

Proximal bundle methods for hybrid weakly convex composite optimization problems

Jiaming Liang^{*} Renato D.C. Monteiro[†] Honghao Zhang[†]

March 26, 2023

(1st revision: Dec 26, 2024; 2nd revision: May 16, 2026; 3rd revision: Oct 13, 2026)

Abstract

This paper establishes the iteration-complexity of proximal bundle methods for solving hybrid (i.e., a blend of smooth and nonsmooth) weakly convex composite optimization (HWC-CO) problems. This is done in a unified manner by considering a proximal bundle framework (PBF), which includes various well-known bundle update schemes. In contrast to hard-to-check stationary conditions (e.g., the Moreau stationarity) used by other methods for solving HWC-CO, PBF uses a stationarity measure that is easily verifiable.

Key words. hybrid weakly convex composite optimization, iteration-complexity, proximal bundle method, regularized stationary point, Moreau envelope.

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

1 Introduction

Let $h : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ be a proper lower semi-continuous convex function and $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ be a proper lower semi-continuous m -weakly convex function (i.e., $f + (m/2)\|\cdot\|^2$ is convex) such that $\text{dom } f \supseteq \text{dom } h$ and consider the composite optimization (CO) problem

$$\min \{\phi(x) := f(x) + h(x) : x \in \mathbb{R}^n\}. \quad (1)$$

It is said that (1) is a hybrid weakly convex CO (HWC-CO) problem if there exist nonnegative scalars M and L and a first-order oracle $f' : \text{dom } h \rightarrow \mathbb{R}^n$ (i.e., $f'(x) \in \partial f(x)$ for every $x \in \text{dom } h$) satisfying the (M, L) -hybrid condition, namely: $\|f'(u) - f'(v)\| \leq 2M + L\|u - v\|$ for every $u, v \in \text{dom } h$. This problem class includes the class of weakly convex non-smooth (resp., smooth) CO problems, i.e., the one with $L = 0$ (resp., $M = 0$). Throughout our complexity analysis, we assume $m > 0$. Thus, the convex case $m = 0$ is not formally covered by the complexity results.

This problem class appears in various applications in modern data science where f is usually a loss function and h is either the indicator function of some set (e.g., the set of points satisfying some functional constraints) or a regularization function that imposes sparsity or some special structure on the solution being sought. Examples of such applications are robust phase retrieval, covariance matrix estimation, and sparse dictionary learning (see Subsection 2.1 of [5] and the references therein). In these applications, the loss is often weakly convex while the regularizer is convex but nonsmooth, making the HWC-CO framework a natural modeling setting.

^{*}Goergen Institute for Data Science and Artificial Intelligence and Department of Computer Science, University of Rochester, Rochester, NY 14620 (email: jiaming.liang@rochester.edu). This work was partially supported by AFOSR grant FA9550-25-1-0182.

[†]School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, 30332-0205. (email: renato.monteiro@isye.gatech.edu and hzhang906@gatech.edu). This work was partially supported by AFOSR Grants FA9550-22-1-0088 and FA9550-25-1-0131.

The main goal of this paper is to study the complexity of a unified framework (referred to as PBF) of proximal bundle (PB) methods for solving the HWC-CO problem (1). More specifically, like other proximal bundle methods, a PBF iteration solves a prox bundle subproblem of the form

$$x = \operatorname{argmin}_{u \in \mathbb{R}^n} \left\{ \Gamma(u) + \frac{1}{2\lambda} \|u - x^c\|^2 \right\} \quad (2)$$

where λ is the prox stepsize, x^c is the current prox-center, and (usually simple) lower semi-continuous convex function Γ denotes the current bundle function. Moreover, the bundle function is updated in every iteration but the prox-center x^c is updated in some (i.e., serious) iterations and left the same in the other (i.e., null) ones. Instead of choosing the bundle function underneath ϕ as in the proximal bundle methods for solving the convex version of (1), PBF uses the idea of choosing Γ underneath $\phi(\cdot) + (m/2)\|\cdot - x^c\|^2$. An interesting feature of PBF is that it terminates based on an easily verifiable stopping criterion.

Literature Review: We first discuss the development of PB methods for convex CO problems. PB methods were introduced in [21, 33] and subsequently developed in, for example, [12, 26, 28]. Their convergence analyses for nonsmooth convex problems are discussed extensively in the textbooks [29, 30], while their iteration complexity has been studied in, for example, [8, 10, 19, 23, 24]. For broader treatments of bundle methods, including methods for nonconvex nonsmooth optimization, we also refer to the recent textbooks [2, 31].

We next discuss PB methods for weakly convex CO problems, as introduced at the beginning of this section, focusing first on the nonsmooth case $L = 0$. Extensions of PB methods to this setting can be found in [13, 14, 15, 16, 20, 25, 27, 32]. These works primarily establish asymptotic convergence results. In particular, [14, 15, 16] already considered constructing convex bundle (more precisely, cutting-plane) models underneath the regularized function $\phi(\cdot) + (m/2)\|\cdot - x^c\|^2$, which is also the basic convexification principle used by PBF. Thus, the distinction of PBF does not lie in this convexification principle itself. Rather, PBF develops a general proximal bundle framework based on this principle and establishes explicit overall iteration-complexity bounds for obtaining regularized stationary points under the hybrid condition considered in this paper. The more recent work [1] develops a unified model for analyzing descent-type bundle methods and establishes local convergence rates for the serious iterate and function value sequences under a strong stationary growth condition. In contrast, the focus of the present work is on global nonasymptotic iteration-complexity guarantees.

Weakly convex problems also admit a natural difference-of-convex (DC) representation, and hence PBF is related to bundle methods developed for nonsmooth DC programming; see, for example, [7, 17, 18]. We discuss this connection, and the distinction between PBF and these methods, in more detail in Subsection 2.3.

Main contribution. As already mentioned above, this paper establishes the (serious and overall) iteration-complexity of PB methods for solving the HWC-CO problem described at the beginning of this introduction. An interesting feature of our analysis is that it is based on a stationarity measure which, in contrast to other well-known stationary conditions in the context of HWC-CO (including the Moreau stationary one), is easily and directly verifiable. Moreover, by proper choice of tolerances, it is shown that the new measure implies these other well-known near stationarity conditions (including the Moreau stationary one), so that any complexity result based on the first one can be easily translated to the latter ones.

As a consequence of our analysis, it is shown that the iteration-complexity for PBF to find a ρ -Moreau stationary point of ϕ is similar to that of the deterministic versions of the stochastic proximal subgradient (PS) methods studied in [5, 6], i.e., $\mathcal{O}(\rho^{-4})$, but the first complexity bound has the advantage that its constant (in its $\mathcal{O}(\cdot)$) is never worse and is generally better than the one which appears in the bound for the PS method of [5]. The latter feature of PBF is due to its bundle nature which allows it to use a considerably larger prox stepsize that is determined by the weakly convex parameter. This contrasts with the nature of proximal subgradient-type methods which use relatively small prox stepsizes (e.g., depending on a pre-specified iteration count or decreasing towards zero as the latter grows).

Other advantages of our method are: 1) PBF assumes that f satisfies the more general hybrid condition (both [5, 6] only consider the case where $L = 0$); 2) as opposed to the methods of [5, 6], PBF does not require knowledge of the constants M and L (see the first paragraph above), which are usually hard to estimate, however, PBF still requires knowledge of a weak convexity parameter m ; and 3) it can obtain an estimate of the size of the gradient of the Moreau envelope whenever the prox subproblem is solved according to an

easily verifiable stopping criterion; in contrast, Algorithm 2 of [6] needs to perform a pre-specified number of subgradient iterations to the prox subproblem (see Section 3 of [6]) to estimate the size of the gradient of the Moreau envelope.

Organization of the paper. Section 2 contains three subsections. Subsection 2.1 presents basic definitions and notation used throughout the paper. Subsection 2.2 presents three stationary notions and clarifies the relationships among these three notions. Subsection 2.3 reviews related methods for weakly convex nonsmooth optimization. Section 3 contains three subsections. Subsection 3.1 formally describes problem (1) and the assumptions made on it. Subsection 3.2 describes PBF and states the iteration-complexity result of PBF. Subsection 3.3 presents two concrete instances of PBF. Section 4 contains three subsections. Subsection 4.1 provides a preliminary bound on the length of a cycle in PBF and Subsection 4.2 bounds the number of cycles generated by PBF. Subsection 4.3 presents the proof of the main complexity result. Section 5 presents computational results for three problems: phase retrieval, blind deconvolution, and constrained MAXQUAD. Section 6 gives some concluding remarks and potential directions for future research. Appendix A describes two useful technical results about subdifferentials. Appendix B provides proofs for the results in Subsection 2.2. Appendix C presents results about an important recursive formula. Appendix D presents some useful technical results used in Section 4. Finally, Appendix E presents a complementary convergence guarantee in terms of the Moreau envelope using PBF.

2 Background and Related Methods

This section provides the background material used throughout the paper and reviews the PS method for weakly convex optimization. Subsection 2.1 introduces the basic definitions and notation. Subsection 2.2 presents the stationarity notion considered in this paper and two other related notions and describes their relationships. Finally, Subsection 2.3 reviews the deterministic version of the stochastic proximal subgradient method of [5] and briefly discusses the relation between PBF and bundle methods for DC programming.

2.1 Basic definitions and notation

The sets of real numbers and positive real numbers are denoted by $\mathbb{R}$ and $\mathbb{R}_{++}$, respectively. Let $\mathbb{R}^n$ denote the standard n -dimensional Euclidean space equipped with inner product and norm denoted by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$, respectively. Let $\log(\cdot)$ denote the natural logarithm and $\log^+(\cdot)$ denote $\max\{\log(\cdot), 0\}$. Let $\mathcal{O}$ denote the standard big-O notation.

For a given function $\varphi : \mathbb{R}^n \rightarrow (-\infty, +\infty]$, let $\text{dom } \varphi := \{x \in \mathbb{R}^n : \varphi(x) < \infty\}$ denote the effective domain of φ and φ is proper if $\text{dom } \varphi \neq \emptyset$. A proper function $\varphi : \mathbb{R}^n \rightarrow (-\infty, +\infty]$ is μ -convex for some $\mu \geq 0$ if

$$\varphi(\lambda x + (1 - \lambda)y) \leq \lambda\varphi(x) + (1 - \lambda)\varphi(y) - \frac{\mu\lambda(1 - \lambda)}{2}\|x - y\|^2$$

for every $x, y \in \text{dom } \varphi$ and $\lambda \in [0, 1]$. Denote the set of all proper lower semi-continuous convex functions by $\overline{\text{Conv}}(\mathbb{R}^n)$. For $\varepsilon \geq 0$, the ε -subdifferential of φ at $x \in \text{dom } \varphi$ is denoted by

$$\partial_\varepsilon \varphi(x) := \{s \in \mathbb{R}^n : \varphi(y) \geq \varphi(x) + \langle s, y - x \rangle - \varepsilon, \forall y \in \mathbb{R}^n\}. \quad (3)$$

For simplicity, the subdifferential of φ at $x \in \text{dom } \varphi$, i.e., $\partial_0 \varphi(x)$, is denoted by $\partial \varphi(x)$. The set of proper lower semi-continuous convex functions Γ such that $\Gamma \leq \varphi$ is denoted by $\mathcal{B}(\varphi)$ and any such Γ is called a bundle for φ .

2.2 Three notions of stationary points

This subsection introduces the subdifferential and directional derivative of a general proper lower semi-continuous function. It then defines the regularized stationary point sought by the main algorithm of this paper. In addition, it presents two alternative stationarity conditions, formulated in terms of the directional derivative and the Moreau envelope, respectively, and clarifies the relationships among these three notions.

We start by giving the definitions of directional derivative of a proper lower semi-continuous function.

Definition 2.1 The directional derivative $\phi'(x; d)$ of ϕ at x along d is

$$\phi'(x; d) := \liminf_{t \downarrow 0} \frac{\phi(x + td) - \phi(x)}{t}.$$

Motivated by the first-order directional optimality condition, we introduce the following intrinsic notion of approximate stationarity.

Definition 2.2 (Directional stationary point) For a pair $(\varepsilon_D, \delta_D) \in \mathbb{R}_{++}^2$, a point $x \in \text{dom } \phi$ is called a $(\varepsilon_D, \delta_D)$ -directional stationary point if there exists $\tilde{x} \in \text{dom } \phi$ such that

$$\|x - \tilde{x}\| \leq \delta_D, \quad \inf_{\|d\| \leq 1} \phi'(\tilde{x}; d) \geq -\varepsilon_D. \quad (4)$$

When $(\varepsilon_D, \delta_D) = (0, 0)$, then (4) reduces to the condition that $\phi'(x; d) \geq 0$ for all $d \in \mathbb{R}^n$, a condition which is known to be equivalent to $0 \in \partial\phi(x)$, which denotes the Frechet subdifferential and formally defined in Definition 2.3 below.

Definition 2.3 The Frechet subdifferential of a proper lower semi-continuous function $\phi : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$ is defined as

$$\partial\phi(x) = \left\{ v \in \mathbb{R}^n : \liminf_{y \rightarrow x} \frac{\phi(y) - \phi(x) - \langle v, y - x \rangle}{\|y - x\|} \geq 0 \quad \forall x \in \mathbb{R}^n \right\}.$$

For every function $g(\cdot)$, $m \in \mathbb{R}_+$, $z \in \mathbb{R}^n$, let

$$g_m(\cdot; z) := g(\cdot) + \frac{m}{2} \|\cdot - z\|^2. \quad (5)$$

It is well known (see, for example, Lemma 2.1 of [5]) that for an m -weakly convex function ϕ ,

$$\partial\phi(x) = \left\{ v \in \mathbb{R}^n : \phi(y) \geq \phi(x) + \langle v, y - x \rangle - \frac{m}{2} \|y - x\|^2, \quad \forall y \in \mathbb{R}^n \right\} = \partial[\phi_m(\cdot; x)](x). \quad (6)$$

We now introduce the definition of a regularized stationary point of a proper lower semi-continuous m -weakly convex function ϕ which is the one used by the main algorithm of this paper.

Definition 2.4 (Regularized stationary point) For a pair $(\bar{\eta}, \bar{\varepsilon}) \in \mathbb{R}_{++}^2$, a point $x \in \text{dom } \phi$ is called a $(\bar{\eta}, \bar{\varepsilon}; m)$ -regularized stationary point of ϕ if there exists a pair $(w, \varepsilon) \in \mathbb{R}^n \times \mathbb{R}_{++}$ such that

$$w \in \partial_\varepsilon [\phi_m(\cdot; x)](x), \quad \|w\| \leq \bar{\eta}, \quad \varepsilon \leq \bar{\varepsilon}. \quad (7)$$

We make two trivial remarks about the above definition. First, if $(\bar{\eta}, \bar{\varepsilon}) = (0, 0)$ then it follows from (6) and Definition 2.4 that x is a $(\bar{\eta}, \bar{\varepsilon}; m)$ -regularized stationary point if and only if x is an exact stationary point of (1), i.e., it satisfies $0 \in \partial\phi(x)$. Second, when $m = 0$ (and hence ϕ is convex), the inclusion in (7) reduces to $w \in \partial_\varepsilon \phi(x)$, and Definition 2.4 reduces to Definition 6.1 of [23], which has already been used in the analysis of proximal bundle methods for solving the convex version of (1) (see Section 6 of [23]).

The verification that x is a $(\bar{\eta}, \bar{\varepsilon}; m)$ -regularized stationary point requires exhibiting a pair of residuals (w, ε) satisfying the inclusion in (7), which is generally not an immediate task. However, many algorithms for solving the convex version of (1) and the one in this paper, are able to generate not only a sequence of iterates $\{x_k\}$ but also a sequence of corresponding residuals $\{(w_k, \varepsilon_k)\}$ such that $(x, w, \varepsilon) = (x_k, w_k, \varepsilon_k)$ satisfies the inclusion in (7), so that verification that x_k is a $(\bar{\eta}, \bar{\varepsilon}; m)$ -regularized stationary point simply amounts to checking the two inequalities in (7).

Before stating the next notion of a stationary point based on the Moreau envelope, we introduce a slightly different notation for the Moreau envelope which is more suitable for our presentation, namely: for any $\lambda > 0$ and $x \in \mathbb{R}^n$, let

$$\hat{M}^\lambda(x) := \min_u \left\{ \phi(u) + \frac{\lambda^{-1} + m}{2} \|u - x\|^2 \right\}. \quad (8)$$

Note that (8) is the Moreau envelope of ϕ_m and therefore depends on m . Since m is fixed throughout the paper, we omit this dependence from the notation for simplicity.

