Mathematical Foundations of Data Sciences

Gabriel Peyré
CNRS & DMA
École Normale Supérieure
gabriel.peyre@ens.fr
https://mathematical-tours.github.io
www.numerical-tours.com

March 19, 2020

Chapter 9

Sparse Regularization

Ref [17, 25, 23]

9.1 Sparsity Priors

9.1.1 Ideal sparsity prior.

As detailed in Chapter **??**, it is possible to use an orthogonal basis $\mathcal{B} = \{\psi_m\}_m$ to efficiently approximate an image f in a given class $f \in \Theta$ with a few atoms from $\mathcal{B}$.

To measure the complexity of an approximation with $\mathcal{B}$, we consider the ℓ^0 prior, which counts the number of non-zero coefficients in $\mathcal{B}$

$$J_0(f) \overset{\text{def.}}{=} \#\{m\,;\,\langle f,\,\psi_m\rangle \neq 0\} \quad \text{where} \quad x_m = \langle f,\,\psi_m\rangle.$$

One often also denote it as the ℓ^0 "pseudo-norm"

$$\|x\|_0 \overset{\text{def.}}{=} J_0(f).$$

which we treat here as an ideal sparsity measure for the coefficients x of f in $\mathcal{B}$.

Natural images are not exactly composed of a few atoms, but they can be well approximated by a function f_M with a small ideal sparsity $M = J_0(f)$. In particular, the best M-term approximation defined in (4.3) is defined by

$$f_M = \sum_{|\langle f,\,\psi_m\rangle| > T} \langle f,\,\psi_m\rangle \psi_m \quad \text{where} \quad M = \#\{m\,;\,|\langle f,\,\psi_m\rangle| > T\}\,.$$

As detailed in Section 4.2, discontinuous images with bounded variation have a fast decay of the approximation error $\|f - f_M\|$. Natural images f are well approximated by images with a small value of the ideal sparsity prior J_0.

Figure 9.1 shows an examples of decomposition of a natural image in a wavelet basis, $\psi_m = \psi^{\omega}_{j,n}$ $m = (j, n, \omega)$. This shows that most $\langle f,\,\psi_m\rangle$ are small, and hence the decomposition is quite sparse.

9.1.2 Convex relaxation

Unfortunately, the ideal sparsity prior J_0 is difficult to handle numerically because $J_0(f)$ is not a convex function of f. For instance, if f and g have non-intersecting supports of there coefficients in $\mathcal{B}$, then $J_0((f+g)/2) = J_0(f) + J_0(g)$, which shows the highly non-convex behavior of J_0.

This ideal sparsity J_0 is thus not amenable to minimization, which is an issue to solve general inverse problems considered in Section **??**.

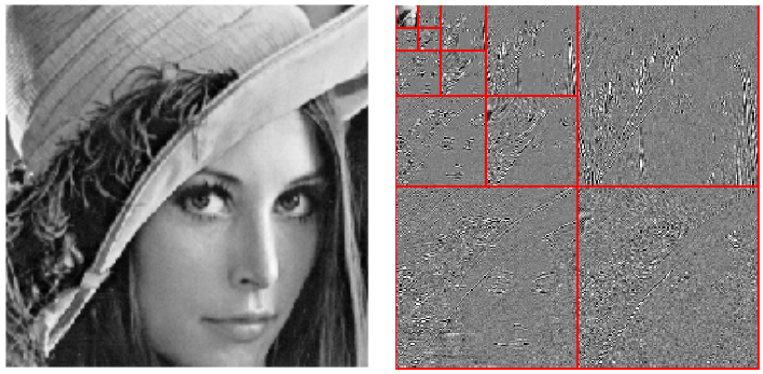

Image f Coefficients $\langle f, \psi_m \rangle$

Figure 9.1: Wavelet coefficients of natural images are relatively sparse.

We consider a family of ℓ^q priors for $q > 0$, intended to approximate the ideal prior J_0

$$J_q(f) = \sum_m |\langle f, \psi_m \rangle|^q.$$

As shown in Figure 9.2, the unit balls in $\mathbb{R}^2$ associated to these priors are shrinking toward the axes, which corresponds to the unit ball for the ℓ^0 pseudo norm. In some sense, the J_q priors are becoming closer to J_0 as q tends to zero, and thus J_q favors sparsity for small q.

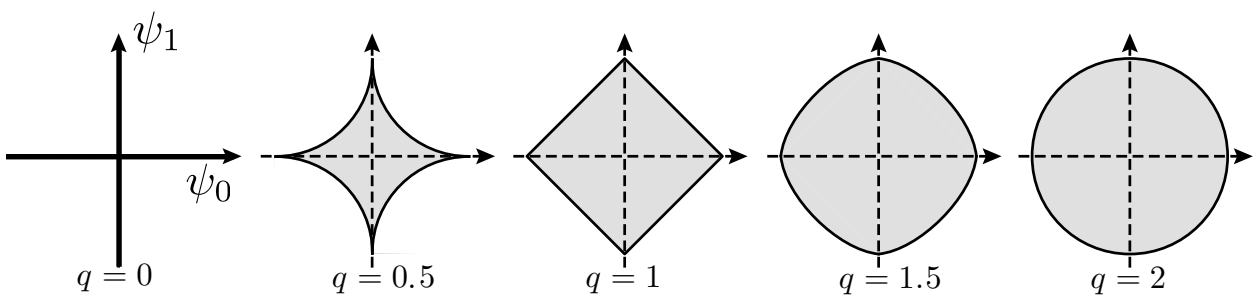

Figure 9.2: ℓ^q balls $\{x \,;\, J_q(x) \leqslant 1\}$ for varying q.

