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

Vincent Guigues [*] Jiaming Liang [†] Renato D.C. Monteiro [‡]

July 14, 2024 (1st revision: August 2, 2024)

**Abstract**

This paper develops two parameter-free methods for solving convex and strongly convex hybrid composite optimization problems, namely, a composite subgradient type method and a proximal bundle type method. Both functional and stationary complexity bounds for the two methods are established in terms of the unknown strong convexity parameter. To the best of our knowledge, the two proposed methods are the first universal methods for solving hybrid strongly convex composite optimization problems that do not rely on any restart scheme nor require the knowledge of the optimal value.

**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

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

where $f, h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ are proper lower semi-continuous convex functions such that $\operatorname{dom} h \subseteq \operatorname{dom} f$ and the following conditions hold: there exist scalars $M_f \geq 0$ and $L_f \geq 0$ and a first-order oracle $f' : \operatorname{dom} h \to \mathbb{R}^n$ (i.e., $f'(x) \in \partial f(x)$ for every $x \in \operatorname{dom} h$) satisfying the $(M_f, L_f)$-hybrid condition that $\|f'(x) - f'(y)\| \leq 2M_f + L_f\|x - y\|$ for every $x, y \in \operatorname{dom} h$. Moreover, assume that $\mu \geq 0$ is the largest scalar such that $\phi(\cdot) - \mu\|\cdot\|^2/2$ is convex, i.e., $\mu$ is the intrinsic convex parameter of $\phi$.

This work is concerned with parameter-free (PF) methods for solving (1), i.e., ones that do not require knowledge of any of parameters associated with the instance $(f, h)$, such as the parameter pair $(M_f, L_f)$ or the intrisic convexity parameter $\mu$ of $\phi$. More specifically, it considers PF methods whose complexities for solving (1) are expressed in terms of $\mu$ (in addition to other parameters associated with $(f, h)$). We refer to them as $\mu$-universal methods. PF methods whose (provable) complexities do not depend on $\mu$ are called universal ones (even if $\mu > 0$). Moreover, PF methods whose complexities are given in terms of the intrinsic convex parameter $\mu_f$ for $f$ (resp., $\mu_h$ for $h$) are called $\mu_f$-universal (resp., $\mu_h$-universal). It is worth noting that $\mu$ 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 $\mu = 2\alpha \gg 0 = \mu_f + \mu_h$). Hence, complexities for $\mu$-universal methods are usually better than $\mu_f$-universal and $\mu_h$-universal methods, and even (universal or non-universal) methods whose complexities depend on $\mu_f + \mu_h$.

**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 [21] and [10]. Specifically, the first paper develops universal variants of the primal gradient, the dual gradient, and the accelerated gradient methods, while the second one shows

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

[†] Goergen Institute for Data Science and Department of Computer Science, University of Rochester, Rochester, NY 14620 (email: jiaming.liang@rochester.edu).

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

that acceleration versions of the bundle-level and the prox-level methods are universal. Additional universal methods for solving (1) have been studied in [14, 16, 19, 25] under the condition that $f$ is smooth, and in [13, 15, 17] for the case where $f$ is either smooth or nonsmooth. The methods in [15, 17] (resp., [19, 25]) are also shown to be $\mu_h$-universal (resp., $\mu_f$-universal under the condition that $h = 0$). The papers [14, 16] 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 [14, 16] also apply to the convex case and yield complexity bounds similar to the ones of [19, 25].

*$\mu$-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, 12, 20, 23] have developed $\mu$-universal (or $\mu_f$-universal) methods for (1) with a $\tilde{\mathcal{O}}(\sqrt{L_f/\mu})$ (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), [20] 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 $\mu$-universal variant along this venue); moreover, using restart ACG schemes motivated by previous works such as [11, 22, 24], paper [23] develops $\mu$-universal methods under the assumption that $\phi_*$ is known.

Among papers concerned with finding an $\varepsilon$-stationary solution of (1), [1, 2, 3, 4, 6] (resp., [12]) develop $\mu$-universal (resp., $\mu_f$-universal) methods based on restart ACG (resp., S-ACG) schemes that estimate the condition number $L_f/\mu$ (resp., $\mu_f$), or some related quantity; [3, 4, 6] then use the estimation to determine the number of iterations of each ACG call while the SCAR method of [12] uses a stationary termination to end each S-ACG call. Moreover, under the assumption that $\phi_*$ and $L_f$ are known, [5] develops a $\mu$-universal method that performs only one call to an ACG variant (for convex CO).

Under the assumption that $f$ is non-smooth (i.e., $M_f > 0$), [23] (see also [9] for an extension) proposes $\mu$-universal methods under the assumption that $\phi_*$ is known. Specifically, the $\mu$-universal method of [23] repeatedly invokes a universal oracle that halves the primal gap $\phi(x) - \phi_*$ on each call.[1]

**Our contribution.** The goal of this work is to present two $\mu$-universal methods for problem (1), namely: a composite subgradient (U-CS) type method and a proximal bundle (U-PB) type method. The first method is a variant of the universal primal gradient method of [21] (see also Appendix C.2 of [17]), which is still not known to be $\mu$-universal. The second one is a variant of the generic proximal bundle (GPB) method of [17] 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 sufficient conditions for its PF instances to be $\mu$-universal. Both functional and stationary complexities are established for FSCO in terms of $\mu$, which are then used to obtain complexity bounds for both U-CS and U-PB. Interestingly, in contrast to previous $\mu$-universal methods, both U-CS and U-PB do not perform any restart scheme nor require $\phi_*$ to be known.

**Organization of the paper.** Subsection 1.1 presents basic definitions and notation used throughout the paper. Section 2 formally describes FSCO and the assumptions on the problem of interest, and provides both functional and stationarity complexity analyses of FSCO. Section 3 presents U-CS and U-PB, as two instances of FSCO, for solving problem (1) and establishes their corresponding complexity bounds. Section 4 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 inner product and norm denoted by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$, respectively. Let $\log(\cdot)$ denote the natural logarithm.

For given $\Phi : \mathbb{R}^n \to (-\infty, +\infty]$, let $\operatorname{dom} \Phi := \{x \in \mathbb{R}^n : \Phi(x) < \infty\}$ denote the effective domain of $\Phi$ and $\Phi$ is proper if $\operatorname{dom} \Phi \neq \emptyset$. A proper function $\Phi : \mathbb{R}^n \to (-\infty, +\infty]$ 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 $x, y \in \operatorname{dom} \Phi$ and $\alpha \in [0, 1]$. Let $\overline{\operatorname{Conv}}_\mu(\mathbb{R}^n)$ denote the set of all proper lower semicontinuous $\mu$-convex functions. We simply denote $\overline{\operatorname{Conv}}_\mu(\mathbb{R}^n)$ by $\overline{\operatorname{Conv}}(\mathbb{R}^n)$ when $\mu = 0$. For $\varepsilon \geq 0$, the *$\varepsilon$-subdifferential*

[1] Paper [23] removes the assumption that $\phi_*$ is known but forces its method to make multiple parallel calls to the universal oracle.

of $\Phi$ at $x \in \operatorname{dom}\Phi$ 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 $x \in \operatorname{dom}\Phi$ 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

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

## 2 A framework for strongly convex optimization

This section presents a general framework, namely FSCO, for convex optimization problems and establishes both functional and stationary complexity bounds for any of its instances. These results will then be used in Subsections 3.1 and 3.2 to analyze the complexities of two specific algorithms, namely: U-CS and U-PB. This section is divided into two subsections. Subsection 2.1 focuses on the functional complexity analysis, while Subsection 2.2 provides the stationary complexity analysis.

FSCO is presented in the context of the convex optimization problem

$$\phi_* := \min\{\phi(x) : x \in \mathbb{R}^n\} \tag{3}$$

for which the following conditions are assumed:

(A1) the set of optimal solutions $X_*$ of problem (3) is nonempty;

(A2) $\phi \in \overline{\operatorname{Conv}}_\mu(\mathbb{R}^n)$ for some $\mu \geq 0$.

Clearly, the HCO problem (1) with the assumptions described underneath it is a special case of (3) where $\phi = f + h$.

We now describe FSCO.

```
FSCO

0. Let $\chi \in [0,1)$, $\varepsilon > 0$, and $\hat{x}_0 \in \operatorname{dom}\phi$ be given, and set $k = 1$;

1. Compute $\lambda_k > 0$, $\hat{\Gamma}_k \in \overline{\operatorname{Conv}}(\mathbb{R}^n)$, $\hat{\Gamma}_k \leq \phi$, and $\hat{y}_k \in \operatorname{dom}\phi$ satisfying

   $\phi(\hat{y}_k) + \frac{\chi}{2\lambda_k}\|\hat{y}_k - \hat{x}_{k-1}\|^2 - \min_{u\in\mathbb{R}^n}\left\{\hat{\Gamma}_k(u) + \frac{1}{2\lambda_k}\|u - \hat{x}_{k-1}\|^2\right\} \leq \varepsilon,$   (4)

   and set

   $\hat{x}_k := \operatorname{argmin}_{u\in\mathbb{R}^n}\left\{\hat{\Gamma}_k(u) + \frac{1}{2\lambda_k}\|u - \hat{x}_{k-1}\|^2\right\};$   (5)

2. Check whether a termination criterion holds and if so stop; else go to step 3;

3. Set $k \leftarrow k + 1$ and go to step 1.
```

FSCO does not specify how sequences $\{\hat{x}_k\}$ and $\{\hat{y}_k\}$ are generated, how models $\{\hat{\Gamma}_k\}$ are updated, and how stepsizes $\{\lambda_k\}$ are computed. Rather, it provides sufficient conditions on these sequences to ensure that its instances are $\mu$-universal.

The complexity analysis of FSCO requires two additional assumptions, namely:

(F1) there exists $\nu \in [0, \mu]$ such that $\hat{\Gamma}_k \in \overline{\operatorname{Conv}}_\nu(\mathbb{R}^n)$;

(F2) there exists $\underline{\lambda} > 0$ such that $\lambda_k \geq \underline{\lambda}$ for every iteration $k$ of the FSCO.

### 2.1 Functional complexity analysis

This subsection studies the iteration complexity for the framework FSCO to obtain an iterate $\hat{y}_k$ such that $\phi(\hat{y}_k) - \phi_* \le \bar{\varepsilon}$. Its main result is stated in Theorem 2.3.

The following lemma will be useful in the sequel.

**Lemma 2.1** *Consider sequences $\{\hat{x}_k\}$, $\{\hat{y}_k\}$, $\{\hat{\Gamma}_k\}$, and $\{\lambda_k\}$ generated by FSCO. Define for $k \ge 1$,*

$$\tilde{s}_k := \frac{\hat{x}_{k-1} - \hat{x}_k}{\lambda_k} - \nu \hat{x}_k, \tag{6}$$

$$\tilde{\eta}_k := \tilde{\phi}(\hat{y}_k) - \tilde{\Gamma}_k(\hat{x}_k) - \langle \tilde{s}_k, \hat{y}_k - \hat{x}_k \rangle, \tag{7}$$

*where*

$$\tilde{\phi} := \phi - \frac{\nu}{2}\|\cdot\|^2, \quad \tilde{\Gamma}_k := \hat{\Gamma}_k - \frac{\nu}{2}\|\cdot\|^2. \tag{8}$$

*Then for $k \ge 1$, we have:*

*a) $\tilde{s}_k \in \partial \tilde{\Gamma}_k(\hat{x}_k)$ and for every $u \in \mathbb{R}^n$,*

$$\tilde{\Gamma}_k(u) \ge \tilde{\Gamma}_k(\hat{x}_k) + \langle \tilde{s}_k, u - \hat{x}_k \rangle; \tag{9}$$

*b) $\tilde{s}_k \in \partial_{\tilde{\eta}_k} \tilde{\phi}(\hat{y}_k)$ and*

$$0 \le 2\lambda_k \tilde{\eta}_k \le 2\lambda_k \varepsilon - (1 + \nu\lambda_k)\|\hat{y}_k - \hat{x}_k\|^2 + (1 - \chi)\|\hat{y}_k - \hat{x}_{k-1}\|^2; \tag{10}$$

*c) for any $\chi \in [0, 1)$ and every $u \in \operatorname{dom} \phi$, we have*

$$\tilde{\phi}(u) \ge \tilde{\phi}(\hat{y}_k) + \langle \tilde{s}_k, u - \hat{y}_k \rangle + \frac{\chi(\mu - \nu)}{2}\|u - \hat{y}_k\|^2 - \frac{\tilde{\eta}_k}{1 - \chi}. \tag{11}$$

**Proof**: (a) The optimality condition of (5) yields $0 \in \partial \hat{\Gamma}_k(\hat{x}_k) + (\hat{x}_k - \hat{x}_{k-1})/\lambda_k$. This inclusion and the fact that $\partial \hat{\Gamma}_k(u) = \partial \tilde{\Gamma}_k(u) + \nu u$ for every $u \in \mathbb{R}^n$ imply that

$$0 \in \partial \tilde{\Gamma}_k(\hat{x}_k) + \nu \hat{x}_k + \frac{\hat{x}_k - \hat{x}_{k-1}}{\lambda_k} \overset{(6)}{=} \partial \tilde{\Gamma}_k(\hat{x}_k) - \tilde{s}_k$$

where the identity is due to the definition of $\tilde{s}_k$ in (6). Hence, the inclusion in a) holds. Relation (9) immediately follows from the inclusion in a) and the definition of the subdifferential.

