Inverse Problems and Imaging
Volume 1, No. 4, 2007, 661–677 Web site: http://www.aimSciences.org

ANALYTICAL BOUNDS ON THE MINIMIZERS OF (NONCONVEX) REGULARIZED LEAST-SQUARES

Mila Nikolova
CMLA, ENS Cachan, CNRS, PRES UniverSud
61 Av. President Wilson, F-94230 Cachan, FRANCE

(Communicated by Jackie (Jianhong) Shen)

Abstract. This is a theoretical study on the minimizers of cost-functions composed of an ℓ_2 data-fidelity term and a possibly nonsmooth or nonconvex regularization term acting on the differences or the discrete gradients of the image or the signal to restore. More precisely, we derive general *nonasymptotic* analytical bounds characterizing the local and the global minimizers of these cost-functions. We first derive bounds that compare the restored data with the noisy data. For edge-preserving regularization, we exhibit a tight data-independent bound on the ℓ_∞ norm of the residual (the estimate of the noise), even if its ℓ_2 norm is being minimized. Then we focus on the smoothing incurred by the (local) minimizers in terms of the differences or the discrete gradient of the restored image (or signal).

1. **Introduction.** We consider the classical inverse problem of the finding of an estimate $\hat{x} \in \mathbb{R}^p$ of an unknown image or signal $x \in \mathbb{R}^p$ based on data $y \in \mathbb{R}^q$ corresponding to $y = Ax + n$, where $A \in \mathbb{R}^{q \times p}$ models the data-acquisition system and n accounts for the noise. For instance, A can be a point spread function accounting for optical blurring, a distortion wavelet in seismic imaging and non-destructive evaluation, a Radon transform in X-ray tomography, a Fourier transform in diffraction tomography, or it can be the identity in denoising and segmentation problems. Such problems are customarily solved using regularized least-squares methods: the solution $\hat{x} \in \mathbb{R}^p$ minimizes a cost-function $\mathcal{F}_y : \mathbb{R}^p \to \mathbb{R}$ of the form

$$\mathcal{F}_y(x) = \|Ax - y\|^2 + \beta\Phi(x) \tag{1}$$

where Φ is the regularization term and $\beta > 0$ is a parameter which controls the trade-off between the fidelity to data and the regularization [3, 7, 2]. The role of Φ is to push $\hat{x}$ to exhibit some a priori expected features, such as the presence of edges and smooth regions. Since [3, 9], a useful class of regularization functions is

$$\Phi(x) = \sum_{i \in I} \varphi(\|G_i x\|), \; I = \{1, \dots, r\}, \tag{2}$$

where $G_i \in \mathbb{R}^{s \times p}$, $i \in I$, are linear operators with $s \geq 1$, $\|.\|$ is the ℓ_2 norm and $\varphi : \mathbb{R}_+ \to \mathbb{R}$ is called a potential function. When x is an image, the usual choices for $\{G_i : i \in I\}$ are either that $\{G_i\}$ correspond to the discrete approximation of the gradient operator with $s = 2$, or that $\{G_i x\}$ are the first-order differences between each pixel and its 4 or 8 nearest neighbors along with $s = 1$. In the following, the

2000 *Mathematics Subject Classification.* Primary: 47J20, 47A52, 47H99, 68U10; Secondary: 90C26, 49J52, 41A25, 26B10.
Key words and phrases. Image restoration, Signal restoration, Regularization, Variational methods, Edge restoration, Inverse problems, Non-convex analysis, Non-smooth analysis.

Convex PFs					
$\varphi(\lvert t\rvert)$ is smooth at zero			$\varphi(\lvert t\rvert)$ is nonsmooth at zero		
(f1)	$\varphi(t)=t^\alpha,\ 1<\alpha\le 2$	[5]	(f3)	$\varphi(t)=t$	[3, 22]
(f2)	$\varphi(t)=\sqrt{\alpha+t^2}$	[25]			
Nonconvex PFs					
$\varphi(\lvert t\rvert)$ is smooth at zero			$\varphi(\lvert t\rvert)$ is nonsmooth at zero		
(f4)	$\varphi(t)=\min\{\alpha t^2,1\}$	[17, 4]	(f8)	$\varphi(t)=t^\alpha,\ 0<\alpha<1$	[23]
(f5)	$\varphi(t)=\dfrac{\alpha t^2}{1+\alpha t^2}$	[11]	(f9)	$\varphi(t)=\dfrac{\alpha t}{1+\alpha t}$	[10]
(f6)	$\varphi(t)=\log(\alpha t^2+1)$	[12]	(f10)	$\varphi(t)=\log(\alpha t+1)$	
(f7)	$\varphi(t)=1-\exp(-\alpha t^2)$	[14, 21]	(f11)	$\varphi(0)=0,\ \varphi(t)=1$ if $t\neq 0$	[14]

TABLE 1. Commonly used PFs φ where $\alpha>0$ is a parameter.

letter G will denote the $rs\times p$ matrix obtained by vertical concatenation of the matrices G_i for $i\in I$, i.e. $G=[G_1^T,G_2^T,\dots,G_r^T]^T$ where ^{T} means transposed. A basic requirement to have regularization is

$$\ker(A)\cap\ker(G)=\{0\}. \tag{3}$$

Notice that (3) is trivially satisfied when A^TA is invertible (i.e. $\operatorname{rank}(A)=p$). In most of the cases $\ker(G)=\operatorname{span}\{\mathbb{1}\}$, where $\mathbb{1}$ is the p-length vector composed of ones, whereas usually $A\mathbb{1}\neq 0$, so (3) holds again. Many different potential functions (PFs) have been used in the literature. The most popular PFs are given in Table 1. Although PFs differ in convexity, boundedness, differentiability, etc., they share some common features. Based on them, we systematically assume the following:

H1. φ increases on $\mathbb{R}_+$ so that $\varphi(0)=0$ and $\varphi(t)>0$ for any $t>0$.

According to the smoothness of $t\to\varphi(|t|)$ at zero, we will consider either H2 or H3:

H2. φ is $\mathcal{C}^1$ on $\mathbb{R}_+\setminus\mathcal{T}$ where the set $\mathcal{T}=\{t>0:\varphi'(t^-)>\varphi'(t^+)\}$ is at most finite and $\varphi'(0)=0$.

The conditions on $\mathcal{T}$ in this assumption allows us to address the PF given in (f4) which corresponds to the discrete version of the Mumford-Shah functional. The alternative assumption is

H3. φ is $\mathcal{C}^1$ on $\mathbb{R}_+$ and $\varphi'(0)>0$.

Notice that our assumptions address convex and nonconvex functions φ. A particular attention is devoted to *edge-preserving* functions φ because of their ability to give rise to solutions $\hat{x}$ involving sharp edges and homogeneous regions. Based on various conditions for edge-preservation in the literature [10, 22, 5, 15, 2, 20], a common requirement is that for t large, φ is upper bounded by a (nearly) affine function.

The aim of this paper is to give *nonasymptotic* analytical bounds on the local and the global minimizers $\hat{x}$ of $\mathcal{F}_y$ in (1)-(2) that hold for all functions φ described above. To our knowledge, related questions have mainly been considered in particular situations, such as A the identity, or a particular φ, or when y is a special noise-free function, or in the context of the regularization of wavelet coefficients, or in asymptotic conditions when one of the terms in (1) vanishes—let us cite among others [24, 16, 1]. An outstanding paper [8] explores the mean and the variance of the minimizers $\hat{x}$ for strictly convex and differentiable functions φ. When φ is nonconvex, $\mathcal{F}_y$ may have numerous local minima and it is crucial to have reliable bounds on its local and global minimizers. The bounds we provide can be of practical interest for the initialization and the convergence analysis of numerical schemes. This paper constitutes a continuation of a previous work on the properties of the minimizers relevant to non-convex regularization [20].

Content of the paper. The focus in section (2) is on restored data $A\hat{x}$ in comparison with noisy data y. More precisely, their norms are compared as well as their means. The cases of smooth and nonsmooth regularization are analyzed separately. Section 3 is devoted to the residual $A\hat{x} - y$ which can be seen as an estimate of the noise. Tight upper bounds on the ℓ_∞ norm that are independent of the data are derived in the wide context of edge-preserving regularization. Restored differences or gradients are compared to those of the least-squares solution in section 4. Concluding remarks are given in section 5.

