

## Research Article

# Inexact Version of Bregman Proximal Gradient Algorithm

S. Kabbadj 

Department of Mathematics, Faculty of Sciences of Meknes, B.P. 11201, Meknes, Morocco

Correspondence should be addressed to S. Kabbadj; kabbajsaid63@yahoo.com

Received 31 August 2019; Revised 19 January 2020; Accepted 21 January 2020; Published 1 April 2020

Academic Editor: Allan Peterson

Copyright © 2020 S. Kabbadj. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The Bregman Proximal Gradient (BPG) algorithm is an algorithm for minimizing the sum of two convex functions, with one being nonsmooth. The supercoercivity of the objective function is necessary for the convergence of this algorithm precluding its use in many applications. In this paper, we give an inexact version of the BPG algorithm while circumventing the condition of supercoercivity by replacing it with a simple condition on the parameters of the problem. Our study covers the existing results, while giving other.

## 1. Introduction

We consider the following minimization problem:

$$\inf \{ \Psi(x) := f(x) + g(x) : x \in \mathbb{R}^d \}. \quad (1)$$

where  $f$  is a convex proper lower-semicontinuous (l.s.c.) function and  $g$  is a convex continuously differentiable function. This problem arises in many applications including compressed sensing [1], signal recovery [2], and phase retrieve problem [3]. One classical algorithm for solving this problem is the proximal gradient (PG) method:

$$x^n := \operatorname{argmin} \left\{ f(u) + \langle \nabla g(x^{n-1}), u \rangle + \frac{1}{2\lambda_n} \|u - x^{n-1}\|^2 \right\} \\ n \in \mathbb{N}^*. \quad (2)$$

where  $\lambda_n$  is the stepsize on each iteration. The Proximal Gradient Method and its variants [4–14] have been one hot topic in optimization field for a long time due to their simple forms. A central property required in the analysis of gradient methods is that of the Lipschitz continuity of the gradient of the smooth part  $g$ . However, in many applications, the differentiable function does not have such a property, e.g., in the broad class of Poisson inverse problems. In [15], by

introducing the Bregman distance [16] generated by some reference convex function  $h$  defined by

$$D_h(x, y) = h(x) - h(y) - \langle x - y, \nabla h(y) \rangle, \quad (3)$$

the authors could replace the intricate question of Lipschitz continuity of gradients by a convex condition easy to verify, which we call below LC property. Thereby, they proposed and studied the algorithm called NoLips defined by

$$x^n = \operatorname{argmin} \left\{ f(u) + \langle \nabla g(x^{n-1}), u \rangle + \frac{1}{\lambda_n} D_h(u, x^{n-1}) \right\} \\ n \in \mathbb{N}^*. \quad (4)$$

where  $g = 0$ . Equation (4) is the Bregman Proximal (BP) studied in [17–21].

In this article, we give an inexact version of the (BPG) defined by

$$x^n \in \varepsilon_n - \operatorname{argmin} \left\{ f(u) + \langle \nabla g(x^{n-1}), u \rangle + \frac{1}{\lambda_n} D_h(u, x^{n-1}) \right\}. \quad (5)$$

While circumventing the condition of supercoercivity required in [15, 22] by replacing it with a simple condition on the parameters of the problem, our study covers the existing results, while giving others.

Our notation is fairly standard,  $\langle \cdot, \cdot \rangle$  is the scalar product on  $R^d$ , and the associated norm  $\|\cdot\|$ . The closure of the set  $C$  (relative interior) is denoted by  $\overline{C}$  ( $\text{ri}C$ , respectively). For any convex function  $f$ , we denote by

- (1)  $\text{dom } f = \{x \in R^d, f(x) < +\infty\}$  its effective domain
- (2)  $\partial_\varepsilon f(\cdot) = \{v, f(y) \geq f(\cdot) + \langle v, y - \cdot \rangle - \varepsilon, \forall y\}$  its  $\varepsilon$ -subdifferential
- (3)  $\text{argmin } f = \{x \in R^d, f(x) = \inf f\}$  its argmin  $f$
- (4)  $\varepsilon - \text{argmin } f = \{x \in R^d, f(x) \leq \inf f + \varepsilon\}$  its  $\varepsilon$ -argmin  $f$

## 2. Preliminary

In this section, we present the main results of the convergence of NoLips.

*Definition 1* (see [23]). Let  $C$  be a convex, not empty of  $R^d$ .

- (i) A convex function  $h: R^d \rightarrow ]-\infty, +\infty]$  is of Legendre on  $C$  if it verifies the three following conditions:
  - (a)  $C = \text{int}(\text{dom } h)$
  - (b)  $h$  is differentiable on  $C$
  - (c)  $\lim_{\|x_i\| \rightarrow \infty} \|\nabla h(x_i)\| = +\infty$ , for any sequence  $\{x_i\}$  of  $C$  that converges towards a boundary point of  $C$
- (ii) The class of strictly convex functions verifying a, b, and c is called the class of Legendre's functions on  $C$  and denoted by  $\mathcal{L}(C)$ .

*Definition 2* (see [23]). Let  $F: R^d \rightarrow ]-\infty, +\infty]$ ; we say that  $F$  is supercoercive if

$$\liminf_{\|x\| \rightarrow \infty} \frac{F(x)}{\|x\|} = \infty. \quad (6)$$

Consider the following assumptions.

*Assumption 1*

- (i)  $h: X \subset R^d \rightarrow ]-\infty, +\infty]$  is of Legendre type
- (ii)  $g: X \rightarrow ]-\infty, +\infty]$  is convex proper l.s.c. with  $\text{dom } h \subset \text{dom } g$ , which is differentiable on  $\text{int}(\text{dom } h)$
- (iii)  $f: X \rightarrow ]-\infty, +\infty]$  is convex proper lower semi-continuous (l.s.c.)
- (iv)  $\text{dom } f \cap \text{int}(\text{dom } h) \neq \emptyset$
- (v)  $\inf\{\Psi(x), x \in \overline{\text{dom } h}\} > -\infty$

We consider the following minimization problem:  $(P)$ :  $\inf\{\Psi(x) := f(x) + g(x) : x \in \overline{\text{dom } h}\}$ .

Let the operator  $T_\lambda$  be defined by

$$T_\lambda(x) = \text{argmin}\left\{f(u) + \langle \nabla g(x), u - x \rangle + \frac{1}{\lambda} D_h(u, x)\right\}. \quad (7)$$

**Lemma 1** (well-posedness of the method). *Under Assumption 1, suppose one of the following assumptions holds:*

- (i)  $\text{argmin}\{\Psi(x), x \in \overline{\text{dom } h}\}$  is nonempty and compact
- (ii)  $\forall \lambda > 0, h + \lambda f$  is supercoercive and the map  $T_\lambda$  defined in (7) is nonempty and single-valued from  $\text{int}(\text{dom } h)$  to  $\text{int}(\text{dom } h)$ .

*Definition 3.* The couple  $(g, h)$  verified a Lipschitz-like/Convexity Condition (LC) if  $\exists L > 0$  with Lh-g convex on  $\text{int}(\text{dom } h)$ .