The gradient formula for the Moreau envelope (see Section 1 in [5]) is as follows:

$$\nabla \hat{M}^\lambda(x) = \left(\frac{1}{\lambda} + m \right) (x - \hat{x}^\lambda(x)) \quad (9)$$

where

$$\hat{x}^\lambda(x) := \operatorname{argmin}_{u \in \mathbb{R}^n} \left\{ \phi(u) + \frac{\lambda^{-1} + m}{2} \|u - x\|^2 \right\}. \quad (10)$$

The Moreau envelope gradient provides a commonly used stationarity measure for weakly convex optimization; see, for example, [5, 6]. Indeed, (9) shows that $\|\nabla \hat{M}^\lambda(x)\|$ is the scaled residual between x and the solution $\hat{x}^\lambda(x)$ of the corresponding proximal subproblem. Moreover, the optimality condition for (10) implies

$$\operatorname{dist}(0, \partial\phi(\hat{x}^\lambda(x))) \leq \|\nabla \hat{M}^\lambda(x)\|.$$

Hence, a small Moreau envelope gradient means that x is close to a nearby point that is nearly stationary for ϕ .

This interpretation motivates the following notion of stationarity based on the Moreau envelope.

Definition 2.5 (Moreau stationary point) *For any $\varepsilon_M > 0$ and $\lambda > 0$, a point $x \in \operatorname{dom} \phi$ is called a $(\varepsilon_M; \lambda)$ -Moreau stationary point if $\|\nabla \hat{M}^\lambda(x)\| \leq \varepsilon_M$.*

It is worth noting that among the three notions of stationary points, the directional stationary one is the only one which does not depend on the weak convexity parameter m .

The next result, whose proof is given in Appendix B, describes the relationship between directional stationary points and Moreau stationary points.

Proposition 2.6 *Assume $\lambda > 0$ and ϕ is a m -weakly convex function. Then, the following statements hold:*

a) *if x is a $(\varepsilon_D, \delta_D)$ -directional stationary point then x is a $(\varepsilon_M; \lambda)$ -Moreau stationary point where*

$$\varepsilon_M = \left(m + \frac{1}{\lambda} \right) [(3 + 2\lambda m)\delta_D + 2\lambda\varepsilon_D];$$

b) *if x is a $(\varepsilon_M; \lambda)$ -Moreau stationary point then x is a $(\varepsilon_M, \varepsilon_M/(m + \lambda^{-1}))$ -directional stationary point.*

The following result, whose proof is given in Appendix B, provides an equivalent characterization of a directional stationary point in terms of the subdifferential of ϕ .

Proposition 2.7 *Assume that ϕ is an m -weakly convex function. Then, x is a $(\varepsilon_D, \delta_D)$ -directional stationary point if and only if there exists $\tilde{x} \in \operatorname{dom} \phi$ such that*

$$\|x - \tilde{x}\| \leq \delta_D, \quad \operatorname{dist}(0; \partial\phi(\tilde{x})) \leq \varepsilon_D.$$

The following result, whose proof is given in Appendix B, shows that a regularized stationary point is both a directional stationary point and a Moreau stationary point.

Proposition 2.8 *Assume that ϕ is an m -weakly convex function. If x is a $(\bar{\eta}, \bar{\varepsilon}; m)$ -regularized stationary point, then the following statements hold:*

a) *x is a $(\bar{\eta} + 2\sqrt{2m\bar{\varepsilon}}, \sqrt{2\bar{\varepsilon}/m})$ -directional stationary point;*

b) *x is a $(18\sqrt{2m\bar{\varepsilon}} + 4\bar{\eta}; 1/m)$ -Moreau stationary point.*

2.3 Related methods for weakly convex nonsmooth optimization

This subsection reviews two related approaches for weakly convex optimization. We first recall the deterministic version of the PS method of [5], and then briefly discuss the relation between PBF and bundle methods for DC programming.

More specifically, this subsection considers the PS method described below under the assumptions stated below (1) except that the constant L is assumed to be zero.

PS

Input: $\hat{x}_0 \in \text{dom } h$, a sequence of $\{\alpha_t\}_{t \geq 0} \subset \mathbb{R}_+$ and iteration count T .

Step: For $t = 0, 1, \dots, T$, compute

$$\hat{x}_{t+1} = \operatorname{argmin}_{u \in \mathbb{R}^n} \left\{ \ell_f(u; \hat{x}_t) + h(u) + \frac{1}{2\alpha_t} \|u - \hat{x}_t\|^2 \right\}. \quad (11)$$

The following result (see Theorem 3.4 in [5]) describes the rate of convergence of the PS method.

Proposition 2.9 *Assume (f, h) are functions satisfying conditions below (1) with $L = 0$ and assume that $\alpha_t \in (0, 1/\bar{m}]$ for every $t \geq 0$ for some $\bar{m} \in (m, 2m]$. Then, the iterate $\hat{x}_t$ of PS satisfies*

$$\frac{\sum_{t=0}^T \alpha_t \left\| \nabla \hat{M}^{1/(\bar{m}-m)}(\hat{x}_t) \right\|^2}{\sum_{t=0}^T \alpha_t} \leq \frac{\bar{m}}{\bar{m} - m} \cdot \frac{\left(\hat{M}^{1/(\bar{m}-m)}(\hat{x}_0) - \phi^* \right) + 2\bar{m}M^2 \sum_{t=0}^T \alpha_t^2}{\sum_{t=0}^T \alpha_t}. \quad (12)$$

As a consequence, for any given tolerance $\rho > 0$ and constant γ such that $\gamma \in (0, 1/(2m)]$, if the stepsizes α_t are chosen according to $\alpha_t = \gamma/\sqrt{T+1}$ for every $t \geq 0$ and the iteration count T satisfies

$$T \geq \frac{\left[\left(\hat{M}^{1/m}(\hat{x}_0) - \phi^* \right) + 4mM^2\gamma^2 \right]^2}{\gamma^2 \rho^4}, \quad (13)$$

then one of the iterates $\hat{x}_t \in \{\hat{x}_0, \dots, \hat{x}_T\}$ of PS satisfies $\left\| \nabla \hat{M}^{1/m}(\hat{x}_t) \right\| \leq \rho$.

We now make some remarks about PS and Proposition 2.9. First, PS always performs $T+1$ iterations and the second part of Proposition 2.9 gives a sufficient condition on T which guarantees that one of its iterates is a $(\rho; 1/m)$ -Moreau stationary point. An termination condition based on the magnitude of $\nabla \hat{M}^{1/m}(\hat{x}_t)$ is not doable since this quantity is generally expensive to compute. Second, a drawback of the estimate (13) on T is that it is generally not computable as it depends on ϕ^* . Third, the method of [5] actually outputs an iterate $\hat{x}_{t^*}$ where t^* is sampled from $\{0, \dots, T\}$ according to the probability mass function $\mathbb{P}(t^* = t) = \alpha_t / \sum_{i=0}^T \alpha_i$ for every $t = 0, \dots, T$. It turns out that $\mathbb{E}[\|\nabla \hat{M}^{1/m}(\hat{x}_{t^*})\|^2]$ is equal to the left-hand side of (12) and hence is bounded by the right-hand side of (12). The advantage of this approach is that, without performing any evaluation of $\|\nabla \hat{M}^{1/m}(\cdot)\|^2$, it returns an iterate $\hat{x}_{t^*}$ such that the expected value of $\|\nabla \hat{M}^{1/m}(\hat{x}_{t^*})\|^2$ is bounded by the right-hand side of (12). However, the authors are unaware of any technique for derandomizing this output strategy due to the fact that the function $\|\nabla \hat{M}^{1/m}(\cdot)\|^2$ is generally nonconvex and hard to compute.

Additionally, the third remark in the Concluding Remarks discusses how the alternative inexact proximal point method developed in [6] can estimate the gradient of the Moreau envelope evaluated at the center of each prox subproblem after it performs a pre-specified number of inner iterations to solve it approximately.

Finally, we conclude this subsection with a brief discussion of the relation between PBF and bundle methods for DC programming. In particular, weak convexity naturally induces a DC representation of (1). Indeed, letting

$$G(x) = f(x) + h(x) + \frac{m}{2} \|x\|^2, \quad H(x) = \frac{m}{2} \|x\|^2,$$

we have $\phi = G - H$, where both G and H are convex. Moreover, linearizing H at a point y gives

$$G(x) - H(y) - \langle \nabla H(y), x - y \rangle = \phi(x) + \frac{m}{2} \|x - y\|^2 = \phi_m(x; y).$$

Thus, PBF is closely related to proximal bundle methods for DC programming and can be viewed as exploiting the particular quadratic DC decomposition induced by weak convexity; see, e.g., [7, 17, 18]. In contrast to general DC bundle methods, PBF does not require handling a general nonsmooth second DC component: the latter is the quadratic function H , which is linearized exactly. Our analysis exploits this structure to derive explicit iteration-complexity bounds for obtaining regularized stationary points, together with their connections to Moreau stationarity.

3 Proximal Bundle Framework and Its Complexity

This section contains three subsections. The first one describes the main problem and the assumptions made on it. The second one describes the proximal bundle framework (PBF) and describes its main complexity result. The last one presents two special instances of PBF.

3.1 Problem description and main assumptions

The main problem of this paper is described in (1) where the functions f and h are assumed to satisfy:

- (A1) functions $f, h : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$ and a scalar $m > 0$ such that $h \in \overline{\text{Conv}}(\mathbb{R}^n)$, f is m -weakly convex and $\text{dom } h \subset \text{dom } f$; (Hence, all complexity results in this paper are stated for $m > 0$.)
- (A2) there exist constants $M, L \geq 0$ and a subgradient oracle, i.e., a function $f' : \text{dom } h \rightarrow \mathbb{R}^n$ satisfying $f'(u) \in \partial f(u)$ for every $u \in \text{dom } h$, such that

$$\|f'(u) - f'(v)\| \leq 2M + L\|u - v\| \quad \forall u, v \in \text{dom } h; \quad (14)$$

- (A3) $\phi^* := \inf\{\phi(u) : u \in \mathbb{R}^n\}$ is finite.

Let $\mathcal{C}(M, L)$ denote the class of functions f satisfying Assumption (A2). Even though $\mathcal{C}(M, L)$ depends on $\text{dom } h$, we have omitted this dependence from its notation. For any function $g \in \mathcal{C}(M, L)$ and $x \in \text{dom } h$, we denote the linearization of g at x by

$$\ell_g(\cdot; x) := g(x) + \langle g'(x), \cdot - x \rangle. \quad (15)$$

To analyze the bundle models used by PBF, we need to control the error of linearizing the convexified function $f_m(\cdot; z)$. The next lemma shows that $f_m(\cdot; z)$ remains in the hybrid class $\mathcal{C}(M, L + m)$, which leads to the linearization-error estimate in (17) below.

Lemma 3.1 *Assume that $f \in \mathcal{C}(M, L)$ for some $(M, L) \in \mathbb{R}_+^2$ and f is m -weakly convex on $\text{dom } h$. Then, for every $z \in \mathbb{R}^n$ and $\tilde{z} \in \text{dom } h$, we have*

$$f'(\tilde{z}) + m(\tilde{z} - z) \in \partial[f_m(\cdot; z)](\tilde{z}), \quad f_m(\cdot; z) \in \overline{\text{Conv}}(\mathbb{R}^n) \cap \mathcal{C}(M, L + m).$$

Proof: The convexity of $f_m(\cdot; z)$ follows from the m -weak convexity of f , while the first inclusion follows from Lemma A.1. Moreover, (14) implies that, for every $x, y \in \text{dom } h$,

$$\|f'_m(x; z) - f'_m(y; z)\| \leq \|f'(x) - f'(y)\| + m\|x - y\| \leq 2M + (L + m)\|x - y\|,$$

and hence $f_m(\cdot; z) \in \mathcal{C}(M, L + m)$. ■

In view of the first inclusion of Lemma 3.1, it follows that for any $z \in \mathbb{R}^n$ and $\tilde{z} \in \text{dom } h$, function $f_m(\cdot; z)$ admits the linearization given by

$$\ell_{f_m(\cdot; z)}(\cdot; \tilde{z}) := f_m(\tilde{z}; z) + \langle f'(\tilde{z}) + m(\tilde{z} - z), \cdot - \tilde{z} \rangle. \quad (16)$$

Moreover, Lemma 3.1 implies that $f_m(\cdot; z)$ is convex and belongs to $\mathcal{C}(M, L + m)$. Hence, the convexity of $f_m(\cdot; z)$ gives the lower bound below, while the $(M, L + m)$ -hybrid condition gives the upper bound:

$$0 \leq f_m(\cdot; z) - \ell_{f_m(\cdot; z)}(\cdot; \tilde{z}) \leq 2M\|\cdot - \tilde{z}\| + \frac{L + m}{2}\|\cdot - \tilde{z}\|^2 \quad \forall z \in \mathbb{R}^n, \tilde{z} \in \text{dom } h. \quad (17)$$

Indeed, the second inequality in (17) follows from a standard argument based on the fundamental theorem of calculus; see, for example, Lemma A.3 of [22]. Note that the above observations with $\tilde{z} = z$ implies that for every $z \in \text{dom } h$, we have

$$\ell_{f_m(\cdot; z)}(\cdot; z) = \ell_f(\cdot; z), \quad (18)$$

and

$$0 \leq f_m(\cdot; z) - \ell_f(\cdot; z) \leq 2M\|\cdot - z\| + \frac{L + m}{2}\|\cdot - z\|^2 \quad \forall z \in \text{dom } h. \quad (19)$$

3.2 The proximal bundle framework

This subsection describes PBF and its main complexity result. It also compares the complexity of PBF with that of the deterministic version of the PS method of [5] described in Subsection 2.3.

Before presenting PBF, we first provide a brief outline of the ideas behind it. PBF consists of solving a sequence of subproblems of the form

$$x_j = \underset{u \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ \Gamma_j(u) + \frac{1}{2\lambda}\|u - x^c\|^2 \right\} \quad (20)$$

where Γ_j is a relatively simple convex function minorizing the convexification $\phi_m(\cdot; x^c)$ of ϕ defined in (5) and x^c is a prox-center. As any classical proximal bundle approach, it updates the center x^c during some iterations (called the serious ones) and keeps the center the same in the other ones (called the null ones).

Following the description of PBF below, we also argue that it can be viewed as a specific implementation of an inexact proximal point method applied to (1). The parameters m , L and M that appear on its description are as in Assumptions (A1) and (A2). We start by introducing the following notion of shadow function.

Definition 3.2 *For any convex function φ , given $x^c \in \mathbb{R}^n$, $\lambda > 0$, and $\Gamma(\cdot) \in \mathcal{B}(\varphi)$, the function $\bar{\Gamma}(\cdot) \in \mathcal{B}(\varphi)$ is called a shadow of Γ for (2) if it satisfies*

$$\bar{\Gamma}(x) = \Gamma(x), \quad x = \underset{u \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ \bar{\Gamma}(u) + \frac{1}{2\lambda}\|u - x^c\|^2 \right\},$$

where x is the optimal solution of (2).

PBF

0. Let an initial point $\hat{x}_0 \in \text{dom } h$, a tolerance pair $(\bar{\eta}, \bar{\varepsilon}) \in \mathbb{R}_{++}^2$, and a prox stepsize $\lambda > 0$ be given, and set $y_0 = \hat{y}_0 = \hat{x}_0$, $j = 1$, $k = 1$, and

$$\delta = \min \left\{ \frac{\bar{\varepsilon}}{16}, \frac{\lambda \bar{\eta}^2}{64(m\lambda + 2)} \right\}, \quad \Gamma_1(\cdot) = \ell_f(\cdot; \hat{x}_0) + h(\cdot); \quad (21)$$

1. Compute the optimal solution x_j and optimal value θ_j of (20) with $x^c = \hat{y}_{k-1}$; if

$$\phi_m(x_j; \hat{y}_{k-1}) + \frac{1}{2\lambda}\|x_j - \hat{y}_{k-1}\|^2 < \phi_m(y_{j-1}; \hat{y}_{k-1}) + \frac{1}{2\lambda}\|y_{j-1} - \hat{y}_{k-1}\|^2, \quad (22)$$

then set $y_j = x_j$; else, set $y_j = y_{j-1}$;

2. Compute

$$w_j := \frac{1}{\lambda}(\hat{y}_{k-1} - x_j) - m(y_j - \hat{y}_{k-1}), \quad \delta_j := \delta + \frac{\lambda}{8(m\lambda + 1)}\|w_j\|^2; \quad (23)$$

3. If

$$t_j := \phi_m(y_j; \hat{y}_{k-1}) + \frac{1}{2\lambda}\|y_j - \hat{y}_{k-1}\|^2 - \theta_j > \delta_j, \quad (24)$$

then go to step 3a; else, go to step 3b;

3a) (null update) find $\Gamma_{j+1}(\cdot) \in \mathcal{B}(\phi_m(\cdot; \hat{y}_{k-1}))$ such that

$$\Gamma_{j+1}(\cdot) = \max \{ \ell_{f_m(\cdot; \hat{y}_{k-1})}(\cdot; x_j) + h(\cdot), \bar{\Gamma}_j(\cdot) \} \quad (25)$$

where $\bar{\Gamma}_j$ is a shadow of Γ_j for (20) with $x^c = \hat{y}_{k-1}$ and go to step 4;

3b) (serious update) set $\hat{x}_k = x_j$, $\hat{y}_k = y_j$, $\hat{\Gamma}_k(\cdot) = \Gamma_j(\cdot)$, $\hat{w}_k = w_j$, $\hat{\delta}_k = \delta_j$ and

$$\hat{v}_k = \frac{1}{\lambda}(\hat{y}_{k-1} - \hat{x}_k), \quad (26)$$

and compute

$$\hat{\varepsilon}_k = \phi_m(\hat{y}_k; \hat{y}_{k-1}) - \hat{\Gamma}_k(\hat{x}_k) - \langle \hat{v}_k, \hat{y}_k - \hat{x}_k \rangle \quad (27)$$

if $\|\hat{w}_k\| \leq \bar{\eta}$ and $\hat{\varepsilon}_k \leq \bar{\varepsilon}$, then stop; else find a bundle $\Gamma_{j+1}(\cdot) \in \mathcal{B}(\phi_m(\cdot; \hat{y}_k))$ such that

$$\Gamma_{j+1}(\cdot) \geq \ell_f(\cdot; \hat{y}_k) + h(\cdot), \quad (28)$$

where $\ell_f(\cdot; \cdot)$ is defined in (15), set $k \leftarrow k + 1$, and go to step 4;

4. set $j \leftarrow j + 1$, and go to step 1.

