Linear Convergence of Proximal Incremental Aggregated Gradient Methods under Quadratic Growth Condition

Hui Zhang*

March 31, 2017

Abstract

Under the strongly convex assumption, several recent works studied the global linear convergence rate of the proximal incremental aggregated gradient (PIAG) method for minimizing the sum of a large number of smooth component functions and a non-smooth convex function. In this paper, under *the quadratic growth condition*–a strictly weaker condition than the strongly convex assumption, we derive a new convergence result, which implies that the PIAG method attains global linear convergence rates in both the function value and iterate point errors. Moreover, by using the relative smoothness (recently proposed to weaken the traditional gradient Lipschitz continuity) and defining the Bregman distance growth condition (that generalizes the quadratic growth condition), we further analyze the PIAG method with general distance functions. Finally, we propose a new variant of the PIAG method with improved linear convergence rates.

Our theory recovers many very recent results under strictly weaker assumptions, but also provides new results for both PIAG methods and the proximal gradient method. Besides, if the strongly convex assumption indeed holds, then our theory shows that one can improve the corresponding rates derived under the quadratic growth condition. The key idea behind our theory is to construct certain Lyapunov functions.

Keywords. linear convergence, strong convexity, quadratic growth condition, incremental aggregated gradient, Lyapunov function, Bregman distance.

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

1 Introduction

A fundamental generic optimization model arises in many problems in machine learning, signal processing, image science, communication systems, and distributed optimization etc. This model consists in minimizing the sum of a differentiable function $F(x)$ and a possibly non-smooth regularization function $h(x)$:

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\, \Phi(x) \triangleq F(x) + h(x). \tag{1}$$

The well-known method to solve this model is the forward-backward splitting (FBS) scheme, which is often called the proximal gradient method in the community of convex optimization. This method

*College of Science, National University of Defense Technology, Changsha, Hunan, 410073, P.R.China. Corresponding author. Email: h.zhang1984@163.com

consists of the composition of a gradient (forward) step of $F(x)$ with a proximal (backward) step on $h(x)$. It can be expressed as follows:

$$y_k = x_k - \alpha \cdot \nabla F(x_k), \tag{2}$$

$$x_{k+1} = \arg\min_x \{h(x) + \frac{1}{2\alpha}\|x - y_k\|^2\}, \tag{3}$$

where $\alpha > 0$ is some step size. The remarkable merit of this method lies in exploiting the smooth plus smooth structure of model (1). However, if the differentiable function $F(x)$ is a sum of N component functions $f_n(x)$, i.e.,

$$F(x) = \sum_{n=1}^{N} f_n(x) \tag{4}$$

with large N (this indeed appears in many applications), then evaluating the full gradient of $F(x)$, i.e., $\nabla F(x) = \sum_{n=1}^{N} \nabla f_n(x)$ is costly and even prohibitive. Hence, a natural idea to overcome this difficulty is to modify the standard FBS method by utilizing the additional structure (4). Following this line of thought, the proximal incremental aggregated gradient (PIAG) method was proposed and studied in several recent papers [14, 1, 15]. At each iteration $k \geq 0$, the PIAG method first constructs a vector that aggregates the gradients of all components functions, possibly evaluated at the $k - \tau_k^n$ iteration,

$$g_k = \sum_{n=1}^{N} \nabla f_n(x_{k-\tau_k^n})$$

where τ_k^n are some nonnegative integers. This vector is used to exploit the additive structure (4) of the N component functions, but also to approximate the full gradient of $F(x)$ at the current iteration. In fact, if $\tau_k^n \equiv 0$, then $g_k = \nabla F(x_k)$. After obtaining g_k, the PIAG method then performs a proximal step on the sum of the non-smooth term $h(x)$ and a linear term $\langle g_k, x - x_k\rangle$ as follows:

$$x_{k+1} = \arg\min_x \{h(x) + \langle g_k, x - x_k\rangle + \frac{1}{2\alpha}\|x - x_k\|^2\}. \tag{5}$$

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

$$y_k = x_k - \alpha \cdot g_k, \tag{6}$$

$$x_{k+1} = \arg\min_x \{h(x) + \frac{1}{2\alpha}\|x - y_k\|^2\}. \tag{7}$$

One can see that the PIAG method differs from its mother scheme–the FBS method mainly at the gradient step, and reduces to it as $\tau_k^n \equiv 0$.

In this paper, we will focus on global linear convergence of the PIAG method and its variants under weak conditions. Beforehand, we review several very recent works around this topic.

1.1 Related work

Work [14] is the first study that establishes a global linear convergence rate for the PIAG method in function value error, i.e., $\Phi(x_k) - \Phi(x^*)$, where x^* denotes the minimizer point of $\Phi(x)$. Work [1] used a different analysis and showed a global linear convergence rate in iterate point error, i.e., $\|x_k - x^*\|$. The authors of [14] combined the results presented in [14] and [1] and provided a stronger

linear convergence rate for the PIAG method in the recent paper [15]. However, all of the work mentioned above are built on the strongly convex assumption, which is actually not satisfied by many application problems and hence motives lots of research to find weaker alternatives. Influential weaker conditions include the error bound property, the restricted strongly convex property, the quadratic growth condition, and the Polyak-Łojasiewicz inequality; interested readers could refer to [16, 17, 13, 10, 8, 6, 18] for more information. Works [16, 8] studied the linear convergence of the proximal gradient method under these weaker conditions. But to our knowledge, there is no work of studying the global linear convergence of the PIAG method under these weaker conditions.

On the other hand, the recent work [2] and work [12] independently introduced a new notion (it was called Lipschitz-like/convexity condition in [2] and relative smoothness in [12]) to weaken the gradient Lipschitz continuity, which is a central property required in the analysis of gradient methods. The (sub)linear convergence of the traditional gradient and proximal gradient methods with general distance functions was studied under these weaker notions in [2] and [12]. But the global linear convergence of the PIAG method without Lipschitz gradient continuity has not been investigated until now.

1.2 Main contribution

Our contribution of this paper is four-fold. First, we obtain a global linear convergence result of the PIAG method under the quadratic growth condition, which is strictly weaker than the strongly convex assumption. Second, we develop an extended theory of the PIAG method, which shows that the PIAG method with general distance functions still has global linear convergence under very weak conditions. Third, we propose a new variant of the PIAG method with improved linear convergence rates. At last, we return to the case where the strongly convex assumption indeed holds, and find that an efficient exploitation of strong convexity can help us improve linear convergence rates.

Although we employ an important lemma presented in [1], which was also used in [15], our proof strategy is essentially different from that of [1] and [15]. The key idea behind is to construct certain Lyapunov functions that simultaneously include the function value and iterate point errors. Hence, our linear convergence results simultaneously characterize the global linear convergence rate in both the function value and iterate point errors; see Theorem 1 in Section 3, Theorem 2 in Section 4, Theorem 3 in Section 5, and Theorem 4 in Section 6. This perspective is new even for the (proximal) gradient method.

The global linear convergence rate of the PIAG method derived in this study has a linear dependence on the condition number of the problem and a quadratic dependence on the delay parameter. This is consistent with the results in [1] and [14]. Moreover, if the delay parameter is less than 47, then we recover the result from [15], which has the best (linear) dependence on the condition number and the delay parameter. Although the constraint of the delay parameter being less than 47 is not required in [15], the linear convergence result in [15], which relies on the lemmas in [14], was derived in an essentially more complicated way under strictly stronger assumptions.

When specialized to the traditional gradient and proximal gradient methods with general distance functions, our extended theory provides complementary results to that in [2] and [12]. Our theory in this paper can also be viewed as a further development of linear convergence theory studied in [16, 13, 10, 8].

