Proximal bundle methods for hybrid weakly convex composite optimization problems

Jiaming Liang * Renato D.C. Monteiro † Honghao Zhang †

March 26, 2023 (first revision: December 26, 2024; second revision: May 16, 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 $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semi-continuous convex function and $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semi-continuous m-weakly convex function (i.e., $f + (m/2)\|\cdot\|^2$ is convex) such that $\mathrm{dom}\, f \supseteq \mathrm{dom}\, h$ and consider the composite optimization (CO) problem

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

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 $f' : \mathrm{dom}\, h \to \mathbb{R}^n$ (i.e., $f'(x) \in \partial f(x)$ for every $x \in \mathrm{dom}\, h$) satisfying the (M, L)-hybrid condition, namely: $\|f'(u) - f'(v)\| \le 2M + L\|u - v\|$ for every $u, v \in \mathrm{dom}\, h$. This problem class includes the class of weakly convex non-smooth (resp., smooth) CO problems, i.e., the one with $L = 0$ (resp., $M = 0$).

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 [4] and the references therein).

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

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

*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 AFORS Grants FA9550-22-1-0088 and FA9550-25-1-0131.

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. But 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: This paragraph discusses the development of proximal bundle methods in the context of convex CO problems. They have been proposed in [17, 27], and then further studied for example in [9, 21, 23]. Their convergence analyses for nonsmooth (i.e., $L = 0$ and $M > 0$) CO problems are broadly discussed for example in the textbooks [24, 25]. Moreover, their complexity analyses are studied for example in [6, 8, 14, 18, 19].

We now discuss the development of PB methods in the context of weakly convex CO problems as mentioned at the beginning of this introduction but with $L = 0$. We start with papers that deal only with their asymptotic convergence. PB extensions to this context can be found in [10, 11, 12, 13, 15, 20, 22, 26]. In particular, papers [11, 12, 13] already considered the idea of constructing convex bundle (more precisely, cutting-plane) models underneath the regularized function $\phi(\cdot)+(m/2)\|\cdot-x^c\|^2$, which, as already mentioned above, is the one adopted in this paper. The more recent paper [1] proposes a model to analyze descent-type bundle methods and establishes local convergence rate for the (serious) iteration sequence as well as function value sequence under some strong stationary growth condition. None of the aforementioned papers establish (either serious or overall) iteration complexities for their methods.

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 [4, 5], 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 [4]. 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 [4, 5] only consider the case where $L = 0$); 2) as opposed to the methods of [4, 5], PBF does not require knowledge of the constants M and L (see the first paragraph above) which are usually hard to estimate; but as the first two methods, PBF still requires knowledge of a weakly convex 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 [5] needs to perform a pre-specified number of subgradient iterations to the prox subproblem (see Section 3 of [5]) to estimate the size of the gradient of the Moreau envelope.

Organization of the paper. Subsection 1.1 presents basic definitions and notation used throughout the paper. Section 2 presents three stationary conditions and clarifies the relationships among these three notions. Section 3 contains four subsections. Subsection 3.1 formally describes problem (1) and the assumptions made on it. Subsection 3.2 reviews the deterministic version of the stochastic composite subgradient method of [4]. Subsection 3.3 describes PBF and states the iteration-complexity result of PBF. Subsection 3.4 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 contains two subsections. Subsection 5.1 presents computational results comparing PBF against the PS method for the phase retrieval problem. Subsection 5.2 showcases computational results comparing PBF against the PS method for the blind deconvolution problem. 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 Section 2. Appendix C describes one important result about 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.

1.1 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 $\varphi : \mathbb{R}^n \to (-\infty, +\infty]$, let $\operatorname{dom} \varphi := \{x \in \mathbb{R}^n : \varphi(x) < \infty\}$ denote the effective domain of φ and φ is proper if $\operatorname{dom} \varphi \neq \emptyset$. A proper function $\varphi : \mathbb{R}^n \to (-\infty, +\infty]$ 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 $x, y \in \operatorname{dom} \varphi$ and $\lambda \in [0, 1]$. Denote the set of all proper lower semicontinuous convex functions by $\overline{\operatorname{Conv}}\,(\mathbb{R}^n)$. For $\varepsilon \geq 0$, the *ε-subdifferential* of φ at $x \in \operatorname{dom} \varphi$ is denoted by

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

For simplicity, the subdifferential of φ at $x \in \operatorname{dom} \varphi$, i.e., $\partial_0 \varphi(x)$, is denoted by $\partial \varphi(x)$. The set of proper closed 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 Basic Definitions and Background

This section introduces the subdifferential and directional derivative of a general closed 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 closed function.

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

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

The next definition for ε-subdifferential can be found in Definition 1.10 of [16].

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

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

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 $0 \in \partial \phi(x)$. 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.

Definition 2.3 *The Frechet subdifferential of a proper closed function $\phi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is defined as*

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

Before stating the next result, we introduce the following notation which is used not only here but also throughout the paper: for every function $g(\cdot)$, $m \in \mathbb{R}_+$, $z \in \mathbb{R}^n$, let

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

The following result provides a characterization of the Frechet subdifferential for a weakly convex function.

Proposition 2.4 *Assume that $\phi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is a closed m-weakly convex function. Then we have*

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

$$= \partial\left[\phi_m(\cdot; x)\right](x). \tag{7}$$

Proof: The proof of (6) can be found for example in Lemma 2.1 of [4]. Moreover, it follows from the definition of the subdifferential in (3) with $\varepsilon = 0$ and the definition of $\phi_m(\cdot; x)$ in (5) that (7) is equivalent to (6). ■

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

Definition 2.5 (Regularized stationary point) *For a pair $(\bar\eta, \bar\varepsilon) \in \mathbb{R}^2_{++}$, a point $x \in \operatorname{dom}\phi$ 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*

$$w \in \partial_\varepsilon\left[\phi_m(\cdot; x)\right](x), \quad \|w\| \leq \bar\eta, \quad \varepsilon \leq \bar\varepsilon. \tag{8}$$

We make two trivial remarks about the above definition. First, if $(\bar\eta, \bar\varepsilon) = (0, 0)$ then it follows from (7) and Definition 2.5 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 (8) reduces to $w \in \partial_\varepsilon\phi(x)$, and the above notion reduces to a familiar one which has already been used in the analysis of several algorithms, including proximal bundle ones (e.g., see Section 6 of [18]), for solving the convex version of (1).

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 (8), 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 (8), so that verification that x_k is a $(\bar\eta, \bar\varepsilon; m)$-regularized stationary point simply amounts to checking the two inequalities in (8).

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

$$\hat M^\lambda(x) := \min_u\left\{\phi(u) + \frac{\lambda^{-1} + m}{2}\|u - x\|^2\right\}. \tag{9}$$

Note that the above definition depends on m but, for simplicity, we have omitted this dependence from its notation since the parameter m is assumed constant throughout our analysis in this paper.

The gradient formula for the Moreau envelope (see Section 1 in [4]) is as follows:

$$\nabla\hat M^\lambda(x) = \left(\frac{1}{\lambda} + m\right)\left(x - \hat x^\lambda(x)\right) \tag{10}$$

where

$$\hat x^\lambda(x) := \operatorname{argmin}_{u \in \mathbb{R}^n}\left\{\phi(u) + \frac{\lambda^{-1} + m}{2}\|u - x\|^2\right\}. \tag{11}$$

The following definition describes another notion of a stationary point that is based on the aforementioned Moreau envelope.

Definition 2.6 (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 $\|\nabla\hat M^\lambda(x)\| \leq \varepsilon_M$.*

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

Proposition 2.7 *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 2.8 *Assume that ϕ is a m-weakly convex function. Then, x is a $(\varepsilon_D, \delta_D)$-directional stationary point if and only if there exists $\tilde{x} \in \mathrm{dom}\,\phi$ such that*

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

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 2.9 *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.

3 Algorithm

This section contains four subsections. The first one describes the main problem and the assumptions made on it. The second one reviews the deterministic version of stochastic proximal subgradient method of [4] and its main complexity result. The third 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 $f, h : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ and a scalar $m > 0$ such that $h \in \overline{\mathrm{Conv}}\,(\mathbb{R}^n)$, f is m-weakly convex and $\mathrm{dom}\,h \subset \mathrm{dom}\,f$;

(A2) there exist constants $M, L \ge 0$ and a subgradient oracle, i.e., a function $f' : \mathrm{dom}\,h \to \mathbb{R}^n$ satisfying $f'(u) \in \partial f(u)$ for every $u \in \mathrm{dom}\,h$, such that

$$\|f'(u) - f'(v)\| \le 2M + L\|u - v\| \quad \forall u, v \in \mathrm{dom}\,h; \tag{12}$$

(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 $\mathrm{dom}\,h$, we have omitted this dependence from its notation. For any function $g \in \mathcal{C}(M, L)$ and $x \in \mathrm{dom}\,h$, we denote the linearization of g at x by

$$\ell_g(\cdot; x) := g(x) + \langle g'(x), \cdot - x\rangle. \tag{13}$$

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* $\operatorname{dom} h$. *Then, for every $z \in \mathbb{R}^n$ and $\tilde{z} \in \operatorname{dom} h$, we have*

