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

November 18, 2020

Chapter 10

Theory of Sparse Regularization

We now apply the basics elements of convex analysis from the previous chapter to perform a theoretical analysis of the properties of the Lasso, in particular its performances to recover sparse vectors.

10.1 Existence and Uniqueness

10.1.1 Existence

We consider problems (9.10) and (9.11), that we rewrite here as

$$\min_{x\in\mathbb{R}^N} f_\lambda(x) \overset{\text{def.}}{=} \frac{1}{2\lambda}\|y-Ax\|^2+\lambda\|x\|_1 \tag{$\mathcal{P}_\lambda(y)$}$$

and its limit as $\lambda\to 0$

$$\min_{Ax=y} \|x\|_1 = \min_x f_0(x) \overset{\text{def.}}{=} \iota_{\mathcal{L}_y}(x)+\|x\|_1. \tag{$\mathcal{P}_0(y)$}$$

where $A\in\mathbb{R}^{P\times N}$, and $\mathcal{L}_y \overset{\text{def.}}{=} \{x\in\mathbb{R}^N \;;\; Ax=y\}$.

We recall that the setup is that one observe noise measures

$$y = Ax_0 + w$$

and we would like conditions to ensure for x_0 to solution to $(\mathcal{P}_0(Ax_0))$ (i.e. when $w=0$) and to be close (in some sense to be defined, and in some proportion to the noise level $\|w\|$) to the solutions of $(\mathcal{P}_0(y=Ax_0+w))$ when λ is wisely chosen as a function of $\|w\|$.

First let us note that since $(\mathcal{P}_\lambda(y))$ is unconstrained and coercive (because $\|\cdot\|_1$ is), this problem always has solutions. Since A might have a kernel and $\|\cdot\|_1$ is not strongly convex, it might have non-unique solutions. If $y\in\mathrm{Im}(A)$, the constraint set of $(\mathcal{P}_0(y))$ is non-empty, and it also has solutions, which might fail to be unique.

Figure 10.1 gives the intuition of the theory that will be developed in this chapter, regarding the exact or approximated recovery of sparse vectors x_0, and the need for a careful selection of the λ parameter.

10.1.2 Polytope Projection for the Constraint Problem

The following proposition gives a geometric description of those vectors which are recovered by ℓ^1 minimization when there is no noise.

Proposition 28. *We denote* $B \overset{\text{def.}}{=} \{x\in\mathbb{R}^N \;;\; \|x\|_1\leqslant 1\}$. *Then, assuming* $\mathrm{Im}(A)=\mathbb{R}^P$,

$$x_0 \text{ is a solution to } \mathcal{P}_0(Ax_0) \quad\Longleftrightarrow\quad A\frac{x_0}{\|x_0\|_1}\in\partial(AB) \tag{10.1}$$

where "∂" denoted the boundary and $AB=\{Ax \;;\; x\in B\}$.

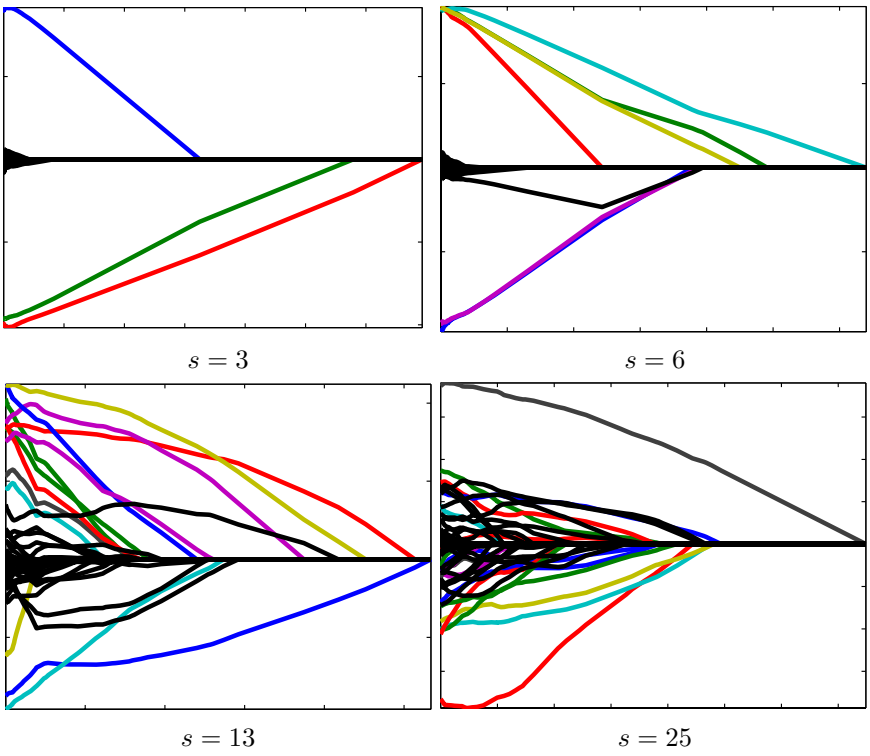

Figure 10.1: Display of the evolution $\lambda \mapsto x_\lambda$ of the solutions of $(\mathcal{P}_\lambda(y))$.

Proof. We first prove " $\Rightarrow$". We suppose that x_0 is not a solution, and aim at showing that $A\frac{x_0}{\|x_0\|_1} \in \mathrm{int}(AB_\rho)$. Since it is not a solution, there exists z such that $Ax_0 = Az$ and $\|z\|_1 = (1-\delta)\|x_0\|_1$ with $\delta > 0$. Then for any displacement $h = A\varepsilon \in \mathrm{Im}(A)$, where one can impose $\varepsilon \in \ker(A)^\perp$, i.e. $\varepsilon = A^+h$, one has $Ax_0 + h = A(z+\varepsilon)$ and

$$\|z+\varepsilon\|_1 \leqslant \|z\|_1 + \|\Phi^+ h\| \leqslant (1-\delta)\|x_0\|_1 + \|\Phi^+\|_{1,1}\|h\|_1 < \frac{\delta}{\|A^+\|_{1,1}}\|x_0\|.$$

This means that choosing $\|h\|_1 < \frac{\delta}{\|A^+\|_{1,1}}\|x_0\|_1$ implies that $A\frac{x_0}{\|x_0\|_1} \in \mathrm{int}(AB)$.

We now prove "$\Leftarrow$". We suppose that $A\frac{x_0}{\|x_0\|_1} \in \mathrm{int}(AB)$. Then there exists z such that $Ax_0 = (1-\delta)Az$ and $\|z\|_1 < \|x_0\|_1$. This implies $\|(1-\delta)z\|_1 < \|x_0\|_1$ so that $(1-\delta)z$ is better than x_0 which is thus not a solution. □

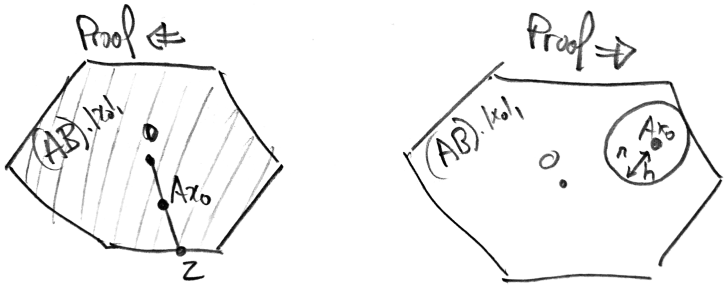

Figure 10.2: Graphical display of the proof for the polytope analysis of ℓ^1 exact recovery.

