# Primal-dual proximal bundle and conditional gradient methods for convex problems

Jiaming Liang *

November 30, 2024 (1st revision: December 23, 2024)

**Abstract**

This paper studies the primal-dual convergence and iteration-complexity of proximal bundle methods for solving nonsmooth problems with convex structures. More specifically, we develop a family of primal-dual proximal bundle methods for solving convex nonsmooth composite optimization problems and establish the iteration-complexity in terms of a primal-dual gap. We also propose a class of proximal bundle methods for solving convex-concave nonsmooth composite saddle-point problems and establish the iteration-complexity to find an approximate saddle-point. This paper places special emphasis on the primal-dual perspective of the proximal bundle method. In particular, we discover an interesting duality between the conditional gradient method and the cutting-plane scheme used within the proximal bundle method. Leveraging this duality, we further develop novel variants of both the conditional gradient method and the cutting-plane scheme.

**Key words.** convex nonsmooth composite optimization, saddle-point problem, proximal bundle method, conditional gradient method, iteration-complexity, primal-dual convergence

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

## 1 Introduction

This paper considers two nonsmooth problems with convex structures: 1) the convex nonsmooth composite optimization (CNCO) problem

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

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

where $f, h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ are proper lower semi-continuous convex functions such that $\mathrm{dom}\, h \subset \mathrm{dom}\, f$; and 2) the convex-concave nonsmooth composite saddle-point problem (SPP)

$$\min_{x\in\mathbb{R}^n} \max_{y\in\mathbb{R}^m} \left\{\phi(x, y) := f(x, y) + h_1(x) - h_2(y)\right\}, \tag{2}$$

where $f(x, y)$ is convex in $x$ and concave in $y$, and $h_1 : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and $h_2 : \mathbb{R}^m \to \mathbb{R}\cup\{+\infty\}$ are proper lower semi-continuous convex functions such that $\mathrm{dom}\, h_1 \times \mathrm{dom}\, h_2 \subset \mathrm{dom}\, f$. The main goal of this paper is to study the primal-dual convergence and iteration-complexity of proximal bundle (PB) methods for solving CNCO and SPP.

Classical PB methods, first proposed in [13, 28] and further developed in [14, 20], are known to be efficient algorithms for solving CNCO problems. At the core of classical PB methods is the introduction of a proximal regularization term to the standard cutting-plane method (or Kelly's method) and a sufficient descent test. Those methods update the prox center (i.e., perform a serious step) if there is a sufficient descent in the function value; otherwise, they keep the prox center and refine the cutting-plane model (i.e., perform a null step). Various bundle management policies (i.e., update schemes on cutting-plane models) have been discussed in [7, 9, 12, 23, 24, 27]. The textbooks [24, 25] provide a comprehensive discussion of the convergence analysis of classical PB methods for CNCO problems. Iteration-complexity bounds have been established in [1, 6, 7, 12] for classical PB methods for solving CNCO problems (1) with $h \equiv 0$ or being the indicator function of a nonempty closed convex set. Notably, the first complexity of classical PB methods is given by [12] as $\mathcal{O}(\bar\varepsilon^{-3})$ to find a $\bar\varepsilon$-solution of (1) (i.e., a point $\bar x \in \mathrm{dom}\, h$ satisfying $\phi(\bar x) - \phi_* \le \bar\varepsilon$).

Since the lower complexity bound of CNCO is $\Omega(\bar\varepsilon^{-2})$ (see for example Subsection 7.1 of [16]), it is clear that the bound $\mathcal{O}(\bar\varepsilon^{-3})$ given by [12] is not optimal. Recent papers [16,17] establish the optimal complexity bound $\mathcal{O}(\bar\varepsilon^{-2})$ for a large range of prox stepsizes by developing modern PB methods, where the sufficient descent test in classical PB methods is replaced by a different serious/null decision condition motivated by the proximal point method (PPM) (see Subsection 3.1 of [16] and Subsection 3.2 of [17]). Moreover, [17] studies the cutting-plane (i.e., multi-cuts) model, the cut-aggregation (i.e., two-cuts) model, and a newly proposed one-cut model under a generic bundle update scheme, and provides a unified analysis for all models encompassed within this general update scheme.

This paper investigates the modern PB methods for solving CNCO problems from the primal-dual perspective. More specifically, it shows that a cycle (consecutive null steps between two serious steps) of the methods indeed finds an approximate primal-dual solution to a proximal subproblem, and further establishes the iteration-complexity of the modern PB methods in terms of a primal-dual gap of (1), which is a stronger convergence guarantee than the $\bar\varepsilon$-solution considered in [16,17]. Furthermore, the paper reveals an interesting dual relationship between the conditional gradient (CG) method and the cutting-plane scheme for solving proximal subproblems within PB. Extending upon this duality, the paper also develops novel variants of both CG and the cutting-plane scheme, drawing inspiration from both perspectives of the dual relationship.

An independent study conducted concurrently by [8] examines the same duality under a more specialized assumption that $f$ is piece-wise linear and $h$ is smooth. Building upon the duality and using the convergence analysis of CG, [8] is able to improve the general complexity bound $\mathcal{O}(\bar{\varepsilon}^{-2})$ to $\mathcal{O}(\bar{\varepsilon}^{-4/5})$ in this context. The duality relationship between the subgradient method/mirror descent and CG is first studied in [3]. Related works [2,5,19,30] investigate the duality between Kelly's method/simplicial method and CG across various settings, and also examine the primal and dual simplicial methods.

The second half of the paper is devoted to developing modern PB methods for solving convex-concave nonsmooth composite SPP. While subgradient-type methods have been extensively studied for solving such SPP, for example, [10, 18, 21, 22, 26, 29], PB methods, which generalize subgradient methods by better using the history of subgradients, have received less attention in this context. Inspired by the PPM interpretation of modern PB methods, this paper proposes a generic inexact proximal point framework (IPPF) to solve SPP (2), comprising both a composite subgradient method and a PB method as special instances. The paper finally establishes the iteration-complexity bounds for both methods to find an approximate saddle-point of (2).

**Organization of the paper.** Subsection 1.1 presents basic definitions and notation used throughout the paper. Section 2 describes the primal-dual proximal bundle (PDPB) method and the assumptions on CNCO, and establishes the iteration-complexity of PDPB in terms of a primal-dual gap. In addition, Subsection 2.1 presents the key subroutine, namely a primal-dual cutting-plane (PDCP) scheme, used within PDPB for solving a proximal subproblem and provides the primal-dual convergence analysis of PDCP. Section 3 explores the duality between PDCP and CG by demonstrating that PDCP applied to the proximal subproblem produces the same iterates as CG applied to the dual problem. Subsection 3.1 presents an alternative primal-dual convergence analysis of PDCP using CG duality. Moreover, inspired by the duality, Subsections 3.2 and 3.3 develop novel PDCP and CG variants, respectively. Section 4 extends PB to solving the convex-concave nonsmooth composite SPP. More specifically, Subsection 4.1 introduces the IPPF for SPP and Subsection 4.2 describes the PB method for SPP (PB-SPP) and establishes its iteration-complexity to find an approximate saddle-point. Section 5 presents some concluding remarks and possible extensions. Appendix A provides a few useful technical results and deferred proofs. Appendices B and C are devoted to the complexity analyses of subgradient methods for solving CNCO (1) and SPP (2), respectively.

### 1.1 Basic definitions and notation

Let $\mathbb{R}$ denote the set of real numbers. Let $\mathbb{R}_{++}$ denote 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.

For given $f : \mathbb{R}^n \to (-\infty, +\infty]$, let $\operatorname{dom} f := \{x \in \mathbb{R}^n : f(x) < \infty\}$ denote the effective domain of $f$. We say $f$ is proper if $\operatorname{dom} f \neq \emptyset$. A proper function $f : \mathbb{R}^n \to (-\infty, +\infty]$ is

$\mu$-strongly convex for some $\mu > 0$ if for every $x, y \in \operatorname{dom} f$ and $t \in [0,1]$

$$f(tx + (1-t)y) \le tf(x) + (1-t)f(y) - \frac{t(1-t)\mu}{2}\|x-y\|^2.$$

Let $\overline{\operatorname{Conv}}\,(\mathbb{R}^n)$ denote the set of all proper lower-semicontinuous convex functions. For $\varepsilon \ge 0$, the *$\varepsilon$-subdifferential* of $f$ at $x \in \operatorname{dom} f$ is denoted by

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

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

$$\ell_f(\cdot; x) := f(x) + \langle f'(x), \cdot - x\rangle. \quad (4)$$

## 2 Primal-dual proximal bundle method for CNCO

In this section, we consider the CNCO problem (1). More specifically, we assume the following conditions hold:

(A1) a subgradient oracle, i.e., a function $f' : \operatorname{dom} h \to \mathbb{R}^n$ satisfying $f'(x) \in \partial f(x)$ for every $x \in \operatorname{dom} h$, is available;

