On the Acceleration of Proximal Bundle Methods

David Fersztand[1] and Andy Sun[1,2]

[1]Operations Research Center, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, 02139, MA, USA.
[2]Sloan School of Management, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, 02139, MA, USA.

Abstract

The proximal bundle method (PBM) is a fundamental and computationally effective algorithm for solving nonsmooth optimization problems. In this paper, we present the first variant of the PBM for smooth objectives, achieving an accelerated convergence rate of $\mathcal{O}(\frac{1}{\sqrt{\epsilon}}\log(\frac{1}{\epsilon}))$, where ϵ is the desired accuracy. Our approach addresses an open question regarding the convergence guarantee of proximal bundle type methods, which was previously posed in two recent papers. We interpret the PBM as a proximal point algorithm and base our proposed algorithm on an accelerated inexact proximal point scheme. Our variant introduces a novel *null step test* and oracle while maintaining the core structure of the original algorithm. The newly proposed oracle substitutes the traditional cutting planes with a smooth lower approximation of the true function. We show that this smooth interpolating lower model can be computed as a convex quadratic program. We also examine a second setting where Nesterov acceleration can be effectively applied, specifically when the objective is the sum of a smooth function and a piecewise linear one.

Keywords: Proximal bundle method, smooth optimization, Nesterov acceleration, iteration complexity

1 Introduction

1.1 Problem setup

We consider the Proximal Bundle type algorithms applied to the unconstrained minimization problem

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

where $f : \mathbb{R}^n \to \mathbb{R}$ is convex and L_f-smooth, i.e. f has Lipschitz continuous gradient:

$$\|\nabla f(y) - \nabla f(x)\| \leq L_f \|x - y\|, \quad \forall x, y \in \mathbb{R}^n.$$

1.2 Motivation

In [1], Liang and Monteiro derived a convergence rate of $\mathcal{O}(\frac{1}{\epsilon^2})$ up to logarithmic terms for the problem (1). They conclude their analysis by noting that this convergence rate *is not optimal when $L_f > 0$. It would be interesting to design an accelerated variant of the PBM that is optimal when $L_f > 0$.*

In the same line of work, Diaz and Grimmer in [2] studied the convergence rates of the classic PBM for a variety of non-smooth convex optimization problems. They concluded the paper with the statement that *"it is likely that a variant of the bundle method can achieve accelerated rates, either by enforcing additional assumptions about the models f_k or by modifying the logic of the algorithm. We leave this as an intriguing open question for future research."*

In light of these remarks, it seems that the acceleration of the PBM for a smooth objective remains an open question. The complexity analysis of PBM is notoriously difficult. It is our hope that the investigation of modified bundle methods, such as the one we propose in Algorithm 3, may ultimately lead to a better understanding of PBM. For smooth convex optimization problems, when the value of the smoothness parameter is available, there exist methods with a convergence rate $\mathcal{O}(\frac{1}{\sqrt{\epsilon}})$, such as the accelerated gradient descent algorithm. This rate is known to be optimal for first-order methods [3]. We prove that, with a slight modification of the bundle minimization oracle and the *null step test*, but keeping the overall structure of the algorithm, an accelerated rate can be achieved up to log terms. More precisely, an ϵ-optimal solution can be reached in $\mathcal{O}\left(\frac{1}{\sqrt{\epsilon}} \log\left(\frac{1}{\epsilon}\right)\right)$ iterations improving upon the standard PBM for which the best-known rate is $\mathcal{O}(\frac{1}{\epsilon})$.

When the objective is no longer smooth, but is the sum of a smooth convex function and a piecewise linear convex function, the best-known convergence rate is $\mathcal{O}\left(\frac{1}{\epsilon^{4/5}} \log\left(\frac{1}{\epsilon}\right)\right)$ [4]. This paper proposes an accelerated variant of PBM for this setup that exhibits the same accelerated convergence rate of $\mathcal{O}\left(\frac{1}{\sqrt{\epsilon}} \log\left(\frac{1}{\epsilon}\right)\right)$.

1.3 Contribution

1. We establish the existence of a minimal smooth lower interpolating model, demonstrating that it can be constructed by solving a convex quadratically constrained quadratic program (QCQP).
2. We introduce a novel Accelerated Proximal Bundle algorithm, which incorporates modifications to both the oracle and the *null step test* while preserving the fundamental structure and logic of the traditional proximal bundle algorithm. Moreover, the proposed *null step test* does not rely on the tolerance ϵ, positively answering an open question in [1].

3. We show that our proposed algorithm achieves an iteration complexity of $\mathcal{O}\left(\frac{1}{\sqrt{\epsilon}}\log\left(\frac{1}{\epsilon}\right)\right)$. This result addresses an open question posed by [1] and [2], providing the first accelerated rate for the PBM.
4. We propose an acceleration scheme for composite objectives, consisting of a smooth component and a piecewise linear one, and establish an iteration complexity of $\mathcal{O}\left(\frac{1}{\sqrt{\epsilon}}\log\left(\frac{1}{\epsilon}\right)\right)$.

1.4 Related work

Kelley's method and bundle methods

Kelley's cutting plane method (1960) [5] is a foundational algorithm for minimizing non-smooth convex objectives. It iteratively refines a piecewise linear lower approximation of the true function. The PBM [6, 7], the trust region bundle method [8], and the level bundle method [9, 10] are variants of Kelley's method. Bundle methods were initially designed for non-smooth optimization because solving the proximal bundle subproblem is considered computationally expensive. This expense is justifiable for nonsmooth problems due to their inherently higher complexity. All these variants have been proven to converge to an optimal solution for any choice of parameter, contrary to gradient descent and its accelerated versions that require a stepsize chosen at most inversely proportional to the smoothness level. Likewise, subgradient methods necessitate diminishing stepsize sequences. These less complex algorithms might not converge if the stepsizes are not cautiously handled, thereby offering a persuasive argument to contemplate bundle methods.

Convergence rates of the proximal bundle algorithm

In 2000, Kiwiel [11] established the first convergence rate for the proximal bundle method, proving that an ϵ-minimizer can be found in $\mathcal{O}(\frac{1}{\epsilon^3})$ iterations. [2] improved the complexity analysis in the unconstrained case while restricting the proximity parameter schedule. They provided convergence rates for all combinations of smoothness and strong convexity. Namely, $\mathcal{O}(\frac{1}{\epsilon^2})$ when the function is only Lipschtiz continuous, $\mathcal{O}(\frac{1}{\epsilon})$ when the function is smooth or strongly convex, removing the log term in [12]. When the function is smooth and strongly convex, the convergence rate becomes $\mathcal{O}(\log\left(\frac{1}{\epsilon}\right))$. In [13] and [1], the authors proposed a new type of *null step test* that is discussed in Section 2.2. Their analysis provides an $\mathcal{O}(\frac{1}{\epsilon^2}\log(\frac{1}{\epsilon}))$ convergence rate when f is only Lipschitz continuous, $\mathcal{O}(\frac{1}{\epsilon}\log(\frac{1}{\epsilon}))$ when f is smooth.

Proximal point algorithm

The PBM can also be interpreted as an inexact proximal point method [14, 15]. Güler in [16] first proposed an inexact accelerated version of the proximal point algorithm. It reaches an $\mathcal{O}(\frac{1}{\sqrt{\epsilon}})$ convergence rate by requiring the approximation error of the proximal problem to be summable. Note that it does not require the objective function to be smooth. In [17], Salzo and Villa corrected a subtle error in the convergence proof in [16] with a slight modification of the hypothesis. He and Yuan in [18] proposed a practical inexact criterion for the proximal point algorithm, eligible for Nesterov-type acceleration.

Acceleration of bundle methods

Lan [10] showed that the bundle level method can be modified to incorporate Nesterov-type acceleration. The algorithm is agnostic about the smoothness of the function in the sense that it achieves the optimal convergence rate for both smooth and nonsmooth problems (respectively $\mathcal{O}(\frac{1}{\sqrt{\epsilon}})$ and $\mathcal{O}(\frac{1}{\epsilon^2})$) without requiring the smoothness of the parameter as an input. To the best of our knowledge, no similar acceleration scheme is available for the PBM.

1.5 The Proximal Bundle and the Proximal Point Algorithms

For the sake of completeness, we present the formulations of the classic PBM (Algorithm 1) and an accelerated inexact proximal point algorithm (Algorithm 2). The two algorithms are presented for minimizing a proper, closed, nonsmooth convex objective f.

Algorithm 1 Proximal Bundle Method (PBM)

Require: $x_0 \in \mathbb{R}^n$, $\mathcal{I}_0$ an index set of initial cuts, $\rho > 0$, $\beta \in (0, 1)$

1: **for** $k \geq 0$ **do**

2: Compute y_{k+1} solving the following quadratic program

$$\begin{aligned} \min_{t \in \mathbb{R}, y \in \mathbb{R}^n} \quad & t + \frac{\rho}{2}\|y - x_k\|^2 \\ \text{s.t.} \quad & \forall i \in \mathcal{I}_k \quad t - f(y_i) - \langle v_i, y - y_i \rangle \geq 0 \end{aligned} \tag{2}$$

3: Compute $f(y_{k+1})$ and $v_{k+1} \in \partial f(y_{k+1})$

4: $\mathcal{I}_{k+1} \leftarrow \mathcal{I}_k \cup \{k+1\}$ ▷ (update piecewise linear model)

5: **if** $\beta(f(x_k) - f_k(y_{k+1})) \leq f(x_k) - f(y_{k+1})$ **then** ▷ (null step test)

6: $x_{k+1} \leftarrow y_{k+1}$ ▷ (serious step)

7: **else**

8: $x_{k+1} \leftarrow x_k$ ▷ (null step)

9: **end if**

10: **end for**

At each iteration k, the classic PBM solves a Quadratic Program (QP) in (2) that minimizes the sum of a piecewise linear lower model $f_k(y) := \max_{i \in \mathcal{I}_k} f(y_i) + \langle v_i, y - y_i \rangle$ with subgradient $v_i \in \partial f(y_i)$ and a quadratic proximal term. The solution of this optimization is denoted by y_{k+1}. The proximal center, which will be denoted by the letter x throughout the article, is only updated subject to the *null step test* on line 5 of Algorithm 1. Steps in which the proximal center is updated are called *serious steps*, while those where it remains unchanged are referred to as *null steps*. The proximal term and the *null step test* stabilize the convergence dynamics of the PBM algorithm compared to the classic cutting plane algorithm.

We also give here an accelerated inexact proximal point algorithm (a-IPPA), which applies Nesterov acceleration to the usual PPA [16, 18]. For $x_j \in \mathbb{R}^n$ and a convex function f, the prox operator is defined as $\text{prox}_{\rho,f}(x_j) = \arg\max_{y\in\mathbb{R}^n} f(y) + \frac{\rho}{2}\|y - x_j\|^2$.

Algorithm 2 Accelerated Inexact Proximal Point Algorithm (a-IPPA)

Require: a convex f, $x_0 \in \mathbb{R}^n$, a sequence $\rho_j > 0$

Set $\zeta_0 = x_0$ and $t_0 = 1$

for $j \geq 0$ **do**

$y_{j+1} \approx \text{prox}_{\rho_j,f}(x_j)$ ▷ (inexact proximal step)

$v_{j+1} \in \partial f(y_{j+1})$

$t_{j+1} \leftarrow \frac{1+\sqrt{1+4t_j^2}}{2}$

$\zeta_{j+1} \leftarrow x_j - \frac{1}{\rho_j} v_{j+1}$

$x_{j+1} \leftarrow \zeta_{j+1} + \frac{t_j - 1}{t_{j+1}}(\zeta_{j+1} - \zeta_j)$

end for

Interpreting the PBM as a proximal point algorithm, we develop an accelerated version of the PBM by specifying the inexact proximal step $y_{j+1} \approx \text{prox}_{\rho,f}(x_j)$. We will detail both the method for finding an approximate solution and the criteria denoted $\approx$ that must be satisfied in Algorithm 2 before updating the variables (t, ζ, x).

Paper overview

Section 2 focuses on the problem introduced in (1) with a smooth objective function. This study is organized as follows. In Section 2.1, we introduce our proposed algorithm. Section 2.2 outlines the key ideas behind the algorithm design and convergence proof. The new oracle, which constructs a smooth lower model, is detailed in Section 2.3. In Section 2.4, we show how the proposed *null step test* ensures an accelerated convergence rate for the number of *serious steps*. In Section 2.5, we provide a bound on the number of consecutive *null steps*. The main result of this paper is the convergence rate of our proposed algorithm and is provided in Section 2.6.

Section 3 presents an accelerated scheme in another setup in which the objective is not supposed to be smooth anymore. Instead, it is the sum of a convex smooth function and a convex piecewise linear one.

1.6 Definitions and notations

We define $\|\cdot\|$ as the Euclidean norm and $\langle\cdot,\cdot\rangle$ the corresponding inner product in the n-dimensional Euclidean space $\mathbb{R}^n$.

The relative interior, closure, and convex hull of a set $S \subseteq \mathbb{R}^n$ are denoted as $\text{relint}(S)$, $\text{cl}(S)$, and $\text{conv}(S)$, respectively.

The conjugate [19] of a proper, convex function h is defined as

$$h^*(u) = \sup_{x\in\mathbb{R}^n} \{\langle x, u\rangle - h(x)\}.$$

For h_1, h_2, two real-valued functions of $\mathbb{R}^n$, "$\leq$" denotes the following property

$$h_1 \leq h_2 \Leftrightarrow h_1(x) \leq h_2(x), \quad \forall x \in \mathbb{R}^n.$$

We suppose that the problem (1) has at least one optimal solution, denoted as x^*. This work studies the worst-case iteration complexity for finding an ϵ-*optimal* solution. Namely, the number of first-order oracle queries to find $x \in \mathbb{R}^n$ such that $f(x) - f(x^*) \leq \epsilon$. To simplify the expression of the convergence rates, we assume that $\epsilon < 1$.

2 Accelerated PBM for smooth objectives

2.1 Algorithm description

In this section, we introduce the accelerated Proximal Bundle Method (a-PBM). It builds upon the accelerated inexact proximal point algorithm (Algorithm 2). Although Algorithm 3 and its corresponding analysis are presented with a fixed parameter ρ, the analysis can be adapted to allow for a decreasing proximity parameter between consecutive *null step* sequences.

2.2 Main ideas behind our proposed algorithm

An accelerated inexact proximal point algorithm

In [1], a variant of the PBM is presented that can be interpreted as an inexact proximal point algorithm (Algorithm 2). In [18], He and Yuan present a Nesterov acceleration of the classical proximal point algorithm with a convenient approximation error for the proximal point oracle.

A novel null step test

The usual *null step test* for the PBM compares the expected progress in the objective value to the actual progress made via

$$\beta(f(x_k) - f_k(y_{k+1})) \leq f(x_k) - f(y_{k+1}),$$

with β, a parameter in (0,1). This criterion does not seem to translate easily into an inexact criterion from the proximal point algorithm literature. In [13] and [1], the *null step test* is changed to

$$f(z_{k+1}) - f_k(y_{k+1}) \leq \frac{\epsilon}{2} + \frac{\rho}{2}\|y_{k+1} - x_k\|^2,$$

where z_{k+1} represents the best iterate for the proximal problem. Using this criterion that includes a quadratic term, the number of serious steps for their modified version of PBM is bounded by $\rho\|x_0 - x^*\|^2/\epsilon$. This test makes it difficult to relate to the accelerated inexact proximal point method literature as the error term $\frac{\rho}{2}\|y_{k+1} - x_k\|^2$ may be large. However, imposing a stricter test will likely increase the provable upper

Algorithm 3 Accelerated Proximal Bundle Method (a-PBM)

Require: f an L_f smooth convex function, $x_0 \in \mathbb{R}^n$, $\mathcal{I}_0$ an index set of initial cuts, $\rho > 0$, and $C = \frac{2L_f}{\rho}(\sqrt{2L_f/\rho} + 1)$.

