One of the simplest and surprisingly effective approach for removing noise in images is to expand the signal in an orthogonal wavelet basis, to apply a soft-threshold to the wavelet coefficients, and to reconstruct the “denoised” image by inverse wavelet transformation. The classical justification for the algorithm is that i.i.d. noise is spread out uniformly in the wavelet domain while the signal gets concentrated in a few significant coefficients (sparsity property) so that the smaller values can be primarily attributed to noise and easily suppressed.

In this chapter, we take advantage of our statistical framework to revisit such wavelet-based reconstruction methods. Our first objective is to present some alternative dictionary-based techniques for the resolution of general inverse problems based on the same stochastic models as in Chapter 10. Our second goal is to take advantage of the orthogonality of wavelets to get a deeper understanding of the effect of proximal operators while investigating the possibility of optimizing shrinkage-thresholding functions for better performance. Finally, we shall attempt to bridge the gap between operator-based regularization, as discussed in Sections 10.2-10.3, and the imposition of sparsity constraints in the wavelet domain. Fundamentally, this relates to the dichotomy between an analysis point of view of the problem (typically in the form of the minimization of an energy functional with a regularization term) versus a synthesis point of view where a signal is represented as a sum of elementary constituents (wavelets.)

The chapter is composed of two main parts. The first is devoted to inverse problems in general. Specifically, in Section 11.1, we apply our general discretization and modeling paradigm to the derivation of wavelet-domain MAP estimators for the resolution of linear inverse problems. One of the key difference with the innovation-based formulation of Chapter 10 is the presence of scale-dependent potential functions whose form is specified by the stochastic model. We then address practical issues in Section 11.2 with the presentation of the two primary iterative thresholding algorithms (ISTA and FISTA). These methods are illustrated with the deconvolution of fluorescence micrographs.

The second part of the chapter focuses on the denoising problem with the aim of improving upon simple soft-thresholding and wavelet-domain MAP estimation. Section 11.3 presents a detailed investigation of shrinkage functions in relation to infinitely-divisible laws with the emphasis on pointwise estimators that are optimal in the MMSE sense. In Section 11.4, we show how the performance of wavelet denois-
Wavelet-domain methods

ing can be boosted even further through the use of redundant representations (tight wavelet frames). In particular, we describe the concept of consistent cycle spinning which provides a conceptual bridge with the optimal estimation techniques of Section 10.4. We then close the circle in Section 11.4.4 by combining all ingredients—tight operator-like wavelet frames, MMSE shrinkage functions, and consistent cycle spinning—and present an iterative wavelet-based algorithm that converges empirically to the reference MMSE solution of Section 10.4.2.

11.1 Discretization of inverse problems in a wavelet basis

As alternative to the shift-invariant formulation presented in Chapter 10, we may choose to discretize a linear inverse problem in a wavelet basis. To that end, we consider a biorthogonal wavelet system of the type investigated in Chapter 8 that is matched to the whitening operator $L$. The underlying signal representation is the wavelet counterpart of (10.2) in Section 10.1. It is given by

$$s_1(r) = \sum_{i=1}^{\infty} \sum_{k \in \mathbb{Z}^d \cap D^i} v_i[k] \psi_{i,k}(r) = \sum_{k \in \mathbb{Z}^d} s[k] \beta_L(r - k)$$ (11.1)

where the wavelet coefficients are obtained as

$$v_i[k] = \langle s, \psi_{i,k} \rangle.$$

The mathematical requirement is that the family of analysis/synthesis functions $(\tilde{\psi}_{i,k}, \tilde{\varphi}_{i,k})$ forms a biorthonormal wavelet basis.

Observe that the central wavelet expansion in (11.1) excludes the finer-scale wavelet coefficients with $i < 1$, so that the signal approximation $s_1(r)$, which is the projection of $s(r)$ onto the reference space $V_0$, can also be represented as a linear combination of the integer shifts of the scaling function $\beta_L$.

The crucial ingredient for our formulation (see Section 6.5.3) is that the analysis wavelets are such that

$$\tilde{\psi}_{i,k}(r) = L^* \tilde{\varphi}_i(r - D^{i-1} k)$$ (11.2)

where $\tilde{\varphi}_i \in L_1(\mathbb{R}^d)$ is some suitable (possibly, scale-dependent) smoothing kernel and $D$ the dilation matrix that specifies the multiresolution decomposition. Recalling that $s = L^{-1} w$, this implies that

$$v_i[k] = \langle s, \psi_{i,k} \rangle = \langle L^{-1} w, L^* \tilde{\varphi}_i(\cdot - D^{i-1} k) \rangle = \langle w, \tilde{\varphi}_i(\cdot - D^{i-1} k) \rangle$$

so that it is possible to derive any finite-dimensional joint pdf of the wavelet coefficients $v_i[k]$ by using the general white-noise analysis exposed in Chapters 8 and 9. In particular, Proposition 8.6 tells us that $p_{V_i}$, the pdf of the wavelet coefficients at scale $i$, is infinitely divisible with modified Lévy exponent $f_{\tilde{\varphi}_i}(\omega) = \int_{\mathbb{R}^d} f(\omega \psi_i(r)) \, dr$. 


11.1 Discretization of inverse problems in a wavelet basis

11.1.1 Specification of wavelet-domain MAP estimator

To obtain a practical reconstruction model, we adopt the same strategy as in Section 10.1.2: We truncate the signal over a spatial region $\Omega$ and introduce problem-specific boundary conditions that are enforced by suitable modifications of the basis functions. This yields the finite-dimensional signal model

$$s_1(r) = \sum_{i=1}^{I_{\text{max}}} \sum_{k \in \Omega_i} v_i[k] \psi_{i,k}(r) = \sum_{k \in \Omega} s[k] \hat{\beta}_{1,k}(r)$$  \hspace{1cm} (11.3)$$

where $\Omega_i$ denotes the wavelet-domain index set corresponding to the ROI $\Omega$. Note that the above expansion spans the same signal space as (10.8), provided that we select $\hat{\phi} = \hat{\phi}_L$ as the scaling function of the wavelet system $\{\psi_{i,k}\}$.

The signal in (11.3) is uniquely specified by an $N$-dimensional vector $v$ of pooled wavelet coefficients $v_i[k], k \in \Omega_i, i = 1, \ldots, I_{\text{max}}$. The right-hand side of (11.3) also indicates that there is a linear, one-to-one correspondence between the sequence of wavelet coefficients $v_i[.]$ and the discrete signal $s[.]$. This mapping specifies the discrete wavelet transform which admits a fast filterbank implementation. In vector notation, this translates into

$$v = \tilde{W}s \iff s = Wv$$

with $W = \tilde{W}^{-1}$ where the entries of the $(N \times N)$ wavelet matrices $\tilde{W}$ and $W$ are given by

$$[\tilde{W}]_{i,k},k' = \langle \tilde{\psi}_{i,k}, \hat{\phi}_{i,k'} \rangle$$

$$[W]_{k',i,k} = \langle \hat{\phi}_{i,k'}, \psi_{i,k} \rangle,$$

respectively. Also note that the wavelet basis is orthonormal if and only if $\tilde{\psi}_{i,k} = \psi_{i,k}$ which translates into $\tilde{W} = W^T$ being an orthonormal matrix; this latter property presupposes that the underlying scaling functions are orthogonal, too.

With the above convention, we write the wavelet version of the measurement equation (10.9) as

$$y = H_{\text{wav}} v + n$$

with wavelet-domain system matrix $H_{\text{wav}}$ whose entries are given by

$$[H_{\text{wav}}]_{m,(i,k)} = \langle \eta_m, \psi_{i,k} \rangle$$  \hspace{1cm} (11.4)$$

where $\eta_m$ is the analysis function corresponding to the $m$th measurement. The link with (10.10) in Section 10.1.2 is $H_{\text{wav}} = HW$ with the proper choice of analysis function $\hat{\beta} = \hat{\beta}_L$.

For the purpose of simplification and mathematical tractability, we make now the same kind of decoupling simplification as in Section 10.1.2, treating the wavelet components as if they were independent \footnote{While this approximation is legitimate within a given scale for sufficiently well localized wavelets, it is less so between scales because the wavelet smoothing kernels $\hat{\phi}_i$ and $\hat{\phi}_j$ typically overlap. (A more-refined probabilistic model should take those inter-scale dependencies into consideration.)}. Using Bayes’ rule, we get the corresponding
expression of the posterior probability distribution as

\[ p_{V|Y}(v|y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\|y - H_{\text{wav}}v\|^2}{2\sigma^2}\right) p_Y(v) \approx \exp\left(-\frac{\|y - H_{\text{wav}}v\|^2}{2\sigma^2}\right) \prod_{i = 1}^{\text{max}} p_{V_i}(v_i|k), \]

where \( p_{V_i} \) is the (conjugate) inverse Fourier transform of \( \hat{p}_{V_i}(\omega) = e^{f_{\tilde{\phi}}(\omega)} \). By maximizing \( p_{V|Y} \), we derive the wavelet-domain version of the MAP estimator

\[ v_{\text{MAP}}(y) = \arg\min_v \left\{ \frac{1}{2}\|y - H_{\text{wav}}v\|^2 + \sigma^2 \sum_{k \in \Omega_i} \Phi_{V_i}(v_i|k) \right\}, \quad (11.5) \]

which is similar to (10.12), except that it now involves the series wavelet potentials

\[ \Phi_{V_i}(x) = -\log p_{V_i}(x). \]

The specificity of the present MAP formulation is that the potential functions \( \Phi_{V_i} \) are scale-dependent and tied to the Lévy exponent \( f \) of the continuous-domain innovation \( w \). Since the pdfs \( p_{V_i} \) of the wavelet coefficients are infinitely divisible with Lévy exponent \( f_{\tilde{\phi}} \), we can determine the exact form of the potentials as

\[ \Phi_{V_i}(x) = -\log \int_{\mathbb{R}} \exp\left(f_{\tilde{\phi}}(\omega) - j\omega x\right) \frac{d\omega}{2\pi} \quad (11.6) \]

with

\[ f_{\tilde{\phi}}(\omega) = \int_{\mathbb{R}^d} f(\omega \tilde{\phi}_i(r)) \, dr \]

where \( \tilde{\phi}_i \) is the wavelet smoothing kernel at resolution \( i \) in (11.2). Moreover, we can rely on the theoretical analysis of id potentials in Section 10.2.1, which remains valid in the wavelet domain, to extract the global characteristics of \( \Phi_{V_i} \). The general trend that emerges is that these characteristics are mostly insensitive to the exact shape of \( \tilde{\phi}_i \), and hence to the choice of a particular wavelet basis.

