



**HAL**  
open science

# Partly Smooth Regularization of Inverse Problems

Samuel Vaïter, Gabriel Peyré, Jalal M. Fadili

► **To cite this version:**

Samuel Vaïter, Gabriel Peyré, Jalal M. Fadili. Partly Smooth Regularization of Inverse Problems. 2014. hal-00987293v1

**HAL Id: hal-00987293**

**<https://hal.science/hal-00987293v1>**

Submitted on 5 May 2014 (v1), last revised 20 Nov 2014 (v4)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Partly Smooth Regularization of Inverse Problems

Samuel Vaiter and Gabriel Peyré  
CNRS and CEREMADE,  
Univ. Paris-Dauphine  
{vaiter,peyre}@ceremade.dauphine.fr

Jalal Fadili  
GREYC  
CNRS-ENSICAEN  
Jalal.Fadili@ensicaen.fr

## Abstract

This article studies the regularization of inverse problems with a convex prior promoting some notion of low-complexity. This low-complexity is obtained by using regularizers that are partly smooth functions. Such functions force the solution of variational problems to live in a low-dimension manifold which is stable under small perturbations of the functional. This property is crucial to make the underlying low-complexity model robust to small noise. We show that a simple criterion implies the stability of the active manifold to small noise perturbations of the observation when the regularization parameter is tuned proportionally to the noise level. This unifies and generalizes several previous works, where this theorem is known to hold for sparse, group sparse, total variation and low-rank regularizations.

**Keywords:** Inverse problem, partial smoothness, sensitivity analysis, convexity, sparsity, low-rank.

## 1 Introduction

### 1.1 Inverse Problem Regularization

We consider the following observation model

$$y = y_0 + w \quad \text{where} \quad y_0 = \Phi x_0,$$

where  $\Phi \in \mathbb{R}^{P \times N}$ ,  $x_0 \in \mathbb{R}^N$  is the vector to recover and  $w \in \mathbb{R}^P$  is some additive noise.

We use a convex bounded function  $J$  in order to regularize this inverse problem. The recovery of  $x_0$  is performed using the following convex variational problem

$$\min_{x \in \mathbb{R}^N} \left\{ E(x, \theta) = J(x) + \frac{1}{2\lambda} \|\Phi x - y\|^2 \right\} \quad (\mathcal{P}_\theta)$$

where  $\lambda > 0$ , and for notation simplicity, we denoted

$$\theta = (\lambda, y) \in \Theta = \mathbb{R}^+ \times \mathbb{R}^P$$

the parameters of the problem. When  $\lambda \rightarrow 0^+$ , we consider the constrained problem

$$\min_{x \in \mathbb{R}^N} \{E(x, (0, y)) = J(x) + \iota_{\mathcal{H}_y}(x)\} \quad (\mathcal{P}_{0,y})$$

$$\text{where } \mathcal{H}_y = \{x \in \mathbb{R}^N ; \Phi x = y\}$$

where for a closed convex set, its indicator function is

$$\iota_{\mathcal{C}}(x) = \begin{cases} 0 & \text{if } x \in \mathcal{C}, \\ +\infty & \text{otherwise.} \end{cases}$$

With these notations,  $E$  is defined on the domain  $\mathbb{R}^N \times \Theta$ .

In the following, we assume that

$$\ker(\Phi) \cap \ker(J) = \{0\} \quad (1)$$

so that both problems  $(\mathcal{P}_\theta)$  and  $(\mathcal{P}_{0,y})$  have non-empty bounded sets of solutions.

The goal of this paper is to asses the performance of this generic class of methods, i.e. to understand how close is the recovered solution of  $(\mathcal{P}_\theta)$  to  $x_0$ . We focus here in the low noise regime, i.e. when  $\|w\|$  is small enough, and study not only  $\ell^2$  stability, but also the identifiability of the correct low-dimensional manifold associated to  $x_0$ . This unifies and extend a large body of literature on sparsity and low-rank regularization, which is shown to special cases of the general theory of partly-smooth regularization.

## 1.2 Notations

We recall that a  $C^2$ -manifold  $\mathcal{M}$  around  $x \in \mathbb{R}^n$  (of codimension  $m$ ) is a subset of  $\mathbb{R}^n$  such that there exists an open set  $U$  of  $\mathbb{R}^n$  and a  $C^2$ -function  $F : U \rightarrow \mathbb{R}^m$  satisfying

$$\mathcal{M} \cap U = \{\bar{x} \in U ; F(\bar{x}) = 0\},$$

and  $F$  has surjective derivative throughout  $U$ . We say that  $\mathcal{M}$  is a  $C^2$ -manifold if  $\mathcal{M}$  is a  $C^2$ -manifold around every  $x \in \mathcal{M}$ . If  $\mathcal{M} \subset \mathbb{R}^N$  is a  $C^2$ -manifold around  $x \in \mathbb{R}^n$ , we denote  $\mathcal{T}_x(\mathcal{M})$  the tangent space of  $\mathcal{M}$  at  $x \in \mathbb{R}^N$  defined as

$$\mathcal{T}_{\mathcal{M}}(x) = \text{Ker } DF(p).$$

We define the tangent model subspace as

$$T_x = \text{VectHull}(\partial J(x))^\perp.$$

where the vectorial hull of a convex set  $\mathcal{C} \subset \mathbb{R}^N$  is

$$\text{VectHull}(\mathcal{C}) = \{\rho(x - x') ; (x, x') \in \mathcal{C}^2, \rho \in \mathbb{R}\}.$$

For a convex set  $\mathcal{C} \subset \mathbb{R}^N$ ,  $\text{ri}(\mathcal{C})$  is its relative interior, i.e. its interior for the topology of its affine hull (the smallest affine space containing  $\mathcal{C}$ ).

For a linear space  $T$ , we denote  $P_T$  the orthogonal projection on  $T$  and  $\Phi_T = \Phi \circ P_T$ .

### 1.3 Partly-smooth Functions

Toward the goal of studying the recovery guarantees of problem  $(P_\theta)$ , our central assumption will be that  $J$  is a partly smooth function. Partial smoothness of functions was originally defined [24]. Our definition hereafter specializes it to the case of closed convex functions.

**Definition 1.** *Let  $J$  be a bounded convex function.  $J$  is partly smooth at  $x$  relative to a set  $\mathcal{M}$  containing  $x$  if*

- (i) (Smoothness)  $\mathcal{M}$  is a  $C^2$ -manifold around  $x$  and  $J$  restricted to  $\mathcal{M}$  is  $C^2$  around  $x$ .
- (ii) (Sharpness) The tangent space  $\mathcal{T}_x(\mathcal{M})$  is  $T_x$ .
- (iii) (Continuity) The set-valued mapping  $\partial J$  is continuous at  $x$  relative to  $\mathcal{M}$ .

$J$  is said to be partly smooth relative to a set  $\mathcal{M}$  if  $\mathcal{M}$  is a manifold and  $J$  is partly smooth at each point  $x \in \mathcal{M}$  relative to  $\mathcal{M}$ .  $J$  is said to be locally partly smooth at  $x$  relative to a set  $\mathcal{M}$  if  $\mathcal{M}$  is a manifold and there exists a neighbourhood  $U$  of  $x$  such that  $J$  is partly smooth at each point  $x' \in \mathcal{M} \cap U$  relative to  $\mathcal{M}$ .

Note that in the previous definition,  $\mathcal{M}$  needs only to be defined locally around  $x$ , and it can be shown to be locally unique, see [21, Corollary 4.2]. Hence the notation  $\mathcal{M}$  is unambiguous.

**Remark 1** (Discussion of the properties).

- Since  $J$  is proper convex continuous, the subdifferential of  $\partial J(x)$  is everywhere non-empty and compact and every subgradient is regular. Therefore, the Clarke regularity property [24, Definition 2.7(ii)] is automatically verified.
- In view of [24, Proposition 2.4(i)-(iii)], our sharpness property (ii) is equivalent to that of [24, Definition 2.7(iii)].
- The continuity property (iii) is equivalent to the fact that  $\partial J$  is inner semicontinuous at  $x$  relative to  $\mathcal{M}$ , that is: for any sequence  $x_n$  in  $\mathcal{M}$  converging to  $x$  and any  $\eta \in \partial J(x)$ , there exists a sequence of subgradients  $\eta_n \in \partial J(x_n)$  converging to  $\eta$ . This equivalent characterization will be very useful in the proof of our main result.