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

Proof: Since $f \in \mathcal{C}(M,L)$, there exists an oracle f' satisfying the conditions in (A2), i.e., $f'(x) \in \partial f(x)$ for every $x \in \operatorname{dom} f$ and (12) holds. Since f is m-weakly convex, it follows from the definition of weak convexity that $f_m \in \overline{\mathrm{Conv}}\,(\mathbb{R}^n)$. Moreover, it follows from Lemma A.1 with $(\varepsilon, x, c) = (0, \cdot, z)$ and the fact that $f'(x) \in \partial f(x)$ for every $x \in \operatorname{dom} f$ that $f'_m(\cdot; z) := f'(\cdot) + m(\cdot - z) \in \partial[f_m(\cdot; z)](\cdot)$. The result now follows by noting that (12) and the definition of $f_m + 1/\lambda$ imply that for every $x, y \in \operatorname{dom} f$

$$\begin{aligned} \|f'_m(x;z) - f'_m(y;z)\| &= \|f'(x) - f'(y) + m(x-y)\| \\ &\le \|f'(x) - f'(y)\| + m\|x - y\| \\ &\le 2M + (L+m)\|x - y\| \end{aligned}$$

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

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

$$\ell_{f_m(\cdot;z)}(\cdot; \tilde{z}) := f_m(\tilde{z}; z) + \langle f'(\tilde{z}) + m(\tilde{z} - z)\,,\, \cdot - \tilde{z}\rangle. \tag{14}$$

Moreover, in view of the second inclusion of Lemma 3.1, we have

$$0 \le f_m(\cdot; z) - \ell_{f_m(\cdot;z)}(\cdot; \tilde{z}) \le 2M\|\cdot - \tilde{z}\| + \frac{L+m}{2}\|\cdot - \tilde{z}\|^2 \quad \forall z \in \mathbb{R}^n,\ \tilde{z} \in \operatorname{dom} h. \tag{15}$$

Note that the above observations with $\tilde{z} = z$ implies that for every $z \in \operatorname{dom} h$, we have

$$\ell_{f_m(\cdot;z)}(\cdot; z) = \ell_f(\cdot; z), \tag{16}$$

and

$$0 \le f_m(\cdot; z) - \ell_f(\cdot; z) \le 2M\|\cdot - z\| + \frac{L+m}{2}\|\cdot - z\|^2 \quad \forall z \in \operatorname{dom} h. \tag{17}$$

3.2 Review of the PS method for the weakly convex case

This subsection reviews the deterministic version of the PS method of [4].

More specifically, it considers the PS method described below under the assumptions stated in Subsection 3.1 except that the constant L in condition (A2) is assumed to be zero.

PS

Input: $\hat{x}_0 \in \operatorname{dom} h$, a sequence of $\{\alpha_t\}_{t \ge 0} \subset \mathbb{R}_+$ and iteration count T.
Step: For $t = 0, 1, \cdots, T$, compute

$$\hat{x}_{t+1} = \operatorname{argmin}_{u \in \mathbb{R}^n} \left\{ \ell_f(u; \hat{x}_t) + h(u) + \frac{1}{2\alpha_t}\|u - \hat{x}_t\|^2 \right\}. \tag{18}$$

The following result (see Theorem 3.4 in [4]) describes the rate of convergence of the PS method.

Proposition 3.2 *Assume (f, h) are functions satisfying conditions (A1)-(A3) with $L = 0$ and assume that $\alpha_t \in (0, 1/\bar{m}]$ for every $t \ge 0$ for some $\bar{m} \in (m, 2m]$. Then, the iterates $\hat{x}_t$ of PS satisfies*

$$\frac{\sum_{t=0}^T \alpha_t \left\| \nabla \hat{M}^{1/(\bar{m}-m)}\,(\hat{x}_t) \right\|^2}{\sum_{t=0}^T \alpha_t} \le \frac{\bar{m}}{\bar{m} - m} \cdot \frac{\left(\hat{M}^{1/(\bar{m}-m)}\,(\hat{x}_0) - \phi^*\right) + 2\bar{m}M^2 \sum_{t=0}^T \alpha_t^2}{\sum_{t=0}^T \alpha_t}. \tag{19}$$

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

$$T \geq \frac{\left[\left(\hat{M}^{1/m}\left(\hat{x}_0\right) - \phi^*\right) + 4mM^2\gamma^2\right]^2}{\gamma^2\rho^4}, \tag{20}$$

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

We now make some remarks about PS and Proposition 3.2. First, PS always performs $T + 1$ iterations and the second part of Proposition 3.2 gives a sufficient condition on T which guarantees that one of its iterates is a $(\rho; 1/m)$-Moreau stationary point. An alternative 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 (20) on T is that it is generally not computable as it depends on ϕ^*. Third, the method of [4] actually outputs an iterate $\hat{x}_{t^*}$ where t^* is sampled from $\{0, \ldots, T\}$ according to the probability mass function $\mathbb{P}(t^* = t) = \alpha_t / \sum_{i=0}^T \alpha_t$ for every $t = 0, \ldots, T$. It turns out that $\mathbb{E}[\|\nabla \hat{M}^{1/m}(\hat{x}_{t^*})\|^2]$ is equal to the left hand side of (19) and hence is bounded by the right hand side of (19). 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 (19). However, the authors are unaware of any technique for de-randomizing this output strategy due to the fact that the function $\|\nabla \hat{M}^{1/m}(\cdot)\|^2$ is generally nonconvex and hard to compute.

Finally, the third remark in the Concluding Remarks discusses how the alternative inexact proximal point method developed in [5] 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.

3.3 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 [4] described in Subsection 3.2.

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

$$x_j = \underset{u \in \mathbb{R}^n}{\operatorname{argmin}} \left\{\Gamma_j(u) + \frac{1}{2\lambda}\|u - x^c\|^2\right\} \tag{21}$$

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 3.3 *For any convex function φ, given $x^c \in \mathbb{R}^n$, $\lambda > 0$, and $\Gamma(\cdot) \in \mathcal{B}(\varphi)$, 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 initial point $\hat{x}_0 \in \operatorname{dom} h$, tolerance pair $(\bar{\eta}, \bar{\varepsilon}) \in \mathbb{R}^2_{++}$, and prox stepsize $\lambda > 0$ be given, and set $y_0 = \hat{y}_0 = \hat{x}_0$, $j = 1$, $k = 1$, and

$$\delta = \min\left\{\frac{\bar{\varepsilon}}{16}, \frac{\lambda\bar{\eta}^2}{64(m\lambda+2)}\right\}, \qquad \Gamma_1(\cdot) = \ell_f(\cdot;\hat{x}_0) + h(\cdot); \tag{22}$$

1. Compute the optimal solution x_j and optimal value θ_j of (21) with $x^c = \hat{y}_{k-1}$; if

$$\phi_m(x_j;\hat{y}_{k-1}) + \frac{1}{2\lambda}\|x_j - \hat{y}_{k-1}\|^2 < \phi_m(y_{j-1};\hat{y}_{k-1}) + \frac{1}{2\lambda}\|y_{j-1} - \hat{y}_{k-1}\|^2, \tag{23}$$

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

2. Compute

$$w_j := \frac{1}{\lambda}(\hat{y}_{k-1} - x_j) - m(y_j - \hat{y}_{k-1}), \quad \delta_j := \delta + \frac{\lambda}{8(m\lambda+1)}\|w_j\|^2; \tag{24}$$

3. If

$$t_j := \phi_m(y_j;\hat{y}_{k-1}) + \frac{1}{2\lambda}\|y_j - \hat{y}_{k-1}\|^2 - \theta_j > \delta_j, \tag{25}$$

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

$$\Gamma_{j+1}(\cdot) = \max\left\{\ell_{f_m(\cdot;\hat{y}_{k-1})}(\cdot;x_j) + h(\cdot), \bar{\Gamma}_j(\cdot)\right\} \tag{26}$$

where $\bar{\Gamma}_j$ is a shadow of Γ_j for (21) 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

$$\hat{v}_k = \frac{1}{\lambda}(\hat{y}_{k-1} - \hat{x}_k), \tag{27}$$

and compute

$$\hat{\varepsilon}_k = \phi_m(\hat{y}_k;\hat{y}_{k-1}) - \hat{\Gamma}_k(\hat{x}_k) - \langle \hat{v}_k, \hat{y}_k - \hat{x}_k\rangle \tag{28}$$

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

$$\Gamma_{j+1}(\cdot) \ge \ell_f(\cdot;\hat{y}_k) + h(\cdot), \tag{29}$$

where $\ell_f(\cdot;\cdot)$ is defined in (13), 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 $t_j \le \delta$ is called a serious iteration; otherwise, j is called a null iteration. Let $j_1 \le j_2 \le \ldots$ 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 $j_{k-1} + 1 \le j \le j_k$, i.e.,

$$\mathcal{C}_k := \{i_k, \ldots, j_k\}, \quad i_k := j_{k-1} + 1. \tag{30}$$

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 (26) or (29) 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 $k \ge 1$. Hence, $\hat{y}_k$ is $(\bar{\eta}, \bar{\varepsilon}; m)$-regularized stationary point of ϕ whenever the stopping criterion $\|\hat{w}_k\| \le \bar{\eta}$ and $\hat{\varepsilon}_k \le \bar{\varepsilon}$ is satisfied in step 3b. Third, in view of the definition of y_j in (23) and the above relation, it then follows that

$$y_j \in \operatorname{Argmin}\left\{\phi_m(x;\hat{y}_{k-1}) + \frac{1}{2\lambda}\|x - \hat{y}_{k-1}\|^2 : x \in \{\hat{y}_{k-1}, x_{i_k}, \ldots, x_j\}\right\}. \tag{31}$$

Fourth, two simple ways of choosing the bundle function Γ_{j+1} such that (29) 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 (29). Moreover, using the fact that $\Gamma_j \le \phi_m(\cdot;\hat{y}_{k-1})$ and the definition of $\phi_m(\cdot;\hat{y}_{k-1})$ in (5), it is easy to see that $\tilde{\Gamma}_{j+1} \le \phi_m(\cdot;\hat{y}_k)$, and hence that the second choice for Γ_{j+1} also satisfies (29). 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 (26) is a special case of the one proposed in relations (4.7)-(4.9) of [3] (see also (2.2)-(2.4) of [6]).

