Proximal-Like Incremental Aggregated Gradient Method with Linear Convergence under Bregman Distance Growth Conditions

Hui Zhang*    Yu-Hong Dai†    Lei Guo‡

November 6, 2017

Abstract

We introduce a unified algorithmic framework, called proximal-like incremental aggregated gradient (PLIAG) method, for minimizing the sum of smooth convex component functions and a proper closed convex regularization function that is possibly non-smooth and extended-valued, with an additional abstract feasible set whose geometry can be captured by using the domain of a Legendre function. The PLIAG method includes many existing algorithms in the literature as special cases such as the proximal gradient (PG) method, the incremental aggregated gradient (IAG) method, the incremental aggregated proximal (IAP) method, and the proximal incremental aggregated gradient (PIAG) method. By making use of special Lyapunov functions constructed by embedding growth-type conditions into descent-type lemmas, we show that the PLIAG method is globally convergent with a linear rate provided that the step-size is not greater than some positive constant. Our results recover existing linear convergence results for incremental aggregated methods even under strictly weaker conditions than the standard assumptions in the literature.

Keywords. incremental aggregated gradient, linear convergence, Legendre function, Lyapunov function, Bregman distance growth.

AMS subject classifications. 90C25, 90C22, 90C20, 65K10.

1 Introduction

In this paper, we consider the following convex minimization problem:

\[
\min_{x \in Q} \Phi(x) := F(x) + h(x),
\]

where \( F(x) := \sum_{n=1}^{N} f_n(x) \) is the smooth part whose component functions \( f_n : \mathbb{R}^d \to \mathbb{R} \) are smooth and convex, \( h : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\} \) is a proper closed convex regularization function that is possibly non-smooth and extended-valued, and \( Q \subset \mathbb{R}^d \) is a nonempty closed convex set with a nonempty interior. Many problems in machine learning, signal processing, image science, communication systems, and distributed optimization can be modeled into this form.

*Department of Mathematics and Systems Science, College of Science, National University of Defense Technology, Changsha, Hunan 410073, China. Email: h.zhang1984@163.com
†LSEC, ICMSEC, Chinese Academy of Sciences. Email: dyh@lsec.cc.ac.cn
‡Corresponding author. Sino-US Global Logistics Institute, Shanghai Jiao Tong University, Shanghai 200030, China. Email: guolayne@sjtu.edu.cn
1.1 Existing algorithms

The well-known method to solve \( (1) \) with \( Q = \mathbb{R}^d \) is the proximal gradient (PG) method, which is also called the forward-backward splitting method. This method consists of the composition of a gradient (forward) step of \( F \) with a proximal (backward) step on \( h \). Concretely, it reads as

\[
y_k = x_k - \alpha \cdot \nabla F(x_k),
\]

\[
x_{k+1} = \arg \min_{x \in \mathbb{R}^d} \{ h(x) + \frac{1}{2\alpha} \| x - y_k \|^2 \},
\]

where \( \alpha > 0 \) is some step-size.

The advantage of the PG method lies in exploiting the smooth and non-smooth structure of the model. However, if the number \( N \) is very large, which frequently happens in big data science, then evaluating the full gradient of \( F \) at some point \( x \), \( \nabla F(x) = \sum_{n=1}^{N} \nabla f_n(x) \), is costly and even prohibitive. One efficient approach to overcome this difficulty is to approximate the gradient of \( F \) by using some component functions in a cyclic or randomized order at each iteration \([5, 6]\).

Moreover, the proximal incremental aggregated gradient (PIAG) method was proposed and studied in \([1, 21, 22]\). At each iteration \( k \geq 0 \), the PIAG method first constructs a vector that aggregates the gradients of all component functions, evaluated at the \( (k - \tau^n_k) \)-th iteration, i.e.,

\[
g_k = \sum_{n=1}^{N} \nabla f_n(x_k - \tau^n_k),
\]

where \( \tau^n_k \) are some nonnegative integers. Note that if \( \tau^n_k \equiv 0 \), then \( g_k \) is actually the full gradient of \( F \) at \( x_k \). This vector \( g_k \) is used to approximate the full gradient of \( F \) at the current iteration point \( x_k \) by exploiting the additive structure of the \( N \) component functions. After having obtained \( g_k \), the PIAG method performs a proximal step on the sum of \( h(x) \) and \( \langle g_k, x - x_k \rangle \) as follows:

\[
x_{k+1} = \arg \min_{x \in \mathbb{R}^d} \{ (h(x) + \langle g_k, x - x_k \rangle) + \frac{1}{2\alpha} \| x - x_k \|^2 \}. \tag{2}
\]

By introducing an auxiliary vector, we can rewrite the PIAG method into the following scheme:

\[
y_k = x_k - \alpha \cdot g_k,
\]

\[
x_{k+1} = \arg \min_{x \in \mathbb{R}^d} \{ h(x) + \frac{1}{2\alpha} \| x - y_k \|^2 \).
\]

One can see that the PIAG method differs from the PG method mainly at the choice of gradients of \( F \) at the current iteration point \( x_k \), and these two methods are clearly the same when \( \tau^n_k \equiv 0 \). In this sense, the PIAG method can be viewed as a generalization of the PG method. In addition, when the regularization function \( h \) vanishes, the PIAG method reduces to the incremental aggregated gradient (IAG) method

\[
x_{k+1} = x_k - \alpha \cdot \sum_{n=1}^{N} \nabla f_n(x_k - \tau^n_k), \tag{3}
\]

which is designed for minimizing \( F(x) = \sum_{n=1}^{N} f_n(x) \). Recently, slightly different from the IAG method, Bertsekas in [7] proposed the incremental aggregated proximal (IAP) method

\[
x_{k+1} = x_k - \alpha (\nabla f_{i_k}(x_{k+1}) + \sum_{i \neq i_k} \nabla f_i(x_{k-\tau_i_k})) \tag{4}
\]
which is an implicit scheme since \( x_{k+1} \) is involved in \( \nabla f_{i_k}(x_{k+1}) \). It is straightforward to verify that the IAP method has the following equivalent form

\[
x_{k+1} = \arg \min_{x \in \mathbb{R}^d} \{ f_{i_k}(x) + \langle \sum_{i \neq i_k} \nabla f_i(x_{k-1}^i), x \rangle + \frac{1}{2\alpha} \| x - x_k \|^2 \}.
\]

In other words, one of the component functions, \( f_{i_k} \), is kept in the subproblem at each iteration.

We can include the abstract feasible set \( Q \) in the objective function by using the indicator function \( \delta_Q(x) \) which is defined later on. In this sense, the objective function can be rewritten as \( \Phi(x) = F(x) + \bar{h}(x) \), where \( \bar{h}(x) := h(x) + \delta_Q(x) \). Then problem (1) becomes

\[
\min_{x \in \mathbb{R}^d} \Phi(x) := F(x) + \bar{h}(x).
\]

The PG-type method with Euclidean distance can directly be applied to this formulation. Another approach to handle the abstract feasible set \( Q \) is to choose a Legendre function on \( Q \) and use its associated Bregman distance as a proximity measure; see, e.g., [2]. By using this Bregman distance to replace the Euclidean distance \( \frac{1}{2\alpha} \| x - y_k \|^2 \) in (2), the PIAG method with general distance functions is proposed as follows:

\[
x_{k+1} = \arg \min_{x \in Q} \{ (h(x) + \langle g_k, x - x_k \rangle) + \frac{1}{\alpha} D_w(x, x_k) \}.
\] (5)

A natural question to ask is whether one can develop a unified algorithmic framework to cover all the iteration schemes mentioned above. This is the first motivation of this paper.

1.2 Linear convergence results

It is not trivial to extend linear convergence results from the classic gradient method to IAG method. Recently, by viewing the IAG iteration (3) as a gradient method with errors and using distances of the iterates to the optimal solution as a Lyapunov function, Gurbuzbalaban et al. derived the first globally linear convergence for the IAG method in [14]. Later on, their idea was successfully employed to prove the linear convergence of the IAP method [7]. However, as pointed out by the authors of [7] and [21], the proof method developed in [14] is not readily extended to the constrained or non-smooth composite cases.