1.3 Organization

The rest of the paper is organized as follows. In Section 2, we list all assumptions of this study and discuss some connection between them. In Section 3, we state our main convergence result for the PIAG method. In Section 4, we extend the main result to the PIAG method with general distance function and state a generalized theorem. In Section 5, we consider a new variant of the PIAG method with convergence rate analysis. In Section 6, we show that strong convexity can help us improve the convergence rates derived under the quadratic growth conditions. All proofs are given in Section 7. A short conclusion can be found in Section 8.

2 Assumptions

First, we list the following standard assumptions that are used in both [1] and [15].

A1. Each component function $f_n(x)$ is convex with L_n-continuous gradient; that is

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

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

A2. The regularization function $h(x) : \mathbb{R}^d \to (-\infty, \infty]$ is proper, closed, convex and 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 time-varying delays τ_k^n are bounded, i.e., there is a nonnegative integer τ such that

$$\tau_k^n \in \{0, 1, \cdots, \tau\},$$

hold for all $k \geq 1$ and $n \in \{1, 2, \cdots, N\}$. Such τ is called the delay parameter.

A4′. The sum function $F(x)$ is μ-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.

In order to replace the strongly convex assumption, we state the quadratic growth condition.

A4. The objective function $\Phi(x)$ satisfies the quadratic growth condition, meaning there is a real number $\beta > 0$ such that

$$\Phi(x) - \Phi^* \geq \frac{\beta}{2} d^2(x, \mathcal{X}),$$

where $\mathcal{X}$ is the set of minimizers of $\Phi(x)$, which is assumed to be nonempty, Φ^* is the minimal value of $\Phi(x)$, and $d(x, \mathcal{X})$ is the distance function from points to a fixed set, defined by

$$d(x, \mathcal{X}) = \inf_{y \in \mathcal{X}} \|x - y\|.$$

On one hand, the strong convexity with parameter $\mu > 0$ implies that

$$\Phi(x) - \Phi^* \geq \frac{\mu}{2} \|x - x^*\|^2,$$

where x^* is the unique minimizer of $\Phi(x)$, and hence implies the quadratic growth condition. On the other hand, we can easily construct functions to show that the quadratic growth condition

does not imply any strong convexity. For example, the composition $g(Ax)$, where $g(\cdot)$ is a strongly convex and A is rank deficient, satisfies the quadratic growth condition but fails to be strongly convex. Therefore, the quadratic growth condition is *strictly weaker* than the strongly convex condition. The former has been adopted recently as an efficient alternative of the latter to derive linear convergence rate results for many fundamental algorithms, such as the projected gradient method [13], the proximal gradient method [8, 16], the conditional gradient method [3], and the cyclic block coordinate gradient descent method [17]. Due to the close connection of the quadratic growth condition with previously mentioned weaker notions [16, 17, 13, 10, 8, 6, 18], the methods in [18, 11] can be used to verify whether a given function satisfies the quadratic growth condition.

Let $Q \subset \mathbb{R}^d$ be a closed convex set and $w(\cdot) : Q \to \mathbb{R}$ be any given differential convex function, not needed to be strictly or strongly convex. Associated to $w(\cdot)$, the Bregman distance is defined by:

$$D_w(y, x) = w(y) - w(x) - \langle \nabla w(x), y - x \rangle.$$

A well known fact about the Bregman distance is that $D_w(y, x) = \frac{1}{2}\|y - x\|^2$ if we choose $w(x) = \frac{1}{2}\|x\|^2$ and $D_w(y, x) \geq \frac{1}{2}\|y - x\|^2$ it $w(x)$ is a 1-strongly convex function on Q.

In what follows, we introduce a group of additional assumptions relied on the Bregman distance, which will be used for generalized theoretical analysis.

A5. Each component function $f_n(x)$ is convex and L_n-smooth relative to $w(x)$; that is

$$f_n(y) \leq f_n(x) + \langle \nabla f_n(x), y - x \rangle + L_n D_w(y, x), \quad \forall x, y \in \text{int}Q.$$

This assumption, which we will call L_n-relative smoothness in this article, was independently proposed in the recent papers [2] and [12] with the common purpose of relaxing the gradient Lipschitz continuity assumption. It can be viewed an extended or weaken Lipschitz continuity of gradient. Inspired by the idea behind this assumption, we further relax the quadratic growth condition and propose the following assumption:

A6. The objective function $\Phi(x)$ has a faster growth than the Bregman distance; that is there exists a real number $\rho > 0$ such that

$$\Phi(y) - \Phi^* \geq \rho \cdot \min\{\inf_{z \in \mathcal{X}} D_w(y, z), \inf_{z \in \mathcal{X}} D_w(z, y)\}, \quad \forall y \in \text{int}Q, \tag{8}$$

where $\mathcal{X}$ is the set of minimizers of $\Phi(x)$, which is assumed to be nonempty and compact, Φ^* is the minimal value of $\Phi(x)$. We call this property the Bregman distance growth condition.

Note that Bregman distances are in general not symmetric and hence it usually can not hold that $\inf_{z \in \mathcal{X}} D_w(y, z) = \inf_{z \in \mathcal{X}} D_w(z, y)$. Below, we list two conditions that could ensure the Bregman distance growth condition (8) to hold.

C1. The regularization function $h(x)$ is convex, and the smooth part $F(x)$ is ρ-strongly convex relative to $w(x)$; that is

$$F(y) \geq F(x) + \langle \nabla F(x), y - x \rangle + \rho D_w(y, x), \quad \forall x, y \in \text{int}Q.$$

The latter property, accompanied by the relative smoothness, was proposed [12] and called ρ-relatively strong convexity.

In fact, combining the convexity of $h(x)$ and the relatively strong convexity of $F(x)$, for any $v \in \partial h(x)$ we have

$$F(y) + h(y) \geq F(x) + h(x) + \langle \nabla F(x) + v, y - x \rangle + \rho D_w(y, x), \quad \forall x, y \in \text{int}Q.$$

Take $x \in \mathcal{X}$. Use the optimality condition $\langle \nabla F(x) + v, y - x \rangle \geq 0$ and recall that $F(x) + h(x) = \Phi^*$. We thus have

$$\Phi(y) - \Phi^* \geq \rho D_w(y, x) \geq \rho \cdot \min\{\inf_{z \in \mathcal{X}} D_w(y, z), \inf_{z \in \mathcal{X}} D_w(z, y)\},$$

which is just the Bregman distance growth condition (8).

C2. The objective function $\Phi(x)$ satisfies the quadratic growth condition and the gradient of $w(x)$ is L-Lipschitz continuous.

Using this condition, we derive that

$$\Phi(y) - \Phi^* \geq \frac{\beta}{2} d^2(x, \mathcal{X}) = \frac{\beta}{2} \|y - y'\|^2 \geq \frac{\beta}{L} D_w(y, y') \geq \frac{\beta}{L} \inf_{z \in \mathcal{X}} D_w(y, z),$$

where y' stands for the projective point of y onto $\mathcal{X}$ and the second inequality follows from the gradient L-Lipschitz continuous property. Therefore, the Bregman distance growth condition (8) holds as well.

Besides, we need an analog of the following inequality for the Bregman distance function:

$$\|v_{k+1} - v_1\|^2 = \|\sum_{j=1}^{k} (v_{j+1} - v_j)\|^2 \leq k \sum_{j=1}^{k} \|v_{j+1} - v_j\|^2.$$

So we make the following assumption:

A7. For any sequence $\{v_1, v_2, \cdots, v_{k+1}\} \subset \mathbb{R}^d$, it holds that

$$D_w(v_{k+1}, v_1) \leq \ell(k) \sum_{j=1}^{k} D_w(v_{j+1}, v_j),$$

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

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

$$\frac{\mu_w}{2} \|x - y\|^2 \leq D_w(x, y) \leq \frac{L_w}{2} \|x - y\|^2, \tag{9}$$

then

$$D_w(v_{k+1}, v_1) \leq \frac{L_w}{2} \|v_{k+1} - v_1\|^2 \leq \frac{kL_w}{2} \sum_{j=1}^{k} \|v_{j+1} - v_j\|^2 \leq \frac{kL_w}{\mu_w} \sum_{j=1}^{k} D_w(v_{j+1}, v_j),$$

i.e., the assumption A7 holds with $\ell(k) \leq \frac{kL_w}{\mu_w}$. Note that the condition (9) was used in [1]. From the deduction above, the assumption A7 is not stronger than the condition (9).

