Convergence of the forward-backward algorithm: beyond the worst-case with the help of geometry

Guillaume Garrigos, Lorenzo Rosasco, Silvia Villa

Convergence of the Forward-Backward algorithm: Beyond the worst-case with the help of geometry

Guillaume Garrigos[1], Lorenzo Rosasco[2,3], and Silvia Villa[4]

[1] LPSM, Université de Paris. 75205 Paris CEDEX 13, France.
[2] DIBRIS, Università degli Studi di Genova. Via Dodecaneso 35, 16146, Genova, Italy.
[3] LCSL, Istituto Italiano di Tecnologia and Massachusetts Institute of Technology.
Bldg. 46-5155, 77 Massachusetts Avenue, Cambridge, MA 02139, USA.
[4] Dipartimento di Matematica, Università degli Studi di Genova. Via Dodecaneso 35, 16146, Genova, Italy.

Abstract

We provide a comprehensive study of the convergence of the forward-backward algorithm under suitable geometric conditions, such as conditioning or Łojasiewicz properties. These geometrical notions are usually local by nature, and may fail to describe the fine geometry of objective functions relevant in inverse problems and signal processing, that have a nice behaviour on manifolds, or sets open with respect to a weak topology. Motivated by this observation, we revisit those geometric notions over arbitrary sets. In turn, this allows us to present several new results as well as collect in a unified view a variety of results scattered in the literature. Our contributions include the analysis of infinite dimensional convex minimization problems, showing the first Łojasiewicz inequality for a quadratic function associated to a compact operator, and the derivation of new linear rates for problems arising from inverse problems with low-complexity priors. Our approach allows to establish unexpected connections between geometry and a priori conditions in inverse problems, such as source conditions, or restricted isometry properties.

1 Introduction

Splitting algorithms based on first order descent methods are widely used to solve high dimensional convex optimization problems in signal and image processing [28], compressed sensing [31], and machine learning [84]. Their main advantage is their simplicity and complexity independent of the dimension of the problem. The worst case convergence rates of these methods have been intensively investigated in the last twenty years. The simplest example is the gradient method applied to a smooth convex function, which is known to converge in values as $o(n^{-1})$ [32, 94]. Analogous results are known for the forward-backward splitting algorithm. We refer to these results as *worst case* since no particular assumption is made on the objective function aside from convexity and existence of a solution. Note that these rates are sharp, meaning that there are functions for which these rates are arbitrarily accurate. Clearly such a large class of convex functions allows for functions with wild behaviors around the minimizers [16], behaviors that might hardly appear in practice. It is then natural to ask whether improved rates can be proved under further regularity assumptions.

Contact: G. Garrigos garrigos@lpsm.paris L. Rosasco lrosasco@mit.edu S. Villa silvia.villa@unige.it

Acknowledgements: This material is supported by the Center for Brains, Minds and Machines, funded by NSF STC award CCF-1231216, and the Air Force project FA9550-17-1-0390. L. Rosasco acknowledges the financial support of the Italian Ministry of Education, University and Research FIRB project RBFR12M3AC. S. Villa is supported by the INDAM GNAMPA research project 2017 Algoritmi di ottimizzazione ed equazioni di evoluzione ereditarie.

Previous work on optimization rates with geometry. One classical geometrical assumption is strong convexity, which indeed guarantees linear convergence rates [50, 95]. In practice, strong convexity is often too restrictive, and one would wish to relax it, while retaining fast rates. A relaxation of this condition is given by geometric conditions that, roughly speaking, describe convex functions $f \in \Gamma_0(X)$ that behave like

$$x \mapsto \operatorname{dist}^p(x, \operatorname{argmin} f), \tag{1}$$

for some $p \geq 1$ and on some subset $\Omega \subset X$, which is typically a neighborhood of the minimizers and/or a sub-level set. The intuition behind this kind of assumption required on a neghborhood of the solution is clear: the bigger is p, the more the function is "*flat*" around its minimizers, which in turns means that a gradient descent algorithm will converge slowly. The idea of exploiting geometric conditions to derive convergence rates has a long history dating back to [89, 91], and plenty of similar convergence rates results have been derived under different yet related geometrical properties.

The optimization community focused on several different but related geometric assumptions, namely the p-conditioning, the p-metric subregularity and the p-Łojasiewicz properties (see Section 3 for their definitions). The first[1] result exploting geometry to derive fast convergence rates dates back to Polyak [89, Theorem 4], showing that the gradient method converges linearly (in terms of the values and iterates) when the objective function verifies the 2-Łojasiewicz inequality. Improved convergence rates for first-order descent methods were then obtained in [91], considering notions slightly stronger than p-metric subregularity, and proving finite convergence of the proximal algorithm for $p = 1$, and linear convergence for $p = 2$. These results are improved and extended in [82], analyzing for the first time convergence rates for the iterates of the proximal algorithm using metric subregularity for general $p \in [1, +\infty[$. The results in [82] recover those in [91] (see also [96, 97]), but also derive superlinear rates for $p \in]1, 2[$, and sublinear rates for $p > 2$. Roughly speaking, the results in [82] show that the bigger is p the slower is the algorithm. A related notion, nowadays called the Luo-Tseng error bound condition, has been considered in the seminal paper [81], and implies the linear convergence of several first order methods. Recently, this condition has been shown to be equivalent to 2-conditioning [40, 74]. In the early 90's, some attention was devoted to the study of p-conditioned functions, in particular for $p = 1$ (some authors call this property superlinear conditioning, sharp growth or sharp minima property). In this context, [45, 64, 23] showed that the proximal algorithm terminates after a finite number of iterations. For $p = 1$, Polyak [90, Theorem 7.2.1] obtained the finite termination for the projected gradient method. The 2-conditioning was also used to obtain linear rates for the proximal algorithm in [70]. In [3], it was observed that the p-Łojasiewicz property could be used to derive precise rates for the iterates of the proximal algorithm. The authors obtain finite convergence when $p = 1$, linear rates when $p \in]1, 2]$, and sublinear rates when $p \in]2, +\infty[$. Similar results can be found in [4, 83]. Such convergence rates for the iterates have been extended to the forward-backward algorithm (and its alternating versions) in [18], and similar rates also hold for the convergence of the values in [27, 46]. More recently, various papers focused on conditions equivalent (or stronger) to the 2-conditioning to derive linear rates [67, 75, 41, 78, 40, 61]. Some effort has also been made to show that the Łojasiewicz property and conditioning are equivalent [16, 17], and to relate it to other error bounds appearing in the literature [61]. See also [85] for a refined analysis of linear rates for the projected gradient algorithm under conditions that interpolate between strong convexity and 2-conditioning (see also Subsection 4.3).

A key observation. Our study starts from a basic observation which allows a number of developments. Indeed, motivated by several relevant examples described in Section 5, we require condition (1) to hold on an arbitrary set Ω, which in general is neither a neighborhood of the solution, nor a sublevel set. This extension allows to establish a connection with modeling assumptions considered in different contexts and unveil their role in optimization. As we explain below, modeling assumptions, such as source conditions in inverse problems [42] or the restricted injectivity property in sparse recovery [25], correspond to conditioning assumptions on specific subsets. This ensures global convergence rates for the forward-backward algorithm that are faster compared to those given by a worst case analysis and indeed often observed in practice.

[1] If we discard the "classic" strong convexity assumption.

Geometry and inverse problems. As a first example of the importance of considering arbitrary sets Ω to define geometrical properties, consider linear inverse problems $Ax = y$ for which the operator A is an infinite dimensional compact operator, making the problem severely ill-posed. A common modeling assumption is to suppose that the minimal norm solution of the problem satisfies a *source condition*, which can be seen as a measure of its regularity (see Section 5.1 for a definition). Under this condition, it is shown that the sublinear rate of the gradient algorithm is faster than the worst case one [42]. However, such a behavior cannot be apparently explained in terms of classical geometrical conditions satisfied by the least squares function: indeed, it was shown in [53] that such a least squares function cannot verify any Łojasiewicz inequality (1) in a neighborhood of its minimizers. On the contrary,thanks to the extension of the definition considered in this paper, we show that geometric assumptions are indeed satisfied, but only on specific subsets. More precisely, we show in Theorem 5.9 that the source condition guarantees that the least squares $\|Ax - y\|^2$ is p-Łojasiewicz ($p > 2$) on a dense affine subspace having empty interior. This allows therefore to explain the faster global rates of the gradient algorithm which are typically observed in this context.

As a second example, consider linear inverse problems with a low-complexity prior, such as sparse inverse problems. For these problems, the restricted injectivity condition [25] is a key modeling assumption to guarantee stable recovery: it means that, even if a linear measurement is corrupted by noise, we can hope to reconstruct an approximated solution by solving a regularized optimization problem. In Section 5.2, we show that this assumption implies a 2-conditioning of the problem over a (nonconvex) cone of sparse vectors. Since this set is active, in the sense that it is reached by the algorithm after a finite time, it immediately gives us asymptotic linear rate of the algorithm. For problems with more general low-complexity priors the situation is similar: an active set will be identified by the iterates of the algorithm, and we show that restricted injectivity condition on the tangent cone to this active set induces a 2-conditioning of the problem on this set. Depending on the applications or on the hypothesis made on the problem, this set can be a low-dimensional manifold, or a set with less structure, and can be computed within the partial smoothness framework [54] or the mirror stratification one [43].

Paper contents. Motivated by the estimation problems presented in Section 5, the goal of this paper is to provide a comprehensive study of the convergence rates of the forward-backward algorithm for convex minimization problems satisfying geometric conditions *on arbitrary sets*. We collect in a unified view a variety of results scattered in the literature, and we extend them to this more general setting. In addition, we derive several novel results along the way. The paper is organized as follows.

After reviewing and discussing worst-case convergence results for the forward-backward algorithm in Section 2, we give in Section 3 the definition of different geometric conditions for a proper convex lower semicontinuous function f: p-conditioning, p-metric subregularity, and p-Łojasiewicz property on general subsets $\Omega \subset X$, rather than sublevel sets or open sets, as typically done in the literature. We show that those geometrical notion are equivalent, provided that the set Ω is stable by the semigroup generated by ∂f (see Proposition 3.3). Since establishing p-conditioning of a function may be hard in general, we provide two sum rules for conditioned functions in Theorem 3.15 and Theorem 3.17. The first one establishes that if a strictly convex function remains p-conditioned under linear perturbations, then it is also p-conditioned under convex perturbation. The second one gives conditions under which the sum of two conditioned functions are conditioned. It allows us to show in particular that the ROF model (minimization of the total variation and the Kullback-Leibler divergence) is 2-conditioned on every bounded set.

Section 4 exploits the p-Łojasiewicz property on general sets to study the convergence of the forward-backward algorithm. In Theorem 4.1, we recover and extend results from the literature, getting finite / superlinear / linear / sublinear convergence rates, depending on the value of $p \in [1, +\infty[$ to our more general setting. Along the way, we extend the sharp superlinear rate known for the proximal method to the Forward-Backward algorithm. In addition, our approach allows to derive in a unified setting both nonasymptotic/global and asymptotic/local convergence results, see Corollaries 4.11 and 4.12. We go beyond the classical analysis by introducing a p-Łojasiewicz property with p taking *nonpositive* values. This allows to study convex functions being bounded from below but with no minimizers, a case which has drawn little attention so far, but which can

arise for instance in function approximation [35] or in statistical learning theory [34, Theorem 9] (see also Section 5.1). For such ill-posed problems, we derive new and sharp sublinear rates for the values in Theorem 4.6, interpolating between $o(n^{-1})$ and $o(1)$. We further show in Section 4.3 that the 2-conditioning is essentially equivalent to the linear convergence of the forward-backward algorithm, illustrating the importance of this notion for convergence rate analysis.

In Section 5, we apply the aforementioned results to optimization problems arising from inverse problems, and discuss the interaction between geometry and modeling assumptions. The key results of this section are Theorem 5.9 and Theorem 5.20. Theorem 5.9 establishes that classical source conditions in inverse problems guarantee the Łojasiewicz property on special sets, and therefore give better convergence rates of the gradient method with respect to worst case ones. Theorem 5.20 says that if we have an a priori assumption about the minimizer, which is assumed to belong to a set C, then a restricted injectivity property of the Hessian of the smooth component of the objective function implies that f is 2-conditioned on this set C around the minimizer. This guarantees asymptotic linear rates for forward-backward when combined with Corollary 4.15.

2 The forward-backward algorithm: notation and background

2.1 Notation and basic definitions

We recall a few classic notions and introduce some notation. Throughout the paper X is a Hilbert space. Given $\Omega \subset X$, we note $\operatorname{int}\Omega$ and $\operatorname{cl}\Omega$ its interior and closure. We say that Ω is a cone, if $\Omega =]0,+\infty[\Omega$. We note $\operatorname{cone}(\Omega)$ (resp. $\operatorname{span}(\Omega)$) the smallest cone (resp. linear subspace) in X containing Ω. Let $x \in X$, $\delta \in]0,+\infty[$, and let $\mathbb{B}_X(x,\delta)$ and $\overline{\mathbb{B}}_X(x,\delta)$ denote respectively the open and closed balls of radius δ centered at x. We also use $\mathbb{B}_X$ and $\overline{\mathbb{B}}_X$ to denote $\mathbb{B}_X(0,1)$ and $\overline{\mathbb{B}}_X(0,1)$, and $\mathbb{S}_X$ to denote the unit sphere $\overline{\mathbb{B}}_X \setminus \mathbb{B}_X$. The distance of $x \in X$ from a set $\Omega \subset X$ is $\operatorname{dist}(x,\Omega) = \inf\{\|x-y\| : y \in \Omega\}$, and $\|\Omega\|_-$ stands for $\operatorname{dist}(0,\Omega)$, so, in particular $\|\varnothing\|_- = +\infty$. If Ω is closed and convex, $\operatorname{proj}(x,\Omega)$ is the projection of x onto Ω, and the relative interior and the strong relative interior of Ω are respectively defined as [11, Definition 6.9]: $\operatorname{ri}\Omega = \{x \in \Omega \mid \operatorname{cone}(C-x) = \operatorname{span}(C-x)\}$, $\operatorname{sri} = \{x \in \Omega \mid \operatorname{cone}(C-x) = \operatorname{cl}\operatorname{span}(C-x)\}$. Given a bounded linear operator A between two Hilbert spaces, its *spectrum*, noted $\operatorname{spec}(A)$, is the set of spectral values $\lambda \in \mathbb{R}$ such that $A - \lambda I$ is not boundedly invertible. We also note $\operatorname{spec}^*(A) := \operatorname{spec}(A) \setminus \{0\}$. The set of *singular values* of A, noted $\sigma(A)$, is defined as $\sigma(A) := \sqrt{\operatorname{spec}^*(AA^*)}$, and we note $\sigma_{inf}(A) := \inf\sigma(A)$. Let $\Gamma_0(X)$ be the class of convex, lower semi-continuous, and proper functions from X to $]-\infty,+\infty]$. For $f \in \Gamma_0(X)$ and $x \in X$, $\partial f(x) \subset X$ denotes the (Fenchel) subdifferential of f at x [11, Definition 16.1], and $\operatorname{dom} f$ (resp. $\operatorname{dom}\partial f$) denotes the effective domain of f (resp. of ∂f). Moreover, f^* is the Fenchel conjugate of f, namely $f^*(v) = \sup_{x\in X}\langle x,v\rangle - f(x)$ for all $v \in X$. We introduce the shorthand notation $\operatorname{dom}^* f := \operatorname{dom} f \setminus \operatorname{argmin} f$. We also introduce the following notation for the (strict) sublevel sets of $f \in \Gamma_0(X)$: for every $r \in]-\infty,+\infty]$, $[f < r] := \{x \in X \mid f(x) < r\}$.

The following assumption will be made throughout this paper.

Assumption 2.1. *Let X be a Hilbert space, $g \in \Gamma_0(X)$, and $h\colon X \to \mathbb{R}$ be differentiable and convex, with L-Lipschitz continuous gradient for some $L \in]0,+\infty[$ and set $f = g + h$.*

Splitting methods, such as the forward-backward algorithm, are extremely popular for minimizing an objective function as in Assumption 2.1. To have an implementable procedure, we implicitly assume that the proximal operator of g can be easily computed (see e.g. [28]):

$$(\forall \lambda > 0)(\forall x \in X) \quad \operatorname{prox}_{\lambda g}(x) = \operatorname*{argmin}_{u\in X} \left\{ g(u) + \frac{1}{2\lambda}\|u-x\|^2 \right\}. \tag{2}$$

Remembering Assumption 2.1 is in force, we introduce the Forward-Backward (FB) map for $\lambda \in]0, 2L^{-1}[$:

$$T_\lambda : x \in X \longmapsto T_\lambda x := \operatorname{prox}_{\lambda g}(x - \lambda\nabla h(x)) \in X, \tag{3}$$

so that the FB algorithm can be simply written as $x_{n+1} = T_\lambda x_n$.

2.2 The Forward-Backward algorithm: worst-case analysis

The following theorem collects known results about the convergence of the FB algorithm. This is a "worst-case" analysis, in the sense that it holds for every $f \in \Gamma_0(X)$ satisfying Assumption 2.1. The main goal of Section 4 is to show how these results can be improved taking into account the geometry of f at its infimum.

Theorem 2.2 (Forward-Backward - convex case)**.** *Suppose that Assumption 2.1 is in force, and let $(x_n)_{n\in\mathbb{N}}$ be generated by the FB algorithm with $\lambda \in]0, 2L^{-1}[$. Then:*

i) (Descent property) The sequence $(f(x_n))_{n\in\mathbb{N}}$ is decreasing, and converges to $\inf f$.

ii) (Féjer property) For all $\bar{x} \in \operatorname{argmin} f$, the sequence $(\|x_n - \bar{x}\|)_{n\in\mathbb{N}}$ is decreasing.

iii) (Boundedness) The sequence $(x_n)_{n\in\mathbb{N}}$ is bounded if and only if $\operatorname{argmin} f$ *is nonempty.*

Suppose in addition that f is bounded from below. Then

iv) (Subgradients convergence) The sequence $(\|\partial f(x_n)\|_-)_{n\in\mathbb{N}}$ converges decreasingly to zero, with $\|\partial f(x_{n+1})\|_-^2 = O\left(f(x_n) - \inf f\right)$.

Moreover, if $\operatorname{argmin} f \neq \emptyset$*, we have:*

v) (Weak convergence) The sequence $(x_n)_{n\in\mathbb{N}}$ converges weakly to a minimizer of f.

vi) (Global rates for function values) For all $n \in \mathbb{N}$,

$$f(x_n) - \inf f \leq C\frac{\operatorname{dist}(x_0, \operatorname{argmin} f)^2}{2\lambda n}, \text{ with } C = \begin{cases} 1 & \text{if } \lambda \leq L^{-1}, \\ 1 + 2(\lambda L - 1)(2 - \lambda L)^{-1} & \text{otherwise.} \end{cases}$$

vii) (Asymptotic rates for function values) When $n \to +\infty$, $f(x_n) - \inf f = o\left(n^{-1}\right)$.

Theorem 2.2 collects various convergence results on the FB algorithm. Item i) appears in [94, Theorem 3.22] (see also [52]). Item ii) is a consequence of the nonexpansiveness of the FB map (see (3)) [65, Lemma 3.2]. Item iii), which is a consequence of Opial's Lemma [87, Lem. 5.2], can be found in [94, Theorem 3.12]. Item iv) follows from Lemma A.9.ii) in the Annex. Item v) is also a consequence of Opial's Lemma, see [65, Proposition 3.1]. Items vi) and vii) are proved in [32, Theorem 3] (see also [20, Proposition 2] and [12, Theorem 3.1]).

Remark 2.3 (Sharpness of the results in the worst-case)**.** The convergence results in Theorem 2.2 are sharp, in the following sense. First, the iterates may not converge strongly: see [8, 52] for a counterexample in $\Gamma_0(\ell^2(\mathbb{N}))$. Even in finite dimension, no sublinear rates should be expected for the iterates. To see this, apply the proximal algorithm to the function $x \in \mathbb{R} \mapsto f_p(x) = |x|^p$, whose unique minimizer is zero. When $p \in]2, +\infty[$, there exists a constant $C_p > 0$ depending on $(\|x_0\|, \lambda, p)$ such that (see e.g. the discussion following [83, Proposition 2.5], or Lemma A.1):

$$(\forall n \geq 1) \quad |x_n| \geq C_p n^{-1/(p-2)}, \quad \text{where } \lim_{p\to+\infty} \frac{1}{p-2} = 0. \tag{4}$$

The estimate (4) also provides a lower bound for the rates on the objective values:

$$f_p(x_n) - \inf f_p \geq C_p^p n^{-p/(p-2)}. \tag{5}$$

The above lower bounds imply that the rate in Theorem vii) cannot be improved into a rate $O(n^{-\delta})$, for some $\delta > 1$, because we can always find a p large enough verifying $p/(p-2) > \delta$. It also means that no polynomial rates can be expected for $\|x^n - \bar{x}\|$. This fact was also observed in [32, Theorem 12] on an infinite dimensional counterexample. When f is bounded from below, but has no minimizers, the values $f(x_n) - \inf f$ go to zero but no rates can be obtained in general. To see this, consider for any $\alpha > 0$ the function $f_\alpha \in \Gamma_0(\mathbb{R})$ defined by

$$f_\alpha : \mathbb{R} \to]-\infty, +\infty] \ : \ f_\alpha(x) = |x|^{-\alpha} \text{ if } x < 0, \ +\infty \text{ otherwise.} \tag{6}$$

If $(x_n)_{n\in\mathbb{N}}$ is obtained by applying the proximal algorithm to this function, then (see Lemma A.1) there exists $C_\alpha > 0$ such that:

$$f_\alpha(x_n) - \inf f_\alpha \geq C_\alpha^{-\alpha} n^{-\alpha/(2+\alpha)}, \text{ where } \lim_{\alpha\to 0} \frac{\alpha}{2+\alpha} = 0 \text{ and } \lim_{\alpha\to+\infty} \frac{\alpha}{2+\alpha} = 1. \tag{7}$$

Observe that this lower bound on the objective function values implies that the convergence for those functions is slower than the usual $O(n^{-1})$ rate obtained in Theorem 2.2.vi). It also shows that no polynomial rates can be proven for the values when $\operatorname{argmin} f = \emptyset$.

3 Identifying the geometry of a function

3.1 Definitions

In this section we introduce the main geometrical concepts that will be used throughout the paper to derive precise rates for the FB method. Roughly speaking, these notions characterize functions which behave like (1) on an arbitrary set $\Omega \subset X$.

Definition 3.1. Let $p \in [1, +\infty[$, let $f \in \Gamma_0(X)$ with $\operatorname{argmin} f \neq \emptyset$, and $\Omega \subset X$. We say that:

i) f is *p-conditioned* on Ω if there exists a constant $\gamma_{f,\Omega} > 0$ such that:

$$\forall x \in \Omega \cap \operatorname{dom} f, \quad \frac{\gamma_{f,\Omega}}{p} \operatorname{dist}(x, \operatorname{argmin} f)^p \leq f(x) - \inf f.$$

ii) ∂f is *p-metrically subregular* on Ω if there exists a constant $\gamma_{\partial f,\Omega} > 0$ such that:

$$\forall x \in \Omega \cap \operatorname{dom}^* f, \quad \gamma_{\partial f,\Omega} \operatorname{dist}(x, \operatorname{argmin} f)^{p-1} \leq \|\partial f(x)\|_-.$$

iii) f is *p-Łojasiewicz* on Ω if there exists a constant $c_{f,\Omega} > 0$ such that:

$$\forall x \in \Omega \cap \operatorname{dom}^* f, \quad (f(x) - \inf f)^{1-\frac{1}{p}} \leq c_{f,\Omega} \|\partial f(x)\|_-.$$

We will refer to these notions as *global* if $\Omega = X$, and as *local* if $\Omega = \mathbb{B}_X(\bar{x}; \delta) \cap [f < r]$ for some $\bar{x} \in \operatorname{argmin} f$, and $\delta \in]0, +\infty]$, $r \in]\inf f, +\infty]$.

The notion of conditioning, introduced in [98, 105], is a common tool in the optimization and regularization literature [6, 86, 66, 101, 17]. It is also called the *growth condition* [86], and it is strongly related to the notion of Tikhonov wellposedness [38]. The p-metric subregularity coincides with metric subregularity of the subdifferential at the origin, and it is less used, generally defined for $p = 1$ or 2 with Ω equal to a neighborhood of a specific minimizer [36, 67]. It is also called upper Lipschitz continuity at zero of ∂f^{-1} in [29], or inverse calmness [37]. The Łojasiewicz property goes back to [79], and was initially designed as a tool to guarantee the convergence of trajectories for the gradient flow of analytic functions, before its recent use in convex and nonconvex optimization. It is generally presented with a constant $\theta \in [0, 1]$ which is equal, in our notation, to $1 - 1/p$ [79, 1, 14, 17], or $1/p$ [83, 53, 46]. In the remark below we explain the main difference between our definition and the one usually considered in the literature.