An iteration j such that $t_j \leq \delta_j$ is called a serious iteration; otherwise, j is called a null iteration. Let $j_1 \leq j_2 \leq \dots$ denote the sequence of all serious iterations and let $j_0 := 0$. Define the k -th cycle $\mathcal{C}_k$ to be the iterations j such that $j_{k-1} + 1 \leq j \leq j_k$, i.e.,

$$\mathcal{C}_k := \{i_k, \dots, j_k\}, \quad i_k := j_{k-1} + 1. \quad (29)$$

Hence, only the last iteration of a cycle (which can be the first one if $\mathcal{C}_k$ contains only one iteration) is serious.

We now make some basic remarks about PBF. First, PBF is referred to as a framework since it does not completely specify the details of how Γ_{j+1} in either (25) or (28) is updated. Second, it is shown in Lemma 4.6 that $\hat{w}_k \in \partial_{\hat{\varepsilon}_k} [\phi_m(\cdot; \hat{y}_k)](\hat{y}_k)$ for every $k \geq 1$. Hence, $\hat{y}_k$ is $(\bar{\eta}, \bar{\varepsilon}; m)$ -regularized stationary point of ϕ whenever the stopping criterion $\|\hat{w}_k\| \leq \bar{\eta}$ and $\hat{\varepsilon}_k \leq \bar{\varepsilon}$ is satisfied in step 3b. Third, in view of the definition of y_j in (22) and the above relation, it then follows that

$$y_j \in \text{Argmin} \left\{ \phi_m(x; \hat{y}_{k-1}) + \frac{1}{2\lambda}\|x - \hat{y}_{k-1}\|^2 : x \in \{\hat{y}_{k-1}, x_{i_k}, \dots, x_{j_k}\} \right\}. \quad (30)$$

In words, x_j is the trial point obtained by minimizing the current bundle model, whereas y_j is the best point encountered so far within the current cycle in terms of the regularized objective. Thus, x_j is used to refine the model, while y_j serves as the incumbent candidate for updating the prox-center. Fourth, two simple ways of choosing the bundle function Γ_{j+1} such that $\Gamma_{j+1}(\cdot) \in \mathcal{B}(\phi_m(\cdot; \hat{y}_k))$ and (28) holds are to set $\Gamma_{j+1} = \ell_f(\cdot; \hat{y}_k) + h$ or $\Gamma_{j+1} = \max\{\tilde{\Gamma}_{j+1}, \ell_f(\cdot; \hat{y}_k) + h\}$, where

$$\tilde{\Gamma}_{j+1} := \Gamma_j - m\langle \hat{y}_k - \hat{y}_{k-1}, \cdot - \hat{y}_k \rangle - \frac{m}{2}\|\hat{y}_k - \hat{y}_{k-1}\|^2.$$

Clearly, the first choice for Γ_{j+1} satisfies both conditions and the second one satisfies (28). Moreover, using the fact that $\Gamma_j \leq \phi_m(\cdot; \hat{y}_{k-1})$ and the definition of $\phi_m(\cdot; \hat{y}_{k-1})$ in (5), it is easy to see that $\tilde{\Gamma}_{j+1} \leq \phi_m(\cdot; \hat{y}_k)$,

and hence that $\Gamma_{j+1} \leq \phi_m(\cdot; \hat{y}_k)$. Fifth, PBF still requires knowledge of a parameter m as in Assumption (A1), and hence is not a completely universal method for finding a stationary point of (1). Finally, when f is convex and $h \equiv 0$, the bundle update in (25) is a special case of the one proposed in relations (4.7)-(4.9) of [4] (see also (2.2)-(2.4) of [8]).

Traditional proximal bundle methods often use serious/null decision policies based on sufficient descent conditions. A detailed discussion of these classical descent conditions and their relaxed variants can be found in Subsection 3.1 of [23]; see in particular the discussion following equation (3.7) therein.

PPM Interpretation: We now discuss how PBF can be interpreted as an inexact proximal point method for solving (1). First, the iterations within the k -th cycle can be interpreted as regularized cutting plane iterations applied to the prox subproblem

$$\hat{M}^\lambda(\hat{y}_{k-1}) = \min \left\{ \phi(x) + \frac{1}{2} \left(\frac{1}{\lambda} + m \right) \|x - \hat{y}_{k-1}\|^2 : x \in \mathbb{R}^n \right\}. \quad (31)$$

Second, the pair (y_j, Γ_j) found in the serious (i.e., the last) iteration j of the k -th cycle approximately solves (31) according to (24), in which case the point y_j as in step 1 of PBF becomes the center $\hat{y}_k$ for the next cycle. More specifically, we now discuss how the inexact criterion (24) can be interpreted in terms of the prox subproblem (31). Since $\Gamma_j \leq \phi_m(\cdot; \hat{y}_{k-1})$, it follows from the definitions of θ_j , $\hat{M}^\lambda(\hat{y}_{k-1})$, and $\phi_m(\cdot; \hat{y}_{k-1})$ in step 1 of PBF, (31), and (5), respectively, that $\theta_j \leq \hat{M}^\lambda(\hat{y}_{k-1})$, and hence that

$$\begin{aligned} 0 &\leq \phi(y_j) + \frac{1}{2} \left(\frac{1}{\lambda} + m \right) \|y_j - \hat{y}_{k-1}\|^2 - \hat{M}^\lambda(\hat{y}_{k-1}) \\ &\leq \phi(y_j) + \frac{1}{2} \left(\frac{1}{\lambda} + m \right) \|y_j - \hat{y}_{k-1}\|^2 - \theta_j \stackrel{(24)}{=} t_j. \end{aligned} \quad (32)$$

Hence, if j is an iteration for which (24) does not hold (i.e., a serious one), then it follows from (32) that y_j is a δ_j -solution of (31).

We now state the main complexity result for PBF. Its complexity bound depends on the quantity

$$\bar{\sigma} := \frac{[4\lambda(L+m)+1] \left(M^2 + \beta_2 \left\{ \beta_1 [\hat{M}^\lambda(\hat{x}_0) - \phi^* + 2\delta] + 4\zeta\lambda M^2 \right\} \right)}{8\lambda M^2} \quad (33)$$

where

$$\beta_1 := \left(m + \frac{2}{\zeta\lambda} \right) \left(m + \frac{1}{\lambda} \right)^{-1}, \quad \beta_2 := \left(\frac{L+m}{2} + 1 \right) \zeta^{-2} \left(\frac{1}{4\zeta\lambda} + \frac{m}{2} \right)^{-1}, \quad (34)$$

and

$$\zeta := \begin{cases} \frac{1}{2(L+m)\lambda} & \text{if } \lambda > \frac{1}{2(L+m)}; \\ 1 & \text{if } \lambda \leq \frac{1}{2(L+m)}. \end{cases} \quad (35)$$

Theorem 3.3 (Main Theorem) *PBF generates a $(\bar{\eta}, \bar{\varepsilon}; m)$ -regularized stationary point of ϕ within the iteration sequence $\{\hat{y}_k\}$ (see step 3b of PBF) in*

$$\left[4 + 4\lambda(L+m) + \frac{16\lambda M^2}{\delta} + [8\lambda(L+m) + 3] \log^+ \bar{\sigma} \right] \left[2 \left(\frac{1}{\bar{\varepsilon}} + \frac{8(m\lambda+1)}{\lambda\bar{\eta}^2} \right) (\hat{M}^\lambda(\hat{x}_0) - \phi^*) + 2 \right] \quad (36)$$

total iterations where $(\bar{\eta}, \bar{\varepsilon})$ is as in step 0 of PBF and δ is as in (21).

In terms of the tolerances $\bar{\eta}$ and $\bar{\varepsilon}$ only, it follows from Theorem 3.3 and the definition of δ in (21) that the iteration complexity of PBF to find a $(\bar{\eta}, \bar{\varepsilon}; m)$ -regularized stationary point of ϕ is $\mathcal{O}(\max\{\bar{\eta}^{-2}, \bar{\varepsilon}^{-1}\}^2)$. In particular, by choosing $\lambda = 1/[2(L+m)]$, we have $\zeta = 1$, β_1 and β_2 remain bounded as L increases, and

$$\delta = \Theta \left(\min \left\{ \bar{\varepsilon}, \frac{\bar{\eta}^2}{L+m} \right\} \right).$$

Hence, up to logarithmic factors, the complexity bound in Theorem 3.3 becomes

$$\tilde{\mathcal{O}} \left(\frac{M^2[\hat{M}^\lambda(\hat{x}_0) - \phi^*]}{L+m} \max \left\{ \frac{L+m}{\bar{\eta}^2}, \frac{1}{\bar{\varepsilon}} \right\}^2 \right), \quad (37)$$

which grows linearly with $L+m$ when the remaining problem parameters and tolerances are fixed.

The following result states the iteration complexity for PBF to find an approximate Moreau stationary point.

Corollary 3.4 *For a given tolerance $\rho > 0$, PBF with $\lambda = 1/m$ and $(\bar{\eta}, \bar{\varepsilon})$ given by*

$$\bar{\eta} = \frac{\rho}{8}, \quad \bar{\varepsilon} = \frac{\rho^2}{2592m},$$

applied to an instance of (1) satisfying (14) with $L = 0$ generates a $(\rho; 1/m)$ -Moreau stationary point of ϕ within the iteration sequence $\{\hat{y}_k\}$ in

$$\mathcal{O} \left(\frac{M^2 m [\hat{M}^\lambda(\hat{x}_0) - \phi^*]}{\rho^4} \right) \quad (38)$$

total iterations.

Proof: We first note that the choice of $(\bar{\eta}, \bar{\varepsilon})$ shows that PBF generates a $(\rho; 1/m)$ -Moreau stationary point of ϕ in view of Proposition 2.8(b). Therefore, the complexity bound (38) immediately follows from (37) with $L = 0$ and the choice of $(\bar{\eta}, \bar{\varepsilon})$ in the corollary. ■

We conclude this subsection by comparing PBF with the PS method of Subsection 2.3 in terms of both iteration complexity and prox stepsize selection. First, it is worth noting that the iteration complexity of the PS method in Subsection 2.3 is no better than that of PBF with $\lambda = 1/m$ and $L = 0$, namely, the bound in (38). Indeed, the PS complexity bound in (13), viewed as a function of γ , is minimized when

$$\gamma = \Theta \left(\frac{1}{M} \sqrt{\frac{\hat{M}^\lambda(\hat{x}_0) - \phi^*}{m}} \right),$$

in which case it matches the PBF bound (38) up to constant factors. However, this optimal choice of γ is generally not computable, since both $\hat{M}^\lambda(\hat{x}_0) - \phi^*$ and M may be unknown.

Second, it is also worth noting that PBF uses the constant prox stepsize $\lambda = 1/m$, whereas the PS method in Proposition 2.9 uses

$$\alpha_t = \frac{\gamma}{\sqrt{T+1}}, \quad \gamma \leq \frac{1}{2m}.$$

Thus,

$$\alpha_t \leq \frac{1}{2m\sqrt{T+1}},$$

and hence the prox stepsize used by PBF is at least $2\sqrt{T+1}$ times larger than that used by PS.

3.3 Special instances of PBF

Noting that the shadow $\bar{\Gamma}_j$ in step 3 of PBF is undetermined, PBF has the flexibility to choose $\bar{\Gamma}_j$. Next, we present two specific schemes of constructing $\bar{\Gamma}_j$, and hence two concrete ways to implement step 3a of PBF. The following two schemes are inspired by the proximal bundle methods for convex nonsmooth optimization [24, 29, 30].

- **2-cut:** This scheme sets $\bar{\Gamma}_j = A_j + h(\cdot)$ where $A_0(\cdot) = \ell_{f_m(\cdot; \hat{y}_0)}(\cdot; x_0)$ and A_j is recursively updated as follows

$$A_{j+1}(\cdot) = \tau A_j(\cdot) + (1 - \tau) \ell_{f_m(\cdot; \hat{y}_{k-1})}(\cdot; x_{j-1}) \quad (39)$$

for some $\tau \in [0, 1]$. In fact, (20) with $\Gamma_j(\cdot) = \max\{A_j(\cdot), \ell_{f_m(\cdot; \hat{y}_{k-1})}(\cdot; x_{j-1})\} + h(\cdot)$ and $x^c = \hat{y}_{k-1}$ is equivalent to

$$x_j = \operatorname{argmin}_{u \in \mathbb{R}^n, t \in \mathbb{R}} \left\{ t + h(u) + \frac{1}{2\lambda} \|u - \hat{y}_{k-1}\|^2 : A_j(u) \leq t, \ell_{f_m(\cdot; \hat{y}_{k-1})}(u; x_{j-1}) \leq t \right\}, \quad (40)$$

we denote by τ an optimal Lagrange multiplier associated with the constraint $A_j(u) \leq t$. Moreover, it is easy to check from the optimality conditions of (40) that $\bar{\Gamma}_j$ is a shadow of Γ_j for (20) with $x^c = \hat{y}_{k-1}$.

- **multi-cut:** This scheme sets $\bar{\Gamma}_j(\cdot) = \Gamma(\cdot; C(x_j))$ where

$$\Gamma(\cdot; C(x_j)) := \max\{\ell_{f_m(\cdot; \hat{y}_{k-1})}(\cdot; c) : c \in C(x_j)\} + h(\cdot), \quad (41)$$

and

$$C(x_j) := \{c \in C_j : \ell_{f_m(\cdot; \hat{y}_{k-1})}(x_j; c) + h(x_j) = \Gamma_j(x_j)\}. \quad (42)$$

It can be shown that $\bar{\Gamma}_j$ is a shadow of Γ_j for (20) with $x^c = \hat{y}_{k-1}$ following a similar argument as in the proof of Proposition D.2 of [24].

We conclude this subsection with a few implementation remarks. Aside from the first-order oracle evaluation, the main computational cost of each PBF iteration is solving the regularized cutting-plane subproblem (20). For the multi-cut scheme with ℓ cuts, (20) can be reformulated as a smooth concave maximization problem over the simplex. Each dual gradient evaluation requires one evaluation of the proximal mapping of h , and the dual problem can be solved using Nesterov's accelerated method. This results in $\mathcal{O}(\sqrt{\ell})$ evaluations of the proximal mapping of h . The 2-cut scheme admits a similar dual formulation, but with a one-dimensional dual variable. The resulting problem can be solved more simply by a bisection search, with each derivative evaluation again requiring one evaluation of the proximal mapping of h .

These observations highlight the practical tradeoff between the two schemes: the 2-cut scheme keeps the model size fixed and is attractive when storage and subproblem cost are primary concerns, whereas the multi-cut scheme may be preferable when maintaining a larger bundle is computationally affordable.

4 Proof of Theorem 3.3

This section establishes the iteration complexity for obtaining the regularized stationary point stated in Theorem 3.3. The basic idea is to bound the total number of iterations by analyzing the number of cycles and the number of iterations within each cycle, a standard proof strategy in the analysis of proximal bundle methods; see, for example, [11, 23, 24]. More specifically, this section contains three subsections. The first one derives a preliminary bound on the length of each cycle in terms of the tolerance δ . The second one bounds the number of cycles generated by PBF with a specific choice of δ until it obtains a $(\bar{\eta}, \bar{\varepsilon}; m)$ -regularized stationary point of ϕ . Finally, the third one derives the total iteration complexity of PBF with the aforementioned choice of δ until it obtains a $(\bar{\eta}, \bar{\varepsilon}; m)$ -regularized stationary point of ϕ .

4.1 Bounding the cardinalities of the cycles

This section establishes a preliminary upper bound on the cardinality of each cycle $\mathcal{C}_k$ defined in (29), and hence on the number of null iterations between two consecutive serious ones.

