IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 20, 2019, accepted August 11, 2019, date of publication August 22, 2019, date of current version September 17, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2937005

Bregman Proximal Gradient Algorithm With Extrapolation for a Class of Nonconvex Nonsmooth Minimization Problems

XIAOYA ZHANG[1], ROBERTO BARRIO[2], M. ANGELES MARTÍNEZ[2], HAO JIANG[3], AND LIZHI CHENG[1]

[1]Department of Mathematics, National University of Defense Technology, Changsha 410073, China
[2]Departmento de Matemática Aplicada and IUMA, University of Zaragoza, E50009 Zaragoza, Spain
[3]College of Computer, National University of Defense Technology, Changsha 410073, China

Corresponding author: Xiaoya Zhang (zhangxiaoya09@nudt.edu.cn)

The work of X. Zhang, H. Jiang, and L. Cheng was supported in part by the National Key Research and Development Program of China under Grant 2017YFB0202003, and in part by the National Natural Science Foundation of Hunan under Grant 2018JJ3616. The work of R. Barrio and M. A. Martínez was supported in part by the Spanish Research Projects under Grant MTM2015-64095-P and Grant PGC2018-096026-B-I00, in part by the European Regional Development Fund, and in part by the Diputación General de Aragón under Grant E24-17R.

ABSTRACT In this paper, we consider an accelerated method for solving nonconvex and nonsmooth minimization problems. We propose a Bregman Proximal Gradient algorithm with extrapolation (BPGe). This algorithm extends and accelerates the Bregman Proximal Gradient algorithm (BPG), which circumvents the restrictive global Lipschitz gradient continuity assumption needed in Proximal Gradient algorithms (PG). The BPGe algorithm has a greater generality than the recently introduced Proximal Gradient algorithm with extrapolation (PGe) and, in addition, due to the extrapolation step, BPGe converges faster than the BPG algorithm. Analyzing the convergence, we prove that any limit point of the sequence generated by BPGe is a stationary point of the problem by choosing the parameters properly. Besides, assuming Kurdyka-Łojasiewicz property, we prove that all the sequences generated by BPGe converge to a stationary point. Finally, to illustrate the potential of the new method BPGe, we apply it to two important practical problems that arise in many fundamental applications (and that not satisfy global Lipschitz gradient continuity assumption): Poisson linear inverse problems and quadratic inverse problems. In the tests the accelerated BPGe algorithm shows faster convergence results, giving an interesting new algorithm.

INDEX TERMS Bregman proximal gradient algorithm with extrapolation, bregman distance, proximal gradient algorithm, smooth adaptive condition, relative weakly convexity.

I. INTRODUCTION

In recent years, different numerical methods have been devised to solve large-scale minimization problems, but still the Cauchy gradient method is at the kernel of most of the schemes (for instance, see the recent books [6], [9] and it is assumed that the gradient of the objective function is globally Lipschitz continuous). This assumption is quite restrictive in some real applications and, therefore, new families of methods have recently been designed to solve more generic cases. In this sense, the remarkable paper of Bauschke *et al.* [2] introduced a new method based on the Bregman distance paradigm (BPG algorithm) capable of addressing non-global Lipschitz continuous gradient problems in the convex case, and Bolte *et al.* [8] extended it to the nonconvex case.

On the other hand, a great effort has been made to accelerate the proximal gradient algorithm to reduce the number of iterations. Several techniques have been introduced, such as the fast iterative shrinkage-thresholding algorithm (FISTA) proposed in [4], the use of Nesterov's techniques [26], [27], and recently introduced in [39] a version of the proximal gradient algorithm with extrapolation for some nonconvex nonsmooth minimization problems (but assuming that the gradient of the objective function is globally Lipschitz continuous).

The main goal of this paper is to focus on the introduction of a new scheme, and analyze theoretically its convergence, which combines the power of the method developed in [8]

The associate editor coordinating the review of this article and approving it for publication was Nianqiang Li.

capable of solving non-global Lipschitz continuous gradient problems in the convex and nonconvex case, and that includes extrapolation techniques [39] to accelerate the method.

In this paper, we consider the following minimization problem:

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

where f is a nonconvex continuously differentiable function and g is a proper lower-semi-continuous (l.s.c.) convex function. We assume that the optimal value of (1) is finite, that is, $\Psi^* := \inf\{\Psi(x) : x \in \mathbb{R}^d\} > -\infty$. Problem (1) arises in many applications including compressed sensing [17], signal recovery [3], phase retrieve problem [25]. One classical algorithm for solving this problem is the Proximal Gradient (PG) method [31]:

$$x^{k+1} = \arg\min_x \left\{ g(x) + \langle \nabla f(x^k), x - x^k \rangle + \frac{1}{2\lambda_k}\|x - x^k\|^2 \right\},$$

where $k \in \mathbb{N}, \lambda_k$ is the stepsize on each iteration. Proximal gradient method and its variants [14], [20], [28], [35], [38], [40] have been one hot topic in optimization field for a long time due to their simple forms and lower computation complexity.

One branch of developing new PG methods was devoted to convergence accelerations. Accelerated proximal algorithms [4], [37] on convex problems have shown to be quite efficient. They were also useful for solving nonconvex problems [11], [18], [23], [39]. For solving nonconvex problems (1), one simple and efficient strategy is to perform extrapolation for each $k \in \mathbb{N}$, with the following form (where $x^{-1} = x^0$)

$$\begin{cases} y^k = x^k + \beta_k(x^k - x^{k-1}), \\ x^{k+1} = \arg\min_x \left\{ g(x) + \langle \nabla f(y^k), x - y^k \rangle + \frac{1}{2\lambda_k}\|x - y^k\|^2 \right\}, \end{cases}$$

where λ_k is the stepsize on each iteration, and $\beta_k(x^k - x^{k-1})$ is an extrapolation term. The previous iteration is called the Proximal Gradient algorithm with Extrapolation (PGe), which have been shown in [39] that converges and performs quite well by setting parameters β_k properly. However, PGe has one restriction on solving problem (1): it requires the continuously differentiable part f to be globally Lipschitz gradient continuous on $\mathbb{R}^d$. In fact, this requirement cannot often be satisfied for many practical problems, such as the quadratic inverse problem in phase retrieve [25] and Poisson linear inverse problems [5], that arise in many real world applications.

In this paper, we propose a new algorithm —*Bregman Proximal Gradient algorithm with Extrapolation (BPGe)*— to solve problem (1) without requiring globally Lipschitz gradient continuity of f for each $k \in \mathbb{N}$, from $x^{-1} = x^0$:

$$\begin{cases} y^k = x^k + \beta_k(x^k - x^{k-1}), \\ x^{k+1} = \arg\min_x \left\{ g(x) + \langle \nabla f(y^k), x - y^k \rangle + \frac{1}{\lambda_k} D_h(x, y^k) \right\}, \end{cases}$$

where D_h is a Bregman distance defined in Section II. On the basis of Bregman distance theory, we utilize a smooth adaptive condition introduced in [8], which generalizes Lipschitz gradient continuous condition. This smooth adaptive condition was originally proposed to analyze Bregman Proximal Gradient (BPG) algorithm in [8]. It can also be used to analyze the convergence of BPGe, since BPGe algorithm extends BPG one by performing extrapolation. In particular, we have that:

(i) When $D_h(x, y) = \frac{1}{2}\|x - y\|^2$ and $\beta_k = 0$, BPGe reduces to PG.
(ii) When $D_h(x, y) = \frac{1}{2}\|x - y\|^2$, BPGe reduces to PGe.
(iii) When $\beta_k = 0$ for any $k \geq 0$, BPGe reduces to BPG (no extrapolation).

Therefore, PG, PGe and BPG are particular cases of BPGe algorithm.

From the convergence analysis (Section IV), the BPGe algorithm has to satisfy the condition $D_h(x^k, y^k) \leq \rho\, C_k\, D_h(x^{k-1}, x^k)$ (where $C_k \in (0, 1]$ and $\rho \in (0, 1)$ are two parameters) to guarantee the convergence. In the Lipschitz gradient continuous case $D_h(x, y) = \frac{1}{2}\|x - y\|^2$, and so this condition is easily satisfied just by choosing $\inf_{k\in\mathbb{N}}\{\beta_k\} \leq \sqrt{\rho\, C}$. But when D_h is general, computing a threshold of $\inf_{k\in\mathbb{N}}\{\beta_k\}$ directly may be hard and expensive. Therefore, we modify this idea to achieve this condition through a line search method (Algorithm 2 introduced in Section III).

In the convergence analysis, we prove that any limit point of the sequence generated by BPGe is a stationary point under very general conditions. Moreover, by adding some slightly stronger assumptions and Kurdyka-Łojasiewicz property, we could guarantee the sequence generated by BPGe converges to a stationary point.

The paper is organized as follows. We first introduce in Section II some basic definitions in optimization, smooth adaptive condition, relative weak convexity, and Kurdyka-Łojasiewicz property. In Section III we introduce the new BPGe algorithm. The convergence analysis is done in Section IV, where under some assumptions of the smooth adaptive condition and relative weak convexity of problem (1), we first show a descent-type lemma, from which the fact that any limit point of the sequence generated by BPGe is a critical point follows. Later, we prove that the whole sequence generated by BPGe converges to a critical point using Kurdyka-Łojasiewicz property and some additional assumptions. Several numerical experiments are shown in Section V to show the performance of the BPGe method compared with the BPG one.

II. PRELIMINARIES

Throughout the paper we will use the following basic notations. Let $\mathbb{N} := \{0, 1, 2, \dots\}$ be the set of nonnegative integers. We will always work in the Euclidean space $\mathbb{R}^d$, and the standard Euclidean inner product and the induced norm on $\mathbb{R}^d$ are denoted by $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$, respectively. We denote $B_\rho(\tilde{x}) := \{x \in \mathbb{R}^d : \|x - \tilde{x}\| \leq \rho\}$ as the ball of radius

$\rho > 0$ around $\tilde{x} \in \mathbb{R}^d$, $\text{dist}(x, \mathcal{S}) := \inf_{y\in\mathcal{S}} \|x - y\|$ as the distance from a point $x \in \mathbb{R}^d$ to a nonempty set $\mathcal{S} \subset \mathbb{R}^d$. The domain of the function $f : \mathbb{R}^d \to (-\infty, +\infty]$ is defined by $\text{dom}\, f = \{x \in \mathbb{R}^d : f(x) < +\infty\}$. We say that f is proper if $\text{dom}\, f \neq \emptyset$. For other generalized notions and definitions we refer to [8], [33], [34].

A. SMOOTH ADAPTABLE FUNCTIONS AND RELATIVE WEAKLY CONVEXITY

In this subsection, we define the notion of *smooth adaptable* condition for nonconvex f proposed in [8]. This property was extended from the recent work [2] in which the differentiable functions need to be convex. This condition is similar to the relative smoothness condition introduced in [24], but the relative smoothness is based on the fact that f is convex. As we want also to deal with nonconvex functions, in our paper we use the smooth adaptable condition to generalize Lipschitz gradient continuity and to derive the related convergence results of BPGe.

We first introduce the concept of Bregman distance needed in the definition of smooth adaptable condition. Is is based on the definition of kernel generating distance (also called Bregman function). The standard definition of Bregman function was given by Censor and Lent [12] based on the work of Bregman [10]. Other works on proximal algorithms with Bregman functions are listed in [13], [15], [21].