11.1.2 Evolution of the potential function across scales

The material in this section is currently under review. It is available from the authors upon request.
11.2 Wavelet-based methods for solving linear inverse problems

Having specified the statistical reconstruction problem in a wavelet basis, we now describe numerical methods of solution. To that end, we consider the general optimization problem

$$\min_s \left\{ \frac{1}{2} \|y - Hs\|_2^2 + \tau \Phi(W^T s) \right\}, \quad (11.10)$$

where $\Phi(v) = \sum_{n=1}^{N} \Phi_n(v_n)$ is a separable potential function and $W^T = W^{-1}$ an orthonormal transform matrix. The qualitative effect of the second term in (11.10) is to favor solutions that admit a sparse wavelet expansion; the strength of this “regularization” constraint is controlled by the parameter $\tau \in \mathbb{R}^+$. Clearly, the solution of (11.10) is equivalent to the MAP estimator (11.5) if we set $\tau = \sigma^2$ and $\Phi(v) = \sum_{i} \sum_{k \in \Omega_i} \Phi_{V_i} \left( v_i[k] \right)$.

While a possible approach for solving (11.10) is to apply the ADMM algorithm of Section 10.2.4 with the substitution of $L$ by $W^T$ and a slight adjustment for scale-dependent potentials, we shall present two alternative techniques (ISTA and FISTA) that capitalize on the orthogonality of the matrix $W$. The second algorithm (FISTA) is a modification of the first one that results in faster convergence.

11.2.1 Preliminaries

To exploit the separability of the potential function $\Phi$, we restate the reconstruction problem in terms of the wavelet coefficients $v = (v_1, \ldots, v_N) = W^T s$ as the min-
imization of the cost functional
\[ C(v) = \frac{1}{2} \| y - H_{\text{wav}} v \|_2^2 + \tau \sum_{n=1}^{N} \Phi_n(v_n), \]  
where \( H_{\text{wav}} = HW \). In order to gain insights on the algorithmic components of ISTA, we first investigate two extreme cases for which the solution can be written down explicitly.

**Least-squares estimation**

For \( \tau = 0 \), the minimization of (11.11) reduces to a classical least-squares estimation problem and there is no advantage in expressing the signal in terms of wavelets. The solution of the reconstruction problem is given by
\[ s_{\text{LS}} = (H^T H)^{-1} H^T y \]
under the assumption that \( H^T H \) is invertible. When the underlying matrix is too large to be inverted numerically, the corresponding linear system of equations is solved iteratively. The simplest iterative reconstruction method is the Landweber algorithm
\[ s^{k+1} = s^k + \mu H^T (y - Hs^k) \]  
with \( \mu \in \mathbb{R}^+ \), which progressively builds up the solution by applying a steepest-descent update. It is a first-order optimizer whose efficiency depends on the step size \( \mu \) and the conditioning of \( H \). A classical result is that this iterative scheme will converge to the solution provided that \( 0 < \mu < 2/L \) where \( L = \lambda_{\text{max}}(H^T H) \) is the spectral radius of the iteration matrix \( A = H^T H \).

**Simple denoising problem**

When both \( s \) and \( y \) are expressed in the wavelet basis and \( H = I \), (11.10) reduces to a separable denoising problem. Specifically, by defining \( z = W^T y \), we get
\[ \tilde{v} = \arg\min_v \left\{ \frac{1}{2} \| y - Wv \|_2^2 + \tau \Phi(v) \right\} \]
\[ = \arg\min_v \left\{ \frac{1}{2} \| z - v \|_2^2 + \tau \sum_{n=1}^{N} \Phi_n(v_n) \right\}, \]  
(by Parseval)
so that
\[ \tilde{v} = \text{prox}_{\phi_n} \left( z; \tau \right) = \left( \begin{array}{c} \text{prox}_{\phi_1}(z_1; \tau) \\ \vdots \\ \text{prox}_{\phi_N}(z_N; \tau) \end{array} \right) \]  
(11.14)
where the definition of the underlying proximal operators (vectorial and scalar) is consistent with the formulation of Section 10.2.3. Hence, the solution \( \tilde{v} \) can be computed by applying a series of component-wise shrinkage-thresholding functions to the wavelet coefficients of \( y \). This is the model-based version of the standard denoising algorithm mentioned in the introduction. The relation between \( \text{prox}_{\phi_n} \) and the
underlying probability model is investigated in more details in Section 11.3. The bottom line is that these are scale-dependent nonlinear maps (see examples in Figure 11.5) that can be precomputed and stored in a lookup table which makes the denoising procedure very efficient.

### 11.2.2 Iterative soft-thresholding algorithm (ISTA)

The idea behind ISTA is to solve (11.10) iteratively by alternatively switching between a simple Landweber update and a denoising step.

ISTA produces a sequence \( v^k \) that converges to the minimizer \( v^\ast \) of (11.11) when \( \Phi \) is convex. At each step, it minimizes a simpler auxiliary cost \( \mathcal{C}'(v, v^k) \) that depends on the current estimate \( v^k \). The design constraint is that \( \mathcal{C}'(v, v^k) \leq \mathcal{C}'(v^k) \) with equality when \( v = v^k \). This guarantees that the cost functional decreases monotonically with the iteration number \( k \). The standard choice is

\[
\begin{align*}
\mathcal{C}'(v, v^k) &= \mathcal{C}(v) + \frac{1}{2L} \|v - v^k\|^2 + \frac{1}{2L} \|H\text{wav}(v - v^k)\|^2 \quad (11.15)
\end{align*}
\]

with \( L \) such that

\[
L\|e\|^2 \geq \|H\text{wav}e\|^2,
\]

for all \( e \in \mathbb{R}^N \). The critical value of \( L \) is \( \lambda_{\text{max}}(H\text{wav}^TH\text{wav}) \), which is the same \( L \) as in the Landweber algorithm of Section 11.2.1 since \( W \) is unitary. The derivation of ISTA is based on the rewriting of (11.15) as

\[
\begin{align*}
\mathcal{C}'(v, v^k) &= \frac{1}{2L} \|v - z^k\|^2 + \frac{1}{2L} \|H\text{wav}(v - v^k)\|^2 + C_0(v^k, y) \quad (11.16)
\end{align*}
\]

where \( C_0(v^k, y) \) is a term that does not depend on \( v \) and where the auxiliary variable \( z^k \) is given by

\[
\begin{align*}
z^k &= v^k + \frac{1}{2L} H\text{wav}^T(y - H\text{wav}v^k) \\
&= W^T(s^k + \frac{1}{2L}H^T(y - Hs^k)). \quad (11.17)
\end{align*}
\]

The crucial point is that the minimization of (11.16) with respect to \( v \) is equivalent to the denoising problem (11.13). This implies that

\[
\begin{align*}
\arg\min_{v \in \mathbb{R}^N} \mathcal{C}'(v, v^k) &= \text{prox}_\Phi \left( z^k : \frac{\tau}{L} \right),
\end{align*}
\]

which corresponds to a shrinkage-thresholding of the wavelet coefficients of the signal. The form of the update equation (11.17) is also highly suggestive for it boils down to a Landweber iteration (see (11.12)) followed by a wavelet transform. The resulting ISTA is summarized in Algorithm 1.

The remarkable aspect is that this simple sequence of Landweber updates and wavelet-domain thresholding operations converges to the solution of (11.10). The only subtle point is that the strength of the thresholding \( (\tau / L) \) is tied to the step size of the gradient update.
Algorithm 1: ISTA solves $s^* = \arg\min_s \left\{ \frac{1}{2} \| y - Hs \|_2^2 + r \Phi(W^T s) \right\}$

| input: $A = H^T H$, $a = H^T y$, $s^0$, $r$, and $L$ |
| set: $k \leftarrow 0$ |
| repeat |
| $s^{k+1} = s^k + \frac{1}{L} (a - A s^k)$ (Landweber step) |
| $v^{k+1} = \text{prox}_\Phi \left( W^T s^{k+1} / \frac{1}{L} \right)$ (wavelet-domain denoising) |
| $s^{k+1} = W v^{k+1}$ (inverse wavelet transform) |
| $k \leftarrow k + 1$ |
| until stopping criterion |
| return $s^k$ |

11.2.3 Fast iterative soft-thresholding algorithm (FISTA)

While ISTA converges to a (possibly local) minimum, it may do so rather slowly since the amount of error reduction at each step is dictated by the Landweber update. The latter, which is a basic first-order technique, is known to be quite inefficient when the system matrix is poorly conditioned.

When $\Phi$ is convex, it is possible to characterize the convergence behavior of ISTA. Specifically, Beck and Teboulle [BT09b, Theorem 3.1] have shown that, for any $k > 1$,

$$
\mathcal{E}(v^k_{\text{ISTA}}) - \mathcal{E}(v^*) \leq \frac{L}{2k} \| v^k_{\text{ISTA}} - v^* \|_2^2
$$

which indicates that the cost function decreases linearly with the iteration number $k$.

In the same paper, these authors have proposed a refinement of the scheme called “fast iterative shrinkage-thresholding algorithm” (FISTA), which improves the rate of convergence by one order. This is achieved via a controlled over-relaxation that utilizes the previous iterates to produce a better guess for the next update. A possible implementation of FISTA is shown in Algorithm 2.

The only difference with ISTA is the update of $v^{k+1}$ which is an extrapolation of the two previous ISTA computations $w^{k+1}$ and $w^k$. The variable $\tau_k$ controls the strength of the over-relaxation which increases with $k$ up to some asymptotic limit.

The theoretical justification for FISTA (see [BT09b, Theorem 4.4]) is that the scheme improves the convergence such that, for any $k > 1$,

$$
\mathcal{E}(v^k_{\text{FISTA}}) - \mathcal{E}(v^*) \leq \frac{2L}{(k+1)^2} \| v^k_{\text{FISTA}} - v^* \|_2^2.
$$

Practically, switching from a linear to a quadratic convergence rate can translate in a spectacular speed improvement over ISTA with the advantage that this change of regime essentially comes for free. FISTA therefore constitutes the method of choice for wavelet-based regularization; it typically delivers state-of-the-art performance for the kind of large-scale optimization problems encountered in imaging.
Algorithm 2: FISTA solves $s^* = \operatorname{argmin}_s \{ \frac{1}{2} \|y - Hs\|_2^2 + \tau \Phi(W^T s) \}$

\begin{algorithm}
\begin{algorithmic}
\STATE \textbf{input:} $A = H^T H$, $a = H^T y$, $s^0$, $\tau$ and $L$
\STATE \textbf{set:} $k \leftarrow 0$, $w_0 \leftarrow Ws^0$, $f_0 \leftarrow 0$
\REPEAT
\STATE $w_{k+1} \leftarrow \operatorname{prox}_\mu \left( W^T (s_k^0 + \frac{1}{L} (a - As^k)) ; \frac{\tau}{L} \right)$ \hfill (ISTA step)
\STATE $t_{k+1} \leftarrow \frac{1}{2} \left( 1 + \sqrt{1 + 4 f_k^2} \right)$
\STATE $v_{k+1} \leftarrow w_{k+1} + \frac{t_k - 1}{t_{k+1}} (w_{k+1} - w_k)$
\STATE $s_{k+1} \leftarrow Wv_{k+1}$
\STATE $k \leftarrow k + 1$
\UNTIL {stopping criterion}
\RETURN $s^k$
\end{algorithmic}
\end{algorithm}

Figure 11.1 Comparison of the convergence properties of ISTA (red) and FISTA (yellow) for the image in 11.2(b) as a function of the iteration index.

