Denoising

Removing noise from signals is possible only if some prior information is available. This information is encapsulated in an operator designed to reduce the noise while preserving the signal. Ideally, the joint probability distribution of the signal and the noise is known. Bayesian calculations then derive optimal operators which minimize the average estimation error. However, such probabilistic models are often not available for complex signals such as natural images.

Simpler signal models can be incorporated in the design of a basis or a frame, which takes advantage of known signal properties to build a sparse representation. Efficient non-linear estimators are then computed by thresholding the resulting coefficients. For one-dimensional signals and images, thresholding estimators are studied in wavelet bases, time-frequency representations, and curvelet frames. Block thresholdings are introduced to regularize these operators, which improves the estimation of audio recordings and images.

The optimality of estimators is analyzed in a minimax framework, where the maximum estimation error is minimized over a predefined set of signals. When signals are not uniformly regular, non-linear thresholding estimators in wavelet bases are shown to be much more efficient than linear estimators. They nearly reach the minimax risk over different signal classes, such as bounded variation signals and images.

11.1 Estimation with Additive Noise

Digital acquisition devices, such as cameras or microphones, output noisy measurements of an incoming analog signal \( \tilde{f}(x) \). These measurements can be modeled by a filtering of \( \tilde{f}(x) \) with the sensor responses \( \phi_n(x) \), to which is added a noise \( W[n] \):

\[
X[n] = \langle \tilde{f}, \phi_n \rangle + W[n] \quad \text{for } 0 \leq n < N. \tag{11.1}
\]

The noise \( W \) incorporates intrinsic physical fluctuations of the incoming signal. For example, an image intensity with low illumination has a random variation depending upon the number of photons captured by each sensor. It also includes noises introduced by the measurement device, such as electronic noises or transmission errors. The aggregated noise \( W \) is modeled by a random vector, whose probability distribution is supposed to be known a priori, and is often Gaussian.

Let us denote by \( f[n] = \langle \tilde{f}, \phi_n \rangle \) the discretized signal. This analog to digital acquisition is supposed to be stable, so that according to Section 3.1.3, an analog signal approximation of \( \tilde{f}(x) \) can be recovered from \( f[n] \) with a linear projector. One must then optimize an estimation \( \tilde{F}[n] = DX[n] \) of \( f[n] \) calculated from the noisy measurements (11.1) that are rewritten:

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

The decision operator \( D \) is designed to minimize the estimation error \( f - \tilde{F} \), measured by a loss function.
For audio signals or images, the loss function should measure the perceived audio or visual degradation. A mean-square distance is certainly not a perfect model of perceptual degradation, but it is mathematically simple and sufficiently precise in most applications. Throughout this chapter, the loss function is thus chosen to be a square Euclidean norm. The risk of the estimator \( \hat{F} \) of \( f \) is the average loss, calculated with respect to the probability distribution of the noise \( W \):

\[
r(D, f) = \mathbb{E}\{\|f - DX\|^2\}.
\]

(11.2)

The decision operator \( D \) is optimized with the prior information available on the signal. The Bayes framework supposes that signals are realizations of a random vector whose probability distribution is known, and a Bayes estimator minimizes the expected risk. A major difficulty is to acquire enough information to model this prior probability distribution. The minimax framework uses simpler deterministic models, which define signals as elements of a predefined set \( \Theta \). The expected risk can not be computed, but the maximum risk can be minimized over \( \Theta \). Section 11.1.2 relates minimax and Bayes estimators through the minimax theorem.

### 11.1.1 Bayes Estimation

A Bayesian model considers signals \( f \) are realizations of a random vector \( F \) whose probability distribution \( \pi \) is known a priori. This probability distribution is called the prior distribution. The noisy data are thus rewritten

\[
X[n] = F[n] + W[n] \quad \text{for } 0 \leq n < N.
\]

We suppose that noise and signal values \( W[k] \) and \( F[n] \) are independent for any \( 0 \leq k, n < N \). The joint distribution of \( F \) and \( W \) is thus the product of the distributions of \( F \) and \( W \). It specifies the conditional probability distribution of \( F \) given the observed data \( X \), also called the posterior distribution. This posterior distribution is used to optimize the decision operator \( D \) that computes an estimation \( \hat{F} = DX \) of \( F \) from the data \( X \).

The Bayes risk is the expected risk calculated with respect to the prior probability distribution \( \pi \) of the signal:

\[
r(D, \pi) = \mathbb{E}_\pi\{r(D, F)\}.
\]

By inserting (11.2), it can be rewritten as an expected value for the joint probability distribution of the signal and the noise:

\[
r(D, \pi) = \mathbb{E}\{\|F - \hat{F}\|^2\} = \sum_{n=0}^{N-1} \mathbb{E}\{\|F[n] - \hat{F}[n]\|^2\}.
\]

Let \( \mathcal{O}_n \) be the set of all operators (linear and non-linear) from \( \mathbb{C}^N \) to \( \mathbb{C}^N \). Optimizing \( D \) yields the minimum Bayes risk:

\[
r_n(\pi) = \inf_{D \in \mathcal{O}_n} r(D, \pi).
\]

The following theorem proves that there exist a Bayes decision operator \( D \) and a corresponding Bayes estimator \( \hat{F} \) that achieve this minimum risk.

**Theorem 11.1.** The Bayes estimator \( \hat{F} \) that yields the minimum Bayes risk \( r_n(\pi) \) is the conditional expectation

\[
\hat{F}[n] = \mathbb{E}\{F[n] \mid X[0], X[1], \ldots, X[N-1]\}.
\]

Proof. Let \( \pi_n(y) \) be the probability distribution of the value \( y \) of \( F[n] \). The minimum risk is obtained by finding \( \hat{F}[n] = D_n(X) \) that minimizes \( r(D_n, \pi_n) = \mathbb{E}\{\|F[n] - \hat{F}[n]\|^2\} \), for each \( 0 \leq n < N \). This risk depends on the conditional distribution \( P_n(x|y) \) of the data \( X = x \), given \( F[n] = y \):

\[
r(D_n, \pi_n) = \int (D_n(x) - y)^2 dP_n(x|y) d\pi_n(y).
\]

Let \( P(x) = \int P_n(x|y) d\pi_n(y) \) be the marginal distribution of \( X \) and \( \pi_n(y|x) \) be the posterior distribution of \( F[n] \) given \( X \). The Bayes formula gives

\[
r(D_n, \pi_n) = \int \left[ \int (D_n(x) - y)^2 d\pi_n(y|x) \right] dP(x).
\]
The double integral is minimized by minimizing the inside integral for each $x$. This quadratic form is minimum when its derivative vanishes:

$$
\frac{\partial}{\partial D_n(x)} \int (D_n(x) - y)^2 d\pi_n(y|x) = 2 \int (D_n(x) - y) d\pi_n(y|x) = 0
$$

which implies that

$$
D_n(x) = \int y d\pi_n(y|x) = \mathbb{E}\{F[n] \mid X = x\},
$$

so $D_n(X) = \mathbb{E}\{F[n] \mid X\}$. \hfill \blackslug

**Linear Estimation** The conditional expectation (11.3) is generally a complicated non-linear function of the data $\{X[k]\}_{0 \leq k < N}$, and is difficult to evaluate. To simplify this problem, we restrict the decision operator $D$ to be linear. Let $\mathcal{O}_1$ be the set of all linear operators from $\mathbb{C}^N$ to $\mathbb{C}^N$. The linear minimum Bayes risk is:

$$
r_l(\pi) = \inf_{D \in \mathcal{O}_1} r(D, \pi).
$$

The linear estimator $\tilde{F} = DX$ that achieves this minimum risk is called the Wiener estimator. The following theorem gives a necessary and sufficient condition that specifies this estimator. We suppose that $\mathbb{E}\{F[n]\} = 0$, which can be enforced by subtracting $\mathbb{E}\{F[n]\}$ from $X[n]$ to obtain a zero-mean signal.

**Theorem 11.2.** A linear estimator $\tilde{F}$ is a Wiener estimator if and only if

$$
\mathbb{E}\{(F[n] - \tilde{F}[n]) X[k]\} = 0 \quad \text{for} \ 0 \leq k, n < N. \quad (11.4)
$$

**Proof.** For each $0 \leq n < N$, we must find a linear estimation

$$
\tilde{F}[n] = D_n X = \sum_{k=0}^{N-1} h[n, k] X[k]
$$

which minimizes

$$
r(D_n, \pi_n) = \mathbb{E}\left\{\left( F[n] - \sum_{k=0}^{N-1} h[n, k] X[k] \right) \left( F[n] - \sum_{k=0}^{N-1} h[n, k] X[k] \right) \right\}. \quad (11.5)
$$

The minimum of this quadratic form is reached if and only if for each $0 \leq k < N$,

$$
\frac{\partial r(D_n, \pi_n)}{\partial h[n, k]} = -2 \mathbb{E}\left\{\left( F[n] - \sum_{l=0}^{N-1} h[n, l] X[l] \right) X[k] \right\} = 0,
$$

which verifies (11.4). \hfill \blackslug

If $F$ and $W$ are independent Gaussian random vectors, then the linear optimal estimator is also optimal among non-linear estimators. Indeed, two jointly Gaussian random vectors are independent if they are non-correlated [52]. Since $F[n] - \tilde{F}[n]$ is jointly Gaussian with $X[k]$, the non-correlation (11.4) implies that $F[n] - \tilde{F}[n]$ and $X[k]$ are independent for any $0 \leq k, n < N$. In this case, we can verify that $\tilde{F}$ is the Bayes estimator (11.3): $\tilde{F}[n] = \mathbb{E}\{F[n] \mid X\}$. The following theorem computes the Wiener estimator from the covariance $R_F$ and $R_W$ of the signal $F$ and of the noise $W$. The properties of covariance operators are described in Appendix A.6.

**Theorem 11.3 (Wiener).** If the signal $F$ and the noise $W$ are independent random vectors of covariance $R_F$ and $R_W$ then the linear Wiener estimator $\tilde{F} = DF$ which minimizes $\mathbb{E}\{\|F - \tilde{F}\|^2\}$ is

$$
\tilde{F} = R_F (R_F + R_W)^{-1} X. \quad (11.6)
$$
Proof. Let \( \tilde{F}[n] \) be a linear estimator of \( F[n] \):

\[
\tilde{F}[n] = \sum_{l=0}^{N-1} h[n, l] X[l].
\]  

(11.7)

This equation can be rewritten as a matrix multiplication by introducing the \( N \times N \) matrix \( H = (h[n, l])_{n, l < N} \):

\[
\tilde{F} = H X.
\]  

(11.8)

Theorem 11.2 proves that an optimal linear estimator satisfies the non-correlation condition (11.4), which implies that for \( 0 \leq n, k < N \)

\[
E\{F[n] X[k]\} = E\{\tilde{F}[n] X[k]\} = \sum_{l=0}^{N-1} h[n, l] E\{X[l] X[k]\}.
\]  

Since \( X[k] = F[k] + W[k] \) and \( E\{F[n] W[k]\} = 0 \), it results

\[
E\{F[n] F[k]\} = \sum_{l=0}^{N-1} h[n, l] \left( E\{F[l] F[k]\} + E\{W[l] W[k]\}\right). \]  

(11.9)

Let \( R_F \) and \( R_W \) be the covariance matrices of \( F \) and \( W \), whose entries are respectively \( E\{F[n] F[k]\} \) and \( E\{W[n] W[k]\} \). Equation (11.9) can be rewritten as a matrix equation:

\[
R_F = H (R_F + R_W).
\]

Inverting this equation gives

\[
H = R_F (R_F + R_W)^{-1}.
\]

\[\blacksquare\]

The optimal linear estimator (11.6) is simple to compute since it only depends upon second order covariance moments of the signal and of the noise.

**Estimation in a Karhunen-Loève Basis** Since a covariance operator is symmetric, it is diagonalized in an orthonormal basis which is called a *Karhunen-Loève basis* or a basis of principal components. If the covariance operators \( R_F \) and \( R_W \) are diagonal in the same Karhunen-Loève basis \( B = \{g_m\}_{0 \leq m < N} \) then the following corollary derives from Theorem 11.3 that the Wiener estimator is diagonal in this basis. We write

\[
X_B[m] = \langle X, g_m \rangle \quad F_B[m] = \langle F, g_m \rangle \quad \tilde{F}_B[m] = \langle \tilde{F}, g_m \rangle,
\]

\[
W_B[m] = \langle W, g_m \rangle \quad \sigma_B[m]^2 = E\{|\langle W, g_m \rangle|^2\}.
\]

**Corollary 11.1.** If there exists a Karhunen-Loève basis \( B = \{g_m\}_{0 \leq m < N} \) that diagonalizes the covariance matrices \( R_F \) and \( R_W \) of \( F \) and \( W \), then the Wiener estimator which minimizes \( E\{\|\tilde{F} - F\|^2\} \) is

\[
\tilde{F} = \sum_{m=0}^{N-1} \frac{E\{|F_B[m]|^2\}}{E\{|F_B[m]|^2\} + \sigma_B[m]^2} X_B[m] g_m.
\]  

(11.10)

The resulting minimum linear Bayes risk is

\[
r_B(\pi) = \sum_{m=0}^{N-1} \frac{E\{|F_B[m]|^2\} \sigma_B[m]^2}{E\{|F_B[m]|^2\} + \sigma_B[m]^2}.
\]  

(11.11)
Proof. The diagonal values of \( R_F \) and \( R_W \) are \( \langle R_F g_m, g_m \rangle = E \{ |F_B[m]|^2 \} \) and \( \langle R_W g_m, g_m \rangle = E \{ |W_B[m]|^2 \} = \sigma_B^2 [m] \). Since \( R_F \) and \( R_W \) are diagonal in \( B \), the linear operator \( R_F (R_F + R_W)^{-1} \) in (11.6) is also diagonal in \( B \), with diagonal values equal to \( E \{ |F_B[m]|^2 \} / (E \{ |F_B[m]|^2 \} + \sigma_B^2 [m])^{-1} \). So (11.6) proves that the Wiener estimator is

\[
\hat{F}_B = R_F (R_F + R_W)^{-1} X = \sum_{m=0}^{N-1} \frac{E \{ |F_B[m]|^2 \}}{E \{ |F_B[m]|^2 \} + \sigma_B^2 [m]} X_B[m] g_m . \tag{11.12}
\]

which proves (11.10).