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 cutting plane iterations applied to the prox subproblem

$$\hat{M}^{\lambda}(\hat{y}_{k-1}) = \min\left\{\phi(x) + \frac{1}{2}\left(\frac{1}{\lambda}+m\right)\|x-\hat{y}_{k-1}\|^2 : x \in \mathbb{R}^n\right\}. \tag{32}$$

Second, the pair (y_j, Γ_j) found in the serious (i.e., the last) iteration j of the k-th cycle approximately solves (32) according to (25), in which case the point x_j as in step 1 of PBF becomes the center $\hat{y}_k$ for the next cycle.

We now discuss how the inexact criterion (25) can be interpreted in terms of the prox subproblem (32). Since $\Gamma_j \le \phi_m(\cdot;\hat{y}_{k-1})$, 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 , (32), and (5), respectively, that $\theta_j \le \hat{M}^{\lambda}(\hat{y}_{k-1})$, and hence that

$$\begin{aligned} 0 &\le \phi(y_j) + \frac{1}{2}\left(\frac{1}{\lambda}+m\right)\|y_j - \hat{y}_{k-1}\|^2 - \hat{M}^{\lambda}(\hat{y}_{k-1}) \\ &\le \phi(y_j) + \frac{1}{2}\left(\frac{1}{\lambda}+m\right)\|y_j - \hat{y}_{k-1}\|^2 - \theta_j = t_j, \end{aligned} \tag{33}$$

where the last identity follows from the definitions of t_j in (25). Hence, if j is an iteration for which (25) does not hold (i.e., a serious one), then it follows from (33) that y_j is a δ-solution of (32).

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

$$\bar{\sigma} := \frac{[4\lambda(L+m)+1]\left(M^2 + \beta_2\left\{\beta_1[\hat{M}^{\lambda}(\hat{x}_0) - \phi^* + 2\delta] + 4\zeta\lambda M^2\right\}\right)}{8\lambda M^2} \tag{34}$$

where

$$\beta_1 := \left(m + \frac{2}{\zeta\lambda}\right)\left(m+\frac{1}{\lambda}\right)^{-1}, \quad \beta_2 := \left(\frac{L+m}{2}+1\right)\zeta^{-2}\left(\frac{1}{4\zeta\lambda}+\frac{m}{2}\right)^{-1}, \tag{35}$$

and

$$\zeta := \begin{cases} \frac{1}{2(L+m)\lambda} & \text{if } \lambda > \frac{1}{2(L+m)}; \\ 1 & \text{if } \lambda \le \frac{1}{2(L+m)}. \end{cases} \tag{36}$$

Theorem 3.4 (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*

$$\left[4 + 4\lambda(L+m) + \frac{16\lambda M^2}{\delta} + [8\lambda(L+m)+3]\log^+\bar{\sigma}\right]\left[2\left(\frac{1}{\bar{\varepsilon}} + \frac{8(m\lambda+1)}{\lambda\bar{\eta}^2}\right)(\hat{M}^{\lambda}(\hat{x}_0) - \phi^*) + 2\right] \tag{37}$$

total iterations where $(\bar{\eta}, \bar{\varepsilon})$ is as in step 0 of PBF and δ is as in (22).

In terms of the tolerances $\bar{\eta}$ and $\bar{\varepsilon}$ only, it follows from Theorem 3.4 that the iteration complexity of PBF to find a $(\bar{\eta}, \bar{\varepsilon}; m)$-regularized stationary point of ϕ is $\mathcal{O}(\bar{\eta}^{-2} \max\{\bar{\eta}^{-2}, \bar{\varepsilon}^{-1}\})$.

For the sake of stating an iteration-complexity for PBF to find a $(\rho; 1/m)$-Moreau stationary point, we consider the specific case of PBF and Theorem 3.4 with parameters λ chosen as $\lambda = \Theta(m^{-1})$ and $L = 0$. In this case, it follows from (22) that

$$\delta = \mathcal{O}\left(\min\left\{\frac{\bar{\eta}^2}{m}, \bar{\varepsilon}\right\}\right). \tag{38}$$

Thus, the complexity bound (37) becomes

$$\mathcal{O}\left(\frac{M^2[\hat{M}^\lambda(\hat{x}_0) - \phi^*]}{\delta \bar{\eta}^2}\right) \tag{39}$$

which, in view of (38), is

$$\mathcal{O}\left(\frac{\hat{M}^\lambda(\hat{x}_0) - \phi^*}{\bar{\eta}^2}\left(M^2 \max\left\{\frac{m}{\bar{\eta}^2}, \frac{1}{\bar{\varepsilon}}\right\}\right)\right). \tag{40}$$

We are now ready to state the iteration complexity for PBF to find an approximate Moreau stationary point.

Corollary 3.5 *For any given $\rho > 0$, 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 (12) with $L = 0$ generates a $(\rho; 1/m)$-Moreau stationary point of ϕ within the iteration sequence $\{\hat{y}_k\}$ in

$$\mathcal{O}\left(\frac{M^2 m[\hat{M}^\lambda(\hat{x}_0) - \phi^*]}{\rho^4}\right) \tag{41}$$

total iterations.

Proof: 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 2.9(b). The choice of $(\bar{\eta}, \bar{\varepsilon}, 1)$ and relation (38) imply that

$$\delta = \Theta\left(\frac{\rho^2}{m}\right).$$

The conclusion of the corollary now follows from the above relation and (40). ∎

It is worth noting that the iteration-complexity of the PS method of Subsection 3.2 is not better than the one for PBF with $\lambda = 1/m$ and $L = 0$, namely, bound (41), and the first bound on the right hand side of (20) equals the latter one only when

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

However, this choice of γ is generally not computable because the quantity $\hat{M}^\lambda(\hat{x}_0) - \phi^*$ and the parameter M may not be known.

3.4 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.

- **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
$$A_{j+1}(\cdot) = \theta A_j(\cdot) + (1-\theta)\ell_{f_m(\cdot;\hat{y}_{k-1})}(\cdot;x_{j-1}) \tag{42}$$
for some $\theta \in [0,1]$. In fact, (21) 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
$$x_j = \underset{u\in\mathbb{R}^n, t\in\mathbb{R}}{\operatorname{argmin}} \left\{ t + h(u) + \frac{1}{2\lambda}\|u-\hat{y}_{k-1}\|^2 : A_j(u) \le t,\ \ell_{f_m(\cdot;\hat{y}_{k-1})}(u;x_{j-1}) \le t \right\}, \tag{43}$$
we denote by θ an optimal Lagrange multiplier associated with the constraint $A_j(u) \le t$. Moreover, it is easy to check from the optimality conditions of (43) that $\bar{\Gamma}_j$ is a shadow of Γ_j for (21) with $x^c = \hat{y}_{k-1}$.

- **multi-cut:** This scheme sets $\bar{\Gamma}_j(\cdot) = \Gamma(\cdot; C(x_j))$ where
$$\Gamma(\cdot; C(x_j)) := \max\{\ell_{f_m(\cdot;\hat{y}_{k-1})}(\cdot;c) : c \in C(x_j)\} + h(\cdot), \tag{44}$$
and
$$C(x_j) := \{c \in C_j : \ell_{f_m(\cdot;\hat{y}_{k-1})}(x_j;c) + h(x_j) = \Gamma_j(x_j)\}. \tag{45}$$
It can be shown that $\bar{\Gamma}_j$ is a shadow of Γ_j for (21) with $x^c = \hat{y}_{k-1}$ following a similar argument as in the proof of Proposition D.2 of [19].

4 Proof of Theorem 3.4

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 (30), and hence on the number of null iterations between two consecutive serious ones.