11.2.4 Discussion of wavelet-based image reconstruction

Iterative shrinkage-thresholding algorithms can be applied to the reconstruction of images for a whole variety of biomedical imaging modalities in the same way as we saw in Section 10.3. For illustration purposes, we have applied ISTA and FISTA to the deconvolution of the fluorescence micrographs of Section 10.3.2. In order to mimic the regularizing effect of the gradient operator, we have selected 2D Haar wavelets which qualitatively act as (smoothed) first-order derivatives. We have also used the same type of potential functions: $\Phi_{\text{Gauss}}(x) = A_i |x|^2$, $\Phi_{\text{Laplace}}(x) = B_i |x|$, and $\Phi_{\text{Student}}(x) = C_i \log(x^2 + \epsilon)$ where $A_i$, $B_i$, and $C_i$ are some proper scale-dependent constants. As in the previous experiments, the overall regularization strength $\tau$ was tuned for best performance (maximum SNR with respect to the reference). Here, we are presenting the results for the image of nerve cells (see Figure 10.3b) with the use of $\ell_1$ wavelet-domain regularization.

The plot in Figure 11.1 documents the evolution of the cost functional (11.11) as a
Wavelet-domain methods

Figure 11.2 Results of deconvolution experiment: (a) Blurry and noisy input of the deconvolution algorithm (BSNR=20dB). (b) Ground truth image (nerve cells). (c) Result of MAP deconvolution with TV regularization (SNR=15.23 dB) (d) Result of wavelet-based deconvolution (SNR=12.73dB). (e) Result of wavelet-based deconvolution with cycle spinning (SNR=15.18dB). (f) Zoomed comparison of results for the region marked in (b).

function of the iteration index for both ISTA and FISTA. It illustrates the faster convergence rate of FISTA, in agreement with Beck and Teboulle’s prediction. As far as quality is concerned, a general observation is that the output of the basic version of the wavelet-based reconstruction algorithm is not on par with the results of Section 10.3. The main problem (see Figure 11.2f) is that the reconstructed images suffer from artifacts (in the form of wavelet footprints) that are typically the consequence of the lack of shift invariance of the wavelet representation. Fortunately, there is a simple remedy to correct for this effect via a mechanism called cycle spinning\(^2\). The approach is to randomly shift the signal back and forth during the course of iterations, which is equivalent to cycling through a family of shifted wavelet transforms, as will be described in Section 11.4.2. Incorporating cycle spinning in ISTA does not increase the computational cost but improves the SNR of the reconstruction significantly, as shown in Figure 11.2e. Hence, we end up with a result that is comparable in quality to the output of the MAP reconstruction algorithm of Section 10.2 (see Figure 11.2c). This trend subsists with other images and across imaging modalities.

\(^2\) Cycle spinning is used almost systematically for the wavelet-based reconstructions showcased in literature. However, the method is rarely accounted for in the accompanying theory.
Combining cycle spinning with FISTA is feasible as well, with the advantage that the convergence rate of the latter is typically superior to that of the ADMM technique.

While averaging across shifts appears to be essential for making wavelets competitive, we are left with the conceptual problem that the cycle-spun version of ISTA does not rigorously fit our statistical formulation. It converges to a solution that is not a strict minimizer of (11.10) but, rather, to some kind of average over a family of “shifted” wavelet transforms. While this description is largely empirical, there is a theoretical explanation of the phenomenon for the simpler signal-denoising problem. Specifically, in Section 11.4, we shall demonstrate that cycle spinning necessarily improves denoising performance (see Proposition 11.3) and that it can be seen as an alternative means of computing the “exact” MAP estimators of Section 10.4.3. In other words, cycle spinning somehow compensates for the inter-scale dependencies of wavelet coefficients that were neglected when writing (11.5).

The most favorable aspect of wavelet-domain processing is that it offers direct control over the reconstruction error, thanks to Parseval’s relation. In particular, it allows for a more refined design of thresholding functions based on the minimum mean-square-error (MMSE) principle. This is the reason why we shall now investigate non-iterative strategies for improving simple wavelet-domain denoising.

11.3 Study of wavelet-domain shrinkage estimators

In the remainder of the chapter, we concentrate on the problem of signal denoising with $H = I$ (identity) or, equivalently, $H_{\text{wav}} = W$, under the assumption that the transform matrix $W$ is orthonormal. The latter ensures that any reduction of the quadratic error achieved in the wavelet domain is automatically transferred to the signal domain.

In this particular setting, we can address the important issue of the dependency between the wavelet-domain thresholding functions and the prior probability model. Our practical motivation is to improve the standard algorithm by identifying the solution that minimizes the mean-square estimation error. To specify the underlying scalar estimation problem, we transpose the measurement equation $y = s + n$ into the wavelet domain as

$$z = W^T s + W^T n = v + n'$$

where $v_i$ and $n_i$ are the wavelet coefficients of the noise-free signal $s$ and of the AWGN $n$, respectively. Since the wavelet transform is orthonormal, the transformed noise $n' = W^T n$ remains white, so that $n_i$ is Gaussian i.i.d. with variance $\sigma^2$. Now, when the wavelet coefficients $v_i$ are statistically independent as has been assumed so far, the denoising can be performed in a separable fashion by considering the wavelet coefficients individually. The estimation problem is then to recover $v$ from the noisy coefficient $z = v + n$ where we have dropped the wavelet indices to simplify the notation. Irrespective of the statistical criterion used (MAP vs. MMSE), the estimator...
\( \hat{v}(z) \) will be a function of the (scalar) noisy input \( z \), in agreement with the standard wavelet-denoising procedure.

Next, we develop the theory associated with the statistical wavelet-based estimators. The prior information is provided by the wavelet-domain pdfs \( p_{V_i} \) which are known to be infinitely divisible (see Proposition 8.6). We then make use of those results to characterize and compare the shrinkage-thresholding functions associated with the id distributions of Table 4.1.

### 11.3.1 Pointwise MAP estimators for AWGN

Our baseline is the MAP solution to the denoising problem given by (11.14). For later reference, we give the scalar formulation of this estimator

\[
\hat{v}_{\text{MAP}}(z) = \arg\min_{\hat{v} \in \mathbb{R}} \left\{ \frac{1}{2} |z - v|^2 + \sigma^2 \Phi_{V_i}(v) \right\} = \text{prox}_{\Phi_{V_i}}(z; \sigma^2),
\]

which involves a scale-specific proximity operator of the type investigated in Section 10.2.3. Explicit formulas and graphs of \( \hat{v}_{\text{MAP}}(z) \) for the primary types of probability models/sparsity patterns are presented in Section 11.3.3.

### 11.3.2 Pointwise MMSE estimators for AWGN

From a mean-square-error point of view, performing denoising in the wavelet domain is equivalent to signal-domain processing since the \( \ell_2 \)-error is preserved. This makes the use of MMSE shrinkage functions highly relevant, even when the wavelet coefficients are only approximately independent. The MMSE estimator of \( v_i \) given the noisy coefficient \( z \) is provided by the posterior mean

\[
\hat{v}_{\text{MMSE}}(z) = E[V | Z = z] = \int_{\mathbb{R}} v \cdot p_{V_i|Z}(v|z) \, dv
\]

where \( p_{V_i|Z}(v|z) = \frac{p_{Z|V}(z|v) \cdot p_{V_i}(v)}{p_{Z}(z)} \) by Bayes’ rule. In the present context of AWGN, we have that \( p_{Z|V}(z|v) = g_\sigma(z - v) \) and \( p_{Z} = g_\sigma \ast p_{V_i} \) where \( g_\sigma \) is a centered Gaussian distribution with standard deviation \( \sigma \). Moreover, we can bypass the integration step in (11.20) by taking advantage of the Miyasawa/Stein formula for the posterior mean of a random variable corrupted by Gaussian noise [Miy61, Ste81], which states that

\[
\hat{v}_{\text{MMSE}}(z) = z - \sigma^2 \Phi'_\sigma(z)
\]
where $\Phi'_Z(z) = -\frac{d}{dz} \log p_Z(z) = -\frac{p'_Z(z)}{p_Z(z)}$. This classical formula, which capitalizes on special properties of the Gaussian distribution, is established as follows:

$$\sigma^2 p'_Z(z) = \sigma^2 (g'_\sigma * p_V)(z)$$
$$= \int \sigma^2 (z-v) g_\sigma(z-v)p_V(v) \, dv$$
$$= -z \int \sigma^2 g_\sigma(z-v)p_V(v) \, dv + p_Z(z) \left( \int z g_\sigma(z-v)p_V(v) \, dv \right)$$
$$= -z p_Z(z) + p_Z(z) v_{\text{MMSE}}(z).$$

This means that we can derive the explicit form of $v_{\text{MMSE}}(z)$ for any given $p_V$ via the evaluation of the Gaussian convolution integrals

$$p_Z(z) = (g_\sigma * p_V)(z) = \mathcal{F}^{-1} \left\{ e^{-\frac{z^2}{2}} \hat{p}_V(\omega) \right\}(z) \quad (11.22)$$

$$p'_Z(z) = (g'_\sigma * p_V)(z) = \mathcal{F}^{-1} \left\{ j\omega e^{-\frac{\omega^2 z^2}{2}} \hat{p}_V(\omega) \right\}(z). \quad (11.23)$$

These can be calculated either in the time or frequency domain. The frequency-domain formulation offers more convenience for the majority of iid distributions and is also directly amenable to numerical computation with the help of the FFT. Likewise, we use formula (11.21) to infer the general asymptotic behavior of this estimator.

