An approximation tour
Approximation with pursuits

Mallat 1999 Ch. 9 Section 9.4

Approximations with pursuits

- To optimize non-linear signal approximations, one can adaptively choose the basis depending on the signal
- The set of orthogonal bases is much smaller than the set of non-orthogonal bases that could be constructed by choosing N linearly independent vectors from these P.
- To improve the approximation of complex signals such as music recordings, we study general ***non-orthogonal*** signal decompositions.
- Consider the space of signals of size N. Let

$$D = \{g_p\}_{0 \le p < P}$$

- be a redundant dictionary of P>N vectors which includes at least N linearly independent vectors

Approximations with pursuits

- For $M \geq 1$, an approximation f_M of f can be calculated with a linear combination of any M dictionary vectors

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

- The freedom of choice opens the door to a considerable combinatorial explosion.
- For general dictionaries of P > N vectors, computing the approximation

  f~ that minimizes $||f - f_M||$ is an NP hard problem.

  – This means that there is no known polynomial time algorithm that can solve this optimization.
- Pursuit algorithms *reduce the computational complexity* by searching for *efficient but non-optimal approximations*.

Basis pursuits

- A basis pursuit formulates the search as a linear programming problem, providing remarkably good approximations with "tractable" complexity.
- For large signals, this remains prohibitive. Matching pursuits are faster greedy algorithms that make the problem tractable
- We study the construction of a "best" basis B, not necessarily orthogonal, for efficiently approximating a signal *f*
- The N vectors of the basis

$$\mathcal{B} = \{g_{p_m}\}_{0 \leq m < N}$$

- are selected with a pursuit.

Basis pursuit

- Let us decompose f in the basis

$$f = \sum_{m=0}^{N-1} a[p_m]\, g_{p_m}.$$

- A basis pursuit searches for a basis that minimizes

$$C(f,\mathcal{B}) = \frac{1}{\|f\|} \sum_{m=0}^{N-1} |a[p_m]|.$$

Minimizing the l^1 norm of the decomposition coefficients avoids diffusing the energy of f among many vectors. It reduces cancellations between the vectors $a[p_m]g_{p_m}$, that decompose f, because such cancellations increase $|a[p_m]|$ and thus increase the cost.

The minimization of an l^1 norm is also related to linear programming, which leads to fast computational effective algorithms.

Linear programming

Linear Programming Instead of immediately isolating subsets of N vectors in the dictionary $\mathcal{D}$, a linear system of size P is written with all dictionary vectors

$$\sum_{p=0}^{P-1} a[p]\, g_p[n] = f[n], \tag{9.79}$$

while trying to minimize

$$\sum_{p=0}^{P-1} |a[p]|. \qquad \text{NxP} \tag{9.80}$$

The system (9.79) can be expressed in matrix form with the $P \times N$ matrix $G = \{g_p[n]\}_{0 \leq n < N, 0 \leq p < P}$

$$Ga = f. \tag{9.81}$$

Although the minimization of (9.80) is nonlinear, it can be reformulated as a linear programming problem.

Linear programming

- It can be shown that the solution has at most N non zero coefficients
- In the non degenerate cases, which are most often encountered, the non zero coefficients correspond to N indicis $\{p_m\}_{0 \le m < M}$ such that

$$\{g_{p_m}\}_{0 \le m < N}$$

- are linearly independent.
- This is the best basis of R^N that minimizes the cost.

Matching pursuit

- Despite the linear programming approach, a basis pursuit is computationally expensive because it minimizes a global cost function over all dictionary vectors.
- The matching pursuit introduced by Mallat and Zhang [259] reduces the computational complexity with a greedy strategy.
- Let $\mathcal{D} = \{g_\gamma\}_{\gamma \in \Gamma}$ be a dictionary of P>N vectors having unit norm.
- This dictionary includes N linearly independent vectors that define a basis of the space C^N of signals of size N.
- A matching pursuit begins by projecting f on a vector $g_{\gamma_0} \in \mathcal{D}$ and computing the residue Rf

$$f = \langle f, g_{\gamma_0} \rangle g_{\gamma_0} + Rf.$$

Matching pursuit

Since Rf is orthogonal to g_{γ_0}

$$\|f\|^2 = |\langle f, g_{\gamma_0}\rangle|^2 + \|Rf\|^2. \tag{9.86}$$

To minimize $\|Rf\|$ we must choose $g_{\gamma_0} \in \mathcal{D}$ such that $|\langle f, g_{\gamma_0}\rangle|$ is maximum. In some cases, it is computationally more efficient to find a vector g_{γ_0} that is almost optimal:

$$|\langle f, g_{\gamma_0}\rangle| \geq \alpha \sup_{\gamma\in\Gamma} |\langle f, g_\gamma\rangle|, \tag{9.87}$$

where $\alpha \in (0,1]$ is an optimality factor. The pursuit iterates this procedure by subdecomposing the residue. Let $R^0 f = f$. Suppose that the m^{th} order residue $R^m f$ is already computed, for $m \geq 0$.

