CHAPTER

12 Sparsity in Redundant Dictionaries

Complex signals such as audio recordings or images often include structures that are not well represented by few vectors in any single basis. Indeed, small dictionaries such as bases have a limited capability of sparse expression. Natural languages build sparsity from large redundant dictionaries of words, which evolve in time. Biological perception systems also seem to incorporate robust and redundant representations that generate sparse encodings at later stages. Larger dictionaries incorporating more patterns can increase sparsity and thus improve applications to compression, denoising, inverse problems, and pattern recognition.

Finding the set of M dictionary vectors that approximate a signal with a minimum error is *NP-hard* in redundant dictionaries. Thus, it is necessary to rely on "good" but nonoptimal approximations, obtained with computational algorithms. Several strategies and algorithms are investigated. Best-basis algorithms restrict the approximations to families of orthogonal vectors selected in dictionaries of orthonormal bases. They lead to fast algorithms, illustrated with wavelet packets, local cosine, and bandlet orthonormal bases. To avoid the rigidity of orthogonality, matching pursuits find freedom in greediness. One by one they select the best approximation vectors in the dictionary. But greediness has it own pitfalls. A basis pursuit implements more global optimizations, which enforce sparsity by minimizing the $\mathbf{l}^1$ norm of decomposition coefficients.

Sparse signal decompositions in redundant dictionaries are applied to noise removal, signal compression, and pattern recognition, and multichannel signals such as color images are studied. Pursuit algorithms can nearly reach optimal M-term approximations in *incoherent* dictionaries that include vectors that are sufficiently different. Learning and updating dictionaries are studied by optimizing the approximation of signal examples.

12.1 IDEAL SPARSE PROCESSING IN DICTIONARIES

Computing an optimal M-term approximation in redundant dictionaries is computationally intractable, but it sets a goal that will guide most of the following sections

and algorithms. The resulting compression algorithms and denoising estimators are described in Sections 12.1.2 and 12.1.3.

12.1.1 Best *M*-Term Approximations

Let $\mathcal{D} = \{\phi_p\}_{p\in\Gamma}$ be a dictionary of P unit norm vectors $\|\phi_p\| = 1$ in a signal space $\mathbb{C}^N$. We study sparse approximations of $f \in \mathbb{C}^N$ with vectors selected in $\mathcal{D}$. Let $\{\phi_p\}_{p\in\Lambda}$ be a subset of vectors in $\mathcal{D}$. We denote by $|\Lambda|$ the cardinal of the index set Λ. The orthogonal projection of f on the space $\mathbf{V}_\Lambda$ generated by these vectors is

$$f_\Lambda = \sum_{p\in\Gamma} a[p]\,\phi_p \quad \text{with} \quad a[p] \neq 0 \quad \text{only for} \quad p \in \Lambda. \tag{12.1}$$

The set $\Lambda \subset \Gamma$ is called the support of the approximation coefficients $a[p]$. Its cardinal $|\Lambda| = \|a\|_0$ is the $\mathbf{l}^0$ pseudo-norm giving the number of nonzero coefficients of a. This support carries geometrical information about f relative to $\mathcal{D}$. In a wavelet basis, it gives the multiscale location of singularities and edges. In a time-frequency dictionary, it provides the location of transients and time-frequency evolution of harmonics.

The best M-term approximation f_Λ minimizes the approximation error $\|f - f_\Lambda\|$ with $|\Lambda| = M$ dictionary vectors. If $\mathcal{D}$ is an orthonormal basis, then Section 9.2.1 proves that the best approximation vectors are obtained by thresholding the orthogonal signal coefficients at some level T. This is not valid if $\mathcal{D}$ is redundant, but Theorem 12.1 proves that a best approximation is still obtained by minimizing an $\mathbf{l}^0$ Lagrangian where T appears as a Lagrange multiplier:

$$\mathcal{L}_0(T, f, \Lambda) = \|f - f_\Lambda\|^2 + T^2\,|\Lambda| = \|f - \sum_{p\in\Gamma} a[p]\phi_p\|^2 + T^2\,\|a\|_0. \tag{12.2}$$

This Lagrangian penalizes the approximation error $\|f - f_\Lambda\|^2$ by the number of approximation vectors.

Theorem 12.1. In a dictionary $\mathcal{D} = \{\phi_p\}_{p\in\Gamma}$,

$$\Lambda_T = \underset{\Lambda\subset\Gamma}{\operatorname{argmin}}\, \mathcal{L}_0(T, f, \Lambda) = \underset{\Lambda\subset\Gamma}{\operatorname{argmin}}\, \|f - f_\Lambda\|^2 + |\Lambda|\,T^2 \tag{12.3}$$

is a best approximation support, which satisfies for all $\Lambda \subset \Gamma$,

$$\|f - f_{\Lambda_T}\| \leq \|f - f_\Lambda\|^2 \quad \text{if} \quad |\Lambda| \leq |\Lambda_T|. \tag{12.4}$$

$$\text{If} \quad \mathcal{L}_0(T, f, \Lambda_T) \leq C\,T^{2-1/s} \text{ with } s \geq 1/2, \text{ then } \|f - f_{\Lambda_T}\|^2 \leq C^{2s}|\Lambda_T|^{1-2s}. \tag{12.5}$$

Proof. The minimization (12.3) implies that any $\Lambda \subset \Gamma$ satisfies

$$\|f - f_\Lambda\|^2 + |\Lambda|\,T^2 \geq \|f - f_{\Lambda_T}\|^2 + |\Lambda_T|\,T^2.$$

Therefore, if $|\Lambda| \leq |\Lambda_T|$, then $\|f - f_\Lambda\| \geq \|f - f_{\Lambda_T}\|$, which proves (12.4).

If $\|f - f_{\Lambda_T}\|^2 + |\Lambda_T| T^2 \leq C\, T^{2-1/s}$, then $|\Lambda_T| \leq C\, T^{-1/s}$, so if $s \geq 1/2$,

$$\|f - f_{\Lambda_T}\|^2 \leq C\, T^{2-1/s} \leq C^{2s}\, |\Lambda_T|^{1-2s}.$$

■

This theorem proves in (12.4) that minimizing the $\mathbf{l}^0$ Lagrangian yields a best approximation $f_M = f_{\Lambda_T}$ of f with $M = |\Lambda_T|$ terms in $\mathcal{D}$. The decay of the approximation error is controlled in (12.5) by the Lagrangian decay as a function of T. If $\mathcal{D}$ is an orthonormal basis, Theorem 12.2 derives that the resulting approximation is a thresholding at T.

Theorem 12.2. If $\mathcal{D}$ is an orthonormal basis, then the best approximation support is

$$\Lambda_T = \underset{\Lambda \subset \Gamma}{\operatorname{argmin}} \|f - f_\Lambda\|^2 + |\Lambda|\, T^2 = \{p \in \Gamma \,:\, |\langle f, \phi_p \rangle| \geq T\} \tag{12.6}$$

and

$$\mathcal{L}_0(T, f, \Lambda_T) = \sum_{p \in \Gamma} \min(|\langle f, \phi_p \rangle|^2, T^2). \tag{12.7}$$

Proof. If $\mathcal{D}$ is an orthonormal basis, then $f_\Lambda = \sum_{p \in \Lambda} \langle f, \phi_p \rangle\, \phi_p$, so

$$\begin{aligned} \|f - f_\Lambda\|^2 + |\Lambda|\, T^2 &= \sum_{p \notin \Lambda} |\langle f, \phi_p \rangle|^2 + |\Lambda|\, T^2 \\ &\geq \sum_{p \in \Gamma} \min(|\langle f, \phi_p \rangle|^2, T^2) = \|f - f_{\Lambda_T}\|^2 + |\Lambda_T|\, T^2. \end{aligned}$$

■

NP-Hard Support Covering

In general, computing the approximation support (12.3) which minimizes the $\mathbf{l}^0$ Lagrangian is proved by Davis, Mallat, and Avellaneda [201] to be an NP-hard problem. This means that there exists dictionaries where finding this solution belongs to a class of NP-complete problems, for which it has been conjectured for the last 40 years that the solution cannot be found with algorithms of polynomial complexity.

The proof [201] shows that for particular dictionaries, finding a best approximation is equivalent to a set-covering problem, which is known to be NP-hard. Let us consider a simple dictionary $\mathcal{D} = \{\phi_p\}_{p \in \Lambda}$ with vectors having exactly three nonzero coordinates in an orthonormal basis $\mathcal{B} = \{g_m\}_{0 \leq m < N}$,

$$\phi_p = \sum_{m \in \Omega_p} g_m \quad \text{with} \quad |\Omega_p| = 3.$$

If the sets $\{\Omega_p\}_{p \in \Lambda}$ define an exact partition of a subset Ω of $\{0, \ldots, N-1\}$, then $f_\Omega = \sum_{m \in \Omega} g_m$ has an exact and optimal dictionary decomposition:

$$f = \sum_{p \in \Lambda} \phi_p \quad \text{with} \quad |\Lambda| = |\Omega|/3.$$

Finding such an exact decomposition for any Ω, if it exists, is an NP-hard *three-sets covering problem*. Indeed, the choice of one element in the solution influences the choice of all others, which essentially requires us to try all possibilities. This argument shows that redundancy makes the approximation problem much more complex. In some dictionaries such as orthonormal bases, it is possible to find optimal M-term approximations with fast algorithms, but these are particular cases.

Since optimal solutions cannot be calculated exactly, it is necessary to find algorithms of reasonable complexity that find "good" if not optimal solutions. Section 12.2 describes the search for optimal solutions restricted to sets of orthogonal vectors in well-structured *tree dictionaries*. Sections 12.3 and 12.4 study pursuit algorithms that search for more flexible and thus nonorthogonal sets of vectors, but that are not always optimal. Pursuit algorithms may yield optimal solutions, if the optimal support Λ satisfies exact recovery properties (studied in Section 12.5).

12.1.2 Compression by Support Coding

Chapter 10 describes transform code algorithms that quantize and code signal coefficients in an orthonormal basis. Increasing the dictionary size can reduce the approximation error by offering more choices. However, it also increases the number of bits needed to code which approximation vectors compress a signal. Optimizing the distortion rate is a trade-off between both effects.

We consider a transform code that approximates f by its orthogonal projection f_Λ on the space $\mathbf{V}_\Lambda$ generated by the dictionary vectors $\{\phi_p\}_{p\in\Lambda}$, and that quantizes the resulting coefficients. The quantization error is reduced by orthogonalizing the family $\{\phi_p\}_{p\in\Lambda}$, for example, with a Gram-Schmidt algorithm, which yields an orthonormal basis $\{g_p\}_{p\in\Lambda}$ of $\mathbf{V}_\Lambda$. The orthogonal projection on $\mathbf{V}_\Lambda$ can then be written as

$$f_\Lambda = \sum_{p\in\Lambda} \langle f, g_p\rangle\, g_p. \tag{12.8}$$

These coefficients are uniformly quantized with

$$Q(x) = \begin{cases} 0 & \text{if } |x| < \Delta/2, \\ \operatorname{sign}(x)\, k\, \Delta & \text{if } (k-1/2)\,\Delta \leqslant |x| < (k+1/2)\,\Delta, \end{cases} \tag{12.9}$$

and the signal recovered from quantized coefficients is

$$\tilde{f} = \sum_{p\in\Lambda} Q(\langle f, g_p\rangle)\, g_p\,. \tag{12.10}$$

The set Λ is further restricted to coefficients $|\langle f, g_p\rangle| \geqslant \Delta/2$, and thus $Q(\langle f, g_p\rangle) \neq 0$, which has no impact on $\tilde{f}$.

Distortion Rate

Let us compute the distortion rate as a function of the dictionary size P. The compression distortion is decomposed in an approximation error plus a quantization error:

$$\|f-\tilde{f}\|^2 = \|f-f_\Lambda\|^2 + \|f_\Lambda-\tilde{f}\|^2. \tag{12.11}$$

Since $|x-Q(x)| \leqslant \Delta/2$,

$$\|f_\Lambda-\tilde{f}\|^2 \leqslant \sum_{p\in\Lambda} |\langle f, g_p\rangle - Q(\langle f, g_p\rangle)|^2 \leqslant |\Lambda|\,\frac{\Delta^2}{4}. \tag{12.12}$$

With (12.11), we derive that the coding distortion is smaller than the $\mathbf{l}^0$ Lagrangian (12.2):

$$d = \|f-\tilde{f}\|^2 \leqslant \|f-f_\Lambda\|^2 + |\Lambda|\,T^2 = \mathcal{L}_0(T, f, \Lambda) \quad \text{for} \quad T=\Delta/2. \tag{12.13}$$

This result shows that minimizing the $\mathbf{l}^0$ Lagrangian reduces the compression distortion.

Having a larger dictionary offers more possibilities to choose Λ and further reduce the Lagrangian. Suppose that some optimization process finds an approximation support Λ_T such that

$$\mathcal{L}_0(T, f, \Lambda_T) \leqslant C\,T^{2-1/s}, \tag{12.14}$$

where C and s depend on the dictionary design and size. The number M of nonzero quantized coefficients is $M = |\Lambda_T| \leqslant C\,T^{-1/s}$. Thus, the distortion rate satisfies

$$d(R, f) = \|f-\tilde{f}\|^2 \leqslant C^{2s}\,M^{1-2s}, \tag{12.15}$$

where R is the total number of bits required to code the quantized coefficients of $\tilde{f}$ with a variable-length code.

As in Section 10.4.1, the bit budget R is decomposed into R_0 bits that code the support set $\Lambda_T \subset \Gamma$, plus R_1 bits to code the M nonzero quantized values $Q(\langle f, g_p\rangle)$ for $p \in \Lambda$. Let P be the dictionary size. We first code $M = |\Lambda_T| \leqslant P$ with $\log_2 P$ bits. There are $\binom{M}{P}$ subsets of size M in a set of size P. Coding Λ_T without any other prior geometric information thus requires $R_0 = \log_2 \binom{M}{P} \sim M \log_2(P/M)$ bits. As in (10.48), this can be implemented with an entropy coding of the binary significance map

$$\forall p \in \Gamma,\ b[p] = \begin{cases} 1 & \text{if } p \in \Lambda_T \\ 0 & \text{if } p \notin \Lambda_T. \end{cases} \tag{12.16}$$

The proportion p_k of quantized coefficients of amplitude $|Q_\Delta(\langle f, g_p\rangle)| = k\Delta$ typically has a decay of $p_k = (k^{-1+\varepsilon})$ for $\varepsilon > 0$, as in (10.57). We saw in (10.58) that coding the amplitude of the M nonzero coefficients with a logarithmic variable length $l_k = \log_2(\pi^2/6) + 2\log_2 k$, and coding their sign, requires a total number of

bits $R_1 \sim M$ bits. For $M \ll P$, it results that the total bit budget is dominated by the number of bits R_0 to code the approximation support Λ_T,

$$R = R_0 + R_1 \sim R_0 \sim M \log_2(P/M),$$

and hence that

$$M \sim R \,|\log_2(P/R)|^{-1}.$$

For a distortion satisfying (12.15), we get

$$d(R,f) = O\Big(C^{2s} R^{1-2s} \,|\log_2(P/R)|^{2s-1}\Big). \tag{12.17}$$

When coding the approximation support Λ_T in a large dictionary of size P as opposed to an orthonormal basis of size N, it introduces a factor $\log_2 P$ in the distortion rate (12.17) instead of the $\log_2 N$ factor in (10.8). This is worth it only if it is compensated by a reduction of the approximaton constant C or an increase of the decay exponent s.

Distortion Rate for Analog Signals

A discrete signal $f[n]$ is most often obtained with a linear discretization that projects an analog signal $\bar{f}(x)$ on an approximation space $\mathbf{U}_N$ of size N. This linear approximation error typically decays like $O(N^{-\beta})$. From the discrete compressed signal $\tilde{f}[n]$, a discrete-to-analog conversion restores an analog approximation $\tilde{\bar{f}}_N(x) \in \mathbf{U}_N$ of $\bar{f}(x)$.

Let us choose a discrete resolution $N \sim R^{(2s-1)/\beta}$. If the dictionary has a polynomial size $P = O(N^\gamma)$, then similar to (10.62), we derive from (12.17) that

$$d(R,\bar{f}) = \|\bar{f} - \tilde{\bar{f}}_N\|^2 = O\Big(R^{1-2s} |\log_2 R|^{2s-1}\Big). \tag{12.18}$$

Thus, the distortion rate in a dictionary of polynomial size essentially depends on the constant C and the exponent s of the $\mathbf{l}^0$ Lagrangian decay $\mathcal{L}_0(T, f, \Lambda_T) \leqslant C\, T^{2-1/s}$ in (12.14). To optimize the asymptotic distortion rate decay, one must find dictionaries of polynomial sizes that maximize s. Section 12.2.4 gives an example of a bandlet dictionary providing such optimal approximations for piecewise regular images.

12.1.3 Denoising by Support Selection in a Dictionary

A hard thresholding in an orthonormal basis is an efficient nonlinear projection estimator, if the basis defines a sparse signal approximation. Such estimators can be improved by increasing the dictionary size. A denoising estimator in a redundant dictionary also projects the observed data on a space generated by an optimized set Λ of vectors. Selecting this support is more difficult than for signal approximation or compression because the noise impacts the choice of Λ. The *model selection theory* proves that a nearly optimal set is estimated by minimizing the $\mathbf{l}^0$ Lagrangian, with an appropriate multiplier T.

Noisy signal observations are written as

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

where $W[n]$ is a Gaussian white noise of variance σ^2. Let $\mathcal{D} = \{\phi_p\}_{p\in\Gamma}$ be a dictionary of P-unit norm vectors. To any subfamily of vectors, $\{\phi_p\}_{p\in\Lambda}$ corresponds an orthogonal projection estimator on the space $\mathbf{V}_\Lambda$ generated by these vectors:

$$X_\Lambda = \sum_{p\in\Lambda} a[p]\,\phi_p.$$

The orthogonal projection in $\mathbf{V}_\Lambda$ satisfies $X_\Lambda = f_\Lambda + W_\Lambda$, so

$$\|f - X_\Lambda\|^2 = \|f - f_\Lambda\|^2 + \|W_\Lambda\|^2. \tag{12.19}$$

The bias term $\|f - f_\Lambda\|^2$ is the signal approximation error, which decreases when $|\Lambda|$ increases. On the contrary, the noise energy $\|W_\Lambda\|^2$ in $\mathbf{V}_\Lambda$ increases when $|\Lambda|$ increases. Reducing the risk amounts to finding a projection support Λ that balances these two terms to minimize their sum.

Since $\mathbf{V}_\Lambda$ is a space of dimension $|\Lambda|$, the projection of a white noise of variance σ^2 satisfies $E\{\|W_\Lambda\|^2\} = |\Lambda|\,\sigma^2$. However, there are 2^P possible subsets Λ in Γ, and $\|W_\Lambda\|^2$ may potentially take much larger values than $|\Lambda|\,\sigma^2$ for some particular sets Λ. A concentration inequality proved in Lemma 12.1 of Theorem 12.3 shows that for any subset Λ of Γ,

$$\|W_\Lambda\|^2 \leqslant (\lambda^2\,\sigma^2\,\log_e P)\,|\Lambda| = T^2\,|\Lambda| \quad \text{for} \quad T = \lambda\,\sigma\,\sqrt{\log_e P},$$

with a probability that tends to 1 as P increases for λ large enough. It results that the estimation error is bounded by the approximation Lagrangian:

$$\|f - X_\Lambda\|^2 \leqslant \mathcal{L}_0(T, f, \Lambda) = \|f - f_\Lambda\|^2 + T^2\,|\Lambda|. \tag{12.20}$$

However, the set Λ_T that minimizes this Lagrangian,

$$\Lambda_T = \underset{\Lambda\subset\Gamma}{\operatorname{argmin}} \left(\|f - f_\Lambda\|^2 + T^2\,|\Lambda| \right), \tag{12.21}$$

can only be found by an oracle because it depends on f, which is unknown. Thus, we need to find an estimator that is nearly as efficient as the oracle projector on this subset Λ_T of vectors.

Penalized Empirical Error

Estimating the oracle set Λ_T in (12.21) requires us to estimate $\|f - f_\Lambda\|^2$ for any $\Lambda \subset \Gamma$. A crude estimator is given by the empirical norm

$$\|X - X_\Lambda\|^2 = \|X\|^2 - \|X_\Lambda\|^2.$$

This may seem naive because it yields a large error,

$$\|X - X_\Lambda\|^2 - \|f - f_\Lambda\|^2 = (\|X\|^2 - \|f\|^2) - (\|X_\Lambda\|^2 - \|f_\Lambda\|^2).$$

Since $X = f + W$ and $X_\Lambda = f_\Lambda + W_\Lambda$, the expected error is

$$E\{\|X\|^2 - \|f\|^2\} - E\{\|X_\Lambda\|^2 - \|f_\Lambda\|^2\} = (N - |\Lambda|)\sigma^2,$$

which is of the order of $N\sigma^2$ if $|\Lambda| \ll N$. However, the first large term does not influence the choice of Λ. The component that depends on Λ is the smaller term $\|X_\Lambda\|^2 - \|f_\Lambda\|^2$, which is only of the order $|\Lambda|\sigma^2$.

Thus, we estimate $\|f - f_\Lambda\|^2$ with $\|X - X_\Lambda\|^2$ in the oracle formula (12.21), and define the best empirical estimation $X_{\tilde{\Lambda}_T}$ as the orthogonal projection on a space $\mathbf{V}_{\tilde{\Lambda}_T}$, where $\tilde{\Lambda}_T$ minimizes the penalized empirical risk:

$$\tilde{\Lambda}_T = \underset{\Lambda \in \Gamma}{\operatorname{argmin}} \left(\|X - X_\Lambda\|^2 + T^2 |\Lambda| \right). \tag{12.22}$$

Theorem 12.3 proves that this estimated set $\tilde{\Lambda}_T$ yields a risk that is within a factor of 4 of the risk obtained by the oracle set Λ_T in (12.21). This theorem is a consequence of the more general model selection theory of Barron, Birgé, and Massart [97], where the optimization of Λ is interpreted as a model selection. Theorem 12.3 was also proved by Donoho and Johnstone [220] for estimating a best basis in a dictionary of orthonormal bases.

Theorem 12.3: *Barron, Birgé, Massart, Donoho, Johnstone.* Let σ^2 be the noise variance and $T = \lambda\, \sigma \sqrt{\log_e P}$ with $\lambda \geq \sqrt{32 + \frac{8}{\log_e P}}$. For any $f \in \mathbb{C}^N$, the best empirical set

$$\tilde{\Lambda}_T = \underset{\Lambda \subset \Gamma}{\operatorname{argmin}} \left(\|X - X_\Lambda\|^2 + T^2 |\Lambda| \right) \tag{12.23}$$

yields a projection estimator $\tilde{F} = X_{\tilde{\Lambda}_T}$ of f, which satisfies

$$E\left[\|\tilde{F} - f\|^2\right] \leq 4 \min_{\Lambda \subset \Gamma} \left(\|f - f_\Lambda\|^2 + T^2 |\Lambda| \right) + \frac{32\,\sigma^2}{P}. \tag{12.24}$$

Proof. Concentration inequalities are at the core of this result. Indeed, the penalty $T^2 |\Lambda|$ must dominate the random fluctuations of the projected noise. We give a simplified proof provided in [233]. Lemma 12.1 uses a concentration inequality for Gaussian variables to ensure with high probability that the noise energy is simultaneously small in all the subspaces $\mathbf{V}_\Lambda$ spanned by subsets of vectors in $\mathcal{D}$.

Lemma 12.1. For any $u \geq 0$ and any $\Lambda \subset \Gamma$,

$$\sigma^{-1} \|W_\Lambda\| \leq \sqrt{|\Lambda|} + \sqrt{4|\Lambda| \log_e P + 2u}, \tag{12.25}$$

with a probability greater than $1 - 2e^{-u}/P$.

This lemma is based on Tsirelson's lemma which proves that for any function L from $\mathbb{C}^N$ to $\mathbb{C}$ that is 1-Lipschitz ($|L(f) - L(g)| \leq \|f - g\|$), and for any normalized Gaussian white noise vector W' of variance $\sigma^2 = 1$,

$$\text{Proba}\left(L(W') \geq E\{L(W')\} + t \right) \leq e^{-t^2/2}.$$

The orthogonal projection's norm $L(W) = \|W_\Lambda\|$ is 1-Lipschitz. Applying Tsirelson's lemma to $W' = \sigma^{-1} W$ for $t = \sqrt{4|\Lambda| \log_e P + 2u}$ yields

$$\text{Proba}\left\{\|W'_\Lambda\| \geqslant E\{\|W'_\Lambda\|\} + \sqrt{4|\Lambda| \log_e P + 2u}\right\} \leqslant P^{-2|\Lambda|} e^{-u}.$$

Since $E\{\|W'_\Lambda\|\} \leqslant (E\{\|W'_\Lambda\|^2\})^{1/2} = \sqrt{|\Lambda|}$, it results that

$$\text{Proba}\left\{\|W'_\Lambda\| \geqslant \sqrt{|\Lambda|} + \sqrt{4|\Lambda| \log_e P + 2u}\right\} \leqslant P^{-2|\Lambda|} e^{-u}.$$

Let us now compute the probability of the existence of a set $\tilde{\Lambda}$ that does not satisfy the lemma condition (12.25), by considering each subset of Γ:

$$\begin{aligned}
&\text{Proba}\left\{\exists \tilde{\Lambda} \subset \Gamma,\ \|W'_{\tilde{\Lambda}}\| \geqslant \sqrt{|\tilde{\Lambda}|} + \sqrt{4|\tilde{\Lambda}| \log_e P + 2u}\right\} \\
&\qquad \leqslant \sum_{\tilde{\Lambda} \subset \Gamma} \text{Proba}\left\{\|W'_{\tilde{\Lambda}}\| \geqslant \sqrt{|\tilde{\Lambda}|} + \sqrt{4|\tilde{\Lambda}| \log_e P + 2u}\right\} \\
&\qquad \leqslant \sum_{\tilde{\Lambda} \subset \Gamma} P^{-2|\tilde{\Lambda}|} e^{-u} \leqslant \sum_{n=1}^{P} \binom{P}{n} P^{-2n} e^{-u} \\
&\qquad \leqslant \sum_{n=1}^{P} P^{-n} e^{-u} \leqslant \frac{P^{-1}}{1 - P^{-1}} e^{-u}.
\end{aligned}$$

It results that for $P \geqslant 2$,

$$\text{Proba}\left\{\exists \tilde{\Lambda} \subset \Gamma,\ \|W'_{\tilde{\Lambda}}\| \geqslant \sqrt{|\tilde{\Lambda}|} + \sqrt{4|\tilde{\Lambda}| \log_e P + 2u}\right\} \leqslant \frac{2}{P} e^{-u},$$

from which we get (12.25) by observing that $W'_\Lambda = \sigma^{-1} W_\Lambda$ because $W' = \sigma^{-1} W$. This finishes the proof of Lemma 12.1.

By construction, the best empirical set $\tilde{\Lambda}_T$ compared to the oracle set Λ_T in (12.21) satisfies

$$\|X - X_{\tilde{\Lambda}_T}\|^2 + T^2 |\tilde{\Lambda}_T| \leqslant \|X - X_{\Lambda_T}\|^2 + T^2 |\Lambda_T|.$$

By using $\|X - X_{\tilde{\Lambda}_T}\|^2 = \|X - f\|^2 + \|f - X_{\tilde{\Lambda}_T}\|^2 + 2\langle X - f, f - X_{\tilde{\Lambda}_T}\rangle$ and a similar equality for $\|X - X_{\Lambda_T}\|^2$ together with the equalities $\|f - X_{\Lambda_T}\|^2 = \|f - f_{\Lambda_T}\|^2 + \|W_{\Lambda_T}\|^2$ and $\langle X - f, X_{\tilde{\Lambda}_T} - X_{\Lambda_T}\rangle = \langle X - f, X_{\tilde{\Lambda}_T} - f_{\Lambda_T}\rangle = \|W_{\Lambda_T}\|^2$, we derive that

$$\|f - X_{\tilde{\Lambda}_T}\|^2 + T^2 |\tilde{\Lambda}_T| \leqslant \|f - f_{\Lambda_T}\|^2 + T^2 |\Lambda_T| + 2|\langle X - f, X_{\tilde{\Lambda}_T} - f_{\Lambda_T}\rangle|. \qquad \textbf{(12.26)}$$

The vectors $\{\phi_p\}_{p \in \tilde{\Lambda}_T \cup \Lambda_T}$ generate a space $\mathbf{V}_{\tilde{\Lambda}_T} + \mathbf{V}_{\Lambda_T}$ of dimension smaller or equal to $|\tilde{\Lambda}_T| + |\Lambda_T|$. We denote by $W_{\tilde{\Lambda}_T \cup \Lambda_T}$ the orthogonal projection of the noise W on this space. The inner product is bounded by writing

$$\begin{aligned}
|2\langle X - f, X_{\tilde{\Lambda}_T} - f_{\Lambda_T}\rangle| &= |2\langle W_{\tilde{\Lambda}_T \cup \Lambda_T}, X_{\tilde{\Lambda}_T} - f_{\Lambda_T}\rangle| \\
&\leqslant 2\|W_{\tilde{\Lambda}_T \cup \Lambda_T}\| \, (\|X_{\tilde{\Lambda}_T} - f\| + \|f - f_{\Lambda_T}\|).
\end{aligned}$$

Lemma 12.1 implies

$$|2\langle X-f, X_{\tilde{\Lambda}_T}-f_{\Lambda_T}\rangle| \leqslant 2\sigma\left(\sqrt{|\tilde{\Lambda}_T|+|\Lambda_T|}+\sqrt{4(|\tilde{\Lambda}_T|+|\Lambda_T|)\log_e P+2u}\right)$$
$$\left(\|X_{\tilde{\Lambda}_T}-f\|+\|f-f_{\Lambda_T}\|\right),$$

with a probability greater than $1-\frac{2}{P}e^{-u}$. Applying $2xy \leqslant \beta^{-2}x^2+\beta^2y^2$ successively with $\beta=1/2$ and $\beta=1$ gives

$$|2\langle X-f, X_{\tilde{\Lambda}_T}-f_{\Lambda_T}\rangle| \leqslant (1/2)^{-2}2\sigma^2(|\tilde{\Lambda}_T|+|\Lambda_T|+4(|\tilde{\Lambda}_T|+|\Lambda_T|\log_e P)+2u)$$
$$+(1/2)^2 2(\|X_{\tilde{\Lambda}_T}-f\|^2+\|f-f_{\Lambda_T}\|^2).$$

Inserting this bound in (12.26) yields

$$\frac{1}{2}\|f-X_{\tilde{\Lambda}_T}\|^2 \leqslant \frac{3}{2}\|f-f_{\Lambda_T}\|^2+\sigma^2(\lambda^2\log_e P+8(1+4\log_e P))|\Lambda_T|$$
$$+\sigma^2(8(1+4\log_e P)-\lambda^2\log_e P)|\tilde{\Lambda}_T|+16\sigma^2u,$$

So that if $\lambda^2 \geqslant 32+\frac{8}{\log_e P}$,

$$\|f-X_{\tilde{\Lambda}_T}\|^2 \leqslant 3\|f-f_{\Lambda_T}\|^2+4\sigma^2\lambda^2\log_e P|\Lambda_T|+32\sigma^2u,$$

which implies for $T=\lambda\,\sigma\sqrt{\log_e P}$ that

$$\|f-X_{\tilde{\Lambda}_T}\|^2 \leqslant 4(\|f-f_{\Lambda_T}\|^2+T^2|\Lambda_T|)+32\sigma^2u,$$

where this result holds with probability greater than $1-\frac{2}{P}e^{-u}$.

Since this is valid for all $u \geqslant 0$, one has

$$\text{Proba}\left\{\|f-X_{\tilde{\Lambda}_T}\|^2-4(\|f-f_{\Lambda_T}\|^2+T^2|\Lambda_T|) \geqslant 32\sigma^2u\right\} \leqslant \frac{2}{P}e^{-u},$$

which implies by integration over u that

$$E\left[\|f-X_{\tilde{\Lambda}_T}\|^2-4(\|f-f_{\Lambda_T}\|^2+T^2|\Lambda_T|)\right] \leqslant 32\sigma^2\frac{2}{P},$$

which proves the theorem result (12.24). ■

This theorem proves that the selection of a best-penalized empirical projection produces a risk that is within a factor of 4 of the minimal oracle risk obtained by selecting the best dictionary vectors that approximate f. Birgé and Massart [114] obtain a better lower bound for λ (roughly $\lambda > \sqrt{2}$ and thus $T > \sigma\sqrt{2\log_e P}$) and a multiplicative factor smaller than 4 with a more complex proof using Talgrand's concentration inequalities.

If $\mathcal{D}$ is an orthonormal basis, then Theorem 12.2 proves that the optimal estimator $\tilde{F}=X_{\tilde{\Lambda}_T}$ is a hard-thresholding estimator at T. Thus, this theorem generalizes the thresholding estimation theorem (11.7) of Donoho and Johnstone that computes an upper bound of the thresholding risk in an orthonormal basis with $P=N$.

The minimum Lagrangian value $\mathcal{L}_0(T, f, \Lambda_T)$ is reduced by increasing the size of the dictionary $\mathcal{D}$. However, this is paid by also increasing T proportionally to $\log_e P$, so that the penalization term $T^2|\Lambda|$ is big enough to dominate the impact of the noise on the selection of dictionary vectors. Increasing $\mathcal{D}$ is thus worth it only if the decay of $\mathcal{L}_0(T, f, \Lambda_T)$ compensates the increase of T, as in the compression application of Section 12.1.2.

Estimation Risk for Analog Signals

Discrete signals are most often obtained by discretizing analog signals, and the estimation risk can also be computed on the input analog signal, as in Section 11.5.3. Let $f[n]$ be the discrete signal obtained by approximating an analog signal $\bar{f}(x)$ in an approximation space $\mathbf{U}_N$ of size N. An estimator $\tilde{F}[n]$ of $f[n]$ is converted into an analog estimation $\bar{F}(x) \in \mathbf{U}_N$ of $\bar{f}(x)$, with a discrete-to-analog conversion. We verify as in (11.141) that the total risk is the sum of the discrete estimation risk plus a linear approximaton error:

$$E\{\|\bar{f} - \bar{F}\|^2\} = E\{\|f - \tilde{F}\|^2\} + \|\bar{f} - P_{\mathbf{U}_N}\bar{f}\|^2.$$

Suppose that the dictionary has a polynomial size $P = O(N^\gamma)$ and that the $\mathbf{l}^0$ Lagrangian decay satisfies

$$\mathcal{L}_0(T, f, \Lambda_T) = \min_{\Lambda \subset \Gamma} \|f - f_\Lambda\|^2 + |\Lambda|\, T^2 \leqslant C\, T^{2-1/s}.$$

If the linear approximation error satisfies $\|\bar{f} - P_{\mathbf{U}_N}\bar{f}\| = O(N^{-\beta})$, then by choosing $N \sim \sigma^{(2s-1)/\beta}$, we derive from (12.24) in Theorem 12.3 that

$$E\{\|\bar{F} - \bar{f}\|^2\} = O(\sigma^{2-1/s}\, |\log \sigma|^{2-1/s}). \tag{12.27}$$

When the noise variance σ^2 decreases, the risk decay depends on the decay exponent s of the $\mathbf{l}^0$ Lagrangian. Optimized dictionaries should thus increase s as much as possible for any given class Θ of signals.

12.2 DICTIONARIES OF ORTHONORMAL BASES

To reduce the complexity of sparse approximations selected in a redundant dictionary, this section restricts such approximations to families of orthogonal vectors. Eliminating approximations from nonorthogonal vectors reduces the number of possible approximation sets Λ in $\mathcal{D}$, which simplifies the optimization. In an orthonormal basis, an optimal nonlinear approximation selects the largest-amplitude coefficients. Dictionaries of orthonormal bases take advantage of this property by regrouping orthogonal dictionary vectors in a multitude of orthonormal bases.

Definition 12.1. A dictionary $\mathcal{D}$ is said to be a dictionary of orthonormal bases of $\mathbb{C}^N$ if any family of orthogonal vectors in $\mathcal{D}$ also belongs to an orthonormal basis $\mathcal{B}$ of $\mathbb{C}^N$ included in $\mathcal{D}$.