**Theorem 11.1** Let $z = v + n$ where $v$ is infinitely divisible with symmetric pdf $p_V$ and $n$ is Gaussian-distributed with variance $\sigma^2$. Then, the MMSE estimator of $v$ given $z$ has the linear behavior around the origin given by

$$v_{\text{MMSE}}(z) = z \left( 1 - \sigma^2 \Phi''_Z(0) \right) + O(z^3) \quad (11.24)$$

where

$$\Phi''_Z(0) = \frac{\int_{\mathbb{R}} \omega^2 e^{\frac{-\omega^2 z^2}{2}} \hat{p}_V(\omega) \, d\omega}{\int_{\mathbb{R}} e^{\frac{-\omega^2 z^2}{2}} \hat{p}_V(\omega) \, d\omega} > 0. \quad (11.25)$$

If, in addition, $p_V$ is unimodal and does not decay faster than an exponential, then

$$v_{\text{MMSE}}(z) \sim v_{\text{MAP}}(z) \sim z - \sigma^2 b'_1$$

as $z \to \infty$, where $b'_1 = \lim_{x \to -\infty} \Phi'_Z(x) = \lim_{x \to -\infty} \Phi'_V(x) \geq 0$.

**Proof** The material in this section is currently under review. It is available from the authors upon request.
Several remarks are in order.
First, the linear approximation (11.24) is exact in the Gaussian case. It actually yields the classical linear (LMMSE) estimator
\[ v_{LMMSE}(z) = \sigma_i^2 / (\sigma_i^2 + \sigma^2 z) \]
where \( \sigma_i^2 \) is the variance of the signal contribution in the \( i \)th wavelet channel. Indeed, when \( p_{V_i} \) is a Gaussian distribution, we have that \( \Phi_{Z}(z) = \sigma_i^2 z^2 / (2(\sigma_i^2 + \sigma^2)) \) which, upon substitution in (11.21), yields the \( v_{LMMSE} \) estimator.
Second, by applying Parseval’s relation, we can express the slope of the MMSE estimator at the origin as the ratio of time-domain integrals
\[ 1 - \sigma^2 \Phi_{Z}''(0) = 1 - \sigma^2 \frac{\int_{R} \sigma_i^2 - \frac{x^2}{\sigma_i^2} e^{-\frac{x^2}{2\sigma^2}} p_{V_i}(x) \, dx}{\int_{R} e^{-\frac{x^2}{2\sigma^2}} p_{V_i}(x) \, dx} \]
\[ = \frac{\int_{R} x^2 e^{-\frac{x^2}{2\sigma^2}} p_{V_i}(x) \, dx}{\sigma^2 \int_{R} e^{-\frac{x^2}{2\sigma^2}} p_{V_i}(x) \, dx} \]
which may be simpler to evaluate for some id distributions.

11.3.3 Comparison of shrinkage functions: MAP vs. MMSE
In order to gain practical insights and to make the connection with existing methods, we now investigate solutions that are tied to specific id distributions. We consider the prior models listed in Table 4.1, which cover a broad range of sparsity behaviors. The common feature is that these pdfs are symmetric and unimodal with tails fatter than a Gaussian. Their practical relevance is that they may be used to fit the wavelet-domain statistics of real-world signals or to derive corresponding families of parametric algorithms. Unless stated otherwise, the graphs that follow display series of comparable estimators with a normalized signal input (SNR_0 = 1).

Laplace distribution
The Laplace distribution with parameter \( \lambda \) is defined as
\[ p_{\text{Laplace}}(x; \lambda) = \frac{1}{2} \lambda e^{-|x|} \]
Its variance is given by \( \sigma_0^2 = \frac{\lambda^2}{2} \). The Lévy exponent is \( f_{\text{Laplace}}(\omega; \lambda) = \log p_{\text{Laplace}}(\omega; \lambda) = \log(\frac{\lambda^2}{|\omega|^2 \sigma_0^2}) \), which is \( p \)-admissible with \( p = 2 \). The Laplacian potential is
\[ \Phi_{\text{Laplace}}(x; \lambda) = \lambda |x| - \log(\lambda/2) \]
Since the second term of \( \Phi_{\text{Laplace}} \) does not depend on \( x \), this translates into a MAP estimator that minimizes the \( \ell_1 \)-norm in the corresponding wavelet channel. It is
11.3 Study of wavelet-domain shrinkage estimators

Figure 11.3 Comparison of potential functions $\Phi(z)$ and pointwise estimators $v(z)$ for signals with matched Laplace and sech distributions corrupted by AWGN with $\sigma = 1$: (a) Laplace (dashed) and sech (solid line) potentials. (b) Laplace MAP estimator (red), MMSE (yellow) estimator and its first-order equivalent (dot-dashed line) for $\lambda = 2$. (c) Sech MAP (red) and MMSE (yellow) estimators for $\sigma_0 = \pi/4$.

well-known that the solution of this optimization problem yields the soft-threshold estimator (see [Tib96, CDLL98, ML99])

$$v_{\text{MAP}}(z; \lambda) = \begin{cases} 
    z - \lambda, & z > \lambda \\
    0, & z \in [-\lambda, \lambda] \\
    z + \lambda, & z < \lambda. 
\end{cases}$$

By applying the time-domain versions of (11.22) and (11.23), one can also derive the analytical form of the corresponding MMSE estimator in AWGN. For reference purposes, we give its normalized version with $\sigma^2 = 1$ as

$$v_{\text{MMSE}}(z; \lambda) = z - \frac{\lambda}{\sqrt{2}} \left( \text{erf} \left( \frac{z - \lambda}{\sqrt{2}} \right) - e^{2\lambda^2} \text{erfc} \left( \frac{\lambda + z}{\sqrt{2}} \right) + 1 \right) \frac{\text{erf} \left( \frac{z + \lambda}{\sqrt{2}} \right) + e^{2\lambda^2} \text{erfc} \left( \frac{\lambda + z}{\sqrt{2}} \right) + 1}{\text{erf} \left( \frac{z - \lambda}{\sqrt{2}} \right) + e^{2\lambda^2} \text{erfc} \left( \frac{\lambda + z}{\sqrt{2}} \right) + 1}$$

where $\text{erfc}(t) = 1 - \text{erf}(t)$ denotes the complementary (Gaussian) error function, which is a result that can be traced back to [HY00, Proposition 1]. A comparison of the estimators for the Laplace distribution with $\lambda = 2$ and unit noise variance is given in Figure 11.3b. While the graph of the MMSE estimator has a smoother appearance than that of the soft-thresholding function, it does also exhibit two distinct regimes that are well represented by first-order polynomials: behavior around the origin vs. behavior at $\pm \infty$. However, the transition between the two regimes is much more
progressive in the MMSE case. Asymptotically, the MAP and MMSE estimators are equivalent, as predicted by Theorem 11.1. The key difference occurs around the origin where the MMSE estimator is linear (in accordance with Theorem 11.1) and quite distinct from a thresholding function. This means that the MMSE estimator will never annihilate a wavelet coefficient, which somewhat contradicts the predominant paradigm for recovering sparse signal.

**Hyperbolic-secant distribution**

The hyperbolic secant (reciprocal of the hyperbolic cosine) is a classical example of id distribution [Fel71]. It seems to us as interesting a candidate for regularization as the Laplace distribution. Its generic version with standard deviation \( \sigma_0 \) is given by

\[
p_{\text{sech}}(x; \sigma_0) = \frac{\text{sech} \left( \frac{\pi x}{2\sigma_0} \right)}{2\sigma_0} = \frac{1}{\sigma_0 \left( e^{-\frac{x}{\sigma_0}} + e^{\frac{x}{\sigma_0}} \right)}.
\]

Remarkably, its characteristic function is part of the same class of distributions with

\[
\hat{p}_{\text{sech}}(\omega; \sigma_0) = \text{sech} \left( \frac{\pi \omega}{\sigma_0} \right).
\]

The hyperbolic-secant potential function is

\[
\Phi_{\text{sech}}(x; \sigma_0) = -\log p_{\text{sech}}(x; \sigma_0) = \log \left( e^{-\frac{\pi x}{2\sigma_0}} + e^{\frac{\pi x}{2\sigma_0}} \right) + \log \sigma_0,
\]

which is convex and increasing for \( x \geq 0 \). Indeed, the second derivative of the potential function is

\[
\Phi''_{\text{sech}}(x) = \frac{\pi^2}{4\sigma_0^2} \text{sech}^2 \left( \frac{\pi x}{2\sigma_0} \right),
\]

which is positive. Note that, for large absolute values of \( x \), \( \Phi_{\text{sech}}(x; \sigma_0) \approx \frac{\pi^2}{2\sigma_0^2} |x| + \log \sigma_0 \), suggesting that it is essentially equivalent to the \( \ell_1 \)-type Laplace potential (see Figure 11.3a). However, unlike the latter, it is infinitely differentiable everywhere with a quadratic behavior around the origin.

The corresponding MAP and MMSE estimators with a parameter value that is matched to the Laplace example are shown in Figure 11.3c. An interesting observation is that the sech MAP thresholding functions are very similar to the MMSE Laplacian ones over the whole range of values. This would suggest using hyperbolic-secant-penalized least-squares regression as a practical substitute for the MMSE Laplace solution.

**Symmetric Student family**

We define the symmetric Student distribution with standard deviation \( \sigma_0 \) and algebraic decay parameter \( r > 1 \) as

\[
p_{\text{Student}}(x; r, \sigma_0) = A_{r,\sigma_0} \left( \frac{1}{C_{r,\sigma_0} + x^2} \right)^{\frac{r+1}{2}}
\]

with \( C_{r,\sigma_0} = \sigma_0^2 (2r - 2) > 0 \) and normalizing constant \( A_{r,\sigma_0} = \frac{(C_{r,\sigma_0})^\frac{r}{2}}{B(r, \frac{1}{2})} \) where \( B(r, \frac{1}{2}) \) is the beta function (see Appendix C.2). Despite the widespread use of this distribution...
11.3 Study of wavelet-domain shrinkage estimators

The interest for signal processing is that the Student model offers a fine control of the behavior of the tail, which conditions the level of sparsity of signals. The Student potential is logarithmic

\[
\Phi_{\text{Student}}(x; r, \sigma_0) = a_0 + \left( r + \frac{1}{2} \right) \log \left( C_r \sigma_0 + x^2 \right) \tag{11.28}
\]

with \( a_0 = \log A_{r,\sigma_0} \).

The Student MAP estimator is specified by a third-order polynomial equation that can be solved explicitly. This results in the thresholding functions shown in Figure 10.1b. We have also observed experimentally that the Student MAP and MMSE estimators are rather close to each other with linear trends around the origin that become indistinguishable as \( r \) increases; this can be verified by comparing Figure 10.1b and Figure 11.4a. This finding is also consistent with the distributions becoming more Gaussian-like for larger \( r \).

