HAL
open science

The Degrees of Freedom of Partly Smooth Regularizers

Samuel Vaiter, Charles-Alban Deledalle, Gabriel Peyré, Jalal M. Fadili, Charles H Dossal

To cite this version:

Samuel Vaiter, Charles-Alban Deledalle, Gabriel Peyré, Jalal M. Fadili, Charles H Dossal. The Degrees of Freedom of Partly Smooth Regularizers. 2014. hal-00981634v2

HAL Id: hal-00981634

https://hal.science/hal-00981634v2

Preprint submitted on 23 Jul 2014 (v2), last revised 10 Feb 2016 (v4)

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

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

The Degrees of Freedom of Partly Smooth Regularizers

Samuel Vaiter · Charles Deledalle · Gabriel Peyré · Jalal Fadili · Charles Dossal

Abstract In this paper, we are concerned with regularized regression problems where the prior penalty is a partly smooth function relative to a linear manifold. This encompasses as special cases the Lasso (ℓ^1 regularizer), the group Lasso ($\ell^1 - \ell^2$ regularizer) and the ℓ^∞-norm regularizer penalties. This also includes so-called analysis-type priors, i.e. composition of the previously mentioned penalties with linear operators, typical examples being the total variation or fused Lasso priors. We study the sensitivity of *any* regularized minimizer to perturbations of the observations and provide its precise local parameterization. Our main result shows that, when the observations are outside a set of zero Lebesgue measure, the predictor moves locally stably along the same linear space as the observations undergo small perturbations. This local stability is a consequence of the piecewise smoothness of the regularizer, which in turn plays a pivotal role to get a closed form expression for the variations of the predictor w.r.t. observations which holds almost everywhere. When the perturbation is random (with an appropriate continuous distribution), this allows us to derive an unbiased estimator of the degrees of freedom and of the risk of the estimator prediction. Our results hold true without requiring the design matrix to be full column rank. They generalize those already known in the literature such as the Lasso problem, the general Lasso problem (analysis ℓ^1-penalty), or the group Lasso where existing results for the latter assume that the design is full column rank.

Keywords Degrees of freedom · Sparsity · Model selection · o-minimal structures · Semi-algebraic sets · Group Lasso · Total variation · Partial smoothness

Samuel Vaiter, Gabriel Peyré
CEREMADE, CNRS, Université Paris-Dauphine, Place du Maréchal De Lattre De Tassigny, 75775 Paris Cedex 16, France
E-mail: {samuel.vaiter,gabriel.peyre}@ceremade.dauphine.fr

Charles Deledalle, Charles Dossal
IMB, CNRS, Université Bordeaux 1, 351, Cours de la libération, 33405 Talence Cedex, France
E-mail: {charles.deledalle,charles.dossal}@math.u-bordeaux1.fr

Jalal Fadili
GREYC, CNRS-ENSICAEN-Université de Caen, 6, Bd du Maréchal Juin, 14050 Caen Cedex, France
E-mail: Jalal.Fadili@greyc.ensicaen.fr

1 Introduction

1.1 Regression and Regularization

We consider a model

$$\mathbb{E}(Y|X) = h(X\beta_0), \tag{1}$$

where $Y = (Y_1, \ldots, Y_n)$ is the response vector, $\beta_0 \in \mathbb{R}^p$ is the unknown vector of linear regression coefficients, $X \in \mathbb{R}^{n\times p}$ is the fixed design matrix whose columns are the p covariate vectors, and the expectation is taken with respect to some σ-finite measure. h is a known real-valued and smooth function $\mathbb{R}^n \to \mathbb{R}^n$. The goal is to design an estimator of β_0 and to study its properties. In the sequel, we do not make any specific assumption on the number of observations n with respect to the number of predictors p. Recall that when $n < p$, (1) is underdetermined, whereas when $n \geqslant p$ and all the columns of X are linearly independent, it is overdetermined.

Many examples fall within the scope of model (1). We here review two of them.

Example 1 (GLM) One naturally thinks of generalized linear models (GLMs) (McCullagh and Nelder 1989) which assume that conditionally on X, Y_i are independent with distribution that belongs to a given (one-parameter) standard exponential family. Recall that the random variable $Z \in \mathbb{R}$ has a distribution in this family if its distribution admits a density with respect to some reference σ-finite measure on $\mathbb{R}$ of the form

$$p(z;\theta) = B(z)\exp(z\theta - \varphi(\theta)), \quad \theta \in \Theta \subseteq \mathbb{R}\ ,$$

where Θ is the natural parameter space and θ is the canonical parameter. For model (1), the distribution of Y belongs to the n-parameter exponential family and its density reads

$$f(y|X;\beta_0) = \left(\prod_{i=1}^n B_i(y_i)\right)\exp\left(\langle y,\, X\beta_0\rangle - \sum_{i=1}^n \varphi_i\left((X\beta_0)_i\right)\right), \quad X\beta_0 \in \Theta^n\ , \tag{2}$$

where $\langle\cdot,\,\cdot\rangle$ is the inner product, and the canonical parameter vector is the linear predictor $X\beta_0$. In this case, $h(\mu) = (h_i(\mu_i))_{1\leqslant i\leqslant n}$, where h_i is the *inverse* of the so-called link function in the language of GLM. Each h_i is a monotonic differentiable function, and a typical choice is the canonical link $h_i = \varphi_i'$, where φ_i' is known to be one-to-one if the family is regular (Brown 1986). Well-known examples are the identity link $h_i(t) = t$ (Gaussian distribution, linear model), the reciprocal link $h_i(t) = -1/t$ (Gamma and exponential distributions), and the logit link $h_i(t) = \frac{1}{1+\exp(-t)}$ (Bernoulli distribution, logistic regression).

Example 2 (Transformations) The second example is where h plays the role of a transformation such as variance-stabilizing transformations (VSTs), symmetrizing transformations, or bias-corrected transformations. There is an

enormous body of literature on transformations, going back to the early 1940s. A typical example is when Y_i are indepedent Poisson random variables $\sim \mathcal{P}\left((X\beta_0)_i\right)$, in which case h_i takes the form of the Anscombe bias-corrected VST. See (DasGupta 2008, Chapter 4) for a comprehensive treatment and more examples.

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 to be estimated. This regularization ranges from squared Euclidean or Hilbertian norms (Tikhonov and Arsenin 1997), to non-Hilbertian norms that have sparked considerable interest in the recent years. Of particular interest are sparsity-inducing penalties, such as the ℓ^1 norm, which has been intensively investigated in the recent years, e.g. (Chen et al 1999; Tibshirani 1996; Osborne et al 2000; Donoho 2006; Candès and Plan 2009; Bickel et al 2009); see (Bühlmann and van de Geer 2011) for a comprehensive review. When the covariates are assumed to be clustered in a few active groups/blocks, the group Lasso has been advocated since it promotes sparsity of the groups, i.e. it drives all the coefficients in one group to zero together hence leading to group selection, see (Bakin 1999; Yuan and Lin 2006; Bach 2008; Wei and Huang 2010) to cite a few. Another popular regularization is the total variation seminorm, introduced in the ROF model (Rudin et al 1992), and the fused Lasso penalty (Tibshirani et al 2005).

1.2 Variational Estimators

Given observations $(y_1, \ldots, y_n)$, we consider the class of estimators obtained as the solution of the convex optimization problem

$$\widehat{\beta}(y) \in \underset{\beta \in \mathbb{R}^p}{\mathrm{Argmin}}\; F(\beta, y) + J(\beta)\ . \qquad (\mathcal{P}(y))$$

The fidelity term F is of the following form

$$F(\beta, y) = F_0(X\beta, y) \qquad (3)$$

where $F_0(\cdot, y)$ is a general loss function assumed to be a proper, convex and sufficiently smooth function of its first argument ; see Section 2 for a detailed exposition of the smoothness assumptions. The regularizing penalty J is proper, continuous and convex, and promotes some specific features on the solution $\widehat{\beta}(y)$; see Section 2 for a precise description of the class of regularizing penalties J that we consider in this paper. The type of convex optimization problem in $(\mathcal{P}(y))$ is referred to as a regularized M-estimator in Negahban et al (2012), where J is morever assumed to have a special decomposability property.

We now provide some illustrative examples of loss functions F and regularizing penalty J routinely used in signal processing, imaging sciences and statistical machine learning.

Example 3 (Generalized linear models) Generalized linear models in the exponential family falls into the class of losses we consider. Indeed, taking the negative log-likelihood corresponding to (2) gives[1]

$$F_0(\mu, y) = \sum_{i=1}^{n} \varphi_i(\mu_i) - \langle y, \mu \rangle \ . \tag{4}$$

It is well-known that if the exponential family is regular, then φ_i is proper, infinitely differentiable, its hessian is definite positive, and thus it is strictly convex (Brown 1986). Therefore, $F_0(\cdot, y)$ shares exactly the same properties. We recover the squared loss $F_0(\mu, y) = \frac{1}{2}\|y-\mu\|^2$ for the standard linear models (Gaussian case), and the logistic loss $F_0(\mu, y) = \sum_{i=1}^{n} \log(1 + \exp(\mu_i)) - \langle y, \mu \rangle$ for logistic regression (Bernoulli case).

GLM estimators with losses (4) and ℓ^1 or $\ell^1 - \ell^2$ (group) penalties have been previously considered and some of their properties studied including in (Bunea 2008; van de Geer 2008; de Geer 2008; Meier et al 2008; Bach 2010; Kakade et al 2010); see also (Bühlmann and van de Geer 2011, Chapter 3, 4 and 6).

Example 4 (Lasso) The Lasso regularization is used to promote the sparsity of the solution and corresponds to choosing J as the ℓ^1-norm

$$J(\beta) = \|\beta\|_1 = \sum_{i=1}^{p} |\beta_i|. \tag{5}$$

It is also refered to as ℓ^1-synthesis in the signal processing community, in contrast to the more general ℓ^1-analysis sparsity penalty detailed below.

Example 5 (General Lasso) To allow for general regularization penalties, it may be desirable to promote sparsity through a linear operator $D = (d_1, \ldots, d_q) \in \mathbb{R}^{p \times q}$. This leads to the so-called analysis-type sparsity penalty (a.k.a. general Lasso after Tibshirani and Taylor (2012)) where the ℓ^1-norm is pre-composed by D^*, hence giving

$$J(\beta) = \|D^*\beta\|_1 = \sum_{j=1}^{q} |\langle d_j, \beta \rangle|. \tag{6}$$

This of course reduces to the usual lasso penalty (5) when $D = \mathrm{Id}_p$. The penalty (6) encapsulates several important penalties including that of the 1-D total variation (Rudin et al 1992), and the fused Lasso (Tibshirani et al 2005). In the former, D^* is a finite difference approximation of the derivative, and in the latter, D^* is the concatenation of the identity matrix Id_p and the finite difference matrix to promote both the sparsity of the vector and that of its variations.

[1] Strictly speaking, the minimization may have to be over a convex subset of $\mathbb{R}^p$.

Example 6 (Group Lasso) In some applications, the sparsity of the parameter vector β is structured into groups. One then often wants to promote sparsity at the group level. This is achieved with the group Lasso penalty (Bakin 1999; Yuan and Lin 2006)

$$J(\beta) = \|\beta\|_{1,2} = \sum_{b\in\mathcal{B}} \|\beta_b\|_2. \tag{7}$$

where $\beta_b = (\beta_i)_{i\in b}$ is the sub-vector of β whose entries are indexed by the block $b \in \mathcal{B}$ where $\mathcal{B}$ is a disjoint union of the set of indices i.e. $\bigcup_{b\in\mathcal{B}} = \{1,\ldots,p\}$ such that $b, b' \in \mathcal{B}, b\cap b' = \emptyset$. (7) is a mixed $\ell^1-\ell^2$ norm which has the attractive property to be invariant under (groupwise) orthogonal transformations.

Example 7 (Analysis Group Lasso) One can push the structured sparsity idea one step further by promoting group/block sparsity through a linear operator, i.e. analysis-type group sparsity. Given a collection of linear operators $\{D_b\}_{b\in\mathcal{B}}$, that are not all orthogonal, the analysis group sparsity penalty is

$$J(\beta) = \|D^*\beta\|_{1,2} = \sum_{b\in\mathcal{B}} \|D_b^*\beta\|_2. \tag{8}$$

This encompasses the 2-D isotropic total variation (Rudin et al 1992), where β is a 2-D discretized image, and each $D_b^*\beta \in \mathbb{R}^2$ is a finite difference approximation of the gradient of β at a pixel indexed by b. This point of view allows also to extend the original group penalty (7) to the case where the blocks $b \in \mathcal{B}$ overlap, by taking $D_b^* : \beta \mapsto \beta_b$ to be a block extractor operator (Peyré et al 2011; Chen et al 2010).