A dictionary of orthonormal bases $\mathcal{D} = \{\phi_p\}_{p\in\Gamma}$ is thus a family of $P > N$ vectors that can also be viewed as a union of orthonormal bases, many of which share common vectors. Wavelet packets and local cosine bases in Chapter 8 define dictionaries of orthonormal bases. In Section 12.2.1 we prove that finding a best signal approximation with orthogonal dictionary vectors can be casted as a search for a best orthonormal basis in which orthogonal vectors are selected by a thresholding. Compression and denoising algorithms are implemented in such a best basis. Tree-structured dictionaries are introduced in Section 12.2.2, in order to compute best bases with a fast dynamic programming algorithm.

12.2.1 Approximation, Compression, and Denoising in a Best Basis

Sparse approximations of signals $f \in \mathbb{C}^N$ are constructed with orthogonal vectors selected from a dictionary $\mathcal{D} = \{\phi_p\}_{p\in\Gamma}$ of orthonormal bases with compression and denoising applications.

Best Basis

We denote by $\Lambda_o \subset \Gamma$ a collection of *orthonormal* vectors in $\mathcal{D}$. Sets of nonorthogonal vectors are not considered. The orthogonal projection of f on the space generated by these vectors is then $f_{\Lambda_o} = \sum_{p\in\Lambda_o} \langle f, \phi_p\rangle \phi_p$.

In an orthonormal basis $\mathcal{B} = \{\phi_p\}_{p\in\Gamma_\mathcal{B}}$, Theorem 12.2 proves that the Lagrangian $\mathcal{L}_o(T, f, \Lambda_o) = \|f - f_{\Lambda_o}\|^2 + T^2\,|\Lambda_o|$ is minimized by selecting coefficients above T. The resulting minimum is

$$\mathcal{L}_0(T, f, \mathcal{B}) = \underset{\Lambda_o\subset\Gamma_\mathcal{B}}{\operatorname{argmin}} \|f - f_{\Lambda_o}\|^2 + T^2|\Lambda_o| = \sum_{p\in\Gamma_\mathcal{B}} \min(|\langle f, \phi_p\rangle|^2, T^2). \tag{12.28}$$

Theorem 12.4 derives that a best approximation from orthogonal vectors in $\mathcal{D}$ is obtained by thresholding coefficients in a best basis that minimizes this $\mathbf{l}^0$ Lagrangian.

Theorem 12.4. In the best basis

$$\mathcal{B}_T = \underset{\mathcal{B}\subset\mathcal{D}}{\operatorname{argmin}}\, \mathcal{L}_0(T, f, \mathcal{B}) = \underset{\mathcal{B}\subset\mathcal{D}}{\operatorname{argmin}} \sum_{p\in\Gamma_\mathcal{B}} \min(|\langle f, \phi_p\rangle|^2, T^2), \tag{12.29}$$

the thresholded set $\Lambda_T = \{p \in \Gamma_{\mathcal{B}_T} : |\langle f, \phi_p\rangle| \geqslant T\}$ satisfies

$$\mathcal{L}_0(T, f, \Lambda_T) = \mathcal{L}_0(T, f, \mathcal{B}_T) = \min_{\Lambda_o\subset\Gamma} \|f - f_{\Lambda_o}\|^2 + |\Lambda_o|\,T^2, \tag{12.30}$$

and for all $\Lambda_o \subset \Gamma$,

$$\|f - f_{\Lambda_T}\| \leqslant \|f - f_{\Lambda_o}\| \text{ if } |\Lambda_o| \leqslant |\Lambda_T|. \tag{12.31}$$

Proof. Since any vector in $\mathcal{D}$ belongs to an orthonormal basis $\mathcal{B} \subset \mathcal{D}$, we can write $\Gamma = \cup_{\mathcal{B}\subset\mathcal{D}}\Gamma_\mathcal{B}$, so (12.28) with (12.29) implies (12.30). The optimal approximation result (12.31), like (12.4), comes from the fact that $\|f - f_{\Lambda_T}\|^2 + |\Lambda_T|T^2 \leqslant \|f - f_{\Lambda_o}\|^2 + |\Lambda_o|T^2$. ■

This theorem proves in (12.31) that the thresholding approximation $f_M = f_{\Lambda_T}$ of f in the best orthonormal basis $\mathcal{B}_T$ is the best approximation of f from $M = |\Lambda_T|$ orthogonal vectors in $\mathcal{D}$.

Compression in a Best Orthonormal Basis

Section 12.1.2 shows that quantizing signal coefficients over orthogonal dictionary vectors $\{\phi_p\}_{p\in\Lambda_o}$ yields a distortion

$$d(R,f) \leq \mathcal{L}_0(T,f,\Lambda_o) = \|f - f_{\Lambda_o}\|^2 + T^2\,|\Lambda_o|,$$

where $2T = \Delta$ is the quantization step. A best transform code that minimizes this Lagrangian upper bound is thus implemented in the best orthonormal basis $\mathcal{B}_T$ in (12.29).

With an entropy coding of the significance map (12.16), the number of bits R_0 to code the indexes of the M nonzero quantized coefficients among P dictionary elements is $R_0 = \log_2 \binom{M}{P} \sim M \log_2(P/M)$. However, the number of sets Λ_o of orthogonal vectors in $\mathcal{D}$ is typically much smaller than the number 2^P of subsets Λ in Γ and the resulting number R_0 of bits is thus smaller.

A best-basis search improves the distortion rate $d(R,f)$ if the Lagrangian approximation reduction is not compensated by the increase of R_0 due to the increase of the number of orthogonal vector sets. For example, if the original signal is piecewise smooth, then a best wavelet packet basis does not concentrate the signal energy much more efficiently than a wavelet basis. Despite the fact that a wavelet packet dictionary includes a wavelet basis, the distortion rate in a best wavelet packet basis is then larger than in a single wavelet basis. For geometrically regular images, Section 12.2.4 shows that a dictionary of bandlet orthonormal bases reduces the distortion rate of a wavelet basis.

Denoising in a Best Orthonormal Basis

To estimate a signal f from noisy signal observations

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

where $W[n]$ is a Gaussian white noise, Theorem 12.3 proves that a nearly optimal estimator is obtained by minimizing a penalized empirical Lagrangian,

$$\mathcal{L}_0(T,X,\Lambda) = \|X - X_\Lambda\|^2 + T^2\,|\Lambda|. \tag{12.32}$$

Restricting Λ to be a set Λ_o of orthogonal vectors in $\mathcal{D}$ reduces the set of possible *signal models*. As a consequence, Theorem 12.3 remains valid for this subfamily of models. Theorem 12.4 proves in (12.30) that the Lagrangian (12.32) is minimized by thresholding coefficients in a best basis. A best-basis thresholding thus yields a risk that is within a factor of 4 of the best estimation obtained by an oracle. This best basis can be calculated with a fast algorithm described in Section 12.2.2.

12.2.2 Fast Best-Basis Search in Tree Dictionaries

Tree dictionaries of orthonormal bases are constructed with a recursive split of orthogonal vector spaces and by defining specific orthonormal bases in each subspace. For any additive cost function such as the $\mathbf{l}^0$ Lagrangian (12.2), a fast dynamic

programming algorithm finds a best basis with a number of operations proportional to the size P of the dictionary.

Recursive Split of Vectors Spaces

A tree dictionary $\mathcal{D} = \{\phi_p\}_{p\in\Gamma}$ is obtained by recursively dividing vector spaces into q orthogonal subspaces, up to a maximum recursive depth. This recursive split is represented by a tree. A vector space $\mathbf{W}_d^l$ is associated to each tree node at a depth d and position l. The q children of this node correspond to an orthogonal partition of $\mathbf{W}_d^l$ into q orthogonal subspaces $\mathbf{W}_{d+1}^{ql+i}$ at depth $d+1$, located at the positions $ql+i$ for $0 \leqslant i < q$:

$$\mathbf{W}_d^l = \bigoplus_{i=0}^{q-1} \mathbf{W}_{d+1}^{ql+i}. \tag{12.33}$$

Space $\mathbf{W}_0^0$ at the root of the tree is the full signal space $\mathbb{C}^N$. One or several specific orthonormal bases are constructed for each space $\mathbf{W}_d^l$. The dictionary $\mathcal{D}$ is the union of all these specific orthonormal bases for all the spaces $\mathbf{W}_d^l$ of the tree.

Chapter 8 defines dictionaries of wavelet packet and local cosine bases along binary trees ($q = 2$) for one-dimensional signals and along quad-trees ($q = 4$) for images. These dictionaries are constructed with a single basis for each space $\mathbf{W}_d^l$. For signals of size N, they have $P = N \log_2 N$ vectors. The bandlet dictionary in Section 12.2.4 is also defined along a quad-tree, but each space $\mathbf{W}_d^l$ has several specific orthonormal bases corresponding to different image geometries.

An admissible subtree of a full dictionary tree is a subtree where each node is either a leaf or has its q children. Figure 12.1(b) gives an example of a binary admissible tree. We verify by induction that the vector spaces at the leaves of an admissible tree define an orthogonal partition of $\mathbf{W}_0^0 = \mathbb{C}^N$ into orthogonal subspaces. The union of orthonormal bases of these spaces is therefore an orthonormal basis of $\mathbb{C}^N$.

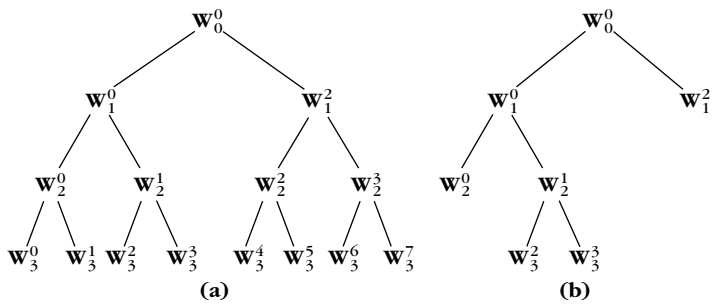


FIGURE 12.1

(a) Full binary tree of depth 3, indexing all possible spaces $\mathbf{W}_d^l$. **(b)** Example of admissible subarea.

Additive Costs

A best basis can be defined with a cost function that is not necessarily the $\mathbf{l}^0$ Lagrangian (12.2). An additive cost function of a signal f in a basis $\mathcal{B} = \{\phi_p\}_{p \in \Gamma_\mathcal{B}}$ is defined as a sum of independent contributions from each coefficient in $\mathcal{B}$:

$$\mathcal{C}(f, \mathcal{B}) = \sum_{p \in \Gamma_\mathcal{B}} C(|\langle f, \phi_p \rangle|). \tag{12.34}$$

A best basis of $\mathbb{C}^N$ in $\mathcal{D}$ minimizes the resulting cost,

$$\widehat{\mathcal{B}} = \operatorname*{argmin}_{\mathcal{B} \subset \mathcal{D}} \mathcal{C}(f, \mathcal{B}). \tag{12.35}$$

The $\mathbf{l}^0$ Lagrangian (12.2) is an example of additive cost function,

$$\mathcal{C}(f, \mathcal{B}) = \mathcal{L}_0(T, f, \mathcal{B}) = \sum_{p \in \Gamma_\mathcal{B}} C(|\langle f,\ \phi_p \rangle|) \quad \text{with} \quad C(x) = \min(T^2, x^2). \tag{12.36}$$

The minimization of an $\mathbf{l}^1$ norm is obtained with

$$\mathcal{C}(f, \mathcal{B}) = \|f_\mathcal{B}\|_1 = \sum_{p \in \Gamma_\mathcal{B}} C(|\langle f, \phi_p \rangle|) \quad \text{for} \quad C(x) = |x|. \tag{12.37}$$

In Section 12.4.1 we introduce a basis pursuit algorithm that also minimizes the $\mathbf{l}^1$ norm of signal coefficients in a redundant dictionary. A basis pursuit selects a best basis but without imposing any orthogonal constraint.

Fast Best-Basis Selection

The fast best-basis search algorithm, introduced by Coifman and Wickerhauser [180], relies on the dictionary tree structure and on the cost additivity. This algorithm is a particular instance of the Classification and Regression Tree (CART) algorithm by Breiman et al. [9]. It explores all tree nodes, from bottom to top, and at each node it computes the best basis $\widehat{\mathcal{B}}_d^l$ of the corresponding space $\mathbf{W}_d^l$.

The cost additivity property (12.34) implies that an orthonormal basis $\mathcal{B} = \cup_{i=0}^{q-1} \mathcal{B}_i$, which is a union of q orthonormal families $\mathcal{B}_i$, has a cost equal to the sum of their cost:

$$\mathcal{C}(f, \mathcal{B}) = \sum_{i=0}^{q-1} \mathcal{C}(f, \mathcal{B}_i).$$

As a result, the best basis $\widehat{\mathcal{B}}_d^l$, which minimizes this cost among all bases of $\mathbf{W}_d^l$, is either one of the specific bases of $\mathbf{W}_d^l$ or a union of the best bases $\widehat{\mathcal{B}}_{d+1}^{ql+i}$ that were previously calculated for each of its subspace $\mathbf{W}_{d+1}^{ql+i}$ for $0 \leqslant i < q$. The decision is thus performed by minimizing the resulting cost, as described in Algorithm 12.1.

ALGORITHM 12.1

Initialization

- Compute all dictionary coefficients $\{\langle f, \phi_p \rangle\}_{p \in \Gamma}$.
- Initialize the cost of each tree space $\mathbf{W}_d^l$ by finding the basis $\mathcal{B}_d^l$ of minimum cost among all specific bases $\mathcal{B}$ of $\mathbf{W}_d^l$:

$$\mathcal{B}_d^l = \underset{\mathcal{B} \text{ of } \mathbf{w}_d^l}{\operatorname{argmin}} C(f, \mathcal{B}) = \underset{\mathcal{B} \text{ of } \mathbf{w}_d^l}{\operatorname{argmin}} \sum_{p \in \Gamma_{\mathcal{B}}} C(|\langle f, \phi_p \rangle|). \tag{12.38}$$

Cost Update

- For each tree node (d, l), visited from the bottom to the top (d decreasing), if we are not at the bottom and if

$$C(f, \mathcal{B}_d^l) > \sum_{i=0}^{q-1} C(f, \widehat{\mathcal{B}}_{d+1}^{ql+i}),$$

then set $\widehat{\mathcal{B}}_d^l = \cup_{i=0}^{q-1} \widehat{\mathcal{B}}_{d+1}^{ql+i}$; otherwise set $\widehat{\mathcal{B}}_d^l = \mathcal{B}_d^l$. ■

This algorithm outputs the best basis $\widehat{\mathcal{B}} = \widehat{\mathcal{B}}_0^0$ of $\mathbb{C}^N = \mathbf{W}_0^0$ that has a minimum cost among all bases of the dictionary. For wavelet packet and local cosine dictionaries, there is a single specific basis per space $\mathbf{W}_d^l$, so (12.38) is reduced to computing the cost in this basis. In a bandlet dictionary there are many specific bases for each $\mathbf{W}_d^l$ corresponding to different geometric image models.

For a dictionary of size P, the number of comparisons and additions to construct this best basis is $O(P)$. The algorithmic complexity is thus dominated by the computation of the P dictionary coefficients $\{\langle f, \phi_p \rangle\}_{p \in \Gamma}$. If implemented with $O(P)$ operations with a fast transform, then the overall computational algorithmic complexity is $O(P)$. This is the case for wavelet packet, local cosine, and bandlet dictionaries.

12.2.3 Wavelet Packet and Local Cosine Best Bases

A best wavelet packet or local cosine basis selects time-frequency atoms that match the time-frequency resolution of signal structures. Therefore, it adapts the time-frequency geometry of the approximation support Λ_T. Wavelet packet and local cosine dictionaries are constructed in Chapter 8. We evaluate these approximations through examples that also reveal their limitations.

Best Orthogonal Wavelet Packet Approximations

A wavelet packet orthogonal basis divides the frequency axis into intervals of varying dyadic sizes 2^j. Each frequency interval is covered by a wavelet packet function that is uniformly translated in time. A best wavelet packet basis can thus be interpreted as a "best" segmentation of the frequency axis in dyadic sizes intervals.

A signal is well approximated by a best wavelet packet basis, if in any frequency interval, the high-energy structures have a similar time-frequency spread. The time translation of the wavelet packet that covers this frequency interval is then well

adapted to approximating all the signal structures in this frequency range that appear at different times. In the best basis computed by minimizing the $\mathbf{l}^0$ Lagrangian in (12.36), Theorem 12.4 proves that the set of Λ_T of wavelet packet coefficients above T correspond to the orthogonal wavelet packet vectors that best approximate f in the whole wavelet packet dictionary. These wavelet packets are represented by Heisenberg boxes, as explained in Section 8.1.2.

Figure 12.2 gives the best wavelet packet approximation set Λ_T of a signal composed of two hyperbolic chirps. The proportion of wavelet packet coefficients that are retained is $M/N = |\Lambda_T|/N = 8\%$. The resulting best M-term orthogonal approximaton f_{Λ_T} has a relative error $\|f - f_{\Lambda_T}\|/\|f\| = 0.11$. The wavelet packet tree was calculated with the symmlet 8 conjugate mirror filter. The time support of chosen wavelet packets is reduced at high frequencies to adapt itself to the chirps' rapid modification of frequency content. The energy distribution revealed by the wavelet packet Heisenberg boxes in Λ_T is similar to the scalogram calculated in Figure 4.17.

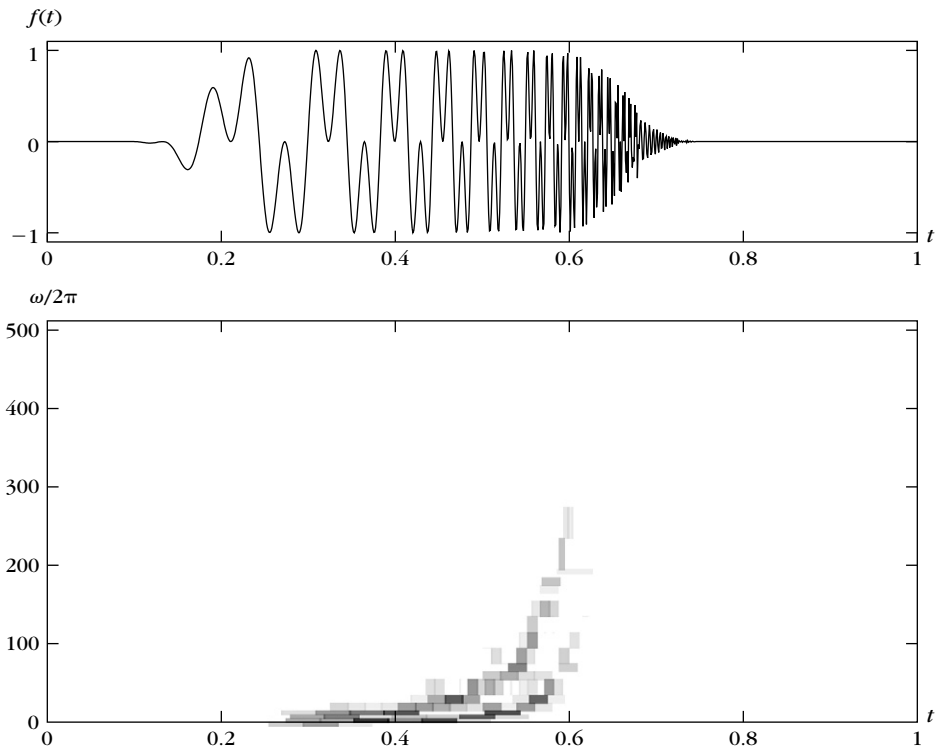


FIGURE 12.2

The top signal includes two hyperbolic chirps. The Heisenberg boxes of the best orthogonal wavelet packets in Λ_T are shown in the bottom image. The darkness of each rectangle is proportional to the amplitude of the corresponding wavelet packet coefficient.

Figure 8.6 gives another example of a best wavelet packet basis for a different multichirp signal, calculated with the entropy cost $C(x) = |x| \log_e |x|$ in (12.34).

Let us mention that the application of best wavelet packet bases to pattern recognition is difficult because these dictionaries are not translation invariant. If the signal is translated, its wavelet packet coefficients are severely modified and the Lagrangian minimization may yield a different basis. This remark applies to local cosine bases as well.

If the signal includes *different types* of high-energy structures, located at different times but in the same frequency interval, there is no wavelet packet basis that is well adapted to all of them. Consider, for example, a sum of four transients centered, respectively, at u_0 and u_1 at two different frequencies ξ_0 and ξ_1:

$$f(t) = \frac{K_0}{\sqrt{s_0}} g\left(\frac{t-u_0}{s_0}\right) e^{i\xi_0 t} + \frac{K_1}{\sqrt{s_1}} g\left(\frac{t-u_1}{s_1}\right) e^{i\xi_0 t} + \frac{K_2}{\sqrt{s_1}} g\left(\frac{t-u_0}{s_1}\right) e^{i\xi_1 t} + \frac{K_3}{\sqrt{s_0}} g\left(\frac{t-u_1}{s_0}\right) e^{i\xi_1 t}. \tag{12.39}$$

The smooth window g has a Fourier transform $\hat{g}$ whose energy is concentrated at low frequencies. The Fourier transform of the four transients have their energy concentrated in frequency bands centered, respectively, at ξ_0 and ξ_1:

$$\hat{f}(\omega) = K_0 \sqrt{s_0}\, \hat{g}\Big(s_0(\omega-\xi_0)\Big) e^{-iu_0(\omega-\xi_0)} + K_1 \sqrt{s_1}\, \hat{g}\Big(s_1(\omega-\xi_0)\Big) e^{-iu_1(\omega-\xi_0)} + K_2 \sqrt{s_1}\, \hat{g}\Big(s_1(\omega-\xi_1)\Big) e^{-iu_0(\omega-\xi_1)} + K_3 \sqrt{s_0}\, \hat{g}\Big(s_0(\omega-\xi_1)\Big) e^{-iu_1(\omega-\xi_1)}.$$

If s_0 and s_1 have different values, the time and frequency spread of these transients is different, which is illustrated in Figure 12.3. In the best wavelet packet basis

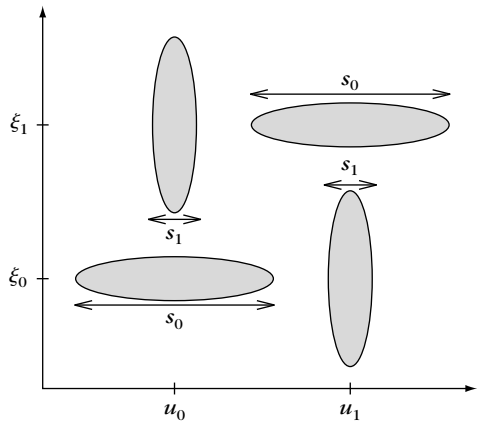


FIGURE 12.3

Time-frequency energy distribution of the four elementary atoms in (12.39).

selection, the first transient $K_0 s_0^{-1/2} g(s_0^{-1}(t-u_0)) \exp(i\xi_0 t)$ "votes" for a wavelet packet basis with a scale 2^j that is of the order s_0 at the frequency ξ_0 whereas $K_1 s_1^{-1/2} g(s_1^{-1}(t-u_1)) \exp(i\xi_0 t)$ "votes" for a wavelet packet basis with a scale 2^j that is close to s_1 at the same frequency. The "best" wavelet packet is adapted to the transient of highest energy. The energy of the smaller transient is then spread across many "best" wavelet packets. The same thing happens for the second pair of transients located in the frequency neighborhood of ξ_1.

Speech recordings are examples of signals that have properties that rapidly change in time. At two different instants in the same frequency neighborhood, the signal may have totally different energy distributions. A best orthogonal wavelet packet basis is not adapted to this time variation and gives poor nonlinear approximations. Sections 12.3 and 12.4 show that a more flexible nonorthogonal approximation with wavelet packets, computed with a pursuit algorithm, can have the required flexibility.

As in one dimension, an image is well approximated in a best wavelet packet basis if its structures within a given frequency band have similar properties across the whole image. For natural scene images, a best wavelet packet often does not provide much better nonlinear approximations than the wavelet basis included in this wavelet packet dictionary. However, for specific classes of images such as fingerprints, one may find wavelet packet bases that significantly outperform the wavelet basis [122].

Best Orthogonal Local Cosine Representations

Tree dictionaries of local cosine bases are constructed in Section 8.5 with $P = N \log_2 N$ local cosine vectors. They divide the time axis into intervals of varying dyadic sizes. A best local cosine basis adapts the time segmentation to the variations of the signal time-frequency structures. It is computed with $O(N \log_2 N)$ operations with the best-basis search algorithm from Section 12.2.2.

In comparison with wavelet packets, we gain time adaptation but we lose frequency flexibility. A best local cosine basis is therefore well adapted to approximating signals with properties that may vary in time, but that do not include structures of very different time and frequency spreads at any given time. Figure 12.4 shows the Heisenberg boxes of the set Λ_T of orthogonal local cosine vectors that best approximate the recording of a bird song, computed by minimizing the $\mathbf{l}^0$ Lagrangian (12.36). The chosen threshold T yields a relative approximation error $\|f - f_{\Lambda_T}\| / \|f\| = 5 \times 10^{-2}$ with $|\Lambda_T|/N = 11\%$ coefficients. The selected local cosine vectors have a time and frequency resolution adapted to the transients and harmonic structures of the signal. Figure 8.19 shows a best local cosine basis that is calculated with an entropy cost function for a speech recording.

The sum of four transients (12.39) is not efficiently represented in a wavelet packet basis but neither is it well approximated in a best local cosine basis. Indeed, if the scales s_0 and s_1 are very different, at u_0 and u_1 this signal includes two transients at the frequency ξ_0 and ξ_1, respectively, that have a very different time-frequency spread. In each time neighborhood, the size of the window is adapted to

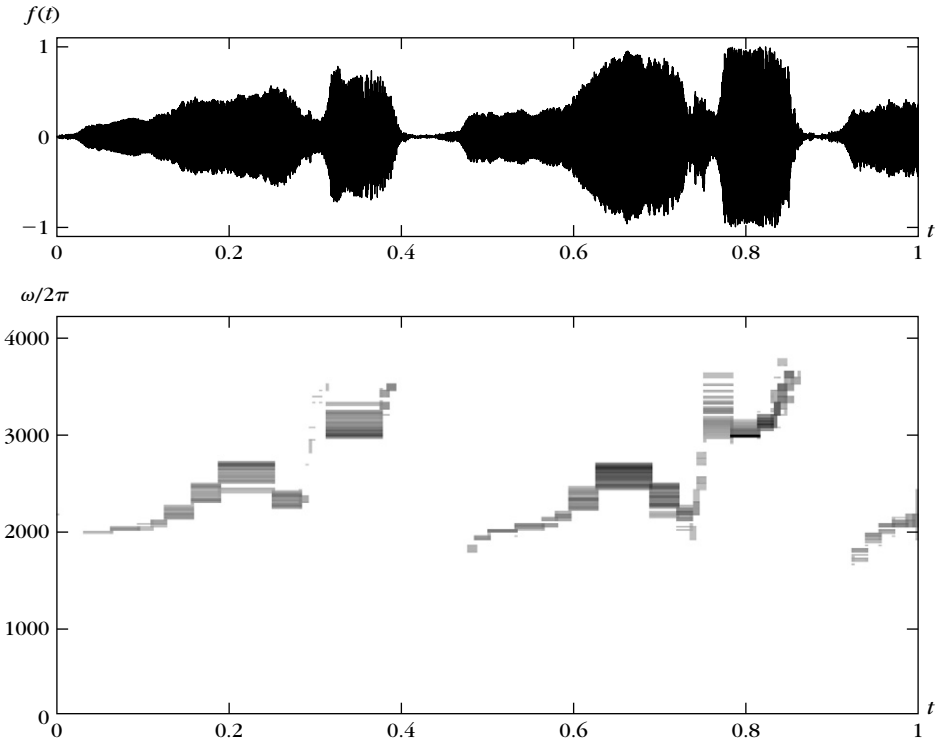


FIGURE 12.4

Recording of a bird song (*top*). The Heisenberg boxes of the best orthogonal local cosine vectors in Λ_T are shown in the bottom image. The darkness of each rectangle is proportional to the amplitude of the local cosine coefficient.

the transient of highest energy. The energy of the second transient is spread across many local cosine vectors. Efficient approximations of such signals require more flexibility, which is provided by the pursuit algorithms from Sections 12.3 and 12.4.

Figure 12.5 gives a denoising example with a best local cosine estimator. The signal in Figure 12.5(b) is the bird song contaminated by an additive Gaussian white noise of variance σ^2 with an SNR of 12 db. According to Theorem 12.3, a best orthogonal projection estimator is computed by selecting a set $\tilde{\Lambda}_T$ of best orthogonal local cosine dictionary vectors, which minimizes an empirical penalized risk. This penalized risk corresponds to the empirical $\mathbf{l}^0$ Lagrangian (12.23), which is minimized by the best-basis algorithm. The chosen threshold of $T = 3.5\,\sigma$ is well below the theoretical universal threshold of $T = \sigma\sqrt{2\log_e P}$, which improves the SNR. The Heisenberg boxes of local cosine vectors indexed by $\tilde{\Lambda}_T$ are shown in boxes of the remaining coefficients in Figure 12.5(c). The orthogonal projection $\tilde{F} = X_{\tilde{\Lambda}}$ is shown in Figure 12.5(d).

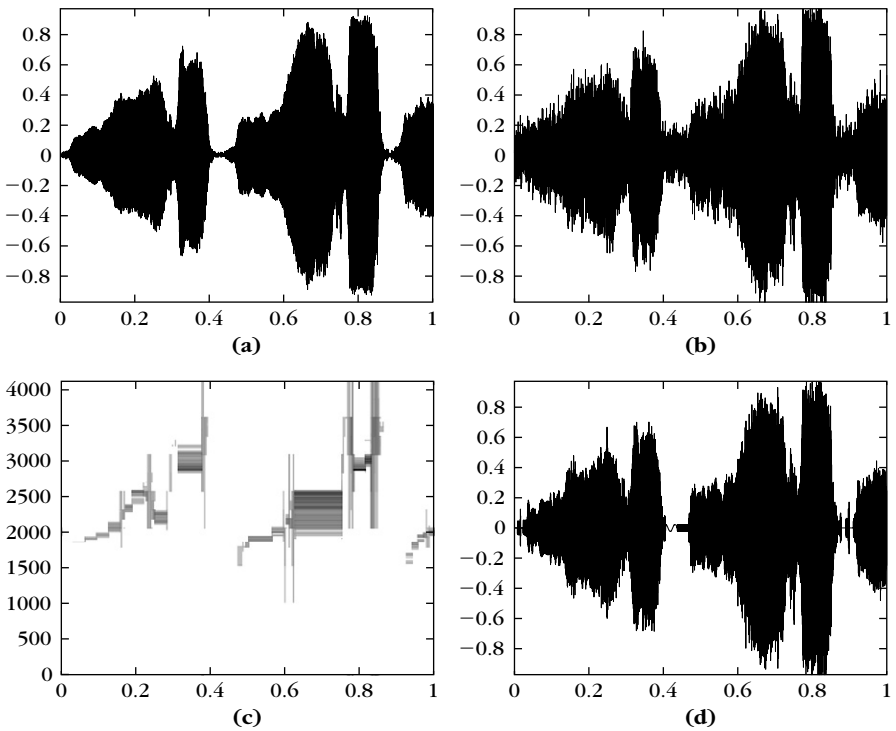


FIGURE 12.5

(a) Original bird song. **(b)** Noisy signal (SNR = 12 db). **(c)** Heisenberg boxes of the set $\tilde{\Lambda}_T$ of estimated best orthogonal local cosine vectors. **(d)** Estimation reconstructed from noisy local cosine coefficients in $\tilde{\Lambda}_T$ (SNR = 15.5 db).

In two dimensions, a best local cosine basis divides an image into square windows that have a size adapted to the spatial variations of local image structures. Figure 12.6 shows the best-basis segmentation of the Barbara image, computed by minimizing the $\mathbf{l}^1$ norm of its coefficients, with the $\mathbf{l}^1$ cost function (12.37). The squares are bigger in regions where the image structures remain nearly the same. Figure 8.22 shows another example of image segmentation with a best local cosine basis, also computed with an $\mathbf{l}^1$ norm.

12.2.4 Bandlets for Geometric Image Regularity

Bandlet dictionaries are constructed to improve image representations by taking advantage of their geometric regularity. Wavelet coefficients are not optimally sparse but inherit geometric image regularity. A bandlet transform applies a directional wavelet transform over wavelet coefficients to reduce the number of large coefficients. This directional transformation depends on a geometric approximation model calculated from the image. Le Pennec, Mallat, and Peyré [342, 365, 396]

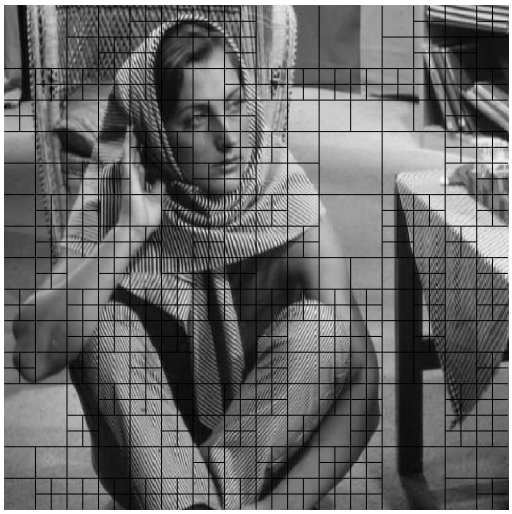


FIGURE 12.6

The grid shows the approximate support of square overlapping windows in the best local cosine basis, computed with an $\mathbf{l^1}$ cost.

introduced dictionaries of orthogonal bandlet bases, where the best-basis selection optimizes the geometric approximation model. For piecewise $\mathbf{C}^\alpha$ images, the resulting M-term bandlet approximations have an optimal asymptotic decay in $O(M^{-\alpha})$.

Approximation of Piecewise $\mathbf{C}^\alpha$ Images

Definition 9.1 defines a piecewise $\mathbf{C}^\alpha$ image f as a function that is uniformly Lipschitz α everywhere outside a set of edge curves, which are also uniformly Lipschitz α. This image may also be blurred by an unknown convolution kernel. If f is uniformly Lipschitz α without edges, then Theorem 9.16 proves that a linear wavelet approximation has an optimal error decay $\varepsilon_l(M,f) = \|f - f_M\|^2 = O(M^{-\alpha})$. Edges produce a larger linear approximation error $\varepsilon_l(M,f) = O(M^{-1/2})$, which is improved by a nonlinear wavelet approximation $\varepsilon_n(M,f) = O(M^{-1})$, but without recovering the $O(M^{-\alpha})$ decay. For $\alpha = 2$, Section 9.3 shows that a piecewise linear approximation over an optimized adaptive triangulation with M triangles reaches the error decay $O(M^{-2})$. Thresholding curvelet frame coefficients also yields a nonlinear approximation error $\varepsilon_n(M,f) = O(M^{-2}(\log M)^3)$ that is nearly optimal. However, curvelet approximations are not as efficient as wavelets for less regular functions such as bounded variation images. If f is piecewise $\mathbf{C}^\alpha$ with $\alpha > 2$, curvelets cannot improve the M^{-2} decay either.

The beauty of wavelet and curvelet approximation comes from their simplicity. A simple thresholding directly selects the signal approximation support. However, for images with geometric structures of various regularity, these approximations do not remain optimal when the regularity exponent α changes. It does not seem

possible to achieve this result without using a redundant dictionary, which requires a more sophisticated approximation scheme.

Elegant adaptive approximation schemes in redundant dictionaries have been developed for images having some geometric regularity. Several algorithms are based on the lifting technique described in Section 7.8, with lifting coefficients that depend on the estimated image regularity [155, 234, 296, 373, 477]. The image can also be segmented adaptively in dyadic squares of various sizes, and approximated on each square by a finite element such as a wedglet, which is a step edge along a straight line with an orientation that is adjusted [216]. Refinements with polynomial edges have also been studied [436], but these algorithms do not provide M-term approximation errors that decay like $O(M^{-\alpha})$ for all piecewise regular $\mathbf{C}^{\alpha}$ images.

