Model Consistency of Partly Smooth Regularizers
Samuel Vaiter, Gabriel Peyré, Jalal M. Fadili

To cite this version:
Samuel Vaiter, Gabriel Peyré, Jalal M. Fadili. Model Consistency of Partly Smooth Regularizers. IEEE Transactions on Information Theory, Institute of Electrical and Electronics Engineers, 2018, 64 (3), pp.1725-1737. 10.1109/TIT.2017.2713822. hal-01658847

HAL Id: hal-01658847
https://hal.archives-ouvertes.fr/hal-01658847
Submitted on 7 Dec 2017

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 Vaiter, Gabriel Peyré and Jalal Fadili

Abstract—This paper studies least-square regression penalized with partly smooth convex regularizers. This class of penalty functions is very large 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 identify 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.

Index Terms—Regularization, regression, inverse problems, model consistency, partial smoothness, sensitivity analysis, sparsity, low-rank.

I. INTRODUCTION

A. Problem Statement

We consider the following observation model

$$y = x_0 + w,$$

where $\Phi \in \mathbb{R}^{p \times n}$ is the design matrix (in statistics or machine learning) or the forward operator (in signal and imaging sciences), $x_0 \in \mathbb{R}^n$ is the vector to recover and $w \in \mathbb{R}^p$ 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 $J : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ 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

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

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

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

and we denote

$$\varepsilon = \frac{\Phi^* w}{p} = u - \Gamma x_0,$$

where $\Phi^*$ is the adjoint operator to the linear operator $\Phi$. In the following, we assume that $y \in \text{Im}(\Phi)$ and thus $u \in \text{Im}(\Gamma)$. Obviously, this does not entail any loss of generality, as the loss term can always be written as

$$\frac{1}{2\lambda} \| \Phi x - P_{\text{Im}(\Phi)} y \|^2 + \frac{1}{2\lambda} \| P_{\text{Im}(\Phi)} y \|^2,$$

where $P_T$ is the orthogonal projection on $T$.

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

$$\min_{x \in \mathbb{R}^n} E(x; \theta) \quad (P_0)$$

where

$$E(x; \theta) = J(x) + \frac{1}{2\mu} \langle \Gamma x, x \rangle - \frac{1}{\mu} \langle x, u \rangle + \frac{1}{2\mu} \langle \Gamma^+ u, u \rangle,$$

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

We also consider the constrained problem

$$\min_{x \in \mathbb{R}^n} \{ E(x, \theta_0) = J(x) + \nu_{\mathcal{H}_u}(x) \} \quad (P_{0\nu})$$

where

$$\mathcal{H}_u = \{ x \in \mathbb{R}^n : \Gamma x = u \},$$

$\theta_0 = (0, u, \Gamma)$, and $\nu$ is the indicator function of the non-empty closed convex set $\mathcal{C}$, i.e., $\nu(x) = 0$ if $x \in \mathcal{C}$ and $\nu(x) = +\infty$ otherwise. Problem $(P_{0\nu})$ can be viewed as a limit of $(P_0)$ as $\mu \to 0^+$. At this stage, it is worth mentioning that though we focus here, for simplicity of exposition, on the squared loss $x \mapsto \frac{1}{2} \| y - \Phi x \|^2$, our results generalize to more general smooth losses, see Section III-E for further details.

The goal of this paper is to assess the recovery performance of $(P_0)$, i.e. to understand how close are the properties of the recovered solution of $(P_0)$ to those of $x_0$. More precisely,
we focus here on the low-noise regime, i.e. when \( \varepsilon \) is small enough, and we investigate stability in \( \ell^2 \) sense, but also, and more importantly, the identifiability of the correct low-dimensional manifold associated to \( x_0 \). This unifies and extends 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.

B. Notations

We recall some basic ingredients from differential geometry and convex analysis that are essential to our exposition. For a function \( J, \text{dom}(J) = \{ x \in \mathbb{R}^p : J(x) < +\infty \} \) is its domain. We denote \( \partial J(x) \) the subdifferential at \( x \) of the proper, lsc and convex function \( J \). Geometrically, when \( x \in \text{dom}(J) \), \( \partial J(x) \) (if non-empty) is the set of gradients of the affine minors of \( J \) supporting it at \( x \). The subdifferential \( \partial J(x) \) is a closed convex set. We denote \( r(C) \) (resp. \( \text{rd}(C) \)) the relative interior of \( C \) (resp. relative boundary), i.e. its interior (boundary) for the topology of its affine hull (the smallest affine space containing \( C \)).

A good source on smooth manifold theory is [32]. A set \( \mathcal{M} \subset \mathbb{R}^n \) is a \( C^2 \)-smooth manifold around a point \( x \in \mathbb{R}^n \), if \( x \in \mathcal{M} \) and \( \mathcal{M} \) consists locally around \( x \) of the solutions of some \( C^2 \)-smooth equations with linearly independent gradients. In this case, the tangent space of \( \mathcal{M} \) at \( x \) is denoted \( T_x(\mathcal{M}) \). We define the tangent model subspace as

\[
T_x = \text{par}(\partial J(x))^\perp,
\]

where \( \text{par}(C) = \mathbb{R}(C - \mathbb{C}) \) is the subspace parallel to the set \( C \subset \mathbb{R}^n \). For a linear space \( T \), we denote \( P_T \) the orthogonal projection on \( T \) and for a matrix \( \Gamma \in \mathbb{R}^{n \times n} \), \( \Gamma_T = P_T \Gamma P_T \).

We use the abbreviation \( O(a_k, b_k) \) to mean \( O(\max(a_k, b_k)) \).

II. PARTLY-SMOOTH FUNCTIONS

Toward the goal of studying the recovery guarantees of problem \((P_{\theta})\), our central assumption will be that \( J \) is a partly smooth function. Partial smoothness of functions was originally defined by [34]. Our definition hereafter specializes it to the case of proper, lsc and convex functions. Rigorously speaking, in the following, one should speak of \( C^2 \)-smooth embedded submanifold of \( \mathbb{R}^p \). Nevertheless, to lighten terminology, we shall state \( C^2 \)-manifold. For a smooth manifold \( \mathcal{M} \) around \( x \in \mathcal{M} \), \( T_x(\mathcal{M}) \) will denote the tangent space to \( \mathcal{M} \) at any point \( x' \) near \( x \) in \( \mathcal{M} \).

**Definition 1.** Let \( J \) be a proper, lsc convex function, and \( x \in \mathbb{R}^p \) such that \( \partial J(x) \neq \emptyset \). \( J \) is partly smooth at \( x \) relative to a set \( \mathcal{M} \) containing \( x \) if

(i) (Smoothness) \( \mathcal{M} \) is a \( C^2 \)-manifold around \( x \) and \( J \) restricted to \( \mathcal{M} \) is \( C^2 \) around \( x \).

(ii) (Sharpness) The tangent space \( T_x(\mathcal{M}) \) is \( T_x(\mathcal{M}) \).

(iii) (Continuity) The set-valued mapping \( \partial J \) is continuous at \( x \) relative to \( \mathcal{M} \).

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

Note that in the previous definition, \( \mathcal{M} \) needs only to be defined locally around \( x \), and it can be shown to be locally unique thanks to prox-regularity of proper, lsc and convex functions, see [27, 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. Examples showing the importance of properties (i)-(iii) and why their individual lack will cause issue are provided in [26, Section 2.2].