Set $\zeta_0 = x_0$ and $t_0 = 1$
for $k \geq 0$ **do**
  Compute y_{k+1} solving the following convex QCQP :

$$\min_{t \in \mathbb{R}, y \in \mathbb{R}^n} \; t + \frac{\rho}{2}\|y - x_k\|^2 \tag{3a}$$

$$\text{s.t.} \quad \forall i \in \mathcal{I}_k \quad t - f(y_i) - \langle \nabla f(y_i), y - y_i \rangle \geq \frac{\rho^2}{2L_f}\|y - x_k + \frac{1}{\rho}\nabla f(y_i)\|^2. \tag{3b}$$

  Compute $f(y_{k+1})$ and $\nabla f(y_{k+1})$
  $\mathcal{I}_{k+1} \leftarrow \mathcal{I}_k \cup \{k+1\}$ ▷ (update piecewise linear model)
  if $C\|y_{k+1} - y_k\| \leq \|x_k - y_{k+1}\|$ **then** ▷ (null step test)
    $t_{k+1} \leftarrow \frac{1+\sqrt{1+4t_k^2}}{2}$
    $\zeta_{k+1} \leftarrow x_k - \frac{1}{\rho}\nabla f(y_{k+1})$ ▷ (serious step)
    $x_{k+1} \leftarrow \zeta_{k+1} + \frac{t_k - 1}{t_{k+1}}(\zeta_{k+1} - \zeta_k)$
  else
    $x_{k+1} \leftarrow x_k$
    $\zeta_{k+1} \leftarrow \zeta_k$ ▷ (null step)
    $t_{k+1} \leftarrow t_k$
  end if
end for

bound on the number of consecutive null steps. Indeed, the number of consecutive null steps is shown to be $\mathcal{O}(\log(\frac{1}{\epsilon}))$ through the geometric convergence of the quantity $m_{k+1} := f_k(y_{k+1}) + \frac{\rho}{2}\|y_{k+1} - x_k\|^2$. Although this linear convergence still holds with the oracle proposed in Section 2.3, it is not sufficient to prove the upper bound on the number of consecutive null steps with a stricter criterion like a summable approximation error, which is the one initially proposed in the first accelerated proximal point algorithms [16].

In order to overcome this difficulty, we prove the geometric decrease of another quantity: the gap between the best value of the current proximal problem encountered so far and the last value of the bundle subproblem. This quantity ξ is introduced in Section 2.5.2. As the quantity $\frac{\rho}{2}\|y_{k+1} - x_k\|^2$ still appears in the study of the number of consecutive null steps, a way to obtain the geometric decrease of ξ is to enforce a *null step test* that relates this quantity to the distance between the two last iterates.

From this result, we still need to show that such a test is sufficiently stringent to ensure the accelerated convergence rate of the corresponding inexact proximal point algorithm. It is indeed the case provided that the lower models f_k are L_f-smooth.

Smooth lower model functions

The cutting plane algorithm iteratively builds a lower approximation of f as the pointwise maximum of the supporting hyperplanes at the previously queried points. This model function is nonsmooth. It is also the lowest convex function that interpolates f in the gradients and function values at the previously queried points [20]. Instead, if the lower models are smooth, the inexact criterion of [18] translates into a comparison between $\|y_{k+1} - y_k\|$, the distance between the two last iterates, and $\|y_{k+1} - x_k\|$ the distance from the last iterate to the proximal center. Provided that this ratio is small enough, the number of serious steps is guaranteed to be $\mathcal{O}(\frac{1}{\sqrt{\epsilon}})$.

Remark 1 (Extending the algorithm to a composite setting)**.** *It is possible to consider a composite minimization setting throughout the proof as in [1]. The problem becomes*

$$\min_{x \in \mathbb{R}^n} f(x) + g(x),$$

with g a L_g-smooth convex function.

To maintain clarity, we have excluded the detailed modifications to the algorithm and the convergence rate proof, as they are straightforward. For instance, the bundle minimization oracle (3) would be replaced by

$$\begin{aligned} &\min_{t \in \mathbb{R}, y \in \mathbb{R}^n, u \in \mathbb{R}^n} t + g(y) + \frac{\rho}{2}\|y - x_k\|^2 \\ s.t. \quad &\forall i \in \mathcal{I}_j \quad t - f(y_i) - \langle \nabla f(y_i), y - y_i \rangle \geq \frac{1}{2L_f}\|u - \nabla f(y_i)\|^2. \end{aligned}$$

The constant C of the null step test would be replaced by $\frac{2L_f}{\rho}\left(\frac{\sqrt{2(L_f+L_g)}}{\sqrt{\rho}} + 1\right)$. In this setup, the convergence rate $\mathcal{O}(\frac{1}{\sqrt{\epsilon}}\log(\frac{1}{\epsilon}))$ still holds.

2.3 Building a smooth model

In this section, we show how the bundle minimization step (3) given in Algorithm 3 leads to lower models that display suitable properties for proving the accelerated convergence rate of the algorithm. It is a convex Quadratically Constrained Quadratic Program (QCQP) which can be efficiently solved.

Drori and Teboulle in [20] first interpreted Kelley's cutting plane method as iteratively minimizing the lowest convex function that interpolates the true function at the previous iterates. The method selects the next iterate to minimize the lowest convex function value. Additionally, this model function is not smooth, meaning it cannot truly represent the function f. However, when f is known to be smooth, we demonstrate that the lowest smooth model functions can be constructed by solving a nonconvex QCQP. We also show that this nonconvex QCQP is equivalent to a more tractable convex QCQP.

We consider a sequence of consecutive null steps j following the serious step k. At iteration j, we consider the bundle set $[k, j] \subset \mathcal{I}_j$. Compared to the classic PBM, we modify the bundle minimization step in order to build a sequence of approximation functions $f_j : \mathbb{R}^n \to \mathbb{R}$ that exhibit the following properties that for all j,

1. f_j is convex and L_f-smooth.
2. f_j interpolates f at the previous iterates, i.e., for all $i \in \mathcal{I}_j$, $f_j(y_i) = f(y_i)$ and $\nabla f_j(y_i) = \nabla f(y_i)$.
3. $f(y_{j+1}) \geq f_j(y_{j+1}) \geq f_{j-1}(y_{j+1})$, where $y_{j+1} = \arg\min_{y \in \mathbb{R}^n} f_j(y) + \frac{\rho}{2}\|y - x_j\|^2$ is the result of the bundle minimization step (3) in Algorithm 3.

The first property is useful to show that the *null step test* given in Algorithm 3 implies the inexact criterion of [18]. The other properties are used when deriving the bound on the number of consecutive null steps.

Definition 2. *For each iteration j of Algorithm 3, $\mathcal{F}_{L_f}(\mathcal{I}_j)$ is defined to be the set of convex, L_f-smooth functions $\hat{f} : \mathbb{R}^n \to \mathbb{R}$ such that $\hat{f}$ interpolates the function values and gradients of f at $(y_i)_{i \in \mathcal{I}_j}$, i.e., $\forall i \in \mathcal{I}_j$, $\hat{f}(y_i) = f(y_i)$ and $\nabla \hat{f}(y_i) = \nabla f(y_i)$.*

Remark 3 (Comparison with the recent approach by Florea and Nesterov [21])**.** *During the final editing of our paper, we became aware of a paper from Florea and Nesterov studying smooth lower models [21]. Their contribution focuses on a different algorithm, specifically the Gradient Method with memory. Although their approach diverges somewhat from ours, the conclusions drawn in Section 2.3 of their paper align with those presented in our Section 2.3. Notably, they demonstrate that a smooth lower bound can be computed by solving a convex QCQP.*

To achieve this, Florea and Nesterov consider a family of functions indexed by $i \in \mathcal{I}_j$, $\phi_i : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ and $p : \mathbb{R}^n \to \mathbb{R}$ such that

$$\forall (y,u) \in \mathbb{R}^n \times \mathbb{R}^n,\ \phi_i(y,u) \stackrel{def}{=} f(y_i) + \langle \nabla f(y_i), y - y_i \rangle + \frac{1}{2L_f}\|u - \nabla f(y_i)\|^2$$

$$p(y) \stackrel{def}{=} \min_{g \in \mathbb{R}^n} \max_{i \in \mathcal{I}_j} \phi_i(y, g). \tag{4}$$

To establish that p is smooth, convex, and interpolates f at the points $(y_i)_{i \in \mathcal{I}_j}$, Florea and Nesterov reformulate the inner maximization in (4) as a maximization over the simplex, leveraging strong duality and Danskin's theorem. In contrast, our approach begins with a nonconvex QCQP known to possess the desired properties outlined at the beginning of this section. We then demonstrate that this problem can be reduced to a more tractable convex QCQP.

2.3.1 A nonconvex QCQP with the desired properties

We now introduce a nonconvex but explicit formulation for the lower bound of smooth functions interpolating f at the previous iterates. In the next subsection, we will show that this nonconvex problem is equivalent to a convex QCQP with a formulation close to the usual proximal bundle minimization subproblem. More precisely, when $L_f = +\infty$ (*i.e.* the function is not smooth), the two bundle minimization problems (2) and (3) match exactly.

At each *null step*, we consider the following problem

$$\min_{t\in\mathbb{R},y\in\mathbb{R}^n,u\in\mathbb{R}^n} \ t+\frac{\rho}{2}\|y-x_k\|^2 \tag{5a}$$

$$\text{s.t.} \quad \forall i\in\mathcal{I}_j, \quad t-f(y_i)-\langle\nabla f(y_i),y-y_i\rangle \geq \frac{1}{2L_f}\|u-\nabla f(y_i)\|^2, \tag{5b}$$

$$\forall i\in\mathcal{I}_j, \quad f(y_i)-t-\langle u,y_i-y\rangle \geq \frac{1}{2L_f}\|u-\nabla f(y_i)\|^2. \tag{5c}$$

Note that (5) is a nonconvex QCQP due to the bilinear term $\langle u, y_i - y\rangle$ in (5c).

Lemma 4. *The above nonconvex QCQP* (5) *is equivalent to*

$$\min_{y\in\mathbb{R}^n,\hat{f}\in\mathcal{F}_{L_f}(\mathcal{I}_j)} \hat{f}(y)+\frac{\rho}{2}\|y-x_k\|^2=\min_{y\in\mathbb{R}^n}\left\{\frac{\rho}{2}\|y-x_k\|^2+\min_{\hat{f}\in\mathcal{F}_{L_f}(\mathcal{I}_j)}\hat{f}(y)\right\}. \tag{6}$$

In particular, this shows that (5) *is always feasible, as* $f\in\mathcal{F}_{L_f}(\mathcal{I}_j)\neq\varnothing$.

Proof. Problem (6) is feasible, because f itself is in $\mathcal{F}_{L_f}(\mathcal{I}_j)$. Then, according to Theorem 4 and Corollary 1 in [22], for any $\hat{f}\in\mathcal{F}_{L_f}(\mathcal{I}_j)$ and $y\in\mathbb{R}^n$, $(\hat{f}(y),y,\nabla\hat{f}(y))$ is feasible for (5). Thus, the optimal objective value of (6) is no less than that of (5). This also shows that (5) is feasible. Conversely, for any feasible solution (t,y,u) of (5), by the same result in [22], there exists an L_f-smooth convex function $\hat{f}$ such that $\hat{f}(y)=t$, $\nabla\hat{f}(y)=u$, and for all $i\in\mathcal{I}_j$, $\hat{f}(y_i)=f(y_i)$ and $\nabla\hat{f}(y_i)=\nabla f(y_i)$, i.e., $\hat{f}\in\mathcal{F}_{L_f}(\mathcal{I}_j)$. Thus, the optimal objective value of (5) is no less than that of (6). This proves the equivalence between (5) and (6). ☐

Following Rockafellar's Convex Analysis book (Theorem 5.6) [19], we define the convex hull of a collection of functions.

Definition 5 (Convex hull of a collection of functions)**.** *The convex hull of an arbitrary collection of functions* $\{h_i : i\in\mathcal{I}\}$ *on* $\mathbb{R}^n$ *is denoted by* $\operatorname{conv}\{h_i : i\in\mathcal{I}\}$*. It is the convex hull of the pointwise infimum of the collection.*

Lemma 6. *Let* f_j *be the closure of the convex hull of all functions in* $\mathcal{F}_{L_f}(\mathcal{I}_j)$*, i.e.,* $f_j=\operatorname{cl}(\operatorname{conv}\{\hat{f}:\hat{f}\in\mathcal{F}_{L_f}(\mathcal{I}_j)\})$*. Then* f_j *is an element of* $\mathcal{F}_{L_f}(\mathcal{I}_j)$*. Thus,* $f_j=\operatorname{conv}\{\hat{f}:\hat{f}\in\mathcal{F}_{L_f}(\mathcal{I}_j)\}$*. Moreover,* f_j *is the lowest element in* $\mathcal{F}_{L_f}(\mathcal{I}_j)$*, i.e.,*

$$\forall\hat{f}\in\mathcal{F}_{L_f}(\mathcal{I}_j),\,\forall x\in\mathbb{R}^n,\, f_j(x)\leq\hat{f}(x).$$

Proof. Let $E : x\mapsto\inf_{\hat{f}\in\mathcal{F}_{L_f}(\mathcal{I}_j)}\hat{f}(x)$. Note that the conjugate $E^*=(\operatorname{cl}(\operatorname{conv}(E)))^*$ [19, Corollary 12.1.1]. Thus, $E^{**}=\operatorname{cl}(\operatorname{conv}(E))=f_j$. In the following, we study the properties of E, E^*, and E^{**}. First, notice that $\forall i\in\mathcal{I}_j, E(y_i)=f(y_i)<+\infty$. Moreover, $\forall\hat{f}\in\mathcal{F}_{L_f}(\mathcal{I}_j), i\in\mathcal{I}_j, y\in\mathbb{R}^n$, $\hat{f}(y)\geq\hat{f}(y_i)+\langle\nabla\hat{f}(y_i),y-y_i\rangle=f(y_i)+\langle\nabla f(y_i),y-y_i\rangle$. Thus, $\forall i\in\mathcal{I}_j, y\in\mathbb{R}^n, E(y)\geq f(y_i)+\langle\nabla f(y_i),y-y_i\rangle>-\infty$. This shows that E is proper. The conjugate E^* of E is given as

$$\begin{aligned}
\forall w \in \mathbb{R}^n, \quad E^*(w) &= \sup_{x\in\mathbb{R}^n} \langle w, x\rangle - E(x) \\
&= \sup_{x\in\mathbb{R}^n} \langle w, x\rangle - \inf_{\hat f \in \mathcal{F}_{L_f}(\mathcal{I}_j)} \hat f(x) \\
&= \sup_{x\in\mathbb{R}^n} \langle w, x\rangle + \sup_{\hat f \in \mathcal{F}_{L_f}(\mathcal{I}_j)} (-\hat f(x)) \\
&= \sup_{\hat f \in \mathcal{F}_{L_f}(\mathcal{I}_j)} \sup_{x\in\mathbb{R}^n} \langle w, x\rangle - \hat f(x) \\
&= \sup_{\hat f \in \mathcal{F}_{L_f}(\mathcal{I}_j)} \hat f^*(w).
\end{aligned}$$

As $\hat f$ is proper, convex, and L_f-smooth, $\hat f^*$ is proper convex and $1/L_f$-strongly convex. The pointwise supremum of convex functions is convex. So $\sup_{\hat f \in \mathcal{F}_{L_f}(\mathcal{I}_j)} \hat f^*(w) - \frac{1}{2L_f}\|w\|^2$ is convex. This shows that E^* is $1/L_f$-strongly convex. Thus, E^{**} is L_f-smooth. Also, E^{**} interpolates f in function value and gradient at $(y_i)_{i\in\mathcal{I}_j}$. This shows that $E^{**} \in \mathcal{F}_{L_f}(\mathcal{I}_j)$ and E^{**} is a lower bound of $\mathcal{F}_{L_f}(\mathcal{I}_j)$.

