

# Universal subgradient and proximal bundle methods for convex and strongly convex hybrid composite optimization

Vincent Guigues <sup>\*</sup>      Jiaming Liang <sup>†</sup>      Renato D.C. Monteiro <sup>‡</sup>

July 14, 2024 (revisions: August 2, 2024; June 25, 2025; November 20, 2025)

## Abstract

This paper develops two parameter-free methods for solving convex and strongly convex hybrid composite universal problems, namely, a composite subgradient type method and a proximal bundle type method. Functional complexity bounds for the two methods are established in terms of the unknown strong convexity parameter. The two proposed methods are universal with respect to all problem parameters, including the strong convexity one, and require no knowledge of the optimal value. Moreover, in contrast to previous works, they do not restart nor use multiple threads.

**Key words.** hybrid composite optimization, iteration complexity, universal method, proximal bundle method

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

## 1 Introduction

This paper considers convex hybrid composite optimization (HCO) problem

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

where  $f, h : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  are proper lower semi-continuous convex functions such that  $\text{dom } h \subseteq \text{dom } f$  and the following conditions hold: there exist scalars  $M_f \geq 0$  and  $L_f \geq 0$  and a first-order oracle  $f' : \text{dom } h \rightarrow \mathbb{R}^n$  (i.e.,  $f'(x) \in \partial f(x)$  for every  $x \in \text{dom } h$ ) satisfying the  $(M_f, L_f)$ -hybrid condition that  $\|f'(x) - f'(y)\| \leq 2M_f + L_f\|x - y\|$  for every  $x, y \in \text{dom } h$ .

The intrinsic convexity parameter  $\mu_\psi$  of a convex function  $\psi$  is defined as the largest scalar  $\mu$  such that  $\psi(\cdot) - \mu\|\cdot\|^2/2$  is convex. A method for solving (1) is called parameter-free if it does not require knowledge of any parameter associated with the instance  $(f, h)$ , such as  $(M_f, L_f)$ , or the intrinsic convexity parameters  $\mu_f$ ,  $\mu_h$ , and  $\mu_\phi$ . Parameter-free methods whose complexities are expressed in terms of  $(M_f, L_f)$  are called universal methods. Moreover, parameter-free methods whose complexities are expressed in terms of  $(M_f, L_f, \mu_\phi)$ ,  $(M_f, L_f, \mu_f)$ , and  $(M_f, L_f, \mu_h)$ , are referred to as  $\mu_\phi$ -universal,  $\mu_f$ -universal, and  $\mu_h$ -universal, respectively. It is worth noting that  $\mu_\phi$  can be substantially larger than  $\mu_f + \mu_h$  (e.g., for  $\alpha \gg 0$ ,  $f(x) = \alpha \exp(x)$ , and  $h(x) = \alpha \exp(-x)$ , we have  $\mu_\phi = 2\alpha \gg 0 = \mu_f + \mu_h$ ).

The main goal of this paper is to develop methods for solving (1) that:

- (1) are  $\mu_\phi$ -universal for any instance  $(f, h)$  of (1) with arbitrary parameter pair  $(M_f, L_f)$ ;
- (2) for any given positive scalars  $\bar{M}$  and  $\bar{\mu}$ , are optimal for the class of instances of (1) such that  $L_f = 0$ ,  $M_f \leq \bar{M}$  and  $\mu_\phi \geq \bar{\mu}$ .

---

<sup>\*</sup>School of Applied Mathematics FGV/EMAp, 22250-900 Rio de Janeiro, Brazil. (email: [vincent.guigues@fgv.br](mailto:vincent.guigues@fgv.br)).

<sup>†</sup>Goergen Institute for Data Science and Artificial Intelligence (GIDS-AI) and Department of Computer Science, University of Rochester, Rochester, NY 14620 (email: [jiaming.liang@rochester.edu](mailto:jiaming.liang@rochester.edu)). This work was partially supported by GIDS-AI seed funding and AFOSR grant FA9550-25-1-0182.

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

It is worth noting that a  $\mu_h$ -universal (resp.,  $\mu_f$ -universal) method that is optimal for the class of instances of (1) such that  $L_f = 0$ ,  $M_f \leq \bar{M}$  and  $\mu_h \geq \bar{\mu}$  (resp.,  $\mu_f \geq \bar{\mu}$ ) is not necessarily optimal for the above class, as the first one (i.e., with  $\mu_\phi \geq \bar{\mu}$ ) is larger than the latter one (i.e., with  $\mu_h \geq \bar{\mu}$  or  $\mu_f \geq \bar{\mu}$ ).

**Related literature.** We divide our discussion here into universal and  $\mu$ -universal methods.

*Universal methods:* The first universal methods for solving (1) under the condition that  $\nabla f$  is Hölder continuous have been presented in [26] and [11]. Specifically, the first paper develops universal variants of the primal and dual gradient methods, and an optimal universal accelerated gradient method, while the second one develops accelerated universal variants of the bundle-level and the prox-level methods which achieve optimal complexity. Additional universal methods for solving (1) have been studied in [17, 19, 23, 31] under the condition that  $f$  is smooth, and in [14, 18, 20] for the case where  $f$  is either smooth or nonsmooth. The methods in [18, 20] (resp., [23, 31]) are also shown to be  $\mu_h$ -universal (resp.,  $\mu_f$ -universal) under the condition that  $h = 0$ . The papers [17, 19] present universal accelerated composite gradients methods for solving (1) under the more general condition that  $f$  is a smooth  $m$ -weakly convex function. Since any convex function is  $m$ -weakly convex for any  $m > 0$ , the results of [17, 19] also apply to the convex case and yield complexity bounds similar to the ones of [23, 31].

*$\mu_\phi$ -Universal methods:* Under the assumption that  $f$  is a smooth function (i.e.,  $M_f = 0$ ), various works [1, 2, 3, 4, 5, 6, 8, 13, 25, 29] have developed  $\mu_\phi$ -universal (or  $\mu_f$ -universal) methods for (1) with  $\tilde{\mathcal{O}}(\sqrt{L_f/\mu_\phi})$  (or  $\tilde{\mathcal{O}}(\sqrt{L_f/\mu_f})$ ) iteration complexity bound. For the sake of our discussion, we refer to a convex (resp., strongly convex) version of an accelerated gradient method as ACG (resp., S-ACG). Among papers concerned with finding an  $\varepsilon$ -solution of (1), [25] proposes the first  $\mu_f$ -universal method based on a restart S-ACG scheme where each iteration adaptively updates an estimate of  $\mu_f$  and calls S-FISTA (see also [8] for a  $\mu_\phi$ -universal variant along this venue); moreover, under the assumption that  $\phi_*$  and  $L_f$  are known, [5] develops a  $\mu_\phi$ -universal method that performs only one call to an ACG variant (for convex CO).

Finally, motivated by previous works such as [10, 12, 21, 24, 25, 28, 30],  $\mu_\phi$ -universal restart schemes, including accelerated ones, were developed in [29] (see also related paper [9]) for both when  $\phi$  is smooth or nonsmooth. Specifically, [29] develops a  $\mu_\phi$ -universal method under the assumption that  $\phi_*$  is known; and, also an alternative one for when  $\phi_*$  is unknown<sup>1</sup>, which solves the original problem for multiple initial points and tolerances in parallel.

**Our contribution.** We present two  $\mu_\phi$ -universal methods for solving (1) that fulfill the two requirements (1) and (2) above, namely: a composite subgradient (U-CS) type method and a proximal bundle (U-PB) type method. In contrast to [9, 29], they are non-restart methods that do not require  $\phi_*$  to be known. Moreover, U-CS and U-PB do not need the use of multiple threads as the parallel version of the restart subgradient universal method of [29] does.

U-CS is a variant of the universal primal gradient method of [26] (see also Appendix C.2 of [20]), which is not known to be  $\mu_\phi$ -universal. U-PB is a variant of the generic proximal bundle (GPB) method of [20] that bounds the number of consecutive null iterations and adaptively chooses the prox stepsize under this policy. Both methods are analyzed in a unified manner using a general framework for strongly convex optimization problems (1) (referred to as FSCO) which specifies minimal conditions for its instances to be  $\mu_\phi$ -universal. A functional complexity bound in terms of  $\mu_\phi$  is established for any FSCO instance which is then used to derive complexities for both U-CS and U-PB.

**Organization of the paper.** Subsection 1.1 presents basic definitions and notation used throughout the paper. Section 2 presents two  $\mu_\phi$ -universal methods, namely U-CS and U-PB, for solving problem (1) and establishes their corresponding complexity bounds. Section 3 formally describes FSCO, which is the framework that include U-CS and U-PB as special instances, and provides the oracle complexity of FSCO. Section 4 is devoted to the proofs of the main results for U-CS and U-PB. Section 5 presents some concluding remarks and possible extensions. Finally, Appendix A provides technical results of FSCO and U-PB.

## 1.1 Basic definitions and notation

Let  $\mathbb{R}$  denote the set of real numbers. Let  $\mathbb{R}_+$  (resp.,  $\mathbb{R}_{++}$ ) denote the set of non-negative real numbers (resp., the set of positive real numbers). Let  $\mathbb{R}^n$  denote the standard  $n$ -dimensional Euclidean space equipped with

<sup>1</sup>For this algorithm, it is assumed that the number of threads is  $\mathcal{O}(\log \varepsilon^{-1})$  and, it is shown that its complexity is better than the usual one by a logarithmic factor. However, if a single thread is used, Corollary 5 of [29] implies that the complexity for its nonsmooth approach (without smoothing) applied to strongly convex instances of (1) reduces to the usual convex complexity (i.e., with  $\mu_\phi = 0$ ), and hence is not  $\mu_\phi$ -universal (see the last paragraph of Subsection 2.1 for more details about this method and its complexity).

inner product and norm denoted by  $\langle \cdot, \cdot \rangle$  and  $\| \cdot \|$ , respectively. Let  $\log(\cdot)$  denote the natural logarithm.

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

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

for every  $x, y \in \text{dom } \Phi$  and  $\alpha \in [0, 1]$ . Let  $\overline{\text{Conv}}_\mu(\mathbb{R}^n)$  denote the set of all proper lower semicontinuous  $\mu$ -convex functions. We simply denote  $\overline{\text{Conv}}_\mu(\mathbb{R}^n)$  by  $\overline{\text{Conv}}(\mathbb{R}^n)$  when  $\mu = 0$ . For  $\varepsilon \geq 0$ , the  $\varepsilon$ -subdifferential of  $\Phi$  at  $x \in \text{dom } \Phi$  is denoted by

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

We denote the subdifferential of  $\Phi$  at  $x \in \text{dom } \Phi$  by  $\partial\Phi(x)$ , which is the set  $\partial_0\Phi(x)$  by definition. For a given subgradient  $\Phi'(x) \in \partial\Phi(x)$ , we denote the linearization of convex function  $\Phi$  at  $x$  by  $\ell_\Phi(\cdot, x)$ , which is defined as

$$\ell_\Phi(\cdot, x) = \Phi(x) + \langle \Phi'(x), \cdot - x \rangle. \quad (2)$$

## 2 Universal composite subgradient and proximal bundle methods

This section presents two completely universal methods for solving (1), namely, U-CS in Subsection 2.1 and U-PB in Subsection 2.2. Their iteration-complexity guarantees are described in this section but their proofs are postponed to Section 4, after the presentation of a general framework for (not necessarily composite) convex optimization problems in Section 3. Viewing U-CS and U-PB as special cases of this framework will enable us to prove the main complexity results of this section in a unified manner in Section 4.

We assume that

- (A1)  $f, h \in \overline{\text{Conv}}(\mathbb{R}^n)$  are such that  $\text{dom } h \subset \text{dom } f$ , and a subgradient oracle, i.e., a function  $f' : \text{dom } h \rightarrow \mathbb{R}^n$  satisfying  $f'(x) \in \partial f(x)$  for every  $x \in \text{dom } h$ , is available;
- (A2) there exists  $(M_f, L_f) \in \mathbb{R}_+^2$  such that for every  $x, y \in \text{dom } h$ ,

$$\|f'(x) - f'(y)\| \leq 2M_f + L_f\|x - y\|;$$

- (A3) the set of optimal solutions  $X_*$  of problem (1) is nonempty.

It is known that that (A2) implies that for every  $x, y \in \text{dom } h$ ,

$$f(x) - \ell_f(x; y) \leq 2M_f\|x - y\| + \frac{L_f}{2}\|x - y\|^2. \quad (3)$$

When  $M_f = 0$ , (A2) implies that  $f$  is  $L_f$ -smooth, and when  $L_f = 0$ , (A2) implies that  $f$  is  $M_f$ -Lipschitz continuous. Hence, our analysis applies to both smooth and nonsmooth, as well as hybrid, instances of (1), where both  $L_f$  and  $M_f$  are positive.

Recall that for a given function  $\phi \in \overline{\text{Conv}}(\mathbb{R}^n)$ , the intrinsic convex parameter  $\mu_\phi$  of  $\phi$  is defined as the largest scalar  $\mu$  such that  $\phi(\cdot) - \mu\|\cdot\|^2/2$  is convex. Clearly,  $\mu_\phi \geq 0$  for any  $\phi \in \overline{\text{Conv}}(\mathbb{R}^n)$ . Even though, for the sake of completeness, the complexity results developed in this paper apply to any  $\phi \in \overline{\text{Conv}}(\mathbb{R}^n)$ , our main interest in this paper is to develop  $\mu_\phi$ -universal methods for instances of (1) such that

$$\mu_\phi > 0.$$

Our goal in this section is to describe two completely universal methods for solving (1) and describe their iteration-complexity guarantees for finding a  $\bar{\varepsilon}$ -solution  $\bar{x}$  of (1), i.e., a point  $\bar{x}$  such that  $\phi(\bar{x}) - \phi_* \leq \bar{\varepsilon}$ , in terms of  $(M_f, L_f, \mu_\phi)$ ,  $\bar{\varepsilon}$ , and  $d_0$ , where  $d_0$  is the distance of the initial iterate  $\hat{x}_0$  with respect to  $X_*$ , i.e.,

$$d_0 := \min\{\|x - \hat{x}_0\| : x \in X_*\}. \quad (4)$$

For any given positive scalars  $\bar{M}$  and  $\bar{\mu}$ , we consider two classes of instances  $(f, h)$  of problem (1), namely: 1)  $L_f = 0$ ,  $M_f \leq \bar{M}$ , and  $\mu_\phi \geq \bar{\mu}$ ; and 2)  $L_f = 0$ ,  $M_f \leq \bar{M}$  and  $\mu_h \geq \bar{\mu}$ . The following remarks about them

hold: i) the second class is a subset of the first one; and, ii) the lower bound  $\Omega(\bar{M}^2/(\bar{\mu}\bar{\varepsilon}))$  has been established for the second class in [18] using an instance where  $f$  is piece-wise linear and  $h(x) = \bar{\mu}\|x\|^2/2$  (see (73) and (74) of [16], which is an extended version of [18]). These two observations imply that  $\Omega(\bar{M}^2/(\bar{\mu}\bar{\varepsilon}))$  is also a lower bound for the first class. Both the adaptive subgradient method and the proximal bundle method studied in Appendix C.2 and Section 3 of [20], respectively, are shown to have the following properties: i) they are  $\mu_h$ -universal for any instance  $(f, h)$  of (1) with arbitrary  $(M_f, L_f)$ ; and ii) they achieve near optimal complexity  $\tilde{O}(\bar{M}^2/(\bar{\mu}\bar{\varepsilon}))$  for the second class mentioned above.