Definition 1 (Kernel Generating Distance and Bregman Distance [8]): Let S be a nonempty, convex and open subset of $\mathbb{R}^d$. Associated with S, a function $h : \mathbb{R}^d \to (-\infty, \infty]$ is called a kernel generating distance *if it satisfies the following:*

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

The function h is also called a Bregman function. We denote the class of kernel generating distances by $\mathcal{G}(S)$. Given $h \in \mathcal{G}(S)$, the Bregman distance *[10] is defined by $D_h : \text{dom}\, h \times \text{int dom}\, h \to [0, +\infty)$*

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

Many kinds of Bregman functions are illustrated in the literature [8], [36], like the Energy $r(x) = \frac{1}{2}x^2$ with $\text{dom}\, r = \mathbb{R}$, the Shannon Entropy $r(x) = x \log x$ with $\text{dom}\, r = [0, \infty]$, the Burg $r(x) = -\log x$ with $\text{dom}\, r = (0, \infty)$. Note that their derived Bregman distances are, obviously, proximity measures that measures the proximity of x and y, and they are widely used in applications.

Next, we list some basic properties of the Bregman distance [15], [36]:

(i) For any $(x, y) \in \text{dom}\, h \times \text{int dom}\, h$, $D_h(x, y) \geq 0$. If in addition h is strictly convex, $D_h(x, y) = 0$ if and only if $x = y$ holds.
(ii) **The three point identity**: For any $y, z \in \text{int dom}\, h$ and $x \in \text{dom}\, h$,

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

(iii) **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.

Throughout the paper we will focus on the pair of functions (f, h) that satisfies the smooth adaptable condition. Next we present the definition introduced in [8].

Definition 2 (L-Smooth Adaptable [8]): A pair of functions (f, h), such that $h \in \mathcal{G}(S)$, $f : \mathbb{R}^d \to (-\infty, +\infty]$ is a proper and lower-semi-continuous function with $\text{dom}\, h \subset \text{dom}\, f$, which is continuously differentiable on $S = \text{int dom}\, h$, is called L-smooth adaptable *(L-smad) on S if there exists $L > 0$ such that $Lh - g$ and $Lh + g$ are convex on S.*

According to [8, Lemma 2.1], the pair of functions (f, h) is L-smad on S if and only if $\|f(x) - f(y) - \langle \nabla f(y), x - y\rangle\| \leq L D_h(x, y)$ for any $(x, y) \in \text{int dom}\, h$. When $h(x) = \frac{1}{2}\|x\|^2$ and consequently $D_h(x, y) = \frac{1}{2}\|x - y\|^2$, the L-smad condition of f would be reduced to Lipschitz gradient continuity: $\|f(x) - f(y) - \langle \nabla f(y), x - y\rangle\| \leq \frac{L}{2}\|x - y\|^2$ for any $(x, y) \in \text{dom}\, h$.

Next, we introduce the definition of a μ-relative weakly convex function, given in [16]. This definition extends the definition of weakly convexity [29], which was employed in the analysis of nonconvex optimization methods.

Definition 3: f is called μ-relative weakly convex *to h on S if there exists $\mu > 0$ such that $f + \mu h$ is convex on S.*

When f is convex, $\mu = 0$. When (f, h) is L-smad on S, obviously f is L-relative weakly convex to h. So, by default, $\mu \leq L$. Now, just to give an example of a relative weakly convex function, we set $f(x) = \frac{1}{4}\sum_{i=1}^m (x^T A_i x - b_i)^2$ and $h(x) = \frac{1}{4}\|x\|_2^4 + \frac{1}{2}\|x\|_2^2$. Then, the pair (f, h) satisfies the L-smad condition when $L \geq \sum_{i=1}^m \left(3\|A_i\|^2 + \|A_i\|\|b_i|\right)$ and f is μ-relative weakly convex h when $\mu \geq \sum_{i=1}^m \|A_i\|\|b_i|$.

B. KURDYKA–ŁOJASIEWICZ PROPERTY

Finally, we introduce the definition of the Kurdyka–Łojasiewicz property proposed in [7]. We need this property to prove the global convergence of the whole sequences generated by BPGe for solving (1).

Definition 4: (Kurdyka–Łojasiewicz property [7]) Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper lower-semi-continuous function.

(i) The function f is said to have the Kurdyka–Łojasiewicz (KL) property at $\bar{x} \in \text{dom}\, \partial f := \{x \in \mathbb{R}^d : \partial f(x) \neq \emptyset\}$ if there exist $\eta \in (0, +\infty]$, a neighborhood U of $\bar{x}$ and a function $\psi : (0, \eta) \to \mathbb{R}_+$ satisfying:

$$\psi(0) = 0,\ \psi \in \mathcal{C}^1(0, \eta) \text{ and continuous at } 0,$$
$$\text{for all } s \in (0, \eta) :\ \psi'(s) > 0,$$

such that for all $x \in U \cap [f(\bar{x}) < f(x) < f(\bar{x}) + \eta]$, the following inequality holds

$$\psi'(f(x) - f(\bar{x})) \cdot dist(0, \partial f(x)) \geq 1.$$

(ii) If f satisfies the KL property at each point of dom ∂f then f is called a KL function.

The KL functions cover a large class of functions and some examples have been listed in the Appendix of [7].

III. BREGMAN PROXIMAL GRADIENT ALGORITHM WITH EXTRAPOLATION (BPGE)

Throughout this paper, we focus on the nonconvex problem (1) of Section I, and we assume that the kernel generating distance function $h \in \mathcal{G}(\mathbb{R}^d)$, (f, h) is L-smad and f is μ-weakly convex relative to h (see Definitions 2 and 3). In addition, we also suppose the following general Assumptions 1 and 2.

Assumption 1 is a quite standard condition [8] to guarantee the existence of the solution to each step of the optimal subproblem of Proximal Gradient (PG) algorithms.

Assumption 1: The function Ψ is supercoercive*, that is,*

$$\lim_{\|u\|\to\infty} \frac{\Psi(u)}{\|u\|} = \infty.$$

Assumptions 2 is a general assumption used in the analysis of Bregman Proximal-type algorithms [2], [15].

Assumption 2: 1) h is strictly convex.

2) If $\{x^k\}_{k\in\mathbb{N}}$ converges to some x in dom h then $D_h(x, x^k) \to 0$.

3) If $\{x^k\}_{k\in\mathbb{N}}, \{y^k\}_{k\in\mathbb{N}}$ defined in dom h are sequences such that $y^k \to x^* \in \overline{\text{dom } h}$, $\{x^k\}_{k\in\mathbb{N}}$ is bounded, and if $D_h(x^k, y^k) \to 0$, then $x^k \to x^*$.

Algorithm 1 BPGe–Bregman Proximal Gradient Algorithm With Extrapolation

Data: A function h defined in Definition 1 such that (f, h) is L-smad holds and f is μ-weakly convex relative to h on $\mathbb{R}^d$. Error tolerance: `TOL`.

Initialization: $x^0 = x^{-1} \in \text{int dom } h$ and $0 < \lambda_k \le 1/L$.

General step:

`For` $k = 0, 1, 2, \ldots, k_{max}$ `repeat`

$$y^k = x^k + \beta_k(x^k - x^{k-1}), \tag{1}$$

where β_k is searched according to **Line Search** in Algorithm 2.

Then compute

$$x^{k+1} \in \arg\min_{x\in\mathbb{R}^d} \left\{ g(x) + \left\langle x - y^k, \nabla f(y^k) \right\rangle + \frac{1}{\lambda_k} D_h(x, y^k) \right\}. \tag{2}$$

until `EXIT(TOL)` received.

We are now ready to introduce our BPGe algorithm, divided in two parts, Algorithm 1 and Algorithm 2. Algorithm 1 is the whole framework for solving Problem (1). And Algorithm 2 is a line search step, which is used to search a proper parameter β_k at every iteration in Algorithm 1. Throughout the whole paper, we make the following notations

$$\overline{\lambda} := \sup_{k\in\mathbb{N}}\{\lambda_k\}, \quad \underline{\lambda} := \inf_{k\in\mathbb{N}}\{\lambda_k\}.$$

By default $0 < \underline{\lambda} \le \overline{\lambda} < \infty$.

Algorithm 2 Line Search for Algorithm 1 at the k-th Iteration

Data: A function h defined in Algorithm 1, fix $0 < \eta < 1$, $\beta_0 \in [0, 1)$, $0 < \rho < 1$.

Input: $x^{k-1}, x^k \in \text{int dom } h$, $C_k = \frac{\lambda_k^{-1}}{\lambda_k^{-1} + \mu}$.

General step:

$\tilde{\beta} = \beta_0$,

`While` $D_h(x^k, x^k + \tilde{\beta}(x^k - x^{k-1})) > \rho\, C_k\, D_h(x^{k-1}, x^k)$ `do`

$$\tilde{\beta} = \eta\tilde{\beta}.$$

Return: Set the feasible step size $\beta_k = \tilde{\beta}$ for iteration k.

We remark that an important point on any iterative process is to define suitable error control techniques. In this paper we consider a quite simple strategy in order to determine the `EXIT` conditions. On one hand we fix a maximum number of iterations k_{max} (in most of our tests 5000 iterations) and `EXIT(TOL)` = `true` if $\|x^k - x^{k-1}\|/\max\{1, \|x^k\|\} \le$ `TOL` (in our tests `TOL` $= 10^{-6}$ as in [39]). Other option is to check the convergence using the objective function, instead of the solution itself, that is $\|\Psi(x^k) - \Psi(x^{k-1})\|/\max\{1, \|\Psi(x^k)\|\} \le$ `TOL`.

We first verify that (2) is well-defined using the following Proposition 1. For all $y \in \text{int dom } h$ and stepsize $0 < \lambda \le 1/L$, we define the Bregman proximal gradient mapping as:

$$T_\lambda(y) := \arg\min_{u\in\mathbb{R}^d} \left\{ g(u) + \langle \nabla f(y), u - y\rangle + \lambda^{-1} D_h(u, y) \right\}.$$

In Proposition 1 we prove that $T_\lambda(y)$ is well posed. Thus by Proposition 1, $x^{k+1} \in T_{\lambda_k}(x^k)$, and fixing $\inf\{\lambda_k\} > 0$, then Step (2) in BPGe algorithm is well-defined.

Proposition 1: Suppose that Assumption 1 holds, let $y \in$ int dom h and $0 < \lambda \le 1/L$. Then, the set $T_\lambda(y)$ is a nonempty and compact set.

Proof: Fix any $y \in \text{int dom } h$ and $0 < \lambda \le 1/L$. For any $u \in \mathbb{R}^d$, we define

$$\Psi_h(u) = g(u) + f(y) + \left\langle u - y, \nabla f(y) \right\rangle + \lambda^{-1} D_h(u, y),$$

so that $T_\lambda(y) = \arg\min_{u\in\mathbb{R}^d} \Psi_h(u)$, It can also be represented as

$$\begin{aligned} \Psi_h(u) &= \Psi(u) - f(u) + f(y) + \left\langle u - y, \nabla f(y) \right\rangle + \lambda^{-1} D_h(u, y) \\ &\ge \Psi(u) + L\, D_h(u, y) - \left[f(u) - f(y) - \left\langle u - y, \nabla f(y) \right\rangle \right] \\ &\ge \Psi(u). \end{aligned}$$

where the second inequality is obtained by taking into account $\lambda^{-1} \ge L$ and in the last inequality that (f, h) is L-smooth adaptable. According to Assumption 1, i.e. $\lim_{\|u\|\to\infty} \Psi(u) = \infty$, there is

$$\lim_{\|u\|\to\infty} \Psi_h(u) \ge \lim_{\|u\|\to\infty} \Psi(u) = \infty.$$

Since Ψ_h is also proper and lower-semi-continuous, invoking the modern form of Weierstrass' theorem (see, e.g., [33, Theorem 1.9, page 11]), it follows that the value $\inf_{\mathbb{R}^d} \Psi_h$ is finite, and the set $\arg\min_{u\in\mathbb{R}^d} \Psi_h(u) \equiv T_\lambda(y)$ is nonempty and compact. □

Secondly, we add an extrapolation step to the BPGe algorithm to choose a suitable β_k at each iteration step through the line search Algorithm 2. On this step it is hard to guarantee directly the decrease of the function value $\Psi(x^k)$. Therefore, we focus on guaranteeing sufficient decrease of the Lyapunov sequences defined in Section IV in the convergence analysis. However, it still requires an extra condition $D_h(x^k, x^k + \beta_k(x^k - x^{k-1})) \le \rho\, C_k\, D_h(x^{k-1}, x^k)$. When $h = \frac{1}{2}\|x\|^2$, BPGe is reduced to the PGe algorithm [39] and this condition is easily satisfied by setting $0 \le \beta_k \le \sqrt{\rho \frac{L}{L+\mu}}$. But when h is more general and complex, computing the threshold of β_k directly may be hard and expensive. So, we try to reach this condition by a line search method introduced in Algorithm 2. Thus, our next step is to verify that Algorithm 2 is well-defined, as the following proposition 2 shows.

