A Unified Approach to Proximal Algorithms using Bregman Distance

Yi Zhou[a,*], Yingbin Liang[a], Lixin Shen[b]

[a]*Department of Electrical Engineering and Computer Science, Syracuse University*
[b]*Department of Mathematics, Syracuse University*

Abstract

We show that proximal gradient algorithm (PGA) with Bregman distance for minimizing the sum of two convex functions can be viewed as proximal point algorithm (PPA) incorporating with another Bregman distance. Consequently, the convergence result of the PGA follows directly from that of PPA, and this leads to a simpler convergence analysis with a tighter convergence rate than existing ones. We further propose and analyze the backtracking line search variant of PGA with Bregman distance.

Keywords: proximal algorithms, Bregman distance, convergence analysis, line search.

1. Introduction

Proximal algorithms have been extensively studied in convex optimization to solve non-smooth problems. These algorithms have been widely applied to solve practical problems including image processing, e.g., [1, 2], distributed statistical learning, e.g., [3], and low rank matrix minimization, e.g., [4].

Consider the following optimization problem:

$$\textbf{(P1)} \quad \min_{\mathbf{x}\in C} r(\mathbf{x}), \tag{1}$$

where $r : \mathbb{R}^n \to (-\infty, +\infty]$ is a proper, lower-semicontinuous convex function, and C is a closed convex set in $\mathbb{R}^n$. The well known proximal point algorithm (PPA) for solving **(P1)** was introduced initially by Martinet in [5]. The algorithm generates a sequence $\{\mathbf{x}_k\}$ via the following iterative step:

$$\text{(PPA-}\mathcal{E}\text{)} \quad \mathbf{x}_{k+1} = \operatorname*{argmin}_{\mathbf{x}\in C} \left\{ r(\mathbf{x}) + \frac{1}{2\lambda_k}\|\mathbf{x}-\mathbf{x}_k\|_2^2 \right\}, \tag{2}$$

where $\lambda_k > 0$ corresponds to the step size at k-th iteration, and we use $\mathcal{E}$ to stand for the choice of the Euclidean distance (i.e., the l_2 term). This algorithm can be interpreted as applying the gradient descent method on the Moreau envelope of r, i.e., a smoothed version of the objective function [6, 7]. With a proper choice of the step size sequence $\{\lambda_k\}$, it was shown in [8, 9] that the sequence $\{\mathbf{x}_k\}$ generated by PPA-$\mathcal{E}$ converges to a solution of **(P1)**, and the rate of convergence was characterized in [10].

A natural generalization of PPA-$\mathcal{E}$ is to replace the Euclidean distance with a more general distance-like term. In existing literature, various choices of distance have been proposed, e.g., [11, 12, 13, 14]. Among them, a popular choice is the Bregman distance [15, 16, 14]. As a consequence, we obtain the following iterative step:

$$\text{(PPA-}\mathcal{B}\text{)} \quad \mathbf{x}_{k+1} = \operatorname*{argmin}_{\mathbf{x}\in C} \left\{ r(\mathbf{x}) + \frac{1}{\lambda_k} D_h(\mathbf{x}, \mathbf{x}_k) \right\}. \tag{3}$$

Here, $D_h(\mathbf{x}, \mathbf{x}_k)$ corresponds to the Bregman distance between the points $\mathbf{x}$ and $\mathbf{x}_k$, and is based on a coninuously differentiable strictly convex function h. We refer to Definition 1 for the detailed definition. We refer to the above method as PPA-$\mathcal{B}$ for using the Bregman distance. The convergence rate of PPA-$\mathcal{B}$ has been extensively discussed in [15, 13], and we refer to [17] for a comprehensive discussion on PPA with different choices of distance metrics.

A generalized model of **(P1)** is the following composite minimization framework:

$$\textbf{(P2)} \quad \min_{\mathbf{x}\in C} \{F(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})\}, \tag{4}$$

where f is usually a smooth and convex loss function that corresponds to the data fitting part, and g is a possibly non-smooth regularizer that promotes structures (such as sparsity) to the solution of the problem. This composite minimization framework generalizes many applications in machine learning, image processing, detection, etc. As an extension of PPA, splitting algorithms are proposed for solving the composite minimization in **(P2)** [9, 18]. In particular, the proximal gradient algorithm (PGA) under Euclidean distance has been developed in [19] to solve **(P2)** efficiently, and the iterative step is given by

