CHAPTER

1 Sparse Representations

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 $N \geqslant 10^6$, 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 $e^{i\omega t}$:

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

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

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

The more regular $f(t)$, the faster the decay of the sinusoidal wave amplitude $|\hat{f}(\omega)|$ when frequency ω 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 $\{e^{i2\pi mt}\}_{m\in\mathbb{Z}}$ of $\mathbf{L^2}[0, 1]$. 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 $\{e^{i2\pi kn/N}\}_{0\leq k<N}$ 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

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

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:

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

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 ψ 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 ψ that generates an orthonormal basis. This attempt was a failure since he ended up constructing a whole family of orthonormal wavelet bases, with functions ψ 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 ψ that generates an orthonormal basis of $\mathbf{L^2}(\mathbb{R})$, 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 $\mathbf{L^2}(\mathbb{R})$, a wavelet basis is constructed by dilating and translating a single function ψ. 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 $\mathbf{L^2}(\mathbb{R}^2)$ of finite energy functions $f(x) = f(x_1, x_2)$:

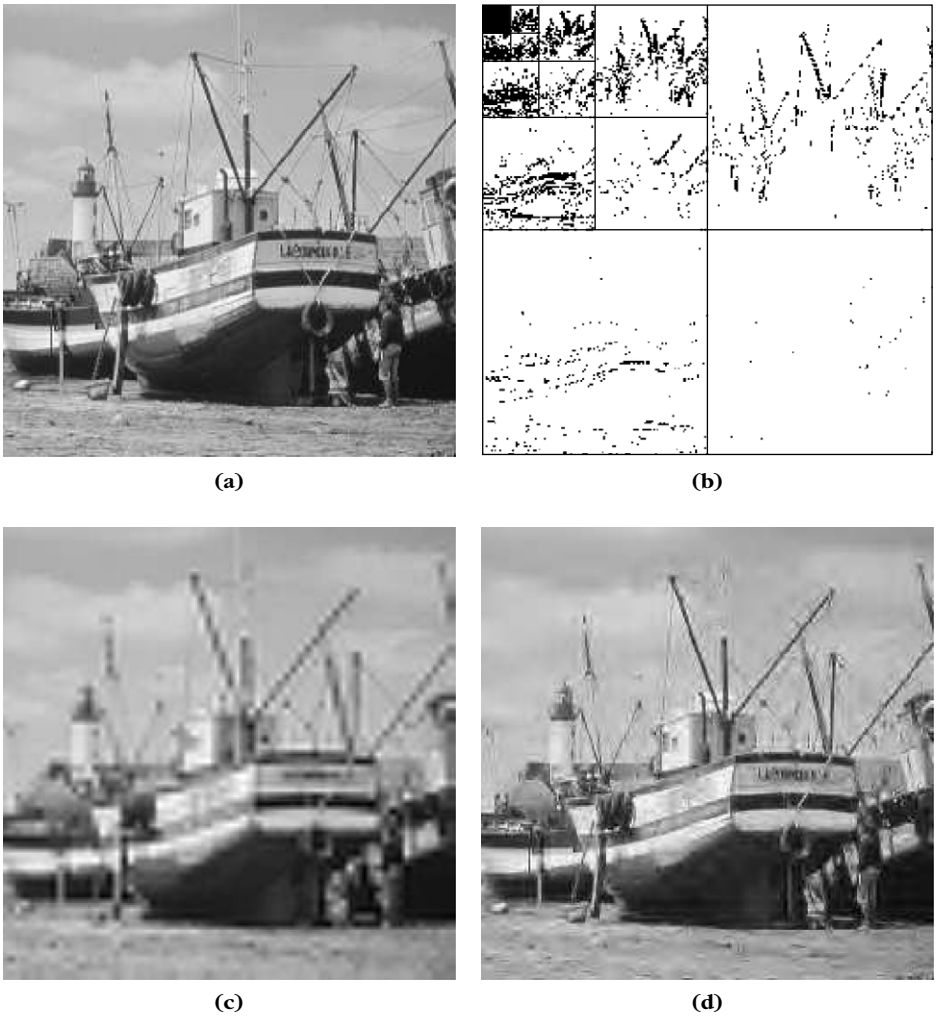
$$\left\{ \psi^k_{j,n}(x) = \frac{1}{2^j} \psi^k \left(\frac{x - 2^j n}{2^j} \right) \right\}_{j \in \mathbb{Z}, n \in \mathbb{Z}^2, 1 \leqslant k \leqslant 3}$$

The support of a wavelet $\psi^k_{j,n}$ 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 $\langle f, \psi^k_{j,n} \rangle$ has a small amplitude if $f(x)$ is regular over the support of $\psi^k_{j,n}$. 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: $|\langle f, \psi^k_{j,n} \rangle| \geqslant T$.