The resulting risk is

\[
E \{ \| F - \hat{F} \|^2 \} = \sum_{m=0}^{N-1} E \{ |F_B[m] - \hat{F}_B[m]|^2 \} . \tag{11.13}
\]

Inserting (11.12) in (11.13) with \( X_B[m] = F_B[m] + W_B[m] \), where \( F_B[m] \) and \( W_B[m] \) are independent, yields (11.11).

This theorem proves that the Wiener estimator is implemented with a diagonal attenuation of each data coefficient \( X_B[m] \) by a factor that depends on the signal to noise ratio \( E \{ |F_B[m]|^2 \}/\sigma_B^2 [m] \) in the direction of \( g_m \). The smaller the signal to noise ratio, the more attenuation is required. If \( F \) and \( W \) are Gaussian processes, then the Wiener estimator is optimal among linear and non-linear estimators of \( F \).

If \( W \) is a white noise then its coefficients are uncorrelated with same variance:

\[
E \{ W[n] W[k] \} = \sigma^2 \delta[n - k] .
\]

Its covariance matrix is therefore \( R_W = \sigma^2 \text{Id} \). It is diagonal in all orthonormal bases and in particular in a Karhunen-Loève basis of \( F \). Theorem 11.1 can thus be applied with \( \sigma_B[m] = \sigma \) for 0 \( \leq m < N \).

**Frequency Filtering** Suppose that \( F \) and \( W \) are zero-mean, wide-sense circular stationary random vectors. The properties of such processes are reviewed in Appendix A.6. Their covariance satisfies

\[
\]

where \( R_F[n] \) and \( R_W[n] \) are \( N \) periodic. These matrices correspond to circular convolution operators and are therefore diagonal in the discrete Fourier basis

\[
\left\{ g_m[n] = \frac{1}{\sqrt{N}} \exp \left( \frac{i2m\pi n}{N} \right) \right\}_{0 \leq m < N} .
\]

The eigenvalues \( E \{ |F_B[m]|^2 \} \) and \( \sigma_B^2 [m] \) are the discrete Fourier transforms of \( R_F[n] \) and \( R_W[n] \), also called *power spectra*:

\[
E \{ |F_B[m]|^2 \} = \sum_{n=0}^{N-1} R_F[n] \exp \left( \frac{-i2m\pi n}{N} \right) = \hat{R}_F[m] ,
\]

\[
\sigma_B^2 [m] = \sum_{n=0}^{N-1} R_W[n] \exp \left( \frac{-i2m\pi n}{N} \right) = \hat{R}_W[m] .
\]

The Wiener estimator (11.10) is then a diagonal operator in the discrete Fourier basis, computed with the frequency filter:

\[
\hat{h}[m] = \frac{\hat{R}_F[m]}{\hat{R}_F[m] + \hat{R}_W[m]} . \tag{11.14}
\]

It is therefore a circular convolution:

\[
\hat{F}[n] = DX = X \circ \hat{h}[n] .
\]
Chapter 11. Denoising

Figure 11.1: (a): Realization of a Gaussian process $F$. (b): Noisy signal obtained by adding a Gaussian white noise (SNR = -0.48 db). (c): Wiener estimation $\hat{F}$ (SNR = 15.2 db).

The resulting risk is calculated with (11.11):

$$r_l(\pi) = \mathbb{E}\{\|F - \hat{F}\|^2\} = \sum_{m=0}^{N-1} \frac{\check{R}_F[m] \check{R}_W[m]}{\check{R}_F[m] + \check{R}_W[m]}.$$  \hfill (11.15)

The numerical value of the risk is often specified by the Signal to Noise Ratio, which is measured in decibels

$$SNR_{db} = 10 \log_{10} \left( \frac{\mathbb{E}\{\|F\|^2\}}{\mathbb{E}\{\|F - \hat{F}\|^2\}} \right).$$  \hfill (11.16)

Example 11.1. Figure 11.1(a) shows a realization of a Gaussian process $F$ obtained as a convolution of a Gaussian white noise $B$ of variance $\beta^2$ with a low-pass filter $g$:

$$F[n] = B \circ g[n],$$

with

$$g[n] = C \cos^2 \left( \frac{\pi n}{2K} \right) \mathbf{1}_{[-K,K]}[n].$$

Theorem A.7 proves that

$$\check{R}_F[m] = \check{R}_B[m] |\hat{g}[m]|^2 = \beta^2 |\hat{g}[m]|^2.$$

The noisy signal $X$ shown in Figure 11.1(b) is contaminated by a Gaussian white noise $W$ of variance $\sigma^2$, so $\check{R}_W[m] = \sigma^2$. The Wiener estimation $\hat{F}$ is calculated with the frequency filter (11.14)

$$\hat{h}[m] = \frac{\beta^2 |\hat{g}[m]|^2}{\beta^2 |\hat{g}[m]|^2 + \sigma^2}.$$  

This linear estimator is also an optimal non-linear estimator because $F$ and $W$ are jointly Gaussian random vectors.
11.1. Bayes Estimation

**Piecewise Regular** The limitations of linear estimators appear clearly for processes whose realizations are piecewise regular signals. A simple example is a random shift process $F$ constructed by translating randomly a piecewise regular signal $f[n]$ of zero mean, $\sum_{n=0}^{N-1} f[n] = 0$:

$$F[n] = f[(n - Q) \mod N].$$  \hspace{1cm} (11.17)

The shift $Q$ is an integer random variable whose probability distribution is uniform on $[0, N - 1]$. It is proved in (9.28) that $F$ is a circular wide-sense stationary process whose power spectrum is calculated in (9.29):

$$\hat{R}_F[m] = \frac{1}{N} |\hat{f}[m]|^2.$$  \hspace{1cm} (11.18)

![Figure 11.2: (a): Piecewise polynomial of degree 3. (b): Noisy signal degraded by a Gaussian white noise (SNR = 21.9 db). (c): Wiener estimation (SNR = 25.9 db).](image)

Figure 11.2 shows an example of a piecewise polynomial signal $f$ of degree $d = 3$ contaminated by a Gaussian white noise $W$ of variance $\sigma^2$. Assuming that we know $|\hat{f}[m]|^2$, the Wiener estimator $\hat{F}$ is calculated as a circular convolution with the filter whose transfer function is (11.14). This Wiener filter is a low-pass filter that averages the noisy data to attenuate the noise in regions where the realization of $F$ is regular, but this averaging is limited to avoid degrading the discontinuities too much. As a result, some noise is left in the smooth regions and the discontinuities are averaged a little. The risk calculated in (11.15) is normalized by the total noise energy $E\{||W||^2\} = N \sigma^2$:

$$\frac{r_1(\pi)}{N \sigma^2} \sim \sum_{m=0}^{N-1} \frac{N^{-1} |\hat{f}[m]|^2}{|\hat{f}[m]|^2 + N \sigma^2}.$$  \hspace{1cm} (11.19)

Suppose that $f$ has discontinuities of amplitude on the order of $C \geq \sigma$ and that the noise energy is not negligible: $N \sigma^2 \geq C^2$. Using the fact that $|\hat{f}[m]|$ decays typically like $CN^{-1/2}$, a direct calculation of the risk (11.19) gives

$$\frac{r_1(\pi)}{N \sigma^2} \sim \frac{C}{\sigma N^{1/2}}.$$  \hspace{1cm} (11.20)
The equivalence \( \sim \) means that upper and lower bounds of the left-hand side are obtained by multiplying the right-hand side by two constants \( A, B > 0 \) that are independent of \( C, \sigma \) and \( N \).

The estimation of \( F \) can be improved by non-linear operators, which average the data \( X \) over large domains where \( F \) is regular but do not make any averaging where \( F \) is discontinuous. Many estimators have been studied [261, 379], to recover the position of the discontinuities of \( f \) in order to adapt the data averaging. These algorithms have long remained ad hoc implementations of intuitively appealing ideas. Wavelet thresholding estimators perform such an adaptive smoothing and Section 11.5.3 proves that the normalized risk decays like \( N^{-1}(\log N)^2 \) as opposed to \( N^{-1/2} \) in (11.20).

### 11.1.2 Minimax Estimation

Although we may have some prior information, it is rare that we know the probability distribution of complex signals. Presently, there exists no stochastic model that takes into account the diversity of natural images. However, many images, such as the one in Figure 2.2, have some form of piecewise regularity, with a bounded total variation. Models are often defined over the original analog signal \( f \) that is measured with sensors having a response \( \tilde{\phi}_n \). The resulting discrete signal \( f[n] = (f, \tilde{\phi}_n) \) then belongs to a particular set \( \Theta \) in \( C^N \) derived from the analog model. This prior information defines a signal set \( \Theta \), but it does not specify the probability distribution of signals in \( \Theta \). The more prior information, the smaller the set \( \Theta \).

Knowing that \( f \in \Theta \), we want to estimate this signal from the noisy data

\[
X[n] = f[n] + W[n].
\]

The risk of an estimation \( \hat{F} = DX \) is \( r(D, f) = E\{\|DX - f\|^2\} \). The expected risk over \( \Theta \) cannot be computed because the probability distribution of signals in \( \Theta \) is unknown. To control the risk for any \( f \in \Theta \), we thus try to minimize the maximum risk:

\[
r(D, \Theta) = \sup_{f \in \Theta} E\{\|DX - f\|^2\}.
\]

The minimax risk is the lower bound computed over all linear and non-linear operators \( D \):

\[
r_n(\Theta) = \inf_{D \in \Omega_n} r(D, \Theta).
\]

In practice, we must find a decision operator \( D \) that is simple to implement and such that \( r(D, \Theta) \) is close to the minimax risk \( r_n(\Theta) \).

As a first step, as for Wiener estimators in the Bayes framework, the problem is simplified by restricting \( D \) to be a linear operator. The linear minimax risk over \( \Theta \) is the lower bound:

\[
r_l(\Theta) = \inf_{D \in \Omega_l} r(D, \Theta).
\]

This strategy is efficient only if \( r_l(\Theta) \) is of the same order as \( r_n(\Theta) \).

#### Bayes Priors

A Bayes estimator supposes that we know the prior probability distribution \( \pi \) of signals in \( \Theta \). If available, this supplement of information can only improve the signal estimation. The central result of game and decision theory shows that minimax estimations are Bayes estimations for a “least favorable” prior distribution.

Let \( F \) be the signal random vector, whose probability distribution is given by the prior \( \pi \). For a decision operator \( D \), the expected risk is \( r(D, \pi) = E_\pi \{r(D, F)\} \). The minimum Bayes risks for linear and non-linear operators are defined by:

\[
r_l(\pi) = \inf_{D \in \Omega_l} r(D, \pi) \quad \text{and} \quad r_n(\pi) = \inf_{D \in \Omega_n} r(D, \pi).
\]

Let \( \Theta^* \) be the set of all probability distributions of random vectors whose realizations are in \( \Theta \). The minimax theorem relates a minimax risk and the maximum Bayes risk calculated for priors in \( \Theta^* \).
11.1. Bayes Estimation

Theorem 11.4 (Minimax). For any subset \( \Theta \) of \( \mathbb{C}^N \)

\[
r_\tau(\Theta) = \sup_{\pi \in \Theta^*} r_\tau(\pi) \quad \text{and} \quad r_n(\Theta) = \sup_{\pi \in \Theta^*} r_n(\pi) .
\]  

(11.21)

Proof. For any \( \pi \in \Theta^* \)

\[
r(D, \pi) \leq r(D, \Theta)
\]

(11.22)

because \( r(D, \pi) \) is an average risk over realizations of \( F \) that are in \( \Theta \), whereas \( r(D, \Theta) \) is the maximum risk over \( \Theta \). Let \( \mathcal{O} \) be a convex set of operators (either \( \mathcal{O}_1 \) or \( \mathcal{O}_n \)). The inequality (11.22) implies that

\[
\sup_{\pi \in \Theta^*} r(\pi) = \sup_{\pi \in \Theta^*} \inf_{D \in \mathcal{O}} r(D, \pi) \leq \inf_{D \in \mathcal{O}} r(D, \Theta) = r(\Theta).
\]  

(11.23)

The main difficulty is to prove the reverse inequality: \( r(\Theta) \leq \sup_{\pi \in \Theta^*} r(\pi) \). When \( \Theta \) is a finite set, the proof gives a geometrical interpretation of the minimum Bayes risk and the minimax risk. The extension to an infinite set \( \Theta \) is sketched.

