Mathematical Foundations of Data Sciences

Gabriel Peyré
CNRS & DMA, École Normale Supérieure

CHAPTER 7

Denoising

September 9, 2026

Denoising estimates a clean signal from measurements corrupted by noise. Its central difficulty is to suppress random fluctuations without erasing edges, textures, or other features of the underlying signal. We model common acquisition noise, compare linear filters with nonlinear thresholding, analyze how regularity and sparsity govern reconstruction error, and adapt the methods to Poisson and multiplicative noise.

7.1 Noise Modeling

7.1.1 Noise in Images

Figure 7.1 illustrates the variety of imaging noise. Different acquisition devices produce different noise statistics, while the images themselves vary in smoothness and geometry.

The model should reflect both the acquisition process and the noise statistics. The regularity and geometry of the clean signal then guide the choice of a basis for its representation.

Since noise perturbs discrete measurements from acquisition devices, we consider only finite-dimensional real signals $f \in \mathbb{R}^N$.

7.1.2 Image Formation

Figure 7.2 summarizes the model: a clean image f_0 combines with noise w to produce the observation $Y := f_0 \oplus w$. The operation $\oplus$ may be addition or multiplication. The observed image Y is modeled as a random vector.

In the statistical model, w is a random vector with a specified distribution. A numerical experiment uses one realization, also written w.

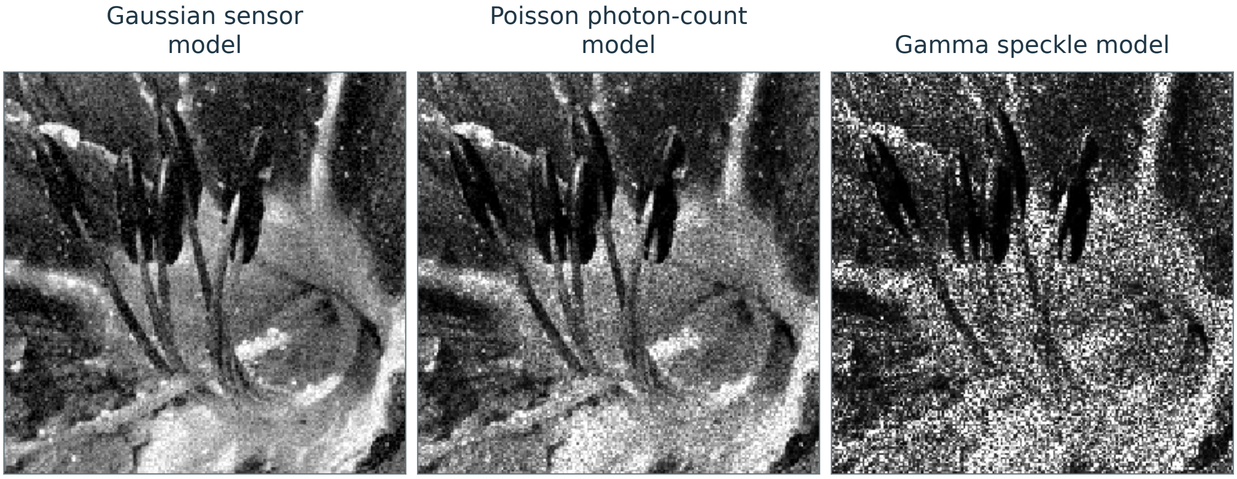

Figure 7.1. Gaussian, Poisson, and multiplicative Gamma noise simulated on the same flower image.

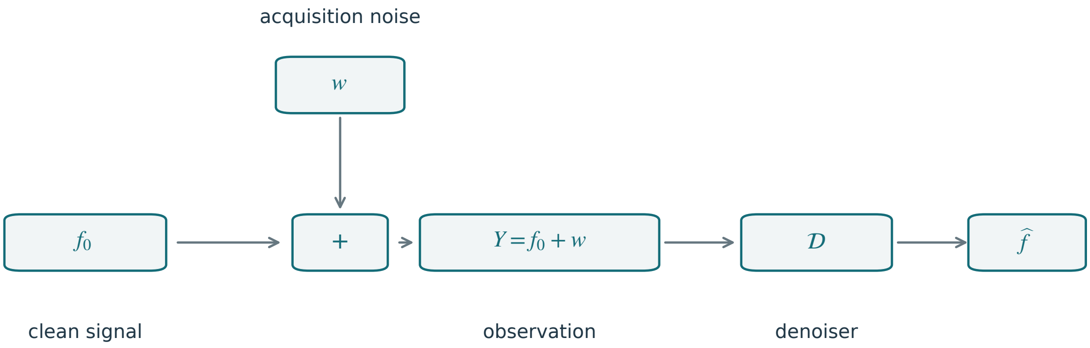

Figure 7.2. Image acquisition, noise, and reconstruction by denoising.

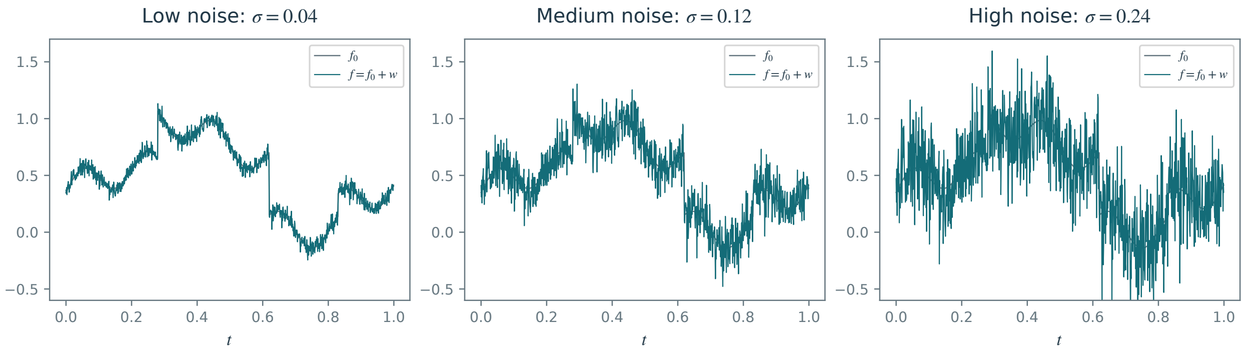

Figure 7.3. One-dimensional additive Gaussian noise at low, medium, and high levels, with a common clean signal and normalized noise realization.

Additive Noise. The simplest image formation model adds noise to a clean signal f_0

$$Y := f_0 + w$$

where w is the noise residual. Both w and Y are random; an acquired image is one realization of Y, denoted by y or f below. Figures 7.3 and 7.4 illustrate additive noise on a signal and an image.

The simplest model assumes independent centered Gaussian entries w_n with variance σ^2: $w \sim \mathcal{N}(0, \sigma^2 \mathrm{Id}_N)$. This is Gaussian white noise.

Depending on the image acquisition device, other noise distributions may be appropriate, including uniform noise $w_n \in [-a, a]$ or heavy-tailed generalized Gaussian noise with density

$$p_w(u) \propto e^{-|u/s|^\alpha}, \qquad s > 0, \quad 0 < \alpha < 2.$$

In many situations, the noise is not additive; its intensity may, for instance, depend on the signal intensity. This is the case with the Poisson and multiplicative noise models considered in Section 7.4.

7.1.3 Denoiser

A denoiser produces an estimate $\tilde{f} = \mathcal{D}(Y)$ of f_0 from the observation Y alone. The map $\mathcal{D}$ is deterministic, but its output is random because it depends on the noise w. With centered additive noise, Y has mean f_0, so denoising estimates that mean from a single realization. Figure 7.5 illustrates the result.

The quality of a denoiser is measured using the mean squared risk $\mathbb{E}_w \|f_0 - \tilde{f}\|^2$, where $\mathbb{E}_w$ averages over the noise w. This risk is a theoretical quantity because f_0 is unknown in practice. In experiments with known clean data, we instead evaluate one realization $y \sim Y = f_0 + w$ using the signal-to-noise ratio $\mathrm{SNR}(\mathcal{D}(y), f_0)$, defined by

$$\mathrm{SNR}(f, g) := -20 \log_{10}(\|f - g\| / \|g\|).$$

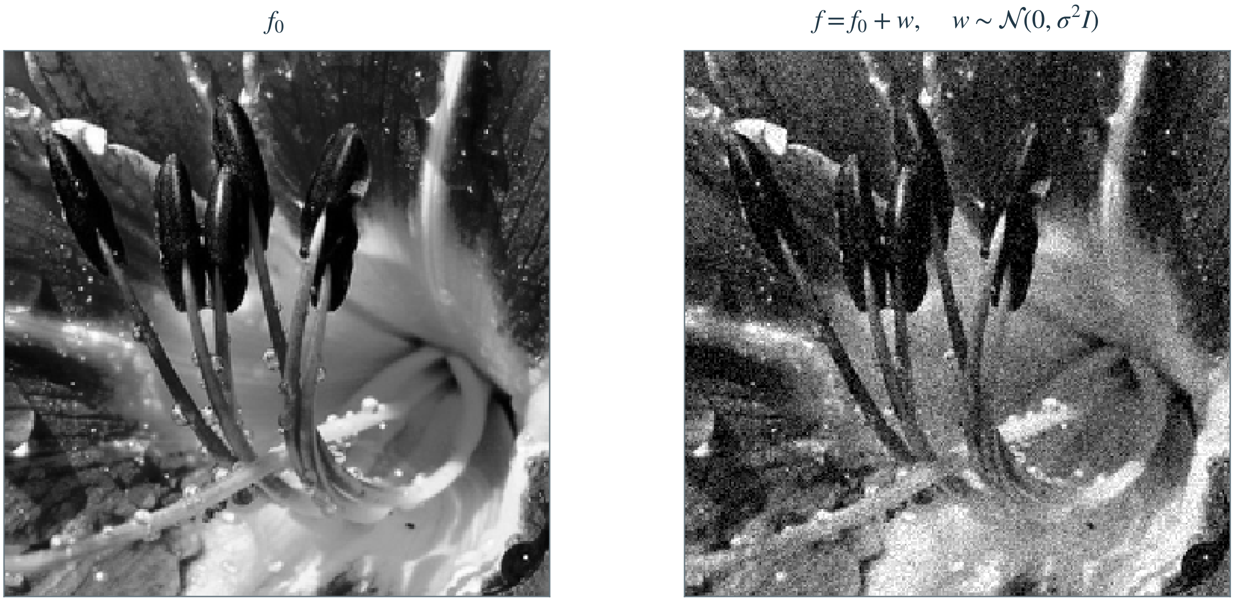

Figure 7.4. Two-dimensional additive noise: the clean image f_0 and the observation $f = f_0 + w$, with $w \sim \mathcal{N}(0, \sigma^2 I)$.

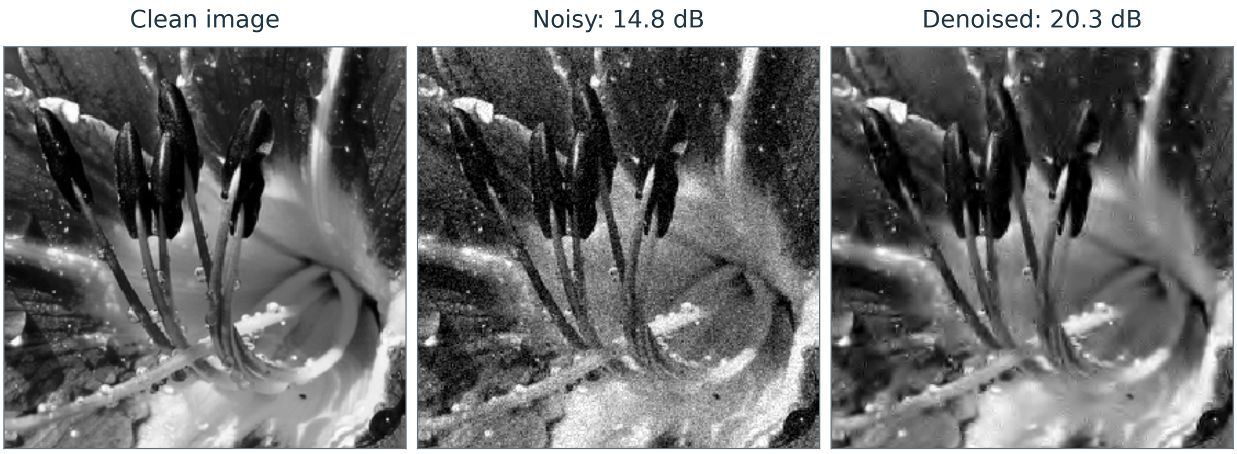

Figure 7.5. Left: clean image, center: noisy image, right: denoised image.

The SNR is measured in decibels (dB). Computing it requires the clean signal f_0, so it serves as a benchmark in controlled experiments. Because an ℓ^2 error does not fully describe perceptual quality, visual inspection should accompany the numerical comparison.