Throughout this section and the next one, we let
$$\hat{f}_k(\cdot) = f_m(\cdot; \hat{y}_{k-1}). \tag{46}$$

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

Lemma 4.1 *For every $j \in \mathcal{C}_k \setminus \{i_k\}$, the following statements hold:*

a) there exists $\bar{\Gamma}_{j-1} \in \overline{\operatorname{Conv}}(\mathbb{R}^n)$ such that
$$\max\{\bar{\Gamma}_{j-1}, \ell_{\hat{f}_k}(\cdot;x_{j-1}) + h\} \le \Gamma_j, \tag{47}$$
$$\bar{\Gamma}_{j-1}(x_{j-1}) = \Gamma_{j-1}(x_{j-1}), \quad x_{j-1} = \underset{u\in\mathbb{R}^n}{\operatorname{argmin}} \left\{ \bar{\Gamma}_{j-1}(u) + \frac{1}{2\lambda}\|u-\hat{y}_{k-1}\|^2 \right\}; \tag{48}$$

b) for every $u \in \operatorname{dom} h$, there holds
$$\bar{\Gamma}_{j-1}(u) + \frac{1}{2\lambda}\|u-\hat{y}_{k-1}\|^2 \ge \theta_{j-1} + \frac{1}{2\lambda}\|u-x_{j-1}\|^2; \tag{49}$$

Proof: a) This statement immediately follows from (21) with $x^c = \hat{y}_{k-1}$, step 3a of PBF, and Definition 3.3.

b) Using the second identity in (48) and the fact that $\bar{\Gamma}_{j-1} \in \overline{\text{Conv}}\,(\mathbb{R}^n)$, we have for every $u \in \text{dom}\, h$,

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

The statement now follows from the above inequality, the first identity in (48), and the definition of θ_{j-1} in (25). ■

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:*

$$\theta_j \geq \theta_{j-1} + \frac{1}{2\lambda}\|x_j - x_{j-1}\|^2, \tag{50}$$

$$t_j + \frac{1}{2\lambda}\|x_j - x_{j-1}\|^2 \leq t_{j-1}. \tag{51}$$

Proof: Using the definition of θ_j in step 1 of PBF , (47), and (49) with $u = x_j$, we have

$$\begin{aligned}\theta_j &= \Gamma_j(x_j) + \frac{1}{2\lambda}\|x_j - \hat{y}_{k-1}\|^2 \\ &\overset{(47)}{\geq} \bar{\Gamma}_{j-1}(x_j) + \frac{1}{2\lambda}\|x_j - \hat{y}_{k-1}\|^2 \overset{(49)}{\geq} \theta_{j-1} + \frac{1}{2\lambda}\|x_j - x_{j-1}\|^2\end{aligned}$$

and thus (50) holds. Using the definition of t_j in (25), (31), and (50), we have

$$\begin{aligned}t_j &\overset{(25)}{=} \phi_m(y_j;\hat{y}_{k-1}) + \frac{1}{2\lambda}\|y_j - \hat{y}_{k-1}\|^2 - \theta_j \overset{(31)}{\leq} \phi_m(y_{j-1};\hat{y}_{k-1}) + \frac{1}{2\lambda}\|y_{j-1} - \hat{y}_{k-1}\|^2 - \theta_j \\ &\overset{(50)}{\leq} \phi_m(y_{j-1};\hat{y}_{k-1}) + \frac{1}{2\lambda}\|y_{j-1} - \hat{y}_{k-1}\|^2 - \theta_{j-1} - \frac{1}{2\lambda}\|x_j - x_{j-1}\|^2 \overset{(25)}{=} t_{j-1} - \frac{1}{2\lambda}\|x_j - x_{j-1}\|^2\end{aligned}$$

and thus (51) holds. ■

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:*

$$t_j\left(1 + \frac{t_j}{4\lambda(2M^2 + (L+m)t_j)}\right) \leq t_{j-1}. \tag{52}$$

Proof: Using the definition of t_j in (25), (31), and the definition of θ_j in step 1 of PBF , we conclude that

$$\begin{aligned}t_j &\overset{(25)}{=} \phi_m(y_j;\hat{y}_{k-1}) + \frac{1}{2\lambda}\|y_j - \hat{y}_{k-1}\|^2 - \theta_j \overset{(31)}{\leq} \phi_m(x_j;\hat{y}_{k-1}) + \frac{1}{2\lambda}\|x_j - \hat{y}_{k-1}\|^2 - \theta_j \\ &= \phi_m(x_j;\hat{y}_{k-1}) - \Gamma_j(x_j).\end{aligned}$$

The above inequality, relations (47) and (15), 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

$$\begin{aligned}0 &\leq \phi_m(x_j;\hat{y}_{k-1}) - (\ell_{\hat{f}_k}(x_j;x_{j-1}) + h(x_j)) - t_j \\ &\overset{(47)}{=} f_m(x_j;\hat{y}_{k-1}) - \ell_{\hat{f}_k}(x_j;x_{j-1}) - t_j \\ &\overset{(15)}{\leq} \frac{L+m}{2}\|x_j - x_{j-1}\|^2 + 2M\|x_j - x_{j-1}\| - t_j,\end{aligned}$$

which, together with $t_j \geq 0$ due to (33), 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 (51) and the above inequality. ■

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

$$N_k := 4 + 4\lambda(L+m) + \frac{16\lambda M^2}{\delta} + [8\lambda(L+m)+3]\log^+ \sigma_k \tag{53}$$

where

$$\sigma_k := \frac{t_{i_k}}{\delta + 8\lambda M^2/[4\lambda(L+m)+1]}. \tag{54}$$

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

$$q_j := t_{j+i_k} > \delta \quad \forall j = 0, \ldots, J. \tag{55}$$

Observe that Lemma 4.3 and (55) imply that assumptions (93) and (94) 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 (53) imply that $J \leq N_k - 2$ which contradicts with the assumption $|\mathcal{C}_k| = J + 2 > N_k$. ■

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

$$\min_{u \in \mathbb{R}^n} \left\{ \hat{\Gamma}_k(u) + \frac{1}{2\lambda}\|u - \hat{y}_{k-1}\|^2 \right\}; \tag{56}$$

hence, if $\hat{\theta}_k$ denotes the optimal value of (56), then

$$\hat{\theta}_k = \hat{\Gamma}_k(\hat{x}_k) + \frac{1}{2\lambda}\|\hat{x}_k - \hat{y}_{k-1}\|^2; \tag{57}$$

b) there hold

$$\hat{\Gamma}_k(\cdot) \in \overline{\text{Conv}}\,(\mathbb{R}^n), \quad \hat{\Gamma}_k(\cdot) \leq \phi_m(\cdot; \hat{y}_{k-1}) \tag{58}$$

and

$$\phi_m(\hat{y}_k; \hat{y}_{k-1}) + \frac{1}{2\lambda}\|\hat{y}_k - \hat{y}_{k-1}\|^2 - \hat{\theta}_k \leq \hat{\delta}_k. \tag{59}$$

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 (56).

b) If $|C_k| = 1$, i.e., j_k is also the first iteration of cycle C_k, then (58) follows from the fact that $\hat{\Gamma}_k(\cdot) = \Gamma_{i_k}(\cdot) \in \mathcal{B}(\phi_m(\cdot; \hat{y}_k))$. If $|C_k| > 1$, then $\hat{\Gamma}_k = \Gamma_{j_k}$ and hence satisfies (58) in view of step 3a of PBF. Moreover, (59) follows from the logic of the prox-center update rule in step 3 of PBF (see (25)). ■

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*

$$\hat{\varepsilon}_k \geq 0, \qquad \hat{v}_k \in \partial_{\hat{\varepsilon}_k}[\phi_m(\cdot; \hat{y}_{k-1})](\hat{y}_k), \tag{60}$$

$$\hat{w}_k \in \partial_{\hat{\varepsilon}_k}[\phi_m(\cdot; \hat{y}_k)](\hat{y}_k). \tag{61}$$

Proof: Since $\hat{x}_k$ is an optimal solution of (56) in view of Lemma 4.5(a), using the optimality condition for (56), the fact that $\hat{\Gamma}_k \in \overline{\text{Conv}}(\mathbb{R}^n)$, and the definition of $\hat{v}_k$ in (27), we have $\hat{v}_k \in \partial \hat{\Gamma}_k(\hat{x}_k)$. This conclusion, (58), the definition of subdifferential in (3), and the definition of $\hat{\varepsilon}_k$ in (28), 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 (60) holds. The inequality in (60) 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 (60), 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 (61) holds. ■

It follows from (61) 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.

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*

$$\hat{\varepsilon}_k + \frac{1}{2\lambda}\|\hat{y}_k - \hat{x}_k\|^2 \leq 2\delta + \frac{1 + m\lambda}{2\lambda}\|\hat{y}_k - \hat{y}_{k-1}\|^2 \tag{62}$$

and

$$\|\hat{w}_k\|^2 \leq \frac{8\delta}{\lambda} + 4\left(m + \frac{1}{\lambda}\right)^2 \|\hat{y}_k - \hat{y}_{k-1}\|^2 \tag{63}$$

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 (28) and (27), respectively, we conclude that

$$\begin{aligned}
\hat{\varepsilon}_k &= \phi_m(\hat{y}_k; \hat{y}_{k-1}) - \hat{\Gamma}_k(\hat{x}_k) - \langle \hat{v}_k, \hat{y}_k - \hat{x}_k \rangle \\
&\overset{(59)}{\leq} \hat{\delta}_k - \frac{1}{2\lambda}\|\hat{y}_k - \hat{y}_{k-1}\|^2 + \hat{\theta}_k - \hat{\Gamma}_k(\hat{x}_k) - \langle \hat{v}_k, \hat{y}_k - \hat{x}_k \rangle \\
&\overset{(57)}{=} \hat{\delta}_k - \frac{1}{2\lambda}\|\hat{y}_k - \hat{y}_{k-1}\|^2 + \frac{1}{2\lambda}\|\hat{x}_k - \hat{y}_{k-1}\|^2 - \langle \hat{v}_k, \hat{y}_k - \hat{x}_k \rangle \\
&\overset{(27)}{=} \hat{\delta}_k - \frac{1}{2\lambda}\|\hat{y}_k - \hat{y}_{k-1}\|^2 + \frac{1}{2\lambda}\|\hat{x}_k - \hat{y}_{k-1}\|^2 + \frac{1}{\lambda}\langle \hat{x}_k - \hat{y}_{k-1}, \hat{y}_k - \hat{x}_k \rangle \\
&= \hat{\delta}_k - \frac{1}{2\lambda}\|\hat{y}_k - \hat{x}_k\|^2.
\end{aligned} \tag{64}$$

