Proximal bundle methods for hybrid weakly convex composite optimization problems

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

March 26, 2023

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) based on a generic bundle update scheme which includes various well-known bundle update schemes. In contrast to other well-known stationary conditions in the context of HWC-CO, PBF uses a new stationarity measure which is easily verifiable and, at the same time, implies any of the former ones.

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 $\operatorname{dom} f \supseteq \operatorname{dom} h$ and consider the composite optimization (CO) problem

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

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 $f' : \operatorname{dom} h \to \mathbb{R}^n$ (i.e., $f'(x) \in \partial f(x)$ for every $x \in \operatorname{dom} h$) satisfying the (M, L)-hybrid condition, namely: $\|f'(u) - f'(v)\| \le 2M + L\|u - v\|$ for every $u, v \in \operatorname{dom} h$. 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 [3] 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) based on a generic bundle update scheme which contains different well-known update schemes as special cases. More specifically, like other proximal bundle methods, a PBF iteration solves a prox bundle subproblem of the form

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

*Department of Computer Science, Yale University, New Haven, CT 06511 (email: jiaming.liang@yale.edu).

†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 Grant FA9550-22-1-0088.

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 which is novel in the context of HWC-CO.

Literature Review This paragraph discusses the development of proximal bundle methods in the context of convex CO problems. They have been proposed in [14, 29], and then further studied for example in [6, 19, 22]. Their convergence analyses for nonsmooth (i.e., $L = 0$ and $M > 0$) CO problems are broadly discussed for example in the textbooks [25, 27]. Moreover, their complexity analyses are studied for example in [4, 5, 11, 15, 16].

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 [7, 8, 9, 10, 12, 18, 20, 28]. In particular, papers [8, 9, 10] 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.

Iteration-complexities for stochastic proximal point and proximal subgradient methods are obtained in [3] using the Moreau envelope as an optimality measure. Moreover, [17, 21, 23, 24, 26] establish iteration-complexities for different types of algorithms for solving the special case where $f(x) = \max_{y\in Y} \Phi(x, y)$, function $x \to \Phi(x, y)$ is weakly convex and differentiable for every $y \in Y$, and $y \to \Phi(x, y)$ is concave for every x.

Main contribution. As already mentioned above, this paper establishes for the first time the (serious and overall) iteration-complexity of PB methods for solving the HWC-CO problem described at the beginning of this introduction. This is done in a unified manner by considering a generic bundle update scheme. An interesting feature of this analysis is the introduction of a new stationarity measure which, in contrast to other well-known stationary conditions in the context of HWC-CO (including the Moreau stationary one), is easily 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 δ-Moreau stationary point of ϕ is similar to that of the deterministic version of the stochastic proximal subgradient (PS) method studied in [3], i.e., $\mathcal{O}(\delta^{-4})$, but the first complexity bound has the major advantage that its constant (in its $\mathcal{O}(\cdot)$) is never worse and is generally much better than the one which appears in the bound for the PS method of [3]. 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).

Organization of the paper. Subsection 1.1 presents basic definitions and notation used throughout the paper. Section 2 contains two subsections. Subsection 2.1 provides a characterization of the subdifferential for a weakly convex function. Subsection 2.2 introduces three different notions of approximate stationary points and discusses their relationship. Section 3 contains four subsections. Subsection 3.1 formally describes the problem (1) and the assumptions made on it. Subsection 3.2 reviews the deterministic version of the stochastic composite subgradient method of [3]. Subsection 3.3 presents a generic bundle update scheme. Subsection 3.4 describes the PBF framework and states the iteration-complexity result 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 gives some concluding remarks and potential directions for future research. Appendix A describes two useful technical results about subdifferentials. Appendix B provides proofs for the results in Subsection 2.2. Finally, Appendix C presents some useful technical results used in Section 4.

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 and $\tilde{\mathcal{O}}_1(\cdot)$ denote $\mathcal{O}(\cdot + 1)$ with the convention that a logarithm factor is neglected.

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

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

for every $x, y \in \mathrm{dom}\,\varphi$ and $\alpha \in [0,1]$. Denote the set of all proper lower semicontinuous convex functions by $\overline{\mathrm{Conv}}\,(\mathbb{R}^n)$. For $\varepsilon \geq 0$, the *ε-subdifferential* of φ at $x \in \mathrm{dom}\,\varphi$ 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 $x \in \mathrm{dom}\,\varphi$, i.e., $\partial_0\varphi(x)$, is denoted by $\partial\varphi(x)$.

2 Basic definitions and background

This section contains two subsections. The first one gives the definition of the subdifferential and the directional derivative of a general closed function. It also relates these two concepts and provides a characterization of the first one for the case of a closed m-weakly convex function. The second subsection introduces the notion of a regularized stationary point which is sought by the main algorithm of this paper. It also introduces two different stationary conditions in terms of the directional derivative and the Moreau envelope, respectively, and provides the relationship between these three different notions.

2.1 Characterization of the subdifferential of weakly convex function

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 [13].

Definition 2.2 *The Frechet subdifferential of a proper closed function $\phi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is defined as*

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

The following result gives a relationship between the above two concepts.

Lemma 2.3 *For any $v \in \partial\phi(x)$ we have*

$$\inf_{\|d\| \leq 1} \phi'(x; d) \geq -\|v\|. \tag{4}$$

Proof: To prove (4), it suffices to show if $v \in \partial\phi(x)$, then $\phi'(x;d) \geq -\|v\|$ for every $d \in \mathbb{R}^n$ such that $\|d\| \leq 1$. Indeed, we may assume that $d \neq 0$ since the case where $d = 0$ is trivial. Then, it follows from the Definition 2.2 and the Cauchy-Schwarz inequality that

$$0 \leq \liminf_{t\downarrow 0} \frac{\phi(x+td) - \phi(x) - t\langle v, d\rangle}{t\|d\|} = -\frac{\langle v, d\rangle}{\|d\|} + \liminf_{t\downarrow 0} \frac{\phi(x+td) - \phi(x)}{t\|d\|}$$
$$= -\frac{\langle v, d\rangle}{\|d\|} + \frac{1}{\|d\|}\phi'(x;d) \leq \|v\| + \frac{1}{\|d\|}\phi'(x;d),$$

and hence that $\phi'(x;d) \geq -\|v\|$ for every $d \in \mathbb{R}^n$ such that $\|d\| \leq 1$. ∎

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

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

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 [3]. 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). ∎

2.2 Notions of approximate stationary points

This subsection introduces three notions of stationary points, including the one adopted by the main algorithm of this paper. It also describes how these notions are related to one another.

We start by introducing 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 now 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 $(\bar\eta, \bar\varepsilon)$-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 [15]), 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).

The following definition uses directional derivatives to define a different notion of a stationary point.

Definition 2.6 (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*

$$\|x - \tilde x\| \leq \delta_D, \quad \inf_{\|d\|\leq 1} \phi'(\tilde x; d) \geq -\varepsilon_D. \tag{9}$$

When $(\varepsilon_D, \delta_D) = (0, 0)$, then (9) 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 $0 \in \partial\phi(x)$.

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

$$\hat{M}^\lambda(x) := \min_u \left\{ \phi(u) + \frac{\lambda^{-1} + m}{2} \|u - x\|^2 \right\}. \tag{10}$$

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 [3]) is as follows:

$$\nabla \hat{M}^\lambda(x) = \left(\frac{1}{\lambda} + m \right) \left(x - \hat{x}^\lambda(x) \right) \tag{11}$$

where

$$\hat{x}^\lambda(x) := \operatorname{argmin}_{u \in \mathbb{R}^n} \left\{ \phi(u) + \frac{\lambda^{-1} + m}{2} \|u - x\|^2 \right\}. \tag{12}$$

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

Definition 2.7 (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 $\|\nabla \hat{M}^\lambda(x)\| \leq \varepsilon_M$.*

We now make some remarks about the above three notions of stationary points. First, 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. Second, among the above three stationary notions, only the directional one imposes a condition on a nearby point instead of the actual point under consideration.

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

Proposition 2.8 *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 2.9 *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 $\tilde{x} \in \operatorname{dom}\phi$ such that*

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

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

Proposition 2.10 *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 composite subgradient (SCS) method of [3] and its main complexity result. The third one presents a generic bundle update scheme and describes three special cases of it. The fourth one describes the proximal bundle framework (PBF) and describes its main complexity result.

