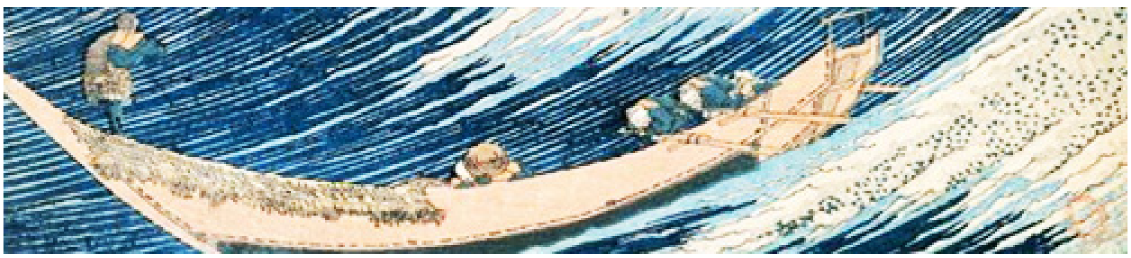
Mathematical Foundations of Data Sciences



Gabriel Peyré
CNRS & DMA
École Normale Supérieure
gabriel.peyre@ens.fr
www.gpeyre.com
www.numerical-tours.com

October 19, 2017

Chapter 13

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.

13.1 Existence and Uniqueness

13.1.1 Existence

We consider problems (11.9) and (11.10), that we rewrite here as

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

and its limit as $\lambda\to 0$

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

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

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.

13.1.2 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 $A_I \stackrel{\text{def.}}{=} (a_i)_{i\in I}\in\mathbb{R}^{P\times|I|}$ the extracted sub-matrix. Similarly, for $x\in\mathbb{R}^N$, we denote $x_I \stackrel{\text{def.}}{=} (x_i)_{i\in I}\in\mathbb{R}^{|I|}$.

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

Proposition 26. *x_λ is a solution to $(\mathcal{P}_\lambda(y))$ for $\lambda>0$ if and only if*

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

where we define

$$I \overset{\text{def.}}{=} \operatorname{supp}(x_\lambda) \overset{\text{def.}}{=} \{i \;;\; x_{\lambda,i} \neq 0\},$$

$$\textit{and} \quad \eta_\lambda \overset{\text{def.}}{=} \frac{1}{\lambda} A^*(y - Ax_\lambda). \tag{13.1}$$

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. □

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

$$\exists \eta \in \mathcal{D}_0(y, x^\star) \overset{\text{def.}}{=} \operatorname{Im}(A^*) \cap \partial \| \cdot \|_1(x^\star). \tag{13.2}$$

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. $\ker(A)^\perp = \operatorname{Im}(A^*)$. □

Writing $I = \operatorname{supp}(x^\star)$, one thus has

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

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).

13.1.3 Uniqueness

The following proposition shows that the Lasso selects a set of linearly independent regressor.

Proposition 28. *There is always a solution x_λ to ($\mathcal{P}_\lambda(y)$) with $I = \operatorname{supp}(x_\lambda)$ such that* $\ker(A_I) = \{0\}$

Proof. TODO. □

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

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

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

Proposition 29. *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*

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

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

Proposition 30. *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 $J \overset{\text{def.}}{=} \operatorname{sat}(\eta)$ then $x^\star$ is the unique solution of ($\mathcal{P}_0(y)$).*

13.2 Consistency and Sparsitency

13.3 Consistency and Convergence Rates

The following theorem studies the convergence rate of the sparse regularized solution, under the same hypothesis has Proposition 30 (with $x^\star = x_0$).

