Unifying restart accelerated gradient and proximal bundle methods

Jiaming Liang *

January 7, 2025

Abstract

This paper presents a novel restarted version of Nesterov's accelerated gradient method and establishes its optimal iteration-complexity for solving convex smooth composite optimization problems. The proposed restart accelerated gradient method is shown to be a specific instance of the accelerated inexact proximal point framework introduced in [28]. Furthermore, this work examines the proximal bundle method within the inexact proximal point framework, demonstrating that it is an instance of the framework. Notably, this paper provides new insights into the underlying algorithmic principle that unifies two seemingly disparate optimization methods, namely, the restart accelerated gradient and the proximal bundle methods.

Key words. convex composite optimization, accelerated gradient method, proximal bundle method, proximal point method, optimal iteration-complexity

AMS subject classifications. 49M37, 65K05, 68Q25, 90C25, 90C30, 90C60

1 Introduction

Nesterov's accelerated gradient methods [5, 30, 31, 33] have been widely employed to solve convex smooth composite optimization (CSCO) problems of the form

$$\phi_* := \min_{x\in\mathbb{R}^n} \left\{\phi(x) := f(x) + h(x)\right\}, \tag{1}$$

where f is typically a convex and L-smooth function, and h is a convex and possibly nonsmooth function with a simple proximal mapping, satisfying $\operatorname{dom} h \subset \operatorname{dom} f$. Extensive research has been dedicated to developing a theoretical understanding of these methods [1, 6, 13, 17, 35, 37], extending their scopes [7, 9, 14, 19, 20, 22, 24], and improving their practical performance [5, 25, 27, 32, 36].

Among many approaches to enhance the convergence of accelerated gradient methods in practice, in particular to suppress the oscillating behavior, the restart technique has shown remarkable improvement in the context of CSCO [2, 3, 4, 8, 10, 27, 29, 34, 35, 36]. The most natural restart scheme is to restart the accelerated gradient method after a fixed number of iterations, and an optimal fixed restart scheme is presented in [29]. Various adaptive restart schemes have also been explored in the literature. Paper [34] proposes a function restart scheme (i.e., it restarts when the function value increases) and a gradient restart scheme (i.e., it restarts when the momentum term and the negative gradient make an obtuse angle). However, this paper only provides a heuristic discussion but no non-asymptotic convergence rate. Inspired by an ODE interpretation of Nesterov's

*Goergen Institute for Data Science and Department of Computer Science, University of Rochester, Rochester, NY 14620 (email: `jiaming.liang@rochester.edu`).

accelerated gradient method, [35] develops speed restart schemes in both continuous and discrete times. In discrete time, the scheme restarts when $\|x_k - x_{k-1}\| < \|x_{k-1} - x_{k-2}\|$. The paper presents a convergence analysis in continuous time, while some constants are just shown to exist. Based on a restart condition for estimating the strong convexity parameter, [36] proposes parameter-free restarted accelerated gradient methods for strongly convex optimization.

Leveraging the A-HPE framework proposed in [28], we develop a novel restart accelerated composite gradient (ACG) method and establish the same optimal iteration-complexity as the accelerated gradient method for CSCO problems (1). A-HPE is a generic framework built on an acceleration scheme. Two specific implementations of A-HPE are given in [28], namely, FISTA [5] (a variant of accelerated gradient method) and an accelerated Newton proximal extragradien method, which is the first optimal second-order method. Our main contribution is to show restart ACG is another instance of A-HPE and establish the optimal complexity bound of restart ACG.

Another contribution of the paper is that it demonstrates the modern proximal bundle (MPB) method [21, 23], an optimal method for solving convex nonsmooth composite optimization (CNCO) problems, is indeed an instance of the HPE framework [11], which can be understood as the non-accelerated counterpart of A-HPE for CNCO. As a result, MPB can be viewed as a restarted version of the cutting-plane method. Building upon the novel perspectives of restart ACG and MPB as multi-step implementations of A-HPE and HPE, respectively, this paper fconcludes by offering a qualitative analysis to elucidate the superior practical performance of restart ACG and MPB in comparison to their corresponding single-step counterparts, FISTA and the subgradient method.

1.1 Basic definitions and notation

A proper function $f : \mathbb{R}^n \to (-\infty, +\infty]$ is μ-strongly convex for some $\mu > 0$ if for every $x, y \in \operatorname{dom} f$ and $\lambda \in [0, 1]$,

$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda)f(y) - \frac{\lambda(1-\lambda)\mu}{2}\|x-y\|^2.$$

For $\varepsilon \ge 0$, the *ε-subdifferential* of f at $x \in \operatorname{dom} f$ is denoted by

$$\partial_\varepsilon f(x) := \{s \in \mathbb{R}^n : f(y) \ge f(x) + \langle s, y-x\rangle - \varepsilon, \forall y \in \mathbb{R}^n\}.$$

We denote the subdifferential of f at $x \in \operatorname{dom} f$ by $\partial f(x)$, which is the set $\partial_0 f(x)$ by definition. For a given subgradient $f'(x) \in \partial f(x)$, we denote the linearization of convex function f at x by $\ell_f(\cdot; x)$, which is defined as

$$\ell_f(\cdot; x) := f(x) + \langle f'(x), \cdot - x\rangle\rangle.$$

2 Restart ACG

This section first reviews an ACG variant used in the paper, then presents the restart ACG method, and finally provides the complexity analysis of restart ACG and shows that it is optimal for CSCO.

2.1 Review of an ACG variant

In this subsection, we consider

$$\min\{\psi(x) := g(x) + h(x) : x \in \mathbb{R}^n\}, \tag{2}$$

where g is μ-strongly convex and $(L+\mu)$-smooth, and h is as in (1). We describe an ACG variant tailored to (2) and present some basic results regarding the ACG variant.