2. **Bounds on the restored data.** In this section we compare the restored data $A\hat{x}$ with the noisy data y. Even though the statements corresponding to $\varphi'(0) = 0$ (H2) and to $\varphi'(0) > 0$ (H3) are similar, the proofs in the latter case are quite different and more intricate. These cases are considered separately.

2.1. **Smooth regularization.** Before to get into the heart of the matter, we restate below a result from [19] saying that even if φ is non-smooth in the sense specified in H2, the function $\mathcal{F}_y$ is smooth at every one of its local minimizers.

Proposition 1. *Let φ satisfy H1 and H2 where $\mathcal{T} \neq \emptyset$. If $\hat{x}$ is a (local) minimizer of $\mathcal{F}_y$, we have $\|G_i\hat{x}\| \neq \tau$, for all $i \in I$, for every $\tau \in \mathcal{T}$. Moreover, $\hat{x}$ satisfies $\nabla\mathcal{F}_y(\hat{x}) = 0$.*

The first statement below corroborates quite an intuitive result. The second statement provides a strict inequality and addresses first-order difference operators or gradients $\{G_i, i \in I\}$ in which case $\ker(G) = \mathrm{span}(\mathbb{1})$. For simplicity, we systematically write $\|.\|$ in place of $\|.\|_2$ for the ℓ_2 norm of vectors.

Theorem 2.1. *Let $\mathcal{F}_y : \mathbb{R}^p \to \mathbb{R}$ be of the form (1)-(2) where φ satisfy H1 and H2.*

(i) Suppose that $\mathrm{rank}(A) = p$ *or* $\left\{ \begin{array}{l} \varphi \textit{ is strictly increasing on } \mathbb{R}_+ \\ \textit{and (3) holds.} \end{array} \right.$

For every $y \in \mathbb{R}^q$, if $\mathcal{F}_y$ reaches a (local) minimum at $\hat{x} \in \mathbb{R}^p$, then

$$\|A\hat{x}\| \leq \|y\| \,. \tag{4}$$

(ii) Assume that $\mathrm{rank}(A) = p \geq 2$, $\ker(G) = \mathrm{span}(\mathbb{1})$ *and φ is strictly increasing on $\mathbb{R}_+$.*

There is a closed subset of Lebesgue measure zero in $\mathbb{R}^q$, denoted N, such that for every $y \in \mathbb{R}^q \setminus N$, if $\mathcal{F}_y$ has a (local) minimum at $\hat{x} \in \mathbb{R}^p$, then

$$\|A\hat{x}\| < \|y\| \,. \tag{5}$$

Hence (5) is satisfied for almost every $y \in \mathbb{R}^q$.

If A is orthonormal (e.g. A is the identity), (4) leads to $\|\hat{x}\| \leq \|y\|$. When A is the identity and x is defined on a convex subset of $\mathbb{R}^d$, $d \geq 1$, it is shown in [2] that $\|\hat{x}\| \leq \sqrt{2}\|y\|$. So (4) provides a bound which is sharper for images on discrete grids and holds for general regularized cost-functions.

The proof of the theorem relies on the lemma stated below whose proof is outlined in the appendix. Given $\theta \in \mathbb{R}^r$, we write $\mathrm{diag}(\theta)$ to denote the diagonal matrix whose main diagonal is θ.

Lemma 2.2. *Let $A \in \mathbb{R}^{q\times p}$, $G \in \mathbb{R}^{r\times p}$ and $\theta \in \mathbb{R}^r_+$. Assume that if* $\mathrm{rank}(A) < p$, *then (3) holds and that $\theta[i] > 0$, for all $1 \le i \le r$. Consider the $q \times q$ matrix C below*

$$C = A\Big(A^TA + G^T\mathrm{diag}(\theta)G\Big)^{-1}A^T. \tag{6}$$

(i) The matrix C is well defined and its spectral norm satisfies $\|\!|C|\!\|_2 \le 1$.

(ii) Suppose that $\mathrm{rank}(A) = p$, *that* $\ker(G) = \mathrm{span}(h)$ *for a nonzero $h \in \mathbb{R}^p$ and that $\theta[i] > 0$ for all $i \in \{1, \dots, r\}$. Then*

$$\|Cy\| < \|y\|, \ \forall y \in \big\{ \ker(A^T) \cup V_h \big\}, \tag{7}$$

where V_h is a vector subspace of $\mathbb{R}^q$ of dimension $q - p + 1$ and reads

$$V_h = \Big\{ y \in \mathbb{R}^q : A^Ty \propto A^TAh \Big\}. \tag{8}$$

Let us remind $\|\!|C|\!\|_2 = \max\{\sqrt{\lambda} : \lambda \text{ is an eigenvalue of } C^TC\} = \sup_{\|u\|=1} \big\|Cu\big\|$. Since C in (6) is symmetric and positive semi-definite, all its eigenvalues are hence contained in $[0, 1]$.

Remark 1. With the notations of Lemma 2.2, the set N in Theorem 2.1 (ii) reads

$$N = \{V_{\mathbb{1}} \cup \ker(A^T)\}$$

since $\ker(G) = \mathrm{span}(\mathbb{1})$.

Proof of Theorem 2.1. If the set $\mathcal{T}$ in H2 is nonempty, Proposition 1 tells us that any minimizer $\hat{x}$ of $\mathcal{F}_y$ satisfies $\|G_i\hat{x}\| \notin \mathcal{T}$, $1 \le i \le r$, in which case $\nabla\Phi$ is well defined at $\hat{x}$. In all cases addressed by H1 and H2, any minimizer $\hat{x}$ of $\mathcal{F}_y$ satisfies $\nabla\mathcal{F}_y(\hat{x}) = 0$ where

$$\nabla\mathcal{F}_y(x) = 2A^TAx + \beta\nabla\Phi(x) - 2A^Ty. \tag{9}$$

For every $i \in I$, the entries of $G_i \in \mathbb{R}^{s\times p}$ are $G_i[j, n]$, for $1 \le j \le s$ and $1 \le n \le p$, so its nth column is $G_i[\cdot, n]$ and its jth row is $G_i[j, \cdot]$. The ith component of a given vector x is denoted $x[i]$.

Since $\|G_ix\| = \sqrt{\sum_{j=1}^s (G_i[j, \cdot]x)^2}$, the entries $\partial_n\Phi(x) = \partial\Phi(x)/\partial x[n]$ of $\nabla\Phi(x)$ read

$$\partial_n\Phi(x) = \sum_{i\in I} \frac{\varphi'(\|G_ix\|)}{\|G_ix\|} \sum_{j=1}^{s} G_i[j, \cdot]x \; G_i[j, n] = \sum_{i\in I} \big(G_i[\cdot, n]\big)^T \frac{\varphi'(\|G_ix\|)}{\|G_ix\|} G_ix. \tag{10}$$

In the case when $s = 1$, we have $G_i^1 = G_i$ and (10) is simplified to

$$\partial_n\Phi(x) = \sum_{i\in I} \varphi'(|G_ix|)\mathrm{sign}(G_ix) \; G_i[1, n]$$

where we can set $\mathrm{sign}(0)$ arbitrarily. Let $\theta \in \mathbb{R}^{rs}$ be defined as a function of $\hat{x}$ as

$$\forall i = 1, \dots, r,$$

$$\theta\big[(i-1)s + 1\big] = \ \cdots \ = \theta\big[is\big] = \begin{cases} \dfrac{\varphi'(\|G_i\hat{x}\|)}{\|G_i\hat{x}\|} & \text{if } \ \|G_i\hat{x}\| \ne 0, \\ 1 & \text{if } \ \|G_i\hat{x}\| = 0. \end{cases} \tag{11}$$

Since φ is increasing on $\mathbb{R}_+$, (11) shows that $\theta[i] \geq 0$ for every $1 \leq i \leq rs$. Introducing θ into (10) leads to

$$\nabla\Phi(\hat{x}) = G^T \mathrm{diag}(\theta) G\hat{x}. \tag{12}$$

Introducing the latter expression into (9) yields

$$\Big(A^T A + \frac{\beta}{2} G^T \mathrm{diag}(\theta) G\Big)\hat{x} = A^T y. \tag{13}$$

Noticing that the requirements for Lemma 2.2(i) are satisfied, we can write that

$$A\hat{x} = Cy, \tag{14}$$

where C is the matrix given in (6). By Lemma 2.2(i), $\|A\hat{x}\| \leq |\!|\!| C |\!|\!|_2 \, \|y\| \leq \|y\|$.