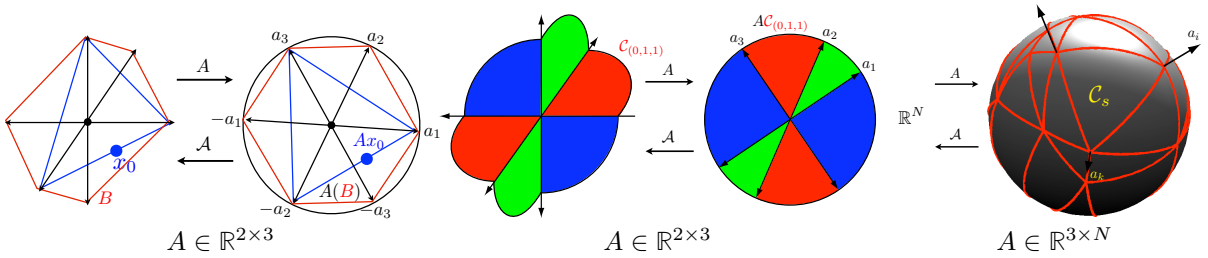

Figure 10.3: Display of the action of the linear map A on the ℓ^1 ball B, and of the inverse non-linear map $\mathcal{A}$ defined by the solution of $(\mathcal{P}_0(y))$.

This results state that "friendly" identifiable vectors (those recovered by ℓ^1) are those who gets projected by A on the boundary of the polytope $\|x_0\|_1 AB$. Intuitively, if P is small in comparison to N, then this projected polytope is small, and most vector will failed to be reconstructed by solving ℓ^1 minimization. This also suggests why using random projections as in Chapter **??**, because somehow they results in a low distortion embedding of the ℓ^1 ball from $\mathbb{R}^N$ to $\mathbb{R}^P$.

Note that if x_0 is identifiable, so is λx_0 for ρx_0 for $\rho > 0$, and in fact, since the recovery condition only depends on the geometry of the faces of B, the obtained condition (10.1) only depends on $\mathrm{sign}(x_0)$. We denote $\mathcal{A} : y \mapsto x^\star$ the map from y to a solution of $(\mathcal{P}_0(y))$, which we assume is unique for simplicity of exposition. Condition (10.1) thus shows that A and $\mathcal{A}$ are inverse bijection on a family of cones $\mathcal{C}_s = \{x \; ; \; \mathrm{sign}(x) = s\}$ and $A\mathcal{C}_s$ for certain "friendly" sign patterns s. These cones $A\mathcal{C}_s$ form a partition of the image space $\mathbb{R}^P$. Assuming for simplicity that the columns $(a_j)_j$ of A have unit norm, for $P = 3$, the interaction of these $A\mathcal{C}_s$ with the unit sphere of $\mathbb{R}^3$ for a so-called Delaunay triangulation of the sphere (this construction extends to higher dimension by replacing triangle by simplexes), see also Figure 10.7. Such Delaunay triangulation is characterized by the empty spherical cap property (each circumcircle associated to a triangle should not contains any columns vector a_j of the matrix). Figure 10.3 illustrate these conclusions in $\mathbb{R}^2$ and $\mathbb{R}^3$.

10.1.3 Optimality Conditions

In the following, given an index set $I \subset \{1, \dots, N\}$, denoting $A = (a_i)_{i=1}^N$ the columns of A, we denote $A_I \overset{\text{def.}}{=} (a_i)_{i\in I} \in \mathbb{R}^{P\times |I|}$ the extracted sub-matrix. Similarly, for $x \in \mathbb{R}^N$, we denote $x_I \overset{\text{def.}}{=} (x_i)_{i\in I} \in \mathbb{R}^{|I|}$.

The following proposition rephrases the first order optimality conditions in a handy way.

Proposition 29. *x_λ is a solution to $(\mathcal{P}_\lambda(y))$ for $\lambda > 0$ if and only if*

$$\eta_{\lambda,I} = \mathrm{sign}(x_{\lambda,I}) \quad \text{and} \quad \|\eta_{\lambda,I^c}\| \leqslant \lambda$$

where we define

$$I \overset{\text{def.}}{=} \mathrm{supp}(x_\lambda) \overset{\text{def.}}{=} \{i \; ; \; x_{\lambda,i} \neq 0\}, \quad \text{and} \quad \eta_\lambda \overset{\text{def.}}{=} \frac{1}{\lambda} A^*(y - Ax_\lambda). \tag{10.2}$$

Proof. Since $(\mathcal{P}_\lambda(y))$ involves a sum of a smooth and a continuous function, its sub-differential reads

$$\partial f_\lambda(x) = \frac{1}{\lambda} A^*(Ax - y) + \lambda \partial \|\cdot\|_1(x).$$

Thus x_λ is solution to $(\mathcal{P}_\lambda(y))$ if and only if $0 \in \partial f_\lambda(x_\lambda)$, which gives the desired result. □

Note that one has in particular that $\mathrm{supp}(x_\lambda) \subset \mathrm{sat}(\eta_\lambda)$.

The following proposition studies the limit case $\lambda = 0$ and introduces the crucial concept of "dual certificates", which are the Lagrange multipliers of the constraint $\mathcal{L}_y$.

Proposition 30. *$x^\star$ being a solution to $(\mathcal{P}_0(y))$ is equivalent to having $Ax^\star = y$ and that*

$$\exists \eta \in \mathcal{D}_0(y, x^\star) \overset{\text{def.}}{=} \mathrm{Im}(A^*) \cap \partial \|\cdot\|_1(x^\star). \tag{10.3}$$

Proof. Since ($\mathcal{P}_0(y)$) involves a sum with a continuous function, one can also computes it sub-differential as

$$\partial f_0(x) = \partial \iota_{\mathcal{L}_y}(x) + \partial \|\cdot\|_1(x).$$

If $x \in \mathcal{L}_y$, then $\partial \iota_{\mathcal{L}_y}(x)$ is the linear space orthogonal to $\mathcal{L}_y$, i.e. $\ker(A)^\perp = \mathrm{Im}(A^*)$. □

Note that one has in particular that $\mathrm{supp}(x^\star) \subset \mathrm{sat}(\eta)$ for any valid vector $\eta \in \mathcal{D}_0(y, x^\star)$.
Writing $I = \mathrm{supp}(x^\star)$, one thus has

$$\mathcal{D}_0(y, x^\star) = \{\eta = A^* p \;;\; \eta_I = \mathrm{sign}(x^\star_I),\ \|\eta\|_\infty \leqslant 1\}.$$

Although it looks like the definition of $\mathcal{D}_0(y, x^\star)$ depends on the choice of a solution $x^\star$, convex duality (studied in the next chapter) shows that it is not the case (it is the same set for all solutions).

10.1.4 Uniqueness

The following proposition shows that the Lasso selects a set of linearly independent regressors (columns of A). This is why this method is also often called "basis pursuit".

Proposition 31. *For $\lambda \geqslant 0$, there is always a solution $x^\star$ to ($\mathcal{P}_\lambda(y)$) with $I = \mathrm{supp}(x^\star)$ such that $\ker(A_I) = \{0\}$*

Proof. Let x be a solution and denote $I = \mathrm{supp}(x)$. If $\ker(A_I) \neq \{0\}$, one selects $h_I \in \ker(A_I)$ and define for $t \in \mathbb{R}$ the vector $x_t \overset{\text{def.}}{=} x + th$. We denote t_0 the smallest $|t|$ such that $\mathrm{sign}(x_t) \neq \mathrm{sign}(x)$, i.e. $\mathrm{supp}(x_t)$ is strictly included in I. For $t < t_0$, since $Ax_t = Ax$ and $\mathrm{sign}(x_t) = \mathrm{sign}(x)$, x_t still satisfies the same first order condition as x_0, and one can apply either Proposition 30 (for $\lambda = 0$) or Proposition 29 (for $\lambda > 0$), so that x_t is a solution of ($\mathcal{P}_\lambda(y)$). Since the minimized function are lower semi continuous, $x_t \to x_{t_0}$ is still a solution. If $\ker(A_J) \neq \{0\}$ with $J = \mathrm{supp}(x_{t_0})$, one is over, otherwise one can iterate this argument on x_{t_0} in place of x and have a sequence of supports which is strictly decaying in size, so it must terminate. □

This results in particular that if columns of A_I are not independent, then the solution of ($\mathcal{P}_\lambda(y)$) is necessarily non-unique.

