

First Order Methods beyond Convexity and Lipschitz Gradient Continuity with Applications to Quadratic Inverse Problems

Jérôme Bolte* Shoham Sabach† Marc Teboulle‡ Yakov Vaisbourd§

June 20, 2017

Abstract

We focus on nonconvex and nonsmooth minimization problems with a composite objective, where the differentiable part of the objective is *freed* from the usual and restrictive global Lipschitz gradient continuity assumption. This longstanding smoothness restriction is pervasive in first order methods (FOM), and was recently circumvented for *convex composite* optimization by Bauschke, Bolte and Teboulle, through a simple and elegant framework which captures, all at once, the geometry of the function and of the feasible set. Building on this work, we tackle genuine *nonconvex* problems. We first complement and extend their approach to derive a full extended descent lemma by introducing the notion of *smooth adaptable functions*. We then consider a Bregman-based proximal gradient methods for the nonconvex composite model with smooth adaptable functions, which is proven to globally converge to a critical point under natural assumptions on the problem's data. To illustrate the power and potential of our general framework and results, we consider a broad class of *quadratic inverse* problems with sparsity constraints which arises in many fundamental applications, and we apply our approach to derive new globally convergent schemes for this class.

2010 Mathematics Subject Classification: Primary 90C25; Secondary 26B25, 49M27, 52A41, 65K05.

Keywords: Composite nonconvex nonsmooth minimization, proximal-gradient algorithms, descent lemma, Non Euclidean distances, Bregman distance, global convergence, Kudyka-Łosiajewicz property, semi-algebraic functions, quadratic inverse problems, phase retrieval.

1 Introduction

The gradient method, forged by Cauchy about 170 years ago, is still at the heart of many fundamental schemes in modern computational optimization through many variants and relatives known as First Order Methods (FOM). These are currently the leading algorithms for solving large scale problems to medium accuracy. The focus and intensity of today research is mostly in the *convex* case: it goes from complexity to decomposition methods, from Lagrangian approaches to stochastic variants. For an appetizer, see for instance [15, 22, 24] and references therein, as well as the

*TSE (Université Toulouse I), Manufacture des Tabacs, 21 allée de Brienne, 31015 Toulouse, France. E-mail: jerome.bolte@tse-fr.eu.

†Faculty of Industrial Engineering, The Technion, Haifa, 32000, Israel. E-mail: ssabach@ie.technion.ac.il.

‡School of Mathematical Sciences, Tel-Aviv University, Ramat-Aviv 69978, Israel. E-mail: teboulle@post.tau.ac.il. This research was partially supported by the Israel Science Foundation, under ISF Grants 998-12 and 1844-16

§School of Mathematical Sciences, Tel-Aviv University, Ramat-Aviv 69978, Israel. E-mail: yakov.vaisbourd@gmail.com. This research is supported by a postdoctoral fellowship under ISF grants 998-12 and 1844-16

recent comprehensive text of Bertsekas [8] with many relevant up-to-date and annotated sources and references.

A crucial and standard assumption common to almost all FOM is that the gradient of the smooth part in a given objective function has to be globally Lipschitz continuous on the entire underlying space. This is a very restrictive assumption which has only been circumvented by line-search approaches and/or quite complex inner loops which are unavoidably distorting the efficiency and the complexity of the initial method.

Recently, Bauschke, Bolte and Teboulle (BBT) [4] solved this longstanding issue and dealt with non globally Lipschitz continuous gradient. Their framework is simple and far reaching, it is based on adapting the geometry to the objective through the Bregman distance paradigm. It allows them to derive a new Descent Lemma, whereby the usual upper quadratic approximation of a smooth function is replaced by a more general proximity measure which captures, all at once, the geometry of the function and the one of the feasible set. The corresponding FOM derived in [4] come with guaranteed complexity estimates and pointwise global convergence results for *convex composite* minimization. Very recently, the publication [26] developed independently similar ideas for analyzing the convergence of a Bregman proximal gradient applied to the convex composite model in Banach spaces.

This work is a major departure from the current convex composite model and hence with far reaching consequences. Our main goal is to extend the BBT framework to the nonconvex setting. We complement and extend their approach to study *nonconvex composite* minimization problems. These consist in minimizing the sum of *two nonconvex* functions: an extended valued function f and a continuously differentiable function g

$$(\mathcal{P}) \quad \inf \{f(x) + g(x) : x \in \overline{C}\},$$

where \overline{C} denotes the closure of C which is a nonempty, convex and open set in \mathbb{R}^d (see Section 3 for a more precise statement).

Thus, here not only the functions f and g are not convex, but we also consider problems where g *does not* admit a globally Lipschitz continuous gradient. The usual restrictive requirement of Lipschitz continuity of the gradient of g in problem (\mathcal{P}) is not needed, and replaced by a convexity condition which adapts to the geometry of the function g . To solve problem (\mathcal{P}) , we focus on a proximal-based gradient method which involves the non Euclidean distances of Bregman type, and which covers standard proximal-based gradient methods. We derive a class of proximal-based gradient algorithms which are proven to globally converge to a critical point of (\mathcal{P}) when $C = \mathbb{R}^d$ and when the data is semi-algebraic. Note that one of the earliest work describing and analyzing the *classical* proximal gradient method in the nonconvex setting goes back to the work of Fukushima and Milne [17]. See also [3, 11] for a full convergence analysis in the semi-algebraic setting, but which also imposed the usual restrictive global Lipschitz continuity of the gradient of g . These results are significantly improved in the present paper within the non Euclidean proximal framework.

To illustrate the power of our general framework and results, we consider a broad class of *quadratic inverse problems* with sparsity constraints which arise in many fundamental applications (see, for instance, [7] and reference therein). We apply our approach to derive new and simple provably convergent schemes, which to the best of our knowledge are the first globally convergent algorithms for this important class of problems.

Outline of the paper. The paper is organized as follows. We first complement and extend the BBT approach to derive a full Descent Lemma by introducing the notion of *smooth adaptable*

functions, see Section 2. The following section presents the mathematical tools necessary to handle the nonconvex setting. It describes the problem and the corresponding proximal- based gradient method for the nonconvex composite model, freed from the usual global Lipschitz gradient continuity restriction. The analysis is developed in Section 4 where the resulting scheme is proven to globally converge to a critical point under natural assumptions on the problem’s data when $C = \mathbb{R}^d$. We demonstrate the potential of our framework by showing in Section 5 how it can be successfully applied to a broad class of quadratic inverse problems with sparsity constraints, resulting in computationally simple and explicit iterative formulas. To make the paper self contained, we provide an appendix (see Appendix 6) which includes the relevant material and results for the convergence analysis of algorithm in the semi-algebraic setting.

Notation. We use standard notation and concepts which, unless otherwise specified can all be found in [23].

2 Smooth Adaptable Functions and a Descent Lemma

We begin by defining the notion of *smooth adaptable* functions. This notion is motivated by the recent work [4] in which a *Lipschitz-like/Convexity condition* was introduced. This condition allows to lift the usual smoothness of the gradient of a given convex function, and derive a *one sided* Descent Lemma, whereby the usual upper quadratic approximation of C^1 functions is given in terms of a more general proximity measure. Here, we extend and complement this notion by accommodating differentiable functions which are not necessarily convex, and we also derive a natural *two sided* Descent Lemma.

2.1 Preliminaries on Proximity Measures

We first introduce our setting with some notations and definitions that will be used throughout the paper.

Definition 2.1. (Kernel Generating Distance) Let C be a nonempty, convex and open subset of \mathbb{R}^d . Associated with C , a function $h : \mathbb{R}^d \rightarrow (-\infty, +\infty]$ is called a kernel generating distance if it satisfies the following:

- (i) h is proper, lower semicontinuous and convex, with $\text{dom } h \subset \overline{C}$ and $\text{dom } \partial h = C$.
- (ii) h is C^1 on $\text{int dom } h \equiv C$.

We denote the class of kernel generating distances by $\mathcal{G}(C)$.

Given $h \in \mathcal{G}(C)$, define the proximity measure $D_h : \text{dom } h \times \text{int dom } h \rightarrow \mathbb{R}_+$

$$D_h(x, y) := h(x) - [h(y) + \langle \nabla h(y), x - y \rangle].$$

The proximity measure D_h is the so-called *Bregman Distance* [12]. It measures the proximity between x and y . Indeed, thanks to the gradient inequality, one has

$$h \text{ is convex if and only if } D_h(x, y) \geq 0, \forall x \in \text{dom } h, y \in \text{int dom } h.$$

If in addition h is strictly convex, equality holds if and only if $x = y$. However, note that D_h is not symmetric in general, unless h is the energy kernel $h = (1/2) \|\cdot\|^2$, which yields the classical

squared Euclidean distance. For early foundation papers and key results on Bregman distances, associated proximal-based algorithms, as well as many examples of kernels h , generating Bregman distances, we refer the reader to [13, 25, 14, 16, 5].