Applying Lemma 2.2(ii) with $h = \mathbb{1}$, we find that the inequality is strict if $y \notin N$ where $N = \{V_{\mathbb{1}} \cup \ker(A^T)\}$. Hence statement (ii) of the theorem. $\square$

2.2. **Nonsmooth regularization.** Now we focus on functions φ such that $t \to \varphi(|t|)$ is nonsmooth at $t = 0$. It is well known that if $\varphi'(0) > 0$ in (2), the minimizers $\hat{x}$ of $\mathcal{F}_y$ typically satisfy $G_i\hat{x} = 0$ for a certain (possibly large) subset of indexes $i \in I$ [18, 19].

Theorem 2.3. *Let $\mathcal{F}_y : \mathbb{R}^p \to \mathbb{R}$ be of the form (1)-(2) where φ satisfy H1 and H3.*

(i) Suppose that $\mathrm{rank}(A) = p$ *or* $\left\{\begin{array}{l}\varphi \textit{ is strictly increasing on } \mathbb{R}_+\\ \textit{and (3) holds.}\end{array}\right.$

For every $y \in \mathbb{R}^q$, if $\mathcal{F}_y$ reaches a (local) minimum at $\hat{x} \in \mathbb{R}^p$, then

$$\|A\hat{x}\| \leq \|y\|\,. \tag{15}$$

(ii) Assume that $\mathrm{rank}(A) = p \geq 2$, *that* $\ker(G) = \mathrm{span}(\mathbb{1})$ *and that φ is strictly increasing on $\mathbb{R}_+$.*

There is a closed subset of Lebesgue measure zero in $\mathbb{R}^q$, denoted N, such that for every $y \in \mathbb{R}^q \setminus N$, if $\mathcal{F}_y$ has a (local) minimum at $\hat{x} \in \mathbb{R}^p$, then

$$\|A\hat{x}\| < \|y\|\,. \tag{16}$$

Hence (16) holds for almost every $y \in \mathbb{R}^q$.

Proof.

Statement (i). Given $\hat{x}$, let us introduce the subsets

$$J = \{i \in I : \|G_i\hat{x}\| = 0\} \text{ and } J^c = I \setminus J, \tag{17}$$

where I was introduced in (2). Since $\mathcal{F}_y$ has a (local) minimum at $\hat{x}$, for every $v \in \mathbb{R}^p$, the one-sided derivative of $\mathcal{F}_y$ at $\hat{x}$ in the direction of v,

$$\delta\mathcal{F}_y(\hat{x})(v) = 2v^T A^T (A\hat{x} - y) + \beta \sum_{i \in J^c} \frac{\varphi'(\|G_i\hat{x}\|)}{\|G_i\hat{x}\|} (G_i\hat{x})^T \, G_i v + \beta\varphi'(0) \sum_{i \in J} \|G_i v\|, \tag{18}$$

satisfies $\delta\mathcal{F}_y(\hat{x})(v) \geq 0$. Let $\mathcal{K}_J \subset \mathbb{R}^p$ be the subspace given below

$$\mathcal{K}_J = \{v \in \mathbb{R}^p : G_i v = 0, \ \forall i \in J\}. \tag{19}$$

In particular, if J is empty then $\mathcal{K}_J = \mathbb{R}^p$ and the sums over J are absent, while if $J = I$, we have $\mathcal{K}_J = \ker(G)$ and the sums over J^c are absent. Notice also that if $\ker(G) = \{0\}$, having a minimizer $\hat{x}$ such that $J = I$ in (17) means that $\hat{x} = 0$ in which case (15) is trivially satisfied while (16) holds for $N = \{0\}$. In the following, consider that the dimension d_J of $\mathcal{K}_J$ satisfies $1 \leq d_J \leq p$. Since the last term on the right side of (18) vanishes if $v \in \mathcal{K}_J$, we can write that

$$\forall v \in \mathcal{K}_J,\ v^T \left(A^T A\hat{x} + \frac{\beta}{2} \sum_{i \in J^c} G_i^T \,\frac{\varphi'(\|G_i\hat{x}\|)}{\|G_i\hat{x}\|} G_i\hat{x} - A^T y \right) = 0. \tag{20}$$

Let B_J be a $p \times d_J$ matrix whose columns form an orthonormal basis of $\mathcal{K}_J$. Then (20) is equivalent to

$$B_J^T \left(A^T A\hat{x} + \frac{\beta}{2} \sum_{i \in J^c} G_i^T \,\frac{\varphi'(\|G_i\hat{x}\|)}{\|G_i\hat{x}\|} G_i\hat{x} \right) = B_J^T A^T y. \tag{21}$$

Let $\theta \in \mathbb{R}^{rs}$ be defined as in (11). Using that $\|G_i\hat{x}\| = 0$ for all $i \in J$ we can write that

$$\sum_{i \in J^c} G_i^T \,\frac{\varphi'(\|G_i\hat{x}\|)}{\|G_i\hat{x}\|} G_i\hat{x} = \sum_{i \in J^c} G_i^T \,\theta[is] G_i\hat{x} + \sum_{i \in J} G_i^T \,\theta[is] G_i\hat{x} = G^T \mathrm{diag}(\theta) G\hat{x}.$$

Then (21) is equivalent to

$$B_J^T \left(A^T A + \frac{\beta}{2} G^T \mathrm{diag}(\theta) G \right) \hat{x} = B_J^T A^T y. \tag{22}$$

Since $\hat{x} \in \mathcal{K}_J$, there is a unique $\tilde{x} \in \mathbb{R}^{d_J}$ such that

$$\hat{x} = B_J \tilde{x}. \tag{23}$$

Define

$$A_J = AB_J \in \mathbb{R}^{q \times d_J} \text{ and } G_J = GB_J \in \mathbb{R}^{r \times d_J}. \tag{24}$$

Then (22) reads

$$\left(A_J^T A_J + \frac{\beta}{2} G_J^T \mathrm{diag}(\theta) G_J \right) \tilde{x} = A_J^T y. \tag{25}$$

If $A^T A$ is invertible, $A_J^T A_J \in \mathbb{R}^{d_J \times d_J}$ is invertible as well. Notice that by (3), $\ker(A_J) \cap \ker(G_J) = \{0\}$. Then the matrix between the parentheses of in (25) is invertible. We can hence write down

$$\begin{aligned} A_J \tilde{x} &= C_J y, \\ C_J &= A_J \left(A_J^T A_J + \frac{\beta}{2} G_J^T \mathrm{diag}(\theta) G_J \right)^{-1} A_J^T. \end{aligned} \tag{26}$$

According to Lemma 2.2 (i) we have $|||C_J||| \leq 1$. Using (23), we deduce that $\|A\hat{x}\| = \|A_J\tilde{x}\| \leq \|y\|$.