□

The previous lemma shows that we can unambiguously define f_j as the minimum of $\mathcal{F}_{L_f}(\mathcal{I}_j)$. The following theorem shows that problem (5) iteratively builds lower models of f with desirable properties for establishing the accelerated convergence rate of Algorithm 3.

Theorem 7. *Problem* (5) *defines a lower model function f_j that is L_f-smooth, convex, and satisfies*

$$f(y_{j+1}) \geq f_j(y_{j+1}) \geq f_{j-1}(y_{j+1}),$$

with $y_{j+1} = \arg\min_{y\in\mathbb{R}^n} f_j(y) + \frac{\rho}{2}\|y - x_k\|^2$.

Proof. Following Lemma 4, (5) is equivalent to (6). We know from Lemma 6 that f_j is well defined and is minimal among all functions of $\mathcal{F}_{L_f}(\mathcal{I}_j)$ for all y, hence

$$\min_{y\in\mathbb{R}^n} \frac{\rho}{2}\|y - x_k\|^2 + \min_{\hat f \in \mathcal{F}_{L_f}(\mathcal{I}_j)} \hat f(y) = \min_{y\in\mathbb{R}^n} \frac{\rho}{2}\|y - x_k\|^2 + f_j(y).$$

This shows that $(f_j(y_{j+1}), y_{j+1}, \nabla f_j(y_{j+1}))$ is optimal for (5).

We have constructed f_j such that $f_j \in \mathcal{F}_{L_f}(\mathcal{I}_j)$ and $\forall \hat f \in \mathcal{F}_{L_f}(\mathcal{I}_j)$, $f_j \leq \hat f$. In particular, as we suppose, during a sequence of null steps, that $\mathcal{I}_{j-1} \subset \mathcal{I}_j$, we deduce $f_j \in \mathcal{F}_{L_f}(\mathcal{I}_{j-1})$ ensuring that $f_j \geq f_{j-1}$. In addition, as $f \in \mathcal{F}_{L_f}(\mathcal{I}_j)$, $f \geq f_j$. □

The following proposition and corollary show that the gradients of f_j are in the convex hull of the gradients of f at the previous iterates. This leads to a bound on the gradients of f_j which will be useful for proving that the second constraint (5c) in the nonconvex QCQP (5) is not necessary.

Intuitively, taken any $\hat{f} \in \mathcal{F}_{L_f}(\mathcal{I}_j)$, we want to construct a transformation K on $\hat{f}$ such that $K(\hat{f})$ "clips" the gradients of $\hat{f} \in \mathcal{F}_{L_f}(\mathcal{I}_j)$ to $\operatorname{conv}(\{\nabla f(y_i)\}_{i\in\mathcal{I}_j})$. Figure 1 illustrates this operation. Suppose the bundle $\mathcal{I}_j$ contains two points y_1 and y_2. Their corresponding cutting planes are dotted black lines. A smooth interpolating function $\hat{f} \in \mathcal{F}_{L_f}(\mathcal{I}_j)$ is given in bold dashed blue. K clips its gradients and transforms $\hat{f}$ into the function represented by a solid red line. If we apply K on $K(\hat{f})$, we get $K(\hat{f})$, i.e., $K(K(\hat{f})) = K(\hat{f})$. Thus, K is an idempotent endomorphism on $\mathcal{F}_{L_f}(\mathcal{I}_j)$.

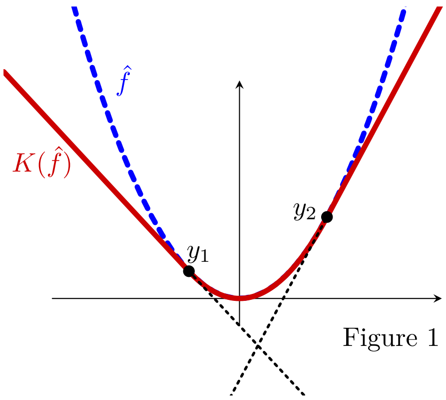

Figure 1

The following proposition makes this precise.

Proposition 8. *We define a mapping K on $\mathcal{F}_{L_f}(\mathcal{I}_j)$ such that*

$$K : \hat{f} \mapsto \left(\hat{f}^* + \delta_{G_j}\right)^*,$$

where δ_{G_j} is the indicator function of $G_j := \operatorname{conv}(\{\nabla f(y_i)\}_{i\in\mathcal{I}_j})$, i.e.,

$$\delta_{G_j}(u) = \begin{cases} 0, & \text{if } u \in G_j, \\ +\infty, & \text{otherwise.} \end{cases}$$

Then, $K(\hat{f}) \in \mathcal{F}_{L_f}(\mathcal{I}_j)$ and

$$\forall \hat{f} \in \mathcal{F}_{L_f}(\mathcal{I}_j), \quad K(\hat{f}) \leq \hat{f}.$$

Proof. Let $\hat{f} \in \mathcal{F}_{L_f}(\mathcal{I}_j)$. As $\hat{f}$ is L_f-smooth, $\hat{f}^*$ is $1/L_f$-strongly convex. So $K(\hat{f})^*$ is $1/L_f$-strongly convex, which implies that $K(\hat{f})$ is L_f-smooth. For $i \in \mathcal{I}_j$, we have

$$\begin{aligned} \partial K(\hat{f})^*(\nabla f(y_i)) &= \partial(\hat{f}^* + \delta_{G_j})(\nabla f(y_i)) \\ &= \partial \hat{f}^*(\nabla f(y_i)) + \partial\delta_{G_j}(\nabla f(y_i)) \\ &\supseteq \partial \hat{f}^*(\nabla f(y_i)) + \{0\} \\ &= \partial \hat{f}^*(\nabla f(y_i)) \ni y_i. \end{aligned}$$

The second equality holds because $\operatorname{relint}(\operatorname{dom}(\hat{f}^*)) \cap \operatorname{relint}(G_j) \neq \varnothing$. The last containment is due to the fact that $x^* \in \partial f(x) \iff x \in \partial f^*(x^*)$ for any closed convex function f (see Theorem 23.5 in [19]). Applying this fact again, we get $y_i \in \partial K(\hat{f})^*(\nabla f(y_i)) \iff \nabla f(y_i) \in \partial K(\hat{f})(y_i)$. Since $K(\hat{f})$ is smooth, $\nabla f(y_i) =$

$\nabla K(\hat{f})(y_i)$ for all $i \in \mathcal{I}_j$. That is, $K(\hat{f})$ interpolates the gradients of f at all the previous iterates. Moreover, we show that $K(\hat{f})$ also interpolates the function values of f at the previous iterates. Indeed, since $K(\hat{f})$ is closed and $\nabla f(y_i) = \nabla K(\hat{f})(y_i)$, applying Theorem 23.5 in [19], we have

$$\begin{aligned} K(\hat{f})(y_i) &= \langle \nabla f(y_i), y_i \rangle - K(\hat{f})^*(\nabla f(y_i)) \\ &= \langle \nabla f(y_i), y_i \rangle - \hat{f}^*(\nabla f(y_i)) \\ &= f(y_i). \end{aligned}$$

The second equality is due to the definition of $K(\hat{f})$ and $\nabla f(y_i) \in G_j$. The last equality again follows from Theorem 23.5 in [19]. Thus, we have shown that $K(\hat{f}) \in \mathcal{F}_{L_f}(\mathcal{I}_j)$. Also, as $K(\hat{f})^* \geq \hat{f}^*$, $K(\hat{f}) \leq \hat{f}$. □

Corollary 9. *The function f_j has a gradient that takes values in the convex hull of the gradient of f at the previous iterates:* $\mathrm{range}(\nabla f_j) \subseteq \mathrm{conv}(\nabla f(y_i)_{i \in \mathcal{I}_j})$.

Proof. $K(f_j) \in \mathcal{F}_{L_f}(\mathcal{I}_j)$ and $K(f_j) \leq f_j$. By minimality of f_j, $K(f_j) = f_j$. $K(f_j)^* = f_j^*$ further implies that $\mathrm{range}(\nabla f_j) = \mathrm{dom}(f_j^*) \subseteq \mathrm{conv}(\nabla f(y_i))$. □

2.3.2 Dropping the second line of constraints

In this section, we show that the nonconvex constraints (5c) of problem (5) are unnecessary and that (5) is equivalent to the same problem without the second line of constraints as

$$\min_{t \in \mathbb{R}, y \in \mathbb{R}^n, u \in \mathbb{R}^n} \; t + \frac{\rho}{2}\|y - x_k\|^2 \tag{7b}$$

$$\text{s.t.} \quad \forall i \in \mathcal{I}_j, \quad t - f(y_i) - \langle \nabla f(y_i), y - y_i \rangle \geq \frac{1}{2L_f}\|u - \nabla f(y_i)\|^2. \tag{7c}$$

Lemma 10. *In problem (7), any optimal solution (t^*, y^*, u^*) satisfies $u^* = \rho(x_k - y^*)$.*

Proof. We write the Lagrangian of this problem:

$$\begin{aligned} \mathcal{L}(t, y, u, \lambda) = t &+ \frac{\rho}{2}\|y - x_k\|^2 \\ &+ \sum_{i \in \mathcal{I}_j} \lambda_i \left(-t + f(y_i) + \langle \nabla f(y_i), y - y_i \rangle + \frac{1}{2L_f}\|u - \nabla f(y_i)\|^2 \right). \end{aligned}$$

By the KKT conditions, for any optimal (t^*, y^*, u^*), there exist λ_i's such that

$$\begin{cases} 0 = 1 - \sum_{i \in \mathcal{I}_j} \lambda_i, \\ 0 = \sum_{i \in \mathcal{I}_j} \lambda_i \frac{1}{L}(u^* - \nabla f(y_i)), \\ 0 = \rho(y^* - x_k) + \sum_{i \in \mathcal{I}_j} \lambda_i \nabla f(y_i). \end{cases} \tag{8}$$

We conclude that $u^* = \rho(x_k - y^*)$. □

Substituting u with its value at optimality given by Lemma 10 in (7) gives the bundle minimization step (3) in Algorithm 3. Compared to the problem (2) solved in the usual PBM, it shares the same objective function, number of variables and number of constraints. However, the constraints are convex quadratic rather than linear.

We will now prove the equivalence between (3) and (5). We consider problem (7) before optimizing over y,

$$\min_{t\in\mathbb{R},u\in\mathbb{R}^n} \; t + \frac{\rho}{2}\|y - x_k\|^2 \tag{9a}$$

$$\text{s.t.} \quad \forall i \in \mathcal{I}_j \quad t \geq f(y_i) + \langle \nabla f(y_i), y - y_i\rangle + \frac{1}{2L_f}\|u - \nabla f(y_i)\|^2. \tag{9b}$$

The quadratic term in the objective only depends on y. We define $\Delta(y)$ as the set of optimal multipliers and $S(y)$ as the set of optimal solutions (t, u), which is a singleton as the problem is strongly convex in u. In problem (9b), the optimal value of t is given by

$$t = \min_{u\in\mathbb{R}^n} \max_{i\in\mathcal{I}_j} f(y_i) + \langle \nabla f(y_i), y - y_i\rangle + \frac{1}{2L_f}\|u - \nabla f(y_i)\|^2. \tag{10}$$

This defines a function $\tilde{f}(y)$. This also shows that $\tilde{f}(y) \geq l(y) := \max_{i\in\mathcal{I}_j} f(y_i) + \langle\nabla f(y_i), y - y_i\rangle$, where $f(y_i) + \langle \nabla f(y_i), y - y_i\rangle$ is the usual cutting plane.

Lemma 11. *$\tilde{f}$ is differentiable with a gradient being the optimal value of u in (9).*

Proof. $\tilde{f}(y)$ is defined as the result of a parametric convex problem in y. The set of optimal solutions is nonempty. The directional regularity condition (Definition 4.8 of [23] and Theorem 4.9) is satisfied. Indeed, the derivative with respect to t of the constraint (9b) is -1 and the ambient set for t is $\mathbb{R}$. Finally, the optimal solution $(t^*(y), u^*(y))$ of (10) is continuous in y as the problem is strongly convex in u. According to Theorem 4.24 in [23], the Hadamard directional derivative of $\tilde{f}$ is

$$\begin{aligned}
\tilde{f}'(y, d) &= \inf_{(t,u)\in S(y)} \sup_{\lambda\in\Delta(y)} D_y\mathcal{L}(t, y, u, \lambda)d \\
&= \sup_{\lambda\in\Delta(y)} D_y\mathcal{L}(t^*(y), y, u^*(y), \lambda)d \\
&= \sup_{\lambda\in\Delta(y)} D_y\left[t + \sum_{i\in\mathcal{I}_j}\lambda_i\left(-t + f(y_i) + \langle\nabla f(y_i), y - y_i\rangle + \frac{1}{2L_f}\|u - \nabla f(y_i)\|^2\right)\right] d \\
&= \sup_{\lambda\in\Delta(y)} \left(\sum_{i\in\mathcal{I}_j}\lambda_i\nabla f(y_i)\right) d \\
&= \langle u, d\rangle.
\end{aligned}$$

The last line follows from (8). This shows that $\tilde{f}$ is differentiable and that its gradient is the optimal value of u in (9). $\square$

Recall that by Lemma 6, $f_j(y) = \min_{\hat{f}\in\mathcal{F}_{L_f}(\mathcal{I}_j)} \hat{f}(y)$ is the lowest smooth interpolating function in $\mathcal{F}_{L_f}(\mathcal{I}_j)$. Consider the function $H_j := f_j - \tilde{f}$. We know that

- H_j is differentiable as Lemma 11 shows that $\tilde{f}$ is differentiable and $f_j \in \mathcal{F}_{L_f}(\mathcal{I}_j)$ is smooth.
- $H_j \geq 0$ as (7) removes a constraint compared to (5).

The following technical result (Proposition 12) provides an elementary real analysis result. The proof can be found in Appendix A.1.