$$\text{(PGA-}\mathcal{E}\text{)}: \quad \mathbf{x}_{k+1} = \operatorname*{argmin}_{\mathbf{x}\in C} \left\{ g(\mathbf{x}) + \langle \mathbf{x}, \nabla f(\mathbf{x}_k)\rangle + \frac{1}{2\gamma_k}\|\mathbf{x}-\mathbf{x}_k\|_2^2 \right\}, \tag{5}$$

where PGA-$\mathcal{E}$ stands for the special choice of Euclidean distance (i.e., the l_2^2 term). It has been shown that PGA-$\mathcal{E}$ has a convergence rate of $\mathcal{O}(1/k)$[1] [1], and the rate can be further improved to be $\mathcal{O}(1/k^2)$ via Nesterov's acceleration technique. Inspired by the way of generalizing PPA-$\mathcal{E}$ to PPA-$\mathcal{B}$, PGA-$\mathcal{E}$ can also be generalized by replacing the Euclidean distance

*Corresponding author
Email address: yzhou35@syr.edu (Yi Zhou)

[1]Here, $f(n) = \mathcal{O}(g(n))$ denotes that $|f(n)| \le \xi|g(n)|$ for all $n > N$, where ξ is a constant and N is a positive integer.

with the Bregman distance $D_h(\cdot,\cdot)$, and correspondingly, the iterative step is given by

(PGA-$\mathcal{B}$) :

$$\mathbf{x}_{k+1} = \operatorname*{argmin}_{\mathbf{x}\in C} \left\{ g(\mathbf{x}) + \langle \mathbf{x}, \nabla f(\mathbf{x}_k)\rangle + \frac{1}{\gamma_k} D_h(\mathbf{x}, \mathbf{x}_k) \right\}. \quad (6)$$

Such formulation naturally extends the mirror descent (MD) method when $g \equiv 0$. Under a Lipschitz condition of ∇f, it has been shown in [20] that the convergence rate of PGA-$\mathcal{B}$ is $\mathcal{O}(1/k)$.

It is clear that PGA-$\mathcal{E}$ and PGA-$\mathcal{B}$ are respectively generalizations of PPA-$\mathcal{E}$ and PPA-$\mathcal{B}$, because they coincide when the function f in **(P2)** is a constant function. Thus, in existing literature, the analysis of PGA is developped by their own as in [1, 20] without resorting to existing analysis of PPA. More recently, a concurrent work [21] to this paper interprets PGA-$\mathcal{B}$ as the composition of mirror descent method and PPA-$\mathcal{B}$. In contrast to this viewpoint, this paper shows that PGA-$\mathcal{B}$ can, in fact, be viewed as PPA-$\mathcal{B}$ under a proper choice of the Bregman distance. Consequently, the analysis of PGA-$\mathcal{B}$ can be mapped to that of PPA-$\mathcal{B}$ in a unified way. We note that the initial version of this paper [22] was posted on arXiv in March, 2015, which already independently developped the aforementioned main result.

We summarize our main contributions as follows. In this paper, we point out that PGA-$\mathcal{B}$ can be viewed as PPA-$\mathcal{B}$ with a special choice of Bregman distance, and thus inherits all the existing convergence results of PPA-$\mathcal{B}$. In particular, we obtain a tighter estimate of the convergence rate of the function value residual, and our result avoids involving the symmetry coefficient in [21]. Lastly, we propose a line search variant of PGA-$\mathcal{B}$ and characterize its convergence rate.

The rest of the paper is organized as follows. In §2, we recall the definition of Bregman distance, unify PGA-$\mathcal{B}$ as a special case of PPA-$\mathcal{B}$ and discuss its convergence results. In §3, we propose a line search variant of PGA-$\mathcal{B}$ and characterize its convergence rate. Finally in §4, we conclude our paper with a few remarks on our results.

2. Unify PGA-$\mathcal{B}$ as PPA-$\mathcal{B}$

2.1. Bregman Distance

We first recall the definition of the Bregman distance [23], see also [24, 15]. Throughout, the interior of a set $C \subset \mathbb{R}^n$ is denoted as $\mathrm{int}C$

Definition 1 (Bregman Distance $\mathcal{B}$)**.** *Let $h : \mathbb{R}^n \to (-\infty, +\infty]$ be a function with $\mathrm{dom}\, h = C$, and satisfies:*

(a) h is continuously differentiable on $\mathrm{int}C$;

(b) h is strictly convex on C.

