Model Consistency of Partly Smooth Regularizers

Samuel Vaiter, Gabriel Peyré, Jalal M. Fadili

Model Consistency of Partly Smooth Regularizers

Samuel Vaiter VAITER@CMAP.POLYTECHNIQUE.FR
CMAP, CNRS-École Polytechnique

Gabriel Peyré PEYRE@CEREMADE.DAUPHINE.FR
CNRS and CEREMADE, Université Paris Dauphine

Jalal Fadili JALAL.FADILI@GREYC.ENSICAEN.FR
GREYC, CNRS-ENSICAEN

Editor: –

Abstract

This paper studies least-square regression penalized with partly smooth convex regularizers. This class of penalty functions is very large and versatile, and allows to promote solutions conforming to some notion of low-complexity. Indeed, such penalties/regularizers force the corresponding solutions to belong to a low-dimensional manifold (the so-called model) which remains stable when the penalty function undergoes small perturbations. Such a good sensitivity property is crucial to make the underlying low-complexity (manifold) model robust to small noise. In a deterministic setting, 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. We also prove that this condition is almost necessary for stable model recovery. We then turn to the random setting where the design matrix and the noise are random, and the number of observations grows large. We show that under our generalized "irrepresentable condition", and a proper scaling of the regularization parameter, the regularized estimator is model consistent. In plain words, with a probability tending to one as the number of measurements tends to infinity, the regularized estimator belongs to the correct low-dimensional model manifold. This work unifies and generalizes a large body of literature, where model consistency was known to hold, for instance for the Lasso, group Lasso, total variation (fused Lasso) and nuclear/trace norm regularizers. We show that under the deterministic model selection conditions, the forward-backward proximal splitting algorithm used to solve the penalized least-square regression problem, is guaranteed to identifiy the model manifold after a finite number of iterations. Lastly, we detail how our results extend from the quadratic loss to an arbitrary smooth and strictly convex loss function. We illustrate the usefulness of our results on the problem of low-rank matrix recovery from random measurements using nuclear norm minimization.

Keywords: Regularization, regression, inverse problems, model consistency, partial smoothness, sensitivity analysis, sparsity, low-rank. [1]

[1] Samuel Vaiter was affiliated with CNRS and CEREMADE, Université Paris Dauphine when this work was completed.

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 proper, lower-semicontinuous (lsc) and convex function $J : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ to promote such a prior. Without loss of generality, we also assume that J is non-negative. This then leads to solving the following convex optimization problem

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

where $\lambda > 0$ is the so-called regularization parameter used to balance the amount of regularization and loss.

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 that $y \in \mathrm{Im}(X)$ and thus $u \in \mathrm{Im}(\Gamma)$. Obviously, this does not entail any loss of generality, as the loss term can always be written as $\frac{1}{2\lambda}\|X\beta - \mathrm{P}_{\mathrm{Im}(X)}\, y\|^2 + \frac{1}{2\lambda}\|\mathrm{P}_{\mathrm{Im}(X)^\perp}\, y\|^2$, where P_T the orthogonal projection on T.

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

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

where A^+ stands for the Moore-Penrose pseudo-inverse of a matrix A. With these notations, E is a function on $\mathbb{R}^p \times \Theta$.

We also consider the constrained problem

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

where $\theta_0 = (0, u, \Gamma)$, and $\iota_{\mathcal{C}}$ is the indicator function of the non-empty closed convex set $\mathcal{C}$, i.e. $\iota_{\mathcal{C}}(\beta) = 0$ if $\beta \in \mathcal{C}$ and $\iota_{\mathcal{C}}(\beta) = +\infty$ otherwise. Problem ($\mathcal{P}_{\theta_0}$) can be viewed as a limit of ($\mathcal{P}_\theta$) as $\mu \to 0^+$.

At this stage, it is worth mentioning that though we focus here, for simplicity of exposition, on the squared loss $x \mapsto \frac{1}{2}\|y - Xx\|^2$, our results generalize to more general smooth losses, see Section 3.4 for further details.

The goal of this paper is to assess the recovery performance of $(\mathcal{P}_\theta)$, i.e. to understand how close are the properties of the recovered solution of $(\mathcal{P}_\theta)$ to those of β_0. More precisely, we focus here on the low-noise regime, i.e. when ε is small enough, and we investigate stability in ℓ^2 sense, but also, and more importantly, the identifiability of the correct low-dimensional manifold associated to β_0. This unifies and extend a large body of literature, including sparsity and low-rank regularization, which turn to be a very special case of the powerful theory of partly smooth regularization.

1.2 Notations

We recall some basic ingredients from differential geometry that are essential to our exposition. A good source on smooth manifold theory is (Lee, 2003). A set $\mathcal{M} \subset \mathbb{R}^p$ is a C^2-smooth manifold around a point $\beta \in \mathbb{R}^p$, if $\beta \in \mathcal{M}$ and $\mathcal{M}$ consists locally around β of the solutions of some C^2-smooth equations with linearly independent gradients. In this case, the tangent space of $\mathcal{M}$ at β is denoted $\mathcal{T}_\beta(\mathcal{M})$. We define the tangent model subspace as

$$T_\beta = \operatorname{VectHull}(\partial J(\beta))^\perp,$$

where $\operatorname{VectHull}(\mathcal{C}) = \left\{\rho(\beta - \beta') \; ; \; (\beta, \beta') \in \mathcal{C}^2, \rho \in \mathbb{R}\right\}$ is the linear hull of the convex set $\mathcal{C} \subset \mathbb{R}^p$. We denote $\operatorname{ri}(\mathcal{C})$ (resp. $\operatorname{rbd}(\mathcal{C})$)) the relative interior of $\mathcal{C}$ (resp. relative boundary), i.e. its interior (boundary) for the topology of its affine hull (the smallest affine space containing $\mathcal{C}$).

For a function J, $\operatorname{dom}(J) = \{\beta \in \mathbb{R}^p \; ; \; J(\beta) < +\infty\}$ is its domain. We denote $\partial J(\beta)$ the subdifferential at β of the proper, lsc and convex function J. Geometrically, when $\beta \in \operatorname{dom}(J)$, $\partial J(\beta)$ (if non-empty) is the set of gradients of the affine minorants of J supporting it at β. The subdifferential $\partial J(\beta)$ is a closed convex set.

For a linear space T, we denote P_T the orthogonal projection on T and for a matrix $\Gamma \in \mathbb{R}^{p \times p}$, $\Gamma_T = \mathrm{P}_T \, \Gamma \, \mathrm{P}_T$.

2. Partly-smooth Functions

Toward the goal of studying the recovery guarantees of problem $(\mathcal{P}_\theta)$, our central assumption will be that J is a partly smooth function. Partial smoothness of functions was originally defined by Lewis (2003a). Our definition hereafter specializes it to the case of proper, lsc and convex functions.

Definition 1 *Let J be a proper, lsc convex function, and $\beta \in \mathbb{R}^p$ such that $\partial J(\beta) \neq \emptyset$. 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 thanks to prox-regularity of proper, lsc and convex functions, see (Hare and Lewis, 2004, Corollary 4.2).

Loosely speaking, a partly smooth function behaves smoothly as we move on the identifiable manifold, and sharply if we move normal to the manifold.

Remark 2 (Discussion of the properties) *Since J is proper, lsc and convex, it is subdifferentially regular at any point in its domain, and in particular at β. Therefore, the Clarke regularity property (Lewis, 2003a, Definition 2.7(ii)) is automatically verified. In view of (Lewis, 2003a, Proposition 2.4(i)-(iii)), the sharpness property* ((ii)) *is equivalent to (Lewis, 2003a, 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, signal and image processing. We first expose basic building blocks (sparsity, group sparsity) and then show how the machinery of partial smoothness enables a powerful calculus to create new priors (using post-composition with a linear operator, spectral lifting, positive linear combinations and separable priors).

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

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

The use of sparse regularizations has been popularized in the signal processing literature under the name basis pursuit method (Chen et al., 1999) and in the statistics literature under the name Lasso (Tibshirani, 1996).

$\ell^1 - \ell^2$ 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 $\bigcup_{b \in \mathcal{B}} b = \{1, \ldots, p\}$. 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 (Yuan and Lin, 2005; Bach, 2008b) 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

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

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 $\beta = U_\beta \operatorname{diag}(\Lambda_\beta) V_\beta^*$ an SVD decomposition of β, where $\Lambda_\beta \in \mathbb{R}_+^{p_0}$. The nuclear norm is defined as

$$J(\beta) = \|\beta\|_* = \|\Lambda_\beta\|_1. \tag{2}$$

It has been used for instance in machine learning applications (Bach, 2008b), matrix completion (Recht et al., 2010; Candès and Recht, 2009) and phase retrieval (Candès et al., 2013). The nuclear norm can be shown to be partly smooth at x relative to the manifold (Lewis and Malick, 2008, Example 2)

$$\mathcal{M} = \left\{ \beta' \; ; \; \operatorname{rank}(\beta') = \operatorname{rank}(\beta) \right\}. \tag{3}$$

More generally, if $j : \mathbb{R}^{p_0} \to \mathbb{R}$ 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 (Lewis, 2003b). 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

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

where $\mathcal{O}_{p_0} \subset \mathbb{R}^{p_0 \times p_0}$ is the group of orthogonal matrices. This result is proved in (Daniilidis et al., 2014a, Theorem 3.19), building upon the work of (Daniilidis et al., 2014b) on manifold smoothness transfer under spectral lifting. This result can be extended to non-symmetric (possibly rectangular) matrices by requiring that j is an absolutely permutation-invariant closed convex function, see (Daniilidis et al., 2014a, Theorem 5.3). The nuclear norm $\|\cdot\|_*$ is a special case where $j(\Lambda) = \|\Lambda\|_1$.