Proposition 2 (Finite Termination of Algorithm 2): Consider Algorithm 1 and fix $k \in \mathbb{N}$. Let $0 < \eta < 1, 0 < \rho < 1$, $\tilde{\beta} \in [0, 1)$, $C_k = \frac{\lambda_k^{-1}}{\lambda_k^{-1}+\mu} > 0$. Then, there exists $J \in \mathbb{N}$ such that $\beta_k := \eta^j \tilde{\beta}$ satisfies

$$D_h(x^k, x^k + \beta_k(x^k - x^{k-1})) \le \rho\, C_k\, D_h(x^{k-1}, x^k)$$

for any $j \ge J$.

Proof: This result is proved by contradiction. Suppose that

$$D_h(x^k, x^k + \eta^j \tilde{\beta}(x^k - x^{k-1})) > \rho\, C_k\, D_h(x^{k-1}, x^k)$$

holds for any $j \in \mathbb{N}$.

When $x^k = x^{k-1}$, Algorithm 2 terminates with $\beta_k = \tilde{\beta}$ directly.

When $x^k \ne x^{k-1}$, since

$$\|x^k - (x^k + \tilde{\beta}(x^k - x^{k-1}))\| = \eta^j \tilde{\beta}\|x^k - x^{k-1}\| \to 0, \quad j \to \infty,$$

according to Assumption 2(2), $D_h\big(x^k, x^k + \eta^j\tilde{\beta}(x^k - x^{k-1})\big) \to 0$. Thus for any $\varepsilon > 0$, there exist a number $J \in \mathbb{N}$ such that

$$D_h(x^k, x^k + \eta^j \tilde{\beta}(x^k - x^{k-1})) < \varepsilon, \quad \text{for all } j \ge J.$$

Since $x^k \ne x^{k-1}$, and due to the strictly convexity of h in Assumption 2(1),

$$D_h(x^{k-1}, x^k) > 0.$$

If we set $\varepsilon = \frac{1}{2}\rho\, C_k\, D_h(x^{k-1}, x^k)$, then

$$\rho\, C_k\, D_h(x^{k-1}, x^k) < D_h\left(x^k, x^k + \eta^j\tilde{\beta}(x^k - x^{k-1})\right) < \frac{1}{2}\rho\, C_k\, D_h(x^{k-1}, x^k),$$

for $j \ge J$, which is a contradiction. □

IV. CONVERGENCE ANALYSIS OF BPGE

In this section we provide the main convergence results of the BPGe algorithm. First of all, following the analysis of Remark 4.1(ii) in [8], we obtain the following Lemma 1. We find that after adding an extrapolation term, it is hard to justify monotonicity of the objective function Ψ directly. But for a special auxiliary sequence, defined by

$$H_{k,M} := \Psi(x^k) + M D_h(x^{k-1}, x^k), \quad M > 0, \quad \forall k \in \mathbb{N}$$

the monotone property will be presented in our settings.

Lemma 1: For any $x \in$ int dom h, and let be a sequence $\{x^k\}_{k\in\mathbb{N}}$ produced by BPGe, then

(i) For any $k \in \mathbb{N}$, we have

$$\Psi(x^{k+1}) - \Psi(x) \le (\lambda_k^{-1} + \mu)\, D_h(x, y^k) - \lambda_k^{-1} D_h(x, x^{k+1}) - (\lambda_k^{-1} - L)\, D_h(x^{k+1}, y^k). \quad (3)$$

(ii) For any $k \in \mathbb{N}$, we have

$$H_{k+1,M} - H_{k,M} \le (M - \lambda_k^{-1})\, D_h(x^k, x^{k+1}) - \left(M - \rho\lambda_k^{-1}\right) D_h(x^{k-1}, x^k). \quad (4)$$

Moreover, assuming there exists some M such that $\rho\,\underline{\lambda}^{-1} \le M \le \overline{\lambda}^{-1}$, then the sequence $\{H_{k,M}\}$ is nonincreasing and convergent for the fixed M.

Proof: (i) According to the first order condition of (2), we get

$$0 \in \partial g(x^{k+1}) + \nabla f(y^k) + \lambda_k^{-1}\big(\nabla h(x^{k+1}) - \nabla h(y^k)\big), \quad \forall k \in \mathbb{N}.$$

Combining with the convexity of g, there is

$$g(x) - g(x^{k+1}) \ge \left\langle -\nabla f(y^k) - \lambda_k^{-1}\big(\nabla h(x^{k+1}) - \nabla h(y^k)\big),\, x - x^{k+1}\right\rangle,$$

for all $k \in \mathbb{N}$. Together with the three point identity of Bregman distance

$$\lambda_k^{-1}\left\langle \nabla h(x^{k+1}) - \nabla h(y^k),\, x - x^{k+1}\right\rangle = \lambda_k^{-1}\left(D_h(x, y^k) - D_h(x, x^{k+1}) - D_h(x^{k+1}, y^k)\right) \quad (5)$$

we have that

$$g(x) - g(x^{k+1}) + f(x) - f(x^{k+1}) \ge f(x) - f(x^{k+1}) - \left\langle \nabla f(y^k),\, x - x^{k+1}\right\rangle - \lambda_k^{-1}\left(D_h(x, y^k) - D_h(x, x^{k+1}) - D_h(x^{k+1}, y^k)\right), \quad (6)$$

for all $k \in \mathbb{N}$. If we take the μ-relative weakly convex property and L-smad property of (f, h) (see Definitions 2 and 3),

$$\begin{aligned} f(x) - f(x^{k+1}) - \left\langle \nabla f(y^k),\, x - x^{k+1}\right\rangle &= f(x) - f(y^k) - \left\langle \nabla f(y^k),\, x - y^k\right\rangle + f(y^k) - f(x^{k+1}) - \left\langle \nabla f(y^k),\, y^k - x^{k+1}\right\rangle \\ &\ge -\mu\, D_h(x, y^k) - L\, D_h(x^{k+1}, y^k), \quad \forall k \in \mathbb{N}. \end{aligned} \quad (7)$$

Thus

$$\Psi(x^{k+1}) - \Psi(x) \le (\lambda_k^{-1} + \mu) D_h(x, y^k) - \lambda_k^{-1} D_h(x, x^{k+1}) - (\lambda_k^{-1} - L) D_h(x^{k+1}, y^k).$$

(ii) For any $k \in \mathbb{N}$, taking $x = x^k$ into (3), together with $L \le \lambda_k^{-1}$, $D_h(x^{k+1}, y^k) \ge 0$ we get

$$\Psi(x^{k+1}) - \Psi(x^k) \le (\lambda_k^{-1} + \mu) D_h(x^k, y^k) - \lambda_k^{-1} D_h(x^k, x^{k+1}).$$

If $x^k = x^{k-1}$, we get $y^k = x^k$, thus $D_h(x^k, y^k) = D_h(x^{k-1}, x^k) = 0$ and

$$\Psi(x^{k+1}) + \lambda_k^{-1} D_h(x^k, x^{k+1}) \le \Psi(x^k) + (\lambda_k^{-1} + \mu)\, \rho\, C_k\, D_h(x^{k-1}, x^k). \quad (8)$$

If $x^k \ne x^{k-1}$, according to Algorithm 2, we have $D_h(x^k, y^k) \le \rho\, C_k\, D_h(x^{k-1}, x^k)$, thus

$$\Psi(x^{k+1}) + \lambda_k^{-1} D_h(x^k, x^{k+1}) \le \Psi(x^k) + (\lambda_k^{-1} + \mu)\, \rho\, C_k\, D_h(x^{k-1}, x^k). \quad (9)$$

Combining these two cases, we obtain

$$\Psi(x^{k+1}) + \lambda_k^{-1} D_h(x^k, x^{k+1}) \le \Psi(x^k) + (\lambda_k^{-1} + \mu)\, \rho\, C_k\, D_h(x^{k-1}, x^k), \quad \forall k \in \mathbb{N}.$$

From the definition of $H_{k,M}$, we see that

$$H_{k+1,M} - H_{k,M} \le (M - \lambda_k^{-1}) D_h(x^k, x^{k+1}) - \left(M - \rho\lambda_k^{-1}\right) D_h(x^{k-1}, x^k), \quad \forall k \in \mathbb{N}.$$

Furthermore, assuming there exists some M such that

$$\rho\, \lambda_k^{-1} \le \rho\, \underline{\lambda}^{-1} \le M \le \overline{\lambda}^{-1} \le \lambda_k^{-1},$$

and fixing one of such values of M, we find that

$$H_{k+1,M} - H_{k,M} \le 0, \quad \forall k \in \mathbb{N},$$

that is, $\{H_{k,M}\}_{k\in\mathbb{N}}$ is nonincreasing for the fixed value of M.

Recall that $H_{k,M} \ge \inf \Psi > -\infty$ and $H_{k,M}$ is nonincreasing. This implies that $\{H_{k,M}\}$ is convergent for some fixed M. □

The next corollary is an obvious result based on Lemma 1. We analyze the boundness of the sequences produced by BPGe algorithm. Since $H_{k,M}$ is nonincreasing according to Lemma 1(ii), it is easy to verify that the sequence $\{x^k\}_{k\in\mathbb{N}}$ generated by BPGe is bounded according to Assumption 1. The boundness would act as a tool in the following analysis, so we present this result as the auxiliary Corollary 1.

Corollary 1: Assume there exists some M such that $\rho\, \underline{\lambda}^{-1} \le M \le \overline{\lambda}^{-1}$, then the sequence $\{x^k\}_{k\in\mathbb{N}}$ generated by BPGe is bounded.

If the stepsize λ_k and parameter ρ in Algorithm 2 satisfy $\rho < \overline{\lambda}^{-1}/\underline{\lambda}^{-1} = \underline{\lambda}/\overline{\lambda}$, then we could get sufficient decrease of the auxiliary sequence $\{H_{k,M}\}_{k\in\mathbb{N}}$ for the fixed M given in Lemma 1. As a consequence, we can bound the sum of Bregman distance between two iteration points generated by BPGe. Moreover, adding stronger assumptions than Assumption 2 on the kernel generating distance h, such as strong convexity, we could get that $\lim_{k\to\infty} \|x^k - x^{k-1}\| = 0$ for the sequence $\{x^k\}_{k\in\mathbb{N}}$ in $\mathbb{R}^d$ by BPGe. In this paper, we just consider the set of weaker blanket Assumptions 1 and 2, that permit us to prove that any limit point of the sequence $\{x^k\}_{k\in\mathbb{N}}$ generated by BPGe, if exists, is a stationary point of the objective function Ψ.

Assume that $\{x^k\}_{k\in\mathbb{N}}$ is generated from a starting point x^0. The set of all limit points of $\{x^k\}_{k\in\mathbb{N}}$ is denoted by

$$\omega(x^0) := \{\overline{x} : \text{ an increasing sequence of integers } \{k_i\}_{i\in\mathbb{N}} \text{ such that } x^{k_i} \to \overline{x} \text{ as } i \to \infty\}.$$

The next technical lemma shows, among other results, that for any $x^0 \in \mathbb{R}^d$, $\omega(x^0) \subseteq \text{crit } \Psi$ holds.