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

$$\begin{aligned}
\|\hat{w}_k\|^2 &\overset{(24)}{\leq} \frac{2}{\lambda^2}\|\hat{y}_k - \hat{x}_k\|^2 + 2\left(m + \frac{1}{\lambda}\right)^2 \|\hat{y}_k - \hat{y}_{k-1}\|^2 \\
&\overset{(64)}{\leq} \frac{4\hat{\delta}_k}{\lambda} + 2\left(m + \frac{1}{\lambda}\right)^2 \|\hat{y}_k - \hat{y}_{k-1}\|^2 \\
&\overset{(24)}{=} \frac{4\delta}{\lambda} + \frac{1}{2(m\lambda + 1)}\|\hat{w}_k\|^2 + 2\left(m + \frac{1}{\lambda}\right)^2 \|\hat{y}_k - \hat{y}_{k-1}\|^2
\end{aligned}$$

where the first inequality is due to the relation $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$. Inequality (63) now follows from the above inequality and simple algebraic manipulation. Using (63), (64) and the definition of $\hat{\delta}_k$ in step 3b of PBF, we have

$$
\begin{aligned}
\hat{\varepsilon}_k + \frac{1}{2\lambda}\|\hat{y}_k - \hat{x}_k\|^2 &\overset{(24),(64)}{\leq} \delta + \frac{\lambda}{8(m\lambda+1)}\|\hat{w}_k\|^2 \\
&\overset{(63)}{\leq} \delta + \frac{\lambda}{8(m\lambda+1)}\left(\frac{8\delta}{\lambda} + 4\left(m + \frac{1}{\lambda}\right)^2 \|\hat{y}_k - \hat{y}_{k-1}\|^2\right) \\
&\leq 2\delta + \frac{1+m\lambda}{2\lambda}\|\hat{y}_k - \hat{y}_{k-1}\|^2,
\end{aligned}
$$

and thus (62) 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 (23) in step 1 of PBF, it follows that $y_j = x_j$ and $\Psi(y_{j-1}) > \Psi(x_j)$, and hence that

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

which yields the desired contradiction.
b) Since $\hat{y}_k = \hat{y}_{k-1}$, (63) implies

$$
\|\hat{w}_k\|^2 \leq \frac{8\delta}{\lambda} \leq \frac{\bar{\eta}^2}{16} \tag{65}
$$

where the last inequality is due the definition of δ in step 0 of PBF. Observe that (62) implies that

$$
\hat{\varepsilon}_k \leq 2\delta \leq \frac{\bar{\varepsilon}}{8} \tag{66}
$$

where the last inequality is due the definition of δ in step 0 of PBF. Finally, (61), (65), (66) and Definition 2.5 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

$$
\phi(\hat{y}_k) + \frac{1}{2}\left(\frac{1}{\lambda} + m\right)\|\hat{y}_k - \hat{y}_{k-1}\|^2 \leq \phi(\hat{y}_{k-1}) \tag{67}
$$

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

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

$$
\frac{1}{2}\left(\frac{1}{\lambda} + m\right)\sum_{l=2}^{k}\|\hat{y}_l - \hat{y}_{l-1}\|^2 \leq \hat{M}^\lambda(\hat{x}_0) - \phi(\hat{y}_k) + 2\delta. \tag{68}
$$

Proof: a) It follows from (5) and (31) 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 (67).
b) Summing (67) from $k=2$ to k, we have

$$\phi(\hat{y}_k) + \frac{1}{2}\left(\frac{1}{\lambda} + m\right)\sum_{l=2}^{k} \|\hat{y}_l - \hat{y}_{l-1}\|^2 \le \phi(\hat{y}_1). \tag{69}$$

Using (59) with $k=1$, the definitions of $\hat{M}^\lambda(\hat{x}_0)$ and θ_j in (9) and step 1 of PBF , respectively, and the fact that $\hat{\Gamma}_1(\cdot) \le \phi_m(\cdot;\hat{x}_0)$, we can conclude that

$$\begin{aligned}\phi(\hat{y}_1) + \frac{1}{2}\left(\frac{1}{\lambda} + m\right)\|\hat{y}_1 - \hat{y}_0\|^2 &\overset{(59)}{\le} \hat{\theta}_1 + \hat{\delta}_1 \overset{(9)}{\le} \hat{M}^\lambda(\hat{x}_0) + \hat{\delta}_1 \\ &\overset{(24)}{=} \hat{M}^\lambda(\hat{x}_0) + \delta + \frac{\lambda}{8(m\lambda+1)}\|\hat{w}_1\|^2,\end{aligned}$$

where the identity is due to (24). It thus follows from (63) that

$$\phi(\hat{y}_1) + \frac{1}{2}\left(\frac{1}{\lambda} + m\right)\|\hat{y}_1 - \hat{y}_0\|^2 \overset{(63)}{\le} \hat{M}^\lambda(\hat{x}_0) + 2\delta + \frac{1}{2}\left(\frac{1}{\lambda} + m\right)\|\hat{y}_1 - \hat{y}_0\|^2,$$

which together with (69) implies that (68). ∎

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*

$$K = K(\bar{\eta},\bar{\varepsilon}) := \left\lceil 2\left(\frac{1}{\bar{\varepsilon}} + \frac{8(m\lambda+1)}{\lambda\bar{\eta}^2}\right)(\hat{M}^\lambda(\hat{x}_0) - \phi^*) + 1\right\rceil \tag{70}$$

where $\hat{M}^\lambda(\cdot)$ is as in (9). Then, PBF with δ as in (22) generates an iteration index $k \le K(\bar{\eta},\bar{\varepsilon})$ such that

$$\|\hat{w}_k\| \le \bar{\eta}, \quad \hat{\varepsilon}_k \le \bar{\varepsilon}. \tag{71}$$

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

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

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

and thus

$$\frac{\|\hat{w}_k\|^2}{\bar{\eta}^2} + \frac{\hat{\varepsilon}_k}{\bar{\varepsilon}} > 1 \quad \forall k = 1,\ldots,K. \tag{72}$$

Observe that the definitions of δ and K in (22) and (70), respectively, imply that

$$\left(\frac{4}{\bar{\varepsilon}} + \frac{16(m\lambda+2)}{\lambda\bar{\eta}^2}\right)\delta \le \frac{1}{2}, \quad \frac{1}{K-1}\left(\frac{1}{\bar{\varepsilon}} + \frac{8(m\lambda+1)}{\lambda\bar{\eta}^2}\right)(\hat{M}^\lambda(\hat{x}_0) - \phi^*) \le \frac{1}{2}. \tag{73}$$

Moreover, using (62), (63) and (72), we have

$$\begin{aligned}1 \overset{(72)}{<} \frac{\|\hat{w}_k\|^2}{\bar{\eta}^2} + \frac{\hat{\varepsilon}_k}{\bar{\varepsilon}} &\le \frac{2\delta + (1+m\lambda)\|\hat{y}_k - \hat{y}_{k-1}\|^2/(2\lambda)}{\bar{\varepsilon}} + \frac{8\delta + 4(m\lambda+1)^2\|\hat{y}_k - \hat{y}_{k-1}\|^2/\lambda}{\lambda\bar{\eta}^2} \\ &= \frac{2\delta}{\bar{\varepsilon}} + \frac{8\delta}{\lambda\bar{\eta}^2} + \left(\frac{1+m\lambda}{2\lambda\bar{\varepsilon}} + \frac{4(m\lambda+1)^2}{\lambda^2\bar{\eta}^2}\right)\|\hat{y}_k - \hat{y}_{k-1}\|^2.\end{aligned}$$

Summing the above inequality from $k = 2$ to K, dividing the resulting inequality by $K - 1$, and using (68), we have

$$\begin{aligned} 1 &< \frac{2\delta}{\bar\varepsilon} + \frac{8\delta}{\lambda\bar\eta^2} + \left(\frac{1+m\lambda}{2\lambda\bar\varepsilon} + \frac{4(m\lambda+1)^2}{\lambda^2\bar\eta^2}\right)\frac{1}{K-1}\sum_{k=2}^{K}\|\hat y_k - \hat y_{k-1}\|^2 \\ &\overset{(68)}{\le} \frac{2\delta}{\bar\varepsilon} + \frac{8\delta}{\lambda\bar\eta^2} + \left(\frac{1+m\lambda}{2\lambda\bar\varepsilon} + \frac{4(m\lambda+1)^2}{\lambda^2\bar\eta^2}\right)\frac{2\lambda}{(\lambda m+1)(K-1)}(\hat M^\lambda(\hat x_0) - \phi(\hat y_K) + 2\delta) \\ &\le \left(\frac{4}{\bar\varepsilon} + \frac{16(m\lambda+2)}{\lambda\bar\eta^2}\right)\delta + \frac{1}{K-1}\left(\frac{1}{\bar\varepsilon} + \frac{8(m\lambda+1)}{\lambda\bar\eta^2}\right)(\hat M^\lambda(\hat x_0) - \phi^*) \le 1 \end{aligned}$$

where the second last inequality is due to the fact that $\phi(\hat y_K) \ge \phi^*$ and the last inequality is due to (73). Hence, the above inequality gives the desired contradiction and thus the statement is proved. ∎

Before ending this subsection, we observe that the quantity $\hat M^\lambda(\hat x_0) - \phi^*$ in (70) 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 (9).