Note that the structural form of D_h is also useful when h is *not* convex. It measures the discrepancy or error, between the value of h at a given point $x \in \text{dom } h$ from its linear approximation around $y \in \text{int dom } h$. In that case, obviously, the distance like property $D_h(x, y) \geq 0$ and equal to zero if and only if $x = y$, is no longer valid. However, D_h still enjoys two simple, but remarkable properties, which follows from elementary algebra:

- **The three point identity** [14, Lemma 3.1] For any $y, z \in \text{int dom } h$ and $x \in \text{dom } h$,

$$D_h(x, z) - D_h(x, y) - D_h(y, z) = \langle \nabla h(y) - \nabla h(z), x - y \rangle.$$

- **Linear Additivity** For any $\alpha, \beta \in \mathbb{R}$, and any functions h_1 and h_2 we have:

$$D_{\alpha h_1 + \beta h_2}(x, y) = \alpha D_{h_1}(x, y) + \beta D_{h_2}(x, y),$$

for all couple $(x, y) \in (\text{dom } h_1 \cap \text{dom } h_2)^2$ such that both h_1 and h_2 are differentiable at y .

2.2 Smooth Adaptivity and an Extended Descent Lemma

Throughout the paper we will work with a pair of functions (g, h) satisfying:

- (i) $h \in \mathcal{G}(C)$.
- (ii) $g : \mathbb{R}^d \rightarrow (-\infty, +\infty]$ is a proper and lower semicontinuous function with $\text{dom } h \subset \text{dom } g$, which is continuously differentiable on $C = \text{int dom } h$

Definition 2.2 (L-smooth adaptable). A pair (g, h) is called L -smooth adaptable on C if there exists $L > 0$ such that $Lh - g$ and $Lh + g$ are convex on C .

Note that the above definition holds for any convex function h which is C^1 on any open subset of $\text{dom } h$. The additional properties required in the class $\mathcal{G}(C)$ are not necessary. Only the convexity of $Lh \pm g$ plays a central role. However, for the sake of consistency with the algorithmic development and results that follow, we will always use $h \in \mathcal{G}(C)$.

The above definition allows for immediately obtaining the promised *two-sided* Descent Lemma which naturally complements and extends the one derived in [4, Lemma 1, p. 333].

Lemma 2.1 (Full Extended Descent Lemma). *The pair of functions (g, h) is L -smooth adaptable on C if and only if:*

$$|g(x) - g(y) - \langle \nabla g(y), x - y \rangle| \leq LD_h(x, y), \quad \forall x, y \in \text{int dom } h. \quad (2.1)$$

Proof. By using Definition 2.2, the pair (g, h) is L -smooth adaptable on C if and only if $Lh - g$ and $Lh + g$ are convex on $\text{int dom } h$. Therefore, thanks to the definition of the Bregman distance and its linear additivity property, this holds if and only if $LD_h(x, y) - D_g(x, y) = D_{Lh-g}(x, y) \geq 0$, and likewise for $D_{Lh+g}(x, y) \geq 0$, from which the result immediately follows. \square

Clearly, in the setting of [4], that is, when the function g is also assumed to be convex, the condition $Lh + g$ is convex, trivially holds, and hence redundant. In this case, the Nolips Descent Lemma given in [4, Lemma 1, p.333] is recovered, i.e., $D_g(x, y) \leq LD_h(x, y)$, though here g needs not be convex.

Using the structural definition of D_g (even though D_g is not necessarily a proximity measure à la Bregman, since here g is not convex) the full Descent Lemma reads compactly:

$$|D_g(x, y)| \leq LD_h(x, y), \quad \forall x, y \in \text{int dom } h.$$

In the special case when the set C is the whole space \mathbb{R}^d , and $h = (1/2) \|\cdot\|^2$, the classical Descent Lemma for a function g with an L -smooth gradient on \mathbb{R}^d , which provides lower and upper *quadratic* approximation for g , is recovered, i.e.,

$$|D_g(x, y)| \equiv |g(x) - g(y) - \langle \nabla g(y), x - y \rangle| \leq \frac{L}{2} \|x - y\|^2, \quad \forall x, y \in \mathbb{R}^d.$$

The new Descent Lemma allows for more general lower and upper approximations for g by exploiting the geometry of the set C through the use of the kernel function $h \in \mathcal{G}(C)$ and its associated Bregman proximity measure D_h .

Remark 2.1. The convexity requirement of $Lh + g$ (and therefore also the left-hand side inequality in (2.1)) can be written with respect to a different parameter $\ell \leq L$. For simplicity, we have used here the same parameter L which appears in the condition $Lh - g$.

Similarly to [4, Proposition 1, p. 334], it is easy to see that the convexity condition on $Lh - g$ (and likewise on $Lh + g$), admit various alternative reformulations which can facilitate its checking. In particular, when both g and h are twice continuously differentiable on C , the usual Hessian test can be used, see the applications section for interesting cases.

For the purposes of this paper, and throughout the rest (unless otherwise specified) it will be enough to consider only the condition $Lh - g$ convex, and its corresponding one-sided Descent Lemma. Thus, for convenience we adopt the following short hand terminology for such a pair (g, h) :

L-smad holds on C (SMooth ADaptable) if and only if

$$\exists L > 0 \text{ such that } Lh - g \text{ is convex on } \text{int dom } h \iff D_g(x, y) \leq LD_h(x, y), \quad \forall x, y \in \text{int dom } h.$$

We end this section by observing that the **L-smad** property is *invariant* when h is assumed σ -strongly convex, where $\sigma > 0$ stands for the strong convexity modulus, namely $h - (\sigma/2) \|x\|^2$ is convex. Indeed, since no convexity is needed/assumed for g , we obviously have with $\omega(\cdot) := (1/2) \|\cdot\|^2$:

$$Lh - g = L(h - \omega) - (g - L\omega) := L\bar{h} - \bar{g},$$

namely, **L-smad** holds on C for the new pair (\bar{g}, \bar{h}) .

3 The Problem and a Bregman Proximal Gradient Algorithm

Our aim is to solve the nonconvex problem (\mathcal{P}) with $C := \text{int dom } h$, that is,

$$\inf \{ \Psi(x) \equiv f(x) + g(x) : x \in \bar{C} \},$$

where the following assumptions on the problem's data are made throughout the paper.

Assumption A. (i) $h \in \mathcal{G}(C)$ with $\overline{C} = \overline{\text{dom } h}$.

(ii) $f : \mathbb{R}^d \rightarrow (-\infty, +\infty]$ is a proper and lower semicontinuous function with $\text{dom } f \cap C \neq \emptyset$.

(iii) $g : \mathbb{R}^d \rightarrow (-\infty, +\infty]$ is a proper and lower semicontinuous function with $\text{dom } h \subset \text{dom } g$, which is C^1 on C .

(iv) $v(\mathcal{P}) := \inf \{ \Psi(x) : x \in \overline{C} \} > -\infty$.

3.1 A Bregman Proximal Gradient Algorithm

To present the algorithm for tackling problem (\mathcal{P}) , we define for all $x \in \text{int dom } h$ and any step-size $\lambda > 0$, the *Bregman proximal gradient* mapping

$$\begin{aligned} T_\lambda(x) &:= \operatorname{argmin} \left\{ f(u) + \langle \nabla g(x), u - x \rangle + \frac{1}{\lambda} D_h(u, x) : u \in \overline{C} \right\} \\ &= \operatorname{argmin} \left\{ f(u) + \langle \nabla g(x), u - x \rangle + \frac{1}{\lambda} D_h(u, x) : u \in \mathbb{R}^d \right\}, \end{aligned} \quad (3.1)$$

where the second inequality follows from the fact that $\text{dom } h \subset \overline{C}$. Note that, since f is *nonconvex*, the mapping T_λ is *not* in general single-valued. This map emerges from the usual approach which consists of linearizing the differential part g around x , and regularize it with a proximal distance from that point. Clearly, with $h \equiv (1/2) \|\cdot\|^2$, the above boils down to the classical set-valued proximal gradient mapping.

We now briefly discuss the properties of the Bregman proximal mapping T_λ in the nonconvex setting. For that purpose, we make the following two additional assumptions.

Assumption B. The function $h + \lambda f$ is supercoercive for all $\lambda > 0$, that is,

$$\lim_{\|u\| \rightarrow \infty} \frac{h(u) + \lambda f(u)}{\|u\|} = \infty.$$

Assumption C. For all $x \in C$ we have

$$T_\lambda(x) \subset C, \quad \forall x \in C.$$

Assumption B is a quite standard coercivity condition, it is for instance automatically satisfied when \overline{C} is compact. On the other hand Assumption C can be shown to hold under a classical qualification condition which ensures the use of a partial sum rule [23, Corollary 10.9, p. 430] and which is formulated through the use of horizon subgradients

$$\partial^\infty f(x) \cap (-\partial^\infty h(x)) = \{0\}, \quad \forall x \in \mathbb{R}^d. \quad (3.2)$$

It also holds automatically when f is convex or when $C = \mathbb{R}^d$. Another approach to warrant Assumption C is to consider extending the notion of prox-boundedness as defined in [23, Chapter 1, p. 19-20] to Bregman proximal-based map and their envelopes. However, to keep our presentation simple and transparent, these technical issues will not be pursued here, since they are irrelevant in the context of this paper.