Then the Bregman distance $D_h : C \times \mathrm{int}C \to \mathbb{R}_+$ associated with function h is defined as

$$D_h(\mathbf{x}, \mathbf{y}) = h(\mathbf{x}) - h(\mathbf{y}) - \langle \mathbf{x} - \mathbf{y}, \nabla h(\mathbf{y})\rangle, \quad (7)$$

for all $\mathbf{x} \in C$ and $\mathbf{y} \in \mathrm{int}C$.

We denote $\mathcal{B}$ as the class of all Bregman distances. Clearly, the Bregman distance D_h is defined as the residual of the first order expansion of function h. Thus, it is in general asymmetric with respect to the two arguments. On the other hand, the convexity of h implies the nonnegativity of D_h, making it behaves like a distance metric. Moreover, the following properties are direct consequences of eq. (7): For any $\mathbf{u} \in C, \mathbf{x}, \mathbf{y} \in \mathrm{int}C$ and any $D_h, D_{h'} \in \mathcal{B}$,

$$D_h(\mathbf{u}, \mathbf{x}) + D_h(\mathbf{x}, \mathbf{y}) - D_h(\mathbf{u}, \mathbf{y}) = \langle \nabla h(\mathbf{y}) - \nabla h(\mathbf{x}), \mathbf{u} - \mathbf{x}\rangle. \quad (8)$$

$$D_h(\mathbf{u}, \mathbf{x}) \pm D_{h'}(\mathbf{u}, \mathbf{x}) = D_{h\pm h'}(\mathbf{u}, \mathbf{x}). \quad (9)$$

The property in eq. (8) establishes a relationship among the Bregman distances of three points, while the property in eq. (9) shows the linearity of the Bregman distance with respect to the function h. In summary, Bregman distances are similar to metrics (but they can be asymmetric), and the following are several popular examples.

Example 1. *(Euclidean Distance) For $h : \mathbb{R}^n \to \mathbb{R}$ with $h(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$, $D_h(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$.*

Example 2. *(KL Relative Entropy) For $h : \mathbb{R}^n_+ \to \mathbb{R}$ with $h(\mathbf{x}) = \sum_{j=1}^n x_j \log x_j - x_j$ (with the convention $0 \log 0 = 0$), $D_h(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^n x_j \log \frac{x_j}{y_j} - x_j + y_j$.*

Example 3. *(Burg's Entropy) For $h : \mathbb{R}^n_{++} \to \mathbb{R}$ with $h(\mathbf{x}) = -\sum_{i=1}^n \log x_i$, $D_h(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^n \frac{x_j}{y_j} - \log \frac{x_j}{y_j} - 1$.*

Clearly, the Euclidean distance in Example 1 is a special case of Bregman distance, and hence the proximal algorithms under Bregman distance naturally generalize the corresponding ones under Euclidean distance. The Bregman distances in Examples 2 and 3 have a non-Euclidean structure. In particular, the Kullback-Liebler (KL) relative entropy is useful when the set C is the simplex [14], and the Burg's Entropy is suitable for optimizing Poisson log-likelihood functions [21].

2.2. Connecting PGA-$\mathcal{B}$ to PPA-$\mathcal{B}$

Consider applying PGA-$\mathcal{B}$ to solve **(P2)**. The following standard assumptions are adopted regarding the functions f, g, h.

Assumption 1. *Regarding $f, g, h : \mathbb{R}^n \to (-\infty, +\infty]$:*

1. *Functions f, g are proper, lower semicontinuous and convex functions, f is differentiable on $\mathrm{int}C$; $\mathrm{dom}\, f \supset C, \mathrm{dom}\, g \cap \mathrm{int}C \neq \emptyset$;*
2. *$F^* := \inf_{\mathbf{x}\in C} F(\mathbf{x}) > -\infty$, and the solution set $\mathcal{X}^* := \{\mathbf{x} \,|F(\mathbf{x}) = F^*\}$ is non-empty;*
3. *Function h satisfies the properties in Definition 1.*

To simplify the analysis, we also assume that the iterative step of PGA-$\mathcal{B}$ is well defined, and refer to [17, 21, 11] for a detailed argument. The following theorem establishes the main result that connects PGA-$\mathcal{B}$ with PPA-$\mathcal{B}$.