The prior J_q is convex if and only if $q \geqslant 1$. To reach the highest degree of sparsity while using a convex prior, we consider the ℓ^1 sparsity prior J_1, which is thus defined as

$$J_1(f) = \|(\langle f, \psi_m \rangle)\|_1 = \sum_m |\langle f, \psi_m \rangle|. \tag{9.1}$$

In the following, we consider discrete orthogonal bases $\mathcal{B} = \{\psi_m\}_{m=0}^{N-1}$ of $\mathbb{R}^N$.

9.1.3 Sparse Regularization and Thresholding

Given some orthogonal basis $\{\psi_m\}_m$ of $\mathbb{R}^N$, the denoising by regularization (7.15) is written using the sparsity J_0 and J_1 as

$$f^\star = \operatorname*{argmin}_{g \in \mathbb{R}^N} \frac{1}{2}\|f - g\|^2 + \lambda J_q(f)$$

for $q = 0$ or $q = 1$. It can be re-written in the orthogonal basis as

$$f^\star = \sum_m x_m^\star \psi_m$$

$$\text{where} \quad x^\star = \operatorname*{argmin}_{y \in \mathbb{R}^N} \sum_m \frac{1}{2}|x_m - y_m|^2 + \lambda |y_m|^q$$

where $x_m \overset{\text{def.}}{=} \langle f, \psi_m \rangle$, $y_m \overset{\text{def.}}{=} \langle g, \psi_m \rangle$, and where we use the following slight abuse of notation for $q = 0$

$$\forall\, u \in \mathbb{R}, \quad |u|^0 = \begin{cases} 0 & \text{if} \quad u = 0, \\ 1 & \text{otherwise.} \end{cases}$$

Each coefficients of the denoised image is the solution of a 1-D optimization problem

$$x_m^\star = \operatorname*{argmin}_{u \in \mathbb{R}} \frac{1}{2}|x_m - u|^2 + \lambda |u|^q \tag{9.2}$$

and the following proposition this optimization is solved exactly in closed form using thresholding.

Proposition 26. *One has*

$$x_m^\star = S_T^q(x_m) \quad \textit{where} \quad \begin{cases} T = \sqrt{2\lambda} \quad \textit{for} \quad q = 0, \\ T = \lambda \quad \textit{for} \quad q = 1, \end{cases} \tag{9.3}$$

where

$$\forall\, u \in \mathbb{R}, \quad S_T^0(u) \overset{\text{def.}}{=} \begin{cases} 0 & \textit{if} \quad |u| < T, \\ u & \textit{otherwise} \end{cases} \tag{9.4}$$

is the hard thresholding introduced in (6.6)*, and*

$$\forall\, u \in \mathbb{R}, \quad S_T^1(u) \overset{\text{def.}}{=} \operatorname{sign}(u)(|u| - T)_+ \tag{9.5}$$

is the soft thresholding introduced in (6.7)*.*

Proof. One needs to solve (9.2). Figure 9.3, left shows the function $\|x - y\|^2 + T^2 \|x\|_0$, and the minimum is clearly at $x = 0$ when $T \geqslant y$, and at $x = y$ otherwise. This is thus a hard thresholding with threshold $T^2 = 2\lambda$. Figure (9.3), right, shows the evolution with λ of the function $\frac{1}{2}\|x - y\|^2 + \lambda |x|$. For $x > 0$, one has $F'(x) = x - y + \lambda$ wich is 0 at $x = y - \lambda$. The minimum is at $x = y - \lambda$ for $\lambda \leqslant y$, and stays at 0 for all $\lambda > y$. □

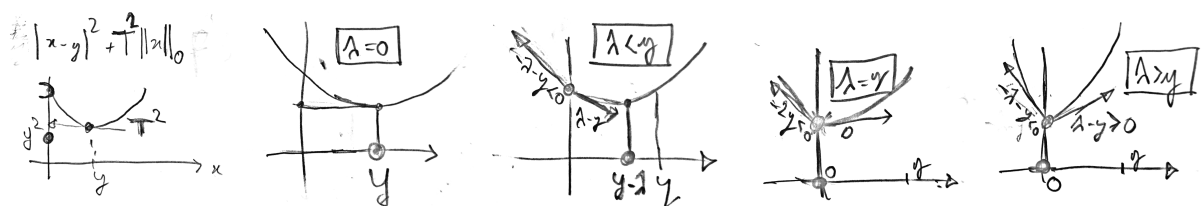

Figure 9.3: Leftmost: function $\| \cdot - y\|^2 + T^2 \| \cdot \|_0$. Others: evolution with λ of the function $F(x) \overset{\text{def.}}{=} \frac{1}{2}\| \cdot - y\|^2 + \lambda | \cdot |$.

One thus has

$$f_{\lambda,q} = \sum_m S_T^q(\langle f, \psi_m \rangle) \psi_m.$$

As detailed in Section 6.3, these denoising methods has the advantage that the threshold is simple to set for Gaussian white noise w of variance σ^2. Theoritical values indicated that $T = \sqrt{2\log(N)}\sigma$ is asymptotically optimal, see Section 6.3.3. In practice, one should choose $T \approx 3\sigma$ for hard thresholding (ℓ^0 regularization), and $T \approx 3\sigma/2$ for soft thresholding (ℓ^1 regularization), see Figure 6.14.

9.2 Sparse Regularization of Inverse Problems

Sparse ℓ^1 regularization in an orthogonal basis $\{\psi_m\}_m$ of $\mathbb{R}^N$ makes use of the J_1 prior defined in (9.1), so that the inversion is obtained by solving the following convex program

$$f_\lambda \in \operatorname*{argmin}_{f \in \mathbb{R}^N} \frac{1}{2}\|y - \Phi f\|^2 + \lambda \sum_m |\langle f, \psi_m \rangle|. \tag{9.6}$$

This corresponds to the basis pursuit denoising for sparse approximation introduced by Chen, Donoho and Saunders in [9]. The resolution of (9.6) can be perform using an iterative thresholding algorithm as detailed in Section 9.3.