7.2 Linear Denoising by Filtering

7.2.1 Translation Invariant Estimators

A linear estimator $\mathcal{D}(Y) = \tilde{f}$ of f_0 preserves addition, $\mathcal{D}(f + g) = \mathcal{D}(f) + \mathcal{D}(g)$, and scalar multiplication. A translation-invariant estimator commutes with translations: $\mathcal{D}(f_\tau) = \mathcal{D}(f)_\tau$, where $f_\tau(t) = f(t - \tau)$. We assume periodic boundary conditions, in either one or two dimensions. An estimator with both properties is a convolution filter:

$$\mathcal{D}(f) = f \star h$$

where $h \in \mathbb{R}^N$ is a filter. Denoising filters are usually low-pass. To preserve constant signals, one imposes

$$\sum_n h_n = \hat{h}_0 = 1$$

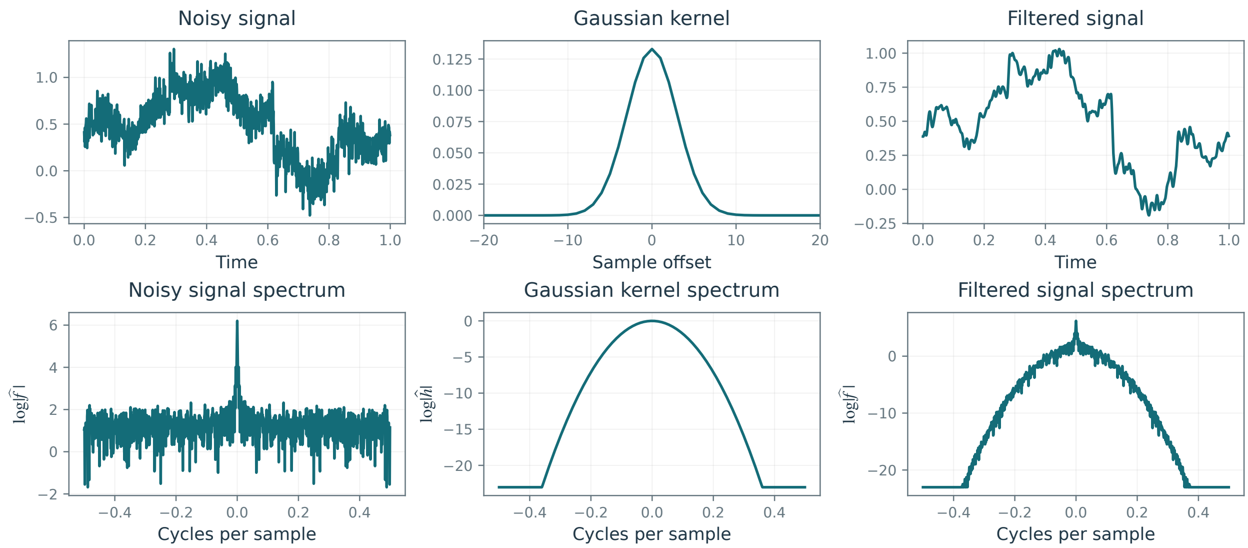

Figure 7.6. Gaussian filtering: noisy signal, kernel, and filtered signal (top), with their Fourier magnitudes below.

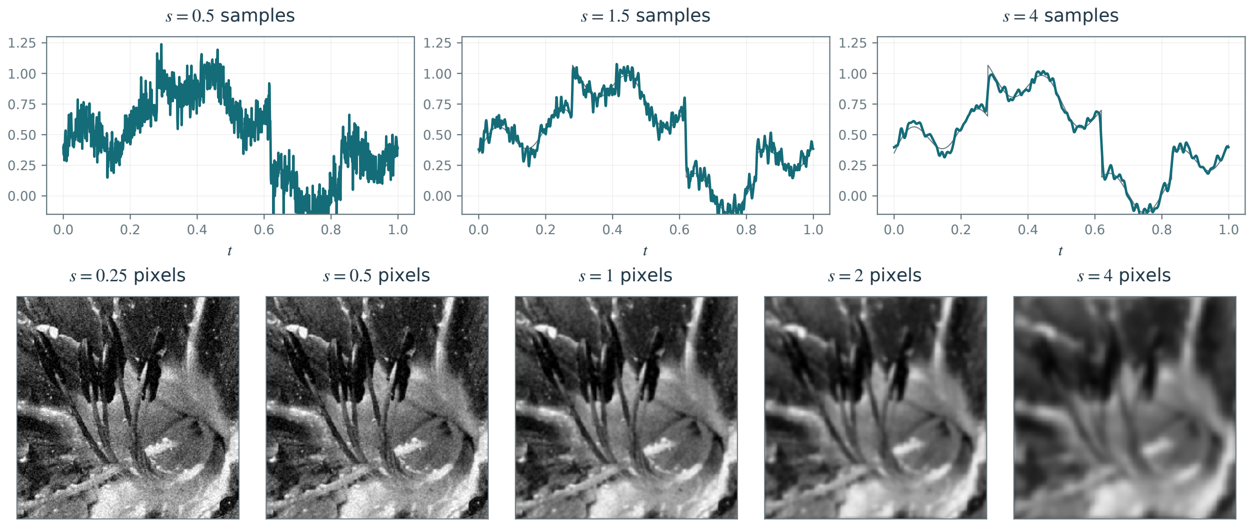

Figure 7.7. Gaussian filtering with increasing width: three signals (top) and five images (bottom).

where $\hat{h}$ is the discrete Fourier transform.

Figure 7.6 illustrates denoising with a low-pass filter.

The filtering strength is usually controlled by the width s of h. A typical example is the (discretized) Gaussian filter

$$\forall -N/2 < i \leqslant N/2, \quad h_{s,i} = \frac{1}{Z_s} \exp\left(-\frac{i^2}{2s^2}\right) \tag{7.1}$$

where Z_s normalizes the filter so that $\sum_i h_{s,i} = 1$, preserving constant signals. Figure 7.6 shows the effect of Gaussian filtering in the spatial and Fourier domains.

Figure 7.7 shows the effect of increasing the filter width s. Linear filtering works well on smooth signals, but removing more noise also blurs edges and other singularities.

7.2.2 Optimal Filter Selection and Bias-Variance Tradeoff

The best filter depends on both the unknown signal f_0 and the noise level σ. Rather than optimize over all filters, we can tune a width parameter s within a family such as the Gaussian filters in (7.1).

The triangle inequality separates the smoothing error from the remaining noise:

$$\|\tilde{f} - f_0\| \leqslant \|h_s \star f_0 - f_0\| + \|h_s \star w\|$$

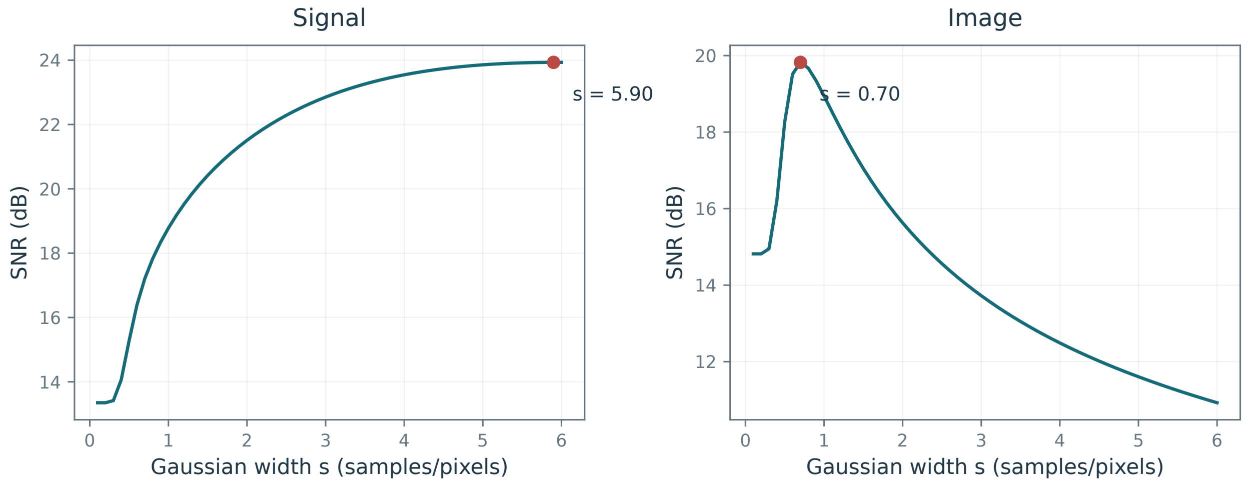

Figure 7.8. SNR as a function of filter width for a signal (left) and an image (right).

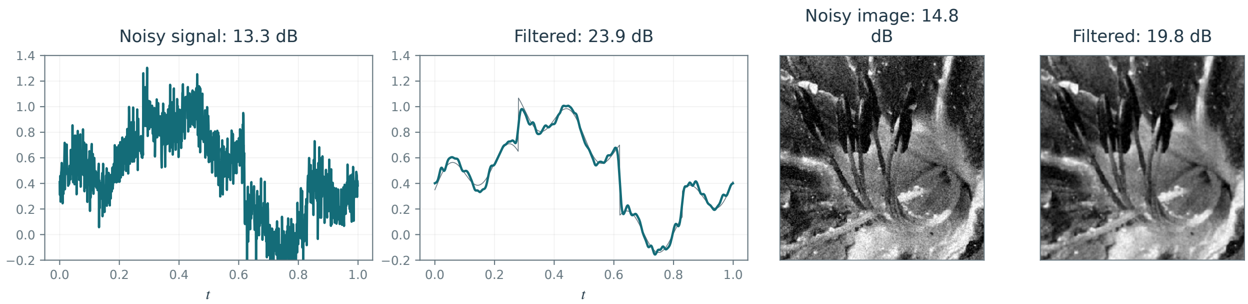

Figure 7.9. Gaussian filtering at the SNR-optimal widths. From left to right: noisy signal, filtered signal, noisy image, and filtered image.

The filter width s balances noise removal against excessive smoothing of singularities. A wider filter typically reduces $\|h_s \star w\|$ while increasing $\|h_s \star f_0 - f_0\|$.

Figure 7.8 shows the SNR as a function of filter width, evaluated using the known clean signal.

Figure 7.9 uses the width $s^\star$ that maximizes SNR for the displayed observation. This choice uses the known clean signal to assess the best achievable result within the filter family.

Some noise remains at the SNR-optimal width. Increasing the filter width introduces strong blur and reduces the SNR, although the result may look more pleasant.

7.2.3 Oracle Denoiser

To study the best possible choice of h, first suppose that an oracle knows the clean signal f_0. Recall the model $Y = f_0 + w$, where w is Gaussian white noise: $w_k \sim \mathcal{N}(0, \sigma^2)$ and the w_k are independent and identically distributed. The clean signal f_0 is fixed. Choose an orthonormal basis $(\psi_k)_k$, typically the Fourier basis, and restrict attention to linear estimators that are diagonal in it:

$$\tilde{f} := \sum_k \lambda_k \langle Y, \psi_k \rangle \, \psi_k. \tag{7.2}$$

For the Fourier basis, this means $\tilde{f} = Y \star h$ where $\hat{h} = \lambda$. Note that $\tilde{f}$ is a random vector. The oracle may use f_0 to select the gains λ, but the estimator must still be linear in Y and have the displayed diagonal form. It cannot simply return f_0.

Denoting $c_k := \langle f_0, \psi_k \rangle$ (for the normalized discrete Fourier basis, $c = \hat{f}_0 / \sqrt{N}$), the expected risk of this denoiser is

$$\mathbb{E}_w \|\tilde{f} - f_0\|^2 = \sum_k \mathbb{E}_w \left| \langle \tilde{f} - f_0, \psi_k \rangle \right|^2 = \sum_k \mathbb{E}_w \left| \lambda_k \left(c_k + \langle w, \psi_k \rangle \right) - c_k \right|^2 .$$

For a real orthonormal basis, the transformed noise is again independent Gaussian noise with variance σ^2. For a complex unitary Fourier basis applied to real noise, each coefficient still has second moment σ^2, which suffices here, although conjugate-frequency coefficients are dependent. By expanding the square and using $\mathbb{E}(\langle w, \psi_k\rangle) = 0$, we obtain

$$\mathbb{E}_w \|\tilde{f} - f_0\|^2 = \sum_k \left(|c_k|^2 |\lambda_k - 1|^2 + |\lambda_k|^2 \sigma^2 \right).$$

Knowing c_k, the oracle minimizes $\mathbb{E}_w \left(\|\tilde{f} - f_0\|^2 \right)$ over λ. The minimization separates over the indices k:

$$\min_{\lambda_k} \left(|c_k|^2 |\lambda_k - 1|^2 + |\lambda_k|^2 \sigma^2 \right).$$

The solution thus satisfies

$$|c_k|^2 (\lambda_k - 1) + \lambda_k \sigma^2 = 0,$$

i.e.,

$$\lambda_k = \frac{|c_k|^2}{|c_k|^2 + \sigma^2}.$$

If $\sigma = 0$, choosing $\lambda = 1$ gives the identity map. When $|c_k|$ decays, increasing σ suppresses the weaker coefficients more strongly, concentrating the gains on the dominant signal components.

7.2.4 Wiener Filter

We now replace access to the clean signal by a known probability model for it. Suppose both the noise $w \sim W$ and the clean signal $f_0 \sim F$ have known distributions. We seek a filter minimizing the error averaged over both sources of randomness, assuming W and F are independent. We assume $\mathcal{D}$ has the form (7.2).

The optimal h thus minimizes

$$\mathbb{E}_{w,F} \|\mathcal{D}(F + w) - F\|^2$$

Independence of W and F allows the same risk calculation:

$$\mathbb{E}_{w,F} \|\mathcal{D}(F + w) - F\|^2 = \sum_k \left(\alpha_k |\lambda_k - 1|^2 + |\lambda_k|^2 \sigma^2 \right) \quad \text{where} \quad \alpha_k := \mathbb{E}_F |\langle F, \psi_k \rangle|^2$$

The optimal filter thus has coefficients

$$\lambda_k = \frac{\alpha_k}{\alpha_k + \sigma^2}. \tag{7.3}$$

Relative to the oracle filter, the individual energy $|c_k|^2$ is replaced by the expected coefficient energy $\alpha_k \geqslant 0$, the signal power spectrum in the Fourier basis. This filter is known as the Wiener filter.

For the normalized discrete Fourier basis, $\alpha_k = N^{-1} \mathbb{E} |\hat{F}_k|^2$. Define the averaged autocorrelation

$$\eta_j = \frac{1}{N} \mathbb{E} \sum_{n=0}^{N-1} F_n \overline{F_{n-j}}, \qquad \alpha_k = \hat{\eta}_k.$$

Here indices are periodic and $\hat{\eta}$ uses the unnormalized DFT of Chapter 2. This normalization is needed because the Fourier atoms used in the risk calculation have unit norm.