(A2) $\|f'(x)\| \le M$ for every $x \in \operatorname{dom} h$ and some $M > 0$;

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

Define the linearization of $f$ at $x \in \operatorname{dom} h$, $\ell_f : \operatorname{dom} h \to \mathbb{R}$ as

$$\ell_f(\cdot; x) := f(x) + \langle f'(x), \cdot - x\rangle.$$

Clearly, it follows from (A2) that for every $x, y \in \operatorname{dom} h$,

$$f(x) - \ell_f(x; y) \le 2M\|x - y\|. \quad (5)$$

For a given initial point $x_0 \in \operatorname{dom} h$, we denote its distance to $X_*$ as

$$d_0 := \|x_0^* - x_0\|, \quad x_0^* := \underset{x_* \in X_*}{\operatorname{argmin}}\{\|x_* - x_0\|\}. \quad (6)$$

The primal-dual subgradient method denoted by $\text{PDS}(x_0, \lambda)$, where $x_0 \in \operatorname{dom} h$ is the initial point and $\lambda > 0$ is the prox stepsize, recursively computes

$$s_k = f'(x_{k-1}) \in \partial f(x_{k-1}), \quad x_k = \underset{u \in \mathbb{R}^n}{\operatorname{argmin}}\left\{\ell_f(u; x_{k-1}) + h(u) + \frac{1}{2\lambda}\|u - x_{k-1}\|^2\right\}. \quad (7)$$

For given tolerance $\bar\varepsilon > 0$, letting $\lambda = \bar\varepsilon/(16M^2)$, then the iteration-complexity for $\text{PDS}(x_0, \lambda)$ to generate a primal-dual pair such that the primal-dual gap of a constrained version of (1) is bounded by $\bar\varepsilon$ is $\mathcal{O}(M^2 d_0^2/\bar\varepsilon^2)$ (see Theorem B.2).

### 2.1 Primal-dual cutting-plane scheme

The PDPB method solves a sequence of proximal subproblems of the form

$$\min_{u\in\mathbb{R}^n}\left\{\phi^\lambda(u):=\phi(u)+\frac{1}{2\lambda}\|u-x_{k-1}\|^2\right\},\tag{8}$$

where $\lambda$ is the prox stepsize and $x_{k-1}$ is the prox center in the $k$-th proximal subproblem (or cycle). We omit the index $k$ in $\phi^\lambda$ when it is clear from the context. Each proximal subproblem invokes the PDCP scheme to find an approximate solution. Hence, PDPB can be viewed as a generalization of PDS, which only takes one proximal subgradient step (i.e., (7)) to solve every proximal subproblem (8). The goal of this subsection is to describe the key subroutine PDCP for solving (8) and present its primal-dual convergence analysis.

At the $j$-th iteration (within a cycle of PDPB), given some prox stepsize $\lambda>0$ and prox center $x_0$, PDCP computes a primal-dual pair $(x_j,s_j)$ as follows

$$x_j=\underset{u\in\mathbb{R}^n}{\operatorname{argmin}}\left\{\Gamma_j(u)+h(u)+\frac{1}{2\lambda}\|u-x_0\|^2\right\},\quad s_j\in\partial\Gamma_j(x_j),\tag{9}$$

where $\Gamma_j$ is a proper, closed and convex function satisfying $\Gamma_j\le f$ for every $j\ge 1$. Starting from $\Gamma_1(\cdot)=\ell_f(\cdot;x_0)$, and for every $j\ge 1$, $\Gamma_{j+1}$ is obtained from the following generic bundle management (GBM), which is motivated by BU given in Subsection 3.1 of [17]. It is easy to verify that one-cut, two-cuts, and multiple-cuts schemes (i.e., (E1)-(E3)) described in Subsection 3.1 of [17] all satisfy GBM.

**Algorithm 1** Generic Bundle Management, GBM$(\lambda,\tau_j,x_0,x_j,\Gamma_j)$

**Initialize:** $(\lambda,\tau_j)\in\mathbb{R}_{++}\times[0,1]$, $(x_0,x_j)\in\mathbb{R}^n\times\mathbb{R}^n$, and $\Gamma_j\in\overline{\text{Conv}}\,(\mathbb{R}^n)$ satisfying $\Gamma_j\le f$

- find a bundle model $\Gamma_{j+1}\in\overline{\text{Conv}}\,(\mathbb{R}^n)$ satisfying $\Gamma_j\le f$ and

$$\Gamma_{j+1}(\cdot)\ge\tau_j\bar\Gamma_j(\cdot)+(1-\tau_j)\ell_f(\cdot;x_j),\tag{10}$$

where $\bar\Gamma_j\in\overline{\text{Conv}}\,(\mathbb{R}^n)$ satisfies $\bar\Gamma_j\le f$ and

$$\bar\Gamma_j(x_j)=\Gamma_j(x_j),\quad x_j=\underset{u\in\mathbb{R}^n}{\operatorname{argmin}}\left\{\bar\Gamma_j(u)+h(u)+\frac{1}{2\lambda}\|u-x_0\|^2\right\}.\tag{11}$$

**Output:** $\Gamma_{j+1}$.

PDCP computes an auxiliary sequence $\{\tilde x_j\}$ to determine termination. It sets $\tilde x_1=x_1$, and for $j\ge 1$, it chooses $\tilde x_{j+1}$ such that

$$\phi^\lambda(\tilde x_{j+1})\le\tau_j\phi^\lambda(\tilde x_j)+(1-\tau_j)\phi^\lambda(x_{j+1})\tag{12}$$

where $\phi^\lambda$ is as in (8). PDCP also computes

$$m_j = \min_{u\in\mathbb{R}^n} \left\{\Gamma_j(u) + h(u) + \frac{1}{2\lambda}\|u - x_0\|^2\right\}, \quad t_j = \phi^\lambda(\tilde{x}_j) - m_j. \tag{13}$$

For given tolerance $\varepsilon > 0$, PDCP terminates the current cycle when $t_j \le \varepsilon$.

PDCP is formally stated below.

---

**Algorithm 2** Primal-Dual Cutting-Plane, PDCP$(x_0, \lambda, \varepsilon)$

---

**Initialize:** given $x_0 \in \operatorname{dom} h$, $\lambda > 0$, $\varepsilon > 0$, set $t_0 = 2\varepsilon$, $\Gamma_1(\cdot) = \ell_f(\cdot; x_0)$, and $j = 1$;

**while** $t_{j-1} > \varepsilon$ **do**

**1.** compute $(x_j, s_j)$ by (9), choose $\tilde{x}_j$ as in (12), and set $t_j$ as in (13);

**2.** select $\tau_j \in [0, 1]$ and update $\Gamma_{j+1}$ by GBM$(\lambda, \tau_j, x_0, x_j, \Gamma_j)$ and $j \leftarrow j + 1$;

**end while**

**Output:** $(x_{j-1}, \tilde{x}_{j-1}, s_{j-1})$.

---

The auxiliary iterate $\tilde{x}_j$ vaguely given in (12) can be explicitly computed by either of the following two formulas:

$$\tilde{x}_{j+1} = \tau_j \tilde{x}_j + (1 - \tau_j)x_{j+1}, \quad \forall j \ge 1,$$

and

$$\tilde{x}_j \in \operatorname{Argmin}\{\phi^\lambda(u) : u \in \{x_1, \ldots, x_j\}\}, \quad \forall j \ge 1.$$

Clearly, $\{\tilde{x}_j\}$ obtained from the second formula above satisfies (12) with any $\tau_j \in [0, 1]$.

The following result proves that $t_j$ is an upper bound on the primal-dual gap for (8) and hence shows that $(\tilde{x}_j, s_j)$ an approximate primal-dual solution pair for (8).

**Lemma 2.1.** *Define $h^\lambda(\cdot) := h(\cdot) + \|\cdot - x_0\|^2/(2\lambda)$. Then, we have for every $j \ge 1$,*

$$\phi^\lambda(\tilde{x}_j) + f^*(s_j) + (h^\lambda)^*(-s_j) \le t_j. \tag{14}$$

**Proof**: It follows from (9) that $s_j \in \partial\Gamma_j(x_j)$ and $-s_j \in \partial h^\lambda(x_j)$. Using Theorem 4.20 of [4], we have

$$\Gamma_j^*(s_j) = -\Gamma_j(x_j) + \langle s_j, x_j\rangle, \quad (h^\lambda)^*(-s_j) = -h^\lambda(x_j) - \langle s_j, x_j\rangle.$$

Combining the above identities and using the definition of $m_j$ in (13), we have

$$-m_j = \Gamma_j^*(s_j) + (h^\lambda)^*(-s_j).$$

It clearly from $\Gamma_j \le f$ that $\Gamma_j^* \ge f^*$. This observation and the above inequality imply that

$$\phi^\lambda(\tilde{x}_j) + f^*(s_j) + (h^\lambda)^*(-s_j) \le \phi^\lambda(\tilde{x}_j) - m_j.$$

Hence, (14) immediately follows from the definition of $t_j$ in (13). Finally, we note that $-f^*(s)-(h^\lambda)^*(-s)$ is the Lagrange dual function of $\phi^\lambda(x)$ in (8). Therefore, the left-hand side of (14) is the primal-dual gap for (8). ■

With regard to Lemma 2.1, it suffices to show the convergence of $t_j$ to develop the primal-dual convergence analysis of PDCP. We begin this analysis by providing some basic properties of GBM. The following lemma is adapted from Lemma 4.4 of [17], and hence we omit the proof.