Proposition 12. *Let $h : \mathbb{R}^n \to \mathbb{R}$ be an upper-bounded differentiable function and $\bar{h} = \sup_{x\in\mathbb{R}^n} h(x)$. For all $\delta > 0$, there exists x such that*

$$\begin{cases} \bar{h} - h(x) \leq \delta, \\ \|\nabla h(x)\| \leq \delta. \end{cases}$$

Theorem 13. *The nonconvex QCQP (5) is equivalent to the convex QCQP (3).*

Proof. We use Corollary 9 to show that H_j is bounded.

$$\begin{aligned} \forall y \in \mathbb{R}^n,\ f_j(y) &= \max_{i\in\mathcal{I}_j} f(y_i) + \langle \nabla f(y_i), y - y_i\rangle + \frac{1}{2L_f}\|\nabla f_j(y) - \nabla f(y_i)\|^2 \\ &\leq l(y) + \frac{1}{2L_f}\operatorname{diam}(\{\nabla f(y_i)\}_{i\in\mathcal{I}_j})^2. \end{aligned}$$

It holds that $H_j = f_j - \tilde{f} \leq f_j - l \leq \frac{1}{2L_f}\operatorname{diam}(\{\nabla f(y_i)\}_{i\in\mathcal{I}_j})^2$. This shows that H_j is bounded from above. Let $\bar{H}_j$ denote its supremum over $\mathbb{R}^n$.

Following Proposition 12, we know that for all $\eta > 0$, there exists $\tilde{y} \in \mathbb{R}^n$ such that $\bar{H}_j - H_j(\tilde{y}) \leq \eta/2$ and with $w = \nabla H_j(\tilde{y})$, satisfying

$$\|w\| \leq \frac{\eta L_f}{2\operatorname{diam}(\{\nabla f(y_i)\}_{i\in\mathcal{I}_j})}.$$

Let $\tilde{u} = \nabla f_j(\tilde{y}) = \nabla\tilde{f}(\tilde{y}) + w$ and $i_0 \in \mathcal{I}_j$ such that

$$f_j(\tilde{y}) = f(y_{i_0}) + \langle \nabla f(y_{i_0}), \tilde{y} - y_{i_0}\rangle + \frac{1}{2L_f}\|\tilde{u} - \nabla f(y_{i_0})\|^2.$$

This translates for $\tilde{f}$ into

$$\begin{aligned} \tilde{f}(\tilde{y}) &= \max_{i\in\mathcal{I}_j} f(y_i) + \langle \nabla f(y_i), \tilde{y} - y_i\rangle + \frac{1}{2L_f}\|\nabla\tilde{f}(\tilde{y}) - \nabla f(y_i)\|^2 \\ &\geq f(y_{i_0}) + \langle \nabla f(y_{i_0}), \tilde{y} - y_{i_0}\rangle + \frac{1}{2L_f}\|\tilde{u} - w - \nabla f(y_{i_0})\|^2 \end{aligned}$$

$$= f(y_{i_0}) + \langle \nabla f(y_{i_0}), \tilde{y} - y_{i_0} \rangle + \frac{1}{2L_f} \left(\|\tilde{u} - \nabla f(y_{i_0})\|^2 + \|w\|^2 - 2\langle w, \tilde{u} - \nabla f(y_{i_0}) \rangle \right)$$
$$\geq f_j(\tilde{y}) + \frac{1}{2L_f} \left(-2\langle w, \tilde{u} - \nabla f(y_{i_0}) \rangle \right)$$
$$\geq f_j(\tilde{y}) - \frac{1}{L_f} \|w\| \cdot \|\tilde{u} - \nabla f(y_{i_0})\|$$
$$\geq f_j(\tilde{y}) - \frac{\eta}{2}.$$

Thus, $\bar{H}_j \leq H_j(\tilde{y}) + \eta/2 = f_j(\tilde{y}) - \tilde{f}(\tilde{y}) + \eta/2 \leq \eta/2 + \eta/2 = \eta$. As this is true for all $\eta > 0$, $\sup_{y \in \mathbb{R}^n} H_j(y) = 0$. But we also know that $H_j \geq 0$. This shows that $H_j = 0$ and proves that problems (7) and (5) are equivalent. Lemma 10 shows the equivalence between (7) and (3).

□

2.4 Translation of the inexactness criterion to our setup

Now, we translate the inexactness criterion originally proposed for the accelerated inexact proximal point algorithm (a-IPPA, Algorithm 2) in [18] to the setting of the accelerated proximal bundle method (a-PBM, Algorithm 3). The idea is that the consecutive null steps (3) after a serious step k in a-PBM can be viewed as inexactly solving the proximal point problem $\operatorname{prox}_{\rho,f}(x_k)$ in a-IPPA until the inexactness criterion is satisfied, then the next serious step $k+1$ starts. We propose the following inexactness criterion for the proximal step

$$\left\langle \nabla f(y_{k+1}) - \nabla f(x_k - \frac{1}{\rho} \nabla f(y_{k+1})), y_{k+1} - (x_k - \frac{1}{\rho} \nabla f(y_{k+1})) \right\rangle \leq \frac{1}{2\rho} \|\nabla f(y_{k+1})\|^2. \tag{11}$$

The following theorem bounds the number of steps in a-IPPA when the inexactness criterion (11) is satisfied by the inexact proximal step. This also gives the number of serious steps in a-PBM, when the serious steps satisfy the inexactness criterion (11).

Theorem 14 (Theorem 5.1 [18])**.** *Suppose that we apply the accelerated proximal point algorithm (a-IPPA, Algorithm 2) to problem (1) with the inexact oracle for solving the proximal step problem satisfying condition (11), then*

$$f(\zeta_k) - f(x^*) \leq \frac{2\rho \|x_0 - x^*\|^2}{(k+1)^2}, \quad \forall k \geq 1.$$

An ϵ-solution can be obtained after at most the following number of iterations

$$\left\lceil \frac{\sqrt{2\rho}\|x_0 - x^*\|}{\sqrt{\epsilon}} \right\rceil.$$

Remark 15. *It is worth noting that Güler in [16] has first introduced an accelerated inexact proximal point algorithm. However, Salzo and villa in [17] have shown that the* *proof is flawed and the result does not hold with the approximation criterion chosen by Güler. In Güler's setup, the correct convergence rate is $\mathcal{O}(\frac{1}{\epsilon})$. Even though [18] builds upon Guler's algorithm, and was published before [17], the convergence proof technique is quite different and does not suffer from the same flaw.*

In order for a serious step in a-PBM to satisfy the inexactness criterion, we need to design a *null step test* that implies the inexactness criterion. This is done by the following theorem.

Theorem 16 (A sufficient condition for satisfying the inexact criterion)**.** *Let k be a serious step in the execution of Algorithm 3 and let x_k be the corresponding proximal center. Let $j > k$ be an index in the sequence of null steps following the serious step k. If*

$$C\|y_{j+1} - y_j\| \leq \|x_k - y_{j+1}\| \qquad \textit{with} \qquad C = \frac{2L_f}{\rho}\left(\frac{\sqrt{2L_f}}{\sqrt{\rho}} + 1\right), \tag{12}$$

then the inexactness criterion (11) is satisfied.

Proof. We bound the left-hand side of (11):

$$\begin{aligned}
&\left\langle \nabla f(y_{j+1}) - \nabla f(x_k - \frac{1}{\rho}\nabla f(y_{j+1})), y_{j+1} - (x_k - \frac{1}{\rho}\nabla f(y_{j+1}))\right\rangle \\
&\leq \|\nabla f(y_{j+1}) - \nabla f(x_k - \frac{1}{\rho}\nabla f(y_{j+1}))\| \cdot \|y_{j+1} - (x_k - \frac{1}{\rho}\nabla f(y_{j+1}))\| \qquad (13)\\
&\leq L_f\|y_{j+1} - (x_k - \frac{1}{\rho}\nabla f(y_{j+1}))\| \cdot \|y_{j+1} - (x_k - \frac{1}{\rho}\nabla f(y_{j+1}))\| \qquad (14)\\
&= \frac{L_f}{\rho^2}\|\rho(y_{j+1} - x_k) + \nabla f(y_{j+1})\|^2.
\end{aligned}$$

Eq. (13) uses the Cauchy-Schwartz and Eq. (14) uses the L_f-smoothness of f.

Note that because of the optimality of y_{j+1} for the minimization step (3),

$$\nabla f_j(y_{j+1}) = -\rho(y_{j+1} - x_k).$$

As f_j is L_f-smooth and interpolates f at y_j, for all $x \in \mathbb{R}^n$, we have

$$\|\nabla f(y_j) - \nabla f_j(x)\| = \|\nabla f_j(y_j) - \nabla f_j(x)\| \leq L_f\|x - y_j\|.$$

Thus,

$$\begin{aligned}
\|\rho(y_{j+1} - x_k) + \nabla f(y_{j+1})\| &= \| - \nabla f_j(y_{j+1}) + \nabla f(y_{j+1})\| \\
&\leq \| - \nabla f_j(y_{j+1}) + \nabla f(y_j)\| + \|\nabla f(y_{j+1}) - \nabla f(y_j)\| \\
&\leq 2L_f\|y_{j+1} - y_j\|.
\end{aligned}$$

We now consider the right-hand side of (11),

$$\frac{1}{\sqrt{2\rho}}\|\nabla f(y_{j+1})\| \geq \frac{1}{\sqrt{2\rho}}(-\|\nabla f_j(y_{j+1}) - \nabla f(y_{j+1})\| + \rho\|x_k - y_{j+1}\|)$$
$$\geq \frac{1}{\sqrt{2\rho}}\left(-2L_f\|y_{j+1} - y_j\| + \rho\|x_k - y_{j+1}\|\right).$$

The following condition and $\rho\|x_k - y_{j+1}\| - 2L_f\|y_{j+1} - y_j\| \geq 0$ imply (11):

$$\left(\frac{2\sqrt{L_f}}{\rho} + \frac{\sqrt{2}L_f}{\sqrt{\rho}}\right)\|y_{j+1} - y_j\| \leq \frac{\sqrt{\rho}}{\sqrt{2}}\|x_k - y_{j+1}\|,$$
$$\frac{2L_f}{\rho}\left(\frac{\sqrt{2L_f}}{\sqrt{\rho}} + 1\right)\|y_{j+1} - y_j\| \leq \|x_k - y_{j+1}\|. \tag{15}$$

Eq. (15) is more stringent than $\rho\|x_k - y_{j+1}\| - 2L_f\|y_{j+1} - y_j\| \leq 0$, showing that the criterion (15) alone suffices to imply (11). □

Remark 17. *In [1] and [13] the authors considered a modified version of the PBM that has the drawback that the null step test (hence the algorithm logic) depends on the value of ϵ. Algorithm 3 (as the classic PBM [2]) does not require the knowledge of the desired precision ϵ. This makes the algorithm more versatile and reliable in diverse settings where the optimal precision might not be predetermined.*

2.5 Bound on the number of null steps

In [1], when the objective function is smooth, the number of consecutive null steps is shown to be $\mathcal{O}(\log(\frac{1}{\epsilon}))$. However, our setup is different in two ways:

1. The update rule for the proximal center is given by the Nesterov acceleration scheme, so the bound involving distance to proximal centers is not valid anymore.
2. Our *null step test* (12) may be harder to satisfy than the one in [1].

The first point will be discussed in the next subsection. In Subsection 2.5.2 we will recall the main inequalities from [1] and introduce a new quantity ξ_j which converges towards zero with a geometric rate. Finding an upper bound on the first term of this sequence, right after a serious step, and a lower bound on this sequence that is a power of ϵ suffices to show that there are $\mathcal{O}(\log(\frac{1}{\epsilon}))$ consecutive null steps.

2.5.1 A Technical result: bound on the distance between the proximal center and the optimal solutions

We first give a technical lemma (Lemma 18) providing a bound derived from a recurrence inequality. The proof can be found in Appendix A.2.

Lemma 18. *If a sequence satisfies the following recurrence inequality*

$$\forall k \geq 0, \quad r_{k+1} \leq (1 + \frac{2}{k+2})r_k + \frac{2C'}{k+2},$$

where $C' > 0$, then it holds that

$$\forall k \geq 1, \quad r_k \leq \left(e^2 \cdot r_0 + \frac{C' \pi^2 e^{(3+\frac{\pi^2}{3})}}{3} \right) \cdot k^2.$$

In the execution of Algorithm 3, denote j_k the overall iteration count for the k-th serious step. Then, we define $x^*_{j_k} := \arg\min_{y \in \mathbb{R}^n} f(y) + \frac{\rho}{2}\|y - x_{j_k}\|^2$. The next lemma bounds two key quantities for the analysis of the number of consecutive null steps.

Lemma 19. *There exists a constant $C'' > 0$, depending only on ρ, x^*, x_1, $f(x^*)$, and $f(x_1)$ such that for all $k > 0$,*

$$\|x_{j_k} - x^*\| \leq \frac{3C''}{\epsilon} \qquad \textit{and} \qquad \|x^* - x^*_{j_k}\| \leq \frac{6C''}{\epsilon}. \tag{16}$$

Proof. Following Lemma 5.1 in [18], considering consecutive serious steps with index j_k for $k = 1, 2, 3, \ldots$,

$$2t_{j_k}^2 v_{j_k} - 2t_{j_{k+1}}^2 v_{j_{k+1}} \geq \rho \|u_{j_{k+1}}\|^2 - \rho \|u_{j_k}\|^2, \quad \forall k \geq 0,$$

where $v_{j_k} := f(\zeta_{j_{k+1}}) - f(x^*)$ and $u_{j_k} := t_{j_k}\zeta_{j_{k+1}} - (t_{j_k} - 1)\zeta_{j_k} - x^*$. By adding these inequalities, we have

$$\rho \|u_{j_k}\|^2 \leq \rho \|u_0\|^2 + 2t_0^2 v_0 \leq \rho \|x_1 - x^*\|^2 + 2(f(x_1) - f(x^*)).$$

We define $C' = \sqrt{\|x_1 - x^*\|^2 + \frac{2}{\rho}(f(x_1) - f(x^*))}$, which is an upper bound on $\|u_{j_k}\|$ for all k. Using the definition of u_{j_k}, we get

$$\|u_{j_k}\| = \|t_{j_k}(\zeta_{j_{k+1}} - \zeta_{j_k}) + (\zeta_{j_k} - x^*)\| \geq t_{j_k}\|(\zeta_{j_{k+1}} - \zeta_{j_k})\| - \|\zeta_{j_k} - x^*\|.$$

This provides an upper bound on $\|\zeta_{j_{k+1}} - \zeta_{j_k}\|$ as

$$\|\zeta_{j_{k+1}} - \zeta_{j_k}\| \leq \frac{1}{t_{j_k}}(\|\zeta_{j_k} - x^*\| + C').$$

From this inequality, we get

$$\begin{aligned} \|\zeta_{j_{k+1}} - x^*\| \leq \|\zeta_{j_k} - x^*\| + \|\zeta_{j_{k+1}} - \zeta_{j_k}\| &\leq \|\zeta_{j_k} - x^*\| + \frac{1}{t_{j_k}}(\|\zeta_{j_k} - x^*\| + C') \\ &= (1 + \frac{1}{t_{j_k}})\|\zeta_{j_k} - x^*\| + \frac{C'}{t_{j_k}}. \end{aligned}$$

We also have the following classic result about the Nesterov acceleration scheme, namely, $t_{j_k} \geq \frac{k+2}{2}$, with k counting serious steps only (see, for instance, Lemma 5.3 in [18] or [24]). This is a non-homogeneous linear recurrence for which we can derive a bound $\|\zeta_{j_{k+1}} - x^*\| \leq C'' \cdot \frac{1}{\epsilon}$ as follows.

