A simple convergence analysis of Bregman proximal gradient algorithm

Yi Zhou · Yingbin Liang · Lixin Shen

Received: 13 December 2017 / Published online: 4 April 2019
© Springer Science+Business Media, LLC, part of Springer Nature 2019

Abstract
In this paper, we provide a simple convergence analysis of proximal gradient algorithm with Bregman distance, which provides a tighter bound than existing result. In particular, for the problem of minimizing a class of convex objective functions, we show that proximal gradient algorithm with Bregman distance can be viewed as proximal point algorithm that incorporates another Bregman distance. Consequently, the convergence result of the proximal gradient algorithm with Bregman distance follows directly from that of the proximal point algorithm with Bregman distance, and this leads to a simpler convergence analysis with a tighter convergence bound than existing ones. We further propose and analyze the backtracking line-search variant of the proximal gradient algorithm with Bregman distance.

Keywords Proximal algorithms · Bregman distance · Convergence analysis · Line-search

The work of Y. Zhou and Y. Liang was supported by the National Science Foundation under Grant CCF-1704169 and ECCS-1609916. The work of L. Shen was supported in part by the National Science Foundation under Grants DMS-1115523 and DMS-1522332, and by the National Research Council.
1 Introduction

Proximal algorithms have been extensively studied in optimization theory, as they are efficient solvers to problems that involve non-smoothness. The proximal algorithms have been widely applied to solve practical problems including image processing, e.g., [3,17], distributed statistical learning, e.g., [6], and low rank matrix minimization, e.g., [19].

Consider the following optimization problem:

\[(P1) \quad \min_{x \in C} r(x), \quad (1)\]

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 [16]. The algorithm generates a sequence \( \{x_k\} \) via the following iterative step:

\[
(PPA - E) \quad x_{k+1} = \arg\min_{x \in C} \left\{ r(x) + \frac{1}{2\lambda_k} \|x - x_k\|^2 \right\}, \quad (2)
\]

where \( \lambda_k > 0 \) corresponds to the step size at the \( k \)-th iteration. We refer to the algorithm as PPA-\(E\) for the choice of the Euclidean distance (i.e., the \( \|\cdot\|^2 \) 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 [4,18]. It was shown in [12,20] that the sequence \( \{x_k\} \) generated by PPA-\(E\) converges to a solution of \((P1)\) with a proper choice of the step size sequence \( \{\lambda_k\} \), and the rate of convergence was characterized in [14].

A natural generalization of PPA-\(E\) 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., [2,8,21,22]. Among them, a popular choice is the Bregman distance [2,9,11]. As a consequence, we obtain the following iterative step of proximal point algorithm with Bregman distance:

\[
(PPA - B) \quad x_{k+1} = \arg\min_{x \in C} \left\{ r(x) + \frac{1}{\lambda_k} D_h(x, x_k) \right\}. \quad (3)
\]

Here, \( D_h(x, x_k) \) corresponds to the Bregman distance between the points \( x \) and \( x_k \), and is based on a continuously differentiable and strictly convex function \( h \). We refer to the algorithm as PPA-\(B\) for the choice of the Bregman distance, which is formally defined in Definition 1. The convergence rate of PPA-\(B\) has been characterized in [9,22], and we refer to [1] for a comprehensive discussion on the PPA with different choices of distance metrics.

A generalized optimization problem of \((P1)\) is the following composite objective minimization problem:

\[(P2) \quad \min_{x \in C} \{ F(x) := f(x) + g(x) \}, \quad (4)\]
where \( f \) is usually a differentiable 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, low-rankness, etc, to the solution of the problem. This composite objective minimization problem generalizes many applications in machine learning, image processing, detection, etc. As an extension of the PPA, splitting algorithms are proposed for solving the composite objective minimization problem in (P2) [12,15].