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 $h \in \overline{\text{Conv}}\,(\mathbb{R}^n)$, f is m-weakly convex and $\text{dom}\, h \subset \text{dom}\, f$;

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

$$\|f'(u) - f'(v)\| \leq 2M + L\|u - v\| \quad \forall u, v \in \text{dom}\, h; \tag{13}$$

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

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

$$\ell_g(\cdot; x) := g(x) + \langle g'(x), \cdot - x\rangle. \tag{14}$$

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

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

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 $x \in \text{dom}\, f$ and (13) holds. Since f is m-weakly convex, it follows from the definition of weak convexity that $f_m \in \overline{\text{Conv}}\,(\mathbb{R}^n)$. 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 $x \in \text{dom}\, f$ that $f'_m(\cdot; z) := f'(\cdot) + m(\cdot - z) \in \partial[f_m(\cdot; z)](\cdot)$. The result now follows by noting that (13) and the definition of $f_m + \chi/\lambda$ imply that for every $x, y \in \text{dom}\, f$

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

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 $\tilde{z} \in \text{dom}\, h$, function $f_m(\cdot; z)$ admits the linearization given by

$$\ell_{f_m(\cdot;z)}(\cdot; \tilde{z}) := f_m(\tilde{z}; z) + \langle f'(\tilde{z}) + m(\tilde{z} - z)\, , \cdot - \tilde{z}\rangle. \tag{15}$$

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

$$0 \leq f_m(\cdot; z) - \ell_{f_m(\cdot;z)}(\cdot; \tilde{z}) \leq 2M\| \cdot - \tilde{z}\| + \frac{L + m}{2}\| \cdot - \tilde{z}\|^2 \quad \forall z \in \mathbb{R}^n,\ \tilde{z} \in \text{dom}\, h. \tag{16}$$

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

$$\ell_{f_m(\cdot;z)}(\cdot; z) = \ell_f(\cdot; z), \tag{17}$$

and

$$0 \leq f_m(\cdot; z) - \ell_f(\cdot; z) \leq 2M\| \cdot - z\| + \frac{L + m}{2}\| \cdot - z\|^2 \quad \forall z \in \text{dom}\, h. \tag{18}$$

3.2 Review of the SCS method for the weakly convex case

This subsection reviews the deterministic version of the SCS method of [3].

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 $\{\alpha_t\}_{t\geq 0} \subset \mathbb{R}_+$ and iteration count T.
Step: $t = 0, 1, \cdots, T$:

$$\text{Set } \hat{x}_{t+1} = \operatorname{argmin}_{u\in\mathbb{R}^n} \left\{ \ell_f(u; \hat{x}_t) + h(u) + \frac{1}{2\alpha_t}\|u - \hat{x}_t\|^2 \right\}. \tag{19}$$

The following result (see Theorem 3.4 in [3]) 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 $t \geq 0$ for some $\bar{m} \in (m, 2m]$. Then, the iterates $\hat{x}_t$ of PS satisfies*

$$\frac{\sum_{t=0}^T \alpha_t \left\| \nabla \hat{M}^{1/(\bar{m}-m)}(\hat{x}_t) \right\|^2}{\sum_{t=0}^T \alpha_t} \leq \frac{\bar{m}}{\bar{m}-m} \cdot \frac{\left(\hat{M}^{1/(\bar{m}-m)}(\hat{x}_0) - \phi^*\right) + 2\bar{m}M^2 \sum_{t=0}^T \alpha_t^2}{\sum_{t=0}^T \alpha_t}. \tag{20}$$

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

$$T \geq \frac{\left[\left(\hat{M}^{1/m}(\hat{x}_0) - \phi^*\right) + 4mM^2\gamma^2\right]^2}{\gamma^2\rho^4}, \tag{21}$$

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

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 PS is that the right hand side of (21) is generally not computable, and hence there is no clear way of choosing T satisfying (21). Third, the method of [3] actually outputs an iterate $\hat{x}_{t^*}$ where t^* is sampled from $\{0, \ldots, T\}$ according to the probability mass function $\mathbb{P}(t^* = t) = \alpha_t / \sum_{i=0}^T \alpha_i$ 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 (20) and hence is bounded by the right hand side of (20). 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 (20). However, there is no simple way of 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.

3.3 The generic bundle update scheme

As mentioned in the Introduction, PBF uses a bundle (convex) function underneath $\phi(\cdot) + m\| \cdot - x^c\|^2/2$, where x^c is the current prox center, to construct subproblem (2) at a given iteration, and then updates Γ to obtain the bundle function Γ^+ for the next iteration. This subsection describes ways of updating the bundle. Instead of focusing on a specific bundle update scheme, this subsection describes a generic update scheme. It also discusses three concrete ways of implementing the generic scheme.

We start by describing the generic bundle update scheme (GBUS) .

GBUS

Input: Pair (g,h) where $g,h \in \overline{\mathrm{Conv}}\,(\mathbb{R}^n)$, $(\lambda,\tau) \in \mathbb{R}_{++} \times (0,1)$, and $(x^c, x, \Gamma) \in \mathbb{R}^n \times \mathbb{R}^n \times \overline{\mathrm{Conv}}\,(\mathbb{R}^n)$ such that $\Gamma \le g+h$ and (2) holds;

- find function Γ^+ such that

$$\Gamma^+ \in \overline{\mathrm{Conv}}\,(\mathbb{R}^n), \qquad \tau\bar\Gamma(\cdot) + (1-\tau)[\ell_g(\cdot;x) + h(\cdot)] \le \Gamma^+(\cdot) \le (g+h)(\cdot), \tag{22}$$

  where $\ell_g(\cdot;\cdot)$ is as in (14) and $\bar\Gamma(\cdot)$ is such that

$$\bar\Gamma \in \overline{\mathrm{Conv}}\,(\mathbb{R}^n), \quad \bar\Gamma(x) = \Gamma(x), \quad x = \underset{u\in\mathbb{R}^n}{\mathrm{argmin}} \left\{ \bar\Gamma(u) + \frac{1}{2\lambda}\|u - x^c\|^2 \right\}. \tag{23}$$

Output: Γ^+.

We will now discuss three concrete ways of implementing GBUS.

(S1) **one-cut scheme**: This scheme sets $\Gamma^+ = \Gamma^+_\tau := \tau\Gamma + (1-\tau)[\ell_g(\cdot;x) + h]$ where x is as in (23).

(S2) **two-cut scheme:** Suppose $\Gamma = \max\{A_g, \ell_g(\cdot;x^-)\} + h$ where $A_g \le g$ is an affine function and x^- is the previous iterate. This scheme sets $\Gamma^+ = \max\{A_g^+, \ell_g(\cdot;x)\} + h$ where $A_g^+ = \theta A_g + (1-\theta)\ell_g(\cdot;x^-)$ for some $\theta \in [0,1]$.

(S3) **multiple-cut scheme:** Suppose $\Gamma = \Gamma(\cdot;C)$ where $C \subset \mathbb{R}^n$ is a finite set (i.e., the current bundle set) and $\Gamma(\cdot;C)$ is defined as

$$\Gamma(\cdot;C) := \max\{\ell_g(\cdot;c) : c \in C\} + h(\cdot). \tag{24}$$

This scheme chooses the next bundle set C^+ so that

$$C(x) \cup \{x\} \subset C^+ \subset C \cup \{x\} \tag{25}$$

where

$$C(x) := \{c \in C : \ell_g(x;c) + h(x) = \Gamma(x)\}, \tag{26}$$

and then output $\Gamma^+ = \Gamma(\cdot;C^+)$.

We now make some remarks to argue that all the update schemes above are special implementations of GBUS. It can be easily seen that the update Γ^+ in (S1), together with $\bar\Gamma = \Gamma$, satisfies (22) and (23), and hence that scheme (S1) is a special way of implementing GBUS. On the other hand, the results below, whose proof can be found in Appendix D in [16], show that (S2) and (S3) are also special implementations of GBUS.

(a) If Γ^+ is obtained according to (S2), then the pair $(\Gamma^+, \bar\Gamma)$ where $\bar\Gamma = A_g^+ + h$ satisfies (22) and (23).

(b) If Γ^+ is obtained according to (S3), then the pair $(\Gamma^+, \bar\Gamma)$ where $\bar\Gamma = \Gamma(\cdot;C(x))$ satisfies (22) and (23).

3.4 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 SCS method of [3] 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

$$x = \underset{u\in\mathbb{R}^n}{\mathrm{argmin}} \left\{ \Gamma(u) + \frac{1}{2\lambda}\|u - \hat x_{k-1}\|^2 \right\} \tag{27}$$

where Γ is a relatively simple convex function minorizing the convexification $\phi_m(\cdot;\hat{x}_{k-1})$ of ϕ defined in (5). As any classical proximal bundle approach, it updates the center $\hat{x}_{k-1}$ 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).