Using Theorem 14, an ϵ-optimal solution is reached after at most $\left\lceil \frac{\sqrt{2\rho}\|x_0-x^*\|}{\sqrt{\epsilon}} \right\rceil$ serious iterations. We apply the result of Lemma 18 with $r_0 = \|\zeta_0 - x^*\| = \|x_0 - x^*\|$ to conclude that at the k^{th} serious step,

$$
\begin{aligned}
\|\zeta_{j_k} - x^*\| &\leq \left(e^2\|x_0 - x^*\| + \frac{C'\pi^2 e^{(3+\frac{\pi^2}{3})}}{3} \right) \cdot k^2 \\
&\leq \left(e^2\|x_0 - x^*\| + \frac{C'\pi^2 e^{(3+\frac{\pi^2}{3})}}{3} \right) \cdot \left(1 + \frac{\sqrt{2\rho}\|x_0 - x^*\|}{\sqrt{\epsilon}}\right)^2 \\
&\leq \left(e^2\|x_0 - x^*\| + \frac{C'\pi^2 e^{(3+\frac{\pi^2}{3})}}{3} \right) \cdot \left(2 + \frac{4\rho\|x_0 - x^*\|^2}{\epsilon}\right) = \mathcal{O}\left(\frac{1}{\epsilon}\right).
\end{aligned}
$$

Let C'' be such that $\forall k \in \mathbb{N}^*$, $\|\zeta_{j_k} - x^*\| \leq C''/\epsilon$. If we suppose that $\epsilon \leq 1$, we can take

$$
C'' = \left(2 + 4\rho\|x_0 - x^*\|^2\right) \cdot \left[e^2\|x_0 - x^*\| + (\rho\|x_1 - x^*\|^2 + 2(f(x_1) - f(x^*))) \frac{\pi^2 e^{(3+\frac{\pi^2}{3})}}{3} \right].
$$

We will conclude the proof by giving a bound on $\|x_{j_{k+1}} - x^*\|$. The first line uses the update of x given in Algorithm 3.

$$
\begin{aligned}
\|x_{j_{k+1}} - x^*\| &= \left\| \zeta_{j_{k+1}} - x^* + \frac{t_{j_k} - 1}{t_{j_{k+1}}} \left((\zeta_{j_{k+1}} - x^*) - (\zeta_k - x^*) \right) \right\| \\
&\leq \|\zeta_{j_{k+1}} - x^*\| + \frac{t_{j_k} - 1}{t_{j_{k+1}}} (\|\zeta_{j_{k+1}} - x^*\| + \|\zeta_k - x^*\|) \\
&\leq (1 + 2\frac{t_{j_k} - 1}{t_{j_{k+1}}}) \frac{C''}{\epsilon} \leq (1 + \frac{4t_{j_k}}{1 + \sqrt{1 + 4t_{j_k}^2}}) \frac{C''}{\epsilon} \\
&\leq (1 + \frac{4t_{j_k}}{2t_{j_k}\sqrt{1 + 1/(4t_{j_k}^2)}}) \frac{C''}{\epsilon} \qquad (17) \\
&\leq (1 + \frac{2}{\sqrt{1}}) \frac{C''}{\epsilon} \leq 3\frac{C''}{\epsilon}.
\end{aligned}
$$

By optimality of x^* and $x^*_{j_k}$ for their respective problems,

$$
f(x^*) + \frac{\rho}{2}\|x^*_{j_k} - x_{j_k}\|^2 \leq f(x^*_{j_k}) + \frac{\rho}{2}\|x^*_{j_k} - x_{j_k}\|^2 \leq f(x^*) + \frac{\rho}{2}\|x^* - x_{j_k}\|^2.
$$

Thus, $\|x^*_{j_k} - x_{j_k}\| \leq \|x^* - x_{j_k}\|$. This leads to

$$
\|x^* - x^*_{j_k}\| \leq \|x^* - x_{j_k}\| + \|x^*_{j_k} - x_{j_k}\| \leq 2\|x^* - x_{j_k}\| \leq 6\frac{C''}{\epsilon}.
$$

□

2.5.2 Main result on the number of consecutive null steps

We consider a serious step k followed by null steps j with $j > k$.

We define

$$m_j = f_{j-1}(y_j) + \frac{\rho}{2}\|y_j - x_k\|^2.$$

This is exactly the value of the bundle problem (3) solved at iteration j. Let

$$\begin{cases} z_j = \underset{z\in\{y_{k+1},y_{k+2},\dots,y_j\}}{\arg\min} \{f(z) + \frac{\rho}{2}\|z - x_k\|^2\} \\ \xi_j = f(z_j) + \frac{\rho}{2}\|z_j - x_k\|^2 - m_j, \end{cases} \tag{18}$$

where ξ_j represents the discrepancy between the best value of the *null step* problem $\min_y\{f(y) + \frac{\rho}{2}\|y - x_k\|^2\}$ observed thus far and the most recent value of this problem when the true function f is substituted with its model f_{j-1}.

Lemma 20. *Let $\tau \in [0, 1)$, a constant such that*

$$\frac{\tau}{1-\tau} \geq \frac{L_f}{\rho} + C^2. \tag{19}$$

Then, the following linear convergence rate on the quantity ξ_j holds

$$\forall j > k, \qquad \xi_{j+1} \leq \tau\xi_j. \tag{20}$$

Remark 21. *Contrary to the case where f is only Lipschitz continuous in [1], τ does not depend on ϵ and enables to derive a linear convergence of ξ_j, hence an $\mathcal{O}(\log(\frac{1}{\epsilon}))$ number of consecutive null steps.*

Proof of Lemma 20. We define

$$l_f(u, v) = f(v) + \langle\nabla f(v), u - v\rangle,$$

which corresponds to the supporting hyperplane of f at v.

As the bundle set, at step $j + 1$, contains the cut made at step j and that we suppose that $\mathcal{I}_j \subset \mathcal{I}_{j+1}$,

$$m_{j+1} \geq l_f(y_{j+1}, y_j), \tag{21}$$

$$m_{j+1} \geq f_{j-1}(y_{j+1}) + \frac{\rho}{2}\|y_{j+1} - x_k\|^2 \geq m_j + \frac{\rho}{2}\|y_j - y_{j+1}\|^2. \tag{22}$$

For (21), we used the convexity of f_j and the fact that $f_j(y_j) = f(y_j)$ and $\nabla f_j(y_j) = \nabla f(y_j)$. The second inequality (22) uses the ρ-strong convexity of the bundle subproblem, the optimality of y_j for the problem $\min_{y\in\mathbb{R}^n} \{f_{j-1}(y) + \rho/2\|y - x_k\|^2\}$ and $f_j \geq f_{j-1}$.

Putting these two inequalities together, we get

$$\begin{aligned} m_{j+1} &\geq (1-\tau)\left(l_f(y_{j+1}, y_j)\right) + \tau(m_j + \frac{\rho}{2}\|y_j - y_{j+1}\|^2) \\ &= \tau m_j + (1-\tau)\left(l_f(y_{j+1}, y_j) + \frac{\tau\rho}{2(1-\tau)}\|y_j - y_{j+1}\|^2\right). \end{aligned}$$

Using the smoothness of f, $l_f(y_{j+1}, y_j) \geq f(y_{j+1}) - \frac{L_f}{2}\|y_j - y_{j+1}\|^2$.

$$m_{j+1} - \tau m_j \geq (1-\tau)f(y_{j+1}) + (1-\tau)\left(\frac{\tau\rho}{2(1-\tau)} - \frac{L_f}{2}\right)\|y_j - y_{j+1}\|^2. \tag{23}$$

Using the definitions of z_j and ξ_j,

$$\begin{aligned} \xi_{j+1} &= f(z_{j+1}) + \frac{\rho}{2}\|z_{j+1} - x_k\|^2 - m_{j+1} && (24) \\ &\leq \tau(f(z_{j+1}) + \frac{\rho}{2}\|z_{j+1} - x_k\|^2) + (1-\tau)(f(y_{j+1}) + \frac{\rho}{2}\|y_{j+1} - x_k\|^2) - m_{j+1} && (25) \\ &\leq \tau(f(z_{j+1}) + \frac{\rho}{2}\|z_{j+1} - x_k\|^2) + (1-\tau)(f(y_{j+1}) + \frac{\rho}{2}\|y_{j+1} - x_k\|^2) \\ &\quad - \tau m_j - (1-\tau)f(y_{j+1}) - (1-\tau)\left(\frac{\tau\rho}{2(1-\tau)} - \frac{L_f}{2}\right)\|y_j - y_{j+1}\|^2 && (26) \\ &\leq \tau(f(z_{j+1}) + \frac{\rho}{2}\|z_{j+1} - x_k\|^2 - m_j) + \frac{\rho(1-\tau)}{2}\|y_{j+1} - x_k\|^2 \\ &\quad - (1-\tau)\left(\frac{\tau\rho}{2(1-\tau)} - \frac{L_f}{2}\right)\|y_j - y_{j+1}\|^2 && (27) \\ &\leq \tau\xi_j + \frac{1-\tau}{2}\left[\rho C^2 - \left(\frac{\tau\rho}{(1-\tau)} - L_f\right)\right]\|y_j - y_{j+1}\|^2 && (28) \\ &\leq \tau\xi_j + \frac{1-\tau}{2}\underbrace{\left[\rho C^2 + L_f - \frac{\tau\rho}{(1-\tau)}\right]}_{\leq 0}\|y_j - y_{j+1}\|^2 && (29) \\ &\leq \tau\xi_j. && (30) \end{aligned}$$

In (24), we rewrite the definition of ξ. The inequality (25) involves splitting the terms to make the quantities τ and $(1-\tau)$ appear. We also use the optimality of z_{j+1} that ensures that $f(y_{j+1}) + \frac{\rho}{2}\|y_{j+1} - x_k\|^2 \geq f(z_{j+1}) + \frac{\rho}{2}\|z_{j+1} - x_k\|^2$. In (26), we directly apply inequality (23). In (27), the terms $f(y_{j+1})$ cancel. Eq. (28) is a key step in this proof as it leverages the *null step test*. In (29), we use the definition of z_{j+1}. Finally, we conclude with (30), where τ is chosen to be sufficiently close to 1 in accordance with (19), ensuring that the resulting quantity is less than or equal to 0.

This shows the linear convergence of the quantity ξ_j.

□

After proving the geometric decrease of ξ, the following lemma gives a lower bound on this quantity. Supposing no serious step occurs, we know the *null step test* is not satisfied. If ξ_j is small enough, then the iterates y_j will be close to the proximal center x_k. The proximal problem becomes approximately the initial problem (1) and it is possible to show that an ϵ-optimal solution to (1) has been found. The lower bound on ξ, which is polynomial in ϵ (specifically, ϵ^3), is sufficient for our proof. This quantity will appear as an argument of the logarithm function in the convergence rate of the algorithm.

Lemma 22. *There exists a constant $A > 0$ such that, if at iteration j, $\xi_j \leq A\epsilon^3$ and the null step criterion is not satisfied, then $f(y_j) - f(x^*) \leq \epsilon$.*

Proof. Let $\delta > 0$. We consider an iteration j such that $\xi_j \leq \delta$. Suppose that the *null step* criterion is not satisfied, i.e., $C^2\|y_{j+1} - y_j\|^2 \geq \|x_k - y_{j+1}\|^2$. Also, using (22),

$$m_{j+1} \geq m_j + \frac{\rho}{2}\|y_{j+1} - y_j\|^2. \tag{31}$$

We combine these inequalities to get

$$m_{j+1} - m_j \geq \frac{\rho}{2C^2}\|x_k - y_{j+1}\|^2.$$

This provides an upper-bound on $\|x_k - y_{j+1}\|^2$,

$$\begin{aligned}
\|x_k - y_{j+1}\|^2 &\leq \frac{2C^2}{\rho}(m_{j+1} - m_j) && (32)\\
&= \frac{2C^2}{\rho}\left(f_j(y_{j+1}) + \frac{\rho}{2}\|y_{j+1} - x_k\|^2 - m_j\right) && (33)\\
&\leq \frac{2C^2}{\rho}\left(f_j(z_j) + \frac{\rho}{2}\|z_j - x_k\|^2 - m_j\right) && (34)\\
&\leq \frac{2C^2}{\rho}\left(f(z_j) + \frac{\rho}{2}\|z_j - x_k\|^2 - m_j\right) = \frac{2C^2}{\rho}\xi_j && (35)\\
&\leq \frac{2C^2}{\rho}\delta. && (36)
\end{aligned}$$

Eq. (33) is given by definition of m_{j+1}. Eq. (34) uses the optimality of y_{j+1} for (3). Eq. (35) uses $f_j \leq f$ and the definition of ξ_j. Eq. (36) concludes using the upper bound on ξ_j supposed in this lemma.

We use this to bound $\|y_j - x_k\|^2$ as

$$\begin{aligned}
\|y_j - x_k\|^2 &\leq 2\|x_k - y_{j+1}\|^2 + 2\|y_j - y_{j+1}\|^2 && (37)\\
&\leq 2\frac{2C^2}{\rho}\delta + 2\frac{2}{\rho}(m_{j+1} - m_j) && (38)\\
&\leq \frac{4C^2}{\rho}\delta + 4\delta/\rho && (39)
\end{aligned}$$

$$= 4\rho^{-1}\left(C^2+1\right)\delta. \tag{40}$$

To derive (38), we use (31) again. For (39), $\forall z \in \mathbb{R}^n$, $m_{j+1} \le f(z) + \frac{\rho}{2}\|z - x_k\|^2$. In particular, $m_{j+1} - m_j \le f(z_j) + \frac{\rho}{2}\|z_j - x_k\|^2 - m_j = \xi_j \le \delta$.

We want to relate the best value for the *null step* problem seen so far to the current value of this problem, namely $f(z_j) + \frac{\rho}{2}\|z_j - x_k\|^2$ and $f(y_j) + \frac{\rho}{2}\|y_j - x_k\|^2$.

$$f(z_j) + \frac{\rho}{2}\|z_j - x_k\|^2 \ge f_j(z_j) + \frac{\rho}{2}\|z_j - x_k\|^2 \tag{41}$$
$$\ge f_j(y_{j+1}) + \frac{\rho}{2}\|y_{j+1} - x_k\|^2 \tag{42}$$
$$\ge f_j(y_j) + \langle \nabla f_j(y_j), y_{j+1} - y_j\rangle + \frac{\rho}{2}\|y_{j+1} - x_k\|^2 \tag{43}$$
$$\ge f(y_j) - \|\nabla f(y_j)\| \cdot \|y_{j+1} - y_j\| \tag{44}$$
$$= f(y_j) - \|\nabla f(x^*) - \nabla f(y_j)\| \cdot \|y_{j+1} - y_j\| \tag{45}$$
$$\ge f(y_j) - L_f\|x^* - y_j\| \cdot \|y_{j+1} - y_j\| \tag{46}$$
$$\ge f(y_j) - L_f(\|x^* - x_k\| + \|y_j - x_k\|)\|y_{j+1} - y_j\| \tag{47}$$
$$\ge f(y_j) - L_f\left[3C''/\epsilon + 4\rho^{-1}\left(C^2+1\right)\delta\right] \cdot 2\delta/\rho + \left(\frac{\rho}{2}\|y_j - x_k\|^2 - 4\rho^{-1}\left(C^2+1\right)\delta\right) \tag{48}$$
$$= f(y_j) + \frac{\rho}{2}\|y_j - x_k\|^2 - 2\delta L_f/\rho\left[3C''/\epsilon + 4\rho^{-1}\left(C^2+1\right)\delta + 2\left(C^2+1\right)/L_f\right] \tag{49}$$
$$= f(y_j) + \frac{\rho}{2}\|y_j - x_k\|^2 - D\delta/\epsilon. \tag{50}$$

Eq. (41) follows from the fact that $f_j \le f$. Eq. (42) is derived from the optimality of y_{j+1} in the bundle minimization problem described by (3). Eq. (43) utilizes the convexity of f_j. (44) applies the Cauchy-Schwarz inequality and the fact that f_j and f have the same value and gradient at y_j. Eq. (45) relies on the first-order optimality condition at x^*. Eq. (46) uses the smoothness of f. Eq. (47) employs the triangle inequality. Eq. (48) incorporates bounds from (40), (16), and (31). Finally, (49) involves rearranging terms, and (50) introduces a new constant $D > 0$.

