Chapter X
An introduction to wavelets

Nelly Pustelnik

Chapter objectives.

- Understand the limitations of Fourier analysis and the time–frequency viewpoint (STFT, uncertainty principle).
- Define the continuous wavelet transform, then derive the discrete wavelet transform through multiresolution analysis and filter banks (Mallat's algorithm).
- Explain why wavelet coefficients of natural signals are sparse (vanishing moments, regularity), and introduce redundant representations (frames).
- See how orthonormality makes the proximity operators of wavelet-based regularizers explicit, while frames require dual or primal–dual strategies (Chapter 6).
- Use the hierarchy of approximation spaces to build the coarse levels of multilevel algorithms.

Contents

1 Why multiscale representations? 3

2 From Fourier analysis to time–frequency analysis 5

2.1 What Fourier analysis does, and what it misses 5

2.2 Short-time Fourier transform and the uncertainty principle 7

2.3 The wavelet idea: constant relative bandwidth 7

3 The continuous wavelet transform 10

3.1 Definition 10

3.2 Reconstruction and redundancy 11

4 Multiresolution analysis and orthonormal wavelet bases 13

4.1 Approximation spaces 13

4.2 Detail spaces and the mother wavelet 14

4.3 Low and high frequencies 15

5 Discrete wavelet transform: filter banks and Mallat's algorithm 16

5.1 Recursive computation 16

5.2 The DWT as an orthogonal matrix 17

5.3 Images: separable 2D wavelets 18

6 Vanishing moments, regularity and sparsity 18

6.1 Vanishing moments 19

6.2 Regularity and decay of wavelet coefficients 19

6.3 Sparsity and nonlinear approximation 20

7 Frames: redundant wavelet representations 21

7.1 Motivation 21

7.2 Definition and associated operators 21

7.3 Examples of wavelet frames 22

8 Wavelets and proximal operators 22

8.1 Setting 22

8.2 Orthonormal bases: explicit proximal operators 23

8.3 Redundant frames: three ways to handle the proximal step 23

9 Towards multilevel optimization 24

Notation and conventions. The Fourier transform of $f \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ is $\widehat{f}(\omega) = \int_{\mathbb{R}} f(t) e^{-i\omega t}\, \mathrm{d}t$, extended to $L^2(\mathbb{R})$, so that $\|f\|^2 = \frac{1}{2\pi} \left\|\widehat{f}\right\|^2$ (Plancherel) and $\langle f \,|\, g \rangle = \int_{\mathbb{R}} f(t)\overline{g(t)}\, \mathrm{d}t$. In the discrete setting, signals are vectors of $\mathbb{R}^N$ with the canonical inner product, and $\Gamma_0(\mathbb{R}^N)$ denotes the proper, convex, lower semicontinuous functions (Chapter 4). Scales are indexed so that *larger j means coarser*: scale 2^j, i.e. resolution 2^{-j}. The soft-thresholding and clipping operators are $\mathrm{soft}_\lambda(\xi) = \mathrm{sign}(\xi)\max\{|\xi| - \lambda, 0\}$ and $\mathrm{clip}_\lambda(\xi) = \max\{-\lambda, \min\{\xi, \lambda\}\}$, applied componentwise.

1 Why multiscale representations?

Natural signals and images are made of structures living at very different scales: smooth illumination variations, textures, edges, isolated details. Looking at an image at a coarse resolution reveals its global organization; zooming in reveals details that were invisible at the coarse level, and that are themselves organized at finer scales. A good representation should therefore

- separate the information *scale by scale*, i.e. frequency band by frequency band, from a coarse approximation (low frequencies) to successive layers of details (high frequencies);
- remain *localized in space/time*, so that a local event (an edge, a spike) only affects a few coefficients;
- be computable with *fast algorithms* and be *stable*, so that processing the coefficients (thresholding, quantization, optimization) and coming back to the signal domain is cheap and well conditioned.

Wavelets fulfil these requirements. Historically, they emerged in the early 1980s from the work of J. Morlet in geophysics and A. Grossmann in physics [16], and were turned into a mathematical theory by Y. Meyer, S. Mallat and I. Daubechies [21, 19, 9]. Their success in signal and image processing (compression with JPEG2000, denoising, inverse problems) comes from the combination of three properties studied below: time–frequency localization, sparsity, and fast filter-bank implementations.

For this course, wavelets play a double role.

1. **As a regularizer.** Penalizing the ℓ_1-norm of the wavelet coefficients of an image is one of the most classical priors for inverse problems. Its use inside proximal algorithms requires computing proximal operators, which is easy for orthonormal bases and harder for redundant frames (Section 8).
2. **As a hierarchy of levels.** A multiresolution analysis provides a nested sequence of approximation spaces together with explicit restriction and prolongation operators. This is exactly the structure needed to build the coarse models of multilevel optimization methods (Section 9).

> **Remark** (Scale invariance and regularity) A third role of wavelets, not developed in this course but fundamental, is the *characterization of regularity across scales*. Because wavelets are obtained by dilations and translations of a single mother wavelet, the wavelet transform is naturally adapted to *scale-free* (or scale-invariant) signals and images, i.e. objects with no characteristic scale, whose statistics behave as power laws of the scale. Typical examples are fractional Brownian motion, turbulence, natural textures, or many natural images whose power spectrum decays as $|\xi|^{-\beta}$.
> The decay of the wavelet coefficients across scales encodes the smoothness of the analyzed function. For instance, if f is uniformly Hölder with exponent $0 < \alpha < r$ (with r the number of vanishing moments of ψ), then
> $$|\langle f, \psi_{j,k} \rangle| \leq C\, 2^{-j(\alpha + d/2)}$$

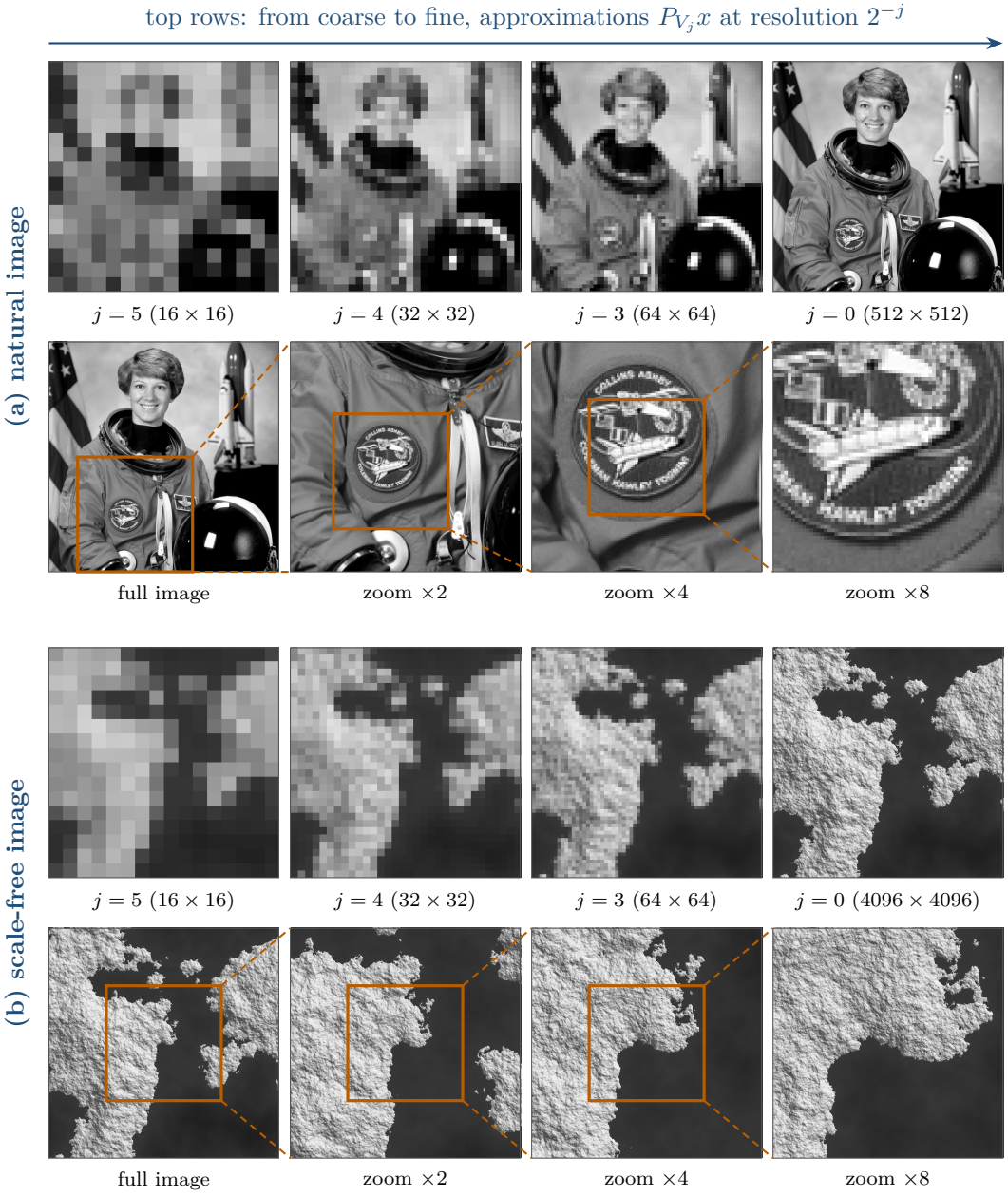


Figure 1: Structures at different scales. In each part, the top row shows approximations of the image at decreasing resolutions (local averages over blocks of $2^j \times 2^j$ pixels of the 512×512 display, i.e. the orthogonal projections P_{V_j} of the Haar multiresolution analysis of Section 4), and the bottom row shows successive zooms by a factor 2 on the full-resolution image. (a) A natural image (512×512 pixels). Even at the coarsest resolution, the global organization is visible (bright background, flag on the left, figure in the centre, helmet at the bottom right); zooming in reveals structures that are invisible at coarse resolutions (the mission patch, its lettering, the drawing of the shuttle), themselves made of edges, textures and fine details. Image: NASA, public domain (`astronaut` test image of scikit-image). (b) A scale-free image: a synthetic coastline (level set of a fractional Brownian field with Hurst exponent $H = 0.75$, 4096×4096 pixels, rendered with relief shading). Each zoom reveals bays, peninsulas and islets that were invisible before, and which look statistically like the whole coastline: the image has no characteristic scale. Wavelets are well suited to such scale-invariant structures, since their atoms are themselves dilations of a single pattern.