1.3 Sensitivity Analysis

A chief goal of this paper is to investigate the sensitivity of any solution $\widehat{\beta}(y)$ to the parameterized problem $(\mathcal{P}(y))$ to (small) perturbations of y. Sensitivity analysis[2] is a major branch of optimization and optimal control theory. Comprehensive monographs on the subject are (Bonnans and Shapiro 2000; Mordukhovich 1992). The focus of sensitivity analysis is the dependence and the regularity properties of the optimal solution set and the optimal values when the auxiliary parameters (e.g. y here) undergo a perturbation. In its simplest form, sensitivity analysis of first-order optimality conditions, in the parametric form of the Fermat rule, relies on the celebrated implicit function theorem.

The set of priors J we consider (coined piecewise regular or partly smooth regularizers relative to a linear manifold, as detailed in Section 2) can be seen as a special case of the broader class of "partly smooth" functions (Lewis 2003). The latter unifies many non-smooth functions known in the literature. The

[2] The meaning of sensitivity is different here from what is usually intended in satistical sensitivity and uncertainty analysis.

notion of partial smoothness (as well as identifiable surfaces (Wright 1993)) captures essential features of the geometry of non-smoothness which are along the so-called "active/identifiable manifold". For convex functions, a closely related idea was developed in (Lemaréchal et al 2000). Loosely speaking, a partly smooth function behaves smoothly as we move on the identifiable manifold, and sharply if we move normal to the manifold. In fact, the behaviour of the function and of its minimizers (or critical points) depend essentially on its restriction to this manifold, hence offering a powerful framework for sensitivity analysis theory. In particular, critical points of partly smooth functions move stably on the manifold as the function undergoes small perturbations (Lewis 2003; Lewis and Zhang 2013).

Getting back to our class of regularizers, it turns out that our active/identifiable manifold is actually a linear subspace. Indeed, the core of our proof strategy relies on the identification of a certain linear subspace (that we coin model subspace), denoted $T = T_{\widehat{\beta}(y)}$ associated to a particular minimizer $\widehat{\beta}(y)$ of $(\mathcal{P}(y))$. We exhibit explicitly a certain set of observations, denoted $\mathcal{H}$ (see Definition 3), outside which the initial non-smooth optimization $(\mathcal{P}(y))$ boils down locally to a smooth optimization constrained by T. This part of the proof strategy is in close agreement with the one developed in (Lewis 2003) for the sensitivity analysis of partly smooth functions. See also (Bolte et al 2011, Theorem 13) for the case of linear optimization over a convex semialgebraic partly smooth feasible set, where the authors proves a sensitivity result with a zero-measure transition space. However, for our special class of regularizers, we were able to go beyond by solving additional key challenges that are important in a statistical context, namely: (i) we provide an analytical description of the set $\mathcal{H}$; (ii) we prove that this set is of zero Lebesgue measure; (iii) we compute the first-order expansion of $\widehat{\beta}(y)$ and provide an analytical form of the weak derivative of $X\widehat{\beta}(y)$. Altogether, this allows to get an unbiased estimator of the risk on the prediction $X\widehat{\beta}(Y)$.

1.4 Degrees of Freedom and Unbiased Risk Estimation

The degrees of freedom (DOF) of an estimator quantifies the complexity of a statistical modeling procedure (Efron 1986). It is at the heart of several risk estimation procedures and thus allows one to perform parameter selection through risk minimization.

In this section, we will assume that F_0 in (3) is strictly convex, so that the response (or the prediction) $\widehat{\mu}(y) = X\widehat{\beta}(y)$ is uniquely defined as a single-valued mapping of y (see Lemma 1). That is, it does not depend on a particular choice of solution $\widehat{\beta}(y)$ of $(\mathcal{P}(y))$. Let $\mu_0 = X\beta_0$.

Suppose that h in (1) is the identity and that the observations $Y \sim \mathcal{N}(\mu_0, \sigma^2 \mathrm{Id}_n)$. Following (Efron 1986), the DOF is defined as

$$df = \sum_{i=1}^{n} \frac{\mathrm{cov}(Y_i, \widehat{\mu}_i(Y))}{\sigma^2} \ .$$

The well-known Stein's lemma (Stein 1981) asserts that, if $y \mapsto \widehat{\mu}(y)$ is weakly differentiable function (i.e. typically in a Sobolev space over an open subset of $\mathbb{R}^n$), such that each coordinate $y \mapsto \widehat{\mu}_i(y) \in \mathbb{R}$ has an essentially bounded weak derivative[3]

$$\mathbb{E}\left(\left|\frac{\partial \widehat{\mu}_i}{\partial y_i}(Y)\right|\right) < \infty, \quad \forall i \ ,$$

then its divergence is an unbiased estimator of its DOF, i.e.

$$\widehat{df} = \mathrm{div}(\widehat{\mu})(Y) = \mathrm{tr}(\mathrm{D}\widehat{\mu}(Y)) \quad \text{and} \quad \mathbb{E}(\widehat{df}) = df \ ,$$

where $\mathrm{D}\widehat{\mu}$ is the Jacobian of $y \mapsto \widehat{\mu}(y)$. In turn, this allows to get an unbiased estimator of the prediction risk $\mathbb{E}(\|\widehat{\mu}(Y) - \mu_0\|^2)$ through the SURE (Stein Unbiased Risk Estimate, Stein 1981).

Extensions of the SURE to independent variables from an exponential family are considered in (Hudson 1978) for the continuous case, and (Hwang 1982) in the discrete case. Eldar (2009) generalizes the SURE principle to continuous multivariate exponential families.

1.5 Contributions

We consider a large class of losses F_0, and of regularizing penalties J which are finite-valued convex and partly smooth functions relative to linear manifold, following the definition of (Vaiter et al 2013). We recall it in Section 2. For this class of regularizers and losses, we first establish in Theorem 1 a general sensivity analysis result, which provides the local parametrization of any solution to $(\mathcal{P}(y))$ as a function of the observation vector y. This is achieved without placing any specific assumption on X, should it be full column rank or not. With such a result at hand, we derive an expression of the divergence of the prediction with respect to the observations (Theorem 2). Using tools from o-minimal geometry, we prove that this divergence formula is valid Lebesgue-a.e.. In turn, this allows us to get an unbiased estimate of the DOF and of the prediction risk (Theorem 3 and Theorem 4) for model (1) under two scenarios: (i) Lipschitz continuous non-linearity h and an additive i.i.d. Gaussian noise; (ii) GLMs with a continuous exponential family. Our results encompass some previous ones in the literature as special cases (see discussion in the next section).

1.6 Relation to prior works

In the case of standard Lasso (i.e. ℓ^1 penalty (5)) with $Y \sim \mathcal{N}(X\beta_0, \sigma^2 \mathrm{Id}_n)$ and X of full column rank, (Zou et al 2007) showed that the number of nonzero coefficients is an unbiased estimate for the DOF. Their work was generalized

[3] We write the same symbol as for the derivative, and rigorously speaking, this has to be understood to hold Lebesgue-a.e.

in (Dossal et al 2013) to any arbitrary design matrix. Under the same Gaussian linear regression model, unbiased estimators of the DOF for the Lasso with ℓ^1-analysis penalty (6), were given independently in (Tibshirani and Taylor 2012; Vaiter et al 2012a).

A formula of an estimate of the DOF for the group Lasso when the design is orthogonal within each group was conjectured in (Yuan and Lin 2006). Kato (2009) studied the DOF of a general shrinkage estimator where the regression coefficients are constrained to a closed convex set C. His work extends that of (Meyer and Woodroofe 2000) which treats the case where C is a convex polyhedral cone. When X is full column rank, (Kato 2009) derived a divergence formula under a smoothness condition on the boundary of C, from which an unbiased estimator of the degrees of freedom was obtained. When specializing to the constrained version of the group Lasso, the author provided an unbiased estimate of the corresponding DOF under the same group-wise orthogonality assumption on X as (Yuan and Lin 2006). Hansen and Sokol (2014) studied the DOF of the metric projection onto a closed set (non-necessarily convex), and gave a precise representation of the bias when the projector is not sufficiently differentiable. An estimate of the DOF for the group Lasso was also given by (Solo and Ulfarsson 2010) using heuristic derivations that are valid only when X is full column rank, though its unbiasedness is not proved.

Vaiter et al (2012b) also derived an estimator of the DOF of the group Lasso and proved its unbiasedness when X is full column rank, but without the orthogonality assumption required in (Yuan and Lin 2006; Kato 2009). When specialized to the group Lasso penalty, our results establish that the DOF estimator formula in (Vaiter et al 2012b) is still valid while removing the full column rank assumption. This of course allows one to tackle the more challenging rank-deficient or underdetermined case $p > n$.

1.7 Notations

In the following, for any subspace $T \subset \mathbb{R}^p$, we denote P_T the orthogonal projection on T and

$$\beta_T = \mathrm{P}_T(\beta) \quad \text{and} \quad X_T = X\,\mathrm{P}_T\,.$$

For any matrix A, A^* denotes its transpose.

For a subspace $T \subset \mathbb{R}^p$, and any function $g \in \mathrm{C}^2(T \times \mathbb{R}^n)$, we denote

$$\mathrm{D}_1^2 g_T(\beta, y) = \mathrm{P}_T \circ \mathrm{D}_1^2 g(\beta, y) \circ \mathrm{P}_T$$

which can be understood as the Hessian of the mapping $\beta \in T \mapsto g(\beta, y)$, i.e. the restriction of $g(\cdot, y)$ to T. Of course, when T is the whole space, we recover the "full" Hessian.

We also denote $\mathrm{D}_{12}^2 g(\beta, y)$ the Jacobian of the mapping $y \in \mathbb{R}^n \mapsto \nabla_1 g(\beta, y)$ with respect to y, and $\nabla_1 g(\beta, y)$ is the gradient of g w.r.t the first variable at (β, y).

We now turn to the notion of *model space*. The interested reader may refer to (Vaiter et al 2013) and references therein for a comprehensive treatment. Consider a convex continuous, hence of full domain, function J. We denote

$$\bar{S}_\beta = \mathrm{aff}(\partial J(\beta))$$

the affine hull of the sub-differential at β (i.e. the smallest affine manifold containing it), and

$$e\,(\beta) = \underset{e \in \bar{S}_\beta}{\mathrm{argmin}}\ \|e\|,$$

i.e. $e\,(\beta)$ is the orthogonal projection of the origin on $\bar{S}_\beta$. We denote the subspaces

$$S_\beta = \bar{S}_\beta - e\,(\beta) \quad \text{and} \quad T_\beta = S_\beta^\perp. \tag{9}$$

We denote the set of all possible subspaces T_β as

$$\mathcal{T} = \{T_\beta \subset \mathbb{R}^p \ : \ \beta \in \mathbb{R}^p\}\,.$$

For any $T \in \mathcal{T}$, we denote $\widetilde{T}$ the set of vectors sharing the same subspace T,

$$\widetilde{T} = \{\beta \in \mathbb{R}^p \ : \ T_\beta = T\}\,.$$

For instance, when $J = \|\cdot\|_1$, $\widetilde{T}$ is the cone of all the vectors sharing the same support.

We give in the following examples of the model subspace associated to some convex partly smooth regularizers that are popular in the literature.

Example 8 (Lasso regularizer) We denote $(a_i)_{1 \leqslant i \leqslant p}$ the canonical basis of $\mathbb{R}^p$. In the case $J(\beta) = \|\beta\|_1$, the model space of a vector $\beta \in \mathbb{R}^p$ is given by

$$T_\beta = \mathrm{Span}\{(a_i)_{i \in \mathrm{supp}(\beta)}\} \quad \text{where} \quad \mathrm{supp}(\beta) = \{i \in \{1, \ldots, p\} \ : \ \beta_i \neq 0\}\,.$$

Example 9 (General Lasso regularizer) Proposition 9 in (Vaiter et al 2013) relates the model subspace and vector associated to a convex partly smooth regularizer $J \circ D^*$, where D is a linear operator, to those of J. Let us illustrate this in the case where $J = \|\cdot\|_1$. For $J(\beta) = \|D^*\beta\|_1$, one has

$$T_\beta = \mathrm{Ker}\, D^*_\Lambda \quad \text{where} \quad \Lambda = \mathrm{supp}(D^*\beta)^c.$$

Example 10 (Group Lasso regularizer) The model space associated to β when the blocks are of size greater than 1 can be defined similarly, but using the notion of block support. Using the block structure $\mathcal{B}$, one has

$$T_\beta = \mathrm{Span}\{(a_i)_{i \in \mathrm{supp}_{\mathcal{B}}\,\beta)}\},$$

where

$$\mathrm{supp}_{\mathcal{B}}(\beta) = \{i \in \{1, \ldots, p\} \ : \ \exists b \in \mathcal{B},\, \beta_b \neq 0 \quad \text{and} \quad i \in b\}\,.$$

2 Partly Smooth Functions with Linear Manifold

2.1 Partial Smoothness

Toward the goal of studying the sensitivity behaviour of $\widehat{\beta}(y)$ and $\widehat{\mu}(y)$ with non-negative finite-valued convex regularizers J, we restrict our attention to a subclass of these functions that fulfill some regularity assumptions according to the following definition.