**Lemma 2.2.** *For every $j \ge 1$, there exists $\bar\Gamma_j \in \overline{\mathrm{Conv}}\,(\mathbb{R}^n)$ such that for every $u \in \mathbb{R}^n$,*

$$\bar\Gamma_j(u) + h^\lambda(u) \ge m_j + \frac{1}{2\lambda}\|u - x_j\|^2. \tag{15}$$

Following Lemma 2.2, we are able to present the convergence rate of $t_j$ under the assumption that $\tau_j = j/(j+2)$ for every $j \ge 1$. The following proposition resembles Lemma 4.6 in [17].

**Proposition 2.3.** *Considering Algorithm 2 with $\tau_j = j/(j+2)$, then for every $j \ge 1$, we have*

$$t_j \le \frac{2t_1}{j(j+1)} + \frac{16\lambda M^2}{j+1}, \tag{16}$$

*where $t_j$ is as in (13). Moreover, the number of iterations for PDCP to obtain $t_j \le \varepsilon$ is at most*

$$\mathcal{O}\left(\frac{\sqrt{t_1}}{\sqrt{\varepsilon}} + \frac{\lambda M^2}{\varepsilon} + 1\right).$$

**Proof**: We first note that for every $j \ge 1$, $\tau_j = A_j/A_{j+1}$ where $A_{j+1} = A_j + j + 1$ and $A_0 = 0$, i.e., $A_j = j(j+1)/2$ for every $j \ge 0$. It follows from this observation, the definition of $m_j$ in (13), and relation (10) that

$$\begin{aligned} A_{j+1}m_{j+1} &\overset{(13)}{=} A_{j+1}(\Gamma_{j+1} + h^\lambda)(x_{j+1}) \\ &\overset{(10)}{\ge} A_j\left[(\bar\Gamma_j + h^\lambda)(x_{j+1})\right] + (j+1)\left[\ell_f(x_{j+1};x_j) + h^\lambda(x_{j+1})\right]. \end{aligned}$$

Using Lemma 2.2(a) and (5), we have

$$\begin{aligned} A_{j+1}m_{j+1} &\overset{(15)}{\ge} A_j\left[m_j + \frac{1}{2\lambda}\|x_{j+1}-x_j\|^2\right] + (j+1)\left[\ell_f(x_{j+1};x_j) + h^\lambda(x_{j+1})\right] \\ &= A_j m_j + (j+1)\left[\ell_f(x_{j+1};x_j) + h^\lambda(x_{j+1}) + \frac{A_j}{2\lambda(j+1)}\|x_{j+1}-x_j\|^2\right] \\ &\overset{(5)}{\ge} A_j m_j + (j+1)\left[\phi^\lambda(x_{j+1}) - 2M\|x_{j+1}-x_j\| + \frac{A_j}{2\lambda(j+1)}\|x_{j+1}-x_j\|^2\right] \\ &\ge A_j m_j + (j+1)\phi^\lambda(x_{j+1}) - \frac{2\lambda M^2(j+1)^2}{A_j} \end{aligned}$$

where the last inequality is due to the Young's inequality $a^2 + b^2 \ge 2ab$. It follows from the fact that $A_j = j(j+1)/2$ that for every $j \ge 1$,

$$A_{j+1} m_{j+1} \ge A_j m_j + (j+1)\phi^\lambda(x_{j+1}) - 8\lambda M^2.$$

Summing the above inequality from $j = 1$ to $j - 1$, and using the definition of $t_j$ in (13) and the fact that $\tilde{x}_1 = x_1$, we obtain

$$\begin{aligned} A_j m_j &\ge A_1 m_1 + 2\phi^\lambda(x_2) + \cdots + j\phi^\lambda(x_j) - 8\lambda M^2(j-1) \\ &\overset{(13)}{=} -A_1 t_1 + A_1 \phi^\lambda(x_1) + 2\phi^\lambda(x_2) + \cdots + j\phi^\lambda(x_j) - 8\lambda M^2(j-1) \\ &\overset{(12)}{\ge} -A_1 t_1 + A_j \phi^\lambda(\tilde{x}_j) - 8\lambda M^2(j-1), \end{aligned}$$

where the last inequality follows from (12) and the fact that $A_j = A_{j-1} + j$. Rearranging the terms and using the definition of $t_j$ in (13) again, we have

$$A_j t_j \le A_1 t_1 + 8\lambda M^2(j-1). \tag{17}$$

Hence, (16) follows from the fact that $A_j = j(j+1)/2$. Finally, the complexity bound immediately follows from (16). ∎

### 2.2 Primal-dual proximal bundle method

Recall the definitions of $d_0$ and $x_0^*$ in (6). Since $x_0^* \in B(x_0, 6d_0)$, which is the ball centered at $x_0$ and with radius $4d_0$, it is easy to see that to solve (1), it suffices to solve

$$\min\left\{\hat{\phi}(x) := f(x) + \hat{h}(x) : x \in \mathbb{R}^n\right\} = \min\left\{\phi(x) : x \in Q\right\}, \tag{18}$$

where $\hat{h} = h + I_Q$ and $I_Q$ is the indicator function of $Q = B(x_0, 6d_0)$.

In what follows, we present the PDPB and establish the complexity for obtaining a primal-dual solution pair of (18). The PDPB is formally stated below.

---

**Algorithm 3** Primal-Dual Proximal Bundle, PDPB$(x_0, \lambda, \bar{\varepsilon})$

---

**Initialize:** given $(x_0, \lambda, \bar{\varepsilon}) \in \operatorname{dom} h \times \mathbb{R}_{++} \times \mathbb{R}_{++}$

**for** $k = 1, 2, \cdots$ **do**

- call oracle $(x_k, \tilde{x}_k, s_k) = \text{PDCP}(x_{k-1}, \lambda, \bar{\varepsilon})$ and compute

$$\bar{x}_k = \frac{1}{k}\sum_{i=1}^{k} \tilde{x}_i, \quad \bar{s}_k = \frac{1}{k}\sum_{i=1}^{k} s_i. \tag{19}$$

**end for**

---

In the $k$-th iteration of PDPB, we are approximately solving the proximal subproblem

$$\min_{u\in\mathbb{R}^n}\left\{\phi^\lambda(u) := \phi(u) + \frac{1}{2\lambda}\|u - x_{k-1}\|^2\right\}. \tag{20}$$

Here we abuse the notation $\phi^\lambda$ (also see (8)), while it should be clear from the context. Recall from Subsection 2.1 that (20) is approximately solved by invoking PDCP. The (global) iteration indices in PDCP are regarded as the $k$-th cycle, denoted by $\mathcal{C}_k = \{i_k, \ldots, j_k\}$, where $j_k$ is the last iteration index of the $k$-th call to PDCP, $j_0 = 0$, and $i_k = j_{k-1} + 1$. Hence, for the $j_k$-th iteration of PDCP, we have

$$x_k = x_{j_k}, \quad \tilde{x}_k = \tilde{x}_{j_k}, \quad s_k = s_{j_k}, \quad \Gamma_k = \Gamma_{j_k}, \quad m_k = m_{j_k}, \tag{21}$$

and (9) becomes

$$x_k = \underset{u\in\mathbb{R}^n}{\operatorname{argmin}}\left\{\Gamma_k(u) + h(u) + \frac{1}{2\lambda}\|u - x_{k-1}\|^2\right\}, \quad s_k \in \partial\Gamma_k(x_k). \tag{22}$$

The following lemma provides basic properties of PDPB and is the starting point of the the complexity analysis of PDPB.

**Lemma 2.4.** *The following statements hold for every $k \geq 1$:*

*(a)* $\Gamma_k \leq f$ *and* $f^* \leq \Gamma_k^*$*;*

*(b)* $s_k \in \partial\Gamma_k(x_k)$ *and* $g_k \in \partial h(x_k)$ *where* $g_k = -s_k + (x_{k-1} - x_k)/\lambda$*;*

*(c)* $\phi^\lambda(\tilde{x}_k) \leq \bar{\varepsilon} + m_k = \bar{\varepsilon} + (\Gamma_k + h)(x_k) + \|x_k - x_{k-1}\|^2/(2\lambda)$.