## 2.1 A universal composite subgradient method

This subsection presents the first of two aforementioned universal methods for solving (1). Specifically, it states a variant of the universal primal gradient method of [26] for solving (1), referred to as the U-CS method, and establishes its complete universality. Moreover, it is shown that its  $\mu_\phi$ -universality does not require any line search on the intrinsic convex parameter  $\mu_\phi$ .

We start by stating U-CS.

---

U-CS( $\hat{x}_0, \chi, \lambda_0, \bar{\varepsilon}$ )

---

**Input:**  $(\hat{x}_0, \chi, \lambda_0, \bar{\varepsilon}) \in \text{dom } h \times [0, 1) \times \mathbb{R}_{++} \times \mathbb{R}_{++}$ .

0. Set  $\lambda = \lambda_0$  and  $k = 1$ ;

1. Compute

$$x = \underset{u \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ \ell_f(u; \hat{x}_{k-1}) + h(u) + \frac{1}{2\lambda} \|u - \hat{x}_{k-1}\|^2 \right\}; \quad (5)$$

if  $\phi(x) - \phi_* \leq \bar{\varepsilon}$ , then **stop**;

2. **If** the inequality

$$f(x) - \ell_f(x; \hat{x}_{k-1}) - \frac{1-\chi}{2\lambda} \|x - \hat{x}_{k-1}\|^2 \leq \frac{(1-\chi)\bar{\varepsilon}}{2}$$

does not hold, **then** set  $\lambda = \lambda/2$  and go to step 1; **else**, set  $\lambda_k = \lambda$ ,  $\hat{x}_k = x$ ,  $k \leftarrow k+1$ , and go to step 1.

---

We now make some remarks about U-CS. First, U-CS does not require any knowledge about the parameter pair  $(M_f, L_f)$  or the intrinsic convex parameter  $\mu_\phi$ . Second, the U-CS method with  $\chi = 0$  is exactly the universal primal gradient method analyzed in [26], which establishes its universality with respect to the parameters of a Hölder condition on  $f$  but not with respect to  $\mu_h$ . The subsequent paper [20] establishes the  $\mu_h$ -universality of the U-CS method with  $\chi = 0$  (see Proposition C.2 of [20]), but left its  $\mu_\phi$ -universality as an open question. Third, Theorem 2.2 below shows that U-CS, with  $\chi \in (0, 1)$  and  $\chi$  not too close to 1, is  $\mu_h$ -universal, and also  $\mu_\phi$ -universal if in addition  $\chi$  is not close to 0. The latter extra condition on  $\chi$  is because its iteration complexity bound, in terms of  $\mu_\phi$  only, depends on  $\chi^{-1}$ , and hence does not apply to the universal primal gradient method of [26] (i.e., U-CS with  $\chi = 0$ ).

We now make some comments about the universality of the U-CS method. First, it performs a line search on  $\lambda$  on step 2 due to the fact that  $(M_f, L_f)$  is not assumed to be known. Lemma 2.1 below shows that when  $(M_f, L_f)$  is known then a constant prox stepsize, which provably satisfies the inequality in step 2 for every  $k$ , can be used throughout the method. Thus, the line search on step 2 is only performed so as to make the U-CS method  $(M_f, L_f)$ -parameter-free. Second, the  $\mu$ -universality of the U-CS method will be established in Theorem 2.2 regardless of whether the line search is performed (i.e., when  $(M_f, L_f)$  is not known) or is not needed (i.e., when  $(M_f, L_f)$  is known).

**Lemma 2.1** *If  $\lambda$  satisfies*

$$\lambda \leq \frac{(1-\chi)^2 \bar{\varepsilon}}{4M_f^2 + \bar{\varepsilon}L_f}, \quad (6)$$

*then the point  $x = x_k(\lambda)$  computed in (5) satisfies the inequality in step 2. As a consequence, if the initial prox stepsize  $\lambda_0$  is chosen to be the right hand side of (6), then  $\lambda$  remains constant throughout the method.*

**Proof:** Assume that  $\lambda$  satisfies (6). Using (3) with  $(M_f, L_f, u, v) = (M_f, L_f, x, \hat{x}_{k-1})$  and the inequality  $2M_f\Delta - a\Delta^2 \leq M_f^2/a$  with  $\Delta = \|x - \hat{x}_{k-1}\|$  and  $a = (1 - \chi - \lambda L_f)/2\lambda$ , we have

$$\begin{aligned} f(x) - \ell_f(x; \hat{x}_{k-1}) + \frac{\chi - 1}{2\lambda} \|x - \hat{x}_{k-1}\|^2 &\stackrel{(3)}{\leq} 2M_f \|x - \hat{x}_{k-1}\| + \frac{L_f}{2} \|x - \hat{x}_{k-1}\|^2 + \frac{\chi - 1}{2\lambda} \|x - \hat{x}_{k-1}\|^2 \\ &= 2M_f \|x - \hat{x}_{k-1}\| - \frac{1 - \chi - \lambda L_f}{2\lambda} \|x - \hat{x}_{k-1}\|^2 \\ &\leq \frac{2\lambda M_f^2}{1 - \chi - \lambda L_f} \stackrel{(6)}{\leq} \frac{(1 - \chi)\bar{\varepsilon}}{2}, \end{aligned}$$

where the last inequality is due to (6). Hence, the lemma is proved.  $\blacksquare$

The main result of this subsection, whose proof is postponed until Subsection 4.1, is stated next. Its goal is to establish the complete universality of U-CS.

**Theorem 2.2** *U-CS terminates in at most*

$$\min \left\{ \min \left[ \frac{1}{\chi} \left( 1 + \frac{Q(\bar{\varepsilon})}{\bar{\varepsilon}\mu_\phi} \right), 1 + \frac{Q(\bar{\varepsilon})}{\bar{\varepsilon}\mu_h} \right] \log \left( 1 + \frac{\mu_\phi d_0^2}{\bar{\varepsilon}} \right), \frac{d_0^2 Q(\bar{\varepsilon})}{\bar{\varepsilon}^2} \right\} + \left\lceil 2 \log \frac{\lambda_0 Q(\bar{\varepsilon})}{\bar{\varepsilon}} \right\rceil \quad (7)$$

iterations and

$$Q(\bar{\varepsilon}) = \frac{8M_f^2}{(1 - \chi)^2} + \bar{\varepsilon} \left( \lambda_0^{-1} + \frac{2L_f}{(1 - \chi)^2} \right). \quad (8)$$

We now make some remarks about Theorem 2.2. First, if  $\lambda_0^{-1} = \mathcal{O}(L_f)$  and  $\chi \in [0, 1)$  is such that  $(1 - \chi)^{-1} = \mathcal{O}(1)$ , then  $Q(\bar{\varepsilon})/\bar{\varepsilon} = \tilde{\mathcal{O}}(M_f^2/\bar{\varepsilon} + L_f)$ , and hence (7) in terms of  $\mu_h$  is

$$\tilde{\mathcal{O}} \left( \frac{M_f^2}{\bar{\varepsilon}\mu_h} + \frac{L_f}{\mu_h} \right), \quad (9)$$

which is the same well-known bound established for the U-CS method with  $\chi = 0$  (e.g., see Proposition C.2 of [20]).

Second, if in addition to the assumptions on the pair  $(\chi, \lambda_0)$  made in the first remark, the parameter  $\chi$  is also assumed to be not close to zero (i.e.,  $\chi^{-1} = \mathcal{O}(1)$ ), then the complexity bound (7) is also

$$\tilde{\mathcal{O}} \left( \frac{Q(\bar{\varepsilon})}{\chi \bar{\varepsilon} \mu_\phi} \right) = \tilde{\mathcal{O}} \left( \frac{M_f^2}{\bar{\varepsilon} \mu_\phi} + \frac{L_f}{\mu_\phi} \right), \quad (10)$$

which is considerably smaller than (9) when  $\mu_\phi \gg \mu_h$ .

Third, in view of the lower complexity remarks made in the paragraph following (4), bound (10) with  $(\mu_\phi, M_f)$  replaced by  $(\bar{\mu}, \bar{M})$  is optimal, up to logarithmic terms, for the class of instances  $(f, h)$  of problem (1) where  $L_f = 0$ ,  $M_f \leq \bar{M}$ , and  $\mu_\phi \geq \bar{\mu}$ .

Fourth, algorithms with optimal complexity for the class of instances  $(f, h)$  such that  $L_f \leq \bar{L}$ ,  $M_f \leq \bar{M}$ , and  $\mu_\phi \geq \bar{\mu}$  have been studied in [11, 26] for the case where  $\bar{\mu} = 0$ , and in [1, 2, 3, 4, 5, 6, 8, 13, 25, 29] for the case where  $\bar{M} = 0$  and  $\bar{\mu} > 0$ .

Finally, we end this subsection with a comparison of the complexity bound of Theorem 2.2 with the one derived in Corollary 5 of [29] for its subgradient variant under the assumption that  $f$  has quadratic growth. Specifically, under the assumption that  $M_f$  is known and  $L_f = 0$ , Corollary 5 of [29] shows that their subgradient variant has (parallel) iteration complexity bound

$$\mathcal{O} \left( \frac{M_f^2}{\bar{\varepsilon}\mu_f} + N + \left( \frac{M_f d_0}{2^N \bar{\varepsilon}} \right)^2 \right) \quad (11)$$

where  $N$  is the number of threads. As a consequence, they conclude that the above bound reduces to

$$\mathcal{O} \left( \frac{M_f^2}{\bar{\varepsilon}\mu_f} + \log \frac{1}{\bar{\varepsilon}} + M_f^2 d_0^2 \right)$$

when  $N = \mathcal{O}(\log \bar{\varepsilon}^{-1})$ . For sufficiently small  $\bar{\varepsilon} > 0$ , the above bound and Theorem 3.2.1 of [27] show that their subgradient variant is exactly optimal for the class of instances  $(f, h)$  of problem (1) where  $L_f = 0$ ,  $M_f \leq \bar{M}$ , and  $\mu_f \geq \bar{\mu}$ . However, bound (11) with  $N = 1$  reduces to  $\mathcal{O}(M_f^2 d_0^2 / \bar{\varepsilon}^2)$ , and hence does not imply that their single-thread subgradient variant is optimal for the aforementioned class.

## 2.2 A universal proximal bundle method

This subsection describes the U-PB method and states a result that ensures its complete universality.

Conditions (A1)-(A3) are assumed to hold in this subsection. We also assume

(A4) the diameter of the domain of  $h$  given by  $\max\{\|x - y\| : x, y \in \text{dom } h\}$  is bounded by  $D < +\infty$ .

The U-PB method is an extension of the GPB method of [20]. In contrast to GPB, we use an adaptive stepsize and introduce a maximal number  $\bar{N}$  (which can be as small as 2) of iterations for all cycles. Similarly to U-CS, U-PB is another instance of FSCO and we establish functional complexity for U-PB using the results of Section 3. Compared with the complexity results in [20], those obtained in this paper are sharper, since they are expressed in terms of  $\mu_\phi$  instead of  $\mu_h$ .

U-PB is based on the following bundle update (BU) blackbox which builds a model  $m_f^+ + h$  for  $f + h$  on the basis of a previous model  $m_f$  of  $f$  and of a new linearization  $\ell_f(\cdot, x)$  of  $f$ . This blackbox  $\text{BU}(x^c, x, m_f, \lambda)$  is given below and takes as inputs a prox-center  $x^c$ , a current approximate solution  $x$ , an initial model  $m_f$  for  $f$ , and a stepsize  $\lambda > 0$ .

---

$\text{BU}(x^c, x, m_f, \lambda)$

---

**Inputs:**  $\lambda \in \mathbb{R}_{++}$  and  $(x^c, x, m_f(\cdot)) \in \mathbb{R}^n \times \mathbb{R}^n \times \overline{\text{Conv}}(\mathbb{R}^n)$  such that  $m_f(\cdot) \leq f(\cdot)$  and

$$x = \underset{u \in \mathbb{R}^n}{\text{argmin}} \left\{ m_f(u) + h(u) + \frac{1}{2\lambda} \|u - x^c\|^2 \right\}.$$

Find function  $m_f^+(\cdot) \in \overline{\text{Conv}}(\mathbb{R}^n)$  such that

$$\max\{\bar{m}_f(\cdot), \ell_f(\cdot; x)\} \leq m_f^+(\cdot) \leq f(\cdot), \quad (12)$$

where  $\ell_f(\cdot; \cdot)$  is as in (2) and  $\bar{m}_f(\cdot)$  is such that

$$\bar{m}_f(\cdot) \leq f(\cdot), \quad \bar{m}_f(x) = m_f(x), \quad x = \underset{u \in \mathbb{R}^n}{\text{argmin}} \left\{ \bar{m}_f(u) + h(u) + \frac{1}{2\lambda} \|u - x^c\|^2 \right\}. \quad (13)$$

**Output:**  $m_f^+$ .

---

In the following, we give two examples of BU, namely two-cuts and multiple-cuts schemes. The proofs for the two schemes belonging to BU can be provided similarly to Appendix D of [20].