The reference [21] is the first to establish the global convergence rate of the PIAG method. In contrast with technique in [14], their analysis uses function values to track the evolution of the iterates generated by the PIAG method, and their results improve upon the best known condition number dependence of the convergence rate of the IAG method. By introducing a key lemma that characterizes the linear convergence of a Lyapunov function, Aytekin et al. obtained a globally linear convergence rate about distances of the iterates to the optimal solution in [1]. Later on, by combining the results presented in [21] and [1], a stronger linear convergence rate for the PIAG method in the sense of function values is given in [22].

In order to guarantee first order methods to converge linearly, one of standard assumptions is the strong convexity of the objective function, which however may be too stringent in practice sometimes. Recently, there has developed a group of interesting results about linear convergence of first order methods for non-strongly convex optimization [8, 9, 12, 15, 18, 19, 23, 24]. We wonder whether one can establish linear convergence for the IAG and PIAG methods without the strong
convexity, which is a common assumption made in \cite{1, 7, 14, 21, 22}. This is our second motivation for this study.

Another required assumption in the literature for analyzing the convergence of first order methods is the gradient Lipschitz continuity of the smooth part of the objective function. To weaken this assumption, Lipschitz-like/convexity condition and relative smoothness condition were introduced in \cite{2} and \cite{18} separately and used to study the (sub)linear convergence of the gradient and PG methods with general distance functions. It should be noted that these two conditions are equivalent. Our last but not least motivation is to analyze the linear convergence of the PIAG method and other more general methods under these new notions.

1.3 Contributions

Our contribution made in this paper is two-fold: The first is that we propose a unified algorithmic framework that includes all the mentioned schemes in Subsection 1.1 as special cases; Moreover, a new proof strategy is developed to derive linear convergence for the proposed algorithmic framework without the strong convexity and Lipschitz gradient continuity of the smooth function. The new proof strategy is based on certain Lyapunov functions, which are constructed by a delicate embedding of growth-type conditions into descent-type lemmas. This makes our proof technically different from existing ones as in \cite{1} and \cite{22}. We point out that when our proposed method reduces to the PIAG method, our results can recover existing linear convergence results of the PIAG method under strictly weaker conditions.

Our perspective is new even for the traditional (proximal) gradient methods. When specialized to the PG method with Bregman distance functions, our framework may be viewed as a complement to that in \cite{2}, and can also be viewed as a further development of the convergence theory developed recently in \cite{12, 15, 19, 23}.

1.4 Organization

The rest of the paper is organized as follows. In Section 2, we list all the assumptions which are useful in the paper. In Section 3, we propose the unified algorithmic framework. In Section 4, we analyze the linear convergence for our proposed algorithm. As an extension, in Section 5, we derive an improved locally linear convergence for the PIAG method under a Hölderian growth condition. Some discussions can be found in Section 6. All proofs are given in Appendix.

Notation. The notation used in this paper is standard as in the literature. $\| \cdot \|$ stands for the Euclidean norm. For any $x \in \mathbb{R}^d$ and any nonempty $\Omega \subset \mathbb{R}^d$, the Euclidean distance from $x$ to $\Omega$ is defined by $d(x, \Omega) := \inf_{y \in \Omega} \| x - y \|$. We let $\delta_{\Omega}(\cdot)$ stand for the indicator function which is equal to 0 if $x \in \Omega$ and $\infty$ otherwise, and $\overline{\Omega}$ denote the closure of $\Omega$. We let $\mathcal{X}$ be the optimal solution set of problem \cite{1}, and $\Phi^*$ be the associated optimal function value. We always assume that $\mathcal{X}$ is nonempty and compact. If $\mathcal{X}$ consists of a solitary point, we let $x^*$ present the unique optimal solution. For simplicity, we let $\mathcal{N} := \{1, 2, \ldots, N\}$ and $\mathcal{T} := \{0, 1, \ldots, \tau\}$.

2 Assumptions

In this section, we give all the assumptions that may be used in this paper. Let the feasible set $\mathcal{Q} \subset \mathbb{R}^d$ be a nonempty, closed, and convex set with a nonempty interior. A standard strategy to
handle $Q$ is to choose a Legendre function on $Q$ and then use its associated Bregman distance as a proximity measure.

First of all, let us recall the definition of a Legendre function; see, e.g., [20]. Let $\varphi : \mathbb{R}^d \to (-\infty, +\infty]$ be a lsc proper convex function. We say that it is essentially smooth if $\text{int dom}\varphi \neq \emptyset$, $\varphi$ is differentiable on $\text{int dom}\varphi$, and $\|\nabla \varphi(x_k)\| \to \infty$ for every sequence $\{x_k\}_{k \geq 0} \subseteq \text{int dom}\varphi$ converging to a boundary point of $\text{dom}\varphi$ as $k \to \infty$. We say that it is of Legendre type if $\varphi$ is essentially smooth and strictly convex on $\text{int dom}\varphi$.

Since $\text{int } Q$ is an open convex set in $\mathbb{R}^d$, there is a Legendre function $w$ such that $\text{int } Q = \text{dom } w$; see, e.g., [10, Theorem 3.5]. Thus, we have that $Q = \text{dom } w$. Associated with $w$, the Bregman distance is defined by:

$$D_w(y, x) = w(y) - w(x) - \langle \nabla w(x), y - x \rangle, \quad \forall y \in \text{dom } w, x \in \text{int dom } w.$$  

In contrast to the Euclidean distance, it is lacking symmetry. If $Q = \mathbb{R}^d$ and we take $w(x) = \frac{1}{2}\|x\|^2$, then $D_w(y, x) = \frac{1}{2}\|y - x\|^2$. In this sense, the Bregman distance generalizes the Euclidean distance.

### 2.1 Standard assumptions

The following are the standard assumptions in the literature which are also used in [1] and [22].

A0. The time-varying delays $\tau_k^n$ are bounded, i.e., there is a nonnegative integer $\tau$ such that

$$\tau_k^n \in \mathcal{T}$$

holds for all $k \geq 1$ and $n \in \mathcal{N}$. Such $\tau$ is called the delay parameter.

A1. Each component function $f_n$ is convex with $L_n$-continuous gradient, i.e.,

$$\|\nabla f_n(x) - \nabla f_n(y)\| \leq L_n\|x - y\|, \quad \forall x, y \in \mathbb{R}^d.$$

This assumption implies that the sum function $F$ is convex with $L$-continuous gradient, where $L = \sum_{n=1}^{N} L_n$.

A2. The function $h$ is subdifferentiable everywhere in its effective domain, i.e., $\partial h(x) \neq \emptyset$ for all $x \in \{y \in \mathbb{R}^d : h(y) < \infty\}$.

A3. The sum function $F$ is $\mu$-strongly convex on $\mathbb{R}^d$ for some $\mu > 0$, i.e., the function $x \mapsto F(x) - \frac{\mu}{2}\|x\|^2$ is convex.

### 2.2 Growth conditions

Recently, more and more researches indicate that the strong convexity assumption is not necessary to obtain linear convergence for first order methods; see, e.g., [24]. Usually, it can be replaced by the following quadratic growth condition:

A3a. The objective function $\Phi(x)$ satisfies the quadratic growth condition, i.e., there is a positive number $\mu > 0$ such that

$$\Phi(x) - \Phi^* \geq \frac{\mu}{2}d^2(x, \mathcal{X}), \quad \forall x \in Q. \quad (6)$$
This condition may be seen as a generalization of the strong convexity of \( \Phi \) (see [19, 24]). Indeed, when \( \Phi \) is differential or \( \mathcal{Q} \) is the whole space, by using the optimality conditions, it is easy to verify that the strong convexity of \( \Phi \) implies that there is a positive number \( \nu > 0 \) such that
\[
\Phi(x) - \Phi^* \geq \frac{\nu}{2} \|x - x^*\|^2 = \frac{\nu}{2} d^2(x, \mathcal{X}), \quad \forall x \in \mathcal{Q}.
\]
However, for the case that \( \Phi \) is not differential or \( \mathcal{Q} \) is a proper subset of \( \mathbb{R}^d \), the above relation may not be true due to the possible failure of the optimality conditions. But when the condition \( \text{rint dom} \Phi \cap \text{rint} \mathcal{Q} \neq \emptyset \) holds where “rint” stands for the relative interior, there always exists \( g \in \partial \Phi(x^*) \) such that
\[
\langle g, x - x^* \rangle \geq 0, \quad \forall x \in \mathcal{Q}.
\]
Based on this, it is easy to show that the strong convexity of \( \Phi \) implies the quadratic growth condition. Obviously the strong convexity of \( F \) and the convexity of \( h \) is stronger than the strong convexity of \( \Phi \). Thus, the assumption A3a is weaker than the assumption A3. Moreover, we can easily construct functions to show that the quadratic growth condition does not imply the strong convexity. We refer to Appendix A for more examples of such type.