Analysis vs. synthesis priors. When the set of atoms $\Psi = \{\psi_m\}_{m=1}^Q$ is non-orthognal (and might even be redundant in the case $Q > N$), there is two distincts way to generalizes problem (9.6), which we formulate as in (**??**), by introducing a generic convex prior J

$$f_\lambda \in \operatorname*{argmin}_{f \in \mathbb{R}^N} \frac{1}{2}\|y - \Phi f\|^2 + \lambda J(f). \tag{9.7}$$

In the following, with a slight abuse of notation, we denote the "analysis" and "synthesis" operator as

$$\Psi : x \in \mathbb{R}^Q \mapsto \Psi x = \sum_m x_m \psi_m \quad \text{and} \quad \Psi^* : f \in \mathbb{R}^N \mapsto (\langle f, \psi_m \rangle)_{m=1}^Q \in \mathbb{R}^Q.$$

The so-called analysis-type prior is simply measuring the sparsity of the correlations with the atoms in the dictionary

$$J_1^{\mathrm{A}}(f) \overset{\text{def.}}{=} \sum_m |\langle f, \psi_m \rangle| = \|\Psi^* f\|_1. \tag{9.8}$$

The so-called synthesis-type priori in contrast measure the sparsity of the sparsest expansion of f in Ψ, i.e.

$$J_1^{\mathrm{S}}(f) \overset{\text{def.}}{=} \min_{x \in \mathbb{R}^q, \Psi x = f} \|x\|_1. \tag{9.9}$$

While the analysis regularization (9.8) seems simpler to handle, it is actually the contrary. Solving (9.7) with $J = J_1^{\mathrm{A}}$ is in fact quite involved, and necessitate typically primal-dual algorithm as detailed in Chapter 12. Furthermore, the theoretical study of the performance of the resulting regularization method is mostly an open problem.

We thus now focus on the synthesis regularization problem $J = J_1^{\mathrm{S}}$, and we re-write (9.7) conveniently as $f_\lambda = \Psi x_\lambda$ where x_λ is any solution of the following Basis Pursuit Denoising problem

$$x_\lambda \in \operatorname*{argmin}_{x \in \mathbb{R}^Q} \frac{1}{2\lambda}\|y - Ax\|^2 + \|x\|_1 \tag{9.10}$$

where we introduced the following matrix

$$A \overset{\text{def.}}{=} \Phi\Psi \in \mathbb{R}^{P \times Q}.$$

As $\lambda \to 0$, we consider the following limit constrained problem

$$x^\star = \operatorname*{argmin}_{Ax = y} \|x\|_1 \tag{9.11}$$

and the signal is recovered as $f^\star = \Psi x^\star \in \mathbb{R}^N$.

9.3 Iterative Soft Thresholding Algorithm

This section details an iterative algorithm that compute a solution of (9.10).

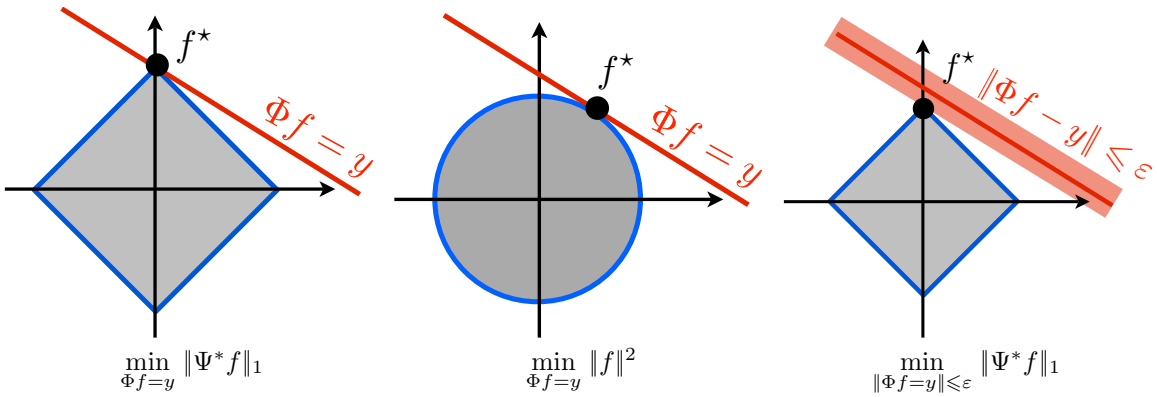

Figure 9.4: Geometry of convex optimizations.

9.3.1 Noiseless Recovery as a Linear Program

Before detailing this methods, which only deal with the case $\lambda > 0$, let us note that in the noiseless settig, $\lambda = 0$ and (9.11) is actually equivalent to a linear program. Indeed, decomposing $a = x_+ - x_-$ with $(x_+, x_-) \in (\mathbb{R}_+^Q)^2$, one has

$$x^\star = \underset{(x_+,x_-)\in(\mathbb{R}_+^Q)^2}{\operatorname{argmin}} \left\{ \langle x_+, \mathbb{1}_Q \rangle + \langle x_-, \mathbb{1}_Q \rangle \;;\; y = A(x_+ - x_-) \right\}. \tag{9.12}$$

which is a linear program. For small to medium scale problem (Q of the order of a few thousands) it can be solved using the simplex algorithm or interior point methods. For large scale problems such as those encountered in imaging or machine learning, this is not possible, and one has to ressort to simpler first order schemes. A possible option is the Douglas-Rachford splitting scheme, which is detailed in Section **??**. Let us however stress that the constrained problem (9.11), because of its polyhedral (linear) nature, is in fact harder to solve than the penalized problem (9.10) that we now target.