### 1.4 Examples in Imaging and Machine Learning

We describe below some popular examples of partly smooth regularizers that are routinely used in signal processing and machine learning.

**$\ell^1$  sparsity.** One of the most popular non-quadratic convex regularization is the  $\ell^1$  norm

$$J(x) = \sum_{i=1}^N |x_i|$$

which enforces sparsity of the solution of the inverse problem. Indeed, it is easy to check that  $J$  is partly smooth at  $x$  for the linear space

$$\mathcal{M} = T_x = \{u \in \mathbb{R}^N ; \text{supp}(u) \subseteq \text{supp}(x)\}.$$

The use of sparse regularizations has been popularized in the signal processing literature under the name basis pursuit method [10] and in the statistics literature under the name Lasso [35].

**$\ell^1 - \ell^2$  group sparsity.** To better capture the sparsity pattern of natural signals and images, it makes sense to structure the sparsity into non-overlapping blocks  $\mathcal{B}$  such that  $\bigcup_{b \in \mathcal{B}} b = \{1, \dots, N\}$ . This structuration is enforced for instance by using a mixed  $\ell^1 - \ell^2$  norm

$$\|x\|_{\mathcal{B}} = \sum_{b \in \mathcal{B}} \|x_b\|$$

where  $x_b = (x_i)_{i \in b} \in \mathbb{R}^{|b|}$ . We refer to [2] and to the references therein for more details about group sparsity.

On contrary to the  $\ell^1$  norm, this norm is in general not polyhedral, but is still partly smooth relative to the linear manifold defined as

$$\mathcal{M} = T_x = \{u ; \text{supp}_{\mathcal{B}}(u) \subseteq \text{supp}_{\mathcal{B}}(x)\} \quad \text{where} \quad \text{supp}_{\mathcal{B}}(x) = \bigcup \{b ; x_b \neq 0\}.$$

**Spectral functions.** The natural extension of sparsity to matrix-valued data  $x \in \mathbb{R}^{n \times n}$  (where  $N = n^2$ ) is to impose a low-rank constraint, which should be understood as imposing sparsity of the singular values. We denote  $x = V_x \text{diag}(\Lambda_x) U_x^*$  an SVD decomposition of  $x$ , where  $\Lambda_x \in \mathbb{R}_+^n$ . Note that this can be extended easily to rectangular matrices. The nuclear norm imposes such a sparsity and is defined as

$$J(x) = \|x\|_* = \|\Lambda_x\|_1.$$

It has been used for instance for machine learning applications and matrix completion [30, 5] and phase retrieval [8].

The nuclear norm can be shown to be partly smooth for the manifold [26, Example 2]

$$\mathcal{M} = \{u ; \text{rank}(u) = \text{rank}(x)\}.$$

More generally, if  $j : \mathbb{R}^n \rightarrow \mathbb{R}$  is a permutation-invariant closed convex function, then one can consider the following function

$$J(x) = j(\Lambda_x)$$

which can be shown to be a convex function as well [25]. When restricted to the linear space of symmetric matrices,  $j$  is partly smooth at  $\Lambda_x$  for a manifold  $m_{\Lambda_x}$ , if and only if  $J$  is partly smooth at  $x$  relative to the manifold

$$\mathcal{M} = \{U \operatorname{diag}(\Lambda)U^* ; \Lambda \in m_{\Lambda_x}, U \in \mathcal{O}_n\},$$

where  $\mathcal{O}_n \subset \mathbb{R}^{n \times n}$  is the group of orthogonal matrices, see [11, Theorem 3.19]. This result can be extended to nonsymmetric matrices by requiring that  $j$  is an absolutely permutation-invariant closed convex function, see [11, Theorem 5.3]. The nuclear norm  $\|\cdot\|_*$  is a special case where  $j(\Lambda) = \|\Lambda\|_1$ .

**Analysis regularizers.** If  $J_0 : \mathbb{R}^Q \rightarrow \mathbb{R}$  is a convex function and  $D \in \mathbb{R}^{N \times Q}$  is a linear operator, one can consider the following analysis regularizer

$$J(x) = J_0(D^*x).$$

A popular example is when taking  $J_0 = \|\cdot\|_1$  and  $D^* = \nabla$  a finite difference approximation of the gradient of an image. This defines the (anisotropic) total variation, which promotes piecewise constant images, and is popular in image processing [33]. It is also possible to define families of sparsity-enforcing prior by using  $J_0 = \|\cdot\|_*$  the nuclear norm, see [18, 31].

If  $J_0$  is partly smooth at  $z = D^*x$  for the manifold  $\mathcal{M}_z^0$ , then it is shown in [24, Theorem 4.2] that  $J$  is partly smooth at  $x$  relative to the manifold

$$\mathcal{M} = \{u \in \mathbb{R}^N ; D^*u \in \mathcal{M}_z^0\}.$$

Note that as  $J_0$  is convex and continuous, so is  $J$ , and there is no need of the transversality/regularity condition in [24, Theorem 4.2].

**Mixed regularization.** Starting from a set of convex functions  $\{J_\ell\}_{\ell \in \mathcal{L}}$ , it is possible to design a convex function as

$$J(x) = \sum_{\ell \in \mathcal{L}} \rho_\ell J_\ell(x)$$

where  $\rho_\ell > 0$  are weights. A popular example is to impose both sparsity and low rank of a matrix, when using  $J_1 = \|\cdot\|_1$  and  $J_2 = \|\cdot\|_*$ , see for instance [28].

If each  $J_\ell$  is partly smooth at  $x$  relative to a manifold  $\mathcal{M}^\ell$ , then it is shown in [24, Corollary 4.8] that  $J$  is partly smooth at  $x$  for

$$\mathcal{M} = \bigcap_{\ell \in \mathcal{L}} \mathcal{M}^\ell.$$

Again, the regularity condition in [24, Corollary 4.8] is in force in our case by convexity and continuity.

## 1.5 Contributions

In the following, we denote  $T = T_{x_0}$ ,  $e = P_T(\partial J(x_0))$ . The following definition introduces the central object of this article, which controls the stability of  $\mathcal{M}$  when the signal to noise ratio is large enough.

**Definition 2** (Astymptotic pre-certificate). *Assuming  $\ker(\Phi) \cap T = \{0\}$ , we define*

$$\eta_F = \Phi^* p_F \quad \text{where} \quad p_F = \Phi_T^{+,*} e,$$

where we have denoted  $\Phi_T^+$  the Moore-Penrose pseudo inverse of  $\Phi_T$ .

Our main contribution is the following theorem, which shows the robustness of the manifold  $\mathcal{M} = \mathcal{M}_{x_0}$  associated to  $x_0$  to small noise, if  $\lambda$  is well chosen.

**Theorem 1.** *We assume that  $J$  is locally partly smooth at  $x_0$  relative to  $\mathcal{M}$  and that*

$$\ker(\Phi) \cap T = \{0\}, \quad \text{and} \quad \eta_F \in \text{ri}(\partial J(x_0)) \quad (2)$$

*Then there exists constants  $(C, C')$  such that if  $\|w\| \leq C$  and if  $\lambda = C'\|w\|$ , then the solution  $x_\theta$  of  $(\mathcal{P}_\theta)$  is unique and satisfies*

$$x_\theta \in \mathcal{M} \quad \text{and} \quad \|x_\theta - x_0\| = O(\|w\|). \quad (3)$$

**Remark 2** (Identification of the manifold). *Theorem 1 guarantees that, under some hypotheses on  $(x_0, w, \lambda)$ ,  $x_\theta$  belongs to  $\mathcal{M}$ . For all the regularizations considered in Section 1.4, one can furthermore show that actually, under these hypotheses,  $\mathcal{M}_{x_\theta} = \mathcal{M}$ . This is because, for any  $(x, x')$  with  $x' \in \mathcal{M}_x$  close enough to  $x$ , one has  $\mathcal{M}_{x'} = \mathcal{M}_x$ .*

The following proposition shows that Theorem 1 is in some sense sharp, since the hypothesis  $\eta_F \in \text{ri}(\partial J(x_0))$  (almost) characterizes the stability of  $\mathcal{M}$ .