Theorem 1. *Let Assumption 1 hold and assume that PGA-$\mathcal{B}$ is well defined. Then the iteration step of PGA-$\mathcal{B}$ in eq. (6) is equivalent to the following PPA-$\mathcal{B}$ step:*

$$\mathbf{x}_{k+1} = \operatorname*{argmin}_{\mathbf{x}\in C} \{F(\mathbf{x}) + D_{\ell_k}(\mathbf{x}, \mathbf{x}_k)\}, \tag{10}$$

where D_{ℓ_k} is the Bregman distance based on the function $\ell_k = \frac{1}{\gamma_k}h - f$.

Proof. By linearity in eq. (9) and the definition of Bregman distance in eq. (7), we obtain

$$\begin{aligned} &F(\mathbf{x}) + D_{\ell_k}(\mathbf{x}, \mathbf{x}_k) \\ &= f(\mathbf{x}) + g(\mathbf{x}) + \tfrac{1}{\gamma_k} D_h(\mathbf{x}, \mathbf{x}_k) - D_f(\mathbf{x}, \mathbf{x}_k) \\ &= g(\mathbf{x}) + \langle \mathbf{x}, \nabla f(\mathbf{x}_k)\rangle + \tfrac{1}{\gamma_k} D_h(\mathbf{x}, \mathbf{x}_k) + f(\mathbf{x}_k) - \langle \mathbf{x}_k, \nabla f(\mathbf{x}_k)\rangle. \end{aligned}$$

Thus, by ignoring the last two constant terms, the minimization problem of PGA-$\mathcal{B}$ is equivalent to eq. (10), which is a PPA-$\mathcal{B}$ step with Bregman distance D_{ℓ_k}. This completes the proof. □

Thus, PGA-$\mathcal{B}$ can be mapped exactly into the form of PPA-$\mathcal{B}$, and the form in eq. (10) provides a new insight of PGA—It is PPA with a special Bregman distance D_{ℓ_k}. In particular, the $-f$ part of the Bregman function ℓ_k linearizes the smooth objective function f, i.e., $f(\mathbf{x}) + D_{-f}(\mathbf{x}, \mathbf{x}_k) = f(\mathbf{x}_k) + \langle \mathbf{x} - \mathbf{x}_k, \nabla f(\mathbf{x}_k)\rangle$. This linearizion simplifies the subproblem at each iteration and usually leads to a update rule with closed form.

The PGA-$\mathcal{B}$ is a general framework that covers several existing algorithms. In particular, it reduces to the mirror descent (MD) method for minimizing a differentiable convex function f when g is a constant function. Furthermore, it recovers the classical gradient descent (GD) method for minimizing f when g is a constant and D_h is the Euclidean distance in Example 1. In summary, we have the following observations regarding the update rules of these first order methods.

$$\text{GD} \subset \text{MD} \subset \text{PGA-}\mathcal{B} \subset \text{PPA-}\mathcal{B}.$$

In order to make PGA-$\mathcal{B}$ be a proper PPA-$\mathcal{B}$, the function ℓ_k in the Bregman distance should be convex and independent of k. Then, we are motivated to make the following assumption.

Assumption 2. *There exists $\bar{\gamma} > 0$ such that for all $\gamma_k \equiv \gamma < \bar{\gamma}$, the function $\ell_k \equiv \ell = \frac{1}{\gamma}h - f$ is convex on C.*

Here, we consider the case $\gamma_k \equiv \gamma$, which corresponds to the choice of constant step size of the algorithm. The case with varying stepsize is discussed in Section 3. Assumption 2 has also been considered in [21] to generalize the following conditions for any $\mathbf{x} \in C, \mathbf{y} \in \text{int}C$

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_* \le \tfrac{1}{\gamma}\|\mathbf{x} - \mathbf{y}\|, \tag{11}$$

$$h(\mathbf{x}) - h(\mathbf{y}) - \langle \mathbf{x} - \mathbf{y}, \nabla h(\mathbf{y})\rangle \ge \tfrac{1}{2}\|\mathbf{x} - \mathbf{y}\|^2, \tag{12}$$

where the first condition corresponds to the Lipschitz continuity of ∇f with respect to the dual norm $\|\cdot\|_*$, and the second condition corresponds to the strong convexity of h with respect to the norm $\|\cdot\|$ (hence $\frac{1}{\gamma}h - f$ is convex). In comparison, Assumption 2 does not require function h and f to have the normed structures specified in eqs. (11) and (12). This is clearly more general and is useful, for instance, when the objective function f is the Kullback-Liebler divergence for the Poisson log-likelihood function [21] and h is chosen to be the Burg's entropy in Example 3.