We have the following basic result.

Lemma 3.1 (Well-Posedness of T_λ). *Suppose that Assumptions A, B and C hold, and let $x \in \text{int dom } h$. Then, the set $T_\lambda(x)$ is a nonempty and compact subset of C .*

Proof. Fix any $x \in \text{int dom } h$, and $\lambda > 0$. For any $u \in \mathbb{R}^d$, we define the function

$$\Psi_h(u) := \lambda f(u) + \lambda \langle \nabla g(x), u - x \rangle + D_h(u, x), \quad (3.3)$$

so that $T_\lambda(x) = \text{argmin}_{u \in \mathbb{R}^d} \Psi_h(u)$. Using the definition of the Bregman distance, note that Ψ_h can be rewritten as:

$$\Psi_h(u) = \lambda f(u) + h(u) + \langle \gamma, u \rangle + \rho,$$

where $\gamma := \lambda \nabla g(x) - \nabla h(x) \in \mathbb{R}^d$ and $\rho := -h(x) - \langle \gamma, x \rangle \in \mathbb{R}$ are constant quantities. We now show that Ψ_h is level bounded on \mathbb{R}^d , i.e., $\lim_{\|u\| \rightarrow \infty} \Psi_h(u) = \infty$. Using the Cauchy-Schwarz inequality in the above definition of Ψ_h we obtain

$$\Psi_h(u) \geq \lambda f(u) + h(u) - \|\gamma\| \cdot \|u\| - |\rho| = \|u\| \left(\frac{\lambda f(u) + h(u)}{\|u\|} - \|\gamma\| - \frac{|\rho|}{\|u\|} \right).$$

Passing to the limit as $\|u\| \rightarrow \infty$, the supercoercivity of $h + \lambda f$ gives $\lim_{\|u\| \rightarrow \infty} \Psi_h(u) = \infty$. Therefore, since Ψ_h is also proper and lower semicontinuous, invoking the modern form of Weierstrass' theorem (see, e.g., [23, Theorem 1.9, p. 11]), it follows that the value $\inf_{\mathbb{R}^d} \Psi_h$ is finite, and the set $\text{argmin}_{u \in \mathbb{R}^d} \Psi_h(u) \equiv T_\lambda(x)$ is nonempty and compact. \square

Remark 3.1 (The case when f is convex). First note that in this case problem (\mathcal{P}) remains a *nonconvex* composite minimization which consists in minimizing the sum of a nonsmooth and convex function f with a nonconvex and continuously differentiable function g . However, in this case, under our Assumption A, the function Ψ_h which is proper and lower semicontinuous is now convex. It can then be shown (see [4, Lemma 2, p. 336]), that in this case the mapping T_λ is also single-valued from $\text{int dom } h$ to $\text{int dom } h$.

We are now ready to describe our algorithm for solving the nonconvex composite problem (\mathcal{P}) .

Bregman Proximal Gradient - BPG

Input. A function $h \in \mathcal{G}(C)$ with $C = \text{int dom } h$ such that **L-smad** holds on C .

Initialization. $x^0 \in \text{int dom } h$ and let $\lambda > 0$.

General Step. For $k = 1, 2, \dots$, compute

$$x^k \in \text{argmin} \left\{ f(x) + \left\langle x - x^{k-1}, \nabla g(x^{k-1}) \right\rangle + \frac{1}{\lambda} D_h(x, x^{k-1}) : x \in \overline{C} \right\}. \quad (3.4)$$

Under our standing Assumptions A, B and C the algorithm is well-defined by Lemma 3.1. In the next section we analyze its properties, and establish a global convergence result to a critical point of Ψ .

4 Convergence Analysis of BPG

Throughout this section, we take the following as our blanket assumption

- (i) **L-smad** holds on C .
- (ii) Assumptions A, B and C hold.

4.1 Properties of the Algorithm

Using the Descent Lemma (see Lemma 2.1), we easily obtain the following key estimation for the composite objective function Ψ , which will play an essential role to derive our main convergence results.

Lemma 4.1 (Sufficient Decrease Property). *For any $x \in \text{int dom } h$, and any $x^+ \in \text{int dom } h$ defined by*

$$x^+ \in \operatorname{argmin} \left\{ f(u) + \langle u - x, \nabla g(x) \rangle + \frac{1}{\lambda} D_h(u, x) : u \in \overline{C} \right\}, \quad (\lambda > 0), \quad (4.1)$$

we have

$$\lambda \Psi(x^+) \leq \lambda \Psi(x) - (1 - \lambda L) D_h(x^+, x). \quad (4.2)$$

In particular, with $0 < \lambda L < 1$, the sufficient decrease in the composite objective function value Ψ is ensured.

Proof. By the definition of global optimality for (4.1), we obtain by taking $u = x \in \text{int dom } h$, that

$$f(x^+) + \langle x^+ - x, \nabla g(x) \rangle + \frac{1}{\lambda} D_h(x^+, x) \leq f(x). \quad (4.3)$$

Invoking the Descent Lemma (see Lemma 2.1) for g we then get with the above,

$$\begin{aligned} g(x^+) + f(x^+) &\leq g(x) + \langle x^+ - x, \nabla g(x) \rangle + L D_h(x^+, x) + f(x^+) \\ &\leq g(x) + L D_h(x^+, x) + f(x) - \frac{1}{\lambda} D_h(x^+, x) \\ &= g(x) + f(x) - \left(\frac{1}{\lambda} - L \right) D_h(x^+, x), \end{aligned}$$

and hence with $\Psi = f + g$, the desired inequality for Ψ is proved, and the last statement immediately follows with $0 < \lambda L < 1$. \square

Remark 4.1. (i) When f is assumed convex, using the global optimality condition for the corresponding convex problem, which defines x^+ through (4.1), followed by using the three points identity property of a Bregman distance, and the **L-smad** property, one easily see that for any $x \in \text{int dom } h$, the inequality (4.2) can be strengthened and reads

$$\lambda (\Psi(x^+) - \Psi(u)) \leq D_h(u, x) - D_h(u, x^+) - (1 - \lambda L) D_h(x^+, x) - \lambda D_g(u, x), \quad \forall u \in \text{dom } h.$$

When g is also convex, then the last term $-\lambda D_g(u, x) \leq 0$, and the above result recovers the key estimation result proven in [4, Lemma 5, p. 340].

(ii) When g is *nonconvex* and the *Full Descent Lemma* (see Lemma 2.1) holds, i.e., there exists $L > 0$ with $Lh + g$ convex, then we have $-D_g(u, x) \leq L D_h(u, x)$ and hence the above inequality reduces to

$$\lambda (\Psi(x^+) - \Psi(u)) \leq (1 + \lambda L) D_h(u, x) - (1 - \lambda L) D_h(x^+, x) - D_h(u, x^+), \quad \forall u \in \text{dom } h.$$

The result given in Lemma 4.1 is valid for any $\lambda > 0$, and as seen, imposing the condition $0 < \lambda L < 1$ ensures a sufficient decrease in the composite objective function value Ψ . This fact yields the following result.

Proposition 4.1. *Let $\{x^k\}_{k \in \mathbb{N}}$ be a sequence generated by BPG with $0 < \lambda L < 1$. Then the following assertions hold:*

- (i) *The sequence $\{\Psi(x^k)\}_{k \in \mathbb{N}}$ is nonincreasing.*
- (ii) *$\sum_{k=1}^{\infty} D_h(x^k, x^{k-1}) < \infty$, and hence the sequence $\{D_h(x^k, x^{k-1})\}_{k \in \mathbb{N}}$ converges to zero.*
- (iii) *$\min_{1 \leq k \leq n} D_h(x^k, x^{k-1}) \leq \frac{\lambda}{n} \left(\frac{\Psi(x^0) - \Psi_*}{1 - \lambda L} \right)$, where $\Psi_* = v(\mathcal{P}) > -\infty$ (by Assumption A(iv)).*

Proof. (i) Fix $k \geq 1$. Under our assumptions and using the iterative step (3.4), we can apply Lemma 4.1 with $x = x^{k-1}$ and $x^+ = x^k$, to obtain

$$\lambda \left(\Psi(x^k) - \Psi(x^{k-1}) \right) \leq -(1 - \lambda L) D_h(x^k, x^{k-1}). \quad (4.4)$$

Thus, with $0 < \lambda L < 1$, we immediately obtain that the sequence $\{\Psi(x^k)\}_{k \in \mathbb{N}}$ is nonincreasing.

- (ii) Let n be a positive integer. Summing the above inequality from $k = 1$ to n we get

$$\sum_{k=1}^n D_h(x^k, x^{k-1}) \leq \frac{\lambda (\Psi(x^0) - \Psi(x^n))}{1 - \lambda L} \leq \frac{\lambda (\Psi(x^0) - \Psi_*)}{1 - \lambda L}, \quad (4.5)$$

where $\Psi_* = v(\mathcal{P}) > -\infty$. Taking the limit as $n \rightarrow \infty$, we obtain the first desired assertion (ii), from which we immediately deduce that $\{D_h(x^k, x^{k-1})\}_{k \in \mathbb{N}}$ converges to zero.