In particular, the proximal gradient algorithm (PGA) with Euclidean distance has been developed in [13] to solve (P2) efficiently, and the iterative step is given by

\[
\text{(PGA-E)}: \quad x_{k+1} = \arg\min_{x \in C} \left\{ g(x) + \langle x, \nabla f(x_k) \rangle + \frac{1}{2\gamma_k} \| x - x_k \|_2^2 \right\},
\]

where we refer to the algorithm as PGA-E for the choice of Euclidean distance. It has been shown that the sequence of function value residual generated by PGA-E has a convergence rate of \( O(1/k) \) \(^1\) [3], and the rate can be further improved to be \( O(1/k^2) \) via Nesterov’s acceleration technique. Inspired by the way of generalizing PPA-E to PPA-B, PGA-E can also be generalized by replacing the Euclidean distance with the Bregman distance, and correspondingly, the iterative step is given by

\[
\text{(PGA-B)}: \quad x_{k+1} = \arg\min_{x \in C} \left\{ g(x) + \langle x, \nabla f(x_k) \rangle + \frac{1}{\gamma_k} D_h(x, x_k) \right\},
\]

where we refer to the algorithm as PGA-B for the choice of Bregman distance. Under Lipschitz continuity of \( \nabla f \), it has been shown in [23] that the convergence rate of PGA-B is \( O(1/k) \).

It is clear that PGA-E and PGA-B are respectively generalizations of PPA-E and PPA-B, because they coincide when the function \( f \) in (P2) is a constant function. Thus, in existing literature, the analysis of PGA is developed by their own as in [3,23] without resorting to existing analysis of PPA. More recently, a concurrent work [5] to this paper interprets PGA-B as the composition of mirror descent method and PPA-B. In contrast to this viewpoint, this paper shows that PGA-B can, in fact, be viewed as PPA-B under a proper choice of the Bregman distance. Consequently, the analysis of PGA-B can be mapped to that of PPA-B, resulting in a much simpler analysis. We note that the initial version of this paper [24] was posted on arXiv in March, 2015, which already independently developed the aforementioned main result.

We summarize our main contributions as follows. In this paper, we show that PGA-B can be viewed as PPA-B with a special choice of Bregman distance, and thus the convergence results of PGA-B inherit existing convergence results of PPA-B. Following this viewpoint, we obtain a tighter bound of the convergence rate of the function value residual, and our result avoids involving the symmetry coefficient in [5]. Lastly, we propose a line-search variant of PGA-B and characterize its convergence rate.

The rest of the paper is organized as follows. In Sect. 2, we recall the definition of Bregman distance, unify PGA-B as a special case of PPA-B and discuss its convergence

\(^1\) Here, \( f(n) = O(g(n)) \) denotes that \( |f(n)| \leq \xi |g(n)| \) for all \( n > N \), where \( \xi \) is a constant and \( N \) is a positive integer.
results. In Sect. 3, we propose a line-search variant of PGA-$B$ and characterize its convergence rate. Finally in Sect. 4, we conclude our paper with a few remarks on our results.

2 Unifying PGA-$B$ as PPA-$B$

2.1 Preliminaries on Bregman distance

We first recall the definition of the Bregman distance [7], see also [9,10]. Throughout, $C$ denotes a closed convex set, and the interior of set $C \subset \mathbb{R}^n$ is denoted as int$C$.

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

(a) $h$ is continuously differentiable on int$C$;
(b) $h$ is strictly convex on $C$.

Then, the Bregman distance $D_h : C \times \text{int}C \to \mathbb{R}^+$ associated with function $h$ is defined as, for all $x \in C$ and $y \in \text{int}C$,

$$D_h(x, y) = h(x) - h(y) - \langle x - y, \nabla h(y) \rangle.$$  \hfill (7)

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

$$D_h(u, x) + D_h(x, y) - D_h(u, y) = \langle \nabla h(y) - \nabla h(x), u - x \rangle.$$  \hfill (8)

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

We note that $h - h'$ in (9) may not be convex in general, but the application of (9) in our proof does satisfy the convexity requirement for properly defining the Bregman distance. The property in (8) establishes a relationship among the Bregman distances of three points, and the property in (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 of Bregman distance.