**Remark 1** (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 \( x \). Therefore, the regularity property [34, Definition 2.7(iii)] is automatically verified. In view of [34, Proposition 2.4(i)-(iii)], the sharpness property (ii) is equivalent to [34, Definition 2.7(iii)]. The continuity property (iii) is equivalent to the fact that \( \partial J \) is inner semicontinuous at \( x \) relative to \( \mathcal{M} \), that is: for any sequence \( x_k \in \mathcal{M} \) converging to \( x \) and any \( \eta \in \partial J(x_k) \), there exists a sequence of subgradients \( \eta_k \in \partial J(x_k) \) converging to \( \eta \). This equivalent characterization will be essential in the proof of our main result.

A. 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 examples (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). It turns out that all these examples are partly smooth functions as has been shown in [54].

a) \( \ell^1 \) sparsity.: One of the most popular non-quadratic convex regularization is the \( \ell^1 \) norm \( J(x) = \sum_{i=1}^n |x_i| \), which promotes sparsity. Indeed, it is easy to check that \( J \) is partly smooth at \( x \) relative to the subspace

\[
\mathcal{M} = T_x = \{ u \in \mathbb{R}^n \mid \text{supp}(u) \subseteq \text{supp}(x) \}.
\]

For any \( x \in \mathbb{R}^n \), \( \mathcal{M} \) is a linear subspace, which is obviously a \( C^2 \)-manifold. Moreover, on a neighborhood of \( x \) in \( \mathcal{M} \), the \( \ell^1 \) norm is locally linear and thus \( C^2 \). These two facts prove property (i). As far as property (ii) is concerned, since again \( \mathcal{M} \) is a linear subspace, its tangent space \( T_x(\mathcal{M}) \) is nothing but \( T_x(\mathcal{M}) \). Finally, the subdifferential of the \( \ell^1 \) norm is a constant set locally around \( x \) along \( \mathcal{M} \), which in turns shows property (iii). The use of sparse regularization has been popularized in the signal processing literature under the name of basis pursuit [13], and in the statistics literature under the name of Lasso [52].

b) \( \ell^1 - \ell^2 \) 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 \( B \) such that \( \bigcup_{b \in B} b = \{1, \ldots, n\} \). This group structure is enforced by
using typically the mixed $\ell^1-\ell^2$ norm $J(x) = \sum_{b \in B} \|x_b\|$, where $x_b = (x_i)_{i \in b} \in \mathbb{R}^{|b|}$. We refer to [58, 3] and references therein for more details. Unlike the $\ell^1$ norm, and except the case $|b| = 1$, the $\ell^1-\ell^2$ norm is not polyhedral, but can be still be shown to be partly smooth at $x$ relative to the linear manifold

$$\mathcal{M} = T_x = \{x' : \text{supp}_B(x') \subseteq \text{supp}_B(x)\},$$

where

$$\text{supp}_B(x) = \bigcup \{b : x_b \neq 0\}.$$

See [54].

c) Spectral functions.: The natural spectral extension of sparsity to matrix-valued data $x \in \mathbb{R}^{n \times n_0}$ (where $n = n_0^2$) is to impose a low-rank prior, which should be understood as sparsity of the singular values. Denote $x = U_x \text{diag}(\Lambda_x) V_x^*$ an SVD decomposition of $x$, where $\Lambda_x \in \mathbb{R}_{++}^{n_0}$. The nuclear norm is defined as

$$J(x) = \|x\|_* = \|\Lambda_x\|_1. \quad (2)$$

It has been used for instance in machine learning applications [3], matrix completion [45, 7] and phase retrieval [11]. The nuclear norm can be shown to be partly smooth at $x$ relative to the manifold [36, Example 2]

$$\mathcal{M} = \{x' : \text{rank}(x') = \text{rank}(x)\}. \quad (3)$$

More generally, if $J : \mathbb{R}^{n_0} \to \mathbb{R}$ is a permutation-invariant closed convex function, then one can consider the function $J(x) = J(\Lambda_x)$ which can be shown to be a convex function as well [35]. When restricted to the linear space of symmetric matrices, $J$ is partly smooth at $\Lambda_x$ for a manifold $m_{\Lambda_x}$, and if only if $J$ is partly smooth at $x$ relative to the manifold

$$\mathcal{M} = \{U \text{diag}(\Lambda) U^* : \Lambda \in m_{\Lambda_x}, U \in \mathcal{O}_{n_0}\},$$

where $\mathcal{O}_{n_0} \subset \mathbb{R}^{n_0 \times n_0}$ is the group of orthogonal matrices. This result is proved in [14, Theorem 3.19], building upon the work of [15] 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 [14, Theorem 5.3]. The nuclear norm $\|\cdot\|_*$ is a special case where $J(\Lambda) = \|\Lambda\|_1$.

d) Analysis regularizers.: If $J_0 : \mathbb{R}^q \to \mathbb{R} \cup \{+\infty\}$ is a proper lsc convex function and $D \in \mathbb{R}^{n \times q}$ is a linear operator, an analysis regularizer (following the terminology introduced in [18]) is of the form

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

Such a prior controls the low-complexity (as measured by $J_0$) of the correlations between the columns of $D$ and $x$. 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 [49]. The fused Lasso [53] corresponds to $J_0$ being the $\ell^1$-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 [25, 46] to use a family of analysis-type priors where $J_0 = \|\cdot\|_1$ is the nuclear norm.

If $J_0$ is partly smooth at $\alpha = D^* x$ for the manifold $\mathcal{M}^0$, then it is shown in [34, Theorem 4.2] that $J$ is partly smooth at $x$ relative to the manifold

$$\mathcal{M} = \{x' \in \mathbb{R}^n : D^* x' \in \mathcal{M}^0\}.$$
pursuit, proposed in [8], 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 $J_2 = \| \cdot \|_s$.

III. MAIN RESULTS

In the following, we denote $T = T_{x_0}$, $e = P_T(\partial J(x_0)) \in \mathbb{R}^n$.

A. Linearized pre-certificate and minimal norm dual certificate

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 $\Gamma \in \mathbb{R}^{n \times n}$, assuming $\ker(\Gamma) \cap T = \{0\}$, we define $\eta_\Gamma = \Gamma \Gamma^T e$.

Recall that $\Gamma$ is proper lsc convex function, and we suppose that $\text{Im}(\Gamma) \cap \partial J(x_0) \neq \emptyset$ (so-called range or source condition in the inverse problem community). The latter is equivalent to the fact that $x_0$ is a minimizer of $(P_0)$. This is straightforward to see by writing the first-order optimality condition of this convex program.

Definition 3 (Minimal norm certificate). The minimal norm certificate is the vector

$$\tilde{\eta}_\Gamma = \tilde{\Gamma} \tilde{z}_\Gamma,$$

where $\tilde{z}_\Gamma = \text{argmin}_{\tilde{z} \in \partial J(x_0)} \tilde{\Gamma} \tilde{z}$. (5)

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

$$\eta_\Gamma \in \text{ri}(\partial J(x_0)) \implies \tilde{\eta}_\Gamma = \eta_\Gamma,$$

$$\eta_\Gamma \in \text{ri}(\partial J(x_0)) \implies \tilde{\eta}_\Gamma = \eta_\Gamma.$$ (6) (7)

Under either of these conditions, $x_0$ is the unique minimizer to $(P_{0, \Gamma_{x_0}, \tilde{\Gamma}})$.

The proof is postponed to Section V-B.

B. Deterministic model consistency

We first consider the case where $\Phi$ and $w$ (or equivalently $\Gamma$ and $u$) are fixed and deterministic. Our main contribution is the following model consistency theorem, which shows the robustness of the manifold $\mathcal{M}$ associated to $x_0$ to small perturbations on both the observations and the design matrix, provided that $\mu$ (or equivalently $\lambda$) is well chosen. As a product, we also get $\ell^2$ stability.