Assuming that x_λ is a solution such that $\ker(A_I) = \{0\}$, then from ($\mathcal{P}_\lambda(y)$), one obtains the following implicit expression for the solution

$$x_{\lambda,I} = A_I^+ y - \lambda (A_I^* A_I)^{-1} \mathrm{sign}(x_{\lambda,I}). \tag{10.4}$$

This expression can be understood as a form of generalized soft thresholding (one retrieve the soft thresholding when $A = \mathrm{Id}_N$).

The following useful lemma shows that while solutions x_λ to ($\mathcal{P}_\lambda(y)$) are not necessarily unique, the associated "predictor" (i.e. denoised version of y) Ax_λ is however always uniquely defined. Note that according to (10.5), one has

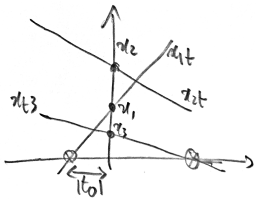

Figure 10.4: Trajectory $(x_t)_t$.

$$\Phi x_\lambda = \mathrm{Proj}_{\mathrm{Im}(A_I)} y - \lambda A_I (A_I^* A_I)^{-1} \mathrm{sign}(x_{\lambda,I}). \tag{10.5}$$

so up to a $O(\lambda)$ bias, this predictor is an orthogonal projection on a low dimensional subspace indexed by I.

Lemma 3. *For $\lambda \geqslant 0$, if (x_1, x_2) are solution to ($\mathcal{P}_\lambda(y)$), then $Ax_1 = Ax_2$.*

Proof. For $\lambda = 0$, this is trivial because $Ax_1 = Ax_2 = y$. Otherwise, let us assume $Ax_1 \neq Ax_2$. Then for $x = (x_1 + x_2)/2$, one has

$$\|x\|_1 \leqslant \frac{\|x_1\|_1 + \|x_2\|_2}{2} \quad \text{and} \quad \|Ax - y\|^2 < \frac{\|Ax_1 - y\|^2 + \|Ax_2 - y\|^2}{2}$$

where the second inequality follows from the strict convexity of the square. This shows that

$$\frac{1}{2\lambda}\|Ax - y\|^2 + \|x\|_1 < \frac{1}{2\lambda}\|Ax_1 - y\|^2 + \|x_1\|_1,$$

which is a contradiction to the optimality of x_1. □

Proposition 32. *For $\lambda > 0$, let x_λ be a solution to ($\mathcal{P}_\lambda(y)$) and denote $\eta_\lambda \overset{\text{def.}}{=} \frac{1}{\lambda} A^*(y - Ax_\lambda)$. We define the "extended support" as*

$$J \overset{\text{def.}}{=} \mathrm{sat}(\eta_\lambda) \overset{\text{def.}}{=} \{i \;;\; |\eta_{\lambda,i}| = 1\}\,.$$

If $\ker(A_J) = \{0\}$ then x_λ is the unique solution of ($\mathcal{P}_\lambda(y)$).

Proof. If $\tilde{x}_\lambda$ is also a minimizer, then by Lemma 3, $Ax_\lambda = A\tilde{x}_\lambda$, so that in particular they share the same dual certificate

$$\eta_\lambda = \frac{1}{\lambda} A^*(y - Ax_\lambda) = \frac{1}{\lambda} A^*(y - A\tilde{x}_\lambda).$$

The first order condition, Proposition 29, shows that necessarily $\mathrm{supp}(x_\lambda) \subset J$ and $\mathrm{supp}(\tilde{x}_\lambda) \subset J$. Since $A_J x_{\lambda,J} = A_J \tilde{x}_{\lambda,J}$, and since $\ker(A_J) = \{0\}$, one has $x_{\lambda,J} = \tilde{x}_{\lambda,J}$, and thus $x_\lambda = \tilde{x}_\lambda$ because of their supports are included in J. □

Proposition 33. *Let $x^\star$ be a solution to ($\mathcal{P}_0(y)$). If there exists $\eta \in \mathcal{D}_0(y, x^\star)$ such that $\ker(A_J) = \{0\}$ where $J \overset{\text{def.}}{=} \mathrm{sat}(\eta)$ then $x^\star$ is the unique solution of ($\mathcal{P}_0(y)$).*

Proof. The proof is the same as for Proposition 32, replacing η_λ by η. □

These propositions can be used to show that if A is drawn according to a distribution having a density over $\mathbb{R}^{p\times n}$, then with probability 1 on the matrix A, the solution to ($\mathcal{P}_\lambda(y)$) is unique. Note that this results is not true if A is non random but y is.

10.1.5 Duality

We now related the first order conditions and "dual certificate" introduced above to the duality theory detailed in Section **??**. This is not strictly needed to derive the theory of sparse regularization, but this offers an alternative point of view and allows to better grasp the role played by the certificates.

Theorem 11. *For any $\lambda \geqslant 0$ (i.e. including $\lambda = 0$), one has strong duality between ($\mathcal{P}_\lambda(y)$) and*

$$\sup_{p\in\mathbb{R}^P} \left\{ \langle y,\, p\rangle - \frac{\lambda}{2}\|p\|^2 \;;\; \|A^*p\|_\infty \leqslant 1 \right\}. \tag{10.6}$$

One has for any $\lambda \geqslant 0$ that $(x^\star, p^\star)$ are primal and dual solutions if and only if

$$A^*p^\star \in \partial\|\cdot\|_1(x^\star) \quad \Leftrightarrow \quad (I \subset \mathrm{sat}(A^*p) \quad \text{and} \quad \mathrm{sign}(x^\star_I) = A^*_I p)\,, \tag{10.7}$$

where we denoted $I = \mathrm{supp}(x^\star)$, and furthermore, for $\lambda > 0$,

$$p^\star = \frac{y - Ax^\star}{\lambda}.$$

while for $\lambda = 0$, $Ax^\star = y$.

Proof. There are several ways to derive the same dual. One can for instance directly use the Fenchel-Rockafeller formula (**??**). But it is instructive to do the computations using Lagrange duality. One can first consider the following re-writing of the primal problem

$$\min_{x\in\mathbb{R}^N} \{f(z) + \|x\|_1 \;;\; Ax = z\} = \min_{x\in\mathbb{R}^N} \sup_{p\in\mathbb{R}^p} \mathcal{L}(x, z, p) \overset{\text{def.}}{=} f_\lambda(z) + \|x\|_1 + \langle z - Ax,\, p\rangle$$

where $f_\lambda(z) \overset{\text{def.}}{=} \frac{1}{2\lambda}\|z - y\|^2$ if $\lambda > 0$ and $f(z) = \iota_{\{y\}}(z)$ if $\lambda = 0$. For $\lambda > 0$ since f_λ and $\|\cdot\|_1$ are continuous, strong duality holds. For $\lambda = 0$, since the constraint appearing in f_0 is linear (actually a singleton), strong duality holds also. Thus using Theorem **??**, one can exchange the min and the max and obtains

$$\max_{p\in\mathbb{R}^P} \;(\min z\; \langle z,\, p\rangle + f_\lambda(z)) + \;(\min x\; \|x\|_1 - \langle x,\, A^*p\rangle) = \max_{p\in\mathbb{R}^P} \; - f^*_\lambda(-p) - (\|\cdot\|_1)^*(A^*p).$$

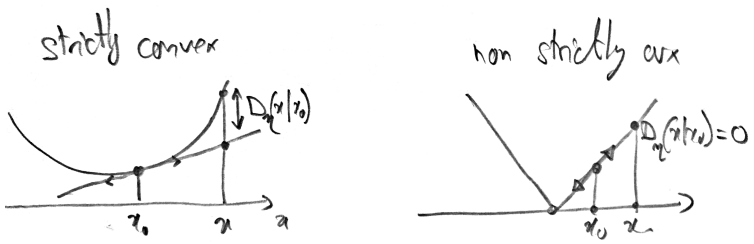

Figure 10.5: Visualization of Bregman divergences.