Statement (ii). Define

$$V_\infty = \Big\{ y \in \mathbb{R}^q : y \propto A\mathbb{1} \Big\}.$$

Consider that for some $y \in \mathbb{R}^q \setminus V_\infty$, there is a (local) minimizer $\hat{x}$ of $\mathcal{F}_y$ such that the subspace $\mathcal{K}_J$ defined by (17) and (19) is of dimension $d_J = 1$. Using that $\ker(G) = \mathbb{1}$, this means that $\mathcal{K}_J = \{\lambda\mathbb{1} : \lambda \in \mathbb{R}\}$ and hence $\hat{x} = \hat{\lambda}\mathbb{1}$ where $\hat{\lambda}$ satisfies $\|A\hat{\lambda}\mathbb{1} - y\|^2 = \min_{\lambda \in \mathbb{R}} \|A\lambda\mathbb{1} - y\|^2$, since $\Phi(\lambda\mathbb{1}) = 0$, $\forall \lambda \in \mathbb{R}$. It is easy to find that $\hat{\lambda} = \frac{y^T A\mathbb{1}}{\|A\mathbb{1}\|^2}$ and that

$$\|A\hat{x}\| = \hat{\lambda}\|A\mathbb{1}\| = \frac{|y^T A\mathbb{1}|}{\|A\mathbb{1}\|}.$$

By Schwarz inequlity, $|y^T A\mathbb{1}| < \|y\| \, \|A\mathbb{1}\|$ since $y \notin V_\infty$. Hence

$$\left.\begin{array}{c} y \in \mathbb{R}^q \setminus V_\infty \\ \mathcal{F}_y \text{ has a (local) minimum at } \hat{x} = \hat{\lambda}\mathbb{1} \end{array}\right\} \Rightarrow \ \|A\hat{x}\| < \|y\|. \tag{27}$$

Let us next define the family of subsets of indexes $\mathcal{J}$ as

$$\mathcal{J} = \{J \subset I \ : \ \dim(\mathcal{K}_J) \geq 2\},$$

where for every J, the set $\mathcal{K}_J$ is defined by (19). For every $J \in \mathcal{J}$, let us denote $d_J = \dim(\mathcal{K}_J) \in [2, p]$ and let the columns of $B_J \in \mathbb{R}^{p \times d_J}$ form an orthonormal basis of $\mathcal{K}_J$. Let A_J, G_J and C_J be defined as in (24) and (26). Notice that $\{\emptyset\} \in \mathcal{J}$ and that $\dim(\mathcal{K}_{\{\emptyset\}}) = p$.

Since for every $J \in \mathcal{J}$ we have $\mathbb{1} \in \mathcal{K}_J$, there exists $h_J \in \mathbb{R}^{d_J}$ such that

$$B_J h_J = \mathbb{1} \in \mathbb{R}^p.$$

Using that $\mathrm{rank}(B_J) = d_J$, this h_J is unique and then $\ker(G_J) = h_J$. For every $J \in \mathcal{J}$ define the subspaces

$$\begin{aligned} V_J &= \Big\{y \in \mathbb{R}^q : A_J^T y \propto A_J^T A_J h_J\Big\}, \\ W_J &= \ker(A_J^T). \end{aligned}$$

Using Lemma 2.2 (ii), for every $J \in \mathcal{J}$ we can write that

$$\left.\begin{array}{l} y \in \mathbb{R}^q \setminus (V_J \cup W_J) \\ \mathcal{F}_y \text{ has a (local) minimum at } \hat{x} \\ \text{ such that } \{i \in I : G_i \hat{x} = 0\} = J \end{array}\right\} \Rightarrow \ \|A\hat{x}\| < \|y\|. \tag{28}$$

Since for every $J \in \mathcal{J}$ we have $\dim(V_J) \leq q-1$ and $\dim(W_J) \leq q-1$, the subset $N \subset \mathbb{R}^q$ defined by

$$N = V_\infty \cup \bigcup_{J \in \mathcal{J}} \Big(V_J \cup W_J\Big)$$

is closed and negligible in $\mathbb{R}^q$ as being a finite union of subspaces of $\mathbb{R}^q$ of dimension strictly smaller than q. Using (27) and (28),

$$\|A\hat{x}\| < \|y\|, \ \forall y \in \mathbb{R}^q \setminus N.$$

The proof is complete. □

2.3. **The mean of restored data.** Under the common assumption that the noise corrupting the data is of mean zero, it is reasonable to require that the restored data $A\hat{x}$ and the observed data y have the same mean. When x is an image defined on a convex subset of $\mathbb{R}^2$ and φ is applied to $\|\nabla x\|$, and when A is the identity, it is well known that the solution $\hat{x}$ and the data y have the same mean [2]. Below we consider this question in our context where x is defined on a discrete grid and A is a general linear operator. For the seek of clarity, the index in $\mathbb{1}_p$ will specify the dimension of $\mathbb{1}_p$.

Proposition 2. *Let $\mathcal{F}_y : \mathbb{R}^p \to \mathbb{R}$ read as in (1)-(2) where φ satisfy H1 combined with one of the assumptions H2 or H3. Assume that $\mathbb{1}_p \in \ker(G)$ and that*

$$A\mathbb{1}_p \propto \mathbb{1}_q. \tag{29}$$

For every $y \in \mathbb{R}^q$, if $\mathcal{F}_y$ has a (local) minimum at $\hat{x}$, then

$$\mathbb{1}^T A\hat{x} = \mathbb{1}^T y. \tag{30}$$

Proof. Consider first that φ satisfies H1-H2. Using that $\mathbb{1}_p^T \nabla \mathcal{F}_y(\hat{x}) = 0$, where $\nabla \mathcal{F}_y$ is given in (9) and (12), we obtain

$$\begin{aligned}(A\mathbb{1}_p)^T A\hat{x} - (A\mathbb{1}_p)^T y &= -\frac{1}{2}(G\mathbb{1}_p)^T \mathrm{diag}(\theta) G\hat{x} \\ &= 0\end{aligned}$$

where the last equality comes from the assumption that $\mathbb{1}_p \in \ker(G)$. Using that $A\mathbb{1}_p = c\mathbb{1}_q$ for some $c \in \mathbb{R}$ leads to (30).

Consider now that φ satisfies H1-H3. From the necessary condition for a minimum, $\delta\mathcal{F}_y(\hat{x})(\mathbb{1}) \geq 0$ and $\delta\mathcal{F}_y(\hat{x})(-\mathbb{1}) \geq 0$, where $\delta\mathcal{F}_y$ is given in (18). Noticing that $G_i \mathbb{1}_p = 0$ for all $i \in I$ leads to $(A\mathbb{1})^T(A\hat{x} - y) = 0$. Applying the assumption (29) gives the result. $\square$

The assumption $A\mathbb{1} \propto \mathbb{1}$ holds in the case of shift-invariant blurring under the periodic boundary condition since then A is block-circulant. However, it does not hold for general operators A. We do not claim that the condition (29) on A is always necessary. Instead, we show next that (29) is necessary in a very simple but important case.

Remark 2. Let $\varphi(t) = t^2$, $\ker G = \mathrm{span}(\mathbb{1})$ and A be square and invertible. We will see that in this case (29) is a necessary and sufficient condition to have (30). Indeed, the minimizer $\hat{x}$ of $\mathcal{F}_y$ reads $\hat{x} = (A^TA + \frac{\beta}{2}G^TG)^{-1}A^Ty$. Taking (30) as a requirement is equivalent to $\mathbb{1}^T A(A^TA + \frac{\beta}{2}G^TG)^{-1}A^Ty = \mathbb{1}^Ty$, for all $y \in \mathbb{R}^q$. Equivalently, $A(A^TA + \frac{\beta}{2}G^TG)^{-1}A^T\mathbb{1} = \mathbb{1}$, and also $A^T\mathbb{1} = A^T\mathbb{1} + \frac{\beta}{2}G^TG(A^TA)^{-1}A^T\mathbb{1}$. Since $\ker G = \mathrm{span}(\mathbb{1})$, we get $A^T\mathbb{1} \propto A^TA\mathbb{1}$. Using that A is invertible, the latter is equivalent to $A\mathbb{1} \propto \mathbb{1}$.

Finding general necessary and sufficient conditions for (30) means ensuring that for every $y \in \mathbb{R}^q$, for every minimizer $\hat{x}$ of $\mathcal{F}_y$, we have $A\hat{x} - y \in \{A\mathbb{1}_p\}^\perp$. This may be tricky while the expected result seems of limited interest. Based on the remarks given above, we can expect that (30) fails for general operators A.