Suppose that \( \Theta = \{ f_i \}_{i \in \mathcal{P}} \) is a finite set of signals. We define a risk set:

\[
R = \{ (y_1, ..., y_p) \in \mathbb{C}^p : \exists D \in \mathcal{O} \text{ with } y_i = r(D, f_i) \text{ for } 1 \leq i \leq p \}.
\]

This set is convex in \( \mathbb{C}^p \) because \( \mathcal{O} \) is convex. We begin by giving geometrical interpretations to the Bayes risk and the minimax risk.

A prior \( \pi \in \Theta^* \) is a vector of discrete probabilities \( (\pi_1, ..., \pi_p) \) and

\[
r(\pi, D) = \sum_{i=1}^{p} \pi_i r(D, f_i).
\]  

(11.24)

The equation \( \sum_{i=1}^{p} \pi_i y_i = b \) defines a hyperplane \( \mathcal{P}_b \) in \( \mathbb{C}^p \). Computing \( r(\pi) = \inf_{D \in \mathcal{O}} r(D, \pi) \) is equivalent to finding the infimum \( b_0 = r(\pi) \) of all \( b \) for which \( \mathcal{P}_b \) intersects \( R \). The plane \( \mathcal{P}_{b_0} \) is tangent to \( R \) as shown in Figure 11.3.

The minimax risk \( r(\Theta) \) has a different geometrical interpretation. Let \( Q_c = \{ (y_1, ..., y_p) \in \mathbb{C}^p : y_i \leq c \} \) One can verify that \( r(\Theta) = \inf_{D \in \mathcal{O}} \sup_{f_i \in \mathcal{P}_b} r(D, f_i) \) is the infimum \( c_0 = r(\Theta) \) of all \( c \) such that \( Q_c \) intersects \( R \).

Figure 11.3: At the Bayes point, a hyperplane defined by the prior \( \pi \) is tangent to the risk set \( R \). The least favorable prior \( \tau \) defines a hyperplane that is tangential to \( R \) at the minimax point.

To prove that \( r(\Theta) \leq \sup_{\pi \in \Theta^*} r(\pi) \) we look for a prior distribution \( \tau \in \Theta^* \) such that \( r(\tau) = r(\Theta) \).

Let \( \tilde{Q}_{c_0} \) be the interior of \( Q_{c_0} \). Since \( \tilde{Q}_{c_0} \cap R = \emptyset \) and both \( \tilde{Q}_{c_0} \) and \( R \) are convex sets, the hyperplane separation theorem says that there exists a hyperplane of equation

\[
\sum_{i=1}^{p} \tau_i y_i = \tau \cdot y = b,
\]  

(11.25)

with \( \tau \cdot y \leq b \) for \( y \in \tilde{Q}_{c_0} \) and \( \tau \cdot y \geq b \) for \( y \in R \). Each \( \tau_i \geq 0 \), for if \( \tau_i < 0 \) then for \( y \in \tilde{Q}_{c_0} \) we obtain a contradiction by taking \( y_j \) to \(-\infty\) with the other coordinates being fixed. Indeed, \( \tau \cdot y \) goes to \(+\infty\).
The extension of this result to an infinite set of signals \( \Theta \) is done with a compactness argument. When \( \mathcal{O} = \mathcal{O}_1 \) or \( \mathcal{O} = \mathcal{O}_c \), for any prior \( \pi \in \Theta^* \) we know from Theorem 11.1 and Theorem 11.2 that \( \inf_{D \in \mathcal{O}} r(D, \pi) \) is reached by some Bayes decision operator \( D \in \mathcal{O} \). One can verify that there exists a subset of operators \( C \) that includes the Bayes operator for any prior \( \pi \in \Theta^* \), and such that \( C \) is compact for an appropriate topology. When \( \mathcal{O} = \mathcal{O}_1 \), one can choose \( C \) to be the set of linear operators of norm smaller than 1, which is compact because it belongs to a finite dimensional space of linear operators. Moreover, the risk \( r(f, D) \) can be shown to be continuous in this topology with respect to \( D \in C \).

Let \( c < r(\Theta) \). For any \( f \in \Theta \) we consider the set of operators \( \mathcal{S}_f = \{ D \in \mathcal{C} : r(D, f) > c \} \). The continuity of \( r \) implies that \( \mathcal{S}_f \) is an open set. For each \( D \in C \) there exists \( f \in \Theta \) such that \( D \in \mathcal{S}_f \), so \( C = \cup_{f \in \Theta} \mathcal{S}_f \). Since \( C \) is compact there exists a finite covering \( C = \cup_{i \in S} \mathcal{S}_{f_i} \). The minimax risk over \( \Theta_c = \{ f_i \}_{i \in S} \) satisfies

\[
r(\Theta_c) = \inf_{D \in C} \sup_{i \in S} r(D, f_i) \geq c.
\]

Since \( \Theta_c \) is a finite set, we proved that there exists \( \tau_c \in \Theta_c \subset \Theta^* \) such that \( r(\tau_c) = r(\Theta_c) \). But \( r(\Theta_c) \geq c \) so letting \( c \) go to \( r(\Theta) \) implies that \( \sup_{\pi \in \Theta^*} r(\pi) \geq r(\Theta) \). Together with (11.23) this shows that \( \inf_{\pi \in \Theta^*} r(\pi) = r(\Theta) \). □

A distribution \( \tau \in \Theta^* \) such that \( r(\tau) = \inf_{\pi \in \Theta^*} r(\pi) \) is called a least favorable prior distribution. The minimax theorem proves that the minimax risk is the minimum Bayes risk for a least favorable prior.

In signal processing, minimax calculations are often hidden behind apparently orthodox Bayes estimations. Let us consider an example involving images. It has been observed that histograms of the wavelet coefficients of “natural” images can be modeled with generalized Gaussian distributions [360, 439]. This means that natural images belong to a certain set \( \Theta \), but it does not specify a prior distribution over this set. To compensate for the lack of knowledge about the dependency of wavelet coefficients spatially and across scales, one may be tempted to create a “simple probabilistic model” where all wavelet coefficients are considered to be independent. This model is clearly simplistic since images have geometrical structures that create strong dependencies both spatially and across scales (see Figure 7.24). However, calculating a Bayes estimator with this inaccurate prior model may give valuable results when estimating images. Why? Because this “simple” prior is often close to a least favorable prior. The resulting estimator and risk are thus good approximations of the minimax optimum. If not chosen carefully, a “simple” prior may yield an optimistic risk evaluation that is not valid for real signals.

On the other hand, the minimax approach may seem very pessimistic since we always consider the maximum risk over \( \Theta \). It is sometimes the case. However, when the set \( \Theta \) is large, one can often verify that a “typical” signal of \( \Theta \) has a risk of the order of the maximum risk over \( \Theta \). For example, a minimax risk calculation for bounded variation signals gives a result whose order of magnitude remains valid as long as a bounded variation signal has a discontinuity. A minimax calculation amounts to isolate the class of signals that are the most difficult to estimate, and one can check if these signals are indeed typically encountered in an application. If it is not the case, then it indicates that the model specified by \( \Theta \) is not well adapted to this application.

### 11.2 Diagonal Estimation in a Basis

It is generally not possible to compute the optimal Bayes or minimax estimator that minimizes the risk among all possible operators. To manage this complexity, the most classical strategy limits the choice of operators among linear operators. This comes at a cost, because the minimum risk among linear estimators may be well above the minimum risk obtained with non-linear estimators.
11.2. Diagonal Estimation in a Basis

Figure 11.2 is an example where the linear Wiener estimation can be considerably improved with a non-linear averaging. This section studies a particular class of non-linear estimators that are diagonal in a basis $B$. If the basis $B$ defines a sparse signal representation, then such diagonal estimators are nearly optimal among all non-linear estimators.

Section 11.2.1 computes a lower bound for the risk when estimating an arbitrary signal $f$ with a diagonal operator. Donoho and Johnstone [220] made a fundamental breakthrough by showing that thresholding estimators have a risk that is close to this lower bound. The general properties of thresholding estimators are introduced in Sections 11.2.2 and 11.2.3.

11.2.1 Diagonal Estimation with Oracles

We consider estimators computed with a diagonal operator in an orthonormal basis $B = \{ g_m \}_{0 \leq m < N}$. Lower bounds for the risk are computed with “oracles,” which simplify the estimation by providing information about the signal that is normally not available. These lower bounds are closely related to errors when approximating signals from a few vectors selected in $B$.

The noisy data

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

(11.26)

is decomposed in $B$. We write

$$X_B[m] = \langle X, g_m \rangle, \quad f_B[m] = \langle f, g_m \rangle \quad \text{and} \quad W_B[m] = \langle W, g_m \rangle.$$ 

The inner product of (11.26) with $g_m$ gives

$$X_B[m] = f_B[m] + W_B[m].$$

We suppose that $W$ is a zero-mean white noise of variance $\sigma^2$, which means

$$\mathbb{E}\{W[n]W[k]\} = \sigma^2 \delta[n-k].$$

The noise coefficients

$$W_B[m] = \sum_{n=0}^{N-1} W[n] g_m^*[n]$$

also define a white noise of variance $\sigma^2$. Indeed,

$$\mathbb{E}\{W_B[m]W_B[p]\} = \sum_{n=0}^{N-1} \sum_{k=0}^{N-1} g_m[n]g_p[k] \mathbb{E}\{W[n]W[k]\} = \sigma^2 \delta[p-m].$$

Since the noise remains white in all bases, it does not influence the choice of basis.

A diagonal operator estimates independently each $f_B[m]$ by multiplying $X_B[m]$ by a factor $a_m(X_B[m])$. The resulting estimator is

$$\tilde{F} = DX = \sum_{m=0}^{N-1} a_m(X_B[m]) X_B[m] g_m.$$ 

(11.27)

The operator $D$ is linear when $a_m(X_B[m])$ is a constant independent of $X_B[m]$.

**Oracle Attenuation** For a given signal $f$ let us find the constant $a_m = a_m(X_B[m])$ that minimizes the risk $r(D, f)$ of the estimator (11.27):

$$r(D, f) = \mathbb{E}\{||f - \tilde{F}||^2\} = \sum_{m=0}^{N-1} \mathbb{E}\{|f_B[m] - a_m X_B[m]|^2\}.$$ 

(11.28)

We shall see that $|a_m| \leq 1$, which means that the diagonal operator $D$ should attenuate the noisy coefficients.
Since \( X_B = f_B + W_B \) and \( \mathbb{E}(|W_B[m]|^2) = \sigma^2 \), and since \( a_m \) is a constant that does not depend upon the noise, it follows that

\[
\mathbb{E}(|f_B[m] - X_B[m]a_m|^2) = |f_B[m]|^2 (1 - a_m)^2 + \sigma^2 a_m^2.
\]  
(11.29)

This risk is minimum for

\[
a_m = \frac{|f_B[m]|^2}{|f_B[m]|^2 + \sigma^2},
\]  
(11.30)
in which case

\[
r_{\text{inf}}(f) = \mathbb{E}\|f - \hat{f}\|^2 = \sum_{m=0}^{N-1} \frac{|f_B[m]|^2 \sigma^2}{|f_B[m]|^2 + \sigma^2}.
\]  
(11.31)

Observe that the attenuation factor \( a_m \) and the resulting risk has the same structure as the Wiener filter (11.10). However, the Wiener filter is a linear operator that depends upon expected signal to noise ratios that are constant values. This attenuation factor depends upon the unknown original signal to noise ratio \( |f_B[m]|^2/\sigma^2 \). Since \( |f_B[m]| \) is unknown, the attenuation factor \( a_m \) cannot be computed. It is considered as an oracle information. The resulting oracle risk \( r_{\text{inf}}(f) \) is a lower bound which is normally not reachable. However, Section 11.2.2 shows that one can get close to \( r_{\text{inf}}(f) \) with a simple thresholding.

**Oracle Projection** The analysis of diagonal estimators can be simplified by restricting \( a_m \in \{0, 1\} \). When \( a_m = 1 \), the estimator \( \hat{f} = DX \) selects the coefficient \( X_B[m] \), and it removes it if \( a_m = 0 \). The operator \( D \) is then an orthogonal projector on a selected subset of vectors of the basis \( B \).