By posing

$$\text{prox}_{\lambda f}^h(x) := \text{argmin}\left\{f(u) + \frac{1}{\lambda} D_h(u, x)\right\}, \quad (8)$$

they showed that

$$T_\lambda(x) = \text{prox}_{\lambda f}^h \circ \text{prox}_{\lambda p}^h(x), \quad (9)$$

where  $p(u) = \langle \nabla g(x), u \rangle$ . The operator  $T_\lambda$  thus appears as composed of two operators prox. The NoLips algorithm then becomes

$$x^n = T_{\lambda_n}(x^{n-1}), \quad \forall n \in N^*. \quad (10)$$

*Assumption 2*

- (i)  $\text{argmin}\{\Psi(x), x \in \overline{\text{dom } h}\}$  is nonempty and compact or  $\forall \lambda > 0, h + \lambda f$  is supercoercive
- (ii) For every  $x \in \text{int}(\text{dom } h)$  and  $r \in R$ , the level set  $L_1(x, r) = \{y \in \text{int}(\text{dom } h) : D_h(x, y) \leq r\}$  is bounded
- (iii) If  $\{x^n\}_n$  converges to some  $x$  in  $\text{int}(\text{dom } h)$ , then  $D_h(x, x^n) \rightarrow 0$
- (iv) Reciprocally, if  $x$  in  $\text{int}(\text{dom } h)$  and if  $\{x^n\}_n$  is such that  $D_h(x, x^n) \rightarrow 0$ , then  $x^n \rightarrow x$
- (v)  $\exists L > 0$  with Lh-g convex on  $\text{int}(\text{dom } h)$  (LC)

**Theorem 1** (Global Convergence). *Assume that*

- (i)  $\overline{\text{dom } h} = \text{dom } h$ .
- (ii)  $\sum \lambda_n = +\infty$  and Assumptions 1 and 2 are satisfied. Then, the sequence  $\{x^n\}_n$  converges to some solution  $x^*$  of  $(P)$ .

Our contribution is resumed in two essential points:

- (1) Improvements of some assumptions:
  - (a) Suppose  $f$  and  $g$  are both are convex (see [15, 22]), we show that we can reduce this hypothesis by supposing only that  $\Psi$  is convex, which allows to distinguish two interesting cases that are still not yet studied neither in the case of the BPG nor in the case PG:

- (i) The nonsmooth part  $f$  is possibly not convex and the smooth part  $g$  is convex
  - (ii) The nonsmooth part  $f$  is convex and the smooth part  $g$  is possibly not convex
- (b) The assumption is as follows:  $\operatorname{argmin}\{\Psi(x), x \in \overline{\operatorname{dom} h}\}$  is compact or  $\forall \lambda > 0, h + \lambda f$  is supercoercive.

This is a condition on  $f$  and  $g$  (see [15]), which precludes the application of NoLips for the functions  $\Psi$  non-supercoercive. In this work, we show that we can circumvent this condition by coupling the LC property with the bounded level sets as follows:

$$L_2(x, r) = \{y \in \bar{S}; D_h(y, x) \leq r\}. \tag{11}$$

It is a condition which relates to the parameter  $h$  and which is verified by most of the interesting Bregman distances.

(2) Inexact version of NoLips.

We propose an inexact version of NoLips called  $\varepsilon$ -NoLips which verifies

$$x^n = \varepsilon_n - \operatorname{argmin} \left\{ f(u) + \langle \nabla g(x^{n-1}), u \rangle + \frac{1}{\lambda_n} D_h(u, x^{n-1}) \right\},$$

$$n \in N^*. \tag{12}$$

The convergence result is established in Section 4. This study covers the convergence results given for PG and BPG, by giving new results, in particular, the convergence of the inexact version of the interior method with Bregman distance studied in [24]; this result has not been established until now.

We also note that the convergence of NoLips is given with the following condition:

$$\overline{\operatorname{dom} h} = \operatorname{dom} h. \tag{13}$$

It is for that and for the clarity of the hypothesis, we suppose in what follows that  $h: \bar{S} \rightarrow R$ , with  $S$  being an open convex set of  $R^d$ .

### 3. Main Results

In order to clarify the status of parameter  $h$ , we give the following definitions. Let  $S$  be a convex open subset of  $R^d$  and  $h: \bar{S} \rightarrow R$ . Let us consider the following hypotheses:

- $H_1$ :  $h$  is continuously differentiable on  $S$ .
- $H_2$ :  $h$  is continuous and strictly convex on  $\bar{S}$ .
- $H_3$ :  $\forall r \geq 0, \forall x \in \bar{S}$ , the sets below are bounded

$$L_2(x, r) = \{y \in S; D_h(y, x) \leq r\}. \tag{14}$$

- $H_4$ :  $\forall r \geq 0, \forall x \in S$ , the sets below are bounded

$$L_2(x, r) = \{y \in \bar{S}; D_h(y, x) \leq r\}. \tag{15}$$

$H_5$ : if  $\{x^n\}_n \subset S$ , then  $x^n \rightarrow x^* \in \bar{S}$ , so

$$D_h(x^*, x^n) \rightarrow 0. \tag{16}$$

$H_6$ : if  $\{x^n\}_n \subset S$ , then  $D_h(x^*, x^n) \rightarrow 0$ , so

$$x^n \rightarrow x^*. \tag{17}$$

*Definition 4*

- (i)  $h: \bar{S} \rightarrow R$  is a Bregman function on  $S$  or “ $D$ -function” if  $h$  verifies  $H_1, H_2, H_3, H_4, H_5$ , and  $H_6$ .
- (ii)  $D_h(\cdot, \cdot): \bar{S} \times S \rightarrow R$  such that  $\forall x \in \bar{S}, \forall y \in S$ :

$$D_h(x, y) = h(x) - h(y) - \langle x - y, \nabla h(y) \rangle. \tag{18}$$

Eq. (18) is called Bregman distance if  $h$  is a Bregman function. We put the following conditions:

- $A(S) = \{h: \bar{S} \rightarrow R \text{ verifying } H_1, H_2\}$
- $B(S) = \{h: \bar{S} \rightarrow R \text{ verifying } H_1, H_2, H_3, H_4, H_5 \text{ and } H_6\}$
- $\mathcal{E}(S) = \{h: \bar{S} \rightarrow R, \text{ the Legendre type of } S\}$

**Proposition 1.** Let  $h$  and  $h'$  verify  $H_1$ .

$$\forall \lambda, D_{\lambda h + h'}(\cdot, \cdot) = \lambda D_h(\cdot, \cdot) + D_{h'}(\cdot, \cdot). \tag{19}$$

**Lemma 2.**  $\forall h \in A(S), \forall a \in \bar{S}$ , and  $\forall b, c \in S$ :

$$D_h(a, b) + D_h(b, c) - D_h(a, c) = \langle a - b, \nabla h(c) - \nabla h(b) \rangle. \tag{20}$$

*Example 1.* If  $S_0 = R^d$  and  $h_0(x) = (1/2)\|x\|^2$ , then

$$D_{h_0}(x, y) = \frac{1}{2}\|x - y\|^2. \tag{21}$$

*Example 2.* If  $S_1 = R_{++}^d := \{x \in R^d / x_i > 0, i = 1, \dots, d\}$  and

$$h_1(x) = \sum_{i=1}^{i=d} x_i \log x_i, \quad \forall x \in \bar{S}_1, \tag{22}$$

with the convention  $0 \log 0 = 0$ , then  $\forall (x, y) \in \bar{S}_1 \times S_1$ :

$$D_{h_1}(x, y) = \sum_{i=1}^d x_i \log \frac{x_i}{y_i} + y_i - x_i. \tag{23}$$

*Example 3.* If  $S_2 = ]-1, 1]^d$  and  $h_2(x) = -\sum_{i=1}^{i=d} \sqrt{1 - x_i^2}$ , then  $D_{h_2}(x, y) = h_2(x) + \sum_{i=1}^d 1 - x_i y_i / y_i^2, \forall (x, y) \in \bar{S}_2 \times S_2$ .

**Proposition 2** (see [19]).  $h_i \in B(S_i) \cap \mathcal{E}(S_i)$ ,  $i = 0, 1, 2$ . We consider the following minimization problem:

$$(p): \inf\{\Psi(x) := f(x) + g(x) : x \in \bar{S}\}. \quad (24)$$

The following assumptions on the problem's data are made throughout the paper (and referred to as the blanket assumptions).

