An accelerated proximal bundle method for convex optimization

Feng-Yi Liao FLIAO@UCSD.EDU
Department of Electrical and Computer Engineering, University of California San Diego

Thomas Madden THMADDEN@UCSD.EDU
Department of Mathematics, University of California San Diego

Yang Zheng ZHENGY@UCSD.EDU
Department of Electrical and Computer Engineering, University of California San Diego

Editors: G. Sukhatme, L. Lindemann, S. Tu, A. Wierman, N. Atanasov

Abstract

The proximal bundle method (PBM) is a powerful and widely used approach for minimizing nonsmooth convex functions. However, for smooth objectives, its best-known convergence rate remains suboptimal, and whether PBM can be accelerated remains open. In this work, we present the first *accelerated proximal bundle method* that achieves the optimal $\mathcal{O}(1/\sqrt{\epsilon})$ iteration complexity for obtaining an ϵ-accurate solution in smooth convex optimization. The proposed method is *conceptually simple*, and differs from Nesterov's accelerated gradient descent by only a single line and retains all key structural properties of the classical PBM. In particular, it relies on the same minimal assumptions on model approximations and preserves the standard bundle testing criterion. Numerical experiments confirm the accelerated $\mathcal{O}(1/\sqrt{\epsilon})$ convergence rate predicted by our theory.

Keywords: Acceleration, proximal bundle method, smooth convex optimization

1. Introduction

Convex optimization plays a fundamental role in various disciplines (Nesterov et al., 2018; Boyd and Vandenberghe, 2004). Among its many subfields, *smooth and convex* optimization is arguably the most fundamental, which considers the unconstrained problem

$$f^\star = \min_{x\in\mathbb{R}^n} \; f(x) \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is convex and M-smooth, i.e., $\|\nabla f(x) - \nabla f(y)\| \leq M\|x - y\|, \forall\, x, y \in \mathbb{R}^n$. In this setting, the simplest yet fundamental algorithm is the gradient descent, which moves along the negative gradient direction at every iteration, i.e., $x_{k+1} = x_k - \alpha_k \nabla f(x_k)$, where α_k denotes the stepsize. A standard textbook result shows that with a constant step size $\alpha_k = \frac{1}{M}$, gradient descent can find an ϵ-solution (i.e., an iterate x_k such that $f(x_k) - f^\star \leq \epsilon$) within $\mathcal{O}(1/\epsilon)$ iterations (Nesterov et al. (2018, Section 1.2.3)). Nevertheless, the rate is only suboptimal. Nesterov introduced an accelerated gradient descent (AGD), which improves the iteration complexity to the optimal complexity $\mathcal{O}(1/\sqrt{\epsilon})$ (Nesterov et al. (2018, Section 2.2)).

Another conceptually simple algorithm is the *proximal point method* (PPM), which follows an implicit (sub)gradient update. Equivalently, PPM can be viewed as applying gradient descent to the Moreau envelope (Parikh and Boyd (2014, Section 4.1.1)). It achieves the same $\mathcal{O}(1/\epsilon)$ iteration complexity for both smooth and nonsmooth convex functions (Güler (1991, Theorem 2.1)). Inspired by Nesterov's AGD, various accelerated PPMs have also been proposed to improve the complexity to $\mathcal{O}(1/\sqrt{\epsilon})$; see e.g., (He and Yuan, 2012; Monteiro and Svaiter, 2013; Salzo and Villa, 2012).

Although PPM can handle both smooth and nonsmooth objectives, its update is often computationally intractable, as it requires solving a proximal subproblem exactly. To address this limitation, the *proximal bundle method* (PBM), originally introduced in (Lemarechal, 1978; Mifflin, 1977) for nonsmooth functions, relaxes the exact proximal step to an inexact implicit (sub)gradient update. This relaxation greatly reduces computational cost while preserving the convergence properties of PPM. The asymptotic convergence of PBM iterates to an optimal solution was first established in Kiwiel (1983). Subsequently, Kiwiel (2000) provided the first nonasymptotic complexity bound of $\mathcal{O}(1/\epsilon^3)$ for obtaining an ϵ-solution of general convex (potentially nonsmooth) functions. More recently, sharper convergence guarantees have been established under additional growth or smoothness assumptions (Du and Ruszczyński, 2017; Díaz and Grimmer, 2023). Nevertheless, even for smooth convex functions, the best-known iteration complexity of PBM remains suboptimal at $\mathcal{O}(1/\epsilon)$, even with adaptive step-size rules (Díaz and Grimmer, 2023, Table 1).

It remains unclear whether and how the classical PBM can be accelerated to achieve the optimal rate $\mathcal{O}(1/\sqrt{\epsilon})$. As noted in two recent studies (Díaz and Grimmer, 2023; Liang and Monteiro, 2024), this question is still open. One difficulty lies in the convergence analysis itself: even for the standard PBM, establishing tight iteration-complexity bounds has been challenging and was only recently clarified in Díaz and Grimmer (2023). Another challenge is algorithmic: unlike gradient-based methods, it is not obvious how to introduce momentum or extrapolation into the traditional single-loop PBM framework. Very recently, Fersztand and Sun (2025) achieved an improved complexity of $\mathcal{O}(\log(1/\epsilon)/\sqrt{\epsilon})$ by employing more intricate bundle model assumptions.

In this work, we introduce the *first* accelerated PBM that achieves the optimal iteration complexity $\mathcal{O}(1/\sqrt{\epsilon})$ for finding an ϵ-solution for smooth convex functions. Algorithmically, our development of accelerated PBM is derived intuitively from Nesterov's famous AGD (Nesterov, 1983). In particular, our accelerated PBM only differs by one line from Nesterov's AGD. Under a proper choice of parameters and under-estimators, our proposed method reduces to Nesterov's AGD exactly. Moreover, our proposed accelerated PBM preserves *all* key features of the classical PBM, as it *does not* change the testing criterion of the classical PBM or impose any additional assumptions on the under-estimators. This is made possible through a crucial observation that the proximal bundle update can be interpreted as an inexact implicit (sub)gradient step, which naturally accommodates Nesterov-type momentum and enables acceleration. Importantly, our proposed accelerated PBM can be viewed as a special realization of the abstract accelerated inexact proximal point framework in Monteiro and Svaiter (2013). We present a detailed comparison with Fersztand and Sun (2025) in Remark 1.

The rest of this paper is organized as follows. Section 2 reviews the proximal point method and the PBM. Section 3 introduces our proposed accelerated PBM. Section 4 establishes the convergence guarantees. Section 5 shows numerical experiments, and Section 6 concludes the paper.

2. Preliminaries and problem statement

2.1. Gradient descent and proximal point methods

Consider the optimization problem (1). If the objective f is a differentiable convex function with M-Lipschitz continuous gradient, the standard *gradient descent* (GD) method performs the update

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k), \tag{2}$$

where α_k denotes the step size. With the constant stepsize $\alpha_k = 1/M$, the GD iterates converge with a rate $f(x_k) - f^\star = \mathcal{O}(1/k)$; see, e.g., Corollary 2.1.2 in Nesterov et al. (2018)

For a general convex (possibly nonsmooth) objective f, the *proximal point method* (PPM) replaces the gradient step (2) with a proximal update

$$y_{k+1} = \arg\min_x \left\{ f(x) + \frac{\rho}{2}\|x - y_k\|^2 \right\}, \qquad \rho > 0. \tag{3}$$

The PPM update (3) is well-defined for any $\rho > 0$ since the subproblem is strongly convex and thus admits a unique solution. It is known that the PPM has the same convergence rate $f(y_k) - f^\star = \mathcal{O}(1/k)$; see e.g., Güler (1991, Theorem 2.1). This rate holds for both smooth and nonsmooth convex objectives. In general, the PPM update (3) may not be realizable efficiently. Nevertheless, the PPM serves as a conceptual foundation for modern proximal and bundle-type algorithms (Drusvyatskiy, 2017; Liang and Monteiro, 2021; Díaz and Grimmer, 2023; Liao and Zheng, 2025a,b).

2.2. Proximal bundle method as an inexact PPM

It is known that the PPM (3) can be interpreted as an algorithm which performs an *implicit* (sub)gradient update (Correa and Lemaréchal, 1993). From the optimality condition, (3) can be equivalently written as

$$y_{k+1} = y_k - \frac{1}{\rho} g_{k+1}, \tag{4}$$

Algorithm 1 Proximal Bundle method (PBM)

Require: $y_0 \in \mathbb{R}^n, \rho > 0, \beta \in (0,1)$, $k_{\max} \in \mathbb{N}$

1: **for** $k = 0, 1, 2, \ldots, k_{\max}$ **do**
2: $\quad y_{k+1} = \texttt{ProxDescent}(y_k, \beta, \rho)$;
3: **end for**

where $g_{k+1} \in \partial f(y_{k+1})$ is a subgradient at the next iterate y_{k+1}. Recall that the usual convex subdifferential is defined as $\partial f(x) := \{s \in \mathbb{R}^n \mid f(y) \geq f(x) + \langle s, y - x\rangle, \forall y \in \mathbb{R}^n\}$. If the function f is differentiable, the subdifferential reduces to the usual gradient, i.e., $\partial f(x) = \{\nabla f(x)\}$.

Since the PPM update (3) or (4) is non-trivial, inexact variants of the PPM have been widely considered (Rockafellar, 1976). We here introduce the class of proximal bundle methods (PBMs) (Kiwiel, 2000) as a special inexact PPM. We list the PBM as a double-loop algorithm in Algorithm 1 involving a subroutine $\texttt{ProxDescent}(y_k, \beta, \rho)$, which we shall define next.

We recall the notion of ϵ-inexact subdifferential for a convex function f:

$$\partial_\epsilon f(x) := \{v \in \mathbb{R}^n \mid f(y) \geq f(x) + \langle v, y - x\rangle - \epsilon, \forall\, y \in \mathbb{R}^n\}, \tag{5}$$

where $\epsilon \geq 0$. It is clear that we have $\partial_\epsilon f(x) = \partial f(x)$ when $\epsilon = 0$. Instead of the true subgradient update (4), the PBM considers an inexact update

$$y_{k+1} = y_k - \frac{1}{\rho}\tilde{g}_{k+1}, \tag{6}$$

where $\tilde{g}_{k+1} \in \partial_{\epsilon_{k+1}} f(y_{k+1})$ is an inexact subgradient at the next iterate y_{k+1} with inexactness $\epsilon_{k+1} \geq 0$ coming from a descent criterion that depends on a pre-defined parameter $\beta \in (0, 1]$; see (8).