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 40×10^3 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.



(a) **(b)** **(c)** **(d)**

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 Θ, 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 Θ 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 Θ. When this distribution is not available, the average error cannot be calculated and is replaced by the maximum error over Θ. 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:

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

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 ϕ_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

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

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 $\mathbf{L^2}[0, 1]^2$, with the first N vector $\{\bar{g}_m(x)\}_{0 \leq m < N}$ 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 $f[n] = \bar{f} \star \bar{\phi}_s(ns)$. 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 $\mathbf{V}_\Lambda$ generated by M vectors $\{g_m\}_{m \in \Lambda}$ in $\mathcal{B}$ is

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

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

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

We write $|\Lambda|$ the size of the set Λ. 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:

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

This approximation is nonlinear because the approximation set Λ_T changes with f. The resulting approximation error is:

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

Figure 1.1(b) shows that the approximation support Λ_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 Λ_T. It increases the approximation resolution where the signal is irregular. The geometry of Λ_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 $\varepsilon_n(M, f) \geqslant \varepsilon_l(N, f)$.

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 Λ_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 Δ, then $|x - Q(x)| \leqslant \Delta/2$; 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 $\Lambda_T = \{m \in \Gamma : |\langle f, g_m\rangle| \geqslant T\}$ with $T = \Delta/2$. For sparse signals, Chapter 10 shows that the bit budget R is dominated by the number of bits to code Λ_T in Γ, 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 Λ_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 $f[n]$:

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

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 π 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 Θ, 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

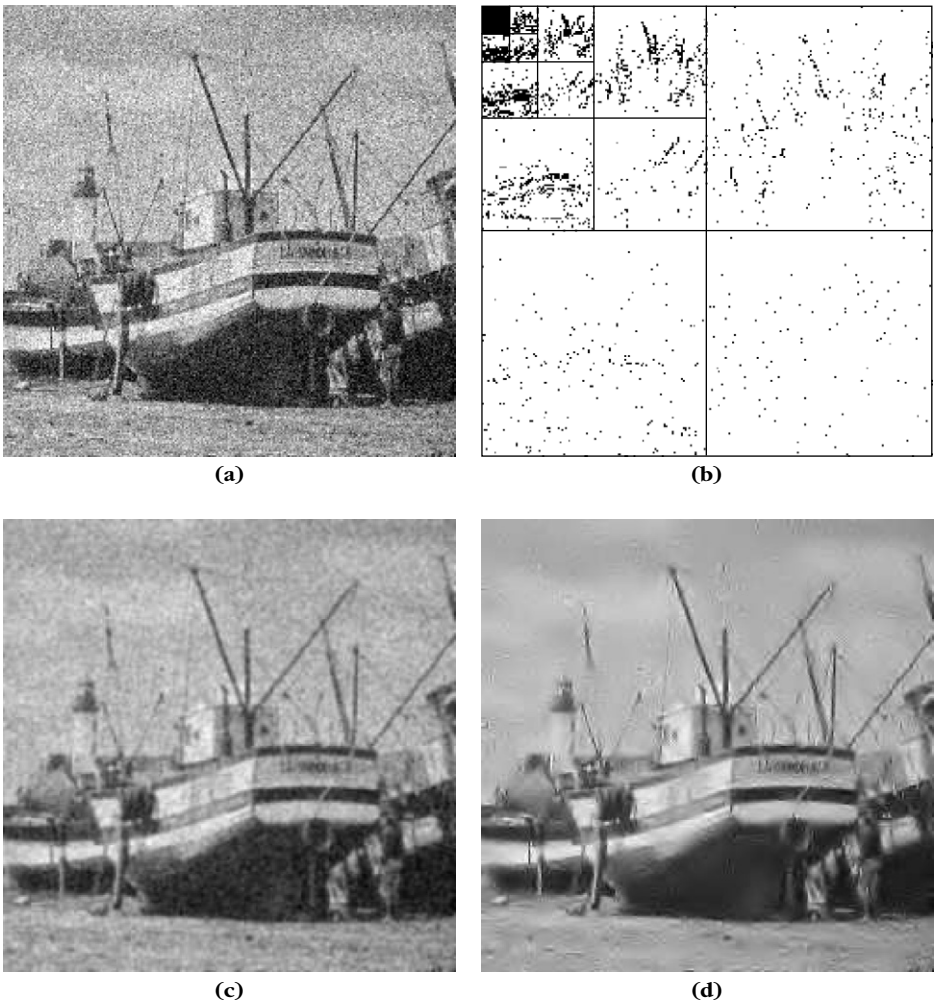


FIGURE 1.2

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

linear filtering estimation $\tilde{F} = X \star h[n]$, 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