3. The residuals for edge-preserving regularization.

In this section we give bounds that characterize the data term at a local minimizer $\hat{x}$ of $\mathcal{F}_y$. More precisely we focus on edge-preserving functions φ which are currently characterized by

$$\|\varphi'\|_\infty = \sup_{0 \leq t < \infty} \left|\varphi'(t)\right| < \infty. \tag{31}$$

A look at Table 1 shows that this condition is satisfied by all the PFs there except for (f1). Let us notice that under H1-H3 we usually have $\|\varphi'\|_\infty = \varphi'(0)$.

Theorem 3.1. *Let $\mathcal{F}_y : \mathbb{R}^p \to \mathbb{R}$ read as in (1)-(2) where $\mathrm{rank}(A) = q \leq p$ and (3) holds. Let φ satisfy H1 combined with one of the assumptions H2 or H3. Assume also that $\|\varphi'\|_\infty$ is finite.*

For every $y \in \mathbb{R}^q$, if $\mathcal{F}_y$ reaches a (local) minimum at $\hat{x}$ then

$$\|y - A\hat{x}\|_\infty \leq \frac{\beta}{2}\, \|\varphi'\|_\infty\, \left\|\!\left|(AA^T)^{-1}A\right|\!\right\|_\infty |\!|\!| G |\!|\!|_1. \tag{32}$$

Remark 3. Notice that (32) and (33) provide tight bounds that are independent of y and hold for any local or global minimizer $\hat{x}$ of $\mathcal{F}_y$.

Let us remind that for any matrix C, we have $|||C|||_1 = \max_j \sum_i |C[i,j]|$ and $|||C|||_\infty = \max_i \sum_j |C[i,j]|$, see e.g. [13, 6]. Two important consequences of Theorem 3.1 are stated next.

Remark 4 (Signal denoising or segmentation)**.** If x is a signal and G corresponds to the differences between consecutive samples, it is easy to see that $|||G|||_1 = 2$. If A is the identity operator, $\left|\left|\left|(AA^T)^{-1}A^T\right|\right|\right|_\infty = 1$, so (32) yields

$$\|y - A\hat{x}\|_\infty \leq \beta\|\varphi'\|_\infty.$$

Remark 5 (Image denoising or segmentation)**.** Let the pixels of an image with m rows be ordered column by column in x. For definiteness, let us consider a forward discretization. If $\{G_i : i \in I\}$ corresponds to the first-order differences between each pixel and its 4 adjacent neighbors ($s = 1$), or to the discrete approximation of the gradient operator ($s = 2$), the matrix G is obtained by translating column by column and row by row, the following $2 \times m$ submatrix:

$$\begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & -1 \end{bmatrix},$$

all other entries being null. Whatever the boundary conditions, each column of G has at most 4 non-zero entries with values in $\{-1, 1\}$, hence

$$|||G|||_1 = 4.$$

If in addition A is the identity,

$$\|y - \hat{x}\|_\infty \leq 2\beta\|\varphi'\|_\infty. \tag{33}$$

Proof of Theorem 3.1. The cases relevant to H2 and H3 are analyzed separately.
Case H1-H2. From the first-order necessary condition for a minimum $\nabla\mathcal{F}_y(\hat{x}) = 0$,

$$2A^T(y - A\hat{x}) = \beta\nabla\Phi(\hat{x}).$$

By assumption, AA^T is invertible. Multiplying both sides of the above equation by $\frac{1}{2}(AA^T)^{-1}A$ yields

$$y - A\hat{x} = \frac{\beta}{2}(AA^T)^{-1}A\ \nabla\Phi(\hat{x})$$

and then

$$\|y - A\hat{x}\|_\infty \leq \frac{\beta}{2}\left|\left|\left|(AA^T)^{-1}A\right|\right|\right|_\infty \|\nabla\Phi(\hat{x})\|_\infty. \tag{34}$$

The entries of $\nabla\Phi(\hat{x})$ are given in (10). Introducing into (10) the assumption (31) and the observation that $\frac{\left|G_i[j,\cdot]x\right|}{\|G_i x\|} \leq 1$, $\forall j, \forall i$, leads to

$$\left|\partial_n\Phi(x)\right| \leq \sum_{i\in I}\left|\varphi'(\|G_i x\|)\right|\sum_{j=1}^{s}\frac{\left|G_i[j,\cdot]x\right|}{\|G_i x\|}\ \left|G_i[j,n]\right| \leq \|\varphi'\|_\infty \sum_{i\in I}\sum_{j=1}^{s}\left|G_i[j,n]\right|.$$

Observe that the double sum on the right is the ℓ_1 norm (denoted $\|.\|_1$) of the nth column of G. Using the expression for ℓ_∞ matrix norm mentioned just before Remark 4, it is found that

$$\|\nabla\Phi(\hat{x})\|_\infty = \max_{n\in I}\left|\partial_n\Phi(x)\right| \leq \|\varphi'\|_\infty \max_{n\in I}\sum_{i\in I}\sum_{j=1}^{s}\left|G_i[j,n]\right| = \|\varphi'\|_\infty |||G|||_1.$$

Inserting this into (34) leads to the result.

Case H1-H3. Using that $\delta\mathcal{F}_y(\hat{x})(v) \geq 0$ and $\delta\mathcal{F}_y(\hat{x})(-v) \geq 0$ for every $v \in \mathbb{R}^p$, where $\delta\mathcal{F}_y(\hat{x})(v)$ is given in (18), we obtain that

$$\left|2(A\hat{x}-y)^T Av + \beta \sum_{i\in J^c} \frac{\varphi'(\|G_i\hat{x}\|)}{\|G_i\hat{x}\|}(G_i\hat{x})^T\, G_i v\right| \leq \beta\varphi'(0)\sum_{i\in J}\|G_i v\|,$$

where J and J^c are defined in (17). Then we have the following inequality chain:

$$\begin{aligned}
\left|2(A\hat{x}-y)^T Av\right| &\leq \beta \sum_{i\in J^c}\varphi'(\|G_i\hat{x}\|)\sum_{j=1}^{s}\frac{\left|G_i[j,\cdot]\hat{x}\right|}{\|G_i\hat{x}\|}\left|G_i[j,\cdot]v\right| + \beta\varphi'(0)\sum_{i\in J}\|G_i v\| \\
&\leq \beta\|\varphi'\|_\infty\left(\sum_{i\in J^c}\sum_{j=1}^{s}\left|G_i[j,\cdot]v\right| + \sum_{i\in J}\|G_i v\|\right) \\
&= \beta\|\varphi'\|_\infty\left(\sum_{i\in J^c}\|G_i v\|_1 + \sum_{i\in J}\|G_i v\|\right).
\end{aligned}$$

Since $\|u\|_2 \leq \|u\|_1$ for every u, we can write that for every $v \in \mathbb{R}^p$,

$$\left|(A\hat{x}-y)^T Av\right| \leq \frac{\beta}{2}\|\varphi'\|_\infty\sum_{i\in I}\|G_i v\|_1 = \frac{\beta}{2}\|\varphi'\|_\infty\|Gv\|_1. \tag{35}$$

Let $\{e_n, 1 \leq n \leq q\}$ denote the canonical basis of $\mathbb{R}^q$. For any $n = 1,\ldots,q$, we apply (35) with

$$v = A^T(AA^T)^{-1}e_n.$$

Then for any $n = 1,\cdots,q$, we have $\left|(A\hat{x}-y)^T Av\right| = \left|(A\hat{x}-y)[n]\right|$ and (35) yields

$$\left|(A\hat{x}-y)[n]\right| \leq \frac{\beta}{2}\|\varphi'\|_\infty\|GA^T(AA^T)^{-1}e_n\|_1 \leq \frac{\beta}{2}\|\varphi'\|_\infty|\!|\!|G|\!|\!|_1\,\left\|A^T(AA^T)^{-1}e_n\right\|_1.$$

It follows that

$$|\!|\!|A\hat{x}-y|\!|\!|_\infty \leq \frac{\beta}{2}\|\varphi'\|_\infty\;|\!|\!|G|\!|\!|_1\;\max_{1\leq n\leq q}\left\|A^T(AA^T)^{-1}e_n\right\|_1.$$

Since $\left\|A^T(AA^T)^{-1}e_n\right\|_1$ is the ℓ_1 norm of the nth column of $A^T(AA^T)^{-1}$, we find that $\left\|A^T(AA^T)^{-1}\right\|_\infty = \max_{1\leq n\leq q}\left\|A^T(AA^T)^{-1}e_n\right\|_1$. Hence (32). □

Remark 6. In a statistical setting, the quadratic data-fidelity term $\|Ax-y\|^2$ in (1) corresponds to white Gaussian noise on the data [3, 7, 2]. Such a noise is unbounded, even if its ℓ_2 norm is finite. It may seem surprising to realize that whenever φ is edge-preserving, i.e. when $\|\varphi'\|_\infty$ is bounded, the minimizers $\hat{x}$ of $\mathcal{F}_y$ give rise to noise estimates $(y-A\hat{x})[i]$, $1 \leq i \leq q$ that are tightly bounded as stated in (32). So the assumption for Gaussian noise is distorted by the solution. The proof of the theorem reveals that this behavior is due to the boundedness of the gradient of the regularization term.