The quadratic growth condition (6) is a special case of the following Hölderian growth condition, under which the linear convergence rate of the PG method depends on the parameter \( \theta \) (see [8,13]). When \( \theta \) becomes smaller, the convergence speed will be faster.

A3b. The objective function \( \Phi(x) \) satisfies the Hölderian growth condition with \( 0 < \theta \leq 1 \), i.e., there is a positive number \( \mu > 0 \) such that
\[
\Phi(x) - \Phi^* \geq \frac{\mu}{2} d^{2\theta}(x, \mathcal{X}), \quad \forall x \in \mathcal{Q}.
\]
As pointed out in [13] that the bigger the parameter \( \theta \) is, the more “flat” the function is around its minimizers which in turns means that a gradient descent algorithm shall converge slowly. In this paper, we will show that this intuition applies to the PIAG method as well, and faster convergence rates of the PIAG method can be expected when \( \theta \) is less than 1.

### 2.3 Modified assumptions

In this subsection, we introduce a group of assumptions that modifies (or say, generalizes) existing conditions:

A1'. Each component function \( f_n \) is proper lsc convex with \( \text{dom} w \subset \text{dom} f_n \), which is differentiable on \( \text{int dom} w \). In addition, it is \( L_n \)-smooth relative to \( w \), i.e.,
\[
f_n(y) \leq f_n(x) + \langle \nabla f_n(x), y - x \rangle + L_n D_w(y, x), \quad \forall x, y \in \text{int dom} w.
\]
A2'. \( \text{dom} h \cap \text{int dom} w \neq \emptyset \). Moreover, \( h \) is subdifferentiable in \( \text{dom} h \cap \text{int dom} w \).
A3'. The objective function \( \Phi \) is assumed to satisfy the Bregman distance growth condition, i.e., there exists a real number \( \mu > 0 \) such that
\[
\Phi(y) - \Phi^* \geq \mu \inf_{z \in \mathcal{X}} D_w(z, y), \quad \forall y \in \text{int dom} w.
\]
A4'. For any sequence \( \{v_1, v_2, \ldots, v_k\} \) with \( v_i \in \text{int} \text{ dom} w \; i = 1, \ldots, k \), it holds that

\[
D_w(v_k, v_j) \leq \ell(k - j) \sum_{i=j}^{k-1} D_w(v_{i+1}, v_i), \quad \forall 1 \leq j < k \leq k_0,
\]

where \( \ell(k) \) is a monotonic increasing function with \( \ell(1) = 1 \).

The relative smoothness, appeared in A1', was independently introduced in the recent papers [2] and [18] for relaxing the global gradient Lipschitz continuity assumption. It can be viewed an extended or weaken Lipschitz continuity of gradient. Very recently, Bolte et al. used this assumption to analyze the convergence of nonconvex composition minimization problems in [9].

The Bregman distance growth condition, appeared in A3', is a new notion. It can be viewed as a generalization of the quadratic growth condition when using the Bregman distance instead of the Euclidean distance. We give two conditions in the Appendix B to ensure it to hold.

The motivation to introduce A4' is that we need an analog of the following inequality for the Bregman distance function:

\[
\|v_k - v_j\|^2 = \| \sum_{i=j}^{k-1} (v_{i+1} - v_i) \|^2 \leq (k - j) \sum_{i=j}^{k-1} \| v_{i+1} - v_i \|^2, \quad 1 \leq j < k.
\]

If there exist positive constants \( \mu_w \) and \( L_w \) such that \( D_w(\cdot, \cdot) \) satisfies the following condition:

\[
\frac{\mu_w}{2} \| x - y \|^2 \leq D_w(x, y) \leq \frac{L_w}{2} \| x - y \|^2,
\]

then

\[
D_w(v_k, v_j) \leq \frac{L_w}{2} \| v_k - v_j \|^2 \leq \frac{(k - j)L_w}{2} \sum_{i=j}^{k-1} \| v_{i+1} - v_i \|^2 \leq \frac{(k - j)L_w}{\mu_w} \sum_{j=1}^{k-1} D_w(v_{j+1}, v_j).
\]

This inequality implies that the assumption A4' holds with \( \ell(k - j) := \frac{L_w}{\mu_w} (k - j) \). Thus, the assumption A4' is not stronger than the condition (9) that was used in [1].

3 Unified algorithmic framework

In this section, we describe our unified algorithmic framework for solving problem (1).

**Proximal-Like Incremental Aggregated Gradient (PLIAG) Method:**

i) Choose a Legendre function \( w \) such that \( Q = \text{dom} w \). Choose \( \tau \geq 0 \), \( x_0 \in \text{int} \text{ dom} w \).

Assume that \( x_i = 0 \) for all \( i < 0 \). Let \( k = 0 \).

ii) Choose \( \alpha_k > 0 \), \( \emptyset \neq J_k \subseteq \mathbb{N} \), and \( \tau_k^n \in \mathcal{T} \) for all \( n \in \mathbb{N} \). Let \( I_k \) be the complement of \( J_k \) with respect to \( \mathbb{N} \). The next iteration point \( x_{k+1} \) is obtained via

\[
x_{k+1} = \arg \min_{x \in Q} \{ \Phi_k(x) := h(x) + \sum_{i \in I_k} f_i(x) + \langle \sum_{j \in J_k} \nabla f_j(x_{k-\tau_k^n}), x \rangle + \frac{1}{\alpha_k} D_w(x, x_k) \}.
\]

iii) Let \( k = k + 1 \) and go to Step ii).
Note that $J_k$ in Step ii) is not required to be the same for different $k$. Moreover, since $\text{dom} w \subset Q$, the subproblem (10) can be equivalently written as:

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^d} \{ \Phi_k(x) \}.$$ 

In the following result, we show that the PLIAG method is well-posed under some modified assumptions.

**Proposition 1.** Assume that the modified assumptions $A1'$ and $A4'$ hold. Let $\alpha_k \leq \frac{1}{\sum_{j \in J_k} L_j}$. If $x_j \in \text{int} \text{dom} w$ for all $j \leq k$, then the problem in (10) has a unique solution in $\text{int} \text{dom} w$.

When $Q = \mathbb{R}^d$, let $w(x) = \frac{1}{2} \|x\|^2$ and hence $D_w(x, x_k) = \frac{1}{2} \|x - x_k\|^2$. Then it is easy to see that a). if $I_k = \emptyset$ and $h(x) \equiv 0$, then the iterate (10) recovers the IAG method, b). if $I_k = \{i_k\}$ and $h(x) \equiv 0$, then the iterate (10) recovers the IAP method, c). if $I_k = \emptyset$, then the iterate (10) recovers the PIAG method, and d). if $N = 1$ and $\tau_n^\alpha \equiv 0$, then the iterate (10) recovers the corresponding algorithms in [2] and [18].

When $Q \neq \mathbb{R}^d$, the iterate (10) not only recovers the PIAG method with general distance function (see iterate (5)), but also provides us with incremental aggregated versions of the algorithms studied in [2, 18]. The PLIAG method is indeed a unified algorithmic framework, which includes all the algorithms mentioned in Subsection 1.1 as special cases. A self-contained linear convergence analysis for the PLIAG method is undoubtedly very important, which will be given in Section 4 and Appendix.

### 4 Key Lemmas and Main Results

Throughout this section, we remind the reader that for simplicity we consider the sequence $\{x_k\}$ generated by the PLIAG method with $\alpha_k \equiv \alpha$. All the obtained results and the proofs are also valid for the PLIAG method with different $\alpha_k$. First of all, we introduce a key result, which was given in [11].