(b) It follows from the relation $\phi \ge \hat{\Gamma}_k$ (see step 1 of FSCO) and the definition of $\tilde{\phi}$ and $\tilde{\Gamma}_k$ in (8) that $\tilde{\phi} \ge \tilde{\Gamma}_k$. This inequality, (9), the definition of $\tilde{\eta}_k$ in (7) imply that for every $u \in \operatorname{dom} \phi$,

$$\tilde{\phi}(u) \ge \tilde{\Gamma}_k(u) \overset{(9)}{\ge} \tilde{\Gamma}_k(\hat{x}_k) + \langle \tilde{s}_k, u - \hat{x}_k \rangle \overset{(7)}{=} \tilde{\phi}(\hat{y}_k) + \langle \tilde{s}_k, u - \hat{y}_k \rangle - \tilde{\eta}_k,$$

which yields the inclusion in b). Taking $u = \hat{y}_k$ in the above inequality gives $\tilde{\eta}_k \ge 0$, and hence the first inequality in (10) holds. Using the definitions of $\tilde{s}_k$ and $\tilde{\eta}_k$ in (6) and (7), respectively, the definitions of $\tilde{\phi}$ and $\tilde{\Gamma}_k$ in (8), and (4) and (5), we have

$$\begin{aligned}
\tilde{\eta}_k &\overset{(7),(8)}{=} \phi(\hat{y}_k) - \hat{\Gamma}_k(\hat{x}_k) + \frac{\nu}{2}\left(\|\hat{x}_k\|^2 - \|\hat{y}_k\|^2\right) - \langle \tilde{s}_k, \hat{y}_k - \hat{x}_k \rangle \\
&\overset{(4),(5)}{\le} \left[\varepsilon - \frac{\chi}{2\lambda_k}\|\hat{y}_k - \hat{x}_{k-1}\|^2 + \frac{1}{2\lambda_k}\|\hat{x}_k - \hat{x}_{k-1}\|^2\right] + \frac{\nu}{2}\left(\|\hat{x}_k\|^2 - \|\hat{y}_k\|^2\right) - \langle \tilde{s}_k, \hat{y}_k - \hat{x}_k \rangle \\
&\overset{(6)}{=} \varepsilon - \frac{\chi}{2\lambda_k}\|\hat{y}_k - \hat{x}_{k-1}\|^2 + \frac{1}{2\lambda_k}\|\hat{x}_k - \hat{x}_{k-1}\|^2 + \frac{\nu}{2}\left(\|\hat{x}_k\|^2 - \|\hat{y}_k\|^2\right) + \left\langle \frac{\hat{x}_k - \hat{x}_{k-1}}{\lambda_k} + \nu \hat{x}_k \, , \hat{y}_k - \hat{x}_k \right\rangle \\
&= \varepsilon + \frac{1 - \chi}{2\lambda_k}\|\hat{y}_k - \hat{x}_{k-1}\|^2 - \frac{1}{2\lambda_k}\|\hat{y}_k - \hat{x}_k\|^2 - \frac{\nu}{2}\|\hat{y}_k - \hat{x}_k\|^2.
\end{aligned}$$

Hence, the second inequality in (10) holds.

(c) Using the inclusion in b), the fact that $\tilde{\phi}$ is $(\mu - \nu)$-convex, and Lemma A.1 of [18], we have for any $\zeta \in (0, \infty]$ and every $u \in \operatorname{dom} \phi$,

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

Now, (11) follows from the above inequality with $\zeta = (1 - \chi)/\chi$ where $\chi$ is as in step 0 of FSCO. ■

Before showing Theorem 2.3 on the complexity of FSCO, we also need the following Proposition 2.2.

**Proposition 2.2** *Consider sequences $\{\hat{x}_k\}$, $\{\hat{y}_k\}$, and $\{\lambda_k\}$ generated by FSCO. For every $u \in \mathbb{R}^n$, we have for every $k \geq 1$, $u \in \mathbb{R}^n$, and $\chi \in [0, 1)$, we have:*

$$\chi(1 + \nu\underline{\lambda})\|\hat{x}_k - \hat{y}_k\|^2 \leq (1 - \chi)\|\hat{x}_{k-1} - x_*\|^2 + 2\lambda_k\varepsilon. \tag{12}$$

*and*

$$2\lambda_k \left[\phi(\hat{y}_k) - \phi(u)\right] \leq \frac{2\lambda_k\varepsilon}{1 - \chi} + \|\hat{x}_{k-1} - u\|^2 - (1 + \sigma)\|\hat{x}_k - u\|^2, \tag{13}$$

*where*

$$\sigma = \sigma(\underline{\lambda}) := \frac{\underline{\lambda}[\nu(1 + \mu\underline{\lambda}) + \chi(\mu - \nu)]}{1 + \mu\underline{\lambda} + \chi\underline{\lambda}(\nu - \mu)}. \tag{14}$$

**Proof**: Let $A(u) := \|\hat{x}_{k-1} - u\|^2 - (1 + \nu\lambda_k)\|\hat{x}_k - u\|^2$. Since $A(u)$ is a quadratic function in $u$, its Taylor's expansion gives

$$A(u) = A(\hat{y}_k) + \langle \nabla A(\hat{y}_k), u - \hat{y}_k\rangle + \frac{1}{2}\langle \nabla^2 A(\hat{y}_k)(u - \hat{y}_k), u - \hat{y}_k\rangle,$$

where $\nabla^2 A(\hat{y}_k) = -2\nu\lambda_k I$ and

$$\nabla A(\hat{y}_k) = 2(\hat{y}_k - \hat{x}_{k-1}) - 2(1 + \nu\lambda_k)(\hat{y}_k - \hat{x}_k) \overset{(6)}{=} -2\nu\lambda_k\hat{y}_k - 2\lambda_k\tilde{s}_k.$$

Using the above formulas, we have for $A(u) - A(\hat{y}_k)$ the expresssion

$$\begin{aligned}
&\|\hat{x}_{k-1} - u\|^2 - (1 + \nu\lambda_k)\|\hat{x}_k - u\|^2 - \left(\|\hat{x}_{k-1} - \hat{y}_k\|^2 - (1 + \nu\lambda_k)\|\hat{x}_k - \hat{y}_k\|^2\right)\\
&= -2\lambda_k\langle \nu\hat{y}_k + \tilde{s}_k, u - \hat{y}_k\rangle - \nu\lambda_k\|u - \hat{y}_k\|^2 = -2\lambda_k\langle \tilde{s}_k, u - \hat{y}_k\rangle - \nu\lambda_k(\|u\|^2 - \|\hat{y}_k\|^2)\\
&\overset{(11)}{\geq} 2\lambda_k\left[\tilde{\phi}(\hat{y}_k) - \tilde{\phi}(u) + \frac{\chi(\mu - \nu)}{2}\|u - \hat{y}_k\|^2 - \frac{\tilde{\eta}_k}{1 - \chi}\right] - \nu\lambda_k(\|u\|^2 - \|\hat{y}_k\|^2)\\
&\overset{(8)}{=} 2\lambda_k\left[\phi(\hat{y}_k) - \phi(u)\right] + \chi(\mu - \nu)\lambda_k\|u - \hat{y}_k\|^2 - \frac{2\lambda_k\tilde{\eta}_k}{1 - \chi},
\end{aligned} \tag{15}$$

where the inequality is due to (11) and the last identity is due to the definition of $\tilde{\phi}$ in (8). Rearranging the terms in (15) and using Lemma 2.1(b), we have

$$\begin{aligned}
&\|\hat{x}_{k-1} - u\|^2 - (1 + \nu\lambda_k)\|\hat{x}_k - u\|^2 - 2\lambda_k\left[\phi(\hat{y}_k) - \phi(u)\right]\\
&\overset{(15)}{\geq} \|\hat{x}_{k-1} - \hat{y}_k\|^2 - (1 + \nu\lambda_k)\|\hat{x}_k - \hat{y}_k\|^2 + \chi(\mu - \nu)\lambda_k\|u - \hat{y}_k\|^2 - \frac{2\lambda_k\tilde{\eta}_k}{1 - \chi}\\
&\overset{(10)}{\geq} -\frac{2\lambda_k\varepsilon}{1 - \chi} + \frac{\chi(1 + \nu\underline{\lambda})}{1 - \chi}\|\hat{x}_k - \hat{y}_k\|^2 + \chi(\mu - \nu)\underline{\lambda}\|u - \hat{y}_k\|^2,
\end{aligned}$$

where in the last inequality we have used Assumption (F1).

Rearranging the terms and using Assumption (F1), we have

$$\begin{aligned}
2\lambda_k\left[\phi(\hat{y}_k) - \phi(u)\right] \leq &\frac{2\lambda_k\varepsilon}{1 - \chi} + \|\hat{x}_{k-1} - u\|^2 - (1 + \nu\underline{\lambda})\|\hat{x}_k - u\|^2\\
&- \chi\left(\frac{1 + \nu\underline{\lambda}}{1 - \chi}\|\hat{x}_k - \hat{y}_k\|^2 + (\mu - \nu)\underline{\lambda}\|u - \hat{y}_k\|^2\right).
\end{aligned} \tag{16}$$

It is clear that (12) follows from (16) with $u = x_*$ and observing that $\phi(\hat{y}_k) - \phi(x_*) \geq 0$. 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 $(a_1, a_2) = (\|\hat{x}_k - \hat{y}_k\|, \|u - \hat{y}_k\|)$ and $(b_1, b_2) = ((1 + \nu\underline{\lambda})/(1 - \chi), (\mu - \nu)\underline{\lambda})$ we have

$$\|\hat{x}_k - u\|^2 \leq \left(\frac{1 - \chi}{1 + \nu\underline{\lambda}} + \frac{1}{(\mu - \nu)\underline{\lambda}}\right)\left(\frac{1 + \nu\underline{\lambda}}{1 - \chi}\|\hat{x}_k - \hat{y}_k\|^2 + (\mu - \nu)\underline{\lambda}\|u - \hat{y}_k\|^2\right). \tag{17}$$

Plugging the above ineqaulity into (16), we have

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

which is the same as (13) after simplification. $\blacksquare$

Before giving the first main complexity result for FSCO, namely, a complexity bound for obtaining a $\bar{\varepsilon}$-solution of (3), we first introduce some terminology used throughout our analysis. Let $x_*$ denote the closest solution of (3) to the initial point $\hat{x}_0$ of FSCO and let $d_0$ denote its distance to $\hat{x}_0$, i.e.,

$$\|\hat{x}_0 - x_*\| = \min\{\|x - \hat{x}_0\| : x \in X_*\}, \quad d_0 = \|\hat{x}_0 - x_*\|. \tag{18}$$

**Theorem 2.3** *For a given tolerance $\bar{\varepsilon} > 0$, consider FSCO with $\varepsilon = (1-\chi)\bar{\varepsilon}/2$, where $\chi \in [0,1)$ is as in step 0 of FSCO. Then, the number of iterations of FSCO to generate an iterate $\hat{y}_k$ satisfying $\phi(\hat{y}_k) - \phi_* \le \bar{\varepsilon}$ is at most*

$$\mathcal{C}_{func}(\bar{\varepsilon}) := \min\left\{ \min\left[\frac{1}{\chi}\left(1 + \frac{1}{\underline{\lambda}\mu}\right), 1 + \frac{1}{\underline{\lambda}\nu}\right] \log\left(1 + \frac{\lambda_0 \mu d_0^2}{\underline{\lambda}\bar{\varepsilon}}\right), \frac{d_0^2}{\underline{\lambda}\bar{\varepsilon}} \right\}. \tag{19}$$

**Proof**: It is easy to see that (13) with $u = x_*$ where $x_*$ is given by (18) satisfies (93) with $\sigma$ is as in (14) and

$$\gamma_k = 2\lambda_k, \quad \eta_k = \phi(\hat{y}_k) - \phi_*, \quad \alpha_k = \|\hat{x}_k - x_*\|^2, \quad \delta = \frac{\varepsilon}{1-\chi}. \tag{20}$$

Also, note that

$$\underline{\gamma} = 2\underline{\lambda}, \quad 2\delta = \frac{2\varepsilon}{1-\chi} = \bar{\varepsilon}.$$

It follows from Lemma A.1(c) with the above parameters and the definition of $\sigma$ in (14) that the complexity to find a $\bar{\varepsilon}$-solution is

$$\min\left\{\frac{1+\sigma}{\sigma}\log\left(1 + \frac{\sigma d_0^2}{\underline{\lambda}\bar{\varepsilon}}\right), \frac{d_0^2}{\underline{\lambda}\bar{\varepsilon}}\right\} \overset{(14)}{=} \min\left\{\frac{(1+\nu\underline{\lambda})(1+\mu\underline{\lambda})}{\underline{\lambda}[\nu(1+\mu\underline{\lambda}) + \chi(\mu-\nu)]}\log\left(1 + \frac{\sigma d_0^2}{\underline{\lambda}\bar{\varepsilon}}\right), \frac{d_0^2}{\underline{\lambda}\bar{\varepsilon}}\right\}. \tag{21}$$

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

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

and hence that

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

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

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

Combining the above two inequalities, we obtain

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

The result now immediately follows from (21), the above inequality, and the fact that

$$\sigma \le \frac{\underline{\lambda}[\nu(1+\mu\underline{\lambda}) + \chi(\mu-\nu)]}{1+\mu\underline{\lambda}} \le \underline{\lambda}\left(\nu + \frac{\chi(\mu-\nu)}{1+\mu\underline{\lambda}}\right) \le \lambda_0\left(\nu + \frac{\chi(\mu-\nu)}{1+\mu\lambda_0}\right) \le \lambda_0\left(\nu + (\mu-\nu)\right) \le \lambda_0\mu.$$

$\blacksquare$

We now comment on the complexity bound obtained in Theorem 2.3. First, the bound

$$\min\left\{\frac{1}{\chi}\left(1 + \frac{1}{\underline{\lambda}\mu}\right), 1 + \frac{1}{\underline{\lambda}\nu}\right\} \log\left(1 + \frac{\lambda_0\mu d_0^2}{\underline{\lambda}\bar{\varepsilon}}\right) \tag{22}$$

implied by (19) is meaningful only when $\chi > 0$ or $\nu > 0$ (otherwise, it should be understood as being infinity). Second, the validity of the second bound in (22) is well-known and can be found for example in [7, 15]. Third, if $\mu \gg \nu$ (see the assumption (F2) in Section 1) and $\chi$ is sufficiently close to one, the smallest term in (22) is the first one, in which case (19) reduces to

$$\frac{1}{\chi}\left(1+\frac{1}{\underline{\lambda}\mu}\right)\log\left(1+\frac{\lambda_0\mu d_0^2}{\underline{\lambda}\bar{\varepsilon}}\right) = \tilde{\mathcal{O}}\left(\frac{1}{\underline{\lambda}\mu}\right).$$

Fourth, since the second bound does not depend on $\nu$, $\mu$ and $\chi$, it holds for any parameters $\mu \geq \nu \geq 0$ and $\chi \in [0,1)$.