So far, $\mathcal{D}$ has been restricted to the basis ψ_m through the diagonal form (7.2). If F is stationary, its law is unchanged by shifts: $F(\cdot - \tau)$ has the same law as F. In that case, the Fourier Wiener filter (7.3) is optimal among all linear denoisers, even without imposing a diagonal form in advance.

An empirical estimate based on one noisy realization is $N^{-1} |\hat{y}_k|^2$. Its expectation is $\alpha_k + \sigma^2$, and its variance can be large. Empirical Wiener methods therefore need noise subtraction, nonnegativity constraints, and usually additional smoothing or a signal model.

7.2.5 Denoising and Linear Approximation

Approximation theory, developed in Chapter 5, provides a different route to denoising bounds. It requires a class of clean signals rather than a probability distribution on that class. Fix an orthonormal basis $\mathcal{B} = (\psi_m)_m$ of $\mathbb{R}^N$. A simple estimator retains only the first M coefficients of the noisy observation in $\mathcal{B}$:

$$\mathcal{D}(f) = \tilde{f} \stackrel{\text{def.}}{=} \sum_{m=1}^{M} \langle f, \psi_m \rangle \psi_m. \tag{7.4}$$

This is a linear projection onto the space spanned by $(\psi_m)_{m=1}^M$. This denoising scheme is parameterized by the number $1 \leqslant M \leqslant N$ of retained coefficients. Increasing M retains more signal detail but also more noise. For the discrete Fourier basis ordered by increasing frequency, this is ideal low-pass filtering by convolution with a discrete Dirichlet kernel. This is the special case of (7.2) in which the gains λ_m are binary, $\lambda_m \in \{0, 1\}$. Although restrictive, this choice yields useful rates under the approximation assumptions of Theorem 7.1. It requires a regularity model for f_0, but does not assume a random signal distribution F.

The proof is finite-dimensional, but its bound has no explicit dependence on N. Extending it to a continuum model requires interpreting white noise as a generalized random process.

Theorem 7.1. *Assume $\sigma > 0$, $\beta > 0$, and*

$$\|f_0 - f_{0,M}^{\text{lin}}\|^2 \leqslant CM^{-2\beta}, \qquad f_{0,M}^{\text{lin}} = \sum_{m=1}^{M} \langle f_0, \psi_m \rangle \psi_m, \quad 1 \leqslant M \leqslant N.$$

The projection denoiser (7.4) has risk

$$\mathbb{E}\|f_0 - \tilde{f}\|^2 = M\sigma^2 + \|f_0 - f_{0,M}^{\text{lin}}\|^2.$$

If $0 < \sigma^2 \leqslant C$, choose

$$t = (C/\sigma^2)^{1/(2\beta+1)}, \qquad M = \min\{N, \lceil t \rceil\}. \tag{7.5}$$

Then

$$\mathbb{E}\|f_0 - \tilde{f}\|^2 \leqslant 3C^{1/(2\beta+1)}\sigma^{4\beta/(2\beta+1)}.$$

Proof. Orthogonality separates retained noise coefficients from discarded signal coefficients. Each retained noise coefficient has variance σ^2, which proves the risk identity. If $t \leqslant N$, then $t \leqslant M \leqslant 2t$ and

$$M\sigma^2 + CM^{-2\beta} \leqslant 2t\sigma^2 + Ct^{-2\beta} = 3C^{1/(2\beta+1)}\sigma^{4\beta/(2\beta+1)}.$$

If $t > N$, the estimator retains all coefficients and its risk is $N\sigma^2 \leqslant t\sigma^2$, which obeys the same bound. □

Several observations help interpret this result:

- The approximation assumption bounds the denoising error $\mathbb{E}(\|f_0 - \tilde{f}\|^2)$ independently of the sample count N, even though the total input noise energy $\mathbb{E}(\|w\|^2) = N\sigma^2$ grows with N.
- A larger approximation exponent β gives faster decay of the error as σ tends to zero. The exponent approaches that of the parametric rate σ^2 as β increases.
- The proof uses a finite sample count N. White noise with constant positive variance in every coordinate has infinite energy in an infinite-dimensional Hilbert space, although other random vectors may have finite energy. The balancing choice applies when $N \geqslant t$; otherwise the estimator is the identity. A continuum interpretation as $\sigma \to 0$ also requires increasing the sampling resolution.
- Section 5.3.1 bounds approximation errors for continuous signal and image models. To transfer these bounds to sampled data, choose N large enough that discretization error is smaller than denoising error.

The Sobolev model in Section 5.2.1 provides a useful example. For discrete signals with an increasing sample count N, especially as σ decreases, a compatible weighted coefficient bound gives the same approximation argument as Proposition 5.3.

Proposition 7.2. *Let $\alpha > 0$. If*

$$\sum_{m=1}^{N} m^{2\alpha} |\langle f_0, \psi_m \rangle|^2 \leqslant C \tag{7.6}$$

then

$$\forall M \in \{1, \ldots, N\}, \quad \|f_0 - f_{0,M}^{lin}\|^2 \leqslant CM^{-2\alpha}$$

Proof.

$$C \geqslant \sum_{m=1}^{N} m^{2\alpha} |\langle f_0, \psi_m \rangle|^2 \geqslant \sum_{m>M} m^{2\alpha} |\langle f_0, \psi_m \rangle|^2 \geqslant M^{2\alpha} \sum_{m>M} |\langle f_0, \psi_m \rangle|^2 \geqslant M^{2\alpha} \|f_0 - f_{0,M}^{\text{lin}}\|^2.$$

□

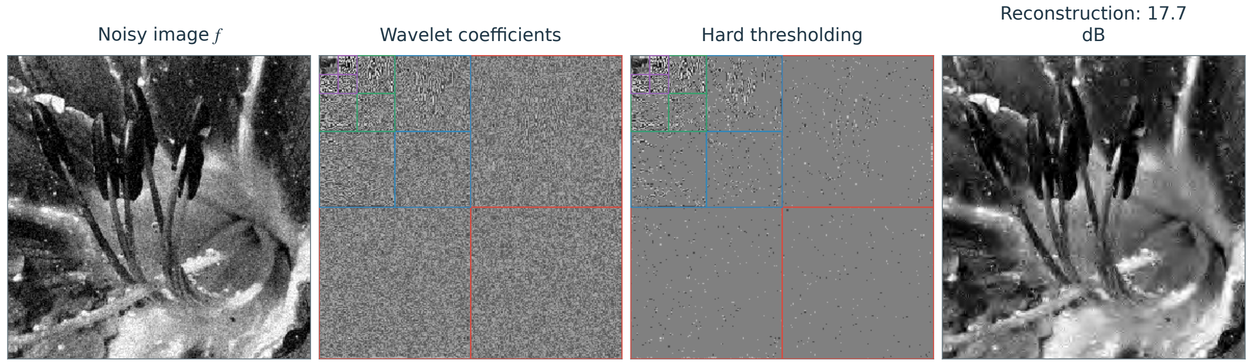

Figure 7.10. Denoising using thresholding of wavelet coefficients.

For normalized Fourier atoms ordered by increasing physical frequency, an appropriately scaled discrete weighted-energy bound approximates the periodic Sobolev model. The sample normalization, frequency weights, and boundary convention must match those of the continuous signal. Under this compatibility, low-frequency projection efficiently denoises smooth signals and images.

7.3 Nonlinear Denoising by Thresholding

7.3.1 Hard Thresholding

We consider a real orthonormal basis $\{\psi_m\}_m$ of $\mathbb{R}^N$, for instance a discrete wavelet basis. The noisy coefficients satisfy

$$\langle f, \psi_m \rangle = \langle f_0, \psi_m \rangle + \langle w, \psi_m \rangle. \tag{7.7}$$

Orthogonal transforms preserve Gaussian white noise, so the coefficients $\langle w, \psi_m \rangle$ remain independent centered Gaussians with variance σ^2. If the basis $\{\psi_m\}_m$ represents f_0 sparsely, most clean coefficients $\langle f_0, \psi_m \rangle$ are small, leaving many noisy coefficients dominated by noise; see Figure 7.10. Donoho and Johnstone systematically developed thresholding estimators based on this observation [1].

Hard thresholding retains only coefficients whose magnitude exceeds the threshold T:

$$\tilde{f} = \sum_{|\langle f, \psi_m \rangle| > T} \langle f, \psi_m \rangle \psi_m = \sum_m S_T(\langle f, \psi_m \rangle) \psi_m.$$

where S_T is defined in (5.4). Retaining the M largest noisy coefficients gives the approximation $\tilde{f} = f_M$ of f. Figure 7.10 illustrates how a suitable T removes much of the noise while preserving sharp features.

7.3.2 Soft Thresholding

We recall that the hard thresholding operator is defined as

$$S_T(x) = S_T^0(x) = \begin{cases} x & \text{if} \quad |x| > T, \\ 0 & \text{if} \quad |x| \leqslant T. \end{cases} \tag{7.8}$$

Hard thresholding makes a discontinuous keep-or-discard decision, which can create artifacts. Soft thresholding is a continuous alternative:

$$S_T^1(x) = \max(1 - T/|x|, 0)x. \tag{7.9}$$

Set $S_T^1(0) = 0$ by continuity. Figure 7.11 compares the two nonlinear maps.

For $q = 0$ and $q = 1$, these thresholding rules define two different estimators

$$\tilde{f}^q = \sum_m S_T^q(\langle f, \psi_m \rangle) \psi_m \tag{7.10}$$

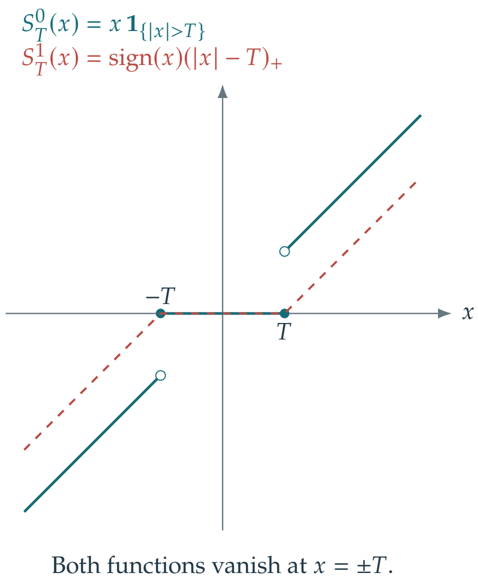

Figure 7.11. Hard and soft thresholding functions.

Preserving coarse-scale coefficients. Soft thresholding S^1_T biases retained coefficients by reducing their magnitudes. Shrinking coarse wavelet coefficients can introduce unwanted low-frequency artifacts. At a coarsest scale 2^{j_0}, it is therefore common to retain the approximation coefficients unchanged. In one dimension, this gives

$$\tilde{f}^1 = \sum_{0 \leqslant n < 2^{-j_0}} \langle f,\, \varphi_{j_0,n} \rangle \varphi_{j_0,n} + \sum_{j=J+1}^{j_0} \sum_{0 \leqslant n < 2^{-j}} S^1_T(\langle f,\, \psi_{j,n} \rangle) \psi_{j,n}.$$

Empirical choice of the threshold. Figure 7.12 plots SNR against the threshold T for a fixed natural image f_0. In this experiment, hard thresholding performs best near $T \approx 3\sigma$, and soft thresholding near $T \approx 3\sigma/2$. The best soft-thresholded estimate has a slightly higher SNR.

These values describe the displayed experiments; the best threshold depends on the image, basis, noise level, and loss function.

7.3.3 Minimax Optimality of Thresholding

Estimating sparse coefficients. To analyze the estimator and guide the choice of T, we first assume that the coefficients

$$a_{0,m} = \langle f_0,\, \psi_m \rangle \in \mathbb{R}, \qquad a_0 \in \mathbb{R}^N$$

are sparse: most entries $a_{0,m}$ vanish, so the ℓ^0 count

$$\|a_0\|_0 = \#\,\{m\ ;\ a_{0,m} \neq 0\}$$

is small. As shown in (7.7), noisy coefficients

$$\langle f,\, \psi_m \rangle = a_m = a_{0,m} + z_m$$

are perturbed by additive Gaussian white noise of variance σ^2. Figure 7.14 shows an example of such a noisy sparse signal.

Universal threshold value. When all nonzero clean coefficients are well separated from the noise level, choosing a threshold comparable to the largest pure-noise coefficient suppresses false detections:

$$T \approx \tau_N, \qquad \tau_N = \max_{0 \leqslant m < N} |z_m|.$$

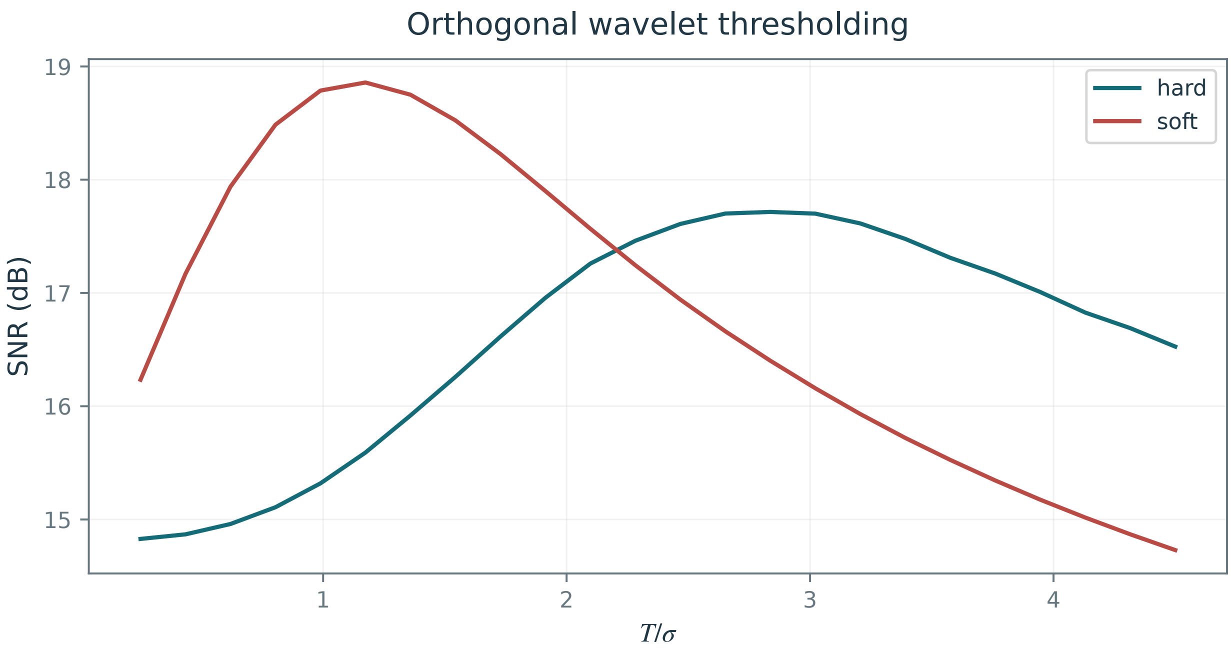

Figure 7.12. SNR as a function of T/σ for hard and soft thresholding.