The non-linear projector that minimizes the risk (11.29) is defined by

\[
a_m = \begin{cases} 
1 & \text{if } |f_B[m]| \geq \sigma \\
0 & \text{if } |f_B[m]| < \sigma.
\end{cases}
\]  
(11.32)

The resulting oracle projector is

\[
DX = \sum_{m \in \Lambda_\sigma} X_B[m] g_m \quad \text{with} \quad \Lambda_\sigma = \{0 \leq m < N : |f_B[m]| \geq \sigma\}.
\]  
(11.33)

It is an orthogonal projection on the set \( \{g_m\}_{m \in \Lambda_\sigma} \) of basis vectors that best approximate \( f \). This “oracle” projector cannot either be implemented because \( a_m \) and \( \Lambda_\sigma \) depend on \( f_B[m] \) instead of \( X_B[m] \). The resulting risk is computed with (11.29):

\[
r_{\text{pr}}(f) = \mathbb{E}\|f - \hat{f}\|^2 = \sum_{m=0}^{N-1} \min(|f_B[m]|^2, \sigma^2).
\]  
(11.34)

Since for any \((x, y) \in \mathbb{R}^2\)

\[
\min(x, y) \geq \frac{xy}{x + y} \geq \frac{1}{2} \min(x, y)
\]

the risk of the oracle projector (11.34) is of the same order as the risk of an oracle attenuation (11.31):

\[
r_{\text{pr}}(f) \geq r_{\text{inf}}(f) \geq \frac{1}{2} r_{\text{pr}}(f).
\]  
(11.35)

The risk of an oracle projector can also be related to the approximation error of \( f \) in the basis \( B \). Let \( M = |\Lambda_\sigma| \) be the number of coefficients such that \( |f_B[m]| \geq \sigma \). The best \( M \)-term approximation of \( f \) defined in Section 9.2.1 is the orthogonal projection on the \( M \) vectors \( \{g_m\}_{m \in \Lambda_\sigma} \) that yield the largest amplitude coefficients:

\[
f_M = \sum_{m \in \Lambda_\sigma} f_B[m] g_m.
\]
The non-linear oracle projector risk can be rewritten

\[ r_{pr}(f) = \sum_{m=0}^{N-1} \min(|f_B[m]|^2, \sigma^2) = \sum_{m \notin \Lambda_x} |f_B[m]|^2 + M \sigma^2 \]  

(11.36)

\[ = \varepsilon_n(M, f) + M \sigma^2 \]  

(11.37)

where

\[ \varepsilon_n(M, f) = \|f - f_M\|^2 = \sum_{m \notin \Lambda_x} |f_B[m]|^2 \]

is the best \(M\)-term approximation error. It is the bias produced by setting signal coefficients to zero, and \(M \sigma^2\) is the variance of the remaining noise on the \(M\) coefficients that are kept. The following theorem proves that when \(\sigma\) decreases, the decay of this risk is characterized the decay of the non-linear approximation error \(\varepsilon_n(M, f)\) as \(M\) increases.

**Theorem 11.5.** If \(\varepsilon_n(M, f) \leq C^2 M^{1-2s}\) with \(1 \leq C/\sigma \leq N^s\), then

\[ r_{pr}(f) \leq 3 C^{1/s} \sigma^{2-1/s}. \]  

(11.38)

**Proof.** Observe that

\[ r_{pr}(f) = \min_{0 \leq M \leq N} \left( \varepsilon_n(M, f) + M \sigma^2 \right) . \]  

(11.39)

Let \(M_0\) be defined by \(2M_0 \sigma^2 > C^2 M_0^{1-2s} \geq M_0 \sigma^2\). Since \(1 \leq C/\sigma \leq N^s\), necessarily \(1 \leq M_0 \leq N\). Inserting \(\varepsilon_n(M_0, f) \leq C^2 M_0^{1-2s}\) with \(s > 1/2\) and \(M_0 \leq C^{1/s} \sigma^{-1/2}\) yields

\[ r_{pr}(f) \leq \varepsilon_n(M_0, f) + M_0 \sigma^2 \leq 3 M_0 \sigma^2 \leq 3 C^{1/s} \sigma^{2-1/s}, \]  

(11.40)

which proves (11.38).

---

**Linear Projection.** Oracle estimators can not be implemented because \(a_m\) is a constant that depends on the unknown signal \(f\). Let us consider linear projectors obtained by setting \(a_m\) to be equal to 1 on the first \(M\) vectors and 0 otherwise:

\[ \tilde{F} = D_M X = \sum_{m=0}^{M-1} X_B[m] g_m . \]  

(11.41)

The risk (11.28) becomes

\[ r(D_M, f) = \sum_{m=M}^{N-1} |f_B[m]|^2 + M \sigma^2 = \varepsilon_l(M, f) + M \sigma^2, \]  

(11.42)

where \(\varepsilon_l(M, f)\) is the linear approximation error computed in (9.3). The two terms \(\varepsilon_l(M, f)\) and \(M \sigma^2\) are respectively the bias and the variance components of the estimator. To minimize \(r(D_M, f)\), the parameter \(M\) is adjusted so that the bias is of the same order as the variance. When the noise variance \(\sigma^2\) decreases, the following theorem proves that resulting risk depends upon the decay of \(\varepsilon_l(M, f)\) as \(M\) increases.

**Theorem 11.6.** If \(\varepsilon_l(M, f) \leq C^2 M^{1-2s}\) with \(1 \leq C/\sigma \leq N^s\) then

\[ r(D_{M_0}, f) \leq 3 C^{1/s} \sigma^{2-1/s} \text{ for } (C/(2\sigma))^{1/s} < M_0 \leq (C/\sigma)^{1/s}. \]  

(11.43)

**Proof.** As in (11.40), we verify that

\[ r(D_{M_0}, f) = \varepsilon_l(M_0, f) + M_0 \sigma^2 \leq 3 C^{1/s} \sigma^{2-1/s}, \]

for \((C/(2\sigma))^{1/s} < M_0 \leq (C/\sigma)^{1/s}\), which proves (11.43).
Theorems 11.5 and 11.6 prove that the performances of oracle projection estimators and optimized linear projectors depend respectively on the precision of non-linear and linear approximations in the basis $\mathcal{B}$. Having an approximation error that decreases quickly means that one can then construct a sparse and precise signal representation with only a few vectors in $\mathcal{B}$. Section 9.2 shows that non-linear approximations can be much more precise, in which case the risk of a non-linear oracle projection is much smaller than the risk of a linear projection. Next section shows that thresholding estimators are non-linear projection estimators whose risk can get close to the oracle projection risk.

### 11.2.2 Thresholding Estimation

In a basis $\mathcal{B} = \{g_m\}_{0 \leq m < N}$, a diagonal estimator of $f$ from $X = f + W$ can be written

$$\tilde{F} = DX = \sum_{m=0}^{N-1} a_m(X_\mathcal{B}[m]) X_\mathcal{B}[m] g_m.$$ \hspace{1cm} (11.44)

We suppose that $W$ is a Gaussian white noise of variance $\sigma^2$. When $a_m$ are thresholding functions, the risk of this estimator is shown to be close to the lower bounds obtained with oracle estimators.

**Hard thresholding** A hard thresholding estimator is implemented with

$$a_m(x) = \begin{cases} 1 & \text{if } |x| \geq T \\ 0 & \text{if } |x| < T \end{cases}.$$ \hspace{1cm} (11.45)

and can thus be rewritten

$$\tilde{F} = DX = \sum_{m \in \tilde{\Lambda}_T} X_\mathcal{B}[m] g_m \quad \text{with } \tilde{\Lambda}_T = \{0 \leq m < N : |X_\mathcal{B}[m]| \geq T \}.$$ \hspace{1cm} (11.46)

It is an orthogonal projection of $X$ on the set of basis vectors $\{g_m\}_{m \in \tilde{\Lambda}_T}$. This estimator can also be rewritten with a hard thresholding function

$$\tilde{F} = \sum_{m=0}^{N-1} \rho_T(X_\mathcal{B}[m]) g_m \quad \text{with } \rho_T(x) = \begin{cases} x & \text{if } |x| > T \\ 0 & \text{if } |x| \leq T \end{cases}.$$ \hspace{1cm} (11.47)

The risk of this thresholding is

$$r_{th}(f) = r(D, f) = \sum_{m=0}^{N-1} \mathbb{E}(|f_\mathcal{B}[m] - \rho_T(X_\mathcal{B}[m])|^2),$$

with $X_\mathcal{B}[m] = f_\mathcal{B}[m] + W_\mathcal{B}[m]$ and hence

$$|f_\mathcal{B}[m] - \rho_T(X_\mathcal{B}[m])|^2 = \begin{cases} |W_\mathcal{B}[m]|^2 & \text{if } |X_\mathcal{B}[m]| > T \\ |f_\mathcal{B}[m]|^2 & \text{if } |X_\mathcal{B}[m]| \leq T \end{cases}.$$  

Since a hard thresholding is a non-linear projector in the basis $\mathcal{B}$, the thresholding risk is larger than the risk (11.34) of an oracle projector:

$$r_{th}(f) \geq r_{pr}(f) = \sum_{m=0}^{N-1} \min(|f_\mathcal{B}[m]|^2, \sigma^2).$$

**Soft Thresholding** An oracle attenuation (11.30) yields a risk $r_{inf}(f)$ that is smaller than the risk $r_{pr}(f)$ of an oracle projection, by slightly decreasing the amplitude of all coefficients in order to reduce the added noise. A soft attenuation, although non-optimal, is implemented by

$$0 \leq a_m(x) = \max \left( 1 - \frac{T}{|x|}, 0 \right) \leq 1.$$ \hspace{1cm} (11.48)
Theorem 11.7

The resulting diagonal estimator \( \tilde{F} \) in (11.44) can be written as in (11.47) with a soft thresholding function, which decreases by \( T \) the amplitude of all noisy coefficients.

\[
\rho_T(x) = \begin{cases} 
  x - T & \text{if } x \geq T \\
  x + T & \text{if } x \leq -T \\
  0 & \text{if } |x| \leq T 
\end{cases} \quad (11.49)
\]

The threshold \( T \) is generally chosen so that there is a high probability that it is just above the maximum level of the noise coefficients \( |W_B[m]| \). Reducing by \( T \) the amplitude of all noisy coefficients thus ensures that the amplitude of an estimated coefficient is smaller than the amplitude of the original one:

\[
|\rho_T(X_B[m])| \leq |f_B[m]|. \quad (11.50)
\]

In a wavelet basis where large amplitude coefficients are created by sharp signal variations, this estimation restores a signal that is at least as regular as the original signal \( f \), without adding sharp transitions due to the noise.

Thresholding Risk

The following theorem \([220]\) proves that for an appropriate choice of \( T \), the risk of a thresholding is close to the risk of an oracle projector \( r_{\text{pr}}(f) = \sum_{m=0}^{N-1} \min(|f_B[m]|^2, \sigma^2) \). We denote by \( \mathcal{O}_q \) the set of all linear or non-linear operators that are diagonal in \( B \).

**Theorem 11.7** (Donoho, Johnstone). Let \( T = \sigma \sqrt{2 \log e N} \). The risk \( r_{\text{th}}(f) \) of a hard or a soft thresholding estimator satisfies for all \( N \geq 4 \)

\[
r_{\text{th}}(f) \leq (2 \log_e N + 1) \left( \sigma^2 + r_{\text{pr}}(f) \right). \quad (11.51)
\]

The factor \( 2 \log_e N \) is optimal among diagonal estimators in \( B \):

\[
\lim_{N \to +\infty} \inf_{D \in \mathcal{O}_q} \sup_{f \in \mathcal{C}^N} \frac{\mathbb{E}[\|f - \tilde{F}\|^2]}{\sigma^2 + r_{\text{pr}}(f)} \leq \frac{1}{2 \log_e N} = 1. \quad (11.52)
\]

**Proof.** The proof of (11.51) is given for a soft thresholding. For a hard thresholding, the proof is similar although slightly more complicated. For a threshold \( \lambda \), a soft thresholding is computed with

\[
\rho_\lambda(x) = (x - \lambda \text{sign}(x)) 1_{|x| > \lambda}.
\]

Let \( X \) be a Gaussian random variable of mean \( \mu \) and variance 1. The risk when estimating \( \mu \) with a soft thresholding of \( X \) is

\[
r(\lambda, \mu) = \mathbb{E}(|\rho_\lambda(X) - \mu|^2) = \mathbb{E}(|(X - \lambda \text{sign}(X)) 1_{|X| > \lambda} - \mu|^2) . \quad (11.53)
\]

If \( X \) has a variance \( \sigma^2 \) and a mean \( \mu \) then by considering \( \hat{X} = X/\sigma \) we verify that

\[
\mathbb{E}(|\rho_\lambda(X) - \mu|^2) = \sigma^2 r \left( \frac{\lambda}{\sigma}, \frac{\mu}{\sigma} \right).
\]

Since \( f_B[m] \) is a constant, \( X_B[m] = f_B[m] + W_B[m] \) is a Gaussian random variable of mean \( f_B[m] \) and variance \( \sigma^2 \). The risk of the soft thresholding estimator \( \tilde{F} \) with a threshold \( T \) is thus

\[
r_{\text{th}}(f) = \sigma^2 \sum_{m=0}^{N-1} r \left( \frac{T}{\sigma}, \frac{f_B[m]}{\sigma} \right). \quad (11.54)
\]

An upper bound of this risk is calculated with the following lemma.