(for L^2-normalized wavelets in dimension d, with j increasing toward fine scales); Here $f : \mathbb{R}^d \to \mathbb{R}$ ($d = 1$ for signals, $d = 2$ for images) and $\psi_{j,k}(x) = 2^{jd/2}\psi^{(i)}(2^j x - k)$, $k \in \mathbb{Z}^d$, where $i \in \{1, \dots, 2^d - 1\}$ indexes the orientations of a separable multiresolution analysis (omitted from the notation). conversely, such a decay essentially characterizes Hölder regularity. Finally, measuring how coefficient magnitudes scale with j (log-log regressions of structure functions $S(j,q) = \frac{1}{n_j}\sum_k |d_{j,k}|^q \sim 2^{-j\zeta(q)}$) yields estimators of self-similarity (Hurst) exponents and, through wavelet leaders, of multifractal spectra describing how local regularity fluctuates in space. These tools are central in scale-free analysis of signals and images but go beyond the scope of this course.

Sources. These notes follow the slides of B. Pascal [22], the chapter by B. Pesquet-Popescu and J.-C. Pesquet [23], Chapter 2 of the PhD thesis of L. Davy [11], and the review [24]. The reference textbook is [20].

2 From Fourier analysis to time–frequency analysis

2.1 What Fourier analysis does, and what it misses

The Fourier coefficient $\widehat{f}(\omega) = \langle f \,|\, e_\omega \rangle$, with $e_\omega(t) = e^{i\omega t}$, measures the correlation between f and a pure oscillation at frequency ω. This is ideal for stationary signals and for operators that commute with translations: a convolution $f \mapsto f * k$ becomes a multiplication $\widehat{f}\,\widehat{k}$, which is why the Fourier basis diagonalizes blur operators (and why the FFT is at the heart of deconvolution algorithms).

However, e_ω has a constant modulus over the whole real line: it has *no time localization*. Consequently:

- a local event changes *all* Fourier coefficients. For instance a jump of f at a single time t_0 makes $|\widehat{f}(\omega)|$ decay only like $1/|\omega|$, whatever the regularity of f elsewhere;
- the regularity of f is encoded in the *global* decay of $\widehat{f}$, so that one cannot tell from $|\widehat{f}|$ *where* the signal is smooth and where it is singular;
- the time location of events is hidden in the phase of $\widehat{f}$: a translation $f(\cdot - \tau)$ has the same modulus $|\widehat{f}|$.

To analyze non-stationary signals (a chirp, a piecewise smooth signal, an image with edges) one needs representations that are localized *jointly* in time and frequency.

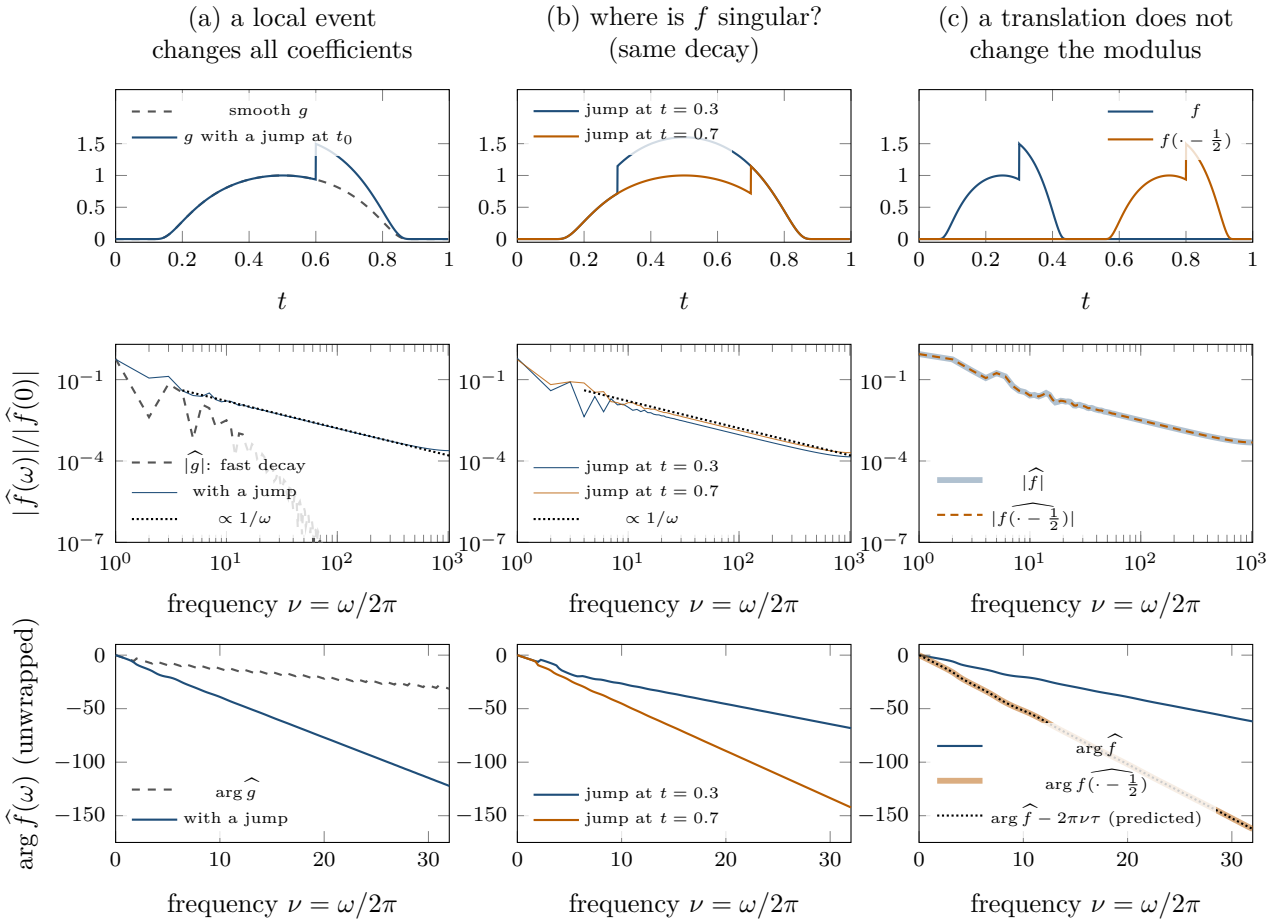


Figure 2: What the modulus of the Fourier transform misses, and what its phase carries (signals sampled on $N = 2048$ points, $|\widehat{f}|$ computed by the FFT, modulus in log–log scale, unwrapped phase on the first frequencies). (a) Adding a single jump at $t_0 = 0.6$ to an infinitely smooth signal g changes *all* Fourier coefficients: the decay of $|\widehat{f}(\omega)|$ drops from faster than any power of ω to $1/\omega$. (b) A jump at $t = 0.3$ or at $t = 0.7$ gives the same $1/\omega$ decay: $|\widehat{f}|$ does not tell *where* the signal is singular, whereas the phases (bottom) differ. (c) A translated signal has exactly the same modulus $|\widehat{f}|$ (the two curves coincide); its location is only encoded in the phase, which is shifted by $-\omega\tau = -2\pi\nu\tau$ (dotted line), since $\widehat{f(\cdot - \tau)}(\omega) = e^{-i\omega\tau}\widehat{f}(\omega)$.

2.2 Short-time Fourier transform and the uncertainty principle

A first solution is to restrict the Fourier analysis to a portion of the signal, using a window $g \in L^2(\mathbb{R})$, real, symmetric, with $\|g\| = 1$, well localized around $t = 0$, which is translated along the time axis. The *short-time Fourier transform* (STFT, or windowed Fourier transform) is

$$Sf(u,\xi) = \int_{\mathbb{R}} f(t)\, g(t-u)\, e^{-i\xi t}\, \mathrm{d}t = \langle f \,|\, g_{u,\xi} \rangle\,, \qquad g_{u,\xi}(t) = g(t-u)e^{i\xi t}. \tag{1}$$

The quantity $|Sf(u,\xi)|^2$ is the *spectrogram*: it measures the energy of f around time u and frequency ξ.

Each atom $g_{u,\xi}$ is characterized by its centre (u,ξ) and its spreads

$$\sigma_t^2 = \int_{\mathbb{R}} (t-u)^2 |g_{u,\xi}(t)|^2\, \mathrm{d}t, \qquad \sigma_\omega^2 = \frac{1}{2\pi} \int_{\mathbb{R}} (\omega-\xi)^2 |\widehat{g}_{u,\xi}(\omega)|^2\, \mathrm{d}\omega,$$

which define a *Heisenberg box* $[u-\sigma_t/2, u+\sigma_t/2] \times [\xi-\sigma_\omega/2, \xi+\sigma_\omega/2]$ in the time–frequency plane. Translating in time or modulating only moves the box: *all STFT atoms have the same box.*

Theorem 2.1 – (Heisenberg uncertainty principle)

For every $g \in L^2(\mathbb{R})$ with $\|g\| = 1$, $\sigma_t\, \sigma_\omega \geq \frac{1}{2}$, with equality if and only if g is a (modulated, translated) Gaussian.

Hence no atom can be perfectly localized in both time and frequency, and the STFT imposes a *single* time–frequency resolution chosen once and for all through the window size: a short window localizes transients but blurs frequencies, a long window resolves frequencies but blurs transients. Gaussian windows reach the optimal localization; one then speaks of *Gabor analysis*.

2.3 The wavelet idea: constant relative bandwidth

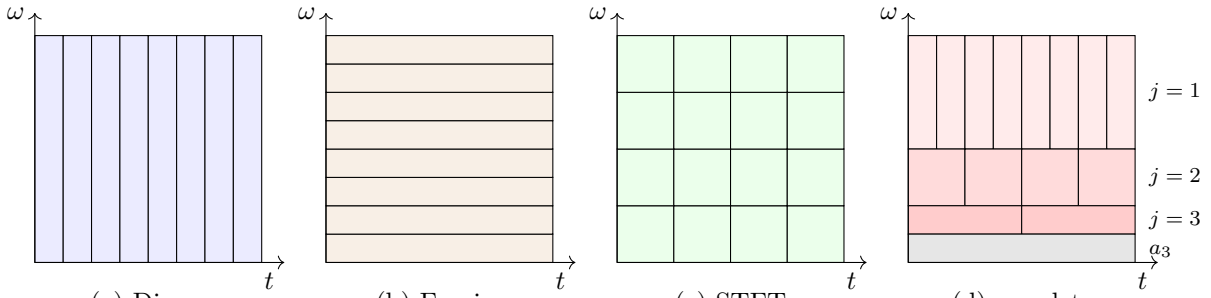
Wavelets replace modulation by *dilation*. From a single oscillating function ψ, centered at $t = 0$ and at frequency $\eta > 0$, with spreads σ_t and σ_ω, one builds

$$\psi_{u,s}(t) = \frac{1}{\sqrt{s}}\, \psi\!\left(\frac{t-u}{s}\right), \qquad u \in \mathbb{R},\ s > 0.$$

Since $\widehat{\psi_{u,s}}(\omega) = \sqrt{s}\, \widehat{\psi}(s\omega) e^{-i\omega u}$, the atom $\psi_{u,s}$ has a Heisenberg box centered at $(u, \eta/s)$ with time width $s\sigma_t$ and frequency width σ_ω/s. The area of the box is constant, but its shape adapts to the scale:

- at **small scales** (high frequencies) the atoms are short: good time resolution, coarse frequency resolution;
- at **large scales** (low frequencies) the atoms are long: coarse time resolution, good frequency resolution.