Algorithm 1 Accelerated Composite Gradient

Initialize: given initial point $x_0 \in \mathbb{R}^n$, set $A_0 = 0$, $\tau_0 = 1/L$, and $y_0 = x_0$

for $j = 0, 1, \cdots$ **do**

1. Compute

$$\tau_{j+1} = \tau_j + \frac{\mu a_j}{L}, \quad a_j = \frac{\tau_j + \sqrt{\tau_j^2 + 4\tau_j A_j}}{2}, \quad A_{j+1} = A_j + a_j, \quad \tilde{x}_j = \frac{A_j}{A_{j+1}} y_j + \frac{a_j}{A_{j+1}} x_j; \tag{3}$$

2. Compute

$$\tilde{y}_{j+1} = \underset{u \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ \ell_g(u; \tilde{x}_j) + h(u) + \frac{L+\mu}{2}\|u - \tilde{x}_j\|^2 \right\}, \tag{4}$$

$$y_{j+1} \in \operatorname{Argmin} \left\{ \psi(u) : u \in \{y_j, \tilde{y}_{j+1}\} \right\}, \tag{5}$$

$$x_{j+1} = \frac{(L+\mu) a_j \tilde{y}_{j+1} - \frac{A_j a_j L}{A_{j+1}} y_j}{A_{j+1}\mu + 1}. \tag{6}$$

end for

The following three lemmas are standard results for ACG. Therefore, we omit their proofs in this subsection but provide them in the Appendix for completeness.

Lemma 2.1. *The following statements hold for every $j \geq 0$:*

(a) $\tau_j = (1 + \mu A_j)/L$;

(b) $A_{j+1}\tau_j = a_j^2$.

Lemma 2.2. *Define $\Gamma_0 \equiv 0$ and*

$$\tilde{\gamma}_j(\cdot) := \ell_g(\cdot; \tilde{x}_j) + h(\cdot) + \frac{\mu}{2}\|\cdot - \tilde{x}_j\|^2, \tag{7}$$

$$\gamma_j(\cdot) := \tilde{\gamma}_j(\tilde{y}_{j+1}) + L\langle \tilde{x}_j - \tilde{y}_{j+1}, \cdot - \tilde{y}_{j+1}\rangle + \frac{\mu}{2}\|\cdot - \tilde{y}_{j+1}\|^2, \tag{8}$$

$$\Gamma_{j+1}(\cdot) := \frac{A_j \Gamma_j(\cdot) + a_j \gamma_j(\cdot)}{A_{j+1}}. \tag{9}$$

Then, the following statements hold for every $j \geq 0$:

(a) $\gamma_j \leq \tilde{\gamma}_j \leq \psi$, $\tilde{\gamma}_j(\tilde{y}_{j+1}) = \gamma_j(\tilde{y}_{j+1})$,

$$\min_{u \in \mathbb{R}^n} \left\{ \tilde{\gamma}_j(u) + \frac{L}{2}\|u - \tilde{x}_j\|^2 \right\} = \min_{u \in \mathbb{R}^n} \left\{ \gamma_j(u) + \frac{L}{2}\|u - \tilde{x}_j\|^2 \right\}, \tag{10}$$

and these minimization problems have $\tilde{y}_{j+1}$ as a unique optimal solution;

(b) γ_j *and* Γ_j *are μ-strongly convex quadratic functions;*

(c) $x_j = \underset{u \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ A_j \Gamma_j(u) + \|u - x_0\|^2/2 \right\}$.

Lemma 2.3. *For every $j \geq 0$, we have*

$$A_j \psi(y_j) \leq \min_{u \in \mathbb{R}^n} \left\{ A_j \Gamma_j(u) + \frac{1}{2}\|u - x_0\|^2 \right\}. \tag{11}$$

The following lemma is the same as Proposition 1(c) of [27] and hence we omit the proof.

Lemma 2.4. *For every $j \geq 1$, we have*

$$A_j \geq \max\left\{\frac{j^2}{4L}, \frac{1}{L}\left(1+\frac{\sqrt{\mu}}{2\sqrt{L}}\right)^{2(j-1)}\right\}.$$

2.2 The algorithm

Subsection 2.1 outlines a single-loop ACG method. Designing a restart ACG requires repeatedly invoking Algorithm 1 as a subroutine within a double-loop algorithm. This approach aligns naturally with the proximal point method (PPM), which iteratively solves a sequence of proximal subproblems using a recursive subroutine. More precisely, we adopt the A-HPE framework from [28] as an inexact PPM. Within each loop of A-HPE, Algorithm 1 is employed to solve a certain proximal subproblem, while between successive loops, an acceleration scheme from A-HPE is applied. Consequently, the proposed restart ACG method (i.e., Algorithm 2) can be described as "doubly accelerated."

Algorithm 2 Restart ACG

Initialize: given initial point $w_0 \in \operatorname{dom} h$ and stepsize $\lambda > 0$, set $z_0 = w_0$ and $B_0 = 0$

for $k = 1, 2, \cdots$ **do**

1. Compute

$$b_k = \frac{\lambda + \sqrt{\lambda^2 + 4\lambda B_{k-1}}}{2}, \quad B_k = B_{k-1} + b_k, \quad \tilde{z}_k = \frac{B_{k-1}}{B_k} w_{k-1} + \frac{b_k}{B_k} z_{k-1};$$

2. Call Algorithm 1 with

$$x_0 = \tilde{z}_k, \quad \psi = \phi + \frac{1}{2\lambda}\|\cdot - \tilde{z}_k\|^2, \quad g = f + \frac{1}{2\lambda}\|\cdot - \tilde{z}_k\|^2 \tag{12}$$

to find a triple $(\tilde{w}_k, u_k, \eta_k)$ satisfying

$$u_k \in \partial_{\eta_k}\phi(\tilde{w}_k), \tag{13}$$

$$\|\lambda u_k + \tilde{w}_k - \tilde{z}_k\|^2 + 2\lambda\eta_k \leq 0.9\|\tilde{z}_k - \tilde{w}_k\|^2; \tag{14}$$

3. Compute $z_k = z_{k-1} - b_k u_k$ and $w_k \in \operatorname{Argmin}\{\phi(u) : u \in \{w_{k-1}, \tilde{w}_k\}\}$.

end for

From the "inner loop" perspective, Algorithm 2 keeps performing ACG iterations until (13) and (14) are satisfied, and then restarts ACG with the initialization as in (12). From the "outer loop" perspective, Algorithm 2 is an instance of the A-HPE framework of [28] for solving (1) with $\lambda_k = \lambda$ for every $k \geq 1$ and ACG as its subroutine for step 2. The constant 0.9 in (14) is not critical and can be any arbitrary number within the interval $(0, 1)$. With minor modification, such as generalizing f to ϕ, the results in Section 3 of [28] are applicable to this paper. Consequently, Theorem 3.8 of [28] also holds. For completeness, we state the theorem below in our context without providing a proof.