- (iii) From (4.5) we also obtain,

$$n \min_{1 \leq k \leq n} D_h(x^k, x^{k-1}) \leq \sum_{k=1}^n D_h(x^k, x^{k-1}) \leq \frac{\lambda (\Psi(x^0) - \Psi_*)}{1 - \lambda L},$$

which after division by n yields the desired result. \square

Clearly, with h being the energy kernel on $C \equiv \mathbb{R}^d$, the above proposition yields all the classical results for the nonconvex proximal gradient method (e.g., set $\lambda L = 1/2$). In particular, in this case one obtains that the corresponding classical gradient mapping, defined by $G(x^{k-1}) := \|x^{k-1} - T_\lambda(x^{k-1})\|$ (see [6, Theorem 2.3, p. 61]) converges to zero at a rate of $O(1/\sqrt{n})$.

Finally, note that thanks to the invariance of L -smad (cf. Section 2.2), we can assume that h is σ -strongly convex on C . In that case, for any nonempty set $S \subset \mathbb{R}^d$, let $\text{dist}(y, S) := \inf \{\|u - y\| : u \in S\}$. Since $x^k \in T_\lambda(x^{k-1})$, using the σ -strong convexity of h combined with Proposition 4.1(iii) we immediately obtain the following rate of convergence result for two successive iterates:

$$\min_{1 \leq k \leq n} \text{dist}^2(x^{k-1}, T_\lambda(x^{k-1})) \leq \min_{1 \leq k \leq n} \|x^k - x^{k-1}\|^2 \leq \frac{1}{n} \cdot \frac{\lambda (\Psi(x^0) - \Psi_*)}{\sigma (1 - \lambda L)}.$$

4.2 Global Convergence of BPG

Throughout this section we consider problem (\mathcal{P}) defined on $C \equiv \mathbb{R}^d$, namely

$$(\mathcal{P}) \quad v(\mathcal{P}) = \inf \left\{ f(x) + g(x) : x \in \mathbb{R}^d \right\}.$$

Here, throughout this subsection, we additionally assume that

Assumption D. (i) $\text{dom } h = \mathbb{R}^d$ and h is σ -strongly convex on \mathbb{R}^d .

(ii) ∇h and ∇g are Lipschitz continuous on any bounded subset of \mathbb{R}^d .

Noting that Assumption C is automatically fulfilled since $C = \mathbb{R}^d$. As usual we use the concept of limiting subdifferential for f so that thanks to Fermat's rule [23, Theorem 10.1, p. 422], the set of critical points of Ψ is given by:

$$\text{crit } \Psi = \left\{ x \in \mathbb{R}^d : 0 \in \partial\Psi(x) \equiv \partial f(x) + \nabla g(x) \right\}.$$

To prove the global convergence of the sequence $\{x^k\}_{k \in \mathbb{N}}$ generated by BPG to a critical point of Ψ , we apply the methodology developed in [11]. For the reader's convenience, and to make this paper self-contained, the main tools and relevant proofs are given in Appendix 6. First, we describe three key ingredients of the methodology [11].

Definition 4.1 (Gradient-like descent sequence). Let $F : \mathbb{R}^d \rightarrow (-\infty, \infty]$ be a proper and lower semicontinuous function. A sequence $\{x^k\}_{k \in \mathbb{N}}$ is called a *gradient-like descent sequence* for F if the following three conditions hold:

(C1) *Sufficient decrease property.* There exists a positive scalar ρ_1 such that

$$\rho_1 \left\| x^{k+1} - x^k \right\|^2 \leq F(x^k) - F(x^{k+1}), \quad \forall k \in \mathbb{N}.$$

(C2) *A subgradient lower bound for the iterates gap.* There exists a positive scalar ρ_2 such that

$$\left\| w^{k+1} \right\| \leq \rho_2 \left\| x^{k+1} - x^k \right\| \text{ for some } w^{k+1} \text{ in } \partial F(x^{k+1}), \quad \forall k \in \mathbb{N}.$$

(C3) Let \bar{x} be a limit point of a subsequence $\{x^k\}_{k \in \mathcal{K}}$, then $\limsup_{k \in \mathcal{K} \subset \mathbb{N}} F(x^k) \leq F(\bar{x})$.

The two conditions (C1) and (C2) are typical for any descent type algorithm (see, e.g., [1]). They are also the main tools to prove subsequential convergence, as recorded below. The condition (C3) is a minimal and very weak continuity requirement, which, in particular, holds when F is continuous. We denote by $\omega(x^0)$ the set of all limit points of $\{x^k\}_{k \in \mathbb{N}}$.

Lemma 4.2 (Abstract subsequence convergence). *Let $\{x^k\}_{k \in \mathbb{N}}$ be a bounded gradient-like descent sequence for F . Then, $\omega(x^0)$ is a nonempty and compact subset of $\text{crit } F$, and we have*

$$\lim_{k \rightarrow \infty} \text{dist} \left(x^k, \omega(x^0) \right) = 0. \tag{4.6}$$

In addition, the function F is finite and constant on $\omega(x^0)$.

Proof. See Appendix 6. □

Remark 4.2 (Boundedness of Sequences). Observe that the mere coercivity of F ensures that any gradient-like sequence is bounded.

The above, together with the so-called nonsmooth Kurdyka-Łojasiewicz (KL) property (see [9] and Appendix 6 for details) allows us to establish our main convergence results of BPG.

Theorem 4.1 (Convergence of BPG). *Let $\{x^k\}_{k \in \mathbb{N}}$ be a sequence generated by BPG which is assumed to be bounded and let $0 < \lambda L < 1$. The following assertions hold.*

- (i) **Subsequential convergence.** *Any limit point of the sequence $\{x^k\}_{k \in \mathbb{N}}$ is a critical point of Ψ .*
- (ii) **Global convergence.** *Suppose that Ψ satisfies the KL property on $\text{dom } \Psi$. Then, the sequence $\{x^k\}_{k \in \mathbb{N}}$ has finite length and converges to a critical point x^* of Ψ .*

Proof. (i) We first prove that the sequence $\{x^k\}_{k \in \mathbb{N}}$ is a gradient-like descent sequence; namely it satisfies the three conditions (C1), (C2) and (C3) as described in Definition 6.1, and the result will then be established by invoking Lemma 4.2.

From Proposition 4.1 (see (4.4)) we have that

$$\Psi(x^k) - \Psi(x^{k+1}) \geq \left(\frac{1}{\lambda} - L\right) D_h(x^{k+1}, x^k) \geq \left(\frac{1}{\lambda} - L\right) \frac{\sigma}{2} \|x^{k+1} - x^k\|^2, \quad (4.7)$$

where the last inequality follows from the σ -strong convexity of h (see Assumption D(i)). This proves that condition (C1) holds true.

Writing the optimality condition of the optimization problem which defines x^{k+1} (see (3.4)) yields that

$$0 \in \partial f(x^{k+1}) + \nabla g(x^k) + \frac{1}{\lambda} \left(\nabla h(x^{k+1}) - \nabla h(x^k) \right).$$

Therefore, by defining

$$w^{k+1} \equiv \nabla g(x^{k+1}) - \nabla g(x^k) + \frac{1}{\lambda} \left(\nabla h(x^k) - \nabla h(x^{k+1}) \right),$$

we obviously obtain that $w^{k+1} \in \partial \Psi(x^{k+1})$. Since $\{x^k\}_{k \in \mathbb{N}}$ is a bounded sequence and both ∇h and ∇g are Lipschitz continuous on bounded subsets of \mathbb{R}^d (see Assumption D(ii)), there exists $M > 0$ such that

$$\|w^{k+1}\| \leq \|\nabla g(x^{k+1}) - \nabla g(x^k)\| + \frac{1}{\lambda} \|\nabla h(x^k) - \nabla h(x^{k+1})\| \leq M \left(1 + \frac{1}{\lambda}\right) \|x^{k+1} - x^k\|.$$

This proves that condition (C2) also holds true.

Consider a subsequence $\{x^{n_k}\}_{n_k \in \mathbb{N}}$ which converges to some point x^* (there exists such a subsequence since the sequence $\{x^k\}_{k \in \mathbb{N}}$ is assumed to be bounded). Using (4.7) and Proposition 4.1(iii) we obtain that $\lim_{k \rightarrow \infty} \|x^k - x^{k-1}\| = 0$. Therefore, the sequence $\{x^{n_k-1}\}_{n_k \in \mathbb{N}}$ also converges to x^* . In addition, since h is continuously differentiable on \mathbb{R}^d we have that $\lim_{k \rightarrow \infty} D_h(x^*, x^{k-1}) = 0$. Now, from (3.4), it follows that

$$\begin{aligned} f(x^k) + \left\langle x^k - x^{k-1}, \nabla g(x^{k-1}) \right\rangle + \frac{1}{\lambda} D_h(x^k, x^{k-1}) &\leq f(x^*) + \left\langle x^* - x^{k-1}, \nabla g(x^{k-1}) \right\rangle \\ &\quad + \frac{1}{\lambda} D_h(x^*, x^{k-1}), \end{aligned}$$

that is,

$$f(x^k) \leq f(x^*) + \left\langle x^* - x^k, \nabla g(x^{k-1}) \right\rangle + \frac{1}{\lambda} D_h(x^*, x^{k-1}) - \frac{1}{\lambda} D_h(x^k, x^{k-1}).$$

Substituting k by n_k and letting $k \rightarrow \infty$, we obtain from Proposition 4.1(iii) and the facts mentioned above, that

$$\limsup_{k \rightarrow \infty} f(x^{n_k}) \leq f(x^*).$$

This proves condition (C3), and thanks to Lemma 4.2, the first item of the theorem follows. Moreover, we also obtain that Ψ is constant on $\omega(x^0)$.