Theorem 32. *If there exists*

$$\eta \in \mathcal{D}_0(Ax_0, x_0) \tag{13.3}$$

and $\ker(A_J) = \{0\}$ *where* $J \overset{\text{def.}}{=} \operatorname{sat}(\eta)$ *then there exists* (C, C') *such that choosing* $\lambda = C'\|w\|$ *for any solution* x_λ *of* $\mathcal{P}(Ax_0 + w)$ *satisfy*

$$\|x_\lambda - x_0\| \leqslant C\|w\|. \tag{13.4}$$

Note that this theorem does not imply that x_λ is a unique solution, only x_0 is unique in general. The condition (13.3) 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 (13.4). The result (13.4) 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 31 on linear methods for inverse problem regularization, which only gives sub-linear rate. The sources conditions in the linear (10.12) and non-linear (13.3) cases are however very different. In the linear case, for $\beta = 1/2$, it reads $x_0 \in \operatorname{Im}(A^*) = \ker(A)^\perp$, 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 $\operatorname{Im}(A^*)$, and is able (in the favorable cases of course) to recover information in $\ker(A)$.

13.3.1 Sparsistency

Theorem 32 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 (13.3). We now give a powerful "recipe" which – when it works – not only give a sufficient condition for linear rate, but also provides "support stability".

There are several ways to detail this construction. One way is to consider, for any solution x_λ of $(\mathcal{P}_\lambda(y))$,

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

Assuming $y = Ax_0$ where x_0 is a solution to $(\mathcal{P}_\lambda(y = Ax_0))$, one can show that

$$p_\lambda \to p_0 \overset{\text{def.}}{=} \underset{p \in \mathbb{R}^P}{\operatorname{argmin}} \ \{\|p\| \ ; \ A^* p \in \mathcal{D}_0(y, x_0)\}\,. \tag{13.5}$$

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

A major difficulty in computing (13.14) 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"

$$\eta_F \overset{\text{def.}}{=} A^* p_F \quad \text{where} \quad p_F \overset{\text{def.}}{=} \underset{p \in \mathbb{R}^P}{\operatorname{argmin}} \ \{\|p\| \ ; \ A_I^* p = \operatorname{sign}(x_{0,I})\}\,. \tag{13.6}$$

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 (21)). 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 $C \overset{\text{def.}}{=} A^* A$ the "correlation" matrix, one has the nice formula

$$\eta_F = C_{\cdot,I} C_{I,I}^{-1} \operatorname{sign}(x_{0,I}). \tag{13.7}$$

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

Proposition 31. *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.

Remark 7 (Operator norm). In the proof, we use the $\ell^p - \ell^q$ matrix operator norm, which is defined as

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

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 33. *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*

$$x_{\lambda,I} = x_{0,I} + A_I^+ w - \lambda(A_I^* A_I)^{-1}\operatorname{sign}(x_{0,I}^\star). \tag{13.8}$$

In particular, $\|x_\lambda - x_0\| = O(\|w\|)$.

Proof. In the following we denote $T \stackrel{\text{def.}}{=} \min_{i\in I}|x_{0,i}|$ the signal level, and $\delta \stackrel{\text{def.}}{=} \|A^* w\|_\infty$ which is the natural way to measure the noise amplitude in the sparse setting. We define $s \stackrel{\text{def.}}{=} \operatorname{sign}(x_0)$, and consider the "ansatz" (13.8) and thus define the following candidate solution

$$\hat{x}_I \stackrel{\text{def.}}{=} x_{0,I} + A_I^+ w - \lambda(A_I^* A_I)^{-1} s_I, \tag{13.9}$$

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 implies by

$$\|x_{0,I} - \hat{x}_I\|_\infty \leqslant K\|A_I^* w\|_\infty + K\lambda < T \quad \text{where} \quad K \stackrel{\text{def.}}{=} \|(A_I^* A_I)^{-1}\|_\infty, \tag{13.10}$$

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 29, 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

$$\|A_{I^c}^* A_I (A_I^* A_I)^{-1}\|_\infty \|A_I^* w\|_\infty + \|A_{I^c}^* w\|_\infty + \lambda\|\eta_{F,I^c}\|_\infty \leqslant R\|A_I^* w\|_\infty - S\lambda < 0 \tag{13.11}$$

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

where we denoted $L \stackrel{\text{def.}}{=} \|A_{I^c}^* A_I\|_\infty$, and also we used the hypothesis $\|\eta_{F,I^c}\|_\infty < 1$.