Lemma 2: Suppose $\rho < \underline{\lambda}/\overline{\lambda}$ and let $\{x^k\}_{k\in\mathbb{N}}$ be a sequence generated from x^0 by BPGe. Then the following statements hold:

(i) $\sum_{k=0}^{\infty} D_h(x^{k-1}, x^k) < \infty$ and $\lim_{k\to\infty} D_h(x^{k-1}, x^k) = 0$.

(ii) Any limit point of $\{x^k\}_{k\in\mathbb{N}}$ is a critical point of Ψ ($\omega(x^0) \subseteq$ crit Ψ).

(iii) $\zeta := \lim_{k\to\infty} \Psi(x^k)$ exists and $\Psi \equiv \zeta$ on $\omega(x^0)$.

Proof: (i) Since $\rho < \underline{\lambda}/\overline{\lambda}$, we have that $\rho\, \lambda_k^{-1} \le \rho\, \underline{\lambda}^{-1} < \overline{\lambda}^{-1}$, and we choose $M \in (\rho\, \underline{\lambda}^{-1}, \overline{\lambda}^{-1}]$. From (4), together with the nonnegativeness of $D_h(x^k, x^{k+1})$ and $M \le \lambda_k^{-1}$, we have $\forall k \in \mathbb{N}$

$$\left(M - \rho\underline{\lambda}^{-1}\right) D_h(x^{k-1}, x^k) \le \left(M - \rho\lambda_k^{-1}\right) D_h(x^{k-1}, x^k) \le H_{k,M} - H_{k+1,M}, \quad (10)$$

which implies, $\forall K \in \mathbb{N}$, that

$$0 \le \sum_{i=0}^{K} \left(M - \rho\underline{\lambda}^{-1}\right) D_h(x^{k-1}, x^k) \le H_{0,M} - H_{K+1,M}, \quad (11)$$

by summing both sides of (10) from 0 to K. Since $\{H_{k,M}\}$ is convergent by Lemma 1(ii), letting $K \to \infty$, we conclude that the infinite sum exists and is finite, i.e.,

$$\sum_{i=0}^{K} \left(M - \rho\underline{\lambda}^{-1}\right) D_h(x^{k-1}, x^k) < \infty.$$

Since $M - \rho\underline{\lambda}^{-1} > 0$, we obtain directly that $\sum_{i=0}^{K} D_h(x^{k-1}, x^k) \le \infty$ and $\lim_{k\to\infty} D_h(x^{k-1}, x^k) = 0$.

(ii) Let $\overline{x}$ be a limit point of $\{x^k\}_{k\in\mathbb{N}}$. Let $\{x^{k_i}\}$ be a subsequence such that $\lim_{i\to\infty} x^{k_i} = \overline{x}$. Since $D_h(x^{k_i-1}, x^{k_i}) \to 0$, and we know $\{x^{k_i-1}\}_{i\in\mathbb{N}}$ is bounded according to Corollary 1, Assumption 1(ii) implies $x^{k_i-1} \to \overline{x}$. Similarly, we get $x^{k_i-2} \to \overline{x}$. By the representation of $y^{k_i-1} = x^{k_i-1} + \beta_{k_i-1}(x^{k_i-1} - x^{k_i-2})$ or $y^{k_i-1} = x^{k_i-1}$ (if $x^{k_i-1} = x^{k_i-2}$),

we obtain

$$\begin{aligned}\|y^{k_i-1}-x^{k_i}\| &\le \|x^{k_i-1}-x^{k_i}\|+\|x^{k_i-1}-x^{k_i-2}\|\\ &\le \|x^{k_i}-\bar{x}\|+2\|x^{k_i-1}-\bar{x}\|+\|x^{k_i-2}-\bar{x}\|\to 0.\end{aligned} \tag{12}$$

On one hand, we prove that there exists $v^{k_i}\in\partial\Psi(x^{k_i})$ such that $v^{k_i}\to 0$. By using the first-order optimality condition of the minimization problem (2), we obtain

$$0\in\lambda_{k_i-1}\partial g(x^{k_i})+\lambda_{k_i-1}\nabla f(y^{k_i-1})+\nabla h(x^{k_i})-\nabla h(y^{k_i-1}),$$

for all $k_i\in\mathbb{N}$. Therefore, we observe that

$$\nabla f(x^{k_i})-\nabla f(y^{k_i-1})-\lambda_{k_i-1}^{-1}\big(\nabla h(x^{k_i})-\nabla h(y^{k_i-1})\big)\in\partial\Psi(x^{k_i}), \tag{13}$$

for all $k_i\in\mathbb{N}$. Taking limits on the left hand in (13) we have that

$$\begin{aligned}&\|\nabla f(x^{k_i})-\nabla f(y^{k_i-1})-\lambda_{k_i-1}^{-1}(\nabla h(x^{k_i})-\nabla h(y^{k_i-1}))\|\\ &\le\|\nabla f(x^{k_i})-\nabla f(y^{k_i-1})\|+\underline{\lambda}^{-1}\|\nabla h(x^{k_i})-\nabla h(y^{k_i-1})\|\to 0,\end{aligned} \tag{14}$$

as $k_i\to\infty$. where the limit can be got according to (12) and the continuity of ∇f and ∇h. Thus, we get that there exist $v^{k_i}\in\partial\Psi(x^{k_i})$ such that $\|v^{k_i}\|\to 0$ as $k_i\to\infty$.

On the other hand, we derive that $\Psi(x^{k_i})\to\Psi(\bar{x})$, $k_i\to\infty$. From the lower-semi-continuity of Ψ, we have

$$\Psi(\bar{x})\le\liminf_{i\to\infty}\Psi(x^{k_i}). \tag{15}$$

According to the iteration step (2) of BPGe, for $k_i\ge 1$, we have

$$\begin{aligned}\lambda_{k_i-1}\,g(x^{k_i})+\left\langle x^{k_i}-\bar{x},\lambda_{k_i-1}\nabla f(y^{k_i-1})\right\rangle+D_h(x^{k_i},y^{k_i-1})\\ \le\lambda_{k_i-1}g(\bar{x})+D_h(\bar{x},y^{k_i-1}).\end{aligned}$$

Adding $\lambda_{k_i-1}f(x^{k_i})$ to both sides, we have

$$\begin{aligned}\lambda_{k_i-1}\Psi(x^{k_i})+\left\langle x^{k_i}-\bar{x},\lambda_{k_i-1}\nabla f(y^{k_i-1})\right\rangle+D_h(x^{k_i},y^{k_i-1})\\ \le\lambda_{k_i-1}g(\bar{x})+\lambda_{k_i-1}f(x^{k_i})+D_h(\bar{x},y^{k_i-1}),\end{aligned} \tag{16}$$

for all $k_i\in\mathbb{N}$. After rearranging terms, for all $k_i\in\mathbb{N}$, it follows

$$\begin{aligned}\Psi(x^{k_i})\le\Psi(\bar{x})+f(x^{k_i})-f(\bar{x})-\left\langle x^{k_i}-\bar{x},\nabla f(y^{k_i-1})\right\rangle\\ -\lambda_{k_i-1}^{-1}D_h(x^{k_i},y^{k_i-1})+\lambda_{k_i-1}^{-1}D_h(\bar{x},y^{k_i-1}).\end{aligned} \tag{17}$$

L-smad property and μ-relative weakly convexity of (f,h) imply that for all $k_i\in\mathbb{N}$

$$\begin{aligned}&f(x^{k_i})-f(\bar{x})-\left\langle x^{k_i}-\bar{x},\nabla f(y^{k_i-1})\right\rangle\\ &\le L\,D_h(x^{k_i},\bar{x})+\left\langle x^{k_i}-\bar{x},\ \nabla f(\bar{x})-\nabla f(y^{k_i-1})\right\rangle\\ &= L\,D_h(x^{k_i},\bar{x})+D_f(x^{k_i},y^{k_i-1})-D_f(x^{k_i},\bar{x})-D_f(\bar{x},y^{k_i-1}).\\ &\le L\,D_h(x^{k_i},\bar{x})+L\,D_h(x^{k_i},y^{k_i-1})+\mu\,D_h(x^{k_i},\bar{x})\\ &\quad+\mu\,D_h(\bar{x},y^{k_i-1})\end{aligned} \tag{18}$$

Plugging (18) in (17), passing to the limit, together with the relationship $\underline{\lambda}\le\lambda_{k_i}\le\bar{\lambda}$, we have

$$\begin{aligned}&\lim_{i\to\infty}\Psi(x^{k_i})\\ &\le\Psi(\bar{x})+\lim_{i\to\infty}\big[(-\bar{\lambda}^{-1}+L)D_h(x^{k_i},y^{k_i-1})\\ &\quad+(\underline{\lambda}^{-1}+\mu)D_h(\bar{x},y^{k_i-1})+(L+\mu)D_h(x^{k_i},\bar{x})\big]\\ &\le\Psi(\bar{x})+\lim_{i\to\infty}(\underline{\lambda}^{-1}+\mu)\Big[D_h(\bar{x},y^{k_i-1})+D_h(x^{k_i},\bar{x})\Big],\end{aligned}$$

where the second inequality is based on $L\le\bar{\lambda}^{-1}\le\underline{\lambda}^{-1}$ in BPGe. From (12), together with the continuity of ∇h, we obtain

$$\begin{aligned}&\lim_{i\to\infty}D_h(\bar{x},y^{k_i-1})+D_h(x^{k_i},\bar{x})\\ &\le\lim_{i\to\infty}D_h(\bar{x},y^{k_i-1})+D_h(y^{k_i-1},\bar{x})+D_h(x^{k_i},\bar{x})+D_h(\bar{x},x^{k_i})\\ &\le\lim_{i\to\infty}\|\nabla h(y^{k_i-1})-\nabla h(\bar{x})\|\|y^{k_i-1}-\bar{x}\|\\ &\quad+\|\nabla h(x^{k_i})-\nabla h(\bar{x})\|\|x^{k_i}-\bar{x}\|\\ &=0.\end{aligned}$$

Hence we have

$$\limsup_{i\to\infty}\Psi(x^{k_i})\le\Psi(\bar{x}). \tag{19}$$

Combining (15) and (19) yields $\Psi(x^{k_i})\to\Psi(\bar{x})$, $k_i\to\infty$.

Thus, according to these results, and the closedness of $\partial\Psi$ (see, Exercise 8 in Page 80 [9]), we have $0\in\partial\Psi(\bar{x})$.

(iii) In view of Lemma 1 and (i), the sequence $\{H_{k,M}\}$ is convergent and $D_h(x^{k-1},x^k)\to 0$. These, together with the definition of $H_{k,M}$, implies $\lim_{k\to\infty}\Psi(x^k)$ exists, denoted as ζ. According to the last part of the proof in (ii), and taking $\bar{x}\in\omega(x^0)$ with a convergent subsequence $\{x^{k_i}\}$ such that $\lim_{i\to\infty}x^{k_i}=\bar{x}$, we know that

$$\zeta=\lim_{i\to\infty}\Psi(x^{k_i})=\Psi(\bar{x}).$$

Thus the conclusion is completed since $\bar{x}$ is arbitrary. □

Next, we prove a global $\mathcal{O}(\frac{1}{K})$ sublinear convergence rate for the sequence $\min_{k\in\mathbb{N}}D_h(x^{k-1},x^k)$ of the algorithm. In fact, the linear convergence rate can also be got if we add more assumptions, like KL property and concrete KL exponent (we refer to [22]), based on similar deductions as in [8, Theorem 6.3].

Corollary 2: Suppose $\rho<\underline{\lambda}/\bar{\lambda}$ *and* $\{x^k\}_{k\in\mathbb{N}}$ *be a sequence generated from* x^0 *by BPGe. Then for all* $K\ge 1$, $\min_{1\le k\le K}D_h(x^{k-1},x^k)$ *converges with a sublinear rate as* $\mathcal{O}(\frac{1}{K})$.

