Low Complexity Regularizations of Inverse Problems

Samuel Vaiter

Unversité Paris-Dauphine
École Doctorale de Dauphine

Régularisations de Faible Complexité pour les Problèmes Inverses

Low Complexity Regularizations of Inverse Problems

Thèse

Pour l'obtention du titre de

Docteur en Sciences
Spécialité Mathématiques Appliquées

Présentée par

Samuel VAITER

Soutenue publiquement le 10 juillet 2014 devant le jury composé de

Francis	Bach	INRIA, École Normale Supérieure	Examinateur
Emmanuel	Candès	Stanford University	Rapporteur
Antonin	Chambolle	CNRS, École Polytechnique	Examinateur
Albert	Cohen	Université Pierre et Marie Curie	Examinateur
Charles	Dossal	Université Bordeaux 1	Invité
Jalal	Fadili	ENSICAEN	Examinateur
Rémi	Gribonval	INRIA Rennes	Rapporteur
Gabriel	Peyré	CNRS, Paris Dauphine	Directeur
Vincent	Rivoirard	Paris Dauphine	Examinateur

L'université n'entend donner aucune approbation ni improbation aux opinions émises dans les thèses : ces opinions doivent être considérées comme propres à leurs auteurs.

À Brigitte, Jacques, Nathalie, Patrick et Simona.

Abstract
This thesis is concerned with recovery guarantees and sensitivity analysis of variational regularization for noisy linear inverse problems. This is cast as a convex optimization problem by combining a data fidelity and a regularizing functional promoting solutions conforming to some notion of low complexity related to their non-smoothness points. Our approach, based on partial smoothness, handles a variety of regularizers including analysis/structured sparsity, antisparsity and low-rank structure. We first give an analysis of the noise robustness guarantees, both in terms of the distance of the recovered solutions to the original object, as well as the stability of the promoted model space. We then turn to sensivity analysis of these optimization problems to observation perturbations. With random observations, we build unbiased estimator of the risk which provides a parameter selection scheme.

Keywords: inverse problem, variational regularization, low complexity prior, sparsity, robustness, sensitivity, risk estimation, degrees of freedom, parameter selection, partly smooth function.

Résumé
Cette thèse se consacre aux garanties de reconstruction et de l'analyse de sensibilité de régularisation variationnelle pour des problèmes inverses linéaires bruités. Il s'agit d'un problème d'optimisation convexe combinant un terme d'attache aux données et un terme de régularisation promouvant des solutions vivant dans un espace dit de faible complexité. Notre approche, basée sur la notion de fonctions partiellement lisses, permet l'étude d'une grande variété de régularisations comme par exemple la parcimonie de type analyse ou structurée, l'antiparcimonie et la structure de faible rang. Nous analysons tout d'abord la robustesse au bruit, à la fois en termes de distance entre les solutions et l'objet original, ainsi que la stabilité de l'espace modèle promu. Ensuite, nous étudions la stabilité de ces problèmes d'optimisation à des perturbations des observations. À partir d'observations aléatoires, nous construisons un estimateur non biaisé du risque afin d'obtenir un schéma de sélection de paramètre.

Mots-clés : problème inverse, régularisation variationnelle, a priori de faible complexité, parcimonie, robustesse, sensibilité, estimation du risque, degrés de liberté, sélection de paramètre, fonction partiellement lisse.

Remerciements

Gabriel m'a reçu à Cachan pour la première fois le 8 décembre 2010. Dès cet entretien, j'ai compris que celui-ci était enthousiaste, passionné et pédagogue, ce que tout individu le connaissant pourra confirmer. Scientifiquement, mon respect est total envers Gabriel. Je n'ai appris que par la suite qu'il était aussi quelqu'un de très humain et généreux à travers les différents aléas pouvant survenir sur trois ans. J'ai eu la chance de rencontrer, non seulement un excellent directeur de thèse, mais également un mentor. Merci Gabriel.

Je remercie Emmanuel Candès et Rémi Gribonval d'avoir accepter la tâche de rapporteur pour la présente thèse. Merci également aux membres de ce jury, Francis Bach, Antonin Chambolle, Albert Cohen, Charles Dossal, Jalal Fadili et Vincent Rivoirard. Leur présence est un véritable honneur au vu du respect scientifique que chacun m'inspire.

Je me dois de remercier toute spécialement Jalal Fadili et Charles Dossal, deux collaborateurs qui m'ont suivi tout au long de mon travail. Que ce soit humainement ou scientifiquement, travailler avec eux a été un véritable plaisir, qui j'espère sincèrement se répétera dans le futur. Jalal et Charles ont été des co-encadrants de fait, chacun avec son style propre, qui m'ont permis d'aborder les problèmes sous des angles différents.

Certains chercheurs ont eu un impact notable dans mon orientation en thèse. Je pense notamment à Renaud Keriven, Stéphane Mallat et Yves Meyer. Je tiens à remercier spécialement Lionel Moisan pour m'avoir suivi depuis le Master 1 avec attention. Lionel m'a toujours encouragé dans les différentes évolutions de mon projet professionnel.

Les discussions avec les *imagiciens* ont également façonné cette thèse. Merci à Hugo, Nicolas, JB, John, Fang, Sira, Giacomo, Mohammad, FX, Jean-Marie, Laurent, Joseph, Vincent et Charles. Travailler avec ces trois derniers a été enrichissant et plaisant. Mais il n'y a pas que l'image à Dauphine. Merci également à Loïc, Vito, Rémi, Maxime, et tous les autres thésards du CEREMADE durant ces trois années. L'environnement de travail au CEREMADE doit beaucoup à Isabelle, Marie et Gilles, je les en remercie. J'en profite pour également remercier les autres bordelais, notamment Nicolas et Jean-François, pour mes différents séjours chez eux. Merci aux Lillois pour l'accueil durant cette dernière année, notamment Stephan, Vianney et Fanny.

Je voudrais remercier mes amis, PAG, ami de longue date, Nico et Jerem. Je remercie également Michel pour ses conseils avisés et son amitié.

Les remerciements qui suivent ne sont que la partie émergée de l'iceberg. Cette thèse ne fait sens que par leur présence. Merci à mon père et sa belle histoire. Merci à ma mère d'avoir passer tant de temps à écouter cette belle histoire. Merci à eux de m'avoir nourri, dans tous les sens que ce verbe possède. Merci à mes grands-parents pour leur soutien. Merci à ma soeur pour notre complicité. *Infine*, merci à Simona pour son sourire qui me rend chaque jour un peu plus heureux.

Contents

1

Introduction

Contents

1.1 Inverse Problems and Regularization

Consider the following challenges:

- You are given an image where half of the sensors in your CCD camera are defective ! Could one recover the original image up to a given accuracy ?
- You work for a major entertainment company which is willing to build a recommender system to provide recommendations on movies based on the user's preferences. However, the data is quite incomplete since users typically rate only a few movies in the database. Could one infer the preference of any user for any movie, including the unrated ones ?
- You were recording the best performance with your rock band. Unfortunately, someone near the microphone was talking during the recording. Can you remove the voice of this uncivil ?
- You want to build a search engine for large-scale images, whose goal is to retrieve images based on a semantic query. Can one build efficient compact descriptors/features on which efficient retrieval can be based ?

Several strategies have been proposed in the past decades to solve these problems (image inpainting, matrix completion, source separation, large-scale nearest neighbor search). All these problems can be cast in the same framework, where one has access to recover an object of interest (signal, image, video, matrix, etc.) while only partial, indirect and possibly imperfect information of it is available. To handle this class of problems within the same setting, we hinge on the following triad:

(i) *Forward model :* One has to model the degradation process underlying the incomplete and corrupted observations. Throughout this thesis, we consider the case of linear forward models where both the original object and the observations live in finite-dimensional vector spaces.

(ii) *A priori :* While recovering a vector from an underdetermined system of linear equations seems hopeless by basic arguments, the situation radically changes if some information is available in the form of a prior. Here, we consider a variational formulation of this prior encoded into a convex function. More precisely, we focus on functions promoting *low complexity objects*, for instance piecewise constant, sparse or low rank.

(iii) *Computational algorithm :* In practice, it is necessary to be able to compute quickly a solution of a convex optimization problem casted as a trade-off between data fidelity (item (i)) and prior (item (ii)), hopefully unique, up to a good accuracy. It is thus important to propose an efficient algorithm, which is the case of the majority of the regularization considered here, using the structure of the problem.

1.1.1 Forward Model

This thesis is concerned with linear inverse problems in finite dimension. This framework is used in many applications in the fields of signal processing, imaging sciences, statistics and machine learning. Although one may object that this does not always conform to real world applications, where the corresponding objects may be infinite-dimensional or even continuous, our setting is sufficiently large to covers a wide spectrum of problems and practical applications in imaging or statistics. It also lends to a unified, generic and rigorous mathematical analysis.

We model physically the observed data with functions defined on a subspace $\Omega \subseteq \mathbb{R}^d$, where $d = 1$ for an audio signal, $d = 2$ for an image, etc. Let us take the example of images. Intrinsically, a physical image is the projection of an object on an optical system. Thus, the image is a function f_0 defined by the quantity of energy $f_0(\nu)$ received by the focal plane at the point ν, defining a function $\Omega \to \mathbb{R}$, where Ω corresponds to a sub-domain of the focal plane. From a mathematical point of view, one assumes that f_0 belongs to some functional space $\mathbb{H}$. Typically, we consider f_0 as a finite energy function, i.e. $\mathbb{H} = L^2(\Omega)$.

In order to take into account properties of these signals or images (smooth, piecewise smooth, oscillating) other richer functional spaces are used. For instance, one can consider the space of functions with bounded variation, a Besov or Sobolev space. Sometimes, it is more meaningful to consider f_0 as a distribution. One may think for instances of point sources in an astronomical image, e.g. stars, which can be seen as a sum of Dirac masses. Another alternative, which is not considered in this thesis, is to place a random model on the signals, which corresponds to the Bayesian approach.

In many modern digital systems, the physical quantity (light) available at the focal plane, is directly sampled on a discrete cartesian grid (by construction on CCD or CMOS camera), hence giving directly equi-spaced samples $y \in \mathbb{R}^q$ of the acquired scene. A general forward model relating the original image f_0 to the observations reads

$$y = \Psi(f_0) \odot b, \tag{1.1}$$

where Ψ is degradation operator from the signal space $\mathbb{H}$ to the observation space $\mathbb{R}^q$, b is a noise term and $\odot$ is some composition operator between the degraded data $\Psi(f_0)$ and the noise. Typically, this composition is additive or multiplicative depending on the nature of the acquisition device. The operator Ψ model the acquisition device (digital camera, scanner, etc) and typically entails some sort of degradation and loss of resolution (blurring, missing pixels, etc). The noise term b may originate from several causes. It models the fluctuations (deterministic or random) that contaminate the observations (such as thermal noise).

In the overwhelming majority of applications in image and signal processing, the forward operator Ψ is considered as linear, either exactly or to a good approximation, see e.g. (Mallat 2009). Thus, we leave aside the case of non-linear observations, such as the magnitude of complex measurements, for instance Fourier in interferometric or diffraction imaging (Hofmann et al. 1993). Moreover, the noise is considered additive in many cases, so that the forward model (1.1) reduces to the following:

$$y = \Psi f_0 + b. \tag{1.2}$$

In practice, the goal of recovering a continuous function f_0 is in many cases hopeless a numerical point of view. Our goal is thus to find a discrete approximation of this function. To achieve that, we set some basis $\mathcal{B}(L)$ of a subspace L of $\mathbb{H}$ with dimension n, for instance taking finite elements, e.g. piecewise constants on a square grid or piecewise affine on a triangulation. Thus, we obtain an approximation Bx_0 of our original signal f_0, where $x_0 \in \mathbb{R}^n$ are the coefficients of f_0 in the basis $\mathcal{B}(L)$ and B is the matrix whose columns are the atoms of the basis. It leads us to consider the following forward model (cf. Figure 1.1):

$$y = \Phi x_0 + w \tag{1.3}$$

where

$$\Phi = \Psi B : \mathbb{R}^n \to \mathbb{R}^q \quad \text{and} \quad w = b + \Psi(f_0 - Bx_0). \tag{1.4}$$

In general, the observation domain $\mathbb{R}^q$ and the computational one $\mathbb{R}^n$ are different ($q \neq n$). Indeed, q is dictated by the acquisition device, whereas n is a choice made by the numerical user, resulting from a trade-off between computational cost, precision and theoretical limit. This is the forward model that we will consider throughout this manuscript. Typically, Φ is not invertible, or badly conditioned. Beyond signal processing, the linear model is also used in statistics and machine learning under the name of regression. One can find its history in the paper of Seal (1967).

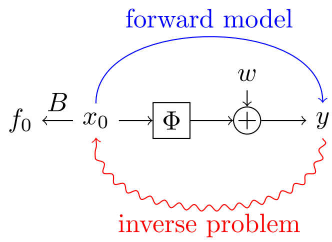


Figure 1.1: *Forward and Inverse Problem*

From now on, we focus on the problem of recovering Bx_0. This thus corresponds to a finite dimensional problem: finding a good approximation of x_0 from the observation y alone. The behavior when the grid size tends to zero raises many important and difficult issues, which will be not treated here.

For some problem, it is important to take in account the random nature of the noise, and thus to consider the stochastic forward model

$$Y = \Phi x_0 + W, \tag{1.5}$$

where the noise W is a random vector with realizations taking values in $\mathbb{R}^q$. Supposing that the noise follows a centered Gaussian density, $W \sim \mathcal{N}(0, \sigma^2 \mathrm{Id})$, we obviously have $Y \sim \mathcal{N}(\Phi x_0, \sigma^2 \mathrm{Id})$. This classical model is studied in details in (Trevor et al. 2009). Others noise models are considered in image processing, such as Poisson noise for short noise (e.g. CCD cameras, computerized tomography), and multiplicative noise in SAR imaging. We refer to (Refregier et al. 2004) and (Boncelet 2005) for a more comprehensive account on noise models in imaging systems.

When no noise corrupts the data, which is hardly the case for real life applications, the forward model becomes

$$y = \Phi x_0. \tag{1.6}$$

We now list some classical examples of the forward operator Φ used in image processing.

Denoising. The denoising problem is among the most intensively studied in the image processing literature. This step may prove crucial prior to more high-level image analysis and processing tasks, e.g. object segmentation or detection. The model (1.3) thus reduces to

$$y = x_0 + w.$$

In other words, the operator Φ is nothing more than the identity $\Phi = \mathrm{Id}$.

Deconvolution. In the case of photography, we observe a blur when the camera is not adequately stabilized (motion blur), but also a blur due to the point spread function (PSF) of the acquisition system. A reasonable approximation allows to model this degradation as a convolution operator, i.e. $\Phi x = \mathcal{K}_\Phi \star x$, where $\mathcal{K}_\Phi$ is the blurring kernel. In particular, the high frequency content of x_0 may be seriously damaged. An important property of the convolution is the fact that it is shift invariant. Estimating both x_0 and $\mathcal{K}_\Phi$, a.k.a blind-deconvolution, is a difficult problem, but we are solely here concerned with the case where $\mathcal{K}_\Phi$ is known. The deconvolution procedure is popular in many fields in science and engineering (Biemond et al. 1990), for instance in astronomy (Starck et al. 2002), in geophysics (Santosa et al. 1986) or microscopy (Agard 1984).

Inpainting. In presence of occlusion or damages pixels, the inpainting procedure aims at recovering such parts. In this case, Φ is a binary diagonal operator such that Φ_{ii} is 1 if the data are preserved, 0 otherwise. This operator can be deterministic, or the realization of a random mask. Inpainting is commonly used in many applications, such as medical fluroscopy (Chan et al. 1993), in colorization (Sapiro 2005) or in data compression (Liu et al. 2007).

Compressed Sensing. The conventional wisdom in signal processing is that for a continuous band-limited signal to be reconstructed perfectly from its equi-spaced samples, it has to be acquired at a frequency at least twice its bandwidth; this is the celebrated Shannon (1948)–Nyquist (1928) theorem. This theory however precludes many signals of interests that are not band-limited, but whose intrinsic dimension is small, think for instance of a sparse signal, or of a smooth signal away from a few singularities. The compressed sensing theory (Candès et al. 2006a; Donoho 2006) asserts that for such signals, exact and stable reconstruction is possible, hence allowing to break the Shannon-Nyquist limit. The reconstruction is moreover achieved by solving a computationally tractable convex optimization problem. The sampling operator can be modeled with a matrix Φ which is the realization of an appropriate random ensemble, such i.i.d. Gaussian or Bernoulli entries, or partial random Fourier or Hadamard matrices. The corresponding inverse problem can be shown to be efficiently regularized by some popular low complexity priors discussed in the next section. This theory has sparkled a whole research field, and hardware proofs of concept have been also developed. The first one is the single pixel camera (Wakin et al. 2006) at Rice University, which measures directly random projections on a single CCD element with a binary reflector composed of micro-mirrors. Figure 1.2 illustrates this process. Compressed sensing has been also used for Dirac train recovery in ultrasonic imagery (Tur et al. 2011). Introducing a partial randomization of the measurements (Lustig et al. 2007) is also promising in medical imaging applications such as fMRI.

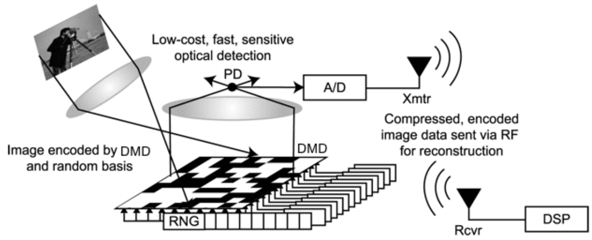


Figure 1.2: *One Pixel Camera project. Source: Rice University*

Compressed sensing is also used in astronomy (Bobin et al. 2008), in particular on the telescope Herschel.

Tomography. Tomography is commonly used in medical imaging (Newton et al. 1981). Popular CT scanners are X-ray and PET modalities. In this case, the operator Φ is a discrete Radon transform (Herman 2009), possibly with a sub-sampling to model incomplete measurements. In practice, only a few measurements can be collected, leading to an increase of the ill-posedness of the (continuous) forward operator Ψ.

1.1.2 Variational Regularizations

Solving an inverse problem from the observations (1.3) corresponds to computing an approximation of x_0 from the knowledge of y alone. This problem is said to be well posed (in the sense of Hadamard (1902)) in a space $\mathcal{S}$ if $\Phi x = y$ has x_0 as unique solution on $\mathcal{S}$, and if this solution depends continuously on y. This means that one recovers exactly $x = x_0$ when there is no noise, and a good approximation if w is small. In general, the matrix Φ is rank deficient or ill-conditionned, so that the problem is not well posed on the whole space $\mathcal{S} = \mathbb{R}^n$. In order to recover well-posedness it is thus necessary to restrict the inversion process to a well-chosen space $\mathcal{S}$ that includes x_0. A closely related procedure, that we describe next, is to set-up a variational inversion process which is penalized by a well-chosen prior.

A first line of works has considered imposing a random model on the signal x_0. This corresponds to the Bayesian formalism, see for instance the monograph (Hunt 1977) for an introduction to these methods. We do not explore these strategies in this thesis. We rather directly impose some prior on the (deterministic) x_0 through some penalty function J. This corresponds to the usual notion of variational regularization, which was initially introduced in (Tikhonov et al. 1977) as a way to recover well-posedness of the inverse problem under investigation.

Within this framework, the computation of an approximation $x^\star$ to x_0 is obtained by solving the following optimization problem

$$x^\star \in \underset{x\in\mathbb{R}^n}{\operatorname{Argmin}}\ F(x,y) + \lambda J(x). \qquad (\mathcal{P}^F_{y,\lambda})$$

Here, $F : \mathbb{R}^n \times \mathbb{R}^q \to \mathbb{R}_+$ is a data fidelity term. Typically, it is a smooth non-negative convex function. Thus, we expect that F is small when the prediction Φx is close enough from y. The factorization $F(x,y) = F_0(\Phi x, y)$ is commonly used, where $F_0 : \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R}$ is smooth, non-negative and strongly convex. This data fidelity term can be the quadratic loss, the Poisson antilog likelyhood or the logistic loss. Statistically, one may interpret $(\mathcal{P}^F_{y,\lambda})$ as a Maximum A Posteriori (MAP). This interpretation can however be misleading, as exemplified in (Gribonval 2011), where failures of the MAP approach are analyzed for sparse distributions.

The function $J : \mathbb{R}^n \to \mathbb{R}_+$ is the regularization term imposing some prior on the signal class. We assume in this thesis that J is a convex function. Convexity is important to ensure the ability to compute global optima of $(\mathcal{P}^F_{y,\lambda})$ with fast algorithms, and also enables a fine theoretical analysis of the properties of $x^\star$. It is however important to realize that non-convex penalties, as well as non-variational methods (e.g. greedy algorithms) are routinely used and often outperform their convex counterparts. This is however beyond the scope of this thesis, and we focus here on convex regularizers. Section 1.1.4 details the basic properties of these regularizers and sketch some important examples.

The scalar λ is the regularization parameter (or hyper-parameter) allowing a trade-off between fidelity and regularization. The choice of λ is an important question, which is treated in the second part of this thesis, and discussed in Section 1.3.

Since Φ is generally not injective, note that the objective function of the problem $(\mathcal{P}^F_{y,\lambda})$ is not stricly convex. Thus, it may admits several solutions. It is also possible to use the constrained version, in opposition to $(\mathcal{P}^F_{y,\lambda})$ qualified as penalized or Lagrangian form, coined as the Ivanov form in the inverse problem litterature (Ivanov et al. 1978). It reads

$$x^\star \in \operatorname*{Argmin}_{x \in \mathbb{R}^n} J(x) \quad \text{subject to} \quad F(x, y) \leqslant \varepsilon. \qquad (\bar{\mathcal{P}}^F_{y,\varepsilon})$$

We mainly focus on the Lagrangian version in this dissertation. However, problems $(\bar{\mathcal{P}}^F_{y,\varepsilon})$ and $(\mathcal{P}^F_{y,\lambda})$ are equivalent in some sense (Ivanov et al. 1978; Poljak 1987), but one may take care that the mapping between ε and λ is generally not explicit. Some recent work (Ciak et al. 2012) in this direction exploit the Fenchel–Rockafellar duality to overcome this difficulty in some particular cases.

When there is no noise, i.e. when the observations follow (1.6), we consider the constrained version of $(\mathcal{P}_{y,0})$ which reads

$$x^\star \in \operatorname*{Argmin}_{x \in \mathbb{R}^n} J(x) \quad \text{subject to} \quad \Phi x = y. \qquad (\mathcal{P}_{y,0})$$

As it will be proved formally in Chapter 5 (more precisely Proposition 5.2), problems $(\mathcal{P}_{y,\lambda})$ and $(\bar{\mathcal{P}}^F_{y,\varepsilon})$ converge (in an appropriate sense) to $(\mathcal{P}_{y,0})$.

1.1.3 Data Fidelity

The data fidelity is linked to the forward model (1.3). We find in the statistical literature several data fidelity term for the problem $(\mathcal{P}^F_{y,\lambda})$. Note that many of them does not assume that the forward model is of the form (1.3). This will discussed in details in Chapter 8. One naturally thinks to generalized linear models (GLMs) introduced by Nelder et al. (1972) which assume that conditionally on Φ, Y_i are independent with distribution that belongs to a given (one-parameter) standard exponential family. Well-known examples are Gaussian distribution (linear model), the reciprocal link (Gamma and exponential distributions), and the logit link (Bernoulli distribution, logistic regression).

When the noise is the realization of a white Gaussian noise, it is common to use the quadratic loss as a data fidelity term.

$$F_0(\mu, y) = \frac{1}{2}||\mu - y||_2^2.$$

The functional F_0 can also be chosen for instance as the logistic loss

$$F_0(\mu, y) = \sum_{i=1}^{q} \log(1 + \exp(\mu_i)) - \langle y, \mu \rangle,$$

or Huber loss (smoothed) or a ℓ^p loss. Note that ℓ^p loss is not smooth for $p < 1$. From a deterministic point of view, F_0 can be chosen from the prior on the noise in the continuous case (for instance a noise in a Banach space) or in the discrete setting considered, as a prior of ℓ^p-boundness.

In the case where F_0 is the quadratic loss, the problem $(\mathcal{P}^F_{y,\lambda})$ reads

$$x^\star \in \underset{x \in \mathbb{R}^n}{\operatorname{Argmin}} \frac{1}{2}||y - \Phi x||_2^2 + \lambda J(x). \tag{$\mathcal{P}_{y,\lambda}$}$$

This variational formulation is at the core of the first part of this thesis.

1.1.4 Low Complexity Priors

1.1.4.1 Combinatorial Model Selections

Penalizing in accordance to some notion of complexity is a key idea, whose roots can be traced back for instance to the statistics literature (Mallows 1973; Akaike 1973). This complexity is measured using a functional **pen**(T) where T is some linear subspace containing x, and chosen among a fixed collection of spaces $\mathcal{T}$. This approach typically makes use of hierarchy of models of increasing complexity, which should be designed in accordance to some prior knowledge about the data x_0 to recover. A union of linear models is a collection $\mathcal{T}$ of subspaces of $\mathbb{R}^n$ which is usually finite but very large, in the case of finite dimensional problems. These subspaces typically account for some kind of smoothness or simplicity of the signal. A key example is sparsity, which, in its simplest form, corresponds to a problem of selecting few of active variables in the data. In this setting, a subspace T has the form $T = \{x \mid \operatorname{supp}(x) = I\}$ for some set of indexes I indicating the active variables. With such a set of model at hand, one can use the following prior

$$J(x) = \inf_{x \in T} \mathbf{pen}(T). \tag{1.7}$$

The problem $(\mathcal{P}_{y,\lambda})$ can be recast as a model selection problem

$$\inf_{T \in \mathcal{T}, x \in T} ||y - \Phi x||^2 + \lambda \mathbf{pen}(T).$$

The model selection literature (Birgé et al. 1997; Barron et al. 1999; Birgé et al. 2007) proposes many significant results to quantify the performance of these approaches. A major bottleneck of this class of approaches is that the corresponding J function defined in (1.7) is non-convex, thus typically leading to intractable, often NP-hard problems. For instance, the sparsity of coefficients $x \in \mathbb{R}^n$ is measured using the ℓ^0 pseudo-norm

$$J_0(x) = ||x||_0 = |\operatorname{supp}(x)|.$$

Minimizing $(\mathcal{P}_{y,\lambda})$ or $(\mathcal{P}_{y,0})$ with $J = J_0$ is known to be NP-hard, see for instance (Natarajan 1995). There is a wide variety of approaches to tackles directly non-convex optimization problems. A line of research considers greedy algorithms. The most popular ones are Matching Pursuit (Mallat et al. 1993) and Orthogonal Matching Pursuit (Pati et al. 1993; Davis et al. 1994), see also the comprehensive reviews Needell et al. (2008) and references therein. Another line of research, which is the one under study in this thesis, consists in considering convexified versions of (1.7).

1.1.4.2 Convex Encoding of Models

For any subspace T of a real vector space E, we denote P_T the orthogonal vector on T, $x_T = P_T(x)$ and $\Phi_T = \Phi P_T$. We now introduce the model tangent subspace at a point x for some finite-valued convex functional J.

Definition 1.1 — Model Tangent Subspace For any vector $x \in \mathbb{R}^N$, we denote e_x its *model vector*,

$$e_x = \underset{e \in \operatorname{aff} \partial J(x)}{\operatorname{argmin}} \ \|e\|,$$

where $\operatorname{aff} \partial J(x)$ is the affine hull of the subdifferential (see Definition 2.12) of J at x, and

$$T_x = \operatorname{span}(\partial J(x))^{\perp}.$$

T_x is coined the *model tangent subspace* of x associated to J.

This terminology will be clear after we define partly smooth function in Section 1.2.2. When J is Gâteaux-differentiable at x, i.e. $\partial J(x) = \{\nabla J(x)\}$, $e_x = \nabla J(x)$ and $T_x = \mathbb{R}^N$. On the contrary, when J is not smooth at x, the dimension of T_x is of smaller dimension, and the regularizing functional J essentially promotes elements living on or close to the affine space $x + T_x$. Table 1.1 exemplifies Definition 1.1 on several regularizers that are popular in the literature. The details of the exact derivations is provided in Chapter 3.

J	T_x	e_x	Comment
$\|\cdot\|_1$	$\{\eta \mid \forall j \notin I, \eta_j = 0\}$	$\operatorname{sign}(x)$	$I = \operatorname{supp}(x)$
$\|D^* \cdot\|_1$	$\operatorname{Ker}(D^*_{I^c})$	$P_{\operatorname{Ker}(D^*_{I^c})} \operatorname{sign}(D^* x)$	$I = \operatorname{supp}(D^* x)$
$\|\cdot\|_{1,2}$	$\{\eta \mid \forall j \notin I, \eta_j = 0\}$	$(\mathcal{N}(x_b))_{b \in \mathcal{B}}$	$I = \{g \in \mathcal{B} \mid x_g \neq 0\}$
$\|\cdot\|_*$	$\{Z \mid U_\perp^* Z V_\perp = 0\}$	UV^*	$x = U \Lambda V^*$
$\|\cdot\|_\infty$	$\{\alpha \mid \alpha_I = \rho s_I \text{ for } \rho \in \mathbb{R}\}$	$\operatorname{sign}(x)/\|I\|$	$I = \{i \mid \|x_i\| = \|x\|_\infty\}$

Table 1.1: *Examples of Model Tangent Subspace. The notations are precised in the following sections.*

1.1.4.3 Sparsity

A dictionary $D = (d_i)_{i=1}^p$ is a (possibly redundant, i.e. when $p > n$) collection of p atoms $d_i \in \mathbb{R}^n$. It can also be viewed as a linear mapping from $\mathbb{R}^p$ to $\mathbb{R}^n$

which is used to synthesize a signal $x \in \text{Im}(D) \subseteq \mathbb{R}^n$ as

$$x = D\alpha = \sum_{i=1}^{p} \alpha_i d_i,$$

where α is the coefficient vector that synthesizes x from the dictionary D. Note that if D is redundant, there is an infinite number of coefficients α such that $x = D\alpha$. An issue beyond our work is to build a good dictionary. We may cite the wavelet transform (Mallat 1989) and the curvelet transform (Candès et al. 2000) for images that are piecewise smooth away from smooth edge curves, local Fourier basis for sounds (Allen 1977), or union of dictionaries for image and signal decomposition, see for instance cartoon+texture decomposion in (Elad et al. 2005).

Synthesis sparsity. When considering sparsity in the canonical basis, i.e. $D = \text{Id}$, the model subspace and model vector read

$$T_x = \left\{ x' \mid \text{supp}(x') = \text{supp}(x) \right\} \quad \text{and} \quad e_x = \text{sign}(x).$$

Looking for the sparsest representation of x in the dictionary D amounts to solving

$$\min_{\alpha \in \mathbb{R}^p} ||\alpha||_0 \quad \text{subject to} \quad x = D\alpha.$$

Replacing the ℓ^0 norm by the ℓ^1 norm leads to a convex problem. The sparsest set of coefficients, according to the ℓ^1 norm, defines a signal prior which is the image of $||.||_1$ under D,

$$J_S(x) = \min_{\alpha \in \mathbb{R}^p} ||\alpha||_1 \quad \text{subject to} \quad x = D\alpha.$$

Therefore any solution x of $(\mathcal{P}_{y,\lambda})$ using $J = J_S$ can be written as $x = D\alpha$ where α is a solution of

$$\min_{\alpha \in \mathbb{R}^p} \frac{1}{2}||y - \Phi D\alpha||_2^2 + \lambda||\alpha||_1. \tag{1.8}$$

Note that $\|\cdot\|_1$ is the convexification of $\|\cdot\|_0$ restricted to the ℓ^∞-ball. Problem (1.8) was introduced in the statistical community in (Tibshirani 1996) where it was coined Lasso. Note that it was originally introduced as an ℓ^1-ball constrained optimization and in the over-determined case. It is also known in the signal processing community as Basis Pursuit DeNoising (Chen et al. 1999). Such a problem corresponds to the so-called sparse synthesis regularization, as sparsity is assumed on the coefficients α that synthesize the signal $x = D\alpha$. In the noiseless case, the constrained problem $(\mathcal{P}_{y,0})$ becomes

$$\min_{\alpha \in \mathbb{R}^p} \|\alpha\|_1 \quad \text{subject to} \quad y = \Phi D \alpha, \tag{1.9}$$

which goes by the name of Basis Pursuit after (Chen et al. 1999).

Sparse regularization is a popular class of priors to model natural signals and images, see for instance (Mallat 2009). The idea of ℓ^1 regularization finds its root in the seismic imaging literature (Santosa et al. 1986) for deconvolution. It is also used in many applications, see (Starck et al. 2010) for a comprehensive account.

A key problem of active research is to learn and optimize the dictionary in order to represent optimally a set of given exemplar. We refer to the book of (Elad 2010, Chapter 12) for a recent overview of the relevant literature.

Analysis sparsity. Analysis regularization corresponds to using $J = J_A$ in $(\mathcal{P}_{y,\lambda})$ where

$$J_A(x) = \|D^* x\|_1 = \sum_{i=1}^{p} |\langle d_i, x \rangle|,$$

It imposes the sparsity of the correlations $(\langle d_j, x \rangle)_{j=1,\ldots,p}$ between x and the atoms in a dictionary D. In this case,

$$T_x = \left\{ x' \mid \operatorname{supp}(D^* x') = \operatorname{supp}(D^* x) \right\} \quad \text{and} \quad e_x = \operatorname{sign}(D^* x).$$

Note that synthesis and analysis regularizations are different as soon as D is not an invertible square matrix. Hence, $(\mathcal{P}_{y,\lambda})$ reads

$$\min_{x \in \mathbb{R}^n} \frac{1}{2}||y - \Phi x||^2 + \lambda ||D^* x||_1. \tag{1.10}$$

In the noiseless case, the ℓ^1-analysis equality-constrained problem is

$$\min_{x \in \mathbb{R}^n} ||D^* x||_1 \quad \text{subject to} \quad \Phi x = y. \tag{1.11}$$

In (Nam et al. 2013), the term *cosparse* is used, motivated by the role played by the complement of the support (i.e. cosupport) of the vector $D^* x$ in the theoretical analysis of (1.11).

The adjoint of any synthesis dictionary (see above) can be used to define analysis sparsity prior. Analysis sparsity allows for more intricate operators D^* because D^* is not required to be a stable frame of the signal space. One of the most popular is the finite difference operator used in the total variation seminorm, first introduced for denoising (in a continuous setting) by Rudin et al. (1992). Typically, for 1-D discrete signals, D can be taken as a dictionary of forward finite differences D_{DIF} where

$$D_{\text{DIF}} = \begin{pmatrix} +1 & & & 0 \\ -1 & +1 & & \\ & -1 & \ddots & \\ & & \ddots & +1 \\ 0 & & & -1 \end{pmatrix}. \tag{1.12}$$

The corresponding prior J_A favors piecewise constant signals and images. A comprehensive review of total variation regularization can be found in (Chambolle et al. 2010). One can also use a wavelet dictionary D which is shift-invariant, such that the corresponding regularization J_A can be see as a multi-scale total variation in the case of the Haar wavelet (Steidl et al. 2004) for 1D signals. When using wavelets with m vanishing moment, the corresponding priors favors discrete piecewise polynomial signals of degree m. A numerical exploration of the relative performances of analysis and synthesis regularization is performed in (Elad et al. 2007). Selesnick et al. (2009) report an extensive numerical exploration where they use shift invariant wavelet dictionaries to compare analysis and synthesis sparsity priors for several inverse problems (e.g. deconvolution and inpainting). As a last example of sparse analysis regularization, we would like to mention the Fused Lasso (Tibshirani et al. 2005), where D is the concatenation of D_{DIF} and a weighted identity. The corresponding prior J_A promotes both sparsity of the signal and its derivative, hence favoring the grouping of non-zero coefficients in blocks over which the signal is constant.

Structured sparsity. To further improve the performance of sparse regularization, it is useful to group the coefficients, imposing the sparsity in a block-wise manner. It has been first proposed by Hall et al. (1997, 1999); Cai (1999) for wavelet block shrinkage. For over-determined regression of the form (1.3), it has been introduced by Bakin (1999); Yuan et al. (2005). Block regularization is popular in image processing because wavelet coefficients of a natural image have a group structure (Mallat 2009). Indeed, edges and textures induce strong local dependencies between coefficients. In multi-task learning (Obozinski et al. 2010), it is used to control the sparsity pattern of the covariates. In audio processing, it is also useful to deal with multi-channel data as studied by Gribonval, Rauhut, et al. (2008) which is also known as the multiple measurements vector (MMV) model, see for instance (Cotter et al. 2005; Chen et al. 2006).

Suppose what we split the signal space $\mathbb{R}^n$ into groups without overlapping. We formalize this splitting by a disjoint partition $\mathcal{B}$ of $\{1, \ldots, n\}$, i.e.

$$\bigcup_{b \in \mathcal{B}} b = \{1, \ldots, n\} \quad \text{and} \quad \forall b, b' \in \mathcal{B}, \, b \cap b' = \emptyset.$$

Then, we define the $\ell^1 - \ell^2$ norm as $J = J_{\mathcal{B}}$ where

$$J_{\mathcal{B}}(x) = \sum_{b \in \mathcal{B}} ||x_b||, \tag{1.13}$$

where x_b is a vector of size $|b|$ containing the entries indexed by b. Thus, the model space and model vector reads

$$T_x = \left\{ x' \mid \forall b \in \mathcal{B}, x_b = 0 \Rightarrow x'_b = 0 \right\} \quad \text{and} \quad e_x = \left(\frac{x_b}{||x_b||} \right)_{b \in \mathcal{B}},$$

where we take the convention that if $x_b = 0$ then $\frac{x_b}{||x_b||} = 0$. It is possible to replace the ℓ^2 norm with more general functionals, such as ℓ^p norms for $p > 1$ (Turlach et al. 2005; Negahban et al. 2011; Vogt et al. 2012) or to use analysis block sparsity

$$J_{\mathcal{B}}(x) = \sum_{b \in \mathcal{B}} ||D_b^* x_b||,$$

where D_b^* are linear operators from $\mathbb{R}^{|b|} \to \mathbb{R}^p$. For instance, one can express the 2D isotropic total variation by defining $D_b^* x \in \mathbb{R}^2$ to be an approximation by finite differences of the gradient of the image x at the pixel indexed by b. This block analysis sparsity allows us also to take into account overlapping groups (Jenatton et al. 2011; Cai et al. 2001), or groups structured in a tree (Peyré et al. 2011; Zhao et al. 2009).

1.1.4.4 Beyond Sparsity

While sparsity has become mainstream in imaging sciences and machine learning, there is now a flurry of activity to develop novel priors to take into account various types of low-dimensional structures to model the data.

Low rank prior. The natural extension of sparsity to matrices $x \in \mathbb{R}^{n_1 \times n_2}$, where $n = n_1 n_2$, is to impose a low rank constraint. This should be understood as imposing the sparsity of the singular values. Denoting $x = V_x \operatorname{diag}(\Lambda_x) U_x^*$ a Singular Value Decomposition of x, where $\Lambda_x \in \mathbb{R}^m$ and $m = \min(n_1, n_2)$. Hence the rank reads $\operatorname{rank}(x) = ||\Lambda_x||_0$. Here, the natural models are not linear subspaces T_x but manifolds of matrices with a fixed rank, see Chapter 4. The nuclear norm (or trace, 1-Schatten norm) imposes such a sparsity (Fazel 2002) and is defined as

$$||x||_* = ||\Lambda_x||_1.$$

The nuclear norm is the convexification of the rank function with respect to the spectral norm ball, see (Fazel 2002; Hiriart-Urruty et al. 2012). It has been used for instance to recover low rank matrices by Srebro (2004) (Netflix prize) or for model reduction in (Fazel et al. 2001).

Spread representation. In some cases, one expects to recover flat vectors, i.e such that for most i, $x_i = ||x||_\infty$. A convex function promoting such behavior is the ℓ^∞ norm defined as

$$||x||_\infty = \max_{i \in \{1,\dots,n\}} |x_i|.$$

Such a prior is encoded in a linear model T which is defined w.r.t to the number of saturating coordinates. More precisely,

$$T_x = \left\{ x' \mid x'_I = \rho x_I \text{ for some } \rho \in \mathbb{R} \right\},$$

where $I = \{i \mid x_i = ||x||_\infty\}$. For applications in computer vision such as image retrieval in a database (Jégou et al. 2010), it is useful to have a compact signature of signals, typically with only two values ± 1. An approach proposed in (Jégou et al. 2012) for obtaining this binary quantification is to compute these vectors as spread approximations in a random dictionary. A study of this regularization is done in (Fuchs 2011), where an homotopy-like algorithm is provided. Moreover, the use of ℓ^∞ regularization is connected to Kashin's representation (Lyubarskii et al. 2010), which is known to be useful in stabilizing the quantization error for instance. Others applications such as wireless network optimization (Studer et al. 2012) also rely on ℓ^∞ prior.

1.1.4.5 From Continuous to Discrete, and Vice Versa.

Even if we focus on finite dimensional problems, an important issue is to understand the link between these models with their continuous counterparts. The underlying mathematical problems, for instance the convergence of discrete models to the continuous ones, have practical implications in order to understand the fine structure of signals computed with these methods, which are generally easier to describe in a continuous setting (for instance edges in images). Two typical questions naturally arise:

(i) How to analyze the convergence of the function $Bx^\star$ to f_0 when n tends to $+\infty$, where $x^\star$ is some solution of $(\mathcal{P}_{y,\lambda})$ and B is the basis defined in (1.4) ? It is often treated by a control on both the estimation error $||x^\star - x_0||$ and the approximation error $||f_0 - Bx_0||$.

(ii) How to define a variational problem directly in the continuous setting ? It corresponds to replacing the function J par a function $f \mapsto \mathcal{J}(f)$ such that $\mathcal{J}(Bx^\star)$ is "close enough" to $J(x^\star)$. In this case, we deal with the following optimization problem

$$f^\star \in \underset{f \in \mathbb{H}}{\operatorname{argmin}} \ \frac{1}{2}||y - \Psi f||^2 + \lambda \mathcal{J}(f). \qquad (c\mathcal{P}_{y,\lambda})$$

Note that the choice of $\mathbb{H}$ should be chosen in accordance to the functional $\mathcal{J}$.

Wavelet sparsity and Besov spaces. Let $W(f) \in \mathbb{R}^{\mathbb{N}}$ be the wavelet transform of $f \in \mathbb{H}$ in an orthogonal basis introduced by Mallat (1989); Daubechies (1992); Meyer (1992). Besov Banach spaces are defined using appropriate sparsity inducing norms of $W(f)$. Besov spaces form an important class of Banach spaces since they are a powerful tool to model piecewise regular signals and image with pointwise singularities, see for instance (Meyer 1992; Chambolle et al. 1998). Wavelets are known to provide optimally sparse representations for functions in Besov spaces (Mallat 2009).

These spaces have been widely used in the statistical community to establish minimaxity of wavelet-based estimators for several problems (e.g. regression, inverse problems, etc.), for instance to quantify the optimality of the soft thresholding for denoising (Donoho et al. 1994) or for the wavelet-vaguelettes (Donoho 1995) method which corresponds to applying the soft thresholding to $\Psi^+(y)$. These results have been extended to group sparsity (Hall et al. 1998), for instance in the denoising setting (Chesneau et al. 2010a) or deblurring (Chesneau et al. 2010b).

Sparsity and Radon measures. When we wish to recover highly localized signals, a convenient model is to use sum of Dirac distributions. The finite dimensional problem ($\mathcal{P}_{y,\lambda}$), with $J = \|\cdot\|_1$ and $\Phi = \mathrm{Id}$, is equivalent to considering that these distributions are on a fixed grid whereas ($c\mathcal{P}_{y,\lambda}$) can be seen as its continuous (grid-free) counterpart. In this case, $\mathbb{H}$ is the space of finite Radon measures and $\mathcal{J}(f)$ is the total variation of the measure f which should not be confused with the total variation of a function.

Solving inverse problems on this space of measures has been recently considered by Bredies et al. (2013) and a theoretical study of the performance is proposed by Candès and Fernandez-Granda (2013) for the case of deconvolution (super-resolution). The convergence of the solutions of ($\mathcal{P}_{y,\lambda}$) to those of ($c\mathcal{P}_{y,\lambda}$) is studied in (Duval et al. 2013; Tang et al. 2012).

Sparse gradient and bounded variation functions. Analysis sparsity of the gradient (1.12) can be seen as a discretization by finite difference of the total variation of a function. More precisely, for $f \in L^1_{\mathrm{loc}}(\Omega)$, we denote the total variation $\mathcal{J}$ as

$$\mathcal{J}(f) = \sup \left\{ -\int_\Omega f \operatorname{div} \psi \mid \psi \in C_c^\infty(\Omega, \mathbb{R}^n), \forall x \in \Omega, |\psi(x)| \leqslant 1 \right\}. \qquad (1.14)$$

f has bounded variation if $\mathcal{J}(f) < +\infty$ and we denote $\mathrm{BV}(\Omega)$ the Banach spaces of functions of bounded variations endowed with the norm $\|\cdot\|_{L^1(\Omega)} + \mathcal{J}(\cdot)$. Remark that $W^{1,1}(\Omega)$ is strictly included in $\mathrm{BV}(\Omega)$. In fact, if f is C^1 then $\mathcal{J}(f) = \|\nabla f\|_{L^1(\Omega)}$.

A useful property of this space with respect to any Sobolev space is the fact that the problem ($c\mathcal{P}_{y,\lambda}$) can admit non-continuous solutions when using $\mathbb{H} = \mathrm{BV}(\Omega)$. The denoising problem has been studied in (Caselles et al. 2007) where the discontinuity set of its solution is characterized. We refer to (Ambrosio et al. 2000) for a detailed study of this space. Note that higher order priors have been introduced recently, e.g the total generalized variation (Bredies et al. 2010).

1.1.5 Solving the Optimization Problem

Several algorithms exist in order to solve the problem $(\mathcal{P}_{y,\lambda})$ or $(\mathcal{P}_{y,0})$, depending on the nature of J.

Solving $(\mathcal{P}_{y,\lambda})$ corresponds to the minimization of a convex function f. When J is smooth, one can make use of traditional gradient or Newton descent schemes. However, the class of low-complexity regularizations J considered in this thesis are highly non-smooth. It is possible to adapt the gradient descent scheme when f is convex, lower semi-continuous and proper by replacing the descent direction $\operatorname{grad} f$ by any element of the subdifferential $\partial f(x)$. This scheme is however quite inefficient for the penalties J considered in this thesis, which are highly structured. Making use of this structure is crucial to obtain fast algorithms.

For a large class of J regularizers, such as those introduced in this section (ℓ^1, nuclear norm, total variation, etc), the optimization $(\mathcal{P}_{y,\lambda})$ can be shown to be equivalent to a conic program. This cone constraint can be enforced using a self-concordant barrier function, and the optimization problem can hence be solved using interior point methods, as pioneered by Nesterov et al. (1994), see also the monograph (Boyd et al. 2004). This class of methods enjoys fast convergence rate. Each iteration however is typically quite costly. This class of solvers is a wise choice for problem of medium size, and when high accuracy is needed.

Homotopy methods have been introduced in the case of the sparsity $J = \|\cdot\|_1$ by Osborne et al. (2000), then adapted to the analysis sparsity $J = \|D^* \cdot\|_1$ in (Tibshirani et al. 2011) and spread representations $\|\cdot\|_\infty$ in (Fuchs 2011). The LARS algorithm (Efron et al. 2004) is closely related and compute an approximation of the homotopy path with a faster algorithm. These methods rely on the behavior of $\lambda \mapsto x^\star(\lambda)$, where $x^\star(\lambda)$ is a solution of $(\mathcal{P}_{y,\lambda})$. In the case of a polyhedral regularization, such as ℓ^1 or ℓ^∞, this path turns out to be piecewise polygonal, see Chapter 8.

The cost per iteration of both interior point and homotopy methods scales badly with the dimension, thus preventing them to be used in large scale problems such as those encountered in imaging science. Proximal schemes are attractive alternatives, since they correspond to first order schemes whose iterations are in practice quite cheap. We refer to (Beck et al. 2009; Bauschke et al. 2011; Combettes et al. 2011; Parikh et al. 2013) for comprehensive reviews. Their slow convergence rate is thus generally not a big issue in imaging or machine learning, where one typically does not seek for a high precision solution to the optimization problem ($\mathcal{P}_{y,\lambda}$).

1.2 Robustness: Handling the Impact of Noise

Observations are in general contaminated by noise. It is thus important to study the robustness of ($\mathcal{P}_{y,\lambda}$) to analyze its performance. More precisely, we aim to derive criteria quantifiying how $x^\star$ is close to to x_0. This notion of closeness will be analyzed mathematically through two quality criteria: error distance in the sense of the ℓ^2 norm and model selection.

1.2.1 Linear Convergence Rate

Here, we are seeking sufficient conditions under which any solution of ($\mathcal{P}_{y,\lambda}$) satisfies

$$||x^\star - x_0|| = O(||w||).$$

It depends typically on x_0, while λ should be chosen proportionally to the noise level for the linear convergence to hold. The terminology "linear" in the convergence rate, which stems from the inverse problem community, pertains to the fact that the error is within a factor of the noise level. This rate is made possible by the fact that J is not smooth. For instance, this is not the case for $|| \cdot ||_2^2$, see (Scherzer 2009).

1.2.1.1 Dual Certificate and Non-Degeneracy

We introduce the notion of dual certificate which characterizes the solutions of the noiseless problem ($\mathcal{P}_0(\Phi x_0)$). This notion is a key ingredient of our analysis in the sequel.

Definition 1.2 — Dual Certificates. A *(dual) certificate* for $x \in \mathbb{R}^n$ is a vector $\mathfrak{p} \in \mathbb{R}^q$ such that the *source condition* is verified:

$$\Phi^* \mathfrak{p} \in \partial J(x). \tag{SC_x}$$

If $\mathfrak{p}$ is a certificate, and moreover

$$\Phi^* \mathfrak{p} \in \operatorname{ri} \partial J(x), \tag{$\overline{SC}_x$}$$

we say that $\mathfrak{p}$ is a *non-degenerate certificate*, where ri denotes the relative interior.

A subspace $T \subseteq \mathbb{R}^n$ satisfies the *restricted injectivity condition* (INJ_T) if Φ is injective on T.

In practice, it might be difficult to find such a non-degenerate certificate. A popular strategy in the literature is to single out a particular certificate (that we coined minimal norm) which in some cases can be actually computed in closed form. The *minimal norm certificate* $\mathfrak{p}_0$ for $x \in \mathbb{R}^n$ is defined by

$$\mathfrak{p}_0 = \underset{\mathfrak{p} \in \mathbb{R}^q}{\operatorname{argmin}} \, ||\mathfrak{p}|| \quad \text{subject to} \quad \Phi^* \mathfrak{p} \in \partial J(x).$$

We define also the *linearized precertificate* $\mathfrak{p}_F$ as

$$\mathfrak{p}_F = \underset{\mathfrak{p} \in \mathbb{R}^q}{\operatorname{argmin}} \, ||\mathfrak{p}|| \quad \text{subject to} \quad (\Phi^* \mathfrak{p})_{T_x} = e_x.$$

Now, suppose that (INJ_{T_x}) is satisfied. In this case, $\mathfrak{p}_F = \Phi_{T_x}^{+,*} e_x$, see Lemma 5.5. Then $\Phi^* \mathfrak{p}_F \in \operatorname{ri} \partial J(x)$ or $\Phi^* \mathfrak{p}_0 \in \operatorname{ri} \partial J(x)$ implies that $\mathfrak{p}_F = \mathfrak{p}_0$. Thus the linearized precertificate is the minimal norm certificate if it is indeed a non-degenerate certificate, see Chapter 5 for a precise statement. This is important since $\mathfrak{p}_F$ is simple enough to be computed and analyzed mathematically, leading to an easy way to check if $\mathfrak{p}_0$ is a non-degenerate certificate. Another crucial point is that $\mathfrak{p}_0$ is the certificate that drives the robustness of the model, as detailed in Section 1.2.2.

1.2.1.2 Main Contribution

We prove the following Theorem which establishes a linear convergence rate for any closed convex function, without particular assumption on it, except the fact that it is finite-valued, hence continuous.

Theorem 1 Let $T_0 = T_{x_0}$. Suppose that $(\overline{\mathrm{SC}}_{x_0})$ is verified for $\Phi^* p \in \operatorname{ri} \partial J(x_0)$ and (INJ_{T_0}) holds. If $\lambda = c\varepsilon$, $c > 0$, then for every minimizer $x^\star$ of $(\mathcal{P}_{y,\lambda})$

$$||x^\star - x_0||_2 \leqslant C\varepsilon \ ,$$

where

$$C = C_1\ (2 + c||p||_2) + C_2 \frac{(1 + c||p||_2/2)^2}{cC_p},$$

$C_1 > 0$ and $C_2 > 0$ are two constants independent of p and $0 < C_p < 1$.

This theorem is proved in Chapter 6. This result holds for any finite-valued convex function and holds for any minimizer of $(\mathcal{P}_{y,\lambda})$ (not necessarily unique). However, remark that (INJ_{T_0}) makes sense only if J promotes subspace of low dimension. Note that finding a certificate p is not trivial, and that the constant involved in Theorem 1 depends on it. This leaves a degree of freedom to optimize the constant for the certificate. The closer to 1 the constant C_p is, the better is the robustness. It measures how far from the relative boundary is p. Finally, the constants C_1 and C_2 are not absolute and may depend on the dimension. Hence, this theorem does not extend straightforwardly to the infinite-dimensional problem $(c\mathcal{P}_{y,\lambda})$.

1.2.1.3 Relations to Previous Works

Convergence rates. The monograph (Scherzer 2009) is dedicated to regularization properties of inverse problems in infinite-dimensional Hilbert and Banach spaces with application to imaging. In particular, Chapter 3 of this book treats the case where J is a coercive gauge for the problem $(\mathcal{P}_{y,\lambda})$. In (Burger et al. 2004), the authors consider the case where $\mathcal{J}$ is a proper, convex and l.s.c functional for both the constrained and Lagrangian regularization $(c\mathcal{P}_{y,\lambda})$. Under the source condition and a restricted injectivity assumption, they bound the error in Bregman divergence with a linear rate $O(\|w\|)$. For the classical Thikonov regularization, i.e. $\mathcal{J} = \|\cdot\|_{L^2(\Omega)}$, the estimation is in $O(\sqrt{\|w\|})$, which is not a linear convergence. Extensions of these results have been proved in (Resmerita 2005) and (Hofmann et al. 2007) for the Bregman rate.

Lorenz (2008) treats the case where J is a ℓ^p norm with $1 \leqslant p \leqslant 2$ and provides a prediction error $\Phi x_0 - \Phi x^\star$ in $O(\|w\|)$ and an estimation error $x^\star - x_0$ in $O(\sqrt{\|w\|})$. Grasmair et al. (2011) is concerned with the special case of ℓ^1 regularization, and draws some connection with the restricted isometry property (RIP), see below. The result which is the closed to our appears in (Grasmair 2011). Here, $\mathcal{J}$ is a proper, convex, l.s.c and positively homogeneous functional on some Banach space $\mathbb{H}$. Under a source condition and restricted injectivity on a an appropriate cone, a linear convergence rate is proved with respect to $\mathcal{J}$, i.e.

$$J(x^\star - x_0) = O(\|w\|).$$

This result implies ours, but only if J is injective which precludes many important regularizers, e.g. TV.

Compressed sensing. In a compressed sensing setting, for instance when Φ is drawn from a i.i.d. normal distribution, it was proved (Rudelson et al. 2008) that if the number of measurements q is such that $q \gtrsim \log(n/k)$ where $k = \|x_0\|_0$ then there exists with high probability on Φ a non-degenerate certificate when $J = \|\cdot\|_1$, i.e. $(\overline{SC}_x)$ holds and one can apply the result of Theorem 1.

The performance of compressed sensing recovery has initially been analyzed using the so-called restricted isometry property (RIP) introduced in (Candès et al. 2006a, 2006b; Candès and Tao 2006) for ℓ^1. It is defined for a couple (Φ, k) where k is a targeted sparsity, as the smallest constant δ_k such that

$$(1-\delta_k)\|x\|^2 \leqslant \|\Phi x\|^2 \leqslant (1+\delta_k)\|x\|^2, \qquad (1.15)$$

for any vector x such that $\|x\|_0 \leqslant k$. It is shown (Candès et al. 2006a) that if $\delta_{2k} + \delta_{3k} < 1$, then for *every* vector x_0 of sparsity k, there exists a non-degenerate certificate (Candès et al. 2005, Lemma 2.2) as remarked also by Grasmair et al. (2011). This result thus implies linear convergence rate, and is applied in (Candès et al. 2006b) to show the robustness to noise of compressed sensing. This was generalized to analysis sparsity (i.e. $J = \|D^* \cdot\|_1$ with D tight frame) in (Candès, Eldar, et al. 2011), structured sparsity in (Candès, Eldar, et al. 2011) and matrix completion (Recht et al. 2010; Candès and Plan 2011b) using $J = \|\cdot\|_*$. A major shortcoming of this approach is that available designs of matrices satisfying (1.15) for reasonnably large value of k are essentially random. Indeed, in this case, the constant δ_k can be shown to be small enough with high probability on Φ for a nearly optimal scaling of (n, q, k). For instance, when Φ is drawn for the Gaussian ensemble, it is the case when $q \succcurlyeq k \log(n/k)$. as proved by Candès and Tao (2006) Note that in general, computing the RIP constants for a given matrix is an NP-hard problem (Bandeira et al. 2013).

The golfing scheme introduced by Gross (2011) for the nuclear norm allows to consider non-Gaussian distributions, e.g. partial Fourier measurements. It is based on an iterative scheme starting from the linearized precertificate $\mathfrak{p}_F$ in order to construct an (approximate) certificate with high probability on the matrix for a given vector. It was further studied by Candès and Plan (2011a) for ℓ^1 regularization and Koltchinskii et al. (2011).

1.2.2 Model Selection

So far, we were concerned with ℓ^2-stability/robustness. What can be said about the recovery of the model $T_0 = T_{x_0}$ underlying the original vector itself x_0 ? To be able to state such a result, the regularization has to enjoy some additional structure. This is the goal of partial smoothness that we introduce formally hereafter.

1.2.2.1 Partly Smooth Functions

The notion of partial smoothness (Lewis 2002) unifies many notions of structured non-smooth functions known in the literature The 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". 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 behavior 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 2002; Lewis et al. 2013).

Definition 1.3 — Partly Smooth Function. A finite-valued convex function $J \in \Gamma_c^+(\mathbb{R}^n)$ is said to be *partly smooth* (PSF) at x relative to a set $\mathcal{M} \subseteq \mathbb{R}^n$ if there exists a neighborhood U of x such that

(i) $\mathcal{M} \cap U$ is a C^2-manifold and J restricted to $\mathcal{M}$ is C^2 around x,

(ii) T_x is the tangent plane of $\mathcal{M}$ at x, i.e. $\mathcal{T}_{\mathcal{M}}(x) = T_x$,

(iii) the set-valued mapping ∂J is continuous at x relative to $\mathcal{M}$.

The manifold $\mathcal{M}$ is coined the *model manifold* of $x \in \mathbb{R}^n$. 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. Hence the notation $\mathcal{M} = \mathcal{M}_x$ is unambiguous (locally).

ℓ^1, ℓ^1-ℓ^2 and nuclear norms are partly smooth, where the first two ones are such that $\mathcal{M} = T_x$. This special class of partly smooth functions is dubbed partly smooth with linear manifold functions. Moreover, if $J = J_0 \circ D^*$ and J_0 is partly smooth at $z = D^* x$ for the manifold $\mathcal{M}_z^0$, then it is shown in (Lewis 2002, Theorem 4.2) that J is partly smooth at x for

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

A similar result is also proved for the sum of two partly smooth functions. We detail these results in Section 4.1.

1.2.2.2 Main Contribution

We prove the following theorem.

Theorem 2 Let J a locally partly smooth function at x_0 relative to $\mathcal{M}$. Assume that (INJ_T) holds and $\Phi^* \mathfrak{p}_F \in \operatorname{ri} \partial J(x_0)$. Then there exist positive constants C, C' such that if w and λ obey

$$\|w\| \leqslant C \quad \text{and} \quad \lambda = C' \|w\|, \tag{1.16}$$

the solution $x^\star$ of $(\mathcal{P}_{y,\lambda})$ with noisy measurements y is unique, and satisfies

$$x^\star \in \mathcal{M} \quad \text{and} \quad \|x_0 - x^\star\| = O(\|w\|).$$

This theorem is proved in Chapter 7. Obviously, the assumptions of Theorem 2 imply the conclusion of Theorem 1. Contrary to this same result, this theorem is based on an explicit formulation of the precertificate $\mathfrak{p}_F$, which makes it directly effective. Note that there exist vectors which can be stably recovered in the ℓ^2 sense of Theorem 1, but whose underlying manifold model cannot be stably identified in the sense of Theorem 2, see our numerical experiments in Chapter 10. When J is partly smooth with linear manifold ($\mathcal{M} = T_x$), i.e. the manifold is in fact the model subspace, a more precise statement of Theorem 2 is given with the explicit derivation of the constants C, C' and the one involved in the $O(.)$ term, see Chapter 7.

1.2.2.3 Relation to Previous Works

Special cases. Theorem 2 is a generalization of many previous works that have appeared in the literature. For the ℓ^1 norm, $J = \|\cdot\|_1$, to the best of our knowledge, this result was initially stated by Fuchs (2004). In this setting, the result $x^\star \in \mathcal{M}$ corresponds to the correct identification of the support, i.e. $\mathrm{supp}(x^\star) = \mathrm{supp}(x_0)$. Moving to a setting where both Φ and w are random, the condition $\mathfrak{p}_F \in \mathrm{ri}\,\partial J(x_0)$ implies model consistency (also known as sparsistency for ℓ^1), i.e. the probability that the support is correctly identified tends to 0 when the dimensions of the problem increases. Bach proves respectively in (Bach 2008a) and (Bach 2008b) Theorem 2 (in fact a variant since he considers randomized Φ and w) for $\ell^1 - \ell^2$ and nuclear norm gauges, in the special case where Φ has full rank (i.e. is injective). Our results thus shows that the same condition ensure rank consistency with the additional constraint that $\mathrm{Ker}(\Phi) \cap T = \{0\}$. Theorem 2 for a ℓ^1 analysis prior was proved by Vaiter, Peyré, et al. (2013). A similar result was shown in (Duval et al. 2013) for an infinite dimensional sparse recovery problem over the space of Dirac measures, with J the total variation of a measure.

Compressed sensing. Condition $\Phi^* \mathfrak{p}_F \in \mathrm{ri}\,\partial J(x_0)$ is often used when Φ is drawn from the Gaussian matrix ensemble to asses the performance of compressed sensing recovery with ℓ^1 norm. It has been proved (Wainwright 2009; Dossal et al. 2012) for $J = \|\cdot\|_1$ that if Φ is a random matrix drawn from the Gaussian ensemble, then for $q > 2s \log n$, $\Phi^* \mathfrak{p}_F \in \mathrm{ri}\,\partial J(x)$ with high probability on Φ for $k = \|x_0\|_0$. One may have observed that the bound on q bears similarities with that of Section 1.2.1 except in the scaling in the log term. It was also used to ensure ℓ^2 robustness of matrix completion in a noisy setting by Candès et al. (2010), and our findings show that it also ensures rank consistency for matrix completion at high signal to low noise levels. It generalizes the result proved for a family of decomposable norms (including in particular ℓ^1-ℓ^2 norm and the nuclear norm) by Candès and Recht (2013) when $w = 0$.