**Proof**: (a) It follows from the facts that $\Gamma_j \leq f$ for every $j \geq 1$ and $\Gamma_k = \Gamma_{j_k}$ that the first inequality holds. The second one immediately follows from the first one and the definition of the conjugate function.

(b) This statement follows from (22) and the definitions in (21).

(c) This statement follows from the termination criterion of the $k$-th cycle, that is, $t_{j_k} \leq \bar{\varepsilon}$, and the definitions in (13) and (21). ∎

The following proposition is a key component of our complexity analysis, as it establishes an important primal-dual gap for (1).

**Proposition 2.5.** *For every $k \geq 1$, we have*

$$\phi(\bar{x}_k) + f^*(\bar{s}_k) + \hat{h}^*(-\bar{s}_k) \leq \bar{\varepsilon} + \frac{18 d_0^2}{\lambda k}. \tag{23}$$

*where $\bar{x}_k$ and $\bar{s}_k$ are as in (19).*

**Proof**: It follows from Lemma 2.4(b) and Theorem 4.20 of [4] that for every $k \ge 1$,

$$\Gamma_k(x_k) + \Gamma_k^*(s_k) = \langle x_k, s_k \rangle, \quad h(x_k) + h^*(g_k) = \langle x_k, g_k \rangle.$$

Summing the above two equations yields

$$(\Gamma_k + h)(x_k) + \Gamma_k^*(s_k) + h^*(g_k) = \frac{1}{\lambda}\langle x_k, x_{k-1} - x_k \rangle. \tag{24}$$

Using the above identity and Lemma 2.4(a) and (c), we have for every $k \ge 1$,

$$\begin{aligned} &\phi(\tilde{x}_k) + f^*(s_k) + h^*(g_k) \le \phi(\tilde{x}_k) + \Gamma_k^*(s_k) + h^*(g_k) \\ \le& \bar{\varepsilon} + (\Gamma_k + h)(x_k) + \frac{1}{2\lambda}\|x_k - x_{k-1}\|^2 + \Gamma_k^*(s_k) + h^*(g_k) \\ \overset{(24)}{=}& \bar{\varepsilon} + \frac{1}{2\lambda}(\|x_{k-1}\|^2 - \|x_k\|^2). \end{aligned}$$

Summing the above inequality from $k = 1$ to $k$, and using convexity and the definitions in (19), we obtain

$$\phi(\bar{x}_k) + f^*(\bar{s}_k) + h^*(\bar{g}_k) \le \bar{\varepsilon} + \frac{1}{2\lambda k}\left(\|x_0\|^2 - \|x_k\|^2\right), \tag{25}$$

where $\bar{g}_k = (\sum_{i=1}^k g_i)/k$. Define

$$\eta_k(u) = \frac{1}{2\lambda k}\|u - x_0\|^2, \quad \hat{\eta}_k(u) = \eta_k(u) - I_Q(u). \tag{26}$$

Noting that $\nabla \eta_k(x_k) = (x_k - x_0)/(\lambda k) = -\bar{g}_k - \bar{s}_k$, and hence it follows from Theorem 4.20 of [4] that

$$\eta_k^*(-\bar{g}_k - \bar{s}_k) = \frac{1}{\lambda k}\langle x_k - x_0, x_k \rangle - \eta_k(x_k) = \frac{1}{2\lambda k}\left(\|x_k\|^2 - \|x_0\|^2\right).$$

The above observation and (25) together imply that

$$\phi(\bar{x}_k) + f^*(\bar{s}_k) + h^*(\bar{g}_k) + \eta_k^*(-\bar{g}_k - \bar{s}_k) \le \bar{\varepsilon}. \tag{27}$$

It follows from Theorem 4.17 of [4] that

$$(h + \eta_k)^*(-\bar{s}_k) = (h^* \Box \eta_k^*)(-\bar{s}_k) = \min_{u \in \mathbb{R}^n}\{h^*(u) + \eta_k^*(-\bar{s}_k - u)\} \le h^*(\bar{g}_k) + \eta_k^*(-\bar{g}_k - \bar{s}_k).$$

Noting from (26) that $\hat{h} = h + \eta_k - \hat{\eta}_k$ and applying Theorem 4.17 of [4] again, we obtain

$$\begin{aligned} \hat{h}^*(-\bar{s}_k) &= [(h + \eta_k)^* \Box (-\hat{\eta}_k)^*](-\bar{s}_k) = \min_{u \in \mathbb{R}^n}\{(h + \eta_k)^*(u) + (-\hat{\eta}_k)^*(-\bar{s}_k - u)\} \\ &\le (h + \eta_k)^*(-\bar{s}_k) + (-\hat{\eta}_k)^*(0). \end{aligned}$$

Summing the above two inequalities, we have

$$\hat{h}^*(-\bar{s}_k) \le h^*(\bar{g}_k) + \eta_k^*(-\bar{g}_k - \bar{s}_k) + (-\hat{\eta}_k)^*(0),$$

which together with (27) implies that

$$\phi(\bar{x}_k) + f^*(\bar{s}_k) + \hat{h}^*(-\bar{s}_k) \le \bar{\varepsilon} + (-\hat{\eta}_k)^*(0).$$

It follows from (26) that

$$(-\hat{\eta}_k)^*(0) = \max_{u\in\mathbb{R}^n}\left\{\langle 0,u\rangle - \left(-\frac{\|u-x_0\|^2}{2\lambda k} + I_Q(u)\right)\right\} = \frac{\max_{u\in Q}\|u-x_0\|^2}{2\lambda k} = \frac{18d_0^2}{\lambda k},$$

where the last identity follows from the fact that $Q = B(x_0, 6d_0)$. Therefore, (23) holds in view of the above two relations. ∎

The next lemma is a technical result showing that $x_k \in Q$ and $\tilde{x}_k \in Q$ under mild conditions, where $Q = B(x_0, 6d_0)$.

**Lemma 2.6.** *Given $(x_0, \bar{\varepsilon}) \in \mathbb{R}^n \times \mathbb{R}_{++}$, if $\lambda \le 2d_0^2/\bar{\varepsilon}$ and $k \le 2d_0^2/(\lambda\bar{\varepsilon})$, then the sequences $\{x_k\}$ and $\{\tilde{x}_k\}$ generated by PDPB$(x_0, \lambda, \bar{\varepsilon})$ satisfy*

$$x_k \in Q, \quad \tilde{x}_k \in Q. \tag{28}$$

**Proof**: Noticing that the objective function in (22) is $\lambda^{-1}$-strongly convex, it thus follows from Theorem 5.25(b) of [4] that for every $u \in \operatorname{dom} h$,

$$m_k + \frac{1}{2\lambda}\|u - x_k\|^2 \le \Gamma_k(u) + h(u) + \frac{1}{2\lambda}\|u - x_{k-1}\|^2 \le \phi(u) + \frac{1}{2\lambda}\|u - x_{k-1}\|^2, \tag{29}$$

where the second inequality follows from the first one in Lemma 2.4(a). Taking $u = x_0^*$ in (29), we have

$$m_k + \frac{1}{2\lambda}\|x_k - x_0^*\|^2 \le \phi_* + \frac{1}{2\lambda}\|x_{k-1} - x_0^*\|^2.$$

This inequality and Lemma 2.4(c) then imply that

$$\begin{aligned}\frac{1}{2\lambda}\|x_k - x_0^*\|^2 &\le \phi(\tilde{x}_k) - \phi_* + \frac{1}{2\lambda}\|x_k - x_0^*\|^2 \\ &\le \phi(\tilde{x}_k) - m_k + \frac{1}{2\lambda}\|x_{k-1} - x_0^*\|^2 \le \bar{\varepsilon} + \frac{1}{2\lambda}\|x_{k-1} - x_0^*\|^2.\end{aligned}$$

Summing the above inequality from $k = 1$ to $k$, we have

$$\|x_k - x_0^*\|^2 \le \|x_0 - x_0^*\|^2 + 2k\lambda\bar{\varepsilon}.$$

Using the fact that $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for $a, b \ge 0$ and the assumption that $k \le 2d_0^2/(\lambda\bar{\varepsilon})$, we further obtain

$$\|x_k - x_0^*\| \le d_0 + \sqrt{2k\lambda\bar{\varepsilon}} \le 3d_0. \tag{30}$$

Taking $u = \tilde{x}_k$ in (29) and using Lemma 2.4(c), we have

$$\frac{1}{2\lambda}\|\tilde{x}_k - x_k\|^2 \le \phi(\tilde{x}_k) + \frac{1}{2\lambda}\|\tilde{x}_k - x_{k-1}\|^2 - m_k \le \bar{\varepsilon}.$$

Under the assumption that $\lambda \le 2d_0^2/\bar{\varepsilon}$, using (30), the above inequality, and the triangle inequality, we have

$$\begin{aligned}
\|x_k - x_0\| &\le \|x_k - x_0^*\| + \|x_0^* - x_0\| \overset{(30)}{\le} 4d_0, \\
\|\tilde{x}_k - x_0\| &\le \|x_k - x_0\| + \|\tilde{x}_k - x_k\| \le 4d_0 + \sqrt{2\lambda\bar{\varepsilon}} \le 6d_0.
\end{aligned}$$

Hence, (28) follows immediately. ■

Now we are ready to present the number of oracle calls to PDCP in PDPB (i.e., Algorithm 3).