Proof: Set $M=\bar{\lambda}^{-1}$, recall (11), now for $K\ge 1$,

$$0\le\sum_{i=1}^{K}\left(\bar{\lambda}^{-1}-\rho\,\underline{\lambda}^{-1}\right)D_h(x^{k-1},x^k)\le H_{1,M}-H_{K+1,M}.$$

Hence we obtain

$$\min_{1\le k\le K}D_h(x^{k-1},x^k)\le\frac{H_{1,M}-H_{K+1,M}}{K\left(\bar{\lambda}^{-1}-\rho\,\underline{\lambda}^{-1}\right)}. \tag{20}$$

□

Next, we focus on performing a global convergence analysis. We aim to prove that the sequence $\{x^k\}_{k\in\mathbb{N}}$ generated by BPGe converges to a critical point of the objective function Ψ defined in (1). In order to prove global convergence, we use the proof methodology introduced in [1]. This proof methodology proves global convergence result for several types of nonconvex nonsmooth problems. Other similar forms were referred in many works [30, Section 3.2], [32, Section 4], [8, Section 4.2].

For the reader's convenience, we firstly describe the proof methodology summarized in [30, Theorem 3.7] with a few modifications and then we apply it to prove the convergence of BPGe in Theorem 2.

Theorem 1: [30, Theorem 3.7] Let $F : \mathbb{R}^{2d} \to (-\infty, \infty]$ *be a proper lower-semi-continuous function. Assume that* $\{z^k\}_{k\in\mathbb{N}} := \{(x^k, x^{k-1})\}_{k\in\mathbb{N}}$ *is a sequence generated by a general algorithm from* $z^0 := (x^0, x^0)$*, for which the following three hypotheses are satisfied for any* $k \in \mathbb{N}$.

(H1) *For each* $k \in \mathbb{N}$*, there exists a positive 'a' such that*

$$F(z^{k+1}) + a\,\|x^k - x^{k-1}\|^2 \le F(z^k), \quad \forall k \in \mathbb{N}.$$

(H2) *For each* $k \in \mathbb{N}$*, there exists a positive 'b' such that for some* $v^{k+1} \in \partial F(z^{k+1})$ *we have*

$$\|v^{k+1}\| \le \frac{b}{2}\big(\|x^{k+1} - x^k\| + \|x^k - x^{k-1}\|\big), \quad \forall k \in \mathbb{N}.$$

(H3) *There exists a subsequence* $(z^{k_j})_{j\in\mathbb{N}}$ *such that* $z^{k_j} \to \tilde{z}$ *and* $F(z^{k_j}) \to F(\tilde{z})$.

Moreover, if F *have the Kurdyka-Łojasiewicz property at the limit point* $\tilde{z} = (\tilde{x}, \tilde{x})$ *specified in (H3), then, the sequence* $\{x^k\}_{k\in\mathbb{N}}$ *has finite length, i.e.,* $\sum_{k=1}^{\infty} \|x^k - x^{k-1}\| < \infty$*, and converges to* $\bar{x} = \tilde{x}$ *as* $k \to \infty$*, where* $(\bar{x}, \bar{x})$ *is a critical point of* F.

In our paper, what we need is to verify that the hypotheses given in Theorem 1 are satisfied for $F(x, y) = \Psi(x) + MD_h(y, x)$ and the sequence $(x^k, x^{k-1})_{k\in\mathbb{N}} \in \mathbb{R}^{2d}$ generated by the BPGe algorithm.

In order to guarantee the three hypotheses of the Theorem hold, we need another extra assumption (the following Assumption 3). Note that the first two requirements of the assumption were also required in [8, see Assumption D(ii)], and the third assumption is easily verified.

Assumption 3: 1) h *is* σ*-strongly convex on* $\mathbb{R}^d$.

2) $\nabla h, \nabla f$ *are Lipschitz continuous on any bounded subset of* $\mathbb{R}^d$.

3) *There exists a bounded* u *such that* $u \in \partial^2 h$ *on any bounded subset of* $\mathbb{R}^d$.

In fact, Assumption 3(1-2) can guarantee that Assumption 1(2-3) hold for the bounded sequence $\{x_k\}_{k\in\mathbb{N}}$.

The next task is to verify the three hypotheses one by one. Then, together with Theorem 1, we obtain the result that, under proper parameter selection, the whole sequence generated by BPGe converges to a critical point of the objective function.

Theorem 2: Suppose $\rho < \underline{\lambda}/\overline{\lambda}$*. Let* $\{x^k\}_{k\in\mathbb{N}}$ *be a sequence generated from* x^0 *by BPGe. If* $F(x, y) = \Psi(x) + MD_h(y, x)$*(where* $M \in (\rho\,\underline{\lambda}^{-1}, \overline{\lambda}^{-1}]$*) satisfies the Kurdyka–Łojasiewicz property at some limit point* $\tilde{z} = (\tilde{x}, \tilde{x}) \in \mathbb{R}^{2d}$ *and Assumption 3 holds, then*

(i) The sequence $\{x^k\}_{k\in\mathbb{N}}$ *has finite length, i.e.* $\sum_{k=1}^{\infty} \|x^k - x^{k-1}\| < \infty$.

(ii) $x^k \to \tilde{x}$ *as* $k \to \infty$*, and* $\tilde{x}$ *is a critical point of* Ψ.

Proof: We first verify the three hypotheses of the Theorem 1 for function H and BPGe algorithm.

(H1) According to Assumption 3, since h is strongly convex, assume that h is σ-strongly convex, that is $D_h(x, y) \ge \frac{\sigma}{2}\|x - y\|^2$ for any $x, y \in \mathbb{R}^d$. We denote $a = \frac{\sigma}{2}(M - \rho\,\underline{\lambda}^{-1})$. For any $k \in \mathbb{N}$,

$$\begin{aligned} &F(x^{k+1}, x^k) + a\|x^k - x^{k-1}\|^2 \\ &\le F(x^{k+1}, x^k) + (M - \rho\underline{\lambda}^{-1})D_h(x^{k-1}, x^k) \\ &\le F(x^{k+1}, x^k) + (M - \rho\lambda_k^{-1})D_h(x^{k-1}, x^k) \\ &= H_{k+1,M} + (M - \rho\lambda_k^{-1})D_h(x^{k-1}, x^k) \\ &\le H_{k,M} + (M - \lambda_k^{-1})\,D_h(x^k, x^{k+1}) \\ &\le H_{k,M} = F(x^k, x^{k-1}), \end{aligned}$$

where the first inequality is based on the strongly convexity of h, the second inequality is based on $\underline{\lambda} \le \lambda_k$, the third and the last equality is from the definitions of $H_{k,M}$ and F, the fourth inequality is from Lemma 1(ii), and the fifth inequality is according to the nonnegativeness of $(M - \lambda_k^{-1})\,D_h(x^k, x^{k+1})$. Thus (H1) is verified.

(H2) From the optimal condition (2), there exists $-\nabla f(y^k) + \lambda_k^{-1}(\nabla h(y^k) - \nabla h(x^{k+1})) \in \partial g(x^{k+1})$. Due to Corollary 1, $\{x^k\}_{k\in\mathbb{N}}$ generated by BPGe is bounded, and so also $\{y^k\}_{k\in\mathbb{N}}$ is bounded. Thus, according to Assumption 3(iii), there exists a bounded $u_k \in \partial^2 h(x^k)$, and

$$\begin{aligned} &v_{k+1} \\ &= \Big(\nabla f(x^{k+1}) - \nabla f(y^k) - \lambda_k^{-1}(\nabla h(x^{k+1}) - \nabla h(y^k)) \\ &\quad - M\langle u_k, x^{k+1} - x^k\rangle,\ M\big(\nabla h(x^k) - \nabla h(x^{k+1})\big)\Big), \end{aligned}$$

such that $v_{k+1} \in \partial F(x^{k+1}, x^k)$. According to Assumption 3, there exist L_f, L_h, δ such that for any $k \in \mathbb{N}$, $\|\nabla h(x^{k+1}) - \nabla h(y^k)\| \le L_h\|x^{k+1} - y^k\|$, $\|\nabla f(x^{k+1}) - \nabla f(y^k)\| \le L_f\|x^{k+1} - y^k\|$, $\|u_k\| \le \delta$.
Hence

$$\begin{aligned} \|v_{k+1}\| \le &\left(L_f + \lambda_k^{-1}L_h\right)\|x^{k+1} - y^k\| \\ &+ M(\delta + L_h)\|x^{k+1} - x^k\| \end{aligned}$$

$$\leq \left(L_f + (\lambda_k^{-1} + M)L_h + M\delta\right)\|x^{k+1} - x^k\| + \left(L_f + \lambda_k^{-1}L_h\right)\|x^k - x^{k-1}\| \leq \left(L_f + (\lambda_k^{-1} + M)L_h + M\delta\right)\left(\|x^{k+1} - x^k\| + \|x^k - x^{k-1}\|\right),$$

And so, (H2) is satisfied.

(H3) Hypothesis (H3) follows naturally from Lemma 2(ii).

According to Theorem 1, combining the three hypotheses given in Theorem 1 and KL property at $\tilde{z}$ could guarantee that conclusion (i) holds. Conclusion (ii) is followed by Theorem 2(i). Thus $\{x^k\}_{k\in\mathbb{N}}$ is a Cauchy sequence of $\mathbb{R}^d$ and converges to its limit point $\tilde{x}$. From Theorem 1 $\tilde{x}$ is the critical point. □

V. NUMERICAL RESULTS

In this section we perform several numerical tests in order to show the behaviour and the convergence speed up obtained when using the BPGe algorithm. We consider two important optimization problems in which the differentiable part of the objective *does not admit* a global Lipschitz continuous gradient: a convex Poisson linear inverse problem and a nonconvex quadratic inverse problem (and so the PG and PGe algorithms cannot be applied to these problems). It is important to remark that for cases where the differentiable part of the objective admits a global Lipschitz continuous gradient the BPG and BPGe algorithms become the PG and PGe algorithms, respectively. That is, the BPG and BPGe methods can be applied but the performance in these cases it was already shown in [39].

The main parameters in BPGe algorithm are the stepsizes λ_k in Algorithm 1, and the parameter ρ that gives the extrapolation coefficients β_k in the line search method of Algorithm 2. In our tests we consider fixed stepsizes $\lambda_k = \lambda$. The influence of both parameters $\{\lambda,\ \rho\}$ in order to fix suitable values is studied below in the tests.

All the numerical experiments have been performed in Matlab 2013a on a PC Intel(R) Xeon(R) CPU E5-2697 (2.6 GHz).

A. APPLICATION TO POISSON LINEAR INVERSE PROBLEMS (PLIP)

Poisson Linear Inverse Problems (that is, linear inverse problems in presence of Poisson noise) emerged in many fields, like astronomy, nuclear medicine (e.g., Positron Emission Tomography), inverse problems in fluorescence microscopy [2], [5], [19]. Therefore, the design of methods and estimators for such problems has been studied intensively over the last two decades (for a review, see [5], [19]). Often these problems can be represented as a minimization problem like

$$\min\left\{d(b, Ax) + \theta g(x) : x \in \mathbb{R}^d_+\right\} \qquad \text{(PLIP)}$$

where $\theta > 0$ is used to weigh matching the data fidelity criteria and its regularizer g, and $d(\cdot,\cdot)$ denotes a convex proximity measure between two vectors.