Finally, we need an (nearly) equivalent description of the strongly convex assumption on the sum function.

A8. Each component function is μ_n-strongly convex for some $\mu_n \geq 0$ such that $\sum_{n=1}^{N} \mu_n = \mu$.

Let each component function be convex. Then, the equivalence follows from

$$\mu = \sup_{x \neq y} \sum_{i=1}^{N} \frac{\langle \nabla f_n(x) - \nabla f_n(y), x - y \rangle}{\|x - y\|^2} = \sum_{i=1}^{N} \sup_{x \neq y} \frac{\langle \nabla f_n(x) - \nabla f_n(y), x - y \rangle}{\|x - y\|^2} = \sum_{n=1}^{N} \mu_n,$$

where we use the convexity to ensure each inner product to be nonnegative such that we can change the order of summation and taking supremum. Therefore, if we use assumptions A1 and A8 together, then A8 can be viewed as an equivalent description of the strongly convex assumption on the sum function.

3 Main results

Throughout this section, we remind the reader that, we consider the optimization model (1) with $F(x)$ given by (4) and the sequence $\{x_k\}$ generated by the PIAG method.

First, we introduce the following lemma, which was presented in [1] and will play an important role in forthcoming convergence rate analysis.

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

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

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

$$\frac{c}{1-a} \frac{1 - a^{k_0+1}}{a^{k_0}} \leq b.$$

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

Before presenting the main result of this paper, we state another lemma, which can be viewed as a generalization of the standard descent lemma for the proximal gradient method; see for example Lemma 2.3 in [4].

Lemma 2. *Suppose that the standard assumptions A1-A3 hold. Let*

$$\Delta_k^1 = \frac{L(\tau + 1)}{2} \sum_{j=k-\tau}^{k} \|x_{j+1} - x_j\|^2.$$

Then, the following holds:

$$\Phi(x_{k+1}) \leq \Phi(x) + \frac{1}{2\alpha}\|x - x_k\|^2 - \frac{1}{2\alpha}\|x - x_{k+1}\|^2 - \frac{1}{2\alpha}\|x_{k+1} - x_k\|^2 + \Delta_k^1. \tag{10}$$

Now, we present the main theorem of this paper.

Theorem 1. *Suppose that assumptions A1-A4 hold, and the step-size satisfies:*

$$\alpha \leq \frac{\left(1+\frac{\beta}{L}\frac{1}{\tau+1}\right)^{\frac{1}{(\tau+1)}}-1}{\beta}.$$

Define a Lyapunov function

$$\Psi(x) \triangleq \Phi(x) - \Phi^* + \frac{1}{2\alpha}d^2(x,\mathcal{X}).$$

Then, the PIAG method converges linearly in the sense that

$$\Psi(x_k) \leq \left(1-\frac{\alpha\beta}{1+\alpha\beta}\right)^k \Psi(x_0), \tag{11}$$

for all $k \geq 0$. In particular, the PIAG method attains a global linear convergence in function value error:

$$\Phi(x_k) - \Phi^* \leq \left(1-\frac{\alpha\beta}{1+\alpha\beta}\right)^k \Psi(x_0), \tag{12}$$

and a global linear convergence in iterate point error:

$$d^2(x_k,\mathcal{X}) \leq \Psi(x_0)\frac{2\alpha}{1+\alpha\beta}\left(1-\frac{\alpha\beta}{1+\alpha\beta}\right)^k, \tag{13}$$

for all $k \geq 0$. Furthermore, if

$$\alpha = \frac{\left(1+\frac{\beta}{L}\frac{1}{\tau+1}\right)^{\frac{1}{(\tau+1)}}-1}{\beta},$$

then

$$\Psi(x_k) \leq \left(1-\frac{1}{[1+\eta(1+\tau)](\tau+1)}\right)^k \Psi(x_0), \tag{14}$$

for all $k \geq 0$, where $\eta = L/\beta$ stands for the number condition of optimization problem (1).

Some comments are in order:

- First, since the strongly convex assumption implies the quadratic growth condition with $\beta = \mu$ and $\mathcal{X} = \{x^*\}$ with x^* being the unique minimizer of $\Phi(x)$, noting that

  $$1-\frac{\alpha\beta}{1+\alpha\beta} = \frac{1}{1+\alpha\beta} = \frac{1}{\mu\alpha+1},$$

  we have that the global linear convergence in iterate point error (13) attains the following form:

  $$\|x_k - x^*\|^2 \leq \Psi(x_0)\frac{2\alpha}{1+\alpha\mu}\left(\frac{1}{\mu\alpha+1}\right)^k. \tag{15}$$

  This recovers the main result from [1].

- Second, if the delay parameter $\tau \leq 47$, then following from (14) we have the following linear convergence in function value error:

  $$\Phi(x_k) - \Phi^* \leq \left(1-\frac{1}{49\eta(\tau+1)}\right)^k \Psi(x_0). \tag{16}$$

  This recovers the main result from [15]. Note that the constraint of $\tau \leq 47$ is not required in [15]. But there used the strongly convex conditions as an assumption.

4 Extension via general distance functions

In this section, we consider the constrained optimization problem:

$$\underset{x\in Q}{\text{minimize}}\, \Phi(x) \triangleq F(x) + h(x), \tag{17}$$

with $F(x) = \sum_{n=1}^{N} f_n(x)$. This problem can be written into the form of (1) if we introduce the indicate function of Q. Here, we explicitly express the set Q as a feasible set so that one can take advantage of its geometry. Based on this point, the authors of [1] proposed the following iterative scheme to solve (17):

$$g_k = \sum_{n=1}^{N} \nabla f_n(x_{k-\tau_k^n}), \tag{18}$$

$$x_{k+1} \leftarrow \arg\min_{x\in Q}\{h(x) + \langle g_k, x - x_k\rangle + \frac{1}{\alpha} D_w(x, x_k)\}, \tag{19}$$

which is an extension of the PIAG method by using $D_w(x, x_k)$ to replace $\frac{1}{2}\|x - x_k\|^2$. We call this scheme E-PIAG method. Its linear convergence rate in iterate point error was studied in [1] under very strong assumptions. Here, we try to use weaker conditions to recover the corresponding result from [1], but also derive new convergence rate results similar to Theorem 1.

First, we state a descent lemma, which is similar to Lemma 2.

Lemma 3. *Suppose that assumptions A2, A3, A5 and A7 hold. Let* $L = \sum_{n=1}^{N} L_n$ *and*

$$\Delta_k^2 = \ell(\tau+1)L \sum_{j=k-\tau}^{k} D_w(x_{j+1}, x_j).$$

Then, the following holds:

$$\Phi(x_{k+1}) \le \Phi(x) + \frac{1}{\alpha} D_w(x, x_k) - \frac{1}{\alpha} D_w(x, x_{k+1}) - \frac{1}{\alpha} D_w(x_{k+1}, x_k) + \Delta_k^2. \tag{20}$$

Now, we state a generalized theorem for the PIAG method with general distance functions.

