



HAL
open science

Model Consistency of Partly Smooth Regularizers

Samuel Vaïter, Gabriel Peyré, Jalal M. Fadili

► **To cite this version:**

Samuel Vaïter, Gabriel Peyré, Jalal M. Fadili. Model Consistency of Partly Smooth Regularizers. 2014. hal-00987293v2

HAL Id: hal-00987293

<https://hal.science/hal-00987293v2>

Submitted on 7 Jun 2014 (v2), 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.

Model Consistency of Partly Smooth Regularizers

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

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

Abstract

This paper studies least-square regression penalized with partly smooth convex regularizers. This class of functions is very large and versatile allowing to promote solutions conforming to some notion of low-complexity. Indeed, they force solutions of variational problems to belong to a low-dimensional manifold (the so-called model) which is stable under small perturbations of the function. This property is crucial to make the underlying low-complexity model robust to small noise. We show that a generalized “irrepresentable condition” implies stable model selection under small noise perturbations in the observations and the design matrix, when the regularization parameter is tuned proportionally to the noise level. This condition is shown to be almost a necessary condition. We then show that this condition implies model consistency of the regularized estimator. That is, with a probability tending to one as the number of measurements increases, the regularized estimator belongs to the correct low-dimensional model manifold. This work unifies and generalizes several previous ones, where model consistency is known to hold for sparse, group sparse, total variation and low-rank regularizations.

1 Introduction

1.1 Problem Statement

We consider the following observation model

$$y = X\beta_0 + w,$$

where $X \in \mathbb{R}^{n \times p}$ is the design matrix (in statistics or machine learning) or the forward operator (in signal and imaging sciences), $\beta_0 \in \mathbb{R}^p$ is the vector to recover and $w \in \mathbb{R}^n$ is the noise. The design can be either deterministic or random, and similarly for the noise w .

Regularization is now a central theme in many fields including statistics, machine learning and inverse problems. It allows one to impose on the set of candidate solutions some prior structure on the object x_0 to be estimated. We therefore consider a positive convex bounded function J to promote such a prior. This then leads to solving the following convex optimization problem

$$\min_{\beta \in \mathbb{R}^p} \left\{ J(\beta) + \frac{1}{2\lambda} \|X\beta - y\|^2 \right\}, \quad (1)$$

where $\lambda > 0$ controls the amount of regularization.

To simplify the notations, we introduce the following “canonical” parameters

$$\theta = (\mu, u, \Gamma) = \left(\frac{\lambda}{n}, \frac{X^*y}{n}, \frac{X^*X}{n} \right) \in \Theta = \mathbb{R}^+ \times \mathbb{R}^p \times \mathbb{R}^{p \times p}$$

and we denote

$$\varepsilon = \frac{X^*w}{n} = u - \Gamma\beta_0.$$

In the following, we assume without loss of generality that $y \in \text{Im } X$ and thus $u \in \text{Im}(\Gamma)$.

With these new parameters, the initial problem (1) now reads

$$\min_{\beta \in \mathbb{R}^p} \left\{ E(\beta, \theta) = J(\beta) + \frac{1}{2\mu} \langle \Gamma\beta, \beta \rangle - \frac{1}{\mu} \langle \beta, u \rangle + \frac{1}{2\mu} \langle \Gamma^+ u, u \rangle \right\}. \quad (\mathcal{P}_\theta)$$

where A^+ stands for the Moore-Penrose pseudo-inverse of a matrix A .

When $\mu \rightarrow 0^+$, we consider the constrained problem

$$\min_{\beta \in \mathbb{R}^p} \{E(\beta, \theta_0) = J(\beta) + \iota_{\mathcal{H}_u}(\beta)\} \quad \text{where } \mathcal{H}_u = \{\beta \in \mathbb{R}^p ; \Gamma\beta = u\} \quad (\mathcal{P}_{\theta_0})$$

where $\theta_0 = (0, u, \Gamma)$ and where the indicator function of some closed convex set \mathcal{C} is $\iota_{\mathcal{C}}(\beta) = 0$ for $\beta \in \mathcal{C}$ and $\iota_{\mathcal{C}}(\beta) = +\infty$ otherwise. With these notations, E is a function on $\mathbb{R}^p \times \Theta$.

The goal of this paper is to assess the recovery performance of (\mathcal{P}_θ) , i.e. to understand how close is the recovered solution of (\mathcal{P}_θ) to β_0 . We focus here on the low noise regime, i.e. when ε is small enough, and study not only ℓ^2 stability, but also the identifiability of the correct low-dimensional manifold associated to β_0 . This unifies and extends a large body of literature, including sparsity and low-rank regularization, which turn to be a special case of the general theory of partly-smooth regularization.

