relative approximation efficiency of compressed sensing and nonlinear approximations.

Figure 13.11(a) gives the average value of  $Q/M$ , computed for  $s = 3/2$  in a Dirac basis  $\mathcal{D}$ , with a normalized Gaussian random measurement matrix and a random Fourier sampling. This ratio has a small relative variation that verifies that the compressed sensing error has the same decay rate as the nonlinear approximation

**FIGURE 13.11**

Ratio  $Q/M$  as a function of  $Q$  to obtain the same error with  $Q$  compressed-sensing measurements and an  $M$ -term nonlinear approximation. (a) Results for a normalized random Gaussian matrix and a Fourier random sampling with  $N = 1024$  and  $s = 3/2$ . (b) Comparison between a matching pursuit with backprojection (MP), an orthogonal matching pursuit, and a basis pursuit with backprojection for  $N = 1024$  and  $s = 3/2$ . (c) Evolution of  $Q/M$  for  $N = 1024$  and  $s$  varying.

error, as predicted by Theorem 13.9. The slow decay is partially explained by the  $\log(Q/N)$  factor for Gaussian random matrices. A Fourier random sampling is not universal, but in this favorable case where the signal is sparse in a Dirac basis, it gives a slightly lower  $Q/M$  ratio. For  $s = 3/2$ , numerical experiments for several signal sizes  $N$  verify that  $Q/M \approx \beta (\log_2 N)$ , which is coherent with the theorem statement. In these experiments,  $\beta \approx 0.75$  for  $s = 3/2$ .

Figure 13.11(b) gives the evolution of the ratio depending on the algorithm used to compute the sparse approximation of  $Y$  in the transformed dictionary  $\mathcal{D}_U$ . An  $\ell^1$  Lagrangian pursuit slightly outperforms an orthogonal matching pursuit that is slightly better than a nonorthogonal matching pursuit, but in this case, the difference of efficiency between these algorithms is not so large relative to the difference of computational complexity.

Theorem 13.9 is valid only for  $s > 1$ . Figure 13.11(c) gives the value of  $Q/M$  when  $s$  decreases with a random Fourier sampling. Estimations are computed with an orthogonal matching pursuit. For each  $s$  the ratio remains nearly constant, but when  $s$  goes below 1 this ratio increases very significantly—up to 19 for  $s = 0.8$ .

### **Compressed Sensing versus Linear Image Approximations**

Numerical experiments verify that the error of a compressed sensing has the same decay rate as a nonlinear approximation error, which is by itself remarkable since all measurements are linear. However, for images the constant factors are quite large and in the same range as standard linear approximations.

In a wavelet orthonormal basis, the sorted coefficients of most images have a decay exponent  $s \leq 1$ . For images of  $N = 512^2$  pixels, the ratio  $Q/M$  is typically below 5 to obtain the same error with an  $M$ -term wavelet approximation and a linear approximation with  $Q$  low frequencies provided by a uniform sampling. This is below the compressed sensing ratio previously computed. However,  $Q/M$  does not remain constant for linear approximations, which have errors that decay more slowly than  $M$ -term approximation errors.

Compressed sensing is improved by using more prior information on the image. Donoho and Tsaig [224] use the distribution of wavelet coefficients across scales by computing a scale-by-scale compressed sensing of wavelet coefficients. The number of wavelet coefficients at a scale  $2^j$  is proportional to  $2^{-2j}$ , but a constant number of random measurements is performed at each scale. This is coherent with image properties, where an edge produces the same number of large-amplitude coefficients at each scale. Incorporating such prior information lowers the ratio  $Q/M$  between 4 and 5 but it depends on the image size and its content. Further reductions are possible by mixing linear and compressed sensing measurements, which can outperform linear approximations [136, 415], depending on the images.

Improving constants is a central challenge for compressive sensing applications that will influence the range of its applications. These algorithms can indeed reduce approximation errors by taking better advantage of new representations or prior information on signal coefficients [96].

#### **13.4.3 Compressive Sensing Applications**

Randomized data acquisition offers the possibility to improve the resolution of measurement devices. For analog signal acquisition, the measurement operator  $\bar{U}$  provides randomized linear measurements of an analog signal  $\bar{f}(x)$ . One can build hardware that implements this measurement randomization. Examples are given with fully randomized sensors such as a single-pixel camera. Applications can also involve a mixture of randomized and structured acquisitions, which both play a role in the signal reconstruction.

Compressive acquisition is a democratic acquisition process where all measurements are equally important. The loss of a particular coefficient introduces an error that is diluted over the whole signal reconstruction and is thus less visible than a localized error created by a dysfunctional sensor such as a camera pixel. This robustness can also be important for signal acquisition with unreliable sensors.

#### **Analog-to-Digital Converters**

Two compressed sensing strategies are studied by Candès and Wakin [144] to improve current analog-to-digital converters. For signals that are sparse in a Fourier

basis, a uniform sampling at a Nyquist rate can be replaced by a random sampling in time. Indeed, the Fourier basis and Dirac bases have a mutual low coherence and Theorem 13.7 proves that a random time sampling yields an incoherent dictionary. If the signal is sufficiently sparse in frequency, the average sampling rate becomes lower than the uniform Nyquist rate.

Ultrawide-band signals in communication have a bandwidth that is limited by hardware analog-to-digital conversion [294]. Whereas it can be difficult to increase the sampling rate, changing the signal polarization at a very high rate may be possible. A random modulation multiplies the signal at a very high rate with  $+1$  or  $-1$  and performs an integration over a time window, which is digitized at regular time intervals. This is implemented over multiple parallel channels that modulate the signal with different random sequences of  $+1$  and  $-1$  in order to provide enough Bernouilli random measurements. This random sampling operator is universal and can thus be applied to any signal having a sparse representation in some dictionary, such as a time-frequency Gabor dictionary.

### ***Single-Pixel Camera***

Takhar et al. [453] built a compressive sensing camera that uses a single pixel photodetector to compute inner products with  $Q$  measurement vectors  $\{\bar{u}_q\}_q$ . A micromirror array located on the focal plane of the camera multiplies the image of the scene  $\tilde{f}(x)$  with a pseudo-random mask  $\bar{u}_q$  that has constant values  $+1$  or  $-1$  on a regular lattice of  $N$  squares. A photoreceptor sums this randomly modulated signal, which computes  $Y[q] = \langle \tilde{f}, \bar{u}_q \rangle + W[q]$ . It implements a random Bernouilli measurement with random signs. The Bernouilli random measurement matrix is universal and can thus take advantage of any dictionary  $\mathcal{D}$  providing a sparse image representation in order to recover a high-resolution image  $f \in \mathbb{C}^N$ .

### ***Tomography and MRI Imaging***

Tomography imaging acquires integrals  $\overline{U}\tilde{f}$  of the analog signal  $\tilde{f}(x)$  along rays, as illustrated in Figure 13.7. It provides a subsampling of the signal Fourier transform  $\widehat{\tilde{f}}(\omega)$  for  $\omega \in \Omega$ . Angles are usually chosen to be uniformly distributed so that  $\Omega$  is located along evenly distributed rays. Randomizing the ray locations can increase the incoherence of the tomography inversion problem, but the ray integration considerably limits the level of incoherence.

Medical resonance imaging (MRI) is another example of acquisition that subsamples Fourier frequencies  $\widehat{\tilde{f}}(\omega)$  for  $\omega \in \Omega$ . The excitation of atoms with a spatially varying magnetic field can select an arbitrary  $\omega \in \Omega$ . In contrast to tomography,  $\Omega$  is therefore not restricted to be located along rays. In principle, a pseudo-random set  $\Omega$  could be chosen to obtain an incoherent inverse problem. However, physical and physiological limitations enforce the magnetic field direction to follow smooth sampling curves so that  $\Omega$  cannot be fully random. For applications to whole-heart coronary MRIs, Lustig et al. [356] designed a set  $\Omega$  composed of spirals distributed radially with a pseudo-random density. This reduces the number of

required measurements and makes it possible to acquire an image of the entire heart in a single held breath of the patient.

### Error Correction

Channel coding adds some information to a message  $f$  to build a longer encoded message  $Y$  that is robust against noisy transmissions. Classical error-correcting methods [39] consider a message  $f$  over a finite alphabet, and construct  $Y$  from  $f$  using arithmetics over finite fields. Compressed sensing provides strategies to design error-correcting codes over the real numbers.

Wyner [491] considers a real-valued message  $f \in \mathbb{R}^N$ , and performs the coding as  $Y = Af \in \mathbb{R}^{N_0}$  where  $A \in \mathbb{R}^{N_0 \times N}$  is a random matrix. The additional  $Q = N_0 - N$  dimensions encode redundant information that is used to detect errors  $e$  of an unreliable transmission  $Y = Af + e$ .

Since most of the entries of the error  $e = Y - Af$  are expected to be zero, the signal  $f$  is recovered using an  $\ell^1$  optimization

$$\tilde{f} = \underset{h \in \mathbb{R}^N}{\operatorname{argmin}} \|Y - Ah\|_1. \quad (13.80)$$

The vectors  $d = Y - Ah$  is the set of vectors such that

$$Ud = U(Y - Ah) = UY,$$

where  $U \in \mathbb{R}^{Q \times N_0}$  in any annihilating matrix that satisfies  $\mathbf{Null}U = \mathbf{Im}A$ , and thus  $UA = 0$ . Thanks to this change of variables, the error is recovered with basis pursuit

$$\tilde{e} = \underset{d \in \mathbb{R}^{N_0}}{\operatorname{argmin}} \|d\|_1 \quad \text{subject to} \quad Ud = UY, \quad (13.81)$$

which can be solved by linear programming, as explained in Section 12.4.1. The signal  $f$  is estimated by the solution  $\tilde{F}$  of the equation  $A\tilde{F}f = Y - \tilde{e}$ .

Donoho and Huo [230] made a first analysis of this algorithm with a mutual incoherence argument. This result was refined by Candès and Tao [138] with a compressive algorithm. If the annihilating matrix  $U$  has a small restricted isometry constant,  $\delta_{3M}(U) < 1/3$ , and if  $\|e\|_0 \leq M$ , then Theorem 13.8 proves that we have an exact signal recovery  $\tilde{F} = f$ . If  $U$  is a Gaussian random matrix, then Theorem 13.6 proves with a large probability that  $\delta_{3M}(U) < 1/3$  for  $M = \beta Q / (\log N_0 / Q)$ . If  $A$  is an encoding matrix that is annihilated by  $U$ , then the  $\ell^1$  basis pursuit (13.81) can thus recover  $CQ / \log(N_0 / Q)$  transmission errors.

---

## 13.5 BLIND SOURCE SEPARATION

Signals are sometimes recorded as a mixture, from which the original sources must be separated. Separating the sounds of  $S$  musical instruments recorded with  $K$  microphones is an example, with  $K = 2$  for stereo recordings. Discriminating the heartbeat of a fetus from the heartbeat of his or her mother with  $K$  electrocardiogram signals is another source separation example with  $S = 2$ . In these *blind source separation* problems, the mixing parameters of the sources are unknown.

Linear mixtures of sources is a particularly appropriate model for sound acquisition. The  $K$  channel measurements of  $S$  sources  $\{f_s\}_{0 \leq s < S}$  can then be written as

$$Y_k[n] = \sum_{s=0}^{S-1} u_{k,s} f_s[n] + W_k[n] \quad \text{for } 0 \leq k < K, \quad (13.82)$$

where  $U = \{u_{k,s}\}_{0 \leq k < K, 0 \leq s < S}$  is the mixing matrix and  $W_k$  are measurement noises.

The number of measurements  $K$  is often smaller than the number  $S$  of sources. Knowing the mixing matrix is then not enough to recover the sources  $f_s$  from the measurements  $Y_k$ . This source separation can be interpreted as a super-resolution problem where  $S$  sources of size  $N$  and thus  $SN$  data values must be recovered from  $Q = KN < SN$  measurements. The situation looks worse than in previous inverse problems since the operator  $U$  to invert is not even known.

A successful method for source separation is based on stochastic source models, which are supposed to be independent. The pioneer work of Herault and Jutten [298] and Comon [187] has established the principles of independent component analysis for source separation. Efficient procedures such as the JADE algorithm of Cardoso [148] are separating sources by optimizing functionals that promote the independence of the recovered sources. As previously explained, it can be difficult to define a stochastic model of complex signals and thus verify that they are independent. This strong independence assumption is also not always valid, for instance, in the recording of musical instruments playing together.

Sparse blind source separation is based on weaker deterministic models. Prior information is used to define a dictionary where the different sources have a sparse representation. Jourjine, Rickard, and Yilmaz [321, 493] as well as Zibulevsky et al. [117, 498, 499] have developed algorithms that estimate the mixing operators and all the sources under the hypothesis that sources have approximation supports that do not overlap too much in an appropriate dictionary. These algorithms have then been further refined by a number of authors [103, 116, 270, 308, 343, 349, 350, 457]. The smaller the approximation support of the sources, the more likely they are to be separated. Support separation is even stronger if the dictionary takes into account differences between the sources to guarantee that the chosen dictionary vectors are different. Following a French discussion is indeed much easier in a cocktail of English speakers.

### 13.5.1 Blind Mixing Matrix Estimation

For blind source separation, the mixing coefficients  $u_{k,s}$  are estimated by constructing a sparse representation of the multichannel measurements. As explained in Section 12.6, a whitening operator may first be applied to the measurement vectors in order to decorrelate and renormalize all channels before further processing.

#### *Sparse Multichannel Signal Decomposition*

Let us represent multichannel measurements and the mixing matrix as vectors in  $\mathbb{R}^K$ :

$$\vec{Y}[n] = (Y_k[n])_{0 \leq k < K}, \quad \vec{W}[n] = (W_k[n])_{0 \leq k < K}, \quad \text{and} \quad \vec{u}_s = (u_{k,s})_{0 \leq k < K}.$$

The measurement equation:

$$Y_k[n] = \sum_{s=0}^{S-1} u_{k,s} f_s[n] + W_k[n] \quad \text{for } 0 \leq k \leq K \quad (13.83)$$

is rewritten as a signal vector equation:

$$\vec{Y}[n] = \sum_{s=0}^{S-1} \vec{u}_s f_s[n] + \vec{W}[n]. \quad (13.84)$$

Let  $\mathcal{D} = \{\phi_p\}_{p \in \Gamma}$  be a dictionary of unit vectors in which each source  $f_s$  has a sparse approximation. The measurement vector  $\vec{Y}$  also has a sparse approximation support, which is the union of the supports of all  $f_s$ . As explained in Section 12.6, the multichannel signal  $\vec{Y}$  can be decomposed in  $\mathcal{D}$  by calculating the inner product vectors:

$$\langle \vec{Y}, \phi_p \rangle = \left( \langle Y_k, \phi_p \rangle \right)_{0 \leq k \leq K} \in \mathbb{R}^K.$$

The Euclidean norm of a vector  $\vec{a} \in \mathbb{R}^K$  is written as  $\|\vec{a}\| = \sqrt{\sum_{k=0}^{K-1} |a[k]|^2}$ . The noise is reduced by thresholding the inner product norms  $\|\langle \vec{Y}, \phi_p \rangle\|$ . The resulting approximation support is

$$\tilde{\Lambda} = \left\{ p \in \Gamma : \|\langle \vec{Y}, \phi_p \rangle\| \geq T \right\}.$$

If the noise is not white, then  $T$  depends on  $p$  and is proportional to the noise variance in the direction of  $\phi_p$  with a large multiplicative factor.

Let us write

$$\vec{b}[p] = \langle \vec{Y}, \phi_p \rangle, \quad \vec{w}[p] = \langle \vec{W}, \phi_p \rangle \quad \text{and} \quad a_s[p] = \langle f_s, \phi_p \rangle. \quad (13.85)$$

Computing the inner product of the measurement vector equation (13.83) with  $\phi_p$  gives

$$\vec{b}[p] = \sum_{s=0}^{S-1} a_s[p] \vec{u}_s + \vec{w}[p] \quad \text{for } p \in \tilde{\Lambda}. \quad (13.86)$$

The vectors  $\vec{b}[p]$  thus define a cloud of points in  $\mathbb{R}^K$  that are a combination of the  $S$  mixing vectors  $\vec{u}_s$  plus a noise perturbation.

### Support Separation

If the source supports are mostly disjoint, then the mixing vectors can be identified. Indeed, suppose that the supports of each  $a_s[p]$  are strictly disjoint:

$$a_s[p] \neq 0 \Rightarrow a_{s'}[p] = 0 \quad \text{for } s' \neq s.$$

Since for each  $p \in \tilde{\Lambda}$  there exists a single  $s$  for which  $a_s[p] \neq 0$ , (13.86) implies

$$\vec{b}[p] = a_s[p] \vec{u}_s + \vec{w}[p] \in \mathbb{R}^K, \quad (13.87)$$

which provides the direction of the mixing vector up to a noise perturbation

$$\frac{\vec{b}[p]}{\|\vec{b}[p]\|} = \frac{\vec{u}_s}{\|\vec{u}_s\|}(1 + \varepsilon_1) + \frac{\vec{w}[p]}{\|\vec{b}[p]\|} \quad \text{with} \quad |\varepsilon_1| \leq \frac{\|\vec{w}[p]\|}{\|\vec{b}[p]\|}.$$

The noise perturbation is small if we only keep coefficients  $\vec{b}[p]$  with norms that are larger than a threshold that is well above the noise variance, so that

$$\frac{\|\vec{w}[p]\|}{\|\vec{b}[p]\|} \leq \varepsilon \ll 1 \quad (13.88)$$

with a high probability.

The strict support disjoint hypothesis is typically not satisfied. However, if a signal coefficient is much larger than the others and satisfies

$$\varepsilon |\alpha_s[p]| \|\vec{u}_s\| \geq \sum_{s \neq s'} \|\vec{u}_{s'}\| |\alpha_{s'}[p]|, \quad (13.89)$$

then we verify with (13.86) by inserting (13.88) and (13.89) that the normalized coefficients give the mixing direction with a small error that is of the order of  $\varepsilon$ :

$$\frac{\vec{b}[p]}{\|\vec{b}[p]\|} = \frac{\vec{u}_s}{\|\vec{u}_s\|} + \vec{\varepsilon}_2 \quad \text{with} \quad \|\vec{\varepsilon}_2\| \leq 3\varepsilon. \quad (13.90)$$

If there are enough such coefficients for all  $s$ , then the mixing direction is identified with a voting procedure using a histogram. Norm  $\|\vec{u}_s\|$  is not recovered, which means that ultimately sources are computed up to a multiplicative factor. However, this global amplitude is most often an arbitrary factor that is normalized afterward.

The accuracy of this algorithm relies on the near separation of the source supports. Source supports are more likely to be separated if each  $f_s$  has a sparse approximation in  $\mathcal{D}$ , with a support  $\Lambda_s$  that has size  $M_s$ , which is small relative to  $N$ . Indeed, these sets are then unlikely to intersect often. Figure 13.12 gives a synthetic simulation illustrating the impact of sparsity. We consider three sources having  $M_s = M$  nonzero coefficients in an orthonormal basis  $\mathcal{D}$ . In this simulation, these  $M$  coefficients are randomly distributed among the  $N$  vectors of the basis. Figure 13.12 shows that for a relatively low sparsity  $M/N = 0.4$ , the three mixing vector directions can still be identified.

### ***Identification of Mixing Directions***

The source directions  $\vec{u}_s/\|\vec{u}_s\|$  are identified with a voting implemented by local maxima detection in a histogram of directions. The source directions belong to the unit sphere of  $\mathbb{R}^K$  and are thus characterized by  $K - 1$  parameters that can be a vector  $\vec{\theta}_s$  of  $K - 1$  angles. For  $K = 2$  measurements, there is a single angle.

According to (13.90), a normalized coefficient  $\vec{b}[p]/\|\vec{b}[p]\|$  gives the direction of a dominating source  $\vec{u}_s$  up to an error that depends on the noise and the relative energy of other sources in this direction. Let  $\vec{\theta}(p)$  be the angle vector of  $\vec{b}[p]/\|\vec{b}[p]\|$ .

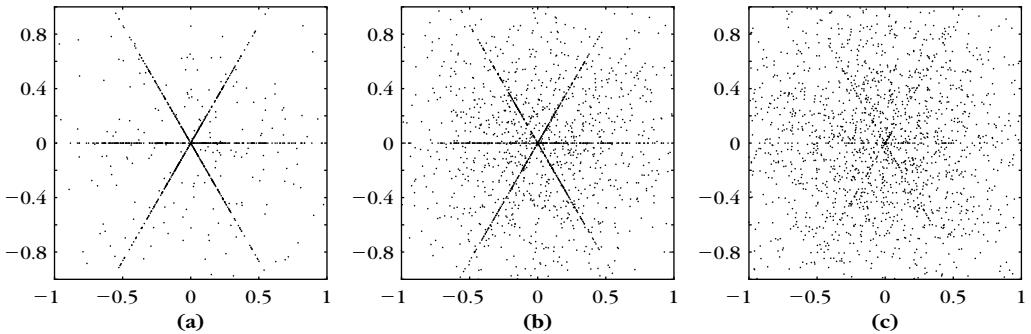


FIGURE 13.12

Distribution of synthetic coefficients  $\{\vec{b}[p]\}_p$  computed with  $S = 3$  sources having  $M$  nonzero coefficients each, and that are randomly distributed among the  $N$  coordinates of a basis: (a)  $M/N = 0.1$ , (b)  $M/N = 0.4$ , and (c)  $M/N = 0.7$ .