Theorem 1. Assume that $J$ is locally partially smooth at $x_0$ relative to $\mathcal{M}$ and that there exists $\tilde{\Gamma} \in \mathbb{R}^{n \times n}$ such that

$$\ker(\tilde{\Gamma}) \cap T = \{0\}, \quad \eta_\Gamma \in \text{ri}(\partial J(x_0)).$$ (8)

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

$$\text{max} \left( \|\Gamma - \tilde{\Gamma}\|, \|\epsilon\|\mu^{-1}, \mu \right) \leq C,$$

the solution $x_0$ of $(P_\theta)$ is unique and satisfies

$$x_0 \in \mathcal{M} \quad \text{and} \quad \|x_\theta - x_0\| = O(\|\epsilon\|).$$ (10)

This theorem is proved in Section V-C.

Remark 2 (Stability constants). Observe that the non-degeneracy and restricted injectivity conditions in (8) can be viewed as a geometric generalization of the so-called irreprezentable condition in statistics (see Section III-F for further details). They guarantee in particular that $x_0$ is identifiable in the exact case, i.e. a unique solution to $(P_{0, \Gamma_{x_0}, \Gamma})$. Theorem 1 ensures that under (8), solving $(P_\theta)$ indeed recovers a unique solution $x_\theta$ having the correct model (i.e. $x_\theta \in \mathcal{M}$). In our way of proving model consistency, we also get $\ell^2$ stability to such a small noise. When the underlying model has a low complexity (typically $\mathcal{M}$ has a small dimension), this means that the recovery will be highly stable, which is reflected both in the constant $C$ in (9) and the (local) Lipschitz constant hidden behind the bound $\|x_\theta - x_0\| = O(\|\epsilon\|)$. Obtaining sharp estimates of these constants for general low complexity models (manifolds) is rather challenging and will necessitate even more involved arguments from differential geometry. The case of regularizers $J$ where the partial smoothness manifold is affine was deeply investigated in [54]. There, the authors derived explicit (though quite involved) formulae for $C$ and the Lipschitz constant. It is however easier to see how these constants behave in the small noise limit, i.e. when $\|\epsilon\| \to 0$ with $\mu = C_0\|\epsilon\|$ for some large enough constant $C_0$ ($C_0 \geq 1/C$ from (9)), and where we assume for simplicity that $\Gamma = \tilde{\Gamma}$, indeed recovers a unique minimizer $x_\theta$ to $(P_{0, \Gamma_{x_0}, \Gamma})$.

The triangle inequality, (8), convexity and closedness of $\partial J(x)$
In view of which is valid since uniqueness of $\eta$. This implies $\text{dist}(c, \text{rbd}(\partial J(x_0))) 
lessdist\|\eta - \eta_1\|_\Gamma + c|x_0 - x_0||\eta_1|| < \text{dist}(\eta_1, \text{rbd}(\partial J(x_0)))$

for $c \geq 0$. To simplify the discussion, we suppose that $\text{rbd}(\partial J(x_0)) \subset \text{rbd}(\partial J(x_0))$ for small enough noise (this is true for instance if $J$ is locally polyhedral around $x_0$). This implies $\text{dist}(\eta_1, \text{rbd}(\partial J(x_0))) \leq \text{dist}(\eta_1, \text{rbd}(\partial J(x_0)))$. Consequently, to have $\eta_1 \in \text{r}(\partial J(x_0))$, as required to show uniqueness of $x_0$ in Theorem 1, it is sufficient that

$$||\eta - \eta_1|| + c|x_0 - x_0||\eta_1|| < \text{dist}(\eta_1, \text{rbd}(\partial J(x_0)))$$

which is valid since $\text{dist}(\eta_1, \text{rbd}(\partial J(x_0))) > 0$ owing to (8). In view of (31) and (35), this holds true if

$$C_1 \text{dist}(\eta_1, \text{rbd}(\partial J(x_0))) \leq \mu \leq C_2 \text{dist}(\eta_1, \text{rbd}(\partial J(x_0)))$$

and

$$||\Gamma - \hat{\Gamma}|| \leq C_3 \text{dist}(\eta_1, \text{rbd}(\partial J(x_0))),$$

where $C_1$, $C_2$ and $C_3$ are positive constants that depend on $\eta_1$, $c$ and on the constants in the $O(\cdot)$ terms in (31) and (35). These constants encode again the complexity of $M$ and $x_0$ in them. One can clearly see that the closer $\eta_1$ to $\text{rbd}(\partial J(x_0))$, the smaller the noise that can be tolerated for model stability. Moreover, $\mu$ must be chosen large enough but not too large. In particular, if $\mu = C_1 \text{dist}(\eta_1, \text{rbd}(\partial J(x_0)))^{-1}$, we have from the previous remark that

$$||x_0 - x_0|| \sim \left(1 + \frac{C_1||\eta_1||}{\text{dist}(\eta_1, \text{rbd}(\partial J(x_0)))}\right)||\hat{\Gamma}||,$$

which in turn shows the influence of non-degeneracy on $\ell^2$ stability. The same reasoning can be extended to the scenario where $J$ is the separable sum of sublinear regularizers, in which case the assumption $\text{rbd}(\partial J(x_0)) \subset \text{rbd}(\partial J(x_0))$ can be removed.

Remark 4 (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 $\Phi$ is generally fixed and known, and one then takes $\Gamma = \Gamma = \Phi^\dagger \Phi/p$.

Remark 5 (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 1. This scenario has been considered for sparse recovery (i.e. $J$ the $\ell^1$-norm) by several authors for sparse linear regression and compressed sensing, see e.g. [28, 48, 39].

Remark 6 (Random setting). In statistics or machine learning, one considers a regression problem where the design $\Phi$ and the noise $w$ are random, under the asymptotic regime where the number of observations $p$, i.e. number of rows $\Phi$, grows large, so that $\Gamma$ only reach $\hat{\Gamma}$ in the limit $p \rightarrow +\infty$. See Theorem 2 for details.

Remark 7 (Identification of the manifold). Theorem 1 guarantees that, under some hypotheses on $x_0$ and $\theta$, $x_0$ belongs to $M$. For all the regularizations considered in Section II-A, it turns out that actually $M_{x_0} = M$. This is because, for any $(x, x')$ with $x' \in M_x$ close enough to $x$, one has $M_{x'} = M_x$.

The following proposition, proved in Section V-F, shows that Theorem 1 is in some sense sharp, since the hypothesis $\eta_1 \in \text{r}(\partial J(x_0))$ (almost) characterizes the stability of $M$.

Proposition 2. Suppose that $x_0$ is the unique solution of $(P_{(0, \Gamma_0, \Gamma)})$ and that

$$\ker(\hat{\Gamma}) \cap T = \{0\}, \text{ and } \eta_1 \notin \partial J(x_0). \tag{11}$$

Then there exists $C > 0$ such that if (9) holds, then any solution $x_0$ of $(P_\mu)$ for $\mu > 0$ satisfies $x_0 \notin M$.

In the particular case where $\varepsilon = 0$ (no noise) and $\hat{\Gamma} = \Gamma$, this result shows that the manifold $M$ cannot be correctly identified by solving $(P_\mu)$ for any $\mu > 0$ small enough.