1.2 Notations

If $\mathcal{M} \subset \mathbb{R}^p$ is a C^2 -manifold around $\beta \in \mathbb{R}^p$, we denote $\mathcal{T}_\beta(\mathcal{M})$ the tangent space of \mathcal{M} at $\beta \in \mathbb{R}^p$. We define the tangent model subspace as

$$T_\beta = \text{VectHull}(\partial J(\beta))^\perp.$$

where the linear hull of a convex set $\mathcal{C} \subset \mathbb{R}^p$ is $\text{VectHull}(\mathcal{C}) = \{\rho(\beta - \beta') ; (\beta, \beta') \in \mathcal{C}, \rho \in \mathbb{R}\}$. For a convex set $\mathcal{C} \subset \mathbb{R}^p$, $\text{ri}(\mathcal{C})$ 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 for a matrix $\Gamma \in \mathbb{R}^{p \times p}$, $\Gamma_T = P_T \Gamma P_T$.

2 Partly-smooth Functions

Toward the goal of studying the recovery guarantees of problem (\mathcal{P}_θ) , our central assumption will be that J is a partly smooth function. Partial smoothness of functions was originally defined [12]. Our definition hereafter specializes it to the case of bounded convex functions.

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

- (i) (Smoothness) \mathcal{M} is a C^2 -manifold around β and J restricted to \mathcal{M} is C^2 around β .
- (ii) (Sharpness) The tangent space $\mathcal{T}_\beta(\mathcal{M})$ is T_β .
- (iii) (Continuity) The set-valued mapping ∂J is continuous at β 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 $\beta \in \mathcal{M}$ relative to \mathcal{M} . J is said to be locally partly smooth at β relative to a set \mathcal{M} if \mathcal{M} is a manifold and there exists a neighbourhood U of β such that J is partly smooth at each point $\beta' \in \mathcal{M} \cap U$ relative to \mathcal{M} .

Note that in the previous definition, \mathcal{M} needs only to be defined locally around β , and it can be shown to be locally unique, see [9, Corollary 4.2].

Remark 1 (Discussion of the properties). *Since J is proper convex continuous, the subdifferential of $\partial J(\beta)$ is everywhere non-empty and compact and every subgradient is regular. Therefore, the*

Clarke regularity property [12, Definition 2.7(ii)] is automatically verified. In view of [12, Proposition 2.4(i)-(iii)], our sharpness property (ii) is equivalent to that of [12, Definition 2.7(iii)]. The continuity property (iii) is equivalent to the fact that ∂J is inner semicontinuous at β relative to \mathcal{M} , that is: for any sequence β_n in \mathcal{M} converging to β and any $\eta \in \partial J(\beta)$, there exists a sequence of subgradients $\eta_n \in \partial J(\beta_n)$ converging to η . This equivalent characterization will be very useful in the proof of our main result.

2.1 Examples in Imaging and Machine Learning

We describe below some popular examples of partly smooth regularizers that are routinely used in machine learning, statistics and imaging sciences.

ℓ^1 sparsity. One of the most popular non-quadratic convex regularization is the ℓ^1 norm $J(\beta) = \sum_{i=1}^p |\beta_i|$, which promotes sparsity. Indeed, it is easy to check that J is partly smooth at β relative to the subspace

$$\mathcal{M} = T_\beta = \{u \in \mathbb{R}^p ; \text{supp}(u) \subseteq \text{supp}(\beta)\}.$$

The use of sparse regularizations has been popularized in the signal processing literature under the name basis pursuit method [5] and in the statistics literature under the name Lasso [20].

$\ell^1 - \ell^2$ group sparsity. To better capture the sparsity pattern of natural signals and images, it is useful to structure the sparsity into non-overlapping blocks/groups \mathcal{B} such that $\bigcup_{b \in \mathcal{B}} b = \{1, \dots, p\}$. This group structure is enforced by using typically the mixed $\ell^1 - \ell^2$ norm $J(\beta) = \sum_{b \in \mathcal{B}} \|\beta_b\|$, where $\beta_b = (\beta_i)_{i \in b} \in \mathbb{R}^{|b|}$. We refer to [23, 2] and references therein for more details. Unlike the ℓ^1 norm, and except the case $|b| = 1$, the $\ell^1 - \ell^2$ norm is not polyhedral, but is still partly smooth at β relative to the linear manifold defined as

$$\mathcal{M} = T_\beta = \{\beta' ; \text{supp}_{\mathcal{B}}(\beta') \subseteq \text{supp}_{\mathcal{B}}(\beta)\} \quad \text{where} \quad \text{supp}_{\mathcal{B}}(\beta) = \bigcup \{b ; \beta_b \neq 0\}.$$

Spectral functions. The natural spectral extension of sparsity to matrix-valued data $\beta \in \mathbb{R}^{p_0 \times p_0}$ (where $p = p_0^2$) is to impose a low-rank prior, which should be understood as sparsity of the singular values. Denote $\beta = V_\beta \text{diag}(\Lambda_\beta) U_\beta^*$ an SVD decomposition of β , where $\Lambda_\beta \in \mathbb{R}_+^{p_0}$. Note that this can be extended easily to rectangular matrices. The nuclear norm is defined as $J(\beta) = \|\beta\|_* = \|\Lambda_\beta\|_1$. It has been used for instance in machine learning applications [2], matrix completion [17, 3] and phase retrieval [4]. The nuclear norm can be shown to be partly smooth at x relative to the manifold [14, Example 2] $\mathcal{M} = \{\beta' ; \text{rank}(\beta') = \text{rank}(\beta)\}$. More generally, if $j : \mathbb{R}^{p_0} \rightarrow \mathbb{R}$ is a permutation-invariant closed convex function, then one can consider the function $J(\beta) = j(\Lambda_\beta)$ which can be shown to be a convex function as well [13]. When restricted to the linear space of symmetric matrices, j is partly smooth at Λ_β for a manifold m_{Λ_β} , if and only if J is partly smooth at β relative to the manifold