Bandletization of Wavelet Coefficients

A bandlet transform takes advantage of the geometric regularity captured by a wavelet transform. The decomposition coefficients of f in an orthogonal wavelet basis can be written as

$$\langle f,\ \psi^k_{j,n}\rangle = f \star \bar{\psi}^k_j(2^j n) \quad \text{with} \quad \bar{\psi}^k_j(x) = 2^{-j}\psi^k(-2^{-j}x), \tag{12.40}$$

for $x = (x_1, x_2)$ and $n = (n_1, n_2)$. The function $f \star \bar{\psi}^k_j(x)$ has the directional regularity of f, for example along an edge, and it is regularized by the convolution with $\bar{\psi}^k_j(x)$. Figure 12.7 shows a zoom on wavelet coefficients near an edge.

Bandlets retransform wavelet coefficients to take advantage of their directional regularity. This is implemented with a directional wavelet transform applied over wavelet coefficients, which creates new vanishing moments in appropriate directions. The resulting bandlets are written as

$$\phi_p(x) = \sum_n \tilde{\psi}_{i,l,m}[n]\, \psi^k_{j,n}(x) \quad \text{with} \quad p = (k, j, l, i, m), \tag{12.41}$$

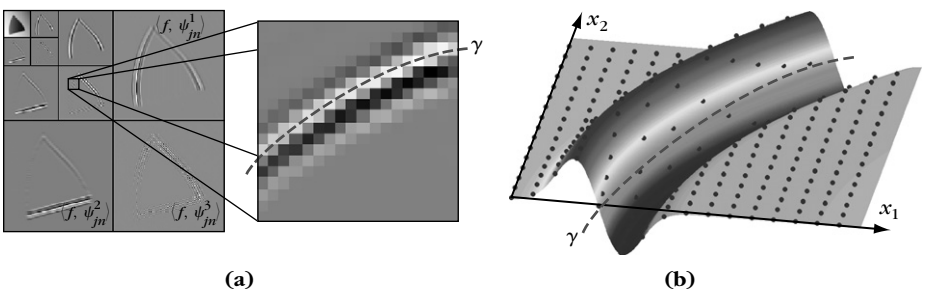


FIGURE 12.7

Orthogonal wavelet coefficients at a scale 2^j are samples of a function $f \star \bar{\psi}^k_j(x)$, shown in **(a)**. The filtered image $f \star \bar{\psi}^k_j(x)$ varies regularly when moving along an edge γ **(b)**.

where $\tilde{\psi}_{i,l,m}[n]$ is a directional wavelet of length 2^i, of width 2^l, and that has a position indexed by m in the wavelet coefficient array. The bandlet function $\phi_p(x)$ is a finite linear combination of wavelets $\psi^k_{j,n}(x)$ and thus has the same regularity as these wavelets. Since $\psi^k_{j,n}(x)$ has a square support of width proportional to 2^j, the bandlet ϕ_p has a support length proportional to 2^{j+i} and a support width proportional to 2^{j+l}.

If the regularity exponent α is known then in a neighborhood of an edge, one would like to have elongated bandlets with an aspect ratio defined by $2^{l+j} = (2^{i+j})^\alpha$, and thus $l = \alpha i + (\alpha - 1)j$. Curvelets satisfy this property for $\alpha = 2$. However, when α is not known in advance and may change, the scale parameters i and l must be adjusted adaptively.

As a result of (12.41), the bandlet coefficients of a signal $\bar{f}$ can be written as

$$\langle \bar{f}, \phi_p \rangle = \sum_n \tilde{\psi}_{i,l,m}[n] \, \langle \bar{f}, \psi^k_{j,n} \rangle .$$

They are computed by applying a discrete directional wavelet tranform on the signal wavelet coefficients $\langle \bar{f}, \psi^k_{j,n} \rangle$ for each k and 2^j. This is also a called a *bandletization* of wavelet coefficients.

Geometric Approximation Model

The discrete directional wavelets $\{\tilde{\psi}_{i,l,m}[n]\}_{i,l,m}$ are defined with a geometric approximation model providing information about the directional image regularity. Many constructions are possible [359]. We describe here a geometric approximation model that is piecewise parallel and yields orthogonal bandlet bases.

For each scale 2^j and direction k, the array of wavelet transform coefficients $\{\langle f, \psi^k_{j,n} \rangle\}_n$ is divided into squares of various sizes, as shown in Figure 12.8(b). In regular image regions, wavelet coefficients are small and do not need to be retransformed. Near junctions, the image is irregular in all directions and these few wavelet coefficients are not retransformed either. It is in the neighborhood of edges and directional image structures that an appropriate retransformation can improve the wavelet sparsity.

A *geometric flow* is defined over each *edge square*. It provides the direction along which the discrete bandlets $\tilde{\psi}_{i,l,m}[n]$ are constructed. It is a vector field, which is parallel horizontally or vertically and points in local directions in which $f \star \bar{\psi}^k_j(x)$ is the most regular. Figure 12.9(a) gives an example. The segmentation of wavelet coefficients in squares and the specification of a geometric flow in each square defines a geometric approximation model that is used to construct a bandlet basis.

Bandlets with Alpert Wavelets

Let us consider a square of wavelet coefficients where a geometric flow is defined. We suppose that the flow is parallel vertically. Its vectors can thus be written $\vec{\tau}(x) = (1, \tilde{\gamma}'(x_1))$. Let $\tilde{\gamma}(x_1)$ be a primitive of $\tilde{\gamma}'(x_1)$. Wavelet coefficients are translated vertically with a warping operator $w(x_1, x_2) = (x_1, x_2 - \tilde{\gamma}(x_1))$ so that the resulting

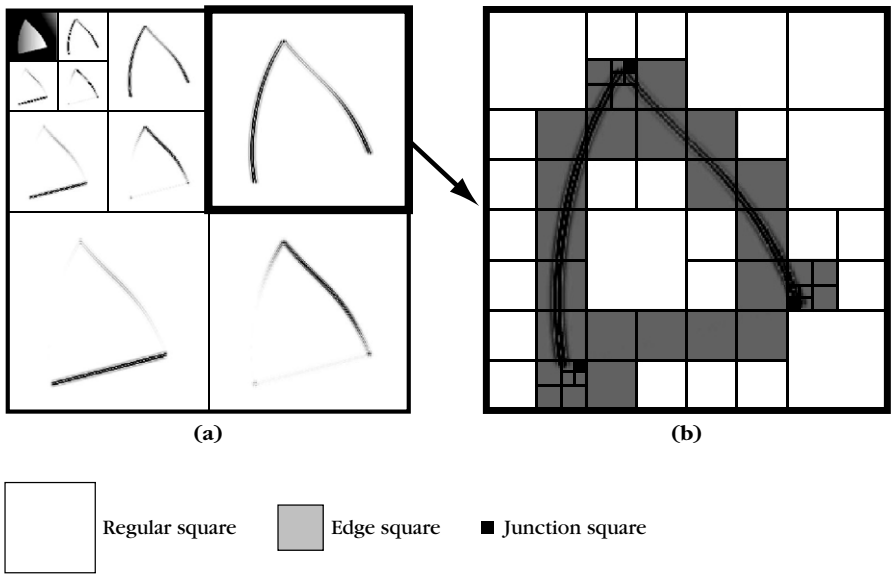


FIGURE 12.8

(a) Wavelet coefficients of the image. **(b)** Example of segmentation of an array of wavelet coefficients $\langle f, \psi_{j,n}^{k} \rangle$ for a particular direction k and scale 2^j.

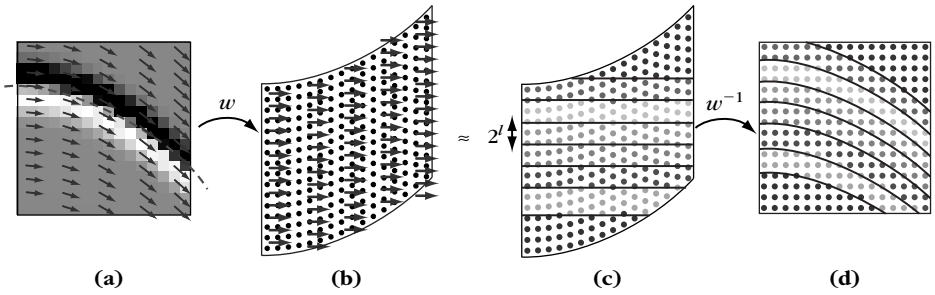


FIGURE 12.9

(a) Square of wavelet coefficients including an edge. A geometric flow nearly parallel to the edge is shown with arrows. **(b)** A vertical warping w maps the flow onto a horizontal flow. **(c)** Support of directional wavelets $\tilde{\psi}_{i,l,m}[w(n)]$ of length 2^i and width 2^l in the warped domain. **(d)** Directional wavelets $\tilde{\psi}_{i,l,m}[n]$ in the square of wavelet coefficients.

geometric flow becomes horizontal, as shown in Figure 12.9(b). In the warped domain, the regularity of $f \star \bar{\psi}_j^k(w(x))$ is now horizontal.

Warped directional wavelets $\tilde{\psi}_{i,l,m}[w(n)]$ are defined to take advantage of this horizontal regularity over the translated orthogonal wavelet coefficients, which are

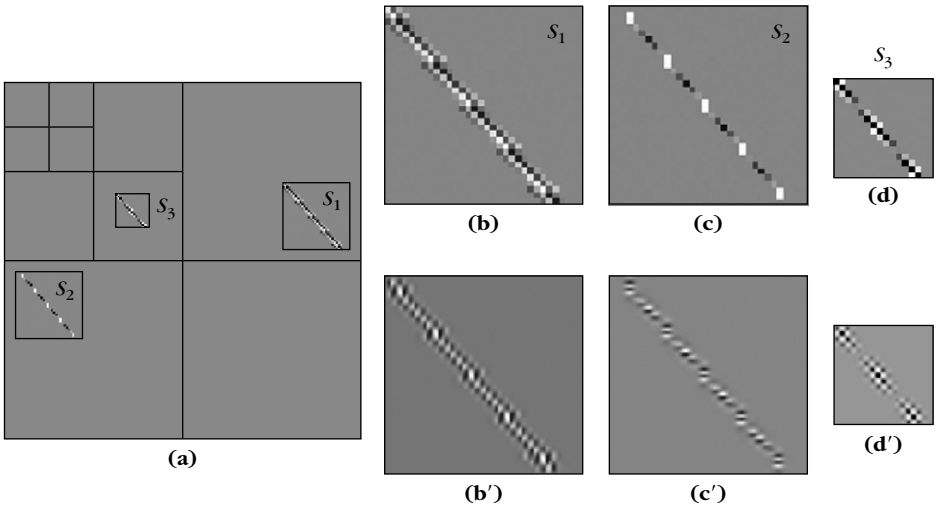


FIGURE 12.10

(a) Squares of wavelet coefficients on which bandlets are computed. **(b–d)** Directional Alpert wavelets $\tilde{\psi}_{i,l,m}$ of different length 2^i and width 2^i. **(b′–d′)** Bandlet functions $\phi_p(x)$ computed from the directional wavelets in (b–d), and the wavelets $\psi_j^k(x)$ corresponding to the squares in (a).

not located on a square grid anymore. Directional wavelets can be constructed with piecewise polynomial Alpert wavelets [84], which are adapted to nonuniform sampling grids [365] and have q vanishing moments. Over a square of width 2^i, a discrete Alpert wavelet $\tilde{\psi}_{i,l,m}[w(n)]$ has a length 2^i, a total of $2^i \times 2^l$ coefficients on its support, and thus a width of the order of 2^l and a position $m2^l$. These directional wavelets are horizontal in the warped domain, as shown in Figure 12.9(c). After inverse warping, $\tilde{\psi}_{i,l,m}[n]$ is parallel to the geometric flow in the original wavelet square, and $\{\tilde{\psi}_{i,l,m}[n]\}_{i,l,m}$ is an orthonormal basis over the square of 2^{2i} wavelet coefficients. The fast Alpert wavelet transform computes 2^{2i} bandlet coefficients in a square of 2^{2i} coefficients with $O(2^{2i})$ operations.

Figure 12.10 shows in (b), (c), and (d) several directional Alpert wavelets $\tilde{\psi}_{i,l,m}[n]$ on squares of different lengths 2^i, and for different width 2^l. The corresponding bandlet functions $\phi_p(x)$ are computed in (b′), (c′), and (d′), with the wavelets $\psi_j^k(x)$ corresponding to the squares shown in Figure 12.10(a).

Dictionary of Bandlet Orthonormal Bases

A bandlet orthonormal basis is defined by segmenting each array of wavelet coefficients $\langle f, \psi_{j,n}^k \rangle$ in squares of various sizes, and by applying an Alpert wavelet transform along the geometric flow defined in each square. A dictionary of bandlet

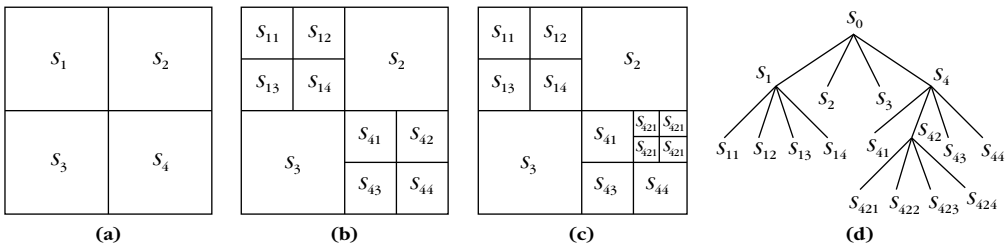


FIGURE 12.11

(a–c) Construction of a dyadic segmentation by successive subdivisions of squares. **(d)** Quad-tree representation of the segmentation. Each leaf of the tree corresponds to a square in the final segmentation.

orthonormal bases is associated to a family of geometric approximation models corresponding to different segmentations and different geometric flows. Choosing a best basis is equivalent to finding an image's best geometric approximation model.

To compute a best basis with the fast algorithm in Section 12.2.2, a tree-structured dictionary is constructed. Each array of wavelet coefficients is divided in squares obtained with a dyadic segmentation. Figure 12.11 illustrates such a segmentation. Each square is recursively subdivided into four squares of the same size until the appropriate size is reached. This subdivision is represented by a quad-tree, where the division of a square appears as the subdivision of a node in four children nodes. The leafs of the quad-tree correspond to the squares defining the dyadic segmentation, as shown in Figure 12.11. At a scale 2^j, the size 2^i of a square defines the length 2^{j+i} of its bandlets. Optimizing this segmentation is equivalent to locally adjusting this length. The resulting bandlet dictionary has a tree structure. Each node of this tree corresponds to a space $\mathbf{W}_d^l$ generated by a square of 2^{2i} orthogonal wavelets at a given wavelet scale 2^j and orientation k. A bandlet orthonormal basis of $\mathbf{W}_d^l$ is associated to each geometric flow.

The number of different geometric flows depends on the geometry's required precision. Suppose that the edge curve in the square is parametrized horizontally and defined by $(x_1, \gamma(x_1))$. For a piecewise $\mathbf{C}^\alpha$ image, $\gamma(x_1)$ is uniformly Lipschitz α. Tangent vectors to the edge are $(1, \gamma'(x_1))$ and $\gamma'(x_1)$ is uniformly Lipschitz $\alpha - 1$. If $\alpha \leqslant q$, then it can be approximated by a polynomial $\tilde{\gamma}'(x_1)$ of degree $q-2$ with

$$\forall (x_1, x_2) \in S, \quad \|\tilde{\gamma}'(x_1) - \gamma'(x_1)\|_\infty = O(2^{i(\alpha-1)}). \tag{12.42}$$

The polynomial $\tilde{\gamma}'(x_1)$ is specified by $q-1$ parameters that must be quantized to limit the number of possible flows. To satisfy (12.42), these parameters are quantized with a precision 2^{-i}. The total number of possible flows in the square of width 2^i is thus $O(2^{i(q-1)})$. A bandlet dictionary $\mathcal{D} = \{\phi_p\}_{p \in \Gamma}$ of order q is constructed

with $\mathbf{C}^q$ wavelets having q vanishing moments, and with polynomial flows of degree $q-2$.

Bandlet Approximation

A best M-term bandlet signal approximation is computed by finding a best basis $\mathcal{B}_T$ and the corresponding best approximation support Λ_T, which minimize the $\mathbf{l}^0$ Lagrangian

$$\mathcal{L}_0(T, f, \Lambda_T) = \mathcal{L}_0(T, f, \mathcal{B}_T) = \underset{\mathcal{B}\subset\mathcal{D}}{\operatorname{argmin}} \sum_{p\in\Gamma_{\mathcal{B}}} \min(|\langle f, \phi_p\rangle|^2, T^2). \tag{12.43}$$

This minimization chooses a best dyadic square segmentation of each wavelet coefficient array, and a best geometric flow in each square. It is implemented with the best-basis algorithm from Section 12.2.2.

An image $\bar{f}\in\mathbf{L}^2[0,1]^2$ is first approximated by its orthogonal projection in an approximation space $\mathbf{V}_L$ of dimension $N=2^{-2L}$. The resulting discrete signal $f[n]=\langle \bar{f}, \phi^2_{L,n}\rangle$ has the same wavelet coefficients as $\bar{f}$ at scales $2^j>2^L$, and thus the same bandlet coefficients at these scales. A best approximation support Λ_T calculated from f yields an $M=|\Lambda_T|$ term approximation of $\bar{f}$:

$$\bar{f}_M(x) = \bar{f}_{\Lambda_T} = \sum_{p\in\Lambda_T} \langle \bar{f}, \phi_p\rangle\, \phi_p(x).$$

Theorem 12.5, proved in [342, 365], computes the nonlinear approximation error $\|\bar{f}-\bar{f}_M\|$ for piecewise regular images.

Theorem 12.5: *Le Pennec, Mallat, Peyré.* Let $\bar{f}\in\mathbf{L}^2[0,1]^2$ be a piecewise $\mathbf{C}^\alpha$ regular image. In a bandlet dictionary of order $q\geqslant\alpha$, for $T>0$ and $2^L=N^{-1/2}\sim T^2$,

$$\mathcal{L}_0(T, f, \Lambda_T) = O(T^{2-2/(\alpha+1)}). \tag{12.44}$$

For $M=|\Lambda_T|$, the resulting best bandlet approximation $\bar{f}_M$ has an error

$$\|\bar{f}-\bar{f}_M\| = O(M^{-\alpha}). \tag{12.45}$$

Proof. The proof finds a bandlet orthogonal basis $\mathcal{B}=\{\phi_p\}_{p\in\Gamma_{\mathcal{B}}}$ such that

$$\mathcal{L}_0(T, f, \mathcal{B}) = \sum_{p\in\Gamma_{\mathcal{B}}} \min(|\langle f, \phi_p\rangle|^2, T^2) = O(T^{2-2/(\alpha+1)}). \tag{12.46}$$

Since $\mathcal{L}_0(T, f, \Lambda_T) = \mathcal{L}_0(T, f, \mathcal{B}_T) \leqslant \mathcal{L}_0(T, f, \mathcal{B})$, it implies (12.44). Theorem 12.1 derives in (12.5) that $\|f-f_{\Lambda_T}\|^2 = O(M^{-\alpha})$ with $M=O(T^{-2/(\alpha+1)})$. A piecewise regular image has a bounded total variation, so Theorem 9.18 proves that a linear approximation error with N larger-scale wavelets has an error $\|\bar{f}-\bar{f}_N\|^2=O(N^{-1/2})$. Since $N^{-1/2}\sim T^2 = O(M^{-\alpha})$, it results that

$$\|\bar{f}-\bar{f}_M\|^2 = \|\bar{f}-\bar{f}_N\|^2 + \|f-f_{\Lambda_T}\|^2 = O(M^{-\alpha}),$$

which proves (12.45).

We give the main ideas for constructing a bandlet basis $\mathcal{B}$ that satisfies (12.46). Detailed derivations can be found in [365]. Following Definition 9.1, a function $\bar{f}$ is piecewise $\mathbf{C}^\alpha$

with a blurring scale $s \geqslant 0$ if $\bar{f} = \tilde{f} \star h_s$ where $\tilde{f}$ is uniformly Lipschitz α on $\Omega = [0,1]^2 - \{e_r\}_{1 \leqslant r < K}$, where the edge curves e_r are uniformly Lipschitz α and do not intersect tangentially. Since h_s is a regular kernel of size s, the wavelet coefficients of $\bar{f}$ at a scale 2^j behave as the wavelet coefficients of $\tilde{f}$ at a scale $2^{j'} \sim 2^j + s$ multiplied by s^{-1}. Thus, it has a marginal impact on the proof. We suppose that $s = 0$ and consider a signal $\bar{f}$ that is not blurred.

Wavelet coefficients $\langle \bar{f}, \psi^k_{j,n} \rangle = \langle f, \psi^k_{j,n} \rangle$ are computed at scales $2^j > 2^L = T^2$. A dyadic segmentation of each wavelet coefficient array $\{\langle \bar{f}, \psi^k_{j,n} \rangle\}_n$ is computed according to Figure 12.8, at each scale $2^j > 2^L$ and orientation $k = 1, 2, 3$. Wavelet arrays are divided into three types of squares. In each type of square a geometric flow is specified, so that the resulting bandlet basis $\mathcal{B}$ has a Lagrangian that satisfies $\mathcal{L}_o(T, f, \mathcal{B}) = O(T^{2-2/(\alpha+1)})$. This is proved by verifying that the number of coefficients above T is $O(T^{-2/(\alpha+1)})$ and that the energy of coefficients below T is $O(T^{2-2/(\alpha+1)})$.

- *Regular squares* correspond to coefficients $\langle \bar{f}, \psi^k_{j,n} \rangle$, such that f is uniformly Lipschitz α over the support of all $\psi^k_{j,n}$.
- *Edge squares* include coefficients corresponding to wavelets with support that intersects a single edge curve. This edge curve can be parametrized horizontally or vertically in each square.
- *Junction squares* include coefficients corresponding to wavelets with support that intersects at least two different edge curves.

Over regular squares, since $\bar{f}$ is uniformly Lipschitz α, Theorem 9.15 proves in (9.15) that $|\langle \bar{f}, \psi^k_{j,n} \rangle| = O(2^{-j(\alpha+1)})$. These small wavelet coefficients do not need to be retransformed and no geometric flow is defined over these squares. The number of coefficients above T in such squares is indeed $O(T^{-2/(\alpha+1)})$ and the energy of coefficients below T is $O(T^{2-2/(\alpha+1)})$.

Since edges do not intersect tangentially, one can construct junction squares of width $2^i \leqslant C$ where C does not depend on 2^j. As a result, over the $|\log_2 T^2|$ scales $2^j \geqslant T^2$, there are only $O(|\log_2 T|)$ wavelet coefficients in these junction squares, which thus have a marginal impact on the approximation.

At a scale 2^j, an edge square S of width 2^i has $O(2^i 2^{-j})$ large coefficients having an amplitude $O(2^j)$ along the edge. Bandlets are created to reduce the number of these large coefficients that dominate the approximation error. Suppose that the edge curve in S is parametrized horizontally and defined by $(x_1, \gamma(x_1))$. Following (12.42), a geometric flow of vectors $(1, \tilde{\gamma}'(x_1))$ is defined over the square, where $\tilde{\gamma}'(x_1)$ is a polynomial of degree $q - 2$, which satisfies

$$\|\tilde{\gamma}'(x_1) - \gamma'(x_1)\|_\infty = O(2^{i(\alpha-1)}).$$

Let $w(x_1, x_2) = (x_1, x_2 - \tilde{\gamma}(x_1))$ be the warping that maps this flow to a horizontal flow, as illustrated in Figure 12.9. One can prove [365] that the warped wavelet transform satisfies

$$\left| \frac{\partial^{p_1+p_2} f \star \bar{\psi}^k_j(w(x))}{\partial^{p_1} x_1 \, \partial^{p_2} x_2} \right| = O(2^j \, 2^{-j(p_1/\alpha + p_2)}) \text{ for any } 0 \leqslant p_1, p_2 \geqslant q.$$

The bandlet transform takes advantage of the regularity of $f \star \bar{\psi}_j^k(w(x))$ along x_1 with Alpert directional wavelets having q vanishing moments along x_1. Computing the amplitude of the resulting bandlet coefficients shows that there are $O(2^i T^{-2/(\alpha+1/2)})$ bandlet coefficients of amplitude larger than T and the error of all coefficients below T is $O(2^i T^{2-2/(\alpha+1)})$. The total length of edge squares is proportional to the total length of edges in the image, which is $O(1)$. Summing the errors over all squares gives a total number of bandlet coefficients, which is $O(T^{-1/(\alpha+1/2)})$, and a total error, which is $(O(T^{2-2/(\alpha+1)}).)$

As a result, the bandlet basis $\mathcal{B}$ defined over the three types of squares satisfies $\mathcal{L}_0(T, f, \mathcal{B}) = O(T^{2-2/(\alpha+1)})$, which finishes the proof. ■

The best-basis algorithm finds a best geometry to approximate each image. This theorem proves that the resulting approximation error decays as quickly as if the image was uniformly Lipschitz α over its whole support $[0, 1]^2$. Moreover, this result is adaptive in the sense that it is valid for all $\alpha \leqslant q$.

The downside of bandlet approximations is the dictionary size. In a square of width 2^i, we need $O(2^{i(q-1)})$ polynomial flows, each of which defines a new bandlet family. As a result, a bandlet dictionary of order q includes $P = O(N^{1+(q-1)^3/q})$ different bandlets. The total number of operations to compute a best bandlet approximation is $O(P)$, which becomes very large for $q > 2$. A fast implementation is described in [398] for $q = 2$ where $P = O(N^{3/2})$. Theorem 12.5 is then reduced to $\alpha \leqslant 2$. It still recovers the $O(M^{-2})$ decay for $\mathbf{C}^2$ images obtained in Theorem 9.19 with piecewise linear approximations over an adaptive triangulation.

Figure 12.12 shows a comparison of the approximation of a piecewise regular image with M largest orthogonal wavelet coefficients and the best M orthogonal bandlet coefficients. Wavelet approximations exhibit more ringing artifacts along edges because they do not capture the anisotropic regularity of edges.

(a) **(b)** **(c)**

FIGURE 12.12

(a) Original image. **(b)** Approximation with $M/N = 1\%$ largest-amplitude wavelet coefficients (SNR = 21.8 db). **(c)** Approximation with $M/N = 1\%$ best bandlet vectors computed in a best bandlet basis (SNR = 23.2 db).

Bandlet Compression

Following the results of Section 12.1.2, a bandlet compression algorithm is implemented by quantizing the best bandlet coefficients of f with a bin size $\Delta = 2T$. According to Section 12.1.2, the approximation support Λ_T is coded on $R_0 = M \log_2(P/N)$ bits and the amplitude of nonzero quantized coefficients with $R_1 \sim M$ bits [365]. If f is the discretization of a piecewise $\mathbf{C}^\alpha$ image $\bar{f}$, since $\mathcal{L}_0(T, f, \Lambda_T) = O(T^{2-2/(\alpha+1)})$, we derive from (12.17) with $s = (\alpha+1)/2$ that the distortion rate satisfies

$$d(R, f) = O(R^{-\alpha} \, |\log_2(P/R)|^{-\alpha}).$$

Analog piecewise $\mathbf{C}^\alpha$ images are linearly approximated in a multiresolution space of dimension N with an error $\|\bar{f} - \bar{f}_N\|^2 = O(N^{-1/2})$. Taking this into account, we verify that the analog distortion rate satisfies the asymptotic decay rate (12.18)

$$d(R, \bar{f}) = O\left(R^{-\alpha} |\log_2 R|^{-\alpha}\right).$$

Although bandlet compression improves the asymptotic decay of wavelet compression, such coders are not competitive with a JPEG-2000 wavelet image coder, which requires less computations. Moreover, when images have no geometric regularity, despite the fact that the decay rate is the same as with wavelets, bandlets introduce an overhead because of the large dictionary size.

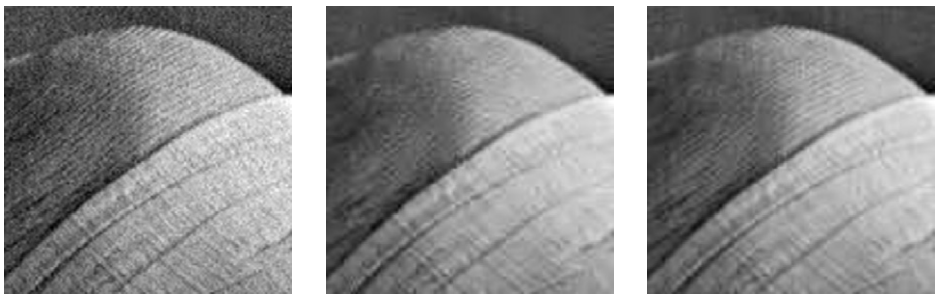
Bandlet Denoising

Let W be a Gaussian white noise of variance σ^2. To estimate f from $X = f + W$, a best bandlet estimator $\tilde{F} = X_{\tilde{\Lambda}_T}$ is computed according to Section 12.2.1 by projecting X on an optimized family of orthogonal bandlets indexed by $\tilde{\Lambda}_T$. It is obtained by thresholding at T the bandlet coefficients of X in the best bandlet basis $\tilde{\mathcal{B}}_T$, which minimizes $\mathcal{L}_0(T, X, \mathcal{B})$, for $T = \sigma\sqrt{2\log_e P}$.

An analog estimator $\bar{F}$ of $\bar{f}$ is reconstructed from the noisy signal coefficients in $\tilde{\Lambda}_T$ with the analog bandlets $\{\phi_p(x)\}_{p \in \tilde{\Lambda}_T}$. If $\bar{f}$ is a piecewise $\mathbf{C}^\alpha$ image $\bar{f}$, then Theorem 12.5 proves that $\mathcal{L}_0(T, f, \Lambda_T) = O(T^{2-2/(\alpha+1)})$. The computed risk decay (12.27) thus applies for $s = (\alpha+1)/2$:

$$E\{\|\bar{F} - \bar{f}\|^2\} = O(\sigma^{2-2/(\alpha+1)} \, |\log \sigma|^{2-2/(\alpha+1)}). \tag{12.47}$$

This decay rate [233] shows that a bandlet estimation over piecewise $\mathbf{C}^\alpha$ images nearly reaches the minimax risk $r_n \sim \sigma^{2-2/(\alpha+1)}$ calculated in (11.152) for uniformly $\mathbf{C}^\alpha$ images. Figure 12.13 gives a numerical example comparing a best bandlet estimation and a translation-invariant wavelet thresholding estimator for an image including regular geometric structures. The threshold is $T = 3\sigma$ instead of $T = \sigma\sqrt{2\log_e P}$, because it improves the SNR.



(a) (b) (c)

FIGURE 12.13

(a) Noisy image (SNR = 22 db). **(b)** Translation-invariant wavelet hard thresholding (SNR = 25.3 db). **(c)** Best bandlet thresholding estimation (SNR = 26.4 db).

12.3 GREEDY MATCHING PURSUITS

Computing an optimal M-term approximation f_M of a signal f with M vectors selected in a redundant dictionary $\mathcal{D}$ is NP-hard. Pursuit strategies construct nonoptimal yet efficient approximations with computational algorithms. Matching pursuits are greedy algorithms that select the dictionary vectors one by one, with applications to compression, denoising, and pattern recognition.

12.3.1 Matching Pursuit

Matching pursuit introduced by Mallat and Zhang [366] computes signal approximations from a redundant dictionary, by iteratively selecting one vector at a time. It is related to projection pursuit algorithms used in statistics [263] and to shape–gain vector quantizations [27].

Let $\mathcal{D} = \{\phi_p\}_{p\in\Gamma}$ be a dictionary of $P > N$ vectors having a unit norm. This dictionary is supposed to be complete, which means that it includes N linearly independent vectors that define a basis of the signal space $\mathbb{C}^N$. A matching pursuit begins by projecting f on a vector $\phi_{p_0} \in \mathcal{D}$ and by computing the residue Rf:

$$f = \langle f, \phi_{p_0}\rangle\, \phi_{p_0} + Rf. \tag{12.48}$$

Since Rf is orthogonal to ϕ_{p_0},

$$\|f\|^2 = |\langle f, \phi_{p_0}\rangle|^2 + \|Rf\|^2. \tag{12.49}$$

To minimize $\|Rf\|$, we must choose $\phi_{p_0} \in \mathcal{D}$ such that $|\langle f, \phi_{p_0}\rangle|$ is maximum. In some cases it is computationally more efficient to find a vector ϕ_{p_0} that is almost optimal:

$$|\langle f, \phi_{p_0}\rangle| \geqslant \alpha \sup_{p\in\Gamma} |\langle f, \phi_p\rangle|, \tag{12.50}$$

where $\alpha \in (0, 1]$ is a relaxation factor. The pursuit iterates this procedure by sub-decomposing the residue. Let $R^0 f = f$. Suppose that the mth-order residue $R^m f$ is already computed for $m \geqslant 0$. The next iteration chooses $\phi_{p_m} \in \mathcal{D}$ such that

$$|\langle R^m f, \phi_{p_m} \rangle| \geqslant \alpha \sup_{p \in \Gamma} |\langle R^m f, \phi_p \rangle|, \tag{12.51}$$

and projects $R^m f$ on ϕ_{p_m}:

$$R^m f = \langle R^m f, \phi_{p_m} \rangle \, \phi_{p_m} + R^{m+1} f. \tag{12.52}$$

The orthogonality of $R^{m+1} f$ and ϕ_{p_m} implies

$$\|R^m f\|^2 = |\langle R^m f, \phi_{p_m} \rangle|^2 + \|R^{m+1} f\|^2. \tag{12.53}$$

Summing (12.52) from m between 0 and $M-1$ yields

$$f = \sum_{m=0}^{M-1} \langle R^m f, \phi_{p_m} \rangle \, \phi_{p_m} + R^M f. \tag{12.54}$$

Similarly, summing (12.53) from m between 0 and $M-1$ gives

$$\|f\|^2 = \sum_{m=0}^{M-1} |\langle R^m f, \phi_{p_m} \rangle|^2 + \|R^M f\|^2. \tag{12.55}$$

Convergence of Matching Pursuit

A matching pursuit has an exponential decay if the residual $\|R^m f\|$ has a minimum rate of decay. The conservation of energy (12.53) implies

$$\frac{\|R^{m+1} f\|^2}{\|R^m f\|^2} = 1 - \left| \left\langle \frac{R^m f}{\|R^m f\|}, \phi_{p_m} \right\rangle \right|^2 \leqslant 1 - \mu^2(R^m f, \mathcal{D}), \tag{12.56}$$

where $\mu(r, \mathcal{D})$ is the coherence of a vector relative to the dictionary, defined by

$$\mu(r, \mathcal{D}) = \max_{p \in \Gamma} \left| \left\langle \frac{r}{\|r\|}, \phi_p \right\rangle \right| \leqslant 1.$$

Theorem 12.6 proves that

$$\mu_{\min}(\mathcal{D}) = \inf_{r \in \mathbb{C}^N, r \neq 0} \mu(r, \mathcal{D}) > 0,$$

and thus that matching pursuits converge exponentially.