**Lemma 1.** Assume that the nonnegative sequences $\{V_k\}$ and $\{w_k\}$ satisfy

$$V_{k+1} \leq aV_k - bw_k + c \sum_{j=k-k_0}^{k} w_j, \quad \forall k \geq 0,$$

for some real numbers $a \in (0, 1)$, $b \geq 0$, $c \geq 0$, and some nonnegative integer $k_0$. Assume also that $w_k = 0$ for $k < 0$, and the following holds:

$$\frac{c}{1 - a} \frac{1 - a^{k+1}_0}{a^{k_0}} \leq b. \quad (11)$$

Then, $V_k \leq a^k V_0$ for all $k \geq 0$. 

8
In addition, we need another crucial result, which can be viewed as a generalization of the standard descent lemma (i.e., [4, Lemma 2.3]) for the PG method.

**Lemma 2.** Suppose that the modified assumptions $A1'$, $A2'$, and $A4'$ hold. Let $x_j \in \text{int dom} w$ for all $j \leq k$ and $x_{k+1}$ be obtained via (10). Let $\tilde{L}_k := \sum_{j \in J_k} L_j$, $L := \max_{k \geq 0} \tilde{L}_k$, and

$$\Delta_k = L \cdot \ell(\tau + 1) \sum_{j = k - \tau}^k D_w(x_{j+1}, x_j).$$

Then, the following holds:

$$\Phi(x_{k+1}) \leq \Phi(x) + \frac{1}{\alpha} D_w(x, x_k) - \frac{1}{\alpha} D_w(x, x_{k+1}) - \frac{1}{\alpha} D_w(x_{k+1}, x_k) + \Delta_k, \quad \forall x \in \text{dom} w.$$

Now, we can state the main result of this paper.

**Theorem 1.** Let $\tilde{L}_k := \sum_{j \in J_k} L_j$ and $L := \max_{k \geq 0} \tilde{L}_k$, where $L_j$ are constants in the assumption $A1'$. Suppose that the modified assumptions $A1'$-$A4'$ hold, and the step-size $\alpha$ satisfies:

$$0 < \alpha \leq \alpha_0 := \left(1 + \ell(\tau + 1)\right)^{\frac{1}{\mu} - 1},$$

where $\ell$ is the function in the assumption $A4'$. Define a Lyapunov function as

$$\Gamma_\alpha(x) := \Phi(x) - \Phi^* + \frac{1}{\alpha} \inf_{z \in X} D_w(z, x).$$

Then, the PLIAG method converges linearly in the sense that

$$\Gamma_\alpha(x_k) \leq \left(1 - \frac{\alpha \mu}{1 + \alpha \mu}\right)^k \Gamma_\alpha(x_0), \quad \forall k \geq 0.$$  \hspace{1cm} (14)

In particular, the PLIAG method attains a globally linear convergence in function values:

$$\Phi(x_k) - \Phi^* \leq \left(1 - \frac{\alpha \mu}{1 + \alpha \mu}\right)^k \Gamma_\alpha(x_0), \quad \forall k \geq 0,$$  \hspace{1cm} (15)

and a globally linear convergence in Bregman distances of the iterates to the optimal solution set:

$$\inf_{z \in X} D_w(z, x_k) \leq \alpha \Gamma_\alpha(x_0) \left(1 - \frac{\alpha \mu}{1 + \alpha \mu}\right)^{k+1}, \quad \forall k \geq 0.$$  \hspace{1cm} (16)

Furthermore, if $\alpha = \alpha_0$, then

$$\Gamma_{\alpha_0}(x_k) \leq \left(1 - \frac{1}{[\ell(\tau + 1)Q + 1][\tau + 1]}\right)^k \Gamma_{\alpha_0}(x_0), \quad \forall k \geq 0,$$  \hspace{1cm} (17)

where $Q = L/\mu$ stands for the condition number.
A key ingredient in Theorem 1 is the Lyapunov function (13), which includes two terms: the function value difference \( \Phi(x) - \Phi^* \) and the Bregman distance to the optimal solution set \( \inf_{z \in X} D_w(z, x) \). It is different from the Lyapunov function, using only one of the two terms, constructed in the existing convergence analysis for PIAG. Our new Lyapunov function can thus provide us with a unified analysis for globally linear convergence in both the function values and Bregman distances of the iterates to the optimal solution set. If we let \( I_k = \emptyset \), \( Q = \mathbb{R}^d \), and \( w(x) = \frac{1}{2} \|x\|^2 \), then PLIAG method reduces to the PIAG method whose linear convergence of the objective function values and the distance of the iterates to the optimal solution set are studied respectively in [21,22] and [1] under the standard assumptions A1-A3. Our results can recover these two classes of convergence under strictly weaker assumptions (See the comparison of assumptions in Appendix). In particular,

- Our result (15) can recover the rate of linear convergence in [21] and moreover, if the delay parameter is chosen as \( \tau \leq 47 \) and \( L \geq \mu \), then by choosing \( \ell \) as the identity function, our result (17) can recover a better rate of linear convergence given in [22] as follows:
  \[
  \Phi(x_k) - \Phi^* \leq \left(1 - \frac{1}{49Q(\tau + 1)}\right)^k \Gamma_{\alpha_0}(x_0), \quad \forall k \geq 0. \tag{18}
  \]
  Here we need to point out that when \( \tau > 47 \), the result (18) given in [22] is better than our result (17). This leaves us a question whether one can derive the stronger result (18) for any nonnegative \( \tau \) under the modified assumptions A1'-A4'. We will consider it for future work.

- In the setting of the problem considered in [1], our result (16) reads as
  \[
  \|x_k - x^*\| \leq 2\alpha \Gamma_{\alpha}(x_0) \left(\frac{1}{\mu \alpha + 1}\right)^k, \quad \forall k \geq 0,
  \]
  which is actually the linear convergence result in [1].

When specialized to the PG method with Bregman distance functions, Bauschke et al. in [2] proved a globally sublinear convergence based on a descent lemma beyond Lipschitz gradient continuity (or equivalently say, the smooth part in the objective function satisfies the relative smoothness). As a complement, with the help of the Bregman distance growth condition, Theorem 1 gives the globally linear convergence for the PG method with Bregman distance functions.

5 Linear convergence under Hölderian growth condition

The convergence results developed in the last section also apply to the PIAG method since the PLIAG method includes the PIAG method as a special case. However, we find that when analyzing directly the convergence of the PIAG method, we can show that it converges linearly with a slightly improved rate under the Hölderian growth condition (17). To this end, we need the following lemma, which can be viewed as a generalization of Lemma 1, due to the fact that Lemma 3 reduces to Lemma 1 when \( \theta = 1 \).

Lemma 3. Assume that the nonnegative sequences \( \{V_k\} \) and \( \{w_k\} \) satisfy
  \[
  dV_{k+1}^\theta + aV_{k+1} \leq aV_k - bw_k + c \sum_{j=k-k_0}^k w_j, \quad \forall k \geq 0, \tag{19}
  \]
for some real numbers $a \in (0, 1)$, $b, c \geq 0$, $d \geq 0$, $a+d = 1$, $\theta \in (0, 1]$, $V_0 \leq 1$, and some nonnegative integer $k_0$. Assume also that $w_k = 0$ for $k < 0$ and condition (11) holds. Then, $V_k \leq V_0^\frac{1}{\rho} a^\frac{k}{\rho}$ for all $k \geq 0$, where $\rho = (1-a)\theta + a$.

The following is the main result of this section.