Throughout this section and the next one, we let

$$\hat{f}_k(\cdot) = f_m(\cdot; \hat{y}_{k-1}). \quad (43)$$

The first result below presents a few basic properties of the null iterations between two consecutive serious ones.

Lemma 4.1 *For every $j \in \mathcal{C}_k \setminus \{i_k\}$, there exists $\bar{\Gamma}_{j-1} \in \overline{\operatorname{Conv}}(\mathbb{R}^n)$ such that the following statements hold:*

a)

$$\max\{\bar{\Gamma}_{j-1}, \ell_{\hat{f}_k}(\cdot; x_{j-1}) + h\} \leq \Gamma_j, \quad (44)$$

$$\bar{\Gamma}_{j-1}(x_{j-1}) = \Gamma_{j-1}(x_{j-1}), \quad x_{j-1} = \operatorname{argmin}_{u \in \mathbb{R}^n} \left\{ \bar{\Gamma}_{j-1}(u) + \frac{1}{2\lambda} \|u - \hat{y}_{k-1}\|^2 \right\}; \quad (45)$$

b) for every $u \in \operatorname{dom} h$, there holds

$$\bar{\Gamma}_{j-1}(u) + \frac{1}{2\lambda} \|u - \hat{y}_{k-1}\|^2 \geq \theta_{j-1} + \frac{1}{2\lambda} \|u - x_{j-1}\|^2; \quad (46)$$

Proof: a) This statement follows directly from step 3a of PBF and Definition 3.2. In particular, the 2-cut and multi-cut schemes described in Subsection 3.3 provide two concrete constructions of such a shadow function $\bar{\Gamma}_{j-1}$.

b) Using the second identity in (45) and the fact that $\bar{\Gamma}_{j-1} \in \overline{\operatorname{Conv}}(\mathbb{R}^n)$, we have for every $u \in \operatorname{dom} h$,

$$\bar{\Gamma}_{j-1}(u) + \frac{1}{2\lambda} \|u - \hat{y}_{k-1}\|^2 \geq \bar{\Gamma}_{j-1}(x_{j-1}) + \frac{1}{2\lambda} \|x_{j-1} - \hat{y}_{k-1}\|^2 + \frac{1}{2\lambda} \|u - x_{j-1}\|^2.$$

The statement now follows from the above inequality, the first identity in (45), and the definition of θ_{j-1} in step 1 of PBF. $\blacksquare$

The next lemma presents some basic facts about $\{\theta_j\}$ and $\{t_j\}$.

Lemma 4.2 For every $j \in \mathcal{C}_k \setminus \{i_k\}$, the following inequalities hold:

$$\theta_j \geq \theta_{j-1} + \frac{1}{2\lambda} \|x_j - x_{j-1}\|^2, \quad (47)$$

$$t_j + \frac{1}{2\lambda} \|x_j - x_{j-1}\|^2 \leq t_{j-1}. \quad (48)$$

Proof: Using the definition of θ_j in step 1 of PBF, (44), and (46) with $u = x_j$, we have

$$\begin{aligned} \theta_j &= \Gamma_j(x_j) + \frac{1}{2\lambda} \|x_j - \hat{y}_{k-1}\|^2 \\ &\stackrel{(44)}{\geq} \bar{\Gamma}_{j-1}(x_j) + \frac{1}{2\lambda} \|x_j - \hat{y}_{k-1}\|^2 \stackrel{(46)}{\geq} \theta_{j-1} + \frac{1}{2\lambda} \|x_j - x_{j-1}\|^2 \end{aligned}$$

and thus (47) holds. Using the definition of t_j in (24), (30), and (47), we have

$$\begin{aligned} t_j &\stackrel{(24)}{=} \phi_m(y_j; \hat{y}_{k-1}) + \frac{1}{2\lambda} \|y_j - \hat{y}_{k-1}\|^2 - \theta_j \stackrel{(30)}{\leq} \phi_m(y_{j-1}; \hat{y}_{k-1}) + \frac{1}{2\lambda} \|y_{j-1} - \hat{y}_{k-1}\|^2 - \theta_j \\ &\stackrel{(47)}{\leq} \phi_m(y_{j-1}; \hat{y}_{k-1}) + \frac{1}{2\lambda} \|y_{j-1} - \hat{y}_{k-1}\|^2 - \theta_{j-1} - \frac{1}{2\lambda} \|x_j - x_{j-1}\|^2 \stackrel{(24)}{=} t_{j-1} - \frac{1}{2\lambda} \|x_j - x_{j-1}\|^2 \end{aligned}$$

and thus (48) holds. $\blacksquare$

The following technical result provides an important recursive formula for $\{t_j\}$.

Lemma 4.3 For every $j \in \mathcal{C}_k \setminus \{i_k\}$, we have

$$t_j \left(1 + \frac{t_j}{4\lambda[2M^2 + (L+m)t_j]} \right) \leq t_{j-1}.$$

Proof: The key idea is to combine the decrease estimate in Lemma 4.2 with the linearization error bound to obtain a quantitative lower bound on $\|x_j - x_{j-1}\|$, which in turn yields the stated recursion for t_j .

Using the definition of t_j in (24), (30), and the definition of θ_j in step 1 of PBF, we conclude that

$$\begin{aligned} t_j &\stackrel{(24)}{=} \phi_m(y_j; \hat{y}_{k-1}) + \frac{1}{2\lambda} \|y_j - \hat{y}_{k-1}\|^2 - \theta_j \stackrel{(30)}{\leq} \phi_m(x_j; \hat{y}_{k-1}) + \frac{1}{2\lambda} \|x_j - \hat{y}_{k-1}\|^2 - \theta_j \\ &= \phi_m(x_j; \hat{y}_{k-1}) - \Gamma_j(x_j). \end{aligned}$$

The above inequality, relations (44) and (17), and the fact that $\phi_m(x_j; \hat{y}_{k-1}) = f_m(x_j; \hat{y}_{k-1}) + h(x_j)$, imply that

$$\begin{aligned} 0 &\leq \phi_m(x_j; \hat{y}_{k-1}) - (\ell_{\hat{f}_k}(x_j; x_{j-1}) + h(x_j)) - t_j \\ &\stackrel{(44)}{=} f_m(x_j; \hat{y}_{k-1}) - \ell_{\hat{f}_k}(x_j; x_{j-1}) - t_j \\ &\stackrel{(17)}{\leq} \frac{L+m}{2} \|x_j - x_{j-1}\|^2 + 2M \|x_j - x_{j-1}\| - t_j, \end{aligned}$$

which, together with $t_j \geq 0$ due to (32), can be easily seen to imply that

$$\|x_j - x_{j-1}\| \geq \frac{-2M + \sqrt{4M^2 + 2(L+m)t_j}}{L+m} = \frac{2t_j}{2M + \sqrt{4M^2 + 2(L+m)t_j}} \geq \frac{t_j}{\sqrt{4M^2 + 2(L+m)t_j}}.$$

The statement now follows from (48) and the above inequality. $\blacksquare$

Proposition 4.4 *For every cycle index $k \geq 1$ generated by PBF, its size $|\mathcal{C}_k|$ is bounded by*

$$N_k := 4 + 4\lambda(L+m) + \frac{16\lambda M^2}{\delta} + [8\lambda(L+m) + 3] \log^+ \sigma_k \quad (49)$$

where

$$\sigma_k := \frac{t_{i_k}}{\delta + 8\lambda M^2 / [4\lambda(L+m) + 1]}. \quad (50)$$

Proof: Suppose for contradiction that $|\mathcal{C}_k| > N_k$ and define $J = |\mathcal{C}_k| - 2$. Since the k th cycle does not stop at iteration $j + i_k$ for any $j = 0, \dots, J$, we have that

$$q_j := t_{j+i_k} > \delta \quad \forall j = 0, \dots, J. \quad (51)$$

Observe that Lemma 4.3 and (51) imply that assumptions (90) and (91) of Lemma C.1 is satisfied with $a = 8\lambda M^2$, $b = 4\lambda(L+m)$, and $\{q_j\}_{j=0}^J$. Hence, the conclusion of Lemma C.1 and the definition of N_k in (49) imply that $J \leq N_k - 2$ which contradicts with the assumption $|\mathcal{C}_k| = J + 2 > N_k$. $\blacksquare$

4.2 Bounding total number of cycles

The goal of this subsection is to establish a bound on the total number of cycles generated by PBF.

The following technical result provides the main properties of the sequences $\{\hat{x}_k\}$, $\{\hat{y}_k\}$ and $\{\hat{\Gamma}_k\}$ generated by PBF. Recall that $\hat{x}_k$, $\hat{y}_k$, and $\hat{\Gamma}_k$ denote the last x_j , y_j , and Γ_j , respectively, generated within the cycle $\mathcal{C}_k$, i.e., the ones with $j = j_k$.

Lemma 4.5 *The following statements hold for every $k \geq 1$:*

a) $\hat{x}_k$ is the optimal solution of

$$\min_{u \in \mathbb{R}^n} \left\{ \hat{\Gamma}_k(u) + \frac{1}{2\lambda} \|u - \hat{y}_{k-1}\|^2 \right\}; \quad (52)$$

hence, if $\hat{\theta}_k$ denotes the optimal value of (52), then

$$\hat{\theta}_k = \hat{\Gamma}_k(\hat{x}_k) + \frac{1}{2\lambda} \|\hat{x}_k - \hat{y}_{k-1}\|^2; \quad (53)$$

b) there hold

$$\hat{\Gamma}_k(\cdot) \in \overline{\text{Conv}}(\mathbb{R}^n), \quad \hat{\Gamma}_k(\cdot) \leq \phi_m(\cdot; \hat{y}_{k-1}) \quad (54)$$

and

$$\phi_m(\hat{y}_k; \hat{y}_{k-1}) + \frac{1}{2\lambda} \|\hat{y}_k - \hat{y}_{k-1}\|^2 - \hat{\theta}_k \leq \hat{\delta}_k. \quad (55)$$

Proof: a) The definition of $\hat{x}_k$ in step 3b of PBF and step 1 of PBF imply that $\hat{x}_k$ is an optimal solution of (52).

b) If $|C_k| = 1$, i.e., j_k is also the first iteration of cycle C_k , then (54) follows from the fact that $\hat{\Gamma}_k(\cdot) = \Gamma_{i_k}(\cdot) \in \mathcal{B}(\phi_m(\cdot; \hat{y}_{k-1}))$. If $|C_k| > 1$, then $\hat{\Gamma}_k = \Gamma_{j_k}$ and hence satisfies (54) in view of step 3a of PBF. Moreover, (55) follows from the logic of the prox-center update rule in step 3 of PBF (see (24)). ■

The following result presents an important inclusion involving $(\hat{y}_k, \hat{w}_k, \hat{\varepsilon}_k)$ which implies that $\hat{y}_k$ is a $(\|\hat{w}_k\|, \hat{\varepsilon}_k; m)$ -regularized stationary point of ϕ .

Lemma 4.6 *For every $k \geq 1$, the quantities $\hat{x}_k$, $\hat{y}_k$, $\hat{v}_k$, $\hat{w}_k$ and $\hat{\varepsilon}_k$ as in step 3b of PBF satisfy*

$$\hat{\varepsilon}_k \geq 0, \quad \hat{v}_k \in \partial_{\hat{\varepsilon}_k}[\phi_m(\cdot; \hat{y}_{k-1})](\hat{y}_k), \quad (56)$$

$$\hat{w}_k \in \partial_{\hat{\varepsilon}_k}[\phi_m(\cdot; \hat{y}_k)](\hat{y}_k). \quad (57)$$

Proof: Since $\hat{x}_k$ is an optimal solution of (52) in view of Lemma 4.5(a), using the optimality condition for (52), the fact that $\hat{\Gamma}_k \in \overline{\text{Conv}}(\mathbb{R}^n)$, and the definition of $\hat{v}_k$ in (26), we have $\hat{v}_k \in \partial \hat{\Gamma}_k(\hat{x}_k)$. This conclusion, (54), the definition of subdifferential in (3), and the definition of $\hat{\varepsilon}_k$ in (27), then imply that for every $u \in \text{dom } h$,

$$\phi_m(u; \hat{y}_{k-1}) \geq \hat{\Gamma}_k(u) \geq \hat{\Gamma}_k(\hat{x}_k) + \langle \hat{v}_k, u - \hat{x}_k \rangle = \phi_m(\hat{y}_k; \hat{y}_{k-1}) + \langle \hat{v}_k, u - \hat{y}_k \rangle - \hat{\varepsilon}_k$$

and hence that the inclusion in (56) holds. The inequality in (56) follows from the above inequality with $u = \hat{y}_k$. Moreover, the definition of $\hat{w}_k$ (see step 3b of PBF), the inclusion in (56), and Lemma A.1 imply that

$$\hat{w}_k \in \partial_{\hat{\varepsilon}_k}[\phi_m(\cdot; \hat{y}_{k-1})](\hat{y}_k) - m(\hat{y}_k - \hat{y}_{k-1}) = \partial_{\hat{\varepsilon}_k}[\phi_m(\cdot; \hat{y}_k)](\hat{y}_k),$$

and hence that (57) holds. ■

It follows from (57) that $\hat{y}_k$ is a $(\|\hat{w}_k\|, \hat{\varepsilon}_k; m)$ -regularized stationary point of ϕ where the pair $(\hat{w}_k, \hat{\varepsilon}_k)$ can be easily computed according to step 3b of PBF. Our remaining effort from now on will be to analyze the number of iterations it takes to obtain an index k such that $\|\hat{w}_k\| \leq \bar{\eta}$ and $\hat{\varepsilon}_k \leq \bar{\varepsilon}$, and hence a $(\bar{\eta}, \bar{\varepsilon}; m)$ -regularized stationary point $\hat{y}_k$ of ϕ .

The purpose of the following three results is to establish a recursive formula (see Lemma 4.9 below). Lemma 4.7 and Corollary 4.8 are technical results that are needed to prove Lemma 4.9. The next lemma bounds the stationarity residuals at a serious iteration in terms of the displacement between two consecutive prox-centers. These estimates will allow us to control the residuals using the cumulative descent of the prox-centers.

Lemma 4.7 *For every $k \geq 1$, the quantities $\hat{x}_k$, $\hat{y}_k$, $\hat{w}_k$ and $\hat{\varepsilon}_k$, as in step 3b of PBF, satisfy*

$$\hat{\varepsilon}_k + \frac{1}{2\lambda} \|\hat{y}_k - \hat{x}_k\|^2 \leq 2\delta + \frac{1 + m\lambda}{2\lambda} \|\hat{y}_k - \hat{y}_{k-1}\|^2 \quad (58)$$

and

$$\|\hat{w}_k\|^2 \leq \frac{8\delta}{\lambda} + 4 \left(m + \frac{1}{\lambda} \right)^2 \|\hat{y}_k - \hat{y}_{k-1}\|^2 \quad (59)$$

where λ and δ are as in step 0 of PBF.