The ratio bandwidth/central frequency is constant ("constant-Q" analysis), a logarithmic splitting of the spectrum that matches well the structure of natural signals and of human perception [23]. Figure 3 compares the resulting tilings of the time–frequency plane.



(a) Diracs (b) Fourier (c) STFT (d) wavelets

Figure 3: Schematic tilings of the time–frequency plane by the atoms of (a) the canonical basis, (b) the Fourier basis, (c) a short-time Fourier transform with a fixed window and (d) a dyadic wavelet basis with three levels of details and the coarse approximation a_3. All boxes have the same area (uncertainty principle) but only wavelets adapt their shape to the frequency.

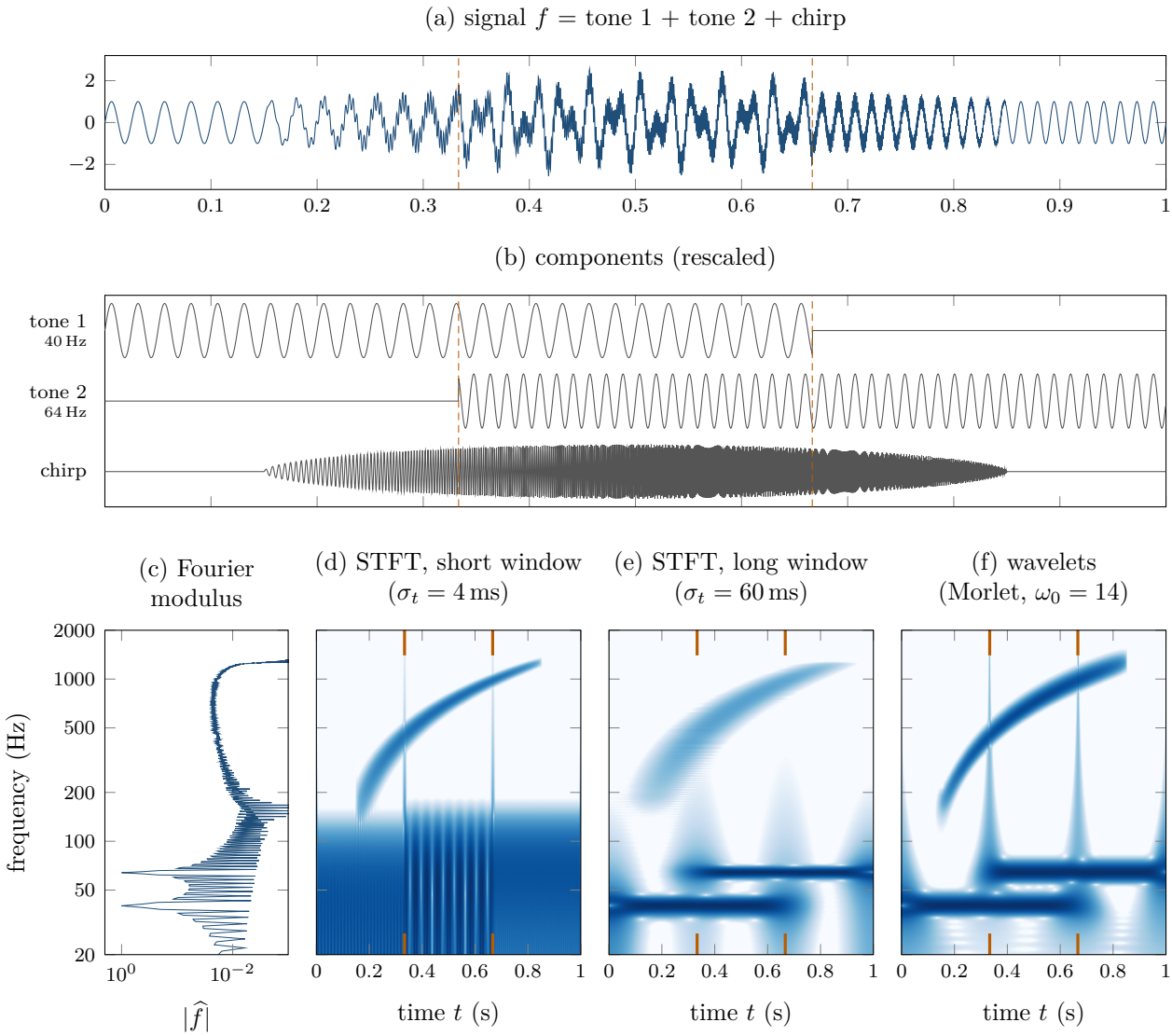


Figure 4: Fourier, short-time Fourier and wavelet analyses of the same signal (1 s sampled at 4096 Hz): a tone at 40 Hz switched on at $t = 0$ and off at $t = 2/3$, a tone at 64 Hz switched on at $t = 1/3$ until the end, and a linear chirp ($150 \to 1270$ Hz); the dashed lines in (a)–(b) and the orange ticks in (d)–(f) mark the switching times. (a) Signal. (b) Its components. Bottom row: all panels share the same logarithmic frequency axis. (c) The Fourier modulus (rotated by 90°) shows the two tones and the band of the chirp, but not *when* each component is present: the two tones seem to last the whole second. (d)–(f) Time–frequency maps (modulus in dB, 45 dB range). (d) A short window locates the switching times precisely but cannot separate the two tones on $[1/3, 2/3]$, where they overlap. (e) A long window separates the two tones but blurs their switching times and widens the chirp: the STFT has a single resolution, fixed by the window (Figure 3(c)). (f) The wavelet transform adapts its resolution to the frequency: long atoms at low frequencies separate the two tones, short atoms at high frequencies follow the chirp and reveal the switching transients as thin vertical lines (Figure 3(d)).

3 The continuous wavelet transform

3.1 Definition

Definition 3.1 – (Wavelet)

A *wavelet* is a function $\psi \in L^2(\mathbb{R}) \cap L^1(\mathbb{R})$ with $\|\psi\| = 1$, localized around $t = 0$, oscillating, and with zero mean:

$$\int_{\mathbb{R}} \psi(t)\,\mathrm{d}t = 0 \qquad \Longleftrightarrow \qquad \widehat{\psi}(0) = 0. \tag{2}$$

Definition 3.2 – (Continuous wavelet transform)

The continuous wavelet transform (CWT) of $f \in L^2(\mathbb{R})$ is

$$\mathcal{W}f(u,s) = \langle f \,|\, \psi_{u,s} \rangle = \int_{\mathbb{R}} f(t)\, \frac{1}{\sqrt{s}}\, \overline{\psi\left(\frac{t-u}{s}\right)}\,\mathrm{d}t, \qquad (u,s) \in \mathbb{R} \times \mathbb{R}_+^*. \tag{3}$$

Remark The quantity $|\mathcal{W}f(u,s)|^2$ is the *scalogram* of f.

The factor $1/\sqrt{s}$ ensures $\|\psi_{u,s}\| = \|\psi\| = 1$. Introducing $\bar{\psi}_s(t) = \frac{1}{\sqrt{s}}\overline{\psi(-t/s)}$, the CWT is a convolution:

$$\mathcal{W}f(u,s) = (f * \bar{\psi}_s)(u), \qquad \widehat{\bar{\psi}_s}(\omega) = \sqrt{s}\,\overline{\widehat{\psi}(s\omega)}. \tag{4}$$

Since $\widehat{\psi}(0) = 0$ and $\widehat{\psi}$ decays at infinity, $\bar{\psi}_s$ is a *band-pass filter* whose passband is centred at η/s: **the CWT is a bank of band-pass filters with dyadically shrinking bandwidths**. This filtering view is the key to the discrete algorithms of Section 5. Figure 5 illustrates it for the Mexican hat wavelet.

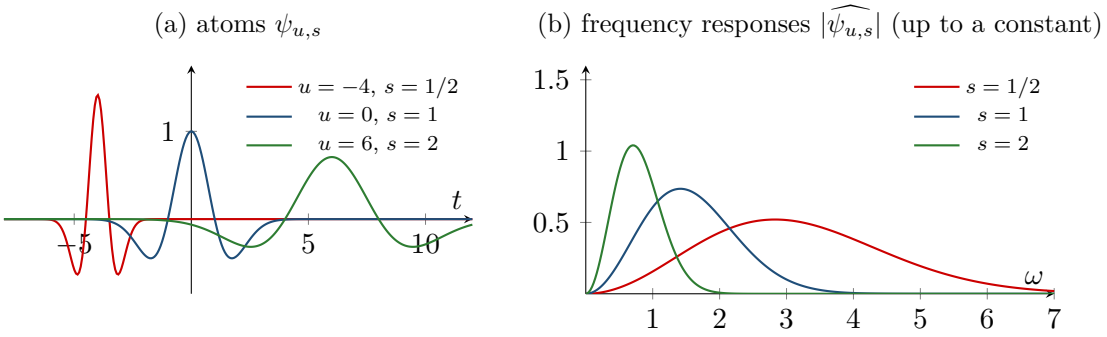


Figure 5: The Mexican hat wavelet $\psi(t) = (1-t^2)e^{-t^2/2}$ (up to normalization), second derivative of a Gaussian, with two vanishing moments. (a) Translated and dilated atoms: small scales give short atoms. (b) In the Fourier domain, each atom is a band-pass filter centred at $\sqrt{2}/s$: small scales analyze high frequencies with a large bandwidth, large scales analyze low frequencies with a narrow bandwidth.