Our view point of PGA-$\mathcal{B}$ is very different from that developed in a concurrent independent work [21]. There, they view PGA-$\mathcal{B}$ as a mirror descent step composed with a PPA-$\mathcal{B}$ step, and develop a generalized descent lemma based on Assumption 2 to analyze the algorithm. Our approach, however, is straightforward — we simply map the PGA-$\mathcal{B}$ step exactly to a PPA-$\mathcal{B}$ step under Assumption 2. This provides a unified view of PGA-$\mathcal{B}$ as a special case of PPA-$\mathcal{B}$, and consequently, the convergence analysis of PGA-$\mathcal{B}$ naturally follows from those of PPA-$\mathcal{B}$. In particular, Lemma 3.3 of [15] proposed the following properties of PPA-$\mathcal{B}$.

Lemma 1. *[15, Lemma 3.3] Consider the problem **(P1)** with optimal solution set $\mathcal{X}^*$ and a corresponding Bregman distance $D_h(\cdot,\cdot) \in \mathcal{B}$. Let $\{\lambda_k\}$ be a sequence of positive numbers and denote $\sigma_k := \sum_{l=1}^{k} \lambda_l$. Then the sequence $\{\mathbf{x}_k\}$ generated by PPA-$\mathcal{B}$ given in eq. (3) satisfy*

$$r(\mathbf{x}_{k+1}) - r(\mathbf{x}_k) \le -D_h(\mathbf{x}_k, \mathbf{x}_{k+1}), \tag{13}$$

$$D_h(\mathbf{x}^*, \mathbf{x}_{k+1}) \le D_h(\mathbf{x}^*, \mathbf{x}_k), \quad \forall \mathbf{x}^* \in \mathcal{X}^*, \tag{14}$$

$$r(\mathbf{x}_k) - r(\mathbf{u}) \le \frac{D_h(\mathbf{u}, \mathbf{x}_0)}{k}, \quad \forall \mathbf{u} \in C, \tag{15}$$

By the connection between PPA-$\mathcal{B}$ and PGA-$\mathcal{B}$ that established in Theorem 1, we now identify $r = F, \lambda_k \equiv 1 (\sigma_k = k), h = \ell$ in Lemma 1, and directly obtain the following results of PGA-$\mathcal{B}$.

Corollary 1. *Under Assumption 1 and Assumption 2, the sequence $\{\mathbf{x}_k\}$ generated by PGA-$\mathcal{B}$ satisfies:*

$$F(\mathbf{x}_{k+1}) - F(\mathbf{x}_k) \le -D_\ell(\mathbf{x}_k, \mathbf{x}_{k+1}), \tag{16}$$

$$D_\ell(\mathbf{x}^*, \mathbf{x}_{k+1}) \le D_\ell(\mathbf{x}^*, \mathbf{x}_k), \quad \forall \mathbf{x}^* \in \mathcal{X}^*, \tag{17}$$

$$F(\mathbf{x}_k) - F(\mathbf{u}) \le \frac{D_\ell(\mathbf{u}, \mathbf{x}_0)}{k}, \quad \forall \mathbf{u} \in C, \tag{18}$$

The result in eq. (16) implies that the sequence of function value is non-increasing, and hence PGA-$\mathcal{B}$ is a descent method. Also, eq. (17) shows that the Bregman distance between $\mathbf{x}_k$ and the optimal solution point $\mathbf{x}^* \in \mathcal{X}^*$ is non-increasing. Moreover, eq. (18) with $\mathbf{u} = \mathbf{x}^* \in \mathcal{X}^*$ implies that the function value sequence $\{F(\mathbf{x}_k)\}$ converges to optimum at a rate $\mathcal{O}(1/k)$.

