Accelerated proximal gradient algorithm for weakly convex function*

Milan Barik[†] Suvendu Ranjan Pattanaik[‡]

Abstract

In this work, we investigate the accelerated proximal gradient algorithm (APG_α) for weakly convex composite optimization problems. Building upon the framework of Böhm and Wright [8], and additionally assuming that f is convex and coercive while g is bounded below, we establish an objective residual convergence rate of $\mathcal{O}\left(\frac{1}{j^2}\right)$ for $\alpha \geq 3$. Moreover, when $\alpha > 3$, this rate improves to $o\left(\frac{1}{j^2}\right)$, and the sequence of iterates $\{x_j\}_{j\in\mathbb{N}}$ converges globally to a minimum point. In addition, we prove that the squared subdifferential residual satisfies the convergence estimate $\min\limits_{1\leq i\leq j} \text{dist}^2\big(0, \partial F(x_i)\big) = o\left(\dfrac{1}{j^2}\right)$ for $\alpha \geq 3$.

Keywords. *Composite minimization, Proximal gradient method, Lyapunov analysis, convergence rates, weakly convex function*

MSC codes. *90C26,90C30,49J52*

1 Introduction

We consider the following optimization problem:

$$\min_{x\in\mathbb{R}^m} F(x) = f(x) + g(x), \tag{1}$$

where $f : \mathbb{R}^m \to \mathbb{R}$ is a convex coercive differentiable function with ∇f is $L_{\nabla f}-$ Lipschitz continuous, and $g : \mathbb{R}^m \to \mathbb{R}$ is proper, lower semi-continuous (l.s.c.) and $\rho-$ weakly convex function with bounded below.

The proximal gradient method (PGM) originates from the proximal point method, which was first introduced by Martinet [17] and later extended to composite optimization

*Submitted to the editors DATE.

†Department of Mathematics, NIT Rourkela, India (milanbarik7008@gmail.com).

‡Department of Mathematics, NIT Rourkela, India (suvendu.pattanaik@gmail.com).

problems by Rockafellar [21]. The forward–backward splitting method was subsequently developed in [16]. The ℓ_1-regularized optimization problem $\min\{f(x)+\lambda\|x\|_1 : x\in\mathbb{R}^m\}$ was studied using the Iterative Shrinkage–Thresholding Algorithm (ISTA), introduced in [10] for the case where f is smooth and convex, as an extension of the classical gradient descent method. ISTA generates a sequence $\{x_j\}_{j\in\mathbb{N}}$ according to $x_{j+1} = \mathrm{Prox}_{\lambda g}\big(x_j - s\nabla f(x_j)\big)$, for some stepsize $s>0$. When f is smooth and convex and $g(x) = \|x\|_1$, ISTA achieves the convergence rate $F(x_j) - F(x^*) = \mathcal{O}\left(\frac{1}{j}\right)$. Later, Beck and Teboulle [6] proposed the Fast Iterative Shrinkage–Thresholding Algorithm (FISTA) by incorporating Nesterov's acceleration technique [18] into ISTA for non-smooth convex optimization problem where the objective function $F = f+g$ with f is smooth convex and g is possibly non-smooth convex. Their method improves the convergence rate of the objective residual to $\mathcal{O}\left(\frac{1}{j^2}\right)$, which represents a significant improvement over ISTA. More recently, Bauschke and Moursi [5] established the convergence of the iterates generated by the classical FISTA algorithm.

Attouch and Peypouquet [3] established an improved convergence rate of $o\left(\frac{1}{j^2}\right)$ of objective residual and proved weak convergence of the iterates generated by the accelerated forward-backward method (AFB) for non-smooth composite convex functions. Later, Attouch et al. [2] incorporated perturbation terms into the analysis and obtained analogous convergence results in both continuous and discrete frameworks through Lyapunov techniques. Attouch and Cobot [1] further investigated the inertial forward-backward method (IFB) for general non-negative sequences α_j, deriving convergence properties comparable to those in [3].

Considerable attention has been devoted to optimization problems involving weakly convex functions. The proximal gradient method for weakly convex optimization has been investigated in [12, 11]. Liao and Zheng [15] proposed a proximal descent framework that combines the inexact proximal point method with classical convex bundle techniques, and established non-asymptotic convergence guarantees in terms of (n,ϵ)-inexact stationarity. More recently, Khanh et al. [13] developed an inexact proximal gradient method for composite weakly convex problems of the form $F = f+g$, where f is smooth and g is weakly convex, and proved the global convergence of the generated iterates. Böhm and Wright [8] introduced a variable smoothing approach for composite weakly convex problems of the form $F(x) = f(x)+g(Ax)$, where f is smooth, g is weakly convex, and A is a matrix. They established a complexity bound of $\mathcal{O}(1/\epsilon^3)$ for obtaining an ϵ-approximate solution. In the same work [8], they also proposed a proximal gradient method for the composite problem $F = f+g$, where f is $L_{\nabla f}$-smooth and g is a possibly nonsmooth weakly convex function, and obtained the convergence rate $\mathcal{O}\left(\frac{1}{\sqrt{j}}\right)$ for $\min\limits_{1\le i\le j} \mathrm{dist}(0,\partial F(x_i))$. In [7], Bednarczuk et al. studied both exact and inexact versions of the proximal gradient method for weakly convex composite problems of the form $F = f+g$, where f is $L_{\nabla f}$-smooth and convex, and g is nonsmooth and weakly convex. When the objective function F satisfies the sharpness condition:

$$F(x) - \inf_{x\in\mathbb{R}^m} F(x) \geq \mu \operatorname{dist}(x, S), \tag{2}$$

where $S = \arg\min_{x\in\mathbb{R}^m} F(x)$, and $\mu > 0$, they proved that the exact proximal gradient method converges locally to a global minimizer.

From the above discussion, it follows that accelerated variants of the proximal gradient method like AFB [3] for composite weakly convex optimization problems have not yet been thoroughly investigated through Lyapunov analysis. Moreover, the global convergence of the iterates generated by such accelerated proximal gradient schemes remains largely unexplored.