**Theorem 2.** Suppose that assumptions A0, A1, A2, and A3b hold. Assume that $d(x_0, X) \leq 1$. If the step-size $\alpha$ satisfies:

$$0 < \alpha \leq \alpha_0 := \left(1 + \frac{\mu}{(\tau+1)L}\right)^{\frac{1}{\tau+1}} - 1,$$

then the PIAG method converges linearly in the sense that

$$d^2(x_k, X) \leq \left(1 - \frac{\alpha\mu}{1+\alpha\mu}\right)^k d^2(x_0, X), \quad \forall k \geq 0,$$

where $\rho = \frac{\alpha\mu}{1+\alpha\mu} + \frac{1}{1+\alpha\mu}$. Moreover, if $\alpha = \alpha_0$, then

$$d^2(x_k, X) \leq \left(1 - \frac{1}{(\tau+1)\rho_0 + (\tau+1)^2\rho_0Q}\right)^k d^2(x_0, X), \quad \forall k \geq 0,$$

where $Q = \frac{L}{\mu}$ and $\rho_0 = \frac{\alpha\mu}{1+\alpha\mu} + \frac{1}{1+\alpha\mu}$.

We remark that as for the PG method in [13], Theorem 2 shows that for the PIAG method, the linear convergence rate also depends on $\theta$. When $0 < \theta < 1$, we have slightly improved the convergence rate of the PIAG method from $\frac{1}{1+\mu}$ to $\left(\frac{1}{1+\mu}\right)^\frac{1}{\rho}$. However, it remains unclear whether one can get improved rates in function values under the Hölderian growth condition.

### 6 Discussions

In this paper, we propose a unified algorithmic framework for minimizing the sum of a large number of smooth convex component functions and a possibly non-smooth convex regularization function, with an additional abstract feasible set. Our proposed algorithm includes the IAG method, the IAP method, the PG method, and the PIAG method as special cases. By introducing the Bregman distance growth condition and using the so-called relative smoothness, we derive a group of linear convergence properties for the unified algorithm. The key idea behind is to construct certain Lyapunov functions by embedding the Bregman distance growth condition into a descent-type lemma. Our theory can recover existing rates of linear convergence for the PIAG method under strictly weaker assumptions. Moreover, we analyze the convergence of the PIAG method under the Hölderian growth condition, and obtain improved rate of linear convergence.

We believe that the method developed in this paper will find more applications in other types of incremental methods, including randomized and accelerated versions of the PIAG method. In fact, some colleagues of the first author have shown that our method can be modified to analyze inertial PIAG methods [25]. However, it should not be trivial to extend our method to analyze the
corresponding inertial version of the PLIAG method:

\[ z_{k+1} = \arg\min_{x \in \mathbb{R}^d} \left\{ h(x) + \sum_{i \in I_k} f_i(x) + \langle \sum_{j \in J_k} \nabla f_j(x_{k-\tau_j}), x \rangle + \frac{1}{\alpha_k} D_w(x, x_k) \right\} , \]

\[ x_{k+1} = z_{k+1} + \eta_k \cdot \left( z_{k+1} - z_k \right) , \]

where \( \eta_k \) are positive inertial parameters. The potential difficulty is that some inequalities for the Euclidean distance fail to hold for the Bregman distance. As one of the future studies, we will first study the convergence of the inertial version of the PG method with Bregman distance functions and then that of the inertial version of our proposed algorithm.

The last but not least avenue of future research is to study the convergence of the PLIAG method under assumptions A1’, A2’, and A4’. In other words, we wonder how the PLIAG method behaves when the Bregman distance growth condition fails to hold. It seems infeasible to extend the methods presented in [2,9,11] to analyze the convergence because for the PLIAG method, the sequence \( \{ \Phi(x_k) \}_{k \in \mathbb{N}} \) may not be monotonically decreasing due to the existence of the term \( \triangle_k \) in Lemma \[2\].

Acknowledgements

The work of H. Zhang is supported by the National Science Foundation of China (No.11501569 and No. 61601488 ). The work of Y.H. Dai is partially supported by the Key Project of Chinese National Programs for Fundamental Research and Development Grant 2015CB856002, and NSFC Grants(No.11631013, No. 11331012, and No. 71331001). The work of L. Guo is supported by the National Science Foundation of China (No.11401379, No.11771287).

References

[1] A. Aytekin, H. R. Feyzmahdavian, and M. Johansson. Analysis and implementation of an asynchronous optimization algorithm for the parameter server. CoRR, abs/1610.05507, 2016.

[2] H. H. Bauschke, J. Bolte, and M. Teboulle. A descent lemma beyond lipschitz gradient continuity: First-order methods revisited and applications. Mathematics of Operations Research, 42(2):330-348, 2016.

[3] A. Beck and S. Shtern. Linearly convergent away-step conditional gradient for non-strongly convex functions. to appear in Math. Program., Ser. A, 164(4):1-27, 2017.

[4] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imag. Sci., 17(4):183–202, 2009.

[5] D. P. Bertsekas. Incremental gradient, subgradient, and proximal methods for convex optimization: a survey. Optimization for Machine Learning, 2010:1-38, 2011.

[6] D. P. Bertsekas. Incremental proximal methods for large scale convex optimization. Mathematical Programming, 129(2):163-195, 2011.

[7] D. P. Bertsekas. Incremental aggregated proximal and augmented lagrangian algorithms. arXiv:1509.09257, 2015.
[8] J. Bolte, T. P. Nguyen, J. Peypouquet, and B. W. Suter. From error bounds to the complexity of first-order descent methods for convex functions. Math. Program., Ser. A, 165(2):471-507, 2017.

[9] J. Bolte, S. Sabach, M. Teboulle, and Y. Vaisbourd. First Order Methods beyond Convexity and Lipschitz Gradient Continuity with Applications to Quadratic Inverse Problems. arXiv:1706.06461v1 [math.OC] 20 Jun 2017.

[10] J. M. Borwein and D. J. Vanderwerff. Convex functions of Legendre type in general Banach spaces. Journal of Convex Analysis, 8(2): 569-581, 2001.

[11] G. Chen and M. Teboulle. Convergence analysis of a proximal-like minimization algorithm using bregman functions. SIAM J. Optim., (3):538–543, 1993.

[12] D. Drusvyatskiy and A. S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. arXiv:1602.06661v1 [math.OC] 22 Feb 2016.

[13] G. Garrigos, L. Rosasco, and S. Villa. Convergence of the forward-backward algorithm: beyond the worst-case with the help of geometry. arXiv:1703.09477v2 [math.OC] 29 Mar 2017.

[14] M. Gurbuzbalaban, A. Ozdaglar, and P. Parrilo. On the convergence rate of incremental aggregated gradient algorithms. SIAM J. Optim. 27(2):1035-1048, 2017.

[15] H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of proximal-gradient methods under the polyak-losajesiewicz condition. arXiv:1608.04636v1 [cs.LG] 16 Aug 2016.

[16] M. J. Lai and W. Yin. Augmented ℓ_1 and nuclear-norm models with a globally linearly convergent algorithm. SIAM. J. Imaging Sci. 6(2):1059-1091, 2013.

[17] G. Li and T. K. Pong. Calculus of the exponent of kurdyka-losajesiewicz inequality and its applications to linear convergence of first-order methods. Found Comput Math, https://doi.org/10.1007/s10208-017-9366-8, 2017.

[18] H. Lu, R. M. Freund, and Y. N. Nesterov. Relatively-smooth convex optimization by first-order methods, and applications. accepted by SIAM Journal on Optimization, arXiv:1610.05708v1 [math.OC] 18 Oct 2016.

[19] I. Necoara, Y. Nesterov, and F. Glineur. Linear convergence of first order methods for non-strongly convex optimization. arXiv:1504.06298v2 [math.OC] 23 Apr 2015.

[20] R. T. Rockafellar. Convex Analysis. Princeton University Press, Princeton, NJ, 1970.

[21] N. Vanli, M. Gurbuzbalaban, and A. Ozdaglar. Global convergence rate of proximal incremental aggregated gradient methods. CoRR, abs/1608.01713, 2016.

[22] N. Vanli, M. Gurbuzbalaban, and A. Ozdaglar. A stronger convergence result on the proximal incremental aggregated gradient method. arXiv:1611.08022v1 [math.OC] 23 Nov 2016.

[23] H. Zhang. The restricted strong convexity revisited: analysis of equivalence to error bound and quadratic growth. Optimization Letter, 11:817-833, 2017.

[24] H. Zhang. New analysis of linear convergence of gradient-type methods via unifying error bound conditions. arXiv:1606.00269v6 [math.OC], 2017.