4.3 Proof of Theorem 3.4

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 term of σ_k. The following result refines the latter result by providing a uniform bound on σ_k.

Lemma 4.11 *Define*

$$\bar t := M^2 + \beta_2\left(\beta_1[\hat M^\lambda(\hat x_0) - \phi^* + 2\delta] + 4\zeta\lambda M^2\right), \tag{74}$$

where δ is as in (22), ζ is as in (36) , and β_1 and β_2 are as in (35). Then, the following statements hold for every $k \ge 1$:

a) $t_{i_k} \le \bar t$;

b) $\sigma_k \le \bar\sigma$ where σ_k and $\bar\sigma$ are as in (54) and (34), respectively.

Proof: a) It follows from (25) with $j = i_k$ and the definition of θ_j in step 1 of PBF that

$$\begin{aligned} t_{i_k} &\overset{(25)}{=} \phi_m(y_{i_k};\hat y_{k-1}) + \frac{1}{2\lambda}\|y_{i_k} - \hat y_{k-1}\|^2 - \theta_{i_k} \\ &= \phi_m(y_{i_k};\hat y_{k-1}) + \frac{1}{2\lambda}\|y_{i_k} - \hat y_{k-1}\|^2 - \Gamma_{i_k}(x_{i_k}) - \frac{1}{2\lambda}\|x_{i_k} - \hat y_{k-1}\|^2. \end{aligned} \tag{75}$$

Noting from the definition of y_j given below (23) that

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

Hence, the above inequality with $j = i_k$ and (75) imply that

$$t_{i_k} \le \phi_m(x_{i_k};\hat y_{k-1}) - \Gamma_{i_k}(x_{i_k}) \overset{(22),(29)}{\le} f_m(x_{i_k};\hat y_{k-1}) - \ell_f(x_{i_k};\hat y_{k-1}),$$

where the second inequality follows from (22), (29), and the definition of ϕ_m in (5). Moreover, applying (17) with $z = \hat y_{k-1}$, we have

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

Combining the above two inequalities, we have

$$t_{i_k} \le 2M\|x_{i_k} - \hat y_{k-1}\| + \frac{L+m}{2}\|x_{i_k} - \hat y_{k-1}\|^2 \le M^2 + \left(\frac{L+m}{2} + 1\right)\|x_{i_k} - \hat y_{k-1}\|^2, \tag{76}$$

where the last inequality is due to the fact that $2ab \le a^2 + b^2$ 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 (68) that

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

which, in view of the definition of β_1 in (35), is equivalent to

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

It is easy to verify that $\beta_1 \ge 1$ from its definition in (35). Clearly, this observation, the fact that $\phi(\hat{y}_k) \ge \phi^*$, and the above inequality imply that

$$\phi(\hat{y}_k) - \phi^* + \frac{1}{2}\left(m + \frac{2}{\zeta\lambda}\right)\|\hat{y}_k - \hat{y}_{k-1}\|^2 \le \beta_1[\hat{M}^\lambda(\hat{x}_0) - \phi^* + 2\delta]. \tag{77}$$

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

$$\begin{aligned}\zeta^2\left(\frac{1}{4\zeta\lambda} + \frac{m}{2}\right)\|x_{i_k} - \hat{y}_{k-1}\|^2 &\overset{(100)}{\le} \phi(\hat{y}_k) - \phi^* + \frac{1}{2}\left(m + \frac{2}{\zeta\lambda}\right)\|\hat{y}_k - \hat{y}_{k-1}\|^2 + 4\zeta\lambda M^2 \\ &\overset{(77)}{\le} \beta_1[\hat{M}^\lambda(\hat{x}_0) - \phi^* + 2\delta] + 4\zeta\lambda M^2.\end{aligned}$$

The statement now follows by combining (76) and the above inequality, and using the definitions of β_2 and $\bar{t}$ in (35) and (74), respectively.

b) This statement follows from a) and the definitions of σ_k and $\bar{\sigma}$ in (54) and (34), respectively. ∎

We end this section by noting that the proof of Theorem 3.4 now follows immediately from Propositions 4.4 and 4.10, and Lemma 4.11(b).

5 Computational Results

This section reports the computational results for the PBF method of Subsection 3.3 against the PS method of Subsection 3.2. It contains two subsections. The first one presents the computational results for the phase retrieval problem. The second one showcases the computational results for the blind deconvolution problem.

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

$$f(x) = g(c(x)) \tag{78}$$

for some convex and L-Lipschitz function g and smooth map c with β-Lipschitz Jacobian. It follows from Lemma 4.2 in [7] 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 2.6) which is hard to verify in view of the remark below Proposition 3.2. 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 2.5). Hence for the sake of our numerical experiments, we adopt the termination criterion

$$\phi(x_k) - \phi_* \le \tilde{\varepsilon}[\phi(x_0) - \phi_*] \tag{79}$$

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 two subsections. We implement the constant stepsize version of PS method, i.e., $\alpha_t = \alpha$ for every $t \ge 0$. 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 3.2 where T is the total number of PS iterations; (ii) T is determined by (79) rather than being pre-specified. PBF sets $\lambda = 1/2m$ and $\delta = \tilde{\varepsilon}[\phi(x_0) - \phi_*]$. Two variants of PBF are implemented: one based on the 2-cut scheme and the other one based on the multi-cut scheme (see Subsection 3.4). Due to the fact that $h = 0$, the subproblem (21) in the 2-cut scheme has a closed-form solution . In the multi-cut scheme, we utilize Mosek 10.2 [1] to solve subproblem (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 on the phase retrieval problem.

Consider the Phase Retrieval problem (see Subsection 5.1 of [4])

$$\min_{x\in\mathbb{R}^d}\left\{f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x)\right\},\quad f_i(x) := \left|\langle a_i, x\rangle^2 - b_i\right| \tag{80}$$

where $a_i \in \mathbb{R}^d$ and $b_i \in \mathbb{R}$. Observe that $f_i(\cdot)$ is a special case of (78) with $g(\cdot) = |\cdot|$ and $c(\cdot) = \langle a_i, \cdot\rangle^2 - b_i$. Thus f_i is $2\|a_i a_i^T\|$-weakly convex, and we set

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

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 $i = 1, \ldots, n$. Thus the optimal value $\phi_* = 0$ for (80). 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 (rounded) 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 two hours time limit.

-	$L_1 : (100, 300)$	$L_2 : (200, 600)$	$L_3 : (500, 1500)$	$L_4 : (1000, 3000)$
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: Twocut	5088/0.16	24486/2.08	316782/154.21	87514/242.62
PBF: Multicut	4716/0.19	21345/2.31	278234/172.32	723475/274.38

Table 1: PS versus two variants of PBF on (80). 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: Twocut	15228/0.42	125603/10.61	1056034/454.93	742553/1.86E+03
PBF: Multicut	13762/0.56	104572/12.41	923202/511.21	621001/2.02E+03

Table 2: PS versus two variants of PBF on (80). A relative tolerance of $\tilde{\varepsilon} = 10^{-4}$ is set.

[1] https://docs.mosek.com/latest/toolbox/index.html

The two tables show that the PS method is sensitive to the choice of stepsize, and the optimal stepsize in terms of CPU running time varies with different tolerances. In contrast, PBF with one stepsize demonstrates performance comparable to the best-performing stepsize of PS across all four stepsizes for both tolerance levels.

5.2 Blind deconvolution

This subsection reports computational results comparing PBF against PS on the blind deconvolution problem. Consider the blind deconvolution problem (see Subsection 5.2 of [4])

$$\min_{x,y} \left\{ f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x) \right\}, \quad f_i(x) := |\langle u_i, x\rangle \langle v_i, y\rangle - b_i| \tag{81}$$

where $u_i, v_i \in \mathbb{R}^d$ and $b_i \in \mathbb{R}$. Note that (81) is a special instance of (1) with $h \equiv 0$. Observe that $f_i(\cdot)$ is a special case of (78) 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 $\|u_i v_i^T\|$-weakly convex, and we set

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

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 $i = 1, \ldots, n$. Thus the optimal value $\phi_* = 0$ for (80). 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: Twocut	5021/0.39	344221/72.68	101299/280.67	77511/212.78
PBF: Multicut	4396/0.45	293115/80.18	93231/310.45	69932/232.65

Table 3: PS versus two variants of PBF on (81). 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: Twocut	35227/1.21	394212/89.32	188823/690.21	1142572/2.65E+03
PBF: Multicut	31444/1.51	378932/111.34	177012/739.32	1023990/2.93E+03

Table 4: PS versus two variants of PBF on (81). A relative tolerance of $\tilde{\varepsilon} = 10^{-4}$ is set.