Remark 8 (Critical case). The only case not covered by either Theorem 1 or Proposition 2 is when $\eta_1 \in \text{rbd}(\partial J(x_0))$. In this case, one cannot conclude in general, since depending on the noise $w$, one can have either stability or non-stability of $M$. We refer to [55] where an example illustrates this situation for the 1-D total variation $J = \|D_{\text{DIF}} \cdot 1\|$ (here $D_{\text{DIF}}$ is a discretization of the 1-D derivative operator).

C. Probabilistic model consistency

We now turn to study consistency of our estimator. In this section, we work under the classical setting where $n$ and $x_0$ are fixed as the number of observations $p \rightarrow \infty$. We consider that the design matrix and the noise are random. More precisely, the data $(\varphi_i, w_i)$ are random vectors in $\mathbb{R}^{p} \times \mathbb{R}$, $i = 1, \ldots, n$, where $\varphi_i$ is the $i$-th row of $\Phi$, are assumed independent and identically distributed (i.i.d.) samples from a joint probability distribution such that $E(w_i|\varphi_i) = 0$, finite fourth-order moments, i.e. $E(w_i^4) < +\infty$ and $E(||\varphi_i||^4) < +\infty$. Note that in general, $w_i$ and $\varphi_i$ 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. [31, 59, 3]. Let’s denote $\hat{\Gamma} = E(\xi^\dagger \xi) \in \mathbb{R}^{n \times n}$, where $\xi$ is any row of $\Phi$. We do not make any assumption on invertibility of $\hat{\Gamma}$.

To make the discussion clearer, the canonical parameters $\theta$ will be indexed by $p$. The estimator $x_{\theta_p}$ obtained by solving $(P_{\theta_p})$ for a sequence $\theta_p$ is said to be consistent for $x_0$ if

$$\lim_{p \rightarrow +\infty} \text{Pr}(x_{\theta_p} \text{ is unique}) \rightarrow 1 \text{ and } x_{\theta_p} \text{ converges to } x_0 \text{ in probability.}$$

The estimator is said to be model consistent if

$$\lim_{p \rightarrow +\infty} \text{Pr}(x_{\theta_p} \in M) \rightarrow 1, \text{ where } M \text{ is the manifold associated to } x_0.$$
Theorem 2. If conditions (8) hold and
\[ \mu_p = o(1) \quad \text{and} \quad \mu_p^{-1} = o(p^{1/2}). \] (12)
Then the estimator \( x_{\theta_p} \) of \( x_0 \) obtained by solving (P\( _{\theta_p} \)) is model consistent.

Remark 9 (Sharpness of the criterion). One can also state a probabilistic equivalent to Proposition 2. That is, if \( x_0 \) is the unique solution of (P\( _{\theta_0,F,x_0} \)) and conditions (11) and (12) hold, then the estimator \( x_{\theta_p} \) of \( x_0 \) defined by solving (P\( _{\theta_p} \)) cannot be model consistent.

D. Algorithmic Implications

A popular iterative scheme to compute a solution of (P\( _\theta \)) is the Forward-Backward (F-B) splitting algorithm. A comprehensive treatment of the convergence properties of this algorithm, and other proximal splitting schemes, can be found in the monograph [4]. Starting from some \( x^0 \in \mathbb{R}^n \), the algorithm implements the following iteration
\[ x^{k+1} = \text{Prox}_\gamma J(x^k + \tau_k (u - \Pi x^k)) , \] (13)
where the step size satisfies \( 0 < \tau \leq \tau_k \leq \tau < 2/\|\Pi\| \), and the proximity operator is defined, for \( \gamma > 0 \), as
\[ \text{Prox}_\gamma J(x) = \arg\min_{x'} \frac{1}{2} \|x - x'\|^2 + \gamma J(x') . \] (14)

The following theorem shows that the F-B algorithm correctly identifies the manifold \( \mathcal{M} \) after a finite number of iterations.

Theorem 3. Suppose that the assumptions of Theorem 1 hold. Then, there exists \( k_0 \) large enough, such that for all \( k \geq k_0 \), the F-B iterates satisfy \( x^k \in \mathcal{M} \).

Proof. Inspection of the proof of Theorem 1 shows that the solution \( x_0 \) of (P\( _\theta \)), which is unique, is such that the vector \( \eta_0 = \frac{u - \Pi x_0}{\gamma} \) satisfies \( \eta_0 \in \text{ri}(\partial J(x_0)) \) when (8) and (9) hold. Moreover, as \( x_0 \in \mathcal{M} \), \( x_0 \) is near \( x_0 \), and \( J \) is locally partly smooth at \( x_0 \) relative to \( \mathcal{M} \), it is also partly smooth at \( x_0 \) relative to the same manifold \( \mathcal{M} \). Altogether, this implies that the assumptions of [38, 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 (8) holds and \( (\|\Gamma - \Gamma\|, \|\epsilon\|/\mu, \mu) \) are sufficiently small.

E. General Loss Functions

For the sake of simplicity, we have described our contributions with the squared loss function \( u \in \mathbb{R}^p \rightarrow \frac{1}{2} \|y - u\|^2 \). 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
\[ \min_x F(\Phi x, y) + \lambda J(x) , \]
where the loss function \( F : \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R} \) fulfills the following assumptions:

(A.1) For any \( y \in \mathbb{R}^p \), \( F(\cdot, y) \in C^2(\mathbb{R}^p) \) is \( \sigma_m \)-strongly convex and \( \nabla_1 F(\cdot, y) \) is \( \sigma_M \)-Lipschitz continuous, \( \sigma_m > 0 \) and \( \sigma_M > 0 \).

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

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 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 1 and in a similar way our other contributions) remains valid, and one simply needs to replace condition (9) by
\[ \max \{ \|\Gamma - \tilde{\Gamma}\|, \|\epsilon\|\mu^{-1}, \mu \} \leq C , \] (15)
where now
\[ \tilde{\Gamma} = \frac{1}{p} \Phi^* \nabla_1^2 F(y,y) \Phi \quad \text{and} \quad \tilde{\epsilon} = \frac{1}{p} \Phi^* \nabla_1^2 F(y,y) w , \]
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 V-D.

F. Relation to Previous Works

a) Works on linear convergence rates.: Following the pioneer work [6] (who study convergence in terms of Bregman divergence), there is a large amount of works on the study conditions under which \( \|x_0 - x_0\| = O(\|\epsilon\|) \) (so-called linear convergence rate) where \( x_0 \) is any solution of (P\( _\theta \)), see for instance the book [50] for an overview of these results. The initial work of [24] proves a sharp criteria to ensure linear convergence rate for the \( \ell^1 \) norm, and this approach is further extended to arbitrary convex functions by [23] and [19], who respectively proved linear convergence rates in terms of the penalty \( J \) and \( \ell^2 \)-norm.

These works show that if
\[ \ker(\Gamma) \cap T = \{0\} \quad \text{and} \quad \exists \eta \in \text{Im}(\Gamma) \cap \text{ri}(\partial J(x_0)) \] (16)
(which is often called the source condition), then linear convergence rate holds. Note that condition (8) implies (16), but it is stronger. Indeed, condition (16) does not ensure model consistence (10), which is a stronger requirement. Model consistency requires, as we show in our work, the use of a special certificate, the minimal norm certificate \( \eta_\Gamma \), which is equal to \( \eta_\Gamma \) if \( \eta_\Gamma \in \text{ri}(\partial J(x_0)) \) (see Proposition 1).

b) Works on model consistencies.: 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 [21]. In this setting, the result (10) corresponds to the correct identification of the support, i.e. \( \text{supp}(x_0) = \text{supp}(x_0) \). Condition (8) for \( J = \| \cdot \|_1 \) is known in the statistics literature under the name “irrepresentable condition”, see e.g. [59]. [31] have shown estimation consistency for Lasso for fixed \( n \) and \( x_0 \).
and asymptotic normality of the estimates. The authors in [59] proved Theorem 2 for $J = \| \cdot \|_1$, though under slightly different assumptions on the covariance and noise distribution. A similar result was established in [30] for the elastic net, namely the group Lasso nuclear/trace norm minimization, under a specialization of (8) to these two penalties and in an asymptotic setting. Note that these previous works assume that the asymptotic covariance $\Gamma$ is invertible. We do not impose such an assumption, and only require the weaker restricted injectivity condition $\Gamma = \tilde{\Gamma}$ that these previous works assume that the asymptotic covariance problems seeks conditions to ensure smoothness of the mapping $\theta \mapsto x_\theta$ where $x_\theta$ is a minimizer of $f(\cdot, \theta)$, see for instance [40, 47, 5].