Analysis regularizers. If $J_0 : \mathbb{R}^q \to \mathbb{R} \cup \{+\infty\}$ is a proper lsc convex function and $D \in \mathbb{R}^{p \times q}$ is a linear operator, an analysis regularizer (following the terminology introduced in (Elad et al., 2007)) is of the form

$$J(\beta) = J_0(D^* \beta).$$

Such a prior controls the low-complexity (as measured by J_0) of the correlations between the columns of D and β. A popular example is when taking $J_0 = \|\cdot\|_1$ and D^* 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 (Rudin et al., 1992). The fused Lasso (Tibshirani et al., 2005) corresponds to J_0 being the ℓ^1-norm and D^* is the concatenation of the identity and a finite-difference operator. To cope with correlated covariates in linear regression, it was devised in (Grave et al., 2011; Richard et al., 2013) to use a family of analysis-type priors where $J_0 = \|\cdot\|_*$ is the nuclear norm.

If J_0 is partly smooth at $\alpha = D^* \beta$ for the manifold $\mathcal{M}_\alpha^0$, then it is shown in (Lewis, 2003a, Theorem 4.2) that J is partly smooth at β relative to the manifold

$$\mathcal{M} = \left\{ \beta' \in \mathbb{R}^p \; ; \; D^* \beta' \in \mathcal{M}_\alpha^0 \right\}.$$

provided that the following transversality condition holds (Lee, 2003, Theorem 6.30(a))

$$\operatorname{Ker}(D) \cap \mathcal{T}_\alpha(\mathcal{M}_\alpha^0)^\perp = \{0\} \iff \operatorname{Im}(D^*) + \mathcal{T}_\alpha(\mathcal{M}_\alpha^0) = \mathbb{R}^N .$$

Moreover, the co-dimension of $\mathcal{M}$ in $\mathbb{R}^p$ equals the co-dimension of $\mathcal{M}_\alpha^0$ in $\mathbb{R}^q$.

Mixed regularization. Starting from a family of proper, lsc and convex functions $\{J_\ell\}_{\ell\in\mathcal{L}}$, $\mathcal{L} = \{1,\dots,L\}$, 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, by using $J_1 = \|\cdot\|_1$ and $J_2 = \|\cdot\|_*$, see for instance (Golbabaee and Vandergheynst, 2012; Oymak et al., 2012).

Suppose that $\bigcap_{\ell\in\mathcal{L}} \operatorname{ri}(\operatorname{dom}(J_\ell)) \neq \emptyset$. Let $\mathcal{S} \subseteq \mathbb{R}^p$ be a C^2-manifold. If each J_ℓ is partly smooth at β relative to a submanifold $\mathcal{M}^\ell \subseteq \mathcal{S}$, then it can be shown that J is also partly smooth at β for

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

with the proviso that the submanifolds $\mathcal{M}^\ell$ intersect transversally, i.e.

$$\sum_{\ell\in\mathcal{L}} z_\ell = 0 \quad \text{and} \quad z_\ell \in \mathcal{T}_\beta(\mathcal{M}^\ell)^\perp \text{ for each } \ell\in\mathcal{L} \Rightarrow z_\ell \in \mathcal{T}_\beta(\mathcal{S})^\perp \text{ for each } \ell\in\mathcal{L} .$$

Moreover, the co-dimension of $\mathcal{M}$ (in $\mathcal{S}$) equals the sum of the co-dimensions of $\mathcal{M}^\ell$. This assertion is a weaker version of (Lewis, 2003a, Corollary 4.8), since we use convexity and closedness of the functions J_ℓ. For the case where $\mathcal{L} = 2$, the above transversality condition reads (Lee, 2003, Theorem 6.30(b))

$$\mathcal{T}_\beta(\mathcal{M}_1)^\perp \cap \mathcal{T}_\beta(\mathcal{M}_2)^\perp = \mathcal{T}_\beta(\mathcal{S})^\perp \iff \mathcal{T}_\beta(\mathcal{M}_1) + \mathcal{T}_\beta(\mathcal{M}_2) = \mathcal{T}_\beta(\mathcal{S}) . \tag{4}$$

Separable Regularization. Let $\{J_\ell\}_{\ell\in\mathcal{L}}$, $\mathcal{L} = \{1,\dots,L\}$, be a family of proper lsc convex functions. If J_ℓ is partly smooth at β_ℓ relative to a manifold $\mathcal{M}^\ell_{\beta_\ell}$, then the separable function

$$J\left(\{\beta_\ell\}_{\ell\in\mathcal{L}}\right) = \sum_{\ell\in\mathcal{L}} J_\ell(\beta_\ell)$$

is partly smooth at $(\beta_1,\dots,\beta_L)$ relative to $\mathcal{M}^1_{\beta_1} \times \dots \times \mathcal{M}^L_{\beta_L}$ (Lewis, 2003a, Proposition 4.5).

One fundamental problem that has attracted a lot of interest in the recent years in data processing involves decomposing an observed object into a linear combination of components/constituents β_ℓ, $\ell\in\mathcal{L}$. One instance of such a problem is image decomposition into texture and piece-wise-smooth (cartoon) parts, see e.g. (Starck et al., 2005; Aujol et al., 2005; Peyré et al., 2010) and references therein. Another example of decomposition is principal component pursuit, proposed in (Candès et al., 2011), to decompose a matrix which is the superposition of a low-rank component and a sparse component. In this case $J_1 = \|\cdot\|_1$ and $J_2 = \|\cdot\|_*$.

3. Main results

In the following, we denote $T = T_{\beta_0}$, $e = \mathrm{P}_T(\partial J(\beta_0)) \in \mathbb{R}^p$. 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 3 (Linearized pre-certificate) *For some matrix $\Gamma \in \mathbb{R}^{p\times p}$, assuming $\ker(\Gamma) \cap T = \{0\}$, we define $\eta_\Gamma = \Gamma\Gamma_T^+ e$.*

In Section 5.2, we will investigate the connection between η_Γ and the so-called minimal norm certificate.

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 (or equivalently λ) is well chosen.

Theorem 4 *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*

$$\ker(\tilde{\Gamma}) \cap T = \{0\}, \quad \textit{and} \quad \eta_{\tilde{\Gamma}} \in \mathrm{ri}(\partial J(\beta_0)). \tag{5}$$

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

$$\max\Big(\|\Gamma - \tilde{\Gamma}\|, \|\varepsilon\|\mu^{-1}, \mu\Big) \leqslant C, \tag{6}$$

the solution β_θ of $(\mathcal{P}_\theta)$ is unique and satisfies

$$\beta_\theta \in \mathcal{M} \quad \textit{and} \quad \|\beta_\theta - \beta_0\| = O(\|\varepsilon\|). \tag{7}$$

This theorem is proved in Section 5.3.

Remark 5 (Inverse problems) *A typical case of application of this result is in inverse problems that are encountered in various disciplines in science and engineering, such as in signal and image processing. In such a setting, the forward operator X is generally fixed and known, and one then takes $\tilde{\Gamma} = \Gamma = X^*X/n$.*

Remark 6 (Uncertain design/forward operator) *If only a noisy version of the forward operator (in inverse problems) or the design (in regression) is available then this can also be handled by Theorem 4. This scenario has been considered for sparse recovery (i.e. J the ℓ^1-norm) by several authors for sparse linear regression and compressed sensing, see e.g. (Herman and Strohmer, 2010; Rosenbaum and Tsybakov, 2010; Loh and Wainwright, 2012).*

Remark 7 (Random setting) *In statistics or machine learning, one considers a regression problem where the design X and the noise w are random, under the asymptotic regime where the number of observations n, i.e. number of rows X, grows large, so that Γ only reach $\tilde{\Gamma}$ in the limit $n \to +\infty$. See Theorem 11 for details.*

Remark 8 (Identification of the manifold) *Theorem 4 guarantees that, under some hypotheses on β_0 and θ, β_θ belongs to $\mathcal{M}$. For all the regularizations considered in Section 2.1, it turns out that actually $\mathcal{M}_{\beta_\theta} = \mathcal{M}$. This is because, for any (β, β') with $\beta' \in \mathcal{M}_\beta$ close enough to β, one has $\mathcal{M}_{\beta'} = \mathcal{M}_\beta$.*

The following proposition, proved in Section 5.6, shows that Theorem 4 is in some sense sharp, since the hypothesis $\eta_\Gamma \in \mathrm{ri}(\partial J(\beta_0))$ (almost) characterizes the stability of $\mathcal{M}$.

Proposition 9 *Suppose that β_0 is the unique solution of $(\mathcal{P}_{(0,\tilde{\Gamma}\beta_0,\tilde{\Gamma})})$ and that*

$$\ker(\tilde{\Gamma}) \cap T = \{0\}, \quad \textit{and} \quad \eta_{\tilde{\Gamma}} \notin \partial J(\beta_0). \tag{8}$$

Then there exists $C > 0$ such that if (6) *holds, then any solution β_θ of $(\mathcal{P}_\theta)$ 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 $\mathcal{M}$ cannot be correctly identified by solving $\mathcal{P}_{(\mu,\Gamma\beta_0,\Gamma)}$ for any $\mu > 0$ small enough.