(E1) **two-cuts scheme:** We assume that  $m_f$  is of the form  $m_f = \max\{A_f, \ell_f(\cdot; x^-)\}$  where  $A_f$  is an affine function satisfying  $A_f \leq f$ . The scheme then sets  $A_f^+(\cdot) := \theta A_f(\cdot) + (1 - \theta)\ell_f(\cdot; x^-)$  and updates  $m_f^+$  as  $m_f^+(\cdot) := \max\{A_f^+(\cdot), \ell_f(\cdot; x)\}$ , where  $\theta \in [0, 1]$  satisfies

$$\begin{aligned} \frac{1}{\lambda}(x - x^c) + \partial h(x) + \theta \nabla A_f + (1 - \theta)f'(x^-) &\ni 0, \\ \theta A_f(x) + (1 - \theta)\ell_f(x; x^-) &= \max\{A_f(x), \ell_f(x; x^-)\}. \end{aligned}$$

(E2) **multiple-cuts scheme:** We assume that  $m_f$  has the form  $m_f = m_f(\cdot; B)$  where  $B \subset \mathbb{R}^n$  is the current bundle set and  $m_f(\cdot; B)$  is defined as  $m_f(\cdot; B) := \max\{\ell_f(\cdot; b) : b \in B\}$ . This scheme selects the next bundle set  $B^+$  so that  $B(x) \cup \{x\} \subset B^+ \subset B \cup \{x\}$  where  $B(x) := \{b \in B : \ell_f(x; b) = m_f(x)\}$ , and then outputs  $m_f^+ = m_f(\cdot; B^+)$ .

Before giving the motivation of U-PB, we briefly review the GPB method of [20]. GPB is an inexact proximal point method (PPM, with fixed stepsize) in that, given a prox-center  $\hat{x}_{k-1} \in \mathbb{R}^n$  and a prox stepsize  $\lambda > 0$ , it computes the next prox-center  $\hat{x}_k$  as a suitable approximate solution of the prox subproblem

$$\hat{x}_k \approx \underset{u \in \mathbb{R}^n}{\text{argmin}} \left\{ (f + h)(u) + \frac{1}{2\lambda} \|u - \hat{x}_{k-1}\|^2 \right\}. \quad (14)$$

More specifically, a sequence of prox bundle subproblems of the form

$$x_j = \operatorname{argmin}_{u \in \mathbb{R}^n} \left\{ (f_j + h)(u) + \frac{1}{2\lambda} \|u - \hat{x}_{k-1}\|^2 \right\}, \quad (15)$$

where  $f_j \leq f$  is a bundle approximation of  $f$ , is solved until for the first time an iterate  $x_j$  as in (15) approximately solves (14), and such  $x_j$  is then set to be  $\hat{x}_k$ . The bundle approximation  $f_j$  is sequentially updated, for example, according to either one of the schemes (E1) and (E2) described above.

U-PB is also an inexact PPM but with variable prox stepsizes (i.e., with  $\lambda$  in (14) replaced by  $\lambda_k$ ) instead of a constant one as in GPB. Given iteration upper limit  $\bar{N} \geq 1$  and prox-center  $\hat{x}_{k-1}$ , it adaptively computes  $\lambda_k > 0$  as follows: starting with  $\lambda = \lambda_{k-1}$ , it solves at most  $\bar{N}$  prox subproblems of the form (15) in an attempt to obtain an approximate solution of (14) and, if it fails, repeats this procedure with  $\lambda$  divided by 2; otherwise, it sets  $\lambda_k$  to be the first successful  $\lambda$  and  $\hat{x}_k$  as described in the previous paragraph.

We are now ready to state U-PB.

---

U-PB( $\hat{x}_0, \chi, \lambda_0, \bar{\varepsilon}, \bar{N}$ )

---

**Input:**  $(\hat{x}_0, \lambda_0, \chi, \bar{\varepsilon}, \bar{N}) \in \operatorname{dom} h \times \mathbb{R}_{++} \times [0, 1) \times \mathbb{R}_{++} \times \mathbb{N}_{++}$ .

0. Set  $\lambda = \lambda_0$ ,  $y_0 = \hat{x}_0$ ,  $N = 0$ ,  $j = 1$ , and  $k = 1$ . Find  $f_1 \in \overline{\operatorname{Conv}}(\mathbb{R}^n)$  such that  $\ell_f(\cdot; \hat{x}_0) \leq f_1(\cdot) \leq f(\cdot)$ ;

1. Compute  $x_j$  as in (15); if  $\phi(x_j) - \phi_* \leq \bar{\varepsilon}$  stop;

2. Compute

$$\bar{\phi}_j = \begin{cases} \phi(x_j) + \frac{\chi}{2\lambda} \|x_j - \hat{x}_{k-1}\|^2, & \text{if } N = 0, \\ \min \left\{ \bar{\phi}_{j-1}, \phi(x_j) + \frac{\chi}{2\lambda} \|x_j - \hat{x}_{k-1}\|^2 \right\}, & \text{otherwise,} \end{cases} \quad (16)$$

set  $N = N + 1$  and

$$t_j = \bar{\phi}_j - \left( (f_j + h)(x_j) + \frac{1}{2\lambda} \|x_j - \hat{x}_{k-1}\|^2 \right); \quad (17)$$

3. **If**  $t_j > (1 - \chi)\bar{\varepsilon}/2$  and  $N < \bar{N}$  **then**

3.a. perform a **null update**, i.e.: set  $f_{j+1} = \operatorname{BU}(\hat{x}_{k-1}, x_j, f_j, \lambda)$ ;

**else**

**if**  $t_j > (1 - \chi)\bar{\varepsilon}/2$  and  $N = \bar{N}$

3.b. perform a **reset update**, i.e., set  $\lambda \leftarrow \lambda/2$ ;

**else** (i.e.,  $t_j \leq (1 - \chi)\bar{\varepsilon}/2$  and  $N \leq \bar{N}$ )

3.c. perform a **serious update**, i.e., set  $\hat{x}_k = x_j$ ,  $\hat{\Gamma}_k = f_j + h$ ,  $\hat{y}_k = y_j$ ,  $\lambda_k = \lambda$ ;  
 $k \leftarrow k + 1$ ;

**end if**

3.d. set  $N = 0$  and find  $f_{j+1} \in \overline{\operatorname{Conv}}(\mathbb{R}^n)$  such that  $\ell_f(\cdot; \hat{x}_{k-1}) \leq f_{j+1} \leq f$ ;

**end if**

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

---

We now give further explanation about U-PB. U-PB performs three types of iterations, namely, null, reset, and serious, corresponding to the types of updates performed at the end. A reset (resp., serious) cycle of U-PB consists of a reset (resp., serious) iteration and all the consecutive null iterations preceding it. The index  $j$  counts the total number of iterations including null, reset, and serious ones. The index  $k$  counts the serious cycles which, together with the quantities  $\hat{x}_k$ ,  $\hat{y}_k$ , and  $\hat{\Gamma}_k$  computed at the end of cycle  $k$ , is used to cast U-PB as an instance of FSCO. All iterations within a cycle are referred to as inner iterations. The quantity  $N$  counts the number of inner iterations performed in the current cycle. Each cycle of U-PB performs at most  $\bar{N}$  iterations. A serious cycle successfully finds  $t_j \leq (1 - \chi)\bar{\varepsilon}/2$  within  $\bar{N}$  iterations, while a reset cycle fails to do so. In both cases, U-PB resets the counter  $N$  to 0 and starts a new cycle. The differences between the two

cases are: 1) the stepsize  $\lambda$  is halved at the end of a reset cycle, while it is kept as is at the end of a serious cycle; and 2) the prox-center is kept the same at the end of a reset cycle, but it is updated to the latest  $x_j$  at the end of a serious cycle.

We now make three remarks about U-PB. First, U-CS is a special case of U-PB with  $\bar{N} = 1$  and  $f_{j+1} = \ell_f(\cdot; \hat{x}_{k-1})$  in step 3.d. Second, it follows from the fact that  $f_j \leq f$  and the definition of  $t_j$  in (17) that the primal gap of the prox subproblem in (14) at the iterate  $y_j$  is upper bounded by  $t_j + (1 - \chi)\|y_j - \hat{x}_{k-1}\|^2/(2\lambda)$ . Hence, if  $t_j \leq (1 - \chi)\bar{\varepsilon}/2$ , then  $y_j$  is an  $\varepsilon_j$ -solution of (14) where  $\varepsilon_j = (1 - \chi)[\bar{\varepsilon}/2 + \|y_j - \hat{x}_{k-1}\|^2/(2\lambda)]$ . Third, the iterate  $y_j$  computed in step 2 of U-PB satisfies

$$y_j \in \text{Argmin} \left\{ \phi(x) + \frac{\chi}{2\lambda} \|x - \hat{x}_{k-1}\|^2 : x \in \{x_i, \dots, x_j\} \right\}, \quad (18)$$

where  $i$  denotes the first iteration index of the cycle containing  $j$ . In other words,  $y_j$  is the best point in terms of  $\phi(\cdot) + \chi\|\cdot - \hat{x}_{k-1}\|^2/(2\lambda)$  among all the points obtained in the course of solving (15) and the point  $\hat{y}_{k-1}$  obtained at the end of the previous cycle.

We now state the first main result of this subsection where the functional iteration complexity of U-PB is established.

**Theorem 2.3** *Assuming that  $\text{dom } h$  has a finite diameter  $D > 0$ . Let*

$$B = 8 + 12 \log \left( 1 + \frac{L_f^2 D^2 \bar{N}}{16M_f^2} \right) \quad (19)$$

and define  $U(\bar{\varepsilon})$  by

$$U(\bar{\varepsilon}) := \frac{4(M_f^2 + \bar{\varepsilon}L_f)B}{(1 - \chi)^2} + \frac{\bar{N}\bar{\varepsilon}}{\lambda_0}. \quad (20)$$

The total number of inner iterations (i.e., the ones indexed by  $j$ ) performed by U-PB is bounded by

$$\min \left\{ \min \left[ \frac{1}{\chi} \left( \bar{N} + \frac{U(\bar{\varepsilon})}{\bar{\varepsilon}\mu_\phi} \right), \bar{N} + \frac{U(\bar{\varepsilon})}{\bar{\varepsilon}\mu_h} \right] \log \left( 1 + \frac{\mu_\phi d_0^2}{\bar{\varepsilon}} \right), \frac{d_0^2 U(\bar{\varepsilon})}{\bar{\varepsilon}^2} \right\} + \bar{N} \left\lceil 2 \log \frac{\lambda_0 U(\bar{\varepsilon})}{\bar{N}\bar{\varepsilon}} \right\rceil. \quad (21)$$

The complexity result of Theorem 2.3, under the mild conditions that  $\chi$  is neither close to one nor zero (i.e.,  $\max\{\chi^{-1}, (1 - \chi)^{-1}\} = \mathcal{O}(1)$ ),

$$\bar{N} = \mathcal{O} \left( \frac{M_f^2 + \bar{\varepsilon}L_f}{\bar{\varepsilon}\mu_\phi} \min \{1, \lambda_0\mu_\phi\} \right),$$

and  $\mu_\phi \gg \mu_h$ , reduces to

$$\tilde{\mathcal{O}} \left( \frac{M_f^2}{\bar{\varepsilon}\mu_\phi} + \frac{L_f}{\mu_\phi} \right), \quad (22)$$

which is similar to the functional complexity bound (10) obtained for U-CS. Moreover, similar to U-CS, bound (22) with  $(\mu_\phi, M_f)$  replaced by  $(\bar{\mu}, \bar{M})$  is nearly optimal for the class of instances  $(f, h)$  of problem (1) where  $L_f = 0$ ,  $M_f \leq \bar{M}$ , and  $\mu_\phi \geq \bar{\mu}$ , because of the lower complexity remarks made in the paragraph following (4).

### 3 A functional framework for strongly convex optimization

This section presents a general framework, namely FSCO, for (strongly) convex optimization problems and establishes an oracle-complexity bound for it.

FSCO is presented in the context of the convex optimization problem

$$\phi_* := \min \{ \phi(x) : x \in \mathbb{R}^n \} \quad (23)$$

where  $\phi \in \overline{\text{Conv}}(\mathbb{R}^n)$ . It will be shown in this section that any instance of FSCO is  $\mu_\phi$ -universal. Moreover, its oracle-complexity bound will be used in Subsections 4.1 and 4.2 to establish the iteration-complexities of two important instances, namely, the U-CS method described in Subsection 2.1, and the U-PB method described

in Subsection 2.2. It is worth mentioning, though, that FSCO is only used to establish universality relative to  $\mu_\phi$ . Complete universality of the two specific instances mentioned above, and hence relative to the parameter pair  $(M_f, L_f)$ , is established by performing line searches on the prox stepsize  $\lambda_k$ , as described in step 2 of U-CS and in step 3 of U-PB.

Given a prox-center  $x^- \in \text{dom } \phi$ , every iteration of FSCO approximately solves a prox subproblem of the form

$$\min_{u \in \mathbb{R}^n} \left\{ \phi(u) + \frac{1}{2\lambda} \|u - x^-\|^2 \right\}, \quad (24)$$

for some  $\lambda > 0$ , to obtain an inexact solution  $y$  whose primal gap size is used to terminate FSCO. The black-box below, which is repeatedly invoked by FSCO, describes how the above subproblem is approximately solved.

---

Black-box (BB)  $(x^-, \chi, \varepsilon)$

---

**Input:**  $(x^-, \chi, \varepsilon) \in \text{dom } \phi \times [0, 1) \times \mathbb{R}_{++}$ .

**Output:**  $(x, y, \Gamma, \lambda) \in \text{dom } \phi \times \text{dom } \phi \times \overline{\text{Conv}}(\mathbb{R}^n) \times \mathbb{R}_{++}$  satisfying  $\Gamma \leq \phi$ ,

$$\phi(y) + \frac{\chi}{2\lambda} \|y - x^-\|^2 - \min_{u \in \mathbb{R}^n} \left\{ \Gamma(u) + \frac{1}{2\lambda} \|u - x^-\|^2 \right\} \leq \varepsilon, \quad (25)$$

and

$$x = \underset{u \in \mathbb{R}^n}{\text{argmin}} \left\{ \Gamma(u) + \frac{1}{2\lambda} \|u - x^-\|^2 \right\}. \quad (26)$$


---

We now make some remarks about BB. First,  $x^-$  is the current prox center and  $(\chi, \bar{\varepsilon})$  are tolerances used to quantify the inexactness of  $y$  as a solution of (24). Second, (25) quantifies the inexactness of  $y$  in terms of a functional certificate  $\Gamma$ , which is a bundle for  $\phi$ . More precisely,  $y$  is a  $\hat{\varepsilon}$ -solution of (24), where  $\hat{\varepsilon} = \varepsilon + (1 - \chi)\|y - x^-\|^2/(2\lambda)$ . Indeed, this follows from (25) and the fact that  $\Gamma \leq \phi$ . Finally, even though the next prox-center  $x$  is viewed as an output of BB, it is uniquely determined by  $\Gamma$  as in (26), and hence could alternatively be removed from the BB output. In other words, the relevant output produced by BB is  $(\lambda, y, \Gamma)$ .