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

Thresholding these noisy coefficients yields an orthogonal projection estimator

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

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

Figure 1.2(b) shows the estimated approximation set $\tilde{\Lambda}_T$ of noisy-wavelet coefficients, $|\langle X, \psi_{j,n}| \geqslant T$, that can be compared to the optimal approximation support Λ_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 σ^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 Λ_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 Λ_T in Figure 1.1(b) because it has more isolated points. This indicates that some prior information on the geometry of Λ_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 Θ 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 ϕ_γ 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

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

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, ω) in a rectangle centered at (u, ξ), of size $\sigma_{t,\gamma} \times \sigma_{\omega,\gamma}$. This rectangle is illustrated by Figure 1.3 in this time-frequency plane (t, ω). It can be interpreted as a "quantum of information" over an elementary

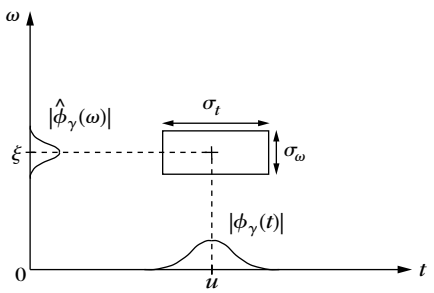


FIGURE 1.3

Heisenberg box representing an atom ϕ_γ.

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

$$\sigma_{t,\gamma}\,\sigma_{\omega,\gamma} \geqslant \frac{1}{2}. \tag{1.11}$$

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 $t = 0$:

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

The atom $g_{u,\xi}$ is translated by u in time and by ξ in frequency. The time-and-frequency spread of $g_{u,\xi}$ is independent of u and ξ. 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, ξ), as shown by Figure 1.4.

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

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

It can be interpreted as a Fourier transform of f at the frequency ξ, 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, ξ). It is thus necessary to understand how to select many fewer time-frequency coefficients that represent the signal efficiently.

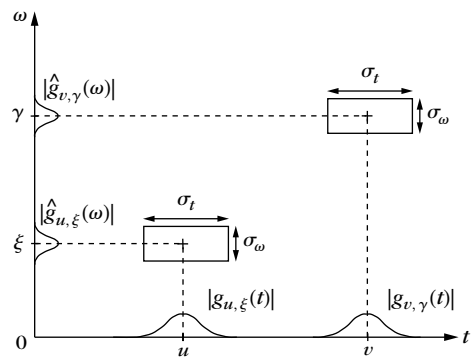


FIGURE 1.4

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 Λ 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 ψ of zero average

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

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

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

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

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

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 ψ whose Fourier transform $\hat{\psi}(\omega)$ is nonzero in a positive frequency interval centered at η. 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 ψ 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.

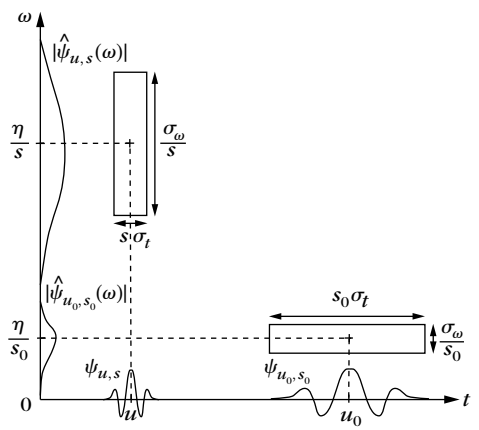


FIGURE 1.5

Heisenberg time-frequency boxes of two wavelets, $\psi_{u,s}$ and ψ_{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 ψ 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 ψ. 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.