**Lemma 11.1.** If \( \mu \geq 0 \) then

\[
r(\lambda, \mu) \leq r(\lambda, 0) + \min(1 + \lambda^2, \mu^2). \quad (11.55)
\]
To prove (11.55), we first verify that if $\mu \geq 0$ then
\[
0 \leq \frac{\partial r(\lambda, \mu)}{\partial \mu} = 2\mu \int_{-\lambda}^{\lambda} \phi(x) \, dx \leq 2\mu ,
\] (11.56)
where $\phi(x)$ is the normalized Gaussian probability density
\[
\phi(x) = \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{x^2}{2} \right).
\]
Indeed (11.53) shows that
\[
r(\lambda, \mu) = \mu^2 \int_{-\lambda}^{-\lambda+\mu} \phi(x) \, dx + \int_{-\lambda}^{+\lambda} (x - \lambda)^2 \phi(x) \, dx + \int_{-\infty}^{-\lambda+\mu} (x + \lambda)^2 \phi(x) \, dx.
\] (11.57)
We obtain (11.56) by differentiating with respect to $\mu$. Since $\int_{-\infty}^{+\infty} \phi(x) \, dx = f^+ \phi(x) \, dx = 1$ and $\frac{\partial r(\lambda, \mu)}{\partial \mu} \geq 0$, necessarily
\[
r(\lambda, \mu) \leq \lim_{\mu \to +\infty} r(\lambda, \mu) = 1 + \lambda^2.
\] (11.58)
Moreover, since $\frac{\partial r(\lambda, s)}{\partial s} \leq 2s$
\[
r(\lambda, \mu) - r(\lambda, 0) = \int_{0}^{\mu} \frac{\partial r(\lambda, s)}{\partial s} \, ds \leq \mu^2.
\] (11.59)
The inequality (11.55) of the lemma is finally derived from (11.58) and (11.59):
\[
r(\lambda, \mu) \leq \min(r(\lambda, 0) + \mu^2, 1 + \lambda^2) \leq r(\lambda, 0) + \min(1 + \lambda^2, \mu^2).
\]
By inserting the inequality (11.55) of the lemma in (11.54), we get
\[
r_{\inf}(f) \leq N\sigma^2 r \left( \frac{T}{\sigma}, 0 \right) + \sigma^2 \sum_{m=-\infty}^{\infty} \min \left( \frac{T^2 + \sigma^2}{\sigma^2}, \frac{|f_B[m]|^2}{\sigma^2} \right). 
\] (11.60)
The expression (11.57) shows that $r(\lambda, 0) = 2 \int_{0}^{\infty} x^2 \phi(x + \lambda) \, dx$. For $T = \sigma \sqrt{2 \log N}$ and $N \geq 4$, one can verify that
\[
N r \left( \frac{T}{\sigma}, 0 \right) \leq 2 \log N + 1.
\] (11.61)
Moreover,
\[
\sigma^2 \min \left( \frac{T^2 + \sigma^2}{\sigma^2}, \frac{|f_B[m]|^2}{\sigma^2} \right) = \min(2\sigma^2 \log N + \sigma^2, |f_B[m]|^2) \leq \left( 2 \log N + 1 \right) \min(\sigma^2, |f_B[m]|^2).
\] (11.62)
Inserting (11.61) and (11.62) in (11.60) proves (11.51).

Since the soft and hard thresholding estimators are particular instances of diagonal estimators, the inequality (11.51) implies that
\[
\lim_{N \to +\infty} \inf_{B \in \mathcal{D}_2} \sup_{f \in \mathcal{C}_N} \frac{E[|f - \hat{F}|^2]}{\sigma^2 + r_{pr}(f)} \leq \frac{1}{2 \log N} \leq 1.
\] (11.63)
To prove that the limit is equal to 1, for $N$ fixed we compute a lower bound by replacing the sup over all signals $f$ by an expected value over the distribution of a particular signal process $F$. The coefficients $F_B[m]$ are chosen to define a very sparse sequence. They are independent random variables having a high probability $1 - \alpha_N$ to be equal to 0 and a low probability $\alpha_N$ to be equal to a value $\mu_N$ that is on the order of $\sigma \sqrt{2 \log N}$, but smaller. By adjusting $\mu_N$ and $\alpha_N$, Donoho and Johnstone [220] prove that the Bayes estimator $\hat{F}$ of $F$ tends to zero as $N$ increases and they derive a lower bound of the left-hand side of (11.63) that tends to 1.

The upper bound (11.51) proves that the risk $r_{\inf}(f)$ of a thresholding estimator is at most $2 \log N$ times larger than the risk $r_{pr}(f)$ of an oracle projector. Moreover, (11.52) proves that the $2 \log N$ factor cannot be improved by any other diagonal estimator. For $r_{pr}(f)$ to be small, (11.37) shows that $f$ must be well approximated by a few vectors in $B$. One can verify [220] that the theorem remains valid if $r_{pr}(f)$ is replaced by the risk $r_{\inf}(f)$ of an oracle attenuation, which is smaller.
Choice of Threshold The threshold $T$ must be chosen just above the maximum level of the noise. Indeed, if $f = 0$ and thus $X_B = W_B$, then to ensure that $\tilde{F} \approx 0$ the noise coefficients $|W_B[m]|$ must have a high probability of being below $T$. However, if $f \neq 0$ then $T$ must not be too large, so that we do not set to zero too many coefficients such that $|f_B[m]| \geq \sigma$. Since $W_B$ is a vector of $N$ independent Gaussian random variables of variance $\sigma^2$, one can prove [8] that the maximum amplitude of the noise has a very high probability of being just below $T = \sigma \sqrt{2 \log_e N}$:

$$\lim_{N \to \infty} \Pr \left( T - \frac{\sigma \log_e \log_e N}{\log_e N} \leq \max_{0 \leq m < N} |W_B[m]| \leq T \right) = 1.$$  

(11.64)

This explains why the theorem chooses this value. That the threshold $T$ increases with $N$ may seem counter-intuitive. This is due to the tail of the Gaussian distribution, which creates larger and larger amplitude noise coefficients when the sample size increases. The threshold $T = \sigma \sqrt{2 \log_e N}$ is not optimal and in general a lower threshold reduces the risk.

A soft thresholding computed for $T = \sigma \sqrt{2 \log_e N}$ often produces a risk that is larger than with a hard thresholding. A soft thresholding reduces to nearly 0 the amplitude of coefficients just above $T$ or just below $-T$, whereas a hard thresholding leave them as is. To obtain nearly the same risk for a hard thresholding and a soft thresholding, it is often necessary to reduce by two the threshold of the soft thresholding. Section 11.2.3 explains how to adapt the threshold to the data $X$.

Upper-Bound Interpretation Despite the technicality of the proof, the factor $2 \log_e N$ of the upper bound (11.51) can be easily explained. The ideal coefficient selection (11.32) sets $X_B[m]$ to zero if and only if $|f_B[m]| \leq \sigma$, whereas a hard thresholding sets $X_B[m]$ to zero when $|X_B[m]| \leq T$. If $|f_B[m]| \leq \sigma$ then it is very likely that $|X_B[m]| \leq T$, because $T$ is above the noise level. In this case the hard thresholding sets $X_B[m]$ to zero as the oracle projector (11.32) does. If $|f_B[m]| \geq 2T$ then it is likely that $|X_B[m]| \geq T$ because $|W_B[m]| \leq T$. In this case the hard thresholding and the oracle projector retain $X_B[m]$.

The hard thresholding may behave differently from the ideal coefficient selection when $|f_B[m]|$ is on the order of $T$. The ideal selection yields a risk: $\min(\sigma^2, |f_B[m]|^2) = \sigma^2$. If we are unlucky and $|X_B[m]| \leq T$, then the thresholding sets $X_B[m]$ to zero, which produces a risk

$$|f_B[m]|^2 \sim T^2 = 2 \log_e N \sigma^2.$$  

In this worst case, the thresholding risk is $2 \log_e N$ times larger than the ideal selection risk. Since the proportion of coefficients $|f_B[m]|$ on the order of $T$ is often small, the ratio between the hard thresholding risk and the oracle projection risk is generally significantly smaller than $2 \log_e N$.

Colored Noise Thresholding estimators can be adapted when the noise $W$ is not white. We suppose that $E[W[m]] = 0$. Since $W$ is not white, $\sigma_B[m]^2 = E[|W_B[m]|^2]$ depends on each vector $g_m$ of the basis. As in (11.32) and (11.34), we verify that an oracle projector which keeps all coefficients such that $|f_B[m]| \geq \sigma_B[m]$ and sets to zero all others has a risk

$$r_{pr}(f) = \sum_{m=0}^{N-1} \min(|f_B[m]|^2, \sigma_B[m]^2).$$  

(11.65)

Any linear or non-linear projector in the basis $\mathcal{B}$ has a risk larger than $r_{pr}(f)$.

Since the noise variance depends on $m$, a thresholding estimator must vary the threshold $T_m$ as a function of $m$. Such a hard or soft thresholding estimator can be written

$$\tilde{F} = DX = \sum_{m=0}^{N-1} \rho_{T_m}(X_B[m]) g_m.$$  

(11.66)

The following theorem generalizes Theorem 11.7 to compute the thresholding risk $r_{th}(f) = E[\|f - \tilde{F}\|^2]$.
Chapter 11. Denoising

Theorem 11.8 (Donoho, Johnstone). Let $\hat{F}$ be a hard or soft thresholding estimator with

$$T_m = \sigma_B[m] \sqrt{2 \log e N} \quad \text{for } 0 \leq m < N.$$ 

Let $\hat{\sigma}^2 = N^{-1} \sum_{m=0}^{N-1} \sigma_B[m]^2$. For any $N \geq 4$

$$r_{th}(f) \leq (2 \log e N + 1) \left( \hat{\sigma}^2 + r_{pr}(f) \right). \quad (11.67)$$

The proof of (11.67) is identical to the proof of (11.51). The thresholds $T_m$ are chosen to be just above the amplitude of each noisy coefficient $W_B[m]$. Section 13.2 studies an application to the restoration of blurred signals.

Frame Thresholding Estimators The properties of thresholding estimators remain valid for non-orthogonal Riesz bases and frames. The redundancy of frames often produces a smaller risk than with an orthogonal basis, thanks to their redundancy. They are thus most often used in numerical applications.

Let us recall that $\{\phi_p\}_{0 \leq p < P}$ with $P \geq N$ is a frame of $C^N$ if there exists $0 < A \leq B$ such that

$$A \|f\|^2 \leq \sum_{p=0}^{P-1} |\langle f, \phi_p \rangle|^2 \leq B \|f\|^2.$$ 

When $P = N$, the frame is a Riesz basis, otherwise it is redundant. In the following we consider that all frame vectors are normalized $\|\phi_p\| = 1$. Theorem 5.2 then proves that $A \leq P/N \leq B$.

Theorem 5.5 in Section 5.1.2 proves that there exists a dual frame $\{\tilde{\phi}_p\}_{0 \leq p < P}$ such that

$$f = \sum_{p=0}^{P-1} \langle f, \phi_p \rangle \tilde{\phi}_p.$$ 

A signal $f$ can be estimated from noisy coefficients $X = f + W$ by thresholding its frame coefficients

$$\hat{F} = \sum_{p=0}^{P-1} \rho_T(\langle X, \phi_p \rangle) \tilde{\phi}_p. \quad (11.68)$$

The resulting risk is $r_{th}(f) = \mathbb{E}\{\|\hat{F} - f\|^2\}$. Let us write

$$r_{pr}(f) = \sum_{p=0}^{P-1} \min(|\langle f, \phi_p \rangle|^2, \sigma^2).$$

Using an oracle, we verify in Exercise 11.12 that $r_{th}(f) \geq r_{pr}(f)/B$. Moreover, for a threshold $T = \sigma \sqrt{2 \log e P}$, with the same derivation steps as in the proof of Theorem 11.8, one can prove that for any $P \geq 4$

$$r_{th}(f) \leq \frac{2 \log e P + 1}{A} \left( \sigma^2 + r_{pr}(f) \right). \quad (11.69)$$

This proves that thresholding estimators in frames behave like thresholding estimators in orthogonal bases. The threshold $T = \sigma \sqrt{2 \log e P}$ is a conservative upper bound that is too large in most numerical experiments. For a tight frame, $A = B = P/N$. The thresholding estimator then behaves as the average of $A$ estimators in $A$ orthogonal bases. This averaging often reduces the resulting risk.

11.2.3 Thresholding Improvements

The thresholding risk is often reduced by choosing a threshold smaller than $\sigma \sqrt{2 \log e N}$. A threshold adapted to the data is calculated by minimizing an estimation of the risk. Different thresholding functions are also considered, and an important improvement is introduced with a translation invariant thresholding algorithm.
11.2. Diagonal Estimation in a Basis

Sure Thresholds  To study the impact of the threshold on the risk, we denote by $r_{th}(f,T)$ the risk of a soft thresholding estimator calculated with a threshold $T$. An estimate of $r_{th}(f,T)$ is calculated from the noisy data $X$, and $T$ is optimized by minimizing the estimated risk.