The drawback of using the termination criterion $\phi(\bar{x}) - \phi_* \leq \bar{\varepsilon}$ is that it requires knowledge of $\phi_*$. The next subsection presents the complexity of FSCO to obtain a point satisfying a stopping criterion that is easily checkable for any instance of (3).

## 2.2 Stationarity complexity analysis

This subsection studies the iteration complexity for FSCO to obtain a near-stationary solution of (3) (see the definition below). Its main result is stated in Theorem 2.8.

We start by defining the notion of near-stationary solutions considered in this subsection.

**Definition 2.4** *A triple $(x, v, \eta)$ is called $\phi$-compatible if it satisfies the inclusion $v \in \partial_\eta \phi(x)$. For a given tolerance pair $(\hat{\rho}, \hat{\varepsilon})$, a $\phi$-compatible triple $(x, v, \eta)$ is called a $(\hat{\rho}, \hat{\varepsilon})$-stationary solution of (3) if it satisfies $\|v\| \leq \hat{\rho}$ and $\eta \leq \hat{\varepsilon}$.*

We now comment on the benefits of using near-stationary solutions as a way to terminate an algorithm. First, many algorithms, including the ones considered in this paper, naturally generate a sequence of $\phi$-compatible triples $\{(\bar{y}_k, \bar{s}_k, \bar{\varepsilon}_k)\}$ where the sequence of residual pairs $\{(\bar{s}_k, \bar{\varepsilon}_k)\}$ can be made arbitrarily small (see Proposition 2.7 below). As a consequence, some $(\bar{y}_k, \bar{s}_k, \bar{\varepsilon}_k)$ will eventually become a $(\hat{\rho}, \hat{\varepsilon})$-stationary solution of (3). Moreover, verifying this only requires checking whether the two inequalities $\|\bar{s}_k\| \leq \hat{\rho}$ and $\bar{\varepsilon}_k \leq \hat{\varepsilon}$ hold as the inclusion $\bar{s}_k \in \partial_{\bar{\varepsilon}_k}\phi(\bar{y}_k)$ is guaranteed to hold for every $k \geq 1$. Second, this notion is related to the one considered in Subsection 2.1 as follows. If $(\bar{y}_k, \bar{s}_k, \bar{\varepsilon}_k)$ is a $(\hat{\rho}, \hat{\varepsilon})$-stationary solution and $\operatorname{dom}\phi$ has a finite diameter $D$, then it follows that $\bar{y}_k$ is a $(\hat{\varepsilon} + D\hat{\rho})$-solution of (3).

The following lemma will be useful to derive such complexity for stationary conditions:

**Lemma 2.5** *For every $k \geq 1$, define*

$$\bar{y}_k = \operatorname{argmin}\{\phi(y) : y \in \{\hat{y}_1, \ldots, \hat{y}_k\}\} \tag{23}$$

*and*

$$S_k = \sum_{j=1}^{k} (1+\sigma)^{j-1}\lambda_j \tag{24}$$

*where $\sigma$ is as in (14). Then, for every $u \in \operatorname{dom}\phi$, we have:*

$$\phi(\bar{y}_k) - \phi(u) \leq \frac{\|\hat{x}_0 - u\|^2 - (1+\sigma)^k\|\hat{x}_k - u\|^2}{2S_k} + \frac{\varepsilon}{1-\chi}, \tag{25}$$

$$\|\hat{x}_k - x_*\|^2 \leq \frac{d_0^2}{(1+\sigma)^k} + \frac{2\varepsilon S_k}{(1-\chi)(1+\sigma)^k}, \tag{26}$$

*where $x_*$ and $d_0$ are as in (18).*

**Proof**: First note that (13) is a special case of (93) with

$$\gamma_k = 2\lambda_k, \quad \eta_k = \phi(\hat{y}_k) - \phi(u), \quad \alpha_k = \|\hat{x}_k - u\|^2, \quad \delta = \frac{\varepsilon}{1-\chi}.$$

Then, (25) follows from Lemma A.1(a). It is also easy to verify that (13) with $u = x_*$ satisfies (93) with parameters as in (20). Then, (26) follows from Lemma A.1(b). ∎

**Lemma 2.6** *For every $k \geq 1$, we have*

$$\|\hat{x}_0 - \hat{x}_k\|^2 \leq 4\left(d_0^2 + \frac{S_k \varepsilon}{1-\chi}\right), \tag{27}$$

$$\|\hat{x}_0 - \bar{y}_k\|^2 \leq \frac{1}{\chi}\left(5d_0^2 + \frac{8S_k\varepsilon}{1-\chi}\right). \tag{28}$$

**Proof**: Using (26) and the fact that $\sigma \geq 0$, we have for every $i \in \{1, 2, \ldots, k\}$,

$$\|\hat{x}_i - x_*\|^2 \overset{(26)}{\leq} \frac{d_0^2}{(1+\sigma)^i} + \frac{2\varepsilon S_i}{(1-\chi)(1+\sigma)^i} \leq d_0^2 + \frac{2\varepsilon S_i}{1-\chi} \leq d_0^2 + \frac{2\varepsilon S_k}{1-\chi}. \tag{29}$$

It follows from the triangle inequality that

$$\|\hat{x}_0 - \hat{x}_k\|^2 \leq (\|\hat{x}_0 - x_*\| + \|\hat{x}_k - x_*\|)^2 \leq 2\left[d_0^2 + \|\hat{x}_k - x_*\|^2\right].$$

which together with (29) implies (27). 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 $(a_1, a_2) = (\|\hat{x}_0 - \hat{x}_k\|, \|\hat{x}_k - \hat{y}_k\|)$ and $(b_1, b_2) = (1, \chi(1+\nu\underline{\lambda})/(1-\chi))$, we have

$$\begin{aligned}
\|\hat{x}_0 - \hat{y}_k\|^2 &\leq (\|\hat{x}_0 - \hat{x}_k\| + \|\hat{x}_k - \hat{y}_k\|)^2 \\
&\leq \left(1 + \frac{1-\chi}{\chi(1+\nu\underline{\lambda})}\right)\left[\|\hat{x}_0 - \hat{x}_k\|^2 + \frac{\chi(1+\nu\underline{\lambda})}{1-\chi}\|\hat{x}_k - \hat{y}_k\|^2\right] \\
&\overset{(12)}{\leq} \frac{1}{\chi}\left(\|\hat{x}_0 - \hat{x}_k\|^2 + \|\hat{x}_{k-1} - x_*\|^2 + \frac{2\lambda_k\varepsilon}{1-\chi}\right),
\end{aligned} \tag{30}$$

where the last inequality is due to (12). Noting that $\bar{y}_k = \hat{y}_i$ for some $i \in \{1, 2, \ldots, k\}$, and using (27), (29), and (30), we have

$$\begin{aligned}
\|\hat{x}_0 - \bar{y}_k\|^2 = \|\hat{x}_0 - \hat{y}_i\|^2 &\overset{(30)}{\leq} \frac{1}{\chi}\left[\|\hat{x}_0 - \hat{x}_i\|^2 + \|\hat{x}_{i-1} - x_*\|^2 + \frac{2\lambda_i\varepsilon}{1-\chi}\right] \\
&\overset{(27),(29)}{\leq} \frac{1}{\chi}\left(4d_0^2 + \frac{4\varepsilon S_k}{1-\chi} + d_0^2 + \frac{2\varepsilon S_k}{1-\chi} + \frac{2\lambda_i\varepsilon}{1-\chi}\right) \leq \frac{1}{\chi}\left(5d_0^2 + \frac{8\varepsilon S_k}{1-\chi}\right),
\end{aligned}$$

where the last inequality is due to the fact that $S_k \geq (1+\sigma)^{i-1}\lambda_i \geq \lambda_i$. Hence, (28) is proved. ■

The following results shows that FSCO naturally generates a sequence of residual pairs $\{(\hat{v}_k, \hat{\varepsilon}_k)\}$ such that $(\hat{z}_k, \hat{v}_k, \hat{\varepsilon}_k)$ is $\phi$-compatible for every $k \geq 1$ and provides suitable bounds for it.

**Proposition 2.7** *For every $k \geq 1$, define*

$$\bar{s}_k = \frac{\hat{x}_0 - \hat{x}_k}{S_k}, \quad \bar{\varepsilon}_k = \frac{\|\hat{x}_0 - \bar{y}_k\|^2 - \|\hat{x}_k - \bar{y}_k\|^2}{2S_k} + \frac{\varepsilon}{1-\chi} \tag{31}$$

*where $\bar{y}_k$ is as in (23). Then, the following statements hold for every $k \geq 1$:*

*a)* $\bar{s}_k \in \partial\phi_{\bar{\varepsilon}_k}(\bar{y}_k)$;

*b) the residual pair $(\bar{s}_k, \bar{\varepsilon}_k)$ is bounded by*

$$\|\bar{s}_k\| \leq \frac{2d_0}{S_k} + \frac{\sqrt{2\varepsilon}}{\sqrt{(1-\chi)S_k}}, \quad \bar{\varepsilon}_k \leq \frac{\|\hat{x}_0 - \bar{y}_k\|^2}{2S_k} + \frac{\varepsilon}{1-\chi}, \tag{32}$$

*where $S_k$ is as in (24).*

**Proof**: (a) Using (25) and the fact that $\sigma \geq 0$, we have for every $u \in \mathbb{R}^n$,

$$\begin{aligned}
\phi(\bar{y}_k) - \phi(u) &\leq \frac{\|\hat{x}_0 - u\|^2 - \|\hat{x}_k - u\|^2}{2S_k} + \frac{\varepsilon}{1-\chi} \\
&= \frac{\|\hat{x}_0 - \bar{y}_k\|^2 - \|\hat{x}_k - \bar{y}_k\|^2 + 2\langle \hat{x}_0 - \hat{x}_k, \bar{y}_k - u\rangle}{2S_k} + \frac{\varepsilon}{1-\chi} \\
&\overset{(31)}{=} \langle \bar{s}_k, \bar{y}_k - u\rangle + \bar{\varepsilon}_k,
\end{aligned}$$

where the last identity is due to the definitions of $\bar{s}_k$ and $\bar{\varepsilon}_k$ in (31). Hence, the statement holds.

(b) Using the triangle inequality, (26), and the fact that $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$, we have

$$\|\bar{s}_k\| \le \frac{d_0 + \|\hat{x}_k - x_*\|}{S_k} \overset{(26)}{\le} \frac{d_0}{S_k} + \frac{1}{S_k}\left(\frac{d_0}{\sqrt{(1+\sigma)^k}} + \frac{\sqrt{2\varepsilon S_k}}{\sqrt{(1-\chi)(1+\sigma)^k}}\right).$$

Hence, the first inequality in (32) follows from the fact that $\sigma \ge 0$. Moreover, the second inequality in (32) follows immediately from the definition of $\bar{\varepsilon}_k$ in (31). ■

The following Theorem 2.8 provides the complexity for stationary conditions announced in the beginning of this section.