$$\mathcal{M} = \{U \text{diag}(\Lambda) U^* ; \Lambda \in m_{\Lambda_\beta}, U \in \mathcal{O}_{p_0}\},$$

where $\mathcal{O}_{p_0} \subset \mathbb{R}^{p_0 \times p_0}$ is the group of orthogonal matrices, see [6, Theorem 3.19]. This result can be extended to non-symmetric matrices by requiring that j is an absolutely permutation-invariant closed convex function, see [6, 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 $D \in \mathbb{R}^{p \times q}$ is a linear operator, one can consider the analysis regularizer $J(\beta) = J_0(D^* \beta)$. 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 [19]. It is also possible to define families of sparsity-enforcing prior by using $J_0 = \|\cdot\|_*$ the nuclear norm, see [8, 18]. If J_0 is partly smooth at $z = D^* \beta$ for the manifold \mathcal{M}_z^0 , then it is shown in [12, Theorem 4.2] that J is partly smooth at β relative to the manifold

$$\mathcal{M} = \{\beta' \in \mathbb{R}^p ; D^* \beta' \in \mathcal{M}_z^0\}.$$

Mixed regularization. Starting from a set of convex functions $\{J_\ell\}_{\ell \in \mathcal{L}}$, it is possible to design a convex function as $J_\ell(\beta) = \sum_{\ell \in \mathcal{L}} \rho_\ell J_\ell(\beta)$, 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 [15]. If each J_ℓ is partly smooth at β relative to a manifold \mathcal{M}^ℓ , then it is shown in [12, Corollary 4.8] that J is also partly smooth at β for $\mathcal{M} = \bigcap_{\ell \in \mathcal{L}} \mathcal{M}^\ell$.

3 Main results

In the following, we denote $T = T_{\beta_0}$, $e = P_T(\partial J(\beta_0)) \in \mathbb{R}^p$. Before stating our main contributions, we first introduce a central object of this paper, which controls the stability of \mathcal{M} when the signal to noise ratio is large enough.

Definition 2 (Linearized pre-certificate). *For some matrix $\Gamma \in \mathbb{R}^{p \times p}$, assuming $\ker(\Gamma) \cap T = \{0\}$, we define $\eta_\Gamma = \Gamma \Gamma_T^+ e$.*

3.1 Deterministic model consistency.

We first consider the case where X and w (or equivalently Γ and u) are fixed and deterministic. Our main contribution is the following theorem, which shows the robustness of the manifold \mathcal{M} associated to β_0 to small perturbations on both the observations and the design matrix, provided that μ is well chosen.

Theorem 1. *We assume that J is locally partly smooth at β_0 relative to \mathcal{M} and that there exists $\tilde{\Gamma} \in \mathbb{R}^{p \times p}$ such that*

$$\ker(\tilde{\Gamma}) \cap T = \{0\}, \quad \text{and} \quad \eta_{\tilde{\Gamma}} \in \text{ri}(\partial J(\beta_0)). \quad (2)$$

Then, there exists a constant $C > 0$ such that if

$$\max\left(\|\Gamma - \tilde{\Gamma}\|, \|\varepsilon\| \mu^{-1}, \mu\right) \leq C, \quad (3)$$

the solution β_θ of (\mathcal{P}_θ) is unique and satisfies

$$\beta_\theta \in \mathcal{M} \quad \text{and} \quad \|\beta_\theta - \beta_0\| = O(\|\varepsilon\|). \quad (4)$$

This theorem is proved in Section 4.1.

The following proposition, proved in Section 4.3, shows that Theorem 1 is in some sense sharp, since the hypothesis $\eta_\Gamma \in \text{ri}(\partial J(\beta_0))$ (almost) characterizes the stability of \mathcal{M} .

Proposition 1. *We suppose that β_0 is the unique solution of $\mathcal{P}_{(0, \tilde{\Gamma}\beta_0, \tilde{\Gamma})}$ and that*

$$\ker(\tilde{\Gamma}) \cap T = \{0\}, \quad \text{and} \quad \eta_{\tilde{\Gamma}} \notin \partial J(\beta_0). \quad (5)$$

Then there exists $C > 0$ such that if (3) holds, then any solution β_θ of (\mathcal{P}_θ) for $\mu > 0$ satisfies $\beta_\theta \notin \mathcal{M}$.

In the particular case where $\varepsilon = 0$ (no noise) and $\tilde{\Gamma} = \Gamma$, this result shows that the manifold θ is not correctly identified when solving $\mathcal{P}_{(\mu, \Gamma\beta_0, \Gamma)}$ for any $\mu > 0$ small enough.