Using this inequality, we derive a bound on $\|x_k^* - y_j\|$ as

$$\begin{aligned}
f(y_j) + \frac{\rho}{2}\|y_j - x_k\|^2 &\le f(z_j) + \frac{\rho}{2}\|z_j - x_k\|^2 + D\delta/\epsilon \\
&= f(z_j) + \frac{\rho}{2}\|z_j - x_k\|^2 - m_j + m_j + D\delta/\epsilon \\
&= D\delta/\epsilon + \xi_j + f_{j-1}(y_j) + \frac{\rho}{2}\|y_j - x_k\|^2 && (51)\\
&\le \delta + D\delta/\epsilon + f_{j-1}(y_j) + \frac{\rho}{2}\|y_j - x_k\|^2 && (52)\\
&\le \delta + D\delta/\epsilon + f_{j-1}(x_k^*) + \frac{\rho}{2}\|x_k^* - x_k\|^2 - \frac{\rho}{2}\|x_k^* - y_j\|^2 && (53)\\
&\le \delta + D\delta/\epsilon + f(x_k^*) + \frac{\rho}{2}\|x_k^* - x_k\|^2 - \frac{\rho}{2}\|x_k^* - y_j\|^2 && (54)
\end{aligned}$$

$$\leq \delta + D\delta/\epsilon + f(y_j) + \frac{\rho}{2}\|y_j - x_k\|^2 - \rho\|x_k^* - y_j\|^2 \tag{55}$$

$$\iff \|x_k^* - y_j\|^2 \leq \delta(1 + D/\epsilon)/\rho. \tag{56}$$

Eq. (51) employs the definition of ξ_j, while (52) utilizes the condition $\xi_j \leq \delta$. (53) leverages both the optimality of y_j in solving the bundle subproblem at iteration j and the strong convexity of the corresponding function, whereas (54) relies on the inequality $f_{j-1} \leq f$. Eq. (55) depends on the optimality of x_k^* for the proximal problem centered at x_k, alongside the strong convexity of the corresponding function. In (56), we rearrange the terms.

Starting from (54) we get

$$\begin{aligned}\delta(1 + D/\epsilon) &\geq f(y_j) + \frac{\rho}{2}\|y_j - x_k\|^2 - (f(x_k^*) + \frac{\rho}{2}\|x_k^* - x_k\|^2 - \frac{\rho}{2}\|y_j - x_k^*\|^2) \\ &\geq f(y_j) + \frac{\rho}{2}\|y_j - x_k\|^2 \\ &\quad - \left(f(x^*) + \frac{\rho}{2}\|x^* - x_k\|^2 - \frac{\rho}{2}\|y_j - x_k^*\|^2 - \frac{\rho}{2}\|x^* - x_k^*\|^2\right). \end{aligned} \tag{57}$$

Eq. (57) uses the optimality of x_k^* for the proximal subproblem and the strong convexity of this subproblem.

Rearanging, using $f_j \leq f$ and writing $x^* - x_k = (x^* - x_k^*) + (x_k^* - y_j) + (y_j - x_k)$,

$$\begin{aligned}&f(y_j) - f(x^*) \\ &\leq \delta(1 + D/\epsilon) - \frac{\rho}{2}\|y_j - x_k\|^2 + \frac{\rho}{2}\|x^* - x_k\|^2 - \frac{\rho}{2}\|y_j - x_k^*\|^2 - \frac{\rho}{2}\|x^* - x_k^*\|^2 \\ &\leq \delta(1+D/\epsilon)+\rho\left(\|y_j-x_k^*\|\cdot\|x^*-x_k^*\|+\|y_j-x_k\|\cdot\|x^*-x_k^*\|+\|y_j-x_k\|\cdot\|y_j-x_k^*\|\right) &(58)\\ &\leq \delta(1+D/\epsilon)+\delta(1+D/\epsilon)\|x^*-x_k^*\|+4\left(C^2+1\right)\delta\|x^*-x_k^*\|+4\delta^2\left(C^2+1\right)(1+D/\epsilon) &(59)\\ &= \delta\left[(1 + D/\epsilon) + \left(4C^2 + 5 + D/\epsilon\right)\|x^* - x_k^*\| + 4\delta\left(C^2 + 1\right)(1 + D/\epsilon)\right]. &(60)\end{aligned}$$

We use the Cauchy-Schwarz inequality to derive (58). For (59), we apply (40) and (56).

Using Lemma 19, we derive

$$\begin{aligned}&f(y_j) - f(x^*) \\ &\leq \delta\left[(1 + D/\epsilon) + \left(4C^2 + 2 + D/\epsilon\right)6C''/\epsilon + 4\delta\left(C^2 + 1\right)(1 + D/\epsilon)\right].\end{aligned} \tag{61}$$

We get $f(y_j) - f(x^*) = \mathcal{O}(\delta\epsilon^{-2})$. This concludes the proof of the lemma. □

Remark 23. *The conclusion of Lemma 22 is used in its contrapositive form in Theorem 3. Namely, if y_j is not yet ϵ-optimal (i.e., $f(y_j) - f(x^*) > \epsilon$) and iteration j is a null step (i.e., the null step test is not satisfied), then ξ_j is bounded from below as $\xi_j \geq A\epsilon^3$.*

The following lemma gives an upper bound on ξ_j for serious steps.

Lemma 24. *There exists a constant $M > 0$ such that for all serious step k,*

$$\xi_{k+1} \leq \frac{M}{\epsilon^2}. \tag{62}$$

Proof. We use the smoothness of f to show the claimed bound

$$\begin{aligned}
\xi_{k+1} &= f(y_{k+1}) + \frac{\rho}{2}\|y_{k+1} - x_k\|^2 - \left(f_k(y_{k+1}) + \frac{\rho}{2}\|y_{k+1} - x_k\|^2\right) \\
&= f(y_{k+1}) - f_k(y_{k+1}) \\
&\leq f(y_{k+1}) - l_f(y_{k+1}, y_k) \\
&\leq \frac{L_f}{2}\|y_{k+1} - y_k\|^2.
\end{aligned} \tag{63}$$

The first equality is by setting $j = k + 1$ in the definition (18). The first inequality holds because f_k and f have the same value and gradient at y_k and f_k is convex. The last inequality is given by the smoothness of f.

We bound the distance between an iterate and the corresponding proximal center. To do so, we first use the optimality of y_{k+1} for (3) and $f_k \leq f$,

$$f_k(y_{k+1}) + \frac{\rho}{2}\|y_{k+1} - x_k\|^2 \leq f_k(x_0) + \frac{\rho}{2}\|x_0 - x_k\|^2 \leq f(x_0) + \frac{\rho}{2}\|x_0 - x_k\|^2.$$

Rearranging,

$$\begin{aligned}
\frac{\rho}{2}\|y_{k+1} - x_k\|^2 &\leq f(x_0) - f_k(y_{k+1}) + \frac{\rho}{2}\|x_0 - x_k\|^2 \\
&\leq f(x_0) - l_f(y_{k+1}, x_0) + \frac{\rho}{2}\|x_0 - x_k\|^2 && (64) \\
&\leq \|\nabla f(x_0)\| \cdot \|x_0 - y_{k+1}\| + \frac{\rho}{2}\|x_0 - x_k\|^2 && (65) \\
&\leq \|\nabla f(x_0)\| \cdot \|y_{k+1} - x_k\| + \|\nabla f(x_0)\|\|x_0 - x_k\| + \frac{\rho}{2}\|x_0 - x_k\|^2. && (66)
\end{aligned}$$

Eq. (64) follows from the fact that $f_k \geq l_f(\cdot, x_0)$. Eq. (65) uses the Cauchy-Schwarz inequality, and (66) employs the triangle inequality. Note that (66) is a quadratic inequality in $\|y_{k+1} - x_k\|$. This gives an upper bound

$$\begin{aligned}
\|y_{k+1} - x_k\| &\leq \frac{\|\nabla f(x_0)\| + \sqrt{\Delta_k}}{\rho} \\
&\leq \frac{\|\nabla f(x_0)\| + \sqrt{\Delta}}{\rho},
\end{aligned} \tag{67}$$

where Δ_k and Δ are defined and related as

$$\Delta_k := \|\nabla f(x_0)\|^2 + 2\rho\|\nabla f(x_0)\|\|x_0 - x_k\| + \rho^2\|x_0 - x_k\|^2$$

$$\leq \|\nabla f(x_0)\|^2 + 2\rho\|\nabla f(x_0)\|(\|x_0 - x^*\| + \|x^* - x_k\|) + \rho^2(\|x_0 - x^*\| + \|x^* - x_k\|)^2$$
$$\leq \Delta := \|\nabla f(x_0)\|^2 + 2\rho\|\nabla f(x_0)\|(\|x_0 - x^*\| + 3C''/\epsilon) + \rho^2(\|x_0 - x^*\| + 3C''/\epsilon)^2.$$

The first inequality uses the triangle inquality and the second inequality applies (16). Now we use this to bound $\|y_{k+1} - y_k\|$ as

$$\begin{aligned}\|y_{k+1} - y_k\| &\leq \|y_{k+1} - x_k\| + \|y_k - x_{k-1}\| + \|x_{k-1} - x^*\| + \|x^* - x_k\| \\ &\leq 2\frac{\|\nabla f(x_0)\| + \sqrt{\Delta}}{\rho} + 2\frac{3C''}{\epsilon}.\end{aligned}$$

Starting from (63),

$$\begin{aligned}\xi_{k+1} &\leq \frac{L_f}{2}\|y_{k+1} - y_k\|^2 \\ &\leq \frac{L_f}{2}2\left(2\frac{\|\nabla f(x_0)\| + \sqrt{\Delta}}{\rho}\right)^2 + 4L_f\frac{9C''^2}{\epsilon^2} \quad (68)\\ &\leq 8L_f\frac{\|\nabla f(x_0)\|^2 + \Delta}{\rho^2} + 4L_f\frac{9C''^2}{\epsilon^2} \quad (69)\\ &\leq \frac{8L_f}{\rho^2}\left(\rho^2\left(\|x_0 - x^*\| + \frac{3C''}{\epsilon}\right)^2\right. \quad (70)\\ &\quad \left. + 2\rho\|\nabla f(x_0)\|\left(\frac{3C''}{\epsilon} + \|x_0 - x^*\|\right) + 2\|\nabla f(x_0)\|^2\right) + 4L_f\frac{9C''^2}{\epsilon^2} \\ &= 8L_f\left(\|x_0 - x^*\| + \frac{3C''}{\epsilon}\right)^2 + \frac{48L_fC''\|\nabla f(x_0)\|}{\rho\epsilon} \\ &\quad + \frac{16L_f}{\rho}\|\nabla f(x_0)\|\|x_0 - x^*\| + \frac{16L_f}{\rho^2}\|\nabla f(x_0)\|^2 + \frac{36L_fC''^2}{\epsilon^2} \quad (71)\\ &= \mathcal{O}(\frac{1}{\epsilon^2}).\end{aligned}$$

We get (68) and (69) from the simple inequality $(a + b)^2 \leq 2a^2 + 2b^2$. We apply (16) to get (70). In (71), we simply rearrange the terms.

□

In the above three lemmas, we have shown the following properties of ξ that directly imply a $\mathcal{O}(\log(\frac{1}{\epsilon}))$ number of consecutive null steps :

- $\xi_{j+1} \leq \tau\xi_j$ (with $\tau < 1$ independent of ϵ).
- $\xi_j \geq A\epsilon^3$ for a null step j.
- $\xi_{k+1} \leq M\epsilon^{-2}$ for a serious step k.

Remark 25. *For both the lower bound and the upper bound on ξ_j, it is certainly possible to get tighter inequalities (and maybe an upper bound on ξ independent of ϵ). However, this would only affect the constant of the leading term of the convergence* *rate of at most a factor corresponding to the exponents of the ϵ in Lemma 22 and Lemma 24, that is,* $3 + 2 = 5$*.*

2.6 Overall convergence rate

The following theorem is the main result of this paper and provides a convergence rate for Algorithm 3.

Theorem 26. *Algorithm 3 reaches an ϵ-optimal solution in at most the following number of iterations*

$$\left(\frac{\sqrt{2\rho}\|x_0 - x^*\|}{\sqrt{\epsilon}} + 1\right)\left(\frac{5\log(\frac{1}{\epsilon}) + \log\left(\frac{M}{A}\right)}{\log\left(1 + \left(\frac{L_f}{\rho} + \frac{8L_f^2}{\rho^2} + \frac{16L_f^3}{\rho^3}\right)^{-1}\right)} + 1\right).$$

Proof. Let k be the index of a serious step. We have shown in Lemmas 22 and 24 that there exist constants M and A, which depend polynomially on the problem parameters and their inverse, such that $\xi_{k+1} \leq M/\epsilon^2$, and for all *null step* index $j > k$ in the sequence of null steps following the serious step k, $\xi_j > A\epsilon^3$. Also, Lemma 20 gives the linear convergence of ξ_j with rate $\tau = \frac{L_f + \rho C^2}{\rho + L_f + \rho C^2}$. This leads to the following inequality for $j = k + T_k$ in the sequence of null steps,

$$\begin{aligned}
& A\epsilon^3 \leq \xi_j \leq \tau^{T_k - 1}\xi_{k+1} \leq \tau^{T_k - 1} M/\epsilon^2 \\
\Rightarrow \quad T_k & \leq \frac{5\log(\frac{1}{\epsilon}) + \log\left(\frac{M}{A}\right)}{-\log(\tau)} + 1 \\
& = \frac{5\log(\frac{1}{\epsilon}) + \log\left(\frac{M}{A}\right)}{\log\left(1 + \left(\frac{L_f}{\rho} + C^2\right)^{-1}\right)} + 1 \\
& \leq \frac{5\log(\frac{1}{\epsilon}) + \log\left(\frac{M}{A}\right)}{\log\left(1 + \left(\frac{L_f}{\rho} + \frac{8L_f^2}{\rho^2} + \frac{16L_f^3}{\rho^3}\right)^{-1}\right)} + 1.
\end{aligned}$$

We now bound the total number of iterations by the maximum number of consecutive null steps times the number of serious steps given in Theorem 14. This gives the following bound on the total number of iterations

$$\left(\frac{\sqrt{2\rho}\|x_0 - x^*\|}{\sqrt{\epsilon}} + 1\right)\left(\frac{5\log(\frac{1}{\epsilon}) + \log\left(\frac{M}{A}\right)}{\log\left(1 + \left(\frac{L_f}{\rho} + \frac{8L_f^2}{\rho^2} + \frac{16L_f^3}{\rho^3}\right)^{-1}\right)} + 1\right).$$

□

Remark 27. *If we choose $\rho = L_f$, we get the following upper bound on the total number of iterations*

$$\begin{aligned} &\left(\frac{\sqrt{2L_f}\|x_0 - x^*\|}{\sqrt{\epsilon}} + 1\right)\left(\frac{5\log(\frac{1}{\epsilon}) + \log\left(\frac{M}{A}\right)}{\log\left(1 + (1+16+8)^{-1}\right)} + 1\right) \\ &= \left(\frac{\sqrt{2L_f}\|x_0 - x^*\|}{\sqrt{\epsilon}} + 1\right)\left(\frac{5}{\log(1+1/25)}\log(\frac{1}{\epsilon}) + K_1 \log\left(L_f\right) + K_2\right), \end{aligned} \tag{72}$$

where K_1 is a numerical constant and K_2 is a constant that depends only on the initial gap in function value and distance to an optimal solution and is independent of (ϵ, ρ, L_f).