Using (**??**), one has that $(\|\cdot\|_1^* = \iota_{\|\cdot\|_\infty \leqslant 1}$. For $\lambda > 0$, one has using Proposition **??** that

$$f_\lambda^* = (\frac{1}{2\lambda}\|\cdot - y\|^2)^* = \frac{1}{\lambda}(\frac{1}{2}\|\cdot - y\|^2)^*(\lambda\cdot) = \frac{1}{2\lambda}\|\lambda\cdot\|^2 + \langle\cdot, y\rangle$$

which gives the desired dual problem. The first order optimality conditions read $Ax^\star = z^\star$ and

$$0 \in \partial\|\cdot\|_1(x^\star) - A^*p^\star \quad \text{and} \quad 0 \in \partial f_\lambda(z^\star) + p^\star.$$

The first condition is equivalent to (10.7). For $\lambda > 0$, f_λ is smooth, and the second condition is equivalent to

$$p^\star = \frac{y - A^*x^\star}{\lambda} \quad \text{and} \quad A^*p^\star \in \partial\|\cdot\|_1(x^\star)$$

which are the desired formula. For $\lambda = 0$, the second condition holds as soon as $z^\star = Ax^\star = y$. □

Note that in the case $\lambda > 0$, (10.6) is strongly convex, and in fact the optimal solution p_λ is computed as an orthogonal projection

$$p_\lambda \in \operatorname*{argmin}_{p\in\mathbb{R}^P} \{\|p - y/\lambda\| \;;\; \|A^*p\|_\infty \leqslant 1\}.$$

The sup in (10.6) is thus actually a max if $\lambda > 0$. If $\lambda > 0$, in case $\ker(A^*) = \mathrm{Im}(A)^\perp = \{0\}$, the constraint set of the dual is bounded, so that the sup is also a max.

10.2 Consistency and Sparsitency

10.2.1 Bregman Divergence Rates for General Regularizations

Here we consider the case of a general regularization of the form

$$\min_{x\in\mathbb{R}^N} \frac{1}{2\lambda}\|Ax - y\|^2 + J(x) \tag{10.8}$$

for a convex regularizer J.

For any $\eta \in \partial J(x_0)$, we define the associated Bregman divergence as

$$D_\eta(x|x_0) \overset{\text{def.}}{=} J(x) - J(x_0) - \langle\eta, x - x_0\rangle.$$

One has $D_\eta(x_0|x_0)$, and since J is convex, one has $D_\eta(x|x_0) \geqslant 0$ **[ToDo: put here drawings]**.

In the case where J is differentiable, since $\partial J(x_0) = \{\nabla J(x_0)\}$, this divergence simply reads

$$D(x|x_0) \overset{\text{def.}}{=} J(x) - J(x_0) - \langle\nabla J(x_0), x - x_0\rangle.$$

If furthermore J is strictly convex, then $D(x|x_0) = 0$ if and only if $x = x_0$, so that $D(\cdot|\cdot)$ is similar to a distance function (but it does not necessarily satisfies the triangular inequality.

If $J = \|\cdot\|^2$, then $D(x|x_0) = \|x - x_0\|^2$ is the Euclidean norm. If $J(x) = \sum_i x_i(\log(x_i) - 1) + \iota_{\mathbb{R}^+}(x_i)$ is the entropy, then

$$D(x|x_0) = \sum_i x_i \log\left(\frac{x_i}{x_{0,i}}\right) + x_{0,i} - x_i$$

is the so-called Kulback-Leibler divergence on $\mathbb{R}^N_+$.

The following theorem, which is due to Burger-Osher, state a linear rate in term of this Bregman divergence.

Theorem 12. *If there exists*

$$\eta = A^*p \in \mathrm{Im}(A^*) \cap \partial J(x_0), \tag{10.9}$$

then one has for any x_λ solution of (10.8)

$$D_\eta(x_\lambda|x_0) \leqslant \frac{1}{2}\left(\frac{\|w\|}{\sqrt{\lambda}} + \sqrt{\lambda}\|p\|\right)^2. \tag{10.10}$$

Futhermore, one has the useful bound

$$\|Ax_\lambda - y\| \leqslant \|w\| + (\sqrt{2}+1)\|p\|\lambda. \tag{10.11}$$

Proof. The optimality of x_λ for (10.8) implies

$$\frac{1}{2\lambda}\|Ax_\lambda - y\|^2 + J(x_\lambda) \leqslant \frac{1}{2\lambda}\|Ax_0 - y\|^2 + J(x_0) = \frac{1}{2\lambda}\|w\|^2 + J(x_0).$$

Hence, using $\langle \eta,\, x_\lambda - x_0\rangle = \langle p,\, Ax_\lambda - Ax_0\rangle = \langle p,\, Ax_\lambda - y + w\rangle$, one has

$$\begin{aligned} D_\eta(x_\lambda|x_0) &= J(x_\lambda) - J(x_0) - \langle \eta,\, x_\lambda - x_0\rangle \leqslant \frac{1}{2\lambda}\|w\|^2 - \frac{1}{2\lambda}\|Ax_\lambda - y\|^2 - \langle p,\, Ax_\lambda - y\rangle - \langle p,\, w\rangle \\ &= \frac{1}{2\lambda}\|w\|^2 - \frac{1}{2\lambda}\|Ax_\lambda - y + \lambda p\|^2 + \lambda\|p\|^2 - \langle p,\, w\rangle \\ &\leqslant \frac{1}{2\lambda}\|w\|^2 + \frac{\lambda}{2}\|p\|^2 + \|p\|\|w\| = \frac{1}{2}\left(\frac{\|w\|}{\sqrt{\lambda}} + \sqrt{\lambda}\|p\|\right)^2. \end{aligned}$$

From the second line above, since $D_\eta(x_\lambda|x_0) \geqslant 0$, one has using Cauchy-Schwartz

$$\|Ax_\lambda - y + \lambda p\|^2 \leqslant \|w\|^2 + 2\lambda^2\|p\|^2 + 2\lambda\|p\|\|w\| \leqslant \|w\|^2 + 2\sqrt{2}\|p\|\|w\|\lambda + 2\lambda^2\|p\|^2 = \left(\|w\| + \sqrt{2}\lambda\|p\|\right)^2.$$

Hence

$$\|Ax_\lambda - y\| \leqslant \|Ax_\lambda - y + \lambda p\| + \lambda\|p\| \leqslant \|w\| + \sqrt{2}\lambda\|p\| + \lambda\|p\|.$$

□

Choosing $\lambda = \|w\|/\|p\|$ in (10.10), one thus obtain a linear rate in term of Bregman divergence $D_\eta(x_\lambda|x_0) \leqslant 2\|w\|\|p\|$. For the simple case of a quadratic regularized $J(x) = \|x\|^2/2$, as used in Section **??**, one sees that the source conditions (10.9) simply reads

$$x_0 \in \mathrm{Im}(A^*)$$

which is equivalent to (8.12) with exponent $\beta = \frac{1}{2}$, and under this condition, (10.10) gives the following sub-linear rate in term of the ℓ^2 norm

$$\|x_0 - x_\lambda\| \leqslant 2\sqrt{\|w\|\|p\|}.$$

[ToDo: This seems inconsistent, this should corresponds to $\beta = 1$ to obtain the same rates in both theorems!]

Note that the "source condition" (10.9) is equivalent to x_0 such that $Ax_0 = y$ is a solution to the constraint problem

$$\min_{Ax=y} J(x).$$

So simply being a solution of the constraint noiseless problem thus implies a linear rate for the resolution of the noisy problem in term of the Bregman divergence.

10.2.2 Linear Rates in Norms for ℓ^1 Regularization

The issue with the control (10.10) of the error in term of Bregman divergence is that it is not "distance-like" for regularizers J which are not strictly convex. This is in particular the case for the ℓ^1 norm $J = \|\cdot\|_1$ which we now study.