Stronger criteria for ℓ^1. Many sufficient conditions can be formulated to ensure that $\mathfrak{p}_F$ is a non-degenerate certificate, and hence to guarantee the model stability. The strongest criterion to ensure a noise robustness for ℓ^1 regularization is the mutual coherence, introduced by Donoho et al. (2001). Finer criteria based on Babel functions have been proposed in (Gribonval and Nielsen 2008; Borup et al. 2008). The Exact Recovery Condition introduced by Tropp (2006) is weaker than the coherence which in turns is greater that the weak-ERC (Dossal 2012).

1.3 Sensitivity Analysis and Parameter Selection

Beside studying stability, the second goal of this thesis is to investigate the sensitivity of any solution $x^\star(y)$ to the parameterized problem $(\mathcal{P}_{y,\lambda})$ to (small) perturbations of y. This sensivity analysis is central to construct an unbiased estimator of the quadratic risk, as described in Section 1.3.2. We suppose here that J is a partly smooth gauge with linear manifold, i.e. such that $\mathcal{M}_x = T_x$ and J is 1-homogeneous. We conjecture that this statement remains true for any finite-valued convex partly smooth function, though this has not been formally proved yet. The technical obstacles faced by this generalization will be discussed in Chapter 9.

1.3.1 Local Differentiability of the Optimal Solutions

The objective here is find a formula of the derivative of $x^\star(y)$ with respect to the observations, valid on the biggest set possible. Moreover, since $x^\star(y)$ is not uniquely defined, it has to be interpreted as a multivalued mapping. Sensitivity analysis[1] is a major branch of optimization and optimal control theory. Comprehensive monographs on the subject are (Bonnans et al. 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.

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

1.3.1.1 Main Contribution

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[2] $y \mapsto x^\star(y)$. Rather, the sensitivity behavior will be local. This is the reason why we need to introduce the following transition space $\mathcal{H}$, which will be shown to contain points of non-smoothness of $y \mapsto x^\star(y)$.

We introduce the *transition space* $\mathcal{H}$ defined as

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

where $\Pi_{n+p,n}$ is the canonical projection onto the first n components, $\mathrm{bd}\, C$ is the boundary of C, and

$$\mathcal{A}_T = \left\{ (y, x_T) \in \mathbb{R}^n \times \widetilde{T} \mid \frac{1}{\lambda} \Phi_T^*(\Phi x_T - y) \in \mathrm{rbd}\, \partial J(x_T) \right\}.$$

Here, $\mathrm{rbd}\, \partial J(x_T)$ is the relative boundary of $\partial J(x_T)$ relatively to its affine hull and $\widetilde{T} = \{x \in \mathbb{R}^n \mid T_x = T\}$. This set corresponds exactly to the observations y such that the model space associated to a solution of $(\mathcal{P}_{y,\lambda})$ is not stable with respect to small perturbations. In particular, when $J = \|\cdot\|_1$, we show that this set is in fact a union of hyperplanes and when $J = \|\cdot\|_{1,2}$ it is a semi-algebraic set.

Our main sensitivity analysis is the following.

Theorem 3 Let $y \notin \mathcal{H}$ and $x^\star$ a solution of $(\mathcal{P}_{y,\lambda})$ such that

$$\mathrm{Ker}\, \Phi_T \cap \mathrm{Ker}\, D^2 J_T(x^\star) = \{0\} \qquad (I_{x^\star})$$

where $T = T_{x^\star}$. Then, there exists an open neighborhood $\mathcal{V} \subset \mathbb{R}^n$ of y, and a mapping $\tilde{x} : \mathcal{V} \to T$ such that

(i) For every $\bar{y} \in \mathcal{V}$, $\tilde{x}(\bar{y})$ is a solution of $(\mathcal{P}_{\bar{\lambda}}(\bar{y}))$, and $\tilde{x}(y) = x^\star$.

2. To be understood here as a set-valued mapping.

(ii) The mapping $\tilde{x}$ is $C^1(\mathcal{V})$ and

$$\forall \bar{y} \in \mathcal{V}, \quad D_1\tilde{x}(\bar{y}) = -(\Phi_T^*\Phi_T + \lambda D^2 J_T(x^\star))^{-1}\Phi_T.$$

The mapping $y \mapsto \widehat{\mu}(y) = \Phi x^\star$ is single-valued and $C^1(\mathbb{R}^n \setminus \mathcal{H})$. For every $y \notin \mathcal{H}$, there exists a solution $x^\star$ of $\mathcal{P}_\lambda(y)$ such that $(I_{x^\star})$ is satisfied. Moreover, for any $y \notin \mathcal{H}$,

$$\mathrm{div}(\widehat{\mu})(y) = \mathrm{tr}(\Delta(y))$$

where

$$\Delta(y) = -\Phi_T(\Phi_T{}^*\Phi_T + \lambda D^2 J_T(x^\star))^{-1} \circ \Phi_T{}^*.$$

This theorem is proved in Chapter 8.

1.3.1.2 Relation to Previous Works

Sensitivity analysis is a major branch of optimization and optimal control theory. Comprehensive monographs on the subject are (Bonnans et al. 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.

For the Lasso problem, the above divergence formula implies that

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

where $x^\star$ is a solution of $(\mathcal{P}_{y,\lambda})$ such that $(I_{x^\star})$ holds, i.e. $\Phi_{\mathrm{supp}(x^\star)}$ has full rank. This result was proved in (Dossal et al. 2013), see also (Tibshirani et al. 2012) where a similar result is proved without the condition $(I_{x^\star})$.

The case of analysis sparsity was investigated in (Vaiter, Deledalle, et al. 2013) and (Tibshirani et al. 2012). In this case, one has

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

where $x^\star$ is such that $(I_{x^\star})$ holds.

The originality of our contribution in this direction is the following:

(i) We formulate the set $\mathcal{H}$ of non-smoothness points, which is crucial for the application to risk estimation exposed bellow.

(ii) We give an explicit formula of the divergence of the prediction.

(iii) Our sensitivity result deals with a set-valued mapping (even if its image by Φ is single-valued).

1.3.2 Unbiased Risk Estimation

The degrees of freedom (DOF) of a statistical procedure quantifies its complexity (Efron 1986). Among possible applications is the computation of efficient risk estimators. These estimator allows an objectively guided choice of the hyperparameters associated to the statistical procedure.

Let $\mu_0 = \Phi x_0$. Suppose that the observations $Y \sim \mathcal{N}(\mu_0, \sigma^2 \mathrm{Id}_n)$. Following (Efron 1986), the DOF is defined as

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

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]

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

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

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

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

where $D\widehat{\mu}$ 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).

1.3.2.1 Main Contribution

To apply Stein's lemma we need to provide a closed-form of the Jacobian of $y \mapsto \widehat{\mu}(y)$ which holds true almost everywhere. Roughly speaking, to be able to control the size of $\mathbb{R}^q \setminus \mathcal{H}$, the functions J cannot be too oscillating in order to prevent pathological behaviors. In order to do this, we use arguments of o-minimal geometry. More precisely, we ask that the function J is definable in such a structure and that $\mathcal{T} = (T_x)_{x \in \mathbb{R}^n}$ is finite. These assumptions exclude the nuclear norm. Under such assumptions, we prove the following theorem.

Theorem 4 Let $Y = \Phi x_0 + W$ with $W \sim \mathcal{N}(0, \sigma^2 \mathrm{Id}_n)$. Then,

(i) $\mathcal{H}$ is of Lebesgue measure zero;

(ii) $\widehat{\mu}$ is Lipschitz continuous, hence weakly differentiable, with an essentially bounded gradient.

(iii) $\widehat{\mathrm{df}} = \mathrm{tr}(\Delta(Y))$ is an unbiased estimate of $\mathrm{df} = \mathbb{E}(\mathrm{div}(\widehat{\mu}(Y))$.

(iv) The SURE

$$\mathrm{SURE}(\widehat{\mu})(Y) = ||Y - \widehat{\mu}(Y)||^2 + 2\sigma^2 \widehat{\mathrm{df}} - n\sigma^2 \tag{1.17}$$

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

This theorem is proved in Chapter 9. This result holds true for the SURE within an exponential family, see Chapter 9.

1.3.2.2 Relation to Previous Works

In the case of standard Lasso (i.e. ℓ^1 penalty) with $Y \sim \mathcal{N}(\Phi x_0, \sigma^2 \mathrm{Id}_n)$ and Φ 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 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, were given independently in (Tibshirani et al. 2012; Vaiter, Deledalle, et al. 2013).

A formula of an estimate of the DOF for the group Lasso when the design is orthogonal within each group was conjectured in (Yuan et al. 2005). 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 et al. (2000) which treats the case where C is a convex polyhedral cone. When Φ 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 Φ as (Yuan et al. 2005). An estimate of the DOF for the group Lasso was also given by Solo et al. (2010) using heuristic derivations that are valid only when Φ is full column rank, though its unbiasedness is not proved.

1.4 Reading Guide

This thesis is organized in 11 chapters. Figure 1.3 provides a description of the dependencies between them. A summary in French of the thesis is provided in appendix.

Chapter 2: Mathematical Background. This chapter provides the necessary common material used in this thesis. In particular, we recall basic definitions of convex analysis (in particular gauges), o-minimality and smooth manifolds.

This chapter contains mainly well known definitions and properties, but we would like to emphasize that some new results are established. The reader is invited to take in account the List of Notations in appendix.

1.4.1 Part I: Models, Partial Smoothness and Dual Certificates

This part lays down the three main concepts used in this manuscrit: tangent model space, partly smooth functions and dual certificates. Each of these tools has its dedicated chapter.

Chapter 3: Model Tangent Subpace. In Chapter 3, we define the model tangent space and model vector in Definition 3.1 and the subdifferential gauge in Definition 3.2. This allows us to prove Theorem 3.1 which provides a pointwise decomposition of the subdifferential of any continuous convex function.

Chapter 4: Partial Smoothness. Chapter 4 introduces partly smooth functions (Definition 4.1) specialized to convex functions, and partly smooth functions relative to a linear manifold (Definition 4.2). In particular, we provide a derivation of explicit partial smoothness Lipschitz-constants for the latters.

Chapter 5: Certificates and Uniqueness. In Chapter 5, we introduce the (non-degenerate) dual certificates (Definition 5.1), minimal norm certificate (Definition 5.2), linearized precertificate (Definition 5.4) and its associated identifiability criterion (Definition 5.5). We also define the restricted injectivity assumption (Definition 5.3). The main result of this chapter is Theorem 5.3 which gives a sufficient condition for uniqueness for $(\mathcal{P}_{y,\lambda})$ or $(\mathcal{P}_{y,0})$.

1.4.2 Part II: Robustness

Chapter 6: Noise ℓ^2 Robustness. In this chapter, we prove Theorem 6.1 showing that if both the non-degenerate source condition and the restricted injectivity hold, then $(\mathcal{P}_{y,\lambda})$ enjoys a linear convergence rate with respect to the estimation error.

Chapter 7: Model Selection. In Chapter 7, we prove Theorem 7.2 which ensures that for a partly smooth function J, if the restricted injectivity holds and that the linearized precertificate is a non-degenerate certificate, then for a certain regime of small noise, $(\mathcal{P}_{y,\lambda})$ has a unique solution and x_0 is an element of the manifold relative to x_0. Theorem 7.3 proves a similar result for partly smooth function with linear manifold with more explicit constants.

1.4.3 Part III: Sensitivity

Chapter 8: Local Differentiability of the Optimal Solutions. In this chapter, we introduce the transition space (Definition 8.2) and the restricted injectivity for $(\mathcal{P}^F_{y,\lambda})$. Theorem 8.1 constructs a smooth solution mapping of $(\mathcal{P}^F_{y,\lambda})$ on an open neighborhood of some solution $x^\star$. Theorem 8.2 shows that the prediction map is well-defined outside the transition space and gives its local behavior.

Chapter 9: Unbiased Risk Estimation. In this chapter, we prove Proposition 9.1 stating that that the transition space has zero measure w.r.t Lebesgue measure. Proposition 9.2 proves that the prediction is Lipschitz continuous. Theorems 9.1 and 9.2 prove that the (G)SURE is an unbiased estimator of the risk for non-linear Gaussian regression and generalized linear model.

1.4.4 Numerical Considerations and Conclusion

Chapter 10: Numerical Considerations. This chapter recaps our results from a numerical point of view. We prove in Theorem 10.1 that under the same hypothesis of non-degeneracy and partial smoothness of Theorem 7.2, the forward-backward algorithm identifies the correct manifold after a finite number of steps. We discuss in Sections 10.2 and 10.3 how the linearized precertificate behaves in different concrete scenarios. We investigate further the noiseless behavior of total variation denoising in Theorem 10.2 and the compressed sensing with ℓ^∞ regularization in Theorem 10.3. We also show how in practice one can use our sensitivity result (Theorem 9.1) to select the best hyperparameter λ for ℓ^1-analysis regularization in Section 10.4.

Chapter 11: Conclusion. This last chapter summarizes our contributions. We also discuss several open problems.

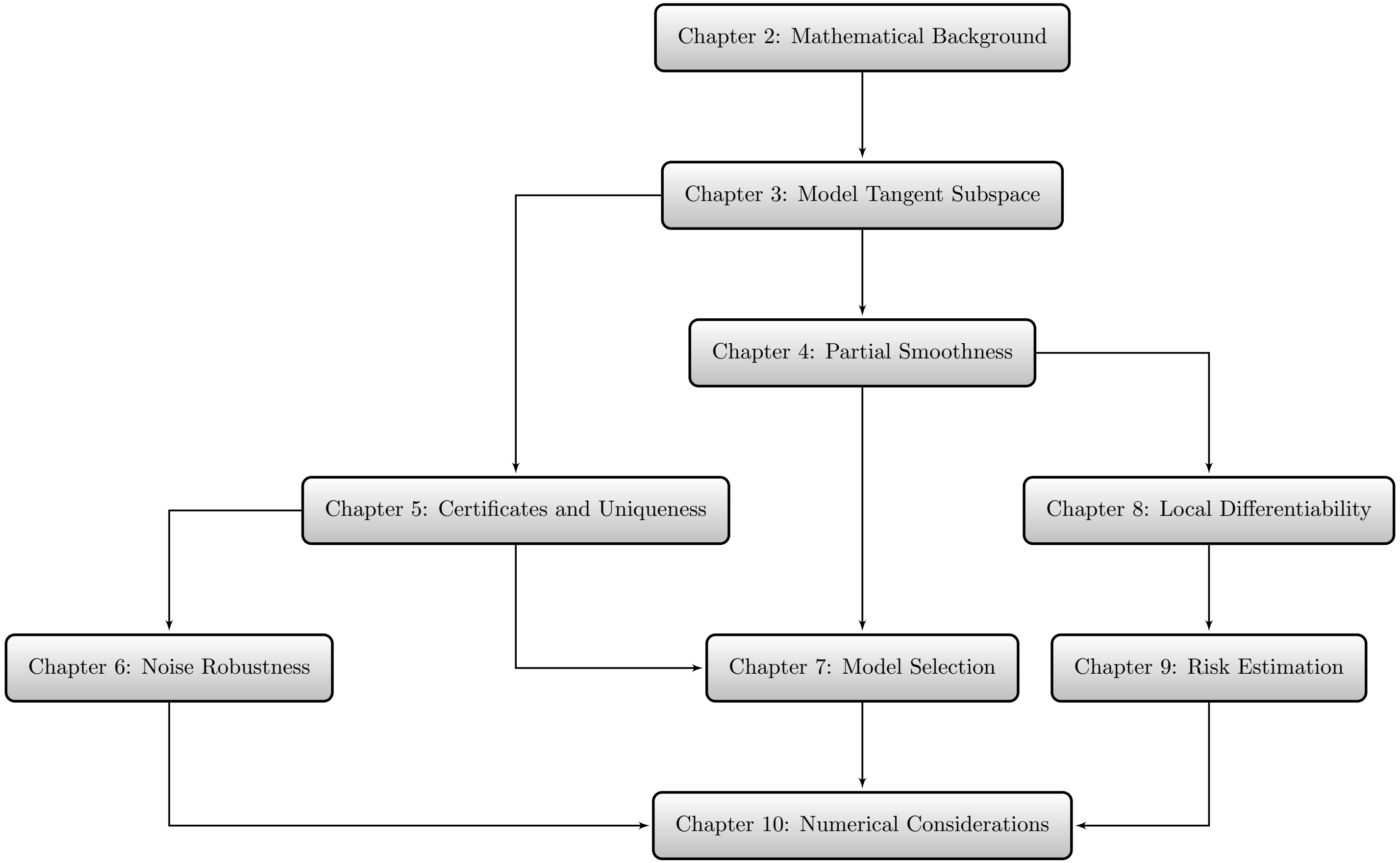


Figure 1.3: *Dependencies between chapters.*

2

Mathematical Background

Contents

This chapter lays the mathematical foundation of our work. In particular, we recap important results of convex analysis, o-minimal geometry and some properties of smooth manifolds. An emphasis is done on the notion of gauge which will be important in order to capture the structure of the subdifferential of a convex function.

In all the following, $\mathbb{R}^n$ will be the signal space, $\mathbb{R}^q$ the observation space and $\mathbb{R}^p$ the analysis space. The space $\mathbb{R}^n$ will be endowed with its canonical Euclidian structure and its associated inner product is denoted $\langle \cdot, \cdot \rangle$, i.e.

$$\forall x, x' \in \mathbb{R}^n, \langle x, x' \rangle = \sum_{i=1}^{n} x_i x'_i,$$

and the associated norm, the ℓ^2 norm[1] is denoted

$$||x|| = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

For any subspace T of a real vector space E, we denote P_T the orthogonal projection on T, and

$$x_T = P_T(x) \quad \text{and} \quad \Phi_T = \Phi \, P_T \, .$$

For a subset I of $\{1, \ldots, n\}$, we denote by I^c its complement with respect to $\{1, \ldots, n\}$, $|I|$ its cardinality, $x_{(I)}$ is the subvector whose entries are those of x restricted to the indices in I, and $\Phi_{(I)}$ the submatrix whose columns are those of Φ indexed by I. For any matrix A, A^* denotes its adjoint matrix and A^+ its Moore–Penrose pseudo-inverse. We denote the right-completion of the real line by $\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$.

Section 2.1 recalls basics of convex analysis, and Section 2.2 is concerned with differential properties. Then, Section 2.3 details properties of gauges.

1. Any other norm has its own subscript in this manuscrit.

Finally, in Section 2.4 we sketch some properties of o-minimal geometry, a generalization of semi-algebraic geometry.

2.1 Convex Analysis

In this section, we recall useful concepts from convex analysis in finite dimension. The definitive reference book on this subject is (Rockafellar 1996). One may also refers to (Zalinescu 2002; Hiriart-Urruty et al. 2001) for more details, or (Ekeland et al. 1974) for the infinite dimensional case.

2.1.1 Functions

We recall basic definitions of real analysis.

Definition 2.1 — Epigraph and Domain. The *epigraph* of a function $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ is the set

$$\operatorname{epi} f = \{(x, \alpha) \in \mathbb{R}^n \times \mathbb{R} \mid \alpha \geqslant f(x)\} \subseteq \mathbb{R}^{n+1}.$$

The *(effective) domain* of f is the set the projection of $\operatorname{epi} f$ under the mapping $(x, \alpha) \mapsto x$, i.e.

$$\operatorname{dom} f = \left\{x \in \mathbb{R}^N \mid f(x) < +\infty\right\}.$$

The function f is *proper* if $\operatorname{dom} f \neq \emptyset$.

The epigraph is the set of points lying above its graph. Coercivity and lower semicontinuity will play an important role.

Definition 2.2 — Coercivity. A function $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ is *coercive* if

$$\lim_{||x|| \to +\infty} f(x) = +\infty.$$

An example of coercive function is any norm over $\mathbb{R}^n$. However, this is not the case of $||D^* \cdot ||$ as soon as D^* has a non trivial kernel.

Definition 2.3 — Lower Semicontinuity. A function $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ is *lower semi-continuous* (l.s.c.) at $x \in \mathbb{R}^n$ if

$$\liminf_{z \to x} f(z) \geqslant f(x).$$

The fact that a function f is l.s.c. is equivalent to $\operatorname{epi} f$ closed in $\mathbb{R}^n \times \mathbb{R}$, see (Rockafellar et al. 1998, Theorem 1.6). For this reason, we also say that f is *closed*. We recall that

$$\liminf_{z \to x} f(z) = \sup_{\delta > 0} \inf_{\{z | \, ||z-x||\} < \delta} f(z).$$

Definition 2.4 — Kernel. The *kernel* of a function is defined as

$$\operatorname{Ker} f = \{x \in \mathbb{R}^n \mid f(x) = 0\}.$$

Note that the kernel of a function is not necessarily a linear subspace. However, if f is convex, $\operatorname{Ker} f$ is a convex cone.

Definition 2.5 — Sublevel Set. The *sublevel set* $\operatorname{slev}_x J$ of J passing through x is defined as

$$\operatorname{slev}_x J = \{z \in \mathbb{R}^n \mid J(z) \leqslant J(x)\} .$$

2.1.2 Convexity

All functionals considered in this manuscrit are convex. We recall the definition of convexity and give several examples.

Definition 2.6 — Convexity. A set $C \subseteq \mathbb{R}^n$ is said to be *convex* if

$$\forall x, x' \in C, \forall \mu \in [0,1], \quad \mu x + (1-\mu)x' \in C.$$

A function $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ is said to be *convex* if its epigraph is convex, i.e.

$$\forall x, x' \in \mathbb{R}^n, \forall \mu \in [0,1], \quad f(\mu x + (1-\mu)x') \leqslant \mu f(x) + (1-\mu) f(x').$$

It is *strictly convex* if

$$\forall x, x' \in \mathbb{R}^n, \forall \mu \in [0,1], \quad x \neq x' \Rightarrow f(\mu x + (1-\mu)x') < \mu f(x) + (1-\mu) f(x').$$

It is *strongly convex of modulus* τ if for every $x, x' \in C$ and every $\mu \in [0,1]$,

$$f(\mu x + (1-\mu)x') \leqslant \mu f(x) + (1-\mu) f(x') - \frac{\tau}{2}\mu(1-\mu)\|x' - x\|^2.$$

The set of all convex, proper and closed functions is denoted $\Gamma_0(\mathbb{R}^n)$. The set of all finite-valued, bounded from below, convex, proper (hence continuous) functions is denoted $\Gamma_c^+(\mathbb{R}^n)$.

Definition 2.7 — Indicator Function. Let C a nonempty closed convex subset of $\mathbb{R}^n$. The *indicator function* $\iota_C \in \Gamma_0(\mathbb{R}^n)$ of C is

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

Definition 2.8 — Conjugate. The Legendre–Fenchel *conjugate* $f^* \in \Gamma_0(\mathbb{R}^n)$ of a proper, closed and convex function $f \in \Gamma_0(\mathbb{R}^n)$ is

$$f^*(u) = \sup_{x \in \operatorname{dom} f} \langle u, x \rangle - f(x)\ .$$

Here, f^* is proper, closed and convex, thus $f^{**} = f$. For instance, the conjugate of the indicator function ι_C is the support function of C defined as:

Definition 2.9 — Support function The *support function* of a nonempty closed convex subset C of $\mathbb{R}^n$ is

$$\sigma_C(u) = \sup_{x \in C} \langle u, x \rangle \ .$$

σ_C is sublinear, is non-negative if $0 \in C$, and is finite everywhere if, and only if, C is a bounded set. We have the following property.

Proposition 2.1 Let C_1, C_2 two nonempty closed convex subsets. Then,

(i) $C_1 \subseteq C_2 \Leftrightarrow \sigma_{C_1} \leqslant \sigma_{C_2}$,

(ii) $\sigma_{C_1+C_2} = \sigma_{C_1} + \sigma_{C_2}$,

(iii) For any $\rho \in \mathbb{R}$, $\sigma_{\rho C_1} = \rho \sigma_{C_1}$.

Definition 2.10 — Infimal convolution. Let f and g be two proper closed convex functions from $\mathbb{R}^n$ to $\overline{\mathbb{R}}$. Their infimal convolution is the function

$$(f \overset{+}{\vee} g)(x) = \inf_{x_1+x_2=x} f(x_1) + g(x_2) = \inf_{z \in \mathbb{R}^N} f(z) + g(x-z) \ .$$

An important property of optimization with convex function is recapped below.

PROPOSITION 2.2 Let F_0 be a strictly convex function and J a closed convex function. Then every solutions of the problem

$$\min_{x \in \mathbb{R}^n} F_0(\Phi x) + J(x) \tag{2.1}$$

share the same image by Φ and the same value J. Moreover, given two solutions of (2.1) ${x^\star}_0, {x^\star}_1$, there exists δ such that ${x^\star}_0 = {x^\star}_1 + \delta$.

PROOF Let ${x^\star}_0, {x^\star}_1$ be two solutions of $\mathcal{P}(y)$ such that $\Phi {x^\star}_0 \neq \Phi {x^\star}_1$. Take any convex combination ${x^\star}_t = (1-t){x^\star}_0 + t{x^\star}_1$, $t \in]0,1[$. Strict convexity of $\mu \mapsto F_0(\mu)$ implies that the Jensen inequality is strict, i.e.

$$F_0(\Phi {x^\star}_t) < (1-t)F_0(\Phi {x^\star}_0) + tF_0(\Phi {x^\star}_1).$$

The convexity of the regularization implies

$$J({x^\star}_t) \leqslant (1-t)J({x^\star}_0) + tJ({x^\star}_1) \ .$$

Summing these two inequalities we arrive at

$$F_0(\Phi {x^\star}_t) + J({x^\star}_t) < F_0(\Phi {x^\star}_0) + J({x^\star}_0)$$

a contradiction since ${x^\star}_0$ is a minimizer of (2.1). ∎

2.1.3 Special Convex Sets

DEFINITION 2.11 — CONVEX HULL AND ITS CLOSURE. The *convex hull* of a non-empty set $C \subset \mathbb{R}^n$ is the intersection of all convex sets containing C. We denote $\overline{\text{co}}\,(C)$ the *closure of its convex hull*.

Definition 2.12 — Affine Hull. Its *affine hull* $\operatorname{aff} C$ is the smallest affine manifold containing it, i.e.

$$\operatorname{aff} C = \left\{ \sum_{i=1}^{k} \rho_i x_i \mid k > 0, \rho_i \in \mathbb{R}, x_i \in C, \sum_{i=1}^{k} \rho_i = 1 \right\}.$$

It is included in the *linear hull* $\operatorname{span} C$ which is the smallest subspace containing C.

Definition 2.13 — Interior and Relative Interior. The interior of C is denoted $\operatorname{int} C$. The *relative interior* $\operatorname{ri} C$ of a convex set C is the interior of C for the topology relative to its affine full.

Definition 2.14 — Closed Conical Hull. The *closed conical hull* of a nonempty set $C \subset \mathbb{R}^n$ is

$$\overline{\operatorname{cone}}(C) = \operatorname{cl} \left\{ \sum_{i=1}^{k} \rho_i x_i \mid k > 0, \rho_i \geqslant 0, x_i \in C \right\} .$$

Note that the closure operation is necessary. In general, the argument of the closure is neither compact nor closed, even if C is a convex compact set.

In the following, we give a handy expression of the tangent cone to a closed convex set, see (Hiriart-Urruty et al. 2001, Proposition III.5.2.1),

Definition 2.15 — Tangent and Normal Cones. The *tangent cone* to a nonempty closed convex set $C \subset \mathbb{R}^n$ at x is

$$T_C(x) = \overline{\operatorname{cone}}(C - x) = \operatorname{cl} \bigcup_{t \geqslant 0} t(C - x) .$$

The *normal cone* to C at x is the polar of $T_C(x)$, i.e.

$$N_C(x) = \{z \in \mathbb{R}^n \mid \forall c \in C, \langle z, x - c\rangle \leqslant 0\} .$$

2.1.4 Multivalued Mappings

We refer to (Aubin et al. 2009) for more details about multivalued mappings. We need the definition of continuity and Lipschitz-property in this work.

DEFINITION 2.16 — MULTIVALUED MAPPING. A *multivalued mapping* $S : X \rightrightarrows Y$ from X to Y is a mapping from X to the subsets of Y.

DEFINITION 2.17 — CONTINUITY. Let $S : X \rightrightarrows Y$ a multivalued mapping. We say that S is

- *outer semicontinuous* at x if $\limsup_{z\to x} S(z) \subseteq S(x)$ where

$$\limsup_{z\to x} S(z) = \{u \mid \exists x^\nu \to x, \exists u^\nu \to u \text{ with } u^\nu \in S(x^\nu)\}.$$

- *inner semicontinuous* at x if $\liminf_{z\to x} S(z) \supseteq S(x)$ where

$$\liminf_{z\to x} S(z) = \{u \mid \forall x^\nu \to x, \text{ with } u^\nu \in S(x^\nu)\}.$$

- *continuous* if both conditions holds.

DEFINITION 2.18 — LIPSCHITZ MAP. Let $S : X \subseteq \mathbb{R}^n \rightrightarrows Y$ a multivalued mapping. We say that S is β-*Lipschitz* around $x \in X$ if there exists a neighborhood U of x such that

$$\forall x_1, x_2 \in U, S(x_1) \subseteq S(x_2) + \beta\|x_1 - x_2\|B_X,$$

where B_X is the unit ball of X.

The following lemma is important in the study of partly smooth functions in Chapter 4.

LEMMA 2.1 Let $C : X \rightrightarrows Y$ be a β-Lipschitz multivalued mapping, such that $C(x)$ is a compact convex set for every $x \in X$. Then, for every $x_1, x_2 \in X$ and $y \in Y$,

$$\sigma_{C(x_1)}(y) - \sigma_{C(x_2)}(y) \leqslant \beta \|x_1 - x_2\| \|y\|.$$

PROOF Since $C(x_1) \subseteq C(x_2) + \beta \|x_1 - x_2\| B_X$, we have by 2.1 (i),

$$\sigma_{C(x_1)} \leqslant \sigma_{\subseteq C(x_2) + \beta \|x_1 - x_2\| B_X}.$$

By Proposition 2.1 (ii) and (iii), we obtain

$$\sigma_{C(x_1)}(y) \leqslant \sigma_{C(x_2)}(y) + \beta \|x_1 - x_2\| \sigma_{B_X}(y).$$

Since $\sigma_{B_X}(y) = \|y\|$, we obtain our claim. ∎

A proof of this statement can also be found in (Hiriart-Urruty et al. 2001, Theorem V.3.3.8).

2.1.5 Asymptotic Cone and Function

DEFINITION 2.19 Let C be a non-empty closed convex set in $\mathbb{R}^n$. Its asymptotic cone, or recession cone, C_∞ is the set

$$C_\infty = \{d \in \mathbb{R}^n \mid x + td \in C, \ \forall t > 0\} = \bigcap_{t>0} \frac{C - x}{t}, \quad \forall x \in C .$$

The closure assumption on the convex set C is crucial and cannot be removed.

The importance of the asymptotic cone is revealed by the following key properties, in particular property (iii).

PROPOSITION 2.3 Let C be a non-empty closed convex set in $\mathbb{R}^n$.

(i) C_∞ is independent of x.

(ii) C_∞ is a closed convex cone.

(iii) C is compact if and only if $C_\infty = \{0\}$.

(iv) If C is non-empty closed convex cone, then $C_\infty = C$.

PROOF (i) (Auslender et al. 2003, Proposition 2.1.5).

(ii) (Auslender et al. 2003, Proposition 2.1.5).

(iii) (Auslender et al. 2003, Proposition 2.1.2).

(iv) (Auslender et al. 2003, Proposition 2.1.1(c) and Proposition 2.1.5). ∎

DEFINITION 2.20 For any function $f \in \Gamma_0(\mathbb{R}^n)$, there exists a unique function $f_\infty : \mathbb{R}^n \to \overline{\mathbb{R}}$ associated with f, called the asymptotic function, or the recession function, such that $\operatorname{epi} f_\infty = (\operatorname{epi} f)_\infty$.

In (Hiriart-Urruty et al. 2001), the notation f'_∞ is used which is justified by the properties hereafter.

The epigraph of f_∞ is a closed convex cone, see Proposition 2.3. Moreover, f_∞ enjoys many important properties some of which we summarize as follows.

PROPOSITION 2.4 Let $f \in \Gamma_0(\mathbb{R}^n)$.

(i) $f_\infty \in \Gamma_0(\mathbb{R}^n)$ and positively homogeneous,

$$f_\infty(d) = \sup_{x \in \mathrm{dom}(f)} f(x+d) - f(x) ,$$

and

$$f_\infty(d) = \lim_{t \to +\infty} \frac{f(x+td) - f(x)}{t} = \sup_{t>0} \frac{f(x+td) - f(x)}{t}, \quad \forall x \in \mathrm{dom}(f) .$$

(ii) In particular, if $0 \in \mathrm{dom}(f)$, then $\forall d \in \mathbb{R}^n$

$$f_\infty(d) = \lim_{t \to +\infty} \frac{f(td)}{t} .$$

(iii) $(\iota_C)_\infty = \iota_{C_\infty}$, for C a non-empty closed convex set.

(iv) If f is a gauge of C containing the origin, then $\mathrm{Ker}(f) = C_\infty$.

(v) Let $f_i \in \Gamma_0(\mathbb{R}^n)$, $i = 1, \cdots, m$, $f := \sum_{i=1}^p f_i$ and $\bigcap_{i=1}^m \mathrm{dom}(f_i) \neq \emptyset$. Then, $f \in \Gamma_0(\mathbb{R}^n)$ and

$$f_\infty = \sum_{i=1}^p (f_i)_\infty .$$

(vi) Let $A : \mathbb{R}^n \to \mathbb{R}^p$ be a linear map such that $\mathrm{Im}(A) \cap \mathrm{dom}(f) \neq \emptyset$. Then

$$(f \circ A)_\infty(d) = f_\infty(Ad) .$$

PROOF (i) The first statement is a consequence of convexity and (Auslender et al. 2003, Proposition 2.5.1(a)), which is in turn uses Proposition 2.3(ii). The equivalent expressions of f_∞ follow from (Auslender et al. 2003, Proposition 2.5.2).

(ii) Since $0 \in \mathrm{dom}(f)$, $f(0) < \infty$ and the formula follows from (i).

(iii) (Auslender et al. 2003, Corollary 2.5.1).

(iv) (Auslender et al. 2003, Proposition 2.6.1).

(v) (Auslender et al. 2003, Proposition 2.6.3). ∎

2.2 Differential Properties

The set of continuously differentiable functions from a set $X \subseteq \mathbb{R}^n$ to $\mathbb{R}^q$ is denoted $C^1(X, Y)$ and the Jacobian of a function $f : X \to \mathbb{R}^q$ at a point $x \in X$ is denoted $Df(x)$.

First of all, we should recall the classical implicit function theorem.

THEOREM 2.1 — IMPLICIT FUNCTION. Let $f : \mathbb{R}^n \times \mathbb{R}^q \to \mathbb{R}^n$ be a C^1 function in a neighborhood of $(\bar{x}, \bar{y})$ such that

$$f(\bar{x}, \bar{y}) = 0.$$

Assume that the Jacobian matrix $D_1 f(\bar{x}, \bar{y})$ with respect to the first variable is non-singular at $(\bar{x}, \bar{y})$. Then, there exists an open neighborhood U of $\bar{y}$ and a mapping $\tilde{x} : U \to \mathbb{R}^n$ such that $\tilde{x}$ is C^1 on U,

$$\forall y \in U,\ f(\tilde{x}(y), y) = 0 \quad \text{and} \quad \tilde{x}(\bar{y}) = \bar{x}.$$

Moreover, its Jacobian reads

$$\forall y \in U,\ D\tilde{x}(y) = -(D_1 f(\tilde{x}(y), y))^{-1} D_2 f(\tilde{x}(y), y).$$

We draw the attention of the reader to the fact that this theorem admits several generalizations such as for instance in the context of multivalued mappings, see the monograph of Dontchev et al. (2009).

2.2.1 Subdifferential

Definition 2.21 — Subdifferential. The *subdifferential* $\partial f(x)$ of a convex function f at x is the set

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

An element of $\partial f(x)$ is a subgradient. If the convex function f is Gâteaux-differentiable at x, then its only subgradient is its gradient, i.e. $\partial f(x) = \{\nabla f(x)\}$.

Proposition 2.5 Let $f \in \Gamma_0(\mathbb{R}^n)$. Then ∂f is outer semicontinuous.

PROOF See (Hiriart-Urruty et al. 2001, Theorem 6.2.4). ■

This result can be stated as

$$\forall \varepsilon > 0, \exists \delta > 0, \|x' - x\| \leqslant \delta \Rightarrow \partial f(x') \subseteq \partial f(x) + B(0, \varepsilon).$$

Note that without additional constraints, ∂f is *not* inner semicontinuous. This fact will motivate us to introduce the notion of partial smoothness in Chapter 4.

Definition 2.22 — Directional Derivative. The *directional derivative* $f'(x, \delta)$ of a finite-valued closed function f at the point $x \in \operatorname{dom} f$ in the direction $\delta \in \mathbb{R}^n$ is

$$f'(x, \delta) = \lim_{t \downarrow 0} \frac{f(x + t\delta) - f(x)}{t}.$$

When f is convex, then the function $\delta \mapsto f'(x, \cdot)$ exists and is sublinear. The subdifferential $\partial f(x)$ is a non-empty compact convex set of $\mathbb{R}^n$ whose support function is $f'(x, \cdot)$, i.e.

$$f'(x, \delta) = \sigma_{\partial f(x)}(\delta) = \sup_{\eta \in \partial f(x)} \langle \eta, \delta \rangle.$$

We also recall the fundamental first-order minimality condition of a convex function.

> Proposition 2.6 A vector $x^\star$ is the global minimizer of a convex function f if, and only if, $0 \in \partial f(x)$.

We define the Bregman divergence, a classical tool in convex analysis.

> Definition 2.23 The *Bregman divergence* $D^J_\eta(x, x_0)$ associated to a convex function $J \in \Gamma_0(\mathbb{R}^n)$ and a vector $\eta \in \partial J(x_0)$ between two points x and $x_0 \in \mathbb{R}^n$ is defined as
>
> $$D^J_\eta(x, x_0) = J(x) - J(x_0) - \langle \eta, x - x_0 \rangle .$$

It is obvious that by convexity, the Bregman divergence is non-negative. When J is differentiable at x_0, the unique Bregman divergence is then associated to $\eta = \nabla J(x_0)$ and we recover the standard smooth case where

$$D^J(x, x_0) = J(x) - J(x_0) - \langle \nabla J(x_0), x - x_0 \rangle .$$

Note that the Bregman divergence is not a distance, since it does not satisfy the triangle inequality nor the symmetry axioms. However, it is common in the litterature to find the term Bregman distance. We will drop the exponent J if the context allows it.

2.2.2 Minimizers Gradients for Composite Problems

The following lemma shows that for a minimization problem $\min f + g$ such that $f \in \Gamma_0(\mathbb{R}^n)$ is C^2 and $g \in \Gamma_0(\mathbb{R}^n)$, the solutions share the same gradient.

LEMMA 2.2 Let $x^{\star}_0$ and $x^{\star}_1$ be two solutions of

$$\min_{x \in \mathbb{R}^p} f(x) + g(x) \tag{2.2}$$

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

$$\nabla f(x^{\star}_0) = \nabla f(x^{\star}_1).$$

PROOF Let $x^{\star}_0$ and $x^{\star}_1$ be two distinct solutions of (2.2), otherwise, there is nothing to prove. We denote $x^{\star}_t = x^{\star}_0 + th$ where $h = x^{\star}_1 - x^{\star}_0$, $t \in [0, 1]$. By convexity, $x^{\star}_t$ is also a minimizer of (2.2). We have $-\nabla f(x^{\star}_t) \in \partial g(x^{\star}_t)$. Convexity of g then yields

$$\langle \nabla f(x^{\star}_t) - \nabla f(x^{\star}_0), th \rangle \leqslant 0.$$

Similarly, convexity of f entails

$$\langle \nabla f(x^{\star}_t) - \nabla f(x^{\star}_0), th \rangle \geqslant 0.$$

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

$$\langle \nabla f(x^{\star}_t) - \nabla f(x^{\star}_0), h \rangle = 0. \tag{2.3}$$

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

$$\nabla f(x^{\star}_1) - \nabla f(x^{\star}_0) = \int_0^1 D^2 f(x^{\star}_t) h dt \ , \tag{2.4}$$

which, after taking the inner product of both sides with h and using (2.3), yields

$$\langle \nabla f(x^{\star}_1) - \nabla f(x^{\star}_0), h \rangle = \int_0^1 \langle D^2 f(x^{\star}_t) h, h \rangle dt = 0. \tag{2.5}$$

By convexity, the Hessian $D^2 f(x^{\star}_t)$ is semidefinite positive, and (2.5) implies that

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

or equivalently

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

Inserting this again in (2.4) yields the desired claim. ∎

2.2.3 Smooth Manifolds

In this thesis, we will not use advanced results of differential geometry. However, we need the structure of smooth manifold to define the central notion of partial smoothness. This section aims to to recall basic notion on smooth manifolds. The reader may refer to (Lee 2003).

> DEFINITION 2.24 — SMOOTH MANIFOLD. Let $k \geqslant 1$. A C^k*-manifold* $\mathcal{M}$ *around* $x \in \mathbb{R}^n$ *of codimension* m is a subset of $\mathbb{R}^n$ such that there exists an open set U of $\mathbb{R}^n$ and a C^k-function $g : U \to \mathbb{R}^m$ satisfying
>
> $$\mathcal{M} \cap U = \{\bar{x} \in U \mid \ g(\bar{x}) = 0\},$$
>
> and g has surjective derivative throughout U. We say that $\mathcal{M}$ is a C^k*-manifold* if $\mathcal{M}$ is a C^k-manifold around every $x \in \mathcal{M}$ of codimension m.

Note that every linear subspace H of $\mathbb{R}^n$ is a manifold around each point $x \in H$, and this is in particular true for $H = \mathbb{R}^n$. Another example, which will be used in this thesis, is the set of matrices of fixed rank (Lee 2003).

> DEFINITION 2.25 — TANGENT SPACE. Let $\mathcal{M}$ be a C^k-manifold around $x \in \mathcal{M}$ of codimension m associated to a C^k function g. The *tangent space* of $\mathcal{M}$ at x is defined as
>
> $$\mathcal{T}_x(\mathcal{M}) = \operatorname{Ker} Dg(x).$$

We now introduce the Grassmann manifold, which will be used in Chapter 7.

PROPOSITION 2.7 — GRASSMANN MANIFOLD. Let $G_{k,n}$ be the set of all linear subspaces of $\mathbb{R}^n$ of dimension k. Then, $G_{k,n}$ is a smooth manifold of dimension $k(n-k)$, coined the Grassman manifold of k-planes. Moreover, $(G_{k,n}, d)$ endowed with

$$d(V, V') = \|P_V - P_{V'}\|$$

is a compact metric space, where $\|\cdot\|$ is an operator norm.

PROOF This property is a consequence of the isomorphism between $G_{k,n}$ and $\mathcal{O}_n/(\mathcal{O}_k \times \mathcal{O}_{n-k})$, see (Lee 2003). ∎

An important property is the fact that the projection onto a manifold is locally well-defined as a single-valued mapping.

LEMMA 2.3 Let $\mathcal{M}$ be a C^k-manifold with $k \geqslant 2$ around a point $x \in \mathcal{M}$. Then, there exists a neigborhood $\mathcal{U}$ of x such that for every $\bar{x} \in \mathcal{U}$, $\bar{x}$ has a unique projection $P_{\mathcal{M}}(\bar{x})$ onto $\mathcal{M}$. Moreover, the function $P_{\mathcal{M}} : \mathcal{U} \to \mathcal{M}$ is C^{k-1}, with derivative

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

PROOF See (Lewis et al. 2008, Lemma 2.1). ∎

2.3 Gauges

This section gives some general results on gauges. Again, we refer to (Rockafellar 1996) for more insight on this notion. Gauges are equivalently defined as nonnegative, convex and positively homegeneous functions or are parameterized by convex set containing 0. They are the natural extension of norms or seminorms, which are indeed gauges. The classical duality is then replaced by the polarity of convex set parameterizing the gauges.

2.3.1 Gauge and its Polar

2.3.1.1 Definitions and Main Properties.

We start by defining formally a gauge, and prove the associated Lemma 2.4 stating the equivalence between gauges and convex sets containing zero.

We begin with the definition of a gauge.

Definition 2.26 — Gauge Let $C \subseteq \mathbb{R}^n$ be a non-empty closed convex set containing the origin. The *gauge* of C is the function $\gamma_C \in \Gamma_0(\mathbb{R}^n)$ defined on $\mathbb{R}^n$ by

$$\gamma_C(x) = \inf\{\lambda > 0 \mid x \in \lambda C\}.$$

As usual, $\gamma_C(x) = +\infty$ if the infimum is not attained.

We say that γ_C is bounded (or finite-valued) if, for every $x \in \mathbb{R}^n, \gamma_C(x) < +\infty$. This is typically not the case if the gauge is of the form $\gamma_C(x) = f(x) + \iota_{\mathcal{C}}(x)$ where $\iota_{\mathcal{C}}$ is the indicator function of a convex set $\mathcal{C}$. Some important properties are stated below. In particular, Lemma 2.4((ii)) is a fundamental result of convex analysis that states that there is a one-to-one correspondence between gauge functions and closed convex sets containing the origin. This allows to identify sets from their gauges, and vice versa.

Lemma 2.4 Let $C \subseteq \mathbb{R}^n$ and γ_C the associated gauge.

(i) γ_C is a non-negative, lsc and sublinear function.

(ii) Suppose C is a closed convex set containing the origin. Then, f is the gauge associated to C if, and only if, f is positively homegeneous and

$$C = \{x \in \mathbb{R}^n \mid f(x) \leqslant 1\}.$$

(iii) γ_C is bounded if, and only if, $0 \in \operatorname{int} C$, in which case γ_C is continuous.

(iv) $\operatorname{Ker} \gamma_C = \{0\}$, or equivalently γ_C is coercive if, and only if, C is compact.

(v) γ_C is bounded and coercive on $\operatorname{dom} \gamma_C = \operatorname{span} C$ if, and only if, C is compact and $0 \in \operatorname{ri} C$. In particular, γ_C is bounded and coercive if, and only if, C is compact and $0 \in \operatorname{int} C$.

PROOF (i)-(iii) are obtained from (Hiriart-Urruty et al. 2001, Theorem V.1.2.5). (iv) is obtained by combining (Hiriart-Urruty et al. 2001, Corollary V.1.2.6 and Proposition IV.3.2.5). (v): the second statement follows by combining (iii)-(iv), while the first part is the second one written in $\operatorname{dom} \gamma_C = \operatorname{aff} C = \operatorname{span} C$ since $0 \in C$. ■

2.3.1.2 Polar Set and Gauges.

Let us now turn to the polar of a convex set and a gauge.

DEFINITION 2.27 — POLAR SET Let C be a non-empty convex set. The set C° given by

$$C^\circ = \{v \in \mathbb{R}^n \mid \forall x \in C, \langle v, x \rangle \leqslant 1\}$$

is called the *polar* of C.

C° is a closed convex set containing the origin. When the set C is also closed and contains the origin, then it coincides with its bipolar, i.e. $C^{\circ\circ} = C$.

We are now in position to define the polar gauge.

Definition 2.28 — Polar Gauge The polar of a gauge γ_C is the function γ_C° defined by

$$\gamma_C^\circ(u) = \inf\{\mu \geqslant 0 \mid \forall x \in \mathbb{R}^n, \langle x, u\rangle \leqslant \mu\gamma_C(x)\} \ .$$

Observe that gauges polar to each other have the property

$$\forall (x, u) \in \operatorname{dom}\gamma_C \times \operatorname{dom}\gamma_C^\circ, \quad \langle x, u\rangle \leqslant \gamma_C(x)\gamma_C^\circ(u) \ ,$$

just as dual norms satisfy a duality inequality. In fact, polar pairs of gauges correspond to the best inequalities of this type. The following Lemma 2.5 shows the relation between polar sets and polar gauges.

Lemma 2.5 Let $C \subseteq \mathbb{R}^n$ be a closed convex set containing 0. Then,

(i) γ_C° is a gauge function and $\gamma_C^{\circ\circ} = \gamma_C$.

(ii) $\gamma_C^\circ = \gamma_{C^\circ}$, or equivalently

$$C^\circ = \{x \in \mathbb{R}^n \mid \gamma_C^\circ(x) \leqslant 1\} = \{x \in \mathbb{R}^n \mid \gamma_{C^\circ}(x) \leqslant 1\}.$$

(iii) The gauge of C and the support function of C are mutually polar, i.e.

$$\gamma_C = \sigma_{C^\circ} \quad \text{and} \quad \gamma_{C^\circ} = \sigma_C \ .$$

Proof (i) follows from (Rockafellar 1996, Theorem 15.1). (ii) (Rockafellar 1996, Corollary 15.1.1) or (Hiriart-Urruty et al. 2001, Proposition V.3.2.4). (iii) (Rockafellar 1996, Corollary 15.1.2) or (Hiriart-Urruty et al. 2001, Proposition V.3.2.5). ■

2.3.1.3 Subdifferential of a Gauge

The subdifferential of a gauge γ_C at a point x is completely characterized by the face of its polar set C° exposed by x. Put formally, we have, PROPOSITION 2.8 Let C be a convex set containing 0. Then,

$$\partial\gamma_C(x) = \mathbf{F}_{C^\circ}(x) = \left\{\eta \in \mathbb{R}^N \mid \eta \in C^\circ \quad \text{and} \quad \langle \eta, x\rangle = \gamma_C(x)\right\},$$

where $\mathbf{F}_{C^\circ}(x)$ is the face of C° exposed by x. The latter is the intersection of C° and the supporting hyperplane $\{\eta \in \mathbb{R}^N \mid \langle \eta, x\rangle = \gamma_C(x)\}$.

PROOF See (Hiriart-Urruty et al. 2001, Proposition 3.1.4). ■

The special case of $x = 0$ has a much simpler structure; it is the polar set C° from Lemma 2.5(ii)-(iii), i.e.

$$\partial\gamma_C(0) = \left\{\eta \in \mathbb{R}^N \mid \gamma_{C^\circ}(\eta) \leqslant 1\right\} = C^\circ.$$

2.3.2 Polar Calculus

We here derive the expression of the gauge function of the Minkowski sum of two sets, as well as that of the image of a set by a linear operator. These results play an important role in Chapter 7.

First of all, we prove that if a multivalued mapping is Lipschitz, then the polar mapping is also Lipschitz continuous.

LEMMA 2.6 Let $C : X \rightrightarrows \mathbb{R}^n$ be a β_C-Lipschitz multivalued mapping, such that $C(x)$ is a compact convex set containing 0 for every $x \in X$. Then C° defined by as $x \mapsto C(x)^\circ$ is β_C-Lipschitz and the mapping $x \mapsto \gamma_{C(x)}$ is β_C-Lipschitz.

PROOF Using the Lipschitz continuity of C, there exists $\beta_{C(x)}$ such that

$$C(x') \subseteq C(x) + \beta_C \|x' - x\| B_X,$$

Using the symmetry of B_X, we get that

$$C(x') + \beta_C \|x' - x\| B_X \subseteq C(x).$$

Since the polarity reverse the order for the inclusion, we have

$$(C(x') + \beta_C \|x' - x\| B_X)^\circ \supseteq C(x)^\circ.$$

Hence,

$$\sigma_{(C(x')+\beta_C\|x'-x\|B_X)^\circ} \geqslant \sigma_{C(x)^\circ},$$

or equivalently

$$\gamma_{C(x')+\beta_C\|x'-x\|B_X} \geqslant \gamma_{C(x)}. \tag{2.6}$$

According to Lemma 2.7, one has

$$\gamma_{C(x')+\beta_C\|x'-x\|B_X}(u) = \inf_{z\in\mathbb{R}^n} \max(\gamma_{C(x')}(u), \gamma_{\beta_C\|x'-x\|B_X}(u-z)).$$

Hence,

$$\begin{aligned}\gamma_{C(x')+\beta_C\|x'-x\|B_X}(u) &\leqslant \gamma_{C_(x')}(u) + \gamma_{\beta_C\|x'-x\|B_X}(u)\\ &= \gamma_{C(x')}(u) + \beta_C\|x'-x\|\|u\|.\end{aligned}$$

Hence, combining with (2.6), we get

$$\gamma_{C(x')}(u) + \beta_C\|x'-x\|\|u\| \geqslant \gamma_{C(x)},$$

or equivalently,

$$\gamma_{C(x)^\circ} \leqslant \gamma_{C(x')^\circ}(u) + \beta_C\|x'-x\|\|u\|,$$

which concludes the proof. ∎

2.3.2.1 Minkowski Sum

Recall that the Minkowski sum of two sets A and B, subsets of $\mathbb{R}^n$ is defined as

$$A + B = \{a + b \in \mathbb{R}^n \mid a \in A, b \in B\}.$$

In particular, the (Minkowski) sum of two convex sets containing zero is a convex set containing zero. The following Lemma 2.7 makes a connection between the gauge of $C_1 + C_2$ with gauges of C_1 and C_2 through the inf-convolution operator.

LEMMA 2.7 Let C_1 and C_2 be nonempty closed convex sets containing the origin. Then

$$\gamma_{C_1+C_2}(x) = \sup_{\rho\in[0,1]} \rho\gamma_{C_1} \overset{+}{\vee} (1-\rho)\gamma_{C_2}(x)\ .$$

If x is such that $\gamma_{C_1}(x_1) + \gamma_{C_2}(x_2)$ is continuous and bounded on $\{(x_1, x_2) \mid x_1 + x_2 = x\}$, then

$$\gamma_{C_1+C_2}(x) = \inf_{z\in\mathbb{R}^n} \max(\gamma_{C_1}(z), \gamma_{C_2}(x - z))\ .$$

PROOF We have from Lemma 2.5 and calculus rules on support functions,

$$\gamma_{(C_1+C_2)^\circ} = \sigma_{C_1+C_2} = \sigma_{C_1} + \sigma_{C_2}\ .$$

Thus,

$$(C_1 + C_2)^\circ = \{u \mid \sigma_{C_1}(u) + \sigma_{C_2}(u) \leqslant 1\}\ . \tag{2.7}$$

Using the fact that the gauge of a set C is the support function of its polar, we have

$$\gamma_{C_1+C_2}(x) = \sigma_{(C_1+C_2)^\circ}(x).$$

Using (2.7),

$$\gamma_{C_1+C_2}(x) = \sigma_{\sigma_{C_1}(u)+\sigma_{C_2}(u)\leqslant 1}(x).$$

By definition of the support function,

$$\gamma_{C_1+C_2}(x) = \sup_{\sigma_{C_1}(u)+\sigma_{C_2}(u)\leqslant 1} \langle u, x\rangle.$$

Introduction $\rho = \sigma_{C_1}(u) + \sigma_{C_2}(u)$, we rewrite it as

$$\gamma_{C_1+C_2}(x) = \sup_{\rho\in[0,1]} \sup_{\sigma_{C_1}(u)\leqslant\rho, \sigma_{C_2}(u)\leqslant 1-\rho} \langle u, x\rangle.$$

This yields

$$\begin{aligned}\gamma_{C_1+C_2}(x) &= \sup_{\rho\in[0,1]} \sigma_{\sigma_{C_1}(u)\leqslant\rho} \overset{+}{\vee} \sigma_{\sigma_{C_2}(u)\leqslant 1-\rho}(x)\\ &= \sup_{\rho\in[0,1]} \rho\sigma_{\sigma_{C_1}(u)\leqslant 1} \overset{+}{\vee} (1-\rho)\sigma_{\sigma_{C_2}(u)\leqslant 1}(x).\end{aligned}$$

By definition of the polarity,

$$\begin{aligned}\gamma_{C_1+C_2}(x) &= \sup_{\rho\in[0,1]} \rho\sigma_{C_1^\circ} \overset{+}{\vee} (1-\rho)\sigma_{C_2^\circ}(x)\\ &= \sup_{\rho\in[0,1]} \sigma_{\rho C_1^\circ} \overset{+}{\vee} \sigma_{(1-\rho)C_2^\circ}(x)\\ &= \sup_{\rho\in[0,1]} \rho\gamma_{C_1} \overset{+}{\vee} (1-\rho)\gamma_{C_2}(x)\ ,\end{aligned}$$

which is the first assertion.

The last identity can be rewritten

$$\gamma_{C_1+C_2}(x) = \sup_{\rho\in[0,1]} \inf_{x_1+x_2=x} \rho\gamma_{C_1}(x_1) + (1-\rho)\gamma_{C_2}(x_2)\ .$$

Under the boundedness and continuity assumption of the lemma, the objective[2] in

2. The objective function is the function to be optimized.

the sup inf is a continuous bounded concave-convex function on the set $[0,1] \times \{(x_1, x_2) \mid x_1 + x_2 = x\}$. Since the latter sets are non-empty, closed and convex, and $[0,1]$ is obviously bounded, we have from using (Rockafellar 1996, Corollary 37.3.2)

$$\begin{aligned}\gamma_{C_1+C_2}(x) &= \inf_{z\in\mathbb{R}^n} \sup_{\rho\in[0,1]} \rho\gamma_{C_1}(z) + (1-\rho)\gamma_{C_2}(x-z) \\ &= \inf_{z\in\mathbb{R}^n} \max(\gamma_{C_1}(z), \gamma_{C_2}(x-z)) \ ,\end{aligned}$$

which concludes the proof. ∎

2.3.2.2 Image of a Set by a Linear Operator

Considering a linear operator $D : \mathbb{R}^p \to \mathbb{R}^n$, one constructs the image $D(C)$ of a convex set $C \subseteq \mathbb{R}^p$ by

$$D(C) = \{Dx \in \mathbb{R}^n \mid x \in C\} .$$

Lemma 2.8 connects the gauge associated to C to the one associated to $D(C)$ by an optimization problem over the kernel of D.

> LEMMA 2.8 Let C be a closed convex set such that $0 \in \operatorname{ri} C$, and D a linear operator. Then, for every $x \in \operatorname{Im}(D)$
>
> $$\gamma_{D(C)}(x) = \inf_{z\in\operatorname{Ker}(D)} \gamma_C(D^+x + z) \ .$$

PROOF It is immediate to see that $D(C)$ is a closed convex set containing the origin. Moreover, we have $\operatorname{Im}(D^*) \cap \operatorname{dom}(\sigma_C) \neq \emptyset$, since the origin is in both of them. Thus, using (Hiriart-Urruty et al. 2001, Theorem X.2.1.1) and Lemma 2.5, we have

$$\gamma_{(D(C))^\circ} = \sigma_{D(C)} = \left(\iota_{D(C)}\right)^* = \sigma_C \circ D^* \ .$$

Now, as by assumption $0 \in \operatorname{ri} C$, we have $0 \in \operatorname{ri}(C^\circ)$, and therefore $\operatorname{Im}(D^*) \cap \operatorname{ri}(C^\circ) \neq \emptyset$. By virtue of (Hiriart-Urruty et al. 2001, Theorem X.2.2.3) and Lemma 2.5, we get

$$\begin{aligned}
\gamma_{D(C)}(x) &= \sigma_{(D(C))^\circ}(x) \\
&= \sigma_{\sigma_C \circ D^*(u) \leqslant 1}(x) \\
&= \left(\iota_{\sigma_C(w) \leqslant 1} \circ D^*\right)^*(x) \\
&= \inf_v \sigma_{\sigma_C(w) \leqslant 1}(v) \quad \text{s.t.} \quad Dv = x \\
&= \inf_{z \in \mathrm{Ker}(D)} \sigma_{\sigma_C(w) \leqslant 1}(D^+ x + z) \\
&= \inf_{z \in \mathrm{Ker}(D)} \sigma_{\sigma_C(w) \leqslant 1}(D^+ x + z) \\
&= \inf_{z \in \mathrm{Ker}(D)} \gamma_C(D^+ x + z) \ ,
\end{aligned}$$

which concludes the proof. ∎

In particular, if $\mathrm{Ker}\, D = 0$, then $\gamma_{D(C)}(x) = \gamma_C(D^+ x)$. Using Lemma 2.4(v), one can observe that the infimum is bounded if $(D^+ x + \mathrm{Ker}(D)) \cap \mathrm{span}\, C \neq \emptyset$.

2.3.3 Lift to Matrix Spaces

We recall the singular value decomposition theorem.

> PROPOSITION 2.9 For any matrix $A \in \mathbb{R}^{n_1 \times n_2}$, there exists three matrices $U \in \mathbb{R}^{n_1 \times n_1}$, $\Sigma \in \mathbb{R}^{n_1 \times n_2}$, $V \in \mathbb{R}^{n_2 \times n_2}$ such that U and V are orthogonal matrices, Σ is empty outside its main diagonal and $A = U \Sigma V^*$. The matrix Σ is unique, up to permutation.

PROOF See (Horn et al. 2012, Theorem 7.3.3). ∎

Denoting $n = \min(n_1, n_2)$, we call the diagonal elements of Σ the singular values of A, denoted $(\sigma_i(A))_{1 \leqslant i \leqslant n}$. Thus we define a function $\sigma : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^n$ such that

$$\sigma_1(A) \geqslant \sigma_2(A) \geqslant \ldots \geqslant \sigma_n(A).$$

We start by the following definition

DEFINITION 2.29 A function $f : \mathbb{R}^n \to \mathbb{R}$ is an *absolutely symmetric gauge* if f is a gauge and is absolutely symmetric, i.e.

$$\forall x \in \mathbb{R}^n, \forall P \in \mathbb{P}_n, \quad f(Px) = f(x),$$

where $\mathbb{P}_n$ is the set of all signed permutation matrices of $\{1, \dots, n\}$,

The following proposition makes a connection between absolutely symmetric gauges and unitarily invariant norms, i.e. norms F such that $F(U\Lambda_x V^*) = F(\Lambda_x)$.

PROPOSITION 2.10 There is a one-to-one correspondance between absolutely symmetric gauges and unitarily invariant norms. More precisely,

(i) If f is an absolutely symmetric gauge, then $F = f \circ \sigma$ is a unitarily invariant norm.

(ii) If F is a unitarily invariant norm, then $f = F \circ \text{diag}$ is an absolutely symmetric gauge.

PROOF The proof might be found in (Von Neumann 1961) or (Horn et al. 2012, Theorem 7.4.7.2, p 464). ∎

For instance, the nuclear norm is nothing more than a unitarily invariant norm induced by the ℓ^1-norm $\|\cdot\|_1$ and the spectral norm is induced by the ℓ^∞-norm.

2.3.4 Operator Bounds

Since we use more general regularizers than norms, we have to generalize the concept of operator norm. Recall that if $(V, ||\cdot||)$ is a normed vector space, we embedded the set of continuous linear operators from V to W using the operator norm, i.e.

$$|||A||| = \sup_{||x|| \leqslant 1} ||Ax|| = \sup_{||x||=1} ||Ax|| = \sup_{x \in V} \frac{||Ax||}{x}.$$

This motivates the following Definition 2.30.

> DEFINITION 2.30 Let J_1 and J_2 be two gauges defined on two vector spaces V_1 and V_2, and $A : V_1 \to V_2$ a linear map. The *operator bound* $\|A\|_{J_1 \to J_2}$ of A between J_1 and J_2 is given by
>
> $$\|A\|_{J_1 \to J_2} = \sup_{J_1(x) \leqslant 1} J_2(Ax).$$

Note that $\|A\|_{J_1 \to J_2} < +\infty$ if, and only if $A \operatorname{Ker}(J_1) \subseteq \operatorname{Ker}(J_2)$. In particular, if J_1 is coercive (i.e. $\operatorname{Ker} J_1 = \{0\}$ from Lemma 2.4(iv)), then $\|A\|_{J_1 \to J_2}$ is finite. As a convention, $\|A\|_{J_1 \to ||\cdot||_p}$ is denoted as $\|A\|_{J_1 \to \ell^p}$. An easy consequence of this definition is the fact that for every $x \in V_1$,

$$J_2(Ax) \leqslant \|A\|_{J_1 \to J_2} J_1(x).$$

2.4 O-minimality

The goal of o-minimal geometry is to prevent pathological behavior with respect to the common operations on sets, such as addition and projection. To expose our motivation, we take the example of (Coste 1999). Consider the function $f : x \mapsto \sin \frac{1}{x}$ defined on $\mathbb{R}_+^*$ and $\mathcal{G}$ its graph on $\mathbb{R}^2$. Denote $\bar{\mathcal{G}}$ the closure of $\mathcal{G}$ in $\mathbb{R}^2$. Then, $\dim(\bar{\mathcal{G}} \setminus \mathcal{G}) = \dim \mathcal{G}$ in the Hausdorff sense. This is typically this kind of behavior that we wish to avoid.