(ii) Since Ψ is assumed to have the KL property, and we have just shown that Ψ is constant on $\omega(x^0)$, we can then apply Theorem 6.2 given in Appendix 6, with $F := \Psi$. This proves that the sequence $\{x^k\}_{k \in \mathbb{N}}$ has finite length, and is globally convergent to a critical point x^* of Ψ . \square

We end by noting that, when the data f and g are semi-algebraic, the convergence rate of the generated sequence described in Theorem 4.1 follows by invoking Theorem 6.3, see the Appendix 6 for details.

5 Application to Quadratic Inverse Problems

5.1 Motivation and Problem's Statement

This section illustrates the potential of our approach and results. To this end, we consider the broad class of problems which consists of solving approximately a system of *quadratic* equations and can be described as follows. Given a finite number of symmetric matrices $A_i \in \mathbb{R}^{d \times d}$, $i = 1, 2, \dots, m$, describing the model, and a vector of possibly noisy measurements $b \in \mathbb{R}^m$, the goal is to find $x \in \mathbb{R}^d$, that solves the following system

$$x^T A_i x \simeq b_i, \quad i = 1, 2, \dots, m.$$

This type of problems is a natural extension of the classical linear inverse problem which arise in the broad area of signal recovery, whereby linear measurements are replaced by *quadratic* measurements, see [7] and references therein. It also includes the well-known and fundamental class of phase retrieval problems as a special case, which has been, and is still intensively studied in the literature; see [20] for a very recent review on this problem and references therein.

Commonly, the system under consideration is underdetermined, thus some prior information on the original signal is incorporated into the model by means of some regularizer represented by a function f , possibly nonconvex, nonsmooth, and extended valued (to accommodate constraints). Adopting the usual least-squares model to measure the error, the problem can then be reformulated as the nonconvex optimization problem

$$(\text{QIP}) \quad \min \left\{ \Psi(x) := \frac{1}{4} \sum_{i=1}^m (x^T A_i x - b_i)^2 + \theta f(x) : x \in \mathbb{R}^d \right\},$$

where $\theta > 0$ plays the role of a penalty parameter controlling the trade-off between matching the data fidelity criteria versus its regularizer f .

Clearly, the nonconvex function $g : \mathbb{R}^d \rightarrow (-\infty, +\infty]$ defined by $g(x) := (1/4) \sum_{i=1}^m (x^T A_i x - b_i)^2$ is continuously differentiable on \mathbb{R}^d , and it *does not* admit a global Lipschitz continuous gradient, thus precluding the use of the classical proximal gradient method.

In the following, we demonstrate the power of our approach, illustrating how new and simple globally convergent schemes can be derived for the broad class of problems (QIP). In particular, we design two new algorithms: one for the unconstrained ℓ_1 regularized (QIP) model, and the other for the sparsity ℓ_0 constrained (QIP); both models being known to promote sparse solutions. Our algorithms are given through an explicit *closed form formula*, and are thus straightforward to implement. To the best of our knowledge, they appear to be the first globally convergent schemes for this class of problems.

5.2 Two Algorithms for Sparse Quadratic Inverse Problems

Throughout this section, the underlying space is $C \equiv \mathbb{R}^d$. Given the nonconvex function $g : \mathbb{R}^d \rightarrow \mathbb{R}$ defined by

$$g(x) = \frac{1}{4} \sum_{i=1}^m (x^T A_i x - b_i)^2,$$

we consider solving the nonconvex model (QIP) in the following two cases:

- (a) **A convex ℓ_1 -norm regularization.** $f : \mathbb{R}^d \rightarrow \mathbb{R}$ with $f(x) = \|x\|_1$.
- (b) **A nonconvex ℓ_0 -ball constraint.** $f : \mathbb{R}^d \rightarrow (-\infty, +\infty]$ with $f(x) = \delta_{\mathbb{B}_0^s}(x)$, the indicator of the ℓ_0 -ball, where for a positive integer $s < d$,

$$\mathbb{B}_0^s \equiv \{x : \|x\|_0 \leq s\},$$

and $\|x\|_0$ is the “ ℓ_0 norm” which counts the number of nonzero components in x .

To apply our approach on the (QIP) model in both cases (a) and (b), we first need to identify a suitable function $h \in \mathcal{G}(\mathbb{R}^d)$ such that **L-smad** holds for the pair (g, h) . Here, we use $h : \mathbb{R}^d \rightarrow \mathbb{R}$ given by

$$h(x) = \frac{1}{4} \|x\|_2^4 + \frac{1}{2} \|x\|_2^2.$$

Equipped with this h , we now show that **L-smad** holds, i.e., there exists $L > 0$ such that $Lh - g$ is convex on \mathbb{R}^d .

Lemma 5.1. *Let g and h as defined above. Then, for any L satisfying*

$$L \geq \sum_{i=1}^m \left(3 \|A_i\|^2 + \|A_i\| |b_i| \right),$$

the function $Lh - g$ is convex on \mathbb{R}^d .

Proof. Let $x \in \mathbb{R}^d$. Since g and h are C^2 on \mathbb{R}^d , in order to warrant the convexity of $Lh - g$ it is sufficient to find $L > 0$ such that $L\lambda_{\min}(\nabla^2 h(x)) \geq \lambda_{\max}(\nabla^2 g(x))$, where $\lambda_{\min}(M)$ and $\lambda_{\max}(M)$ denote the minimal and maximal eigenvalue of a matrix M , respectively.

By a straightforward computation we obtain that

$$\nabla^2 g(x) = \sum_{i=1}^m (2A_i x x^T A_i + (x^T A_i x - b_i) A_i) \quad \text{and} \quad \nabla^2 h(x) = \left(\|x\|_2^2 + 1 \right) I_d + 2x x^T,$$

and due to the well-known fact that $\lambda_{\max}(M) \leq \|M\|$ we can write

$$\lambda_{\max}(\nabla^2 g(x)) \leq \|\nabla^2 g(x)\| \leq \sum_{i=1}^m \left(3 \|A_i\|^2 \|x\|^2 + \|A_i\| |b_i| \right).$$

On the other hand, since $\nabla^2 h(x) \succeq (\|x\|_2^2 + 1) I_d$, we obtain that $\lambda_{\min}(\nabla^2 h(x)) \geq \|x\|_2^2 + 1$. Therefore, taking $L \geq \sum_{i=1}^m (3\|A_i\|^2 + \|A_i\| |b_i|)$ yields

$$\lambda_{\max}(\nabla^2 g(x)) \leq \sum_{i=1}^m (3\|A_i\|^2 \|x\|^2 + \|A_i\| |b_i|) \leq L (\|x\|_2^2 + 1) \leq L \lambda_{\min}(\nabla^2 h(x)),$$

which proves the desired result. \square

In order to apply our results (cf. Section 4.2) for both scenarios (a) and (b), we observe that h given above is 1-strongly convex on \mathbb{R}^d and it easy to see that AssumptionsA-D hold. Furthermore, the function g is a real polynomial, hence semi-algebraic, and both functions $\|x\|_0$ and $\|x\|_1$ is also semi-algebraic (see, [11, Appendix 5, p. 490]). Therefore, since the addition of semi-algebraic functions results in a semi-algebraic function, it follows that for both models (a) and (b), the objective Ψ is semi-algebraic, and BPG can be applied on the (QIP) model in each case, to produce a globally convergent sequence which converges to a critical point of Ψ .

To apply BPG, the main computational step require us to compute the Bregman proximal gradient map (see (3.1)):

$$T_\lambda(x) = \operatorname{argmin} \left\{ f(u) + \langle \nabla g(x), u - x \rangle + \frac{1}{\lambda} D_h(u, x) : u \in \mathbb{R}^d \right\} \quad (\lambda > 0).$$

We now show that for both scenarios (a) and (b), this step yields an *explicit* closed form formula.

Before doing so, we introduce some convenient notations and recall some well known operators that will be used in the rest of this section. Let $\lambda > 0$ and fix any $x \in \mathbb{R}^d$. Define

$$p \equiv p_\lambda(x) = \lambda \nabla g(x) - \nabla h(x) \quad (\text{for simplicity we often drop } \lambda, x). \quad (5.1)$$

For the pair (g, h) given above, a direct computation of their gradients gives $p_\lambda(x)$. Now, omitting constant terms in the above expression defining T_λ , we can write

$$T_\lambda(x) = \operatorname{argmin} \left\{ \lambda f(u) + \langle p_\lambda(x), u \rangle + h(u) : u \in \mathbb{R}^d \right\}. \quad (5.2)$$

We now recall two very well-known operators that will be used to compute the resulting T_λ .