Remark 3.2. There is a subtle but crucial difference in the terminology used in Definition 3.1 with respect to the one commonly used for the Łojasiewicz property. It is usually said that a function has the Łojasiewicz property at $\bar{x}$ if there exist $\delta > 0$, $c > 0$, and $r > \inf f$ such that $f(x) - f(\bar{x}) \leq c\|\partial f(x)\|_-$ holds on $\Omega = \mathbb{B}_X(\bar{x}; \delta) \cap [f < r]$. If the latter property holds for every $\bar{x} \in S \subset X$, the function is said to have the Łojasiewicz property on S. This is a different requirement with respect to the one in Definition 3.1. Indeed, we require the inequality to hold uniformly on Ω, while the above definition must hold locally around every point of interest in a given set, and typically only allows for asymptotic convergence rates (see Corollary 4.12). This change of viewpoint is motivated by the fact that for many *convex* functions, we have more than just a local information about the geometry

(see Sections 3.3 and 4). More importantly, it is actually necessary for the analysis of the problems discussed in Section 5, which motivated this paper. Beyond that, it also allows to understand in a unified framework both global (Corollary 4.11) and local (Corollary 4.12) convergence rates.

The notions introduced in Definition 3.1 are closely related to each other. Indeed, for convex functions, p-conditioning implies metric subregularity, which implies the Łojasiewicz property. Under some additional assumptions, it is possible to show that the reverse implications hold. For instance, metric subregularity implies conditioning when $\Omega = \operatorname{argmin} f + \delta\mathbb{B}_X$, $\delta > 0$ [102, Theorem 4.3]. Similar results can also be found in [2, 7, 41, 39], and [29, Theorem 5.2] (for $\Omega = X$). Also, it is shown in [17, Theorem 5] that the local Łojasiewicz property implies local conditioning. The next result, proved in Annex A.2, extends the mentioned ones, and states the equivalence between conditioning, metric subregularity, and Łojasiewicz property on ∂f-invariant sets (see Definition A.2 in Annex A.2).

Proposition 3.3. Let $p \in [1, +\infty[$, let $\Omega \subset X$, and let $f \in \Gamma_0(X)$ be such that $\operatorname{argmin} f \neq \emptyset$. Consider the following properties:

i) f is p-conditioned on Ω ,

ii) ∂f is p-metrically subregular on Ω,

iii) f is p-Łojasiewicz on Ω.

Then i) $\implies$ ii) $\implies$ iii). One can respectively take $\gamma_{\partial f,\Omega} = \gamma_{f,\Omega}/p$ and $c_{f,\Omega} = \gamma_{\partial f,\Omega}^{-1/p}$. Assuming in addition that Ω is ∂f-invariant, we also have iii) $\implies$ i) with $\gamma_{f,\Omega} = c_{f,\Omega}^{-p} p^{1-p}$.

The two next propositions show that these geometric notions are stronger when p is smaller, and are meaningful only on sets containing minimizers (their proof follow directly from Definition 3.1 and are left to the reader).

Proposition 3.4. Let $f \in \Gamma_0(X)$ be such that $\operatorname{argmin} f \neq \emptyset$, $\Omega \subset X$, and $p' \geq p \geq 1$.

i) If f is p-conditioned (resp. ∂f is p-metrically subregular) on Ω, then f is p'-conditioned (resp. ∂f is p'-metrically subregular) on $\Omega \cap \delta\mathbb{B}_X$ for any $\delta \in]0, +\infty[$.

ii) If f is p-Łojasiewicz on Ω, then f is p'-Łojasiewicz on $\Omega \cap [f < r]$ for any $r > \inf f$.

Proposition 3.5. Let $f \in \Gamma_0(X)$ be such that $\operatorname{argmin} f \neq \emptyset$. If $\Omega \subset X$ is a weakly compact set for which $\Omega \cap \operatorname{argmin} f = \emptyset$, then f is p-conditioned on Ω for any $p \in [1, +\infty[$.

3.2 Examples

In this section, we collect some relevant examples.

Example 3.6 (Uniformly convex functions)**.** Suppose that $f \in \Gamma_0(X)$ is uniformly convex of order $p \in [2, +\infty[$ [11, Definition 10.7]. Then, there exists $\gamma > 0$ such that [101, Corollary 3.5.11.iv]:

$$(\forall (x_1, x_2) \in \operatorname{dom} \partial f^2)(\forall x_1^* \in \partial f(x_1)) \quad f(x_2) - f(x_1) - \langle x_1^*, x_2 - x_1 \rangle \geq \frac{\gamma}{p}\|x_2 - x_1\|^p.$$

Such function is globally p-conditioned, with $\gamma_{f,X} = \gamma$, and globally p-Łojasiewicz, with $c_{f,X} = (1 - 1/p)^{1-1/p}\gamma^{-1/p}$ (see Lemma A.4). In the strongly convex case, when $p = 2$, the 2-Łojasiewicz inequality holds with the constant $c_{f,X} = 1/\sqrt{2\gamma}$, which is sharp. Examples of uniformly convex functions of order p are $x \mapsto \|x\|^p$ [11, Example 10.16].

Example 3.7 (Least squares)**.** Let $A : X \to Y$ be a nonzero bounded linear operator between Hilbert spaces, and $f(x) = (1/2)\|Ax - y\|^2$, for some $y \in Y$. Then, the conditioning, metric subregularity, and Łojasiewicz properties, with $p = 2$ and $\Omega = X$, are equivalent to verify on $\operatorname{Ker} A^\perp$, respectively:

$$\gamma_{f,X}\|x\|^2 \leq \langle A^*Ax, x\rangle, \quad \gamma_{\partial f,X}\|x\| \leq \|A^*Ax\|, \quad \text{and} \quad \langle A^*Ax, x\rangle \leq 2c_{f,X}^2\|A^*Ax\|^2.$$

If $\sigma_{\inf}(A^*A) > 0$ holds, one can see that the above inequalities hold with

$$\gamma_{f,X} = \gamma_{\partial f,X} = 1/(2c_{f,X}^2) = \sigma_{\inf}(A^*A),$$

meaning in particular that f is globally 2-conditioned. Since $\sigma_{\inf}(A^*A) > 0$ is equivalent for $R(A^*A)$ to be closed (see Proposition 5.2), it is in particular always true when Y has finite dimension. If instead $\sigma_{\inf}(A^*A) = 0$ holds, [53, Theorem 2.1] shows that f cannot satisfy any local p-Łojasiewicz property, for any $p \geq 1$. This is for instance the case for infinite dimensional compact operators. Nevertheless, we will show in Section 5, that the least squares always satisfies a p-Łojasiewicz property on the so-called regularity sets, for any $p > 2$.

Example 3.8 (Convex piecewise polynomials)**.** A convex continuous function $f : \mathbb{R}^N \to \mathbb{R}$ is said to be *convex piecewise polynomial* if $\mathbb{R}^N$ can be partitioned in a finite number of polyhedra $P_1, ..., P_s$ such that for all $i \in \{1, ..., s\}$, the restriction of f to P_i is a convex polynomial, of degree $d_i \in \mathbb{N}$. The degree of f is defined as $\deg(f) := \max\{d_i \mid i \in \{1, ..., s\}\}$. Assume $\deg(f) > 0$. Convex piecewise polynomial functions are conditioned [71, Corollary 3.6]. More precisely, for all $r > \inf f$, f is p-conditioned on its sublevel set $\Omega = [f < r]$, with $p = 1 + (\deg(f) - 1)^N$. In general, the constant $\gamma_{f,\Omega}$ (which depends on r) cannot be explicitly computed. This result implies that polyhedral functions ($\deg(f) = 1$) are 1-conditioned (in agreement with [23, Corollary 3.6]), and that convex piecewise quadratic functions ($\deg(f) = 2$) are 2-conditioned (in agreement with [70, Theorem 2.7]). More generally, convex semi-algebraic functions are locally p-conditioned [15].

Example 3.9 (L1 regularized least squares)**.** Let $f(x) = \alpha\|x\|_1 + (1/2)\|Ax - y\|^2$, for some linear operator $A : \mathbb{R}^N \to \mathbb{R}^M$, $y \in \mathbb{R}^M$ and $\alpha > 0$. As observed in [17, Section 3.2.1], f is convex piecewise polynomial of degree 2, thus it is 2-conditioned on every nonempty level set $\Omega = [f < r]$. The computation of the conditioning constant $\gamma_{f,\Omega}$ is rather difficult. In [17, Lemma 10] an estimate of $\gamma_{f,\Omega}$ is provided, by means of Hoffman's bound [58]. Extensions of this result to the infinite dimensional setting can be found in [49].

Example 3.10 (Regularized problems)**.** Let X be an Euclidean space, $f(x) := g(x) + h(Ax)$, where $A : X \to \mathbb{R}^M$ is a linear operator, $g \in \Gamma_0(X)$, and $h \in \Gamma_0(\mathbb{R}^M)$ is a strongly convex $C^{1,1}$ function, and $\operatorname{argmin} f \neq \emptyset$. Then f is 2-conditioned on any level set $\Omega = [f < r]$, for $r > \inf f$, if

i) $g(x) = \|x\|_p$ with $p \in]1, 2]$, (see [104, Corollary 2]),

ii) $g(x) = \|x\|_p^p$ with $p \in]1, 2]$, (use [40, Theorem 4.2]; the details are left to the reader as an exercise, and can be checked in the Appendix),

iii) $g(x) = \|x\|_*$ is the nuclear norm of the matrix $x \in X$, provided the following qualification condition holds[2] (see [103]): $\exists \bar{x} \in \operatorname{argmin} f$ such that $-A^*\nabla h(A\bar{x}) \in \operatorname{ri} \partial\|\cdot\|_*(\bar{x})$.

iv) g is polyhedral (see [103, Proposition 6]).

Note that in [103, 104], the authors do not prove directly that the functions are 2-conditioned, but that they verify the so-called Luo-Tseng error bound, that is known to be equivalent to 2-conditioning on sublevel sets [40, Corollary 3.6]. Note also that in items ii-iv), the strong convexity and $C^{1,1}$ assumptions on h can be weakened (see [103] and [40, Theorem 4.2]).

Example 3.11 (Distance to an intersection)**.** Let C, D be two closed convex sets in X such that $C \cap D \neq \emptyset$, and for which the intersection is sufficiently regular, i.e. $0 \in \operatorname{sri}(C - D)$. Let $f(\cdot) = \max\{\operatorname{dist}(\cdot, C), \operatorname{dist}(\cdot, D)\}$. Clearly, $f \in \Gamma_0(X)$, and $\operatorname{argmin} f = C \cap D$. Then f is 1-conditioned on bounded sets [10, Theorem 4.3]. Let $p \in [1, +\infty[$. From $\|\cdot\|_\infty \leq \|\cdot\|_p$, it follows that the function $x \mapsto \operatorname{dist}(x, C)^p + \operatorname{dist}(x, D)^p$ is p-conditioned on bounded sets. The regularity condition $0 \in \operatorname{sri}(C - D)$ is not necessary if the two sets are polyhedral, as proved by Hoffman [58].

[2] We mention that this result was originally announced in [60, Theorem 3.1] without the qualification condition, but then corrected in [103, Proposition 12 & following remarks], in which the authors show that such condition is necessary.

Example 3.12 (Minimum of Łojasiewicz functions)**.** If $f = \min_{i=1,\dots,m} f_i$, with $f_i \in \Gamma_0(\mathbb{R}^N)$ being continuous on its domain, and locally p-Łojasiewicz at $\bar{x} \in \operatorname{argmin} f$, then f is locally p-Łojasiewicz at $\bar{x}$ [74, Theorem 3.1]. It is important to notice that this result do not need the f_i's to be convex.

The next section presents new sum rules for conditioned functions.

3.3 A sum rule for p-conditioned functions

Since verifying conditioning directly with the definition can be difficult, it is very useful to establish which basic operations preserve conditioning. In this section we present two new sum rules for conditioned functions in a setting where $f = g + h \circ A$, where g and h are convex and A is a bounded linear operator. Theorem 3.15 states that if g strictly convex and p-conditioned up to linear perturbations then also f is p-conditioned. Theorem 3.17 provides an alternative where the assumption of strict convexity of g is replaced by a stable conditioning assumption on h, which we formalise in the next definition, inspired by the terminology used in [88, 41, 40].

Definition 3.13. Let $f \in \Gamma_0(X)$, $\Omega \subset X$, and $p \in [1, +\infty[$. We say that f is *p-tilt-conditioned* if, for every $u \in X$, the tilted function $f + \langle u, \cdot \rangle$ has no minimizers, or is p-conditioned on Ω.

Note that a similar notion is already present in the literature: if f is p-tilt-conditioned (in our sense) on every compact set, then it is *firmly convex* in the sense of [40, Definition 4.1].

Example 3.14 (Tilt-conditioned functions)**.** Many conditioned functions relevant for inverse problems are also tilt-conditioned:

- The 1-norm $\|\cdot\|_1$, and more generally every polyhedral function, are 1-tilt-conditioned on Euclidean spaces [23, Cor. 3.6].
- Convex piecewise polynomials of degree 2 are 2-tilt-conditioned on their sublevel sets. This is due to Example 3.8 and the fact that this class of functions is stable up to linear perturbations.
- For the same reasons as above, p-uniformly convex functions are p-tilt-conditioned on X, for $p \geq 2$.
- If $KL(x_1; x_2)$ denotes the Kullback-Leibler divergence between two vectors in $]0, +\infty[^N$, then the divergence $KL(x_1; \cdot)$ is 2-tilt-conditioned on bounded sets. This result is new, and its proof can be found in Lemma A.6.
- The nuclear norm is 2-tilt-conditioned on bounded sets [103, Proposition 11].
- See [40, Section 4] for more examples and properties of 2-tilt-conditioned functions on compact sets.

In this first theorem, we show that if a strictly convex function remains conditioned up to linear perturbations, then it is also stable up to *convex* perturbations:

Theorem 3.15 (Sum rule involving a strictly convex tilt-conditioned function)**.** *Let $f = g + h \circ A$, where $g \in \Gamma_0(X)$, let Y be a Hilbert space, $h \in \Gamma_0(Y)$ and $A : X \to Y$ a bounded linear operator. Suppose that* $\operatorname{argmin} f \neq \emptyset$*. Let $\Omega \subset X$, and assume that:*

a) the nondegeneracy condition $0 \in \operatorname{sri}(\operatorname{dom} h - A(\operatorname{dom} g))$ holds,

b) g is strictly convex on its domain,

c) g is p-tilt conditioned on Ω for some $p \in [1, +\infty[$.

Then, f is p-conditioned on Ω. We have $\gamma_{f,\Omega} = \gamma_{\tilde{g},\Omega}$, where $\tilde{g} = g + \langle \cdot, u \rangle$, for some $u \in X$.

Proof. Let $\bar{x} \in \operatorname{argmin} f$; Fermat's rule implies that $0 \in \partial f(\bar{x})$. Using assumption a) with [11, Thm. 16.47], we can write $0 \in \partial g(\bar{x}) + A^*\partial h(A\bar{x})$. Let $\bar{v} \in -\partial h(A\bar{x})$ be such that $0 \in \partial g(\bar{x}) - A^*\bar{v}$, i.e., $\bar{x} \in \partial g^*(A^*\bar{v})$. Let $x \in \Omega \cap \operatorname{dom} f$, and set $\tilde{g} = g - \langle A^*\bar{v}, \cdot \rangle$. Using the fact that linear forms are continuous, we can use again Fermat's rule together with a sum rule [87, Thm. 3.30] to write

$$u \in \operatorname{argmin} \tilde{g} \Leftrightarrow 0 \in \partial(g - \langle A^*\bar{v}, \cdot \rangle)(u) = \partial g(u) - A^*\bar{v} \Leftrightarrow A^*\bar{v} \in \partial g(u) \Leftrightarrow u \in \partial g^*(A^*\bar{v}), \quad (8)$$

meaning that $\operatorname{argmin} \tilde{g} = \partial g^*(A^*\bar{v}) \neq \emptyset$. It follows then from assumption c) that $\tilde{g}$ is p-conditioned on Ω. Moreover, because g is strictly convex, we have $\partial g^*(A^*\bar{v}) = \{\bar{x}\}$ [11, Prop. 16.37.i], and $\operatorname{argmin} f = \{\bar{x}\}$ [11, Cor 11.9]. These facts mean that $\operatorname{argmin} \tilde{g} = \operatorname{argmin} f$. We can now write the conditioning of $\tilde{g}$ evaluated at x, together with the convexity of h (remember that $-\bar{v} \in \partial h(A\bar{x})$):

$$\begin{aligned} g(x) &\geq g(\bar{x}) + \langle A^*\bar{v}, x - \bar{x} \rangle + (\gamma_{\tilde{g},\Omega}/p)\operatorname{dist}^p(x, \operatorname{argmin} f), \\ h(Ax) &\geq h(A\bar{x}) + \langle -\bar{v}, Ax - A\bar{x} \rangle. \end{aligned}$$

Observe that we are allowed to use the conditioning of $\tilde{g}$ at x, because $x \in \operatorname{dom} f \subset \operatorname{dom} g = \operatorname{dom} \tilde{g}$. Summing these two last inequalities gives

$$f(x) - \inf f \geq \frac{\gamma_{f,\Omega}}{p} \operatorname{dist}^p(x, \operatorname{argmin} f),$$

with $\gamma_{f,\Omega} := \gamma_{\tilde{g},\Omega}$, which concludes the proof. □

Remark 3.16 (On the nondegeneracy condition a) of Theorem 3.15)**.** This condition is very mild, and is satisfied under any of the following sufficient conditions (we note $\bar{x}$ a minimizer of f):

- h is continuous at $A\bar{x}$ (see [11, Prop. 16.27 & Prop. 6.19.vii]).
- h has a full domain.
- $\dim Y < +\infty$, $\bar{x} \in \operatorname{qri} \operatorname{dom} g$ and $A\bar{x} \in \operatorname{ri} \operatorname{dom} h$ (see [11, Def. 6.9 & Prop. 6.19.ix]). These inclusions hold for instance if g and h have open domains.

Theorem 3.15 is useful, but proves to be impractical when g is not strictly convex, which typically happens when g corresponds to some low-complexity-inducing regularizer used in inverse problems (ℓ^1 norm, group lasso, nuclear norm, total variation, etc). The next theorem provides a setting for those functions; in exchange for the strict convexity of g, we will require h to also be tilt-conditioned, and to some strong qualification condition to hold.

Theorem 3.17 (Sum rule for tilt-conditioned functions)**.** *Let $f = g + h \circ A$, where $g \in \Gamma_0(X)$, $h \in \Gamma_0(Y)$ and $A : X \to Y$ is a bounded linear operator with closed range. Suppose that* $\operatorname{argmin} f \neq \emptyset$*, and let $\Omega \subset X$. If $\psi \in \Gamma_0(Y)$ denotes the corresponding Fenchel-Rockafellar dual problem $\psi(v) = g^*(A^*v) + h^*(-v)$, and*

a) the nondegeneracy condition $0 \in \operatorname{sri}\,(\operatorname{dom} h - A(\operatorname{dom} g))$ holds,

then $\operatorname{argmin} \psi \neq \emptyset$*. Moreover, if*

b) there is $\bar{v} \in \operatorname{argmin} \psi$ for which the following qualification conditions are satisfied:

$$0 \in \operatorname{sri}\,(\partial g^*(A^*\bar{v}) - A^{-1}\partial h^*(-\bar{v})), \quad (9)$$

$$0 \in \operatorname{sri}\,(R(A) - \partial h^*(-\bar{v})), \quad (10)$$

c) g is p_1-tilt-conditioned on Ω, and h is p_2-tilt-conditioned on $A\Omega$ for some $p_1, p_2 \geq 1$,

then f is p-conditioned on every bounded subset of Ω, with $p := \max\{p_1, p_2\}$.

Proof. The beginning of this proof starts as in the proof of Theorem 3.15: we use the nondegeneracy assumption a) with [11, Thm. 16.47] to get some $\bar{x} \in \operatorname{argmin} f$ and $\bar{v} \in -\partial h(A\bar{x})$ such that $\bar{x} \in$

$\partial g^*(A^*\bar{v})$. So the condition [11, Thm. 19.1.iii] is verified, meaning that strong duality holds (in the sense that $\inf f = -\inf \psi$). This allows to use [11, Cor. 19.2] to obtain

$$\bar{x} \in \operatorname{argmin} f = \partial g^*(A^*\bar{v}) \cap A^{-1}\partial h^*(-\bar{v}). \tag{11}$$

We can use again [11, Cor. 19.2], this time on the dual problem, to also obtain

$$\bar{v} \in \operatorname{argmin} \psi = -\partial h(A\bar{x}) \cap A^{*-1}\partial g(\bar{x}).$$

The above equality allows us to assume, without loss of generality, that $\bar{v}$ is the element of $\operatorname{argmin} \psi$ satisfying b). So, it remains to prove that, for all $\delta > 0$, there exists $\gamma > 0$ such that:

$$(\forall x \in \Omega \cap \delta\mathbb{B}_X \cap \operatorname{dom} f) \quad f(x) - \inf f \geq \gamma \operatorname{dist}^p(x, \partial g^*(A^*\bar{v}) \cap A^{-1}\partial h^*(-\bar{v})). \tag{12}$$

Fix $\delta > 0$, let $x \in \Omega_\delta := \Omega \cap \delta\mathbb{B}_X \cap \operatorname{dom} f$, and set $\tilde{g} = g - \langle A^*\bar{v}, \cdot\rangle$ and $\tilde{h} = h + \langle \bar{v}, \cdot\rangle$. Setting $p = \max\{p_1, p_2\}$, we see from assumption c) and Proposition 3.4 that $\tilde{g}$ and $\tilde{h}$ are p-conditioned on the bounded sets Ω_δ and $A\Omega_\delta$, respectively. Using the same arguments as in (8), we obtain that $\operatorname{argmin} \tilde{g} = \partial g^*(A^*\bar{v}) \ni \bar{x}$ and $\operatorname{argmin} \tilde{h} = \partial h^*(-\bar{v}) \ni A\bar{x}$. Therefore, the conditioning of $\tilde{g}$ (resp. $\tilde{h}$) evaluated at $x \in \operatorname{dom} f \subset \operatorname{dom} g = \operatorname{dom} \tilde{g}$ (resp. $Ax \in A \operatorname{dom} f \subset \operatorname{dom} h = \operatorname{dom} \tilde{h}$) writes as

$$\begin{aligned} g(x) &\geq g(\bar{x}) + \langle A^*\bar{v}, x - \bar{x}\rangle + (\gamma_{\tilde{g},\Omega_\delta}/p)\operatorname{dist}^p(x, \partial g^*(A^*\bar{v})), \\ h(Ax) &\geq h(A\bar{x}) + \langle -\bar{v}, Ax - A\bar{x}\rangle + (\gamma_{\tilde{h},A\Omega_\delta}/p)\operatorname{dist}^p(Ax, \partial h^*(-\bar{v})). \end{aligned}$$

Summing these two last inequalities gives,

$$f(x) - \inf f \geq C_1 \left(\operatorname{dist}^p(x, \partial g^*(A^*\bar{v})) + \operatorname{dist}^p(Ax, \partial h^*(-\bar{v}))\right), \tag{13}$$

with $C_1 = p^{-1}\min\{\gamma_{\tilde{g},\Omega_\delta}, \gamma_{\tilde{h},A\Omega_\delta}\}$. Since $\|\cdot\|_\infty \leq \|\cdot\|_p$ on $\mathbb{R}^2$, we deduce that

$$f(x) - \inf f \geq C_1 \max\left\{\operatorname{dist}(x, \partial g^*(A^*\bar{v})), \operatorname{dist}(Ax, \partial h^*(-\bar{v}))\right\}^p,$$

It remains to lower bound the right hand side by the distance to $\operatorname{argmin} f$. By Example 3.11, thanks to the qualification condition (9) and the fact that Ω_δ is bounded, we derive from (11) that there exists $C_2 > 0$ independent of x such that

$$\operatorname{dist}(x, \operatorname{argmin} f) \leq C_2 \max\{\operatorname{dist}(x, \partial g^*(A^*\bar{v})), \operatorname{dist}(x, A^{-1}\partial h^*(-\bar{v}))\}. \tag{14}$$

Define $y := \operatorname{proj}(Ax, R(A) \cap \partial h^*(-\bar{v})))$, which is well defined since we assumed $R(A)$ to be closed. Let $\phi_y \in \Gamma_0(X)$ be defined by $\phi_y(u) := (1/2)\|Au - y\|^2$. Since $y \in R(A)$, necessarily $\inf \phi_y = 0$, so we deduce from Example 3.7 that

$$(\forall u \in X) \quad \phi_y(u) \geq (\sigma_{\inf}(A^*A)/2)\operatorname{dist}^2(u, \operatorname{argmin} \phi_y). \tag{15}$$

On the one hand, we have $\operatorname{argmin} \phi_y = A^{-1}y \subset A^{-1}\partial h^*(-\bar{v})$. On the other hand, the definition of y implies $\phi_y(x) = (1/2)\operatorname{dist}^2(Ax, R(A) \cap \partial h^*(-\bar{v}))$. Thus, it follows from (15) that

$$\operatorname{dist}(Ax, R(A) \cap \partial h^*(-\bar{v})) \geq \sigma_{\inf}(A) \operatorname{dist}(x, A^{-1}\partial h^*(-\bar{v}))).$$

Since this is true for any $x \in \Omega_\delta$, we can combine it with (14) to get for all $x \in \Omega_\delta$

$$\operatorname{dist}(x, \operatorname{argmin} f) \leq C_3 \max\{\operatorname{dist}(x, \partial g^*(A^*\bar{v})), \operatorname{dist}(Ax, R(A) \cap \partial h^*(-\bar{v}))\}, \tag{16}$$

with $C_3 = C_2 \max\{1, \sigma_{\inf}(A)^{-1}\}$. To end the proof, use the qualification condition (10) with Example 3.11 again to get some $C_4 > 0$ such that for all $x \in \Omega_\delta$,

$$\begin{aligned} \operatorname{dist}(Ax, R(A) \cap \partial h^*(-\bar{v})) &\leq C_4 \max\{\operatorname{dist}(Ax, R(A)), \operatorname{dist}(Ax, \partial h^*(-\bar{v}))\} \\ &= C_4 \operatorname{dist}(Ax, \partial h^*(-\bar{v})). \end{aligned} \tag{17}$$

The above inequality, combined with (16) and (12), concludes the proof. □

Remark 3.18 (On the qualification conditions)**.** When g is not strictly convex, the conclusion of Theorem 3.17 may not hold if the qualification conditions (9) and (10) are removed, as proved in [103, Section 4.4.4] with $g = \|\cdot\|_*$. Let us give some sufficient conditions for (9) and (10) to hold:

- If X and Y have finite dimension, b) is equivalent to

$$0 \in \operatorname{ri} A\partial g^*(A^*\bar{v}) - \operatorname{ri}\partial h^*(-\bar{v}).$$

  To prove this, use [11, Cor. 6.15] and [92, Thm. 6.7] to see that the above condition is equivalent to (9), which implies (10). This condition is for instance satisfied if $0 \in \operatorname{ri}\partial\psi(\bar{v})$ and $0 \in \operatorname{ri}\operatorname{dom} g^* + A^*(\operatorname{ri}\operatorname{dom} h^*)$ (see [11, Thm. 16.47]). Those are the two conditions needed in [40, Theorem 4.2].