3.2 Reconstruction and redundancy

Theorem 3.3 – (Calderón; Grossmann–Morlet [16])

Let ψ be a real wavelet satisfying the *admissibility condition*

$$C_\psi = \int_0^{+\infty} \frac{|\widehat{\psi}(\omega)|^2}{\omega}\,\mathrm{d}\omega < +\infty. \tag{5}$$

Then every real $f \in L^2(\mathbb{R})$ satisfies

$$f(t) = \frac{1}{C_\psi}\int_0^{+\infty}\!\!\int_{\mathbb{R}} \mathcal{W}f(u,s)\,\frac{1}{\sqrt{s}}\,\psi\!\left(\frac{t-u}{s}\right)\mathrm{d}u\,\frac{\mathrm{d}s}{s^2}, \tag{6}$$

$$\|f\|^2 = \frac{1}{C_\psi}\int_0^{+\infty}\!\!\int_{\mathbb{R}} |\mathcal{W}f(u,s)|^2\,\mathrm{d}u\,\frac{\mathrm{d}s}{s^2}. \tag{7}$$

Remark Admissibility forces $\widehat{\psi}(0) = 0$, i.e. the zero-mean condition (2). Conversely, if ψ has zero mean and decays fast enough, e.g. $\int_{\mathbb{R}}(1+|t|)|\psi(t)|\,\mathrm{d}t < +\infty$, then $\widehat{\psi}$ is continuously differentiable, $\widehat{\psi}(\omega) = O(\omega)$ near 0, and (5) holds [11, 20]. This is the case for all bounded compactly supported wavelets used in practice.

Translation and dilation covariance. If the signal is shifted, its wavelet transform is shifted by the same amount: for $f_\tau(t) = f(t-\tau)$,

$$\mathcal{W}f_\tau(u,s) = \mathcal{W}f(u-\tau,s),$$

which follows from the change of variable $t' = t - \tau$ in the integral defining the CWT. Likewise, if the signal is dilated, $f_\lambda(t) = \lambda^{-1/2} f(t/\lambda)$, its transform is dilated in both variables: $\mathcal{W}f_\lambda(u,s) = \mathcal{W}f(u/\lambda, s/\lambda)$. In words, an event (a singularity, an oscillation, a chirp) always leaves the *same pattern* in the time-scale plane, wherever it occurs and whatever its size: the pattern is simply moved or zoomed. This is what makes the CWT a natural tool to detect, localize and track events. We speak of *covariance* rather than *invariance*: the transform is not unchanged, it moves along with the signal.

Redundancy and reproducing kernel. The CWT represents a function of *one* variable t by a function of *two* variables (u,s): it contains much more numbers than needed, so its values cannot be independent of each other. To make this precise, start from the reconstruction formula $f = \frac{1}{C_\psi}\int_0^{+\infty}\!\int_{\mathbb{R}} \mathcal{W}f(u,s)\,\psi_{u,s}\,\mathrm{d}u\,\frac{\mathrm{d}s}{s^2}$ and take its inner product with a given wavelet $\psi_{u',s'}$. Since $\mathcal{W}f(u',s') = \langle f \,|\, \psi_{u',s'}\rangle$, we obtain

$$\mathcal{W}f(u',s') = \frac{1}{C_\psi}\int_0^{+\infty}\!\!\int_{\mathbb{R}} K(u',s';u,s)\,\mathcal{W}f(u,s)\,\mathrm{d}u\,\frac{\mathrm{d}s}{s^2}, \qquad K(u',s';u,s) = \langle \psi_{u,s} \,|\, \psi_{u',s'}\rangle\,.$$

The function K is called the *reproducing kernel*. It measures how much two wavelets resemble each other: it is close to 1 when $\psi_{u,s}$ and $\psi_{u',s'}$ almost coincide, and it decays when they are far apart in position or in scale. The formula says that each coefficient is a weighted average of its neighbours, with weights given by K: knowing the CWT in a neighbourhood of (u',s') essentially determines its value at (u',s'). Therefore, neighbouring coefficients are strongly correlated. The width of K in u is proportional to the scale s, so coefficients are redundant over a distance of order s in time, and over a constant ratio in scale.

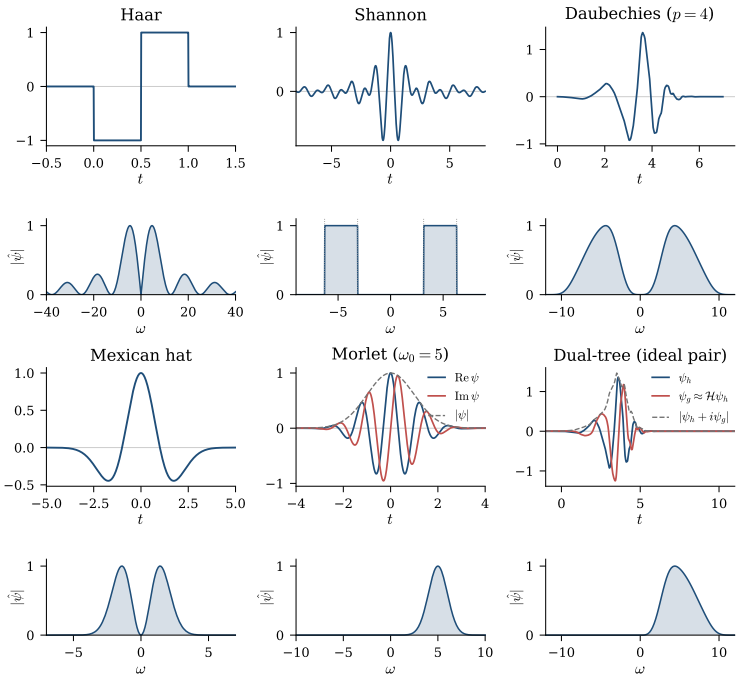
The CWT is therefore excellent for *analysis* (visualization, detection, estimation of regularity), where redundancy makes patterns easy to see, but it is wasteful for *processing* and storage. The

previous remark also indicates how to discretize it without losing information: sample the scales geometrically, $s = 2^{-j}$, and the positions on a grid of step proportional to the scale, $u = k\,2^{-j}$. This is the starting point of the discrete wavelet transform.

Classical wavelets. Table 1 summarizes a few classical choices (adapted from [11]). The number n_ψ of vanishing moments, defined in Section 6, measures the ability of ψ to ignore smooth (polynomial) components.

Wavelet	Definition	n_ψ	Localization	Typical use
Haar (real)	$\psi = \mathbf{1}_{[1/2,1)} - \mathbf{1}_{[0,1/2)}$	1	compact support, poor in frequency	fast transforms, piecewise constant signals
Shannon (real)	$\widehat{\psi} = \mathbf{1}_{\pi \le \lvert\omega\rvert \le 2\pi}$	∞	ideal in frequency, slow decay in time	theoretical reference
Daubechies	orthogonal filters of minimal length	p	compact support of length $2p-1$	standard orthonormal bases
Mexican hat	$(1-t^2)e^{-t^2/2}$	2	fast decay	CWT, edge and ridge detection
Morlet (complex)	$\simeq e^{-t^2/2}e^{i\omega_0 t}$	≈ 1	near-optimal time–frequency	scalograms, chirps
Dual-tree (complex)	pair of orthogonal filter banks	depends	compact support, nearly analytic	oriented 2D analysis

Table 1: Some classical wavelets and their properties.



4 Multiresolution analysis and orthonormal wavelet bases

To obtain a non-redundant representation, the parameters are sampled on a dyadic grid, $s = 2^j$ and $u = 2^j k$ with $(j,k) \in \mathbb{Z}^2$:

$$\psi_{j,k}(t) = \frac{1}{\sqrt{2^j}}\,\psi\left(\frac{t - 2^j k}{2^j}\right) = 2^{-j/2}\psi(2^{-j}t - k).$$

With this convention, j increases toward *coarse* scales: $\psi_{j,k}$ has width of order 2^j and is centred near $2^j k$. The question is whether one can choose ψ such that $\{\psi_{j,k}\}_{(j,k)\in\mathbb{Z}^2}$ is an *orthonormal basis* of $L^2(\mathbb{R})$. Multiresolution analysis gives a constructive answer, together with a fast algorithm.

Strategy: from ψ to φ. Constructing directly a function ψ whose dilations and translations are orthonormal is difficult: the conditions mix all scales at once. Multiresolution analysis takes a detour. Instead of the *details* of a signal, it first describes its successive *approximations* at scales 2^j, using a second function φ, called the *scaling function* (or father wavelet). The two functions play complementary roles:

- φ computes **local averages**: normalized so that $\int \varphi = 1$, it is a low-pass function, and $\langle f \,|\, \varphi_{j,k}\rangle$ is a blurred sample of f at scale 2^j around position $2^j k$;
- ψ computes **local differences**: $\int \psi = 0$, it is a band-pass function, and $\langle f \,|\, \psi_{j,k}\rangle$ measures what is lost when f is blurred from scale 2^{j-1} to the coarser scale 2^j.

The construction proceeds in three steps:

1. define nested approximation spaces V_j spanned by the translates of φ (Section 4.1);
2. encode φ in a discrete low-pass filter h_0;
3. obtain ψ from φ through a high-pass filter h_1, as the function generating the *difference* between two consecutive approximations (Section 4.2).

The wavelet is therefore a by-product of the scaling function, not its starting point.

4.1 Approximation spaces

Definition 4.1 – (Multiresolution analysis [19, 21])

A multiresolution analysis (MRA) of $L^2(\mathbb{R})$ is a sequence of closed subspaces $(V_j)_{j\in\mathbb{Z}}$ such that

(i) **nesting**: $\{0\} \subset \cdots \subset V_2 \subset V_1 \subset V_0 \subset V_{-1} \subset \cdots \subset L^2(\mathbb{R})$;

(ii) **translation invariance**: $f \in V_j \iff f(\cdot - 2^j k) \in V_j$ for all $k \in \mathbb{Z}$;

(iii) **dilation**: $f \in V_j \iff f(\cdot/2) \in V_{j+1}$;

(iv) **completeness**: $\bigcap_j V_j = \{0\}$ and $\bigcup_j V_j$ is dense in $L^2(\mathbb{R})$;

(v) **scaling function**: there exists $\varphi \in L^2(\mathbb{R})$ such that $\{\varphi(\cdot - k)\}_{k\in\mathbb{Z}}$ is an orthonormal basis of V_0.

V_j is the space of approximations at scale 2^j (resolution 2^{-j}): the larger j, the coarser the approximation. Properties (iii) and (v) imply that $\{\varphi_{j,k} = 2^{-j/2}\varphi(2^{-j}\cdot - k)\}_{k\in\mathbb{Z}}$ is an orthonormal basis of V_j, so that the best approximation of f at scale 2^j is its orthogonal projection

$$P_{V_j} f = \sum_{k\in\mathbb{Z}} a_j[k]\,\varphi_{j,k}, \qquad a_j[k] = \langle f \,|\, \varphi_{j,k}\rangle \quad \text{(approximation coefficients)}. \tag{8}$$

For the Haar example below, V_j is the space of functions that are constant on the intervals $[2^j k, 2^j(k+1))$, and $P_{V_j} f$ replaces f by its mean on each of them.