2.4.1 Definition

We briefly recall here the definition and the main properties of o-minimal structures, that are used for our proof. We refer to (Dries 1998; Coste 1999) for more details about o-minimal structures.

O-minimal geometry can been seen as a generalization of the notion of semi-algebraicity.

DEFINITION 2.31 — SEMI-ALGEBRAIC SUBSETS. The *semi-algebraic* subsets of $\mathbb{R}^n$ are the smallest set $\mathcal{SA}_n$ of subsets of $\mathbb{R}^n$ such that:

(i) For every real polynomial $P \in \mathbb{R}[X_1, \dots, X_n]$,

$$\{x \in \mathbb{R}^n \mid P(x) = 0\} \in \mathcal{SA}_n \quad \text{and} \quad \{x \in \mathbb{R}^n \mid P(x) > 0\} \in \mathcal{SA}_n.$$

(ii) If $A, B \in \mathcal{SA}_n$, then $A \cup B, A \cap B, \mathbb{R}^n \setminus A \in \mathcal{SA}_n$.

The following result is central in the study of semi-algebraic sets.

THEOREM 2.2 — TARSKI-SEIDENBERG. The set $\mathcal{SA}_n$ of semi-algebraic sets is closed under projection.

We now define o-minimal structures.

DEFINITION 2.32 — STRUCTURE. An *o-minimal structure* $\mathcal{O}$ expanding $\mathbb{R}$ is a sequence of sets $(\mathcal{O}_n)_{n \in \mathbb{N}}$ which satisfies the following axioms:

(i) Each $\mathcal{O}_n$ is a Boolean algebra of subsets of $\mathbb{R}^n$, with $\mathbb{R}^n \in \mathcal{O}_n$.

(ii) Every semi-algebraic subset of $\mathbb{R}^n$ is in $\mathcal{O}_n$, i.e. $\mathcal{SA}_n \subseteq \mathcal{O}_n$.

(iii) If $A \in \mathcal{O}_n$ and $B \in \mathcal{O}_{n'}$, then $A \times B \in \mathcal{O}_{n+n'}$.

(iv) If $A \in \mathcal{O}_{n+1}$, then $\Pi(A) \in \mathcal{O}_n$, where $\Pi : \mathbb{R}^{n+1} \to \mathbb{R}^n$ is the projection on the first n.

(v) o-minimality: $\mathcal{O}_1$ is precisely the finite unions of intervals and points.

Definition 2.33 — 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 $\mathcal{G}(f) = \{(x, u) \in \Omega \times \mathbb{R}^p \mid u = f(x)\} \subseteq \mathbb{R}^n \times \mathbb{R}^p$ is a definable subset of $\mathbb{R}^n \times \mathbb{R}^p$.

Note that in this case, the application p times of axiom (iv) implies that Ω is definable. The fundamental example of o-minimal structure is the set of semi-algebraic sets, which is in some sense the smallest o-minimal structure. For example, note that in the special case where q is a rational number, the functionals $|| \cdot ||_q$ are actually semi-algebraic. When $q \in \mathbb{R}$ is not rational, then $|| \cdot ||_q$ is not semi-algebraic, but it can be shown to be definable in a o-minimal structure.

2.4.2 Properties

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 (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) and (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 translate the first order formula into operations on the involved subsets, in particular projections, and invoke the above stability axioms of o-minimal structures.

LEMMA 2.9 — ADDITION AND MULTIPLICATION. Let Ω a subset of $\mathbb{R}^n$. Let $f : \Omega \to \mathbb{R}^p$ 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 $S = \{(x, u, y, v, z, w) \mid x = y = z, u = v + w\}$ is obviously an algebraic (in fact linear) subset, hence definable by axiom 2. Property 1 implies 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 2.10 — INEQUALITIES IN DEFINABLE SETS. Let $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}$ be a definable function. Then $\{x \in \Omega \mid f(x) > 0\}$, 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 $B = \{(x, y) \in \mathbb{R} \times \mathbb{R} \mid f(x) = y\} \cap (\Omega \times (0, +\infty)$, which is definable thanks to axioms 1 and 3, and that the level sets of a definable function are also definable. Thus

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

and we conclude using again axiom 4. ∎

Yet another (simpler) proof.

PROOF (2) It is sufficient to remark that $\{x \in \Omega \mid f(x) > 0\}$ is the projection of the set $\left\{(x,t) \in \Omega \times \mathbb{R} \mid t^2 f(x) - 1 = 0\right\}$, where the latter is definable owing to Lemma 2.9. ∎

LEMMA 2.11 — 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 2.9. We now write the graph of f' as

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

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

$$B = \left\{(x,y,\nu,t,\varepsilon,\delta) \mid t^2 < \delta^2, (\nu - ty)^2 < \varepsilon^2 t^2\right\} \cap C \ .$$

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

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

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 2.11 — 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.

Lemma 2.12 — Marginal Function. Let $g : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ be a definable function, and Ω a definable subset of $\mathbb{R}^m$. The function

$$f(x) = \sup_{y \in \Omega} g(x, y)$$

is definable. The same conclusion holds true with inf instead of sup.

PROOF Let the subset

$$B = \{(x, u, y) \in \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^m \mid g(x, y) > u\} \cap (\mathbb{R}^n \times \mathbb{R} \times \Omega) \ .$$

B is definable thanks to Lemma 2.10 and axiom 1. Projecting on the components (x, u), we get the set

$$\Pi_{n+1+m,n+1}(B) = \{(x, u) \in \mathbb{R}^n \times \mathbb{R} \mid \exists y \in \Omega, g(x, y) > u\}$$

whose complement is

$$\mathbb{R}^{n+1} \setminus \Pi_{n+1+m,n+1}(B) = \{(x, u) \in \mathbb{R}^n \times \mathbb{R} \mid \forall y \in \Omega, g(x, y) \leqslant u\} = \text{epi}(f) \ ,$$

and therefore, $\text{epi}(f)$ is definable using axioms 4 and 1. Similarly, replacing $>$ with $<$ in B, we get that the hypograph $\text{hypo}(f)$ is definable. Thus $\mathcal{G}(f) = \text{hypo}(f) \cap \text{epi}(f)$ is definable by axiom 1. The proof for the inf is similar. ∎

As applications of this result, it follows that the Legendre-Fenchel conjugate of a definable function is definable, that the support function of a definable set is definable, and that the infimal convolution of definable functions is also definable.

Proposition 2.12 — Polars and Gauges. Let Ω be a closed convex subset of $\mathbb{R}^n$ containing the origin. Then the following assertions are equivalent:

(i) Ω is definable.

(ii) The polar set Ω° is definable.

(iii) The gauge f of Ω is definable.

(iv) The polar gauge f° is definable.

PROOF
(i) $\iff$ (ii): We have

$$\Omega^\circ = \{u \in \mathbb{R}^n \mid \forall x \in \Omega, \langle x, u\rangle \leqslant 1\} = \mathbb{R}^n \setminus \Pi_{2n,n}(\{(u,x) \in \mathbb{R}^n \times \Omega \mid \langle x, u\rangle > 1\}) .$$

We conclude that Ω° is definable using axioms 1-4. The converse statement follows by exchanging the roles of Ω and Ω°.

(i) $\iff$ (iii): f is the support function of Ω°

$$f(x) = \sigma_{\Omega^\circ}(x) = \sup_{u \in \Omega^\circ} \langle x, u\rangle ,$$

We get that f is definable using (i)⇒(ii), and applying Lemma 2.12 with $g(x,u) = \langle x, u\rangle$, which is obviously definable by axiom 2. The converse statement follows from Lemma 2.10 since $\Omega^\circ = \{x \in \mathbb{R}^n \mid f(x) \leqslant 1\}$ and (ii)⇒(i).

(ii) $\iff$ (iv): The proof follows exactly the lines of the previous item replacing f and Ω with their polars. ■

Proposition 2.13 — Subdifferential of a Gauge. Let f be the gauge of a closed convex subset Ω of $\mathbb{R}^n$ containing the origin as an interior point. Suppose that f is definable. Then for any $x \in \mathbb{R}^n$, the subdifferential $\partial f(x)$ is definable.

PROOF Let Ω° be the polar set of Ω. We have $\forall x \in \mathbb{R}^n$

$$\partial f(x) = \mathrm{Argmax}_{u \in \Omega^\circ} \langle x, u\rangle = \{u \in \Omega^\circ \mid \langle x, u\rangle = f(x)\},$$

i.e. the exposed face of Ω° associated with x, which is a non-empty convex compact set for all x. Thus, since f is definable, so is Ω° by Proposition 2.12, and $\partial f(x)$ is also definable by axioms 1-2 (the intersection of Ω° and a linear set). ■

Lemma 2.13 — Graph of the Relative Interior. Let f be the gauge of a closed convex subset Ω of $\mathbb{R}^n$ containing the origin in its interior. Suppose that f is definable. Then, the set

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

is definable.

PROOF Denote $C = \{(x,\eta) \mid \eta \in \operatorname{ri} \partial f(x)\}$. Combining the characterization of the relative interior of a convex set (Rockafellar 1996, Theorem 6.4) and the structure of the subdifferential of a bounded gauge, which is non-empty convex, see Proposition 2.13), we rewrite C in the more convenient form

$$C = \left\{ (\beta,\eta) \mid \forall u \in \Omega^\circ \text{ and } \langle u,\, \beta\rangle = f(\beta), \exists t > 1 \text{ s.t } \begin{cases} (1-t)u + t\eta \in \Omega^o \\ \langle \eta,\, \beta\rangle = f(\beta) \end{cases} \right\}.$$

Let $B_u = (\mathbb{R}^n \setminus \Omega^\circ) \cup (\mathbb{R}^n \setminus \{u \mid \langle u,\, \beta\rangle = f(\beta)\})$ and

$$\begin{aligned} B =& (\mathbb{R}^n \times \mathbb{R}^n \times B_u \times]1, +\infty[\times \Omega^\circ) \\ & \cap \{(\beta,\eta,u,t,\xi) \mid \langle \eta,\, \beta\rangle = f(\beta),\ \xi = (1-t)u + t\eta\} \ . \end{aligned}$$

B_u is definable by virtue of Proposition 2.12, axiom 1, and the fact it involves algebraic relations and the level sets of a definable function. In turn, B is definable owing to Proposition 2.12 and axioms 1-3. It then results that

$$C = \mathbb{R}^n \setminus \Pi_{4n+1,2n}(B) \ ,$$

which is then definable after axioms 4 and 1. ∎

Part I

Models, Partial Smoothness and Dual Certificates

3

Model Tangent Subpace

Main contributions of this chapter

- Introduction of the model tangent space and model vector in Definition 3.1 and the subdifferential gauge in Definition 3.2.
- Theorem 3.1 provides a pointwise decomposition of the subdifferential of any function in $\Gamma_c^+(\mathbb{R}^n)$.

Contents

THE purpose of this chapter is to introduce one of the main concepts used throughout this thesis, the model tangent subspace associated to a convex function. The main result, Theorem 3.1 of this chapter, proves that the subdifferential of any convex function exhibits some kind of decomposability property.

It is known that the subdifferential, see for instance (Fuchs 2004), of the ℓ^1-norm exhibits a particular structure. More precisely, for any $x \in \mathbb{R}^n$,

$$\partial\|\cdot\|_1(x) = \{\eta \in \mathbb{R}^n \mid \eta_I = \operatorname{sign}(x)_I \quad \text{and} \quad \|\eta_{I^c}\|_\infty \leqslant 1\},$$

where $I = \operatorname{supp}(x)$, the support of x. Such a structure is very convenient when dealing with $(\mathcal{P}_{y,\lambda})$ or $(\mathcal{P}_{y,0})$, since it allows to split the analysis between "active" components of x and non-active. One may ask if such a splitting is always possible for a convex function. More precisely, we aim to split the subdifferential onto an orthogonal decomposition of $\mathbb{R}^n$ in a structured way. We answer positively in the following sections, replacing the support I by what we coined the *tangent model subspace*, the sign pattern $\operatorname{sign}(x)$ by the *model vector* and the ℓ^∞-norm by the *subdifferential gauge*. However, this decomposition will be pointwise, an issue that is solved in Chapter 4.

3.1 Model Tangent Subpace Associated to a Convex Function

This section introduces the model tangent subspace associated to a convex function, and its associated model vector. We can compute the model tangent space of a sum of convex functions at some point, and also by precomposition by a linear operator. This section is illustrated by two examples, the $\ell^1 - \ell^2$ norm and the ℓ^∞ norm. Section 3.3 provides several other examples to connection these definition to practical applications.

The terminology of model tangent subspace is partly explained in this chapter. Indeed, the tangent aspect is a consequence of a further property, partial smoothness, that is studied in Chapter 4. Nevertheless, we stick with this name right now.

3.1.1 Model Tangent Subspace

Let $J \in \Gamma_c^+(\mathbb{R}^n)$ a regularizer[1], i.e. a continuous, bounded from below, proper, and convex function. We now introduce the model tangent subspace at a point x.

Definition 3.1 — Model Tangent Subspace. For any vector $x \in \mathbb{R}^n$, we denote by $\bar{S}_x$ the affine hull of the subdifferential of J at x

$$\bar{S}_x = \operatorname{aff} \partial J(x),$$

and e_x, its *model vector*, the orthogonal projection of 0 onto $\bar{S}_x$

$$e_x = \operatorname*{argmin}_{e \in \bar{S}_x} ||e||.$$

We denote

$$S_x = \bar{S}_x - e_x = \operatorname{span}(\partial J(x)) \quad \text{and} \quad T_x = S_x^\perp.$$

T_x is coined the *model tangent subspace* of x associated to J.

1. The boundness assumption does not play any role in section. It will however be important in our results in Chapters 6–9.

When J is Gâteaux-differentiable at x, i.e. $\partial J(x) = \{\nabla J(x)\}$, $e_x = \nabla J(x)$ and $T_x = \mathbb{R}^N$.

On the contrary, when J is not smooth at x, the dimension of T_x is of smaller dimension, and the regularizing function J essentially promotes elements living on or close to this model subspace.

We start by summarizing some key properties of e_x and T_x.

Proposition 3.1 Let $J \in \Gamma_c^+(\mathbb{R}^n)$. For any $x \in \mathbb{R}^N$, one has

(i) $e_x \in T_x \cap \bar{S}_x$,

(ii) $\bar{S}_x = \left\{\eta \in \mathbb{R}^N \mid \eta_{T_x} = e_x\right\}$.

Moreover, if J is a gauge, then

(iii) For every $u \in \bar{S}_x$, $J(x) = \langle u, x \rangle$,

(iv) $x \in T_x$.

PROOF (i) This is due to the fact that e_x is the orthogonal projection of 0 on the affine space $\bar{S}_x$. It is therefore an element of $\bar{S}_x \cap (\bar{S}_x - e_x)^\perp$, i.e. $e_x \in \bar{S}_x \cap T_x$.

(ii) This is straightforward from the fact that $S_x = \left\{\eta \in \mathbb{R}^N \mid \eta_{T_x} = 0\right\}$, $\bar{S}_x = S_x + e_x$ and $e_x \in T_x$ from (i).

(iii) Each element of $\bar{S}_x$ can be written as $u = \sum_{i=1}^k \rho_i \eta_i$, for $k > 0$, where $\eta_i \in \partial J(x)$ and $\sum_{i=1}^k \rho_i = 1$. By Fenchel identity applied to the gauge J, and using Lemma 2.5(iii), we have

$$\langle x, \eta_i \rangle = J(x) + \iota_{C^\circ}(\eta_i), \quad \forall i \; .$$

Since $\eta_i \in \partial J(x) \subseteq C^\circ$, we get

$$\langle x, \eta_i \rangle = J(x), \quad \forall i \; .$$

Multiplying by ρ_i and summing this identity over i and using the fact that $\sum_{i=1}^{k} \rho_i = 1$ we obtain the desired result.

(iv) For any $v \in S_x$, we have $v + e_x \in \bar{S}_x$ since $e_x \in \bar{S}_x$. Thus applying (i), we get $\langle x, e_x + v \rangle = J(x)$ and $\langle x, e_x \rangle = J(x)$. Combining both identities implies that $\langle x, v \rangle = 0$, $\forall v \in S_x$, or equivalently that $x \in S_x^{\perp} = T_x$. ■

In general $e_x \notin \partial J(x)$, which is the situation displayed on Figure 3.1. In this figure, x is an element of T_x which is not the case for a convex function. This is however always the case if J is a gauge, as stated in Proposition 3.1(iv).

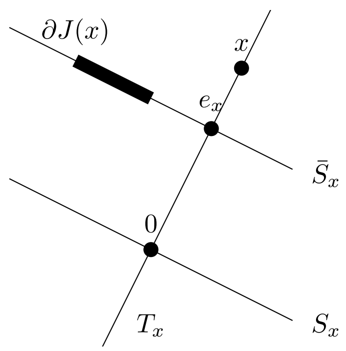


Figure 3.1: *Illustration of the geometrical elements* (S_x, T_x, e_x) *for a gauge.*

3.1.2 Algebraic Stability

The following proposition determines the model tangent subspace of the sum of two functions.

$$H = J + G$$

in terms of those associated to J and G.

Proposition 3.2 Let $J, G \in \Gamma_c^+(\mathbb{R}^n)$. Denote T^J and e_J (resp. T^G and e_G) the model tangent subspace and vector at a point x corresponding to J (resp. G). Then the model tangent subspace at x of $H = J + G$ reads

(i) $T^H = T^J \cap T^G$, or equivalently $S^H = (T^H)^{\perp} = \text{span}\left(S^J \cup S^G\right)$.

(ii) $e_H = P_{T^H}(e_J + e_G)$.

PROOF (i) First, we have

$$\partial H(x) = \partial J(x) + \partial G(x),$$

Let $S^J = \text{span}(\partial J(x) - \eta^J)$ and $S^G = \text{span}(\partial G(x) - \eta^G)$, for any pair $\eta^J \in \partial J(x)$ and $\eta^G \in \partial G(x)$. Choosing $\eta^H = \eta^J + \eta^G \in \partial H(x)$ we have

$$\begin{aligned} S^H &= \text{span}(\partial H(x) - \eta^H) \\ &= \text{span}\left((\partial J(x) - \eta^J) + (\partial G(x) - \eta^G)\right) \\ &= \text{span}\left(\text{span}(\partial J(x) - \eta^J) + \text{span}(\partial G(x) - \eta^G)\right) \\ &= \text{span}(S^J \cup S^G). \end{aligned}$$

As a consequence we have $T^H = (S^H)^\perp = T^J \cap T^G$.

(ii) Moreover, since $T^H \perp S^J \cup S^G$ we have from Proposition 3.1(iii) that

$$\begin{aligned} e_H = P_{T^H}(\partial H(x)) &= P_{T^H}(\partial J(x) + \partial G(x)) \\ &= P_{T^H}\left(e_J + P_{S^J}\,\partial J(x) + e_G + P_{S^G}\,\partial G(x)\right) \\ &= P_{T^H}(e_J + e_G). \end{aligned}$$

■

Functions of the form $J_0 \circ D^*$, where $J_0 \in \Gamma_c^+(\mathbb{R}^n)$ is a bounded regularizing convex function, correspond to the so-called analysis-type regularizers. In the following, we denote $T = T_x = S^\perp$ and $e = e_x$ the subspace and vector in the decomposition of the subdifferential of J at a given $x \in \mathbb{R}^N$. Analogously, $T_0 = S_0^\perp$ and e_0 are those of the function J_0 at D^*x. The following proposition details the decomposability structure of analysis-type functions.

PROPOSITION 3.3 With the above notation, the model tangent subspace of $J = J_0 \circ D^*$ reads

(i) $T = \text{Ker}(D^*_{S_0}) = D^* T_0$, or equivalently $S = \text{Im}(D_{S_0}) = DS_0$.

(ii) $e = P_T\, De_0$.

PROOF

(i) One has $\partial J = D \circ \partial J_0 \circ D^*$, hence $S = DS_0 = \mathrm{Im}(D_{S_0})$ and $T = S^\perp = \mathrm{Ker}(D^*_{S_0})$.

(ii) As $S = D\bar{S_0} = De_0 + S$, we get from Proposition 3.1

$$\begin{aligned} e \in \operatorname*{argmin}_{z \in \bar{S}} \|z\| &= \operatorname*{argmin}_{z - De_0 \in S} \|z\| \\ &= De_0 + \operatorname*{argmin}_{h \in S} \|h + De_0\|. \end{aligned}$$

The second term is the projection of $-De_0$ onto the linear subspace S. Thus,

$$\begin{aligned} e &= De_0 + P_S(-De_0) \\ &= (\mathrm{Id} - P_S)De_0 \\ &= P_T\, De_0 \ , \end{aligned}$$

which is the result stated. ■

It is common in the litterature (Zou et al. 2005) to find regularization of the form $J_\varepsilon(x) = J(x) + \frac{\varepsilon}{2}\|x\|_2^2$ in order to stabilize the numerical optimization. More generally, we consider a function G which is Gâteaux-differentiable.

COROLLARY 3.1 Let $J \in \Gamma_c^+(\mathbb{R}^n)$, $x \in \mathbb{R}^n$ and $G \in \Gamma_c^+(\mathbb{R}^n)$ a function which is Gâteaux-differentiable at x. Then,

$$T_x^{J+G} = T^J \quad \text{and} \quad e_x^{J+G} = e_x^J + P_{T_x^J}\nabla G(x).$$

PROOF Indeed, since $\partial G(x) = \{\nabla G(x)\}$, we obtain that $T_x^G = \mathbb{R}^n$ and $e_x^G = \nabla G(x)$. Applying Proposition 3.2, we get the result. ■

Hence, the model tangent space does not vary with the pertubation G (unlike e_x). Remark that the function $G : x \mapsto \frac{\varepsilon}{2}\|x\|_2^2$ is C^∞ everywhere. If J is a gauge, hence $x \in T_x$, we get that the model vector reads $e_x^{J+G} = e_x^J + \varepsilon x$.

3.1.3 Examples

We illustrate the definition of model tangent subspace with two norms, the ℓ^1-ℓ^2 norm used in structured sparsity and the ℓ^∞-norm used for spread representations as discussed in the introduction.

The ℓ^1-ℓ^2 norm. We consider a uniform disjoint partition $\mathcal{B}$ of $\{1, \cdots, n\}$,

$$\{1, \ldots, N\} = \bigcup_{b \in \mathcal{B}} b, \quad b \cap b' = \emptyset, \ \forall b \neq b' \ .$$

The $\ell^1 - \ell^2$ norm of x is

$$J(x) = ||x||_{\mathcal{B}} = \sum_{b \in \mathcal{B}} ||x_b||.$$

PROPOSITION 3.4 Let $J = ||\cdot||_{\mathcal{B}}$. The tangent model space of J at $x \neq 0$ reads

$$T_x = \left\{\eta \in \mathbb{R}^N \mid \ \forall b \notin I, \eta_b = 0\right\},$$

where $I = \{b \in \mathcal{B} \mid \ x_b \neq 0\}$, and its orthogonal S_x reads

$$S_x = \bar{S}_x - e_x = \{\eta \in \mathbb{R}^n \mid \ \forall b \in I, \eta_b = 0\}.$$

It model vector reads

$$e_x = (\mathcal{N}(x_b))_{b \in \mathcal{B}},$$

where $\mathcal{N}(a) = a/||a||$ if $a \neq 0$, and $\mathcal{N}(0) = 0$.

PROOF The subdifferential of J at $x \in \mathbb{R}^n$ is

$$\partial J(x) = \left\{\eta \in \mathbb{R}^n \mid \ \forall b \in I, \eta_b = \frac{x_b}{||x_b||} \quad \text{and} \quad \forall g \notin I, ||\eta_g|| \leqslant 1\right\}.$$

Thus, the affine hull of $\partial J(x)$ reads

$$\bar{S}_x = \left\{\eta \in \mathbb{R}^n \mid \ \forall b \in I, \eta_b = \frac{x_b}{||x_b||}\right\}.$$

Hence the projection of 0 onto $\bar{S}_x$ is

$$e_x = (\mathcal{N}(x_b))_{b \in \mathcal{B}}$$

and

$$S_x = \bar{S}_x - e_x = \{\eta \in \mathbb{R}^n \mid \forall b \in I, \eta_b = 0\},$$

which completes the proof. ∎

The ℓ^∞ norm. The ℓ^∞ norm is $J(x) = ||x||_\infty = \max_{1 \leqslant i \leqslant n} |x_i|$.

PROPOSITION 3.5 Let $J = ||\cdot||_\infty$. The tangent model space of J at $x \neq 0$ reads

$$T_x = \left\{\eta \mid \eta_{(I)} = \rho(\text{sign}(x))_{(I)} \quad \text{for} \quad \rho \in \mathbb{R}\right\},$$

where $I = \{i \in \{1, \dots, n\} \mid |x_i| = ||x||_\infty\}$ and e_x is defined by $s_{(I)} = (\text{sign}(x))_{(I)}$ and $s_{(I^c)} = 0$.

PROOF For $x = 0$, $\partial J(x)$ is the unit ℓ^1 ball, hence $\bar{S}_x = S_x = \mathbb{R}^N$, $T_x = \{0\}$ and $e_x = 0$.

For $x \neq 0$, we have

$$\partial J(x) = \{\eta \mid \forall i \in I^c, \eta_i = 0, \langle \eta, s \rangle = 1, \eta_i s_i > 0 \, \forall i \in I\} .$$

It is clear that $\bar{S}_x$ is the affine hull of an $|I|$-dimensional face of the unit ℓ^1 ball exposed by the sign subvector $s_{(I)}$. Thus e_x is the barycenter of that face, i.e.

$$e_x = s/|I| \quad \text{and} \quad S_x = \left\{\eta \mid \eta_{(I^c)} = 0 \quad \text{and} \quad \langle \eta_{(I)}, s_{(I)} \rangle = 0\right\} .$$

In turn, we have the expression of T_x. ∎

3.2 The Decomposability Property

In the previous section, we defined the model tangent subspace and the model vector. They are going to play a key role in structuring the subdifferential of J.

The following proposition gives an equivalent convenient description of the subdifferential of a *gauge* γ_C at x in terms of a particular supporting hyperplane to C°: the affine hull $\bar{S}_x$.

PROPOSITION 3.6 Let γ_C be a finite-valued gauge. Then for $x \in \mathbb{R}^N$, one has

$$\partial\gamma_C(x) = \bar{S}_x \cap C^\circ.$$

PROOF Let $x \in \mathbb{R}^N$. We have

$$\partial\gamma_C(x) = \mathbf{F}_{C^\circ}(x) = H \cap C^\circ,$$

where $H = \left\{\eta \in \mathbb{R}^N \mid \langle \eta, x \rangle = J(x)\right\}$ is the supporting hyperplane of C° at x. By Proposition 3.1(iii), we have

$$\bar{S}_x = \operatorname{aff} \partial\gamma_C(x) \subseteq H,$$

which implies that

$$\bar{S}_x \cap C^\circ \subseteq H \cap C^\circ.$$

The converse inclusion is true since $\partial\gamma_C(x) = H \cap C^\circ \subseteq \bar{S}_x$. ∎

Note that this property holds only for gauges. In the following, we propose an alternative for any kind of convex function.

3.2.1 The Subdifferential Gauge

3.2.1.1 Definition

Before providing an equivalent description of the subdifferential of J at x in terms of the geometrical objects e_x, T_x and S_x, we introduce a gauge that plays a prominent role in this description.

DEFINITION 3.2 — SUBDIFFERENTIAL GAUGE. Let $J \in \Gamma_c^+(\mathbb{R}^n)$. Let $x \in \mathbb{R}^N \setminus \{0\}$ and $f_x \in \operatorname{ri} \partial J(x)$. The *subdifferential gauge* associated to f_x is the gauge $J_{f_x}^{x,\circ} = \gamma_{\partial J(x) - f_x}$.

Since $\partial J(x) - f_x$ is a closed (in fact compact) convex set containing the origin as a relative interior point, it is uniquely characterized by the subdifferential gauge $J_{f_x}^{x,\circ}$ (see Lemma 2.4(i)).

The following proposition states the main properties of the gauge $J_{f_x}^{x,\circ}$.

PROPOSITION 3.7 The subdifferential gauge $J_{f_x}^{x,\circ}$ is such that $\operatorname{dom} J_{f_x}^{x,\circ} = S_x$, and is coercive on S_x. Moreover, if J is a gauge, then

$$J_{f_x}^{x,\circ}(\eta) = \inf_{\tau \geqslant 0} \max(J^\circ(\tau f_x + \eta), \tau) + \iota_{S_x}(\eta) \ .$$

PROOF The first assertion follows from Lemma 2.4(v) since $0 \in \operatorname{ri}(\partial J(x) - f_x)$. Let's now turn to the second part. Since $f_x \in \operatorname{ri} \partial J(x) \subset \bar{S}_x$, Proposition 3.1 implies that $f_x = P_{S_x}(f_x) + P_{T_x}(f_x) = P_{S_x}(f_x) + e_x$. Hence, using Proposition 3.6, we get

$$\begin{aligned} \partial J(x) - f_x &= (C^\circ - f_x) \cap (\bar{S}_x - f_x) \\ &= (C^\circ - f_x) \cap (S_x - \{P_{S_x}(f_x)\}) \\ &= (C^\circ - f_x) \cap S_x \ . \end{aligned}$$

We therefore obtain

$$\begin{aligned}
J_{f_x}^{x,\circ}(\eta) &= \gamma_{(C^\circ - f_x)\cap S_x}(\eta) \\
&= \max(\gamma_{C^\circ - f_x}(\eta), \gamma_{S_x}(\eta)) \\
&= \max(\gamma_{C^\circ - f_x}(\eta), \iota_{S_x}(\eta)) \\
&= \gamma_{C^\circ - f_x}(\eta) + \iota_{S_x}(\eta) \ .
\end{aligned}$$

At this stage, Lemma 2.7 does not apply straightforwardly since $0 \in C^\circ$ but $f_x \neq 0$ in general. However, proceeding as in the proof of that lemma, we arrive at

$$\gamma_{C^\circ + \{-f_x\}}(\eta) = \sup_{\rho \in [0,1]} \rho J^\circ \overset{+}{\vee} (1-\rho)\sigma_{\{-f_x\}^\circ}(\eta)$$

where, from Definition 2.27, $\{-f_x\}^\circ = \{\eta \mid \langle \eta, f_x \rangle \geqslant -1\}$, which indeed contains the origin as an interior point. Continuing from the last equality, we get

$$\begin{aligned}
\gamma_{C^\circ + \{-f_x\}}(\eta) &= \sup_{\rho \in [0,1]} \rho J^\circ \overset{+}{\vee} (1-\rho)\gamma_{\{-f_x\}^{\circ\circ}}(\eta) \\
&= \sup_{\rho \in [0,1]} \rho J^\circ \overset{+}{\vee} (1-\rho)\gamma_{\overline{\mathrm{co}}(\{-f_x\}\cup\{0\})}(\eta) \\
&= \sup_{\rho \in [0,1]} \rho J^\circ \overset{+}{\vee} (1-\rho)\gamma_{\{-\mu f_x \mid \mu \in [0,1]\}}(\eta) \ .
\end{aligned}$$

It is easy to see that

$$\gamma_{\{-\mu f_x \mid \mu\in[0,1]\}}(-\eta) = \begin{cases} \tau & \text{if } \eta \in \tau f_x, \tau \in \mathbb{R}_+ \ , \\ +\infty & \text{otherwise} \ . \end{cases}$$

Thus

$$\gamma_{C^\circ + \{-f_x\}}(\eta) = \sup_{\rho \in [0,1]} \inf_{\tau \geqslant 0} \rho J^\circ(\tau f_x + \eta) + (1-\rho)\tau \ .$$

Recalling that J° is a finite-valued gauge, hence continuous, the objective in the $\sup\inf$ fulfills the assumption of the second assertion of Lemma 2.7, whence we get

$$\gamma_{C^\circ+\{-f_x\}}(\eta) = \inf_{\tau\geqslant 0} \max(J^\circ(\tau f_x + \eta), \tau) \ ,$$

which completes the proof. ■

The second claim gives a formula which links $J_{f_x}^{x,\circ}$ to the polar gauge J°. But they are not equal in general unless some additional assumptions are imposed on J, as we will see shortly.

3.2.1.2 The Polar of the Subdifferential Gauge

We now turn to the gauge polar to the subdifferential gauge defined by the relation $(J_{f_x}^{x,\circ})^\circ = J_{f_x}^x$. $J_{f_x}^x$ comes into play in several results in the rest of the manuscript. The following proposition summarizes its most important properties.

PROPOSITION 3.8 The gauge $J_{f_x}^x$ is such that:

(i) Its has a full domain.

(ii) $J_{f_x}^x(d) = J_{f_x}^x(d_S) = \sup_{J_{f_x}^{x,\circ}(\eta_{S_x})\leqslant 1} \langle \eta_{S_x}, d\rangle$, where $S = S_x$.

(iii) $\operatorname{Ker} J_{f_x}^x = T_x$ and $J_{f_x}^x$ is coercive on S_x.

Moreover, if J is a gauge,

(iv) $J_{f_x}^x(d) = J(d_{S_x}) - \langle f_{S_x}, d_{S_x}\rangle$

PROOF The gauge $J_{f_x}^x$ is the support function of the set

$$\mathcal{K}_x \stackrel{\text{def.}}{=} \partial J(x) - f_x = \left\{\eta \in \mathbb{R}^N \mid J_{f_x}^{x,\circ}(\eta) \leqslant 1\right\} \subset S_x \ ,$$

where the inclusion follows from Proposition 3.7.

(i) Since $\mathcal{K}_x$ is a bounded set, its support function is also finite-valued (Hiriart-Urruty et al. 2001, Proposition V.2.1.3). It follows that $\operatorname{dom} J_{f_x}^x = \mathbb{R}^n$.

(ii) We have

$$\begin{aligned} J^x_{f_x}(d) &= \sup_{\eta \in \mathcal{K}_x} \langle \eta, d \rangle = \sup_{J^\circ_{f_x}(\eta) \leqslant 1} \langle \eta, d \rangle = \sup_{J^{x,\circ}_{f_x}(\eta_{S_x}) \leqslant 1} \langle \eta_{S_x}, d \rangle \\ &= \sup_{\eta \in \mathcal{K}_x} \langle \eta, d_{T_x} \rangle + \langle \eta, d_{S_x} \rangle = \sup_{\eta \in \mathcal{K}_x} \langle \eta, d_{S_x} \rangle \\ &= J^x_{f_x}(d_{S_x}) \ , \end{aligned}$$

where we used the fact that $\langle \eta, d_{T_x} \rangle = 0$ on $\mathcal{K}_x$.

(iii) As a consequence of (ii), $J^x_{f_x}(d_{T_x}) = 0$. Clearly, $T_x \subset \mathrm{Ker}(J^x_{f_x})$ and $J^x_{f_x}$ is constant along all affine subspaces parallel to T_x. But, since $0 \in \mathrm{ri}\,\mathcal{K}_x$, excluding the origin, the supremum in J_{f_x} is always attained at some nonzero $\eta \in \mathcal{K}_x \subset S_x$. Thus $J^x_{f_x}(d) > 0$ for all d such that $d \notin T_x$. This shows that actually $\mathrm{Ker}(J^x_{f_x}) = T_x$. In particular, this yields that on S_x, the gauge $J^x_{f_x}$ is coercive.

(iv) Using some calculus rules with support functions and assertion (ii), we have

$$\begin{aligned} J^x_{f_x}(d) = J^x_{f_x}(d_{S_x}) &= \sigma_{(C^\circ + \{-f_x\}) \cap S_x}(d_{S_x}) \\ &= \overline{\mathrm{co}} \left(\inf(\sigma_{C^\circ + \{-f_x\}}(d_{S_x}), \sigma_S(d_{S_x})) \right) \\ &= \overline{\mathrm{co}} \left(\inf(\sigma_{C^\circ + \{-f_x\}}(d_{S_x}), \iota_T(d_{S_x})) \right) \\ &= \sigma_{C^\circ + \{-f_x\}}(d_{S_x}) \\ &= \sigma_{C^\circ}(d_{S_x}) - \langle P_{S_x}(f_x), d_{S_x} \rangle \\ &= J(d_{S_x}) - \langle P_{S_x}(f_x), d_{S_x} \rangle \ , \end{aligned}$$

which completes the proof. ∎

3.2.2 Main Result

Piecing together the above ingredients yields a fundamental pointwise decomposition of the subdifferential of the regularizer J.

Theorem 3.1 — Decomposability. Let $J \in \Gamma_c^+(\mathbb{R}^n)$. Let $x \in \mathbb{R}^N \setminus \{0\}$ and $f_x \in \operatorname{ri} \partial J(x)$. Then the subdifferential of J at x reads

$$\partial J(x) = \left\{ \eta \in \mathbb{R}^N \mid \eta_{T_x} = e_x \quad \text{and} \quad J_{f_x}^{x,\circ}(P_{S_x}(\eta - f_x)) \leqslant 1 \right\} .$$

PROOF Invoking Proposition 3.1, we get that for every $\eta \in \partial J(x)$, $\eta_{T_x} = e_x$, and $P_{T_x}(f_x) = e_x$. It remains now to uniquely characterize the part of the subdifferential lying in S_x, i.e. $\partial J(x) - e_x$. Since $f_x \in \operatorname{ri} \partial J(x)$, we have from the one-to-one correspondence of Lemma 2.4(i) and the definition of the subdifferential gauge,

$$\begin{aligned} \eta \in \left\{ \eta \in \mathbb{R}^N \mid J_{f_x}^{x,\circ}(\eta_{S_x} - P_{S_x}(f_x)) \leqslant 1 \right\} &\iff \eta_{S_x} - P_{S_x}(f_x) \in \partial J(x) - f_x \\ &\iff \eta_{S_x} \in \partial J(x) - e_x \\ &\iff \eta \in \partial J(x) \ , \end{aligned}$$

which completes the proof. ∎

Capitalizing on Theorem 3.1, we are now able to deduce a convenient necessary and sufficient first-order (global) minimality condition of $(\mathcal{P}_{y,\lambda})$ and $(\mathcal{P}_{y,0})$.

Proposition 3.9 Let $x \in \mathbb{R}^n$, and denote for short $T = T_x$ and $S = S_x$. The two following propositions hold.

(i) The vector x is a global minimizer of $(\mathcal{P}_{y,\lambda})$ if, and only if,

$$\Phi_T^*(y - \Phi x) = \lambda e_x \quad \text{and} \quad J_{f_x}^{x,\circ}(\lambda^{-1}\Phi_S^*(y - \Phi x) - P_S(f_x)) \leqslant 1.$$

(ii) The vector x is a global minimizer of $(\mathcal{P}_{y,0})$ if, and only if, there exists a dual vector $\mathfrak{p} \in \mathbb{R}^q$ such that

$$\Phi_T^* \mathfrak{p} = e_x \quad \text{and} \quad J_{f_x}^{x,\circ}(\Phi_S^* \mathfrak{p} - P_S(f_x)) \leqslant 1.$$

PROOF This is a convenient rewriting of the fact that x is a global minimizer if, and only if, 0 is a subgradient of the objective function at x.

(i) For problem $(\mathcal{P}_{y,\lambda})$, this is equivalent to

$$\frac{1}{\lambda}\Phi^*(y - \Phi x) \in \partial J(x).$$

Projecting this relation on T and S yields the desired result.

(ii) Let's turn to problem $(\mathcal{P}_{y,0})$. We have at any global minimizer x

$$0 \in \partial J(x) + \Phi^* N_{\{p|\ p=y\}}(\Phi x)$$

where $N_{\{p|\ p=y\}}(x)$ is the normal cone of the constraint set $\{p \mid\ p = y\}$ at x, which is obviously the whole space $\mathbb{R}^q$. Thus, this monotone inclusion is equivalent to the existence of $p \in \mathbb{R}^q$ such that

$$\Phi^* p \in \partial J(x)\ .$$

Projecting again this on T and S proves the assertion. ∎

These results can be extended easily when $\frac{1}{2}||y - \Phi x||^2$ is replaced by an other data fidelity term.

3.2.3 Decomposability of the Sum and Precomposition by a Linear Operator

Following the same path as for the model space, we establish the subdifferential gauge in the case of the sum and the precomposition by a linear operator. We recall that (Proposition 3.2) denoting T^J and e_J (resp. T^G and e_G) the model tangent subspace and vector at a point x corresponding to J (resp. G), we proved that $T^H = T^J \cap T^G$ and $e_H = P_{T^H}(e_J + e_G)$. The following proposition describes the subdifferential gauge of $H = J + G$.

Proposition 3.10 Let $J, G \in \Gamma_c^+(\mathbb{R}^n)$. Let $J^{x,\circ}_{f^J_x}$ and $G^{x,\circ}_{f^G_x}$ denote the subdifferential gauges for the pairs $(J, f^J_x \in \operatorname{ri}\partial J(x))$ and $(G, f^G_x \in \operatorname{ri}\partial G(x))$, correspondingly. Then, for the particular choice of

$$f^H_x = f^J_x + f^G_x$$

we have $f_x^H \in \operatorname{ri} \partial H(x)$, and for a given $\eta \in S^H$, the subdifferential gauge of H reads

$$H^{x,\circ}_{f_x^H}(\eta) = \inf_{\eta_1+\eta_2=\eta} \max(J^{x,\circ}_{f_x^J}(\eta_1), G^{x,\circ}_{f_x^G}(\eta_2)) \ .$$

PROOF As $f_x^J \in \operatorname{ri} \partial J(x)$ and $f_x^G \in \operatorname{ri} \partial G(x)$, it follows from (Rockafellar 1996, Corollary 6.6.2) that

$$f_x^H = f_x^J + f_x^G \in \operatorname{ri} \partial J(x) + \operatorname{ri} \partial G(x) = \operatorname{ri} (\partial J(x) + \partial G(x)) = \operatorname{ri} \partial H(x) \ .$$

The subdifferential gauge associated to H is then

$$H^{x,\circ}_{f_x^H} = \gamma_{\partial H(x) - f_x^H} = \gamma_{(\partial J(x) - f_x^J) + (\partial G(x) - f_x^G)} \ ,$$

which is coercive and finite-valued on S^H according to Proposition 3.7. Invoking Lemma 2.7, we get the desired result since for any $\rho \geqslant 0$,

$$u \mapsto \rho J^{x,\circ}_{f_x^J}(u) + (1-\rho) G^{x,\circ}_{f_x^G}(\eta - u) = \rho \gamma_{\partial J(x) - f_x^J}(u) + (1-\rho) \gamma_{\partial G(x) - f_x^G}(\eta - u)$$

is finite-valued and continuous on $S^J \cap (S^G + \eta)$, for $\eta \in S^H = \operatorname{span}(S^J + S^G)$. ■

Similarly, we derive the expression of the subdifferential gauge for an analysis-type prior. In this case, according to Proposition 3.3, the tangent model space reads $T = \operatorname{Ker}(D^*_{S_0}) = D^* T_0$ and its model vector $e = P_T D e_0$.

PROPOSITION 3.11 Let $J_0 \in \Gamma_c^+(\mathbb{R}^p)$. Let $J^{D^*x,\circ}_{0,f_{D^*x}}$ denote the subdifferential gauge for the pair $(J_0, f_{0,D^*x} \in \operatorname{ri} \partial J_0(x))$. Then, for the particular choice of

$$f_x = D f_{0,D^*x}$$

we have $f_x \in \operatorname{ri} \partial J(x)$, $\operatorname{dom} J^{x,\circ}_{f_x} = S$ and for every $\eta \in S$

$$J^{x,\circ}_{f_x}(\eta) = \inf_{z \in \operatorname{Ker}(D_{S_0})} J^{D^*x,\circ}_{0,f_{D^*x}}(D^+_{S_0}\eta + z) \ .$$

The infimum can be equivalently taken over $\operatorname{Ker}(D) \cap S_0$.

PROOF With such a choice of f_x, we have

$$f_{0,D^*x} \in \operatorname{ri}\partial J_0(D^*x) \Rightarrow Df_{0,D^*x} \in D\operatorname{ri}\partial J_0(D^*x)$$

$$\iff f_x \in \operatorname{ri} D\partial J_0(D^*x) \iff f_x \in \operatorname{ri}\partial J(x) \ .$$

We follow the same lines as in the proof of Lemma 2.8, where we additionally invoke Proposition 3.8(ii) to get

$$\begin{aligned} J_{f_x}^x(d) &= \sigma_{\partial J(x)-f_x}(d) \\ &= \sigma_{D\left(\partial J_0(D^*x)-f_{0,D^*x}\right)}(d) \\ &= \sigma_{\partial J_0(D^*x)-f_{0,D^*x}}(D^*d). \end{aligned}$$

Identifying $\sigma_{\partial J_0(D^*x)-f_{0,D^*x}}$ with the gauge $\gamma_{\partial J_0(D^*x)-f^\circ_{0,D^*x}}$, we get

$$\begin{aligned} J_{f_x}^x(d) &= J_{0,f_{0,D^*x}}^{D^*x}(D^*d) \\ &= J_{0,f_{0,D^*x}}^{D^*x}(D_{S_0}^*d) \ . \end{aligned}$$

Note that $J_{f_x}^x$ is indeed constant along affine subspaces parallel to $\operatorname{Ker}(D_{S_0}^*) = S^\perp = T$. We now get that for every $\eta \in S = \operatorname{Ker}(D_{S_0}^+)^\perp$

$$\begin{aligned} J_{f_x}^{x,\circ}(\eta) &= \sigma_{J_{f_x}(d)\leqslant 1}(\eta) \\ &= \sigma_{J_{0,f_{0,D^*x}}^{D^*x}(D_{S_0}^*d)\leqslant 1}(\eta) \\ &= \left(\iota_{J_{0,f_{0,D^*x}}^{D^*x}(w)\leqslant 1} \circ D_{S_0}^*\right)^*(\eta) \\ &= \inf_{v} \sigma_{J_{0,f_{0,D^*x}}^{D^*x}(w)\leqslant 1}(v) \quad \text{s.t.} \quad D_{S_0}v = \eta \\ &= \inf_{z\in\operatorname{Ker}(D_{S_0})} J_{0,f_{D^*x}}^{D^*x,\circ}(D_{S_0}^+\eta + z) \ . \end{aligned}$$

The infimum is bounded and is attained necessarily at some $z \in \operatorname{Ker}(D_{S_0}) \cap S_0 \neq \emptyset$ since $\operatorname{dom} J_{0,f_{D^*x}}^{D^*x,\circ} = S_0$ and $\operatorname{Im}(D_{S_0}^+) = \operatorname{Im}(D_{S_0}^*) \subset S_0$. Moreover, $\operatorname{Ker}(D_{S_0}) \cap S_0 = \operatorname{Ker}(D) \cap S_0$. ∎

We get the following corollary for smooth perturbation G, see Section 3.1. We recall that in this case, the model space $T_x^{J+G} = T_x^J$ and $e_x^{J+G} = e_x^J + P_{T_x^J}\nabla G(x)$.

Corollary 3.2 Let $J \in \Gamma_c^+(\mathbb{R}^n)$, $x \in \mathbb{R}^n$ and $G \in \Gamma_c^+(\mathbb{R}^n)$ a function which is Gâteaux-differentiable at x.. Then, for the particular choice of

$$f_x^{J+G} = f_x^J + \nabla G(x),$$

we have $f_x^{J+G} \in \mathrm{ri}(J+G)(x)$ and for a given $\eta \in S_x^J$, the subdifferential gauge of $J+G$ reads

$$(J+G)^{x,\circ}_{f_x^{J+G},x} = J^{x,\circ}_{f_x^J,x}.$$

PROOF It is sufficient to remark that the smooth perturbation G translates the subdifferential $\partial J(x)$ by $\nabla G(x)$. Hence, using our choice of f_x^{J+G}, we find the same subdifferential gauge. ∎

3.3 Special Cases

3.3.1 Strong Gauge

In this section, we study a particular subclass of convex functions that we dub strong gauges. We start with some definitions.

Definition 3.3 A bounded regularizing gauge J is *separable* with respect to $T = S^\perp$ if

$$\forall\,(x, x') \in T \times S, \quad J(x + x') = J(x) + J(x').$$

Separability of J is equivalent to the following property on the polar J°.

Lemma 3.1 Let J be a bounded gauge. Then, J is separable w.r.t. to $T = S^\perp$ if, and only if its polar J° satisfies

$$J^\circ(x + x') = \max\left(J^\circ(x), J^\circ(x')\right), \quad \forall\,(x, x') \in T \times S\ .$$

PROOF Let $J = \gamma_C$, $x \in T$ and $x' \in S$.

$\Rightarrow$: By virtue of Lemma 2.5, we have

$$\begin{aligned} J^\circ(x+x') &= \sup_{u \in C} \langle x+x',\, u\rangle \\ &= \sup_{J(u) \leqslant 1} \langle x+x',\, u\rangle \\ &= \sup_{J(u_T+u_S) \leqslant 1} \langle x,\, u_T\rangle + \langle x',\, u_S\rangle \\ &= \sup_{J(u_T)+J(u_S) \leqslant 1} \langle x,\, u_T\rangle + \langle x',\, u_S\rangle \qquad \text{using the separability.} \end{aligned}$$

Rewriting the condition $J(u_T + u_S) \leqslant 1$ as $J(u_T) \leqslant \rho, J(u_S) \leqslant 1 - \rho$ for $\rho \in [0, 1]$, we arrive to

$$\begin{aligned} J^\circ(x+x') &= \sup_{\rho \in [0,1]} \sup_{J(u_T) \leqslant \rho, J(u_S) \leqslant 1-\rho} \langle x,\, u_T\rangle + \langle x',\, u_S\rangle \\ &= \sup_{\rho \in [0,1]} \rho \sup_{J(u_T) \leqslant 1} \langle x,\, u_T\rangle + (1-\rho) \sup_{J(u_S) \leqslant 1} \langle x',\, u_S\rangle \\ &= \sup_{\rho \in [0,1]} \rho \sup_{v \in C \cap T} \langle x,\, v\rangle + (1-\rho) \sup_{w C \cap T} \langle x',\, w\rangle \\ &= \sup_{\rho \in [0,1]} \rho \sigma_{C \cap T}(x) + (1-\rho)\sigma_{C \cap S}(x') \\ &= \max(\sigma_{C \cap T}(x), \sigma_{C \cap S}(x'))\ . \end{aligned}$$

Since

$$\sigma_{C \cap T}(x) = \overline{\mathrm{co}}\,(\inf(\sigma_C(x), \iota_S(x))) = \sigma_C(x) = J^\circ(x)$$

and

$$\sigma_{C \cap S}(x') = \overline{\mathrm{co}}\,\big(\inf(\sigma_C(x'), \iota_T(x'))\big) = \sigma_C(x') = J^\circ(x')\ ,$$

the implication follows.

$\Leftarrow$: Using again Lemma 2.5, we get

$$\begin{aligned} J(x+x') &= \sup_{u \in C^\circ} \langle x+x',\, u\rangle \\ &= \sup_{J^\circ(u_T+u_S) \leqslant 1} \langle x,\, u_T\rangle + \langle x',\, u_S\rangle. \end{aligned}$$

Using the separability of the polar,

$$\begin{aligned} J(x+x') &= \sup_{\max(J^\circ(u_T),J^\circ(u_S))\leqslant 1} \langle x, u_T\rangle + \langle x', u_S\rangle \\ &= \sup_{J^\circ(u_T)\leqslant 1, J^\circ(u_S)\leqslant 1} \langle x, u_T\rangle + \langle x', u_S\rangle \\ &= \sup_{v\in C^\circ\cap T} \langle x, v\rangle + \sup_{w\in C^\circ\cap S} \langle x', w\rangle \\ &= \sigma_{C^\circ\cap T}(x) + \sigma_{C^\circ\cap S}(x') \\ &= \overline{\mathrm{co}}\,(\inf(\sigma_{C^\circ}(x), \iota_S(x))) + \overline{\mathrm{co}}\,\big(\inf(\sigma_{C^\circ}(x'), \iota_T(x'))\big) \\ &= \sigma_{C^\circ}(x) + \sigma_{C^\circ}(x') \\ &= J(x) + J(x')\ . \end{aligned}$$

This concludes the proof. ∎

The decomposability of $\partial J(x)$ as described in Theorem 3.1 depends on the particular choice of the map $x \mapsto f_x \in \mathrm{ri}\,\partial J(x)$. An interesting situation is encountered when $e_x \in \mathrm{ri}\,\partial J(x)$, in which case, one can just choose $f_x = e_x$, hence implying that $f_{S_x} = 0$. Strong gauges are precisely a class of gauges for which this situation occurs.

In the sequel, for a given model subspace T, we denote $\widetilde{T}$ the set of vectors sharing the same T,

$$\widetilde{T} = \left\{x \in \mathbb{R}^N \mid T_x = T\right\}.$$

Using positive homogeneity, it is easy to show that $T_{\rho x} = T_x$ and $e_{\rho x} = e_x\ \forall \rho > 0$. Thus $\widetilde{T}$ is a non-empty cone which is contained in T by Proposition 3.1(iv).

> Definition 3.4 — Strong Gauge A *strong gauge* on T is a bounded gauge J such that
>
> (i) For every $x \in \widetilde{T}$, $e_x \in \mathrm{ri}\,\partial J(x)$.
>
> (ii) J is separable with respect to T and $S = T^\perp$.

The following result shows that the decomposability property of Theorem 3.1 has a simpler form when J is a strong gauge.

Proposition 3.12 Let J be a strong gauge on T_x. Then, the subdifferential of J at x reads

$$\partial J(x) = \left\{ \eta \in \mathbb{R}^N \mid \eta_{T_x} = e_x \quad \text{and} \quad J^\circ(\eta_{S_x}) \leqslant 1 \right\}.$$

PROOF Let $J = \gamma_C$. We only need to show that $J^{x,\circ}_{e_x}(\eta_{S_x}) = J^\circ(\eta_{S_x})$. This follows from Proposition 3.7, Lemma 3.1 and Lemma 2.5(ii). Indeed,

$$\begin{aligned}
J^{x,\circ}_{e_x}(\eta_{S_x}) &= \inf_{\tau \geqslant 0} \max(J^\circ(\tau e_x + \eta_{S_x}), \tau) && \text{from Proposition 3.7,} \\
&= \inf_{\tau \geqslant 0} \max(\tau J^\circ(e_x), J^\circ(\eta_{S_x}), \tau) && \text{from Lemma 3.1,} \\
&= \inf_{\tau \geqslant 0} \max(J^\circ(\eta_{S_x}), \tau) && \text{from } e_x \in \partial J(x) \subset C^\circ, \\
&= J^\circ(\eta_{S_x}) \ ,
\end{aligned}$$

which concludes the proof. ■

When J is in addition a norm, this coincides with the decomposability definition of (Candès and Recht 2013). Note however that the last part of assertion (ii) in Proposition 3.8 is an intrinsic property of gauges, while it is stated as an assumption in their definition. A notion of decomposability closely related to that of (Candès and Recht 2013), but different, was proposed in (Negahban et al. 2009). Typical examples of (strongly) decomposable norms are the ℓ^1, $\ell^1 - \ell^2$ and nuclear norms. However, strong decomposability excludes many important cases. One can think of analysis-type semi-norms since strong decomposability is not preserved under pre-composition by a linear operator, or the ℓ^∞ norm among many others.

3.3.2 Examples

3.3.2.1 ℓ^1 Norm

The norm $J(x) = ||x||_1$ is a symmetric (bounded) strong gauge. More precisely, we have the following result.

Proposition 3.13 $J = \|\cdot\|_1$ is a symmetric strong gauge with

$$T_x = \left\{\eta \in \mathbb{R}^N \mid \forall j \notin I, \eta_j = 0\right\}, \quad S_x = \left\{\eta \in \mathbb{R}^N \mid \forall i \in I, \eta_i = 0\right\},$$
$$e_x = \operatorname{sign}(x), \quad f_x = e_x, \quad J^{x,\circ}_{f_x} = \|\cdot\|_\infty + \iota_{S_x},$$

where $I = I(x) = \{i \mid x_i \neq 0\}$.

PROOF The subdifferential of $\|\cdot\|_1$ reads

$$\partial\|\cdot\|_1(x) = \left\{\eta \in \mathbb{R}^N \mid \eta_{(I)} = \operatorname{sign}(x_{(I)}) \quad \text{and} \quad \|\eta_{(I^c)}\|_\infty \leqslant 1\right\}.$$

The expressions of S_x, T_x, e_x and f_x follow immediately. Since $e_x \in \operatorname{ri}\partial\|\cdot\|_1(x)$ and $\|\cdot\|_1$ is separable, it follows from Definition 3.4 that the ℓ^1-norm is a strong gauge. Therefore $J^{x,\circ}_{f_x} = \|\cdot\|_\infty$, and Proposition 3.12 specializes to the stated subdifferential. ∎

Figure 3.2 shows the underlying geometry of the ℓ^1 regularization in two dimensions. Note that $\partial J(x)$ is included in the dual closed ball.

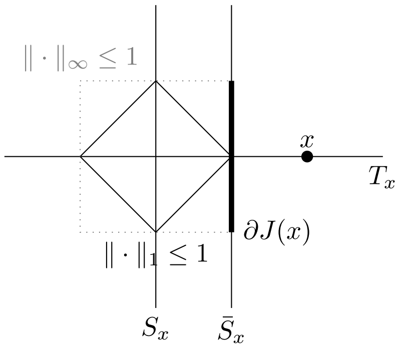


Figure 3.2: ℓ^1 *geometry.*

3.3.2.2 Analysis-ℓ^1 Seminorm

The semi-norm $J(x) = \|D^* x\|_1$ is a symmetric gauge.

PROPOSITION 3.14 $J = ||D^* \cdot ||_1$ is a symmetric (bounded) gauge with

$$T_x = \mathrm{Ker}(D^*_{(I^c))}) = \left\{\eta \in \mathbb{R}^N \mid \forall j \notin I, \langle d_j, \eta_j \rangle = 0\right\}, \quad S_x = \mathrm{Im}(D_{I^c}),$$
$$e_x = P_{\mathrm{Ker}(D^*_{I^c})} D \,\mathrm{sign}(D^* x), \quad f_x = D \,\mathrm{sign}(D^* x),$$
$$J^{x,\circ}_{f_x}(\eta) = \inf_{z \in \mathrm{Ker}(D_{(I^c)})} ||D^+_{(I^c)} \eta + z||_\infty, \quad \text{for} \quad \eta \in S_x \;,$$

where $I = I(x) = \{i \mid \langle d_i, x_i \rangle \neq 0\}$.

PROOF This is a direct consequence of Proposition 3.11 and Proposition 3.13.■

3.3.2.3 ℓ^∞ Norm

The norm $J(x) = ||x||_\infty$ is a symmetric gauge, but unlike the ℓ^1-norm, it is not strongly so (except for $n = 2$). In the following proposition, we rule out the trivial case $x = 0$.

PROPOSITION 3.15 $J = || \cdot ||_\infty$ is a symmetric (bounded) gauge for $x \neq 0$ with

$$S_x = \left\{\eta \mid \eta_{(I^c)} = 0 \quad \text{and} \quad \langle \eta_{(I)}, s_{(I)} \rangle = 0\right\},$$
$$T_x = \left\{\alpha \mid \alpha_{(I)} = \rho s_{(I)} \quad \text{for} \quad \rho \in \mathbb{R}\right\},$$
$$e_x = \frac{s}{|I|}, \quad f_x = e_x, \quad J^{x,\circ}_{f_x}(\eta) = \max_{i \in I} \,(-|I| s_i \eta_i)_+ \quad \text{for} \quad \eta \in S_x \;,$$

where $s = \mathrm{sign}(x)$ and $I = I(x) = \{i \mid |x_i| = ||x||_\infty\}$.

PROOF Recall that for $J = || \cdot ||_\infty$, $f_x = e_x = s/|I|$, with $s = \mathrm{sign}(x)$. Let $\mathcal{K}_x = \partial J(x) - e_x$. It can be straightforwardly shown that in this case,

$$\mathcal{K}_x = \left\{\nu \mid \forall (i,j) \in I \times I^c, \nu_j = 0, \langle \nu_{(I)}, s_{(I)} \rangle = 0, -|I| \nu_i s_i \leqslant 1\right\}$$

This is rewritten as

$$\mathcal{K}_x = S_x \cap \underbrace{\{v \mid \forall i \in I, -|I| v_i s_i \leqslant 1\}}_{=\mathcal{K}'_x}.$$

Thus the subdifferential gauge reads

$$J^{x,\circ}_{f_x}(\eta) = \gamma_{\mathcal{K}_x}(\eta) = \max(\gamma_{S_x}(\eta), \gamma_{\mathcal{K}'_x}(\eta)).$$

We have $\gamma_{S_x}(\eta) = \iota_{S_x}(\eta)$ and $\gamma_{\mathcal{K}'_x}(\eta) = \max_{i \in I} (-|I| s_i \eta_i)_+$, where $(\cdot)_+$ is the positive part, hence we obtain

$$J^{x,\circ}_{f_x}(\eta) = \begin{cases} \max_{i \in I} (-|I| s_i \eta_i)_+ & \text{if } \eta \in S_x \\ +\infty & \text{otherwise.} \end{cases}$$

Therefore the subdifferential of $\|\cdot\|_\infty$ at x takes the form

$$\partial J(x) = \left\{ \eta \in \mathbb{R}^N \mid \eta_{T_x} = e_x = \frac{s}{|I|} \quad \text{and} \quad \max_{i \in I} (-|I| s_i \eta_i)_+ \leqslant 1 \right\},$$

which conclude the proof. ∎

Figure 3.3 shows the underlying geometry of the ℓ^∞ regularization in three dimensions.

3.3.2.4 $\ell^1 - \ell^2$ Norm

The $\ell^1 - \ell^2$ norm is a symmetric strong gauge.

Proposition 3.16 The $\ell^1 - \ell^2$ norm associated to the partition $\mathcal{B}$, as defined in (1.13), is a symmetric (bounded) strong gauge with

$$T_x = \left\{ \eta \mid \forall j \notin I, \eta_j = 0 \right\}, \quad S_x = \{\eta \mid \forall i \in I, \eta_i = 0\},$$
$$e_x = (\mathcal{N}(x_b))_{b \in \mathcal{B}}, \quad f_x = e_x, \quad J^{x,\circ}_{f_x} = \|\cdot\|_{\infty,2} + \iota_{S_x},$$

where $I = I(x) = \{b \mid x_b \neq 0\}$, and $\mathcal{N}(a) = a/\|a\|$ if $a \neq 0$, and $\mathcal{N}(0) = 0$.