Soft-thresholding (with parameter τ). For any $y \in \mathbb{R}^d$,

$$\mathcal{S}_\tau(y) = \operatorname{argmin}_{x \in \mathbb{R}^d} \left\{ \tau \|x\|_1 + \frac{1}{2} \|x - y\|^2 \right\} = \max\{|y| - \tau, 0\} \operatorname{sgn}(y), \quad (5.3)$$

with the absolute value understood to be component-wise.

Hard-thresholding (with parameter τ). For any $y \in \mathbb{R}^d$,

$$\mathcal{H}_\tau(y) = \operatorname{argmin}_{x \in \mathbb{R}^d} \left\{ \|x - y\|^2 : x \in \mathbb{B}_0^\tau \right\} = \begin{cases} y_i, & i \leq \tau, \\ 0, & \text{otherwise,} \end{cases} \quad (5.4)$$

where we assumed, without the loss of generality, that the vector y is sorted in a decreasing order according to the absolute value of its entries (ties may be broken arbitrarily), i.e., $|y_1| \geq |y_2| \geq \dots \geq |y_n|$.

We are now ready to establish the promised formula of T_λ for the two cases (a) and (b).

Proposition 5.1 (Bregman Proximal Formula for the ℓ_1 -Norm Regularization). *Let $f = \|\cdot\|_1$, and for $x \in \mathbb{R}^d$, let $v(x) := \mathcal{S}_{\lambda\theta}(p_\lambda(x))$. Then, $x^+ = T_\lambda(x)$ is given by*

$$x^+ = -t^*v(x) = t^*\mathcal{S}_{\lambda\theta}(\nabla h(x) - \lambda\nabla g(x)),$$

where t^* is the unique positive real root of

$$t^3 \|v(x)\|_2^2 + t - 1 = 0,$$

which admits an explicit formula.¹

Proof. From the formulation of T_λ given in (5.2), we have

$$x^+ = \operatorname{argmin} \left\{ \lambda\theta \|u\|_1 + \langle p, u \rangle + \frac{1}{4} \|u\|_2^4 + \frac{1}{2} \|u\|_2^2 : u \in \mathbb{R}^d \right\}, \quad (5.5)$$

where we recall that $p = \lambda\nabla g(x) - \nabla h(x)$. By the first order global optimality condition for the strongly convex problem (5.5) we obtain for each $i = 1, 2, \dots, d$,

$$x_i^+ \left(1 + \|x^+\|_2^2 \right) + p_i + \lambda\theta\gamma_i = 0,$$

where $\gamma_i = \operatorname{sgn}(x_i^+)$ if $x_i^+ \neq 0$, and $\gamma_i \in [-1, 1]$ if $x_i^+ = 0$. For each $i = 1, 2, \dots, d$, set $v_i = p_i + \lambda\theta\gamma_i$. Then, from the above equation, $x_i^+ = -tv_i$ with $t > 0$, and

$$v_i \left(t^3 \|v\|_2^2 + t - 1 \right) = 0, \quad i = 1, 2, \dots, d.$$

Now consider the following two cases for any $i = 1, 2, \dots, d$:

- If $v_i = 0$, then $x_i^+ = 0$, and $\gamma_i \in [-1, 1]$, i.e., we obtain $x_i^+ = 0$ whenever $|p_i| \leq \lambda\theta$.
- If $v_i \neq 0$, then $x_i^+ \neq 0$, and $\gamma_i = \operatorname{sgn}(x_i^+) = \operatorname{sgn}(-tv_i) = -\operatorname{sgn}(v_i)$. Therefore, using $\gamma_i = (v_i - p_i) / (\lambda\theta)$, the latter relation reads $\lambda\theta\operatorname{sgn}(v_i) + v_i - p_i = 0$, which is nothing else but the following optimality condition for v :

$$v = \operatorname{argmin} \left\{ \lambda\theta \|\xi\|_1 + \frac{1}{2} \|\xi - p\|_2^2 : \xi \in \mathbb{R}^d \right\} \equiv \mathcal{S}_{\lambda\theta}(p).$$

where the last equality is from the definition of the soft threshold operator (see (5.3)).

To summarize, in all cases, we have thus obtained that $x^+ = -t^*\mathcal{S}_{\lambda\theta}(p)$, where t^* is the positive root of the cubic equation $t^3 \|v\|_2^2 + t - 1 = 0$, which is given by a closed form expression. \square

Now, we turn to the ℓ_0 -norm constrained version. First we recall the following useful result established in [21, Proposition 4.3 p. 79].

Lemma 5.2. *Given $0 \neq a \in \mathbb{R}^d$ and a positive integer $s < d$, then*

$$\max \{ \langle a, z \rangle : \|z\|_2 = 1, \|z\|_0 \leq s \} = \|\mathcal{H}_s(a)\|_2,$$

with optimal solution obtained at $z^* = \mathcal{H}_s(a) / \|\mathcal{H}_s(a)\|_2$.

¹The general solution of a cubic equation was apparently discovered first by Scipione del Ferro and published by Gerolamo Cardano in his *Ars magna* in 1545, see <https://www.britannica.com/topic/cubic-equation>

Proposition 5.2 (Bregman Proximal Formula for the ℓ_0 -Ball Constraint). *Let $f = \delta_{\mathbb{B}_0^s}$ and $x \in \mathbb{R}^d$. Then, $x^+ = T_\lambda(x)$ is given by*

$$x^+ = \sqrt{t^*} \|\mathcal{H}_s(p_\lambda(x))\|_2^{-1} \mathcal{H}_s(p_\lambda(x)),$$

where $\sqrt{t^*} \equiv \eta^*$ is the unique nonnegative real root of the cubic equation

$$\eta^3 + \eta - \|\mathcal{H}_s(p_\lambda(x))\|_2 = 0. \quad (5.6)$$

Proof. We use the same notations as in Proposition 5.1. Here $x^+ = T_\lambda(x)$ consists in finding x^+ which solves the nonconvex problem:

$$\min_{u \in \mathbb{R}^d} \left\{ \langle p, u \rangle + \frac{1}{4} \|u\|_2^4 + \frac{1}{2} \|u\|_2^2 : \|u\|_0 \leq s \right\}.$$

Introducing the new variable $t := \|u\|_2^2 \geq 0$, the latter is equivalent to solving

$$\min_{t \in \mathbb{R}_+} \left\{ \frac{1}{4} t^2 + \frac{1}{2} t + \min_{u \in \mathbb{R}^d} \left\{ \langle p, u \rangle : \|u\|_0 \leq s, \|u\|_2^2 = t \right\} \right\}. \quad (5.7)$$

Invoking Lemma 5.2 for the inner minimization with respect to u , one easily obtains:

$$\min_{u \in \mathbb{R}^d} \left\{ \langle p, u \rangle : \|u\|_0 \leq s, \|u\|_2^2 = t \right\} = -\sqrt{t} \mathcal{H}_s(-p) \equiv -\sqrt{t} \mathcal{H}_s(p),$$

with solution, $u^* = \sqrt{t} \|\mathcal{H}_s(p)\|_2^{-1} \mathcal{H}_s(p)$ where we used the fact that for the hard threshold operator: $\mathcal{H}_s(p) = \mathcal{H}_s(-p)$. Therefore, problem (5.7) reduces to solve

$$\min_{t \in \mathbb{R}_+} \left\{ \psi(t) := -\|\mathcal{H}_s(p)\|_2 \sqrt{t} + \frac{1}{4} t^2 + \frac{1}{2} t \right\}, \quad (5.8)$$

and it follows that $x^+ = \sqrt{t^*} \|\mathcal{H}_s(p)\|_2^{-1} \mathcal{H}_s(p)$, where t^* is an optimal solution of (5.8), thus proving the first part of the proposition. Now, if $\|\mathcal{H}_s(p)\|_2 = 0$ then clearly $t^* = 0$. Otherwise, with $\|\mathcal{H}_s(p)\|_2 > 0$, since $\lim_{t \rightarrow 0^+} \psi'(t) = -\infty$ and the above objective ψ is strongly convex over \mathbb{R}_+ , the problem (5.8) admits a unique minimizer $t^* > 0$ which satisfies the optimality condition:

$$\sqrt{t}(t+1) = \|\mathcal{H}_s(p)\|_2.$$

Thus, in either cases, the unique optimal solution t^* for problem (5.8) is the non-negative real root of the last equation, which can be rewritten as (5.6), and hence the desired result is proved. \square

Equipped with Lemma 5.1 together with Propositions 5.1 and 5.2, thanks to Theorem 4.1, we have now all the ingredients to formulate simple explicit algorithms in terms of the problem's data for solving a broad class of sparse quadratic inverse problems (SQIP). A thorough computational study of (SQIP), and some other related extensions will be developed in a separate paper.