Two-scale equation and low-pass filter. Since $2^{-1/2}\varphi(t/2) \in V_1 \subset V_0$, it decomposes on the basis of V_0:

$$\frac{1}{\sqrt{2}}\,\varphi\!\left(\frac{t}{2}\right) = \sum_{n\in\mathbb{Z}} h_0[n]\,\varphi(t-n), \qquad h_0[n] = \left\langle 2^{-1/2}\varphi(\cdot/2) \,\middle|\, \varphi(\cdot - n)\right\rangle. \tag{9}$$

In words: a coarse scaling function is a weighted *average* of finer ones. In the Fourier domain, with $H_0(\omega) = \sum_n h_0[n] e^{-i\omega n}$ (2π-periodic),

$$\widehat{\varphi}(2\omega) = \frac{1}{\sqrt{2}} H_0(\omega)\,\widehat{\varphi}(\omega) \qquad \Longrightarrow \qquad \widehat{\varphi}(\omega) = \prod_{p=1}^{+\infty} \frac{H_0(2^{-p}\omega)}{\sqrt{2}}\,\widehat{\varphi}(0). \tag{10}$$

The whole MRA is thus encoded in the discrete filter h_0.

Proposition 4.2 – (Conjugate mirror filter)

If φ is the scaling function of an MRA, then

$$H_0(0) = \sqrt{2} \qquad \text{and} \qquad |H_0(\omega)|^2 + |H_0(\omega+\pi)|^2 = 2 \quad \forall \omega \in \mathbb{R}. \tag{11}$$

In particular $H_0(\pi) = 0$: h_0 is a *low-pass* filter.

The first condition comes from (10) at $\omega = 0$ (with $\widehat{\varphi}(0) \neq 0$); the second translates the orthonormality of the translates $\varphi(\cdot - k)$. Conversely, under mild regularity conditions, any filter satisfying (11) defines a scaling function through the infinite product (10) [20, 9].

4.2 Detail spaces and the mother wavelet

Going from V_{j-1} to the coarser space V_j loses information: the details visible at scale 2^{j-1} but not at scale 2^j. This information is stored in the orthogonal complement W_j:

$$V_{j-1} = V_j \oplus W_j, \qquad W_j \perp V_j, \qquad P_{V_{j-1}} f = P_{V_j} f + P_{W_j} f.$$

We look for a function ψ playing for W_j the role that φ plays for V_j, i.e. whose translates form an orthonormal basis of W_0. Since $W_1 \subset V_0$, the function $2^{-1/2}\psi(t/2)$ must, like $2^{-1/2}\varphi(t/2)$, be a combination of the $\varphi(t-n)$: only the filter changes. Where h_0 *averages* neighbouring translates of φ, the filter h_1 takes their *differences*. It is obtained from h_0 by shifting its frequency response by π (so that low and high frequencies are exchanged) and alternating signs, which ensures $W_j \perp V_j$.

Theorem 4.3 – (Mallat, Meyer)

Let $h_1[n] = (-1)^{1-n} h_0[1-n]$, i.e. $H_1(\omega) = e^{-i\omega}\,\overline{H_0(\omega+\pi)}$, and define the *mother wavelet* ψ by

$$\frac{1}{\sqrt{2}}\,\psi\!\left(\frac{t}{2}\right) = \sum_{n\in\mathbb{Z}} h_1[n]\,\varphi(t-n), \qquad \widehat{\psi}(2\omega) = \frac{1}{\sqrt{2}} H_1(\omega)\widehat{\varphi}(\omega). \tag{12}$$

Then $\{\psi_{j,k}\}_{k\in\mathbb{Z}}$ is an orthonormal basis of W_j for every j, and $\{\psi_{j,k}\}_{(j,k)\in\mathbb{Z}^2}$ is an orthonormal basis of $L^2(\mathbb{R})$.

Note that $H_1(0) = \overline{H_0(\pi)} = 0$, so that $\widehat{\psi}(0) = 0$: the wavelet automatically has zero mean, as required by the admissibility condition (2).

Example: Haar. Let $\varphi = \mathbf{1}_{[0,1)}$. The two-scale equation (9) reads

$$\varphi(t/2) = \mathbf{1}_{[0,2)}(t) = \varphi(t) + \varphi(t-1), \qquad \text{i.e.} \qquad h_0 = \left(\tfrac{1}{\sqrt{2}}, \tfrac{1}{\sqrt{2}}\right) \text{ (average).}$$

The formula of the theorem gives $h_1[0] = -h_0[1] = -\frac{1}{\sqrt{2}}$ and $h_1[1] = h_0[0] = \frac{1}{\sqrt{2}}$ (difference), hence

$$\psi(t/2) = \varphi(t-1) - \varphi(t) \qquad \Longleftrightarrow \qquad \psi = \mathbf{1}_{[1/2,1)} - \mathbf{1}_{[0,1/2)}.$$

On each interval of length 2^j, the approximation at scale 2^j keeps the mean of the two half-intervals, while the wavelet coefficient keeps (up to a factor) their difference: together, they recover the two values exactly.

Finite-depth decomposition. Applying $V_{j-1} = V_j \oplus W_j$ repeatedly, from V_0 down to V_J, gives for every $J \geq 1$ and every $f \in V_0$

$$f = \underbrace{\sum_k a_J[k]\,\varphi_{J,k}}_{\text{coarse approximation } P_{V_J} f} + \sum_{j=1}^{J} \underbrace{\sum_k d_j[k]\,\psi_{j,k}}_{\text{details } P_{W_j} f}, \qquad d_j[k] = \langle f \,|\, \psi_{j,k} \rangle \quad \text{(wavelet coefficients)}, \tag{13}$$

and, by orthonormality, $\|f\|^2 = \sum_k |a_J[k]|^2 + \sum_{j=1}^{J} \sum_k |d_j[k]|^2$. A signal is thus a coarse approximation (one function φ, averages) plus a hierarchy of details (one function ψ, differences), one level per octave.

4.3 Low and high frequencies

The pair of filters (h_0, h_1) splits the frequency content of a signal in two complementary halves:

- $H_0(0) = \sqrt{2}$ and $H_0(\pi) = 0$: h_0 keeps the **low frequencies**, i.e. the slowly varying, coarse content (local averages);
- $H_1(0) = 0$ and $|H_1(\pi)| = \sqrt{2}$: h_1 keeps the **high frequencies**, i.e. the fast variations (local differences, edges, textures, noise);
- the split is *energy preserving*:

$$|H_0(\omega)|^2 + |H_1(\omega)|^2 = 2 \qquad \forall \omega. \tag{14}$$

Iterating the split on the low-pass branch only yields a *dyadic partition of the frequency axis*: at the first level, details capture roughly $[\pi/2, \pi]$, at the second level $[\pi/4, \pi/2]$, ..., and the coarse approximation a_J keeps $[0, 2^{-J}\pi]$ (Figure 6). In other words, V_j contains (approximately) the frequencies below $2^{-j}\pi$ and W_j the band $[2^{-j}\pi, 2^{-j+1}\pi]$. The bands are sharp for the Shannon wavelet and overlap for compactly supported wavelets, but they always sum to the identity in energy, which is why perfect reconstruction holds.

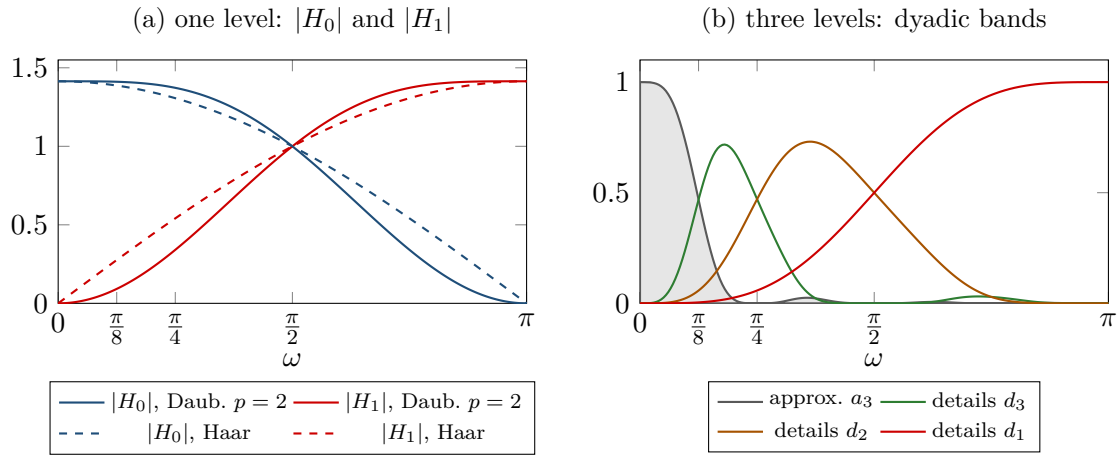


Figure 6: Frequency splitting by orthogonal wavelet filters. (a) Low-pass h_0 and high-pass h_1 filters for Haar (dashed) and Daubechies with two vanishing moments (solid); they are power complementary (14). (b) Squared moduli of the equivalent filters of a three-level decomposition (normalized by 2^{-j}): the details d_1, d_2, d_3 occupy successive octaves and a_3 keeps the lowest frequencies; the four curves sum to 1.

5 Discrete wavelet transform: filter banks and Mallat's algorithm

5.1 Recursive computation

Computing (8) and (13) by numerical integration would be expensive. The two-scale relations provide recursions between successive levels. Inserting (9) and (12) at level j in the definitions of a_j and d_j gives [19]:

Theorem 5.1 – (Mallat's algorithm)

For every $j \geq 1$ and $p \in \mathbb{Z}$,

$$\text{(analysis)} \quad a_j[p] = \sum_{n\in\mathbb{Z}} h_0[n-2p]\, a_{j-1}[n] = (a_{j-1} * \bar{h}_0)[2p],$$

$$d_j[p] = \sum_{n\in\mathbb{Z}} h_1[n-2p]\, a_{j-1}[n] = (a_{j-1} * \bar{h}_1)[2p], \tag{15}$$

$$\text{(synthesis)} \quad a_{j-1}[p] = \sum_{n\in\mathbb{Z}} h_0[p-2n]\, a_j[n] + \sum_{n\in\mathbb{Z}} h_1[p-2n]\, d_j[n], \tag{16}$$

where $\bar{h}[n] = h[-n]$.