*Assumption 3*

- (i)  $h \in A(S) \cap \mathcal{E}(S)$
- (ii)  $g: R^d \rightarrow ]-\infty, +\infty]$  is proper (l.s.c.) with  $\bar{S} \subset \text{dom } g$ , which is and continuously differentiable on  $\bar{S}$
- (iii)  $f: R^d \rightarrow ]-\infty, +\infty]$  is proper (l.s.c.)
- (iv)  $\text{ri}(\text{dom } \Psi) \cap S \neq \emptyset$
- (v)  $\Psi^* := \inf\{\Psi(x), x \in \bar{S}\} > -\infty$

We consider the operator  $T_\lambda$  defined by  $\forall x \in S$ :

$$T_\lambda(x) = \underset{u \in \bar{S}}{\text{argmin}} \left\{ f(u) + \langle \nabla g(x), u - x \rangle + \frac{1}{\lambda} D_h(u, x), u \in \bar{S} \right\}. \quad (25)$$

We give in the following a series of lemmas allowing establishment of Theorem 2, which assures the well-posedness of the method proposed in Section 4.

**Lemma 3.**  $\forall \lambda > 0$  and  $\forall x \in S$ :

$$T_\lambda(x) = \underset{u \in \bar{S}}{\text{argmin}} \left[ \Psi(u) + \frac{1}{\lambda} D_{h-\lambda g}(u, x), u \in \bar{S} \right]. \quad (26)$$

*Proof.*  $\forall x \in S \subset \text{dom } g$ :

$$(25) \implies T_\lambda(x) = \underset{u \in \bar{S}}{\text{argmin}} \left\{ f(u) + g(x) + \langle \nabla g(x), u - x \rangle + \frac{1}{\lambda} D_h(u, x), u \in \bar{S} \right\}. \quad (27)$$

When  $\forall u \in \bar{S} \subset \text{dom } g$ , we have

$$\begin{aligned} f(u) + g(x) + \langle \nabla g(x), u - x \rangle + \frac{1}{\lambda} D_h(u, x) &= \Psi(u) + g(x) - g(u) + \langle \nabla g(x), u - x \rangle + \frac{1}{\lambda} D_h(u, x) \\ &= \Psi(u) - D_g(u, x) + \frac{1}{\lambda} D_h(u, x) \\ &= \Psi(u) + \frac{1}{\lambda} D_{h-\lambda g}(u, x). \end{aligned} \quad (28)$$

□

**Lemma 4.** If the pair  $(g, h)$  verified the condition (LC), then  $\exists L > 0$ ,  $\forall x \in S$ ,  $\forall u \in \bar{S}$ :

- (i)  $D_g(u, x) \leq LD_h(u, x)$
- (ii)  $\forall \lambda \in ]0, 1/L[$ ,  $D_{h-\lambda g}(u, x) \geq 0$

*Proof*

- (i) If  $Lh-g$  is convex on  $S$ , then

$$D_{Lh-g}(u, x) \geq 0, \quad \forall x \in S, \forall u \in S. \quad (29)$$

Let  $u \in (\bar{S}/S)$ . There exists a sequence  $\{u^n\}_n \subset S$  such that  $u^n \rightarrow u$ ; then, we have

$$D_{Lh-g}(u^n, x) \geq 0, \quad (30)$$

$Lh-g$  is continuous in  $\bar{S}$ ; then,  $D_{Lh-g}(u, x) \geq 0$ .

- (ii) Let  $\lambda \in ]0, 1/L[$ :

$$\begin{aligned} D_{Lh-g}(u, x) \geq 0 &\implies LD_h(u, x) \geq D_g(u, x) \\ &\implies \frac{1}{\lambda} D_h(u, x) \geq D_g(u, x) \\ &\implies D_{h-\lambda g}(u, x) \geq 0. \end{aligned} \quad (31)$$

□

**Lemma 5.** If  $h$  is the Legendre on  $S$ , then  $h - \lambda g$  is also the Legendre on  $S$ , for all  $\lambda$  such that  $0 < \lambda < (1/L)$ .

*Proof.* Conditions (a) and (b) of Definition 1 being verified, let us demonstrate that the condition (c) is verified too. Let  $\{x_i\}$  be  $x_i \rightarrow x^* \in \text{Fr}(S) = (\bar{S}/S)$ :

$$\begin{aligned} \|\nabla h(x_i) - \lambda \nabla g(x_i)\|^2 &\geq \|\nabla h(x_i)\| (\|\nabla h(x_i)\| - 2\lambda \|\nabla g(x_i)\|) \\ &\quad + \lambda^2 \|\nabla g(x_i)\|^2. \end{aligned} \quad (32)$$

Then,

$$\lim_{x_i \rightarrow x^* \in \text{Fr}(S)} \|\nabla(h - \lambda g)(x_i)\| = +\infty. \quad (33)$$

$h - \lambda g$  is strictly convex in  $S$ . Indeed, let  $x, y \in S$ ,  $x \neq y$ ; we have  $D_{h-\lambda g}(x, y) \geq 0$ :

$$\begin{aligned} D_{h-\lambda g}(x, y) = 0 &\implies D_h(x, y) \\ &= \lambda D_g(x, y) \implies D_h(x, y) \leq \lambda LD_h(x, y), \end{aligned} \quad (34)$$

$h$  is strongly convex on  $S$ ,  $x, y \in S$ ,  $x \neq y$ ; then,  $D_h(x, y) \neq 0 \implies 1 \leq \lambda L$  which is absurd. Hence,  $h - \lambda g$  is strictly convex in  $S$ . □

**Lemma 6.** Consider the following:

$$\forall \lambda \in ]0, \frac{1}{L}[, \quad \forall u \in S, \quad \partial(D_{h-\lambda g}(\cdot, u))(x^*) = \begin{cases} \{\nabla(h-\lambda g)(x^*) - \nabla(h-\lambda g)(u)\}, & \text{if } x^* \in S, \\ \emptyset, & \text{if not.} \end{cases} \quad (35)$$

*Proof.* Since  $h - \lambda g$  is a Legendre function on  $S$ ,  $D_{h-\lambda g}(\cdot, u)$  is also a Legendre. By application of Theorem 26.1 in [23],  $\partial(D_{h-\lambda g}(\cdot, u))$  verifies the following:

- (i) If  $x^* \in \text{int}(\text{dom } D_{h-\lambda g}(\cdot, u)) = S$ , then
 
$$\partial(D_{h-\lambda g}(\cdot, u))(x^*) = \{\nabla D_{h-\lambda g}(\cdot, u)(x^*)\}. \quad (36)$$
- (ii) If  $x^* \notin S$ , then  $\partial(D_{h-\lambda g}(\cdot, u))(x^*) = \emptyset$ . □

**Theorem 2** (well-posedness of the method). *We assume that*

- (i)  $\Psi$  is convex.
- (ii) The pair  $(g, h)$  verified the condition (LC).
- (iii)  $\forall r \geq 0, \forall x \in S$ , the sets below are bounded:

$$L_2(x, r) = \{y \in S; D_h(y, x) \leq r\}. \quad (37)$$

Then,  $\forall \lambda \in ]0, (1/L)[$  and the map  $T_\lambda$  defined in (25) is nonempty and single-valued from  $S$  to  $S$ .