In [3], it shows that the lower bound for algorithms using only first-order information such as Algorithm 3 is asymptotically,

$$\frac{\sqrt{L_f}\|x_0 - x^*\|}{\sqrt{\epsilon}}. \tag{73}$$

Our rate (72) *is suboptimal because it includes an additional* $\log(\frac{1}{\epsilon})$ *term. The dependence on the smoothness parameter L_f appears under the form $\sqrt{L_f}\log(L_f)$ compared to $\sqrt{L_f}$ in the lower bound* (73).

3 Acceleration of the PBM in a nonsmooth setting

Unlike the gradient descent algorithm, accelerated convergence rates for Nesterov acceleration applied to the proximal point algorithm are not restricted to smooth objective functions. Notably, [16–18] establish a convergence rate of $\mathcal{O}(\frac{1}{\sqrt{\epsilon}})$ for their algorithms without requiring smoothness of the objective function.

When the proximal subproblem can be solved in a logarithmic number of consecutive null steps, acceleration becomes feasible. This occurs when the objective function is smooth. As demonstrated by [4], the same holds when dealing with composite problems, combining a smooth term with a polyhedral component [4]. In this section, we prove that applying Nesterov's acceleration to the updates of the proximal center in this framework yields a convergence rate of $\mathcal{O}\left(\frac{1}{\sqrt{\epsilon}}\log\left(\frac{1}{\epsilon}\right)\right)$. This rate, however, depends on the problem's dimension, as does the bound on the number of consecutive null steps.

3.1 Assumptions

We study the following composite problem

$$\min_{x\in R^n} h(x) := g(x) + f(x). \tag{74}$$

We assume that g is a convex function and L_g-smooth. No longer do we require f to be differentiable; instead, we now assume that f is a convex piecewise linear function. That is, there exists a finite subset $\mathcal{V}$ of $\mathbb{R}^{n+1}$ such that

$$\forall x \in \mathbb{R}^n, \quad f(x) = \max_{(v,b)\in\mathcal{V}} \langle v, x\rangle + b.$$

Let D and γ respectively denote the diameter and pyramidal width of $\mathcal{V}$ [4, 25].

We also define $D_b = \max\{|b_1 - b_2| : (v_1, b_1), (v_2, b_2) \in \mathcal{V}\}$ and M_f the Lipschitz continuity constant of f such that $M_f = \max\{\|v\| : (v, b) \in \mathcal{V}\}$.

We suppose that an oracle can solve the bundle subproblem

$$\min_{y\in R^n} g(y) + f_k(y) + \frac{\rho}{2}\|y - x_j\|^2, \tag{75}$$

with x_j the proximal center and f_k a lower model of f.

Compared to the previous setting we add the assumption that the set of optimal solutions is bounded.

3.2 Algorithm description

The ϵ-subdifferential of h at the point $z \in \mathbb{R}^n$ is the set

$$\partial_\epsilon h(z) = \{\xi \in \mathbb{R}^n : h(x) \geq h(z) + \langle \xi, x - z\rangle - \epsilon, \forall x \in \mathbb{R}^n\},$$

and we recall the notation $\text{prox}_{\rho,h}(x) = \arg\min_{y\in R^n} h(y) + \frac{\rho}{2}\|y - x\|^2$.

Definition 28 (Definition 2 in [17])**.** *We say that $y \in \mathbb{R}^n$ is a type 2 approximation of $\text{prox}_{\rho,h}(x_j)$ with ϵ_j-precision and write $y \approx_2 \text{prox}_{\rho,h}(x_j)$ with ϵ_j-precision if and only if*

$$\rho(x_j - y) \in \partial_{\frac{\epsilon_j^2 \rho}{2}} h(y).$$

We consider the accelerated inexact proximal point algorithm given in [17]. It specifies the approximation criterion of Algorithm 2 using $\approx_2$ and removes the additional sequence ζ_j.

Algorithm 4 Accelerated Proximal Point Algorithm with Type 2 Approximation

Require: h, $x_0 \in \mathbb{R}^n$
  Initialize ϵ_0. Set $y_0 = x_0$, and $t_0 = 1$
  for $j \geq 0$ **do**
    Compute $y_{j+1} \approx_2 \text{prox}_{\rho,h}(x_j)$ with ϵ_j-precision
    $t_{j+1} \leftarrow \frac{1+\sqrt{1+4t_j^2}}{2}$
    $x_{j+1} \leftarrow y_{j+1} + \frac{t_j - 1}{t_{j+1}}(y_{j+1} - y_j)$
    Update ϵ_{j+1}
  end for

Using a piecewise linear model such as in the classic PBM (Algorithm 1), we derive the following accelerated PBM.

Algorithm 5 Accelerated Proximal Bundle Method

Require: f, g, $x_0 \in \mathbb{R}^n$, $\mathcal{I}_0$ an index set of initial cuts, $B > 0$ and $\rho > 0$

Set $y_0 = \zeta_0 = x_0$, $t_0 = 1$ and $j = 0$

for $k \geq 0$ **do**

Compute y_{k+1} solving the following convex QCQP

$$\begin{aligned} \min_{t\in\mathbb{R}, y\in\mathbb{R}^n} \quad & t + g(y) + \frac{\rho}{2}\|y - x_j\|^2 \\ \text{s.t.} \quad & \forall i \in \mathcal{I}_k, \quad t - f(y_i) - \langle v_i, y - y_i\rangle \geq 0. \end{aligned} \tag{76}$$

Compute $f(y_{k+1})$ and $v_{k+1} \in \partial f(y_{k+1})$

$\mathcal{I}_{k+1} \leftarrow \mathcal{I}_k \cup \{k+1\}$

if $f(y_{k+1}) - f_k(y_{k+1}) \leq \frac{\rho\epsilon_j^2}{2}$ **then** ▷ (null step test)

$t_{j+1} \leftarrow \frac{1+\sqrt{1+4t_j^2}}{2}$

$\zeta_{j+1} \leftarrow y_{k+1}$

$x_{j+1} \leftarrow \zeta_{j+1} + \frac{t_j - 1}{t_{j+1}}(\zeta_{j+1} - \zeta_j)$ ▷ (serious step)

$\epsilon_{j+1} \leftarrow \frac{\sqrt{6B}}{\pi\sqrt{\rho}(j+3)^2}$

$j \leftarrow j + 1$

end if

end for

Comparison between Algorithm 3 and Algorithm 5

Both algorithms share a similar structure and can be viewed as accelerated versions of the classic PBM (Algorithm 1). The Nesterov acceleration is implemented using the same sequence, t_k. Additionally, the bundle minimization step (76) in Algorithm 5 can be interpreted as the bundle minimization step (3) in Algorithm 3, where the smoothness constant $L_f = +\infty$ since f is not assumed to be differentiable.

Despite these similarities, the two algorithms differ in their null step tests. In Algorithm 5, the test requires knowledge of the desired accuracy, ϵ. Furthermore, Algorithm 3 keeps track of the sequence ζ which is updated as $\zeta_{k+1} \leftarrow x_k - \frac{1}{\rho}\nabla f(y_{k+1})$.

3.3 Analysis of the algorithm

3.3.1 Number of serious steps

The following proposition establishes a connection between the two criteria: the gap between the value of the lower model and the true function, as typically considered in the cutting-plane literature, and the inexactness criterion $\approx_2$ from the proximal point algorithm literature.

Proposition 29. *Let y_{k+1} be an iterate of the Algorithm 5 such that the null step test $f(y_{k+1}) - f_k(y_{k+1}) \le \frac{\rho\epsilon_j^2}{2}$ is satisfied for some $j \ge 0$. Then,*

$$y_{k+1} \approx_2 \operatorname{prox}_{\rho,h}(x_j) \quad \textit{with } \epsilon_j\textit{-precision.}$$

Proof. Let $u \in R^n$.

$$\begin{aligned} g(u) + f(u) + \frac{\rho}{2}\|u - x_j\|^2 &\ge g(u) + f_k(u) + \frac{\rho}{2}\|u - x_j\|^2 \\ &\ge g(y_{k+1}) + f_k(y_{k+1}) + \frac{\rho}{2}\|y_{k+1} - x_j\|^2 + \frac{\rho}{2}\|y_{k+1} - u\|^2. \\ &\ge g(y_{k+1}) + f(y_{k+1}) + \frac{\rho}{2}\|y_{k+1} - x_j\|^2 + \frac{\rho}{2}\|y_{k+1} - u\|^2 - \frac{\rho\epsilon_j^2}{2}. \end{aligned}$$

The first line inequality uses $f \ge f_k$. The second inequality leverages the optimality of y_{k+1} for the bundle minimization step (75) and the strong convexity of this problem. The last inequality uses the hypothesis that $f(y_{k+1}) - f_k(y_{k+1}) \le \frac{\rho\epsilon_j^2}{2}$.

Rearranging, we get

$$\begin{aligned} g(u) + f(u) &\ge g(y_{k+1}) + f(y_{k+1}) + \frac{\rho}{2}\left(\|y_{k+1} - x_j\|^2 + \|y_{k+1} - u\|^2 - \|u - x_j\|^2\right) - \frac{\rho\epsilon_j^2}{2} \\ &= g(y_{k+1}) + f(y_{k+1}) + \rho\langle u - y_{k+1}, x_j - y_{k+1}\rangle - \frac{\rho\epsilon_j^2}{2}. \end{aligned}$$

As this inequality holds for all $u \in \mathbb{R}^n$, this proves that $\rho(x_j - y_{k+1}) \in \partial_{\frac{\epsilon_j^2\rho}{2}} h(y_{k+1})$. Following the definition of $\approx_2$ with ϵ_j-precision, this concludes the proof. ☐

Corollary 30. *Algorithm 5 is an instance of Algorithm 4 with $h = f + g$.*

Proof. After noticing that Algorithm 5 contains an implicit loop in which an approximate solution of the proximal problem is computed and during which the proximal center is not updated, this corollary follows directly from Proposition 29. ☐

Theorem 31 (Number of serious steps of Algorithm 5)**.** *At the j^{th} serious step, the iterate ζ_j is such that*

$$h(\zeta_j) - h^* \le \beta_j(h(x_0) - h^* + \frac{\rho}{2}\|x_0 - x^*\|^2) + \delta_j \tag{77}$$

with $\beta_j \le \frac{1}{(1+\frac{j}{2})^2}$ and $\delta_j \le \frac{\beta_j}{2}\sum_{i=0}^{j-1}\rho\epsilon_i^2\left(1 + \sqrt{2}(i+1)\right)^2$.

Moreover, for $B > 0$ choosing $\epsilon_j \le \frac{\sqrt{2B}\cdot\sqrt{6}/\pi}{\sqrt{\rho}(j+1)\left(1+\sqrt{2}(j+1)\right)}$ leads to $\delta_j \le B\beta_j$. We can then rewrite the inequality (77) as

$$h(\zeta_j) - h^* \le \frac{4}{(j+2)^2}(h(x_0) - h^* + \frac{\rho}{2}\|x_0 - x^*\|^2 + B). \tag{78}$$

Proof. A more general proof is given in Remark 4.7 of [17]. Compared to their notations, we have $\forall i, \lambda_i = \rho^{-1}$, $a = 1$, $A_0 = \rho$, and $p = 1$. □

This theorem shows that the number of serious steps to reach an ϵ-optimal solution is at most

$$\frac{2\sqrt{h(x_0) - h^* + \frac{\rho}{2}\|x_0 - x^*\|^2 + B}}{\sqrt{\epsilon}}. \tag{79}$$

Lemma 32. *While an ϵ-optimal solution has not been found, choosing $\epsilon_j = \frac{\sqrt{6B}}{\pi\sqrt{\rho}(j+2)^2}$ satisfies the upper bound on ϵ_j in Theorem 31; moreover, such ϵ_j is lower bounded as*

$$\forall j, \quad \epsilon_j \geq \frac{\sqrt{6B}\epsilon}{4\pi\sqrt{\rho}\left(h(x_0) - h^* + \frac{\rho}{2}\|x_0 - x^*\|^2 + B\right)}. \tag{80}$$

Proof. We suppose that $h(\zeta_j) - h^* \geq \epsilon$. Following (78),

$$\epsilon \leq \frac{4}{(j+2)^2}(h(x_0) - h^* + \frac{\rho}{2}\|x_0 - x^*\|^2 + B).$$

Rearranging and plugging the maximum value of j (given in (79)) in the expression of ϵ_j leads to the claimed bound. □

3.3.2 Number of consecutive null steps

Theorem 33 (Lemma 3.6 in [4])**.** *In Algorithm 5, at most the following number of consecutive null steps occur between two serious steps*

$$\left[1 + \max\left\{2, \frac{D^2}{\mu_{\psi,\rho,j}\rho\gamma^2}\right\}\log\left(\frac{4D^4}{\rho^4\epsilon_j^2}\right)\right], \tag{81}$$

with $\mu_{\psi,\rho}$ such that

$$\frac{1}{2}\mu_{\psi,\rho,j}^{-1} = D_b + 3\frac{8M_f^2}{\rho} + 6M_f\|x_j^*\| + 2L_g\left[\left(\frac{4M_f}{\rho} + \|x_j^*\|\right)^2 + 1\right], \tag{82}$$

where $x_j^* = \arg\min_x h(x) + \frac{\rho}{2}\|x - x_j\|^2$.

Definition 34 (Level-Boundedness)**.** *A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be* level-bounded *if for every $c \in \mathbb{R}$, the level set*

$$L_c = \{x \in \mathbb{R}^n : f(x) \leq c\}$$

is bounded.

We recall the following known result (Lemma 35) that derives the level-boundedness of the convex function h based on the boundedness of its minimizers. For completeness, a proof is given in the Appendix A.3.

Lemma 35 (Level-Boundedness of Convex Functions with Bounded Minimizers)**.** *If* $\{x \in \mathbb{R}^n : f(x) = \min_{y\in\mathbb{R}^n} f(y)\}$ *is non-empty and bounded, then* f *is level bounded.*

The constant $\mu_{\psi,\rho,j}$ defined in Theorem 33 depends on $\|x_j^*\|$, the norm of the proximal solution $\text{prox}_{\rho,h}(x_j)$. We leverage the level-boundedness of h to bound $\|x_j^*\|$ uniformly in j.

Proposition 36. *The optimal solutions of the proximal problems* $(x_j^*)_{j\in\mathbb{N}}$ *are bounded as*

$$\forall j, \quad \|x_j^*\| \leq 3\|x^*\| + 6R_{h,x_0}, \tag{83}$$

where R_{h,x_0} *is the radius of a Euclidean ball centered at zero that contains the level set of* h *at level* $h(x_0) + \frac{\rho}{2}\|x_0 - x^*\|^2 + B$*. Note that* R_{h,x_0} *does not depend on* ϵ*.*

Proof. By Theorem 31, $\beta_j \leq 1$ for all j. Following (78), all the ζ_j are in the level set of h with level $h(x_0) + \frac{\rho}{2}\|x_0 - x^*\|^2 + B$.