A very well-known measure of proximity of two nonnegative vectors Ax and b is based on the Kullback-Liebler divergence:

$$d(b, Ax) := \sum_{i=1}^{m}\left\{b_i \log\frac{b_i}{(Ax)_i} + (Ax)_i - b_i\right\}.$$

which corresponds to noise of the negative Poisson loglikelihood function. It is easy to find that $f := d(b, Ax)$ has no globally Lipschitz continuous gradient [2], but satisfies L-smad condition with a kernel generating distance called Burg's entropy, denoted as

$$h(x) = -\sum_{j=1}^{d}\log x_j,\ \text{dom}\, h = \mathbb{R}^d_+,$$

and so now the Bregman distance is given by

$$D_h(x, y) = \sum_{j=1}^{d}\left\{\frac{x_j}{y_j} - \log\left(\frac{x_j}{y_j}\right) - 1\right\}.$$

Therefore, we have that

(i) (f, h) is L-smad, where $L \geq \|b\|_1$ (according to Lemma 7 in [2]), and f is 0-relative weakly convex to h since f is convex;

(ii) Assumptions 1 and 2 hold, but Assumption 3 does not hold.

So, from the convergence Section IV, we can solve this problem using the BPGe algorithm and it is guaranteed that any limit point of the sequence generated by BPGe is a stationary point of the objective function Ψ.

An important point in any iterative method is to define suitable error control techniques. As discussed in Section III, `EXIT` conditions of the experiments are set when iterations exceed 5000 times or $\|x^k - x^{k-1}\|/\max\{1, \|x^k\|\} \leq 10^{-6}$ (as in [39]).

In the tests, the entries of $A \in \mathbb{R}^{m\times d}_+$ and $x \in \mathbb{R}^d_+$ are generated following independent uniform distributions over the interval [0, 1]. We consider the case $g(x) \equiv 0$, i.e., we solve the inverse problem without regularization, so now the minimization problem is the standard Poisson type maximum likelihood estimation problem (modulo change of sign to pass to a minimization problem).

As these methods (BPG and BPGe) can be applied to both, *overdetermined* ($m > d$) and *underdetermined* ($m < d$) problems, we have performed numerical tests on both cases. First, we present the results obtained in the *overdetermined* case. As commented before, the main parameters in BPGe algorithm are the stepsize λ and the parameter ρ. In order to study briefly the most suitable set of parameters, we analyze the influence of both parameters $\{\lambda,\ \rho\}$ in Figure 1. In all the pictures we show the evolution of $\|\Psi(x_k) - \Psi(x^*)\|$ (being x^* the approximate solution obtained at termination of each

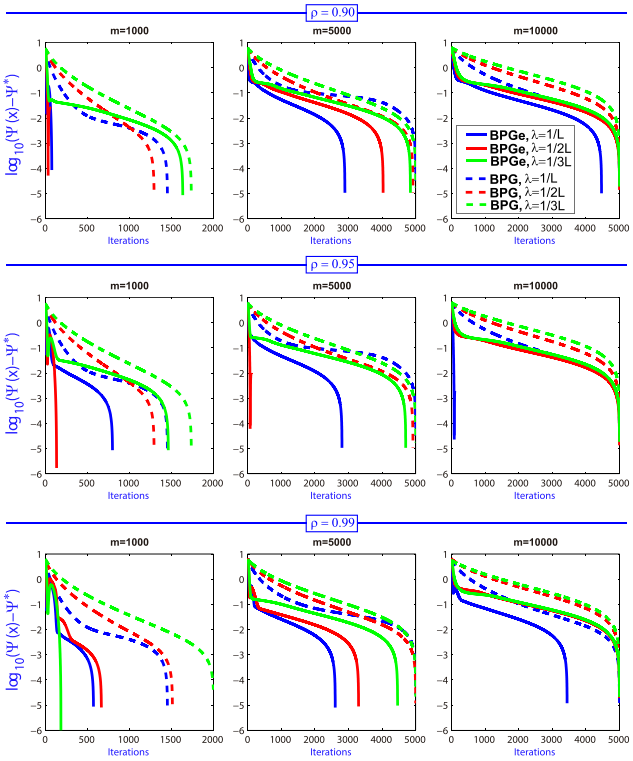

FIGURE 1. Poisson Linear Inverse Problems tests (*overdetermined* case $m > d$): evolution of the difference $\|\Psi(x_k) - \Psi(x^*)\|$ vs. iteration number, changing the parameters $\{\lambda, \rho\}$ and for several problem sizes (measurements m) with fixed vector dimension $d = 100$.

respective algorithm) with respect to the iteration number k. With this figure we can study the influence of the parameters with respect to the size of the problem (measurements m) with fixed dimension $d = 100$. Globally, we observe that the value $\rho = 0.99$ has the best results, even if for some cases, the set of initial conditions gives rise to a very fast convergence (as in the cases of using $\lambda = 1/(2L)$ for $m = 5000$ and $\rho = 0.95$, where we have a fast linear convergence instead of sublinear). Note that this kind of differences can be observed on other situations, but the average behaviour tells us that the best performance occurs when we take $\rho = 0.99$. On the other hand, similar comments can be said with respect to the stepsize parameter λ. The general situation also recommends us to take the highest value $\lambda = 1/L$ (also for both algorithms BPGe and BPG).

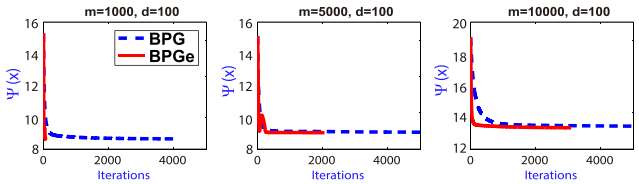

FIGURE 2. Poisson Linear Inverse Problems tests (*overdetermined* case $m > d$): evolution of the objective function $\Psi(x_k)$ vs. iteration number, using the parameter values $\{\lambda = 1/L, \rho = 0.99\}$ and for several problem sizes (measurements m) with fixed vector dimension $d = 100$.

In Figure 2, now with the fixed parameter values $\{\lambda = 1/L, \rho = 0.99\}$ and for the *overdetermined* $(m > d)$ case,

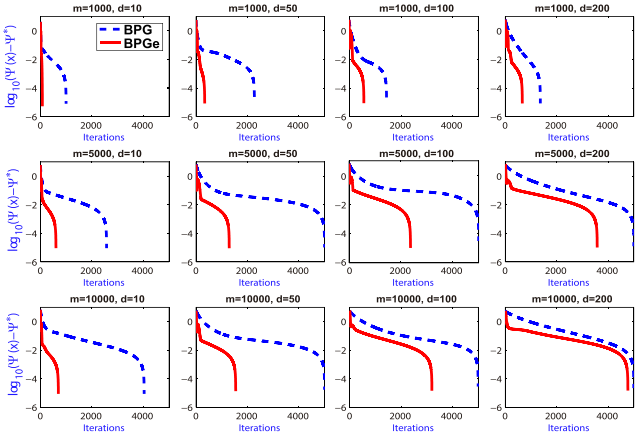

FIGURE 3. Poisson Linear Inverse Problems tests (*overdetermined* case $m > d$): evolution of the difference $\|\Psi(x_k) - \Psi(x^*)\|$ vs. iteration number, using the parameter values $\{\lambda = 1/L, \rho = 0.99\}$ and for several problem sizes (measurements m and vector dimensions d).

we show the evolution of the objective function $\Psi(x_k)$ vs. iteration number and for several problem sizes (measurements m) with fixed vector dimension $d = 100$. We observe that always the BPGe algorithm is much faster than the BPG one. In order to observe more clearly the faster convergence, we present in Figure 3 much more simulations but now showing the evolution of $\|\Psi(x_k) - \Psi(x^*)\|$. We note that the differences of both methods are bigger for low dimension d problems, in fact for the most overdetermined problems $m \gg d$.

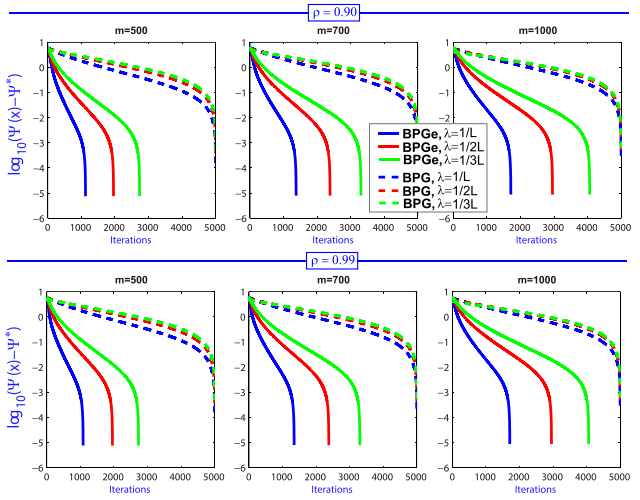

FIGURE 4. Poisson Linear Inverse Problems tests (*underdetermined* case $m < d$): evolution of the difference $\|\Psi(x_k) - \Psi(x^*)\|$ vs. iteration number, changing the parameters $\{\lambda, \rho\}$ and for several problem sizes (measurements m) with fixed vector dimension $d = 5000$.

In the *underdetermined* case we also analyze the influence of both parameters $\{\lambda, \rho\}$ in Figure 4 with respect to the size of the problem (measurements m) with fixed dimension $d = 5000$. Now, we observe that the value of the parameter ρ seems to not affect too much on the global performance of the method, so we will take the value $\rho = 0.99$ when we fix

TABLE 1. Poisson Linear Inverse Problems tests: CPU-time and number of iterations for different cases of m (number of data) and d (dimension) for two different values of the λ parameter for *overdetermined* (top) and *underdetermined* (bottom) cases. T_{BPGe} and T_{BPG} denote the CPU-time of BPGe and BPG algorithms, and N_{BPGe} and N_{BPG} the number of iterations to reach the EXIT criteria. Superscript –a– points out discordant cases related with a fast linear convergence.

Overdetermined case

		$\lambda = 1/L$				$\lambda = 1/(3L)$			
m	d	T_{BPGe}	$\frac{T_{BPGe}}{T_{BPG}}$	N_{BPGe}	$\frac{NI_{BPGe}}{NI_{BPG}}$	T_{BPGe}	$\frac{T_{BPGe}}{T_{BPG}}$	N_{BPGe}	$\frac{NI_{BPGe}}{NI_{BPG}}$
1000	10	0.08	0.07	74	0.07	0.21	0.22	279	0.21
	50	0.40	0.15	336	0.15	0.14	0.16	155	0.15^a
	100	1.13	0.41	574	0.40	0.32	0.10	187	0.09^a
	200	1.68	0.63	665	0.49	0.44	0.07	226	0.07
5000	10	0.77	0.24	605	0.23	0.83	0.22	745	0.21
	50	3.32	0.26	1291	0.26	4.16	0.34	1353	0.32
	100	7.50	0.53	2602	0.52	13.97	0.96	4460	0.89
	200	13.43	0.72	3577	0.72	20.26	1.12	5000	1.00
10000	10	2.53	0.18	699	0.17	0.50	0.03	141	0.03^a
	50	6.68	0.33	1543	0.31	15.36	0.68	3255	0.65
	100	16.75	0.70	3441	0.69	23.90	1.02	5000	1.00
	200	30.32	0.99	4770	0.95	30.20	1.05	5000	1.00

Underdetermined case

		$\lambda = 1/L$				$\lambda = 1/(3L)$			
m	d	T_{BPGe}	$\frac{T_{BPGe}}{T_{BPG}}$	N_{BPGe}	$\frac{NI_{BPGe}}{NI_{BPG}}$	T_{BPGe}	$\frac{T_{BPGe}}{T_{BPG}}$	N_{BPGe}	$\frac{NI_{BPGe}}{NI_{BPG}}$
100	1000	0.60	0.15	369	0.14	2.19	0.25	1314	0.26
200		5.03	0.89	1754	0.67	3.56	0.29	1298	0.26
300		4.50	0.78	1760	0.67	2.81	0.26	1315	0.26
500	5000	9.17	0.23	1085	0.22	70.00	1.49	5000	1.00
700		12.85	0.28	1378	0.28	115.16	1.34	5000	1.00
1000		27.51	0.32	1565	0.31	345.13	1.18	5000	1.00
1000	10000	210.06	0.66	3284	0.66	549.52	1.03	5000	1.00
2000		643.71	0.89	4271	0.85	886.94	1.07	5000	1.00
3000		967.90	1.02	5000	1.00	1084.82	1.04	5000	1.00

the parameter. On the other hand, similar comments as in the *overdetermined* case can be said with respect to the stepsize parameter λ. Now the behaviour is quite regular, and no cases of very fast convergence have been observed, and the fastest convergence is obtained for the highest value $\lambda = 1/L$ (also for both algorithms BPGe and BPG). Therefore, in the rest of tests on this paper we fix the parameter values $\{\lambda = 1/L, \rho = 0.99\}$.