A histogram is defined over all angles with a weighting function that depends on  $\|\vec{b}[p]\|$ :

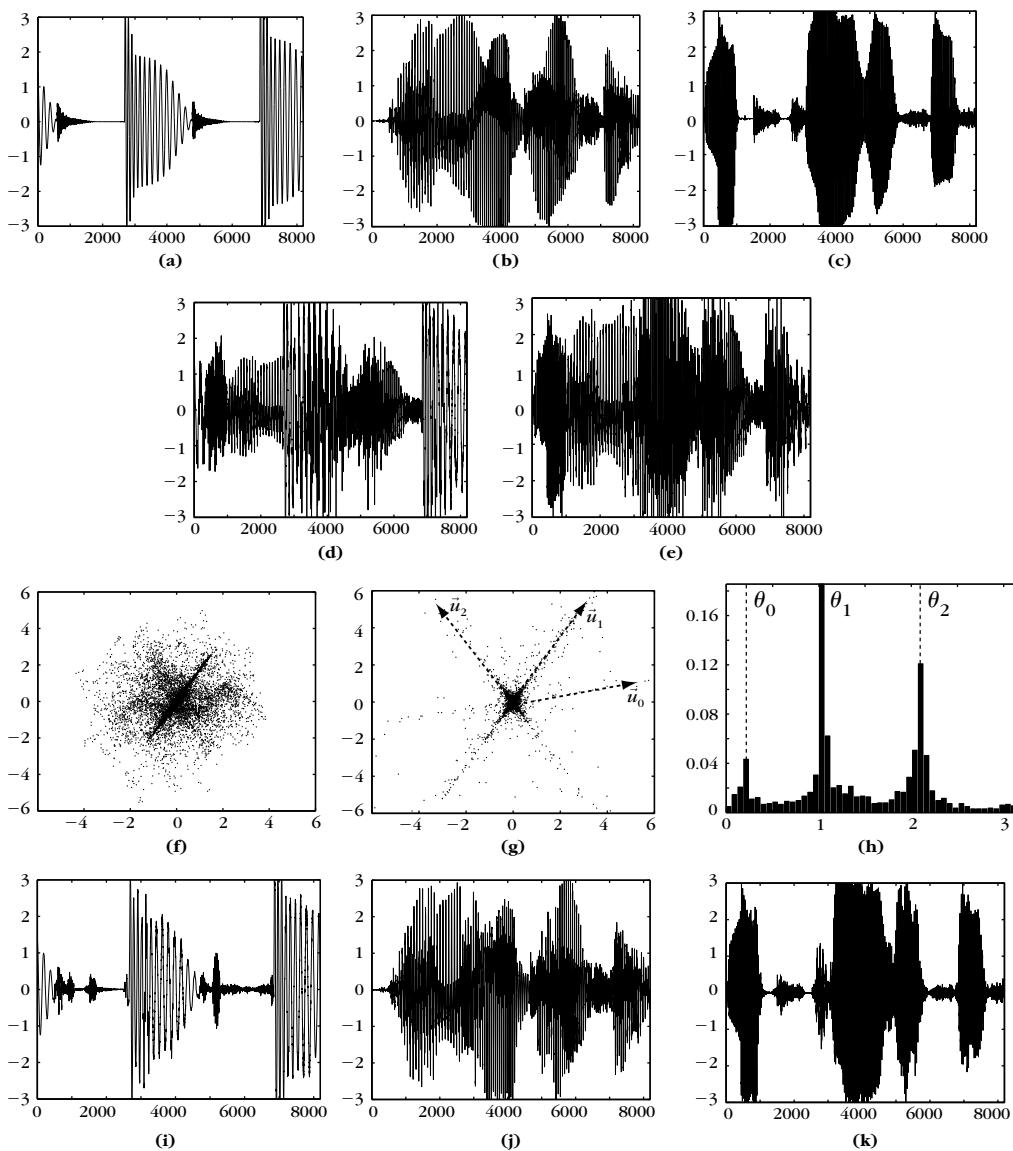
$$H(\vec{\theta}) = \sum_{p \in \tilde{\Lambda}} \rho(\|\vec{b}[p]\|) P(\vec{\theta} - \vec{\theta}[p]), \quad (13.91)$$

where  $P(\vec{x})$  is a Parzen window that regularizes the histogram. The weighting function  $\rho(\|\vec{b}[p]\|)$  reduces the influence of smaller-amplitude coefficients that are more affected by the noise. An appropriate weighting is

$$\rho(\|\vec{b}[p]\|) = \|\vec{b}[p]\|^2,$$

but other weighting schemes are also possible [493]. The Parzen window  $P(\vec{x})$  is typically separable along the  $K - 1$  directions. The histogram  $H(\vec{\theta})$  is sampled along a  $K - 1$  dimensional array at intervals proportional to the window size in each direction. The number of sources and directions of the sources are identified as local maxima of the histogram (13.91). Other classification algorithms such as  $K$ -mean algorithms may also be used to identify the mixing directions  $\vec{\theta}_s$ .

Figure 13.13 shows an example of stereo sound separation with  $K = 2$  measurements for  $S = 3$  audio sources  $f_s[n]$  that are shown in the top now. The cloud of  $N$  vectors  $\vec{Y}[n] \in \mathbb{R}^2$  is shown in the figure. Since the sources  $f_s[n]$  are not sparse in time, this cloud has no preferential direction. In this example, the dictionary  $\mathcal{D}$  is a local cosine orthonormal basis over windows of constant time duration, defined in Section 8.4.3. Figure 13.13(g) gives the cloud of local cosine coefficient vectors  $\{\vec{b}[p]\}_{p \in \Gamma}$  in  $\mathbb{R}^2$ . This cloud is clearly elongated along 3 preferential directions corresponding to the directions of the  $S = 3$  mixing vectors  $\vec{u}_s$ . All cosine coefficients have been kept. As a consequence, a large group of many small noisy vectors is at the center of the cloud. Figure 13.13(h) shows an angle histogram  $H(\theta)$  computed with a Parzen window  $P(\theta)$ , which is the indicator of an interval, with

**FIGURE 13.13**

Example of blind-source separation of  $S = 3$  sound sources from  $K = 2$  measurements.

- (a, b, c) Sources of  $f_0$ ,  $f_1$ , and  $f_2$ .
- (d, e) Measurements  $Y_0$  and  $Y_1$ .
- (f) Cloud  $\{\vec{Y}[n]\}_n$  in  $\mathbb{R}^2$ .
- (g) Cloud  $\{\vec{b}[p]\}_p$  in  $\mathbb{R}^2$ .
- (h) Angle histogram  $H(\theta)$ .
- (i, j, k) Estimations  $\tilde{F}_0$ ,  $\tilde{F}_1$ , and  $\tilde{F}_2$ .

a weighting  $\rho(\|\vec{b}[p]\|)$  that is a thresholding that eliminates the smallest-amplitude coefficients and keeps the 5% vectors of largest norm  $\|\vec{b}[p]\|$ . This histogram exhibits three local maxima that correspond to the  $S = 3$  mixing directions  $\theta_s$ . The identification of mixing directions is clearly more difficult when they are close. It is then particularly important to improve the separation of the source approximation supports by decomposing the measurements over richer redundant dictionaries.

### **Improved Separation with Pursuits in Redundant Dictionaries**

Making sure that coefficient vectors  $\vec{b}[p]$  of a large norm are mostly influenced by a single source requires us to build sparse source representations or to construct a dictionary where different sources have a tendency to choose different approximation vectors. When a drum plays with a guitar, the impulsive sounds are easy to discriminate from the guitar's narrow harmonics. In a windowed Fourier dictionary or a local cosine basis where windows sizes are chosen a priori, these sounds may be difficult to discriminate. Indeed, they both occur at the same time and the impulsive sounds of the drum have a spread frequency that overlaps the guitar frequencies. To clearly separate these sounds, it is necessary to use a larger and redundant dictionary including time-frequency atoms of different scales. In the multiscale Gabor dictionary  $\mathcal{D}_\Delta$  in (12.78), impulsive sounds are better represented by narrow Gabor atoms, where the guitar harmonics have a more sparse representation with elongated atoms having a better frequency resolution. Redundant dictionaries improve source separation by providing more sparse approximations and by approximating the sources with different types of atoms depending on their properties [101]. Section 12.5.3 gives an example where edges and textures are separated in an image by separating the wavelet and local cosine vectors selected in a large dictionary.

A sparse representation of  $\vec{Y}$  is computed in a redundant dictionary  $\mathcal{D}$  with multichannel extensions of pursuit algorithms, described in Section 12.6. We want to define a sparse representation of the measurement vector  $\vec{Y}$  with coefficients  $\vec{b}[p]$ , which are related to the mixing directions by the same equation as (13.86):

$$\vec{b}[p] = \sum_{s=0}^{S-1} a_s[p] \vec{u}_s + \vec{w}[p] \quad \text{for } p \in \tilde{\Lambda}. \quad (13.92)$$

For this purpose, after selecting the dictionary vectors  $\{\phi_p\}_{p \in \tilde{\Lambda}}$  with a multichannel pursuit on  $\vec{Y}$ , pursuit coefficients are computed with an orthogonal projection. A matching pursuit or an  $\mathbf{l}^1$  pursuit are therefore followed by a backprojection, described in Section 12.6. It computes the orthogonal projection of  $\vec{Y}$  on the space  $\mathbf{V}_{\tilde{\Lambda}}$  generated by  $\{\phi_p\}_{p \in \tilde{\Lambda}}$ .

The resulting coefficients are inner products with the dual frame  $\{\tilde{\phi}_{p,\tilde{\Lambda}}\}_{p \in \tilde{\Lambda}}$ :

$$\vec{b}[p] = \langle \vec{Y}, \tilde{\phi}_{p,\tilde{\Lambda}} \rangle, \quad \vec{w}[p] = \langle \vec{W}, \tilde{\phi}_{p,\tilde{\Lambda}} \rangle, \quad \text{and} \quad a_s[p] = \langle f_s, \tilde{\phi}_{p,\tilde{\Lambda}} \rangle \quad \text{for } p \in \tilde{\Lambda}. \quad (13.93)$$

Equation (13.92) then results from the inner product of the measurement vector equation (13.84) with  $\phi_{p,\tilde{\Lambda}}$ . The identification of the source directions  $\vec{u}_s / \|\vec{u}_s\|$  proceeds as previously described, by using the angle histogram (13.91).

### 13.5.2 Source Separation

Let us suppose that we know the mixing vectors  $\vec{u}_s$ . They are either provided by some a priori knowledge on the mixing system, or they are calculated with the identification algorithm previously described. The measurement vector  $\vec{Y}[n]$  is represented by coefficient vectors  $\vec{b}[p]$  that are either computed by projecting  $\vec{Y}$  on dictionary elements as in (13.86) or with a pursuit in a redundant dictionary as in (13.92):

$$\vec{b}[p] = \sum_{s=0}^{S-1} a_s[p] \vec{u}_s + \vec{w}[p] \quad \text{for } p \in \tilde{\Lambda}. \quad (13.94)$$

#### Cone Classification

A simple but effective masking algorithm introduced by Jourjine, Rickard, and Yilmaz [321, 493] divides the space  $\mathbb{R}^K$  into cones  $C_s$  corresponding to vectors that have directions that are the closest to each mixing direction  $\vec{u}_s/\|\vec{u}_s\|$ ,

$$C_s = \left\{ \vec{c} \in \mathbb{R}^K : s = \underset{0 \leq s' < S}{\operatorname{argmax}} \frac{|\langle \vec{c}, \vec{u}_{s'} \rangle|}{\|\vec{u}_{s'}\|} \right\},$$

where  $\langle \vec{a}, \vec{b} \rangle$  is the usual inner product in  $\mathbb{R}^K$ .

Each source is estimated by projecting  $\vec{b}[p]$  over the mixing direction of the cone  $C_{s_0}$  where it belongs:

$$\tilde{a}_s[p] = \begin{cases} \langle \vec{b}[p], \vec{u}_{s_0} \rangle / \|\vec{u}_{s_0}\| & \text{if } s_0 = \underset{0 \leq s' < S}{\operatorname{argmax}} |\langle \vec{b}[p], \vec{u}_{s'} \rangle| / \|\vec{u}_{s'}\| \\ 0 & \text{otherwise} \end{cases}. \quad (13.95)$$

If the dictionary  $\mathcal{D}$  is an orthogonal basis or a tight frame and the  $\vec{b}[p]$  have been computed with a decomposition operator (13.85), then a signal estimation is recovered with

$$\tilde{F}_s = \sum_{p \in \Lambda} \tilde{a}_s[p] \phi_p. \quad (13.96)$$

If  $\mathcal{D}$  is a redundant dictionary and the  $\vec{b}[p]$  are calculated with a dual family in (13.93), then the reconstruction formula (13.96) remains valid.

As previously explained,  $\tilde{F}_s$  is an estimator of  $f_s$  up to the unknown multiplicative constant  $\|\vec{u}_s\|$ . Figure 13.13 (i, j, k) display the  $S = 3$  estimated sources  $\tilde{F}_s$  computed from orthogonal local cosine coefficients  $\vec{b}[n]$  with the cone classification (13.96). The source directions are calculated with the histogram shown in Figure 13.13(h).

#### Source Demixing with Pursuits

When the number of sources  $S$  becomes relatively large, it is more likely that a coefficient  $\vec{b}[p]$  is the superposition of several nonnegligible source coefficients  $a_s[p]$ . As shown by Zibulevsky et al. [117, 498, 499] and analyzed by Gribonval and

Nielsen [280], these sources can still be identified with a pursuit algorithm that finds a sparse approximation of  $\vec{b}[p]$  in the dictionary of normalized mixing directions  $\mathcal{D}_u = \{\vec{u}_s / \|\vec{u}_s\|\}_{0 \leq s < S}$ :

$$\vec{b}_{\Lambda_p}[p] = \sum_{s \in \Lambda_p} \tilde{a}_s[p] \frac{\vec{u}_s}{\|\vec{u}_s\|}.$$

For a fixed  $p$ ,  $\tilde{a}_s[p] \neq 0$  only if  $s \in \Lambda_p$  and a source estimation is recovered with (13.96).

The coefficient classification (13.95) can be interpreted as a first step of a matching pursuit that projects  $\vec{b}[p]$  on the direction  $\vec{u}_{s_0} / \|\vec{u}_{s_0}\|$  of the best match. If the residue

$$R^1 \vec{b}[p] = \vec{b}[p] - \frac{\langle \vec{b}[p], \vec{u}_{s_0} \rangle}{\|\vec{u}_{s_0}\|} \vec{u}_{s_0}$$

is large  $\|R^1 \vec{b}[p]\| \geq T$ , then it is further decomposed by finding a next direction  $\vec{u}_{s_1} / \|\vec{u}_{s_1}\|$  in the dictionary  $\mathcal{D}_u$  of mixing directions, which best matches  $R^1 \vec{b}[p]$ , and so on. It is preferable to implement an orthogonal matching pursuit that orthogonalizes the projection directions and computes the decomposition coefficients  $\tilde{a}_{s_m}[p]$  from the orthogonalized residues  $R^m \vec{b}[p]$ , as explained in Section 12.3.2. The iterations can be stopped with a threshold  $T$  on the norm of the residue, which is typically proportional to the noise standard deviation  $(E\{\|\vec{w}[p]\|^2\})^{1/2}$ .

The sparse decomposition of  $\vec{b}[p]$  in the dictionary of mixing directions can also be implemented with an  $\mathbf{l}^1$  pursuit that computes

$$(\tilde{a}_s[p])_{0 \leq s < S} = \underset{(a_s)_s \in \mathbb{R}^S}{\operatorname{argmin}} \frac{1}{2} \|\vec{b}[p] - \sum_{s=0}^{S-1} a_s \frac{\vec{u}_s}{\|\vec{u}_s\|}\|^2 + T \sum_{s=0}^{S-1} |a_s|.$$

This minimization can be solved with the iterative thresholding algorithm in Section 12.4.3.

## 13.6 EXERCISES

For this chapter's exercises, see the Web site at <http://wavelet-tour.com>.

# Mathematical Complements

Important mathematical concepts are reviewed without proof. Sections A.1 through A.5 we present results of real and complex analysis, including properties of Hilbert spaces, bases, and linear operators [59]. Random vectors and Dirac distributions are covered in the last two sections.

---

## A.1 FUNCTIONS AND INTEGRATION

Analog signals are modeled by measurable functions. We first give the main theorems of Lebesgue integration. A function  $f$  is said to be *integrable* if  $\int_{-\infty}^{+\infty} |f(t)| dt < +\infty$ . The space of integrable functions is written as  $L^1(\mathbb{R})$ . Two functions  $f_1$  and  $f_2$  are equal in  $L^1(\mathbb{R})$  if

$$\int_{-\infty}^{+\infty} |f_1(t) - f_2(t)| dt = 0.$$

This means that  $f_1(t)$  and  $f_2(t)$  can differ only on a set of points of measure 0. We say that they are *almost everywhere* equal.

The Fatou lemma (A.1) gives an inequality when taking a limit under the Lebesgue integral of positive functions.

**Lemma A.1:** *Fatou.* Let  $\{f_n\}_{n \in \mathbb{N}}$  be a family of positive functions  $f_n(t) \geq 0$ . If  $\lim_{n \rightarrow +\infty} f_n(t) = f(t)$  almost everywhere, then

$$\int_{-\infty}^{+\infty} f(t) dt \leq \lim_{n \rightarrow +\infty} \int_{-\infty}^{+\infty} f_n(t) dt.$$

The dominated convergence theorem (A.1) supposes the existence of an integrable upper bound to obtain an equality when taking a limit under a Lebesgue integral.

**Theorem A.1:** *Dominated Convergence.* Let  $\{f_n\}_{n \in \mathbb{N}}$  be a family such that almost everywhere  $\lim_{n \rightarrow +\infty} f_n(t) = f(t)$ . If

$$\forall n \in \mathbb{N} \quad |f_n(t)| \leq g(t) \quad \text{and} \quad \int_{-\infty}^{+\infty} g(t) dt < +\infty, \tag{A.1}$$

then  $f$  is integrable and

$$\int_{-\infty}^{+\infty} f(t) dt = \lim_{n \rightarrow +\infty} \int_{-\infty}^{+\infty} f_n(t) dt.$$

The Fubini theorem (A.2) gives a sufficient condition for inverting the order of integrals in multidimensional integrations.

**Theorem A.2:** *Fubini.* If  $\int_{-\infty}^{+\infty} \left( \int_{-\infty}^{+\infty} |f(x_1, x_2)| dx_1 \right) dx_2 < +\infty$ , then

$$\begin{aligned} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x_1, x_2) dx_1 dx_2 &= \int_{-\infty}^{+\infty} \left( \int_{-\infty}^{+\infty} f(x_1, x_2) dx_1 \right) dx_2 \\ &= \int_{-\infty}^{+\infty} \left( \int_{-\infty}^{+\infty} f(x_1, x_2) dx_2 \right) dx_1. \end{aligned}$$

### Convexity

A function  $f(t)$  is said to be *convex* if for all  $p_1, p_2 > 0$  with  $p_1 + p_2 = 1$  and all  $(t_1, t_2) \in \mathbb{R}^2$ ,

$$f(p_1 t_1 + p_2 t_2) \leq p_1 f(t_1) + p_2 f(t_2).$$

The function  $-f$  satisfies the reverse inequality and is said to be *concave*. If  $f$  is convex, then the Jensen inequality generalizes this property for any  $p_k \geq 0$  with  $\sum_{k=1}^K p_k = 1$  and any  $t_k \in \mathbb{R}$ :

$$f \left( \sum_{k=1}^K p_k t_k \right) \leq \sum_{k=1}^K p_k f(t_k). \quad (\text{A.2})$$

Theorem A.3 relates the convexity to the sign of the second-order derivative.

**Theorem A.3.** If  $f$  is twice differentiable, then  $f$  is convex if and only if  $f''(t) \geq 0$  for all  $t \in \mathbb{R}$ .

The notion of convexity also applies to sets  $\Omega \subset \mathbb{R}^n$ . This set is convex if for all  $p_1, p_2 > 0$  with  $p_1 + p_2 = 1$  and all  $(x_1, x_2) \in \Omega^2$ , then  $p_1 x_1 + p_2 x_2 \in \Omega$ . If  $\Omega$  is not convex, then its convex hull is defined as the smallest convex set that includes  $\Omega$ .

## A.2 BANACH AND HILBERT SPACES

### Banach Space

Signals are often considered as vectors. To define a distance, we work within a vector space  $\mathbf{H}$  that admits a norm. A norm satisfies the following properties:

$$\forall f \in \mathbf{H}, \quad \|f\| \geq 0 \quad \text{and} \quad \|f\| = 0 \Leftrightarrow f = 0, \quad (\text{A.3})$$

$$\forall \lambda \in \mathbb{C} \quad \|\lambda f\| = |\lambda| \|f\|, \quad (\text{A.4})$$

$$\forall f, g \in \mathbf{H}, \quad \|f + g\| \leq \|f\| + \|g\|. \quad (\text{A.5})$$

With such a norm, the convergence of  $\{f_n\}_{n \in \mathbb{N}}$  to  $f$  in  $\mathbf{H}$  means that

$$\lim_{n \rightarrow +\infty} f_n = f \Leftrightarrow \lim_{n \rightarrow +\infty} \|f_n - f\| = 0.$$

To guarantee that we remain in  $\mathbf{H}$  when taking such limits, we impose a completeness property, using the notion of *Cauchy sequences*. A sequence  $\{f_n\}_{n \in \mathbb{N}}$  is a Cauchy sequence if for any  $\varepsilon > 0$ , if  $n$  and  $p$  are large enough, then  $\|f_n - f_p\| < \varepsilon$ . The space  $\mathbf{H}$  is said to be *complete* if every Cauchy sequence in  $\mathbf{H}$  converges to an element of  $\mathbf{H}$ .