[25] X. Y. Zhang, W. Peng, H. Zhang, and L.Z. Cheng. Inertial proximal incremental aggregated gradient methods. In preparation, 2017.

[26] Z. Zhou and M. C. So. A unified approach to error bounds for structured convex optimization problems. Math. Program., Ser. A, 165:689-728, 2017.
A  Examples of quadratic growth condition

We introduce two examples whose objective function fails to be strongly convex but satisfies the quadratic growth condition. Then our results show that the PIAG method is a suitable algorithm for solving them; see the discussions after Theorem 1. More examples can be found in [12].

Example 1. [8, Lemma 10] Consider the least squares objective with \( \ell^1 \)-norm regularization problem:

\[
\min_{x \in \mathbb{R}^d} \Phi(x) := \frac{1}{2} \|Ax - b\|^2 + \lambda \|x\|_1,
\]

where \( \lambda > 0 \) is some regularization parameter, \( A \) and \( b \) are given appropriate matrix and vector. Let \( R > \frac{\|b\|^2}{2\lambda} \). Then, there exists a constant \( \mu > 0 \) such that

\[
\Phi(x) - \Phi^* \geq \mu d^2(x, \mathcal{X}) \quad \text{for all } x \in \mathbb{R}^d \text{ such that } \|x\|_1 \leq R. \quad (23)
\]

Let \( w(x) = \frac{1}{2} \|x\|^2, \ Q = \mathbb{R}^d, \ \Omega = \{x \in \mathbb{R}^d : \|x\|_1 \leq R\}, \ F(x) = \frac{1}{2} \|Ax - b\|^2, \) and \( h(x) = \|x\|_1 + \delta_\Omega(x) \). In this setting, the quadratic growth condition is actually the condition (23).

Example 2. The following is a dual model appeared in compressed sensing

\[
\min_{x \in \mathbb{R}^d} F(x) := -b^T x + \frac{\alpha}{2} \|\text{shrink}_\mu(A^T x)\|^2.
\]

Here the parameter \( \alpha, \) matrix \( A, \) and vector \( b \) are given, and the operator shrink is defined by

\[
\text{shrink}_\mu(s) = \text{sign}(s) \max\{|s| - \mu, 0\},
\]

where \( \text{sign}(\cdot), |\cdot|, \) and \( \max\{\cdot, \cdot\} \) are component-wise operations if \( s \) is a vector. Note that the smooth part \( F \) in the problem above can be written into the sum of many component functions.

The objective function \( F \) was proved in [16] to be restricted strongly convex on \( \mathbb{R}^d, \) and hence satisfies the quadratic growth condition with \( Q = \mathbb{R}^d \) due to their equivalence [23, 24].

B  Conditions guaranteeing Bregman distance growth

Below, we list two cases, where the Bregman distance growth condition holds.

C1. \( \Phi \) satisfies the quadratic growth condition and \( w \) has a Lipschitz continuous gradient over \( \text{int dom} \ w. \)

C2. \( h \) is convex, \( \text{rint dom} h \cap \text{rint dom} w \neq \emptyset, \) and \( F \) satisfies

\[
F(y) \geq F(x) + \langle \nabla F(x), y - x \rangle + \mu \cdot D_w(x, y), \quad \forall x, y \in \text{int dom} w. \quad (24)
\]

Assume that C1 holds. From the quadratic growth condition of \( \Phi, \) it follows that there exists \( \mu > 0 \) such that

\[
\Phi(y) - \Phi^* \geq \frac{\mu}{2} d^2(y, \mathcal{X}) = \frac{\mu}{2} \|y - y'\|^2 \quad \forall y \in \text{dom} w,
\]

where \( y' \) stands for the projection of \( y \) onto \( \mathcal{X}. \) Since \( w \) has a Lipschitz continuous gradient, it is easy to verify that there exists \( L > 0 \) such that

\[
\frac{\mu}{2} \|y - y'\|^2 \geq \frac{\mu}{L} D_w(y', y), \quad \forall y, y' \in \text{int dom} w.
\]
Therefore, the Bregman distance growth condition follows from the last two inequalities immediately.

Assume that C2 holds. Together with the convexity of \( h \), we have that for any \( v \in \partial h(x) \),
\[
F(y) + h(y) \geq F(x) + h(x) + \langle \nabla F(x) + v, y - x \rangle + \mu \cdot D_w(x, y), \quad \forall x, y \in \text{int dom} w.
\] (25)

Take an optimal solution \( \bar{x} \in \mathcal{X} \). Using Fermat’s rule and the condition \( \text{rint dom} h \cap \text{rint dom} w \neq \emptyset \), it follows that there exists \( \bar{v} \in \partial h(\bar{x}) \) such that
\[
\langle \nabla F(\bar{x}) + \bar{v}, y - \bar{x} \rangle \geq 0, \quad \forall y \in \text{dom} w.
\]
Then using (25) with \( x = \bar{x} \) and \( v = \bar{v} \) and recalling that \( F(x) + h(x) = \Phi^* \), we thus have
\[
\Phi(y) - \Phi^* \geq \mu \cdot D_w(x, y) \geq \mu \cdot \inf_{z \in \mathcal{X}} D_w(z, y), \quad \forall y \in \text{int dom} w
\]
which is just the Bregman distance growth condition.

Note that condition (24) is different from the \( \mu \)-strong convexity relative to \( w \) introduced in [15]
\[
F(y) \geq F(x) + \langle \nabla F(x), y - x \rangle + \mu \cdot D_w(y, x), \quad \forall x, y \in \text{int dom} w.
\]
These two conditions are different due to the nonsymmetry of the Bregman distance. To link these two conditions, we need a measure for the lack of symmetry in \( D_w \), which was introduced by Bauschke et al. in [2].

**Definition 1.** Given a Legendre function \( w : \mathbb{R}^d \to (-\infty, +\infty) \), its symmetry coefficient is defined by
\[
\alpha(w) := \inf \left\{ \frac{D_w(x, y)}{D_w(y, x)} : x, y \in \text{int dom} w, x \neq y \right\} \in [0, 1].
\]
If \( \alpha(w) \neq 0 \), then by the definition,
\[
\alpha(w) D_w(y, x) \leq D_w(x, y) \leq \alpha(w)^{-1} D_w(y, x), \quad \forall x, y \in \text{int dom} w,
\]
which implies that condition (24) and the relatively strong convexity are equivalent up to a constant.

**C Proof of Proposition 1**

First we observe that the objective function \( \Phi_k \) in [10] is strictly convex. Thus this problem may have at most one minimizer. Let \( x_j \in \text{int dom} w \) for any \( j \leq k \) and \( x \in \text{int dom} w \). Then by the modified assumptions A1’ and A4’, we derive that
\[
f_j(x) \leq f_j(x_{k-j}) + \langle \nabla f_j(x_{k-j}), x - x_{k-j} \rangle + L_j \cdot D_w(x, x_{k-j}) \leq f_j(x_{k-j}) + \langle \nabla f_j(x_{k-j}), x - x_{k-j} \rangle + L_j \cdot D_w(x, x_{k}) + L_j \cdot D_w(x_{k}, x_{k-j}). \] (26)

Noting that \( \alpha_k \leq \frac{1}{\sum_{j \in J_k} L_j} \), we can further get
\[
\sum_{j \in J_k} f_j(x) \leq C_k + \left( \sum_{j \in J_k} \nabla f_j(x_{k-j}), x \right) + \sum_{j \in J_k} L_j \cdot D_w(x, x_k) \leq C_k + \left( \sum_{j \in J_k} \nabla f_j(x_{k-j}), x \right) + \frac{1}{\alpha_k} \cdot D_w(x, x_k),
\]
where
\[ C_k := \sum_{j \in J_k} f_j(x_k - \tau_k) - \sum_{j \in J_k} \nabla f_j(x_k - \tau_k), x_{k-1} - x_k + \sum_{j \in J_k} L_j \cdot D_w(x_k, x_{k-1}). \]

This above inequality leads to the following relationship between \( \Phi \) and \( \Phi_k \):
\[ \Phi(x) \leq \Phi_k(x) + C_k, \quad x \in \text{int dom } w. \quad (27) \]

Due to the nonemptiness and compactness of the optimal solution set \( \mathcal{X} \), we can have that \( \Phi + \delta_Q \) is level-bounded; see, e.g., [20] Corollary 8.7.1. Then by (27) and the fact that \( \text{dom } w \subseteq Q \), we can have that \( \Phi_k \) is also level-bounded. Hence the optimal solution set of minimizing \( \Phi_k \) is nonempty. So far we have shown that the objective (16) has only one solution, say \( x_{k+1} \). Next we show that \( x_{k+1} \in \text{int dom } w \). This can be seen from the optimality condition of \( x_{k+1} \) and the fact that \( \partial w(z) = \emptyset \) for any \( z \notin \text{int dom } w \) (see [20] Theorem 26.1).