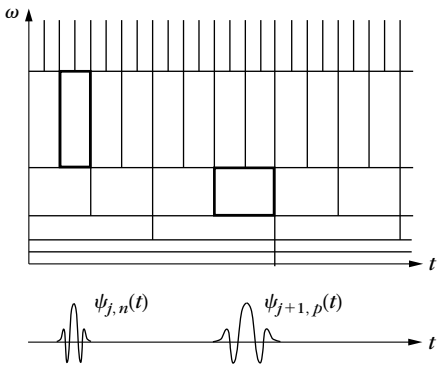


FIGURE 1.6

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

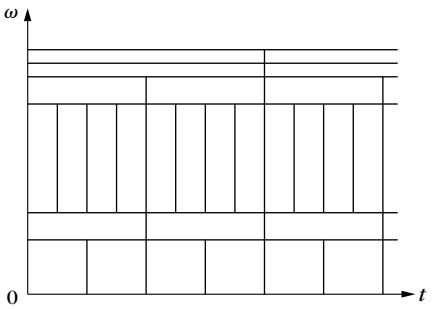


FIGURE 1.7

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,

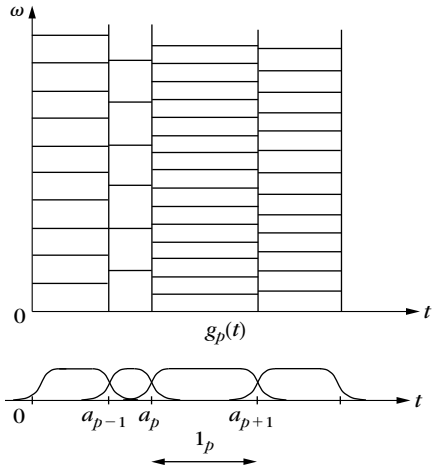


FIGURE 1.8

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 ξ 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_Λ 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_Λ 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 $B \geqslant A > 0$ such that

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

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:

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

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

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

The frame bounds A and B are redundancy factors. If the vectors $\{\phi_p\}_{p\in\Gamma}$ are normalized and linearly independent, then $A \leqslant 1 \leqslant B$. 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 Λ with $M = |\Lambda|$ vectors, which minimize the error $\|f - f_\Lambda\|^2$. Chapter 12 proves that it is equivalent to find Λ_T, which minimizes the corresponding approximation Lagrangian

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

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 Λ_T in Γ.

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 $\tilde{\Lambda}_T$, 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 ϕ_{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 ϕ_{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 $\mathbf{l}^0$ norm $\|a\|_0$, which gives the number of nonzero coefficients. This $\mathbf{l}^0$ norm is highly nonconvex, which explains why the resulting minimization is NP-hard. Donoho and Chen [158] thus proposed replacing the $\mathbf{l}^0$ norm by the $\mathbf{l}^1$ norm $\|a\|_1 = \sum_{p\in\Gamma} |a[p]|$, which is convex. The resulting basis pursuit algorithm computes a synthesis operator

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

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 ε. The optimization (1.18) is thus relaxed by finding a synthesis such that

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

This is a convex minimization problem, with a solution calculated by minimizing the corresponding $\mathbf{l}^1$ 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 ε. This is called an $\mathbf{l}^1$ 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 $\mathbf{l}^1$ Lagrangian pursuits are optimal if they recover the approximation support Λ_T, which minimizes the approximation Lagrangian

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

where f_Λ 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 Λ_T. An *Exact Recovery Criteria* proved by Tropp [464] guarantees that pursuit algorithms do recover the optimal support Λ_T if

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

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 ϕ_q outside Λ_T should have a small inner product with vectors in Λ_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 Λ_T is possible if vectors in Λ_T are incoherent with respect to other dictionary vectors and are incoherent between themselves. It depends on the geometric configuration of Λ_T in Γ.

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 $\mathbf{ImU} = \{Uh : h \in \mathbb{C}^N\}$ 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 **ImU**, 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 **NullU**, all signals h are set to zero $Uh = 0$ and thus disappear in the measured data Y. Recovering the projection of f in **NullU** 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 **ImU** 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 **ImU**, while the others are set to zero. With an appropriate renormalization, $\{\tilde{\lambda}_m^{-1}\, Ug_m\}_{m\in\Gamma_Q}$ 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 nonlinear projector remains in a space of dimension Q. 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 **ImU** of dimension $Q < P$ and thus define a redundant dictionary. Vectors in the approximation support Λ of f are not restricted a priori to a particular subspace of $\mathbb{C}^N$. Super-resolution is possible if the approximation support Λ 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 Λ of f in Γ.

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 Λ 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 Λ 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 Λ 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 Λ in Γ, 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 $|\Lambda| \leq M$:

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

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 $\mathbf{Im}\mathbf{U}$.

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 $f \in C^N$ 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 acquisiton 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 $Q = KN \leq SN$ measurements. Not knowing the operator U makes it even more complicated.

If each source f_s has a sparse approximation support Λ_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 Λ_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*.

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.