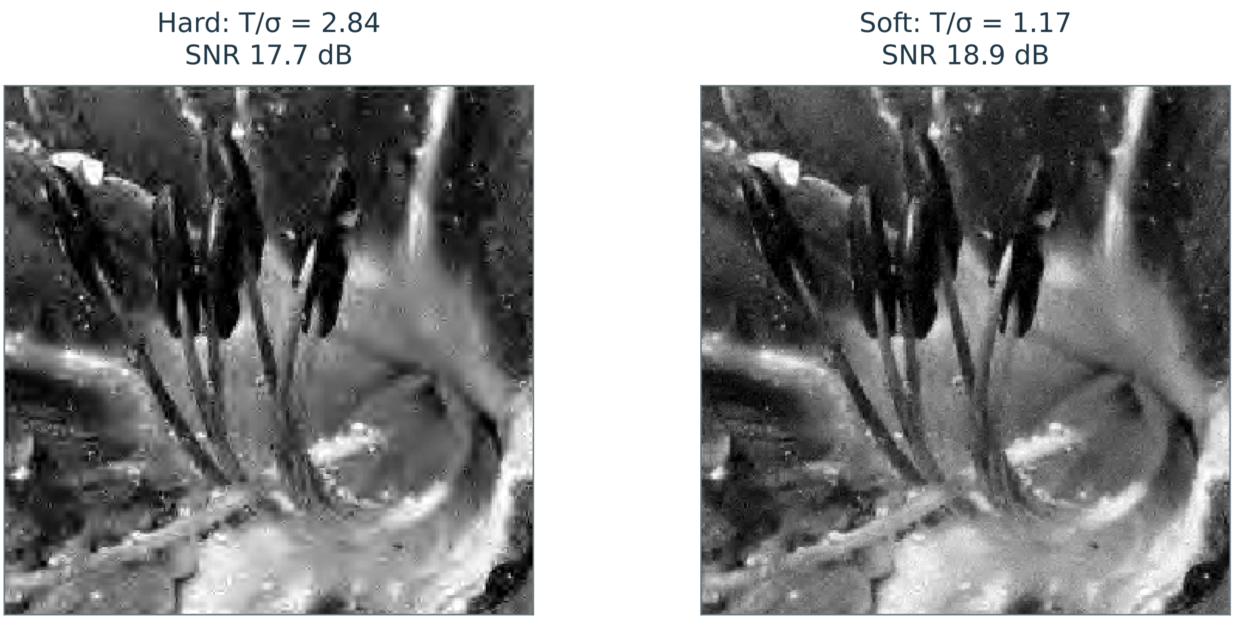

Figure 7.13. Comparison of hard (left) and soft (right) thresholding.

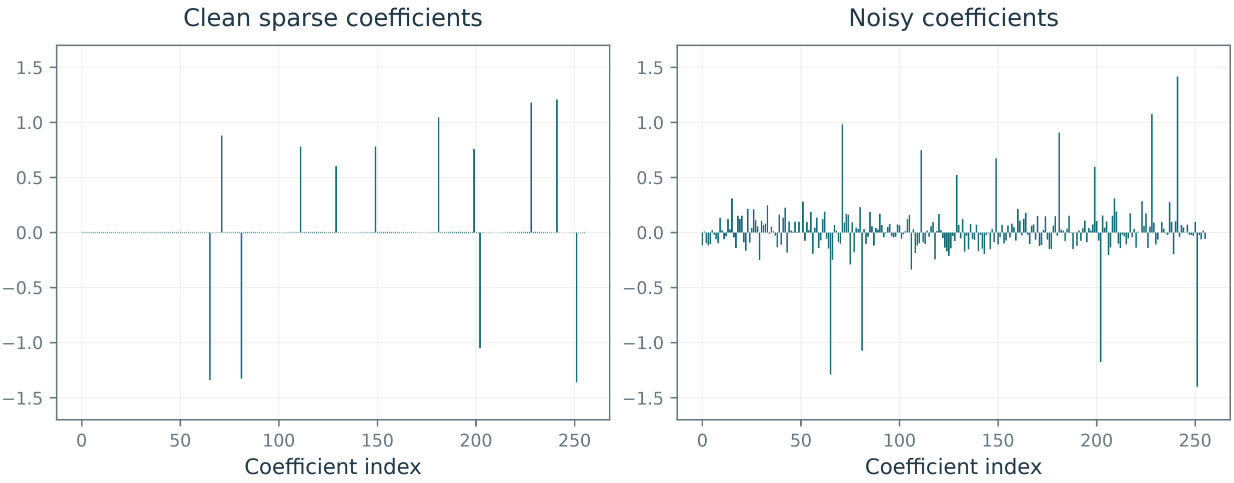

Figure 7.14. Left: clean sparse coefficients a_0; right: noisy coefficients a.

This motivates a deterministic threshold based on the typical size of τ_N; it is not an exact minimization formula for the realized denoising error. The random maximum τ_N depends on N. For fixed σ, its mean is asymptotic to $\sigma\sqrt{2\log N}$ as $N \to \infty$. Its variance tends to zero with N, so τ_N concentrates increasingly near its mean. Figure 7.15 illustrates these trends numerically.

Asymptotic optimality. Donoho and Johnstone [1] showed that the universal threshold $T = \sigma\sqrt{2\log(N)}$ gives a risk bound for signals that admit accurate nonlinear approximations in $\{\psi_m\}_m$. The risk is within a logarithmic factor of an ideal coefficient-selection risk.

Theorem 7.3. *Let $N \geqslant 2$ and $Y = f_0 + w$, with $w \sim \mathcal{N}(0, \sigma^2 I_N)$, in a real orthonormal basis. Hard or soft thresholding* (7.10) *with*

$$T = \sigma\sqrt{2\ln N}$$

satisfies an oracle inequality of the form

$$\mathbb{E}\|f_0 - \tilde{f}^q\|^2 \leqslant C_0(1 + \ln N)\left[\sigma^2 + \min_{0\leqslant M\leqslant N}\left(\|f_0 - f_{0,M}^{\mathrm{nlin}}\|^2 + M\sigma^2\right)\right],$$

where C_0 is an absolute constant and $f_{0,M}^{\mathrm{nlin}}$ retains the M largest clean coefficients. In particular, if $\|f_0 - f_{0,M}^{\mathrm{nlin}}\|^2 \leqslant CM^{-2\beta}$ for $1 \leqslant M \leqslant N$, with $\beta > 0$, then for $0 < \sigma \leqslant 1$,

$$\mathbb{E}\|f_0 - \tilde{f}^q\|^2 \leqslant C'(1 + \ln N)\sigma^{4\beta/(2\beta+1)},$$

where C' depends on C and β.

The oracle inequality is a standard Gaussian thresholding bound; see [1]. The rate follows by balancing $CM^{-2\beta}$ and $M\sigma^2$, with integer rounding and truncation at N as in Theorem 7.1. The added σ^2 term accounts for residual noise even at the zero signal.

The universal threshold $T = \sigma\sqrt{2\ln N}$ is conservative: it suppresses nearly all pure-noise coefficients. The smaller thresholds illustrated in Figure 7.13 often give better SNR in these experiments.

7.3.4 Translation Invariant Thresholding Estimators

Translation invariance. Let $f \mapsto \tilde{f} = \mathcal{D}(f)$ be a denoising method, and $f_\tau(x) = f(x - \tau)$ be a translated signal or image for $\tau \in \mathbb{R}^d$ ($d = 1$ or $d = 2$). The denoiser is translation invariant on the shift lattice Δ if

$$\forall \tau \in \Delta, \quad \mathcal{D}(f) = \mathcal{D}(f_\tau)_{-\tau}$$

where Δ is a lattice of $\mathbb{R}^d$. A denser shift lattice imposes equivariance at finer spatial resolution. This corresponds to the fact that $\mathcal{D}$ commutes with the translation operator.

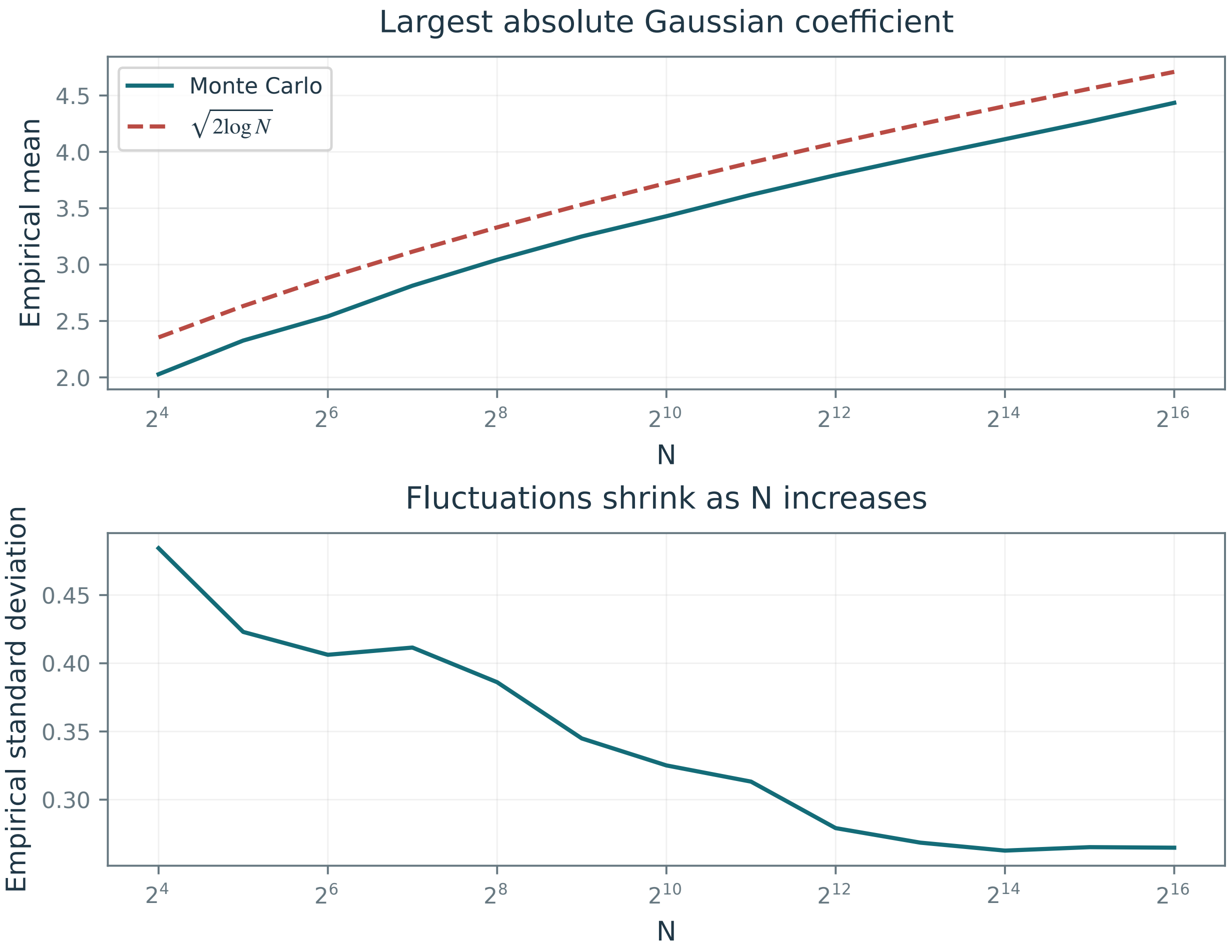

Figure 7.15. Empirical mean (top) and standard deviation (bottom) of the largest absolute Gaussian noise coefficient.

$$\begin{tikzcd}
f \arrow[r, "\mathcal{D}"] \arrow[d, "T_\tau"'] & \mathcal{D}(f) \arrow[d, "T_\tau"] \\
T_\tau f \arrow[r, "\mathcal{D}"'] & \mathcal{D}(T_\tau f) = T_\tau \mathcal{D}(f)
\end{tikzcd}$$

Translation invariance on a sufficiently fine set Δ prevents the denoiser from favoring particular feature locations. Without it, the same feature may be reconstructed differently depending on where it appears, producing visible artifacts.

For denoising by thresholding

$$\mathcal{D}(f) = \sum_m S_T(\langle f, \psi_m \rangle)\psi_m.$$

a sufficient condition is invariance of the basis $\{\psi_m\}_m$ under shifts in Δ, up to unit-modulus factors:

$$\forall\, m', \ \forall\, \tau \in \Delta, \ \exists m, \ \exists \lambda \in \mathbb{C}, \qquad (\psi_{m'})_\tau = \lambda \psi_m$$

where $|\lambda| = 1$.