Theorem 2.5. *For every $k \geq 1$, we have*

$$\phi(w_k) - \phi_* \leq \frac{2d_0^2}{\lambda k^2}.$$

We next provide some PPM interpretations of conditions (13) and (14). Each call to ACG in step 2 of Algorithm 2 approximately solves the proximal subproblem

$$\hat{z}_k = \underset{u\in\mathbb{R}^n}{\operatorname{argmin}} \left\{\phi(u) + \frac{1}{2\lambda}\|u - \tilde{z}_k\|^2\right\} \tag{15}$$

such that criteria (13) and (14) are satisfied. They can be equivalently written as

$$\lambda u_k + \tilde{w}_k - \tilde{z}_k \in \partial_{\lambda\eta_k}\left(\lambda\phi(\cdot) + \frac{1}{2}\|\cdot - \tilde{z}_k\|^2\right)(\tilde{w}_k)$$
$$\|\lambda u_k + \tilde{w}_k - \tilde{z}_k\|^2 + 2\lambda\eta_k \le 0.9\|\tilde{z}_k - \tilde{w}_k\|^2.$$

Alternatively, we can derive a relative solution accuracy guarantee on (15). It follows from (13) that for every $u \in \mathbb{R}^n$,

$$\phi(u) \ge \phi(\tilde{w}_k) + \langle u_k, u - \tilde{w}_k\rangle - \eta_k,$$

which together with (14) implies that

$$\begin{aligned} 0.9\|\tilde{z}_k - \tilde{w}_k\|^2 &\ge \|\lambda u_k + \tilde{w}_k - \tilde{z}_k\|^2 + 2\lambda\eta_k \\ &\ge \|\lambda u_k + \tilde{w}_k - \tilde{z}_k\|^2 + 2\lambda[\phi(\tilde{w}_k) + \langle u_k, u - \tilde{w}_k\rangle - \phi(u)]. \end{aligned}$$

Taking $u = \hat{z}_k$, which is the exact solution to (15), we have

$$\begin{aligned} 0.9\|\tilde{z}_k - \tilde{w}_k\|^2 &\ge \|\lambda u_k\|^2 + 2\lambda\left[\phi(\tilde{w}_k) + \frac{1}{2\lambda}\|\tilde{w}_k - \tilde{z}_k\|^2 - \phi(\hat{z}_k) + \langle u_k, \hat{z}_k - \tilde{z}_k\rangle\right] \\ &\ge \|\lambda u_k + \hat{z}_k - \tilde{z}_k\|^2 + 2\lambda\left[\phi(\tilde{w}_k) + \frac{1}{2\lambda}\|\tilde{w}_k - \tilde{z}_k\|^2 - \phi(\hat{z}_k) - \frac{1}{2\lambda}\|\hat{z}_k - \tilde{z}_k\|^2\right]. \end{aligned}$$

Therefore,

$$\phi(\tilde{w}_k) + \frac{1}{2\lambda}\|\tilde{w}_k - \tilde{z}_k\|^2 - \phi(\hat{z}_k) - \frac{1}{2\lambda}\|\hat{z}_k - \tilde{z}_k\|^2 \le \frac{0.9}{2\lambda}\|\tilde{z}_k - \tilde{w}_k\|^2,$$

indicating that $\tilde{w}_k$ is an approximate solution to (15) with respect to the relative accuracy given above.

2.3 Complexity analysis

This subsection provides the complexity analysis of Algorithm 2. It first establishes the iteration-complexity for ACG to find a triple $(\tilde{w}_k, u_k, \eta_k)$ satisfying (13) and (14).

To set the stage, recall the initialization in (12), we note that ACG invoked at step 2 of Algorithm 2 solves (2) with

$$\psi(\cdot) = \phi(\cdot) + \frac{1}{2\lambda}\|\cdot - x_0\|^2, \quad g(\cdot) = f(\cdot) + \frac{1}{2\lambda}\|\cdot - x_0\|^2, \quad \mu = \frac{1}{\lambda}. \tag{16}$$

Lemma 2.6. *Define*

$$\phi_j(\cdot) := \Gamma_j(\cdot) - \frac{1}{2\lambda}\|\cdot - x_0\|^2, \tag{17}$$

where Γ_j is as in (9). Then, for every $j \ge 1$, we have

$$\hat{v}_j \in \partial_{\varepsilon_j}\phi(y_j), \quad \|\lambda\hat{v}_j + y_j - x_0\|^2 + 2\lambda\varepsilon_j = \|\lambda v_j\|^2 + 2\lambda[\psi(y_j) - \Gamma_j(x_j)] \tag{18}$$

where

$$v_j := \frac{x_0 - x_j}{A_j}, \quad \hat{v}_j := \frac{x_0 - x_j}{A_j} + \frac{x_0 - x_j}{\lambda}, \quad \varepsilon_j := \phi(y_j) - \phi_j(x_j) - \langle\hat{v}_j, y_j - x_j\rangle. \tag{19}$$

Proof: It follows from the optimality condition of (6) and the definition of ϕ_j in (17) that

$$v_j = \frac{x_0 - x_j}{A_j} \overset{(6)}{\in} \partial\Gamma_j(x_j) \overset{(17)}{=} \partial\phi_j(x_j) + \frac{1}{\lambda}(x_j - x_0),$$

which together with (19) implies that $\hat{v}_j \in \partial\phi_j(x_j)$. This inclusion, the fact that $\phi \geq \phi_j$, and the definition of ε_j in (19) yields that for every $u \in \mathbb{R}^n$,

$$\phi(u) \geq \phi_j(u) \geq \phi_j(x_j) + \langle \hat{v}_j, u - x_j\rangle = \phi(y_j) + \langle \hat{v}_j, u - y_j\rangle - \varepsilon_j,$$

and hence that the inclusion in (18) holds. Noting that $\hat{v}_j = v_j + (x_0 - x_j)/\lambda$, we have

$$\begin{aligned}
&\|\lambda\hat{v}_j + y_j - x_0\|^2 + 2\lambda\varepsilon_j \\
=&\|\lambda v_j + y_j - x_j\|^2 + 2\lambda\left[\phi(y_j) - \phi_j(x_j)\right] - 2\lambda\langle v_j, y_j - x_j\rangle - 2\langle x_0 - x_j, y_j - x_j\rangle \\
=&\|\lambda v_j\|^2 + 2\lambda\left[\phi(y_j) + \frac{1}{2\lambda}\|y_j - x_0\|^2 - \phi_j(x_j) - \frac{1}{2\lambda}\|x_j - x_0\|^2\right].
\end{aligned}$$