Theorem 2. *Suppose that assumptions A2, A3, A5-A7 hold, and the step-size satisfies:*

$$\alpha \le \frac{\left(1 + \frac{\rho}{L}\frac{1}{\ell(\tau+1)}\right)^{\frac{1}{(\tau+1)}} - 1}{\rho},$$

where $L = \sum_{n=1}^{N} L_n$ *with* L_n *being constants appeared in the assumption A5. Define a new Lyapunov function*

$$\Gamma(x) \triangleq \Phi(x) - \Phi^* + \frac{1}{\alpha} \inf_{z\in\mathcal{X}} D_w(z, x).$$

Then, the E-PIAG method converges linearly in the sense that

$$\Gamma(x_k) \le \left(1 - \frac{\alpha\rho}{1+\alpha\rho}\right)^k \Gamma(x_0), \tag{21}$$

for all $k \geq 0$. In particular, the E-PIAG method attains a global linear convergence in function value error:

$$\Phi(x_k) - \Phi^* \leq \left(1 - \frac{\alpha\rho}{1+\alpha\rho}\right)^k \Gamma(x_0), \tag{22}$$

and a global linear convergence in iterate point error:

$$\inf_{z\in\mathcal{X}} D_w(z, x_k) \leq \alpha\Gamma(x_0)\left(1 - \frac{\alpha\rho}{1+\alpha\rho}\right)^k, \tag{23}$$

for all $k \geq 0$. Furthermore, if

$$\alpha = \frac{\left(1 + \frac{\rho}{L}\frac{1}{\ell(\tau+1)}\right)^{\frac{1}{(\tau+1)}} - 1}{\beta},$$

then

$$\Gamma(x_k) \leq \left(1 - \frac{1}{[\ell(\tau+1)+1](\tau+1)\theta}\right)^k \Gamma(x_0), \tag{24}$$

for all $k \geq 0$, where $\theta = L/\rho$ stands for the number condition of optimization problem (17).

Some comments are in order:

- First, under assumptions A1-A3, the strongly convex assumption, the condition (9), and requiring the reference function $w(x)$ to be strongly convex, the authors in [1] derived the following convergence rate result:

  $$D_w(x^*, x_k) \leq \left(1 - \frac{\mu\alpha}{\mu\alpha + L_w}\right)^k D_w(x^*, x_0), \tag{25}$$

  where μ is the strongly convex parameter of the objective function and x^* is the unique minimizer of $\Phi(x)$ over Q. The result (23) recovers this result with $L_w \geq 1$ under weaker assumptions, as discussed in Section 2. The remained results of Theorem 2 are new to our knowledge.

- Second, work [12] studied the gradient method with general distance functions for solving

  $$\underset{x\in Q}{\text{minimize}}\, f(x). \tag{26}$$

  Under the L-relative smoothness and μ-relatively strong convexity of $f(x)$, they showed that the sequence $\{x_k\}$ generated by

  $$x_{k+1} \leftarrow \arg\min_x \{\langle \nabla f(x_k), x - x_k\rangle + LD_w(x, x_k)\}, \tag{27}$$

  converges linearly in the function value error with the rate $1 - \frac{\mu}{L}$. When specialized to the problem (26) and the iterate (27), the convergence result (22) with $\alpha = \frac{1}{L}$ and $\rho = \mu$ reads

  $$f(x_k) - f^* \leq C \cdot \left(1 - \frac{\mu}{L+\mu}\right)^k, \tag{28}$$

  where $C > 0$ is a constant and f^* is the optimal objective value. Although this is slightly worse than the rate $1 - \frac{\mu}{L}$, the conditions used here is not stronger than the L-relative smoothness and μ-relatively strong convexity, as discussed in Section 2, and our results are more general.

- At last, the authors of [2] also studied the proximal gradient method with general distance functions. But they only discussed the sublinear convergence under the L-relative smoothness property. Theorem 2 provides complementary results to that in [2] and [12].

5 Variants of PIAG with convergence rate analysis

Recently, the author of [5] proposed a new variant of the PIAG method for solving

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\, F(x) = \sum_{n=1}^{N} f_n(x). \tag{29}$$

The new variant has the form

$$x_{k+1} = x_k - \alpha_k(\nabla f_{i_k}(x_{k+1}) + \sum_{i\neq i_k} \nabla f_i(x_{k-\tau_k^i})), \tag{30}$$

which is an incremental aggregated version of the proximal algorithm, and hence called incremental aggregated proximal (abbreviated IAP) algorithm. It is straightforward to verify the IAP method has the following equivalent form

$$x_{k+1} = \arg\min_x\{f_{i_k}(x) + \langle \sum_{i\neq i_k} \nabla f_i(x_{k-\tau_k^i}), x\rangle + \frac{1}{2\alpha_k}\|x - x_k\|^2\}. \tag{31}$$

The author of [5] then proved that the IAP method attains a global convergence rate in iterate point error under the strongly convex assumption of $F(x)$ and the gradient Lipschitz continuous property of each $f_n(x)$. Following closely the one of [9], the proof idea is to view the IAP iteration (30) as a gradient method with errors. However, this proof method is very limited and can only show the existence of linear convergence rate and does not readily extend to the constrained case. In this section, we will show that our method in this paper can avoid all these difficulties. To this end, we propose the following variant scheme of the PIAG method:

$$x_{k+1} = \arg\min_x\{h(x) + \sum_{i\in\mathcal{I}_k} f_{i_k}(x) + \langle \sum_{j\in\mathcal{J}_k} \nabla f_j(x_{k-\tau_k^j}), x\rangle + \frac{1}{2\alpha_k}\|x - x_k\|^2\}, \tag{32}$$

where $\mathcal{I}_k, \mathcal{J}_k$ are index sets satisfying that

$$\mathcal{I}_k \bigcup \mathcal{J}_k = \{1, 2, \cdots, N\}, \quad \mathcal{I}_k \bigcap \mathcal{J}_k = \emptyset, \quad \mathcal{J}_k \neq \emptyset.$$

Note that for different k, $\mathcal{I}_k$ does not need to be the same. It is easy to see that if $\mathcal{I}_k = \{i_k\}$ and $h(x) = 0$, then the iterate (32) recovers the original IAP iterate (30), and if $\mathcal{I}_k = \emptyset$, then the iterate (32) recovers the PIAG method. In this sense, the proposed scheme (32) can be viewed as a generalization of the PIAG method, and hence named G-PIAG method. In what follows, we show that the G-PIAG method converges linearly as well under very weak conditions. Actually, as should be predicted, its linear convergence rate can be better than that of the PIAG method since some of component functions are minimized directly.

Below, we remind the reader that, we consider the optimization model (1) with $F(x)$ given by (4) and the sequence $\{x_k\}$ generated by the G-PIAG method. First, we state a descent lemma.

Lemma 4. *Suppose that the standard assumptions A1-A4 hold. Let* $\alpha_k \equiv \alpha$, $L^{(k)} = \sum_{j\in\mathcal{J}_k} L_j$, $\widetilde{L} = \max_k\{L^{(k)}, \beta\}$, *and*

$$\Delta_k^3 = \frac{\widetilde{L}(\tau+1)}{2} \sum_{j=k-\tau}^{k} \|x_{j+1} - x_j\|^2.$$

Then, the following holds:

$$\Phi(x_{k+1}) \leq \Phi(x) + \frac{1}{2\alpha}\|x - x_k\|^2 - \frac{1}{2\alpha}\|x - x_{k+1}\|^2 - \frac{1}{2\alpha}\|x_{k+1} - x_k\|^2 + \Delta_k^3. \tag{33}$$

Combining Lemma 1 and Lemma 4 and repeating the argument used in the proof of Theorem 1, we can derive the following convergence result for the G-PIAG method, whose proof is omitted.