9.3.2 Projected Gradient Descent for ℓ^1.

As a first practical example to solve (9.10), we will show how to use the projected gradient descent method, which is analyzed in detailed in Section 12.1.3. Similarly to (9.12), we remap (9.10) as the resolution of a constraint minimization problem of the form (12.4) where here $\mathcal{C}$ is a positivity constraint and

$$u = (u_+, u_-) \in (\mathbb{R}^Q)^2, \quad \mathcal{C} = (\mathbb{R}_+^Q)^2, \quad \text{and} \quad \mathcal{E}(u) = \frac{1}{2}\|\Phi(u_+ - u_-) - y\|^2 + \lambda\langle u_+, \mathbb{1}_Q\rangle + \lambda\langle u_-, \mathbb{1}_Q\rangle.$$

The projection on $\mathcal{C}$ is here simple to compute

$$\operatorname{Proj}_{(\mathbb{R}_+^Q)^2}(u_+, u_-) = ((u_+)_\oplus, (u_-)_\oplus) \quad \text{where} \quad (r)_\oplus \overset{\text{def.}}{=} \max(r, 0),$$

and the gradient reads

$$\nabla\mathcal{E}(u_+, u_-) = (\eta + \lambda\mathbb{1}_Q, -\eta + \lambda\mathbb{1}_Q) \quad \text{where} \quad \eta = \Phi^*(\Phi(u_+ - u_-) - y)$$

Denoting $u^{(\ell)} = (u_+^{(\ell)}, u_-^{(\ell)})$ and $x^{(\ell)} \overset{\text{def.}}{=} u_+^{(\ell)} - u_-^{(\ell)}$, the iterate of the projected gradient descent algorithm (12.5) read

$$u_+^{(\ell+1)} \overset{\text{def.}}{=} \left(u_+^{(\ell)} - \tau_\ell(\eta^{(\ell)} + \lambda)\right)_\oplus \quad \text{and} \quad u_-^{(\ell+1)} \overset{\text{def.}}{=} \left(u_-^{(\ell)} - \tau_\ell(-\eta^{(\ell)} + \lambda)\right)_\oplus$$

where $\eta^{(\ell)} \overset{\text{def.}}{=} \Phi^*(\Phi x^{(\ell)} - y)$.

Theorem 17 ensures that $u^{(\ell)} \to u^\star$ a solution to (12.4) if

$$\forall \ell, \quad 0 < \tau_{\min} < \tau_\ell < \tau_{\max} < \frac{2}{\|\Phi\|^2},$$

and thus $x^{(\ell)} \to x^\star = u_+^\star - u_-^\star$ which is thus a solution to (9.10).

9.3.3 Iterative Soft Thresholding and Forward Backward

A drawback of this projected gradient descent scheme is that it necessitate to store $2Q$ coefficients. A closely related method, which comes with exactly the same convergence guarantees and rate, is the so called "iterative soft thresholding algorithm" (ISTA). This algorithm was derived by several authors, among which [15, 12], and belongs to the general family of forward-backward splitting in proximal iterations [11], which we detail in Section 12.3.2.

For the sake of simplicity, let us derive this algorithm in the special case of ℓ^1 by surrogate function minimization. We aim at minimizing (9.6)

$$\mathcal{E}(x) \overset{\text{def.}}{=} \frac{1}{2}\|y - Ax\|^2 + \lambda\|x\|_1$$

and we introduce for any fixed x' the function

$$\mathcal{E}_\tau(x, x') \overset{\text{def.}}{=} \mathcal{E}(x) - \frac{1}{2}\|Ax - Ax'\|^2 + \frac{1}{2\tau}\|x - x'\|.$$

We notice that $\mathcal{E}(x, x) = 0$ and one has

$$K(x, x') \overset{\text{def.}}{=} -\frac{1}{2}\|Ax - Ax'\|^2 + \frac{1}{2\tau}\|x - x'\| = \frac{1}{2}\langle \left(\frac{1}{\tau}\mathrm{Id}_N - A^*A\right)(x - x'),\, x - x'\rangle.$$

This quantity $K(x, x')$ is positive if $\lambda_{\max}(A^*A) \leqslant 1/\tau$ (maximum eigenvalue), i.e. $\tau \leqslant 1/\|A\|_{\mathrm{op}}^2$, where we recall that $\|A\|_{\mathrm{op}} = \sigma_{\max}(A)$ is the operator (algebra) norm. This shows that $\mathcal{E}_\tau(x, x')$ is a valid surrogate functional, in the sense that

$$\mathcal{E}(x) \leqslant \mathcal{E}_\tau(x, x'), \quad \mathcal{E}_\tau(x, x') = 0, \quad \text{and} \quad \mathcal{E}(\cdot) - \mathcal{E}_\tau(\cdot, x') \text{ is smooth.}$$

This leads to define

$$x^{(\ell+1)} \overset{\text{def.}}{=} \operatorname*{argmin}_x \mathcal{E}_{\tau_\ell}(x, x^{(\ell)}) \tag{9.13}$$

which by construction satisfies

$$\mathcal{E}(x^{(\ell+1)}) \leqslant \mathcal{E}(x^{(\ell)}).$$

Proposition 27. *The iterates $x^{(\ell)}$ defined by (9.13) satisfy*

$$x^{(\ell+1)} = \mathcal{S}^1_{\lambda\tau_\ell}\left(x^{(\ell)} - \tau_\ell A^*(Ax^{(\ell)} - y)\right) \tag{9.14}$$

where $\mathcal{S}^1_\lambda(x) = (S^1_\lambda(x_m))_m$ where $S^1_\lambda(r) = \mathrm{sign}(r)(|r| - \lambda)_\oplus$ is the soft thresholding operator defined in (9.5).

