Approximation and Processing in Bases[*]

Stephane Mallat

Abstract

This collection comprises Chapter 1 of the book A Wavelet Tour of Signal Processing, The Sparse Way (third edition, 2009) by Stéphane Mallat. The book's website at Academic Press is http://www.elsevier.com/wps/find/bookdescrip
The book's complementary materials are available at http://wavelet-tour.com

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.

1 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 mini-mizing this maximum error, which is called a *minimax* optimization.

[*] Version 1.2: Sep 18, 2009 2:19 pm -0500

[†] http://creativecommons.org/licenses/by/3.0/

2 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.

2.1 Sampling Theorems

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

$$f[n] = \overline{f}\,\texttt{[U+2606]}\,\overline{\phi}_s(ns) \quad \text{for} \quad 0 \leq n < N. \tag{1}$$

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

$$\overline{f}\,\texttt{[U+2606]}\,\overline{\phi}_s(ns) = \int f(u)\ \overline{\phi}_s(ns-u)\ du =< f(x), \phi_s(x-ns) > \tag{2}$$

with $\phi_s(x) = \overline{\phi}_s(-x)$. 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 U_N. The best linear approximation of $\overline{f}$ in U_N recovered from these samples is the orthogonal projection $\overline{f}_N$ of f in U_N, and if the basis is orthonormal, then

$$\overline{f}_N(x) = \textstyle\sum_{n=0}^{N-1} f[n]\ \phi_s(x-ns). \tag{3}$$

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

2.2 Linear Approximation Error

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

$$\overline{f}_N(x) = \sum_{m=0}^{N-1} < \overline{f}, \overline{g}_m > \overline{g}_m(x). \tag{4}$$

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

$$\varepsilon_l(N, f) =\| \overline{f} - \overline{f}_N\|^2 = \sum_{m=N}^{+\infty} |< \overline{f}, \overline{g}_m >|^2. \tag{5}$$

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

Figure (a) shows a discrete image $f[n]$ approximated with $N = 256^2$ pixels. Figure (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 U_N corresponds to a uniform grid that approximates precisely uniform regular signals. Since images $\overline{f}$ 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.

3 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 (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 $\overline{f}$ with N samples written $f[n] = \overline{f}$[U+2606]$\overline{\phi}_s (ns)$. 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

$$< f, g >= \sum_{n=0}^{N-1} f[n]\, g^*[n] \quad \text{and} \quad \| f \|^2 = \sum_{n=0}^{N-1} | f[n] |^2. \tag{6}$$

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 $\{< f, g_m >\}_{m\in\Gamma}$ 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.

3.1 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_λ generated by M vectors $\{g_m\}_{m\in\Lambda}$ in $\mathcal{B}$ is

$$f_\lambda = \sum_{m\in\lambda} < f, g_m > g_m. \tag{7}$$

Since $f = \sum_{m\in\gamma} < f, g_m > g_m$, the resulting error is

$$\| f - f_\lambda\|^2 = \sum_{m\notin\lambda} | < f, g_m > |^2. \tag{8}$$

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:

$$f_M = f_{\lambda_T} = \sum_{m\in\lambda_T} < f, g_m > g_m \quad \text{with} \quad \lambda_T = \{m \in \gamma \ : \ | < f, g_m > | \geq T\}. \tag{9}$$

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

$$\varepsilon_n (M, f) =\| f - f_M\|^2 = \sum_{m\notin\Lambda_T} | < f, g_m > |^2. \tag{10}$$

(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.

(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 approxima-tion f_M:

$$\overline{f}_M (x) = \sum_{n=0}^{N-1} f_M [n]\ \phi_s (x - ns) . \tag{11}$$

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

$$\| \overline{f} - \overline{f}_M\|^2 = \| \overline{f} - \overline{f}_N\|^2 + \| f - f_M \|^2 = \varepsilon_l (N, f) + \varepsilon_n (M, f) . \tag{12}$$

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

3.2 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.

4 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 $\mathbb{C}^N$:

$$f = \sum_{m\in\Gamma} < f, g_m > g_m. \tag{13}$$

Coefficients $< f, g_m >$ are approximated by quantized values $Q (< f, g_m >)$. If Q is auniform quantizer of step Δ, 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

$$\tilde{f} = \sum_{m \in \Gamma} Q\left(< f, g_m >\right) g_m. \tag{14}$$

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

The coefficients not quantized to zero correspond to the set $\lambda_T = \{m \in \gamma : | < f, g_m > | \geq T\}$ with $T = \Delta/2$. For sparse signals, Chapter 10 shows that the bit budget R is dominated by the number of bits to code λ_T in γ, which is nearly proportional to its size $|\lambda_T|$. This means that the "information" about a sparse representation ismostly geometric. Moreover, the distortion is dominated by the nonlinear approximation error $\| f - f_{\Lambda_T} \|^2$, for $f_{\Lambda_T} = \sum_{m \in \lambda_T} < f, g_m > g_m$. Compression is thus a sparse approximation problem. For a given distortion $d\left(R, f\right)$, 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. (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.

5 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 andJohnstone (DonohoJ:94), 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 $W\left[n\right]$ to the original signal $f\left[n\right]$:

$$X\left[n\right] = f\left[n\right] + W\left[n\right] \quad \text{for} \quad 0 \leq n < N. \tag{15}$$

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

$$\tilde{F} = DX. \tag{16}$$

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

$$r\left(D, f\right) = E\{\| f - DX \|^2\}. \tag{17}$$

5.1 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:

$$r\left(D, \pi\right) = E_\pi\{r\left(D, F\right)\}. \tag{18}$$

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

$$r_n\left(\pi\right) = \underset{all\ D}{inf}\, r\left(D, \pi\right). \tag{19}$$

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:

$$r\left(D, \Theta\right) = \underset{f \in \Theta}{sup} r\left(D, f\right). \tag{20}$$

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

$$r_n\left(\Theta\right) = \underset{all\ D}{inf} r\left(D, \Theta\right). \tag{21}$$

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.