Theorem 3. *Suppose that assumptions A1-A4 hold. Let* $\alpha_k \equiv \alpha$, $L^{(k)} = \sum_{j\in\mathcal{J}_k} L_j$, $\widetilde{L} = \max_k\{L^{(k)}, \beta\}$, *and the step-size satisfy:*

$$\alpha \leq \frac{\left(1 + \frac{\beta}{\widetilde{L}}\frac{1}{\tau+1}\right)^{\frac{1}{(\tau+1)}} - 1}{\beta}.$$

Recall that

$$\Psi(x) = \Phi(x) - \Phi^* + \frac{1}{2\alpha}d^2(x, \mathcal{X}).$$

Then, the G-PIAG method converges linearly in the sense that

$$\Psi(x_k) \leq \left(1 - \frac{\alpha\beta}{1+\alpha\beta}\right)^k \Psi(x_0), \tag{34}$$

for all $k \geq 0$. *In particular, the G-PIAG method attains a global linear convergence in function value error:*

$$\Phi(x_k) - \Phi^* \leq \left(1 - \frac{\alpha\beta}{1+\alpha\beta}\right)^k \Psi(x_0), \tag{35}$$

and a global linear convergence in iterate point error:

$$d^2(x_k, \mathcal{X}) \leq \Psi(x_0)\frac{2\alpha}{1+\alpha\beta}\left(1 - \frac{\alpha\beta}{1+\alpha\beta}\right)^k, \tag{36}$$

for all $k \geq 0$. *Furthermore, if*

$$\alpha = \frac{\left(1 + \frac{\beta}{\widetilde{L}}\frac{1}{\tau+1}\right)^{\frac{1}{(\tau+1)}} - 1}{\beta}, \tag{37}$$

then

$$\Psi(x_k) \leq \left(1 - \frac{1}{(\tau+1)(\tau+2)\widetilde{\eta}}\right)^k \Psi(x_0), \tag{38}$$

for all $k \geq 0$, *where* $\widetilde{\eta} = \widetilde{L}/\beta$.

Compared Theorems 1 and 3, we find that the only difference is taking different constants L and $\widetilde{L}$. The latter can be strictly less than the former if $\mathcal{I}_k \neq \emptyset$ for all $k \geq 0$ (means there are at lease one component function minimized directly at each iterate), and hence the G-PIAG method can take a larger step-size in (37) to obtain a smaller convergence rate in (38), as predicted before.

6 Improving rates via exploiting strong convexity

Since the strongly convex assumption is strictly stronger than the quadratic growth condition, a natural question arises: Whether the former can help us improve linear convergence rates. In this section, we will positively answer this question. Denote

$$\hat{f}_n(x) \triangleq f_n(x) - \frac{\mu_n}{2}\|x\|^2, \quad \hat{h}(x) \triangleq h(x) + \frac{\mu}{2}\|x\|^2.$$

Then the optimization model (1) with $F(x)$ given by (4) can be equivalently written into the following form

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\, \Phi(x) = \sum_{n=1}^{N} \hat{f}_n(x) + \hat{h}(x). \tag{39}$$

With this reformulated form, we suggest the following iterative scheme

$$\hat{g}_k = \sum_{n=1}^{N} (\nabla f_n(x_{k-\tau_k^n}) - \mu_n x_{k-\tau_k^n}), \tag{40}$$

$$x_{k+1} = \arg\min_x \{h(x) + \frac{\mu}{2}\|x\|^2 + \langle \hat{g}_k, x - x_k\rangle + \frac{1}{2\alpha}\|x - x_k\|^2\}, \tag{41}$$

which is obtained by applying the PIAG method to (39). Note that under assumptions A1, A2, and A8, each new component function $\hat{f}_n(x)$ is convex with $(L_n - \mu_n)$-continuous gradient and the new regularization function $\hat{h}(x)$ is μ-strongly convex. Together with the $(\mu + \frac{1}{\alpha})$-strongly convex property of the objective function in (41), we can get an improved descent lemma, whose proof is omitted. We remind the reader that the sequence of this section is generated by the PIAG method applying to (39).

Lemma 5. *Suppose that assumptions A1-A3, A8 hold. Let*

$$\Delta_k^4 = \frac{(L-\mu)(\tau+1)}{2} \sum_{j=k-\tau}^{k} \|x_{j+1} - x_j\|^2.$$

Then, the following holds:

$$\Phi(x_{k+1}) \le \Phi(x) + \frac{1}{2\alpha}\|x - x_k\|^2 - (\frac{1}{2\alpha} + \frac{\mu}{2})\|x - x_{k+1}\|^2 - \frac{1}{2\alpha}\|x_{k+1} - x_k\|^2 + \Delta_k^4. \tag{42}$$

Now, we present the improved rate result under the strongly convex assumption.

Theorem 4. *Suppose that assumptions A1-A3, A8 hold, and the step-size satisfies:*

$$\alpha \le \frac{\left(1 + \frac{2\mu}{L-\mu}\frac{1}{\tau+1}\right)^{\frac{1}{(\tau+1)}} - 1}{2\mu}.$$

Let x^* be the unique minimizer of $\Phi(x)$ and Φ^* its minimum. Define a new Lyapunov function

$$\Omega(x) \triangleq \Phi(x) - \Phi^* + (\frac{1}{2\alpha} + \frac{\mu}{2})\|x - x^*\|^2.$$

Then, the PIAG method converges linearly in the sense that

$$\Omega(x_k) \leq \left(1 - \frac{2\alpha\mu}{1+2\alpha\mu}\right)^k \Omega(x_0), \tag{43}$$

for all $k \geq 0$. In particular, the PIAG method attains a global linear convergence in function value error:

$$\Phi(x_k) - \Phi^* \leq \left(1 - \frac{2\alpha\mu}{1+2\alpha\mu}\right)^k \Omega(x_0), \tag{44}$$

and a global linear convergence in iterate point error:

$$\|x_k - x^*\|^2 \leq \Omega(x_0)\frac{2\alpha}{1+2\alpha\mu}\left(1 - \frac{2\alpha\mu}{1+2\alpha\mu}\right)^k, \tag{45}$$

for all $k \geq 0$. Furthermore, if

$$\alpha = \frac{\left(1 + \frac{2\mu}{L-\mu}\frac{1}{\tau+1}\right)^{\frac{1}{(\tau+1)}} - 1}{2\mu},$$

then

$$\Omega(x_k) \leq \left(1 - \frac{2}{[2+(\tau+1)(Q-1)](\tau+1)}\right)^k \Omega(x_0), \tag{46}$$

for all $k \geq 0$, where $Q = L/\mu$ stands for the number condition of optimization problem (39).

Some comments are in order:

- First, it is fair to compare the convergence rates in (14) and (46) since both of them are obtained by taking the largest allowed step-sizes. For comparison, we let $\mu = \beta$ and hence $Q = \eta$. Then, the rates in (14) and (46) reads, respectively,

  $$r_1(\tau) \triangleq 1 - \frac{1}{[1+Q(\tau+1)](\tau+1)}, \quad r_2(\tau) \triangleq 1 - \frac{1}{[1+\frac{Q-1}{2}(\tau+1)](\tau+1)},$$

  from which we can see that the latter is smaller than the former. In particular, for $\tau = 1$ we have

  $$r_1(1) = 1 - \frac{1}{1+Q} > 1 - \frac{2}{1+Q} = r_2(1).$$

  In other words, the strongly convex assumption indeed helps us improve the linear convergence rates derived under the quadratic growth condition.