**Theorem 2.8** *For a given tolerance pair $(\hat{\varepsilon}, \hat{\rho}) \in \mathbb{R}^2_{++}$, FSCO with*

$$\chi \in (0,1), \quad \varepsilon = \frac{\chi(1-\chi)\hat{\varepsilon}}{10}, \tag{33}$$

*generates a triple $(\bar{y}_k, \bar{s}_k, \bar{\varepsilon}_k)$ satisfying*

$$\bar{s}_k \in \partial\phi_{\bar{\varepsilon}_k}(\bar{y}_k), \quad \|\bar{s}_k\| \le \hat{\rho}, \quad \bar{\varepsilon}_k \le \hat{\varepsilon} \tag{34}$$

*in at most*

$$\min\left\{\min\left[\frac{1}{\chi}\left(1+\frac{1}{\underline{\lambda}\mu}\right), 1+\frac{1}{\underline{\lambda}\nu}\right]\log\left[1+\lambda_0\mu\beta(\hat{\varepsilon},\hat{\rho})\right], \beta(\hat{\varepsilon},\hat{\rho})\right\} \tag{35}$$

*iterations where*

$$\beta(\hat{\varepsilon},\hat{\rho}) = \frac{1}{\underline{\lambda}}\left(\frac{4\chi\hat{\varepsilon}}{5\hat{\rho}^2} + \frac{5d_0^2}{\chi\hat{\varepsilon}}\right). \tag{36}$$

**Proof**: First, it follows from Proposition 2.7(a) that the inclusion $\bar{s}_k \in \partial\phi_{\bar{\varepsilon}_k}(\bar{y}_k)$ holds for every $k \ge 1$. We next show that the number of iterations required for (34) is at most

$$k_0 := \min\left\{\frac{1+\sigma}{\sigma}\log\left[1+\sigma\beta(\hat{\varepsilon},\hat{\rho})\right], \beta(\hat{\varepsilon},\hat{\rho})\right\}. \tag{37}$$

Hence, it suffices to show that $\|\bar{s}_k\| \le \hat{\rho}$ and $\bar{\varepsilon}_k \le \hat{\varepsilon}$ for every $k \ge k_0$. Using the definition of $S_k$ in (24), assumption (F2), and (98), we have

$$S_k \ge \underline{\lambda}\sum_{j=1}^{k}(1+\sigma)^{j-1} \overset{(98)}{\ge} \underline{\lambda}\max\left\{\frac{e^{\sigma k/(1+\sigma)}-1}{\sigma}, k\right\}. \tag{38}$$

It is easy to verify that for every $k \ge k_0$, we have

$$S_k \ge \underline{\lambda}\beta(\hat{\varepsilon},\hat{\rho}). \tag{39}$$

It immediately follows from the inequality that $a^2+b^2 \ge 2ab$ and the definition of $\beta(\hat{\varepsilon},\hat{\rho})$ in (36) that

$$\beta(\hat{\varepsilon},\hat{\rho}) \ge \frac{4d_0}{\underline{\lambda}\hat{\rho}}. \tag{40}$$

Using Lemma 2.6, Proposition 2.7(b), (33), the definition of $\beta(\hat{\varepsilon},\hat{\rho})$ in (36), and the above observation, we have

$$\|\bar{s}_k\| \overset{(32),(33)}{\le} \frac{2d_0}{S_k} + \frac{\sqrt{\chi\hat{\varepsilon}}}{\sqrt{5S_k}} \overset{(36),(40)}{\le} \frac{\beta(\hat{\varepsilon},\hat{\rho})\underline{\lambda}\hat{\rho}}{2S_k} + \frac{\sqrt{\beta(\hat{\varepsilon},\hat{\rho})\underline{\lambda}}\hat{\rho}}{2\sqrt{S_k}}, \tag{41}$$

$$\bar{\varepsilon}_k \overset{(28),(32)}{\le} \frac{5d_0^2}{2\chi S_k} + \frac{4\varepsilon}{\chi(1-\chi)} + \frac{\varepsilon}{1-\chi} \overset{(33)}{\le} \frac{5d_0^2}{2\chi S_k} + \frac{\hat{\varepsilon}}{2} \overset{(36)}{\le} \frac{\beta(\hat{\varepsilon},\hat{\rho})\underline{\lambda}\hat{\varepsilon}}{2S_k} + \frac{\hat{\varepsilon}}{2}. \tag{42}$$

Finally, plugging (39) into (41) and (42), we conclude that $\|\bar{s}_k\| \le \hat{\rho}$ and $\bar{\varepsilon}_k \le \hat{\varepsilon}$ for every $k \ge k_0$, where $k_0$ is as in (37). It follows from the same argument as in the proof of Theorem 2.3 that (35) is an upper bound on $k_0$. Finally, we conclude that the theorem holds. ■

We now make some remarks about Theorem 2.8. In contrast to Theorem 2.3, it does not apply to the case where $\chi = 0$ because the quantity $\beta(\hat{\varepsilon},\hat{\rho})$ that appears in (35) depends on $\chi^{-1}$ (see in (36)). The following result considers two special cases which cover the case $\chi = 0$.

**Theorem 2.9** *For a given tolerance pair $(\hat{\varepsilon}, \hat{\rho}) \in \mathbb{R}^2_{++}$, the following statements hold for FSCO with $\chi \in [0, 1)$:*

*a) if $\operatorname{dom}\phi$ is bounded with diameter $D > 0$, then FSCO with $\varepsilon = (1-\chi)\hat{\varepsilon}/2$ generates a triple $(\bar{y}_k, \bar{s}_k, \bar{\varepsilon}_k)$ satisfying (34) within a number iterations bounded by (35) but with $\beta(\hat{\varepsilon}, \hat{\rho})$ now given by*

$$\beta(\hat{\varepsilon}, \hat{\rho}) = \frac{2}{\underline{\lambda}} \left( \frac{2\hat{\varepsilon}}{\hat{\rho}^2} + \frac{d_0^2 + D^2}{\hat{\varepsilon}} \right); \tag{43}$$

*b) the special case of FSCO, where $\varepsilon = (1-\chi)\hat{\varepsilon}/6$ and $\hat{y}_k = \hat{x}_k$ for every $k \geq 1$, generates a triple $(\bar{y}_k, \bar{s}_k, \bar{\varepsilon}_k)$ satisfying (34) within a number iterations bounded by (35) but with $\beta(\hat{\varepsilon}, \hat{\rho})$ now given by*

$$\beta(\hat{\varepsilon}, \hat{\rho}) = \frac{4}{\underline{\lambda}} \left( \frac{\hat{\varepsilon}}{3\hat{\rho}^2} + \frac{d_0^2}{\hat{\varepsilon}} \right). \tag{44}$$

**Proof**: a) Since $\hat{y}_k$ and $x_*$ are in $\operatorname{dom}\phi$, it follows from the boundedness assumption that $\|\hat{y}_k - x_*\| \leq D$. Using the inequality and the triangle inequality, we have

$$\|\hat{x}_0 - \bar{y}_k\|^2 \leq (\|\hat{x}_0 - x_*\| + \|x_* - \bar{y}_k\|)^2 \leq 2(d_0^2 + D^2).$$

Thus, it follows the second inequality in (32) and the fact that $\varepsilon = (1-\chi)\hat{\varepsilon}/2$ that

$$\bar{\varepsilon}_k \overset{(32)}{\leq} \frac{d_0^2 + D^2}{S_k} + \frac{\hat{\varepsilon}}{2} \overset{(43)}{\leq} \frac{\beta(\hat{\varepsilon}, \hat{\rho})\underline{\lambda}\hat{\varepsilon}}{2S_k} + \frac{\hat{\varepsilon}}{2},$$

where the last inequality is due to the definition of $\beta(\hat{\varepsilon}, \hat{\rho})$ in (43). Similarly, using the first inequality in (32), the fact that $\varepsilon = (1-\chi)\hat{\varepsilon}/2$, and (43), we have

$$\|\bar{s}_k\| \overset{(32)}{\leq} \frac{2d_0}{S_k} + \frac{\sqrt{\hat{\varepsilon}}}{\sqrt{S_k}} \overset{(43)}{\leq} \frac{\beta(\hat{\varepsilon}, \hat{\rho})\underline{\lambda}\hat{\rho}}{2S_k} + \frac{\sqrt{\beta(\hat{\varepsilon}, \hat{\rho})\underline{\lambda}}\hat{\rho}}{2\sqrt{S_k}},$$

where the second inequality is also due to the observation that $\beta(\hat{\varepsilon}, \hat{\rho}) \geq 4d_0/(\underline{\lambda}\hat{\rho})$ in view of (43). Finally, the rest of the proof follows from the same argument as in the proof of Theorem 2.8.

b) Since $\hat{y}_k = \hat{x}_k$ for every $k \geq 1$, we know $\bar{y}_k = \hat{y}_i = \hat{x}_i$ for some $i \in \{1, 2, \ldots, k\}$. This observation, (27), and the second inequality in (32) imply that

$$\bar{\varepsilon}_k \overset{(32)}{\leq} \frac{\|\hat{x}_0 - \hat{x}_i\|^2}{2S_k} + \frac{\varepsilon}{1-\chi} \overset{(27)}{\leq} \frac{2d_0^2}{S_k} + \frac{2\varepsilon S_i}{(1-\chi)S_k} + \frac{\varepsilon}{1-\chi} \leq \frac{2d_0^2}{S_k} + \frac{3\varepsilon}{1-\chi},$$

where the last inequality is due to the fact that $S_i \leq S_k$ for $i \leq k$. It follows from the fact that $\varepsilon = (1-\chi)\hat{\varepsilon}/6$ and the definition of $\beta(\hat{\varepsilon}, \hat{\rho})$ in (44) that

$$\bar{\varepsilon}_k \leq \frac{2d_0^2}{S_k} + \frac{\hat{\varepsilon}}{2} \overset{(44)}{\leq} \frac{\beta(\hat{\varepsilon}, \hat{\rho})\underline{\lambda}\hat{\varepsilon}}{2S_k} + \frac{\hat{\varepsilon}}{2}.$$

Similarly, using the first inequality in (32), the fact that $\varepsilon = (1-\chi)\hat{\varepsilon}/6$, and (44), we have

$$\|\bar{s}_k\| \overset{(32)}{\leq} \frac{2d_0}{S_k} + \frac{\sqrt{\hat{\varepsilon}}}{\sqrt{3S_k}} \overset{(44)}{\leq} \frac{\beta(\hat{\varepsilon}, \hat{\rho})\underline{\lambda}\hat{\rho}}{2S_k} + \frac{\sqrt{\beta(\hat{\varepsilon}, \hat{\rho})\underline{\lambda}}\hat{\rho}}{2\sqrt{S_k}},$$

where the second inequality is also due to the observation that $\beta(\hat{\varepsilon}, \hat{\rho}) \geq 4d_0/(\underline{\lambda}\hat{\rho})$ in view of (44). Finally, the rest of the proof follows from the same argument as in the proof of Theorem 2.8. ■

Compared to the complexity result of Theorem 2.8 (which does not apply to $\chi = 0$) where term $\chi$ appears in the denominator, the one of Theorem 2.9 (which applies to $\chi = 0$) involves a term $\beta(\hat{\varepsilon}, \hat{\rho})$ where the term $\chi$ is now removed from the denominator.

# 3 Universal composite subgradient and proximal bundle methods

This section presents the two $\mu$-universal methods for solving (1), namely, U-CS in Subsection 3.1 and U-PB in Subsection 3.2. We prove that both methods are instances of FSCO and establish their iteration complexities in terms of function value and stationarity based on the analysis in Section 2.

We assume that conditions (A1) and (A2) hold for (1). We also assume that

(A3) $h \in \overline{\mathrm{Conv}}_\nu(\mathbb{R}^n)$ for some $0 \le \nu \le \mu$;

(A4) $f \in \overline{\mathrm{Conv}}(\mathbb{R}^n)$ is such that $\mathrm{dom}\, h \subset \mathrm{dom}\, f$, and a subgradient oracle, i.e., a function $f' : \mathrm{dom}\, h \to \mathbb{R}^n$ satisfying $f'(x) \in \partial f(x)$ for every $x \in \mathrm{dom}\, h$, is available;

(A5) there exists $(M_f, L_f) \in \mathbb{R}^2_+$ such that for every $x, y \in \mathrm{dom}\, h$,

$$\|f'(x) - f'(y)\| \le 2M_f + L_f\|x - y\|.$$

It is well known that (A5) implies that for every $x, y \in \mathrm{dom}\, h$,

$$f(x) - \ell_f(x; y) \le 2M_f\|x - y\| + \frac{L_f}{2}|x - y|2. \tag{45}$$

Also, for a given tolerance $\varepsilon > 0$, the fact that the set $\Omega := \{(M_f, L_f) \in \mathbb{R}^2_+ : \text{(A5) holds with } (M_f, L_f)\}$ is a (nonempty) closed convex set implies that there exists a unique pair $(\overline{M}_f, \overline{L}_f) := (\overline{M}_f(\varepsilon), \overline{L}_f(\varepsilon))$ that minimizes $M_f^2 + \varepsilon L_f$ over $\Omega$ (referred to as the $\varepsilon$-best pair).

## 3.1 A universal composite subgradient method

The U-CS method is a variant of the universal primal gradient method of [21] by introducing a parameter $\chi \in [0, 1)$. It can also be shown as an instance of FSCO, and we thus establish its complexity using the analysis in Section 2. The method is described below.

U-CS

```
0. Let $\hat{x}_0 \in \mathrm{dom}\, h$, $\chi \in [0,1)$, $\lambda_0 > 0$, and $\varepsilon > 0$ be given, and set $\lambda = \lambda_0$ and $j = 1$;
1. Compute
       $x = \underset{u \in \mathbb{R}^n}{\mathrm{argmin}} \left\{ \ell_f(u; \hat{x}_{j-1}) + h(u) + \frac{1}{2\lambda}\|u - \hat{x}_{j-1}\|^2 \right\};$
2. If $f(x) - \ell_f(x; \hat{x}_{j-1}) - (1-\chi)\|x - \hat{x}_{j-1}\|^2/(2\lambda) \le \varepsilon$ does not hold, then set $\lambda = \lambda/2$ and go to step 1;
   else, go to step 3;
3. Set $\lambda_j = \lambda$, $\hat{x}_j = x$, $j \leftarrow j+1$, and go to step 1.
```

We now make some remarks about U-CS. First, no stopping criterion is added to it since our goal is to analyze its iteration-complexity for obtaining two types of approximate solutions, i.e., either a $\bar{\varepsilon}$-solution (see Theorem 3.2 below) or a $(\hat{\rho}, \hat{\varepsilon})$-stationary solution (see Theorem 3.3 below). Second, U-CS with $\chi = 0$ and $\varepsilon = \bar{\varepsilon}/2$ is exactly the universal primal gradient method analyzed in [21]. Hence, with the introduction of the damping parameter $\chi$, U-CS can be viewed as a generalization of the method of [21].