PBF

---

0. Let initial point $\hat{x}_0 \in \operatorname{dom} h$, tolerance pair $(\bar{\eta}, \bar{\varepsilon}) \in \mathbb{R}^2_{++}$, cycle tolerance $\delta > 0$, prox stepsize $\lambda > 0$, and parameters $\chi \in (0,1]$ and $\tau \in (0,1)$ such that

$$\frac{\tau}{1-\tau} \geq \lambda\left(\frac{4M^2}{\delta} + L + m\right), \qquad \chi(2 + m\lambda) - 1 > 0, \tag{28}$$

be given, and set $y_0 = \hat{x}_0$, $j = 1$, $k = 1$, and

$$\Gamma_1(\cdot) = \ell_f(\cdot;\hat{x}_0) + h(\cdot), \qquad \tilde{m} = m + \frac{\chi}{\lambda}; \tag{29}$$

1. Compute

$$x_j = \underset{u\in\mathbb{R}^n}{\arg\min}\left\{\Gamma_j(u) + \frac{1}{2\lambda}\|u - \hat{x}_{k-1}\|^2\right\}, \tag{30}$$

$$\theta_j = \Gamma_j(x_j) + \frac{1}{2\lambda}\|x_j - \hat{x}_{k-1}\|^2, \tag{31}$$

choose $y_j \in \{x_j, y_{j-1}\}$ such that

$$\phi_{\tilde{m}}(y_j;\hat{x}_{k-1}) = \min\left\{\phi_{\tilde{m}}(x_j;\hat{x}_{k-1}), \phi_{\tilde{m}}(y_{j-1};\hat{x}_{k-1})\right\}; \tag{32}$$

2. If

$$t_j := \phi_{\tilde{m}}(y_j;\hat{x}_{k-1}) - \theta_j > \delta, \tag{33}$$

then go to step 2a; else, go to step 2b;

2a) (null update) let $\Gamma_{j+1} \in \overline{\text{Conv}}\,(\mathbb{R}^n)$ denote the output Γ^+ of GBUS with input (λ, τ, h), $g = f_m(\cdot;\hat{x}_{k-1})$, and $(x^c, x, \Gamma) = (\hat{x}_{k-1}, x_j, \Gamma_j)$, and go to step 3;

2b) (serious update) set $\hat{x}_k = x_j$, $\hat{y}_k = y_j$, $\hat{\Gamma}_k(\cdot) = \Gamma_j(\cdot)$, and

$$\hat{v}_k = \frac{1}{\lambda}(\hat{x}_{k-1} - \hat{x}_k), \tag{34}$$

and compute

$$\hat{w}_k = \hat{v}_k - m(\hat{y}_k - \hat{x}_{k-1}), \quad \hat{\varepsilon}_k = \phi_m(\hat{y}_k;\hat{x}_{k-1}) - \hat{\Gamma}_k(\hat{x}_k) - \langle \hat{v}_k, \hat{y}_k - \hat{x}_k\rangle \tag{35}$$

if $\|\hat{w}_k\| \leq \bar{\eta}$ and $\hat{\varepsilon}_k \leq \bar{\varepsilon}$, then stop; else find function $\Gamma_{j+1}(\cdot)$ such that

$$\Gamma_{j+1} \in \overline{\text{Conv}}\,(\mathbb{R}^n), \qquad \ell_f(\cdot;\hat{x}_k) + h(\cdot) \leq \Gamma_{j+1}(\cdot) \leq \phi_m(\cdot;\hat{x}_k) \tag{36}$$

where $\ell_f(\cdot;\cdot)$ is defined in (14), set $k \leftarrow k+1$, and go to step 3;

3. 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.,

$$\mathcal{C}_k := \{i_k, \ldots, j_k\}, \quad i_k := j_{k-1}+1. \tag{37}$$

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 GBUS is implemented nor how Γ_{j+1} in (36) is chosen. Second, in view of (36) or the fact that the output of GBUS satisfies (22), it follows that

$$\Gamma_j \le \phi_m(\cdot;\hat{x}_{k-1}), \qquad \Gamma_j \in \overline{\mathrm{Conv}}\,(\mathbb{R}^n), \qquad \forall j \in \mathcal{C}_k,\ k \ge 1. \tag{38}$$

Third, it is shown in Lemma 4.6 that $\hat{w}_k \in \partial_{\hat{\varepsilon}_k}\left[\phi_m(\cdot;\hat{y}_k)\right](\hat{y}_k)$ 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 2b. Fourth, in view of the definition of y_j in (32) and the above relation, it then follows that

$$y_j \in \mathrm{Argmin}\,\left\{\phi_m(x;\hat{x}_{k-1}) : x \in \{\hat{y}_{k-1}, x_{i_k}, \ldots, x_j\}\right\}.$$

Two simple ways of choosing the bundle function Γ_{j+1} such that (36) holds are to set $\Gamma_{j+1} = \ell_f(\cdot;\hat{x}_k)+h$ or $\Gamma_{j+1} = \max\{\tilde{\Gamma}_{j+1}, \ell_f(\cdot;\hat{x}_k)+h\}$, where

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

Clearly, the first choice for Γ_{j+1} satisfies (36). Moreover, using the fact that $\Gamma_j \le \phi_m(\cdot;\hat{x}_{k-1})$ and the definition of $\phi_m(\cdot;\hat{x}_{k-1})$ in (5), it is easy to see that $\tilde{\Gamma}_{j+1} \le \phi_m(\cdot;\hat{x}_k)$, and hence that the second choice for Γ_{j+1} also satisfies (36).

We now discuss the role played by the parameter τ of PBF. First, the scalar τ satisfying the first inequality in (28) is only used in step 2a as input to GBUS to obtain Γ_{j+1}. Second, even though the analysis of PBF requires a parameter τ satisfying the first inequality in (28) (whose right-hand side depends on M and L), the implementations of some specific instances of PBF do not require knowledge of such τ. For instance, since the updates (E2) and (E3) in Subsection 3.3 do not depend on τ, the two PBF instances based on these updates do not depend on τ either. In summary, even though the analysis of PBF requires a τ satisfying (28), it contains important variants which do not require knowledge of such τ, and hence of parameters M and L satisfying (A2).

It is worth noting however that 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).

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

$$\hat{M}^{\lambda}(\hat{x}_{k-1}) = \min\left\{\phi(x) + \frac{1}{2}\left(\frac{1}{\lambda}+m\right)\|x-\hat{x}_{k-1}\|^2 : x \in \mathbb{R}^n\right\}. \tag{39}$$

Second, the pair (y_j, Γ_j) found in the serious (i.e., the last) iteration j of the k-th cycle approximately solves (39) according to (33), in which case the point x_j as in (30) becomes the center $\hat{x}_k$ for the next cycle.

We now discuss how the inexact criterion (33) can be interpreted in terms of the prox subproblem (39). Since $\Gamma_j \le \phi_m(\cdot;\hat{x}_{k-1})$, it follows from the definitions of θ_j, $\hat{M}^{\lambda}(\hat{x}_{k-1})$, and $\phi_m(\cdot;\hat{x}_{k-1})$ in (31), (39), and (5), respectively, that $\theta_j \le \hat{M}^{\lambda}(\hat{x}_{k-1})$, and hence that

$$\begin{aligned} 0 &\le \phi(y_j) + \frac{1}{2}\left(\frac{1}{\lambda}+m\right)\|y_j-\hat{x}_{k-1}\|^2 - \hat{M}^{\lambda}(\hat{x}_{k-1}) \\ &\le \phi(y_j) + \frac{1}{2}\left(\frac{1}{\lambda}+m\right)\|y_j-\hat{x}_{k-1}\|^2 - \theta_j = t_j + \frac{1-\chi}{2\lambda}\|y_j-\hat{x}_{k-1}\|^2, \end{aligned} \tag{40}$$

where the last identity follows from the definitions of t_j in (33). Hence, if j is an iteration for which (33) does not hold (i.e., a serious one), then it follows from (40) that y_j is a δ_j-solution of (39) where

$$\delta_j := \delta + \frac{1-\chi}{2\lambda}\|y_j - \hat{x}_{k-1}\|^2.$$

Clearly, the smaller $\chi \in (0,1]$ is, the looser (40) becomes. However, our analysis only holds under the condition that the second (strict) inequality in (28) holds. On the other hand, our analysis also requires the left hand side of this strict inequality to be bounded away from zero. Hence a convenient way of choosing χ is as

$$\chi = \frac{1+\gamma}{2+m\lambda}$$

for some $\gamma \in (0, 1+m\lambda]$ such that $\gamma^{-1} = \mathcal{O}(1)$.

We now state the main complexity result for PBF.