Finally, the identity in (18) follows from the above one and the definitions of ψ and ϕ_j in (16) and (17), respectively. ■

Lemma 2.7. *Define*

$$\hat{x}_j := \text{argmin}\,\{\Gamma_j(u) : u \in \mathbb{R}^n\}, \tag{20}$$

where Γ_j is as in (9). *Assuming that $A_j \geq 3\lambda$, then the following statements hold for every $j \geq 1$:*

a)

$$\psi(y_j) - \Gamma_j(\hat{x}_j) \leq \frac{1}{2A_j}\|\hat{x}_j - x_0\|^2 \leq \frac{1}{A_j - 2\lambda}\|y_j - x_0\|^2; \tag{21}$$

b)

$$\|v_j\| \leq \frac{3\|y_j - x_0\|}{2A_j}. \tag{22}$$

Proof: a) Using Lemma 2.3, we have

$$\psi(y_j) \overset{(11)}{\leq} \min_{u\in\mathbb{R}^n}\left\{\Gamma_j(u) + \frac{1}{2A_j}\|u - x_0\|^2\right\} \leq \Gamma_j(\hat{x}_j) + \frac{1}{2A_j}\|\hat{x}_j - x_0\|^2,$$

and hence the first inequality in (21) holds. Since Γ_j is as in (9), it follows from Lemma 2.2(b) and the fact that $\mu = 1/\lambda$ (see (16)) that Γ_j is λ^{-1}-strongly convex. This observation, the above inequality, and the definition of $\hat{x}_j$ in (20) thus imply that for every $u \in \mathbb{R}^n$,

$$\psi(y_j) - \frac{1}{2A_j}\|\hat{x}_j - x_0\|^2 \leq \Gamma_j(\hat{x}_j) \leq \Gamma_j(u) - \frac{1}{2\lambda}\|u - \hat{x}_j\|^2.$$

Taking $u = y_j$ in the above inequality and using the fact that $\Gamma_j \leq \psi$, we obtain

$$\|y_j - \hat{x}_j\|^2 \leq \frac{\lambda}{A_j}\|\hat{x}_j - x_0\|^2.$$

Using the above inequality, the triangle inequality, and the fact that $(a+b)^2 \leq 2(a^2 + b^2)$, we have

$$\|\hat{x}_j - x_0\|^2 \leq 2(\|\hat{x}_j - y_j\|^2 + \|y_j - x_0\|^2) \leq \frac{2\lambda}{A_j}\|\hat{x}_j - x_0\|^2 + 2\|y_j - x_0\|^2,$$

and hence the second inequality in (21) follows.

b) It follows from Lemma 2.3 and the fact that Γ_j is λ^{-1}-strongly convex that

$$\psi(y_j) + \frac{1}{2}\left(\frac{1}{\lambda} + \frac{1}{A_j}\right)\|u - x_j\|^2 \leq \min_{u\in\mathbb{R}^n}\left\{\Gamma_j(u) + \frac{1}{2A_j}\|u - x_0\|^2\right\} + \frac{1}{2}\left(\frac{1}{\lambda} + \frac{1}{A_j}\right)\|u - x_j\|^2$$
$$\leq \Gamma_j(u) + \frac{1}{2A_j}\|u - x_0\|^2.$$

Taking $u = y_j$ in the above inequality and using the fact $\Gamma_j \leq \psi$, we have

$$\frac{1}{2}\left(\frac{1}{\lambda} + \frac{1}{A_j}\right)\|y_j - x_j\|^2 \leq \Gamma_j(y_j) - \psi(y_j) + \frac{1}{2A_j}\|y_j - x_0\|^2 \leq \frac{1}{2A_j}\|y_j - x_0\|^2,$$

and hence

$$\|y_j - x_j\|^2 \leq \frac{\lambda}{\lambda + A_j}\|y_j - x_0\|^2 \leq \frac{1}{4}\|y_j - x_0\|^2$$

where the second inequality is due to $A_j \geq 3\lambda$. Finally, (22) immediately follows from the definition of v_j in (19), the triangle inequality, and the above inequality. ■

Proposition 2.8. *The number of iterations performed by ACG to find a triple* $(\tilde{w}_k, u_k, \eta_k)$ *satisfying* (13) *and* (14) *is at most*

$$\min\left\{2\sqrt{6\lambda L}, \left(\frac{1}{2} + \sqrt{\lambda L}\right)\ln(6\lambda L)\right\}. \tag{23}$$

Proof: It is easy to verify that (23) and Lemma 2.4 imply that $A_j \geq 6\lambda$. Consider the first j such that $A_j \geq 6\lambda$, then we prove that

$$u_k = \hat{v}_j, \quad \eta_k = \varepsilon_j, \quad \tilde{w}_k = y_j \tag{24}$$

satisfy (13) and (14). First, it follows from Lemma 2.6 that the inclusion in (18) is equivalent to (13) with the assignment (24). Using the identity in (18) and Lemma 2.7 we have

$$\|\lambda\hat{v}_j + y_j - x_0\|^2 + 2\lambda\varepsilon_j = \|\lambda v_j\|^2 + 2\lambda[\psi(y_j) - \Gamma_j(x_j)]$$
$$\overset{(21),(22)}{\leq} \frac{9\lambda^2\|y_j - x_0\|^2}{4A_j^2} + \frac{2\lambda}{A_j - 2\lambda}\|y_j - x_0\|^2$$
$$\leq \left(\frac{1}{16} + \frac{1}{2}\right)\|y_j - x_0\|^2,$$

where the last inequality is due to the fact that $A_j \geq 6\lambda$. Hence, (14) also holds in view of (24). ■

Now we are ready to present the main result of the paper.