The following result describes a key contraction inequality that holds for the two prox-centers  $x^-$  and  $x$  of BB. Its proof is postponed to Subsection 3.1 and uses a key result stated in Appendix A.1.

**Proposition 3.1** *Assume that  $\phi \in \overline{\text{Conv}}(\mathbb{R}^n)$  and let  $\mu = \mu_\phi$ . Assume also that  $(x, y, \Gamma, \lambda) = \text{BB}(x^-, \chi, \varepsilon)$  for some  $(x^-, \chi, \varepsilon) \in \text{dom } \phi \times [0, 1) \times \mathbb{R}_{++}^n$ , and  $\Gamma \in \overline{\text{Conv}}_\nu(\mathbb{R}^n)$  for some  $\nu \in [0, \mu]$ . Then, for every  $u \in \mathbb{R}^n$ , we have*

$$2\lambda [\phi(y) - \phi(u)] \leq \frac{2\lambda\varepsilon}{1 - \chi} + \|x^- - u\|^2 - (1 + \sigma)\|x - u\|^2, \quad (27)$$

where

$$\sigma = \sigma(\mu, \nu) := \frac{\lambda[\nu(1 + \mu\lambda) + \chi(\mu - \nu)]}{1 + \mu\lambda + \chi\lambda(\nu - \mu)}. \quad (28)$$

We now motivate the assumption made by Proposition 3.1 on  $\Gamma$  by considering the special case (1) of problem (23) where  $\phi = f + h$  and  $f$  and  $h$  are as in (A1)-(A3). In this setting, many algorithms (see e.g., U-CS and U-PB in Section 2) generate  $\Gamma$  functions of the form  $\Gamma_f + h$  where  $\Gamma_f$  is a convex (possibly, affine) function underneath  $f$ . Clearly, letting  $\nu = \mu_h$ , it follows that bundles  $\Gamma$  generated in this form are  $\nu$ -convex, since so is  $h$ . In conclusion, the  $\nu$ -convexity of  $\Gamma$  is inherited from the  $\nu$ -convexity assumption on the composite component  $h$  of  $\phi$ , i.e., the part of  $\phi$  that is not approximated by simpler functions in the prox subproblems generated while solving (1).

We make some remarks about Proposition 3.1. If  $\sigma$  in (28) is such that  $\sigma = \Theta(\lambda\mu)$ , then it will be shown in Theorem 3.2 that an algorithm that repeatedly calls BB finds an  $\bar{\varepsilon}$ -solution of (23) in  $\mathcal{O}((\lambda\mu)^{-1} \log(\bar{\varepsilon}^{-1}))$  calls. We mention two cases where  $\sigma = \Theta(\lambda\mu)$ . The first case is when  $\nu \approx \mu$  and  $\chi$  is arbitrary, but this case does not hold for many instances of (23). In particular, the ‘‘contraction’’ formula (27) for the pair  $(\nu, \chi) = (\mu, 0)$  is well-known (e.g., see Lemma 4.1(a) of [20]). The second case is when  $\chi^{-1} = \mathcal{O}(1)$  and  $\nu \in [0, \mu]$ . Indeed, it

follows from the first inequality in Lemma A.3(a) that  $\sigma = \Theta(\lambda\mu)$  when  $\chi^{-1} = \mathcal{O}(1)$ . Finally, even though  $\chi$  chosen very close to one satisfies the condition in the second case, such choice of  $\chi$  is not desirable as it makes the term  $2\lambda\bar{\varepsilon}/(1-\chi)$  in (27) too large. Thus, in view of the previous remarks, a good choice of  $\chi$  is one such that  $\max\{\chi^{-1}, (1-\chi)^{-1}\} = \mathcal{O}(1)$ .

We now describe FSCO.

---

FSCO( $\hat{x}_0, \chi, \bar{\varepsilon}$ )

---

0. Let  $(\hat{x}_0, \chi, \bar{\varepsilon}) \in \text{dom } \phi \times [0, 1) \times \mathbb{R}_+$  be given and set  $k = 1$ ;
  1. Compute  $(\hat{x}_k, \hat{y}_k, \hat{\Gamma}_k, \lambda_k) = \text{BB}(\hat{x}_{k-1}, \chi, (1-\chi)\bar{\varepsilon}/2)$ ;
  2. Check whether  $\phi(\hat{y}_k) - \phi_* \leq \bar{\varepsilon}$  and if so **stop**; else set  $k \leftarrow k + 1$  and go to step 1.
- 

We now make some comments about FSCO.

First, FSCO generates two sequences, i.e.,  $\{\hat{x}_k\}$  and  $\{\hat{y}_k\}$ . If FSCO stops, then the last iterate of the second one is guaranteed to be a  $\bar{\varepsilon}$ -solution of (23) due to step 2 of FSCO. Hence, the second sequence ensures that FSCO stops, while the first one should be viewed as the sequence of prox-centers generated by the bundle sequence  $\{\hat{\Gamma}_k\}$ .

Second, since BB does not provide a specific recipe for computing  $(\hat{x}_k, \hat{y}_k, \hat{\Gamma}_k, \lambda_k)$  in step 1, FSCO is not a completely specified algorithm but rather a framework built with the sole purpose of studying  $\mu_\phi$ -universality of its instances. Specifically, FSCO provides minimal conditions on its generated sequence  $(\hat{x}_k, \hat{y}_k, \hat{\Gamma}_k, \lambda_k)$  (i.e., only that is generated by iteratively calling BB) to ensure  $\mu_\phi$ -universality of its instances.

Finally, for some instances of FSCO, such as U-CS and U-PB,  $\lambda_k$  is computed by performing a line search on  $\lambda$  within specific implementations of BB, because some parameters associated with  $\phi$  (e.g.,  $L_f$  and  $M_f$  in the setting of Section 2) are assumed to be unknown. When these parameters are known, the line search can be avoided by choosing  $\lambda_k$  in terms of these parameters.

The main complexity result for FSCO is relative to the instances that satisfy the following condition:

(F1) there exists  $\underline{\lambda} > 0$  such that  $\lambda_k \geq \underline{\lambda}$  for every  $k \geq 0$ .

**Theorem 3.2** *The number of iterations performed by any instance of FSCO that satisfies condition (F1) is bounded by*

$$\min \left\{ \min \left[ \frac{1}{\chi} \left( 1 + \frac{1}{\underline{\lambda}\mu} \right), 1 + \frac{1}{\underline{\lambda}\nu} \right] \log \left( 1 + \frac{\mu d_0^2}{\bar{\varepsilon}} \right), \frac{d_0^2}{\underline{\lambda}\bar{\varepsilon}} \right\}, \quad (29)$$

where the scalars  $\chi$  and  $\bar{\varepsilon}$  are input to FSCO,  $d_0$  is as in (4), and

$$\mu = \mu_\phi, \quad \nu := \min \left\{ \mu_\phi, \inf_k \mu(\hat{\Gamma}_k) \right\}. \quad (30)$$

(By convention, (29) should be understood as being equal to  $d_0^2/(\underline{\lambda}\bar{\varepsilon})$  when  $\mu = 0$ .)

**Proof:** It is straightforward to see from Lemma A.3(b) that  $\sigma(\lambda) \geq \underline{\sigma} := \sigma(\underline{\lambda})$  for  $\lambda \geq \underline{\lambda}$ , where  $\sigma(\lambda)$  denotes the  $\sigma$  defined in (28), considered as a function of  $\lambda$ . Note that we invoke the oracle  $(\hat{x}_k, \hat{y}_k, \hat{\Gamma}_k, \lambda_k) = \text{BB}(\hat{x}_{k-1}, \chi, (1-\chi)\bar{\varepsilon}/2)$  in step 1. Using (27), assumption (F1), and the observation that  $\sigma(\lambda_k) \geq \underline{\sigma}$ , we have

$$2\lambda_k [\phi(\hat{y}_k) - \phi(u)] \leq \lambda_k \bar{\varepsilon} + \|\hat{x}_{k-1} - u\|^2 - (1 + \underline{\sigma}) \|\hat{x}_k - u\|^2, \quad (31)$$

where

$$\underline{\sigma} = \frac{\underline{\lambda}[\nu(1 + \mu\underline{\lambda}) + \chi(\mu - \nu)]}{1 + \mu\underline{\lambda} + \chi\underline{\lambda}(\nu - \mu)}, \quad (32)$$

and  $\mu$  and  $\nu$  are as in (30). It is easy to see that (31) with  $u = x_*$ , where  $x_*$  denotes the closest solution of (23) to the initial point  $\hat{x}_0$ , satisfies

$$\gamma_k \eta_k \leq \alpha_{k-1} - (1 + \sigma) \alpha_k + \gamma_k \delta$$

with

$$\gamma_k = 2\lambda_k, \quad \eta_k = \phi(\hat{y}_k) - \phi_*, \quad \alpha_k = \|\hat{x}_k - x_*\|^2, \quad \delta = \frac{\bar{\varepsilon}}{2}, \quad \sigma = \underline{\sigma}, \quad \underline{\gamma} = 2\underline{\lambda}, \quad (33)$$

where  $\underline{\gamma}$  satisfies  $\gamma_k \geq \underline{\gamma}$ .

Using Lemma A.2(b) with the above parameters, Lemma A.3(c), and the definition of  $\underline{\sigma}$  in (32) that the complexity to find a  $\bar{\varepsilon}$ -solution is

$$\min \left\{ \frac{1 + \underline{\sigma}}{\underline{\sigma}} \log \left( 1 + \frac{\underline{\sigma} d_0^2}{\underline{\lambda} \bar{\varepsilon}} \right), \frac{d_0^2}{\underline{\lambda} \bar{\varepsilon}} \right\} \leq \min \left\{ \min \left[ \frac{1}{\chi} \left( 1 + \frac{1}{\underline{\lambda} \mu} \right), 1 + \frac{1}{\underline{\lambda} \nu} \right] \log \left( 1 + \frac{\underline{\sigma} d_0^2}{\underline{\lambda} \bar{\varepsilon}} \right), \frac{d_0^2}{\underline{\lambda} \bar{\varepsilon}} \right\}. \quad (34)$$

Moreover, we note that  $\underline{\sigma} \leq \underline{\lambda} \mu$  in view of the second inequality in Lemma A.3(a). Finally, this observation and (34) immediately imply the complexity in (29).  $\blacksquare$

We now comment on the complexity bound obtained in Theorem 3.2. First, under the assumption that  $\mu > 0$ , the bound

$$\min \left\{ \frac{1}{\chi} \left( 1 + \frac{1}{\underline{\lambda} \mu} \right), 1 + \frac{1}{\underline{\lambda} \nu} \right\} \log \left( 1 + \frac{\mu d_0^2}{\bar{\varepsilon}} \right) \quad (35)$$

implied by (29) is meaningful only when either  $\chi > 0$  or  $\nu > 0$  (otherwise, it should be understood as being infinity). Second, the validity of the second bound in (35) has been established for some composite subgradient and proximal bundle methods (see for example [7, 18]). Third, if  $\mu \gg \nu$  and  $\chi$  is sufficiently bounded away from zero, the smallest term in (35) is the first one, in which case (29) reduces to

$$\frac{1}{\chi} \left( 1 + \frac{1}{\underline{\lambda} \mu} \right) \log \left( 1 + \frac{\mu d_0^2}{\bar{\varepsilon}} \right) = \tilde{\mathcal{O}} \left( \frac{1}{\underline{\lambda} \mu} \right).$$

Fourth, since the second bound  $d_0^2/(\underline{\lambda} \bar{\varepsilon})$  does not depend on  $\nu$ ,  $\mu$  and  $\chi$ , it holds for any parameters  $\mu \geq \nu \geq 0$  and  $\chi \in [0, 1)$ .

### 3.1 Proof of Proposition 3.1

We need the following technical result before proving Proposition 3.1.

**Lemma 3.3** *Define*

$$s := \frac{x^- - x}{\lambda}, \quad (36)$$

$$\eta := \phi(y) - \Gamma(x) - \langle s, y - x \rangle. \quad (37)$$

*Then, the following statements hold:*

a)  $s \in \partial \Gamma(x)$ , i.e., for every  $u \in \mathbb{R}^n$ ,

$$\Gamma(u) \geq \Gamma(x) + \langle s, u - x \rangle; \quad (38)$$

b)  $s \in \partial_\eta \phi(y)$  and

$$0 \leq 2\lambda\eta \leq 2\lambda\varepsilon - \|y - x\|^2 + (1 - \chi)\|y - x^-\|^2. \quad (39)$$

**Proof:** (a) The optimality condition of (26) yields  $0 \in \partial \Gamma_k(x) + (x - x^-)/\lambda$ , which together with the definition of  $s$  in (36) implies that the inclusion in a) holds. Relation (38) immediately follows from the inclusion in a) and the definition of the subdifferential.

(b) The relation  $\phi \geq \Gamma$  (see step 1 of FSCO), (38), the definition of  $\eta$  in (37) imply that for every  $u \in \text{dom } \phi$ ,

$$\phi(u) \geq \Gamma(u) \stackrel{(38)}{\geq} \Gamma(x) + \langle s, u - x \rangle \stackrel{(37)}{=} \phi(y) + \langle s, u - y \rangle - \eta,$$

which yields the inclusion in b). Taking  $u = y$  in the above inequality gives  $\eta \geq 0$ , and hence the first inequality in (39) holds. Using the definitions of  $s$  and  $\eta$  in (36) and (37), respectively, and (25) and (26), we have

$$\begin{aligned} \eta &\stackrel{(37)}{=} \phi(y) - \Gamma(x) - \langle s, y - x \rangle \\ &\stackrel{(25), (26)}{\leq} \left[ \varepsilon - \frac{\chi}{2\lambda} \|y - x^-\|^2 + \frac{1}{2\lambda} \|x - x^-\|^2 \right] - \langle s, y - x \rangle \\ &\stackrel{(36)}{=} \varepsilon - \frac{\chi}{2\lambda} \|y - x^-\|^2 + \frac{1}{2\lambda} \|x - x^-\|^2 + \left\langle \frac{x - x^-}{\lambda}, y - x \right\rangle \\ &= \varepsilon + \frac{1 - \chi}{2\lambda} \|y - x^-\|^2 - \frac{1}{2\lambda} \|y - x\|^2. \end{aligned}$$

Hence, the second inequality in (39) holds.  $\blacksquare$

Using a technical result, Lemma A.1 from Appendix A.1, we are able to partially recover the intrinsic convexity  $\mu_\phi$ , even though it is not assumed to be known from either the objective  $\phi$  or the model  $\Gamma$ .