Each analysis step is a *convolution followed by a decimation by 2*; each synthesis step is an *upsampling by 2 (inserting zeros) followed by a convolution* and a sum. This is a two-channel perfect reconstruction filter bank (Figure 7); cascading it on the low-pass branch gives the multilevel discrete wavelet transform (DWT).

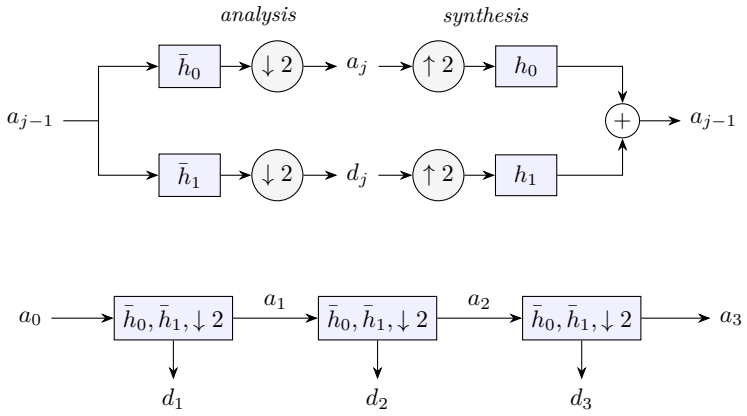


Figure 7: Top: one level of the two-channel filter bank (analysis then synthesis); with conjugate mirror filters the output equals the input. Bottom: three-level DWT obtained by cascading the analysis bank on the approximation coefficients.

Initialization. The recursion needs the coefficients $a_0[n] = \langle f \,|\, \varphi(\cdot - n)\rangle$ at the finest level. In practice one only has samples and sets $a_0[n] \approx f(n)$. This approximation propagates through all scales; refined initializations exist when the continuous interpretation matters, e.g. for regularity estimation [11]. In image restoration, the image (a vector of pixels) is simply identified with a_0, and the DWT is regarded as a linear operator on $\mathbb{R}^N$.

5.2 The DWT as an orthogonal matrix

Consider signals of length $N = 2^L$ with periodic boundary conditions (indices modulo N) and define the decimated filtering operators $H_0, H_1 \in \mathbb{R}^{N/2\times N}$,

$$(H_0 x)[p] = \sum_{n=0}^{N-1} h_0[n-2p]\, x[n], \qquad (H_1 x)[p] = \sum_{n=0}^{N-1} h_1[n-2p]\, x[n].$$

By (16), their adjoints $H_0^\top, H_1^\top$ are "upsample then filter". One level of decomposition is the square matrix $W^{(1)} = \begin{bmatrix} H_0 \\ H_1 \end{bmatrix} \in \mathbb{R}^{N\times N}$.

Proposition 5.2 – (The DWT is orthogonal)

If (h_0, h_1) satisfy (11) and h_1 is defined as in Theorem 4.3, then (for filters shorter than N)

$$H_0 H_0^\top = H_1 H_1^\top = I_{N/2}, \qquad H_0 H_1^\top = 0, \qquad H_0^\top H_0 + H_1^\top H_1 = I_N,$$

i.e. $W^{(1)}$ is an orthogonal matrix. The J-level DWT

$$W = \begin{bmatrix} W^{(1)}_{N/2^{J-1}} & 0 \\ 0 & I \end{bmatrix} \cdots \begin{bmatrix} W^{(1)}_{N/2} & 0 \\ 0 & I_{N/2} \end{bmatrix} W^{(1)}_N$$

(each factor acts on the current approximation only) is orthogonal as a product of orthogonal matrices: $W^\top W = W W^\top = I_N$, and $W^{-1} = W^\top$ is the inverse DWT.

Cost and structure of the DWT

- With filters of length ℓ, level j costs about $2\ell\, N/2^{j-1}$ operations, hence a total $\le 4\ell N = O(N)$, faster than the FFT ($O(N \log N)$).

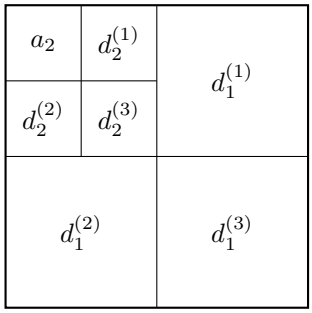
- W is never stored: Wx and $W^\top c$ are computed by (15)–(16).
- W is orthogonal: $\|Wx\| = \|x\|$, analysis and synthesis are adjoint, and the condition number is 1.

5.3 Images: separable 2D wavelets

For images, the simplest construction uses tensor products of the 1D functions [20, 23]:

$$\varphi^{(0)}(x_1, x_2) = \varphi(x_1)\varphi(x_2), \quad \psi^{(1)} = \psi(x_1)\varphi(x_2), \quad \psi^{(2)} = \varphi(x_1)\psi(x_2), \quad \psi^{(3)} = \psi(x_1)\psi(x_2),$$

and $\{2^{-j}\psi^{(m)}(2^{-j}\boldsymbol{x} - \boldsymbol{n})\}_{m\in\{1,2,3\},\, j\in\mathbb{Z},\, \boldsymbol{n}\in\mathbb{Z}^2}$ is an orthonormal basis of $L^2(\mathbb{R}^2)$. Numerically, one level of the 2D DWT applies the 1D filter bank to the rows, then to the columns, which yields four subbands: one approximation and three detail subbands, sensitive to variations along x_1, along x_2 and along diagonals. The decomposition is iterated on the approximation (Figure 8). The transform remains orthogonal, with an $O(N)$ cost for an image of N pixels.



$d_j^{(1)}$: high-pass along x_1, low-pass along x_2
$d_j^{(2)}$: low-pass along x_1, high-pass along x_2
$d_j^{(3)}$: high-pass along both directions (diagonal)
a_J: low-pass along both directions
Each level divides the size by 2×2;
the total number of coefficients is N.

Figure 8: Layout of a two-level separable 2D wavelet decomposition.

Remark (Limitations of separable real wavelets) Whatever the 1D wavelet, separable real wavelets offer only three orientations, and the diagonal subband cannot distinguish $+45°$ from $-45°$ (the Fourier transform of a real function is Hermitian, so the spectral support of $\psi^{(3)}$ is spread over the four quadrants: *checkerboard* effect). Moreover, because of decimation, the DWT is *not translation invariant*: shifting the image by one pixel can deeply change its coefficients, which creates artefacts after thresholding (ringing near edges). These limitations motivate redundant transforms (frames, Section 7) such as the undecimated wavelet transform or the dual-tree complex wavelet transform [25, 11].

6 Vanishing moments, regularity and sparsity

6.1 Vanishing moments

Definition 6.1 – (Vanishing moments)

ψ has n_ψ *vanishing moments* if $\int_{\mathbb{R}} t^m \psi(t)\, \mathrm{d}t = 0$ for $m = 0, \ldots, n_\psi - 1$.

Theorem 6.2 – (Characterizations of vanishing moments [20])

Let φ and ψ be the scaling function and wavelet of an orthonormal basis, with $|\varphi(t)|, |\psi(t)| = O((1+t^2)^{-n_\psi/2-1})$. The following are equivalent:

(i) ψ has n_ψ vanishing moments;

(ii) $\widehat{\psi}$ and its first $n_\psi - 1$ derivatives vanish at $\omega = 0$;

(iii) (Strang–Fix) for $0 \le m < n_\psi$, $t \mapsto \sum_{k\in\mathbb{Z}} k^m \varphi(t-k)$ is a polynomial of degree m: the approximation spaces reproduce polynomials of degree $< n_\psi$.

A wavelet with n_ψ vanishing moments is orthogonal to all polynomials of degree $< n_\psi$: it is "blind" to locally polynomial behaviour and only responds to what deviates from it. In the Fourier domain, (ii) says that $\widehat{\psi}$ is very flat at 0: ψ is a high-pass analysis that discards smooth trends.

6.2 Regularity and decay of wavelet coefficients

Theorem 6.3 – (Local regularity)

Let ψ be compactly supported with n_ψ vanishing moments, and let f be α-Hölder on an interval I, with $0 < \alpha < n_\psi$: for every $t_0 \in I$ there is a polynomial p_{t_0} of degree $< n_\psi$ with $|f(t) - p_{t_0}(t)| \le K|t - t_0|^\alpha$ for $t \in I$. Then, for every (j,k) such that $\operatorname{supp}\psi_{j,k} \subset I$,

$$|d_j[k]| = |\langle f \,|\, \psi_{j,k}\rangle| \le C\,K\,2^{j(\alpha+1/2)}, \qquad C = \int_{\mathbb{R}} |u|^\alpha |\psi(u)|\,\mathrm{d}u.$$

Proof. Let $t_0 = 2^j k$. Since $\psi_{j,k}$ is orthogonal to p_{t_0}, $\langle f \,|\, \psi_{j,k}\rangle = \langle f - p_{t_0} \,|\, \psi_{j,k}\rangle$, so that

$$|d_j[k]| \le \int K|t - 2^j k|^\alpha\, 2^{-j/2} |\psi(2^{-j}t - k)|\,\mathrm{d}t = K\,2^{j(\alpha+1/2)} \int |u|^\alpha |\psi(u)|\,\mathrm{d}u,$$

with the change of variable $t = 2^j(u + k)$. ∎

6.3 Sparsity and nonlinear approximation

Let $(c_k)_k = Wx$ be the coefficients of x in an orthonormal basis and let x_M be obtained by keeping the M largest $|c_k|$ and setting the others to zero (best M-term approximation). By orthonormality,

$$\|x - x_M\|^2 = \sum_{k \notin M \text{ largest}} |c_k|^2,$$

so the approximation error is directly governed by the decay of the *sorted* coefficients: the sparser the representation, the better the approximation. For a piecewise C^α function with finitely many discontinuities, one can show [20, Chap. 9] that the squared error decays like M^{-1} in a Fourier basis but like $M^{-2\alpha}$ in a wavelet basis with $n_\psi > \alpha$ vanishing moments.