### EXAMPLE A.1

For any integer  $p \geq 1$ , we define over discrete sequences  $f[n]$

$$\|f\|_p = \left( \sum_{n=-\infty}^{+\infty} |f[n]|^p \right)^{1/p}.$$

The space  $\ell^p = \{f : \|f\|_p < +\infty\}$  is a Banach space with the norm  $\|f\|_p$ .

### EXAMPLE A.2

The space  $\mathbf{L}^p(\mathbb{R})$  is composed of the measurable functions  $f$  on  $\mathbb{R}$  for which

$$\|f\|_p = \left( \int_{-\infty}^{+\infty} |f(t)|^p dt \right)^{1/p} < +\infty.$$

This integral defines a norm for  $p \geq 1$  and  $\mathbf{L}^p(\mathbb{R})$  is a Banach space, provided one identifies functions that are equal almost everywhere.

### Hilbert Space

Whenever possible, we work in a space that has an inner product to define angles and orthogonality. A *Hilbert space*  $\mathbf{H}$  is a Banach space with an inner product. The inner product of two vectors  $\langle f, g \rangle$  is linear with respect to its first argument:

$$\forall \lambda_1, \lambda_2 \in \mathbb{C}, \quad \langle \lambda_1 f_1 + \lambda_2 f_2, g \rangle = \lambda_1 \langle f_1, g \rangle + \lambda_2 \langle f_2, g \rangle. \quad (\text{A.6})$$

It has an Hermitian symmetry:

$$\langle f, g \rangle = \langle g, f \rangle^*.$$

Moreover,

$$\langle f, f \rangle \geq 0 \quad \text{and} \quad \langle f, f \rangle = 0 \Leftrightarrow f = 0.$$

One can verify that  $\|f\| = \langle f, f \rangle^{1/2}$  is a norm. The positivity (A.3) implies the Cauchy-Schwarz inequality:

$$|\langle f, g \rangle| \leq \|f\| \|g\|, \quad (\text{A.7})$$

which is an equality if and only if  $f$  and  $g$  are linearly dependent.

We write  $\mathbf{V}^\perp$  the orthogonal complement of a subspace  $\mathbf{V}$  of  $\mathbf{H}$ . All vectors of  $\mathbf{V}$  are orthogonal to all vectors of  $\mathbf{V}^\perp$  and  $\mathbf{V} \oplus \mathbf{V}^\perp = \mathbf{H}$ .

---

**EXAMPLE A.3**

An inner product between discrete signals  $f[n]$  and  $g[n]$  can be defined by

$$\langle f, g \rangle = \sum_{n=-\infty}^{+\infty} f[n]g^*[n].$$

It corresponds to an  $\ell^2(\mathbb{Z})$  norm:

$$\|f\|^2 = \langle f, f \rangle = \sum_{n=-\infty}^{+\infty} |f[n]|^2.$$

The space  $\ell^2(\mathbb{Z})$  of finite-energy sequences is therefore a Hilbert space. The Cauchy-Schwarz inequality (A.7) proves that

$$\left| \sum_{n=-\infty}^{+\infty} f[n]g^*[n] \right| \leq \left( \sum_{n=-\infty}^{+\infty} |f[n]|^2 \right)^{1/2} \left( \sum_{n=-\infty}^{+\infty} |g[n]|^2 \right)^{1/2}.$$


---

---

**EXAMPLE A.4**

Over analog signals  $f(t)$  and  $g(t)$ , an inner product can be defined by

$$\langle f, g \rangle = \int_{-\infty}^{+\infty} f(t)g^*(t) dt.$$

The resulting norm is

$$\|f\| = \left( \int_{-\infty}^{+\infty} |f(t)|^2 dt \right)^{1/2}.$$

The space  $L^2(\mathbb{R})$  of finite-energy functions is thus also a Hilbert space. In  $L^2(\mathbb{R})$ , the Cauchy-Schwarz inequality (A.7) is

$$\left| \int_{-\infty}^{+\infty} f(t)g^*(t) dt \right| \leq \left( \int_{-\infty}^{+\infty} |f(t)|^2 dt \right)^{1/2} \left( \int_{-\infty}^{+\infty} |g(t)|^2 dt \right)^{1/2}.$$

Two functions  $f_1$  and  $f_2$  are equal in  $L^2(\mathbb{R})$  if

$$\|f_1 - f_2\|^2 = \int_{-\infty}^{+\infty} |f_1(t) - f_2(t)|^2 dt = 0,$$

which means that  $f_1(t) = f_2(t)$  for almost all  $t \in \mathbb{R}$ .

---

---

## A.3 BASES OF HILBERT SPACES

### *Orthonormal Basis*

A family  $\{e_n\}_{n \in \mathbb{N}}$  of a Hilbert space  $\mathbf{H}$  is orthogonal if for  $n \neq p$ ,

$$\langle e_n, e_p \rangle = 0.$$

If for  $f \in \mathbf{H}$  there exists a sequence  $a[n]$  such that

$$\lim_{N \rightarrow +\infty} \|f - \sum_{n=0}^N a[n] e_n\| = 0,$$

then  $\{e_n\}_{n \in \mathbb{N}}$  is said to be an *orthogonal basis* of  $\mathbf{H}$ . The orthogonality implies that necessarily  $a[n] = \langle f, e_n \rangle / \|e_n\|^2$ , and we write

$$f = \sum_{n=0}^{+\infty} \frac{\langle f, e_n \rangle}{\|e_n\|^2} e_n. \quad (\text{A.8})$$

A Hilbert space that admits an orthogonal basis is said to be *separable*.

The basis is *orthonormal* if  $\|e_n\| = 1$  for all  $n \in \mathbb{N}$ . Computing the inner product of  $g \in \mathbf{H}$  with each side of (A.8) yields a Parseval equation for orthonormal bases:

$$\langle f, g \rangle = \sum_{n=0}^{+\infty} \langle f, e_n \rangle \langle g, e_n \rangle^*. \quad (\text{A.9})$$

When  $g = f$ , we get an energy conservation called the *Plancherel formula*:

$$\|f\|^2 = \sum_{n=0}^{+\infty} |\langle f, e_n \rangle|^2. \quad (\text{A.10})$$

The Hilbert spaces  $\ell^2(\mathbb{Z})$  and  $\mathbf{L}^2(\mathbb{R})$  are separable. For example, the family of translated Diracs  $\{e_n[k] = \delta[k - n]\}_{n \in \mathbb{Z}}$  is an orthonormal basis of  $\ell^2(\mathbb{Z})$ . Chapters 7 and 8 construct orthonormal bases of  $\mathbf{L}^2(\mathbb{R})$  with wavelets, wavelet packets, and local cosine functions.

### *Riesz Bases*

In an infinite-dimensional space, if we loosen up the orthogonality requirement, we must still impose a partial-energy equivalence to guarantee the stability of the basis. A family of vectors  $\{e_n\}_{n \in \mathbb{N}}$  is said to be a Riesz basis of  $\mathbf{H}$  if it is linearly independent and if there exist  $B \geq A > 0$  such that

$$\forall f \in \mathbf{H}, \quad A \|f\|^2 \leq \sum_{n=0}^{+\infty} |\langle f, e_n \rangle|^2 \leq B \|f\|^2. \quad (\text{A.11})$$

Section 5.1.2 proves that there exists a unique dual basis  $\{\tilde{e}_n\}_{n \in \mathbb{N}}$  characterized by biorthogonality relations

$$\forall (n, p) \in \mathbb{N}^2, \quad \langle e_n, \tilde{e}_p \rangle = \delta[n - p], \quad (\text{A.12})$$

and that satisfies

$$\forall f \in \mathbf{H}, \quad f = \sum_{n=0}^{+\infty} \langle f, \tilde{e}_n \rangle e_n = \sum_{n=0}^{+\infty} \langle f, e_n \rangle \tilde{e}_n.$$

## A.4 LINEAR OPERATORS

Classical signal-processing algorithms are mostly based on linear operators. An operator  $U$  from a Hilbert space  $\mathbf{H}_1$  to another Hilbert space  $\mathbf{H}_2$  is linear if

$$\forall \lambda_1, \lambda_2 \in \mathbb{C}, \quad \forall f_1, f_2 \in \mathbf{H}, \quad U(\lambda_1 f_1 + \lambda_2 f_2) = \lambda_1 U(f_1) + \lambda_2 U(f_2).$$

The null space and image spaces of  $U$  are defined by

$$\text{Null } U = \{h \in \mathbf{H}_1 : Uh = 0\} \quad \text{and} \quad \text{Im } U = \{g \in \mathbf{H}_2 : \exists h \in H_1, g = Uh\}.$$

### Supremum Norm

The supremum operator norm of  $U$  is defined by

$$\|U\|_S = \sup_{f \in \mathbf{H}_1} \frac{\|Uf\|}{\|f\|}. \quad (\text{A.13})$$

If this norm is finite, then  $U$  is continuous. Indeed,  $\|Uf - Ug\|$  becomes arbitrarily small if  $\|f - g\|$  is sufficiently small.

### Adjoint

The *adjoint* of  $U$  is the operator  $U^*$  from  $\mathbf{H}_2$  to  $\mathbf{H}_1$  such that for any  $f \in \mathbf{H}_1$  and  $g \in \mathbf{H}_2$ ,

$$\langle Uf, g \rangle = \langle f, U^*g \rangle.$$

The null and image spaces of adjoint operators are orthogonal complement:

$$\text{Null } U = (\text{Im } U^*)^\perp \quad \text{and} \quad \text{Im } U = (\text{Null } U^*)^\perp.$$

When  $U$  is defined from  $\mathbf{H}$  into itself, it is *self-adjoint* if  $U = U^*$ . It is also said to be *symmetric*.

A nonzero vector  $f \in \mathbf{H}$  is called an *eigenvector* if there exists an *eigenvalue*  $\lambda \in \mathbb{C}$  such that

$$Uf = \lambda f.$$

In a finite-dimensional Hilbert space (Euclidean space), a self-adjoint operator is always diagonalized by an orthogonal basis  $\{e_n\}_{0 \leq n < N}$  of eigenvectors

$$Ue_n = \lambda_n e_n.$$

When  $U$  is self-adjoint, the eigenvalues  $\lambda_n$  are real. For any  $f \in \mathbf{H}$ ,

$$Uf = \sum_{n=0}^{N-1} \langle Uf, e_n \rangle e_n = \sum_{n=0}^{N-1} \lambda_n \langle f, e_n \rangle e_n.$$

For any  $U$ , the operators  $U^*U$  and  $UU^*$  are self-adjoint and have the same eigenvalues. These eigenvalues are called *singular values* of  $U$ .

In an infinite-dimensional Hilbert space, the eigenvalues of symmetric operators are generalized by introducing the spectrum of the operator.

### **Orthogonal Projector**

Let  $\mathbf{V}$  be a subspace of  $\mathbf{H}$ . A *projector*  $P_V$  on  $\mathbf{V}$  is a linear operator that satisfies

$$\forall f \in \mathbf{H}, P_V f \in \mathbf{V} \quad \text{and} \quad \forall f \in \mathbf{V}, P_V f = f.$$

The projector  $P_V$  is *orthogonal* if

$$\forall f \in \mathbf{H}, \forall g \in \mathbf{V}, \langle f - P_V f, g \rangle = 0.$$

The properties in Theorem A.4 are often used.

**Theorem A.4.** If  $P_V$  is a projector on  $\mathbf{V}$ , then the following statements are equivalent:

1.  $P_V$  is orthogonal.
2.  $P_V$  is self-adjoint.
3.  $\|P_V\|_S = 1$ .
4.  $\forall f \in \mathbf{H}, \|f - P_V f\| = \min_{g \in \mathbf{V}} \|f - g\|$ .
5. If  $\{e_n\}_{n \in \mathbb{N}}$  is an orthogonal basis of  $\mathbf{V}$ , then

$$P_V f = \sum_{n=0}^{+\infty} \frac{\langle f, e_n \rangle}{\|e_n\|^2} e_n. \quad (\text{A.14})$$

6. If  $\{e_n\}_{n \in \mathbb{N}}$  and  $\{\tilde{e}_n\}_{n \in \mathbb{N}}$  are biorthogonal Riesz bases of  $\mathbf{V}$ , then

$$P_V f = \sum_{n=0}^{+\infty} \langle f, e_n \rangle \tilde{e}_n = \sum_{n=0}^{+\infty} \langle f, \tilde{e}_n \rangle e_n. \quad (\text{A.15})$$

### **Limit and Density Argument**

Let  $\{U_n\}_{n \in \mathbb{N}}$  be a sequence of linear operators from  $\mathbf{H}$  to  $\mathbf{H}$ . Such a sequence converges weakly to a linear operator  $U_\infty$  if

$$\forall f \in \mathbf{H}, \lim_{n \rightarrow +\infty} \|U_n f - U_\infty f\| = 0.$$

To find the limit of operators it is often preferable to work in a well-chosen subspace  $\mathbf{V} \subset \mathbf{H}$  that is dense. A space  $\mathbf{V}$  is *dense* in  $\mathbf{H}$  if for any  $f \in \mathbf{H}$  there exist  $\{f_m\}_{m \in \mathbb{N}}$  with  $f_m \in \mathbf{V}$  such that

$$\lim_{m \rightarrow +\infty} \|f - f_m\| = 0.$$

Theorem A.5 justifies this approach.

**Theorem A.5: Density.** Let  $\mathbf{V}$  be a dense subspace of  $\mathbf{H}$ . Suppose that there exists  $C$  such that  $\|U_n\|_S \leq C$  for all  $n \in \mathbb{N}$ . If

$$\forall f \in \mathbf{V}, \quad \lim_{n \rightarrow +\infty} \|U_n f - U_\infty f\| = 0,$$

then

$$\forall f \in \mathbf{H}, \quad \lim_{n \rightarrow +\infty} \|U_n f - U_\infty f\| = 0.$$

## A.5 SEPARABLE SPACES AND BASES

### Tensor Product

Tensor products are used to extend spaces of one-dimensional signals into spaces of multidimensional signals. A tensor product  $f_1 \otimes f_2$  between vectors of two Hilbert spaces  $\mathbf{H}_1$  and  $\mathbf{H}_2$  satisfies the following properties.

$$\text{Linearity: } \forall \lambda \in \mathbb{C}, \quad \lambda(f_1 \otimes f_2) = (\lambda f_1) \otimes f_2 = f_1 \otimes (\lambda f_2). \quad (\text{A.16})$$

$$\begin{aligned} \text{Distributivity: } (f_1 + g_1) \otimes (f_2 + g_2) &= (f_1 \otimes f_2) + (f_1 \otimes g_2) \\ &\quad + (g_1 \otimes f_2) + (g_1 \otimes g_2). \end{aligned} \quad (\text{A.17})$$

This tensor product yields a new Hilbert space  $\mathbf{H} = \mathbf{H}_1 \otimes \mathbf{H}_2$  that includes all vectors of the form  $f_1 \otimes f_2$  where  $f_1 \in \mathbf{H}_1$  and  $f_2 \in \mathbf{H}_2$ , as well as linear combinations of such vectors. An inner product in  $\mathbf{H}$  is derived from inner products in  $\mathbf{H}_1$  and  $\mathbf{H}_2$  by

$$\langle f_1 \otimes f_2, g_1 \otimes g_2 \rangle_{\mathbf{H}} = \langle f_1, g_1 \rangle_{\mathbf{H}_1} \langle f_2, g_2 \rangle_{\mathbf{H}_2}. \quad (\text{A.18})$$

### Separable Bases

Theorem A.6 proves that orthonormal bases of tensor product spaces are obtained with separable products of two orthonormal bases. It provides a simple procedure for transforming bases for one-dimensional signals into separable bases for multidimensional signals.

**Theorem A.6.** Let  $\mathbf{H} = \mathbf{H}_1 \otimes \mathbf{H}_2$ . If  $\{e_n^1\}_{n \in \mathbb{N}}$  and  $\{e_n^2\}_{n \in \mathbb{N}}$  are two Riesz bases, respectively, of  $\mathbf{H}_1$  and  $\mathbf{H}_2$ , then  $\{e_n^1 \otimes e_m^2\}_{(n,m) \in \mathbb{N}^2}$  is a Riesz basis of  $\mathbf{H}$ . If the two bases are orthonormal, then the tensor product basis is also orthonormal.

**EXAMPLE A.5**

A product of functions  $f \in \mathbf{L}^2(\mathbb{R})$  and  $g \in \mathbf{L}^2(\mathbb{R})$  defines a tensor product:

$$f(x_1)g(x_2) = f \otimes g(x_1, x_2).$$

Let  $\mathbf{L}^2(\mathbb{R}^2)$  be the space of  $h(x_1, x_2)$  such that

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} |h(x_1, x_2)|^2 dx_1 dx_2 < +\infty.$$

One can verify that  $\mathbf{L}^2(\mathbb{R}^2) = \mathbf{L}^2(\mathbb{R}) \otimes \mathbf{L}^2(\mathbb{R})$ . Theorem A.6 proves that if  $\{\psi_n(t)\}_{n \in \mathbb{N}}$  is an orthonormal basis of  $\mathbf{L}^2(\mathbb{R})$ , then  $\{\psi_{n_1}(x_1) \psi_{n_2}(x_2)\}_{(n_1, n_2) \in \mathbb{N}^2}$  is an orthonormal basis of  $\mathbf{L}^2(\mathbb{R}^2)$ .

**EXAMPLE A.6**

A product of discrete signals  $f \in \ell^2(\mathbb{Z})$  and  $g \in \ell^2(\mathbb{Z})$  also defines a tensor product:

$$f[n_1]g[n_2] = f \otimes g[n_1, n_2].$$

The space  $\ell^2(\mathbb{Z}^2)$  of images  $h[n_1, n_2]$  such that

$$\sum_{n_1=-\infty}^{+\infty} \sum_{n_2=-\infty}^{+\infty} |h[n_1, n_2]|^2 < +\infty$$

is also decomposed as a tensor product  $\ell^2(\mathbb{Z}^2) = \ell^2(\mathbb{Z}) \otimes \ell^2(\mathbb{Z})$ . Thus, orthonormal bases can be constructed with separable products.

## A.6 RANDOM VECTORS AND COVARIANCE OPERATORS

A class of signals can be modeled by a random process (random vector) with realizations that are the signals in the class. Finite discrete signals  $f$  are represented by a random vector  $Y$  where  $Y[n]$  is a random variable for each  $0 \leq n < N$ . For a review of elementary probability theory for signal processing, the reader may consult [53, 56].

### **Covariance Operator**

If  $p(x)$  is the probability density of a random variable  $X$ , the expected value is

$$E\{X\} = \int x p(x) dx,$$

and the variance is  $\sigma^2 = E\{|X - E\{X\}|^2\}$ . The covariance of two random variables  $X_1$  and  $X_2$  is

$$\text{Cov}(X_1, X_2) = E\left\{ (X_1 - E\{X_1\})(X_2 - E\{X_2\})^* \right\}. \quad (\text{A.19})$$

The covariance matrix of a random vector  $Y$  is composed of the  $N^2$  covariance values

$$R_Y[n, m] = \text{Cov}(Y[n], Y[m]).$$

It defines the covariance operator  $K_Y$ , which transforms any  $h[n]$  into

$$K_Y h[n] = \sum_{m=0}^{N-1} R_Y[n, m] h[m].$$

For any  $h$  and  $g$ ,

$$\langle Y, h \rangle = \sum_{n=0}^{N-1} Y[n] h^*[n] \quad \text{and} \quad \langle Y, g \rangle = \sum_{n=0}^{N-1} Y[n] g^*[n]$$

are random variables, and

$$\text{Cov}(\langle Y, h \rangle, \langle Y, g \rangle) = \langle K_Y g, h \rangle. \quad (\text{A.20})$$

Thus, the covariance operator specifies the covariance of linear combinations of the process values. If  $E\{Y[n]\} = 0$  for all  $0 \leq n < N$ , then  $E\{\langle Y, h \rangle\} = 0$  for all  $h$ .

### **Karhunen-Loève Basis**

The covariance operator  $K_Y$  is self-adjoint because  $R_Y[n, m] = R_Y^*[m, n]$  and positive because

$$\langle K_Y h, h \rangle = E\{|\langle Y, h \rangle - E\{\langle Y, h \rangle\}|^2\} \geq 0. \quad (\text{A.21})$$

This guarantees the existence of an orthogonal basis  $\{e_k\}_{0 \leq k < N}$  that diagonalizes  $K_Y$ :

$$K_Y e_k = \sigma_k^2 e_k.$$

This basis is called a *Karhunen-Loève basis* of  $Y$ , and the vectors  $e_k$  are the *principal directions*. The eigenvalues are the variances

$$\sigma_k^2 = \langle K_Y e_k, e_k \rangle = E\{|\langle Y, e_k \rangle - E\{\langle Y, e_k \rangle\}|^2\}. \quad (\text{A.22})$$

### **Wide-Sense Stationarity**

We say that  $Y$  is *wide-sense stationary* if

$$\text{Cov}(Y[n], Y[m]) = R_Y[n, m] = R_Y[n - m]. \quad (\text{A.23})$$

The covariance between two points depends only on the distance between these points. The operator  $K_Y$  is then a convolution with a kernel  $R_Y[k]$  that is defined for  $-N < k < N$ . A wide-sense stationary process is *circular stationary* if  $R_Y[n]$  is  $N$  periodic:

$$R_Y[n] = R_Y[N + n] \quad \text{for} \quad -N \leq n \leq 0. \quad (\text{A.24})$$

This condition implies that a periodic extension of  $Y[n]$  on  $\mathbb{Z}$  remains wide-sense stationary on  $\mathbb{Z}$ . The covariance operator  $K_Y$  of a circular stationary process is a discrete circular convolution. Section 3.3.1 proves that the eigenvectors of circular convolutions are the discrete Fourier vectors

$$\left\{ e_k[n] = \frac{1}{\sqrt{N}} \exp\left(\frac{i2\pi kn}{N}\right) \right\}_{0 \leq k < N}.$$

The discrete Fourier basis is therefore the Karhunen-Loëve basis of circular stationary processes. The eigenvalues (A.22) of  $K_Y$  are the discrete Fourier transform of  $R_Y$  and are called the *power spectrum*,

$$\sigma_k^2 = \hat{R}_Y[k] = \sum_{n=0}^{N-1} R_Y[n] \exp\left(\frac{-i2k\pi n}{N}\right). \quad (\text{A.25})$$

Theorem A.7 computes the power spectrum after a circular convolution.