Matching pursuit

The next iteration chooses $g_{\gamma_m} \in \mathcal{D}$ such that

$$|\langle R^m f, g_{\gamma_m} \rangle| \geq \alpha \sup_{\gamma \in \Gamma} |\langle R^m f, g_\gamma \rangle|,$$

and projects $R^m f$ on g_{γ_m}:

$$R^m f = \langle R^m f, g_{\gamma_m} \rangle \, g_{\gamma_m} + R^{m+1} f. \qquad (9.89)$$

The orthogonality of $R^{m+1} f$ and g_{γ_m} implies

$$\|R^m f\|^2 = |\langle R^m f, g_{\gamma_m} \rangle|^2 + \|R^{m+1} f\|^2. \qquad (9.90)$$

Matching pursuit

Summing (9.89) from m between 0 and $M-1$ yields

$$f = \sum_{m=0}^{M-1} \langle R^m f, g_{\gamma_m} \rangle \, g_{\gamma_m} + R^M f. \tag{9.91}$$

Similarly, summing (9.90) from m between 0 and $M-1$ gives

$$\|f\|^2 = \sum_{m=0}^{M-1} |\langle R^m f, g_{\gamma_m} \rangle|^2 + \|R^M f\|^2. \tag{9.92}$$

The following theorem proves that $\|R^m f\|$ converges exponentially to 0 when m tends to infinity.

Theorem

Theorem 9.10 *There exists $\lambda > 0$ such that for all $m \geq 0$*

$$\|R^m f\| \leq 2^{-\lambda m} \|f\|. \tag{9.93}$$

As a consequence

$$f = \sum_{m=0}^{+\infty} \langle R^m f, g_{\gamma_m} \rangle \, g_{\gamma_m}, \tag{9.94}$$

and

$$\|f\|^2 = \sum_{m=0}^{+\infty} |\langle R^m f, g_{\gamma_m} \rangle|^2. \tag{9.95}$$

Matching pursuit

- The convergence rate X decreases when the size N of the signal space increases.
- In the limit of infinite dimensional spaces, Jones' theorem proves that the algorithm still converges but the convergence is not exponential [230,259].
- Observe that even in finite dimensions, an infinite number of iterations is necessary to completely reduce the residue.
- In most signal processing applications, this is not an issue because many fewer than N iterations are needed to obtain sufficiently precise signal approximations.

Fast network calculations

- A matching pursuit is implemented with a fast algorithm that computes $\left\langle R^{m+1} f, g_\gamma \right\rangle$ from $\left\langle R^{m} f, g_\gamma \right\rangle$ with a simple updating formula

$$R^m f = \left\langle R^m f, g_{\gamma_m} \right\rangle g_{\gamma_m} + R^{m+1} f$$

➔ $$\left\langle R^{m+1} f, g_\gamma \right\rangle = \left\langle R^m f, g_\gamma \right\rangle - \left\langle R^m f, g_{\gamma_m} \right\rangle \left\langle g_{\gamma_m}, g_\gamma \right\rangle$$

To reduce the computational load, it is necessary to construct dictionaries with vectors having a sparse interaction. This means that each g_γ has non-zero inner products with only a small fraction of all other dictionary vectors

- Dictionaries are designed so that non-zero weights $\left\langle g_\alpha, g_\gamma \right\rangle$ can be retrieved from memory or computed with O(1) operations

Matching pursuit

- A matching pursuit with a relative precision ε is implemented as follows

1. *Initialization* Set $m = 0$ and compute $\{\langle f, g_\gamma \rangle\}_{\gamma \in \Gamma}$.