In this work, we investigate an accelerated proximal gradient method obtained by incorporating a momentum term into the algorithm proposed in [8]. In addition, we assume that f is convex and coercive, while g is bounded below. These assumptions on f and g are natural in many optimization settings. Under these conditions, we establish the following results for composite non-smooth weakly convex optimization problems.

Contribution

- We study the weakly convex composite optimization problem (1) using the APG_α method (3), which will be defined later. By employing a Lyapunov-based analysis, we establish that the objective residual converges at the rate $\mathcal{O}\left(\frac{1}{j^2}\right)$ when $\alpha \geq 3$, and this rate further improves to $o\left(\frac{1}{j^2}\right)$ for $\alpha > 3$. To the best of our knowledge, these are the first convergence results obtained for this class of weakly convex functions via Lyapunov analysis.

- The iterates $\{x_j\}_{j\geq 1}$ produced by APG_α converge to some minimum point when $\alpha > 3$ globally, this represents an improvement over the results found in [7].

- Böhm and Wright [8] established the convergence rate $\mathcal{O}\left(\frac{1}{\sqrt{j}}\right)$ for the norm of the subgradient operator. By additionally assuming that f is convex and coercive, and that g is bounded below, we obtain the improved convergence result $\min_{1\leq i\leq j} \operatorname{dist}^2\big(0, \partial F(x_{i+1})\big) = o\left(\dfrac{1}{j^2}\right)$ for $\alpha \geq 3$.

This paper is organized as follows: in Sect. (2), we introduce some definitions and state the required results. In Sect. (3), we prove the rate of convergence of objective gap is $o\left(\frac{1}{j^2}\right)$ using Lyapunov analysis. The iterates generated by APG_α converge to some minimal point, as has been shown in Sect. (4). It has been analyzed, the rate of the norm of the subgradient operator in Sect. (5).

2 Basic definitions and preliminaries

The basic definitions and notation required below are reviewed in this section, along with the characteristics that are utilized in the following.

Definition 1. *[8] A proper l.s.c function $g : \mathbb{R}^m \to \mathbb{R}$ is called a $\rho-$weakly convex for some $\rho \geq 0$ if the quadratically shifted function $g(\cdot) + \frac{\rho}{2}\|\cdot\|^2$ is convex.*

Definition 2. *[20] Suppose $f : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ is a extended real-valued function is called coercive if*

$$|f(x)| \to +\infty \qquad \text{as } \|x\| \to +\infty.$$

Lemma 1. *[4] Suppose $\phi \neq S \subseteq \mathbb{R}^m$ and $\{x_n\}_{n\in\mathbb{N}}$ is a sequence in $\mathbb{R}^m$. Assume*

(i) For all $x^* \in S$, $\lim\limits_{n\to+\infty} \|x_n - x^*\|$ exists.

(ii) Each cluster point of the sequence x_n belongs to S.

Then, (x_n) converges to a point in S as $n \to \infty$.

Lemma 2. *[4] Suppose $\{(a_n, \beta_n, \gamma_n, \xi_n)\}_{n\in\mathbb{N}}$ are sequences in positive real numbers such that $\sum\limits_{n\in\mathbb{N}} \gamma_n < +\infty$ and $\sum\limits_{n\in\mathbb{N}} \xi_n < +\infty$. If*

$$a_{n+1} \leq (1 + \gamma_n)a_n - \beta_n + \xi_n, \quad (\forall n \in \mathbb{N}).$$

Then $\{a_n\}_{n\in\mathbb{N}}$ converges and $\sum\limits_{n\in\mathbb{N}} \beta_n < +\infty$.

Let us define the composite gradient mapping $T_\lambda : \mathbb{R}^m \to \mathbb{R}^m$ as $\lambda \in (0, \frac{1}{\rho})$ with $0^{-1} = \infty$ by [13]

$$\begin{aligned} T_\lambda(x) &= \arg\min_{y\in\mathbb{R}^m} \left\{ g(y) + \frac{1}{2\lambda}\|y - (x - \lambda\nabla f(x))\|^2 \right\} \\ &= \arg\min_{y\in\mathbb{R}^m} \left\{ \frac{\lambda}{2}\|\nabla f(x)\|^2 + \langle \nabla f(x), y - x\rangle + \frac{1}{2\lambda}\|y - x\|^2 + g(y) \right\} \\ &= \arg\min_{y\in\mathbb{R}^m} \left\{ f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2\lambda}\|y - x\|^2 + g(y) \right\}. \end{aligned}$$

Define further the gradient mapping $G_\lambda : \mathbb{R}^m \to \mathbb{R}^m$ by

$$G_\lambda(x) = \frac{1}{\lambda}(x - T_\lambda(x)).$$

The accelerated proximal gradient method is defined as

$$\mathrm{APG}_\alpha : \quad \begin{cases} y_j = x_j + \left(1 - \dfrac{\alpha}{j}\right)(x_j - x_{j-1}), \\ x_{j+1} = \mathrm{Prox}_{\lambda g}\big(y_j - \lambda \nabla f(y_j)\big) = y_j - \lambda G_\lambda(y_j). \end{cases} \tag{3}$$

We make the following hypothesis

$$H : \quad \begin{cases} f \text{ is a } \mathcal{C}^1 \text{ convex coercive function with } \nabla f \text{ being } L_{\nabla f}\text{-Lipschitz continuous;} \\ g \text{ is a proper, l.s.c. and } \rho\text{-weakly convex with bounded below;} \\ S = \arg\min F \neq \phi. \end{cases}$$

3 Lyapunov function

We define the Lyapunov function in the following way:

$$\mathcal{E}(j) := t_j^2(F(x_j) - F(x^*)) + \frac{(1-\rho\lambda)}{2\lambda}\|z_j - x^*\|^2, \tag{4}$$

where

$$z_j := x_{j-1} + t_j(x_j - x_{j-1}), \qquad \text{and } t_j := \frac{j-1}{\alpha - 1}. \tag{5}$$