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

**Example 2 (KL Relative Entropy)** For $h : \mathbb{R}_+^n \to \mathbb{R}$ with $h(x) = \sum_{j=1}^n x_j \log x_j - x_j$ (with the convention $0 \log 0 = 0$), $D_h(x, y) = \sum_{j=1}^n x_j \log \frac{x_j}{y_j} - x_j + y_j$.
**Example 3** (*Burg’s Entropy*) For $h : \mathbb{R}^n_+ \to \mathbb{R}$ with $h(x) = -\sum_{i=1}^{n} \log x_i$, $D_h(x, y) = \sum_{j=1}^{n} \frac{x_j}{y_j} - \log \frac{x_j}{y_j} - 1$.

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 [2], and the Burg’s Entropy is suitable for optimizing Poisson log-likelihood functions [5].

### 2.2 Connecting PGA-$B$ to PPA-$B$

Consider applying PGA-$B$ 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 int$C$; dom $f \supset C$, dom $g \cap$ int$C \neq \emptyset$.
2. $F^* := \inf_{x \in C} F(x) > -\infty$, and the solution set $\mathcal{X}^* := \{x : F(x) = F^*\}$ is non-empty;
3. Function $h$ satisfies the properties in Definition 1.

To simplify the analysis, we also assume that the iteration step of PGA-$B$ is well defined, and refer to [1,5,8] for a detailed discussion. The following theorem establishes the main result that connects PGA-$B$ with PPA-$B$.

**Theorem 1** Assume the iteration steps of PGA-$B$ are well defined. Then, the iteration step of PGA-$B$ in (6) is equivalent to the following PPA-$B$ step:

$$x_{k+1} = \arg\min_{x \in C} \left\{ F(x) + D_{\ell_k}(x, x_k) \right\},$$

(10)

where the function $\ell_k = \frac{1}{\gamma_k} h - f$.

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

$$F(x) + D_{\ell_k}(x, x_k) = f(x) + g(x) + \frac{1}{\gamma_k} D_h(x, x_k) - D_f(x, x_k) = g(x) + (x, \nabla f(x_k)) + \frac{1}{\gamma_k} D_h(x, x_k) + f(x_k) - (x_k, \nabla f(x_k)).$$

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

Thus, PGA-$B$ can be mapped exactly into the form of PPA-$B$, and the form in (10) provides a new insight of PGA—It is PPA with a special Bregman distance $D_{\ell_k}$. In particular, the $-f$ part of the function $\ell_k$ linearizes the objective function $f$, i.e., $f(x) + D_{-f}(x, x_k) = f(x_k) + (x - x_k, \nabla f(x_k))$, which is a linear function. The
linearizion simplifies the subproblem at each iteration, and leads to an update rule with closed form for a simple regularizer $g$.

We can further understand the class of objective functions that can be solved by PGA-$B$ with a theoretical guarantee by leveraging the above equivalence viewpoint. In particular, to make the equivalent PPA-$B$ step in (10) be proper, the function $\ell_k$ in the Bregman distance should be convex and independent of the iteration $k$. Then, we are motivated to make the following assumption on the composite objective function.

**Assumption 2** For the function $f$ in problem (P2), there exists $\gamma > 0$ and function $h$ in Definition 1 such that for $\gamma_k = \gamma$, $k = 1, 2, \ldots$ with $0 < \gamma < \gamma$, the function $\ell_k := \ell = \frac{1}{\gamma} h - f$ is convex on $C$.