This is usually guaranteed by the non-degenerate source condition and restricted injectivity condition (16), which, as already reviewed above, ensures linear convergence rate, and hence Lipschitz behaviour of this mapping. The result captured by Theorem 1 goes one step further, by assessing that $M_{x_\theta}$ is a stable manifold (in the sense of [57]), since the minimizer $x_\theta$ is unique and remains in $M_{x_\theta}$ for $\theta$ close to $\theta_0$. Our starting point for establishing Theorem 1 is the inspiring work of [34] 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 $U-V$-decompositions developed in [33]. 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 [34, 37]. A important and distinctive feature of Theorem 1 is that, partial smoothness of $J$ at $x_0$ relative to $M$ transfers to $E(\cdot, \theta)$ for $\lambda > 0$, but not when $\lambda = 0$ in general. In particular, [34, Theorem 5.7] does not apply to prove our claim.

IV. 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 III-F, previous consistency results due to [3] only apply to the overdetermined setting, while our result tackles arbitrary design $\Phi$ by only requiring the weaker injectivity condition (8). For simplicity of exposition, we consider recovery of square matrices of size $n = n_\circ \times n_\circ$, but the same holds for arbitrary rectangular matrices.

A. Irrepresentability Criterion IC

The nuclear norm, defined in (2), turns out to be the tightest convex relaxation of the rank function on the spectral ball. It is then the best convex candidate to enforce a low-rank prior [20]. It is moreover partly smooth at any $x_0 \in \mathbb{R}^{n_\circ \times n_\circ}$ relative to the manifold $M$ of fixed rank $r = \text{rank}(x_0)$ defined in (3).

Let $x_0 = U \text{diag}(\sigma(x_0)) V^*$ be a reduced rank-$r$ SVD decomposition of $x_0$, where $V, U \in \mathbb{R}^{n_\circ \times r}$ have orthonormal columns and $\sigma(x_0) \in (\mathbb{R}_+^*)^r$ is the vector of singular values of $x_0$. The subdifferential of the nuclear norm at $x_0$ reads (see for instance [10])

$$\partial \| \cdot \|_* (x_0) = \{ \eta \in \mathbb{R}^{n_\circ \times n_\circ} : \eta T = e \quad \text{and} \quad \| \eta S \| \leq 1 \} ,$$

where $\| \cdot \|$ is the operator norm, $T = T_+(M)$, $S = T_-^\perp$ and $e = P_{T^\perp}(\partial J(x_0))$, with

$$T = \{ UA^* + BV^* : A, B \in \mathbb{R}^{n_\circ \times r} \} \quad \text{and} \quad e = UV^*.$$
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 \(\partial \| \cdot \|_+ (x_0)\) is formed by subgradients \(\eta\) satisfying the inequality in (17) strictly. Thus, condition (8) in the case \(\Gamma = \Gamma^+\) takes the analytical form

\[
\eta^r \in \mathfrak{r}(\partial J(x_0)) \iff \text{IC}(x_0) < 1 , \tag{18}
\]

where

\[
\text{IC}(x_0) = \| P_s \Gamma^+ e \| .
\]

The value of \(\text{IC}(x_0)\) can then be easily computed. Loosely speaking, the smaller the quantity \(1 - \text{IC}(x_0)\) is, the further \(\eta^r\) is from the relative boundary of \(\partial J(x_0)\), and in turn the smaller the stability constant controlling \(\| x_0 - x_0^\|_E \|\) in (10) is.

**B. Recovery from Gaussian Measurements**

Bounding \(\text{IC}\) for an arbitrary operator \(\Phi\) and matrix \(x_0\) is in general difficult. It is however possible to leverage tools from random matrix theory to obtain sharp upper-bounds when \(\Phi\) is drawn from certain matrix ensembles. This strategy has been deployed to study matrix completion problems, see for instance [7, 12]. Another problem on which we now focus is when \(\Phi\) is drawn from the standard Gaussian ensemble, i.e. its entries are independent identically distributed from \(N(0,1)\). The following result, proved by [10], shows that \(\text{IC}(x_0) < 1\) with high probability as soon as \(p\) is larger than \(6rn_0\) (up to negligible terms).

**Proposition 3** ([10], Theorem 1.2). Let \(x_0 \in \mathbb{R}^{n_0 \times n_0}\) such that \(\text{rank}(x_0) = r\). If

\[
p \geq \delta r(6n_0 - 5r) \tag{19}
\]

for some \(\delta > 1\), then \(\text{IC}(x_0) < 1\) with probability at least \(1 - 2e^{(1 - \delta)n_0/8}\).

Combining this result with Theorem 1, this shows that under the scaling (19) of \((p,n_0,r)\), one obtains with high probability on the design matrix a rank-consistent estimation of the unknown matrix \(x_0\), 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 \(\text{IC}(x_0)\) for either a varying \(p\) 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 (19). This suggests numerically that the upper-bound (19) is indeed sharp.

**C. Forward-Backward Model Consistency and Unconsistency**

As detailed in Section III-D, our theoretical analysis of model consistency also sheds light on the behavior of proximal splitting algorithms, and in particular of the celebrated F-B scheme (13). In the special case \(J = \| \cdot \|_s\) considered in this section, the proximal mapping (14) at \(x \in \mathbb{R}^{n_0 \times n_0}\) is computed by simply soft-thresholding the singular values

\[
\text{Prox}_{\gamma \| \cdot \|_s}(x) = U \text{diag}(\text{Prox}_{\gamma \| \cdot \|_s}(|\sigma(x)|))V^* , \tag{20}
\]

where

\[
\text{Prox}_{\gamma \| \cdot \|_s}(s) = (\text{sign}(s_i) \max(0, |s_i| - \gamma))_i , \tag{21}
\]

where \(x = U \text{diag}(\sigma(x))V^*\) is a reduced SVD decomposition of \(x\).

As in the previous section, we consider here a compressed sensing scenario, where again \(\Phi \in \mathbb{R}^{p \times n_0^2}\) is drawn from the standard Gaussian ensemble, and \(x_0 \in \mathbb{R}^{n_0 \times n_0}\) has low-rank. The observations \(y = \Phi x_0 + w \in \mathbb{R}^p\) are generated with an additive zero-mean white Gaussian noise \(w\) of standard deviation \(10^{-3} |\Phi x_0|\) (but the same conclusion holds for a noise of arbitrary small amplitude). We then compute (approximately) a minimizer of (1) using the F-B iterations (13), tuning the regularization parameter \(\mu = C_0|w|\) in accordance to the noise level, as prescribed by Theorems 1 and 3. As detailed in these theorems, the value of \(C_0\) is chosen large enough to obtain the desired denoising effect (otherwise the solution does not have a low complexity), but its precise value does not affect the observed identification results we describe below.