Theorem 2.9. *Given* $\bar{\varepsilon} > 0$*, assuming that* λ *satisfies* $1/L \leq \lambda \leq d_0^2/\bar{\varepsilon}$*, then the total iteration-complexity of Algorithm 2 to find a* $\bar{\varepsilon}$*-solution to* (1) *is* $\mathcal{O}(\sqrt{L}d_0/\sqrt{\bar{\varepsilon}})$.

Proof: It follows from Theorem 2.5 that to find a $\bar{\varepsilon}$-solution, the number of calls to ACG is at most $\sqrt{2}d_0/\sqrt{\lambda\bar{\varepsilon}}$. Therefore, the conclusion of the theorem immediately follows from Proposition 2.8 and the assumption on λ. ■

3 Connections between restart ACG and MPB

Inspired by the interpretation of restart ACG as an instance of A-HPE, we revisit the MPB method and show that it is indeed an instance of the HPE framework for CNCO. To begin with, we present below the HPE framework, adapted from Framework 1 of [11], for solving (1) where f is M-Lipschitz continuous instead of being smooth and h is as in (1).

Algorithm 3 HPE framework

Initialize: given initial point $w_0 \in \operatorname{dom} h$, stepsize $\lambda > 0$, and tolerance $\delta > 0$

for $k = 1, 2, \cdots$ **do**

1. Find a triple $(\tilde{w}_k, u_k, \eta_k)$ satisfying

$$u_k \in \partial_{\eta_k} \phi(\tilde{w}_k), \tag{25}$$

$$\|\lambda u_k + \tilde{w}_k - w_{k-1}\|^2 + 2\lambda\eta_k \le 2\lambda\delta; \tag{26}$$

2. Compute $w_k = w_{k-1} - \lambda u_k$.

end for

3.1 MPB as an instance of HPE

This subsection shows that the MPB method is an instance of the HPE framework. We begin with a brief review of MPB. A key distinction of MPB from classical PB methods [15, 16, 26, 38] lies in its incorporation of the PPM. MPB approximately solves a sequence of proximal subproblem of the form

$$\min_{u\in\mathbb{R}^n} \left\{ \psi(u) := \phi(u) + \frac{1}{2\lambda} \|u - w_{k-1}\|^2 \right\}. \tag{27}$$

Letting $x_0 = w_{k-1}$ be the initial point of the subroutine for solving (27), MPB iteratively solves

$$x_j = \operatorname*{argmin}_{u\in\mathbb{R}^n} \left\{ \Gamma_j(u) + h(u) + \frac{1}{2\lambda}\|u - x_0\|^2 \right\}, \tag{28}$$

where Γ_j is a bundle model underneath f. Details about various models and a unifying framework underlying them are discussed in [23]. MPB keeps refining Γ_j and solving x_j through (28), until a criterion $t_j = \psi(\tilde{x}_j) - m_j \le \delta$ is met, where

$$m_j = \Gamma_j(x_j) + h(x_j) + \frac{1}{2\lambda}\|x_j - x_0\|^2, \quad \tilde{x}_j \in \operatorname{Argmin}\{\psi(u) : u \in \{x_0, x_1, \ldots, x_j\}\}. \tag{29}$$

As explained in [18], the criterion $t_j \le \delta$ indicates that a primal-dual solution to (27) with primal-dual gap bounded by δ is obtained. It also implies that $\tilde{x}_j$ is a δ-solution to (27) (see also [21]). Once the condition $t_j \le \delta$ is met, MPB updates the prox center to $w_k = x_j$, resets the bundle model Γ_j from scratch, and proceeds to solve (27) with w_{k-1} replaced by w_k.

Lemma 3.1. *Given $x_0 = w_{k-1}$, the MPB method is an instance of the HPE framework with*

$$w_k = x_j, \quad \tilde{w}_k = \tilde{x}_j, \quad u_k = \frac{x_0 - x_j}{\lambda}, \quad \eta_k = \phi(\tilde{x}_j) - (\Gamma_j + h)(x_j) + \frac{1}{\lambda}\langle x_0 - x_j, x_j - \tilde{x}_j\rangle, \tag{30}$$

where j is the first iteration index such that the condition $t_j \le \delta$ is met.

Proof: Let $\varepsilon_j = \phi(\tilde{x}_j) - (\Gamma_j + h)(x_j) + \langle x_0 - x_j, x_j - \tilde{x}_j \rangle/\lambda$, we first prove that

$$\frac{x_0 - x_j}{\lambda} \in \partial_{\varepsilon_j} \phi(\tilde{x}_j), \quad \|x_j - \tilde{x}_j\|^2 + 2\lambda\varepsilon_j = 2\lambda t_j. \tag{31}$$

It follows from the optimality condition of (28) that

$$\frac{x_0 - x_j}{\lambda} \in \partial(\Gamma_j + h)(x_j).$$

This inclusion and the assumption that $\Gamma_j \le f$ imply that for every $u \in \mathbb{R}^n$,

$$\phi(u) \ge (\Gamma_j + h)(u) \ge (\Gamma_j + h)(x_j) + \frac{1}{\lambda}\langle x_0 - x_j, u - x_j \rangle.$$

Using the definition of ε_j, we thus have for every $u \in \mathbb{R}^n$,

$$\phi(u) \ge \phi(\tilde{x}_j) + \frac{1}{\lambda}\langle x_0 - x_j, u - \tilde{x}_j \rangle - \varepsilon_j,$$

and hence the inclusion in (31) follows. Next, we show the identity in (31). Recalling that

$$2\lambda t_j = 2\lambda[\psi(\tilde{x}_j) - m_j] \overset{(29)}{=} 2\lambda\phi(\tilde{x}_j) + \|\tilde{x}_j - x_0\|^2 - 2\lambda(\Gamma_j + h)(x_j) - \|x_j - x_0\|^2,$$

after simple algebraic manipulation, we obtain

$$\|x_j - \tilde{x}_j\|^2 + 2\lambda\varepsilon_j = \|x_j - \tilde{x}_j\|^2 + 2\langle x_0 - x_j, x_j - \tilde{x}_j \rangle + 2\lambda[\phi(\tilde{x}_j) - (\Gamma_j + h)(x_j)] = 2\lambda t_j.$$