- Second, we would like to highlight that the improved rate result is obtained by efficiently exploiting the strong convexity of the sum function. Traditionally, the strong convexity was buried in the gradient step and merely used to derive linear convergence rates. In contrast, we utilize the strong convexity in the proximal step to get fast descent. We take the regularized sparse binary classification problem as an example to illustrate our point of view. The problem can be described by (1) with $F(x)$ given by (4), where

  $$f_n(x) = \frac{1}{N}\left(\log(1+\exp(-b_n\langle a_n, x\rangle)) + \frac{N}{2}\mu_n\|x\|^2\right), \quad h(x) = \lambda\|x\|_1.$$

Here, $a_n \in \mathbb{R}^d$ are the feature vectors, $b_n \in \{-1, 1\}$ are the corresponding binary labels, $\mu_n \geq 0$ and $\lambda > 0$ are regularization parameters. All of them are given. We can easily verify that assumptions A1-A3, A8 hold for this certain problem. Now, there are two ways of applying the PIAG method: One is the direct and traditional way as done like (5); The other is the new way as done like (40)-(41). Our theory shows that the latter has faster convergence than the former. Moreover, in terms of the discovery made in Section 5, we can also explain why in the new way one can improve the convergence rate. In fact, by letting $g_n(x) = \frac{1}{N}\log(1 + \exp(-b_n\langle a_n, x\rangle))$ and $g_{N+1}(x) = \frac{\mu}{2}\|x\|^2$ with $\mu = \sum_{n=1}^{N}\mu_n$, we can see that the iterative scheme (40)-(41) is just an application of the PIAG variant (32) with $\mathcal{I}_k \equiv \{N+1\}$ to the problem of minimizing $\Phi(x) = \sum_{n=1}^{N+1} g_n(x) + h(x)$. Hence, faster convergence can be expected as discussed below Theorem 3.

- At last, it should be noted that the iterative scheme (40)-(41) includes all the strongly convex parameters μ_n such that it becomes unfeasible in some practical cases.

7 Detailed Proofs

7.1 Proof of Lemma 2

We divide the proof into two parts. The first part is a slight modification of the method used for Theorem 1 in [1]. The second part is a standard argument.

Part 1. Since each component function $f_n(x)$ is convex with L_n-continuous gradient, we have the following upper bound estimations:

$$\begin{aligned} f_n(x_{k+1}) \leq & f_n(x_{k-\tau_k^n}) + \langle \nabla f_n(x_{k-\tau_k^n}), x_{k+1} - x_{k-\tau_k^n}\rangle + \frac{L_n}{2}\|x_{k+1} - x_{k-\tau_k^n}\|^2 \\ \leq & f_n(x) + \langle \nabla f_n(x_{k-\tau_k^n}), x_{k+1} - x\rangle + \frac{L_n}{2}\|x_{k+1} - x_{k-\tau_k^n}\|^2, \end{aligned} \tag{47}$$

where the second inequality follows from the convexity of $f_n(x)$. Summing (47) over all components functions and using the expression of g_k, we obtain

$$F(x_{k+1}) \leq F(x) + \langle g_k, x_{k+1} - x\rangle + \sum_{n=1}^{N}\frac{L_n}{2}\|x_{k+1} - x_{k-\tau_k^n}\|^2. \tag{48}$$

The last term of the inequality above can be upper-bounded using Jensen's inequality as follows:

$$\begin{aligned} \sum_{n=1}^{N}\frac{L_n}{2}\|x_{k+1} - x_{k-\tau_k^n}\|^2 = & \sum_{n=1}^{N}\frac{L_n}{2}\|\sum_{j=k-\tau_k^n}^{k}(x_{j+1} - x_j)\|^2 \\ \leq & \frac{L(\tau+1)}{2}\sum_{j=k-\tau}^{k}\|x_{j+1} - x_j\|^2 = \Delta_k^1. \end{aligned} \tag{49}$$

Therefore,

$$F(x_{k+1}) \leq F(x) + \langle g_k, x_{k+1} - x\rangle + \Delta_k^1. \tag{50}$$

Part 2. By the definition of x_{k+1}, x_{k+1} is the minimizer of the $\frac{1}{\alpha}$-strongly convex function

$$x \mapsto h(x) + \langle g_k, x - x_k\rangle + \frac{1}{2\alpha}\|x - x_k\|^2,$$

hence for all $x \in \mathbb{R}^d$, we have

$$h(x_{k+1}) \leq h(x) + \langle g_k, x - x_{k+1}\rangle + \frac{1}{2\alpha}\|x - x_k\|^2 - \frac{1}{2\alpha}\|x - x_{k+1}\|^2 - \frac{1}{2\alpha}\|x_{k+1} - x_k\|^2. \tag{51}$$

Adding (50) to (51) yields the announced inequality.

7.2 Proof of Lemma 3

The proof idea is similar to that of Lemma 2 and can be found for example in [1] and [2]. However, for the ease of the reader, we will sketch it below.

First, by the assumption A5, we can derive that

$$\begin{aligned} f_n(x_{k+1}) \leq & f_n(x_{k-\tau_k^n}) + \langle \nabla f_n(x_{k-\tau_k^n}), x_{k+1} - x_{k-\tau_k^n}\rangle + L_n D_w(x_{k+1}, x_{k-\tau_k^n}) \\ \leq & f_n(x) + \langle \nabla f_n(x_{k-\tau_k^n}), x_{k+1} - x\rangle + L_n D_w(x_{k+1}, x_{k-\tau_k^n}), \end{aligned} \tag{52}$$

where the two inequalities follow from the L_n-relatively smooth property and the convexity of $f_n(x)$, respectively. Summing (52) over all components functions and using the expression of g_k, we obtain

$$F(x_{k+1}) \leq F(x) + \langle g_k, x_{k+1} - x\rangle + \sum_{n=1}^{N} L_n D_w(x_{k+1}, x_{k-\tau_k^n}). \tag{53}$$

Applying the assumptions A3 and A7 and noting the monotonic increasing property of $\ell(\cdot)$, we derive that

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

Thus,

$$\begin{aligned} F(x_{k+1}) \leq & F(x) + \langle g_k, x_{k+1} - x\rangle + \sum_{n=1}^{N} L_n \ell(\tau + 1) \sum_{j=k-\tau}^{k} D_w(x_{j+1}, x_j) \\ = & F(x) + \langle g_k, x_{k+1} - x\rangle + \Delta_k^2. \end{aligned} \tag{54}$$

Second, by the optimality condition, x_{k+1} satisfies the following inclusion:

$$0 \in \partial h(x_{k+1}) + g_k + \frac{1}{\alpha}(\nabla w(x_{k+1}) - \nabla w(x_k)).$$

Substituting the latter in the subgradient inequality for the convex function $h(x)$, we have that

$$\begin{aligned} h(x_{k+1}) \leq & h(x) + \langle g_k + \frac{1}{\alpha}(\nabla w(x_{k+1}) - \nabla w(x_k)), x - x_{k+1}\rangle \\ = & h(x) + \langle g_k, x - x_{k+1}\rangle + \frac{1}{\alpha}\langle \nabla w(x_{k+1}) - \nabla w(x_k), x - x_{k+1}\rangle \\ = & h(x) + \langle g_k, x - x_{k+1}\rangle + \frac{1}{\alpha} D_w(x, x_k) - \frac{1}{\alpha} D_w(x, x_{k+1}) - \frac{1}{\alpha} D_w(x_{k+1}, x_k), \end{aligned} \tag{55}$$

where the last line follows from the three points identity of Bregman distance [7]; that is

$$D_w(x, z) - D_w(x, y) - D_w(y, z) = \langle \nabla w(y) - \nabla w(z), x - y\rangle.$$

Finally, adding (55) to (54) yields the announced result.