Theorem 3.3 (Main Theorem) *Given $(\bar{\eta}, \bar{\varepsilon}) \in \mathbb{R}^2_{++}$, define*

$$\delta = \delta(\bar{\eta}, \bar{\varepsilon}) := \min\left\{\frac{\bar{\varepsilon}\alpha}{2\left(\alpha + (1-\chi)(3+m\lambda)\right)}, \frac{\lambda\bar{\eta}^2}{8 + N(3+m\lambda)}\right\}, \tag{41}$$

$$N := \frac{8\left[1-\chi+(m\lambda+1)^2\right]}{\alpha}, \quad \alpha := \chi(2+m\lambda) - 1 > 0, \tag{42}$$

and let $\tau > 0$ be such that

$$\frac{\tau}{1-\tau} = \lambda\left(\frac{4M^2}{\delta} + L + m\right). \tag{43}$$

Then, PBF with the above choice of δ and τ generates in at most

$$\left[\left\{1 + \lambda\left(\frac{4M^2}{\delta} + L + m\right)\right\}\log^+\left(\frac{2\bar{t}}{\delta}\right) + 2\right]\left[(\hat{M}^\lambda(\hat{x}_0) - \phi^*)\max\left\{\frac{2(1-\chi)}{\alpha\bar{\varepsilon}}, \frac{N}{\lambda\bar{\eta}^2}\right\} + 1\right] \tag{44}$$

total iterations an iterate within the sequence $\{\hat{y}_k\}$ (as in step 2b of PBF) which is a $(\bar{\eta}, \bar{\varepsilon}; m)$-regularized stationary point of ϕ.

We now make two remarks about Theorem 3.3. First, the denominator of (42) is positive in view of the inequality in (28). Second, in terms of the tolerances $\bar{\eta}$ and $\bar{\varepsilon}$ only, it follows from Theorem 3.3 that the iteration complexity of PBF to find a $(\bar{\eta}, \bar{\varepsilon}; m)$-regularized stationary point of ϕ is $\tilde{\mathcal{O}}_1(\max\{\bar{\eta}^{-4}, \bar{\varepsilon}^{-2}\})$. Third, if $\chi = 1$ then (44) is $\tilde{\mathcal{O}}_1(\max\{\bar{\eta}^{-4}, \bar{\eta}^{-2}\bar{\varepsilon}^{-1}\})$, which is better than the previous estimate when $\bar{\varepsilon} \ll \bar{\eta}^2$.

For the sake of stating an iteration-complexity for PBF to find a $(\rho; 1/m)$-Moreau stationary point, we now consider the specific case of PBF and Theorem 3.3 with parameters χ and λ chosen as $\chi = 1$ and $\lambda = \Theta(m^{-1})$. In this case, it follows that

$$\alpha = 1 + m\lambda = \Theta(1), \quad N = 8(1+m\lambda) = \Theta(1),$$

and hence that

$$\delta = \mathcal{O}\left(\min\left\{\frac{\bar{\eta}^2}{m}, \bar{\varepsilon}\right\}\right). \tag{45}$$

Thus, the complexity bound (44) becomes

$$\mathcal{O}\left(\left[\left(\frac{M^2}{m\delta} + \frac{L}{m} + 1\right)\log_1^+\left(\frac{\bar{t}}{\delta}\right) + 1\right]\left[\frac{\hat{M}^\lambda(\hat{x}_0) - \phi^*}{\lambda\bar{\eta}^2} + 1\right]\right) \tag{46}$$

which, upon disregarding all the $\mathcal{O}(1)$ and $\mathcal{O}(1/\delta)$ terms, becomes

$$\tilde{\mathcal{O}}_1\left(\frac{\hat{M}^\lambda(\hat{x}_0) - \phi^*}{\bar{\eta}^2}\left(M^2\max\left\{\frac{m}{\bar{\eta}^2}, \frac{1}{\bar{\varepsilon}}\right\} + L\right)\right). \tag{47}$$

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

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

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

finds in

$$\tilde{\mathcal{O}}_1\left(\frac{(\hat M^{1/m}(\hat x_0) - \phi^*)}{\rho^2}\left(\frac{M^2 m}{\rho^2} + L\right)\right) \tag{48}$$

total iterations an iterate within the sequence $\{\hat y_k\}$ which is a $(\rho; 1/m)$-Moreau stationary point of ϕ.

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 2.10(b). The choice of $(\bar\eta, \bar\varepsilon, \chi)$ and relation (45) imply that

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

The conclusion of the corollary now follows from the above relation and (47). ∎

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 $\chi = 1$, $\lambda = 1/m$ and $L = 0$, namely, bound (48) with $L = 0$, and the first bound in the right hand side of (21) equals the latter one only when

$$\gamma = \Theta\left(\frac{1}{M}\sqrt{\frac{\hat M^{1/m}(\hat x_0) - \phi^*}{m}}\right).$$

However, this choice of γ is generally not computable due to the fact that $\hat M^{1/m}(\hat x_0) - \phi^*$ and, most likely, the parameter M is not known.

4 Proof of Theorem 3.3

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

Throughout this section and the next one, we let

$$\hat f_k(\cdot) = f_m(\cdot; \hat x_{k-1}). \tag{49}$$

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

Lemma 4.1 *For every $j \in \mathcal{C}_k \setminus \{j_k\}$ and $u \in \operatorname{dom} h$ the following statements hold:*

a) there exists $\bar\Gamma_j \in \overline{\operatorname{Conv}}(\mathbb{R}^n)$ such that $\bar\Gamma_j \le \hat f_k + h$ and

$$\tau\bar\Gamma_j + (1-\tau)[\ell_{\hat f_k}(\cdot; x_j) + h] \le \Gamma_{j+1}, \tag{50}$$

$$\bar\Gamma_j(x_j) = \Gamma_j(x_j), \quad x_j = \underset{u\in\mathbb{R}^n}{\operatorname{argmin}}\left\{\bar\Gamma_j(u) + \frac{1}{2\lambda}\|u - \hat x_{k-1}\|^2\right\}; \tag{51}$$

b) there holds

$$\bar{\Gamma}_j(u) + \frac{1}{2\lambda}\|u - \hat{x}_{k-1}\|^2 \geq \theta_j + \frac{1}{2\lambda}\|u - x_j\|^2.$$

Proof: a) This statement immediately follows from (22), (23), and the facts that iteration j is a null one and Γ_{j+1} is the output of the GBUS with input (λ, τ) and $(x^c, x, \Gamma) = (\hat{x}_{k-1}, x_j, \Gamma_j)$ (see the null update in step 2 of GPB) .

b) Using (51) and the fact that $\bar{\Gamma}_j \in \overline{\text{Conv}}\,(\mathbb{R}^n)$, we have for every $u \in \text{dom}\, h$,

$$\bar{\Gamma}_j(u) + \frac{1}{2\lambda}\|u - \hat{x}_{k-1}\|^2 \geq \bar{\Gamma}_j(x_j) + \frac{1}{2\lambda}\|x_j - \hat{x}_{k-1}\|^2 + \frac{1}{2\lambda}\|u - x_j\|^2.$$

The statement now follows from the above inequality, (51) and the definition of θ_j in (33). ■

The following technical result provides an important recursive formula for $\{\theta_j\}$ which is used in Lemma 4.3 to give a recursive formula for $\{t_j\}$. It is worth observing that its proof uses for the first time the condition that τ is chosen to satisfy (28).

Lemma 4.2 *For every $j \in \mathcal{C}_k \setminus \{j_k\}$, we have*

$$\theta_{j+1} \geq \tau\theta_j + (1-\tau)\left[\ell_{\hat{f}_k}(x_{j+1}; x_j) + h(x_{j+1}) + \frac{1}{2\lambda}\|x_{j+1} - \hat{x}_{k-1}\|^2 + \left(\frac{2M^2}{\delta} + \frac{L+m}{2}\right)\|x_{j+1} - x_j\|^2\right] \quad (52)$$

where $\hat{f}_k$ is as in (49).

Proof: Using (50) and Lemma 4.1(b) with $u = x_{j+1}$, and the definitions of θ_{j+1} and Γ_{j+1} in (33) and (30), respectively, we have

$$\begin{aligned}
\theta_{j+1} &= \Gamma_{j+1}(x_{j+1}) + \frac{1}{2\lambda}\|x_{j+1} - \hat{x}_{k-1}\|^2 \\
&\geq (1-\tau)\left[\ell_{\hat{f}_k}(x_{j+1}; x_j) + h(x_{j+1}) + \frac{1}{2\lambda}\|x_{j+1} - \hat{x}_{k-1}\|^2\right] + \tau\left(\bar{\Gamma}_j(x_{j+1}) + \frac{1}{2\lambda}\|x_{j+1} - \hat{x}_{k-1}\|^2\right) \\
&\geq (1-\tau)\left[\ell_{\hat{f}_k}(x_{j+1}; x_j) + h(x_{j+1}) + \frac{1}{2\lambda}\|x_{j+1} - \hat{x}_{k-1}\|^2\right] + \tau\left(\theta_j + \frac{1}{2\lambda}\|x_{j+1} - x_j\|^2\right)
\end{aligned}$$

which, together with the definition of τ in (28), implies the conclusion of the lemma. ■

The next result establishes a key recursive formula for $\{t_j\}$ which plays an important role in the analysis of the null iterates.