Proof: Using both statements (a) and (b) of Lemma 4.5, the definitions of $\hat{\varepsilon}_k$ and $\hat{v}_k$ in (27) and (26), respectively, we conclude that

$$\begin{aligned} \hat{\varepsilon}_k &= \phi_m(\hat{y}_k; \hat{y}_{k-1}) - \hat{\Gamma}_k(\hat{x}_k) - \langle \hat{v}_k, \hat{y}_k - \hat{x}_k \rangle \\ &\stackrel{(55)}{\leq} \hat{\delta}_k - \frac{1}{2\lambda} \|\hat{y}_k - \hat{y}_{k-1}\|^2 + \hat{\theta}_k - \hat{\Gamma}_k(\hat{x}_k) - \langle \hat{v}_k, \hat{y}_k - \hat{x}_k \rangle \\ &\stackrel{(53)}{=} \hat{\delta}_k - \frac{1}{2\lambda} \|\hat{y}_k - \hat{y}_{k-1}\|^2 + \frac{1}{2\lambda} \|\hat{x}_k - \hat{y}_{k-1}\|^2 - \langle \hat{v}_k, \hat{y}_k - \hat{x}_k \rangle \\ &\stackrel{(26)}{=} \hat{\delta}_k - \frac{1}{2\lambda} \|\hat{y}_k - \hat{y}_{k-1}\|^2 + \frac{1}{2\lambda} \|\hat{x}_k - \hat{y}_{k-1}\|^2 + \frac{1}{\lambda} \langle \hat{x}_k - \hat{y}_{k-1}, \hat{y}_k - \hat{x}_k \rangle \\ &= \hat{\delta}_k - \frac{1}{2\lambda} \|\hat{y}_k - \hat{x}_k\|^2. \end{aligned} \quad (60)$$

The above inequality together with the definition of $\hat{w}_k$ and $\hat{\delta}_k$ in step 3b of PBF implies that

$$\begin{aligned}
\|\hat{w}_k\|^2 &\stackrel{(23)}{\leq} \frac{2}{\lambda^2} \|\hat{y}_k - \hat{x}_k\|^2 + 2 \left(m + \frac{1}{\lambda}\right)^2 \|\hat{y}_k - \hat{y}_{k-1}\|^2 \\
&\stackrel{(60)}{\leq} \frac{4\hat{\delta}_k}{\lambda} + 2 \left(m + \frac{1}{\lambda}\right)^2 \|\hat{y}_k - \hat{y}_{k-1}\|^2 \\
&\stackrel{(23)}{=} \frac{4\delta}{\lambda} + \frac{1}{2(m\lambda + 1)} \|\hat{w}_k\|^2 + 2 \left(m + \frac{1}{\lambda}\right)^2 \|\hat{y}_k - \hat{y}_{k-1}\|^2
\end{aligned}$$

where the first inequality is due to the relation $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$. Inequality (59) now follows from the above inequality and simple algebraic manipulation. Using (59), (60) and the definition of $\hat{\delta}_k$ in step 3b of PBF, we have

$$\begin{aligned}
\hat{\varepsilon}_k + \frac{1}{2\lambda} \|\hat{y}_k - \hat{x}_k\|^2 &\stackrel{(23),(60)}{\leq} \delta + \frac{\lambda}{8(m\lambda + 1)} \|\hat{w}_k\|^2 \\
&\stackrel{(59)}{\leq} \delta + \frac{\lambda}{8(m\lambda + 1)} \left(\frac{8\delta}{\lambda} + 4 \left(m + \frac{1}{\lambda}\right)^2 \|\hat{y}_k - \hat{y}_{k-1}\|^2 \right) \\
&\leq 2\delta + \frac{1 + m\lambda}{2\lambda} \|\hat{y}_k - \hat{y}_{k-1}\|^2,
\end{aligned}$$

and thus (58) holds. ■

Corollary 4.8 *If for some $k \geq 1$, we have $\hat{y}_k = \hat{y}_{k-1}$, then the following statements hold:*

- a) $y_j = \hat{y}_{k-1}$ for all $j \in \mathcal{C}_k$;
- b) $\|\hat{w}_k\| \leq \bar{\eta}/4$ and $\hat{\varepsilon}_k \leq \bar{\varepsilon}/8$, and hence $\hat{y}_k$ is a $(\bar{\eta}/4, \bar{\varepsilon}/8; m)$ -regularized stationary point of problem (1).

Proof: a) Denote $\Psi(x) = \phi_m(x; \hat{y}_{k-1}) + \frac{1}{2\lambda} \|x - \hat{y}_{k-1}\|^2$. Assume that for contradiction that j is the first index in $\mathcal{C}_k$ such that $y_j \neq \hat{y}_{k-1}$, and hence that $y_j \neq y_{j-1} = \hat{y}_{k-1}$. Because of the test (22) in step 1 of PBF, it follows that $y_j = x_j$ and $\Psi(y_{j-1}) > \Psi(x_j)$, and hence that

$$\Psi(\hat{y}_{k-1}) = \Psi(y_{j-1}) > \Psi(x_j) = \Psi(y_j) \stackrel{(30)}{\geq} \Psi(\hat{y}_k) = \Psi(\hat{y}_{k-1}),$$

which yields the desired contradiction.

b) Since $\hat{y}_k = \hat{y}_{k-1}$, (59) implies

$$\|\hat{w}_k\|^2 \leq \frac{8\delta}{\lambda} \leq \frac{\bar{\eta}^2}{16} \tag{61}$$

where the last inequality is due the definition of δ in step 0 of PBF. Observe that (58) implies that

$$\hat{\varepsilon}_k \leq 2\delta \leq \frac{\bar{\varepsilon}}{8} \tag{62}$$

where the last inequality is due the definition of δ in step 0 of PBF. Finally, (57), (61), (62) and Definition 2.4 then imply that $\hat{y}_k$ is a $(\bar{\eta}/4, \bar{\varepsilon}/8; m)$ -regularized stationary point of problem (1). ■

The above corollary indicates that whenever PBF has a repeated cycle which yields $\hat{y}_k = \hat{y}_{k-1}$ for some $k \geq 1$, the method will terminate and return a $(\bar{\eta}/4, \bar{\varepsilon}/8; m)$ -regularized stationary point of problem (1).

Lemma 4.9 *For every $k \geq 1$, the following statements are true:*

- a) there holds

$$\phi(\hat{y}_k) + \frac{1}{2} \left(\frac{1}{\lambda} + m \right) \|\hat{y}_k - \hat{y}_{k-1}\|^2 \leq \phi(\hat{y}_{k-1}) \tag{63}$$

where $\{\hat{y}_k\}$ is as in step 3b of PBF;

b) for every $k \geq 2$, there holds

$$\frac{1}{2} \left(\frac{1}{\lambda} + m \right) \sum_{l=2}^k \|\hat{y}_l - \hat{y}_{l-1}\|^2 \leq \hat{M}^\lambda(\hat{x}_0) - \phi(\hat{y}_k) + 2\delta. \quad (64)$$

Proof: a) It follows from (5) and (30) that

$$\phi(\hat{y}_k) + \frac{1}{2} \left(\frac{1}{\lambda} + m \right) \|\hat{y}_k - \hat{y}_{k-1}\|^2 \leq \phi(\hat{y}_{k-1}) + \frac{1}{2} \left(\frac{1}{\lambda} + m \right) \|\hat{y}_{k-1} - \hat{y}_{k-1}\|^2 = \phi(\hat{y}_{k-1}),$$

which proves (63).

b) Summing (63) from $k = 2$ to k , we have

$$\phi(\hat{y}_k) + \frac{1}{2} \left(\frac{1}{\lambda} + m \right) \sum_{l=2}^k \|\hat{y}_l - \hat{y}_{l-1}\|^2 \leq \phi(\hat{y}_1). \quad (65)$$

Using (55) with $k = 1$, the definitions of $\hat{M}^\lambda(\hat{x}_0)$ and θ_j in (8) and step 1 of PBF, respectively, and the fact that $\hat{\Gamma}_1(\cdot) \leq \phi_m(\cdot; \hat{x}_0)$, we can conclude that

$$\begin{aligned} \phi(\hat{y}_1) + \frac{1}{2} \left(\frac{1}{\lambda} + m \right) \|\hat{y}_1 - \hat{y}_0\|^2 &\stackrel{(55)}{\leq} \hat{\theta}_1 + \hat{\delta}_1 \stackrel{(8)}{\leq} \hat{M}^\lambda(\hat{x}_0) + \hat{\delta}_1 \\ &\stackrel{(23)}{=} \hat{M}^\lambda(\hat{x}_0) + \delta + \frac{\lambda}{8(m\lambda + 1)} \|\hat{w}_1\|^2, \end{aligned}$$

where the identity is due to (23). It thus follows from (59) that

$$\phi(\hat{y}_1) + \frac{1}{2} \left(\frac{1}{\lambda} + m \right) \|\hat{y}_1 - \hat{y}_0\|^2 \stackrel{(59)}{\leq} \hat{M}^\lambda(\hat{x}_0) + 2\delta + \frac{1}{2} \left(\frac{1}{\lambda} + m \right) \|\hat{y}_1 - \hat{y}_0\|^2,$$

which together with (65) implies that (64). ■

We are now ready to bound the total number of cycles generated by PBF.

Proposition 4.10 For a given tolerance pair $(\bar{\eta}, \bar{\varepsilon}) \in \mathbb{R}_{++}^2$, define

$$K = K(\bar{\eta}, \bar{\varepsilon}) := \left\lceil 2 \left(\frac{1}{\bar{\varepsilon}} + \frac{8(m\lambda + 1)}{\lambda \bar{\eta}^2} \right) (\hat{M}^\lambda(\hat{x}_0) - \phi^*) + 1 \right\rceil \quad (66)$$

where $\hat{M}^\lambda(\cdot)$ is as in (8). Then, PBF with δ as in (21) generates an iteration index $k \leq K(\bar{\eta}, \bar{\varepsilon})$ such that

$$\|\hat{w}_k\| \leq \bar{\eta}, \quad \hat{\varepsilon}_k \leq \bar{\varepsilon}. \quad (67)$$

As a consequence, $\hat{y}_k$ is a $(\bar{\eta}, \bar{\varepsilon}; m)$ -regularized stationary point of problem (1).

Proof: The proof is based on contradiction. If PBF does not terminate within the first K cycles, then each cycle has a sufficiently large normalized stationarity residual. Summing these residuals and using the cumulative prox-center displacement bound in Lemma 4.9 leads to a contradiction.

Assuming by contradiction that PBF does not stop at the K -th cycle. Then, the stopping criterion in step 3b of PBF implies that

$$\max \left\{ \frac{\|\hat{w}_k\|^2}{\bar{\eta}^2}, \frac{\hat{\varepsilon}_k}{\bar{\varepsilon}} \right\} > 1 \quad \forall k = 1, \dots, K$$

and thus

$$\frac{\|\hat{w}_k\|^2}{\bar{\eta}^2} + \frac{\hat{\varepsilon}_k}{\bar{\varepsilon}} > 1 \quad \forall k = 1, \dots, K. \quad (68)$$

Observe that the definitions of δ and K in (21) and (66), respectively, imply that

$$\left(\frac{4}{\bar{\varepsilon}} + \frac{16(m\lambda + 2)}{\lambda\bar{\eta}^2}\right)\delta \leq \frac{1}{2}, \quad \frac{1}{K-1} \left(\frac{1}{\bar{\varepsilon}} + \frac{8(m\lambda + 1)}{\lambda\bar{\eta}^2}\right) (\hat{M}^\lambda(\hat{x}_0) - \phi^*) \leq \frac{1}{2}. \quad (69)$$

Moreover, using (58), (59) and (68), we have

$$\begin{aligned} 1 &\stackrel{(68)}{<} \frac{\|\hat{w}_k\|^2}{\bar{\eta}^2} + \frac{\hat{\varepsilon}_k}{\bar{\varepsilon}} \leq \frac{2\delta + (1+m\lambda)\|\hat{y}_k - \hat{y}_{k-1}\|^2/(2\lambda)}{\bar{\varepsilon}} + \frac{8\delta + 4(m\lambda + 1)^2\|\hat{y}_k - \hat{y}_{k-1}\|^2/\lambda}{\lambda\bar{\eta}^2} \\ &= \frac{2\delta}{\bar{\varepsilon}} + \frac{8\delta}{\lambda\bar{\eta}^2} + \left(\frac{1+m\lambda}{2\lambda\bar{\varepsilon}} + \frac{4(m\lambda + 1)^2}{\lambda^2\bar{\eta}^2}\right) \|\hat{y}_k - \hat{y}_{k-1}\|^2. \end{aligned}$$

Summing the above inequality from $k = 2$ to K , dividing the resulting inequality by $K - 1$, and using (64), we have

$$\begin{aligned} 1 &< \frac{2\delta}{\bar{\varepsilon}} + \frac{8\delta}{\lambda\bar{\eta}^2} + \left(\frac{1+m\lambda}{2\lambda\bar{\varepsilon}} + \frac{4(m\lambda + 1)^2}{\lambda^2\bar{\eta}^2}\right) \frac{1}{K-1} \sum_{k=2}^K \|\hat{y}_k - \hat{y}_{k-1}\|^2 \\ &\stackrel{(64)}{\leq} \frac{2\delta}{\bar{\varepsilon}} + \frac{8\delta}{\lambda\bar{\eta}^2} + \left(\frac{1+m\lambda}{2\lambda\bar{\varepsilon}} + \frac{4(m\lambda + 1)^2}{\lambda^2\bar{\eta}^2}\right) \frac{2\lambda}{(\lambda m + 1)(K-1)} (\hat{M}^\lambda(\hat{x}_0) - \phi(\hat{y}_K) + 2\delta) \\ &\leq \left(\frac{4}{\bar{\varepsilon}} + \frac{16(m\lambda + 2)}{\lambda\bar{\eta}^2}\right)\delta + \frac{1}{K-1} \left(\frac{1}{\bar{\varepsilon}} + \frac{8(m\lambda + 1)}{\lambda\bar{\eta}^2}\right) (\hat{M}^\lambda(\hat{x}_0) - \phi^*) \leq 1 \end{aligned}$$

where the second last inequality is due to the fact that $\phi(\hat{y}_K) \geq \phi^*$ and the last inequality is due to (69). Hence, the above inequality gives the desired contradiction and thus the statement is proved. $\blacksquare$

Before ending this subsection, we observe that the quantity $\hat{M}^\lambda(\hat{x}_0) - \phi^*$ in (66) can be majorized by the more standard initial primal gap $\phi(\hat{x}_0) - \phi^*$ due to the definition of $\hat{M}^\lambda(\cdot)$ in (8).

4.3 Proof of Theorem 3.3

Recall that Proposition 4.10 bounds the total number of cycles while Proposition 4.4 provides a bound on the cardinality of every cycle $\mathcal{C}_k$ in terms of σ_k . The following result refines the latter result by providing a uniform bound on σ_k .

Lemma 4.11 *Define*

$$\bar{t} := M^2 + \beta_2 \left(\beta_1 [\hat{M}^\lambda(\hat{x}_0) - \phi^* + 2\delta] + 4\zeta\lambda M^2 \right), \quad (70)$$

where δ is as in (21), ζ is as in (35), and β_1 and β_2 are as in (34). Then, the following statements hold for every $k \geq 1$:

- a) $t_{i_k} \leq \bar{t}$ where t_{i_k} is as in (24);
- b) $\sigma_k \leq \bar{\sigma}$ where σ_k and $\bar{\sigma}$ are as in (50) and (33), respectively.

Proof: a) It follows from (24) with $j = i_k$ and the definition of θ_j in step 1 of PBF that

$$\begin{aligned} t_{i_k} &\stackrel{(24)}{=} \phi_m(y_{i_k}; \hat{y}_{k-1}) + \frac{1}{2\lambda} \|y_{i_k} - \hat{y}_{k-1}\|^2 - \theta_{i_k} \\ &= \phi_m(y_{i_k}; \hat{y}_{k-1}) + \frac{1}{2\lambda} \|y_{i_k} - \hat{y}_{k-1}\|^2 - \Gamma_{i_k}(x_{i_k}) - \frac{1}{2\lambda} \|x_{i_k} - \hat{y}_{k-1}\|^2. \end{aligned} \quad (71)$$

Noting from the observation (30) that

$$\phi_m(y_j; \hat{y}_{k-1}) + \frac{1}{2\lambda} \|y_j - \hat{y}_{k-1}\|^2 \leq \phi_m(x_j; \hat{y}_{k-1}) + \frac{1}{2\lambda} \|x_j - \hat{y}_{k-1}\|^2.$$

Hence, the above inequality with $j = i_k$ and (71) imply that

$$t_{i_k} \leq \phi_m(x_{i_k}; \hat{y}_{k-1}) - \Gamma_{i_k}(x_{i_k}) \stackrel{(21),(28)}{\leq} f_m(x_{i_k}; \hat{y}_{k-1}) - \ell_f(x_{i_k}; \hat{y}_{k-1}),$$

where the second inequality follows from (21), (28) with $j + 1$ replaced by $j = i_k$, and the definition of ϕ_m in (5). Moreover, applying (19) with $z = \hat{y}_{k-1}$, we have

$$f_m(x_{i_k}; \hat{y}_{k-1}) - \ell_f(x_{i_k}; \hat{y}_{k-1}) \leq 2M\|x_{i_k} - \hat{y}_{k-1}\| + \frac{L+m}{2}\|x_{i_k} - \hat{y}_{k-1}\|^2.$$

Combining the above two inequalities, we have

$$t_{i_k} \leq 2M\|x_{i_k} - \hat{y}_{k-1}\| + \frac{L+m}{2}\|x_{i_k} - \hat{y}_{k-1}\|^2 \leq M^2 + \left(\frac{L+m}{2} + 1\right)\|x_{i_k} - \hat{y}_{k-1}\|^2, \quad (72)$$

where the last inequality is due to the fact that $2ab \leq a^2 + b^2$ with $a = M$ and $b = \|x_{i_k} - \hat{y}_{k-1}\|$.