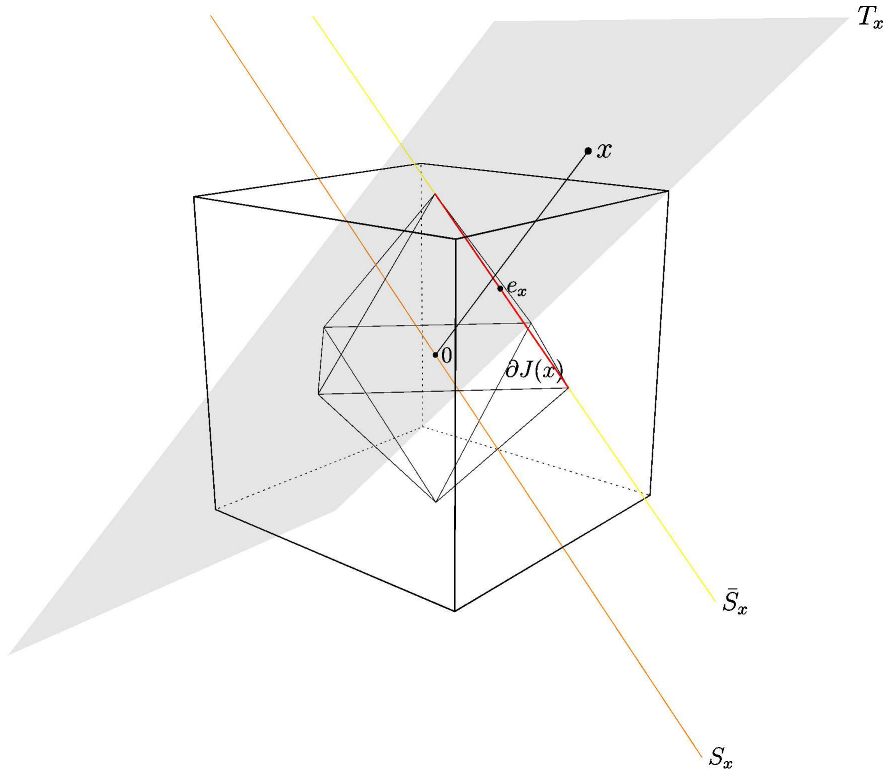


Figure 3.3: ℓ^∞ *geometry.*

Figure 3.4 shows the underlying geometry of the $\ell^1 - \ell^2$ regularization in three dimensions. We take $J(x) = \sqrt{x_1^2 + x_2^2} + |x_3|$.

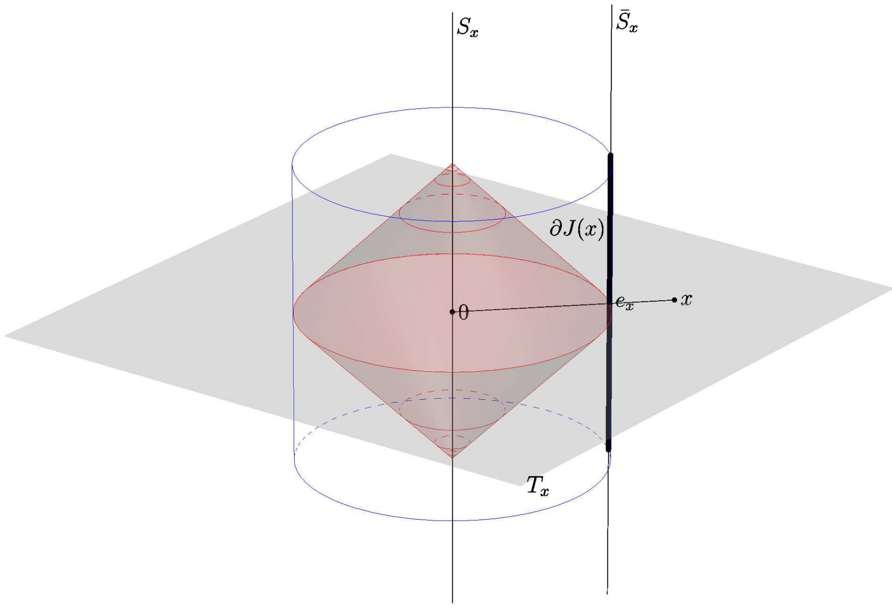


Figure 3.4: $\ell^1 - \ell^2$ *geometry. In red, the* $\ell^1 - \ell^2$ *ball. In blue, the dual ball.*

3.3.2.5 Nuclear Norm

We show that the nuclear norm is a symmetric strong gauge. Since $\|\cdot\|_1$ is an absolutely symmetric gauge, it is immediate to see that $\|\cdot\|_*$ is an unitarily invariant norm according to Proposition 2.10.

Proposition 3.17 The nuclear norm is a symmetric strong gauge with

$$S_x = \left\{ U_\perp^* C V_\perp \mid C \in \mathbb{R}^{(N_1 - r) \times (N_2 - r)} \right\},$$

$$\begin{aligned} T_x &= \{ UA^* + BV^* \mid A \in \mathbb{R}^{N_2 \times r}, B \in \mathbb{R}^{N_1 \times r} \} \\ &= \left\{ Z \in \mathbb{R}^{(N_1 - r) \times (N_2 - r)} \mid U_\perp^* Z V_\perp = 0 \right\}, \end{aligned}$$

$$e_x = UV^*, \quad f_x = e_x, \quad J_{f_x}^{x,\circ}(x) = \max_i \sigma_i + \iota_{S_x} \ ,$$

where $U_\perp, V_\perp$ span the orthogonal of the ranges of U, V.

It can be observed that $\dim(T_x) = r(N_1 + N_2 - r)$ and $\dim(S_x) = N_1 N_2 - \dim(T_x) = N_1 N_2 - r(N_1 + N_2) + r^2$.

PROOF The subdifferential of the nuclear norm is a classical result in convex analysis of spectral functions, see e.g. (Watson 1992; Lewis 1995). More precisely, let $x \in \mathbb{R}^{N_1 \times N_2}$ be a matrix and $x = U\Sigma V^*$ its singular value decomposition. Then, the subdifferential $\partial J(x)$ reads

$$\partial J(x) = \{UV^* + M \mid \ ||M|| \leqslant 1, \quad U^* M = 0 \quad \text{and} \quad MV = 0\}.$$

The expressions of the subspaces T_x, S_x and e_x follow immediately. Since the nuclear norm is a strong gauge, we get from Proposition 3.12 that the subdifferential gauge is the spectral norm. ■

3.3.2.6 Polyhedral Gauges

The ℓ^1 and ℓ^∞ norms are special cases of polyhedral priors. There are two alternative ways to define a polyhedral gauge. The H-representation encodes the gauge through the hyperplanes that support the polygonal facets of its unit level set. The V-representation encodes the gauge through the vertices that are the extreme points of this unit level set. We focus here on the H-representation.

A polyhedral gauge in the H-representation is defined as

$$J(x) = \max_{1 \leqslant i \leqslant p} (\langle x, \, d_i \rangle)_+ = J_0(D^* x) \quad \text{where} \quad J_0(u) = \max_{1 \leqslant i \leqslant p} (u_i)_+,$$

and we have defined $D = (d_i)_{i=1}^p \in \mathbb{R}^{n \times p}$. For instance, $J = || \cdot ||_1$ can be recovered using the matrix $D \in \mathbb{R}^{n \times 2^n}$ enumerating all sign patterns and

$J = \|\cdot\|_\infty$ corresponds to taking $D = [-\mathrm{Id}_n \quad \mathrm{Id}_n]$. Observe that the polar of a polyhedral gauge is again a polyhedral gauge.

Such a polyhedral gauge can also be thought as an analysis gauge. One can then characterize decomposability of J_0 and then invoke Proposition 3.11 to derive those of J. This is what we are about to do. In the following, we denote $(a^i)_{1 \leqslant i \leqslant p}$ the standard basis of $\mathbb{R}^p$. Figure 3.5 shows the geometry of this regularization when u is on the positive ray $\mathbb{R}_+(1,1)$ in two dimensions. Note that the level-set $\{J_0(\cdot) \leqslant 1\}$ to 1 is unbounded.

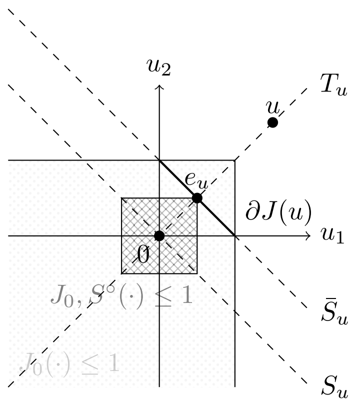


Figure 3.5: *Polyhedral geometry.*

PROPOSITION 3.18 $J_0(u) = \max_{1 \leqslant i \leqslant p} (u_i)_+$ is a (bounded) gauge and,

- If $u_i \leqslant 0$, $\forall i \in \{1, \cdots, p\}$, then

$$S_u = \mathrm{span}\left(a^i\right)_{i \in I_0}, \quad T_u = \mathrm{span}\left(a^i\right)_{i \notin I_0},$$
$$e_u = 0, \quad f_u = \mu \sum_{i \in I_0} a^i, \text{ for any } 0 < \mu < 1,$$
$$J_{f_u}^{u,\circ}(\eta) = \inf_{\tau \geqslant \max_{i \in I_0} (-\eta_i)_+/\mu} \max\left(\tau\mu|I_0| + \sum_{i \in I_0} \eta_i, \tau\right) \quad \text{for} \quad \eta \in S_u,$$

where
$$I_0 = \{i \in \{1, \cdots, p\} \mid u_i = J_0(u) = 0\}.$$

- If $\exists i \in \{1, \cdots, p\}$ such that $u_i > 0$, then

$$S_u = \left\{ \eta \mid \eta_{(I_+^c)} = 0 \quad \text{and} \quad \langle \eta_{(I_+)}, s_{(I_+)} \rangle = 0 \right\},$$
$$T_u = \left\{ \alpha \mid \alpha_{(I_+)} = \mu s_{(I_+)} \quad \text{for} \quad \mu \in \mathbb{R} \right\},$$
$$e_u = \frac{s}{|I_+|}, \quad f_u = e_u, \quad J_{f_u}^{u,\circ}(\eta) = \max_{i \in I_+} (-|I_+|\eta_i)_+ \quad \text{for} \quad \eta \in S_u \ ,$$

where

$$s = \sum_{i \in I_+} a^i \quad \text{and} \quad I_+ = \{ i \in \{1, \cdots, p\} \mid u_i = J_0(u) \quad \text{and} \quad u_i > 0 \} \ .$$

PROOF In general, the subdifferential of J_0 reads

$$\partial J_0(u) = \left\{ \sum_{i \in I} \rho_i s_i a^i \mid \rho \in \Sigma_I, s_i \in \begin{cases} \{1\} & \text{if } u_i > 0 \\ [0, 1] & \text{if } u_i = 0 \\ \{0\} & \text{if } u_i < 0 \end{cases} \right\},$$

where Σ_I is the canonical simplex in $\mathbb{R}^{|I|}$, and $I = \{ i \in \{1, \cdots, p\} \mid (x_i)_+ = J_0(x) \}$.

- If $u_i \leqslant 0$, $\forall i \in \{1, \cdots, p\}$, the above expression becomes

$$\partial J_0(u) = \left\{ \sum_{i \in I_0} \rho_i s_i a^i \mid \rho \in \Sigma_{I_0}, s_i \in [0, 1] \right\} \ ,$$

where $I_0 = \{ i \in \{1, \cdots, p\} \mid u_i = J_0(u) = 0 \}$. Equivalently, $\partial J_0(u)$ is the intersection of the unit ℓ_1 ball and the positive orthant on $\mathbb{R}^{|I_0|}$. The expressions of S_u, T_u and e_u then follow immediately. $\partial J_0(u)$ then contains $e_u = 0$, but not in its relative interior. Choosing any f_u as advocated, we have $f_u \in \operatorname{ri} \partial J_0(u)$. To get the subdifferential gauge, we some calculus rules on gauges and apply Lemma 3.7 to get

$$J_{f_u}^{u,\circ}(\eta_{(I_0)}) = \inf_{\tau \geqslant 0, \quad \tau (f_u)_i \geqslant -\eta_i \ \ \forall i \in I_0} \max(\|\tau f_u + \eta\|_1, \tau) \ ,$$

where the extra-constraints on τ come from the fact that $\partial J_0(u)$ is in the positive orthant, and the ℓ^1 norm is the gauge of the unit ℓ^1-ball. We then have

$$J^{u,\circ}_{f_u}(\eta_{(I_0)}) = \inf_{\tau \geqslant 0,\ \mu\tau \geqslant \max_{i \in I_0} -\eta_i} \max(\tau \sum_{i \in I_0} (\mu a^i + \eta_i), \tau)$$
$$= \inf_{\tau \geqslant \max_{i \in I_0}(-\eta_i)_+/\mu} \max(\tau\mu|I_0| + \sum_{i \in I_0} \eta_i, \tau) \ .$$

- Assume now that $u_i > 0$ for at least one $i \in \{1, \cdots, p\}$. In such a case, $J_0(u) = ||u||_\infty$, and the subdifferential becomes

$$\partial J_0(u) = \Sigma_{I_+} \ ,$$

where $I_+ \{i \in \{1, \cdots, p\} \mid \ u_i = J_0(u) \quad \text{and} \quad u_i > 0\}$. The forms of S_u, T_u, e_u, f_u and the subdifferential gauge can then be retrieved from those of the ℓ^∞-norm with $s_{(I_+)} = 1$ and $s_{(I_+^c)} = 0$. ■

4

Partial Smoothness

Main contributions of this chapter

- Specialization and application of the theory of partial smoothness (Definition 4.1) to popular gauges in imaging and statistics.
- Derivation of explicit partial smoothness Lipschitz-constants for a particular sub-class of partly smooth functions (Definition 4.2).

Contents

THEOREM 3.1 provides a pointwise decomposition of the subdifferential of a convex function. It actually says nothing about the stability of such a formula at points $\tilde{x}$ close enough from x. In order to obtain this stability, one needs to restrict the set of finite-valued convex functions used as regularizers. We propose two different classes of such regularizers, coined partly smooth functions and partly smooth functions relative to a linear manifold . The first one comes directly from the optimization litterature (Lewis 2002), whereas the second is introduced in order to be able to provide explicit constants in our robustness results.

4.1 Partly Smooth Functions

The notion of "partly smooth" functions (Lewis 2002) unifies many non-smooth functions known in the literature. The 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". Loosely speaking, a partly smooth function behaves smoothly as we move on the partial smoothness 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 2002; Lewis et al. 2013).

4.1.1 Definition

Specialized to convex functions, the definition of partly smooth functions reads as follows[1].

DEFINITION 4.1 A function $J \in \Gamma_c^+(\mathbb{R}^n)$ is said to be *partly smooth* (PSF) at x relative to a set $\mathcal{M} \subseteq \mathbb{R}^n$ if there exists a neighborhood U of x such that

(i) **Smoothness.** $\mathcal{M} \cap U$ is a C^2-manifold and J restricted to $\mathcal{M} \cap U$ is C^2,

$$J_{|U \cap \mathcal{M}} \in C^2(\mathcal{M} \cap U).$$

(ii) **Sharpness.** The tangent space of $\mathcal{M}$ at x is the model tangent space T_x,

$$\mathcal{T}_{\mathcal{M}}(x) = T_x.$$

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

The manifold $\mathcal{M}$ is coined the *model manifold* of $x \in \mathbb{R}^n$. 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 $\mathcal{M}$.

1. Again, we could define this notion without assumption of boundness from below.

We denote the set of all partly smooth functions at x relative to a manifold $\mathcal{M}$ as $\mathcal{S}_x(\mathcal{M})$ and the set of all partly smooth functions relative to a manifold $\mathcal{M}$ as $\mathcal{S}(\mathcal{M})$. The definition of continuity of ∂J is to be understood according to Definition 2.17. Since J is proper convex continuous, the subdifferential of $\partial J(x)$ is everywhere non-empty and compact and every subgradient is regular. Therefore, the Clarke regularity property (Lewis 2002, Definition 2.7(ii)) is automatically verified. In view of (Lewis 2002, Proposition 2.4(i)-(iii)), our sharpness property is equivalent to that of (Lewis 2002, Definition 2.7(iii)).

Obviously, any smooth function $J : \mathbb{R}^n \to \mathbb{R}$ is partly smooth relative to the manifold $\mathbb{R}^n$. Moreover, any indicator function $\iota_{\mathcal{M}}$ of a manifold $\mathcal{M}$ is partly smooth relative to $\mathcal{M}$.

Remark that in the previous definition, $\mathcal{M}$ needs only to be defined locally around x, and it can be shown to be locally unique. Hence the notation $\mathcal{M}$ is unambiguous.

Lemma 4.1 Let $J \in \mathcal{S}_x(\mathcal{M})$ be a partly smooth function at $x \in \mathbb{R}^n$ relative to both $\mathcal{M}$ and $\bar{\mathcal{M}}$. Then, there exists a neighborhood U of x such that

$$U \cap \mathcal{M} = U \cap \bar{\mathcal{M}}.$$

PROOF This is proved in Corollary 4.2 of (Hare et al. 2007). ∎

4.1.2 Partial Smoothness Calculus

Partial smoothness is preserved under addition, pre-composition by a linear operator and matrix lift. These results are proved in (Lewis 2002; Daniilidis et al. 2013).

Proposition 4.1 Let $J_0 \in \mathcal{S}_z(\mathcal{M}_z^0)$ be a partly smooth function at $z = D^* x$ relative to the manifold $\mathcal{M}_z^0$. Then $J = J_0 \circ D^* \in \mathcal{S}_x(\mathcal{M})$ for the manifold

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

PROOF This is proved in (Lewis 2002, Theorem 4.2). Contrary to the general case, the transversality hypothesis is automatically satisfied since J_0 is convex and continuous. ■

PROPOSITION 4.2 Let J and G two partly smooth functions at $x \in \mathbb{R}^n$ relative to the manifolds $\mathcal{M}^J$ and $\mathcal{M}^G$. Then $J + G \in \mathcal{S}_x(\mathcal{M})$ for the manifold

$$\mathcal{M} = \mathcal{M}^J \cap \mathcal{M}^G.$$

PROOF This is proved in (Lewis 2002, Corollary 4.6). Contrary to the general case, the transversality hypothesis is automatically satisfied since J and G are convex and continuous. ■

PROPOSITION 4.3 Let j be an absolutely symmetric gauge and $x \in \mathbb{R}^{n_1 \times n_2}$ a symmetric matrix. Then the two following statements are equivalent:

(i) j is a partly smooth function at Λ_x relative to the manifold m_{Λ_x}.

(ii) $J = j \circ \sigma \in \mathcal{S}_x(\mathcal{M})$ for the manifold $\mathcal{M} = \sigma^{-1}(m_{\Lambda_x})$.

PROOF See (Daniilidis et al. 2013, Theorem 5.3). ■

4.2 Partly Smooth Functions With Linear Manifolds

In practice, many of the partly smooth functions we consider are associated to linear manifolds (i.e. the tangent model subspace is the model manifold $\mathcal{M} = T_x$). These functions, coined partly smooth functions with linear manifolds, encompass most of the knowns regularizations in the image processing and statistics, such as the ℓ^1, $\ell^1 - \ell^2$, ℓ^∞ norm and their sums and compositions by a linear operator, with the noticeable exception of the nuclear norm.

4.2.1 Definition

We restrict our interest in this section to convex funtions J which are partly smooth at $x \in \mathbb{R}^n$ with respect to a linear subspace. In this case, this subspace is T_x. The following theorem proves that such functions, which enjoy the fact that ∂J is Lipschitz on T_x are characterized by a set of parameters.

THEOREM 4.1 Let Γ be any coercive gauge bounded on T_x for $x \in \mathbb{R}^n$. Let $J \in \mathcal{S}_x(\mathcal{M}_x)$ for the manifold $\mathcal{M}_x = T_x$ and we assume that $\partial J : T_x \rightrightarrows \mathbb{R}^n$ is Lipschitz around x. Then for any Lipschitz-mapping.

$$f : \begin{cases} T_x & \to \mathbb{R}^n \\ \tilde{x} & \mapsto f_{\tilde{x}} \in \operatorname{ri} \partial J(\tilde{x}), \end{cases}$$

there exist four non-negative reals $\nu_x, \mu_x, \tau_x, \xi_x$ such that

$$\forall x' \in T, \Gamma(x - x') \leqslant \nu_x \Rightarrow T_x = T_{x'} \tag{4.1}$$

and for every $x' \in T$ with $\Gamma(x - x') < \nu_x$, one has

$$\Gamma(e_x - e_{x'}) \leqslant \mu_x \Gamma(x - x'), \tag{4.2}$$

$$J_{f_x}^{x,\circ}(P_S(f_x - f_{x'})) \leqslant \tau_x \Gamma(x - x'), \tag{4.3}$$

$$\sup_{\substack{u \in S \\ u \neq 0}} \frac{J_{f_{x'}}^{x',\circ}(u) - J_{f_x}^{x,\circ}(u)}{J_{f_x}^{x,\circ}(u)} \leqslant \xi_x \Gamma(x - x'). \tag{4.4}$$

Moreover, there exists such a mapping f.

PROOF We prove this result for $\Gamma = \|\cdot\|$. This is not restrictive, since for every $x \in \mathbb{R}^n$,

$$\Gamma(x) \leqslant |||\mathrm{Id}|||_{\Gamma \to \ell^2} \|x\|.$$

We start from the hypotheses $J \in \mathcal{S}_x(\mathcal{M}_x)$ for the manifold $\mathcal{M}_x = T_x$ and $\partial J : T_x \rightrightarrows \mathbb{R}^n$ is Lipschitz around x.

- *Existence of* f_x. Such a mapping exists according to (Aubin et al. 2009, Theorem 9.4.3).
- ν-*stability.* Using (Lewis 2002, Proposition 2.10) the sharpness property (ii) is locally stable. Hence, for $x' \in T_x$ in a neighborhood of x, $T_{x'} = T_x$. The radius of this neighborhood can be taken as ν_x.
- μ-*stability.* Using (Hiriart-Urruty et al. 2001, Corollary VI.2.1.3), we write for any $h \in T_x$
$$J(x + th) = J(x) + t\langle s, h\rangle + o(t),$$
where $s \in F_{\partial J(x)}(h)$. Since J restricted to $T_x \cap U$ is C^2 according to the smoothness property, repeating this argument at order 2 let us concludes that the mapping $z \in T_x \cap U \mapsto e_z$ is C^1. Hence, this map is Lipschitz.
- τ-*stability.* One has
$$J^{x,\circ}_{f_x}(P_S(f_x - f_{x'})) \leqslant |||P_{S_x}|||_{J^{x,\circ}_{f_x} \to \ell^2} ||f_x - f_{x'}|| \leqslant \tau_x ||x - x'||,$$
where $\tau_x = |||P_{S_x}|||_{J^{x,\circ}_{f_x} \to \ell^2} \beta$ and β is the Lipschitz constant associated to f_x, proving (4.3).
- ξ-*stability.* ∂J is Lipschitz around x and $x \mapsto f_x$ is Lipschitz. Hence, the application $x \mapsto (\partial J(x) - f_x)$ is Lipschitz on T_x. Using Lemma 2.6, we get that
$$J^{x',\circ}_{f_{x'}}(u) - J^{x,\circ}_{f_x}(u) \leqslant \beta ||x' - x|| ||u||.$$
Since $||u|| \leqslant |||\mathrm{Id}|||_{\ell^2 \to J^{x,\circ}_{f_x}} J^{x,\circ}_{f_x}(u)$, we get bound (4.4) where $\xi_x = \beta |||\mathrm{Id}|||_{\ell^2 \to J^{x,\circ}_{f_x}}$.
■

This result motivates the following definition.

> DEFINITION 4.2 — PSF RELATIVE TO A LINEAR MANIFOLD. A finite-valued convex function J is said to be *partly smooth relative to a linear manifold* at $x \in \mathbb{R}^n$, if J is partly smooth at x for the manifold $\mathcal{M} = T_x$. The set of all partly smooth functions with linear manifolds at x, such that ∂J is Lipschitz around x relative to T_x, with parameters $(\Gamma, f_x, \nu_x, \mu_x, \tau_x, \xi_x)$ is denoted $\mathbb{SL}_x(\Gamma, f_x, \nu_x, \mu_x, \tau_x, \xi_x)$.

4.2.2 Stability under the Sum and Precomposition by a Linear Operator

Partial smoothness with linear manifold property is preserved under addition and pre-composition by a linear operator, and one can give explicit bound on the corresponding Lipschitz constants.

4.2.2.1 Addition

PROPOSITION 4.4 Let $x \in \mathbb{R}^n$, J and G two partly smooth functions with linear manifolds such that

$$J \in \mathbb{SL}_x(\Gamma^J, f_x^J, \nu_x^J, \mu_x^J, \tau_x^J, \xi_x^J)$$
$$G \in \mathbb{SL}_x(\Gamma^G, f_x^G, \nu_x^G, \mu_x^G, \tau_x^G, \xi_x^G).$$

Then, $H = J + G$ is also partly smooth with linear manifold at x, for the choice $f_x^H = f_x^J + f_x^G$ and $\Gamma^H = \max(\Gamma^J, \Gamma^G)$, with ∂H Lipschitz and the parameters

$$\begin{aligned}
\nu_x^H &= \min(\nu_x^J, \nu_x^G) \\
\mu_x^H &= \mu_x^J \, |\!|\!| P_{T^H} |\!|\!|_{\Gamma^J \to \Gamma^H} + \mu_x^G \, |\!|\!| P_{T^H} |\!|\!|_{\Gamma^G \to \Gamma^H} \\
\tau_x^H &= \tau_x^J + \tau_x^G + \mu_x^J \, |\!|\!| P_{S^H \cap T^J} |\!|\!|_{\Gamma^J \to H_{f_x^H}^{x,\circ}} + \mu_x^G \, |\!|\!| P_{S^H \cap T^G} |\!|\!|_{\Gamma^G \to H_{f_x^H}^{x,\circ}} \\
\xi_x^H &= \max(\xi_x^J, \xi_x^G).
\end{aligned}$$

PROOF In the following, all operator bounds that appear are finite owing to the coercivity assumption on the involved gauges in Definition 4.2 of a PSFL.

It is straightforward to see that the function $\Gamma^H = \max(\Gamma^J, \Gamma^G)$ is indeed a gauge, which is bounded and coercive on $T^H = T^J \cap T^G$. Moreover, given that both J and G are PSFL at x with corresponding parameters ν_x^J and ν_x^G, we have with the advocated choice of Γ^H and ν_x^H,

$$\Gamma^J(x - x') \leqslant \nu_x^J \quad \text{and} \quad \Gamma^G(x - x') \leqslant \nu_x^G,$$

for every $\forall x' \in T_x^H$ such that $\Gamma^H(x-x') \leqslant \nu_x^H$. It follows that:

- Since J and G are both PSFL, then we have $T_x^J = T_{x'}^J$ and $T_x^G = T_{x'}^G$, and thus by Proposition 3.10(i)

$$T_x^H = T_x^J \cap T_x^G = T_{x'}^J \cap T_{x'}^G = T_{x'}^H = T^H.$$

- μ_x^H**-stability:** we have from Proposition 3.10(ii)

$$\begin{aligned}
\Gamma^H(e_x^H - e_{x'}^H) &= \Gamma^H\left(P_{T^H}(e_x^J + e_x^G - e_{x'}^J - e_{x'}^G)\right)\\
&\leqslant \Gamma^H\left(P_{T^H}(e_x^J - e_{x'}^J)\right) + \Gamma^H\left(P_{T^H}(e_x^G - e_{x'}^G)\right)\\
&\leqslant |||P_{T^H}|||_{\Gamma^J\to\Gamma^H}\,\Gamma^J\left(e_x^J - e_{x'}^J\right) + |||P_{T^H}|||_{\Gamma^G\to\Gamma^H}\,\Gamma^G\left(e_x^G - e_{x'}^G\right)\\
&\leqslant \left(\mu_x^J\,|||P_{T^H}|||_{\Gamma^J\to\Gamma^H} + \mu_x^G\,|||P_{T^H}|||_{\Gamma^G\to\Gamma^H}\right)\Gamma^H(x-x')\ ,
\end{aligned}$$

  where we used μ_x^J- and μ_x^G-stability of J and G in the last inequality.

- τ_x^H**-stability:** the fact that $S^J \subseteq S^H$ and $S^G \subseteq S^H$ and subadditivity of gauges lead to

$$\begin{aligned}
&H_{f_x^H}^{x,\circ}\left(P_{S^H}(f_x^H - f_{x'}^H)\right)\\
&= H_{f_x^H}^{x,\circ}\left(P_{S^J}(f_x^J - f_{x'}^J) + P_{S^G}(f_x^G - f_{x'}^G) + P_{S^H}(e_x^J - e_{x'}^J) + P_{S^H}(e_x^G - e_{x'}^G)\right)\\
&\leqslant H_{f_x^H}^{x,\circ}\left(P_{S^J}(f_x^J - f_{x'}^J)\right) + H_{f_x^H}^{x,\circ}\left(P_{S^G}(f_x^G - f_{x'}^G)\right)\\
&\quad + H_{f_x^H}^{x,\circ}\left(P_{S^H}(e_x^J - e_{x'}^J)\right) + H_{f_x^H}^{x,\circ}\left(P_{S^H}(e_x^G - e_{x'}^G)\right)\ . \qquad (4.5)
\end{aligned}$$

  According to Proposition 3.10(iii), we have

$$H_{f_x^H}^{x,\circ}\left(P_{S^J}(f_x^J - f_{x'}^J)\right) = \inf_{\eta_1+\eta_2 = P_{S^J}(f_x^J - f_{x'}^J)} \max(J_{f_x^J}^{x,\circ}(\eta_1), G_{f_x^G}^{x,\circ}(\eta_2))\ .$$

  Since $\operatorname{dom} J_{f_x^J}^{x,\circ} = S^J$, $(\eta_1,\eta_2) = (P_{S^J}(f_x^J - f_{x'}^J), 0)$ is a feasible point of the last problem, and we get

$$H_{f_x^H}^{x,\circ}\left(P_{S^J}(f_x^J - f_{x'}^J)\right) \leqslant J_{f_x^J}^{x,\circ}\left(P_{S^J}(f_x^J - f_{x'}^J)\right).$$

Moreover, as $e^J_x, e^J_{x'} \in T^J$ (see Proposition 3.1(ii)) and $S^J \subseteq S^H$, we have

$$\begin{aligned}
&\min_{\eta_1 \in T^J, \eta_2 S^J, \eta_1+\eta_2 \in S^H} \|\eta_1 + \eta_2 - (e^J_x - e^J_{x'})\|^2 \\
&= \min_{\eta_1 \in T^J, \eta_2 S^J, \eta_1+\eta_2 \in S^H} \|\eta_1 - (e^J_x - e^J_{x'})\|^2 + \|\eta_2\|^2 \\
&= \min_{\eta_1 \in T^J, \eta_2 S^J, \eta_1 \in S^H} \|\eta_1 - (e^J_x - e^J_{x'})\|^2 + \|\eta_2\|^2 \\
&= \min_{\eta_1 \in S^H \cap T^J} \|\eta_1 - (e^J_x - e^J_{x'})\|^2 \ .
\end{aligned}$$

That is

$$P_{S^H}(e^J_x - e^J_{x'}) = P_{S^H \cap T^J}(e^J_x - e^J_{x'}) \ .$$

Thus

$$H^{x,\circ}_{f^H_x}\left(P_{S^H}(e^J_x - e^J_{x'})\right) \leqslant |\!|\!| P_{S^H \cap T^J} |\!|\!|_{\Gamma^J \to H^{x,\circ}_{f^H_x}} \Gamma^J\left(e^J_x - e^J_{x'}\right) \ .$$

Similar reasoning leads to the following bounds

$$\begin{aligned}
H^{x,\circ}_{f^H_x}\left(P_{S^G}(f^G_x - f^G_{x'})\right) &\leqslant G^{x,\circ}_{f^G_x}\left(P_{S^G}(f^G_x - f^G_{x'})\right), \\
H^{x,\circ}_{f^H_x}\left(P_{S^H}(e^G_x - e^G_{x'})\right) &\leqslant |\!|\!| P_{S^H \cap T^G} |\!|\!|_{\Gamma^J \to H^{x,\circ}_{f^H_x}} \Gamma^G\left(e^G_x - e^G_{x'}\right) \ .
\end{aligned}$$

Having this, we can continue to bound (4.5) as

$$\begin{aligned}
&H^{x,\circ}_{f^H_x}\left(P_{S^H}(f^H_x - f^H_{x'})\right) \\
&\leqslant J^{x,\circ}_{f^J_x}\left(P_{S^J}(f^J_x - f^J_{x'})\right) + G^{x,\circ}_{f^G_x}\left(P_{S^G}(f^G_x - f^G_{x'})\right) \\
&\quad + |\!|\!| P_{S^H \cap T^J} |\!|\!|_{\Gamma^J \to H^{x,\circ}_{f^H_x}} \Gamma^J\left(e^J_x - e^J_{x'}\right) + |\!|\!| P_{S^H \cap T^G} |\!|\!|_{\Gamma^J \to H^{x,\circ}_{f^H_x}} \Gamma^G\left(e^G_x - e^G_{x'}\right) \\
&\leqslant \tau^J_x \, \Gamma^J(x - x') + \tau^G_x \, \Gamma^G(x - x') + \mu^J_x \, |\!|\!| P_{S^H \cap T^J} |\!|\!|_{\Gamma^J \to H^{x,\circ}_{f^H_x}} \Gamma^J\left(x - x'\right) \\
&\quad + \mu^G_x \, |\!|\!| P_{S^H \cap T^G} |\!|\!|_{\Gamma^G \to H^{x,\circ}_{f^H_x}} \Gamma^G\left(x - x'\right) \\
&\leqslant \left(\tau^J_x + \tau^G_x + \mu^J_x \, |\!|\!| P_{S^H \cap T^J} |\!|\!|_{\Gamma^J \to H^{x,\circ}_{f^H_x}} + \mu^G_x \, |\!|\!| P_{S^H \cap T^G} |\!|\!|_{\Gamma^G \to H^{x,\circ}_{f^H_x}}\right) \Gamma^H(x - x') \ ,
\end{aligned}$$

where the last two inequalities J and G follow from μ^J_x-, τ^J_x-, μ^G_x- and τ^G_x-stability of J and G.

- ξ_x^H**-stability:** Proposition 3.10(iii) again yields that for any $\eta \in S^H$

$$\begin{aligned} H_{f_{x'}^H}^{x',\circ}(\eta) &= \inf_{\eta_1+\eta_2=\eta} \max(J_{f_{x'}^J}^{x',\circ}(\eta_1), G_{f_{x'}^G}^{x',\circ}(\eta_2)) \\ &\leqslant \max(J_{f_{x'}^J}^{x',\circ}(\bar\eta_1), G_{f_{x'}^G}^{x',\circ}(\bar\eta_2)) \end{aligned}$$

for any feasible $(\bar\eta_1,\bar\eta_2) \in S^J \times S^G \cap \{(\eta_1,\eta_2 \mid \eta_1+\eta_2=\eta\}$. Now both J and G are PRF, hence respectively ξ_x^J- and ξ_x^G-stable. Therefore, with the form of Γ^H we have

$$\begin{aligned} J_{f_{x'}^J}^{x',\circ}(\bar\eta_1) &\leqslant (1+\xi_x^J\Gamma^J(x-x'))J_{f_x^J}^{x,\circ}(\bar\eta_1) \leqslant \beta J_{f_x^J}^{x,\circ}(\bar\eta_1) \\ G_{f_{x'}^G}^{x',\circ}(\bar\eta_2) &\leqslant (1+\xi_x^G\Gamma^G(x-x'))G_{f_x^G}^{x,\circ}(\bar\eta_2) \leqslant \beta G_{f_x^G}^{x,\circ}(\bar\eta_2) \ , \end{aligned}$$

where $\beta = 1+\max\left(\xi_x^J,\xi_x^G\right)\Gamma^H(x-x')$. Whence we get

$$\max(J_{f_{x'}^J}^{x',\circ}(\eta_1), G_{f_{x'}^G}^{x',\circ}(\eta_2)) \leqslant \beta \max(J_{f_x^J}^{x,\circ}(\bar\eta_1), G_{f_x^G}^{x,\circ}(\bar\eta_2)) \ .$$

Taking in particular

$$(\bar\eta_1,\bar\eta_2) \in \operatorname*{Argmin}_{\eta_1+\eta_2=\eta} \max(J_{f_x^J}^{x,\circ}(\eta_1), G_{f_x^G}^{x,\circ}(\eta_2))$$

we arrive at

$$H_{f_{x'}^H}^{x',\circ}(\eta) \leqslant \beta \inf_{\eta_1+\eta_2=\eta} \max(J_{f_x^J}^{x,\circ}(\eta_1), G_{f_x^G}^{x,\circ}(\eta_2)) = \beta H_{f_x^H}^{x,\circ}(\eta) \ .$$

This completes the proof. ■

4.2.2.2 Precomposition by a Linear Operator

Proposition 4.5 Let J_0 be a partly smooth function with linear manifold at $u = D^*x$ with parameter

$$J_0 \in \mathbb{SL}_u(\Gamma_0, f_{0,u}, \nu_{0,u}, \mu_{0,u}, \tau_{0,u}, \xi_{0,u}).$$

Then $J = J_0 \circ D^*$ is partly smooth with linear manifold at x, with the choice $f_x = Df_{0,u}$ and Γ any bounded coercive gauge on T, with ∂J Lipschitz and the parameters

$$\begin{aligned}
\nu_x &= \frac{1}{\|D^*\|_{\Gamma\to\Gamma_0}}\nu_{0,u}\\
\mu_x &= \mu_{0,u}\|P_T D\|_{\Gamma\to\Gamma_0}\|D^*\|_{\Gamma\to\Gamma_0}\\
\tau_x &= \left(\tau_{0,u}\left\|D_{S_0}^+ P_S D\right\|_{J^{u,\circ}_{0,f_{0,u}}\to J^{u,\circ}_{0,f_{0,u}}} + \mu_{0,u}\left\|D_{S_0}^+ P_S D\right\|_{\Gamma_0\to J^{u,\circ}_{0,f_{0,u}}}\right)\|D^*\|_{\Gamma\to\Gamma_0}\\
\xi_x &= \xi_{0,u}\|D^*\|_{\Gamma\to\Gamma_0} .
\end{aligned}$$

PROOF In the following, all operator bounds that appear are finite owing to the coercivity assumption on the involved gauges in Definition 4.2 of a PSFL.

- Let x' such that
$$\Gamma(x-x') \leqslant \frac{1}{\|D^*\|_{\Gamma\to\Gamma_0}}\nu_{0,D^*x}.$$
Hence,
$$\Gamma_0(D^*x - D^*x') \leqslant \|D^*\|_{\Gamma\to\Gamma_0}\Gamma(x-x') \leqslant \nu_{0,D^*x}$$
As J_0 is a PSFL at D^*x, we have $T_{0,D^*x} = T_{0,D^*x'} = T_0$ and consequently, using Proposition 3.11(i), $T_x = \mathrm{Ker}(D^*_{S_{0,D^*x}}) = \mathrm{Ker}(D^*_{S_{0,D^*x'}}) = T_{x'} = T = S^\perp$.

- μ_x**-stability:** we now have
$$\begin{aligned}
\Gamma(e_x - e'_x) &= \Gamma(P_T D(e_{0,D^*x} - e_{0,D^*x'})) && \text{Proposition 3.11(ii)}\\
&\leqslant \|P_T D\|_{\Gamma_0\to\Gamma}\Gamma_0(e_{0,D^*x} - e_{0,D^*x'})\\
&\leqslant \mu_{0,D^*x}\|P_T D\|_{\Gamma_0\to\Gamma}\Gamma_0(D^*x - D^*x') && \text{using } \mu_{0,D^*x}\text{-stability of } J_0\\
&\leqslant \mu_{0,D^*x}\|P_T D\|_{\Gamma_0\to\Gamma}\|D^*\|_{\Gamma\to\Gamma_0}\Gamma(x-x').
\end{aligned}$$

- τ_x**-stability:** since $f_{0,D^*x} \in \partial J_0(D^*x)$ and $f_{0,D^*x'} \in \partial J_0(D^*x')$, one has
$$f_{0,D^*x} - f_{0,D^*x'} = P_{S_0}(f_{0,D^*x} - f_{0,D^*x'}) + e_{0,D^*x} - e_{0,D^*x'}.$$

Thus, subadditivity yields

$$\begin{aligned}
J_{f_x}^{x,\circ}(P_S(f_x - f_{x'})) &= J_{f_x}^{x,\circ}(P_S\, D(f_{0,D^*x} - f_{0,D^*x'})) \\
&\leqslant J_{f_x}^{x,\circ}(P_S\, D\, P_{S_0}(f_{0,D^*x} - f_{0,D^*x'})) + J_{f_x}^{x,\circ}(P_S\, D(e_{0,D^*x} - e_{0,D^*x'})).
\end{aligned}$$

Using Proposition 3.11(iii), we get the following bound on the first term

$$\begin{aligned}
& J_{f_x}^{x,\circ}(P_S\, D\, P_{S_0}(f_{0,D^*x} - f_{0,D^*x'})) \\
&= \inf_{z \in \mathrm{Ker}(D) \cap S_0} J_{0,f_{D^*x}}^{D^*x,\circ}(D_{S_0}^+\, P_S\, D\, P_{S_0}(f_{0,D^*x} - f_{0,D^*x'}) + z) \\
&\leqslant J_{0,f_{D^*x}}^{D^*x,\circ}(D_{S_0}^+\, P_S\, D\, P_{S_0}(f_{0,D^*x} - f_{0,D^*x'})) \\
&\leqslant \left|\!\left|\!\left| D_{S_0}^+\, P_S\, D \right|\!\right|\!\right|_{J_{0,f_{D^*x}}^{D^*x,\circ} \to J_{0,f_{D^*x}}^{D^*x,\circ}} J_{0,f_{D^*x}}^{D^*x,\circ}(P_{S_0}(f_{0,D^*x} - f_{0,D^*x'}))
\end{aligned}$$

Using τ_{0,D^*x}-stability of J_0, we get

$$\begin{aligned}
& J_{f_x}^{x,\circ}(P_S\, D\, P_{S_0}(f_{0,D^*x} - f_{0,D^*x'})) \\
&\leqslant \tau_{0,D^*x} \left|\!\left|\!\left| D_{S_0}^+\, P_S\, D \right|\!\right|\!\right|_{J_{0,f_{D^*x}}^{D^*x,\circ} \to J_{0,f_{D^*x}}^{D^*x,\circ}} \Gamma_0(D^*x - D^*x') \\
&\leqslant \tau_{0,D^*x} \left|\!\left|\!\left| D_{S_0}^+\, P_S\, D \right|\!\right|\!\right|_{J_{0,f_{D^*x}}^{D^*x,\circ} \to J_{0,f_{D^*x}}^{D^*x,\circ}} |\!|\!| D^* |\!|\!|_{\Gamma \to \Gamma_0} \Gamma(x - x').
\end{aligned}$$

Now, combining Proposition 3.11(iii) and μ_{0,D^*x}-stability of J_0, we obtain the following bound on the second term

$$\begin{aligned}
J_{f_x}^{x,\circ}(P_S\, D(e_{0,D^*x} - e_{0,D^*x'})) &\leqslant J_{0,f_{D^*x}}^{D^*x,\circ}(D_{S_0}^+\, P_S\, D(e_{0,D^*x} - e_{0,D^*x'})) \\
&\leqslant \left|\!\left|\!\left| D_{S_0}^+\, P_S\, D \right|\!\right|\!\right|_{\Gamma_0 \to J_{0,f_{D^*x}}^{D^*x,\circ}} \Gamma_0(e_{0,D^*x} - e_{0,D^*x'}) \\
&\leqslant \mu_{0,D^*x} \left|\!\left|\!\left| D_{S_0}^+\, P_S\, D \right|\!\right|\!\right|_{\Gamma_0 \to J_{0,f_{D^*x}}^{D^*x,\circ}} |\!|\!| D^* |\!|\!|_{\Gamma \to \Gamma_0} \Gamma(x - x').
\end{aligned}$$

Combining these inequalities, we arrive at

$$\begin{aligned}
J_{f_x}^{x,\circ}(P_S(f_x - f_{x'})) \leqslant \Big(&\tau_{0,D^*x} \left|\!\left|\!\left| D_{S_0}^+\, P_S\, D \right|\!\right|\!\right|_{J_{0,f_{D^*x}}^{D^*x,\circ} \to J_{0,f_{D^*x}}^{D^*x,\circ}} \\
&+ \mu_{0,D^*x} \left|\!\left|\!\left| D_{S_0}^+\, P_S\, D \right|\!\right|\!\right|_{\Gamma_0 \to J_{0,f_{D^*x}}^{D^*x,\circ}} \Big) |\!|\!| D^* |\!|\!|_{\Gamma \to \Gamma_0} \Gamma(x - x'),
\end{aligned}$$

whence we get τ_x-stability.

- ξ_x**-stability:** from Proposition 3.11(iii), we can write for any $\eta \in S$

$$\begin{aligned} J_{f_{x'}}^{x',\circ}(\eta) &= \inf_{z \in \mathrm{Ker}(D) \cap S_0} J_{f_{0,D^* x'}}^{\circ}(D_{S_0}^{+}\eta + z) \\ &\leqslant J_{0,f_{D^* x'}}^{D^* x',\circ}(D_{S_0}^{+}\eta + \bar{z}) \end{aligned}$$

for any $\bar{z} \in \mathrm{Ker}(D) \cap S_0$.

Owing to ξ_{0,D^*x}-stability of J_0, and since $D_{S_0}^{+}\eta \in S_0$, we have for any feasible $\bar{z} \in \mathrm{Ker}(D) \cap S_0$

$$J_{0,f_{D^* x'}}^{D^* x',\circ}(D_{S_0}^{+}\eta + \bar{z}) \leqslant \left(1 + \xi_{0,D^*x}\Gamma_0(D^*x - D^*x')\right) J_{0,f_{D^* x}}^{D^* x,\circ}(D_{S_0}^{+}\eta + \bar{z}) \ .$$

Taking in particular

$$\bar{z} \in \operatorname*{Argmin}_{z \in \mathrm{Ker}(D) \cap S_0} J_{0,f_{D^* x}}^{D^* x,\circ}(D_{S_0}^{+}\eta + z)$$

we get the bound

$$\begin{aligned} J_{f_{x'}}^{x',\circ}(\eta) &\leqslant \left(1 + \xi_{0,D^*x}\Gamma_0(D^*x - D^*x')\right) \inf_{z \in \mathrm{Ker}(D) \cap S_0} J_{0,f_{D^* x}}^{D^* x,\circ}(D_{S_0}^{+}\eta + z) \\ &= \left(1 + \xi_{0,D^*x}\Gamma_0(D^*x - D^*x')\right) J_{f_{x'}}^{x',\circ}(\eta) \\ &= \left(1 + \xi_{0,D^*x} \| D^* \|_{\Gamma \to \Gamma_0} \Gamma(x - x')\right) J_{f_{x'}}^{x',\circ}(\eta) \ , \end{aligned}$$

where we used again Proposition 3.11(iii) in the first equality. ■

4.3 Examples

4.3.1 Synthesis Sparsity

The norm $J(x) = \|x\|_1$ is a strong partly smooth function.

> Proposition 4.6 $J = \|\cdot\|_1$ is a strong partly smooth function with
>
> $$\Gamma = \|\cdot\|_\infty, \quad \nu_x = \min_{i \in I} |x_i| \quad \text{and} \quad \mu_x = \tau_x = \xi_x = 0,$$

where $I = I(x) = \{i \mid x_i \neq 0\}$.

PROOF Let $x' \in T$, i.e. $I(x') \subseteq I(x)$, and assume that

$$\|x - x'\|_\infty < \nu_x = \min_{i \in I} |x_i| \ .$$

This implies that $\forall i \in I(x)$, $|x'_i| > \nu_x - \|x - x'\|_\infty \geqslant 0$, which in turn yields $I(x') = I(x)$, and thus $T_{x'} = T_x$. Since the sign is also locally constant on the restriction to T of the ℓ^∞-ball centered at x of radius ν_x, one can choose $\mu_x = 0$. Finally $\tau_x = \xi_x = 0$ because $f_x = e_x$. ∎

4.3.2 Analysis Sparsity

PROPOSITION 4.7 $J = \|D^* \cdot\|_1$ is a strong partly smooth function with parameters

$$\nu_x = \min_{i \in I} |\langle d_i,\, x_i \rangle| \quad \text{and} \quad \mu_x = \tau_x = \xi_x = 0.$$

PROOF This is a consequence of Proposition 4.5 with $J_0 = \|\cdot\|_1$. ∎

4.3.3 Antisparsity

PROPOSITION 4.8 $J = \|\cdot\|_\infty$ is a partly smooth function with linear manifold with

$$\Gamma = \|\cdot\|_1, \quad \nu_x = \left(\|x\|_\infty - \max_{j \notin I} |x_j|\right) \quad \text{and} \quad \mu_x = \tau_x = \xi_x = 0.$$

PROOF Let $x' \in T$, and assume that

$$\|x - x'\|_1 < \nu_x = \left(\|x\|_\infty - \max_{j \notin I} |x_j|\right) .$$

This means that x' lies in the relative interior of the ℓ^1-ball (relatively to T) centered at x of radius $\|x\|_\infty - \max_{j \notin I} |x_j|$. Within this ball, the support and the sign pattern restricted to the support are locally constant, i.e. $I(x) = I(x')$ and $\mathrm{sign}(x_{(I(x))}) = \mathrm{sign}(x'_{(I(x'))})$. Thus $T_{x'} = T_x = T$ and $e_{x'} = e_x$, and from the latter we deduce that $\mu_x = 0$. As $f_x = e_x$ we also conclude that $\tau_x = \xi_x = 0$, which completes the proof. ∎

4.3.4 Group Sparsity

The $\ell^1 - \ell^2$ norm is a strong partly smooth function. We start by the following lemma

LEMMA 4.2 Given any pair of non-zero vectors u and v where, $\|u - v\| \leqslant \rho\|u\|$, for $0 < \rho < 1$, we have

$$\left\| \frac{u}{\|u\|} - \frac{v}{\|v\|} \right\| \leqslant C_\rho \frac{\|u - v\|}{\|u\|},$$

where $C_\rho = \frac{\sqrt{2}}{\rho}\sqrt{1 - \sqrt{1 - \rho^2}} \in]1, \sqrt{2}[$.

PROOF Let $d = v - u$ and $\beta = \frac{\langle u, d \rangle}{\|u\|\|d\|} \in [-1, 1]$. We then have the following identities

$$\left\| \frac{u}{\|u\|} - \frac{v}{\|v\|} \right\|^2 = 2 - 2\frac{\langle u, v \rangle}{\|u\|\|v\|} = 2 - 2\frac{\|u\|^2 + \|u\|\|d\|\beta}{\|u\|\sqrt{\|u\|^2 + \|d\|^2 + 2\|u\|\|d\|\beta}}, \tag{4.6}$$

for non-zero vectors u and d, the unique maximizer of (4.6) is $\beta^\star = -\|d\|/\|u\|$. Note that the assumption $\|d\|/\|u\| \leqslant \rho < 1$ assures $\beta^\star$ to comply with the admissible range of β and further, the argument of the square root will be always positive.

Now, inserting $\beta^\star$ in (4.6), using concavity of $\sqrt{\cdot}$ on $\mathbb{R}_+$, and that $\|d\|/\|u\| \leqslant \rho$, we can deduce the following bound

$$\begin{aligned}\left\|\frac{u}{\|u\|}-\frac{v}{\|v\|}\right\|^2 \leqslant 2-2\sqrt{1-\frac{\|d\|^2}{\|u\|^2}} &= 2-2\sqrt{\left(1-\frac{\|d\|^2}{\rho^2\|u\|^2}\right)+\frac{\|d\|^2}{\rho^2\|u\|^2}(1-\rho^2)} \\ &\leqslant 2-2\left(\left(1-\frac{\|d\|^2}{\rho^2\|u\|^2}\right)+\frac{\|d\|^2}{\rho^2\|u\|^2}\sqrt{1-\rho^2}\right) \\ &= 2-2\left(1-\frac{1-\sqrt{1-\rho^2}}{\rho^2}\frac{\|d\|^2}{\|u\|^2}\right) \\ &= 2\frac{1-\sqrt{1-\rho^2}}{\rho^2}\frac{\|d\|^2}{\|u\|^2} \ .\end{aligned}$$

Proposition 4.9 The $\ell^1-\ell^2$ norm associated to the partition $\mathcal{B}$ is a strong partly smooth function with

$$\Gamma = \|\cdot\|_{\infty,2}, \quad \nu_x = \min_{b\in I}\|x_b\|, \quad \mu_x = \frac{\sqrt{2}}{\nu_x} \quad \text{and} \quad \tau_x = \xi_x = 0.$$

where $I = I(x) = \{b \mid \ x_b \neq 0\}$.

Proof Let $x' \in T$, i.e. $I(x') \subseteq I(x)$, and $\nu_x = \min_{b\in I}\|x_b\|$. First, observe that the condition

$$\|x-x'\|_{\infty,2} = \max_{b\in\mathcal{B}}\|x_b - x'_b\| < \nu_x$$

ensures that for all $b \in I$

$$\|x'_b\| \geqslant \|x_b\| - \|x_b - x'_b\| > \nu_x - \|x-x'\|_{\infty,2} \geqslant 0,$$

and thus $I(x') = I(x)$, i.e. $T_{x'} = T_x$. Moreover, since the gauge is strong, one has $\tau_x = \xi_x = 0$. To establish the μ_x-stability we use Lemma 4.2.

By definition of ν_x, we have $\|x_b\| > \nu_x$, $\forall b \in I$, and thus $\|x_b - x'_b\| < \nu_x < \|x_b\|$. Lemma 4.2 then applies, and it follows that, $\forall b \in I$

$$\|\mathcal{N}(x_b) - \mathcal{N}(x'_b)\| \leqslant C_\rho \frac{\|x'_b - x_b\|}{\|x_b\|} \leqslant C_\rho \frac{\|x'_b - x_b\|}{\nu_x},$$

and therefore we get

$$||\mathcal{N}(x) - \mathcal{N}(x')||_{\infty,2} \leqslant \frac{C_\rho}{\nu_x} ||x' - x||_{\infty,2},$$

which implies μ_x-stability for $\mu_x = C_\rho/\nu_x$. ∎

4.3.5 Polyhedral Regularizations

PROPOSITION 4.10 $J_0(u) = \max_{1 \leqslant i \leqslant N_H} (u_i)_+$ is a partly smooth function with linear manifold with parameters (assuming $I_+ \neq \emptyset$)

$$\nu_u = \big(\max_{i \in I_+} u_i - \max_{j \notin I_+, u_j > 0} u_j\big), \delta \in]0,1] \quad \text{and} \quad \mu_u = \tau_u = \xi_u = 0.$$

PROOF The parameters are derived following the same lines as for the ℓ^∞-norm. Let $u' \in T$, and assume that

$$||u - u'||_1 < \nu_u = \left(\max_{i \in I_+} u_i - \max_{j \notin I_+, u_j > 0} u_j \right).$$

This means that x' lies in the relative interior of the ℓ^1-ball (relatively to T) centered at x of radius

$$\max_{i \in I_+} u_i - \max_{j \notin I_+, u_j > 0} u_j = ||u||_\infty - \max_{j \notin I_+, u_j > 0} |u_j|$$

Within this set, one can observe that the set I_+ associated to u is constant. Moreover, the sign pattern is also constant leading to the fact that $T_{u'} = T_u = T$. Hence, we deduce as in the ℓ^∞-case that $\mu_u = \tau_u = \xi_u = 0$. ∎

4.3.6 Nuclear Norm

Proposition 4.11 Let $x \in \mathbb{R}^{n \times n}$ The nuclear norm is partly smooth at x for the manifold

$$\mathcal{M} = \left\{ u \in \mathbb{R}^{n \times n} \mid \operatorname{rank}(u) = \operatorname{rank}(x) \right\}.$$

PROOF This is a direct consequence of Proposition 4.3 using $j(\Lambda_x) = \|\Lambda_x\|_1$. ■

However, one should note that the nuclear norm is not a partly smooth function with linear manifold. Indeed, according to Proposition 4.11, the model subspace reads

$$T_x = \left\{ UA^* + BV^* \mid A \in \mathbb{R}^{n \times r}, B \in \mathbb{R}^{n \times r} \right\},$$

which in particular contains matrices of ranks larger than r.

J	$\mathcal{M}$	T_x	e_x	$J_{f_x}^{x,\circ}$ on S_x	ν_x	μ_x	Comment
$\|\cdot\|_1$	$\{\eta \mid \forall j \notin I, \eta_j = 0\}$	$\mathcal{M}_x$	$\mathrm{sign}(x)$	$\|\cdot\|_\infty$	$\min_{i \in I} \lvert x_i \rvert$	0	$I = \mathrm{supp}(x)$
$\|D^* \cdot\|_1$	$\mathrm{Ker}(D_{I^c}^*)$	$\mathcal{M}_x$	$P_{\mathrm{Ker}(D_{I^c}^*)} \mathrm{sign}(D^* x)$	$\inf_{z \in \mathrm{Ker}(D_{I^c})} \|D_{I^c}^+ \eta + z\|_\infty$	$\min_{i \in I} \lvert\langle d_i, x_i \rangle\rvert$	0	$I = \mathrm{supp}(D^* x)$
$\|\cdot\|_{1,2}$	$\{\eta \mid \forall j \notin I, \eta_j = 0\}$	$\mathcal{M}_x$	$(\mathcal{N}(x_b))_{b \in \mathcal{B}}$	$\|\cdot\|_{\infty,2}$	$\min_{b \in I} \|x_b\|$	$\frac{\sqrt{2}}{\nu_x}$	$I = \{g \in \mathcal{B} \mid x_g \neq 0\}$
$\|\cdot\|_\infty$	$\{\alpha \mid \alpha_I = \rho s_I \text{ for } \rho \in \mathbb{R}\}$	$\mathcal{M}_x$	$\mathrm{sign}(x)/\lvert I \rvert$	$\max_{i \in I} (-\lvert I \rvert s_i \eta_i)_+$	$\left(\|x\|_\infty - \max_{j \notin I} \lvert x_j \rvert\right)$	0	$I = \{i \mid \lvert x_i \rvert = \|x\|_\infty\}$
$\|\cdot\|_*$	$\{Z \mid \mathrm{rank}(Z) = \mathrm{rank}(x)\}$	$\{Z \mid U_\perp^* Z V_\perp = 0\}$	UV^*	$\|\cdot\|_{sp}$			$x = U \Lambda V^*$

Table 4.1: *Examples of Partly Smooth Functions. For all these regularizations,* $\tau_x = \xi_x = 0$.

5

Certificates and Uniqueness

Main contributions of this chapter

- Introduction of (non-degenerate) dual certificates (Definition 5.1), minimal norm certificate (Definition 5.2), linearized precertificate (Definition 5.4) and its associated identifiability criterion (Definition 5.5).
- Introduction of the restricted injectivity assumption (Definition 5.3).
- Theorem 5.3 gives a sufficient condition for uniqueness for $(\mathcal{P}_{y,\lambda})$ or $(\mathcal{P}_{y,0})$.

Contents

THIS chapter introduces in particular the notion of dual certificates which is connected to the solution of the dual problem of $(\mathcal{P}_{y,0})$. Moreover, we provide a sufficient condition for uniqueness of problem $(\mathcal{P}_{y,\lambda})$ and $(\mathcal{P}_{y,0})$.

5.1 Primal Problem

We consider $J \in \Gamma_c^+(\mathbb{R}^n)$. Let us split $y = y_0 + w$ where $y_0 = \Phi x_0$. We supose that

$$\mathrm{Ker}(\Phi) \cap \mathrm{Ker}(J_\infty) = \{0\}, \tag{5.1}$$

We rewrite problems $(\mathcal{P}_{y,\lambda})$ and $(\mathcal{P}_{y,0})$ as a common regularization problem

$$\min_{x \in \mathbb{R}^n} f(x,\theta) \quad \text{where} \quad f(x,\theta) = \begin{cases} J(x) + \iota_{\mathcal{H}_y}(x) & \text{if } \theta = (y,0) \\ J(x) + \frac{1}{2\lambda}\|\Phi x - y\|^2 & \text{otherwise,} \end{cases} \tag{$\mathcal{P}_\theta$}$$

where $\mathcal{H}_y = \{x \in \mathbb{R}^n \mid \Phi x = y\}$ and $\theta = (y,\lambda) \in \mathbb{R}^q \times \mathbb{R}_+$.

5.1.1 Existence of Solutions

We recall that in general ($\mathcal{P}_\theta$) might have multiple solutions. Here, based on classical compactness arguments, we show that the set of minimizers of ($\mathcal{P}_\theta$) is non-empty, compact and convex.

PROPOSITION 5.1 Let $f \in \Gamma_0(\mathbb{R}^n)$. Then,

(i) If $\inf f > -\infty$, then $f_\infty(d) \geqslant 0, \quad \forall d$.

(ii) The set of minimizers of f is non-empty and compact $\iff$ f is coercive $\iff$ the sublevel sets of f are bounded $\iff$ $f_\infty(d) > 0, \forall d \neq 0$.

PROOF (i) The statement follows from the equivalent analytic representation of f_∞ in Proposition 2.4(i).

(ii) (Auslender et al. 2003, Proposition 3.1.2 and Proposition 3.1.3). ∎

Let us now turn to the minimization problem

$$\min_{x \in \mathbb{R}^n} F(\Phi x) + J(x) \tag{5.2}$$

where $F \in \Gamma_0(\mathbb{R}^p)$ and strictly convex, $J \in \Gamma_0(\mathbb{R}^n)$ and continuous on $\mathbb{R}^n$, $\inf J > -\infty$, and $\Phi : \mathbb{R}^n \to \mathbb{R}^p$.

LEMMA 5.1 The set of minimizers of (5.2) is non-empty and compact if and only if

$$\mathrm{Ker}(J_\infty) \cap \mathrm{Ker}(\Phi) = \{0\} ,$$

where J_∞ is given by either expressions of Proposition 2.4(i) or (ii).

PROOF By strict convexity of F, all minimizers of (5.2) share the same image under Φ. Let $x^\star$ any minimizer. Thus, (5.2) can be equivalently rewritten

$$\min_{\delta \in \mathrm{Ker}(\Phi)} J(x^\star + \delta) .$$

∎

Let $f = J(x^\star + \cdot) + \iota_{\mathrm{Ker}(\Phi)}$. Thus, owing to Proposition 5.1(ii)

$$\begin{aligned}
\text{the set of minimizers is compact} &\iff f_\infty(d) > 0 \quad \forall d \neq 0 \\
\text{Proposition 2.4(v)} &\iff J_\infty(d) + (\iota_{\mathrm{Ker}(\Phi)})_\infty(d) > 0 \quad \forall d \neq 0 \\
\text{Proposition 2.4(iii)} &\iff J_\infty(d) + \iota_{\mathrm{Ker}(\Phi)_\infty}(d) > 0 \quad \forall d \neq 0 \\
\text{Proposition 2.3(iv)} &\iff J_\infty(d) + \iota_{\mathrm{Ker}(\Phi)}(d) > 0 \quad \forall d \neq 0 \\
&\iff J_\infty(d) > 0 \quad \forall d \in \mathrm{Ker}(\Phi) \setminus \{0\} \\
\text{Proposition 5.1(i)} &\iff \mathrm{Ker}(J_\infty) \cap \mathrm{Ker}(\Phi) = \{0\}\ .
\end{aligned}$$

This condition (5.1) is from now on assumed in all propositions.

5.1.2 Convergence of the Primal Problem

We first show the convergence of the solutions of the primal problem toward x_0 when $(\mathcal{P}_{y_0,0})$ has a unique solution x_0.