Remark 10 (Critical case) *The only case not covered by either Theorem 4 or Proposition 9 is when $\eta_{\tilde{\Gamma}} \in \mathrm{rbd}(\partial J(\beta_0))$. In this case, one cannot conclude in general, since depending on the noise w, one can have either stability or non-stability of $\mathcal{M}$. We refer to (Vaiter et al., 2013b) where an example illustrates this situation for the 1-D total variation $J = \|D_{\mathrm{DIF}}^* \cdot\|_1$ (here D_{DIF}^* 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 $n \to \infty$. We consider that the design matrix and the noise are random. More precisely, the data (ξ_i, w_i) are random vectors in $\mathbb{R}^p \times \mathbb{R}$, $i = 1, \cdots, n$, 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 $\mathbb{E}\left(w_i | \xi_i\right) = 0$, finite fourth-order moments, i.e. $\mathbb{E}\left(w_i^4\right) < +\infty$ and $\mathbb{E}\left(\|\xi_i\|^4\right) < +\infty$. 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 strenghthening others, see e.g. (Knight and Fu, 2000; Zhao and Yu, 2006; Bach, 2008b). 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 $(\mathcal{P}_{\theta_n})$ for a sequence θ_n is said to be consistent for β_0 if, $\lim_{n \to +\infty} \Pr\left(\beta_{\theta_n} \text{ is unique}\right) \to 1$ and β_{θ_n} converges to β_0 in probability. The estimator is said to be model consistent if $\lim_{n \to +\infty} \Pr\left(\beta_{\theta_n} \in \mathcal{M}\right) \to 1$, 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 5.5

Theorem 11 *If conditions* (5) *hold and*

$$\mu_n = o(1) \quad \textit{and} \quad \mu_n^{-1} = o(n^{1/2}). \tag{9}$$

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

Remark 12 (Sharpness of the criterion) *One can also state a probabilistic equivalent to Proposition 8. That is, if β_0 is the unique solution of $(\mathcal{P}_{0,\tilde{\Gamma}\beta_0,\tilde{\Gamma}})$, and conditions* (8) *and* (9) *hold, then the estimator β_{θ_n} of β_0 defined by solving $(\mathcal{P}_{\theta_n})$ cannot be model consistent.*

3.3 Algorithmic Implications

A popular iterative 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 (Bauschke and Combettes, 2011). Starting from some $\beta_0 \in \mathbb{R}^p$, the algorithm implements the following iteration

$$\beta_{k+1} = \operatorname{Prox}_{\tau_k \mu J}\left(\beta_k + \tau_k(u - \Gamma\beta_k)\right),$$

where the step size satisfies $0 < \underline{\tau} \leqslant \tau_k \leqslant \overline{\tau} < 2/\|\Gamma\|$, and the proximity operator is defined, for $\gamma > 0$, as

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

The following theorem shows that the Forward-Backward algorithm correctly identifies the manifold $\mathcal{M}$ after a finite number of iterations.

Theorem 13 *Suppose that the assumptions of Theorem 4 hold. Then, there exists k_0 large enough, such that for all $k \geqslant k_0$, the Forward-Backward iterates satisfy $\beta_k \in \mathcal{M}$.*

Proof Inspection of the proof of Theorem 4 shows that the solution β_θ of $(\mathcal{P}_\theta)$, which is unique, is such that the vector $\eta_\theta = \frac{u - \Gamma\beta_\theta}{\lambda}$ satisfies $\eta_\theta \in \operatorname{ri}(\partial J(\beta_\theta))$ when (5) and (6) hold. Moreover, as $\beta_\theta \in \mathcal{M}$, β_θ is near β_0, and J is locally partly smooth at β_0 relative to $\mathcal{M}$, it is also partly smooth at β_θ relative to the same manifold $\mathcal{M}$. Altogether, this implies that the assumptions of (Liang et al., 2014, Theorem 3.1) are fulfilled and the manifold identification claim follows. ■

This result sheds light on the convergence behaviour of this algorithm in the favourable case where condition (5) holds and $(\|\Gamma - \tilde{\Gamma}\|, \|\varepsilon\|/\mu, \mu)$ are sufficiently small.

3.4 General Loss Functions

For the sake of simplicity, we have described our contributions with the squared loss function $u \in \mathbb{R}^n \mapsto \frac{1}{2}\|y - u\|^2$. Our results, however, extend readily to the case of more general loss functions of the form $F(u, y)$. In the following $\nabla_1^2 F(u, y)$ denotes the Hessian of F with respect to the first variable evaluated at (u, y).

We thus consider the variational problem

$$\min_\beta F(X\beta, y) + \lambda J(\beta),$$

where the loss function $F : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ fulfills the following assumptions:

(A.1) For any $y \in \mathbb{R}^n$, $F(\cdot, y) \in C^2(\mathbb{R}^n)$ is σ_{m}-strongly convex and $\nabla_1 F(\cdot, y)$ is σ_{M}-Lipschitz continuous, $\sigma_{\mathrm{m}} > 0$ and $\sigma_{\mathrm{M}} > 0$.

(A.2) The gradient of F with respect to the first variable, $\nabla_1 F(u, y)$, is such that $\nabla_1 F(u, u) = 0$.

Any loss function of the form $F(u,y) = G(u) - \langle u,\, y\rangle$, where G is a C^2 strongly convex function and and its gradient is Lipschitz-continuous, satisfies assumptions **(A.1)**. Assumption **(A.2)** is quite natural for a data fidelity term, and is fulfilled for instance for some losses in the exponential family.

In this setting, Theorem 4 (and in a similar way our other contributions) remains valid, and one simply needs to replace condition (6) by

$$\max\left(\|\check\Gamma - \tilde\Gamma\|, \|\check\varepsilon\|\mu^{-1}, \mu\right) \leqslant C, \tag{10}$$

where now

$$\check\Gamma = \frac{1}{n} X^* \nabla_1^2 F(y,y) X \quad \text{and} \quad \check\varepsilon = \frac{1}{n} X^* \nabla_1^2 F(y,y) w \ ,$$

where $\nabla_1^2 F(y,y)$ is the Hessian with respect to the first variable (assumed to be positive definite by assumption **(A.1)**) taken at (y,y). A detailed treatment on the way to adapt the proofs to handle such a generic loss is provided in Section 5.4.

3.5 Relation to Previous Works

Works on linear convergence rates. Following the pioneer work (Burger and Osher, 2004) (who study convergence in terms of Bregman divergence), there is a large amount of works on the study conditions under which $\|\beta_\theta - \beta_0\| = O(\|\varepsilon\|)$ (so-called linear convergence rate) where β_θ is any solution of $(\mathcal{P}_\theta)$, see for instance the book (Scherzer et al., 2009) for an overview of these results. The initial work of (Grasmair et al., 2011) proves a sharp criteria to ensure linear convergence rate for the ℓ^1 norm, and this approach is further extended to arbitrary convex functions by (Grasmair, 2011) and (Fadili et al., 2013), who respectively proved linear convergence rates in terms of the penalty J and ℓ^2-norm.

These works show that if

$$\ker(\Gamma) \cap T = \{0\} \quad \text{and} \quad \exists \eta \in \mathrm{Im}(\Gamma) \cap \mathrm{ri}(\partial J(\beta_0)) \tag{11}$$

(which is often called the source condition), then linear convergence rate holds. Note that condition (5) implies (11), but it is stronger. Indeed, condition (11) does not ensure model consistency (7), which is a stronger requirement. Model consistency requires, as we show in our work, the use of a special certificate, the minimal norm certificate $\mathring{\eta}_\Gamma$, which is equal to η_Γ if $\eta_\Gamma \in \mathrm{ri}(\partial J(\beta_0))$ (see Proposition 17).

Works on model consistency. Theorem 4 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 (Fuchs, 2004). In this setting, the result (7) corresponds to the correct identification of the support, i.e. $\mathrm{supp}(\beta_\theta) = \mathrm{supp}(\beta_0)$. Condition (5) for $J = \|\cdot\|_1$ is known in the statistics literature under the name "irrepresentable condition", see e.g. (Zhao and Yu, 2006). Knight and Fu (2000) have shown estimation consistency for Lasso for fixed p and β_0 and asymptotic normality of the estimates. The authors in (Zhao and Yu, 2006) proved Theorem 11 for $J = \|\cdot\|_1$, though under slightly different assumptions on the covariance and noise distribution. A similar result was established in (Jia and Yu, 2010) for the elastic net, i.e. $J = \|\cdot\|_1 + \rho\|\cdot\|_2^2$ for $\rho > 0$. In (Bach, 2008a) and (Bach, 2008b), the author has shown Theorem 11 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 (Vaiter et al., 2013b), we have proved an instance of Theorem 4 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 (Vaiter et al., 2013a) when $\Gamma = \tilde{\Gamma}$, and J belongs to the class of partly smooth functions relative to affine manifolds $\mathcal{M}$, i.e. $\mathcal{M} = \beta + T_\beta$. Typical instances encompassed in this class are the $\ell^1 - \ell^2$ norm, or its analysis version, as well as non-negative polyhedral functions including the ℓ^∞ norm. Note that the nuclear norm (and composition of it with linear operators as studied for instance in (Grave et al., 2011; Richard et al., 2013)), whose manifold is not affine, does not fit into the framework of (Vaiter et al., 2013a), while it is covered by Theorem 4. J. D. Lee and Taylor (2013) investigated a class of geometrically decomposable penalties, for which they extended the irrepresentable condition and used it to establish ℓ^2-consistency and model consistency. This class of penalties turns out to be a very special case of ours. Lastly, a similar result was proved in (Duval and Peyré, 2013) for an infinite dimensional sparse recovery problem over space of Radon measures, when J is the total variation of a measure (not to be confused with the total variation semi-norm mentioned above). In this setting, an interesting finding is that, when $\mathring{\eta}_{X^*X} \in \mathrm{ri}(\partial J(\beta_0))$, $\mathring{\eta}_{X^*X}$ is not equal to η_{X^*X} but to a different certificate (called "vanishing derivative" certificate by Duval and Peyré (2013)) that can also be computed by solving a linear system.