Here, we consider the case $\gamma_k \equiv \gamma$, which corresponds to the choice of constant step size of PGA-$B$. We further provide a backtracking line-search rule for choosing the stepsize in Sect. 3. Assumption 2 generalizes the Lipschitz gradient condition that is adopted in the analysis of the PGA. For example, traditional analysis of PGA under the Euclidean distance assumes that $\nabla f$ is $L$-Lipschitz. Such Lipschitz condition is equivalent to that $\frac{1}{2} \|x\|_2^2 - f(x)$ is convex, i.e., Assumption 2 is satisfied with $h(x) = \frac{1}{2} \|x\|_2^2$, $\gamma = \frac{1}{L}$. Assumption 2 has also been considered in [5] to generalize the assumptions that $\nabla f$ is Lipschitz continuous respect to certain norm. In comparison, Assumption 2 does not require function $h$ and $\nabla f$ to satisfy these structures under certain norm.

Our view point of PGA-$B$ is very different from that developed in a concurrent independent work [5]. There, they view PGA-$B$ as a mirror descent step composed with a PPA-$B$ 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-$B$ step exactly to a PPA-$B$ step under Assumption 2. This provides a unified view of PGA-$B$ as a special case of PPA-$B$, and consequently, the convergence analysis of PGA-$B$ naturally follows from those of PPA-$B$. In particular, Lemma 3.3 of [9] proposed the following properties of PPA-$B$.

**Lemma 1** [9, Lemma 3.3] Consider the problem (P1) with optimal solution set $X^*$. Let $\{\lambda_k\}$ be a sequence of positive numbers and denote $\sigma_k := \sum_{l=1}^k \lambda_l$. Then the sequence $\{x_k\}$ generated by PPA-$B$ given in (3) satisfy

$$r(x_{k+1}) - r(x_k) \leq -D_h(x_k, x_{k+1}),$$

$$D_h(x^*, x_{k+1}) \leq D_h(x^*, x_k), \quad \forall x^* \in X^*,$$

$$r(x_k) - r(u) \leq \frac{D_h(u, x_0)}{\sigma_k}, \quad \forall u \in C,$$

By the connection between PPA-$B$ and PGA-$B$ 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 on the iterate sequence $\{x_k\}$ generated by PGA-$B$.

 Springer
Corollary 1 Under Assumptions 1 and 2, the sequence \( \{x_k\} \) generated by PGA-B for solving problem (P2) satisfies:

\[
F(x_{k+1}) - F(x_k) \leq -D_\ell(x_k, x_{k+1}), \tag{14}
\]

\[
D_\ell(x^*, x_{k+1}) \leq D_\ell(x^*, x_k), \quad \forall x^* \in \mathcal{X}^*, \tag{15}
\]

\[
F(x_k) - F(u) \leq \frac{D_\ell(u, x_0)}{k}, \quad \forall u \in C, \tag{16}
\]

The result in (14) implies that the sequence of function value is non-increasing, and hence PGA-B is a descent method. Also, (15) shows that the Bregman distance between \( x_k \) and the optimal solution point \( x^* \in \mathcal{X}^* \) is non-increasing. Moreover, (16) with \( u = x^* \in \mathcal{X}^* \) implies that the function value sequence \( \{F(x_k)\} \) converges to optimum at a rate \( O(1/k) \).

Similar results to those in Corollary 1 are established in [5], but they are in terms of the Bregman distance \( D_h \) (not \( D_\ell \)). Moreover, their analysis crucially depends on a symmetry coefficient \( \alpha := \inf_{x \neq y} \{ \frac{D_h(x, y)}{D_h(y, x)} \} \in [0, 1] \), which is avoided in our result through the unified point of view. Thus, our unification of PGA-B as PPA-B provides much simplicity of the analysis and avoids introducing the symmetry coefficient \( \alpha \). Moreover, our global estimate in (16) is tighter than the result in [5, Theorem 1, (iv)], since for all \( x \in C \) and \( y \in \text{int}C \)