The following fundamental lemma shows however that this Bregman divergence for ℓ^1 behave like a distance (and in fact controls the ℓ^1 norm) on the indexes where η does not saturate.

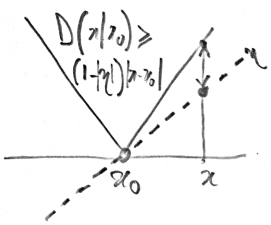

Figure 10.6: Controlling Bregman divergence with the ℓ^1 norm when η is not saturating.

Lemma 4. *For $\eta \in \partial\|\cdot\|_1(x_0)$, let $J \overset{\text{def.}}{=} \mathrm{sat}(\eta)$. Then*

$$D_\eta(x|x_0) \geqslant (1 - \|\eta_{J^c}\|_\infty)\|(x - x_0)_{J^c}\|_1. \tag{10.12}$$

Proof. Note that $x_{0,J^c} = 0$ since $\mathrm{supp}(x_0) \subset \mathrm{sat}(\eta)$ by definition of the sub-differential of the ℓ^1 norm. Since the ℓ^1 norm is separable, each term in the sum defining $D_\eta(x|x_0)$ is positive, hence

$$\begin{aligned} D_\eta(x|x_0) &= \sum_i |x_i| - |x_{0,i}| - \eta_i(x_i - x_{0,i}) \geqslant \sum_{i\in J^c} |x| - |x_0| - \eta_i(x_i - x_{0,i}) \\ &= \sum_{i\in J^c} |x_i| - \eta_i x_i \geqslant \sum_{i\in J^c} (1 - |\eta_i|)|x_i| \geqslant (1 - \|\eta_{J^c}\|_\infty) \sum_{i\in J^c} |x_i| = (1 - \|\eta_{J^c}\|_\infty) \sum_{i\in J^c} |x_i - x_{0,i}|. \end{aligned}$$

□

The quantity $1 - \|\eta_{J^c}\|_\infty > 0$ controls how much η is "inside" the sub-differential. The larger this coefficients, the better is the control of the Bregman divergence.

The following theorem uses this lemma to state the convergence rate of the sparse regularized solution, under the same hypothesis has Proposition 33 (with $x^\star = x_0$).

Theorem 13. *If there exists*

$$\eta \in \mathcal{D}_0(Ax_0, x_0) \tag{10.13}$$

and $\ker(A_J) = \{0\}$ where $J \overset{\text{def.}}{=} \mathrm{sat}(\eta)$ then choosing $\lambda = c\|w\|$, there exists C (depending on c) such that any solution x_λ of $\mathcal{P}(Ax_0 + w)$ satisfies

$$\|x_\lambda - x_0\| \leqslant C\|w\|. \tag{10.14}$$

Proof. We denote $y = Ax_0 + w$. The optimality of x_λ in ($\mathcal{P}_\lambda(y)$) implies

$$\frac{1}{2\lambda}\|Ax_\lambda - y\|^2 + \|x_\lambda\|_1 \leqslant \frac{1}{2\lambda}\|Ax_0 - y\|^2 + \|x_0\|_1 = \frac{1}{2\lambda}\|w\|^2 + \|x_0\|_1$$

and hence

$$\|Ax_\lambda - y\|^2 \leqslant \|w\|^2 + 2\lambda\|x_0\|_1$$

Using the fact that A_J is injective, one has $A_J^+ A_J = \mathrm{Id}_J$, so that

$$\begin{aligned} \|(x_\lambda - x_0)_J\|_1 &= \|A_J^+ A_J (x_\lambda - x_0)_J\|_1 \leqslant \|A_J^+\|_{1,2}\|A_J x_{\lambda,J} - y + w\| \leqslant \|A_J^+\|_{1,2}\left(\|A_J x_{\lambda,J} - y\| + \|w\|\right) \\ &\leqslant \|A_J^+\|_{1,2}\left(\|Ax_\lambda - y\| + \|A_{J^c} x_{\lambda,J^c}\| + \|w\|\right) \\ &\leqslant \|A_J^+\|_{1,2}\left(\|Ax_\lambda - y\| + \|A_{J^c}\|_{2,1}\|x_{\lambda,J^c} - x_{0,J^c}\|_1 + \|w\|\right) \\ &\leqslant \|A_J^+\|_{1,2}\left(\|w\| + (\sqrt{2}+1)\|p\|\lambda + \|A_{J^c}\|_{2,1}\|x_{\lambda,J^c} - x_{0,J^c}\|_1 + \|w\|\right) \end{aligned}$$

where we used $x_{0,J^c} = 0$ and (10.11). One plug this bound in the decomposition, and using (10.12) and (10.10)

$$\begin{aligned}
\|x_\lambda - x_0\|_1 &= \|(x_\lambda - x_0)_J\|_1 + \|(x_\lambda - x_0)_{J^c}\|_1 \\
&\leqslant \|(x_\lambda - x_0)_{J^c}\|_1 \left(1 + \|A_J^+\|_{1,2}\|A_{J^c}\|_{2,1}\right) + \|A_J^+\|_{1,2}\left((\sqrt{2}+1)\|p\|\lambda + 2\|w\|\right) \\
&\leqslant \frac{D_\eta(x|x_0)}{1-\|\eta_{J^c}\|_\infty}\left(1 + \|A_J^+\|_{1,2}\|A_{J^c}\|_{2,1}\right) + \|A_J^+\|_{1,2}\left((\sqrt{2}+1)\|p\|\lambda + 2\|w\|\right) \\
&\leqslant \frac{\frac{1}{2}\left(\frac{\|w\|}{\sqrt{\lambda}} + \sqrt{\lambda}\|p\|\right)^2}{1-\|\eta_{J^c}\|_\infty}\left(1 + \|A_J^+\|_{1,2}\|A_{J^c}\|_{2,1}\right) + \|A_J^+\|_{1,2}\left((\sqrt{2}+1)\|p\|\lambda + 2\|w\|\right).
\end{aligned}$$

Thus setting $\lambda = c\|w\|$, one obtains the constant

$$C \overset{\text{def.}}{=} \frac{\frac{1}{2}\left(\frac{1}{\sqrt{c}} + \sqrt{c}\|p\|\right)^2}{1-\|\eta_{J^c}\|_\infty}\left(1 + \|A_J^+\|_{1,2}\|A_{J^c}\|_{2,1}\right) + \|A_J^+\|_{1,2}\left((\sqrt{2}+1)\|p\|c + 2\right).$$

□

Note that this theorem does not imply that x_λ is a unique solution, only x_0 is unique in general. The condition (10.13) is often called a "source condition", and is strengthen by imposing a non-degeneracy $\ker(A_J) = \{0\}$. This non-degeneracy imply some stability in ℓ^2 sense (10.14). The result (10.14) shows a linear rate, i.e. the (possibly multi-valued) inverse map $y \mapsto x_\lambda$ is Lipschitz continuous.

It should be compared with Theorem 10 on linear methods for inverse problem regularization, which only gives sub-linear rate. The sources conditions in the linear (8.12) and non-linear (10.13) cases are however very different. In the linear case, for $\beta = 1/2$, it reads $x_0 \in \mathrm{Im}(A^*) = \ker(A)^\perp$, which is mandatory because linear method cannot recover anything in $\ker(A)$. On contrary, the non-linear source condition only requires that η to be in $\mathrm{Im}(A^*)$, and is able (in the favorable cases of course) to recover information in $\ker(A)$.

10.2.3 Sparsistency for Low Noise

Theorem 13 is abstract in the sense that it rely on hypotheses which are hard to check. The crux of the problem, to be able to apply this theorem, is to be able to "construct" a valid certificate (10.13). We now give a powerful "recipe" which – when it works – not only give a sufficient condition for linear rate, but also provides "support stability".