Compressed sensing Condition (5) is often used when X is drawn from the Gaussian matrix ensemble to asses the performance of compressed sensing recovery with ℓ^1 norm, see (Wainwright, 2009; Dossal et al., 2012). This is extended to a more general family of decomposable norms (including in particular $\ell^1 - \ell^2$ norms and the nuclear norm) in (Candès and Recht, 2013), 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 X is a random masking operator) in a noiseless setting (Candès and Recht, 2009; Candès and Tao, 2009). It was also used to ensure ℓ^2 robustness of matrix completion in a noisy setting (Candès and Plan, 2010), and our findings shows that these results also ensure rank consistency for matrix completion at high signal to low noise levels.

Sensitivity analysis. Sensitivity analysis is a central theme in variational analysis. Theorems 4 can be understood as a sensitivity analysis of the minimizers of E at the point (β_0, θ_0). Classical sensitivity analysis of non-smooth optimization problems seeks conditions to ensure smoothness of the mapping $\theta \mapsto \beta_\theta$ where β_θ is a minimizer of $f(\cdot, \theta)$, see for instance (Mordukhovich, 1992; Rockafellar and Wets, 1998; Bonnans and Shapiro, 2000).

This is usually guaranteed by the non-degenerate source condition and restricted injectivity condition (11), which, as already reviewed above, ensures linear convergence rate, and hence Lipschitz behaviour of this mapping. The result captured by Theorem 4 goes one step further, by assessing that $\mathcal{M}_{\beta_0}$ is a stable manifold (in the sense of (Wright, 1993)), since the minimizer β_θ is unique and remains in $\mathcal{M}_{\beta_0}$ for θ close to θ_0. Our starting point for establishing Theorem 4 is the inspiring work of Lewis (2003a) who first introduced the notion of partial smoothness and showed that this broad class of functions enjoys a powerful calculus and sensitivity theory. For convex functions (which is the setting considered in our work), partial smoothness is closely related to $\mathcal{U}-\mathcal{V}$-decompositions developed in (Lemaréchal et al., 2000). In fact, the behaviour of a partly smooth 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 (Lewis, 2003a; Lewis and Zhang, 2013). A important and distinctive feature of Theorem 4 is that, partial smoothness of J at β_0 relative to $\mathcal{M}$ transfers to $E(\cdot, \theta)$ for $\lambda > 0$, but not when $\lambda = 0$ in general. In particular, (Lewis, 2003a, Theorem 5.7) does not apply to prove our claim.

4. Case Study: Nuclear Norm Regularization

In this section, we illustrate the usefulness of our model consistency results to derive a sharp manifold stability analysis for the nuclear norm (a.k.a trace norm) regularization. As detailed in Section 3.5, previous consistency results due to Bach (2008b) only apply to the overdetermined setting, while our result tackles arbitrary design X by only requiring the weaker injectivity condition (5). For simplicity of exposition, we consider recovery of square matrices of size $p = p_0 \times p_0$, but the same holds for arbitrary rectangular matrices.

4.1 Irrepresentability Criterion IC

The nuclear norm, defined in (2), turns out to be the tightest convex relation of the rank function on the spectral ball. It is then the best convex candidate to enforce a low-rank prior (Fazel, 2002). It is moreover partly smooth at any $\beta_0 \in \mathbb{R}^{p_0 \times p_0}$ relative to the manifold $\mathcal{M}$ of fixed rank $r = \operatorname{rank}(\beta_0)$ defined in (3).

Let $\beta = U \operatorname{diag}(\Lambda) V^*$ be a reduced rank-r SVD decomposition of β_0, with $V, U \in \mathbb{R}^{p_0 \times r}$ with orthonormal columns and $\Lambda \in (\mathbb{R}_+^*)^r$. The subdifferential of the nuclear norm at β_0 reads (see for instance Candès and Recht (2013))

$$\partial \| \cdot \|_*(\beta_0) = \left\{ \eta \in \mathbb{R}^{p_0 \times p_0} \; ; \; \eta_T = e \quad \text{and} \quad \|\eta_S\| \leqslant 1 \right\} , \tag{12}$$

where $\|\eta\|$ is the operator norm, $T = \mathcal{T}_\beta(\mathcal{M})$, $S = T^\perp$ and $e = \mathrm{P}_T(\partial J(\beta_0))$, with

$$T = \left\{ U A^* + B V^* \; ; \; A, B \in \mathbb{R}^{p_0 \times r} \right\} \quad \text{and} \quad e = U V^*$$

and S is the subspace of matrices spanned by the family (wz^*), where w (resp. z) is any vector orthogonal to U (resp. V).

The relative interior of $\partial \| \cdot \|_*(\beta_0)$ is formed by subgradients η satisfying the inequality in (12) strictly. Thus, condition (5) in the case $\tilde{\Gamma} = \Gamma$ takes the analytical form

$$\eta_\Gamma \in \operatorname{ri}(\partial J(\beta_0)) \quad \iff \quad \mathbf{IC}(\beta_0) < 1 \quad \text{where} \quad \mathbf{IC}(\beta_0) = \| \mathrm{P}_S \, \Gamma \Gamma_T^+ e \| . \tag{13}$$

The value of $\mathbf{IC}(\beta_0)$ can then be easily computed. Loosely speaking, the smaller the quantity $1 - \mathbf{IC}(\beta_0)$ is, the further η_Γ is from the relative boundary of $\partial J(\beta_0)$, and in turn the smaller the stability constant controlling $\|\beta_\theta - \beta_0\| / \|\varepsilon\|$ in (7) is.

4.2 Recovery from Gaussian Measurements

Bounding **IC** for an arbitrary operator X and matrix β_0 is in general difficult. It is however possible to leverage tools from random matrix theory to obtain sharp upper-bounds when X is drawn from certain matrix ensembles. This strategy has been deployed to study matrix completion problems, see for instance Candès and Recht (2009); Candès and Tao (2009). Another problem on which we now focus is when X is drawn from the standard Gaussian ensemble, i.e. its entries are independent identically distributed from $\mathcal{N}(0,1)$. The following result, proved by Candès and Recht (2013), shows that $\mathbf{IC}(\beta_0) < 1$ with high probability as soon as n is larger than $6rp_0$ (up to negligible terms).

Proposition 14 ((Candès and Recht, 2013), Theorem 1.2) *Let $\beta_0 \in \mathbb{R}^{p_0 \times p_0}$ such that $\operatorname{rank}(\beta_0) = r$. If*

$$n \geqslant \delta r(6p_0 - 5r) \tag{14}$$

for some $\delta > 1$, then $\mathbf{IC}(\beta_0) < 1$ with probability at least $1 - 2e^{(1-\delta)p_0/8}$.

Combining this result with Theorem 4, this shows that under the scaling (14) of (n, p_0, r), one obtains with high probability on the design matrix a rank-consistent estimation of the unknown matrix β_0, which is (to the best of our knowledge) a novel result.

Figure 1 illustrates this result by computing the average (over 25 Monte Carlo replications) values of $\mathbf{IC}(\beta_0)$ for either a varying n or rank r. The shaded area corresponds to $\pm 3\times$ standard deviation across the 25 replications, and the dashed vertical line indicates the transition predicted by (14). This suggests numerically that the upper-bound (14) is indeed sharp.

5. Proofs

5.1 Uniqueness sufficient condition

Proposition 15 *Let J be a proper lsc convex function. For a point β, assume that*

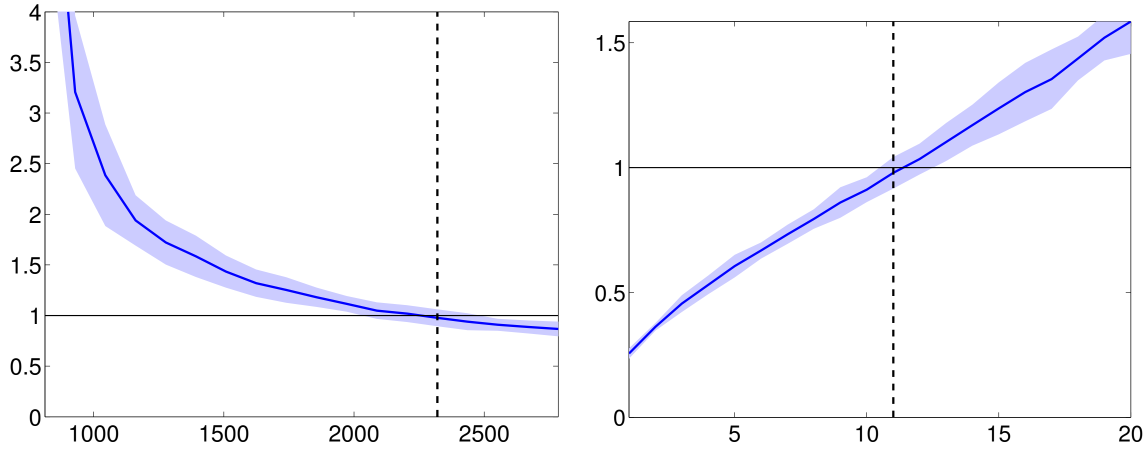
$$\ker(\Gamma) \cap T_\beta = \{0\}, \quad \textit{and} \quad \eta_\Gamma \in \operatorname{ri}(\partial J(\beta)).$$

Then β is the unique minimizer of $(\mathcal{P}_\theta)$ (resp. $(\mathcal{P}_{0,\Gamma\beta,\Gamma})$).