Definition 1 *A finite-valued convex function J is said to be* partly smooth *at β relative to a linear manifold $\mathcal{M} \subseteq \mathbb{R}^p$ if*

1. *Smoothness:*

$$J \text{ restricted to } \mathcal{M} \text{ is } \mathrm{C}^2 \text{ around } \beta. \qquad (C_{\mathrm{sm}})$$

2. *Sharpness: $\mathcal{M} = T_\beta$.*
3. *Continuity:*

$$\text{The set-valued mapping } \partial J \text{ is continuous at } \beta \text{ relative to } \mathcal{M}. \qquad (C_{\mathrm{cont}})$$

J is said to be partly smooth relative to the linear manifold $\mathcal{M} \in \mathcal{T}$ *if J is partly smooth at each point $\beta \in \mathcal{M}$ relative to $\mathcal{M}$.*

It turns out that the sharpness property is locally stable (Lewis 2003, Proposition 2.10), meaning that if it holds at β implies that it also holds at all nearby points in $\mathcal{M}$. This can be formally written as

$$\exists \varepsilon > 0, \forall\, \beta' \in T_\beta \cap \mathbb{B}(\beta, \varepsilon) \Rightarrow T_\beta = T_{\beta'}\ . \qquad (C_{\mathrm{sharp}})$$

We will also assume in the following that

$$\text{The set } \mathcal{T} \text{ is finite.} \qquad (C_{\mathcal{T}})$$

Some remarks are in order. Assumption (C_{sharp}) amounts to saying that there exists a neighbourhood of β on T_β on which this subspace model is constant. The above class of partly smooth functions is closed under addition and pre-composition by a linear operator, see see (Vaiter et al 2013). Many well-studied regilarizing penalties are partly smooth relative to a linear manifold, including the ℓ^1, the $\ell^1 - \ell^2$, the ℓ^∞ norms, and their analysis-type versions and/or positive combinations, see (Vaiter et al 2013, Section 6) for a detailed discussion of the examples.

Assumption $(C_{\mathcal{T}})$ holds in many important cases, including the Lasso (ℓ^1-norm) and group Lasso ($\ell^1 - \ell^2$) penalties, the ℓ^∞-norm, as well as their analysis-type counterparts (composition with linear operators). However, our definition precludes the case of the nuclear norm (also known as the trace norm). Indeed, in this case, the function is still partly smooth bu relatively to non-linear smooth manifold composed of matrices with fixed rank. We refer to (Vaiter et al 2014) for a detailed discussion of the model manifold properties of this regularizer. Our results thus do not cover the latter.

2.2 Restriction and Second-Order Derivative of the Regularizer

We denote

$$J_T : \beta_T \in T \mapsto J(\beta_T) \in \mathbb{R}^+$$

the restriction of J to T for some subspace $T \subset \mathcal{T}$. Hence the hessian of J_T is well-defined on $\mathcal{T}$. We illustrate this definition on several examples.

Example 11 (Lasso and general Lasso) For $J = \|\cdot\|_1$, one has

$$\forall\, \beta_T \in T, \quad \nabla J_T(\beta_T) = \mathrm{sign}(\beta_T)\ ,$$

and thus, $\mathrm{D}^2 J_T(\beta_T) = 0$. This is also the case for the analysis ℓ^1-penalty (general Lasso), see for instance (Vaiter et al 2012a). This property basically reflects the fact that these regularizers are polyhedral, hence piecewise affine.

Example 12 (Group Lasso) For $J = \|\cdot\|_{1,2}$ as defined in (8), we have

$$\mathrm{D}^2 J_T(\beta_T) = \delta_\beta \circ \mathrm{P}_{\beta^\perp},$$

where, for $I = \mathrm{supp}_{\mathcal{B}}(\beta)$,

$$\begin{aligned} &\delta_\beta : v \in \mathbb{R}^{|I|} \mapsto (v_b/\|\beta_b\|)_{b\subset I} \in \mathbb{R}^{|I|} \\ \text{and} \quad &\mathrm{P}_{\beta^\perp} : v \in \mathbb{R}^{|I|} \mapsto (\mathrm{P}_{\beta_b^\perp}\, v_b)_{b\subset I} \in \mathbb{R}^{|I|}\ , \end{aligned}$$

where

$$\mathrm{P}_{\beta_b^\perp}\, v_b = v_b - \frac{\langle \beta_b,\, v_b\rangle}{\|\beta_b\|^2}\beta_b$$

is the orthogonal projector on $\beta_b^\perp$.

3 Sensitivity Analysis of $\widehat{\beta}(y)$

In all the following, we consider a variational regularized problem of the form of $(\mathcal{P}(y))$. We assume that the fidelity term enjoys the following properties.

$$\forall\, (y,\beta) \in \mathbb{R}^n \times \mathbb{R}^p, \quad F(\cdot, y) \in \mathrm{C}^2(\mathbb{R}^p) \quad \text{and} \quad \nabla_1 F(\beta, \cdot) \in \mathrm{C}^1(\mathbb{R}^n). \qquad (C_F)$$

In this section, we aim at computing the derivative of the map $y \mapsto \widehat{\beta}(y)$ whenever this is possible. The following condition plays a pivotal role in this analysis.

Definition 2 (Restricted Injectivity) *A vector $\beta \in \mathbb{R}^p$ with $T = T_\beta$ is said to satisfy the* restricted injectivity condition *if, and only if,*

$$T \cap \ker(\mathrm{D}_1^2 F_T(\beta, y)) \cap \ker(\mathrm{D}^2 J_T(\beta)) = \{0\}. \qquad (\mathcal{C}_{\beta,y})$$

Example 13 (Lasso) For the Lasso problem, i.e. $J = \|\cdot\|_1$ and F_0 is the squared loss, condition $(\mathcal{C}_{\beta,y})$ reads $\ker(X_I) = \{0\}$, where I is the support of the vector β. This condition is already known in the literature, see for instance (Dossal et al 2013) in the context of DOF estimation.

Example 14 (Group Lasso) For the group Lasso, i.e. $J = \|\cdot\|_{1,2}$ and F_0 is the squared loss, condition $(\mathcal{C}_{\beta,y})$ amounts to assuming that the collection of vectors $(X_b\beta_b)_{b\subset I}$ is linearly independent, where $I = \mathrm{supp}_{\mathcal{B}}(\beta)$. This condition appears in (Liu and Zhang 2009) to establish ℓ^2-consistency of the group Lasso. It goes without saying that condition $(\mathcal{C}_{\beta,y})$ is much weaker than imposing that X_I is full column rank, which is standard when analyzing the Lasso.

Let us now turn to the sensitivity of a minimizer $\widehat{\beta}(y)$ of $(\mathcal{P}(y))$ to perturbations of y. Because of non-smoothness of the regularizer J, it is a well-known fact in sensitivity analysis that one cannot hope for a global claim, i.e. an everywhere smooth mapping[4] $y \mapsto \widehat{\beta}(y)$. Rather, the sensitivity behaviour is local. This is why the reason we need to introduce the following transition space $\mathcal{H}$, which will be shown to contain points of non-smoothness of $\widehat{\beta}(y)$.

Definition 3 *The* transition space $\mathcal{H}$ *is defined as*

$$\mathcal{H} = \bigcup_{T\in\mathcal{T}} \mathcal{H}_T, \quad \textit{where} \quad \mathcal{H}_T = \mathrm{bd}(\Pi_{n+p,n}(\mathcal{A}_T)),$$

where we have denoted

$$\Pi_{n+p,n} : \begin{cases} \mathbb{R}^n \times \widetilde{T} \longrightarrow \mathbb{R}^n \\ (y,\beta_T) \longmapsto y \end{cases}$$

the canonical projection on the first n coordinates, $\mathrm{bd}\, C$ *is the boundary of the set C, and*

$$\mathcal{A}_T = \left\{ (y,\beta_T) \in \mathbb{R}^n \times \widetilde{T} \ : \ -\nabla_1 F(\beta_T, y) \in \mathrm{rbd}\, \partial J(\beta_T) \right\}.$$

Here, $\mathrm{rbd}\, \partial J(\beta_T)$ *is the relative boundary of* $\partial J(\beta_T)$ *defined as its boundary in the topology of its affine hull.*

In the particular case of the Lasso (resp. general Lasso), i.e. F_0 is the squared loss, $J = \|\cdot\|_1$ (resp. $J = \|D^*\cdot\|_1$), the transition space specializes to the one introduced in (Dossal et al 2013) (resp. (Vaiter et al 2012a)). In these specific cases, since J is a polyhedral gauge, $\mathcal{H}$ is in fact a union of affine hyperplanes. The geometry of this set can be significantly more complex for other regularizers. For instance, for $J = \|\cdot\|_{1,2}$, it can be shown to be a semi-algebraic set (union of algebraic hyper-surfaces). Section 5 is devoted to a detailed analysis of this set $\mathcal{H}$.

We are now equipped to state our main sensitivity analysis result, whose proof is deferred to Section 6.2.

[4] To be understood here as a set-valued mapping.

Theorem 1 *Let $y \notin \mathcal{H}$, and $\beta^\star$ a solution of $(\mathcal{P}(y))$ such that $(\mathcal{C}_{\beta^\star,y})$ holds. Then, there exists an open neighborhood $\mathcal{V} \subset \mathbb{R}^n$ of y, and a mapping $\tilde{\beta} : \mathcal{V} \to T$ such that*

1. *For all $\bar{y} \in \mathcal{V}$, $\tilde{\beta}(\bar{y})$ is a solution of $(\mathcal{P}(\bar{y}))$, and $\tilde{\beta}(y) = \beta^\star$.*
2. *the mapping $\tilde{\beta}$ is $\mathrm{C}^1(\mathcal{V})$ and*

$$\forall\, \bar{y} \in \mathcal{V}, \quad \mathrm{D}\tilde{\beta}(\bar{y}) = -(\mathrm{D}_1^2 F_T(\beta^\star, \bar{y}) + \mathrm{D}^2 J_T(\beta^\star))^{-1} \circ \mathrm{P}_T \circ \mathrm{D}_{12}^2 F(\beta^\star, \bar{y}), \tag{10}$$

where $T = T_{\beta^\star}$.

Theorem 1 can be extended to the case where the data fidelity is of the form $F(\beta, \theta)$ for some parameter θ, with no particular role of y here. One now may wonder whether condition $(\mathcal{C}_{\beta^\star,y})$ is restrictive, and in particular, whether there exists always a solution $\beta^\star$ such that it holds. In the following section, we give an affirmative answer with the proviso that the loss F_0 is strictly convex.

4 Sensitivity Analysis of $\widehat{\mu}(y)$

We assume in this section that F takes the form (3) with

$$\forall\, (\mu, y) \in \mathbb{R}^n \times \mathbb{R}^n, \quad \mathrm{D}_1^2 F_0(\mu, y) \text{ is definite positive.} \tag{C_{dp}}$$

This in turn implies that $F_0(\cdot, y)$ is strictly convex for any y (the converse is obviously not true). Recall that this condition is mild and holds in many situations, in particular for losses (4) in the exponential family, see Section 1.2 for details.

Under this condition, the following immediate lemma (proved in Section 6.3) gives a convenient re-writing of condition $(\mathcal{C}_{\beta^\star,y})$.

Lemma 1 *Assume that condition (C_{dp}) holds. For $\beta \in \mathbb{R}^p$, and $T = T_\beta$, the two following are equivalent.*

(i) $(\mathcal{C}_{\beta,y})$ *holds.*
(ii) $\ker(X_T) \cap \ker(\mathrm{D}^2 J_T(\beta)) = \{0\}$.

Furthermore, all minimizers of $(\mathcal{P}(y))$ share the same image under X.

Owing to this lemma, we can now define the prediction

$$\widehat{\mu}(y) = X\widehat{\beta}(y) \tag{11}$$

without ambiguity given any solution $\widehat{\beta}(y)$, which in turn defines a single-valued mapping $\widehat{\mu}$. The following theorem provides a closed-form expression of the local variations of $\widehat{\mu}$ as a function of perturbations of y.