Conclusion. Putting (13.10) and (13.11) 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

$$\tilde{\mathcal{R}} = \left\{(\delta, \lambda) \;;\; \frac{\delta}{\lambda} < \frac{S}{R} \quad \text{and} \quad \lambda < \lambda_{\max}\right\} \quad \text{where} \quad \lambda_{\max} \stackrel{\text{def.}}{=} \frac{T(KL+1)}{K(R+S)}. \tag{13.12}$$

□

A nice feature of this proof is that it gives access to explicit constant, involving the three key parameter K, L, S, which controls:
– K accounts for the contininoing 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 (13.12). 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)

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

13.4 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.

Another interesting case study, which shows the limitation of this approach, is the case of super-resolution. To simplify the analysis, we assume "continuous measurement", and replace the measurement space $\mathbb{R}^P$ by functions $L^2(\mathbb{R}^d)$ (for simplicity, we treat here $d = 1$). The measurements reads

$$Ax = \sum_{i=1}^N x_i a_i(\cdot) \in L^2(\mathbb{R})$$

where the $a_i : \mathbb{R} \to \mathbb{R}$ are smooth functions. A typical example is the deconvolution problem, where $a_i = \varphi(\cdot - z_i)$ where $\varphi : \mathbb{R} \to \mathbb{R}$ is the smoothing kernel and $(z_i)_{i=1}^N$ is a discretization grid, and for simplicity we assume $x_i = i/N \in [0, 1]$.

In this case, this forward model correspond to convolution of measures supported on the grid

$$Ax = \varphi \star m_{z,x} \quad \text{where} \quad m_{z,x} \stackrel{\text{def.}}{=} \sum_{i=1}^N x_i \delta_{z_i}.$$

The dual pre-certificate $\eta_F(t)$ is thus a function defined on $\mathbb{R}$.

We denote the "continuous" covariance

$$\mathcal{C}(z, z') \stackrel{\text{def.}}{=} \langle \varphi(\cdot - z),\, \varphi(\cdot - z') \rangle_{L^2(\mathbb{R})} = \int_{\mathbb{R}} \varphi(t - z)\varphi(t - z)\mathrm{d}t = (\varphi \star \bar{\varphi})(z - z')$$

where $\bar{\varphi}(t) = \varphi(-t)$, so that the discrete covariance is $C = (\mathcal{C}(z_i, z_i')))_{(i,i')} \in \mathbb{R}^{N\times N}$ and $C_{I,I} = (\mathcal{C}(z_i, z_i')))_{(i,i')\in I^2} \in \mathbb{R}^{I\times I}$.

Using (13.7), one sees that η_F is obtained as a sampling on the grid of a "continuous' ' certificate $\tilde{\eta}_F$

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

$$\text{where} \quad \tilde{\eta}_F(x) = \sum_{i\in I} b_i \mathcal{C}(x, z_i) \quad \text{where} \quad b_I = C_{I,I}^{-1} \operatorname{sign}(x_{0,I}), \tag{13.13}$$

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 **??** 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

$$\eta_V \stackrel{\text{def.}}{=} A^* p_V \quad \text{where} \quad p_V \underset{p \in L^2(\mathbb{R})}{\operatorname{argmin}} \left\{ \|p\|_{L^2(\mathbb{R})} \; ; \; \tilde{\eta}(z_I) = \operatorname{sign}(x_{0,I}), \tilde{\eta}'(z_I) = \mathbf{0}_I \right\}. \tag{13.14}$$

where we denoted $\tilde{\eta} = \bar{\psi} \star p$. Similarly to (13.13), 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 **??** shows that this pre-certificate η_V is much better behaved than η_F. If $\|\eta_V\|_\infty \leqslant 1$, one can apply (32) 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 33: 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).