The following result shows that U-CS is an instance of FSCO and that assumptions (F1) and (F2) of Section 2 are satisfied.

**Proposition 3.1** *The following statements hold for U-CS:*

*a)* $\{\lambda_k\}$ *is a non-increasing sequence;*

*b) for every $k \geq 1$, we have*

$$\hat{x}_k = \underset{u \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ \ell_f(u; \hat{x}_{k-1}) + h(u) + \frac{1}{2\lambda_k}\|u - \hat{x}_{k-1}\|^2 \right\}, \tag{46}$$

$$f(\hat{x}_k) - \ell_f(\hat{x}_k; \hat{x}_{k-1}) + \frac{\chi - 1}{2\lambda_k}\|\hat{x}_k - \hat{x}_{k-1}\|^2 \leq \varepsilon, \tag{47}$$

$$\lambda_k \geq \underline{\lambda}(\varepsilon) := \min \left\{ \frac{(1-\chi)\varepsilon}{4\left(\overline{M}_f^2 + \varepsilon \overline{L}_f\right)}, \lambda_0 \right\}. \tag{48}$$

*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) assumptions (F1) and (F2) are satisfied with $\underline{\lambda} = \underline{\lambda}(\varepsilon)$ given by (48) and $\nu$ from assumption (A3).*

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

b) Relations (46) and (47) directly follow from the description of U-CS. Using (45) with $(M_f, L_f, u, v) = (\overline{M}_f, \overline{L}_f, \hat{x}_k, \hat{x}_{k-1})$ and the inequality $a^2 + b^2 \geq 2ab$ for $a, b \in \mathbb{R}$, we have

$$\begin{aligned} f(\hat{x}_k) - \ell_f(\hat{x}_k; \hat{x}_{k-1}) + \frac{\chi - 1}{2\lambda_{k-1}}\|\hat{x}_k - \hat{x}_{k-1}\|^2 &\overset{(45)}{\leq} 2\overline{M}_f\|\hat{x}_k - \hat{x}_{k-1}\| + \frac{\overline{L}_f}{2}\|\hat{x}_k - \hat{x}_{k-1}\|^2 + \frac{\chi - 1}{2\lambda_{k-1}}\|\hat{x}_k - \hat{x}_{k-1}\|^2 \\ &= 2\overline{M}_f\|\hat{x}_k - \hat{x}_{k-1}\| - \frac{1 - \chi - \lambda_{k-1}\overline{L}_f}{2\lambda_{k-1}}\|\hat{x}_k - \hat{x}_{k-1}\|^2 \\ &\leq \frac{2\lambda_{k-1}\overline{M}_f^2}{1 - \chi - \lambda_{k-1}\overline{L}_f}. \end{aligned} \tag{49}$$

Observe that $\lambda_{k-1} \leq (1-\chi)\varepsilon/(2(\overline{M}_f^2 + \overline{L}_f\varepsilon))$ implies that $\lambda_{k-1} \leq (1-\chi)\varepsilon/(2\overline{M}_f^2 + \overline{L}_f\varepsilon)$, and hence that

$$\frac{2\lambda_{k-1}\overline{M}_f^2}{1 - \chi - \lambda_{k-1}\overline{L}_f} \leq \varepsilon.$$

The above inequality and (49) thus imply that (47) 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 $\underline{\lambda}(\varepsilon)$ as in (48).

c) Relations (46) and (47) are the analogues of relations (4) and (5) of FSCO with $\hat{y}_k = \hat{x}_k$ and $\hat{\Gamma}_k$ as in (i). Inequality (48) shows that Assumption (F2) is satisfied with $\underline{\lambda} = \underline{\lambda}(\varepsilon)$. Finally, in the $k$-th iteration of U-CS, the model $\hat{\Gamma}_k$ for $\phi = f + h$ being simply the linearization $\ell_f(\cdot, \hat{x}_{k-1}) + h(\cdot)$ with $h \in \overline{\text{Conv}}_\nu(\mathbb{R}^n)$ for some $0 \leq \nu \leq \mu$ (from Assumption (A3)), we obtain that Assumption (F1) is satisfied. ■

We are now in a position to state the main result for the functional complexity of U-CS.

**Theorem 3.2** *Let $\bar{\varepsilon} > 0$ be given and consider U-CS with $\varepsilon = (1-\chi)\bar{\varepsilon}/2$, where $\chi \in [0, 1)$ is as in step 0 of U-CS. Then, the number of iterations of U-CS to generate an iterate $\hat{x}_k$ satisfying $\phi(\hat{x}_k) - \phi_* \leq \bar{\varepsilon}$ is at most*

$$\min\left\{ \min\left[\frac{1}{\chi}\left(1 + \frac{Q_f(\bar{\varepsilon})}{\mu\bar{\varepsilon}}\right), 1 + \frac{Q_f(\bar{\varepsilon})}{\nu\bar{\varepsilon}}\right] \log\left(1 + \frac{\lambda_0 \mu Q_f(\bar{\varepsilon}) d_0^2}{\bar{\varepsilon}^2}\right), \frac{d_0^2 Q_f(\bar{\varepsilon})}{\bar{\varepsilon}^2} \right\} + \left\lceil 2\log\frac{\lambda_0 Q_f(\bar{\varepsilon})}{\bar{\varepsilon}} \right\rceil \tag{50}$$

*where*

$$Q_f(\bar{\varepsilon}) = \frac{8\overline{M}_f^2}{(1-\chi)^2} + \bar{\varepsilon}\left(\lambda_0^{-1} + \frac{8\overline{L}_f}{(1-\chi)^2}\right) \tag{51}$$

**Proof**: Define

$$\bar{k} = \left\lceil 2\log\max\left\{\frac{4\lambda_0(\overline{M}_f^2 + \varepsilon\overline{L}_f)}{(1-\chi)\varepsilon}, 1\right\} \right\rceil. \tag{52}$$

Observe that $\lambda_0/2^k \le \underline{\lambda}(\varepsilon)$ for $k \ge \bar{k}$, and from Lemma 3.1 that we cannot halve $\lambda$ more than $\bar{k}$ iterations. It follows from $\varepsilon \le \bar{\varepsilon}$ that $\overline{M}_f^2 + \varepsilon \overline{L}_f \le \overline{M}_f^2 + \bar{\varepsilon}\overline{L}_f$, which together with $\varepsilon = (1-\chi)\bar{\varepsilon}/2$ implies that

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

Therefore, the second term on the right-hand side of (50) 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. Theorem 2.3 (which can be applied since we have shown that U-CS is a special case of FSCO) gives for U-CS the upper bound

$$\min\left\{ \min\left[ \frac{1}{\chi}\left(1 + \frac{1}{\underline{\lambda}(\varepsilon)\mu}\right), 1 + \frac{1}{\underline{\lambda}(\varepsilon)\nu} \right] \log\left(1 + \frac{\lambda_0 \mu d_0^2}{\underline{\lambda}(\varepsilon)\bar{\varepsilon}}\right), \frac{d_0^2}{\underline{\lambda}(\varepsilon)\bar{\varepsilon}} \right\} \tag{53}$$

on the number of the remaining iterations (where $\lambda$ is not halved) required to find an $\varepsilon$-optimal solution of (1) with $\underline{\lambda}(\varepsilon)$ given by (48). Using the inequality

$$\frac{1}{\underline{\lambda}(\varepsilon)} = \max\left\{ \frac{4(\overline{M}_f^2 + \varepsilon\overline{L}_f)}{(1-\chi)\varepsilon}, \frac{1}{\lambda_0} \right\} \le \frac{4(\overline{M}_f^2 + \varepsilon\overline{L}_f)}{(1-\chi)\varepsilon} + \frac{1}{\lambda_0}, \tag{54}$$

the assumption that $\varepsilon = (1-\chi)\bar{\varepsilon}/2$, and the definition of $Q_f(\bar{\varepsilon})$ in (51), we conclude that $1/\underline{\lambda}(\varepsilon) \le Q_f(\bar{\varepsilon})/\bar{\varepsilon}$. This observation and (53) thus imply that the first term in (50) is an upper bound on the number of the remaining iterations (where $\lambda$ is not halved). This completes the proof. ■

We now make some comments about Theorem 3.2. First, Theorem 3.2 applies to any $\chi \in [0,1)$, and hence to the universal primal gradient method of [21]. In this case, if $\lambda_0^{-1} = \mathcal{O}(1)$, then the strong convexity part of the bound in (50) is

$$\tilde{\mathcal{O}}\left(\frac{Q_f(\bar{\varepsilon})}{\nu\bar{\varepsilon}}\right) = \tilde{\mathcal{O}}\left(\frac{\overline{M}_f^2}{\nu\bar{\varepsilon}} + \frac{\overline{L}_f}{\nu}\right), \tag{55}$$

which is identical to the one in Proposition C.3 of [17]. Second, if $\chi > 0$ and $\lambda_0^{-1} = \mathcal{O}(1)$, then the complexity bound (50) is also

$$\tilde{\mathcal{O}}\left(\frac{Q_f(\bar{\varepsilon})}{\chi\mu\bar{\varepsilon}}\right) = \tilde{\mathcal{O}}\left(\frac{\overline{M}_f^2}{\chi\mu\bar{\varepsilon}} + \frac{\overline{L}_f}{\chi\mu}\right), \tag{56}$$

which is smaller than (55) whenever $\chi\mu \ge \nu$. For example, if $\chi = 1/2$ and $\mu \gg \nu$, then (56) is quite smaller than (55). In summary, the performance of the U-CS with the damping parameter $\chi > 0$ depends on the strong convexity parameter $\mu$ of the overall objective function $\phi$ and hence is potentially more universal than the universal primal gradient method of [21] whose complexity bound depends only on the strong convexity parameter $\nu$ of the composite function $h$.

The following result states the complexity of stationary complexity of U-CS.

**Theorem 3.3** *For a given tolerance pair $(\hat{\varepsilon}, \hat{\rho}) \in \mathbb{R}^2_{++}$, consider U-CS with $\chi \in [0,1)$ and $\varepsilon = (1-\chi)\hat{\varepsilon}/6$. Define $\bar{y}_k$ as in (23) with sequence $\{\hat{y}_k\}$ replaced by $\{\hat{x}_k\}$, and let $\bar{s}_k$ and $\bar{\varepsilon}_k$ be as in (31) where the sequence $\{\hat{x}_k\}$ is generated by U-CS. Then for every $k \ge 1$, U-CS generates a triple $(\bar{y}_k, \bar{s}_k, \bar{\varepsilon}_k)$ satisfying (34) within a number of iterations bounded by*

$$\min\left\{ \min\left[ \frac{1}{\chi}\left(1 + \frac{Q_s(\hat{\varepsilon})}{\mu\hat{\varepsilon}}\right), 1 + \frac{Q_s(\hat{\varepsilon})}{\nu\hat{\varepsilon}} \right] \log C(\hat{\varepsilon}, \hat{\rho}), \frac{4Q_s(\hat{\varepsilon})}{\hat{\varepsilon}}\left(\frac{\hat{\varepsilon}}{3\hat{\rho}^2} + \frac{d_0^2}{\hat{\varepsilon}}\right) \right\} + \left\lceil 2\log\frac{\lambda_0 Q_s(\hat{\varepsilon})}{\hat{\varepsilon}} \right\rceil \tag{57}$$

*where*

$$C(\hat{\varepsilon}, \hat{\rho}) = 1 + \frac{4\lambda_0\mu Q_s(\hat{\varepsilon})}{\hat{\varepsilon}}\left(\frac{\hat{\varepsilon}}{3\hat{\rho}^2} + \frac{d_0^2}{\hat{\varepsilon}}\right), \quad Q_s(\hat{\varepsilon}) = \frac{24\overline{M}_f^2}{(1-\chi)^2} + \hat{\varepsilon}\left(\frac{1}{\lambda_0} + \frac{24\overline{L}_f}{(1-\chi)^2}\right).$$

**Proof**: Same as in the proof of Theorem 3.2, integer $\bar{k}$ given by (52) gives an upper bound on the number of iterations where $\lambda$ is halved. Using $\varepsilon = (1-\chi)\hat{\varepsilon}/6$ and $\overline{M}_f^2 + \varepsilon\overline{L}_f \le \overline{M}_f^2 + \hat{\varepsilon}\overline{L}_f$, we have that

$$\bar{k} \le \left\lceil 2\log\frac{\lambda_0 Q_s(\hat{\varepsilon})}{\hat{\varepsilon}} \right\rceil,$$

which gives the second term on the right-hand side of (57). Next, using Theorem 2.9(b) (observe that this theorem can be applied to U-CS since we have $\hat{y}_k = \hat{x}_k$ for U-CS and we consider the possibility for $\chi$ to be 0), the number of remaining iterations to satisfy (34) is upper bounded by

$$\min\left\{\min\left[\frac{1}{\chi}\left(1+\frac{1}{\underline{\lambda}(\varepsilon)\mu}\right), 1+\frac{1}{\underline{\lambda}(\varepsilon)\nu}\right]\log\left[1+\lambda_0\mu\beta(\hat{\varepsilon},\hat{\rho})\right],\ \beta(\hat{\varepsilon},\hat{\rho})\right\} \tag{58}$$

where

$$\beta(\hat{\varepsilon},\hat{\rho}) = \frac{4}{\underline{\lambda}(\varepsilon)}\left(\frac{\hat{\varepsilon}}{3\hat{\rho}^2}+\frac{d_0^2}{\hat{\varepsilon}}\right)$$

with $\underline{\lambda}(\varepsilon)$ given by (48). Now, using the inequality (54), the assumption that $\varepsilon = (1-\chi)\hat{\varepsilon}/6$, and the definition of $Q_s(\hat{\varepsilon})$ in (51), we conclude that $1/\underline{\lambda}(\varepsilon) \le Q_s(\hat{\varepsilon})/\hat{\varepsilon}$. This observation and (58) then imply that the first term in (57) is an upper bound on the number of the remaining iterations (where $\lambda$ is not halved) required to satisfy (34). This completes the proof. ■

It is easy to see that when $\lambda_0$ is large, $\chi$ is close to 1, and $\mu \gg \nu$, the upper bound (57) on the number of iterations for U-CS to generate a triple $(\bar{y}_k, \bar{s}_k, \bar{\varepsilon}_k)$ satisfying (34) reduces again to (56).