Theorem 12.6. The residual $R^m f$ computed by a matching pursuit with relaxation parameter $\alpha \in (0, 1]$ satisfies

$$\|R^m f\|^2 \leqslant (1 - \alpha^2 \mu_{\min}(\mathcal{D})^2)^m \|f\|^2 \quad \text{with} \quad 1 \geqslant \mu_{\min}(\mathcal{D}) > 0. \tag{12.57}$$

As a consequence,

$$f = \sum_{m=0}^{+\infty} \langle R^m f,\ \phi_{p_m} \rangle\, \phi_{p_m}, \quad \text{and} \quad \|f\|^2 = \sum_{m=0}^{+\infty} |\langle R^m f,\ \phi_{p_m} \rangle|^2. \tag{12.58}$$

Proof. The atom ϕ_{p_m} selected by a matching pursuit satisfies $|\langle R^m f,\ \phi_{p_m} \rangle| \geqslant \alpha \sup_{p \in \Gamma} |\langle R^m f,\ \phi_p \rangle|$. It results from (12.56) that

$$\frac{\|R^{m+1} f\|^2}{\|R^m f\|^2} \leqslant 1 - \alpha^2 {\mu_{\min}}^2(\mathcal{D}).$$

Iterating on this equation proves that

$$\|R^m f\|^2 \leqslant (1 - \alpha^2 {\mu_{\min}}^2(\mathcal{D}))^m\, \|f\|^2. \tag{12.59}$$

To verify that $\mu_{\min}(\mathcal{D}) > 0$, a contrario lets us suppose that $\mu_{\min}(\mathcal{D}) = 0$. There exist $\{f_m\}_{m \in \mathbb{N}}$ with $\|f_m\| = 1$ such that

$$\lim_{m \to +\infty} \sup_{p \in \Gamma} |\langle f_m, \phi_p \rangle| = 0. \tag{12.60}$$

Since the unit sphere of $\mathbb{C}^N$ is compact, there exists a subsequence $\{f_{m_k}\}_{k \in \mathbb{N}}$ that converges to a unit vector $f \in \mathbb{C}^N$. It follows that

$$\sup_{p \in \Gamma} |\langle f, \phi_p \rangle| = \lim_{k \to +\infty} \sup_{p \in \Gamma} |\langle f_{m_k}, \phi_p \rangle| = 0, \tag{12.61}$$

so $\langle f, \phi_p \rangle = 0$ for all $\phi_p \in \mathcal{D}$. Since $\mathcal{D}$ contains a basis of $\mathbb{C}^N$, necessarily $f = 0$, which is not possible because $\|f\| = 1$. It results that, necessarily, $\mu_{\min}(\mathcal{D}) > 0$.

This proves that $1 - \alpha^2\, {\mu_{\min}}^2(\mathcal{D}) < 1$ and thus that $\lim_{m \to +\infty} \|R^m f\| = 0$. Inserting this in (12.54) and (12.55) proves (12.58). ■

Matching pursuits often converge more slowly when the size N of the signal space increases because $\mu_{\min}(\mathcal{D})$ can become close to 0. In the limit of infinite-dimensional spaces, Jones' theorem proves that the matching pursuit still converges but the convergence is not exponential [319, 366]. Section 12.3.2 describes an orthogonalized matching pursuit that converges in fewer than N iterations.

Backprojection

A matching pursuit computes an approximation $\tilde{f}_M = \sum_{m=0}^{M-1} \langle R^m f, \phi_{p_m} \rangle\, \phi_{p_m}$ that belongs to space $\mathbf{V}_M$ generated by M vectors $\{\phi_{p_m}\}_{0 \leqslant m < M}$. However, in general $\tilde{f}_M$ is not equal to the orthogonal projection f_M on f in $\mathbf{V}_M$, and thus $\|f - \tilde{f}_M\| \geqslant \|f - f_M\|$. In finite dimension, an infinite number of matching pursuit iterations is typically necessary to completely remove the error $\|f - \tilde{f}_M\|$, although in most applications this approximation error becomes sufficiently small for $M \ll N$. To reduce the matching pursuit error, Mallat and Zhang [366] introduced a backprojection that computes the coefficients $\tilde{a}[m]$ of the orthogonal projection

$$f_M = \sum_{m=0}^{M-1} \tilde{a}[m]\, \phi_{p_m}.$$

Let $y[m] = \langle f, \phi_{p_m} \rangle$. Section 5.1.3 shows that the decomposition coefficients of this dual-analysis problem are obtained by inverting the Gram operator

$$\tilde{a} = L^{-1} y \quad \text{with} \quad La[m] = \sum_{n=0}^{M-1} a[n] \langle \phi_{p_n}, \phi_{p_m} \rangle.$$

This inversion can be computed with a conjugate-gradient algorithm or with a Richardson gradient descent from the initial coefficients $a_0[m] = \langle R^m f, \phi_{p_m} \rangle$ provided by the matching pursuit. Let γ be a relaxation parameter that satisfies

$$\delta = \max \{|1 - \gamma A_M|, |1 - \gamma B_M|\} < 1,$$

where $B_M \geqslant A_M > 0$ are the frame bounds of $\{\phi_{p_m}\}_{0 \leqslant m < N}$ in $\mathbf{V}_M$. Theorem 5.7 proves that

$$a_k = a_{k-1} + \gamma (y - La_{k-1})$$

converges to the solution: $\lim_{k \to +\infty} a_k = \tilde{a}$. A safe choice is $\gamma = 2/B$ where $B \geqslant B_M$ is the upper frame bound of the overall dictionary $\mathcal{D}$.

Fast Network Calculations

A matching pursuit is implemented with a fast algorithm that computes $\langle R^{m+1} f, \phi_p \rangle$ from $\langle R^m f, \phi_p \rangle$ with an *updating* formula. Taking an inner product with ϕ_p on each side of (12.52) yields

$$\langle R^{m+1} f, \phi_p \rangle = \langle R^m f, \phi_p \rangle - \langle R^m f, \phi_{p_m} \rangle \langle \phi_{p_m}, \phi_p \rangle. \tag{12.62}$$

In neural network language, this is an inhibition of $\langle R^m f, \phi_p \rangle$ by the selected pattern ϕ_{p_m} with a weight $\langle \phi_{p_m}, \phi_p \rangle$ that measures its correlation with ϕ_p. To reduce the computational load, it is necessary to construct dictionaries with vectors having a sparse interaction. This means that each $\phi_p \in \mathcal{D}$ has nonzero inner products with only a small fraction of all other dictionary vectors. It can also be viewed as a network that is not fully connected. Dictionaries can be designed so that nonzero weights $\langle \phi_\alpha, \phi_p \rangle$ are retrieved from memory or computed with $O(1)$ operations. A matching pursuit with a relative precision ε is implemented with the following steps:

1. *Initialization*. Set $m = 0$ and compute $\{\langle f, \phi_p \rangle\}_{p \in \Gamma}$ in $\mathcal{D}$.
2. *Best match*. Find $\phi_{p_m} \in \mathcal{D}$ such that

$$|\langle R^m f, \phi_{p_m} \rangle| = \max_{p \in \Gamma} |\langle R^m f, \phi_p \rangle|. \tag{12.63}$$

3. *Update*. For all $\phi_p \in \mathcal{D}$ with $\langle \phi_{p_m}, \phi_p \rangle \neq 0$,

$$\langle R^{m+1} f, \phi_p \rangle = \langle R^m f, \phi_p \rangle - \langle R^m f, \phi_{p_m} \rangle \langle \phi_{p_m}, \phi_p \rangle. \tag{12.64}$$

4. *Stopping rule.* If

$$\|R^{m+1}f\|^2 = \|R^m f\|^2 - |\langle R^m f, \phi_{p_m}\rangle|^2 \leq \varepsilon^2 \|f\|^2,$$

then stop. Otherwise, $m = m + 1$ and go to 2.

If $\mathcal{D}$ is very redundant, computations at steps 1, 2, and 3 are reduced by performing the calculations in a subdictionary $\mathcal{D}_\Delta = \{\phi_p\}_{p\in\Gamma_\Delta} \subset \mathcal{D}$. The subdictionary $\mathcal{D}_\Delta$ is constructed so that if $\phi_{\tilde{p}_m} = \underset{\phi_p\in\mathcal{D}_\Delta}{\operatorname{argmax}} |\langle R^m f, \phi_p\rangle|$, then

$$|\langle R^m f, \phi_{\tilde{p}_m}\rangle| \geq \alpha \max_{p\in\Gamma} |\langle R^m f, \phi_p\rangle|. \tag{12.65}$$

The selected atom $\phi_{\tilde{p}_m} \in \mathcal{D}_\Delta$ is improved with a local search in the larger dictionary $\mathcal{D}$, among all atoms ϕ_p "close" to $\phi_{\tilde{p}_m}$, in the sense that $|\langle \phi_p, \phi_{\tilde{p}_m}\rangle| > C$ for a predefined constant C. This local search finds ϕ_{p_m}, which locally maximizes the residue correlation

$$|\langle R^m f, \phi_{p_m}\rangle| = \max_{p\in\Gamma, |\langle \phi_p, \phi_{\tilde{p}_m}\rangle| > C} |\langle R^m f, \phi_p\rangle|.$$

The updating (12.64) is restricted to vectors $\phi_p \in \mathcal{D}_\Delta$. The construction of hierarchical dictionaries can also reduce the calculations needed to compute inner products in $\mathcal{D}$ from inner products in $\mathcal{D}_\Delta$ [387].

The dictionary must incorporate important signal features, which depend on the signal class. Section 12.3.3 studies dictionaries of Gabor atoms. Section 12.3.4 describes applications to noise reduction. Specific dictionaries for inverse electromagnetic problems, face recognition, and data compression are constructed in [80, 374, 399]. Dictionary learning is studied in Section 12.7.

Wavelet Packets and Local Cosines Dictionaries

Wavelet packet and local cosine trees constructed in Sections 8.2.1 and 8.5.3 are dictionaries with $P = N \log_2 N$ vectors. For each dictionary vector, there are few other dictionary vectors having nonzero inner products that can be stored in tables to compute the updating formula (12.64). Each matching pursuit iteration then requires $O(N \log_2 N)$ operations.

In a dictionary of wavelet packet bases calculated with a Daubechies 8 filter, the best basis shown in Figure 12.14(c) optimizes the division of the frequency axis, but it has no flexibility in time. It is, therefore, not adapted to the time evolution of the signal components. The matching pursuit flexibility adapts the wavelet packet choice to local signal structures; Figure 12.14(d) shows that it better reveals its time-frequency properties than the best wavelet packet basis.

Translation Invariance

Representing a signal structure independently from its location is a form of translation invariance that is important for pattern recognition. Decompositions in

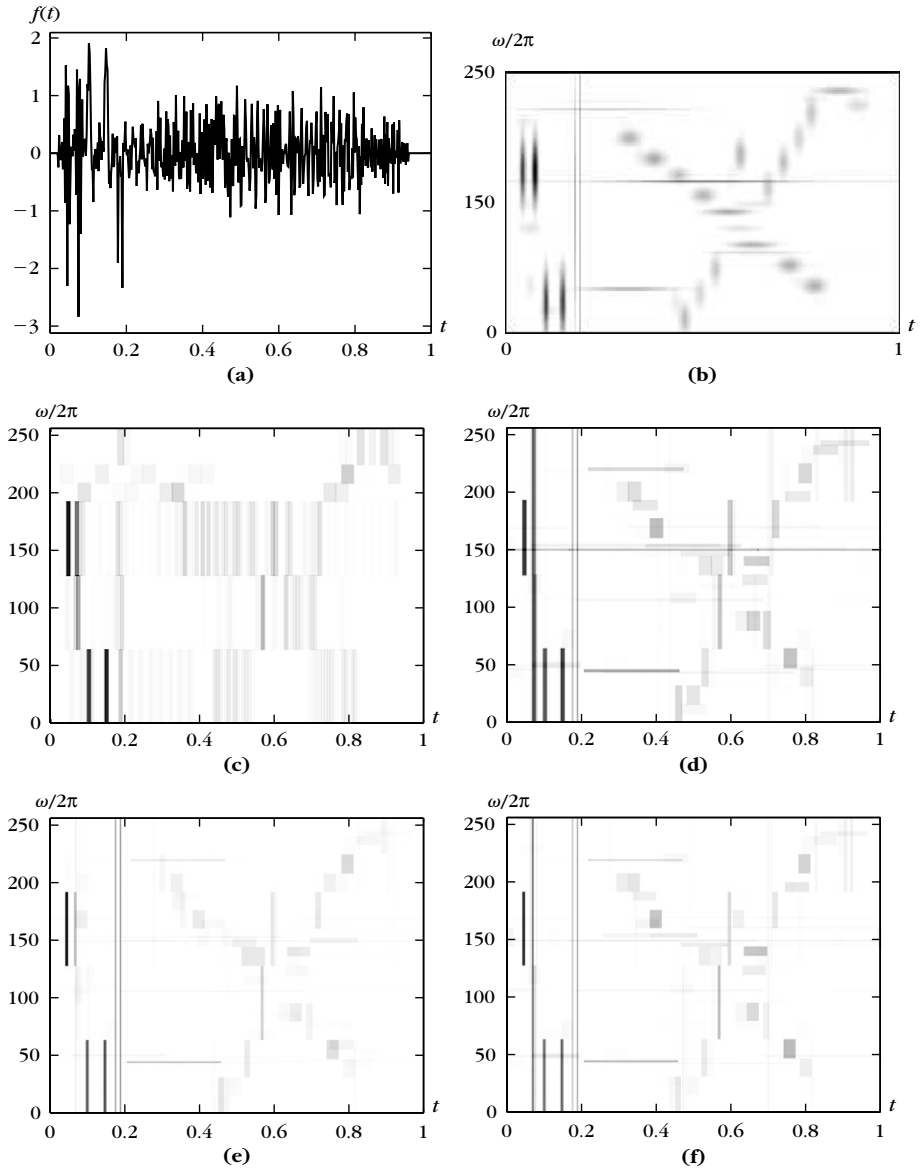


FIGURE 12.14

(a) Signal synthesized with a sum of chirps, truncated sinusoids, short time transients, and Diracs. The time-frequency images display the atoms selected by different adaptive time-frequency transforms. The darkness is proportional to the coefficient amplitude. **(b)** Gabor matching pursuit. Each dark blob is the Wigner-Ville distribution of a selected Gabor atom. **(c)** Heisenberg boxes of a best wavelet packet basis calculated with a Daubechies 8 filter. **(d)** Wavelet packets selected by a matching pursuit. **(e)** Wavelet packets of a basis pursuit. **(f)** Wavelet packets of an orthogonal matching pursuit.

orthonormal bases lack this translation invariance. Matching pursuits are translation invariant if calculated in translation-invariant dictionaries. A dictionary $\mathcal{D}$ is *translation invariant* if for any $\phi_p \in \mathcal{D}$, then $\phi_p[n-p] \in \mathcal{D}$ for $0 \leqslant p < N$. Suppose that the matching decomposition of f in $\mathcal{D}$ is [201]

$$f[n] = \sum_{m=0}^{M-1} \langle R^m f, \phi_{p_m} \rangle \, \phi_{p_m}[n] + R^M f[n]. \tag{12.66}$$

Then the matching pursuit of $f_p[n] = f[n-p]$ selects a translation by p of the same vectors ϕ_{p_m} with the same decomposition coefficients

$$f_p[n] = \sum_{m=0}^{M-1} \langle R^m f, \phi_{p_m} \rangle \, \phi_{p_m}[n-p] + R^M f_p[n].$$

Thus, patterns can be characterized independently of their position.

Translation invariance is generalized as an invariance with respect to any group action [201]. A frequency translation is another example of a group operation. If the dictionary is invariant under the action of a group, then the pursuit remains invariant under the action of the same group. Section 12.3.3 gives an example of a Gabor dictionary, which is translation invariant in time and frequency.

12.3.2 Orthogonal Matching Pursuit

Matching pursuit approximations are improved by orthogonalizing the directions of projection with a Gram-Schmidt procedure. The resulting orthogonal pursuit converges with a finite number of iterations. This orthogonalization was introduced by Mallat and Zhang together with the nonorthogonal pursuit algorithm in Zhang thesis [74]. The higher computational cost of the Gram-Schmidt algorithm may seem discouraging (reviewers suppressed it from the first publication in [366]), but the improved precision of this orthogonalization becomes important for the inverse problems studied in Chapter 13. It appeared in [202] and was proposed independently by Pati, Rezaifar, and Krishnaprasad [395].

The vector ϕ_{p_m} selected by the matching algorithm is a priori not orthogonal to the previously selected vectors $\{\phi_{p_l}\}_{0 \leqslant l < m}$. When subtracting the projection of $R^m f$ over ϕ_{p_m}, the algorithm reintroduces new components in the directions of $\{\phi_{p_l}\}_{0 \leqslant l < m}$. This is avoided by projecting the residues on an orthogonal family $\{u_l\}_{0 \leqslant l < m}$ computed from $\{\phi_{p_l}\}_{0 \leqslant l < m}$.

Let us initialize $u_0 = \phi_{p_0}$. For $m \geqslant 0$, an orthogonal matching pursuit selects ϕ_{p_m} that satisfies

$$|\langle R^m f, \phi_{p_m} \rangle| \geqslant \alpha \sup_{p \in \Gamma} |\langle R^m f, \phi_p \rangle|. \tag{12.67}$$

The Gram-Schmidt algorithm orthogonalizes ϕ_{p_m} with respect to $\{\phi_{p_l}\}_{0 \leqslant l < m}$ and defines

$$u_m = \phi_{p_m} - \sum_{l=0}^{m-1} \frac{\langle \phi_{p_m}, u_l \rangle}{\|u_l\|^2} \, u_l. \tag{12.68}$$

The residue $R^m f$ is projected on u_m instead of ϕ_{p_m}:

$$R^m f = \frac{\langle R^m f, u_m \rangle}{\|u_m\|^2} u_m + R^{m+1} f. \tag{12.69}$$

Summing this equation for $0 \leq m < k$ yields

$$\begin{aligned} f &= \sum_{m=0}^{k-1} \frac{\langle R^m f, u_m \rangle}{\|u_m\|^2} u_m + R^k f \\ &= P_{\mathbf{V}_k} f + R^k f, \end{aligned} \tag{12.70}$$

where $P_{\mathbf{V}_k}$ is the orthogonal projector on the space $\mathbf{V}_k$ generated by $\{u_m\}_{0 \leq m < k}$. The Gram-Schmidt algorithm ensures that $\{\phi_{p_m}\}_{0 \leq m < k}$ is also a basis of $\mathbf{V}_k$. For any $k \geq 0$ the residue $R^k f$ is the component of f that is orthogonal to $\mathbf{V}_k$. For $m = k$, (12.68) implies that

$$\langle R^m f, u_m \rangle = \langle R^m f, \phi_{p_m} \rangle. \tag{12.71}$$

Since $\mathbf{V}_k$ has dimension k there exists $M \leq N$, but most often $M = N$, such that $f \in \mathbf{V}_M$, so $R^M f = 0$ and inserting (12.71) in (12.70) for $k = M$ yields

$$f = \sum_{m=0}^{M-1} \frac{\langle R^m f, \phi_{p_m} \rangle}{\|u_m\|^2} u_m. \tag{12.72}$$

The algorithm stops after $M \leq N$ iterations. The energy conservation resulting from this decomposition in a family of orthogonal vectors is

$$\|f\|^2 = \sum_{m=0}^{M-1} \frac{|\langle R^m f, \phi_{p_m} \rangle|^2}{\|u_m\|^2}. \tag{12.73}$$

The exponential convergence rate of the matching pursuit in Theorem 12.6 remains valid for an orthogonal matching pursuit, but it also converges in less than N iterations.

To expand f over the original dictionary vectors $\{\phi_{p_m}\}_{0 \leq m < M}$, we must perform a change of basis. The triangular Gram-Schmidt relations (12.68) are inverted to expand u_m in $\{\phi_{p_k}\}_{0 \leq k \leq m}$:

$$u_m = \sum_{k=0}^{m} b[k, m] \, \phi_{p_k}. \tag{12.74}$$

Inserting this expression into (12.72) gives

$$f = \sum_{k=0}^{M-1} a[p_k] \, \phi_{p_k}, \tag{12.75}$$

with

$$a[p_k] = \sum_{m=k}^{M-1} b[k, m] \, \frac{\langle R^m f, \phi_{p_m} \rangle}{\|u_m\|^2}.$$

The Gram-Schmidt summation (12.68) must be carefully implemented to avoid numerical instabilities [29]. A Gram-Schmidt orthogonalization of M vectors requires $O(NM^2)$ operations. In wavelet packet, local cosine, and Gabor dictionaries, M matching pursuit iterations are calculated with $O(MN \log_2 N)$ operations.

For M large, the Gram-Schmidt orthogonalization very significantly increases the computational complexity of a matching pursuit. A final matching pursuit orthogonal backprojection requires at most $O(M^3)$ operators, but both algorithms may not give the same results because they do not necessarily select the same vectors. An orthogonal pursuit can improve the approximation precision as shown in Sections 13.3 and 13.4 for the resolution of inverse problems.

Figure 12.14(f) displays the wavelet packets selected by an orthogonal matching pursuit. A comparison with Figure 12.14(d) shows that the orthogonal and nonorthogonal pursuits select nearly the same wavelet packets having a high-amplitude inner product. These vectors are called *coherent structures*. They are selected during the first few iterations. A mathematical interpretation of these coherent structures is given in Section 12.5.2. Most often, during the first few iterations, a matching pursuit selects nearly orthogonal vectors, so orthogonal and nonorthogonal pursuits are nearly identical. When the number of iterations increases and gets close to N, the residues of an orthogonal pursuit have norms that decrease faster than for a nonorthogonal pursuit. For large-size signals, where the number of iterations is a small fraction of N, the nonorthogonal pursuit is more often used, but the orthogonalization or a backprojection becomes important if a high-approximation precision is needed.

12.3.3 Gabor Dictionaries

Gabor dictionaries are constructed with Gaussian windows, providing optimal time and frequency energy concentration. For images, directional Gabor dictionaries lead to efficient representations, particularly for video compression.

Time-Frequency Gabor Dictionary

A time and frequency translation-invariant Gabor dictionary is constructed by Qian and Chen [405] as well as Mallat and Zhang [366], by scaling, modulating, and translating a Gaussian window on the signal-sampling grid. For each scale 2^j, a discrete Gaussian window is defined by

$$g_j[n] = K_j\, 2^{-j/2+1/4}\, \exp\left(-\pi(2^{-j}n)^2\right), \tag{12.76}$$

where the constant $K_j \approx 1$ is adjusted so that $\|g_j\| = 1$. A Gabor time-frequency frame is derived with time intervals $u_j = 2^j\Delta^{-1}$ and frequency intervals $\xi_j = 2\pi\Delta^{-1}2^{-j}$:

$$\mathcal{D}_{j,\Delta} = \left\{\phi_p[n] = g_j[n - qu_j]\, \exp(i\xi_j k n)\right\}_{0\leq q<\Delta N 2^{-j}, 0\leq k<\Delta 2^j}. \tag{12.77}$$

It includes $P = \Delta^2 N$ vectors. Asymptotically for N large, this family of Gabor signals has the same properties as the frames of the Gabor functions studied in Section 5.4.

Theorem 5.19 proves that a necessary condition to obtain a frame is that $u_j\,\xi_j = 2\pi\Delta^{-2} < 2\pi$, and thus $\Delta > 1$. Table 5.3 shows that for $\Delta \geqslant 2$, this Gabor dictionary is nearly a tight frame with $A \approx B \approx \Delta^2$.

A multiscale Gabor dictionary is a union of such tight frames

$$\mathcal{D}_\Delta = \bigcup_{j=k}^{\log_2 N - k} \mathcal{D}_{j,\Delta}, \tag{12.78}$$

with typically $k \geqslant 2$ to avoid having too-small or too-large windows. Its size is thus $P \leqslant \Delta^2 N \log_2 N$, and for $\Delta \geqslant$, 2 it is nearly a tight frame with frame bounds $\Delta^2 \log_2 N$. A translation-invariant dictionary is a much larger dictionary obtained by setting $u_j = 1$ and $\xi_j = 2\pi/N$ in (12.77), and it thus includes $P \approx N^2 \log_2 N$ vectors.

A matching pursuit decomposes real signals in the multiscale Gabor dictionary (12.78) by grouping atoms ϕ_{p^+} and ϕ_{p^-} with $p^\pm = (qu_j, \pm k\xi_j, 2^j)$. At each iteration, instead of projecting $R^m f$ over an atom ϕ_p, the matching pursuit computes its projection on the plane generated by (ϕ_{p^+}, ϕ_{p^-}). Since $R^m f[n]$ is real, one can verify that this is equivalent to projecting $R^m f$ on a real vector that can be written as

$$\phi_p^\gamma[n] = K_{j,\gamma}\, g_j[n - qu_j]\, \cos\bigl(k\xi_j n + \gamma\bigr)\,.$$

The constant $K_{j,\gamma}$ sets the norm of this vector to 1 and the phase γ is optimized to maximize the inner product with $R^m f$. Matching pursuit iterations yield

$$f = \sum_{m=0}^{+\infty} \langle R^m f, \phi_{p_m}^{\gamma_m} \rangle\, \phi_{p_m}^{\gamma_m}. \tag{12.79}$$

The time-frequency signal geometry is characterized by the time-frequency and scale support $\Lambda_M = \{p_m = (q_m u_{j_m}, k\xi_{j_m}, 2^{j_m})\}_{0 \leqslant m < M}$ of the M selected Gabor atoms. It is more easily visualized with a time-frequency energy distribution obtained by summing the Wigner-Ville distribution $P_V \phi_{p_m}[n, k]$ of the complex atoms ϕ_{p_m}:

$$P_M f[n, k] = \sum_{m=0}^{+\infty} |\langle R^m f, \phi_{p_m}^{\gamma_m} \rangle|^2\, P_V \phi_{p_m}[n, k]. \tag{12.80}$$

Since the window is Gaussian, $P_V \phi_{p_m}$ is a two-dimensional Gaussian blob centered at $(q_m u_{j_m}, k_m \xi_{j_m})$ in the time-frequency plane. It is scaled by 2^{j_m} in time and by $N 2^{-j_m}$ in frequency.

Computations

A matching pursuit in a translation-invariant Gabor dictionary of size $P = N^2 \log_2 N$ is implemented by restricting most computations in the multiscale dictionary $\mathcal{D}_\Delta$ of smaller size $P \leqslant 4N \log_2 N$ for $\Delta = 2$. At each iteration, a Gabor atom $\phi_{\tilde{p}_m}$ that best matches $R^m f$ is selected in $\mathcal{D}_\Delta$. The position and frequency of this atom are

then refined with (12.65). It finds a close time-frequency atom ϕ_{p_m} in the larger translation-invariant dictionary, which has a better correlation with $R^m f$. The inner product update (12.64) is then computed only for atoms in $\mathcal{D}_\Delta$, with an analytical formula. Two Gabor atoms that have positions, frequencies, and scales $p_1 = (u_1, \xi_1, s_1)$ and $p_2 = (u_2, \xi_2, s_2)$ have an inner product

$$\langle g_{p_1}, g_{p_2} \rangle = \frac{\sqrt{2 s_1 s_2}}{\sqrt{s_1^2 + s_2^2}} \exp\left(- \frac{i\,(s_1^2 u_2 + s_2^2 u_1)(\xi_2 - \xi_1) + \pi\,(u_2 - u_1)^2}{s_1^2 + s_2^2} - \frac{(\xi_2 - \xi_1)^2}{4\pi (s_1^{-2} + s_2^{-2})} \right)$$
$$+ O\left(\exp\left(\frac{-\pi N^2}{s_1^2 + s_2^2} \right) + \exp\left(\frac{-\pi}{s_1^{-2} + s_2^{-2}} \right) \right). \tag{12.81}$$

The error terms can be neglected if the scales s_1 and s_2 are not too small or too close to N. The resulting matching pursuit in a translation-invariant Gabor dictionary requires marginally more computations than a matching pursuit in $\mathcal{D}_\Delta$.

Figure 12.14(b) shows the matching pursuit decomposition of a signal having localized time-frequency structures. This representation is more sparse than the matching pursuit decomposition in the wavelet packet dictionary shown in Figure 12.14(d). Indeed, Gabor dictionary atoms are translated on a finer time-frequency grid than wavelet packets, and they have a better time-frequency localization. As a result, the matching pursuits find Gabor atoms that better match the signal structures.

Figure 12.15 gives the Gabor matching pursuit decomposition of the word "greasy," sampled at 16 kHz. The time-frequency energy distribution shows the low-frequency component of the "g" and the quick-burst transition to the "ea." The "ea" has many harmonics that are lined up. The "s" is a noise with a time-frequency energy spread over a high-frequency interval. Most of the signal energy is characterized by a few time-frequency atoms. For $m = 250$ atoms, $\|R^m f\| / \|f\| = 0.169$, even though the signal has 5782 samples, and the sound recovered from these atoms is of good audio quality.

Matching pursuits in Gabor dictionaries provide sparse representation of oscillatory signals, with frequency and scale parameters that are used to characterize the signal structures. For example, studies have been carried in cognitive neurophysiology for the analysis of gamma and high-gamma oscillations in electroencephalogram (EEG) signals [406], which are highly nonstationary. Matching pursuit decompositions are also used to predict epilepsy patterns [22], allowing physicians to identify periods of seizure initiation by analyzing the selected atom properties [259, 320].

In Figure 12.14(b), the two chirps with frequencies that increase and decrease linearly are decomposed in many Gabor atoms. To improve the representation of signals having time-varying spectral lines, the dictionary can include Gabor chirps having an instantaneous frequency that varies linearly in time:

$$\phi_p[n] = g_j[n - q u_j]\, \exp\left(i \xi_j\, (k + c n)\, n \right).$$

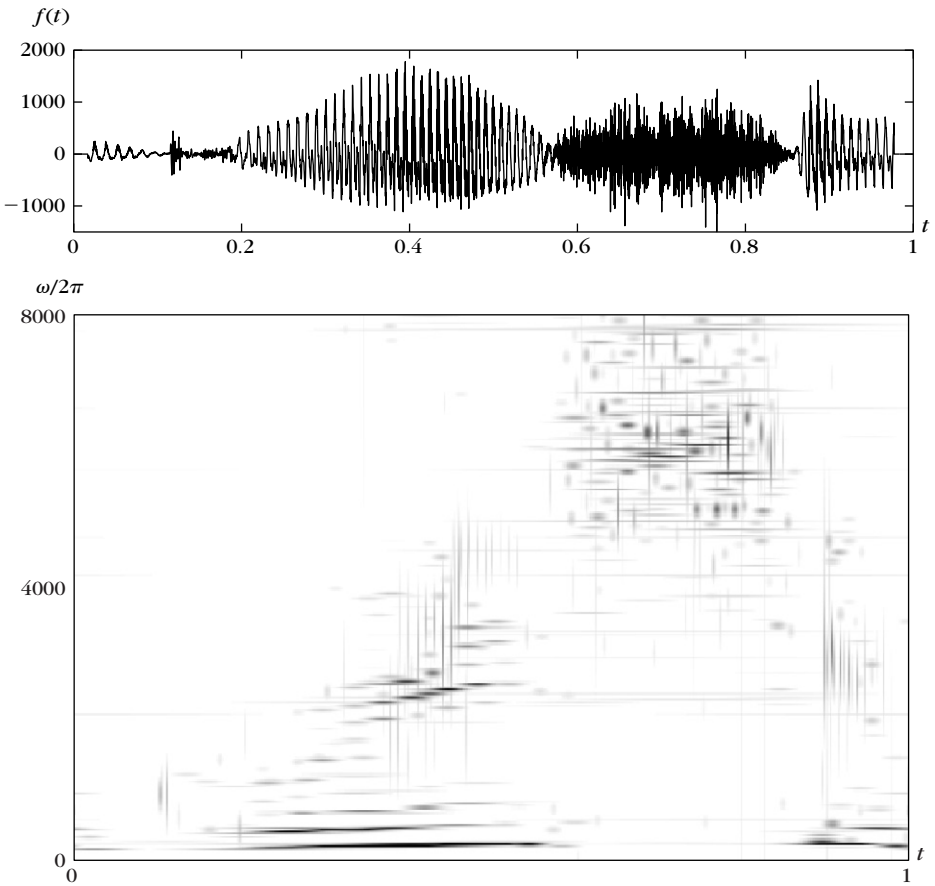


FIGURE 12.15

Speech recording of the word "greasy" sampled at 16 kHz. In the time-frequency image, the dark blobs of various sizes are the Wigner-Ville distributions of Gabor functions selected by the matching pursuit.

Their Wigner-Ville distribution $P_V\phi_P[n, k]$ is localized around an oriented segment in the time-frequency plane. Such atoms can more efficiently represent progressive frequency variations of the signal spectral components. However, increasing the dictionary size also increases intermediate memory storage and computational complexity. To incorporate Gabor chirps, Gribonval [278] reduces the matching pursuit complexity by first optimizing the scale-time-frequency parameters $(2^j, q, k)$ for $c = 0$, and then adjusting c instead of jointly optimizing all parameters.

Directional Image Gabor Dictionaries

A sparse representation of image structures such as edges, corners, and textures requires using a large dictionary of vectors. Section 5.5.1 describes redundant

dictionaries of directional wavelets and curvelets. Matching pursuit decompositions over two-dimensional directional Gabor wavelets are introduced in [105]. They are constructed with a separable product of Gaussian windows $g_j[n]$ in (12.76), with angle directions $\theta = k\pi/C$ where C is typically 4 or 8:

$$\mathcal{D}_\Delta = \left\{ g_j[n_1 - q_1 2^j \Delta^{-1}]\, g_j[n_2 - q_2 2^j \Delta^{-1}] \exp\left(-i 2^{-j} \eta (n_1 \cos\theta + n_2 \sin\theta)\right) \right\}_{q_1, q_2, j, \theta},$$

with $\Delta \geq 2$ and $\eta < 2\pi$. This dictionary is a redundant directional wavelet frame. As opposed to the frame decompositions in Section 5.5.1, a matching pursuit yields a sparse image representation by selecting a few Gabor atoms best adapted to the image.

Figure 12.16 shows the atoms selected by a matching pursuit on the Lena image. Each selected atom is displayed as an ellipse at a position $2^j \Delta^{-1}(q_1, q_2)$, of width proportional to 2^j and oriented in direction θ, with a grayscale amplitude proportional to the matching pursuit coefficient.

To better capture the anisotropic regularity of edges, more Gabor atoms are incorporated in the dictionary, with an anisotropic stretching of their support. This redundant dictionary, which includes directional wavelets and curvelets, can be applied to low-bit rate image compression [257].

Video Compression

In MPEG-1, -2, -4 video compression standards, motion vectors are coded to predict an image from a previous one, with a motion compensation [156, 437]. Figure 12.17(b) shows a prediction error image. It is the difference between the image in Figure 12.17(a) and a prediction obtained by moving pixel blocks of a previous image by using computed motion vectors. When the motion is not accurate, it yields errors along edges and sharp structures. These errors typically define oriented oscillatory structures. In MPEG compression standards, prediction error images are

(a) **(b)** **(c)**

FIGURE 12.16

Directional Gabor wavelets selected by a matching pursuit at several scales 2^j: **(a)** 2^1, **(b)** 2^2, and **(c)** 2^3. Each ellipse gives the direction, scale, and position of a selected Gabor atom.

(a) (b)

FIGURE 12.17

(a) Image of video sequences with three cars moving on a street. **(b)** Motion compensation error.

compressed with the discrete cosine transform (DCT) introduced in Section 8.3.3. The most recent MPEG-4 H.264 standard adapts the size and shape of the DCT blocks to optimize the distortion rate.

Neff and Zakhor [386] introduced a video matching pursuit compression in two-dimensional Gabor dictionaries that efficiently compresses prediction error images. Section 12.1.2 explains that an orthogonalization reduces the quantization error, but the computational complexity of the orthogonalization is too important for real-time video calculations. Compression is thus implemented with a nonorthogonal matching pursuit iteration (12.52), modified to quantize the selected inner product with $Q(x)$:

$$R^{m+1}f = R^m f - Q(\langle R^m f, \phi_{p_m}\rangle)\,\phi_{p_m}.$$

Initially implemented in a separable Gabor dictionary [386], this procedure is refined in hierarchical dictionaries providing fast algorithms for larger directional Gabor dictionaries, which improves the compression efficiency [387]. Other dictionaries reducing computations have been proposed [80, 191, 318, 351, 426], with distortion rate models to adjust the quantizer to the required bit budget [388]. This lead to a video coder, recognized in 2002 by the MPEG-4 standard expert group as having the best distortion rate with a realistic implementation among all existing solutions. However, industrial priorities have maintained a DCT solution for the new MPEG-4 standard.

12.3.4 Coherent Matching Pursuit Denoising

If we cannot interpret the information carried by a signal component, it is often misconstrued as noise. In a crowd speaking a foreign language, we perceive surrounding conversations as background noise. In contrast, our attention is easily attracted by

a remote conversation spoken in a known language. What is important here is not the information content but whether this information is in a coherent format with respect to our system of interpretation. The decomposition of a signal in a dictionary of vectors can similarly be considered as a signal interpretation. Noises are then defined as signal components that do not have a strong correlation with any vector of the dictionary. In the absence of any knowledge concerning the noise, Mallat and Zhang [366] introduced a coherent matching pursuit denoising that selects coherent structures having a high correlation with vectors in the dictionary. These coherent structures typically correspond to the approximation support of f that can be identified in $\mathcal{D}$.