To estimate the risk $r_{th}(f,T)$, observe that if $|X_B[m]| < T$ then the soft thresholding sets this coefficient to zero, which produces a risk equal to $|f_B[m]|^2$. Since

$$E\{|X_B[m]|^2\} = |f_B[m]|^2 + \sigma^2,$$

one can estimate $|f_B[m]|^2$ with $|X_B[m]|^2 - \sigma^2$. If $|X_B[m]| \geq T$, the soft thresholding subtracts $T$ from the amplitude of $X_B[m]$. The expected risk is the sum of the noise energy plus the bias introduced by the reduction of the amplitude of $X_B[m]$ by $T$. It is estimated by $\sigma^2 + T^2$. The resulting estimator of $r_{th}(f,T)$ is

$$Sure(X,T) = \sum_{m=0}^{N-1} C(X_B[m])$$

with

$$C(u) = \begin{cases} u^2 - \sigma^2 & \text{if } u \leq T \\ \sigma^2 + T^2 & \text{if } u > T. \end{cases}$$

The following theorem [221] proves that $Sure(X,T)$ is an unbiased risk estimator. It is called a Stein Unbiased Risk Estimator (SURE) [444].

**Theorem 11.9** (Donoho, Johnstone). For a soft thresholding, the risk estimator $Sure(X,T)$ is unbiased:

$$E\{Sure(X,T)\} = r_{th}(f,T).$$

**Proof.** A soft thresholding estimator performs a soft thresholding of each noisy coordinate. As in (11.54), we thus derive that the resulting risk is the sum of the soft thresholding risk for each coordinate

$$r_{th}(f,T) = E\{\|f - \tilde{F}\|^2\} = \sigma^2 \sum_{m=0}^{N-1} r(T, f_B[m], \sigma),$$

where $r(\lambda, \mu, \sigma)$ is the risk when estimating $\mu$ by soft thresholding a Gaussian random variable $X$ of mean $\mu$ and variance $\sigma^2$:

$$r(\lambda, \mu, \sigma) = E\{|\rho_{\lambda}(X) - \mu|^2\} = E\{|(X - \lambda \text{sign}(X) 1_{|X| > \lambda} - \mu|^2\}.$$

Let us rewrite

$$r(T, \mu, \sigma) = E\{(X - g(X) - \mu)^2\},$$

where $g(x) = T \text{sign}(x) + (x - T \text{sign}(x) 1_{|x| < T}$ is a weakly differentiable function (in the sense of distributions). This risk is calculated by the following Stein formula [444].

**Lemma 11.2** (Stein). Let $g(x)$ be a weakly differentiable function. If $X$ is a Gaussian random vector of mean $\mu$ and variance $\sigma^2$ then

$$E\{|X + g(X) - \mu|^2\} = \sigma^2 + E\{|g(X)|^2\} + 2\sigma^2 E\{g'(X)\}.$$

To prove this lemma, let us develop (11.76)

$$E\{|X + g(X) - \mu|^2\} = E\{|X - \mu|^2\} + E\{|g(X)|^2\} - 2E\{(X - \mu) g(X)\}.$$

The probability density of $X$ is the Gaussian $\phi_{\sigma}(y - \mu)$. The change of variable $x = y - \mu$ shows that

$$E\{(X - \mu) g(X)\} = \int_{-\infty}^{\infty} x g(x + \mu) \phi_{\sigma}(x) dx.$$

Since $x \phi_{\sigma}(x) = -\sigma^2 \phi_{\sigma}'(x)$, an integration by parts gives

$$E\{(X - \mu) g(X)\} = -\sigma^2 \int_{-\infty}^{\infty} g(x + \mu) \phi_{\sigma}'(x) dx$$

$$= \sigma^2 \int_{-\infty}^{\infty} g'(x) \phi_{\sigma}(x - \mu) dx = E\{g'(X)\}.$$
Inserting this result in (11.77) proves (11.76).

For the soft thresholding risk, \( r(f, T) = T \cdot \text{sign}(x) + (x - T \cdot \text{sign}(x)) \cdot 1_{|x| < T} \) and hence \( g'(x) = 1_{|x| < T} \). Using the fact that \( \mathbb{E}\{1_{|X| < T}\} = 1 \), the Stein unbiased risk formula (11.76) implies that

\[
    r(T, \mu, \sigma) = (\sigma^2 + T^2) \mathbb{E}\{1_{|X| < T}\} + \mathbb{E}\{(\|X\|^2 - \sigma^2) 1_{|X| < T}\} = \mathbb{E}\{C(\|X\|^2)\},
\]

where \( C(x) \) is defined in (11.71). Inserting this expression in (11.73) proves (11.72).

This results suggests choosing the threshold which minimizes the Sure estimator

\[
    \hat{T} = \arg\min_T \text{Sure}(X, T) .
\]

Although the estimator \( \text{Sure}(X, T) \) of \( r_{th}(f, T) \) is unbiased, its variance may induce errors leading to a threshold \( T \) that is too small. This happens if the signal energy is small relative to the noise energy: \( \|f\|^2 \ll \mathbb{E}\{\|W\|^2\} = N\sigma^2 \). In this case, one must impose \( T = \sigma \sqrt{2\log_2 N} \) in order to remove all the noise. Since \( \mathbb{E}\{\|X\|^2\} = \|f\|^2 + N\sigma^2 \), we estimate \( \|f\|^2 \) with \( \|X\|^2 - N\sigma^2 \) and compare this value with a minimum energy level \( \varepsilon_N = \sigma^2 N^{1/2}(\log_2 N)^{3/2} \). The resulting Sure threshold is

\[
    T = \begin{cases} 
        \sigma \sqrt{2\log_2 N} & \text{if } \|X\|^2 - N\sigma^2 \leq \varepsilon_N \\
        \frac{\sigma}{T} \|X\|^2 - N\sigma^2 > \varepsilon_N 
    \end{cases}
\]

(11.79)

Let \( \Theta \) be a signal set and \( \min_T r_{th}(\Theta) \) be the minimax risk of a soft thresholding obtained by optimizing the choice of \( T \) depending on \( \Theta \). Donoho and Johnstone [221] prove that the threshold computed empirically with (11.79) yields a risk \( r_{th}(\Theta) \) equal to \( \min_T r_{th}(\Theta) \) plus a corrective term that decreases rapidly when \( N \) increases, if \( \varepsilon = \sigma^2 N^{1/2}(\log_2 N)^{3/2} \).

Exercise 11.13 studies a similar risk estimator for hard thresholding. However, this risk estimator is biased. We cannot guarantee that the threshold that minimizes this estimated risk is nearly optimal for hard thresholding estimations.

Other Thresholdings and Masking Noise

Besides hard and soft thresholdings, other diagonal attenuation functions in (11.66) can improve a diagonal signal estimation. A whole family of diagonal attenuations is defined by

\[
    a_m(x) = \max\left(1 - \frac{T^\beta}{|x|^\beta}, 0\right) \quad \text{with} \quad \beta > 0 .
\]

For \( \beta = 1 \) it corresponds to the soft thresholding (11.48). When \( \beta \) tends to \( +\infty \) it yields the hard thresholding (11.45).

If \( \beta = 2 \) then \( a_m(x) \) is a James-Stein shrinkage [444]. It is intermediate between a hard and a soft attenuation. Since \( \mathbb{E}\{|X_B[m]|^2\} = |f_B[m]|^2 + \sigma^2 \), for \( T = \sigma \) the attenuation

\[
    a_m(X_B[m]) = \max\left(|X_B[m]|^2 - \sigma^2, 0\right) \frac{|X_B[m]|^2 - \sigma^2}{|X_B[m]|^2}
\]

can be interpreted as an empirical estimation of the oracle attenuation factor \( a_m = |f_B[m]|^2/(|f_B[m]|^2 + \sigma^2) \). Is also called an empirical Wiener attenuation.

Thresholding signal coefficients can introduce perceptual artifacts which reduce the perceived quality of the estimation. Next section shows that in wavelet bases, thresholding noisy image coefficient removes fine texture and can produce cartoon-like images with no textures. Other artifacts can be created by thresholding estimators. Leaving some noise reduces our perceptual sensitivity to these artifacts and can thus improve the perceived signal quality, although it may increase the mean-square norm of the error. It is implemented with attenuation factors that remain strictly positive:

\[
    a_m(x) = \max\left(1 - \frac{T^\beta}{|x|^\beta}, \varepsilon\right) \quad \text{with} \quad \varepsilon > 0 .
\]

(11.80)

If \( |X_B[m]| < T \) then \( a_m(X_B[m]) = \varepsilon \) so this thresholding leaves a reduced noise of variance \( \varepsilon^2 \sigma^2 \), which masks potential artifacts.
11.3 Thresholding Sparse Representations

Translation Invariant Thresholding

In many applications, signal models are translation invariant. This is often the case for audio signals, where the recording beginning may be arbitrarily shifted, or for images which are translated by changing the camera position. For stochastic signal models with random processes, translation invariance means that the process is stationary. For a deterministic model that specifies a set \( \Theta \) where the signal belongs, it means that any \( f \in \Theta \) remains in \( \Theta \) after a translation. For signals embedded in additive noise, if the noise is stationary and \( \Theta \) is translation invariant then the minimization of the maximum risk over \( \Theta \) is achieved with translation invariant estimators. Theorem 11.12 proves this result for linear minimax estimators, which is also valid for non-linear estimators.

Coifman and Donoho [178] have introduced translation invariant thresholding estimators, which reduce the risk for translation invariant sets \( \Theta \). For signals of finite length \( N \), to avoid boundary issues, we consider circular translations modulo \( N \): \( f_p[n] = f[(n - p) \mod N] \). Observe that if \( B = \{g_m\}_{0 \leq m < N} \) is an orthonormal basis of \( \mathbb{C}^N \), then the translated basis \( B_p = \{g_{p,m}[n] = g_m[(n - p) \mod N]\}_{0 \leq m < N} \) is also an orthogonal basis of \( \mathbb{C}^N \), for any \( 0 \leq p < N \). If there is no translation information on the signal, all the bases \( \{B_p\}_{0 \leq p < N} \) are a priori equivalent for a thresholding estimations. Coifman and Donoho [178] thus proposed to average the thresholding estimations obtained in these \( N \) bases. This is equivalent to decompose the signal in a translation invariant dictionary that is a union of these \( N \) translated orthonormal bases

\[
\mathcal{D} = \bigcup_{p=0}^{N-1} B_p = \{g_{p,m}\}_{0 \leq m, p < N}.
\]  

(11.81)

The energy conservation in each orthogonal basis \( B_p \) implies a global energy conservation over the \( N^2 \) dictionary vectors

\[
\|f\|^2 = \frac{1}{N} \sum_{m=0}^{N-1} \sum_{p=0}^{N-1} |\langle f, g_{p,m} \rangle|^2,
\]

which proves that this dictionary is a tight frame.

The resulting translation invariant estimator of \( f \) from noisy data \( X = f + W \) is obtained by thresholding the translation invariant tight frame coefficients of \( X \):

\[
\hat{F}[n] = \frac{1}{N} \sum_{p=0}^{N-1} \sum_{m=0}^{N-1} \rho_T(\langle X, g_{p,m} \rangle) \ g_{p,m}[n],
\]

(11.82)

where \( \rho_T \) is a hard or a soft thresholding operator. Since this estimator is obtained by averaging \( N \) translation invariant estimators in orthogonal bases, the resulting thresholding risk satisfies the same upper bound as in Theorem 11.7.

A priori, this translation invariant thresholding requires \( N \) times more operation than a thresholding estimation in the original basis \( B \). However, this is not the case when the original basis \( B \) includes translated vectors. In this case the translation invariant dictionary \( D \) has less than \( N^2 \) different vectors. For example, a wavelet orthogonal basis yields a translation invariant dyadic wavelet dictionary which has only \( N \log_2 N \) different wavelets.

Translation invariant tight frames are not necessarily derived from an orthogonal basis. Theorem 5.12 in Section 5.1.5 proves that a translation invariant dictionary obtained by translating \( Q \) generators \( \{g_q\}_{0 \leq q < Q} \) is a tight frame if and only if their discrete Fourier transforms satisfy

\[
\sum_{q=0}^{Q-1} |\hat{g}_q[k]|^2 = Q
\]

for all \( 0 \leq k < N \). The simplicity of this condition offers more flexibility to build translation invariant thresholding estimators then from orthogonal bases. 

11.3 Thresholding Sparse Representations

Thresholding estimators are particularly efficient in a basis which can precisely approximate signals with few non-zero coefficients. The basis must therefore be chosen from prior information on signal properties, to obtain sparse representations.

Wavelet bases are particularly efficient to estimate piecewise regular signals. Noise removal from images is studied in Section 11.3.2, with wavelet bases and curvelet frames. For audio signals, sparse representations are obtained with localized time-frequency transforms. An important limitation of diagonal thresholding operators is illustrated in Section 11.3.3, with the creation of a “musical noise” when thresholding windowed Fourier coefficients for audio noise removal.
11.3.1 Wavelet Thresholding

Thresholding wavelet coefficients implements an adaptive signal averaging, with a kernel that is locally adapted to the signal regularity [4]. Numerical examples illustrate the properties of these estimators for piece-wise regular one-dimensional signals. The minimax optimality of wavelet thresholding estimators is studied in Section 11.5.3.