For any solution x_λ of $(\mathcal{P}_\lambda(y))$, as already done in (10.2), we define the (unique, independent of the chosen solution) dual certificate

$$\eta_\lambda \overset{\text{def.}}{=} A^* p_\lambda \quad \text{where} \quad p_\lambda \overset{\text{def.}}{=} \frac{y - Ax_\lambda}{\lambda}.$$

The following proposition shows that p_λ converge to a very specific dual certificate of the constrained problem, which we coined "minimal norm" certificate.

Proposition 34. *If $y = Ax_0$ where x_0 is a solution to $(\mathcal{P}_\lambda(y = Ax_0))$, one has*

$$p_\lambda \to p_0 \overset{\text{def.}}{=} \operatorname*{argmin}_{p \in \mathbb{R}^P} \left\{\|p\| \;;\; A^*p \in \mathcal{D}_0(y, x_0)\right\}. \tag{10.15}$$

The vector $\eta_0 \overset{\text{def.}}{=} A^*p_0$ is called the "minimum norm certificate".

Proof. This follows from the fact that p_λ is the unique solution to (10.6) and then applying the same proof as the one done in Proposition 25 to study the small λ limit of penalized problems. □

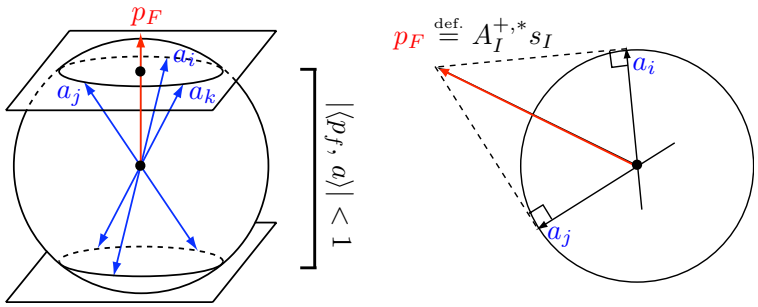

Figure 10.7: Visualization of the condition that $\|\eta_F\|_\infty \leqslant 1$ as a spherical Delaunay triangulation constraint that all Delaunay spherical caps indexes by identifiable vector should be empty of $(\pm a_i)_i$.

This proposition shows that, while dual certificate $\mathcal{D}_0(y, x_0)$ for $\lambda = 0$ are non-unique, taking the limit as $\lambda \to 0$ singles-out a specific one, which is of paramount importance to study stability of the support when the noise and λ are small.

A major difficulty in computing (10.24) is that it should satisfy the non-linear constraint $\|\eta_0\|_\infty$. One thus can "simplify" this definition by removing this ℓ^∞ constraint and define the so-called "minimum norm certificate"

$$\eta_F \stackrel{\text{def.}}{=} A^* p_F \quad \text{where} \quad p_F \stackrel{\text{def.}}{=} \operatorname*{argmin}_{p \in \mathbb{R}^P} \{\|p\| \;;\; A_I^* p = \operatorname{sign}(x_{0,I})\}. \tag{10.16}$$

The notation "η_F" refers to the "Fuchs" certificate, which we named in honour of J-J. Fuchs who first used it to study ℓ^1 minimization.

We insists that p_F is not necessarily a valid certificate (hence the naming "pre-certificate") since one does not have in general $\|\eta_F\|_\infty \leqslant 1$. The vector p_F is a least square solution to the linear system $A_I^* p = \operatorname{sign}(x_{0,I})$, and it can thus be compute in closed form using the pseudo-inverse $p_F = A_I^{*,+} \operatorname{sign}(x_{0,I})$ (see Proposition (23)). In case $\ker(A_I) = \{0\}$, one has the simple formula

$$p_F = A_I (A_I^* A_I)^{-1} \operatorname{sign}(x_{0,I}).$$

Denoting $C \stackrel{\text{def.}}{=} A^* A$ the "correlation" matrix, one has the nice formula

$$\eta_F = C_{\cdot,I} C_{I,I}^{-1} \operatorname{sign}(x_{0,I}). \tag{10.17}$$

The following proposition relates η_F to η_0, and shows that η_F can be used as a "proxy" for η_0

Proposition 35. *If $\|\eta_F\|_\infty \leqslant 1$, then $p_F = p_0$ and $\eta_F = \eta_0$.*

The condition $\|\eta_F\|_\infty \leqslant 1$ implies that x_0 is solution to ($\mathcal{P}_0(y)$). The following theorem shows that if one strengthen this condition to impose a non-degeneracy on η_F, then one has linear rate with a stable support in the small noise regime.

Before proceeding to the proof, let us note that the constraint $\|\eta_F\|_\infty \leqslant 1$ corresponds to the definition of the spherical Delaunay triangulation, as highlighted by Figure 10.7. This remark was made to us by Charles Dossal.

Remark 1 (Operator norm). In the proof, we use the $\ell^p - \ell^q$ matrix operator norm, which is defined as

$$\|B\|_{p,q} \stackrel{\text{def.}}{=} \max \{\|Bu\|_q \;;\; \|u\|_p \leqslant 1\}.$$

For $p = q$, we denote $\|B\|_p \stackrel{\text{def.}}{=} \|B\|_{p,p}$. For $p = 2$, $\|B\|_2$ is the maximum singular value, and one has

$$\|B\|_1 = \max_j \sum_i |B_{i,j}| \quad \text{and} \quad \|B\|_\infty = \max_i \sum_j |B_{i,j}|.$$

Theorem 14. *If*

$$\|\eta_F\|_\infty \leqslant 1 \quad \textit{and} \quad \|\eta_{F,I^c}\|_\infty < 1,$$

and $\ker(A_I) = \{0\}$*, then there exists* C, C' *such that if* $\max(\|w\|, \|w\|/\lambda) \leqslant C$*, then the solution* x_λ *of* ($\mathcal{P}_\lambda(y)$) *is unique, is supported in* I*, and in fact*

$$x_{\lambda,I} = x_{0,I} + A_I^+ w - \lambda (A_I^* A_I)^{-1} \operatorname{sign}(x^\star_{0,I}). \tag{10.18}$$

In particular, $\|x_\lambda - x_0\| = O(\|A^* w\|_\infty) = O(\|w\|)$.

Proof. In the following we denote $T \overset{\text{def.}}{=} \min_{i \in I} |x_{0,i}|$ the signal level, and $\delta \overset{\text{def.}}{=} \|A^* w\|_\infty$ which is the natural way to measure the noise amplitude in the sparse setting. We define $s \overset{\text{def.}}{=} \operatorname{sign}(x_0)$, and consider the "ansatz" (10.18) and thus define the following candidate solution

$$\hat{x}_I \overset{\text{def.}}{=} x_{0,I} + A_I^+ w - \lambda (A_I^* A_I)^{-1} s_I, \tag{10.19}$$

and $\hat{x}_{I^c} = 0$. The goal is to show that $\hat{x}$ is indeed the unique solution of ($\mathcal{P}_\lambda(y)$).
Step 1. The first step is to show sign consistency, i.e. that $\operatorname{sign}(\hat{x}) = s$. This is true if $\|x_{0,I} - \hat{x}_I\|_\infty < T$, and is thus implied by

$$\|x_{0,I} - \hat{x}_I\|_\infty \leqslant K \|A_I^* w\|_\infty + K\lambda < T \quad \text{where} \quad K \overset{\text{def.}}{=} \|(A_I^* A_I)^{-1}\|_\infty, \tag{10.20}$$

where we used the fact that $A_I^+ = (A_I^* A_I)^{-1} A_I^*$.
Step 2. The second step is to check the first order condition of Proposition 32, i.e. $\|\hat{\eta}_{I^c}\|_\infty < 1$, where $\lambda \hat{\eta} = A^*(y - A\hat{x})$. This implies indeed that $\hat{x}$ is the unique solution of ($\mathcal{P}_\lambda(y)$). One has

$$\begin{aligned} \lambda \hat{\eta} &= A^*(A_I x_{0,I} + w - A_I \left(x_{0,I} + A_I^+ w - \lambda (A_I^* A_I)^{-1} s_I \right)) \\ &= A^*(A_I A_I^+ - \mathrm{Id}) w + \lambda \eta_F. \end{aligned}$$