4. **Bounds on the reconstructed differences.**

The regularization term Φ being defined on the discrete gradients or differences Gx, we wish to find bounds characterizing how they behave in the solution $\hat{x}$. This problem is intricate even when A is the identity. In this section we systematically suppose that A^TA is invertible. The considerations are quite different according to the smoothness at zero of $t \to \varphi(|t|)$.

4.1. **Smooth regularization.** When the data y involve a linear transform A from the original x, it does not make sense to consider the (discrete) gradient of differences of y. Instead, we can compare $G\hat{x}$—which combines information from the data and from the prior—with a solution that is built based only on the data, without any prior. We will compare $G\hat{x}$ with $G\hat{z}$ where $\hat{z}$ is the least-squares solution, i.e. the minimizer of $\|Ax-y\|^2$:

$$\hat{z} = (A^TA)^{-1}A^Ty. \tag{36}$$

Theorem 4.1. *Let $\mathcal{F}_y : \mathbb{R}^p \to \mathbb{R}$ be of the form (1)-(2) where* $\operatorname{rank}(A) = p$ *and φ satisfy H1-H2. For every $y \in \mathbb{R}^q$, if $\mathcal{F}_y$ has a (local) minimum at $\hat{x}$, then*

(i) there is a linear operator $H_y : \mathbb{R}^r \to \mathbb{R}^r$ such that

$$G\hat{x} = H_y\, G\, \hat{z}, \tag{37}$$

$$\text{Spectral Radius}(H_y) \quad \leq \quad 1; \tag{38}$$

(ii) if $\varphi'(t) > 0$ on $(0, +\infty)$ and $\{G_i : i \in I\}$ is a set of linearly independent vectors of $\mathbb{R}^q$, then the linear operator H_y in (37) satisfies

$$\text{Spectral Radius}(H_y) < 1;$$

(iii) if A is orthonormal, then $\|G\hat{x}\| \leq \|G\|\ \|A^Ty\|$.

It can be useful to remind that for any $\varepsilon > 0$ there exists a matrix norm $|\!|\!|.|\!|\!|$ such that $|\!|\!|H_y|\!|\!| \leq \text{Spectral Radius}(H_y) + \varepsilon$—see e.g. [6]—and hence $|\!|\!|H_y|\!|\!| \leq 1 + \varepsilon$.

Proof of Theorem 4.1. Multiplying both sides of (13) by $G(A^TA)^{-1}$ yields

$$G\hat{x} + \frac{\beta}{2}G(A^TA)^{-1}G^T\text{diag}(\theta)G\hat{x} = G\hat{z}.$$

Then the operator H_y introduced in (37) reads

$$H_y = \left(I + \frac{\beta}{2}G(A^TA)^{-1}G^T\text{diag}(\theta)\right)^{-1}. \tag{39}$$

Let λ and u with $\|u\| = 1$ be an eigenvalue and the relevant eigenvector of H_y, respectively. Starting with $H_yu = \lambda u$ we derive

$$u = \lambda u + \lambda\frac{\beta}{2}G(A^TA)^{-1}G^T\text{diag}(\theta)u.$$

If $\text{diag}(\theta)u = 0$, we have $\lambda = 1$, hence (38) holds. Consider now that $\text{diag}(\theta)u \neq 0$, then $u^T\text{diag}(\theta)u > 0$. Multiplying both sides of the above equation from the left by $u^T\text{diag}(\theta)$ leads to

$$\lambda = \frac{u^T\text{diag}(\theta)u}{u^T\text{diag}(\theta)u + \frac{\beta}{2}u^T\text{diag}(\theta)G(A^TA)^{-1}G^T\text{diag}(\theta)u}.$$

Using that $u^T\text{diag}(\theta)G(A^TA)^{-1}G^T\text{diag}(\theta)u \geq 0$ shows that $\lambda \leq 1$ which proves (38).

Under the conditions of (ii), $\theta[i] > 0$ for all i, then $G^T\text{diag}(\theta)u \neq 0$ since the rows of G^T are linearly independent. It follows that $\frac{\beta}{2}u^T\text{diag}(\theta)G(A^TA)^{-1}G^T\text{diag}(\theta)u > 0$, hence the result stated in (ii).

In the case when A is orthonormal we have $A^TA = I$ so (13) yields

$$G\hat{x} = G\big(I + \frac{\beta}{2}G^T\text{diag}(\theta)G\big)^{-1}A^Ty$$

and hence

$$\begin{aligned}\|G\hat{x}\| &\leq \|G\|_2 \left\|\left(I + \frac{\beta}{2}G^T\mathrm{diag}(\theta)G\right)^{-1}A^Ty\right\| \\ &\leq \|G\|_2 \|A^Ty\|\end{aligned}$$

where the last inequality is obtained by applying Lemma 2.2(i). □

The condition on $\{G_i : i \in I\}$ required in (ii) holds for instance if x is a p-length signal and $G_i x = x_i - x_{i+1}$ for all $i = 1, \ldots, p-1$.

4.2. **Nonsmooth regularization.** With any local minimizer $\hat{x}$ of $\mathcal{F}_y$ we associate the subsets J and J^c defined in (17) as well as the subspace $\mathcal{K}_J$ given in (19), along with its orthonormal basis given by the columns of $B_J \in \mathbb{R}^{p\times d_J}$ where $d_J = \dim(\mathcal{K}_J)$.

Theorem 4.2. *Let $\mathcal{F}_y : \mathbb{R}^p \to \mathbb{R}$ be of the form (1)-(2) where* $\mathrm{rank}(A) = p$ *and φ satisfy H1-H3. For every $y \in \mathbb{R}^q$, if $\mathcal{F}_y$ has a (local) minimum at $\hat{x}$, then*

(i) there is a linear operator $H_y : \mathbb{R}^r \to \mathbb{R}^r$ such that

$$G\hat{x} = H_y\, G\hat{z}_J, \tag{40}$$

$$\text{Spectral Radius}(H_y) \leq 1, \tag{41}$$

where $\hat{z}_J$ is the least-squares solution constrained to $\mathcal{K}_J$, i.e. the point yielding

$$\min_{x\in\mathcal{K}_J} \|Ax - y\|^2; \tag{42}$$

(ii) if $\varphi'(t) > 0$ on $\mathbb{R}_+$ and $\{G_i : i \in I\}$ is a set of linearly independent vectors of $\mathbb{R}^p$, then the linear operator in (40) satisfies

$$\text{Spectral Radius}(H_y) < 1;$$

(iii) if in particular A is orthonormal then $\|G\hat{x}\| \leq \|G\|_2\, \|A^Ty\|$.