- If X and Y have finite dimension and h is strictly convex, then a sufficient condition for b) is $\bar{x} \in \operatorname{ri}\partial g^*(A^*\bar{v})$ [11, Prop. 18.9].

- If X and Y have finite dimension, g is polyhedral and h is strictly convex, then assumption b) is not needed. As pointed out in [40, Cor. 4.3], this is due to the fact that the subdifferentials of h^* and g^* are polyhedral, which allows the use of Hoffman's bound [58] instead of [10, Theorem 4.3] in the proof.

Remark 3.19 (On the closedness of the range)**.** In Theorem 3.15 we assume $R(A)$ to be closed. To see how important this hypothesis is in infinite dimension, take $g = 0$ (which is not strictly convex), $h = \|\cdot\|^2$ and A an operator with a nonclosed range. Then, for this example, the qualification conditions cannot be satisfied. Indeed, even if (9) is automatically satisfied (because $\partial g^*(0) = X$), condition (10) reduces to $0 \in \operatorname{sri} R(A)$, which is equivalent by definition to $R(A) = \operatorname{cl} R(A)$, which is impossible. Worse, even if we could get rid of this qualification condition, and if the conclusion of the theorem were true, we would obtain that $x \mapsto \|Ax\|^2$ is 2-conditioned on bounded sets, which was proven to be impossible in [53, Theorem 2.1] (combine it with Proposition 3.3).

Remark 3.20 (Previous results)**.** Our results can be seen as extensions and refinements of arguments from [40], where the authors introduce the ideas of exploiting the 2-conditioning of tilted functions on compact sets, together with the description of argmin f as an intersection (11). Theorem 3.17 improves on [40, Thm. 4.2] and [40, Cor. 4.3] which require the argmin f to be bounded, and h to be in C^1 with $\operatorname{dom} h = Y$ (we only ask for a compatibility condition which is satisfied if h is continuous at $A\bar{x}$, see Remark 3.16). As far as we know, Theorem 3.15 is the first sum rule of this kind with such weak assumptions on g.

To illustrate the interest of these sum rules, we provide a new result for regularized inverse problems where the loss function is the Kullback-Leibler divergence, and the regularizer is a polyhedral function, such as the ℓ^1 norm, or the Total Variation, which are commonly used in the signal and image processing literature.

Proposition 3.21. Let $f(x) = g(x) + KL(y; Ax)$, where $g \in \Gamma_0(\mathbb{R}^N)$ is polyhedral, $A \in \mathcal{M}_{M,N}(\mathbb{R})$, and $y \in]0, +\infty[^M$. If $\operatorname{argmin} f \neq \emptyset$, then f is 2-conditioned on bounded sets.

Proof. We just have to verify the hypotheses of Theorem 3.17, by noting $h := KL(y; \cdot)$. First, the nondegeneracy condition a) is verified because $\operatorname{dom} h$ is open, and h is continuous on its domain (see Remark 3.16). Second, the qualification conditions b) are not needed because we are in a finite dimensional setting, g is polyhedral and h is strictly convex (see Remark 3.18). Finally, g being polyhedral implies that it is globally 1-tilt-conditioned (see [23, Corollary 3.6]), and we prove in Lemma A.6 that h is 2-tilt-conditioned on bounded sets, so c) is verified. □

4 Sharp convergence rates for the Forward-Backward algorithm

In this section, we present sharp convergence results for the forward-backward algorithm applied to p-Łojasiewicz functions on a subset Ω, building on the ideas in [5]. We extend the analysis to the case

where Ω is an arbitrary set, which will allow us to deal with infinite dimensional inverse problems (see Section 5.1), or structured problems for which all the information is encoded in a manifold (see Section 5.2). We also provide explicit rates of convergence, for both the iterates and the values. The proofs of Section 4.1 are left in the Annex A.3.

4.1 Refined analysis with p-Łojasiewicz functions

Theorem 4.1 (Strong convergence and rates, $p \geq 1$)**.** *Suppose that Assumption 2.1 is in force, and that f is bounded from below. Let $(x_n)_{n\in\mathbb{N}}$ be generated by the FB algorithm. Assume that:*

a) (Localization) *for all $n \in \mathbb{N}$, $x_n \in \Omega \subset X$,*

b) (Geometry) *f is p-Łojasiewicz on Ω, for some $p \geq 1$.*

Then the sequence $(x_n)_{n\in\mathbb{N}}$ has finite length in X, meaning that $\sum_{n\in\mathbb{N}} \|x_{n+1} - x_n\| < +\infty$, and converges strongly to some $x_\infty \in \operatorname{argmin} f \neq \emptyset$. Moreover, there exist some constants $C_p, C'_p > 0$ with explicit expressions (see equations (53) and (55)), such that the following convergence rates hold, depending on the value of p, and of $\kappa := \lambda(2-\lambda L)[2c^2_{f,\Omega}]^{-1}$:

i) If $p = 1$, then $x_n = x_\infty$ for every $n \geq (f(x_0) - \inf f)/\kappa$.

ii) If $p \in]1, 2[$, the convergence is superlinear: for all $n \in \mathbb{N}$,

$$f(x_{n+1}) - \inf f \leq \left(\frac{f(x_n) - \inf f}{\kappa}\right)^{\frac{p}{2(p-1)}} \quad \text{and} \quad \|x_{n+1} - x_\infty\| \leq C_p (f(x_n) - \inf f)^{1/2},$$

iii) If $p = 2$, the convergence is linear: for all $n \in \mathbb{N}$,

$$f(x_{n+1}) - \inf f \leq \frac{1}{1+\kappa}(f(x_n) - \inf f) \text{ and } \|x_{n+1} - x_\infty\| \leq C_2 (f(x_0) - \inf f)^{1/2} (1+\kappa)^{-n/2}.$$

iv) If $p \in]2, +\infty[$, the convergence is sublinear: for all $n \in \mathbb{N}$,

$$f(x_n) - \inf f \leq (C'_p)^{p/(p-2)} n^{-\frac{p}{p-2}} \quad \text{and} \quad \|x_{n+1} - x_\infty\| \leq C_p (C'_p)^{1/(p-2)} n^{-\frac{1}{p-2}}.$$

Note that the rates range from the finite termination, for $p = 1$, to the worst-case rates seen in Theorem 2.2, when p tends to $+\infty$. The bigger is p, the more the function is ill-conditioned, in the sense that the rates of its values become closer to $o(n^{-1})$, and the rates of its iterates become arbitrarily slow.

Remark 4.2 (Related work)**.** Theorem 4.1 collects known and new results. We present a simple proof of this theorem, focusing on the analysis of a real sequence satisfying (51) (see [27, Theorem 3.2] or [46, Theorem 3.4] for previous results). The superlinear rates in ii), which were known for the proximal point algorithm [82], are new for the Forward-Backward algorithm. Moreover, the case $p = 2$ was giving R-linear rates for the values in [27, 46], while we prove here Q-linear rates. Also, the quantification of the number of steps in the case $p = 1$ involving κ is new.

Remark 4.3 (On the sharpness of the rates I)**.** Let $f = \|\cdot\|^p$. According to (4) and (5), the order of the sublinear rates for the forward-backward algorithm that we obtain for both iterates and values are sharp when $p \in]2, +\infty[$, see Remark 2.3. When $p = 2$, we see that the proximal algorithm verifies $x_{n+1} = (1+2\lambda)^{-1} x_n$, and the algorithm converges linearly. Finally, when $p \in]1, 2[$, the order of superlinearity that we obtain is not sharp, since for this function the proximal algorithm has a Q-superlinear rate of order $(p-1)^{-1}$. It is shown in [82, Theorem 3.1] that $\operatorname{dist}(x_n, \operatorname{argmin} f)$ converges with this order for the proximal algorithm. For this, the author uses the stronger notion of metric subregularity, and we will extend this result in Theorem 4.21 to the FB algorithm.

Remark 4.4 (Best stepsize and condition number)**.** When $p \in [1,2]$, we directly see that the bigger is κ, the better are the constants in the rates for the values. This is true also for $p > 2$, by looking in the proof of Theorem 4.1 to the definition of the constant C'_p. The constant κ is maximal when we take $\lambda = L^{-1}$, in which case $\kappa = (L2c^2_{f,\Omega})^{-1}$. When f is a γ-strongly convex function, $\kappa = \gamma/L$ is the condition number of f (see Example 3.6) . So $(L2c^2_{f,\Omega})^{-1}$ can be seen as a generalized condition number, extending this notion from strongly convex functions to p-Łojasiewicz ones.

In Theorem 4.1 the p-Łojasiewicz assumption with $p \in [1, +\infty[$ implies that the argmin f is nonempty. In what follows we will derive convergence rates for the objective function values, even in the case where f is bounded from below but has no minimizers. Such results are of interest for instance in function approximation theory, where the goal is to find the best approximation of a target function within a specified function class [35]. Since in general the considered classes are not closed in the ambient space, the minimizer of the error does not exist, but convergence rates in objective function values are useful. A similar problem appears also in supervised statistical learning theory, where some convergence results can still be obtained are available (see e.g. [34, Theorem 9] and [33, Theorem A.1]).

We show below that the p-Łojasiewicz notion can be extended to *nonpositive* values of p, which allows to describe thegeometry of problems without minimizers. Based on this new definition, we then derive sharp convergence rates for the objective function values.

Definition 4.5. Let $p \in]-\infty, 0[$, let $f \in \Gamma_0(X)$ be bounded from below, and let $\Omega \subset X$. We say that f is *p-Łojasiewicz* on Ω if $\exists c_{f,\Omega} > 0$ such that the Łojasiewicz inequality holds:

$$\forall x \in \Omega \cap \operatorname{dom}^* f, \quad (f(x) - \inf f)^{1-\frac{1}{p}} \leq c_{f,\Omega} \|\partial f(x)\|_-.$$

Similarly to the case $p \geq 1$, where this property describes the behavior of f around its minimizers, here it describes the decay of $f(x)$ when $\|x\|$ goes to $+\infty$. This assumption leads to convergence rates, interpolating between $o(1)$ and $o(n^{-1})$, depending on the value of $p < 0$. We will see in Section 5.1 that this result applies to ill-posed linear problems involving a compact operator between infinite dimensional spaces.

Theorem 4.6 (Rates of convergence, $p < 0$)**.** *Let $f \in \Gamma_0(X)$ be bounded from below and satisfying Assumption 2.1, $(x_n)_{n\in\mathbb{N}}$ be generated by the FB algorithm. Assume that:*

a) (Localization) *for all $n \in \mathbb{N}$, $x_n \in \Omega \subset X$,*

b) (Geometry) *f is p-Łojasiewicz on Ω, for some $p < 0$.*

Then the values converge sublinearly (with C'_p defined as in (53)):

$$\forall n \in \mathbb{N},\ f(x_n) - \inf f \leq C_p'^{\frac{p}{p-2}} n^{\frac{p}{2-p}}.$$

Remark 4.7 (On the sharpness of the rates II)**.** The rates obtained in Theorem 4.6 are sharp. Indeed, the function defined in (6) is p-Łojasiewicz on $\mathbb{R}$ with $p = -\alpha$, and our rates match the lower bounds obtained in Remark 2.3.

Theorem 4.6, together with Theorem 4.1, give a complete (and sharp) picture of the asymptotic behavior of the FB algorithm. In fact, looking at the proofs of the mentioned results, we see that the only properties of forward-backward algorithm that are used are (47) and (48). We can then extend the previous theorems to a broader class of first-order descent methods, which encompasses block coordinate descent methods, and/or variable metric extensions of the FB algorithm [5, 18, 46].

Theorem 4.8 (General first-order descent method)**.** *The statements of Theorems 4.1 and 4.6 remain true if the sequence $(x_n)_{n\in\mathbb{N}}$ is generated by any algorithm satisfying:*

$$(\exists a > 0) \qquad a\|x_{n+1} - x_n\|^2 \leq f(x_{n+1}) - f(x_n) \tag{18}$$

$$(\exists b > 0) \qquad \|\partial f(x_{n+1})\|_- \leq b\|x_{n+1} - x_n\|. \tag{19}$$

In that case the constant appearing in Theorem 4.1 becomes $\kappa := ab^{-2}c^{-2}_{f,\Omega}$.

4.2 How to localize the sequence of iterates

One of the two assumptions we do in Theorems 4.1 and 4.6 is that the sequence belongs to a set Ω on which the geometry of f is known. We discuss here some possible choices. One first simple case is when Ω remains invariant under the action of T_λ (see also Annex A.2).

Definition 4.9. We say that $\Omega \subset X$ is *FB-invariant* if for all $\lambda \in]0, 2L^{-1}[$, $T_\lambda \Omega \subset \Omega$.

Example 4.10 (FB-invariant sets)**.** Theorem 2.2i)-ii) and Lemma A.9.ii) imply that these sets are FB-invariant (as well as any of their intersection):

- $\mathbb{B}_X(\bar{x}, \delta)$ and $\overline{\mathbb{B}}_X(\bar{x}, \delta)$ for every $\bar{x} \in \operatorname{argmin} f$, and for every $\delta \in]0, +\infty]$,
- $[f < r]$ for every $r > \inf f$,
- $\{x \in X \mid \|\partial f(x)\|_- < M\}$ and $\{x \in X \mid \|\partial f(x)\|_- \leq M\}$, for every $M \in]0, +\infty]$,
- $\Omega = \{x_n\}_{n\in\mathbb{N}}$ if $(x_n)_{n\in\mathbb{N}}$ is generated by the FB algorithm.

Assuming that Ω is FB-invariant, the localization property becomes a simple assumption on the initialization of the algorithm. The proof of the next corollary is immediate:

Corollary 4.11 (Geometry on stable sets gives global rates)**.** Let $f \in \Gamma_0(X)$ be bounded from below and satisfying Assumption 2.1, and $(x_n)_{n\in\mathbb{N}}$ be generated by the FB algorithm. Assume that $\Omega \subset X$ is FB-invariant and that:

a) *(Initialization)* $x_0 \in \Omega$,

b) *(Geometry)* f is p-Łojasiewicz on Ω, for some $p \in]-\infty, 0[\cup[1, +\infty[$.

Then the results of Theorems 4.1 and 4.6 apply for the sequence $(x_n)_{n\in\mathbb{N}}$.

In some cases, it is possible to remove the assumption $x_0 \in \Omega$, to the price of having only *asymptotic rates*. Indeed, it suffices to prove that the sequence will enter in Ω at a certain iteration, which is the argument used in [5, 46], in a non-convex setting. This happens for instance with the local level sets, under a slight compactness assumption (see below).

Corollary 4.12 (Local geometry gives asymptotical rates)**.** Let $f \in \Gamma_0(X)$ be such that $\operatorname{argmin} f \neq \emptyset$ and satisfying Assumption 2.1. Let $(x_n)_{n\in\mathbb{N}}$ be generated by the FB algorithm and assume that:

a) *(Compactness)* $(x_n)_{n\in\mathbb{N}}$ admits a subsequence strongly converging to $\bar{x}$ in X,

b) *(Local geometry)* for some $p \in [1, +\infty[$:

$$(\exists (\delta, r) \in]0, +\infty]) \text{ such that } f \text{ is } p\text{-Łojasiewicz on } \mathbb{B}_X(\bar{x}, \delta) \cap [f < r + \inf f].$$

Then there exists $n_0 \in \mathbb{N}$ such that the rates of Theorem 4.1 apply for the sequence $(x_{n_0+n})_{n\in\mathbb{N}}$.

Proof. Let $(x_{n_k})_{k\in\mathbb{N}}$ be a subsequence strongly converging to some x_∞, which belongs to $\operatorname{argmin} f$ according to Theorem 2.2. Therefore, f is p-Łojasiewicz on $\Omega := \mathbb{B}_X(x_\infty, \delta) \cap [f < r + \inf f]$, for some $(\delta, r) \in]0, +\infty]$. Since $x_{n_k} \to x_\infty$ and $f(x_{n_k}) \downarrow \inf f$, there exists $K \in \mathbb{N}$ such that $x_{n_K} \in \Omega$. Since Ω is FB-invariant, we conclude that $(x_n)_{n\geq N} \subset \Omega$. □

Remark 4.13 (On the compactness assumption)**.** The compactness assumption made in Corollary 4.12 is always satisfied in finite dimension. Indeed Theorem 2.2 guarantees that the sequence is bounded under the assumption that $\operatorname{argmin} f \neq \emptyset$. If X has infinite dimension, this assumption can be verified provided that f has compact level sets, due to the decreasing property of $f(x_n)$.

The property that a sequence $(x_n)_{n\in\mathbb{N}}$ generated by an algorithm reaches a set of interest Ω after a finite number of iterations, is usually called identifiability, or finite identification of Ω [100, 68, 54], and Ω is therefore called an *active set*. For instance, the so-called *active manifolds* can be identified in finite time, under the assumption that f is partially smooth with respect to this manifold [54, 55]. An alternative approach, recently introduced in [43], shows that the *strata* of mirror-stratifiable functions are identifiable. We will use this notion of active strata to derive another asymptotic convergence result.

Before introducing the notion of mirror-stratifiability, we recall that a set $M \subset \mathbb{R}^N$ is said to be stratified by $\{M_i\}_{i=1}^s \subset M$ if this family is a finite partition $\sqcup M_i = M$ such that $M_i \cap \operatorname{cl} M_j \neq \varnothing \Leftrightarrow M_i \subset \operatorname{cl} M_j$. The latter inclusion endows the family of strata with an order relation $M_i \preceq M_j \Leftrightarrow M_i \subset \operatorname{cl} M_j$. Given a point $x \in M$, it will be useful to note M_x the unique strata which contains x.

Definition 4.14 (Mirror-stratifiable function)**.** We say that a function $f \in \Gamma_0(\mathbb{R}^N)$ is *mirror-stratifiable* if

a) $\operatorname{dom} \partial f$ (resp. $\operatorname{dom} \partial f^*$) is stratified by $\{M_i\}_{i=1}^s$ (resp. $\{M_i^*\}_{i=1}^s$),

b) the map $J_f : M \longmapsto \bigcup_{x\in M} \operatorname{ri} \partial f(x)$ realizes a bijection between $\{M_i\}_{i=1}^s$ and $\{M_i^*\}_{i=1}^s$,

c) the map J_f is decreasing, in the sense that $M_i \preceq M_j \Leftrightarrow J_f(M_j) \preceq J_f(M_i)$.

Both notions appear naturally in most sparsity-based inverse problems such as the 1-norm, group-lasso norm, nuclear norm, or the total variation, or any polyhedral function, see [43] for more details and many examples.

Corollary 4.15. Suppose that Assumption 2.1 is in force, that $X = \mathbb{R}^N$, and let $(x_n)_{n\in\mathbb{N}}$ be the sequence generated by the FB algorithm converging to some $\bar{x} \in \operatorname{argmin} f$. Assume that:

a) g is mirror-stratifiable, and we define $C_{\bar{x}} := \cup\{M \mid M_{\bar{x}} \preceq M \preceq J_g^{-1}(M^*_{-\nabla h(\bar{x})})\}$,

b) f is p-Łojasiewicz on $C_{\bar{x}} \cap \mathbb{B}_X(\bar{x}, \delta)$ for some $\delta \in]0, +\infty]$ and $p \in [1, +\infty[$.

Then there exists $n_0 \in \mathbb{N}$ such that the rates of Theorem 4.1 apply for the sequence $(x_{n_0+n})_{n\in\mathbb{N}}$. Note that $C_{\bar{x}} = M_{\bar{x}}$ holds whenever $0 \in \operatorname{ri} \partial f(\bar{x})$.

Proof. It follows from [43, Theorem 4] that there exists $n_0 \in \mathbb{N}$ for which $x_{n_0+n} \in C_{\bar{x}}$ for every $n \in \mathbb{N}$. Since $(x_n)_{n\in\mathbb{N}}$ converges to $\bar{x}$, we can assume that n_0 is such that $x_{n_0+n} \in C_{\bar{x}} \cap \mathbb{B}_X(\bar{x}, \delta)$ for every $n \in \mathbb{N}$. This, together with b), allows to apply Theorem 4.1 to the sequence $(x_{n_0+n})_{n\in\mathbb{N}}$. The equality $C_{\bar{x}} = M_{\bar{x}}$ follows directly from the bijectivity of J_g, and the fact that $-\nabla h(\bar{x}) \in \operatorname{ri} \partial g(\bar{x})$. □

The reader not familiar with the notion of mirror-stratifiability might wonder what is the active set $C_{\bar{x}}$ appearing in Corollary 4.15. Here are a few example of interest:

Example 4.16. We keep here the notations of Corollary 4.15:

- If $g(x) = \|x\|_1$, we can choose a stratification based on sets with prescribed support, which gives
$$C_{\bar{x}} = \{x \in \mathbb{R}^N \mid \operatorname{supp}(\bar{x}) \subset \operatorname{supp}(x) \subset \operatorname{act}(-\nabla h(\bar{x}))\}, \tag{20}$$
where $\operatorname{supp}(x)$ is the support of x, and $\operatorname{act}(x^*) = \{i \mid |x_i| = 1\}$ is the set of active indices of x^* in $[-1, 1]^N$. Some authors call $\operatorname{act}(-\nabla h(\bar{x}))$ the extended support of $\bar{x}$. In the case that $0 \in \operatorname{ri} \partial f(\bar{x})$, we have $\operatorname{supp}(\bar{x}) = \operatorname{act}(-\nabla h(\bar{x}))$.

- If $g(x) = \|x\|_*$ is the nuclear norm, we can choose a stratification based on sets of matrices with prescribed rank, which gives
$$C_{\bar{x}} = \{x \in \mathcal{M}_{M,N}(\mathbb{R}) \mid \operatorname{rank}(\bar{x}) \leq \operatorname{rank}(x) \leq \#\operatorname{act}(\sigma(-\nabla h(\bar{x})))\}, \tag{21}$$
where $\sigma(x^*)$ denotes the set of singular values of the matrix x^*. If $0 \in \operatorname{ri} \partial f(\bar{x})$, we have $\operatorname{rank}(\bar{x}) = \#\operatorname{act}(\sigma(-\nabla h(\bar{x})))$.

Remark 4.17 (Partial smoothness)**.** Even if there is no direct relation between mirror stratification and partial smoothness, all the above mentioned functions are both mirror-stratifiable and partially smooth, and it would be immediate to provide an analogue result to Corollary 4.15 for partially smooth functions. Note that when using the identification theorems for partially smooth functions, it is necessary to assume the qualification condition $0 \in \operatorname{ri} \partial f(\bar{x})$ to hold. In this case, the active manifold coincide with the active set $C_{\bar{x}} = M_{\bar{x}}$ for most practical cases (polyhedral functions, spectral norms), meaning that those cases are already covered by Corollary 4.15.