The condition $\|\hat{\eta}_{I^c}\|_\infty < 1$ is thus implied by

$$\|A_{I^c}^* A_I (A_I^* A_I)^{-1}\|_\infty \|A_I^* w\|_\infty + \|A_{I^c}^* w\|_\infty + \lambda \|\eta_{F,I^c}\|_\infty \leqslant R \|A_I^* w\|_\infty - S\lambda < 0 \tag{10.21}$$

$$R \overset{\text{def.}}{=} KL + 1 \quad \text{and} \quad S \overset{\text{def.}}{=} 1 - \|\eta_{F,I^c}\|_\infty > 0$$

where we denoted $L \overset{\text{def.}}{=} \|A_{I^c}^* A_I\|_\infty$, and also we used the hypothesis $\|\eta_{F,I^c}\|_\infty < 1$.
Conclusion. Putting (10.20) and (10.21) together shows that $\hat{x}$ is the unique solution if (λ, w) are such that the two linear inequations are satisfies

$$\mathcal{R} = \left\{ (\delta, \lambda) \;;\; \delta + \lambda < \frac{T}{K} \quad \text{and} \quad R\delta - S\lambda < 0 \right\}$$

This region $\mathcal{R}$ is trianglular-shaped, and includes the following "smaller" simpler triangle

$$\tilde{\mathcal{R}} = \left\{ (\delta, \lambda) \;;\; \frac{\delta}{\lambda} < \frac{S}{R} \quad \text{and} \quad \lambda < \lambda_{\max} \right\} \quad \text{where} \quad \lambda_{\max} \overset{\text{def.}}{=} \frac{T(KL+1)}{K(R+S)}. \tag{10.22}$$

□

It is important to realize that Theorem 14 operates in a "small noise" regime, i.e. $\|w\|$ (and hence λ) needs to be small enough for the support to be identifiable (otherwise small amplitude comment of x_0 will be killed by the regularization). In contrast, Theorem 13 is "global" and holds for any noise level $\|w\|$. The price to pay is that one has no controls about the support (and one does not even knows whether x_λ is unique) and than the constant involved are more pessimistic.

A nice feature of this proof is that it gives access to explicit constant, involving the three key parameter K, L, S, which controls:

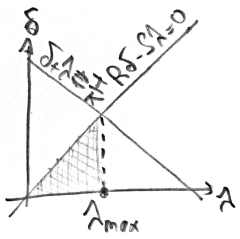

Figure 10.8: Zone in the (λ, δ) where sign consistency occurs.

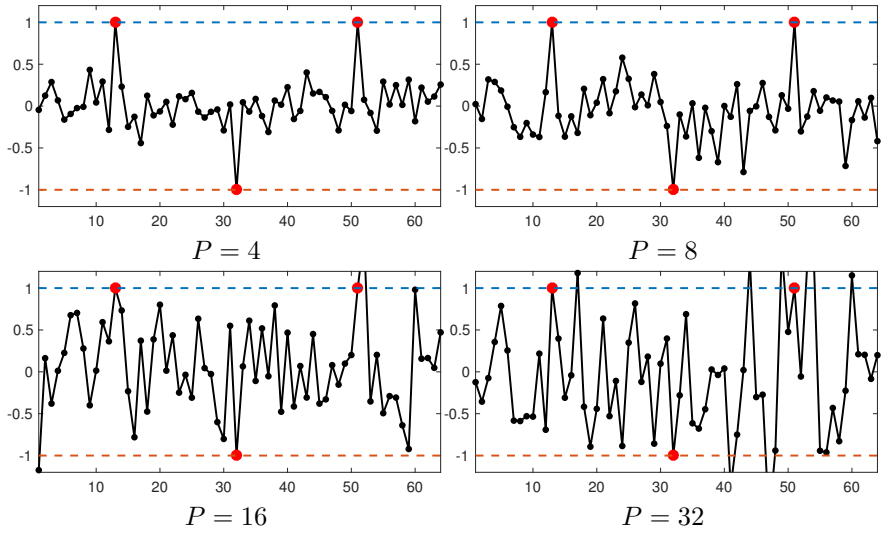

Figure 10.9: Display of certificate η_F for a $A \in \mathbb{R}^{p \times n}$, $n = 64$, with independent Gaussian entries.

– K accounts for the conditionning of the operator on the support I ;
– L accounts for the worse correlation between atoms inside and outside the support ;
– S accounts for how much the certificates η_F is non-degenerate.

The constant on $\|A^* w\|/\lambda$ and on λ are given by (10.22). Choosing (which is in practice impossible, because it requires knowledge about the solution) the smallest possible λ gives $\lambda = \delta \frac{S}{R}$ and in this regime the error is bounded in ℓ^∞ (using other error norms would simply leads to using other matrix norm)

$$\|x_0 - x_\lambda\|_\infty \leqslant \left(1 + \frac{KL+1}{S}\right) K\delta.$$

The crux of the analysis of the performance (in term of support stability) of ℓ^1 regularization is to be able to say wether, for some class of signal x_0 of interest, η_F is a valid certificate, i.e. $\|\eta_F\|_\infty \leqslant 1$. Figure 10.9 displays numerically what one obtains when A is random. One see that η_F is non-degenerate when P is large enough. Section **??** performs a mathematical analysis of this phenomena.

10.2.4 Sparsistency for Arbitrary Noise

A flow of the previous approach is that it provides only asymptotic sparsistency when the noise is small enough with respect to the signal. In particular, it cannot be used to asses the performance of sparse recovery for approximately sparse (e.g. compressible signal), for which the residual error is of the error of the signal itself (and thus not small).

This can be aleviated by controlling all possible certificate associated to all the sign pattern of a given support. This is equivalent to the ERC condition of Tropp. **[ToDo: write me]**

The proof proceeds by restricting the optimization to the support, which is still a convex program, and then showing that this candidate solution is indeed the correct one. Since one does not know in advance the sign of this candidate, this is why one needs to control all possible certificates.

10.3 Sparse Deconvolution Case Study

Chapter **??** studies the particular case where A is random, in which case it is possible to make very precise statement about wether η_F is a valid certificate.

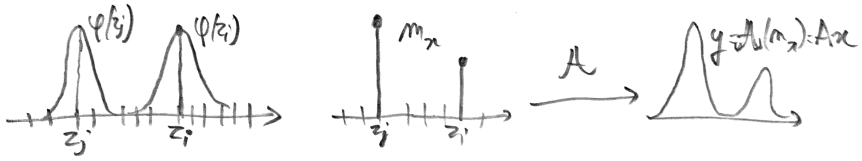

Figure 10.10: Convolution operator.

Another interesting case study, which shows the limitation of this approach, is the case of "super-resolution". It corresponds to inverse problems where the columns $(a_i)_i$ of A are highly correlated, since typically they are obtained by sampling a smooth kernel.

We thus consider the case where $a_i = \varphi(z_i)$ where the $(z_i)_i \subset \mathbb{X}$ is a sampling grid of a domain $\mathbb{X}$ and $\varphi : \mathbb{X} \to \mathcal{H}$ is a smooth map. One has

$$Ax = \sum_i x_i \varphi(z_i).$$

Since we seek for sparse x, one can view x as representing the weights of a discrete measure $m_x \overset{\text{def.}}{=} \sum_{i=1}^N x_i \delta_{z_i}$ where the dirac masses are constraint to be on the sampling grid.

The matrix A is a discretized version of an infinite dimensional operator mapping Radon measures to vectors of observations $\mathcal{A} : m \in \mathcal{M}(\mathbb{X}) \mapsto y = \mathcal{A}m \in \mathcal{H}$

$$\mathcal{A}(m) \overset{\text{def.}}{=} \int_{\mathbb{X}} \varphi(x) \mathrm{d}m(x).$$

Indeed, one has for discrete measure $\mathcal{A}(m_x) = Ax$.