On the periodic domain, the Fourier basis is invariant under every shift $\Delta = \mathbb{R}^d$ of $[0,1]^d$. The discrete Fourier basis is invariant under all integer shifts $\Delta = \{0, \ldots, N_0 - 1\}^d$, where $N = N_0$ counts samples in one dimension and $N = N_0 \times N_0$ counts pixels in two dimensions.

Unfortunately, an orthogonal wavelet basis

$$\{\psi_m = \psi_{j,n}\}_{j,n}$$

is generally not invariant under arbitrary shifts in either the continuous or discrete setting. For instance, in 1-D,

$$(\psi_{j',n'})_\tau \notin \{\psi_{j,n}\} \quad \text{for} \quad \tau = 2^{j'}/2.$$

Cycle spinning. For a finite group Δ of periodic shifts, average the shifted, denoised, and realigned outputs of $\mathcal{D}$ to obtain a translation-invariant estimator:

$$\mathcal{D}_{\text{inv}}(f) = \frac{1}{|\Delta|} \sum_{\tau \in \Delta} \mathcal{D}(f_\tau)_{-\tau}. \tag{7.11}$$

One checks that

$$\forall\, \tau \in \Delta, \quad \mathcal{D}_{\text{inv}}(f) = \mathcal{D}_{\text{inv}}(f_\tau)_{-\tau}$$

To obtain translation invariance at pixel precision for data with N samples, one should use a set of $|\Delta| = N$ translation vectors. For wavelets, this requires $O(N^2)$ operations.

Figure 7.16 applies cycle spinning to hard thresholding in an orthogonal wavelet basis, using the following shifts: $\Delta = \{0, 1/N_0, 2/N_0, 3/N_0\}^2$ on an $N_0 \times N_0$ image, corresponding to shifts by $\{0,1,2,3\}^2$ pixels. This requires 16 forward and inverse wavelet-transform pairs. Averaging this subset of shifts does not in general guarantee exact pixel-shift equivariance. In this example, partial averaging reduces sensitivity to shifts, substantially improves the SNR, and suppresses oscillatory artifacts. These artifacts move as τ changes, so averaging reduces them.

In Figure 7.17, translation-invariant hard thresholding slightly outperforms translation-invariant soft thresholding, reversing their ranking in the orthogonal-basis experiment.

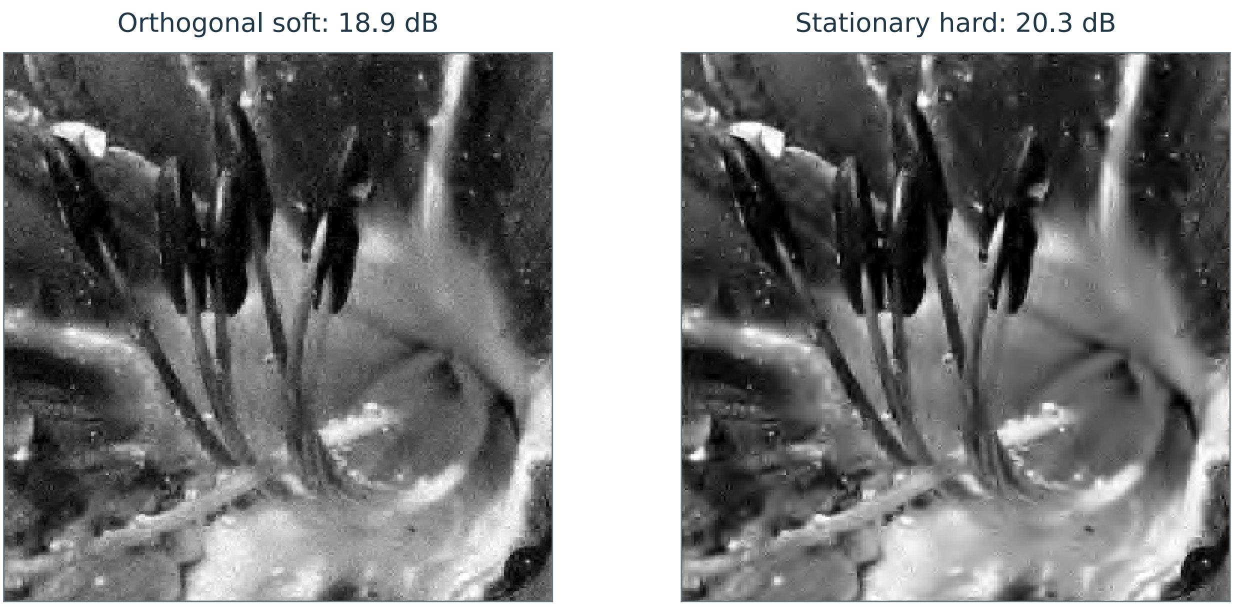

Figure 7.16. Soft thresholding in an orthogonal wavelet basis (left) and translation-invariant hard thresholding (right).

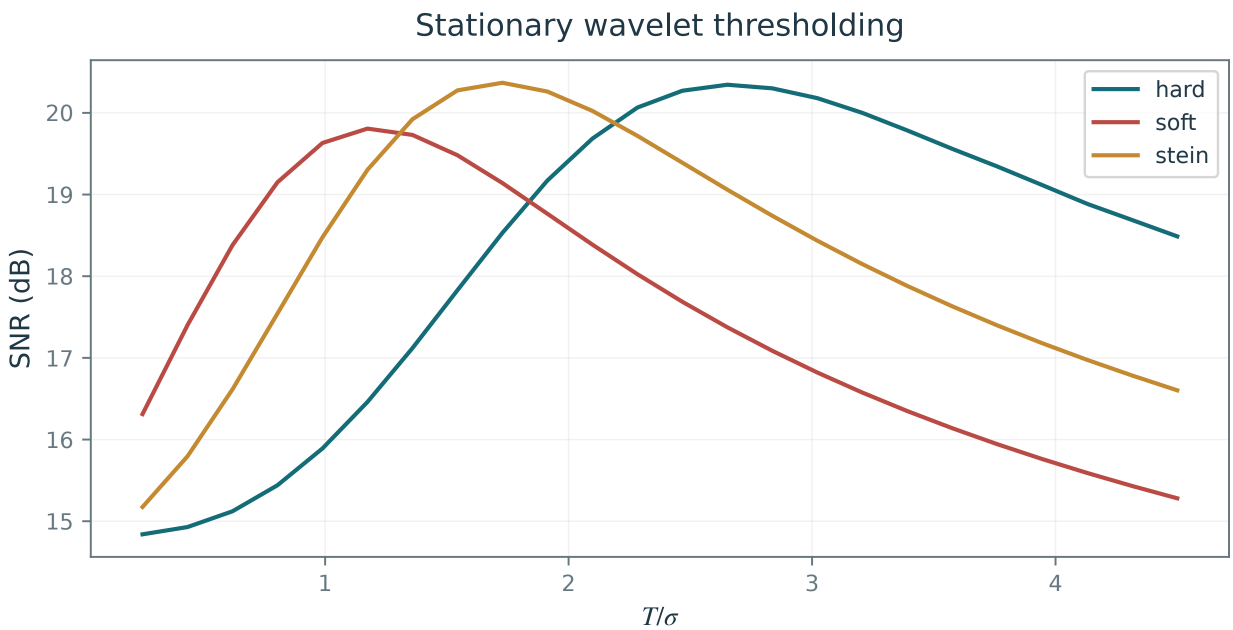

Figure 7.17. SNR as a function of T/σ for translation-invariant thresholding.

Translation invariant wavelet frame. Cycle spinning also has a frame interpretation: replace the orthogonal basis $\mathcal{B} = \{\psi_m\}$ by its translated copies,

$$\mathcal{B}_{\text{inv}} = \{(\psi_m)_\tau\}_{m,\tau\in\Delta}. \tag{7.12}$$

Although $\mathcal{B}_{\text{inv}}$ is no longer an orthogonal basis, orthonormality of each translated copy gives the energy and reconstruction identities

$$\|f\|^2 = \frac{1}{|\Delta|}\sum_{m,\tau\in\Delta} |\langle f, (\psi_m)_\tau\rangle|^2 \quad \text{and} \quad f = \frac{1}{|\Delta|}\sum_{m,\tau\in\Delta} \langle f, (\psi_m)_\tau\rangle(\psi_m)_\tau.$$

Such redundant families are called tight frames.

One can then define a translation invariant thresholding denoiser

$$\mathcal{D}_{\text{inv}}(f) = \frac{1}{|\Delta|}\sum_{m,\tau\in\Delta} S_T(\langle f, (\psi_m)_\tau\rangle)(\psi_m)_\tau. \tag{7.13}$$

This estimator equals the cycle spinning estimator defined in (7.11).

Counted with multiplicity, the translated frame has $|\Delta||\mathcal{B}|$ elements. For a basis of signals with N samples and a lattice of $|\Delta| = N$ shifts, this gives up to N^2 vectors in $\mathcal{B}_{\text{inv}}$. In a hierarchical construction such as a wavelet basis, however, different shifts can produce the same atom:

$$(\psi_m)_\tau = (\psi_{m'})_{\tau'} \quad \text{for} \quad m \neq m' \quad \text{and} \quad \tau \neq \tau',$$

The number of distinct vectors in $\mathcal{B}_{\text{inv}}$ can therefore be much smaller than N^2. For instance, for an orthogonal wavelet basis, one has

$$(\psi_{j,n})_{k2^j} = \psi_{j,n+k},$$

so many translated atoms coincide. For a signal of length N, an undecimated transform has $\log_2 N$ detail arrays of length N, plus a coarse array. For an $N_0 \times N_0$ image ($N = N_0^2$), it has $3\log_2 N_0$ detail arrays of N entries, plus a coarse array. Appropriate multiplicity weights must be retained when replacing the full translated family by distinct atoms.

The fast undecimated wavelet transform, often called the "à trous" transform, computes these coefficients in $O(N \log N)$ operations. Each detail array is indexed directly by every pixel location:

$$d_j^\omega[n] = \langle f, (\psi_{j,0}^\omega)_n\rangle, \qquad n \in \{0, \dots, N_0 - 1\}^2, \quad \omega \in \{V, H, D\},$$

where the wavelets and shifts are discrete and appropriately normalized. Each array is a band-pass filtered image, as illustrated in Figure 7.18.

Figure 7.19 illustrates hard thresholding of these translated coefficients.

7.3.5 Other Shrinkage Rules

Other shrinkage rules offer different compromises between retaining large coefficients and suppressing noise. We describe two examples whose performance depends on the coefficient distribution of f in the chosen basis.

Semi-soft thresholding. Semi-soft thresholding interpolates between hard and soft thresholding through a parameter $\mu > 1$:

$$S_T^\theta(x) = g_{\frac{1}{1-\theta}}(x) \quad \text{where} \quad g_\mu(x) = \begin{cases} 0 & \text{if} \quad |x| < T \\ x & \text{if} \quad |x| > \mu T \\ \operatorname{sign}(x)\frac{\mu(|x|-T)}{\mu-1} & \text{otherwise.} \end{cases}$$

For $0 < \theta < 1$, set $\mu = 1/(1-\theta)$. The limits $\theta \downarrow 0$ and $\theta \uparrow 1$ give hard and soft thresholding, respectively, away from the threshold boundary. Figure 7.20 shows an example of this nonlinearity.

Figure 7.21 shows an SNR improvement for a suitable choice of μ, although the visual difference from hard and soft thresholding is small.

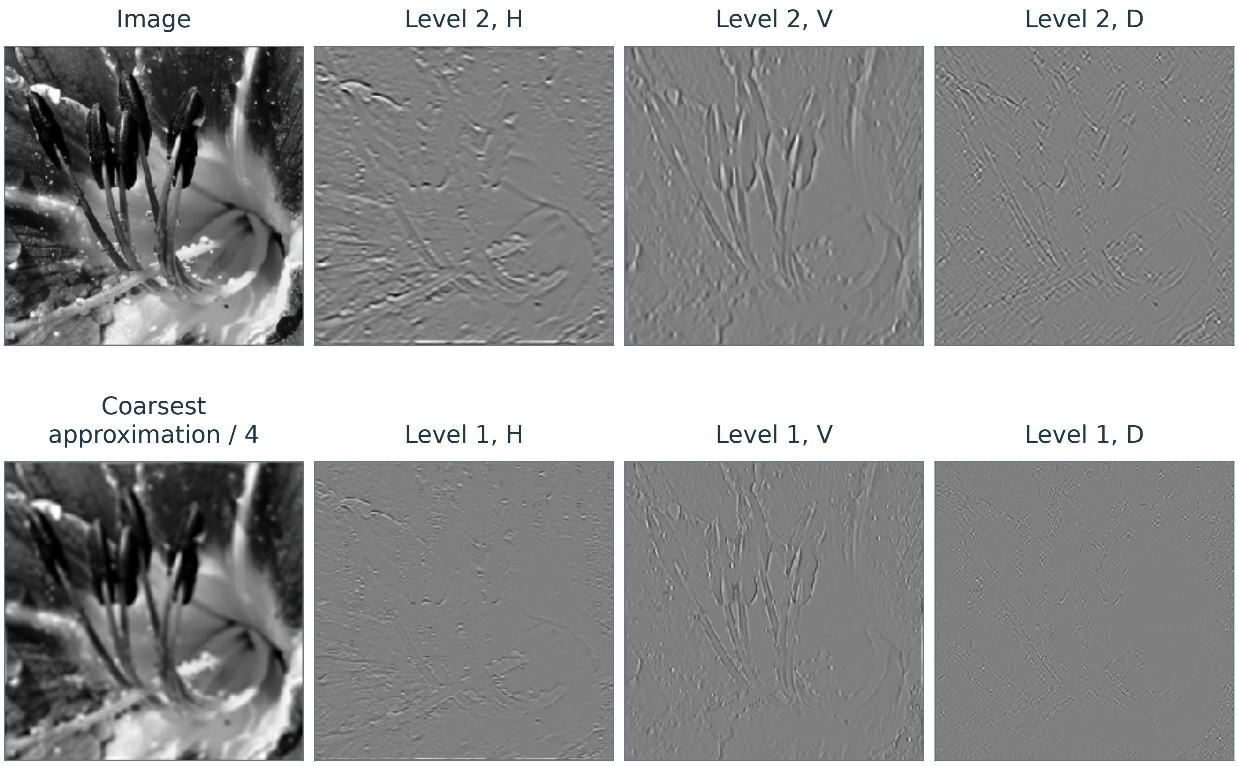

Figure 7.18. Translation invariant wavelet coefficients.