2. *Best match* Find $g_{\gamma_m} \in \mathcal{D}$ such that

$$|\langle R^m f, g_{\gamma_m} \rangle| \geq \alpha \sup |\langle R^m f, g_\gamma \rangle|. \quad (9.102)$$

3. *Update* For all $g_\gamma \in \mathcal{D}$ with $\langle g_{\gamma_m}, g_\gamma \rangle \neq 0$

$$\langle R^{m+1} f, g_\gamma \rangle = \langle R^m f, g_\gamma \rangle - \langle R^m f, g_{\gamma_m} \rangle \langle g_{\gamma_m}, g_\gamma \rangle.$$

4. *Stopping rule* If

$$\|R^{m+1} f\|^2 = \|R^m f\|^2 - |\langle R^m f, g_{\gamma_m} \rangle|^2 \leq \epsilon^2 \|f\|^2$$

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

Matching pursuit

- If D is highly redundant, computations at steps 2 and 3 are reduced by performing the calculation on a subdictionary Ds

$$\mathcal{D}_s = \{g_\gamma\}_{\gamma \in \Gamma_s} \subset \mathcal{D}.$$

- The sub-dictionary Ds is constructed so that

$$\text{if } g_{\tilde{\gamma}_m} \in \mathcal{D}_s \text{ maximizes } |\langle f, g_\gamma \rangle| \text{ in } \mathcal{D}_s$$

- then there exists $g_{\gamma_m} \in \mathcal{D}$ which minimizes (9.102) and whos γ_m is close to $\tilde{\gamma}_m$
- The index γ_m is found by a local search
- This is done in time-frequency dictionaries where a sub-dictionary can be sufficient to indicate a time-frequency region where an almost best match is located.

Translation invariance

- Decompositions in orthogonal bases *lack translation invariance* and are thus difficult to use for pattern recognition.
- Matching pursuits are translation invariant *if calculated in translation invariant dictionaries*
- A dictionary is translation invariant if for any

$$g_\gamma \in D \text{ and } n \in [0, N-1] \rightarrow g_\gamma[n-p] \in D$$

- Suppose that the matching decomposition of f in D is

$$f[n] = \sum_{m=0}^{M-1} \langle R^m f, g_{\gamma_m} \rangle \, g_{\gamma_m}[n] + R^M f[n].$$

Translation invariance

One can verify [151] that the matching pursuit of $f_p[n] = f[n-p]$ selects a translation by p of the same vectors g_{γ_m} with the same decomposition coefficients

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

Patterns can thus be characterized independently of their position. The same translation invariance property is valid for a basis pursuit. However, translation invariant dictionaries are necessarily very large, which often leads to prohibitive calculations. Wavelet packet and local cosine dictionaries are not translation invariant because at each scale 2^j the waveforms are translated only by $k2^j$ with $k \in \mathbb{Z}$.

Gabor dictionaries

- A time and frequency translation invariant Gabor dictionary is constructed by Qian and Chen **[287]** as well as Mallat and Zhong [259], by scaling, translating and modulating a Gaussian window.
- Gaussian windows are used because of their optimal time and frequency energy concentration, proved by the uncertainty theorem.

For each scale 2^j, a discrete window of period N is designed by sampling and periodizing a Gaussian $g(t) = 2^{1/4}\exp(-\pi t^2)$:

$$g_j[n] = K_j \sum_{p=-\infty}^{+\infty} g\left(\frac{n-pN}{2^j}\right).$$

Gabor dictionaries

The constant K_j is adjusted so that $\|g_j\| = 1$. This window is then translated in time and frequency. Let Γ be the set of indexes $\gamma = (p, k, 2^j)$ for $(p, k) \in [0, N-1]^2$ and $j \in [0, \log_2 N]$. A discrete Gabor atom is

$$g_\gamma[n] = g_j[n-p] \exp\left(\frac{i2\pi kn}{N}\right). \tag{9.105}$$

The resulting Gabor dictionary $\mathcal{D} = \{g_\gamma\}_{\gamma\in\Gamma}$ is time and frequency translation invariant modulo N. A matching pursuit decomposes real signals in this dictionary by grouping atoms g_{γ^+} and g_{γ^-} with $\gamma^\pm = (p, \pm k, 2^j)$. At each iteration, instead of projecting $R^m f$ over an atom g_γ, the matching pursuit computes its projection on the plane generated by $(g_{\gamma^+}, g_{\gamma^-})$. 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

$$g_\gamma^\phi[n] = K_{j,\phi}\, g_j[n-p] \cos\left(\frac{2\pi kn}{N} + \phi\right).$$

Gabor dictionaries

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

$$f = \sum_{m=0}^{+\infty} \langle R^m f, g_{\gamma_m}^{\phi_m} \rangle \, g_{\gamma_m}^{\phi_m}. \tag{9.106}$$

Orthogonal MP

- The approximations of a matching pursuit are improved by orthogonalizing the directions of projection, with a Gram-Schmidt procedure
- The resulting orthogonal pursuit converges with a finite number of iterations, which is not the case for a non-orthogonal pursuit.
- The price to be paid is the important computational cost of the Gram-Schmidt orthogonalization.