We combine this bound with the update of the proximal center given in Algorithm 5

$$\begin{aligned}\|x_{j+1}\| &= \left\|\zeta_{j+1} + \frac{t_j - 1}{t_{j+1}}(\zeta_{j+1} - \zeta_j)\right\| \\ &\leq \|\zeta_{j+1}\| + \frac{t_j - 1}{t_{j+1}}(\|\zeta_{j+1}\| + \|\zeta_j\|) \\ &\leq 3R_{h,x_0}.\end{aligned}$$

For the last inequality we use $\frac{t_j-1}{t_{j+1}} \leq 1$ as in (17).

By optimality of x^* and x_j^* for their respective problems,

$$h(x^*) + \frac{\rho}{2}\|x_j^* - x_j\|^2 \leq h(x_j^*) + \frac{\rho}{2}\|x_j^* - x_j\|^2 \leq h(x^*) + \frac{\rho}{2}\|x^* - x_j\|^2.$$

Thus, $\|x_j^* - x_j\| \leq \|x^* - x_j\|$. We combine these two inequalities,

$$\begin{aligned}\|x_j^*\| &\leq \|x^*\| + \|x_j^* - x^*\| \leq \|x^*\| + \|x_j^* - x_j\| + \|x^* - x_j\| \\ &\leq \|x^*\| + 2\|x^* - x_j\| \leq 3\|x^*\| + 2\|x_j\| \\ &\leq 3\|x^*\| + 6R_{h,x_0}. \end{aligned} \tag{84}$$

□

Theorem 37. *Algorithm 5 provides an* ϵ*-optimal solution in at most the following number of iterations*

$$\frac{2\sqrt{h(x_0) - h^* + \frac{\rho}{2}\|x_0 - x^*\|^2 + B}}{\sqrt{\epsilon}} \cdot \left[1 + \max\left\{2, \frac{D^2}{\mu_{\psi,\rho}\rho\gamma^2}\right\} \log\left(\frac{4D^4}{\rho^4\sigma\epsilon^2}\right)\right] \tag{85}$$

with

$$\frac{1}{2}\mu_{\psi,\rho}^{-1} \stackrel{def}{=} D_b+3\frac{8M_f^2}{\rho}+6M_f(3\|x^*\|+6R_{h,x_0})+2L_g\left[\left(\frac{4M_f}{\rho}+3\|x^*\|+6R_{h,x_0}\right)^2+1\right],$$

and

$$\sigma = \frac{6B}{\pi^2\rho\left(h(x_0)-h^*+\frac{\rho}{2}\|x_0-x^*\|^2+B\right)^2}. \tag{86}$$

Proof. We upper bound the total number of iterations by the bound on the number of serious steps given in (79) and the bound on the number of consecutive null steps given in Theorem 33.

Combining the definition of $\mu_{\psi,\rho,j}$ in (82) and (84), we get $\forall j$, $\mu_{\psi,\rho} \leq \mu_{\psi,\rho,j}$.

Lemma 32 shows that for all serious step index j such that an ϵ-optimal solution has not been found yet, $\sigma\epsilon^2 \leq \epsilon_j^2$.

□

4 Remarks and conclusion

Bundle management and varying ρ parameter

For simplicity, throughout this article, we have assumed that the parameter ρ remains fixed and that all cuts are retained in memory during sequences of null steps. However, the proposed algorithms can be readily adapted to more general settings where ρ may be updated at serious steps following any nonincreasing sequence as Theorem 14 still holds. Also, selective cut management strategies may be employed, without significant changes to the convergence properties.

Conclusion

In this paper, we have proposed an Accelerated Proximal Bundle algorithm designed to improve the convergence rate of the proximal bundle methods for smooth objectives. By incorporating a novel null step test and a smooth lower model, we demonstrated that our algorithm achieves an improved convergence rate of $\mathcal{O}(\frac{1}{\sqrt{\epsilon}}\log(\frac{1}{\epsilon}))$. This enhancement addresses an open question about achieving accelerated rates with proximal bundle methods. We have further proposed an acceleration of the proximal bundle method for composite convex optimization with piecewise linear nonsmoothness. The resulting Algorithm 5 achieves a similar rate of $\mathcal{O}(\frac{1}{\sqrt{\epsilon}}\log(\frac{1}{\epsilon}))$, although the complexity bound also depends the problem's dimension through some geometric quantity such as the pyramidal width of the subdifferentials of the piecewise linear function.

Appendix A Additional Proofs

A.1 Proof of Proposition 12

Proof. Let $x_\delta \in \mathbb{R}^n$ such that $\bar{h} - h(x_\delta) \leq \delta$. We consider the problem

$$\max_{x \in \mathbb{R}^n} h(x) - \frac{\delta}{2}\|x - x_\delta\|. \tag{A1}$$

As h is upper-bounded, (A1) can be restricted to a compact feasible set. By continuity of the objective, this problem admits an optimal solution x_δ^*. In particular, as $h(x_\delta^*) \geq h(x_\delta)$, $\|\nabla h(x_\delta^*)\| \leq \delta$ would conclude the proof. We assume the opposite. h is differentiable at x_δ^* so there exists $t > 0$ such that $z = x_\delta^* + t\nabla h(x_\delta^*)$ satisfies:

$$\begin{aligned}
h(z) = h(x_\delta^* + t\nabla h(x_\delta^*)) &\geq h(x_\delta^*) + \frac{3}{4}\left(t\nabla h(x_\delta^*)^T\right)\nabla h(x_\delta^*) \\
&= h(x_\delta^*) + \frac{3}{4}\|x_\delta^* - z\| \cdot \|\nabla h(x_\delta^*)\| \\
\implies \quad h(z) - \frac{\delta}{2}\|z - x_\delta\| &\geq h(x_\delta^*) + \delta\frac{3}{4}\|x_\delta^* - z\| - \frac{\delta}{2}\|z - x_\delta\| \\
&\geq h(x_\delta^*) + \delta\frac{3}{4}\|x_\delta^* - z\| - \frac{\delta}{2}(\|x_\delta^* - z\| + \|x_\delta^* - x_\delta\|) \\
&\geq h(x_\delta^*) - \frac{\delta}{2}\|x_\delta^* - x_\delta\| + \frac{\delta}{4}\|x_\delta^* - z\|.
\end{aligned}$$

This contradicts the maximality of x_δ^* and concludes the proof. □

A.2 Proof of Lemma 18

Proof. We first consider the homogeneous recurrence $s_0 = r_0$ and

$$\begin{aligned}
s_{k+1} &= (1 + \frac{2}{k+2})s_k \\
\Leftrightarrow \log(s_{k+1}) - \log(s_k) &= \log(1 + \frac{2}{k+2}) \\
&\leq \frac{2}{k+2}. \qquad \text{(concavity of the log)}
\end{aligned}$$

We sum up these inequalities,

$$\begin{aligned}
\log(s_{k+1}) - \log(s_0) &\leq 2\sum_{i=0}^{k}\frac{1}{i+2} \\
&\leq 2\sum_{i=1}^{k+1}\frac{1}{i} \\
&\leq 2 + 2\log(k+1).
\end{aligned}$$

This yields : $s_{k+1} \leq s_0 \cdot e^2 \cdot (k+1)^2$, for $k \geq 0$.

Using the inequality $\forall x > -1, \log(1+x) \geq x - x^2/2$:

$$\begin{aligned}\log(s_{k+1}) - \log(s_0) &\geq 2\sum_{i=0}^{k} \frac{1}{i+2} - \frac{1}{2}\sum_{i=0}^{k}\left(\frac{2}{i+2}\right)^2 \\ &= 2\sum_{i=2}^{k+2}\frac{1}{i} - 2\sum_{i=2}^{k+2}\left(\frac{1}{i}\right)^2 \\ &\geq -1 + 2\log(k+3) - 2\frac{\pi^2}{6}.\end{aligned}$$

This gives $s_k \geq s_0 \exp(-1 - \frac{\pi^2}{3})(k+2)^2$ for $k > 0$. We now show that the inhomogeneous term $\frac{2C'}{k+2}$ vanishes at infinity.

For $k \geq 0$, let $w_k = \frac{r_k}{s_k}$. We have $w_0 = 1$.

$$\begin{aligned}w_{k+1} = \frac{r_{k+1}}{s_{k+1}} &\leq \frac{(1+\frac{2}{k+2})r_k + \frac{2C'}{k+2}}{(1+\frac{2}{k+2})s_k} = w_k + \frac{\frac{2C'}{k+2}}{(1+\frac{2}{k+2})s_k} = w_k + \frac{2C'}{(k+4)s_k} \\ &\leq w_k + \frac{2C'}{(k+4)s_0\exp(-1-\frac{\pi^2}{3})(k+2)^2} \leq w_k + \frac{2C'}{s_0 \cdot \exp(-1-\frac{\pi^2}{3}) \cdot (k+1)^3}.\end{aligned}$$

Solving this simple linear recurrence leads to $\forall k \geq 0$,

$$w_k \leq w_0 + \frac{2C'}{s_0 \cdot \exp(-1-\frac{\pi^2}{3})} \cdot \sum_{i=1}^{+\infty}\frac{1}{k^3} \leq 1 + \frac{2C'\pi^2}{6s_0 \cdot \exp(-1-\frac{\pi^2}{3})}.$$

Combining the bounds on w_k and s_k, we get

$$\forall k \geq 1, \quad r_k = w_k s_k \leq \left(e^2 \cdot r_0 + \frac{C'\pi^2 e^{(3+\frac{\pi^2}{3})}}{3}\right) \cdot k^2.$$

□

A.3 Proof of Lemma 35

Proof. Since the set of minimizers of f is non-empty, the minimum $c_1 := \min_{y\in\mathbb{R}^n} f(y)$ is finite. We proceed by contradiction. By the assumption of the lemma, the level set $L_1 = \{x : f(x) \leq c_1\}$ is bounded. Suppose there exists $c_2 > c_1$ such that $L_2 = \{x : f(x) \leq c_2\}$ is unbounded. Note that $L_1 \subseteq L_2$. Since L_2 is unbounded, it follows that the recession cone of L_2, denoted by 0^+L_2, is nontrivial; that is, there exists a nonzero direction $d \in 0^+L_2$ (see, for instance, [19] Theorem 8.4).

In particular, for any $x_1 \in L_1$, we have $x_1 + td \in L_2$ for all $t \geq 0$. Define the function $\varphi(t) = f(x_1+td)$. Since $x_1+td \in L_2$ for all $t \geq 0$, $\varphi(t)$ must be nonincreasing.

Otherwise, for some $t > 0$, we would have $f(x_1 + td) > c_2$, which contradicts the assumption that $x_1 + td \in L_2$ for all $t \geq 0$.

Moreover, since $\varphi(t)$ is nonincreasing and $f(x_1) \leq c_1$, it follows that $f(x_1+td) \leq c_1$ for all $t \geq 0$, implying that $x_1+td \in L_1$ for all $t \geq 0$. This contradicts the boundedness of L_1. Therefore, all level sets of f must be bounded. □

References

[1] Liang, J., Monteiro, R.D.C.: A unified analysis of a class of proximal bundle methods for solving hybrid convex composite optimization problems. Mathematics of Operations Research **0**(0) (2023)

[2] Diaz, M., Grimmer, B.: Optimal convergence rates for the proximal bundle method. SIAM Journal on Optimization **33**(2), 424–454 (2023) https://doi.org/10.1137/21M1428601

[3] Drori, Y.: The exact information-based complexity of smooth convex minimization. Journal of Complexity **39**, 1–16 (2017) https://doi.org/10.1016/j.jco.2016.11.001 . Accessed 2024-07-22

[4] Fersztand, D., Sun, A.: A modified proximal bundle method under a Frank-Wolfe perspective (2024). https://arxiv.org/abs/2411.15926

[5] Kelley, J.E. Jr.: The cutting-plane method for solving convex programs. Journal of the Society for Industrial and Applied Mathematics **8**(4), 703–712 (1960)

[6] Kiwiel, K.C.: Proximity control in bundle methods for convex nondifferentiable minimization. Mathematical Programming **46**(1), 105–122 (1990)

[7] Oliveira, W., Sagastizábal, C., Lemaréchal, C.: Convex proximal bundle methods in depth: a unified analysis for inexact oracles. Mathematical Programming **148**(1), 241–277 (2014)

[8] Schramm, H., Zowe, J.: Bundle method for nonsmooth optimization: Concept, convergence, and computation. SIAM Journal on Optimization **2**(1), 121–152 (1992)

[9] Lemaréchal, C., Nemirovskii, A., Nesterov, Y.: New variants of bundle methods. Mathematical Programming **69**(1), 111–147 (1995)

[10] Lan, G.: Bundle-level type methods uniformly optimal for smooth and nonsmooth convex optimization. Mathematical Programming **149**(1), 1–45 (2015)

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

[12] Du, Y., Ruszczyński, A.: Rate of convergence of the bundle method. Journal of

Optimization Theory and Applications **173**(3), 908–922 (2017)

[13] Liang, J., Monteiro, R.D.C.: A proximal bundle variant with optimal iteration-complexity for a large range of prox stepsizes. SIAM Journal on Optimization **31**(4), 2955–2986 (2021) https://doi.org/10.1137/20M1327513 https://doi.org/10.1137/20M1327513

[14] Martinet, B.: Algorithmes pour la résolution de problèmes d'optimisation et de minimax. PhD thesis, Université Scientifique et Médicale de Grenoble (1972)

[15] Rockafellar, R.T.: Monotone operators and the proximal point algorithm. SIAM Journal on Control and Optimization **14**(5), 877–898 (1976)

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

[17] Salzo, S., Villa, S.: Inexact and accelerated proximal point algorithms. Optimization Online (2012). Mathematical Subject Classification: 90C25 (49D37, 65K10)

[18] He, B., Yuan, X.: An accelerated inexact proximal point algorithm for convex minimization. Journal of Optimization Theory and Applications **154**(2), 536–548 (2012) https://doi.org/10.1007/s10957-011-9948-6

[19] Rockafellar, R.: Convex Analysis. Princeton Mathematics. Princeton University Press, Princeton, New Jersey (1970)

[20] Drori, Y., Teboulle, M.: An optimal variant of Kelley's cutting-plane method. Mathematical Programming, Series A **160**, 321–351 (2016) https://doi.org/10.1007/s10107-016-0985-7

[21] Florea, M.I., Nesterov, Y.: An optimal lower bound for smooth convex functions (2024). https://arxiv.org/abs/2404.18889

[22] Taylor, A.B., Hendrickx, J.M., Glineur, F.: Smooth strongly convex interpolation and exact worst-case performance of first-order methods. Mathematical Programming, Series A **161**, 307–345 (2017)

[23] Bonnans, J.F., Shapiro, A.: Perturbation Analysis of Optimization Problems. Springer, New York, NY (2000). https://doi.org/10.1007/978-1-4612-1394-9 . http://link.springer.com/10.1007/978-1-4612-1394-9 Accessed 2024-07-17

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

[25] Lacoste-Julien, S., Jaggi, M.: On the global linear convergence of frank-wolfe optimization variants, 496–504 (2015)