**Theorem A.7.** Let  $Z$  be a wide-sense circular stationary random vector. The random vector  $Y[n] = Z \circledast h[n]$  is also wide-sense circular stationary and its power spectrum is

$$\hat{R}_Y[k] = \hat{R}_Z[k] |\hat{h}[k]|^2. \quad (\text{A.26})$$

## A.7 DIRACS

Diracs are useful in making the transition from functions of a real variable to discrete sequences. Symbolic calculations with Diracs simplify computations, without worrying about convergence issues. This is justified by the theory of distributions [61, 64]. A Dirac  $\delta$  has a support reduced to  $t = 0$  and associates to any continuous function  $\phi$  its value at  $t = 0$ ,

$$\int_{-\infty}^{+\infty} \delta(t) \phi(t) dt = \phi(0). \quad (\text{A.27})$$

### Weak Convergence

A Dirac can be obtained by squeezing an integrable function  $g$  such that  $\int_{-\infty}^{+\infty} g(t) dt = 1$ . Let  $g_s(t) = s^{-1}g(s^{-1}t)$ . For any continuous function  $\phi$ ,

$$\lim_{s \rightarrow 0} \int_{-\infty}^{+\infty} g_s(t) \phi(t) dt = \phi(0) = \int_{-\infty}^{+\infty} \delta(t) \phi(t) dt. \quad (\text{A.28})$$

Thus, a Dirac can be formally defined as the limit  $\delta = \lim_{s \rightarrow 0} g_s$ , which must be understood in the sense of (A.28). This is called *weak convergence*. A Dirac is not a function since it is zero at  $t \neq 0$ , although its “integral” is equal to 1. The integral at the right of (A.28) is only a symbolic notation, which means that a Dirac applied to a continuous function  $\phi$  associates its value at  $t = 0$ .

General distributions are defined over the space  $C_0^\infty$  of *test functions* that are infinitely continuously differentiable with a compact support. A distribution  $d$  is a

linear form that associates to any  $\phi \in \mathbf{C}_0^\infty$  a value that is written as  $\int_{-\infty}^{+\infty} d(t)\phi(t)dt$ . It must also satisfy some weak continuity properties [61, 64] that we do not discuss here, and that are satisfied by a Dirac. Two distributions  $d_1$  and  $d_2$  are equal if

$$\forall \phi \in \mathbf{C}_0^\infty, \quad \int_{-\infty}^{+\infty} d_1(t) \phi(t) dt = \int_{-\infty}^{+\infty} d_2(t) \phi(t) dt. \quad (\text{A.29})$$

### Symbolic Calculations

The symbolic integral over a Dirac is a useful notation because it has the same properties as a usual integral, including change of variables and integration by parts. A translated Dirac  $\delta_\tau(t) = \delta(t - \tau)$  has a mass concentrated at  $\tau$  and

$$\int_{-\infty}^{+\infty} \phi(t) \delta(t - u) dt = \int_{-\infty}^{+\infty} \phi(t) \delta(u - t) dt = \phi(u).$$

This means that  $\phi \star \delta(u) = \phi(u)$ . Similarly,  $\phi \star \delta_\tau(u) = \phi(u - \tau)$ .

A Dirac can also be multiplied by a continuous function  $\phi$  and since  $\delta(t - \tau)$  is zero outside  $t = \tau$ , it follows that

$$\phi(t) \delta(t - \tau) = \phi(\tau) \delta(t - \tau).$$

The derivative of a Dirac is defined with an integration by parts. If  $\phi$  is continuously differentiable, then

$$\int_{-\infty}^{+\infty} \phi(t) \delta'(t) dt = - \int_{-\infty}^{+\infty} \phi'(t) \delta(t) dt = -\phi'(0).$$

The  $k$ th derivative of  $\delta$  is similarly obtained with  $k$  integrations by parts. It is a distribution that associates to  $\phi \in \mathbf{C}^k$ ,

$$\int_{-\infty}^{+\infty} \phi(t) \delta^{(k)}(t) dt = (-1)^k \phi^{(k)}(0).$$

The Fourier transform of  $\delta$  associates to any  $e^{-i\omega t}$  its value at  $t = 0$ :

$$\hat{\delta}(\omega) = \int_{-\infty}^{+\infty} \delta(t) e^{-i\omega t} dt = 1,$$

and after translation,  $\hat{\delta}_\tau(\omega) = e^{-i\tau\omega}$ . The Fourier transform of the Dirac comb  $c(t) = \sum_{n=-\infty}^{+\infty} \delta(t - nT)$  is therefore  $\hat{c}(\omega) = \sum_{n=-\infty}^{+\infty} e^{-inT\omega}$ . The Poisson formula (2.4) proves that

$$\hat{c}(\omega) = \frac{2\pi}{T} \sum_{k=-\infty}^{+\infty} \delta\left(\omega - \frac{2\pi k}{T}\right).$$

This distribution equality must be understood in the sense of (A.29).

# Bibliography

---

## BOOKS

- [1] A. Akansu and R. Haddad. *Multiresolution Signal Decomposition*. Academic Press, 1993.
- [2] A. Akansu and M. J. Smith, editors. *Subband and Wavelet Transforms*. Kluwer, 1995.
- [3] A. Aldroubi and M. Unser, editors. *Wavelets in Medicine and Biology*. CRC Press, 1996.
- [4] A. Antoniadis and G. Oppenheim, editors. *Wavelets and Statistics*. Springer-Verlag, 1995.
- [5] A. Arneodo, F. Argoul, E. Bacry, J. Elezgaray, and J. F. Muzy. *Ondelettes, Multifractales et Turbulences*. Diderot, 1995.
- [6] J. J. Benedetto and M. W. Frazier, editors. *Wavelets: Mathematics and Applications*. CRC Press, 1994.
- [7] S. M. Berman. *Sojourns and Extremes of Stochastic Processes*. Wadsworth, 1989.
- [8] B. Boashash, editor. *Time-Frequency Signal Analysis*. Wiley Halsted Press, 1992.
- [9] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, 1984.
- [10] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
- [11] C. S. Burrus and T. W. Parks. *DFT/FFT and Convolution Algorithms: Theory and Implementation*. John Wiley and Sons, 1985.
- [12] C. K. Chui, editor. *Wavelets: A Tutorial in Theory and Applications*. Academic Press, 1992.
- [13] M. Cannone. *Ondelettes, Paraproducts et Navier-Stokes*. Diderot, 1995.
- [14] W. K. Chen. *Passive and Active Filters*. John Wiley and Sons, 1986.
- [15] C. K. Chui. *An Introduction to Wavelets*. Academic Press, 1992.
- [16] A. Cohen and R. D. Ryan. *Wavelets and Multiscale Signal Processing*. Chapman and Hall, 1995.
- [17] J. M. Combes, A. Grossmann, and P. Tchamitchian, editors. *Wavelets Time-Frequency Methods and Phase Space*. Springer-Verlag, 1989.
- [18] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 1991.

- [19] I. Daubechies. *Ten Lectures on Wavelets*. SIAM, 1992.
- [20] R. DeVore and G. Lorentz. *Constructive Approximation*. Vol. 303 of *Comprehensive Studies in Mathematics*. Springer-Verlag, 1993.
- [21] D. E. Dudgeon and R. M. Mersereau. *Multidimensional Digital Signal Processing*. Prentice-Hall, 1984.
- [22] P. Durka. *Matching Pursuit and Unification in EEG Analysis*. Lavoisier, 2007.
- [23] H. Dym and H. P. McKean. *Fourier Series and Integrals*. Academic Press, 1972.
- [24] M. B. Ruskai et al., editor. *Wavelets and Their Applications*. Jones and Bartlett, 1992.
- [25] F. Feder. *Fractals*. Pergamon, 1988.
- [26] P. Flandrin. *Temps-Fréquence*. Hermès, 1993.
- [27] A. Gersho and R. M. Gray. *Vector Quantization and Signal Compression*. Kluwer/Academic Publishers, 1992.
- [28] P. E. Gill, W. Murray, and M. H. Wright. *Numerical Linear Algebra and Optimization*. Addison-Wesley, 1991.
- [29] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 1989.
- [30] E. Hernández and G. Weiss. *A First Course on Wavelets*. CRC Press, 1996.
- [31] B. Burke Hubbard. *The World According to Wavelets*. A. K. Peters, 1996.
- [32] S. Jaffard and Y. Meyer. *Wavelet Methods for Pointwise Regularity and Local Oscillations of Functions*, Vol. 123. American Mathematical Society, 1996.
- [33] A. K. Jain. *Fundamentals of Digital Image Processing*. Prentice-Hall, 1989.
- [34] N. J. Jayant and P. Noll. *Digital Coding of Waveforms*. Prentice-Hall, 1984.
- [35] F. John. *Partial Differential Equations*. Springer-Verlag, 1975.
- [36] G. Kaiser. *A Friendly Guide to Wavelets*. Birkhäuser, 1994.
- [37] G. Kanizsa. *Organization in Vision*. Praeger Scientific, 1979.
- [38] M. Barlaud, editor. *Wavelets in Image Communication*. Elsevier, 1995.
- [39] F. J. MacWilliams and N. J. A. Sloane. *The Theory of Error-Correcting Codes*, Vol. 16 of North-Holland Mathematical Library. North-Holland, 1977.
- [40] H. S. Malvar. *Signal Processing with Lapped Transforms*. Artech House, 1992.
- [41] B. B. Mandelbrot. *The Fractal Geometry of Nature*. W. H. Freeman and Co., 1982.
- [42] D. Marr. *Vision*. W. H. Freeman and Co., 1982.
- [43] Y. Meyer. *Ondelettes et Algorithmes Concurrents*. Hermann, 1992.
- [44] Y. Meyer. *Wavelets and Operators: Advanced Mathematics*. Cambridge University Press, 1992.

- [45] Y. Meyer. *Wavelets:Algorithms and Applications*. Translated and revised by R. D. Ryan. SIAM, 1993.
- [46] Y. Meyer. *Wavelets, Vibrations and Scalings*. CRM, Université de Montréal, Cours de la chaire Aisenstadt, 1997.
- [47] Y. Meyer. *Oscillating Patterns in Image Processing and Nonlinear Evolution Equations*. American Mathematical Society, 2001.
- [48] Y. Meyer and S. Roques, editors. *Progress in Wavelet Analysis and Applications*. Frontières, 1993.
- [49] H. J. Nussbaumer. *Fast Fourier Transform and Convolution Algorithms*. Springer-Verlag, 1982.
- [50] A. V. Oppenheim, A. S. Willsky, and I. T. Young. *Signals and Systems*. Prentice-Hall, 1997.
- [51] A. V. Oppenheim and R. W. Shafer. *Discrete-Time Signal Processing*. Prentice-Hall, 1989.
- [52] P. G. Lemarié, editor. *Les Ondelettes en 1989*. Lecture Notes in Mathematics, no. 1438. Springer-Verlag, 1990.
- [53] A. Papoulis. *Probability, Random Variables and Stochastic Processes*, 2nd edition. McGraw-Hill, 1984.
- [54] A. Papoulis. *The Fourier Integral and Its Applications*, 2nd edition. McGraw-Hill, 1987.
- [55] A. Papoulis. *Signal Analysis*. McGraw-Hill, 1988.
- [56] B. Porat. *Digital Processing of Random Signals: Theory and Method*. Prentice-Hall, 1994.
- [57] M. B. Priestley. *Spectral Analysis and Time Series*. Academic Press, 1981.
- [58] A. Rosenfeld, editor. *Multiresolution Techniques in Computer Vision*. Springer-Verlag, 1984.
- [59] W. Rudin. *Real and Complex Analysis*. McGraw Hill, 1987.
- [60] D. J. Sakrison. *Communication Theory: Transmission of Waveforms and Digital Information*. John Wiley and Sons, 1968.
- [61] L. Schwartz. *Théorie Des Distributions*. Hermann, 1970.
- [62] J. J. Slotine and W. Li. *Applied Nonlinear Control*. Prentice-Hall, 1991.
- [63] G. Strang and T. Nguyen. *Wavelets and Filter Banks*. Wellesley-Cambridge Press, 1996.
- [64] R. Strichartz. *A Guide to Distribution Theory and Fourier Transforms*. CRC Press, 1994.
- [65] D. Taubman and M. Marcellin. *JPEG2000: Image Compression Fundamentals, Standards and Practice*. Kluwer Academic Publishers, 2001.

- [66] B. Torrésani. *Analyse Continue par Ondelettes*. CNRS Editions, 1995.
- [67] H. Triebel. *Theory of Function Spaces*. Birkhäuser Verlag, 1992.
- [68] P. P. Vaidyanathan. *Multirate Systems and Filter Banks*. Prentice-Hall, 1993.
- [69] M. Vetterli and J. Kovacevic. *Wavelets and Subband Coding*. Prentice-Hall, 1995.
- [70] H. Weyl. *The Theory of Groups and Quantum Mechanics*. Dutton, 1931.
- [71] M. V. Wickerhauser. *Adapted Wavelet Analysis from Theory to Software*. A. K. Peters, 1994.
- [72] J. W. Woods, editor. *Subband Image Coding*. Kluwer, 1991.
- [73] G. W. Wornell. *Signal Processing with Fractals: A Wavelet-Based Approach*. Prentice-Hall, 1995.
- [74] Z. Zhang. *Matching Pursuit*. Ph.D Thesis, Courant Institute, New York University, 1993.
- [75] W. P. Ziemer. *Weakly Differentiable Functions*. Springer-Verlag, 1989.

## ARTICLES

- [76] E. H. Adelson, E. Simoncelli, and R. Hingorani. Orthogonal pyramid transforms for image coding. In *Proc. SPIE*, 845:50–58, October 1987.
- [77] J. C. Aguilar and J. B. Goodman. Anisotropic mesh refinement for finite element methods based on error reduction. *J. Comput. Appl. Math.*, 193(2):497–515, 2006.
- [78] M. Aharon, M. Elad, and A. M. Bruckstein. The k-svd: An algorithm for designing of overcomplete dictionaries for sparse representation. *IEEE Trans. Signal Proc.*, 2006.
- [79] A. N. Akansu, R. A. Haddad, and H. Caglar. The binomial QMF-wavelet transform for multiresolution signal decomposition. *IEEE Trans. Signal Proc.*, 40, 1992.
- [80] O. K. Al Shaykh, E. Miloslavsky, T. Nomura, R. Neff, and A. Zakhor. Video compression using matching pursuits. *IEEE Trans. Circ. Syst. Video Technol.*, 9(1):123–143, 1999.
- [81] A. Aldroubi and H. Feichtinger. Complete iterative reconstruction algorithms for irregularly sampled data in spline-like spaces. *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, April 1997.
- [82] A. Aldroubi and M. Unser. Families of multiresolution and wavelet spaces with optimal properties. *Numer. Funct. Anal. Optim.*, 14:417–446, 1993.
- [83] J. Aloimonos and A. Rosenfeld. Computer vision. *Science*, 253:1249–1253, 1991.

- [84] B. Alpert, G. Beylkin, R. Coifman, and V. Rokhlin. Wavelet-like bases for the fast solutions of second-kind integral equations. *SIAM J. Sci. Comput.*, 14(1):159–184, 1993.
- [85] R. Ansari and C. Guillemont. Exact reconstruction filter banks using diamond FIR filters. In *Proc. Bilkent Intl. Conf.*, pp. 1412–1424, July 1990.
- [86] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies. Image coding using wavelet transform. *IEEE Trans. Image Proc.*, 2(1):205–220, 1992.
- [87] A. Averbuch, G. Aharoni, R. Coifman, and M. Israeli. Local cosine transform—a method for the reduction of the blocking effect in JPEG. *J. Math. Imaging Vision*, 3:7–38, 1993.
- [88] A. Averbuch, D. Lazar, and M. Israeli. Image compression using wavelet decomposition. *IEEE Trans. Image Proc.*, 5(1):4–15, 1996.
- [89] P. M. Aziz, H. V. Sorensen, and J. Van Der Spiegel. An overview of sigma-delta converters. *IEEE Sig. Proc. Mag.*, 13(1):61–84, 1996.
- [90] I. Babuška and A. K. Aziz. On the angle condition in the finite element method. *SIAM J. Numer. Anal.*, 13(2):214–226, 1976.
- [91] E. Bacry, J. F. Muzy, and A. Arneodo. Singularity spectrum of fractal signals: exact results. *J. Stat. Phys.*, 70(3/4):635–674, 1993.
- [92] R. Bajcsy. Computer description of textured surfaces. *JCAI*, August 1973.
- [93] R. Balian. Un principe d'incertitude en théorie du signal ou en mécanique quantique. *C. R. Acad. Sci. Paris, Série*, 2:292, 1981.
- [94] C. Ballester, M. Bertalmio, V. Caselles, G. Sapiro, and J. Verdera. Filling-in by joint interpolation of vector fields and gray levels. *IEEE Trans. Image Proc.*, 10(8):1200–1211, 2001.
- [95] R. Baraniuk and M. Wakin. Random projections of smooth manifolds. In *Foundations of Computational Mathematics*, 2007.
- [96] R. G. Baraniuk, V. Cevher, M. Duarte, and C. Hegde. Model-based compressive sensing. *IEEE Trans. Inf. Theory*, 2008.
- [97] A. R. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113:301–415, 1999.
- [98] M. Basseville, A. Benveniste, and A. S. Willsky. Multiscale autoregressive processes: Shur-Levinson parametrizations. *IEEE Trans. Signal Process.*, 1992.
- [99] G. Battle. A block spin construction of ondelettes. Part I: Lemarié functions. *Comm. Math. Phys.*, 110:601–615, 1987.
- [100] J. Bect, L. Blanc Féraud, G. Aubert, and A. Chambolle. A  $\ell_1$ -unified variational framework for image restoration. In *Proc. of ECCV04*, IV:1–13. Springer-Verlag, 2004.

- [101] L. Benaroya, F. Bimbot, and R. Gribonval. Audio source separation with a single sensor. *IEEE Trans. Audio, Speech & Lang. Process.*, 14(1):191–199, 2006.
- [102] J. J. Benedetto. Irregular sampling and frames. In C. K. Chui, editor, *Wavelets: A Tutorial in Theory and Applications*. Academic Press, 1992.
- [103] B. Benichou and N. Saito. *Sparsity vs. Statistical Independence in Adaptive Signal Representations: A Case Study of the Spike Process*, Vol. 10, pp. 225–257. Academic Press, 2003.
- [104] E. van den Berg and M. P. Friedlander. Probing the pareto frontier for basis pursuit solutions. Tech. Rep. TR-2008-01, Department of Computer Science, January 2008.
- [105] F. Bergeaud and S. Mallat. Matching pursuit:Adaptative representations of images. *Comput.Appl.Math.*, 15(2):97–109, 1996.
- [106] T. Berger and J. O. Stromberg. Exact reconstruction algorithms for the discrete wavelet transform using spline wavelets. *J.Appl.Comput.Harmon.Anal.*, 2:392–397, 1995.
- [107] Z. Berman and J. S. Baras. Properties of the multiscale maxima and zero-crossings representations. *IEEE Trans. Signal Proc.*, 41(12):3216–3231, 1993.
- [108] C. Bernard. Discrete wavelet analysis: A new framework for fast optical flow computations. *J.Appl.Comput.Harmon.Anal.*, 11(1):32–63, 2001.
- [109] M. Berouti, R. Schwartz, and J. Makhoul. Enhancement of speech corrupted by acoustic noise. *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 4:208–211, 1979.
- [110] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester. Image inpainting. *Siggraph 2000*, pp. 417–424, 2000.
- [111] M. Bertero. Linear inverse and ill-posed problems. In *Advances in Electronics and Electron Physics*. Academic Press, 1989.
- [112] G. Beylkin, R. Coifman, and V. Rokhlin. Fast wavelet transforms and numerical algorithms. *Comm. Pure Appl. Math.*, 44:141–183, 1991.
- [113] P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.*, 2008.
- [114] L. Birgé and P. Massart. Gaussian model selection. *J. Eur. Math. Soc.*, 3:203–268, 2001.
- [115] T. Blumensath and M. E. Davies. Gradient pursuits. *IEEE Trans. Signal Proc.*, 56(6):2370–2382, 2008.
- [116] J. Bobin, J. L. Starck, J. Fadili, and Y. Moudden. Sparsity and morphological diversity in blind source separation. *IEEE Trans. Image Proc.*, 16(11):2662–2674, 2007.
- [117] P. Bofill and M. Zibulevsky. Underdetermined blind source separation using sparse representations. *Signal Process.*, 81(11):2353–2362, 2001.