Proof This is a consequence of (Vaiter et al., 2013a, Corollary 1). Though their result was stated for J finite-valued convex, it remains valid when it is proper lsc and convex. Indeed, in this case, J is subdifferentially regular at β (Rockafellar and Wets, 1998, Example 7.27). Moreover, $\partial J(\beta) \neq \emptyset$ by assumption, and thus the directional derivative at β is proper, sublinear and closed, and it is the support of $\partial J(\beta)$ (Rockafellar and Wets, 1998, Theorem 8.30). Continuing the proof as in (Vaiter et al., 2013a, Corollary 1) shows the claim. ■

5.2 Minimal norm certificate and linearized pre-certificate

In this section, we establish the connections between the minimal norm certificate and the linearized pre-certificate $\eta_{\tilde{\Gamma}}$. Recall that J is proper lsc convex function, and we suppose that $\operatorname{Im}(\tilde{\Gamma}) \cap \partial J(\beta_0) \neq \emptyset$ (so-called range or source condition in the inverse problem



(a) $\mathbf{IC}(\beta_0)$ as function of n **(b)** $\mathbf{IC}(\beta_0)$ as function of r

Figure 1: Curves of $\mathbf{IC}(\beta_0)$ (central solid line: average, blue shaded band: $\pm 3\times$ standard deviation) for 25 realizations of $X \in \mathbb{R}^{n \times p_0^2}$ from the standard Gaussian ensemble, where $p_0 = 10^3$, and a random $\beta_0 = AB^*$ of rank r where $A, B \in \mathbb{R}^{p_0 \times r}$ are Gaussian matrices. **(a)** $\mathbf{IC}(\beta_0)$ as a function of n for a fixed $r = 4$. The vertical dashed line shows the threshold $n = r(6p_0 - 5r)$ indicated by (14). **(b)** $\mathbf{IC}(\beta_0)$ as a function of r for a fixed $n = 0.6p_0^2$. The vertical dashed line shows the threshold $r = (3p_0 - \sqrt{9p_0^2 - 5n})/5$ indicated by (14).

community). The latter is equivalent to the fact that β_0 is a minimizer of $(\mathcal{P}_{\theta_0})$. This is straightforward to see by writing the first-order optimality condition of this convex program.

Definition 16 (Minimal norm certificate) *The minimal norm certificate is the vector*

$$\mathring{\eta}_{\tilde{\Gamma}} = \tilde{\Gamma} \mathring{z}_{\tilde{\Gamma}}, \quad \textit{where} \quad \mathring{z}_{\tilde{\Gamma}} = \underset{\tilde{\Gamma} z \in \partial J(\beta_0)}{\operatorname{argmin}} \; \|z\|. \tag{15}$$

This certificate is uniquely defined as the constraint set is non-empty closed and convex, and the solution of the minimization problem, which is the projection of the origin on it, is obviously unique.

Proposition 17 *Assume that* $\ker(\tilde{\Gamma}) \cap T = \{0\}$*. Then,*

$$\eta_{\tilde{\Gamma}} \in \operatorname{ri}(\partial J(\beta_0)) \quad \Longrightarrow \quad \mathring{\eta}_{\tilde{\Gamma}} = \eta_{\tilde{\Gamma}}, \tag{16}$$

$$\mathring{\eta}_{\tilde{\Gamma}} \in \operatorname{ri}(\partial J(\beta_0)) \quad \Longrightarrow \quad \mathring{\eta}_{\tilde{\Gamma}} = \eta_{\tilde{\Gamma}}. \tag{17}$$

Under either of these conditions, β_0 *is the unique minimizer to* $(\mathcal{P}_{0,\tilde{\Gamma}\beta_0,\tilde{\Gamma}})$*.*

Proof *Proof of* (16) Under condition $\ker(X) \cap T = \{0\}$, we have from the definition of $\tilde{\Gamma}_T^+$, that

$$z_{\tilde{\Gamma}} = \tilde{\Gamma}_T^+ e = \underset{z}{\operatorname{argmin}} \; \|z\| \quad \text{subject to} \quad \tilde{\Gamma}_T z = e \tag{18}$$

and thus

$$\eta_{\tilde{\Gamma}} = \tilde{\Gamma} z_{\tilde{\Gamma}} \; .$$

Clearly, the constraint set of problem (18) includes that of (15), which entails

$$\|z_{\tilde{\Gamma}}\| \leqslant \|\mathring{z}_{\tilde{\Gamma}}\| \ .$$

If $\eta_{\tilde{\Gamma}} \in \mathrm{ri}(\partial J(\beta_0))$, then $z_{\tilde{\Gamma}}$ is also a feasible point of problem (15) and thus

$$\|\mathring{z}_{\tilde{\Gamma}}\| \leqslant \|z_{\tilde{\Gamma}}\| \ .$$

Altogether, we get that $\|\mathring{z}_{\tilde{\Gamma}}\| = \|z_{\tilde{\Gamma}}\|$ and, since $\mathring{z}_{\tilde{\Gamma}}$ is the unique minimizer of (15), we get that $\mathring{z}_{\tilde{\Gamma}} = z_{\tilde{\Gamma}}$, which implies that $\mathring{\eta}_{\tilde{\Gamma}} = \eta_{\tilde{\Gamma}}$.
Proof of (17) Let $S = T^{\perp}$. Problem (15) can be conviniently rewritten as

$$\mathring{z}_{\tilde{\Gamma}} = \underset{z}{\mathrm{argmin}}\ \|z\| \quad \text{subject to} \quad \tilde{\Gamma}_T z = e \quad \text{and} \quad \tilde{\Gamma}_S z \in \mathrm{P}_S(\partial J(\beta_0)) \ .$$

The fact that $\mathring{\eta}_{\tilde{\Gamma}} = \tilde{\Gamma}\mathring{z}_{\tilde{\Gamma}} \in \mathrm{ri}(\partial J(\beta_0))$ implies $\mathrm{P}_S\, \mathring{\eta}_{\tilde{\Gamma}} = \mathrm{P}_S\, \tilde{\Gamma}\mathring{z}_{\tilde{\Gamma}} \in \mathrm{ri}(\mathrm{P}_S\, \partial J(\beta_0))$, and thus, the second constraint in the last problem is inactive. We then recover problem (18), which in turn implies that $\mathring{\eta}_{\tilde{\Gamma}} = \eta_{\tilde{\Gamma}}$.
Proof of uniqueness. See Proposition 15. ■

5.3 Proof of Theorem 4

In order to prove Theorem 4, 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

$$\left(\Gamma_k, \varepsilon_k\, \mu_k^{-1}, \mu_k\right) \longrightarrow (\tilde{\Gamma}, 0, 0) \ . \tag{19}$$

Then proving Theorem 4 boils down to showing that for k large enough, the solution β_k of $(\mathcal{P}_{\theta_k})$ is unique and satisfies $\beta_k \in \mathcal{M}$.

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

$$\beta_k \in \underset{\beta \in \mathcal{M} \cap \mathcal{K}}{\mathrm{Argmin}}\ E(\beta, \theta_k) \tag{20}$$

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

The following key lemma establishes the convergence of β_k to β_0.

Lemma 18 *Under conditions* (5) *and* (19)*,* $\beta_k \to \beta_0$*.*

Proof We denote $\|u\|_{\Gamma}^2 = \langle \Gamma u,\, u\rangle$ for any positive semidefinite matrix Γ. Under condition (5), Proposition 15 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 $E(\beta_k, \theta_k) \leqslant E(\beta_0, \theta_k)$ and hence

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

which is equivalently stated as

$$\frac{1}{2}\|\beta_k - \beta_0\|_{\Gamma_k}^2 - \langle \beta_k - \beta_0,\, \varepsilon_k \rangle + \mu_k J(\beta_k) \leqslant \mu_k J(\beta_0). \tag{21}$$

Since $\beta_k \in \mathcal{K}$, the sequence $(\beta_k)_k$ is bounded, and we let $\beta^\star$ be any accumulation point. Using (19), that J is non-negative and $J(\beta_k)$ are bounded, we have

$$\begin{aligned}
&\limsup_{k\to\infty} \left(\mu_k J(\beta_k)\right) \leqslant \lim_{k\to\infty} \mu_k \limsup_{k\to\infty} J(\beta_k) = 0 \quad \text{and} \\
&\liminf_{k\to\infty} \left(\mu_k J(\beta_k)\right) \geqslant \lim_{k\to\infty} \mu_k \liminf_{k\to\infty} J(\beta_k) = J(\beta^\star) \lim_{k\to\infty} \mu_k = 0 \ ,
\end{aligned}$$

and thus $\lim_{k\to\infty} \left(\mu_k J(\beta_k)\right) = 0$. Consequently, passing to the limit in (21), using (19), and continuity of the inner product and the norm, shows that $\|\beta^\star - \beta_0\|_{\tilde\Gamma}^2 \leqslant 0$, or equivalently $\tilde\Gamma \beta^\star = \tilde\Gamma \beta_0$, i.e. $\beta^\star$ is a feasible point of $(\mathcal{P}_{0,\tilde\Gamma x_0,\tilde\Gamma})$. Furthermore, since $\frac{1}{2}\|\beta_k - \beta_0\|_{\Gamma_k}^2 \geqslant 0$, (21) yields

$$-\langle \beta_k - \beta_0,\, \frac{\varepsilon_k}{\mu_k} \rangle + J(\beta_k) \leqslant J(\beta_0).$$

Passing again to the limit, using lower semicontinuity of J, (19) and continuity of the inner product, we then get

$$\begin{aligned}
J(\beta^\star) \leqslant \liminf_{k\to\infty} J(\beta_k) &= \liminf_{k\to\infty} \left(-\langle \beta_k - \beta_0,\, \frac{\varepsilon_k}{\mu_k} \rangle + J(\beta_k) \right) \\
&\leqslant \limsup_{k\to\infty} \left(-\langle \beta_k - \beta_0,\, \frac{\varepsilon_k}{\mu_k} \rangle + J(\beta_k) \right) = \limsup_{k\to\infty} J(\beta_k) \leqslant J(\beta_0) \ .
\end{aligned}$$

Combining this with the previous claim on feasibility of $\beta^\star$ for $(\mathcal{P}_{0,\tilde\Gamma x_0,\tilde\Gamma})$ allows to conclude that $\beta^\star$ is a solution of $(\mathcal{P}_{0,\tilde\Gamma x_0,\tilde\Gamma})$. Since β_0 is unique, this leads to $\beta^\star = \beta_0$. ■

We now aim at showing that for k large enough, β_k is the unique solution of $(\mathcal{P}_{\theta_k})$.