**Proposition 1.** *We suppose that  $x_0$  is the unique solution of  $\mathcal{P}_{0,y_0}$  and that*

$$\ker(\Phi) \cap T = \{0\}, \quad \text{and} \quad \eta_F \notin \partial J(x_0).$$

*Then there exists  $C > 0$  such that for  $\|w\| \leq C\lambda$  and any  $\lambda$  small enough, then any solution  $x_\theta$  of  $(\mathcal{P}_\theta)$  satisfies  $x_\theta \notin \mathcal{M}$ .*

In the particular case where  $w = 0$  (no noise), this result shows that the manifold  $\theta$  is not correctly identified when solving  $\mathcal{P}_{\lambda,y_0}$  for any  $\lambda > 0$  small enough.

The only case not covered by either Theorem 1 or Proposition 1 is when  $\eta_F \in \text{rbound}(\partial J(x_0))$  (the relative boundary). In this case, one cannot conclude, since depending on the noise  $w$ , one can have either stability or non-stability of  $\mathcal{M}$ . We refer to [37] where an example illustrates this situation for the 1-D total variation  $J = \|\nabla \cdot\|_1$  (here  $\nabla$  is a discretization of the 1-D derivative operator).

## 1.6 Algorithmic Implications

A popular scheme to compute a solution of  $(\mathcal{P}_\theta)$  is the Forward-Backward splitting algorithm. A comprehensive treatment of the convergence properties of this algorithm, and other proximal splitting schemes, can be found in the monograph [3]. Starting from some  $x_0 \in \mathbb{R}^N$ , the algorithm implements the following iteration

$$x_{k+1} = \text{Prox}_{\tau\lambda J}(x_k - \tau\Phi^*(y - \Phi x_k)),$$

where the step size satisfies  $0 < \tau < 2/\|\Phi\|^2$ , and the proximity operator is defined as, for  $\gamma > 0$

$$\text{Prox}_{\gamma J}(x) = \underset{x' \in \mathbb{R}^N}{\text{argmin}} \frac{1}{2}\|x - x'\|^2 + \gamma J(x').$$

A close inspection of the proof of Theorem 1 tells us more about the properties of the vector  $\eta_\theta = \Phi^* \frac{y - x_\theta}{\lambda}$ . Indeed, we have  $\eta_\theta \in \text{ri}(\partial J(x_\theta))$  for the assumed regime of  $(\|w\|, \lambda)$ . This in turn implies that the assumptions of [20, Theorem 13.7], see also [19, Theorem 2]<sup>1</sup>, are fulfilled and thus shows the following manifold identification result of the Forward-Backward splitting algorithm.

**Theorem 2.** *Suppose that the assumptions of Theorem 1 hold. Then, for  $k$  large enough,  $x_k \in \mathcal{M}$ .*

In plain words, the Forward-Backward algorithm correctly identifies the manifold  $\mathcal{M}$  after a finite number of iterations. This result sheds some light on the convergence behavior of this algorithm in the favorable case where condition (2) holds and  $(\|w\|, \lambda)$  are sufficiently small.

## 1.7 Relation to Previous Works

**Works on linear convergence rates.** Following the pioneer work [4] (who study convergence in term of Bregman divergence), there is a large amount of works on the study conditions under which  $\|x_\theta - x_0\| = O(\|w\|)$  (so-called linear convergence rate) where  $x_\theta$  is any solution of  $(\mathcal{P}_\theta)$ , see for instance the book [34] for an overview of these results. The initial work of [17] proves a sharp criteria to ensure linear convergence rate for the  $\ell^1$  norm, and this approach is further extended to arbitrary convex functions by [16] and [14] who proves respectively convergence rate in term of  $J$  functional and  $\ell^2$  norm.

These works show that if

$$\ker(\Phi) \cap T = \{0\} \quad \text{and} \quad \exists \eta \in \text{Im}(\Phi^*) \cap \text{ri}(\partial J(x_0)) \quad (4)$$

---

<sup>1</sup>The result of [20] applies more generally to variable metric (Newton-like) Forward-Backward when the smooth term is assumed to be  $C^2$ . This can be easily adapted to our case by taking the metric as the identity. Observe also that the result of [19] applies to the projected gradient algorithm, i.e. when  $J$  is the indicator function of a closed convex partly smooth set, and the proof easily extends also to an arbitrary partly smooth closed convex function.



(which is often called the source condition), then linear convergence rate holds. Note that condition (2) implies (4), but it is stronger. Indeed, condition (4) does not ensure model consistency (3), which is a stronger requirement. Model consistency requires, as we show in our work, the use of a special certificate, the minimal norm certificate  $\eta_0$ , which is equal to  $\eta_F$  if  $\eta_F \in \text{ri}(\partial J(x_0))$  (see Proposition 7).

**Works on model consistency.** Theorem 1 is a generalization of many previous works that have appear in the literature. For the  $\ell^1$  norm,  $J = \|\cdot\|_1$ , to the best of our knowledge, this result was initially stated by Fuchs [15]. In this setting, the result (3) corresponds to the correct identification of the support, i.e.  $\text{supp}(x_\theta) = \text{supp}(x_0)$ . Condition (2) is known in the statistics literature under the name “irrepresentability condition” [40], and moving to a setting where both  $\Phi$  and  $w$  are random, it implies model consistency (also known as sparsistency for  $\ell^1$ ), i.e. the probability that the support is correctly identified tends to 0 when the dimensions of the problem increases.

Bach proves respectively in [1] and [2] Theorem 1 (in fact a variant since he considers randomized  $\Phi$  and  $w$ ) for  $\ell^1 - \ell^2$  and nuclear norm gauges, in the special case where  $\Phi$  has full rank (i.e. is injective). Our results thus shows that the same condition ensures rank consistency with the additional constraint that  $\ker(\Phi) \cap T = \{0\}$ .

Vaiter et al. proves Theorem 1 in [37] for a  $\ell^1$  analysis prior, i.e.  $J(x) = \|D^*x\|_1$  where  $D \in \mathbb{R}^{N \times Q}$  is an arbitrary linear operator. This includes, as a special case, the discrete anisotropic total variation prior, when using  $D^* = \nabla$  a finite difference approximation of the gradient operator. This was further generalized in [36] for partly smooth function  $J$  where the manifolds  $\mathcal{M}$  are linear spaces (and hence  $\mathcal{M} = T_x$ ). This includes as special cases analysis  $\ell^1 - \ell^2$  gauges  $J(x) = \|D^*x\|_{1,2}$  such as the isotropic discrete total variation. This also includes polyhedral functions such as  $\ell^\infty$ . Note that the nuclear norm (and composition of it with linear operators as proposed for instance in [18, 31]) does not fit into this framework, but is partly smooth, and thus is covered by Theorem 1. Lastly, a similar result was proved in [13] for an infinite dimensional sparse recovery problem over space of measures, when  $J$  the total variation of a measure. In this setting, a interesting finding is that, when  $\eta_0 \in \text{ri}(\partial J(x_0))$ ,  $\eta_0$  is not equal to  $\eta_F$  but to a difference certificate (called “vanishing derivative” certificate in [13]) that can also be computed by solving a linear system.

Condition (2) is often used when  $\Phi$  is drawn from the Gaussian matrix ensemble to asses the performance of compressed sensing recovery with  $\ell^1$  norm, see [38, 12]. This is extended to a more general family of decomposable norms (including in particular  $\ell^1 - \ell^2$  norms and the nuclear norm) in [7], but only in the noiseless setting. Our result shows that this analysis extends to the noisy setting as well, and ensures model consistency at high signal to low noise levels. The same condition is used to asses the performance of matrix completion (i.e. the operator  $\Phi$  is a random masking operator) in a noiseless setting [5, 9]. It was also used to ensure  $\ell^2$  robustness of matrix completion in a noisy setting [6],

and our findings shows that these results also ensure rank consistency for matrix completion at high signal to low noise levels.