*Proof.*  $\forall x \in S$  and  $T_\lambda(x)$  is nonempty; for this, it is enough to demonstrate that  $\forall r \in R$ :

$$L(x, r) = \{u \in \bar{S} : \Psi(u) + \lambda^{-1} D_{h-\lambda g}(u, x) \leq r\}, \quad (38)$$

which is closed and is bounded when it is nonempty:

$$\begin{aligned} u \in L(x, r) &\implies \Psi(u) + \lambda^{-1} D_{h-\lambda g}(u, x) \leq r \\ &\implies D_{h-\lambda g}(u, x) \leq \lambda(r - \Psi^*) \\ &\implies D_h(u, x) \leq \lambda(r - \Psi^*) + \lambda D_g(u, x) \\ &\implies D_h(u, x) \leq \lambda(r - \Psi^*) \\ &\quad + \lambda D_h(u, x) \\ &\implies D_h(u, x) \leq \frac{\lambda(r - \Psi^*)}{1 - \lambda L}. \end{aligned} \quad (39)$$

It follows that

$$L(x, r) \subset L_2\left(x, \frac{\lambda(r - \Psi^*)}{1 - \lambda L}\right), \quad (40)$$

thanks to  $H_4$ ;  $L_2(x, (\lambda(r - \Psi^*)/(1 - \lambda L)))$  is bounded, which leads that  $L(x, r)$  is bounded too, which shows that

$$T_\lambda(x) \neq \emptyset. \quad (41)$$

Let  $x^* \in T_\lambda(x)$ . Let us suppose that  $x^* \in S(26) \implies 0 \in \partial(\Psi(\cdot) + 1/\lambda D_{h-\lambda g}(\cdot, x))(x^*)$ , since  $\text{ri}(\text{dom } \Psi) \cap \text{ri}(\text{dom } D_{h-\lambda g}$

$(\cdot, x)) = \text{ri}(\text{dom } \Psi) \cap \text{ri}(\bar{S}) = \text{ri}(\text{dom } \Psi) \cap S \neq \emptyset$ , from [10], which allows to write that

$$0 \in \partial\Psi(x^*) + \partial\left(\frac{1}{\lambda} D_{h-\lambda g}(\cdot, x)\right)(x^*). \quad (42)$$

It follows that

$$\exists u \in \partial\Psi(x^*), \quad (43)$$

such that

$$-\lambda u \in \partial D_{h-\lambda g}(\cdot, x)(x^*) \quad (44)$$

is in contradiction with Lemma 6. Then,  $T_\lambda(x) \subset S$ .

On the other hand,  $h - \lambda g$  is strictly convex in  $S$  and  $\Psi$  is convex, so  $\Psi(\cdot) + D_{h-\lambda g}(\cdot, x)$  is strongly convex in  $S$ . Then,  $T_\lambda(x)$  has a unique value for all  $x \in S$ . □

*Remark 1.* This result is liberated from the supercoercivity of  $\Psi$  and the simultaneous convexity of  $f$  and  $g$ , as required by Lemma 2 [15].

**Proposition 3.**  $\forall x \in S, \forall \lambda \in ]0, (1/L)[$

$$\frac{\nabla(h-\lambda g)(x) - \nabla(h-\lambda g)(T_\lambda(x))}{\lambda} \in \partial\Psi(T_\lambda(x)), \quad (45)$$

$$\exists \bar{x} \in S, \frac{\nabla(h-\lambda g)(x) - \nabla(h-\lambda g)(\bar{x})}{\lambda} \in \partial_\varepsilon \Psi(\bar{x}). \quad (46)$$

*Proof.* Since  $T_\lambda(x) \in S$ , we have

$$\begin{aligned} 0 &\in \partial\left(\Psi(\cdot) + \frac{1}{\lambda} D_{h-\lambda g}(\cdot, x)\right)(T_\lambda(x)) \\ &\implies -\nabla\left(\frac{1}{\lambda} D_{h-\lambda g}(\cdot, x)\right)((T_\lambda(x))) \in \partial(\Psi(T_\lambda(x))) \\ &\implies (45). \end{aligned} \quad (47)$$

For (46), just take  $\bar{x} = T_\lambda(x)$ , since

$$\partial\Psi(T_\lambda(x)) \subset \partial_\varepsilon \Psi(T_\lambda(x)). \quad (48) \quad \square$$

**Proposition 4.**  $\forall x \in S, \forall \lambda \in ]0, (1/L)[$ :

$$T_\lambda(x) = \text{prox}_{\lambda\Psi}^{h-\lambda g}(x) = \text{prox}_{\lambda f}^h \circ \text{prox}_{\lambda p}^h(x), \quad (49)$$

where  $p(u) = \langle \nabla g(x), u \rangle$ .

*Proof.* The first equality is due to Lemma 3. The second is established in [15]. □

The first equality played a decisive role in the development of this paper.

#### 4. Analysis of the $\varepsilon$ -NoLips Algorithm

In this section, we propose an Inexact Bregman Proximal Gradient Algorithm (IBPG), which is an inexact version of the BPG algorithm described in [15, 22]; the IBPG

framework allows an error in the subgradient inclusion by using the error  $\varepsilon_n$ . We study two algorithms:

- (i) Algorithm 1: inexact Bregman Proximal Gradient (IBPG) algorithm without relative error criterion
- (ii) Algorithm 2: inexact Bregman Proximal Gradient (IBPG) algorithm with relative error criterion which we call  $\varepsilon$ -NoLips

We establish the main convergence properties of the proposed algorithms. In particular, we prove its global rate of convergence, showing that it shares the claimed sublinear rate  $O(1/n)$  of basic first-order methods such as the classical PG and BPG. We also derive a global convergence of the sequence generated by NoLips to a minimizer of  $(P)$ .

*Assumptions 4*

- (i)  $h \in B(S) \cap \mathcal{E}(S)$
- (ii)  $\Psi$  is convex
- (iii) The pair  $(g, h)$  verified the condition (LC)
- (iv)  $\text{Argmin } \Psi \neq \emptyset$

In our analysis,  $\Psi$  is supposed to be a convex function; it allows to distinguish two interesting cases:

- (i) The nonsmooth part  $f$  is possibly not convex, and the smooth part  $g$  is convex
- (ii) The nonsmooth part  $f$  is convex, and the smooth part  $g$  is possibly not convex

In what follows, the choice of the sequence  $\{\lambda_n\}$  depends of the convexity of  $g$ .

Let  $\bar{\lambda}$  such that  $0 < \bar{\lambda} < (1/L)$ ,  $\lambda_0 := 0$ .

If  $g$  is not convex, then we choose

$$\lambda_n = \lambda, \quad (0 < \lambda \leq \bar{\lambda}), \quad n = 1 \dots \quad (50)$$

If  $g$  is convex, then we choose

$$\lambda_n \leq \lambda_{n+1} \leq \bar{\lambda}, \quad n = 1 \dots \quad (51)$$

In those conditions, we easily show that  $\forall x, y \in S, \forall n \in N, 0 < \lambda_n \leq \bar{\lambda}$ , such that

$$(\lambda_{n+1} - \lambda_n)D_g(x, y) \geq 0, \quad (52)$$

We pose

$$h_n = h - \lambda_n g, \quad n = 1 \dots \quad (53)$$

**Proposition 5.** *The sequence  $\{x^n\}_n$  defined by (IBPG) exists and is verified for all  $n \in N^*$ :*

$$x^n \in \varepsilon_n - \text{argmin} \left\{ \Psi(u) + \frac{1}{\lambda_n} D_{h_n}(u, x^{n-1}), u \in \bar{S} \right\}. \quad (54)$$

*Proof.* Existence is deduced trivially from (45):

$$\Omega^n := \frac{\nabla h_n(x^{n-1}) - \nabla h_n(x^n)}{\lambda_n} \in \partial_{\varepsilon_n} \Psi(x^n) \quad (55)$$

$$\implies \Psi(u) \geq \Psi(x^n) + \langle u - x^n, \Omega^n \rangle - \varepsilon_n.$$