Theorem 2 *Under assumption* (C_{dp})*, the mapping* $y \mapsto \widehat{\mu}(y)$ *is* $\mathrm{C}^1(\mathbb{R}^n \setminus \mathcal{H})$*. For all* $y \notin \mathcal{H}$*, there exists a solution* $\beta^\star$ *of* $(\mathcal{P}(y))$ *such that* $(\mathcal{C}_{\beta^\star,y})$ *is satisfied. Moreover, for all* $y \notin \mathcal{H}$*,*

$$\mathrm{D}\widehat{\mu}(y) = \Delta(y) \tag{12}$$

where

$$\Delta(y) = -X_T \circ (X_T^* \circ \mathrm{D}_1^2 F_0(X\beta^\star, y) \circ X_T + \mathrm{D}^2 J_T(\beta^\star))^{-1} \circ X_T^* \circ \mathrm{D}_{12}^2 F_0(X\beta^\star, y)$$

where $\beta^\star$ *is any solution of* $(\mathcal{P}(y))$ *such that* $(\mathcal{C}_{\beta^\star,y})$ *holds and* $T = T_{\beta^\star}$*.*

This Theorem is proved in Section 6.3.

Example 15 (Lasso) For the Lasso problem, the above divergence formula boils down to

$$\mathrm{div}(\widehat{\mu})(y) = |\operatorname{supp}(\beta^\star)|,$$

where $\beta^\star$ is a solution of $(\mathcal{P}(y))$ such that $(\mathcal{C}_{\beta^\star,y})$ holds. Indeed, the ℓ^1-norm is affine on the model subspace $T = T_{\beta^\star}$, and thus $\mathrm{D}^2 J_T(\beta^\star) = 0$, as already remarked in Section 2.2. This result was proved in (Dossal et al 2013), see also (Tibshirani and Taylor 2012).

Example 16 (General Lasso) The general Lasso case was investigated in (Vaiter et al 2012a) and (Tibshirani and Taylor 2012). In this case, using again the fact that $\mathrm{D}^2 J_T(\beta^\star) = 0$, one has

$$\mathrm{div}(\widehat{\mu})(y) = \dim \operatorname{Ker} D_\Lambda^*, \quad \Lambda = \operatorname{supp}(D^*\beta^\star)^c \ ,$$

where $\beta^\star$ is such that $(\mathcal{C}_{\beta^\star,y})$ holds.

5 Degrees of Freedom and Unbiased Risk Estimation

Throughout this section, we use the same symbols to denote weak derivatives (whenever they exist) as for derivatives. Rigorously speaking, the identities have to be understood to hold Lebesgue-a.e. (Evans and Gariepy 1992).

So far, we have shown that outside the transition space $\mathcal{H}$, the mapping $\widehat{\mu}(y)$ enjoys (locally) nice smoothness properties, which in turn gives closed-form formula of its divergence. To establish that such formal hold Lebesgue a.e., a key argument that we need to show is that $\mathcal{H}$ is of negligible Lebesgue measure. This is where o-minimal geometry enters the picture. In turn, for Y drawn from some appropriate probability measures with density with respect to the Lebesgue measure, this will allow us to establish unbiasedness of quadratic risk estimators.

5.1 O-minimal Geometry

Roughly speaking, to be able to control the size of $\mathcal{H}$, the function J cannot be too oscillating in order to prevent pathological behaviours. We now briefly recall here the definition. Some important properties of o-minimal structures that are relevant to our context together with their proofs are collected in Section A. The interested reader may refer to (van den Dries 1998; Coste 1999) for a comprehensive account and further details on o-minimal structures.

Definition 4 (Structure) *A* structure $\mathcal{O}$ *expanding* $\mathbb{R}$ *is a sequence* $(\mathcal{O}_n)_{n\in\mathbb{N}}$ *which satisfies the following axioms:*

1. *Each* $\mathcal{O}_n$ *is a Boolean algebra of subsets of* $\mathbb{R}^n$*, with* $\mathbb{R}^n \in \mathcal{O}_n$*.*
2. *Every semi-algebraic subset of* $\mathbb{R}^n$ *is in* $\mathcal{O}_n$*.*
3. *If* $A \in \mathcal{O}_n$ *and* $B \in \mathcal{O}_n$*, then* $A \times B \in \mathcal{O}_{n+n'}$*.*
4. *If* $A \in \mathcal{O}_{n+1}$*, then* $\Pi_{n+1,n}(A) \in \mathcal{O}_n$*, where* $\Pi_{n+1,n} : \mathbb{R}^{n+1} \to \mathbb{R}^n$ *is the projection on the first* n *components.*

The structure $\mathcal{O}$ *is said to be* o-minimal *if, moreover, it satisfies*

5. *(o-minimality) Sets in* $\mathcal{O}_1$ *are precisely the finite unions of intervals and points of* $\mathbb{R}$*.*

In the following, a set $A \in \mathcal{O}_n$ is said to be definable.

Definition 5 (Definable set and function) *Let* $\mathcal{O}$ *be an o-minimal structure. The elements of* $\mathcal{O}_n$ *are called the* definable subsets *of* $\mathbb{R}^n$*, i.e.* $\Omega \subset \mathbb{R}^n$ *is definable if* $\Omega \in \mathcal{O}_n$*. A map* $f : \Omega \to \mathbb{R}^p$ *is said to be definable if its graph* $\mathcal{G}(f) = \{(x,u) \in \Omega \times \mathbb{R}^p \,:\, u = f(x)\} \subseteq \mathbb{R}^n \times \mathbb{R}^p$ *is a definable subset of* $\mathbb{R}^n \times \mathbb{R}^p$ *(in which case* p *times applications of axiom 4 implies that* Ω *is definable).*

A fundamental class of o-minimal structures is the collection of semi-algebraic sets, in which case axiom 4 is actually a property known as the Tarski-Seidenberg theorem (Coste 2002). For example, in the special case where q is a rational number, $J = \|\cdot\|_q$ is semi-algebraic. When $q \in \mathbb{R}$ is not rational, $\|\cdot\|_q$ is not semi-algebraic, however, it can be shown to be definable in an o-minimal structure. To see this, we recall from (van den Dries and Miller 1996, Example 5 and Property 5.2) that there exists a (polynomially bounded) o-minimal structure that contains the family of functions $\{t > 0 \,:\, t^q, q \in \mathbb{R}\}$ and restricted analytic functions. Functions F_0 that correspond to the exponential family losses introduced in Example 3 are also definable.

Our o-minimality assumptions requires the existence of an o-minimal structure $\mathcal{O}$ such that

$$\text{the functionals } F \text{ and } J \text{ are definable in } \mathcal{O}. \qquad (C_{\mathcal{O}})$$

5.2 Degrees of Freedom and Unbiased Risk Estimation

We assume in this section that F takes the form (3) and that

$$\forall\, y \in \mathbb{R}^n, \quad F_0(\cdot, y) \text{ is strongly convex with modulus } \tau \tag{C_τ}$$

and

$$\exists L > 0, \quad \sup_{(\mu,y)\in\mathbb{R}^n\times\mathbb{R}^n} \|\mathrm{D}^2_{12} F_0(\mu, y)\| \leqslant L. \tag{C_L}$$

Obviously, assumption (C_τ) implies (C_{dp}), and thus the claims of the previous section remain true. Moreover, this assumption holds for the squared loss, but also for some losses of the exponential family (4), possibly adding a small quadratic term in β. As far as assumption (C_L) is concerned, it is easy to check that it is fulfilled with $L = 1$ for any loss of the exponential family (4), since $\mathrm{D}^2_{12} F_0(\mu, y) = \mathrm{Id}$.

Non-linear Gaussian regression. Assume that the observation model (1) specializes to $Y \sim \mathcal{N}(h(X\beta_0), \sigma^2 \mathrm{Id}_n)$, where h is Lipschitz continuous.

Theorem 3 *Suppose that conditions $(C_{\mathcal{O}})$, (C_τ) and (C_L) hold. Then,*

(i) $\mathcal{H}$ is of Lebesgue measure zero;
(ii) $h \circ \widehat{\mu}$ is Lipschitz continuous, hence weakly differentiable, with an essentially bounded gradient.
(iii) $\widehat{df} = \mathrm{tr}(\mathrm{D}h(\widehat{\mu}(Y))\Delta(Y))$ is an unbiased estimate of $df = \mathbb{E}(\mathrm{div}(h \circ \widehat{\mu}(Y)))$.
(iv) The SURE

$$\mathrm{SURE}(h \circ \widehat{\mu})(Y) = \|Y - h(\widehat{\mu}(Y))\|^2 + 2\sigma^2 \widehat{df} - n\sigma^2 \tag{13}$$

is an unbiased estimator of the risk $\mathbb{E}\left(\|h(\widehat{\mu}(Y)) - h(\mu_0)\|^2\right)$.

This theorem is proved in Section 6.4.

GLM with the continuous exponential family. Assume that the observation model (1) corresponds to the GLM with a distribution which belongs to a continuous standard exponential family as parameterized in (2). We denote

$$\nabla \log B(y) = \left(\frac{\partial \log B_i(y_i)}{\partial y_i}\right)_i.$$

Theorem 4 *Suppose that conditions $(C_{\mathcal{O}})$, (C_τ) and (C_L) hold. Then,*

(i) $\mathcal{H}$ is of Lebesgue measure zero;
(ii) $\widehat{\mu}$ is Lipschitz continuous, hence weakly differentiable, with an essentially bounded gradient.
(iii) $\widehat{df} = \mathrm{tr}(\Delta(Y))$ is an unbiased estimate of $df = \mathbb{E}(\mathrm{div}(\widehat{\mu}(Y)))$.
(iv) The GSURE

$$\mathrm{GSURE}(\widehat{\mu})(Y) = \|\nabla \log B(Y) - \widehat{\mu}(Y)\|^2 + 2\widehat{df} - (\|\nabla \log B(Y)\|^2 - \|\mu_0\|^2) \tag{14}$$

is an unbiased estimator of the risk $\mathbb{E}\left(\|\widehat{\mu}(Y) - \mu_0\|^2\right)$.

This theorem is proved in Section 6.4.

Though $\mathrm{GSURE}(\widehat{\mu})(Y)$ depends on μ_0, which is obviously unknown, it is only through an additive constant, which makes it suitable for parameter selection by risk minimization. Moreover, even if it is not stated here explicitly, Theorem 4 can be extended to unbiasedly estimate other measures of the risk, including the *projection* risk, or the *estimation* risk (in the full rank case) through the Generalized Stein Unbiased Risk Estimator as proposed in (Eldar 2009, Section IV), see also (Vaiter et al 2012a) in the Gaussian case.

6 Proofs

This section details the proofs of our results.

6.1 Preparatory Lemmata

We first collect some results that are used in the sequel, and whose proof can be found in Vaiter et al (2013).

Proposition 1 (Decomposability property) *Let J be a finite-valued convex function. Let $\beta \in \mathbb{R}^N \setminus \{0\}$. Then any subgradient $\eta \in \partial J(\beta)$ is such that*

$$\eta_{T_\beta} = e(\beta) \ .$$

Moreover, the affine hull $\bar{S}_\beta$ equivalently reads

$$\bar{S}_\beta = \left\{\eta \in \mathbb{R}^p \ : \ \eta_{T_\beta} = e(\beta)\right\} .$$

Separation theory in convex analysis and the subdifferential structure in Proposition 1 yield the following useful equivalent characterization of the relative interior of the subdifferential.