**Sensitivity analysis.** Theorem 1 can be seen as a sensitivity analysis of the minimizers of the function  $f$  at the point  $(x, \theta) = (x_0, \theta_0)$  where  $\theta_0 = (0, y_0)$ . Classical sensitivity analysis of non-smooth functions seeks condition to ensure continuity of the map  $\theta \mapsto x_\theta$ , see for instance [32]. This is usually guaranteed by the source condition (4), which, as already exposed, ensures linear convergence rate, and hence Lipschitz behavior of this map. The analysis proposed by Theorem 1 goes one step further, by assessing that  $\mathcal{M}_{x_0}$  is a stable manifold (in the sense of [39]), since the minimizer  $x_\theta$  is unique and stays in  $\mathcal{M}_{x_0}$  for small  $\theta$ . Our main source of inspiration for this analysis is the notion of partly smooth function introduced by Lewis [24] in order to ensure the existence of stable manifolds. For convex functions (which is the setting considered in our work) this corresponds to the notion of  $\mathcal{U}$ -Lagrangian, introduced in [23]. Loosely speaking, a partly smooth function behaves smoothly as we move on the identifiable manifold, and sharply if we move normal to the manifold. In fact, the behaviour of the function and of its minimizers (or critical points) depend essentially on its restriction to this manifold, hence offering a powerful framework for sensitivity analysis theory. In particular, critical points of partly smooth functions move stably on the manifold as the function undergoes small perturbations [24, 27]. A important and distinctive feature of our result is that, while the regularized  $J$  is assumed to be partly smooth, the function  $f$  is not partly smooth at  $(x, \theta) = (x_0, \theta_0)$  relative to the manifold  $\mathcal{M}_{x_0} \times \Theta$  because of the indicator function (constrained problem) appearing in  $E(\cdot, (0, y))$  when  $\lambda = 0$ . Thus one cannot apply Theorem 5.7 of [24]. We refer to Section 3.1 for a discussion about this point.

## 2 Primal and Dual Problems

In this section, we do not assume that  $J$  is partly smooth. We first recall some fundamental properties of both the primal and dual problems. In particular, we study the convergence of the solutions to these problems when  $\lambda \rightarrow 0$ . While these properties are not used for the proof of Theorem 1 (Section 3), they are crucial to understand the rationale behind the hypotheses of this theorem, and in particular to introduce the key concept of minimal norm certificate.

### 2.1 Convergence of the Primal Problem

We first show the convergence of the solution of the primal problem toward  $x_0$ .

**Proposition 2.** *We denote  $x_\theta$  any solution of  $(\mathcal{P}_\theta)$ . If  $x_0$  is the unique solution*

of  $(\mathcal{P}_{0,y_0})$ , then

$$x_\theta \longrightarrow x_0 \quad \text{when} \quad \begin{cases} \lambda \longrightarrow 0, \\ \|y - y_0\|^2/\lambda \longrightarrow 0. \end{cases} \quad (5)$$

**Remark 3.** The notations in statement (5) should be understood as follow: if  $\theta_n = (\lambda_n, y_n)$  is a sequence such that  $\lambda_n \rightarrow 0$  and  $\|y_n - y_0\|^2/\lambda_n \rightarrow 0$ , then  $x_{\theta_n} \rightarrow x_0$ . We will make use of this slight abuse of notation in the remaining part of this section to ease the exposition.

*Proof.* This is a classical result, whose proof can be found for instance in [22, Theorem 3.5]. We recall it by sake of clarity.

By optimality of  $x_\theta$  one has  $f(x_\theta, \theta) \leq f(x_0, \theta)$  and hence

$$\|\Phi(x_\theta - x_0) - w\|^2 \leq \|w\|^2 + 2\lambda J(x_0), \quad (6)$$

$$J(x_\theta) \leq \frac{\|w\|^2}{2\lambda} + J(x_0). \quad (7)$$

Thanks to (1), these bounds shows that the sequence  $\{x_\theta\}_\theta$  is bounded if  $\|w\|^2/\lambda$  and  $\lambda$  are bounded. We let  $x^*$  be any accumulation point.

For the considered asymptotics, (6) implies that  $\Phi x^* = \Phi x_0$ , while (7) implies that  $J(x^*) \leq J(x_0)$ . This shows that  $x^*$  is a solution of  $(\mathcal{P}_{0,y_0})$  and hence  $x^* = x_0$ .  $\square$

## 2.2 Convergence of the Dual Problem

For  $\lambda > 0$ , the Fenchel-Rockafellar dual problem associated to  $(\mathcal{P}_\theta)$  reads

$$p_\theta = \underset{p \in \mathbb{R}^P}{\operatorname{argmin}} g(p, \theta) = J^*(\Phi^* p) - \langle y, p \rangle + \frac{\lambda}{2} \|p\|^2 \quad (\mathcal{D}_\theta)$$

where  $J^*$  is the Legendre transform of  $J$ , defined as

$$\forall u, \quad J^*(u) = \max_x \langle x, u \rangle - J(x).$$

Note that  $p_\theta$  is uniquely defined because  $g(\cdot, \theta)$  is strictly convex.

The following proposition recalls the primal-dual relationships.

**Proposition 3.** For any  $x_\theta$  solution of  $(\mathcal{P}_\theta)$ , one has

$$p_\theta = \frac{y - \Phi^* x_\theta}{\lambda} \quad \text{and} \quad \eta_\theta = \Phi^* p_\theta \in \partial J(x_\theta).$$

*Proof.* Since  $J$  is a bounded function, strong duality holds, hence the result using Fenchel-Rockafellar duality.  $\square$

For  $\lambda = 0$ , the dual problem associated to  $(\mathcal{P}_{0,y})$  reads

$$\mathcal{S}_{0,y} = \underset{p \in \mathbb{R}^P}{\operatorname{Argmin}} g(p, (0, y)) = J^*(\Phi^* p) - \langle y, p \rangle. \quad (\mathcal{D}_{0,y})$$

Note that in general  $\mathcal{S}_{0,y}$  is not a singleton. The following proposition shows how  $\mathcal{S}_{0,y}$  relates to solution of the primal problem.

**Proposition 4.** For any solution  $x_{0,y}$  of  $(\mathcal{P}_{0,y})$ , one has

$$\mathcal{S}_{0,y} = \{p \in \mathbb{R}^P ; \Phi^* p \in \partial J(x_{0,y})\}.$$

*Proof.* Similarly to Proposition 3, strong duality holds between  $(\mathcal{P}_{0,y})$  and  $(\mathcal{D}_{0,y})$ , and the primal-dual relationships states that  $(x_{0,y}, p_{0,y})$  form a solution to these problems if and only if  $\Phi^* p_{0,y} \in \partial J(x_{0,y})$ .  $\square$

The following definition singles out a particular solution of the dual problem.

**Definition 3** (Minimal norm certificate). We define

$$p_{0,y} = \operatorname{argmin}_{p \in \mathcal{S}_{0,y}} \|p\|. \quad (8)$$

Note that  $p_{0,y}$  is well defined because it is the projection of 0 on the closed convex set  $\mathcal{S}_{0,y}$ . In the following, we make use of the following notations

$$p_0 = p_{0,y_0} \quad \text{and} \quad \eta_0 = \Phi^* p_0.$$

The following proposition gives the limit of  $p_\theta$  when  $\theta \rightarrow 0$  in a certain sense.

**Proposition 5.** One has

$$\|p_\theta - p_0\| \leq \frac{\|y - y_0\|}{\lambda} + \varepsilon(\lambda)$$

where  $\varepsilon(\lambda) \rightarrow 0$  when  $\lambda \rightarrow 0$ .