Remark 4.18 (On the assumptions)**.** Note that our assumptions do not require or imply that f has unique minimizer; we only require f to be Łojasiewicz on the active set. In Section 5.2, we will show how this geometrical assumption can be guaranteed, provided that $\nabla^2 h(\bar{x})$ is injective when restricted to the tangent cone of the active set. In [74, Thm. 3.7] the authors provide a sufficient condition for the Łojasiewicz inequality to hold locally when g is a partially smooth function.

4.3 Linear rates of convergence for the Forward-Backward algorithm

In this Section we give more insights on the linear rates for the FB algorithm. According to Theorem 4.1, $f(x_n) - \inf f$ and $\|x_n - x_\infty\|$ converge linearly when a 2-Łojasiewicz property is verified. Another decreasing quantity of interest is $\operatorname{dist}(x_n, \operatorname{argmin} f)$, and its Q-linear convergence is equivalent to asking that the forward-backward map T_λ satisfies

$$(\exists \varepsilon_{f,\Omega} \in]0,1[)(\forall x \in \Omega \cap \operatorname{dom} f) \quad \operatorname{dist}(T_\lambda x, \operatorname{argmin} f) \leq \varepsilon_{f,\Omega} \operatorname{dist}(x, \operatorname{argmin} f). \tag{22}$$

If such property holds on a set Ω containing $(x_n)_{n\in\mathbb{N}}$, the sequence $(\operatorname{dist}(x_n, \operatorname{argmin} f))_{n\in\mathbb{N}}$ will converge Q-linearly. In fact, it is possible to show that (22) is *equivalent* to the 2-conditioning of f on Ω, provided this set is FB-invariant (see Definition 4.9). This fact has been observed in [85] for the projected gradient method, with $\Omega = X$ and $\lambda = L^{-1}$, and below we extend the argument to our more general setting.

Proposition 4.19 (Linear rates and 2-conditioning)**.** Suppose that Assumption 2.1 is in force and assume that $\operatorname{argmin} f \neq \emptyset$. Let $\Omega \subset X$ and $\lambda \in]0, 2L^{-1}[$.

i) If f verifies (22) on Ω, then it is 2-conditioned on Ω with $\gamma_{f,\Omega} = \lambda^{-1}(2 - \lambda L)(1 - \varepsilon_{f,\Omega})^2$.

ii) If f is 2-conditioned on $T_\lambda \Omega$, then it verifies (22) on Ω with $\varepsilon_{f,\Omega} = (1 + \lambda \gamma_{f,\Omega})^{-1/2}$ for stepsizes $\lambda \in \left]0, L^{-1}\right]$.

Then, on FB-invariant sets, the 2-conditioning is equivalent to (22), for stepsizes $\lambda \in \left]0, L^{-1}\right]$.

Proof. Let $S = \operatorname{argmin} f$, and let $x \in \Omega$. It follows from the triangular inequality that

$$\operatorname{dist}(x, S) \leq \|x - \operatorname{proj}(T_\lambda x, S)\| \leq \|x - T_\lambda x\| + \operatorname{dist}(T_\lambda x, S). \tag{23}$$

Lemma A.9.i) implies that

$$\|x - T_\lambda x\|^2 \leq 2\lambda(2 - \lambda L)^{-1}(f(x) - \inf f) \tag{24}$$

For item i), combine (22), (23), and (24):

$$(1 - \varepsilon_{f,\Omega})^2 \operatorname{dist}(x; S)^2 \leq \|T_\lambda x - x\|^2 \leq 2\lambda(2 - \lambda L)^{-1}(f(x) - \inf f).$$

For item ii), Lemma A.9.i) with $u = \operatorname{proj}(x; S)$, and the fact that $\lambda \leq 1/L$ implies

$$\|T_\lambda x - \operatorname{proj}(x; S)\|^2 \leq \operatorname{dist}(x; S)^2 - 2\lambda(f(T_\lambda x) - \inf f).$$

Then, since f is 2-conditioned on $T_\lambda \Omega \ni T_\lambda x$, we can conclude from

$$\operatorname{dist}(T_\lambda x; S)^2 \leq \|T_\lambda x - \operatorname{proj}(x; S)\|^2 \leq \operatorname{dist}(x; S)^2 - \lambda \gamma_{f,\Omega} \operatorname{dist}(T_\lambda x; S)^2. \qquad \square$$

Let us assume that f is a γ-strongly convex function, with $\gamma > 0$ as in Example 3.6, and let $\bar{x}$ be its unique minimizer. Let $(x_n)_{n\in\mathbb{N}}$ be generated by the FB algorithm, for which we take $\lambda = 1/L$, and define the condition number of f as $\kappa := \gamma/L$. We compare the different linear rates that we can get for $\|x_n - \bar{x}\|$ by using different theorems, relying on more or less strong assumptions. Using that f is 2-Łojasiewicz (with $c_{f,X} = (2\gamma)^{-1/2}$, see Example 3.6), Theorem 4.1 yields R-linear rates of the form

$$\|x_n - \bar{x}\| \leq C\varepsilon^n,\ C > 0,$$

where $\varepsilon = 1/\sqrt{1+\kappa}$. If instead we exploit 2-conditioning (recall that in general this is a stronger notion than 2-Łojasiewicz , Proposition 3.3), we obtain Q-linear rates from Proposition 4.19 with exactly the same constant ε. If we use directly the strong convexity of f, we obtain in this case Q-linear rates with $\varepsilon = 1 - \kappa$ (see e.g. [95, Proposition 3]). So, the more information we use, the better rates we derive. In [85], the authors investigate different notions belonging between strong convexity and the 2-conditioning. For instance, under an assumption of "quasi strong convexity", they obtain $\varepsilon = \sqrt{(1-\kappa)/(1+\kappa)}$, which is smaller than $(1+\kappa)^{-1/2}$, but not as good as $1-\kappa$. In conclusion, two aspects are crucial in the linear convergence of forward-backward. First, to have Q-linear rates for the iterates, it is necessary and sufficient to require the 2-conditioning of the function, due to the equivalence result of Proposition 4.19. Second, just assuming 2-conditioning is not a guarantee of having a *fast computation* of the solution, since linear rates can be arbitrarily slow on any finite number of iterations. Indeed two constants play a key role: the condition number κ, which is directly related to $\gamma_{f,\Omega}$ (some extra assumptions on f could improve the value of $\gamma_{f,\Omega}$, see e.g. the discussion in Subsection 5.2), and ε (see also [85]).

4.4 Superlinear rates and finite termination

In this section, we refine the convergence analysis for the case $p \in]1,2[$, replacing the p-Łojasiewicz property with p-metric subregularity (or p-conditioning). As discussed in Remark 4.3, the order of superlinear convergence that we derive for the FB algorithm in the case $p \in]1,2[$ is not sharp. In Theorem 4.21, using p-metric subregularity (or p-conditioning) instead of p-Łojasiewicz , we derive better (and indeed sharp, see Remark 4.3) superlinear rates. Keep in mind these three notions are only equivalent via Proposition 3.3 if Ω verifies a stability condition. The proof of Theorem 4.21 below follows directly from the next lemma, which is a partial analogue of Proposition 4.19-ii).

Lemma 4.20. Suppose that Assumption 2.1 is in force and assume that $\operatorname{argmin} f \neq \emptyset$.

i) If ∂f is p-metrically subregular on $\Omega \subset X$, then for all $p \in]1,2[$, and $x \in \operatorname{dom}^* f$:

$$T_\lambda x \in \Omega \Rightarrow \operatorname{dist}(T_\lambda x, \operatorname{argmin} f)^{p-1} \leq 2/(\lambda\gamma_{\partial f,\Omega})^{-1}\operatorname{dist}(x, \operatorname{argmin} f).$$

ii) If f is p-conditioned on Ω, then for all $p \in]1,2[$, and $x \in \operatorname{dom}^* f$:

$$(x, T_\lambda x) \in \Omega^2 \Rightarrow (f(T_\lambda x) - \inf f)^{p-1} \leq \left(p/\gamma_{f,\Omega}\right)^2 (2/\lambda)^p\,(f(x) - \inf f).$$

Proof. Let $S = \operatorname{argmin} f$. Lemma A.9.ii), the triangular inequality, and Theorem 2.2-ii) yield

$$\lambda\|\partial f(T_\lambda x)\|_- \leq \|T_\lambda x - x\| \leq \|T_\lambda x - \operatorname{proj}(x,S)\| + \|\operatorname{proj}(x,S) - x\| \leq 2\operatorname{dist}(x,S). \tag{25}$$

For i), use the hypothesis with (25) to derive $\gamma_{\partial f,\Omega}\operatorname{dist}(T_\lambda x, S)^{p-1} \leq (2/\lambda)\operatorname{dist}(x,S)$. For ii), use the p-Łojasiewicz inequality via Proposition 3.3 , together with (25) and the p-conditioning:

$$(f(T_\lambda x) - \inf f)^{p-1} \leq (p/\gamma_{f,\Omega})\|\partial f(T_\lambda x)\|_-^p \leq (p/\gamma_{f,\Omega})^2(2/\lambda)^p(f(x) - \inf f) \qquad \square$$

Theorem 4.21. *Assume that $p \in]1,2[$ and that the hypotheses of Theorem 4.1 hold. If the p-Łojasiewicz hypothesis is replaced by p-metric subregularity (resp. p-conditioning), then* $\operatorname{dist}(x_n, \operatorname{argmin} f)$ *(resp.* $f(x_n) - \inf f$*) Q-superlinearly converges with order* $(p-1)^{-1}$.

We now discuss the relevance of these fast rates when f is p-Łojasiewicz with $p \in [1,2[$. While superlinear rates are well-known for the proximal algorithm applied to sharp functions, it is not observed for the gradient method. The apparent contradiction between this result and practice is in fact related to a quite intuitive fact, stated in the following Proposition: the more a function is smooth, the less it can be sharp. This means that the gradient algorithm cannot be applied to p-Łojasiewicz function, with $p < 2$, because it is incompatible with ∇f being Lipschitz continuous. A similar statement, under different assumptions, can be found in [13, Proposition 2.8].

Proposition 4.22. Let $f \in \Gamma_0(X)$ be such that $\operatorname{dom} f$ has a nonempty interior. Assume f to be differentiable on Ω, where $\Omega \subset X$ is convex and such that[3] $\operatorname{proj}(\Omega; \operatorname{argmin} f) \subsetneq \Omega$. Assume that f is p-conditioned on Ω, and that ∇f is α-Hölder continuous on Ω, i.e.

$$(\exists L_{\nabla f,\Omega,\alpha} > 0)(\exists \alpha > 0)(\forall (x,y) \in \Omega^2) \quad \|\nabla f(x) - \nabla f(y)\| \leq L_{\nabla f,\Omega,\alpha}\|x-y\|^{\alpha}.$$

Then $p \in [\alpha+1, +\infty[$. In the case that $p = \alpha + 1$, we have moreover that $\gamma_{f,\Omega} \leq L_{\nabla f,\Omega,\alpha}$.

Proof. Let $x \in \Omega \cap \operatorname{dom}^* f$, and $\bar{x} := \operatorname{proj}(x, \operatorname{argmin} f)$. Then $\bar{x} \in \Omega$ and $\bar{x} \neq x$. For all $t \in]0,1]$, let $x_t := tx + (1-t)\bar{x}$. Then $x_t \in \Omega \setminus \operatorname{argmin} f$ and $\bar{x} = \operatorname{proj}(x_t, \operatorname{argmin} f)$. From the p-conditioning assumption and the Descent Lemma A.10 applied at $(\bar{x}, x_t) \in \Omega^2$, we see that:

$$(\forall t \in]0,1]) \quad 0 < \frac{\gamma_{f,\Omega}}{p}\|x_t - \bar{x}\|^p \leq f(x_t) - f(\bar{x}) \leq \frac{L_{\nabla f,\Omega}}{\alpha+1}\|x_t - \bar{x}\|^{\alpha+1}. \tag{26}$$

If we suppose that $p < \alpha + 1$, then by passing to the limit for $t \to 0$, we get $\gamma_{f,\Omega}/p \leq 0$ which is impossible. So $p \geq \alpha + 1$, and if equality holds, $\gamma_{f,\Omega} \leq L_{\nabla f,\Omega}$ follows from (26). □

As a consequence of Proposition 4.22, we should not expect more than linear rates for the gradient method applied to a $C^{1,1}$ convex function. Such a result cannot be extended straightforwardly to the Forward-backward algorithm. For instance, the function $f(x) = \|x\|^2 + \|x\|$ has a nontrivial smooth term in its decomposition, but is still sharp at its minimizer.

5 Linear inverse problems: from modeling assumptions to convergence rates

Throughout this section, X and Y are Hilbert spaces and $A : X \longrightarrow Y$ is a bounded linear operator. X is called the parameter space and Y is the data space. Given the linear inverse problem $Ax = y$, for some $y \in Y$, we are interested in the (possibly regularized) convex optimization problem

$$\min_{x \in X} f(x) = g(x) + D(Ax; y), \tag{27}$$

where $g \in \Gamma_0(X)$ and $h = D(\cdot, y) \in \Gamma_0(Y)$. The goal of this section is to show that typical modeling assumptions made in the inverse problem literature can be interpreted as geometric assumptions on (27), which are often not local, in the sense of Definition 3.1. First, we show that the classical source conditions are equivalent to a Łojasiewicz condition on suitable subsets, that we call source sets. Second, we show that the restricted isometry property, which is the key for exact recovery in sparsity based regularization, induces a 2-conditioning of the problem over a cone of sparse vectors, which is identified in finite time by the algorithm. This result extends to general inverse problems with mirror-stratifiable regularizing functions, for which the restricted isometry property entails a 2-conditioning of the problem over an active set (introduced in Corollary 4.15).

[3] Note that $\operatorname{proj}(\Omega; \operatorname{argmin} f) \subset \Omega$ holds when $\Omega = \mathbb{B}_X(\bar{x}, \delta) \cap [f < r]$, for $\bar{x} \in \operatorname{argmin} f$, because $\operatorname{proj}(\cdot; \operatorname{argmin} f)$ is nonexpansive.

5.1 Łojasiewicz property of quadratic functions via source conditions in Hilbert spaces

All across this Section 5.1, we assume that $A : X \to Y$ is a bounded linear operator, that $y \in Y$, and that $f(x) := (1/2)\|Ax - y\|^2$ is the associated least squares function. We will also note $y^\dagger := \mathrm{proj}(y, \mathrm{cl}\, R(A))$, and, whenever $\mathrm{argmin}\, f \neq \emptyset$, we will note $x^\dagger := \mathrm{proj}(0, \mathrm{argmin}\, f)$, which verifies $Ax^\dagger = y^\dagger$.

5.1.1 Elements of linear algebra

Before going further into the topic, let us recall some basic (but not necessarily well-known) facts about bounded linear operators in Hilbert spaces. A first important difference with the finite-dimensional setting is that the set of minimizers of f can be empty:

Proposition 5.1 ([51, Theorem 3.1.1])**.** Let $A : X \to Y$ be a bounded linear operator, $y \in Y$ and $f(x) := \|Ax - y\|^2/2$. Then $\mathrm{argmin}\, f \neq \emptyset \Leftrightarrow y \in R(A) + R(A)^\perp \Leftrightarrow y^\dagger \in R(A)$.

We see that $\mathrm{argmin}\, f \neq \emptyset$ is guaranteed when $R(A)$ is closed, which for instance cannot happen for compact operators with infinite-dimensional range [51, Theorem 3.1.3]. Observe that the closedness of $R(A)$ can be checked by means of its singular values:

Proposition 5.2. Let $A : X \to Y$ be a bounded linear operator. Then $R(A)$ is closed if and only if $\sigma_{inf}(A) > 0$.

Proof. Use the fact that $R(A) = R((AA^*)^{1/2})$ [42, Proposition 2.18] together with [53, Remark 2.3] and the fact that $\mathrm{spec}((AA^*)^{1/2}) = \mathrm{spec}(AA^*)^{1/2}$ [56, §32 Theorem 3]. □

5.1.2 Known results about the Landweber algorithm

The quadratic function f can be minimized by means of a gradient method, defined as

$$(\forall n \in \mathbb{N}) \quad x_{n+1} = x_n - \lambda A^*(Ax_n - y), \text{ with } x_0 \in X \text{ and } \lambda \in \left]0, 2\|A^*A\|^{-1}\right[. \tag{28}$$

A vast literature is devoted to this algorithm, which is often called in this context the Landweber algorithm. It is well-known that whenever $\mathrm{argmin}\, f \neq \emptyset$, the sequence $(x_n)_{n\in\mathbb{N}}$ generated by the Landweber algorithm converges strongly to the projection of x_0 onto $\mathrm{argmin}\, f$ (see e.g. [42, Theorem 6.1], or [51, Theorem 3.3.2] for varying stepsizes). When the range $R(A)$ is closed, the algorithm behaves exactly as in finite dimensions: both iterates and values converge linearly, see Example 3.7 and Theorem 4.1. If the $R(A)$ is not closed, instead, the rates for $\|x_n - \bar{x}_0\|$ can be arbitrarily slow without additional assumptions [32, Theorem 12]. Moreover, [53, Theorem 2.1] shows that *no local Łojasiewicz property* can be satisfied by such quadratic function when $R(A)$ is not closed. This could suggest that it is not possible to rely on geometrical assumptions to obtain convergence rates. Nevertheless, as we will see below, this is not true. Indeed, in the inverse problem literature, this worst-case scenario is avoided by making an extra assumption on the problem. For instance, if the following *source condition* is verified

$$(\exists \mu \in]0, +\infty[) \quad x^\dagger \in R(A^*A)^\mu, \tag{29}$$

the Landweber algorithm initialized with $x_0 = 0$ is known [42] to have the rates

$$f(x_n) - \inf f = O(n^{-(1+2\mu)}), \text{ and } \|x_n - x^\dagger\| = O(n^{-\mu}). \tag{30}$$

Also, when $\mathrm{argmin}\, f = \emptyset$, a source condition in Y can be made:

$$(\exists \nu \in]0, +\infty[) \quad y^\dagger \in R(AA^*)^\nu, \tag{31}$$

so that the Landweber algorithm initialized with $x_0 = 0$ verifies [34, Theorem 2.10]:

$$f(x_n) - \inf f = O(n^{-2\nu}). \tag{32}$$

The source condition (31) can be understood in light of Proposition 5.1. Indeed, this proposition says that the problem is well posed (in the sense that $\operatorname{argmin} f \neq \emptyset$) when $y^\dagger \in R(A)$. So it is reasonable to think that the "deeper" $y^\dagger$ is in $R(A)$, and the easier the problem is. In the ill-posed case $y^\dagger \in \operatorname{cl} R(A) \setminus R(A)$, we could also imagine that the "further away" $y^\dagger$ is from $R(A)$, and the more difficult the problem is. Estimating the location of $y^\dagger$ can be done thanks to the spaces $R(AA^*)^\nu$, because they form a sequence of nonincreasing dense subsets of $\operatorname{cl} R(A)$ (see Lemma A.14 and [42, Proposition 2.8]):

$$\operatorname{cl} R(A) = \operatorname{cl} \bigcup_{\nu>0} R(AA^*)^\nu \quad \text{and} \quad R(AA^*)^{1/2} = R(A).$$

The aim of this section is to highlight how the rates (30) and (32) can be simply explained using the results of Section 4. We show that the source conditions (29) and (31) are equivalent to assume that the initialization x_0 of the algorithm belongs to a so-called *source set*. Our main result in this section consists in showing that the function f satisfies a Łojasiewicz inequality on these source sets, which are FB-invariant. As a by-product of Corollary 4.11, we will obtain a new and simple geometrical interpretation of the rates in (30) and (32).

5.1.3 Regularity spaces and source sets

Definition 5.3 (Regularity space and source set)**.**

1. Given $(\nu, \delta) \in]0, +\infty[\times]0, +\infty[$, the data regularity space and the data source set are respectively defined as:

$$Y_\nu := y^\dagger + R(AA^*)^\nu \quad \text{and} \quad Y_{\nu,\delta} := y^\dagger + \{(AA^*)^\nu \omega \mid \omega \in \operatorname{cl} R(A), \|\omega\| \leq \delta\}.$$

2. Given $(\mu, \delta) \in]-1/2, +\infty[\times]0, +\infty[$, the regularity space and the source set are respectively defined as:

$$X_\mu := A^{-1} Y_{\mu+1/2} \quad \text{and} \quad X_{\mu,\delta}(y) := A^{-1} Y_{\mu+1/2,\delta},$$

where A^{-1} denotes the preimage of a set under the application A.

Proposition 5.4.

i) $\operatorname{argmin} f = \emptyset$ if and only if $X_\mu = \emptyset$ for all $\mu \in [0, +\infty[$.

ii) $\operatorname{argmin} f \neq \emptyset$ if and only if $X_\mu = X$ for all $\mu \in]-1/2, 0]$.

iii) Assume $R(A)$ is closed. Then $X_\mu = X$ for all $\mu \in]-1/2, +\infty[$.

Proof. Given any $x \in X$, observe that $x \in X_0$ is, by definition, equivalent to $Ax \in Y_{1/2}$. Since $R(A) = R(AA^*)^{1/2}$, the latter is equivalent to $Ax \in y^\dagger + R(A)$. We can then easily deduce, using also Proposition 5.1, that $X_0 = X \Leftrightarrow X_0 \neq \emptyset \Leftrightarrow y^\dagger \in R(A) \Leftrightarrow \operatorname{argmin} f \neq \emptyset$. For items i) and ii), the claim follows directly from the nonincreasingness of $\{X_\mu\}_{-1/2<\mu<+\infty}$. For item iii), observe that for all $\nu > 0$, $\operatorname{spec}((AA^*)^\nu) = \operatorname{spec}(AA^*)^\nu$ [56, §32 Theorem 3]. As a consequence of Proposition 5.2, we deduce that $R(AA^*)^\nu$ is closed, and therefore $R(AA^*)^\nu = R(A)$ (see Lemma A.14 in the Annex). In particular, $Y_\nu = y^\dagger + R(A)$ for all $\nu \in]0, +\infty[$, and the result follows from item ii). □

For well-posed problems, for which $\operatorname{argmin} f \neq \emptyset$ (and $x^\dagger$ exists), the source sets can be expressed with a simpler expression (the proof is left in the Annex):

Lemma 5.5 (Source sets for well-posed problems)**.** Assume that $\operatorname{argmin} f \neq \emptyset$. Then, for all $\mu > 0$ and $\delta > 0$:

$$X_\mu = \{x^\dagger\} + \operatorname{Ker} A + R(A^*A)^\mu \text{ and } X_{\mu,\delta} = \{x^\dagger\} + \operatorname{Ker} A + \{(A^*A)^\mu w \mid w \in \operatorname{Ker} A^\perp, \|w\| \leq \delta\}.$$

Remark 5.6. Given that $x^\dagger \in \ker A^\perp$, we see that the classical conditions in (29) and (31) are equivalent, with our notations, to $0 \in X_\mu$ and $0 \in X_{\nu-1/2}$. This means in particular that (29) is just a particular case of (31).

Remark 5.7 (Source sets as balls)**.** Assume that A is injective and $y \in R(A)$. For all $\mu > 0$, $R(A^*A)^\mu$ is a dense subspace of X (Lemma A.14), and we can endow it with the norm induced by the unbounded operator $(A^*A)^{-\mu}$, defined as $\|x\|_\mu := \inf\{\|w\| \mid w \in X \text{ and } x = (A^*A)^\mu w\}$. Then, we see that the source sets $X_{\mu,\delta}$ are nothing but balls centered at the solution $x^\dagger$, with respect to this norm:

$$X_{\mu,\delta} = \{x \in X \mid \|x - x^\dagger\|_\mu \leq \delta\},$$

while X_μ is the affine space spanned by these balls. By doing an analogy with the following example, the reader can think about this norm $\|\cdot\|_\mu$ in X as if it was a Sobolev norm in an L^2 space. Note that these balls may have an empty interior with respect to the topology of X.

Example 5.8 (Regularity spaces as Sobolev spaces)**.** Assume that X is the space of zero mean L^2-functions on $[0, 2\pi]$:

$$X = \left\{\varphi \in L^2([0,2\pi]),\ \int_0^{2\pi} \varphi(t)\,\mathrm{d}t = 0\right\}.$$

If A is the linear integration operator defined on X, then $R(A^*A)^\mu$ coincides with the Sobolev space $H^{2\mu}([0,2\pi]) \cap X$ [59, Theorem 6.4], so that the regularity space is here

$$X_\mu = \{x^\dagger\} + H^{2\mu}([0,2\pi]) \cap X.$$

5.1.4 Properties of quadratic functions on source sets

Here is the main result of this section: on each source set $X_{\mu,\delta}$, the least squares functional f is p-Łojasiewicz with $p = 2 + \mu^{-1}$.

Theorem 5.9 (Geometry of least squares on source sets)**.** *Let $\mu \in]-1/2, 0[\cup]0, +\infty[$ and $\delta \in]0, +\infty[$. Then $f(x) = \frac{1}{2}\|Ax - y\|^2$ is p-Łojasiewicz on $\Omega = X_{\mu,\delta}$, with*

$$p = 2 + \mu^{-1} \text{ and } c_{f,\Omega} = 2^{-(\mu+1)/(2\mu+1)}\delta^{1/(1+2\mu)}. \tag{33}$$

Moreover, these two constants are sharp.