**Proposition 2.7.** *Given $(x_0, \bar{\varepsilon}) \in \mathbb{R}^n \times \mathbb{R}_{++}$, if $\lambda \le 2d_0^2/\bar{\varepsilon}$, then the number of iterations for PDPB$(x_0, \lambda, \bar{\varepsilon})$ to generate $(\bar{x}_k, \bar{s}_k)$ satisfying*

$$\hat{\phi}(\bar{x}_k) + f^*(\bar{s}_k) + \hat{h}^*(-\bar{s}_k) \le 10\bar{\varepsilon} \tag{31}$$

*is at most $2d_0^2/(\lambda\bar{\varepsilon})$.*

**Proof**: Since $Q$ is a convex set, it follows from the definition of $\bar{x}_k$ in (19) and Lemma 2.6 that $\bar{x}_k \in Q$ for every $k \le 2d_0^2/(\lambda\bar{\varepsilon})$. This observation and the fact that $\hat{h} = h + I_Q$ imply that $\hat{h}(\bar{x}_k) = h(\bar{x}_k)$. Hence, using Proposition 2.5, we have for every $k \le 2d_0^2/(\lambda\bar{\varepsilon})$,

$$\hat{\phi}(\bar{x}_k) + f^*(\bar{s}_k) + \hat{h}^*(-\bar{s}_k) \le \bar{\varepsilon} + \frac{18d_0^2}{\lambda k}.$$

Therefore, the conclusion of the proposition follows immediately. ■

The following lemma is a technical result providing a universal bound on the first gap $t_{i_k}$ for each cycle $\mathcal{C}_k$.

**Lemma 2.8.** *For $k \le 2d_0^2/(\lambda\bar{\varepsilon})$, we have*

$$t_{i_k} \le \bar{t} := 4M(3d_0 + \lambda M), \tag{32}$$

*where $i_k$ is the first iteration index in the cycle $\mathcal{C}_k$.*

**Proof**: Using (5), definitions of $m_j$ and $t_j$ in (13), and the facts that $\tilde{x}_{i_k} = x_{i_k}$ and $\Gamma_{i_k} = \ell_f(\cdot; x_{k-1})$, we have

$$\begin{aligned}
t_{i_k} &\overset{(13)}{=} \phi^\lambda(\tilde{x}_{i_k}) - m_{i_k} = \phi^\lambda(x_{i_k}) - m_{i_k} \\
&\overset{(13)}{=} f(x_{i_k}) - \ell_f(x_{i_k}; x_{k-1}) \overset{(5)}{\le} 2M\|x_{i_k} - x_{k-1}\|.
\end{aligned} \tag{33}$$

In view of (9) and the fact that $\Gamma_{i_k} = \ell_f(\cdot; x_{k-1})$, we know the first iteration of PDCP is the same as PDS$(x_0, \lambda)$ (see (7)). Hence, following an argument similar to the proof of Lemma B.1, we can prove for every $u \in \mathrm{dom}\, h$,

$$\phi(x_{i_k}) - \ell_f(u; x_{k-1}) - h(u) \stackrel{(101)}{\leq} 2\lambda M^2 + \frac{1}{2\lambda}\|u - x_{k-1}\|^2 - \frac{1}{2\lambda}\|u - x_{i_k}\|^2.$$

It follows from the above inequality with $u = x_0^*$ and the convexity of $f$ that

$$\begin{aligned} 0 \leq \phi(x_{i_k}) - \phi_* &\leq \phi(x_{i_k}) - \ell_f(x_0^*; x_{k-1}) - h(x_0^*) \\ &\leq 2\lambda M^2 + \frac{1}{2\lambda}\|x_0^* - x_{k-1}\|^2 - \frac{1}{2\lambda}\|x_0^* - x_{i_k}\|^2. \end{aligned}$$

Rearranging the terms and using the inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for any $a, b \geq 0$, we have

$$\|x_0^* - x_{i_k}\| \leq \|x_0^* - x_{k-1}\| + 2\lambda M.$$

This inequality and the triangle inequality then imply that

$$\|x_{i_k} - x_{k-1}\| \leq \|x_{i_k} - x_0^*\| + \|x_0^* - x_{k-1}\| \leq 2\|x_{k-1} - x_0^*\| + 2\lambda M.$$

Recall from the proof of Lemma 2.6 that (30) gives $\|x_k - x_0^*\| \leq 3d_0$ for $k \leq 2d_0^2/(\lambda\bar{\varepsilon})$. Hence, we have

$$\|x_{i_k} - x_{k-1}\| \leq 2(3d_0 + \lambda M).$$

Therefore, (32) follows from (33) and the above inequality. ■

Finally, we are ready to establish the total iteration-complexity of PDPB.

**Theorem 2.1.** *Given $(x_0, \bar{\varepsilon}) \in \mathbb{R}^n \times \mathbb{R}_{++}$, assuming that $\lambda$ satisfies*

$$\frac{\sqrt{\bar{\varepsilon}} d_0}{M^{3/2}} \leq \lambda \leq \frac{2d_0^2}{\bar{\varepsilon}}, \tag{34}$$

*then the total iteration-complexity of PDPB$(x_0, \lambda, \bar{\varepsilon})$ to find $(\bar{x}_k, \bar{s}_k)$ satisfying (31) is*

$$\mathcal{O}\left(\frac{M^2 d_0^2}{\bar{\varepsilon}} + 1\right). \tag{35}$$

**Proof**: In view of Proposition 2.7, PDPB takes

$$\mathcal{O}\left(\frac{d_0^2}{\lambda\bar{\varepsilon}} + 1\right) \tag{36}$$

cycles to find $(\bar{x}_k, \bar{s}_k)$ satisfying (31). It follows from Proposition 2.3 and Lemma 2.8 that for every cycle in PDPB before termination, the number of iterations in the cycle is

$$\mathcal{O}\left(\frac{\sqrt{Md_0 + \lambda M^2}}{\sqrt{\bar{\varepsilon}}} + \frac{\lambda M^2}{\bar{\varepsilon}} + 1\right) = \mathcal{O}\left(\frac{\sqrt{Md_0}}{\sqrt{\bar{\varepsilon}}} + \frac{\lambda M^2}{\bar{\varepsilon}} + 1\right),$$

which together with the assumption that $\sqrt{\bar{\varepsilon} d_0}/M^{3/2} \le \lambda$ becomes

$$\mathcal{O}\left(\frac{\lambda M^2}{\bar{\varepsilon}} + 1\right). \tag{37}$$

Combining (36) and (37), and using (34), we conclude that (35) holds. ∎

# 3 Duality between PDCP and CG

The dual problem of the proximal subproblem (8) can be written as

$$\min_{z\in\mathbb{R}^n} \left\{\psi(z) := (h^\lambda)^*(-z) + f^*(z)\right\}, \tag{38}$$

where $-\psi$ is the dual function of $\phi^\lambda$ and $h^\lambda$ is as in Lemma 2.1. Since $h^\lambda$ is $\lambda^{-1}$-strongly convex, $(h^\lambda)^*$ is $\lambda$-smooth and one possible algorithm to solve (38) is the CG method.

We describe CG for solving (38) below.

---

**Algorithm 4** Conditional Gradient for (38), CG($z_1$)

---

**Initialize:** given $z_1 \in \mathrm{dom}\, f^*$

**for** $j = 1, 2, \cdots$ **do**

$$\bar{z}_j = \underset{z\in\mathbb{R}^n}{\mathrm{argmin}} \left\{\langle -\nabla (h^\lambda)^*(-z_j), z\rangle + f^*(z)\right\}, \tag{39}$$

$$z_{j+1} = \tau_j z_j + (1-\tau_j)\bar{z}_j. \tag{40}$$

**end for**

---

Motivated by the duality between the mirror descent/subgradient method and CG studied in [3], we prove the nice connection between CG (i.e., Algorithm 4) and PDCP (i.e., Algorithm 2) via duality. More specifically, we consider a specific implementation of GBM within PDCP, that is $\Gamma_j$ is updated as

$$\Gamma_{j+1}(\cdot) = \tau_j \Gamma_j(\cdot) + (1-\tau_j)\ell_f(\cdot; x_j). \tag{41}$$

Since $\Gamma_1(\cdot) = \ell_f(\cdot; x_0)$, $\Gamma_j$ is always affine and $s_j = \nabla \Gamma_j$ in view of (9).