- [118] J. M. Bony. Two-microlocalization and propagation of singularities for semilinear hyperbolic equations. In *Proc. of Tanaguchi Symp., HERT Katata*, pp. 11–49, October 1984.
- [119] A. C. Bovik, N. Gopal, T. Emmoth, and A. Restrepo. Localized measurement of emergent image frequencies by Gabor wavelets. *IEEE Trans. Info. Theory*, 38(2):691–712, 1992.
- [120] A. C. Bovik, P. Maragos, and T. F. Quatieri. AM-FM energy detection and separation in noise using multiband energy operators. *IEEE Trans. Signal Proc.*, 41(12):3245–3265, 1993.
- [121] K. Brandenburg, G. Stoll, F. Dehery, and J. D. Johnstone. The ISO-MPEG-Audio codec: A generic-standard for coding of high quality digital audio. *J. Audio Eng. Soc.*, 42(10):780–792, 1994.
- [122] C. Brislawn. Fingerprints go digital. *Notices of the AMS*, 42(11):1278–1283, November 1995.
- [123] O. Bryt and M. Elad. Compression of facial images using the K-SVD algorithm. *J. Vis. Comm. Image Representation*, 19(4):270–282, May 2008.
- [124] J. B. Buckheit and D. L. Donoho. WAVELET and reproducible research. In *Wavelets and Statistics*, ed. A. Antoniadis, pp. 53–81. Springer-Verlag, 1995.
- [125] P. J. Burt. Smart sensing within a pyramid vision machine. *Proc. IEEE*, 76(8):1006–1015, 1988.
- [126] P.J. Burt and E. H. Adelson. The Laplacian pyramid as a compact image code. *Proc. IEEE Int. Conf. Commun.*, 31(4):532–540, 1983.
- [127] C. A. Cabrelli and U. M. Molter. Wavelet transform of the dilation equation. *J. Austral. Math. Soc.*, 37, 1996.
- [128] C. A. Cabrelli and U. M. Molter. Generalized self-similarity. *J. Math. Anal. Appl.*, 230:251–260, 1999.
- [129] T. Cai. Adaptive wavelet estimation: A block thresholding and oracle inequality approach. *Ann. Statist.*, 27(3):898–924, 1999.
- [130] T. Cai and B. W. Silverman. Incorporate information on neighboring coefficients into wavelet estimation. *Sankhyā*, 63:127–148, 2001.
- [131] T. Cai and H. Zhou. A data-driven block thresholding approach to wavelet estimation. Technical report, *Ann. Statist.*, 36, 2008.
- [132] A. P. Calderón. Intermediate spaces and interpolation: The complex method. *Stud. Math.*, 24:113–190, 1964.
- [133] E. Candès. Compressive sampling. *Proc. International Congress of Mathematicians*, Madrid, 2006.

- [134] E. Candès and D. Donoho. Curvelets: A surprisingly effective nonadaptive representation for objects with edges. *Tech. Report Statistics Depart.*, Stanford University, 1999.
- [135] E. Candès and D. Donoho. New tight frames of curvelets and optimal representations of objects with piecewise  $C^2$  singularities. *Comm. Pure Appl. Math.*, 57(2):219–266, 2004.
- [136] E. Candès and J. Romberg. Practical signal recovery from random projections. *Wavelet Applications in Signal and Image Processing XI, Proc. SPIE Conf.* 5914, 2004.
- [137] E. Candès, J. Romberg, and T. Tao. Signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59(8):1207–1223, 2005.
- [138] E. Candès and T. Tao. Decoding by linear programming. *IEEE Trans. Info. Theory*, 51(12):4203–4215, 2005.
- [139] E. Candès and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans. Info. Theory*, 52(12):5406–5425, 2006.
- [140] E. J. Candès, L. Demanet, D. L. Donoho, and L. Ying. Fast discrete curvelet transforms. *SIAM Multiscale Model. Simul.*, 3(5):861–899, 2003.
- [141] E. J. Candès and D. L. Donoho. Recovering edges in ill-posed inverse problems: Optimality of curvelet frames. *Ann. Statist.*, 30(3):784–842, 2000.
- [142] E. J. Candès and D. L. Donoho. Continuous curvelet transform:  $I_i$ , discretization and frames. *J. Appl. Comput. Harmon. Anal.*, 19(3):198–222, 2003.
- [143] E. J. Candès and T. Tao. Rejoinder: The dantzig selector: Statistical estimation when  $p$  is much larger than  $n$ . *Ann. Statist.*, 35(6):2392–2404, 2007.
- [144] E. J. Candès and M. Wakin. An introduction to compressive sampling. *IEEE Sig. Proc. Mag.*, pp. 21–30, March 2008.
- [145] M. Cannon and J. J. Slotine. Space-frequency localized basis function networks for nonlinear system estimation and control. *Neurocomputing*, 9(3), 1995.
- [146] J. Canny. A computational approach to edge detection. *IEEE Trans. Patt. Anal. and Mach. Intell.*, 36:961–1005, 1986.
- [147] O. Cappè. Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. *IEEE Trans. Speech Audio Process.*, 2:345–349, 1994.
- [148] J.-F. Cardoso. High-order contrasts for independent component analysis. *Neural Comput.*, 11(1):157–192, 1999.
- [149] L. Carleson. On the convergence and growth of partial sums of Fourier series. *Acta Math.*, 116:135–157, 1966.

- [150] R. Carmona. Extrema reconstruction and spline smoothing: Variations on an algorithm of Mallat and Zhong. In *Wavelets and Statistics*, ed. A. Antoniadis, pp. 96–108. Springer-Verlag, 1995.
- [151] R. Carmona, W. L. Hwang, and B. Torrésani. Identification of chirps with continuous wavelet transform. In *Wavelets and Statistics*. Springer-Verlag, 1995.
- [152] A. Chambolle. An algorithm for total variation minimization and applications. *J. Math. Imaging Vision*, 20:89–97, 2004.
- [153] T. F. Chan, G. H. Golub, and P. Mulet. A nonlinear primal-dual method for total variation-based image restoration. *SIAM J. Sci. Comput.*, 20(6):1964–1977, 1999.
- [154] T. F. Chan and J. H. Shen. Mathematical models for local nontexture inpaintings. *SIAM J. Appl. Math.*, 62(3):1019–1043, 2001.
- [155] V. Chappelier and C. Guillemot. Oriented wavelet transform for image compression and denoising. *IEEE Trans. Image Process.*, 15(10):2892–2903, 2006.
- [156] C. T. Chen. Video compression: Standards and applications. *J. Visual Comm. Image Representation*, 4(2):103–111, 1993.
- [157] J. Chen and X. Huo. Theoretical results on sparse representations of multiple measurement vectors. *IEEE Trans. Signal Process.*, 54(12):4634–4643, 2006.
- [158] S. Chen and D. Donoho. Atomic decomposition by basis pursuit. *SPIE International Conference on Wavelets*, July 1995.
- [159] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61, 1998.
- [160] L. P. Chew. Guaranteed-quality mesh generation for curved surfaces. *SCG '93: Proceedings of the Ninth Annual Symposium on Computational Geometry*, pp. 274–280, 1993.
- [161] H. I. Choi and W. J. Williams. Improved time-frequency representation of multi-component signals using exponential kernels. *IEEE Trans. Acoust. Speech Signal Process.*, 37(6):862–871, 1989.
- [162] C. K. Chui and X. Shi. Characterization of fundamental scaling functions and wavelets. *Approx. Theory and Its Appl.*, 1993.
- [163] C. K. Chui and X. Shi. Inequalities of Littlewood-Paley type for frames and wavelets. *SIAM J. Math. Anal.*, 24(1):263–277, 1993.
- [164] C. K. Chui and J. Z. Wang. A cardinal spline approach to wavelets. *Proc. Amer. Math. Soc.*, 113:785–793, 1991.
- [165] T. C. Claassen and W. F. Mecklenbrauker. The aliasing problem in discrete-time Wigner distribution. *IEEE Trans. Acoust. Speech Signal Process.*, 31:1067–1072, 1983.
- [166] J. F. Clearbout and F. Muir. Robust modeling of erratic data. *Geophysics*, 38(5): 826–844, 1973.

- [167] A. Cohen. Ondelettes, analyses multirésolutions et filtres miroir en quadrature. *Ann. Inst. H. Poincaré, Anal. Non Linéaire*, 7:439–459, 1990.
- [168] A. Cohen and J. P. Conze. Régularité des bases d'ondelettes et mesures ergodiques. Technical report, CEREMADE, Université Paris Dauphine, 1991.
- [169] A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and best  $k$ -term approximation. Technical report, 2006.
- [170] A. Cohen and I. Daubechies. Non-separable bidimensional wavelet bases. *Rev. Mat. Iberoamericana*, 9(1):51–137, 1993.
- [171] A. Cohen and I. Daubechies. On the instability of arbitrary biorthogonal wavelet packets. *SIAM J. Math. Anal.*, 24(5):1340–1354, 1993.
- [172] A. Cohen, I. Daubechies, and J.-C. Feauveau. Biorthogonal bases of compactly supported wavelets. *Comm. Pure Appl. Math.*, 45:485–560, 1992.
- [173] A. Cohen, I. Daubechies, O. Guleryuz, and M. Orchard. On the importance of combining wavelet-based nonlinear approximation with coding strategies. *IEEE Trans. Info. Theory*, 48(7):1895–1921, 2002.
- [174] A. Cohen, I. Daubechies, and P. Vial. Wavelet bases on the interval and fast algorithms. *J. Appl. Comput. Harmon. Anal.*, 1:54–81, 1993.
- [175] A. Cohen, R. DeVore, P. Petrushev, and H. Xu. Non-linear approximation and the space  $BV(\mathbb{R}^2)$ . *Amer. J. Math.*, 1998.
- [176] I. Cohen. Speech enhancement using a noncausal a priori snr estimator. *IEEE Signal Process. Lett.*, 11(9):725–728, September 2004.
- [177] L. Cohen. Generalized phase-space distribution functions. *J. Math. Phys.*, 7(5): 781–786, 1966.
- [178] L. Cohen. Time-frequency distributions: A review. *Proc. IEEE*, 77(7):941–981, 1989.
- [179] R. R. Coifman and D. Donoho. Translation invariant de-noising. Technical Report 475, Dept. of Statistics, Stanford University, May 1995.
- [180] R. R. Coifman, G. Matviyenko, and Y. Meyer. Modulated Malvar-Wilson bases. *J. Appl. Comput. Harmon. Anal.*, 4(1):58–61, 1997.
- [181] R. R. Coifman and Y. Meyer. Remarques sur l'analyse de Fourier à fenêtre. *C. R. Acad. Sci.*, pp. 259–261, 1991.
- [182] R. R. Coifman, Y. Meyer, and M. V. Wickerhauser. Wavelet analysis and signal processing. In *Wavelets and Their Applications*. Jones and Barlett, pp. 153–178, 1992.
- [183] P. L. Combettes. Solving monotone inclusions via compositions of nonexpansive averaged operators. *Optimization*, 53(5–6):475–504, 2004.

- [184] P. L. Combettes and J. C. Pesquet. A douglas-rachford splitting approach to non-smooth convex variational signal recovery. *IEEE Trans. Signal Process.*, 1(4), 2007.
- [185] P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *SIAM J. Multiscale Model. Simul.*, 4(4), 2005.
- [186] P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Model. Simul.*, 4(4):1168–1200, 2005.
- [187] P. Comon. Independent component analysis—a new concept? *Signal Process.*, 36(3):287–314, 1994.
- [188] S. F. Cotter, B. D. Rao, K. Engan, and K. Kreutz-Delgado. Sparse solutions to linear inverse problems with multiple measurement vectors. *IEEE Trans. Signal Process.*, 53(7):2477–2488, 2005.
- [189] A. Croisier, D. Esteban, and C. Galand. Perfect channel splitting by use of interpolation/decimation/tree decomposition techniques. *Int. Conf. on Info. Sciences and Systems*, Patras, Greece, pp. 443–446, August 1976.
- [190] Z. Cvetkovic and M. Vetterli. Consistent reconstruction of signals from wavelet extrema/zero-crossings representation. *IEEE Trans. Signal Process.*, March 1995.
- [191] P. J. Czerepinski, C. Davies, N. Canagarajah, and D. R. Bull. Matching pursuits video coding: Dictionaries and fast implementation. *IEEE Trans. Circ. Syst. Video Technol.*, 10(7):1103–1115, 2000.
- [192] R. von Sachs, D. Donoho, S. Mallat, and Y. Samuelides. Signal and covariance estimation with macrotiles. *IEEE Trans. Signal Process.*, 53(3):614–627, 2003.
- [193] S. Mallat, D. Donoho, and R. von Sachs. Estimating covariances of locally stationary processes: Consistency of best basis methods. *Proc. of Time-Freq. and Time-Scale Symp.*, Paris, July 1996.
- [194] I. Daubechies. Orthonormal bases of compactly supported wavelets. *Comm. Pure Appl. Math.*, 41:909–996, 1988.
- [195] I. Daubechies. The wavelet transform, time-frequency localization and signal analysis. *IEEE Trans. Info. Theory*, 36(5):961–1005, 1990.
- [196] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. Pure Appl. Math.*, 57:1413–1541, 2004.
- [197] I. Daubechies, A. Grossmann, and Y. Meyer. Painless nonorthogonal expansions. *J. Math. Phys.*, 27:1271–1283, 1986.
- [198] I. Daubechies and J. Lagarias. Two-scale difference equations: II. Local regularity, infinite products of matrices and fractals. *SIAM J. Math. Anal.*, 24, 1992.
- [199] I. Daubechies and W. Sweldens. Factoring wavelet transforms into lifting steps. *J. Fourier Anal. Appl.*, 4(3):245–267, 1998.

- [200] J. G. Daugmann. Two-dimensional spectral analysis of cortical receptive field profile. *Vision Res.*, 20:847–856, 1980.
- [201] G. M. Davis, S. Mallat, and M. Avellaneda. Greedy adaptive approximations. *J. Constr. Approx.*, 13:57–98, 1997.
- [202] G. M. Davis, S. Mallat, and Z. Zhang. Adaptive time-frequency decompositions. *SPIE J. Opt. Engin.*, 33(7):2183–2191, 1994.
- [203] Y. F. Dehery, M. Lever, and P. Urcum. A MUSICAM source codec for digital audio broadcasting and storage. *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 3605–3608, 1991.
- [204] N. Delprat, B. Escudié, P. Guillemain, R. Kronland-Martinet, P. Tchamitchian, and B. Torrésani. Asymptotic wavelet and Gabor analysis: Extraction of instantaneous frequencies. *IEEE Trans. Info. Theory*, 38(2):644–664, 1992.
- [205] L. Demaret, N. Dyn, and A. Iske. Image compression by linear splines over adaptive triangulations. Preprint, Technische Universität München, 2004.
- [206] G. Deslauriers and S. Dubuc. Symmetric iterative interpolation. *Constr. Approx.*, 5: 49–68, 1989.
- [207] R. A. DeVore. Nonlinear approximation. *Acta Numer.*, pp. 51–150, 1998.
- [208] R. A. DeVore, B. Jawerth, and B. J. Lucier. Image compression through wavelet transform coding. *IEEE Trans. Info. Theory*, 38(2):719–746, 1992.
- [209] R. A. DeVore, B. Jawerth, and V. Popov. Compression of wavelet decompositions. *Amer. J. Math.*, 114:737–785, 1992.
- [210] R. A. DeVore, G. Kyriazis, and D. Leviatan. Wavelet compression and nonlinear  $n$ -widths. *Adv. Comput. Math.*, 1:197–214, 1993.
- [211] J. M. Bioucas Dias and M. A. T. Figueiredo. A new TWIST: Two-step iterative shrinkage/thresholding algorithms for image restoration. *IEEE Trans. Image Process.*, 16(12):2992–3004, 2007.
- [212] M. N. Do and M. Vetterli. The contourlet transform: An efficient directional multiresolution image representation. *IEEE Trans. Image Process.*, 14(12):2091–2106, 2005.
- [213] D. Donoho. Interpolating wavelet transforms. *J. Appl. Comput. Harmon. Anal.*, 1994.
- [214] D. Donoho. Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. *J. Appl. Comput. Harmon. Anal.*, 2(2):101–127, 1995.
- [215] D. Donoho. Unconditional bases and bit-level compression. *J. Appl. Comput. Harmon. Anal.*, 3:388–392, 1996.
- [216] D. Donoho. Wedgelets: Nearly-minimax estimation of edges. *Ann. Statist.*, 27(3): 859–897, 1999.

- [217] D. Donoho. Compressed sensing. *IEEE Trans. Info. Theory*, 52(4):1289–1306, 2006.
- [218] D. Donoho. For most large underdetermined systems of linear equations, the minimal  $\ell_1$  norm solution is also the sparsest solution. *Comm. Pure Appl. Math.*, 59(7):797–829, 2006.
- [219] D. Donoho. For most large underdetermined systems of linear equations, the minimal  $\ell_1$  norm near solution is also the sparsest near-solution. *Comm. Pure Appl. Math.*, 59(7):797–829, 2006.
- [220] D. Donoho and I. Johnstone. Ideal denoising in an orthonormal basis chosen from a library of bases. *C. R. Acad. Sci. Paris, Série I*, 319:1317–1322, 1994.
- [221] D. Donoho and I. Johnstone. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.
- [222] D. Donoho and I. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, 90:1200–1224, 1995.
- [223] D. Donoho and I. Johnstone. Minimax estimation via wavelet shrinkage. *Ann. Statist.*, 26(3):879–921, 1998.
- [224] D. Donoho and Y. Tsaig. Extensions of compressed sensing. *Signal Process.*, 86(3):549–571, 2006.
- [225] D. L. Donoho. Superresolution via sparsity constraints. *SIAM J. Math. Anal.*, 23(5):1309–1331, 1992.
- [226] D. L. Donoho and B. F. Logan. Signal recovery and the large sieve. *SIAM J. Appl. Math.*, 52(2):577–591, 1992.
- [227] D. L. Donoho and P. B. Stark. Uncertainty principles and signal recovery. *SIAM J. Appl. Math.*, 49(3):906–931, 1989.
- [228] D. L. Donoho and Y. Tsaig. Fast solution of  $\ell^1$ -norm minimization problems when the solution may be sparse. *J. Math. Imaging Vision*, 2006.
- [229] D. L. Donoho and M. Elad. Optimally-sparse representation in general (non-orthogonal) dictionaries via  $\ell_1$  minimization. *Proc. Natl. Acad. Sci.*, 5(100):2197–2202, 2003.
- [230] D. L. Donoho and X. Huo. Uncertainty principles and ideal decomposition. *IEEE Trans. Info. Theory*, 47:2845–2862, 2001.
- [231] C. Dossal. Estimations de fonctions géométriques et déconvolution. Ph.D. Thesis, <http://www.cmap.polytechnique.fr/dossal/these.pdf>, 2005.
- [232] C. Dossal and S. Mallat. Sparse spike deconvolution with minimum scale. *Proc. Signal Processing with Adaptive Sparse Structured Representations*, pp. 123–126, 2005.
- [233] C. Dossal, E. Le Pennec, and S. Mallat. Bandlet image estimation with model selection. *Submitted to Test*, 2008.

- [234] P. L. Dragotti and M. Vetterli. Wavelet footprints: Theory, algorithms and applications. *IEEE Trans. Signal Proc.*, 51(5):1306–1323, 2003.
- [235] R. J. Duffin and A. C. Schaeffer. A class of nonharmonic Fourier series. *Trans. Amer. Math. Soc.*, 72:341–366, 1952.
- [236] P. Duhamel, Y. Mahieux, and J. Petit. A fast algorithm for the implementation of filter banks based on time domain aliasing cancellation. *Proc. IEEE Int. Conf. Acoust. Speech Signal Process*, pp. 2209–2212, May 1991.
- [237] P. Duhamel and M. Vetterli. Fast Fourier transforms: A tutorial review and a state of the art. *Signal Proc.*, 19(4):259–299, April 1990.
- [238] N. Dyn, D. Levin, and S. Rippa. Data dependent triangulations for piecewise linear interpolation. *IMA J. Numer. Anal.*, 10(1):137–154, 1990.
- [239] N. Dyn and S. Rippa. Data-dependent triangulations for scattered data interpolation and finite element approximation. *Appl. Numer. Math.*, 12:89–105, 1993.
- [240] W. Yin, E. T. Hale, and Y. Zhang. A fixed-point continuation method for  $\ell^1$ -regularized minimization with applications to compressed sensing. *CAAM Technical Report TR07-07*, 2008.
- [241] M. Elad. Why simple shrinkage is still relevant for redundant representations? *IEEE Trans. Info. Theory*, 52(12):5559–5569, 2006.
- [242] M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Trans. Image Process.*, 15(12):3736–3745, 2006.
- [243] M. Elad, P. Milanfar, and R. Rubinstein. Analysis versus synthesis in signal priors. *Inverse Problems*, 23(3):947–968, 2007.
- [244] M. Elad, J.-L. Starck, D. Donoho, and P. Querre. Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA). *J. Appl. Comput. Harmon. Anal.*, 19:340–358, 2005.
- [245] E. Karoui. New results about random covariance matrices and statistical applications. Ph.D. thesis, Stanford University, 2004.
- [246] K. Engan, S. O. Aase, and J. Hakon Husoy. Method of optimal directions for frame design. In *Proc. ICASSP '99*, pp. 2443–2446, Washington, DC, 1999. IEEE Computer Society.
- [247] Y. Ephraim and D. Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing]*, 32(6):1109–1121, 1984.
- [248] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing]*, 33(2):443–445, 1985.