Lemma 2 *For the subdifferenial $\partial J(\beta)$, there holds*

$$\eta \in \operatorname{ri} \partial J(\beta) \iff \forall u \in S \setminus \{0\}, \exists \eta' \in \partial J(\beta) \text{ such that } \langle u, \eta' - \eta \rangle > 0 \ .$$

Proof First, recall that the directional derivative $J'(\beta, u)$ of J at β in the direction u is

$$J'(\beta, u) = \lim_{t \downarrow 0} \frac{J(\beta + tu) - J(\beta)}{t}.$$

Since J is proper, convex and continuous, the subdifferential $\partial J(\beta)$ is a non-empty compact convex set of $\mathbb{R}^p$ whose support function is $J'(\beta, \cdot)$ (Rockafellar 1996, Theorem 23.4), i.e.

$$J'(\beta, u) = \max_{\eta \in \partial J(\beta)} \langle \eta, u \rangle \ ,$$

and the maximum is attained at some η'. From the charcaterization of the relative interior of a non-empty closed convex set (Lemaréchal and Hiriart-Urruty 1996, Theorem V.2.2.3) or (Rockafellar 1996, Theorem 13.1), and sublinearity we deduce that

$$\eta \in \operatorname{ri} \partial J(\beta) \iff J'(\beta, u) > \langle u,\, \eta \rangle \quad \forall u \text{ such that } J'(\beta, u) + J'(-\beta, u) > 0 \ .$$

Using Proposition 1 shows that

$$J'(\beta, u) = \langle e(\beta),\, u \rangle + \max_{\eta \in \mathrm{P}_S(\partial J(\beta))} \langle \eta,\, u \rangle \ .$$

Sublinearity implies that (Lemaréchal and Hiriart-Urruty 1996, Corollary V.1.1.5)

$$J'(\beta, u) + J'(\beta, -u) \geqslant 0 \ .$$

Thus

$$J'(\beta, u) + J'(\beta, -u) = \max_{\eta \in \mathrm{P}_S(\partial J(\beta))} \langle \eta,\, u \rangle - \min_{\eta \in \mathrm{P}_S(\partial J(\beta))} \langle \eta,\, u \rangle \ ,$$

whence we obtain

$$J'(\beta, u) + J'(\beta, -u) > 0 \iff u \notin T \ .$$

Piecing everything together, we get

$$\begin{aligned}
\eta \in \operatorname{ri} \partial J(\beta) &\iff \forall u \notin T, \quad J'(\beta, u) > \langle u,\, \eta \rangle \\
&\iff \forall u \notin T, \exists \eta' \in \partial J(\beta) \text{ such that } \langle u,\, \eta' \rangle > \langle u,\, \eta \rangle \\
&\iff \forall u \notin T, \exists \eta' \in \partial J(\beta) \text{ such that } \langle u,\, \eta' - \eta \rangle > 0 \\
&\iff \forall u \notin T, \exists \eta' \in \partial J(\beta) \text{ such that } \langle u,\, \eta'_S - \eta_S \rangle > 0 \\
&\iff \forall u \in S \setminus \{0\}, \exists \eta' \in \partial J(\beta) \text{ such that } \langle u,\, \eta' - \eta \rangle > 0 \ .
\end{aligned}$$

□

By standard arguments of convex analysis and using again the subdifferential structure of Proposition 1, the following lemma gives the first-order sufficient and necessary optimality condition of a minimizer of $(\mathcal{P}(y))$.

Lemma 3 *A vector $\beta^\star \in \mathbb{R}^p$ is a minimizer of $(\mathcal{P}(y))$ if, and only if,*

$$-\nabla_1 F(\beta^\star, y) \in \partial J(\beta^\star).$$

In particular, if $\beta^\star \in \mathbb{R}^p$ is a minimizer of $(\mathcal{P}(y))$, then

$$-\nabla_1 F(\beta^\star, y)_T = e\,(\beta^\star)\,,$$

where we have denoted $T_{\beta^\star} = T$.

Proof The first monotone inclusion is just the first-oder necessary and sufficient minimality condition for our convex program. Using the structure of the subdifferential of Proposition 1 for J at $\beta^\star$, this is equivalent to

$$-\nabla_1 F(\beta^\star, y)_T = e\,(\beta^\star) \quad \text{and} \quad -\nabla_1 F(\beta^\star, y)_S \in (\partial J(x))_S.$$

□

Lemma 4 *Let $\beta \in \mathbb{R}^p$, and $T = T_\beta$. Assume that $(\mathcal{C}_{\beta,y})$ holds. Then the linear operator $\mathrm{D}_1^2 F_T(\beta, y) + \mathrm{D}^2 J_T(\beta) : T \to T$ is invertible on T.*

Proof Since $F(\cdot, y)$ and J are convex and in $\mathrm{C}^2(T)$ by assumptions (C_F) and (C_{sm}), the (restricted) hessians $\mathrm{D}_1^2 F_T(\beta, y)$ and $\mathrm{D}^2 J_T(\beta)$ are symmetric semidefinite positive on T. To ensure invertibility of their sum on T, it is necessary and sufficient that their kernels have a trivial intersection, which is exactly what assumption $\mathcal{C}_{\beta,y}$ states. □

Lemma 5 *Let $\beta_0^\star$ and $\beta_1^\star$ be two solutions of*

$$\min_{\beta \in \mathbb{R}^p} f(\beta) + g(\beta) \tag{15}$$

where f is proper, convex and $\mathrm{C}^2(\mathbb{R}^p)$ function, and g is proper, convex and lower semicontinuous with a non-necessarily full-domain. Then

$$\nabla f(\beta_0^\star) = \nabla f(\beta_1^\star).$$

Recall that the subdifferential of a proper, lower semicontinous and convex function $g : \beta \in \mathbb{R}^p \mapsto \mathbb{R} \cup \{+\infty\}$ is a maximal monotone (set-valued) operator (Lemaréchal and Hiriart-Urruty 1996), i.e. for every $\beta_1, \beta_2 \in \mathrm{dom}(f)$, and $\eta_1 \in \partial f(\beta_1)$ and $\eta_2 \in \partial g(\beta_2)$, the following holds

$$\langle \beta_1 - \beta_2,\, \eta_1 - \eta_2 \rangle \geqslant 0.$$

Moreover, if g is (Gâteaux) differentiable at β then $\nabla g(\beta)$ is its unique subgradient, i.e. $\partial g(\beta) = \{\nabla g(\beta)\}$.

Proof Let $\beta_0^\star$ and $\beta_1^\star$ be two distinct solutions of (15), otherwise, there is nothing to prove. We denote $\beta_t^\star = \beta_0^\star + th$ where $h = \beta_1^\star - \beta_0^\star$, $t \in [0, 1]$. By convexity, $\beta_t^\star$ is also a minimizer of (15). Similarily to Lemma 3, we have $-\nabla f(\beta_t^\star) \in \partial g(\beta_t^\star)$. Convexity of g then yields

$$\langle \nabla f(\beta_t^\star) - \nabla f(\beta_0^\star),\, th \rangle \leqslant 0.$$

Similarly, convexity of f entails

$$\langle \nabla f(\beta_t^\star) - \nabla f(\beta_0^\star),\, th \rangle \geqslant 0.$$

Combining these inequalities yields, for any $t \in [0, 1]$

$$\langle \nabla f(\beta_t^\star) - \nabla f(\beta_0^\star),\, h \rangle = 0. \tag{16}$$

Since f is $\mathrm{C}^2(\mathbb{R}^p)$, Taylor expansion gives

$$\nabla f(\beta_1^\star) - \nabla f(\beta_0^\star) = \int_0^1 \mathrm{D}^2 f(\beta_t^\star) h \mathrm{d}t \ , \tag{17}$$

which, after taking the inner product of both sides with h and using (16), yields

$$\langle \nabla f(\beta_1^\star) - \nabla f(\beta_0^\star),\, h\rangle = \int_0^1 \langle \mathrm{D}^2 f(\beta_t^\star) h,\, h\rangle \mathrm{d}t = 0. \tag{18}$$

By convexity, the Hessian $\partial^2 f(\beta_t^\star)$ is semidefinite positive, and (18) implies that

$$\forall t \in [0,1], \quad \langle \mathrm{D}^2 f(\beta_t^\star) h,\, h\rangle = 0,$$

or equivalently

$$\|\mathrm{D}^2 f(\beta_t^\star)^{1/2} h\| = 0 \Leftrightarrow h \in \operatorname{Ker} \mathrm{D}^2 f(\beta_t^\star) \ .$$

Inserting this again in (17) yields the desired claim. □

6.2 Proof of Theorem 1

Let $y \notin \mathcal{H}$ and $\beta^\star$ be a solution of $(\mathcal{P}(y))$ such that $(\mathcal{C}_{\beta^\star,y})$ holds. We denote $T_{\beta^\star} = T = S^\perp$.

We define the following mapping

$$\mathit{\Gamma} : (\beta_T, y) \in T \times \mathbb{R}^n \mapsto \nabla_1 F(\beta_T, y)_T + e\,(\beta_T)\,.$$

Observe that owing to Proposition 1, the first equation of Lemma 3 is equivalent to $\mathit{\Gamma}(\beta_T^\star, y) = 0$.

Note that any $\beta_T \in \widetilde{T}$ such that $\mathit{\Gamma}(\beta_T, y) = 0$ is a solution of the constrained problem

$$\min_{\alpha \in T} F(\alpha, y) + J(\alpha) \ . \tag{$(\mathcal{P}(y)_T)$}$$

It comes from the fact that $\mathit{\Gamma}(\beta_T, y) = 0$ is the first-order minimality condition over the subspace T.

We split the proof in three steps. We first show that there exists a mapping $\bar{y} \mapsto \tilde{\beta}(\bar{y}) \in T$ and an open neighborhood $\mathcal{V}$ of y such that every element $\bar{y}$ of $\mathcal{V}$ satisfies $\mathit{\Gamma}(\tilde{\beta}(\bar{y})_T, \bar{y}) = 0$ and $\tilde{\beta}(\bar{y})_S = 0$. Then, we prove that $\tilde{\beta}(\bar{y})$ is a solution of $(\mathcal{P}(\bar{y}))$ for $\bar{y} \in \mathcal{V}$. Finally, we obtain (10) from the implicit function theorem.

Step 1: construction of $\tilde{\beta}(\bar{y})$. The Jacobian of Γ with respect to the first variable reads

$$\mathrm{D}_1\Gamma(\beta_T^\star, \bar{y}) = \mathrm{D}_1^2 F_T(\beta_T^\star, \bar{y})_T + \mathrm{D}_1 e\,(\beta_T^\star)\,,$$

where D_1 denotes the derivative with respect to the first variable. Moreover, since $\beta^\star \in \widetilde{T}$, Assumption (C_{sm}) yields $\mathrm{D}_1 e\,(\beta_T^\star) = \mathrm{D}^2 J_T(\beta_T^\star)$. Thus, we get

$$\mathrm{D}_1\Gamma(\beta_T^\star, \bar{y}) = \mathrm{D}_1^2 F_T(\beta_T^\star, \bar{y}) + \mathrm{D}^2 J_T(\beta_T^\star).$$

The linear operator mapping $\mathrm{D}_1\Gamma(\beta_T^\star, y)$ is invertible on T according to Lemma 4. Hence, using the implicit function theorem restricted to T, there exists a neighborhood $\widetilde{\mathcal{V}}$ of y such that we can define a mapping $\tilde{\beta}_T : \widetilde{\mathcal{V}} \to T$ which is $\mathrm{C}^1(\widetilde{\mathcal{V}})$, and satisfies for $\bar{y} \in \widetilde{\mathcal{V}}$

$$\Gamma(\tilde{\beta}_T(\bar{y}), \bar{y}) = 0 \quad \text{and} \quad \tilde{\beta}_T(y) = \beta_T^\star.$$

We then extend $\tilde{\beta}(\bar{y})$ on S as $\tilde{\beta}_S(\bar{y}) = 0$. which defines a continuous mapping $\tilde{\beta} : \widetilde{\mathcal{V}} \to T \subset \mathbb{R}^p$.

Step 2: checking the first-order minimality condition on S. We now have to check the first order conditions on S, i.e. to check that $-\nabla_1 F(\tilde{\beta}(\bar{y}), \bar{y}) \in \partial J(\tilde{\beta}(\bar{y}))$. We distinguish two cases.