PROPOSITION 5.2 Assume that x_0 is the unique solution of $(\mathcal{P}_{y_0,0})$. Let $\theta_k = (\lambda_k, y_k)$ a sequence such that $0 < \lambda_k \to 0$ and $||y_k - y_0||^2/\lambda_k \to 0$. Then, for any sequence $(x_{\theta_k})_k$ of minimizers to $(\mathcal{P}_{\theta_k})$,

$$x_{\theta_k} \longrightarrow x_0.$$

In order to ease the exposition, we will write in the following this convergence statement with the following slight abuse of notation.

$$x_\theta \to x_0 \quad \text{when} \quad \begin{cases} \lambda \to 0 \\ \frac{||w||^2}{\lambda} \to 0. \end{cases}$$

PROOF This is a classical result, whose proof can be found for instance in (Hofmann et al. 2007, Theorem 3.5). We recall it by sake of clarity.

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

$$||\Phi(x_\theta - x_0) - w||^2 \leqslant ||w||^2 + 2\lambda J(x_0), \tag{5.3}$$

$$J(x_\theta) \leqslant \frac{||w||^2}{2\lambda} + J(x_0). \tag{5.4}$$

Thanks to (5.1), these bounds show that the sequence $\{x_\theta\}_\theta$ is bounded if $||w||^2/\lambda$ and λ are bounded. We let $x^\star$ be any accumulation point.

For the considered asymptotics, (5.3) implies that $\Phi x^\star = \Phi x_0$, while (5.4) implies that $J(x^\star) \leqslant J(x_0)$. This shows that $x^\star$ is a solution of $(\mathcal{P}_{y_0,0})$ and hence $x^\star = x_0$. ■

5.2 Certificates and Restricted Injectivity

This section introduces the two main objects of the noise stability study. The first one is the dual certificate, which characterizes the set of solutions of $(\mathcal{P}_{y,0})$. The second one is the restricted injectivity condition, to be able to ensure the uniqueness.

5.2.1 Fenchel–Rockafellar Duality

We characterize the dual problem in the following Lemma.

LEMMA 5.2 Let $\theta = (\lambda, y)$ with $\lambda \geqslant 0$. The dual problem of $(\mathcal{P}_\theta)$ reads

$$\min_{p \in \mathbb{R}^q} g(p, \theta) \tag{$\mathcal{D}_\theta$}$$

where

$$g(p,\theta) = \begin{cases} J^*(\Phi^* p) - \langle y,\, p\rangle & \text{if } \lambda = 0 \\ J^*(\Phi^* p) - \langle y,\, p\rangle + \frac{\lambda}{2}||p||^2 & \text{otherwise.} \end{cases}$$

Moreover, there is no duality gap, i.e.

$$\min_{x \in \mathbb{R}^n} f(x, \theta) = -\min_{p \in \mathbb{R}^q} g(p, \theta).$$

Observe that domain qualification conditions (on their relative interiors) to ensure closedness of the dual objective (i.e. the min is attained) are verified for the penalized problem since $\frac{1}{2}\|\cdot\|_2^2$ has full domain.

PROOF The proof of this result is a simple application of the calculus rules on Fenchel–Rockafellar duality. The case $\lambda = 0$ is the Fenchel–Rockafellar duality for linear constraints, see for instance (Borwein et al. 2010, Corollary 3.3.11). The case $\lambda > 0$ is due to the fact that $(1/2\|\cdot\|^2)^* = 1/2\|\cdot\|^2$. ∎

We now relate the solutions of the primal problem $(\mathcal{P}_\theta)$ to those of the dual $(\mathcal{D}_\theta)$.

PROPOSITION 5.3 Let $\theta = (\lambda, y)$ with $\lambda \geqslant 0$ and x_θ any solution of $(\mathcal{P}_\theta)$. Then,

(i) if $\lambda > 0$, then $(\mathcal{D}_\theta)$ has a unique solution $\mathcal{S}_\theta = \{\mathfrak{p}_\theta\}$ and

$$\mathfrak{p}_\theta = \frac{y - \Phi x_\theta}{\lambda} \quad \text{and} \quad \alpha_\theta = \Phi^* \mathfrak{p}_\theta \in \partial J(x_\theta).$$

(ii) if $\lambda = 0$, then the set of solutions of $(\mathcal{D}_\theta)$ is

$$\mathcal{S}_\theta = \{\mathfrak{p} \in \mathbb{R}^q \mid \Phi^* \mathfrak{p} \in \partial J(x_\theta)\}.$$

PROOF For the first statement, since J is finite-valued, strong duality holds, hence the result using Fenchel-Rockafellar duality. Similarly, strong duality holds between $(\mathcal{P}_{0,y_0})$ and $(\mathcal{D}_{0,y_0})$, and the primal-dual relationships states that (x_{0,y_0}, p_{0,y_0}) form a solution to these problems if and only if $\Phi^* p_{0,y_0} \in \partial J(x_{0,y_0})$. ∎

5.2.2 Dual Certificates

These observations lead us to consider the notion of dual certificate, a terminology introduced in (Candès et al. 2006a) and revitalized in (Candès and

Recht 2013), which corresponds to Lagrange multipliers, which are solution of the dual problem.

Definition 5.1 A *(dual) certificate* for $x \in \mathbb{R}^n$ is a vector $p \in \mathbb{R}^q$ such that the *source condition* is verified:

$$\Phi^* p \in \partial J(x). \tag{SC_x}$$

If p is a certificate, and moreover

$$\Phi^* p \in \operatorname{ri} \partial J(x), \tag{$\overline{SC}_x$}$$

we say that p is a *non-degenerate certificate*.

Hence, according to Proposition 5.3, being a dual certificate is equivalent to be a solution of the dual problem $(\mathcal{D}_0)$ where $x_0 = x, y = y_0$. One important certificate is the minimal norm certificate defined as follow

Definition 5.2 The *minimal norm certificate* for $x_0 \in \mathbb{R}^n$ is defined by

$$p_0 = \underset{p \in \mathcal{S}_{0,y_0}}{\operatorname{argmin}} \|p\|.$$

Since $\mathcal{S}_0$ is a convex set, and p_0 is the projection of 0 onto it, p_0 is well-defined as a single-valued mapping. Moreover, we prove the following proposition, related to the convergence of the dual vectors associated to a solution of $(\mathcal{P}_{y,\lambda})$ to the minimal norm certificate.

Proposition 5.4 One has

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

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

PROOF This result is already proved by Duval et al. (2013) in the special case where J is the TV norm of a Radon measure (an infinite dimensional Banach space). By extension of the Definition 5.2, we denote

$$p_{0,y} = \underset{p \in \mathcal{S}_{0,y}}{\operatorname{argmin}} ||p||.$$

Formulation $(\mathcal{D}_\theta)$ shows that p_θ is the output of proximal operator of the function $J^*(\Phi^* \cdot)/\lambda$ applied at the point y/λ, This shows that $y/\lambda \mapsto p_\theta$ is 1-Lipschitz, see Proposition 10.1, and hence

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

We now prove that

$$p_\theta \overset{\lambda \to 0}{\longrightarrow} p_{0,y},$$

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

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

$$-\langle p_{0,y}, y \rangle \leqslant -\langle p_\theta, y \rangle. \tag{5.5}$$

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

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

or equivalently

$$||p_\theta|| \leqslant ||p_{0,y}||. \tag{5.6}$$

This shows that $\{p_\theta\}_\theta$ is bounded. Let $p^\star$ be any cluster point. Operating as in the proof of Proposition 5.2, we have $\forall \bar{x} \in \{x \mid y = \Phi x\}$

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

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

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

or equivalently, that $x_{0,y}$ is a minimizer of $(\mathcal{P}_{y,0})$. Morever, from the primal-dual extremality relationships, we have $\Phi^* p_\theta \in \partial J(x_\theta)$. Since J is a proper closed convex function, the graph of ∂J is sequentially closed, which yields $\Phi^* p^\star \in \partial J(x_{0,y})$, i.e. $p^\star \in \mathcal{S}_{0,y}$. Now (5.6) implies that $\|p^\star\| \leqslant \|p_{0,y}\|$ and hence $p^\star = p_{0,y}$, which shows that p_θ is converging to $p_{0,y}$. ∎

The following lemma gives a useful characterization of non-degenerate dual vectors.

Lemma 5.3 Let $J \in \Gamma_c^+(\mathbb{R}^n)$ and $x \in \mathbb{R}^n$. Then,

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

Note that if J is a gauge, u can be normalized in the lemma, e.g. by restricting it to the unit sphere.

PROOF First, recall that the directional derivative $J'(x,u)$ of J at x in the direction u is

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

From the characterization of the relative interior of a non-empty closed convex set (Hiriart-Urruty et al. 2001, Theorem V.2.2.3) or (Rockafellar 1996, Theorem 13.1), and sublinearity we deduce that

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

Using Theorem 3.1 shows that

$$J'(x,u) = \langle e_x, u \rangle + \max_{\eta \in P_S(\partial J(x))} \langle \eta, u \rangle \ .$$

Sublinearity implies that (Hiriart-Urruty et al. 2001, Corollary V.1.1.5)

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

Thus

$$J'(x,u) + J'(x,-u) = \max_{\eta \in P_S(\partial J(x))} \langle \eta, u \rangle - \min_{\eta \in P_S(\partial J(x))} \langle \eta, u \rangle ,$$

whence we obtain

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

Piecing everything together, we get

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

which is the statement announced. ■

5.2.3 Restricted Injectivity

Let us consider $(\mathcal{P}_{y,0})$ when $J = \|\cdot\|_1$. Thus, we want to recover some vector $x_0 \in \mathbb{R}^n$ from the observations $y = \Phi x_0$. Assume that we know the support I_0 of x_0. Remark that $x_0 \in \operatorname{span}(u_i)_{i \in I_0} \cap \{x \mid y = \Phi x_0\}$ where $(u_i)_{i \in \{1,\dots,n\}}$ is the canonical basis. Hence, to be uniquely recovered, one needs that $\Phi_{(I_0)}$ has full rank. Conversely, if $\Phi_{(I_0)}$ has not full rank, then any vector of the form $x_0 + h$ with $h \in \operatorname{Ker} \Phi_{(I_0)}$ will be solution of $(\mathcal{P}_{y,0})$.

In general, this idea leads us to consider the following condition.

> Definition 5.3 A subspace $T \subseteq \mathbb{R}^n$ satisfies the *restricted injectivity condition* (INJ_T) if Φ is injective on T.

For instance, (INJ_T) is equivalent to $\Phi_{(I)}$ being full rank in the case of the ℓ^1-norm, or equivalent to $\operatorname{Ker} \Phi \cap \operatorname{Ker} D^*_{(J)}$ for the analysis ℓ^1-norm, where J is some cosupport.

5.3 Uniqueness

In this section, we provide several results on the uniqueness of the solutions of $(\mathcal{P}_{y,\lambda})$ and $(\mathcal{P}_{y,0})$.

5.3.1 Sublevel Set and its Cones

In the following, we draw a connection between the sublevel sets of a convex function and the uniqueness of the problem $(\mathcal{P}_{y,\lambda})$.

The following proposition summarizes some key properties of the above cones when generated from the sublevel set of a continuous convex function. It will play a pivotal role in our proof of uniqueness (see Theorem 5.1 and Theorem 5.2).

PROPOSITION 5.5 Let J be a continuous convex function on $\mathbb{R}^n$. Then,

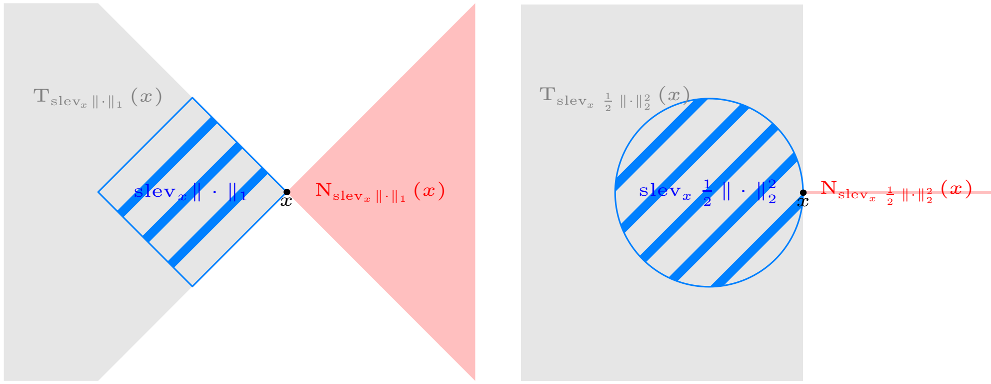
$$T_{\mathrm{slev}_x\, J}(x) \subset \left\{ \delta \mid J'(x,\delta) \leqslant 0 \right\} . \tag{5.7}$$

PROOF See (Hiriart-Urruty et al. 2001, Proposition III.5.3.1). ∎

Figure 5.1 illustrates the tangent cone for the ℓ^1-norm and a quadratic regularization $\frac{1}{2}||x||_2^2$.

THEOREM 5.1 Let J be a continuous convex function on $\mathbb{R}^n$. If Φ is injective on $T_{\mathrm{slev}_{x^\star}\, J}(x^\star)$ then $x^\star$ is the unique minimizer of $(\mathcal{P}_\theta)$. In particular, If Φ is injective on $T_{\mathrm{slev}_{x_0}\, J}(x_0)$ x_0 is the unique minimizer of $(\mathcal{P}_{y,0})$.

PROOF We provide the proof for $(\mathcal{P}_\theta)$ when $\lambda > 0$, the proof is similar when $\lambda = 0$. According to Proposition 2.2, any other minimizer different from $x^\star$ can be written



(a) (b)

Figure 5.1: *Tangent cone generated at* $x = (1, 0)$ *and its polar for two functions* J*: a)* $\|x\|_1$ *and b)* $\frac{1}{2}\|x\|_2^2$*. In both cases, we have the equality* (5.7) *since* $(1, 0) \notin (\partial J)^{-1}(0) = \{(0, 0)\}$.

as $x^\star + \delta$, where $\delta \in \mathrm{Ker}(\Phi) \setminus \{0\}$, and $J(x^\star + \delta) = J(x^\star)$. Therefore, we have

$$
\begin{aligned}
&\delta \notin T_{\mathrm{slev}_{x^\star} J}(x^\star), \quad \forall \delta \in \mathrm{Ker}(\Phi) \setminus \{0\}\,, \\
\Rightarrow &J(x^\star + \delta) > J(x^\star), \quad \forall \delta \in \mathrm{Ker}(\Phi) \setminus \{0\} \\
\Rightarrow &J \text{ has a unique minimizer } x^\star,
\end{aligned}
$$

which concludes our proof. ∎

The last statement coincides with that of (Chandrasekaran et al. 2012, Proposition 2.1) for atomic norms.

5.3.2 The Strong Nullspace Property

We are going to restate the previous Theorem 5.1 in a more meaningful way. We now compute the directional derivative of a bounded convex function J.

Lemma 5.4 Let $J \in \Gamma_c^+(\mathbb{R}^n)$. The directional derivative $J'(x,\delta)$ at point $x \in \mathbb{R}^N$ in the direction δ reads

$$J'(x,\delta) = \langle e_x, \delta_{T_x}\rangle + \langle P_{S_x}(f_x), \delta_{S_x}\rangle + J^x_{f_x}(\delta_{S_x}).$$

PROOF This comes directly from the structure of $J^x_{f_x}$. Indeed, one has

$$\begin{aligned} J^x_{f_x}(\delta_{S_x}) &= J^x_{f_x}(\delta) && \text{Using Proposition 3.8(ii)} \\ &= \sup_{\eta \in \partial J(x) - \{f_x\}} \langle \eta, \delta\rangle \\ &= -\langle \delta, f_x\rangle + \sup_{\eta \in \partial J(x)} \langle \eta, d\rangle \\ &= -\langle \delta, f_x\rangle + J'(x,\delta) \\ &= -\langle e_x, \delta_{T_x}\rangle - \langle P_{S_x}(f_x), \delta_{S_x}\rangle + J'(x,\delta)\ , \end{aligned}$$

which concludes our proof. ■

The following condition is a generalization of the Null Space Property well-known for ℓ_1 regularization (Donoho et al. 2001).

Theorem 5.2 Let $J \in \Gamma_c^+(\mathbb{R}^n)$. For a minimizer $x^\star$ of $(\mathcal{P}_{y,\lambda})$ (resp. a feasible point of $(\mathcal{P}_{y,0})$), let $T = S^\perp$, e and f the subspace and vectors associated to it. If the *Strong Null Space Property* holds

$$\forall \delta \in \mathrm{Ker}(\Phi) \setminus \{0\}, \quad \langle e, \delta_T\rangle + \langle P_S(f), \delta_S\rangle < J^x_f(-\delta_S), \qquad (\mathrm{NSP}^S)$$

then $x^\star$ is the unique minimizer of $(\mathcal{P}_{y,\lambda})$ (resp. $(\mathcal{P}_{y,0})$).

PROOF From Lemma 5.4, the directional derivative $J'(x,\delta)$ at $x \in \mathbb{R}^N$ in the direction δ reads

$$J'(x,\delta) = \langle e, \delta_T\rangle + \langle P_S f, \delta_S\rangle + J^x_{f_x}(\delta_S)\ .$$

Combining (5.7) in Proposition 5.5(i) and (68), applied at $x^\star$, together with the fact that $\mathrm{Ker}(\Phi)$ is a subspace yield

$$\begin{aligned} &\forall \delta \in \mathrm{Ker}(\Phi)\setminus\{0\}, \langle e,\, \delta_T\rangle + \langle P_S f,\, \delta_S\rangle < J^x_{f_x}(-\delta_S) \\ \iff &\forall \delta \in \mathrm{Ker}(\Phi)\setminus\{0\}, \quad J'(x,\delta) > 0 \\ \implies &T_{\mathrm{slev}_{x^\star} J}(x^\star) \cap \mathrm{Ker}(\Phi) = \{0\} \ . \end{aligned}$$

We then conclude using Theorem 5.1. ∎

5.3.3 Topological Conditions

A direct consequence of Theorem 5.2 above is the following corollary.

> THEOREM 5.3 Let $J \in \Gamma_c^+(\mathbb{R}^n)$. For a minimizer $x^\star$ of $(\mathcal{P}_{y,\lambda})$ (resp. a feasible point of $(\mathcal{P}_{y,0})$), let $T = S^\perp$, e and f the subspace and vector associated to it. Assume that $(\overline{\mathrm{SC}}_{x^\star})$ is verified with $\eta = \Phi^* p \in \mathrm{ri}\, \partial J(x^\star)$, and that (INJ_T) holds. Then, $x^\star$ is the unique minimizer of $(\mathcal{P}_{y,\lambda})$ (resp. $(\mathcal{P}_{y,0})$).

PROOF The source condition $(\mathrm{SC}_{x^\star})$ implies that $\forall\ \delta \in \mathrm{Ker}(\Phi)\setminus\{0\}$

$$\langle \delta,\, \eta\rangle = \langle \delta,\, \Phi^* p\rangle = \langle \Phi\delta,\, p\rangle = 0\ .$$

Moreover

$$\langle \delta,\, \eta\rangle = \langle \delta_T,\, e\rangle + \langle \delta_S,\, \eta_S\rangle = \langle \delta_T,\, e\rangle + \langle \delta_S,\, P_S\, f\rangle + \langle \delta_S,\, \eta_S - P_S\, f\rangle\ .$$

Thus, applying the duality inequality of gauges we get

$$\langle \delta_T,\, e\rangle + \langle \delta_S,\, P_S\, f\rangle \leqslant J^x_{f_x}(-\delta_S) J^\circ_f(\eta_S - P_S\, f) < J^x_{f_x}(-\delta_S)\ ,$$

where the last inequality is strict since δ_S does not vanish owing to (INJ_T), and $\alpha \in \mathrm{ri}\, \partial J(x^\star)$ is equivalent to $\eta_T = e$ and $J^{x,\circ}_{f_x}(\eta_S - P_S\, f) < 1$. ∎

The topological condition on the dual certificate required in Theorem 5.3 can be weakened to hold only on a subspace $V \subset S$ and the conclusions of the corollary remain valid, and assuming a stronger restricted injectivity assumption. We have the following corollary of of Theorem 5.2.

Corollary 5.1 With the same notations as in Theorem 5.2, suppose that J is such that $J^x_{f_x}$ is separable on $S = V \oplus W$. Assume that $(\overline{\text{SC}}_{x^\star})$ is verified with $J^{x,\circ}_{f_x}(\eta_V - P_V(f)) < 1$, and (INJ_V) holds. Then, $x^\star$ is the unique minimizer of $(\mathcal{P}_{y,\lambda})$ (resp. $(\mathcal{P}_{y,0})$).

PROOF We follow the same lines as the proof of Corollary 5.3 and get

$$\langle \delta_T, \eta \rangle = \langle \delta_T, e \rangle + \langle \delta_S, P_S\, f \rangle + \langle \delta_V, \eta_V - P_V\, f \rangle + \langle \delta_W, \eta_W - P_W\, f \rangle \ .$$

Moreover, by separability of J_f on S, we have

$$\begin{aligned}
J^{x,\circ}_{f_x}(\eta_S) &= \sup_{J^x_{f_x}(d) \leqslant 1} \langle d, \eta_S \rangle \\
&= \sup_{J^x_{f_x}(d_S) \leqslant 1} \langle d_S, \eta_S \rangle \\
&= \sup_{J^x_{f_x}(d_V) + J^x_{f_x}(d_W) \leqslant 1} \langle d_V, \eta_V \rangle + \langle d_W, \eta_W \rangle \\
&= \sup_{\rho \in [0,1]} \sup_{J^x_{f_x}(d_V) \leqslant \rho, J^x_{f_x}(d_W) \leqslant 1-\rho} \langle d_V, \eta_V \rangle + \langle d_W, \eta_W \rangle \\
&= \sup_{\rho \in [0,1]} \rho \sup_{J^x_{f_x}(d_V) \leqslant 1} \langle d_V, \eta_V \rangle + (1-\rho) \sup_{J^x_{f_x}(d_W) \leqslant 1} \langle d_W, \eta_W \rangle \\
&= \sup_{\rho \in [0,1]} \rho J^{x,\circ}_{f_x}(\eta_V) + (1-\rho) J^{x,\circ}_{f_x}(\eta_W) \\
&= \max(J^{x,\circ}_{f_x}(\eta_V), J^{x,\circ}_{f_x}(\eta_W)) \ .
\end{aligned}$$

This implies in particular that

$$J^{x,\circ}_{f_x}(\eta_W - P_W\, f) \leqslant \max(J^{x,\circ}_{f_x}(\eta_V), J^\circ_f(\eta_W)) = J^{x,\circ}_{f_x}(\eta_S - P_S) \leqslant 1 \ .$$

We therefore obtain

$$\begin{aligned}\langle \delta_T, e\rangle + \langle \delta_S, P_S f\rangle &\leqslant J_{f_x}^x(-\delta_V) J_{f_x}^{x,\circ}(\eta_V - P_V f) + J_{f_x}^x(-\delta_W) J_{f_x}^{x,\circ}(\eta_W - P_W f) \\ &< J_{f_x}^x(-\delta_V) + J_{f_x}^x(-\delta_W) = J_{f_x}^x(-\delta_S)\ ,\end{aligned}$$

where we used that $\delta \notin T$, $J_{f_x}^{x,\circ}(\eta_V - P_V f) < 1$ and separability of $J_{f_x}^x$ on S. ∎

5.4 Construction of Non-Degenerate Certificates

5.4.1 Linearized Precertificate

Let us first introduce the definition of the linearized precertificate.

DEFINITION 5.4 The *linearized precertificate* $\mathfrak{p}_F$ for $x \in \mathbb{R}^n$ is defined by

$$\mathfrak{p}_F = \underset{(\Phi^*\mathfrak{p})_{T_x} = e_x}{\operatorname{argmin}} \ ||\mathfrak{p}||.$$

The intuition behind this definition is well-understood if one realizes that the existence of a dual certificate $\mathfrak{p}$ is equivalent to $\eta = \Phi^*\mathfrak{p}$ for some $\mathfrak{p}$ such that $\eta_T = e_x$ and $J_{f_x}^\circ(\eta_S - f_S) \leqslant 1$. Dropping the last constraint, we recover the definition of $\mathfrak{p}_F$. A nice property of this vector, is that under the restricted injectivity condition, it has a closed form expression.

LEMMA 5.5 Let $x \in \mathbb{R}^n$ and suppose that (INJ_{T_x}) is verified. Then $\mathfrak{p}_F$ is well-defined and

$$\mathfrak{p}_F = \Phi_{T_x}^{+,*} e_x.$$

PROOF The vector $\mathfrak{p}_F$ is in fact the projection of 0 to the set $\{\mathfrak{p} \mid (\Phi^*\mathfrak{p})_{T_x} = e_x\}$. In particular,

$$\Phi_{T_x}^* \mathfrak{p}_F = e_x$$

Using hypothesis (INJ_{T_x}), we multiply both sides by $\Phi_{T_x}^{+,*}$ to get the result. ∎

In fact, one can show that $\mathfrak{p}_F$ or $\mathfrak{p}_0$ being non-degenerate certificates are equivalent in some sense.

PROPOSITION 5.6 Under the hypothesis (INJ_T), one has

$$\Phi^*\mathfrak{p}_F \in \text{ri}(\partial J(x_0)) \implies \mathfrak{p}_0 = \mathfrak{p}_F, \tag{5.8}$$

$$\Phi^*\mathfrak{p}_0 \in \text{ri}(\partial J(x_0)) \implies \mathfrak{p}_0 = \mathfrak{p}_F, \tag{5.9}$$

These conditions implies that x_0 is the unique solution of $(\mathcal{P}_0)$.

PROOF Owing to Corollary 5.3, this shows that the left hand side conditions of both (5.8) and (5.9) implies that x_0 is a solution of $\mathcal{P}_0$.

Proof of (5.8) Under the condition $\text{Ker}(\Phi) \cap T = \{0\}$, one has, from the definition of $\Phi_T^{*,+}$, that

$$\mathfrak{p}_F = \underset{\mathfrak{p}}{\text{argmin}} \ \{||\mathfrak{p}|| \mid \ \Phi_T^*\mathfrak{p} = e\} \tag{5.10}$$

Using Proposition 5.3 for $w = 0$ with x_0 being solution of $(\mathcal{P}_0)$, one sees that the constraint of problem (5.10) includes the constraint of the Definition 5.2. Indeed, one has

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

If $\eta_F \in \text{ri}(\partial J(x_0))$, then it is a feasible point in the definition of $\mathfrak{p}_{0,w}$ when $w = 0$. Hence, necessarily $\mathfrak{p}_0 = \mathfrak{p}_F$.

Proof of (5.9) Since x_0 is a solution of $(\mathcal{P}_0)$, according to Proposition 5.3, one has that

$$\mathfrak{p}_0 = \underset{\mathfrak{p}}{\text{argmin}} \ \left\{||\mathfrak{p}||^2 \mid \ \Phi_T^*\mathfrak{p} = e, \Phi_S^*\mathfrak{p} \in \mathcal{U}\right\}$$

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

$$\mathfrak{p}_0 + \Phi_T q + u = 0 \quad \text{where} \quad \begin{cases} \Phi_T^*\mathfrak{p}_0 = e, \\ u \in \mathcal{N}_{\mathcal{U}}(\mathfrak{p}_0). \end{cases}$$

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

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

which leads to $p_0 = (\Phi_T)^{+,*} e = p_F$. ∎

Beside condition (INJ_{T_x}) stated above, the following Identifiability Criterion will play a pivotal role.

> Definition 5.5 For $x \in \mathbb{R}^N$ such that (INJ_{T_x}) holds, we define the *Identifiability Criterion* at x as
>
> $$\mathbf{IC}(x) = J_{f_x}^{x,\circ}(\Phi_{S_x}^* \Phi_{T_x}^{+,*} e_x - P_{S_x} f_x).$$

The fact that $\mathbf{IC}(x) < 1$ is totaly equivalent to $\Phi^* p_F \in \text{ri}\,\partial J(x)$ but stated in analytical form. Note that if J is a strong gauge on T, then it becomes $\mathbf{IC}(x) = J_{f_x}^{x,\circ}(\Phi_{S_x}^* \Phi_{T_x}^{+,*} e_x)$. The Identifiability Criterion clearly brings into play the promoted subspace T_{x_0} and the interaction between the restriction of Φ to T_{x_0} and S_{x_0}. It is a generalization of the irrepresentable condition that has been studied in the literature for some popular regularizers, including the ℓ^1-norm (Fuchs 2004), analysis-ℓ^1 (Vaiter, Peyré, et al. 2013), ℓ^1-ℓ^2 (Bach 2008a) and nuclear (Bach 2008b).

It turns out that in such a setting, $\mathbf{IC}(x_0) < 1$ is a sufficient condition for identifiability without any any other particular assumption on the finite-valued function J, such as partial smoothness. By identifiability, we mean the fact that x_0 is the unique solution of $(\mathcal{P}_{y,0})$.

> Proposition 5.7 Let $x_0 \in \mathbb{R}^N$ and $T = T_{x_0}$. We assume that ($\text{INJ}_{T_{x_0}}$) holds and $\mathbf{IC}(x_0) < 1$. Then x_0 is the unique solution of $(\mathcal{P}_{y,0})$.

PROOF This is a straightforward consequence of the first order condition using $\mathfrak{p}_0$ as a dual certificate. Denote $e = e_{x_0}$, $f = f_{x_0}$ and $S = T^\perp$. Taking the dual vector $\mathfrak{p} = \Phi_T^{+,*} e$, we have on the one hand

$$\Phi_T^* \Phi_T^{+,*} e = e$$

since $e \in \mathrm{Im}(\Phi_T^*)$. On the other hand,

$$J_{f_x}^{x,\circ}(\Phi_S^* \Phi_T^{+,*} e - P_S f) = \mathbf{IC}(x_0) < 1.$$

We conclude thanks to Theorem 5.3. ∎

5.4.2 Analysis Precertificate

In the special case of analysis ℓ^1 regularization, Nam et al. (2013) introduced a different precertificate. We extend their idea to any function of the form $J = J_0 \circ D^*$

> DEFINITION 5.6 — ANALYSIS PRECERTIFICATE Let $x \in \mathbb{R}^n$ and $T = T_{D^*x}$. The *analysis precertificate* $\mathfrak{p}_A$ reads
>
> $$\mathfrak{p}_A = D \operatorname*{argmin}_{\omega \in \mathbb{R}^p} \|\omega\| \quad \text{subject to} \quad D\omega \in \mathrm{Im}\,\Phi^* \quad \text{and} \quad \omega_T = e_{D^*x}.$$

We can give an explicit form of this certificate using a basis of $\mathrm{Ker}\,\Phi$.

> PROPOSITION 5.8 Let N^* be a basis of $\mathrm{Ker}\,\Phi$. Then,
>
> $$\mathfrak{p}_A = -D(ND_S)^+ ND e_{D^*x}.$$

PROOF Since N^* be a basis of $\mathrm{Ker}\,\Phi$, one has $\mathrm{Im}\,\Phi^* = \mathrm{Ker}\,N$. Thus,

$$D\omega \in \mathrm{Im}\,\Phi^* \Leftrightarrow ND\omega = 0.$$

Hence,

$$D\omega \in \operatorname{Im} \Phi^* \quad \text{and} \quad \omega_T = e_{D^*x} \Leftrightarrow ND(e_{D^*x} + P_S \omega) = 0.$$

The least-square solution to this linear equation, which coincides with its minimal ℓ^2 norm solution ω, yields

$$\mathfrak{p}_A = -D(ND_S)^+ NDe_{D^*x},$$

which concludes the proof. ∎

We explore numerically the difference between this precertificate and the linearized in Section 10.2. We draw the attention of the reader to the fact that in (Nam et al. 2013), the authors look after

$$\exists \omega \text{ s.t. } D\omega \in \operatorname{Im} \Phi^* \quad \text{and} \quad \omega \in \partial J_0(D^*x),$$

and its non-degenerate version, which implies the source condition.

Part II

Robustness

6

Noise ℓ^2 Robustness

Main contributions of this chapter

- Theorem 6.1 shows that if both the non-degenerate source condition and the restricted injectivity hold, then $(\mathcal{P}_{y,\lambda})$ enjoys a linear convergence rate with respect to the estimation error.

Contents

This chapter is dedicated to seek sufficient conditions under which for any solution of $(\mathcal{P}_{y,\lambda})$, one has

$$||x^\star - x_0|| = O(||w||).$$

This condition depends on x_0, while λ must be chosen proportionaly to the noise level $||w||$. The terminology "linear" in the convergence rate, which stems from the inverse problems community, pertains to the fact that the error is within a factor of the noise level.

In Section 6.1, we establish the rate of convergence of any solution $x^\star$ to x_0 with respect to the Bregman divergence, introduced in Definition 2.23. Section 6.2 states our main result, namely the fact that any if both the source condition and the restricted injectivity hold, then $(\mathcal{P}_{y,\lambda})$ enjoys a linear convergence rate with respect to the estimation error. Finally, in Section 6.3, we draw connections with previous works on this subject.

6.1 Bregman Rate

The following Lemma 6.1 gives the prediction error and Bregman distance rates for $(\mathcal{P}_{y,\lambda})$. Such results can be found in (Scherzer 2009).

Lemma 6.1 Suppose that (SC_{x_0}) is satisfied with $\eta = \Phi^* p \in \partial J(x_0)$. Then, for any minimizer $x^\star$ of $(\mathcal{P}_{y,\lambda})$, and with $\lambda = c\varepsilon$ for some $c > 0$ and $\varepsilon = ||w||$, we have

$$\begin{aligned} D_\eta(x^\star, x_0) &\leqslant \varepsilon \frac{(1 + c\|p\|_2/2)^2}{c} && \text{Bregman divergence rate,} \\ \|\Phi x^\star - \Phi x_0\|_2 &\leqslant \varepsilon(2 + c\|p\|_2) && \text{Prediction error.} \end{aligned}$$

PROOF The proof follows the same lines as (Scherzer 2009). For the sake of completness, we provide here a proof.

By definition of $x^\star$, one has

$$\|\Phi x^\star - y\|^2 + \lambda J(x^\star) \leqslant \|\Phi x_0 - y\|^2 + \lambda J(x_0).$$

Since $\Phi x_0 - y = -w$, one has

$$\|\Phi x^\star - y\|^2 + \lambda J(x^\star) \leqslant \varepsilon^2 + \lambda J(x^\star). \tag{6.1}$$

Now, by definition of $D_\eta(x^\star, x_0)$, one has

$$\begin{aligned} D_\eta(x^\star, x_0) &= J(x^\star) - J(x_0) - \langle \Phi^* p,\, x^\star - x_0 \rangle \\ &= J(x^\star) - J(x_0) - \langle p,\, \Phi(x^\star - x_0) \rangle. \end{aligned} \tag{6.2}$$

Using the Cauchy-Schwarz inequality, we have that

$$\langle p,\, \Phi(x^\star - x_0) \rangle \leqslant \|p\|\, \|\Phi(x^\star - x_0)\|.$$

By the fact that $\Phi(x^\star - x_0) = \Phi x^\star - y + w$ and the triangle inequality, one has

$$\langle p,\, \Phi(x^\star - x_0) \rangle \leqslant \|p\| \left(\|\Phi x^\star - y\| + \|w\| \right).$$

Injecting this in (6.2), we get

$$D_\eta(x^\star, x_0) \leqslant J(x^\star) - J(x_0) + \|p\| \|\Phi x^\star - y\| + \|p\|\varepsilon.$$

Starting from (6.1), we have

$$\varepsilon^2 \geqslant \|\Phi x^\star - y\|^2 + \lambda \left(D_\eta(x^\star, x_0) - \|p\| \|\Phi x^\star - y\| - \|p\|\varepsilon \right). \tag{6.3}$$

Using that $a^2 + b^2 \geqslant 2ab$,

$$\lambda||p||||\Phi x^\star - y|| \leqslant ||\Phi x^\star - y||^2 + \frac{\lambda^2}{4}||p||^2.$$

Thus, we get

$$\varepsilon^2 \geqslant \lambda D_\eta(x^\star, x_0) - \lambda\varepsilon||p|| - \frac{1}{4}\lambda^2||p||^2.$$

Finally, we have

$$\begin{aligned}\lambda D_\eta(x^\star, x_0) &\leqslant \varepsilon^2 + \varepsilon||p||^2\lambda^2 + \frac{1}{4}\lambda^2||p||^2 \\ &= \left(\varepsilon + \frac{\lambda||p||}{2}\right)^2,\end{aligned}$$

which proves the first inequality (Bregman divergence rate).

Now, for the second one, we start from

$$||\Phi x^\star - \Phi x_0|| \leqslant ||\Phi x^\star - y|| + ||y - \Phi x_0|| = ||\Phi x^\star - y|| + \varepsilon.$$

Using bound (6.3) and the fact that $D_\eta(x^\star, x_0)$ is nonnegative, we have that

$$||\Phi x^\star - y|| \leqslant \varepsilon + \lambda||p||.$$

Hence,

$$||\Phi x^\star - \Phi x_0|| \leqslant 2\varepsilon + \lambda||p||,$$

which concludes our proof. ∎

6.2 Linear Convergence Rate

6.2.1 Main Result

We are now ready to state our main convergence results. We denote $x_0 \in \mathbb{R}^n$ and denote $T_0 = T_{x_0}$, $f_0 = f_{x_0}$.

Theorem 6.1 Assume that ($\overline{\text{SC}}_{x_0}$) holds at x_0 with $\eta = \Phi^* \mathfrak{p} \in \text{ri}\, \partial J(x_0)$, and that ($\text{INJ}_{T_0}$) holds. Choosing $\lambda = c\varepsilon$, $c > 0$, where $\varepsilon = ||w||$, the following holds for any minimizer $x^\star$ of ($\mathcal{P}_{y,\lambda}$)

$$||x^\star - x_0||_2 \leqslant C\varepsilon \ ,$$

where

$$C = C_1 \left(2 + c||\mathfrak{p}||_2\right) + C_2 \frac{(1 + c||\mathfrak{p}||_2/2)^2}{c\left(1 - J_{f_0}^{x_0,\circ}(\eta_{S_0} - P_{S_0} f_0)\right)} \tag{6.4}$$

and $C_1 > 0$ and $C_2 > 0$ are constants independent of $\mathfrak{p}$ and η.

This result holds for any finite-valued convex function and holds for any minimizer of ($\mathcal{P}_{y,\lambda}$) (not necessarily unique). However, remark that (INJ_{T_0}) makes sense only if J promotes subspaces of low dimension. Note that finding a certificate $\mathfrak{p}$ is not trivial, and that the constant involved in Theorem 1 depends on it. This leaves a degree of freedom to optimize the constant for the certificate. The closer to 1 the constant $C_\mathfrak{p} = 1 - J_{f_0}^{x_0,\circ}(\eta_{S_0} - P_{S_0} f_0)$ is, the better is the robustness. It measures how far from the relative boundary of the subdifferential of J at x_0 is $\mathfrak{p}$. Finally, the constants C_1 and C_2 are not absolute and may depend on the dimension. Hence, this theorem does not extend straightforwardly to the infinite-dimensional problem ($c\mathcal{P}_{y,\lambda}$).

The constants read as follows.

$$C_1 = C_\Phi^{-1} \quad \text{and} \quad C_2 = \frac{||\Phi||_{2,2} + C_\Phi}{C_J C_\Phi},$$

where C_Φ is the coercivity constant associated to the hypothesis (INJ_{T_0}), i.e.

$$\exists\, C_\Phi > 0 \quad \text{s.t.} \quad ||\Phi x||_2 \geqslant C_\Phi ||x||_2, \quad \forall x \in T_0 \ ,$$

and C_J is defined by the coercivity of $J_{f_0}^{x_0}$ on S_0:

$$\exists\, C_J > 0 \quad \text{s.t.} \quad \forall x \in \mathbb{R}^n,\ J_{f_0}^{x_0}(\eta) \geqslant C_J ||\eta||_2.$$

When the decomposable norm is also separable (see Corollary 5.1), the stability result of Theorem 6.1 remains true assuming that $J_f^\circ(\eta_V - P_V f) < 1$ for $V \subset S_0$. This however comes at the price of the stronger restricted injectivity assumption ($\text{INJ}_{V^\perp}$). To show this, the only thing to modify is the statement and the proof of Lemma 6.2 which can be done easily using similar arguments to those in the proof of Corollary 5.1.

6.2.2 Proof of Theorem 6.1

Let T_0 and e_0 be the subspace and generalized sign vector associated to x_0, and denote $S_0 = {T_0}^\perp$. We choose some $f_0 \in \text{ri}\,\partial J(x_0)$. Now as $J_{f_0}^{x_0}$ is coercive and bounded on S_0 (see Lemma 3.8), we get

$$\exists\, C_J > 0 \quad \text{s.t.} \quad \forall x \in \mathbb{R}^n,\ J_{f_0}^{x_0}(\eta) \geqslant C_J \|\eta\|_2.$$

We obtain the following bound on the projected distance between $x^\star$ and x_0.

LEMMA 6.2 Suppose that ($\overline{\text{SC}}_{x_0}$) holds at x_0 with $\eta \in \text{ri}\,\partial J(x_0)$. Then,

$$\|P_{S_0}(x^\star - x_0)\|_2 \leqslant \frac{D_\eta(x^\star, x_0)}{C_J\left(1 - J_{f_0}^{x_0,\circ}(\eta_{S_0} - P_{S_0} f_0)\right)} .$$

PROOF From the properties of J_{f_0} (see Lemma 3.8), there exists $v \in S_0$ such that

$$J_{f_0}^{x_0,\circ}(v) \leqslant 1 \quad \text{and} \quad J_{f_0}^{x_0}(x^\star - x_0) = J_{f_0}^{x_0}(P_{S_0}(x^\star - x_0)) = \langle P_{S_0}(x^\star - x_0), v\rangle .$$

Moreover, $\nu + P_{S_0} f_0 + e_0 \in \partial J(x_0)$. Thus

$$\begin{aligned}
D_\eta(x^\star, x_0) &\geqslant D_\eta(x^\star, x_0) - D_{\nu + P_{S_0} f_0 + e_0}(x^\star, x_0) \\
&= \langle \nu + P_{S_0} f_0 + e_0 - \eta, \, x^\star - x_0 \rangle \\
&= \langle \nu - (\eta_{S_0} - P_{S_0} f_0), \, x^\star - x_0 \rangle \\
&= J_{f_0}^{x_0}(P_{S_0}(x^\star - x_0)) - \langle \eta_{S_0} - P_{S_0} f_0, \, P_{S_0}(x^\star - x_0) \rangle \\
&\geqslant J_{f_0}^{x_0}(P_{S_0}(x^\star - x_0)) \left(1 - J_{f_0}^{x_0,\circ}(\eta_{S_0} - P_{S_0} f_0)\right) \\
&\geqslant C_J \| P_{S_0}(x^\star - x_0) \|_2 \left(1 - J_{f_0}^{x_0,\circ}(\eta_{S_0} - P_{S_0} f_0)\right) \ ,
\end{aligned}$$

where in the last two inequalities, we used the duality inequality on $\mathrm{dom}\, J_{f_0}^{x_0,\circ} \times \mathrm{dom}\, J_{f_0}^{x_0}$ with $\mathrm{dom}\, J_{f_0}^{x_0} = \mathbb{R}^N$ and $\mathrm{dom}\, J_{f_0}^{x_0,\circ} = S_0$. ∎

We now give the proof of Theorem 6.1.

PROOF

$$\begin{aligned}
\|x^\star - x_0\|_2 &\leqslant \| P_{T_0}(x^\star - x_0) \|_2 + \| P_{S_0}(x^\star - x_0) \|_2 \\
&\leqslant {C_\Phi}^{-1} \| \Phi \, P_{T_0}(x^\star - x_0) \|_2 + \| P_{S_0}(x^\star - x_0) \|_2 \\
&\leqslant {C_\Phi}^{-1} \| \Phi (x^\star - x_0) \|_2 + (1 + {C_\Phi}^{-1} \|\Phi\|_{2,2}) \| P_{S_0}(x^\star - x_0) \|_2 \ ,
\end{aligned}$$

where we used assumption (INJ_{T_0}), *i.e.*,

$$\exists \ C_\Phi > 0 \quad \text{s.t.} \quad \|\Phi x\|_2 \geqslant C_\Phi \|x\|_2, \quad \forall x \in T_0 \ .$$

We finally apply Lemma 6.2 to get

$$\|x^\star - x_0\|_2 \leqslant {C_\Phi}^{-1} \|\Phi(x^\star - x_0)\|_2 + \frac{\|\Phi\|_{2,2} + C_\Phi}{C_J C_\Phi \left(1 - J_{f_0}^{x_0,\circ}(\eta_{S_0} - P_{S_0} f_0)\right)} D_\eta(x^\star, x_0) \ .$$

Using Lemma 6.1 yields the assertion. ∎

6.3 Relation to Previous Works

Convergence rates. The monograph (Scherzer 2009) is dedicated to regularization properties of inverse problems in infinite-dimensional Hilbert and Banach spaces with application to imaging. In particular, Chapter 3 of this book treats the case where J is a coercive gauge for the problem $(\mathcal{P}_{y,\lambda})$. In (Burger et al. 2004), the authors consider the case where $\mathcal{J}$ is a proper, convex and l.s.c functional for both the constrained and Lagrangian regularization $(c\mathcal{P}_{y,\lambda})$. Under the source condition and a restricted injectivity assumption, they bound the error in Bregman divergence with a linear rate $O(||w||)$. For the classical Thikonov regularization, i.e. $\mathcal{J} = || \cdot ||_{L^2(\Omega)}$, the estimation is in $O(\sqrt{||w||})$, which is not a linear convergence. Extensions of these results have been proved in (Resmerita 2005) and (Hofmann et al. 2007) for the Bregman rate.

Lorenz (2008) treats the case where J is a ℓ^p norm with $1 \leqslant p \leqslant 2$ and provides a prediction error $\Phi x_0 - \Phi x^\star$ in $O(||w||)$ and an estimation error $x^\star - x_0$ in $O(\sqrt{||w||})$. Grasmair et al. (2011) is concerned with the special case of ℓ^1 regularization, and draws some connection with the restricted isometry property (RIP), see below. The results that are the closest to our are contained in (Grasmair 2011). Here, $\mathcal{J}$ is a proper, convex, l.s.c and positively homogeneous functional on some Banach space $\mathbb{H}$. Under a source condition and restricted injectivity on a an appropriate cone, a linear convergence rate is proved with respect to $\mathcal{J}$, i.e.

$$J(x^\star - x_0) = O(||w||).$$

This result implies ours, but only if J is injective which precludes many important regularizers, e.g. TV.

Compressed sensing. In a compressed sensing setting, for instance when Φ is drawn from a i.i.d. normal distribution, it was proved (Rudelson et al. 2008) that if the number of measurements q is such that $q \succcurlyeq k \log(n/k)$ where $k = ||x_0||_0$ then there exists with high probability on Φ a non-degenerate certificate when $J = || \cdot ||_1$, i.e. $(\overline{SC}_x)$ holds and one can apply the result of Theorem 6.1.

The performance of compressed sensing recovery has initially been analyzed using the so-called restricted isometry property (RIP) introduced in (Candès et al. 2006a, 2006b; Candès and Tao 2006) for ℓ^1. It is defined for a couple

(Φ, k) where k is a targeted sparsity, as the smallest constant δ_k such that

$$(1-\delta_k)||x||^2 \leqslant ||\Phi x||^2 \leqslant (1+\delta_k)||x||^2, \tag{6.5}$$

for any vector x such that $||x||_0 \leqslant k$. It is shown (Candès et al. 2006a) that if $\delta_{2k} + \delta_{3k} < 1$, then for *every* vector x_0 of sparsity k, there exists a non-degenerate certificate (Candès et al. 2005, Lemma 2.2) as remarked also by Grasmair et al. (2011). This result thus implies linear convergence rate, and is applied in (Candès et al. 2006b) to show the robustness to noise of compressed sensing. This was generalized to analysis sparsity (i.e. $J = ||D^* \cdot||_1$ with D tight frame) in (Candès, Eldar, et al. 2011), structured sparsity in (Candès, Eldar, et al. 2011) and matrix completion (Recht et al. 2010; Candès and Plan 2011b) using $J = || \cdot ||_*$. A major shortcoming of this approach is that available designs of matrices satisfying (6.5) for reasonnably large value of k are essentially random. Indeed, in this case, the constant δ_k can be shown to be small enough with high probability on Φ for nearly optimal scaling of (n, q, k). For instance, when Φ is drawn for the Gaussian ensemble, it is the case when $q \gtrsim k \log(n/k)$. as proved by Candès and Tao (2006) Note that in general, computing the RIP constants for a given matrix is an NP-hard problem (Bandeira et al. 2013).

The golfing scheme introduced by Gross (2011) for the nuclear norm allows to consider structured non-Gaussian measurements, e.g. partial Fourier measurements. It is based on an iterative scheme starting from the linearized precertificate $\mathfrak{p}_F$ in order to construct an (approximate) certificate with high probability on the matrix for a given vector. It was further studied by Candès and Plan (2011a) for ℓ^1 regularization.

7
Model Selection

Main contributions of this chapter

- Theorem 7.2 ensures that for a partly smooth function J, if the restricted injectivity holds and that the linearized precertificate is a non-degenerate certificate, then for a certain regime of small noise, $(\mathcal{P}_{y,\lambda})$ has a unique solution which belongs in the model manifold of x_0.
- Theorem 7.3 proves a similar result for partly smooth functions with linear manifold with explicit constants.

Contents

So far, we were concerned with ℓ^2-stability/robustness. What can be said about the recovery of the model T_{x_0} underlying the original vector itself x_0 ? To be able to state such a result, the regularization has to enjoy some additional regularity assumption. This is the goal of partial smoothness that we introduced in Definition 4.1.

Section 7.1 states our main result. It ensures that for a partly smooth function J, if the restricted injectivity holds and that the linearized precertificate $\mathfrak{p}_F$ is a non-degenerate certificate, then in a small noise regime, $(\mathcal{P}_{y,\lambda})$ has a unique solution and it belongs to the same model manifold $\mathcal{M}$ as x_0. In Section 7.2, we specialize this result to partly smooth functions with linear manifold. This specialization does not cover for instance the nuclear norm regularization, but provides more explicit constants involved in the robustness. Finally, we draw connections with previous works in Section 7.3.

7.1 Selection Against Small Noise: General Case

7.1.1 Sensitivity of the Lagrangian Problem

Before diving into our main result, we first show of the theory of partly smooth functions introduced in (Lewis 2002) can be directly applied to study the sensitivity of $(\mathcal{P}_\theta)$ for $\lambda > 0$.

Theorem 7.1 Let x_θ be a solution of $(\mathcal{P}_\theta)$, with $\lambda > 0$, and suppose that J is locally partly smooth at x_θ relative to $\mathcal{M}$. If

$$\mathrm{Ker}(\Phi) \cap T_{x_\theta} = \{0\} \quad \text{and} \quad \eta_\theta \in \mathrm{ri}(\partial J(x_\theta)) \tag{7.1}$$

where we have denoted

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

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

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

PROOF It suffices to apply Theorem 5.7 of (Lewis 2002). Indeed, the function f

$$f(x,\theta) = J(x) + \frac{1}{2\lambda}||\Phi x - y||^2$$

is partly smooth at (x_θ, θ) relative to the manifold $\mathcal{M} \times \Theta$, where $\Theta = \mathbb{R}^q \times \mathbb{R}_+$ and condition (7.1) is exactly equivalent to x_θ being a strong minimizer of $f(\cdot,\theta)$, see (Lewis 2002, Definition 5.6). ■

Condition (7.1) is not very useful because it depends on the solution x_θ and not on the data to recover x_0. The rationale behind Theorem 7.2 is to make θ tends to 0, and under the conditions

$$\lambda \to 0 \quad \text{and} \quad \frac{||w||}{\lambda} \to 0,$$

Propositions 5.2 and 5.4 ensure that

$$x_\theta \to x_0 \quad \text{and} \quad p_\theta \to p_0.$$

7.1.2 Main Result

We now state our main result, which performs a sensivity analysis at $\lambda = 0$.

THEOREM 7.2 Let $J \in \mathcal{S}_{x_0}(\mathcal{M})$ a locally partly smooth function at x_0 relative to $\mathcal{M}$ such that (INJ_T) with $T = T_{x_0}$ holds and $\mathfrak{p}_F$ is a non-degenerate certificate, i.e.

$$\text{Ker}(\Phi) \cap T = \{0\}, \quad \text{and} \quad \Phi^* \mathfrak{p}_F \in \text{ri}(\partial J(x_0)). \tag{7.2}$$

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

$$x_\theta \in \mathcal{M} \quad \text{and} \quad ||x_\theta - x_0|| = O(||w||). \tag{7.3}$$

The heuristic underlying the hypotheses of Theorem 7.2 is that the conditions in (7.1) converge toward those of (7.2). Indeed, according to Proposition 5.6, (7.2) implies $\mathfrak{p}_0 = \mathfrak{p}_F$. This is precisely what we need to show in order to prove Theorem 7.2.

Obviously, the assumptions of Theorem 7.2 imply the conclusion of Theorem 6.1. Contrary to the latter, the former is based on an explicit formulation of the precertificate $\mathfrak{p}_F$. Note that there exist vectors which can be stably recovered in the ℓ^2 sense of Theorem 6.1, but whose underlying manifold model cannot be stably identified in the sense of Theorem 7.2, see our numerical experiments in Chapter 10.

The following proposition shows that Theorem 7.2 is in some sense sharp, since the hypothesis $\Phi^* \mathfrak{p}_F \in \text{ri}(\partial J(x_0))$ (almost) characterizes the stability of $\mathcal{M}$.

PROPOSITION 7.1 We suppose that x_0 is the unique solution of $(\mathcal{P}_{y,0})$ and that

$$\text{Ker}(\Phi) \cap T = \{0\}, \quad \text{and} \quad \Phi^* \mathfrak{p}_F \notin \text{ri}(\partial J(x_0))$$

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

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

The only case not covered by neither Theorem 7.2 nor Proposition 7.1 is when $\Phi^* p_F \in \text{rbound}(\partial J(x_0))$ (the relative boundary). In this case, one cannot conclude, since depending on the noise w, one can have either stability or non-stability of $\mathcal{M}$. We refer to Chapter 10 where an example illustrates this situation for the 1-D total variation $J = ||\nabla \cdot ||_1$ (here ∇ is a discretization of the 1-D derivative operator).

7.1.3 Proof of Theorem 7.2

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

$$\tilde{x}_\theta \in \underset{x \in \mathcal{M}}{\text{Argmin}}\ f(x, \theta). \tag{7.4}$$

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

The proof of Proposition 5.2 carries over verbatim to this constrained problem, which shows that

$$\tilde{x}_\theta \to x_0 \quad \text{when} \quad \begin{cases} \lambda \longrightarrow 0, \\ ||w||^2/\lambda \longrightarrow 0. \end{cases} \tag{7.5}$$

In the following, to lighten the notations, we denote $T_{\tilde{x}_\theta} = \hat{T}$.

Convergence of the tangent model subspaces. By definition of the constrained problem (7.4), $\tilde{x}_\theta \in \mathcal{M}$. Moreover, since $f(\cdot, \theta)$ is partly smooth at x_0 relative to $\mathcal{M}$, the sharpness property Definition 4.1(ii) holds at all nearby points in the manifold $\mathcal{M}$, see (Lewis 2002, Proposition 2.10). Thus, as soon as $(||w||^2/\lambda, \lambda)$ is small enough, we have that $\mathcal{M}$ is a C^2-manifold around $\tilde{x}_\theta$

and $\hat{T} = \mathcal{T}_{\mathcal{M}}(\tilde{x}_\theta)$. Using the fact that $\mathcal{M}$ is of class C^2, we get the following convergence

$$\hat{T} = \mathcal{T}_{\mathcal{M}}(\tilde{x}_\theta) \longrightarrow \mathcal{T}_{\mathcal{M}}(x_0) = T \quad \text{when} \quad \begin{cases} \lambda \longrightarrow 0, \\ ||w||^2/\lambda \longrightarrow 0, \end{cases} \tag{7.6}$$

where the convergence should be understood over the Grasmanian of linear spaces with the same dimension (or equivalently, as the convergence of the projection operators $P_{\hat{T}} \to P_T$), see Section 2.2. Since $\text{Ker}(\Phi) \cap T = \{0\}$, (7.6) implies that for $(||w||^2/\lambda, \lambda)$ small enough,

$$\text{Ker}(\Phi) \cap \hat{T} = \{0\}, \tag{7.7}$$

which we also assume now.

First order conditions. By partial smoothness, the restriction of J to $\mathcal{M}$ is smooth at $\tilde{x}_\theta$ for θ small enough. Hence, since $x \mapsto \frac{1}{2\lambda}||y - \Phi x||^2$ is smooth everywhere, the smooth perturbation rule (Lewis 2002, Corollary 4.7) implies that $f(\cdot, \theta)$ is also partly smooth at $\tilde{x}_\theta$ for $\mathcal{M}$, and thus its restriction to $\mathcal{M}$ is smooth at $\tilde{x}_\theta$. Therefore, Lewis (2002, Proposition 2.4(b)) applies, and it follows that $\tilde{x}_\theta$ is a critical point of (7.5) if, and only if,

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

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

$$\Phi^*_{\hat{T}}(\Phi\tilde{x}_\theta - y) + \lambda e_{\tilde{x}_\theta} = 0, \tag{7.8}$$

Convergence of primal variables. Since $\tilde{x}_\theta$ and x_0 belongs to the same active manifold, and $\mathcal{M}$ is a manifold of class C^2 around them, using Lemma 2.3, each point in their neighbourhoods has unique projection on $\mathcal{M}$. In particular, $\tilde{x}_\theta = P_{\mathcal{M}}(\tilde{x}_\theta)$ and $x_0 = P_{\mathcal{M}}(x_0)$. Moreover, $P_{\mathcal{M}}$ is of class C^1 near $\tilde{x}_\theta$. Thus, continuous differentiability shows

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

where $DP_{\mathcal{M}}(\tilde{x}_\theta)$ is the derivative of $P_{\mathcal{M}}$ at $\tilde{x}_\theta$. Combining (Lewis et al. 2008, Lemma 2.3) and (Lewis 2002, Proposition 2.4(i)), the derivative $DP_{\mathcal{M}}(\tilde{x}_\theta)$ is given by

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

Inserting this in (7.8), we get

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

Using (7.7), $\Phi_{\hat{T}}$ has full rank, and thus

$$\tilde{x}_\theta - x_0 = \Phi_{\hat{T}}^+ w - \lambda(\Phi_{\hat{T}}^* \Phi_{\hat{T}})^{-1} e_{\tilde{x}_\theta} + o\left(||\tilde{x}_\theta - x_0||\right) \ . \tag{7.9}$$

Altogether, we obtain the bound

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

Convergence of dual variables. We define

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

Arguing as above, and using (7.9) we have

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

We thus arrive at

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

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

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

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

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

This implies

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

Altogether, we get the bound

$$\|\tilde{\eta}_\theta - \eta_F\| = O(\|w\|/\lambda, \lambda). \tag{7.10}$$

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

$$\tilde{p}_\theta \in \mathrm{ri}(\partial J(\tilde{x}_\theta)). \tag{7.11}$$

We follow the line of proof of (Lewis 2002).

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

$$\tilde{p}_n \in \mathrm{rbound}(\partial J(\tilde{x}_n)) \tag{7.12}$$

where we used the shorthand notations

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

According to (7.10) and (7.5),

$$(\tilde{x}_n, \tilde{p}_n) \to (x_0, p_F). \tag{7.13}$$

Condition (7.12) is equivalently stated as, for each n

$$\exists z_n \in T_{\tilde{x}_n}^{\perp}, \quad \forall \mathfrak{p} \in \partial J(\tilde{x}_n), \quad \langle z_n,\, \mathfrak{p} - \tilde{\mathfrak{p}}_n \rangle \geqslant 0, \tag{7.14}$$

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

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

$$\forall \nu \in \partial J(x_0), \quad \langle z^\star,\, \mathfrak{p} - \mathfrak{p}_F \rangle \geqslant 0. \tag{7.15}$$

Indeed, let us consider any $\nu \in \partial J(x_0)$. By condition 3 of Definition 4.1, ∂J is continuous along $\mathcal{M}$, so that since $\tilde{x}_n \to x_0$ there exists $\nu_n \in \partial J(\tilde{x}_n)$ with $\nu_n \to \nu$. Applying (7.14) with $\mathfrak{p} = \nu_n$ gives

$$\langle z_n,\, \nu_n - \tilde{\mathfrak{p}}_n \rangle \geqslant 0.$$

Taking the limit $n \to +\infty$ in this inequality leads to (7.15), which contradicts the fact that $\mathfrak{p}_F \in \mathrm{ri}(\partial J(x_0))$.

Conditions (7.11) and (7.7) implies, using Theorem 5.3, that $\tilde{x}_\theta = x_\theta$ is the unique solution of $(\mathcal{P}_\theta)$.

7.1.4 Proof of Proposition 7.1

Let x_θ be a solution of $(\mathcal{P}_{y,\lambda})$. Suppose that $x_\theta \in \mathcal{M}$. In particular, x_θ is a solution of the non-convex minimization (7.4). Arguing as in the proof of Theorem 7.2, we get the bound (7.10), i.e.

$$\|\eta_\theta - \eta_F\| = O(\|w\|/\lambda, \lambda) \quad \text{where} \quad \eta_\theta = \Phi^* \frac{y - \Phi x_\theta}{\lambda}. \tag{7.16}$$

Since x_0 is the unique solution of $(\mathcal{P}_{y,0})$, $\mathfrak{p}_0$ is well defined, hence $\eta_0 = \Phi^* \mathfrak{p}_0 \in \partial J(x)$. Thus, there exists $K > 0$ (for instance $K = d(\eta_F, \partial J(x))$) such that $\|\eta_F -$

$\eta_0\| > K$. Moreover,

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

According to (7.16) and (5.4), one has

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

This leads to a contradiction since by assumption $\eta_F \notin \partial J(x_0)$, hence $x_\theta \notin \mathcal{M}$.

7.2 Selection of Linear Manifold

When J is partly smooth with linear manifold ($\mathcal{M} = T_x$), see Definition 4.2, i.e. the manifold is in fact the model subspace, we derive a more precise result with explicit constants.

7.2.1 Main Result

THEOREM 7.3 Let $x_0 \in \mathbb{R}^n$ and $T = T_{x_0}$. We suppose that J is a partly smooth function with linear manifold at x_0 with the corresponding parameters $(\Gamma, \nu_{x_0}, \mu_{x_0}, \tau_{x_0}, \xi_{x_0})$ where the constants are defined in (4.2), (4.3) and (4.4). Assume that (INJ_T) holds and $\mathbf{IC}(x_0) < 1$. Then there exist positive constants (A_T, B_T) that solely depend on T and a constant $C(x_0)$ such that if w and λ obey

$$\frac{A_T}{1 - \mathbf{IC}(x_0)} \|w\| \leqslant \lambda \leqslant \nu_{x_0} \min\left(B_T, C(x_0)\right) \tag{7.17}$$

the solution $x^\star$ of ($\mathcal{P}_{y,\lambda}$) with noisy measurements y is unique, and satisfies $T_{x^\star} = T$. Furthermore, one has

$$\|x_0 - x^\star\| = O\Big(\max(\|w\|, \lambda)\Big).$$

Clearly this result asserts that exact recovery of T_{x_0} from noisy partial measurements is possible with the proviso that the regularization parameter λ lies in the interval (7.17). The value λ should be large enough to reject noise, but small enough to recover the entire subspace T_{x_0}. In order for the constraint (7.17) to be non-empty, the noise-to-signal level $||w||/\nu_{x_0}$ should be small enough, i.e.

$$\frac{||w||}{\nu_{x_0}} \leqslant \frac{1-\mathbf{IC}(x_0)}{A_T} \min\left(B_T, C(x_0)\right) \ .$$

The constant $C(x_0)$ involved in this bound depends on x_0 and has the form

$$C(x_0) = \frac{1-\mathbf{IC}(x_0)}{\xi_{x_0}\nu_{x_0}} H\left(\frac{D_T\,\mu_{x_0} + \tau_{x_0}}{\xi_{x_0}}\right)$$
$$\text{where} \quad H(\beta) = \frac{\beta + 1/2}{E_T\,\beta}\,\varphi\left(\frac{2\beta}{(\beta+1)^2}\right) \quad \text{and} \quad \varphi(u) = \sqrt{1+u} - 1 \ .$$

The constants (D_T, E_T) only depend on T. $C(x_0)$ captures the influence of the parameters $\pi_{x_0} = (\mu_{x_0}, \tau_{x_0}, \xi_{x_0})$, where the latter reflect the geometry of the regularizing function J at x_0. More precisely, the larger $C(x_0)$, the more tolerant the recovery is to noise. Thus favorable regularizers are those where $C(x_0)$ is large, or equivalently where π_{x_0} has small entries, since H is a strictly decreasing function.