- [249] S. Esedoglu and J. H. Shen. Digital inpainting based on the mumford-shah-euler image model. *Eur. J. Appl. Math.*, 13:353–370, 2002.
- [250] M. J. Fadili, J. L. Starck, and F. Murtagh. Inpainting and zooming using sparse representations. *Comput. J.*, 2007.
- [251] M. Farge and M. Holschneider. Interpretation of two-dimensional turbulence spectrum in terms of singularity in the vortex cores. *Europhys. Lett.*, 15(7):737–743, 1990.
- [252] M. Farge, N. Kevlahan, V. Perrier, and E. Goirand. Wavelets and turbulence. *Proc. IEEE*, 84(4):639–669, 1996.
- [253] J. C. Feauveau. Analyze multirésolution avec un facteur de résolution  $\sqrt{2}$ . *J. Traitement du Signal*, 7(2):117–128, 1990.
- [254] M. Feilner, D. Van De Ville, and M. Unser. An orthogonal family of quincunx wavelets with continuously adjustable order. *IEEE Trans. Image Process.*, 14(4):499–510, 2005.
- [255] M. Figueiredo and R. Nowak. An EM algorithm for wavelet-based image restoration. *IEEE Trans. Image Process.*, 12(8):906–916, 2003.
- [256] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE J. Sel. Topics Signal Process.*, 1(4):586–598, 2007.
- [257] R. M. Figueras i Ventura, P. Vandergheynst, and P. Frossard. Low-rate and flexible image coding with redundant representations. *IEEE Trans. Image Process.*, 15(3):726–739, 2006.
- [258] P. Flandrin. Wavelet analysis and synthesis of fractional Brownian motion. *IEEE Trans. Info. Theory*, 38(2):910–916, 1992.
- [259] P. J. Franaszczuk, G. K. Bergye, P. J. Durka, and H. M. Eisenberg. Time-frequency analysis using matching pursuit algorithm applied to seizures originating from the mesial temporal lobe. *Electroenceph. Clin. Neurophysiology*, 106:513–521, 1998.
- [260] M. Frazier and B. Jawerth. Decomposition of Besov spaces. *Indiana Univ. Math. J.*, 34:777–789, 1985.
- [261] J. Friedman, T. Hastie, H. Hofling, and R. Tibshirani. Pathwise coordinate optimization. *Ann. Appl. Statist.*, 1(2):302–332, 2007.
- [262] J. H. Friedman. Multivariate adaptive regression splines. *Ann. Statist.*, 19(1):1–141, 1991.
- [263] J. H. Friedman and W. Stuetzle. Projection pursuit regression. *J. Amer. Statist. Assoc.*, 76:817–823, 1981.
- [264] U. Frisch and G. Parisi. Turbulence and predictability in geophysical fluid dynamics and climate dynamics. In *Fully Developed Turbulence and Intermittency*, North-Holland, 1985.

- [265] J. Froment and S. Mallat. Second generation compact image coding with wavelets. In *Wavelets: A Tutorial in Theory and Applications*. Academic Press, 1992.
- [266] J.-J. Fuchs. On sparse representations in arbitrary redundant bases. *IEEE Trans. Info. Theory*, 50(6):1341–1344, 2004.
- [267] D. Gabor. Theory of communication. *J. IEE*, 93:429–457, 1946.
- [268] M. Garland and P. Heckbert. Surface simplification using quadric error metrics. *Proc. of SIGGRAPH 1997*, pp. 209–215, 1997.
- [269] D. Geiger and K. Kumaran. Visual organization of illusory surfaces. In *4th Eur. Conf. Comp. Vision*, 1996.
- [270] P. Georgiev, F. J. Theis, and A. Cichocki. Sparse component analysis and blind source separation of underdetermined mixtures. *IEEE Trans. Neural Networks*, 16(4):992–996, 2005.
- [271] J. Geronimo, D. Hardin, and P. R. Massupust. Fractal functions and wavelet expansions based on several functions. *J. Approx. Theory*, 78:373–401, 1994.
- [272] A. C. Gilbert, M. J. Strauss, and J. A. Tropp. A tutorial on fast fourier sampling. *IEEE Sig. Proc. Mag*, pp. 57–66, 2008.
- [273] H. Gish and J. Pierce. Asymptotically efficient quantizing. *IEEE Trans. Info. Theory*, 14:676–683, 1968.
- [274] D. Goldfarb and W. Yin. Second-order cone programming methods for total variation-based image restoration. *SIAM J. Sci. Comput.*, 27(2):622–645, 2005.
- [275] T. Goldstein and S. Osher. The split bregman algorithm for L1 regularized problems. UCLA CAM Report 08-29, 2008.
- [276] P. Goupillaud, A. Grossman, and J. Morlet. Cycle-octave and related transforms in seismic signal analysis. *Geoexploration*, 23:85–102, 1984/85.
- [277] R. M. Gray. Quantization noise spectra. *IEEE Trans. Info. Theory*, pp. 1220–1240, June 1990.
- [278] R. Gribonval. Fast matching pursuit with a multiscale dictionary of Gaussian chirps. *IEEE Trans. Signal Process.*, 49(5):994–1001, 2001.
- [279] R. Gribonval, P. Depalle, X. Rodet, E. Bacry, and S. Mallat. Sound signals decomposition using a high-resolution matching pursuit. In *Proc. Int. Computer Music Conf. (ICMC'96)*, pp. 293–296, August 1996.
- [280] R. Gribonval and M. Nielsen. Beyond sparsity: Recovering structured representation by  $l^1$ -minimization and greedy algorithms. *Adv. Comput. Math.*, 2007.
- [281] R. Gribonval and M. Nielsen. Beyond sparsity: Recovering structured representations by  $l^1$  minimization and greedy algorithms. *Adv. Comput. Math.*, 2008.
- [282] R. Gribonval, H. Rauhut, K. Schnass, and P. Vandergheynst. Atoms of all channels, unite! average case analysis of multi-channel sparse recovery using greedy algorithms. *Technical Report IRISA*, 1848, 2007.

- [283] R. Gribonval and P. Vandergheynst. On the exponential convergence of matching pursuits in quasi-incoherent dictionaries. *IEEE Trans. Info. Theory*, 52(1):255–261, 2006.
- [284] W. Grimson. Computational experiments with a feature based stereo algorithm. *IEEE Trans. Patt. Anal. and Mach. Intell.*, 7:17–34, 1985.
- [285] K. Gröchenig. Sharp results on random sampling of band-limited function. In *NATO ASI 1991 on Probabilistic and Stochastic Methods in Analysis and Applications*, 1992.
- [286] K. Gröchenig. Irregular sampling of wavelet and short-time Fourier transforms. *Constr. Approx.*, 9:283–297, 1993.
- [287] K. Gröchenig. Acceleration of the frame algorithm. *IEEE Trans. Signal Process.*, 41(12):3331–3340, 1993.
- [288] A. Grossmann and J. Morlet. Decomposition of Hardy functions into square integrable wavelets of constant shape. *SIAM J. Math. Anal.*, 15(4):723–736, 1984.
- [289] P. Guillemin and R. Kronland-Martinet. Characterization of acoustic signals through continuous linear time-frequency representations. *Proc. IEEE*, 84(2):561–585, 1996.
- [290] Y. Guoshen and S. Mallat. Super-resolution image zooming. Technical report, 2008.
- [291] A. Haar. Zur theorie der orthogonalen funktionensysteme. *Math. Annal.*, 69:331–371, 1910.
- [292] T. Halsey, M. Jensen, L. Kadanoff, I. Procaccia, and B. Shraiman. Fractal measures and their singularities: The characterization of strange sets. *Phys. Rev. A*, 33(2):1141–1151, 1986.
- [293] F.J. Harris. On the use of windows for harmonic analysis with the discrete Fourier transform. *Proc. IEEE*, pp. 11–33, 1978.
- [294] D. Healy and D. J. Brady. Compression at the physical interface. *IEEE Sig. Proc. Mag.*, pp. 67–71, 2008.
- [295] D. M. Healy and J. B. Weaver. Two applications of wavelet transforms in magnetic resonance imaging. *IEEE Trans. Info. Theory*, 38(2):840–860, 1992.
- [296] H. J. Heijmans, B. Pesquet-Popescu, and G. Piella. Building nonredundant adaptive wavelets by update lifting. *J. Appl. Comput. Harmon. Anal.*, 18(3):252–281, 2005.
- [297] C. Heil and D. Walnut. Continuous and discrete wavelet transforms. *SIAM Rev.*, 31:628–666, 1989.
- [298] J. Herault and C. Jutten. Space or time adaptive signal processing by neural network models. In *AIP Conference Proceedings 151 on Neural Networks for Computing*, pp. 206–211. American Institute of Physics Inc., Woodbury, NY, 1987.

- [299] C. Herley, J. Kovačević, K. Ramchandran, and M. Vetterli. Tilings of the time-frequency plane: Construction of arbitrary orthogonal bases and fast tiling algorithms. *IEEE Trans. Signal Process.*, 41(12):3341–3359, 1993.
- [300] C. Herley and M. Vetterli. Wavelets and recursive filter banks. *IEEE Trans. Signal Process.*, 41(8):2536–2556, 1993.
- [301] F. Hlawatsch and F. Boudreux-Bartels. Linear and quadratic time-frequency signal representations. *IEEE Sig. Proc. Mag.*, 9(2):21–67, 1992.
- [302] F. Hlawatsch and P. Flandrin. The interference structure of the Wigner distribution and related time-frequency signal representations. In *The Wigner Distribution Theory and Applications in Signal Processing*, 1993.
- [303] M. Holschneider, R. Kronland-Martinet, J. Morlet, and P. Tchamitchian. A real-time algorithm for signal analysis with the help of the wavelet transform. In *Wavelets, Time-Frequency Methods and Phase Space*, pp. 289–297. Springer-Verlag, 1989.
- [304] M. Holschneider and P. Tchamitchian. Pointwise analysis of Riemann’s nondifferentiable function. *Inventiones Mathematicae*, 105:157–176, 1991.
- [305] H. Hoppe. Progressive meshes. *Proc. of SIGGRAPH 1996*, pp. 99–108, 1996.
- [306] D. Hubel and T. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex. *J. Physio.*, 160, 1962.
- [307] D. Huffman. A method for the construction of minimum redundancy codes. *Proc. of IRE*, 40:1098–1101, 1952.
- [308] M. V. Hulle. Clustering approach to square and non-square blind source separation. *Proc. IEEE Workshop Neural Netw. Signal Process.*, pp. 315–323, 1999.
- [309] B. Hummel and R. Moniot. Reconstruction from zero-crossings in scale-space. *IEEE Trans. Acoust. Speech Signal Process.*, 37(12), December 1989.
- [310] W. L. Hwang and S. Mallat. Characterization of self-similar multifractals with wavelet maxima. *J. Appl. Comput. Harmon. Anal.*, 1:316–328, 1994.
- [311] I. A. Ibragimov and R. Z. Khas’minskii. Bounds for the risks of non parametric regression estimates. *Theory Probab. Appl.*, 27:84–99, 1984.
- [312] S. Jaffard. Pointwise smoothness, two-microlocalization and wavelet coefficients. *Publicacions Matemàtiques*, 35:155–168, 1991.
- [313] S. Jaffard. Multifractal formalism for functions parts I and II. *SIAM J. Math. Anal.*, 28(4):944–998, 1997.
- [314] S. Jaggi, W. C. Karl, S. Mallat, and A. S. Willsky. High resolution pursuit for feature extraction. *J. Appl. Comput. Harmon. Anal.*, 5:428–449, 1998.
- [315] A. K. Jain and F. Farrokhnia. Unsupervised texture segmentation using Gabor filters. *Patt. Recogn.*, 24(12):1167–1186, 1991.
- [316] N. Jayant. Signal compression: technology targets and research directions. *IEEE J. Sel. Areas Comm.*, 10(5):796–818, 1992.

- [317] N. J. Jayant, J. Johnstone, and B. Safranek. Signal compression based on models of human perception. *Proc. IEEE*, 81(10):1385–1422, 1993.
- [318] B. W. Jeon and S. B. Oh. Fast matching pursuit with vector norm comparison. *IEEE Trans. Circ. Syst. Video Technol.*, 13(4):338–342, 2003.
- [319] L. K. Jones. On a conjecture of Huber concerning the convergence of projection pursuit regression. *Ann. Statist.*, 15(2):880–882, 1987.
- [320] C. C. Jouny, B. Adamolekun, P. J. Franaszczuk, and G. K. Bergye. Intrinsic ictal dynamics at the seizure focus: Effects of secondary generalization revealed by complexity measures. *Epilepsia*, 48:297–304, 2007.
- [321] A. Jourjine, S. Rickard, and O. Yilmaz. Blind separation of disjoint orthogonal signals: Demixing  $n$  sources from two mixtures. *IEEE Conf. Acoustics Speech Signal Process.*, Vol. 5, pp. 2985–2988, 2000.
- [322] B. Julesz. Textons, the elements of texture perception and their interactions. *Nature*, 290, March 1981.
- [323] J. Kalifa and S. Mallat. Minimax restoration and deconvolution. In *Bayesian Inference in Wavelet Based Models*, ed. B. Vidatovic. Springer-Verlag, 1999.
- [324] J. Kalifa and S. Mallat. Thresholding estimators for inverse problems and deconvolutions. *Ann. Statist.*, 31(1):58–109, 2003.
- [325] N. K. Karmarkar. A new polynomial-time algorithm for linear programming. *Combinatorica*, 4, 1984.
- [326] G. Kerkyacharian and D. Picard. Estimation in inverse problems and second-generation wavelets. *Proc. of the ICM*, 2006.
- [327] A. Khodakovsky, P. Schröder, and W. Sweldens. Progressive geometry compression. *Proc. of SIGGRAPH-00*, pp. 271–278, New York, July 23–28, 2000.
- [328] C. J. Kicey and C. J. Lennard. Unique reconstruction of band-limited signals by a Mallat-Zhong wavelet transform algorithm. *Fourier Anal. Appl.*, 3(1):63–82, 1997.
- [329] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky. A method for large-scale  $\ell^1$ -regularized least squares. *IEEE J. Sel. Topics Signal Process.*, 1(4):606–617, 2007.
- [330] J. J. Koenderink. The structure of images. *Biological Cybernetics*, 50:363–370, 1984.
- [331] A. N. Kolmogorov. The local structure of turbulence in incompressible viscous fluid for very large Reynolds numbers. *C. R. Acad. Sc. USSR*, 31(4):538–540, 1941.
- [332] A. N. Kolmogorov. A refinement of previous hypotheses concerning the local structure of turbulence in a viscous incompressible fluid at high Reynolds number. *J. Fluid Mech.*, 13:82–85, 1962.
- [333] J. Kovacevic and W. Sweldens. Wavelet families of increasing order in arbitrary dimensions. *IEEE Trans. Image Process.*, 9(3):480–496, 2000.

- [334] J. Kovačević and M. Vetterli. Nonseparable multidimensional perfect reconstruction filter banks and wavelet bases for  $\mathcal{R}^n$ . *IEEE Trans. Info. Theory*, 38(2):533–555, 1992.
- [335] M. A. Krasner. The critical band coder-digital encoding of speech signals based on the perceptual requirements of the auditor system. *Proc. IEEE Int. Conf. Acoust. Speech Signal Process*, pp. 327–331, 1990.
- [336] K. Kreutz-Delgado, J. F. Murray, B. D. Rao, K. Engan, T.-W. Lee, and T. J. Sejnowski. Dictionary learning algorithms for sparse representation. *Neural Comput.*, 15(2):349–396, 2003.
- [337] F. Labelle and J. R. Shewchuk. Anisotropic voronoi diagrams and guaranteed-quality anisotropic mesh generation. *Proc. of the Nineteenth Conference on Computational Geometry (SCG-03)*, pp. 191–200, June 2003.
- [338] A. Laine and J. Fan. Frame representations for texture segmentation. *IEEE Trans. Image Proc.*, 5(5):771–780, 1996.
- [339] J. Laroche, Y. Stylianos, and E. Moulines. HNS: speech modification based on a harmonic plus noise model. *Proc. IEEE Int. Conf. Acoust. Speech Signal Process*, April 1993.
- [340] W. Lawton. Tight frames of compactly supported wavelets. *J. Math. Phys.*, 31:1898–1901, 1990.
- [341] W. Lawton. Necessary and sufficient conditions for constructing orthonormal wavelet bases. *J. Math. Phys.*, 32:57–61, 1991.
- [342] E. Le Pennec and S. Mallat. Bandelet Image Approximation and Compression. *SIAM Multiscale Model. Simul.*, 4(3):992–1039, 2005.
- [343] T. W. Lee, M. S. Lewicki, M. Girolami, and T. J. Sejnowski. Blind source separation of more sources than mixtures using overcomplete representations. *IEEE Signal Process. Lett.*, 6(4):87–90, 1999.
- [344] D. LeGall. MPEG: A video compression standard for multimedia applications. *Comm. ACM*, 34(4):46–58, 1991.
- [345] P. G. Lemarié. Ondelettes à localisation exponentielle. *J. Math. Pures et Appl.*, 67:227–236, 1988.
- [346] D. Leviatan and V. N. Temlyakov. Simultaneous approximation by greedy algorithms. *Adv. Comput. Math.*, 25(1–3):73–90, 2006.
- [347] M. S. Lewicki and T. J. Sejnowski. Learning overcomplete representations. *Neural Comput.*, 12(2):337–365, 2000.
- [348] A. S. Lewis and G. Knowles. Image compression using the 2D wavelet transform. *IEEE Trans. Image Proc.*, 1(2):244–250, 1992.
- [349] Y. Li, S. Amari, A. Cichocki, D. W. C. Ho, and S. Xie. Underdetermined blind source separation based on sparse representation. *IEEE Trans. Signal Proc.*, 54(2):423–437, 2006.

- [350] J. K. Lin, D. G. Grier, and J. D. Cowan. Faithful representation of separable distributions. *Neural Comput.*, 9(6):1305–1320, 1997.
- [351] J. L. Lin, W. L. Hwang, and S. C. Pei. Fast matching pursuit video coding by combining dictionary approximation and atom extraction. *IEEE Trans. Circ. Syst. Video Technol.*, 17(12):1679–1689, 2007.
- [352] J. M. Lina and M. Mayrand. Complex Daubechies wavelets. *J. Appl. Comput. Harmon. Anal.*, 2:219–229, 1995.
- [353] A. E. Litvak, A. Pajor, M. Rudelson, and N. Tomczak-Jaegermann. Beyond sparsity: Smallest singular value of random matrices and geometry of random polytopes. *Adv. Math.*, 195:491–523, 2005.
- [354] M. Lounsbery, T. D. DeRose, and J. Warren. Multiresolution analysis for surfaces of arbitrary topological type. *ACM Trans. Graph.*, 16(1):34–73, 1997.
- [355] I. J. Lustig, R. E. Marsten, and D. F. Shanno. Interior point methods for linear programming: Computational state of the art. *ORSA J. Comput.*, 6(1):1–14, 1994.
- [356] D. L. Donoho, M. Lustig, and J. M. Pauly. Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magn. Reson. Med.*, 58(6):1182–1195, 2007.
- [357] J. Mairal, M. Elad, and G. Sapiro. Sparse representation for color image restoration. *IEEE Trans. Image Process.*, 17(1):53–69, 2008.
- [358] S. Mallat. Super-resolution bandlet upconversion for HDTV. White paper in [www.cmap.polytechnique.fr/mallat/biblio.html](http://www.cmap.polytechnique.fr/mallat/biblio.html), 2006.
- [359] S. Mallat. Geometrical grouplets. *J. Appl. Comput. Harmon. Anal.*, 2008.
- [360] S. Mallat. An efficient image representation for multiscale analysis. *Proc. Machine Vision Conference*, February 1987.
- [361] S. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. Patt. Anal. Mach. Intell.*, 11(7):674–693, 1989.
- [362] S. Mallat. Multiresolution approximations and wavelet orthonormal bases of  $L^2$ . *Trans. Amer. Math. Soc.*, 315:69–87, 1989.
- [363] S. Mallat and F. Falzon. Analysis of low bit rate image transform coding. *IEEE Trans. Signal Process.*, 46(4), April 1998.
- [364] S. Mallat and W. L. Hwang. Singularity detection and processing with wavelets. *IEEE Trans. Info. Theory*, 38(2):617–643, 1992.
- [365] S. Mallat and G. Peyré. Orthogonal bandlets for geometric image approximation. *Comm. Pure Appl. Math.*, 61(9):1173–1212, 2008.
- [366] S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signal Process.*, 41(12):3397–3415, 1993.
- [367] S. Mallat and S. Zhong. Characterization of signals from multiscale edges. *IEEE Trans. Patt. Anal. Mach. Intell.*, 14(7):710–732, 1992.