Note that Definition (11.27) remains valid in the super-sparse regimes with \( r \in (0,1] \), provided that the normalization constant \( C > 0 \) is no longer tied to \( r \) and \( \sigma_0 \). The catch is that the variance of the signal is unbounded for \( r \leq 1 \) which tends to flatten the shrinkage function around the origin, but maintains continuity since \( \Phi_{\text{Student}} \) is infinitely differentiable.

**Compound Poisson family**

We have already mentioned that the Poisson case results in pdfs that exhibit a Dirac distribution at the origin and are therefore unsuitable for MAP estimation. A compound Poisson variable is typically generated by integration of a random sequence of Dirac impulses with some amplitude distribution \( p_A \) and a density parameter \( \lambda \) corresponding to the average number of impulses within the integration window. The generic form of a compound Poisson pdf is given by (4.9). It can be
written as \( p_{\text{Poisson}}(x) = e^{-\lambda} \delta(x) + (1 - e^{-\lambda}) p_{A,\lambda}(x) \) where the pdf \( p_{A,\lambda} \) describes the distribution of the non-zero values.

The determination of the MMSE estimator from (11.21) requires the computation of \( \Phi'_Z(z) = -p'_Z(z)/p_Z(z) \). The most convenient approach is to evaluate the required factors using the right-hand-side expressions in (11.22) and (11.23) where \( \hat{p}_{V_i} \) is specified by its Poisson parameters as in Table 4.1. This leads to

\[
\hat{p}_{V_i}(\omega) = \exp\{\lambda_i(\hat{\rho}_{A_i}(\omega) - 1)\},
\]

where \( \lambda_i \in \mathbb{R}^+ \) and \( \hat{\rho}_{A_i} : \mathbb{R} \to \mathbb{C} \) are the Poisson rate and the characteristic function of the Poisson amplitude distribution at resolution \( i \), respectively. Moreover, due to the multiscale structure of the analysis, the wavelet-domain Poisson parameters are related to each other by

\[
\lambda_i = \lambda_0 2^{id}
\]

\[
\hat{\rho}_{A_i}(\omega) = \hat{\rho}_{A_0}(2^{i(\gamma-d/2)}\omega),
\]

which follows directly from (11.7). The first formula highlights the fact that the sparseness of the wavelet distributions, as measured by the proportion \( e^{-\lambda_i} \) of zero coefficients, decreases substantially as the scale gets coarser. Also note that the strength of this effect increases with the number of dimensions.

Some examples of MMSE thresholding functions corresponding to a sequence of compound Poisson signals with Gaussian amplitude distributions are shown in Figure 11.4b. Not too surprisingly, the smaller \( \lambda \) (red curve), the stronger the thresholding behavior at the origin. In that experiment, we have considered a wavelet-like progression of the rate parameter \( \lambda \), while keeping the signal-to-noise ratio constant to facilitate the comparison. For larger values of \( \lambda \) (yellow), the estimator converges to the LMSE solution (thin black line), which is consistent with the fact that the distribution becomes more and more Gaussian-like.

**Evolution of the estimators across wavelet scales**

The increase of \( \lambda_i \) in Figure 11.4b is consistent with the one predicted for a wavelet-domain analysis. Nonetheless, this graph does only account for part of the story because we enforced a constant signal-to-noise-ratio. In a realistic wavelet-domain scenario, another effect that predominates as the scale gets coarser must also be accounted for: the amplification of the quadratic signal-to-noise ratio that follows from (9.23) and that results in

\[
\text{SNR}_i = \frac{\text{Var}(Z_i)}{\sigma^2} = (2^{2i})^\gamma \text{SNR}_0,
\]

where \( \gamma \) is the scaling order of the stochastic process. Consequently, the potential function \( \Phi \) is dilated by \( b_i = (2^{(\gamma-d/2)} \cdot i) \). The net effect is to make the estimators more identity-like as \( i \) increases, both around the origin and at infinity because of the corresponding decrease of the magnitude of \( \Phi'(0) \) and \( \lim_{x \to \infty} \Phi'(x) \), respectively. This progressive convergence to the identity map is easiest to describe for a Gaussian signal where the sequence of estimators is linear—this is illustrated in Figure 11.5a for \( \gamma = 1 \) and \( d = 1 \) (Brownian motion) so that \( b_1 = 2^{1/2} \).
11.3 Study of wavelet-domain shrinkage estimators

Figure 11.5 Sequence of wavelet-domain estimators $v_i(z)$ for a Laplace-type Lévy process corrupted by AWGN with $\sigma = 1$ and wavelet resolutions $i = 0$ (red) to 4 (yellow). (a) LMMSE (or Brownian-motion MMSE) estimators. (b) Sym-gamma MAP estimators. (c) Sym-gamma MMSE estimators. The reference (fine-scale) parameters are $\sigma_0^2 = 1$ (SNR_0 = 1) and $r_0 = 1$ (Laplace distribution). The scale progression is dyadic.

In the non-Gaussian case, the sequence of wavelet-domain estimators will be part of some specific family that is completely determined by $p_{\psi_i}$, the wavelet pdf at scale 0, as described in Section 11.1.2. When the variance of the signal is finite, the implication of the underlying semigroup structure and iterated convolution relations is that the MAP and MMSE estimators both converge to the Gaussian solution (LMMSE estimator) as the scale gets coarser (yellow curves). Thus, the global picture remains very similar to the Gaussian one, as illustrated in Figure 11.5b-c. Clearly, the most significant nonlinearities can be found at the finer scale (red curves) where the sparsity effect is prominent.

The example shown (wavelet analysis of a Lévy process in a Haar basis) was setup such that the fine-level MAP estimator is a soft-threshold. As discussed next, the wavelet-domain estimators are all part of the sym-gamma family, which is the semigroup extension of the Laplace distribution. An interesting observation is that the thresholding behavior fades with a coarsening of the scale. Again, this points to the fact that the non-Gaussian effects (nonlinearities) are the most significant at the finer
levels of the wavelet analysis where the signal-to-noise ratio is also the least favorable.

Symmetric gamma and Meixner distributions
Several authors have proposed to represent wavelet statistics using Bessel K-forms [SLG02, FB05]. The Bessel K-forms are in fact equivalent to the symmetric gamma distributions specified by Table 4.1. Here, we would like to emphasize a further theoretical advantage: the semigroup structure of this family ensures compatibility across wavelet scales, provided that one properly links the distribution parameters.

The symmetric gamma distribution with bandwidth $\lambda$ and order parameter $r$ is best specified in terms of its characteristic function

$$\hat{p}_{\text{gamma}}(\omega; \lambda, r) = \left( \frac{\lambda^2}{\lambda^2 + \omega^2} \right)^r.$$  

The inverse Fourier transform of this expression yields a Bessel function of the second kind (a.k.a. Bessel K-form), as discussed in Appendix C. The variance of the distribution is $2r/\lambda^2$. We also note that $p_{\text{gamma}}(x; \lambda, 1)$ is equivalent to the Laplace pdf and that $p_{\text{gamma}}(x; \lambda, r)$ is the $r$-fold convolution of the former (semigroup property). In
the case of a dyadic wavelet analysis, we invoke (11.7) to show that the evolution of
the sym gamma parameters across scales is given by

$$\lambda_i = \frac{\lambda_0}{(2^{-d/2})^i}$$

$$r_i = r_0(2^d)^i$$

where \((\lambda_0, r_0)\) are the parameters at resolution \(i = 0\). These relations underly the
generation of the graphs in Figure 11.5b-c with \(y = 1, d = 1, \lambda_0 = \sqrt{2}, \text{and } r_0 = 1\).

Some further examples of sym-gamma MAP and MMSE estimators over a range of
orders are shown in Figure 11.6 under constant signal-to-noise ratio to highlight the
differences in sparsity behavior. We observe that the MAP estimators have a hard-
to-soft-threshold behavior for \(r < 3/2\), which is consistent with the discontinuity of
the potential at the origin. For larger values of \(r\), the trend becomes more linear. By
contrast, the MMSE estimator is much closer to the LMMSE (thin black line) around
the origin. For larger signal values, both estimators result into a more-or-less pro-
gressive transition between the two extreme lines of the cone (Identity and LMMSE)
that is controlled by \(r\)—the smaller values of \(r\) correspond to the sparser scenarios
with \(\nu_{\text{MMSE}}\) being closer to identity.

The Meixner family in Table 4.1 with order \(r > 0\) and scale parameter \(s_0 \in \mathbb{R}^+\)
provides the same type of extension for the hyperbolic secant distribution with es-
tentially the same functionality. Mathematically, it is closely linked to the gamma
function whose relevant properties are summarized in Appendix C. As shown in Table
10.1, the Meixner potential has the same asymptotic behavior as the sym-gamma po-
tential at infinity, with the advantage of it being much smoother (infinitely differenti-
able) at the origin. This implies that the curves of the gamma and Meixner estimators
are globally quite similar. The main difference is that the Meixner MAP estimator is
guaranteed to be linear around the origin, irrespective of the value of \(r\), and in better
agreement with the MMSE solution than its gamma counterpart.

**Cauchy distribution**

The prototypical example of a heavy-tail distribution is the symmetric Cauchy dis-
tribution with dispersion parameter \(s_0\), which is given by

$$p_{\text{Cauchy}}(x; s_0) = \frac{s_0}{\pi (s_0^2 + x^2)}.$$  \hspace{1cm} (11.29)

It is a special case of a \(\alpha\)S distribution (with \(\alpha = 1\)) as well as a symmetric Student
with \(r = \frac{1}{2}\).

Since the Cauchy distribution is stable, we can invoke Proposition 9.8, which en-
sures that the wavelet coefficients of a Cauchy process are Cauchy-distributed, too.
For illustration purposes, we consider the analysis of a stable Lévy process (a.k.a.
Lévy flight) in an orthonormal Haar wavelet basis with \(\psi = D^* \phi\) where \(\phi\) is a trian-
gular smoothing kernel. The corresponding wavelet-domain Cauchy parameters may
be determined from (9.27) with \(y = 1, d = 1, \text{and } \alpha = 1\), which yields \(s_i = s_0(2\sqrt{2})^i\).
While the variance of the Cauchy distribution is unbounded, an analytical characterization of the corresponding MAP estimator can be obtained by solving a cubic equation. The MMSE solution is then described by a cumbersome formula that involves exponentials and the error function erf. In particular, we can evaluate (11.24) to linearize its behavior around the origin as

\[ v_{\text{MMSE}}(z_i; s_i) = z_i \left( 1 - \sigma^2 \left( \frac{\sqrt{2\pi} e^{-\frac{s_i^2}{2}}}{\text{erfc} \left( \frac{s_i}{\sqrt{2}} \right)} - s_i^2 - 1 \right) \right) + O(z_i^3). \] (11.30)

The corresponding MAP and MMSE shrinkage functions with \( s_0 = \frac{1}{4} \) and resolution levels \( i = 0, \ldots, 4 \) are shown in Figure 11.7. The difference between both types of estimators is striking around the origin and is much more dramatic at finer scales \( (i = 0) \) (red) and \( (i = 1) \). As expected, all estimators converge to the identity map for large input values due to the slow (algebraic) decay of the Cauchy distribution. We observe that the effect of processing (deviation from identity) becomes less-and-less significant at coarser scales (yellow curves). This is consistent with the relative increase of the signal contribution while the power of the noise remains constant across wavelet channels.