Remark 2 (Critical case). *The only case not covered by either Theorem 1 or Proposition 1 is when $\eta_{\tilde{\Gamma}} \in \text{rbound}(\partial J(\beta_0))$ (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 [22] where an example illustrates this situation for the 1-D total variation $J = \|\nabla \cdot\|_1$ (here ∇ is a discretization of the 1-D derivative operator).*

3.2 Probabilistic model consistency.

We now turn to study consistency of our estimator. In this section, we work under the classical setting where p and β_0 are fixed as the number of observations $n \rightarrow \infty$. We consider that the design matrix and the noise are random. More precisely, the data (ξ_i, w_i) are random vectors in $\mathbb{R}^p \times \mathbb{R}$, $i = 1, \dots, n$, where ξ_i is the i -th row of X , are assumed independent and identically distributed (i.i.d.) samples from a joint probability distribution such that $\mathbb{E}(W_i | \xi_i) = 0$, finite

fourth-order moments, i.e. $\mathbb{E}(W_i^4) < +\infty$ and $\mathbb{E}(\|\xi_i\|^4) < +\infty$. Note that in general, W_i and ξ_i are not necessarily independent. It is possible to extend our result to other distribution models by weakening some of the assumptions and strengthening others, see e.g. [11, 24, 2]. Let's denote $\tilde{\Gamma} = \mathbb{E}(\xi^*\xi) \in \mathbb{R}^{p \times p}$, where ξ is any row of X . We do not make any assumption on invertibility of $\tilde{\Gamma}$.

To make the discussion clearer, the canonical parameters θ will be indexed by n . The estimator β_{θ_n} obtained by solving (\mathcal{P}_{θ_n}) for a sequence θ_n is said to be consistent for β_0 if, $\lim_{n \rightarrow +\infty} \Pr(\beta_{\theta_n} \text{ is unique}) \rightarrow 1$ and β_{θ_n} converges to β_0 in probability. The estimator is said to be model consistent if $\lim_{n \rightarrow +\infty} \Pr(\beta_{\theta_n} \in \mathcal{M}) \rightarrow 1$, where \mathcal{M} is the manifold associated to β_0 .

The following result ensures model consistency for certain scaling of μ_n . It is proved in Section 4.2

Theorem 2. *If conditions (2) hold and*

$$\mu_n = o(1) \quad \text{and} \quad \mu_n^{-1} = o(n^{1/2}). \quad (6)$$

Then the estimator β_{θ_n} of β_0 obtained by solving (\mathcal{P}_{θ_n}) is model consistent.

3.3 Relation to Previous Works

Theorem 1 is a generalization of a large body of results in the literature. For the Lasso, i.e. $J = \|\cdot\|_1$, and when $\Gamma = \tilde{\Gamma}$, to the best of our knowledge, this result was initially stated in [7]. In this setting, the result (4) corresponds to the correct identification of the support, i.e. $\text{supp}(\beta_\theta) = \text{supp}(\beta_0)$. Condition (2) for $J = \|\cdot\|_1$ is known in the statistics literature under the name ‘‘irrepresentable condition’’, see e.g. [24]. [11] have shown estimation consistency for Lasso for fixed p and β_0 and asymptotic normality of the estimates. The authors in [24] proved Theorem 2 for $J = \|\cdot\|_1$, though under slightly different assumptions on the covariance and noise distribution. A similar result was established in [10] for the elastic net, i.e. $J = \|\cdot\|_1 + \rho \|\cdot\|_2^2$ for $\rho > 0$. In [1] and [2], the author have shown Theorem 2 for two special cases, namely the group Lasso nuclear/trace norm minimization. Note that these previous works assume that the asymptotic covariance $\tilde{\Gamma}$ is invertible. We do not impose such an assumption, and only require the weaker restricted injectivity condition $\ker(\tilde{\Gamma}) \cap T = \{0\}$. In a previous work [22], we have proved an instance of Theorem 1 when $\Gamma = \tilde{\Gamma}$ and $J(\beta) = \|D^*\beta\|_1$, where $D \in \mathbb{R}^{p \times q}$ is an arbitrary linear operator. This covers as special cases the discrete anisotropic total variation or the fused Lasso. This result was further generalized in [21] when $\Gamma = \tilde{\Gamma}$, and J belongs to the class of partly smooth functions relative to linear manifolds \mathcal{M} , i.e. $\mathcal{M} = T_\beta$. Typical instances encompassed in this class are the $\ell^1 - \ell^2$ norm, or its analysis version, as well as polyhedral gauges including the ℓ^∞ norm. Note that the nuclear norm (and composition of it with linear operators as proposed for instance in [8, 18]), whose manifold is not linear, does not fit into the framework of [21], while it is covered by Theorem 1.

4 Proofs

4.1 Proof of Theorem 1

In order to prove Theorem 1, we consider any sequence $\theta_k = (\mu_k, u_k = \Gamma_k x_0 + \varepsilon_k, \Gamma_k)_k$ where $X_k \in \mathbb{R}^{n_k \times p}$. Assume that

$$(\Gamma_k, \varepsilon_k \mu_k^{-1}, \mu_k) \longrightarrow (\tilde{\Gamma}, 0, 0). \quad (7)$$

Then proving Theorem 1 boils down to showing that for k large enough, the solution β_k of (\mathcal{P}_{θ_k}) is unique and satisfies $\beta_k \in \mathcal{M}$.