In general, finding such an inexact subgradient $\tilde{g}_{k+1}$ requires an iterative search. For brevity, we denote such a subroutine as $y_{k+1} = \texttt{ProxDescent}(y_k, \beta, \rho)$, since the iterate y_{k+1} comes from simple proximal updates (see (7)) and satisfies a certain descent performance (see (8)). In particular, $\texttt{ProxDescent}(y_k, \beta, \rho)$ is realized efficiently via an iterative bundle strategy, as detailed below.

Proximal bundle strategy: A unique feature of the PBM is to approximate the original function f such that the proximal update becomes much simpler. Starting with $j = 1$ and $z_1 = y_k$, the PBM generates a sequence of candidate points

$$z_{j+1} = \operatorname*{argmin}_y \left\{ \tilde{f}_j(y) + \frac{\rho}{2}\|y - y_k\|^2 \right\}, \quad (7)$$

where $\tilde{f}_j$ is a convex under-estimator of f. Since f is convex, such an under-estimator $\tilde{f}_j$ is easy to construct, e.g., using its (sub)gradient $s \in \partial f(y_k)$. If this candidate point z_{j+1} satisfies

$$f(y_k) - f(z_{j+1}) \geq \beta(f(y_k) - \tilde{f}_j(z_{j+1})), \quad (8)$$

where $\beta \in (0, 1]$, then we set $y_{k+1} = z_{j+1}$; otherwise, we construct a new convex under-estimator $\tilde{f}_{j+1}$ satisfying Assumption 1, and repeat (7) with $\tilde{f}_{j+1}$ until (8) is satisfied. We list this procedure in Algorithm 2. The criterion (8) indicates that the candidate z_{j+1} achieves at least β fraction of the objective value decrease that the model $f_j(\cdot)$ predicts. If (8) holds, this is known as a *descent* step (and Algorithm 2 terminates); otherwise, it is a *null* step. At this stage, it is not obvious how such a criterion (8) is connected with the inexact update (6). We will detail their relationship in Section 4, which turns out to be one key observation that leads to our accelerated PBM algorithm.

Algorithm 2 `ProxDescent`(y_k, β, ρ)

1: Initialize $z_1 = y_k$
2: **for** $j = 1, 2, \ldots$ **do**
3: Construct $\tilde{f}_j$ satisfying Assumption 1;
4: Compute z_{j+1} using (7);
5: **if** (8) holds **then** Break;
6: **end if**
7: **end for**
8: **Return** z_{j+1};

Assumption 1 *Let $z_1 = y_k$. For $j \geq 1$, the convex function f_j satisfies three conditions:*

1. ***Lower approximation:*** $\tilde{f}_j(y) \leq f(y), \forall y \in \mathbb{R}^n$.
2. ***Subgradient:*** *There exists $g_j \in \partial f(z_j)$ such that $\tilde{f}_j(y) \geq f(z_j) + \langle g_j, y - z_j \rangle, \forall y \in \mathbb{R}^n$.*
3. ***Aggregation:*** *If $j > 1$ (equivalently, (8) does not hold for z_j), we require $\tilde{f}_j(y) \geq \tilde{f}_{j-1}(z_j) + \langle s_j, y - z_j \rangle, \forall y \in \mathbb{R}^n$, where $s_j = \rho(y_k - z_j) \in \partial \tilde{f}_{j-1}(z_j)$.*

In Assumption 1, the first requirement means that $\tilde{f}_j$ is a global under-estimator of f, the second requirement means that $\tilde{f}_j$ should be a better model of f than the subgradient lower-bound of f at z_j, and the last requirement asks that $\tilde{f}_j$ is also better than the linearization of the previous model $\tilde{f}_{j-1}$ at z_j. The second and the third requirements together guarantee that the value $\eta_j := \min_y\{\tilde{f}_j(y) + \frac{\rho}{2}\|y - y_k\|^2\}$ increases monotonically as j increases, and the first requirement guarantees that this value η_j is upper-bounded as $\eta_j \leq \eta^* := \min_y\{f(y) + \frac{\rho}{2}\|y - y_k\|^2\}$ for all $j \geq 1$.

Any approximation model $\tilde{f}_j$ satisfying Assumption 1 can ensure that Algorithm 2 terminates with a finite number of iterations for general convex functions; see e.g., Díaz and Grimmer (2023). Specifically, we can choose $\tilde{f}_j$ as the maximum of two linear lower bounds:

$$\tilde{f}_j(y) = \max\{f(z_j) + \langle g_j, y - z_j \rangle, \tilde{f}_{j-1}(z_j) + \langle s_j, y - z_j \rangle\}. \quad (9)$$

In this case, the proximal update (7) admits an analytically closed-form solution, which is almost as efficient as the simple (sub)gradient update (2). For specific applications, we may choose other approximating models (Helmberg and Rendl, 2000; Liao et al., 2025). If f is a smooth convex function, Algorithm 2 will terminate within a constant number of iterations.

Lemma 1 (Díaz and Grimmer (2023, Lemma 5.2 and Theorem 2.2)) *Suppose f is convex and M-smooth. Let $\beta \in (0, 1)$ and $\rho > 0$. Then, for any $y_k \in \mathbb{R}^n$, `ProxDescent`(y_k, β, ρ) in Algorithm 2 terminates in at most $\frac{16(M+\rho)^3}{(1-\beta)^2\rho^3}$ iterations. Accordingly, the total iteration complexity of Algorithm 1 (including all inner iterations) to find an ϵ-optimal solution is $\mathcal{O}(1/\epsilon)$.*

The PBM described in Algorithm 1 is presented as a double-loop algorithm. Historically, however, the PBM is often formulated in a single-loop form; see, e.g., Kiwiel (2000, Algorithm 2.1) and Díaz and Grimmer (2023, Algorithm 1). These two formulations are in fact equivalent; see Liao and Zheng (2025b, Section 4.3) or Appendix A.5. The double-loop form in Algorithm 1 will prove to be particularly convenient for developing the accelerated PBM in this work.

2.3. Problem statement

The $\mathcal{O}(1/\epsilon)$ complexity of PBM established in Lemma 1 matches that of the exact PPM (Güler (1991, Theorem 2.1)) and vanilla gradient descent with a constant stepsize (Nesterov et al. (2018, Corollary 2.1.2)). However, this $\mathcal{O}(1/\epsilon)$ rate is not optimal for smooth convex functions. For this class, it is well-known that gradient descent with momentum (commonly referred to as *accelerated gradient descent* (Nesterov, 1983)) achieves the optimal rate $\mathcal{O}(1/\sqrt{\epsilon})$. Similarly, the PPM can be accelerated to achieve the same optimal rate, with the first such result due to (Güler, 1992). Building on this foundation, several accelerated *inexact* PPMs have been developed in (He and Yuan, 2012; Monteiro and Svaiter, 2013; Salzo and Villa, 2012; Lin et al., 2018).

In this work, we aim to develop an accelerated PBM that attains the optimal convergence rate $\mathcal{O}(1/\sqrt{\epsilon})$ for smooth convex problems. The proposed method will rely only on the standard bundle assumptions in Assumption 1 and the conventional descent criterion (8), without requiring any additional bundle modifications. Our key observation is that the inexact (sub)gradient update (6) and the double-loop structure in Algorithm 1 can naturally incorporate Nesterov's momentum mechanism. Importantly, the resulting accelerated PBM can be interpreted as a particular instance of the accelerated inexact proximal point framework introduced by Monteiro and Svaiter (2013).

3. An accelerated proximal bundle method

This section introduces our accelerated PBM that integrates Nesterov's extrapolation mechanism. We present its convergence guarantees and identify an interesting connection with Nesterov's AGD.

3.1. Integrating Nesterov's extrapolation with PBM

Among various acceleration techniques for first-order methods, the most fundamental and influential is *Nesterov's accelerated gradient descent* (AGD) (Nesterov, 1983; Nesterov et al., 2018). Note that Nesterov's AGD has multiple equivalent formulations (see e.g., d'Aspremont et al. (2021, Section 4.4)). We list a particular Nesterov's AGD in Algorithm 3.

Unlike the standard gradient descent (2), Nesterov's AGD maintains three coupled sequences of iterates, denoted by $\{x_k\}$, $\{y_k\}$, and $\{z_k\}$. At each iteration, it performs a gradient step at the extrapolated point y_k (see line 6 in Algorithm 3), followed by a momentum-based extrapolation to generate the next iterate y_{k+1} in lines 5 and 7. This additional extrapolation step introduces a form of momentum that substantially accelerates the convergence performance. Indeed, Algorithm 3 improves the iteration complexity from $\mathcal{O}(1/\epsilon)$ to $\mathcal{O}(1/\sqrt{\epsilon})$ in achieving an ϵ-accurate solution for smooth convex functions (Nesterov et al., 2018, Theorem 2.17).

Our key observation is that the PBM can be interpreted as an inexact (sub)gradient update (6). It is then possible to integrate this Nesterov's extrapolation mechanism within the standard PBM framework. In particular, leveraging the double-loop structure in Algorithm 1, we propose the accelerated PBM listed in Algorithm 4. Notably, Algorithms 3 and 4 share the same extrapolation

and momentum updates; they differ only in the sixth line, where the exact gradient step in AGD $x_{k+1} = y_k - \nabla f(y_k)/M$ is replaced by the PBM oracle

$$x_{k+1} = \texttt{ProxDescent}(y_k, \beta, \rho).$$

This oracle corresponds to the inexact proximal step described in (6) and ensures the descent and model-approximation conditions defined in (8) and Assumption 1.

The proposed accelerated PBM can be viewed as an accelerated inexact PPM realized through standard bundle updates, which preserves the classical PBM structure. We show next that Algorithm 4 not only achieves the optimal $\mathcal{O}(1/\sqrt{\epsilon})$ convergence rate for smooth convex problems, but also naturally includes Algorithm 3 as a special case when choosing the proximal parameter ρ and descent parameter β appropriately.