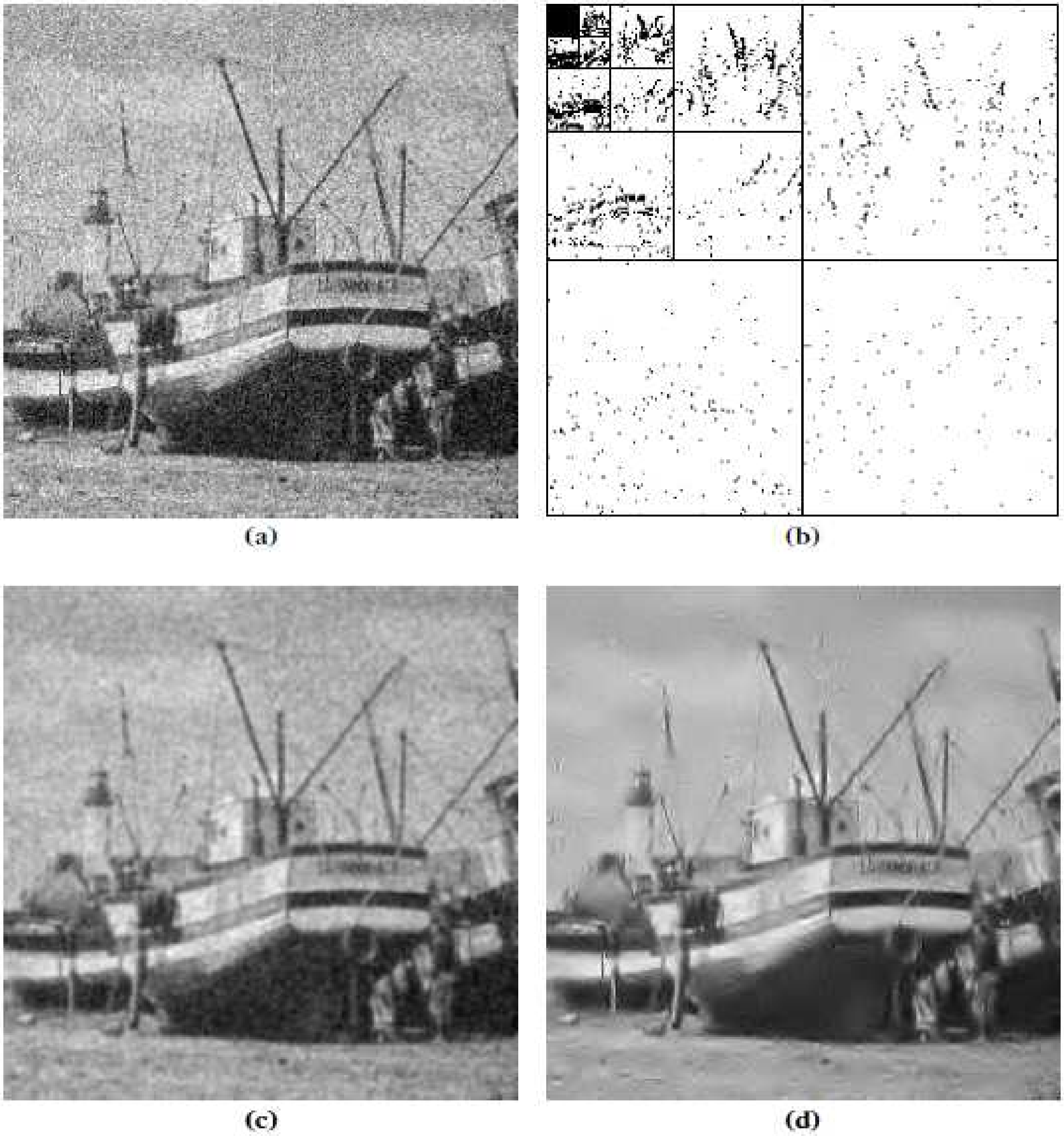

Figure 1

5.2 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 (a) is an image contaminated by Gaussian white noise. Figure (b) shows an optimized linear filtering estimation $\tilde{F} = X$ [U+2606] $h[n]$, 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 (DonohoJ:94) 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

$$< X, g_m >=< f, g_m > + < W, g_m > \quad \text{for} \quad m \in \gamma. \tag{22}$$

Thresholding these noisy coefficients yields an orthogonal projection estimator

$$\tilde{F} = X_{\tilde{\Lambda}_T} = \textstyle\sum_{m\in\tilde{\Lambda}_T} < X, g_m > g_m \quad \text{with} \quad \tilde{\Lambda}_T = \{m \in \gamma \ :\ | < X, g_m > | \geq T\}. \tag{23}$$

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(b) shows the estimated approximation set $\tilde{\lambda}_T$ of noisy-wavelet coefficients, $| < X, \psi_{j,n}| \geq T$, that can be compared to the optimal approximation support Λ_T shown in (b). The estimation in Figure 1(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 $T = \sigma\sqrt{2log_eN}$ yields a risk $E\{\| f - \tilde{F}\|^2\}$ of the order of $\| f - f_{\Lambda_T}\|^2$, up to a log_eN 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(b) "looks" different from the λ_T in (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.