Convergence of the tangent model subspace. By definition of the constrained problem (20), $\beta_k \in \mathcal{M}$. 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 $\mathcal{M}$ (Lewis, 2003a, 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

$$T_k = \mathcal{T}_{\beta_k}(\mathcal{M}) \longrightarrow \mathcal{T}_{\beta_0}(\mathcal{M}) = T \tag{22}$$

when (19) 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 $\mathrm{P}_{T_k} \to \mathrm{P}_T$). Since $\ker(\tilde\Gamma) \cap T = \{0\}$, (22) implies that for k large enough, when (19) holds,

$$\ker(\Gamma_k) \cap T_k = \{0\}, \tag{23}$$

which we assume from now on.

First order condition. Let $\mathbb{B}$ be the Euclidean unit ball in $\mathbb{R}^n$. Take $\mathcal{K} = \beta_0 + r\mathbb{B}$ for $r > 0$ sufficiently large. For any $\delta > 0$, $\exists k_\delta$ such that $\forall k \geqslant k_\delta$, $\beta_k \in \beta_0 + \delta\mathbb{B}$ according to Lemma 18. Thus, for k large enough, i.e. δ sufficiently small, we indeed have $\beta_k \in \mathrm{int}(\mathcal{K})$. Furthermore, it is easy to see that $\iota_\mathcal{K}$ is locally partly smooth at β_0 relative to $\mathcal{K}$, and thus is partly smooth at β_k relative to $\mathcal{K}$ for k large enough. Moreover, local partial smoothness of J at β_0 relative to $\mathcal{M}$ entails that J is also partly smooth at β_k relative to $\mathcal{M}$. Therefore, the sum rule (Lewis, 2003a, Corollary 4.6) (the transversality condition is satisfied as $\mathcal{K}$ is full-dimensional and $\beta_k \in \mathrm{int}(\mathcal{K})$, see (4)) shows that, for all sufficiently large 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 (Lewis, 2003a, Corollary 4.7). Therefore, (Lewis, 2003a, Proposition 2.4(a)-(b)) applies, and it follows that β_k is a critical point of (20) if, and only if,

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

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 $e_{\beta_k} \in T_k$,

$$\mathrm{P}_{T_k}(\Gamma_k \beta_k - u_k) + \mu_k e_{\beta_k} = 0. \tag{24}$$

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 $\mathcal{M}$ Poliquin et al. (2000). In particular, $\beta_k = \mathrm{P}_\mathcal{M}(\beta_k)$ and $\beta_0 = \mathrm{P}_\mathcal{M}(\beta_0)$. Moreover, $\mathrm{P}_\mathcal{M}$ is of class C^1 near β_k (Lewis and Malick, 2008, Lemma 4). Thus, C^2 differentiability shows that

$$\beta_k - \beta_0 = \mathrm{P}_\mathcal{M}(\beta_k) - \mathrm{P}_\mathcal{M}(\beta_0) = \mathrm{D}\,\mathrm{P}_\mathcal{M}(\beta_k)(\beta_k - \beta_0) + R(\beta_k)$$

where $R(\beta_k) = O(\|\beta_k - \beta_0\|^2)$ and where $\mathrm{D}\,\mathrm{P}_\mathcal{M}(\beta_k)$ is the derivative of $\mathrm{P}_\mathcal{M}$ at β_k. Using (Lewis and Malick, 2008, Lemma 4), and recalling that $T_k = \mathcal{T}_{\beta_k}(\mathcal{M})$ by the sharpness property, the derivative $\mathrm{D}\,\mathrm{P}_\mathcal{M}(\beta_k)$ is given by $\mathrm{D}\,\mathrm{P}_\mathcal{M}(\beta_k) = \mathrm{P}_{T_k}$. Inserting this in (24), we get

$$\mathrm{P}_{T_k}\,\Gamma_k\left(\mathrm{P}_{T_k}(\beta_k - \beta_0) + R(\beta_k)\right) - \mathrm{P}_{T_k}\,\varepsilon_k + \mu_k e_{\beta_k} = 0. \tag{25}$$

Using (23), Γ_{k,T_k} has full rank, and thus

$$\beta_k - \beta_0 = \Gamma^+_{k,T_k}\left(\varepsilon_k - \mu_k e_{\beta_k} - \Gamma_k R(\beta_k)\right), \tag{26}$$

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

$$\|\beta_k - \beta_0\| = O\left(\|\varepsilon_k\|, \mu_k\right). \tag{27}$$

Convergence of the dual variables. We define $\eta_k = \frac{u_k - \Gamma_k \beta_k}{\mu_k}$. Arguing as above, and using (26) we have

$$\begin{aligned}
\mu_k \eta_k &= \varepsilon_k + \Gamma_k(\beta_0 - \beta_k) = \varepsilon_k - \Gamma_k \Gamma^+_{k,T_k}\left(\varepsilon_k - \mu_k e_{\beta_k} - \Gamma_k R(\beta_k)\right) \\
&= \varepsilon_k - \Gamma_k\,\mathrm{P}_{T_k}\,\Gamma^+_{k,T_k}\left(\varepsilon_k - \mu_k e_{\beta_k} - \Gamma_k R(\beta_k)\right) \\
&= \mathrm{P}_{V^\perp_{T_k}}\,\varepsilon_k + \mathrm{P}_{V_{T_k}}\,\Gamma_k R(\beta_k) + \mu_k \Gamma_k \Gamma^+_{k,T_k} e_{\beta_k},
\end{aligned}$$

where we denoted $V_{T_k} = \mathrm{Im}(\Gamma_k \mathrm{P}_{T_k})$, and used that $\mathrm{Im}(\Gamma^+_{k,T_k}) \subset T_k$. 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

$$\|e_{\beta_k} - e\| = O(\|\beta_k - \beta_0\|). \tag{28}$$

Using the triangle inequality, we get

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

Again, since $\Gamma^+_{k,T_k} \to \tilde\Gamma^+_T$, 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

$$\|\Gamma^+_{k,T_k} - \tilde\Gamma^+_T\| = O(\|\Gamma_{k,T_k} - \tilde\Gamma_T\|) = O(\|\Gamma_k - \tilde\Gamma\|, \|\mathrm{P}_{T_k} - \mathrm{P}_T\|) = O(\|\Gamma_k - \tilde\Gamma\|, \|\beta_k - \beta_0\|).$$

This shows that

$$\|\Gamma_k \Gamma^+_{k,T_k} - \tilde\Gamma \tilde\Gamma^+_T\| = O(\|\Gamma_k - \tilde\Gamma\|, \|\beta_k - \beta_0\|). \tag{29}$$

Putting (28) and (29) 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 (27), we arrive at

$$\|\eta_k - \eta_{\tilde\Gamma}\| = O\left(\|\Gamma_k - \tilde\Gamma\|, \|\varepsilon_k\| \mu_k^{-1}, \mu_k \right) . \tag{30}$$

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

$$\eta_k \in \mathrm{ri}(\partial J(\beta_k)). \tag{31}$$

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

$$\eta_k \in \mathrm{rbd}(\partial J(\beta_k)) \ . \tag{32}$$

According to (30) and Lemma 18, under (19), $(\beta_k, \eta_k) \to (\beta_0, \eta_{\tilde\Gamma})$. Condition (32) is equivalently stated as, for each k

$$\exists z_k \in T^\perp_{\beta_k}, \quad \forall \eta \in \partial J(\beta_k), \quad \langle z_k,\, \eta - \eta_k \rangle \geqslant 0, \tag{33}$$

where one can impose the normalization $\|z_k\| = 1$ by positive-homogeneity. Up to a subsequence (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 $z^\star$.

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

$$\forall\, v \in \partial J(\beta_0), \quad \langle z^{\star},\, \eta - \eta_{\tilde{\Gamma}} \rangle \geqslant 0. \tag{34}$$

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 $\beta_k \to \beta_0$ there exists $v_k \in \partial J(\beta_k)$ with $v_k \to v$. Applying (33) with $\eta = v_k$ gives $\langle z_k,\, v_k - \eta_k \rangle \geqslant 0$. Taking the limit $k \to +\infty$ in this inequality leads to (34), which contradicts the fact that $\eta_{\tilde{\Gamma}} \in \mathrm{ri}(\partial J(\beta_0))$. In view of (31) and (23), using Proposition 15 shows that β_k is the unique solution of $(\mathcal{P}_{\theta_k})$.