7.2.2 Proof of Theorem 7.3

The proof is similar to Theorem 7.2. To lighten the notations, we let $\varepsilon = ||w||$, $\nu = \nu_{x_0}, \mu = \mu_{x_0}, \tau = \tau_{x_0}, \xi = \xi_{x_0}, f = f_{x_0}$ and $T = T_{x_0}$.

The strategy is to construct a vector which is the unique solution to

$$\min_{x \in T} \frac{1}{2}||y - \Phi x||^2 + \lambda J(x) \ , \qquad (\mathcal{P}_\theta^T)$$

and then to show that it is actually the unique solution to $(\mathcal{P}_\theta)$ under the assumptions of Theorem 7.3.

The following lemma gives a convenient implicit equation satisfied by the unique solution to $(\mathcal{P}_\theta^T)$.

LEMMA 7.1 Assume that (INJ_T) holds. Then ($\mathcal{P}_\theta^T$) has exactly one minimizer $\hat{x}$, and the latter satisfies

$$\hat{x} = x_0 + \Phi_T^+ w - \lambda(\Phi_T^* \Phi_T)^{-1} \tilde{e} \quad \text{where} \quad \tilde{e} \in P_T(\partial J(\hat{x})). \tag{7.18}$$

PROOF Assumption (INJ_T) implies that the objective in ($\mathcal{P}_\theta^T$) is strongly convex on the feasible set T, whence uniqueness follows immediately. By a change of variable, ($\mathcal{P}_\theta^T$) be also rewritten in the unconstrained form

$$\hat{x} = \underset{x \in \mathbb{R}^N}{\operatorname{argmin}} \frac{1}{2}||y - \Phi_T x||^2 + \lambda J(P_T\, x) \ .$$

Thus, using Proposition 3.9(i), $\hat{x}$ has to satisfy

$$\Phi_T^*(y - \Phi_T \hat{x}) + \lambda \tilde{e} = 0,$$

for any $\tilde{e} \in P_T(\partial J(\hat{x}))$. Owing to the invertibility of Φ on T, i.e. (INJ_T), we obtain (7.18). ∎

We are now in position to prove Theorem 7.3. This is be achieved in three steps:

Step 1: We show that in fact $T_{\hat{x}} = T$.

Step 2: Then, we prove that $\hat{x}$ is the unique solution of ($\mathcal{P}_{y,\lambda}$) using Theorem 5.3.

Step 3: We finally exhibit an appropriate regime on λ and ε for the above two statements to hold.

Step 1: Subspace equality. By construction of $\hat{x}$ in ($\mathcal{P}_\theta^T$), it is clear that $\hat{x} \in T$. The key argument now is to use that J is PRG at x_0, and to show that

$$\Gamma(x_0 - \hat{x}) \leqslant \nu, \tag{7.19}$$

which in turn will imply subspace equality, i.e. $T_{\hat{x}} = T$ (see Definition 4.2).

We have from (7.18) and subadditivity that

$$\begin{aligned}\Gamma(x_0 - \hat{x}) &\leqslant \Gamma(-\Phi_T^+ w) + \lambda\Gamma((\Phi_T^*\Phi_T)^{-1}\tilde{e}) \\ &\leqslant \left|\left|\left|(\Phi_T^*\Phi_T)^{-1}\right|\right|\right|_{\Gamma\to\Gamma}\{\Gamma(-\Phi_T^* w) + \lambda\Gamma(\tilde{e})\} \\ &\leqslant \left|\left|\left|(\Phi_T^*\Phi_T)^{-1}\right|\right|\right|_{\Gamma\to\Gamma}\{\|\Phi_T^*\|_{\ell^2\to\Gamma}\varepsilon + \alpha_0\lambda\}. \end{aligned} \tag{7.20}$$

where $\alpha_0 = \Gamma(\tilde{e})$. Consequently, to show that (7.19) is verified, it is sufficient to prove that

$$A\varepsilon + B\lambda \leqslant \nu, \tag{C_1}$$

where we set the positive constants

$$\begin{aligned} A &= \left|\left|\left|(\Phi_T^*\Phi_T)^{-1}\right|\right|\right|_{\Gamma\to\Gamma}\|\Phi_T^*\|_{\ell^2\to\Gamma}, \\ B &= \alpha_0\left|\left|\left|(\Phi_T^*\Phi_T)^{-1}\right|\right|\right|_{\Gamma\to\Gamma}. \end{aligned}$$

Suppose for now that (C_1) holds and consequently, $T_{\hat{x}} = T$. Then decomposability of J on T (Theorem 3.1) implies that

$$\hat{e} = P_{T_{\hat{x}}}(\partial J(\hat{x})) = P_T(\partial J(\hat{x})) = \tilde{e},$$

where we have denoted $\hat{e} = e_{\hat{x}}$. Thus (7.18) yields the following implicit equation

$$\hat{x} = x_0 + \Phi_T^+ w - \lambda(\Phi_T^*\Phi_T)^{-1}\hat{e}. \tag{7.21}$$

Step 2: $\hat{x}$ is the unique solution of ($\mathcal{P}_{y,\lambda}$). Recall that under condition (C_1), J is decomposable at $\hat{x}$ and x_0 with the same model subspace T. To deduce that $\hat{x}$ is the unique solution of ($\mathcal{P}_{y,\lambda}$), it remains to show that

$$J^\circ_{\hat{f}}(\lambda^{-1}\Phi_S^*(y - \Phi\hat{x}) - \hat{f}_S) < 1. \tag{7.22}$$

where we use the shorthand notations $\hat{f} = f_{\hat{x}}$ and $\hat{f}_S = P_S\,\hat{f}$.

Under condition (C_1), the ξ-stability property (4.4) of J at x_0 yields

$$J_f^\circ(\lambda^{-1}\Phi_S^*(y-\Phi\hat{x})-\hat{f}_S) \leqslant \big(1+\xi\Gamma(x_0-\hat{x})\big)J_f^\circ(\lambda^{-1}\Phi_S^*(y-\Phi\hat{x})-\hat{f}_S). \tag{7.23}$$

Furthermore, from (7.21), we can derive

$$\lambda^{-1}\Phi_S^*(y-\Phi\hat{x})-\hat{f}_S = \Phi_S^*\Phi_T^{+,*}\hat{e}+\lambda^{-1}\Phi_S^*Q_T w-\hat{f}_S, \tag{7.24}$$

where $Q_T = \mathrm{Id}-\Phi_T\Phi_T^+ = P_{\mathrm{Ker}(\Phi_T^*)}$. Inserting(7.24) in (7.23), we obtain

$$J_f^\circ(\lambda^{-1}\Phi_S^*(y-\Phi\hat{x})-\hat{f}_S) \leqslant \big(1+\xi\Gamma(x_0-\hat{x})\big)J_f^\circ(\Phi_S^*\Phi_T^{+,*}\hat{e}+\lambda^{-1}\Phi_S^*Q_T w-\hat{f}_S).$$

Moreover, subadditivity yields

$$\begin{aligned} J_f^\circ(\Phi_S^*\Phi_T^{+,*}\hat{e}+\lambda^{-1}\Phi_S^*Q_T w-\hat{f}_S) &\leqslant J_f^\circ(\Phi_S^*\Phi_T^{+,*}e-f_S)+J_f^\circ(\Phi_S^*\Phi_T^{+,*}(\hat{e}-e)) \\ &\quad + J_f^\circ(P_S(f-\hat{f}))+J_f^\circ(\lambda^{-1}\Phi_S^*Q_T w). \end{aligned} \tag{7.25}$$

We now bound each term of (7.25). In the first term, one recognizes

$$J_f^\circ(\Phi_S^*\Phi_T^{+,*}e-f_S) \leqslant \mathbf{IC}(x_0). \tag{7.26}$$

Appealing to the μ-stability property, we get

$$\begin{aligned} J_f^\circ(\Phi_S^*\Phi_T^{+,*}(\hat{e}-e)) &\leqslant \left|\!\left|\!\left| -\Phi_S^*\Phi_T^{+,*} \right|\!\right|\!\right|_{\Gamma\to J_f^\circ}\Gamma(e-\hat{e}) \\ &\leqslant \mu \left|\!\left|\!\left| -\Phi_S^*\Phi_T^{+,*} \right|\!\right|\!\right|_{\Gamma\to J_f^\circ}\Gamma(x_0-\hat{x}). \end{aligned} \tag{7.27}$$

From τ-stability, we have

$$J_f^\circ(f_S-\hat{f}_S) \leqslant \tau\Gamma(x_0-\hat{x}). \tag{7.28}$$

Finally, we use a simple operator bound to get

$$J_f^\circ(\lambda^{-1}\Phi_S^*Q_T w) \leqslant \frac{1}{\lambda}\|\Phi_S^*Q_T\|_{\ell^2\to J_f^\circ}\varepsilon. \tag{7.29}$$

Following the same steps as for the bound (7.20), except using $\tilde{e}=\hat{e}$ here, gives

$$\Gamma\big(x_0-\hat{x}\big) \leqslant \left|\!\left|\!\left| (\Phi_T^*\Phi_T)^{-1} \right|\!\right|\!\right|_{\Gamma\to\Gamma}\{\|\Phi_T^*\|_{\ell^2\to\Gamma}\varepsilon+\lambda\Gamma(\hat{e})\}. \tag{7.30}$$

Plugging inequalities (7.26)-(7.30) into (7.23) we get the upper-bound

$$
\begin{aligned}
&J^{\circ}_{\hat{f}}(\Phi_S^* \Phi_T^{+,*}\hat{e} + \lambda^{-1}\Phi_S^* Q_T w - \hat{f}_S) \\
&\leqslant (1+\xi\Gamma(x_0-\hat{x}))\Big(\mathbf{IC}(x_0) + \Gamma(x_0-\hat{x})\left(\mu|||-\Phi_S^*\Phi_T^{+,*}|||_{\Gamma\to J^\circ_f} + \tau\right) \\
&\quad + \frac{1}{\lambda}|||\Phi_S^* Q_T|||_{\ell^2\to J^\circ_f}\varepsilon\Big) \\
&\leqslant (1+\xi(c_1\varepsilon+\lambda c_2))\Big(\mathbf{IC}(x_0) + (c_1\varepsilon+\lambda c_2)\bar{\mu} + \frac{1}{\lambda}c_4\varepsilon\Big) < 1,
\end{aligned}
$$

where we have introduced

$$\bar{\mu} = \mu c_3 + \tau \quad \text{and} \quad \alpha_1 = \Gamma(\hat{e}) = \Gamma(\tilde{e}) = \alpha_0$$

and

$$
\begin{aligned}
c_1 &= A, & c_2 &= \alpha_1|||(\Phi_T^*\Phi_T)^{-1}|||_{\Gamma\to\Gamma}, \\
c_3 &= |||-\Phi_S^*\Phi_T^{+,*}|||_{\Gamma\to J^\circ_f}, & c_4 &= |||\Phi_S^* Q_T|||_{\ell^2\to J^\circ_f}.
\end{aligned}
$$

If is then sufficient that

$$(1+\xi(c_1\varepsilon+\lambda c_2))\Big(\mathbf{IC}(x_0) + (c_1\varepsilon+\lambda c_2)\bar{\mu} + \frac{1}{\lambda}c_4\varepsilon\Big) < 1. \tag{7.31}$$

In particular, if

$$C\varepsilon \leqslant \lambda$$

holds for some constant $C > 0$ to be fixed later, then inequality (7.31) is true if

$$P(\lambda) = a\lambda^2 + b\lambda + c > 0 \tag{7.32}$$

where

$$
\begin{cases}
a = & -\xi\bar{\mu}\,(c_1/C + c_2)^2 \\
b = & -(c_1/C + c_2)\,(\xi\mathbf{IC}(x_0) + \xi c_4/C + \bar{\mu}) \\
c = & 1 - \mathbf{IC}(x_0) - c_4/C
\end{cases}
$$

Let us set the value of C to

$$C = \frac{2c_4}{1-\mathbf{IC}(x_0)},$$

which, for $0 \leqslant \mathbf{IC}(x_0) < 1$, it ensures that $c = \frac{1-\mathbf{IC}(x_0)}{2}$ is bounded and positive, and thus, the polynomial P has a negative and a positive root $\lambda_{\max}$ equal to

$$\lambda_{\max} = \frac{b}{2a}\varphi\left(-4\frac{ac}{b^2}\right)$$

where

$$\begin{cases} a = & -\xi\bar{\mu}((1-\mathbf{IC}(x_0))c_1/(2c_4)+c_2)^2 \\ b = & -((1-\mathbf{IC}(x_0))c_1/(2c_4)+c_2)\,(\bar{\mu}+(1+\mathbf{IC}(x_0))\xi/2) \\ c = & (1-\mathbf{IC}(x_0))/2. \end{cases}$$

Hence,

$$\begin{aligned}\lambda_{\max} &= \frac{\bar{\mu}+(1+\mathbf{IC}(x_0))\xi/2}{\xi\bar{\mu}((1-\mathbf{IC}(x_0))c_1/c_4+2c_2)}\varphi\left(\frac{2\xi(1-\mathbf{IC}(x_0))\bar{\mu}}{(\bar{\mu}+(1+\mathbf{IC}(x_0))\xi/2)^2}\right) \\ &\geqslant \frac{1-\mathbf{IC}(x_0)}{\xi}H(\bar{\mu}/\xi),\end{aligned}$$

where

$$\varphi(\beta) = \sqrt{1+\beta}-1, \quad \text{and} \quad H(\beta) = \frac{\beta+1/2}{\beta(c_1/c_4+2c_2)}\varphi\left(\frac{2\beta}{(\beta+1)^2}\right).$$

Consequently, we can conclude that the bounds

$$\frac{2c_4}{1-\mathbf{IC}(x_0)}\varepsilon \leqslant \lambda \leqslant \frac{1-\mathbf{IC}(x_0)}{\xi}H(\bar{\mu}/\xi) \tag{C$_2$}$$

imply condition (7.31), which in turn yields (7.22).

Step 3: (C_1) and (C_2) are in agreement. It remains now that show the compatibility of (C_1) and (C_2), i.e. to provide appropriate regimes of λ and ε such that both conditions hold simultaneously. We first observe that (C_1) and the left-hand-side of (C_2) both hold for λ fulfilling

$$\lambda \leqslant C_0\nu \quad \text{where} \quad C_0 = \left(\frac{A}{2c_4}+B\right)^{-1} \leqslant \left(\frac{1-\mathbf{IC}(x_0)}{2c_4}A+B\right)^{-1}.$$

This updates (C_2) to the following ultimate range on λ

$$\frac{2c_4}{1-\mathbf{IC}(x_0)}\varepsilon \leqslant \lambda \leqslant \min\left(C_0\nu, \frac{1-\mathbf{IC}(x_0)}{\xi}H(\bar{\mu}/\xi)\right) .$$

Now in order to have an admissible non-empty range for λ, the noise level ε must be upper-bounded as

$$\varepsilon \leqslant \frac{1-\mathbf{IC}(x_0)}{2c_4}\min\left(C_0\nu, \frac{1-\mathbf{IC}(x_0)}{\xi}H(\bar{\mu}/\xi)\right).$$

Finally, the constants provided in the statement of the theorem (and subsequent discussion) are as follows

$$A_T = 2c_4,\ B_T = C_0,\ D_T = c_3, \text{and } E_T = c_1/c_4 + 2c_2 ,$$

which completes the proof.

7.3 Relation to Previous Works

Special cases. Theorems 7.2 and 7.3 are generalizations of many previous works that have appeared in the literature. For the ℓ^1 norm, $J = \|\cdot\|_1$, to the best of our knowledge, this result was initially stated by Fuchs (2004). In this setting, the result $T_x^\star = T_{x_0}$ corresponds to the correct identification of the support, i.e. $\operatorname{supp}(x^\star) = \operatorname{supp}(x_0)$. Moving to a setting where both Φ and w are random, the condition $\mathfrak{p}_F \in \operatorname{ri}\partial J(x_0)$ implies model consistency (also known as sparsistency for ℓ^1), i.e. the probability that the support is correctly identified tends to 1 as the number of measurements grows large. Bach proves respectively in (Bach 2008a) and (Bach 2008b) Theorem 7.2 (in fact a variant since he considers randomized Φ and w) for $\ell^1 - \ell^2$ and nuclear norm gauges, in the special case where Φ has full rank (i.e. is injective). Our result thus shows that the same condition ensures rank consistency with the additional constraint that $\operatorname{Ker}(\Phi) \cap T = \{0\}$. Theorem 7.3 for a ℓ^1 analysis prior was proved by Vaiter, Peyré, et al. (2013). Theorem 7.2 is extended in (Duval et al. 2013) to the TV norm that endows the infinite dimensional Banach space of Radon measures, and where Φ has a finite-dimensional range. In this setting, they show that $\mathfrak{p}_F$ must be replaced by a different pre-certificate.

Compressed sensing. Condition $\Phi^* \mathfrak{p}_F \in \operatorname{ri} \partial J(x_0)$ is often used when Φ is drawn from the Gaussian matrix ensemble to asses the performance of compressed sensing recovery with ℓ^1 norm, see (Wainwright 2009; Dossal et al. 2012). It has been proved (Wainwright 2009; Dossal et al. 2012) for $J = \|\cdot\|_1$ that if Φ is a random matrix drawn from the Gaussian ensemble, then for $q > 2k \log n$, $\Phi^* \mathfrak{p}_F \in \operatorname{ri} \partial J(x)$ with high probability on Φ for $k = \|x_0\|_0$. One may have observed that the bound on q bears similarities with that of Chapter 6.3 except in the scaling in the log term, but induces stronger conclusion. It is also used to ensure ℓ^2 robustness of matrix completion in a noisy setting by Candès et al. (2010), and our findings show that it also ensures rank consistency for matrix completion at high signal to noise levels. It generalizes a result proved for a family of decomposable norms (including in particular ℓ^1-ℓ^2 norm and the nuclear norm) by Candès and Recht (2013) when $w = 0$.

Stronger criteria for ℓ^1. Many sufficient conditions can be formulated to ensure that $\mathfrak{p}_F$ is a non-degenerate certificate, and hence to guarantee the model stability. The strongest criterion to ensure a noise robustness for ℓ^1 regularization is the coherence, introduced by Donoho et al. (2001). Finer criteria based on Babel functions have been proposed in (Gribonval and Nielsen 2008; Borup et al. 2008). The Exact Recovery Condition introduced by Tropp (2006) is weaker than the coherence which in turns is greater that the weak-ERC (Dossal 2012). More precisely, the coherence of a matrix with unit-norm is defined as

$$\mu = \max_{i \neq j} |\langle \Phi_i, \Phi_j \rangle|,$$

and the associated coherence criterion reads

$$\operatorname{coh}(x_0) = \frac{\|x_0\|_0 \mu}{1 - (\|x_0\|_0 - 1)\mu}.$$

The Fuchs' criterion reads

$$\mathbf{IC}(x_0) = \|\Phi_{I^c}^* \Phi_I^{+,*} \operatorname{sign}(x_0)_I\|_\infty = \|\Phi^* \mathfrak{p}_F\|_\infty,$$

where $I = \text{supp}(x_0)$. The Exact Recovery Condition reads

$$\text{ERC}(x_0) = \|\Phi_{I^c}^* \Phi_I^{+,*}\|_{\infty,\infty}.$$

The weak-ERC reads

$$\text{wERC}(x_0) = \frac{\max\limits_{j \notin I} \sum_{i \in I} |\langle \Phi_i, \Phi_J \rangle|}{1 - \max\limits_{j \in I} \sum_{i \neq j \in I} |\langle \Phi_i, \Phi_J \rangle|}.$$

These quantity obey the following inequality:

$$\mathbf{IC}(x_0) \leqslant \text{ERC}(x_0) \leqslant \text{wERC}(x_0) \leqslant \text{coh}(x_0).$$

In particular, if any of these quantity is less than 1, then $\mathfrak{p}_F$ is a non-degenerate certificate.

Part III

Sensitivity

8

Local Differentiability of the Optimal Solutions

Main contributions of this chapter

- Theorem 8.1 constructs a smooth map of solutions to $(\mathcal{P}^{F}_{y,\lambda})$ on an open neighborhood of some solution $x^\star$, and computes its derivative.
- Theorem 8.2 shows that the prediction map is well-defined outside the transition space and gives its derivative.

Contents

THIS chapter is concerned with differentiability of an optimal map of solution to $(\mathcal{P}^F_{y,\lambda})$. Moreover, we prove that the prediction map is well-defined outside the transition space and gives its derivative. The core of our proof strategy relies on the identification of a certain linear subspace $T = T_{x^\star(y)}$ associated to a particular minimizer $x^\star(y)$ of $(\mathcal{P}^F_y)$. We exhibit explicitly a certain set of observations, denoted $\mathcal{H}$ (see Definition 8.2), outside which the initial non-smooth optimization $(\mathcal{P}^F_y)$ 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 2002) for the sensitivity analysis of partly smooth functions. The robustness analysis of Chapter 7 relies on the manifold stability when $\lambda = 0$. In contrast, we provide in this chapter a sensitivity analysis when $\lambda > 0$. Even if we state our result only w.r.t to small variations of y, our result can be extended to analyze the sensibility with respect to other variable parameterizing F, which could be useful for homotopy-like results. This sensivity analysis is central to construct an unbiased estimator of the quadratic risk. We suppose here that J is a partly smooth gauge with linear manifold, i.e. such that $\mathcal{M}_x = T_x$ and J is 1-homogeneous. We conjecture that this statement remains true for any convex partly smooth function, though this has not been formally proved yet. The technical obstacles faced by this generalization will be discussed in Chapter 9.

8.1 Main Assumptions

This section details our assumptions on both the data fidelity term F and the regularizer J. We also introduce the restriction and second order derivative of the regularizer J.

8.1.1 Assumptions on the Regularizer

We assume in this chapter that J is a **partly smooth gauge with linear manifold**, see Definitions 4.1 and 4.2. More precisely, we need the following assumption.

$$\forall T \in \mathcal{T}, \quad J \in C^2(\widetilde{T}) \; . \tag{C_{sm}}$$

$$\forall x \in \mathbb{R}^n, \exists \nu > 0, \forall x' \in T_x, \|x' - x\| < \tau \Rightarrow T_x = T_{x'} \; . \tag{C_{reg}}$$

$$\text{The set } \mathcal{T} \text{ is finite.} \tag{$C_{\mathcal{T}}$}$$

$$J \text{ is positively homogeneous.} \tag{C_{hom}}$$

We recall that the set $\mathcal{T}$ is defined as

$$\mathcal{T} = \{T_x \mid x \in \mathbb{R}^n\}.$$

Some remarks are in order. Assumption (C_{reg}) amounts to saying that there exists a neighbourhood of x on T_x on which this subspace model is constant. This condition is a part of the assumptions defining the class of partly smooth function with linear manifold introduced in Definition 4.1. 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.

8.1.2 Assumption on the Data Fidelity

In all the following, we consider a variational regularized problem of the form of $(\mathcal{P}^F_{y,\lambda})$, i.e. of the form[1]

$$x^\star(y) \in \underset{x\in\mathbb{R}^n}{\operatorname{Argmin}}\ F(x,y) + J(x). \tag{$\mathcal{P}^F_y$}$$

The fidelity term F is of the following form

$$F(x,y) = F_0(\Phi x, y) \tag{8.1}$$

where $F_0(\cdot, y)$ is a general loss function assumed to be a proper, convex and sufficiently smooth function of its first argument $\forall y$. We assume that the fidelity term enjoys the following properties.

$$\forall\,(y,x) \in \mathbb{R}^q \times \mathbb{R}^n, \quad F(\cdot, y) \in C^2(\mathbb{R}^n) \quad \text{and} \quad \nabla_1 F(x,\cdot) \in C^1(\mathbb{R}^q). \tag{C_F}$$

Generalized linear models in the exponential family falls into the class of losses we consider. Indeed, taking the negative log-likelihood corresponding to (9.2) gives[2]

$$F_0(\mu, y) = \frac{1}{\lambda}\sum_{i=1}^{q} \varphi_i\,(\mu_i) - \langle y,\, \mu\rangle\ . \tag{8.2}$$

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 $F_0(\mu, y) = \frac{1}{2\lambda}\|y-\mu\|^2$ for the standard linear models (Gaussian case), and the logistic loss $F_0(\mu, y) = \sum_{i=1}^q \log\left(1 + \exp(\mu_i)\right) - \langle y,\, \mu\rangle$ for logistic regression (Bernoulli case). GLM estimators with losses (8.2) 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; Meier et al. 2008; Bach 2010; Kakade et al. 2010); see also (Bühlmann et al. 2011, Chapter 3, 4 and 6).

1. Note that here the parameter λ is absorbed within the fidelity term F.
2. Strictly speaking, the minimization may have to be over a convex subset of $\mathbb{R}^n$.

8.1.3 Restriction and Second-Order Derivative of the Regularizer

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

$$D_1^2 g_T(x,y) = P_T \circ D_1^2 g(x,y) \circ P_T$$

which can be understood as the Hessian of the mapping $x \in T \mapsto g(x,y)$, 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 $D_{12}^2 g(x,y)$ the Jacobian of the mapping $y \in \mathbb{R}^q \mapsto \nabla_1 g(x,y)$ with respect to y, and $\nabla_1 g(x,y)$ is the gradient of g w.r.t the first variable at (x,y).

We denote

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

the restriction of J to T for some subspace $T \subset \mathcal{T}$. Hence the hessian of J_T is well-defined on $\mathcal{T}$. Observe that $\nabla J_T(x) = e_x$ for $x \in T$. We illustrate this definition on several examples.

Lasso and general Lasso. For $J = \|\cdot\|_1$, one has

$$\forall x_T \in T, \quad \nabla J_T(x_T) = \operatorname{sign}(x_T) ,$$

and thus, $D^2 J_T(x_T) = 0$. This is also the case for the analysis ℓ^1-penalty (general Lasso), see for instance (Vaiter, Deledalle, et al. 2013). This property basically reflects the fact that these regularizers are polyhedral, hence piecewise affine.

Group Lasso. For $J = \|\cdot\|_{1,2}$ as defined in (1.13), we have

$$D^2 J_T(x_T) = \delta_x \circ P_{x^\perp},$$

where, for $I = \text{supp}_{\mathcal{B}}(x)$,

$$\begin{aligned} \delta_x &: v \in \mathbb{R}^{|I|} \mapsto (v_b/\|x_b\|)_{b \subset I} \in \mathbb{R}^{|I|} \\ \text{and} \quad P_{x^\perp} &: v \in \mathbb{R}^{|I|} \mapsto (P_{x_b^\perp} v_b)_{b \subset I} \in \mathbb{R}^{|I|} \ , \end{aligned}$$

where

$$P_{x_b^\perp} v_b = v_b - \frac{\langle x_b, v_b \rangle}{\|x_b\|^2} x_b$$

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

8.2 Local Behavior of a Solution Mapping

8.2.1 Restricted Injectivity

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

Definition 8.1 — Restricted Injectivity A vector $x \in \mathbb{R}^n$ with $T = T_x$ is said to satisfy the *restricted injectivity condition* if, and only if,

$$T \cap \text{Ker}(D_1^2 F_T(x, y)) \cap \text{Ker}(D^2 J_T(x)) = \{0\}. \qquad (\mathcal{C}_{x,y})$$

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

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

Under this condition, the derivative of the objective function is invertible on T.

Lemma 8.1 Let $x \in \mathbb{R}^n$, and $T = T_x$. Assume that ($\mathcal{C}_{x,y}$) holds. Then the linear operator $D_1^2 F_T(x, y) + D^2 J_T(x) : T \to T$ is invertible on T.

Proof Since $F(\cdot, y)$ and J are convex and $C^2(T)$ by assumptions (C_F) and (C_{sm}), the (restricted) hessians $D_1^2 F_T(x, y)$ and $D^2 J_T(x)$ 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 $\mathcal{C}_{x,y}$ states. ■

8.2.2 Transition Space

Let us now turn to the sensitivity of a minimizer $x^\star(y)$ of ($\mathcal{P}_y^F$) 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[3] $y \mapsto x^\star(y)$. 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 $x^\star(y)$.

Definition 8.2 The *transition space* $\mathcal{H}$ is defined as

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

where we have denoted

$$\Pi_{q+n,q} : \begin{cases} \mathbb{R}^q \times \widetilde{T} & \longrightarrow \quad \mathbb{R}^q \\ (y, x_T) & \longmapsto \quad y \end{cases}$$

3. To be understood here as a set-valued mapping.

the canonical projection on the first q coordinates, $\operatorname{bd} C$ is the boundary of the set C, and

$$\mathcal{A}_T = \left\{ (y, x_T) \in \mathbb{R}^q \times \widetilde{T} \mid -\nabla_1 F(x_T, y) \in \operatorname{rbd} \partial J(x_T) \right\}.$$

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

In the particular case where F is the square loss, $J = \|\cdot\|_1$ (synthesis sparsity) and $J(x) = \|D^* x\|_1$ (analysis sparsity), the same transition set is introduced in (Dossal et al. 2013) and (Vaiter, Deledalle, et al. 2013). In these specific cases, since J is a polyhedral gauge, $\mathcal{H}$ is a union of affine hyperplane. The geometry of this set can be significantly more complex for other gauges. For instance, for $J = \|\cdot\|_{1,2}$, it can be shown to be a semi-algebraic set (union of algebraic hyper-surfaces).

8.2.3 Main Result

We are now equipped to state our main sensitivity analysis result.

Theorem 8.1 Let $y \notin \mathcal{H}$, and $x^\star$ a solution of $\mathcal{P}_\lambda(y)$ such that $(\mathcal{C}_{x^\star,y})$ holds. Then, there exists an open neighborhood $\mathcal{V} \subset \mathbb{R}^q$ of y, and a mapping $\tilde{x} : \mathcal{V} \to T$ such that

(i) For all $\bar{y} \in \mathcal{V}$, $\tilde{x}(\bar{y})$ is a solution of $(\mathcal{P}_{\bar{\lambda}}(\bar{y}))$, and $\tilde{x}(y) = x^\star$.

(ii) the mapping $\tilde{x}$ is $C^1(\mathcal{V})$ and for every $\bar{y} \in \mathcal{V}$,

$$\partial_1 \tilde{x}(\bar{y}) = -(D_1^2 F_T(x^\star, \bar{y}) + D^2 J_T(x^\star))^{-1} \circ P_T \circ D_{12}^2 F(x^\star, \bar{y}), \qquad (8.3)$$

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

One now may wonder whether condition $(\mathcal{C}_{x^\star,y})$ is restrictive, and in particular, whether there exists always a solution $x^\star$ such that it holds. In Section 8.3, we give an affirmative answer with the proviso that the loss F_0 is strictly convex.

The above result can be extended to the case where the data fidelity is of the form $F(x, \theta)$ for some parameter θ, with no particular role of y here. One may think for instance to consider $\theta = (y, \lambda)$. The variations with respect to λ are important for developing homotopy-like algorithm. Vaiter, Deledalle, et al. (2013) proved that if J is polyhedral, then the path is locally affine.

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

We define the following mapping

$$\Gamma : (x_T, y) \in T \times \mathbb{R}^q \mapsto \nabla_1 F(x_T, y)_T + e_{x_T}.$$

Observe that owing to Proposition 3.1(iv), the first equation of the first-order condition is equivalent to $\Gamma({x^\star}_T, y) = 0$.

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

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

It comes from the fact that $\Gamma(x_T, y) = 0$ 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 $\bar{y} \mapsto \tilde{x}(\bar{y}) \in T$ and an open neighborhood $\mathcal{V}$ of y such that every element $\bar{y}$ of $\mathcal{V}$ satisfies $\Gamma(\tilde{x}(\bar{y})_T, \bar{y}) = 0$ and $\tilde{x}(\bar{y})_S = 0$. Then, we prove that $\tilde{x}(\bar{y})$ is a solution of $(\mathcal{P}_{\bar{y}}^F)$ for $\bar{y} \in \mathcal{V}$. Finally, we obtain (8.3) from the implicit function theorem.

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

$$D_1 \Gamma({x^\star}_T, \bar{y}) = D_1^2 F_T({x^\star}_T, \bar{y})_T + D_1 e_{{x^\star}_T},$$

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

$$D_1 \Gamma({x^\star}_T, \bar{y}) = D_1^2 F_T({x^\star}_T, \bar{y}) + D^2 J_T({x^\star}_T).$$

The linear operator mapping $D_1\Gamma({x^\star}_T, y)$ is invertible on T according to Lemma 8.1. Hence, using the implicit function theorem (Theorem 2.1) restricted to T, there exists a neighborhood $\widetilde{\mathcal{V}}$ of y such that we can define a mapping $\tilde{x}_T : \widetilde{\mathcal{V}} \to T$ which is $C^1(\widetilde{\mathcal{V}})$, and satisfies for $\bar{y} \in \widetilde{\mathcal{V}}$

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

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

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 $-\nabla_1 F(\tilde{x}(\bar{y}), \bar{y}) \in \partial J(\tilde{x}(\bar{y}))$. We distinguish two cases.

(i) Assume that $-\nabla_1 F(x^\star, y) \in \operatorname{ri} \partial J(x^\star)$: we show that for a sufficiently small neighbourhood of y, we also have $-\nabla_1 F(\tilde{x}(\bar{y}), \bar{y}) \in \operatorname{ri} \partial J(\tilde{x}(\bar{y}))$. First, since $\tilde{x}$ is continuous on T, for any $\varepsilon > 0$, there exists a neighborhood $\bar{\mathcal{V}} \subset \widetilde{\mathcal{V}}$ of y such that

$$\|\tilde{x}(\bar{y}) - x^\star\| \leqslant \varepsilon \quad \forall \bar{y} \in \bar{\mathcal{V}} .$$

By virtue of Assumption (C_{reg}), one can then choose ε sufficiently small to conclude that $S_{\tilde{x}(\bar{y})} = S$ for any $\bar{y} \in \bar{\mathcal{V}}$.

Suppose that there is a sequence $(y_\ell)_\ell$ approaching y such that

$$-\nabla_1 F(\tilde{x}(y_\ell), y_\ell) \notin \operatorname{ri} \partial J(\tilde{x}(y_\ell))$$

for all ℓ. This can be equivalently written, owing to Lemma 5.3, as

$$\exists u_\ell \in S_{\tilde{x}(y_\ell)}, \quad \forall v \in \partial J(\tilde{x}(y_\ell)) \quad \langle u_\ell, v + \nabla_1 F(x^\star, y)\rangle \leqslant 0, \forall \ell ,$$

or

$$\exists u_\ell \in S_{\tilde{x}(y_\ell)}, \quad \sup \ \langle u_\ell, \partial J(\tilde{x}(y_\ell)) + \nabla_1 F(\tilde{x}(y_\ell), y_\ell)\rangle \leqslant 0, \forall \ell . \tag{8.4}$$

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 $S_{\tilde{x}(y_\ell)}$ converges to S. We now claim that

$$\sup \langle u, \partial J(x^\star) + \nabla_1 F(x^\star, y)\rangle \leqslant 0 \ .$$

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

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

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

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

we get that

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

The latter inequality holds for any $\eta \in \partial J(x^\star)$, which, in view of Lemma 5.3, means that $-\nabla_1 F(x^\star, y) \notin \operatorname{ri} \partial J(x^\star)$. But this contradicts our initial assumption.

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

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

Applying Lemma 2.2 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. Thus, we also have that $\nabla_1 F(\bar{x}, \bar{y}) = \nabla_1 F(\tilde{x}(\bar{y}), \bar{y})$. This implies in particular that $e(\bar{x}) = e_{\tilde{x}(\bar{y})}$. Since $\widetilde{T} \subset T$ is an open set for the topology relative to T and $\tilde{x}(y) = x^\star \in \widetilde{T}$, for $\bar{y}$ sufficiently close to y, Assumption (C_{reg}) allows to deduce that

$$\tilde{x}(\bar{y}) \in \widetilde{T} \quad \Longrightarrow \quad T_{\tilde{x}(\bar{y})} = T.$$

Thus, we have $T_{\tilde{x}(\bar{y})} = T_{\tilde{x}}$, hence $S_{\tilde{x}(\bar{y})} = S_{\tilde{x}}$. Combining this with Proposition 3.1 and the claim that both vectors have the same image under $e_.$, yields that they also share the same affine hull, i.e. $\bar{S}_{\tilde{x}(\bar{y})} = \bar{S}_{\tilde{x}}$. In turn, this implies the equality of the subdifferential by virtue of Proposition 3.6, i.e. $\partial J(\tilde{x}(\bar{y})) = \partial J(\tilde{x})$. We conclude that

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

Moreover, by definition of the mapping $\tilde{x}_T$, one has for all $\bar{y} \in \mathcal{V} \cap \bar{\mathcal{V}}$

$$\nabla_1 F(\tilde{x}_T(\bar{y}), \bar{y})_T + e_{\tilde{x}_T(\bar{y})} = 0 \ .$$

According to Lemma 3.9, the vector $\tilde{x}(\bar{y})$ is a solution of $\mathcal{P}^F_{\bar{y}}$.

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 $\tilde{x}_T$ as

$$D\tilde{x}_T(\bar{y}) = -\big(D_1\Gamma(\tilde{x}_T(\bar{y}), \bar{y})\big)^{-1}\big(D_2\Gamma(\tilde{x}_T(\bar{y}), \bar{y})\big)$$

where

$$D_2\Gamma(x_T, \bar{y}) = P_T \circ D^2_{12} F(x_T, \bar{y}),$$

which leads us to (8.3). ∎

8.3 Local Behavior of the Prediction Mapping

In this section, we aim to provide a closed-form expression of the local variations of $\widehat{\mu}(y)$ with respect to the observation y. Our result is Theorem 8.2. We assume in this section that F takes the form (8.1) and that

$$\forall\, y \in \mathbb{R}^q, \quad F_0(\cdot, y) \text{ is strictly convex.} \qquad (C_{\text{strict}})$$

8.3.1 Single-Valued Prediction Mapping

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

LEMMA 8.2 We assume that condition (C_{strict}) holds. For $x \in \mathbb{R}^n$, and $T = T_x$, the two following conditions are equivalent.

(i) ($\mathcal{C}_{x^\star,y}$) holds.

(ii) $\text{Ker}(\Phi_T) \cap \text{Ker}(D^2 J_T(x)) = \{0\}$.

Furthermore, if ${x^\star}_0$ and ${x^\star}_1$ are two solutions of $\mathcal{P}(y)$, then $\Phi {x^\star}_0 = \Phi {x^\star}_1$.

PROOF The first part of the lemma come from the following equivalent statements.

$$\begin{aligned} & z \in \text{Ker}(D_1^2 F_T(x,y)) \cap T \\ \iff & \langle z, D_1^2 F_T(x,y) z\rangle = \langle \Phi_T z, D_1^2 F_0(\Phi x, y)\Phi_T z\rangle = 0 \\ \iff & z \in \text{Ker}(\Phi_T) \ . \end{aligned}$$

The second part is contained in Proposition 2.2. ∎

This lemma allows us to define the prediction

$$\widehat{\mu} : \begin{cases} \mathbb{R}^q \to \mathbb{R}^q \\ y \mapsto \Phi x^\star(y) \end{cases}$$

without ambiguity given any solution $x^\star(y)$, which in turn defines a single-valued mapping $\widehat{\mu}$.

8.3.2 Well-Posedness of the Restricted Injectivity Condition

The following lemma proves that ($\mathcal{C}_{x^\star,y}$) is not restrictive, and in particular, there exists always a solution $x^\star$ such that it holds.

LEMMA 8.3 There exists a solution $x^\star$ of $(\mathcal{P}_y^F)$ such that $(\mathcal{C}_{x^\star,y})$ holds.

PROOF Let $x^\star$ a solution of $(\mathcal{P}_y^F)$ such that $(\mathcal{C}_{x^\star,y})$ does not hold. Consider the associated subspace $T = T_{x^\star}$. Thus, for any $h \in \big(\mathrm{Ker}(\Phi) \cap T \cap \mathrm{Ker}(D^2 J_T(x^\star))\big) \setminus \{0\}$, we have $\Phi_T h = 0$ and $D^2 J_T(x^\star)\, h = 0$. Let $v_t = x^\star + th, \forall\, t > 0$. By Proposition 3.1, $v_t \in T$. Moreover, $\Phi_T v_t = \Phi_T x^\star$, and thus $F(\Phi_T v_t, y) = F(\Phi_T x^\star, y)$.

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

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

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

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

where we used Proposition 3.1. From Assumption (C_{sm}), Taylor expansion gives

$$e(v_t) = e(x^\star) + tD^2 J_T(x^\star)h + t\varepsilon(th)||h|| = e(x^\star) + t\varepsilon(th)||h|| \ ,$$

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

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

Suppose now that there exists no $x^\star$ such that $(\mathcal{C}_{x^\star,y})$ holds. Then, we can always find a solution $x^\star$ such that[4] $e(x^\star) \notin \big(\mathrm{Ker}(\Phi) \cap T \cap \mathrm{Ker}(D^2 J_T(x^\star))\big)^\perp$, and therefore there is some $h \in \big(\mathrm{Ker}(\Phi) \cap T \cap \mathrm{Ker}(D^2 J_T(x^\star))\big) \setminus \{0\}$ such that

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

and thus

$$F(\Phi_T v_t, y) + J(v_t) < F(\Phi_T x^\star, y) + J(x^\star) \ ,$$

for t sufficiently small, leading to a contradiction. ∎

4. Recall that $e(x^\star)$ is always different from the origin unless $x^\star = 0$.

8.3.3 Main Result

The following theorem provides a closed-form expression of the local variations of $\widehat{\mu}(y)$ with respect to the observation y.

THEOREM 8.2 We assume that condition (C_{strict}) holds. The mapping $y \mapsto \widehat{\mu}(y)$ is $C^1(\mathbb{R}^n \setminus \mathcal{H})$. For all $y \notin \mathcal{H}$, there exists a solution $x^\star$ of $(\mathcal{P}_y^F)$ such that $(\mathcal{C}_{x^\star,y})$ is satisfied. Moreover, for all $y \notin \mathcal{H}$,

$$D\widehat{\mu}(y) = \Delta(y) \tag{8.5}$$

where

$$\Delta(y) = -\Phi_T \circ (\Phi_T{}^* \circ D_1^2 F_0(\Phi x^\star, y) \circ \Phi_T + D^2 J_T(x^\star))^{-1} \circ \Phi_T{}^* \circ D_{12}^2 F_0(\Phi x^\star, y)$$

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

PROOF We can now prove Theorem 8.2. At any $y \notin \mathcal{H}$, using the previous Lemma 8.3 we consider $x^\star$ a solution of $(\mathcal{P}_y^F)$ such that $(\mathcal{C}_{x^\star,y})$ holds. According to Theorem 8.1, one can construct a mapping $\tilde{x}(\bar{y})$ which coincides with $x^\star$ at y, and is C^1 for $\bar{y}$ in a neighborhood of y. Since $\widehat{\mu}(\bar{y}) = \Phi\tilde{x}(\bar{y})$ on this neighborhood, this shows that $\widehat{\mu}$ is in turn C^1 at y, and its divergence is equal to $\text{tr}(\partial_y \widehat{\mu}(y))$. Note that this shows that this computation is independent of the particular choice of $x^\star$ provided that $(\mathcal{C}_{x^\star,y})$ holds. ■

8.4 Relation to Previous Works

Sensitivity analysis[5] is a major branch of optimization and optimal control theory. Comprehensive monographs on the subject are (Bonnans et al. 2000; Mordukhovich 1992). The focus of sensitivity analysis is the dependence and the regularity properties of the optimal solution set and the optimal values

5. The meaning of sensitivity is different here from what is usually intended in statistical sensitivity and uncertainty analysis.

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.

For the Lasso problem, the above differential formula (8.5) implies that

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

where $x^\star$ is any solution of $(\mathcal{P}_{y,\lambda})$ such that $(\mathcal{C}_{x^\star,y})$ holds, i.e. $\Phi_{\operatorname{supp}(x^\star)}$ has full rank. This result is proved in (Dossal et al. 2013), see also (Tibshirani et al. 2012). The analysis sparsity case was investigated in (Vaiter, Deledalle, et al. 2013) and (Tibshirani et al. 2012). In this case, one has $J = \|D^* \cdot\|_1$ and

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

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

The originality of our contribution in this direction is the following

(i) We formulate the set $\mathcal{H}$ of non-smoothness points, which is crucial for the application to risk estimation exposed in the next chapter.

(ii) We give an explicit formula of the differential of the prediction.

(iii) Our sensitivity result deals with a set-valued mapping (even if its image by Φ is single-valued).

9

Unbiased Risk Estimation

Main contributions of this chapter

- Proposition 9.2 proves that the prediction is Lipschitz continuous with respect to the observation.
- Theorems 9.1 and 9.2 prove that the (G)SURE is an unbiased estimator of the risk for non-linear Gaussian regression and generalized linear model.

Contents

THIS chapter is concerned with unbiased risk estimation for the variational problem $(\mathcal{P}_y^F)$. More precisely, given an estimator family $x^\star{}_\lambda(y)$ of x_0 defined as solution of this problem, we aim to find the best parameter λ. Using tools from o-minimal geometry, we prove that the divergence formula (8.5) is valid Lebesgue-a.e. In turn, this allows us to define an unbiased estimate of the DOF and of the prediction risk (Theorem 9.1 and Theorem 9.2) for model (9.1) under two scenarios: (i) Lipschitz continuous non-linearity h and an additive i.i.d. Gaussian noise; (ii) Generalized Linear Models (GLMs) with a continuous exponential family. Our results encompass some previous ones in the literature as special cases.

9.1 Generalized Linear Models and Degrees of Freedom

9.1.1 Generalized Linear Models

We consider a more general model than a simple linear regression (1.3)

$$\mathbb{E}(Y|\Phi) = h(\Phi x_0), \tag{9.1}$$

where $Y \in \mathbb{R}^q$ is the response vector, $x_0 \in \mathbb{R}^n$ is the unknown vector, $\Phi \in \mathbb{R}^{q \times n}$ is the fixed design matrix whose columns are the n covariate vectors, and the expectation is taken with respect to some σ-finite measure. h is a known smooth function $\mathbb{R}^q \to \mathbb{R}^q$. The goal is to design an estimator of x_0 and to study its properties. In the sequel, we do not make any specific assumption on the number of observations q with respect to the number of predictors n. Recall that when $q < n$, (9.1) is underdetermined, whereas when $q \geqslant n$ and all the columns of Φ are linearly independent, it is overdetermined. Many examples fall within the scope of model (9.1). We here review two of them.

Generalized Linear Models One naturally thinks of generalized linear models (GLMs) introduced by Nelder et al. (1972) which assume that conditionally on Φ, 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 (9.1), the distribution of Y belongs to the n-parameter exponential family and its density reads

$$f(y|\Phi; x_0) = \left(\prod_{i=1}^{n} B_i(y_i)\right) \exp\left(\langle y, \Phi x_0\rangle - \sum_{i=1}^{n} \varphi_i\left((\Phi x_0)_i\right)\right), \quad \Phi x_0 \in \Theta^n , \tag{9.2}$$

where the canonical parameter vector is the linear predictor Φx_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).

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 $\sim \mathcal{P}\left((\Phi x_0)_i\right)$, 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.

9.1.2 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 (8.1) is strictly convex, so that the response (or the prediction) $\widehat{\mu}(y) = \Phi x^\star(y)$ is uniquely defined as a single-valued mapping of y (see Lemma 8.2). That is, it does not depend on a particular choice of solution $x^\star(y)$ of $(\mathcal{P}_y^F)$. More generally, the degrees of freedom could be defined for any estimator of the prediction. Let $\mu_0 = \Phi x_0$.

Suppose that h in (9.1) 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

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

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[1]

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

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

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

where $D\widehat{\mu}$ 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.

9.2 GSURE for Gaussian Observations

9.2.1 Definition

The Stein's lemma is the foundation of risk estimation using the SURE.

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

Lemma 9.1 — Stein's Lemma. Let $Y = \Phi x_0 + W$ with $W \sim \mathcal{N}(0, \sigma^2 \mathrm{Id}_q)$. Assume that $g : y \mapsto g(y)$ is weakly differentiable (and a fortiori a single-valued mapping), then

$$\mathbb{E}_W \langle W, g(Y) \rangle = \sigma^2 \mathbb{E}_W \operatorname{tr} \left[\frac{\partial g(Y)}{\partial Y} \right] .$$

PROOF This result is proved in (Stein 1981). ∎

We here develop an extended version of GSURE, defined by Eldar (2009), that unbiasedly estimates the risk of reconstructing $A\mu_0$ where $A \in \mathbb{R}^{m \times q}$ is an arbitrary matrix. This allows us to cover in a unified framework unbiased estimation of several classical risks including the prediction risk (with $A = \mathrm{Id}$), the projection risk when Φ is rank deficient (with $A = \Phi^*(\Phi\Phi^*)^+$), and the estimation risk when Φ has full rank (with $A = \Phi^+ = (\Phi^*\Phi)^{-1}\Phi^*$). A quantity that will enter into play in the risk of estimating $A\mu_0$ is the degrees of freedom defined as

$$\mathrm{df}^A = \sum_{i=1}^{q} \frac{\mathrm{cov}_W((Ay)_i, (A\widehat{\mu}(y))_i)}{\sigma^2} .$$

Definition 9.1 Let $A \in \mathbb{R}^{m \times q}$. We define the Generalized Stein Unbiased Risk Estimate (GSURE) associated to A as

$$\mathrm{GSURE}^A(y) = \|A(y - \widehat{\mu}(y))\|_2^2 - \sigma^2 \operatorname{tr}(A^*A) + 2\sigma^2 \widehat{\mathrm{df}}^A(y) ,$$

where

$$\widehat{\mathrm{df}}^A(y) = \operatorname{tr}\left(A \frac{\partial \widehat{\mu}(y)}{\partial y} A^* \right) .$$

9.2.2 Unbiasedness

The next result shows that GSURE^A is an unbiased estimator of an appropriate ℓ^2 risk, and $\widehat{\mathrm{df}}^A(y)$ is an unbiased estimator of df^A

THEOREM 9.1 Let $A \in \mathbb{R}^{m \times q}$. Suppose that $y \mapsto \widehat{\mu}(y)$ is weakly differentiable, so that its divergence is well-defined in the weak sense. If $Y = \Phi x_0 + W$ with $W \sim \mathcal{N}(0, \sigma^2 \mathrm{Id}_q)$, then

$$\mathbb{E}_W \mathrm{GSURE}^A(Y) = \mathbb{E}_W \left(\|A\mu_0 - A\widehat{\mu}(Y)\|_2^2 \right) \quad \text{and} \quad \mathbb{E}_W \widehat{\mathrm{df}}^A(Y) = \mathrm{df}^A \ .$$

PROOF Since $y \mapsto \widehat{\mu}(y) = \Phi x^\star(y)$ is weakly differentiable, so is $A^* A \widehat{\mu}(y)$ and we have

$$\frac{\partial A^* A \widehat{\mu}(y)}{\partial y} = A^* A \frac{\partial \widehat{\mu}(y)}{\partial y} \ .$$

Then, using Lemma 9.1, we get

$$\mathbb{E}_W \langle w, A^* A \widehat{\mu}(Y) \rangle = \sigma^2 \mathbb{E}_W \operatorname{tr} \left(A^* A \frac{\partial \widehat{\mu}(Y)}{\partial y} \right) = \sigma^2 \mathbb{E}_W \widehat{\mathrm{df}}^A(Y) \ .$$

Using the decomposition $AY = A\Phi x_0 + AW$, we obtain

$$\begin{aligned}
\mathbb{E}_W \|AY - A\widehat{\mu}(Y)\|_2^2 &= \mathbb{E}_W \|A\Phi x_0 + AW\|_2^2 - 2\mathbb{E}_W \langle A\Phi x_0 + AW, A\widehat{\mu}(Y) \rangle \\
&\quad + \mathbb{E}_W \|A\widehat{\mu}(Y)\|_2^2 \\
&= \mathbb{E}_W \|A\Phi x_0\|_2^2 + \sigma^2 \operatorname{tr}(A^* A) - 2\mathbb{E}_W \langle A\Phi x_0, A\widehat{\mu}(Y) \rangle \\
&\quad - 2\mathbb{E}_W \langle W, A^* A \widehat{\mu}(Y) \rangle + \mathbb{E}_W \|A\widehat{\mu}(Y)\|_2^2 \\
&= \mathbb{E}_W \|A\Phi x_0 - A\widehat{\mu}(Y)\|_2^2 \\
&\quad + \sigma^2 \operatorname{tr}(A^* A) - 2\sigma^2 \mathbb{E}_W \widehat{\mathrm{df}}^A(Y) \ .
\end{aligned}$$

Moreover, $\sum_i \mathrm{cov}_W((AY)_i, (A\widehat{\mu}(Y))_i) = \mathbb{E}_W \langle AW, A\widehat{\mu}(Y) \rangle$, which shows that $\widehat{\mathrm{df}}^A(Y)$ is indeed an unbiased estimator of df^A. ∎

Theorem 9.1 can be straightforwardly adapted to deal with any white Gaussian noise with a non-singular covariance matrix Σ. It is sufficient to consider the change of variable $y \mapsto \Sigma^{-1/2} y$ and $\Phi \mapsto \Sigma^{-1/2}\Phi$. This is similar to the work of Eldar (2009).

9.2.3 Prediction, Projected and Estimation Risk

All estimators of the form GSURE^{B} with B such that $B\Phi = A\Phi$ share the same expectation given by Theorem 9.1. Hence, there are several ways to estimate the risk in reconstructing $A\mu_0$. For the estimation of the prediction, projection and estimation risks, we now give the corresponding expressions and associated estimators (with subscript notations) as direct consequences of Theorem 9.1:

- $A = \mathrm{Id}$: in which case $\mathrm{GSURE}^{\mathrm{Id}}$ becomes

  $$\mathrm{GSURE}_{\Phi}(y) = \|y - \widehat{\mu}(y)\|_2^2 - q\sigma^2 + 2\sigma^2 \operatorname{tr}\left(\frac{\partial \widehat{\mu}(y)}{\partial y}\right)$$

  which provides an unbiased estimate of the prediction risk

  $$\mathrm{Risk}_{\Phi}(x_0) = \mathbb{E}_W \|\Phi x^\star(Y) - \Phi x_0\|_2^2 \ .$$

  This coincides with the classical SURE.

- $A = \Phi^*(\Phi\Phi^*)^+$: when Φ is rank deficient, $\Pi = \Phi^*(\Phi\Phi^*)^+\Phi$ is the orthogonal projector on $\mathrm{Ker}(\Phi)^\perp = \mathrm{Im}(\Phi^*)$. Denoting $x_{\mathrm{ML}}(y) = \Phi^*(\Phi\Phi^*)^+ y$ the maximum likelihood estimator (MLE), $\mathrm{GSURE}^{\Phi^*(\Phi\Phi^*)^+}$ becomes

  $$\begin{aligned}\mathrm{GSURE}_{\Pi}(y) = &\|x_{\mathrm{ML}}(y) - \Pi x^\star(y)\|_2^2 - \sigma^2 \operatorname{tr}\big((\Phi\Phi^*)^+\big) \\ &+ 2\sigma^2 \operatorname{tr}\left((\Phi\Phi^*)^+ \frac{\partial \widehat{\mu}(y)}{\partial y}\right).\end{aligned}$$

  It provides an unbiased estimate of the projection risk

  $$\mathrm{Risk}_{\Pi}(x_0) = \mathbb{E}_W \|\Pi x^\star(Y) - \Pi x_0\|_2^2 \ .$$

  If Φ is the synthesis operator of a Parseval tight frame, i.e. $\Phi\Phi^* = \mathrm{Id}$, the projection risk coincides with the prediction risk and so do the corresponding GSURE estimates

  $$\mathrm{Risk}_{\Pi}(x_0) = \mathrm{Risk}_{\Phi}(x_0) \quad \text{and} \quad \mathrm{GSURE}_{\Pi}(y) = \mathrm{GSURE}_{\Phi}(y) \ .$$

It is also worth noting that if $x^\star(y)$ never lies in $\mathrm{Ker}(\Phi)$, then $\mathrm{Risk}_\Pi(x_0)$ coincides with the estimation risk up to the additive constant $\|(\mathrm{Id} - \Pi)x_0\|_2^2$.

- $A = (\Phi^*\Phi)^{-1}\Phi^*$: in this case Φ has full rank, and the mapping $y \mapsto x^\star(y)$ is single-valued and weakly differentiable. The maximum likelihood estimator is now $x_{\mathrm{ML}}(y) = (\Phi^*\Phi)^{-1}\Phi^* y$, and $\mathrm{GSURE}^{(\Phi^*\Phi)^{-1}\Phi^*}$ takes the form

$$\begin{aligned}\mathrm{GSURE}_{\mathrm{Id}}(y) = \|x_{\mathrm{ML}}(y) - x^\star(y)\|_2^2 - \sigma^2 \operatorname{tr}\left((\Phi^*\Phi)^{-1}\right) \\ +2\sigma^2 \operatorname{tr}\left(\Phi(\Phi^*\Phi)^{-1}\frac{\partial x^\star(y)}{\partial y}\right) .\end{aligned}$$

This is an unbiased estimator of the estimation risk given by

$$\mathrm{Risk}_{\mathrm{Id}}(x_0) = \mathbb{E}_W\|x^\star(Y) - x_0\|_2^2 .$$

9.3 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 et al. 1992).

So far, we have shown that ouside 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 a formula holds 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 measure with density with respect to the Lebesgue measure, this allows us to establish unbiasedness of quadratic risk estimators.

Our o-minimality assumptions requires the existence of an o-minimal structure $\mathcal{O}$, see Definition 2.32, such that

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

Section 2.4 argues that this condition is not restrictive.

We assume in this section that F takes the form (8.1) and that

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

and

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

Obviously, assumption (C_τ) implies (C_{strict}), 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 (8.2), 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 (8.2), since $D_{12}^2 F_0(\mu, y) = \text{Id}$.

9.3.1 The Transition Space has Zero-Measure

Proposition 9.1 Suppose that conditions ($C_{\mathcal{O}}$) and ($C_{\mathcal{T}}$) hold. Then, $\mathcal{H}$ is of Lebesgue measure zero.

PROOF 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 ($C_{\mathcal{T}}$). Let $C \subset \mathbb{R}^n$ be the set whose gauge is J, and C° its polar.

- Since J is definable by ($C_{\mathcal{O}}$), $\nabla_1 F(x, y)$ is also definable by virtue of Proposition 2.11.
- Given $T \in \mathcal{T}$, $\widetilde{T}$ is also definable. Indeed, $\widetilde{T}$ can be equivalently written

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

  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 $\mathbf{D} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ the set-valued mapping whose graph is

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

From Lemma 2.13, $\mathcal{G}(\mathbf{D})$ is definable. Since the graph ∂J is closed (Hiriart-Urruty et al. 2001), and definable (Lemma 2.13), the set

$$\{(\beta, \eta) \mid \eta \in \operatorname{rbound} \partial J(\beta)\} = \mathcal{G}(\partial J) \setminus \mathcal{G}(\mathbf{D}) \ ,$$

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

$$\begin{aligned}\mathcal{A}_T = (\mathbb{R}^q \times \widetilde{T} \times \mathbb{R}^n) \cap \{(y, x, \eta) \mid \exists \eta, \eta = -\nabla_1 F(x_T, y)\} \\ \cap \{(\beta, \eta) \mid \eta \in \operatorname{rbound} \partial J(\beta)\} \ .\end{aligned}$$

- By axiom 4, the canonical projection $\Pi_{q+n,q}(\mathcal{A}_T)$ is definable, and its boundary $\mathcal{H}_T = \operatorname{bd}(\Pi_{q+n,q}(\mathcal{A}_T))$ is also definable by (Coste 1999, Proposition 1.12) with a strictly smaller dimension than $\Pi_{q+n,q}(\mathcal{A}_T)$ (Coste 1999, Theorem 3.22).

- We recall now from (Coste 1999, Theorem 2.10) that any definable subset $A \subset \mathbb{R}^q$ 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))

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

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

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

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

9.3.2 The Prediction is Lipschitz Continuous

PROPOSITION 9.2 Suppose that conditions (C_τ) and (C_L) hold. Then, $\widehat{\mu}$ is Lipschitz continuous.

PROOF $F_0(\cdot, y)$ 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, $(\mathcal{P}_y^F)$ amounts to solving

$$\min_{x \in \mathbb{R}^n} \frac{\tau}{2}||\Phi x||^2 + G(\Phi x, y) + \lambda J(x).$$

It can be recasted as a constrained optimization problem

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

Introducing the image (ΦJ) of J under the linear mapping Φ, it is equivalent to

$$\min_{\mu \in \mathbb{R}^q} \frac{\tau}{2}||\mu||^2 + G(\mu, y) + \lambda(\Phi J)(\mu) \ , \tag{9.3}$$

where $(\Phi J)(\mu) = \min_{\{x \in \mathbb{R}^n \mid \mu = \Phi x\}} \lambda J(x)$. This is a proper closed convex function, which is finite on $\mathrm{Im}(\Phi)$. The minimization problem amounts to computing the proximal point at 0 of $G(\cdot, y) + \lambda(\Phi J)$, 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 $\mathbb{R}^q$ with Lipschitz constant $L||y_1 - y_2||$. It then follows from (Bonnans et al. 2000, Proposition 4.32) that $\widehat{\mu}(\cdot)$ is Lipschitz continuous with constant $2L/\tau$. ∎

9.3.3 A Closed Form Expression of the DOF

We now arrive at our main contribution. The following theorem prove that the quantity $\Delta(y)$ defined in (8.5) allows us to define an unbiased estimate of the degrees of freedom, which is computable in closed form.

THEOREM 9.2 Suppose that conditions $(C_{\mathcal{O}})$, (C_{τ}) and (C_L) hold. Then,

$$\widehat{df}(y) = \operatorname{tr} \Delta(y) \text{ a.e.},$$

where Δ is defined in (8.5). Hence, $\operatorname{tr} \Delta(y)$ is an unbiased estimate of $df(y)$.

PROOF By Proposition 9.2, $\widehat{\mu}$ is Lipschitz continuous. From (Evans et al. 1992, Theorem 5, Section 4.2.3), weak differentiability follows. Rademacher theorem asserts that a Lipschitz continuous function is differentiable Lebesgue a.e. and its derivative and weak derivative coincide Lebesgue a.e., (Evans et al. 1992, Theorem 2, Section 6.2). Its weak derivative, whenever it exists, is upper-bounded by the Lipschitz constant. Thus

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

This formula is valid everywhere except on the set $\mathcal{H}$ which is of Lebesgue measure zero as shown in Proposition 9.1. We conclude by invoking (i) and Stein's lemma (Stein 1981) to establish unbiasedness of the estimator $\widehat{df}$ of the DOF. ∎

9.3.4 The (G)SURE is an Unbiased Estimator of the Risk

Gaussian Regression. Assume that the observation model (9.1) specialises to $Y \sim \mathcal{N}(\Phi x_0, \sigma^2 \mathrm{Id}_n)$.