*Proof.* This result is already proved in [13] in a special case of functional  $J$ . Formulation  $(\mathcal{D}_\theta)$  shows  $p_\theta$  is the proximal operator of the function  $J^*(\Phi^*\cdot)/\lambda$  applied at the point  $y/\lambda$ . This shows that  $y/\lambda \mapsto p_\theta$  is 1-Lipschitz, and hence

$$\|p_\theta - p_0\| \leq \|p_\theta - p_{\lambda,y_0}\| + \|p_{\lambda,y_0} - p_0\| \leq \frac{\|w\|}{\lambda} + \|p_{\lambda,y_0} - p_0\|.$$

We now prove that

$$p_\theta \xrightarrow{\lambda \rightarrow 0} p_{0,y},$$

which gives the desired result when setting  $y = y_0$  in the previous formula.

Since  $p_{0,y}$  is a solution of  $(\mathcal{D}_{0,y})$ , one has

$$-\langle p_{0,y}, y \rangle \leq -\langle p_\theta, y \rangle. \quad (9)$$

By optimality of  $p_\theta$ , one has  $g(p_\theta, \theta) \leq g(p_{0,y}, \theta)$ , and thus

$$-2\langle p_\theta, y \rangle + \lambda \|p_\theta\|^2 \leq -2\langle p_{0,y}, y \rangle + \lambda \|p_{0,y}\|^2 \leq -2\langle p_\theta, y \rangle + \lambda \|p_{0,y}\|^2$$

and thus

$$\|p_\theta\| \leq \|p_{0,y}\|. \quad (10)$$

This shows that  $\{p_\theta\}_\lambda$  is bounded. Let  $p^*$  be any cluster point. Operating as in the proof of Proposition 2, we have  $\forall \bar{x} \in \{x; y = \Phi x\}$

$$\|y - \Phi x_\theta\|^2 \leq 2\lambda J(\bar{x}) \quad \text{and} \quad J(x_\theta) \leq J(\bar{x}).$$

Letting  $\lambda \rightarrow 0$ , we get by continuity that

$$x_{0,y} \in \{x; y = \Phi x\} \quad \text{and} \quad J(x_\theta) \leq J(\bar{x}),$$

or equivalently, that  $x_{0,y}$  is a minimizer of  $(\mathcal{P}_{0,y})$ . Moreover, from the primal-dual extremality relationships, we have  $\Phi^* p_\theta \in \partial J(x_\theta)$ . Since  $J$  is a proper closed convex function, the graph of  $\partial J$  is sequentially closed, which yields  $\Phi^* p^* \in \partial J(x_{0,y})$ , i.e.  $p^* \in \mathcal{S}_{0,y}$ . Now (10) implies that  $\|p^*\| \leq \|p_{0,y}\|$  and hence  $p^* = p_{0,y}$ , which shows that  $p_\theta$  is converging to  $p_{0,y}$ .  $\square$

### 2.3 Uniqueness Sufficient Conditions

The following result is the extension of a classical result from  $\ell^1$  minimization to arbitrary  $J$  functionals, see for instance [15].

The result below follows the strategy of proof already used in [36]. However, to simplify the exposition of the proof, we assume that  $J$  is partly smooth at  $x_\theta$  a solution of  $(\mathcal{P}_\theta)$ . We start by the following lemma.

**Lemma 1.** *Let  $C$  be a non-empty closed convex set and  $f$  a proper lsc convex function. Let  $x$  be a minimizer of  $\min_{z \in C} f(z)$ . If*

$$f'(x, z - x) > 0 \quad \forall z \in C, z \neq x,$$

*then,  $x$  is the unique solution of  $f$  on  $C$ .*

The proof is deferred to Appendix A. We now state the uniqueness result.

**Proposition 6.** *Let  $x_\theta$  be a solution of  $(\mathcal{P}_\theta)$  (resp. a feasible point of  $(\mathcal{P}_{0,y})$ ). Assume that there exists a dual vector  $p$  such that  $\eta = \Phi^* p \in \text{ri}(\partial J(x_\theta))$ , and such that*

$$\ker(\Phi) \cap T = \{0\},$$

*where  $T = T_x$ . Assume also that  $J$  is partly smooth at  $x_\theta$  relative to the manifold  $\mathcal{M}$ . Then  $x_\theta$  is the unique solution of  $(\mathcal{P}_\theta)$  (resp.  $(\mathcal{P}_{0,y})$ ).*

*Proof.* One has

$$\eta \in \text{ri}(\partial J(x_\theta)) \Leftrightarrow J'(x_\theta, \delta) > \langle \eta, \delta \rangle \quad \forall \delta \text{ such that } J'(x_\theta, \delta) + J'(x_\theta, -\delta) > 0.$$

Applying this with  $\eta = \Phi^* p \in \text{ri}(\partial J(x_\theta))$ , we obtain

$$\Phi^* p \in \text{ri}(\partial J(x_\theta)) \Leftrightarrow J'(x_\theta, \delta) > \langle p, \Phi \delta \rangle \quad \forall \delta \text{ such that } J'(x_\theta, \delta) + J'(x_\theta, -\delta) > 0.$$

Now, since  $J$  is sharp (Definition 1(ii)) at  $x$ , using [24, Proposition 2.4], we have

$$J'(x_\theta, -\delta) > J'(x_\theta, \delta) \quad \text{for} \quad \forall \delta \in N_{\mathcal{M}}(x_\theta) \setminus \{0\}.$$

Hence,

$$\Phi^* p \in \text{ri}(\partial J(x_\theta)) \Rightarrow J'(x_\theta, \delta) > \langle p, \Phi \delta \rangle \quad \forall \delta \in N_{\mathcal{M}}(x_\theta) \setminus \{0\}.$$

Since  $N_{\mathcal{M}}(x_\theta) = T_x^\perp$ , we have

$$\Phi^* p \in \text{ri}(\partial J(x_\theta)) \Rightarrow J'(x_\theta, \delta) > 0 \quad \forall \delta \in \text{Ker } \Phi \setminus \{0\},$$

which concludes the proof using Lemma 1.  $\square$

**Proposition 7.** *Suppose  $J$  is partly smooth at  $x_0$  for the manifold  $\mathcal{M}$ . Under the hypothesis  $\text{ker}(\Phi) \cap T = \{0\}$ , one has*

$$\eta_{\text{F}} \in \text{ri}(\partial J(x_0)) \implies \eta_0 = \eta_{\text{F}}, \quad (11)$$

$$\eta_0 \in \text{ri}(\partial J(x_0)) \implies \eta_0 = \eta_{\text{F}}. \quad (12)$$

*These conditions implies that  $x_0$  is the unique solution of  $(\mathcal{P}_{0,y_0})$ .*

*Proof.* We first remark that the first order condition of  $(\mathcal{P}_{0,y_0})$  reads

$$x_0 \text{ solution of } (\mathcal{P}_{0,y_0}) \iff \exists \eta \in \text{Im}(\Phi^*) \cap \partial J(x_0).$$

This shows that the left hand side conditions of both (11) and (12) implies that  $x_0$  is a solution of  $(\mathcal{P}_{0,y_0})$ .

*Proof of (11)* Under the condition  $\text{ker}(\Phi) \cap T = \{0\}$ , one has, from the definition of  $\Phi_T^{*,+}$ , that

$$p_{\text{F}} = \underset{p}{\text{argmin}} \{ \|p\| ; \Phi_T^* p = e \} \quad (13)$$

Using Proposition 4 for  $w = 0$  with  $x_0$  being solution of  $(\mathcal{P}_{0,y_0})$ , one sees that the constraint of problem (13) includes the constraint of problem (8). Indeed, one has

$$\forall \eta \in \partial J(x), \quad P_{T_x}(\eta) = e_x.$$

If  $p_{\text{F}} \in \text{ri}(\partial J(x_0))$ , then it is a feasible point of problem (8) when  $w = 0$ . Hence, necessarily  $p_0 = p_{\text{F}}$ .