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

3 PGA-B with line search

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

Algorithm 1: PGA-B with backtracking line-search

1. Initialize \( \gamma_0 > 0, \ell_0 = \frac{1}{\gamma_0} h - f, 0 < \beta < 1; \)
2. for \( k = 0, 1, \cdots \) do
3. \( x_{k+1} = \arg\min_{x \in C} \{ F(x) + D_\ell_k(x, x_k) \}; \)
4. while \( D_\ell_k(x_{k+1}, x_k) < 0 \) do
5. \( \gamma_k \leftarrow \beta \gamma_k; \)
6. Repeat the \( k \)-th iteration with \( \gamma_k; \)
7. end
8. end
We note that the above $D_{\ell_k}$ may not be a proper Bregman distance since $\ell_k$ may not be globally convex when $\gamma_k$ is large. Intuitively, at each iteration we search for a small enough $\gamma_k$ that guarantees the non-negativity of the Bregman distance $D_{\ell_k}$ between the successive iterates $x_{k+1}$ and $x_k$. Intuitively, this implies that $\ell_k$ 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 $D_{\ell_k}(x_{k+1}, x_k) \geq 0$ reduces to that of the PGA-$E$ with line-search in [3]. Next we characterize the convergence rate of PGA-$B$ with line-search.

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

$$F(x_k) - F^* \leq \frac{D_h(x^*, x_0)}{\beta \bar{\gamma} k}.$$  

**Proof** Since the line-search method reduces $\gamma_k$ by a factor of $\beta$ whenever the Bregman distance is negative, we must have

$$\beta \bar{\gamma} \leq \gamma_k \leq \bar{\gamma}, \quad \forall k = 1, 2, \ldots$$  

At the $k$-th iteration, from PGA-$B$ we have that

$$F(x_{k+1}) + D_{\ell_k}(x_{k+1}, x_k) \leq F(x_k),$$  

which, combines with the line-search criterion $D_{\ell_k}(x_{k+1}, x_k) \geq 0$, guarantees that $F(x_{k+1}) \leq F(x_k)$, i.e., the method is a descent algorithm. Now by the convexity of $f$ and $g$, for any $x^* \in X^*$ and any $u_{k+1} \in \partial g(x_{k+1})$ we have

$$F(x^*) \geq f(x_k) + \langle x^* - x_k, \nabla f(x_k) \rangle + g(x_{k+1}) + \langle x^* - x_{k+1}, u_{k+1} \rangle$$

$$= F(x_{k+1}) + \langle x^* - x_{k+1}, \nabla f(x_{k+1}) + u_{k+1} \rangle + D_f(x^*, x_{k+1}) - D_f(x^*, x_k).$$  

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

$$\nabla \ell_k(x_k) - \nabla \ell_k(x_{k+1}) \in \partial F(x_{k+1}).$$

Therefore, one can choose $u_{k+1} \in \partial g(x_{k+1})$ such that $\nabla f(x_{k+1}) + u_{k+1} = \nabla \ell_k(x_k) - \nabla \ell_k(x_{k+1})$, which together with the property in (8) further implies that

$$\langle x^* - x_{k+1}, \nabla f(x_{k+1}) + u_{k+1} \rangle = D_{\ell_k}(x^*, x_{k+1}) - D_{\ell_k}(x^*, x_k) + D_{\ell_k}(x_{k+1}, x_k) \geq D_{\ell_k}(x^*, x_{k+1}) - D_{\ell_k}(x^*, x_k).$$
Substituting the above inequality into (20), we obtain that

\[ 0 \geq F(x^*) - F(x_{k+1}) \]
\[ \geq \frac{1}{\gamma_k} \left[ D_h(x^*, x_{k+1}) - D_h(x^*, x_k) \right] \]
\[ \geq \frac{1}{\beta \gamma} \left[ D_h(x^*, x_{k+1}) - D_h(x^*, x_k) \right], \]

where the last inequality follows from negativity and the fact that \( \beta \gamma \leq \gamma_k \). Telescoping the above inequality from 0 to \( k - 1 \) and applying the fact that \( F(x_{k+1}) \leq F(x_k) \), we obtain that

\[ kF(x^*) - kF(x_k) \geq \sum_{l=1}^{k} F(x_l) \]
\[ \geq \frac{1}{\beta \gamma} \sum_{l=0}^{k-1} \left[ D_h(x^*, x_{l+1}) - D_h(x^*, x_l) \right] \]
\[ \geq -\frac{1}{\beta \gamma} D_h(x^*, x_0). \]

The result follows by rearranging the inequality. \( \square \)

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 viewpoint 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.