## 3.2 A universal proximal bundle method

This subsection describes the U-PB method and establishes its iteration complexities. More specifically, § 3.2.1 describes U-PB and states the main results, namely Theorems 3.6 and 3.7, and § 3.2.2 is devoted to the proof of a technical result (i.e., Proposition 3.4) that is crucial to the proof of the main results. Conditions (A1)-(A5) are assumed to hold in this subsection.

### 3.2.1 Description of U-PB and related complexity results

The U-PB method is an extension of the GPB method of [17]. In contrast to GPB, we use an adaptive stepsize and introduce a maximal number $\overline{N}$ (which can be as small as one) of iterations for all cycles. Similarly to U-CS, U-PB is another instance of FSCO and we establish both functional and stationary complexities for U-PB using the results of Section 2. Compared with the complexity results in [17], those obtained in this paper are sharper, since they are expressed in terms of $\mu = \mu_\phi$ instead of $\mu_h$.

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

BU$(x^c, x, f_M, \lambda)$

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

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

Find function $f_M^+$ such that

$$f_M^+ \in \overline{\text{Conv}}(\mathbb{R}^n), \qquad \max\{\overline{f}, \ell_f(\cdot; x)\} \le f_M^+ \le f, \tag{59}$$

where $\ell_f(\cdot;\cdot)$ is as in (2) and $\overline{f}$ is such that

$$\overline{f} \le f, \quad \overline{f} \in \overline{\text{Conv}}(\mathbb{R}^n), \quad \overline{f}(x) = f(x), \quad x = \underset{u\in\mathbb{R}^n}{\operatorname{argmin}}\left\{\overline{f}(u) + h(u) + \frac{1}{2\lambda}\|u - x^c\|^2\right\}. \tag{60}$$

**Output:** $f_M^+$.

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 [17].

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

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

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

Before giving the motivation of U-PB, we briefly review the GPB method of [17]. 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

$$\hat{x}_k \approx \underset{u \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ (f+h)(u) + \frac{1}{2\lambda}\|u - \hat{x}_{k-1}\|^2 \right\}. \tag{61}$$

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

$$x_j = \underset{u \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ (f_j+h)(u) + \frac{1}{2\lambda}\|u - \hat{x}_{k-1}\|^2 \right\}, \tag{62}$$

where $f_j \le f$ is a bundle approximation of $f$, is solved until for the first time an iterate $x_j$ as in (62) approximately solves (61), 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 (61) replaced by $\lambda_k$) instead of a constant one as in GPB. Given iteration upper limit $\overline{N} \ge 1$ 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 $\overline{N}$ prox subproblems of the form (62) in an attempt to obtain an approximate solution of (61) 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.

U-PB is given below.

---

U-PB

---

0. Let $\hat{x}_0 \in \operatorname{dom} h$, $\lambda_1 = \lambda > 0$, $\chi \in [0,1)$, $\varepsilon > 0$, and integer $\overline{N} \ge 1$ be given, and set $y_0 = \hat{x}_0$, $N = 0$, $j = 1$, and $k = 1$. Find $f_1 \in \overline{\text{Conv}}(\mathbb{R}^n)$ such that $\ell_f(\cdot; \hat{x}_0) \le f_1 \le f$;

1. Compute $x_j$ as in (62);

2. Choose $y_j \in \{x_j, y_{j-1}\}$ such that

$$\phi(y_j) + \frac{\chi}{2\lambda}\|y_j - \hat{x}_{k-1}\|^2 = \min\left\{ \phi(x_j) + \frac{\chi}{2\lambda}\|x_j - \hat{x}_{k-1}\|^2, \phi(y_{j-1}) + \frac{\chi}{2\lambda}\|y_{j-1} - \hat{x}_{k-1}\|^2 \right\}, \tag{63}$$