Proof. One has

$$\begin{aligned}
\mathcal{E}_\tau(x, x') &= \frac{1}{2}\|Ax - y\|^2 - \frac{1}{2}\|Ax - Ax'\|^2 + \frac{1}{2\tau}\|x - x'\| + \lambda\|x\|_1 \\
&= C + \frac{1}{2}\|Ax\|^2 - \frac{1}{2}\|Ax\|^2 + \frac{1}{2\tau}\|x\|^2 - \langle Ax,\, y\rangle + \langle Ax,\, Ax'\rangle - \frac{1}{\tau}\langle x,\, x'\rangle + \lambda\|x\|_1 \\
&= C + \frac{1}{2\tau}\|x\|^2 + \langle x,\, -A^*y + AA^*x' - \frac{1}{\tau}x'\rangle + \lambda\|x\|_1 \\
&= C' + \frac{1}{\tau}\ \left(\|x - (x' - \tau A^*(Ax' - y))\,\|^2 + \tau\lambda\|x\|_1.\right)
\end{aligned}$$

Proposition (26) shows that the minimizer of $\mathcal{E}_\tau(x, x')$ is thus indeed $\mathcal{S}^1_{\lambda\tau}(x' - \tau_\ell A^*(Ax' - y))$. □

Of course, these iterations (9.14) are the same as the FB iterates (12.15), when, for the special case (9.6), one can consider a splitting of the form (12.15) defining

$$\mathcal{F} = \frac{1}{2}\|A \cdot -y\|^2 \quad \text{and} \quad \mathcal{G} = \lambda \|\cdot\|_1. \tag{9.15}$$

In the case (9.15), Proposition (26) shows that $\mathrm{Prox}_{\rho J}$ is the soft thresholding.

9.4 Example: Sparse Deconvolution

9.4.1 Sparse Spikes Deconvolution

Sparse spikes deconvolution makes use of sparsity in the spacial domain, which corresponds to the orthogonal basis of Diracs $\psi_m[n] = \delta[n-m]$. This sparsity was first introduced in the seismic imaging community [], where the signal f_0 represent the change of density in the underground and is assumed to be composed of a few Diracs impulse.

In a simplified linearized 1D set-up, ignoring multiple reflexions, the acquisition of underground data f_0 is modeled as a convolution $y = h \star f_0 + w$, where h is a so-called "wavelet" signal sent in the ground. This should not be confounded with the construction of orthogonal wavelet bases detailed in Chapter **??**, although the term "wavelet" originally comes from seismic imaging.

The wavelet filter h is typically a band pass signal that perform a tradeoff between space and frequency concentration especially tailored for seismic exploration. Figure (9.5) shows a typical wavelet that is a second derivative of a Gaussian, together with its Fourier transform. This shows the large amount of information removed from f during the imaging process.

The sparse ℓ^1 regularization in the Dirac basis reads

$$f^\star = \underset{f \in \mathbb{R}^N}{\mathrm{argmin}} \frac{1}{2}\|f \star h - y\|^2 + \lambda \sum_m |f_m|.$$

Figure 9.5 shows the result of ℓ^1 minimization for a well chosen λ parameter, that was optimized in an oracle manner to minimize the error $\|f^\star - f_0\|$.

The iterative soft thresholding for sparse spikes inversion iterates

$$\tilde{f}^{(k)} = f^{(k)} - \tau h \star (h \star f^{(k)} - y)$$

and

$$f_m^{(k+1)} = S^1_{\lambda\tau}(\tilde{f}_m^{(k)})$$

where the step size should obeys

$$\tau < 2/\|\Phi^*\Phi\| = 2/\max_\omega |\hat{h}(\omega)|^2$$

to guarantee convergence. Figure 9.6 shows the progressive convergence of the algorithm, both in term of energy minimization and iterates. Since the energy is not strictly convex, we note that convergence in energy is not enough to guarantee convergence of the algorithm.

9.4.2 Sparse Wavelets Deconvolution

Signal and image acquired by camera always contain some amount of blur because of objects being out of focus, movements in the scene during exposure, and diffraction. A simplifying assumption assumes a spatially invariant blur, so that Φ is a convolution

$$y = f_0 \star h + w.$$

In the following, we consider h to be a Gaussian filter of width $\mu > 0$. The number of effective measurements can thus be considered to be $P \sim 1/\mu$, since Φ nearly set to 0 large enough Fourier frequencies. Table **??** details the implementation of the sparse deconvolution algorithm.