References

1. Auslender, A., Teboulle, M.: Interior gradient and proximal methods for convex and conic optimization. SIAM J. Optim. 16(3), 697–725 (2006)
2. Beck, A., Teboulle, M.: Mirror descent and nonlinear projected subgradient methods for convex optimization. Oper. Res. Lett. 31(3), 167–175 (2003)
3. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. 2(1), 183–202 (2009)
4. Beck, A., Teboulle, M.: Smoothing and first order methods: a unified framework. SIAM J. Optim. 22(2), 557–580 (2012)
5. Bolte, J., Bauschke, H., Teboulle, M.: A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. Math. Oper. Res. 42, 330–348 (2016)
6. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. 3(1), 1–122 (2011)
7. Bregman, L.M.: 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. 7(3), 200–217 (1967)
8. Censor, Y., Zenios, S.A.: Proximal minimization algorithm with d-functions. J. Optim. Theory Appl. 73(3), 451–464 (1992)
9. Chen, G., Teboulle, M.: Convergence analysis of a proximal-like minimization algorithm using Bregman functions. SIAM J. Optim. 3(3), 538–543 (1993)
10. De Pierro, A.R., Iusem, A.N.: A relaxed version of Bregman’s method for convex programming. J. Optim. Theory Appl. 51(3), 421–440 (1986)
11. Eckstein, J.: Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming. Math. Oper. Res. 18(1), 202–226 (1993)
12. Eckstein, J., Bertsekas, D.P.: On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators. Math. Program. 55, 293–318 (1992)
13. Fukushima, M., Mine, H.: A generalized proximal point algorithm for certain non-convex minimization problems. Int. J. Syst. Sci. 12(8), 989–1000 (1981)
14. Güler, O.: On the convergence of the proximal point algorithm for convex minimization. SIAM J. Control Optim. 29(2), 403–419 (1991)
15. Lions, P., Mercier, B.: Splitting algorithms for the sum of two nonlinear operators. SIAM J. Numer. Anal. 16(6), 964–979 (1979)
16. Martinet, B.: Bréve communication: régularisation d’inéquations variationnelles par approximations successives. ESAIM Math. Modell. Numer. Anal. 4(R3), 154–158 (1970)
17. Micchelli, C.A., Shen, L., Xu, Y.: Proximity algorithms for image models: denoising. Inverse Probl. 27, 045009 (2011)
18. Nesterov, Y.: Smooth minimization of non-smooth functions. Math. Program. 103(1), 127–152 (2005)
19. Recht, B., Fazel, M., Parrilo, P.A.: Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. SIAM Rev. 52(3), 471–501 (2010)
20. Rockafellar, R.: Monotone operators and the proximal point algorithm. SIAM J. Control Optim. 14(5), 877–898 (1976)
21. Teboulle, M.: Entropic proximal mappings with applications to nonlinear programming. Math. Oper. Res. 17(3), 670–690 (1992)
22. Teboulle, M.: Convergence of proximal-like algorithms. SIAM J. Optim. 7(4), 1069–1083 (1997)
23. Tseng, P.: Approximation accuracy, gradient methods, and error bound for structured convex optimization. Math. Program. 125(2), 263–295 (2010)
24. Zhou, Y., Liang, Y., Shen, L.: A new perspective of proximal gradient algorithms. arXiv: 1503.05601 (2015)

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.