and set $N = N + 1$ and

$$t_j = \phi(y_j) + \frac{\chi}{2\lambda}\|y_j - \hat{x}_{k-1}\|^2 - \left( (f_j + h)(x_j) + \frac{1}{2\lambda}\|x_j - \hat{x}_{k-1}\|^2 \right); \tag{64}$$

```
3. If $t_j > \varepsilon$ and $N < \overline{N}$ then
        perform a null update, i.e.: set $f_{j+1} = \text{BU}(\hat{x}_{k-1}, x_j, f_j, \lambda)$;
   else
        if $t_j > \varepsilon$ and $N = \overline{N}$
            perform a reset update, i.e., set $\lambda \leftarrow \lambda/2$;
        else (i.e., $t_j \le \varepsilon$ and $N \le \overline{N}$)
```

```
            perform a serious update, i.e., set $\hat{x}_k = x_j$, $\hat{\Gamma}_k = f_j + h$, $\hat{y}_k = y_j$, $\lambda_k = \lambda$, and $k \leftarrow k+1$;
        end if
        set $N = 0$ and find $f_{j+1} \in \overline{\mathrm{Conv}}(\mathbb{R}^n)$ such that $\ell_f(\cdot;\hat{x}_{k-1}) \le f_{j+1} \le f$;
    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 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 $\overline{N}$ iterations. A serious cycle successfully finds $t_j \le \varepsilon$ within $\overline{N}$ 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 some remarks about U-PB. First, it follows from the fact that $f_j \le f$ and the definition of $t_j$ in (64) that the primal gap of the prox subproblem in (61) is upper bounded by $t_j + (1-\chi)\|y_j - \hat{x}_{k-1}\|^2/(2\lambda)$. Hence, if $t_j \le \varepsilon$, then $y_j$ is an $\varepsilon_j$-solution of (61) where $\varepsilon_j = \varepsilon + (1-\chi)\|y_j - \hat{x}_{k-1}\|^2/(2\lambda)$. Second, the GPB method of [17] (resp., [15]) computes $y_j$ using (63) with $\chi = 0$ (resp., $\chi = 1$). In the case where $\chi = 1$, it can be easily seen that $t_j$ is an upper bound on the primal gap of the prox subproblem in (61), and hence that $y_j$ is an $\varepsilon$-solution of (61) if $t_j \le \varepsilon$. Third, the iterate $y_j$ computed in step 2 of U-PB satisfies

$$y_j \in \mathrm{Argmin}\left\{\phi(x) + \frac{\chi}{2\lambda}\|x - \hat{x}_{k-1}\|^2 : x \in \{\hat{y}_{k-1}, x_{\ell_0}, \ldots, x_j\}\right\}, \tag{65}$$

where $\ell_0$ 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 (62) and the point $\hat{y}_{k-1}$ obtained at the end of the previous cycle.

The next proposition shows some useful relations about the sequences $\{\hat{x}_k\}$, $\{\hat{y}_k\}$ and $\{\lambda_k\}$ generated at the end of the serious cycles of U-PB. This proposition will be proved in § 3.2.2.

**Proposition 3.4** *Define*

$$\overline{U}(\varepsilon) = \varepsilon\left[\frac{1}{\lambda_0} + \frac{40\overline{L}_f}{1-\chi}\right] + \frac{32\overline{M}_f^2}{(1-\chi)\overline{N}}\left(1 + \log(\overline{N})\right). \tag{66}$$

*The following statements hold for U-PB:*

*a) every cycle in U-PB has at most $\overline{N}$ inner iterations;*

*b) each stepsize $\lambda_k$ generated by U-PB satisfies*

$$\lambda_k \ge \frac{\varepsilon}{\overline{U}(\varepsilon)}; \tag{67}$$

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

$$\left\lceil 2\log\frac{\lambda_0\overline{U}(\varepsilon)}{\varepsilon}\right\rceil; \tag{68}$$

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

$$\hat{x}_k = \underset{u\in\mathbb{R}^n}{\mathrm{argmin}}\left\{\hat{\Gamma}_k(u) + \frac{1}{2\lambda_k}\|u - \hat{x}_{k-1}\|^2\right\}, \tag{69}$$

$$\phi(\hat{y}_k) + \frac{\chi}{2\lambda_k}\|\hat{y}_k - \hat{x}_{k-1}\|^2 - \left[\hat{\Gamma}_k(\hat{x}_k) + \frac{1}{2\lambda_k}\|\hat{x}_k - \hat{x}_{k-1}\|^2\right] \le \varepsilon. \tag{70}$$

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 3.5** *U-PB is a special case of FSCO where:*

*a) the pair $(\hat{x}_k, \hat{y}_k)$ satisfies relations (4) and (5);*

*b) conditions (F1) and (F2) used in the analysis of FSCO are satisfied.*

**Proof**: a) Relations (4) and (5) are given in Proposition 3.4(d).

b) By Assumption (A3) we have that $\hat{\Gamma}_k \in \overline{\text{Conv}}_\nu(\mathbb{R}^n)$ and therefore Assumption (F1) of FSCO is satisfied for U-PB. Assumption (F2) is given for U-PB in Proposition 3.4(b). ■

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

**Theorem 3.6** *Given tolerance $\bar{\varepsilon} > 0$, consider U-PB and $\varepsilon = (1-\chi)\bar{\varepsilon}/2$, where $\chi \in [0,1)$ is as in step 0 of U-PB. Let $\{\hat{x}_k\}$ and $\{\hat{y}_k\}$ be the sequences generated by U-PB. Then, the number of iterations of U-PB to generate an iterate $\hat{y}_k$ satisfying $\phi(\hat{y}_k) - \phi_* \le \bar{\varepsilon}$ is at most*

$$\min\left\{\min\left[\frac{1}{\chi}\left(\overline{N} + \frac{R_f(\bar{\varepsilon})}{\mu\bar{\varepsilon}}\right), \overline{N} + \frac{R_f(\bar{\varepsilon})}{\nu\bar{\varepsilon}}\right]\log\left(1 + \frac{\lambda_0 \mu R_f(\bar{\varepsilon}) d_0^2}{\bar{\varepsilon}^2 \overline{N}}\right), \frac{d_0^2 R_f(\bar{\varepsilon})}{\bar{\varepsilon}^2 \overline{N}}\right\} + \overline{N}\left\lceil 2\log\frac{\lambda_0 R_f(\bar{\varepsilon})}{\bar{\varepsilon}\overline{N}}\right\rceil \tag{71}$$

*where*

$$R_f(\bar{\varepsilon}) = \bar{\varepsilon}\overline{N}\left[\frac{1}{\lambda_0} + \frac{40\overline{L}_f}{1-\chi}\right] + \frac{64\overline{M}_f^2}{(1-\chi)^2}\left(1 + \log(\overline{N})\right).$$

**Proof**: Since every serious cycle of U-PB has at most $\overline{N}$ inner iterations (by Proposition 3.4(a)), the complexity of serious cycles of U-PB is that of FSCO, given by (19) in Theorem 2.3, multiplied by $\overline{N}$ (observe that the functional complexity of FSCO can be applied to U-PB since we have shown in Proposition 3.5 that serious iterates of U-PB follow the FSCO framework). In this expression, by Proposition 3.4(b), we can bound from above $1/\underline{\lambda}$ by $\overline{U}(\varepsilon)/\varepsilon \le R_f(\bar{\varepsilon})/(\bar{\varepsilon}\overline{N})$ which gives the first term in (71). By Proposition 3.4(c), the number of reset cycles is at most (68) which is bounded from above by

$$\left\lceil 2\log\frac{\lambda_0 \overline{U}(\varepsilon)}{\varepsilon}\right\rceil \le \left\lceil 2\log\frac{\lambda_0 R_f(\bar{\varepsilon})}{\bar{\varepsilon}\overline{N}}\right\rceil,$$

and each of these cycles also has at most $\overline{N}$ inner iterations (by Proposition 3.4(a)). This gives the second term in (71) and the result follows. ■

The complexity result of Theorem 3.6 for the case where $\lambda_0$ is not too small, $\chi$ is neither close to one nor to zero, and $\mu \gg \nu$ reduces to

$$\tilde{\mathcal{O}}\left(\frac{R_f(\bar{\varepsilon})}{\mu\bar{\varepsilon}}\right) = \tilde{\mathcal{O}}\left(\frac{\overline{M}_f^2}{\mu\bar{\varepsilon}}\right) \tag{72}$$

and we obtain the functional complexity (56) of U-CS. For the case where $\overline{L}_f = 0$, the above complexity is optimal up to logarithmic terms.

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

**Theorem 3.7** *For a given tolerance pair $(\hat{\varepsilon}, \hat{\rho}) \in \mathbb{R}^2_{++}$, U-PB with*

$$\chi \in (0,1), \quad \varepsilon = \frac{\chi(1-\chi)\hat{\varepsilon}}{10}, \tag{73}$$

*generates a triple $(\bar{y}_k, \bar{s}_k, \bar{\varepsilon}_k)$ satisfying (34) in at most*

$$\min\left\{\min\left[\frac{1}{\chi}\left(\overline{N} + \frac{R_s(\hat{\varepsilon})}{\hat{\varepsilon}\mu}\right), \overline{N} + \frac{R_s(\hat{\varepsilon})}{\hat{\varepsilon}\nu}\right]\log C(\hat{\varepsilon}, \hat{\rho}), \frac{R_s(\hat{\varepsilon})}{\hat{\varepsilon}\overline{N}}\left(\frac{4\chi\hat{\varepsilon}}{5\hat{\rho}^2} + \frac{5d_0^2}{\chi\hat{\varepsilon}}\right)\right\} + \overline{N}\left\lceil 2\log\frac{\lambda_0 R_s(\hat{\varepsilon})}{\hat{\varepsilon}\overline{N}}\right\rceil \tag{74}$$

*iterations where*

$$\begin{aligned} C(\hat{\varepsilon}, \hat{\rho}) &= 1 + \frac{\lambda_0 \mu R_s(\hat{\varepsilon})}{\hat{\varepsilon}\overline{N}}\left(\frac{4\chi\hat{\varepsilon}}{5\hat{\rho}^2} + \frac{5d_0^2}{\chi\hat{\varepsilon}}\right) \quad \text{and} \\ R_s(\hat{\varepsilon}) &= \hat{\varepsilon}\overline{N}\left[\frac{1}{\lambda_0} + \frac{40\overline{L}_f}{1-\chi}\right] + \frac{320\overline{M}_f^2}{\chi(1-\chi)^2}\left(1 + \log(\overline{N})\right). \end{aligned} \tag{75}$$

**Proof**: The complexity of serious cycles of U-PB to satisfy stationarity conditions (34) is that of FSCO, given by (35) in Theorem 2.8, multiplied by $\overline{N}$. In this expression

$$\min\left\{\min\left[\frac{1}{\chi}\left(1+\frac{1}{\underline{\lambda}\mu}\right), 1+\frac{1}{\underline{\lambda}\nu}\right]\log\left[1+\lambda_0\mu\beta(\hat{\varepsilon},\hat{\rho})\right]\,,\,\beta(\hat{\varepsilon},\hat{\rho})\right\}$$

where

$$\beta(\hat{\varepsilon},\hat{\rho}) = \frac{1}{\underline{\lambda}}\left(\frac{4\chi\hat{\varepsilon}}{5\hat{\rho}^2}+\frac{5d_0^2}{\chi\hat{\varepsilon}}\right), \tag{76}$$

using the expression (73) for $\varepsilon$, we can bound from above $1/\underline{\lambda}$ by $\overline{U}(\varepsilon)/\varepsilon \le R_s(\hat{\varepsilon})/(\hat{\varepsilon}\overline{N})$ which gives the first term in (74). By Proposition 3.4(c), the number of reset cycles is at most (68) which is bounded from above by

$$\left\lceil 2\log\frac{\lambda_0 R_s(\hat{\varepsilon})}{\hat{\varepsilon}\overline{N}}\right\rceil,$$

and each of these cycles also has at most $\overline{N}$ inner iterations (by Proposition 3.4(a)). This gives the second term in (74) and the result follows. ■

Theorem 3.7 does not hold for $\chi = 0$ since the definitions of $C(\hat{\varepsilon},\hat{\rho})$ and $R_s(\hat{\varepsilon})$ in (75) depend on $\chi^{-1}$. However, if $\chi = 0$ and the domain of $\phi$ is bounded with diameter $D$, then we can apply Theorem 2.9(b) to conclude that U-PB with $\varepsilon = (1-\chi)\hat{\varepsilon}/2$ generates a triple $(\bar{y}_k, \bar{s}_k, \bar{\varepsilon}_k)$ satisfying (34) within a number iterations bounded by (35) but with $\beta(\hat{\varepsilon},\hat{\rho})$ now given by (43) and $1/\underline{\lambda}$ bounded from above by $W_s(\hat{\varepsilon})/\hat{\varepsilon}$ where

$$W_s(\hat{\varepsilon}) = \hat{\varepsilon}\left[\frac{1}{\lambda_0}+\frac{40\overline{L}_f}{1-\chi}\right]+\frac{64\overline{M}_f^2}{(1-\chi)^2\overline{N}}\left(1+\log(\overline{N})\right).$$

### 3.2.2 Proof of Proposition 3.4

To prove Proposition 3.4, we first state Lemmas 3.8 and 3.9. Lemma 3.8 will be used to show Lemma 3.9. Lemma 3.9 plays an important role in the analysis of the null iterates and establishes a key recursive formula for the sequence $\{t_j\}$ defined in (64).

**Lemma 3.8** *For the $j$-th iteration of a cycle with prox stepsize $\lambda$ and prox-center $\hat{x}_{k-1}$, define*

$$m_j := (f_j+h)(x_j)+\frac{1}{2\lambda}\|x_j-\hat{x}_{k-1}\|^2. \tag{77}$$

*If $j$ is not the last iteration of the cycle, then*

$$\begin{aligned} &m_{j+1}-\tau m_j \\ &\ge (1-\tau)\left[\ell_f(x_{j+1};x_j)+h(x_{j+1})+\frac{1}{2\lambda}\|x_{j+1}-\hat{x}_{k-1}\|^2+\left(\frac{\overline{L}_f}{2}+\frac{2\overline{M}_f^2}{\varepsilon}\right)\|x_{j+1}-x_j\|^2\right], \end{aligned} \tag{78}$$

*where*

$$\tau = 1-\left(1+\frac{4\lambda(\overline{M}_f^2+\varepsilon\overline{L}_f)}{\varepsilon}\right)^{-1}. \tag{79}$$

**Proof**: It follows from the definitions of $\tau$ in (79) that

$$\frac{\tau}{2\lambda(1-\tau)} \overset{(79)}{=} 2\overline{L}_f+\frac{2\overline{M}_f^2}{\varepsilon} \ge \frac{\overline{L}_f}{2}+\frac{2\overline{M}_f^2}{\varepsilon}. \tag{80}$$

Using the definition of $m_j$ in (77), and relations (99) and (101) with $u = x_{j+1}$, we have

$$\begin{aligned} m_{j+1} &\overset{(77)}{=} (f_{j+1}+h)(x_{j+1})+\frac{1}{2\lambda}\|x_{j+1}-\hat{x}_{k-1}\|^2 \\ &\overset{(99)}{\ge} (1-\tau)\left[\ell_f(x_{j+1};x_j)+h(x_{j+1})+\frac{1}{2\lambda}\|x_{j+1}-\hat{x}_{k-1}\|^2\right]+\tau\left((\overline{f}_j+h)(x_{j+1})+\frac{1}{2\lambda}\|x_{j+1}-\hat{x}_{k-1}\|^2\right) \\ &\overset{(101)}{\ge} (1-\tau)\left[\ell_f(x_{j+1};x_j)+h(x_{j+1})+\frac{1}{2\lambda}\|x_{j+1}-\hat{x}_{k-1}\|^2\right]+\tau\left(m_j+\frac{1}{2\lambda}\|x_{j+1}-x_j\|^2\right). \end{aligned}$$

This inequality and (80) then imply (78). ■

**Lemma 3.9** *The following statements about U-PB hold:*

*a) for every iteration $j$ that is not the last one of the cycle, we have*

$$t_{j+1} - \frac{\varepsilon}{2} \le \tau \left( t_j - \frac{\varepsilon}{2} \right) \tag{81}$$

*where $\tau$ is as in (79);*

*b) if $i$ is the first iteration of the cycle and $\lambda \le (1-\chi)/(2\overline{L}_f)$, then*

$$t_i \le \frac{4\lambda \overline{M}_f^2}{1-\chi}. \tag{82}$$

**Proof**: a) In what follows, we use the notation

$$\psi(\cdot) := \phi(\cdot) + \frac{\chi}{2\lambda}\| \cdot - \hat{x}_{k-1}\|^2. \tag{83}$$

It follows from the above notation and the definitions of $t_j$ and $m_j$ in (64) and (77), respectively, that

$$t_j = \psi(y_j) - m_j \quad \text{and} \quad \psi(y_{j+1}) \overset{(83),(63)}{=} \min\{\psi(x_{j+1}), \psi(y_j)). \tag{84}$$

Using (45) with $(M_f, L_f, u, v) = (\overline{M}_f, \overline{L}_f, x_{j+1}, x_j)$ and the fact that $\phi = f + h$, we have

$$\ell_f(x_{j+1}; x_j) + h(x_{j+1}) + \frac{\overline{L}_f}{2}\|x_{j+1} - x_j\|^2 \ge \phi(x_{j+1}) - 2\overline{M}_f \|x_{j+1} - x_j\|. \tag{85}$$

This inequality, the definition of $\psi$ in (83), and relation (78), imply that

$$\begin{aligned}
& m_{j+1} - \tau m_j \\
& \overset{(78)}{\ge} (1-\tau)\left[ \ell_f(x_{j+1}; x_j) + h(x_{j+1}) + \frac{1}{2\lambda}\|x_{j+1} - \hat{x}_{k-1}\|^2 + \left( \frac{\overline{L}_f}{2} + \frac{2\overline{M}_f^2}{\varepsilon} \right) \|x_{j+1} - x_j\|^2 \right] \\
& \overset{(85)}{\ge} (1-\tau)\Big( \phi(x_{j+1}) + \frac{1}{2\lambda}\|x_{j+1} - \hat{x}_{k-1}\|^2 \Big) + \frac{1-\tau}{\varepsilon}\Big( 2\overline{M}_f^2 \|x_{j+1} - x_j\|^2 - 2\overline{M}_f \varepsilon \|x_{j+1} - x_j\| \Big) \\
& \overset{(83)}{=} (1-\tau)\Big( \psi(x_{j+1}) + \frac{1-\chi}{2\lambda}\|x_{j+1} - \hat{x}_{k-1}\|^2 \Big) + \frac{1-\tau}{\varepsilon}\Big( 2\overline{M}_f^2 \|x_{j+1} - x_j\|^2 - 2\overline{M}_f \varepsilon \|x_{j+1} - x_j\| \Big) \\
& \ge (1-\tau)\psi(x_{j+1}) - \frac{(1-\tau)\varepsilon}{2},
\end{aligned} \tag{86}$$