6 Appendix: Global Convergence for KL Functions

The objective of this appendix is twofold. First, to make the paper self-contained, and second to outline here the general convergence mechanism as recently described in [11] on the so-called PALM algorithm, so that it can be used and applied to any given algorithm as described below. To this

end, let $F : \mathbb{R}^d \rightarrow (-\infty, +\infty]$ be a proper and lower semicontinuous function which is bounded from below and consider the problem

$$\inf \left\{ F(x) : x \in \mathbb{R}^d \right\}.$$

Consider a generic algorithm \mathcal{A} which generates a sequence $\{x^k\}_{k \in \mathbb{N}}$ via the following:

$$\text{start with any } x^0 \in \mathbb{R}^d \text{ and set } x^{k+1} \in \mathcal{A}(x^k), \quad k = 0, 1, \dots$$

The main goal is to prove that the *whole sequence* $\{x^k\}_{k \in \mathbb{N}}$, generated by the algorithm \mathcal{A} , converges to a critical point of F . For that purpose we first outline the three key ingredients of the forthcoming methodology.²

Definition 6.1 (Gradient-like descent sequence). A sequence $\{x^k\}_{k \in \mathbb{N}}$ is called a *gradient-like descent sequence* for F if the following three conditions hold:

(C1) *Sufficient decrease property.* There exists a positive scalar ρ_1 such that

$$\rho_1 \left\| x^{k+1} - x^k \right\|^2 \leq F(x^k) - F(x^{k+1}), \quad \forall k \in \mathbb{N}.$$

(C2) *A subgradient lower bound for the iterates gap.* There exists a positive scalar ρ_2 such that

$$\left\| w^{k+1} \right\| \leq \rho_2 \left\| x^{k+1} - x^k \right\|, \quad w^{k+1} \in \partial F(x^{k+1}), \quad \forall k \in \mathbb{N}.$$

(C3) Let \bar{x} be a limit point of a subsequence $\{x^k\}_{k \in \mathcal{K}}$, then $\limsup_{k \in \mathcal{K} \subset \mathbb{N}} F(x^k) \leq F(\bar{x})$.

The two conditions (C1) and (C2) are typical for any descent type algorithm (see, e.g., [1]), and basic to prove subsequential convergence. The condition (C3) is a minimal and weak requirement, which, in particular, holds when F is continuous.

The set of all limit points of $\{x^k\}_{k \in \mathbb{N}}$ is defined by

$$\omega(x^0) := \left\{ \bar{x} \in \mathbb{R}^d : \exists \text{ an increasing sequence of integers } \{k_l\}_{l \in \mathbb{N}} \text{ such that } x^{k_l} \rightarrow \bar{x} \text{ as } l \rightarrow \infty \right\}.$$

The next result establish the promised subsequential convergence.

Lemma 6.1 (Subsequence Convergence). *Let $\{x^k\}_{k \in \mathbb{N}}$ be a gradient-like descent sequence for F which is assumed to be bounded. Then, $\omega(x^0)$ is a nonempty and compact subset of $\text{crit } F$, and we have*

$$\lim_{k \rightarrow \infty} \text{dist} \left(x^k, \omega(x^0) \right) = 0. \quad (6.1)$$

In addition, the objective function F is finite and constant on $\omega(x^0)$.

Proof. Since $\{x^k\}_{k \in \mathbb{N}}$ is bounded there is $x^* \in \mathbb{R}^d$ and a subsequence $\{x^{k_q}\}_{q \in \mathbb{N}}$ such that $x^{k_q} \rightarrow x^*$ as $q \rightarrow \infty$ and hence $\omega(x^0)$ is nonempty. Moreover, the set $\omega(x^0)$ is compact since it can be viewed as an intersection of compact sets. Now, from condition (C3) and the lower semicontinuity of F , we obtain

$$\lim_{q \rightarrow \infty} F(x^{k_q}) = F(x^*). \quad (6.2)$$

²To keep this appendix as a completely independent unit, we repeat here a definition and results already stated in Section 4.

On the other hand, from conditions (C1) and (C2), we know that there is $w^k \in \partial F(x^k)$, $k \in \mathbb{N}$, such that $w^k \rightarrow 0$ as $k \rightarrow \infty$. The closedness property³ of ∂F implies thus that $0 \in \partial F(x^*)$. This proves that x^* is a critical point of F , and hence (6.1) is valid.

To complete the proof, let $\lim_{k \rightarrow \infty} F(x^k) = l \in \mathbb{R}$. Then $\{F(x^{k_q})\}_{q \in \mathbb{N}}$ converges to l and from (6.2) we have $F(x^*) = l$. Hence the restriction of F to $\omega(x^0)$ equals l . \square

To achieve our main goal, i.e., to establish global convergence of the *whole* sequence, we need an additional assumption on the class of functions F : it must satisfy the so-called nonsmooth Kurdyka-Łojasiewicz (KL) property [9](see [18, 19] for smooth cases). We refer the reader to [10] for an in depth study of the class of KL functions, as well as references therein. We provide now the formal definition of the KL property and two important results.

Denote $[\alpha < F < \beta] := \{x \in \mathbb{R}^d : \alpha < F(x) < \beta\}$. Let $\eta > 0$, and set

$$\Phi_\eta = \{\varphi \in C^0[0, \eta) \cap C^1(0, \eta) : \varphi(0) = 0, \varphi \text{ concave and } \varphi' > 0\}.$$

Definition 6.2 (The nonsmooth KL property). A proper and lower semicontinuous function $F : \mathbb{R}^d \rightarrow (-\infty, +\infty]$ has the Kurdyka-Łojasiewicz (KL) property locally at $\bar{u} \in \text{dom } F$ if there exist $\eta > 0$, $\varphi \in \Phi_\eta$, and a neighborhood $U(\bar{u})$ such that

$$\varphi'(F(u) - F(\bar{u})) \text{dist}(0, \partial F(u)) \geq 1,$$

for all $u \in U(\bar{u}) \cap [F(\bar{u}) < F(u) < F(\bar{u}) + \eta]$.

Verifying the KL property of a given function might often be a difficult task. However, thanks to a fundamental result established in [9], it holds for the broad class of *semi-algebraic* functions.

Theorem 6.1. *Let $F : \mathbb{R}^d \rightarrow (-\infty, +\infty]$ be a proper and lower semicontinuous function. If F is semi-algebraic then it satisfies the KL property at any point of $\text{dom } F$.*

Our last ingredient is a key uniformization of the KL property proven in [11, Lemma 6, p. 478], which we record below.

Lemma 6.2 (Uniformized KL property). *Let Ω be a compact set and let $F : \mathbb{R}^d \rightarrow (-\infty, +\infty]$ be a proper and lower semicontinuous function. Assume that F is constant on Ω and satisfies the KL property at each point of Ω . Then, there exist $\varepsilon > 0$, $\eta > 0$ and $\varphi \in \Phi_\eta$ such that for all \bar{x} in Ω one has,*

$$\varphi'(F(x) - F(\bar{x})) \text{dist}(0, \partial F(x)) \geq 1, \tag{6.3}$$

and all $x \in \{x \in \mathbb{R}^d : \text{dist}(x, \Omega) < \varepsilon\} \cap [F(\bar{x}) < F(x) < F(\bar{x}) + \eta]$.

We can now conveniently summarize the convergence results of [11] through the following abstract convergence result.

Theorem 6.2 (Global convergence). *Let $\{x^k\}_{k \in \mathbb{N}}$ be a bounded gradient-like descent sequence for F . If F satisfies the KL property, then the sequence $\{x^k\}_{k \in \mathbb{N}}$ has finite length, i.e., $\sum_{k=1}^{\infty} \|x^{k+1} - x^k\| < \infty$ and it converges to $x^* \in \text{crit } F$.*

³Let $\{(x^k, u^k)\}_{k \in \mathbb{N}}$ be a sequence in $\text{graph}(\partial F)$ that converges to (x, u) as $k \rightarrow \infty$. By the very definition of $\partial F(x)$, if $F(x^k)$ converges to $F(x)$ as $k \rightarrow \infty$, then $(x, u) \in \text{graph}(\partial F)$.

Proof. Since $\{x^k\}_{k \in \mathbb{N}}$ is bounded there exists a subsequence $\{x^{k_q}\}_{q \in \mathbb{N}}$ such that $x^{k_q} \rightarrow \bar{x}$ as $q \rightarrow \infty$. In a similar way as in Lemma 4.2 we get that

$$\lim_{k \rightarrow \infty} F(x^k) = F(\bar{x}). \quad (6.4)$$

If there exists an integer \bar{k} for which $F(x^{\bar{k}}) = F(\bar{x})$ then condition (C1) would imply that $x^{\bar{k}+1} = x^{\bar{k}}$. A trivial induction show then that the sequence $\{x^k\}_{k \in \mathbb{N}}$ is stationary and the announced results are obvious. Since $\{F(x^k)\}_{k \in \mathbb{N}}$ is a nonincreasing sequence, it is clear from (6.4) that $F(\bar{x}) < F(x^k)$ for all $k > 0$. Again from (6.4) for any $\eta > 0$ there exists a nonnegative integer k_0 such that $F(x^k) < F(\bar{x}) + \eta$ for all $k > k_0$. From Lemma 4.2 we know that $\lim_{k \rightarrow \infty} \text{dist}(x^k, \omega(x^0)) = 0$. This means that for any $\varepsilon > 0$ there exists a positive integer k_1 such that $\text{dist}(x^k, \omega(x^0)) < \varepsilon$ for all $k > k_1$.