Algorithm 3 Nesterov's AGD

1: Set $z_0 = x_0$ and $A_0 = 0$.
2: **for** $k = 0, 1, \ldots, T$ **do**
3: $\quad a_k = (1 + \sqrt{1 + 4A_k})/2$
4: $\quad A_{k+1} = A_k + a_k$
5: $\quad y_k = \frac{A_k}{A_{k+1}} x_k + \frac{a_k}{A_{k+1}} z_k$
6: $\quad x_{k+1} = y_k - \nabla f(y_k)/M$
7: $\quad z_{k+1} = z_k - a_k(y_k - x_{k+1})$
8: **end for**

Algorithm 4 Accelerated PBM

1: Set $z_0 = x_0, A_0 = 0$
2: **for** $k = 0, 1, \ldots, T$ **do**
3: $\quad a_k = (1 + \sqrt{1 + 4A_k})/2$
4: $\quad A_{k+1} = A_k + a_k$
5: $\quad y_k = \frac{A_k}{A_{k+1}} x_k + \frac{a_k}{A_{k+1}} z_k$
6: $\quad x_{k+1} = \texttt{ProxDescent}(y_k, \beta, \rho)$
7: $\quad z_{k+1} = z_k - a_k(y_k - x_{k+1})$
8: **end for**

3.2. Accelerated rate for smooth convex problems

We now present our main convergence guarantee. Similar to Nesterov's AGD, our accelerated PBM also achieves the $\mathcal{O}(1/\sqrt{\epsilon})$ optimal iteration complexity for smooth convex problems.

Theorem 1 *Consider the problem (1). Let $S = \operatorname{argmin}_x f(x)$, and assume $S \neq \emptyset$. If f is convex and M-smooth, then Algorithm 4 with parameters $\rho \geq \frac{M}{c}$ for some $c > 0$, and $\beta \in [\frac{c+2\sqrt{c}+2}{c+2\sqrt{c}+3}, 1)$ generates a sequence of iterates $\{x_k\}$ satisfying*

$$f(x_k) - f^\star \leq \frac{2\rho \mathrm{dist}^2(x_0, S)}{k^2}, \ \forall k \geq 1. \tag{10}$$

Moreover, the total iteration count (including the inner loop) required to obtain an ϵ-optimal solution, i.e., a point x_k such that $f(x_k) - f^\star \leq \epsilon$, is bounded by

$$\frac{16(M+\rho)^3}{(1-\beta)^2\rho^3} \sqrt{\frac{2\rho \cdot \mathrm{dist}^2(x_0, S)}{\epsilon}}. \tag{11}$$

The proof builds on the interpretation of the PBM update as an inexact (sub)gradient step (6), realized by the subroutine $\texttt{ProxDescent}(y_k, \beta, \rho)$. In particular, we show that the bundle-model assumptions in Assumption 1 and the descent test (8) jointly control the inexactness parameter ϵ_{k+1} in (6). This connection allows us to interpret our accelerated PBM in Algorithm 4 as a special realization of the general accelerated inexact proximal point framework in Monteiro and Svaiter

(2013). Consequently, the accelerated rate in (10) follows directly from Monteiro and Svaiter (2013, Theorem 3.8). The total iteration bound in (11) then follows from Lemma 1, which ensures that each call to `ProxDescent`(y_k, β, ρ) terminates after a constant number of inner iterations. The proof details are given in Section 4.

Although the convergence proof builds upon the general inexact acceleration framework in Monteiro and Svaiter (2013), we believe the underlying observation is nontrivial. The possibility of accelerating PBM had remained open in the literature (Díaz and Grimmer, 2023; Liang and Monteiro, 2024). To our best knowledge, Algorithm 4 provides the first accelerated PBM that achieves the optimal convergence rate for smooth convex functions. In the past, PBMs have been primarily studied for nonsmooth optimization problems, and this reflects their origins in cutting-plane and bundle methods. Only recently, an $\mathcal{O}(1/\epsilon)$ iteration complexity of PBM for smooth functions was established in Díaz and Grimmer (2023). The result in Theorem 1 advances this line of work by improving the rate from $\mathcal{O}(1/\epsilon)$ to $\mathcal{O}(1/\sqrt{\epsilon})$ through the incorporation of Nesterov-type momentum and extrapolation, paralleling the developments of Nesterov's AGD and the accelerated PPM.

We note that Nesterov's AGD in Algorithm 3 always requires the small stepsize choice $1/M$. In contrast, as guaranteed in Theorem 1, the accelerated PBM in Algorithm 4 allows a larger stepsize $1/\rho$ since $\rho > 0$ can be chosen arbitrarily, as long as we choose an appropriate β to control the solution quality of a descent step.

Remark 1 *Here, we compare our accelerated PBM (Algorithm 4) with the recent development Fersztand and Sun (2025, Algorithm 3). First, our Algorithm 4 does not change the descent criterion for the PBM, whereas Fersztand and Sun (2025) proposes to use a more stringent testing criterion. Second, our Assumption 1 for the under-estimators $\tilde{f}_j$ are standard in the classical PBM framework. In contrast, Fersztand and Sun (2025) requires a smooth under-estimator, which is not satisfied by the usual cutting-plane approximation (e.g. $\tilde{f}_j(y) = \max_{i=1,\ldots,j}\{f(y_i)+\langle\nabla f(y_i), y - y_i\rangle\}$). Third, our result is more general as the under-estimator assumption in Fersztand and Sun (2025) naturally satisfies our Assumption 1. As a result, our algorithm enables the use of finite-memory models, where the difficulty of performing a null step is limited by a (sub)gradient evaluation. Finally, our $\mathcal{O}(1/\sqrt{\epsilon})$ iteration complexity in Theorem 1 is better than $\mathcal{O}(\log(1/\epsilon)/\sqrt{\epsilon})$ in Fersztand and Sun (2025, Theorem 26) by a logarithmic term.*

3.3. Connections with Nesterov's AGD

Finally, we point out an interesting connection with Nesterov's AGD. For smooth convex functions, the oracle `ProxDescent`(y_k, β, ρ) can be tuned to produce an iterate in a single inner iteration when the parameters β and ρ are chosen appropriately. Similar observations have been made in (Ding and Grimmer, 2023; Liao et al., 2025) in the context of applying PBM to solve semidefinite programs.

Proposition 1 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a M-smooth function. Given a center point $y_k \in \mathbb{R}^n$, suppose that we choose $\beta \in (0, \frac{1}{2}]$ and $\rho \geq M$ in Algorithm 2. Then, Algorithm 2 terminates in one iteration.*

Due to the page limit, the proof is provided in Appendix A.1. Proposition 1 implies that if the first under-estimator is constructed as $\tilde{f}_1(\cdot) = f(y_k) + \langle\nabla f(y_k), \cdot - y_k\rangle$ and parameters are chosen as $\beta \in (0, 1/2]$ and $\rho \geq M$, then Algorithm 2 will directly output the usual gradient update

$$z_2 = y_k - \tfrac{1}{\rho}\nabla f(y_k)$$

since $z_2 = \operatorname{argmin}_y\{\tilde{f}_1(y) + \frac{\rho}{2}\|y - y_k\|^2\}$. In this case, our Algorithm 4 reduces exactly to Nesterov's AGD in Algorithm 3. When choosing $\beta \in (1/2, 1)$, the oracle $\texttt{ProxDescent}(y_k, \beta, \rho)$ may take more (but a constant number of) gradient-like iterations before the extrapolation step.

In this sense, our proposed Algorithm 4 gracefully generalizes Nesterov's AGD by allowing multiple gradient-like updates within each proximal step before applying the extrapolation mechanism. Such a tunable integration of multiple gradient-like iterations within a Nesterov-type scheme seems not to have been explicitly established before.

4. Technical proofs

In this section, we prove the accelerated convergence of our proposed Algorithm 4, stated in Theorem 1. In particular, we show that Algorithm 4 can be viewed as a special realization of the accelerated hybrid proximal extragradient (A-HPE) in Monteiro and Svaiter (2013).

4.1. A simplified A-HPE algorithm

Let us first review a simplified and reformulated version of A-HPE, listed in Algorithm 5. We note that the update of Algorithm 5 is slightly different from the original version in Monteiro and Svaiter (2013). We believe this reformulated version may provide a clearer picture of the connection with Nesterov's AGD.

Algorithm 5 Simplified A-HPE

1: Set $z_0 = x_0, A_0 = 0, \rho > 0$.
2: **for** $k = 0, 1, \ldots, T$ **do**
3: $\quad a_k = \frac{1+\sqrt{1+4A_k}}{2}$
4: $\quad A_{k+1} = A_k + a_k$
5: $\quad y_k = \frac{A_k}{A_{k+1}} x_k + \frac{a_k}{A_{k+1}} z_k$
6: $\quad x_{k+1} = y_k - g_{k+1}/\rho$ satisfying (12)
7: $\quad z_{k+1} = z_k - a_k(y_k - x_{k+1})$
8: **end for**

The only difference between Algorithms 3 and 5 is the update in the sixth line. In Algorithm 5, instead of restricting to a simple gradient update, Algorithm 5 uses an update direction $g_{k+1} \in \mathbb{R}^n$ with an inexactness $\epsilon_{k+1} \geq 0$ such that

$$x_{k+1} = y_k - \frac{1}{\rho} g_{k+1}, \quad g_{k+1} \in \partial_{\epsilon_{k+1}} f(x_{k+1}), \quad 2\epsilon_{k+1} \leq \rho \|x_{k+1} - y_k\|^2. \tag{12}$$

This scheme can be viewed as an inexact PPM update, as the inexactness ϵ_{k+1} is expressed in terms of the subdifferential at the next iterate. Below, we state the convergence of Algorithm 5.

Theorem 2 *Monteiro and Svaiter (2013, Theorem 3.8) Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function and set $S = \operatorname{argmin}_x f(x)$. Assume $S \neq \emptyset$. The sequence $\{x_k\}$ in Algorithm 5 satisfies*

$$f(x_k) - \min_x f(x) \leq \frac{2\rho \cdot \operatorname{dist}(x_0, S)^2}{k^2}, \ \forall k \geq 1.$$

For self-completeness, we provide the proof of Theorem 2 in Appendix B. It is worth pointing out that Theorem 2 holds for both smooth and nonsmooth convex functions. However, for a nonsmooth function, condition (12) may be difficult to satisfy unless a strong oracle is assumed, which will inevitably require more computational resources.