We will now bound $\|x_{i_k} - \hat{y}_{k-1}\|^2$. It follows from (64) that

$$\frac{1}{2} \left(m + \frac{1}{\lambda}\right) \|\hat{y}_k - \hat{y}_{k-1}\|^2 \leq \hat{M}^\lambda(\hat{x}_0) - \phi(\hat{y}_k) + 2\delta.$$

which, in view of the definition of β_1 in (34), is equivalent to

$$\phi(\hat{y}_k) - \phi^* + \frac{1}{2\beta_1} \left(m + \frac{2}{\zeta\lambda}\right) \|\hat{y}_k - \hat{y}_{k-1}\|^2 \leq \hat{M}^\lambda(\hat{x}_0) - \phi^* + 2\delta.$$

It is easy to verify that $\beta_1 \geq 1$ from its definition in (34). Clearly, this observation, the fact that $\phi(\hat{y}_k) \geq \phi^*$, and the above inequality imply that

$$\phi(\hat{y}_k) - \phi^* + \frac{1}{2} \left(m + \frac{2}{\zeta\lambda}\right) \|\hat{y}_k - \hat{y}_{k-1}\|^2 \leq \beta_1 [\hat{M}^\lambda(\hat{x}_0) - \phi^* + 2\delta]. \quad (73)$$

Using this inequality and Lemma D.3 with $(\Gamma, z_0, u) = (\Gamma_{i_k}, \hat{y}_{k-1}, \hat{y}_k)$, we obtain

$$\begin{aligned} \zeta^2 \left(\frac{1}{4\zeta\lambda} + \frac{m}{2}\right) \|x_{i_k} - \hat{y}_{k-1}\|^2 &\stackrel{(98)}{\leq} \phi(\hat{y}_k) - \phi^* + \frac{1}{2} \left(m + \frac{2}{\zeta\lambda}\right) \|\hat{y}_k - \hat{y}_{k-1}\|^2 + 4\zeta\lambda M^2 \\ &\stackrel{(73)}{\leq} \beta_1 [\hat{M}^\lambda(\hat{x}_0) - \phi^* + 2\delta] + 4\zeta\lambda M^2. \end{aligned}$$

The statement now follows by combining (72) and the above inequality, and using the definitions of β_2 and $\bar{\epsilon}$ in (34) and (70), respectively.

b) This statement follows from a) and the definitions of σ_k and $\bar{\sigma}$ in (50) and (33), respectively. $\blacksquare$

We end this section by noting that the proof of Theorem 3.3 now follows immediately from Propositions 4.4 and 4.10, and Lemma 4.11(b).

5 Computational Results

This section reports computational results comparing the PBF method of Subsection 3.2 with the PS method of Subsection 2.3 and the Redistributed Proximal Bundle (RPB) method of [15]. It contains three subsections. The first presents computational results for the phase retrieval problem, the second for the blind deconvolution problem, and the third for the constrained MAXQUAD problem.

Both phase retrieval and blind deconvolution are special cases of (1) where $h \equiv 0$ and $f(\cdot)$ has the form

$$f(x) = g(c(x)) \quad (74)$$

for some convex and L -Lipschitz function g and smooth map c with β -Lipschitz Jacobian. It follows from Lemma 4.2 in [9] that such a function $f(\cdot)$ is $L\beta$ -weakly convex and hence we set $m = L\beta$. Recall that the

PS method terminates based on Moreau stationary point (see Definition 2.5) which is hard to verify in view of the remark below Proposition 2.9. On the other hand, it is still an open problem for the PS method to establish the convergence to find a regularized stationary point (see Definition 2.4). Hence for the sake of our numerical experiments, we adopt the termination criterion

$$\phi(x_k) - \phi_* \leq \tilde{\varepsilon}[\phi(x_0) - \phi_*] \quad (75)$$

that requires the knowledge of ϕ_* where either $\tilde{\varepsilon} = 10^{-3}$ or 10^{-4} .

Now we provide details of the algorithms used in the following three subsections. We implement the constant stepsize version of PS method, i.e., $\alpha_t = \alpha$ for every $t \geq 0$. In our experiment, we choose four values of α , namely, $\alpha = 1/(32m)$, $1/(8m)$, $1/(2m)$, and $1/m$. Hence in our experiment: (i) the choice of α for the PS method does not follow $\mathcal{O}(1/\sqrt{T})$ recipe of Proposition 2.9 where T is the total number of PS iterations; (ii) T is determined by (75) rather than being pre-specified. PBF sets $\lambda = 1/2m$ and $\delta = \tilde{\varepsilon}[\phi(x_0) - \phi_*]$. PBF is implemented based on the 2-cut scheme (see Subsection 3.3). For phase retrieval and blind deconvolution problems, due to the fact that $h = 0$, the subproblem (20) in the 2-cut scheme has a closed-form solution. We implement RPB according to [15] and set $M_0 = 10$, $R_0 = 10$, the Armijo parameter to 0.05, and the convexification growth parameter to 2. It is implemented using the three-cut scheme, which retains the newly generated cut, the cut associated with the current serious point, and the aggregate cut obtained from the preceding bundle. The resulting proximal subproblem is formulated as a convex quadratic program and solved using MOSEK 10.2¹.

Finally, all experiments were performed in MATLAB 2023a and run on a PC with a 16-core Intel Core i9 processor and 32 GB of memory.

5.1 Phase retrieval

This subsection reports computational results comparing PBF against PS and RPB on the phase retrieval problem.

Consider the Phase Retrieval problem (see Subsection 5.1 of [5])

$$\min_{x \in \mathbb{R}^d} \left\{ f(x) := \frac{1}{n} \sum_{i=1}^n f_i(x) \right\}, \quad f_i(x) := \left| \langle a_i, x \rangle^2 - b_i \right| \quad (76)$$

where $a_i \in \mathbb{R}^d$ and $b_i \in \mathbb{R}$. Observe that $f_i(\cdot)$ is a special case of (74) with $g(\cdot) = |\cdot|$ and $c(\cdot) = \langle a_i, \cdot \rangle^2 - b_i$. Thus f_i is $2\|a_i a_i^\top\|$ -weakly convex, and we set

$$m = \frac{2}{n} \sum_{i=1}^n \|a_i a_i^\top\|.$$

Now we describe the data generation process. Vectors $\{a_i\}$ are i.i.d. generated according to standard Gaussian measurements $N(0, I_{d \times d})$. Then we generate the target signal $\bar{x}$ and initial point x_0 uniformly on the unit sphere; and set $b_i = \langle a_i, \bar{x} \rangle^2$ for each $i = 1, \dots, n$. Thus the optimal value $\phi_* = 0$ for (76). We perform four sets of experiments corresponding to $(d, n) = (100, 300), (200, 600), (500, 1500), (1000, 3000)$ and record the results in Tables 1 and 2.

We now describe some details about the tables that appear in this subsection. The second to fifth columns provide numbers of iterations and running times for the six variants. An entry in each table has two numbers where the first one expresses the number of iterations and the second one expresses the CPU running time in seconds. An entry marked as */* indicates that the CPU running time exceeds the one hour time limit.

¹<https://docs.mosek.com/latest/toolbox/index.html>

-	$L_1 : (100, 300)$	$L_2 : (200, 600)$	$L_3 : (500, 1500)$	$L_4 : (1000, 3000)$
PS: $\alpha = 1/(32m)$	50852/1.48	203416/17.82	*/*	*/*
PS: $\alpha = 1/(8m)$	14852/0.47	64294/5.58	432176/201.12	235700/650.21
PS: $\alpha = 1/(2m)$	4436/0.12	20011/1.76	345346/167.21	79165/216.67
PS: $\alpha = 1/m$	*/*	*/*	*/*	67095/183.56
PBF	5088/0.16	24486/2.08	316782/154.21	87514/242.62
RPB	9318/17.08	*/*	*/*	*/*

Table 1: PS versus PBF and RPB on (76). A relative tolerance of $\tilde{\varepsilon} = 10^{-3}$ is set.

-	$L_1 : (100, 300)$	$L_2 : (200, 600)$	$L_3 : (500, 1500)$	$L_4 : (1000, 3000)$
PS: $\alpha = 1/(32m)$	72376/2.11	380165/31.32	653256/305.65	*/*
PS: $\alpha = 1/(8m)$	20328/0.61	118589/9.91	*/*	424433/1.03E+03
PS: $\alpha = 1/(2m)$	*/*	*/*	*/*	*/*
PS: $\alpha = 1/m$	*/*	*/*	*/*	*/*
PBF	15228/0.42	125603/10.61	1056034/454.93	742553/1.86E+03
RPB	*/*	*/*	*/*	*/*

Table 2: PS versus PBF and RPB on (76). A relative tolerance of $\tilde{\varepsilon} = 10^{-4}$ is set.

The two tables show that PS is sensitive to the choice of stepsize, with the fastest stepsize varying across tolerance levels and problem dimensions. In contrast, PBF uses a single fixed stepsize and achieves performance comparable to the best-performing PS stepsize among the four tested choices. RPB reaches the prescribed tolerance only for the smallest instance $\tilde{\varepsilon} = 10^{-3}$ and reaches the time limit in all other instances.

5.2 Blind deconvolution

This subsection reports computational results comparing PBF against PS and RPB on the blind deconvolution problem. Consider the blind deconvolution problem (see Subsection 5.2 of [5])

$$\min_{x,y} \left\{ f(x) := \frac{1}{n} \sum_{i=1}^n f_i(x) \right\}, \quad f_i(x) := |\langle u_i, x \rangle \langle v_i, y \rangle - b_i| \quad (77)$$

where $u_i, v_i \in \mathbb{R}^d$ and $b_i \in \mathbb{R}$. Note that (77) is a special instance of (1) with $h \equiv 0$. Observe that $f_i(\cdot)$ is a special case of (74) with $g(\cdot) = |\cdot|$ and $c(x, y) = \langle u_i, x \rangle \langle v_i, y \rangle - b_i$. Thus $f_i(\cdot)$ is $\|u_i v_i^\top\|$ -weakly convex, and we set

$$m = \frac{1}{n} \sum_{i=1}^n \|u_i v_i^\top\|.$$

Now we describe the data generation process. Vectors $\{u_i\}, \{v_i\}$ are i.i.d. generated Gaussian measurements $N(0, I_{d \times d})$. Then we generate the target signals $\bar{x}, \bar{y}$ and initial points x_0, y_0 uniformly on the unit sphere; and set $b_i = \langle u_i, \bar{x} \rangle \langle v_i, \bar{y} \rangle$ for each $i = 1, \dots, n$. Thus the optimal value $\phi_* = 0$ for (77). We perform four sets of experiments corresponding to $(d, n) = (100, 300), (200, 600), (500, 1500), (1000, 3000)$ and record the results in Tables 3 and 4. The explanations of the tables are the same as in Subsection 5.1.

-	$L_1 : (100, 300)$	$L_2 : (200, 600)$	$L_3 : (500, 1500)$	$L_4 : (1000, 3000)$
PS: $\alpha = 1/(32m)$	76282/2.21	812578/143.48	*/*	*/*
PS: $\alpha = 1/(8m)$	16852/4.41	224893/39.62	121091/350.21	343700/740.55
PS: $\alpha = 1/(2m)$	5425/0.42	*/*	92137/241.41	83125/288.21
PS: $\alpha = 1/m$	*/*	*/*	*/*	64035/172.12
PBF	5021/0.39	344221/72.68	101299/280.67	77511/212.78
RPB	*/*	*/*	*/*	*/*

Table 3: PS versus PBF and RPB on (77). A relative tolerance of $\tilde{\varepsilon} = 10^{-3}$ is set.

-	$L_1 : (100, 300)$	$L_2 : (200, 600)$	$L_3 : (500, 1500)$	$L_4 : (1000, 3000)$
PS: $\alpha = 1/(32m)$	152976/6.12	1701581/300.33	*/*	*/*
PS: $\alpha = 1/(8m)$	41328/1.39	283353/50.32	173291/670.81	824121/1.93E+03
PS: $\alpha = 1/(2m)$	*/*	*/*	*/*	*/*
PS: $\alpha = 1/m$	*/*	*/*	*/*	*/*
PBF	35227/1.21	394212/89.32	188823/690.21	1142572/2.65E+03
RPB	*/*	*/*	*/*	*/*

Table 4: PS versus PBF and RPB on (77). A relative tolerance of $\tilde{\varepsilon} = 10^{-4}$ is set.

The two tables show that PS is sensitive to the choice of stepsize, with the best-performing stepsize in terms of CPU time varying across tolerance levels and problem dimensions. In contrast, PBF uses a single fixed stepsize and performs comparably to the best of the four tested PS stepsizes. RPB does not reach the prescribed tolerance within the time limit.

5.3 MAXQUAD

This subsection reports computational results comparing PBF, PS, and RPB method on the constrained MAXQUAD problem

$$\min_{x \in \mathbb{R}^d} \{\phi(x) := f(x) + h(x)\}, \quad f(x) := \max_{1 \leq j \leq p} \{q_j(x) := x^\top A_j x + b_j^\top x\}, \quad (78)$$

where h is the indicator function of the ball $\{x : \|x\| \leq D\}$. We set $p = 5$ and $D = 2$.

We now describe the data generation. Let $u_1, \dots, u_p$ be the vertices of a regular simplex in $\mathbb{R}^d$, satisfying $\sum_{j=1}^p u_j = 0$, $\|u_j\| = 1$, and $\langle u_i, u_j \rangle = -1/(p-1)$ for $i \neq j$. Let $P := (p-1) \sum_{j=1}^p u_j u_j^\top / p$, which is the orthogonal projector onto $\text{span}\{u_1, \dots, u_p\}$, and let x^* be a randomly generated unit vector orthogonal to this subspace. For $\beta > 0$, we set

$$\kappa = \frac{2\beta(p-1)}{p-2}, \quad A_j = \beta I + \kappa \left(\frac{P}{p-1} - u_j u_j^\top \right), \quad b_j = u_j - 2\beta x^*.$$

Since $\lambda_{\min}(A_j) = -\beta$, every q_j , and hence f , is 2β -weakly convex. We set $\beta = 1$, and hence $m = 2$. Moreover, $\sum_j A_j / p = \beta I$ and $\sum_j b_j / p = -2\beta x^*$, which imply

$$f(x) \geq \frac{1}{p} \sum_{j=1}^p q_j(x) = \beta \|x - x^*\|^2 - \beta.$$

Since all $q_j(x^*) = -\beta \|x^*\|^2 = -\beta$ and $\|x^*\| = 1 < D$, we have $\phi_* = \phi(x^*) = -\beta = -1$.

The initial point x_0 is generated from a standard Gaussian vector, scaled by $1/\sqrt{d}$, and projected onto the feasible ball if necessary. We test $d \in \{20, 50, 100, 200\}$. Each method is given an independent one-hour time limit. Each table entry reports the number of iterations and CPU time in seconds; */* indicates that the time limit was reached before the target accuracy was attained. Tables 5 and 6 report the results for $\tilde{\varepsilon} = 10^{-3}$ and 10^{-4} , respectively.

-	$L_1 : (20)$	$L_2 : (50)$	$L_3 : (100)$	$L_4 : (200)$
PS: $\alpha = 1/(32m)$	16636/0.29	17183/0.57	17260/1.27	17257/11.25
PS: $\alpha = 1/(8m)$	3385/0.06	3482/0.11	3502/0.26	3507/3.02
PS: $\alpha = 1/(2m)$	1732/0.03	1775/0.06	1784/0.13	1793/1.41
PS: $\alpha = 1/m$	*/*	*/*	*/*	*/*
PBF	818/0.09	750/0.07	776/0.10	855/0.78
RPB	*/*	*/*	*/*	*/*

Table 5: PS versus PBF and RPB on (78). A relative tolerance of $\tilde{\varepsilon} = 10^{-3}$ is set.

-	$L_1 : (20)$	$L_2 : (50)$	$L_3 : (100)$	$L_4 : (200)$
PS: $\alpha = 1/(32m)$	23101/0.37	23518/0.75	23620/1.72	23697/20.57
PS: $\alpha = 1/(8m)$	*/*	*/*	*/*	*/*
PS: $\alpha = 1/(2m)$	*/*	*/*	*/*	*/*
PS: $\alpha = 1/m$	*/*	*/*	*/*	*/*
PBF	8435/0.41	7491/0.51	7702/0.95	8571/8.76
RPB	*/*	*/*	*/*	*/*

Table 6: PS versus PBF and RPB on (78). A relative tolerance of $\tilde{\varepsilon} = 10^{-4}$ is set.

The two tables show that PBF requires fewer iterations than PS across the tested dimensions. Moreover, PS is sensitive to the choice of stepsize, whereas PBF uses the same stepsize in all instances. RPB does not reach the prescribed tolerance within the time limit.

6 Concluding Remarks

In this section, we provide some further remarks and directions for future research.

First, from the point of view of the sequences of serious iterates $\{\hat{x}_k\}$ and $\{\hat{y}_k\}$, the complexity result for PBF is point-wise since it is about a single iterate from $\{\hat{y}_k\}$. It would be also interesting to establish an ergodic complexity result about a weighted average of such sequence.

Second, as already observed in the fifth remark following PBF, PBF requires the knowledge of a weakly convex parameter, i.e., a scalar m as in Assumption (A1), and hence is not a universal method. It would be interesting to develop an adaptive method which do not require a scalar m as above, but instead generates adaptive estimates for it which possibly violate the condition on (A1).