The following result reveals the duality between PDCP with update scheme (41) and CG. Since the tolerance $\bar{\varepsilon}$ is not important in the discussion below, we will ignore it as input to PDCP. Assuming $\lambda$ in both PDCP and CG are the same, we only focus on the initial points of the two methods. Hence, we denote them by PDCP($x_0$) and CG($z_1$).

**Theorem 3.1.** *Given $x_0 \in \mathbb{R}^n$, $z_1 = f'(x_0)$, and the sequence $\{\tau_j\}$, then PDCP($x_0$) with update scheme (41) for solving (8) and CG($z_1$) for solving (38) have the following correspondence for every $j \ge 1$,*

$$s_j = z_j, \quad x_j = \nabla (h^\lambda)^*(-z_j), \quad f'(x_j) = \bar{z}_j. \tag{42}$$

**Proof**: We first show that the first relation in (42) implies the other two in (42). Using the definition of $x_j$ in (9), the fact that $s_j = \nabla\Gamma_j$, and the first relation in (42), we have $x_j$ from PDCP is equivalent to

$$x_j \overset{(9)}{=} \underset{x\in\mathbb{R}^n}{\operatorname{argmin}} \{\langle s_j, x\rangle + h^\lambda(x)\} \overset{(42)}{=} \underset{x\in\mathbb{R}^n}{\operatorname{argmax}} \{\langle -z_j, x\rangle - h^\lambda(x)\},$$

which implies that the second relation in (42) holds. The last one in (42) similarly follows from the second relation and (39).

We next prove the first relation in (42) by induction. For the case $j = 1$, it is easy to see from $\Gamma_1(\cdot) = \ell_f(\cdot; x_0)$ that

$$s_1 = \nabla\Gamma_1 = f'(x_0) = z_1.$$

Assume that the first relation in (42) holds for some $j \geq 1$. By the argument above, we know that the second and third relations in (42) also hold for $j$. Using the fact that $s_j = \nabla\Gamma_j$, the definition of $\Gamma_{j+1}$ in (41), and the last two relations in (42), we obtain

$$\begin{aligned} s_{j+1} = \nabla\Gamma_{j+1} &\overset{(41)}{=} \tau_j\nabla\Gamma_j + (1-\tau_j)f'(x_j) \\ &\overset{(42)}{=} \tau_j s_j + (1-\tau_j)\bar{z}_j \overset{(42)}{=} \tau_j z_j + (1-\tau_j)\bar{z}_j \overset{(40)}{=} z_{j+1}, \end{aligned}$$

where the last identity is due to (40). Hence, the first relation in (42) also holds for the case $j+1$. We thus complete the proof. ∎

### 3.1 Alternative primal-dual convergence analysis of PDCP

Theorem 3.1 demonstrates that PDCP and CG represent primal and dual perspectives for solving the equivalent problems (8) and (38), respectively. Recall that Proposition 2.3 establishes the primal-dual convergence rate of PDCP for solving (8), and hence it is worth studying the primal-dual convergence of CG for solving (38) as well. Thanks to the duality connection illustrated by Theorem 3.1, the convergence analysis of CG also serves as an alternative approach to study PDCP from the dual perspective.

Recall from (13.4) of [4] that the Wolfe gap $S:\mathbb{R}^n \to \mathbb{R}$ for problem (38) is defined by

$$S(w) = \max_{z\in\mathbb{R}^n} \left\{ -\langle \nabla(h^\lambda)^*(-w), w - z\rangle + f^*(w) - f^*(z) \right\}. \tag{43}$$

In the following lemma, we show that $S(z_j)$ is a primal-dual gap for (38). This result is an analog of Lemma 2.1, which also shows that $t_j$ is a primal-dual gap for (8).

**Lemma 3.1.** *Suppose that the assumptions in Theorem 3.1 hold, then for every $j \geq 1$, we have*

$$S(z_j) = \phi^\lambda(x_j) + \psi(z_j). \tag{44}$$

**Proof**: Since the assumptions in Theorem 3.1 hold, it follows from Theorem 3.1 that (42) holds for every $j \ge 1$. Using the second relation in (42) and the definition of $S(w)$ in (43), we have

$$\begin{aligned} S(z_j) &\overset{(43)}{=} \max_{z\in\mathbb{R}^n} \left\{ -\langle \nabla (h^\lambda)^*(-z_j), z_j - z\rangle + f^*(z_j) - f^*(z) \right\} \\ &\overset{(42)}{=} f^*(z_j) - \langle x_j, z_j\rangle + \max_{z\in\mathbb{R}^n} \{\langle x_j, z\rangle - f^*(z)\} \\ &= f^*(z_j) + \langle x_j, -z_j\rangle + f(x_j) \\ &\overset{(42)}{=} f^*(z_j) + (h^\lambda)^*(-z_j) + h^\lambda(x_j) + f(x_j), \end{aligned}$$

where we use the second relation in (42) again in the last identity. Finally, (44) immediately follows from the definitions of $\phi^\lambda$ and $\psi$ in (8) and (38), respectively. ∎

Recalling from Lemma 2.1 and using the first relation in (42) and the definition of $\psi$ in (38), we know

$$t_j \ge \phi^\lambda(\tilde{x}_j) + \psi(z_j), \tag{45}$$

i.e., $t_j$ an upper bound on a primal-dual gap for (38). On the other hand, Lemma 3.1 shows that $S(z_j)$ is a primal-dual gap for (38). We also note that the primal iterate used in $S(z_j)$ is $x_j$, while the one used in $t_j$ is $\tilde{x}_j$.

The following lemma gives a basic inequality used in the analysis of CG, which is adapted from Lemma 13.7 of [4]. For completeness, we present Lemma 13.7 of [4] as Lemma A.1 in Appendix A.

**Lemma 3.2.** *For every $j \ge 1$ and $\tau_j \in [0,1]$, the iterates $z_j$ and $\bar{z}_j$ generated by Algorithm 4 satisfy*

$$\psi(z_{j+1}) \le \psi(z_j) - (1-\tau_j)S(z_j) + \frac{(1-\tau_j)^2\lambda}{2}\|\bar{z}_j - z_j\|^2. \tag{46}$$

**Proof**: It is easy to see that (38) as an instance of (95) with

$$F = \psi, \quad f = (h^\lambda)^*, \quad g = f^*, \quad L_f = \lambda.$$

Therefore, (46) immediately follows from (96) with

$$x = z_j, \quad t = 1 - \tau_j, \quad p(x) = \bar{z}_j, \quad x + t(p(x) - x) = z_{j+1}.$$

∎

Define

$$u_j = \begin{cases} x_1, & \text{if } j = 1; \\ \tau_{j-1}u_{j-1} + (1-\tau_{j-1})x_{j-1}, & \text{otherwise.} \end{cases} \tag{47}$$

We are now ready to prove the primal-dual convergence of CG in terms of gap $\phi^\lambda(u_j) + \psi(z_j)$ in the following theorem, which resembles Proposition 2.3 for PDCP. An implicit assumption is that we are solving (38) as the dual to the proximal subproblem (8) within PDPB. Consequently, the iteration count $k$ in PDPB satisfies $k \le 2d_0^2/(\lambda\bar{\varepsilon})$, in accordance with the assumption in Lemma 2.8.

**Theorem 3.2.** *Suppose that the assumptions in Theorem 3.1 hold, and $\tau_j = j/(j+2)$, then for every $j \ge 1$,*

$$\phi^\lambda(u_j) + \psi(z_j) \le \frac{8M(3d_0 + \lambda M)}{j(j+1)} + \frac{8\lambda M^2}{j+1}. \tag{48}$$

**Proof**: Using Lemma 3.1, the convexity of $\phi^\lambda$, and definition of $u_j$ in (47), we have for every $j \ge 1$,

$$\begin{aligned} -(1-\tau_j)S(z_j) &\overset{(44)}{=} -(1-\tau_j)\phi^\lambda(x_j) - (1-\tau_j)\psi(z_j) \\ &\overset{(47)}{\le} -\phi^\lambda(u_{j+1}) + \tau_j\phi^\lambda(u_j) - (1-\tau_j)\psi(z_j). \end{aligned}$$

This inequality and Lemma 3.2 imply that

$$\phi^\lambda(u_{j+1}) + \psi(z_{j+1}) \overset{(46)}{\le} \tau_j[\phi^\lambda(u_j) + \psi(z_j)] + 2(1-\tau_j)^2\lambda M^2,$$

where we also use the facts that $\|\bar{z}_j\| \le M$ and $\|z_j\| \le M$ due to (A2) and $\bar{z}_j, z_j \in \operatorname{dom} f^*$. Note that for every $j \ge 1$, $\tau_j = A_j/A_{j+1}$ where $A_{j+1} = A_j + j + 1$ and $A_0 = 0$, i.e., $A_j = j(j+1)/2$ for every $j \ge 0$. It thus follows from the above inequality that

$$A_{j+1}[\phi^\lambda(u_{j+1}) + \psi(z_{j+1})] \le A_j[\phi^\lambda(u_j) + \psi(z_j)] + 4\lambda M^2$$

Summing the above inequality from $j = 1$ to $j$ and using the fact that $A_1 = 1$, we have

$$A_j[\phi^\lambda(u_j) + \psi(z_j)] \le \phi^\lambda(u_1) + \psi(z_1) + 4\lambda M^2 j.$$

In view of (47), it is easy to see that $u_1 = x_1 = \tilde{x}_1$, which together with Lemma 2.8 and (45) yields that

$$\phi^\lambda(u_1) + \psi(z_1) = \phi^\lambda(\tilde{x}_1) + \psi(z_1) \overset{(45)}{\le} t_1 \overset{(32)}{\le} 4M(3d_0 + \lambda M).$$

Therefore, (48) immediately follows from the above two inequalities and the fact that $A_j = j(j+1)/2$. ■

The results in this subsection justify the implementation of proximal subproblem (8) using CG from the dual point of view. In other words, PDPB can be also understood as the inexact PPM with CG as a subroutine.