4.2. Proof of Theorem 1

We now establish the convergence of Algorithm 4. We only need to show that the iterate $x_{k+1} = \texttt{ProxDescent}(y_k, \beta, \rho)$ satisfies (12). Recall that $\texttt{ProxDescent}(y_k, \beta, \rho)$ is realized by repeating

$$z_{j+1} = \operatorname*{argmin}_y \left\{ \tilde{f}_j(y) + \frac{\rho}{2}\|y - y_k\|^2 \right\} \tag{13}$$

until a point z_{j+1} satisfies (8). Our first observation is that z_{j+1} can be rewritten in the form of (12).

Proposition 2 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex. For every $j \geq 1$ iteration of* `ProxDescent`*(y_k, β, ρ) (i.e. Algorithm 2) with $\rho > 0$ and $\beta \in (0, 1)$, it holds that $z_{j+1} = y_k - \frac{1}{\rho}\tilde{g}_{j+1},\ \tilde{g}_{j+1} \in \partial f_{\tilde{\epsilon}_{j+1}}(z_{j+1})$ with the inexactness satisfying $\tilde{\epsilon}_{j+1} = f(z_{j+1}) - \tilde{f}_j(z_{j+1})$.*

Proof From the optimality condition, (13) can be written as

$$z_{j+1} = y_k - \tilde{g}_{j+1}/\rho,\ \tilde{g}_{j+1} \in \partial \tilde{f}_j(z_{j+1}).$$

The fact that $\tilde{g}_{j+1} \in \partial \tilde{f}_j(z_{j+1})$ implies that for all y, we have the following:

$$\begin{aligned} f(y) &\geq \tilde{f}_j(z_{j+1}) + \langle \tilde{g}_{j+1}, y - z_{j+1} \rangle \\ &= f(z_{j+1}) + \langle \tilde{g}_{j+1}, y - z_{j+1} \rangle - (f(z_{j+1}) - \tilde{f}_j(z_{j+1})). \end{aligned}$$

This is the same as saying that $\tilde{g}_{j+1} \in \partial_{\tilde{\epsilon}_{j+1}} f(z_{j+1})$ with $\tilde{\epsilon}_{j+1} = f(z_{j+1}) - \tilde{f}_j(z_{j+1}) \geq 0$. ■

We next show that the error $\tilde{\epsilon}_{j+1}$ can be upper bounded by the decrease in function value whenever a descent step happens, i.e., the descent criterion (8) holds. The proof is provided in Appendix A.2.

Proposition 3 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex. The iterate $x_{k+1} =$* `ProxDescent`*(y_k, β, ρ) with $\rho > 0$ and $\beta \in (0, 1)$ satisfies*

$$x_{k+1} = y_k - g_{k+1}/\rho,\ g_{k+1} \in \partial_{\epsilon_{k+1}} f(x_{k+1}),\ \epsilon_{k+1} \leq \frac{1-\beta}{\beta}\left(f(y_k) - f(x_{k+1})\right). \tag{14}$$

Note that we replaced z_{j+1} by x_{k+1} in Proposition 3 since z_{j+1} satisfies the descent criterion (8) by assumption. Proposition 3 bounds the inexactness of the subgradient. It is not immediate, however, that the inexactness in Proposition 3 satisfies the acceleration condition (12). Fortunately, under the smoothness assumption and with a correctly chosen β, we can guarantee the condition (12).

The key insight is to use smoothness to relate the function difference $f(y_k) - f(x_{k+1})$ to the inexact subgradient interpretation in Proposition 2. Then, notice that we can choose β so that the $\frac{1-\beta}{\beta}$ term in (14) contributes a damping factor to the bound. This damping mechanism enables the enlarged step size choice of Algorithm 4 (i.e., we allow $\rho < M$) when compared to Nesterov's AGD that always requires a small step size $1/M$. We formalize this result in the next lemma. The proof is postponed to Appendix A.3.

Lemma 2 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be M-smooth and convex, and choose $\rho \geq \frac{M}{c}$, $\beta \in \left[\frac{c+2\sqrt{c}+2}{c+2\sqrt{c}+3}, 1\right)$ for $c > 0$ arbitrary. The iterate $x_{k+1} =$* `ProxDescent`*(y_k, β, ρ) satisfies*

$$x_{k+1} = y_k - g_{k+1}/\rho,\ g_{k+1} \in \partial_{\epsilon_{k+1}} f(x_{k+1}),\ 2\epsilon_{k+1} \leq \rho \|x_{k+1} - y_k\|^2.$$

Combining Lemma 2 with Theorem 2 confirms that the iterate $\{x_k\}$ in Algorithm 4 satisfies

$$f(x_k) - f^\star \leq \frac{2\rho \cdot \mathrm{dist}(x_0, S)^2}{k^2},\ \forall k \geq 1.$$

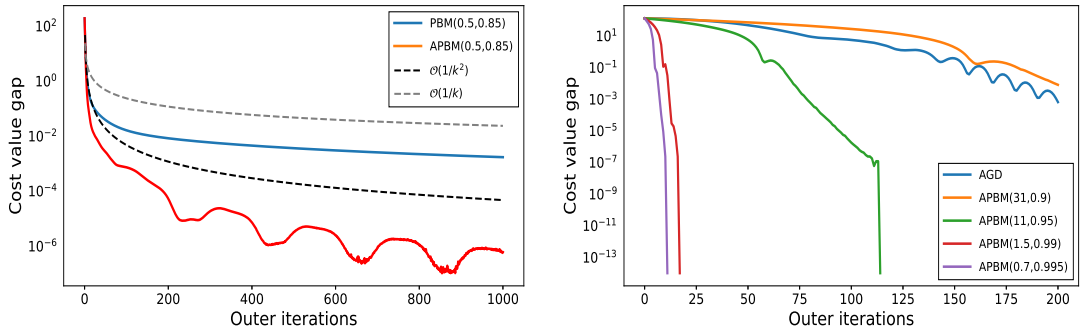
Finally, Lemma 1 ensures that each call of `ProxDescent`(y_k, β, ρ) terminates in at most $\frac{16(M+\rho)^3}{(1-\beta)^2\rho^3}$ iterations. Combining these two results completes the proof of Theorem 1.

5. Numerical experiments

In this section, we perform two numerical experiments to test the numerical performance of our proposed Algorithm 4. We use the essential model (9) to construct the under-estimator in the oracle `ProxDescent`(y_k, β, ρ) in Algorithm 2. In this case, the subproblem in (7) admits an analytical solution; see e.g., Díaz and Grimmer (2023, Claim 1), or Liao and Zheng (2025b, Appendix B.6).

Our first experiment is to verify the $\mathcal{O}(1/k^2)$ convergence guaranteed by Theorem 1. To demonstrate the worst convergence rate, we consider the function $f(x) = \frac{1}{8}x^\top Lx - \frac{1}{4}\langle x, e_1\rangle$, where $L \in \mathbb{R}^{200\times 200}$ is the matrix which is 2 on the diagonal and -1 on the off-diagonals, and e_1 is the standard basis vector of the first coordinate. This function f is 1-smooth and convex but not strongly convex. It is also used in Nesterov et al. (2018, Section 2.1.7) to show the $\Omega(1/k^2)$ complexity bounds for smooth and convex functions under a first-order oracle. We run both the classical PBM and accelerated PBM for 1000 iterations. The numerical result is presented in Figure 1-(a). In Figure 1-(a), we see that the classical PBM only has the usual $\mathcal{O}(1/k)$ convergence behavior. In contrast, the accelerated PBM enjoys a much faster $\mathcal{O}(1/k^2)$ convergence rate, validating our theoretical findings in Theorem 1.

Our second experiment considers the logistic regression objective $f(w) = \frac{1}{m}\sum_{i=1}^m \log(1 + e^{-y_i\langle x_i, w\rangle})$ with synthetic feature-label pairs $\{(x_i, y_i)\}_{i=1}^m$ such that $x_i \in \mathbb{R}^{200}$, $y_i = \pm 1$, and $m = 200$. We generate the problem data such that the objective has a smoothness constant bounded by 1000 and a minimizer $w^\star = 0$. We compare Nesterov's AGD (Algorithm 3) with our proposed accelerated PBM (Algorithm 4) with different choices of parameters. The result of the experiment is presented in Figure 1-(b). We see that the accelerated PBM instances that use a larger step size (smaller ρ) greatly speed up the convergence. For example, the accelerated PBM with $\rho = 0.7$ and $\beta = 0.995$ achieves the accuracy of 10^{-14} in 20 iterations, while the AGD only achieves the accuracy of 10^{-3} in 200 iterations. This numerical experiment also validates our theoretical finding in Theorem 1 that the convergence is guaranteed with a larger step size.



(*a*) Worst-case function (*b*) Logistic regression

Figure 1: Numerical experiments. The worst-case function is taken from Nesterov et al. (2018). The notation PBM(x, y) and APBM(x, y) denotes the PBM and the accelerated PBM with parameters $\rho = x$ and $\beta = y$.

6. Conclusion

We have introduced a new accelerated proximal bundle method (PBM), which naturally integrates Nesterov's acceleration scheme with the classical PBM framework. The proposed method achieves the optimal $\mathcal{O}(1/\sqrt{\epsilon})$ convergence rate for smooth convex functions, and the theory is supported by numerical results. Since PBM is inherently designed for nonsmooth optimization, an interesting future direction is to extend these convergence guarantees beyond the smooth setting.

Acknowledgments

This work is supported by NSF ECCS-2154650, NSF CMMI 2320697, and NSF CAREER 2340713.

References

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Rafael Correa and Claude Lemaréchal. Convergence of some algorithms for convex minimization. *Mathematical Programming*, 62:261–275, 1993.

Mateo Díaz and Benjamin Grimmer. Optimal convergence rates for the proximal bundle method. *SIAM Journal on Optimization*, 33(2):424–454, 2023.

Lijun Ding and Benjamin Grimmer. Revisiting spectral bundle methods: Primal-dual (sub) linear convergence rates. *SIAM Journal on Optimization*, 33(2):1305–1332, 2023.

Dmitriy Drusvyatskiy. The proximal point method revisited. *arXiv preprint arXiv:1712.06038*, 2017.

Yu Du and Andrzej Ruszczyński. Rate of convergence of the bundle method. *Journal of Optimization Theory and Applications*, 173(3):908–922, 2017.