Denoising by Thresholding

Let $X[n] = f[n] + W[n]$ be noisy measurements. The dictionary estimator of Theorem 12.3 in Section 12.1.3 projects X on a best set of dictionary vectors $\Lambda \subset \Gamma$, which minimizes the $\mathbf{l}^0$ Lagrangian $\|X - X_\Lambda\|^2 + T^2\,|\Lambda|$.

For an orthogonal matching pursuit approximation that selects one by one the vectors that are orthogonalized, this is equivalent to thresholding at T the resulting decomposition (12.72):

$$\tilde{F} = \sum_{m=0}^{N-1} \rho_T \left(\frac{\langle R^m X, \phi_{p_m} \rangle}{\|u_m\|^2} \right) u_m.$$

Since the amplitude of residual coefficients $|\langle R^m X, \phi_{p_m} \rangle| / \|u_m\|^2$ almost systematically decreases as m increases, it is nearly equivalent to stop the matching pursuit decomposition at the first iteration M such that $|\langle R^M X, \phi_{p_M} \rangle| / \|u_M\|^2 < T$. Thus, the threshold becomes a stopping criteria.

For a nonorthogonal matching pursuit, despite the nonorthogonality of selected coefficients, a denoising algorithm can also be implemented by stopping the decomposition (12.58) as soon as $|\langle R^m, \phi_{p_m} \rangle| < T$. The resulting estimator is

$$\tilde{F} = \sum_{m=0}^{M-1} \langle R^m X,\ \phi_{p_m} \rangle\, \phi_{p_m} \quad \text{with} \quad |\langle R^M, \phi_{p_M} \rangle| < T,$$

and can be optimized with a back projection computing the orthogonal projection of X in $\{\phi_{p_m}\}_{0 \le m < M}$. Coherent denoising provides a different approach that does not rely on a particular noise model and does not set in advance the threshold T.

Coherent Denoising

A coherent matching pursuit denoising selects signal structures having a correlation with dictionary vectors that is above an average defined over a matching pursuit attractor. A matching pursuit behaves like a nonlinear chaotic map, and it has been proved by Davis, Mallat, and Avellaneda [201] that for particular dictionaries, the normalized residues $R^m f / \|R^m f\|$ converge to an attractor. This attractor

is a set of normalized signals h that do not correlate well with any $\phi_p \in \mathcal{D}$ because all coherent structures of f producing maximum inner products with vectors in $\mathcal{D}$ are progressively removed by the pursuit. Signals on the attractor do not correlate well with any dictionary vector and are thus considered as an incoherent noise with respect to $\mathcal{D}$. The coherence of f relative to $\mathcal{D}$ is defined in (12.3.1) by $\mu(f, \mathcal{D}) = \max_{p \in \Gamma} |\langle f/\|f\|, \phi_p \rangle|$. For signals in the attractor, this coherence has a small amplitude, and we denote the average coherence of this attractor as $\bar{\mu}$, which depends on $\mathcal{D}$ [201]. This average coherence is defined by

$$\bar{\mu} = \lim_{m \to +\infty} E\{\mu(R^m W', \mathcal{D})\},$$

where W' is a Gaussian white noise of variance $\sigma^2 = 1$. The bottom regular curve C in Figure 12.18 gives the value of $E\{\mu(R^m W', \mathcal{D})\}$ that is nearly equal to $\bar{\mu} = 0.06$ for $m \geq 40$ in a Gabor dictionary.

The convergence of the pursuit to the attractor implies that for $m \geq M$ iterations, the residue $R^m f$ has a normalized correlation $\mu(R^m f, \mathcal{D})$ that is nearly equal to $\bar{\mu}$. Curve A in Figure 12.18 gives the decay of $\mu(R^m f, \mathcal{D})$ as a function of m for the "greasy" signal f in Figure 12.19(a). After about $M = 1400$ iterations, it reaches the average coherence level of the attractor. The corresponding 1400 time-frequency atoms are shown in Figure 12.15. Curve B in Figure 12.18 shows the decay of $\mu(R^m X, \mathcal{D})$ for the noisy signal $X = f + W$ in Figure 12.19(b), which has an SNR of

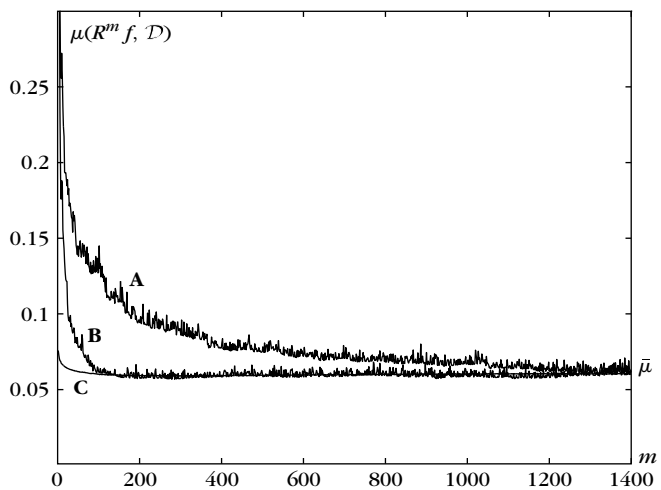


FIGURE 12.18

Decay of the correlation $\mu(R^m f, \mathcal{D})$ as a function of the number of iterations m for two signals decomposed in a Gabor dictionary. **A**: f is the recording of "greasy" shown in Figure 12.19(a). **B**: f is the noisy "greasy" signal shown in Figure 12.19(b). **C**: $E\{\mu(R^m W', \mathcal{D})\}$ that is for a normalized Gaussian white noise W'.

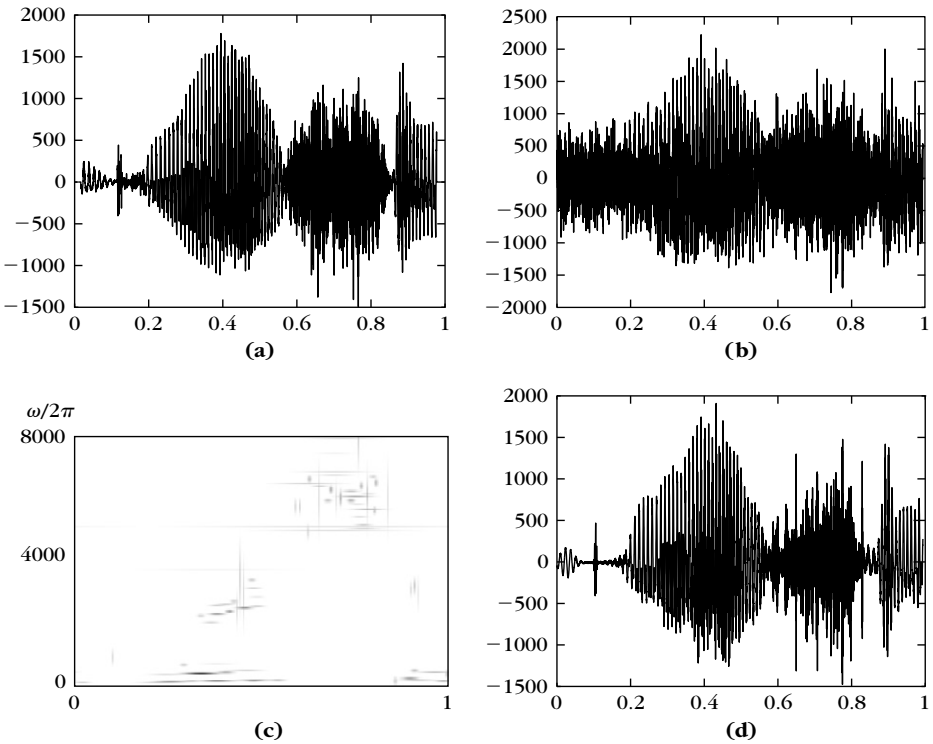


FIGURE 12.19

(a) Speech recording of "greasy." **(b)** Recording of "greasy" plus a Gaussian white noise (SNR = 1.5 db). **(c)** Time-frequency distribution of the $M = 76$ coherent Gabor structures. **(d)** Estimation $\tilde{F}$ reconstructed from the 76 coherent structures (SNR = 6.8 db).

1.5 db. The high-amplitude noise destroys most coherent structures and the attractor is reached after $M = 76$ iterations.

A coherent matching pursuit denoising with a relaxation parameter $\alpha = 1$ decomposes a signal as long as the coherence of the residue is above $\bar{\mu}$ and stops after:

$$\tilde{F} = \sum_{m=0}^{M-1} \langle R^m X, \phi_{p_m} \rangle \, \phi_{p_m} \quad \text{with} \quad \mu(R^M X, \mathcal{D}) = \frac{|\langle R^M X, \phi_{p_M} \rangle|}{\|R^M X\|} < \bar{\mu}.$$

This estimator can thus also be interpreted as a thresholding of the matching pursuit of X with a threshold that is adaptively adjusted to

$$T = \bar{\mu} \, \|R^M X\|.$$

The time-frequency energy distribution of the $M = 76$ coherent Gabor atoms of the noisy signal is shown in Figure 12.19(c). The estimation $\tilde{F}$ calculated from the

76 coherent structures is shown in Figure 12.19(d). The SNR of this estimation is 6.8 db. The white noise has been removed with no estimation of the variance, and the restored speech signal has a good intelligibility because its main time-frequency components are retained.

12.4 $\mathbf{l}^1$ PURSUITS

To reduce inefficiencies produced by the greediness of matching pursuits, $\mathbf{l}^1$ pursuits perform a more global optimization, which replaces the $\mathbf{l}^0$ norm minimization of a best M-term approximation by an $\mathbf{l}^1$ norm. The $\mathbf{l}^0$ Lagrangian studied from Section 12.1.1 is thus replaced by the $\mathbf{l}^1$ Lagrangian from Section 12.4.2. Although they are not optimal in general, Section 12.5 proves that matching pursuits and basis pursuits can compute nearly optimal M-term approximations, depending on the signal approximation support and the dictionary.

12.4.1 Basis Pursuit

Each step of a matching pursuit performs a local optimization that can be fooled. A basis pursuit minimizes a global criterion that avoids some mistakes made by greedy pursuits. A simple but typical example of failure happens when a linear combination of two vectors $f = \phi_m + \phi_q$ happens to be highly correlated with a third vector $\phi_r \in \mathcal{D}$. A matching pursuit may choose ϕ_r instead of ϕ_m or ϕ_q, and many other vectors are then needed to correct this wrong initial choice, which produces a nonoptimal representation.

Figure 12.20 illustrates this phenomenon with a dictionary $\mathcal{D}_{j,\Delta} = \{\phi_p\}_{p\in\Gamma}$ of one-dimensional Gabor atoms specified in (12.77). Each Gabor function is a Gaussian translated in time and frequency with an oversampled time-frequency grid

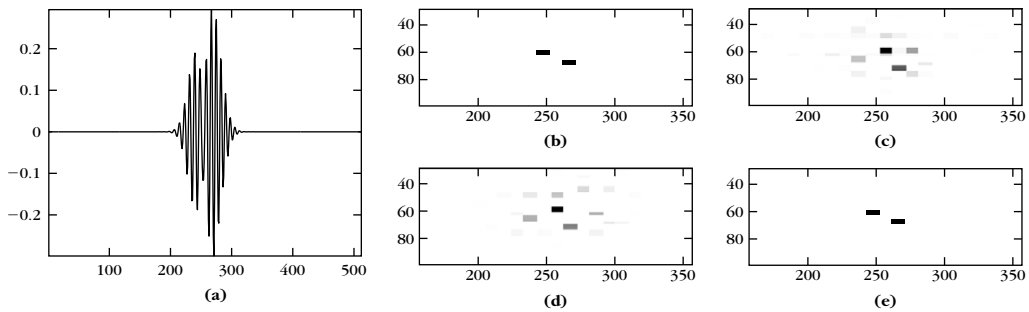


FIGURE 12.20

(a) Signal $f = \phi_m + \phi_q$. **(b)** Reduced Heisenberg boxes of the two Gabor atoms ϕ_m and ϕ_q, shown in the time-frequency plane. **(c)** Atoms selected by a matching pursuit. The darkness of each box is proportional to selected coefficients' amplitude. **(d)** Atoms selected by an orthogonal matching pursuit. **(e)** A basis pursuit recovers the two original atoms.

calculated with $\Delta = 1/4$. It has $P = 16N$ vectors of size N. Figure 12.20 shows a signal $f = \phi_m + \phi_q$ where ϕ_m and ϕ_q are two Gabor functions having nearly the same position and frequency. Let σ_t and σ_ω be the time and frequency variance of these Gabor atoms. Figure 12.20(b) represents these atoms in the time-frequency plane, with two reduced Heisenberg rectangles, of time width σ_t/Δ and frequency width σ_ω/Δ, so that all dictionary coefficients can be visualized. The full-size Heisenberg boxes $\sigma_t \times \sigma_\omega$ of ϕ_m and ϕ_q overlap widely, which makes it difficult to discriminate them in f. Figures 12.20(c, d) show that a matching pursuit and an orthogonal matching pursuit select a first time-frequency atom with a time and frequency location intermediate between these two atoms, and then other subsequent vectors to compensate for this initial mistake. Such nonoptimal greedy choices are observed on real signals decomposed in redundant dictionaries. High-resolution greedy pursuits can reduce the loss of resolution in time with nonlinear correlation measures [279, 314], but the greediness can still have adverse effects.

l^1 Minimization

Avoiding this greediness suboptimality requires using a global criterion that enforces the sparsity of the decomposition coefficients of f in $\mathcal{D} = \{\phi_p\}_{p\in\Gamma}$. Let $\Phi f[p] = \langle f, \phi_p \rangle$ be the decomposition operator in $\mathcal{D}$. The reconstruction from dictionary vectors is implemented by the adjoint (5.3)

$$f[n] = \Phi^* a[n] = \sum_{p=0}^{P-1} a[p]\, \phi_p[n]. \tag{12.82}$$

Since the dictionary is redundant, there are many possible reconstructions. The basis pursuit introduced by Chen and Donoho [159] finds the vector $\tilde{a}$ of coefficients having a minimum $\mathbf{l}^1$ norm

$$\tilde{a} = \underset{a\in\mathbb{R}^P}{\operatorname{argmin}} \|a\|_1 \quad \text{subject to} \quad \Phi^* a = f. \tag{12.83}$$

This is a convex minimization that can be written as a linear programming and is thus calculated with efficient algorithms, although computationally more intensive than a matching pursuit. If the solution of (12.83) is not unique, then any valid solution may be used.

The signal in Figure 12.20 can be written exactly as a sum of two dictionary vectors and the basis pursuit minimization recovers this best representation by selecting the appropriate dictionary vectors, as opposed to matching pursuit algorithms. Minimizing the $\mathbf{l}^1$ norm of the decomposition coefficients $a[p]$ avoids cancellation effects when selecting an inappropriate vector in the representation, which is then canceled by other redundant dictionary vectors. Indeed, these cancellations increase the $\mathbf{l}^1$ norm of the resulting coefficients. As a result, the global optimization of a basis pursuit can provide a more accurate representation of sparse signals than matching pursuits for highly correlated and redundant dictionaries. This is further studied in Section 12.5.

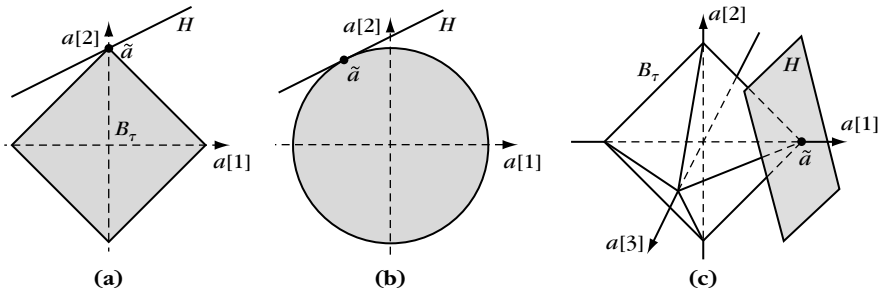


FIGURE 12.21

(a, b) Comparison of the minimum $\mathbf{l}^1$ and $\mathbf{l}^2$ solutions of $\Phi^* a = f$ for $P = 2$. **(c)** Geometry of $\mathbf{l}^1$ minimization for $P = 3$. Solution $\tilde{a}$ typically has fewer nonzero coefficients for $\mathbf{l}^1$ than for $\mathbf{l}^2$.

Sparsity

A geometric interpretation of basis pursuit also explains why it can recover sparse solutions. For a dictionary of size P the decomposition coefficients $a[p]$ define a vector in $a \in \mathbb{R}^P$. Let H be the affine subspace of $\mathbb{R}^P$ of coordinate vectors that recover $f \in \mathbb{R}^N$,

$$H = \left\{ a \in \mathbb{R}^P : \Phi^* a = f \right\} = a_0 + \mathbf{Null}(\Phi^*) \subset \mathbb{R}^P, \quad \text{where} \quad \Phi^* a_0 = f. \tag{12.84}$$

The dimension of H is $P - N$. A basis pursuit (12.83) finds in H an element $\tilde{a}$ of minimum $\mathbf{l}^1$ norm. It can be found by inflating the $\mathbf{l}^1$ ball

$$B_\tau = \left\{ a \in \mathbb{R}^P : \|a\|_1 \leqslant \tau \right\} \subset \mathbb{R}^P, \tag{12.85}$$

by increasing τ until it intersects H. This geometric configuration is depicted for $P = 2$ and $P = 3$ in Figure 12.21.

The $\mathbf{l}^1$ ball remains closer to the coordinate axes of $\mathbb{R}^P$ than the $\mathbf{l}^2$ ball. When the dimension P increases, the volume of the $\mathbf{l}^1$ ball becomes much smaller than the volume of the $\mathbf{l}^2$ ball. Thus, the optimal solution $\tilde{a}$ is likely to have more zeros or coefficients close to zero when it is computed by minimizing an $\mathbf{l}^1$ norm rather than an $\mathbf{l}^2$ norm. This is illustrated by Figure 12.21.

Basis Pursuit and Best-Basis Selection

Theorem 12.7 proves that a basis pursuit selects vectors that are independent, unless it is a degenerated case where the solution is not unique, which happens rarely. We denote by $\tilde{\Lambda} = \{ p : \tilde{a}[p] \neq 0 \}$ the support of $\tilde{a}$.

Theorem 12.7. A basis pursuit (12.83) admits a solution $\tilde{a}$ with support $\tilde{\Lambda}$ that corresponds to a family $\{\phi_p\}_{p \in \tilde{\Lambda}}$ of linearly independent dictionary vectors.

Proof. If $\{\phi_p\}_{p \in \tilde{\Lambda}}$ is linearly dependent, then there exists $h \in \mathbf{Null}(\Phi^*)$ with $h \neq 0$ and $h[p] = 0$ for $p \notin \tilde{\Lambda}$. For λ small enough such that $\mathrm{sign}(\tilde{a} + \lambda h) = \mathrm{sign}(\tilde{a})$, the mapping $\lambda \mapsto \|\tilde{a} + \lambda h\|_1$

is locally affine until at least one of the components of $\tilde{a}+\lambda h$ vanishes. Since $\|\tilde{a}\|_1$ is minimum, $\|\tilde{a}+\lambda h\|_1$ is constant for λ small enough, and thus $\|\tilde{a}+\lambda h\|_1 = \|\tilde{a}\|_1$ for all such λ. The minimization of $\|a\|_1$ with $\Phi^* a = f$ is therefore nonunique.

Furthermore, for a critical value of λ, one of the components of $\tilde{a}+\lambda h$ vanishes. The support of $\tilde{a}+\lambda h$ is strictly included in $\tilde{\Lambda}$ and $\|\tilde{a}+\lambda h\|_1$ is minimum. Setting $\tilde{a}_1 = \tilde{a}+\lambda h$ and iterating this argument shows that there exists a solution supported inside $\tilde{\Lambda}$ that indexes vectors that are linearly independent. ■

Signals of size N can rarely be written exactly as a sum of less than N dictionary vectors, and the N independent vectors selected by a basis pursuit thus define a basis of $\mathbb{C}^N$. A basis pursuit can therefore be interpreted as a *best-basis* algorithm. Among all possible bases of $\mathcal{D}$, it selects a basis $\mathcal{B} = \{\phi_{p_m}\}_{0 \leq m < N}$, which yields decomposition coefficients $\{a[p_m]\}_{0 \leq m < N}$ of minimum $\mathbf{l}^1$ norm. Unlike the best-basis selection algorithm in Section 12.2.2, it does not restrict the search to orthonormal bases, which provides much more flexibility.

Signal denoising or compression applications can be implemented by thresholding or quantizing the decomposition coefficients of a basis pursuit. However, there is no control on the stability of the selected basis $\mathcal{B}$. The potential instabilities of the basis do not provide a good control on the resulting error, but results are typically slightly better than with a matching pursuit.

Linear Programming for the Resolution of Basis Pursuit

The basis pursuit minimization of (12.83) is a convex optimization problem that can be reformulated as a linear programming problem. A standard-form linear programming problem [28] is a constrained optimization over positive vectors $d[p]$ of size L. Let $b[n]$ be a vector of size $N < L$, $c[p]$ a nonzero vector of size L, and $A[n,p]$ an $L \times N$ matrix. A linear programming problem finds $d[p] \in \mathbb{R}^L$ such that $d[p] \geqslant 0$, which is the solution of the minimization problem

$$\tilde{d} = \underset{d \in (\mathbb{R}^+)^L}{\operatorname{argmin}} \sum_{p=0}^{L-1} d[p]\, c[p] \quad \text{subject to} \quad Ad = b. \tag{12.86}$$

The basis pursuit optimization

$$\tilde{a} = \underset{a \in \mathbb{R}^P}{\operatorname{argmin}} \|a\|_1 \quad \text{subject to} \quad \Phi^* a = f \tag{12.87}$$

is recast as linear programming by introducing slack variables $u[p] \geqslant 0$ and $v[p] \geqslant 0$ such that $a = u - v$. One can then define

$$A = (\Phi^*, -\Phi^*) \in \mathbb{R}^{N \times 2P} \quad c = 1, \quad d = (u, v) \in \mathbb{R}^{2P}, \text{ and } \quad b = f.$$

Since

$$\|a\|_1 = \sum_{p=0}^{L-1} d[p]\, c[p] \quad \text{and} \quad Ad = \Phi^* u - \Phi^* v = f,$$

this shows that (12.87) is written as a linear programming problem (12.86) of size $L = 2P$. The matrix A of size $N \times L$ has rank N because the dictionary $\mathcal{D}$ includes N linearly independent vectors.

The collection of feasible points $\{d : Ad = b,\ d \geqslant 0\}$ is a convex polyhedron in $\mathbb{R}^L$. Theorem 12.7 proves there exists a solution of the linear programming problem with at most N nonzero coefficients. The linear cost (12.86) can thus be minimized over the subpolyhedron of vectors having N nonzero coefficients. These N nonzero coefficients correspond to N column vectors $\mathcal{B} = \{\phi_{p_m}\}_{0 \leqslant m < N}$ that form a basis.

One can also prove [28] that if the cost is not minimum at a given vertex, then there exists an adjacent vertex with a cost that is smaller. The simplex algorithm takes advantage of this property by jumping from one vertex to an adjacent vertex while reducing the cost (12.86). Going to an adjacent vertex means that one of the zero coefficients of $d[p]$ becomes nonzero while one nonzero coefficient is set to zero. This is equivalent to modifying the basis $\mathcal{B}$ by replacing one vector by another vector of $\mathcal{D}$. The simplex algorithm thus progressively improves the basis by appropriate modifications of its vectors, one at a time. In the worst case, all vertices of the polyhedron will be visited before finding the solution, but the average case is much more favorable.

Since the 1980s, more effective interior point procedures have been developed. Karmarkar's interior point algorithm [325] begins in the middle of the polyhedron and converges by iterative steps toward the vertex solution, while remaining inside the convex polyhedron. For finite precision calculations, when the algorithm has converged close enough to a vertex, it jumps directly to the corresponding vertex, which is guaranteed to be the solution. The middle of the polyhedron corresponds to a decomposition of f over all vectors of $\mathcal{D}$, typically with $P > N$ nonzero coefficients. When moving toward a vertex some coefficients progressively decrease while others increase to improve the cost (12.86). If only N decomposition coefficients are significant, jumping to the vertex is equivalent to setting all other coefficients to zero. Each step requires computing the solution of a linear system. If A is an $N \times L$, matrix, then Karmarkar's algorithm terminates with $O(L^{3.5})$ operations. Mathematical work on interior point methods has led to a large variety of approaches that are summarized in [355].

Besides linear programming, let us also mention that simple iterative algorithms can also be implemented to compute the basis pursuit solution [184].

Application to Wavelet Packet and Local Cosine Dictionaries

Dictionaries of wavelet packets and local cosines have $P = N \log_2 N$ time-frequency atoms. A straightforward implementation of interior point algorithms thus requires $O(N^{3.5} \log_2^{3.5} N)$ operations. By using the fast wavelet packet and local cosine transforms together with heuristic computational rules, the number of operations is considerably reduced [158]. The algorithm still remains computationally intense.

Figure 12.14 is an example of a synthetic signal with localized time-frequency structures, decomposed on a wavelet packet dictionary. The flexibility of a basis pursuit decomposition in Figure 12.14(e) gives a much more sparse representation

than a best orthogonal wavelet packet basis in Figure 12.14(c). In this case, the resulting representation is very similar to the matching pursuit decompositions in Figures 12.14(d, f). Figure 12.20 shows that a basis pursuit can improve matching pursuit decompositions when the signal includes close time-frequency structures that are not distinguished by a matching pursuit that selects an intermediate time-frequency atom [158].

12.4.2 l^1 Lagrangian Pursuit

Compression and denoising applications of basis pursuit decompositions create sparse representations by quantizing or thresholding the resulting coefficients. This is not optimal because the selected basis $\mathcal{B} \subset \mathcal{D}$ is not orthogonal and may even be unstable. It is then more efficient to directly solve the sparse denoising or approximation problem with a Lagrangian approach.

Basis Pursuit Approximation and Denoising

To suppress an additive noise or to approximate a signal with an error ε, instead of thresholding the coefficients obtained with a basis pursuit, Chen, Donoho, and Saunders [159] compute the solution

$$\tilde{f} = \sum_{p=0}^{P-1} \tilde{a}[p]\, \phi_p = \Phi^* \tilde{a}$$

with decomposition coefficients $\tilde{a}$ that are a solution of a minimization problem that incorporates the precision parameter ε:

$$\tilde{a} = \underset{a \in \mathbb{R}^P}{\operatorname{argmin}} \|a\|_1 \quad \text{subject to} \quad \|\Phi^* a - f\|^2 \leqslant \varepsilon. \tag{12.88}$$

In a denoising problem, f is replaced by the noisy signal $X = f + W$ where W is the additive noise. It is then called a *basis pursuit denoising* algorithm.

Computing the solution of this quadratically constrained problem is more complicated than the linear programming problem corresponding to a basis pursuit. However, it can be recast as a second-order cone program, which is solved using interior point methods and, in particular, log-barrier methods [10] that extend the interior point algorithms for linear programming problems. These general-purpose algorithms can also be optimized to take into account the separability of the $\mathbf{l}^1$ norm. Iterative algorithms converging to the solution have also been developed [183].

The minimization problem (12.88) is convex and thus can also be solved through its Lagrangian formulation:

$$\tilde{a} = \underset{a \in \mathbb{C}^P}{\operatorname{argmin}}\, \mathcal{L}_1(T, f, a) = \underset{a \in \mathbb{C}^P}{\operatorname{argmin}} \frac{1}{2} \|f - \Phi^* a\|^2 + T \|a\|_1. \tag{12.89}$$

In the following, this Lagrangian minimization will be called a *Lagrangian basis pursuit* or *$\mathbf{l}^1$ pursuit*. For each $\varepsilon > 0$, there exists a Lagrangian multiplier T such that convex minimization (12.88) and the Lagrangian minimization (12.89) have

a common solution [266, 463]. In a denoising problem, where f is replaced by $X = f + W$, it is easier to adjust T then ε to the noise level. Indeed, we shall see that a Lagrangian pursuit performs a generalized soft thresholding by T. For a Gaussian white noise σ, one can typically set $T = \lambda\,\sigma$ with $\lambda \leqslant \sqrt{2\log_e P}$, where P is the dictionary size. Section 12.4.3 describes algorithms for computing a solution of the Lagrangian minimization (12.89).

Figure 12.22 shows an example of basis pursuit denoising of an image contaminated by a Gaussian white noise. The dictionary includes a translation-invariant dyadic wavelet frame and a tight frame of local cosine vectors with a redundancy factor of 16. The resulting estimation is better than in a translation-invariant wavelet dictionary. Indeed, local cosine vectors provide more sparse representations of the image oscillatory textures.

Convexification of l^0 with l^1

Theorem 12.1 proves that a best M-term approximation f_{Λ_T} in a dictionary $\mathcal{D}$ is obtained by minimizing the $\mathbf{l}^0$ Lagrangian (12.30)

$$\mathcal{L}_0(T, f, \Lambda) = \|f - f_\Lambda\|^2 + T^2\,|\Lambda| = \|f - \sum_{p\in\Gamma} a[p]\phi_p\|^2 + T^2\,\|a\|_0. \qquad (12.90)$$

Since the $\mathbf{l}^0$ pseudo norm is not convex, Section 12.1.1 explains that the minimization of $\mathcal{L}_0$ is intractable. An $\mathbf{l}^0$ norm can be approximated by an $\mathbf{l}^q$ pseudo-norm

$$\|a\|_q = \left(\sum_{p\in\Lambda} |a[p]|^q\right)^{1/q},$$

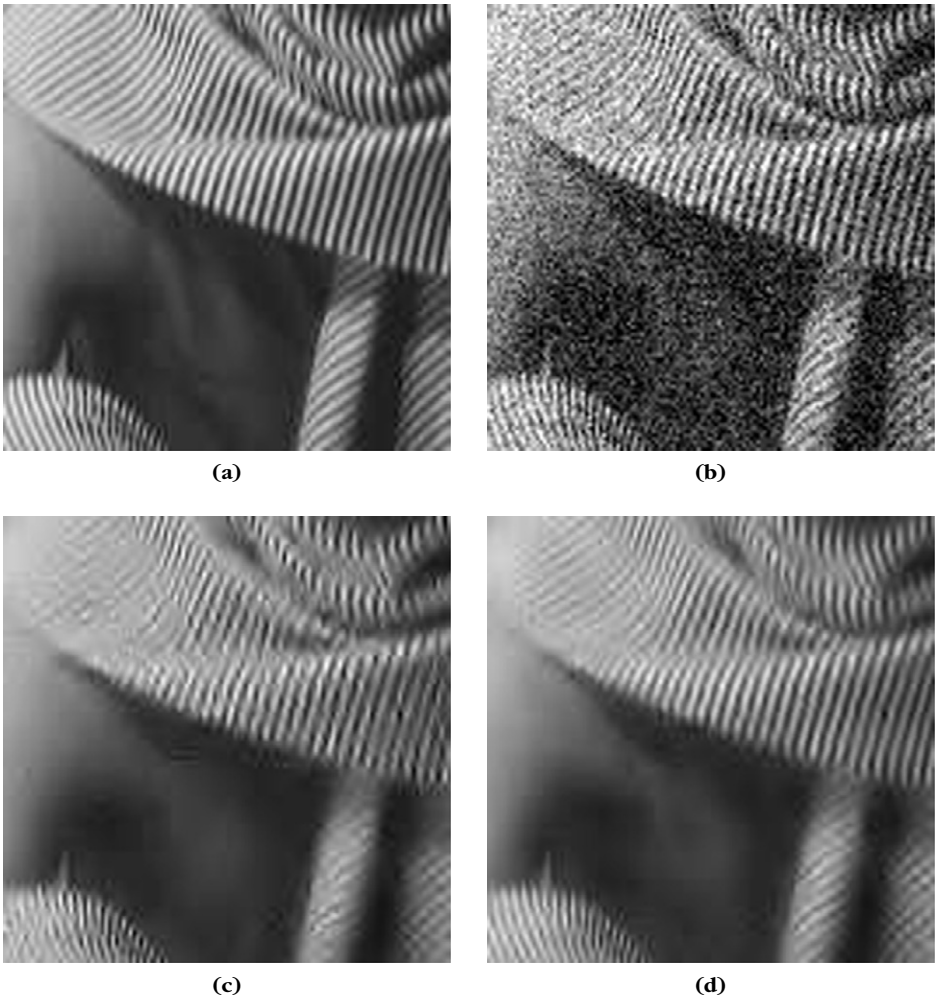
which is nonconvex for $0 \leqslant q < 1$, and convex and thus a norm for $q \geqslant 1$. As q decreases, Figure 12.23 shows in $P = 2$ dimensions that the unit ball of vectors $\|a\|_q \leqslant 1$ approaches the $\mathbf{l}^0$ unit "ball," which corresponds to the two axes. The $\mathbf{l}^1$ Lagrangian minimization (12.89) can thus be interpreted as a convexification of the $\mathbf{l}^0$ Lagrangian minimization.

Let $\tilde{\Lambda} = \{p \in \Gamma : \tilde{a}[p] \neq 0\}$ be the support of the $\mathbf{l}^1$ Lagrangian pursuit solution $\tilde{a}$. For $|\tilde{\Lambda}| = M$, the support $\tilde{\Lambda}$ is typically not equal to the best M-term approximation support Λ_T obtained by minimizing the $\mathbf{l}^0$ Lagrangian and $\|f - \tilde{f}\| \geqslant \|f - f_{\Lambda_T}\|$. However, Section 12.5.3 proves that $\tilde{\Lambda}$ may include the main approximation vectors of Λ_T and $\|f - \tilde{f}\|$ can be of the same order as $\|f - f_{\Lambda_T}\|$ if Λ_T satisfies an exact recovery condition that is specified.

Generalized Soft Thresholding

A Lagrangian pursuit computes a sparse approximation $\tilde{f} = \Phi^* \tilde{a}$ of f by minimizing

$$\tilde{a} = \underset{a\in\mathbb{C}^P}{\operatorname{argmin}}\, \mathcal{L}_1(T, f, a) \quad \text{where} \quad \mathcal{L}_1(T, f, a) = \frac{1}{2}\|f - \Phi^* a\|^2 + T\|a\|_1. \qquad (12.91)$$



(a) **(b)** **(c)** **(d)**

FIGURE 12.22

(a) Original image f. **(b)** Noisy image $X = f + W$ (SNR = 12.5 db). **(c)** Translation-invariant wavelet denoising, (SNR = 18.6 db). **(d)** Basis pursuit denoising in a dictionary that is a union of a translation-invariant wavelet frame and a frame of redundant local cosine vectors (SNR = 19.8 db).

Increasing T often reduces the support $\tilde{\Lambda}$ of $\tilde{a}$ but increases the approximation error $\|f - \Phi^* \tilde{a}\|$. The restriction of the dictionary operator and its adjoint to $\tilde{\Lambda}$ is written as

$$\Phi_{\tilde{\Lambda}} = \{ \langle f, \phi_p \rangle \}_{p \in \tilde{\Lambda}} \quad \text{and} \quad \Phi^*_{\tilde{\Lambda}} a = \sum_{p \in \tilde{\Lambda}} a[p] \, \phi_p .$$

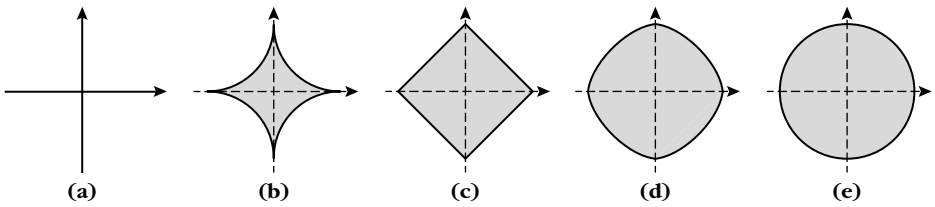


FIGURE 12.23

Unit balls of l^q functionals: **(a)** $q = 0$, **(b)** $q = 0.5$, **(c)** $q = 1$, **(d)** $q = 1.5$, and **(e)** $q = 2$.