- Assume that $-\nabla_1 F(\beta^\star, y) \in \operatorname{ri} \partial J(\beta^\star)$: we show that for a sufficiently small neighbourhood of y, we also have $-\nabla_1 F(\tilde{\beta}(\bar{y}), \bar{y}) \in \operatorname{ri} \partial J(\tilde{\beta}(\bar{y}))$. First, since $\tilde{\beta} : \widetilde{\mathcal{V}} \to T$ is continuous, for any $\varepsilon > 0$, there exists a neighborhood $\bar{\mathcal{V}} \subset \widetilde{\mathcal{V}}$ of y such that
$$\|\tilde{\beta}(\bar{y}) - \beta^\star\| \leqslant \varepsilon \quad \forall\, \bar{y} \in \bar{\mathcal{V}}\ .$$
By virtue of Assumption (C_{sharp}), one can then choose ε sufficiently small to conclude that $S_{\tilde{\beta}(\bar{y})} = S$ for any $\bar{y} \in \bar{\mathcal{V}}$.
Suppose that there is a sequence $(y_\ell)_\ell$ approaching y such that
$$-\nabla_1 F(\tilde{\beta}(y_\ell), y_\ell) \notin \operatorname{ri} \partial J(\tilde{\beta}(y_\ell))$$
for all ℓ. This can be equivalently written, owing to Lemma 2, as
$$\exists u_\ell \in S_{\tilde{\beta}(y_\ell)}, \quad \forall v \in \partial J(\tilde{\beta}(y_\ell)) \quad \langle u_\ell,\, v + \nabla_1 F(\beta^\star, y)\rangle \leqslant 0, \forall \ell\ ,$$
or
$$\exists u_\ell \in S_{\tilde{\beta}(y_\ell)}, \quad \sup\ \langle u_\ell,\, \partial J(\tilde{\beta}(y_\ell)) + \nabla_1 F(\tilde{\beta}(y_\ell), y_\ell)\rangle \leqslant 0, \forall \ell\ .$$
Recall that the sequence u_ℓ can be taken on the unit sphere, and therefore has a non-zero cluster point, say u, which belongs to S as $S_{\tilde{\beta}(y_\ell)}$ converges to S. We now claim that
$$\sup\ \langle u,\, \partial J(\beta^\star) + \nabla_1 F(\beta^\star, y)\rangle \leqslant 0\ .$$

Consider any $\eta \in \partial J(\beta^\star)$. Since $\tilde{\beta}(y_\ell)$ converges to $\beta^\star$ in T, we have from the argument above that $T_{\tilde{\beta}(y_\ell)} = T$ for ℓ sufficiently large. This together with Assumption (C_{cont}), which means that $\partial J(\beta)$ is continuous on $\widetilde{T}$, allow to deduce that $\partial J(\tilde{\beta}(y_\ell))$ converges to $\partial J(\beta^\star)$. Thus, there exists a sequence $\eta_\ell \in \partial J(\tilde{\beta}(y_\ell))$ converging to η. Now, continuity of the mapping

$$y_\ell \in \widetilde{\mathcal{V}} \mapsto \nabla_1 F(\tilde{\beta}(y_\ell), y_\ell) \in \mathbb{R}^p$$

(since $\tilde{\beta}$ and $\nabla_1 F$ are both continuous on T and $\mathbb{R}^p \times \mathbb{R}^n$) yields also that $\nabla_1 F(\tilde{\beta}(y_\ell), y_\ell)$ converges to $\nabla_1 F(\beta^\star, y)$. Since

$$\langle u_\ell,\, \eta_\ell + \nabla_1 F(\tilde{\beta}(y_\ell), y_\ell)\rangle \leqslant \sup\ \langle u_\ell,\, \partial J(\tilde{\beta}(y_\ell)) + \nabla_1 F(\tilde{\beta}(y_\ell), y_\ell)\rangle \leqslant 0\ , \forall \ell$$

we get that

$$\langle u,\, \eta + \nabla_1 F(\beta^\star, y)\rangle \leqslant 0\ .$$

The latter inequality holds for any $\eta \in \partial J(\beta^\star)$, which, in view of Lemma 2, means that $-\nabla_1 F(\beta^\star, y) \notin \operatorname{ri} \partial J(\beta^\star)$. But this contradicts our initial assumption.

- We now turn to the case where $-\nabla_1 F(\beta^\star, y) \in \operatorname{rbd} \partial J(\beta^\star)$. Observe that $(y, \beta^\star) \in \mathcal{A}_T$. In particular $y \in \Pi_{n+p,n}(\mathcal{A}_T)$. Since by assumption $y \notin \mathcal{H}$, one has $y \notin \operatorname{bd}(\Pi_{n+p,n}(\mathcal{A}_T))$. Hence, there exists an open ball $\mathbb{B}(y, \varepsilon)$ for some $\varepsilon > 0$ such that $\mathbb{B}(y, \varepsilon) \subset \Pi_{n+p,n}(\mathcal{A}_T)$. Thus for every $\bar{y} \in \mathbb{B}(y, \varepsilon)$, there exists $\bar{\beta} \in \widetilde{T}$ such that

$$-\nabla_1 F(\bar{\beta}, \bar{y}) \in \operatorname{rbd} \partial J(\bar{\beta}).$$

Since $\partial J(\bar{\beta}) \subset \bar{S}_{\bar{\beta}}$ and $\bar{\beta} \in T$, $\bar{\beta}$ is a solution of $(\mathcal{P}(\bar{y})_T)$. Thus, applying Lemma 5 with $f = F(\cdot, y)$ and $g = J + \iota_T$, where ι_T is the indicator function of T, we deduce that all solutions of $(\mathcal{P}(\bar{y})_T)$ share the same gradient, whence we get $\nabla_1 F(\bar{\beta}, \bar{y}) = \nabla_1 F(\tilde{\beta}(\bar{y}), \bar{y})$. Since $\tilde{\beta}(\bar{y}) \in T$, for $\bar{y}$ sufficiently close to y, Assumption (C_{sharp}) allows to deduce that

$$T_{\tilde{\beta}(\bar{y})} = T.$$

In view of Proposition 1 and by definition of the mapping $\tilde{\beta}_T$, we deduce that for all $\bar{y} \in \mathcal{V} \cap \bar{\mathcal{V}}$, $-\nabla_1 F(\tilde{\beta}(\bar{y}), \bar{y})_T = e(\tilde{\beta}(\bar{y})) = \eta_T$, for any $\eta \in \partial J(\tilde{\beta}(\bar{y}))$. Combining this with convexity of J and the fact that $\bar{\beta} - \tilde{\beta}(\bar{y}) \in$

T, implies that $\forall \alpha \in \mathbb{R}^p$

$$
\begin{aligned}
J(\alpha) &\geqslant J(\bar{\beta}) - \langle \nabla_1 F(\bar{\beta}, \bar{y}),\, \alpha - \bar{\beta} \rangle \\
&= J(\tilde{\beta}(\bar{y})) - \langle \nabla_1 F(\tilde{\beta}(\bar{y}), \bar{y}),\, \alpha - \tilde{\beta}(\bar{y}) \rangle \\
&\quad + J(\bar{\beta}) - J(\tilde{\beta}(\bar{y})) + \langle \nabla_1 F(\tilde{\beta}(\bar{y}), \bar{y}),\, \alpha - \tilde{\beta}(\bar{y}) \rangle - \langle \nabla_1 F(\bar{\beta}, \bar{y}),\, \alpha - \bar{\beta} \rangle \\
&= J(\tilde{\beta}(\bar{y})) - \langle \nabla_1 F(\tilde{\beta}(\bar{y}), \bar{y}),\, \alpha - \tilde{\beta}(\bar{y}) \rangle \\
&\quad + J(\bar{\beta}) - J(\tilde{\beta}(\bar{y})) + \langle \nabla_1 F(\tilde{\beta}(\bar{y}), \bar{y}),\, \bar{\beta} - \tilde{\beta}(\bar{y}) \rangle \\
&= J(\tilde{\beta}(\bar{y})) - \langle \nabla_1 F(\tilde{\beta}(\bar{y}), \bar{y}),\, \alpha - \tilde{\beta}(\bar{y}) \rangle \\
&\quad + J(\bar{\beta}) - J(\tilde{\beta}(\bar{y})) + \langle \nabla_1 F(\tilde{\beta}(\bar{y})_T, \bar{y}),\, \bar{\beta} - \tilde{\beta}(\bar{y}) \rangle \\
&= J(\tilde{\beta}(\bar{y})) - \langle \nabla_1 F(\tilde{\beta}(\bar{y}), \bar{y}),\, \alpha - \tilde{\beta}(\bar{y}) \rangle \\
&\quad + J(\bar{\beta}) - J(\tilde{\beta}(\bar{y})) - \langle e(\tilde{\beta}(\bar{y})),\, \bar{\beta} - \tilde{\beta}(\bar{y}) \rangle \\
&= J(\tilde{\beta}(\bar{y})) - \langle \nabla_1 F(\tilde{\beta}(\bar{y}), \bar{y}),\, \alpha - \tilde{\beta}(\bar{y}) \rangle \\
&\quad + J(\bar{\beta}) - J(\tilde{\beta}(\bar{y})) - \langle \eta,\, \bar{\beta} - \tilde{\beta}(\bar{y}) \rangle \quad \forall \eta \in \partial J(\tilde{\beta}(\bar{y})) \\
&\geqslant J(\tilde{\beta}(\bar{y})) - \langle \nabla_1 F(\tilde{\beta}(\bar{y}), \bar{y}),\, \alpha - \tilde{\beta}(\bar{y}) \rangle \ ,
\end{aligned}
$$

which in turn is equivalent to $-\nabla_1 F(\tilde{\beta}(\bar{y}), \bar{y}) \in \partial J(\tilde{\beta}(\bar{y}))$.
We conclude that

$$\forall \bar{y} \in \mathbb{B}(y, \varepsilon), \quad -\nabla_1 F(\tilde{\beta}(\bar{y}), \bar{y}) \in \partial J(\tilde{\beta}(\bar{y})).$$

According to Lemma 3, the vector $\tilde{\beta}(\bar{y})$ is a solution of $(\mathcal{P}(\bar{y}))$.

Step 3: computing the differential. By virtue of step 1., we are in position to use the implicit function theorem, and we get the Jacobian of $\tilde{\beta}_T$ as

$$\mathrm{D}\tilde{\beta}_T(\bar{y}) = -\big(\mathrm{D}_1 \Gamma(\tilde{\beta}_T(\bar{y}), \bar{y})\big)^{-1} \big(\mathrm{D}_2 \Gamma(\tilde{\beta}_T(\bar{y}), \bar{y})\big)$$

where

$$\mathrm{D}_2 \Gamma(\beta_T, \bar{y}) = \mathrm{P}_T \circ \mathrm{D}^2_{12} F(\beta_T, \bar{y}),$$

which leads us to (10). □

6.3 Proof of Theorem 2

We first show that all solutions of $(\mathcal{P}(y))$ share the same image under the action of X, which in turn implies that the prediction/response vector $\widehat{\mu}$ is a single-valued mapping of y.

Proof (of Lemma 1) The first part of the lemma comes from the following equivalent statements:

$$
\begin{aligned}
& z \in \ker(\mathrm{D}^2_1 F_T(\beta, y)) \cap T \\
\iff & \langle z_T,\, \mathrm{D}^2_1 F_T(\beta, y) z_T \rangle = \langle X_T z,\, \mathrm{D}^2_1 F_0(X\beta, y) X_T z \rangle = 0 \\
\iff & z \in \ker(X_T) \ .
\end{aligned}
$$

Let $\beta_0^\star, \beta_1^\star$ be two solutions of $(\mathcal{P}(y))$ such that $X\beta_0^\star \neq X\beta_1^\star$. Take any convex combination $\beta_t^\star = (1-t)\beta_0^\star + t\beta_1^\star$, $t \in]0,1[$. Strict convexity of $\mu \mapsto F_0(\mu, y)$ implies that the Jensen inequality is strict, i.e.

$$F_0(X\beta_t^\star, y) < (1-t)F_0(X\beta_0^\star, y) + tF_0(X\beta_1^\star, y).$$

The convexity of the regularization implies

$$J(\beta_t^\star) \leqslant (1-t)J(\beta_0^\star) + tJ(\beta_1^\star) \ .$$

Summing these two inequalities we arrive at

$$F_0(X\beta_t^\star, y) + J(\beta_t^\star) < F_0(X\beta_0^\star, y) + J(\beta_0^\star)$$

a contradiction since $\beta_0^\star$ is a minimizer of $(\mathcal{P}(y))$. $\square$

Lemma 6 *There always exists a solution $\beta^\star$ of $(\mathcal{P}(y))$ such that $(\mathcal{C}_{\beta^\star,y})$ holds.*

Proof Let $\beta^\star$ a solution of $(\mathcal{P}(y))$ such that $(\mathcal{C}_{\beta^\star,y})$ does not hold. Consider the associated subspace $T = T_{\beta^\star}$. Thus, for any $h \in \big(\ker(X) \cap T \cap \ker(\mathrm{D}^2 J_T(\beta^\star))\big) \setminus \{0\}$, we have $X_T h = 0$ and $\mathrm{D}^2 J_T(\beta^\star)\, h = 0$. Let $v_t = \beta^\star + th$, $\forall\, t > 0$. Obviously, $v_t \in T$ since J is partly smooth at $\beta^\star$ relative to the linear manifold T. Moreover, $X_T v_t = X_T \beta^\star$, and thus $F(X_T v_t, y) = F(X_T \beta^\star, y)$.