11.3.4 Conclusion on simple wavelet-domain shrinkage estimators

The general conclusions that can be drawn from the present statistical analysis are as follows:

- The thresholding functions should be tuned to the wavelet-domain statistics, which necessarily involve \textit{infinitely divisible} distributions. In particular, this excludes the generalized Gaussian models with \( 1 < p < 2 \) which have been invoked in the past to justify \( \ell_p \)-minimization algorithms. The specification of wavelet-domain estimators can be carried out explicitly—at least, numerically—for the primary fam-
ilies of sparse processes characterized by their Lévy exponent \( f(\omega) \) and scaling order \( \gamma \).

- Pointwise MAP and MMSE estimators can differ quite substantially, especially for small input values. The MAP estimator sometimes acts as a soft-threshold, setting small wavelet coefficients to zero, while the MMSE solution is always linear around the origin where it essentially replicates the traditional Wiener solution. On the other hand, the two estimators are indistinguishable for large input values: they exhibit a shrinkage behavior with an offset that depends on the decay (exponential vs. algebraic) of the canonical id distribution.

- Wavelet-domain shrinkage functions must be adapted to the scale. In particular, the present statistical formulation does not support the use of a single, universal denoising function—such as the fixed soft-threshold dictated by a global \( \ell_1 \)-minimization argument—that could be applied to all coefficients in a nondiscriminative manner.

- Two effects come into play as the scale gets coarser. The first is a progressive increase of the signal-to-noise ratio which results in a shrinkage function that becomes more-and-more identity-like. This justifies the heuristic strategy to leave the coarser-scale coefficients untouched. The second is a Gaussianization of the wavelet-domain statistics (under the finite-variance hypothesis) due to the summation of a large number of random components (generalized version of the central limit theorem.) Concretely, this means that the estimator ought to progressively switch to a linear regime when the scale gets coarser, which is not what is currently done in practice.

- The present analysis did not take into account the statistical dependencies of wavelet coefficients across scales. While these dependencies affect neither the performance of pointwise estimators nor the present conclusions, their existence clearly suggests that the basic application of wavelet-domain shrinkage functions is suboptimal. A possible refinement is to specify higher-order estimators (e.g., bivariate shrinkage functions.) Such designs could benefit from a tree-like structure where each wavelet coefficient is statistically linked to its parents. The other alternative is the algorithmic solution described next, which constitutes a promising mechanism for turning a suboptimal solution into an optimal one.

### 11.4 Improved denoising by consistent cycle spinning

A powerful strategy for improving the performance of the basic wavelet-based denoisers described in Section 11.3 is through the use of an overcomplete representation. Here, we formalize the idea of cycle spinning by expanding the signal in a wavelet frame. In essence, this is equivalent to considering a series of “shifted” orthogonal wavelet transforms in parallel. The denoising task thereby reduces to finding a consensus solution. We show that this can be done either through simple averaging or by
constructing a solution that is globally consistent by way of an iterative refinement procedure.

To demonstrate the concept and the virtues of an optimized design, we concentrate on the model-based scenario of Section 10.4. The first important ingredient is the proper choice of basis functions which is discussed in Section 11.4.1. Then, in Section 11.4.2, we switch to a redundant representation (tight wavelet frame) with a demonstration of its benefits for noise reduction. In Section 11.4.3, we introduce the idea of consistent cycle spinning which results in an iterative variant of the basic denoising algorithm. The impact of each of these refinements, including the use of the MMSE shrinkage functions of Section 11.3, is evaluated experimentally in Section 11.4.4. The final outcome is an optimized wavelet-based algorithm that is able to replicate the MMSE results of Chapter 10.

11.4.1 First-order wavelets: Design and implementation

In line with the results of Sections 8.5 and 10.4, we focus on the first-order (or Markov) processes, which lend themselves to an analytical treatment. The underlying statistical model is characterized by the first-order whitening operator $L = D - \alpha_1 I$ with $\alpha_1 \in \mathbb{R}$ and the Lévy exponent $f$ of the innovation. We then apply the design procedure of Section 6.5 to determine the operator-like wavelet at resolution level $i = 1$, which is given by $\psi_{01}(t) = L^* \varphi_{01}(t - 1)$ where $L^* = -D - \alpha_1 I$. Here, $\varphi_{01}$ is the unique spline interpolant in the space of cardinal $L^*$-spline which is calculated as

$$
\varphi_{01}(t) = \frac{1}{(\beta_{i1}^* \ast \beta_{i1})(0)} (\beta_{i1}^* \ast \beta_{i1})(t)
$$

$$
= \begin{cases} 
\frac{e^{-\alpha_1 t - e^{2\alpha_1} + \alpha_1 |t|}}{1-e^{2\alpha_1}}, & \text{for } t \in [-1,1] \text{ and } \alpha_1 \neq 0 \\
1 - |t|, & \text{for } t \in [-1,1] \text{ and } \alpha_1 = 0 \\
0, & \text{otherwise.}
\end{cases}
$$

where $\beta_{i1}$ is the first-order exponential spline defined by (6.21).

Examples of the functions $\varphi_{01} \sim \beta_{01,0}$ (B-spline), $\varphi_{01} = \psi$ (wavelet smoothing kernel), and $\psi_{01}$ (operator-like wavelet) are shown in Figure 11.8. The B-spline $\beta_{01,1}$ in Figure 11.8(b) is an extrapolated version of $\beta_{01}$; it generates the coarser-resolution space $V_1 = \text{span}(\beta_{01,1}(-2k))_{k \in \mathbb{Z}}$ which is such that $V_0 = \text{span}(\beta_{01,1}(-k))_{k \in \mathbb{Z}} = V_1 + W_1$ with $W_1 = \text{span}(\psi_{01,0}(-2k))_{k \in \mathbb{Z}}$ and $W_1 \perp V_1$. A key property of the first-order model is that these basis functions are orthogonal and non-overlapping, as a result of the construction (B-spline of unit support).

The fundamental ingredient for the implementation of the wavelet transform is that the scaling function (exponential B-splines) and wavelets at resolution $i$ satisfy the two-scale relation

$$
\begin{bmatrix}
\beta_{i1,i}(t - 2^i k) \\
\psi_{i1,i}(t - 2^i k)
\end{bmatrix}
\propto
\begin{bmatrix}
1 & e^{2\alpha_1} \\
-e^{2\alpha_1} & 1
\end{bmatrix}
\begin{bmatrix}
\beta_{i1,i-1}(t - 2^i k) \\
\beta_{i1,i-1}(t - 2^i k - 2^{i-1})
\end{bmatrix}
$$

(11.31)

which involves two filters of length 2 (row vectors of the $(2 \times 2)$ transition matrix) since the underlying B-splines and wavelets are non-overlapping for distinct values
11.4 Improved denoising by consistent cycle spinning

1. Improved denoising by consistent cycle spinning

(a) B-splines at fine level

(b) B-splines at coarse level

(c) Operator-like wavelets

(d) Smoothing kernels

Figure 11.8 Operator-like wavelets and exponential B-splines for the first-order operator \( L = D - \alpha_1 \text{Id} \) with \( \alpha_1 = 0 \) (yellow) and \( \alpha_1 = -1 \) (red). (a) Fine-level exponential B-splines \( \beta_{a_1,0}(t) \). (b) Coarse-level exponential B-splines \( \beta_{a_1,1}(t) \). (c) Wavelet smoothing kernels \( \psi_{\text{int}}(t-1) \). (d) Operator-like wavelets \( \sqrt{\alpha_1} \sqrt[2]{\beta_{a_1,1}(t)} = L \sqrt{\psi_{\text{int}}(t-1)} \).

of \( k \). In particular, for \( i = 1 \) and \( k = 0 \), we get that \( \beta_{a_1,1}(t) \propto \beta_{a_1}(t) + a_1 \beta_{a_1}(t-1) \) and \( \psi_{a_1,1}(t) \propto -a_1 \beta_{a_1}(t) + \beta_{a_1}(t-1) \) with \( a_1 = e^{\alpha_1} \). These relations can be visualized in Figures 11.8(b) and 11.8(d), respectively. Also, for \( \alpha_1 = 0 \), we recover the Haar system for which the underlying filters \( h \) and \( g \) (sum and difference with \( e^{2\alpha_1} = 1 = a_1 \)) do not depend upon the scale \( i \) (see (6.6) and (6.7) in Section 6.1). Finally, we note that the proportionality factor in (11.31) is set by renormalizing the basis functions on both sides such that their norm is unity, which results in an orthonormal transform.

The corresponding fast wavelet-transform algorithm is derived by assuming that the fine-scale expansion of the input signal is \( s(t) = \sum_{k \in Z} s[k] \beta_{a_1,0}(t-k) \). To specify the first iteration of the algorithm, we observe that the support of \( \sqrt{\alpha_1} \sqrt[2]{\beta_{a_1,1}(t)} \) (resp., \( \beta_{a_1,1}(t-2k) \)) overlaps with the fine-scale B-splines at locations \( 2k \) and \( 2k+1 \) only. Due to the orthonormality of the underlying basis functions, this results into

\[
\begin{pmatrix}
   s[k] \\
   v_1[k]
\end{pmatrix}
= \begin{pmatrix}
   \langle s, \beta_{a_1,1}(t-2k) \rangle \\
   \langle s, \psi_{a_1,1}(t-2k) \rangle
\end{pmatrix}
= \frac{1}{\sqrt{1 + |e^{a_1}|^2}} \begin{pmatrix}
   1 \\
   -e^{a_1}
\end{pmatrix} \cdot \begin{pmatrix}
   s[2k] \\
   s[2k+1]
\end{pmatrix}
\]

(11.32)

which is consistent with (11.31) and \( i = 1 \). The reconstruction algorithm is obtained by straightforward matrix inversion

\[
\begin{pmatrix}
   s[2k] \\
   s[2k+1]
\end{pmatrix}
= \frac{1}{\sqrt{1 + |e^{a_1}|^2}} \begin{pmatrix}
   1 \\
   e^{a_1}
\end{pmatrix} \cdot \begin{pmatrix}
   s_1[k] \\
   v_1[k]
\end{pmatrix}
\]

(11.33)
Wavelet-domain methods

The final key observation is that the computation of the first level of wavelet coefficients is analogous to the determination of the discrete increment process \( u[k] = s[k] - \alpha_1 s[k-1] \) (see Section 8.5.1) in the sense that \( \nu_1[k] \propto u[2k+1] \) is a subsampled version of the latter.