Proof. Let $x \in X_{\mu,\delta}$ and remind that $y^\dagger = \mathrm{proj}(y, \mathrm{cl}\, R(A))$. From Definition 5.3 and the definition of $y^\dagger$, we get

$$Ax = y^\dagger + (AA^*)^{\mu+1/2}\omega, \text{ where } \omega \in \ker A^{*\perp} \text{ with } \|\omega\| \leq \delta, \tag{34}$$

$$f(x) - \inf f = (1/2)\|Ax - y^\dagger\|^2 \text{ and } \|\nabla f(x)\| = \|A^*(Ax - y^\dagger)\|. \tag{35}$$

We first prove that f verifies the Łojasiewicz inequality by using the interpolation inequality (see Lemma A.13 in the Annex) with $\alpha = \mu + (1/2)$ and $\beta = \mu + 1$, together with (34):

$$\|Ax - y^\dagger\| = \|(AA^*)^{\mu+1/2}\omega\| \leq \|(AA^*)^{1+\mu}\omega\|^{\frac{2\mu+1}{2\mu+2}}\|\omega\|^{\frac{1}{2\mu+2}} \leq \|(AA^*)^{1+\mu}\omega\|^{\frac{2\mu+1}{2\mu+2}}\delta^{\frac{1}{2\mu+2}}. \tag{36}$$

We use (34) in the right member of (36), to write

$$\|(AA^*)^{1+\mu}\omega\|^2 = \|(AA^*)^{1/2}(AA^*)^{\mu+1/2}\omega\|^2 = \|(AA^*)^{1/2}(Ax - y^\dagger)\|^2 = \|A^*(Ax - y^\dagger)\|^2. \tag{37}$$

By combining (34), (35), (36) and (37), we obtain the following inequality

$$f(x) - \inf f = (1/2)\|Ax - y^\dagger\|^2 \leq (1/2)\delta^{\frac{1}{\mu+1}}\|A^*(Ax - y^\dagger)\|^{\frac{2\mu+1}{\mu+1}} = (1/2)\delta^{\frac{1}{\mu+1}}\|\nabla f(x)\|^{\frac{2\mu+1}{\mu+1}}.$$

Then the desired Łojasiewicz inequality holds by taking $p := 2 + \mu^{-1}$. Now we verify that the obtained constants in (33) are sharp. For this, let $X = \ell^2(\mathbb{N})$, and let $(e_k)_{k\in\mathbb{N}} \subset X$ be its canonical basis. Let $(\sigma_k)_{k\in\mathbb{N}}$ be a strictly positive sequence converging to zero, and define $A : X \longrightarrow X$ as follows: $\forall x = (x_k)_{k\in\mathbb{N}} \in X$, $Ax := \sum_{k\in\mathbb{N}} \sigma_k x_k e_k$. Let $f(x) = (1/2)\|Ax\|^2$, $y = 0$, and let us assume that f is p-Łojasiewicz on $X_{\mu,\delta}$ for some $p \geq 1$:

$$(\forall x \in X_{\mu,\delta}) \quad [(1/2)\|Ax\|^2]^{1-(1/p)} \leq c_{f,X_{\mu,\delta}}\|A^*Ax\|. \tag{38}$$

Let $v^k := \delta\sigma_k^{2\mu} e_k \in X_{\mu,\delta}$, which satisfies $\|A^*Av^k\| = \delta\sigma_k^{2+2\mu}$, and deduce from (38) that

$$(\forall k \in \mathbb{N}) \quad 2^{(1-p)/p}\delta^{(2p-2)/p} \leq c_{f,X_{\mu,\delta}}\sigma_k^{(1/p)(4\mu+2)-2\mu}\delta. \tag{39}$$

It follows from $\sigma_k \to 0$ that $4\mu - 2\mu p + 2 \leq 0$, which is equivalent to $\mu p \geq 2\mu + 1$. If $\mu > 0$, it means that $p \geq 2 + \mu^{-1} > 0$, which is a regime in which the smallest is p, the better. If $\mu \in]-1/2, 0[$, then $p \leq 2 + \mu^{-1} < 0$, which is a regime in which the largest is p, the better. In both cases we see that $2 + \mu^{-1}$ is the best possible exponent. Moreover, when $p = 2 + \mu^{-1}$, (39) becomes $2^{-\frac{1+\mu}{1+2\mu}}\delta^{\frac{1}{1+2\mu}} \leq c_{f,X_{\mu,\delta}}$, which implies the sharpness of the constant obtained in (33). □

Remark 5.10. The result of Theorem 5.9 contrasts with [53, Theorem 2.1], in which the authors show that no local Łojasiewicz property can be satisfied by a quadratic function when $R(A)$ is not closed. The key difference here is that we look at the Łojasiewicz property on specific *dense* sets with empty interior (see Remark 5.7).

Let us now verify that the source sets are invariant under the action of the Landweber algorithm (28). As mentioned at the beginning of the section, the Landweber algorithm is the gradient decent algorithm applied to a quadratic function, and therefore it is an instance of the FB algorithm. We can thus apply the convergence rates of Section 4 once we prove that the source sets are invariant.

Proposition 5.11 (Invariance of source sets)**.** For all $(\mu, \delta) \in]-1/2, \infty[\times]0, +\infty[^2$, the source set $X_{\mu,\delta}$ is FB-invariant.

Proof. Let $x \in X_{\mu,\delta}$, $\lambda \in]0, 2/\|A^*A\|[$, and let us prove that $T_\lambda x = x - \lambda A^*(Ax - y)$ belongs to $X_{\mu,\delta}$. By using Lemma 5.5, we deduce that $Ax = y^\dagger + (AA^*)^\nu\omega$, $\nu := \mu + 1/2$, and $\omega \in \operatorname{cl} R(A)$ with $\|\omega\| \leq \delta$. Since $A^*(Ax - y) = A^*(Ax - y^\dagger)$, this implies that

$$AT_\lambda x = Ax - \lambda AA^*(Ax - y^\dagger) = y^\dagger + (AA^*)^\nu(I - \lambda AA^*)\omega$$

The above equality shows that $T_\lambda x \in X_\mu$. It remains only to prove that $\hat{T}_\lambda\omega := (I - \lambda AA^*)\omega$ verifies $\hat{T}_\lambda\omega \in \operatorname{cl} R(A)$ and $\|\hat{T}_\lambda\omega\| \leq \delta$. The condition $\hat{T}_\lambda\omega \in \operatorname{cl} R(A)$ immediately follows from $\omega \in \operatorname{cl} R(A)$ and $AA^*\omega \in R(A)$. Next, observe that $\hat{T}_\lambda\omega$ is obtained by applying a gradient descent step to ω with respect to the function $u \mapsto (1/2)\|A^*u\|^2$. Since this function has zero as a minimizer, and is differentiable with a $\|A^*A\|$-Lipschitz gradient, the Fejér property (see Theorem 2.2-ii)) implies that $\|\hat{T}_\lambda\omega\| \leq \|\omega\| \leq \delta$. □

Next we combine all the results of this section to derive convergence rates of the Landweber algorithm under source conditions from Łojasiewicz conditions.

Corollary 5.12 (Convergence rates for Landweber algorithm)**.** Let $(x_n)_{n\in\mathbb{N}}$ be a sequence generated by the Landweber algorithm (28). Assume that for some $\mu \in]-1/2, +\infty[$, the source condition $x_0 \in X_\mu$ is satisfied. Then:

i) $f(x_n) - \inf f = O(n^{-(1+2\mu)})$,

ii) If $\mu > 0$, then $\|x_n - \bar{x}_0\| = O(n^{-\mu})$, where $\bar{x}_0 := \operatorname{proj}(x_0, \operatorname{argmin} f)$.

Proof. For item i), the source condition together with Proposition 5.11 imply $(x_n)_{n\in\mathbb{N}} \subset X_{\mu,\delta}$ for some $\delta > 0$. If $\mu \neq 0$, we derive from Theorem 5.9 that f is $2 + \mu^{-1}$-Łojasiewicz on $X_{\mu,\delta}$. Depending on the sign of $2 + \mu^{-1}$, the rates on $f(x_n) - \inf f$ follow from Theorems 4.1 and 4.6. If $\mu = 0$, then the source condition and Proposition 5.4 ensures that $y^\dagger \in R(A)$, meaning that $\operatorname{argmin} f \neq \emptyset$, so the rate $O(n^{-1})$ follows from Theorem 2.2. For item ii), the convergence and rates on the iterates follows from Theorem 4.1. To show that the limit of the sequence (let us note it x_∞) is $\bar{x}_0$, it is enough to verify that $x_\infty - x_0 \in \ker A^\perp$, since $\operatorname{argmin} f$ is an affine space parallel to $\ker A$. Because of the definition of the algorithm, it is easy to show by recurrence that $x_n - x_0 \in R(A^*)$. This being true for all $n \in \mathbb{N}$, we can pass to the limit and deduce that $x_\infty - x_0 \in \operatorname{cl} R(A^*) = \ker A^\perp$. □

5.2 Sparsity based regularization, partial smoothness, and restricted injectivity

In this section we turn to the general case of optimization problems coming from a regularized inverse problem (27). In particular, we focus on the case where $\nabla^2 h$ verifies a restricted injectivity condition at a solution, a situation which typically arises when g is mirror-stratifiable, and typical modeling assumptions from the inverse problems/compressed sensing literature hold. In this setting we will be able to derive the 2-conditioning of the objective function in (27). In what follows, we will use the notation $\mathcal{S}_+(X)$ to refer to the set of bounded selfadjoint positive linear operators on X.

5.2.1 Coercive linear operators on a cone

Definition 5.13. We say that $K \subset X$ is a cone if it is a union of rays: $[0, +\infty[K \subset K$.

Note that we do not require a cone to be convex. This is important for certain applications in which we have geometrical information about a function over a union of linear spaces, see for instance (40) in the context of sparse regularization problems.

Definition 5.14. Let $S \in \mathcal{S}_+(X)$, let $\gamma \in]0, +\infty[$, and let $K \subset X$ be a cone. We say that S is γ-coercive on K if, for all $d \in K$, $\langle Sd, d\rangle \geq \gamma \|d\|^2$.

Example 5.15 (coercivity for positive symmetric matrices)**.** A matrix $S \in \mathcal{S}_+(\mathbb{R}^N)$ is coercive on a closed cone $K \subset \mathbb{R}^N$ if and only if S is injective when restricted on K (see Proposition A.15 for a proof):

$$K \cap \operatorname{Ker} S = \{0\}.$$

Example 5.16. Any operator $S \in \mathcal{S}_+(X)$ is $\sigma_{\inf}(S)$-coercive on $\operatorname{Ker} S^{\perp}$ (see e.g. the proof in [30, Thm. 4]). In particular, if S is positive definite then it is $\sigma_{\inf}(S)$-coercive on X.

In the next proposition we relate the coercivity of the Hessian of a function f on a cone to the 2-conditioning of f on this cone. This relation can be seen as a weakened analogue of the well known fact (see [11, Prop. 10.8 & 17.7.(iii)]) that, for $f \in C^2(X)$:

$$f \text{ is } \gamma\text{-strongly convex} \Leftrightarrow (\forall x \in X)\ \nabla^2 f(x) \text{ is } \gamma\text{-coercive on } X.$$

Strong convexity is a global notion, which requires the function to have a *positive definite* quadratic-like geometry at each $x \in X$. On the contrary, the 2-conditioning requires the function to have a *positive* quadratic-like geometry, on a given set Ω. We now state our result (its proof is left in the Annex A.5). For similar results, see also [19, Section 3.3.1] and [41].

Proposition 5.17 (Coercivity of the Hessian implies 2-conditioning)**.** Let $f = g + h$ with $g, h \in \Gamma_0(X)$ and $\operatorname{argmin} f \neq \emptyset$. Assume that h is of class C^2 in a neighbourhood of $\bar{x} \in \operatorname{argmin} f$, and that $\nabla^2 h(\bar{x})$ is γ-coercive on a closed cone $K \subset X$. Then,

$$(\forall \gamma' \in]0, \gamma[)\ (\exists \delta \in]0, +\infty])\ \text{s.t. } f \text{ is 2-conditioned on } \Omega := \bar{x} + (K \cap \delta \mathbb{B}_X) \text{ with } \gamma_{f,\Omega} = \gamma',$$

and $\Omega \cap \operatorname{argmin} f = \{\bar{x}\}$. If $h \in C^2(X)$ and $\nabla^2 h$ is L-Lipschitz, we can take $\delta = \frac{\gamma - \gamma'}{L}$.

5.2.2 Conditioning on prox-regular sets via restricted injectivity of the Hessian

Let us define some useful tools from variational analysis. The notion of reached set (or set with positive reach) was introduced by Federer [44, Def. 4.1], and later extended to *prox-regularity* (see Proposition A.19 and [93]).

Definition 5.18. Let $C \subset \mathbb{R}^N$. The (Bouligand) tangent cone to C at $\bar{x} \in C$ is defined as

$$T_C(\bar{x}) := \{d \in \mathbb{R}^N \mid (\exists t_n \downarrow 0)(\exists d_n \to d) \quad \bar{x} + t_n d_n \in C\}.$$

The normal cone to C at $\bar{x}$ is $N_C(\bar{x}) := \{\eta \in \mathbb{R}^N \mid (\forall d \in T_C(\bar{x}))\ \langle \eta, d\rangle \leq 0\}$.

Definition 5.19. Let $C \subset \mathbb{R}^N$, and $\rho > 0$. We say that C is ρ-reached at $\bar{x} \in C$, if it is locally closed at $\bar{x}$, and verifies

$$(\forall \eta \in N_C(\bar{x}) \cap \mathbb{S}_{\mathbb{R}^N}) \quad \mathbb{B}(\bar{x} + \frac{1}{\rho}\eta, \frac{1}{\rho}) \cap C = \emptyset.$$

We say that C is prox-regular at $\bar{x}$ if there exists $\rho > 0$ and a closed neighbourhood U of $\bar{x}$ such that $C \cap U$ is ρ-reached at any $x \in U$. We say further that C is prox-regular if it is prox-regular at every $\bar{x} \in C$.

Convex sets, and in particular affine spaces, are prox-regular. Manifolds of class C^2 are locally prox-regular (see Proposition A.19).

We now provide the result at the core of this section, which says that if a minimizer $\bar{x}$ belongs to some prox-regular set, and if the Hessian $\nabla^2 h(\bar{x})$ is injective when restricted to the tangent cone of this set, then f is 2-conditioned on this set around $\bar{x}$. This will guarantee asymptotic linear rates when combined with Corollary 4.15.

Theorem 5.20 (Injective Hessian on tangent cone implies 2-conditioning)**.** *Let $g, h \in \Gamma_0(\mathbb{R}^N)$, and $f = g + h$. Assume that there exists some $\bar{x} \in \operatorname{argmin} f$ such that:*

a) $\bar{x}$ belongs to some $C \subset \mathbb{R}^N$ which is ρ-reached at $\bar{x}$,

b) h is of class C^2 in a neighbourhood of $\bar{x}$,

c) Ker $\nabla^2 f(\bar{x})$ is γ-coercive on $T_C(\bar{x})$.

Then $\operatorname{argmin} f_{|C} = \{\bar{x}\}$*, and for every $\gamma' \in]0, \gamma[$, there exists $\delta \in]0, +\infty]$ such that f is 2-conditioned on $\Omega := C \cap \mathbb{B}(\bar{x}, \delta)$, with $\gamma_{f,\Omega} = \gamma'$. If we assume moreover that $\nabla^2 h$ is L-Lipschitz continuous, then we can take $\delta = \frac{2(\gamma - \gamma')}{2L + \rho \|\nabla^2 h(\bar{x})\|}$.*

Proof. Let $K := T_C(\bar{x})$. Using Proposition A.20, we see that for every $\gamma' < \gamma$ there exists a $\theta \in]0, \frac{\pi}{2}[$ such that the enlarged cone K_θ (see Definition A.16) contains $(C - \bar{x}) \cap \mathbb{B}(0, \delta)$ for $\delta > 0$ small enough, and such that $\nabla^2 h(\bar{x})$ is γ'-coercive on K_θ. The conclusion of the claim follows from Proposition 5.17 applied to h and K_θ. Under the additional assumption that $\nabla^2 h$ is L-Lipschitz, take any $\gamma' \in]0, \gamma[$, and let $\gamma'' := \alpha\gamma + (1 - \alpha)\gamma'$, with $\alpha = 2L/(2L + \rho\|\nabla^2 h(\bar{x})\|)$. Using again Proposition A.20, we obtain that $\nabla^2 h(\bar{x})$ is γ''-coercive on some cone K_θ, with $\bar{x} + K_\theta \supset C \cap \mathbb{B}(\bar{x}, \delta_1)$ and $\delta_1 = 2(\gamma - \gamma'')/(\rho\|\nabla^2 h(\bar{x})\|)$. Then, Proposition 5.17 shows that f is 2-conditioned on $\Omega = \bar{x} + K_\theta \cap \mathbb{B}(\bar{x}, \delta_2)$, with $\delta_2 = (\gamma'' - \gamma')/L$ and $\gamma_{f,\Omega} = \gamma'$. The conclusion follows by seeing that $\delta_1 = \delta_2$ with our choice of γ''. □

Theorem 5.20 can be used in combination with Corollary 4.15: in this case we obtain that the restricted injectivity of the Hessian on the tangent cone to the active set $C_{\bar{x}}$ guarantees asymptotic linear rates. In the example below, we detail what our assumptions mean for the examples in Example 4.16.

Example 5.21.

- If $g(x) = \|x\|_1$, the active set (20) is an open and dense subset of the vector space $X_I = \{x \in \mathbb{R}^N \mid \operatorname{supp}(x) \subset I\}$ with $I = \operatorname{act}(-\nabla h(\bar{x}))$. It is therefore ρ-reached for every $\rho > 0$, and $T_{C_{\bar{x}}}(\bar{x}) = X_I$.
- If $g(x) = \|x\|_*$, let $r = \#\operatorname{act}(\sigma(-\nabla h(\bar{x})))$ and let $\mathcal{M}_r$ be the manifold of matrices with rank equal to r. If $0 \in \operatorname{ri} \partial f(\bar{x})$, the active set $C_{\bar{x}}$ (see (21)) is equal to $\mathcal{M}_r$. In particular, it is prox-regular (see Proposition A.19), and an expression for its tangent space can be found in [69, Example 2.2]. More generally, $C_{\bar{x}}$ is locally prox-regular at $\bar{x}$ if $\operatorname{rank}(\bar{x}) = r$. To see this, use the same arguments as in [80, Prop. 3.1]: the fact that the singular values depend continuously on the matrix allows to find a neighbourhood U of $\bar{x}$ where the matrices have a rank greater or equal to r. This means that $C_{\bar{x}} \cap U = \mathcal{M}_r \cap U$, which is prox-regular.

Remark 5.22 (Related results with partial smoothness)**.** While our results are new in the setting of mirror-stratifiable functions (where no condition $0 \in \operatorname{ri} \partial f(\bar{x})$ is required), they intersect with existing results when g is partially smooth with respect to an active manifold $\mathcal{M}$. It is shown in [75] that the γ-coercivity of $\nabla^2 h(\bar{x})$ on the tangent space $T_{\mathcal{M}}(\bar{x})$ guarantees asymptotic linear rates. We recover a similar result by combining Theorem [54, Theorem 5.3] with Theorem 5.20 and Theorem 2.2. For a fixed stepsize $\lambda = 1/L$, [75, Thm. 3.1] predicts a Q-linear rate arbitrarily close to $\sqrt{2(1-\kappa)}$ (where $\kappa = \gamma/L$) provided that $\kappa \geq 1/2$. Instead, our results predict a R-linear rate arbitrarily close to $(1 + (\kappa/4))^{-1/2}$, without condition on κ. Note that our constant is worse (resp. better) than $\sqrt{2(1-\kappa)}$ when κ is close to 1 (resp. $1/2$). Note also that the partial smoothness of g together with [54, Theorem 6.2.ii)] ensures that f is 2-conditioned on a neighbourhood Ω of the solution, with $\gamma_{f,\Omega} = \gamma'$, meaning that we can use Proposition 4.19 to obtain Q-linear rates arbitrarily close to $(1+\kappa)^{-1/2}$.

5.2.3 Application to low-complexity inverse problems

Consider $f \in \Gamma_0(\mathbb{R}^N)$ be defined by, for every $x \in \mathbb{R}^N$, $f(x) = \alpha\|x\|_1 + (1/2)\|Ax - y\|^2$. f is the sum of a smooth function, with Hessian equal to A^*A, and a nonsmooth function $\alpha\|x\|_1$. Example 3.9 ensures that f is locally 2-conditioned on its sublevel sets *without any assumption* on A. This means, according to Theorem 4.1, that for any $r > \inf f$, and any $x_0 \in [f < r]$, there exists a constant $\varepsilon \in]0,1[$ such that the iterative soft-thresholding initialized at x_0 verifies $f(x_{n+1}) - \inf f \leq \varepsilon(f(x_n) - \inf f)$. Nevertheless, expressing the 2-conditioning constant, or ε, in terms of the components of the problems is far to be easy [17]. One way to recover a meaningful constant is to exploit modeling assumptions which are usually made to ensure the stability and recovery of the inverse problem $Ax = y$.

Suppose that we are given the sequence generated by the iterative soft-thresholding, which converges to a minimizer of f, $x_n \to \bar{x}$. It is known that, after some iterations, the support of the sequence is stable [76, 49]:

$$(\exists I \subset \{1,\dots,N\})(\exists n_0 \in N)(\forall n \geq n_0) \quad \operatorname{supp}(x_n) \subset I.$$

In particular, if the qualification condition $0 \in \operatorname{ri} \partial f(\bar{x})$ holds, we can take $I = \operatorname{supp}(\bar{x})$ [76, Prop. 3.6]. To estimate the rates of convergence for the sequence, it is then sufficient to make a *restricted injectivity* assumption on the matrix A, depending on the knowledge we have on I.

In the case we have access to I, suppose that on the space $X_I := \{x \in \mathbb{R}^N \mid \operatorname{supp}(x) \subset I\}$ the matrix A is injective, i.e. $\operatorname{Ker} A \cap X_I = \{0\}$ holds. Then, there exists a constant $\gamma_I > 0$ such that A^*A is γ-coercive on X_I (see Example 5.15), which implies via Proposition 5.17 that f is 2-conditioned on X_I, with $\gamma_{f,X_I} = \gamma_I$. We deduce then that, asymptotically, the rates are governed by $\varepsilon = (1 + \gamma_I\|A^*A\|^{-1})^{-1}$. It might happen that instead of knowing I, we have only access to a partial information via the sparsity level $s := |I|$. We can then follow the same reasoning with the (nonconvex) cone $K_s := \{x \in \mathbb{R}^N \mid |\operatorname{supp}(x)| \leq s\}$ instead of X_I. In that case, the constant γ_s of coercivity of A^*A on K_s is defined by

$$(\forall x \in K_s) \quad \gamma_s\|x\|^2 \leq \|Ax\|^2, \tag{40}$$

and guarantees linear rates governed by $\varepsilon = (1 + \gamma_s\|A^*A\|^{-1})^{-1}$, using again Proposition 5.17. Such assumption is classical in sparsity based regularization, and it is related to the so-called Restricted Isometry Property [25], to ensure uniqueness of the minimizer and guarantee the robustness or recovery [99, 26]. Observe that while the computation of γ_s remains impracticable [9], it is *meaningful* with respect to the properties of our problem, and, more importantly, can be estimated when the matrix A is random [47, Section 9]. Of course, this whole discussion can be extended to other regularized inverse problems, in particular if $\|\cdot\|_1$ is replaced by a mirror-stratifiable function. In this case we will use Theorem 5.20 instead of Proposition 5.17 to derive linear rates.

6 Conclusion and perspectives

In this paper, we dicussed in details how geometry can be used to improve the rates of the FB method, or more general first-order descent schemes. We characterized the geometry, using tools that are often encountered in practice, like the p-conditioning, and we provided a new sum rule for it. In Figure 6.1 we recall the various rates obtained for the FB method, from the worst case scenario (no minimizers, no assumptions) to the best one (sharp functions).

	$f(x_n) - \inf f$	$\|x_n - x_\infty\|$
$\inf f > -\infty$	$o(1)$	—
$p \in]-\infty, 0[$	$O(n^{p/(2-p)})$	—
$\operatorname{argmin} f \neq \emptyset$	$o(n^{-1})$	decreasing, $o(1)$ in finite dimension
$p \in]2, +\infty[$	$O(n^{-p/(p-2)})$	$O(n^{-1/(p-2)})$
$p = 2$	Q-linear with $\varepsilon = 1/(1+\kappa)$	R-linear with $\varepsilon = 1/(1+\kappa)$
$p \in]1, 2[$	Q-superlinear of order $1/(p-1)$	R-superlinear of order $1/(p-1)$
$p = 1$	finite	finite

Figure 6.1: Convergence rates of the FB algorithm for locally p-Łojasiewicz functions (with the constant κ defined in Theorem 4.1).