Constrained problem. We consider the following non-smooth, in general non-convex, constrained minimization problem

$$\beta_k \in \underset{\beta \in \mathcal{M} \cap \mathcal{K}}{\text{Argmin}} E(\beta, \theta_k) \quad (8)$$

where \mathcal{K} is an arbitrary fixed convex compact neighbourhood of β_0 .

The following lemma first show the convergence of β_k .

Lemma 1. *Under condition (7), $\beta_k \rightarrow \beta_0$.*

Proof. We denote $\|u\|_{\tilde{\Gamma}}^2 = \langle \Gamma u, u \rangle$ for any non-negative definite matrix Γ . We first note that (2) implies that β_0 is the unique solution of $(\mathcal{P}_{0, \tilde{\Gamma}\beta_0, \tilde{\Gamma}})$. By optimality of β_k one has $E(\beta_k, \theta_k) \leq E(\beta_0, \theta_k)$ and hence

$$\frac{1}{2}\|\beta_k\|_{\Gamma_k} - \langle \beta_k, \Gamma_k\beta_0 + \varepsilon_k \rangle + \mu_k J(\beta_k) \leq \frac{1}{2}\|\beta_0\|_{\Gamma_k} - \langle \beta_0, \Gamma_k\beta_0 + \varepsilon_k \rangle + \mu_k J(\beta_0)$$

which is equivalently stated as

$$\frac{1}{2}\|\beta_k - \beta_0\|_{\Gamma_k}^2 - \langle \beta_k - \beta_0, \varepsilon_k \rangle + \mu_k J(\beta_k) \leq \mu_k J(\beta_0). \quad (9)$$

Since $\beta_k \in \mathcal{K}$, the sequence $(x_k)_k$ is bounded, and we let β^* be any accumulation point. Taking the limit $k \rightarrow +\infty$ in (9) and using (7) and continuity of the inner product shows that $\tilde{\Gamma}\beta^* = \tilde{\Gamma}\beta_0$. Furthermore, since $\frac{1}{2}\|\beta_k - \beta_0\|_{\Gamma_k}^2 \geq 0$, (9) yields $-\langle \beta_k - \beta_0, \frac{\varepsilon_k}{\mu_k} \rangle + J(\beta_k) \leq J(\beta_0)$. Taking the limit $k \rightarrow +\infty$ shows that $J(\beta^*) \leq J(\beta_0)$. Combining this with the previous claim that β^* is a feasible point of $(\mathcal{P}_{0, \tilde{\Gamma}x_0, \tilde{\Gamma}})$ allows to conclude that β^* is a solution of $(\mathcal{P}_{0, \tilde{\Gamma}x_0, \tilde{\Gamma}})$. Since β_0 is unique, this leads to $\beta^* = \beta_0$. \square

We now aim at showing that for k large enough, β_k is the unique solution of (\mathcal{P}_{θ_k}) . In order to do so, we make use of the following classical result, whose proof can be found for instance in [22].

Proposition 2. *Let $\beta \in \mathbb{R}^p$ such that $\frac{u - \Gamma\beta}{\mu} \in \text{ri}(\partial J(\beta))$ and $\ker(\Gamma) \cap T_\beta = \{0\}$. Then β is the unique solution of (\mathcal{P}_θ) .*

Convergence of the tangent model subspace. By definition of the constrained problem (8), $\beta_k \in \mathcal{M}$. Moreover, since $E(\cdot, \theta_k)$ is partly smooth at β_0 relative to \mathcal{M} , the sharpness property Definition 1(ii) holds at all nearby points in the manifold \mathcal{M} [12, Proposition 2.10]. Thus as soon as k is large enough, we have $T_k = \mathcal{T}_{\beta_k}(\mathcal{M})$. Using the fact that \mathcal{M} is of class C^2 , we get

$$T_k = \mathcal{T}_{\beta_k}(\mathcal{M}) \longrightarrow \mathcal{T}_{\beta_0}(\mathcal{M}) = T \quad (10)$$

when (7) holds, 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 $P_{T_k} \rightarrow P_T$). Since $\ker(\tilde{\Gamma}) \cap T = \{0\}$, (10) implies that for k large enough, when (7) holds,

$$\ker(\Gamma_k) \cap T_k = \{0\}, \quad (11)$$

which we assume from now on.

First order condition. Let's take $\mathcal{K} = \mathbb{B}_r(\beta_0)$ for r sufficiently large. For any $\delta > 0$, $\exists K_\delta > 0$ such that $\forall k > K_\delta, \beta_k \in \mathbb{B}_\delta(\beta_0)$. Thus, for k large enough, i.e. δ sufficiently small, we indeed have $\beta_k \in \text{int}(\mathcal{K})$. Furthermore, it is easy to see that $\iota_{\mathcal{K}}$ is locally partly smooth at β_0 relative to \mathcal{K} . Since J is also locally partly smooth at β_0 relative to \mathcal{M} , the sum rule [12, Corollary 4.6] shows that, for all sufficiently large k , when (7) holds and $\beta_k \in \text{int}(\mathcal{K})$, $J + \iota_{\mathcal{K}}$ is locally partly smooth at β_k relative to $\mathcal{M} \cap \mathcal{K}$, and then so is $E(\cdot, \theta_k) + \iota_{\mathcal{K}}$ by the smooth perturbation rule [12, Corollary 4.7]. Therefore, [12, Proposition 2.4(a)-(b)] applies, and it follows that β_k is a critical point of (8) if, and only if,

$$0 \in \text{Aff}(\partial E(\beta_k, \theta_k) + N_{\mathcal{K}}(\beta_k)) = \frac{\Gamma_k\beta_k - u_k}{\mu_k} + \text{Aff}(\partial J(\beta_k)) = \frac{\Gamma_k\beta_k - u_k}{\mu_k} + e_{\beta_k} + T_k^\perp.$$