By applying Lemma 2, we have

$$\begin{aligned} \Psi(u) &\geq \Psi(x^n) + \lambda_n^{-1} [D_{h_n}(u, x^n) + D_{h_n}(x^n, x^{n-1}) - D_{h_n}(u, x^{n-1})] - \varepsilon_n, \\ \implies \Psi(x^n) + \lambda_n^{-1} D_{h_n}(x^n, x^{n-1}) &\leq \Psi(u) + \lambda_n^{-1} D_{h_n}(u, x^{n-1}) + \varepsilon_n, \quad \forall u \in \bar{S} \\ \implies x^n &\in T_{\lambda_n}^{\varepsilon_n}(x^{n-1}), \end{aligned} \quad (56)$$

where  $T_{\lambda}^{\varepsilon}(x) = \varepsilon - \text{argmin} \{ \Psi(u) + (1/\lambda)D_{h-\lambda g}(u, x) \}$ .  $\square$

*Remark 2.* This result shows that IBPG is an inexact version of BPG and this is exactly the BPG when  $\varepsilon_n = 0$ , i.e.:

- (i)  $x^n \in T_{\lambda_n}^{\varepsilon_n}(x^{n-1}) \implies x^n \simeq T_{\lambda_n}(x^{n-1})$
- (ii)  $\varepsilon_n = 0 \implies x^n = T_{\lambda_n}(x^{n-1})$

**Proposition 6.** *For all  $n \in N^*$ ,*

$$(i) (\Psi(x^n) - \Psi(x^{n-1})) \leq -D_{h_n}(x^n, x^{n-1}) + \lambda_n \varepsilon_n, \quad (57)$$

*Proof*

$$\begin{aligned} (ii) D_h(x^n, x^{n-1}) + D_h(x^{n-1}, x^n) &\leq \frac{\bar{\lambda}}{1 - \bar{\lambda}L} (\Psi(x^{n-1}) \\ &\quad - \Psi(x^n)) + \frac{\lambda_n \varepsilon_n}{1 - \bar{\lambda}L}. \end{aligned} \quad (58)$$

$$(55) \implies \lambda_n (\Psi(x^n) - \Psi(u)) \leq [D_{h_n}(u, x^{n-1}) - D_{h_n}(u, x^n) - D_{h_n}(x^n, x^{n-1})] + \varepsilon_n \lambda_n. \quad (59)$$

We put  $u = x^{n-1}$  in (59), we get (57)

(ii) Put  $u = x^{n-1}$  in (59), we have

$$\begin{aligned} D_{h_n}(x^n, x^{n-1}) + D_{h_n}(x^{n-1}, x^n) &\leq \lambda_n(\Psi(x^{n-1}) - \Psi(x^n)) + \lambda_n \varepsilon_n \\ \implies D_h(x^n, x^{n-1}) + D_h(x^{n-1}, x^n) &\leq \lambda_n(\Psi(x^{n-1}) - \Psi(x^n)) + \lambda_n \varepsilon_n + \lambda_n D_g(x^n, x^{n-1}) + \lambda_n D_g(x^{n-1}, x^n) \\ \implies D_h(x^n, x^{n-1}) + D_h(x^{n-1}, x^n) &\leq \bar{\lambda}(\Psi(x^{n-1}) - \Psi(x^n)) + \lambda_n \varepsilon_n + \bar{\lambda}L(D_h(x^n, x^{n-1}) + D_h(x^{n-1}, x^n)) \\ \implies D_h(x^n, x^{n-1}) + D_h(x^{n-1}, x^n) &\leq \frac{\bar{\lambda}}{1 - \bar{\lambda}L}(\Psi(x^{n-1}) - \Psi(x^n)) + \frac{\lambda_n \varepsilon_n}{1 - \bar{\lambda}L}. \end{aligned} \tag{60}$$

**Corollary 1**

- (i) If  $\varepsilon_n = 0$ , the sequence  $\{\Psi(x^n)\}_n$  is nonincreasing
- (ii) Summability: if  $\sum_{n=1}^\infty \lambda_n \varepsilon_n < +\infty$ , then  $\sum_{n=1}^\infty D_h(x^n, x^{n-1}) < +\infty$  and  $\sum_{n=1}^\infty D_h(x^{n-1}, x^n) < +\infty$

*Proof*

- (i) From (57),  $D_{h_n}(x^n, x^{n-1}) \geq 0 \implies \Psi(x^n) \leq \Psi(x^{n-1})$ ,  $\forall n \in N^*$ .
- (ii) From (58),  $\sum_{n=1}^{n=p} D_h(x^n, x^{n-1}) + D_h(x^{n-1}, x^n) \leq$

$$\begin{aligned} &\frac{\bar{\lambda}}{1 - \bar{\lambda}L} \sum_{n=1}^{n=p} (\Psi(x^{n-1}) - \Psi(x^n)) + \frac{1}{1 - \bar{\lambda}L} \sum_{n=1}^{n=p} \lambda_n \varepsilon_n \\ \implies &\sum_{n=1}^{n=p} D_h(x^n, x^{n-1}) + D_h(x^{n-1}, x^n) \tag{61} \\ &\leq \frac{\bar{\lambda}}{1 - \bar{\lambda}L} (\Psi(x^0) - \Psi^*) + \frac{1}{1 - \bar{\lambda}L} \sum_{n=1}^{n=p} \lambda_n \varepsilon_n. \end{aligned}$$

In the following, we pose

$$\begin{aligned} t_p &:= \sum_{n=1}^{n=p} \lambda_n, \quad \forall p \in N^*, t_0 = \lambda_0 = 0, \\ A_n(u) &= D_{h_n}(u, x^n), \quad \forall u \in \bar{S}, \forall n \in N^*, \\ \Phi(x^p) &:= \min\{\Psi(x^k), 1 \leq k \leq p\}, \quad \forall p \in N^*. \end{aligned} \tag{62}$$

**Proposition 7** (Global Estimate in Function Values). (a) For all  $u \in \bar{S}$  and  $\forall p \in N^*$ ,

$$\Phi(x^p) - \Psi(u) \leq \frac{1}{t_p} \left[ D_h(u, x^0) + \sum_{n=1}^\infty \lambda_n \varepsilon_n \right]. \tag{63}$$

*Proof.* We have  $D_{h_n}(u, x^{n-1}) = A_{n-1}(u) - (\lambda_n - \lambda_{n-1})D_g(u, x^{n-1})$ ; from (59), we have

$$\begin{aligned} \lambda_n(\Psi(x^n) - \Psi(u)) &\leq A_{n-1}(u) - A_n(u) - D_{h_n}(x^n, x^{n-1}) \\ &\quad - (\lambda_n - \lambda_{n-1})D_g(u, x^{n-1}) + \lambda_n \varepsilon_n. \end{aligned} \tag{64}$$

From (52), we have

$$\begin{aligned} \lambda_n(\Psi(x^n) - \Psi(u)) &\leq A_{n-1}(u) - A_n(u) + \lambda_n \varepsilon_n \\ \implies \sum_{n=1}^{n=p} \lambda_n(\Psi(x^n) - \Psi(u)) &\leq \sum_{n=1}^{n=p} A_{n-1}(u) - A_n(u) + \sum_{n=1}^{n=p} \lambda_n \varepsilon_n \\ \implies \sum_{n=1}^{n=p} \lambda_n(\Psi(x^n) - \Psi(u)) &\leq A_0(u) - A_p(u) + \sum_{n=1}^{n=p} \lambda_n \varepsilon_n \\ \implies (\Phi(x^p) - \Psi(u)) \sum_{n=1}^{n=p} \lambda_n &\leq A_0(u) - A_p(u) + \sum_{n=1}^{n=p} \lambda_n \varepsilon_n, \end{aligned} \tag{65}$$

$A_0(u) = D_h(u, x^0)$  ( $\lambda_0 = 0$ ), so

$$\Phi(x^p) - \Psi(u) \leq \frac{1}{t_p} \left[ D_h(u, x^0) + \sum_{n=1}^\infty \lambda_n \varepsilon_n \right]. \tag{66}$$

This theorem covers the evaluation of the global convergence rate given in [6, 12], as shown by the following corollary.