- [368] H. S. Malvar. The LOT: A link between block transform coding and multirate filter banks. *Proc. IEEE Int. Symp. Circ. and Syst.*, Espoo, Finland, pp. 835–838, June 1988.
- [369] H. S. Malvar and D. H. Staelin. The LOT: Transform coding without blocking effects. *IEEE Trans. Acoust. Speech Signal Process.*, 37(4):553–559, 1989.
- [370] B. B. Mandelbrot. Intermittent turbulence in self-similar cascades: divergence of high moments and dimension of carrier. *J. Fluid. Mech.*, 62:331–358, 1975.
- [371] B. B. Mandelbrot and J. W. Van Ness. Fractional Brownian motions, fractional noises and applications. *SIAM Rev.*, 10:422–437, 1968.
- [372] S. Masnou. Disocclusion: A variational approach using level lines. *IEEE Trans. Image Process.*, 11(2):68–76, 2002.
- [373] B. Matei and A. Cohen. Nonlinear subdivision schemes: Applications to image processing, in tutorials on multiresolution. In *Geometric Modeling*, pp. 93–97, A. Iske, E. Quak, and M. S. Floater eds. Springer-Verlag, 2002.
- [374] M. R. McClure and L. Carin. Matching pursuits with a wave-based dictionary. *IEEE Trans. Signal Process.*, 45(12):2912–2927, 1997.
- [375] Y. Meyer. Principe d'incertitude, bases hilbertiennes et algèbres d'opérateurs. *Séminaire Bourbaki*, 662, 1986.
- [376] F. Mintzer. Filters for distortion-free two-band multirate filter banks. *IEEE Trans. Acoust. Speech Signal Process.*, 33(3):626–630, 1985.
- [377] A. Moffat. Linear time adaptive arithmetic coding. *IEEE Trans. Info. Theory*, 36(2): 401–406, 1990.
- [378] P. Moulin. A wavelet regularization method for diffuse radar target imaging and speckle noise reduction. *J. Math. Imaging Vision*, pp. 123–134, 1993.
- [379] J. E. Moyal. Quantum mechanics as a statistical theory. *Proc. Cambridge Phi. Soci.*, 45:99–124, 1949.
- [380] H. G. Müller and U. Stadtmuller. Variable bandwidth kernel estimators of regression curves. *Ann. Statist.*, 15:182–201, 1987.
- [381] J. F. Muzy, E. Bacry, and A. Arneodo. The multifractal formalism revisited with wavelets. *Int. J. Bifurcation Chaos*, 4:245, 1994.
- [382] G. Narkiss and M. Zibulevsky. Sequential subspace optimization method for large-scale unconstrained problems. Technical report CCIT, No. 559, EE Dept., Technion, 2005.
- [383] D. Needell and J. A. Tropp. Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *Appl. Comp. Harmon. Anal.*, 2008.
- [384] D. Needell and R. Vershynin. Signal recovery from inaccurate and incomplete measurements via regularized orthogonal matching pursuit. *Appl. Comp. Harmon. Anal.*, 2008.

- [385] R. Neelamani, H. Choi, and R. Baraniuk. Forward: Fourier-wavelet regularized deconvolution for ill-conditioned systems. *IEEE Trans. Image Process.*, 52(2):418–433, 2004.
- [386] R. Neff and A. Zakhori. Very-low bit-rate video coding based on matching pursuits. *IEEE Trans. Circ. Syst. Video Technol.*, 7(1):158–171, 1997.
- [387] R. Neff and A. Zakhori. Matching pursuit video coding. I: Dictionary approximation. *IEEE Trans. Circ. Syst. Video Technol.*, 12(1):13–26, 2002.
- [388] R. Neff and A. Zakhori. Matching-pursuit video coding. II: Operational models for rate and distortion. *IEEE Trans. Circ. Syst. Video Technol.*, 12(1):27–39, 2002.
- [389] Y. Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1, Ser. A):127–152, 2005.
- [390] Y. Nesterov. Gradient methods for minimizing composite objective function. *Preprint UCL*, 2007.
- [391] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive-field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.
- [392] A. Ortego and K. Ramchandran. Rate-distortion methods for image and video compression. *IEEE Sig. Proc. Mag.*, 15(6):23–50, 1998.
- [393] M. R. Osborne, Brett Presnell, and B. A. Turlach. A new approach to variable selection in least squares problems. *IMA J. Numer. Anal.*, 20(3):389–403, 2000.
- [394] F. O’Sullivan. A statistical perspective on ill-posed inverse problems. *Statist. Sci.*, 1:502–527, 1986.
- [395] Y. C. Pati, R. Rezaifar, and P. S. Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. *27th Asilomar Conf. on Signals, Systems and Comput.*, November 1993.
- [396] E. Le Pennec and S. Mallat. Sparse geometric image representations with bandelets. *IEEE Trans. Image Process.*, 14(4):423–438, 2005.
- [397] J. C. Pesquet, H. Krim, H. Carfantan, and J. G. Proakis. Estimation of noisy signals using time-invariant wavelet packets. *Asilomar Conf. on Signals, Systems and Comput.*, November 1993.
- [398] G. Peyré and S. Mallat. Surface compression with geometric bandelets. *ACM Transactions on Graphics*, 24(3), August 2005.
- [399] P. Jonathon Phillips. Matching pursuit filters applied to face identification. *IEEE Trans. Image Process.*, 7(8):1150–1164, 1998.
- [400] M. Pinsker. Optimal filtering of square integrable signals in gaussian white noise. *Problems in Information Transmission*, 16:120–133, 1980.
- [401] D. A. Paffen and S. F. Ronner. Visual cortical neurons as localized spatial frequency filter. *IEEE Trans. Syst., Man, Cybern.*, 13, September 1983.

- [402] M. Porat and Y. Zeevi. Localized texture processing in vision: Analysis and synthesis in Gaborian space. *IEEE Trans. Biomed. Eng.*, 36(1):115–129, 1989.
- [403] M. Porat and Y. Zeevi. The generalized Gabor scheme of image representation in biological and machine vision. *IEEE Trans. Pattern. Anal. Mach. Intell.*, 10(4):452–468, 1988.
- [404] J. Portilla, V. Strela, M. Wainwright, and E. P. Simoncelli. Image denoising using a scale mixture of Gaussians in the wavelet domain. *IEEE Trans. Image Process.*, 12(11):1338–1351, 2003.
- [405] S. Qian and D. Chen. Signal representation via adaptive normalized Gaussian functions. *IEEE Trans. Signal Process.*, 36(1), January 1994.
- [406] S. Ray, C. C. Jouny, N. E. Crone, D. Boatman, N. V. Thakor, and P. J. Franaszczuk. Human ecography analysis during speech perception using matching pursuit: a comparison between stochastic and dyadic dictionaries. *IEEE Trans. Biomed. Eng.*, 50:1371–1373, 2003.
- [407] O. Rioul. Regular wavelets: A discrete-time approach. *IEEE Trans. Signal Process.*, 41(12):3572–3578, 1993.
- [408] O. Rioul and P. Duhamel. Fast algorithms for discrete and continuous wavelet transforms. *IEEE Trans. Info. Theory*, 38(2):569–586, 1992.
- [409] O. Rioul and M. Vetterli. Wavelets and signal processing. *IEEE Sig. Proc. Mag.*, 8(4):14–38, 1991.
- [410] Shmuel Rippa. Long and thin triangles can be good for linear interpolation. *SIAM J. Numer. Anal.*, 29(1):257–270, 1992.
- [411] J. Rissanen and G. Langdon. Arithmetic coding. *IBM J. Research Development*, 23(2):149–162, 1979.
- [412] J. Rissanen and G. Langdon. Universal modeling and coding. *IEEE Trans. Info. Theory*, 27(1):12–23, 1981.
- [413] X. Rodet. Time-domain formant-wave function synthesis. *Spoken Language Generation and Understanding*, ed. J.C. Simon. Reidel Publishing, 1980.
- [414] X. Rodet and P. Depalle. A new additive synthesis method using inverse Fourier transform and spectral envelopes. *Proc. ICMC*, October 1992.
- [415] J. Romberg. Imaging via compressive sensing. *IEEE Sig. Proc. Mag.*, pp. 14–20, March 2008.
- [416] A. Rosenfeld and M. Thurston. Edge and curve detection for visual scene analysis. *IEEE Trans. Comput.*, C(29), 1971.
- [417] B. Roug  . Remarks about space-frequency and space-scale representations to clean and restore noisy images in satellite frameworks. In *Progress in Wavelet Analysis and Applications*, 1993.

- [418] B. Roug . Th orie de la chaine image optique et restauration. Th se d'habilitation, Universit  Paris-Dauphine, 1997.
- [419] M. Rudelson and R. Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Comm. Pure Appl. Math.*, 61(8):1025–1045, 2008.
- [420] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Phys. D*, 60(1–4):259–268, 1992.
- [421] J. Ruppert. A delaunay refinement algorithm for quality two-dimensional mesh generation. *J. Algorithms*, 18(3):548–585, 1995.
- [422] A. Said and W. A. Pearlman. A new, fast, and efficient image codec based on set partitioning in hierarchical trees. *IEEE Trans. Circ. Syst. Video Technol.*, 6(3):243–250, 1996.
- [423] N. Saito and G. Beylkin. Multiresolution representation using the auto-correlation functions of compactly supported wavelets. *IEEE Trans. Signal Proc.*, 41(12):3584–3590, 1993.
- [424] F. Santosa and W. W. Symes. Linear inversion of band-limited reflection seismograms. *SIAM J. Sci. Statist. Computing*, 7(4):1307–1330, 1986.
- [425] B. Scharf. Critical bands. In *Foundations in Modern Auditory Theory*, pp. 150–202. Academic Press, 1970.
- [426] P. Schmid Saugeon and A. Zakhor. Dictionary design for matching pursuit and application to motion-compensated video coding. *IEEE Trans. Circ. Syst. Video Technol.*, 14(6):880–886, 2004.
- [427] P. Schr der and W. Sweldens. Spherical wavelets: Efficiently representing functions on the sphere. *Proc. of SIGGRAPH 95*, pp. 161–172, 1995.
- [428] E. Schwartz. Computational anatomy and functional architecture of striate cortex: A spatial mapping approach to perceptual coding. *Vision Research*, 20:665, 1980.
- [429] C. E. Shannon. Communications in the presence of noise. *Proc. IRE*, 37:10–21, 1949.
- [430] Y. Shao and C. H. Chang. A generalized time-frequency subtraction method for robust speech enhancement based on wavelet filter bank modeling of human auditory system. *IEEE Trans. Syst. Man Cybern.*, 4(37):877–889, 2007.
- [431] J. Shapiro. Adaptive mcclellan transformations for quincunx filter banks. *IEEE Trans. Image Process.*, 42(3):642–648, 1994.
- [432] J. M. Shapiro. Embedded image coding using zero-trees of wavelet coefficients. *IEEE Trans. Signal Process.*, 41(12):3445–3462, 1993.
- [433] M. J. Shensa. The discrete wavelet transform: Wedding the *  trous* and Mallat algorithms. *IEEE Trans. Signal Process.*, 40(10):2464–2482, 1992.

- [434] J. R. Shewchuk. What is a good linear element? Interpolation, conditioning, and quality measures. In *International Meshing Roundtable*, pp. 115–126, 2002.
- [435] Y. Shoham and A. Gersho. Efficient bit allocation for an arbitrary set of quantizers. *IEEE Trans. Acoust. Speech Signal Process.*, 36(9):1445–1453, 1988.
- [436] R. Shukla, P. L. Dragotti, M. Do, and M. Vetterli. Rate distortion optimized tree structured compression algorithms for piecewise smooth images. *IEEE Trans. Image Process.*, 14(3), 2005.
- [437] T. Sikora. Mpeg digital video coding standards. In R. Jurgens, editor, *Digital Electronics Consumer Handbook*. McGraw-Hill, 1997.
- [438] E. Simoncelli and J. Portilla. Texture characterization via second-order statistics of wavelet coefficient amplitudes. *Proc. IEEE Int. Conf. Image*, October 1998.
- [439] E. P. Simoncelli and E. H. Adelson. Nonseparable extensions of quadrature mirror filters to multiple dimensions. *Proc. IEEE*, 78(4):652–664, 1990.
- [440] E. P. Simoncelli and R. W. Buccigrossi. Embedded wavelet image compression based on joint probability model. *Proc. IEEE Int. Conf. Image*, October 1997.
- [441] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger. Shiftable multi-scale transforms. *IEEE Trans. Info. Theory*, 38(2):587–607, 1992.
- [442] D. Sinha and A. H. Tewfik. Low bit rate transparent audio compression using adapted wavelets. *IEEE Trans. Signal Proc.*, 41(12):3463–3479, 1993.
- [443] M. J. Smith and T. P. Barnwell III. Exact reconstruction for tree-structured subband coders. *IEEE Trans. Acoust. Speech Signal Process.*, 34(3):431–441, 1986.
- [444] M. J. Smith and T. P. Barnwell III. A procedure for designing exact reconstruction filter banks for tree structured sub-band coders. *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, March 1984.
- [445] C. Stein. Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, 9:1135–1151, 1981.
- [446] G. Strang and G. Fix. A Fourier analysis of the finite element variational method. *Construct. Asp. Funct. Anal.*, pp. 796–830, 1971.
- [447] G. Strang and V. Strela. Short wavelets and matrix dilation equations. *IEEE Trans. Signal Proc.*, 43:108–115, 1995.
- [448] T. Strohmer and R. W. Heath. Grassmannian frames with applications to coding and communication. *J. Appl. Comput. Harmon. Anal.*, 14(19):257–275, 2003.
- [449] J. O. Strömberg. A modified Franklin system and higher order spline systems on  $\mathbb{R}^n$  as unconditional bases for Hardy spaces. *Proc. Conf. in Honor of Antoni Zygmund*. Vol. II, Wadsworth, pp. 475–493, 1981.
- [450] W. Sweldens. The lifting scheme: A custom-design construction of biorthogonal wavelets. *J. Appl. Comput. Harmon. Anal.*, 3(2):186–200, 1996.
- [451] W. Sweldens. The lifting scheme: A custom-design construction of biorthogonal wavelets. *J. Appl. Comput. Harmon. Anal.*, 3(2):186–200, 1996.

- [452] W. Sweldens. The lifting scheme: A construction of second-generation wavelets. *SIAM J. Math. Anal.*, 29(2):511–546, 1997.
- [453] D. Takhar, J. Laska, M. Wakin, M. Duarte, D. Baron, S. Sarvotham, K. Kelly, and R. Baraniuk. A new compressive imaging camera architecture using optical-domain compression. *Proc. Computational Imaging IV at SPIE Electronic Imaging*, 2006.
- [454] P. Tchamitchian. Biorthogonalité et théorie des opérateurs. *Revista Matemática Iberoamericana*, 3(2):163–189, 1987.
- [455] P. Tchamitchian and B. Torrésani. Ridge and skeleton extraction from the wavelet transform. In *Wavelets and Their Applications*, Jones and Bartlett, pp. 123–151, 1992.
- [456] N. T. Thao and M. Vetterli. Deterministic analysis of oversampled a/d conversion and decoding improvement based on consistent estimates. *IEEE Trans. Signal Proc.*, 42:519–531, 1994.
- [457] F. J. Theis, A. Jung, C. G. Puntonet, and E. W. Lang. Linear geometric ICA: Fundamentals and algorithms. *Neural Comput.*, 15(2):419–439, 2003.
- [458] P. Thevenaz, T. Blu, and M. Unser. Interpolation revisited. *IEEE Trans. Medical Imaging*, 19(7):739–758, 2000.
- [459] B. Torrésani. Wavelets associated with representations of the affine Weil-Heisenberg group. *J. Math. Physics*, 32:1273, 1991.
- [460] T. Tremain. The government standard linear predictive coding algorithm: LPC-10. *Speech Technol.*, 1(2):40–49, 1982.
- [461] J. A. Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE Trans. Info. Theory*, 50(10):2231–2242, 2004.
- [462] J. A. Tropp. Algorithms for simultaneous sparse approximation. part II: Convex relaxation. *Signal Process.*, 86(3):589–602, 2006.
- [463] J. A. Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Trans. Info. Theory*, 52(3):1030–1051, 2006.
- [464] J. A. Tropp, A. C. Gilbert, and M. J. Strauss. Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit. *Signal Process.*, 86(3):572–588, 2006.
- [465] D. Tschumperlé and R. Deriche. Vector-valued image regularization with PDEs: A common framework for different applications. *IEEE Trans. Patt.Anal.Mach.Intell.*, 27(4):506–517, 2005.
- [466] P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optimization Theory and Applications*, 109(3):475–494, 2001.
- [467] M. Unser. Texture classification and segmentation using wavelet frames. *IEEE Trans. Image Process.*, 4(11):1549–1560, 1995.
- [468] M. Unser and A. Aldroubi. A general sampling theory for nonideal acquisition device. *IEEE Trans. Signal Process.*, 42(11):2915–2925, 1994.

- [469] P. P. Vaidyanathan. Quadrature mirror filter banks, M-band extensions and perfect reconstruction techniques. *IEEE ASSP Mag.*, 4(3):4–20, 1987.
- [470] P. P. Vaidyanathan and P. Q. Hoang. Lattice structures for optimal design and robust implementation of two-channel perfect reconstruction filter banks. *IEEE Trans. Acoust. Speech Signal Process.*, 36(1):81–94, 1988.
- [471] M. Vetterli. Filter banks allowing perfect reconstruction. *Signal Proc.*, 10(3):219–244, 1986.
- [472] M. Vetterli. Splitting a signal into subsampled channels allowing perfect reconstruction. *Proc. IASTED Conf. on Appl. Sig. Proc. and Dig. Filt.*, June 1985.
- [473] M. Vetterli and C. Herley. Wavelets and filter banks: Theory and design. *IEEE Trans. Signal Process.*, 40(9):2207–2232, 1992.
- [474] M. Vetterli and K. M. Uz. Multiresolution coding techniques for digital video: A review. *Video Signals Multidimensional Systems Signal Process.*, 3:161–187, 1992.
- [475] J. Ville. Theorie et applications de la notion de signal analytique. *Cables et Transm.*, 2A(1):61–74, 1948.
- [476] C. R. Vogel and M. E. Oman. Iterative methods for total variation denoising. *SIAM J. Sci. Comput.*, 17(1):227–238, 1996.
- [477] M. Wakin, J. Romberg, H. Choi, and R. Baraniuk. Wavelet-domain approximation and compression of piecewise smooth images. *IEEE Trans. Image Proc.*, 15(5):1071–1087, 2006.
- [478] G. K. Wallace. The JPEG still picture compression standard. *Comm. ACM*, 34(4):30–44, 1991.
- [479] A. Wang. Fast algorithms for the discrete wavelet transform and for the discrete Fourier transform. *IEEE Trans. Acoust. Speech Signal Process.*, 32:803–816, August 1984.
- [480] Y. Wang, W. Yin, and Y. Zhang. A fast algorithm for image deblurring with total variation regularization. CAAM Technical report TR07-10, Rice University, 2007.
- [481] A. Watson, G. Yang, J. Solomon, and J. Villasenor. Visual thresholds for wavelet quantization error. *Proceedings of the SPIE*, 2657:382–392, 1996.
- [482] J. Whittaker. Interpolatory function theory. *Cambridge Tracts Math. Math. Phys.*, 33, 1935.
- [483] M. V. Wickerhauser. Acoustic signal compression with wavelet packets. In C. K. Chui, editor, *Wavelets: A Tutorial in Theory and Applications*. Academic Press, 1992.
- [484] E. P. Wigner. On the quantum correction for thermodynamic equilibrium. *Phys. Rev.*, 40:749–759, 1932.

- [485] E. P. Wigner. Quantum-mechanical distribution functions revisited. In *Perspective in Quantum Theory*, 1971.
- [486] K. G. Wilson. Generalized Wannier functions. Technical report, Cornell University, 1987.
- [487] A. Witkin. Scale space filtering. *Proc. Int. Joint. Conf. Artificial Intell.*, June 1983.
- [488] I. Witten, R. Neal, and J. Cleary. Arithmetic coding for data compression. *Comm. ACM*, 30(6):519–540, 1987.
- [489] J. W. Woods and S. D. O’Neil. Sub-band coding of images. *IEEE Trans. Acoust. Speech Signal Process.*, 34(5):1278–1288, 1986.
- [490] G. W. Wornell and A. V. Oppenheim. Wavelet-based representations for a class of self-similar signals with application to fractal modulation. *IEEE Trans. Info. Theory*, 38(2):785–800, 1992.
- [491] A. D. Wyner. An analog scrambling scheme which does not expand bandwidth, part 1. *IEEE Trans. Info. Theory*, 25(3):261–274, 1979.
- [492] Z. X. Xiong, O. Guleryuz, and M. T. Orchard. Embedded image coding based on DCT. *VCIP in Europeen Image Proc. Conf.*, 1997.
- [493] O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Trans. Signal Proc.*, 52(7):1830–1847, 2004.
- [494] W. Yin, S. Osher, J. Darbon, and D. Goldfarb. Bregman iterative algorithms for compressed sensing and related problems. UCLA CAM Technical report, TR07-13, 2007.
- [495] G. Yu, S. Mallat, and E. Bacry. Audio denoising by time-frequency block thresholding. *IEEE Trans. Signal Process.*, 56(3):11820–1839, 2008.
- [496] A. Yuille and T. Poggio. Scaling theorems for zero-crossings. *IEEE Trans. Pattern. Anal. Mach. Intell.*, 8, 1986.
- [497] M. Zhu and T. Chan. An efficient primal-dual hybrid gradient algorithm for total variation image restoration. UCLA CAM report 08-34, 2007.
- [498] M. Zibulevsky and B. A. Pearlmutter. Blind source separation by sparse decomposition in a signal dictionary. *Neural Comput.*, 13(4):863–882, 2001.
- [499] M. Zibulevsky and Y. Y. Zeevi. Extraction of a source from multichannel data using sparse decomposition. *Neurocomputing*, 49(1–4):163–173, 2002.
- [500] M. Zibulski, V. Segalescu, N. Cohen, and Y. Zeevi. Frame analysis of irregular periodic sampling of signals and their derivatives. *J. Fourier Anal. Appl.*, 42:453–471, 1996.
- [501] M. Zibulski and Y. Zeevi. Frame analysis of the discrete Gabor-scheme analysis. *IEEE Trans. Signal Process.*, 42:942–945, 1994.