11.4.2 From wavelet bases to tight wavelet frames

From now on, we shall pool the computed wavelet and approximation coefficients of a signal \( s \in \mathbb{R}^N \) in the wavelet vector \( v \) and formally represent the decomposition/reconstruction process (Equations (11.32) and (11.33)) by their vector-matrix counterparts \( v = W^T s \) and \( s = Wv \), respectively. Moreover, since the choice of the origin of the signal is arbitrary, we shall consider a series of “\( m \)-shifted” versions of the wavelet transform matrix \( W_m = Z^m W \) where \( Z \) (resp., \( Z^m \)) is the unitary matrix that circularly shifts the samples of the vector to which is applied by one (resp., by \( m \)) to the left. With this convention, the solution to the denoising problem (11.13) in the orthogonal basis \( W_m \) is given by

\[
\tilde{v}_m = \arg\min_v \left\{ \frac{1}{2} \| v - \frac{1}{m} W_m^T y \|_2^2 + \tau \Phi(v) \right\} = \text{prox}_\tau(W_m^T y),
\]

which amounts to a component-wise shrinkage of the wavelet coefficients. For reference, we also give the equivalent signal-domain (or analysis) formulation of the algorithm

\[
\tilde{s}_m = \arg\min_s \left\{ \frac{1}{2} \| s - \frac{1}{m} W_m^T y \|_2^2 + \tau \Phi(W_m^T s) \right\} = W_m \text{prox}_\tau(W_m^T y),
\]

Next, instead of a single orthogonal wavelet transform, we shall consider a wavelet frame expansion which is build from the concatenation of \( M \) shifted orthonormal transforms. The corresponding \((MN \times N)\) transformation matrix is denoted by

\[
A = \begin{bmatrix}
W_1^T \\
\vdots \\
W_M^T
\end{bmatrix}
\]

while the augmented wavelet vector is

\[
z = As = \begin{bmatrix}
v_1 \\
\vdots \\
v_M
\end{bmatrix}
\]

where \( v_m = W_m^T s \).

**Proposition 11.2** The transformation matrix \( A : \mathbb{R}^N \to \mathbb{R}^{MN} \), which is formed from
11.4 Improved denoising by consistent cycle spinning

the concatenation of $M$ orthonormal matrices $W_m$ as in (11.36), defines a tight frame of $\mathbb{R}^N$ in the sense that

$$\|Ax\|^2 = M\|x\|^2$$

for all $x \in \mathbb{R}^N$. Moreover, its pseudo-inverse $A^\dagger : \mathbb{R}^{MN} \rightarrow \mathbb{R}^N$ is given by

$$A^\dagger = \frac{1}{M} [W_1 \cdots W_M] = \frac{1}{M} A^T$$

with the property that

$$\arg\min_{x \in \mathbb{R}^N} \{\|z - Ax\|^2\} = A^\dagger z$$

for all $z \in \mathbb{R}^{MN}$ and $A^\dagger A = I$.

Proof. The frame expansion of $x$ is $z = Ax = (v_1, \ldots, v_M)$. The energy preservation then follows from

$$\|z\|^2 = \sum_{m=1}^{M} \|v_m\|^2 = \sum_{m=1}^{M} \|W_m^T x\|^2 = M \|x\|^2$$

where the equality on the right-hand side results from the application of Parseval's identity for each individual basis. Next, we express the quadratic error between an arbitrary vector $z = (z_1, \ldots, z_M) \in \mathbb{R}^{MN}$ and its approximation by $Ax$ as

$$\|z - Ax\|^2 = \sum_{m=1}^{M} \|z_m - W_m^T x\|^2 = \sum_{m=1}^{M} \|W_m z_m - x\|^2.$$ (by Parseval)

This error is minimized by setting its gradient with respect to $x$ to zero; that is,

$$\frac{\partial}{\partial x} \|z - Ax\|^2 = - \sum_{m=1}^{M} (W_m z_m - x) = 0,$$

which yields

$$x_{LS} = \frac{1}{M} \sum_{m=1}^{M} W_m z_m = A^\dagger z.$$

Finally, we check the left-inverse property

$$A^\dagger A = \frac{1}{M} \sum_{m=1}^{M} W_m W_m^T = I,$$

which follows from the orthonormality of the matrices $W_m$. $\square$

In practice, when the wavelet expansion is performed over $I$ resolution levels, the number of distinct shifted wavelet transforms is at most $M = 2^I$. In direct analogy with (11.13), the transposition of the wavelet-denoising problem to the context of
wavelet frames is then
\[
\hat{z} = \arg\min_z \left\{ \frac{1}{2} \| A^\dagger \tilde{z} - y \|^2_2 + \frac{1}{2M} \Phi(z) \right\}
\]
(11.37)
\[
= \arg\min_z \left\{ \frac{1}{2} \| z - Ay \|^2_2 + r \Phi(z) \right\}
\]
(due to the tight-frame property)
\[
= \text{prox}_r(Ay; r) = (\tilde{v}_1, \ldots, \tilde{v}_M),
\]
which simply amounts to performing \(M\) basic wavelet-denoising operations in parallel since the cost function is separable. We refer to (11.37) as the \textit{synthesis-with-cycle-spinning} formulation of the wavelet-denoising problem. The corresponding signal reconstruction is given by
\[
\tilde{s} = A^\dagger \hat{z} = \frac{1}{M} \sum_{m=1}^{M} \tilde{s}_m
\]
(11.38) which is the average of the solutions (11.34) obtained with each individual wavelet basis.

A remarkable property is that the cycle-spun version of wavelet denoising is guaranteed to improve upon the non-redundant version of the algorithm.

**Proposition 11.3** Let \(y = s + n\) be the samples of a signal \(s\) corrupted by zero-mean i.i.d. noise \(n\) and \(\tilde{s}_m\) the corresponding signal estimates given by the wavelet-based denoising algorithm (11.35) with \(m = 1, \ldots, M\). Then, under the assumption that the mean-square errors of the individual wavelet denoisers are equivalent, the averaged signal estimate (11.38) satisfies
\[
E\{\| \tilde{s} - s \|^2 \} \leq E\{\| \tilde{s}_m - s \|^2 \}
\]
for any \(m = 1, \ldots, M\).

**Proof** The residual noise in the orthonormal wavelet basis \(W_m\) is \((\tilde{v}_m - E\{v_m\})\) where \(\tilde{v}_m = W^\dagger_m \tilde{s}_m\) and \(E\{v_m\} = W^\dagger_m E\{y\} = W^\dagger_m s\) because of the assumption of zero-mean noise. This allows us to express the total noise power over the \(M\) wavelet bases as
\[
\| \tilde{z} - As \|^2 = \sum_{m=1}^{M} \| \tilde{v}_m - E\{v_m\} \|^2.
\]
The favorable aspect of considering a redundant representation is that the inverse frame operator \(A^\dagger\) is an orthogonal projector onto the signal space \(\mathbb{R}^N\) with the property that
\[
\| A^\dagger w \|^2 \leq \frac{1}{M^2} \| w \|^2
\]
for all \(w \in \mathbb{R}^{MN}\). This follows from the Pythagorean relation \(\| w \|^2 = \| AA^\dagger w \|^2 + \| (I - AA^\dagger) w \|^2\) (projection theorem) and the tight-frame property which is equivalent to \(\| AA^\dagger w \|^2 = M\| A^\dagger w \|^2\). By applying this result to \(w = \tilde{z} - As\), we obtain
\[
\| A^\dagger (\tilde{z} - As) \|^2 = \| \tilde{s} - s \|^2 \leq \frac{1}{M} \sum_{m=1}^{M} \| \tilde{v}_m - E\{v_m\} \|^2.
\]
(11.39)
Next, we take the statistical expectation of (11.39) which yields
\[
E(\|s - \hat{s}\|^2) \leq \frac{1}{M} \sum_{m=1}^{M} E \{ \|v_m - W_m^T s\|^2 \}. \tag{11.40}
\]

The final result then follows from Parseval’s relation (norm preservation of individual wavelet transforms) and the weak stationarity hypothesis (MSE equivalence of shifted-wavelet denoisers).

Note that this general result does not depend on the type of wavelet-domain processing—MAP vs. MMSE, or even, scalar vs. vectorial—as long as the nonlinear mapping \( \hat{v}_m = f(v_m) \) is fixed and applied in a consistent fashion. The inequality in Proposition 11.3 also suggests that one can push the denoising performance further by optimizing the MSE globally in the signal domain which is not the same as minimizing the error for each individual wavelet denoiser. The only downside of the redundant synthesis formulation (11.37) is that the underlying cost function looses its statistical interpretation (e.g., MAP criterion) because of the inherent coupling that results from considering multiple series of wavelet coefficients. The fundamental limitation there is that it is impossible to specify a proper innovation model in an overcomplete system.

### 11.4.3 Iterative MAP denoising

The alternative way of making use of wavelet frames is the dual analysis formulation of the denoising problem
\[
\hat{s} = \arg\min_s \left\{ \frac{1}{2} \|s - y\|^2 + \frac{r}{M} \Phi(As) \right\}. \tag{11.41}
\]

The advantage there is that the cost function is compatible with the statistical innovation based formulation of Section 10.4.3 provided that the weights and components of the potential function are properly chosen. In light of the comment at the end of Section 11.4.1, the equivalence with MAP estimation is exact if we perform a single level of wavelet decomposition with \( M = 2 \) and if we do not apply any penalty to the lowpass coefficients \( s_1 \). Yet, the price to pay is that we are now facing a harder optimization problem.