Similar results are established in [21], but they are in terms of the Bregman distance D_h (not D_ℓ). Moreover, their analysis crucially depends on a symmetry coefficient $\alpha := \inf_{\mathbf{x}\neq\mathbf{y}}\{\frac{D_h(\mathbf{x},\mathbf{y})}{D_h(\mathbf{y},\mathbf{x})}\} \in [0, 1]$, which is avoided via our unified point of view. Thus, our unification of PGA-$\mathcal{B}$ as PPA-$\mathcal{B}$ provides much simplicity of the analysis and avoids introducing the symmetry coefficient α. Moreover, our global estimate in eq. (18) is tighter than the result in [21, Theorem 1, (iv)], since

for all $\mathbf{x} \in C$ and $\mathbf{y} \in \text{int}C$

$$D_\ell(\mathbf{x},\mathbf{y}) \le \frac{1}{\gamma} D_h(\mathbf{x},\mathbf{y}) \le \frac{2}{(1+\alpha)\gamma} D_h(\mathbf{x},\mathbf{y}), \quad \forall \alpha \in [0,1].$$

To further ensure the convergence of $\{\mathbf{x}_k\}$ generated by PGA-$\mathcal{B}$ to a minimizer $\mathbf{x}^* \in \mathcal{X}^*$, the following additional conditions on the Bregman distance D_ℓ are needed, and they are parallel to the conditions introduced in [15, Def 2.1, (iii)-(v)] to analyze the convergence of the sequence that generated by PPA-$\mathcal{B}$.

Corollary 2. *Under Assumption 1 and Assumption 2, the sequence $\{\mathbf{x}_k\}$ generated by PGA-$\mathcal{B}$ converges to some $\mathbf{x}^* \in \mathcal{X}^*$ if the Bregman distance D_ℓ additionaly satisfies*

1. *For every $\mathbf{x} \in C$ and every $\alpha \in \mathbb{R}$, the level set $\{\mathbf{y} \in \text{int}C \mid D_\ell(\mathbf{x},\mathbf{y}) \le \alpha\}$ is bounded;*
2. *If $\{\mathbf{x}_k\} \in \text{int}C$ and $\mathbf{x}_k \to \mathbf{x}^* \in C$, then $D_\ell(\mathbf{x}^*,\mathbf{x}_k) \to 0$;*
3. *If $\{\mathbf{x}_k\} \in \text{int}C$ and $\mathbf{x}^* \in C$ is such that $D_\ell(\mathbf{x}^*,\mathbf{x}_k) \to 0$, then $\mathbf{x}_k \to \mathbf{x}^*$.*

Proof. The proof follows the argument in [15, Theorem 3.4]. □

3. PGA-$\mathcal{B}$ with line search

Assumption 2 requires a constant parameter $\gamma_k \equiv \gamma$, which is usually unknown a priori in practical applications. In particular, it corresponds to the global Lipschitz parameter of ∇f when $h(\cdot) = \frac{1}{2}\|\cdot\|_2^2$. Next, we propose an adaptive version of PGA-$\mathcal{B}$ that searches for a proper γ_k in each step via the backtracking line search method. The algorithm is referred to as PGA-$\mathcal{B}$ with backtracking line search, and the details are summarized in Algorithm 1.

Algorithm 1: PGA-$\mathcal{B}$ with backtracking line search

```
Initialize γ_0 > 0 ,ℓ_0 = (1/γ_0)h − f, 0 < β < 1;
for k = 0, 1, ··· do
    x_{k+1} = argmin_{x∈C}{F(x) + D_{ℓ_k}(x, x_k)};
    while D_{ℓ_k}(x_{k+1}, x_k) < 0 do
        Set γ_k ← βγ_k;
        Repeat the k-th iteration with γ_k;
    end
end
```

We note that the above D_{ℓ_k} may not be a proper Bregman distance since ℓ_k may not be globally convex when γ_k is large. Intuitively, at each iteration we search for a small enough γ_k that guarantees the non-negativity of the Bregman distance D_{ℓ_k} between the successive iterates $\mathbf{x}_{k+1}$ and $\mathbf{x}_k$. This, equivalently, implies that ℓ_k behaves like a convex function between these two successive iterates. Note that in the special case where $h(\cdot) = \frac{1}{2}\|\cdot\|_2^2$, the line search criterion $D_{\ell_k}(\mathbf{x}_{k+1},\mathbf{x}_k) \ge 0$ reduces to that of the PGA-$\mathcal{E}$ with line search in [1]. Next we characterize the convergence rate of PGA-$\mathcal{B}$ with line search.