Alexandre d'Aspremont, Damien Scieur, Adrien Taylor, et al. Acceleration methods. *Foundations and Trends® in Optimization*, 5(1-2):1–245, 2021.

David Fersztand and Xu Andy Sun. On the acceleration of proximal bundle methods. *arXiv preprint arXiv:2504.20351*, 2025.

Osman Güler. On the convergence of the proximal point algorithm for convex minimization. *SIAM journal on control and optimization*, 29(2):403–419, 1991.

Osman Güler. New proximal point algorithms for convex minimization. *SIAM Journal on Optimization*, 2(4):649–664, 1992. doi: 10.1137/0802032.

Bingsheng He and Xiaoming Yuan. An accelerated inexact proximal point algorithm for convex minimization. *Journal of Optimization Theory and Applications*, 154(2):536–548, 2012.

Christoph Helmberg and Franz Rendl. A spectral bundle method for semidefinite programming. *SIAM Journal on Optimization*, 10(3):673–696, 2000.

Krzysztof C Kiwiel. Efficiency of proximal bundle methods. *Journal of Optimization Theory and Applications*, 104(3):589–603, 2000.

Krzysztof Czesław Kiwiel. An aggregate subgradient method for nonsmooth convex minimization. *Mathematical Programming*, 27(3):320–341, 1983.

Claude Lemarechal. Nonsmooth optimization and descent methods. 1978.

Jiaming Liang and Renato DC 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.

Jiaming Liang and Renato DC 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.

Feng-Yi Liao and Yang Zheng. A bundle-based augmented lagrangian framework: Algorithm, convergence, and primal-dual principles. *arXiv preprint arXiv:2502.08835*, 2025a.

Feng-Yi Liao and Yang Zheng. A proximal descent method for minimizing weakly convex optimization. *arXiv preprint arXiv:2509.02804*, 2025b.

Feng-Yi Liao, Lijun Ding, and Yang Zheng. An overview and comparison of spectral bundle methods for primal and dual semidefinite programs. *Computational Optimization and Applications*, pages 1–44, 2025.

Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. Catalyst acceleration for first-order convex optimization: from theory to practice. *Journal of Machine Learning Research*, 18(212):1–54, 2018.

Robert Mifflin. An algorithm for constrained optimization with semismooth functions. *Mathematics of Operations Research*, 2(2):191–207, 1977.

Renato DC Monteiro and Benar Fux 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.

Yurii Nesterov. A method for solving the convex programming problem with convergence rate o (1/k2). In *Dokl akad nauk Sssr*, volume 269, page 543, 1983.

Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends® in Optimization*, 1 (3):127–239, 2014. ISSN 2167-3888.

R Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976.

Saverio Salzo and Silvia Villa. Inexact and accelerated proximal point algorithms for minimizing a proper lower semicontinuous and convex function. *Journal of Convex Analysis*, 19(4):1167–1192, 2012.

Appendix A. Additional properties of the proximal bundle method

A.1. Proof of Proposition 1

To prove Proposition 1, we first recall a result from (Liao and Zheng, 2025a).

Theorem 1 (Liao and Zheng (2025a, Theorem 8)) *Let $\alpha > 0$, $y_k \in \mathbb{R}^n$, and $f : \mathbb{R}^n \to \mathbb{R}$. Suppose that $\tilde{f} : \mathbb{R}^n \to \mathbb{R}$ is a convex function that satisfies*

$$\tilde{f}(x) \leq f(x) \leq \tilde{f}(x) + \frac{\alpha}{2}\|x - y_k\|^2, \quad \forall\, x \in \mathbb{R}^n. \tag{15}$$

If $\theta \geq \alpha$ and $z = \operatorname{argmin}_x \left(\tilde{f}(x) + \frac{\theta}{2}\|x - y_k\|^2\right)$, then the following holds:

$$0 \leq f(z) - \tilde{f}(z) \leq f(y_k) - f(z). \tag{16}$$

As Proposition 1 considers a convex and M-smooth function f, it is clear that for any $y_k \in \mathbb{R}^n$, the function $\tilde{f}(\cdot) = f(y_k) + \langle \nabla f(y_k), \cdot - y_k \rangle$ and the constant $\alpha = M$ satisfy (15). With this theorem in hand, we are ready to prove Proposition 1.

Proof of Proposition 1: Given a center point y_k, by Assumption 1, we know that $\tilde{f}_1(y) \leq f(y)$ for all y, and

$$\begin{aligned} & f(y_k) + \langle \nabla f(y_k), y - y_k \rangle \leq \tilde{f}_1(y), \ \forall y \in \mathbb{R}^n \\ \implies & f(y) \leq f(y_k) + \langle \nabla f(y_k), y - y_k \rangle + \frac{\rho}{2}\|y - y_k\|^2 \leq \tilde{f}_1(y) + \frac{\rho}{2}\|y - y_k\|^2, \ \forall y \in \mathbb{R}^n, \rho \geq M, \end{aligned}$$

where the implication is due to the M-smoothness. Thus, we see that $\tilde{f}_1$ satisfies (15). Invoking Theorem 1 on the update $z_2 = \operatorname{argmin}_y\{\tilde{f}_1(y) + \frac{\rho}{2}\|y - y_k\|^2\}$, we obtain

$$0 \leq f(z_2) - \tilde{f}(z_2) \leq f(y_k) - f(z_2). \tag{17}$$

Finally, we calculate that

$$\begin{aligned} \beta(f(y_k) - f_k(z_2)) &= \beta(f(y_k) - f(z_2) + f(z_2) - f_k(z_2)) && (18) \\ &\leq \beta(f(y_k) - f(z_2) + f(y_k) - f(z_2)) && (19) \\ &= 2\beta(f(y_k) - f(z_2) \\ &\leq f(y_k) - f(z_2) && (20) \end{aligned}$$

where we added and subtracted $f(z_2)$ within the parentheses in (18), applied (17) in (19), and used that $\beta \leq \frac{1}{2}$ in (20). Thus, the acceptance criterion (8) is satisfied for z_2, so Algorithm 2 terminates in one iteration. □

A.2. Proof of Proposition 3

For completeness, we prove Proposition 3 now, a result which allows us to bound differences in function values as opposed to bounding the model gap.

Proof of Proposition 3: From Proposition 2, we can write $x_{k+1} = \texttt{ProxDescent}(y_k, \beta, \rho)$ as $x_{k+1} = y_k - \frac{1}{\rho}\tilde{g}_J$ where J is the last iteration index in Algorithm 2 and

$$\tilde{g}_J \in \partial f_{\tilde{\epsilon}_{J+1}}(x_{k+1}),\ \tilde{\epsilon}_{J+1} = f(x_{k+1}) - \tilde{f}_J(x_{k+1}).$$

Moreover, the error $f(x_{k+1}) - \tilde{f}_J(x_{k+1})$ can be further bounded as

$$\begin{aligned} f(x_{k+1}) - \tilde{f}_J(x_{k+1}) &\leq f(y_k) - \beta\big(f(y_k) - \tilde{f}_J(x_{k+1})\big) - \tilde{f}_J(x_{k+1}) && (21)\\ &= (1-\beta)\big(f(y_k) - \tilde{f}_J(x_{k+1})\big) && (22)\\ &\leq \frac{1-\beta}{\beta}\big(f(y_k) - f(x_{k+1})\big), && (23) \end{aligned}$$

where (21) and (23) apply the definition of the test in (8), and (22) combines like terms. Finally, letting $g_{k+1} = \tilde{g}_J$ and $\epsilon_{k+1} = \tilde{\epsilon}_{J+1}$ completest the proof.

□

A.3. Proof of Lemma 2

An important ingredient in the proof of Lemma 2 is the fact that, given a smooth function, the distance between a true gradient and an inexact subgradient can be controlled by the inexactness. This is formalized in the following lemma.

Lemma 3 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be M-smooth and convex. Suppose that $g \in \partial_\epsilon f(x)$ with an arbitrary $\epsilon > 0$. Then, we have*

$$\|\nabla f(x) - g\| \leq \sqrt{2M\epsilon}.$$

Proof By definition, $g \in \partial_\epsilon f(x)$ ensures that the following lower bound holds for all $y \in \mathbb{R}^n$:

$$f(y) \geq f(x) + \langle g, x - y\rangle - \epsilon. \tag{24}$$

Specifying (24) to the choice $y = x - \frac{1}{M}(\nabla f(x) - g)$ gives the lower bound

$$f(x - \frac{1}{M}(\nabla f(x) - g)) \geq f(x) - \frac{1}{M}\langle g, \nabla f(x) - g\rangle - \epsilon. \tag{25}$$

Meanwhile, smoothness gives an upper bound on $f(x - \frac{1}{M}(\nabla f(x) - g))$ via

$$f(x - \frac{1}{M}(\nabla f(x) - g)) \leq f(x) - \frac{1}{M}\langle \nabla f(x), \nabla f(x) - g\rangle + \frac{M}{2}\| - \frac{1}{M}(\nabla f(x) - g)\|^2. \tag{26}$$

Therefore, combining the bounds in (25) and (26) gives

$$\begin{aligned} & -\frac{1}{M}\langle g, \nabla f(x) - g\rangle - \epsilon \leq -\frac{1}{M}\langle \nabla f(x), \nabla f(x) - g\rangle + \frac{1}{2M}\|\nabla f(x) - g\|^2 \\ \Rightarrow\quad & \frac{1}{M}\langle \nabla f(x) - g, \nabla f(x) - g\rangle - \epsilon \leq \frac{1}{2M}\|\nabla f(x) - g\|^2 && (27)\\ \Rightarrow\quad & \|\nabla f(x) - g\|^2 \leq 2M\epsilon && (28) \end{aligned}$$

where (27) and (28) rearrange terms. Taking square roots on both sides of (28) gives the claim. ■

With Lemma 3, we now prove Lemma 2. The idea is to bound the backwards difference $f(y_k) - f(x_{k+1})$ using smoothness, which is possible due to the inexact PPM interpretation.