Lemma 3. *Suppose that (H) holds, and $\lambda \in \left(0, \min\{\frac{1}{\rho}, \frac{1}{L_{\nabla f}}\}\right)$. If $\alpha > 3$, then*

$$\sum_{j=1}^{\infty} j(F(x_j) - F(x^*)) \leq \frac{(\alpha-1)^2 \mathcal{E}(1)}{(\alpha - 3)} < +\infty.$$

Proof. By the definition of the proximal map, this map

$$z \longmapsto g(z) + f(\bar{x}) + \langle \nabla f(\bar{x}), z - \bar{x}\rangle + \frac{1}{2\gamma}\|z - \bar{x}\|^2.$$

is $(\gamma^{-1} - \rho)$-strongly convex function, thus for all $x, y \in \mathbb{R}^m$, we have

$$\begin{aligned} g(x_{j+1}) + f(y) + \langle \nabla f(y), x_{j+1} - y\rangle + \frac{1}{2\lambda}\|x_{j+1} - y\|^2 + \frac{\lambda^{-1} - \rho}{2}\|x - x_{j+1}\|^2 \\ \leq g(x) + f(y) + \langle \nabla f(y), x - y\rangle + \frac{1}{2\lambda}\|x - y\|^2. \end{aligned} \tag{6}$$

From [19], as f is convex and $\nabla f(x)$ is $L_{\nabla f}$ Lipschitz continuous, we have

$$f(x_{j+1}) \leq f(y) + \langle \nabla f(y), x_{j+1} - y\rangle + \frac{L_{\nabla f}}{2}\|x_{j+1} - y\|^2, \tag{7}$$

since $\lambda < \frac{1}{L_{\nabla f}}$, we get

$$f(x_{j+1}) \leq f(y) + \langle \nabla f(y), x_{j+1} - y \rangle + \frac{1}{2\lambda}\|x_{j+1} - y\|^2, \tag{8}$$

Utilizing Equation (8) in (6), we obtain

$$g(x_{j+1}) + f(x_{j+1}) + \frac{\lambda^{-1} - \rho}{2}\|x - x_{j+1}\|^2 \leq g(x) + f(y) + \langle \nabla f(y), x - y \rangle + \frac{1}{2\lambda}\|x - y\|^2. \tag{9}$$

Since f is convex [19], we have $f(y) + \langle \nabla f(y), x - y \rangle \leq f(x)$ and $F = f + g$, Equation (9) reduces to

$$F(x_{j+1}) + \frac{\lambda^{-1} - \rho}{2}\|x - x_{j+1}\|^2 \leq F(x) + \frac{1}{2\lambda}\|x - y\|^2. \tag{10}$$

Putting $x = y = x_j$ in Equation (10), we obtain

$$F(x_{j+1}) + \frac{\lambda^{-1} - \rho}{2}\|x_j - x_{j+1}\|^2 \leq F(x_j). \tag{11}$$

From APG$_\alpha$ (3), we have $x_{j+1} = y_j - \lambda G_\lambda(y_j)$, then

$$\begin{aligned} \|x_j - x_{j+1}\|^2 &= \|x_j - y_j + \lambda G_\lambda(y_j)\|^2 \\ &= \|x_j - y_j\|^2 + \lambda^2\|G_\lambda(y_j)\|^2 + 2\lambda\langle G_\lambda(y_j), x_j - y_j \rangle. \end{aligned} \tag{12}$$

Applying Equation (12) in (11), we obtain

$$F(x_{j+1}) + \frac{\lambda^{-1} - \rho}{2}\|x_j - y_j\|^2 + \frac{\lambda(1 - \rho\lambda)}{2}\|G_\lambda(y_j)\|^2 + (1 - \rho\lambda)\langle G_\lambda(y_j), x_j - y_j \rangle \leq F(x_j), \tag{13}$$

also

$$F(x_{j+1}) \leq F(x_j) + (1 - \rho\lambda)\langle G_\lambda(y_j), y_j - x_j \rangle - \frac{\lambda(1 - \rho\lambda)}{2}\|G_\lambda(y_j)\|^2. \tag{14}$$

Putting $x = y = x^*$ in Equation (10), we obtain

$$F(x_{j+1}) + \frac{\lambda^{-1} - \rho}{2}\|x^* - x_{j+1}\|^2 \leq F(x^*). \tag{15}$$

From APG$_\alpha$ (3), we have $x_{j+1} = y_j - \lambda G_\lambda(y_j)$, then

$$\begin{aligned} \|x^* - x_{j+1}\|^2 &= \|x^* - y_j + \lambda G_\lambda(y_j)\|^2 \\ &= \|x^* - y_j\|^2 + \lambda^2\|G_\lambda(y_j)\|^2 + 2\lambda\langle G_\lambda(y_j), x^* - y_j \rangle. \end{aligned} \tag{16}$$

Applying Equation (16) in (15), we obtain

$$F(x_{j+1}) + \frac{\lambda^{-1} - \rho}{2}\|x^* - y_j\|^2 + \frac{\lambda(1 - \rho\lambda)}{2}\|G_\lambda(y_j)\|^2 + (1 - \rho\lambda)\langle G_\lambda(y_j), x^* - y_j \rangle \leq F(x^*), \tag{17}$$

also

$$F(x_{j+1}) \leq F(x^*) + (1-\rho\lambda)\langle G_\lambda(y_j), y_j - x^*\rangle - \frac{\lambda(1-\rho\lambda)}{2}\|G_\lambda(y_j)\|^2. \tag{18}$$

We have $t_j = \frac{j-1}{\alpha-1}$ and $\alpha_j = 1 - \frac{\alpha}{j} = \frac{t_j - 1}{t_{j+1}}$. Then multiplying $(t_{j+1}-1) \geq 0$ in the equation (14) and after that adding with the equation (18), we obtain

$$\begin{aligned} t_{j+1}F(x_{j+1}) \leq (t_{j+1}-1)F(x_j) + F(x^*) - \frac{\lambda t_{j+1}}{2}(1-\rho\lambda)\|G_\lambda(y_j)\|^2 \\ + (1-\rho\lambda)\langle G_\lambda(y_j), (t_{j+1}-1)(y_j - x_j) + (y_j - x^*)\rangle. \end{aligned} \tag{19}$$