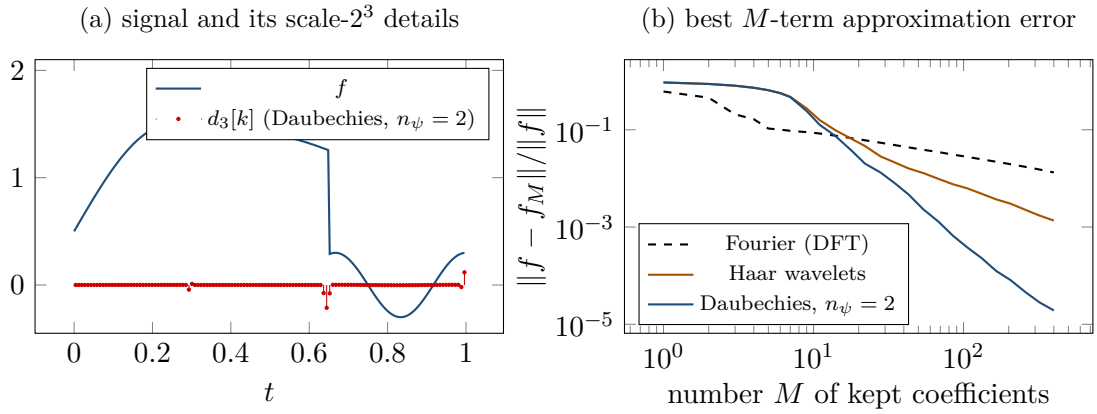


Figure 9: Sparsity of wavelet representations ($N = 1024$, $J = 6$ levels, periodic boundary). (a) A piecewise smooth signal (the periodisation adds a jump at $t = 1 \equiv 0$) and its detail coefficients at scale 2^3: they are negligible except around the discontinuities. (b) Relative error of the best M-term approximation (keep the M largest coefficients in modulus). With the same budget, wavelets approximate the signal far better than Fourier, and more vanishing moments help.

Statistical view and ℓ_1 priors. The histograms of the detail coefficients of natural images are sharply peaked at 0 with heavy tails, well modelled by Laplacian or generalized Gaussian distributions [24]. The corresponding negative log-likelihood is the ℓ_1 norm, which leads to the classical regularizer

$$g(x) = \lambda \|Wx\|_1 = \lambda \sum_{j,k} |(Wx)_{j,k}|, \tag{17}$$

possibly with scale-dependent weights λ_j and without penalizing the coarse approximation a_J. This sparsity is also what makes wavelets efficient for compression (JPEG2000) and denoising: *small coefficients are mostly noise, large ones mostly signal* [12]. We will see in Section 8 that wavelet thresholding is nothing but a proximal operator.

7 Frames: redundant wavelet representations

7.1 Motivation

The CWT is very redundant and expensive; orthonormal bases are economical but suffer from the limitations discussed at the end of Section 5.3 (shift variance, few orientations). Frames are the intermediate solution: redundant, but discrete and stable [13, 23, 24]. Redundancy brings flexibility (more atoms, better adapted to edges and textures) and approximate shift invariance, which often improves reconstruction quality in inverse problems.

7.2 Definition and associated operators

We work in finite dimension: $\mathcal{H} = \mathbb{R}^N$ and a family $(e_k)_{1 \le k \le K}$ of vectors of $\mathbb{R}^N$, with $K \ge N$.

Definition 7.1 – (Frame)

$(e_k)_{1 \le k \le K}$ is a *frame* if there exist $0 < \underline{\mu} \le \overline{\mu}$ such that

$$\underline{\mu} \|x\|^2 \le \sum_{k=1}^{K} |\langle x \,|\, e_k \rangle|^2 \le \overline{\mu} \|x\|^2 \qquad \forall x \in \mathbb{R}^N. \tag{18}$$

It is a *tight frame* if $\underline{\mu} = \overline{\mu} = \nu$ (a *Parseval frame* if $\nu = 1$).

- The **analysis operator** is $F : \mathbb{R}^N \to \mathbb{R}^K$, $x \mapsto (\langle x \,|\, e_k \rangle)_k$, i.e. the matrix whose rows are the $e_k^\top$.
- The **synthesis operator** is its adjoint $F^\top : \mathbb{R}^K \to \mathbb{R}^N$, $c \mapsto \sum_k c_k e_k$.
- The **frame operator** is $S = F^\top F$. Condition (18) means $\underline{\mu} I \preceq S \preceq \overline{\mu} I$: the optimal bounds are the extreme eigenvalues of S, and $\overline{\mu} = \|F\|^2$.

In finite dimension, a family is a frame if and only if it spans $\mathbb{R}^N$ (so F is injective and $F^\top$ surjective). The vector x is recovered from its coefficients by the *canonical dual frame* $\tilde{e}_k = S^{-1} e_k$:

$$x = \sum_k \langle x \,|\, e_k \rangle \, \tilde{e}_k = S^{-1} F^\top F x = \sum_k \langle x \,|\, \tilde{e}_k \rangle \, e_k.$$

For a tight frame $S = \nu I$, so that $x = \nu^{-1} F^\top F x$: *the pseudo-inverse of F is $\nu^{-1} F^\top$*. An orthonormal basis is a Parseval frame with $K = N$, for which $F^{-1} = F^\top$.

Non-uniqueness of the coefficients. When $K > N$, $\ker F^\top \neq \{0\}$ and $x = F^\top c$ holds for every $c \in S^{-1} F x + \ker F^\top$ (or $\nu^{-1} F x + \ker F^\top$ for a tight frame). The analysis coefficients Fx are only one possible choice. This degree of freedom is exploited by sparse synthesis models, which select, among all coefficient vectors representing x, a sparse one.

Proposition 7.2 – (A crucial asymmetry)

Let $F \in \mathbb{R}^{K \times N}$ be the analysis operator of a tight frame with constant ν. Then $F^\top F = \nu I_N$, but $F F^\top = \nu P_{\operatorname{ran} F}$, where $P_{\operatorname{ran} F}$ is the orthogonal projector onto the N-dimensional range of F. Hence $F F^\top = \nu I_K$ if and only if $K = N$, i.e. the frame is (up to the factor $\sqrt{\nu}$) an orthonormal basis.

Proof. $(F F^\top)^2 = F (F^\top F) F^\top = \nu F F^\top$ and $F F^\top$ is symmetric, so $\nu^{-1} F F^\top$ is an orthogonal projector, whose range is $\operatorname{ran} F$ (of dimension N since F is injective). It equals I_K if and only if $N = K$. ■

This seemingly innocent asymmetry is exactly what will make proximal operators of analysis priors $x \mapsto g(Fx)$ non-explicit for redundant frames (Section 8).

7.3 Examples of wavelet frames

- **Union of orthonormal bases.** If $W_1, \ldots, W_\nu$ are orthogonal matrices, $F = [W_1; \ldots; W_\nu]$ satisfies $F^\top F = \sum_i W_i^\top W_i = \nu I$: a tight frame with constant ν. A real (resp. complex) dual-tree wavelet transform is a union of 2 (resp. 4) orthonormal wavelet bases [25, 24]; in 2D it provides six oriented, nearly analytic subbands per scale and near shift invariance [11].
- **Undecimated (stationary, *à trous*) wavelet transform.** Removing the decimations of Mallat's algorithm and upsampling the filters by inserting $2^{j-1} - 1$ zeros between their taps at level j [17] yields a translation invariant tight frame with $K = (J+1)N$ coefficients (in 1D). It is equivalent to averaging the DWT over all circular shifts of the signal (*cycle spinning* [4]). The undecimated Haar transform is closely related to the discrete gradient, hence to total variation [24].

- **Discretized CWT.** Sampling the CWT on $(s,u) = (\gamma^j, b\gamma^j k)$ with $\gamma > 1$, $b > 0$ small enough gives a frame of $L^2(\mathbb{R})$ under mild conditions on ψ [23], with much more freedom on the shape of ψ.
- Other redundant multiscale transforms: curvelets, contourlets, steerable pyramids, Gabor frames [24].

8 Wavelets and proximal operators

8.1 Setting

We consider the restoration of $\bar{x} \in \mathbb{R}^N$ from $z = \mathrm{A}\bar{x} + \varepsilon$, with a linear degradation A (blur, tomographic projection, masking) and Gaussian noise ε. With a wavelet (or frame) transform $F \in \mathbb{R}^{K\times N}$, two formulations coexist [14, 24]:

$$\text{(analysis)} \qquad \hat{x} \in \operatorname*{Argmin}_{x\in\mathbb{R}^N} \tfrac{1}{2}\|\mathrm{A}x - z\|^2 + \lambda \|Fx\|_1 , \tag{19}$$

$$\text{(synthesis)} \qquad \hat{x} = F^\top \hat{c}, \quad \hat{c} \in \operatorname*{Argmin}_{c\in\mathbb{R}^K} \tfrac{1}{2}\left\|\mathrm{A}F^\top c - z\right\|^2 + \lambda \|c\|_1 . \tag{20}$$

Both have the structure $\min h + g$ with h smooth (gradient Lipschitz) and g convex non-smooth, and are naturally solved by proximal algorithms such as forward–backward (ISTA/FISTA). Their practicality depends on our ability to compute prox_g.

Recall that for $g \in \Gamma_0(\mathbb{R}^N)$, $\mathrm{prox}_g(x) = \arg\min_y \frac{1}{2}\|y - x\|^2 + g(y)$, that the proximal operator of a separable function $g(y) = \sum_k g_k(y_k)$ acts componentwise, and that

$$\mathrm{prox}_{\lambda|\cdot|}(\xi) = \mathrm{soft}_\lambda(\xi) = \mathrm{sign}(\xi)\max\{|\xi| - \lambda, 0\}. \tag{21}$$

Thus $\mathrm{prox}_{\lambda\|\cdot\|_1}$ is the componentwise soft thresholding: the proximal operator of the ℓ_1-norm *on the coefficients* is always explicit. The difficulty lies in the composition with F in the analysis formulation.

8.2 Orthonormal bases: explicit proximal operators

Proposition 8.1 – (Proximal operator in an orthonormal basis)

Let $W \in \mathbb{R}^{N\times N}$ be orthogonal and $g \in \Gamma_0(\mathbb{R}^N)$. Then

$$\mathrm{prox}_{g\circ W}(x) = W^\top \mathrm{prox}_g(Wx) \qquad \forall x \in \mathbb{R}^N.$$

In particular $\mathrm{prox}_{\lambda\|W\cdot\|_1}(x) = W^\top \mathrm{soft}_\lambda(Wx)$.