Theorem 12.8, proved by Fuchs [266], gives necessary and sufficient conditions on $\tilde{a}$ and its support $\tilde{\Lambda}$ to be a minimizer of $\mathcal{L}_1(T, f, a)$.

Theorem 12.8: *Fuchs.* A vector $\tilde{a}$ is a minimum of $\mathcal{L}_1(T, f, a)$ if and only if there exists $h \in \mathbb{R}^P$ such that

$$\Phi(\Phi^* \tilde{a} - f) + Th = 0 \quad \text{where} \quad \begin{cases} h[p] = \text{sign}(\tilde{a}[p]) & \text{if} \quad \tilde{a}[p] \neq 0 \\ |h[p]| \leqslant 1 & \text{if} \quad \tilde{a}[p] = 0. \end{cases} \tag{12.92}$$

There exists a solution $\tilde{a}$ with support $\tilde{\Lambda}$ that corresponds to a family $\{\phi_p\}_{p \in \tilde{\Lambda}}$ of linearly independent dictionary vectors. The restriction $\tilde{a}_{\tilde{\Lambda}}$ over its support satisfies

$$\tilde{a}_{\tilde{\Lambda}} + T(\Phi_{\tilde{\Lambda}}\, \Phi_{\tilde{\Lambda}}^*)^{-1} \text{sign}(\tilde{a}_{\tilde{\Lambda}}) = \Phi_{\tilde{\Lambda}}^{*\,+} f\ , \tag{12.93}$$

where $\Phi_{\tilde{\Lambda}}^{*\,+} f = (\Phi_{\tilde{\Lambda}}\, \Phi_{\tilde{\Lambda}}^*)^{-1} \Phi_{\tilde{\Lambda}}$ is the pseudo inverse of $\Phi_{\tilde{\Lambda}}^*$.

Proof. If $\tilde{a}[p] \neq 0$, then $\mathcal{L}_1(T, f, a)$ is differentiable along the coordinate $a[p]$ at the point $\tilde{a}[p]$. Setting this derivative to 0 shows that $\tilde{a}$ is minimum if and only if

$$\langle \phi_p,\ \Phi^* \tilde{a} - f \rangle + T\, \text{sign}(\tilde{a}[p]) = 0, \quad \text{so} \quad h[p] = \text{sign}(\tilde{a}[p]). \tag{12.94}$$

If $\tilde{a}[p] = 0$, let us consider the vector $a[q] = \tilde{a}[q] + \tau\, \delta[q - p]$ for $\tau \in \mathbb{R}$. The corresponding Lagrangian value is

$$\mathcal{L}_1(T, f, a) = \mathcal{L}_1(T, f, \tilde{a}) + \tau \langle \phi_p,\ \Phi^* \tilde{a} - f \rangle + \frac{\tau^2}{2} + T|\tau| \geqslant \mathcal{L}_1(\tilde{a}).$$

Since $\tilde{a}$ is a minimizer

$$\forall\, \tau, \quad \tau \langle \phi_p,\ \Phi^* \tilde{a} - f \rangle + T|\tau| + \frac{\tau^2}{2} \geqslant 0.$$

By separately considering the cases $\tau > 0$, and $\tau < 0$, we verify that when τ goes to zero it implies that

$$|\langle \phi_p,\ \Phi^* \tilde{a} - f \rangle| \leqslant T. \tag{12.95}$$

Conditions (12.94) and (12.95) are equivalent to (12.92).

Conversely, if h satisfies (12.92), then for any a we verify that

$$|a[p]| \geqslant |\tilde{a}[p]| + h[p](a[p] - \tilde{a}[p]),$$

and thus

$$\|a\|_1 \geqslant \|\tilde{a}\|_1 + \sum_p h[p](a[p] - \tilde{a}[p]) = \|\tilde{a}\|_1 + \langle h,\ a - \tilde{a}\rangle.$$

Since $\|\Phi^* \tilde{a} - f\|^2$ is differentiable and convex, this leads to

$$\mathcal{L}_1(T, f, a) \geqslant \frac{1}{2}\|\Phi^* \tilde{a} - f\|^2 + \langle a - \tilde{a},\ \Phi(\Phi^* \tilde{a} - f)\rangle + T\left(\|\tilde{a}\|_1 + \langle h,\ a - \tilde{a}\rangle\right)$$
$$\geqslant \mathcal{L}_1(T, f, \tilde{a}) + \langle a - \tilde{a},\ \Phi(\Phi^* \tilde{a} - f) + Th\rangle = \mathcal{L}_1(T, f, \tilde{a})$$

because of (12.92), and thus $\tilde{a}$ minimizes $\mathcal{L}_1(T, f, a)$. The conditions (12.92) therefore are necessary and sufficient.

The proof of the existence of a solution $\tilde{a}$ corresponding to linear independent vectors $\{\phi_p\}_{p \in \Lambda(\tilde{a})}$ is identical to the proof of Theorem 12.7 in the basis pursuit context. Over the support $\tilde{\Lambda}$, the necessary and sufficient condition (12.92) can be rewritten as

$$\Phi_{\tilde{\Lambda}}(\Phi^*_{\tilde{\Lambda}} \tilde{a}_{\tilde{\Lambda}} - f) + T\,\mathrm{sign}(\tilde{a}_{\tilde{\Lambda}}) = 0,$$

which implies (12.93). ■

Let $f_{\tilde{\Lambda}} = \Phi^*_{\tilde{\Lambda}} a_{\tilde{\Lambda}}$ be the orthogonal projection of f over the space $\mathbf{V}_{\tilde{\Lambda}}$ generated by $\{\phi_p\}_{\tilde{\Lambda}}$. Its coefficients are $a_{\tilde{\Lambda}} = \Phi^{*\,+}_{\tilde{\Lambda}} f_{\tilde{\Lambda}} = \Phi^{*\,+}_{\tilde{\Lambda}} f$. Theorem 12.8 proves in (12.93) that

$$\tilde{a}_{\tilde{\Lambda}} = a_{\tilde{\Lambda}} - T(\Phi_{\tilde{\Lambda}} \Phi^*_{\tilde{\Lambda}})^{-1}\mathrm{sign}(\tilde{a}_{\tilde{\Lambda}}), \tag{12.96}$$

which shows that the $\mathbf{l}^1$ minimization attenuates by an amount proportional to T the amplitude of the coefficients $a_{\tilde{\Lambda}}$ of the orthogonal projection $f_{\tilde{\Lambda}}$. This can be interpreted as a generalized soft thresholding. If $\mathcal{D}$ is an orthonormal basis, then $\Phi\Phi^* = \mathrm{Id}$ and $\tilde{a}$ is a classical soft thresholding of the coefficient Φf of f in this orthonormal basis.

Since $\tilde{f} = \Phi^* \tilde{a} \in \mathbf{V}_{\tilde{\Lambda}}$, we know that $\|f - \tilde{f}\| \geqslant \|f - f_{\tilde{\Lambda}}\|$. Once the optimal Lagrangian support $\tilde{\Lambda}$ is recovered, similar to the matching pursuit backprojection, the sparse approximation $\tilde{f} = \Phi^*_{\tilde{\Lambda}} \tilde{a}_{\tilde{\Lambda}}$ of f can be improved by computing the orthogonal projection $f_{\tilde{\Lambda}}$ and its coefficients $a_{\tilde{\Lambda}}$.

12.4.3 Computations of $\mathbf{l}^1$ Minimizations

The relaxed formulation of an $\mathbf{l}^1$ Lagrangian pursuit

$$\tilde{a} = \underset{a \in \mathbb{C}^P}{\mathrm{argmin}}\ \mathcal{L}_1(T, f, a) = \underset{a \in \mathbb{C}^P}{\mathrm{argmin}}\ \frac{1}{2}\|f - \Phi^* a\|^2 + T\|a\|_1 \tag{12.97}$$

cannot be formulated as a linear program, as opposed to the basis pursuit minimization (12.83). Several approaches have been developed to compute this minimization.

Iterative Thresholding

Several authors [100, 185, 196, 241, 255, 466] have proposed an iterative algorithm that solves (12.89) with a soft thresholding to decrease the $\mathbf{l}^1$ norm of the

coefficients a, and a gradient descent to decrease the value of $\|f - \Phi^* a\|$. The coefficient vector a may be complex, and $|a[p]|$ is then the complex modulus.

1. *Initialization*. Choose a_0 (e.g., 0), set $k = 0$, and compute $b = \Phi f$.
2. *Gradient step*. Update
$$\bar{a}_k = a_k + \gamma\,(b - \Phi\,\Phi^* a_k), \tag{12.98}$$
where $\gamma < 2\,\|\Phi\Phi^*\|_S^{-1}$.
3. *Soft thresholding*. Compute
$$a_{k+1}[p] = \rho_{\gamma T}(\bar{a}_k[p]) \quad \text{where} \quad \rho_{\gamma T}(x) = x\,\max\left(1 - \frac{\gamma T}{|x|},\, 0\right). \tag{12.99}$$
4. *Stop*. If $\|a_k - a_{k+1}\|$ is smaller than a fixed-tolerance criterion, stop the iterations, otherwise set $k \leftarrow k+1$ and go back to 2.

This algorithm includes the same gradient step as the Richardson iteration algorithm in Theorem 5.7, which inverts the symmetric operator $L = \Phi\Phi^*$. The convergence condition $\gamma < 2\,\|\Phi\Phi^*\|_S^{-1}$ is identical to the convergence condition (5.35) of the Richardson algorithm. Theorem 12.9 proves that the convergence is guaranteed for any a_0.

Theorem 12.9. The sequence a_k defined by
$$a_{k+1}[p] = \rho_{\gamma T}\left(a_k[p] + \gamma\,\Phi(f - \Phi^* a_k)[p]\right) \quad \text{with} \quad \gamma < 2\,\|\Phi\,\Phi^*\|_S^{-1} \tag{12.100}$$
converges to a solution of (12.89) for any $a_0 \in \mathbb{C}^P$.

Proof. The following proof is due to Daubechies, Defries, and DeMol [196], showing that a_k converges to a minimum of the energy
$$\mathcal{L}_1(T, f, a) = \frac{1}{2}\|\Phi^* a - f\|^2 + T\|a\|_1 \tag{12.101}$$
for $\gamma < \|\Phi\Phi^*\|_S^{-1}$. A proof of convergence for $\gamma < 2\,\|\Phi\Phi^*\|_S^{-1}$ can be found in [185].

To simplify notations, the dependencies of the Lagrangian on T and f are dropped, and it is written $\mathcal{L}_1(a)$. Lemma 12.2 proves that a_{k+1} is the minimum of a modified surrogate Lagrangian $\tilde{\mathcal{L}}_1(a, a_k)$ that approximates $\mathcal{L}_1(a)$ and depends on the previous iterate.

Lemma 12.2. Let ξ be the operator that associates to a vector b the vector
$$\xi(b) = \{\rho_{\gamma T}(\bar{b}[p])\}_{p \in \Gamma} \quad \text{with} \quad \bar{b} = b + \gamma\Phi(f - \Phi^* b). \tag{12.102}$$
For any $b \in \mathbb{C}^N$, $\xi(b)$ is the unique minimizer of $\tilde{\mathcal{L}}_1(a, b)$ over all a, where
$$\tilde{\mathcal{L}}_1(a, b) = \mathcal{L}_1(a) + \frac{1}{2\gamma}\|a - b\|^2 - \frac{1}{2}\|\Phi^* a - \Phi^* b\|^2 \text{ and } \mathcal{L}_1(a) = \frac{1}{2}\|f - \Phi^* a\|^2 + T\|a\|_1.$$

Furthermore,

$$\forall\, h \in \mathbb{C}^N, \quad \tilde{\mathcal{L}}_1(\xi(b)+h, b) \geqslant \tilde{\mathcal{L}}_1(\xi(b), b) + \frac{1}{2\gamma}\|h\|^2. \tag{12.103}$$

Proof. The modified Lagrangian $\tilde{\mathcal{L}}_1(a, b)$ is expanded as

$$\tilde{\mathcal{L}}_1(a, b) = \frac{1}{2\gamma}\|a\|^2 - \frac{1}{\gamma}\langle a,\ b\rangle + \langle \Phi^* a,\ \Phi^* b - f\rangle + T\|a\|_1 + C_1$$
$$\gamma\tilde{\mathcal{L}}_1(a, b) = \frac{1}{2}\|a - \bar{b}\|^2 + T\gamma\|a\|_1 + C_2,$$

where C_1 and C_2 are two constants that do not depend on a. This proves that $\tilde{\mathcal{L}}_1(a, b)$ is a strictly convex function with respect to a. Since

$$\frac{1}{2}\|a - \bar{b}\|^2 + T\gamma\|a\|_1 = \sum_{p\in\Gamma}(|a[p] - \bar{b}[p]|^2 + T\gamma|a[p]|),$$

each term of this sum can be minimized independently, and Lemma 12.3 proves that the minimum is reached by $\rho_{\gamma T}(\bar{b}[p])$. Moreover, at the minimum, the result (12.104) of Lemma 12.3 implies (12.103). Lemma 12.3 is proved by verifying (12.104) with a direct algebraic calculation.

Lemma 12.3. The scalar energy $e(\alpha) = |\beta - \alpha|^2/2 + T|\alpha|$ is minimized by $\alpha = \rho_T(\beta)$ and

$$e(h + \rho_T(\beta)) - e(\rho_T(\beta)) \geqslant \frac{|h|^2}{2}. \tag{12.104}$$

■

We now prove that the a_k defined in equation (12.100) converge to a fixed point of ξ, and that this fixed point is a minimizer of $\mathcal{L}_1(a)$. The difficulty is that ξ is not strictly contracting. Without loss of generality, we assume that $\gamma = 1$ and that $\|\Phi^* b\| < \|b\|$. Otherwise, Φ^* is replaced by $\sqrt{\gamma}\Phi^*$, f by $\sqrt{\gamma}f$, and T by γT. The thresholding operator ρ_T satisfies

$$\forall\, \alpha, \alpha' \in \mathbb{C}, \quad |\rho_T(\alpha) - \rho_T(\alpha')| \leqslant |\alpha - \alpha'|.$$

This implies the contractions of the mapping ξ:

$$\|\xi(a) - \xi(a')\| \leqslant \|(\mathrm{Id} - \Phi\Phi^*)(a - a')\| \leqslant \|\mathrm{Id} - \Phi\Phi^*\|\,\|a - a'\| \leqslant \|a - a'\|,$$

because $\|\Phi^*\|_S < 1$.

In the following, we write $L = \sqrt{\mathrm{Id} - \Phi\Phi^*}$, which satisfies $\|La\|^2 = \|a\|^2 - \|\Phi^* a\|^2$. Lemma 12.2 shows that a_{k+1} is the minimizer of $a \mapsto \tilde{\mathcal{L}}_1(a, a_k)$, and thus

$$\mathcal{L}_1(a_{k+1}) \leqslant \mathcal{L}_1(a_{k+1}) + \frac{1}{2}\|L(a_{k+1} - a_k)\|^2 = \tilde{\mathcal{L}}_1(a_{k+1}, a_k) \leqslant \tilde{\mathcal{L}}_1(a_k, a_k) = \mathcal{L}_1(a_k), \tag{12.105}$$

so $\{\mathcal{L}_1(a_k)\}_k$ is a nonincreasing sequence. Similarly,

$$\tilde{\mathcal{L}}_1(a_{k+2}, a_{k+1}) \leqslant \mathcal{L}_1(a_{k+1}) \leqslant \mathcal{L}_1(a_{k+1}) + \frac{1}{2}\|L(a_{k+1} - a_k)\|^2 = \tilde{\mathcal{L}}_1(a_{k+1}, a_k),$$

so $\{\tilde{\mathcal{L}}_1(a_{k+1}, a_k)\}_k$ is also a nonincreasing sequence.

Since $\|\Phi^*\|_S < 1$, the operator L is positive definite, and if one denotes $A > 0$ the smallest eigenvalue of L,

$$\sum_{k=0}^{K} \|a_{k+1} - a_k\|^2 \leqslant \frac{1}{A} \sum_{k=0}^{K} \|L(a_{k+1} - a_k)\|^2.$$

It results from (12.105) that $1/2\|L(a_{k+1} - a_k)\|^2 \leqslant \mathcal{L}_1(a_k) - \mathcal{L}_1(a_{k+1})$, and thus

$$\begin{aligned}\sum_{k=0}^{K} \|a_{k+1} - a_k\|^2 &\leqslant \frac{1}{2A} \sum_{k=0}^{K} (\mathcal{L}_1(a_k) - \mathcal{L}_1(a_{k+1})) \\ &= \frac{1}{2A}(\mathcal{L}_1(a_0) - \mathcal{L}_1(a_K)) < \frac{1}{2A}\mathcal{L}_1(a_0).\end{aligned}$$

It follows that the series $\sum_k \|a_{k+1} - a_k\|^2$ converges, and thus

$$\|\xi(a_k) - a_k\| \to 0 \quad \text{when} \quad k \to +\infty. \tag{12.106}$$

Since

$$\|a_k\|_1 \leqslant \frac{1}{T}\mathcal{L}_1(a_k) \leqslant \frac{1}{T}\mathcal{L}_1(a_0),$$

the sequence $\{a_k\}_k$ is bounded. As the vectors are in the finite-dimensional space $\mathbb{C}^N$, there exists a subsequence $\{a_{\gamma(k)}\}_k$ that converges to some $\tilde{a}$. Equation (12.106) proves that $\xi(a_{\gamma(k)})$ also converges to $\tilde{a}$, which is thus a fixed point of ξ.

Since ξ is contracting,

$$\|a_{k+1} - \tilde{a}\| = \|\xi(a_k - \tilde{a})\| \leqslant \|a_k - \tilde{a}\|.$$

The sequence $\{\|a_k - \tilde{a}\|\}_k$ is decreasing and thus convergent. But $\{\|a_{\gamma(k)} - \tilde{a}\|\}_k$ converges to 0 so the whole sequence a_k is converging to $\tilde{a}$.

Given that $\xi(\tilde{a}) = \tilde{a}$, Lemma 12.2 with $\gamma = 1$ proves that

$$\mathcal{L}_1(\tilde{a} + h) + \frac{1}{2}\|h\|^2 - \frac{1}{2}\|\Phi^* h\|^2 = \tilde{\mathcal{L}}_1(\tilde{a} + h, \tilde{a}) \geqslant \tilde{\mathcal{L}}_1(\tilde{a}, \tilde{a}) + \frac{1}{2}\|h\|^2 = \mathcal{L}_1(\tilde{a}) + \frac{1}{2}\|h\|^2,$$

which proves that $\mathcal{L}_1(\tilde{a} + h) \geqslant \mathcal{L}_1(\tilde{a}) + \frac{1}{2}\|\Phi^* h\|^2$, and thus $\tilde{a}$ is a minimizer of $\mathcal{L}_1$. ■

This theorem proves the convergence of the algorithm but provides no bound on the decay rate. It incorporates in the loop the Richardson gradient descent so its convergence is slower than the convergence of a Richardson inversion. Theorem 5.7 proves that this convergence depends on frame bounds. In this iterative thresholding algorithm, it depends on the frame bounds $B_\Lambda \geqslant A_\Lambda > 0$ of the family of vectors $\{\phi_p\}_{p \in \Lambda}$ over the support Λ of the solution as it evolves during the convergence. For a gradient descent alone, Theorem 5.7 proves that error decreases by a factor $\delta = \max\{|1 - \gamma A_\Lambda|, |1 - \gamma B_\Lambda|\}$. This factor is small if A_Λ is small, which indicates that this family of vectors defines a nearly unstable frame. Once the final support $\tilde{\Lambda}$ is recovered, the convergence is fast only if $A_{\tilde{\Lambda}}$ is also not too small. This property guarantees the numerical stability of the solution.

Backprojection

The amplitudes of the coefficients $\tilde{a}$ on the support $\tilde{\Lambda}$ are reduced relative to the coefficients $a_{\tilde{\Lambda}}$ of the orthogonal projection $f_{\tilde{\Lambda}}$ of f on the space generated by $\{\phi_p\}_{p\in\tilde{\Lambda}}$. The approximation error is reduced by a backprojection that recovers $a_{\tilde{\Lambda}}$ from $\tilde{a}$ as in a matching pursuit.

Implementing this backprojection with the Richardson algorithm is equivalent to continuing the iterations with the same gradient step (12.98) and by replacing the soft thresholding (12.99) by an orthogonal projector on $\tilde{\Lambda}$:

$$a_{k+1}[p] = \begin{cases} 0 & \text{if } p \notin \tilde{\Lambda} \\ \tilde{a}_k[p] & \text{if } p \in \tilde{\Lambda}. \end{cases} \tag{12.107}$$

The convergence is then guaranteed by the Richardson theorem (5.7) and depends on $A_{\tilde{\Lambda}}$.

Automatic Threshold Updating

To solve the minimization under an error constraint of ε,

$$\tilde{a} = \underset{a\in\mathbb{R}^P}{\operatorname{argmin}} \|a\|_1 \quad \text{subject to} \quad \|\Phi^* a - f\| \leqslant \varepsilon, \tag{12.108}$$

the Lagrange multiplier T must be adjusted to ε. A sequence of $\tilde{a}_l = \underset{a\in\mathbb{C}^P}{\operatorname{argmin}} \frac{1}{2}\|f - \Phi^* a\|^2 + T_l\|a\|_1$ can be calculated so that $\|\Phi^*\tilde{a}_l - f\|$ converges to ε. The error $\|f - \Phi^*\tilde{a}_l\|$ is an increasing function of T_l but not strictly. A possible threshold adjustment proposed by Chambolle [152] is

$$T_{l+1} = T_l \frac{\varepsilon}{\|f - \Phi^*\tilde{a}_l\|}. \tag{12.109}$$

One can also update a threshold T_k with (12.109) at each step k of the soft-thresholding iteration, which works numerically well, although there is no proof of convergence.

Other Algorithms

Several types of algorithms can solve the $\mathbf{l}^1$ Lagrangian minimization (12.97). It includes primal-dual schemes [497], specialized interior points with preconditioned conjugate gradient [329], Bregman iterations [494], split Bregman iterations [275], two-step iterative thresholding [211], SGPL1 [104], gradient pursuit [115], gradient projection [256], fixed-point continuation [240], gradient methods [389, 390], coordinate-wise descent [261], and sequential subspace optimization [382].

Continuation methods like homotopy [228, 393] keep track of the solution $\tilde{a}_l$ of (12.97) for a decreasing sequence of thresholds T_l. For a given T_l, one can compute the next smallest T_{l+1} where the support of the optimal solution includes a new component (or more rarely where a new component disappears) and the position of this component. At each iteration, the solution is then computed with the implicit equation (12.93) by calculating the pseudo inverse $\Phi_\Lambda^{*\,+} f$. These algorithms are thus

quite fast if the final solution $\tilde{a}$ is highly sparse, because only a few iterations are then necessary and the size of all matrices remains small.

It is difficult to compare all these algorithms because their speed of convergence depends on the sparsity of the solution and on the frame bounds of the dictionary vectors over the support solution. The iterative thresholding algorithm of Theorem 12.9 has the advantage of simplicity.

12.4.4 Sparse Synthesis versus Analysis and Total Variation Regularization

Matching pursuit and basis pursuit algorithms assume that the signal has a *sparse synthesis* in a dictionary $\mathcal{D} = \{\phi_p\}_{p\in\Gamma}$ and compute this representation with as few vectors as possible. The sparse synthesis hypothesis should not be confused with a *sparse analysis* hypothesis, which assumes that a linear signal transform $\Phi f = \{\langle f, \phi_p\rangle\}_{p\in\Gamma}$ is sparse. A sparse analysis can often be related to some form of regularity of f. For example, piecewise regular signals have a sparse wavelet transform. Similarly, total variation regularizations make a sparse analysis assumption on the sparsity of the gradient vector field. Sparse analysis and synthesis assumptions are equivalent in an orthonormal basis, but are very different when the dictionary is redundant. These aspects are clarified and algorithms are provided to solve sparse analysis problems, including total variation regularizations.

Sparse Synthesis and Analysis with l^1 Norms

A basis pursuit incorporates the sparse synthesis assumption by minimizing the $\mathbf{l}^1$ norm of the synthesis coefficients. To approximate f, the Lagrangian formulation computes

$$\tilde{f}_s = \Phi^* \tilde{a} \quad \text{with} \quad \tilde{a} = \underset{a\in\mathbb{C}^P}{\operatorname{argmin}} \frac{1}{2}\|f - \Phi^* a\|^2 + T\|a\|_1. \tag{12.110}$$

In a denoising problem, f is replaced by the input noisy signal $X = f + W$, and if W is a Gaussian white noise of variance σ^2, then T is proportional to σ.

A sparse analysis approximation of f computes

$$\tilde{f}_a = \underset{f\in\mathbb{R}^N}{\operatorname{argmin}} \|\Phi f\|_1 \quad \text{with} \quad \|f - y\| \leqslant \varepsilon, \tag{12.111}$$

where ε is the approximation precision. This problem is convex. Similarly, in a denoising problem, f is replaced by the input noisy signal $X = f + W$. A solution $\tilde{f}_a$ of (12.111) can be computed as a solution of the Lagrangian formulation

$$\tilde{f}_a = \underset{h\in\mathbb{R}^N}{\operatorname{argmin}} \frac{1}{2}\|f - h\|^2 + T\|\Phi h\|_1, \tag{12.112}$$

where T depends on ε. In a denoising problem, f is also replaced by the noisy signal $X = f + W$, and T is proportional to the noise standard deviation σ.

If $\mathcal{D}$ is an orthonormal basis, then $\Phi^* = \Phi^{-1}$ so $\tilde{f}_s = \tilde{f}_a$. The analysis and synthesis prior assumptions are then equivalent. This is, however, not the case when the dictionaries are redundant [243]. In a redundant dictionary, a sparse synthesis assumes that a signal is well approximated by few well-chosen dictionary vectors. However, it often has a nonsparse representation with badly chosen dictionary vectors. For example, a high-frequency oscillatory texture has a sparse synthesis in a dictionary of wavelet and local cosine because it is well represented by local cosine vectors, but it does not have a sparse analysis representation in this dictionary because the wavelet coefficients of these textures are not sparse.

Thresholding algorithms in redundant translation-invariant dictionaries such as translation-invariant wavelet frames rely on a sparse analysis assumption. They select all large frame coefficients, and the estimation is precise if there are relatively few of them. Sparse analysis constraints can often be interpreted as imposing some form of signal regularity condition—for example, with a total variation norm.

In Section 12.4.3 we describe an iterative thresholding that solves the sparse synthesis Lagrangian minimization (12.110). When the dictionary is redundant, the sparse analysis minimization (12.112) must integrate that the $\mathbf{l}^1$ norm is carried over a redundant set of coefficients $a[p] = \Phi h[p]$ that cannot be adjusted independently. If Φ is a frame, then Theorem 5.9 proves that a satisfies a reproducing kernel equation $a = \Phi\Phi^+ a$ where $\Phi\Phi^+$ is an orthogonal projector on the space of frame coefficients. One could think of solving (12.112) with the iterative soft-thresholding algorithm that minimizes (12.110), and project with the projector $\Phi\Phi^+ a$ at each iteration on the constraints, but this algorithm is not guaranteed to converge.

Total Variation Denoising

Rudin, Osher, and Fatemi's [420] total variation regularization algorithm assumes that the image gradient $\vec{\nabla} f$ is sparse, which is enforced by minimizing its $\mathbf{L}^1$ norm, and thus the total image variation $\int\int |\vec{\nabla} f(x)|\, dx$. Over discrete images, the gradient vector is computed with horizontal and vertical finite differences. Let $\tau_1 = (1, 0)$ and $\tau_2 = (0, 1)$:

$$D_k f[p] = f[p] - f[p - \tau_k] = \langle f, \phi_p^k \rangle \quad \text{with} \quad \phi_p^k = \delta[n-p] - \delta[n-p-\tau_k] \quad \text{for} \quad k = 1, 2.$$

The discrete total variation norm is the complex $\mathbf{l}^1$ norm

$$\|f\|_V = \sum_p \sqrt{|D_1 f[p]|^2 + |D_2 f[p]|^2} = \|\Phi f\|_1,$$

where Φ is a complex valued analysis operator

$$\Phi f = D_1 f + i D_2 f = \{\langle f, \phi_p \rangle\}_p \quad \text{for} \quad \phi_p = \phi_p^1 + i\, \phi_p^2.$$

The discrete image gradient $\Phi f = D_1 f + i D_2 f$ has the same size N as the image but has twice the scalar values because of the two partial derivative coordinates. Thus, it defines a redundant representation in $\mathbb{R}^N$. The finite difference D_1 and D_2 are convolution operators with transfer functions that vanish at the 0 frequency. As

(a) (b) (c)

FIGURE 12.24

(a) Noisy image $X = f + w$ (SNR = 16 db). **(b)** Translation-invariant wavelet thresholding estimation (SNR = 22.9 db). **(c)** Estimation by total variation regularization (SNR = 21.9 db).

a result, Φ is not a frame of $\mathbb{R}^N$. However, it is invertible on the subspace $\mathbf{V} = \{f \in \mathbb{R}^N : \sum_n f[n] = 0\}$ of dimension $N - 1$. Since we are in finite dimension, it defines a frame of $\mathbf{V}$, but the pseudo inverse Φ^+ in (5.21) becomes numerically unstable when N increases because the frame bound ratio A/B tends to zero. Theorem 5.9 proves that $\Phi\Phi^+$ remains an orthogonal projector on the space of gradient coefficients.

Figure 12.24 shows an example of image denoising with a total variation regularization. The total variation minimization recovers an image with a gradient vector that is as sparse as possible, which has a tendency to produce piecewise constant regions when the thresholding increases. As a result, the total variation image denoising most often yields a smaller SNR than wavelet thresholding estimators, unless the image is piecewise constant.

Computation of Sparse Analysis Approximations and Denoising with l^1 Norms

A sparse analysis approximation is defined by

$$\tilde{f}_a = S_T(f) = \underset{h \in \mathbb{R}^N}{\operatorname{argmin}} \frac{1}{2}\|f - h\|^2 + T\|\Phi h\|_1, \tag{12.113}$$

and f is replaced by the noisy data $X = f + W$ in a denoising problem. Chambolle [152] sets a dual problem by verifying that the regularization term $\|\Phi h\|_1$ can be replaced by a set of dual variables:

$$\|\Phi h\|_1 = \max_{r \in \mathcal{K}} \langle r,\ h \rangle \quad \text{where} \quad \mathcal{K} = \left\{ \Phi^* a : \|a\|_\infty \leqslant 1 \right\}, \tag{12.114}$$

with $\|a\|_\infty = \max_p |a[p]|$, and where $|\cdot|$ is the modulus of complex numbers. Because of (12.114), the regularization $J(h) = \|\Phi h\|_1$ is "linearized" by taking a maximum over inner products with vectors in a convex set $r \in \mathcal{K}$. This decomposition remains valid for any positively homogeneous functional J, which satisfies $J(\lambda h) = |\lambda| J(h)$.

The minimization (12.113) is thus rewritten as a minimization and maximization:

$$\tilde{f}_a = \operatorname*{argmin}_{h \in \mathbb{R}^N} \frac{1}{2}\|f-h\|^2 + T\|\Phi h\|_1 = \operatorname*{argmin}_{h \in \mathbb{R}^N} \max_{r \in \mathcal{K}} \frac{1}{2}\|f-h\|^2 + T\langle r,\ h\rangle.$$

Inverting the min and max and computing the solution of the min problem shows that $\tilde{f}_a$ is written as

$$\tilde{f}_a = f - T\tilde{r} \quad \text{where} \quad \tilde{r} = \operatorname*{argmin}_{r \in \mathcal{K}} \|f - Tr\|,$$

which means that $\tilde{r}$ is the projection of f/T on the convex set $\mathcal{K}$.

The solution of (12.113) is thus also the solution of the following convex problem:

$$\tilde{f}_a = f - T\Phi^*\tilde{a} \quad \text{where} \quad \tilde{a} = \operatorname*{argmin}_{\|a\|_\infty \leqslant 1, a \in \mathbb{C}^P} \|\Phi^* a - f/T\|. \tag{12.115}$$

The solution $\tilde{a}$ is not necessarily unique, but both $\tilde{r} = \Phi^*\tilde{a}$ and $\tilde{f}_a$ are unique. With this dual formulation, the redundancy of $a = \Phi f$ in the $\mathbf{l}^1$ norm of (12.113) does not appear anymore.

Starting from a choice of a_0, for example $a_0 = 0$, Chambolle [152] computes $\tilde{a}$ by iterating between a gradient step to minimize $\|\Phi^* a - f/T\|^2$,

$$b^{(k+1)} = \Phi(\Phi^* a^{(k)} - f/T),$$

and a "projection" step to ensure that $|a[p]| \leqslant 1$ for all p. An orthogonal projection at each iteration on the constraint gives

$$\tilde{a}^{(k+1)} = a^{(k)}[p] + \gamma\, b^{(k+1)}[p] \quad \text{and} \quad a^{(k+1)}[p] = \frac{\tilde{a}^{(k+1)}[p]}{\max(|\tilde{a}^{(k+1)}[p]|\ ,\ 1)}.$$

One can verify [152] that the convergence is guaranteed if $\gamma < 2\,\|\Phi\Phi^*\|_S^{-1}$. To satisfy the constraints, another possibility is to set

$$a^{(k+1)}[p] = \frac{a^{(k)}[p] + \gamma\, b^{(k+1)}[p]}{1 + \gamma\,|b^{(k+1)}[p]|}. \tag{12.116}$$

For the gradient operator $\Phi^* = \nabla$ discretized with finite differences in 2D, $\|\Phi\Phi^*\|_S = 1/8$ so one can choose $\gamma = 1/4$. The convergence is often fast during the first few iterations, leading to a visually satisfying result, but the remaining iterations tend to converge slowly to the final solution.

Other iterative algorithms have been proposed to solve (12.113), for instance, fixed-point iterations [476], second-order cone programming [274], splitting [480], splitting with Bregman iterations [275], and primal-dual methods [153, 497]. Primal-dual methods tend to have a faster convergence than the Chambolle algorithm.

Chambolle [152] proves that the sparse analysis minimization (12.111) with a precision ε is obtained by iteratively computing the solution $\tilde{a}_l$ of Lagrangian problems (12.113) for thresholds T_l that are adjusted with

$$T_{l+1} = T_l\, \frac{\varepsilon}{\|f - \Phi^*\tilde{a}_l\|}. \tag{12.117}$$

The convergence proof relies on the fact that $\|f - S_T(f)\|$ is a strictly increasing function of T for the analysis problem, which is not true for the synthesis problem.

Computation of Sparse Analysis Inverse Problems with l^1 Norms

In inverse problems studied in Chapter 13, f is estimated from $Y = Uf + W$, where U is a linear operator and W an additive noise. A sparse analysis estimation of f computes

$$\tilde{F}_a = \underset{h \in \mathbb{R}^N}{\operatorname{argmin}} \frac{1}{2}\|Y - Uh\|^2 + T\|\Phi h\|_1. \tag{12.118}$$

The sparse Lagrangian analysis approximation operator S_T in (12.113) can be used to replace the wavelet soft thresholding in the iterative thresholding algorithm (12.100), which leads to the iteration

$$\tilde{F}_{k+1} = S_{\gamma T}\left(\tilde{F}_k + \mu\, U^*(Y - U\tilde{F}_k)\right).$$

One can prove with the general machinery of proximal iterations that $\tilde{F}_k$ converges to a solution $\tilde{F}$ of (12.118) if $\mu < \|U^*U\|_S^{-1}$ [186]. The algorithm is implemented with two embedded loops. The outer loop on k computes $\tilde{F}_k + \mu\, U^*(Y - U\tilde{F}_k)$, followed by the inner loop, which computes the Lagrangian approximation S_T, for example, with Chambolle algorithm.

12.5 PURSUIT RECOVERY

Matching pursuits and basis pursuits are nonoptimal sparse approximation algorithms in a redundant dictionary $\mathcal{D}$, but are computationally efficient. However, pursuit approximations can be nearly as precise as optimal M-term approximations, depending on the properties of the approximation supports in $\mathcal{D}$.