### 3.2 GBM implementations inspired by CG

The discussion in this section so far is based on a particular implementation of GBM within PDCP, i.e., the one-cut scheme (41) with $\tau_j = j/(j+2)$ for every $j \geq 1$. Note that $\tau_j = j/(j+2)$ is also a standard choice in CG but not the only option. Inspired by alternative choices of $\tau_j$ used in CG (e.g., Section 13.2.3 of [4]), we also consider

$$\alpha_j = \max\left\{0, 1 - \frac{S(z_j)}{\lambda\|z_j - \bar{z}_j\|^2}\right\} \tag{49}$$

and

$$\beta_j \in \text{Argmin}\left\{\psi(\beta z_j + (1-\beta)\bar{z}_j) : \beta \in [0,1]\right\} \tag{50}$$

in this subsection and establish convergence rates of CG as in Theorem 3.2 but with $\alpha_j$ and $\beta_j$. As a consequence of the duality result (i.e., Theorem 3.1), this means that the one-cut scheme (41) can use also $\tau_j$ different from $j/(j+2)$. It is worth noting that these new choices of $\tau_j$ and their corresponding convergence proofs are only made possible by the duality connection discovered in this section.

The following theorem is a counterpart of Theorem 3.2 in the case of choosing $\tau_j$ of CG as in (49) or (50). An implicit assumption is that we are solving (38) as the dual to the proximal subproblem (8) within PDPB. Consequently, the iteration count $k$ in PDPB satisfies $k \leq 2d_0^2/(\lambda\bar{\varepsilon})$, in accordance with the assumption in Lemma 2.8.

**Theorem 3.3.** *Consider Algorithm 4 with $\tau_j$ as in (49) or (50), then for every $j \geq 1$, (48) holds where $u_j$ is as in (47) with $\tau_j = j/(j+2)$ and $z_j$ is as in (40) with $\tau_j$ as in (49) or (50) correspondingly.*

**Proof**: First, it follows from Lemma 3.2 and the definition of $z_{j+1}$ in (40) that for any $\tau_j \in [0,1]$,

$$\psi(\tau_j z_j + (1-\tau_j)\bar{z}_j) \overset{(46)}{\leq} \psi(z_j) - (1-\tau_j)S(z_j) + \frac{(1-\tau_j)^2\lambda}{2}\|\bar{z}_j - z_j\|^2. \tag{51}$$

Claim: In either case of Algorithm 4 with $\tau_j$ as in (49) or (50), we have for any $\tau_j \in [0,1]$,

$$\psi(z_{j+1}) \leq \psi(z_j) - (1-\tau_j)S(z_j) + \frac{(1-\tau_j)^2\lambda}{2}\|\bar{z}_j - z_j\|^2. \tag{52}$$

In the case of $\alpha_j$ in (49), it is easy to see from (40) that $z_{j+1} = \alpha_j z_j + (1-\alpha_j)\bar{z}_j$, which together with (51) with $\tau_j = \alpha_j$ implies that

$$\psi(z_{j+1}) \leq \psi(z_j) - (1-\alpha_j)S(z_j) + \frac{(1-\alpha_j)^2\lambda}{2}\|\bar{z}_j - z_j\|^2. \tag{53}$$

Noting from (49) that

$$1 - \alpha_j = \min\left\{1, \frac{S(z_j)}{\lambda\|z_j - \bar{z}_j\|^2}\right\},$$

which minimizes the right-hand side of (52) as a quadratic function of $1 - \tau_j$ over the interval [0,1]. Hence, (53) immediately implies that (52) holds for any $\tau_j \in [0,1]$. In the case of $\beta_j$ in (50), it is clear that for any $\tau_j \in [0,1]$,

$$\psi(z_{j+1}) \overset{(40)}{=} \psi(\beta_j z_j + (1-\beta_j)\bar{z}_j) \overset{(50)}{\leq} \psi(\tau_j z_j + (1-\tau_j)\bar{z}_j).$$

Hence, (52) immediately follows from this observation and (51). We have thus proved the claim. Except for $z_{j+1}$ in (52) is computed as in (40) with $\tau_j$ replaced by $\alpha_j$ or $\beta_j$, the claim is the same as Lemma 3.2. Finally, the conclusion of the theorem holds as a consequence of Theorem 3.2. ∎

### 3.3 New variants of CG inspired by GBM implementations

Motivated by possible $\tau_j$'s used in CG, we develop in Subsection 3.2 new implementations of GBM, i.e., the one-cut scheme (41) with $\alpha_j$ and $\beta_j$ in (49) and (50), respectively. In this subsection, we further exploit the duality between PDCP and CG from the other direction by developing novel CG variants with inspiration from other GBM implementations used in PDCP.

Apart from the one-cut scheme (41), Subsection 3.1 of [17] also provides two other candidates for GBM, i.e., two-cuts and multiple-cuts schemes, which are standard cut-aggregation and cutting-plane models, respectively.

To begin with, we first briefly review the two-cuts scheme. It starts from $\Gamma_1(\cdot) = \bar{\Gamma}_0(\cdot) = \ell_f(\cdot; x_0)$. For $j \geq 1$, given

$$\Gamma_j(\cdot) = \max\{\bar{\Gamma}_{j-1}(\cdot), \ell_f(\cdot; x_{j-1})\} \tag{54}$$

where $\bar{\Gamma}_{j-1}$ is an affine function, the two-cuts scheme recursively updates $\Gamma_{j+1}$ as in (54), i.e., $\Gamma_{j+1}(\cdot) = \max\left\{\bar{\Gamma}_j(\cdot), \ell_f(\cdot; x_j)\right\}$, which always maintains two cuts. The auxiliary bundle model $\bar{\Gamma}_j$ is updated as

$$\bar{\Gamma}_j(\cdot) = \theta_{j-1}\bar{\Gamma}_{j-1}(\cdot) + (1-\theta_{j-1})\ell_f(\cdot; x_{j-1}), \tag{55}$$

where $\theta_{j-1}$ is the Lagrange multiplier associated with the first constraint in the problem below

$$\min_{(u,r)\in\mathbb{R}^n\times\mathbb{R}} \left\{ r + h^\lambda(u) : \bar{\Gamma}_{j-1}(u) \leq r,\ \ell_f(u, x_{j-1}) \leq r \right\}. \tag{56}$$

Proposition D.1 in [17] shows that the above two-cuts scheme satisfies GBM.

Recall the previous options of $\tau_j$ in CG (see (40)), i.e., $j/(j+2)$, (49), and (50), are all determined once we know $z_j$ and $\bar{z}_j$. One natural way to generalize CG is to leave $\tau_j$ and, consequently, $z_{j+1}$ undetermined, deferring their computation to the subsequent iteration. Therefore, (39) and (40) are insufficient to determine $\tau_j$ and $z_{j+1}$, and more conditions are needed. For instance, motivated by the two-cuts scheme above, we additionally require

$$x_j = \nabla(h^\lambda)^*(-z_j), \quad \theta_{j-1}\bar{\Gamma}_{j-1}(x_j) + (1-\theta_{j-1})\ell_f(x_j; x_{j-1}) = \Gamma_j(x_j), \tag{57}$$

where $z_j = \theta_{j-1} z_{j-1} + (1-\theta_{j-1})\bar{z}_{j-1}$ following from (40). Note that (56) is equivalent to (9) with $\Gamma_j$ as in (54), and hence the optimal solution to (56) is $(x_j, \Gamma_j(x_j))$. As a result, with the understanding that $z_j = \nabla\bar{\Gamma}_j$ and $\bar{z}_j = f'(x_j)$, the first identity in (57) corresponds to the optimality of (56), and the second one in (57) corresponds to the complementary slackness of (56). Moreover, it follows from (58) that $\partial\Gamma_j$ is the convex hull of $\nabla\bar{\Gamma}_{j-1}$ and $f'(x_{j-1})$, and hence that

$$z_j = \nabla\bar{\Gamma}_j = \theta_{j-1}\nabla\bar{\Gamma}_{j-1} + (1-\theta_{j-1})f'(x_{j-1}) \in \partial\Gamma_j(x_j).$$

The discussion above verifies that Theorem 3.1 also holds in the context of the two-cuts scheme. In other words, in the spirit of Theorem 3.1, this new CG variant is the dual method of PDCP with the two-cuts implementation of GBM.