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

$$t_{j+1} - \frac{\delta}{2} \leq \tau\left(t_j - \frac{\delta}{2}\right)$$

where t_j is as in (33).

Proof: Applying (16) with $z = \hat{x}_{k-1}$ and $\tilde{z} = x_j$, together with the definition of $\hat{f}_k$ in (49), we have

$$\hat{f}_k(x_{j+1}) - 2M\|x_{j+1} - x_j\| \leq \ell_{\hat{f}_k}(x_{j+1}; x_j) + \frac{L+m}{2}\|x_{j+1} - x_j\|^2 \quad (53)$$

This inequality, relations (5), (29) and (49), Lemma 4.2, and the fact that $\chi \leq 1$, then imply

$$\begin{aligned}
\theta_{j+1} - \tau\theta_j &\overset{(52)}{\geq} (1-\tau)\left[\ell_{\hat{f}_k}(x_{j+1}; x_j) + h(x_{j+1}) + \frac{1}{2\lambda}\|x_{j+1} - \hat{x}_{k-1}\|^2 + \left(\frac{2M^2}{\delta} + \frac{L+m}{2}\right)\|x_{j+1} - x_j\|^2\right] \\
&\overset{(53),1\geq\chi}{\geq} (1-\tau)\left[(\hat{f}_k + h)(x_{j+1}) - 2M\|x_{j+1} - x_j\| + \frac{2M^2}{\delta}\|x_{j+1} - x_j\|^2 + \frac{\chi}{2\lambda}\|x_{j+1} - \hat{x}_{k-1}\|^2\right] \\
&\geq (1-\tau)\left[(\hat{f}_k + h)(x_{j+1}) + \frac{\chi}{2\lambda}\|x_{j+1} - \hat{x}_{k-1}\|^2 - \frac{\delta}{2}\right] \\
&\overset{(5),(29),(49)}{=} (1-\tau)\left[\phi_{\tilde{m}}(x_{j+1}; \hat{x}_{k-1}) - \frac{\delta}{2}\right]
\end{aligned} \quad (54)$$

where the third inequality is due to the inequality $a^2 - 2ab \geq -b^2$ with $a = \sqrt{2}M\|x_{j+1} - x_j\|$ and $b = \delta/\sqrt{2}$. Hence, it follows from (33), (54) and (32) that

$$\begin{aligned} t_{j+1} - \tau t_j &\overset{(33)}{=} [\phi_{\tilde{m}}(y_{j+1}; \hat{x}_{k-1}) - \theta_{j+1}] - \tau[\phi_{\tilde{m}}(y_j; \hat{x}_{k-1}) - \theta_j] \\ &= [\phi_{\tilde{m}}(y_{j+1}; \hat{x}_{k-1}) - \tau\phi_{\tilde{m}}(y_j; \hat{x}_{k-1})] - [\theta_{j+1} - \tau\theta_j] \\ &\overset{(54)}{\leq} [\phi_{\tilde{m}}(y_{j+1}; \hat{x}_{k-1}) - \tau\phi_{\tilde{m}}(y_j; \hat{x}_{k-1})] - (1-\tau)\left[\phi_{\tilde{m}}(x_{j+1}; \hat{x}_{k-1}) - \frac{\delta}{2}\right] \overset{(32)}{\leq} \frac{(1-\tau)\delta}{2}, \end{aligned}$$

and hence that the conclusion of the lemma holds. ■

We are now ready to present the main result of this subsection where a bound on $|\mathcal{C}_k|$ is obtained in terms of τ, t_{i_k} and δ.

Proposition 4.4 *The set $\mathcal{C}_k$ is finite and*

$$|\mathcal{C}_k| \leq \frac{1}{1-\tau}\log^+\left(\frac{2t_{i_k}}{\delta}\right) + 2$$

where τ is as in step 0 of PBF.

Proof: If $t_{i_k} \leq \delta$, then $|\mathcal{C}_k| = 1$ and the conclusion of the lemma trivially holds. Assume now that $t_{i_k} > \delta$ and let $j \in \mathcal{C}_k$ be such that $t_j > \delta$. Then, using the inequality $\tau \leq e^{\tau-1}$ and Lemma 4.3, we conclude that

$$\frac{\delta}{2} < t_j - \frac{\delta}{2} \leq \tau^{j-i_k}\left(t_{i_k} - \frac{\delta}{2}\right) \leq \tau^{j-i_k} t_{i_k} \leq e^{(\tau-1)(j-i_k)} t_{i_k}.$$

and hence that

$$j - i_k \leq \frac{1}{1-\tau}\log^+\left(\frac{2t_{i_k}}{\delta}\right).$$

This inequality together with the definition of $\mathcal{C}_k$ immediately implies the conclusion of the proposition. ■

4.2 Bounding total number of cycles

Recall that $\hat{x}_k$, $\hat{y}_k$, $\hat{\Gamma}_k$ denote the last x_j, y_j and Γ_j generated within a cycle, i.e., the ones with $j = j_k$.

The following technical result provides the main properties of the sequences $\{\hat{x}_k\}$, $\{\hat{y}_k\}$ and $\{\hat{\Gamma}_k\}$, which are sufficent to establish a bound on the total number of cycles.

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

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

$$\min_{u \in \mathbb{R}^n} \hat{\Gamma}_k(u) + \frac{1}{2\lambda}\|u - \hat{x}_{k-1}\|^2; \tag{55}$$

hence, if $\hat{\theta}_k$ denotes the optimal value of (55), then

$$\hat{\theta}_k = \hat{\Gamma}_k(\hat{x}_k) + \frac{1}{2\lambda}\|\hat{x}_k - \hat{x}_{k-1}\|^2; \tag{56}$$

b) there hold

$$\hat{\Gamma}_k(\cdot) \in \overline{\text{Conv}}\,(\mathbb{R}^n), \quad \hat{\Gamma}_k(\cdot) \leq \phi_m(\cdot; \hat{x}_{k-1}) \tag{57}$$

and

$$\phi_{\tilde{m}}(\hat{y}_k; \hat{x}_{k-1}) - \hat{\theta}_k \leq \delta \tag{58}$$

where $\tilde{m}$ is as in (29).

Proof: a) The definition of $\hat{x}_k$ in step 2b of PBF and (30) imply that $\hat{x}_k$ is an optimal solution of (55).

b) If $|C_k| = 1$, i.e., j_k is also the first iteration of cycle C_k, then (57) follows from (36). If $|C_k| > 1$, then $\hat{\Gamma}_k = \Gamma_{j_k}$ is the output of GBUS with $(g, h) = (f_m(\cdot; \hat{x}_{k-1}), h)$, and hence satisfies (57) in view of (22). Moreover, (58) follows from the logic of the prox-center update rule in step 2 of PBF (see (33)). ∎

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 2b of PBF satisfy*

$$\hat{\varepsilon}_k \geq 0, \quad \hat{v}_k \in \partial_{\hat{\varepsilon}_k}[\phi_m(\cdot; \hat{x}_{k-1})](\hat{y}_k), \tag{59}$$

$$\hat{w}_k \in \partial_{\hat{\varepsilon}_k}[\phi_m(\cdot; \hat{y}_k)](\hat{y}_k). \tag{60}$$

Proof: Since $\hat{x}_k$ is an optimal solution of (55) in view of Lemma 4.5(a), using the optimality condition for (55), the fact that $\hat{\Gamma}_k \in \overline{\text{Conv}}\,(\mathbb{R}^n)$, and the definition of $\hat{v}_k$ in (34), 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 (35), then imply that for every $u \in \operatorname{dom} h$,

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

and hence that the inclusion in (59) holds. The inequality in (59) follows from the above inequality with $u = \hat{y}_k$. Moreover, the definition of $\hat{w}_k$ in (35), the inclusion in (59), and Lemma A.1 imply that

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

and hence that (60) holds. ∎

It follows from (60) 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 2b 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) involving $\phi_{\tilde{m}}(\hat{y}_k; \hat{x}_{k-1})$, a quantity which will be used as a potential in the analysis of this subsection. The two results preceding Lemma 4.9 are technical ones that are needed to prove the aforementioned lemma.