**Corollary 2.** We assume that

- (a)  $\varepsilon_n = 0, n = 1, \dots$
- (b)  $\lambda_n = \bar{\lambda} = 1/2L, n = 1, \dots$
- (c)  $x^* \in \arg \min \Psi$

Then,

$$\Psi(x^p) - \Psi^* \leq \frac{2L}{p} D_h(x^*, x^0), \tag{67}$$

i.e.,  $\Psi(x^p) - \Psi^* = O(1/p)$

*Proof.* Immediate consequence of Proposition 7. □

Now, we derive a global convergence of the sequence generated by Algorithm 1 to a minimizer of (P).

**Theorem 3.** We assume that  $\sum \lambda_n \varepsilon_n < +\infty$ , if one of the following assumptions holds:

- (i)  $\exists \underline{\lambda} > 0$  such that  $\underline{\lambda} \leq \lambda_n \leq \bar{\lambda}$ ,  $n = 1, \dots$
- (ii) The sequence  $\{\Psi(x^n)\}$  is nonincreasing and  $\sum \lambda_n = +\infty$ ; then, (a)  $\Psi(x^n) \rightarrow \inf \Psi$  and (b)  $x^n \rightarrow x^* \in \arg \min \Psi$

*Proof*

(a) Suppose

- (i) Let  $x^* \in \arg \min \Psi$  and we put  $u = x^*$  in (69), we have

$$\underline{\lambda}(\Psi(x^n) - \Psi(x^*)) \leq A_{n-1}(x^*) - A_n(x^*) + \lambda_n \varepsilon_n, \quad (68)$$

$$A_n(x^*) \leq A_{n-1}(x^*) + \lambda_n \varepsilon_n, \quad (69)$$

$$\sum \lambda_n \varepsilon_n < +\infty \implies A_n(x^*) \rightarrow l \in \mathbb{R}. \quad (70)$$

(68) and (suppose (69))  $\implies \Psi(x^n) \rightarrow \inf \Psi$ .

Suppose (ii)

$$\Phi(x^p) = \min\{\Psi(x^k), 1 \leq k \leq p\} = \Psi(x^p). \quad (71)$$

For  $u = x^* \in \arg \min \Psi$  in (63), we have  $0 \leq \Psi(x^p) - \Psi(x^*) \leq (1/t_p)[D_h(x^*, x^0) + \sum_{n=1}^{\infty} \lambda_n \varepsilon_n]$ , so  $\sum \lambda_n = +\infty \implies \Psi(x^n) \rightarrow \inf \Psi$ .

- (b) (69)  $\implies \exists \alpha > 0$ ,  $A_n(x^*) \leq \alpha A_n(x^*) = D_{h_n}(x^*, x^n) = D_h(x^*, x^n) - \lambda_n D_g(x^*, x^n)$ , so

$$\begin{aligned} D_h(x^*, x^n) &\leq \alpha + \lambda_n D_g(x^*, x^n) \\ &\Downarrow \\ D_h(x^*, x^n) &\leq \frac{\alpha}{1 - \bar{\lambda}L}. \end{aligned} \quad (72)$$

Then,  $\{D_h(x^*, x^n)\}$  is bounded, and from  $H_3$ ,  $\{x^n\}$  is bounded as well. Let  $u^* \in \text{Adh}\{x^n\}$ ; there exists then a subsequence  $\{x^{n_i}\}$  of  $\{x^n\}$  such that  $x^{n_i} \rightarrow u^* \in S$ . From  $H_5$ ,  $D_h(u^*, x^{n_i}) \rightarrow 0$ . On the other hand,

$$0 \leq D_g(u^*, x^{n_i}) \leq L \cdot D_h(u^*, x^{n_i}), \quad (73)$$

so  $D_g(u^*, x^{n_i}) \rightarrow 0$ ;  $u^* \in \arg \min \Psi$ . Indeed,

$$\begin{aligned} \inf \Psi &\leq \Psi(u^*) \leq \underline{\lim} \Psi(x^{n_i}) = \inf \Psi \\ &\Downarrow \\ \inf \Psi &= \Psi(u^*), \end{aligned} \quad (74)$$

which shows that  $\implies u^* \in \arg \min \Psi$ . Then,

$$D_{h_{n_i}}(u^*, x^{n_i}) = D_h(u^*, x^{n_i}) - \lambda_{n_i} D_g(u^*, x^{n_i}) \rightarrow 0. \quad (75)$$

Since  $D_{h_{n_i}}(u^*, x^{n_i}) \rightarrow 0$  and  $u^* \in \arg \min \Psi$ , we have

$$D_{h_n}(u^*, x^n) \rightarrow 0. \quad (76)$$

We have

$$\begin{aligned} D_h(u^*, x^n) &= D_{h_n}(u^*, x^n) + \lambda_n D_g(u^*, x^n) \\ &\leq D_{h_n}(u^*, x^n) + L \cdot \bar{\lambda} D_h(u^*, x^n), \end{aligned} \quad (77)$$

so  $(1 - L \cdot \bar{\lambda})D_h(u^*, x^n) \leq D_{h_n}(u^*, x^n)$ ; then,

$$D_h(u^*, x^n) \rightarrow 0, \quad (78)$$

And from  $H_6$ , we have  $x^n \rightarrow u^* \in \arg \min \Psi$ .  $\square$

The IBPG algorithm generates a sequence such that  $\{\Psi(x^n)\}_n$  does not necessarily be nonincreasing; for this reason and for improvement of the global estimate in function values, we now propose  $\varepsilon$ -NoLips which is an inexact version of BPG with a relative error criterion. Let  $\sigma$  such that  $0 \leq \sigma < 1$  be given as follows.

In what follows, we will derive a convergence rate result (Theorem 4) for the  $\varepsilon$ -NoLips framework. First, we need to establish a few technical lemmas.  $\square$

In the following,  $\{x^n\}_n$  denotes the sequence generated by  $\varepsilon$ -NoLips.

**Lemma 7.** For every  $u \in \bar{S}$ , for all  $n \in N^*$ ,

$$\begin{aligned} \lambda_n(\Psi(x^n) - \Psi(u)) &\leq A_{n-1}(u) - A_n(u) \\ &\quad - (1 - \sigma)(1 - \bar{\lambda}L)D_h(x^n, x^{n-1}). \end{aligned} \quad (79)$$

*Proof.* Since  $(\lambda_n - \lambda_{n-1})D_g(u, x^{n-1}) \geq 0$ , we have from (62),

$$\begin{aligned} \lambda_n(\Psi(x^n) - \Psi(u)) &\leq -D_{h_n}(x^n, x^{n-1}) + A_{n-1}(u) \\ \&\&9; \quad -A_n(u) + \lambda_n \varepsilon_n. \end{aligned} \quad (80)$$

From Algorithm 2, we have

$$\lambda_n(\Psi(x^n) - \Psi(u)) \leq A_{n-1}(u) - A_n(u) + (\sigma - 1)D_{h_n}(x^n, x^{n-1}). \quad (81)$$

From the condition LC, we have

$$\begin{aligned} \lambda_n(\Psi(x^n) - \Psi(u)) &\leq A_{n-1}(u) - A_n(u) \\ &\quad - (1 - \sigma)(1 - \bar{\lambda}L)D_h(x^n, x^{n-1}). \end{aligned} \quad (82)$$

$\square$

*Remark 3.* We now notice that  $\{\Psi(x^n)\}_n$  is nonincreasing. Just replace  $u$  with  $x^{n-1}$  in (79).