*Proof of (12)* Since  $x_0$  is a solution of  $(\mathcal{P}_{0,y_0})$ , according to Proposition 4, one has that

$$p_0 = \underset{p}{\text{argmin}} \{ \|p\|^2 ; \Phi_T^* p = e, \Phi_S^* p \in \mathcal{U} \}$$

where we have denoted  $S = T^\perp$  and  $\mathcal{U} = P_S(\partial J(x_0))$ . The first order conditions of this problem state the existence of  $q \in \mathbb{R}^N$  and  $u \in \mathbb{R}^N$  such that

$$p_0 + \Phi_T q + \Phi_S u = 0 \quad \text{where} \quad \begin{cases} \Phi_T^* p_0 = e, \\ u \in \mathcal{N}_{\mathcal{U}}(\Phi_S^* p_0). \end{cases}$$

The condition  $\Phi^* p_0 \in \text{ri}(\partial J(x_0))$  implies that  $\Phi_S^* p_0 \in \text{ri}(\mathcal{U})$  and thus  $\mathcal{N}_{\mathcal{U}}(\Phi_S^* p_0) = T$ . This implies  $\Phi_S u = 0$  and hence one has the equation

$$\Phi_T^* p_0 + \Phi_T^* \Phi_T q = e + \Phi_T^* \Phi_T q = 0$$

which leads to  $p_0 = (\Phi_T)^{+,*} e = p_{\text{F}}$ .

*Proof of uniqueness.* See Proposition 6.  $\square$

## 3 Proofs

### 3.1 Sensitivity of the Lagrangian Problem

Before diving into the proof of Theorem 1, we first show how the theory of partly smooth functions introduced in [24] can be directly applied to study the sensitivity of  $(\mathcal{P}_\theta)$ .

**Theorem 3.** *Let  $x_\theta$  be a solution of  $(\mathcal{P}_\theta)$ . We assume that  $J$  is locally partly smooth at  $x_\theta$  relative to a set  $\mathcal{M}$ . If*

$$\ker(\Phi) \cap T_{x_\theta} = \{0\} \quad \text{and} \quad \eta_\theta \in \text{ri}(\partial J(x_\theta)) \quad (14)$$

where we have denoted

$$\eta_\theta = \Phi^* p_\theta = \frac{1}{\lambda} \Phi^*(y - \Phi x_\theta),$$

then for  $\theta'$  close enough from  $\theta$ , the solution  $x_{\theta'}$  of  $(\mathcal{P}_{\theta'})$  is unique and satisfies

$$x_{\theta'} \in \mathcal{M}.$$

*Proof.* This is an straightforward application of [24, Theorem 5.7]. Indeed, by the smooth perturbation rule [24, Corollary 4.7], the function  $E$  is partly smooth at  $(x_\theta, \theta)$  relative to the manifold  $\mathcal{M} \times \Theta$ , and condition (14) is exactly equivalent to  $x_\theta$  being a strong minimizer of  $E(\cdot, \theta)$ , see [24, Definition 5.6].  $\square$

Condition (14) is not very useful because it depends on the solution  $x_\theta$  and not on the data to recover  $x_0$ . The rationale behind Theorem 1 is to make  $\theta$  tends to 0, and under the constraints

$$\lambda \rightarrow 0 \quad \text{and} \quad w/\lambda \rightarrow 0,$$

Propositions 2 and 5 ensure that

$$x_\theta \rightarrow x_0 \quad \text{and} \quad \eta_\theta \rightarrow \eta_0.$$

The heuristic underlying the hypotheses of Theorem 1 is that the conditions in (14) converge toward those of (2). Indeed, according to Proposition 7, condition (2) implies  $\eta_0 = \eta_F$ . This is precisely what we need to prove to make the statement of the theorem correct.

### 3.2 Proof of of Theorem 1

**Constrained problem.** We consider the following non-convex constrained minimization problem

$$\tilde{x}_\theta \in \underset{x \in \mathcal{M}}{\text{Argmin}} E(x, \theta) \quad (15)$$

We aim at showing that for  $(\|w\|/\lambda, \lambda)$  small enough,  $\tilde{x}_\theta$  is the unique solution of  $(\mathcal{P}_\theta)$ .

The proof of Proposition 2 carries over verbatim to this constrained problem, which shows that

$$\tilde{x}_\theta \rightarrow x_0 \quad \text{when} \quad \begin{cases} \lambda \rightarrow 0, \\ \|w\|^2/\lambda \rightarrow 0, \end{cases} \quad (16)$$

where convergence is in  $\mathcal{M}$ . In the following, for ease of notations, we denote  $\tilde{T} = T_{\tilde{x}_\theta}$ .

**Convergence of the tangent model subspace.** By definition of the constrained problem (15),  $\tilde{x}_\theta \in \mathcal{M}$ . Moreover, since  $E(\cdot, \theta)$  is partly smooth at  $x_0$  relative to  $\mathcal{M}$ , the sharpness property Definition 1(ii) holds at all nearby points in the manifold  $\mathcal{M}$  [24, Proposition 2.10]. Thus as soon as  $(\|w\|/\lambda, \lambda)$  is small enough, we have  $\tilde{T} = \mathcal{T}_{\tilde{x}_\theta}(\mathcal{M})$ . Using the fact that  $\mathcal{M}$  is of class  $C^2$ , we get

$$\tilde{T} = \mathcal{T}_{\tilde{x}_\theta}(\mathcal{M}) \rightarrow \mathcal{T}_{x_0}(\mathcal{M}) = T \quad \text{when} \quad \begin{cases} \lambda \rightarrow 0, \\ \|w\|^2/\lambda \rightarrow 0, \end{cases} \quad (17)$$

where the convergence should be understood over the Grassmannian of linear subspaces with the same dimension (or equivalently, as the convergence of the projection operators  $P_{\tilde{T}} \rightarrow P_T$ ). Since  $\ker(\Phi) \cap T = \{0\}$ , (17) implies that for  $(\|w\|^2/\lambda, \lambda)$  small enough,

$$\ker(\Phi) \cap \tilde{T} = \{0\}, \quad (18)$$

which we assume from now on.

**First order condition.** Local partial smoothness of  $J$ , hence of  $E(\cdot, \theta)$  by the smooth perturbation rule, at  $x_0$  relative to  $\mathcal{M}$  implies that for  $(\|w\|^2/\lambda, \lambda)$  sufficiently small,  $E(\cdot, \theta)$  is partially smooth at  $\tilde{x}_\theta \in \mathcal{M}$  relative to  $\mathcal{M}$ . Therefore, [24, Proposition 2.4(a)-(b)] applies, and it follows that  $\tilde{x}_\theta$  is a critical point of (15) if, and only if,

$$\begin{aligned} 0 \in \text{Aff}(\partial E(\tilde{x}_\theta, \theta)) &= \frac{1}{\lambda} \Phi^*(\Phi \tilde{x}_\theta - y) + \text{Aff}(\partial J(\tilde{x}_\theta)) \\ &= \frac{1}{\lambda} \Phi^*(\Phi \tilde{x}_\theta - y) + e_{\tilde{x}_\theta} + \tilde{T}^\perp. \end{aligned}$$

The first equality comes from the fact that  $E(\cdot, \theta)$  is a closed convex function and the second one from the decomposability of the subdifferential. Projecting this relation onto  $\tilde{T}$ , we get

$$\Phi_{\tilde{T}}^*(\Phi \tilde{x}_\theta - y) + \lambda e_{\tilde{x}_\theta} = 0, \quad (19)$$

since  $e_{\tilde{x}_\theta} \in \tilde{T}$ .