A filter bank of conjugate mirror filters decomposes a discrete signal in a discrete orthogonal wavelet basis defined in Section 7.3.3. The discrete wavelets $\psi_{j,m}[n] = 2^{-j/2} \psi[2^n-m]$ are translated modulo modifications near the boundaries, which are explained in Section 7.5. The support of the signal is normalized to $[0,1]$ with $N$ samples spaced by $N^{-1}$. The scale parameter $2^j$ thus varies from $2^L = N^{-1}$ up to $2^J < 1$:

$$B = \{\{\psi_{j,m}[n]\}_{L < j < J, 0 \leq m < 2^{-j}}, \{\phi_{j,m}[n]\}_{0 \leq m < 2^{-j}}\}. \quad (11.83)$$

A thresholding estimator in this wavelet basis can be written

$$\tilde{F} = \sum_{j=L+1}^{J} \sum_{m=0}^{2^{-j}} \rho_T (\langle X, \psi_{j,m} \rangle) \psi_{j,m} + \sum_{m=0}^{2^{-J}} \rho_T (\langle X, \phi_{j,m} \rangle) \phi_{j,m}, \quad (11.84)$$

where $\rho_T$ is a hard thresholding (11.47) or a soft thresholding (11.49). The upper bound (11.51) proves that the estimation risk is small if the energy of $f$ is absorbed by a few wavelet coefficients.

Adaptive Smoothing The thresholding sets to zero all coefficients $|\langle X, \psi_{j,m} \rangle| \leq T$. This performs an adaptive smoothing that depends on the regularity of the signal $f$. Since $T$ is above the maximum amplitude of the noise coefficients $|\langle W, \psi_{j,m} \rangle|$, if

$$|\langle X, \psi_{j,m} \rangle| = |\langle f, \psi_{j,m} \rangle + \langle W, \psi_{j,m} \rangle| \geq T,$$

then $|\langle f, \psi_{j,m} \rangle|$ has a high probability of being at least of the order $T$. At fine scales $2^j$, these coefficients are in the neighborhood of sharp signal transitions, as shown by Figure 11.4(b). By keeping them, we avoid smoothing these sharp variations. In the regions where $|\langle X, \psi_{j,m} \rangle| < T$, the coefficients $|\langle f, \psi_{j,m} \rangle|$ are likely to be small, which means that $f$ is locally regular. Setting wavelet coefficients to zero is equivalent to locally averaging the noisy data $X$, which is done only if the underlying signal $f$ appears to be regular.

Noise Variance Estimation To estimate the variance $\sigma^2$ of the noise $W[n]$ from the data $X[n] = W[n] + f[n]$, we need to suppress the influence of $f[n]$. When $f$ is piecewise smooth, a robust estimator is calculated from the median of the finest scale wavelet coefficients [220].

The signal $X$ of size $N$ has $N/2$ wavelet coefficients $\{\langle X, \psi_{l,m} \rangle\}_{0 \leq m < N/2}$ at the finest scale $2^J = 2N^{-1}$. The coefficient $|\langle f, \psi_{l,m} \rangle|$ is small if $f$ is smooth over the support of $\psi_{l,m}$, in which case $\langle X, \psi_{l,m} \rangle \approx \langle W, \psi_{l,m} \rangle$. In contrast, $|\langle f, \psi_{l,m} \rangle|$ is large if $f$ has a sharp transition in the support of $\psi_{l,m}$. A piecewise regular signal has few sharp transitions, and hence produces a number of large coefficients that is small compared to $N/2$. At the finest scale, the signal $f$ thus influences the value of a small portion of large amplitude coefficients $\langle X, \psi_{l,m} \rangle$ that are considered to be “outliers.” All others are approximately equal to $\langle W, \psi_{l,m} \rangle$, which are independent Gaussian random variables of variance $\sigma^2$.

A robust estimator of $\sigma^2$ is calculated from the median of $\{\langle X, \psi_{l,m} \rangle\}_{0 \leq m < N/2}$. The median of $P$ coefficients $\text{Med}(\alpha_p)_{0 \leq p < P}$ is the value of the middle coefficient $\alpha_{m_0}$ of rank $P/2$. As opposed to an average, it does not depend on the specific values of coefficients $\alpha_p > \alpha_{m_0}$. If $M$ is the median of the absolute value of $P$ independent Gaussian random variables of zero-mean and variance $\sigma^2$, then one can show that

$$E(M) \approx 0.6745 \sigma_0.$$  

The variance $\sigma^2$ of the noise $W$ is estimated from the median $M_X$ of $\{\langle X, \psi_{l,m} \rangle\}_{0 \leq m < N/2}$ by neglecting the influence of $f$:

$$\sigma = \frac{M_X}{0.6745}. \quad (11.85)$$

Indeed $f$ is responsible for few large amplitude outliers, and these have little impact on $M_X$. 

### 11.3. Thresholding Sparse Representations

#### Hard or Soft Thresholding

If \( T = \sigma \sqrt{2 \log N} \) then (11.50) shows that a soft thresholding guarantees with a high probability that

\[
|\langle \tilde{F}, \psi_{j,m} \rangle| = |\rho_T(\langle X, \psi_{j,m} \rangle)| \lesssim |\langle f, \psi_{j,m} \rangle|.
\]

The estimator \( \tilde{F} \) is at least as regular as \( f \) because its wavelet coefficients have a smaller amplitude. This is not true for the hard thresholding estimator, which leaves unchanged the coefficients above \( T \), and which can therefore be larger than those of \( f \) because of the additive noise component.

Figure 11.4(a) shows a piecewise polynomial signal of degree at most 3, whose wavelet coefficients are calculated with a Symlet 4. Figure 11.4(c) gives an estimation computed with a hard thresholding of the noisy wavelet coefficients in Figure 11.4(b). An estimator \( \hat{\sigma}^2 \) of the noise variance \( \sigma^2 \) is calculated with the median (11.85) and the threshold is set to \( T = \hat{\sigma} \sqrt{2 \log N} \). Thresholding wavelet coefficients removes the noise in the domain where \( f \) is regular but some traces of the noise remain in the neighborhood of singularities. The resulting SNR is 30.8 db. The soft thresholding estimation of Figure 11.4(d) attenuates the noise effect at the discontinuities but the reduction by \( T \) of the coefficient amplitude is much too strong, which reduces the SNR to 23.8 db. As already explained, to obtain comparable SNR values, the threshold of the soft thresholding must be about half the size of the hard thresholding one. In this example, reducing by two the threshold increases the SNR of the soft thresholding to 28.6 db.

#### Multiscale Sure Thresholds

Piecewise regular signals have a proportion of large coefficients \(|\langle f, \psi_{j,m} \rangle|\) that increases when the scale \( 2^j \) increases. Indeed, a singularity creates the same number of large coefficients at each scale, whereas the total number of wavelet coefficients increases when the scale decreases. To use this prior information, one can adapt the threshold choice to the scale \( 2^j \). At large scale \( 2^j \), the threshold \( T_j \) should be smaller in order to avoid setting to zero too many large amplitude signal coefficients, which would increase the risk. Section 11.2.3 explains how to compute the threshold value for a soft thresholding, from the coefficients of the noisy data. We first compute an estimate \( \hat{\sigma}^2 \) of the noise variance \( \sigma^2 \) with the median formula (11.85) at the finest scale. At each scale \( 2^j \), a different threshold is calculated from the \( 2^{-j} \) noisy coefficients \(|\langle X, \psi_{j,m} \rangle|_{0 \leq m < 2^{-j}}\) with the algorithm of Section 11.2.3. A Sure threshold \( T_j \) is calculated with (11.79) at each scale \( 2^j \). A soft thresholding is then applied at each scale \( 2^j \), with a threshold \( T_j \). For a hard thresholding, we have no reliable formula to estimate the risk and hence compute an adapted threshold by minimizing the estimated risk. However, ad-hoc hard thresholds may be computed by multiplying by 2 the Sure threshold calculated for a soft thresholding.

Figure 11.5(c) is a hard thresholding estimation calculated with the same threshold \( T = \hat{\sigma} \sqrt{2 \log N} \) at all scales \( 2^j \). The SNR is 23.3 db. Figure 11.5(d) is obtained by a soft thresholding with Sure thresholds \( T_j \) adapted at each scale \( 2^j \). The SNR is 24.1 db. A soft thresholding with the threshold \( T = \hat{\sigma}/2 \sqrt{2 \log N} \) at all scales gives a smaller SNR equal to 21.7 db. The adaptive calculation of thresholds clearly improves the estimation.

#### Translation Invariance

Thresholding noisy wavelet coefficients creates small ripples near discontinuities, as seen in Figures 11.4(c,d) and 11.5(c,d). Indeed, setting to zero a coefficient \(|\langle f, \psi_{j,m} \rangle|\) subtracts \( \langle f, \psi_{j,m} \rangle \psi_{j,m} \) from \( f \), which introduces oscillations whenever \(|\langle f, \psi_{j,m} \rangle|\) is non-negligible. Figure 11.4(e) and Figures 11.5(e,f) shows that averaging the signal estimation over translated wavelet bases reduce these oscillations, significantly improving the SNR.

A translation invariant wavelet thresholding estimator decomposes the noisy data \( X \) over a dictionary obtained by translating each orthogonal wavelet \( \psi_{j,m}[n] = \psi_j[n - N/2^j] \) by any factor \( 0 \leq p < N \) modulo \( N \). Suppose that each of the \( J - L \) wavelet \( \psi_j[n] \) is \( N \) periodic. This yields a translation invariant dyadic wavelet tight frame including \( \{J - L\} N \) wavelets:

\[
D = \{\psi_{j}[n-p], \phi_{j}[n-p]\}_{L < j < J, 0 \leq p < N}.
\]

and the resulting translation invariant thresholding estimator can be written:

\[
\tilde{F}[n] = \sum_{j=L+1}^{J-N} \sum_{p=0}^{N-1} \rho_T(\langle X[q], \psi_{j}[q-p]\rangle) \psi_{j}[n-p] + \sum_{j=0}^{N-1} \rho_T(\langle X[q], \phi_{j}[q-p]\rangle) \phi_{j}[n-p].
\]
Figure 11.4: (a): Piecewise polynomial signal and its wavelet transform on the right. (b): Noisy signal (SNR = 21.9 db) and its wavelet transform. (c): Estimation reconstructed from the wavelet coefficients above threshold, shown on the right (SNR = 30.8 db). (d): Estimation with a wavelet soft thresholding (SNR = 23.8 db). (e): Estimation with a translation invariant hard thresholding (SNR = 33.7 db).
11.3. Thresholding Sparse Representations

Figure 11.5: (a): Original signal. (b): Noisy signal (SNR = 13.1 db). (c): Estimation by a hard thresholding in a wavelet basis (Symmlet 4), with $T = \hat{\sigma} \sqrt{2 \log N}$ (SNR = 23.3 db). (d): Soft thresholding calculated with Sure thresholds $T_j$ adapted to each scale $2^j$ (SNR = 24.5 db). (e): Translation invariant hard thresholding with $T = \hat{\sigma} \sqrt{2 \log N}$ (SNR = 25.7 db). (f): Translation invariant soft thresholding with Sure thresholds (SNR = 25.6 db).
The decomposition coefficients of $X$ in this dictionary are provided by the dyadic wavelet transform defined in Section 5.2:

$$W X[2^j, p] = \langle X[n], \psi_2[n - p] \rangle \text{ for } 0 \leq p < N.$$ 

The “algorithme à trous” of Section 5.2.2 computes these $(J - L)N$ coefficients for $L < j < J$ with $O(N (J - L))$ operations and reconstructs a signal from the thresholded coefficients with the same number of operations. Since $(J - L) \leq \log_2 N$ the total number of operations is bounded by $O(N \log_2 N)$.

### 11.3.2 Wavelet and Curvelet Image Denoising

Reducing noise by thresholding wavelet coefficients is particularly effective for piecewise regular images, which have sparse wavelet representations. When images includes edges or textures having a regular geometry, then curvelet frames can improve wavelet thresholding estimators.

#### Wavelet Bases

Figure 11.6(b) shows an example of image contaminated by an additive Gaussian white noise of variance $\sigma^2$. This image is decomposed in a separable two-dimensional biorthogonal wavelet basis, generated by a $7/9$ mother wavelet. As in one-dimension, an estimator $\tilde{\sigma}$ of $\sigma$ is computed from the median $M_X$ of the finest scale noisy wavelet coefficient amplitudes with (11.85). For images of $N = 512^2$ pixels, the universal threshold of Theorem 11.7 is $T = 3\tilde{\sigma} \sqrt{2 \log_2 N} \approx 5\tilde{\sigma}$. Wavelet hard thresholding estimators are improved by choosing $T = 3\tilde{\sigma}$, which increases significantly the SNR and the image visual quality. Figure 11.6(c) gives an example. Figure 11.6(d) shows a soft thresholding estimation with $T = 3\tilde{\sigma}/2$, from the same wavelet coefficients. For hard and soft thresholdings estimations, low-frequency scaling coefficients are not thresholded. A hard thresholding at $T$ and a soft thresholding at $T/2$ set to zero the same wavelet coefficients, which are shown in white in Figure 11.6(f). A hard thresholding does not modify the other coefficients shown in black, where as a soft thresholding reduces their amplitude by $T/2$. Coefficients are mostly kept near edges, but some isolated noise coefficients above $3\tilde{\sigma}$ remain in regular regions. These isolated noise wavelet coefficients above threshold produce small wavelet oscillation artifacts, that are more visible with a hard thresholding. The visual quality of edges is also affected by small Gibbs-like oscillations, that also appear in the one-dimensional estimations in Figure 11.4(c) and Figure 11.5(c). The soft thresholding improves the SNR by 1 db relatively to the hard thresholding, and for most images an improvement between 0.5 db and 1 db is observed, with or without finer optimizations of thresholds. A Sure optimization of thresholds at each scale with (11.79), further increases the soft thresholding SNR by over 0.5 db.