**Lemma 8.** For every  $n \in N^*$  and  $x^* \in \arg \min \Psi$ , we have

$$\begin{aligned} t_n(\Psi(x^n) - \Psi^*) &+ \lambda_n^{-1} t_n (1 - \sigma)(1 - \bar{\lambda}L)D_h(x^n, x^{n-1}) \\ &\leq t_{n-1}(\Psi(x^{n-1}) - \Psi^*) + A_{n-1}(x^*) - A_n(x^*). \end{aligned} \quad (83)$$

*Proof.* Replacing  $u$  by  $x^{n-1}$  in (79), and since  $A_n(x^{n-1}) \geq 0$  and  $A_{n-1}(x^{n-1}) = 0$ , we have



(1) Input:  $x^0 \in S \cap \text{ri}(\text{dom } \Psi)$   
 (2) For  $n = 1, 2, \dots$  with  $\varepsilon_n > 0$ , we obtain  $(\nabla h_n(x^{n-1}) - \nabla h_n(x^n))/\lambda_n \in \partial_{\varepsilon_n} \Psi(x^n)$

ALGORITHM 1: Inexact Bregman Proximal Gradient (IBPG).

$$\begin{aligned} & t_{n-1}(\Psi(x^n) - \Psi^*) + t_{n-1}\lambda_n^{-1}(1 - \sigma)(1 - \bar{\lambda}L)D_h(x^n, x^{n-1}) \\ & \leq t_{n-1}(\Psi(x^{n-1}) - \Psi^*). \end{aligned} \tag{84}$$

Replacing  $u$  by  $x^*$  in (79), we have

$$\begin{aligned} & \lambda_n(\Psi(x^n) - \Psi(x^*)) + (1 - \sigma)(1 - \bar{\lambda}L)D_h(x^n, x^{n-1}) \\ & + \leq A_{n-1}(x^*) - A_n(x^*). \end{aligned} \tag{85}$$

Since  $t_{n-1} + \lambda_n = t_n$ , by adding (84) and (85), we have (83).  $\square$

**Lemma 9.** For every  $k \in N^*$

$$\begin{aligned} & t_k(\Psi(x^k) - \Psi^*) + (1 - \sigma)(1 - \bar{\lambda}L) \sum_{n=1}^{n=k} \lambda_n^{-1} t_n D_h(x^n, x^{n-1}) \\ & \leq A_0(x^*) - A_k(x^*). \end{aligned} \tag{86}$$

*Proof.* This result is obtained by adding inequality (83) from 1 to  $k$  ( $t_0 = \lambda_0 = 0$ ).  $\square$

We are now ready to state the convergence rate result for the  $\varepsilon$ -NoLips framework. This result improves and completes the one given in the Proposition 7.

**Theorem 4.** For every  $k \in N^*$ , the following statements hold: we pose  $\rho_k := \sum_{n=1}^{n=k} \lambda_n^{-1} t_n$ . Consider

$$(a) \Psi(x^k) - \Psi^* \leq \left( \frac{D_h(x^*, x^0)}{t_k} \right), \tag{87}$$

$$(b) \gamma_k := \min_{1 \leq n \leq k} D_h(x^n, x^{n-1}) \leq \left( \frac{D_h(x^*, x^0)}{(1 - \sigma)(1 - \bar{\lambda}L)\rho_k} \right). \tag{88}$$

*Proof.* From (86) and since  $A_0(x^*) = D_h(x^*, x^0)(\lambda_0 = 0)$  and  $A_k(x^*) \geq 0$ , we immediately have (87) and (88).  $\square$

**Corollary 3.** Consider an instance of the  $\varepsilon$ -NoLips framework with  $\lambda_n = \lambda$  for every  $n \in N^*$ . Then, for every  $k \in N^*$ , the following statements hold:

$$(a) \Psi(x^k) - \Psi^* = O\left(\frac{1}{k}\right), \tag{89}$$

$$(b) \gamma_k = \min_{1 \leq n \leq k} D_h(x^n, x^{n-1}) = O\left(\frac{1}{k^2}\right). \tag{90}$$

*Proof.*  $t_k = k\lambda$  and (87)  $\implies$  (90):

$$\rho_k = \frac{k(k+1)}{2} \geq \frac{k^2}{2}. \tag{91}$$

and (88)  $\implies$  (90).  $\square$

*Remark 4.* (87) and (89) represent exactly the convergence rate established in [15, 22]. (90) is a new result not established in [15, 22]; this result shows that  $\gamma_k$  converges to zero at a rate of  $O(1/k^2)$ .

**Theorem 5.** If  $\sum \lambda_n = +\infty$ , then

- (a)  $\Psi(x^n) \longrightarrow \inf \Psi$
- (b)  $x^n \longrightarrow x^* \in \arg \min \Psi$

*Proof.* Replacing  $u$  by  $x^*$  in (81), we have

$$(1 - \sigma)D_{h_n}(x^n, x^{n-1}) \leq A_{n-1}(x^*) - A_n(x^*). \tag{92}$$

By adding the inequality (92) from 1 to  $k$ , we have

$$(1 - \sigma) \sum_{n=1}^{n=k} D_{h_n}(x^n, x^{n-1}) \leq A_0(x^*) - A_k(x^*). \tag{93}$$

$A_k(x^*) \geq 0$ , so  $\sum_{n=1}^{\infty} D_{h_n}(x^n, x^{n-1}) < +\infty$ . From (69), we have

$$\sum_{n=1}^{\infty} \lambda_n \varepsilon_n < +\infty. \tag{94}$$

Since  $\{\Psi(x^n)\}_n$  is nonincreasing and (94), by applying the Theorem 3 (ii), we have the results of Theorem 5.  $\square$

## 5. Application to Nonnegative Linear Inverse Problem

In Poisson inverse problems (e.g., [25, 26]), we are given a nonnegative observation matrix  $A \in R_+^{m \times d}$  and a noisy measurement vector  $b \in R_+^m$ , and the goal is to reconstruct the signal  $x \in R_+^d$  such that  $Ax \approx b$ . We can naturally adopt the distance  $D(Ax, b)$  to measure the residuals between two nonnegative points, with

(1) Input:  $x^0 \in S \cap \text{ri}(\text{dom } \Psi)$   
 (2) Choose  $\lambda_n > 0$  and find  $x_n \in S, \sigma_n \in [0, \sigma]$  and  $\varepsilon_n \geq 0$  such that  
 $(\nabla h_n(x^{n-1}) - \nabla h_n(x^n))/\lambda_n \in \partial_{\varepsilon_n} \Psi(x^n)$ ,  
 $\lambda_n \varepsilon_n \leq \sigma_n D_{h_n}(x^n, x^{n-1})$   
 (3) Set  $n \leftarrow n + 1$  and go to step 1

ALGORITHM 2:  $\varepsilon$ -NoLips.

$$D(Ax, b) = \sum_{i=1}^m \langle a_i, x \rangle \log \frac{\langle a_i, x \rangle}{b_i} + b_i - \langle a_i, x \rangle, \quad (95)$$

where  $a_i$  denotes the  $i$ th line of  $A$ .