**Convergence of the primal variables.** Since both  $\tilde{x}_\theta$  and  $x_0$  belong to  $\mathcal{M}$ , and partial smoothness implies that  $\mathcal{M}$  is a manifold of class  $C^2$  around each of them, we deduce that each point in their respective neighbourhoods has a unique



projection on  $\mathcal{M}$  [29]. In particular,  $\tilde{x}_\theta = P_{\mathcal{M}}(\tilde{x}_\theta)$  and  $x_0 = P_{\mathcal{M}}(x_0)$ . Moreover,  $P_{\mathcal{M}}$  is of class  $C^1$  near  $\tilde{x}_\theta$  [26, Lemma 4]. Thus, continuous differentiability shows

$$\tilde{x}_\theta - x_0 = P_{\mathcal{M}}(\tilde{x}_\theta) - P_{\mathcal{M}}(x_0) = DP_{\mathcal{M}}(\tilde{x}_\theta)(\tilde{x}_\theta - x_0) + o(\|\tilde{x}_\theta - x_0\|) .$$

where  $DP_{\mathcal{M}}(\tilde{x}_\theta)$  is the derivative of  $P_{\mathcal{M}}$  at  $\tilde{x}_\theta$ . Using [26, Lemma 4], and recalling that  $\tilde{T} = \mathcal{T}_{\tilde{x}_\theta}(\mathcal{M})$ , the derivative  $DP_{\mathcal{M}}(\tilde{x}_\theta)$  is given by

$$DP_{\mathcal{M}}(\tilde{x}_\theta) = P_{\tilde{T}} .$$

Inserting this in (19), we get

$$\Phi_{\tilde{T}}^* \Phi(\tilde{x}_\theta - x_0) = \Phi_{\tilde{T}}^* \Phi_{\tilde{T}}(\tilde{x}_\theta - x_0) + o(\|\tilde{x}_\theta - x_0\|) = \Phi_{\tilde{T}}^* w - \lambda e_{\tilde{x}_\theta} .$$

Using (18),  $\Phi_{\tilde{T}}$  has full rank, and thus

$$\tilde{x}_\theta - x_0 = \Phi_{\tilde{T}}^+ w - \lambda (\Phi_{\tilde{T}}^* \Phi_{\tilde{T}})^{-1} e_{\tilde{x}_\theta} + o(\|\tilde{x}_\theta - x_0\|) . \quad (20)$$

Altogether, we obtain the bound

$$\|\tilde{x}_\theta - x_0\| = O(\|w\|, \lambda) .$$

**Convergence of the dual variables.** We define

$$\tilde{\eta}_\theta = \Phi^* \tilde{p}_\theta \quad \text{where} \quad \tilde{p}_\theta = \frac{y - \Phi \tilde{x}_\theta}{\lambda} .$$

Arguing as above, and using (20) we have

$$\begin{aligned} \lambda \tilde{p}_\theta &= \Phi(x_0 - \tilde{x}_\theta) + w \\ &= \Phi_{\tilde{T}}(x_0 - \tilde{x}_\theta) + w + o(\|\tilde{x}_\theta - x_0\|) \\ &= P_{\text{Im}(\Phi_{\tilde{T}})^\perp} w + \lambda \Phi_{\tilde{T}}^{+,*} e_{\tilde{x}_\theta} + o(\|\tilde{x}_\theta - x_0\|) . \end{aligned}$$

We thus arrive at

$$\|\tilde{p}_\theta - p_{\text{F}}\| = O\left(\frac{\|w\|}{\lambda}, \|\Phi_{\tilde{T}}^{+,*} e_{\tilde{x}_\theta} - \Phi_T^{+,*} e\|\right) .$$

Since  $\mathcal{M}$  is a  $C^2$  manifold, and by partial smoothness  $x \mapsto e_x$  is  $C^1$  on  $\mathcal{M}$  (recall that  $J$  is  $C^2$  on  $\mathcal{M}$ ), one has

$$\|e_{\tilde{x}_\theta} - e\| = O(\|\tilde{x}_\theta - x_0\|) .$$

Since  $A \mapsto A^{+,*}$  is smooth at  $A = \Phi_T$  along the manifold of matrices of constant rank, one has

$$\|\Phi_{\tilde{T}}^{+,*} - \Phi_T^{+,*}\| = O(\|\Phi_{\tilde{T}} - \Phi_T\|) = O(\|P_{\tilde{T}} - P_T\| \|\Phi\|) = O(\|\tilde{x}_\theta - x_0\|) .$$

This implies

$$\|\Phi_{\tilde{T}}^{+,*} e_{\tilde{x}_\theta} - \Phi_T^{+,*} e\| \leq \|\Phi_{\tilde{T}}^{+,*} - \Phi_T^{+,*}\| \|e_{\tilde{x}_\theta}\| + \|e_{\tilde{x}_\theta} - e\| \|\Phi_T^{+,*}\| = O(\|\tilde{x}_\theta - x_0\|) .$$

Altogether, we get the bound

$$\|\tilde{\eta}_\theta - \eta_{\text{F}}\| = O(\|w\|/\lambda, \lambda) . \quad (21)$$

**Convergence inside the relative interior.** Using the hypothesis that  $\eta_F \in \text{ri}(\partial J(x_0))$ , we will show that for  $(\|w\|/\lambda, \lambda)$  small enough,

$$\tilde{\eta}_\theta \in \text{ri}(\partial J(\tilde{x}_\theta)). \quad (22)$$

We follow the line of proof of [24].

Let us suppose this does not hold. Then there exists a sequence  $(\theta_n = (\lambda_n, w_n))_n$ , with  $(w_n/\lambda_n, \lambda_n)$  tending to 0, such that

$$\tilde{\eta}_n \in \text{rbound}(\partial J(\tilde{x}_n)) \quad (23)$$

where we used the shorthand notations

$$\tilde{x}_n = \tilde{x}_{\theta_n} \quad \text{and} \quad \tilde{\eta}_n = \tilde{\eta}_{\theta_n}.$$

According to (21) and (16),

$$(\tilde{x}_n, \tilde{\eta}_n) \rightarrow (x_0, \eta_F). \quad (24)$$

Condition (23) is equivalently stated as, for each  $n$

$$\exists z_n \in T_{\tilde{x}_n}^\perp, \quad \forall \eta \in \partial J(\tilde{x}_n), \quad \langle z_n, \eta - \tilde{\eta}_n \rangle \geq 0, \quad (25)$$

where one can impose the normalization  $\|z_n\| = 1$ . Up to a sub-sequence (that for simplicity we still denote  $z_n$  with a slight abuse of notation), since  $z_n$  is in a compact set, we can suppose that  $z_n \rightarrow z^*$ .

Since  $T_{\tilde{x}_n}^\perp \rightarrow T^\perp$  because  $\mathcal{M}$  is a  $C^2$  manifold, one has that  $z^* \in T^\perp$ . We now show that

$$\forall v \in \partial J(x_0), \quad \langle z^*, \eta - \eta_F \rangle \geq 0. \quad (26)$$

Indeed, let us consider any  $v \in \partial J(x_0)$ . In view of the continuity property in Definition 1(iii) (see also the discussion in Remark 1),  $\partial J$  is continuous at  $x_0$  along  $\mathcal{M}$ , so that since  $\tilde{x}_n \rightarrow x_0$  there exists  $v_n \in \partial J(\tilde{x}_n)$  with  $v_n \rightarrow v$ . Applying (25) with  $\eta = v_n$  gives

$$\langle z_n, v_n - \tilde{\eta}_n \rangle \geq 0.$$

Taking the limit  $n \rightarrow +\infty$  in this inequality leads to (26), which contradicts the fact that  $\eta_F \in \text{ri}(\partial J(x_0))$ .

In view of (22) and (18), using Proposition 6 shows that  $\tilde{x}_\theta = x_\theta$  is the unique solution of  $(\mathcal{P}_\theta)$ .  $\square$

### 3.3 Proof of Proposition 1

Let  $x_\theta$  be a solution of  $(\mathcal{P}_\theta)$ . Suppose that  $x_\theta \in \mathcal{M}$ . In particular,  $x_\theta$  is a solution of the non-convex minimization (15). Arguing as in the proof of Theorem 1, we get the bound (21), i.e.

$$\|\eta_\theta - \eta_F\| = O(\|w\|/\lambda, \lambda) \quad \text{where} \quad \eta_\theta = \Phi^* \frac{y - \Phi x_\theta}{\lambda}. \quad (27)$$

Since  $x_0$  is the unique solution of  $(\mathcal{P}_{0,y})$ ,  $p_0$  is well defined, hence  $\Phi^*p_0 \in \partial J(x)$ . Thus, there exists  $K > 0$  (for instance  $K = d(\eta_F, \partial J(x))$ ) such that  $\|\eta_F - \eta_0\| > K$ . Moreover,

$$\|\eta_F - \eta_0\| \leq \|\eta_F - \eta_\theta\| + \|\eta_\theta - \eta_0\|.$$

According to (27) and (5), one has

$$\|\eta_F - \eta_\theta\| \rightarrow 0 \quad \text{and} \quad \|\eta_\theta - \eta_0\| \rightarrow 0.$$

This leads to a contradiction, hence  $x_\theta \notin \mathcal{M}$ . □

## 4 Conclusion

In this paper, we provided a unified analysis of the recovery performances of partly smooth regularizations of linear inverse problems. A distinctive feature of our analysis is that we provided for the first time a criterion to ensure the correct identification of the low-complexity manifold associated to the input data. This criterion requires that a specific certificate (that we coined “minimal norm”) is non-degenerate. This condition can be checked by making use of a pre-certificate (that we coined “linearized”), which can be computed in closed form from the input data.