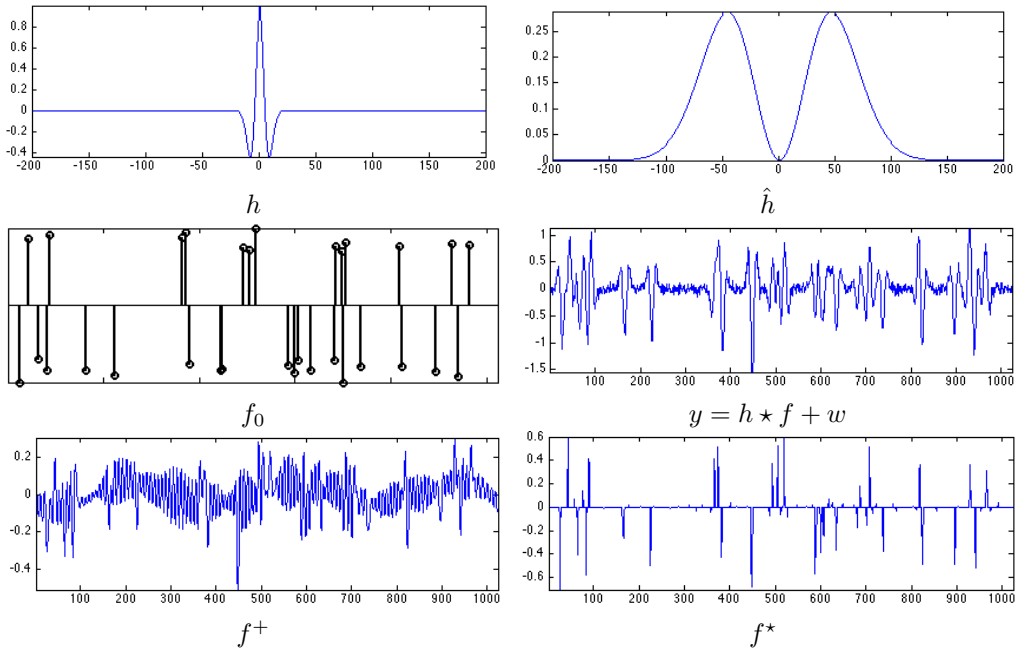

Figure 9.5: Pseudo-inverse and ℓ^1 sparse spikes deconvolution.

Figures 9.7 and 9.8 shows examples of signal and image acquisition with Gaussian blur.

Sobolev regularization (7.17) improves over ℓ^2 regularization (**??**) because it introduces an uniform smoothing that reduces noise artifact. It however fail to recover sharp edge and thus does a poor job in inverting the operator. To recover sharper transition and edges, one can use either a TV regularization or a sparsity in an orthogonal wavelet basis.

Figure 9.7 shows the improvement obtained in 1D with wavelets with respect to Sobolev. Figure 9.8 shows that this improvement is also visible for image deblurring. To obtain a better result with fewer artifact, one can replace the soft thresholding in orthogonal wavelets in during the iteration (**??**) by a thresholding in a translation invariant tight frame as defined in (6.10).

Figure 9.9 shows the decay of the SNR as a function of the regularization parameter λ. This SNR is computed in an oracle manner since it requires the knowledge of f_0. The optimal value of λ was used in the reported experiments.

9.4.3 Sparse Inpainting

This section is a follow-up of Section 8.5.2.

To inpaint using a sparsity prior without noise, we use a small value for λ. The iterative thresholding algorithm (**??**) is written as follow for $\tau = 1$,

$$f^{(k+1)} = \sum_m S_\lambda^1(\langle P_y(f^{(k)}),\, \psi_m \rangle)\psi_m$$

Figure 9.10 shows the improvevement obtained by the sparse prior over the Sobolev prior if one uses soft thresholding in a translation invariant wavelet frame.

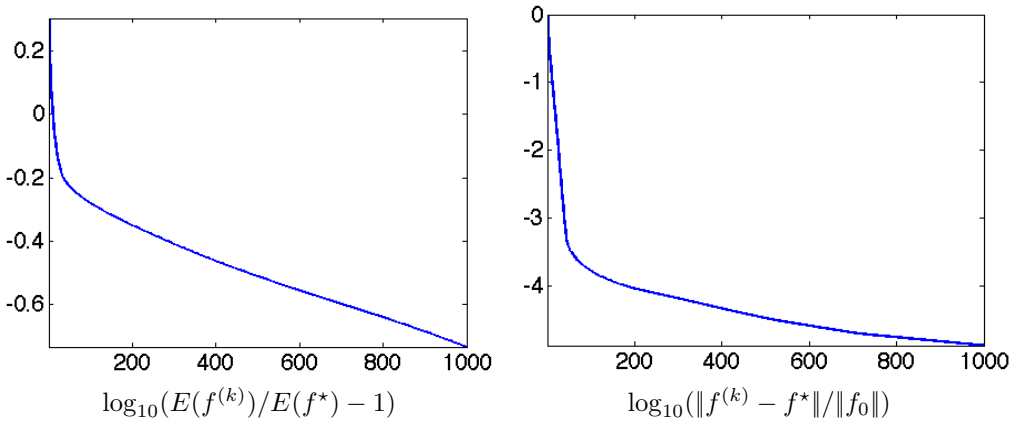

Figure 9.6: Decay of the energy and convergence through the iterative thresholding iterations.

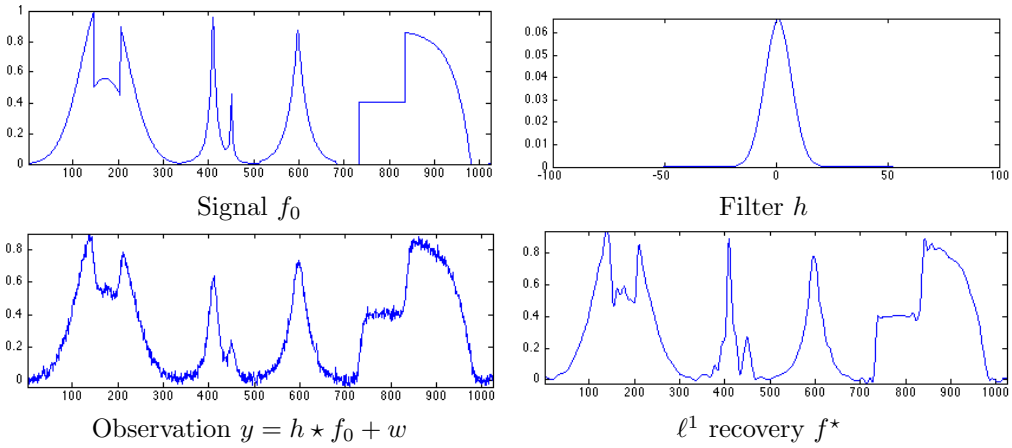

Figure 9.7: Sparse 1D deconvolution using orthogonal wavelets.