D Proof of Lemma [2]

Since each component function \( f_j(x) \) is convex and \( L_j \)-smooth relative to \( w \), we derive that
\[ f_j(x_{k+1}) \leq f_j(x_{k-\tau_k}) + \langle \nabla f_j(x_{k-\tau_k}), x_{k+1} - x_{k-\tau_k} \rangle + L_j \cdot D_w(x_{k+1}, x_{k-\tau_k}) \]
\[ \leq f_j(x) + \langle \nabla f_j(x_{k-\tau_k}), x_{k+1} - x \rangle + L_j \cdot D_w(x_{k+1}, x_{k-\tau_k}), \quad (28) \]
where the second inequality follows from the convexity of \( f_j(x) \). For simplicity, we denote \( s_k := \sum_{j \in J_k} \nabla f_j(x_{k-\tau_k}) \).

Sum (28) over all \( j \in J_k \) and use the expression of \( s_k \) to yield
\[ \sum_{j \in J_k} f_j(x_{k+1}) \leq \sum_{j \in J_k} f_j(x) + \langle s_k, x_{k+1} - x \rangle + \sum_{j \in J_k} L_j \cdot D_w(x_{k+1}, x_{k-\tau_k}). \quad (29) \]

On the other hand, by the optimality condition, \( x_{k+1} \) satisfies the inclusion relationship
\[ -s_k - \frac{1}{\alpha} (\nabla w(x_{k+1}) - \nabla w(x_k)) \in \partial h(x_{k+1}) + \sum_{i \in I_k} \nabla f_i(x_{k+1}). \quad (30) \]

Substituting (30) in the subgradient inequality for the convex function \( h(x) + \sum_{i \in I_k} \nabla f_i(x) \), we derive that
\[ h(x_{k+1}) + \sum_{i \in I_k} \nabla f_i(x_{k+1}) \leq h(x) + \sum_{i \in I_k} \nabla f_i(x) + \langle s_k + \frac{1}{\alpha} (\nabla w(x_{k+1}) - \nabla w(x_k)), x - x_{k+1} \rangle \]
\[ = h(x) + \sum_{i \in I_k} \nabla f_i(x) + \langle s_k, x - x_{k+1} \rangle + \frac{1}{\alpha} (\nabla w(x_{k+1}) - \nabla w(x_k), x - x_{k+1}) \]
\[ = h(x) + \sum_{i \in I_k} \nabla f_i(x) + \langle s_k, x - x_{k+1} \rangle + \frac{1}{\alpha} D_w(x, x_k) - \frac{1}{\alpha} D_w(x_{k+1}, x_k). \quad (31) \]
where the last equality follows from the three points identity of Bregman distance $[11]$, i.e.,

$$D_w(x, z) - D_w(x, y) - D_w(y, z) = \langle \nabla w(y) - \nabla w(z), x - y \rangle, \quad \forall y, z \in \text{int dom} w, x \in \text{dom} w.$$

Adding (31) to (29), we obtain

$$\Phi(x_{k+1}) \leq \Phi(x) + \frac{1}{\alpha} D_w(x, x_k) - \frac{1}{\alpha} D_w(x, x_{k+1}) - \frac{1}{\alpha} D_w(x_{k+1}, x_k) + \sum_{j \in J_k} L_j \cdot D_w(x_{k+1}, x_{k-\tau_k}).$$

(32)

Recalling that $\tau^n_k$ is bounded above by $\tau$ and using the assumption A4', we can have

$$D_w(x_{k+1}, x_{k-\tau^n_k}) \leq \ell(\tau_k^n + 1) \sum_{j = k - \tau_k^n}^{k} D_w(x_{j+1}, x_j) \leq \ell(\tau + 1) \sum_{j = k - \tau}^{k} D_w(x_{j+1}, x_j).$$

This implies

$$\sum_{j \in J_k} L_j \cdot D_w(x_{k+1}, x_{k-\tau_k}) \leq \ell(\tau + 1) \sum_{j \in J_k} \sum_{j = k - \tau}^{k} D_w(x_{j+1}, x_j) \leq L \cdot \ell(\tau + 1) \sum_{j = k - \tau}^{k} D_w(x_{j+1}, x_j).$$

This together with (32) leads to the desired result.

E Proof of Theorem 1

First we show that the whole sequence $\{x^k\}$ is well-defined and moreover, it belongs to $\text{int dom} w$. Using the definition of $L$ and the Bernoulli inequality, i.e., $(1 + x)^r \leq 1 + rx$ for any $x \geq -1$ and $r \in [0, 1]$, we have

$$\alpha_0 \leq \frac{1}{\ell(\tau + 1) \cdot (\tau + 1) \cdot L} \leq \frac{1}{\sum_{j \in J_k} L_j}, \quad \forall k \geq 0.$$

Then the desired result follows from Proposition 1 and the choice of the step-size (12) immediately.

We next show the convergence. For simplicity, we denote

$$Y_k := \arg \min_{x \in \mathcal{X}} D_w(x, x_k).$$

Recall that $\mathcal{X}$ is assumed to be nonempty and compact, and note that $D_w(x, x_k)$ is a lsc function. By Weierstrass’ theorem, we have that $Y_k$ is nonempty and compact. Pick $\tilde{x}_k \in Y_k$ and invoke Lemma 2 at $x = \tilde{x}_k$ to yield

$$\Phi(x_{k+1}) \leq \Phi^* + \frac{1}{\alpha} D_w(\tilde{x}_k, x_k) - \frac{1}{\alpha} D_w(\tilde{x}_k, x_{k+1}) - \frac{1}{\alpha} D_w(x_{k+1}, x_k) + \Delta_k.$$

(33)

Since $\tilde{x}_k \in Y_k \subset \mathcal{X}$, it holds that

$$\inf_{x \in \mathcal{X}} D_w(x, x_{k+1}) \leq D_w(\tilde{x}_k, x_{k+1}).$$

(34)
Using the Bregman distance growth condition \( \| \), we have
\[
D_w(\tilde{x}_k, x_k) = \inf_{x \in \mathcal{X}} D_w(x, x_k) \leq \frac{1}{\mu} (\Phi(x_k) - \Phi^*),
\]
and hence
\[
D_w(\tilde{x}_k, x_k) \leq p \cdot D_w(\tilde{x}_k, x_k) + \frac{q}{\mu} (\Phi(x_k) - \Phi^*),
\]
for any \( p, q > 0 \) with \( p + q = 1 \). Pick
\[
q = \frac{\alpha \mu}{1 + \alpha \mu}, \, p = \frac{1}{1 + \alpha \mu}.
\]
Then using (33), (34), and (35), we derive that
\[
\Phi(x_{k+1}) - \Phi^* + \frac{1}{\alpha} \cdot \inf_{x \in \mathcal{X}} D_w(x, x_{k+1}) \leq \frac{q}{\alpha \mu} \left( \Phi(x_k) - \Phi^* + \frac{1}{\alpha} \cdot \inf_{x \in \mathcal{X}} D_w(x, x_k) \right) - \frac{1}{\alpha} D_w(x_{k+1}, x_k) + \Delta_k,
\]
where the last equality follows from the fact that \( q = \alpha \mu p \). In terms of the expressions of \( \Gamma_\alpha(x) \) and \( \Delta_k \), we have
\[
\Gamma_\alpha(x_{k+1}) \leq \frac{1}{1 + \alpha \mu} \Gamma_\alpha(x_k) - \frac{1}{\alpha} D_w(x_{k+1}, x_k) + L \cdot \ell(\tau + 1) \cdot \sum_{j=k-\tau}^{k} D_w(x_{j+1}, x_j).
\]
In order to apply Lemma \( \| \) let \( V_k = \Gamma_\alpha(x_k) \), \( w_k = D_w(x_{k+1}, x_k) \), \( a = \frac{1}{1 + \alpha \mu} \), \( b = \frac{1}{\alpha} \), \( c = L \ell(\tau + 1) \), and \( k_0 = \tau \). When the step-size \( \alpha \) satisfies the condition \( \| \), it is not hard to verify that the condition \( \| \) holds. Then by Lemma \( \) the desired result \( \| \) follows immediately. The result \( \| \) is a direct consequence of \( \| \), and the result \( \| \) follows from the Bregman distance growth condition and \( \| \).