The first equality comes from the fact that $E(\cdot, \theta)$ is a closed convex function, and that the normal cone of \mathcal{K} at β_k vanishes on the interior points of \mathcal{K} , and the second one from the decomposability of the subdifferential. Projecting this relation onto T_k , we get, since $e_{\beta_k} \in T_k$,

$$P_{T_k}(\Gamma_k\beta_k - u_k) + \mu_k e_{\beta_k} = 0. \quad (12)$$

Convergence of the primal variables. Since both β_k and β_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} [16]. In particular, $\beta_k = P_{\mathcal{M}}(\beta_k)$ and $\beta_0 = P_{\mathcal{M}}(\beta_0)$. Moreover, $P_{\mathcal{M}}$ is of class C^1 near β_k [14, Lemma 4]. Thus, C^2 differentiability shows that

$$\beta_k - \beta_0 = P_{\mathcal{M}}(\beta_k) - P_{\mathcal{M}}(\beta_0) = \text{DP}_{\mathcal{M}}(\beta_k)(\beta_k - \beta_0) + R(\beta_k)$$

where $R(\beta_k) = O(\|\beta_k - \beta_0\|^2)$ and where $DP_{\mathcal{M}}(\beta_k)$ is the derivative of $P_{\mathcal{M}}$ at β_k . Using [14, Lemma 4], and recalling that $T_k = \mathcal{T}_{\beta_k}(\mathcal{M})$ by the sharpness property, the derivative $DP_{\mathcal{M}}(\beta_k)$ is given by $DP_{\mathcal{M}}(\beta_k) = P_{T_k}$. Inserting this in (12), we get

$$P_{T_k} \Gamma_k (P_{T_k}(\beta_k - \beta_0) + R(\beta_k)) - P_{T_k} \varepsilon_k + \mu_k e_{\beta_k} = 0.$$

Using (11), Γ_{k, T_k} has full rank, and thus

$$\beta_k - \beta_0 = \Gamma_{k, T_k}^+ (\varepsilon_k - \mu_k e_{\beta_k} - \Gamma_k R(\beta_k)), \quad (13)$$

where we also used that $T_k^\perp \subset \ker(\Gamma_{k, T_k}^+)$. One has $\Gamma_{k, T_k}^+ \rightarrow \tilde{\Gamma}^+$ so that $\Gamma_{k, T_k}^+ \Gamma_k = O(1)$ and $\Gamma_{k, T_k}^+ = O(1)$. Altogether, we thus obtain the bound

$$\|\beta_k - \beta_0\| = O(\|\varepsilon_k\|, \mu_k). \quad (14)$$

Convergence of the dual variables. We define $\eta_k = \frac{u_k - \Gamma_k \beta_k}{\mu_k}$. Arguing as above, and using (13) we have

$$\begin{aligned} \mu_k \eta_k &= \varepsilon_k + \Gamma_k (\beta_0 - \beta_k) = \varepsilon_k - \Gamma_k \Gamma_{k, T_k}^+ (\varepsilon_k - \mu_k e_{\beta_k} - \Gamma_k R(\beta_k)) \\ &= \varepsilon_k - \Gamma_k P_{T_k} \Gamma_{k, T_k}^+ (\varepsilon_k - \mu_k e_{\beta_k} - \Gamma_k R(\beta_k)) \\ &= P_{V_{T_k}^\perp} \varepsilon_k + P_{V_{T_k}} \Gamma_k R(\beta_k) + \mu_k \Gamma_k \Gamma_{k, T_k}^+ e_{\beta_k}, \end{aligned}$$

where we denoted $V_{T_k} = \text{Im}(\Gamma_k P_{T_k})$, and used that $\text{Im}(\Gamma_{k, T_k}^+) \subset T_k$. We thus arrive at

$$\|\eta_k - \eta_{\tilde{\Gamma}}\| = O\left(\|\varepsilon_k\| \mu_k^{-1}, \|\Gamma_k \Gamma_{k, T_k}^+ e_{\beta_k} - \eta_{\tilde{\Gamma}}\|, \|\Gamma_k\| \|\beta_k - \beta_0\|^2 \mu_k^{-1}\right).$$