7.3 Proof of Lemma 4

Again, we sketch the proof below. Let

$$g_k = \sum_{j\in\mathcal{J}_k} \nabla f_j(x_{k-\tau_k^j}).$$

On one hand, using the convexity and gradient Lipshitz continuity of each $f_n(x)$, we can conclude that

$$\sum_{j\in\mathcal{J}_k} f_j(x_{k+1}) \leq \sum_{j\in\mathcal{J}_k} f_j(x) + \langle g_k, x_{k+1} - x\rangle + \sum_{j\in\mathcal{J}_k} \frac{L_j}{2}\|x_{k+1} - x_{k-\tau_k^j}\|^2. \tag{56}$$

On the other hand, using the strongly convex property of the objective in (32), we can conclude that

$$\begin{aligned}\sum_{i\in\mathcal{I}_k} f_i(x_{k+1}) + h(x_{k+1}) \leq & \sum_{i\in\mathcal{I}_k} f_i(x) + h(x) + \langle g_k, x - x_{k+1}\rangle \\ & + \frac{1}{2\alpha}\|x - x_k\|^2 - \frac{1}{2\alpha}\|x - x_{k+1}\|^2 - \frac{1}{2\alpha}\|x_{k+1} - x_k\|^2.\end{aligned} \tag{57}$$

Now, adding (57) to (56) and noting that

$$\begin{aligned}\sum_{j\in\mathcal{J}_k} \frac{L_j}{2}\|x_{k+1} - x_{k-\tau_k^j}\|^2 \leq & \frac{\sum_{j\in\mathcal{J}_k} L_j(\tau+1)}{2} \sum_{j=k-\tau}^{k} \|x_{j+1} - x_j\|^2 \\ \leq & \frac{\widetilde{L}(\tau+1)}{2} \sum_{j=k-\tau}^{k} \|x_{j+1} - x_j\|^2 = \Delta_k^3,\end{aligned} \tag{58}$$

we obtain the announced inequality.

7.4 Proof of Theorem 1

Below, we use x' to stand for the projection of x onto the set $\mathcal{X}$. Let us write successively the inequality in Lemma 2 at $x = x_k$ and then at $x = x'_k$. Note that $\Phi(x'_k) = \Phi^*$. We obtain

$$\Phi(x_{k+1}) - \Phi^* \leq \Phi(x_k) - \Phi^* - \frac{1}{\alpha}\|x_k - x_{k+1}\|^2 + \Delta_k^1 \tag{59}$$

and

$$\Phi(x_{k+1}) \leq \Phi^* + \frac{1}{2\alpha}\|x'_k - x_k\|^2 - \frac{1}{2\alpha}\|x'_k - x_{k+1}\|^2 - \frac{1}{2\alpha}\|x_{k+1} - x_k\|^2 + \Delta_k^1. \tag{60}$$

By the definition of projection, we have

$$\|x'_k - x_{k+1}\|^2 \geq \|x'_{k+1} - x_{k+1}\|^2 = d^2(x_{k+1}, \mathcal{X}). \tag{61}$$

We split the term $\frac{1}{2\alpha}\|x'_k - x_k\|^2$ into two terms as follows:

$$\frac{1}{2\alpha}\|x'_k - x_k\|^2 = \mu\|x'_k - x_k\|^2 + \nu\|x'_k - x_k\|^2,$$

where $\mu + \nu = \frac{1}{2\alpha}$ and $\mu, \nu > 0$. By the quadratic growth condition, we relax the second term to introduce the function value error $\Phi(x_k) - \Phi^*$. Thus,

$$\nu\|x_k' - x_k\|^2 = \nu d^2(x_k, \mathcal{X}) \leq \frac{2\nu}{\beta}(\Phi(x_k) - \Phi^*). \tag{62}$$

Now, using (60) together with (61) and (62), we obtain the following relation

$$\Phi(x_{k+1}) - \Phi^* + \frac{1}{2\alpha}d^2(x_{k+1}, \mathcal{X}) \leq \frac{2\nu}{\beta}(\Phi(x_k) - \Phi^*) + \mu d^2(x_k, \mathcal{X}) - \frac{1}{2\alpha}\|x_{k+1} - x_k\|^2 + \Delta_k^1. \tag{63}$$

As we will show later, this relation is sufficient for us to derive the desired result. But we now proceed in a slightly complicated way to include possible benefit brought by the relation (59).

Multiplying the relation (59) by a parameter $\lambda \geq 0$ and adding the resulting inequality to (63), we obtain

$$\begin{aligned}
&\Phi(x_{k+1}) - \Phi^* + \frac{1}{2\alpha(1+\lambda)}d^2(x_{k+1}, \mathcal{X}) \\
\leq &\frac{\lambda + \frac{2\nu}{\beta}}{1+\lambda}\left(\Phi(x_k) - \Phi^* + \frac{\mu}{\lambda + \frac{2\nu}{\beta}}d^2(x_k, \mathcal{X})\right) - \frac{\lambda + 0.5}{\alpha(1+\lambda)}\|x_{k+1} - x_k\|^2 + \Delta_k^1.
\end{aligned} \tag{64}$$

Denote $a(\lambda, \nu) = \frac{\lambda + \frac{2\nu}{\beta}}{1+\lambda}$. In order to apply Lemma 1 to derive linear convergence rate results, we require $a(\lambda, \nu) \in (0, 1)$, which can be guaranteed by letting $0 < \nu < \frac{\beta}{2}$. Besides, we require

$$\frac{\mu}{\lambda + \frac{2\nu}{\beta}} \leq \frac{1}{2\alpha(1+\lambda)}$$

to relax the first term on the right-hand side of (64). This can be guaranteed by letting

$$\nu \geq \frac{1}{\frac{2}{\beta} + 2(1+\lambda)\alpha} \triangleq \nu_0.$$

Now, for any fixed $\lambda \geq 0$, we use this lower bound of ν to yield the smallest rate:

$$a(\lambda, \nu_0) = 1 - \frac{\alpha\beta}{1 + \alpha\beta(1+\lambda)}.$$

Thus, the smallest rate for varying parameter $\lambda \geq 0$ is $1 - \frac{\alpha\beta}{1+\alpha\beta}$. In other words, the relation (59) can not help to improve the linear convergence rate. Now, we set $\lambda = 0$ and $\nu = \nu_0$ in (64). Note that for such setting,

$$\frac{\mu}{\lambda + \frac{2\nu}{\beta}} = \frac{1}{2\alpha}, \quad \frac{\lambda + \frac{2\nu}{\beta}}{1+\lambda} = 1 - \frac{\alpha\beta}{1+\alpha\beta} \triangleq a.$$

Thus, we use (64) to obtain

$$\begin{aligned}
&\Phi(x_{k+1}) - \Phi^* + \frac{1}{2\alpha}d^2(x_{k+1}, \mathcal{X}) \\
\leq &a\left(\Phi(x_k) - \Phi^* + \frac{1}{2\alpha}d^2(x_k, \mathcal{X})\right) - \frac{1}{2\alpha}\|x_{k+1} - x_k\|^2 + \Delta_k^1,
\end{aligned} \tag{65}$$

i.e.,

$$\Psi(x_{k+1}) \leq a\Psi(x_k) - \frac{1}{2\alpha}\|x_{k+1} - x_k\|^2 + \frac{L(\tau+1)}{2}\sum_{j=k-\tau}^{k}\|x_{j+1} - x_j\|^2, \tag{66}$$

where we use the expressions of $\Psi(x)$ and Δ_k^1. Note that $\|x_{j+1} - x_j\|^2 = 0$ for all $j < 0$. We are ready to apply Lemma 1 with $V_k = \Psi_k$, $w_k = \|x_{k+1} - x_k\|^2$, $a = 1 - \frac{\alpha\beta}{1+\alpha\beta}$, $b = \frac{1}{2\alpha}$, $c = \frac{L(\tau+1)}{2}$, and $k_0 = \tau$. To ensure Lemma 1 hold, we need

$$\frac{c}{\alpha\beta(1+\alpha\beta)^{-1}}\frac{1-(1+\alpha\beta)^{-\tau-1}}{(1+\alpha\beta)^{-\tau}} \leq \frac{1}{2\alpha}$$

to hold. Simplifying and rearranging terms we obtain the desired upper bound for the step-size α. The linear convergence result (11) follows from Lemma 1. In particular, (12) is a direct consequence of (11), and (13) follows from the quadratic growth condition and (11).