The two tables above show that the PS method is sensitive to the choice of stepsize, with the optimal stepsize for CPU running time varying depending on the tolerance. On the other hand, PBF performs comparably to the best-performing PS stepsize across all four choices, for both tolerance levels.

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 [5] presents an inexact proximal point method (see Algorithm 2 of [5]) for solving (1) where the subgradient method (SM) is used to obtain an approximate solution $\hat{x}_k$ of (32). More specifically, under the condition that $j_k \ge 11(1+\lambda m)^2$, it is shown that j_k iterations of the SM initialized from $\hat{y}_{k-1}$ yield a point $\hat{x}_k$ satisfying

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

where x_k^* is the minimizer of (32). 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 (10), we conclude that

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

and hence that

$$\inf_{k\le K}\|\nabla \hat{M}^\lambda(\hat{y}_{k-1})\|^2 \le \frac{4(1+\lambda m)^2}{\lambda K}[\phi(x_0) - \phi(\hat{x}_K)] + \frac{\varepsilon^2}{2}. \tag{82}$$

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 $j_k \ge \bar{j}$. In contrast, it is shown in Appendix E that the iteration sequence $\{(\hat{x}_k, \hat{y}_k)\}$ satisfies a bound similar to (82) even though it solves (32) only according to the stopping criterion (25), 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. Beck. *First-order methods in optimization*, volume 25. SIAM, 2017.

[3] Rafael Correa and Claude Lemaréchal. Convergence of some algorithms for convex minimization. *Mathematical Programming*, 62(1):261–275, 1993.

[4] D. Davis and D. Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

[5] D. Davis and B. Grimmer. Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems. *SIAM Journal on Optimization*, 29(3):1908–1930, 2019.

[6] M. Díaz and B. Grimmer. Optimal convergence rates for the proximal bundle method. *SIAM Journal on Optimization*, 33(2):424–454, 2023.

[7] D. Drusvyatskiy and C. Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, 178:503–558, 2019.

[8] Y. Du and A. Ruszczyński. Rate of convergence of the bundle method. *Journal of Optimization Theory and Applications*, 173:908–922, 2017.

[9] A. Frangioni. Generalized bundle methods. *SIAM Journal on Optimization*, 13(1):117–156, 2002.

[10] 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.

[11] W. Hare and C. Sagastizábal. Computing proximal points of nonconvex functions. *Mathematical Programming*, 116(1-2):221–258, 2009.

[12] W. Hare and C. Sagastizábal. A redistributed proximal bundle method for nonconvex optimization. *SIAM Journal on Optimization*, 20(5):2442–2473, 2010.

[13] 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.

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

[15] K. C. Kiwiel. *Methods of descent for nondifferentiable optimization*, volume 1133. Springer, 2006.

[16] A Ya Kruger. On fréchet subdifferentials. *Journal of Mathematical Sciences*, 116(3):3325–3358, 2003.

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

[18] 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.

[19] 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.

[20] M. M. Makela and P. Neittaanmaki. *Nonsmooth optimization: analysis and algorithms with applications to optimal control*. World Scientific, 1992.

[21] 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.

[22] 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.

[23] 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.

[24] A. Ruszczyński. *Nonlinear optimization*. Princeton university press, 2011.

[25] J.-B. H. Urruty and C. Lemaréchal. *Convex analysis and minimization algorithms II*. Springer-Verlag, 1993.

[26] 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.

[27] 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 $\phi : \mathbb{R}^n \to \mathbb{R}\cup\{\infty\}$ is an m-weakly convex function, then $\phi_m(\cdot;c)$ is convex for every $c \in \mathbb{R}^n$. Moreover, for every $x \in \operatorname{dom}\phi$, $c \in \mathbb{R}^n$, and $\varepsilon \geq 0$, we have*

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

Proof: Let $x \in \operatorname{dom}\phi$, $c \in \mathbb{R}^n$, 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). ■

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

Lemma A.2 *If g is a closed 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*

$$0 \in \partial g(\hat{y}), \quad \|y - \hat{y}\| \leq \sqrt{\frac{2\varepsilon}{\mu}}. \tag{83}$$

Proof: The inclusion in (83) 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 (83) follows. ■

B Relationships Between Notions of Stationary Points

This section contains three proofs for the results in Section 2.

Proof of Proposition 2.7 For this proof only, we let

$$\psi(\cdot) := \phi(\cdot) + \frac{\lambda^{-1}+m}{2}\|\cdot - x\|^2 \tag{84}$$

and denote $\hat{x}^\lambda(x)$ as in (11). Observe that ψ is a $(1/\lambda)$-strongly convex function in view of the fact that ϕ is m-weakly convex. Thus, for any $u \in \operatorname{dom}\psi$, it holds that

$$\psi(u) - \psi(\hat{x}) \geq \frac{1}{2\lambda}\|u-\hat{x}\|^2. \tag{85}$$

We start with the proof of a).

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

$$\|x - \tilde{x}\| \leq \delta_D \qquad \inf_{\|d\|\leq 1} \phi'(\tilde{x};d) \geq -\varepsilon_D. \tag{86}$$

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

$$\psi(\hat{x}) - \psi(\tilde{x}) \geq \psi'(\tilde{x}; \hat{x} - \tilde{x}) \geq - \left[\varepsilon_D + \left(\frac{1}{\lambda} + m\right)\delta_D\right] \|\hat{x} - \tilde{x}\|. \tag{87}$$

Then using the above inequality and (85) 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 (86) 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 (10) we can get

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

Now the statement follows from the above inequality and the definition of Moreau stationary point in Definition 2.6.

b) Since x is a $(\varepsilon_M; \lambda)$-Moreau stationary point, (10) 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. ■

Proof of Proposition 2.8 Fix $x \in \operatorname{dom}\phi$. For simplicity, we denote

$$\phi_m(\cdot) = \phi_m(\cdot; x) \overset{(5)}{=} \phi(\cdot) + \frac{m}{2}\|\cdot - x\|^2. \tag{88}$$

Since ϕ_m is convex, it follows from Theorem 3.26 of [2] that

$$-\inf_{\|d\|\leq 1}\phi_m'(x; d) = -\inf_{\|d\|\leq 1}\sigma_{\partial\phi_m(x)}(d) = -\inf_{\|d\|\leq 1}\sup_{s\in\partial\phi_m(x)}\langle s, d\rangle = \inf_{s\in\partial\phi_m(x)}\|s\| = \operatorname{dist}(0; \partial\phi_m(x)). \tag{89}$$

In view of (88), Proposition 2.4 shows that $\partial\phi(x) = \partial\phi_m(x)$, which together with (89) implies that

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

Moreover, it follows from Definition 2.1 and (88) that

$$\phi_m'(x; d) = \liminf_{t\downarrow 0}\frac{\phi_m(x + td) - \phi_m(x)}{t} \overset{(88)}{=} \liminf_{t\downarrow 0}\frac{\phi(x + td) + \frac{m}{2}\|td\|^2 - \phi(x)}{t} = \phi'(x; d).$$

Therefore, the above two relations indicate that

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

and the conclusion directly follows from Definition 2.2. ■

Proof of Proposition 2.9 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 (8). Using the fact that $\phi_m(x;x) = \phi_{2m}(x;x)$ and $\phi_m(\cdot;x) \le \phi_{2m}(\cdot;x)$, we easily see that the inclusion in (8) 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

$$w \in \partial\left[\phi_{2m}(\cdot;x)\right](\tilde{x}), \quad \|x - \tilde{x}\| \le \sqrt{\frac{2\varepsilon}{m}} \le \sqrt{\frac{2\bar{\varepsilon}}{m}} \tag{90}$$

where the last inequality is due to the last inequality in (8). Now, letting $\hat{w} := w - 2m(\tilde{x} - x)$, it follows from (7) with $m = 2m$ and Lemma A.1 with $(\varepsilon, x, c) = (0, \tilde{x}, x)$ that

$$\hat{w} \in \partial\left[\phi_{2m}(\cdot;\tilde{x})\right](\tilde{x}) = \partial\phi(\tilde{x}). \tag{91}$$

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

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

where the second inequality is due to the first inequality in (8) and the inequality in (90). Then the above inequality and (91) imply that

$$\operatorname{dist}(0, \partial\phi(\tilde{x})) \le \bar{\eta} + 2\sqrt{2m\bar{\varepsilon}}. \tag{92}$$

Now statement (a) follows from (90), (92), and Proposition 2.8.

b) Statement (b) immediately follows from the first statement and Proposition 2.7(a) with $\lambda = 1/m$. ■

C Important Results about Recursive Formula

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

$$q_j > \delta \quad \forall j \in \{0, \ldots, J\}, \tag{93}$$

and

$$q_j\left(1 + \frac{q_j}{a + bq_j}\right) \le q_{j-1}, \quad \forall j \in \{1, \ldots, J\}. \tag{94}$$

Then,

$$J \le \bar{N} := (2b + 3)\log^+\left(\frac{q_0}{\delta + a/(b+1)}\right) + \frac{2a}{\delta} + b + 2. \tag{95}$$

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

$$\bar{j} := \max\{j \le J : q_j \ge \tilde{\delta}\} \le \mathcal{T} := \frac{\log^+(q_0/\tilde{\delta})}{\log(1+\theta)} \le (1 + \theta^{-1})\log^+\left(\frac{q_0}{\tilde{\delta}}\right) \le \bar{N}. \tag{96}$$

Assume for contradiction that there exists an index $j > \mathcal{T}$ such that $q_j \ge \tilde{\delta}$. Using the fact $\{q_j\}$ is non-increasing (see (94)) and the definitions of $\tilde{\delta}$ and θ, we have that for every $l = 1, \ldots, j$,

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

Hence, in view of (94), we conclude that $(1+\theta)^j q_j \le (1+\theta)^{j-1} q_{j-1} \le \cdots \le q_0$. On the other hand, using the fact that $j > \mathcal{T}$ and the definition of $\mathcal{T}$ in (96), 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 $q_j \ge \tilde{\delta}$, the claim follows.