In this section, we propose an approach for solving the nonnegative linear inverse problem defined by

$$(P_\alpha): \inf \{ \alpha \|x\|_1 + D(Ax, b) : x \in R_+^d \}, \quad (\alpha > 0). \quad (96)$$

We take

$$\begin{aligned} f(x) &:= \alpha \|x\|_1, \\ g(x) &:= D(Ax, b) \\ h(x) &:= h_1(x) = \sum_{i=1}^d x_i \log x_i. \end{aligned} \quad (97)$$

It is shown in [15] that the couple  $(g, h)$  verified a Lipschitz-like/Convexity Condition (LC) on  $R_+^d$  for any  $L$  such that

$$L \geq \max_{1 \leq j \leq d} \sum_{i=1}^m a_{ij} = \max_{1 \leq j \leq d} \|A_j\|_1, \quad (98)$$

where  $A_j$  denotes the  $j$ th column of  $A$ .

For  $\lambda_n := \lambda, \forall n$ , Theorem 3 is applicable and global warrant of Algorithm 1 convergence to an optimal solution of  $(P_\alpha)$ .

Given  $x^n \in R_{++}^d$ , the iteration

$$x^{n+1} \simeq T_\lambda(x^n) \quad (99)$$

amounts to solving the one-dimensional problem:

For  $j = 1, \dots, d$ ,

$$x_j^{n+1} \simeq \arg \min \left\{ \alpha t + \gamma_j t + \frac{1}{\lambda} \left( t \log \frac{t}{x_j^n} + x_j^n - t \right), t > 0 \right\}, \quad (100)$$

where  $\gamma_j$  is the  $j$ th component of  $\nabla g(x^n)$ .

### 6. Conclusion

The proposed algorithms constitute a unified frame for the existing algorithms BPG, BP, and PG, by giving others, in particular, the inexact version of the interior method with Bregman distance studied in [24]. More precisely,

- (i) When  $\varepsilon_n = 0, \forall n \in N^*$ , our algorithm is the NoLips studied in [15, 22]

- (ii) When  $g = 0$ , our algorithm is the inexact version of Bregman Proximal (BP) studied in [19]
- (iii) When  $\varepsilon_n = 0, \forall n \in N^*$ , and  $g = 0$ , our algorithm is the Bregman Proximal (BP) studied in [17, 21]
- (iv) When  $f = 0$ , our algorithm is the inexact version of the interior method with Bregman distance studied in [24]
- (v) When  $f = 0$  and  $\varepsilon_n = 0, \forall n \in N^*$ , our algorithm is the interior method with Bregman distance studied in [24]
- (vi) When  $h = (1/2)\|\cdot\|^2$ , our algorithm is the proximal gradient method (PG) and its variants [4–14, 27]

Our analysis is different and more simple than the one given in [15, 22] and allows to reduce some hypothesis, in particular, the supercoercivity of  $\Psi$  as well as the simultaneous convexity of  $f$  and  $g$ .

### Data Availability

No data were used to support this study.

### Conflicts of Interest

The authors declare that they have no conflicts of interest.

### References

- [1] D. L. Donoho, “Compressed sensing,” *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.
- [2] A. Beck and Y. C. Eldar, “Sparsity constrained nonlinear optimization: optimality conditions and algorithms,” *SIAM Journal on Optimization*, vol. 23, no. 3, pp. 1480–1509, 2013.
- [3] D. R. Luke, “Phase retrieval, what’s new,” *SIAG/OPT Views and News*, vol. 25, no. 1, pp. 1–5, 2017.
- [4] A. Auslender, “Numerical methods for non-differentiable convex optimization,” in *Mathematical Programming Studies*, J. P. Vial and C. B. Nguyen, Eds., vol. 30, pp. 102–126, Springer, Berlin, Germany, 1987.
- [5] A. I. Chen and A. Ozdaglar, “A fast distributed proximal-gradient method,” in *Proceedings of the 2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 601–608, IEEE, Monticello, IL, USA, October 2012.
- [6] J. B. Hiriart-Urruty, “ $\varepsilon$ -subdifferential calculus. Convex analysis and optimization,” in *Research Notes in Mathematics. Series, Vol. 57*, Pitman Publishers, London, UK, 1982.
- [7] K. Jiang, D. Sun, and K.-C. Toh, “An inexact accelerated proximal gradient method for large scale linearly constrained convex SDP,” *SIAM Journal on Optimization*, vol. 22, no. 3, pp. 1042–1064, 2012.

- [8] B. Lemaire, "About the convergence of the proximal method , proceedings 6th French German conference on optimisation," in *1991 Advances in Optimization Lecture Notes*, pp. 39–51, Springer, Berlin, Germany, 1992.
- [9] B. Lemaire, "The proximal algorithm," *International Series of Numerical Mathematics*, vol. 87, pp. 73–87, 1989.
- [10] B. Martinet, "Perturbation des méthodes d'optimisation-Application," *Modélisation Mathématique et Analyse Numérique*, vol. 12, no. 2, pp. 153–171, 1976.
- [11] N. Parikh, "Proximal algorithms," *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 127–239, 2014.
- [12] M. Schmidt, N. L. Roux, and F. R. Bach, "Convergence rates of inexact proximal-gradient methods for convex optimization," in *Advances in Neural Information Processing Systems*, vol. 28, pp. 1458–1466, MIT Press, Cambridge, MA, USA, 2011.
- [13] N. D. Vanli, M. Gürbüzbalaban, and A. Ozdaglar, "Global convergence rate of proximal incremental aggregated gradient methods," *SIAM Journal on Optimization*, vol. 28, no. 2, pp. 1282–1300, 2018.
- [14] L. Xiao and T. Zhang, "A proximal stochastic gradient method with progressive variance reduction," *SIAM Journal on Optimization*, vol. 24, no. 4, pp. 2057–2075, 2014.
- [15] J. Bolte, S. Sabach, M. Teboulle, and Y. Vaisbourd, "First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems," *SIAM Journal on Optimization*, vol. 28, no. 3, pp. 2131–2151, 2018.
- [16] 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 finding the common point of convex sets and its application to the solution of problems in conve programming," *USSR Computational Mathematics and Mathematical Physics*, vol. 7, no. 3, pp. 200–217, 1967.
- [17] J. Eckstein, "Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming," *Mathematics of Operations Research*, vol. 18, no. 1, pp. 202–226, 1993.
- [18] S. Kabbadj, "Theoretical aspect of diagonal bregman proximal methods," *Journal of Applied Mathematics*, vol. 2020, Article ID 3108056, 9 pages, 2020.
- [19] S. Kabbadj, "Méthodes proximales entropiques," Thèse, Université Montpellier, II, Montpellier, France, 1994.
- [20] R. T. Rockafellar, "Monotone operators and the proximal point algorithm," *SIAM Journal on Control and Optimization*, vol. 14, no. 5, pp. 877–898, 1975.
- [21] M. Teboulle and C. Gong, "Convergence analysis of a proximal-like minimization algorithm using Bregman function," *SIAM Journal on Optimization*, vol. 3, no. 3, pp. 538–543, 1993.
- [22] D. H. Gutman and J. F. Pena, "Unified framework for bregman proximal methods, subgradient, gradient, and accelerated gradient schemes," 2018, <https://arxiv.org/pdf/1812.10198.pdf>.
- [23] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, USA, 1970.
- [24] A. Auslender and M. Teboulle, "Interior gradient and proximal methods for convex and conic optimization," *SIAM Journal on Optimization*, vol. 16, no. 3, pp. 697–725, 2006.
- [25] M. Bertero, P. Boccacci, G. Desiderà, and G. Vicidomini, "Image deblurring with Poisson data: from cells to galaxies," *Inverse Problems*, vol. 25, no. 12, 2009.
- [26] I. Csiszar, "Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems," *The Annals of Statistics*, vol. 19, no. 4, pp. 2032–2066, 1991.
- [27] A. Nitanda, "Stochastic proximal gradient descent with acceleration techniques," in *Advances in Neural Information Processing Systems*, pp. 1574–1582, MIT Press, Cambridge, MA, USA, 2014.