# Index

## A

- Adaptive
  - basis, 620
  - grid, 457, 466
  - smoothing, 563
- Adjoint operator, 758
  - frame, 156
- Admissible tree, 381, 427, 431, 624
- Affine invariance, 147
- Algorithme à trous*, 176, 240
- Aliasing, 61, 69, 81, 303
- Ambiguity function, 98, 146
- Amplitude modulation, 57, 117
- Analog digital conversion, 7, 65, 168, 742
- Analytic
  - discrete signal, 88, 108
  - function, 108, 116
  - wavelet, 109, 129
- Approximation
  - adaptive grid, 457
  - bounded variation, 463, 468
  - image, 464
  - in wavelet bases, 442, 455
  - linear, 8, 436, 442, 468, 551
  - nonlinear, 9, 451, 512, 551
  - support, 9, 23, 455, 615
  - thresholding, 9
  - uniform grid, 442
- Arithmetic code, 491, 494, 510, 527
- Atom
  - time-frequency, 15, 89
  - wavelet, 92, 102, 109
  - windowed Fourier, 92

## Audio

- masking, 502
- scaling, 117, 124, 132
- transposition, 118, 125, 132

## B

- Backprojection, 163
  - $\mathbf{l}^1$  pursuit, 672
  - matching pursuit, 644
  - Radon, 55
- Balian-Low theorem, 185, 410
- Banach space, 754
- Bandlets, 631
- Basis
  - biorthogonal, 161, 306, 309, 757
  - orthogonal, 757
  - pursuit, 660
  - Riesz, 161, 265, 757
- Basis pursuit, 25
  - Lagrangian, 25, 664, 665, 684
  - wavelet packets, 663
- Battle-Lemarié wavelet, 281, 291, 457
- Bayes
  - estimation, 12, 536, 545
  - risk, 12, 536, 545
- Bernouilli random matrix, 732
- Bernstein inequality, 454
- Besov
  - norm, 460
  - space, 455, 459, 514
- Best approximation, 612
- Best basis, 393, 431, 504, 662
  - approximation, 622
  - compression, 623

- Best basis (*continued*)
  - local cosine, 629
  - search, 623
  - wavelet packet, 626
- Bezout theorem, 174, 293
- Binary tree, 624
- Biorthogonal wavelets
  - basis, 310
  - fast transform, 311
  - lifting, 350
  - ordering, 312
  - regularity, 312
  - splines, 314, 369
  - support, 311
  - symmetry, 312, 369
  - two-dimensional, 345
  - vanishing moments, 311
- Blind source separation, 744
- Block basis, 401, 519
  - cosine, 404, 407
  - Fourier, 401
  - two-dimensional, 402
- Block thresholding, 576
  - risk, 577
- Boundary conditions, 442, 444
- Boundary wavelets, 317, 322, 369
- Bounded variation
  - discrete signal, 47
  - function, 46, 440, 446, 460, 461, 601, 711
  - image, 50, 467, 514, 603, 712
- Box spline, 70, 174, 266
- Butterworth filter, 56
  
- C**
- Canny edge detector, 230
- Cantor
  - measure, 245
  - set, 242
  - spectrum, 251
- Capacity dimension, 243
- CART algorithm, 623
- Cartoon image, 570
- Cauchy-Schwarz inequality, 755
- Chambolle algorithm, 675
- Channel coding, 744
- Chirp
  - hyperbolic, 129, 134
  - linear, 94, 126, 133
  - quadratic, 94
- Choi-William distribution, 148
- Co-area formula, 50
- Coarse to fine, 340
- Code
  - adaptive, 492, 507
  - arithmetic, 491
  - block, 490
  - conditional, 527
  - embedded, 516, 527
  - Huffman, 488
  - prefix, 485
  - Shannon, 488
  - variable length, 485, 506
- Coding gain, 500
- Cohen's class, 145
  - discrete, 150
  - marginals, 146
- Coherent
  - matching pursuit, 658
  - structure, 656
- Coiflets, 296
- Color image(s), 478, 689, 692
- Compact support, 286, 292, 311
- Compression
  - audio, 482
  - dictionary, 614
  - image, 482, 506
  - speech, 482
  - video, 483, 654

- Compressive sensing, 29, 728  
 Concave function, 754  
 Concentration inequality, 618  
 Conditional expectation, 537  
 Cone of influence, 215, 457  
 Conjugate gradient, 165  
 Conjugate mirror filters, 4,  
     276, 306  
     choice, 284, 505  
 Daubechies, 292  
 Smith-Barnwell, 298  
 Vaidyanath-Hoang, 298  
 Continuous wavelet transform,  
     102  
 Convex  
     function, 754  
     hull, 587, 592, 754  
     quadratic, 587  
 Convolution  
     circular, 76, 83, 301, 320, 394,  
     450, 541  
     continuous, 34  
     discrete, 71  
     fast FFT algorithm, 79  
     fast overlap-add, 80  
     integral, 34  
     separable, 83  
 Convolution theorem  
     circular, 77  
     discrete, 74  
     Fourier integral, 37  
 Cosine I basis, 403, 418, 519  
     discrete, 406, 425  
 Cosine IV basis, 404, 418  
     discrete, 407, 422  
 Cost function, 642  
 Covariance, 447, 762  
     operator, 447, 539, 762  
 Cubic spline, 270, 277  
 Curvelets, 194, 476  
     denoising, 570  
     tight frame, 197
- D**
- Daubechies wavelets, 292  
 DCT-I, 406, 409  
 DCT-IV, 407, 408  
 Decibels, 99, 541  
 Deinterlacing, 724  
 Devil's staircases, 246  
 DFT, *see* Discrete Fourier transform  
 Diagonal estimation,  
     552  
 Dictionary, 612  
     denoising, 616  
     Gabor, 650  
     local cosine, 646, 663  
     orthonormal bases, 621  
     wavelet packet, 646, 663  
 Dirac, 33, 40, 94, 719, 763  
     comb, 41, 60, 764  
 Discrete Fourier transform, 76, 540,  
     589  
     inversion, 77  
     Plancherel formula, 77  
     two-dimensional, 83  
 Discrete wavelet basis, 308, 563  
 Distortion rate, 11, 484, 517, 520  
 Dolby, 504  
 Dominated convergence, 274, 753  
 Dual  
     analysis, 22  
     frame, 159  
     synthesis, 22, 162  
 Dyadic wavelet transform, 170, 190,  
     568  
     maxima, 224  
     splines, 174  
     two-dimensional, 189

**E**

## Edges

- curve, 232, 471
- detection, 230
- illusory, 236
- image reconstruction, 235, 236
- multiples, 230
- Eigenvector, 37, 71, 76
- Embedded code, 516, 527
- Energy conservation
  - discrete Fourier transform, 77
  - discrete windowed Fourier, 101
- Fourier integral, 39
- Fourier series, 73
- matching pursuit, 643
- tight frame, 155
- wavelet transform, 105, 111
- windowed Fourier, 96

## Entropy, 486

- differential, 495

## Error correcting code, 744

## Estimation, 12

- adaptive, 544
- block thresholding, 578
- multiples edges, 236
- noise variance, 565
- oracle, 550, 551, 590, 705
- orthogonal projection, 550
- thresholding, 553

Wiener, 539

## Exact Recovery Criteria, 25, 679

**F**

## Fast Fourier transform, 78

- two-dimensional, 85

## Fast wavelet transform

- biorthogonal, 310
- continuous, 114

dyadic, 175

initialization, 301

multidimensional, 349

orthogonal, 298

two-dimensional, 346

Fatou lemma, 753

FFT, *see* Fast Fourier transform

Filter, 34

analog, 37

causal, 34, 71

discrete, 71

interpolation, 337

low-pass, 40, 74

recursive discrete, 74, 87

separable, 83

stable, 34, 71

two-dimensional discrete, 82

varying, 351

Filter bank, 4, 176, 298

perfect reconstruction, 304

separable, 346, 399

Finite elements, 361, 442, 471

Fix-Strang condition, 286, 330, 370

Folded wavelet basis, 320

lifting, 369

Fourier integral, 2

amplitude decay, 42

convolution theorem, 37

in  $L^2(\mathbb{R})$ , 38

in  $L^1(\mathbb{R})$ , 35

inverse, 36

Parseval formula, 39

Plancherel formula, 39

properties, 38

rotation, 53

sampling, 60

slice theorem, 54, 726

support, 45

two-dimensional, 51  
 uncertainty principle, 44  
 Fourier series, 72, 438  
     approximation, 438  
     inversion, 73  
     Parseval formula, 73  
     pointwise convergence, 73  
     random measurements, 733  
 Fractal  
     dimension, 243  
     noise, 258  
 Fractional Brownian, 254, 261  
 Frame  
     algorithm, 164  
     analysis, 155  
     definition, 22, 156  
     dual, 160, 187  
     dual wavelet, 180  
     projector, 166  
     synthesis, 156  
     tight, 156, 183, 197, 476  
     wavelet, 178  
     windowed Fourier, 182  
 Frequency modulation, 117  
 Frequency ridges, 17  
 Frobenius norm, 688  
 Fubini's theorem, 754

**G**

Gabor, 14  
     dictionary, 650  
     wavelet, 111, 190  
 Gaussian  
     function, 41, 45, 126, 137  
     matrix, 731  
     process, 484, 499, 501, 540  
     white noise, 548  
 Geometry, 510  
 Gibbs oscillations, 47, 69, 440

Gram matrix, 157  
 Gram-Schmidt orthogonalization, 648  
 Gray code, 386

**H**

Hölder  
     exponent, 206  
     norm, 464  
     space, 445, 464  
 Haar wavelet, 2, 3, 291  
 Hard thresholding, 668  
 Hausdorff dimension, 243  
 Heat diffusion, 221  
 Heisenberg  
     box, 16, 90, 109, 388, 420, 628  
     uncertainty, 15, 43, 89, 90, 98  
 Hilbert space, 755  
 Histogram, 491, 506, 509  
 Huffman code, 488, 494  
 Hurst exponent, 254  
 Hyperrectangle, 587, 590

**I**

Illusory contours, 236  
 Impulse response, 34, 82  
     discrete, 70, 82  
 Incoherence, 730  
 Inpainting, 722  
 Instantaneous frequency, 94, 115, 138  
 Interpolation, 61, 472, 725  
     Deslauriers-Dubuc, 332, 337  
     function, 328  
     Lagrange, 337  
     spline, 331  
     wavelets, 335  
 Inverse problem, 700  
     compressive sensing, 728  
     super-resolution, 713

Inverse problem (*continued*)  
 thresholding, 27  
 Iterative thresholding, 668

**J**

Jackson inequality, 454  
 Jensen inequality, 754  
 JPEG, 11, 519  
 JPEG-2000, 11, 523

**K**

Karhunen-Loëve  
 approximation, 447  
 basis, 447, 450, 499, 539, 762  
 Kraft inequality, 486, 507

**L**

Lagrangian  
 approximation, 612, 665  
 basis pursuit, 664, 665, 684  
 Lapped  
 fast transform, 424  
 frequency transform, 418  
 orthogonal basis, 416  
 orthogonal transform, 410  
 projector, 411  
 Lazy wavelet, 352  
 Least favorable distribution, 547  
 Left inverse, 159  
 Legendre transform, 248  
 Level set, 50, 232, 467, 471, 728  
 Lifting, 350  
 dual, 355  
 factorization, 367  
 prediction, 353  
 update, 355  
 Linear  
 Bayes risk, 543

estimation, 12, 537  
 programming, 662  
 Lipschitz  
 exponent, 205, 456, 460  
 Fourier condition, 206  
 in two dimensions, 230  
 regularity, 206  
 wavelet condition, 211, 212  
 wavelet maxima, 219

Littlewood-Paley sum, 212

Local cosine  
 basis, 20, 418, 440, 501  
 discrete, 423, 429  
 quad-tree, 430  
 tree, 426, 429  
 two-dimensional, 630  
 Local stationarity, 501  
 Loss function, 536  
 LOT, *see* Lapped

**M**

*M*-band wavelets, 390  
 Mallat algorithm, 298  
 Markov chain, 532  
 Masking noise, 561, 570  
 Matching pursuit, 24, 642, 679  
 denoising, 656  
 fast calculation, 645  
 orthogonal, 648  
 wavelet packets, 646  
 Maxima  
 curves, 232  
 of wavelet transform, 218, 231, 245  
 propagation, 221  
 Median filter, 565  
 Mesh, 361, 472  
 Mexican hat wavelet, 103, 180  
 Meyer  
 wavelet, 289  
 wavelet packets, 418

Minimax, 7  
 estimation, 12, 544  
 risk, 12, 544, 545, 586, 590, 606  
 theorem, 545  
 Mirror wavelet basis, 711  
 Missing data, 722  
 Model selection, 617  
 Modulus maxima, 218, 230  
 Modulus of continuity, 334  
 Mother wavelet, 92  
 Moyal formula, 139  
 MP3, 503  
 MPEG, 483  
 MRI imaging, 743  
 Multichannel signals, 688  
 Multifractal, 19, 242  
 partition function, 248  
 scaling exponent, 248  
 Multiresolution approximations  
 definition, 264  
 piecewise constant, 265, 277, 339  
 Shannon, 265, 266, 277, 339  
 splines, 266, 277, 340  
 Multiscale derivative, 208  
 Multiwavelets, 287, 373  
 MUSICAM, 502  
 Mutual coherence, 678

**N**

Neural network, 645  
 Norm, 754  
 $L^2(\mathbb{R})$ , 756  
 $\ell^2(\mathbb{Z})$ , 756  
 $\ell^p$ , 454, 460, 755  
 $\mathbf{l}^1$ , 660  
 $\mathbf{l}^0$ , 665  
 sup for operators, 758  
 weighted, 498, 520  
 NP-hard, 613

**O**

Operator  
 adjoint, 758  
 projector, 759  
 sup norm, 758  
 time-invariant, 33, 70  
 Oracle  
 attenuation, 550, 590  
 estimation, 549  
 projection, 551, 557  
 Orthogonal  
 basis, 757  
 projector, 759  
 Orthosymmetric set, 592, 606

**P**

Parseval formula, 39, 757  
 Partition function, 248  
 Penalized estimation, 617  
 Piecewise  
 constant, 265, 277, 339  
 polynomial, 543  
 Piecewise regular  
 in 1D, 456, 599  
 in 2D, 471  
 Pixel, 80  
 Plancherel formula, 39, 757  
 Poisson formula, 41, 285  
 Polynomial  
 approximation, 330  
 spline, *see* Spline  
 Posterior distribution, 536  
 Power spectrum, 541, 763  
 Pre-echo, 502  
 Prediction, 606  
 Prefix code, 485  
 Principal directions, 449, 762  
 Prior distribution, 536  
 Prior set, 544

- Pseudo inverse, 159
- PSNR, 508
- Pursuit
  - basis, 660
  - matching, 642, 679
  - orthogonal matching, 648
- Q**
  - Quad-tree, 396, 430, 624
  - Quadratic
    - convex hull, 587, 592
    - convexity, 587
  - Quadrature mirror filters, 302, 371
  - Quantization, 11, 483, 493
    - adaptive, 502
    - bin, 493
    - high resolution, 493, 496, 507
    - low resolution, 510
    - uniform, 494
    - vector, 484
    - weighted, 526
  - Quincunx
    - sampling, 359
    - wavelets, 359
- R**
  - Radon transform, 53, 726, 743
  - Random sensing, 731
  - Random shift process, 449, 542
  - Rate distortion, 529
  - Real wavelet transform, 103
    - energy conservation, 105
    - inverse, 105
  - Regularization
    - Tikhonov, 700, 722
    - total variation, 728
  - Reproducing kernel
    - frame, 167
- wavelet, 106
- windowed Fourier, 97
- Residue, 643, 648
- Restoration, 700
- Restricted isometry constant, 730
- Richardson iteration, 163
- Ridges
  - wavelet, 129
  - windowed Fourier, 122
- Riemann function, 260
- Riemann-Lebesgue lemma, 56
- Riesz basis, 22, 65, 161, 265, 757
- Rihaczek distribution, 147
- Risk, 12, 536
- Run-length code, 519
- S**
  - Sampling
    - Block, 69
    - generalized theorems, 69, 328
    - irregular, 158
    - redundant, 168
    - spline, 70
    - two-dimensional, 81
  - Whittaker, 68
  - Whittaker theorem, 61, 81
- Sampling theorems, 7
- Satellite image, 712
- Scaling equation, 270, 330
- Scaling function, 106, 267
- Scaling images, 724
- Scalogram, 109
- Segmentation, 192
- Seismic imaging, 719
- Self-similar
  - function, 19, 244
  - set, 242
- Separable
  - basis, 84, 760
  - block basis, 402

- convolution, 83
  - decomposition, 84
  - filter, 83
  - filter bank, 399
  - local cosine basis, 431
  - multiresolution, 338
  - wavelet basis, 338, 341
  - wavelet packet basis, 399
  - Shannon
    - code, 488
    - entropy theorem, 486
    - multiresolution, 266
    - sampling theorem, 61
  - Sigma-Delta, 168
  - Signal to Noise Ratio, 541
  - Significance map, 510, 516, 519, 526
  - Singular value decomposition, 27, 701
  - Singular values, 156, 759
  - Singularity, 19, 205
    - spectrum, 246
  - SNR, 541
  - Sobolev
    - differentiability, 438, 443
    - space, 439, 443, 459
  - Soft thresholding, 553
  - Sonar, 126
  - Sound
    - model, 117, 744
    - separation, 744
  - Source separation, 29, 687, 744
  - Sparse spike deconvolution, 719, 733
  - Spectrogram, 92
  - Spectrum
    - of singularity, 246
    - operator, 759
    - power, 763
  - Speech, 117, 482
  - Spline
    - approximation, 457
    - multiresolution, 266
  - sampling, 70
  - wavelet basis, 281
  - Stationary process, 450
    - circular, 540
    - locally, 501
  - Stein Estimator, 559
  - Super-resolution, 28, 713, 724
  - Support
    - approximation, 23
    - recovery, 25
  - Sure threshold, 558, 566
  - Symmetric filters, 313
  - Symmetric operator, 758
  - Symmlets, 294, 565
- T**
- Tensor product, 339, 760
  - Texture discrimination, 191
  - Thresholding
    - block, 576
    - estimation, 14, 568, 705
    - hard, 552, 565, 668
    - inverse problem, 27
    - iterative, 668
    - risk, 552, 592
    - soft, 553, 565
    - Sure, 558, 566
    - threshold choice, 556, 705
    - translation invariant, 561, 566
    - wavelets, 566, 606
  - Tikhonov regularization, 701, 723
  - Time-frequency
    - atom, 15, 89
    - plane, 15, 90
    - resolution, 90, 98, 109, 124, 135, 140, 146, 388
  - Tomography, 53, 726, 743
    - backprojection, 55
  - Tonality, 502

Total variation, 440, 461, 728

discrete signal, 47

function, 46

image, 50

Transfer function, 83

analog, 37

discrete, 71

Transform code, 11, 482, 483

JPEG, 11, 519

with wavelets, 11, 514

Transient, 628

Translation invariance, 168, 226, 422,

561, 566, 589, 646

Transposition, 118, 125, 132

Triangulation, 361, 472

Delaunay, 475

Turbulence, 258

## U

Uncertainty principle, 16, 43, 89, 90, 98

Uniform sampling, 60

## V

Vanishing moments, 208, 284, 330, 342,  
352, 358, 443, 455, 524

Variance estimation, 565

Video compression, 483, 654

Vision, 189

Von Koch fractal, 244

## W

Walsh wavelet packets, 387

Wavelet

directional, 189

seismic, 719

Wavelet basis, 278, 281

Battle-Lemarié, 291, 457

boundary, 301, 322

choice, 284, 524

Coiflets, 296

Daubechies, 3, 292

discrete, 306

folded, 320

graphs, 302

Haar, 291

interval, 317, 369, 442

lazy, 352

lifting, 350

M-band, 390, 504

Meyer, 289

mirror, 711

non-separable, 359

on surfaces, 361

orthogonal, 3

periodic, 318

quincunx, 359

regularity, 287

separable, 341

Shannon, 289

spherical, 365

Symplets, 294

Wavelet packet basis, 19, 382, 504,

626, 710

quad-tree, 430

tree, 379

two-dimensional, 395

Walsh, 387

Wavelet transform, 17

admissibility, 106, 179

analytic, 109

continuous, 17, 102

decay, 211, 212

dyadic, 170

frame, 178

lifting, 356

- maxima, 218, 232
- multiscale differentiation, 208
- real, 103
- ridges, 129, 216
- Weak convergence, 763
- White noise, 540, 548
- Wiener estimator, 538, 539, 542, 589
- Wigner-Ville
  - cross terms, 140
  - discrete, 149
  - distribution, 89, 136, 140, 651
  - instantaneous frequency, 138
  - interferences, 140
  - marginals, 139
  - positivity, 143
- Window
  - Blackman, 99
  - design, 75, 99, 419
  - discrete, 75
- Gaussian, 99
- Hamming, 99
- Hanning, 75, 99
- rectangle, 75
- scaling, 98
- side-lobes, 75, 99, 125
- Windowed Fourier transform, 16, 92
  - discrete, 101
  - energy conservation, 96
  - frame, 182
  - inverse, 96
  - reproducing kernel, 97
  - ridges, 122
- Z
  - Zak transform, 203
  - Zero-tree, 526
  - Zygmund class, 212