The vector g_{γ_m} selected by the matching algorithm is a priori not orthogonal to the previously selected vectors $\{g_{\gamma_p}\}_{0\leq p<m}$. When subtracting the projection of $R^m f$ over g_{γ_m} the algorithm reintroduces new components in the directions of $\{g_{\gamma_p}\}_{0\leq p<m}$. This is avoided by projecting the residues on an orthogonal family $\{u_p\}_{0\leq p<m}$ computed from $\{g_{\gamma_p}\}_{0\leq p<m}$.

Orthogonal MP

Let us initialize $u_0 = g_{\gamma_0}$. For $m \geq 0$, an orthogonal matching pursuit selects

g_{γ_m} that satisfies

$$|\langle R^m f, g_{\gamma_m} \rangle| \geq \alpha \sup_{\gamma \in \Gamma} |\langle R^m f, g_\gamma \rangle|. \tag{9.108}$$

The Gram-Schmidt algorithm orthogonalizes g_{γ_m} with respect to $\{g_{\gamma_p}\}_{0 \leq p < m}$ and defines

$$u_m = g_{\gamma_m} - \sum_{p=0}^{m-1} \frac{\langle g_{\gamma_m}, u_p \rangle}{\|u_p\|^2} u_p. \tag{9.109}$$

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

$$R^m f = \frac{\langle R^m f, u_m \rangle}{\|u_m\|^2} u_m + R^{m+1} f. \tag{9.110}$$

Orthogonal MP

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

$$f = \sum_{m=0}^{k-1} \frac{\langle R^m f, u_m \rangle}{\|u_m\|^2} u_m + R^k f \quad (9.111)$$
$$= P_{\mathbf{V}_k} f + R^k f,$$

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 $\{g_{\gamma_m}\}_{0 \leq m < k}$ 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 $m = k$ (9.109) implies that

$$\langle R^m f, u_m \rangle = \langle R^m f, g_{\gamma_m} \rangle. \quad (9.112)$$

Orthogonal MP

Since $\mathbf{V}_k$ has dimension k there exists $M \leq N$ such that $f \in \mathbf{V}_M$, so $R^M f = 0$ and inserting (9.112) in (9.111) for $k = M$ yields

$$f = \sum_{m=0}^{M-1} \frac{\langle R^m f, g_{\gamma_m} \rangle}{\|u_m\|^2} u_m. \tag{9.113}$$

The convergence is obtained with a finite number M of iterations. This is a decomposition in a family of orthogonal vectors so

$$\|f\|^2 = \sum_{m=0}^{M-1} \frac{|\langle R^m f, g_{\gamma_m} \rangle|^2}{\|u_m\|^2}. \tag{9.114}$$

To expand f over the original dictionary vectors $\{g_{\gamma_m}\}_{0 \leq m < M}$, we must perform a change of basis. The triangular Gram-Schmidt relations (9.109) are inverted to expand u_m in $\{g_{\gamma_p}\}_{0 \leq p \leq m}$:

$$u_m = \sum_{p=0}^{m} b[p, m]\, g_{\gamma_p}. \tag{9.115}$$

Orthogonal MP

Inserting this expression into (9.113) gives

$$f = \sum_{p=0}^{M-1} a[\gamma_p]\, g_{\gamma_p} \tag{9.116}$$

with

$$a[\gamma_p] = \sum_{m=p}^{M-1} b[p,m]\, \frac{\langle R^m f, g_{\gamma_m} \rangle}{\|u_m\|^2}.$$

Orthogonal MP

- During the first few iterations, the pursuit often selects nearly orthogonal vectors, so the Gram-Schmidt orthogonalization is not needed.
- The orthogonal and nonorthogonal pursuits are then nearly the same.
- 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 non-orthogonal pursuit.