Since \mathcal{M} is a C^2 manifold, and by partial smoothness (J is C^2 on \mathcal{M}), we have $\beta \mapsto e_\beta$ is C^1 on \mathcal{M} , one has

$$\|e_{\beta_k} - e\| = O(\|\beta_k - \beta_0\|). \quad (15)$$

Using the triangle inequality, we get

$$\|\Gamma_k \Gamma_{k, T_k}^+ - \tilde{\Gamma} \tilde{\Gamma}_T^+\| \leq \|\Gamma_{k, T_k}^+\| \|\Gamma_k - \tilde{\Gamma}\| + \|\tilde{\Gamma}\| \|\Gamma_{k, T_k}^+ - \tilde{\Gamma}_T^+\|.$$

Again, since $\Gamma_{k, T_k}^+ \rightarrow \tilde{\Gamma}_T^+$, we have $\|\Gamma_{k, T_k}^+\| = O(1)$. Moreover, $A \mapsto A^+$ is smooth at $A = \Gamma_T$ along the manifold of matrices of constant rank, and \mathcal{M} is a C^2 manifold near β_0 . Thus

$$\|\Gamma_{k, T_k}^+ - \tilde{\Gamma}_T^+\| = O(\|\Gamma_{k, T_k} - \tilde{\Gamma}_T\|) = O(\|\Gamma_k - \tilde{\Gamma}\|, \|P_{T_k} - P_T\|) = O(\|\Gamma_k - \tilde{\Gamma}\|, \|\beta_k - \beta_0\|).$$

This shows that

$$\|\Gamma_k \Gamma_{k, T_k}^+ - \tilde{\Gamma} \tilde{\Gamma}_T^+\| = O(\|\Gamma_k - \tilde{\Gamma}\|, \|\beta_k - \beta_0\|). \quad (16)$$

Putting (15) and (16) together implies $\|\Gamma_k \Gamma_{k, T_k}^+ e_{\beta_k} - \eta_{\tilde{\Gamma}}\| = O(\|\Gamma_k - \tilde{\Gamma}\|, \|\beta_k - \beta_0\|)$. Altogether, we get the bound

$$\|\eta_k - \eta_{\tilde{\Gamma}}\| = O\left(\|\varepsilon_k\| \mu_k^{-1}, \|\beta_k - \beta_0\|, \|\Gamma_k - \tilde{\Gamma}\|, \|\Gamma_k\| \|\beta_k - \beta_0\|^2 \mu_k^{-1}\right).$$

Since $\|\beta_k - \beta_0\|$ is bounded according to (14), we arrive at

$$\|\eta_k - \eta_{\tilde{\Gamma}}\| = O\left(\|\Gamma_k - \tilde{\Gamma}\|, \|\varepsilon_k\| \mu_k^{-1}, \mu_k\right). \quad (17)$$

Convergence inside the relative interior. Using the hypothesis that $\eta_{\tilde{\Gamma}} \in \text{ri}(\partial J(\beta_0))$, we will show that for k large enough,

$$\eta_k \in \text{ri}(\partial J(\beta_k)). \quad (18)$$

Let us suppose this does not hold. Then there exists a sub-sequence of η_k , that we do not relabel for the sake of readability of the proof, such that

$$\eta_k \in \text{rbound}(\partial J(\beta_k)). \quad (19)$$

According to (17) and Lemma 1, under (7), $(\beta_k, \eta_k) \rightarrow (\beta_0, \eta_{\tilde{\Gamma}})$. Condition (19) is equivalently stated as, for each k

$$\exists z_k \in T_{\beta_k}^\perp, \quad \forall \eta \in \partial J(\beta_k), \quad \langle z_k, \eta - \eta_k \rangle \geq 0, \quad (20)$$

where one can impose the normalization $\|z_k\| = 1$ by positive-homogeneity. Up to a sub-sequence (that for simplicity we still denote z_k with a slight abuse of notation), since z_k is in a compact set, we can assume z_k approaches a non-zero cluster point z^* .

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

$$\forall v \in \partial J(\beta_0), \quad \langle z^*, \eta - \eta_{\tilde{\Gamma}} \rangle \geq 0. \quad (21)$$

Indeed, let us consider any $v \in \partial J(\beta_0)$. In view of the continuity property in Definition 1(iii) ∂J is continuous at β_0 along \mathcal{M} , so that since $\beta_k \rightarrow \beta_0$ there exists $v_k \in \partial J(\beta_k)$ with $v_k \rightarrow v$. Applying (20) with $\eta = v_k$ gives $\langle z_k, v_k - \eta_k \rangle \geq 0$. Taking the limit $k \rightarrow +\infty$ in this inequality leads to (21), which contradicts the fact that $\eta_{\tilde{\Gamma}} \in \text{ri}(\partial J(\beta_0))$. In view of (18) and (11), using Proposition 2 shows that β_k is the unique solution of (\mathcal{P}_θ) . \square