Beyond approximation, this section studies the ability of pursuit algorithms to recover a specific set Λ of vectors providing a sparse signal approximation in a redundant dictionary. The stability of this recovery is important for pattern recognition when selected dictionary vectors are used to analyze the signal information. It is also at the core of the sparse super-resolution algorithms introduced in Section 13.3.

The stability of sparse signal approximations in redundant dictionaries is related to the dictionary coherence in Section 12.5.1. The exact recovery of signals with matching pursuits and basis pursuits are studied in Sections 12.5.2 and 12.5.3, together with the precision of M-term approximations.

12.5.1 Stability and Incoherence

Let $\mathcal{D} = \{\phi_p\}_{p \in \Gamma}$ be a redundant dictionary of normalized vectors. Given a family of linearly independent approximation vectors $\{\phi_p\}_{p \in \Lambda}$ selected by some algorithm,

the best approximation of f is its orthogonal projection f_Λ in the space $\mathbf{V}_\Lambda$ generated by these vectors

$$f_\Lambda = \sum_{p\in\Lambda} a[p]\,\phi_p.$$

The calculation of the coefficients $a[p]$ is stable if $\{\phi_p\}_{p\in\Lambda}$ is a Riesz basis of $\mathbf{V}_\Lambda$, with Riesz bounds $B_\Lambda \geqslant A_\Lambda > 0$, which satisfy:

$$\forall a \in \mathbb{C}^{|\Lambda|}, \quad A_\Lambda \|a\|^2 \leqslant \| \sum_{p\in\Lambda} a[p]\phi_p\|^2 \leqslant B_\Lambda \|a\|^2. \tag{12.119}$$

The closer A_Λ/B_Λ to 1, the more stable the basis. Gradient descent algorithms compute the coefficients $a[p]$ with a convergence rate that also depends on A_Λ/B_Λ. Theorem 12.10 relates the Riesz bounds to the dictionary *mutual coherence*, introduced by Donoho and Huo [230],

$$\mu(\mathcal{D}) = \sup_{(q,q)\in\Gamma^2, p\neq q} |\langle \phi_p, \phi_q\rangle|.$$

Theorem 12.10. The Riesz bounds of $\{\phi_p\}_{p\in\Lambda}$ satisfy

$$\delta_\Lambda = \max(1 - A_\Lambda, B_\Lambda - 1) \leqslant \max_{p\in\Lambda} \sum_{q\in\Lambda, p\neq q} |\langle \phi_p, \phi_q\rangle| \leqslant (|\Lambda| - 1)\,\mu(\mathcal{D}). \tag{12.120}$$

Proof. Theorem 5.1 proves that the constants A_Λ and B_Λ are lower and upper bounds of the eigenvalues of the Gram matrix $G_\Lambda = \{\langle \phi_p, \phi_q\rangle\}_{(p,q)\in\Lambda^2}$. Let $v \neq 0$ be an eigenvector satisfying $G_\Lambda v = \lambda v$. Let $|v[p]| = \max_{q\in\Lambda} |v[q]|$. Since $\|\phi_p\|^2 = 1$,

$$v[p] + \sum_{q\in\Lambda, q\neq p} \langle \phi_p, \phi_q\rangle v[q] = \lambda\, v[p] \quad \Longrightarrow \quad |1-\lambda| \leqslant \sum_{q\in\Lambda, q\neq p} |\langle \phi_p, \phi_q\rangle| \frac{|v[q]|}{|v[p]|}.$$

It results that

$$\delta_\Lambda = \max_\lambda |1-\lambda| \leqslant \sum_{q\in\Lambda, q\neq p} |\langle \phi_p, \phi_q\rangle| \leqslant (|\Lambda| - 1)\mu(\mathcal{D}),$$

which proves (12.120). ■

The Riesz bound ratio A_Λ/B_Λ is close to 1 if δ_Λ is small and thus if the vectors in Λ have a small correlation. The upper bound (12.120) proves that sufficiently small sets $|\Lambda| < \mu(\mathcal{D})$ are Riesz bases with $A_\Lambda > 0$. To increase the maximum size of these sets, one should construct dictionaries that are as *incoherent* as possible. The upper bounds (12.120) are simple but relatively crude, because they only depend on inner products of pairs of vectors, whereas the Riesz stability of $\{\phi_p\}_{p\in\Lambda}$ depends on the distribution of this whole group of vectors on the unit sphere of $\mathbb{C}^N$. Section 13.4 proves that much better bounds can be calculated over dictionaries of random vectors.

The mutual coherence of an orthonormal basis $\mathcal{D}$ is 0, and one can verify (Exercise 12.5) that if we add a vector g, then $\mu(\mathcal{D} \cup \{g\}) \geqslant \frac{1}{\sqrt{N}}$. However, this level of incoherence can be reached with larger dictionaries. Let us consider the dictionary of $P = 2N$ vectors, which is a union of a Dirac basis and a discrete Fourier basis:

$$\mathcal{D} = \{\delta[n-p]\}_{0 \leqslant p < N} \cup \{N^{-1/2} e^{-i2\pi pn/N}\}_{0 \leqslant p < N}. \tag{12.121}$$

Its mutual coherence is $\mu(\mathcal{D}) = N^{-1/2}$. The right upper bound of (12.120) thus proves that any family of $|\Lambda| \leqslant \sqrt{N}/2$ Dirac and Fourier vectors defines a basis with a Riesz bound ratio $A_\Lambda / B_\Lambda \geqslant 1/3$. One can construct larger frames $\mathcal{D}$ of $P = N^2$ vectors, called Grassmannian frames, that have a coherence in $O(1/\sqrt{N})$ [448].

Dictionaries often do not have a very small coherence, so the inequality $\delta_\Lambda \leqslant (|\Lambda| - 1)\mu(\mathcal{D})$ applies to relatively small sets Λ. For example, in the Gabor dictionary $\mathcal{D}_{j,\Delta}$ defined in (12.77), the mutual coherence $\mu(\mathcal{D})$ is maximized by two neighboring Gabor atoms, and the inner product formula (12.81) proves that $\mu(\mathcal{D}) = e^{-\pi \Delta^{-2}/2} = 0.67$ for $\Delta = 2$. The mutual coherence upper bound is therefore useless in this case, but the first upper bound of (12.120) can be used with (12.81) to verify that Gabor atoms that are sufficiently far in time and frequency generate a Riesz basis.

12.5.2 Support Recovery with Matching Pursuit

We first study the reconstruction of signals f that have an exact sparse representation in $\mathcal{D}$,

$$f = \sum_{p \in \Lambda} a[p] \phi_p.$$

An exact recovery condition on Λ is established to guarantee that a matching pursuit selects only approximation vectors in Λ. The optimality of matching pursuit approximations is then analyzed for more general signals. We suppose that the matching pursuit relaxation factor is $\alpha = 1$.

An orthogonal matching pursuit, or a matching pursuit with backprojection, computes the orthogonal projection $f_{\tilde{\Lambda}}$ of f on a family of vectors $\{\phi_p\}_{p \in \tilde{\Lambda}}$ selected one by one. At a step m, a matching pursuit selects an atom in Λ if and only if the correlation of the residual $R^m f$ with vectors in the complement Λ^c of Λ is smaller than the correlation with vectors in Λ:

$$C(R^m f, \Lambda^c) = \frac{\max_{q \in \Lambda^c} |\langle R^m f, \phi_q \rangle|}{\max_{p \in \Lambda} |\langle R^m f, \phi_p \rangle|} < 1. \tag{12.122}$$

The relative correlation of a vector h with vectors in Λ^c relative to Λ is defined by

$$C(h, \Lambda^c) = \frac{\max_{q \in \Lambda^c} |\langle h, \phi_q \rangle|}{\max_{p \in \Lambda} |\langle h, \phi_p \rangle|}. \tag{12.123}$$

Theorem 12.11, proved by Tropp [461], gives a condition that guarantees the recovery of Λ with a matching pursuit.

Theorem 12.11: *Tropp.* If $\{\tilde{\phi}_{p,\Lambda}\}_{p\in\Lambda}$ is the dual basis of $\{\phi_p\}_{p\in\Lambda}$ in the space $\mathbf{V}_\Lambda$, then

$$\mathrm{ERC}(\Lambda) = \max_{q\in\Lambda^c} \sum_{p\in\Lambda} |\langle \tilde{\phi}_{p,\Lambda},\, \phi_q\rangle| = \sup_{h\in\mathbf{V}_\Lambda} C(h,\Lambda^c). \tag{12.124}$$

If $f \in \mathbf{V}_\Lambda$ and the following exact recovery condition (ERC) is satisfied,

$$\mathrm{ERC}(\Lambda) < 1, \tag{12.125}$$

then a matching pursuit of f selects only vectors in $\{\phi_p\}_{p\in\Lambda}$ and an orthogonal matching pursuit recovers f with at most $|\Lambda|$ iterations.

Proof. Let us first prove that $\sup_{h\in\mathbf{V}_\Lambda} C(h,\Lambda^c) \leqslant \mathrm{ERC}(\Lambda)$. Let Φ_Λ^+ be the pseudo inverse of Φ_Λ with $\Phi_\Lambda f[p] = \langle f, \phi_p\rangle$ for $p\in\Lambda$. Theorem 5.6 proves that ${\Phi_\Lambda}^+\Phi_\Lambda$ is an orthogonal projector in $\mathbf{V}_\Lambda$, so if $h\in\mathbf{V}_\Lambda$ and $q\in\Lambda^c$,

$$\begin{aligned} |\langle h,\, \phi_q\rangle| &= |\langle \Phi_\Lambda^+\Phi_\Lambda h,\, \phi_q\rangle| = |\langle \Phi_\Lambda h, \\ &\qquad (\Phi_\Lambda^+)^*\phi_q\rangle| \leqslant \|\Phi_\Lambda h\|_\infty \max_{q\in\Lambda^c} \|(\Phi_\Lambda^+)^*\phi_q\|_1. \end{aligned} \tag{12.126}$$

Theorem 5.5 proves that the dual-frame operator satisfies $\tilde{\Phi}_\Lambda^* = \Phi_\Lambda^+$ and thus that $(\Phi_\Lambda^+)^* f[p] = \tilde{\Phi}_\Lambda[p] = \langle f, \tilde{\phi}_{p,\Lambda}\rangle$ for $p\in\Lambda$. It results that

$$\mathrm{ERC}(\Lambda) = \max_{q\in\Lambda^c} \|(\Phi_\Lambda^+)^*\phi_q\|_1 = \max_{q\in\Lambda^c} \sum_{p\in\Lambda} |\langle \tilde{\phi}_{p,\Lambda},\, \phi_q\rangle|, \tag{12.127}$$

and (12.126) implies that

$$\forall h\in\mathbf{V}_\Lambda, \quad \max_{q\in\Lambda^c} |\langle h,\, \phi_q\rangle| \leqslant \max_{p\in\Lambda} |\langle h,\, \phi_p\rangle|\ \mathrm{ERC}(\Lambda),$$

which proves that $\mathrm{ERC}(\Lambda) \geqslant \sup_{h\in\mathbf{V}_\Lambda} C(h,\Lambda^c)$.

We now prove the reverse inequality. Let $q_0\in\Lambda^c$ be the index such that

$$\sum_{p\in\Lambda} |\langle \tilde{\phi}_{p,\Lambda},\, \phi_{q_0}\rangle| = \max_{q\in\Lambda^c} \sum_{p\in\Lambda} |\langle \tilde{\phi}_{p,\Lambda},\, \phi_q\rangle|.$$

Introducing

$$h = \sum_{p\in\Lambda} \mathrm{sign}(\langle \tilde{\phi}_{p,\Lambda},\, \phi_{q_0}\rangle)\, \tilde{\phi}_{p,\Lambda} \in \mathbf{V}_\Lambda \tag{12.128}$$

leads to

$$\begin{aligned} \mathrm{ERC}(\Lambda) &= \max_{q\in\Lambda^c} \sum_{p\in\Lambda} |\langle \tilde{\phi}_{p,\Lambda},\, \phi_q\rangle| = |\langle h,\, \phi_{q_0}\rangle| \\ &\leqslant \max_{q\in\Lambda^c} |\langle h,\, \phi_q\rangle| \leqslant C(h,\Lambda^c) \max_{p\in\Lambda} |\langle h,\, \phi_p\rangle|. \end{aligned}$$

Since $|\langle h,\phi_p\rangle| = |\mathrm{sign}(\langle \tilde{\phi}_{p,\Lambda},\, \phi_{q_0}\rangle)| = 1$, it results that $C(h,\Lambda^c) \geqslant \mathrm{ERC}(\Lambda)$ and thus $\sup_{h\in\mathbf{V}_\Lambda} C(h,\Lambda^c) \geqslant \mathrm{ERC}(\Lambda)$, which finishes the proof of (12.124).

Suppose now that $f = R^0 f \in \mathbf{V}_\Lambda$ and $\mathrm{ERC}(\Lambda) < 1$. We prove by induction that a matching pursuit selects only vectors in $\{\phi_p\}_{p\in\Lambda}$. Suppose that the first $m < M$ matching pursuit

vectors are in $\{\phi_p\}_{p\in\Lambda}$ and thus that $R^m f \in \mathbf{V}_\Lambda$. If $R^m f \neq 0$, then (12.125) implies that $C(R^m f, \Lambda^c) < 1$ and thus that the next vector is selected in Λ. Since $\dim(\mathbf{V}_\Lambda) \leqslant |\Lambda|$, an orthogonal pursuit converges in less than $|\Lambda|$ iterations. ■

This theorem proves that if a signal can be exactly decomposed over $\{\phi_p\}_{p\in\Lambda}$, then $\mathrm{ERC}(\Lambda) < 1$ guarantees that a matching pursuit reconstructs f with vectors in $\{\phi_p\}_{p\in\Lambda}$. A nonorthogonal pursuit may, however, require more than $|\Lambda|$ iterations to select all vectors in this family.

If $\mathrm{ERC}(\Lambda) > 1$, then there exists $f \in \mathbf{V}_\Lambda$ such that $C(f, \Lambda^c) > 1$. As a result, there exists a vector ϕ_q with $q \in \Lambda^c$, which correlates f better than any other vectors in Λ, and that will be selected by the first iteration of a matching pursuit. This vector may be removed, however, from the approximation support at the end of the decomposition. Indeed, if the remaining iterations select all vectors $\{\phi_p\}_{p\in\Lambda}$, then an orthogonal matching pursuit decomposition or a backprojection will associate a coefficient 0 to ϕ_q because $f = P_{\mathbf{V}_\Lambda} f$. In particular, we may have $\mathrm{ERC}(\Lambda) > 1$ but $\mathrm{ERC}(\tilde{\Lambda}) < 1$ with $\Lambda \subset \tilde{\Lambda}$, in which case an orthogonal pursuit recovers exactly the support of any $f \in \mathbf{V}_\Lambda$ with $|\tilde{\Lambda}| \geqslant |\Lambda|$ iterations.

Theorem 12.12 gives an upper bound proved by Tropp [461], which relates $\mathrm{ERC}(\Lambda)$ to the support size $|\Lambda|$. A tighter bound proved by Gribonval and Nielsen [281] and Dossal [231] depends on inner products of dictionary vectors in Λ relative to vectors in the complement Λ^c.

Theorem 12.12: *Tropp, Gribonval, Nielsen, Dossal.* For any $\{\phi_p\}_{p\in\Lambda} \subset \mathcal{D}$,

$$\mathrm{ERC}(\Lambda) \leqslant \frac{\max\limits_{q\in\Lambda^c} \sum_{p\in\Lambda} |\langle \phi_p, \phi_q\rangle|}{1 - \max\limits_{q\in\Lambda} \sum_{p\in\Lambda, p\neq q} |\langle \phi_p, \phi_q\rangle|} \leqslant \frac{|\Lambda|\mu(\mathcal{D})}{1 - (|\Lambda| - 1)\mu(\mathcal{D})}. \tag{12.129}$$

Proof. It is shown in (12.127) that $\mathrm{ERC}(\Lambda) = \max\limits_{q\in\Lambda^c} \|(\Phi_\Lambda^+)^* \phi_q\|_1$. We verify that

$$(\Phi_\Lambda^+)^* h = \Phi_\Lambda (\Phi_\Lambda^* \Phi_\Lambda)^{-1} h = (\Phi_\Lambda^*)^+ h = (\Phi_\Lambda \Phi_\Lambda^*)^{-1} \Phi_\Lambda h, \tag{12.130}$$

and thus that we can also write $\mathrm{ERC}(\Lambda) = \max\limits_{q\in\Lambda^c} \|(\Phi_\Lambda \Phi_\Lambda^*)^{-1} \Phi_\Lambda \phi_q\|_1$. Introducing the operator norm associated to the $\mathbf{l}^1$ norm

$$\|A\|_{1,1} = \max_{b\neq 0} \frac{\|Ab\|_1}{\|b\|_1} = \max_j \sum_i |a_{i,j}|$$

for a matrix $A = (a_{i,j})_{i,j}$, leads to

$$\mathrm{ERC}(\Lambda) = \max_{q\in\Lambda^c} \|(\Phi_\Lambda \Phi_\Lambda^*)^{-1} \Phi_\Lambda \phi_q\|_1 \leqslant \max_{q\in\Lambda^c} \|(\Phi_\Lambda \Phi_\Lambda^*)^{-1}\|_{1,1} \max_{q\in\Lambda^c} \|\Phi_\Lambda \phi_q\|_1. \tag{12.131}$$

The second term is the numerator of (12.129),

$$\max_{q\in\Lambda^c} \|\Phi_\Lambda \phi_q\|_1 = \max_{q\in\Lambda^c} \sum_{p\in\Lambda} |\langle \phi_p, \phi_q\rangle|. \tag{12.132}$$

The Gram matrix is rewritten as $G = \Phi_\Lambda \Phi_\Lambda^* = \mathrm{Id} + H$, and a Neumann expansion of G^{-1} gives

$$\|(\Phi_\Lambda \Phi_\Lambda^*)^{-1}\|_{1,1} \leqslant \sum_{k \geqslant 0} (\|H\|_{1,1})^k \leqslant \frac{1}{1 - \|H\|_{1,1}}, \tag{12.133}$$

with

$$\|H\|_{1,1} = \max_{q \in \Lambda} \sum_{p \in \Lambda, p \neq q} |\langle \phi_p, \phi_q \rangle|. \tag{12.134}$$

Inserting this result in (12.133) and inserting (12.132) and (12.133) in (12.131) prove the first inequality of (12.129).

The second inequality is derived from the fact that $|\langle \phi_p, \phi_q \rangle| \leqslant \mu(\mathcal{D})$ for any $p \neq q$. ■

This theorem gives upper bounds of $\mathrm{ERC}(\Lambda)$ that can easily be computed. It proves that $\mathrm{ERC}(\Lambda) < 1$ if the vectors $\{\phi_p\}_{p \in \Lambda}$ are not too correlated between themselves and with the vectors in the complement Λ^c. In a Gabor dictionary $\mathcal{D}_{j,\Delta}$ defined in (12.77), the upper bound (12.129) with the Gabor inner product formula (12.81) proves that $\mathrm{ERC}(\Lambda) < 1$ for families of sufficiently separated time-frequency Gabor atoms indexed by Λ. Theorem 12.11 proves that any combination of such Gabor atoms is recovered by a matching pursuit. Suppose that the Gabor atoms are defined with a Gaussian window that has a variance in time and frequency that is σ_t^2 and σ_ω^2, respectively. Their Heisenberg box has a size $\sigma_t \times \sigma_\omega$. If $|\Lambda| = 2$, then one can verify that $\mathrm{ERC}(\Lambda) > 1$ if the Heisenberg boxes of these two atoms intersect. If the time distance of the two Gabor atoms is larger than $1.5\,\sigma_t$, or if the frequency distance is larger than $1.5\,\sigma_\omega$, then $\mathrm{ERC}(\Lambda) < 1$.

The second upper bound in (12.129) proves that $\mathrm{ERC}(\Lambda) < 1$ for any sufficiently small set Λ

$$|\Lambda| < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathcal{D})}\right). \tag{12.135}$$

Very sparse approximation sets are thus more easily recovered. The Dirac and Fourier dictionary (12.121) has a low mutual coherence $\mu(\mathcal{D}) = N^{-1/2}$. Condition (12.135) together with Theorem 12.11 proves that any combination of $|\Lambda| \leqslant N^{1/2}/2$ Fourier and Dirac vectors are recovered by a matching pursuit. The upper bound (12.135) is, however, quite brutal, and in a Gabor dictionary where $\mu(\mathcal{D}_{j,\Delta}) = \mathrm{e}^{-\pi\Delta^{-2}/2}$ and $\Delta \geqslant 2$, it applies only to $|\Lambda| = 1$, which is useless.

Nearly Optimal Approximations with Matching Pursuits

Let f_Λ be the best approximation of f from $M = |\Lambda|$ vectors in $\mathcal{D}$. If $f_\Lambda = f$ and $\mathrm{ERC}(\Lambda) < 1$, then Theorem 12.11 proves that f_Λ is recovered by a matching pursuit, but it is rare that a signal is exactly a combination of few dictionary vectors. Theorem 12.13, proved by Tropp [461], shows that if $\mathrm{ERC}(\Lambda) < 1$, then $|\Lambda|$ orthogonal

matching pursuit iterations recover the "main approximation vectors" in Λ, and thus produce an error comparable to $\|f - f_\Lambda\|$.

Theorem 12.13: *Tropp.* Let $A_\Lambda > 0$ be the lower Riesz bound of $\{\phi_p\}_{p\in\Lambda}$. If $\mathrm{ERC}(\Lambda) < 1$, then an orthogonal matching pursuit approximation $\tilde{f}_M$ on f computed with $M = |\Lambda|$ iterations satisfies

$$\|f - \tilde{f}_M\|^2 \leq \left(1 + \frac{|\Lambda|}{A_\Lambda\,(1 - \mathrm{ERC}(\Lambda))^2}\right) \|f - f_\Lambda\|^2. \tag{12.136}$$

Proof. The residual at step $m < |\Lambda|$ is denoted $R^m f = f - \tilde{f}_m$ where $\tilde{f}_m$ is the orthogonal projection of f on the space generated by the dictionary vectors selected by the first m iterations. The theorem is proved by induction by showing that either $\tilde{f}_m$ satisfies (12.136) or $\tilde{f}_m \in \mathbf{V}_\Lambda$, which means that the first m vectors selected by the orthogonal pursuit are indexed in Λ. If (12.136) is satisfied for some $m \leq M = |\Lambda|$, since $\|R^M f\| \leq \|R^m f\|$, it results that $\tilde{f}_M$ satisfies (12.136). If this is not the case, then the induction proof will show that the $M = |\Lambda|$ selected vectors are in $\{\phi_p\}_{p\in\Lambda}$. Since an orthogonal pursuit selects linearly independent vectors, it implies that $\tilde{f}_M = \tilde{f}_\Lambda$ satisfies (12.136).

The induction step is proved by supposing that $\tilde{f}_m \in \mathbf{V}_\Lambda$, and verifying that either $C(R^m f, \Lambda^c) < 1$, in which case the next selected vector is indexed in Λ, or that (12.136) is satisfied. We write $\Phi_\Lambda^* a = \sum_{p\in\Lambda} a[p]\phi_p$. Since $f - f_\Lambda$ is orthogonal to $\mathbf{V}_\Lambda$ and $f = R^m f + \tilde{f}_m$ for $p \in \Lambda$, we have $\langle R^m f,\ \phi_p\rangle = \langle f_\Lambda - \tilde{f}_m,\ \phi_p\rangle$, so

$$C(R^m f, \Lambda^c) = \frac{\max\limits_{q\in\Lambda^c} |\langle R^m f,\ \phi_q\rangle|}{\max\limits_{p\in\Lambda} |\langle R^m f,\ \phi_p\rangle|} \leq C_1 + C(f_\Lambda - \tilde{f}_m, \Lambda^c), \tag{12.137}$$

with

$$C_1 = \frac{\max\limits_{q\in\Lambda^c} |\langle f - f_\Lambda,\ \phi_q\rangle|}{\max\limits_{p\in\Lambda} |\langle f_\Lambda - \tilde{f}_m,\ \phi_p\rangle|} \quad \text{and} \quad C(f_\Lambda - \tilde{f}_m, \Lambda^c) = \frac{\max\limits_{q\in\Lambda^c} |\langle f_\Lambda - \tilde{f}_m,\ \phi_q\rangle|}{\max\limits_{p\in\Lambda} |\langle f_\Lambda - \tilde{f}_m,\ \phi_p\rangle|}.$$

Since $f_\Lambda - \tilde{f}_m \in \mathbf{V}_\Lambda$, Theorem 12.11 proves that $C(f_\Lambda - \tilde{f}_m, \Lambda^c) \leq \mathrm{ERC}(\Lambda)$. Since $\Phi_\Lambda(f_\Lambda - \tilde{f}_m)[p] = \langle f_\Lambda - \tilde{f}_m, \phi_p\rangle$ and $\|\Phi_\Lambda h\|^2 \geq A_\Lambda \|h\|^2$, we get

$$\max_{p\in\Lambda} |\langle f_\Lambda - \tilde{f}_m,\ \phi_p\rangle| \geq \frac{1}{\sqrt{|\Lambda|}} \|\Phi_\Lambda(f_\Lambda - \tilde{f}_m)\| \geq \frac{\sqrt{A_\Lambda}}{\sqrt{|\Lambda|}} \|f_\Lambda - \tilde{f}_m\|. \tag{12.138}$$

Since $\max\limits_{q\in\Lambda^c} |\langle f - f_\Lambda,\ \phi_q\rangle| \leq \|f - f_\Lambda\|$, inserting these inequalities in (12.137) gives

$$C(R^m f, \Lambda^c) \leq \frac{\sqrt{|\Lambda|}}{\sqrt{A_\Lambda}} \frac{\|f - f_\Lambda\|}{\|f_\Lambda - \tilde{f}_m\|} + \mathrm{ERC}(\Lambda).$$

If $C(R^m f, \Lambda^c) \geq 1$ then

$$\|f_\Lambda - \tilde{f}_m\|^2 \leq \frac{|\Lambda|}{A_\Lambda (1 - \mathrm{ERC}(\Lambda))^2} \|f - f_\Lambda\|^2.$$

Since $f - f_\Lambda$ is orthogonal to $f_\Lambda - \tilde{f}_m$ this condition is equivalent to

$$\|f - \tilde{f}_m\|^2 = \|f - f_\Lambda\|^2 + \|f_\Lambda - \tilde{f}_m\|^2 \leq \left(1 + \frac{|\Lambda|}{A_\Lambda (1 - \mathrm{ERC}(\Lambda))^2}\right) \|f - f_\Lambda\|^2,$$

which proves that (12.136) is satisfied, and thus finishes the induction proof. ■

The proof shows that an orthogonal matching pursuit selects the few first vectors in Λ. These are the "coherent signal structures" having a strong correlation with the dictionary vectors, observed in the numerical experiments in Section 12.3. At some point, the remaining vectors in Λ may not be sufficiently well correlated with f relative to other dictionary vectors; Theorem 12.13 computes the approximation error at this stage. This result is thus conservative since it does not take into account the approximation improvement obtained by the other vectors. Gribonval and Vandergheynst proved [283] that a nonorthogonal matching pursuit satisfies a similar theorem, but with more than $M = |\Lambda|$ iterations. Orthogonal and nonorthogonal matching pursuits select the same "coherent structures." Theorem 12.14 derives an approximation result that depends on only the number of M-term approximations and on the dictionary mutual coherence.

Theorem 12.14. Let f_M be the best M-term approximation of f from M dictionary vectors. If $M \leq \frac{1}{3\mu(\mathcal{D})}$, then an orthogonal matching pursuit approximation $\tilde{f}_M$ with M iterations satisfies

$$\|f - \tilde{f}_M\|^2 \leq (1 + 6M) \|f - f_M\|^2. \tag{12.139}$$

Proof. Let Λ be the approximation support of the best M-term dictionary approximation f_M with $|\Lambda| = M$. If $|\Lambda| \leq \frac{1}{3\mu(\mathcal{D})}$, then (12.129) proves that $(1 - \mathrm{ERC}(\Lambda))^{-1} < 2$. Theorem 12.10 shows that if $M \leq \frac{1}{3\mu(\mathcal{D})}$, then $A_\Lambda \geq 2/3$. Theorem 12.13 derives in (12.136) that $\tilde{f}_M$ satisfies (12.139). ■

In the dictionary (12.121) of Fourier and Dirac vectors where $\mu(\mathcal{D}) = N^{-1/2}$, this theorem proves that M orthogonal matching pursuit iterations are nearly optimal if $M \leq N^{1/2}/3$. This result is attractive because it is simple, but in practice the condition $M \leq \frac{1}{3\mu(\mathcal{D})}$ is very restrictive because the dictionary coherence is often not so small. As previously explained, the mutual coherence of a Gabor dictionary is typically above $1/2$, and this theorem thus does not apply.

12.5.3 Support Recovery with l^1 Pursuits

An $\mathbf{l}^1$ Lagrangian pursuit has properties similar to an orthogonal matching pursuit, with some improvements that are explained. If the best M-term approximation f_Λ of f satisfies $\mathrm{ERC}(\Lambda) < 1$, then an $\mathbf{l}^1$ Lagrangian pursuit also computes a signal approximation with an error comparable to the minimum M-term error.

An $\mathbf{l}^1$ Lagrangian pursuit (12.89) computes a sparse approximation $\tilde{f} = \Phi^* \tilde{a}$ of f, which satisfies

$$\tilde{a} = \underset{a \in \mathbb{C}^P}{\mathrm{argmin}} \frac{1}{2} \|f - \Phi^* a\|^2 + T \|a\|_1. \tag{12.140}$$

Let $\tilde{\Lambda}$ be the support of $\tilde{a}$. Theorem 12.8 proves that $\tilde{a}$ is the solution of (12.140) if and only if there exists $h \in \mathbb{R}^P$ such that

$$\Phi(\Phi^* \tilde{a} - f) + Th = 0 \quad \text{where} \quad \begin{cases} h[p] = \text{sign}(\tilde{a}[p]) & \text{if} \quad p \in \tilde{\Lambda} \\ |h[p]| \leqslant 1 & \text{if} \quad p \notin \tilde{\Lambda}. \end{cases} \tag{12.141}$$

Theorem 12.15, proved by Tropp [463] and Fuchs [266], shows that the resulting approximation satisfies nearly the same error upper bound as an orthogonal matching pursuit in Theorem 12.13.

Theorem 12.15: *Fuchs, Tropp.* Let $A_\Lambda > 0$ be the lower Riesz bound of $\{\phi_p\}_{p \in \Lambda}$. If $\text{ERC}(\Lambda) < 1$ and

$$T = \lambda \frac{\|f - f_\Lambda\|}{1 - \text{ERC}(\Lambda)} \quad \text{with} \quad \lambda > 1, \tag{12.142}$$

then there exists a unique solution $\tilde{a}$ with support that satisfies $\tilde{\Lambda} \subset \Lambda$, and $\tilde{f} = \Phi^* \tilde{a}$ satisfies

$$\|f - \tilde{f}\|^2 \leqslant \left(1 + \frac{\lambda^2 |\Lambda|}{A_\Lambda (1 - \text{ERC}(\Lambda))^2}\right) \|f - f_\Lambda\|^2. \tag{12.143}$$

Proof. The proof begins by computing a solution $\tilde{a}$ with support that is in Λ, and then proves that it is unique. We denote by a_Λ a vector defined over the index set Λ. To compute a solution with a support in Λ, we consider a solution $\tilde{a}_\Lambda$ of the following problem:

$$\tilde{a}_\Lambda = \underset{a_\Lambda \in \mathbb{C}^{|\Lambda|}}{\text{argmin}} \frac{1}{2} \|f - \Phi_\Lambda^* a_\Lambda\|^2 + T \|a_\Lambda\|_1. \tag{12.144}$$

Let $\tilde{a}$ be defined by $\tilde{a}[p] = \tilde{a}_\Lambda[p]$ for $p \in \Lambda$ and $\tilde{a}[p] = 0$ for $p \in \Lambda^c$. It has a support $\tilde{\Lambda} \subset \Lambda$. We prove that $\tilde{a}$ is also the solution of the $\mathbf{l}^1$ Lagrangian minimization (12.140) if (12.142) is satisfied.

Let h be defined by

$$Th = \Phi(f - \Phi^* \tilde{a}_\Lambda) = \Phi(f - \Phi_\Lambda^* \tilde{a}_\Lambda). \tag{12.145}$$

The optimality condition (12.141) applied to the minimization (12.144) implies that

$$\forall p \in \Lambda, \quad h[p] = \text{sign}(\tilde{a}_\Lambda[p]).$$

To prove that $\tilde{a}_\Lambda$ is the solution of (12.140), we must verify that $|h[q]| \leqslant 1$ for $q \in \Lambda^c$. Equation (12.145) shows that the coefficients h_Λ of h inside Λ satisfy

$$Th_\Lambda = \Phi_\Lambda(f - \Phi_\Lambda^* \tilde{a}_\Lambda).$$

Since $A_\Lambda > 0$, the vectors indexed by Λ are linearly independent and

$$\tilde{a}_\Lambda = \Phi_\Lambda^{*\,+} f - T(\Phi_\Lambda \Phi_\Lambda^*)^{-1} h_\Lambda = a_\Lambda - T(\Phi_\Lambda \Phi_\Lambda^*)^{-1} h_\Lambda, \tag{12.146}$$

and (12.145) implies that

$$h_{\Lambda^c} = \frac{1}{T} \Phi_{\Lambda^c} \left(f - f_\Lambda + T \Phi_\Lambda^* (\Phi_\Lambda \Phi_\Lambda^*)^{-1} h_\Lambda\right). \tag{12.147}$$

The expression (12.147) of h_{Λ^c} shows that

$$\|h_{\Lambda^c}\|_\infty = T^{-1} \max_{q \in \Lambda^c} |\langle \phi_q,\ f - f_\Lambda + T\Phi_\Lambda^* (\Phi_\Lambda \Phi_\Lambda^*)^{-1} h_\Lambda \rangle| \tag{12.148}$$

$$\leq T^{-1} \max_{q \in \Lambda^c} |\langle \phi_q,\ f - f_\Lambda \rangle| + \max_{q \in \Lambda^c} |\langle (\Phi_\Lambda \Phi_\Lambda^*)^{-1} \Phi_\Lambda \phi_q,\ h_\Lambda \rangle|. \tag{12.149}$$

We saw in (12.130) that $(\Phi_\Lambda \Phi_\Lambda^*)^{-1} \Phi_\Lambda = \Phi_\Lambda (\Phi_\Lambda^* \Phi_\Lambda)^{-1} = (\Phi_\Lambda^+)^*$. Since $\mathrm{ERC}(\Lambda) = \max_{q \in \Lambda^c} \|(\Phi^+)^* \phi_q\|_1$ and $\|h_\Lambda\|_\infty \leq 1$, we get

$$\|h_{\Lambda^c}\|_\infty \leq T^{-1} \max_{q \in \Lambda^c} |\langle \phi_q,\ f - f_\Lambda \rangle| + \mathrm{ERC}(\Lambda) \leq T^{-1}\, \|f - f_\Lambda\| + \mathrm{ERC}(\Lambda). \tag{12.150}$$

But $T > \|f - f_\Lambda\| (1 - \mathrm{ERC}(\Lambda))^{-1}$, so $\|h_{\Lambda^c}\|_\infty < 1$, which proves that $\tilde{a}$ is indeed a solution with $\tilde{\Lambda} \subset \Lambda$.

To prove that $\tilde{a}$ is the unique solution of (12.140), suppose that $\tilde{a}_1$ is another solution. Then, necessarily, $\Phi^* \tilde{a} = \Phi^* \tilde{a}_1$ because otherwise the coefficients $(\tilde{a} + \tilde{a}_1)/2$ would have a strictly smaller Lagrangian. This proves that

$$\forall\, p \notin \Lambda, \quad |\langle f - \Phi^* \tilde{a},\ \phi_p \rangle| = |\langle f - \Phi^* \tilde{a}_1,\ \phi_p \rangle| < T,$$

and thus that $\tilde{a}_1$ is also supported inside Λ. Since $\Phi_\Lambda^* \tilde{a}_\Lambda = \Phi_\Lambda^* \tilde{a}_{1,\Lambda}$ and Φ_Λ^* is invertible, $\tilde{a} = \tilde{a}_1$, which proves that the solution of (12.140) is unique.