Finally, it remains to show (14). Taking the certain value $\alpha = \frac{\left(1+\frac{\beta}{L}\frac{1}{\tau+1}\right)^{\frac{1}{(\tau+1)}}-1}{\beta}$ in (11), $\eta = \frac{L}{\beta}$, and noting that

$$\left(1 + \frac{1}{\eta(\tau+1)}\right)^{-1} = 1 - \frac{1}{1+\eta(1+\tau)},$$

we derive that

$$\begin{aligned}
\Psi(x_k) &\leq \left(1 + \frac{1}{\eta(\tau+1)}\right)^{\frac{-k}{\tau+1}}\Psi(x_0) \\
&= \left(1 - \frac{1}{1+\eta(1+\tau)}\right)^{\frac{k}{\tau+1}}\Psi(x_0) \\
&\leq \left(1 - \frac{1}{[1+\eta(1+\tau)](\tau+1)}\right)^{k}\Psi(x_0),
\end{aligned} \tag{67}$$

where in the third line we use the Bernoulli inequality, i.e., $(1+x)^r \leq 1 + rx$ for any $x \geq -1$ and $r \in [0,1]$. This completes the proof.

7.5 Proof of Theorem 2

First, we denote

$$\mathcal{Y}_k \triangleq \arg\min_{x\in\mathcal{X}} D_w(x, x_k).$$

Recall that we have assumed that $\mathcal{X}$ is nonempty and compact in the assumption A6, and note that $D_w(x, x_k)$ is a smooth function. Applying Weierstrass' Theorem, we can see that $\mathcal{Y}_k$ is nonempty and compact. Take $\widetilde{x}_k \in \mathcal{Y}$ and write down the inequality in Lemma 3 at $x = \widetilde{x}_k$. Note that $\Phi(\widetilde{x}_k) = \Phi^*$. We obtain

$$\Phi(x_{k+1}) \leq \Phi^* + \frac{1}{\alpha}D_w(\widetilde{x}_k, x_k) - \frac{1}{\alpha}D_w(\widetilde{x}_k, x_{k+1}) - \frac{1}{\alpha}D_w(x_{k+1}, x_k) + \Delta_k^2. \tag{68}$$

Since $\widetilde{x}_k \in \mathcal{Y} \subset \mathcal{X}$, it holds that

$$\inf_{x\in\mathcal{X}} D_w(x, x_{k+1}) \leq D_w(\widetilde{x}_k, x_{k+1}). \tag{69}$$

Using the Bregman distance growth condition, we have

$$D_w(\widetilde{x}_k, x_k) = \inf_{x\in\mathcal{X}} D_w(x, x_k) \le \frac{1}{\rho}(\Phi(x_k) - \Phi^*). \tag{70}$$

Utilizing the inequalities above and repeating the arguments in the proof of Theorem 1, we obtain

$$\Gamma(x_{k+1}) \le \frac{1}{1+\alpha\rho}\Gamma(x_k) - \frac{1}{\alpha}D_w(x_{k+1}, x_k) + \ell(\tau+1)L\sum_{j=k-\tau}^{k} D_w(x_{j+1}, x_j). \tag{71}$$

Finally, applying Lemma 1 with $V_k = \Gamma(x_k)$, $w_k = D_w(x_{k+1}, x_k)$, $a = \frac{1}{1+\alpha\rho}$, $b = \frac{1}{\alpha}$, $c = \ell(\tau+1)L$, and $k_0 = \tau$ yields the desired result (21), and hence other results follow.

7.6 Proof of Theorem 4

The proof is similar to that of Theorem 1. We only sketch it below.

First, we write down the inequality in Lemma 5 at $x = x^*$ to obtain

$$\Phi(x_{k+1}) \le \Phi^* + \frac{1}{2\alpha}\|x^* - x_k\|^2 - (\frac{1}{2\alpha} + \frac{\mu}{2})\|x^* - x_{k+1}\|^2 - \frac{1}{2\alpha}\|x_{k+1} - x_k\|^2 + \Delta_k^4. \tag{72}$$

Note that the strongly convex assumption A8 implies that

$$\|x^* - x_k\|^2 \le \frac{2}{\mu}(\Phi(x_k) - \Phi^*). \tag{73}$$

Then, repeating the arguments in the proof of Theorem 1, we get

$$\Omega(x_{k+1}) \le \frac{1}{1+2\alpha\mu}\Omega(x_k) - \frac{1}{2\alpha}\|x_{k+1} - x_k\|^2 + \frac{(L-\mu)(\tau+1)}{2}\sum_{j=k-\tau}^{k}\|x_{j+1} - x_j\|^2, \tag{74}$$

Thus, applying Lemma 1 with $V_k = \Omega(x_k)$, $w_k = \|x_{k+1} - x_k\|^2$, $a = \frac{1}{1+2\alpha\mu}$, $b = \frac{1}{2\alpha}$, $c = \frac{(L-\mu)(\tau+1)}{2}$, and $k_0 = \tau$ yields the announced result (43), and hence the results (44) and (45) follow.

Finally, it remains to show (46). Taking the certain value $\alpha = \frac{\left(1+\frac{2\mu}{(L-\mu)}\frac{1}{\tau+1}\right)^{\frac{1}{(\tau+1)}}-1}{2\mu}$ in (43), $Q = \frac{L}{\mu}$, and noting that

$$\left(1 + \frac{2}{(Q-1)(\tau+1)}\right)^{-1} = 1 - \frac{2}{2+(Q-1)(\tau+1)},$$

we derive that

$$\begin{aligned}
\Psi(x_k) &\le \left(1 + \frac{2}{(Q-1)(\tau+1)}\right)^{\frac{-k}{\tau+1}}\Psi(x_0) \\
&= \left(1 - \frac{2}{2+(Q-1)(\tau+1)}\right)^{\frac{k}{\tau+1}}\Psi(x_0) \\
&\le \left(1 - \frac{2}{[2+(Q-1)(\tau+1)](\tau+1)}\right)^{k}\Psi(x_0),
\end{aligned} \tag{75}$$

where in the third line we use the Bernoulli inequality again. This completes the proof.

8 Concluding Remarks

In this paper, we develop a new and powerful strategy to study proximal incremental aggregated gradient methods. This new strategy can be divided into three steps: first, prove a descent lemma (as done as Lemmas 2, 3, and 4); then, construct a certain Lyapunov function to deduce an iterate relationship (as done in proofs of Theorems 1 and 2); finally, invoking Lemma 1 to conclude. All these steps can be done under very weak assumptions so that the proposed strategy can be used widely to cover the PIAG method and its variants.

We believe that the strategy developed in this study will find more applications to other types of incremental methods, including randomized versions of the PIAG method.

Acknowledgements

This work is supported by the National Science Foundation of China (No.11501569, No.61571008, and No.6147139). The main idea of this research was carried out while the author was visiting BeiJing International Center for Mathematical Research by invitation of Professor Zaiwen, Wen. Special thanks to my cousin Boya Ouyang who helped me with my English writing.

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*, 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*, 2016.

[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 aggregated proximal and augmented lagrangian algorithms. *arXiv:1509.09257, 2015.*

[6] 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, DOI 10.1007/s10107-016-1091-6*, 2016.

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

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

[9] M. Gurbuzbalaban, A. Ozdaglar, and P. Parillo. On the convergence rate of incremental aggregated gradient algorithms. *arXiv:1506.02081, 2015.*

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

[11] G. Li and T. K. Pong. Calculus of the exponent of kurdyka-lojasiewicz inequality and its applications to linear convergence of first-order methods. *arXiv:1602.02915v1 [math.OC] 9 Feb 2016.*

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

[13] 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.*

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

[15] 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.*

[16] H. Zhang. The restricted strong convexity revisited: analysis of equivalence to error bound and quadratic growth. *Optimization Letter, DOI:10.1007/s11590-016-1058-9, 2016.*

[17] H. Zhang. New analysis of linear convergence of gradient-type methods via unifying error bound conditions. *arXiv:1606.00269v4 [math.OC]*, 2016.

[18] Z. Zhou and M. C. So. A unified approach to error bounds for structured convex optimization problems. *Math. Program., Ser. A, DOI 10.1007/s10107-016-1100-9*, 2017.