4.2 Proof of Theorem 2

It is sufficient to check that (3) is in force with probability 1 as $n \rightarrow +\infty$. Owing to classical results on convergence of sample covariances, which apply thanks to the assumption that the fourth order moments are finite, we get $\Gamma_n - \tilde{\Gamma} = O_P(n^{-1/2})$ and $\frac{1}{n} \langle \xi_i, w \rangle = O_P(n^{-1/2})$, where used the assumption that $\mathbb{E}(\langle \xi_i, w \rangle) = 0$. As p is fixed, it follows that $\|\Gamma_n - \tilde{\Gamma}\| = O_P(n^{-1/2})$ and $\|\varepsilon_n\| = O_P(n^{-1/2})$. Thus under the scaling (6), we get

$$\begin{aligned} \left(\|\Gamma_n - \tilde{\Gamma}\|, \|\varepsilon_n\| \mu_n^{-1}, \mu_n \right) &= \left(O_P(n^{-1/2}), \frac{1}{\mu_n n^{1/2}} O_P(1), o(1) \right) \\ &= \left(O_P(n^{-1/2}), o(1) O_P(1), o(1) \right) = \left(O_P(n^{-1/2}), o(1), o(1) \right), \end{aligned}$$

which indeed converges to 0 in probability. This concludes the proof. \square

4.3 Proof of Proposition 1

Let β_k be a solution of (\mathcal{P}_{θ_k}) . Suppose that $\beta_k \in \mathcal{M}$. In particular, β_k is a solution of the non-convex minimization (8). Arguing as in the proof of Theorem 1, we get the bound (17), i.e.

$$\|\eta_k - \eta_{\tilde{\Gamma}}\| = O(\|\Gamma_k - \tilde{\Gamma}\|, \|\varepsilon_k\| / \mu_k, \mu_k) \quad \text{where} \quad \eta_k = \frac{u_k - \Gamma_k \beta_k}{\mu_k}. \quad (22)$$

In particular, $\|\eta_k - \eta_{\tilde{\Gamma}}\| \rightarrow 0$. Defining $K = d(\eta_{\tilde{\Gamma}}, \partial J(\beta))$, one has $K > 0$ since $\eta_{\tilde{\Gamma}} \notin \text{ri} \partial J(\beta_0)$. Choosing k large enough, the convergence of η_k to $\eta_{\tilde{\Gamma}}$ implies that

$$d(\eta_k, \partial J(\beta_0)) > K/2 \quad (23)$$

where 2 can be changed to any arbitrary value. Using the outer semi-continuity of the subdifferential, we get that

$$\forall \varepsilon, \exists k_0, \forall k \geq k_0, \quad \partial J(x_k) \subseteq \partial J(\beta_0) + B(0, \varepsilon).$$

In particular, $\eta_k \in \partial J(\beta_0) + B(0, \varepsilon)$ which implies that $d(\eta_k, \partial J(\beta_0)) \leq \varepsilon$, which is a contradiction to (23). Hence, $\beta_k \notin \mathcal{M}$.

5 Conclusion

In this paper, we provided a unified analysis of the recovery performance when partly smooth functions are used to regularize linear inverse problems. This class of functions encompass all popular regularizers used in the literature. A distinctive feature of our work is that we provided for the first time a unified analysis together with a generalized ‘‘irrepresentable condition’’ to guarantee stable and correct identification of the low-complexity manifold underlying the original object.

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).

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] E. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.
- [4] 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.
- [5] 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.
- [6] A. Daniilidis, D. Drusvyatskiy, and A. S. Lewis. Orthogonal invariance and identifiability. *to appear in SIAM J. Matrix Anal. Appl.*, 2014.
- [7] J.J. Fuchs. On sparse representations in arbitrary redundant bases. *IEEE Transactions on Information Theory*, 50(6):1341–1344, 2004.
- [8] 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.
- [9] W.L. Hare and A.S. Lewis. Identifying active constraints via partial smoothness and prox-regularity. *J. Convex Anal.*, 11(2):251–266, 2004.
- [10] J. Jia and B. Yu. On model selection consistency of the elastic net when $p \gg n$. *Statistica Sinica*, 20:595–611, 2010.
- [11] K. Knight and W. Fu. Asymptotics for Lasso-Type Estimators. *The Annals of Statistics*, 28(5):1356–1378, 2000.
- [12] A. S. Lewis. Active sets, nonsmoothness, and sensitivity. *SIAM Journal on Optimization*, 13(3):702–725, 2003.
- [13] A. S. Lewis. The mathematics of eigenvalue optimization. *Mathematical Programming*, 97(1–2):155–176, 2003.
- [14] A. S. Lewis and J. Malick. Alternating projections on manifolds. *Mathematics of Operations Research*, 33(1):216–234, 2008.
- [15] 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.
- [16] R.A. Poliquin, R.T. Rockafellar, and L. Thibault. Local differentiability of distance functions. *Trans. Amer. Math. Soc.*, 352:5231–5249, 2000.
- [17] 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.
- [18] 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.
- [19] L.I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992.
- [20] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B. Methodological*, 58(1):267–288, 1996.
- [21] S. Vaïter, M. Golbabaee, J. Fadili, and G. Peyré. Model selection with piecewise regular gauges. Technical report, arXiv:1307.2342, 2013.
- [22] S. Vaïter, G. Peyré, C. Dossal, and M.J. Fadili. Robust sparse analysis regularization. *IEEE Transactions on Information Theory*, 59(4):2001–2016, 2013.
- [23] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2005.
- [24] P. Zhao and B. Yu. On model selection consistency of Lasso. *J. Mach. Learn. Res.*, 7:2541–2563, December 2006.