Note that the definition of $\bar{j}$ in (96) implies that $\tilde{\delta} > q_{\bar{j}+1} \geq q_j$ for any $j \geq \bar{j}+1$. This fact, the assumption that $a \geq 0$ and $b \geq 0$, and relation (94) then imply that

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

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

$$q_{\bar{N}} \leq \frac{a+(b+1)\tilde{\delta}}{\bar{N}-\bar{j}} = \frac{2a+(b+1)\delta}{\bar{N}-\bar{j}} \leq \delta \tag{97}$$

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 (95) and (96), respectively. Hence, the above inequality contradicts (93) and thus the statement follows. ■

D Technical Results for the Proof of Theorem 3.4

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 [19].

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*

$$z_\lambda = \underset{u\in\mathbb{R}^n}{\text{argmin}} \left\{\Gamma(u) + \frac{1}{2\lambda}\|u-z_0\|^2\right\}, \quad z_{\zeta\lambda} = \underset{u\in\mathbb{R}^n}{\text{argmin}} \left\{\Gamma(u) + \frac{1}{2\zeta\lambda}\|u-z_0\|^2\right\}. \tag{98}$$

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 $u \in \text{dom}\, h$, we have

$$\frac{1}{2\lambda}\|u-z_\lambda\|^2 + (\tilde{f}+h)(z_\lambda) - (\tilde{f}+h)(u) \leq \frac{1}{2\lambda}\|u-z_0\|^2 + \frac{2\lambda M^2}{1-\lambda L}. \tag{99}$$

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

$$\zeta^2\left(\frac{1}{4\zeta\lambda} + \frac{m}{2}\right)\|z_\lambda - z_0\|^2 \leq (f+h)(u) - (f+h)(z_{\zeta\lambda}) + \left(\frac{m}{2} + \frac{1}{\zeta\lambda}\right)\|u-z_0\|^2 + 4\zeta\lambda M^2, \tag{100}$$

where z_λ and $z_{\zeta\lambda}$ are as in (98) and ζ is defined as (36).

Proof: We first prove (100) under the assumption that $\lambda \in (0, (2(L+m))^{-1}]$, and hence $\zeta = 1$ in view of (36). 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 $u \in \text{dom}\, h$:

$$\frac{1}{2\lambda}\|u-z_\lambda\|^2 + (f_m(\cdot;z_0)+h)(z_\lambda) - (f_m(\cdot;z_0)+h)(u) - \frac{1}{2\lambda}\|u-z_0\|^2 \overset{(99)}{\leq} \frac{2\lambda M^2}{1-\lambda(L+m)} \leq 4\lambda M^2,$$

and hence that

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

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

$$(f + h)(u) - (f + h)(z_\lambda) + 4\lambda M^2 \geq \left(\frac{1}{4\lambda} + \frac{m}{2}\right)\|z_\lambda - z_0\|^2 - \left(\frac{1}{\lambda} + \frac{m}{2}\right)\|u - z_0\|^2 \tag{101}$$

which, in view of the fact that $\zeta = 1$, immediately implies (100). Next, we show that (100) also holds for $\lambda > 1/[2(L + m)]$. Noting that (36) implies that $\zeta \in (0, 1)$ and $\zeta\lambda = 1/[2(L + m)]$, it then follows from (101) 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, (100) immediately follows from the above inequality and Lemma D.1. ∎

E Further Discussion Regarding (82)

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 (32). Hence, using the definition of $\phi_m(\cdot; \hat{y}_{k-1})$ in (5) and the fact that the objective function of (32) and is λ^{-1}-strongly convex, we have

$$\phi_m(\hat{y}_k; \hat{y}_{k-1}) + \frac{1}{2\lambda}\|\hat{y}_k - \hat{y}_{k-1}\|^2 \geq \phi_m(x_k^*; \hat{y}_{k-1}) + \frac{1}{2\lambda}\|x_k^* - \hat{y}_{k-1}\|^2 + \frac{1}{2\lambda}\|\hat{y}_k - x_k^*\|^2. \tag{102}$$

Moreover, using Lemma 4.5(a) and the fact that the objective function in (56) is λ^{-1}-strongly convex, we for every $u \in \operatorname{dom} h$,

$$\hat{\Gamma}_k(\hat{x}_k) + \frac{1}{2\lambda}\|\hat{x}_k - \hat{y}_{k-1}\|^2 \leq \hat{\Gamma}_k(u) + \frac{1}{2\lambda}\|u - \hat{y}_{k-1}\|^2 - \frac{1}{2\lambda}\|u - \hat{x}_k\|^2. \tag{103}$$

Combining (57) and (59), and using (103), we have for every $u \in \operatorname{dom} h$,

$$\begin{aligned}
\hat{\delta}_k &\overset{(57),(59)}{\geq} \phi_m(\hat{y}_k; \hat{y}_{k-1}) + \frac{1}{2\lambda}\|\hat{y}_k - \hat{y}_{k-1}\|^2 - \hat{\Gamma}_k(\hat{x}_k) - \frac{1}{2\lambda}\|\hat{x}_k - \hat{y}_{k-1}\|^2 \\
&\overset{(103)}{\geq} \phi_m(\hat{y}_k; \hat{y}_{k-1}) + \frac{1}{2\lambda}\|\hat{y}_k - \hat{y}_{k-1}\|^2 - \hat{\Gamma}_k(u) - \frac{1}{2\lambda}\|u - \hat{y}_{k-1}\|^2 + \frac{1}{2\lambda}\|u - \hat{x}_k\|^2 \\
&\overset{(58)}{\geq} \phi_m(\hat{y}_k; \hat{y}_{k-1}) + \frac{1}{2\lambda}\|\hat{y}_k - \hat{y}_{k-1}\|^2 - \phi_m(u; \hat{y}_{k-1}) - \frac{1}{2\lambda}\|u - \hat{y}_{k-1}\|^2 + \frac{1}{2\lambda}\|u - \hat{x}_k\|^2,
\end{aligned} \tag{104}$$

where the last inequality is due to the inequality in (58). Taking $u = \hat{y}_{k-1}$ in (104) yields

$$\hat{\delta}_k + \phi(\hat{y}_{k-1}) - \phi(\hat{y}_k) \geq \hat{\delta}_k + \phi(\hat{y}_{k-1}) - \phi_m(\hat{y}_k; \hat{y}_{k-1}) \geq \frac{1}{2\lambda}\|\hat{y}_k - \hat{y}_{k-1}\|^2 + \frac{1}{2\lambda}\|\hat{x}_k - \hat{y}_{k-1}\|^2. \tag{105}$$

Taking $u = x_k^*$ in (104) and using (102), we have

$$\hat{\delta}_k \geq \frac{1}{2\lambda}\|\hat{y}_k - x_k^*\|^2 + \frac{1}{2\lambda}\|x_k^* - \hat{x}_k\|^2. \tag{106}$$

Combining (105) and (106), we obtain

$$
\begin{aligned}
2\hat{\delta}_k + \phi(\hat{y}_{k-1}) - \phi(\hat{y}_k) &\geq \left(\frac{1}{2\lambda}\|\hat{x}_k - \hat{y}_{k-1}\|^2 + \frac{1}{2\lambda}\|x_k^* - \hat{x}_k\|^2\right) + \left(\frac{1}{2\lambda}\|\hat{y}_k - \hat{y}_{k-1}\|^2 + \frac{1}{2\lambda}\|\hat{y}_k - x_k^*\|^2\right) \\
&\geq \frac{1}{4\lambda}\|x_k^* - \hat{y}_{k-1}\|^2 + \frac{1}{4\lambda}\|x_k^* - \hat{y}_{k-1}\|^2 = \frac{1}{2\lambda}\|x_k^* - \hat{y}_{k-1}\|^2, \qquad (107)
\end{aligned}
$$

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 (24), and using (63), we have

$$
\hat{\delta}_k \overset{(24)}{=} \delta + \frac{\lambda}{8(m\lambda+1)}\|\hat{w}_k\|^2 \overset{(63)}{\leq} 2\delta + \frac{1}{2}\left(m + \frac{1}{\lambda}\right)\|\hat{y}_k - \hat{y}_{k-1}\|^2 \overset{(67)}{\leq} 2\delta + \phi(\hat{y}_{k-1}) - \phi(\hat{y}_k),
$$

where the last inequality is due to (67). The above inequality and (107) 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 (32), and hence $x_k^* = \hat{x}^\lambda(\hat{y}_{k-1})$ in view of (11). It thus follows from (10) 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. ∎