Let us now prove the approximation result (12.143). The optimality conditions (12.141) prove that

$$\|\Phi_\Lambda (\Phi^* \tilde{a} - f)\|_\infty = \max_{p \in \Lambda} |\langle \phi_p, \tilde{f} - f \rangle| \tilde{f} - f \rangle| \leq T.$$

For any $p \in \Lambda$ $\langle \phi_p, f \rangle = \langle \phi_p, f_\Lambda \rangle$, so

$$\max_{p \in \Lambda} |\langle \phi_p, \tilde{f} - f_\Lambda \rangle| \leq T.$$

Since $\tilde{\Lambda} \subset \Lambda$, it results that $\tilde{f} - f_\Lambda \in \mathbf{V}_\Lambda$, and since $\|\Phi_\Lambda h\|^2 \geq A_\Lambda \|h\|^2$, we get

$$T \geq \max_{p \in \Lambda} |\langle f_\Lambda - \tilde{f},\ \phi_p \rangle| \geq \frac{1}{\sqrt{|\Lambda|}} \|\Phi_\Lambda (f_\Lambda - \tilde{f})\| \geq \frac{\sqrt{A_\Lambda}}{\sqrt{|\Lambda|}} \|f_\Lambda - \tilde{f}\|. \tag{12.151}$$

Since $T = \lambda\, \|f - f_\Lambda\| (1 - \mathrm{ERC}(\Lambda))^{-1}$, it results that

$$\|f_\Lambda - \tilde{f}\| \leq \frac{\lambda \sqrt{|\Lambda|}}{\sqrt{A_\Lambda}\, (1 - \mathrm{ERC}(\Lambda))} \|f - f_\Lambda\|. \tag{12.152}$$

Moreover, $f - f_\Lambda$ is orthogonal to $f_\Lambda - \tilde{f} \in \mathbf{V}_\Lambda$, so

$$\|f - \tilde{f}\|^2 = \|f - f_\Lambda\|^2 + \|f_\Lambda - \tilde{f}_{\Lambda_0}\|^2.$$

Inserting (12.152) proves (12.143). ■

This theorem proves that if T is sufficiently large, then the $\mathbf{l}^1$ Lagrangian pursuit selects only vectors in Λ. At one point it stops selecting vectors in Λ but the error has already reached the upper bound (12.143). If $f = f_\Lambda$ and $\mathrm{ERC}(\Lambda) < 1$, then

(12.143) proves that a basis pursuit recovers all the atoms in Λ and thus reconstructs f. It is, therefore, an Exact Recovery Criterion for a basis pursuit.

An exact recovery is obtained by letting T go to zero in a Lagrangian pursuit, which is equivalent to solve a basis pursuit

$$\tilde{a} = \underset{a \in \mathbb{C}^P}{\operatorname{argmin}} \|a\|_1 \quad \text{subject to} \quad \Phi^* a = f. \tag{12.153}$$

Suppose that $f_M = f_\Lambda$ is the best M-term approximation of f from $M \leq \frac{1}{3\mu(\mathcal{D})}$ dictionary vectors. Similar to Theorem 12.14, Theorem 12.15 implies that the $\mathbf{l}^1$ pursuit approximation computed with $T = \|f - f_M\|/2$ satisfies

$$\|f - \tilde{f}\|^2 \leq (1 + 6M)\, \|f - f_M\|^2. \tag{12.154}$$

Indeed, as in the proof of Theorem 12.14, we verify that $(\mathrm{ERC}(\Lambda) - 1)^{-1} < 2$ and $A_\Lambda \geq 2/3$.

Although $\mathrm{ERC}(\Lambda) > 1$, the support Λ of $f \in \mathbf{V}_\Lambda$ may still be recovered by a basis pursuit. Figure 12.20 gives an example with two Gabor atoms with Heisenberg boxes, which overlap and that are recovered by a basis pursuit despite the fact that $\mathrm{ERC}(\Lambda) > 1$. For an $\mathbf{l}^1$ Lagrangian pursuit, the condition $\mathrm{ERC}(\Lambda) < 1$ can be refined with a more precise sufficient criterion introduced by Fuchs [266]. It depends on the sign of the coefficients $a[p]$ supported in Λ, which recover $f = \Phi^* a$. In (12.150) as well as in all subsequent derivations and thus in the statement of Theorem 12.15, $\mathrm{ERC}(\Lambda) = \max_{q \in \Lambda^c} \sum_{p \in \Lambda} |\langle \tilde{\phi}_{p,\Lambda}, \phi_q \rangle|$ can be replaced by

$$F(f, \Lambda) = \max_{q \in \Lambda^c} \sum_{p \in \Lambda} \langle \tilde{\phi}_{p,\Lambda}, \phi_q \rangle \operatorname{sign}(a[p]). \tag{12.155}$$

In particular, if $F(f, \Lambda) < 1$, then the support Λ of f is recovered by a basis pursuit. Moreover, Dossal [231] showed that if there exists $\tilde{\Lambda}$ with $\Lambda \subset \tilde{\Lambda}$ such that $F(f, \tilde{\Lambda}) < 1$, with arbitrary signs for $a[p]$ when $p \in \tilde{\Lambda} - \Lambda$, then f is also recovered by a basis pursuit. By using this criteria, one can verify that if $|\Lambda|$ is small, even though Λ may include very correlated vectors such as Gabor atoms of close time and frequency, we are more likely to recover Λ with a basis pursuit then with an orthogonal matching pursuit, but this recovery can be unstable.

Image Source Separation

The ability to recover sparse approximation supports in redundant dictionaries has applications to source separation with a single measurement, as proposed by Elad et al. [244]. Let $f = f_0 + f_1$ be a mixture of two signals f_0 and f_1 that have a sparse representation over different dictionaries $\mathcal{D}_0 = \{\phi_p\}_{p \in \Gamma_0}$ and $\mathcal{D}_1 = \{\phi_p\}_{p \in \Gamma_1}$. If a sparse representation $\tilde{a}$ of f in $\mathcal{D} = \mathcal{D}_0 \cup \mathcal{D}_1$ nearly recovers the approximation support of f_0 in $\mathcal{D}_0$ and of f_1 in $\mathcal{D}_1$, then both signals are separately approximated with

$$\tilde{f}_0 = \sum_{p \in \Gamma_0} \tilde{a}[p]\, \phi_p \quad \text{and} \quad \tilde{f}_1 = \sum_{p \in \Gamma_1} \tilde{a}[p]\, \phi_p. \tag{12.156}$$

An application is given to separate edges from oscillatory textures in images.

(a) (b) (c)

FIGURE 12.25

Image separation $f = f_0 + f_1$ in a dictionary that is a union of a wavelet and a local cosine dictionary: **(a)** image f, **(b)** piecewise regular component f_0, and **(c)** oscillatory texture f_1.

To take into account the differences between edges and textures, Meyer [47] introduced an image model $f = f_0 + f_1$, where f_0 is a bounded variation function including edges, and f_1 is an oscillatory texture function that belongs to a different functional space. Theorem 9.17 proves that bounded variation images f_0 are sparse in a translation-invariant dictionary $\mathcal{D}_0$ of wavelets. Dictionaries of curvelets in Section 5.5.2 or bandlets in Section 12.2.4 can also improve the approximations of geometrically regular edges in f_0.

The oscillatory image f_1 has well-defined local frequencies and is therefore sparse in a dictionary $\mathcal{D}_1$ of two-dimensional local cosine bases, defined in Section 8.5.3. A dictionary $\mathcal{D} = \mathcal{D}_0 \cup \mathcal{D}_1$ is defined as a union of a wavelet dictionary and a local cosine dictionary [244]. A sparse representation $\tilde{a}$ of f in $\mathcal{D}$ is computed with an $\mathbf{l}^1$ basis pursuit, and approximations $\tilde{f}_0$ and $\tilde{f}_1$ of f_0 and f_1 are computed with (12.156). Figure 12.25 shows that this algorithm can indeed separate oscillating textures from piecewise regular image variations in such a dictionary.

12.6 MULTICHANNEL SIGNALS

Multiple channel measurements often have strong dependencies that a representation should take into account. For color images, the green, blue, and red (RGB) channels are highly correlated. Indeed, edges and sharp variations typically occur at the same location in each color channel. Stereo audio recordings or multiple point recordings of EEGs also output dependent measurement vectors. Taking into account the structural dependancies of these channels improves compression or denoising applications, but also provides solutions to the source separation problems studied in Section 13.5.

A signal with K channels is considered as a signal vector $\vec{f}[n] = (f_k[n])_{0 \leq k < K}$. The Euclidean norm of a vector $\vec{a} = (a_k)_{0 \leq k < K} \in \mathbb{C}^K$ is written as $\|\vec{a}\|^2 = \sum_{k=0}^{K-1} |a_k|^2$.

The Frobenius norm of a signal vector is

$$\|\vec{f}\|_F^2 = \sum_{k=0}^{K-1} \|f_k\|^2.$$

Whitening with Linear Channel Decorrelation

A linear decorrelation and renormalization of the signal channels can be implemented with an operator O, which often improves further multichannel processing. The empirical covariance of the K channels is computed from L signal vector examples $\vec{x}_l = (x_{l,k})_{0 \leq k < K}$:

$$\mu_k = \frac{1}{LN} \sum_{l=0}^{L-1} \sum_{n=0}^{N-1} x_{l,k}[n]$$

and

$$c_{k,k'} = \frac{1}{LN} \sum_{l=0}^{L-1} \sum_{n=0}^{N-1} (x_{l,k}[n] - \mu_k)\,(x_{l,k'}[n] - \mu_{k'}).$$

Let $C = (c_{k,k'})_{0 \leq k,k' \leq K}$ be the empirical covariance matrix. The whitening operator $O = C^{-1/2}$ performs a decorrelation and a renormalization of all channels.

For color images, the change of coordinates from (R, G, B) to (Y, U, V) typically implements such a decorrelation. In noise-removal applications, the noise can be decorrelated across channels by computing C from recordings $\vec{x}_l$ of the noise.

12.6.1 Approximation and Denoising by Thresholding in Bases

We consider multichannel signals over which a whitening operator may already have been applied. Approximation and denoising operators are defined by simultaneously thresholding all the channel coefficients in a dictionary. Let $\mathcal{D} = \{\phi_p\}_{p \in \Gamma}$ be a basis of unit vectors. For any $\vec{f} = (f_k)_{0 \leq k < K}$, we write an inner product vector:

$$\langle \vec{f}, \phi_p \rangle = \Big(\langle f_k, \phi_p \rangle \Big)_{0 \leq k < K} \in \mathbb{C}^K.$$

If $\mathcal{D}$ is an orthonormal basis, then one can verify (Exercise 9.4) that a best M-term approximation $\vec{f}_M$ that minimizes $\|\vec{f} - \vec{f}_M\|_F^2$ is obtained by selecting the M inner product vectors having the largest norm $\|\langle \vec{f}, \phi_p \rangle\|$. Such nonlinear approximations are thus calculated by thresholding the norm of these multichannel inner product vectors:

$$\vec{f}_{\Lambda_T} = \sum_{p \in \Lambda_T} \langle \vec{f}, \phi_p \rangle\, \phi_p \quad \text{with} \quad \Lambda_T = \{p \in \Gamma \,:\, \|\langle \vec{f}, \phi_p \rangle\| \geqslant T\}. \tag{12.157}$$

Let $\vec{W}$ be a random noise vector. We suppose that a whitening operator has been applied so that $\vec{W}[n] = (W_k[n])_{0 \leq k < K}$ is decorrelated across channels, and that each $W_k[n]$ is a Gaussian white noise. A multichannel estimation of $\vec{f}$ from

noisy measurements $\vec{X} = \vec{f} + \vec{W}$ is implemented with a block thresholding, as defined in Section 11.4.1. A hard-block thresholding estimation is an orthogonal projector

$$\tilde{F} = \sum_{p \in \tilde{\Lambda}_T} \langle \vec{X}, \phi_p \rangle \, \phi_p \quad \text{with} \quad \tilde{\Lambda}_T = \{ p \in \Gamma \, : \, \| \langle \vec{X}, \phi_p \rangle \| \geqslant T \}.$$

A James-Stein soft-block thresholding attenuates the amplitude of each inner product vector:

$$\tilde{F} = \sum_{p \in \Gamma} \max \left(1 - \frac{T^2}{\| \langle \vec{X}, \phi_p \rangle \|^2} \, , \, 0 \right) \langle \vec{X}, \phi_p \rangle \, \phi_p .$$

The risk properties of such block thresholding estimators are studied in Section 11.4.1. Vector thresholding of color images improves the SNR and better preserves colors by attenuating all color channels with the same factors.

12.6.2 Multichannel Pursuits

Multichannel signals decomposition in redundant dictionaries are implemented with pursuit algorithms that simultaneously approximate all the channels. Several studies describe the properties of multichannel pursuits and their generalizations and applications [157, 282, 346, 464].

Matching Pursuit

The matching pursuit algorithm in Section 12.3.1 is extended by searching for dictionary elements that maximize the norm of the multichannel inner product vector. We set the relaxation parameter $\alpha = 1$. Let $\mathcal{D} = \{\phi_p\}_{p \in \Gamma}$ be a dictionary of unit vectors. The matching pursuit algorithm is initialized with $R^0 \vec{f} = \vec{f}$. At each iteration m, it selects a best vector $\phi_{p_m} \in \mathcal{D}$ such that

$$\| \langle R^m \vec{f}, \phi_{p_m} \rangle \| = \operatorname*{argmax}_{p \in \Gamma} \| \langle R^m \vec{f}, \phi_p \rangle \| . \tag{12.158}$$

The orthogonal projection on this vector defines a new residue

$$R^{m+1} \vec{f} = R^m \vec{f} - \langle R^m \vec{f}, \phi_{p_m} \rangle \, \phi_{p_m}$$

with an energy conservation

$$\| R^{m+1} \vec{f} \|_F^2 = \| R^m \vec{f} \|_F^2 + \| \langle R^m \vec{f}, \phi_{p_m} \rangle \|^2 .$$

Theorem 12.6 thus remains valid by using the Frobenius norm over signal vectors, which proves the exponential convergence of the algorithm:

$$\vec{f} = \sum_{m=0}^{+\infty} \langle \vec{f}, \phi_{p_m} \rangle \phi_{p_m}$$

After M matching pursuit iterations, a backprojection algorithm recovers the orthogonal projection of $\vec{f}$ on the selected atoms $\{\phi_{p_m}\}_{0 \leq m < M}$. It can be computed with a Richardson gradient descent described afterwards for backprojecting an $\mathbf{l}^1$ pursuit, which is initialized with $\vec{a}_0[p_m] = \langle R^m \vec{f}, \phi_{p_m} \rangle$.

The orthogonal matching pursuit of Section 12.3.2 is similarly extended. At a step m, a vector ϕ_{p_m} is selected as in (12.158). A Gram-Schmidt orthogonalization decomposes ϕ_{p_m} into its projection over the previously selected vectors $\{\phi_{p_k}\}_{0 \leq k < m}$ plus an orthogonal complement u_m. The orthogonalized residue is then

$$R^{m+1}\vec{f} = R^m \vec{f} - \frac{\langle R^m \vec{f}, u_m \rangle}{\|u_m\|^2} u_m.$$

It decomposes $\vec{f}$ over an orthogonal family $\{u_m\}_m$ and for signals of size N:

$$\vec{f} = \sum_{m=0}^{N-1} \frac{\langle R^m \vec{f}, u_m \rangle}{\|u_m\|^2} u_m.$$

The orthogonal matching pursuit properties thus remain essentially the same.

Multichannel l^1 Pursuits

Sparse multichannel signal representations in redundant dictionaries can also be computed with $\mathbf{l}^1$ pursuits, which minimize an $\mathbf{l}^1$ norm of the coefficients. An $\mathbf{l}^1$ norm over multichannel vectors of coefficients is defined by

$$\|\vec{a}\|_1 = \sum_{p \in \Gamma} \|\vec{a}[p]\|.$$

We denote

$$\Phi^* \vec{a} = \sum_{p \in \Gamma} \vec{a}[p]\, \phi_p.$$

A sparse $\mathbf{l}^1$ pursuit approximation of a vector $\vec{f}$ at a precision ε is defined by $\tilde{\vec{f}} = \Phi^* \tilde{\vec{a}}[p]$ with

$$\tilde{\vec{a}} = \underset{\vec{a} \in \mathbb{C}^{PK}}{\operatorname{argmin}} \|\vec{a}\|_1 \quad \text{subject to} \quad \|\vec{f} - \Phi^* \vec{a}\|_F \leq \varepsilon.$$

A solution of this convex optimization is computed with an $\mathbf{l}^1$ Lagrangian minimization

$$\tilde{\vec{a}} = \underset{\vec{a} \in \mathbb{C}^{PK}}{\operatorname{argmin}} \frac{1}{2} \|\vec{f} - \Phi^* \vec{a}\|_F^2 + T \|\vec{a}\|_1, \tag{12.159}$$

where T depends on ε. Several authors have studied the approximation properties of $\mathbf{l}^1$ vector pursuits and their generalizations [157, 188, 462].

The $\mathbf{l}^1$ Lagrangian minimization (12.159) is numerically solved with the iterative thresholding algorithm of Section 12.4.3, which is adjusted as follows. We write $\Phi\vec{f}[p] = \langle \vec{f}, \phi_p \rangle$.

1. *Initialization*. Choose $\vec{a}_0$, set $k = 0$, and compute $\vec{b} = \Phi\vec{f}$.
2. *Gradient step*. Update

$$\bar{\vec{a}}_k = \vec{a}_k + \gamma\,(\vec{b} - \Phi\,\Phi^*\vec{a}_k), \tag{12.160}$$

   where $\gamma < 2\,\|\Phi\Phi^*\|_S^{-1}$.
3. *Soft thresholding*. Compute

$$\vec{a}_{k+1}[p] = \rho_{\gamma T}(\bar{\vec{a}}_k[p]), \tag{12.161}$$

   where $\rho_{\gamma T}(\vec{x}) = \vec{x}\,\max\left(1 - \frac{\gamma T}{\|\vec{x}\|},\ 0\right)$.
4. *Stop*. If $\|\vec{a}_k - \vec{a}_{k+1}\|_F$ is smaller than a fixed-tolerance criterion, stop the iterations; otherwise, set $k \leftarrow k+1$ and go back to 2.

If $\tilde{\Lambda}$ is the support of the computed solution after convergence, then like for a matching pursuit, a backprojection algorithm computes the orthogonal projection of $\vec{f}$ on the selected atoms $\{\phi_p\}_{p\in\tilde{\Lambda}}$. It can be implemented with a Richardson gradient descent. It continues the gradient descent iterations of step 2, and replaces the soft thresholding in step 3 by a projector defined by $\vec{a}_{k+1}[p] = \bar{\vec{a}}_k[p]$ if $p \in \tilde{\Lambda}$ and $\vec{a}_{k+1}[p] = 0$ if $p \notin \tilde{\Lambda}$.

Multichannel Dictionaries

Multichannel signals $\vec{f}$ have been decomposed over dictionaries of scalar signals $\{\phi_p\}_{p\in\Gamma}$, thus implying that the same dictionary elements $\{\phi_p\}_{p\in\Lambda}$ are appropriate to approximate all signal channels $(f_k)_{0\leq k<K}$. More flexibility can be provided by dictionaries of multiple channel signals $\mathcal{D} = \{\vec{\phi}_p\}_{p\in\Gamma}$ where each $\vec{\phi}_p = (\phi_{p,k})_{0\leq k<K}$ includes K channels. In the context of color images, this means constructing a dictionary of color vectors having three color channels. Applying color dictionaries to color images can indeed improve the restitution of colors in noise-removal applications [357].

Inner product vectors and projectors over dictionary vectors are written as

$$\langle \vec{f}, \vec{\phi}_p \rangle = \left(\langle f_k, \phi_{p,k} \rangle\right)_{0\leq k<K} \quad \text{and} \quad \langle \vec{f}, \vec{\phi}_p \rangle\, \vec{\phi}_p = \left(\langle f_k, \phi_{p,k} \rangle\, \phi_{p,k}\right)_{0\leq k<K}.$$

The thresholds and pursuit algorithms of Sections 12.6.1 and 12.6.2 decompose $\vec{f}$ in a dictionary of scalar signals $\{\phi_p\}_{p\in\Gamma}$. They are directly extended to decompose $\vec{f}$ in a dictionary of signal vectors $\mathcal{D} = \{\vec{\phi}_p\}_{p\in\Gamma}$. In all formula and algorithms

$$\langle \vec{f}, \phi_p \rangle \quad \text{is replaced by} \quad \langle \vec{f}, \vec{\phi}_p \rangle$$

and

$$\langle \vec{f}, \phi_p \rangle \, \phi_p \quad \text{is replaced by} \quad \langle \vec{f}, \vec{\phi}_p \rangle \, \vec{f}, \vec{\phi}_p \rangle \, \vec{\phi}_p .$$

For a fixed index p, instead of decomposing all signal channels f_k on the same ϕ_p, the resulting algorithms decompose each f_k on a potentially different $\phi_{p,k}$, but all channels make a simultaneous choice of a dictionary vector $\vec{\phi}_p$. Computations and mathematical properties are otherwise the same. The difficulty introduced by this flexibility is to construct dictionaries specifically adapted to each signal channel. Dictionary learning provides an approach to solve this issue.

12.7 LEARNING DICTIONARIES

For a given dictionary size P, the dictionary should be optimized to best approximate signals in a given set Θ. Prior information on signals can lead to appropriate dictionary design, for example, with Gabor functions, wavelets, or local cosine vectors. These dictionaries, however, can be optimized by better taking into account the signal properties derived from examples. Olshausen and Field [391] argue that such a learning process is part of biological evolution, and could explain how the visual pathway has been optimized to extract relevant information from visual scenes. Many open questions remain on dictionary learning, but numerical algorithms show that learning is possible and can improve applications.

Let us consider a family of K signal examples $\{f_k\}_{0 \leq k < K}$. We want to find a dictionary $\mathcal{D} = \{\phi_p\}_{p \in \Gamma}$ of size $|\Gamma| = P$ in which each f_k has an "optimally" sparse approximation

$$\tilde{f}_k = \sum_{p \in \Gamma} a[k, p] \, \phi_p , \tag{12.162}$$

given a precision $\|f_k - \tilde{f}_k\|^2 \leq \varepsilon$. This sparse decomposition may be computed with a matching pursuit, an orthogonal matching pursuit, or an $\mathbf{l}^1$ pursuit. Let $\Phi f_k = \{\langle f_k, \phi_p \rangle\}_{p \in \Gamma}$ be the dictionary operator with rows equal to the dictionary vectors ϕ_p. The learning process iteratively adjusts $\mathcal{D}$ to optimize the sparse representation of all examples.

Dictionary Update

Following the work of Olshausen and Field [391], several algorithms have been proposed to optimize $\mathcal{D}$, and thus Φ, from a family of examples [78, 246, 336, 347]. It is a highly non-convex optimization problem that therefore can be trapped in local minima.

The approach of Engan, Aase, and Husoy [246] performs alternate optimizations, similar to the Lloyd-Max algorithm for learning code books in vector

quantization [27]. Let us write $\vec{f} = \{f_k\}_{0 \leq k < K}$. Its Frobenius norm is

$$\|\vec{f}\|_F^2 = \sum_{k=0}^{K-1} \|f_k\|^2.$$

According to (12.162), the approximation vector $\tilde{\vec{f}} = \{\tilde{f}_k\}_{0 \leq k < K}$ can be written as $\tilde{\vec{f}} = A\Phi$. The algorithm alternates between the calculation of the matrix of sparse signal coefficients $A = (a[k,p])_{0 \leq k < K, p \in \Gamma}$ and a modification of the dictionary vectors ϕ_p to minimize the Frobenius norm of the residual error

$$\|\vec{f} - A\Phi\|_F^2 = \sum_{k=0}^{K-1} \|f_k - \sum_{p \in \Gamma} a[k,p]\, \phi_p\|^2.$$

The matrix Φ can be considered as a vector transformed by the operator A. As explained in Section 5.1.3, the error $\|\vec{f} - A\Phi\|_F^2$ is minimum if $A\Phi$ is the orthogonal projection of $\vec{f}$ in the image space of A. It results that Φ is computed with the pseudo inverse A^+ of A:

$$\Phi = A^+ \vec{f} = (A^* A)^{-1} A^* \vec{f}.$$

The inversion of the operator $L = A^* A$ can be implemented with the conjugate-gradient algorithm in Theorem 5.8. The resulting learning algorithm proceeds as follows:

1. *Initialization*. Each vector ϕ_p is initialized as a white Gaussian noise with a norm scaled to 1.
2. *Sparse approximation*. Calculation with a pursuit of the matrix $A = (a[k,p])_{0 \leq k < K, p \in \Gamma}$ of sparse approximation coefficients

$$\|f_k - \sum_{p \in \Gamma} a[k,p] \phi_p\| \leq \varepsilon \quad \text{for} \quad 0 \leq k < K. \tag{12.163}$$

3. *Dictionary update*. Minimization of the residual error (12.163) with

$$\Phi = A^+ \vec{f} = (A^* A)^{-1} A^* \vec{f}. \tag{12.164}$$

4. *Dictionary normalization*. Each resulting row ϕ_p of Φ is normalized: $\|\phi_p\| = 1$.
5. *Stopping criterion*. After a fixed number of iterations, or if Φ is marginally modified, then stop; otherwise, go back to 2.

This algorithm is computationally very intensive since it includes a sparse approximation at each iteration and requires the inversion of the $P \times P$ matrix $A^* A$. The sparse approximations are often calculated with an orthogonal pursuit that provides a good precision versus calculation trade-off. When the sparse coefficients $a[k,p]$ are computed with an $\mathbf{l}^1$ pursuit, then one can prove that the algorithm converges to a stationary point [466].

To compress structured images such as identity photographs, Bryt and Elad showed that such learning algorithms are able to construct highly efficient dictionaries [123]. It is also used for video compression with matching pursuit by optimizing a predefined dictionary, which improves the distortion rate [426].

Translation-Invariant Dictionaries

For noise removal or inverse problems, the estimation of stationary signals is improved with translation-invariant dictionaries. As a result, the algorithm must only learn the P generators of this translation-invariant dictionary. A maximum support size for these dictionary vectors is set to a relatively small value of typically $N = 16 \times 16$ pixels. The examples $\{f_k\}_{0 \leq k < K}$ are then chosen to be small patches of N pixels, extracted from images.

Optimized translation-invariant dictionaries lead to high-quality, state-of-the-art noise-removal results [242]. For color images, it can also incorporate intrinsic redundancy between the different color channels by learning color vectors [357]. Section 12.6 explains how to extend pursuit algorithms to multiple channel signals such as color images. Besides denoising, these dictionaries are used in super-resolution inverse problems—for example, to recover missing color pixels in high-resolution color demosaicing [357].

Figure 12.26 shows an example of a dictionary learned with this algorithm with $N = 16^2$ and $P = 2N$. The calculated dictionary vectors ϕ_p look similar to the directional wavelets of Section 5.5. Olshausen and Field [391] observed that these dictionary vectors learned from natural images also look similar to the impulse response of *simple cell neurons* in the visual cortical area V1. It supports the idea of a biological adaptation to process visual scenes.

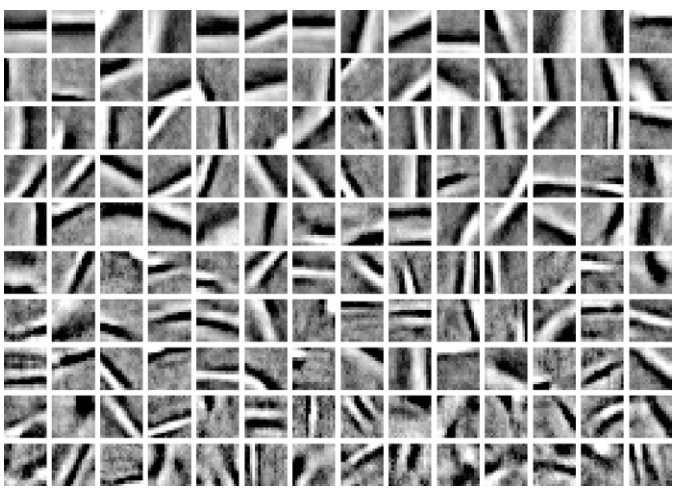


FIGURE 12.26

Dictionary $\{\phi_p\}_{p \in \Gamma}$ learned from examples extracted from natural images.

12.8 EXERCISES

12.1 [2] *Best wavelet packet and local cosine approximations*:

(a) Synthesize a discrete signal that is well approximated by a few vectors in a best wavelet packet basis, but that requires many more vectors to obtain an equivalent approximation in a best local cosine basis. Test your signal numerically.

(b) Design a signal that is well approximated in a best local cosine basis but requires many more vectors to approximate it efficiently in a best wavelet packet basis. Verify your result numerically.

12.2 [3] Describe a coding algorithm that codes the position of M nonzero orthogonal coefficients in a best wavelet packet or local cosine dictionary tree of size $P = N \log_2 N$, and that requires less than $R_0 = \log_2 \binom{M}{P}$ bits. How many bits does your algorithm require?

12.3 [2] A double tree of block wavelet packet bases is defined in Exercise 8.12. Describe a fast best-basis algorithm that requires $O(N(\log_2 N)^2)$ operations to find the block wavelet packet basis that minimizes an additive cost (12.34) [299].

12.4 [2] In a dictionary $\mathcal{D} = \{\phi_p\}_{p\in\Gamma}$ of orthonormal bases, we write $a_o[p]$ a vector of orthogonal coefficients with a support Λ_0 corresponding to an orthonormal family $\{\phi_p\}_{p\in\Lambda_o}$. We want to minimize the $\mathbf{l}^1$ Lagrangian among orthogonal coefficients:

$$\mathcal{L}_1(T, f, \tilde{a}) = \min_{a_o} \frac{1}{2}\|f - \sum_{p\in\Gamma} a_0[p]\,\phi_p\|^2 + T \sum_{p\in\Gamma} |a_0[p]|.$$

(a) Verify that $\mathcal{L}_1(T, f, a_0) = 1/2 \sum_{p\in\Lambda_o} |\langle f, \phi_p\rangle - a_o[p]|^2 + T \sum_{p\in\Lambda_o} |a_0[p]|$.

(b) Prove that $e(\alpha) = (\beta-\alpha)^2/2 + T|\alpha|$ is minimized by $\alpha = \rho_T(\beta)$ where $\rho_T(x) = x \max(1 - T/|x|, 0)$ is a soft thresholding.

(c) Prove that if the dictionary is reduced to a single orthonormal basis $\mathcal{B} = \{\phi_p\}_{p\in\Gamma_\mathcal{B}}$, then the minimum $\mathbf{l}^1$ Lagrangian is

$$\mathcal{L}_1(T, f, \mathcal{B}) = \mathcal{L}_1(T, f, \tilde{a}) = \sum_{p\in\Gamma_\mathcal{B}} C(|\langle f, \phi_p\rangle|)$$

$$\text{with} \quad C(x) = \frac{1}{2}\min(T^2, x^2) + T|x|.$$

(d) In a dictionary of orthonormal bases, describe a best-basis algorithm that finds the minimizer $\tilde{a}$ of $\mathcal{L}_1(T, f, a_o)$ among all vectors of orthogonal coefficients selected in the dictionary.

12.5 [2] Prove that if we add a vector h to an orthonormal basis $\mathcal{B}$, we obtain a redundant dictionary with a mutual coherence that satisfies $\mu(\mathcal{B}\cup\{h\}) \geqslant N^{-1/2}$.

12.6 [2] Let $\mathcal{D}_{j,\Delta}$ be a Gabor dictionary as defined in (12.77). Let Λ be an index set of two Gabor atoms in $\mathcal{D}_{j,\Delta}$.

(a) Compute ERC(Λ) as a function of the distance in time and frequency of both Gabor atoms with (12.81).

(b) Compute the first upper bound in (12.12) of ERC(Λ) and compare its value with ERC(Λ).

(c) Compute the minimum time distance (for atoms at the same frequency) and the minimum frequency distance (for atoms at the same time) so that ERC(Λ) < 1.

12.7 [2] Let $\mathcal{D} = \{\delta[n-k]\,,\ e^{i2\pi kn/N}\}_{0\leq k<N}$ be a Dirac-Fourier dictionary.

(a) Prove that a matching pursuit residue calculated with a relaxation factor $\alpha = 1$ satisfies $\|R^m f\| \leq \|f\|\ \exp(-m/(2N))$.

(b) Prove that if f is a combination of $M \leq N^{1/2}/2$ Fourier and Dirac vectors, then the matching pursuit reconstructs f exactly as a combination of these M Fourier and Dirac vectors.

12.8 [3] *Uncertainty principle* [227]. We consider two orthogonal bases $\mathcal{B}_0, \mathcal{B}_1$ of $\mathbb{R}^N$.

(a) Prove that for any $f \neq 0$,

$$\|f_{\mathcal{B}_0}\|_0\ \|f_{\mathcal{B}_1}\|_0 \geq \frac{1}{\mu(\mathcal{B}_0 \cup \mathcal{B}_1)^2}. \tag{12.165}$$

Hint: Show that $\|f_{\mathcal{B}_0}\|_\infty \leq \mu(\mathcal{B}_0 \cup \mathcal{B}_1)\|f_{\mathcal{B}_1}\|_1$ and use the Cauchy-Schwartz inequality.

(b) Prove that for any $f \neq 0$,

$$\|f_{\mathcal{B}_0}\|_0 + \|f_{\mathcal{B}_1}\|_0 \geq \frac{2}{\mu(\mathcal{B}_0 \cup \mathcal{B}_1)}. \tag{12.166}$$

12.9 [3] *Cumulated coherence.* Let $\mathcal{D} = \{\phi_p\}_{p\in\Gamma}$ be a dictionary with $\|\phi_p\| = 1$. The cumulated mutual coherence of $\mathcal{D}$ is

$$\mu_M(\mathcal{D}) = \max_{|\Lambda|\leq M} \max_{q\in\Lambda^c} \sum_{p\in\Lambda} |\langle \phi_p,\ \phi_q\rangle|.$$

(a) Prove that $\mu_M(\mathcal{D}) \leq M\mu(\mathcal{D})$.

(b) Prove that $\mu_{M-1}(\mathcal{D}) + \mu_M(\mathcal{D}) < 1$ implies that ERC(Λ) < 1 for any Λ such that $|\Lambda| \leq M$. *Hint:* Use Theorem 12.12.

(c) For $\beta < 1$ and for any $p \geq 0$, let

$$\forall\, n \in \mathbb{Z},\quad \phi_p[n] = \begin{cases} 0 & \text{if} \quad n < p \\ \beta^{n-p}\sqrt{1-\beta 2} & \text{if} \quad n \geq p \end{cases}$$

Show that $\mathcal{D}$ spans $\mathbf{l}^2(\mathbb{Z})$ and that $|\langle \phi_p,\ \phi_{p'}\rangle| = \beta^{|p-p'|}$. *Hint:* Consider the case $p > p'$

(d) Show that $\mu_M(\mathcal{D}) < 2\beta/(1-\beta)$ while $\mu(\mathcal{D})M$ grows unbounded with M.

12.10 [2] *Spark*. Let $\mathcal{D} = \{\phi_p\}_{p\in\Gamma}$ be a dictionary with $\|\phi_p\| = 1$. The spark of $\mathcal{D}$, introduced in [229], is

$$\text{spark}(\mathcal{D}) = \min_{b\in\textbf{Null}(\Phi^*), b\neq 0} \|b\|_0.$$

(a) Show that if $\|a\|_0 < \text{spark}(\mathcal{D})/2$, then a is the unique solution of

$$\min_b \|b\|_0 \quad \text{subject to} \quad \Phi^* b = \Phi^* a. \tag{12.167}$$

Hint: If b is a solution of this problem, $a - b \in \textbf{Null}(\Phi^*)$.

(b) Show that $\text{spark}(\mathcal{D}) \geqslant 1 + 1/\mu(\mathcal{D})$. Next, deduce that if $\|a\|_0 < (1 + 1/\mu(\mathcal{D}))/2$, then a is the unique solution of (12.167). *Hint:* Use Theorem 12.10.