Theorem 2. *Under Assumption 1 and Assumption 2, the sequence $\{\mathbf{x}_k\}$ generated by PGA-$\mathcal{B}$ with backtracking line search satisfies: for all k and all $\mathbf{x}^* \in \mathcal{X}^*$*

$$F(\mathbf{x}_k) - F^* \le \frac{D_h(\mathbf{x}^*,\mathbf{x}_0)}{\beta\bar{\gamma}k}. \tag{19}$$

Proof. Since the line search method reduces γ_k by a factor of β whenever the Bregman distance is negative, we must have

$$\beta\bar{\gamma} \le \gamma_k \le \bar{\gamma}, \quad \forall k. \tag{20}$$

At the k-th iteration, from PGA-$\mathcal{B}$ we have that

$$F(\mathbf{x}_{k+1}) + D_{\ell_k}(\mathbf{x}_{k+1},\mathbf{x}_k) \le F(\mathbf{x}_k), \tag{21}$$

which, combines with the line search criterion $D_{\ell_k}(\mathbf{x}_{k+1},\mathbf{x}_k) \ge 0$, guarantees that $F(\mathbf{x}_{k+1}) \le F(\mathbf{x}_k)$, i.e., the method is a descent algorithm. Now by the convexity of f and g, for any $\mathbf{x}^* \in \mathcal{X}^*$ we have

$$\begin{aligned} F(\mathbf{x}^*) &\ge f(\mathbf{x}_k) + \langle \mathbf{x}^* - \mathbf{x}_k, \nabla f(\mathbf{x}_k)\rangle \\ &\quad + g(\mathbf{x}_{k+1}) + \langle \mathbf{x}^* - \mathbf{x}_{k+1}, \partial g(\mathbf{x}_{k+1})\rangle \\ &= F(\mathbf{x}_{k+1}) + \langle \mathbf{x}^* - \mathbf{x}_{k+1}, \partial F(\mathbf{x}_{k+1})\rangle \\ &\quad + D_f(\mathbf{x}^*,\mathbf{x}_{k+1}) - D_f(\mathbf{x}^*,\mathbf{x}_k). \end{aligned} \tag{22}$$

On the other hand, the optimality condition of the $(k+1)$-th iteration of PGA-$\mathcal{B}$ implies that

$$\nabla \ell_k(\mathbf{x}_k) - \nabla \ell_k(\mathbf{x}_{k+1}) \in \partial F(\mathbf{x}_{k+1}),$$

which together with the property in eq. (8) further implies that

$$\begin{aligned} &\langle \mathbf{x}^* - \mathbf{x}_{k+1}, \partial F(\mathbf{x}_{k+1})\rangle \\ &\quad = D_{\ell_k}(\mathbf{x}^*,\mathbf{x}_{k+1}) - D_{\ell_k}(\mathbf{x}^*,\mathbf{x}_k) + D_{\ell_k}(\mathbf{x}_{k+1},\mathbf{x}_k) \\ &\quad \ge D_{\ell_k}(\mathbf{x}^*,\mathbf{x}_{k+1}) - D_{\ell_k}(\mathbf{x}^*,\mathbf{x}_k). \end{aligned}$$

Now plug the above inequality into eq. (22), we further obtain that

$$\begin{aligned} 0 &\ge F(\mathbf{x}^*) - F(\mathbf{x}_{k+1}) \\ &\ge \frac{1}{\gamma_k}\left[D_h(\mathbf{x}^*,\mathbf{x}_{k+1}) - D_h(\mathbf{x}^*,\mathbf{x}_k)\right] \\ &\ge \frac{1}{\beta\bar{\gamma}}\left[D_h(\mathbf{x}^*,\mathbf{x}_{k+1}) - D_h(\mathbf{x}^*,\mathbf{x}_k)\right], \end{aligned}$$

where the last inequality follows from negativity and the fact that $\beta\bar{\gamma} \le \gamma_k$. Telescoping the above inequality from 0 to $k-1$ and applying the fact that $F(\mathbf{x}_{k+1}) \le F(\mathbf{x}_k)$, we further obtain that

$$\begin{aligned} kF(\mathbf{x}^*) - kF(\mathbf{x}_k) &\ge kF(\mathbf{x}^*) - \sum_{l=1}^{k} F(\mathbf{x}_l) \\ &\ge \frac{1}{\beta\bar{\gamma}} \sum_{l=0}^{k-1} \left[D_h(\mathbf{x}^*,\mathbf{x}_{l+1}) - D_h(\mathbf{x}^*,\mathbf{x}_l)\right] \\ &\ge -\frac{1}{\beta\bar{\gamma}} D_h(\mathbf{x}^*,\mathbf{x}_0). \end{aligned}$$

The result follows by rearranging the inequality. □

4. Conclusion

In this paper, we point out that PGA-$\mathcal{B}$ can be viewed as PPA-$\mathcal{B}$ with a special choice of Bregman distance. Consequently, the convergence analysis of PGA-$\mathcal{B}$ follows directly from that of PPA-$\mathcal{B}$. Moreover, this unified view point leads to a tighter convergence rate of the function value residual than existing results, and avoids involving the symmetry coefficient. Lastly, we provide a general line search variant of PGA-$\mathcal{B}$ and characterize its convergence rate.