In view of (30), it is easy to verify that (25) is equivalent to the inclusion in (31). Moreover, it follows from (30) and the identity in (31) that $\|w_k - \tilde{w}_k\|^2 + 2\lambda\eta_k = 2\lambda t_j$, which together with step 2 of Algorithm 3 implies that $\|\lambda u_k + \tilde{w}_k - w_{k-1}\|^2 + 2\lambda\eta_k = 2\lambda t_j$. Finally, MPB satisfies (26) since it terminates solving (27) once the condition $t_j \le \delta$ is met. ∎

3.2 Restart schemes via PPM

A standard scheme of updating the bundle model Γ_j in (28) is a cutting-plane scheme, the approximate solutions x_j and $\tilde{x}_j$ to (27) are obtained via the cutting-plane method. Similar to restart ACG, MPB is also a double-loop algorithm and can be viewed as a restarted version of the cutting-plane method from the "inner loop" perspective.

On the other hand, from the "outer loop" perspective, restart ACG (resp., MPB) is an instance of A-HPE (resp., HPE) employing a multi-step subroutine for solving the proximal subproblem (15) (resp., (27)). As illustrated by Algorithm 1 of [28], a triple $(\tilde{w}_k, u_k, \eta_k)$ satisfying (13) and (14) can be obtained via one step (i.e., a proximal mapping of h) given the stepsize $\lambda \approx 1/L$ is small enough. It is noted at the end of Section 5 of [28] that its Algorithm 1 is equivalent to the well-known FISTA. In contrast, Algorithm 2 of this paper admits relatively large stepsize, i.e., $1/L \le \lambda \le d_0^2/\bar{\varepsilon}$ (see Theorem 2.9), and results in a multi-step subroutine, namely, Algorithm 1, for solving (15). A common feature between Algorithm 2 and FISTA is that they both share the optimal complexity for CSCO problems (1), namely, $\mathcal{O}(\sqrt{L}d_0/\sqrt{\bar{\varepsilon}})$ as in Theorem 2.9.

A similar comparison can be drawn between the MPB method and the subgradient method. MPB allows relatively large stepsize, i.e., $\bar{\varepsilon}/M^2 \le \lambda \le d_0^2/\bar{\varepsilon}$, while the subgradient method only takes small stepsize $\lambda = \bar{\varepsilon}/M^2$. As a result, MPB solves the proximal subproblem (27) via the cutting-plane method as in (28), while the subgradient method always performs only one iteration to solve (27). For CNCO problems, MPB and the subgradient method both have optimal complexity bound $\mathcal{O}(M^2 d_0^2/\bar{\varepsilon}^2)$, however, MPB substantially outperforms the subgradient method in practice.

In summary, the relationship between FISTA (specifically, Algorithm 1 of [28]) and the restart ACG method is analogous to the relationship between the subgradient method and MPB. It is thus understandable that, while restart ACG and FISTA share the same optimal iteration-complexity, the restarted version demonstrates superior performance compared with the latter one. This aligns with the general observation that, in the context of both the A-HPE and HPE frameworks, multi-step implementations consistently outperform their single-step counterparts.

4 Concluding remarks

This paper proposes a novel restarted version of accelerated gradient method, i.e., restart ACG, and establishes it optimal iteration-complexity for solving CSCO. It also demonstrates that the MPB method as an instance of the HPE framework, revealing interesting connections between two seemingly distinct optimal methods, restart ACG and MPB.

Several related questions merit future investigation. It is interesting to develop a restart accelerated gradient method with optimal complexity under the strong convexity assumption. If the strong convexity μ is known, it suffices to study the convergence analysis of A-HPE where h is μ-strongly convex. However, in the absence of prior knowledge about μ, the focus shifts to designing μ-universal methods based on improved analysis of A-HPE, utilizing possible techniques from recent works [12, 36].

References

[1] K. Ahn and S. Sra. Understanding nesterov's acceleration via proximal point method. In *Symposium on Simplicity in Algorithms (SOSA)*. Society for Industrial and Applied Mathematics, 2022.

[2] T. Alamo, P. Krupa, and D. Limon. Gradient based restart FISTA. In *58th IEEE Conference on Decision and Control (CDC)*, pages 3936–3941. IEEE, 2019.

[3] T. Alamo, P. Krupa, and D. Limon. Restart FISTA with global linear convergence. In *18th IEEE European Control Conference (ECC)*, pages 1969–1974. IEEE, 2019.

[4] T. Alamo, P Krupa, and D. Limon. Restart of accelerated first order methods with linear convergence under a quadratic functional growth condition. *IEEE Transactions on Automatic Control*, 67(10):5200–5214, 2022.

[5] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[6] S. Bubeck, Y. T. Lee, and M. Singh. A geometric alternative to nesterov's accelerated gradient descent. *arXiv preprint arXiv:1506.08187*, 2015.

[7] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford. Accelerated methods for nonconvex optimization. *SIAM Journal on Optimization*, 28(2):1751–1772, 2018.

[8] O. Fercoq and Z. Qu. Adaptive restart of accelerated gradient methods under local quadratic growth condition. *IMA Journal of Numerical Analysis*, 39(4):2069–2095, 2019.

[9] S. Ghadimi and G. Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156:59–99, 2016.

[10] P. Giselsson and S. Boyd. Monotonicity and restart in fast gradient methods. In *53rd IEEE Conference on Decision and Control*, pages 5058–5063. IEEE, 2014.

[11] M. L. N. Gonçalves, J. G. Melo, and R. D. C. Monteiro. Improved pointwise iteration-complexity of a regularized ADMM and of a regularized non-Euclidean hpe framework. *SIAM Journal on Optimization*, 27(1):379–407, 2017.

[12] V. Guigues, J. Liang, and R. D. C. Monteiro. Universal subgradient and proximal bundle methods for convex and strongly convex hybrid composite optimization. *arXiv preprint arXiv:2407.10073*, 2024.