Third, paper [6] presents an inexact proximal point method (see Algorithm 2 of [6]) for solving (1) where the subgradient method (SM) is used to obtain an approximate solution $\hat{x}_k$ of (31). More specifically, under the condition that $j_k \geq 11(1 + \lambda m)^2$, it is shown that j_k iterations of the SM initialized from $\hat{y}_{k-1}$ yield a point $\hat{x}_k$ satisfying

$$\phi(\hat{x}_k) - \phi(\hat{y}_{k-1}) \leq \frac{72\lambda M^2}{j_k + 1} - \frac{1}{4\lambda} \|\hat{y}_{k-1} - x_k^*\|^2$$

where x_k^* is the minimizer of (31). Hence, if j_k is chosen as

$$j_k = \bar{j} := \left\lceil \max \left\{ \frac{576\lambda^2 M^2}{\varepsilon^2}, \frac{576(1 + \lambda m)^2 M^2}{\varepsilon^2}, 11(1 + \lambda m)^2 \right\} \right\rceil,$$

then using (9), we conclude that

$$\frac{\lambda^2}{(1 + \lambda m)^2} \|\nabla \hat{M}^\lambda(\hat{y}_{k-1})\|^2 = \|\hat{y}_{k-1} - x_k^*\|^2 \leq 4\lambda[\phi(\hat{y}_{k-1}) - \phi(\hat{x}_k)] + \frac{\lambda^2 \varepsilon^2}{2(1 + \lambda m)^2},$$

and hence that

$$\inf_{k \leq K} \|\nabla \hat{M}^\lambda(\hat{y}_{k-1})\|^2 \leq \frac{4(1 + \lambda m)^2}{\lambda K} [\phi(x_0) - \phi(\hat{x}_K)] + \frac{\varepsilon^2}{2}. \quad (79)$$

The above conclusion gives an easily computable bound on the gradient of the Moreau envelope at $\hat{y}_{k-1}$. A drawback of the above bound is that it only holds if SM performs a prespecified number of iterations $j_k \geq \bar{j}$. In contrast, it is shown in Appendix E that the iteration sequence $\{(\hat{x}_k, \hat{y}_k)\}$ satisfies a bound similar to (79) even though it solves (31) only according to the stopping criterion (24), and hence without performing a prespecified number of inner iterations as SM does.

References

- [1] F. Atenas, C. Sagastizábal, P. JS Silva, and M. Solodov. A unified analysis of descent sequences in weakly convex optimization, including convergence rates for bundle methods. *SIAM Journal on Optimization*, 33(1):89–115, 2023.
- [2] A. M. Bagirov, M. Gaudioso, N. Karmita, M. M. Mäkelä, and S. Taheri. *Numerical nonsmooth optimization*. Springer, 2020.
- [3] A. Beck. *First-order methods in optimization*, volume 25. SIAM, 2017.
- [4] R. Correa and C. Lemaréchal. Convergence of some algorithms for convex minimization. *Mathematical Programming*, 62(1):261–275, 1993.
- [5] D. Davis and D. Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.
- [6] D. Davis and B. Grimmer. Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems. *SIAM Journal on Optimization*, 29(3):1908–1930, 2019.
- [7] W. de Oliveira. Proximal bundle methods for nonsmooth DC programming. *Journal of Global Optimization*, 75(2):523–563, 2019.
- [8] M. Díaz and B. Grimmer. Optimal convergence rates for the proximal bundle method. *SIAM Journal on Optimization*, 33(2):424–454, 2023.
- [9] D. Drusvyatskiy and C. Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, 178:503–558, 2019.
- [10] Y. Du and A. Ruszczyński. Rate of convergence of the bundle method. *Journal of Optimization Theory and Applications*, 173:908–922, 2017.
- [11] D. Fersztand and X. A. Sun. A modified proximal bundle method under a frank-wolfe perspective. *arXiv preprint arXiv:2411.15926*, 2024.
- [12] A. Frangioni. Generalized bundle methods. *SIAM Journal on Optimization*, 13(1):117–156, 2002.
- [13] A. Fuduli, M. Gaudioso, and G. Giallombardo. Minimizing nonconvex nonsmooth functions via cutting planes and proximity control. *SIAM Journal on Optimization*, 14(3):743–756, 2004.
- [14] W. Hare and C. Sagastizábal. Computing proximal points of nonconvex functions. *Mathematical Programming*, 116(1-2):221–258, 2009.
- [15] W. Hare and C. Sagastizábal. A redistributed proximal bundle method for nonconvex optimization. *SIAM Journal on Optimization*, 20(5):2442–2473, 2010.
- [16] W. Hare, C. Sagastizábal, and M. Solodov. A proximal bundle method for nonsmooth nonconvex functions with inexact information. *Computational Optimization and Applications*, 63(1):1–28, 2016.
- [17] K. Joki, A. M Bagirov, N. Karmita, and M. M. Mäkelä. A proximal bundle method for nonsmooth DC optimization utilizing nonconvex cutting planes. *Journal of Global Optimization*, 68(3):501–535, 2017.

- [18] C. Kanzow and T. Neder. A bundle-type method for nonsmooth DC programs. *Journal of Global Optimization*, 88(2):285–326, 2024.
- [19] K. C. Kiwiel. Efficiency of proximal bundle methods. *Journal of Optimization Theory and Applications*, 104(3):589, 2000.
- [20] K. C. Kiwiel. *Methods of descent for nondifferentiable optimization*, volume 1133. Springer, 2006.
- [21] C. Lemaréchal. An extension of davidon methods to non differentiable problems. In *Nondifferentiable optimization*, pages 95–109. Springer, 1975.
- [22] J. Liang and Y. Chen. Proximal oracles for optimization and sampling. *Journal of Optimization Theory and Applications*, 208(3):110, 2026.
- [23] J. Liang and R. D. C. Monteiro. A proximal bundle variant with optimal iteration-complexity for a large range of prox stepsizes. *SIAM Journal on Optimization*, 31(4):2955–2986, 2021.
- [24] J. Liang and R. D. C. Monteiro. A unified analysis of a class of proximal bundle methods for solving hybrid convex composite optimization problems. *Mathematics of Operations Research*, 49(2):832–855, 2024.
- [25] M. M. Makela and P. Neittaanmaki. *Nonsmooth optimization: analysis and algorithms with applications to optimal control*. World Scientific, 1992.
- [26] R. Mifflin. A modification and an extension of Lemaréchal’s algorithm for nonsmooth minimization. In *Nondifferential and variational techniques in optimization*, pages 77–90. Springer, 1982.
- [27] D. Noll, O. Prot, and A. Rondepierre. A proximity control algorithm to minimize nonsmooth and nonconvex functions. *Pacific Journal of Optimization*, 4(3):569–602, 2008.
- [28] W. de Oliveira, C. Sagastizábal, and C. Lemaréchal. Convex proximal bundle methods in depth: a unified analysis for inexact oracles. *Mathematical Programming*, 148(1-2):241–277, 2014.
- [29] A. Ruszczyński. *Nonlinear optimization*. Princeton university press, 2011.
- [30] J.-B. H. Urruty and C. Lemaréchal. *Convex analysis and minimization algorithms II*. Springer-Verlag, 1993.
- [31] W. van Ackooij and W. de Oliveira. *Methods of nonsmooth optimization in stochastic programming: from conceptual algorithms to real-world applications*. Springer, 2025.
- [32] J. Vlček and L. Lukšan. Globally convergent variable metric method for nonconvex nondifferentiable unconstrained minimization. *Journal of Optimization Theory and Applications*, 111:407–430, 2001.
- [33] P. Wolfe. A method of conjugate subgradients for minimizing nondifferentiable functions. In *Nondifferentiable optimization*, pages 145–173. Springer, 1975.

A Technical Results about Subdifferentials

This section presents two technical results about ε -subdifferentials that will be used in our analysis.

The first result describes a simple relationship involving ε -subdifferentials of $\phi_m(\cdot; x)$ for different points x .

Lemma A.1 *If $\phi : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$ is an m -weakly convex function, then $\phi_m(\cdot; c)$ is convex for every $c \in \mathbb{R}^n$. Moreover, for every $x \in \text{dom } \phi$, $c \in \mathbb{R}^n$, and $\varepsilon \geq 0$, we have*

$$\partial_\varepsilon [\phi_m(\cdot; x)](x) = \partial_\varepsilon [\phi_m(\cdot; c)](x) - m(x - c).$$

Proof: Let $x \in \text{dom } \phi$, $c \in \mathbb{R}^n$, and $\varepsilon \geq 0$ be given. Then, using the definition of $\phi_m(\cdot; \cdot)$ in (5), we easily see that

$$\phi_m(u; c) - \phi_m(x; c) - \langle v + m(x - c), u - x \rangle = \phi_m(u; x) - \phi(x) - \langle v, u - x \rangle \quad \forall u, v \in \mathbb{R}^n.$$

The result now follows from the above identity and the definition of the ε -subdifferential in (3). $\blacksquare$

The second result describes the relationship between an ε -solution and its global minimizer for a strongly convex function.

Lemma A.2 *If g is a proper lower semi-continuous μ -strongly convex function and y is an ε -solution of g , i.e., $0 \in \partial_\varepsilon g(y)$, then its global minimizer $\hat{y}$ satisfies*

$$0 \in \partial g(\hat{y}), \quad \|y - \hat{y}\| \leq \sqrt{\frac{2\varepsilon}{\mu}}. \quad (80)$$

Proof: The inclusion in (80) follows from the fact that $\hat{y}$ is a global minimizer of g . Since g is μ -strongly convex and $\hat{y}$ is its global minimizer, we have

$$\frac{\mu}{2} \|y - \hat{y}\|^2 \leq g(y) - g(\hat{y}) \leq \varepsilon$$

where the second inequality is due to $0 \in \partial_\varepsilon g(y)$. hence, the inequality in (80) follows. $\blacksquare$

B Relationships Between Notions of Stationary Points

This section contains three proofs for the results in Subsection 2.2.

Proof of Proposition 2.6 For this proof only, we let

$$\psi(\cdot) := \phi(\cdot) + \frac{\lambda^{-1} + m}{2} \|\cdot - x\|^2 \quad (81)$$

and denote $\hat{x}^\lambda(x)$ as in (10). Observe that ψ is a $(1/\lambda)$ -strongly convex function in view of the fact that ϕ is m -weakly convex. Thus, for any $u \in \text{dom } \psi$, it holds that

$$\psi(u) - \psi(\hat{x}) \geq \frac{1}{2\lambda} \|u - \hat{x}\|^2. \quad (82)$$

We start with the proof of a).

a) Since x is a $(\varepsilon_D, \delta_D)$ -directional stationary point, there exists a $\tilde{x}$ such that

$$\|x - \tilde{x}\| \leq \delta_D \quad \inf_{\|d\| \leq 1} \phi'(\tilde{x}; d) \geq -\varepsilon_D. \quad (83)$$

Then for any $d \in \mathbb{R}^n$, we have

$$\psi'(\tilde{x}; d) = \phi'(\tilde{x}; d) + \left(\frac{1}{\lambda} + m\right) \langle \tilde{x} - x, d \rangle \geq -\varepsilon_D \|d\| - \left(\frac{1}{\lambda} + m\right) \delta_D \|d\| = -\left[\varepsilon_D + \left(\frac{1}{\lambda} + m\right) \delta_D\right] \|d\|.$$

Using the convexity of ψ and the above relation with $d = \hat{x}^\lambda(x) - \tilde{x}$, we conclude that

$$\psi(\hat{x}) - \psi(\tilde{x}) \geq \psi'(\tilde{x}; \hat{x} - \tilde{x}) \geq -\left[\varepsilon_D + \left(\frac{1}{\lambda} + m\right) \delta_D\right] \|\hat{x} - \tilde{x}\|. \quad (84)$$

Then using the above inequality and (82) with $u = \tilde{x}$ we can conclude that

$$\varepsilon_D + \left(\frac{1}{\lambda} + m\right) \delta_D \geq \frac{1}{2\lambda} \|\tilde{x} - \hat{x}\|.$$

Thus (83) further implies that

$$\|x - \hat{x}\| \leq \|x - \tilde{x}\| + \|\tilde{x} - \hat{x}\| \leq \delta_D + 2\lambda \left[\varepsilon_D + \left(\frac{1}{\lambda} + m \right) \delta_D \right].$$

and hence using (9) we can get

$$\|\nabla \hat{M}^\lambda(x)\| = \left(m + \frac{1}{\lambda} \right) \|\hat{x}^\lambda(x) - x\| \leq \left(m + \frac{1}{\lambda} \right) [(3 + 2\lambda m)\delta_D + 2\lambda\varepsilon_D].$$

Now the statement follows from the above inequality and the definition of Moreau stationary point in Definition 2.5.

b) Since x is a $(\varepsilon_M; \lambda)$ -Moreau stationary point, (9) thus imply that

$$\left(\frac{1}{\lambda} + m \right) \|x - \hat{x}\| \leq \varepsilon_M.$$

Then for any $d \in \mathbb{R}^n$ such that $\|d\| \leq 1$, we have

$$0 \leq \psi'(\hat{x}; d) = \phi'(\hat{x}; d) + \left(\frac{1}{\lambda} + m \right) \langle \hat{x} - x, d \rangle \leq \phi'(\hat{x}; d) + \left(\frac{1}{\lambda} + m \right) \|x - \hat{x}\| \leq \phi'(\hat{x}; d) + \varepsilon_M$$

and

$$\|x - \hat{x}\| \leq \frac{\varepsilon_M}{m + 1/\lambda}.$$

Choose $\tilde{x} = \hat{x}$, $\varepsilon_D = \varepsilon_M$ and $\delta_D = \varepsilon_M/(m + 1/\lambda)$. Then the statement follows from the above relation and the definition of $(\varepsilon_D, \delta_D)$ -directional stationary point. $\blacksquare$

Proof of Proposition 2.7 Fix $x \in \text{dom } \phi$. For simplicity, we denote

$$\phi_m(\cdot) = \phi_m(\cdot; x) \stackrel{(5)}{=} \phi(\cdot) + \frac{m}{2} \|\cdot - x\|^2. \quad (85)$$

Since ϕ_m is convex, it follows from Theorem 3.26 of [3] that

$$-\inf_{\|d\| \leq 1} \phi'_m(x; d) = -\inf_{\|d\| \leq 1} \sigma_{\partial \phi_m(x)}(d) = -\inf_{\|d\| \leq 1} \sup_{s \in \partial \phi_m(x)} \langle s, d \rangle = \inf_{s \in \partial \phi_m(x)} \|s\| = \text{dist}(0; \partial \phi_m(x)). \quad (86)$$

In view of (85), (6) becomes that $\partial \phi(x) = \partial \phi_m(x)$, which together with (86) implies that

$$\text{dist}(0; \partial \phi(x)) = \text{dist}(0; \partial \phi_m(x)) = -\inf_{\|d\| \leq 1} \phi'_m(x; d).$$

Moreover, it follows from Definition 2.1 and (85) that

$$\phi'_m(x; d) = \liminf_{t \downarrow 0} \frac{\phi_m(x + td) - \phi_m(x)}{t} \stackrel{(85)}{=} \liminf_{t \downarrow 0} \frac{\phi(x + td) + \frac{m}{2} \|td\|^2 - \phi(x)}{t} = \phi'(x; d).$$

Therefore, the above two relations indicate that

$$\text{dist}(0; \partial \phi(x)) = -\inf_{\|d\| \leq 1} \phi'(x; d),$$

and the conclusion directly follows from Definition 2.2. $\blacksquare$

Proof of Proposition 2.8 a) Since x is a $(\bar{\eta}, \bar{\varepsilon}; m)$ -regularized stationary point of ϕ , there exists a pair $(w, \varepsilon) \in \mathbb{R}^n \times \mathbb{R}_{++}$ satisfying (7). Using the fact that $\phi_m(x; x) = \phi_{2m}(x; x)$ and $\phi_m(\cdot; x) \leq \phi_{2m}(\cdot; x)$, we easily see that the inclusion in (7) implies that $w \in \partial_\varepsilon[\phi_{2m}(\cdot; x)](x)$. Since ϕ_m is convex in view of assumption

(A1), we easily see that $\phi_{2m}(\cdot; x)$ is m -strongly convex, and hence the function $\phi_{2m}(\cdot; x) - \langle w, \cdot \rangle$ has a global minimizer $\tilde{x}$ which, in view of Lemma A.2 with $g = \phi_{2m}(\cdot; x) - \langle w, \cdot \rangle$ and $\mu = m$, satisfies

$$w \in \partial[\phi_{2m}(\cdot; x)](\tilde{x}), \quad \|x - \tilde{x}\| \leq \sqrt{\frac{2\varepsilon}{m}} \leq \sqrt{\frac{2\bar{\varepsilon}}{m}} \quad (87)$$

where the last inequality is due to the last inequality in (7). Now, letting $\hat{w} := w - 2m(\tilde{x} - x)$, it follows from (6) with $m = 2m$ and Lemma A.1 with $(\varepsilon, x, c) = (0, \tilde{x}, x)$ that

$$\hat{w} \in \partial[\phi_{2m}(\cdot; \tilde{x})](\tilde{x}) = \partial\phi(\tilde{x}). \quad (88)$$

Moreover, using the definition of $\hat{w}$ and the triangle inequality, we have

$$\|\hat{w}\| \leq \|w\| + 2m\|x - \tilde{x}\| \leq \bar{\eta} + 2\sqrt{2m\bar{\varepsilon}},$$

where the second inequality is due to the first inequality in (7) and the inequality in (87). Then the above inequality and (88) imply that

$$\text{dist}(0, \partial\phi(\tilde{x})) \leq \bar{\eta} + 2\sqrt{2m\bar{\varepsilon}}. \quad (89)$$