A translation invariant wavelet thresholding estimator is computed by decomposing the image in a two-dimensional translation invariant dyadic wavelet tight frame. A fast dyadic wavelet transform is implemented with a separable filter bank, similar to the two-dimensional fast orthogonal transform described in Section 7.7.3. The one-dimensional filterings and subsamplings along the image rows and columns are replaced by the filterings of the “algorithme à trous” in Section 5.2.2. It requires $O(N \log_2 N)$ operations. Figure 11.6(e) is calculated with a translation invariant hard thresholding, which gives a much higher SNR of 24.7 db and a better visual quality. A translation invariant soft thresholding, gives an SNR of 23.6 db. Although a soft thresholding is typically better than a hard thresholding in an orthogonal or biorthogonal basis, a hard thresholding improves a soft thresholding SNR in a translation invariant wavelet frame, and yields the best results. A translation invariant hard thresholding often removes fine textures which affects the visual image quality. By maintaining a small masking noise with $\rho_T(x) = |x|$ if $|x| \geq T$ and $\rho_T(x) = \varepsilon |x|$ if $|x| > T$, the restored image can look more natural.

Section 11.5.3 proves that a thresholding in a wavelet basis has a nearly minimax risk for bounded variation images. When the noise variance $\sigma^2$ decreases to zero, the wavelet thresholding risk is bounded by $O(\sigma \log \sigma)$. Irregular or oscillatory textures are not as well estimated because they do not have a sparse wavelet representation, and create many non-negligible wavelet coefficients. Block thresholding algorithms presented in Section 11.4.2, can improve texture estimation with wavelets.
11.3. Thresholding Sparse Representations

Figure 11.6: (a): Original image. (b): Noisy image (SNR = 18 db). (c): Hard thresholding in a 7/9 separable wavelet basis (SNR = 21.6 db). (d): Soft thresholding (SNR = 22.6 db). (e): Translation invariant hard thresholding (SNR = 24.7 db). (f): Wavelet coefficients above $T = 3\sigma$ are shown in black. All other coefficients are set to zero by the hard and soft thresholding.
Curvelet Frames  Section 9.3.3 shows that images including structures that are geometrically regular, such as \( C^2 \) piecewise regular images, have a representation that is asymptotically more sparse with curvelets than with wavelets. Thresholding curvelet coefficients can then improve wavelet thresholding estimators. This is also valid for textures including geometrically regular structures.

Curvelet tight frames \( \{c^\alpha_{j,m}\}_{j,m,\alpha} \) presented in Section 5.5.2 are composed of anisotropic waveforms with different scales and directions. Curvelets have an elongated support proportional to \( 2^{j/2} \) in a direction \( \alpha \in [0, \pi) \) and a width proportional to \( 2^{j} \) in the perpendicular direction. They are translated along a grid whose intervals are respectively \( 2^{j/2} \) and \( 2^{j} \) in the direction \( \alpha + \pi/2 \). In numerical implementations, normalized curvelets have a frame bound \( A = B \geq 5 \), which corresponds to a minimum redundancy factor \( 5 \). A thresholding curvelet estimation of \( f \) from a noisy observation \( X = f + W \) can be written:

\[
\hat{F} = \sum_{j,m,\alpha} \rho_T(|\langle X, c^\alpha_{j,m} \rangle|) c^\alpha_{j,m}.
\]

Suppose that \( f \) is obtained by discretizing a \( C^2 \) piecewise regular image with \( C^2 \) edge curves, as specified by Definition 9.1. Theorem 9.20 in Section 9.3.3 proves that a non-linear curvelet approximation error has a decay bounded by \( O(M^{-2}(\log M)^3) \), which improves the asymptotic error decay of wavelet approximations. Theorem 11.5 for \( s = 3/2 \) proves that a non-linear approximation error \( \hat{e}_n(M, f) = O(M^{-2}) \) yields an oracle projection risk which satisfies \( r_T(f) = O(\sigma^{4/3}) \).

A thresholding estimator has the same decay up to a \( \log P \) factor, where \( P = AN \) is the total number of curvelets. Taking into account the \( (\log M)^3 \) factor, Candès and Donoho [140] derive that the risk of a curvelet thresholding of a \( C^2 \) piecewise regular image satisfies

\[
E(||\hat{F} - f||^2) = O(||\log \sigma||^2 \sigma^{4/3}) ,
\]

when the noise variance \( \sigma \) decreases to zero. This improves the risk decay \( O(||\log \sigma|| \sigma) \) of a wavelet thresholding estimator. We later prove in (11.152), with \( \alpha = 2 \), that the non-linear minimax risk over uniformly \( C^2 \) Lipschitz images decays like \( \sigma^{4/3} \). A curvelet thresholding estimator nearly achieves this decay despite the presence of edges, and is therefore asymptotically minimax for the class of \( C^2 \) piecewise regular images, up to the \( || \log \sigma ||^2 \) factor.

The threshold \( T = \sigma \sqrt{\log E} \langle \hat{F} \rangle \) is conservative and Figure 11.7 gives an example with \( T = 3\sigma \), which improves the SNR. When the image has textures with regular stripes as in Lena’s hat, a curvelet thresholding gives a better SNR than a translation invariant wavelet thresholding. However, despite the asymptotic improvements of curvelets on \( C^2 \) piecewise regular images, the pepper image in Figure 11.6 is better estimated with wavelets than with curvelets. As opposed to wavelets, curvelets do not have a compact spatial support and their decay is not exponential because their Fourier transform has a compact support. This increases the number of high amplitude curvelet coefficients created by edges, which impacts the image estimation.

Irregular textures or pointwise singularities have a representation that is more sparse with wavelets than with curvelets, and are thus better estimated by a wavelet thresholding. Other image representations may also be used. A thresholding in a best bandlet basis, presented in Section 12.2.4, adapts the basis to the geometric image regularity, and chooses wavelets when there is no such regularity.

11.3.3 Audio Denoising by Time-Frequency Thresholding

Audio signals, whether music or speech, often have a sparse time-frequency representation. Such signals are well approximated by relatively few coefficients in appropriate time-frequency bases or frames. One may thus expect that thresholding these time-frequency representations yield effective noise removal algorithms. Although this is true from a signal to noise ratio point of view, diagonal thresholding algorithms degrades the audio signal quality by introducing a “musical noise”. This musical noise is produced by isolated noisy time-frequency coefficients above threshold.

Sparse audio representations are obtained in wavelet packet and local cosine orthogonal bases, whose time-frequency localization can be adapted to the signal properties. Window Fourier frames also have a time-frequency localization that can be adjusted by choosing an appropriate window size. Thresholding the complex modulus of windowed Fourier frame coefficients seems to better
11.3. Thresholding Sparse Representations

Figure 11.7: (a): Original image. (b): Noisy image (SNR=22 db). (c): Translation invariant wavelet hard thresholding (SNR=25.3 db). (d): Curvelet tight frame hard thresholding (SNR=26 db)

preserve perceptual sound quality than thresholding real wavelet packet or local cosine coefficients. This could be explained by a better restoration of the phase, which is perceptually important for sounds. We shall thus concentrate on windowed Fourier frame thresholding.

A discrete windowed Fourier tight frame of \( C^N \) is constructed in Section 5.4 by translating and modulating a window \( g[n] \), whose support is included in \( [-K/2, K/2 - 1] \). If \( M \) divides \( N \) and

\[
\sum_{m=0}^{N/M-1} |g[n - mM]|^2 = \frac{A}{K} \quad \text{for } 0 \leq n < N
\]

then Theorem 5.18 proves that

\[
g_{m,k}[n] = g[n - mM] e^{i2\pi kn/K}
\]

is a tight frame of \( C^N \), with a frame bound equal to \( A \). Numerical experiments in Figure 11.8 are performed using a square root Hanning window \( g[n] = \sqrt{2/K} \cos(\pi n/K) \) with \( M = K/2 \) and hence \( A = 2 \). The resulting windowed Fourier frame coefficients for \( 0 \leq k < K \), \( 0 \leq m < N/M \) are

\[
Sf[m, k] = \langle f, g_{m,k} \rangle = \sum_{n=-K/2}^{K/2-1} f[n] g[n - mM] e^{-i2\pi kn/K}.
\]

Audio noise are often stationary but not necessarily white. The time-frequency noise variance thus only depends on the frequency and depends upon the noise power spectrum \( \sigma_B^2[m, k] = \sigma_B^2[k] \). A windowed Fourier thresholding estimator can then be written

\[
\tilde{F} = \sum_{m=0}^{N/M-1} \sum_{k=0}^{K-1} \rho_T(\langle X, g_{m,k} \rangle) g_{m,k}
\]

with a threshold \( T^2_k = \lambda \sigma_B^2[k] \). Since the early work on time-frequency audio denoising [107], many types of thresholding functions have been studied for time-frequency audio noise removal [429].
The James-Stein estimator, called empirical Wiener estimator or “power subtraction” in audio noise removal, is often used

\[ a_{k,m}(\langle X, g_{m,k} \rangle) = \frac{\rho_k(\langle X, g_{m,k} \rangle)}{\langle X, g_{m,k} \rangle} = \max \left( 1 - \frac{T_k^2}{\langle X, g_{m,k} \rangle^2}, \varepsilon \right), \tag{11.87} \]

with a masking noise factor \( \varepsilon \) that is often non-zero.

To illustrate the musical noise produced by a spectrogram thresholding, Figure 11.8 shows the denoising of a short recording of a Mozart oboe concerto with a white Gaussian noise. Figure 11.8(c) and 11.8(d) give respectively the log spectrograms \( \log |Sf[m,k]| \) and \( \log |SX[m,k]| \) of the original signal \( f \) and of the noisy sound \( X \). Figure 11.8(g) displays the attenuation factors \( a_{k,m} \) in (11.87) with \( \varepsilon = 0 \). Black points correspond to \( a_{k,m} = 1 \) and white points to \( a_{k,m} = 0 \). For this Mozart recording, when the noisy signal has an SNR that ranges between –2 db up to 15 db, the SNR improvement of this time-frequency soft thresholding estimator is between 8 db and 10 db, which is important. However, as it can be observed in the zoom in Figure 11.8(h), there are isolated attenuation coefficients \( a_{k,m} \approx 1 \), corresponding to black points, which retain noise coefficients in time-frequency regions where the signal has no energy. Similar isolated points appear in the estimation support \( \Lambda_T \) of Figure 11.6(f), for a wavelet image estimation. Because of these isolated attenuation coefficients \( a_{k,m} \approx 1 \), the estimator (11.86) restores windowed Fourier vectors \( g_{m,k} \) that are perceived as a “musical noise”. Despite its small energy, this musical noise is clearly perceived because it is not masked by a sound component at a close frequency and time. Audio masking properties are explained in Section 10.3.3. Despite the signal to noise ratio improvement, this “musical noise” can be more annoying than the original white noise. Translation invariant spectrogram thresholding barely improves the musical noise problem. It can be reduced by increasing thresholds, but this attenuates too much audio signal information, and thus also degrades the sound quality. A non-zero masking noise factor \( \varepsilon \), which maintains a background noise, can be used to reduce the perception of musical noises.

Next section shows that effective musical noise reduction requires using non-diagonal time-frequency estimators, which regularize the time-frequency estimation by processing coefficients in groups.

## 11.4 Non-Diagonal Block Thresholding

A diagonal estimator in a basis processes each coefficient independently and thus does not take advantage of potential dependencies between neighbor coefficients. Ideally, an optimized representation takes advantage of all structural signal correlations to improve the signal sparsity. In practice this is not the case. When a coefficient has a large amplitude, it is likely that some other neighborhood coefficients are also non-negligible, because of signal dependencies that are not fully taken into account by the representation. For example, wavelet image transforms do not capture the geometric regularity of edges, which produce large wavelet coefficients along curves.

Block thresholding estimators introduced by Cai [128] take advantage of such properties by grouping coefficients in blocks and by taking a decision over these blocks. This grouping regularizes thresholding estimators which improves the resulting risk. It also avoids leaving isolated noise coefficients above threshold, perceived as “musical noises” in audio signals and which appear as isolated oscillations in images.

Block thresholding estimators are introduced in Section 11.4.1 together with their mathematical properties. Section 11.4.2 studies the improvements of block thresholding estimations in wavelet bases, for piecewise regular signals and images. For audio noise, Section 11.4.3 shows that time-frequency block thresholdings are effective estimators that avoid introducing musical noises.

### 11.4.1 Block Thresholding in Bases and Frames

A block thresholding estimator implements thresholding decisions over groups of coefficients. The input noisy signal \( X = f + W \) is decomposed in an orthonormal basis \( B = \{ g_m \}_{0 \leq m < M} \) of \( \mathbb{C}^N \), and we write

\[ X_B[m] = \langle X, g_m \rangle, \quad f_B[m] = \langle f, g_m \rangle, \quad W_B[m] = \langle W, g_m \rangle \quad \text{and} \quad \sigma_B^2[m] = E[|W_B[m]|^2]. \]