[13] W. Krichene, A. Bayen, and P. L. Bartlett. Accelerated mirror descent in continuous and discrete time. *Advances in neural information processing systems*, 28, 2015.

[14] G. Lan, Z. Lu, and R. D. C. Monteiro. Primal-dual first-order methods with $\mathcal{O}(1/\epsilon)$ iteration-complexity for cone programming. *Mathematical Programming*, 126(1):1–29, 2011.

[15] C. Lemaréchal. An extension of davidon methods to non differentiable problems. In *Nondifferentiable optimization*, pages 95–109. Springer, 1975.

[16] C. Lemaréchal. Nonsmooth optimization and descent methods. 1978.

[17] L. Lessard, B. Recht, and A. Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

[18] J. Liang. Primal-dual proximal bundle and conditional gradient methods for convex problems. *arXiv preprint arXiv:2412.00585*, 2024.

[19] J. Liang and R. D. C. Monteiro. A doubly accelerated inexact proximal point method for nonconvex composite optimization problems. *Available on arXiv:1811.11378*, 2018.

[20] J. Liang and R. D. C. Monteiro. An average curvature accelerated composite gradient method for nonconvex smooth composite optimization problems. *SIAM Journal on Optimization*, 31(1):217–243, 2021.

[21] J. Liang and R. D. C. Monteiro. A proximal bundle variant with optimal iteration-complexity for a large range of prox stepsizes. *SIAM Journal on Optimization*, 31(4):2955–2986, 2021.

[22] J. Liang and R. D. C. Monteiro. Average curvature FISTA for nonconvex smooth composite optimization problems. *Computational Optimization and Applications*, 86(1):275–302, 2023.

[23] J. Liang and R. D. C. Monteiro. A unified analysis of a class of proximal bundle methods for solving hybrid convex composite optimization problems. *Mathematics of Operations Research*, 49(2):832–855, 2024.

[24] J. Liang, Renato D. C. Monteiro, and C.-K. Sim. A FISTA-type accelerated gradient algorithm for solving smooth nonconvex composite optimization problems. *Computational Optimization and Applications*, 79(3):649–679, 2021.

[25] H. Lin, J. Mairal, and Z. Harchaoui. A universal catalyst for first-order optimization. *Advances in Neural Information Processing Systems*, 28, 2015.

[26] R. Mifflin. A modification and an extension of Lemaréchal's algorithm for nonsmooth minimization. In *Nondifferential and variational techniques in optimization*, pages 77–90. Springer, 1982.

[27] R. D. C. Monteiro, C. Ortiz, and B. F. Svaiter. An adaptive accelerated first-order method for convex optimization. *Computational Optimization and Applications*, 64:31–73, 2016.

[28] R. D. C. Monteiro and B. F. Svaiter. An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods. *SIAM Journal on Optimization*, 23(2):1092–1125, 2013.

[29] I. Necoara, Y. Nesterov, and F. Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, 175(1-2):69–107, 2019.

[30] Y. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $\mathrm{O}(1/k^2)$. *Doklady AN SSSR*, 269:543–547, 1983.

[31] Y. Nesterov. *Introductory lectures on convex optimization : a basic course.* Kluwer Academic Publ., Boston, 2004.

[32] Y. Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152:381–404, 2015.

[33] Y. Nesterov. *Lectures on Convex Optimization*, volume 137 of *Springer Optimization and Its Applications*. Springer, 2018.

[34] B. O'Donoghue and E. J. Candès. Adaptive restart for accelerated gradient schemes. *Foundations of Computational Mathematics*, 15(3):715–732, 2015.

[35] W. Su, S. Boyd, and E. J. Candès. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016.

[36] A. Sujanani and R. D. C. Monteiro. Efficient parameter-free restarted accelerated gradient methods for convex and strongly convex optimization. *arXiv preprint arXiv:2410.04248*, 2024.

[37] A. Wibisono, A. C. Wilson, and M. I. Jordan. A variational perspective on accelerated methods in optimization. *proceedings of the National Academy of Sciences*, 113(47):E7351–E7358, 2016.

[38] P. Wolfe. A method of conjugate subgradients for minimizing nondifferentiable functions. In *Nondifferentiable optimization*, pages 145–173. Springer, 1975.

A Deferred proofs

Proof of Lemma 2.1: (a) This statement immediately follows from the recessions of τ_j and A_j in (3) and the facts that $\tau_0 = 1/L$ and $A_0 = 0$.

(b) It is easy to verify that a_j in (3) is the root of equation $a_j^2 - \tau_j a_j - \tau_j A_j = 0$, which is equivalent to statement b) in view of the third identity in (3). ∎

Proof of Lemma 2.2: (a) It follows from (4) and definition of $\tilde{\gamma}_j$ in (7) that

$$\tilde{y}_{j+1} = \underset{u\in\mathbb{R}^n}{\operatorname{argmin}} \left\{ \tilde{\gamma}_j(u) + \frac{L}{2}\|u - \tilde{x}_j\|^2 \right\}. \tag{32}$$

Since $\tilde{\gamma}_j$ is μ-strongly convex, (32) implies that

$$\tilde{\gamma}_j(u) + \frac{L}{2}\|u - \tilde{x}_j\|^2 \geq \tilde{\gamma}_j(\tilde{y}_{j+1}) + \frac{L}{2}\|\tilde{y}_{j+1} - \tilde{x}_j\|^2 + \frac{\mu + L}{2}\|u - \tilde{y}_{j+1}\|^2.$$

Hence, using the definition of γ_j in (8) and rearranging the terms, we have $\gamma_j \leq \tilde{\gamma}_j$. Using the definitions of ψ and $\tilde{\gamma}_j$ in (2) and (7), respectively, and the assumption that g is μ-strongly convex, we obtain $\tilde{\gamma}_j \leq \psi$, and thus prove the inequalities in (a). By the definition of γ_j in (8), it is easy to verify that $\tilde{\gamma}_j(\tilde{y}_{j+1}) = \gamma_j(\tilde{y}_{j+1})$ and

$$\tilde{y}_{j+1} = \underset{u \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ \gamma_j(u) + \frac{L}{2}\|u - \tilde{x}_j\|^2 \right\}. \tag{33}$$

Finally, (10) is an immediate consequence of (32), (33) and $\tilde{\gamma}_j(\tilde{y}_{j+1}) = \gamma_j(\tilde{y}_{j+1})$.