Proof. With the change of variable $c = Wy$ ($y = W^\top c$) and $\|y - x\| = \|Wy - Wx\|$,

$$\min_y \tfrac{1}{2}\|y - x\|^2 + g(Wy) = \min_c \tfrac{1}{2}\|c - Wx\|^2 + g(c),$$

whose minimizer is $\hat{c} = \mathrm{prox}_g(Wx)$, hence $\hat{y} = W^\top \hat{c}$. ∎

Consequences of orthonormality

- **Cost**: computing $\operatorname{prox}_{\lambda\|W\cdot\|_1}$ costs one DWT, one thresholding and one inverse DWT, i.e. $O(N)$.
- **Denoising** ($\mathrm{A} = \mathrm{Id}$): the solution of (19) is exactly *wavelet soft thresholding* [12]: transform, threshold, reconstruct.
- **Analysis = synthesis**: since $c = Wx \iff x = W^\top c$, (19) and (20) are equivalent.
- **Forward–backward / ISTA** [15, 10, 7]: for $\tau \in]0, 2/\|\mathrm{A}\|^2[$,

$$x^{[k+1]} = W^\top \operatorname{soft}_{\tau\lambda}\Big(W\big(x^{[k]} - \tau\mathrm{A}^\top(\mathrm{A}x^{[k]} - z)\big)\Big).$$

- The same holds for any separable g: scale-dependent weights λ_j, generalized Gaussian penalties $\lambda|\cdot|^p$, group penalties over subbands or colour channels, box constraints on coefficients [3, 24].

Proposition 8.1 is a special case of a more general rule.

Proposition 8.2 – ([6], [1, Prop. 24.14])

Let $L \in \mathbb{R}^{K\times N}$ be such that $LL^\top = \nu I_K$ for some $\nu > 0$, and $g \in \Gamma_0(\mathbb{R}^K)$. Then

$$\operatorname{prox}_{g\circ L} = \mathrm{Id} + \nu^{-1}L^\top(\operatorname{prox}_{\nu g} - \mathrm{Id})L.$$

For $L = W$ orthogonal ($\nu = 1$) we recover Proposition 8.1. For a redundant tight frame, however, the assumption concerns $FF^\top$, which is *not* a multiple of the identity (Proposition 7.2): the formula does not apply and, in general, $\operatorname{prox}_{\lambda\|F\cdot\|_1}$ has no closed form.

Warning **A common pitfall.** For a tight frame, the operator $x \mapsto \nu^{-1}F^\top \operatorname{soft}_\lambda(Fx)$ (analysis, thresholding, pseudo-inverse synthesis; e.g. cycle spinning) is an excellent denoiser, but it is *not* $\operatorname{prox}_{\lambda\|F\cdot\|_1}$. See **??** for a one-dimensional counterexample.

8.3 Redundant frames: three ways to handle the proximal step

(a) Switch to the synthesis formulation. In (20) the non-smooth term is $\lambda\|c\|_1$, whose proximal operator is soft thresholding, and the smooth term involves $AF^\top$. Forward–backward reads

$$c^{[k+1]} = \operatorname{soft}_{\tau\lambda}\left(c^{[k]} - \tau F\mathrm{A}^\top(\mathrm{A}F^\top c^{[k]} - z)\right), \qquad \tau \in]0, 2/\left\|\mathrm{A}F^\top\right\|^2[,$$

with $\left\|\mathrm{A}F^\top\right\|^2 \leq \nu\|\mathrm{A}\|^2$ for a tight frame; the step size can be estimated by power iterations [24]. This is cheap, but for redundant frames (20) is *not* equivalent to (19) in general: the analysis problem is a particular case of the synthesis one, and both coincide for orthonormal bases [14, 24]. Analysis priors often perform better on natural images.

(b) Compute the proximal operator iteratively, by duality. The proximal operator of an analysis term is the solution of a small optimization problem, which can be solved through its dual.

(c) Avoid the proximal operator of the composition: primal–dual algorithms. Primal–dual splitting methods [2, 8, 27] only require F, $F^\top$ and the proximal operator of g

Transform	$F^\top F$	$FF^\top$	$\operatorname{prox}_{\lambda\|F\cdot\|_1}$	Typical algorithms
Orthonormal basis ($K = N$)	I_N	I_N	$F^\top \operatorname{soft}_\lambda(F\,\cdot)$, $O(N)$	ISTA / FISTA
Tight frame ($K > N$)	νI_N	$\nu P_{\operatorname{ran} F}$	no closed form; dual inner loop	synthesis FB, inexact FB, primal–dual
General frame	S	–	no closed form; dual inner loop	primal–dual (Condat–Vũ)

Table 2: Proximal computations for wavelet-based ℓ_1 regularization.

Take-home message

Orthonormality makes the proximal operator of wavelet-sparsity priors explicit and $O(N)$: this is the reason why orthonormal wavelets are the workhorse of fast restoration algorithms. Redundant frames give better images but break this property, because $FF^\top \neq \nu I$; their proximal step is then either moved to the coefficient domain (synthesis), computed iteratively by duality, or bypassed by primal–dual algorithms.

9 Towards multilevel optimization

The MRA provides, for free, the ingredients of a multilevel method. In the discrete setting of Section 5.2, define

$$R = H_0 \in \mathbb{R}^{N/2\times N} \text{ (restriction)}, \qquad P = R^\top = H_0^\top \text{ (prolongation)}.$$

- R maps a fine signal to its approximation coefficients at the next coarser scale: the coarse level is V_1. In 2D, $R = H_0 \otimes H_0$ keeps the a subband, with $N/4$ unknowns. For Haar, R computes (normalized) averages of pairs of samples, or of 2×2 blocks of pixels.
- By Proposition 5.2, $RR^\top = I$ and $PR = H_0^\top H_0$ is the orthogonal projector onto the coarse space, while $I - PR = H_1^\top H_1$ is the projector onto the details: a fine vector splits exactly into a low-frequency part, representable at the coarse level, and a high-frequency part.
- The construction is recursive: applying R again yields $V_2, V_3, \ldots$, i.e. a whole hierarchy of levels, with costs divided by 2 (1D) or 4 (2D) at each level.
- The wavelet regularizer is consistent across levels: by Mallat's algorithm, the $(J-1)$-level DWT of the coarse vector Rx consists exactly of the coarse-scale coefficients $(d_2, \ldots, d_J, a_J)$ of x. A natural coarse version of $\lambda \|Wx\|_1$ is therefore $\lambda \|W_{J-1} x_H\|_1$, whose proximal operator is again explicit.

Multilevel algorithms exploit this structure: they compute cheap corrections on coarse levels, where most of the energy of natural images lives, prolongate them with P, and refine on the fine level. How the fine iterations and the coarse corrections complement each other depends on the spectral properties of the problem (e.g. blur operators act as low-pass filters), and will be studied in the next chapters. Several articles of the course project build on this wavelet hierarchy, notably the multilevel algorithm of Vonesch and Unser for wavelet-regularized restoration [26] and multilevel plug-and-play methods [18].

References

[1] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, 2nd ed., Springer, Cham, 2017.

[2] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging", *J. Math. Imaging Vis.*, 40(1):120–145, 2011.

[3] C. Chaux, P. L. Combettes, J.-C. Pesquet and V. R. Wajs, "A variational formulation for frame-based inverse problems", *Inverse Problems*, 23(4):1495–1518, 2007.

[4] R. R. Coifman and D. L. Donoho, "Translation-invariant de-noising", in *Wavelets and Statistics*, Springer, 1995, pp. 125–150.

[5] P. L. Combettes, Đ. Dũng and B. C. Vũ, "Dualization of signal recovery problems", *Set-Valued Var. Anal.*, 18(3–4):373–404, 2010.

[6] P. L. Combettes and J.-C. Pesquet, "A Douglas–Rachford splitting approach to nonsmooth convex variational signal recovery", *IEEE J. Sel. Topics Signal Process.*, 1(4):564–574, 2007.

[7] P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forward-backward splitting", *Multiscale Modeling & Simulation*, 4(4):1168–1200, 2005.

[8] L. Condat, "A primal–dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms", *J. Optim. Theory Appl.*, 158(2):460–479, 2013.

[9] I. Daubechies, *Ten Lectures on Wavelets*, SIAM, Philadelphia, 1992.

[10] I. Daubechies, M. Defrise and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint", *Comm. Pure Appl. Math.*, 57(11):1413–1457, 2004.

[11] L. Davy, *Segmentation de textures anisotropes et invariantes d'échelles : ondelettes complexes multibandes, optimisation variationnelle et biniveau, réseaux de neurones informés*, PhD thesis, ENS de Lyon, 2025 (Chapter 2). HAL: tel-05502833.

[12] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage", *Biometrika*, 81(3):425–455, 1994.

[13] R. J. Duffin and A. C. Schaeffer, "A class of nonharmonic Fourier series", *Trans. Amer. Math. Soc.*, 72:341–366, 1952.

[14] M. Elad, P. Milanfar and R. Rubinstein, "Analysis versus synthesis in signal priors", *Inverse Problems*, 23(3):947–968, 2007.

[15] M. A. T. Figueiredo and R. D. Nowak, "An EM algorithm for wavelet-based image restoration", *IEEE Trans. Image Process.*, 12(8):906–916, 2003.

[16] A. Grossmann and J. Morlet, "Decomposition of Hardy functions into square integrable wavelets of constant shape", *SIAM J. Math. Anal.*, 15(4):723–736, 1984.

[17] M. Holschneider, R. Kronland-Martinet, J. Morlet and Ph. Tchamitchian, "A real-time algorithm for signal analysis with the help of the wavelet transform", in *Wavelets: Time-Frequency Methods and Phase Space*, Springer, 1989, pp. 286–297.

[18] N. Laurent, J. Tachella, E. Riccietti and N. Pustelnik, "Multilevel plug-and-play image restoration", *IEEE Trans. Comput. Imaging*, 2025.

[19] S. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation", *IEEE Trans. Pattern Anal. Mach. Intell.*, 11(7):674–693, 1989.

[20] S. Mallat, *A Wavelet Tour of Signal Processing: The Sparse Way*, 3rd ed., Academic Press, 2009.

[21] Y. Meyer, *Wavelets and Operators*, Cambridge University Press, 1992.

[22] B. Pascal, "Multi-resolution in image processing: preliminaries on wavelets", lecture slides, 2026.

[23] B. Pesquet-Popescu and J.-C. Pesquet, "Ondelettes et applications", *Techniques de l'Ingénieur*, TE 5215, 2001.

[24] N. Pustelnik, A. Benazza-Benhayia, Y. Zheng and J.-C. Pesquet, "Wavelet-based image deconvolution and reconstruction", *Wiley Encyclopedia of Electrical and Electronics Engineering*, 2016. HAL: hal-01164833.

[25] I. W. Selesnick, R. G. Baraniuk and N. G. Kingsbury, "The dual-tree complex wavelet transform", *IEEE Signal Process. Mag.*, 22(6):123–151, 2005.

[26] C. Vonesch and M. Unser, "A fast multilevel algorithm for wavelet-regularized image restoration", *IEEE Trans. Image Process.*, 18(3):509–523, 2009.

[27] B. C. Vũ, "A splitting algorithm for dual monotone inclusions involving cocoercive operators", *Adv. Comput. Math.*, 38(3):667–681, 2013.