Now,

$$\begin{aligned} (t_{j+1}-1)(y_j - x_j) + y_j &= x_j + t_{j+1}(y_j - x_j) \\ &= x_j + t_{j+1}\alpha_j(x_j - x_{j-1}) \\ &= x_j + (t_j - 1)(x_j - x_{j-1}) \\ &= x_{j-1} + t_j(x_j - x_{j-1}). \end{aligned} \tag{20}$$

Since $z_j = x_{j-1} + t_j(x_j - x_{j-1})$, then the Equation (19) becomes

$$\begin{aligned} t_{j+1}F(x_{j+1}) &\leq (t_{j+1}-1)F(x_j) + F(x^*) \\ &\quad - \frac{\lambda t_{j+1}}{2}(1-\rho\lambda)\|G_\lambda(y_j)\|^2 + (1-\rho\lambda)\langle G_\lambda(y_j), z_j - x^*\rangle. \end{aligned}$$

By subtracting $t_{j+1}F(x^*)$ on both sides in the above equation, we get

$$\begin{aligned} t_{j+1}(F(x_{j+1}) - F(x^*)) \leq (t_{j+1}-1)(F(x_j) - F(x^*)) - \frac{\lambda t_{j+1}}{2}(1-\rho\lambda)\|G_\lambda(y_j)\|^2 \\ + (1-\rho\lambda)\langle G_\lambda(y_j), z_j - x^*\rangle. \end{aligned} \tag{21}$$

Now

$$\begin{aligned} z_{j+1} - z_j &= x_j + t_{j+1}(x_{j+1} - x_j) - x_{j-1} - t_j(x_j - x_{j-1}) \\ &= t_{j+1}(x_{j+1} - x_j) - (t_j - 1)(x_j - x_{j-1}) \\ &= t_{j+1}(x_{j+1} - x_j - \alpha_j(x_j - x_{j-1})) \\ &= t_{j+1}(x_{j+1} - y_j) = -\lambda t_{j+1}G_\lambda(y_j). \end{aligned} \tag{22}$$

Then $z_{j+1} - x^* = z_j - x^* - \lambda t_{j+1}G_\lambda(y_j)$ and we have

$$\|z_{j+1} - x^*\|^2 = \|z_j - x^*\|^2 - 2\lambda t_{j+1}\langle G_\lambda(y_j), z_j - x^*\rangle + \lambda^2 t_{j+1}^2\|G_\lambda(y_j)\|^2.$$

Using the above equation in (21), we obtain

$$\begin{aligned} t_{j+1}(F(x_{j+1}) - F(x^*)) &\leq (t_{j+1}-1)(F(x_j) - F(x^*)) \\ &\quad + \frac{(1-\rho\lambda)}{2\lambda t_{j+1}}(\|z_j - x^*\|^2 - \|z_{j+1} - x^*\|^2) \end{aligned}$$

and

$$t_{j+1}^2(F(x_{j+1}) - F(x^*)) \le (t_{j+1}^2 - t_{j+1})(F(x_j) - F(x^*)) + \frac{(1-\rho\lambda)}{2\lambda}(\|z_j - x^*\|^2 - \|z_{j+1} - x^*\|^2).$$

Equivalently,

$$\begin{aligned} t_{j+1}^2(F(x_{j+1}) - F(x^*)) + \frac{(1-\rho\lambda)}{2\lambda}\|z_{j+1} - x^*\|^2 \\ \le t_j^2(F(x_j) - F(x^*)) + \frac{(1-\rho\lambda)}{2\lambda}\|z_j - x^*\|^2 \\ + (t_{j+1}^2 - t_{j+1} - t_j^2)(F(x_j) - F(x^*)). \end{aligned} \tag{23}$$

From Equation (4), Equation (23) becomes

$$\mathcal{E}(j+1) - (t_{j+1}^2 - t_{j+1} - t_j^2)(F(x_j) - F(x^*)) \le \mathcal{E}(j). \tag{24}$$

Since $t_{j+1}^2 - t_{j+1} - t_j^2 = -\frac{j(\alpha-3)+1}{(\alpha-1)^2} \le 0$ for $\alpha \ge 3$, we obtain from Equation (24),

$$\mathcal{E}(j+1) + \frac{j(\alpha-3)+1}{(\alpha-1)^2}(F(x_j) - F(x^*)) \le \mathcal{E}(j). \tag{25}$$

By summing the above equation, we obtain

$$\sum_{j=1}^{\infty} j(F(x_j) - F(x^*)) \le \frac{(\alpha-1)^2\mathcal{E}(1)}{(\alpha-3)} < +\infty.$$

□

Theorem 4. *Suppose that (H) holds, $\alpha \ge 3$ and $\lambda \in \Big(0, \min\{\frac{1}{\rho}, \frac{1}{L_{\nabla f}}\}\Big)$. Let $\{(y_j, x_j)\}_{j\in\mathbb{N}}$ be the sequences induced by* APG$_\alpha$. *Then, we have*

$$F(x_j) - F(x^*) \le \frac{(\alpha-1)^2}{j^2}\mathcal{E}(1) = \mathcal{O}(\frac{1}{j^2}), \qquad \|x_j - x_{j-1}\| = \mathcal{O}\Big(\frac{1}{j}\Big).$$

Proof. From the Equation (25), we obtain $\mathcal{E}(j+1) \le \mathcal{E}(j)$ and also, we have $\mathcal{E}(j) \le \mathcal{E}(1)$. Equation (4) implies that $t_j^2(F(x_j) - F(x^*)) + \frac{(1-\rho\lambda)}{2\lambda}\|z_j - x^*\|^2 \le \mathcal{E}(1)$, since $t_j = \frac{j-1}{\alpha-1}$ then

$$F(x_j) - F(x^*) \le \frac{(\alpha-1)^2}{(j-1)^2}\mathcal{E}(1) \le \frac{(\alpha-1)^2}{j^2}\mathcal{E}(1) = \mathcal{O}\Big(\frac{1}{j^2}\Big).$$