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 2b of PBF, satisfy*

$$\hat{\varepsilon}_k + \frac{1}{2\lambda}\|\hat{y}_k - \hat{x}_k\|^2 \leq \delta + \frac{1-\chi}{2\lambda}\|\hat{y}_k - \hat{x}_{k-1}\|^2 \tag{61}$$

and

$$\|\hat{w}_k\|^2 \leq \frac{4\delta}{\lambda} + \frac{\alpha N}{4\lambda^2}\|\hat{y}_k - \hat{x}_{k-1}\|^2 \tag{62}$$

where α and N are as in (42)*, and λ, χ and δ are as in step 0 of PBF.*

Proof: Using both statements (a) and (b) of Lemma 4.5, the definitions of $\hat{\varepsilon}_k$, $\tilde{m}$ and $\hat{v}_k$ in (35), (29) and (34), respectively, we conclude that

$$\begin{aligned}
\hat{\varepsilon}_k &= \phi_m(\hat{y}_k; \hat{x}_{k-1}) - \hat{\Gamma}_k(\hat{x}_k) - \langle \hat{v}_k, \hat{y}_k - \hat{x}_k \rangle \\
&\overset{(58),(29)}{\leq} \delta - \frac{\chi}{2\lambda}\|\hat{y}_k - \hat{x}_{k-1}\|^2 + \hat{\theta}_k - \hat{\Gamma}_k(\hat{x}_k) - \langle \hat{v}_k, \hat{y}_k - \hat{x}_k \rangle \\
&\overset{(56)}{=} \delta - \frac{\chi}{2\lambda}\|\hat{y}_k - \hat{x}_{k-1}\|^2 + \frac{1}{2\lambda}\|\hat{x}_k - \hat{x}_{k-1}\|^2 - \langle \hat{v}_k, \hat{y}_k - \hat{x}_k \rangle \\
&= \delta - \frac{\chi}{2\lambda}\|\hat{y}_k - \hat{x}_{k-1}\|^2 + \frac{1}{2\lambda}\left(\|\hat{x}_k - \hat{x}_{k-1}\|^2 + 2\langle \hat{x}_k - \hat{x}_{k-1}, \hat{y}_k - \hat{x}_k \rangle\right) \\
&= \delta - \frac{\chi}{2\lambda}\|\hat{y}_k - \hat{x}_{k-1}\|^2 + \frac{1}{2\lambda}\left(\|\hat{y}_k - \hat{x}_{k-1}\|^2 - \|\hat{y}_k - \hat{x}_k\|^2\right),
\end{aligned}$$

and hence that (61) holds. It follows from the definition of $\hat{w}_k$ in (35) that

$$\hat{w}_k = \hat{v}_k - m(\hat{y}_k - \hat{x}_{k-1}) = \frac{1}{\lambda}(\hat{y}_k - \hat{x}_k) - \left(m + \frac{1}{\lambda}\right)(\hat{y}_k - \hat{x}_{k-1})$$

and hence that

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

where the first inequality is due to the relation $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$ and the last inequality is due to (61). The above inequality and the definitions of α and N in (42) imply (62). ∎

Lemma 4.8 *Define $\hat{y}_0 = \hat{x}_0$. Then, for every $k \ge 1$, we have*

$$\phi_{\tilde{m}}(\hat{y}_k; \hat{x}_{k-1}) \le \delta + \phi_m(\hat{y}_{k-1}; \hat{x}_{k-1}) + \frac{1}{2\lambda}\|\hat{y}_{k-1} - \hat{x}_{k-1}\|^2 \tag{63}$$

where $\tilde{m}$ is as in (29)*, and $\{\hat{y}_k\}$ and $\{\hat{x}_k\}$ are as in step 2b of PBF.*

Proof: Using the definitions of $\phi_{\tilde{m}}(\cdot;\cdot)$ and $\tilde{m}$ in (5) and (29), respectively, and statements (a) and (b) of Lemma 4.5, we conclude that

$$\begin{aligned} \phi_{\tilde{m}}(\hat{y}_k; \hat{x}_{k-1}) - \delta \le \hat{\theta}_k &= \hat{\Gamma}_k(\hat{x}_k) + \frac{1}{2\lambda}\|\hat{x}_k - \hat{x}_{k-1}\|^2 \\ &\le \hat{\Gamma}_k(\hat{y}_{k-1}) + \frac{1}{2\lambda}\|\hat{y}_{k-1} - \hat{x}_{k-1}\|^2 \\ &\le \phi_m(\hat{y}_{k-1}; \hat{x}_{k-1}) + \frac{1}{2\lambda}\|\hat{y}_{k-1} - \hat{x}_{k-1}\|^2, \end{aligned}$$

which shows (63). ∎

Lemma 4.9 *For every $k \ge 1$, define*

$$\Delta_k := \phi_{\tilde{m}}(\hat{y}_k; \hat{x}_{k-1}). \tag{64}$$

Then, for every $k \ge 1$, we have

$$\Delta_{k+1} + \frac{\alpha}{2\lambda}\|\hat{y}_k - \hat{x}_{k-1}\|^2 \le \Delta_k + (2 + m\lambda)\delta. \tag{65}$$

and

$$\frac{\alpha}{2\lambda}\sum_{l=1}^{k}\|\hat{y}_l - \hat{x}_{l-1}\|^2 \le \hat{M}^\lambda(\hat{x}_0) - \phi(\hat{y}_{k+1}) - \frac{1}{2}\left(m + \frac{\chi}{\lambda}\right)\|\hat{y}_{k+1} - \hat{x}_k\|^2 + (3 + m\lambda)k\delta. \tag{66}$$

Proof: Now, using (42), (61), (63), and (64), and the definition of $\phi_m(\cdot;\cdot)$ in (5), we conclude that

$$\begin{aligned} \Delta_{k+1} &\overset{(63)}{\le} \phi(\hat{y}_k) + \frac{1+\lambda m}{2\lambda}\|\hat{y}_k - \hat{x}_k\|^2 + \delta \\ &\overset{(61)}{\le} \phi(\hat{y}_k) + (1 + m\lambda)\left[\delta + \frac{1-\chi}{2\lambda}\|\hat{y}_k - \hat{x}_{k-1}\|^2\right] + \delta \\ &\overset{(64)}{=} \left[\Delta_k - \frac{\lambda m + \chi}{2\lambda}\|\hat{y}_k - \hat{x}_{k-1}\|^2\right] + \frac{(1+m\lambda)(1-\chi)}{2\lambda}\|\hat{y}_k - \hat{x}_{k-1}\|^2 + (2 + m\lambda)\delta \\ &\overset{(42)}{=} \Delta_k - \frac{\alpha}{2\lambda}\|\hat{y}_k - \hat{x}_{k-1}\|^2 + (2 + m\lambda)\delta, \end{aligned}$$

and hence that (65) holds. Now, adding (65) from $k = 1$ to $k = k$ we conclude that

$$\frac{\alpha}{2\lambda}\sum_{l=1}^{k}\|\hat{y}_l - \hat{x}_{l-1}\|^2 \le \Delta_1 - \Delta_{k+1} + (2 + m\lambda)\delta k \tag{67}$$

Using (58) with $k = 1$, the definition of Δ_1 in (64), the definition of $\hat{M}^\lambda(\hat{x}_0)$ in (10), and the fact that $\hat{\Gamma}_1(\cdot) \le \phi_m(\cdot; \hat{x}_0)$, we can conclude that

$$\Delta_1 \le \hat{\theta}_1 + \delta \le \hat{M}^\lambda(\hat{x}_0) + \delta.$$

Inequality (66) now follows from the above relation, (67) and the definition of Δ_k in (64). ■

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

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

$$K = K(\bar{\eta}, \bar{\varepsilon}) := \left\lceil (\hat{M}^\lambda(\hat{x}_0) - \phi^*) \max\left\{ \frac{2(1-\chi)}{\alpha\bar{\varepsilon}}, \frac{N}{\lambda\bar{\eta}^2} \right\} \right\rceil \tag{68}$$

where $\hat{M}^\lambda(\cdot)$ is as in (10), and α and N are as in (42). Then, PBF with δ as in (41) generates an iteration index $k \le K(\bar{\eta}, \bar{\varepsilon})$ such that

$$\|\hat{w}_k\| \le \bar{\eta}, \quad \hat{\varepsilon}_k \le \bar{\varepsilon}. \tag{69}$$

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

Proof: If PBF stops for some $k' \le K$ iterations, then the stopping criterion in step 2b of PBF ensures that (69) holds. Then, (60), (69), and Definition 2.5 imply that $\hat{y}_k$ is a $(\bar{\eta}, \bar{\varepsilon}; m)$-regularized stationary point of problem (1). Now assume PBF runs for $k' \ge K$ iterations. Using relation (66) with $k = K$ and the fact that (A3) implies that $\phi(\hat{y}_{K+1}) \ge \phi^*$, we have

$$\frac{\alpha}{2\lambda}\sum_{l=1}^{K}\|\hat{y}_l - \hat{x}_{l-1}\|^2 \le \hat{M}^\lambda(\hat{x}_0) - \phi^* + (3 + m\lambda)K\delta. \tag{70}$$