In the numerical results reported hereafter, we used \(n_0 = 20, P = 3n_0^2/4 = 300\) and \(\text{rank}(x_0) = 3\).

Figure 2 shows how the F-B iterations behave radically differently depending on whether the non-degeneracy condition \(\text{IC}(x_0) < 1\) holds or not. Each curve shows the evolution of \(\text{rank}(x^k)\) during the course of iterations, for different randomized instances of the low-rank matrix \(x_0\) to recover. As predicted by Theorem 3, one can see that for those \(x_0\) where model consistency holds (i.e. \(\text{IC}(x_0) < 1\)), plotted in red, F-B converges with the correct rank, i.e. \(\text{rank}(x^k) = \text{rank}(x_0)\).
Proof of (7) Let $S = T^\perp$. Problem (5) can be conveniently rewritten as

$$\hat{z}_k = \arg\min_{\hat{z}} \|z\|^2 \quad \text{subject to} \quad \tilde{\Gamma} T z = e \quad \tilde{\Gamma} S z \in P_S(\partial J(x_0)).$$

The fact that $\hat{\eta}_k = \tilde{\Gamma} z_k \in \text{ri}(\partial J(x_0))$ implies $P_S \hat{\eta}_k = P_S \tilde{\Gamma} z_k \in \text{ri}(P_S \partial J(x_0))$, and thus, the second constraint in the last problem is inactive. We then recover problem (22), which in turn implies that $\hat{\eta}_k = \hat{\eta}_k$.

Proof of uniqueness. See Proposition 4.

C. 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 $\Phi_k \in \mathbb{R}^{p_k \times n}$. Assume that

$$(\Gamma_k, \varepsilon_k, \mu_k^-1, \mu_k) \rightarrow (\hat{\Gamma}, 0, 0).$$

Then proving Theorem 1 boils down to showing that for $k$ large enough, the solution $x_k$ of $(P_{\theta_k})$ is unique and satisfies $x_k \in S$.

a) Constrained problem. Consider the following non-smooth, generalized convex, minimization problem

$$x_k \in \text{Argmin}_{x \in M} E(x, \theta_k)$$

where $K$ is an arbitrary fixed convex compact neighborhood of $x_0$.

The following key lemma establishes the convergence of $x_k$ to $x_0$.

Lemma 1. Under conditions (8) and (23), $x_k \rightarrow x_0$.

Proof. We denote $\|u\|_2^2 = \langle u, u \rangle$ for any positive semidefinite matrix $\Gamma$. Under condition (8), Proposition 4 implies that $x_0$ is the unique solution of $(P_{\theta_0})$. By optimality of $x_k$ one has $E(x_k, \theta_k) \leq E(x_0, \theta_k)$ and hence

$$\frac{1}{2} \|x_k\|^2_{\tilde{\Gamma} k} - \langle x_k, \Gamma_k x_0 + \varepsilon_k \rangle + \mu_k J(x_k) \leq \frac{1}{2} \|x_0\|^2_{\tilde{\Gamma} k} - \langle x_0, \Gamma_k x_0 + \varepsilon_k \rangle + \mu_k J(x_0)$$

which is equivalently stated as

$$\frac{1}{2} \|x_k - x_0\|^2_{\tilde{\Gamma} k} - \langle x_k - x_0, \varepsilon_k \rangle + \mu_k J(x_k) \leq \mu_k J(x_0).$$

Since $x_k \in K$, the sequence $(x_k)_k$ is bounded, and we let $x^*$ be any cluster point. Using (23), that $J$ is non-negative and lsc, and $J(x_k)$ are bounded, we have

$$\lim \sup_{k \rightarrow \infty} (\mu_k J(x_k)) \leq \lim_{k \rightarrow \infty} \mu_k \lim \sup_{k \rightarrow \infty} J(x_k) = 0 \quad \text{and} \quad \lim \inf_{k \rightarrow \infty} (\mu_k J(x_k)) \geq \lim_{k \rightarrow \infty} \mu_k \lim \inf_{k \rightarrow \infty} J(x_k) \geq J(x^*) \lim_{k \rightarrow \infty} \mu_k = 0,$$

and thus $\lim_{k \rightarrow \infty} (\mu_k J(x_k)) = 0$. Consequently, passing to the limit in (25), using (23), and continuity of the inner product and the norm, shows that $\|x^* - x_0\|_{\tilde{\Gamma} k}^2 \leq 0$, or equivalently

Proof of (6) Under condition $\ker(\Phi) \cap T = \{0\}$, we have from the definition of $\tilde{\Gamma}_T^+$, that

$$z_k = \tilde{\Gamma}_T^+ \varepsilon = \arg\min_{z} \|z\| \quad \text{subject to} \quad \tilde{\Gamma}_T z = e$$

and thus

$$\hat{\eta}_k = \tilde{\Gamma}_T z_k .$$

Clearly, the constraint set of problem (22) includes that of (5), which entails

$$\|z_k\| \leq \|z_k\| .$$

If $\hat{\eta}_k \in \text{ri}(\partial J(x_0))$, then $z_k$ is also a feasible point of problem (5) and thus

$$\|z_k\| \leq \|z_k\| .$$

Altogether, we get that $\|z_k\| = \|z_k\|$ and, since $\hat{z}_k$ is the unique minimizer of (5), we get that $\hat{z}_k = z_k$, which implies that $\hat{\eta}_k = \hat{\eta}_k$.

Fig. 2. Evolution of rank($x_k$) as a function of $k$ during the progress of Forward-Backward iterations (13) to solve (1) using observations $y = \Phi x_0 + \varepsilon$. The light red (resp. blue) curves represent the evolution for an input $x_0$ satisfying IC($x_0$) < 1 (resp. IC($x_0$) > 1). The bold red (resp. blue) curve is the average of the light red (resp. blue) curves.
$\tilde{x}^* = \tilde{x}_{x_0}$, i.e. $x^*$ is a feasible point of $(\mathcal{P}_{0,\tilde{x}_{x_0},\tilde{t}})$. Furthermore, since $\frac{1}{2} \|x_k - x_0\|_{T_k}^2 \geq 0$, (25) yields

$$-\langle x_k - x_0, \frac{\varepsilon_k}{\mu_k} \rangle + J(x_k) \leq J(x_0).$$

Passing again to the limit, using lower semicontinuity of $J$, (23) and continuity of the inner product, we then get

$$J(x^*) \leq \liminf_{k \to \infty} J(x_k)$$

$$= \liminf_{k \to \infty} \left( -\langle x_k - x_0, \frac{\varepsilon_k}{\mu_k} \rangle + J(x_k) \right)$$

$$\leq \limsup_{k \to \infty} \left( -\langle x_k - x_0, \frac{\varepsilon_k}{\mu_k} \rangle + J(x_k) \right)$$

$$= \limsup_{k \to \infty} J(x_k) \leq J(x_0).$$

Combining this with the previous claim on feasibility of $x^*$ for $(\mathcal{P}_{0,\tilde{x}_{x_0},\tilde{t}})$ allows to conclude that $x^*$ is a solution of $(\mathcal{P}_{0,\tilde{x}_{x_0},\tilde{t}})$. Since $x_0$ is unique, this leads to $x^* = x_0$. \ □