It remains to show \( \| \). Taking the certain value \( \alpha = \alpha_0 \) in \( \| \), \( Q = \frac{L}{\mu} \), and noting that
\[
1 - \frac{\alpha_0 \mu}{1 + \alpha_0 \mu} = \frac{1}{1 + \alpha_0 \mu} = \left( 1 + \frac{1}{\ell(\tau + 1)Q} \right)^{-\frac{1}{\tau + 1}}
\]
and
\[
\left( 1 + \frac{1}{\ell(\tau + 1)Q} \right)^{-1} = 1 - \frac{1}{1 + \ell(1 + \tau)Q},
\]
we have that
\[
\Gamma_{\alpha_0}(x_k) \leq \left( 1 + \frac{1}{\ell(\tau + 1)Q} \right)^{\frac{1}{\tau + 1}} \Gamma_{\alpha_0}(x_0)
\]
\[
= \left( 1 - \frac{1}{1 + \ell(\tau + 1)Q} \right)^{\frac{1}{\tau + 1}} \Gamma_{\alpha_0}(x_0)
\]
\[
\leq \left( 1 - \frac{1}{1 + \ell(\tau + 1)(\tau + 1)Q} \right)^k \Gamma_{\alpha_0}(x_0),
\]
where the last inequality follows from the Bernoulli inequality. The proof is completed.
F Proof of Lemma 3

First we observe that the following inequality holds:

\[
\sum_{k=0}^{K} \frac{1}{a^{k+1}} \sum_{j=k-k_0}^{k} w_j = \frac{1}{a} (w_{-k_0} + w_{-k_0+1} + \ldots + w_0) \\
+ \frac{1}{a^2} (w_{-k_0+1} + w_{-k_0+2} + \ldots + w_1) + \ldots \\
+ \frac{1}{a^{k_0+1}} (w_0 + w_1 + \ldots + w_{k_0}) + \ldots \\
+ \frac{1}{a^{K+1}} (w_{K-k_0} + w_{K-k_0} + \ldots + w_K) \\
\leq (1 + \frac{1}{a} + \ldots + \frac{1}{a^{k_0}}) \sum_{k=0}^{K} \frac{w_k}{a^{k+1}} \\
= \frac{1}{1-a} \left(1 - \frac{1-a^{k_0+1}}{a^{k_0}}\right) \sum_{k=0}^{K} \frac{w_k}{a^{k+1}}.
\]

Then divide both sides of (19) by \(a^{k+1}\) and take the sum from \(k = 1\) to \(k = K\) to yield

\[
\sum_{k=0}^{K} \frac{dV_{k+1}^\theta}{a^{k+1}} + aV_{k+1} \leq \left(\frac{c}{1-a} \frac{1-a^{k_0+1}}{a^{k_0}} - b\right) \sum_{k=0}^{K} \frac{w_k}{a^{k+1}} + \sum_{k=0}^{K} V_k, \quad \forall K \geq 0.
\]

Taking (11) into account, we then have

\[
\sum_{k=0}^{K} \frac{dV_{k+1}^\theta}{a^{k+1}} + aV_{k+1} \leq \sum_{k=0}^{K} \frac{V_k}{a^k}, \quad \forall K \geq 0.
\]

(37)

Using the convexity of \(V_{k+1}^x\) with respect to \(x\) and noting that \(d\theta + a = \rho\) and \(a + d = 1\), we have

\[
V_{k+1}^\rho = V_{k+1}^{d\theta+a} \leq dV_{k+1}^\theta + aV_{k+1}.
\]

This together with (37) implies

\[
\sum_{k=0}^{K} \frac{V_{k+1}^\rho}{a^{k+1}} \leq \sum_{k=0}^{K} \frac{V_k}{a^k}, \quad \forall K \geq 0.
\]

(38)

Now, we prove by induction that \(V_k \leq 1\) for all \(k \geq 0\). First we note that \(V_0 \leq 1\). We next show that \(V_{k_0+1} \leq 1\) by assuming that \(V_k \leq 1\) for \(k \leq k_0\). It is clear that \(V_k^\rho \geq V_k\) for all \(k \leq k_0\) since \(\rho \in (0,1]\). Then it follows from (38) and the fact that \(a^0 = 1\) that

\[
\frac{V_{k_0+1}^\rho}{a^{k_0+1}} + \sum_{k=1}^{k_0} \frac{V_k}{a^k} \leq \sum_{k=0}^{k_0} \frac{V_{k+1}^\rho}{a^{k+1}} \leq \sum_{k=1}^{k_0} \frac{V_k}{a^k} + V_0.
\]

(39)

Thus we have \(V_{k_0+1}^\rho \leq a^{k_0+1}V_0\). Since \(a < 1\) and \(V_0 \leq 1\), it follows immediately that \(V_{k_0+1} \leq 1\) and hence \(V_{k_0+1} \leq 1\). In conclusion, we have shown that \(V_k \leq 1\) for all \(k \geq 0\). Thus, it is easy to see that (39) holds when \(k_0\) is replaced by any nonnegative integer. We then obtain that \(V_k^\rho \leq a^kV_0\) for all \(k \geq 0\). The desired result follows immediately.
G Proof of Theorem 2

Let \( Q = \mathbb{R}^d \) and \( w(x) = \frac{1}{2} \| x \|^2 \). In this case, the assumptions A1’, A2’ and A4’ hold with \( \ell(\cdot) \) being the identity function. By Lemma 2, it follows that

\[
\Phi(x_{k+1}) \leq \Phi(x) + \frac{1}{2\alpha} \| x - x_k \|^2 - \frac{1}{2\alpha} \| x - x_{k+1} \|^2 - \frac{1}{2\alpha} \| x_{k+1} - x_k \|^2 + \Delta_k^1,
\]  

where \( \Delta_k^1 = \frac{(\tau+1)L}{2} \sum_{j=k-\tau}^k \| x_{j+1} - x_j \|^2 \). The Hölderian growth condition (7) at \( x = x_{k+1} \) reads as

\[
\Phi(x_{k+1}) - \Phi^* \geq \frac{\mu}{2} d^2(x_{k+1}, X). \tag{41}
\]  

Let \( x'_k \) be the unique projection point of \( x_k \) onto \( X \). Using (40) with \( x = x'_k \) and (41), we obtain

\[
dV_{k+1}^\theta + aV_{k+1} \leq aV_k - bw_k + c \sum_{j=k-\tau}^k w_k. \tag{43}
\]  

It is not difficult to verify that

\[
a + d = 1, a \in (0, 1), b, c, d > 0, V_0 \leq 1.
\]  

Moreover, when the step-size \( \alpha \) satisfies (12), condition (11) holds. Then, we apply Lemma 3 to conclude that

\[
d^2(x_k, X) \leq d^2(x_0, X) a^k = d^2(x_0, X) \left( \frac{1}{1 + \alpha \mu} \right)^k \quad \forall k \geq 0,
\]  

where

\[
\rho = (1 - a)\theta + a = (1 - \frac{1}{2\alpha \delta})\theta + \frac{1}{2\alpha \delta} = \frac{\alpha \mu}{1 + \alpha \mu} \theta + \frac{1}{1 + \alpha \mu}.
\]  

When we take \( \alpha = \alpha_0 \), the desired result (22) can be derived by using the same argument for obtaining (36). The proof is completed.