Again, since $z_j = x_{j-1} + t_j(x_j - x_{j-1})$, we obtain

$$\|x_j - x_{j-1}\| = \mathcal{O}\Big(\frac{1}{j}\Big).$$

□

Lemma 5. *Let assume the hypothesis* (H), *if* $\alpha > 3$, *then*

$$\sum_{j=1}^{\infty} j\|x_{j+1}-x_j\|^2 \leq \frac{\lambda}{1-\rho\lambda}\Big[(\alpha-1)^2\mathcal{E}(1) + \frac{3(\alpha-1)^2\mathcal{E}(1)}{\alpha-3}\Big] < +\infty.$$

Proof. From Equation (10) and from the algorithm $x_{j+1} = y_j - \lambda G_\lambda(y_j)$ and $y_j - x_j = \alpha_j(x_j - x_{j-1})$, we obtain

$$\begin{aligned} F(x_{j+1}) &\leq F(x_j) + (1-\rho\lambda)\langle G_\lambda(y_j), y_j - x_j\rangle - \frac{\lambda(1-\rho\lambda)}{2}\|G_\lambda(y_j)\|^2 \\ &= F(x_j) + \frac{(1-\rho\lambda)}{\lambda}\langle y_j - x_{j+1}, y_j - x_j\rangle - \frac{(1-\rho\lambda)}{2\lambda}\|x_{j+1}-y_j\|^2 \\ &= F(x_j) + \frac{(1-\rho\lambda)}{\lambda}\langle x_j - x_{j+1} + \alpha_j(x_j - x_{j-1}), \alpha_j(x_j - x_{j-1})\rangle \\ &- \frac{(1-\rho\lambda)}{2\lambda}\|x_{j+1} - x_j - \alpha_j(x_j - x_{j-1})\|^2, \end{aligned}$$

and then we get

$$\begin{aligned} F(x_{j+1}) &\leq F(x_j) - \frac{\alpha_j(1-\rho\lambda)}{s}\langle x_{j+1}-x_j, x_j - x_{j-1}\rangle + \frac{\alpha_j^2(1-\rho\lambda)}{\lambda}\|x_j - x_{j-1}\|^2 \\ &- \frac{(1-\rho\lambda)}{2\lambda}\|x_{j+1}-x_j\|^2 - \frac{\alpha_j^2(1-\rho\lambda)}{2\lambda}\|x_j - x_{j-1}\|^2 \\ &+ \frac{\alpha_j(1-\rho\lambda)}{\lambda}\langle x_{j+1}-x_j, x_j - x_{j-1}\rangle. \end{aligned}$$

Hence, we obtain

$$F(x_{j+1}) + \frac{(1-\rho\lambda)}{2}\|x_{j+1}-x_j\|^2 \leq F(x_j) + \frac{\alpha_j^2(1-\rho\lambda)}{2\lambda}\|x_j - x_{j-1}\|^2. \tag{26}$$

Since $\alpha_j \leq \alpha_{j+1}$, then Equation (26) reduces to

$$F(x_{j+1}) + \frac{(1-\rho\lambda)}{2}\|x_{j+1}-x_j\|^2 \leq F(x_j) + \frac{\alpha_{j+1}^2(1-\rho\lambda)}{2\lambda}\|x_j - x_{j-1}\|^2.$$

Now subtract $F(x^*)$ on both sides of the above equation, we get

$$\begin{aligned} F(x_{j+1}) - F(x^*) &+ \frac{(1-\rho\lambda)}{2\lambda}\|x_{j+1}-x_j\|^2 \\ &\leq F(x_j) - F(x^*) + \frac{\alpha_{j+1}^2(1-\rho\lambda)}{2\lambda}\|x_j - x_{j-1}\|^2. \end{aligned} \tag{27}$$

Let $\theta_j = F(x_j) - F(x^*)$, $d_j = \frac{(1-\rho\lambda)}{2\lambda}\|x_{j+1}-x_j\|^2$.Then Equation (27) becomes

$$\theta_{j+1} + d_j \leq \theta_j + \alpha_{j+1}^2 d_{j-1} = \theta_j + \Big(\frac{j+1-\alpha}{j+1}\Big)^2 d_{j-1}. \tag{28}$$

As $\alpha > 3$ then $j + 1 - \alpha \leq j + 1 - 2 = j - 1$, Equation (28) becomes

$$\theta_{j+1} + d_j \leq \theta_j + \Big(\frac{j-1}{j+1}\Big)^2 d_{j-1}, \tag{29}$$

and then

$$(j+1)^2 d_j - (j-1)^2 d_{j-1} \leq (j+1)^2(\theta_j - \theta_{j+1}).$$

Now, for $j \geq 1$

$$(j+1)^2(\theta_j - \theta_{j+1}) = j^2\theta_j - (j+1)^2\theta_{j+1} + (2j+1)\theta_j \leq j^2\theta_j - (j+1)^2\theta_{j+1} + 3j\theta_j.$$

Furthermore,

$$(j+1)^2 d_j - (j-1)^2 d_{j-1} = j^2 d_j - (j-1)^2 d_{j-1} + (2j+1)d_j.$$

From the above calculation, we get

$$\begin{aligned} 2jd_j + j^2 d_j - (j-1)^2 d_{j-1} &\leq (j+1)^2 d_j - (j-1)^2 d_{j-1} \\ &\leq (j+1)^2(\theta_j - \theta_{j+1}) \\ &\leq j^2\theta_j - (j+1)^2\theta_{j+1} + 3j\theta_j. \end{aligned} \tag{30}$$

Summing the Equation (30) for $t = 1, 2, \cdots, N$, we obtain

$$2\sum_{j=1}^{N} jd_j + N^2 d_N \leq \theta_1 - (N+1)^2\theta_{N+1} + 3\sum_{j=1}^{N} j\theta_j.$$

If $\alpha > 3$, from Lemma (3) we obtain

$$2\sum_{j=1}^{N} jd_j \leq (\alpha-1)^2\mathcal{E}(1) + \frac{3(\alpha-1)^2\mathcal{E}(1)}{\alpha-3}.$$

Now, we take as $N \to \infty$, we claim our result. □

Lemma 6. *Let assume the hypothesis* (H)*, if* $\alpha > 3$*, then*

$$\lim_{j\to\infty}\big[j^2\|x_{j+1} - x_j\|^2 + (j+1)^2(F(x_{j+1}) - F(x^*))\big]$$

exists.