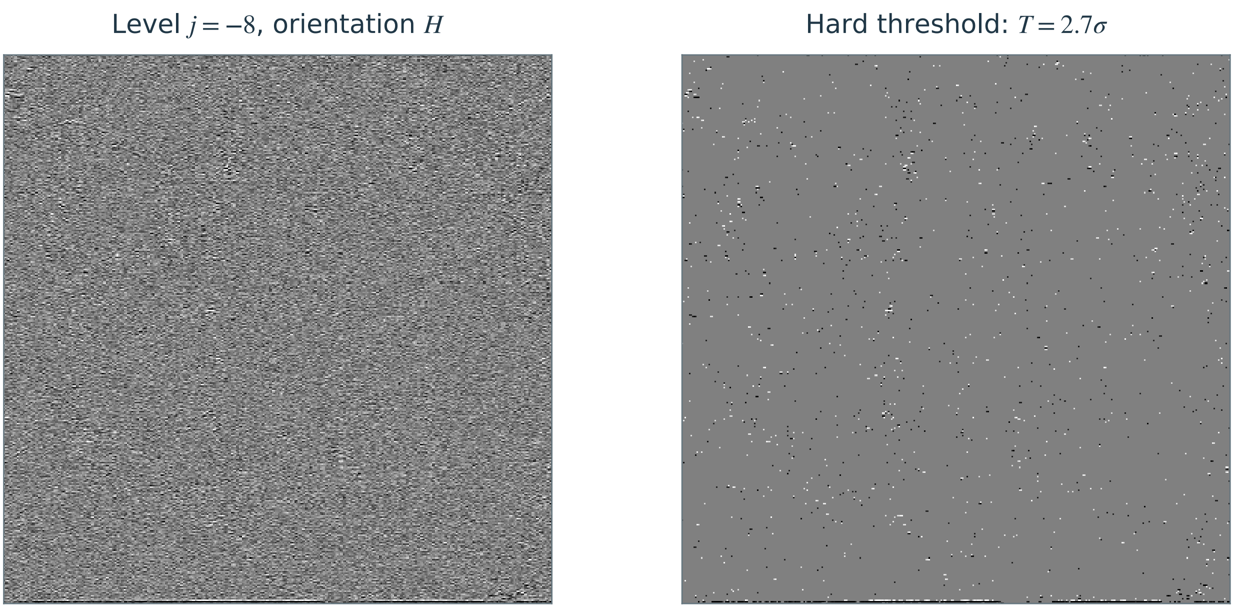

Figure 7.19. Left: translation invariant wavelet coefficients, for $j = -8, \omega = H$, right: thresholded coefficients.

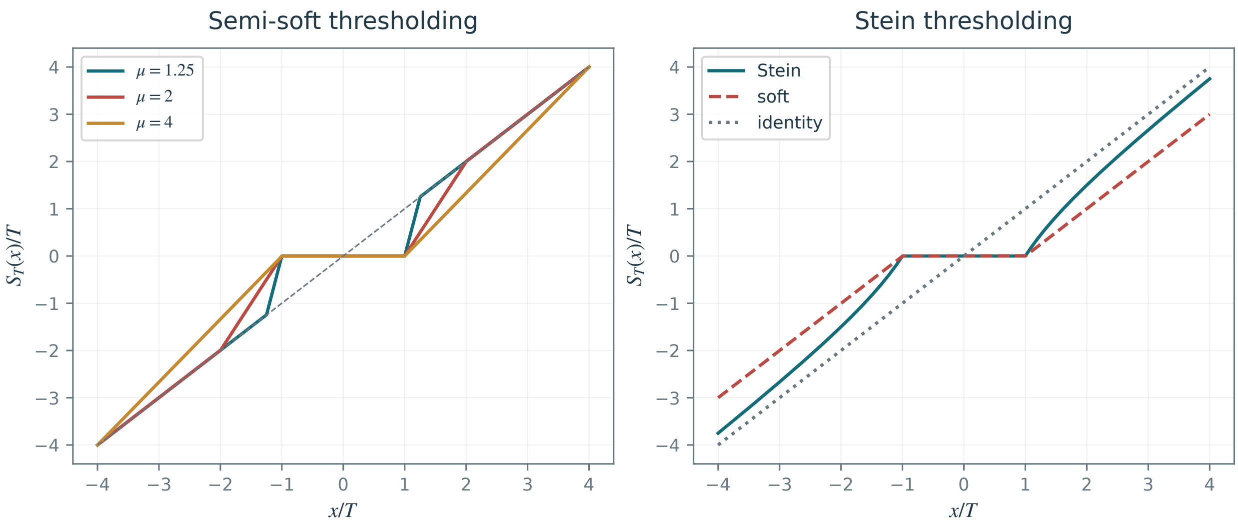

Figure 7.20. Left: semi-soft thresholder, right: Stein thresholder.

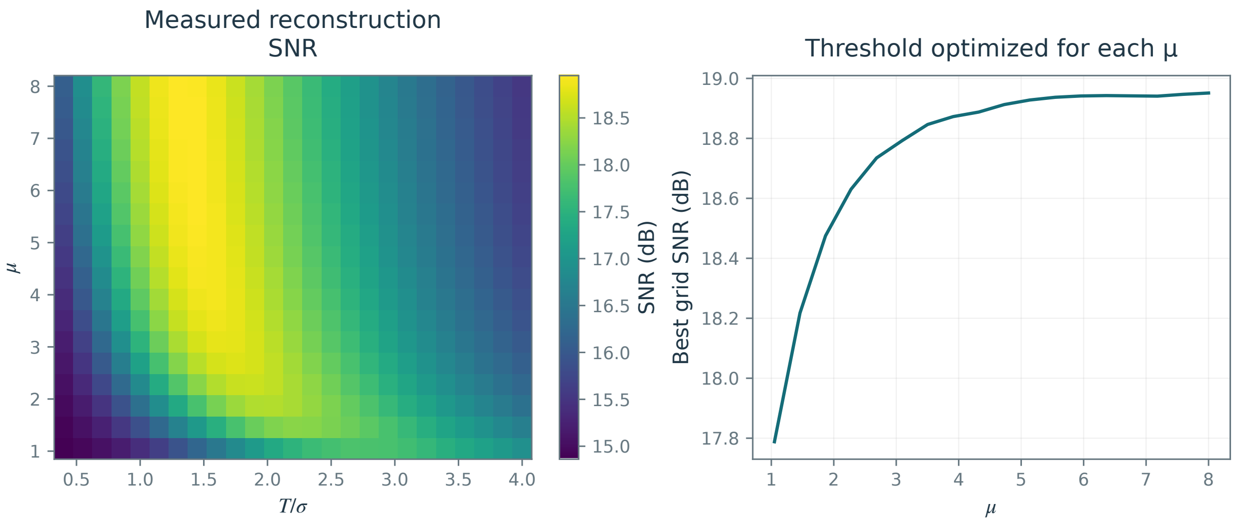

Figure 7.21. Left: SNR as a function of μ and T/σ. Right: SNR versus μ, with T/σ optimized separately for each μ.

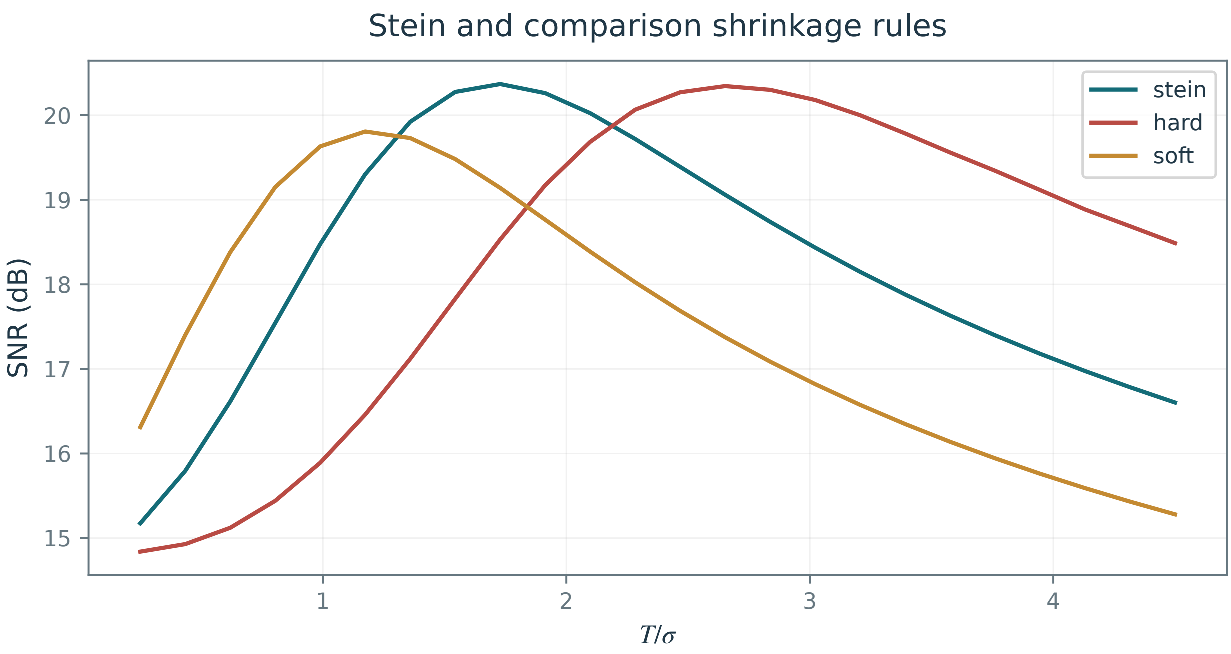

Figure 7.22. SNR as a function of T/σ for Stein thresholding.

Stein thresholding. Stein thresholding is defined using a quadratic attenuation of large coefficients

$$S_T^{\text{Stein}}(x) = \max\left(1 - \frac{T^2}{|x|^2}, 0\right) x.$$

The value at $x = 0$ is defined to be zero. This should be compared with the linear attenuation of soft thresholding

$$S_T^1(x) = \max\left(1 - \frac{T}{|x|}, 0\right) x.$$

For large coefficients, Stein thresholding has the following shrinkage behavior:

$$|S_T^{\text{Stein}}(x) - x| \longrightarrow 0 \quad \text{whereas} \quad |S_T^1(x) - x| \longrightarrow T,$$

as $x \longrightarrow \pm\infty$. Its deterministic shrinkage tends to zero, whereas soft thresholding retains a fixed shrinkage. This limiting property does not imply unbiased estimation at finite amplitudes.

Stein and hard thresholding give similar results in the translation-invariant natural-image experiments shown here.

7.3.6 Block Thresholding

The nonlinear thresholding methods above are diagonal estimators, since they attenuate each coefficient separately

$$\tilde{f} = \sum_m A_T^q(\langle f, \psi_m \rangle)\langle f, \psi_m \rangle \psi_m$$

where

$$A_T^q(x) = \begin{cases} \max(1 - T^2/|x|^2, 0) & \text{for} \quad q = \text{Stein} \\ \max(1 - T/|x|, 0) & \text{for} \quad q = 1 \text{ (soft)} \\ 1_{|x|>T} & \text{for} \quad q = 0 \text{ (hard)} \end{cases}$$

Define every attenuation factor to be zero at $x = 0$ when $T > 0$. Block thresholding shares an attenuation factor across a group of coefficients, exploiting their statistical dependence. This is useful for natural images, where edges create clusters of large wavelet coefficients. Grouping also helps suppress isolated noise coefficients in otherwise smooth regions.

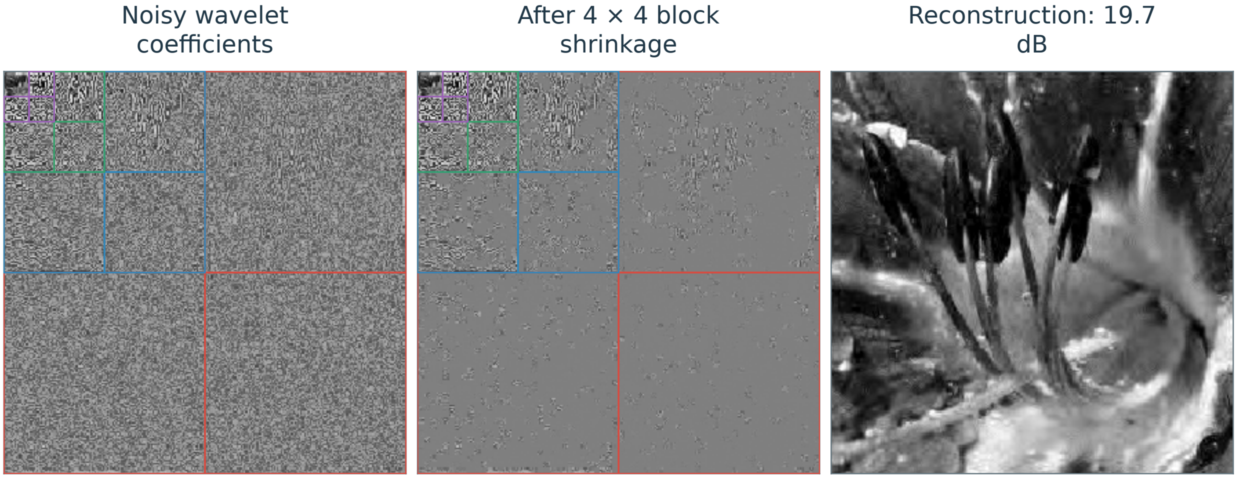

Figure 7.23. Left: wavelet coefficients, center: block thresholded coefficients, right: denoised image.