where the last inequality follows from the fact that $\chi < 1$ and the inequality $a^2 - 2ab \ge -b^2$ with $a = 2\overline{M}_f \|x_{j+1} - x_j\|$ and $b = \varepsilon$. Using relations (84) and (86), we conclude that

$$\begin{aligned}
t_{j+1} - \tau t_j & \overset{(84)}{=} \psi(y_{j+1}) - m_{j+1} - \tau t_j \\
& \overset{(86)}{\le} \psi(y_{j+1}) - \tau(m_j + t_j) - (1-\tau)\psi(x_{j+1}) + \frac{(1-\tau)\varepsilon}{2} \\
& \overset{(84)}{=} \psi(y_{j+1}) - \tau\psi(y_j) - (1-\tau)\psi(x_{j+1}) + \frac{(1-\tau)\varepsilon}{2} \\
& \overset{(84)}{\le} \frac{(1-\tau)\varepsilon}{2},
\end{aligned}$$

and hence that (81) holds.

b) Using relations (77), (83), and (84), we have

$$\begin{aligned}
t_i & \overset{(77),(84)}{=} \psi(y_i) - \Gamma_i(x_i) - \frac{1}{2\lambda}\|x_i - \hat{x}_{k-1}\|^2 \overset{(84)}{\le} \psi(x_i) - \Gamma_i(x_i) - \frac{1}{2\lambda}\|x_i - \hat{x}_{k-1}\|^2 \\
& \overset{(83)}{=} \phi(x_i) - \Gamma_i(x_i) + \frac{\chi - 1}{2\lambda}\|x_i - \hat{x}_{k-1}\|^2 = f(x_i) - f_i(x_i) + \frac{\chi - 1}{2\lambda}\|x_i - \hat{x}_{k-1}\|^2,
\end{aligned}$$

where the last equality follows from the facts that $\phi = f + h$ and $\Gamma_i = f_i + h$. It follows from U-PB that if $i$ is the first iteration of the cycle, then

$$f_i(\cdot) \ge \ell_f(\cdot; \hat{x}_{k-1}). \tag{87}$$

Combining the above two inequalities and using (45), we obtain

$$\begin{aligned} t_i &\overset{(87)}{\le} f(x_i) - \ell_f(x_i; \hat{x}_{k-1}) + \frac{\chi - 1}{2\lambda}\|x_i - \hat{x}_{k-1}\|^2 \\ &\overset{(45)}{\le} 2\overline{M}_f\|x_i - \hat{x}_{k-1}\| + \frac{\overline{L}_f}{2}\|x_i - \hat{x}_{k-1}\|^2 + \frac{\chi - 1}{2\lambda}\|x_i - \hat{x}_{k-1}\|^2 \\ &= 2\overline{M}_f\|x_i - \hat{x}_{k-1}\| - \frac{1 - \chi - \lambda\overline{L}_f}{2\lambda}\|x_i - \hat{x}_{k-1}\|^2. \end{aligned}$$

By maximizing the right-hand side with respect to $\|x_i - \hat{x}_{k-1}\|$ we deduce $t_i \le (2\lambda_k \overline{M}_f^2)/(1 - \chi - \lambda\overline{L}_f)$ and using the fact that $\lambda \le (1-\chi)/(2\overline{L}_f)$, we obtain (82). ■

Before presenting the proof of Proposition 3.4, we also need the following technical lemma.

**Lemma 3.10** *Assume that the prox stepsize $\lambda$ of some cycle of U-PB is such that $\lambda \le \tilde{\lambda}(\varepsilon)$ where*

$$\tilde{\lambda}(\varepsilon) = \begin{cases} \min\left\{ \dfrac{(\overline{N}-1)\varepsilon}{8(\overline{M}_f^2 + \varepsilon\overline{L}_f)\log(\overline{N})}, \dfrac{(1-\chi)\exp(-1/2)\overline{N}\varepsilon}{8\overline{M}_f^2}, \dfrac{1-\chi}{2\overline{L}_f} \right\} & \text{if } \overline{N} \ge 2, \\ \min\left\{ \dfrac{(1-\chi)\varepsilon}{4\overline{M}_f^2}, \dfrac{1-\chi}{2\overline{L}_f} \right\} & \text{if } \overline{N} = 1. \end{cases} \tag{88}$$

*Then, this cycle must be a serious one.*

**Proof**: Let $i$ denote the first iteration of a cycle whose prox stepsize $\lambda$ satisfies (88). It suffices to prove that $t_\ell \le \varepsilon$ where $\ell = i + \overline{N} - 1$ is the last iteration of the cycle. We consider two cases: $\overline{N} = 1$ and $\overline{N} \ge 2$. If $\overline{N} = 1$, using Lemma 3.9(b) (which can be applied since $\lambda < (1-\chi)/(2\overline{L}_f)$) and (88), we have

$$t_\ell = t_i \le \frac{4\lambda\overline{M}_f^2}{1-\chi} \overset{(88)}{\le} \varepsilon$$

which shows that the cycle is a serious one (with one iteration only). Let us now consider the case $\overline{N} \ge 2$. Lemma 3.9 implies that

$$t_\ell - \frac{\varepsilon}{2} \overset{(81)}{\le} \tau^{\overline{N}-1}\left(t_i - \frac{\varepsilon}{2}\right) \le \tau^{\overline{N}-1} t_i \overset{(82)}{\le} \frac{4\lambda\overline{M}_f^2}{1-\chi}\tau^{\overline{N}-1}. \tag{89}$$

It then suffices to show that the right-hand side of (89) is bounded above by $\varepsilon/2$, or equivalently, that

$$\log\left(\frac{8\lambda\overline{M}_f^2}{(1-\chi)\varepsilon}\right) \le (\overline{N}-1)\log(\tau^{-1}). \tag{90}$$

Using (88), we have

$$\begin{aligned} \lambda \le \frac{\varepsilon}{4(\overline{M}_f^2 + \varepsilon\overline{L}_f)}\left(\frac{\overline{N}-1}{2\log(\overline{N})}\right) &\le \frac{\varepsilon}{4(\overline{M}_f^2 + \varepsilon\overline{L}_f)}\left(\frac{\overline{N}-1}{2\log(\overline{N})-1}\right) \\ &= \frac{\varepsilon}{4(\overline{M}_f^2 + \varepsilon\overline{L}_f)}\left(\frac{\overline{N}-1}{2\log(\exp(-1/2)\overline{N})}\right) \\ &\le \frac{\varepsilon}{4(\overline{M}_f^2 + \varepsilon\overline{L}_f)}\left(\frac{\overline{N}-1}{\log(\exp(-1/2)\overline{N})} - 1\right) \end{aligned} \tag{91}$$

where the last inequality above is easily checked for $\overline{N} \geq 2$. Using (88), (79), and (91), we have

$$\frac{\tau^{-1}}{\tau^{-1}-1} = (1-\tau)^{-1} \overset{(79)}{=} 1 + \frac{4(\overline{M}_f^2 + \varepsilon \overline{L}_f)\lambda}{\varepsilon} \overset{(91)}{\leq} \frac{\overline{N}-1}{\log(\exp(-1/2)\overline{N})} \overset{(88)}{\leq} \frac{\overline{N}-1}{\log\left(\frac{8\overline{M}_f^2\lambda}{(1-\chi)\varepsilon}\right)},$$

which, because of the inequality $\log(\tau^{-1}) \geq (\tau^{-1}-1)/\tau^{-1}$, implies inequality (90). ■

We are now ready to prove Proposition 3.4.

**Proof of Proposition 3.4.** a) This statement immediately follows from the description of U-PB.

b) By Lemma 3.10, if for a cycle we have $\lambda \leq \tilde{\lambda}(\varepsilon)$ with $\tilde{\lambda}(\varepsilon)$ given in (88), then the cycle ends with a serious step and $\lambda$ is kept unchanged for all subsequent cycles and all subsequent cycles are serious cycles. Therefore, if $\lambda_0 \leq \tilde{\lambda}(\varepsilon)/2$ we have $\lambda_k = \lambda_0 = \min\{\lambda_0, \tilde{\lambda}(\varepsilon)/2\} \geq (\varepsilon/\overline{U}(\varepsilon))$ for all $k$ (where the last inequality can be easily checked using the definitions of $\tilde{\lambda}(\varepsilon)$ and $\overline{U}(\varepsilon)$). Otherwise, if $\lambda_0 > \tilde{\lambda}(\varepsilon)/2$, we cannot have $\lambda_k < \tilde{\lambda}(\varepsilon)/2$ for some $k$, i.e., $\lambda_k \geq \tilde{\lambda}(\varepsilon)/2 = \min\{\lambda_0, \tilde{\lambda}(\varepsilon)/2\} \geq \varepsilon/\overline{U}(\varepsilon)$ for all $k$.

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

d) Relations (69) and (70) (which are (5) and (4) in FSCO, respectively) follow from (62) and $t_j \leq \varepsilon$ 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 of U-PB). ■

# 4 Concluding remarks

In this paper, we present two $\mu$-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 both functional and stationary 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. One interesting property of our proposed methods is that they do not rely on any restart scheme based on estimating $\mu$ or knowing $\phi_*$.

Some papers about universal methods (see for example [10, 21]) 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 $x, y \in \operatorname{dom} h$. It is shown in [21] 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

$$\tilde{\mathcal{O}}\left(\frac{d_0^2 L_\alpha^{\frac{2}{\alpha+1}}}{\bar{\varepsilon}^{\frac{2}{\alpha+1}}}\right) \tag{92}$$

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 (50) for the U-CS method is $\tilde{\mathcal{O}}(d_0^2(\overline{M}_f^2 + \bar{\varepsilon}\overline{L}_f)/\bar{\varepsilon}^2)$. Second, Proposition 2.1 of [17] implies that there exists a pair $(M_f, L_f)$ as in (A5) and that the $\bar{\varepsilon}$-best pair $(\overline{M}_f, \overline{L}_f)$ defined below (45) satisfies

$$\overline{M}_f^2 + \bar{\varepsilon}\overline{L}_f \leq 2\bar{\varepsilon}^{\frac{2\alpha}{\alpha+1}} L_\alpha^{\frac{2}{\alpha+1}}.$$

Hence, it follows from these two observations that (50) is sharper than bound (92) obtained in [21].

We finally discuss some possible extensions of our analysis in this paper. First, it is shown in Theorems 3.2 and 3.3 (resp., Theorem 3.6) that U-CS (resp., U-PB) is $\mu$-universal if $\chi > 0$ and is $\nu$-universal if $\chi = 0$. It would be interesting to investigate whether they are also $\mu$-universal for $\chi = 0$ too. Note that this question is related to whether the universal primal gradient of [21] (which is the same as U-CS with $\chi = 0$) is $\mu$-universal. Finally, 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 $\mu$-universal.

# 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] G. Lan. Bundle-level type methods uniformly optimal for smooth and nonsmooth convex optimization. *Mathematical Programming*, 149(1-2):1–45, 2015.

[11] G. Lan and R. D. C. Monteiro. Iteration-complexity of first-order penalty methods for convex programming. *Mathematical Programming*, 138(1):115–139, 2013.

[12] G. Lan, Y. Ouyang, and Z. Zhang. Optimal and parameter-free gradient minimization methods for convex and nonconvex optimization. *arXiv: 2310.12139*, 2023.

[13] T. Li and G. Lan. A simple uniformly optimal method without line search for convex optimization. *arXiv:2310.10082*, 2023.

[14] J. Liang and R. D. C. Monteiro. An average curvature accelerated composite gradient method for nonconvex smooth composite optimization problems. *SIAM Journal on Optimization*, 31(1):217–243, 2021.

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

[16] J. Liang and R. D. C. Monteiro. Average curvature fista for nonconvex smooth composite optimization problems. *Computational Optimization and Applications*, 86(1):275–302, 2023.

[17] 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.

[18] 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.

[19] K. Mishchenko and Y. Malitsky. Adaptive gradient descent without descent. In *37th International Conference on Machine Learning (ICLM 2020)*, 2020.

[20] Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

[21] Y. Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152(1):381–404, 2015.

[22] B. O'donoghue and E. Candes. Adaptive restart for accelerated gradient schemes. *Foundations of Computational Mathematics*, 15:715–732, 2015.

[23] 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.

[24] V. Roulet and A. d'Aspremont. Sharpness, restart and acceleration. *SIAM Journal on Optimization*, 30:262–289, 2020.

[25] D. Zhou, S. Ma, and J. Yang. AdaBB: Adaptive Barzilai-Borwein method for convex optimization. *arXiv:2401.08024*, 2024.

# A Technical results

## A.1 Technical results for FSCO

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

$$\gamma_j \eta_j \le \alpha_{j-1} - (1+\sigma)\alpha_j + \gamma_j \delta \tag{93}$$

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

*a) for every $k \ge 1$,*

$$\min_{1 \le j \le k} \eta_j \le \frac{\alpha_0 - (1+\sigma)^k \alpha_k}{\sum_{j=1}^k (1+\sigma)^{j-1}\gamma_j} + \delta; \tag{94}$$

*b) if the sequence $\{\eta_j\}$ is nonnegative, then for every $k \ge 1$,*

$$\alpha_k \le \frac{\alpha_0}{(1+\sigma)^k} + \frac{\sum_{j=1}^k (1+\sigma)^{j-1}\gamma_j \delta}{(1+\sigma)^k}; \tag{95}$$

*c) if the sequence $\{\alpha_j\}$ is nonnegative, then $\min_{1\le j\le k} \eta_j \le 2\delta$ for every $k \ge 1$ such that*

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

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

$$\begin{aligned} \sum_{j=1}^k (1+\sigma)^{j-1}\gamma_j \left[ \min_{1\le j\le k} \eta_j \right] &\le \sum_{j=1}^k (1+\sigma)^{j-1}\gamma_j\eta_j \le \sum_{j=1}^k (1+\sigma)^{j-1} \left(\alpha_{j-1} - (1+\sigma)\alpha_j + \gamma_j\delta\right) \\ &= \alpha_0 - (1+\sigma)^k\alpha_k + \sum_{j=1}^k (1+\sigma)^{j-1}\gamma_j\delta. \end{aligned} \tag{96}$$

Inequality (94) follows immediately from the above inequality.

b) This statement follows immediately from (96) and the fact that $\eta_j \ge 0$.

c) It follows from (94), and the facts that $\alpha_k \ge 0$ and $\gamma_j \ge \underline{\gamma}$ that

$$\min_{1\le j\le k} \eta_j \le \frac{\alpha_0}{\underline{\gamma}\sum_{j=1}^k (1+\sigma)^{j-1}} + \delta. \tag{97}$$

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

$$\sum_{j=1}^k (1+\sigma)^{j-1} = \max\left\{ \frac{(1+\sigma)^k - 1}{\sigma}, k \right\} \ge \max\left\{ \frac{e^{\sigma k/(1+\sigma)} - 1}{\sigma}, k \right\}. \tag{98}$$

Plugging the above inequality into (97), 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 (c). ■

## A.2 Technical results for U-PB

**Lemma A.2** *The following statements hold for a cycle of U-PB with stepsize $\lambda$:*

*a) for every iteration $j$ that is not the last iteration of the cycle, there exists a function $\overline{f}_j(\cdot)$ such that*

$$\tau(\overline{f}_j + h) + (1-\tau)[\ell_f(\cdot; x_j) + h] \leq f_{j+1} + h \leq \phi, \tag{99}$$

$$\overline{f}_j + h \in \overline{\mathrm{Conv}}_\nu(\mathbb{R}^n), \quad \overline{f}_j(x_j) = f_j(x_j), \quad x_j = \underset{u \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ \overline{f}_j(u) + h(u) + \frac{1}{2\lambda}\|u - \hat{x}_{k-1}\|^2 \right\}, \tag{100}$$

*where $\tau$ is as in (79);*

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

$$\overline{f}_j(u) + h(u) + \frac{1}{2\lambda}\|u - \hat{x}_{k-1}\|^2 \geq m_j + \frac{1}{2\lambda}\|u - x_j\|^2. \tag{101}$$

**Proof**: a) Since $j$ is not the last iteration of a cycle, we have $f_{j+1} = \mathrm{BU}(\hat{x}_{k-1}, x_j, f_j, \lambda)$. Using the properties of the BU blackbox, it follows that there exists $\overline{f}_j$ such that $\overline{f}_j + h \in \overline{\mathrm{Conv}}_\nu(\mathbb{R}^n)$, $\overline{f}_j(x_j) = f_j(x_j)$,

$$\max\{\overline{f}_j + h, \ell_f(\cdot; x_j) + h\} \leq f_{j+1} + h \leq \phi, \tag{102}$$

and

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

We have therefore checked that (100) holds while (99) is an immediate consequence of (102).

b) Since the objective function in the last identity of (100) is $\lambda^{-1}$-strongly convex, we have

$$\overline{f}_j(u) + h(u) + \frac{1}{2\lambda}\|u - \hat{x}_{k-1}\|^2 \geq \overline{f}_j(x_j) + h(x_j) + \frac{1}{2\lambda}\|x_j - \hat{x}_{k-1}\|^2 + \frac{1}{2\lambda}\|u - x_j\|^2. \tag{103}$$

The statement follows from (103), the first identity in (100), and the definition of $m_j$ in (77). ■