In Figure 5, now with the fixed parameter values $\{\lambda = 1/L,\ \rho = 0.99\}$ and for the *underdetermined* ($m < d$) case, we observe that always the BPGe algorithm is much faster than the BPG one. But, similarly as in the *overdetermined* case, the differences are bigger when we use the methods for larger ratios d/m, that is, for the most underdetermined problems $m \ll d$.

Finally, in Table 1 we give the CPU-time and number of iterations for different sizes of problems (number of data m and dimension d) for two values of the λ parameter ($\lambda = 1/L$ and $1/(3L)$) for *overdetermined* (top) and *underdetermined* (bottom) cases. From the simulations we observe that when the problem has not a very big size (probably because in these other cases longer simulations are needed) the ratios among both methods provide an interesting speed-up, and in most cases the EXIT strategy stops the BPGe algorithm before the maximum number of iterations is reached. On the other hand, we observe that the CPU-time and iteration number ratios are

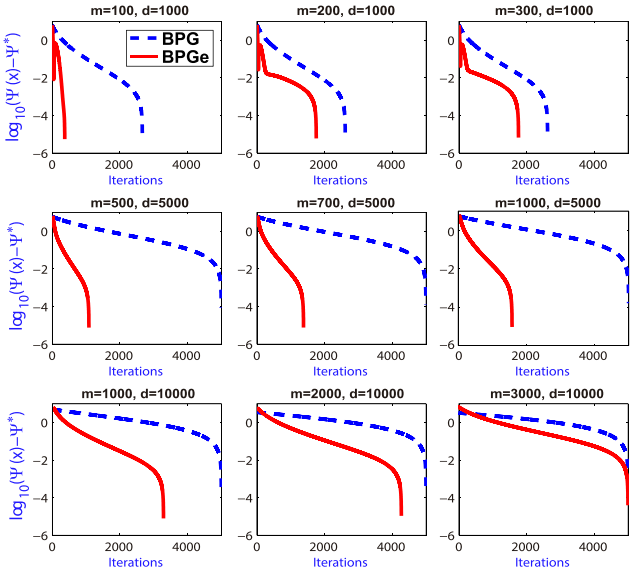

FIGURE 5. Poisson Linear Inverse Problems tests (*underdetermined* case $m < d$): evolution of the difference $\|\Psi(x_k) - \Psi(x^*)\|$ vs. iteration number, using the parameter values $\{\lambda = 1/L,\ \rho = 0.99\}$ and for several problem sizes (measurements m and vector dimensions d).

quite similar, and so there are little differences between them. Note that the BPGe algorithm has an extra step, the line search method of Algorithm 2, but it increments quite a few the final

CPU-time. On the table we have remarked three discordant cases (superscript –a–) related with a fast linear convergence, instead of sublinear. This is illustrated, for example, on the left bottom plot of Figure 1 ($\rho = 0.99, m = 1000$) where the green curve, corresponding to $\lambda = 1/(3L)$ converges faster than the other colours (as it also occurs in other plots of the same figure). Note that for an *overdetermined* problem with random data some initial conditions and data may be led to a faster convergence. For the *underdetermined* problem there is a regular behaviour in all the simulations.

Therefore, in the Poisson Linear Inverse Problems tests the BPGe algorithm presents a faster performance compared with the BPG algorithm, giving an interesting option for real problems.

B. APPLICATION TO QUADRATIC INVERSE PROBLEMS

In the second test (taken from [8]) we show that BPGe algorithm can deal with a nonconvex Quadratic Inverse Problem (QIP) in which the differentiable term has no globally gradient Lipschitz continuous property. This problem is a natural extension of the linear inverse problem, but now using quadratic measurements. It appears in many popular applications, such as signal recovery [3] and phase retrieve [25] from the knowledge of the amplitude of complex signals.

A general description of the Quadratic Inverse Problem is to find the vector $x \in \mathbb{R}^d$ that solves the system

$$x^T A_i x \simeq b_i, \quad i = 1, \ldots, m$$

being $\{A_i \in \mathbb{R}^{d\times d} \mid i = 1, \ldots, m\}$ a set of symmetric matrices that describes the model, and $b = (b_1, \ldots, b_m) \in \mathbb{R}^m$ a vector of usually noisy measurements.

Following the formalism given in [8, section 5.1], this problem can be formulated as a nonconvex minimization problem as:

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

where $\theta > 0$ is used to weigh matching the data fidelity criteria and its regularizer g. In our experiments, we take a convex l_1-norm regularization function $g(x) = \|x\|_1$. Note that the first function $f(x)$ is a nonconvex differentiable function but that does not admit a global Lipschitz continuous gradient.

The main quality of the BPG and BPGe algorithms (as noted to the BPG in [8]) is that these methods can solve the broad class of problems (QIP). To apply BPG and BPGe on the QIP model properly, we first need to identify a suitable function h (Definition 1). In [8], a proper choice has been given as:

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

and so now the Bregman distance is given by

$$D_h(x, y) = \{h(x) - h(y) - (\|y\|^2 y + y)^T (x - y)\}.$$

When L is chosen such that $L \geq \sum_{i=1}^{m} \left(3\|A_i\|^2 + \|A_i\||b_i|\right)$ then by [8, Lemma 5.1], *L-smad* condition (Definition 2) holds for the selected functions $f(x)$, $g(x)$ and $h(x)$. Besides, according to the same analysis in [8, Lemma 5.1], we could derive the relative weakly convex parameter as $\mu \geq \sum_{i=1}^{m} \|A_i\||b_i|$. In conclusion, we have that:

(i) (f, h) is *L-smad*, f is μ-relative weakly convex to h.
(ii) Assumptions 1 and 2 are easily verified.
(iii) f, g, D_h are all semi-algebraic, (see for example [7]). One can show inductively that $H_M(x, y) = \Psi(x) + MD_h(x, y)$ is semi-algebraic, thus it has KL property (Definition 4) at any point (x, x). Besides, we could verify that Assumption 3 holds.

It means, from the convergence Section IV, that the sequences generated by BPGe algorithm converge to a critical point of the objective function Ψ.

Here, we perform several numerical tests to compare the behaviour of the BPGe and BPG algorithms. As we did with the previous problem (PLIP), we have designed two main families of experiments, considering *overdetermined* ($m > d$) and *underdetermined* ($m < d$) cases. To that goal we set different values of m and d, and we generate m random rank-1 matrices $A_i = a_i a_i^T$ in $\mathbb{R}^{d\times d}$, where the entries of the vectors a_i are generated following independent Gaussian distributions with zero mean and unit variance. The accurate $x^* := \arg\min\{\Psi(x) : x \in \mathbb{R}^d\}$ is chosen as a sparse vector (the sparsity is 5%) and $b_i = x^T A_i x^*$, $i = 1, \ldots, m$. We set the weight parameter $\theta = 1$ as default.

As a first performance comparison, in Table 2 we give the CPU-time and number of iterations for different sizes of problems (number of data m and dimension d) for two values of the λ parameter ($\lambda = 1/L$ and $1/(3L)$) for the *overdetermined* case. The values T_{BPGe} and T_{BPG} denote the CPU-time of BPGe and BPG algorithms, and N_{BPGe} and N_{BPG} the number of iterations to reach the `EXIT` criteria, respectively. From the simulations we observe that the ratios among both methods provide an interesting speed-up, and the `EXIT` strategy stops the BPGe algorithm before the maximum number of iterations ($k_{max} = 5000$ in this case) is reached. On the other hand, we observe that the CPU-time and iteration number ratios are quite similar, and so there are little differences between them. Therefore, we note again that although the BPGe algorithm has an extra step (the line search method of Algorithm 2), it increments quite a few the final CPU-time. Also, from the data we observe that although the ratio for the BPGe and BPG algorithms for $\lambda = 1/(3L)$ is quite good, the option BPGe with $\lambda = 1/L$ performs many fewer iterations, and so it is the recommended option.

In Figure 6, with the fixed parameter values $\{\lambda = 1/L, \rho = 0.99\}$ and for the *overdetermined* ($m > d$) and *underdetermined* ($m < d$) cases, we show the evolution of $\|\Psi(x_k) - \Psi(x^*)\|$. In this problem we observe that the performance of the accelerated BPGe algorithm for the *overdetermined* case is quite good, giving a linear convergence. In the *underdetermined* case the behaviour seems to

TABLE 2. Quadratic Inverse Problems tests: CPU-time and number of iterations for different cases of m (number of data) and d (dimension) for two different values of the λ parameter for the *overdetermined* case. T_{BPGe} and T_{BPG} denote the CPU-time of BPGe and BPG algorithms, and N_{BPGe} and N_{BPG} the number of iterations to reach the EXIT criteria.

		$\lambda = (1/L)$				$\lambda = 1/(3L)$			
m	d	T_{BPGe}	$\frac{T_{BPGe}}{T_{BPG}}$	N_{BPGe}	$\frac{NI_{BPGe}}{NI_{BPG}}$	T_{BPGe}	$\frac{T_{BPGe}}{T_{BPG}}$	N_{BPGe}	$\frac{NI_{BPGe}}{NI_{BPG}}$
10000	10	0.29	0.53	146	0.35	0.48	0.28	248	0.20
	50	0.57	0.14	271	0.14	4.41	0.10	480	0.10
	100	1.16	0.10	339	0.08	8.73	0.19	655	0.13
	200	10.15	0.15	608	0.12	17.24	0.31	1668	0.33
20000	10	0.24	0.34	143	0.34	0.39	0.26	304	0.26
	50	4.09	0.14	266	0.14	6.80	0.16	465	0.09
	100	1.79	0.09	323	0.09	9.39	0.16	605	0.12
	200	66.97	0.18	602	0.12	40.74	0.28	1413	0.28
30000	10	0.40	0.44	145	0.35	3.22	0.27	231	0.20
	50	1.48	0.15	261	0.15	10.42	0.10	472	0.10
	100	32.79	0.09	331	0.09	15.06	0.12	594	0.12
	200	153.17	0.12	554	0.11	487.62	0.27	1341	0.27

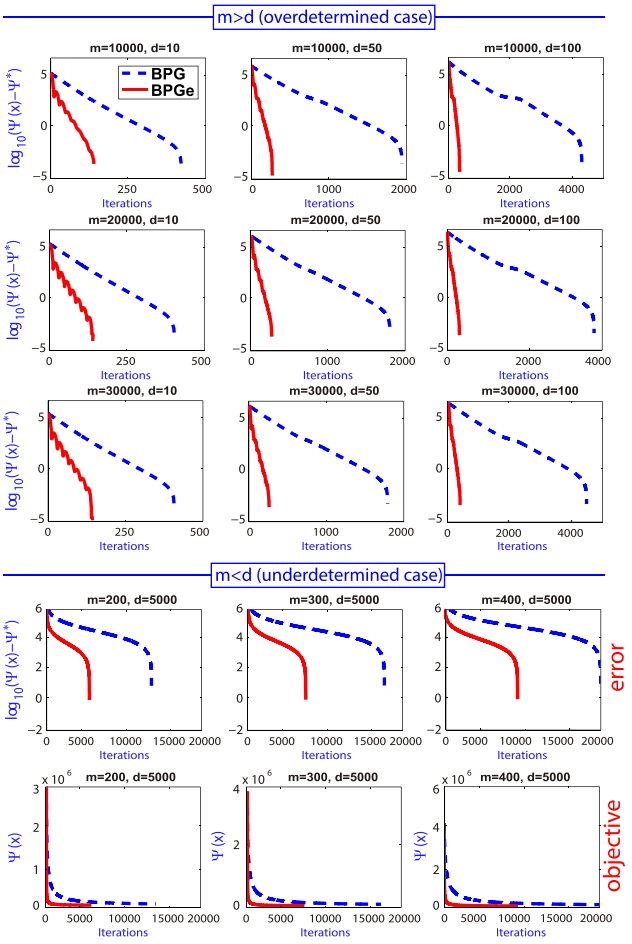

FIGURE 6. Quadratic Inverse Problems tests (*overdetermined* case $m > d$) and (*underdetermined* case $m < d$): evolution of the difference $\|\Psi(x_k) - \Psi(x^*)\|$ vs. iteration number, using the parameter values $\{\lambda = 1/L,\ \rho = 0.99\}$ and for several problem sizes (measurements m and vector dimensions d) and evolution of the objective function $\Psi(x_k)$.

be sublinear, and it needs more iterations to reach the desired value (in this simulations $k_{max} = 20000$). In both cases the BPGe algorithms performs much better than the BPG one.