Using convexity of J and $h \in T$, we have $\forall\, \eta \in \partial J(v_t)$

$$\begin{aligned} J(v_t) &\leqslant J(\beta^\star) + t\langle \eta,\, h\rangle \\ &= J(\beta^\star) + t\langle \eta_T,\, h\rangle \ . \end{aligned}$$

Since J obeys Assumption (C_{sharp}) and $v_t \in T$, for t sufficiently small, we have $T_{v_t} = T$, whence we get

$$J(v_t) \leqslant J(\beta^\star) + t\langle e(v_t),\, h\rangle \ .$$

where we used Proposition 1. From Assumption (C_{sm}), Taylor expansion gives

$$e(v_t) = e(\beta^\star) + t\mathrm{D}^2 J_T(\beta^\star)h + t\varepsilon(th)\|h\| = e(\beta^\star) + t\varepsilon(th)\|h\| \ ,$$

with $\lim_{t\to 0} \varepsilon(th) = 0$. Altogether, we arrive at

$$J(v_t) \leqslant J(\beta^\star) + t\left(\langle e(\beta^\star),\, h\rangle + t\|\varepsilon(th)\|\|h\|^2\right) \ .$$

Suppose now that there exists no $\beta^\star$ such that $(\mathcal{C}_{\beta^\star,y})$ holds. Then, we can always find a solution $\beta^\star$ such that[5] $e(\beta^\star) \notin \big(\ker(X) \cap T \cap \ker(\mathrm{D}^2 J_T(\beta^\star))\big)^\perp$, and therefore there is some $h \in \big(\ker(X) \cap T \cap \ker(\mathrm{D}^2 J_T(\beta^\star))\big) \setminus \{0\}$ such that

$$\langle e(\beta^\star),\, h\rangle < 0$$

and thus

$$F(X_T v_t, y) + J(v_t) < F(X_T \beta^\star, y) + J(\beta^\star) \ ,$$

for t sufficiently small, leading to a contradiction. $\square$

[5] Recall that $e(\beta^\star)$ is always different from the origin unless $\beta^\star = 0$.

We can now prove Theorem 2. At any $y \notin \mathcal{H}$, we consider $\beta^\star$ a solution of $(\mathcal{P}(y))$ such that $(\mathcal{C}_{\beta^\star,y})$ holds, which is always verified owing to Lemma 6. According to Theorem 1, one can construct a mapping $\tilde{\beta}(\bar{y})$ which coincides with $\beta^\star$ at y, and is C^1 for $\bar{y}$ in a neighborhood of y. Since $\widehat{\mu}(\bar{y}) = X\tilde{\beta}(\bar{y})$ on this neighborhood, this shows that $y \mapsto \widehat{\mu}(y)$ is in turn C^1 in a neighbourhood of y, and its differential is equal to $\Delta(y)$. Note that this shows that this computation is independent of the particular choice of $\beta^\star$ provided that $(\mathcal{C}_{\beta^\star,y})$ holds.

6.4 Proof of Theorem 3

(i) We obtain this assertion by proving that all $\mathcal{H}_T$ are of zero measure for all T and that the union is over a finite set, because of $(C_\mathcal{T})$.

- Since J is definable by $(C_\mathcal{O})$, $\nabla_1 F(\beta, y)$ is also definable by virtue of Proposition 2.
- Given $T \in \mathcal{T}$, $\widetilde{T}$ is also definable. Indeed, $\widetilde{T}$ can be equivalently written

$$\widetilde{T} = \{\beta \ : \ \forall \xi \in T \text{ and } \langle d_i,\, \alpha\rangle = 0 \ \forall i \text{ s.t. } \langle d_i,\, \beta\rangle = 0 \Rightarrow \xi = \alpha\} \ .$$

  which involves algebraic (in fact linear) sets, whence definability follows after interpreting the logical notations (conjunction and universal quantifiers) in the first-order formula in terms of set operations, and using axioms 1-4 of definability in an o-minimal structure.
- Let $\boldsymbol{D} : \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ the set-valued mapping whose graph is

$$\mathcal{G}(\boldsymbol{D}) = \{(\beta, \eta) \ : \ \eta \in \operatorname{ri} \partial J(\beta)\} \ .$$

  From Lemma 10, $\mathcal{G}(\boldsymbol{D})$ is definable. Since the graph ∂J is closed (Lemaréchal and Hiriart-Urruty 1996), and definable (Lemma 3), the set

$$\{(\beta, \eta) \ : \ \eta \in \operatorname{rbd} \partial J(\beta)\} = \mathcal{G}(\partial J) \setminus \mathcal{G}(\boldsymbol{D}) \ ,$$

  is also definable by axiom 1. This entails that $\mathcal{A}_T$ is also a definable subset of $\mathbb{R}^n \times \widetilde{T}$ since

$$\begin{aligned}\mathcal{A}_T = (\mathbb{R}^n \times \widetilde{T} \times \mathbb{R}^n) &\cap \{(y, \beta, \eta) \ : \ \eta = -\nabla_1 F(\beta_T, y)\} \\ &\cap (\mathbb{R}^n \times \{(\beta, \eta) \ : \ \eta \in \operatorname{rbd} \partial J(\beta)\}) \ .\end{aligned}$$

- By axiom 4, the canonical projection $\Pi_{n+p,n}(\mathcal{A}_T)$ is definable, and its boundary $\mathcal{H}_T = \operatorname{bd}(\Pi_{n+p,n}(\mathcal{A}_T))$ is also definable by (Coste 1999, Proposition 1.12) with a strictly smaller dimension than $\Pi_{n+p,n}(\mathcal{A}_T)$ (Coste 1999, Theorem 3.22).
- We recall now from (Coste 1999, Theorem 2.10) that any definable subset $A \subset \mathbb{R}^n$ in $\mathcal{O}$ can be decomposed (stratified) in a disjoint finite union of q subsets C_i, definable in $\mathcal{O}$, called cells. The dimension of A is (Coste 1999, Proposition 3.17(4))

$$d = \max_{i \in \{1,\ldots,q\}} d_i \leqslant n \ ,$$

where $d_i = \dim(C_i)$. Altogether we get that

$$\dim \mathcal{H}_T = \dim \operatorname{bd}(\Pi_{n+p,n}(\mathcal{A}_T)) < \dim \Pi_{n+p,n}(\mathcal{A}_T) = d \leqslant n$$

whence we deduce that $\mathcal{H}$ is of zero measure with respect to the Lebesgue measure on $\mathbb{R}^n$ since the union is taken over the finite set $\mathcal{T}$ by $(C_\mathcal{T})$.

(ii) $F_0(\cdot, y)$ is strongly convex with modulus τ if, and only if,

$$F_0(\mu, y) = G(\mu, y) + \frac{\tau}{2}\|\mu\|^2$$

where $G(\cdot, y)$ is convex and satisfies (C_F), and in particular its domain in μ is full-dimensional. Thus, $(\mathcal{P}(y))$ amounts to solving

$$\min_{\beta \in \mathbb{R}^p} \frac{\tau}{2}\|X\beta\|^2 + G(X\beta, y) + \lambda J(\beta).$$

It can be recasted as a constrained optimization problem

$$\min_{\mu \in \mathbb{R}^n, \beta \in \mathbb{R}^p} \frac{\tau}{2}\|\mu\|^2 + G(\mu, y) + \lambda J(\beta) \text{ s.t. } \mu = X\beta.$$

Introducing the image (XJ) of J under the linear mapping X, it is equivalent to

$$\min_{\mu \in \mathbb{R}^n} \frac{\tau}{2}\|\mu\|^2 + G(\mu, y) + \lambda (XJ)(\mu) \ , \tag{19}$$

where $(XJ)(\mu) = \min_{\{\beta \in \mathbb{R}^p \,:\, \mu = X\beta\}} \lambda J(\beta)$. This is a proper closed convex function, which is finite on $\operatorname{Im}(X)$. The minimization problem amounts to computing the proximal point at 0 of $G(\cdot, y) + \lambda(XJ)$, which is a proper closed and convex function. Thus this point exists and is unique.
Furthermore, by assumption on F_0, the difference function $F_0(\cdot, y_1) - F_0(\cdot, y_2) = G(\cdot, y_1) - G(\cdot, y_2)$ is Lipschitz continuous on $\mathbb{R}^p$ with Lipschitz constant $L\|y_1 - y_2\|$. It then follows from (Bonnans and Shapiro 2000, Proposition 4.32) that $\widehat{\mu}(\cdot)$ is Lipschitz continuous with constant $2L/\tau$. Moreover, h is Lipschitz continuous, and thus so is the composed mapping $h \circ \widehat{\mu}(\cdot)$. From (Evans and Gariepy 1992, Theorem 5, Section 4.2.3), weak differentiability follows.
Rademacher theorem asserts that a Lipschitz continuous function is differetiable Lebesgue a.e. and its derivative and weak derivative coincide Lebesgue a.e., (Evans and Gariepy 1992, Theorem 2, Section 6.2). Its weak derivative, whenever it exsist, is upper-bounded by the Lipschitz constant. Thus

$$\mathbb{E}\left(\left|\frac{\partial (h \circ \widehat{\mu})_i}{\partial y_i}(Y)\right|\right) < +\infty \ .$$

(iii) Now, by the chain rule (Evans and Gariepy 1992, Remark, Section 4.2.2), the weak derivative of $h \circ \widehat{\mu}(\cdot)$ at y is precisely

$$\mathrm{D}(h \circ \widehat{\mu})(y)) = \mathrm{D}h\,(\widehat{\mu}(y))\,\Delta(y) \ .$$

This formula is valid everywhere except on the set $\mathcal{H}$ which is of Lebesgue measure zero as shown in (i). We conclude by invoking (ii) and Stein's lemma (Stein 1981) to establish unbiasedness of the estimator $\widehat{df}$ of the DOF.

(iv) Plugging the DOF expression (iii) into that of the SURE (Stein 1981, Theorem 1), the statement follows.

□

6.5 Proof of Theorem 4

For (i)-(iii), the proof is exactly the same as in Theorem 3. For (iv): combining the DOF expression (iii) and (Eldar 2009, Theorem 1), and rearranging the expression yields the stated result.

7 Conclusion

In this paper, we proposed a detailed sensitivity analysis of a class of estimators obtained by minimizing a general convex optimization problem with a regularizing penalty promoting some low complexity models. This allowed us to derive an analytical expression of the local variations of these estimators to perturbations of the observations, and also to prove that the set where the estimator behaves non-smoothly as a function of the obervations is of zero Lebesgue measure. Both results paved the way to derive unbiased estimators of the prediction risk in two random scenarios, one of which covers the (continuous) exponential family. This analysis covers a large set of convex variational estimators routinely used in statistics and imaging (most notably group sparsity and multidimensional total variation penalty). It is also important to note that our proof strategy carries over to more exotic regularizations that are not just of block type, such as generic finite-valued polyhedral regularizers. The key here is that the underlying solutions promoted by these regularizers live on some low-dimensional subspace. An important research program is to extend this analysis to regularizers that do not promote locally solutions belonging to some subspace, but rather to a smooth manifold. This is for instance the case of the nuclear norm (also known as the trace norm), which locally promotes matrices having a fixed (hopefully low) rank.

Acknowledgements This work has been supported by the European Research Council (ERC project SIGMA-Vision).

A Basic Properties of o-minimal Structures

In the following results, we collect some important stability properties of o-minimal structures. To be self-contained, we also provide proofs. To the best of our knowledge, these

proofs, although simple, are not reported in the literature or some of them are left as exercices in the authoritative references van den Dries (1998); Coste (1999). Moreover, in most proofs, to show that a subset is definable, we could just write the appropriate first-order formula (see (Coste 1999, Page 12)(van den Dries 1998, Section Ch1.1.2)), and conclude using (Coste 1999, Theorem 1.13). Here, for the sake of clarity and avoid cryptic statements for the non-specialist, we will translate the first order formula into operations on the involved subsets, in particular projections, and invoke the above stability axioms of o-minimal structures. In the following, n denotes an arbitrary (finite) dimension which is not necessarily the number of observations used previously the paper.

Lemma 7 (Addition and multiplication) *Let $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}^p$ and $g : \Omega \subset \mathbb{R}^n \subset \mathbb{R}^p$ be definable functions. Then their pointwise addition and multplication is also definable.*

Proof Let $h = f + g$, and

$$B = (\Omega \times \mathbb{R} \times \Omega \times \mathbb{R} \times \Omega \times \mathbb{R}) \cap (\Omega \times \mathbb{R} \times \mathcal{G}(f) \times \mathcal{G}(h)) \cap S$$

where $S = \{(x, u, y, v, z, w) \ : \ x = y = z, u = v + w\}$ is obviously an algebraic (in fact linear) subset, hence definable by axiom 2. Axiom 1 and 2 then imply that B is also definable. Let $\Pi_{3n+3p,n+p} : \mathbb{R}^{3n+3p} \to \mathbb{R}^{n+p}$ be the projection on the first $n + p$ coordinates. We then have

$$\mathcal{G}(h) = \Pi_{3n+3p,n+p}(B)$$

whence we deduce that h is definable by applying $3n + 3p$ times axiom 4. Definability of the pointwise mutplication follows the same proof taking $u = v \cdot w$ in S. □

Lemma 8 (Inequalities in definable sets) *Let $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}$ be a definable function. Then $\{x \in \Omega \ : \ f(x) > 0\}$, is definable. The same holds when replacing $>$ with $<$.*

Clearly, inequalities involving definable functions are accepted when defining definable sets.