We now aim at showing that for $k$ large enough, $x_k$ is the unique solution of $(\mathcal{P}_{0_k})$. Since $(\mathcal{P}_{0_k})$ 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 $\mathcal{M}$ [44]. In particular, $x_k = P_M(x_k)$ and $x_0 = P_M(x_0)$. Moreover, $P_M$ is of class $C^1$ near $x_k$ [36, Lemma 4]. Thus, $C^2$ differentiability shows that

$$x_k - x_0 = P_M(x_k) - P_M(x_0) = D P_M(x_k)(x_k - x_0) + R(x_k)$$

where $R(x_k) = O(\|x_k - x_0\|^2)$ and where $D P_M(x_k)$ is the derivative of $P_M$ at $x_k$. Using [36, Lemma 4], and recalling that $T_k = T_{x_k}(\mathcal{M})$ by the sharpness property, the derivative $D P_M(x_k)$ is given by $D P_M(x_k) = P_{T_k}$. Inserting this in (30), we get

$$\Gamma_k P_{T_k} (P_{T_k} (x_k - x_0) + R(x_k)) - P_{T_k} \varepsilon_k + \mu_k \varepsilon_k x_0 = 0.$$  

Using (27), $\Gamma_k$ has full rank, and thus

$$x_k - x_0 = \Gamma_k^+ \varepsilon_k (\varepsilon_k - \mu_k \varepsilon_k x_0 - \Gamma_k R(x_k)),$$

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

$$x_k - x_0 = O(\|x_k\|, \mu_k).$$

Convergence of the dual variables: We define $\eta_k = \frac{\eta_k - \mu_k x_k}{\mu_k}$. Arguing as above, and using (30) we have

$$\mu_k \eta_k = \varepsilon_k + \Gamma_k (x_k - x_0)$$

$$= \varepsilon_k - \Gamma_k \mu_k \varepsilon_k x_0 - \Gamma_k R(x_k))$$

$$= \varepsilon_k - \Gamma_k P_{T_k} \Gamma_k^+ \varepsilon_k (\varepsilon_k - \mu_k \varepsilon_k x_0 - \Gamma_k R(x_k))$$

$$= \Gamma_k^+ \varepsilon_k + P_{V_{T_k^+}} \Gamma_k R(x_k) + \mu_k \Gamma_k \varepsilon_k x_0,$$

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

$$\|\eta_k - \tilde{\eta}\| = O\left(\left\|\varepsilon_k\right\| \mu_k^{-1}, \left|\Gamma_k^+ \varepsilon_k x_0 - \tilde{\eta}\right|, \left\|\Gamma_k\right\| \|x_k - x_0\|^2 \mu_k^{-1}\right).$$

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

$$\|e_{x_k} - e\| = O(\|x_k - x_0\|).$$
Using the triangle inequality, we get
\[ |\Gamma_k^+ \Gamma_k^+ - \tilde{\Gamma}^+_T| = O\left(\|\Gamma_k - \tilde{\Gamma}\|, \|\tilde{\Gamma}^+_T - \tilde{\Gamma}\|\right).\]
Again, since \( \Gamma_k^+ \rightarrow \tilde{\Gamma}^+_T \), we have \( \|\Gamma_k^+ \Gamma_k^+ - \tilde{\Gamma}^+_T\| = O(1) \). Moreover, \( A \rightarrow A^+ \) is smooth at \( A = \Gamma_T \) along the manifold of matrices of constant rank, and \( M \) is a \( C^2 \) manifold near \( x_0 \). Thus
\[
|\Gamma_k^+ \Gamma_k^+ - \tilde{\Gamma}^+_T| = O\left(\|\Gamma_k - \tilde{\Gamma}\|, \|\Gamma_k - \tilde{\Gamma}\|\right).
\]
This shows that
\[
|\Gamma_k \Gamma_k^+ - \tilde{\Gamma}^+_T| = O\left(\|\Gamma_k - \tilde{\Gamma}\|, \|\Gamma_k - \tilde{\Gamma}\|\right). \tag{33}
\]
Putting (32) and (33) together implies
\[
|\Gamma_k \Gamma_k^+ - \tilde{\Gamma}^+_T| = O\left(\|\Gamma_k - \tilde{\Gamma}\|, \|\Gamma_k - \tilde{\Gamma}\|\right). \tag{34}
\]
Altogether, we get the bound
\[
|\eta_k - \eta| = O\left(\|\Gamma_k - \tilde{\Gamma}\|, \|\Gamma_k - \tilde{\Gamma}\|\right) \tag{35}
\]
Since \( \|\Gamma_k - \tilde{\Gamma}\| \) is bounded according to (31), we arrive at
\[
|\eta_k - \eta| = O\left(\|\Gamma_k - \tilde{\Gamma}\|, \|\Gamma_k - \tilde{\Gamma}\|\right) \tag{36}
\]
According to (35) and Lemma 1, under (23), \((x_k, \eta_k) \rightarrow (x_0, \eta)\). Condition (37) is equivalent stated as, for each \( k \)
\[
\exists z_k \in T_{x_k}^\perp, \forall \eta \in \partial J(x_k), \langle z_k, \eta - \eta_k \rangle \geq 0, \tag{38}
\]
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 \( T_{x_k}^\perp \rightarrow T^\perp \) because \( M \) is a \( C^2 \) manifold, one has that \( z^* \in T_{\tilde{\Gamma}}^\perp \). We now show that
\[
\forall v \in \partial J(x_0), \langle z^*, v - \eta \rangle \geq 0. \tag{39}
\]
Indeed, let us consider any \( v \in \partial J(x_0) \). In view of the continuity property in Definition 1(iii) \( \partial J \) is continuous at \( x_0 \) along \( M \), so that since \( x_k \rightarrow x_0 \) there exists \( v_k \in \partial J(x_k) \) with \( v_k \rightarrow v \). Applying (38) with \( \eta = v_k \) gives \( \langle z_k, v_k - \eta \rangle \geq 0 \). Taking the limit \( k \rightarrow \infty \) in this inequality leads to (39), which contradicts the fact that \( \eta \in \partial J(x_0) \). In view of (36) and (27), using Proposition 4 shows that \( x_k \) is the unique solution of \((P_{\theta_k})\).

### D. General Loss Function

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

**a) Proof of Proposition 4:** It follows from assumption (A.1) that \( F(x, y) \) is strictly convex, and the uniqueness follows from [38, Theorem A.1].

**b) Proof of Lemma 1:** Problem (24) now reads
\[
x_k \in \text{Argmin}_{x \in M \cap K} F(\Phi_k(x, y), \lambda_k J(x)) \tag{A.2}
\]
Optimality of \( x_k \) entails
\[
F(\Phi_k(x_k, y), \lambda_k J(x_k)) \leq F(\Phi_k(x_0, y) + \lambda_k J(x_0)) \tag{A.1}
\]
By assumptions (A.1)-(A.2), we have the following useful inequalities for any \( u \in \mathbb{R}^p \), see e.g. [41, p. 57 and 64]
\[
\frac{\sigma_m}{2} |y - u|^2 \leq F(u, y) - F(y, y) = \langle \nabla F(y, y), u - y \rangle \leq \frac{\sigma_M}{2} |y - u|^2.
\]
It then follows that
\[
F(\Phi_k(x_k, y), \lambda_k J(x_k)) - F(\Phi_k(x_0, y), \lambda_k J(x_0)) \geq \frac{\sigma_m}{2} |y_k - \Phi_k(x_k)|^2 - \frac{\sigma_M}{2} |w_k|^2 \tag{A.2}
\]
and therefore
\[
\lambda_k J(x_0) \geq \frac{\sigma_m}{2} |y_k - \Phi_k(x_k)|^2 - \frac{\sigma_M}{2} |w_k|^2 + \lambda_k J(x_k)
\]
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_m}{2} |x_k - x_0|^2 \geq \frac{\sigma_m}{2} |y_k - \Phi_k(x_k)|^2 - \frac{\sigma_m}{2} |x_k - x_0| \ Price of the lemma allows to prove that \( x_k \to x_0 \).
that the fourth order moments are finite, we get
of sample covariances, which apply thanks to the assumption
1 as

\[ P_{T_k} \Phi_k^* \nabla_1 F(\Phi_k x_k, y_k) + \lambda_k e_{x_k} = 0. \]