Partition the coefficients into disjoint blocks; for example, in a 2-D wavelet transform,

$$\{(j,n,\omega)\}_{j,n,\omega} = \bigcup_k B_k,$$

where each B_k contains $s \times s$ neighboring coefficients at a fixed scale and orientation. The block size s is a parameter of the method. Figure 7.23 shows an example of such a block.

The mean squared coefficient magnitude in a block is

$$E_k = \frac{1}{|B_k|} \sum_{m \in B_k} |\langle f, \psi_m \rangle|^2.$$

The block estimator

$$\tilde{f} = \sum_m S_T^{\text{block},q}(\langle f, \psi_m \rangle)\psi_m$$

applies the same attenuation to every coefficient in a block,

$$\forall\, m \in B_k, \quad S_T^{\text{block},q}(\langle f, \psi_m \rangle) = A_T^q(\sqrt{E_k})\langle f, \psi_m \rangle,$$

for $q \in \{0, 1, \text{stein}\}$. Figure 7.23 shows the attenuated blocks and the reconstructed image.

Figure 7.24, left, compares the rules for $q \in \{0, 1, \text{stein}\}$. Stein block thresholding gives the highest SNR in this comparison. Figure 7.24, right, compares block sizes for Stein thresholding. The choice $s = 4$ performs well in the displayed experiment.

Figure 7.25 compares the reconstructions. Stein block thresholding in an orthogonal wavelet basis approaches the quality of translation-invariant hard thresholding here, at lower computational cost. The block thresholding strategy can also be applied to wavelet coefficients in a translation invariant tight frame, which gives the best result among the methods compared in this experiment.

The same block attenuation can be implemented directly after the wavelet transform.

More advanced denoisers use richer statistical models to improve on the methods presented here; see, for instance, [2].

7.4 Signal-Dependent Noise

In many imaging systems, the variance of the noise affecting $f_{0,n}$ depends on the intensity $f_{0,n}$ itself. We now study Poisson counts and multiplicative noise. Both can be expressed as additive residuals, but their residual distributions depend on the signal.

7.4.1 Poisson Noise

Many imaging devices measure intensity by counting photons. Examples include digital cameras, confocal microscopy, PET and SPECT tomography.

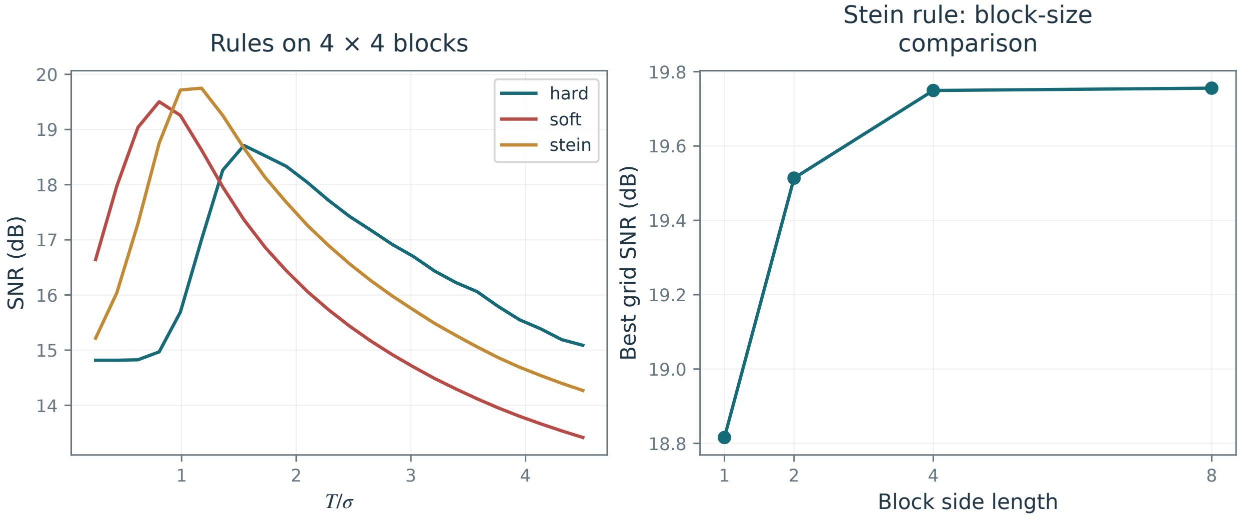

Figure 7.24. SNR as a function of T/σ (left) and comparison of different block sizes (right).

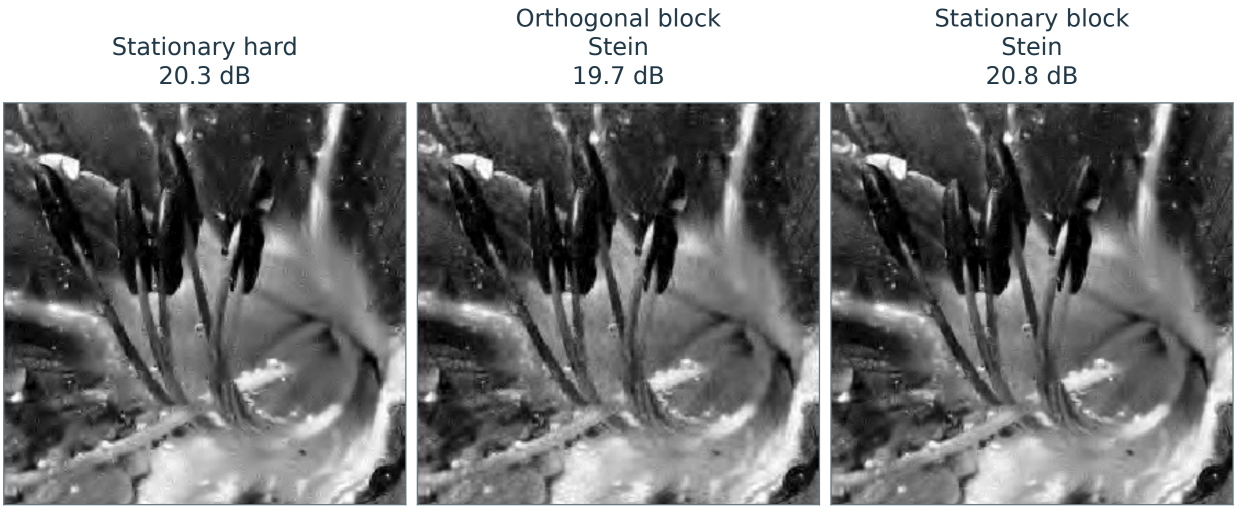

Figure 7.25. Left: translation invariant wavelet hard thresholding, center: block orthogonal Stein thresholding, right: block translation invariant Stein thresholding.

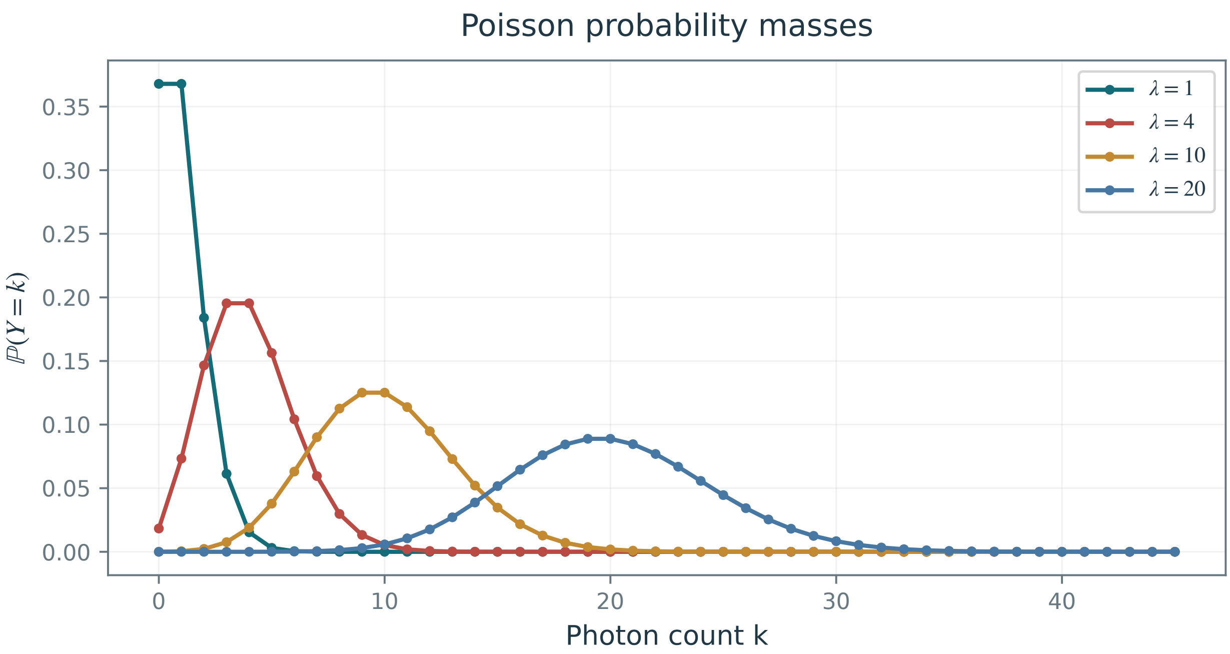

Figure 7.26. Poisson distributions for various λ.

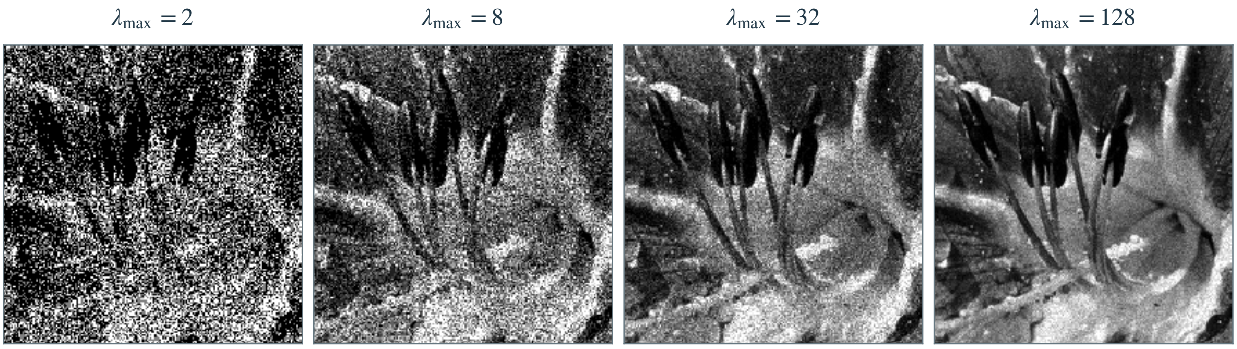

Figure 7.27. Poisson-corrupted flower images at increasing photon-count scales $\lambda_{\max}$.

Poisson model. For an unknown mean intensity $f_{0,n} \geqslant 0$, a Poisson model describes the observed photon count:

$$f_n \sim \mathcal{P}(\lambda) \quad \text{where} \quad \lambda = f_{0,n} \in [0, \infty),$$

The Poisson distribution $\mathcal{P}(\lambda)$ has probability mass function

$$\mathbb{P}(f_n = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

and its parameter varies across pixels with the underlying intensity. Figure 7.26 shows several Poisson distributions.

One has

$$\mathbb{E}(f_n) = \lambda = f_{0,n} \quad \text{and} \quad \mathrm{Var}(f_n) = \lambda = f_{0,n}$$

Thus denoising again estimates a mean from one observation, but the variance now varies with intensity. The absolute noise level increases as more photons reach the sensor. For positive intensity, the relative standard deviation $f_{0,n}^{-1/2}$ instead decreases.

Figure 7.27 shows Poisson-corrupted versions of a clean image f_0 at different maximum mean intensities $\lambda_{\max}$.

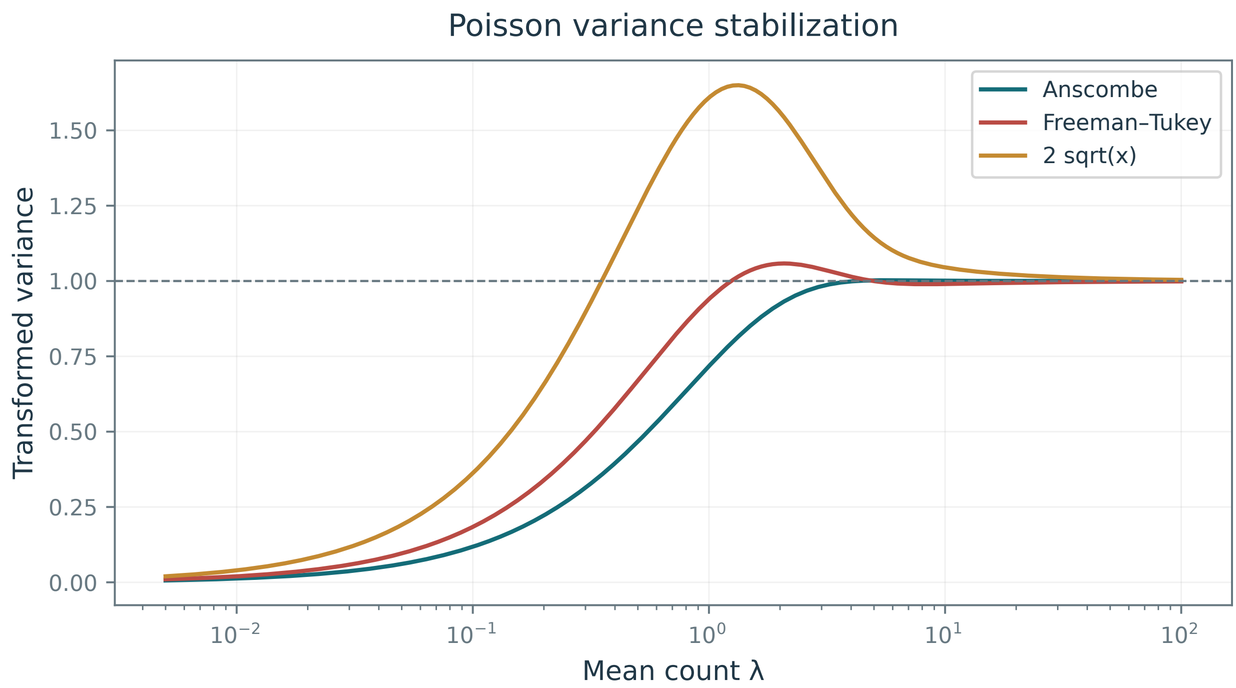

Figure 7.28. Exact Poisson variances of the Anscombe, Freeman–Tukey, and $2\sqrt{x}$ transforms as functions of the mean intensity.