For the *underdetermined* case we also show the evolution of the objective function $\Psi(x_k)$ vs. iteration number to see that in this case the objective function takes large values, and therefore, when applying the EXIT strategy the required precision is obtained (a relative error $< 10^{-6}$) giving not too small absolute values.

Therefore, again in the Quadratic Inverse Problems tests the BPGe algorithm presents a faster performance compared with the BPG algorithm, giving an interesting option for real problems.

VI. CONCLUSION

We have introduced a new accelerated Bregman proximal gradient algorithm (BPGe) useful for nonconvex and nonsmooth minimization problems. This algorithm combines two powerful methods to solve large-scale minimization problems. On one hand, we have taken the BPG algorithm [2] able to deal with non-globally Lipschitz continuous gradient problems (firstly defined for the convex case [2] and later extended to the nonconvex case by [8]). And on the other hand, the accelerated extrapolation algorithm (used for instance in the PG algorithm [39]). The use of the Bregman distance paradigm permits to enlarge the number of problems to work with, because we do not need the assumption of global Lipschitz gradient continuity. And with the extrapolation technique the convergence of the method is accelerated.

The convergence of the new method is studied, and we have proven that any limit point of the sequence generated by BPGe algorithm is a stationary point of the problem by choosing parameters properly. Besides, assuming Kurdyka-Łojasiewicz property, we have proven the whole sequences generated by BPGe converges to a stationary point.

Finally, we have applied it to two important practical problems that arise in many fundamental applications (and that not satisfy global Lipschitz gradient continuity assumption): Poisson linear inverse problems and quadratic inverse problems, for both, *overdetermined* and *underdetermined* cases.

In these tests the BPGe algorithm has shown faster convergence results than the BPG algorithm, and so the new BPGe algorithm seems to be an interesting methodology.

REFERENCES

[1] H. Attouch, J. Bolte, and B. F. Svaiter, "Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods," *Math. Program.*, vol. 137, nos. 1–2, pp. 91–129, 2013.

[2] H. H. Bauschke, J. Bolte, and M. Teboulle, "A descent lemma beyond Lipschitz gradient continuity: First-order methods revisited and applications," *Math. Oper. Res.*, vol. 42, no. 2, pp. 330–348, 2017.

[3] A. Beck and Y. C. Eldar, "Sparsity constrained nonlinear optimization: Optimality conditions and algorithms," *SIAM J. Optim.*, vol. 23, no. 3, pp. 1480–1509, 2013.

[4] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[5] M. Bertero, P. Boccacci, G. Desiderà, and G. Vicidomini, "Image deblurring with Poisson data: From cells to galaxies," *Inverse Problems*, vol. 25, no. 12, pp. 123006-1–123006-27, 2009.

[6] D. Bertsekas, *Convex Optimization Theory*. Belmont, MA, USA: Athena Scientific, 2009.

[7] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Math. Program.*, vol. 146, nos. 1–2, pp. 459–494, 2014.

[8] 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 J. Optim.*, vol. 28, no. 3, pp. 2131–2151, 2018.

[9] J. Borwein and A. S. Lewis, *Convex Analysis and Nonlinear Optimization: Theory and Examples*. Springer, 2010.

[10] 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," *USSR Comput. Math. Math. Phys.*, vol. 7, no. 3, pp. 200–217, 1967.

[11] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford, "Accelerated methods for nonconvex optimization," *SIAM J. Optim.*, vol. 28, no. 2, pp. 1751–1772, 2018.

[12] Y. Censor and A. Lent, "An iterative row-action method for interval convex programming," *J. Optim. Theory Appl.*, vol. 34, no. 3, pp. 321–353, 1981.

[13] Y. Censor and S. A. Zenios, "Proximal minimization algorithm with D-functions," *J. Optim. Theory Appl.*, vol. 73, no. 3, pp. 451–464, 1992.

[14] A. I. Chen and A. Ozdaglar, "A fast distributed proximal-gradient method," in *Proc. 50th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Oct. 2012, pp. 601–608.

[15] G. Chen and M. Teboulle, "Convergence analysis of a proximal-like minimization algorithm using Bregman functions," *SIAM J. Optim.*, vol. 3, no. 3, pp. 538–543, 1993.

[16] D. Davis, D. Drusvyatskiy, and K. J. MacPhee, "Stochastic model-based minimization under high-order growth," 2018, *arXiv:1807.00255*. [Online]. Available: https://arxiv.org/abs/1807.00255

[17] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[18] S. Ghadimi and G. Lan, "Accelerated gradient methods for nonconvex nonlinear and stochastic programming," *Math. Program.*, vol. 156, no. 1, pp. 59–99, 2016.

[19] T. Hohage and F. Werner, "Inverse problems with Poisson data: Statistical regularization theory, applications and algorithms," *Inverse Problems*, vol. 32, no. 9, p. 093001, 2016.

[20] K. Jiang, D. Sun, and K.-C. Toh, "An inexact accelerated proximal gradient method for large scale linearly constrained convex SDP," *SIAM J. Optim.*, vol. 22, no. 3, pp. 1042–1064, 2012.

[21] K. C. Kiwiel, "Proximal minimization methods with generalized Bregman functions," *SIAM J. Control Optim.*, vol. 35, no. 4, pp. 1142–1168, 1997.

[22] G. Li and T. K. Pong, "Calculus of the exponent of Kurdyka–Łojasiewicz inequality and its applications to linear convergence of first-order methods," *Found. Comput. Math.*, vol. 18, no. 5, pp. 1199–1232, 2017.

[23] H. Li and Z. Lin, "Accelerated proximal gradient methods for nonconvex programming," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 379–387.

[24] H. Lu, R. M. Freund, and Y. Nesterov, "Relatively smooth convex optimization by first-order methods, and applications," *SIAM J. Optim.*, vol. 28, no. 1, pp. 333–354, 2018.

[25] D. R. Luke, "Phase retrieval, what's new," *SIAG/OPT Views News*, vol. 25, no. 1, pp. 1–5, 2017.

[26] Y. Nesterov, "A method of solving a convex programming problem with convergence rate $O(1/k^2)$," *Soviet Math. Doklady*, vol. 27, no. 2, pp. 372–376, 1983.

[27] Y. Nesterov, "Dual extrapolation and its applications to solving variational inequalities and related problems," *Math. Program.*, vol. 109, no. 2, pp. 319–344, 2007.

[28] A. Nitanda, "Stochastic proximal gradient descent with acceleration techniques," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1574–1582.

[29] E. Nurminskii, "The quasigradient method for the solving of the nonlinear programming problems," *Cybern. Syst. Anal.*, vol. 9, no. 1, pp. 145–150, 1973.

[30] P. Ochs, Y. Chen, T. Brox, and T. Pock, "iPiano: Inertial proximal algorithm for nonconvex optimization," *SIAM J. Imag. Sci.*, vol. 7, no. 2, pp. 1388–1419, 2014.

[31] N. Parikh and S. Boyd, "Proximal algorithms," *Found. Trends Optim.*, vol. 1, no. 3, pp. 127–239, Jan. 2014.

[32] T. Pock and S. Sabach, "Inertial proximal alternating linearized minimization (iPALM) for nonconvex and nonsmooth problems," *SIAM J. Imag. Sci.*, vol. 9, no. 4, pp. 1756–1787, 2016.

[33] R. T. Rockafellar, *Convex Analysis*. Princeton, NJ, USA: Princeton Univ. Press,, 2015.

[34] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, vol. 317. Springer, 2009.

[35] M. Schmidt, N. L. Roux, and F. R. Bach, "Convergence rates of inexact proximal-gradient methods for convex optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 1458–1466.

[36] M. Teboulle, "A simplified view of first order methods for optimization," *Math. Program.*, vol. 170, no. 1, pp. 67–98, 2018.

[37] K.-C. Toh and S. Yun, "An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems," *Pacific J. Optim.*, vol. 6, no. 3, pp. 615–640, Nov. 2010.

[38] N. D. Vanli, M. Gurbuzbalaban, and A. Ozdaglar, "Global convergence rate of proximal incremental aggregated gradient methods," *SIAM J. Optim.*, vol. 28, no. 2, pp. 1282–1300, 2018.

[39] B. Wen, X. Chen, and T. K. Pong, "Linear convergence of proximal gradient algorithm with extrapolation for a class of nonconvex nonsmooth minimization problems," *SIAM J. Optim.*, vol. 27, no. 1, pp. 124–145, 2017.

[40] L. Xiao and T. Zhang, "A proximal stochastic gradient method with progressive variance reduction," *SIAM J. Optim.*, vol. 24, no. 4, pp. 2057–2075, 2014.

XIAOYA ZHANG received the M.S. degree in mathematics from the National University of Defense Technology, China, in 2013, where she is currently pursuing the Ph.D. degree. Her research interests include optimization and control, image science, and machine learning.

ROBERTO BARRIO received the B.S. and Ph.D. degrees in applied mathematics from the University of Zaragoza, Spain, in 1992 and 1997, respectively, where he is currently a Full Professor with the Department of Applied Mathematics. His current research interests include optimization techniques, numerical analysis, mathematical neuroscience, dynamical systems, and computational dynamics. He has authored or coauthored over 100 technical articles in different fields of applied mathematics in renowned international journals and conferences. He is also an Editor of the *Communications in Nonlinear Science and Numerical Simulation*, *Applied Mathematics and Computation*, *Frontiers in Applied Mathematics and Statistics*, and *Abstract and Applied Analysis*.

M. ANGELES MARTÍNEZ received the B.S. degree in physics from the University of Zaragoza, Spain, in 2002, and the Ph.D. degree from the Doctoral Program of Physics and Mathematics, University of Granada, in 2008. Since 2016, she has been a Lecturer with the Department of Applied Mathematics, Engineering School, University of Zaragoza. She has been engaged in different projects on applied physics and applied mathematics. Her current research interests include dynamical systems, biomathematics, and computational dynamics.

HAO JIANG was born in 1983. He received the Ph.D. degree from the National University of Defense Technology, Changsha, China, where he is currently an Assistant Researcher. His research interests include high-performance computing, rounding error analysis, and numerical computation.

LIZHI CHENG received the Ph.D. degree from the National University of Defense Technology, Changsha, China, in 2002, where he is currently a Professor with the Department of Mathematics. He has authored or coauthored over 100 technical articles in different fields, including optimization and control, image processing, and wavelet analysis. His research interests include the mathematical foundation of signal analysis and wavelet analysis with applications to image compression.