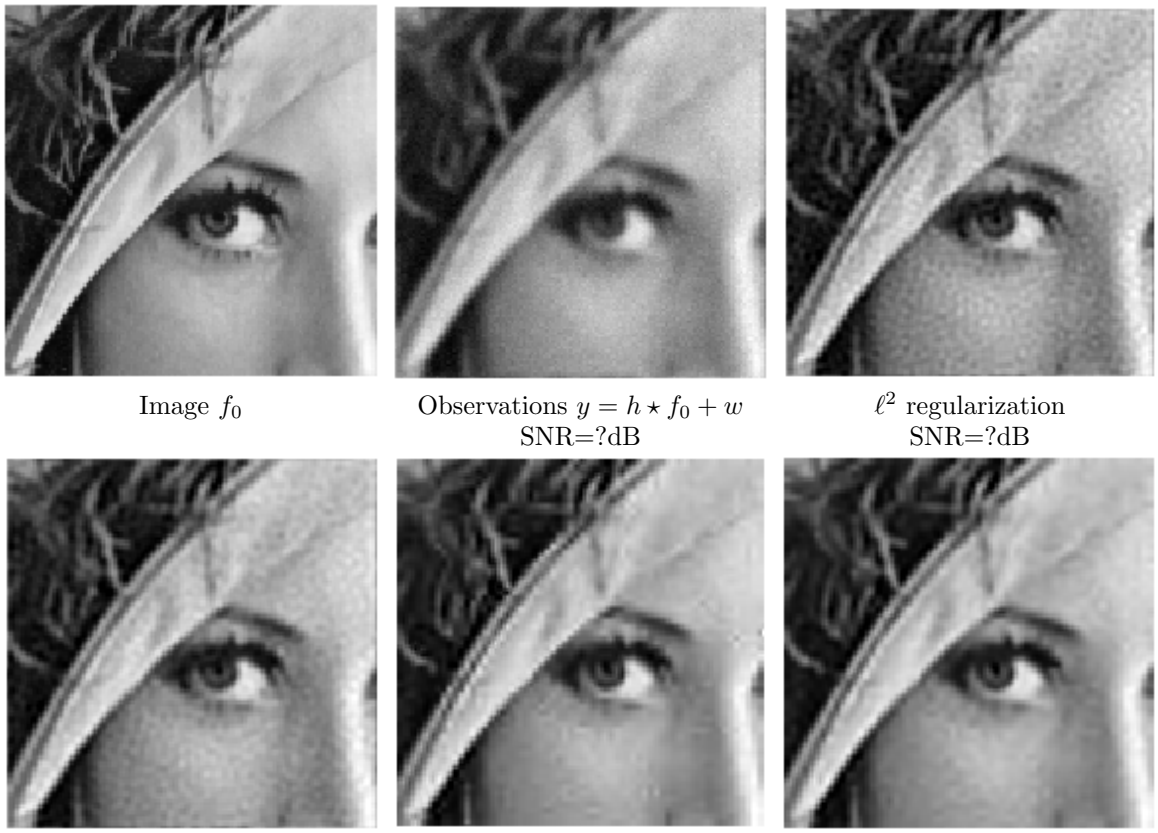

Image f_0 | Observations $y = h \star f_0 + w$ SNR=?dB | ℓ^2 regularization SNR=?dB

Sobolev regularization SNR=?dB | ℓ^1 regularization SNR=?dB | ℓ^1 invariant SNR=?dB

Figure 9.8: Image deconvolution.

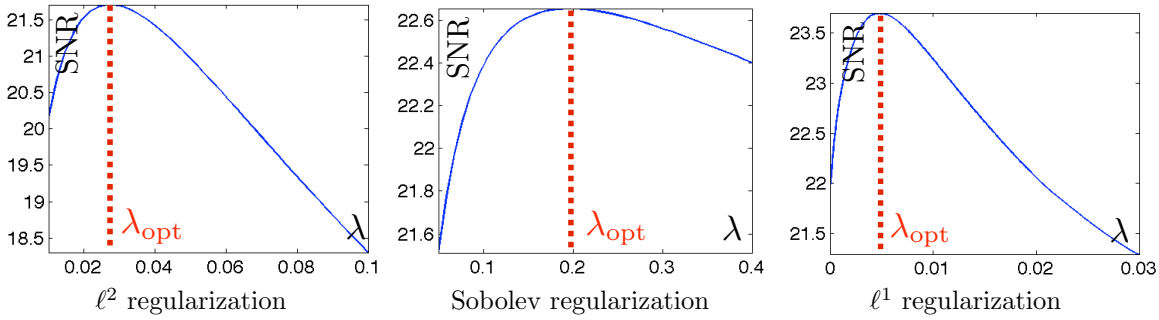

Figure 9.9: SNR as a function of λ.