We also have discussed how those refined results can be obtained by decoupling the geometrical information we have on the function and the localization of the sequence we are looking at. This geometry-based analysis reduces then the gap between theory and practice, where the observed rates are often better than the ones resulting from a worst case analysis. It moreover shows that linear rates are tightly linked to 2-conditioned function. In addition, we showed how our analysis can be specialized to the inverse problems setting, and allows to explain typical modeling assumptions in this context, such as source conditions and restricted injectivity property. It is worth noting that the geometrical information such as conditioning or Łojasiewicz property can be exploited to derive sharper convergence rates for a broader class of functions and/or algorithms than just forward-backward algorithm [5]. We also emphasize that convexity plays no role in the proofs of Theorems 4.1 and 4.6. Indeed, some of these results were already known for non-convex functions [18, 27, 46]. One of the challenges in the future is to have quantitative results concerning the geometry of classes of nonconvex functions. For instance, what can be said about "simple" nonconvex piecewise polynomial functions (see [73] for an answer about maximum of finitely many polynomials)? Can we estimate the Łojasiewicz exponent of semialgebraic functions, depending on the degree of the polynomials defining their graph? Finally, a last challenge is the application of such geometrical tools to derive precise rates for nondescent methods. First results in this direction, using 2-conditioning are known for inertial methods [85, 77] or stochastic gradient methods [61]. It would be of interest to understand the behavior of these algorithms for other geometries.

A Appendix

A.1 Worst case analysis: proofs of Section 2

The following Lemma contains a detailed proof for the lower bound (7) in Example 2.3, which can also be applied to (5) by using a symmetry argument.

Lemma A.1 (Lower bounds for the proximal algorithm)**.** Let $p \in]-\infty, 0[\cup]2, +\infty[$, and let $f_p \in \Gamma_0(\mathbb{R})$ be the function defined by

$$\text{if } p < 0,\ f_p(x) = \begin{cases} |x|^p & \text{if } x < 0, \\ +\infty & \text{if } x \geq 0, \end{cases} \quad \text{and if } p > 2,\ f_p(x) = \begin{cases} 0 & \text{if } x < 0, \\ |x|^p & \text{if } x \geq 0. \end{cases}$$

If $x_0 \in \operatorname{dom} f \setminus \operatorname{argmin} f$, and $x_{n+1} = \operatorname{prox}_{\lambda f}(x_n)$, then for all $n \geq 1$:

$$f_p(x_n) - \inf f_p \geq C_p^p n^{\frac{p}{2-p}} \quad \text{with} \quad C_p = \left(|x_0|^{2-p} + p(p-2)\lambda\right)^{\frac{1}{2-p}}.$$

Proof. Note that $\operatorname{dom} f_p$ is an open interval, and that f_p is infinitely derivable there. We can then see that f_p, f_p' and f_p'' are non-negative. In particular, we deduce that f_p and f_p' are non-decreasing on $\operatorname{dom} f$.

Let us now take some $x_0 \in \operatorname{dom} f \setminus \operatorname{argmin} f$, and consider the following continuous trajectory

$$(\forall t \geq 0) \quad x(t) := \operatorname{sgn}(p)\left(|x_0|^{2-p} + p(p-2)t\right)^{\frac{1}{2-p}}.$$

It is a simple exercise to verify that $x(\cdot)$ is a solution of this differential equation:

$$x(0) = x_0, \quad \dot{x}(t) + f'(x(t)) = 0, \quad x(t) \in \operatorname{dom} f_p.$$

The main step towards proving our lower bound is to show, by induction, that for every $n \in \mathbb{N}$, $x_n \geq x(n\lambda)$. This is clearly true for $n = 0$, so, let us assume now that this is true for $n \in \mathbb{N}$, and show that this implies $x_{n+1} \geq x((n+1)\lambda)$. Start by writing

$$x((n+1)\lambda) = x(n\lambda) + \int_{n\lambda}^{(n+1)\lambda} \dot{x}(t)\, dt = x(n\lambda) + \int_{n\lambda}^{(n+1)\lambda} (-f_p' \circ x)(t)\, dt.$$

On the one hand, f_p' is non-negative on $\operatorname{dom} f$, and $\dot{x}(t) = -f_p'(x(t))$, which means that $x(\cdot)$ is increasing. On the other hand, f_p' is non-decreasing, which means that $(-f_p' \circ x)$ is increasing. This fact, together with our induction assumption, allows us to write

$$\begin{aligned} x((n+1)\lambda) &\leq x_n + \int_{n\lambda}^{(n+1)\lambda} (-f_p' \circ x)((n+1)\lambda))\, dt = x_n - \lambda f_p'(x((n+1)\lambda)), \\ \Leftrightarrow \quad x((n+1)\lambda) + \lambda f_p'(x((n+1)\lambda)) &\leq x_n. \end{aligned}$$

Consider now the function $\phi : \operatorname{dom} f_p \to]0, +\infty[$ defined by $\phi(t) = t + \lambda f_p'(t)$. It is clearly increasing and bijective on its image, so its inverse ϕ^{-1} is also increasing. We observe moreover that, by definition, the proximal sequence satisfies $x_{n+1} = \phi^{-1}(x_n)$. This allows us to write

$$\phi(x((n+1)\lambda)) \leq x_n \quad \Leftrightarrow \quad x((n+1)\lambda) \leq \phi^{-1}(x_n) = x_{n+1}.$$

This ends the proof of the induction argument.

Observe that, given non-negative numbers $a, b > 0$, the following inequality holds

$$(\forall n \geq 1) \quad \operatorname{sgn}(p)(a+bn)^{\frac{1}{2-p}} \geq \operatorname{sgn}(p)(a+b)^{\frac{1}{2-p}} n^{\frac{1}{2-p}}.$$

This means that, for all $n \geq 1$,

$$x_n \geq \operatorname{sgn}(p)\left(|x_0|^{2-p} + p(p-2)\lambda n\right)^{\frac{1}{2-p}} \geq \operatorname{sgn}(p)\left(|x_0|^{2-p} + p(p-2)\lambda\right)^{\frac{1}{2-p}} n^{\frac{1}{2-p}} = \operatorname{sgn}(p) C_p n^{\frac{1}{2-p}}.$$

Passing this inequality through f_p (which is non-decreasing) yields the desired result. □

A.2 Proofs of Section 3

A.2.1 Invariant sets and proofs of Section 3.1

We provide here a result concerning the equivalence between all the notions in Definition 3.1, for a large class of sets $\Omega \subset X$. The sets Ω we will consider are directly related to the gradient flow induced by ∂f. Given $u_0 \in \operatorname{dom} f$, it is known[4] that there exists a unique absolutely continuous trajectory noted $u(\cdot; u_0) : [0, +\infty[\longrightarrow X$, called the steepest descent trajectory, which satisfies:

$$(\text{for a.e. } t > 0) \quad \frac{\mathrm{d}}{\mathrm{d}t} u(t; u_0) + \partial f(u(t; u_0)) \ni 0, \text{ and } u(0; u_0) = u_0. \tag{41}$$

Following [21], we introduce the notion of invariant sets for the flow of ∂f:

[4] See [21, Thm 3.1] when $u_0 \in \operatorname{dom} \partial f$, and [21, Thm. 3.2] with [11, Cor. 16.39] when $u_0 \in \operatorname{cl} \operatorname{dom} f$.

Definition A.2. A set $\Omega \subset X$ is *∂f-invariant* if for any $x \in \Omega \cap \operatorname{dom} \partial f$ and a.e. $t > 0$, $u(t;x) \in \Omega$ holds.

In other words, Ω is said to be ∂f-invariant if any steepest descent trajectory starting in Ω remains therein. It is straightforward to see that the intersection of two ∂f-invariant sets is still ∂f-invariant.

Example A.3. An easy way to construct a ∂f-invariant set is to consider the sublevel set of a *Lyapunov* function $\psi : X \to \mathbb{R} \cup \{+\infty\}$ for the gradient flow induced by ∂f. A function is said to be Lyapunov if for any $x \in \operatorname{dom} f$, $\psi(u(\cdot;x)) : [0,+\infty[\to \mathbb{R}$ is decreasing. Classical examples of this kind are:

- $\Omega = X$, which is $[\psi < 1]$ with $\psi = 0$.
- $\Omega = [f < r]$ for $r > \inf f$, which is $[\psi < r]$ with $\psi = f$ (see [21, Thm. 3.2.17]).
- $\Omega = \mathbb{B}(\bar{x}, \delta)$ for $\bar{x} \in \operatorname{argmin} f$, $\delta > 0$, which is $[\psi < \delta]$ with $\psi(x) = \|x - \bar{x}\|$ (see [21, Thm. 3.1.7]).
- $\Omega = \{x \in X \mid \|\partial f(x)\|_- < M\}$ for $M > 0$, which is $[\psi < M]$ with $\psi(x) = \|\partial f(x)\|_-$ (see [21, Thm. 3.1.6]).

See [21, Section IV.4] for more details on the subject, as well as [22, 63]. It is also a good exercise to verify that the source sets considered in Proposition 5.11 are ∂f-invariant.

We next prove Proposition 3.3, stating the equivalence between conditioning, metric subregularity and Łojasiewicz on ∂f-invariant sets. The proof is based on an argument used in [17, Theorem 5], which relies essentially on the following convergence rate property for the continuous steepest descent dynamic (41).

Proof of Proposition 3.3. Convexity of f and the Cauchy-Schwartz inequality imply

$$(\forall x \in \operatorname{dom} f) \quad f(x) - \inf f \leq \|\partial f(x)\|_- \operatorname{dist}(x, \operatorname{argmin} f),$$

and so i) $\implies$ ii) $\implies$ iii). Next, we just have to prove that the Łojasiewicz property implies the conditioning one. So let us assume that f is p-Łojasiewicz on Ω, which is ∂f-invariant, and fix $x \in \Omega \cap \operatorname{dom}^* f$. Define, for all $t \geq 0$, $\varphi(t) := (pc_{f,\Omega})^{-1} t^{1/p}$, which is derivable on $]0,+\infty[$, and for all $u \in \operatorname{dom} f$, $r(u) = f(u) - \inf f$. Let us lighten the notations by noting $u(\cdot)$ instead of $u(\cdot;x)$, so that $u(0) = x$. Because we will need to distinguish the case in which the trajectory converges in finite time, we introduce $T := \inf\{t \geq 0 \mid u(t) \in \operatorname{argmin} f\} \in [0,+\infty]$. Since $x \in \operatorname{dom}^* f$ and $u(\cdot)$ is continuous, we see that $T > 0$. For every $t \in [0,T[$, we have $u(t) \notin \operatorname{argmin} f$, so $u(t) \in \Omega \cap \operatorname{dom}^* f$ and $r(u(t)) \neq 0$. If $T < +\infty$, we also have for every $t > T$ that $u(t) = u(T)$ and $\dot{u}(t) = 0$. Now, we write:

$$(\forall t \in]0,T[) \quad \varphi(r(x)) \geq \varphi(r(x)) - \varphi(r(u(t))) \quad = \quad \int_t^0 (\varphi \circ r \circ u)'(\tau)\, \mathrm{d}\tau = \int_t^0 \varphi'((r \circ u)(\tau)) \cdot (r \circ u)'(\tau)\, \mathrm{d}\tau.$$

But $\frac{\mathrm{d}}{\mathrm{d}\tau}(r \circ u)(\tau) = -\|\dot{u}(\tau)\|^2 = -\|\partial f(u(\tau))\|_-^2$ (see [21]), so that the above equality becomes

$$(\forall t \in]0,T[) \quad \varphi(r(x)) \geq \int_0^t \varphi'((r \circ u)(\tau)) \|\partial f(u(\tau))\|_-^2\, \mathrm{d}\tau. \tag{42}$$

Since we assume Ω to be ∂f-invariant, we can apply the Łojasiewicz inequality at $u(\tau) \in \Omega \cap \operatorname{dom}^* f$ for all $\tau \in]0,t[$, which can be rewritten in this case as $1 \leq \varphi'(r(u(\tau))) \|\partial f(u(\tau))\|_-$. This applied to (42) gives us:

$$(\forall t \in]0,T[) \quad \varphi(r(x)) \geq \int_0^t \|\dot{u}(\tau)\|\, \mathrm{d}\tau. \tag{43}$$

From (43) and the definition of T, we see that $\int_0^{+\infty} \|\dot{u}(\tau)\|\, \mathrm{d}\tau \leq \varphi(r(x)) < +\infty$, meaning that the trajectory $u(\cdot)$ has finite length. As a consequence, it converges strongly to some $\bar{u}$ when t tends to $+\infty$. Finally, we use on (43) the fact that $\|u(0) - u(t)\| \leq \int_0^t \|\dot{u}(\tau)\|\, \mathrm{d}\tau$, together with the fact that $\bar{u} \in \operatorname{argmin} f$ (see [21, Thm. 3.11]) to conclude that

$$\frac{1}{pc_{f,\Omega}} \operatorname{dist}(x, \operatorname{argmin} f) \leq \frac{1}{pc_{f,\Omega}} \|x - \bar{u}\| \leq (f(x) - \inf f)^{1/p}. \qquad \square$$

Proof of Proposition 3.4. i): let $S := \operatorname{argmin} f \neq \emptyset$. Given $\delta > 0$, there exists $M \in]0, +\infty[$ such that

$$\sup\{\operatorname{dist}(x, S) \mid x \in \Omega \cap \delta \mathbb{B}_X\} \leq M$$

Since f is p-conditioned on Ω, we deduce that:

$$(\forall x \in \Omega \cap \delta \mathbb{B}_X) \quad \operatorname{dist}(x, S)^{p'} = \operatorname{dist}(x, S)^{p} \operatorname{dist}(x, S)^{p'-p} \leq (pM^{p'-p}/\gamma_{f,\Omega})(f(x) - \inf f),$$

meaning that f is p'-conditioned on $\Omega \cap \delta \mathbb{B}_X$.

ii): the proof follows the same lines as in i). □

Proof of Proposition 3.5. Assume by contradiction that there exists a sequence $(z^n)_{n\in\mathbb{N}} \subset \Omega$ such that

$$n^{-1} \operatorname{dist}^p(z^n, \operatorname{argmin} f) > f(z^n) - \inf f. \tag{44}$$

Since Ω is weakly compact, we can assume without loss of generality that z^n weakly converges to some $z^\infty \in \Omega$ when $n \to +\infty$. Then, it follows from (44), the boundedness of $(z^n)_{n\in\mathbb{N}} \subset \Omega$ and the weak lower semi-continuity of f that $f(z^\infty) - \inf f \leq 0$, meaning that $z^\infty \in \operatorname{argmin} f$, contradicting $\Omega \cap \operatorname{argmin} f = \emptyset$. □

A.2.2 Proofs of Section 3.2

Lemma A.4 (The Łojasiewicz constant for uniformly convex functions)**.** Let $f \in \Gamma_0(X)$ be uniformly convex, of order $p \geq 2$, with constant γ. Then f is p-Łojasiewicz on X, with $c_{f,X} = q^{-1/q}\gamma^{-1/p}$, where $1 = (1/p) + (1/q)$.

Proof. Let $x \in \operatorname{dom} \partial f$, $\bar{x} \in \operatorname{argmin} f$, and $x^* \in \partial f(x)$. By definition of uniformly convex functions

$$f(x) - \inf f = \sup_{u\in X} f(x) - f(u) \leq - \inf_{u\in X} \left(\langle x^*, u - x\rangle + (\gamma/p)\|u - x\|^p\right). \tag{45}$$

The right member of the above inequality involves a strictly convex optimization problem, whose unique optimal value $\bar{u}$ can be determined by using Fermat's rule:

$$0 = x^* + \gamma\|\bar{u} - x\|^{p-2}(\bar{u} - x) \Leftrightarrow \bar{u} = x - \gamma^{-1/(p-1)}\|x^*\|^{(2-p)/(p-1)}x^*.$$

Injecting this optimal value in (45) gives, after rearranging the terms,

$$f(x) - \inf f \leq (1 - 1/p)\gamma^{-1/(p-1)}\|x^*\|^{p/(p-1)},$$

and, since x^* is arbitrary in $\partial f(x)$, the result follows after passing this inequality to the power $1 - 1/p$. □

Proof of Example 3.10.ii). To prove the claim, it is enough to verify the three conditions of [40, Theorem 4.2]. The first condition (boundedness of argmin f) is guaranteed by the fact that f is coercive. Indeed, h is strongly convex, therefore bounded from below, and g is itself coercive. The second condition (dual qualification conditions) follows immediately from the fact that both h^* and g^*, and are continuously differentiable. To see this, observe that in this example g^* is (up to a constant) $\|\cdot\|_q^q$, where q is the conjugate number of p: $(1/p) + (1/q) = 1$. Moreover, h being strongly convex means that h^* is also continuously differentiable, with $\operatorname{dom} h^* = \mathbb{R}^M$. The third condition (firm convexity) is easy to check for h because it is strongly convex; for g the proof is left in the following Lemma. We can then apply [40, Theorem 4.2], which ensures that f is 2-conditioned on every compact set. Using again the fact that f is coercive, and therefore has bounded sublevel sets, we conclude that f is 2-conditioned on every sublevel set. □

A.2.3 Proofs of Section 3.3

Lemma A.5 (p-powers are 2-tilt conditioned when $p \in]1,2]$)**.** Let $p \in]1,2]$, $u \in \mathbb{R}^N$, and $f : \mathbb{R}^N \to \mathbb{R}$ be defined as $f(x) = \frac{1}{p}\|x\|_p^p - \langle u, x\rangle$. Then f is 2-conditioned on every bounded subset of $\mathbb{R}^N$.

Proof. This function is a separable sum, so, without loss of generality, we can assume from here that $N = 1$ (see [40, Lemma 4.4]). Given a real $t \in \mathbb{R}$, we will note its sign with $s(t)$, which is equal to -1 (resp. $+1$) if $t < 0$ (resp. $t > 0$), or 0 if $t = 0$. Using the convexity, the differentiability of f, and the Fermat's rule, we see that f admits a unique minimizer $\bar{x}$, defined by the relations

$$0 = s(\bar{x})|\bar{x}|^{p-1} - u \Leftrightarrow \bar{x} = s(u)|u|^{\frac{1}{p-1}} \Leftrightarrow u = s(\bar{x})|\bar{x}|^{p-1}.$$

If $u = 0$, it is immediate to see that f is 2-conditioned on $]-1,1[$, where the relation $|t|^2 \leq |t|^p$ holds. We therefore assume from now that $u \neq 0$, which also means that $\bar{x} \neq 0$. We now compute (we note $q = p/(p-1)$)

$$\inf f = f(\bar{x}) = \frac{1}{p}|\bar{x}|^p - u\bar{x} = \frac{1}{p}|\bar{x}|^p - s(\bar{x})|\bar{x}|^{p-1}\bar{x} = \frac{1}{p}|\bar{x}|^p - |\bar{x}|^p = -\frac{1}{q}|\bar{x}|^p,$$

meaning that we are looking for an inequality like

$$\gamma|x - \bar{x}|^2 \leq \frac{1}{p}|x|^p - ux - \inf f = \frac{1}{p}|x|^p - s(\bar{x})|\bar{x}|^{p-1}x + \frac{1}{q}|\bar{x}|^p.$$

Using the L'Hôpital rule twice allows us to study the following limit:

$$\lim_{x\to\bar{x}} \frac{\frac{1}{p}|x|^p - s(\bar{x})|\bar{x}|^{p-1}x + \frac{1}{q}|\bar{x}|^p}{|x-\bar{x}|^2} = \lim_{x\to\bar{x}} \frac{s(x)|x|^{p-1} - s(\bar{x})|\bar{x}|^{p-1}}{2(x-\bar{x})} = \lim_{x\to\bar{x}} \frac{(p-1)|x|^{p-2}}{2} = \frac{(p-1)}{2}|\bar{x}|^{p-2}.$$

Note that our assumption that $\bar{x} \neq 0$ ensures that we can take the derivative of the second numerator around $\bar{x}$. Since this limit is well-defined, and nonnegative, it means that f is 2-conditioned on a small enough neighbourhood of $\bar{x}$. To conclude the proof, it remains to verify that f is 2-conditioned on *any* bounded set. This follows immediately from Proposition 3.5 and the fact that $\operatorname{argmin} f = \{\bar{x}\}$. □

Lemma A.6 (Kullback-Leibler divergences are 2-tilt conditioned)**.** Let $\bar{x} \in]0,+\infty[^N$, and $f \in \Gamma_0(\mathbb{R}^N)$ be the Kullback-Leibler divergence to $\bar{x}$:

$$f(x) = KL(\bar{x}; x) = \sum_{i=1}^N kl(\bar{x}_i; x_i) \quad \text{where} \quad kl(\bar{t}; t) = \begin{cases} \bar{t}\log(\frac{\bar{t}}{t}) - \bar{t} + t & \text{if } t > 0, \\ +\infty & \text{else.} \end{cases}$$

Then f is 2-tilt-conditioned on every bounded set of $\mathbb{R}^N$.

Proof. Let $d \in \mathbb{R}^N$, and define the tilted function $\tilde{f} = f + \langle d\cdot\rangle$. Using Fermat's rule, we see that $\operatorname{argmin} f = \partial f^*(-d)$. It is a simple exercice to verify that $\operatorname{dom}\partial f^* =]-\infty, 1[^N$, so $\operatorname{argmin}\tilde{f} \neq \emptyset$ if and only if $d \in]-1,+\infty[^N$. Let d be such vector, and write, for any $x_i > 0$:

$$\tilde{f}_i(x_i) = \bar{x}_i\log(\frac{\bar{x}_i}{x_i}) - \bar{x}_i + x_i + d_i x_i = (1+d_i)\left(\frac{\bar{x}_i}{1+d_i}\log(\frac{\bar{x}_i}{x_i}) - \frac{\bar{x}_i}{1+d_i} + x_i\right).$$

Let $X_i := \frac{\bar{x}_i}{1+d_i}$, which is well defined under our assumption that $d_i > -1$. Then

$$\tilde{f}_i(x_i) = (1+d_i)\left(X_i\log(\frac{X_i}{x_i}) - X_i + x_i + X_i\log(1+d_i)\right) = (1+d_i)kl(X_i; x_i) + a_i,$$

where $a_i = X_i(1+d_i)\log(1+d_i) > 0$. We then observe that $\operatorname{argmin}\tilde{f}_i = \{X_i\}$, from which we deduce that $\operatorname{argmin}\tilde{f} = \{X\}$ with $X = (X_i)_{i=1}^N$.

Now, let $\delta > 0$ be fixed, and let $x \in \mathbb{B}(X, \delta)$. Let $\underline{d} := \min_i d_i > -1$, $c := N\|X\|_\infty$, and

$$C := \frac{1}{\delta^2 c}\left(\frac{\delta}{c} - \ln\left(1 + \frac{\delta}{c}\right)\right) \quad \text{which is nonnegative because } t > \ln(1+t) \text{ on }]0,+\infty[.$$

For each $i \in \{1, \dots, N\}$, we have $|x_i - X_i| \leq \delta$, so we can use [24, Lem. A.2] on $\tilde{f}_i$ to write

$$\begin{aligned} \tilde{f}(x) - \inf \tilde{f} &= \sum_{i=1}^{N} \tilde{f}_i(x) - \tilde{f}_i(X_i) = \sum_{i=1}^{N} (1+d_i) kl(X_i; x_i) \\ &\geq \sum_{i=1}^{N} (1+d_i) C |X_i - x_i|^2 \geq (1+\underline{d}) C \|X - x\|^2. \end{aligned}$$

This proves that $\tilde{f}$ is 2-conditioned on $\mathbb{B}(X, \delta)$, which conludes the proof. □

A.3 The Forward-Backward algorithm and proofs of Section 4

Definition A.7. Given a positive real sequence $(r_n)_{n\in\mathbb{N}}$ converging to zero, we say that r_n converges:

- *sublinearly* (of order $\alpha \in]0, +\infty[$) if $\exists C \in]0, +\infty[$ such that $\forall n \in \mathbb{N}$, $r_n \leq C n^{-\alpha}$,
- *Q-linearly* if $\exists \varepsilon \in]0, 1[$ such that $\forall n \in \mathbb{N}$, $r_{n+1} \leq \varepsilon r_n$,
- *R-linearly* if $\exists (s_n)_{n\in\mathbb{N}}$ Q-linearly converging such that $\forall n \in \mathbb{N}$, $r_n \leq s_n$,
- *Q-superlinearly* (of order $\beta \in]1, +\infty[$) if $\exists C \in]0, +\infty[$ such that $\forall n \in \mathbb{N}$, $r_{n+1} \leq C r_n^{\beta}$,
- *R-superlinearly* if $\exists (s_n)_{n\in\mathbb{N}}$ Q-superlinearly convergent such that $\forall n \in \mathbb{N}$, $r_n \leq s_n$.