**Lemma 3.4** *Let  $\mu = \mu_\phi$  and assume that  $\Gamma \in \overline{\text{Conv}}_\nu(\mathbb{R}^n)$  for some  $\nu \in [0, \mu]$ . For any  $\zeta \in [0, 1)$  and every  $u \in \text{dom } \phi$ , we have*

$$\phi(u) \geq \phi(y) + \langle s, u - y \rangle + \frac{\zeta(\mu - \nu)}{2} \|u - y\|^2 - \frac{\eta}{1 - \zeta} + \frac{\zeta}{1 - \zeta} \frac{\nu}{2} \|y - x\|^2 + \frac{\nu}{2} \|u - x\|^2, \quad (40)$$

where  $s$  and  $\eta$  are as in (36) and (37), respectively.

**Proof:** Define

$$\tilde{s} = s + \nu(x^- - x), \quad \tilde{\eta} = \eta - \frac{\nu}{2} \|y - x\|^2, \quad \tilde{\phi} = \phi - \frac{\nu}{2} \|\cdot - x^-\|^2, \quad \tilde{\Gamma} = \Gamma - \frac{\nu}{2} \|\cdot - x^-\|^2. \quad (41)$$

It follows from Lemma 3.3(a) and the above definitions of  $\tilde{s}$  and  $\tilde{\Gamma}$  that  $\tilde{s} \in \partial \tilde{\Gamma}(x)$ . In view of (41) and the assumption that  $\Gamma \in \overline{\text{Conv}}_\nu(\mathbb{R}^n)$  for some  $\nu \in [0, \mu]$ , we observe that  $\tilde{\Gamma}$  is convex and  $\tilde{\phi} \geq \tilde{\Gamma}$ . Hence, following a similar argument as in Lemma 3.3(b), we conclude that  $\tilde{s} \in \partial_{\tilde{\eta}} \tilde{\phi}(y)$ . Using this inclusion, the fact that  $\tilde{\phi}$  is  $(\mu - \nu)$ -convex, and Lemma A.1 with  $(\psi, \xi, y, v, \eta, Q) = (\tilde{\phi}, \mu - \nu, y, \tilde{s}, \tilde{\eta}, I)$ , we have for any  $\tau \in (0, \infty]$  and every  $u \in \text{dom } \phi$ ,

$$\tilde{\phi}(u) \stackrel{(66)}{\geq} \tilde{\phi}(y) + \langle \tilde{s}, u - y \rangle + \frac{\mu - \nu}{2(1 + \tau)} \|u - y\|^2 - (1 + \tau^{-1})\tilde{\eta}.$$

Taking  $\tau = (1 - \zeta)/\zeta$  for  $\zeta \in [0, 1)$  in the above inequality, we obtain

$$\tilde{\phi}(u) \geq \tilde{\phi}(y) + \langle \tilde{s}, u - y \rangle + \frac{\zeta(\mu - \nu)}{2} \|u - y\|^2 - \frac{\tilde{\eta}}{1 - \zeta}.$$

Finally, (40) follows from using the definitions of  $\tilde{\phi}$ ,  $\tilde{s}$ , and  $\tilde{\eta}$  in (41), and rearranging the terms in the above inequality.  $\blacksquare$

Noting that in the proof of Lemma 3.4, we apply Lemma A.1 to  $\tilde{\phi}$  in (41). If we instead apply Lemma A.1 to  $\phi$  directly, we obtain for every  $u \in \text{dom } \phi$ ,

$$\phi(u) \geq \phi(y) + \langle s, u - y \rangle + \frac{\zeta\mu}{2} \|u - y\|^2 - \frac{\eta}{1 - \zeta}. \quad (42)$$

This inequality is weaker than the one in (40). Indeed, applying the Cauchy-Schwarz inequality, we have

$$\left( \frac{1 - \zeta}{\zeta} + 1 \right) \left( \frac{\zeta}{1 - \zeta} \frac{\nu}{2} \|y - x\|^2 + \frac{\nu}{2} \|u - x\|^2 \right) \geq \frac{\nu}{2} (\|y - x\| + \|u - x\|)^2 \geq \frac{\nu}{2} \|u - y\|^2,$$

where the last inequality is due to the triangle inequality. Hence, we have

$$\frac{\zeta}{1 - \zeta} \frac{\nu}{2} \|y - x\|^2 + \frac{\nu}{2} \|u - x\|^2 \geq \frac{\zeta\nu}{2} \|u - y\|^2,$$

and thus (40) implies (42).

If  $\zeta$  is chosen to be 0, then (42) becomes

$$\phi(u) \geq \phi(y) + \langle s, u - y \rangle - \eta \quad \forall u \in \text{dom } \phi,$$

which is equivalent to the inclusion in Lemma 3.3(b). This means that if we use (42), then we cannot recover the intrinsic convex parameter  $\mu_\phi$  when  $\zeta = 0$ . However, even if  $\zeta = 0$ , (40) still preserves a  $\nu$ -strongly convex lower approximation to  $\phi$ . This justifies the necessity of the assumption that  $\Gamma \in \overline{\text{Conv}}_\nu(\mathbb{R}^n)$ .

Now we are ready to present the proof of Proposition 3.1.

**Proof of Proposition 3.1.** It follows from the definition of  $s$  in (36) and Lemma 3.4 with  $\zeta = \chi$ ,

$$\begin{aligned} & \|x^- - u\|^2 - \|x - u\|^2 - (\|x^- - y\|^2 - \|x - y\|^2) = -2\lambda \langle s, u - y \rangle \\ & \stackrel{(40)}{\geq} 2\lambda [\phi(y) - \phi(u)] + \chi(\mu - \nu)\lambda \|u - y\|^2 - \frac{2\lambda\eta}{1 - \chi} + \frac{\chi\nu\lambda}{1 - \chi} \|y - x\|^2 + \nu\lambda \|u - x\|^2. \end{aligned} \quad (43)$$

Rearranging the terms in (43) and using Lemma 3.3(b), we have

$$\begin{aligned}
& \|x^- - u\|^2 - (1 + \nu\lambda)\|x - u\|^2 - 2\lambda[\phi(y) - \phi(u)] \\
& \stackrel{(43)}{\geq} \|x^- - y\|^2 + \left(\frac{\chi\nu\lambda}{1-\chi} - 1\right)\|x - y\|^2 + \chi(\mu - \nu)\lambda\|u - y\|^2 - \frac{2\lambda\eta}{1-\chi} \\
& \stackrel{(39)}{\geq} -\frac{2\lambda\varepsilon}{1-\chi} + \frac{\chi(1+\nu\lambda)}{1-\chi}\|x - y\|^2 + \chi(\mu - \nu)\lambda\|u - y\|^2.
\end{aligned} \tag{44}$$

Rearranging the terms in the above inequality, we have

$$\begin{aligned}
2\lambda[\phi(y) - \phi(u)] & \leq \frac{2\lambda\varepsilon}{1-\chi} + \|x^- - u\|^2 - (1 + \nu\lambda)\|x - u\|^2 \\
& \quad - \chi\left(\frac{1+\nu\lambda}{1-\chi}\|x - y\|^2 + (\mu - \nu)\lambda\|u - y\|^2\right).
\end{aligned} \tag{45}$$

Using the triangle inequality and the fact that  $(a_1 + a_2)^2 \leq (b_1^{-1} + b_2^{-1})(a_1^2 b_1 + a_2^2 b_2)$  with  $(a_1, a_2) = (\|x - y\|, \|u - y\|)$  and  $(b_1, b_2) = ((1 + \nu\lambda)/(1 - \chi), (\mu - \nu)\lambda)$  we have

$$\|x - u\|^2 \leq \left(\frac{1-\chi}{1+\nu\lambda} + \frac{1}{(\mu-\nu)\lambda}\right) \left(\frac{1+\nu\lambda}{1-\chi}\|x - y\|^2 + (\mu - \nu)\lambda\|u - y\|^2\right). \tag{46}$$

Plugging the above inequality into (45), we have

$$2\lambda[\phi(y) - \phi(u)] \leq \frac{2\lambda\varepsilon}{1-\chi} + \|x^- - u\|^2 - \left[1 + \nu\lambda + \chi\left(\frac{1-\chi}{1+\nu\lambda} + \frac{1}{(\mu-\nu)\lambda}\right)^{-1}\right]\|x - u\|^2,$$

which is the same as (27) after simplification. ■

## 4 Proofs of Theorems 2.2 and 2.3

### 4.1 Proof of Theorem 2.2

The following result shows that U-CS is an instance of FSCO and that assumption (F1) of Section 3 is satisfied.

**Proposition 4.1** *The following statements hold for U-CS:*

- a)  $\{\lambda_k\}$  is a non-increasing sequence;
- b) for every  $k \geq 1$ , we have

$$\hat{x}_k = \operatorname{argmin}_{u \in \mathbb{R}^n} \left\{ \ell_f(u; \hat{x}_{k-1}) + h(u) + \frac{1}{2\lambda_k} \|u - \hat{x}_{k-1}\|^2 \right\}, \tag{47}$$

$$f(\hat{x}_k) - \ell_f(\hat{x}_k; \hat{x}_{k-1}) + \frac{\chi - 1}{2\lambda_k} \|\hat{x}_k - \hat{x}_{k-1}\|^2 \leq \frac{(1 - \chi)\bar{\varepsilon}}{2}, \tag{48}$$

$$\lambda_k \geq \min \left\{ \frac{(1 - \chi)^2 \bar{\varepsilon}}{8M_f^2 + 2\bar{\varepsilon}L_f}, \lambda_0 \right\}. \tag{49}$$

- c) U-CS is a special case of FSCO where:

- i)  $\hat{y}_k = \hat{x}_k$  and  $\hat{\Gamma}_k(\cdot) = \ell_f(\cdot; \hat{x}_{k-1}) + h(\cdot)$  for every  $k \geq 1$ ;
- ii) assumption (F1) is satisfied with  $\underline{\lambda}$  given by (49).

**Proof:** a) This statement directly follows from the description of U-CS.

b) Relations (47) and (48) directly follow from the description of U-CS. Next, we prove (49). Supposing  $\lambda_{k-1} \leq (1 - \chi)^2 \bar{\varepsilon} / (4M_f^2 + \bar{\varepsilon}L_f)$ , then it follows from Lemma 2.1 that (48) holds with  $\lambda_k$  replaced by  $\lambda_{k-1}$ .

This indicates that if  $\lambda$  is small enough, then it will remain unchanged. Therefore, following from the update scheme of  $\lambda$  in step 2 of U-CS, there is a lower bound as in (49).

c) Relations (47) and (48) are the analogues of relations (25) and (26) of FSCO with  $\hat{y}_k = \hat{x}_k$  and  $\hat{\Gamma}_k$  as in (i). Inequality (49) shows that Assumption (F1) is satisfied. ■

**Proof of Theorem 2.2.** Define

$$\bar{k} = \left\lceil 2 \log \max \left\{ \frac{\lambda_0(8M_f^2 + 2\bar{\varepsilon}L_f)}{(1-\chi)^2\bar{\varepsilon}}, 1 \right\} \right\rceil. \quad (50)$$

Observe that  $\lambda_0/2^k \leq \underline{\lambda}$  for  $k \geq \bar{k}$  where  $\underline{\lambda}$  is given by the right-hand side of (49), and from Proposition 4.1 that we cannot halve  $\lambda$  more than  $\bar{k}$  iterations. It follows from (50) and the definition of  $Q(\bar{\varepsilon})$  in (8) that

$$\bar{k} \leq \left\lceil 2 \log \frac{\lambda_0 Q(\bar{\varepsilon})}{\bar{\varepsilon}} \right\rceil.$$

Therefore, the second term on the right-hand side of (7) gives an upper bound on the number of iterations with backtracking of  $\lambda$ . We now provide a bound on the number of remaining iterations to obtain a  $\bar{\varepsilon}$ -solution with U-CS. In the  $k$ -th iteration of U-CS, the model  $\hat{\Gamma}_k$  for  $\phi = f + h$  is simply  $\ell_f(\cdot, \hat{x}_{k-1}) + h(\cdot)$  where  $\ell_f$  is the linearization of  $f$  as in (2), hence  $\mu(\hat{\Gamma}_k) = \mu_h \leq \mu_\phi$  and  $\nu = \mu_h$  in (30). Since Proposition 4.1 shows that U-CS is a special case of FSCO, Theorem 3.2 immediately gives for U-CS the upper bound (29) with  $\underline{\lambda}$  given by (49) on the number of the remaining iterations (where  $\lambda$  is not halved) required to find a  $\bar{\varepsilon}$ -solution of (1). Using the inequality

$$\frac{1}{\underline{\lambda}} \stackrel{(49)}{=} \max \left\{ \frac{8M_f^2 + 2\bar{\varepsilon}L_f}{(1-\chi)^2\bar{\varepsilon}}, \frac{1}{\lambda_0} \right\} \leq \frac{8M_f^2 + 2\bar{\varepsilon}L_f}{(1-\chi)^2\bar{\varepsilon}} + \frac{1}{\lambda_0} \quad (51)$$

and the definition of  $Q(\bar{\varepsilon})$  in (8), we conclude that  $1/\underline{\lambda} \leq Q(\bar{\varepsilon})/\bar{\varepsilon}$ . This observation and (29) with  $\underline{\lambda}$  given by (49) thus imply that the first term in (7) is an upper bound on the number of the remaining iterations (where  $\lambda$  is not halved). This completes the proof. ■

## 4.2 Proof of Theorem 2.3

We begin with a technical result, i.e., Lemma 4.2, uses arguments similar to ones used in Subsection 4.2 of [20]. For the sake of completeness, its proof is given in Appendix A.2. These two results together will be used in the proof of Lemma 4.4 to show that if stepsize  $\lambda$  in a cycle is sufficiently small, then the cycle must be a serious one.