Proof. We adopt the notations introduced in (24). The least-squares solution constrained to $\mathcal{K}_J$ is of the form $\hat{z}_J = B_J\tilde{z}$ where $\tilde{z} \in \mathbb{R}^{d_J}$ yields the minimum of $\|AB_Jz - y\|^2$. Then we can write that

$$\tilde{z} = (A_J^TA_J)^{-1}A_J^Ty \tag{43}$$

and hence

$$\hat{z}_J = B_J(A_J^TA_J)^{-1}A_J^Ty.$$

Let $\tilde{x} \in \mathbb{R}^{d_J}$ be the unique element such that $\hat{x} = B_J\tilde{x}$. Multiplying both sides of (25) on the left by $(A_J^TA_J)^{-1}$ and using the expression for $\tilde{z}$ yields

$$\tilde{x} + \frac{\beta}{2}(A_J^TA_J)^{-1}G_J^T\mathrm{diag}(\theta)G_J\tilde{x} = \tilde{z}. \tag{44}$$

Multiplying both sides of the last equation on the left by GB_J, then using the expression for $\hat{z}_J$ and reminding that $GB_J\tilde{x} = G\hat{x}$ shows that

$$\left(I + \frac{\beta}{2}G_J(A_J^TA_J)^{-1}G_J^T\mathrm{diag}(\theta)\right)G\hat{x} = G\hat{z}_J.$$

The operator H_y evoked in (40) reads $H_y = \left(I + \frac{\beta}{2}G_J(A_J^TA_J)^{-1}G_J^T\mathrm{diag}(\theta)\right)^{-1}$. It has the same structure as the operator in (39). By the same arguments, it is found that (41) holds.

The case (ii) is similar to Theorem 4.1(ii) and its proof uses the same arguments.

When A is orthonormal, $A_J^T A_J = B_J^T A^T A B_J = B_J^T B_J = I$ with I the identity on $\mathbb{R}^{d_J}$. Using (43) and (44) we obtain

$$\Big(I + \frac{\beta}{2} G_J^T \mathrm{diag}(\theta) G_J\Big)\tilde{x} = A_J^T y$$

and then $\tilde{x} = \Big(I + \frac{\beta}{2} G_J^T \mathrm{diag}(\theta) G_J\Big)^{-1} A_J^T y$. Multiplying the latter equality on the left by GB_J and using (23) yields

$$G\hat{x} = GB_J\Big(I + \frac{\beta}{2} G_J^T \mathrm{diag}(\theta) G_J\Big)^{-1} B_J^T A^T y.$$

Since B_J is an orthonormal basis of $\mathcal{K}_J$, the latter is equivalent to

$$G\hat{x} = GB_J\Big(B_J^T B_J + \frac{\beta}{2} G_J^T \mathrm{diag}(\theta) G_J\Big)^{-1} B_J^T A^T y.$$

By using Lemma 2.2(i), we find the following:

$$\begin{aligned} \|G\hat{x}\| &\leq \|G\|_2 \left\| B_J\Big(B_J^T B_J + \frac{\beta}{2} G_J^T \mathrm{diag}(\theta) G_J\Big)^{-1} B_J^T A^T y \right\| \\ &\leq \|G\|_2 \, \|A^T y\|. \end{aligned}$$

$\square$

5. **Conclusion.**

We provide simple bounds characterizing the minimizers of regularized least-squares. These bounds are for arbitrary signals and images of a finite size and they hold for possibly nonsmooth or nonconvex regularization terms and hold for local and global minimizers. They do not involve asymptotic assumptions nor simplifications and address practical situations.

6. **Appendix.**

Proof of Lemma 2.2.

Statement (i). Denote $M = G^T \mathrm{diag}(\theta) G$ and observe that M is positive semi-definite. Using (3), the matrix $A^T A + M$ is positive definite so C is well defined. Let the scalar λ and $v \in \mathbb{R}^q$ be such that $\|v\| = 1$ and

$$A\left(A^T A + M\right)^{-1} A^T v = \lambda v. \tag{45}$$

Since $A\left(A^T A + M\right)^{-1} A^T$ is positive semi-definite and symmetric, $\lambda \geq 0$. We clearly have $\lambda = 0$ if and only if $A\left(A^T A + M\right)^{-1} A^T v = 0$. In the following, consider that $A\left(A^T A + M\right)^{-1} A^T v \neq 0$ in which case $A^T v \neq 0$ and $\lambda > 0$.

1. *Case* $\mathrm{rank}(A) = p \leq q$. Using that $A^T A$ is invertible, we deduce that

$$\left(A^T A + M\right)^{-1} A^T v = \lambda (A^T A)^{-1} A^T v.$$

Multiplying both sides of this equation by $v^T A (A^T A)^{-1}\left(A^T A + M\right)$ yields

$$\begin{aligned} v^T A (A^T A)^{-1} A^T v &= \lambda v^T A (A^T A)^{-1}\left(A^T A + M\right)(A^T A)^{-1} A^T v \\ &= \lambda v^T A (A^T A)^{-1} A^T v + \lambda v^T A (A^T A)^{-1} M (A^T A)^{-1} A^T v. \end{aligned}$$

The latter equation also reads

$$(1-\lambda) c_1 = \lambda c_2$$

for

$$\begin{aligned} c_1 &= v^T A(A^T A)^{-1} A^T v, \\ c_2 &= v^T A(A^T A)^{-1} M (A^T A)^{-1} A^T v. \end{aligned}$$

Notice that $c_1 > 0$ because $A^T v \neq 0$ and that $c_2 \geq 0$. Combining this with the fact that $\lambda \geq 0$ shows that $1 - \lambda \geq 0$.

2. *Case* $\text{rank}(A) = p' < p$. Consider the singular value decomposition of A

$$A = USV^T$$

where $U \in \mathbb{R}^{q \times q}$ and $V \in \mathbb{R}^{p \times p}$ are unitary, and $S \in \mathbb{R}^{q \times p}$ is diagonal with

$$S[i,i] > 0,\ 1 \leq i \leq p', \tag{46}$$

since $\text{rank}(A) = p'$. Using that $U^{-1} = U^T$ and $V^{-1} = V^T$, we can write that

$$\left(A^T A + M\right)^{-1} = V\Big(S^T S + V^T M V\Big)^{-1} V^T$$

and then the equation in (45) becomes

$$US\Big(S^T S + V^T M V\Big)^{-1} S^T U^T v = \lambda v. \tag{47}$$

Define $u = U^T v$. Then $u \neq 0$ because $Av \neq 0$ and moreover $\|u\| = 1$ because U is unitary. Multiplying on the left the two sides of (47) by U^T yields

$$S\Big(S^T S + V^T M V\Big)^{-1} S^T u = \lambda u. \tag{48}$$

Based on (46), the $q \times p$ diagonal matrix S can be put into the form

$$S = \begin{bmatrix} S_1 & \vdots & O \\ \cdots & & \cdots \\ O & \vdots & O \end{bmatrix}$$

where S_1 is diagonal of size $p' \times p'$ with $S_1[i,i] > 0$, $1 \leq i \leq p'$ while the other submatrices are null. If $p' = q < p$, then the matrices on the second row are void, i.e. $S = \Big[S_1 \vdots O\Big]$. Accordingly, let us consider the partitioning

$$\Big(S^T S + V^T M V\Big)^{-1} = \begin{bmatrix} L_{11} & \vdots & L_{12} \\ \cdots & & \cdots \\ L_{21} & \vdots & L_{22} \end{bmatrix}$$

where L_{11} is $p' \times p'$ and L_{22} is $p - p' \times p - p'$. Then (48) reads

$$\begin{bmatrix} S_1 L_{11} S_1 & \vdots & O \\ \cdots & & \cdots \\ O & \vdots & O \end{bmatrix} u = \lambda u \tag{49}$$

where the null matrices are absent if $p' = q < p$. If $p' < q$, then $u[i] = 0$ for all $i = p'+1, \ldots, q$. Define $u_1 \in \mathbb{R}^{p'}$ by

$$u_1[i] = u[i],\ 1 \leq i \leq p'$$

and notice that $\|u_1\| = \|u\| = 1$. Then (49) leads to

$$S_1 L_{11} S_1 u_1 = \lambda u_1. \tag{50}$$

In order to compute L_{11} we partition V as

$$V = \left[V_1 \vdots V_0 \right]$$

where V_1 is $p \times p'$ and V_0 has $p \times p - p'$. Furthermore,

$$S^T S + V^T M V = \begin{bmatrix} S_1^2 + V_1^T M V_1 & \vdots & V_1^T M V_0 \\ \cdots & & \cdots \\ V_0^T M V_1 & \vdots & V_0^T M V_0 \end{bmatrix}$$

where we denote

$$S_1^2 = S_1^T S_1 = \mathrm{diag}\Big((S_1[1,1])^2, \ldots, (S_1[p',p'])^2 \Big).$$