There are many possible proofs of this statement.

Proof (1) Let $B = \{(x, y) \in \mathbb{R} \times \mathbb{R} \ : \ f(x) = y\} \cap (\Omega \times (0, +\infty)$, which is definable thanks to axioms 1 and 3, and that the level sets of a definable function are also definable. Thus

$$\{x \in \Omega \ : \ f(x) > 0\} = \{x \in \Omega \ : \ \exists y, f(x) = y, y > 0\} = \Pi_{n+1,n}(B) \ ,$$

and we conclude using again axiom 4. □

Yet another (simpler) proof.

Proof (2) It is sufficient to remark that $\{x \in \Omega \ : \ f(x) > 0\}$ is the projection of the set $\big\{(x, t) \in \Omega \times \mathbb{R} \ : \ t^2 f(x) - 1 = 0\big\}$, where the latter is definable owing to Lemma 7. □

Lemma 9 (Derivative) *Let $f : I \to \mathbb{R}$ be a definable differentiable function on an open interval I of $\mathbb{R}$. Then its devivative $f' : I \to \mathbb{R}$ is also definable.*

Proof Let $g : (x, t) \in I \times \mathbb{R} \mapsto g(x, t) = f(x + t) - f(x)$. Note that g is definable function on $I \times \mathbb{R}$ by Lemma 7. We now write the graph of f' as

$$\mathcal{G}(f') = \{(x, y) \in I \times \mathbb{R} \ : \ \forall \varepsilon > 0, \exists \delta > 0, \forall t \in \mathbb{R}, |t| < \delta, |g(x, t) - yt| < \varepsilon |t|\} \ .$$

Let $C = \big\{(x, y, v, t, \varepsilon, \delta) \in I \times \mathbb{R}^5 \ : \ ((x, t), v) \in \mathcal{G}(g)\big\}$, which is definable since g is definable and using axiom 3. Let

$$B = \big\{(x, y, v, t, \varepsilon, \delta) \ : \ t^2 < \delta^2, (v - ty)^2 < \varepsilon^2 t^2\big\} \cap C \ .$$

The first part in B is semi-algebraic, hence definable thanks to axiom 2. Thus B is also definable using axiom 1. We can now write

$$\mathcal{G}(f') = \mathbb{R}^3 \setminus \big(\Pi_{5,3}\left(\mathbb{R}^5 \setminus \Pi_{6,5}(B)\right)\big) \cap (I \times \mathbb{R}) \ ,$$

where the projectors and completions translate the actions of the existential and universal quantifiers. Using again axioms 4 and 1, we conclude. □

With such a result at hand, this proposition follows immediately.

Proposition 2 (Differential and Jacobian) *Let $f = (f_1, \cdots, f_p) : \Omega \to \mathbb{R}^p$ be a differentiable function on an open subset Ω of $\mathbb{R}^n$. If f is definable, then so its differential mapping and its Jacobian. In particular, for each $i = 1, \cdots, n$ and $j = 1, \cdots, p$, the partial derivative $\partial f_i / \partial x_j : \Omega \to \mathbb{R}$ is definable.*

We provide below some results concerning the subdifferential.

Proposition 3 (Subdifferential) *Suppose that f is a finite-valued convex definable function. Then for any $x \in \mathbb{R}^n$, the subdifferential $\partial f(x)$ is definable.*

Proof For every $x \in \mathbb{R}^n$, the subdifferential $\partial f(x)$ reads

$$\partial f(x) = \left\{ \eta \in \mathbb{R}^n \ : \ f(x') \geqslant f(x) + \langle \eta, \, x' - x \rangle \quad \forall x' \in \mathbb{R}^n \right\}.$$

Let $K = \{(\eta, x') \in \mathbb{R}^n \times \mathbb{R}^n \ : \ f(x') < f(x) + \langle \eta, \, x' - x \rangle\}$. Hence, $\partial f(x) = \mathbb{R}^n \setminus \Pi_{2n,n}(K)$. Since f is definable, the set K is also definable using Lemma 7 and 8, whence definability of $\partial f(x)$ follows using axiom 4. □

Lemma 10 (Graph of the relative interior) *Suppose that f is a finite-valued convex definable function. Then, the set*

$$\{(x, \eta) \ : \ \eta \in \operatorname{ri} \partial f(x)\}$$

is definable.

Proof Denote $C = \{(\beta, \eta) \ : \ \eta \in \operatorname{ri} \partial f(\beta)\}$. Using the characterization of the relative interior of a convex set (Rockafellar 1996, Theorem 6.4), we rewrite C in the more convenient form

$$\begin{aligned} C = \{(x, \eta) \ : & \forall u \in \mathbb{R}^n, \forall z \in \mathbb{R}^n, f(z) - f(x) \geqslant \langle u, \, z - x \rangle, \\ & \exists t > 1, \forall x' \in \mathbb{R}^n, f(x') - f(x) \geqslant \langle (1-t)u + t\eta, \, x' - x \rangle\}. \end{aligned}$$

Let $D = \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \times (1, +\infty) \times \mathbb{R}^n$ and K defined as

$$K = \left\{ (x, \eta, u, z, t, x') \in D \ : \ f(z) - f(x) \geqslant \langle u, \, z - x \rangle), f(x') - f(x) \geqslant \langle (1-t)u + t\eta, \, x' - x \rangle \right\}.$$

Thus,

$$C = \mathbb{R}^{2n} \setminus \Pi_{3n,2n} \left(\mathbb{R}^{3n} \setminus \Pi_{4n,3n} \left(\Pi_{4n+1,4n} \left(\mathbb{R}^{4n} \times (1, +\infty) \setminus \Pi_{5n+1,4n+1}(K) \right) \right) \right),$$

where the projectors and completions translate the actions of the existential and universal quantifiers. Using again axioms 4 and 1, we conclude. □

References

Bach F (2008) Consistency of the group lasso and multiple kernel learning. Journal of Machine Learning Research 9:1179–1225

Bach F (2010) Self-concordant analysis for logistic regression. Electronic Journal of Statistics 4:384–414

Bakin S (1999) Adaptive regression and model selection in data mining problems. Thesis (Ph.D.)–Australian National University, 1999

Bickel PJ, Ritov Y, Tsybakov A (2009) Simultaneous analysis of lasso and Dantzig selector. Annals of Statistics 37(4):1705–1732

Bolte J, Daniilidis A, Lewis AS (2011) Generic optimality conditions for semialgebraic convex programs. Mathematics of Operations Research 36(1):55–70

Bonnans J, Shapiro A (2000) Perturbation analysis of optimization problems. Springer Series in Operations Research, Springer-Verlag, New York

Brown LD (1986) Fundamentals of Statistical Exponential Families with Applica- tions in Statistical Decision Theory, Monograph Series, vol 9. Institute of Mathematical Statistics Lecture Notes, IMS, Hayward, CA

Bühlmann P, van de Geer S (2011) Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer

Bunea F (2008) Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1 + \ell_2$ penalization. Electronic Journal of Statistics 2:1153–1194

Candès E, Plan Y (2009) Near-ideal model selection by ℓ_1 minimization. Annals of Statistics 37(5A):2145–2177

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

Chen X, Lin Q, Kim S, Carbonell JG, Xing EP (2010) An efficient proximal-gradient method for general structured sparse learning. Preprint arXiv:10054717

Coste M (1999) An introduction to o-minimal geometry. Tech. rep., Institut de Recherche Mathematiques de Rennes

Coste M (2002) An introduction to semialgebraic geometry. Tech. rep., Institut de Recherche Mathematiques de Rennes

DasGupta A (2008) Asymptotic Theory of Statistics and Probability. Springer

Donoho D (2006) For most large underdetermined systems of linear equations the minimal ℓ^1-norm solution is also the sparsest solution. Communications on pure and applied mathematics 59(6):797–829

Dossal C, Kachour M, Fadili J, Peyré G, Chesneau C (2013) The degrees of freedom of penalized ℓ_1 minimization. Statistica Sinica 23(2):809–828

Efron B (1986) How biased is the apparent error rate of a prediction rule? Journal of the American Statistical Association 81(394):461–470

Eldar YC (2009) Generalized SURE for exponential families: Applications to regularization. IEEE Transactions on Signal Processing 57(2):471–481

Evans LC, Gariepy RF (1992) Measure theory and fine properties of functions. CRC Press

van de Geer SA (2008) High-dimensional generalized linear models and the lasso. Annals of Statistics 36:614–645

de Geer SV (2008) High-dimensional generalized linear models and the lasso. Annals of Statistics 36(2):614–645

Hansen NR, Sokol A (2014) Degrees of freedom for nonlinear least squares estimation. Tech. rep., arXiv preprint 1402.2997

Hudson H (1978) A natural identity for exponential families with applications in multiparameter estimation. The Annals of Statistics 6(3):473–484

Hwang JT (1982) Improving upon standard estimators in discrete exponential families with applications to poisson and negative binomial cases. Ann Statist 10(3):857–867

Kakade SM, Shamir O, Sridharan K, Tewari A (2010) Learning exponential families in high-dimensions: Strong convexity and sparsity. In: AISTATS

Kato K (2009) On the degrees of freedom in shrinkage estimation. Journal of Multivariate Analysis 100(7):1338–1352

Lemaréchal C, Hiriart-Urruty J (1996) Convex analysis and minimization algorithms: Fundamentals, vol 305. Springer-Verlag

Lemaréchal C, Oustry F, Sagastizábal C (2000) The $\mathcal{U}$-lagrangian of a convex function. Trans Amer Math Soc 352(2):711–729

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

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

Liu H, Zhang J (2009) Estimation consistency of the group lasso and its applications. Journal of Machine Learning Research 5:376–383

McCullagh P, Nelder JA (1989) Generalized Linear Models, second edition edn. Monographs on Statistics & Applied Probability, Chapman & Hall/CRC, URL http://www.worldcat.org/isbn/0412317605

Meier L, Geer SVD, Buhlmann P (2008) The group lasso for logistic regression. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70(1):51–71

Meyer M, Woodroofe M (2000) On the degrees of freedom in shape-restricted regression. Annals of Statistics 28(4):1083–1104

Mordukhovich B (1992) 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

Negahban S, Ravikumar P, Wainwright MJ, Yu B (2012) A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. Statistical Science 27(4):538–557

Osborne M, Presnell B, Turlach B (2000) A new approach to variable selection in least squares problems. IMA journal of numerical analysis 20(3):389–403

Peyré G, Fadili J, Chesneau C (2011) Adaptive Structured Block Sparsity Via Dyadic Partitioning. In: Proc. EUSIPCO 2011, EURASIP, Barcolona, Espagne, URL http://hal.archives-ouvertes.fr/hal-00597772

Rockafellar RT (1996) Convex Analysis. Princeton Landmarks in Mathematics and Physics, Princeton University Press

Rudin L, Osher S, Fatemi E (1992) Nonlinear total variation based noise removal algorithms. Physica D: Nonlinear Phenomena 60(1-4):259–268

Solo V, Ulfarsson M (2010) Threshold selection for group sparsity. In: Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on, IEEE, pp 3754–3757

Stein C (1981) Estimation of the mean of a multivariate normal distribution. The Annals of Statistics 9(6):1135–1151

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

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

Tibshirani RJ, Taylor J (2012) Degrees of freedom in Lasso problems. Ann Statist 40(2):639–1284

Tikhonov AN, Arsenin VY (1997) Solutions of Ill-posed Problems. V. H. Winston and Sons

Vaiter S, Deledalle C, Peyré G, Dossal C, Fadili J (2012a) Local behavior of sparse analysis regularization: Applications to risk estimation. Applied and Computational Harmonic Analysis

Vaiter S, Deledalle C, Peyré G, Fadili J, Dossal C (2012b) Degrees of freedom of the group Lasso. In: ICML'12 Workshops, pp 89–92

Vaiter S, Golbabaee M, Fadili J, Peyré G (2013) Model selection with piecewise regular gauges. Tech. rep., Preprint Hal

Vaiter S, Peyré G, Fadili JM (2014) Model Consistency of Partly Smooth Regularizers. arXiv:1405.1004

van den Dries L (1998) Tame topology and o-minimal structures, Math. Soc. Lecture Note, vol 248. Cambridge Univ Press

van den Dries L, Miller C (1996) Geometric categories and o-minimal structures. Duke Math J 84:497–540

Wei F, Huang J (2010) Consistent group selection in high-dimensional linear regression. Bernoulli 16(4):1369–1384

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

Yuan M, Lin Y (2006) Model selection and estimation in regression with grouped variables. J of The Roy Stat Soc B 68(1):49–67

Zou H, Hastie T, Tibshirani R (2007) On the "degrees of freedom" of the Lasso. The Annals of Statistics 35(5):2173–2192