5.4 General Loss Function

Using , all our results remain valid by appropriately adapting the proofs.

We now detail the necessary arguments to adapt the proof of Theorem 4 to a generic loss function satisfying assumptions **(A.1)-(A.2)**.

Proof of Proposition 15 It follows from assumption **(A.1)** that $F(\cdot, y)$ is strictly convex, and the uniqueness follows from (Liang et al., 2014, Theorem A.1).

Proof of Lemma 18 Problem (20) now reads

$$\beta_k \in \underset{\beta \in \mathcal{M} \cap \mathcal{K}}{\mathrm{Argmin}}\; F(X_k \beta, y_k) + \lambda_k J(\beta)\ .$$

Optimality of β_k entails

$$F(X_k \beta_k, y_k) + \lambda_k J(\beta_k) \leqslant F(X_k \beta_0, y_k) + \lambda_k J(\beta_0)\ .$$

By assumptions **(A.1)-(A.2)**, we have the following useful inequalities for any $u \in \mathbb{R}^n$, see e.g. (Nesterov, 2004, p. 57 and 64)

$$\frac{\sigma_{\mathrm{m}}}{2}\|y - u\|^2 \leqslant F(u, y) - F(y, y) = F(u, y) - F(y, y) - \langle \nabla F(y, y),\, u - y \rangle \leqslant \frac{\sigma_{\mathrm{M}}}{2}\|y - u\|^2\ .$$

It then follows that

$$F(X_k \beta_k, y_k) - F(X_k \beta_0, y_k) \geqslant \frac{\sigma_{\mathrm{m}}}{2}\|y_k - X_k \beta_k\|^2 - \frac{\sigma_{\mathrm{M}}}{2}\|w_k\|^2$$

and therefore

$$\begin{aligned}
\lambda_k J(\beta_0) &\geqslant \frac{\sigma_{\mathrm{m}}}{2}\|y_k - X_k \beta_k\|^2 - \frac{\sigma_{\mathrm{M}}}{2}\|w_k\|^2 + \lambda_k J(\beta_k) \\
&\geqslant \frac{\sigma_{\mathrm{m}}}{2\sigma_{\mathrm{M}}}\|y_k - X_k \beta_k\|^2_{\nabla_1^2 F(y_k, y_k)} - \frac{\sigma_{\mathrm{M}}}{2}\|w_k\|^2 + \lambda_k J(\beta_k) \\
&\geqslant \frac{\sigma_{\mathrm{m}}}{2\sigma_{\mathrm{M}}}\|\beta_k - \beta_0\|^2_{X_k^* \nabla_1^2 F(y_k, y_k) X_k} - \frac{\sigma_{\mathrm{m}}}{\sigma_{\mathrm{M}}}\langle \beta_k - \beta_0,\, X_k^* \nabla_1^2 F(y_k, y_k) w_k \rangle \\
&\quad - \frac{\sigma_{\mathrm{M}}}{2}\left(1 - \frac{\sigma_{\mathrm{m}}^2}{\sigma_{\mathrm{M}}^2}\right)\|w_k\|^2 + \lambda_k J(\beta_k)\ ,
\end{aligned}$$

where we used strong convexity of assumption **(A.1)** in the second and third inequalities. Dividing both sides by $1/P$ we obtain

$$\frac{\sigma_{\mathrm{m}}}{2\sigma_{\mathrm{M}}}\|\beta_k - \beta_0\|^2_{\check{\Gamma}_k} - \frac{\sigma_{\mathrm{m}}}{\sigma_{\mathrm{M}}}\langle \beta_k - \beta_0,\, \check{\varepsilon}_k \rangle - \frac{\sigma_{\mathrm{M}}}{2}\left(1 - \frac{\sigma_{\mathrm{m}}^2}{\sigma_{\mathrm{M}}^2}\right)\|n^{-1/2} w_k\|^2 + \mu_k J(\beta_k) \leqslant \mu_k J(\beta_0)\ ,$$

where now

$$\check{\Gamma}_k = \frac{1}{n} \; X_k^* \nabla_1^2 F(y_k, y_k) X_k \quad \text{and} \quad \check{\varepsilon}_k = \frac{1}{n} \; X_k^* \nabla_{1,2} F(y_k, y_k) w_k \; .$$

Changing (19) to $\left(\check{\Gamma}_k, \check{\varepsilon}_k \, \mu_k^{-1}, \mu_k\right) \longrightarrow (\tilde{\Gamma}, 0, 0)$, which entails implicitly that $n^{-1/2} w_k \to 0$, and arguing as in the rest of the proof of the lemma allows to conclude that $\beta_k \to \beta_0$.

Proof of Theorem 4 C^2-continuity of F allows to use the smooth perturbation rule to conclude that partial smoothness of J is preserved upon adding F. Condition (24) now becomes

$$\mathrm{P}_{T_k} \, X_k^* \nabla_1 F(X_k \beta_k, y_k) + \lambda_k e_{\beta_k} = 0.$$

Using again assumptions **(A.1)-(A.2)** and expanding $\nabla_1 F(X_k \beta_k, y_k)$ at (y_k, y_k) to the first order, we obtain

$$\begin{aligned} \nabla_1 F(X_k \beta_k, y_k) &= \nabla_1^2 F(y_k, y_k) X_k (\beta_k - \beta_0) - \nabla_1^2 F(y_k, y_k) w_k + O\left(\|\beta_k - \beta_0\|^2\right) + O\left(\|w_k\|^2\right) \\ &= \nabla_1^2 F(y_k, y_k) X_k \left(\mathrm{P}_{T_k}(\beta_k - \beta_0) + R(\beta_k)\right) - \nabla_1^2 F(y_k, y_k) w_k \\ &\quad + O\left(\|\beta_k - \beta_0\|^2\right) + O\left(\|w_k\|^2\right) \; . \end{aligned}$$

Dividing by n, plugging this expansion back into the above first-order (criticality) condition, and grouping the $O(.)$ terms, condition (25) becomes

$$\check{\Gamma}_{k,T_k}(\beta_k - \beta_0) - \mathrm{P}_{T_k} \, \check{\varepsilon}_k + \mu_k e_{\beta_k} + \mathrm{P}_{T_k}(n^{-1} X_k^* + \check{\Gamma}_k) R(\beta_k) + \mathrm{P}_{T_k} \, X_k^* Q(n^{-1/2} w_k) = 0 \; ,$$

where $Q(n^{-1/2} w_k) = O(\|n^{-1/2} w_k\|^2)$. Then with the new notations $(\check{\Gamma}_k, \check{\varepsilon}_k)$ in place of $(\Gamma_k, \varepsilon_k)$, one sees that the proof continues unchanged.

5.5 Proof of Theorem 11

It is sufficient to check that (6) is in force with probability 1 as $n \to +\infty$. Owing to classical results on convergence of sample covariances, which apply thanks to the assumption that the fourth order moments are finite, we get $\Gamma_n - \tilde{\Gamma} = O_P\left(n^{-1/2}\right)$ and $\frac{1}{n}\langle \xi_i, \, w \rangle = O_P\left(n^{-1/2}\right)$, where we used the assumption that $\mathbb{E}\left(\langle \xi_i, \, w \rangle\right) = 0$. As p is fixed, it follows that $\|\Gamma_n - \tilde{\Gamma}\| = O_P\left(n^{-1/2}\right)$ and $\|\varepsilon_n\| = O_P\left(n^{-1/2}\right)$. Thus under the scaling (9), 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.

5.6 Proof of Proposition 9

Let $(\beta_k)_k$ be a sequence of solutions to the constrained problem (20). Since β_0 is the unique minimizer to $(\mathcal{P}_{(0, \tilde{\Gamma}\beta_0, \tilde{\Gamma})})$ and (5) is satisfied, $\eta_{\tilde{\Gamma}}$ is well-defined. Moreover, arguing as in the proof of Lemma 18 and Theorem 4, under condition (6), we have $(\beta_k, \eta_k) \to (\beta_0, \eta_{\tilde{\Gamma}})$, and $\eta_k \in \eta_{\tilde{\Gamma}} + C\mathbb{B}$.

Let $\tau = \mathrm{dist}(\eta_{\tilde{\Gamma}}, \partial J(\beta_0)) = \inf_{\eta \in \partial J(\beta_0)} \|\eta - \eta_{\tilde{\Gamma}}\|$. Since $\partial J(\beta_0)$ is a non-empty, closed and convex set, the infimum is attained and one has $\tau > 0$ since $\eta_{\tilde{\Gamma}} \notin \partial J(\beta_0)$.

We now prove the claim by contradiction. Let β_j be a solution of $(\mathcal{P}_{\theta_j})$ such that (6) holds at θ_j for j sufficiently large (taking C smaller if necessary so that $C < \tau$), and suppose that $\beta_j \in \mathcal{M}$. Thus, β_j is also a solution of (20) for θ_j, whence it follows that $\eta_j \in \eta_{\tilde{\Gamma}} + C\mathbb{B}$. Using the triangle inequality, we then get

$$\mathrm{dist}(\eta_j, \partial J(\beta_0)) > \tau - C > 0 \ . \tag{35}$$

Now, in view of the continuity property in Definition 1((iii)), we have $\partial J(\beta_k) \to \partial J(\beta_0)$ along $\mathcal{M}$. This is equivalent, since $\partial J(\beta_0)$ is closed and using (Rockafellar and Wets, 1998, Corollary 4.7), to $\mathrm{dist}(\eta, \partial J(\beta_k)) \to \mathrm{dist}(\eta, \partial J(\beta_0))$ for every $\eta \in \mathbb{R}^n$, i.e.

$$\forall \delta > 0, \exists k_0, \forall k \geqslant k_0, \quad |\,\mathrm{dist}(\eta, \partial J(\beta_k)) - \mathrm{dist}(\eta, \partial J(\beta_0))| < \delta, \quad \forall \eta \in \mathbb{R}^n \ .$$

In particular, as β_j is a minimizer of $(\mathcal{P}_{\theta_j})$ for j large enough, we have $\eta_j \in \partial J(\beta_j)$, and thus $\mathrm{dist}(\eta_j, \partial J(\beta_0)) < \delta$, leading to a contradiction with (35). Hence, $\beta_j \notin \mathcal{M}$.