(b) It clearly follows from (8) that γ_j is μ-strongly convex quadratic. Moreover, it follows from (9) and the fact that $\Gamma_0 \equiv 0$ that Γ_j is also μ-strongly convex quadratic.

(c) It follows from (6), the definition of $\tilde{x}_j$ in (3), and Lemma 2.1(b) that

$$(A_j\mu + 1)x_{j+1} = (L + \mu)a_j\tilde{y}_{j+1} - L(a_j\tilde{x}_j - \tau_j x_j),$$

which together with Lemma 2.1(a) and the third identity in (3) implies that

$$a_j[L(\tilde{x}_j - \tilde{y}_{j+1}) + \mu(x_{j+1} - \tilde{y}_{j+1})] + (A_j\mu + 1)(x_{j+1} - x_j) = 0.$$

In view of the definition of γ_j in (8), the above identity is equivalent to

$$a_j\nabla\gamma_j(x_{j+1}) + (A_j\mu + 1)(x_{j+1} - x_j) = 0. \tag{34}$$

Hence,

$$x_{j+1} = \underset{u \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ a_j\gamma_j(u) + (A_j\mu + 1)\|u - x_j\|^2/2 \right\}.$$

It follows from the definition of γ_j in (8) that

$$\nabla\gamma_j(x_{j+1}) - \nabla\gamma_j(x_j) = \mu(x_{j+1} - x_j),$$

which together with (34) implies that

$$a_j\nabla\gamma_j(x_j) + (A_{j+1}\mu + 1)(x_{j+1} - x_j) = 0. \tag{35}$$

It follows from statement b) and (9) that

$$\begin{aligned} A_{j+1}\nabla\Gamma_{j+1}(x_{j+1}) &= A_{j+1}\nabla\Gamma_{j+1}(x_j) + A_{j+1}\mu(x_{j+1} - x_j) \\ &\overset{(9)}{=} A_j\nabla\Gamma_j(x_j) + a_j\nabla\gamma_j(x_j) + A_{j+1}\mu(x_{j+1} - x_j) \\ &\overset{(35)}{=} A_j\nabla\Gamma_j(x_j) - x_{j+1} + x_j \end{aligned}$$

where the last identity is due to (35). Therefore, for every ≥ 0,

$$A_{j+1}\nabla\Gamma_{j+1}(x_{j+1}) + x_{j+1} - x_0 = A_j\nabla\Gamma_j(x_j) + x_j - x_0 = A_0\nabla\Gamma_0(x_0) + x_0 - x_0 = 0.$$

Therefore, statement c) immediately follows. ∎

Proof of Lemma 2.3: Proof by induction. Since $A_0 = 0$, the case $j = 0$ is trivial. Assume that the claim is true for some $j \geq 0$. Using (9), Lemma 2.2(b), and the induction hypothesis that

$$\begin{aligned} A_{j+1}\Gamma_{j+1}(u) + \frac{1}{2}\|u - x_0\|^2 &\overset{(9)}{\geq} A_j\Gamma_j(u) + a_j\gamma_j(u) + \frac{1}{2}\|u - x_0\|^2 \\ &\geq \min_{u\in\mathbb{R}^n}\left\{A_j\Gamma_j(u) + \frac{1}{2}\|u - x_0\|^2\right\} + \frac{A_j\mu + 1}{2}\|u - x_j\|^2 + a_j\gamma_j(u) \\ &\overset{(11)}{\geq} A_j\psi(y_j) + \frac{A_j\mu + 1}{2}\|u - x_j\|^2 + a_j\gamma_j(u). \end{aligned}$$

It follows from the fact that $\gamma_j \leq \tilde{\gamma}_j \leq \psi$ (see Lemma 2.2(a)) and the definition of $\tilde{x}_j$ in (3) that

$$\begin{aligned} A_{j+1}\Gamma_{j+1}(u) + \frac{1}{2}\|u - x_0\|^2 &\geq A_j\gamma_j(y_j) + a_j\gamma_j(u) + \frac{A_j\mu + 1}{2}\|u - x_j\|^2 \\ &\geq A_{j+1}\gamma_j(\tilde{u}) + \frac{A_j\mu + 1}{2}\frac{A_{j+1}^2}{a_j^2}\|\tilde{u} - \tilde{x}_j\|^2, \end{aligned}$$

where $\tilde{u} = (A_j y_j + a_j u)/A_{j+1}$ and the second inequality is due to the convexity of γ_j. Minimizing both sides of the above inequality over $\mathbb{R}^n$ and using Lemma 2.1(a)-(b), we obtain

$$\begin{aligned} &\min_{u\in\mathbb{R}^n}\left\{A_{j+1}\Gamma_{j+1}(u) + \frac{1}{2}\|u - x_0\|^2\right\} \geq \min_{\tilde{u}\in\mathbb{R}^n}\left\{A_{j+1}\gamma_j(\tilde{u}) + \frac{A_j\mu + 1}{2}\frac{A_{j+1}^2}{a_j^2}\|\tilde{u} - \tilde{x}_j\|^2\right\} \\ =&A_{j+1}\min_{\tilde{u}\in\mathbb{R}^n}\left\{\gamma_j(\tilde{u}) + \frac{L}{2}\|\tilde{u} - \tilde{x}_j\|^2\right\} \overset{(10)}{=} A_{j+1}\min_{\tilde{u}\in\mathbb{R}^n}\left\{\tilde{\gamma}_j(\tilde{u}) + \frac{L}{2}\|\tilde{u} - \tilde{x}_j\|^2\right\} \\ =&A_{j+1}\left(\tilde{\gamma}_j(\tilde{y}_{j+1}) + \frac{L}{2}\|\tilde{y}_{j+1} - \tilde{x}_j\|^2\right) \geq A_{j+1}\psi(\tilde{y}_{j+1}), \end{aligned}$$

where the last inequality follows from the definition of $\tilde{\gamma}_j$ in (7) and the assumption that g is $(L+\mu)$-smooth. Finally, it follows from the definition of y_{j+1} in (5) that (11) holds for $j+1$. Therefore, we complete the proof by induction. ∎