**Lemma 4.2** *If  $j$  is an iteration of a cycle such that  $t_j > (1-\chi)\bar{\varepsilon}/2$ , then*

$$j - i < (1 + u_\lambda) \log \left( \frac{4t_i}{(1-\chi)\bar{\varepsilon}} \right), \quad (52)$$

where  $\lambda$  is the prox stepsize for the cycle,  $i$  is the first iteration of the cycle,  $t_i$  is as in (17), and

$$u_\lambda := \frac{4\lambda[2M_f^2 + (1-\chi)\bar{\varepsilon}L_f]}{(1-\chi)\bar{\varepsilon}}. \quad (53)$$

Noting that the right-hand side of (52) depends on  $t_i$ , the next result provides two upper bounds on this quantity, one that does not require (A4) but only holds for  $\lambda$  sufficiently small, and one that requires (A4) and holds for an arbitrary  $\lambda > 0$ . The proof of the first one is well-known for  $\chi = 0$  (e.g., Lemma 5.5 in [15]) but not for the second one. The proofs for both of them are given in Appendix A.2.

**Lemma 4.3** *Let  $i$  denote the first iteration of a cycle of U-PB. The following statements about the quantity  $t_i$  hold:*

a) *if  $\lambda \leq (1-\chi)/(2L_f)$ , then*

$$t_i \leq \frac{4\lambda M_f^2}{1-\chi}; \quad (54)$$

b) if (A4) holds, then

$$t_i \leq \frac{\lambda(16M_f^2 + L_f^2 D^2)}{4(1 - \chi)}. \quad (55)$$

Recall that U-PB once in a while performs reset cycles, i.e., cycles which perform exactly  $\bar{N}$  inner iterations and terminate without any of their iterates satisfying the condition  $t_j \leq (1 - \chi)\bar{\varepsilon}/2$ . If the parameter pair  $(M_f, L_f)$  is known, then the result below shows that, if  $\lambda$  is chosen sufficiently small, then no reset cycle is performed. On the other hand, if  $(M_f, L_f)$  is not known, then U-PB performs a line search on  $\lambda$  and may generate reset cycles. Reset cycles are thrown away by U-PB, but they signal that the current  $\lambda$  is too large and hence should be decreased (see step 3.b of U-PB).

As already observed in the first remark in the second paragraph following U-PB, U-CS is a special case of U-PB in which  $\bar{N} = 1$ . In this regard, the following technical result is a generalization of Lemma 2.1 to the context of U-PB.

**Lemma 4.4** *Assume that assumption (A4) holds. If for some cycle of U-PB, the prox stepsize  $\lambda$  satisfies*

$$\lambda \leq \frac{(1 - \chi)^2 \bar{N} \bar{\varepsilon}}{2(M_f^2 + \bar{\varepsilon} L_f) B} \quad (56)$$

where  $B$  is as in (19), then the cycle must be a serious one. As a consequence, if the initial prox stepsize  $\lambda_0$  is chosen less than or equal to the right-hand side of (56), then U-PB has the following properties: a) it keeps  $\lambda$  constant throughout its execution; and b) it never performs a reset update.

**Proof:** Assume for contradiction that the cycle is not a serious one and its stepsize  $\lambda$  satisfies (56). Then, this cycle is a reset cycle and hence performs exactly  $\bar{N}$  iterations. Let  $i$  denote the first iteration of the cycle. In this proof, to alleviate notation, we denote  $\varepsilon = (1 - \chi)\bar{\varepsilon}/2$ . Then by inequality (52) in Lemma 4.2 written for  $j = i + \bar{N} - 1$ , we must have  $\bar{N} - 1 < (1 + u_\lambda) \log(2t_i/\varepsilon)$  where  $u_\lambda$  is as in (53). Hence, to obtain a contradiction and prove the lemma, it suffices to show that

$$(1 + u_\lambda) \log \frac{2t_i}{\varepsilon} \leq \bar{N} - 1. \quad (57)$$

To show the above inequality, we consider the following two cases: a)  $8\bar{N} \leq B$  and b)  $8\bar{N} > B$ .

a) Assume that  $8\bar{N} \leq B$ . The assumption (56) and the fact that  $\varepsilon = (1 - \chi)\bar{\varepsilon}/2$  then imply that

$$\lambda \leq \frac{(1 - \chi)\varepsilon}{8(M_f^2 + \varepsilon L_f)} \leq \frac{1 - \chi}{8L_f}$$

and hence that the assumption of Lemma 4.3(a) holds. Thus, statement (a) of this same lemma implies that

$$\frac{2t_i}{\varepsilon} \stackrel{(54)}{\leq} \frac{8\lambda M_f^2}{(1 - \chi)\varepsilon} \stackrel{(56)}{\leq} \frac{8M_f^2 \bar{N}}{(M_f^2 + \varepsilon L_f) B} \leq \frac{M_f^2}{M_f^2 + \varepsilon L_f} \leq 1 \quad (58)$$

where the second last inequality follows from the case assumption  $8\bar{N} \leq B$ . Since the above inequality implies that  $\log(2t_i/\varepsilon) \leq 0$ , inequality (57) immediately holds for case a).

b) Assume now that  $8\bar{N} > B$ . To show (57), we first bound  $1 + u_\lambda$  from above as follows:

$$\begin{aligned} 1 + u_\lambda &\stackrel{(53)}{=} 1 + \frac{4\lambda(M_f^2 + \varepsilon L_f)}{\varepsilon} \stackrel{(56)}{\leq} 1 + \frac{4(1 - \chi)\bar{N}}{B} \\ &\leq \frac{8\bar{N}}{B} + \frac{4\bar{N}}{B} = \frac{12\bar{N}}{B} \end{aligned} \quad (59)$$

where we also use the fact that  $\varepsilon = (1 - \chi)\bar{\varepsilon}/2$  in the first line and the second last inequality is due to the case assumption  $8\bar{N} > B$ . Now observe that condition (56) on  $\lambda$  implies

$$\frac{2t_i}{\varepsilon} \stackrel{(55)}{\leq} \frac{8M_f^2 + L_f^2 D^2/2}{(1 - \chi)\varepsilon} \lambda \stackrel{(56)}{\leq} \frac{8M_f^2 + L_f^2 D^2/2}{(1 - \chi)\varepsilon} \left[ \frac{(1 - \chi)\bar{N}\varepsilon}{(M_f^2 + \varepsilon L_f) B} \right] \leq \frac{(M_f^2 + L_f^2 D^2/16)\bar{N}}{M_f^2 + \varepsilon L_f} \leq 1 + \frac{L_f^2 D^2 \bar{N}}{16M_f^2}$$

where the second last inequality follows from the fact that  $B \geq 8$ . The above inequality and the definition of  $B$  in (19) then imply that

$$\log \frac{2t_i}{\varepsilon} \leq \log \left( 1 + \frac{L_f^2 D^2 \bar{N}}{16M_f^2} \right) \stackrel{(19)}{=} \frac{B-8}{12} \leq \frac{B-B/\bar{N}}{12} = \frac{B(\bar{N}-1)}{12\bar{N}} \quad (60)$$

where the last inequality follows from the case assumption  $8\bar{N} > B$ . Finally, combining (59) and (60), we conclude that (57) also holds in case b).  $\blacksquare$

The next result, which is the analogue of Proposition 4.1, shows that the sequence of stepsizes  $\lambda_k$  is bounded below by some positive constant and provides an upper bound on the number of reset cycles.

**Proposition 4.5** *The following statements hold for U-PB:*

- a) every cycle in U-PB has at most  $\bar{N}$  inner iterations;
- b) each stepsize  $\lambda_k$  generated by U-PB satisfies

$$\lambda_k \geq \frac{\bar{N}\bar{\varepsilon}}{U(\bar{\varepsilon})} \quad (61)$$

where  $U(\bar{\varepsilon})$  is as in (20);

- c) the number of reset cycles is upper bounded by

$$\left\lceil 2 \log \frac{\lambda_0 U(\bar{\varepsilon})}{\bar{N}\bar{\varepsilon}} \right\rceil. \quad (62)$$

**Proof:** a) This statement immediately follows from the description of U-PB (see its step 3).

- b) It follows from Lemma 4.4 and the update rule for  $\lambda$  in U-PB (see its step 3) that

$$\lambda_k \geq \min \left\{ \frac{(1-\chi)^2 \bar{N}\bar{\varepsilon}}{4(M_f^2 + \bar{\varepsilon}L_f)B}, \lambda_0 \right\},$$

and hence that

$$\frac{1}{\lambda_k} \leq \max \left\{ \frac{4(M_f^2 + \bar{\varepsilon}L_f)B}{(1-\chi)^2 \bar{N}\bar{\varepsilon}}, \frac{1}{\lambda_0} \right\} \leq \frac{4(M_f^2 + \bar{\varepsilon}L_f)B}{(1-\chi)^2 \bar{N}\bar{\varepsilon}} + \frac{1}{\lambda_0}.$$

Thus, using the definition of  $U(\bar{\varepsilon})$  in (20), we conclude that

$$\frac{\bar{N}\bar{\varepsilon}}{\lambda_k} \leq \frac{4(M_f^2 + \bar{\varepsilon}L_f)B}{(1-\chi)^2} + \frac{\bar{N}\bar{\varepsilon}}{\lambda_0} = U(\bar{\varepsilon}),$$

and hence that statement b) holds.

- c) This statement immediately follows from b) and the update rule of  $\lambda$  in U-PB.  $\blacksquare$

The following result shows that serious iterations of U-PB generate sequences  $\{\hat{x}_k\}$ ,  $\{\hat{y}_k\}$ ,  $\{\lambda_k\}$ , and  $\{\hat{\Gamma}_k\}$  satisfying the requirements of FSCO.

**Proposition 4.6** *U-PB is a special case of FSCO in the sense that:*

- a) the quadruple  $(\hat{x}_k, \hat{y}_k, \hat{\Gamma}_k, \lambda_k)$  satisfies relations (25) and (26) with  $x^- = \hat{x}_{k-1}$ ;
- b) condition (F1) used in the analysis of FSCO is satisfied.

**Proof:** a) For a serious cycle, the pair  $(\hat{x}_k, \hat{y}_k)$  of U-PB satisfies

$$\hat{x}_k = \operatorname{argmin}_{u \in \mathbb{R}^n} \left\{ \hat{\Gamma}_k(u) + \frac{1}{2\lambda_k} \|u - \hat{x}_{k-1}\|^2 \right\}, \quad (63)$$

$$\phi(\hat{y}_k) + \frac{\chi}{2\lambda_k} \|\hat{y}_k - \hat{x}_{k-1}\|^2 - \left[ \hat{\Gamma}_k(\hat{x}_k) + \frac{1}{2\lambda_k} \|\hat{x}_k - \hat{x}_{k-1}\|^2 \right] \leq \frac{(1-\chi)\bar{\varepsilon}}{2}. \quad (64)$$

Indeed, relations (63) and (64) (which are (26) and (25) in FSCO, respectively) follow from (15) and  $t_j \leq (1 - \chi)\bar{\varepsilon}/2$  with  $\hat{x}_k = x_j$ ,  $\hat{\Gamma}_k = f_j + h$ ,  $\hat{y}_k = y_j$ , and  $\lambda_k = \lambda$  (see the serious update in step 3.c of U-PB).

b) Condition (F1) is satisfied with  $\underline{\lambda} = \bar{N}\bar{\varepsilon}/U(\bar{\varepsilon})$ , in view of Proposition 4.5(b). ■

**Proof of Theorem 2.3** To alleviate notation, denote  $\varepsilon = (1 - \chi)\bar{\varepsilon}/2$  and  $(\mu, \nu) = (\mu_\phi, \mu_h)$ . In view of step 3.c of U-PB, we have  $\mu(\hat{\Gamma}_k) = \mu_h = \nu$  and hence it is the same as the  $\nu$  in (30). Noting that, by Proposition 4.6, the sequences indexed by  $k$  is a special implementation of FSCO, and that, by the proof of Proposition 4.5(b), (F1) holds with  $\underline{\lambda} = \bar{N}\bar{\varepsilon}/U(\bar{\varepsilon})$ , it follows from Theorem 3.2 that the total number of serious cycles generated by U-PB is bounded by

$$\min \left\{ \min \left[ \frac{1}{\chi} \left( 1 + \frac{U(\bar{\varepsilon})}{\mu \bar{\varepsilon} \bar{N}} \right), 1 + \frac{U(\bar{\varepsilon})}{\nu \bar{\varepsilon} \bar{N}} \right] \log \left( 1 + \frac{\mu d_0^2}{\bar{\varepsilon}} \right), \frac{d_0^2 U(\bar{\varepsilon})}{\bar{\varepsilon}^2 \bar{N}} \right\}.$$

The conclusion of the theorem follows from the above bound and statements (a) and (c) of Proposition 4.5. ■

## 5 Concluding remarks

In this paper, we presented two  $\mu_\phi$ -universal methods, namely U-CS and U-PB, to solve HCO (1). We propose FSCO to analyze both methods in a unified manner and establish functional complexity bounds. We then prove that both U-CS and U-PB are instances of FSCO and apply the complexity bounds for FSCO to obtain iteration complexities for the two methods. The two proposed methods are completely universal with respect to all problem parameters, in particular, the strong convexity. Moreover, they do not require knowledge of the optimal value, and do not rely on any restart or parallelization schemes (such as the ones used in [29]).

Some papers about universal methods (see for example [11, 26]) assume that, for some  $\alpha \in [0, 1]$ ,  $f$  in (1) has  $\alpha$ -Hölder continuous gradient, i.e., there exists  $L_\alpha \geq 0$  such that  $\|\nabla f(x) - \nabla f(y)\| \leq L_\alpha \|x - y\|^\alpha$  for every  $x, y \in \text{dom } h$ . It is shown in [26] that the universal primal gradient method proposed on it (i.e., the U-CS method with  $\chi = 0$ ) finds a  $\bar{\varepsilon}$ -solution of (1) in

$$\tilde{\mathcal{O}} \left( \frac{d_0^2 L_{\alpha+1}^{\frac{2}{\alpha+1}}}{\bar{\varepsilon}^{\frac{2}{\alpha+1}}} \right) \quad (65)$$

iterations. This result also follows as a consequence of our results in this paper. Indeed, first note that the dominant term in the iteration complexity (7) for the U-CS method is  $\tilde{\mathcal{O}}(d_0^2(M_f^2 + \bar{\varepsilon}L_f)/\bar{\varepsilon}^2)$ . Second, Proposition 2.1 of [20] implies that there exists a pair  $(M_f, L_f)$  satisfying (A2) and the inequality

$$M_f^2 + \bar{\varepsilon}L_f \leq 2\bar{\varepsilon}^{\frac{2\alpha}{\alpha+1}} L_{\alpha+1}^{\frac{2}{\alpha+1}}.$$

Hence, it follows from these two observations that (7) is sharper than the bound (65) obtained in [26].