COROLLARY 9.1 Suppose that conditions $(C_{\mathcal{O}})$, (C_{τ}) and (C_L) hold. Then, the GSURE

$$\mathrm{GSURE}^A(Y) = ||AY - A\widehat{\mu}(Y)||^2 + 2\sigma^2 \widehat{df}^A(Y) - \sigma^2 \operatorname{tr}(AA^*) \tag{9.4}$$

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

$$\widehat{\mathrm{df}}^A(Y) = \mathrm{tr}(A\Delta(Y)) \text{ a.e.}$$

PROOF By the chain rule (Evans et al. 1992, Remark, Section 4.2.2), the weak derivative of $A \circ \widehat{\mu}(\cdot)$ at y is precisely

$$D(A \circ \widehat{\mu})(y)) = A\left(\widehat{\mu}(y)\right)\Delta(y) \text{ a.e.}$$

This formula is valid everywhere except on the set $\mathcal{H}$ which is of Lebesgue measure zero as shown in Proposition 9.1. We conclude by invoking Proposition 9.2 to establish unbiasedness of the estimator $\widehat{\mathrm{df}}^A(Y)$ and using Theorem 9.1. ∎

GLM with the continuous exponential family. Assume that the observation model (9.1) corresponds to the GLM with a distribution which belongs to a continuous standard exponential family as parameterized in (9.2). Denote $\nabla \log B(y) = \left(\frac{\partial \log B_i(y_i)}{\partial y_i}\right)_i$.

COROLLARY 9.2 Suppose that conditions ($C_{\mathcal{O}}$), (C_τ) and (C_L) hold. Then, the SURE

$$\mathrm{SURE}(Y) = \|\nabla \log B(Y) - \widehat{\mu}(Y)\|^2 + 2\widehat{\mathrm{df}}(Y) - \left(\|\nabla \log B(Y)\|^2 - \|\mu_0\|^2\right) \quad (9.5)$$

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

$$\widehat{\mathrm{df}}(Y) = \mathrm{tr}\,\Delta(Y) \text{ a.e.}$$

PROOF The proof is similar but uses the result (Eldar 2009, Theorem 1) to conclude. ∎

Though $\mathrm{SURE}(Y)$ depends on μ_0, which is obviously unknown, it is only through an additive constant.

9.3.5 A Simple Example: DOF of Block Thresholding

Consider that $\Phi = \mathrm{Id}$, $J = \|\cdot\|_{1,2}$ and $F_0 = \frac{1}{\lambda}\|\cdot - y\|^2$ is the square loss. In this setting, it is known that $(\mathcal{P}^F_y)$ has a unique solution given by the block thresholding operator, i.e. for every $b \in \mathcal{B}$,

$$\widehat{\mu}(y)_b = x^\star_b = \begin{cases} 0 & \text{if } \|y_b\| \leqslant \lambda \\ \left(1 - \dfrac{\lambda}{y_b}\right) y_b & \text{otherwise.} \end{cases}$$

The estimator of the degrees of freedom reads then

$$\widehat{\mathrm{df}}(y) = |\Lambda| - \lambda \sum_{b \subseteq \Lambda} \frac{|b| - 1}{\|y_b\|} \quad \text{where} \quad \Lambda = \bigcup \{ b \in \mathcal{B} \mid \|y_b\| > \lambda \}.$$

Figure 9.1 illustrates this theorem by showing $\widehat{\mathrm{df}}(y)$ as a function of $y \in \mathbb{R}^2$ for a single block $g = \{0, 1\}$ of size 2. Note that the DOF is not constant equal to 2 outside $\mathcal{R}_\lambda = \{y \mid \|y\| > \lambda\}$ (which would be the case for a 1-D soft thresholding). It tends to 2 when $y \to +\infty$ and is equal to 1 on the boundary of $\mathcal{R}_\lambda$.

9.4 Relation to Previous Works

9.4.1 Degrees of Freedom

In the case of standard Lasso (i.e. ℓ^1 penalty) with $Y \sim \mathcal{N}(\Phi x_0, \sigma^2 \mathrm{Id}_n)$ and Φ 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 in (Dossal et al. 2013) to any arbitrary design matrix Φ. Under the same Gaussian linear regression model, unbiased estimators of the DOF for the Lasso with ℓ^1-analysis penalty, were given independently in (Tibshirani et al. 2012; Vaiter, Deledalle, et al. 2013).

A formula of an estimate of the DOF for the group Lasso when the design is orthogonal within each group was conjectured in (Yuan et al. 2005). Kato 2009 studied the DOF of a general shrinkage estimator where the regression

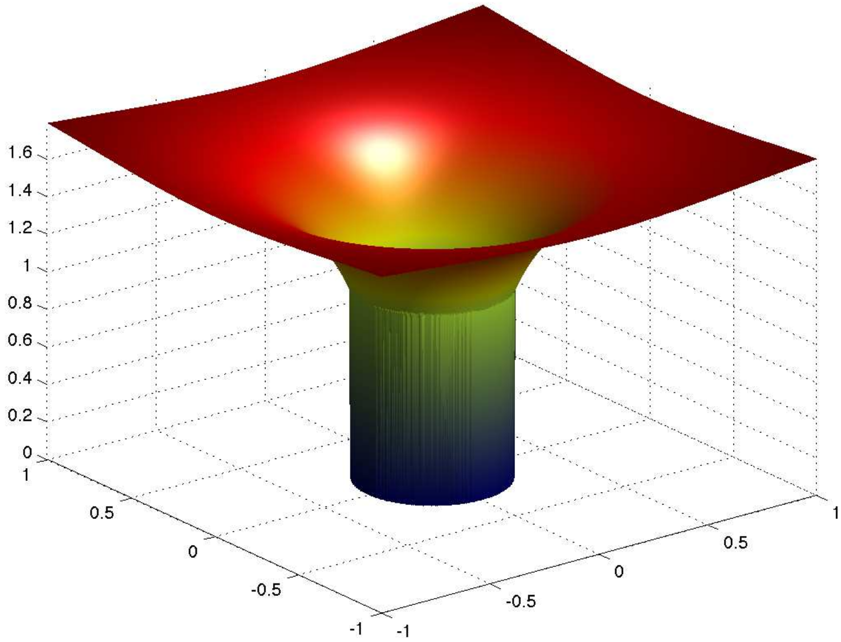


Figure 9.1: *Display of* $\widehat{\mathrm{df}}$ *in 2-D for* $\lambda = 0.3$.

coefficients are constrained to a closed convex set C. His work extends that of Meyer et al. (2000) which treats the case where C is a convex polyhedral cone. When Φ 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 Φ as (Yuan et al. 2005). An estimate of the DOF for the group Lasso was also given by Solo et al. (2010) using heuristic derivations that are valid only when Φ is full column rank, though its unbiasedness is not proved.

9.4.2 Generalized Stein Unbiased Risk Estimator

In (Eldar 2009), the author derived expressions equivalent to GSURE_Π and GSURE_{Id} up to a constant which does not depend on the estimator. However, her expressions were developed separately, whereas we have shown that these GSURE estimates originate from a general result stated in Theorem 9.1. Another distinction between our work and (Eldar 2009) lies in the assumptions imposed. Eldar (2009) supposes $x^\star(y)$ to be a weakly differentiable function of $\Phi^* y/\sigma^2$. In contrast, we just require that the prediction $y \mapsto \widehat{\mu}(y)$ (a single-valued map) is weakly differentiable, as classically assumed in the SURE theory.

Indeed, let $u = \Phi^* y/\sigma^2$, and define $x^\star(y) = z^\star_\theta(u)$. Assume that $u \mapsto z^\star_\theta(u)$ is weakly differentiable (and a fortiori a single-valued mapping).

When Φ is rank deficient, Eldar (2009) proves unbiasedness of the following estimator of the projection risk

$$\begin{aligned}\text{GSURE}_\Pi^{(\text{Eldar})}(z^\star_\theta(u)) =& \|\Pi x_0\|_2^2 + \|\Pi z^\star_\theta(u)\|_2^2 - 2\langle z^\star_\theta(u), x_{\text{ML}}(y)\rangle \\ &+ 2\,\text{tr}\left(\Pi \frac{\partial z^\star_\theta(u)}{\partial u}\right).\end{aligned}$$

Since by assumption $\frac{\partial \Phi z^\star_\theta(u)}{\partial u} = \Phi \frac{\partial z^\star_\theta(u)}{\partial u}$, and using the chain rule, the following holds

$$\sigma^2\,\text{tr}\left((\Phi\Phi^*)^+ \frac{\partial \widehat{\mu}(y)}{\partial y}\right) = \sigma^2\,\text{tr}\left((\Phi\Phi^*)^+ \frac{\partial \Phi z^\star_\theta(u)}{\partial u}\frac{\partial u}{\partial y}\right) = \text{tr}\left(\Pi \frac{\partial z^\star_\theta(u)}{\partial u}\right)$$

whence it follows that

$$\begin{aligned}\text{GSURE}_\Pi(x^\star(y)) - \text{GSURE}_\Pi^{(\text{Eldar})}(x^\star(y)) =& \|x_{\text{ML}}(y)\|_2^2 - \|\Pi x_0\|_2^2 \\ &- \sigma^2\,\text{tr}\big((\Phi\Phi^*)^+\big)\ .\end{aligned}$$

A similar reasoning when Φ has full rank leads to

$$\begin{aligned}\text{GSURE}_{\text{Id}}(x^\star(y)) - \text{GSURE}_{\text{Id}}^{(\text{Eldar})}(x^\star(y)) =& \|x_{\text{ML}}(y)\|_2^2 - \|x_0\|_2^2 \\ &- \sigma^2\,\text{tr}\big((\Phi^*\Phi)^{-1}\big)\ .\end{aligned}$$

Both our estimator GSURE_{Id} and those of (Eldar 2009) are unbiased, but they do not have necessarily the same variance. Given that they only differ by terms that do not depend on $x^\star(y)$, and in particular on a parameter (here λ), selecting the latter by minimizing our GSURE_{Id} expressions or those of (Eldar 2009) leads to the same results.

Let us finally mention that in the context of deconvolution, GSURE_{Π} boils down to the unbiased estimator of the projection risk obtained by Pesquet et al. (2009).

10

Numerical Considerations

Main contributions of this chapter

- We prove in Theorem 10.1 that under the same hypothesis of non-degeneracy and partial smoothness as those of Theorem 7.2, the forward-backward algorithm identifies the correct manifold after a finite number of steps.
- We discuss in Sections 10.2 and 10.3 how the linearized precertificate behaves in different imaging applications.
- We investigate further the behavior of total variation denoising in Theorem 10.2 and the compressed sensing with ℓ^∞ regularization in Theorem 10.3.

Contents

The effective computation of a solution to $(\mathcal{P}_{y,\lambda})$ is not the main scope of this thesis[1]. However, it seems important to give insight on how to numerically solve with such a problem in high dimension and we give basics understanding of optimization in order to compute the linearized precertificate p_F (see Chapter 5).

10.1 Introduction to Proximal Splitting

Suppose that one seeks solutions of

$$x^\star \in \underset{x \in \mathbb{R}^n}{\operatorname{Argmin}}\ f(x), \tag{10.1}$$

1. However, we developped a Python module, coined `pyprox`, available on Github: `http://github.com/svaiter/pyprox` including forward-backward, Douglas-Rachford, ADMM and generalized forward-backward (Raguet et al. 2013) algorithms. It was used for the numerical experiments of this chapter. We also refer the reader to the Numerical Tours website `http://www.numerical-tours.com`.

where f is a differentiable function with a uniformly Lipschitz gradient. Then, the most common algorithm is the gradient descent, see for instance (Boyd et al. 2004), which reads

$$x_{k+1} = x_k - \mu_k \nabla f(x_k),$$

where the step size μ_k should be small enough to ensure convergence. A major issue is that the objective function of $(\mathcal{P}_{y,\lambda})$ is in general not $C^1(\mathbb{R}^n)$ because non-smoothness of J is crucial to induce low complexity models, see Section 1.1.4. Several solutions exist in the litterature. A powerful class of methods to cope with such non-smooth, large scales, optimization problems are so-called proximal splitting schemes. One can refers to (Combettes et al. 2011) for a detailed review.

10.1.1 Proximity Operator

The proximity operator was introduced by (Moreau 1965). Its definition reads as follows.

> Definition 10.1 — Proximity Operator Let $f \in \Gamma_0(\mathbb{R}^n)$. The mapping
>
> $$\operatorname{prox}_f : x \mapsto \underset{z \in \mathbb{R}^n}{\operatorname{argmin}}\ f(z) + \frac{1}{2}\|x - z\|^2,$$
>
> is a well-defined single-valued mapping over $\mathbb{R}^n$, and coined the *proximity operator* of f.

The following proposition recaps the important properties of the proximity operator. A proof can be found in (Bauschke et al. 2011).

> Proposition 10.1 — Main properties of prox_f Let $f \in \Gamma_0(\mathbb{R}^n)$.
>
> (i) prox_f is firmly non-expansive, i.e. for every $x, z \in \mathbb{R}^n$,
>
> $$\|\operatorname{prox}_f(x) - \operatorname{prox}_f(z)\|^2 + \|(x - \operatorname{prox}_f(x)) - (z - \operatorname{prox}_f(z))\|^2 \leqslant \|x - z\|^2.$$

(ii) The set of fixed-points $\{x \in \mathbb{R}^n \mid \operatorname{prox}_f(x) = x\}$ is the set of solutions of (10.1).

(iii) For every $u, x \in \mathbb{R}^n$, one has

$$u = \operatorname{prox}_f(x) \iff x - u \in \partial f(x).$$

(iv) The Moreau identity is satisfied, i.e. for every $x \in \mathbb{R}^n$,

$$x = \operatorname{prox}_f(x) + \operatorname{prox}_{f^*}(x). \tag{10.2}$$

(v) If $g \in \Gamma_0(\mathbb{R}^m)$, for $h(x_1, x_2) = f(x_1) + g(x_2)$, one has

$$\operatorname{prox}_h(x_1, x_2) = (\operatorname{prox}_f(x_1), \operatorname{prox}_g(x_2)).$$

(vi) Define $h(x) = f(tx + a)$ with $t \neq 0$ and $a \in \mathbb{R}^n$. Then,

$$\operatorname{prox}_h(x) = \frac{1}{t}\left(\operatorname{prox}_{t^2 f}(tx + a) - a\right).$$

We shall give some examples of proximity operators.

Indicator function. Let $C \subseteq \mathbb{R}^n$ be a non-empty closed convex set and $f = \iota_C$ its indicator function. Then, $\operatorname{prox}_f = P_C$ the euclidian projection onto C. Note that when C is a linear subspace, the Moreau decomposition (10.2) reads $\mathrm{Id} = P_T + P_{T^\perp}$ which accounts for the decomposition of $\mathbb{R}^n = T + T^\perp$ into orthogonal subspaces

Quadratic objective. Let $f(x) = \frac{1}{2}\langle Ax, x\rangle + \langle b, x\rangle + c$ be a quadratic function, where A is a positive symmetric matrix. Its proximity operator reads

$$\operatorname{prox}_{\gamma f}(x) = (\mathrm{Id} + \gamma A)^{-1}(x - \gamma b).$$

General norms. Let $f(x) = ||x||$ be a norm on $\mathbb{R}^n$. Its conjugate function is the indicator function ι_B of the dual norm ball B defined as

$$B = \{x \in \mathbb{R}^n \mid ||x||^* \leqslant 1\} \quad \text{where} \quad ||x||^* = \max_{||z|| \leqslant 1} \langle x, z \rangle.$$

Using (10.2), its proximity operator reads $\text{prox}_f = \text{Id} - P_B$.

ℓ^1 **norm.** The proximity operator of ℓ^1 is the so-called soft-thresholding operator

$$(\text{prox}_{\gamma||\cdot||_1}(x))_i = \begin{cases} x_i - \gamma & \text{if } x_i \geqslant \gamma \\ 0 & \text{if } |x_i| \leqslant \gamma \\ x_i + \gamma & \text{if } x_i \leqslant \gamma. \end{cases}$$

Nuclear norm. The proximity operator of the nuclear norm is the soft-thresholding operator applies to the singular values. More precisely, if $A = U \,\text{diag}(\sigma_1, \ldots, \sigma_n) V^*$ is the SVD of A, then

$$\text{prox}_{\gamma||\cdot||_*}(A) = U \,\text{diag}(\text{prox}_{\gamma||\cdot||_1}(\sigma_1, \ldots, \sigma_n)) V^*.$$

See (Lewis 1995) for a proof.

ℓ^∞**-norm.** Using Moreau identity, we deduce that

$$\text{prox}_{\gamma||\cdot||_\infty}(x) = x - P_{B_{\ell^1}}\left(\frac{x}{\gamma}\right),$$

where $P_{B_{\ell^1}}$ is the orthogonal projection onto the ℓ^1 unit-ball. This projector can be computed using soft-thresholding and sorting, see (Fadili et al. 2011) for more details.

10.1.2 Algorithms

10.1.2.1 Proximal Point Algorithm

Point (ii) of Proposition 10.1, namely that the set of fixed-points of prox_f coincides with the minimizer of (10.1), suggests to define an algorithm, coined proximal fixed point algorithm, where the iteration are of the form

$$x_{k+1} = \operatorname{prox}_{\gamma f}(x_k),$$

for $\gamma > 0$. Even if such a scheme converges, a major issues with these iterations is that for the functionals f considered, computing prox_f cannot be done in closed form, which makes this algorithm intractable. This however suggests the introduction of more advanced iterations obtained by splitting the functional f in sum of simpler functions.

For many application in machine learning and imaging sciences, one may re-write the problem (10.1) as follows

$$x^\star \in \operatorname*{Argmin}_{x \in \mathbb{R}^n} f(x) + g(x), \tag{10.3}$$

where f and g enjoy some noticeable properties. For instance f or g might be smooth, or one might be able to compute their proximity operator in closed form.

10.1.2.2 Forward-Backward

Suppose that f is $C^1(\mathbb{R}^n)$ with a Lipschitz-continuous derivative and $g \in \Gamma_0(\mathbb{R}^n)$. In this case, one can use forward-backward iterations in order to solve (10.3). We denote by β the Lipschitz modulus of ∇f.

Every sequence $x^{(k)}$ generated by Algorithm 1 converges to a solution to the problem (10.3), see for instance (Combettes et al. 2005). Several versions of the forward-backward algorithm exists, with different relaxation parameters. Again, we refer to (Combettes et al. 2005) for a discussion on this subject. Note that the projection gradient algorithm and the iterative soft-thresholding

Algorithm 1 Forward–Backward

Choose $\varepsilon \in (0, \min(1, 1/\beta))$ and $x^{(0)} \in \mathbb{R}^n$
for $k \geqslant 0$ **do**
  $\mu_k \in [\varepsilon, 2/\beta - \varepsilon]$
  $z^{(k)} = x^{(k)} - \mu_k \nabla f(x^{(k)})$ ▷ forward-step
  $x^{(k+1)} = \text{prox}_{\mu_k g}(z^{(k)})$ ▷ backward-step
end for

algorithm (Daubechies et al. 2004) are special cases of Algorithm 1. A typical case of applications of this algorithm is to solve $(\mathcal{P}_{y,\lambda})$ when f is smooth loss such as the quadratic loss, and $g = J$ a convex regularizer. Cases where this algorithm can be applied is when $J = \|\cdot\|_1$ or $J = \|\cdot\|_*$. Note however that for more complicated regularizers, for instance $J = \|D^* \cdot\|_1$ such as the total variation, it is not possible to compute prox_J in closed form, so one needs to use more advanced algorithms.

10.1.2.3 Douglas–Rachford

The forward-backward algorithm works when one of the two functions is differentiable with a uniformly Lipschitz gradient. Suppose now that $f, g \in \Gamma_0(\mathbb{R}^n)$ such that $\text{ri}\,\text{dom}\, f \cap \text{ri}\,\text{dom}\, g \neq \emptyset$ and $f(x) + g(x) \to +\infty$ when $\|x\| \to +\infty$. For any function f in $\Gamma_0(\mathbb{R}^n)$, we write $\text{rprox}_f(x) = 2\,\text{prox}_f(x) - x$. The Douglas–Rachford has been introduced by Lions et al. (1979), in a special case, and further studied by Eckstein et al. (1992). The algorithm reads as follows. Every sequence $x^{(k)}$ generated by Algorithm 2 converges to a solution to the

Algorithm 2 Douglas–Rachford

Choose $\gamma > 0$, $0 < \mu < 2$ and $z^{(0)} \in \mathbb{R}^n$
for $k \geqslant 0$ **do**
  $z^{(k+1)} = (1 - \mu/2) z^{(k)} + \mu/2(\text{rprox}_{\gamma g} \circ \text{rprox}_{\gamma f})(z^{(k)})$
  $x^{(k+1)} = \text{prox}_{\gamma f}(z^{(k+1)})$
end for

problem (10.3), see for instance (Bauschke et al. 2011). A typical example of application is to solve $(\mathcal{P}_{y,0})$, where $f = \iota_{\{x \mid \Phi x = y\}}$ and $g = J$.

10.1.2.4 Primal Dual Splitting

In the case of analysis models, one has to solve a problem of the form

$$x^\star \in \underset{x \in \mathbb{R}^n}{\mathrm{Argmin}}\ f(x) + g(D^* x), \tag{10.4}$$

where D is a linear operator from $\mathbb{R}^p$ to $\mathbb{R}^n$. Since, in general, there is no easy way to compute the proximity operator of such mappings $g \circ D^*$, it is not possible to apply directly the forward-backward or the Douglas–Rachford algorithms.

Algorithm 3 Relaxed Arrow-Hurwicz primal-dual splitting.

Choose $0 \leqslant \theta \leqslant 1$, $\sigma\tau \|D^*\|^2 < 1$ and $u^{(0)}, x^{(0)}, z^{(0)} \in \mathbb{R}^n$
for $k \geqslant 0$ **do**
  $u^{(k+1)} = \mathrm{prox}_{\sigma g^*}(u^{(k)} + \sigma D^* z^{(k)})$ ▷ dual step
  $x^{(k+1)} = \mathrm{prox}_{\tau f}(x^{(k)} - \tau D u^{(k)})$ ▷ primal step
  $z^{(k+1)} = x^{(k+1)} + \theta(x^{(k+1)} - x^{(k)})$
end for

Every sequence $x^{(k)}$ generated by Algorithm 3 converges to a solution to the problem (10.4), see for instance (Chambolle et al. 2011).

10.1.3 Identifying Activity with Forward-Backward Splitting

While we showed in Chapter 7 that under some mild conditions, the manifold $\mathcal{M}$ is stable, this result only holds for the exact minimizer $x^\star$ of $(\mathcal{P}_{y,\lambda})$. In practice, $x^\star$ is only computed approximately by some iterates $x^{(k)}$ of an optimization scheme. It is thus of practical importance to be able to understand whether the results of Chapter 7 allow to shed some light on the structure of manifolds activated by the sequence of iterates, and their relation to the manifold of the original object to recover. In this section, we answer this question in the case of the forward-backward algorithm when applied to solve $(\mathcal{P}_\theta)$ with J partly smooth.

The following theorem shows that $\mathcal{M}$ is indeed correctly identified by the forward-backward after a finite number of iterations.

THEOREM 10.1 Suppose that the assumptions of Theorem 7.2 hold. Then, for k sufficiently large, $x^{(k)} \in \mathcal{M}$, where $x^{(k)}$ is the sequence generated by Algorithm 1.

PROOF The proof of this result follows the same line as Theorem 7.2 and use (Hare et al. 2007, Theorem 5.3) to concludes. A close inspection of the proof of Theorem 7.2 reveals that $\eta_\theta = \text{ri}(\partial J(x_\theta))$ for the assumed regime of $(||w||, \lambda)$. This in turn implies that the assumptions of (Hare et al. 2007, Theorem 5.3), are fulfilled. We then conclude arguing in a similar way as in (Hare et al. 2007, Theorem 2). ∎

10.2 Robust Sparse Analysis Regularization

We illustrate in this section our theoretical findings on several examples to study the robustness of the 1-D total variation, shift-invariant Haar and Fused Lasso regularizations, which are special cases of analysis ℓ^1 regularization.

10.2.1 Total Variation Denoising

The discrete 1-D total variation (TV) corresponds to taking $D = D_{\text{DIF}}$ as defined in (1.12). We now define a subclass of signals in order to study the stability of there jumps.

DEFINITION 10.2 A signal is said to *contain a staircase* if there exists $i \in \{1 \dots |I| - 1\}$ such that

$$\text{sign}(D_I^* x)_i = \text{sign}(D_I^* x)_{i+1} = \pm 1.$$

Figure 10.1 shows examples of signals with (left) and without (right) staircase subsignals.

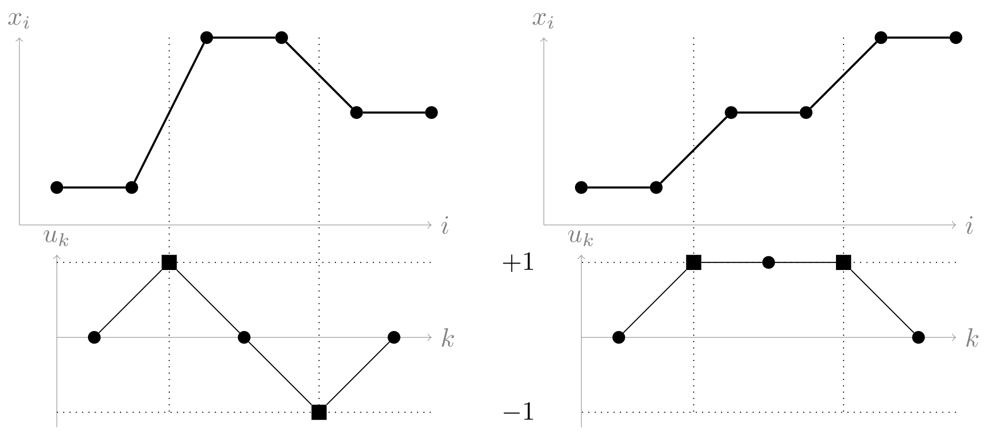


Figure 10.1: *Top row: Two examples of signals* x *having 2 jumps. Bottom row: Associated vector* $\Phi^* p_F = \operatorname{div} u$.

The following result allows to characterize robustness of TV regularization when $\Phi = \mathrm{Id}$, i.e. TV denoising.

> THEOREM 10.2 We consider the case where $\Phi = \mathrm{Id}$. If x_0 does not contain a staircase, then $\mathbf{IC}(x_0) < 1$. Otherwise, $\mathbf{IC}(x_0) = 1$.

PROOF Let $x^\star$ be the unique solution of $(\mathcal{P}_{y,\lambda})$ with $T_x^\star = \operatorname{Ker} D_\Lambda$. We denote Ω the matrix

$$\Omega = D_{(I^c)}^+ \left(\Phi^* \Phi \, P_T (\Phi_T^* \Phi_T)^{-1} \, P_T - \mathrm{Id}\right) D_{(I)} s_{(I)}$$

the vector η defined as

$$\eta : \begin{cases} \eta_I = s_I = \operatorname{sign}(D_I^* x) \\ \eta_J = \sigma = \Omega s_I. \end{cases}$$

The vector σ satisfies $(D_J^* D_J)\sigma = -(D_J^* D_I) s_I$. One can show that this implies that η is the solution of a discrete Poisson equation

$$\forall j \in J, \quad (\Delta \eta)_j = 0 \quad \text{and} \quad \begin{cases} \forall i \in I, \, \eta_i = s_i, \\ \eta_0 = \eta_N = 0. \end{cases}$$

where $\Delta = DD^*$ is a discrete Laplacian operator. This implies that for $i_1 < k < i_2$ where i_1, i_2 are consecutive indices of I, m is obtained by linearly interpolating (see Figure 10.1) the values η_{i_1} and η_{i_2}, i.e

$$\eta_k = \rho\eta_{i_1} + (1-\rho)\eta_{i_2} \quad \text{where} \quad \rho = \frac{k-i_1}{i_2-i_1}.$$

Hence, if x_0 does not contain a staircase subsignal, one has $\mathbf{IC}(x_0) < 1$. On the contrary, if there is i_1 such that $s_{i_1} = s_{i_2}$, where i_1 and i_2 are consecutive indices of I, then for every $i_1 < j < i_2, \eta_j = s_{i_1} = \pm 1$ which implies that $\mathbf{IC}(x_0) = 1$. ■

This theorem together with Theorem 7.2 shows that if a signal x_0 does not have a staircase subsignal, then TV denoising identifies correctly the jump set when the noise is small. This means that if w is small enough, for λ proportional to the noise level, the TV denoised version of y contains the same jumps as x_0.

To gain a better understanding of the latter situation, we build an instructive family of signals x_0 for which the **IC** criterion is equal to 1. It turns out that depending on the structure of the noise w, the D-support of x_0, $\text{supp}(D^*x_0)$, can be either stably identified or not.
For n a multiple of 4, we split $\{1, \ldots, n\}$ into 4 sets $l_k = \{(k-1)M+1, ..., kM\}$ of cardinality $M = n/4$. Let $\mathbf{1}_{l_k}$ be the boxcar signal whose support is l_k . We define the staircase signal $x_0 = -\mathbf{1}_{l_1} + \mathbf{1}_{l_4}$ degraded by a deterministic noise w of the form $w = \varepsilon(\mathbf{1}_{l_3} - \mathbf{1}_{l_2})$, where $\varepsilon \in \mathbb{R}$. The observation vector $y = x_0 + w$ reads

$$y = -\mathbf{1}_{l_1} - \varepsilon\mathbf{1}_{l_2} + \varepsilon\mathbf{1}_{l_3} + \mathbf{1}_{l_4} \ .$$

Suppose that $\varepsilon > 0$, then the solution $x^\star_\lambda$ of $\mathcal{P}_\lambda(y)$ is

$$x^\star_\lambda = \left(-1 + \frac{\lambda}{M}\right)\mathbf{1}_{l_1} - \varepsilon\mathbf{1}_{l_2} + \varepsilon\mathbf{1}_{l_3} + \left(1 - \frac{\lambda}{M}\right)\mathbf{1}_{l_4},$$

if $0 \leqslant \lambda \leqslant \lambda_1 = M(1-\varepsilon)$, and

$$x^\star_\lambda = \left(-\varepsilon + \frac{\lambda - \lambda_1}{2M}\right)(\mathbf{1}_{l_1} + \mathbf{1}_{l_2}) + \left(\varepsilon - \frac{\lambda-\lambda_1}{2M}\right)(\mathbf{1}_{l_3} + \mathbf{1}_{l_4}),$$

if $\lambda_1 \leqslant \lambda \leqslant \lambda_2 = \lambda_1 + 2\varepsilon M$, and 0 if $\lambda > \lambda_2$. Similarly, if $\varepsilon < 0$, the solution $x^\star_\lambda$

reads

$$x^\star_\lambda = \left(-1 + \frac{\lambda}{M}\right) \mathbf{1}_{l_1} - \left(\varepsilon + 2\frac{\lambda}{M}\right) (\mathbf{1}_{l_2} - \mathbf{1}_{l_3}) + \left(1 - \frac{\lambda}{M}\right) \mathbf{1}_{l_4},$$

if $0 \leqslant \lambda \leqslant \bar{\lambda}_1 = -\varepsilon \frac{M}{2}$, and

$$x^\star_\lambda = \left(-1 + \frac{\lambda}{M}\right) \mathbf{1}_{l_1} + \left(1 - \frac{\lambda}{M}\right) \mathbf{1}_{l_4},$$

if $\bar{\lambda}_1 \leqslant \lambda \leqslant \bar{\lambda}_2 = M$, and 0 if $\lambda > \bar{\lambda}_2$. Figure 10.2 displays plots of the the coordinates' paths for both cases. It is worth pointing out that when $\varepsilon > 0$, the

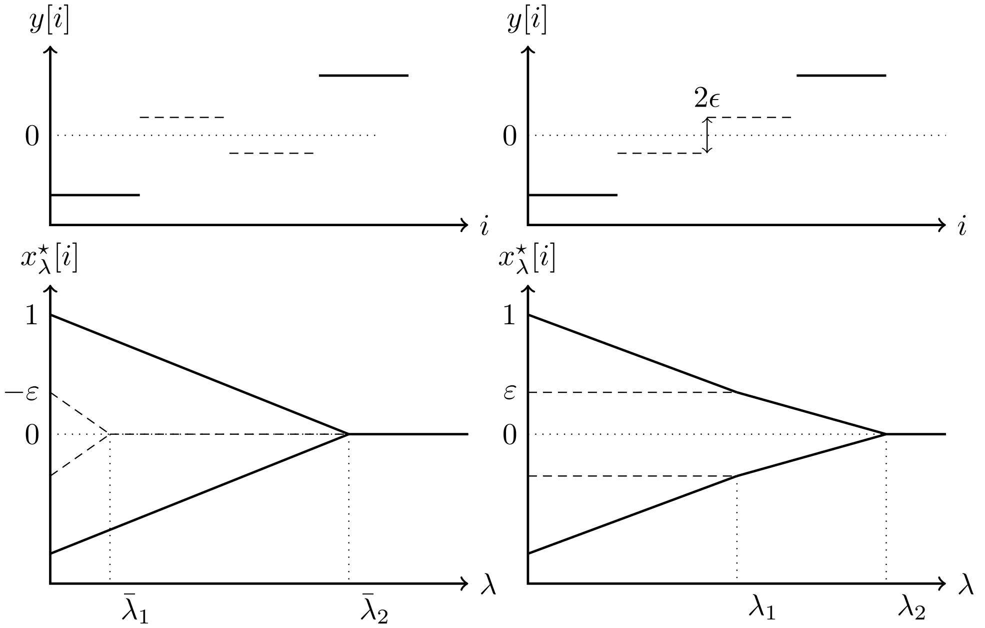


Figure 10.2: *Top row: Signals y for $\varepsilon < 0$ (left) and $\varepsilon > 0$ (right). Bottom row: Corresponding coordinates' path of $x^\star_\lambda$ as a function of λ. The solid lines correspond to the coordinates in l_1 and l_4, and the dashed ones to the coordinates in l_2 and l_3.*

D-support of $x^\star_\lambda$ is always different from that of x_0 whatever the choice of λ, whereas in the case $\varepsilon < 0$, for any $\bar{\lambda}_1 \leqslant \lambda \leqslant \bar{\lambda}_2$, the D-support of $x^\star_\lambda$ and sign of $D^* x^\star_\lambda$ are exactly those of x_0.

10.2.2 Total Variation Compressed Sensing

We compare numerically the difference between the linearized precertificate $\mathfrak{p}_F$ and the analysis precertificate, see Definition 5.6. In this case, the analysis certificate reads

$$\mathfrak{p}_A = -D(ND_{S_x})^+ NDe_x,$$

where N^* is a basis of $\operatorname{Ker}\Phi$. Note that this precertificate cannot be used in Theorem 7.2, but can be used in Theorem 6.1 of ensure ℓ^2 noise robustness. Figure 10.3 shows an example of $\mathfrak{p}_F$ and $\mathfrak{p}_A$ for a single realization of Φ. We consider the realization Φ drawn from the Gaussian ensemble with redundancy $q/n = \frac{1}{3}$ and a signal x with 5 piecewise constant components. In this compressed sensing scenario, one can see that $\mathfrak{p}_A$ behaves much better than $\mathfrak{p}_F$. Indeed, $\mathrm{Phi}^*\mathfrak{p}_A$ is strictly within $\mathrm{ri}(\partial J(x_0))$, which is not the case for $\mathrm{Phi}^*\mathfrak{p}_F$.

10.2.3 Shift-Invariant Haar Deconvolution

Sparse analysis regularization using a 1-D shift invariant Haar dictionary is efficient to recover piecewise constant signals. This dictionary is built using a set of scaled and dilated Haar filters

$$\psi_i^{(j)} = \frac{1}{2^{\tau(j+1)}} \begin{cases} +1 & \text{if } 0 \leqslant i < 2^j \\ -1 & \text{if } -2^j \leqslant i < 0 \\ 0 & \text{otherwise,} \end{cases}$$

where $\tau \geqslant 0$ is a normalization exponent. For $\tau = 1$, the dictionary is said to be *unit-normed*. For $\tau = 1/2$, it corresponds to a *Parseval tight-frame*. The action on a signal x of the analysis operator corresponding to the translation invariant Haar dictionary D_H is

$$D_H^* x = \left(\psi^{(j)} \star x\right)_{0 \leqslant j \leqslant J_{\max}},$$

where $\star$ stands for the discrete convolution (with appropriate boundary conditions) and $J_{\max} < \log_2(n)$, where n is the size of the signal. The analysis regularization $\|D_H^* x\|_1$ can also be written as the sum over scales of the TV

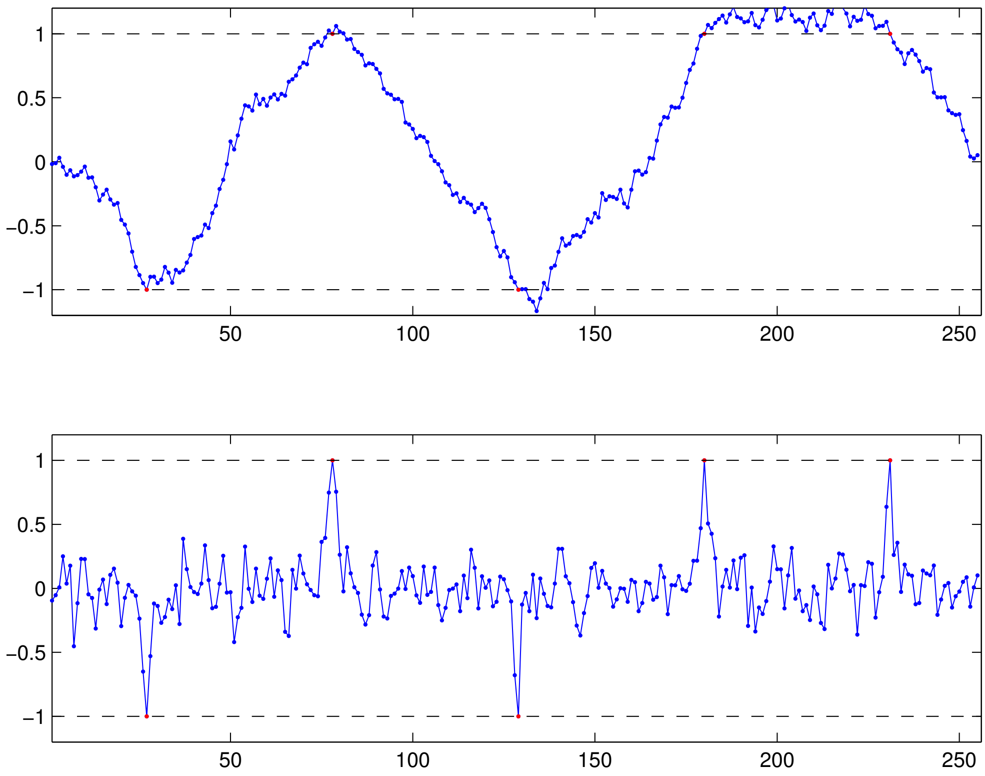


Figure 10.3: *Total Variation Compressed Sensing with* $q/n = 1/3$. *Top:* u_F *such that* $\Phi^* p_F = \operatorname{div} u_F$. *Bottom:* u_A *such that* $\Phi^* p_A = \operatorname{div} u_A$.

semi-norms of filtered versions of the signal. As such, it can be understood as a sort of multiscale total variation regularization. Apart from a multiplicative factor, one recovers Total Variation when $J_{max} = 0$.

We consider a convolution setting (for $n = 256$) where Φ is a circular convolution operator with a Gaussian kernel of standard deviation σ. We first study the impact of σ on the identifiability criterion **IC**. The original signal x_η is a centered boxcar signal with a support of size $2\eta n$

$$x_\eta = \mathbb{1}_{\{\lfloor n/2-\eta n\rfloor,\dots,\lfloor n/2+\eta n\rfloor\}}, \quad \eta \in (0, 1/2] .$$

Figure 10.4 displays the evolution of $\mathbf{IC}(x_0)$ as a function of σ for three dictionaries: the total variation dictionary and the Haar wavelet dictionary with two normalization exponents $\tau = 1$ and $\tau = 0.5$. In this experiment, we chose $\eta = 0.2$. One can observe that the three curves pass through 1 for the same value of σ (near 1 here). This shows that for σ small enough, deconvolv-

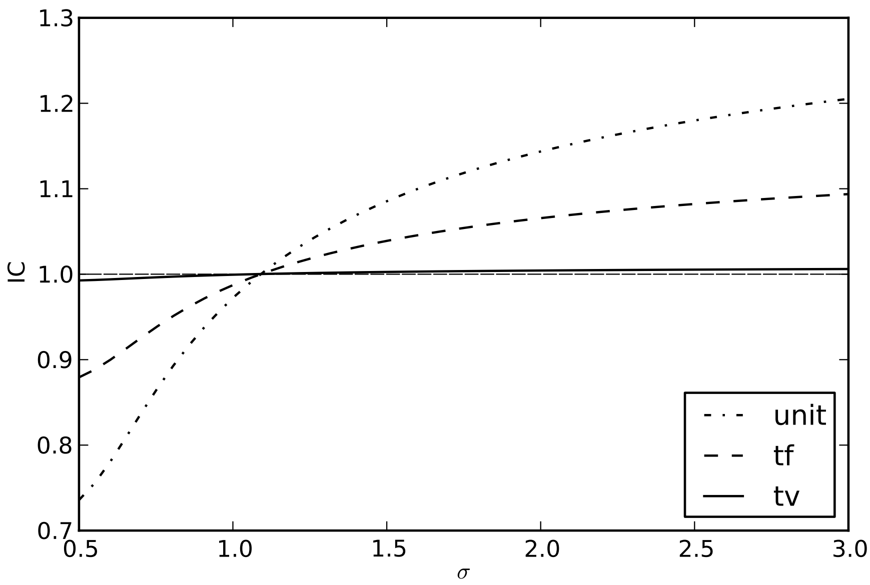


Figure 10.4: *Behavior of* **IC** *for a noiseless deconvolution scenario with a Gaussian blur and ℓ_1-analysis sparsity regularization in a shift invariant Haar dictionary with $J_{max} = 4$.* **IC** *is plotted as a function of the Gaussian blurring kernel size $\sigma \in [0.5, 3.0]$ for the total variation dictionary and the Haar wavelet dictionary with two normalization exponents τ. Dash-dotted line: $\tau = 1$ (unit-normed). Dashed line: $\tau = 1/2$ (tight-frame). Solid line: total variation.*

ing a box signal is stable in the sense that the discontinuities are correctly estimated in the presence of a small additive noise in the observations. In addition, in the identifiability regime, $\mathbf{IC}(x_0)$ appears smaller in the case of the unit-normed normalization (i.e. $\tau = 1$). However, one should avoid to infer stronger conclusions since a detailed computation of the constants involved in Theorem 7.2 would be necessary to completely and fairly compare the stability performance achieved with each of these three dictionaries.

10.2.4 Fused Lasso Compressed Sensing

Fused Lasso was introduced in Tibshirani et al. 2005. It corresponds to taking

$$D = \begin{bmatrix} D_{\text{DIF}} & \varepsilon \text{Id} \end{bmatrix},$$

and $J = ||D^* \cdot ||_1$ in $(\mathcal{P}_{y,\lambda})$, where $\varepsilon > 0$. If $x = \sum_{i=1}^{k} \gamma_i \mathbf{1}_{[a_i, b_i]}$, where $\gamma_i \in \mathbb{R}$ and $a_i \leqslant b_i < a_{i+1}$, then the model space T_x reads

$$T_x = \left\{ \sum_{i=1}^{k} \rho_i \mathbf{1}_{[a_i, b_i]} \;|\; \rho_i \in \mathbb{R} \right\}.$$

This significates that the Fused Lasso favors sparse sums of boxcar signals.

We consider a compressed sensing setting (with the signal size $n = 256$) and examine the behavior of **IC** with respect to the undersampling ratio q/n and the true signal properties. Φ is drawn from the standard Gaussian ensemble, i.e. $\Phi_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, 1)$. The sampled signal $x_{\eta,\rho}$ is the superposition of two boxcars distant from each other by $2\rho N$ and each of support size ηN

$$x_{\eta,\rho} = \mathbf{1}_{\{\lfloor (\frac{1}{2} - \eta - \rho) n \rfloor, \ldots, \lfloor (\frac{1}{2} - \rho) n \rfloor\}} + \mathbf{1}_{\{\lfloor (\frac{1}{2} + \rho) n \rfloor, \ldots, \lfloor (\frac{1}{2} + \eta + \rho) n \rfloor\}}.$$

In our simulations, we fixed $\rho = 0.1$.

Figure 10.5 depicts the evolution of the empirical probability with respect to the sampling of Φ of the event **IC** < 1 as a function of the sampling ratio $Q/N \in [0.5, 1]$ and the boxcar support size $\eta \in [0.025, 0.15]$. This probability is computed from 1000 Monte-Carlo replications of the sampling of Φ. With no surprise, one can clearly see that the probability increases as more measurements are collected. This probability profile also seems to be increasing as η decreases, but this is likely to be a consequence of the choice of the Fused Lasso parameter ε, and the conclusion may be different for other choices.

This is indeed confirmed in our last experiment whose results are displayed in Figure 10.6. It shows the evolution of the empirical probability of the event

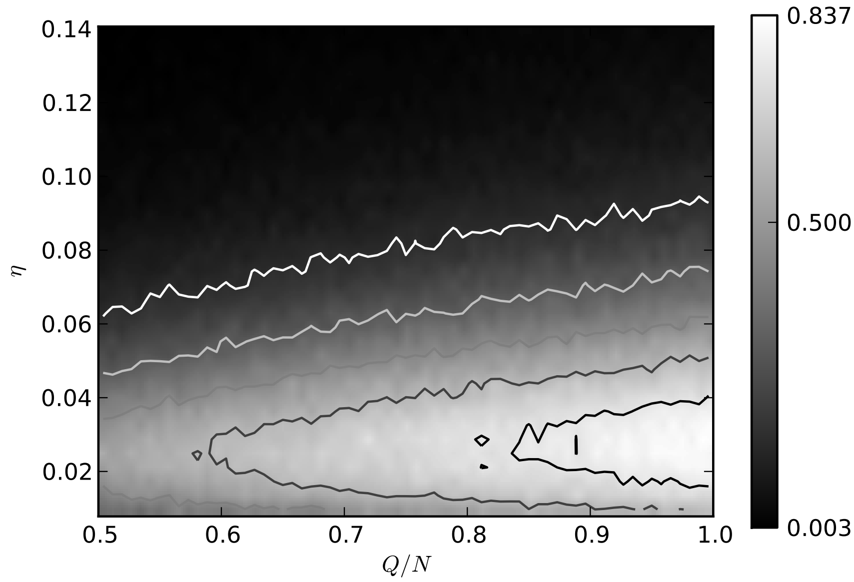


Figure 10.5: *Behavior of* **IC** *for a compressed sensing scenario matrix with a Gaussian measurement matrix and the Fused Lasso regularization. Empirical probability of the event* $\mathbf{IC} < 1$ *as a function of the sampling ratio* $q/n \in [0.5, 1]$ *and the support size* $\eta \in [0.025, 0.15]$ *with* $\varepsilon = 50/N$.

$\mathbf{IC} < 1$ as a function of the Fused Lasso parameter $\varepsilon \in [1/n, 200/n]$ and the support size $\eta \in [0.025, 0.15]$. This probability is again computed from 1000 Monte-Carlo replications. Depending on the choice of ε, the probability profile does not necessarily exhibit a monotonic behavior as a function of η. For large values (more weight on Id in the Fused Lasso dictionary), the probability decreases monotonically as η increases which can be explained by the fact that higher η corresponds to less sparse signals. As ε is lowered, higher weight is put on the TV regularization, and the behavior is not anymore monotonic. Now, the probability reaches a peak at intermediate values of η and then vanishes quickly. The peak probability also decreases with decreasing ε.

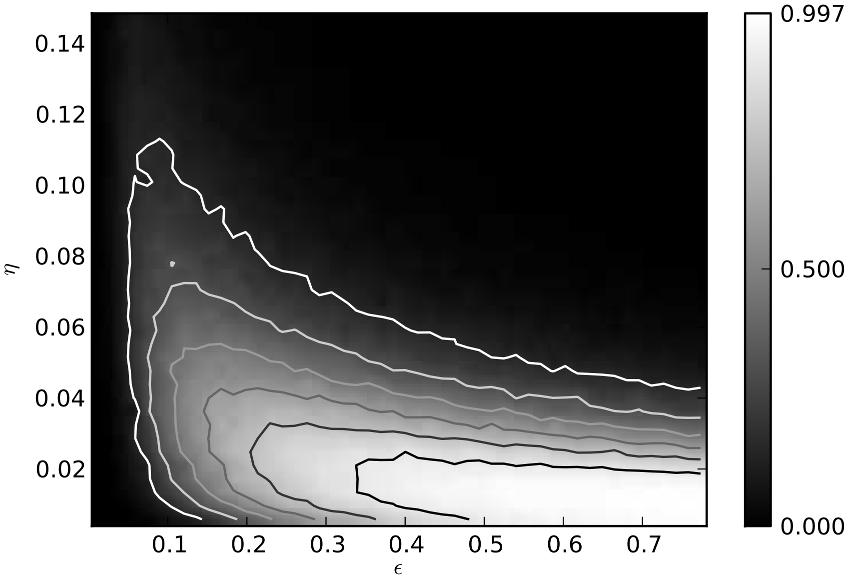


Figure 10.6: *Behavior of* $\mathbf{IC}$ *for a compressed sensing scenario matrix with a Gaussian measurement matrix and the Fused Lasso regularization. Empirical probability of the event* $\mathbf{IC} < 1$ *as a function of the parameter* $\varepsilon \in [1/n, 200/n]$ *and the support size* $\eta \in [0.025, 0.15]$ *with* $q/n = 0.8$.

10.3 Robust Antisparse Regularization

In some cases, one aims at recovering flat vectors, i.e such that for most i, $x_i = ||x||_\infty$. This is for instance the case in computer vision applications when performing quantization of random projections, see (Jégou et al. 2012). One can use as regularizer the ℓ^∞ norm defined as

$$||x||_\infty = \max_{i \in \{1,\ldots,n\}} |x_i|.$$

In this case, one has

$$T_x = \left\{ x' \mid x'_I = \rho x_I \text{ for some } \rho \in \mathbb{R} \right\},$$

where $I = \{i \mid x_i = ||x||_\infty\}$. This reflects that fact that $J = ||\cdot||_\infty$ favors signal having the same saturation pattern as x, see Proposition 3.15.

The following result studies the behavior of the linearized precertificate in a compressed sensing scenario.

THEOREM 10.3 Let x be an arbitrary vector with its saturation support I, its model subspace $T_x = S_x^\perp$ and generalized sign vector e_x as defined above. Let $\beta > 1$. For Φ drawn from the standard Gaussian ensemble with

$$m \geqslant n - |I| + 2\beta |I| \log(|I|/2) \ ,$$

$\mathbf{IC}(x) < 1$ with probability at least $1 - 2(|I|/2)^{-f(\beta,|I|)}$ where

$$f(\beta, |I|) = \left(\sqrt{\frac{\beta}{2|I|} + \beta - 1} - \sqrt{\frac{\beta}{2|I|}} \right)^2 \ .$$

PROOF To lighten the notation, we drop the dependence on x of T, S and e. Without of loss of generality, by symmetry of the norm, we will assume that the entries of x are positive.

We follow the same program as in the CS literature. The key ingredient of the proof is the fact that owing to the isotropy of the Gaussian ensemble, p_F and Φ_S^* are independent. Thus, for some $\tau > 0$

$$\Pr(\mathbf{IC}(x) \geqslant 1) \leqslant \Pr\left(\mathbf{IC}(x) \geqslant 1 \,\middle|\, ||p_F|| \leqslant \tau\right) + \Pr(||p_F|| \geqslant \tau) \ .$$

As soon as $m \geqslant \dim(T) = n - |I| + 1$, Φ_T is full-column rank. Thus

$$||p_F||^2 = \langle e, \, (\Phi_T^* \Phi_T)^{-1} e \rangle \ .$$

$\left(\Phi_T^* \Phi_T\right)^{-1}$ is an inverse Wishart matrix with m degrees of freedom. To estimate the deviation of this quadratic form, we use classical results on inverse χ^2 random variables with $m - n + |I|$ degrees of freedom and we get the tail bound

$$\Pr\left(||p_F|| \geqslant \sqrt{\frac{1}{m - n + |I| - t}} ||e|| \right) \leqslant e^{-\frac{t^2}{4(m-n+|I|)}}$$

for $t > 0$. Now, conditionally on $\mathfrak{p}_F$, the entries of $\alpha_S = P_S\, \Phi^* \mathfrak{p}_F$ are i.i.d. $\mathcal{N}(0, \|\mathfrak{p}_F\|^2)$ and so are those of $-\alpha_S$ by trivial symmetry of the centered Gaussian. Thus, using a union bound, we get

$$\begin{aligned}
\Pr\left(\mathbf{IC}(x) \geqslant 1 \middle| \|\mathfrak{p}_F\| \leqslant \tau\right) &\leqslant \Pr\left(\max_{i\in I} \left(-(\alpha_{S_x})_i\right)_+ \geqslant 1/|I| \middle| \|\mathfrak{p}_F\| \leqslant \tau\right) \\
&\leqslant \Pr\left(\max_{i\in I} \left((\alpha_{S_x})_i\right)_+ \geqslant 1/|I| \middle| \|\mathfrak{p}_F\| \leqslant \tau\right) \\
&\leqslant |I| \Pr\left((z)_+ \geqslant 1/(\tau|I|)\right) \\
&\leqslant |I| \Pr\left(z \geqslant 1/(\tau|I|)\right) \\
&\leqslant |I| e^{-\frac{1}{2\tau^2|I|^2}} .
\end{aligned}$$

Observe that $(\alpha_S)_i = 0$ for all $i \in I^c$. Choosing

$$\tau = \sqrt{\frac{1}{|I|(m-n+|I|-t)}}$$

where we used that $\|e\| = 1/\sqrt{I}$, and inserting in the above probability terms, we get

$$\begin{aligned}
\Pr\left(\|\mathfrak{p}_F\| \geqslant \tau\right) &\leqslant e^{-\frac{t^2}{4(m-n+|I|)}} , \\
\Pr\left(\mathbf{IC}(x) \geqslant 1 \middle| \|\mathfrak{p}_F\| \leqslant \tau\right) &\leqslant e^{-\left(\frac{m-n+|I|-t}{2|I|} - \log(|I|/2)\right)} .
\end{aligned}$$

Equating the arguments of the exponentials and solving

$$\frac{t^2}{4q} + \frac{t}{2|I|} - \left(\frac{q}{2|I|} - \log\left(\tfrac{|I|}{2}\right)\right) = 0$$

for t to get equal probabilities, we get

$$t = \frac{q}{|I|}\left(\sqrt{1 + 2|I|\left(1 - 2\frac{2|I|\log\left(\frac{|I|}{2}\right)}{q}\right)} - 1\right) ,$$

where $q = m - n + |I| \geqslant 1$ by the injectivity assumption. Setting

$$\beta = \frac{q}{2|I| \log\left(\frac{|I|}{2}\right)} ,$$

we get under the bound on m that $\beta > 1$, and

$$t = 2\beta \log\left(\frac{|I|}{2}\right) \left(\sqrt{1 + 2|I|^{\frac{\beta-1}{\beta}}} - 1\right) .$$

Inserting t in one of the probability terms, and after basic algebraic rearrangements, we get the probability of success with the expression of the function $f(\beta, |I|)$. ∎

The above bound and probability bears some similarities to what we get with ℓ^1 minimization, except that now the probability of success scales in a power of $|I|$ and not n directly. The reason underlying such a similarity is the proof technique usual in CS-type bounds and the use of the minimal ℓ^2-norm dual certificate. In particular, a union bound is behind the log factor. If some improvements is sought after, it is on this step that it can be gained.

The map $f(\beta, |I|)$ is an increasing function of $|I|$, so that $\lim_{|I| \to \infty} f(\beta, |I|) = \beta - 1$ and the probability of success increases with increasing size of the saturation support. But this comes at the price of a stronger requirement on the number of measurements.

For the noiseless problem $(\mathcal{P}_{y,0})$, it can be shown using arguments based on the statistical dimension (Amelunxen et al. 2013) of the descent cone of the ℓ^∞-norm that there is a phase transition exactly at $n - |I|/2$, see (Chandrasekaran et al. 2012, Proposition 3.12). The reason is that each face of the descent cone of the hypercube at a point living on its k-dimensional face is the direct sum of a subspace (the linear hull of the face), and of an orthant of dimension $n - k$ (up to an isometry). The statistical dimension is then $(n-k)/2 + k = (n+k)/2 = n - |I|/2$, observing that $k = n - |I|$.

10.4 Efficient GSURE Computation for Analysis ℓ^1

In this section, we exemplify the usefulness of our GSURE estimator, see Definition 9.1, which can serve as a basis for automatically tuning the value of λ in the case of analysis ℓ^1-sparsity, i.e. $J = ||D^* \cdot ||_1$. This is achieved by computing, from a single realization of the noise $w \sim \mathcal{N}(0, \sigma^2 \mathrm{Id})$, the parameter λ that minimizes the value of GSURE when solving $(\mathcal{P}_{y,\lambda})$ from $y = \Phi x_0 + w$ for various scenarios on Φ and x_0. Note that this method can be adapted to other analysis regularizers.

Specializing Theorem 9.1 to this case, we have the following result.

Corollary 10.1 We assume that the observation model is $Y \sim \mathcal{N}(\Phi x_0, \sigma^2 \mathrm{Id}_n)$. In this case,

$$\mathrm{GSURE}^A(Y) = ||A(Y - \widehat{\mu}(Y))||^2 + 2\sigma^2 \widehat{\mathrm{df}}^A(Y) - \sigma^2 \operatorname{tr}(A^* A) \tag{10.5}$$

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

$$\widehat{\mathrm{df}}^A(Y) = \dim T = \dim \operatorname{Ker} D^*_\Lambda, \quad \Lambda = \operatorname{supp}(D^* x^\star)^c \ ,$$

with $x^\star$ is such that (INJ_T), where $T = T_{x^\star}$ i.e.

$$\operatorname{Ker} D^*_\Lambda \cap \operatorname{Ker} \Phi = \{0\}.$$

10.4.1 Computing the GSURE

According to Lemma 8.3, there always exists a solution of $(\mathcal{P}_{y,\lambda})$ such that (INJ_T) holds, and this solution can be computed, see (Vaiter, Deledalle, et al. 2013) for the analysis ℓ^1 prior. With assumption (INJ_T) at hand, we now define the following matrix whose role will be clarified shortly.

DEFINITION 10.3 Let Λ be a D-cosupport. Suppose that (INJ_T) holds. We define the matrix $\Gamma^{[\Lambda]}$ as

$$\Gamma^{[\Lambda]} = U\left(U^*\Phi^*\Phi U\right)^{-1} U^*. \tag{10.6}$$

where U is a matrix whose columns form a basis of $\operatorname{Ker} D_\Lambda^*$.

Observe that the action of $\Gamma^{[\Lambda]}$ can be rewritten as a quadratic optimization under linear constraint

$$\Gamma^{[\Lambda]}u = \underset{D_\Lambda^* x = 0}{\operatorname{argmin}} \frac{1}{2}||\Phi x||^2 - \langle x,\, u\rangle.$$

The remaining obstacle faced when implementing the GSURE formula of Theorem 9.1 is to compute the divergence term. However, for large scale-data as in image and signal processing, the computational storage required for the matrix in the argument of the trace would be prohibitive. Additionally, computing $\Gamma^{[\Lambda]}$ can only be reasonably afforded for small data size. Fortunately, the structure of $\widehat{\mathrm{df}}^A(y)$ and the definition of $\Gamma^{[\Lambda]}$ allows to derive an efficient and principled way to compute the trace term. This is formalized in the next result.

PROPOSITION 10.2 One has

$$\widehat{\mathrm{df}}^A(y) = \mathbb{E}_Z(\langle \nu(Z),\, \Phi^* A^* A Z\rangle) \tag{10.7}$$

where $Z \sim \mathcal{N}(0, \mathrm{Id}_p)$, and where for any $z \in \mathbb{R}^p$, $\nu = \nu(z)$ solves the following linear system

$$\begin{pmatrix} \Phi^*\Phi & D_J \\ D_J^* & 0 \end{pmatrix} \begin{pmatrix} \nu \\ \tilde{\nu} \end{pmatrix} = \begin{pmatrix} \Phi^* z \\ 0 \end{pmatrix}. \tag{10.8}$$

PROOF We have

$$\operatorname{tr}\left[A\Phi\Gamma^{[\Lambda]}\Phi^*A^*\right] = \operatorname{tr}\left[\Phi\Gamma^{[\Lambda]}\Phi^*A^*A\right].$$

Hence denoting $\nu(z) = \Gamma^{[\Lambda]}\Phi^* z$, and using the fact that for any matrix U, $\operatorname{tr} U = \mathbb{E}_Z \langle Z, UZ \rangle$, we arrive at (10.7).

We then use the fact that $\Gamma^{[\Lambda]}\Phi^*$, the inverse of Φ on $\operatorname{Ker} D_\Lambda^*$, is the mapping that solves the following linearly constrained least-squares problem

$$\Gamma^{[\Lambda]}\Phi^* z = \underset{h \in \operatorname{Ker} D_\Lambda^*}{\operatorname{argmin}} \ \|\Phi h - z\|_2^2.$$

Writing the KKT conditions of this problem leads to (10.8), where $\tilde{\nu}$ are the Lagrange multipliers. ∎

In practice, the empirical mean estimator is replaced for the expectation in (10.7), hence giving

$$\frac{1}{k} \sum_{i=1}^{k} \langle \nu(z_i), \Phi^* A^* A z_i \rangle \overset{\text{WLLN}}{\longrightarrow} \widehat{\mathrm{df}}^A(y) \,, \tag{10.9}$$

for k realizations z_i of Z, where WLNN stands for the Weak Law of Large Numbers. Consequently, the computational bulk of computing an estimate of $\widehat{\mathrm{df}}^A(y)$ is invested in solving for each $\nu(z_i)$ the symmetric linear system (10.8) using e.g. a conjugate gradient solver.

10.4.2 Parameter Selection using the GSURE

Super-Resolution with Total Variation Regularization In this example, Φ is a vertical sub-sampling operator of factor two (hence $q/n = 0.5$). The noise level has been set such that the observed image y has a peak signal-to-noise ratio (PSNR) of 27.78 dB. We used an anisotropic total variation regularization; i.e. the sum of the ℓ^1-norms of the partial derivatives in the first and second direction (not to be confused with the isotropic total variation). Fig. 10.7.d depicts the projection risk and its GSURE_Π estimate obtained from (10.9) with $k = 1$ as a function of λ. The curves appear unimodal and coincide even with $k = 1$ and a single noise realization. Consequently, GSURE_Π provides a high-quality selection of λ minimizing the projection risk. Close-up views of the central parts of the degraded, restored (using the optimal λ), and true images are shown in Fig. 10.7(a)-(c) for visual inspection of the restoration quality.

Compressed Sensing with Wavelet Analysis Regularization We consider in this example a compressed sensing scenario where Φ is a random partial DCT measurement matrix with an under-sampling ratio $q/n = 0.5$. The noise is such that input image y has a PSNR set to 27.50 dB. We took D as the shift-invariant Haar wavelet dictionary with 3 scales. Again, we estimate GSURE_Π with $k = 1$ in (10.9). The results observed on the super-resolution example are confirmed in this compressed sensing experiment both visually and qualitatively, see Fig. 10.8.

10.4.3 Relation to Previous Work

In least-squares regression regularized by a sufficiently smooth penalty term, the DOF can be estimated in closed-form (Solo 1996). However even in such simple cases, the computational load and/or storage can be prohibitive for large-scale data.