From Lemma 4.2, we know that $\omega(x^0)$ is nonempty and compact, the function F is finite and constant on $\omega(x^0)$. Hence, we can apply the Uniformization Lemma with $\Omega = \omega(x^0)$. Therefore, for any $k > l := \max\{k_0, k_1\}$, we have

$$\varphi' \left(F(x^k) - F(\bar{x}) \right) \text{dist} \left(0, \partial F(x^k) \right) \geq 1. \quad (6.5)$$

This makes sense since we know that $F(x^k) > F(\bar{x})$ for any $k > l$. From condition (C2) we get that

$$\varphi' \left(F(x^k) - F(\bar{x}) \right) \geq \frac{1}{\rho_2} \left\| x^k - x^{k-1} \right\|^{-1}. \quad (6.6)$$

For convenience, we define for all $p, q \in \mathbb{N}$ and \bar{x} the following quantity

$$\Delta_{p,q} := \varphi \left(F(x^p) - F(\bar{x}) \right) - \varphi \left(F(x^q) - F(\bar{x}) \right).$$

From the concavity of φ we get that

$$\Delta_{k,k+1} \geq \varphi' \left(F(x^k) - F(\bar{x}) \right) \left(F(x^k) - F(x^{k+1}) \right). \quad (6.7)$$

Combining condition (C1) with (6.6) and (6.7) yields, for any $k > l$, that

$$\Delta_{k,k+1} \geq \frac{\|z^{k+1} - z^k\|^2}{\rho \|z^k - z^{k-1}\|}, \quad \text{where } \rho := \rho_2 / \rho_1. \quad (6.8)$$

Using the fact that $2\sqrt{\alpha\beta} \leq \alpha + \beta$ for all $\alpha, \beta \geq 0$, we infer from the later inequality that

$$2 \left\| x^{k+1} - x^k \right\| \leq \left\| x^k - x^{k-1} \right\| + \rho \Delta_{k,k+1}. \quad (6.9)$$

Summing up (6.9) for $i = l + 1, \dots, k$ yields

$$\begin{aligned} 2 \sum_{i=l+1}^k \left\| x^{i+1} - x^i \right\| &\leq \sum_{i=l+1}^k \left\| x^i - x^{i-1} \right\| + \rho \sum_{i=l+1}^k \Delta_{i,i+1} \\ &\leq \sum_{i=l+1}^k \left\| x^{i+1} - x^i \right\| + \left\| x^{l+1} - x^l \right\| + \rho \sum_{i=l+1}^k \Delta_{i,i+1} \\ &= \sum_{i=l+1}^k \left\| x^{i+1} - x^i \right\| + \left\| x^{l+1} - x^l \right\| + \rho \Delta_{l+1,k+1} \end{aligned}$$

where the last inequality follows from the fact that $\Delta_{p,q} + \Delta_{q,r} = \Delta_{p,r}$ for all $p, q, r \in \mathbb{N}$. Since $\varphi \geq 0$, recalling the definition of $\Delta_{l+1,k+1}$, we thus have for any $k > l$ that

$$\sum_{i=l+1}^k \|x^{i+1} - x^i\| \leq \|x^{l+1} - x^l\| + \rho\varphi\left(F(x^{l+1}) - F(\bar{x})\right),$$

which implies that $\sum_{k=1}^{\infty} \|x^{k+1} - x^k\| < \infty$, i.e., $\{x^k\}_{k \in \mathbb{N}}$ is a Cauchy sequence and hence together with Lemma 4.2, we obtain the global convergence to a critical point is established. \square

Finally, we record a generic rate of convergence result (see [1]) which is also valid under the same assumptions as Theorem 6.2.

Theorem 6.3 (Rate of convergence). *Let $\{x^k\}_{k \in \mathbb{N}}$ be a bounded gradient-like descent sequence for F . Assume that F satisfies the KL property where the desingularizing function φ of F is of the following form*

$$\varphi(s) = cs^{1-\theta}, \quad c > 0, \theta \in [0, 1).$$

Let \bar{x} be the limit point of $\{x^k\}_{k \in \mathbb{N}}$. Then the following estimations hold:

- (i) If $\theta = 0$ then the sequence $\{x^k\}_{k \in \mathbb{N}}$ converges to \bar{x} in a finite number of steps.
- (ii) If $\theta \in (0, 1/2]$ then there exist $\omega > 0$ and $\tau \in [0, 1)$ such that

$$\|x^k - \bar{x}\| \leq \omega \tau^k.$$

- (iii) If $\theta \in (1/2, 1)$ then there exist $\omega > 0$ such that

$$\|x^k - \bar{x}\| \leq \omega k^{-\frac{1-\theta}{2\theta-1}}.$$

References

- [1] H. Attouch and J. Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Math. Program.*, 116(1) (2009), 5–16.
- [2] A. Auslender and M. Teboulle. Interior gradient and proximal methods for convex and conic optimization. *SIAM J. Optim.*, 16(3) (2006), 697–725.
- [3] H. Attouch, and J. Bolte, and B. F. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized GaussSeidel methods. *Mathematical Programming*, 137(1-2) (2013), 91–129.
- [4] H. H. Bauschke, and J. Bolte and M. Teboulle. A descent lemma beyond Lipschitz gradient continuity: first order methods revisited and applications. *Math. Oper. Res.*, 42(2) (2016), 330–348.
- [5] H. H. Bauschke and J. M. Borwein. Legendre functions and the method of Bregman projections. *J. Convex Anal.*, 4(1) (1997), 27–67.
- [6] A. Beck and M. Teboulle. Gradient-based algorithms with applications to signal recovery problems. In D. Palomar and Y. C. Eldar, editors, *Convex Optimization in Signal Processing and Communications*, pages 139–162. Cambridge University Press, 2009.

- [7] A. Beck and Y. Eldar. Sparsity constrained nonlinear optimization: optimality conditions and algorithms. *SIAM J. Optimization*, 23 (2012), 1480–1509.
- [8] D. P. Bertsekas. *Convex Optimization Algorithms*. Athena Scientific, Belmont Massachusetts, 2015.
- [9] J. Bolte, and A. Daniilidis, and A. S. Lewis. The Lojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM J. Optim.*, 17(4) (2007), 1205–1223
- [10] J. Bolte, and A. Daniilidis, and O. Ley, and L. Mazet. Characterizations of Lojasiewicz inequalities: subgradient flows, talweg, convexity. *Trans. Am. Math. Soc.*, 362 (2010), 3319–3363.
- [11] J. Bolte, and S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Math. Program.*, 146(1) (2009), 459–494.
- [12] L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *U.S.S.R. Comp. Math. & Math. Phys.*, 7 (1967), 200–217.
- [13] Y. Censor and S. A. Zenios. Proximal minimization algorithm with D-functions. *J. Optim. Theory Appl.*, 73 (1992), 451–464.
- [14] G. Chen and M. Teboulle. Convergence analysis of a proximal-like minimization algorithm using Bregman functions. *SIAM J. Optim.*, 3 (1993), 538–543.
- [15] P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *SIAM Multiscale Model. Simul.*, 4 (2005), 1168–1200.
- [16] J. Eckstein. Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming. *Math. Oper. Res.*, 18 (1993), 202–226.
- [17] M. Fukushima and H. Mine. A generalized proximal point algorithm for certain nonconvex minimization problems. *Internat. J. Systems Sci.*, 12 (1981), 989–1000.
- [18] K. Kurdyka. On gradients of functions definable in o-minimal structures. *Ann. Inst. Fourier*, 48(3) (1998), 769–783.
- [19] S. Lojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. In *Les Équations aux Dérivées Partielles*, pp. 87–89. Éditions du Centre National de la Recherche Scientifique, Paris 1963.
- [20] D. R. Luke. Phase retrieval. What’s new? *SIAG/OPT Views and News*, 25(1) (2017), 1–5.
- [21] R. Luss and M. Teboulle. Conditional gradient algorithms for rank-one matrix approximations with a sparsity constraint. *SIAM Rev.*, 55(1) (2013), 65–98.
- [22] D. P. Palomar and Y. C. Eldar. *Convex optimization in signal processing and communications*. Cambridge University Press, Cambridge, 2010.
- [23] R. T. Rockafellar, and R. Wets. *Variational Analysis*. Grundlehren der Mathematischen Wissenschaften, 317, Springer, 1998.

- [24] S. Sra, and S. Nowozin, and S. J. Wright. *Optimization for Machine Learning*. The MIT Press, Cambridge, 2011
- [25] M. Teboulle. Entropic proximal mappings with application to nonlinear programming. *Math. Oper. Res.*, 17 (1992), 670–690.
- [26] Q. Van Nguyen. Forward-backward splitting with Bregman distances. *Vietnam J. Math.*, (2017). <https://link.springer.com/article/10.1007/s10013-016-0238-3>.