Using again assumptions (A.1)-(A.2) and expanding
\[ \nabla_1 F(\Phi_k x_k, y_k) \]
first to the order, we obtain

\[ \nabla_1 F(\Phi_k x_k, y_k) = \nabla^2_1 F(y_k, y_k) \Phi_k (x_k - x_0) - \nabla^2_1 F(y_k, y_k) w_k + O(\|x_k - x_0\|^2) + O(\|w_k\|^2) \]

Dividing by \( p \), plugging this expansion back into the above
first-order (criticality) condition, and grouping the \( O(.) \) terms, condition (29) becomes

\[ \hat{\Gamma}_{k,T_k} (x_k - x_0) - P_{T_k} \hat{\xi}_k + \mu_k e_{x_k} + P_{T_k} (n^{-1} \Phi_k^* + \hat{\Gamma}_k) R(x_k) + P_{T_k} \Phi_k Q(n^{-1/2} w_k) = 0, \]

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

### E. Proof of Theorem 2

It is sufficient to check that (9) is in force with probability
1 as \( p \to +\infty \). Owing to classical results on convergence of sample covariances, which apply thanks to the assumption
that the fourth order moments are finite, we get \( \Gamma_p - \Gamma = \Gamma_p \) and \( \frac{1}{2} (\hat{\xi}, w) = O_p \left( p^{-1/2} \right) \), where we used the
assumption that \( \mathcal{E} \left( \hat{\xi}, w \right) = 0 \). As \( p \) is fixed, it follows that \( \|\Gamma_p - \hat{\Gamma}\| = O_p \left( p^{-1/2} \right) \) and \( \|\hat{\xi}_p\| = O_p \left( p^{-1/2} \right) \). Thus under the scaling (12), we get

\[ \left( \|\Gamma_p - \hat{\Gamma}\|, \|\hat{\xi}_p\|p^{-1/2}, \mu_p \right) = \left( O_p \left( p^{-1/2} \right), \frac{1}{\mu_p p^{1/2}}, O_p \left( 1 \right) \right) \]

\[ = \left( O_p \left( p^{-1/2} \right), o(1), O_p \left( 1 \right) \right) \]

\[ = \left( O_p \left( p^{-1/2} \right), o(1), o(1) \right) \]

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

### F. Proof of Proposition 2

Let \( (x_k) \) be a sequence of solutions to the constrained
problem (24). Since \( x_0 \) is the unique minimizer to \( (P_{0, \xi}, \hat{\Gamma}) \)
and (8) is satisfied, \( \eta \) is well-defined. Moreover, arguing as
in the proof of Lemma 1 and Theorem 1, under condition (9),
we have \( (x_k, \eta_k) \to (x_0, \eta_0) \), and \( \eta_k \in \eta_0 + C\mathbb{B} \).

Let \( \tau = \text{dist}(\eta_0, \partial J(x_0)) = \inf_{\eta \in \partial J(x_0)} \|\eta - \eta_0\| \). Since
\( \partial J(x_0) \) is a non-empty, closed and convex set, the infimum is attained and one has \( \tau > 0 \) since \( \eta_0 \notin \partial J(x_0) \).

We now prove the claim by contradiction. Let \( x_j \) be a solution of \( (P_{\theta_j}) \) such that (9) holds at \( \theta_j \) for \( j \) sufficiently
large (taking \( C \) smaller if necessary so that \( C < \tau \)), and
suppose that \( x_j \in \mathcal{M} \). Thus, \( x_j \) is also a solution of (24) for \( \theta_j \), whence it follows that \( \eta_j \in \eta_0 + C\mathbb{B} \). Using the triangle
inequality, we then get

\[ \text{dist}(\eta_j, \partial J(x_0)) \geq \tau - C > 0. \]

Now, in view of the continuity property in Definition 1((iii)),
we have \( \partial J(x_k) \to \partial J(x_0) \) along \( \mathcal{M} \). This is equivalent,
since \( \partial J(x_0) \) is closed and using [47, Corollary 4.7], to
\( \text{dist}(\eta, \partial J(x_k)) \to \text{dist}(\eta, \partial J(x_0)) \) for every \( \eta \in \mathbb{R}^p \), i.e.

\[ \forall \delta > 0, \exists k_0, \forall k \geq k_0, \forall \eta \in \mathbb{R}^p, \]

\[ |\text{dist}(\eta, \partial J(x_k)) - \text{dist}(\eta, \partial J(x_0))| < \delta \]

In particular, as \( x_j \) is a minimizer of \( (P_{\theta_j}) \) for \( j \) large enough,
we have \( \eta_j \in \partial J(x_j) \), and thus \( \text{dist}(\eta_j, \partial J(x_0)) < \delta \), leading to a contradiction with (40). Hence, \( x_j \notin \mathcal{M} \).

### VI. 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 under-
lying 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 al-
gorithm 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

variation component and an oscillating component”. In: Journal of Mathematical Imaging and Vision 22 (2005),

ple Kernel Learning”. In: Journal of Machine Learning

In: Journal of Machine Learning Research 9 (2008),
pp. 1019–1048.

and Monotone Operator Theory in Hilbert Spaces.


---

**Samuel Vaiter** is junior researcher (CR2) at the Centre Nationale de Recherche Scientifique (CNRS), working in DMA, École Normale Supérieure, Dijon. His current research interests focus on variational regularization in signal and image processing, convex analysis, sparsity and risk estimation. He has studied Applied Mathematics and Theoretical Computer Science in Lyon and Paris. He has obtained in 2014 a PhD in Applied Mathematics from Université Paris-Dauphine. He has then worked as a post-doc in CMAP, Ecole Polytechnique, Palaiseau.

---

**Gabriel Peyré** is senior researcher at the Centre Nationale de Recherche Scientifique (CNRS), working in DMA, École Normale Supérieure, Paris. His research is focused on developing mathematical and numerical tools in sparse regularization and optimal transport, with applications in computer vision, graphics and neurosciences. Since 2005 Gabriel Peyré has co-authored 65 papers in international journals, 70 conference proceedings in top vision and image processing conferences, and two books. He is the creator of the “Numerical tour of signal processing” (www.numerical-tours.com), a popular online repository of Matlab/Python/Julia resources to teach modern signal and image processing. His research was supported by a ERC starting grant (SIGMA-Vision, 2010-2015) and is now supported by a ERC consolidator grant (NORIA 2017-2021).

---

**Jalal Fadili** is a Full Professor at Ecole National Supérieure d’Ingénieurs de Caen, and Junior member of Institut Universitaire de France since Oct. 2013. He holds several scientific management positions (editorial activities, national excellence research networks). He also held visiting positions at several universities (QUT-Australia, Stanford, CalTech, EPFL-Switzerland, MIT). In the last decade, he has been an invited or plenary speaker at various international events. His research interests include mathematical signal and image processing, mathematical statistics, inverse problems, variational methods and regularization theory, and non-smooth optimization. His areas of application include medical and astronomical imaging. He has published more than 170 papers in the leading journals and conferences of these fields, 7 book chapters and 2 books.