Noticing that the columns of V_0 yield $\ker(A^T A)$, the assumption (3) ensures that $V_0^T M V_0$ is invertible. By the formula for partitioned matrices [13] we get

$$L_{11} = \Big(S_1^2 + V_1^T M V_1 - V_1^T M V_0 \left(V_0^T M V_0 \right)^{-1} V_0^T M V_1 \Big)^{-1}.$$

Using that $\theta \in \mathbb{R}_+^r$, we can define $\theta^{\frac{1}{2}} \in \mathbb{R}^r$ by $\theta^{\frac{1}{2}}[i] = \sqrt{\theta[i]}$ for all $1 \le i \le r$. Observe that then $M = \Big(\mathrm{diag}(\theta^{\frac{1}{2}}) G \Big)^T \mathrm{diag}(\theta^{\frac{1}{2}}) G$. Define

$$\begin{aligned} H_1 &= \mathrm{diag}(\theta^{\frac{1}{2}}) G \, V_1 \\ H_0 &= \mathrm{diag}(\theta^{\frac{1}{2}}) G \, V_0. \end{aligned}$$

Then

$$\begin{aligned} L_{11} &= \Big(S_1^2 + H_1^T H_1 - H_1^T H_0 \left(H_0^T H_0 \right)^{-1} H_0^T H_1 \Big)^{-1} \\ &= \Big(S_1^2 + H_1^T \Big(I - H_0 \left(H_0^T H_0 \right)^{-1} H_0^T \Big) H_1 \Big)^{-1} \\ &= S_1^{-1} \big(I + P \big)^{-1} S_1^{-1} \end{aligned}$$

where P reads

$$P = S_1^{-1} H_1^T \Big(I - H_0 \left(H_0^T H_0 \right)^{-1} H_0^T \Big) H_1 S_1^{-1}.$$

Notice that P is symmetric and positive semi-definite since the matrix between the large parentheses is an orthogonal projector. With this notation, the equation in (50) reads $(I+P)^{-1} u_1 = \lambda u_1$ and hence

$$u_1 = \lambda \big(I + P \big) u_1.$$

Multiplying the two sides of this equation by u_1^T on the left, and recalling that $\|u_1\| = 1$ and that $\lambda \ge 0$, leads to

$$1 - \lambda = \lambda \, u^T P u \ge 0.$$

It follows that $\lambda \in [0, 1]$.

Statement (ii). Using that $A^T A$ is invertible and that $\ker(G^T \mathrm{diag}(\theta) G) = \mathrm{span}(h)$ because $\theta[i] > 0$, $1 \le i \le r$, we have the following chain of implications:

$$\begin{aligned} Cy = y \quad &\Leftrightarrow \quad A(A^T A + G^T \mathrm{diag}(\theta) G)^{-1} A^T y = y \\ &\Rightarrow \quad A^T y = A^T y + G^T \mathrm{diag}(\theta) G (A^T A)^{-1} A^T y \\ &\Rightarrow \quad G^T \mathrm{diag}(\theta) G (A^T A)^{-1} A^T y = 0 \\ &\Leftrightarrow \quad y \in \ker(A^T) \text{ or } (A^T A)^{-1} A^T y \propto h. \end{aligned} \tag{51}$$

It follows that $Cy \neq y$ for all $y \in \mathbb{R} \setminus \{\ker(A^T) \cup V_h\}$ where V_h is defined in (8). Combining this with the result in statement (i) above leads to (7). $\square$

Remark 7. Notice that if A is invertible (i.e. $\text{rank}(A) = p = q$), V_h is of dimension 1 and is spanned by the eigenvector of C corresponding to the unique eigenvalue equal to 1. This comes from the facts that in this case the implication in (51) is an equivalence and that $\ker(A^T) = \{0\}$.

REFERENCES

[1] A. Antoniadis and J. Fan, *Regularization of wavelet approximations*, Journal of Acoustical Society America, **96** (2001), 939–967.

[2] G. Aubert and P. Kornprobst, "Mathematical Problems in Images Processing," Springer-Verlag, Berlin, 2^{nd} ed., 2006.

[3] J. Besag, *Digital image processing : Towards Bayesian image analysis*, Journal of Applied Statistics, **16** (1989), 395–407.

[4] A. Blake and A. Zisserman, "Visual Reconstruction," The MIT Press, Cambridge, 1987.

[5] C. Bouman and K. Sauer, *A generalized Gaussian image model for edge-preserving MAP estimation*, IEEE Transactions on Image Processing, **2** (1993), 296–310.

[6] P.-G. Ciarlet, "Introduction à L'analyse Numérique Matricielle et à L'optimisation," Collection mathématiques appliquées pour la maîtrise, Dunod, Paris, 5e ed., 2000.

[7] G. Demoment, *Image reconstruction and restoration : Overview of common estimation structure and problems*, IEEE Transactions on Acoustics Speech and Signal Processing, ASSP, **37** (1989), 2024–2036.

[8] F. Fessler, *Mean and variance of implicitly defined biased estimators (such as penalized maximum likelihood): Applications to tomography*, IEEE Transactions on Image Processing, **5** (1996), 493–506.

[9] D. Geman, "Random Fields and Inverse Problems in Imaging," **1427**, École d'Été de Probabilités de Saint-Flour XVIII - 1988, Springer-Verlag, lecture notes in mathematics ed., 1990, 117–193.

[10] D. Geman and G. Reynolds, *Constrained restoration and recovery of discontinuities*, IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI, **14** (1992), 367–383.

[11] S. Geman and D. E. McClure, *Statistical methods for tomographic image reconstruction*, in "Proc. of the 46-th Session of the ISI", Bulletin of the ISI, **52**, 1987, 22–26.

[12] T. Hebert and R. Leahy, *A generalized EM algorithm for 3-D Bayesian reconstruction from Poisson data using Gibbs priors*, IEEE Transactions on Medical Imaging, **8** (1989), 194–202.

[13] R. Horn and C. Johnson, "Matrix analysis," Cambridge University Press, 1985.

[14] Y. G. Leclerc, *Constructing simple stable description for image partitioning*, International Journal of Computer Vision, **3** (1989), 73–102.

[15] S. Z. Li, "Markov Random Field Modeling in Computer Vision," Springer-Verlag, New York, 1^{st} ed., 1995.

[16] P. Moulin and J. Liu, *Statistical imaging and complexity regularization*, IEEE Transactions on Information Theory, **46** (2000), 1762–1777.

[17] D. Mumford and J. Shah, *Boundary detection by minimizing functionals*, in Proceedings of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing, (1985), 22–26.

[18] M. Nikolova, *Local strong homogeneity of a regularized estimator*, SIAM Journal on Applied Mathematics, **61** (2000), 633–658.

[19] ———, *Weakly constrained minimization, Application to the estimation of images and signals involving constant regions*, Journal of Mathematical Imaging and Vision, **21** (2004), 155–175.

[20] ———, *Analysis of the recovery of edges in images and signals by minimizing nonconvex regularized least-squares*, SIAM Journal on Multiscale Modeling and Simulation, **4** (2005), 960–991.

[21] P. Perona and J. Malik, *Scale-space and edge detection using anisotropic diffusion*, IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI, **12** (1990), 629–639.

[22] L. Rudin, S. Osher and C. Fatemi, *Nonlinear total variation based noise removal algorithm*, Physica, **60 D** (1992), 259–268.

[23] S. Saquib, C. Bouman and K. Sauer, *ML parameter estimation for Markov random fields, with appications to Bayesian tomography*, IEEE Transactions on Image Processing, **7** (1998), 1029–1044.

[24] U. Tautenhahn, *Error estimates for regularized solutions of non-linear ill-posed problems*, Inverse Problems, (1994), 485–500.
[25] C. R. Vogel and M. E. Oman, *Iterative method for total variation denoising*, SIAM Journal on Scientific Computing, **17** (1996), 227–238.

Received April 2007.

E-mail address: nikolova@cmla.ens-cachan.fr