Proof. From the Equation (28), we obtain

$$\theta_{j+1} + d_j \le \theta_j + \Big(\frac{j+1-\alpha}{j+1}\Big)^2 d_{j-1}.$$

Since $\alpha > 3$,

$$\theta_{j+1} + d_j \le \theta_j + \Big(\frac{j-1}{j}\Big)^2 d_{j-1},$$

equivalently

$$j^2 d_j - (j-1)^2 d_{j-1} \le j^2(\theta_j - \theta_{j+1}). \tag{31}$$

We have

$$(j+1)^2\theta_{j+1} - j^2\theta_j \le j^2(\theta_{j+1} - \theta_j) + 2(j+1)\theta_{j+1}. \tag{32}$$

Adding Equation (31) and (32), we obtain

$$\big[j^2 d_j + (j+1)^2\theta_{j+1}\big] - \big[(j-1)^2 d_{j-1} + j^2\theta_j\big] \le 2(j+1)\theta_{j+1}. \tag{33}$$

From Lemma (3), the series $\sum_{j=1}^{+\infty} j\theta_j < +\infty$. Applying Lemma (2), we get $\lim_{j\to\infty} (j-1)^2 d_{j-1} + j^2\theta_j$ exists. □

Theorem 7. *Suppose that* (H) *holds,* $\alpha > 3$*, and* $\lambda \in \Big(0, \min\{\frac{1}{\rho}, \frac{1}{L_{\nabla f}}\}\Big)$*. Let* $\{(y_j, x_j)\}_{j\in\mathbb{N}}$ *be the sequences induced by* APG_α*. Then, we have*

$$F(x_j) - \min_{\mathbb{R}^m} F = o\left(\frac{1}{j^2}\right), \qquad \|x_{j+1} - x_j\| = o\left(\frac{1}{j}\right).$$

Proof. From Lemma (3) and (5), we deduce that

$$\sum_{j=1}^{\infty} \frac{1}{j}\big[j^2\|x_{j+1} - x_j\|^2 + (j+1)^2(F(x_{j+1}) - F(x^*))\big] < +\infty.$$

Combining this with Lemma (6), we obtain

$$\lim_{j\to\infty}\big[j^2\|x_{j+1} - x_j\|^2 + (j+1)^2(F(x_{j+1}) - F(x^*))\big] = 0.$$

Therefore, we deduce that $j\|x_{j+1} - x_j\| \to 0$ and $(j+1)^2(F(x_{j+1}) - F(x^*)) \to 0$, and we obtain our results. □

4 Iterates convergence

The following statement shows the convergence of the iterates generated by APG_α.

Theorem 8. *Suppose that* (H) *holds,* $\alpha > 3$, *and* $\lambda \in \Big(0, \min\{\frac{1}{\rho}, \frac{1}{L_{\nabla f}}\}\Big)$. *Let* $\{(y_j, x_j)\}_{j\in\mathbb{N}}$ *be the sequences induced by* APG_α. *Then the sequence* $\{x_j\}_{j\in\mathbb{N}}$ *converges to a point in* S.

Proof. From Equation (5), we have $z_j = x_{j-1} + t_j(x_j - x_{j-1})$, therefore

$$\begin{aligned}\|z_j - x^*\|^2 &= \|x_{j-1} + t_j(x_j - x_{j-1}) - x^*\|^2 \\ &= \Big(\frac{j-1}{\alpha-1}\Big)^2 \|x_j - x_{j-1}\|^2 + 2\frac{j-1}{\alpha-1}\langle x_{j-1} - x^*, x_j - x_{j-1}\rangle + \|x_{j-1} - x^*\|^2. \end{aligned} \tag{34}$$

Using

$$\begin{aligned}\|x_j - x^*\|^2 &= \|x_j - x_{j-1} + x_{j-1} - x^*\|^2 \\ &= \|x_j - x_{j-1}\|^2 + \|x_{j-1} - x^*\|^2 + 2\langle x_{j-1} - x^*, x_j - x_{j-1}\rangle. \end{aligned}$$

we get

$$\begin{aligned}\|z_j - x^*\|^2 = &\Big[\Big(\frac{j-1}{\alpha-1}\Big)^2 - \Big(\frac{j-1}{\alpha-1}\Big)\Big]\|x_j - x_{j-1}\|^2 \\ &+ 2\frac{j-1}{\alpha-1}\big[\|x_j - x^*\|^2 - \|x_{j-1} - x^*\|^2\big] + \|x_{j-1} - x^*\|^2. \end{aligned}$$

We require to prove that $\lim_{j\to\infty} \|z_j - x^*\|$ exists. By Lemma (6) and (5), it suffices to prove that

$$\delta_j = (j-1)\big[\|x_j - x^*\|^2 - \|x_{j-1} - x^*\|^2\big] + (\alpha-1)\|x_{j-1} - x^*\|^2$$

has a limit as $j \to \infty$. Since the energy function $\{\mathcal{E}(j)\}_{j\in\mathbb{N}}$ is bounded, which implies that the sequence $\{z_j\}_{j\in\mathbb{N}}$ is bounded. From Lemma (5) and Equation (5), the sequence $\{\delta_j\}_{j\in\mathbb{N}}$ is bounded. Let $h_j = \|x_j - x^*\|^2$, then

$$\delta_j = (j-1)(h_j - h_{j-1}) + (\alpha-1)h_{j-1}$$

and

$$\begin{aligned}\delta_{j+1} - \delta_j &= j(h_{j+1} - h_j) + (\alpha-1)h_j - (j-1)(h_j - h_{j-1}) - (\alpha-1)h_{j-1} \\ &= j(h_{j+1} - h_j) + (\alpha-1)(h_j - h_{j-1}) - (j-1)(h_j - h_{j-1}) \\ &= j(h_{j+1} - h_j) - (j-\alpha)(h_j - h_{j-1}). \end{aligned} \tag{35}$$