Let $k \in \{1, \ldots, K\}$ be such that

$$\|\hat{y}_k - \hat{x}_{k-1}\| = \min_{1\le l\le K}\|\hat{y}_l - \hat{x}_{l-1}\|. \tag{71}$$

The above relation and (70) thus imply that

$$\frac{\alpha}{2\lambda}\|\hat{y}_k - \hat{x}_{k-1}\|^2 \le \frac{\hat{M}^\lambda(\hat{x}_0) - \phi^*}{K} + (3 + m\lambda)\delta. \tag{72}$$

This inequality together with (61) imply that

$$\begin{aligned}\hat{\varepsilon}_k &\le \delta + \frac{1-\chi}{\alpha}\left[\frac{\hat{M}^\lambda(\hat{x}_0)) - \phi^*}{K} + (3 + m\lambda)\delta\right] \\ &= \left(1 + \frac{(1-\chi)(3+m\lambda)}{\alpha}\right)\delta + \frac{(1-\chi)(\phi(\hat{x}_0) - \phi^*)}{\alpha K} \le \bar{\varepsilon}\end{aligned} \tag{73}$$

where the last inequality is due to the definition of δ and K in (41) and (68), respectively. Following the similar reason, inequalities (72) and (62) imply that

$$\begin{aligned}\|\hat{w}_k\|^2 &\le \frac{4\delta}{\lambda} + \frac{\alpha N}{4\lambda^2}\|\hat{y}_k - \hat{x}_{k-1}\|^2 \\ &\le \frac{4\delta}{\lambda} + \frac{N}{2\lambda}\left[\frac{\hat{M}^\lambda(\hat{x}_0) - \phi^*}{K} + (3 + m\lambda)\delta\right] \\ &= \frac{8 + N(3 + m\lambda)}{2\lambda}\delta + \frac{N(\hat{M}^\lambda(\hat{x}_0) - \phi^*)}{2\lambda K} \le \bar{\eta}^2\end{aligned} \tag{74}$$

where the last inequality is due to the definition of δ and K in (41) and (68), respectively. Finally, (60), (73), (74), and Definition 2.5 then imply that $\hat{y}_k$ is a $(\bar{\eta}, \bar{\varepsilon}; m)$-regularized stationary point of problem (1). ∎

Before ending this subsection, we observe that the quantity $\hat{M}^\lambda(\hat{x}_0) - \phi^*$ in (68) 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 (10).

4.3 Proof of Theorem 3.3

Recall that Theorem 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 t_{i_k}. The following result refines the latter result by providing a uniform bound on t_{i_k}.

To simplify the statement of the next result, we introduce the following constants:

$$\zeta := \begin{cases} \frac{1}{2(L+m)\lambda} & \text{if } \lambda > \frac{1}{2(L+m)}; \\ 1 & \text{if } \lambda \le \frac{1}{2(L+m)}, \end{cases} \tag{75}$$

and

$$\beta_1 := \left(m + \frac{2}{\zeta\lambda}\right)\left(m + \frac{\chi}{\lambda}\right)^{-1} > 1, \quad \beta_2 := \left(\frac{L+m}{2} + 1\right)\zeta^{-2}\left(\frac{1}{4\zeta\lambda} + \frac{m}{2}\right)^{-1}. \tag{76}$$

Lemma 4.11 *Let δ and K be as in* (41) *and* (68)*, respectively, and define*

$$\bar{t} := M^2 + \beta_2\left\{\beta_1[\hat{M}^\lambda(\hat{x}_0) - \phi^*] + \beta_1(3 + m\lambda)K\delta + 4\zeta\lambda M^2\right\} \tag{77}$$

where ζ and N are as in (75) *and* (42)*, respectively, and β_1 and γ are as in* (76)*. Then, $t_{i_k} \le \bar{t}$ for every $1 \le k \le K$.*

Proof: Applying (18) with $z = \hat{x}_{k-1}$, we have

$$f_m(x_{i_k}; \hat{x}_{k-1}) - \ell_f(x_{i_k}; \hat{x}_{k-1}) \le 2M\|x_{i_k} - \hat{x}_{k-1}\| + \frac{L+m}{2}\|x_{i_k} - \hat{x}_{k-1}\|^2. \tag{78}$$

Using this inequality, (36), the definition of ϕ_m and $\tilde{m}$ in (5) and (29) respectively, (32) and (33) with $j = i_k$, and the fact that $1 \ge \chi$, we have

$$\begin{aligned} t_{i_k} &\overset{(33)}{=} \phi_{\tilde{m}}(y_{i_k}; \hat{x}_{k-1}) - \theta_{i_k} \\ &\overset{(31)}{=} \phi_{\tilde{m}}(y_{i_k}; \hat{x}_{k-1}) - \Gamma_{i_k}(x_{i_k}) - \frac{1}{2\lambda}\|x_{i_k} - \hat{x}_{k-1}\|^2 \\ &\overset{(32),(29)}{\le} \phi_m(x_{i_k}; \hat{x}_{k-1}) - \Gamma_{i_k}(x_{i_k}) - \frac{1-\chi}{2\lambda}\|x_{i_k} - \hat{x}_{k-1}\|^2 \\ &\overset{(36)}{\le} f_m(x_{i_k}; \hat{x}_{k-1}) - \ell_f(x_{i_k}; \hat{x}_{k-1}) \\ &\overset{(78)}{\le} 2M\|x_{i_k} - \hat{x}_{k-1}\| + \frac{L+m}{2}\|x_{i_k} - \hat{x}_{k-1}\|^2 \\ &\le M^2 + \left(\frac{L+m}{2} + 1\right)\|x_{i_k} - \hat{x}_{k-1}\|^2, \end{aligned} \tag{79}$$

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

We will now bound $\|x_{i_k} - \hat{x}_{k-1}\|^2$. It follows from the inequality in (59) and (66) that

$$0 \le \hat{M}^\lambda(\hat{x}_0) - \phi(\hat{y}_k) - \frac{1}{2}\left(m + \frac{\chi}{\lambda}\right)\|\hat{y}_k - \hat{x}_{k-1}\|^2 + (3 + m\lambda)(k-1)\delta.$$

Thus, using the definition of β_1 in (76) and the fact that $k \le K$, we have

$$\phi(\hat{y}_k) - \phi^* + \frac{1}{2\beta_1}\left(m + \frac{2}{\zeta\lambda}\right)\|\hat{y}_k - \hat{x}_{k-1}\|^2 \le \hat{M}^\lambda(\hat{x}_0) - \phi^* + (3 + m\lambda)K\delta. \tag{80}$$

This inequality, Lemma C.3 with $(\Gamma, z_0, u) = (\Gamma_{i_k}, \hat{x}_{k-1}, \hat{y}_k)$, and the fact that $\phi \geq \phi^*$, imply that

$$\zeta^2 \left(\frac{1}{4\zeta\lambda} + \frac{m}{2} \right) \|x_{i_k} - \hat{x}_{k-1}\|^2 - 4\zeta\lambda M^2 \overset{(92)}{\leq} \phi(\hat{y}_k) - \phi^* + \frac{1}{2}\left(m + \frac{2}{\zeta\lambda} \right) \|\hat{y}_k - \hat{x}_{k-1}\|^2$$
$$\overset{(80)}{\leq} \beta_1 [\hat{M}^\lambda(\hat{x}_0) - \phi^* + (3 + m\lambda) K\delta].$$

The statement now follows by combining (79) and the above inequality and using the definitions of β_2 and $\bar{t}$ in (76) and (77), respectively. ■

Now we are ready to present the proof of Theorem 3.3.

Proof of Theorem 3.3 Using Theorem 4.10, Lemma 4.11, and Proposition 4.4, we know that the total complexity for PBF to obtain a $(\bar{\eta}, \bar{\varepsilon}; m)$-regularized stationary point is

$$\left[\frac{1}{1-\tau} \log^+ \left(\frac{2\bar{t}}{\delta} \right) + 2 \right] \left[(\hat{M}^\lambda(\hat{x}_0) - \phi^*) \max \left\{ \frac{2(1-\chi)}{\alpha\bar{\varepsilon}}, \frac{N}{\lambda\bar{\eta}^2} \right\} + 1 \right].$$

The conclusion of the theorem now follows from the choice of τ in (43). ■

5 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 fourth paragraph 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).

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

[4] M. Díaz and B. Grimmer. Optimal convergence rates for the proximal bundle method. *arXiv preprint arXiv:2105.07874*, 2021.

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

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

[7] 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.

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

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

[10] 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.

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

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

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

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

[15] 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.

[16] J. Liang and R. D. C. Monteiro. A unified analysis of a class of proximal bundle methods for solving hybrid convex composite optimization problems. *arXiv preprint arXiv:2110.01084*, 2021.