The difficulty stems from the fact that we do no longer benefit from Parseval’s norm equivalence between the signal and wavelet domains. A workaround is to re-instate the equivalence by imposing that the wavelet-frame expansion be consistent with the signal. This leads to the reformulation of (11.41) in synthesis form as
\[
\tilde{z} = \arg\min_z \left\{ \frac{1}{2} \|z - Ay\|^2 + r \Phi(z) \right\} \quad \text{s.t.} \quad AA^\dagger z = z \tag{11.42}
\]

which is the consistent cycle-spinning version of denoising. Rather than attempting to solve the constrained optimization problem (11.42) directly, we shall exploit the link with conventional wavelet shrinkage. To that end, we introduce the augmented Lagrangian penalty function
\[
\mathcal{L}_{\text{ad}}(z, x, \lambda; \mu) = \frac{1}{2} \|z - Ay\|^2 + r \Phi(z) + \frac{\mu}{2} \|z - Ax\|^2 - \lambda^T (z - Ax) \tag{11.43}
\]
Wavelet-domain methods

with penalty parameter \( \mu \in \mathbb{R}^+ \) and Lagrangian multiplier vector \( \lambda \in \mathbb{R}^{MN} \). Observe that the minimization of (11.43) over \((z, x, \lambda)\) is equivalent to solving (11.42). Indeed, the consistency condition \( z = Ax \) asserted by (11.43) is equivalent to \( A^\dagger z = z \), while the auxiliary variable \( x = A^\dagger z \) is the sought-after signal.

The standard strategy in the augmented-Lagrangian method of multipliers is to solve the problem iteratively by first minimizing \( L_A(z, x, \lambda; \mu) \) with respect to \((z, x)\) while keeping \( \mu \) fixed and updating \( \lambda \) according to the rule

\[
\lambda_{k+1} = \lambda_k - \mu(z_k^{k+1} - A x^{k+1}).
\]

Here, the task is simplified by applying the alternating-direction method of multipliers; that is, by first minimizing \( L_A(z, x, \lambda; \mu) \) with respect to \( z \) with \( x \) fixed and then the other way around. The link with conventional wavelet denoising is obtained by rewriting (11.43) as

\[
L_A(z, x, \lambda; \mu) = \frac{1}{2\tau} \|z - z\|^2_2 + \tau \Phi(z) + C_0(x, \lambda; \mu)
\]

where

\[
z = \frac{1}{1+\mu} (Ay + \mu Ax + \lambda)
\]

and \( C_0 \) is a term that does not depend on \( z \). Since the underlying cost function is separable, the solution of the minimization of \( L_A \) with respect to \( z \) is obtained by suitable shrinkage of \( z \), leading to

\[
z^{k+1} = \text{prox}_\Phi(z^{k+1}; \frac{1}{1+\mu})
\]

and involving the same kind of pointwise nonlinearity as Algorithm (11.34). The converse task of optimizing \( L_A \) over \( x \) with \( z = z^{k+1} \) fixed is a quadratic problem. The required partial derivatives are obtained as

\[
\frac{\partial L_A(z, x, \lambda; \mu)}{\partial x} = -\mu A^\top (z - Ax) - A^\top \lambda.
\]

This leads to the closed-form solution

\[
x^{k+1} = A^\dagger z^{k+1} - \frac{1}{\mu} A^\dagger \lambda^k,
\]

where we have taken advantage of the tight-frame/pseudo-inverse property \( A^\dagger A = I \) with \( A^\dagger = \frac{1}{\tau A^T} \).

The complete CCS (consistent cycle spinning) denoising procedure is summarized in Algorithm 3. It is an iterative variant of wavelet shrinkage where the thresholding function is determined by the statistics of the signal and applied in a consensus fashion. Its cost per iteration is \( O(N \times M) \) operations which is essentially that of the fast wavelet transform. This makes the method very fast. Since every step is the outcome of an exact minimization, the cost function decreases monotonically until the algorithm reaches a fixed point. The convergence to a global optimum is guaranteed when the potential function \( \Phi \) is convex.

One may also observe that the CCS denoising algorithm is similar to the MAP estimation method of Section 10.2.4 since both rely on ADMM. Besides the fact that
11.4 Improved denoising by consistent cycle spinning

Algorithm 3: CCS denoising solves Problem (11.41) where $A$ is a tight-frame matrix.

```
input: $y, s^0 \in \mathbb{R}^N, r, \mu \in \mathbb{R}^+$
set: $k = 0, A^0 = 0, u = Ay$
repeat
  $z^{k+1} = \text{prox}_{\phi}(\frac{1}{1+\mu}(u + \mu As^k + A^k), \frac{r}{1+\mu})$
  $s^{k+1} = A^{-1}(z^{k+1} - \frac{1}{\mu}A^k)$
  $A^{k+1} = A^k - \mu(z^{k+1} - As^{k+1})$
  $k = k + 1$
until stopping criterion
return $s = s^k$
```

The latter can handle an arbitrary system matrix $H$, the crucial difference is in the choice of the auxiliary variable $u = Ls$ (discrete innovation) vs. $z = As$ (redundant wavelet transform). While the two representations have a significant intersection, the tight wavelet frame has the advantage of resulting in a better conditioned problem (because of the norm-preservation property) and hence a faster convergence to the solution.

11.4.4 Iterative MMSE denoising

The material in this section is currently under review. It is available from the authors upon request.
Wavelet-domain methods

Figure 11.9 SNR improvement as a function of the level of noise for Brownian motion. The wavelet-denosing methods by reverse order of performance are: standard soft-thresholding (ortho-ST), optimal shrinkage in a wavelet basis (ortho-MAP/MMSE), shrinkage in a redundant system (frame-MAP/MMSE), and optimal shrinkage with consistent cycle spinning (CCS-MAP/MMSE).

Figure 11.10 SNR improvement as a function of the level of noise for a Lévy process with Laplace-distributed increments. The wavelet-denosing methods by reverse order of performance are: ortho-MAP (equivalent to soft-thresholding with fixed \( \theta \)), ortho-MMSE, frame-MMSE, frame-MAP, CCS-MAP, and CCS-MMSE. The results of CCS-MMSE are undistinguishable from the ones of the reference MMSE estimator obtained using message passing (see Figure 10.12).
Figure 11.11 SNR improvement as a function of the level of noise for a compound Poisson process (piecewise-constant signal). The wavelet-denoising methods by reverse order of performance are: ortho-ST, ortho-MMSE, frame-MMSE, and CCS-MMSE. The results of CCS-MMSE are undistinguishable from the ones of the reference MMSE estimator obtained using message passing (see Figure 10.10).

Figure 11.12 SNR improvement as a function of the level of noise for a Lévy flight with Cauchy-distributed increments. The wavelet-denoising methods by reverse order of performance are: ortho-MAP, ortho-MMSE, frame-MMSE, frame-MAP, CCS-MAP, and CCS-MMSE. The results of CCS-MMSE are undistinguishable from the ones of the reference MMSE estimator obtained using message passing (See Figure 11.11).

11.5 Bibliographical notes

The earliest instance of wavelet-based denoising is a soft-thresholding algorithm that was developed for magnetic resonance imaging [WYHJC91]. The same algorithm was discovered independently by Donoho and Johnstone for the reconstruction of signals from noisy samples [DJ94]. The key contribution of these authors was to establish the statistical optimality of the procedure in a minimax sense as well as its smoothness-preservation properties [Don95, DJ95]. This series of papers triggered the interest of the statistical and signal-processing communities and got researchers working on extending the technique and applying it to a variety of image reconstruction problems.
Sections 11.1 and 11.2

The discovery of the connection between soft-thresholding and $\ell_1$-minimization [Tib96, CDLL98] was a significant breakthrough. It opened the door to a variety of novel methods for the recovery of sparse signals based on non-quadratic wavelet-domain regularization, while also providing a link with statistical estimation techniques. For instance, Figueiredo and Nowak [FN03] developed an approach for image restoration based on the maximization of a likelihood criterion that is equivalent to (11.5) with a Laplacian prior. These authors also introduced an expectation-maximization algorithm that is one of the earliest incarnations of ISTA. The algorithm was brought into prominence when Daubechies et al. were able to establish its convergence for general linear inverse problems [DDDM04]. This motivated researchers to improve the convergence speed of ISTA through the use of appropriate preconditioning and/or over-relaxation [BDF07, VU08, FR08, BT09b]. A favored algorithm is FISTA because of its ease of deployment and superior convergence guarantees [BT09b]. Biomedical-imaging applications of wavelet-based image reconstruction include image restoration [BDF07], 3-D deconvolution microscopy [VU09], and parallel MRI [GKHPU11]. These works involve accelerated versions of ISTA or FISTA that capitalize on the specificities of the underlying system matrices.

Section 11.3

The signal-processing community's response to the publication of Donoho and Johnstone's work on the optimality of wavelet-domain soft-thresholding was a friendly competition to improve denoising performance. The Bayesian reformulation of the basic signal-denoising problem naturally led to the derivation of thresholding functions that are optimal in the MMSE sense [SA96, Sil99, ASS98]. Moulin and Lui presented a mathematical analysis of pointwise MAP estimators, establishing their shrinkage behavior for heavy-tailed distributions [ML99]. In this statistical view of the problem, the thresholding function is determined by the assumed prior distribution of the wavelet coefficients, the most prominent choices being the generalized Gaussian distribution [Mal89, CYV00, PP06] or a mixture of Gaussians with a peak at the origin [CKM97]. While infinite divisibility is not a property that has been emphasized in the image-processing literature, researchers have considered a number of wavelet-domain models that are compatible with the property and therefore part of the general framework investigated in Section 11.3. These include the Laplace distribution (see [HY00, Mar05] for pointwise MMSE estimator), SåS laws (see [ATB03] and [ABT01, BF06] for pointwise MAP and MMSE estimators, respectively), the Cauchy distribution [BAS07], as well as the sym gamma family [FB05]. The latter choice (a.k.a. Bessel K-form) is supported by a constructive model of images that is reminiscent of generalized Poisson processes [GS01]. There is also experimental evidence that this class of models is able to fit the observed transform-domain histograms well over a variety of natural images [MG01, SLG02).

The multivariate version of (11.21) can be found in [Ste81, Eq. (3.3)]. This formula is also central to the derivation of Stein's unbiased risk estimator (SURE), which provides a powerful data-driven scheme for adjusting the free parameters of a statist-
ical estimator under the AWGN hypothesis. SURE has been applied to the automatic adjustment of the thresholding parameters of wavelet-based denoising algorithms such as the SURE-shrink [DJ95] and SURELET [LBU07] approaches.

The possibility of defining bivariate shrinkage functions for exploiting inter-scale wavelet dependencies is investigated in [SS02].

Section 11.4

The concept of redundant wavelet-based denoising was introduced by Coifman under the name of cycle spinning [CD95]. The fact that this scheme always improves upon non-redundant wavelet-based denoising (see Proposition 11.3) was pointed out by Raphan and Simoncelli [RS08]. A frame is an overcomplete (and stable) generalization of a basis; see for instance [Ald95, Chr03]. For the design and implementation of the operator-like wavelets (including the first-order ones which are orthogonal), we refer to [KU06].

The concept of consistent cycle spinning was developed by Kamilov et al. [KBU12]. The CCS Haar-denoising algorithm was then modified appropriately to provide the MMSE estimator for Lévy processes [KBU12].