Bibliography

[1] P. Alliez and C. Gotsman. Recent advances in compression of 3d meshes. In N. A. Dodgson, M. S. Floater, and M. A. Sabin, editors, *Advances in multiresolution for geometric modelling*, pages 3–26. Springer Verlag, 2005.

[2] P. Alliez, G. Ucelli, C. Gotsman, and M. Attene. Recent advances in remeshing of surfaces. In *AIM@SHAPE repport*. 2005.

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

[4] 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.

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

[6] 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.

[7] E. J. Candès. The restricted isometry property and its implications for compressed sensing. *Compte Rendus de l'Académie des Sciences*, Serie I(346):589–592, 2006.

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

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

[10] 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.

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

[12] 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.

[13] F. R. K. Chung. Spectral graph theory. *Regional Conference Series in Mathematics, American Mathematical Society*, 92:1–212, 1997.

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

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

[16] P. Schroeder et al. D. Zorin. Subdivision surfaces in character animation. In *Course notes at SIGGRAPH 2000*, July 2000.

[17] 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.

[18] I. Daubechies and W. Sweldens. Factoring wavelet transforms into lifting steps. *J. Fourier Anal. Appl.*, 4(3):245–267, 1998.

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

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

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

[22] M. S. Floater and K. Hormann. Surface parameterization: a tutorial and survey. In N. A. Dodgson, M. S. Floater, and M. A. Sabin, editors, *Advances in multiresolution for geometric modelling*, pages 157–186. Springer Verlag, 2005.

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

[24] I. Guskov, W. Sweldens, and P. Schröder. Multiresolution signal processing for meshes. In Alyn Rockwood, editor, *Proceedings of the Conference on Computer Graphics (Siggraph99)*, pages 325–334. ACM Press, August8–13 1999.

[25] A. Khodakovsky, P. Schröder, and W. Sweldens. Progressive geometry compression. In *Proceedings of the Computer Graphics Conference 2000 (SIGGRAPH-00)*, pages 271–278, New York, July 23–28 2000. ACMPress.

[26] L. Kobbelt. $\sqrt{3}$ subdivision. In Sheila Hoffmeyer, editor, *Proc. of SIGGRAPH'00*, pages 103–112, New York, July 23–28 2000. ACMPress.

[27] M. Lounsbery, T. D. DeRose, and J. Warren. Multiresolution analysis for surfaces of arbitrary topological type. *ACM Trans. Graph.*, 16(1):34–73, 1997.

[28] S. Mallat. *A Wavelet Tour of Signal Processing, 3rd edition.* Academic Press, San Diego, 2009.

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

[30] 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.

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

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

[33] Gabriel Peyré and Marco Cuturi. Computational optimal transport. 2017.

[34] 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.

[35] E. Praun and H. Hoppe. Spherical parametrization and remeshing. *ACM Transactions on Graphics*, 22(3):340–349, July 2003.

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

[37] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 2015.

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

[39] P. Schröder and W. Sweldens. Spherical Wavelets: Efficiently Representing Functions on the Sphere. In *Proc. of SIGGRAPH 95*, pages 161–172, 1995.

[40] P. Schröder and W. Sweldens. Spherical wavelets: Texture processing. In P. Hanrahan and W. Purgathofer, editors, *Rendering Techniques '95.* Springer Verlag, Wien, New York, August 1995.

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

[42] A. Sheffer, E. Praun, and K. Rose. Mesh parameterization methods and their applications. *Found. Trends. Comput. Graph. Vis.*, 2(2):105–171, 2006.

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

[44] W. Sweldens. The lifting scheme: A custom-design construction of biorthogonal wavelets. *Applied and Computation Harmonic Analysis*, 3(2):186–200, 1996.

[45] W. Sweldens. The lifting scheme: A construction of second generation wavelets. *SIAM J. Math. Anal.*, 29(2):511–546, 1997.