[17] T. Lin, C. Jin, and M. I Jordan. Near-optimal algorithms for minimax optimization. In *Conference on Learning Theory*, pages 2738–2779. PMLR, 2020.

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

[19] 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.

[20] 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.

[21] M. Nouiehed, M. Sanjabi, T. Huang, J.D. Lee, and M. Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. In *Advances in Neural Information Processing Systems*, pages 14905–14916, 2019.

[22] 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.

[23] D. M Ostrovskii, A. Lowy, and M. Razaviyayn. Efficient search of first-order nash equilibria in nonconvex-concave smooth min-max problems. *SIAM Journal on Optimization*, 31(4):2508–2538, 2021.

[24] H. Rafique, M. Liu, Q. Lin, and T. Yang. Weakly-convex–concave min–max optimization: provable algorithms and applications in machine learning. *Optimization Methods and Software*, 37(3):1087–1121, 2022.

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

[26] K. K. Thekumparampil, P. Jain, P. Netrapalli, and S. Oh. Efficient algorithms for smooth minimax optimization. In *Advances in Neural Information Processing Systems*, pages 12680–12691, 2019.

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

[28] 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.

[29] 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*

$$0 \in \partial g(\hat{y}), \quad \|y-\hat{y}\| \leq \sqrt{\frac{2\varepsilon}{\mu}}. \tag{81}$$

Proof: The inclusion in (81) 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 (81) follows. ■

B Relationships between notions of stationary points

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

Proof of Proposition 2.8 For this proof only, we let

$$\psi(\cdot) := \phi(\cdot) + \frac{\lambda^{-1}+m}{2}\|\cdot - x\|^2 \tag{82}$$

and denote $\hat{x}^\lambda(x)$ as in (12). 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

$$\psi(u) - \psi(\hat{x}) \geq \frac{1}{2\lambda}\|u-\hat{x}\|^2. \tag{83}$$

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

$$\|x-\tilde{x}\| \leq \delta_D \quad \inf_{\|d\|\leq 1} \phi'(\tilde{x};d) \geq -\varepsilon_D. \tag{84}$$

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

$$\psi(\hat{x}) - \psi(\tilde{x}) \geq \psi'(\tilde{x}; \hat{x} - \tilde{x}) \geq -\left[\varepsilon_D + \left(\frac{1}{\lambda} + m\right)\delta_D\right]\|\hat{x} - \tilde{x}\|. \tag{85}$$

Then using the above inequality and (83) 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 (84) 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 (11) 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 2.7.

b) Since x is a $(\varepsilon_M; \lambda)$-Moreau stationary point, (11) 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

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

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 2.9 Since ψ in (82) is convex, it follows from Theorem 3.26 of [2] that

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

It follows from Proposition 2.4 with $m = 2m$ and the definition of ψ in (82) that

$$\partial\phi(x) = \partial\psi(x).$$

Using the above identity and (86), we have

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

where the last identity is due to the definition of ψ in (82). Hence, the conclusion directly follows from Definition 2.6. ■

Proof of Proposition 2.10 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 $\phi_m(\cdot; x) \leq \phi_{2m}(\cdot; x)$, 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

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

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

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

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 (87). Then the above inequality and (88) imply that

$$\operatorname{dist}(0,\partial\phi(\tilde{x}))\le\bar{\eta}+2\sqrt{2m\bar{\varepsilon}}. \tag{89}$$

Now statement (a) follows from (87), (89), and Proposition 2.9.

b) Statement (b) immediately follows from the first statement and Proposition 2.8(a) with $\lambda=1/m$. ∎

C Technical results for the proof of theorem 3.3

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

Before stating and proving Lemma C.3, we first present two technical results whose proof can be found in Appendix A of [16].

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

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

Then, we have $\|z_\lambda-z_0\|\le(\lambda/\zeta\lambda)\|z_{\tilde{\lambda}}-z_0\|$.

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

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

Moreover, define

$$z_\lambda:=\operatorname*{argmin}_{u\in\mathbb{R}^n}\left\{\Gamma(u)+\frac{1}{2\lambda}\|u-z_0\|^2\right\}. \tag{90}$$

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

$$\frac{1}{2\lambda}\|u-z_\lambda\|^2+(\tilde{f}+h)(z_\lambda)-(\tilde{f}+h)(u)\le\frac{1}{2\lambda}\|u-z_0\|^2+\frac{2\lambda M^2}{1-\lambda L}. \tag{91}$$

Lemma C.3 *For some $(M,L)\in\mathbb{R}^2_+$ and $m\ge0$, assume that $(z_0,\lambda)\in\mathbb{R}^n\times\mathbb{R}_{++}$, function $f\in\mathcal{C}(M,L)$ is m-weakly convex, and $\Gamma,h\in\overline{\operatorname{Conv}}(\mathbb{R}^n)$ are such that*

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

where $f_m(\cdot;z_0)$ is as in (90). Then, for every $u\in\mathbb{R}^n$, we have

$$\zeta^2\left(\frac{1}{4\zeta\lambda}+\frac{m}{2}\right)\|z_\lambda-z_0\|^2\le(f+h)(u)-(f+h)(z_{\zeta\lambda})+\frac{1}{2}\left(m+\frac{2}{\zeta\lambda}\right)\|u-z_0\|^2+4\zeta\lambda M^2 \tag{92}$$

where z_λ is as in (90) and ζ is defined as (75).

Proof: We first prove the conclusion of the lemma under the assumption that λ is such that $\lambda \in (0, \frac{1}{2(L+m)}]$, and hence $\zeta = 1$. Indeed, noting that $f_m(\cdot; z_0) \in \mathrm{Conv}(\mathbb{R}^n) \cap \mathcal{C}(M, L+m)$ due to Lemma 3.1, it follows from Lemma C.2 with $\tilde{f} = f_m(\cdot; z_0)$ and $(\Gamma, z_0, \lambda) = (\Gamma, z_0, \lambda)$ and the definition of λ that for any $u \in \mathrm{dom}\, h$:

$$
\begin{aligned}
\frac{1}{2\lambda}\|u - z_\lambda\|^2 + (f_m(\cdot; z_0) + h)(z_\lambda) - (f_m(\cdot; z_0) + h)(u) &\le \frac{1}{2\lambda}\|u - z_0\|^2 + \frac{2\lambda M^2}{1 - \lambda(L+m)} \\
&\le \frac{1}{2\lambda}\|u - z_0\|^2 + 4\lambda M^2,
\end{aligned}
$$

and hence that

$$
\begin{aligned}
(f+h)(u) - (f+h)(z_\lambda) + \frac{m}{2}\|u - z_0\|^2 + 4\lambda M^2 &\ge \frac{1}{2\lambda}\left(\|u - z_\lambda\|^2 - \|u - z_0\|^2\right) + \frac{m}{2}\|z_\lambda - z_0\|^2 \\
&= \frac{1}{2}(\frac{1}{\lambda} + m)\|z_\lambda - z_0\|^2 - \frac{1}{\lambda}\langle z_\lambda - z_0, u - z_0\rangle \\
&\ge \frac{1}{2}\left(\frac{1}{\lambda} + m\right)\|z_\lambda - z_0\|^2 - \frac{1}{2\lambda}\left(\frac{1}{2}\|z_\lambda - z_0\|^2 + 2\|u - z_0\|^2\right)
\end{aligned}
$$

where the last inequality is due to the fact that $\frac{1}{2}a^2 + 2b^2 \ge 2ab$. Rearranging the above inequality, we then conclude that

$$
(f+h)(u) - (f+h)(z_\lambda) + 4\lambda M^2 \ge \frac{1}{2}\left(\frac{1}{2\lambda} + m\right)\|z_\lambda - z_0\|^2 - \left(\frac{1}{\lambda} + \frac{m}{2}\right)\|u - z_0\|^2 \tag{93}
$$

which, in view of the fact that $\zeta = 1$, immediately implies (92). Next, we show that (92) also holds for $\lambda > 1/[2(L+m)]$. Noting that (92) implies that $\zeta\lambda = 1/[2(L+m)]$, it then follows from (93) with $\lambda = \zeta\lambda$ that

$$
\begin{aligned}
(f+h)(u) - (f+h)(z_{\zeta\lambda}) + \left(\frac{m}{2} + \frac{1}{\zeta\lambda}\right)\|u - z_0\|^2 + 4\zeta\lambda M^2 &\ge \left(\frac{1}{4\zeta\lambda} + \frac{m}{2}\right)\|z_{\zeta\lambda} - z_0\|^2 \\
&\ge \zeta^2\left(\frac{1}{4\zeta\lambda} + \frac{m}{2}\right)\|z_\lambda - z_0\|^2
\end{aligned}
$$

where the last inequality is due to Lemma C.1 with $(\lambda, \zeta\lambda) = (\lambda, \zeta\lambda)$. ■