Proof of Lemma 2: By Proposition 3, it suffices to bound the quantity

$$\hat{\epsilon} := \frac{1-\beta}{\beta}\big(f(y_k) - f(x_{k+1})\big)$$

by $\frac{\rho}{2}\|x_{k+1} - y_k\|^2$. We begin to bound $\hat{\epsilon}$ as follows:

$$\hat{\epsilon} \le \frac{1-\beta}{\beta}\left(\langle \nabla f(x_{k+1}), y_k - x_{k+1}\rangle + \frac{M}{2}\|x_{k+1} - y_k\|^2\right) \tag{29a}$$

$$\le \frac{1-\beta}{\beta}\left(\langle \nabla f(x_{k+1}) - g_{k+1}, y_k - x_{k+1}\rangle + \langle g_{k+1}, y_k - x_{k+1}\rangle + \frac{M}{2}\|x_{k+1} - y_k\|^2\right) \tag{29b}$$

$$= \frac{1-\beta}{\beta}\left(\langle \nabla f(x_{k+1}) - g_{k+1}, y_k - x_{k+1}\rangle + \rho\|x_{k+1} - y_k\|^2 + \frac{M}{2}\|x_{k+1} - y_k\|^2\right), \tag{29c}$$

where (29a) applies M-smoothness of the objective function, (29b) adds and subtracts g_{k+1} within the inner product, and (29c) uses that $g_{k+1} = \rho(y_k - x_{k+1})$ as was shown in Proposition 3. Bounding the inner product above and applying Lemma 3 with $x = x_{k+1}$ yields the bound

$$\hat{\epsilon} \le \frac{1-\beta}{\beta}\left(\sqrt{2M\hat{\epsilon}}\|x_{k+1} - y_k\| + \big(\rho + \frac{M}{2}\big)\|x_{k+1} - y_k\|^2\right). \tag{30}$$

We see that (30) is a quadratic inequality in $\sqrt{\hat{\epsilon}}$. As the coefficient of the quadratic term is positive, an upper bound on $\hat{\epsilon}$ is given by the square of the largest solution to the quadratic. We deduce

$$\hat{\epsilon} \le \left(\frac{1-\beta}{\beta}\right)^2\left(M + \frac{\beta}{1-\beta}\big(\rho + \frac{M}{2}\big) + \sqrt{M^2 + \frac{2M\beta}{1-\beta}\big(\rho + \frac{M}{2}\big)}\right)\|x_{k+1} - y_k\|^2. \tag{31}$$

We now finish the proof of Lemma 2 by establishing the fact that whenever

$$\rho \ge \frac{M}{c}, \quad \beta \in \left[\frac{c + 2\sqrt{c} + 2}{c + 2\sqrt{c} + 3}, 1\right) \tag{32a}$$

we have

$$\left(\frac{1-\beta}{\beta}\right)^2\left(M + \frac{\beta}{1-\beta}\big(\rho + \frac{M}{2}\big) + \sqrt{M^2 + \frac{2M\beta}{1-\beta}\big(\rho + \frac{M}{2}\big)}\right) \le \frac{\rho}{2}. \tag{32b}$$

This inequality (32) is not difficult to show, only requiring single variable calculus. For completeness, we provide the details in Appendix A.4.

A.4. The inequality in (32)

Here, we show that the parameters in (32a) always guarantee the bound in (32b). To simplify notation, we write

$$b := \frac{1-\beta}{\beta}, \quad h_b(\rho) := b^2\left(M + \frac{1}{b}\big(\rho + \frac{M}{2}\big) + \sqrt{M^2 + \frac{2M}{b}\big(\rho + \frac{M}{2}\big)}\right) \tag{33}$$

and our goal becomes showing that $h_b(\rho) \le \frac{\rho}{2}$ when (32a) holds. To do this, we will first show that our restriction on $\beta \in [\frac{c+2\sqrt{c}+2}{c+2\sqrt{c}+3}, 1)$ (corresponding to $b \in (0, \frac{1}{c+2\sqrt{c}+2}]$) ensures that, keeping β fixed, $h_b'(\rho) \le \frac{1}{2}$ for $\rho \ge \frac{M}{c}$. Then, we will show that we have $h_b(\rho) = \frac{\rho}{2}$ for the smallest choices of ρ and the largest choice of b (corresponding to the smallest choice of β since $b = \frac{1-\beta}{\beta}$ increases as β decreases). Finally, since $\left(\frac{\rho}{2}\right)' = \frac{1}{2}$, and $h_b(\rho)$ is simultaneously increasing in b and ρ, we conclude that $h_b(\rho) \le \frac{\rho}{2}$ for all $\rho \ge \frac{M}{c}, b \in (0, \frac{1}{c+2\sqrt{c}+2}]$.

Step 1. Show $h_b'(\rho) \le \frac{1}{2}$ for all $\rho \ge \frac{M}{c}$ and $b \in (0, \frac{1}{c+2\sqrt{c}+2}]$: Taking derivatives of $h_b(\rho)$ with respective to ρ gives that

$$h_b'(\rho) = b + b^2 \frac{1}{2\sqrt{M^2 + \frac{2M}{b}(\rho + \frac{M}{2})}} \frac{2M}{b} = b + \frac{Mb}{\sqrt{M^2 + \frac{2M}{b}(\rho + \frac{M}{2})}}. \tag{34}$$

Since $\rho + \frac{M}{2} \ge \rho$ and $\rho \ge \frac{M}{c}$, we can write

$$M^2 + \frac{2M}{b}\left(\rho + \frac{M}{2}\right) \ge M^2 + \frac{2M}{b}\rho \ge M^2 + \frac{2M^2}{bc}$$

so that a specialized upper bound on $h_b'(\rho)$ is given by

$$\tilde{h}(b) := b + \frac{Mb}{\sqrt{M^2 + \frac{2M^2}{bc}}} = b + \frac{b}{\sqrt{1 + \frac{2}{bc}}}. \tag{35}$$

We shift our attention to obtaining $\tilde{h}(b) \le \frac{1}{2}$ from which the claim follows. Observe that $\tilde{h}(b)$ is increasing in b. The range of $\beta \in [\frac{c+2\sqrt{c}+2}{c+2\sqrt{c}+3}, 1)$ corresponds to the range $b \in (0, \frac{1}{c+2\sqrt{c}+2}]$. We aim to show the claim at the extreme choice $b_{\max} := \frac{1}{c+2\sqrt{c}+2}$. Direct computation gives

$$\begin{aligned}
\tilde{h}(b_{\max}) &= \frac{1}{c+2\sqrt{c}+2}\left(1 + \frac{\sqrt{c}}{\sqrt{3c+4\sqrt{c}+4}}\right) \\
&\le \frac{1}{c+2\sqrt{c}+2}\left(1 + \frac{\sqrt{c}}{2+\sqrt{c}}\right) && \text{(36a)} \\
&= \frac{2(\sqrt{c}+1)}{(c+2\sqrt{c}+2)(\sqrt{c}+2)} \\
&\le \frac{2(\sqrt{c}+1)}{4(\sqrt{c}+1)} = \frac{1}{2}, && \text{(36b)}
\end{aligned}$$

where the equations (36a) and (36b) apply the two elementary inequalities, respectively:

- $(\sqrt{c}+2)^2 = c + 4\sqrt{c} + 4 \le 3c + 4\sqrt{c} + 4$
- $4(\sqrt{c}+1) \le c^{3/2} + 4c + 6\sqrt{c} + 4 = (\sqrt{c} + 2\sqrt{c} + 2)(\sqrt{c}+2)$.

Step 2. Show that the choice $\rho = \frac{M}{c}, b_{\max} = \frac{1}{c+2\sqrt{c}+2}$ obtains equality, i.e. $h_{b_{\max}}\left(\frac{M}{c}\right) = \frac{\rho}{2}$:
Under this choice of parameters, we have

$$M = c\rho, \qquad \text{so that} \qquad M + \frac{\rho}{2} = \rho\left(1 + \frac{c}{2}\right), \tag{37}$$

and thus the expression within the radical of $h_{b_{\max}}\left(\frac{M}{c}\right)$ becomes

$$M^2 + 2M(c + 2\sqrt{c} + 2)(\rho + \frac{M}{2}) = c^2\rho^2 + 2c\rho^2(c + 2\sqrt{c} + 2)\big(1 + \frac{c^2}{2}\big) \tag{38a}$$

$$= \rho^2 \left(c^2 + 2c\big(c + 2\sqrt{c} + 2\big)\big(1 + \frac{c^2}{2}\big)\right) \tag{38b}$$

$$= \rho^2 \left(c^2 + c\big(c^2 + 2c^{3/2} + 4c + 4c^{1/2} + 4\big)\right) \tag{38c}$$

$$= \rho^2 c(c + c^{1/2} + 2)^2 \tag{38d}$$

where (38a) applies (37) and our choice of β, (38b) factors ρ^2 out of (38a), (38c) multiplies out the expressions within the parentheses, and (38d) notices that the expression in (38c) a square. We obtain

$$h_{b_{\max}}(\frac{M}{c}) = b_{\max}^2 \left(c\rho + \frac{1}{b_{\max}}\rho\big(1 + \frac{c^2}{2}\big) + \rho c^{1/2}\big(c + c^{1/2} + 2\big)\right) \tag{39a}$$

$$= \rho b_{\max}^2 \left(c + \big(c + 2\sqrt{c} + 2\big)\big(1 + \frac{c^2}{2}\big) + c^{1/2}\big(c + c^{1/2} + 2\big)\right) \tag{39b}$$

$$= \rho b_{\max}^2 \frac{(c + 2\sqrt{c} + 2)^2}{2} \tag{39c}$$

$$= \frac{\rho}{2},$$

where (39a) substitutes (37) and (38d) into the definition of $h_{b_{\max}}\left(\frac{M}{c}\right)$, (39b) factors ρ out, and (39c) recognizes a square once more.