We finally discuss some possible extensions of our analysis in this paper. First, it is shown in Theorem 2.2 (resp., Theorem 2.3) that U-CS (resp., U-PB) is  $\mu_\phi$ -universal if  $\chi > 0$  and is  $\mu_h$ -universal if  $\chi = 0$ . It would be interesting to investigate whether they are also  $\mu_\phi$ -universal for  $\chi = 0$ . Note that this question is related to whether the universal primal gradient of [26] (which is the same as U-CS with  $\chi = 0$ ) is  $\mu_\phi$ -universal. Second, it would also be interesting to study whether the general results obtained for the FSCO framework can also be used to show that other methods for solving the HCO problem (1) are  $\mu_\phi$ -universal. Third, a more challenging question is whether a  $\mu_\phi$ -universal method for solving (1) with a checkable termination condition (and hence which does not depend on  $\phi_*$ ) can be developed. Finally, this work has only dealt with unaccelerated methods for solving (1). It would be interesting to develop accelerated methods such as those for when  $\mu = 0$  (e.g., in [11, 26]), which are  $\mu_\phi$ -universal in the hybrid setting of this paper.

**Data Availability statement.** There is no data related to this publication.

## References

- [1] T. Alamo, P. Krupa, and D. Limon. Gradient based restart FISTA. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 3936–3941. IEEE, 2019.

- [2] T. Alamo, P. Krupa, and D. Limon. Restart of accelerated first-order methods with linear convergence under a quadratic functional growth condition. *IEEE Transactions on Automatic Control*, 68(1):612–619, 2022.
- [3] T. Alamo, D. Limon, and P. Krupa. Restart FISTA with global linear convergence. In *2019 18th European Control Conference (ECC)*, pages 1969–1974. IEEE, 2019.
- [4] J-F Aujol, L. Calatroni, C. Dossal, H. Labarrière, and A. Rondepierre. Parameter-free FISTA by adaptive restart and backtracking. *arXiv preprint arXiv:2307.14323*, 2023.
- [5] J-F Aujol, C. Dossal, and A. Rondepierre. FISTA is an automatic geometrically optimized algorithm for strongly convex functions. *Mathematical Programming*, pages 1–43, 2023.
- [6] J-F Aujol, C. H. Dossal, H. Labarrière, and A. Rondepierre. FISTA restart using an automatic estimation of the growth parameter. *Preprint*, 2022.
- [7] Y. Du and A. Ruszczyński. Rate of convergence of the bundle method. *Journal of Optimization Theory and Applications*, 173(3):908–922, 2017.
- [8] O. Fercoq and Z. Qu. Adaptive restart of accelerated gradient methods under local quadratic growth condition. *IMA Journal of Numerical Analysis*, 39(4):2069–2095, 2019.
- [9] B. Grimmer. On optimal universal first-order methods for minimizing heterogeneous sums. *Optimization Letters*, pages 1–19, 2023.
- [10] A. Iouditski and Y. Nesterov. Primal-dual subgradient methods for minimizing uniformly convex functions. *arXiv*, 2014.
- [11] G. Lan. Bundle-level type methods uniformly optimal for smooth and nonsmooth convex optimization. *Mathematical Programming*, 149(1-2):1–45, 2015.
- [12] G. Lan and R. D. C. Monteiro. Iteration-complexity of first-order penalty methods for convex programming. *Mathematical Programming*, 138(1):115–139, 2013.
- [13] G. Lan, Y. Ouyang, and Z. Zhang. Optimal and parameter-free gradient minimization methods for convex and nonconvex optimization. *arXiv: 2310.12139*, 2023.
- [14] T. Li and G. Lan. A simple uniformly optimal method without line search for convex optimization. *arXiv:2310.10082*, 2023.
- [15] J. Liang, V. Guigues, and R. D. C. Monteiro. A single cut proximal bundle method for stochastic convex composite optimization. *Mathematical Programming*, 208(1):173–208, 2024.
- [16] J. Liang and R. D. C. Monteiro. A proximal bundle variant with optimal iteration-complexity for a large range of prox stepsizes. *Extended version available on arXiv:2003.11457*, 2020.
- [17] J. Liang and R. D. C. Monteiro. An average curvature accelerated composite gradient method for non-convex smooth composite optimization problems. *SIAM Journal on Optimization*, 31(1):217–243, 2021.
- [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. Average curvature fista for nonconvex smooth composite optimization problems. *Computational Optimization and Applications*, 86(1):275–302, 2023.
- [20] 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.
- [21] Q. Lin and L. Xiao. An adaptive accelerated proximal gradient method and its homotopy continuation for sparse optimization. *Proceedings of the 31st International Conference on Machine Learning, Beijing, China*, 32:73–81, 2014.

- [22] J. G. Melo, R. D. C. Monteiro, and H. Wang. A proximal augmented lagrangian method for linearly constrained nonconvex composite optimization problems. *Journal of Optimization Theory and Applications*, pages 1–33, 2023.
- [23] K. Mishchenko and Y. Malitsky. Adaptive gradient descent without descent. In *37th International Conference on Machine Learning (ICML 2020)*, 2020.
- [24] A. Nemirovskii and Y. Nesterov. Optimal methods of smooth convex minimization. *Comput. Math. Math. Phys.*, 2:21–30, 1985.
- [25] Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.
- [26] Y. Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152(1):381–404, 2015.
- [27] Y. Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.
- [28] B. O’donoghue and E. Candes. Adaptive restart for accelerated gradient schemes. *Foundations of Computational Mathematics*, 15:715–732, 2015.
- [29] J. Renegar and B. Grimmer. A simple nearly optimal restart scheme for speeding up first-order methods. *Foundations of Computational Mathematics*, 22(1):211–256, 2022.
- [30] V. Roulet and A. d’Aspremont. Sharpness, restart and acceleration. *SIAM Journal on Optimization*, 30:262–289, 2020.
- [31] D. Zhou, S. Ma, and J. Yang. AdaBB: Adaptive Barzilai-Borwein method for convex optimization. *Mathematics of Operations Research*, 2025.

## A Technical results

This appendix contains two subsections. Subsection A.1 presents some technical results used in the analysis of FSCO while Subsection A.2 proves Lemmas 4.2 and 4.3.

### A.1 Miscellaneous technical results

The following technical result is used in the proof of Lemma 3.4.

**Lemma A.1** Assume that  $\xi > 0, \psi \in \overline{\text{Conv}}(\mathbb{R}^n)$  and  $Q \in \mathcal{S}_{++}^n$  are such that  $\psi - (\xi/2)\|\cdot\|_Q^2$  is convex and let  $(y, v, \eta) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}_+$  be such that  $v \in \partial_\eta \psi(y)$ . Then, for any  $\tau > 0$ ,

$$\psi(u) \geq \psi(y) + \langle v, u - y \rangle - (1 + \tau^{-1})\eta + \frac{(1 + \tau)^{-1}\xi}{2}\|u - y\|_Q^2 \quad \forall u \in \mathbb{R}^n. \quad (66)$$

**Proof:** The proof of this result can be found in Lemma A.1 in [22]. ■

The next technical result is used in the proof of Theorem 3.2.

**Lemma A.2** Assume that sequences  $\{\gamma_j\}$ ,  $\{\eta_j\}$ , and  $\{\alpha_j\}$  satisfy for every  $j \geq 1$ ,  $\gamma_j \geq \underline{\gamma}$  and

$$\gamma_j \eta_j \leq \alpha_{j-1} - (1 + \sigma)\alpha_j + \gamma_j \delta \quad (67)$$

for some  $\sigma \geq 0$ ,  $\delta \geq 0$  and  $\underline{\gamma} > 0$ . Then, the following statements hold:

a) for every  $k \geq 1$ ,

$$\min_{1 \leq j \leq k} \eta_j \leq \frac{\alpha_0 - (1 + \sigma)^k \alpha_k}{\sum_{j=1}^k (1 + \sigma)^{j-1} \gamma_j} + \delta; \quad (68)$$

b) if the sequence  $\{\alpha_j\}$  is nonnegative, then  $\min_{1 \leq j \leq k} \eta_j \leq 2\delta$  for every  $k \geq 1$  such that

$$k \geq \min \left\{ \frac{1+\sigma}{\sigma} \log \left( \frac{\sigma \alpha_0}{\underline{\gamma} \delta} + 1 \right), \frac{\alpha_0}{\underline{\gamma} \delta} \right\}$$

with the convention that the first term is equal to the second term when  $\sigma = 0$ . (Note that the first term converges to the second term as  $\sigma \downarrow 0$ .)

**Proof:** a) Multiplying (67) by  $(1+\sigma)^{j-1}$  and summing the resulting inequality from  $j = 1$  to  $k$ , we have

$$\begin{aligned} \sum_{j=1}^k (1+\sigma)^{j-1} \gamma_j \left[ \min_{1 \leq j \leq k} \eta_j \right] &\leq \sum_{j=1}^k (1+\sigma)^{j-1} \gamma_j \eta_j \leq \sum_{j=1}^k (1+\sigma)^{j-1} (\alpha_{j-1} - (1+\sigma)\alpha_j + \gamma_j \delta) \\ &= \alpha_0 - (1+\sigma)^k \alpha_k + \sum_{j=1}^k (1+\sigma)^{j-1} \gamma_j \delta. \end{aligned} \quad (69)$$

Inequality (68) follows immediately from the above inequality.

b) It follows from (68), and the facts that  $\alpha_k \geq 0$  and  $\gamma_j \geq \underline{\gamma}$  that

$$\min_{1 \leq j \leq k} \eta_j \leq \frac{\alpha_0}{\underline{\gamma} \sum_{j=1}^k (1+\sigma)^{j-1}} + \delta. \quad (70)$$

Using the fact that  $1+\sigma \geq e^{\sigma/(1+\sigma)}$  for every  $\sigma \geq 0$ , we have

$$\sum_{j=1}^k (1+\sigma)^{j-1} = \max \left\{ \frac{(1+\sigma)^k - 1}{\sigma}, k \right\} \geq \max \left\{ \frac{e^{\sigma k/(1+\sigma)} - 1}{\sigma}, k \right\}. \quad (71)$$

Plugging the above inequality into (70), we have for every  $k \geq 1$ ,

$$\min_{1 \leq j \leq k} \eta_j \leq \frac{\alpha_0}{\underline{\gamma}} \min \left\{ \frac{\sigma}{e^{\sigma k/(1+\sigma)} - 1}, \frac{1}{k} \right\} + \delta,$$

which can be easily seen to imply (b). ■

The following technical result is used in the analysis of Section 3.

**Lemma A.3** For every  $\chi \in [0, 1)$ ,  $\lambda > 0$ ,  $\mu > 0$ , and  $\nu \in [0, \mu]$ , the quantity  $\sigma = \sigma(\mu, \nu; \lambda)$  defined in (28) satisfies the following statements:

- a)  $\lambda[\chi\mu + (1-\chi)\nu] \leq \sigma \leq \lambda\mu$ , and hence  $\sigma \geq \lambda\nu$ ;
- b) the function  $\lambda > 0 \mapsto \sigma(\mu, \nu; \lambda)$  is non-decreasing;
- c) there holds

$$\frac{1+\sigma}{\sigma} \leq \min \left\{ \frac{1}{\chi} \left( 1 + \frac{1}{\lambda\mu} \right), 1 + \frac{1}{\lambda\nu} \right\}.$$

**Proof:** a) It follows from (28) and the fact that  $\nu \leq \mu$  that

$$\sigma - \lambda\mu = \frac{\lambda(\mu - \nu)(1 + \lambda\mu)(\chi - 1)}{1 + \mu\lambda + \chi\lambda(\nu - \mu)} \geq \frac{\lambda(\mu - \nu)(1 + \lambda\mu)(\chi - 1)}{1 + \mu\lambda} = \lambda(\mu - \nu)(\chi - 1),$$

and hence that  $\sigma \geq \lambda[\chi\mu + (1-\chi)\nu]$ . Moreover, the second inequality of (a) immediately follows from the first identity above and the facts that  $\chi < 1$  and  $\nu \leq \mu$ .

b) Viewing  $\sigma = \sigma(\lambda)$  in (28) as a function of  $\lambda$ , one can verify that

$$\sigma(\lambda) = \frac{\nu(1 + \mu\lambda) + \chi(\mu - \nu)}{\lambda^{-1} + \mu + \chi(\nu - \mu)}$$

is a non-decreasing function of  $\lambda$ .

c) It follows from the definition of  $\sigma$  in (28) that

$$\frac{1 + \sigma}{\sigma} = \frac{(1 + \nu\lambda)(1 + \mu\lambda)}{\lambda[\nu(1 + \mu\lambda) + \chi(\mu - \nu)]}.$$

Hence, it suffices to prove that

$$\frac{(1 + \nu\lambda)(1 + \mu\lambda)}{\lambda[\nu(1 + \mu\lambda) + \chi(\mu - \nu)]} \leq \min \left\{ \frac{1}{\chi} \left( 1 + \frac{1}{\lambda\mu} \right), 1 + \frac{1}{\lambda\nu} \right\}. \quad (72)$$

Since  $\chi \in [0, 1)$ , it is easy to verify that

$$\frac{1 + \nu\lambda}{\nu(1 + \mu\lambda) + \chi(\mu - \nu)} \leq \frac{1}{\chi\mu},$$

and hence that

$$\frac{(1 + \nu\lambda)(1 + \mu\lambda)}{\lambda[\nu(1 + \mu\lambda) + \chi(\mu - \nu)]} \leq \frac{1 + \mu\lambda}{\chi\lambda\mu} = \frac{1}{\chi} \left( 1 + \frac{1}{\lambda\mu} \right).$$

Moreover, noting that  $\chi \in [0, 1)$  and  $\mu \geq \nu$ , we also have

$$\frac{(1 + \nu\lambda)(1 + \mu\lambda)}{\lambda[\nu(1 + \mu\lambda) + \chi(\mu - \nu)]} \leq 1 + \frac{1}{\lambda\nu}.$$

Combining the above two inequalities, we conclude that (72) holds.  $\blacksquare$

## A.2 Proofs of Lemmas 4.2 and 4.3

Before presenting the proof of Lemma 4.2, we present and prove three technical results.