Acknowledgements

The work of Y. Zhou and Y. Liang was supported by the National Science Foundation under Grant CNS-11-16932. The work of L. Shen was supported in part by the National Science Foundation under grant DMS-1115523 and DMS-1522332, and by the National Research Council.

References

[1] A. Beck, M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, SIAM J. Img. Sci. 2 (1) (2009) 183–202.

[2] C. A. Micchelli, L. Shen, Y. Xu, Proximity algorithms for image models: Denoising, Inverse Problems 27 (2011) 045009(30pp).

[3] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, Found. Trends Mach. Learn. 3 (1) (2011) 1–122.

[4] B. Recht, M. Fazel, P. A. Parrilo, Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization, SIAM Rev. 52 (3) (2010) 471–501.

[5] B. Martinet, Brève communication. régularisation d'inéquations variationnelles par approximations successives, ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique 4 (R3) (1970) 154–158.

[6] Y. Nesterov, Smooth minimization of non-smooth functions, Math. Program. 103 (1) (2005) 127–152.

[7] A. Beck, M. Teboulle, Smoothing and first order methods: A unified framework, SIAM Journal on Optimization 22 (2) (2012) 557–580.

[8] R. Rockafellar, Monotone operators and the proximal point algorithm, SIAM Journal on Control and Optimization 14 (5) (1976) 877–898.

[9] J. Eckstein, D. P. Bertsekas, On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators, Mathematical Programming 55 (1992) 293–318.

[10] O. Güler, On the convergence of the proximal point algorithm for convex minimization, SIAM J. Control Optim. 29 (2) (1991) 403–419.

[11] Y. Censor, S. A. Zenios, Proximal minimization algorithm with d-functions, J. Optim. Theory Appl. 73 (3) (1992) 451–464.

[12] M. Teboulle, Entropic proximal mappings with applications to nonlinear programming, Mathematics of Operations Research 17 (3) (1992) 670–690.

[13] M. Teboulle, Convergence of proximal-like algorithms, SIAM Journal on Optimization 7 (4) (1997) 1069–1083.

[14] A. Beck, M. Teboulle, Mirror descent and nonlinear projected subgradient methods for convex optimization, Operations Research Letters 31 (3).

[15] G. Chen, M. Teboulle, Convergence analysis of a proximal-like minimization algorithm using Bregman functions, SIAM Journal on Optimization 3 (3) (1993) 538–543.

[16] J. Eckstein, Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming, Mathematics of Operations Research 18 (1) (1993) 202–226.

[17] A. Auslender, M. Teboulle, Interior gradient and proximal methods for convex and conic optimization, SIAM Journal on Optimization 16 (3) (2006) 697–725.

[18] P. Lions, B. Mercier, Splitting algorithms for the sum of two nonlinear operators, SIAM Journal on Numerical Analysis 16 (6) (1979) 964–979.

[19] A. A. Goldstein, Convex programming in Hilbert space, Bulletin of the American Mathematical Society 70 (5) (1964) 709–710.

[20] P. Tseng, Approximation accuracy, gradient methods, and error bound for structured convex optimization, Mathematical Programming 125 (2) (2010) 263–295.

[21] J. Bolte, H. Bauschke, M. Teboulle, A descent lemma beyond lipschitz gradient continuity: first-order methods revisited and applications, Mathematics of Operations Research.

[22] Y. Zhou, Y. Liang, L. Shen, A new perspective of proximal gradient algorithms, arXiv.
URL https://arxiv.org/abs/1503.05601

[23] 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 Computational Mathematics and Mathematical Physics 7 (3) (1967) 200 – 217.

[24] A. R. De Pierro, A. N. Iusem, A relaxed version of bregman's method for convex programming, Journal of Optimization Theory and Applications 51 (3) (1986) 421–440.