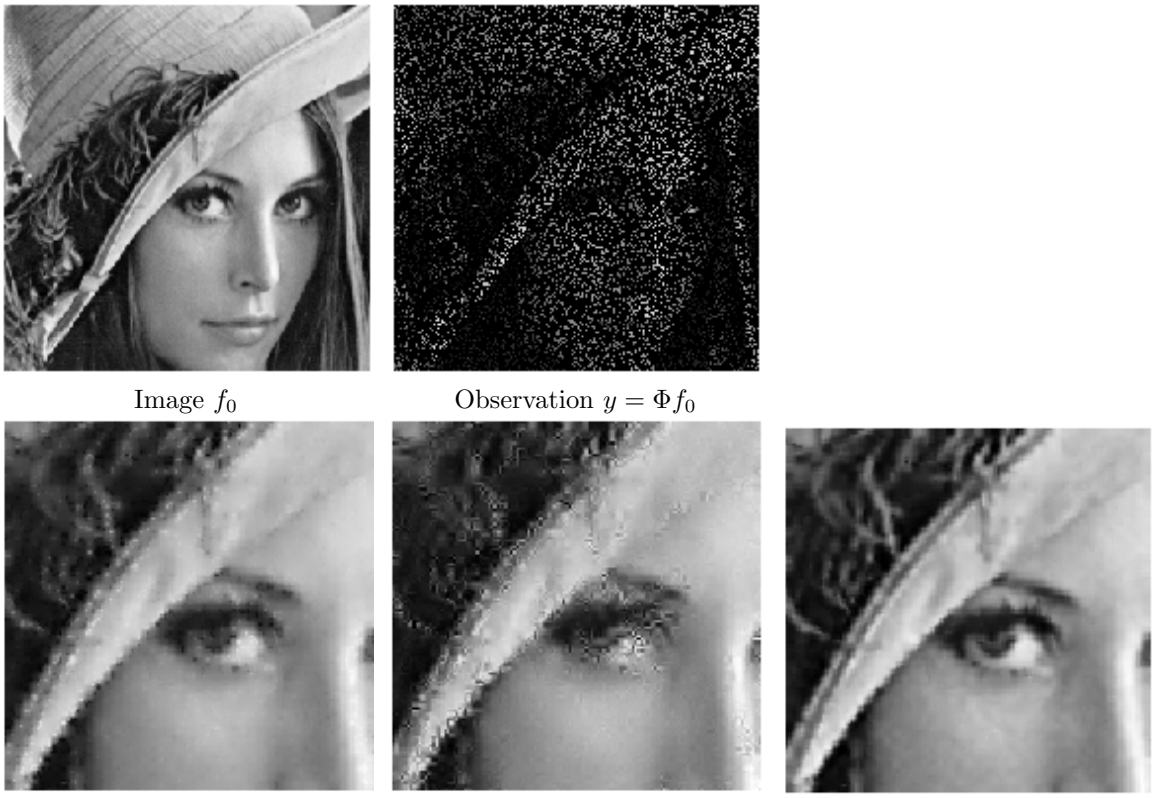

Image f_0 | Observation $y = \Phi f_0$

Sobolev $f^\star$ SNR=?dB | Ortho. wav $f^\star$ SNR=?dB | TI. wav $f^\star$ SNR=?dB

Figure 9.10: Inpainting with Sobolev and sparsity.

Bibliography

[1] Amir Beck. *Introduction to Nonlinear Optimization: Theory, Algorithms, and Applications with MATLAB*. SIAM, 2014.

[2] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[3] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[4] E. Candès and D. Donoho. New tight frames of curvelets and optimal representations of objects with piecewise C^2 singularities. *Commun. on Pure and Appl. Math.*, 57(2):219–266, 2004.

[5] E. J. Candès, L. Demanet, D. L. Donoho, and L. Ying. Fast discrete curvelet transforms. *SIAM Multiscale Modeling and Simulation*, 5:861–899, 2005.

[6] A. Chambolle. An algorithm for total variation minimization and applications. *J. Math. Imaging Vis.*, 20:89–97, 2004.

[7] Antonin Chambolle, Vicent Caselles, Daniel Cremers, Matteo Novaga, and Thomas Pock. An introduction to total variation for image analysis. *Theoretical foundations and numerical methods for sparse recovery*, 9(263-340):227, 2010.

[8] Antonin Chambolle and Thomas Pock. An introduction to continuous optimization for imaging. *Acta Numerica*, 25:161–319, 2016.

[9] S.S. Chen, D.L. Donoho, and M.A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1999.

[10] Philippe G Ciarlet. Introduction à l'analyse numérique matricielle et à l'optimisation. 1982.

[11] P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *SIAM Multiscale Modeling and Simulation*, 4(4), 2005.

[12] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Commun. on Pure and Appl. Math.*, 57:1413–1541, 2004.

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

[14] Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.

[15] M. Figueiredo and R. Nowak. An EM Algorithm for Wavelet-Based Image Restoration. *IEEE Trans. Image Proc.*, 12(8):906–916, 2003.

[16] Simon Foucart and Holger Rauhut. *A mathematical introduction to compressive sensing*, volume 1. Birkhäuser Basel, 2013.

[17] Stephane Mallat. *A wavelet tour of signal processing: the sparse way.* Academic press, 2008.

[18] D. Mumford and J. Shah. Optimal approximation by piecewise smooth functions and associated variational problems. *Commun. on Pure and Appl. Math.*, 42:577–685, 1989.

[19] Neal Parikh, Stephen Boyd, et al. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

[20] Gabriel Peyré. *L'algèbre discrète de la transformée de Fourier.* Ellipses, 2004.

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

[22] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Phys. D*, 60(1-4):259–268, 1992.

[23] Otmar Scherzer, Markus Grasmair, Harald Grossauer, Markus Haltmeier, Frank Lenzen, and L Sirovich. *Variational methods in imaging.* Springer, 2009.

[24] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[25] Jean-Luc Starck, Fionn Murtagh, and Jalal Fadili. *Sparse image and signal processing: Wavelets and related geometric multiscale analysis.* Cambridge university press, 2015.