It is easy to verify that r_n is R-superlinearly convergent of order $\beta > 1$ if and only if

$$(\forall \varepsilon \in]0, 1[)(\exists C > 0)(\forall n \in \mathbb{N}) \quad r_n \leq C \varepsilon^{\beta^n}.$$

Note that *R*-linear and *R*-superlinear convergence ensures only the overall decrease of the sequence, while *Q*-linear and *Q*-superlinear convergence requires the sequence to decrease at a certain speed for each index. It is immediate from the definition that *Q*-convergence implies *R*-convergence.

Lemma A.8 (Estimate for sublinear real sequences)**.** Let $(r_n)_{n\in\mathbb{N}}$ be a real sequence being strictly positive and satisfying, for some $\kappa > 0$, $\alpha > 1$ and all $n \in \mathbb{N}$: $r_n - r_{n+1} \geq \kappa r_{n+1}^{\alpha}$. Define $\tilde{\kappa} := \min\{\kappa, \kappa^{\frac{\alpha-1}{\alpha}}\}$, and $\delta := \max_{s\geq 1} \min\left\{\frac{\alpha-1}{s}, \kappa^{-\frac{\alpha-1}{\alpha}} r_0^{1-\alpha}\left(1 - s^{-\frac{\alpha-1}{\alpha}}\right)\right\} \in]0, +\infty[$. Then, for all $n \in \mathbb{N}$, $r_n \leq (\tilde{\kappa}\delta n)^{-1/(\alpha-1)}$.

Proof. It can be found in [72, Lemma 7.1], see also the proofs of [3, Theorem 2] or [46, Theorem 3.4]. □

Lemma A.9. If Assumption 2.1 holds, then for all $(x, u) \in X^2$ and all $\lambda > 0$:

i) $\|T_\lambda x - u\|^2 - \|x - u\|^2 \leq (\lambda L - 1)\|T_\lambda x - x\|^2 + 2\lambda(f(u) - f(T_\lambda x))$.

ii) $\|\partial f(T_\lambda x)\|_- \leq \lambda^{-1}\|T_\lambda x - x\| \leq \|\partial f(x)\|_-$.

Proof of Lemma A.9. To prove item i), start by writing

$$\|T_\lambda x - u\|^2 - \|x - u\|^2 = -\|T_\lambda x - x\|^2 + 2\langle x - T_\lambda x, u - T_\lambda x\rangle.$$

The optimality condition in (2) gives $x - T_\lambda x \in \lambda \partial g(T_\lambda x) + \lambda \nabla h(x)$ so that, by using the convexity of g:

$$\|T_\lambda x - u\|^2 - \|x - u\|^2 \leq -\|T_\lambda x - x\|^2 + 2\lambda\left(g(u) - g(T_\lambda x) + \langle \nabla h(x), u - T_\lambda x\rangle\right).$$

Since we can write $\langle \nabla h(x), u - T_\lambda x\rangle = \langle \nabla h(x), u - x\rangle + \langle \nabla h(x), x - T_\lambda x\rangle$, we deduce from the convexity of h and the Descent Lemma ([11, Theorem 18.15]) that

$$\langle \nabla h(x), u - T_\lambda x\rangle \leq h(u) - h(x) + h(x) - h(T_\lambda x) + \frac{L}{2}\|T_\lambda x - x\|^2 = h(u) - h(T_\lambda x) + \frac{L}{2}\|T_\lambda x - x\|^2.$$

Item i) is then proved after combining the two previous inequalities. For item ii), the optimality condition in (2), together with a sum rule (see e.g. [87, Theorem 3.30]), to deduce that

$$\forall (u,v) \in X^2, \quad v = \mathrm{prox}_{\lambda g}(u) \Leftrightarrow \lambda^{-1}(u-v) + \nabla h(v) \in \partial f(v). \tag{46}$$

For the first inequality, use (46) with $(u,v) = (x - \lambda \nabla h(x), T_\lambda x)$, together with the contraction property of the gradient map $x \mapsto x - \lambda \nabla h(x)$ when $0 < \lambda \leq 2/L$ (see [11, Cor. 18.17 & Prop. 4.39 & Remark 4.34.i]) to obtain:

$$\|\partial f(T_\lambda x)\|_- \leq \lambda^{-1} \|(x - \lambda \nabla h(x)) - (T_\lambda x - \lambda \nabla h(T_\lambda x))\| \leq \lambda^{-1} \|T_\lambda x - x\|.$$

For the second inequality, consider $x^* := \mathrm{proj}(-\nabla h(x), \partial g(x))$, and use (46) with $(u,v) = (x + \lambda x^*, x)$, together with the nonexpansiveness of the proximal map (see [11, Prop. 12.28]):

$$\|T_\lambda x - x\| = \|\mathrm{prox}_{\lambda g}(x - \lambda \nabla h(x)) - \mathrm{prox}_{\lambda g}(x + \lambda x^*)\| \leq \lambda \|\nabla h(x) + x^*\| = \lambda \|\partial f(x)\|_-. \qquad \square$$

Lemma A.10 (Descent Lemma for Hölder smooth functions)**.** Let $f : X \longrightarrow \mathbb{R}$ and $C \subset X$ be convex. Assume that f is Gateaux differentiable on C, and that there exists $(\alpha, L) \in]0, +\infty[^2$, such that for all $(x,y) \in C^2$, $\|\nabla f(x) - \nabla f(y)\| \leq L\|x-y\|^\alpha$ holds. Then:

$$(\forall (x,y) \in C^2) \quad f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leq \frac{L}{\alpha + 1} \|x - y\|^{\alpha+1}.$$

Proof. The argument used in [101, Remark 3.5.1] for $C = X$ extends directly to convex sets. $\square$

Now we can prove the convergence rate results of Section 4.1:

Proof of Theorem 4.1. We first show that $(x_n)_{n\in\mathbb{N}}$ has finite length. Since $\inf f > -\infty$, $r_n := f(x_n) - \inf f \in [0, +\infty[$, and it follows from Lemma A.9 that

$$a\|x_{n+1} - x_n\|^2 \leq r_n - r_{n+1}, \text{ with } a = \frac{1}{2\lambda}(2 - \lambda L) > 0, \tag{47}$$

$$\|\partial f(x_{n+1})\|_- \leq b\|x_n - x_{n+1}\|, \text{ with } b = \lambda^{-1}. \tag{48}$$

If there exists $n \in \mathbb{N}$ such that $r_n = 0$ then the algorithm would stop after a finite number of iterations (see (47)), therefore it is not restrictive to assume that $r_n > 0$ for all $n \in \mathbb{N}$. We set $\varphi(t) := pt^{1/p}$ and $c := c_{f,\Omega}$, so that the Łojasiewicz inequality at $x_n \in \Omega \cap \mathrm{dom}^* f$ can be rewritten as

$$(\forall n \in \mathbb{N}) \quad 1 \leq c\varphi'(r_n)\|\partial f(x_n)\|_-. \tag{49}$$

Combining (47), (48), and (49), and using the concavity of φ, we obtain for all $n \geq 1$:

$$\|x_{n+1} - x_n\|^2 \leq \frac{bc}{a}\varphi'(r_n)(r_n - r_{n+1})\|x_n - x_{n-1}\| \leq \frac{bc}{a}(\varphi(r_n) - \varphi(r_{n+1}))\|x_n - x_{n-1}\|.$$

By taking the square root on both sides, and using Young's inequality, we obtain

$$(\forall n \geq 1) \quad 2\|x_{n+1} - x_n\| \leq \frac{bc}{a}(\varphi(r_n) - \varphi(r_{n+1})) + \|x_n - x_{n-1}\|. \tag{50}$$

Sum this inequality, and reorder the terms to finally obtain

$$(\forall n \geq 1) \quad \sum_{k=1}^{n} \|x_{k+1} - x_k\| \leq \frac{bc}{a}\varphi(r_1) + \|x_1 - x_0\|.$$

We deduce that $(x_n)_{n\in\mathbb{N}}$ has finite length and converges strongly to some x_∞. Moreover, from (48) and the strong closedness of $\partial f : X \rightrightarrows X$, we conclude that $0 \in \partial f(x_\infty)$.

Now we prove the convergence rates. Let $c = c_{f,\Omega}$ for short. We first derive rates for the sequence of values $r_n := f(x_n) - \inf f$, from which we will derive the rates for the iterates. Equations (47) and (48) yield

$$r_n - r_{n+1} \geq a\|x_{n+1} - x_n\|^2 \geq \frac{a}{b^2}\|\partial f(x_{n+1})\|_-^2.$$

The Łojasiwecz inequality at $x_{n+1} \in \Omega \cap \mathrm{dom}^* f$ implies $c^2 r_{n+1}^{2/p}(r_n - r_{n+1}) \geq ab^{-2} r_{n+1}^2$, so we deduce that

$$(\forall n \in \mathbb{N}) \quad r_{n+1} \neq 0 \;\Rightarrow\; r_{n+1}^{2/p}(r_n - r_{n+1}) \geq \kappa r_{n+1}^2, \quad \text{with } \kappa := a(bc)^{-2}. \tag{51}$$

The rates for the values are derived from the analysis of the sequences satisfying the inequality in (51). Depending on the value of p, we obtain different rates.

• If $p = 1$, then we deduce from (51) that for all $n \in \mathbb{N}, r_{n+1} \neq 0$ implies $r_{n+1} \leq r_n - \kappa$. Since the sequence $(r_n)_{n\in\mathbb{N}}$ is decreasing and positive, $r_{n+1} \neq 0$ implies $n \leq r_0 \kappa^{-1}$.

For the other values of p, we will assume that $r_n > 0$. In particular, we get from (51)

$$(\forall n \in \mathbb{N}) \quad r_n - r_{n+1} \geq \kappa r_{n+1}^{\alpha}, \quad \text{with } \alpha := 2(p-1)p^{-1} \text{ and } \kappa := ab^{-2}c^{-2}. \tag{52}$$

• If $p \in]1,2[$, then $\alpha \in]0,1[$. The positivity of r_{n+1} and (52) imply that for all $n \in \mathbb{N}$, $r_{n+1} \leq \kappa^{-1/\alpha} r_n^{1/\alpha}$, meaning that r_n converges Q-superlinearly.

• If $p = 2$, then $\alpha = 1$ and we deduce from (52) that for all $n \in \mathbb{N}$, $r_{n+1} \leq (1+\kappa)^{-1} r_n$, meaning that r_n converges Q-linearly.

• If $p \in]2, +\infty[$, then $\alpha \in]1,2[$, and the analysis still relies on studying the asymptotic behaviour of a real sequence satisfying (52). Lemma A.8 in the Annex shows that we have $r_{n+1} \leq (C_p')^{p/(p-2)} n^{-p/(p-2)}$, by taking

$$(C_p')^{-1} := \min\left\{\kappa, \kappa^{\frac{p-2}{2p-2}}\right\} \max_{s\geq 1} \min\left\{\frac{p-2}{ps}, \kappa^{\frac{2-p}{2p-2}} r_0^{\frac{2-p}{p}} \left(1 - s^{-\frac{p-2}{2p-2}}\right)\right\}. \tag{53}$$

To end the proof, we will prove that the rates for $\|x_n - x_\infty\|$ are governed by the ones of r_n. Let $1 \leq n \leq N < +\infty$, and sum the inequality in (50) between n and N to obtain (remind that $b = \lambda^{-1}$):

$$\|x_N - x_n\| \leq \sum_{k=n}^{N} \|x_{k+1} - x_k\| \leq \frac{pc}{a\lambda} r_n^{1/p} + \|x_n - x_{n-1}\|.$$

Next, we pass to the limit for $N \to \infty$, we use (47), and the fact that r_n is decreasing to obtain

$$(\forall n \geq 1) \quad \|x_\infty - x_n\| \leq \frac{pc}{a\lambda} r_{n-1}^{1/p} + \frac{1}{\sqrt{a}} \sqrt{r_{n-1}}. \tag{54}$$

Note that $r_{n-1}^{1/2} \leq r_0^{\frac{1}{2}-\frac{1}{p}} r_{n-1}^{1/p}$ if $p \in [2, +\infty[$, and $r_{n-1}^{1/p} \leq r_0^{\frac{1}{p}-\frac{1}{2}} r_{n-1}^{1/2}$ if $p \in [1,2]$. So, by defining

$$C_p := \begin{cases} 2pc(2-\lambda L)^{-1} + (2\lambda r_0)^{1/2}(2-\lambda L)^{-1/2} r_0^{-1/p} & \text{if } p \geq 2, \\ 2pc r_0^{1/p}(2-\lambda L)^{-1} r_0^{-1/2} + (2\lambda)^{1/2}(2-\lambda L)^{-1/2} & \text{if } p \leq 2, \end{cases} \tag{55}$$

we finally conclude from (54) that $\|x_\infty - x_n\| \leq C_p r_{n-1}^{1/\max\{2,p\}}$ when $n \geq 1$. □

Proof of Theorem 4.6. The proof is as for the case $p \in]2, +\infty[$ of Theorem 4.1: the p-Łojasiewicz property implies (51), and the statement follows from Lemma A.8 with $\alpha = 2(p-1)/p \in]2, +\infty[$. □

Proof of Theorem 4.8. The proofs of Theorems 4.1 and 4.6 rely on the combination of the Łojasiewicz inequality with the estimations (47) and (48), which can be replaced by (18) and (19). □

A.4 Linear inverse problems and proofs of Section 5.1

Here we will make use of is the Moore-Penrose *pseudo-inverse* of A. It is a linear operator (not necessarily bounded), whose domain is $D(A^\dagger) := R(A) + R(A)^\perp$, and satisfying

$$(\forall y \in D(A^\dagger)) \quad A^\dagger y := \mathrm{argmin}\,\{\|x\| \mid A^*Ax = A^*y\}.$$

It is easy to see that, whenever $y \in D(A^\dagger)$, the solution set of (27) is $A^\dagger y + \ker A$.

Lemma A.11. Let A be a bounded linear opertator from X to Y. Then, for every continuous function $\phi : [0, +\infty[\to \mathbb{R}$, we have $A\phi(A^*A) = \phi(AA^*)A$.

Proof. A simple induction argument shows that, for every $k \geq 0$, $A(A^*A)^k = (AA^*)^k A$. Taking linear combinations of this equality allows to see that, for every polynomial $P \in \mathbb{R}[X]$, $AP(A^*A) = P(AA^*)A$. Now, if ϕ is continuous on $[0, +\infty[$, it is in particular continuous on $[0, \|A\|^2]$, which is an interval containing the spectrum of both A^*A and AA^*. Thus, ϕ restricted to this interval can be written as the uniform limit of a sequence of polynomials. Passing to the limit (see [56, Thm. VI.32.1]) in the last equality gives the desired result. □

Lemma A.12. For all $b \in Y$, $r \in]0, +\infty[$, the following two properties are equivalent:

1. $(\exists x \in \ker A^\perp) \quad b = Ax, \quad \|x\| = r$
2. $(\exists y \in \operatorname{cl} R(A)) \quad b = \sqrt{AA^*}y, \quad \|y\| = r.$

Proof. It is shown in [42, Proposition 2.18] that $R(A) = R(\sqrt{AA^*})$, so it is enough to verify this implication:

$$(\forall (x,y) \in \ker A^\perp \times \operatorname{cl} R(A)) \quad Ax = \sqrt{AA^*}y \;\Rightarrow\; \|x\| = \|y\|.$$

Let (x, y) be such a pair. Since $Ax = \sqrt{AA^*}y$ and $y \in \operatorname{cl} R(A) = \ker \sqrt{AA^*}^\perp$, we deduce that $y = (\sqrt{AA^*})^\dagger Ax$. Therefore, since AA^* is self-adjoint, $(AA^*)^\dagger Ax = (A^*)^\dagger x$ (see [42, p.35]), and $A^*(A^*)^\dagger x = \operatorname{proj}(x; \ker A^\perp)$, we get

$$\|y\|^2 = \|(\sqrt{AA^*})^\dagger Ax\|^2 = \langle (AA^*)^\dagger Ax, Ax\rangle = \langle A^*(A^*)^\dagger x, x\rangle = \|x\|^2. \qquad \square$$

Proof of Lemma 5.5. Remind that $y^\dagger = Ax^\dagger = AA^\dagger y$ and let $\nu = \mu + 1/2$. Then, Lemma A.12 yields:

$$\begin{array}{rcll}
b \in A^{-1}Y_{\nu,\delta} & \Leftrightarrow & (\exists \omega \in \operatorname{cl} R(A)) \quad \|\omega\| \leq \delta, \quad Ab = y^\dagger + (AA^*)^\mu (AA^*)^{\frac{1}{2}}\omega & \text{with } \nu = \mu + 1/2, \\
& \Leftrightarrow & (\exists w \in \ker A^\perp) \quad \|w\| \leq \delta, \quad Ab = AA^\dagger y + (AA^*)^\mu Aw & \text{with Lemma A.12}, \\
& \Leftrightarrow & (\exists w \in \ker A^\perp) \quad \|w\| \leq \delta, \quad Ab = AA^\dagger y + A(A^*A)^\mu w & \text{with Lemma A.11}, \\
& \Leftrightarrow & (\exists w \in \ker A^\perp) \quad \|w\| \leq \delta, \quad b - x^\dagger - (A^*A)^\mu w \in \ker A & \\
& \Leftrightarrow & b \in X_{\mu,\delta}. &
\end{array}$$

□

Lemma A.13 (Interpolation inequality [42, p. 55])**.** For all $x \in X$ and $0 \leq \alpha < \beta$, we have

$$\|(A^*A)^\alpha x\| \leq \|(A^*A)^\beta x\|^{\frac{\alpha}{\beta}} \, \|x\|^{1-\frac{\alpha}{\beta}}.$$

Lemma A.14 (Powers of self-adjoint operators)**.** Let S be a bounded selfadjoint positive linear operator on a Hilbert space. Then, for all $\alpha > 0$, $\ker S = \ker S^\alpha$, and $\operatorname{cl} R(S^\alpha) = \operatorname{cl} R(S)$.

Proof. Given any $0 < \alpha < \beta$, we can write $S^\beta = S^{\beta-\alpha}S^\alpha$, from which we deduce that $\ker S^\alpha \subset \ker S^\beta$. This means that $(\ker S^\alpha)_{\alpha>0}$ is a nondecreasing family. To prove that this family is constant, it is enough to see that $\ker S^2 \subset \ker S$, which we verify now: If $x \in \ker S^2$, then $\|Sx\|^2 = \langle Sx, Sx\rangle = \langle S^2x, x\rangle = 0$, therefore $x \in \operatorname{Ker} S$. The conclusion follows from the fact that $\ker S^\perp = \operatorname{cl} R(S)$. □

A.5 Regularized inverse problems and proofs of Section 5.2

Proposition A.15. Let $K \subset \mathbb{R}^N$ be a closed cone and $S \in \mathcal{S}_+(\mathbb{R}^N)$. Then S is coercive on K if and only if $K \cap \ker S = \{0\}$.

Proof. The direct implication is immediate from Definition 5.14. For the reverse implication, let K be a closed cone such that $K \cap \ker S = \{0\}$. Since S is linear, we know that $d \mapsto \langle Sd, d\rangle$ is convex and continuous. So, using the compactness of $K \cap \mathbb{S}_{\mathbb{R}^N}$ we deduce that:

$$(\exists \bar{d} \in K \cap \mathbb{S}) \quad \inf_{d \in K \cap \mathbb{S}} \langle Sd, d\rangle = \langle S\bar{d}, \bar{d}\rangle. \tag{56}$$

Because $\bar{d} \in K$ and $\bar{d} \neq 0$, we deduce from our assumption that $\bar{d} \notin \operatorname{Ker} S$. Therefore, $\gamma := \langle S\bar{d}, \bar{d}\rangle > 0$, from which we deduce that S is γ-coercive on K. □

Definition A.16 (Cone enlargement)**.** Let $K \subset \mathbb{R}^N$ be a cone, and $\theta \in [0, \frac{\pi}{2}]$. We define the θ-enlargement of K as

$$K_\theta := \mathbb{R}\left\{x \in \mathbb{S}_{\mathbb{R}^N} \mid (\exists y \in K \cap \mathbb{S}_{\mathbb{R}^N})\ \arccos\left(|\langle x, y\rangle|\right) \leq \theta\right\}.$$

Lemma A.17. If K is a closed cone, then K_θ is a closed cone containing K for all $\theta \in [0, \frac{\pi}{2}]$.

Proof. By definition, K_θ is a cone containing K and $\Delta_\theta := \{x \in \mathbb{S}_{\mathbb{R}^N} \mid (\exists y \in K \cap \mathbb{S}_{\mathbb{R}^N})\ \arccos\left(|\langle x, y\rangle|\right) \leq \theta\}$ is compact, due to the compactness of $K \cap \mathbb{S}$. Since $0 \notin \Delta_\theta$, by compactness of Δ_θ, we deduce that $K_\theta = \mathbb{R}\Delta_\theta$ is a closed cone (see e.g. [48, Proposition A.1.1]). □

Proposition A.18. Let $S \in \mathcal{S}_+(\mathbb{R}^N)$ which is γ-coercive on a closed cone K. Then, for every $\gamma' \in]0, \gamma]$, S is γ'-coercive on K_θ, with $\theta := \arcsin\left(\frac{\gamma - \gamma'}{\|S\|}\right) \in [0, \frac{\pi}{2}[$.

Proof. Let θ and γ be as in the statement. Since S is γ-coercive on K, we see that $\gamma \leq \|S\|$, which guarantees that $\theta \in [0, \frac{\pi}{2}[$. Now, the fact that K_θ is closed (Lemma A.17) implies that $K_\theta \cap \mathbb{S}$ is compact in X, so we can use the same arguments as in (56) to deduce that there exists $\bar{d} \in K_\theta \cap \mathbb{S}_{\mathbb{R}^N}$ such that $\langle S\bar{d}, \bar{d}\rangle = \inf_{d \in K_\theta \cap \mathbb{S}_{\mathbb{R}^N}} \langle Sd, d\rangle$. Since $\bar{d} \in K_\theta$, there exists by definition of K_θ some $\bar{v} \in K \cap \mathbb{S}$ such that $\arccos(|\langle \bar{d}, \bar{v}\rangle|) \leq \theta$. We can use [62, Theorem 1] to write

$$|\langle S\bar{v}, \bar{v}\rangle - \langle S\bar{d}, \bar{d}\rangle| \leq \|S\| \sin \arccos(|\langle \bar{v}, \bar{d}\rangle|) \leq \|S\| \sin\theta. \tag{57}$$

Since $\bar{v} \in K \cap \mathbb{S}_{\mathbb{R}^N} \subset K_\theta \cap \mathbb{S}_{\mathbb{R}^N}$, we have $\langle S\bar{v}, \bar{v}\rangle \geq \langle S\bar{d}, \bar{d}\rangle$. Moreover, $\arccos(|\langle \bar{v}, \bar{d}\rangle|) \leq \theta$, so (57), implies

$$\langle S\bar{d}, \bar{d}\rangle \geq \langle S\bar{v}, \bar{v}\rangle - \|S\| \sin\theta \geq \gamma - \|S\| \sin\theta = \gamma'.$$

We deduce from the definition of $\bar{d}$ that S is γ'-coercive on K_θ. □

Proposition A.19. Let $C \subset \mathbb{R}^N$, and $\bar{x} \in C$.

i) For $\rho > 0$, C is ρ-prox-regular at $\bar{x}$ if and only if :

$$(\forall x \in C)(\forall \eta \in N_C(\bar{x}))\quad \langle \eta, x - \bar{x}\rangle \leq \frac{\rho}{2}\|\eta\|\|x - \bar{x}\|^2. \tag{58}$$

ii) If C is a C^2 manifold, then there exists $\delta, \rho > 0$ such that $C \cap \mathbb{B}(\bar{x}, \delta)$ is ρ-prox-regular.

Proof. Item i) : Definition 5.19 can be rewritten as $(\forall \eta \in N_C(\bar{x}) \cap \mathbb{S}_{\mathbb{R}^N})(\forall x \in C)\quad x \notin \mathbb{B}(\bar{x} + \frac{1}{\rho}\eta, \frac{1}{\rho})$, where the condition $x \notin \mathbb{B}(\bar{x} + \frac{1}{\rho}\eta, \frac{1}{\rho})$ is equivalent to, after developing the square:

$$\frac{1}{\rho^2} \leq \|x - \bar{x} - \frac{1}{\rho}\eta\|^2 = \|x - \bar{x}\|^2 + \frac{1}{\rho^2}\|\eta\|^2 - \frac{2}{\rho}\langle x - \bar{x}, \eta\rangle = \|x - \bar{x}\|^2 + \frac{1}{\rho^2} - \frac{2}{\rho}\langle x - \bar{x}, \eta\rangle.$$

The conclusion follows after cancelling and reorganizing the terms. Item ii) : Every C^2-manifold is prox-regular in the sense of [93, Def. 10.23 & Prop. 13.32]. Therefore, for every $\bar{x} \in C$, there exists $\delta, \rho > 0$ such that for every $x \in C \cap \mathbb{B}(\bar{x}, \delta)$, and for every $\eta \in N_C(\bar{x}) \cap \mathbb{S}_{\mathbb{R}^N}$, the inequality (58) holds [93, Exercice 13.31]. Conclusion follows from the fact that $N_C(\bar{x}) = N_{C \cap \mathbb{B}(\bar{x}, \delta)}(\bar{x})$.