From Equation (18), we have

$$F(x_{j+1}) \leq F(x^*) + (1-\rho\lambda)\langle G_\lambda(y_j), y_j - x^*\rangle - \frac{\lambda(1-\rho\lambda)}{2}\|G_\lambda(y_j)\|^2,$$

and

$$F(x_{j+1}) - F(x^*) \leq \frac{\lambda(\rho\lambda-1)}{2}\|G_\lambda(y_j)\|^2 + (1-\rho\lambda)\langle G_\lambda(y_j), y_j - x^*\rangle.$$

Then

$$\begin{aligned} 0 &\leq -\frac{\lambda}{2}\|G_\lambda(y_j)\|^2 + \langle G_\lambda(y_j), y_j - x^*\rangle \\ &= -\|x_{j+1} - y_j\|^2 + 2\langle y_j - x_{j+1}, y_j - x^*\rangle. \end{aligned} \tag{36}$$

Now

$$\begin{aligned} \|x_{j+1} - x^*\|^2 &= \|(x_{j+1} - y_j) + (y_j - x^*)\|^2 \\ &= \|x_{j+1} - y_j\|^2 + \|y_j - x^*\|^2 + 2\langle x_{j+1} - y_j, y_j - x^*\rangle. \end{aligned}$$

Then Equation (36) becomes

$$0 \leq -\|x_{j+1} - y_j\|^2 + \|x_{j+1} - y_j\|^2 + \|y_j - x^*\|^2 - \|x_{j+1} - x^*\|^2.$$

By solving, we obtain

$$\begin{aligned} \|x_{j+1} - x^*\|^2 &\leq \|y_j - x^*\|^2 \\ &= \|x_j - x^* + \alpha_j^2(x_j - x_{j-1})\|^2 \\ &= \|x_j - x^*\|^2 + \alpha_j^2\|x_j - x_{j-1}\|^2 + 2\alpha_j\langle x_j - x^*, x_j - x_{j-1}\rangle. \end{aligned}$$

Now,

$$\begin{aligned} \|x_{j-1} - x^*\|^2 &= \|(x_j - x^*) - (x_j - x_{j-1})\|^2 \\ &= \|x_j - x^*\|^2 + \|x_j - x_{j-1}\|^2 - 2\langle x_j - x^*, x_j - x_{j-1}\rangle. \end{aligned}$$

From the above calculation, we obtain

$$\begin{aligned} \|x_{j+1} - x^*\|^2 &\leq \|x_j - x^*\|^2 + (\alpha_j^2 + \alpha_j)\|x_j - x_{j-1}\|^2 \\ &\quad + \alpha_j[\|x_j - x^*\|^2 - \|x_{j-1} - x^*\|^2] \\ &\leq \|x_j - x^*\|^2 + 2\|x_j - x_{j-1}\|^2 \\ &\quad + \left(\frac{j-\alpha}{j}\right)[\|x_j - x^*\|^2 - \|x_{j-1} - x^*\|^2]. \end{aligned}$$

Then, we get

$$j[\|x_{j+1}-x^*\|^2-\|x_j-x^*\|^2]-(j-\alpha)[\|x_j-x^*\|^2-\|x_{j-1}-x^*\|^2]\leq 2j\|x_j-x_{j-1}\|^2. \tag{37}$$

From Equation (35) and (37), we get

$$\delta_{j+1}-\delta_j\leq 2j\|x_j-x_{j-1}\|^2. \tag{38}$$

From Lemma (5), the series $\sum_{j=1}^{+\infty} j\|x_j-x_{j-1}\|^2$ is convergent. Applying Lemma (2), we obtain $\lim_{j\to\infty}\delta_j$ exists. It follows that $\lim_{j\to\infty}\|z_j-x^*\|$ exits. From Lemma (6) and Equation (5), we get $\lim_{j\to\infty}\|x_{j-1}-x^*\|$ exists. Since each sequential cluster point of (x_j) belongs to S, applying Lemma (1), we obtain the sequence $\{x_j\}_{j\in\mathbb{N}}$ converge to a some point in S. □

5 Convergence rate of norm of subgradient

Theorem 9. *Suppose that* (H) *holds,* $\alpha>3$*, and* $\lambda\in\Big(0,\min\{\frac{1}{\rho},\frac{1}{L_{\nabla f}}\}\Big)$*. Let* $\{x_j\}_{j\in\mathbb{N}}$ *be the sequences induced from* APG_α*, then*

$$\min_{1\leq i\leq j}\operatorname{dist}^2(0,\partial F(x_{i+1}))=o\left(\frac{1}{j^2}\right).$$

Proof. From Algorithm $x_{j+1}=\operatorname{Prox}_{\lambda g}(y_j-\lambda\nabla f(y_j))$, we can write

$$0\in\nabla f(y_j)+\partial g(x_{j+1})+\lambda^{-1}(x_{j+1}-y_j). \tag{39}$$

Define

$$w_{j+1}:=\frac{1}{\lambda}(y_j-x_{j+1})+\nabla f(x_{j+1})-\nabla f(y_j)\in\partial(f+g)(x_{j+1})=\partial F(x_{j+1}), \tag{40}$$

using Lipschitz continuous of ∇f, we get

$$\begin{aligned}\|w_{j+1}\|^2&=\|\frac{1}{\lambda}(y_j-x_{j+1})+\nabla f(x_{j+1})-\nabla f(y_j)\|^2\\&\leq\frac{2}{\lambda}\|y_j-x_{j+1}\|^2+2\|\nabla f(x_{j+1})-\nabla f(y_j)\|^2\\&\leq 2(\frac{1}{\lambda}+L^2_{\nabla f})\|x_{j+1}-y_j\|^2.\end{aligned} \tag{41}$$