To overcome the analytical difficult for general non-linear estimators, when no closed-form expression is available, first attempts developed bootstrap-based (asymptotically) unbiased estimators of the DOF (Efron 2004). Ye (1998) and Shen et al. (2002) proposed a data perturbation technique to approximate the DOF (and the SURE) when its closed-form expression is not available or numerically expensive to compute. For denoising, a similar Monte-Carlo approach has been used in (Ramani et al. 2008) where it was applied to total-variation denoising, wavelet soft-thresholding, and Wiener filtering/smoothing splines.

Alternatively, an estimate can be obtained by recursively differentiating the sequence of iterates that converges to a solution of the original minimization problem. Initially, it has been proposed by Vonesch et al. (2008) to compute the GSURE of sparse synthesis regularization by differentiating the sequence of iterates of the forward-backward splitting algorithm. We have recently proposed a generalization of this methodology to any proximal splitting algorithm, and exemplified it on ℓ^1-analysis regularization including the isotropic total-variation regularization, and $\ell^1 - \ell^2$ synthesis regularization which promotes block sparsity (Deledalle et al. 2014).

In our case, we have shown that the computation of an accurate estimator of the DOF for analysis ℓ^1 regularization, and therefore of GSURE^A for various risks, can be obtained by solving a linear system. This is more efficient than the previous general-purpose iterative methods that are computationally expensive.

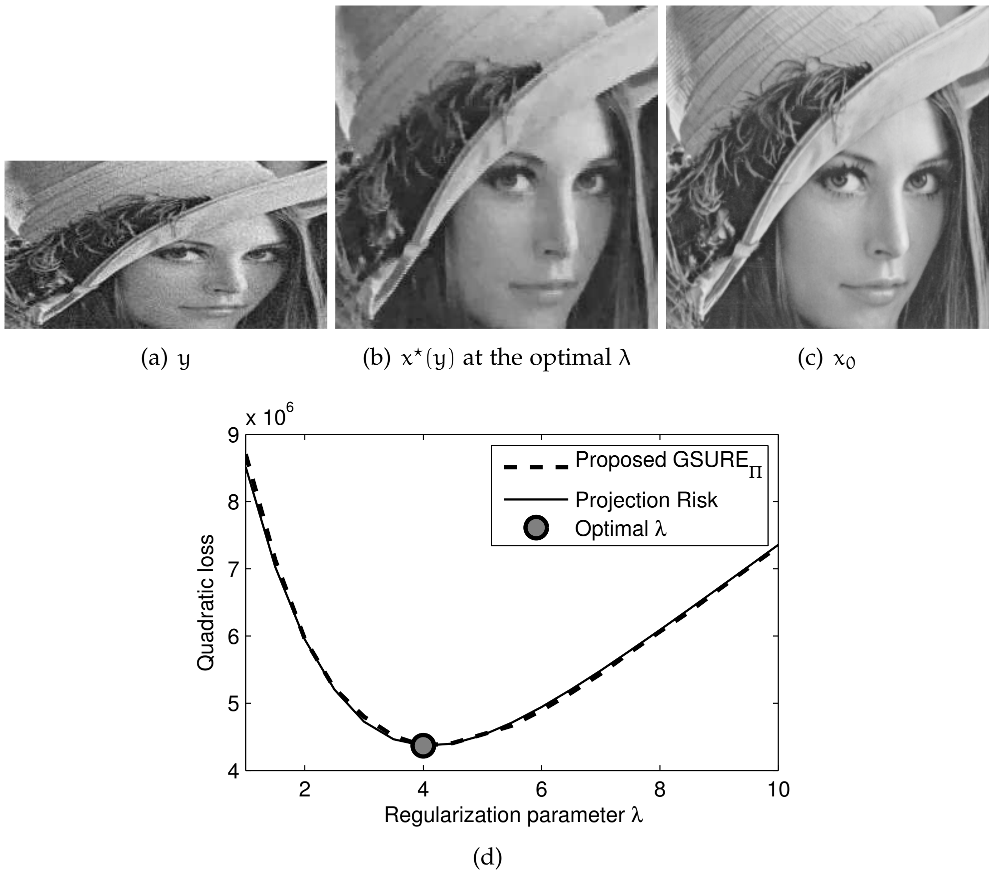
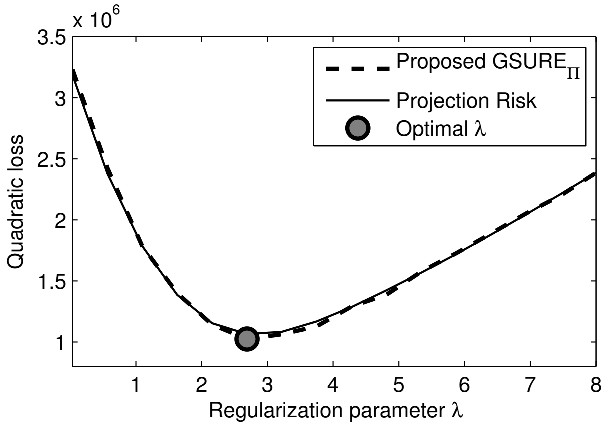


Figure 10.7: *Illustration of the selection of* λ *by minimizing* GSURE_Π *in a super-resolution problem* ($q/n = 0.5$) *with anisotropic total variation regularization. (a) The observed image* y*. (b) A solution* $x^\star(y)$ *of* $(\mathcal{P}_{y,\lambda})$ *at the optimal* λ *(the one minimizing* GSURE_Π*). (c) The underlying true image* x_0*. (d) Projection risk* Risk_Π *and its* GSURE_Π *estimate obtained from* (10.9) *using* $k = 1$ *random realization.*

(a) x_{ML} (b) $x^\star(y)$ at the optimal λ (c) x_0



(d)

Figure 10.8: *Illustration of the selection of* λ *by minimizing* GSURE_Π *in a compressed sensing problem* $(q/n = 0.5)$ *by an* ℓ^1*-analysis regularization in a shift-invariant Haar wavelet dictionary. (a) The MLE* x_{ML}*. (b) A solution* $x^\star(y)$ *of* $(\mathcal{P}_{y,\lambda})$ *at the optimal* λ *(the one minimizing* GSURE_Π*). (c) The underlying true image* x_0*. (d) Projection risk* Risk_Π *and its* GSURE_Π *estimate obtained from* (10.9) *using* $k = 1$ *random realization.*

11

Conclusion

THIS thesis revolves around the theme of *sensitivity analysis* of optimization problems. Each part is a variation on a popular theme of sensitivity analysis, such as Lipschitz continuity of the set of the minimizers of $(\mathcal{P}_{y,\lambda})$ when $\lambda \geqslant 0$ or manifold stability when $\lambda = 0$ and $\lambda > 0$. This leads us to different applications: ℓ^2-robustness (Chapter 6), model identifiability (Chapter 7), local differentiability (Chapter 8), unbiased estimation (Chapter 9) and algorithmic identifiability (Chapter 10).

The theoretical analysis provided by this work draws a connection between these popular applications in imaging, signal processing and machine learning. Partial smoothness allows us to recover results already known in the litterature, within a coherent and unifying framework. It also allows us to significantly extend these results to a larger class of regularizers and to gain a better understanding of the effects of these regularizers.

The research program does not stop here. Many extensions of our work are of interest.

Non convex regularizers and data loss. Non-convex functions are often used in image processing or statistics. There are two differents kinds of non-convexity which arise, i.e. on the data fidelity term and the regularizer.

$$\min_{x \in \mathbb{R}^n} F(x, y) + J(x),$$

For instance, one thinks of the analysis ℓ^p regularization, i.e.

$$\min_{x \in \mathbb{R}^n} \frac{1}{2}||y - \Phi x||^2 + \lambda ||D^* x||_p^p,$$

where $0 < p < 1$ and $||u||_p = (\sum |u_i|^p)^{1/p}$. In practice, using such a non-convex functional seems to produce better result in imaging and computer graphics, most probably because it better fits the high level of sparsity of natural image gradients. We believe that our results can be extended to non-convex partly smooth functions.

Unbounded functions. Our analysis does not cover the case of variational formulation with constraints. A typical example is when one imposes non-negativity constraints. For instance, the problem

$$\min_{x \in \mathbb{R}^n} \frac{1}{2}||y - \Phi x||^2 + \lambda G(x) \quad \text{subject to} \quad \forall i, x_i \geqslant 0,$$

which can be recasted as

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

where $J = G + \iota_{\mathcal{C}}$ and $\mathcal{C} = \{x \mid \forall i, x_i \geqslant 0\}$. Unfortunately, the function J is unbounded, due to the presence of $\iota_{\mathcal{C}}$, hence not covered by our results.

Continuous setting. The continuous problem, defined in $(c\mathcal{P}_{y,\lambda})$,

$$f^\star \in \operatorname*{argmin}_{f \in \mathbb{H}} \frac{1}{2}||y - \Psi f||^2 + \lambda \mathcal{J}(f) \qquad (c\mathcal{P}_\lambda(y))$$

leads to the same questions as $(\mathcal{P}_{y,\lambda})$. For instance, when $\mathcal{J}$ is the total variation of a function (1.14), what can be said about the stability of the discontinuity set of a solution ? This would extend the result of Caselles et al. (2007) obtained when $\Phi = \mathrm{Id}$ and $w = 0$.

Publications

Preprints

(1) M. J. Fadili, G. Peyré and S. Vaiter. Linear Convergence Rates for Gauge Regularization. In preparation.

(2) C. Deledalle, S. Vaiter, G. Peyré, and M. J. Fadili. Stein Unbiased GrAdient Risk estimates (SUGAR) for multiple parameter selection. 2014. arXiv:1405.1164.

(3) S. Vaiter, M. J. Fadili, and G. Peyré. Partly Smooth Regularization of Inverse Problems. 2014. arXiv:1405.1004.

(4) S. Vaiter, C. Deledalle, G. Peyré, M. J. Fadili, and C. Dossal. The Degrees of Freedom of Partly Smooth Gauge Regularizers. 2014. arXiv:1404.5557.

(5) S. Vaiter, M. Golbabaee, M. J. Fadili, and G. Peyré. Model Selection with Piecewise Regular Gauges. 2013. arXiv:1307.2342.

Peer Reviewed International Journals

(6) S. Vaiter, G. Peyré, C. Dossal, and M. J. Fadili. Robust Sparse Analysis Regularization. *IEEE Transactions on Information Theory*, 59(4):2001–2016, 2013.

(7) S. Vaiter, C. Deledalle, G. Peyré, C. Dossal, and M. J. Fadili. Local Behavior of Sparse Analysis Regularization: Applications to Risk Estimation. *Applied and Computational Harmonic Analysis*, 35(3):433–451, 2013.

International Conferences and Workshops

(8) S. Vaiter, G. Peyré, and M. J. Fadili. Robust Polyhedral Regularization. *In* Sampling Theory and Applications (SampTA). 2013.

(9) M. J. Fadili, G. Peyré, S. Vaiter, C. Deledalle, and J. Salmon. Stable Recovery with Analysis Decomposable Priors. *In* Sampling Theory and Applications (SampTA). 2013.

(10) S. Vaiter, C. Deledalle, G. Peyré, M. J. Fadili, and C. Dossal. The degrees of freedom of the Group Lasso for a General Design. *In* Signal Processing with Adaptive Sparse Structured Representations (SPARS). 2013.

(11) C. Deledalle, S. Vaiter, G. Peyré, M. J. Fadili, and C. Dossal. Proximal Splitting Derivatives for Risk Estimation. *In* New Computational Methods for Inverse Problems (NCMIP). 2012.

(12) C. Deledalle, S. Vaiter, G. Peyré, M. J. Fadili, and C. Dossal. Risk estimation for matrix recovery with spectral regularization. *In* International Conference on Machine Learning Workshop (ICML). 2012.

(13) S. Vaiter, C. Deledalle, G. Peyré, M. J. Fadili, and C. Dossal. Degrees of Freedom of the Group Lasso. *In* International Conference on Machine Learning Workshop (ICML). 2012.

(14) C. Deledalle, S. Vaiter, G. Peyré, M. J. Fadili, and C. Dossal. Unbiased Risk Estimation for Sparse Analysis Regularization. *In* International Conference on Image Processing (ICIP). 2012.

National Conferences and Workshops

(15) S. Vaiter, G. Peyré, and M. J. Fadili. Robustesse au bruit des régularisations polyhédrales. *In* GRETSI. 2013.

(16) M. J. Fadili, G. Peyré, S. Vaiter, C. Deledalle, and J. Salmon. Reconstruction Stable par Régularisation Décomposable Analyse. *In* GRETSI. 2013.

List of Notations

γ_C	Gauge associated to a convex set C	59
GSURE^A	Generalized Stein Unbiased Risk Estimate associated to A	206
$\mathcal{H}$	Transition space	192
$\mathbf{IC}(x)$	Identifiability criterion at x	148
$\operatorname{int} C$	Topological interior of C	48
ι_C	Indicator function of C	45
$\operatorname{Ker} f$	Kernel or zeros set of a function f	44
$\operatorname{span} C$	Linear hull of C	48
$\mathcal{M}$	A manifold	113
$\|A\|_{J_1 \to J_2}$	Operator bound of A with respect to J_1 and J_2	69
$\mathcal{O}$	An O-minimal structure	71
$\partial f(x)$	Subdifferential of f at a point x	54
Φ_T	ΦP_T	42
$\Pi_{k,n}$	Projection onto the first k components.	71
P_T	Orthogonal projection of the vector space T	42
$\mathbb{SL}_x(\cdot)$	Set of partly smooth functions at x with linear manifold	117
$\mathbb{S}_x(\mathcal{M})$	Set of partly smooth functions at x relative to $\mathcal{M}$	113
$\mathbb{S}(\mathcal{M})$	Set of partly smooth functions relative to $\mathcal{M}$	113
$\operatorname{ri} C$	Relative interior of C	48
σ_C	Support function of C	46

$\sigma_k(A)$	k-th singular value of A	68
$\text{slev}_x J$	Sublevel set to J at x	44
$T_C(x)$	Tangent cone to C at x	49
$\mathcal{T}_x(\mathcal{M})$	Tangent space of $\mathcal{M}$ at x	58
A^*	Adjoint matrix of A	42
A^+	Moore–Penrose pseudo-inverse of A	42
C°	Polar of a set C	60
$D_\eta^J(x, x_0)$	Bregman divergence between x and x_0 with respect to η	55
e_x	Model vector at x	82
$f \overset{+}{\vee} g$	Infimal convolution of f and g	46
$f'(x, \delta)$	Directional derivative of f at a point x in the direction δ	54
f^*	Legendre–Fenchel conjugate	46
S_x	Orthogonal space to the tangent model space at x	82
T_x	Tangent model space at x	82
x_T	Projection of x on T, $x_T = P_T x$	42
$(\mathcal{C}_{x,y})$	(General) restricted injectivity condition	190
(INJ_T)	Restricted injectivity condition on T	140

Résumé des travaux

Contexte. Cette thèse concerne la résolution de problèmes inverses linéaires en dimension finie. Elle contribue ainsi à l'étude théorique de thématiques centrales en traitement du signal ou d'image, en statistique ainsi qu'en apprentissage. Un tel problème peut être écrit sous la forme

$$y = \Phi x_0 + w,$$

où $y \in \mathbb{R}^q$ est le vecteur d'observations, $x_0 \in \mathbb{R}^n$ les données inconnues à retrouver, Φ un opérateur linéaire de $\mathbb{R}^q$ dans $\mathbb{R}^n$ et w un terme de bruit additif. Ce modèle inclut de nombreux cas typiques en imagerie tels que le débruitage, la déconvolution, l'interpolation, l'échantillonnage compressé ainsi que la tomographie. L'opérateur linéaire Φ est généralement mal conditionné. C'est la raison pour laquelle il s'avère nécessaire de mettre en place une stratégie de reconstruction. Un cadre classique est celui des méthodes variationnelles, pouvant s'écrire sous la forme

$$x^\star \in \underset{x \in \mathbb{R}^n}{\text{Argmin}} \frac{1}{2}\|y - \Phi x\|_2^2 + \lambda J(x), \qquad (\mathcal{P}_{y,\lambda})$$

où J est une fonction de $\mathbb{R}^n$ dans $\mathbb{R}_+$ que l'on considérera dans cette thèse convexe. Il s'agit de réaliser un compromis entre fidélité aux données (terme quadratique ici) et régularisation, représentée ici par J. Ce compromis est dicté par le choix du paramètre λ. L'opérateur Φ n'étant généralement pas injectif, il est important de garder en mémoire le fait que $x^\star$ n'est pas uniquement déterminé. Quand $w = 0$, c'est-à-dire en absence de bruit, $(\mathcal{P}_{y,\lambda})$ se réduit sous la forme

$$x^\star \in \underset{x \in \mathbb{R}^n}{\text{Argmin}}\ J(x) \quad \text{sujet à} \quad \Phi x = y. \qquad (\mathcal{P}_{y,0})$$

Le choix de la pénalité J est un problème de recherche actif. Une des directions possibles est de considérer J comme promouvant des données dites de faible complexité. Plus précisément, en se donnant une collection de sous-espaces vectoriels $\mathcal{T}$ de $\mathbb{R}^n$, nous sommes amenés à considérer le problème de sélection de modèle non-convexe

$$\inf_{T \in \mathcal{T}, x \in T} \|y - \Phi x\|^2 + \lambda \mathbf{pen}(T),$$

où typiquement $\mathbf{pen}(T) = \dim T$. Ce problème étant non seulement non-convexe, mais souvent également NP-difficile, il est nécessaire de considérer une méthode d'approximation. Dans cette thèse, nous nous consacrons aux relaxations convexes. Ainsi la fonction de comptage (taille du support d'un vecteur) est relâchée par la norme ℓ^1, le rang d'une matrice est approché par la norme nucléaire, etc. Ayant fixé une régularisation convexe J, nous définissons le vecteur modèle associé à $x \in \mathbb{R}^n$ comme

$$e_x = \underset{e \in \operatorname{aff} \partial J(x)}{\operatorname{argmin}} \|e\|,$$

où $\operatorname{aff} \partial J(x)$ est l'espace affine engendré par la sous-différentielle de J en x. Nous définissons également l'*espace modèle tangent* comme

$$T_x = \operatorname{span}(\partial J(x))^\perp,$$

où $\operatorname{span} C$ est l'espace linéaire engendré par C. Par exemple, si $J = \|\cdot\|_1$, alors $T_x = \{\eta \mid \operatorname{supp}(\eta) \subseteq \operatorname{supp}(x)\}$ et $e_x = \operatorname{sign}(x)$.

Robustesse. Le premier axe de cette thèse est dédié à l'étude de la robustesse du problème $(\mathcal{P}_{y,\lambda})$. Nous cherchons à estimer combien un vecteur solution $x^\star$ s'approche du vecteur d'origine x_0 à la fois au sens d'une erreur ℓ^2, mais également en terme de sélection de modèle. Nous montrons le théorème suivant de **convergence linéaire**.

Théorème 1 Soit T_0 l'espace modèle tangent de x_0. Supposons que :

- il existe $\alpha = \Phi^* \eta \in \operatorname{ri} \partial J(x_0)$, dite condition source, où $\operatorname{ri} \partial J(x)$ est l'intérieur relatif de $\partial J(x)$ pour la topologie induite par $\operatorname{aff} \partial J(x)$,

- $\operatorname{Ker}\Phi \cap T_0 = \{0\}$ (injectivité restreinte).

Si $\lambda = c\varepsilon$, $c > 0$, alors pour tout minimiseur $x^\star$ de $(\mathcal{P}_{y,\lambda})$,

$$||x^\star - x_0||_2 \leqslant C\varepsilon \ ,$$

où

$$C = C_1\,(2 + c||\eta||_2) + C_2 \frac{(1 + c||\eta||_2/2)^2}{cC_\eta},$$

$C_1 > 0$ et $C_2 > 0$ étant deux constantes indépendantes de η et $0 < C_\eta < 1$.

Nous dirons que J est une *fonction partiellement lisse* (Lewis 2002) pour une variété $\mathcal{M}$ si, pour tout point $x \in \mathcal{M}$, J restreinte à $\mathcal{M}$ est C^2 autour de x, l'espace tangent $\mathcal{T}_{\mathcal{M}}(x)$ à $\mathcal{M}$ en x est T_x et que l'application multivoque ∂J est continue au point x relativement à $\mathcal{M}$. Pour ce type de régularisation, incluant les normes ℓ^1, $\ell^1 - \ell^2$, nucléaire ou encore ℓ^∞, nous montrons le résultat suivant de **sélection de modèle**.

Théorème 2 Soient $x_0 \in \mathbb{R}^n$, $T = T_{x_0}$ et $e = e_{x_0}$. Supposons que :

- J est partiellement lisse pour la variété $\mathcal{M}$ et $x_0 \in \mathcal{M}$,
- $\Phi^*\Phi_T^{+,*}e \in \operatorname{ri}\partial J(x_0)$,
- $\operatorname{Ker}\Phi \cap T = \{0\}$.

Alors il existe des constantes positives C, C' telles que, si w et λ sont choisis tels que

$$||w|| \leqslant C \quad \text{et} \quad \lambda = C'||w||, \tag{.1}$$

la solution $x^\star$ du problème $(\mathcal{P}_{y,\lambda})$ est unique et satisfait

$$x^\star \in \mathcal{M} \quad \text{et} \quad ||x_0 - x^\star|| = O(||w||).$$

Sensibilité. Le second axe de cette thèse porte sur l'analyse de sensibilité de $(\mathcal{P}_{y,\lambda})$. Cette analyse permet, dans le cadre d'observations aléatoires, la construction d'un estimateur du risque quadratique non biaisé. Nous introduisons l'*espace de transition* $\mathcal{H}$, correspondant aux observations y telles que l'espace T associé à une solution de $(\mathcal{P}_{y,\lambda})$ ne soit pas stable vis-à-vis de petites perturbations de y.

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

$\Pi_{n+p,n}$ est la projection canonique sur les n premières composantes, bd C est le bord de C, et

$$\mathcal{A}_T = \left\{ (y, x_T) \in \mathbb{R}^n \times \widetilde{T} \mid \ \Phi_T^*(\Phi x_T - y) \in \text{rbd}\,\partial J(x_T) \right\}.$$

Notre première contribution est de déterminer le **comportement local** des solutions du problème $(\mathcal{P}_{y,\lambda})$ à l'extérieur de cet ensemble. Nous notons J_T la restriction de J à T.

Théorème 3 Soit J une fonction 1-homogène partiellement lisse pour $\mathcal{M} = T_{x_0}$. Soient $y \notin \mathcal{H}$ et $x^\star$ une solution de $(\mathcal{P}_{y,\lambda})$ telle que

$$\text{Ker}\,\Phi_T \cap \text{Ker}\,D^2 J_T(x^\star) = \{0\} \qquad (I_{x^\star})$$

où $T = T_{x^\star}$. Alors il existe un voisinage $\mathcal{V} \subset \mathbb{R}^n$ de y et une application $\tilde{x} : \mathcal{V} \to T$ tels que :

(i) pour tout $\bar{y} \in \mathcal{V}$, $\tilde{x}(\bar{y})$ est une solution de $(\mathcal{P}_{y,\lambda})$, et $\tilde{x}(y) = x^\star$,

(ii) l'application $\tilde{x}$ est $C^1(\mathcal{V})$ et

$$\forall\, \bar{y} \in \mathcal{V}, \quad D_1 \tilde{x}(\bar{y}) = -(\Phi_T^* \Phi_T + D^2 J_T(x^\star))^{-1} \Phi_T,$$

L'application $y \mapsto \widehat{\mu}(y) = \Phi x^\star$ est univoque et $C^1(\mathbb{R}^n \setminus \mathcal{H})$. Pour tout $y \notin \mathcal{H}$, il existe une solution $x^\star$ de $(\mathcal{P}_{y,\lambda})$ telle que $(I_{x^\star})$ est satisfaite. De plus, pour tout $y \notin \mathcal{H}$,

$$\text{div}(\widehat{\mu})(y) = \text{tr}(\Delta(y))$$

où

$$\Delta(y) = -\Phi_T({\Phi_T}^*\Phi_T + D^2 J_T(x^\star))^{-1} \circ {\Phi_T}^*.$$

Soit $Y = \Phi x_0 + W$ avec $W \sim \mathcal{N}(0, \sigma^2 \mathrm{Id}_n)$. Le degré de liberté (DOF) d'une procédure statistique quantifie la complexité de celle-ci. Suivant la définition d'Efron (1986), le DOF est défini comme

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

Dans ce cadre, nous montrons le théorème suivant d'**estimation du risque**.

THÉORÈME 4 Soit J une fonction partiellement lisse pour $\mathcal{M} = T_{x_0}$ définissable dans une structure o-minimale. Alors :

(i) $\mathcal{H}$ est de mesure de Lebesgue nulle,

(ii) $\widehat{\mu}$ est une fonction Lipschitz, donc faiblement différentiable, avec un gradient borné p.p,

(iii) $\widehat{\mathrm{df}} = \mathrm{tr}(\Delta(Y))$ est un estimateur sans biais de $\mathrm{df} = \mathbb{E}(\mathrm{div}(\widehat{\mu}(Y))$,

(iv) le SURE, défini par

$$\mathrm{SURE}(\widehat{\mu})(Y) = ||Y - \widehat{\mu}(Y)||^2 + 2\sigma^2 \widehat{\mathrm{df}} - n\sigma^2, \qquad (.2)$$

est un estimateur non biaisé de $\mathbb{E}\left(||\widehat{\mu}(Y) - \mu_0||^2\right)$.

Publications Cette thèse reprend le contenu d'articles de journaux internationaux ou de prépublications suivants.

(1) M. J. Fadili, G. Peyré et S. Vaiter. Linear Convergence Rates for Gauge Regularization. En préparation.

(2) S. Vaiter, M. J. Fadili, et G. Peyré. Partly Smooth Regularization of Inverse Problems. 2014. arXiv:1405.1004.

(3) S. Vaiter, C. Deledalle, G. Peyré, M. J. Fadili, et C. Dossal. The Degrees of Freedom of Partly Smooth Gauge Regularizers. 2014. arXiv:1404.5557.

(4) S. Vaiter, M. Golbabaee, M. J. Fadili, et G. Peyré. Model Selection with Piecewise Regular Gauges. 2013. arXiv:1307.2342.

(5) S. Vaiter, G. Peyré, C. Dossal, et M. J. Fadili. Robust Sparse Analysis Regularization. *IEEE Transactions on Information Theory*, 59(4):2001–2016, 2013.

(6) S. Vaiter, C. Deledalle, G. Peyré, C. Dossal, et J. Fadili. Local Behavior of Sparse Analysis Regularization: Applications to Risk Estimation. *Applied and Computational Harmonic Analysis*, 35(3):433–451, 2013.

Index

Bibliography

Monographs

Ambrosio, L., N. Fusco, and D. Pallara. 2000. *Functions of bounded variation and free discontinuity problems.* Clarendon Press Oxford. (See p. 21).

Aubin, J.-P., and H. Frankowska. 2009. *Set-Valued Analysis.* Modern Birkhauser Classics. Birkhauser. (See pp. 49, 117).

Auslender, A., and M. Teboulle. 2003. *Asymptotic cones and functions in optimization and variational inequalities.* Springer. (See pp. 51–53, 133).

Bauschke, H. H., and P. L. Combettes. 2011. *Convex analysis and monotone operator theory in Hilbert spaces.* Springer. (See pp. 23, 221, 225).

Bonnans, J. F., and A. Shapiro. 2000. *Perturbation analysis of optimization problems.* Springer Series in Operations Research and Financial Engineering. Springer Verlag. (See pp. 31, 33, 199, 212).

Borwein, J. M., and A. S. Lewis. 2010. *Convex analysis and nonlinear optimization: theory and examples.* Springer. (See p. 136).

Boyd, S. P., and L. Vandenberghe. 2004. *Convex optimization.* Cambridge university press. (See pp. 22, 221).

Brown, L. D. 1986. *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory.* Monograph Series. IMS, Hayward, CA: Institute of Mathematical Statistics Lecture Notes. (See pp. 188, 204).

Bühlmann, P., and S. Van De Geer. 2011. *Statistics for high-dimensional data: methods, theory and applications.* Springer. (See p. 188).

DasGupta, A. 2008. *Asymptotic Theory of Statistics and Probability.* Springer. (See p. 204).

Daubechies, I. 1992. *Ten Lectures on Wavelets.* CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM. (See p. 20).

Dontchev, A. L., and R. T. Rockafellar. 2009. *Implicit functions and solution mappings: A view from variational analysis.* Springer. (See p. 53).

Dries, L. van den. 1998. *Tame topology and o-minimal structures.* Math. Soc. Lecture Note. Cambridge Univ Press. (See pp. 70, 71).

Ekeland, I., and R. Teman. 1974. *Analyse convexe et problemes variationelles.* Paris: Dunod / Gauthier Villars. (See p. 43).

Elad, M. 2010. *Sparse and redundant representations: from theory to applications in signal and image processing.* Springer. (See p. 15).

Evans, L. C., and R. F. Gariepy. 1992. *Measure theory and fine properties of functions.* CRC Press. (See pp. 209, 213, 214).

Herman, G. T. 2009. *Fundamentals of computerized tomography: image reconstruction from projections.* Springer. (See p. 8).

Hiriart-Urruty, J.-B., and C. Lemaréchal. 2001. *Convex Analysis And Minimization Algorithms.* Springer. (See pp. 43, 48, 50, 51, 54, 60–62, 66, 91, 117, 139, 141, 211).

Horn, R.A., and C.R. Johnson. 2012. *Matrix analysis.* Cambridge university press. (See pp. 67, 68).

Lee, J. M. 2003. *Smooth manifolds.* Springer. (See pp. 57, 58).

Mallat, S. G. 2009. *A wavelet tour of signal processing.* Third. Elsevier/Academic Press, Amsterdam. (See pp. 4, 15, 17, 20).

Meyer, Y. 1992. *Wavelets and operators.* Cambridge Univ Press. (See p. 20).

Nesterov, Y., A. Nemirovskii, and Y. Ye. 1994. *Interior-point polynomial algorithms in convex programming.* SIAM. (See p. 22).

Newton, T.H., and D.G. Potts. 1981. *Radiology of the Skull and Brain: Technical aspects of computed tomography.* Radiology of the Skull and Brain. Mosby. (See p. 8).

Poljak, B. T. 1987. *Introduction to optimization.* Optimization Software. (See p. 10).

Refregier, R., and F. Goudail. 2004. *Statistical Image Processing techniques for Noisy Images - An application Oriented Approach.* Kluwer. (See p. 6).

Rockafellar, R. T., and R. Wets. 1998. *Variational analysis.* Berlin: Springer. (See p. 44).

Rockafellar, R.T. 1996. *Convex analysis.* Princeton University Press. (See pp. 43, 58, 61, 66, 76, 95, 139).

Scherzer, O. 2009. *Variational methods in imaging.* Springer. (See pp. 23, 25, 154, 155, 159).

Starck, J.-L., F. Murtagh, and J. M. Fadili. 2010. *Sparse image and signal processing: wavelets, curvelets, morphological diversity.* Cambridge University Press. (See p. 15).

Tikhonov, A. N., and V. Arsenin. 1977. *Solutions of ill-posed problems.* V. H. Winston / Sons. (See p. 9).

Trevor, H., T. Robert, and J. Friedman. 2009. *The elements of statistical learning.* Second. Springer New York. (See p. 6).

Von Neumann, J. 1961. *Collected works.* Edited by A. H. Taub. Pergamon Press. (See p. 68).

Zalinescu, C. 2002. *Convex analysis in general vector spaces.* World Scientific. (See p. 43).

Articles and Book Chapters

Agard, D. A. 1984. "Optical sectioning microscopy: cellular architecture in three dimensions." *Annual review of biophysics and bioengineering* 13 (1): 191–219. (See p. 7).

Allen, J. 1977. "Short-term spectral analysis, and modification by discrete Fourier Transform." *IEEE Transactions on Acoustics Speech and Signal Processing* 25 (3): 235–238. (See p. 14).

Amelunxen, D., M. Lotz, M. B. McCoy, and J. A. Tropp. 2013. "Living on the edge: A geometric theory of phase transitions in convex optimization." *arXiv preprint arXiv:1303.6672.* (See p. 239).

Bach, F. 2008a. "Consistency of the group Lasso and multiple kernel learning." *The Journal of Machine Learning Research* 9 (Jun): 1179–1225. (See pp. 30, 148, 179).

———. 2008b. "Consistency of trace norm minimization." *The Journal of Machine Learning Research* 9 (Jun): 1019–1048. (See pp. 30, 148, 179).

———. 2010. "Self-concordant analysis for logistic regression." *Electronic Journal of Statistics* 4:384–414. (See p. 188).

Bandeira, A. S., E. Dobriban, D. G. Mixon, and W. F. Sawin. 2013. "Certifying the Restricted Isometry Property is Hard." *IEEE transactions on information theory* 59 (6): 3448–3450. (See pp. 27, 161).

Barron, A., L. Birgé, and P. Massart. 1999. "Risk bounds for model selection via penalization." *Probability theory and related fields* 113 (3): 301–413. (See p. 12).

Beck, A., and M. Teboulle. 2009. "Gradient-based algorithms with applications to signal recovery." *Convex Optimization in Signal Processing and Communications.* (See p. 23).

Biemond, J., R. L. Lagendijk, and R. M. Mersereau. 1990. "Iterative methods for image deblurring." *Proceedings of the IEEE* 78 (5): 856–883. (See p. 7).

Birgé, L., and P. Massart. 1997. "From model selection to adaptive estimation." Chap. 4 in *Festschrift for Lucien Le Cam,* edited by D. Pollard, E. Torgersen, and L. Y. Grace, 55–87. Springer New York. (See p. 12).

———. 2007. "Minimal penalties for Gaussian model selection." *Probability theory and related fields* 138 (1-2): 33–73. (See p. 12).

Bobin, J., J.-L. Starck, and R. Ottensamer. 2008. "Compressed sensing in astronomy." *Selected Topics in Signal Processing, IEEE Journal of* 2 (5): 718–726. (See p. 8).

Boncelet, C. 2005. "Handbook of Image and Video Processing." Chap. Image Noise Models in, edited by A. C. Bovik. Academic Press, Inc. (See p. 6).

Borup, L., R. Gribonval, and M. Nielsen. 2008. "Beyond coherence : recovering structured time-frequency representations." *Applied and Computational Harmonic Analysis* 24 (1): 120–128. (See pp. 31, 180).

Bredies, K., K. Kunisch, and T. Pock. 2010. "Total generalized variation." *SIAM Journal on Imaging Sciences* 3 (3): 492–526. (See p. 21).

Bredies, K., and H. K. Pikkarainen. 2013. "Inverse problems in spaces of measures." *ESAIM: Control, Optimisation and Calculus of Variations* 19 (1): 190–218. (See p. 21).

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. (See p. 188).

Burger, M., and S. Osher. 2004. "Convergence rates of convex variational regularization." *Inverse Problems* 20 (5): 1411. (See pp. 26, 160).

Cai, T. T. 1999. "Adaptive wavelet estimation: a block thresholding and oracle inequality approach." *Annals of statistics* 27 (3): 898–924. (See p. 17).

Cai, T.T., and B.W. Silverman. 2001. "Incorporating information on neighbouring coefficients into wavelet estimation." *Sankhya: The Indian Journal of Statistics, Series B:*127–148. (See p. 18).

Candès, E. J., Y. C. Eldar, D. Needell, and P. Randall. 2011. "Compressed sensing with coherent and redundant dictionaries." *Applied and Computational Harmonic Analysis* 31 (1): 59–73. (See pp. 27, 161).

Candès, E. J., and C. Fernandez-Granda. 2013. "Towards a Mathematical Theory of Super-resolution." *Communications on Pure and Applied Mathematics* 67 (6): 906–956. (See p. 21).

Candès, E. J., and Y. Plan. 2010. "Matrix completion with noise." *Proceedings of the IEEE* 98 (6): 925–936. (See pp. 30, 180).

Candès, E. J., and Y. Plan. 2011a. "A Probabilistic and RIPless Theory of Compressed Sensing." *Information Theory, IEEE Transactions on* 57 (11): 7235–7254. (See pp. 27, 161).

———. 2011b. "Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements." *Information Theory, IEEE Transactions on* 57 (4): 2342–2359. (See pp. 27, 161).

Candès, E. J., and B. Recht. 2013. "Simple bounds for recovering low-complexity models." *Mathematical Programming* 141 (1–2): 577–589. (See pp. 30, 100, 136, 180).

Candès, E. J., J. Romberg, and T. Tao. 2006a. "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information." *IEEE Transactions on Information Theory* 52 (2): 489–509. (See pp. 7, 26, 27, 136, 160, 161).

———. 2006b. "Stable signal recovery from incomplete and inaccurate measurements." *Communications on pure and applied mathematics* 59 (8): 1207–1223. (See pp. 26, 27, 160, 161).

Candès, E. J., and T. Tao. 2005. "Decoding by linear programming." *Information Theory, IEEE Transactions on* 51 (12): 4203–4215. (See pp. 27, 161).

———. 2006. "Near-optimal signal recovery from random projections: Universal encoding strategies?" *Information Theory, IEEE Transactions on* 52 (12): 5406–5425. (See pp. 26, 27, 160, 161).

Caselles, V., A. Chambolle, and M. Novaga. 2007. "The Discontinuity Set of Solutions of the TV Denoising Problem and Some Extensions." *Multiscale Modeling and Simulation* 6 (3): 879–894. (See pp. 21, 248).

Chambolle, A., V. Caselles, D. Cremers, M. Novaga, and T. Pock. 2010. "An Introduction to Total Variation for Image Analysis." In *Theoretical Foundations and Numerical Methods for Sparse Recovery.* De Gruyter. (See p. 16).

Chambolle, A., R. A. DeVore, N-Y Lee, and B. J. Lucier. 1998. "Nonlinear wavelet image processing: Variational problems, compression, and noise removal through wavelet shrinkage." *Image Processing, IEEE Transactions on* 7 (3): 319–335. (See p. 20).

Chambolle, A., and T. Pock. 2011. "A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging." *Journal of Mathematical Imaging and Vision* 40 (1): 120–145. (See p. 226).

Chan, C. L., A. K. Katsaggelos, and A. V. Sahakian. 1993. "Image sequence filtering in quantum-limited noise with applications to low-dose fluoroscopy." *Medical Imaging, IEEE Transactions on* 12 (3): 610–621. (See p. 7).

Chandrasekaran, V., B. Recht, P. A. Parrilo, and A. Willsky. 2012. "The Convex Geometry of Linear Inverse Problems." *Foundations of Computational Mathematics* 12 (6): 805–849. (See pp. 142, 239).

Chen, J., and X. Huo. 2006. "Theoretical results on sparse representations of multiple-measurement vectors." *Signal Processing, IEEE Transactions on* 54 (12): 4634–4643. (See p. 17).

Chen, S. S., D. L. Donoho, and M. A. Saunders. 1999. "Atomic decomposition by basis pursuit." *SIAM journal on scientific computing* 20 (1): 33–61. (See p. 15).

Chesneau, C., J. Fadili, and J.-L. Starck. 2010a. "Stein block thresholding for image denoising." *Applied and Computational Harmonic Analysis* 28 (1): 67–88. (See p. 20).

Chesneau, C., J. Fadili, and J-L Starck. 2010b. "Stein block thresholding for wavelet-based image deconvolution." *Electronic Journal of Statistics* 4:415–435. (See p. 20).

Ciak, R., B. Shafei, and G. Steidl. 2012. "Homogeneous penalizers and constraints in convex image restoration." *Journal of Mathematical Imaging and Vision* 47 (3): 210–230. (See p. 10).

Combettes, P. L., and J.-C. Pesquet. 2011. "Proximal splitting methods in signal processing." In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering,* edited by H. H. Bauschke, Burachik R. S., P. L. Combettes, Elser. V., D. R. Luke, and H. Wolkowicz, 185–212. Springer. (See pp. 23, 221).

Combettes, P. L., and V. R. Wajs. 2005. "Signal recovery by proximal forward-backward splitting." *Multiscale Modeling & Simulation* 4 (4): 1168–1200. (See p. 224).

Cotter, S. F., B. D. Rao, J. Engan, and K. Kreutz-Delgado. 2005. "Sparse solutions to linear inverse problems with multiple measurement vectors." *Signal Processing, IEEE Transactions on* 53 (7): 2477–2488. (See p. 17).

Daubechies, I., M. Defrise, and C. De Mol. 2004. "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint." *Communications on Pure and Applied Mathematics* 57 (11): 1413–1457. (See p. 225).

Donoho, D. L. 1995. "Nonlinear solution of linear inverse problems by wavelet–vaguelette decomposition." *Applied and computational harmonic analysis* 2 (2): 101–126. (See p. 20).

———. 2006. "Compressed sensing." *Information Theory, IEEE Transactions on* 52 (4): 1289–1306. (See p. 7).

Donoho, D. L., and X. Huo. 2001. "Uncertainty principles and ideal atomic decomposition." *Information Theory, IEEE Transactions on* 47 (7): 2845–2862. (See pp. 31, 143, 180).

Donoho, D. L., and J. M Johnstone. 1994. "Ideal spatial adaptation by wavelet shrinkage." *Biometrika* 81 (3): 425–455. (See p. 20).

Dossal, C. 2012. "A necessary and sufficient condition for exact sparse recovery by minimization." *Comptes Rendus Mathematique* 350 (1–2): 117–120. (See pp. 31, 180).

Dossal, C., M.-L. Chabanol, G. Peyré, and J. M. Fadili. 2012. "Sharp support recovery from noisy random measurements by ℓ^1-minimization." *Applied and Computational Harmonic Analysis* 33 (1): 24–43. (See pp. 30, 180).

Dossal, C., M. Kachour, M. J. Fadili, G. Peyré, and C. Chesneau. 2013. "The degrees of freedom of the Lasso for general design matrix." *Statistica Sinica* 23:809–828. (See pp. 33, 36, 190, 192, 200, 215).

Eckstein, J., and D. P. Bertsekas. 1992. "On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators." *Mathematical Programming* 55 (1-3): 293–318. (See p. 225).

Efron, B. 1986. "How biased is the apparent error rate of a prediction rule?" *Journal of the American Statistical Association* 81 (394): 461–470. (See pp. 34, 204, 205, 259).

———. 2004. "The estimation of prediction error: Covariance penalties and cross-validation." *Journal of the American Statistical Association* 99 (467): 619–632. (See p. 243).

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani. 2004. "Least angle regression." *The Annals of statistics* 32 (2): 407–451. (See p. 22).

Elad, M., P. Milanfar, and R. Rubinstein. 2007. "Analysis versus synthesis in signal priors." *Inverse problems* 23 (3): 947. (See p. 16).

Elad, M., J.-L. Starck, P. Querre, and D. L. Donoho. 2005. "Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA)." *Applied and Computational Harmonic Analysis* 19 (3): 340–358. (See p. 14).

Eldar, Y. C. 2009. "Generalized SURE for Exponential Families: Applications to Regularization." *Signal Processing, IEEE Transactions on* 57 (2): 471–481. (See pp. 205–207, 214, 217, 218).

Fadili, J. M., and G. Peyré. 2011. "Total variation projection with first order schemes." *Image Processing, IEEE Transactions on* 20 (3): 657–669. (See p. 223).

Fuchs, J.-J. 2004. "On sparse representations in arbitrary redundant bases." *Information Theory, IEEE Transactions on* 50 (6): 1341–1344. (See pp. 30, 80, 148, 179).

Grasmair, M. 2011. "Linear convergence rates for Tikhonov regularization with positively homogeneous functionals." *Inverse Problems* 27 (7): 075014. (See pp. 26, 160).

Grasmair, M., O. Scherzer, and M. Haltmeier. 2011. "Necessary and sufficient conditions for linear convergence of l1-regularization." *Communications on Pure and Applied Mathematics* 64 (2): 161–182. (See pp. 26, 27, 160, 161).

Gribonval, R. 2011. "Should penalized least squares regression be interpreted as Maximum A Posteriori estimation?" *Signal Processing, IEEE Transactions on* 59 (5): 2405–2410. (See p. 9).

Gribonval, R., and M. Nielsen. 2008. "Beyond sparsity : recovering structured representations by ℓ^1-minimization and greedy algorithms." *Advances in Computational Mathematics* 28 (1): 23–41. (See pp. 31, 180).

Gribonval, R., H. Rauhut, K. Schnass, and P. Vandergheynst. 2008. "Atoms of all channels, unite! Average case analysis of multi-channel sparse recovery using greedy algorithms." *Journal of Fourier analysis and Applications* 14 (5-6): 655–687. (See p. 17).

Gross, D. 2011. "Recovering Low-Rank Matrices From Few Coefficients in Any Basis." *Information Theory, IEEE Transactions on* 57 (3): 1548–1566. (See pp. 27, 161).

Hadamard, J. 1902. "Sur les problèmes aux dérivées partielles et leur signification physique." *Princeton University Bulletin* 13 (49-52): 28. (See p. 8).

Hall, P., G. Kerkyacharian, and D. Picard. 1998. "Block threshold rules for curve estimation using kernel and wavelet methods." *The Annals of Statistics* 26 (3): 922–942. (See p. 20).

———. 1999. "On the minimax optimality of block thresholded wavelet estimators." *Statistica Sinica* 9 (1): 33–49. (See p. 17).

Hall, P., S. Penev, G. Kerkyacharian, and D. Picard. 1997. "Numerical performance of block thresholded wavelet estimators." *Statistics and Computing* 7 (2): 115–124. (See p. 17).

Hare, W., and A. S. Lewis. 2007. "Identifying Active Manifolds." *Algorithmic Operations Research* 2 (2). (See pp. 114, 227).

Hiriart-Urruty, J.-B., and H. Y. Le. 2012. "Convexifying the set of matrices of bounded rank: applications to the quasiconvexification and convexification of the rank function." *Optimization Letters* 6 (5): 841–849. (See p. 19).

Hofmann, B., B. Kaltenbacher, C. Poeschl, and O. Scherzer. 2007. "A convergence rates result for Tikhonov regularization in Banach spaces with non-smooth operators." *Inverse Problems* 23 (3): 987. (See pp. 26, 134, 160).

Hofmann, K.-H., and G. Weigelt. 1993. "Iterative image reconstruction from the bispectrum." *Astronomy and Astrophysics* 278:328–339. (See p. 4).

Hudson, H. M. 1978. "A natural identity for exponential families with applications in multiparameter estimation." *Annals of Statistics* 6 (3): 473–484. (See p. 205).

Hunt, B. R. 1977. "Bayesian methods in nonlinear digital image restoration." *Computers, IEEE Transactions on* C26 (3): 219–229. (See p. 9).

Hwang, J. T. 1982. "Improving upon standard estimators in discrete exponential families with applications to Poisson and negative binomial cases." *Ann. Statist.* 10 (3): 857–867. (See p. 205).

Jégou, H., M. Douze, and C. Schmid. 2010. "Improving bag-of-features for large scale image search." *International Journal of Computer Vision* 87 (3): 316–336. (See p. 19).

Jenatton, R., J.Y. Audibert, and F. Bach. 2011. "Structured variable selection with sparsity-inducing norms." *The Journal of Machine Learning Research* 12:2777–2824. (See p. 18).

Kato, K. 2009. "On the degrees of freedom in shrinkage estimation." *Journal of Multivariate Analysis* 100 (7): 1338–1352. (See pp. 36, 215, 216).

Koltchinskii, V., K. Lounici, and A. B. Tsybakov. 2011. "Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion." *The Annals of Statistics* 39, no. 5 (October): 2302–2329. (See p. 27).

Lewis, A. S. 1995. "The Convex Analysis of Unitarily Invariant Matrix Functions." *Journal of Convex Analysis* 2 (1/2): 173–183. (See pp. 106, 223).

———. 2002. "Active Sets, Nonsmoothness, and Sensitivity." *SIAM Journal on Optimization* 13 (3): 702–725. (See pp. 28, 29, 112–115, 117, 164, 165, 167–170, 186, 257).

Lewis, A. S., and J. Malick. 2008. "Alternating Projections on Manifolds." *Mathematics of Operations Research* 33 (1): 216–234. (See pp. 58, 169).

Lewis, A. S., and S. Zhang. 2013. "Partial Smoothness, Tilt Stability, and Generalized Hessians." *SIAM Journal on Optimization* 23 (1): 74–94. (See pp. 28, 113).

Lions, P. L., and B. Mercier. 1979. "Splitting Algorithms for the Sum of Two Nonlinear Operators." *SIAM Journal on Numerical Analysis* 16 (6): 964–979. (See p. 225).

Liu, D., X. Sun, F. Wu, S. Li, and Y.-Q. Zhang. 2007. "Image compression with edge-based inpainting." *Circuits and Systems for Video Technology, IEEE Transactions on* 17 (10): 1273–1287. (See p. 7).

Lorenz, D.A. 2008. "Convergence rates and source conditions for Tikhonov regularization with sparsity constraints." *Journal of Inverse and Ill-Posed Problems* 16 (5): 463–478. (See pp. 26, 160).

Lustig, M., D. L. Donoho, and J. M. Pauly. 2007. "Sparse MRI: The application of compressed sensing for rapid MR imaging." *Magnetic resonance in medicine* 58 (6): 1182–1195. (See p. 7).

Lyubarskii, Y., and R. Vershynin. 2010. "Uncertainty principles and vector quantization." *Information Theory, IEEE Transactions on* 56 (7): 3491–3501. (See p. 19).

Mallat, S. G. 1989. "A theory for multiresolution signal decomposition: the wavelet representation." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 11 (7): 674–693. (See pp. 14, 20).

Mallat, S. G., and Z. Zhang. 1993. "Matching pursuits with time-frequency dictionaries." *Signal Processing, IEEE Transactions on* 41 (12): 3397–3415. (See p. 12).

Mallows, C. L. 1973. "Some Comments on C_p." *Technometrics* 15 (4): 661–675. (See p. 11).

Meier, L., S. Van De Geer, and P. Buhlmann. 2008. "The group lasso for logistic regression." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70 (1): 51–71. (See p. 188).

Meyer, M., and M. Woodroofe. 2000. "On the degrees of freedom in shape-restricted regression." *Annals of Statistics* 28 (4): 1083–1104. (See pp. 36, 216).

Mordukhovich, B. S. 1992. "Sensitivity analysis in nonsmooth optimization." In *Theoretical Aspects of Industrial Design,* edited by D. A. Field and V. Komkov, 58:32–46. SIAM Volumes in Applied Mathematics. (See pp. 31, 33, 199).

Moreau, J.-J. 1965. "Proximité et dualité dans un espace hilbertien." *Bulletin de la Société mathématique de France* 93:273–299. (See p. 221).

Nam, S., M. E. Davies, M. Elad, and R. Gribonval. 2013. "The cosparse analysis model and algorithms." *Applied and Computational Harmonic Analysis* 34 (1): 30–56. (See pp. 16, 149, 150).

Natarajan, B. K. 1995. "Sparse approximate solutions to linear systems." *SIAM Journal on Computing* 24 (2): 227–234. (See p. 12).

Negahban, S. N., and M. J. Wainwright. 2011. "Simultaneous support recovery in high dimensions: Benefits and perils of block-regularization." *Information Theory, IEEE Transactions on* 57 (6): 3841–3863. (See p. 18).

Nelder, J. A., and R. Wedderburn. 1972. "Generalized linear models." *Journal of the Royal Statistical Society. Series A (General)* 135 (3): 370–384. (See pp. 10, 203).

Nyquist, H. 1928. "Certain topics in telegraph transmission theory." *American Institute of Electrical Engineers, Transactions of the* 47 (2): 617–644. (See p. 7).

Obozinski, G., B. Taskar, and M. I. Jordan. 2010. "Joint covariate selection and joint subspace selection for multiple classification problems." *Statistics and Computing* 20 (2): 231–252. (See p. 17).

Osborne, M. R., B. Presnell, and B. A. Turlach. 2000. "A new approach to variable selection in least squares problems." *IMA journal of numerical analysis* 20 (3): 389–403. (See p. 22).

Parikh, N., and S. P. Boyd. 2013. "Proximal algorithms." *Foundations and Trends in Optimization* 1 (3): 123–231. (See p. 23).

Pesquet, J.-C., A. Benazza-Benyahia, and C. Chaux. 2009. "A SURE Approach for Digital Signal/Image Deconvolution Problems." *IEEE Transactions on Signal Processing* 57 (12): 4616–4632. (See p. 218).

Raguet, H., M. J. Fadili, and G. Peyré. 2013. "A Generalized Forward-Backward Splitting." *SIAM Journal on Imaging Sciences* 6 (3): 1199–1226. (See p. 220).

Ramani, S., T. Blu, and M. Unser. 2008. "Monte-Carlo SURE: a black-box optimization of regularization parameters for general denoising algorithms." *Image Processing, IEEE Transactions on* 17 (9): 1540–1554. (See p. 243).

Recht, B., M. Fazel, and P. A. Parrilo. 2010. "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization." *SIAM review* 52 (3): 471–501. (See pp. 27, 161).

Resmerita, E. 2005. "Regularization of ill-posed problems in Banach spaces: convergence rates." *Inverse Problems* 21 (4): 1303. (See pp. 26, 160).

Rudelson, M., and R. Vershynin. 2008. "On sparse reconstruction from Fourier and Gaussian measurements." *Communications on Pure and Applied Mathematics* 61 (8): 1025–1045. (See pp. 26, 160).

Rudin, L.I., S. Osher, and E. Fatemi. 1992. "Nonlinear total variation based noise removal algorithms." *Physica D: Nonlinear Phenomena* 60 (1): 259–268. (See p. 16).

Santosa, F., and W.W. Symes. 1986. "Linear inversion of band-limited reflection seismograms." *SIAM Journal on Scientific and Statistical Computing* 7 (4): 1307–1330. (See pp. 7, 15).

Seal, H. L. 1967. "Studies in the History of Probability and Statistics. XV The historical development of the Gauss linear model." *Biometrika* 54 (1-2): 1–24. (See p. 5).

Shannon, C. 1948. "A Mathematical Theory of Communication." *Bell System Technical Journal* 27:379–423, 623–656. (See p. 7).

Shen, X., and J. Ye. 2002. "Adaptive model selection." *Journal of the American Statistical Association* 97 (457): 210–221. (See p. 243).

Starck, J.-L., E. Pantin, and F. Murtagh. 2002. "Deconvolution in astronomy: A review." *Publications of the Astronomical Society of the Pacific* 114:1051–1069. (See p. 7).

Steidl, G., J. Weickert, T. Brox, P. Mrázek, and M. Welk. 2004. "On the equivalence of soft wavelet shrinkage, total variation diffusion, total variation regularization, and SIDEs." *SIAM Journal on Numerical Analysis* 42 (2): 686–713. (See p. 16).

Stein, C.M. 1981. "Estimation of the Mean of a Multivariate Normal Distribution." *The Annals of Statistics* 9 (6): 1135–1151. (See pp. 34, 35, 205, 206, 213).

Tibshirani, R. 1996. "Regression shrinkage and selection via the Lasso." *Journal of the Royal Statistical Society. Series B. Methodological* 58 (1): 267–288. (See p. 15).

Tibshirani, R. J., and J. Taylor. 2011. "The solution path of the generalized lasso." *The Annals of Statistics* 39 (3): 1335–1371. (See p. 22).

———. 2012. "Degrees of freedom in lasso problems." *The Annals of Statistics* 40 (2): 1198–1232. (See pp. 33, 34, 36, 200, 215).

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Knight. 2005. "Sparsity and smoothness via the fused Lasso." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (1): 91–108. (See pp. 17, 234).

Tropp, J.A. 2006. "Just relax: Convex programming methods for identifying sparse signals in noise." *Information Theory, IEEE Transactions on* 52 (3): 1030–1051. (See pp. 31, 180).

Tur, R., Y. C. Eldar, and Z. Friedman. 2011. "Innovation rate sampling of pulse streams with application to ultrasound imaging." *Signal Processing, IEEE Transactions on* 59 (4): 1827–1842. (See p. 7).

Turlach, B. A., W. N. Venables, and S. J. Wright. 2005. "Simultaneous variable selection." *Technometrics* 47 (3): 349–363. (See p. 18).

Vaiter, S., C.-A. Deledalle, G. Peyré, C. Dossal, and J. Fadili. 2013. "Local behavior of sparse analysis regularization: Applications to risk estimation." *Applied and Computational Harmonic Analysis* 35 (3): 433–451. (See pp. 34, 36, 189, 192, 193, 200, 215, 240).

Vaiter, S., G. Peyré, C. Dossal, and M. J. Fadili. 2013. "Robust Sparse Analysis Regularization." *IEEE Transactions on Information Theory* 59 (4): 2001–2016. (See pp. 30, 148, 179).

Van de Geer, S. 2008. "High-dimensional generalized linear models and the lasso." *Annals of Statistics* 36 (2): 614–645. (See p. 188).

Wainwright, M. J. 2009. "Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ^1-constrained quadratic programming (lasso)." *IEEE Transactions on Information Theory* 55 (5): 2183–2202. (See pp. 30, 180).

Watson, G.A. 1992. "Characterization of the subdifferential of some matrix norms." *Linear Algebra and its Applications* 170:33–45. (See p. 106).

Wright, S. J. 1993. "Identifiable Surfaces in Constrained Optimization." *SIAM Jounal Control Optimisation* (Philadelphia, PA, USA) 31 (4): 1063–1079. (See pp. 28, 112).

Ye, J. 1998. "On measuring and correcting the effects of data mining and model selection." *Journal of the American Statistical Association:*120–131. (See p. 243).

Yuan, M., and Y. Lin. 2005. "Model selection and estimation in regression with grouped variables." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68 (1): 49–67. (See pp. 17, 36, 215, 216).

Zhao, P., G. Rocha, and B. Yu. 2009. "The composite absolute penalties family for grouped and hierarchical variable selection." *The Annals of Statistics* 37 (6A): 3468–3497. (See p. 18).

Zou, H., and T. Hastie. 2005. "Regularization and variable selection via the elastic net." *J. Royal Statistical Soc. B* 67 (2): 301–320. (See p. 85).

Zou, H., T. Hastie, and R. Tibshirani. 2007. "On the "degrees of freedom" of the lasso." *The Annals of Statistics* 35 (5): 2173–2192. (See pp. 36, 215).

Proceedings

Akaike, H. 1973. "Information theory and an extension of the maximum likelihood principle." In *Second international symposium on information theory,* 267–281. Springer Verlag. (See p. 11).

Fazel, M., H. Hindi, and S. P. Boyd. 2001. "A rank minimization heuristic with application to minimum order system approximation." In *American Control Conference, 2001. Proceedings of the 2001,* 6:4734–4739. IEEE. (See p. 19).

Fuchs, J.-J. 2011. "Spread representations." In *Signals, Systems and Computers (ASILOMAR),* 814–817. IEEE. (See pp. 19, 22).

Jégou, H., T. Furon, and J.-J. Fuchs. 2012. "Anti-sparse coding for approximate nearest neighbor search." In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on,* 2029–2032. IEEE. (See pp. 19, 236).

Kakade, S. M., O. Shamir, K. Sridharan, and A. Tewari. 2010. "Learning exponential families in high-dimensions: Strong convexity and sparsity." In *JMLR Workshop and Conference Proceedings,* 9:381–388. (See p. 188).

Liu, H., and J. Zhang. 2009. "Estimation Consistency of the Group Lasso and its Applications." In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS).* (See p. 190).

Needell, D., J. Tropp, and R. Vershynin. 2008. "Greedy signal recovery review." In *Signals, Systems and Computers, Conference on,* 1048–1050. IEEE. (See p. 12).

Negahban, S., P. Ravikumar, M.J. Wainwright, and B. Yu. 2009. "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers." In *Proc. NIPS,* 1348–1356. (See p. 100).

Pati, Y. C., R. Rezaiifar, and P. S. Krishnaprasad. 1993. "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition." In *Signals, Systems and Computers, Conference on,* 40–44. IEEE. (See p. 12).

Peyré, G., J. Fadili, and C. Chesneau. 2011. "Adaptive Structured Block Sparsity Via Dyadic Partitioning." In *Proc. EUSIPCO 2011,* 1455–1459. (See p. 18).

Sapiro, G. 2005. "Inpainting the colors." In *Image Processing, IEEE International Conference on,* 2:II–698. IEEE. (See p. 7).

Selesnick, I. W., and M. A. T. Figueiredo. 2009. "Signal restoration with overcomplete wavelet transforms: comparison of analysis and synthesis priors." In *Proceedings of SPIE,* 7446:74460D. (See p. 16).

Solo, V. 1996. "A SURE-fired way to choose smoothing parameters in ill-conditioned inverse problems." In *IEEE Int. Conf. Image Process. (ICIP),* 3:89–92. IEEE. (See p. 243).

Solo, V., and M. Ulfarsson. 2010. "Threshold selection for group sparsity." In *Acoustics Speech and Signal Processing (ICASSP), IEEE International Conference on,* 3754–3757. IEEE. (See pp. 36, 216).

Studer, C., W. Yin, and R. G. Baraniuk. 2012. "Signal Representations with Minimum ℓ_∞-Norm." In *Communication, Control, and Computing, Proc. 50th Ann. Allerton Conf. on.* (See p. 19).

Tang, G., B. N. Bhaskar, P. Shah, and B. Recht. 2012. "Compressive sensing off the grid." In *Communication, Control, and Computing (Allerton), Annual Allerton Conference on,* 778–785. IEEE. (See p. 21).

Vogt, J. E., and V. Roth. 2012. "A Complete Analysis of the $l_{1,p}$ Group-Lasso." In *Machine Learning, International Conference on.* (See p. 18).

Vonesch, C., S. Ramani, and M. Unser. 2008. "Recursive risk estimation for non-linear image deconvolution with a wavelet-domain sparsity constraint." In *Image Processing, International Conference on,* 665–668. IEEE. (See p. 243).

Wakin, M. B., J. N. Laska, M. F. Duarte, D. Baron, S. Sarvotham, D. Takhar, K. F. Kelly, and R. G. Baraniuk. 2006. "An architecture for compressive imaging." In *Image Processing, IEEE International Conference on,* 1273–1276. IEEE. (See p. 7).

PhD. Thesis and Technical Reports

Bakin, S. 1999. "Adaptive regression and model selection in data mining problems." PhD diss., Australian National University. (See p. 17).

Candès, E. J., and D. L. Donoho. 2000. *Curvelets: A surprisingly effective non-adaptive representation for objects with edges.* Technical report. DTIC Document. (See p. 14).

Coste, M. 1999. "An Introduction to O-minimal Geometry." Lecture notes, November. (See pp. 69–71, 211).

Daniilidis, A., D. Drusvyatskiy, and A. S. Lewis. 2013. *Orthogonal Invariance and Identifiability.* Technical report. arXiv 1304.1198. (See pp. 114, 115).

Davis, G., S.G. Mallat, and Z. Zhang. 1994. *Adaptive Time-Frequency Approximations with Matching Pursuits.* Technical report. Courant Institute of Mathematical Sciences. (See p. 12).

Deledalle, C., S. Vaiter, G. Peyré, and F. Fadili. 2014. *Stein Unbiased GrAdient estimator of the Risk (SUGAR) for multiple parameter selection.* Technical report. Preprint Hal-00987295. (See p. 243).

Duval, V., and G. Peyré. 2013. *Exact Support Recovery for Sparse Spikes Deconvolution.* Technical report. Preprint hal-00839635. (See pp. 21, 30, 138, 179).

Fazel, M. 2002. "Matrix Rank Minimization with Applications." PhD diss., Stanford University. (See pp. 18, 19).

Ivanov, VK, VV Vasin, and VP Tanana. 1978. *Theory of Ill-Posed Linear Problems and Its Applications.* (See p. 10).

Srebro, N. 2004. "Learning with matrix factorizations." PhD diss., MIT. (See p. 19).