The first one summarizes some basic properties of U-PB.

**Lemma A.4** *Assume that  $\lambda$  is the prox stepsize and  $\hat{x}_{k-1}$  is the prox-center of a cycle of U-PB. Then, the following statements about this cycle hold:*

a) *for every null iteration  $j$  of the cycle, there exists a function  $\bar{f}_j(\cdot)$  such that*

$$\tau(\bar{f}_j + h) + (1 - \tau)[\ell_f(\cdot; x_j) + h] \leq f_{j+1} + h \leq \phi, \quad (73)$$

$$\bar{f}_j + h \in \overline{\text{Conv}}_\nu(\mathbb{R}^n), \quad \bar{f}_j(x_j) = f_j(x_j), \quad (74)$$

$$x_j = \underset{u \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ \bar{f}_j(u) + h(u) + \frac{1}{2\lambda} \|u - \hat{x}_{k-1}\|^2 \right\}, \quad (75)$$

where

$$\tau := \frac{u_\lambda}{1 + u_\lambda} \quad (76)$$

and  $u_\lambda$  is as in (53);

b) *for every iteration  $j$  of the cycle and  $u \in \mathbb{R}^n$ , we have*

$$\bar{f}_j(u) + h(u) + \frac{1}{2\lambda} \|u - \hat{x}_{k-1}\|^2 \geq m_j + \frac{1}{2\lambda} \|u - x_j\|^2, \quad (77)$$

where

$$m_j := (f_j + h)(x_j) + \frac{1}{2\lambda} \|x_j - \hat{x}_{k-1}\|^2. \quad (78)$$

**Proof:** a) Since  $j$  is a null iteration of the  $k$ -th serious cycle, it follows from step 3.a of U-PB that  $f_{j+1} = \text{BU}(\hat{x}_{k-1}, x_j, f_j, \lambda)$ . Using the properties of the BU blackbox, this implies the existence of  $\bar{f}_j$  satisfying (74), (75) and

$$\max\{\bar{f}_j + h, \ell_f(\cdot; x_j) + h\} \leq f_{j+1} + h \leq \phi. \quad (79)$$

The statement now follows from the fact that (79) immediately implies (73).

b) Since the objective function in the last identity of (75) is  $\lambda^{-1}$ -strongly convex, we have

$$\bar{f}_j(u) + h(u) + \frac{1}{2\lambda}\|u - \hat{x}_{k-1}\|^2 \geq \bar{f}_j(x_j) + h(x_j) + \frac{1}{2\lambda}\|x_j - \hat{x}_{k-1}\|^2 + \frac{1}{2\lambda}\|u - x_j\|^2, \quad (80)$$

which immediately implies (77) in view of the identity in (74) and the definition of  $m_j$  in (78).  $\blacksquare$

The second technical result provides a recursive bound on the quantity  $m_j$  defined in (78).

**Lemma A.5** *If  $j$  is a null iteration of a cycle, then*

$$m_{j+1} - \tau m_j \geq (1 - \tau) \left[ \phi(x_{j+1}) + \frac{1}{2\lambda}\|x_{j+1} - \hat{x}_{k-1}\|^2 \right] - \frac{(1 - \tau)\varepsilon}{2} \quad (81)$$

where  $\tau$  is defined in (76) and  $\varepsilon = (1 - \chi)\bar{\varepsilon}/2$ .

**Proof:** Using the definitions of  $\tau$  and  $u_\lambda$  in (76) and (53), respectively, we have

$$\frac{\tau}{2\lambda(1 - \tau)} \stackrel{(76)}{=} \frac{u_\lambda}{2\lambda} \stackrel{(53)}{=} 2L_f + \frac{4M_f^2}{(1 - \chi)\bar{\varepsilon}} = 2L_f + \frac{2M_f^2}{\varepsilon} \geq \frac{L_f}{2} + \frac{2M_f^2}{\varepsilon}, \quad (82)$$

where the last identity is due to the relation  $\varepsilon = (1 - \chi)\bar{\varepsilon}/2$ . Using the definition of  $m_j$  in (78), and relations (73) and (77) with  $u = x_{j+1}$ , we have

$$\begin{aligned} m_{j+1} &\stackrel{(78)}{=} (f_{j+1} + h)(x_{j+1}) + \frac{1}{2\lambda}\|x_{j+1} - \hat{x}_{k-1}\|^2 \\ &\stackrel{(73)}{\geq} (1 - \tau) \left[ \ell_f(x_{j+1}; x_j) + h(x_{j+1}) + \frac{1}{2\lambda}\|x_{j+1} - \hat{x}_{k-1}\|^2 \right] + \tau \left( (\bar{f}_j + h)(x_{j+1}) + \frac{1}{2\lambda}\|x_{j+1} - \hat{x}_{k-1}\|^2 \right) \\ &\stackrel{(77)}{\geq} (1 - \tau) \left[ \ell_f(x_{j+1}; x_j) + h(x_{j+1}) + \frac{1}{2\lambda}\|x_{j+1} - \hat{x}_{k-1}\|^2 \right] + \tau \left( m_j + \frac{1}{2\lambda}\|x_{j+1} - x_j\|^2 \right). \end{aligned}$$

This inequality and (82) then imply

$$m_{j+1} - \tau m_j \geq (1 - \tau) \left[ \ell_f(x_{j+1}; x_j) + h(x_{j+1}) + \frac{1}{2\lambda}\|x_{j+1} - \hat{x}_{k-1}\|^2 + \left( \frac{L_f}{2} + \frac{2M_f^2}{\varepsilon} \right) \|x_{j+1} - x_j\|^2 \right]. \quad (83)$$

Using (3) with  $(x, y) = (x_{j+1}, x_j)$  and the fact that  $\phi = f + h$ , we have

$$\ell_f(x_{j+1}; x_j) + h(x_{j+1}) + \frac{L_f}{2}\|x_{j+1} - x_j\|^2 \geq \phi(x_{j+1}) - 2M_f\|x_{j+1} - x_j\|. \quad (84)$$

Using this inequality and relation (83), we conclude that

$$\begin{aligned} m_{j+1} - \tau m_j &\stackrel{(83), (84)}{\geq} (1 - \tau) \left( \phi(x_{j+1}) + \frac{1}{2\lambda}\|x_{j+1} - \hat{x}_{k-1}\|^2 \right) + \frac{2(1 - \tau)}{\varepsilon} \left( M_f^2\|x_{j+1} - x_j\|^2 - M_f\varepsilon\|x_{j+1} - x_j\| \right) \\ &\geq (1 - \tau) \left[ \phi(x_{j+1}) + \frac{1}{2\lambda}\|x_{j+1} - \hat{x}_{k-1}\|^2 \right] - \frac{(1 - \tau)\varepsilon}{2}, \end{aligned}$$

where the last inequality follows the inequality  $a^2 - 2ab \geq -b^2$  with  $a = M_f\|x_{j+1} - x_j\|$  and  $b = \varepsilon/2$ .  $\blacksquare$

The next technical result establishes an important recursive formula for the sequence  $\{t_j\}$  defined in (17).

**Lemma A.6** *Let  $j$  be an iteration of a cycle with step size  $\lambda$  and first iteration  $i$ . Then*

$$t_j - \frac{(1 - \chi)\bar{\varepsilon}}{4} \leq \tau^{j-i} \left( t_i - \frac{(1 - \chi)\bar{\varepsilon}}{4} \right) \quad (85)$$

where  $\tau$  is defined in (76).

**Proof:** If  $j = i$ , (85) trivially holds. We next show that (85) holds for  $j \geq i + 1$ . It suffices to show that for any  $\ell \in \{i, \dots, j - 1\}$ ,

$$t_{\ell+1} - \frac{(1-\chi)\bar{\varepsilon}}{4} \leq \tau \left( t_{\ell} - \frac{(1-\chi)\bar{\varepsilon}}{4} \right). \quad (86)$$

Let  $\ell \in \{i, \dots, j - 1\}$  be given and note that  $\ell$  is a null iteration of the cycle. Using the definition of  $t_{\ell}$  in (17) and Lemma A.5 with  $j = \ell$ , we have

$$\begin{aligned} t_{\ell+1} - \tau t_{\ell} &\stackrel{(17)}{=} \bar{\phi}_{\ell+1} - m_{\ell+1} - \tau[\bar{\phi}_{\ell} - m_{\ell}] = [\bar{\phi}_{\ell+1} - \tau\bar{\phi}_{\ell}] - [m_{\ell+1} - \tau m_{\ell}] \\ &\stackrel{(81)}{\leq} \bar{\phi}_{\ell+1} - \tau\bar{\phi}_{\ell} - (1-\tau) \left[ \phi(x_{\ell+1}) + \frac{1}{2\lambda} \|x_{\ell+1} - \hat{x}_{k-1}\|^2 \right] + \frac{(1-\tau)(1-\chi)\bar{\varepsilon}}{4} \leq \frac{(1-\tau)(1-\chi)\bar{\varepsilon}}{4}, \end{aligned}$$

where the last inequality follows from (16) and the fact that  $\tau \in (0, 1)$ . We have thus shown that (86) holds for any  $\ell \in \{i, \dots, j - 1\}$ .  $\blacksquare$

We are now ready to give the proof of the first main result of this subsection.

**Proof of Lemma 4.2** By (85), we obtain

$$\frac{4t_j}{(1-\chi)\bar{\varepsilon}} - 1 \leq \frac{4}{(1-\chi)\bar{\varepsilon}} \tau^{j-i} t_i. \quad (87)$$

Using relation (76), we then have

$$0 < \log \left( \frac{4t_j}{(1-\chi)\bar{\varepsilon}} - 1 \right) \stackrel{(87)}{\leq} \log \left( \frac{4\tau^{j-i} t_i}{(1-\chi)\bar{\varepsilon}} \right) \stackrel{(76)}{=} \log \left( \frac{4t_i}{(1-\chi)\bar{\varepsilon}} \right) + \log \left( \left[ \frac{u_{\lambda}}{1+u_{\lambda}} \right]^{j-i} \right) \leq \log \left( \frac{4t_i}{(1-\chi)\bar{\varepsilon}} \right) - \frac{j-i}{1+u_{\lambda}}$$

where for the last inequality we have used the  $\log(x) \leq x - 1$  which holds for  $x > 0$ . This completes the proof of (52).  $\blacksquare$

We next prove the second main result of this subsection.

**Proof of Lemma 4.3.** We first show that the inequality

$$t_i \leq 2M_f \|x_i - \hat{x}_{k-1}\| + \frac{L_f}{2} \|x_i - \hat{x}_{k-1}\|^2 - \frac{1-\chi}{2\lambda} \|x_i - \hat{x}_{k-1}\|^2 \quad (88)$$

holds without making the assumptions of either case a) or case b). Indeed, using the definition of  $t_j$  in (17) for  $j = i$  and relation (16), we have

$$\begin{aligned} t_i &\stackrel{(78)}{=} \bar{\phi}_i - (f_i + h)(x_i) - \frac{1}{2\lambda} \|x_i - \hat{x}_{k-1}\|^2 \\ &\stackrel{(16)}{\leq} \phi(x_i) + \frac{\chi}{2\lambda} \|x_i - \hat{x}_{k-1}\|^2 - (f_i + h)(x_i) - \frac{1}{2\lambda} \|x_i - \hat{x}_{k-1}\|^2 \\ &= f(x_i) - f_i(x_i) + \frac{\chi-1}{2\lambda} \|x_i - \hat{x}_{k-1}\|^2, \end{aligned}$$

where the last equality follows from the fact that  $\phi = f + h$ . Since  $i$  is the first iteration of the cycle, it follows from step 3.d of U-PB that  $f_i(\cdot) \geq \ell_f(\cdot; \hat{x}_{k-1})$ . Combining this inequality with the above one, and using (3) with  $(x, y) = (x_i, \hat{x}_{k-1})$ , we conclude that

$$\begin{aligned} t_i &\leq f(x_i) - \ell_f(x_i; \hat{x}_{k-1}) + \frac{\chi-1}{2\lambda} \|x_i - \hat{x}_{k-1}\|^2 \\ &\stackrel{(3)}{\leq} 2M_f \|x_i - \hat{x}_{k-1}\| + \frac{L_f}{2} \|x_i - \hat{x}_{k-1}\|^2 + \frac{\chi-1}{2\lambda} \|x_i - \hat{x}_{k-1}\|^2 \\ &= 2M_f \|x_i - \hat{x}_{k-1}\| - \frac{1-\chi-\lambda L_f}{2\lambda} \|x_i - \hat{x}_{k-1}\|^2 \end{aligned} \quad (89)$$

and hence that (88) holds.

We next prove that a) holds. Assume that  $\lambda > 0$  is such that  $\lambda \leq (1 - \chi)/(2L_f)$ . Note that this implies that  $1 - \chi \geq 2\lambda L_f > 0$ , and hence  $1 - \chi - \lambda L_f > 0$ . Maximizing the right-hand side of (88) with respect to  $\|x_i - \hat{x}_{k-1}\|$ , we conclude that  $t_i \leq (2\lambda M_f^2)/(1 - \chi - \lambda L_f)$ , and hence that (54) due to the inequality  $(1 - \chi)/2 \geq \lambda L_f$ .

We next show that b) holds. Assume that (A4) holds. Using (88), we obtain

$$\begin{aligned} t_i &\leq 2M_f \|x_i - \hat{x}_{k-1}\| + \frac{L_f D}{2} \|x_i - \hat{x}_{k-1}\| - \frac{1 - \chi}{2\lambda} \|x_i - \hat{x}_{k-1}\|^2 \\ &= \left(2M_f + \frac{L_f D}{2}\right) \|x_i - \hat{x}_{k-1}\| - \frac{1 - \chi}{2\lambda} \|x_i - \hat{x}_{k-1}\|^2 \\ &\leq \left(2M_f + \frac{L_f D}{2}\right)^2 \frac{\lambda}{2(1 - \chi)} \leq \frac{\lambda(4M_f^2 + L_f^2 D^2/4)}{1 - \chi}, \end{aligned}$$

where the second last inequality is due to the inequality that  $-a_0 x^2 + b_0 x \leq b_0^2/(4a_0)$  for  $a_0 > 0$ , and the last inequality is due to the fact that  $(a + b)^2 \leq 2(a^2 + b^2)$ .  $\blacksquare$