Variance stabilization. Applying a thresholding estimator

$$\mathcal{D}(f) = \sum_m S_T^q(\langle f, \psi_m \rangle)\psi_m$$

directly to f can perform poorly because a single threshold T cannot adapt to spatially varying noise. A variance-stabilizing transform $\varphi : [0, +\infty) \to \mathbb{R}$, applied pixelwise, maps the counts to $\varphi(f)$. An additive Gaussian white-noise model can then be a useful approximation:

$$\varphi(f) \approx \varphi(f_0) + w \tag{7.14}$$

where $w_n \sim \mathcal{N}(0, \sigma^2)$ are independent centered Gaussians with constant variance σ^2.

The model (7.14) is approximate, and its accuracy depends on the intensity range of $f_{0,n}$. Two common variance-stabilizing transforms for Poisson noise are the Anscombe transform

$$\varphi(x) = 2\sqrt{x + 3/8}$$

and the Freeman–Tukey transform

$$\varphi(x) = \sqrt{x+1} + \sqrt{x}.$$

Figure 7.28 compares the variance of $\varphi(f)$ after these transformations.

A variance-stabilized denoiser is defined as

$$\mathcal{D}^{\mathrm{stab},q}(f) = \varphi^{-1}\Big(\sum_m S_T^q(\langle \varphi(f), \psi_m \rangle)\psi_m\Big)$$

where φ^{-1} is applied after projecting the denoised values onto its domain. The direct inverse generally introduces bias, especially at low counts; bias-corrected inverse transforms can improve the result.

At the moderate count levels in Figure 7.29, variance stabilization improves the denoising result.

7.4.2 Multiplicative Noise

Multiplicative image formation. A multiplicative noise model assumes that

$$f_n = f_{0,n} w_n$$

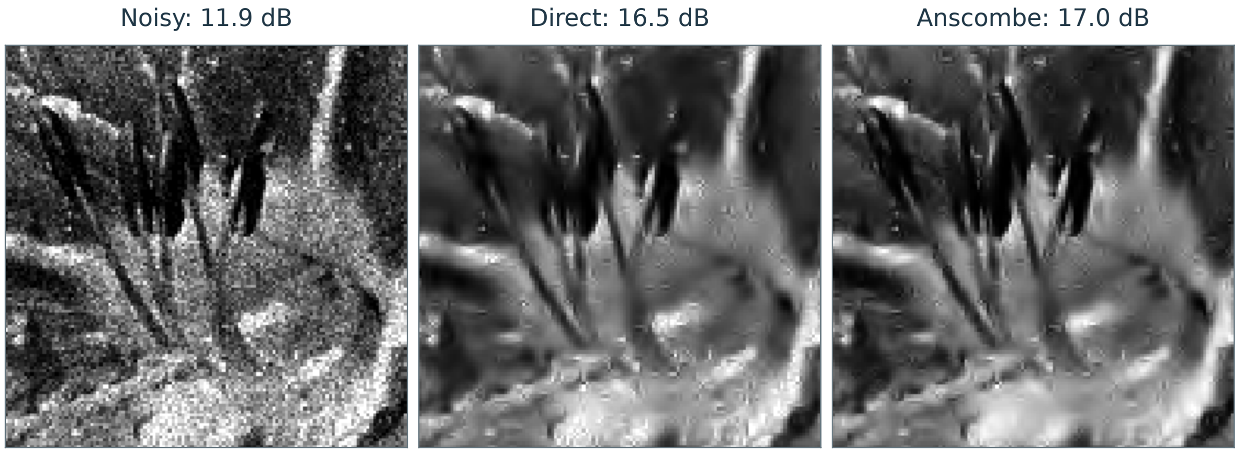

Figure 7.29. Left: noisy image, center: denoising without variance stabilization, right: denoising after variance stabilization.

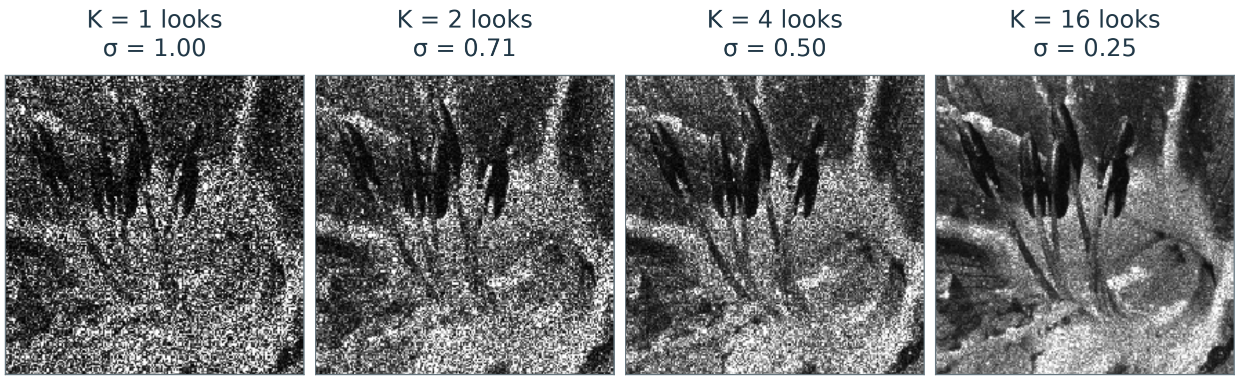

Figure 7.30. Multiplicative Gamma noise for increasing numbers K of averaged looks, with standard deviation $\sigma = K^{-1/2}$.

where w is a random multiplier, observed through one realization, with $\mathbb{E}(w_n) = 1$. Once again, the noise level depends on the pixel value

$$\mathbb{E}(f_n) = f_{0,n} \quad \text{and} \quad \mathrm{Var}(f_n) = f_{0,n}^2 \sigma^2 \quad \text{where} \quad \sigma^2 = \mathrm{Var}(w_n).$$

In SAR imaging, a common model forms f by averaging K independent observations, called looks:

$$\forall\, 1 \leqslant s \leqslant K, \quad f_n^{(s)} = f_{0,n} w_n^{(s)} + r_n^{(s)}$$

where $r^{(s)}$ is additive Gaussian white noise and the multiplier $w_n^{(s)}$ has an exponential distribution:

$$p_{w_n^{(s)}}(x) = e^{-x} \mathbf{1}_{x>0}.$$

Averaging K independent looks divides the additive-noise variance by K. When the remaining additive component is negligible, the average is approximately

$$f_n = \frac{1}{K} \sum_{s=1}^{K} f_n^{(s)} \approx f_{0,n} w_n$$

where the multiplicative factor is distributed according to a Gamma distribution

$$w_n \sim \mathrm{Gamma}(\text{shape} = K, \text{rate} = K), \qquad p_{w_n}(x) = \frac{K^K}{\Gamma(K)} x^{K-1} e^{-Kx} \mathbf{1}_{x>0}.$$

Its mean is one and its variance is $1/K$, so increasing K reduces the relative noise level.

Figure 7.30 illustrates how the noise changes with the number $K = 1/\sigma^2$ of averaged looks.

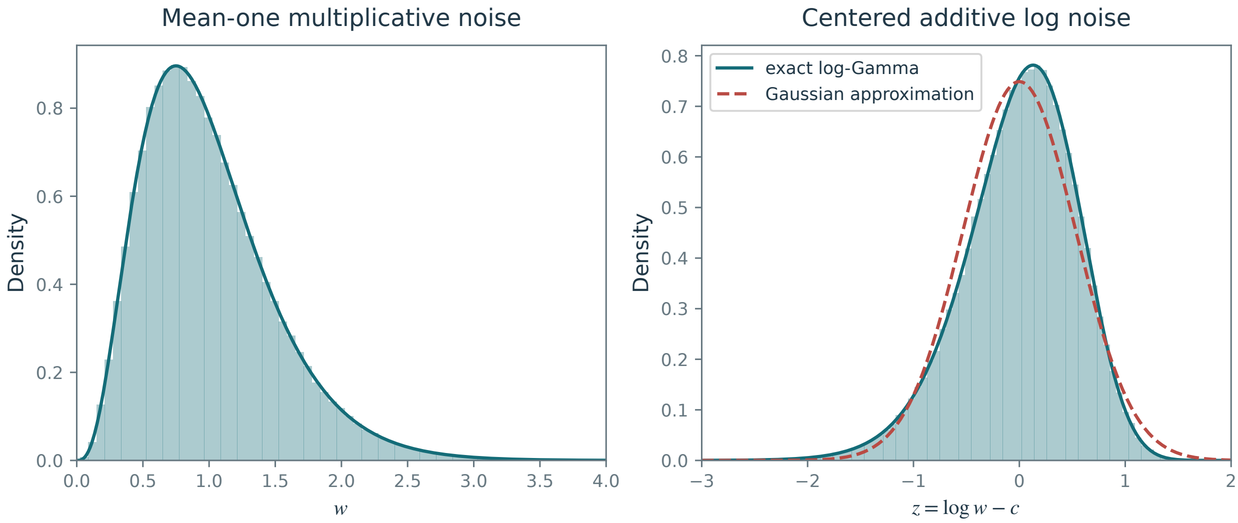

Figure 7.31. Histogram of multiplicative noise before (left) and after (right) stabilization.

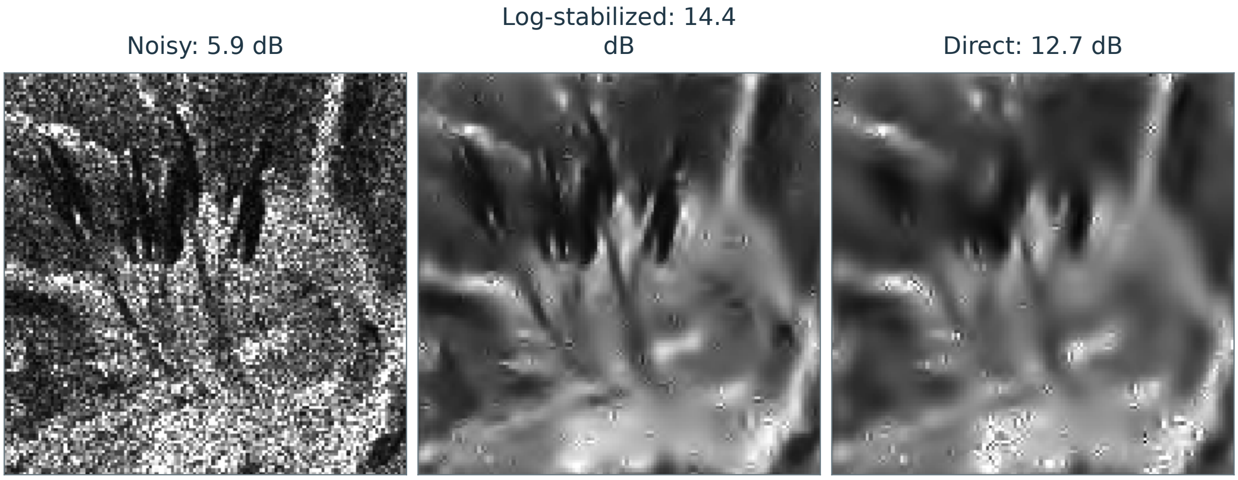

Figure 7.32. Left: noisy image, center: denoising after variance stabilization, right: denoising without variance stabilization.

A simple variance stabilization transform is

$$\varphi(x) = \log(x) - c$$

where

$$c = \mathrm{E}(\log(w)) = \psi(K) - \log(K) \quad \text{where} \quad \psi(x) = \Gamma'(x)/\Gamma(x)$$

Here Γ is the Gamma function and ψ is its logarithmic derivative, the digamma function. The transformed observation satisfies

$$\varphi(f)_n = \log f_{0,n} + z_n,$$

for strictly positive intensities, where $z_n = \log(w_n) - c$ is centered additive noise with variance $\psi^{(1)}(K)$; $\psi^{(1)}$ is the derivative of the digamma function. Thus the transformed target is $\log f_0$, not $\varphi(f_0)$. Exponentiating an estimate of $\log f_0$ introduces a further statistical bias and may require correction.

Figure 7.31 shows the effect of this variance stabilization on the distributions of w and z.

Figure 7.32 compares the two approaches at moderate noise levels σ, where variance stabilization improves the displayed result.

References

[1] D. Donoho and I. Johnstone. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81:425–455, Dec 1994.

[2] J. Portilla, V. Strela, M.J. Wainwright, and Simoncelli E.P. Image denoising using scale mixtures of Gaussians in the wavelet domain. *IEEE Trans. Image Proc.*, 12(11):1338–1351, November 2003.