□

Here is a needed result estimating locally the coercivity of an operator on a prox-regular set via its coercivity on the tangent cone.

Proposition A.20. Let $C \subset X$ be ρ-prox-regular at $\bar{x} \in C$. Let $S : X \to X$ be a bounded positive selfadjoint linear operator, being γ-coercive on $T_C(\bar{x})$. Then, for all $\gamma' \in]0, \gamma[$, there exists a cone $K \subset X$ such that S is γ'-coercive on K, and $C \cap \mathbb{B}_X(\bar{x}, \delta) \subset \bar{x} + K$, with $\delta = \frac{2(\gamma - \gamma')}{\rho\|S\|}$.

Proof. Let $\gamma' \in]0,\gamma[$ be fixed, and define $\theta := \arcsin((\gamma-\gamma')\|S\|^{-1}) \in]0,\frac{\pi}{2}[$. Let K_θ be the θ-enlargement of $T_C(\bar{x})$, then Proposition A.18 guarantees that S is γ'-coercive on K_θ. It remains to prove that there exists $\delta \in]0,+\infty[$ such that $C \cap \mathbb{B}(\bar{x},\delta) \subset \bar{x} + K_\theta$. Let $x \in C$. Because C is ρ-reached at $\bar{x}$, we know that $T_C(\bar{x})$ is a convex cone (use [44, Thm. 4.8.(12)] and the fact that C is locally closed at $\bar{x}$), so we can define $y := \mathrm{proj}(x-\bar{x}, T_C(\bar{x}))$, and $\eta := \mathrm{proj}(x-\bar{x}, N_C(\bar{x}))$. Using Moreau's Theorem [11, Thm. 6.30], we deduce that $\eta = x-\bar{x}-y$ with $\langle \eta, y\rangle = 0$. We define $\delta := \|x-\bar{x}\|$, and look for a condition on it so that $x \in \bar{x} + K_\theta$. For this to happen, it is enough to verify that

$$\langle x-\bar{x}, y\rangle \geq \cos(\theta)\|x-\bar{x}\|\|y\|. \tag{59}$$

Now, use Proposition A.19.i) together with the Cauchy-Schwarz inequality, and the polynomial inequality $X^2 - cX \geq c^2/4$, to write

$$\|y\|^2 = \|x-\bar{x}-\eta\|^2 \geq \|x-\bar{x}\|^2 + \|\eta\|^2 - \rho\|\eta\|\|x-\bar{x}\|^2 \geq \delta^2(1-\rho^2\delta^2/4).$$

We can use this inequality, together with the facts that $x-\bar{x} = y+\eta$ and $\langle y,\eta\rangle = 0$, to write

$$\langle x-\bar{x}, y\rangle^2 = \|y\|^4 \geq \|y\|^2\delta^2(1-\rho^2\delta^2/4).$$

This allows us to conclude that (59) holds as long as:

$$1-\rho^2\delta^2/4 \geq \cos(\theta)^2 \;\Leftrightarrow\; \rho^2\delta^2/4 \leq 1-\cos(\theta)^2 \;\Leftrightarrow\; \rho\delta/2 \leq \sin(\theta) = \frac{\gamma-\gamma'}{\|S\|}.$$

□

Proof of Proposition 5.17. Let $0 < \gamma' < \gamma$, and set $S := \operatorname{argmin} f$. Since h is of class C^2 around $\bar{x} \in S$, there exists some $\delta > 0$ such that for all $u \in \delta\mathbb{B}_X$, $\|\nabla^2 h(\bar{x}+u) - \nabla^2 h(\bar{x})\| \leq \gamma-\gamma'$. Notice that when $\nabla^2 h$ is Lipschitz continuous, we can take $\delta = (\gamma-\gamma')/L$. Also, if it is constant, we can just take $\delta = +\infty$ and $\gamma' = \gamma$. Let us show that f is 2-conditioned on $\Omega := \bar{x} + (K \cap \delta\mathbb{B}_X)$ with the constant $\gamma_{f,\Omega} = \gamma'$. Take $x \in \Omega \cap \operatorname{dom} g$ and use the optimality condition at $\bar{x} \in S$ and the convexity of g to obtain

$$f(x) - \inf f \;=\; g(x) - g(\bar{x}) + \langle \nabla h(\bar{x}), x-\bar{x}\rangle + h(x) - h(\bar{x}) - \langle \nabla h(\bar{x}), x-\bar{x}\rangle \geq h(x) - h(\bar{x}) - \langle \nabla h(\bar{x}), x-\bar{x}\rangle.$$

By Taylor's theorem applied to h, we deduce from the inequality above that there exists $y \in [x,\bar{x}]$ such that:

$$f(x) - \inf f \;\geq\; (1/2)\langle \nabla^2 h(\bar{x})(x-\bar{x}), x-\bar{x}\rangle + (1/2)\langle \left[\nabla^2 h(y) - \nabla^2 h(\bar{x})\right](x-\bar{x}), x-\bar{x}\rangle.$$

On the one hand, since $x \in \Omega$, we have that $x-\bar{x} \in K$. Thus, from the coercivity of $\nabla^2 h(\bar{x})$ we have

$$\langle \nabla^2 h(\bar{x})(x-\bar{x}), x-\bar{x}\rangle \geq \gamma\|x-\bar{x}\|^2.$$

On the other hand, we use the Cauchy-Schwarz inequality together with the definition of δ and the fact that $\|y-\bar{x}\| \leq \|x-\bar{x}\| < \delta$ to obtain

$$\langle \left[\nabla^2 h(y) - \nabla^2 h(\bar{x})\right](x-\bar{x}), x-\bar{x}\rangle \geq -(\gamma-\gamma')\|x-\bar{x}\|^2.$$

By combining the three previous inequalities, we deduce that

$$f(x) - \inf f \geq (\gamma'/2)\|x-\bar{x}\|^2. \tag{60}$$

This implies that $(\bar{x}+K) \cap \operatorname{argmin} f = \{\bar{x}\}$, and the statement follows from $\|x-\bar{x}\| \geq \mathrm{dist}\,(x;S)$. □

References

[1] P.-A. Absil, R. Mahony and B. Andrews, *Convergence of the iterates of descent methods for analytic cost functions*, SIAM Journal on Optimization, **16**, pp. 531–547, 2005.

[2] F.J. Aragón Artacho and M.H. Geoffroy, *Characterization of metric regularity of subdifferentials*, Journal of Convex Analysis, **15**(2), pp.365–380, 2008.

[3] H. Attouch and J. Bolte, *On the convergence of the proximal algorithm for nonsmooth functions involving analytic features*, Mathematical Programming, **116**(1-2), pp. 5–16, 2009.

[4] H. Attouch, J. Bolte, P. Redont and A. Soubeyran, *Proximal alternating minimization and projection methods for nonconvex problems. An approach based on the Kurdyka-Łojasiewicz inequality*, Mathematics of Operations Research, **35**(2), pp. 438–457, 2010.

[5] H. Attouch, J. Bolte and B.F. Svaiter, *Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods*, Mathematical Programming, **137**(1-2), pp. 91–129, 2013.

[6] H. Attouch and R. Wets, *Quantitative stability of variational systems II, a framework for nonlinear conditioning*, SIAM Journal on Optimization, **3**(2), pp. 359–381, 1993.

[7] D. Azé and J.-N. Corvellec, *Nonlinear local error bounds via a change of metric*, Journal of Fixed Point Theory and Applications, **16**(1), pp. 351–372, 2014.

[8] J.-B. Baillon, *Un exemple concernant le comportement asymptotique de la solution du problÃ¨me* $du/dt + \partial\vartheta \ni 0$, Journal of Functional Analysis, **28**(3), pp. 369–376, 1978.

[9] A.S. Bandeira, E. Dobriban, D.G. Mixon, and W.F. Sawin, *Certifying the restricted isometry property is hard*, IEEE Transactions on Information Theory, **59**(6), pp. 3448–3450, 2013.

[10] H.H. Bauschke and J.M. Borwein, *On the convergence of von Neumann's alternating projection algorithm for two sets*, Set-Valued Analysis, **1**(2), pp. 185–212, 1993.

[11] H.H. Bauschke and P. Combettes, *Convex analysis and monotone operator theory*, 2nd Edition, Springer, 2017.

[12] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM Journal on Imaging Sciences, **2**(1), pp. 183–202, 2009.

[13] P. Bégout, J. Bolte and M.A. Jendoubi, *On damped second-order gradient systems*, Journal of Differential Equations, **259**(7), pp. 3115–3143, 2015.

[14] J. Bolte, A. Daniilidis and A. Lewis, *The Łojasiewicz Inequality for Nonsmooth Subanalytic Functions with Applications to Subgradient Dynamical Systems*, SIAM Journal on Optimization, **17**(4), pp. 1205–1223, 2007.

[15] J. Bolte, A. Daniilidis, A.S. Lewis and M. Shiota, *Clarke subgradients of stratifiable functions*, SIAM Journal on Optimization, **18**(2), pp. 556–572, 2007.

[16] J. Bolte, A. Daniilidis, O. Ley and L. Mazet, *Characterizations of Łojasiewicz inequalities: Subgradient flows, talweg, convexity*, Transactions of the American Mathematical Society, **362**, pp. 3319–3363, 2010.

[17] J. Bolte, T.P. Nguyen, J. Peypouquet and B. Suter, *From error bounds to the complexity of first-order descent methods for convex functions*, Mathematical Programming **165**(2), pp. 471–507, 2017.

[18] J. Bolte, S. Sabach and M. Teboulle, *Proximal alternating linearized minimization for nonconvex and nonsmooth problems*, Mathematical Programming, **146**(1-2), pp. 459–494, 2013.

[19] J.F. Bonnans and A. Shapiro, *Perturbation Analysis of Optimization Problems*, Springer-Verlag, New York, 2000.

[20] K. Bredies, and D.A. Lorenz, *Linear convergence of iterative soft-thresholding. Journal of Fourier Analysis and Applications*, **14**(5-6), pp. 813–837, 2008.

[21] H. Brézis, *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*, North-Holland/Elsevier, New-York, 1973.

[22] H. Brézis,*On a characterization of flow-invariant sets*, Communications on Pure and Applied Mathematics, **23**(2), pp. 261–263, 1970.

[23] J. Burke and M.C. Ferris, *Weak Sharp Minima in Mathematical Programming*, SIAM Journal on Control and Optimization, **31**(5), pp. 1340–1359, 1993.

[24] L. Calatroni, G. Garrigos, L. Rosasco, and S. Villa, *Accelerated iterative regularization via dual diagonal descent*, preprint on arXiv:1912.12153, 2019.

[25] E.J. Candès, *The restricted isometry property and its implications for compressed sensing*, Comptes Rendus Mathematique, **346**(9-10), pp. 589-592, 2008.

[26] V. Chandrasekaran, B. Recht, P.A. Parillo and A.S. Willsky, *The Convex Geometry of Linear Inverse Problems*, Foundations of Computational Mathematics, **12**(6), pp. 805–849, 2012.

[27] E. Chouzenoux, J.-C. Pesquet and A. Repetti, *A block coordinate variable metric forward-backward algorithm*, Journal on Global Optimization, **66**, pp. 457–485, 2016.

[28] P.L. Combettes and J.-C. Pesquet, *Proximal splitting methods in signal processing*, in Fixed-point algorithms for inverse problems in science and engineering, Springer New York, 2011.

[29] O. Cornejo, A. Jourani and C. Zalinescu, *Conditioning and Upper-Lipschitz Inverse Subdifferentials in Nonsmooth Optimization Problems*, Journal of Optimization Theory and Applications, **95**(1), pp. 127–148, 1997.

[30] D. K. Crane and M. Gockenbach, *The Singular Value Expansion for Arbitrary Bounded Linear Operators*, Mathematics, **8**(8), 2020.

[31] I. Daubechies, M. Defrise and C. De Mol, *An iterative thresholding algorithm for linear inverse problems with a sparsity constraint*, Communications in Pure and Applied Mathematics, **57**(11), pp. 1413–1457, 2004.

[32] D. Davis and W. Yin, *Convergence rate analysis of several splitting schemes*,in: Splitting Methods in Communication, Imaging, Science, and Engineering, Springer International Publishing, 2014.

[33] E. De Vito, A. Caponnetto and L. Rosasco, *Model selection for regularized least-squares algorithm in learning theory*, Foundations of Computational Mathematics, **5**(1), pp. 59–85, 2005.

[34] Y. Yao, L. Rosasco and A. Caponnetto *On regularizationearly stopping in gradient descent learning*, Constructive Approximation **26**, pp. 289–315, 2007.

[35] R. DeVore, Approximation of functions, Approximation Theory, Proceedings Symp. Applied Mathematics, AMS 36 (1986) 1-20.

[36] A.L. Dontchev, A.S. Lewis and R.T. Rockafellar, *The Radius of Metric Regularity*, Transactions of the American Mathematical Society, **355**(2), pp. 493–517, 2003.

[37] A. Dontchev and T. Rockafellar, *Implicit functions and Solution Mappings*, Springer, New York, 2009.

[38] A. Dontchev and T. Zolezzi, *Well-posed Optimization Problems*, Springer-Verlag, Berlin,1993.

[39] D. Drusvyatskiy and A.D. Ioffe, *Quadratic growth and critical point stability of semi-algebraic functions*, Mathematical Programming, **153**(2) Ser. A, pp. 635–653, 2015.

[40] D. Drusvyatskiy and A.D. Lewis, *Error bounds, quadratic growth, and linear convergence of proximal methods*, Mathematics of Operations Research **43**, pp. 693–1050, 2018.

[41] D. Drusvyatskiy, B.S. Mordukhovich, and T.T.A. Nghia, *Second-order growth, tilt stability, and metric regularity of the subdifferential*, Journal of Convex Analysis, **21**(4), pp. 1165–1192, 2014.

[42] H. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Kluwer, Dordrecht, 1996.

[43] J. Fadili, J. Malick and G. Peyré, *Sensitivity Analysis for Mirror-Stratifiable Convex Functions*, SIAM Journal on Optimization, **28**(4), pp. 2975–3000, 2018.

[44] H. Federer, *Curvature Measures*, Transactions of the American Mathematical Society, **93**(3), pp. 418–491, 1959.

[45] M.C. Ferris, *Finite termination of the proximal point algorithm*, Mathematical Programming, **50**, pp. 359–366, 1991.

[46] P. Frankel, G. Garrigos and J. Peypouquet, *Splitting methods with variable metric for Kurdyka-Łojasiewicz functions and general convergence rates*, Journal of Optimization Theory and Applications, **165**(3), pp. 874–900, 2015.

[47] S. Foucart and H. Rauhut, *A mathematical introduction to compressive sensing*, Springer, 2013.

[48] G. Garrigos, *Descent dynamical systems and algorithms for tame optimization and multi-objective problems*, Ph.D. thesis, 2015. Available on https://tel.archives-ouvertes.fr/tel-01245406

[49] G. Garrigos, L. Rosasco and S. Villa, *Thresholding gradient methods in Hilbert spaces: support identification and linear convergence*, ESAIM Control Optimization and Calculus of Variations, **26**, 28 (20 pages) 2020.

[50] A.A. Goldstein, *Cauchy's method of minimization*, Numerische Mathematik, **4**(1), pp. 146–150, 1962.

[51] C. W. Groetsch, *Generalized inverses of linear operators: representation and approximation*, Dekker, 1977.

[52] O. Güler, *On the Convergence of the Proximal Point Algorithm for Convex Minimization*, SIAM Journal on Control and Optimization, **29**(2), pp. 403–419, 1991.

[53] A. Haraux and M.A. Jendoubi, *The Łojasiewicz gradient inequality in the infinite dimensional Hilbert space framework*, Journal of Functional Analysis, **260**(9), pp. 2826–2842, 2011.

[54] W.L. Hare and A.S. Lewis, *Identifying Active Constraints via Partial Smoothness and Prox-Regularity*, Journal of Convex Analysis, **11**(2), pp. 251–266, 2004.

[55] W.L. Hare and A.S. Lewis, *Identifying active manifolds*, Algorithmic Operations Research, **2**(2), pp. 75–82, 2007.

[56] G. Helmberg, *Introduction to Spectral Theory in Hilbert Space*, Elsevier, 1969.

[57] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex analysis and minimization algorithms I: Fundamentals*, Springer Science & Business Media, 1993.

[58] A.J. Hoffman, *On approximate solutions of systems of linear inequalities*, Journal of Research of the National Bureau of Standards, **49**(4), pp. 263–265, 1952.

[59] T. Hohage, *Lecture notes on inverse problems*, Vorlesungskript, University of Göttingen, Germany, 2002.

[60] K. Hou, Z. Zhou, A. M.-C. So and Z.-Q. Luo, *On the Linear Convergence of the Proximal Gradient Method for Trace Norm Regularization*, in: Advances in Neural Information Processing Systems, pp. 710–718, 2013.

[61] H. Karimi, J. Nutini and M. Schmidt, *Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-Łojasiewicz Condition*, in: Machine Learning and Knowledge Discovery in Databases (ECML PKDD). Lecture Notes in Computer Science, vol 9851. Springer, 2016.

[62] A.V. Knyazev and M.E. Argentati, *On proximity of Rayleigh quotients for different vectors and Ritz values generated by different trial subspaces*, Linear algebra and its applications, **415**(1), pp. 82–95, 2006.

[63] G.S. Ladde and V. Lakshmikantham, *On flow-invariant sets*, Pacific Journal of Mathematics, **51**(1), pp. 215–220, 1974.

[64] B. Lemaire, *About the Convergence of the Proximal Method*, in Advances in Optimization, Lecture Notes in Economics and Mathematical Systems, **382**, pp. 39–51, 1992.

[65] B. Lemaire, *Stability of the iteration method for non expansive mappings*, Serdica Mathematical Journal, **22**(3), pp. 331–340, 1996.

[66] B. Lemaire, *Well-posedness, conditioning and regularization of minimization, inclusion and fixed-point problems*, Pliska Studia Mathematica Bulgarica, **12**(1), pp. 71–84, 1998.

[67] D. Leventhal, *Metric subregularity and the proximal point method*, Journal of Mathematical Analysis and Applications, **360**(2), pp. 681–688, 2009.

[68] A.S. Lewis, *Active sets, nonsmoothness, and sensitivity*, SIAM Journal on Optimization, **13**(3), pp. 702–725, 2002.

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

[70] W. Li, *Error Bounds for Piecewise Convex Quadratic Programs and Applications*, SIAM Journal on Control and Optimization, **33**(5), pp. 1510–1529, 1995.

[71] G. Li, *Global error bounds for piecewise convex polynomials*, Mathematical Programming, **137**(1-2), Ser. A, pp. 37–64, 2013.

[72] G. Li and B. Mordukhovich, *Hölder Metric Subregularity with Applications to Proximal Point Method*, SIAM Journal on Optimization, **22**(4), pp. 1655–1684, 2012.

[73] G. Li, B. S. Mordukhovich and T. S. Pham, *New fractional error bounds for polynomial systems with applications to Holderian stability in optimization and spectral theory of tensors*, Mathematical Programming, **153**, pp. 333–362, 2015.

[74] G. Li and T.K. Pong, *Calculus of the exponent of Kurdika-Łojasiewicz inequality and its applications to linear convergence of first-order methods*, Foundations of Computational Mathematics, **18**, pp.1199-1232, 2018.

[75] J. Liang, J. Fadili and G. Peyré, *Local linear convergence of Forward–Backward under partial smoothness*, in: Advances in Neural Information Processing Systems, pp. 1970–1978, 2014.

[76] J. Liang, J. Fadili and G. Peyré, *Activity identification and local linear convergence of Forward-Backward-type methods*, SIAM Journal on Optimization **27**, pp. 408–437, 2017.

[77] J. Liang, J. Fadili and G. Peyré, *A Multi-step Inertial Forward–Backward Splitting Method for Non-convex Optimization*, in: Advances in Neural Information Processing Systems, pp. 4042–4050, 2016.

[78] J. Liu, S.J. Wright, C. Ré, V. Bittorf and S. Sridhar, *An Asynchronous Parallel Stochastic Coordinate Descent Algorithm*, Journal of Machine Learning Research, **16**(1), pp. 285–322, 2015.

[79] S. Łojasiewicz, *Une propriété topologique des sous-ensembles analytiques réels*, in: Les Équations aux Dérivées Partielles, Éditions du centre National de la Recherche Scientifique, Paris, pp. 87–89, 1963.

[80] R. Luke, *Prox-Regularity of Rank Constraint Sets and Implications for Algorithms*, Journal of Mathematical Imaging and Vision, **47**(3), pp. 231–238, 2013.

[81] Z. Q. Luo and P. Tseng, *Error bounds and convergence analysis of feasible descent methods: a general approach*. Annals of Operations Research, **46**(1), pp. 157–178, 1993.

[82] F. Luque, *Asymptotic Convergence Analysis of the Proximal Point Algorithm*, SIAM Journal on Control and Optimization, **22**(2), pp. 277–293, 1984.

[83] B. Merlet, M. Pierre, *Convergence to equilibrium for the backward Euler scheme and applications*, Commun. Pure Appl. Anal, **9**, pp. 685–702, 2010.

[84] S. Mosci, L. Rosasco,M. Santoro, A. Verri,and S. Villa, *Solving structured sparsity regularization with proximal methods*, in Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 418-433). Springer Berlin Heidelberg.

[85] I. Necoara, Y. Nesterov and F. Glineur, *Linear convergence of first order methods for non-strongly convex optimization*, Mathematical Programming, **175**, pp. 69–107,2019.

[86] J.-P. Penot, *Conditioning convex and nonconvex problems*, Journal of Optimization Theory and Applications, **93**(3), pp. 535–554, 1996.

[87] J. Peypouquet, *Convex optimization in normed spaces. Theory, methods and examples.*, Springer Science & Business media, 2015.

[88] R.A. Poliquin and R.T. Rockafellar, *Prox-regular functions in variational analysis*, Transactions of the American Mathematical Society, **348**(5), pp. 1805–1838, 1996.

[89] B.T. Polyak, *Gradient methods for minimizing functionals*, Zh. Vychisl. Mat. Mat. Fiz., **3**(4), pp. 643–653, 1963.

[90] B.T. Polyak, *Introduction to Optimization*, Optimization Software, New York, 1987.

[91] R.T. Rockafellar, *Monotone Operators and the Proximal Point Algorithm*, SIAM Journal on Control and Optimization, **14**(5), pp. 877–898, 1976.

[92] R.T. Rockafellar, *Convex Analysis*, Princeton University Press, 1996.

[93] R.T. Rockafellar and R. J.-B. Wets *Variational Analysis*, Springer Science & Business Media, 2009.

[94] S. Salzo, *The variable metric forward-backward splitting algorithm under mild differentiability assumptions*, SIAM Journal on Optimization, **27**(4), pp. 2153–2181, 2017.

[95] M. Schmidt, N. Le Roux and F. Bach, *Convergence Rates of Inexact Proximal-Gradient Methods for Convex Optimization*, in Advances in neural information processing systems, pp. 1458–1466, 2011.

[96] J.E. Spingarn, *Applications of the method of partial inverses to convex programming: Decomposition*, Mathematical Programming, **32**(2), pp. 199–223, 1985.

[97] J.E. Spingarn, *A projection method for least-squares solutions to overdetermined systems of linear inequalities*, Linear Algebra and its Applications, **86**, pp. 211–236, 1987.

[98] M. M. Vainberg, *Le problème de la minimisation des fonctionelles non linéaires*, C.I.M.E. IV ciclo (1970).

[99] S. Vaiter, G. Peyré, and J.M. Fadili, *Model consistency of partly smooth regularizers*, IEEE Transactions on Information Theory, **64**(3), pp. 1725–1737, 2017.

[100] S. Wright, *Identifiable Surfaces in Constrained Optimization*, SIAM Journal on Control and Optimization, **31**(4), pp. 1063–1079, 1993.

[101] C. Zalinescu, *Convex Analysis in General Vector Spaces*, Singapore: World Scientific, 2002.

[102] R. Zhang and J. Treiman, *Upper-Lipschitz Multifunction and Inverse Subdifferentials*, Nonlinear Analysis: Theory, Methods, and Applications, **24**, pp. 273–286, 1995.

[103] Z. Zhou, and A.M.-C. So, *A unified approach to error bounds for structured convex optimization problems*, Mathematical Programming **165**, pp. 689–728, 2017.

[104] Z. Zhou, Q. Zhang, and A.M.-C. So, *$\ell_{1,p}$-Norm Regularization: Error Bounds and Convergence Rate Analysis of First-Order Methods.*, in Proceedings of the 32nd International Conference on Machine Learning, pp. 1501–1510, 2015.

[105] T. Zolezzi, *On equiwellset minimum problems*, Appl. Math. Optim, **4**, pp. 209–223, 1978.