Since $\alpha_j\leq 1$, we get

$$\begin{aligned}\|x_{j+1}-y_j\|^2&=\|x_{j+1}-x_j-\alpha_j(x_j-x_{j-1})\|^2\\&\leq 2\|x_{j+1}-x_j\|^2+2\|\alpha_j(x_j-x_{j-1})\|^2\\&\leq 2\|x_{j+1}-x_j\|^2+2\|x_j-x_{j-1}\|^2.\end{aligned} \tag{42}$$

From Equation (40), we obtain

$$\mathrm{dist}^2(0, \partial F(x_{j+1})) \leq \|w_{j+1}\|^2. \tag{43}$$

Then

$$\begin{aligned}\sum_{j=1}^{+\infty} j\mathrm{dist}^2(0, \partial F(x_{j+1})) &\leq \sum_{j=1}^{+\infty} j\|w_{j+1}\|^2 \\ &\overset{(41)}{\leq} \sum_{j=1}^{+\infty} 2j(\frac{1}{\lambda} + L_{\nabla f}^2)\|x_{j+1} - y_j\|^2 \\ &\overset{(42)}{\leq} (\frac{4}{\lambda} + L_{\nabla f}^2)\sum_{j=1}^{+\infty} j(\|x_{j+1} - x_j\|^2 + \|x_j - x_{j-1}\|^2). \end{aligned} \tag{44}$$

From Lemma (5), Equation (44) becomes

$$\sum_{j=1}^{+\infty} j\mathrm{dist}^2(0, \partial F(x_{j+1})) < +\infty. \tag{45}$$

Utilizing [Theorem 3.1, [9]] and [Theorem 5.1, [14]], there exists a constant $C > 0$ such that

$$\lim_{j\to+\infty} \left(j^2 \min_{1\leq i\leq j} \mathrm{dist}^2(0, \partial F(x_{i+1}))\right) \leq C \lim_{j\to+\infty} \sum_{i=\lfloor \frac{j}{2}\rfloor}^{j} i\mathrm{dist}^2(0, \partial F(x_{i+1})) = 0. \tag{46}$$

Thus,

$$\min_{1\leq i\leq j} \mathrm{dist}^2(0, \partial F(x_{i+1})) = o\left(\frac{1}{j^2}\right). \tag{47}$$

□

6 Conclusion

We investigate the APG_α algorithm for a class of weakly convex optimization problems. We prove that the objective residual satisfies the convergence rate $F(x_j) - F(x^*) = \mathcal{O}\left(\frac{1}{j^2}\right)$ whenever $\alpha \geq 3$. Moreover, for $\alpha > 3$, the rate improves to $F(x_j) - F(x^*) = o\left(\frac{1}{j^2}\right)$, and the sequence generated by the APG_α algorithm converges to a minimum point. In addition, we establish the convergence estimate $\min_{1\leq i\leq j} \mathrm{dist}^2\big(0, \partial F(x_{i+1})\big) = o\left(\frac{1}{j^2}\right)$, which characterizes the decay of the squared norm of the subdifferential residual.

References

[1] H. Attouch and A. Cabot. Convergence rates of inertial forward-backward algorithms. *SIAM J. Optim.*, 28:849–874, 2018.

[2] H. Attouch, Z. Chbani, J. Peypouquet, and P. Redont. Fast convergence of inertial dynamics. *Math. Program.*, 168(1):123–175, 2018.

[3] H. Attouch and J. Peypouquet. The rate of convergence of nesterov's method is faster than $1/t^2$. *SIAM J. Optim.*, 26:1824–1834, 2016.

[4] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, Cham, 2017.

[5] Heinz H. Bauschke and Walaa M. Moursi. Understanding fista's weak convergence: A step-by-step introduction to the 2025 milestone. *arXiv preprint arXiv:2601.15398*, 2026.

[6] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2:183–202, 2009.

[7] Ewa Bednarczuk, Giovanni Bruccola, Gabriele Scrivanti, et al. Forward–backward algorithms for weakly convex problems. *Applied Mathematics & Optimization', volume=*.

[8] A. Böhm and S. J. Wright. Variable smoothing for weakly convex composite functions. *J. Optim. Theory Appl.*, 188(3):628–649, 2021.

[9] S. Chen, B. Shi, and Y.-X. Yuan. Gradient norm minimization of nesterov acceleration: $o(1/k^3)$. *arXiv preprint arXiv:2209.08862*, 2022.

[10] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm. *Commun. Pure Appl. Math.*, 57(11):1413–1457, 2004.

[11] Shih-Ping Han and Gang Lou. A parallel algorithm for a class of convex programs. *SIAM Journal on Control and Optimization*, 26(2):345–355, 1988.

[12] T. Hoheisel, M. Laborde, and A. Oberman. A regularization interpretation of the proximal point method for weakly convex functions. *J. Dyn. Games*, 7(1):79–96, 2020.

[13] P. D. Khanh, B. S. Mordukhovich, V. T. Phat, and D. B. Tran. Inexact proximal methods for weakly convex functions. *J. Global Optim.*, 91:611–646, 2025.

[14] B. Li, B. Shi, and Y.-X. Yuan. Proximal subgradient norm minimization of ista and fista. *Appl. Comput. Harmon. Anal.*, page 101848, 2025.

[15] F. Y. Liao and Y. Zheng. A proximal descent method for minimizing weakly convex optimization. *arXiv preprint arXiv:2509.02804*, 2025.

[16] P. L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM J. Numer. Anal.*, 16(6):964–979, 1979.

[17] B. Martinet. Régularisation d'inéquations variationnelles par approximations successives. *Rev. Fr. Inform. Rech. Oper.*, 4:154–158, 1970.

[18] Y. Nesterov. A method for unconstrained convex minimization. *Dokl. Akad. Nauk SSSR*, 269:543–547, 1983.

[19] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer, New York, NY, 2013.

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

[21] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM J. Control Optim.*, 14:877–898, 1976.