A typical example is when using $\mathcal{H} = L^2(\mathbb{X})$ with $\mathbb{X} = \mathbb{R}^d$ or $\mathbb{X} = \mathbb{T}^d$ and $\varphi(z) = \tilde{\varphi}(z - \cdot)$, which corresponds to a convolution

$$(\mathcal{A}m)(z) = \int \tilde{\varphi}(z - x) \mathrm{d}m(x) = (\tilde{\varphi} \star m)(z).$$

Note that here $\mathcal{H}$ is infinite dimensional, and to get finite dimensional observations, it suffices to sample the output and consider $\varphi(z) = (\varphi(z - r_j))_{j=1}^P$ (note that the observation grid $r \in \mathbb{X}^P$ can be different from the recovery grid $z \in \mathbb{X}^N$).

Another example, actually very much related, is when using $\varphi(z) = (e^{\mathrm{i}kz})_{k=-f_c}^{f_c}$ on $\mathbb{X} = \mathbb{T}$, so that $\mathcal{A}$ corresponds to computing the f_c low-frequencies of the Fourier transform of the measure

$$\mathcal{A}(m) = \left(\int_{\mathbb{T}} e^{\mathrm{i}kx} \mathrm{d}m(x) \right)_{k=-f_c}^{f_c}.$$

The operator $\mathcal{A}^*\mathcal{A}$ is a convolution against an ideal low pass (Dirichlet) kernel. By weighting the Fourier coefficients, one can this way model any low pass filtering on the torus.

Yet another interesting example on $\mathbb{X} = \mathbb{R}^+$ is the Laplace transform

$$\mathcal{A}(m) = z \mapsto \int_{\mathbb{R}^+} e^{-xz} \mathrm{d}m(x).$$

We denote the "continuous" covariance as

$$\forall\, (z, z') \in \mathbb{X}^2, \quad \mathcal{C}(z, z') \overset{\text{def.}}{=} \langle \varphi(z), \varphi(z') \rangle_{\mathcal{H}}.$$

Note that this $\mathcal{C}$ is the kernel associated to the operator $\mathcal{A}^*\mathcal{A}$. The discrete covariance, defined on the computational grid is $C = (\mathcal{C}(z_i, z_i')))_{(i,i')} \in \mathbb{R}^{N \times N}$, while its restriction to some support set I is $C_{I,I} = (\mathcal{C}(z_i, z_i')))_{(i,i') \in I^2} \in \mathbb{R}^{I \times I}$.

Using (10.17), one sees that η_F is obtained as a sampling on the grid of a "continuous' " certificate $\tilde{\eta}_F$

$$\eta_F = (\tilde{\eta}_F(z_i))_{i=1}^N \in \mathbb{R}^N,$$

$$\text{where} \quad \tilde{\eta}_F(x) = \sum_{i \in I} b_i \mathcal{C}(x, z_i) \quad \text{where} \quad b_I = C_{I,I}^{-1} \operatorname{sign}(x_{0,I}), \tag{10.23}$$

so that η_F is a linear combination of I basis functions $(\mathcal{C}(x, z_i))_{i \in I}$.

The question is wether $\|\eta_F\|_{\ell^\infty} \leqslant 1$. If the gris is fine enough, i.e. N large enough, this can only hold if $\|\tilde{\eta}_F\|_{L^\infty} \leqslant 1$. The major issue is that $\tilde{\eta}_F$ is only constrained by construction to interpolate $\operatorname{sign}(x_{0,i})$ are points $z_{0,i}$ for $i \in I$. So nothing prevents $\tilde{\eta}_F$ to go outside $[-1, 1]$ around each interpolation point. Figure 10.11 illustrates this fact.

In order to guarantee this property of "local" non-degeneracy around the support, one has to impose on the certificate the additional constraint $\eta'(z_i) = 0$ for $i \in I$. This leads to consider a minimum pre-certificate with vanishing derivatives

$$\eta_V \overset{\text{def.}}{=} A^* p_V \quad \text{where} \quad p_V \in \operatorname*{argmin}_{p \in L^2(\mathbb{R})} \left\{ \|p\|_{L^2(\mathbb{R})} \;;\; \tilde{\eta}(z_I) = \operatorname{sign}(x_{0,I}), \tilde{\eta}'(z_I) = \mathbf{0}_I \right\}. \tag{10.24}$$

where we denoted $\tilde{\eta} = \bar{\psi} \star p$. Similarly to (10.23), this vanishing pre-certificate can be written as a linear combination, but this time of $2|I|$ basis functions

$$\tilde{\eta}_V(x) = \sum_{i \in I} b_i \mathcal{C}(x, z_i) + c_i \partial_2 \mathcal{C}(x, z_i),$$

where $\partial_2 \mathcal{C}$ is the derivative of $\mathcal{C}$ with respect to the second variable, and (b, c) are solution of a $2|I| \times 2|I|$ linear system

$$\begin{pmatrix} b \\ c \end{pmatrix} = \begin{pmatrix} (\mathcal{C}(x_i, x_{i'}))_{i,i' \in I^2} & (\partial_2 \mathcal{C}(x_i, x_{i'}))_{i,i' \in I^2} \\ (\partial_1 \mathcal{C}(x_i, x_{i'}))_{i,i' \in I^2} & (\partial_1 \partial_2 \mathcal{C}(x_i, x_{i'}))_{i,i' \in I^2} \end{pmatrix}^{-1} \begin{pmatrix} \operatorname{sign}(x_{0,I}) \\ \mathbf{0}_I \end{pmatrix}.$$

The associated continuous pre-certificate is $\tilde{\eta}_V = \bar{\psi} \star p_V$, and η_V is a sampling on the grid of $\tilde{\eta}_V$. Figure 10.9 shows that this pre-certificate η_V is much better behaved than η_F. If $\|\eta_V\|_\infty \leqslant 1$, one can apply (13) and thus obtain a linear convergence rate with respect to the ℓ^2 norm on the grid. But for very fine grid, since one is interested in sparse solution, the ℓ^2 norm becomes meaningless (because the L^2 norm is not defined on measures). Since η_V is different from η_F, one cannot directly applies Theorem 14: the support is not stable on discrete grids, which is a fundamental property of super-resolution problems (as opposed to compressed sensing problems). The way to recover interesting results is to use and analyze methods without grids. Indeed, after removing the grid, one can show that η_V becomes the minimum norm certificate (and is the limit of η_λ).

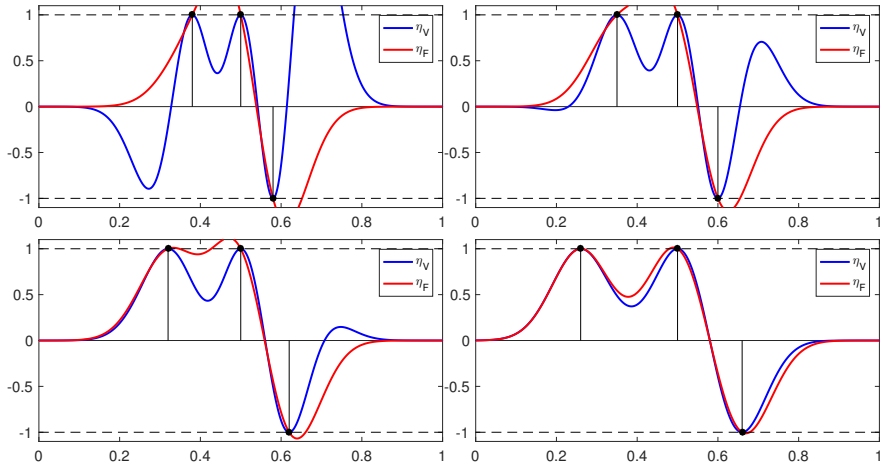

Figure 10.11: Display of "continuous" certificate η_F and η_V for A being a convolution operator.

Bibliography

[1] 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.

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

[3] 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.

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

[5] 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.

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

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

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

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

[10] 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.

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

[12] 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.

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

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

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

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