Step 3. Finally, we note that $h_b(\rho)$ increases when b increases or ρ increases. From step 2, we know that the largest choice of $b = b_{\max}$ and the smallest choice of $\rho = \frac{M}{c}$ gives $h_{b_{\max}}\left(\rho\right) = \frac{\rho}{2}$. To conclude, we only need to argue that for all $\rho \geq \frac{M}{c}$, we have $h_{b_{\max}}\left(\rho\right) \leq \frac{\rho}{2}$. This is indeed the case due to step 1. In other words, step 1 tells us $h_b'(\rho) \leq \frac{1}{2}$ for all $\rho \geq \frac{M}{c}$ and $b \in (0, b_{\max}]$, and we know $(\frac{\rho}{2})' = \frac{1}{2}$. Since $h_{b_{\max}}\left(\rho\right) = \frac{\rho}{2}$ when $\rho = \frac{M}{c}$, we conclude that $h_{b_{\max}}(\rho)$ remains bounded by $\frac{\rho}{2}$ as ρ increases from $\frac{M}{c}$. □

A.5. Equivalence between the double-loop Algorithm 1 and classical single-loop PBM

This subsection discusses the equivalence between Algorithm 1 and the classical single-loop PBM. In other words, we show that Algorithm 1 is a reformulation of the classical PBM. Similar discussion can also be found in Liao and Zheng (2025a, Section 4.3). Let us first review the classical single-loop PBM in Algorithm 6. At every iteration k, it solves the following subproblem

$$z_{k+1} = \operatorname*{argmin}_y \left\{ f_k(y) + \frac{\rho}{2}\|y - y_k\|^2 \right\}$$

to get candidate solution z_{k+1}. To decide if z_{k+1} makes sufficient descent, we adapt the test

$$\beta(f(y_k) - f_k(z_{k+1})) \leq f(y_k) - f(z_{k+1}), \tag{40}$$

where $\beta \in (0, 1)$ is a pre-defined constant. If (40) holds, then we set $y_{k+1} = z_{k+1}$ (known as a descent step). Otherwise, we set $y_{k+1} = y_k$ (this is called a null step). Afterwards, regardless of a descent step or a null step, the algorithm updates the model f_{k+1} following Assumption 2.

Algorithm 6 Classical proximal bundle method

Require: $y_1 \in \mathbb{R}^n, T > 0, \rho > 0, \beta \in (0,1)$
  for $k = 1, 2, \ldots, T$ **do**
    Compute $z_{k+1} = \operatorname{argmin}_y \left\{ f_k(y) + \frac{\rho}{2}\|y - y_k\|^2 \right\}$;
    if $\beta(f(y_k) - f_k(z_{k+1})) \leq f(y_k) - f(z_{k+1})$ **then**
      Set $y_{k+1} = z_{k+1}$; *Descent step*
    else
      Set $y_{k+1} = y_k$; *Null step*
    end if
    Construct f_{k+1} that approximates $f(\cdot)$ satisfying Assumption 2
  end for

Below, we explain the equivalence between Algorithms 1 and 6. First, the test (40) is the same as the test (8). Second, the under-estimator $\tilde{f}_j$ in Algorithm 2 of Algorithm 1 is equivalent to the f_k in Algorithm 6. The $\tilde{f}_j$ in Algorithm 6 resets the index j whenever a new center point is acquired, whereas Algorithm 6 keeps the iteration count k throughout the whole process. Therefore, the only difference is the notation and indices. One can view the oracle $\texttt{ProxDescent}(y_k, \beta, \rho)$ in Algorithm 1 as a cycle of null steps in Algorithm 6 and the satisfaction of the stopping criterion in $\texttt{ProxDescent}(y_k, \beta, \rho)$ corresponds to the fulfillment of the test (40). Third, the assumptions for constructing the under-estimators in both Algorithms 1 and 6 are the same, i.e., Assumptions 1 and 2 are the same.

In summary, Algorithm 1 is a re-interpretation of the classical single-loop PBM Algorithm 6. Algorithm 1, however, provides a cleaner distinction between null and descent steps.

Assumption 2 *The convex function f_{k+1} satisfies three conditions:*

1. ***Lower approximation:*** *Global convex lower approximation, $f_{k+1}(y) \leq f(y), \forall y \in \mathbb{R}^n$.*
2. ***Subgradient lower bound:*** *We have $f_{k+1}(y) \geq f(z_{k+1}) + \langle g_{k+1}, y - z_{k+1} \rangle, \forall y \in \mathbb{R}^n$, where g_{k+1} satisfies $g_{k+1} \in \partial f(z_{k+1})$.*
3. ***Aggregation from the past approximation:*** *If fails, then we require $f_{k+1}(y) \geq f_k(z_{k+1}) + \langle s_{k+1}, y - z_{k+1} \rangle, \forall y \in \mathbb{R}^n$, where $s_{k+1} = \rho(y_k - z_{k+1}) \in \partial f_k(z_{k+1})$.*

Appendix B. Convergence of Algorithm 5

Here, we prove the convergence of Algorithm 5, i.e., Theorem 2. Our proof in this section largely follows Monteiro and Svaiter (2013, Section 3). We provide the proof details for self-containment. For convenience, we restate Theorem 2 below.

Theorem 2 *Monteiro and Svaiter (2013, Theorem 3.8) Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function and set $S = \operatorname{argmin}_x f(x)$. Assume $S \neq \emptyset$. The sequence $\{x_k\}$ in Algorithm 5 satisfies*

$$f(x_k) - \min_x f(x) \leq \frac{2\rho \cdot \operatorname{dist}(x_0, S)^2}{k^2}, \ \forall k \geq 1.$$

Before providing a detailed proof, we first present an outline of the proof, which singles out the key inequalities.

Outline of the proof. The core of the proof is to establish the following two key inequalities

$$A_k \geq \frac{k^2}{4}, \ \forall k \geq 1, \tag{41a}$$

$$A_k f(x_k) + \frac{\rho}{2}\|x - z_k\|^2 \leq A_k f(x) + \frac{\rho}{2}\|x - z_0\|^2, \quad \forall x \in \mathbb{R}^n. \tag{41b}$$

Plugging $x = \operatorname{argmin}_{y \in S} \|y - z_0\|$ into (41b) and using (41a) yields the desired result

$$f(x_k) - \min_x f(x) \leq \frac{\rho \operatorname{dist}(z_0, S)^2}{2A_k} \leq \frac{2\rho \operatorname{dist}(z_0, S)^2}{k^2}, \ \forall k \geq 1.$$

□

In the next two subsections, we prove the two key inequalities (41a) and (41b).

B.1. Proof of (41a)

Recall that the update of a_k gives

$$a_k = \frac{1}{2} + \frac{\sqrt{1 + 4A_k}}{2} \geq \frac{1}{2} + \sqrt{A_k}, \tag{42}$$

The update $A_{k+1} = A_k + a_k$ along with (42) gives

$$A_{k+1} \geq A_k + \frac{1}{2} + \sqrt{A_k} \geq A_k + \sqrt{A_k} + \frac{1}{4} = \left(\sqrt{A_k} + \frac{1}{2}\right)^2$$

$$\implies \quad \sqrt{A_{k+1}} \geq \sqrt{A_k} + \frac{1}{2}, \tag{43}$$

where we took squareroots in (43). Adding (43) from $i = 0$ to $k - 1$ with $A_0 = 0$ shows

$$\sqrt{A_k} \geq \frac{k}{2} \quad \Rightarrow \quad A_k \geq \frac{k^2}{4}.$$

□

B.2. Proof of (41b)

For notational convenience, given $k \geq 0$, we define the following functions

$$\gamma_{k+1}(x) := f(x_{k+1}) + \langle g_{k+1}, x - x_{k+1} \rangle - \epsilon_{k+1}, \tag{44}$$

$$\Gamma_0 \equiv 0, \quad \Gamma_{k+1} = \frac{A_k}{A_{k+1}}\Gamma_k + \frac{a_k}{A_{k+1}}\gamma_{k+1}, \tag{45}$$

as well as the quantity

$$\beta_k = \inf_{z \in \mathbb{R}^n} \left(A_k \Gamma_k(z) + \frac{\rho}{2}\|z - z_0\|^2\right) - A_k f(x_k). \tag{46}$$

The function γ_{k+1} is nothing but an under-estimator of f, i.e., $\gamma_{k+1} \leq f$, as $g_{k+1} \in \partial_{\epsilon_{k+1}} f(x_{k+1})$. The function Γ_{k+1} is a convex combination of Γ_k and γ_{k+1}. The quantity β_k will serve as an important element in the proof.

Below, we establish some useful properties for γ_k and Γ_k.

Lemma 4 (Monteiro and Svaiter (2013, Lemma 3.2)) *For integers $k \geq 0$, the following hold:*

(a) γ_{k+1} *is affine and* $\gamma_{k+1} \leq f$,

(b) Γ_k *is affine and* $A_k\Gamma_k \leq A_k f$,

(c) $z_k = \operatorname{argmin}_{z\in\mathbb{R}^n}\{A_k\Gamma_k(z) + \frac{\rho}{2}\|z - z_0\|^2\}$,

(d) $A_k\Gamma_k(x) + \frac{\rho}{2}\|x - z_0\|^2 = A_k\Gamma_k(z_k) + \frac{\rho}{2}\|z_k - z_0\|^2 + \frac{\rho}{2}\|x - z_k\|^2, \forall x \in \mathbb{R}^n$.

Proof

(a) For $k \geq 0$, γ_{k+1} is affine as it is the sum of a linear function and a constant, and $\gamma_{k+1} \leq f$ as the definition $\gamma_{k+1}(x) = f(x_{k+1}) + \langle g_{k+1}, x - x_{k+1}\rangle - \epsilon_{k+1}$ and $g_{k+1} \in \partial_{\epsilon_{k+1}} f(x_{k+1})$.

(b) As Γ_k is a linear combination of $\{\gamma_i\}_{i=1}^k$ for $k \geq 1$, we know that Γ_k is affine by (a). We now show that $A_k\Gamma_k \leq A_k f$ by induction.

- Note that $\Gamma_0 = 0$ and $A_0 = 0$, so $A_k\Gamma_k \leq A_k f$ for $k = 0$ holds trivially.
- Suppose that $A_k\Gamma_k \leq A_k f$ for some $k > 0$. By the update (45), we have that

$$\begin{aligned} A_{k+1}\Gamma_{k+1} &= A_k\Gamma_k + a_k\gamma_{k+1} \\ &\leq A_k f + a_k f = A_{k+1} f \end{aligned}$$

  where the inequality follows from the induction hypothesis and (a), and the last equality uses the update $A_{k+1} = A_k + a_k$.

(c) From the update of z_k and assumption $z_0 = x_0$, we can rewrite z_k as

$$z_k = z_0 - \sum_{i=1}^{k} a_{i-1} g_{i-1}/\rho. \tag{47}$$

The claim would be established if we can show

$$A_k \nabla\Gamma_k(x) = \sum_{i=1}^{k} a_{i-1} g_{i-1} \quad \text{for } k \geq 1, \tag{48}$$

since (47) and (48) lead to

$$z_k = x_0 - A_k \nabla\Gamma_k(z_k)/\rho,$$

which is exactly the optimality condition of $z_k = \operatorname{argmin}_{z\in\mathbb{R}^n}\{A_k\Gamma_k(z) + \frac{\rho}{2}\|z - z_0\|^2\}$.