## Acknowledgements

The authors would like to thank Vincent Duval for fruitful discussions. This work has been supported by the European Research Council (ERC project SIGMA-Vision).

## A Directional Derivative and Uniqueness

*Proof of Lemma 1.* We first show that  $t \mapsto (f(x + t(z - x)) - f(z))/t$  is non-decreasing on  $(0, 1]$ . Indeed, let  $g : [0, 1] \rightarrow \mathbb{R}$  a convex function such that  $g(0) = 0$ . Let  $(t, s) \in (0, 1]^2$  with  $s > t$ . Then,

$$\begin{aligned} g(t) &= g(s(t/s)) = g(s(t/s) + (1 - t/s)0) \\ &\leq t \frac{g(s)}{s} + (1 - t/s)g(0) \\ &= t \frac{g(s)}{s}, \end{aligned}$$

which proves that  $t \in (0, 1] \mapsto \frac{g(t)}{t}$  is non-decreasing on  $(0, 1]$ . Since  $f$  is convex, applying this result shows that the function

$$t \in (0, 1] \mapsto g(t) = f(x + t(z - x)) - f(z)$$

is such that  $g(0) = 0$  and  $g(t)/t$  is non-decreasing.

Assume now that that  $f'(x, z - x) > 0$ . Then, for every  $x \in C$ ,

$$g(1) = f(z) - f(x) \geq f'(x, z - x) > 0, \quad \forall z \in C, z \neq x,$$

which is equivalent to  $x$  being the unique minimizer of  $f$  on  $C$ .  $\square$

## References

- [1] F.R. Bach. Consistency of the group Lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.
- [2] F.R. Bach. Consistency of trace norm minimization. *Journal of Machine Learning Research*, 9:1019–1048, 2008.
- [3] H. H. Bauschke and P.L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2011.
- [4] M. Burger and S. Osher. Convergence rates of convex variational regularization. *Inverse Problems*, 20(5):1411, 2004.
- [5] E. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.
- [6] E.J. Candès and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.
- [7] E.J. Candès and B. Recht. Simple bounds for recovering low-complexity models. *Math. Program*, 141(1-2):577–589, 2013.
- [8] E.J. Candès, T. Strohmer, and V. Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.
- [9] E.J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2009.
- [10] 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.
- [11] A. Daniilidis, D. Drusvyatskiy, and A. S. Lewis. Orthogonal invariance and identifiability. *to appear in SIAM J. Matrix Anal. Appl.*, 2014.
- [12] C. Dossal, M.L. Chabanol, G. Peyré, and J. Fadili. Sharp support recovery from noisy random measurements by  $\ell_1$  minimization. *Applied and Computational Harmonic Analysis*, 33(1):24–43, 2012.
- [13] V. Duval and G. Peyré. Exact support recovery for sparse spikes deconvolution. Technical report, Preprint hal-00839635, 2013.

- [14] J. Fadili, G. Peyré, S. Vaiter, C. Deledalle, and J. Salmon. Stable recovery with analysis decomposable priors. In *Proc. Sampta'13*, pages 113–116, 2013.
- [15] J.J. Fuchs. On sparse representations in arbitrary redundant bases. *IEEE Transactions on Information Theory*, 50(6):1341–1344, 2004.
- [16] M. Grasmair. Linear convergence rates for Tikhonov regularization with positively homogeneous functionals. *Inverse Problems*, 27:075014, 2011.
- [17] M. Grasmair, O. Scherzer, and M. Haltmeier. Necessary and sufficient conditions for linear convergence of  $\ell_1$ -regularization. *Communications on Pure and Applied Mathematics*, 64(2):161–182, 2011.
- [18] E. Grave, G. Obozinski, and F. Bach. Trace lasso: a trace norm regularization for correlated designs. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger, editors, *Proc. NIPS*, pages 2187–2195, 2011.
- [19] W. Hare and A. S. Lewis. Identifying active manifolds. *Algorithmic Operations Research*, 2(2):75–82, 2007.
- [20] W.L. Hare. Identifying active manifolds in regularization problems. In H. H. Bauschke, R. S., Burachik, P. L. Combettes, V. Elser, D. R. Luke, and H. Wolkowicz, editors, *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, volume 49 of *Springer Optimization and Its Applications*, chapter 13. Springer, 2011.
- [21] W.L. Hare and A.S. Lewis. Identifying active constraints via partial smoothness and prox-regularity. *J. Convex Anal.*, 11(2):251–266, 2004.
- [22] B. Hofmann, B. Kaltenbacher, C. Poschl, and O. Scherzer. A convergence rates result for Tikhonov regularization in Banach spaces with non-smooth operators. *Inverse Problems*, 23(3):987, 2007.
- [23] C. Lemaréchal, F. Oustry, and C. Sagastizábal. The u-lagrangian of a convex function. *Trans. Amer. Math. Soc.*, 352(2):711–729, 2000.
- [24] A. S. Lewis. Active sets, nonsmoothness, and sensitivity. *SIAM Journal on Optimization*, 13(3):702–725, 2003.
- [25] A. S. Lewis. The mathematics of eigenvalue optimization. *Mathematical Programming*, 97(1–2):155–176, 2003.
- [26] A. S. Lewis and J. Malick. Alternating projections on manifolds. *Mathematics of Operations Research*, 33(1):216–234, 2008.
- [27] A. S. Lewis and S. Zhang. Partial smoothness, tilt stability, and generalized Hessians. *SIAM Journal on Optimization*, 23(1):74–94, 2013.

- [28] S. Oymak, A. Jalali, M. Fazel, Y.C. Eldar, and B. Hassibi. Simultaneously structured models with application to sparse and low-rank matrices. *arXiv preprint arXiv:1212.3753*, 2012.
- [29] R.A. Poliquin, R.T. Rockafellar, and L. Thibault. Local differentiability of distance functions. *Trans. Amer. Math. Soc.*, 352:5231–5249, 2000.
- [30] B. Recht, M. Fazel, and P.A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.
- [31] E. Richard, F.R. Bach, and J-P. Vert. Intersecting singularities for multi-structured estimation. In *Proc. ICML*, volume 28 of *JMLR Proceedings*, pages 1157–1165. JMLR.org, 2013.
- [32] R.T. Rockafellar and R. Wets. *Variational analysis*, volume 317. Springer Verlag, 1998.
- [33] L.I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992.
- [34] O. Scherzer, M. Grasmair, H. Grossauer, M. Haltmeier, and F. Lenzen. *Variational Methods in Imaging*. Applied Mathematical Sciences. Springer, 1st edition, 2009.
- [35] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B. Methodological*, 58(1):267–288, 1996.
- [36] S. Vaiteer, M. Golbabaee, J. Fadili, and G. Peyré. Model selection with piecewise regular gauges. Technical report, Preprint Hal, 2013.
- [37] S. Vaiteer, G. Peyré, C. Dossal, and M.J. Fadili. Robust sparse analysis regularization. *IEEE Transactions on Information Theory*, 59(4):2001–2016, 2013.
- [38] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using  $\ell_1$ -constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, 2009.
- [39] S. J. Wright. Identifiable surfaces in constrained optimization. *SIAM Journal on Control and Optimization*, 31(4):1063–1079, 1993.
- [40] P. Zhao and B. Yu. On model selection consistency of Lasso. *J. Mach. Learn. Res.*, 7:2541–2563, December 2006.