6. Conclusion

In this paper, we provided a very general and principled analysis of the recovery performance when partly smooth functions are used to regularize linear inverse/regression problems. This class of functions encompass all popular regularizers used in the literature. The generality of our results is unprecedented since for the first time, a unified analysis is provided together with a generalized “irrepresentable condition” to guarantee consistent identification of the low-complexity manifold underlying the original object. Our work also shows that model consistency is not only of theoretical interest, but also has algorithmic and practical consequences. Indeed, after a finite number of iterations, the iterates of the proximal splitting algorithm used to solve the original optimization problem (here the Forward-Backward), are guaranteed to lie on the original manifold. This opens the door to acceleration by switching to a higher-order smooth optimization method, exploiting the smoothness of the partly smooth objective function along the identified smooth model manifold.

Acknowledgements

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

References

J.-F. Aujol, G. Aubert, L. Blanc-Féraud, and A. Chambolle. Image decomposition into a bounded variation component and an oscillating component. *Journal of Mathematical Imaging and Vision*, 22:71–88, 2005.

F. Bach. Consistency of the group Lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008a.

F. Bach. Consistency of trace norm minimization. *Journal of Machine Learning Research*, 9:1019–1048, 2008b.

H. H. Bauschke and P. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2011.

J. F. Bonnans and A. Shapiro. *Perturbation analysis of optimization problems*. Springer Series in Operations Research and Financial Engineering. Springer Verlag, 2000.

M. Burger and S. Osher. Convergence rates of convex variational regularization. *Inverse Problems*, 20(5):1411, 2004.

E. Candès and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6): 925–936, 2010.

E. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

E. Candès and B. Recht. Simple bounds for recovering low-complexity models. *Math. Program*, 141(1-2):577–589, 2013.

E. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2009.

E. 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. ISSN 1097-0312.

E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *J. ACM*, 58(3):11:1–11:37, June 2011.

S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM journal on scientific computing*, 20(1):33–61, 1999.

A. Daniilidis, D. Drusvyatskiy, and A. S. Lewis. Orthogonal invariance and identifiability. *to appear in SIAM J. Matrix Anal. Appl.*, 2014a.

A. Daniilidis, J. Malick, and H. Sendov. Spectral (isotropic) manifolds and their dimension. *to appear in Journal d'Analyse Mathématique*, 2014b.

C. Dossal, M. Chabanol, G. Peyré, and J. Fadili. Sharp support recovery from noisy random measurements by l1 minimization. *Applied and Computational Harmonic Analysis*, 33(1): 24–43, 2012. doi: 10.1016/j.acha.2011.09.003. URL `http://hal.archives-ouvertes.fr/hal-00553670/`.

V. Duval and G. Peyré. Exact support recovery for sparse spikes deconvolution. Technical report, Preprint hal-00839635, 2013. URL `http://hal.archives-ouvertes.fr/hal-00839635/`.

M. Elad, P. Milanfar, and R. Rubinstein. Analysis versus synthesis in signal priors. *Inverse problems*, 23(3):947, 2007. doi: 10.1088/0266-5611/23/3/007.

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. URL `http://arxiv.org/abs/1304.4407`.

M. Fazel. *Matrix rank minimization with applications.* PhD thesis, Stanford University, 2002.

J. Fuchs. On sparse representations in arbitrary redundant bases. *IEEE Transactions on Information Theory*, 50(6):1341–1344, 2004.

M. Golbabaee and P. Vandergheynst. Hyperspectral Image Compressed Sensing Via Low-Rank And Joint-Sparse Matrix Recovery. In *2012 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP)*, pages 2741–2744. IEEE, 2012.

M. Grasmair. Linear convergence rates for Tikhonov regularization with positively homogeneous functionals. *Inverse Problems*, 27:075014, 2011.

M. Grasmair, O. Scherzer, and M. Haltmeier. Necessary and sufficient conditions for linear convergence of ℓ_1-regularization. *Communications on Pure and Applied Mathematics*, 64 (2):161–182, 2011.

E. Grave, G. Obozinski, and F. Bach. Trace lasso: a trace norm regularization for correlated designs. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. C. N. Pereira, and K. Q. Weinberger, editors, *Proc. NIPS*, pages 2187–2195, 2011.

W. Hare and A. Lewis. Identifying active constraints via partial smoothness and prox-regularity. *J. Convex Anal.*, 11(2):251–266, 2004.

M. Herman and T. Strohmer. General deviants: An analysis of perturbations in compressed sensing. *Selected Topics in Signal Processing, IEEE Journal of*, 4(2):342–349, April 2010.

Y. S. J. D. Lee and Y. E. Taylor. On model selection consistency of m-estimators with geometrically decomposable penalties. Technical report, arXiv:1305.7477, 2013.

J. Jia and B. Yu. On model selection consistency of the elastic net when $p \gg n$. *Statistica Sinica*, 20:595–611, 2010.

K. Knight and W. Fu. Asymptotics for Lasso-Type Estimators. *The Annals of Statistics*, 28(5):1356–1378, 2000.

J. M. Lee. *Smooth manifolds.* Springer, 2003.

C. Lemaréchal, F. Oustry, and C. Sagastizábal. The u-lagrangian of a convex function. *Trans. Amer. Math. Soc.*, 352(2):711–729, 2000.

A. S. Lewis. Active sets, nonsmoothness, and sensitivity. *SIAM Journal on Optimization*, 13(3):702–725, 2003a.

A. S. Lewis. The mathematics of eigenvalue optimization. *Mathematical Programming*, 97 (1–2):155–176, 2003b.

A. S. Lewis and J. Malick. Alternating projections on manifolds. *Mathematics of Operations Research*, 33(1):216–234, 2008.

A. S. Lewis and S. Zhang. Partial smoothness, tilt stability, and generalized hessians. *SIAM Journal on Optimization*, 23(1):74–94, 2013.

J. Liang, M. Fadili, and G. Peyré. Local linear convergence of forward–backward under partial smoothness. Technical report, arxiv preprint arXiv:1407.5611, 2014. appeared in NIPS 2014.

P.-L. Loh and M. J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637–1664, 06 2012. doi: 10.1214/12-AOS1018. URL `http://dx.doi.org/10.1214/12-AOS1018`.

B. Mordukhovich. Sensitivity analysis in nonsmooth optimization. *Theoretical Aspects of Industrial Design (D. A. Field and V. Komkov, eds.), SIAM Volumes in Applied Mathematics*, 58:32–46, 1992.

Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course.* Science & Business Media. Springer, 2004.

S. Oymak, A. Jalali, M. Fazel, Y. Eldar, and B. Hassibi. Simultaneously structured models with application to sparse and low-rank matrices. *arXiv preprint arXiv:1212.3753*, 2012.

G. Peyré, M. Fadili, and J.-L. Starck. Learning the morphological diversity. *SIAM Journal on Imaging Sciences*, 3(3):646–669, 2010.

R. Poliquin, R. Rockafellar, and L. Thibault. Local differentiability of distance functions. *Trans. Amer. Math. Soc.*, 352:5231–5249, 2000.

B. Recht, M. Fazel, and P. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

E. Richard, F. 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.

R. Rockafellar and R. Wets. *Variational analysis*, volume 317. Springer Verlag, 1998.

M. Rosenbaum and A. B. Tsybakov. Sparse recovery under matrix uncertainty. *The Annals of Statistics*, 38(5):2620–2651, 10 2010. doi: 10.1214/10-AOS793. URL `http://dx.doi.org/10.1214/10-AOS793`.

L. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992.

O. Scherzer, M. Grasmair, H. Grossauer, M. Haltmeier, and F. Lenzen. *Variational Methods in Imaging.* Applied Mathematical Sciences. Springer, 1st edition, 2009. ISBN 0387309314.

J.-L. Starck, M. Elad, and D. Donoho. Image decomposition via the combination of sparse representatntions and variational approach. *IEEE Trans. Image Processing*, 14(10):1570–1582, 2005.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B. Methodological*, 58(1):267–288, 1996.

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused Lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

S. Vaiter, M. Golbabaee, M. J. Fadili, and G. Peyré. Model selection with low complexity priors. Technical report, arXiv preprint arXiv:1307.2342, 2013a.

S. Vaiter, G. Peyré, C. Dossal, and M. Fadili. Robust sparse analysis regularization. *IEEE Transactions on Information Theory*, 59(4):2001–2016, 2013b.

M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, 2009.

S. J. Wright. Identifiable surfaces in constrained optimization. *SIAM Journal on Control and Optimization*, 31(4):1063–1079, 1993.

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.

P. Zhao and B. Yu. On model selection consistency of Lasso. *J. Mach. Learn. Res.*, 7: 2541–2563, Dec. 2006. ISSN 1532-4435.