Below, we establish (48) by induction.

- Notice that $a_0 = A_1 = 1$ and $\nabla\Gamma_1(x) = g_0$, so (48) holds for $k = 1$.
- Assume that (48) holds for some $k > 1$. Invoking (45) once more, we have that

$$\begin{aligned} A_{k+1}\nabla\Gamma_{k+1} &= A_k\nabla\Gamma_k + a_k\nabla\gamma_{k+1} \\ &= \sum_{i=1}^{k} a_{i-1} g_{i-1} + a_k g_k = \sum_{i=1}^{k+1} a_{i-1} g_{i-1}. \end{aligned}$$

This proves (c).

(d) By (b), Γ_k is affine so there is a decomposition

$$A_k\Gamma_k(x) + \frac{\rho}{2}\|x - x_0\|^2 = \langle s_k, x\rangle + c_k + \frac{\rho}{2}\|x - x_0\|^2 \quad \text{for some } s_k, c_k \in \mathbb{R}^n. \tag{49}$$

The function above has the minimum z_k by (c), where we necessarily have $z_k = z_0 - s_k/\rho$. It follows that

$$A_k\Gamma_k(x) + \frac{\rho}{2}\|x - z_0\|^2 = \langle s_k, x\rangle + c_k + \frac{\rho}{2}\|x - z_k\|^2 - \langle s_k, x - z_k\rangle + \frac{1}{2\rho}\|s_k\|^2 \tag{50}$$

$$= \langle s_k, z_k\rangle + c_k + \frac{\rho}{2}\|x - z_k\|^2 + \frac{\rho}{2}\|z_k - x_0\|^2 \tag{51}$$

$$= A_k\Gamma_k(z_k) + c_k + \frac{\rho}{2}\|x - z_k\|^2 + \frac{\rho}{2}\|z_k - x_0\|^2, \tag{52}$$

where (50) uses the decomposition (49) and $z_k = z_0 - s_k/\rho$, (51) collects terms and applies the relation $z_k = z_0 - s_k/\rho$ again, and (52) recalls (49). This completes the proof of (d).

■

Points (a) and (b) in Lemma 4 show that the functions γ_{k+1} and Γ_k are affine and lower than the function f. Point (c) reveals that the iterate z_k is in fact the minimizer of the function $A_k\Gamma_k(\cdot) + \frac{\rho}{2}\|\cdot - z_0\|^2$. Point (d) is an identity that will be useful later.

The next lemma establishes another useful identity.

Lemma 5 *The sequences $\{a_k\}$ and $\{A_k\}$ in Algorithm 5 satisfy $A_{k+1} = a_k^2$ for all $k \geq 0$.*

Proof From the update of a_k, it is clear that a_k is a positive root of the equation

$$a^2 - a - A_k = 0,$$

where A_k is given. Plugging a_k into the above equation and rearranging terms gives us $a_k^2 = A_k + a_k = A_{k+1}$, where the last equality comes from the update of A_{k+1}. ■

We next review an identity without proof, as it is independent of the algorithm.

Lemma 6 (Monteiro and Svaiter (2013, Lemma 3.3)) *Take $\tilde{x}, \tilde{y}, \tilde{g} \in \mathbb{R}^n$ and $\rho, \epsilon > 0$. Then the inequality*

$$\|\frac{1}{\rho}\tilde{g} + \tilde{y} - \tilde{x}\|^2 + \frac{2}{\rho}\epsilon \leq \|\tilde{x} - \tilde{y}\|^2 \tag{53}$$

holds if and only if

$$\min_{x\in\mathbb{R}^n}\left\{\langle\tilde{g}, x - \tilde{y}\rangle - \epsilon + \frac{\rho}{2}\|x - \tilde{x}\|^2\right\} \geq 0.$$

Recall the inexactness condition (12) in Algorithm 5:

$$x_{k+1} = y_k - \frac{1}{\rho}g_{k+1},\ g_{k+1} \in \partial_{\epsilon_{k+1}} f(x_{k+1}),\ 2\epsilon_{k+1} \leq \rho\|x_{k+1} - y_k\|^2,\ \forall k \geq 0.$$

Using Lemma 6 with $\tilde{g} = g_{k+1}$, $\tilde{x} = y_k$ and $\tilde{y} = x_{k+1}$, we then know that

$$\min_{x\in\mathbb{R}^n}\left\{\langle g_{k+1}, x - x_{k+1}\rangle - \epsilon_{k+1} + \frac{\rho}{2}\|x - x_{k+1}\|^2\right\} \geq 0. \tag{54}$$

The following lemma establishes the nondecreasing property of the sequence $\{\beta_k\}$.

Lemma 7 *The sequence* $\{\beta_k\}_{k=1}^{\infty}$ *defined in* (46) *satisfies* $0 \leq \beta_k \leq \beta_{k+1}, \forall k \geq 0$.

Proof As $A_0 = 0$, one easily has that

$$\beta_0 = \inf_{x \in \mathbb{R}^n} \left\{ \frac{\rho}{2} \|x - x_0\|^2 \right\} = 0.$$

We move on to proving $\beta_{k+1} \geq \beta_k$ when $k > 0$. For any $y \in \mathbb{R}^n$, define

$$\tilde{y} := \frac{A_k}{A_{k+1}} x_k + \frac{a_k}{A_{k+1}} y.$$

By the update of $y_k = \frac{A_k}{A_{k+1}} x_k + \frac{a_k}{A_{k+1}} z_k$, the update $A_{k+1} = A_k + a_k$, and the fact that γ_{k+1} is affine, the following two facts hold:

$$\tilde{y} - y_k = \frac{a_k}{A_{k+1}} (y - z_k), \tag{55}$$

$$\gamma_{k+1}(\tilde{y}) = \frac{A_k}{A_{k+1}} \gamma_{k+1}(x_k) + \frac{a_k}{A_{k+1}} \gamma_{k+1}(y). \tag{56}$$

By the definition (45), we observe that for all y we have

$$\begin{aligned} A_{k+1}\Gamma_{k+1}(y) + \frac{\rho}{2}\|y - z_0\|^2 &= a_k \gamma_{k+1}(y) + A_k \Gamma_k(y) + \frac{\rho}{2}\|y - z_0\|^2 \\ &= a_k \gamma_{k+1}(y) + A_k \Gamma_k(z_k) + \frac{\rho}{2}\|z_k - z_0\|^2 + \frac{\rho}{2}\|y - z_k\|^2 \qquad (57) \\ &= a_k \gamma_{k+1}(y) + A_k f(x_k) + \beta_k + \frac{\rho}{2}\|y - z_k\|^2, \qquad (58) \end{aligned}$$

where (57) applies (d) in Lemma 4, and (58) is due to (c) of Lemma 4 and the definition of β_k (46). Now, as $\gamma_{k+1} \leq f$ by Lemma 4, we may write

$$\begin{aligned} A_{k+1}\Gamma_{k+1}(y) + \frac{\rho}{2}\|y - z_0\|^2 &\geq \beta_k + a_k \gamma_{k+1}(y) + A_k \gamma_{k+1}(x_k) + \frac{\rho}{2}\|y - z_k\|^2 \\ &= \beta_k + A_{k+1}\gamma_{k+1}(\tilde{y}) + \frac{\rho A_{k+1}^2}{2a_k^2}\|\tilde{y} - y_k\|^2 \\ &= \beta_k + A_{k+1}\gamma_{k+1}(\tilde{y}) + \frac{\rho A_{k+1}}{2}\|\tilde{y} - y_k\|^2, \qquad (59) \end{aligned}$$

where the first equality applies (55) and (56), and the second equality uses $A_{k+1} = a_k^2$ from Lemma 5. Evaluating γ_{k+1} at $\tilde{y}$ and applying (54) shows

$$\begin{aligned} \gamma_{k+1}(\tilde{y}) + \frac{\rho}{2}\|\tilde{y} - y_k\|^2 &= f(x_{k+1}) + \left(\langle g_{k+1}, \tilde{y} - x_{k+1} \rangle - \epsilon_{k+1} + \frac{\rho}{2}\|\tilde{y} - y_k\|^2 \right) \\ &\geq f(x_{k+1}). \qquad (60) \end{aligned}$$

As A_{k+1} is nonnegative, substituting (60) into (59) and taking the infimum over y shows

$$\begin{aligned} \beta_k + A_{k+1} f(x_{k+1}) &\leq \inf_{y \in \mathbb{R}^n} \left(A_{k+1}\Gamma_{k+1}(y) + \frac{\rho}{2}\|y - z_0\|^2 \right) \\ &= \beta_{k+1} + A_{k+1} f(x_{k+1}), \end{aligned}$$

where the equality is from (46). Subtracting $A_{k+1}f(x_{k+1})$ on both sides of the above inequality finishes the proof. ■

We are ready to fully establish (41b). As $\beta_k \geq 0$, it follows that

$$\begin{aligned} A_k f(x_k) &\leq \inf_{x' \in \mathbb{R}^n} \left(A_k \Gamma_k(x') + \frac{\rho}{2}\|x' - z_0\|^2 \right) \\ &= A_k \Gamma_k(z_k) + \frac{\rho}{2}\|z_k - z_0\|^2, \end{aligned}$$

where we applied (c) of Lemma 4 in the second line. Adding the quadratic $\frac{\rho}{2}\|x - z_k\|^2$ on both sides of the equation above yields the relation

$$\begin{aligned} A_k f(x_k) + \frac{\rho}{2}\|x - z_k\|^2 &\leq A_k \Gamma_k(z_k) + \frac{\rho}{2}\|z_k - z_0\|^2 + \frac{\rho}{2}\|x - z_k\|^2 \\ &= A_k \Gamma_k(x) + \frac{\rho}{2}\|x - z_0\|^2 \\ &\leq A_k f(x) + \frac{\rho}{2}\|x - z_0\|^2, \end{aligned}$$

where the equality in the second line applies (d) of Lemma 4, and the last line uses part (b) of Lemma 4.