Now statement (a) follows from (87), (89), and Proposition 2.7.

b) Statement (b) immediately follows from the first statement and Proposition 2.6(a) with $\lambda = 1/m$. ■

C Results about an Important Recursive Formula

Lemma C.1 Assume that for some integer $J > 0$, scalars $\delta > 0$, $a \geq 0$ and $b \geq 0$ such that $a + b > 0$, and scalars $\{q_0, \dots, q_J\}$, we have that

$$q_j > \delta \quad \forall j \in \{0, \dots, J\}, \quad (90)$$

and

$$q_j \left(1 + \frac{q_j}{a + bq_j}\right) \leq q_{j-1}, \quad \forall j \in \{1, \dots, J\}. \quad (91)$$

Then,

$$J \leq \bar{N} := (2b + 3) \log^+ \left(\frac{q_0}{\delta + a/(b+1)} \right) + \frac{2a}{\delta} + b + 2. \quad (92)$$

Proof: Define $\tilde{\delta} := \delta + a/(b+1)$ and $\theta := 1/[2(b+1)]$. Suppose for contradiction that $J > \bar{N}$. We first claim that

$$\bar{j} := \max\{j \leq J : q_j \geq \tilde{\delta}\} \leq \mathcal{T} := \frac{\log^+(q_0/\tilde{\delta})}{\log(1+\theta)} \leq (1+\theta^{-1}) \log^+ \left(\frac{q_0}{\tilde{\delta}} \right) \leq \bar{N}. \quad (93)$$

Assume for contradiction that there exists an index $j > \mathcal{T}$ such that $q_j \geq \tilde{\delta}$. Using the fact $\{q_j\}$ is non-increasing (see (91)) and the definitions of $\tilde{\delta}$ and θ , we have that for every $l = 1, \dots, j$,

$$\frac{q_l}{a + bq_l} \geq \frac{\tilde{\delta}}{a + b\tilde{\delta}} = \frac{\delta + a/(b+1)}{a + b\delta + ab/(b+1)} \geq \frac{\delta + a/(b+1)}{2a + b\delta} \geq \frac{1}{2(b+1)} = \theta.$$

Hence, in view of (91), we conclude that $(1+\theta)^j q_j \leq (1+\theta)^{j-1} q_{j-1} \leq \dots \leq q_0$. On the other hand, using the fact that $j > \mathcal{T}$ and the definition of $\mathcal{T}$ in (93), we easily see that $q_0 < (1+\theta)^j \tilde{\delta}$. Combining the last two conclusions, it follows that $q_j < \tilde{\delta}$. Since the last conclusion contradicts the fact that $q_j \geq \tilde{\delta}$, the claim follows.

It follows from (91) that

$$q_{j-1} \geq q_j \frac{a + (b+1)q_j}{a + bq_j}.$$

Since $a \geq 0$ and $b \geq 0$, taking reciprocals gives

$$\frac{1}{q_{j-1}} \leq \frac{a + bq_j}{q_j(a + (b+1)q_j)}.$$

Hence, we obtain

$$\frac{1}{q_j} - \frac{1}{q_{j-1}} \geq \frac{1}{q_j} - \frac{a + bq_j}{q_j(a + (b+1)q_j)} = \frac{1}{a + (b+1)q_j}. \quad (94)$$

Note that the definition of $\bar{j}$ in (93) implies that $\tilde{\delta} > q_{\bar{j}+1} \geq q_j$ for any $j \geq \bar{j} + 1$. This fact and relation (94) then imply that for every $j \geq \bar{j} + 1$,

$$\frac{1}{q_j} - \frac{1}{q_{j-1}} \geq \frac{1}{a + (b+1)\tilde{\delta}}.$$

Thus, for any given $j \geq \bar{j} + 2$, summing the above inequality from $\bar{j} + 2$ to j and using the fact that $q_{\bar{j}+1} < \tilde{\delta}$, we then conclude that

$$\frac{1}{q_j} \geq \frac{1}{q_{\bar{j}+1}} + \frac{j-1-\bar{j}}{a + (b+1)\tilde{\delta}} > \frac{1}{\tilde{\delta}} + \frac{j-1-\bar{j}}{a + (b+1)\tilde{\delta}} \geq \frac{j-\bar{j}}{a + (b+1)\tilde{\delta}},$$

and hence that

$$q_N \leq \frac{a + (b+1)\tilde{\delta}}{\bar{N} - \bar{j}} = \frac{2a + (b+1)\delta}{\bar{N} - \bar{j}} \leq \delta \quad (95)$$

where the equality is due to the definition of $\tilde{\delta}$ and the last inequality is due to the definition of $\bar{N}$ and $\bar{j}$ in (92) and (93), respectively. Hence, the above inequality contradicts (90) and thus the statement follows. ■

D Technical Results for the Proof of Theorem 3.3

The main result of this section is Lemma D.3 which was used in the proof of Lemma 4.11.

Before stating and proving Lemma D.3, we first present two technical results whose proof can be found in Appendix A of [24].

Lemma D.1 *Let $z_0 \in \mathbb{R}^n$, $0 < \zeta < 1$, $\lambda > 0$, and $\Gamma \in \overline{\text{Conv}}(\mathbb{R}^n)$ be given, and define*

$$z_\lambda = \operatorname{argmin}_{u \in \mathbb{R}^n} \left\{ \Gamma(u) + \frac{1}{2\lambda} \|u - z_0\|^2 \right\}, \quad z_{\zeta\lambda} = \operatorname{argmin}_{u \in \mathbb{R}^n} \left\{ \Gamma(u) + \frac{1}{2\zeta\lambda} \|u - z_0\|^2 \right\}. \quad (96)$$

Then, we have $\|z_{\zeta\lambda} - z_0\| \geq \zeta \|z_\lambda - z_0\|$.

Lemma D.2 *For some $(M, L) \in \mathbb{R}_+^2$, assume that $(z_0, \lambda) \in \mathbb{R}^n \times (0, 1/L)$, function $\tilde{f} \in \overline{\text{Conv}}(\mathbb{R}^n) \cap \mathcal{C}(M, L)$ and $\Gamma, h \in \overline{\text{Conv}}(\mathbb{R}^n)$ are such that*

$$\ell_{\tilde{f}}(\cdot; z_0) + h \leq \Gamma \leq \tilde{f} + h.$$

Then, for every $u \in \operatorname{dom} h$, we have

$$\frac{1}{2\lambda} \|u - z_\lambda\|^2 + (\tilde{f} + h)(z_\lambda) - (\tilde{f} + h)(u) \leq \frac{1}{2\lambda} \|u - z_0\|^2 + \frac{2\lambda M^2}{1 - \lambda L}. \quad (97)$$

Lemma D.3 *For some $(M, L) \in \mathbb{R}_+^2$ and $m \geq 0$, assume that $(z_0, \lambda) \in \mathbb{R}^n \times \mathbb{R}_{++}$, function $f \in \mathcal{C}(M, L)$ is m -weakly convex, and $\Gamma, h \in \overline{\text{Conv}}(\mathbb{R}^n)$ are such that*

$$\ell_f(\cdot; z_0) + h \leq \Gamma \leq f_m(\cdot; z_0) + h,$$

where $f_m(\cdot; z_0)$ is as in (5). Then, for every $u \in \mathbb{R}^n$, we have

$$\zeta^2 \left(\frac{1}{4\zeta\lambda} + \frac{m}{2} \right) \|z_\lambda - z_0\|^2 \leq (f + h)(u) - (f + h)(z_{\zeta\lambda}) + \left(\frac{m}{2} + \frac{1}{\zeta\lambda} \right) \|u - z_0\|^2 + 4\zeta\lambda M^2, \quad (98)$$

where z_λ and $z_{\zeta\lambda}$ are as in (96) and ζ is defined as (35).

Proof: We first prove (98) under the assumption that $\lambda \in (0, (2(L+m))^{-1}]$, and hence $\zeta = 1$ in view of (35). Indeed, noting that $f_m(\cdot; z_0) \in \text{Conv}(\mathbb{R}^n) \cap \mathcal{C}(M, L+m)$ due to Lemma 3.1, it follows from Lemma D.2 with $\tilde{f} = f_m(\cdot; z_0)$ and $\lambda \leq 1/[2(L+m)]$ that for any $u \in \text{dom } h$:

$$\frac{1}{2\lambda} \|u - z_\lambda\|^2 + (f_m(\cdot; z_0) + h)(z_\lambda) - (f_m(\cdot; z_0) + h)(u) - \frac{1}{2\lambda} \|u - z_0\|^2 \stackrel{(97)}{\leq} \frac{2\lambda M^2}{1 - \lambda(L+m)} \leq 4\lambda M^2,$$

and hence that

$$(f_m(\cdot; z_0) + h)(u) - (f_m(\cdot; z_0) + h)(z_\lambda) + 4\lambda M^2 \geq \frac{1}{2\lambda} (\|u - z_\lambda\|^2 - \|u - z_0\|^2) \geq \frac{1}{4\lambda} \|z_\lambda - z_0\|^2 - \frac{1}{\lambda} \|u - z_0\|^2,$$

where the last inequality is due to the fact that $2a^2 + 2b^2 \geq (a+b)^2$ and the triangle inequality. Rearranging the above inequality and using the definition of $f_m(\cdot; z_0)$ in (5), we then conclude that

$$(f + h)(u) - (f + h)(z_\lambda) + 4\lambda M^2 \geq \left(\frac{1}{4\lambda} + \frac{m}{2}\right) \|z_\lambda - z_0\|^2 - \left(\frac{1}{\lambda} + \frac{m}{2}\right) \|u - z_0\|^2 \quad (99)$$

which, in view of the fact that $\zeta = 1$, immediately implies (98). Next, we show that (98) also holds for $\lambda > 1/[2(L+m)]$. Noting that (35) implies that $\zeta \in (0, 1)$ and $\zeta\lambda = 1/[2(L+m)]$, it then follows from (99) with λ replaced by $\zeta\lambda$ that

$$(f + h)(u) - (f + h)(z_{\zeta\lambda}) + \left(\frac{m}{2} + \frac{1}{\zeta\lambda}\right) \|u - z_0\|^2 + 4\zeta\lambda M^2 \geq \left(\frac{1}{4\zeta\lambda} + \frac{m}{2}\right) \|z_{\zeta\lambda} - z_0\|^2.$$

Finally, (98) immediately follows from the above inequality and Lemma D.1. $\blacksquare$

E Further Discussion Regarding (79)

Proposition E.1 *For every $K \geq 1$, then we have*

$$\inf_{k \leq K} \|\nabla \hat{M}^\lambda(\hat{y}_{k-1})\|^2 \leq \frac{2(1 + \lambda m)^2}{\lambda} \left(4\delta + \frac{3}{K} [\phi(\hat{x}_0) - \phi(\hat{y}_K)]\right).$$

Proof: For every $k \geq 1$, let x_k^* be the minimizer of (31). Hence, using the definition of $\phi_m(\cdot; \hat{y}_{k-1})$ in (5) and the fact that the objective function of (31) and is λ^{-1} -strongly convex, we have

$$\phi_m(\hat{y}_k; \hat{y}_{k-1}) + \frac{1}{2\lambda} \|\hat{y}_k - \hat{y}_{k-1}\|^2 \geq \phi_m(x_k^*; \hat{y}_{k-1}) + \frac{1}{2\lambda} \|x_k^* - \hat{y}_{k-1}\|^2 + \frac{1}{2\lambda} \|\hat{y}_k - x_k^*\|^2. \quad (100)$$

Moreover, using Lemma 4.5(a) and the fact that the objective function in (52) is λ^{-1} -strongly convex, we for every $u \in \text{dom } h$,

$$\hat{\Gamma}_k(\hat{x}_k) + \frac{1}{2\lambda} \|\hat{x}_k - \hat{y}_{k-1}\|^2 \leq \hat{\Gamma}_k(u) + \frac{1}{2\lambda} \|u - \hat{y}_{k-1}\|^2 - \frac{1}{2\lambda} \|u - \hat{x}_k\|^2. \quad (101)$$

Combining (53) and (55), and using (101), we have for every $u \in \text{dom } h$,

$$\begin{aligned} \hat{\delta}_k &\stackrel{(53), (55)}{\geq} \phi_m(\hat{y}_k; \hat{y}_{k-1}) + \frac{1}{2\lambda} \|\hat{y}_k - \hat{y}_{k-1}\|^2 - \hat{\Gamma}_k(\hat{x}_k) - \frac{1}{2\lambda} \|\hat{x}_k - \hat{y}_{k-1}\|^2 \\ &\stackrel{(101)}{\geq} \phi_m(\hat{y}_k; \hat{y}_{k-1}) + \frac{1}{2\lambda} \|\hat{y}_k - \hat{y}_{k-1}\|^2 - \hat{\Gamma}_k(u) - \frac{1}{2\lambda} \|u - \hat{y}_{k-1}\|^2 + \frac{1}{2\lambda} \|u - \hat{x}_k\|^2 \\ &\stackrel{(54)}{\geq} \phi_m(\hat{y}_k; \hat{y}_{k-1}) + \frac{1}{2\lambda} \|\hat{y}_k - \hat{y}_{k-1}\|^2 - \phi_m(u; \hat{y}_{k-1}) - \frac{1}{2\lambda} \|u - \hat{y}_{k-1}\|^2 + \frac{1}{2\lambda} \|u - \hat{x}_k\|^2, \end{aligned} \quad (102)$$

where the last inequality is due to the inequality in (54). Taking $u = \hat{y}_{k-1}$ in (102) yields

$$\hat{\delta}_k + \phi(\hat{y}_{k-1}) - \phi(\hat{y}_k) \geq \hat{\delta}_k + \phi(\hat{y}_{k-1}) - \phi_m(\hat{y}_k; \hat{y}_{k-1}) \geq \frac{1}{2\lambda} \|\hat{y}_k - \hat{y}_{k-1}\|^2 + \frac{1}{2\lambda} \|\hat{x}_k - \hat{y}_{k-1}\|^2. \quad (103)$$

Taking $u = x_k^*$ in (102) and using (100), we have

$$\hat{\delta}_k \geq \frac{1}{2\lambda} \|\hat{y}_k - x_k^*\|^2 + \frac{1}{2\lambda} \|x_k^* - \hat{x}_k\|^2. \quad (104)$$

Combining (103) and (104), we obtain

$$\begin{aligned} 2\hat{\delta}_k + \phi(\hat{y}_{k-1}) - \phi(\hat{y}_k) &\geq \left(\frac{1}{2\lambda} \|\hat{x}_k - \hat{y}_{k-1}\|^2 + \frac{1}{2\lambda} \|x_k^* - \hat{x}_k\|^2 \right) + \left(\frac{1}{2\lambda} \|\hat{y}_k - \hat{y}_{k-1}\|^2 + \frac{1}{2\lambda} \|\hat{y}_k - x_k^*\|^2 \right) \\ &\geq \frac{1}{4\lambda} \|x_k^* - \hat{y}_{k-1}\|^2 + \frac{1}{4\lambda} \|x_k^* - \hat{y}_{k-1}\|^2 = \frac{1}{2\lambda} \|x_k^* - \hat{y}_{k-1}\|^2, \end{aligned} \quad (105)$$

where the second inequality is due to the fact that $2a^2 + 2b^2 \geq (a+b)^2$ for every $a, b \in \mathbb{R}$ and the triangle inequality. Noting from step 3) of PBF and (23), and using (59), we have

$$\hat{\delta}_k \stackrel{(23)}{=} \delta + \frac{\lambda}{8(m\lambda + 1)} \|\hat{w}_k\|^2 \stackrel{(59)}{\leq} 2\delta + \frac{1}{2} \left(m + \frac{1}{\lambda} \right) \|\hat{y}_k - \hat{y}_{k-1}\|^2 \stackrel{(63)}{\leq} 2\delta + \phi(\hat{y}_{k-1}) - \phi(\hat{y}_k),$$

where the last inequality is due to (63). The above inequality and (105) yield that

$$\frac{1}{2\lambda} \|x_k^* - \hat{y}_{k-1}\|^2 \leq 4\delta + 3[\phi(\hat{y}_{k-1}) - \phi(\hat{y}_k)].$$

Moreover, recall that x_k^* is the minimizer of (31), and hence $x_k^* = \hat{x}^\lambda(\hat{y}_{k-1})$ in view of (10). It thus follows from (9) that

$$\frac{\lambda}{2(1 + \lambda m)^2} \|\nabla \hat{M}^\lambda(\hat{y}_{k-1})\|^2 \leq 4\delta + 3[\phi(\hat{y}_{k-1}) - \phi(\hat{y}_k)].$$

Therefore, the proposition immediately follows by summing the above inequality from $k = 1$ to K and dividing the resulting inequality by K . $\blacksquare$