We now turn to review the multi-cuts scheme and discuss its implication in generalizing CG. For $j \ge 1$, given an index set $I_j \subseteq \{0, \cdots, j-1\}$, the multi-cuts scheme sets

$$\Gamma_j(\cdot) = \max\{\ell_f(\cdot; x_i) : i \in I_j\}. \tag{58}$$

The index set $I_j$ starts from $I_1 = \{0\}$ and recursively updates as

$$I_{j+1} = \bar{I}_{j+1} \cup \{j\}, \quad \bar{I}_{j+1} = \{i \in I_j : \theta_j^i > 0\},$$

where $\theta_j^i$ is the Lagrange multiplier associated with the constraint $\ell_f(u; x_i) \le r$ in the problem below

$$\min_{(u,r)\in\mathbb{R}^n\times\mathbb{R}} \left\{ r + h^\lambda(u) : \ell_f(u; x_i) \le r, \forall i \in I_j \right\}. \tag{59}$$

Here, $\bar{\Gamma}_j(\cdot) = \max\left\{\ell_f(\cdot; x_i) : i \in \bar{I}_j\right\}$. Proposition D.2 in [17] shows that the above multi-cuts scheme satisfies GBM.

The recursion (40) indicates that $z_j$ in CG is a convex combination of $\{z_1, \bar{z}_1, \ldots, \bar{z}_{j-1}\}$. Hence, a more general candidate of $z_j$ is any point in the convex hull of $\{z_1, \bar{z}_1, \ldots, \bar{z}_{j-1}\}$. Similar to the discussion of the new CG motivated by the two-cuts scheme, we also need to introduce conditions to determine $z_j$ in this generalization. For instance, inspired by the multi-cuts scheme above, we specifically compute

$$z_j = \sum_{i\in I_j} \theta_j^i \bar{z}_i$$

with the convention that $\bar{z}_0 = z_1$, where $\theta_j^i$ is the corresponding Lagrange multiplier for (59). Now, the positive multiplier $\theta_j^i$ (primal perspective) also serves as the convex combination parameter (dual perspective). Note that (59) is equivalent to (9) with $\Gamma_j$ as in (58), and hence the optimal solution to (59) is $(x_j, \Gamma_j(x_j))$. Again, it is easy to verify that

$$z_j \in \partial\Gamma_j(x_j), \quad x_j = \nabla(h^\lambda)^*(-z_j), \quad f'(x_j) = \bar{z}_j,$$

and hence Theorem 3.1 holds in the context of the multi-cuts scheme. In other words, following the spirit of Theorem 3.1, this generalization of CG serves as the dual method of PDCP, implemented with the multi-cuts scheme.

Since the number of nonzero $\theta_j^i$ could be small (compared to $j$), $z_j$ has a sparse representation in terms of $\{\bar z_0, \bar z_1, \ldots, \bar z_{j-1}\}$. Assuming $\{\bar z_j\}$ is a sequence of sparse vectors, then $z_j$ is sparse, and indeed sparser than those generated by CG using (40) with $\tau_j$ being $j/(j+2)$, $\alpha_j$, $\beta_j$, and $\theta_j$.

Leveraging the primal-dual connections between PDCP with two-cuts and multi-cuts schemes and the novel CG variants, we present the following convergence result for the latter. The proof is omitted, as it directly follows from Proposition 2.3 and Lemma 2.8, which establish the convergence of PDCP under the two-cuts and multi-cuts schemes. An implicit assumption is that we are solving (38) as the dual to the proximal subproblem (8) within PDPB. Consequently, the iteration count $k$ in PDPB satisfies $k \le 2d_0^2/(\lambda\bar\varepsilon)$, in accordance with the assumption in Lemma 2.8.

**Theorem 3.4.** *Consider the two CG variants described in this subsection, then $z_j$ generated in each variant satisfies*

$$\phi^\lambda(\tilde x_j) + \psi(z_j) \le \frac{8M(3d_0 + \lambda M)}{j(j+1)} + \frac{16\lambda M^2}{j+1},$$

*where $\tilde x_j$ is as in (12) with $\tau_j = j/(j+2)$.*

# 4 Proximal bundle method for SPP

In this section, we consider the convex-concave nonsmooth composite SPP (2). More specifically, we assume the following conditions hold:

(B1) a subgradient oracle $f_x' : \operatorname{dom} h_1 \times \operatorname{dom} h_2 \to \mathbb{R}^n$ and and a supergradient oracle $f_y' : \operatorname{dom} h_1 \times \operatorname{dom} h_2 \to \mathbb{R}^m$ are available, that is, we have $f_x'(u,v) \in \partial_x f(u,v)$ and $f_y'(u,v) \in \partial_y f(u,v)$ for every $(u,v) \in \operatorname{dom} h_1 \times \operatorname{dom} h_2$;

(B2) both $f_x'$ and $f_y'$ are uniformly bounded by some positive scalar $M$ over $\operatorname{dom} h_1$ and $\operatorname{dom} h_2$, i.e., for every pair $(u,v) \in \operatorname{dom} h_1 \times \operatorname{dom} h_2$,

$$\|f_x'(u,v)\| \le M, \quad \|f_y'(u,v)\| \le M; \tag{60}$$

(B3) $\operatorname{dom} h_1 \times \operatorname{dom} h_2$ is bounded with finite diameter $D > 0$;

(B4) the proximal mappings of $h_1$ and $h_2$ are easy to compute;

(B5) the set of saddle points of problem (2) is nonempty.

Given a pair $(x, y) \in \operatorname{dom} h_1 \times \operatorname{dom} h_2$, for every $(u, v) \in \operatorname{dom} h_1 \times \operatorname{dom} h_2$, define

$$\ell_{f(\cdot,y)}(u; x) = f(x, y) + \langle f'_x(x, y), u - x \rangle, \quad \ell_{f(x,\cdot)}(v; y) = f(x, y) + \langle f'_y(x, y), v - y \rangle.$$

It is easy to see from (B3) that for fixed $(x, y)$ and every $(u, v) \in \operatorname{dom} h_1 \times \operatorname{dom} h_2$,

$$f(u, y) - \ell_{f(\cdot,y)}(u; x) \le 2M\|u - x\|, \quad \ell_{f(x,\cdot)}(v; y) - f(x, v) \le 2M\|v - y\|. \tag{61}$$

We say a pair $(x_*, y_*) \in \operatorname{dom} h_1 \times \operatorname{dom} h_2$ is a saddle-point of (2) if for every $(u, v) \in \operatorname{dom} h_1 \times \operatorname{dom} h_2$,

$$\phi(x_*, v) \le \phi(x_*, y_*) \le \phi(u, y_*). \tag{62}$$

We say a pair $(x, y) \in \operatorname{dom} h_1 \times \operatorname{dom} h_2$ is an $\varepsilon$-saddle-point of (2) if

$$0 \in \partial_\varepsilon [\phi(\cdot, y) - \phi(x, \cdot)](x, y). \tag{63}$$

It is well-known that SPP (2) is equivalent to

$$\min_{x \in \mathbb{R}^n, y \in \mathbb{R}^m} \{\Phi(x, y) := \varphi(x) - \psi(y)\}, \tag{64}$$

where

$$\varphi(x) = \max_{y \in \mathbb{R}^m} \phi(x, y), \quad \psi(y) = \min_{x \in \mathbb{R}^n} \phi(x, y). \tag{65}$$

As a consequence, an equivalent definition of $\varepsilon$-saddle-point is as follows: a pair $(x, y) \in \operatorname{dom} h_1 \times \operatorname{dom} h_2$ satisfying

$$\Phi(x, y) = \varphi(x) - \psi(y) \le \varepsilon. \tag{66}$$

The equivalence between (63) and (66) is given in Lemma A.2. Another related but weaker notion is a pair $(x, y) \in \operatorname{dom} h_1 \times \operatorname{dom} h_2$ satisfying

$$-\varepsilon \le \phi(x, y) - \phi(x_*, y_*) \le \varepsilon. \tag{67}$$

The implication from (63) to (67) is given in Lemma A.3.

The composite subgradient method for SPP (2) denoted by CS-SPP$(x_0, y_0, \lambda)$, where $(x_0, y_0) \in \operatorname{dom} h_1 \times \operatorname{dom} h_2$ is the initial pair and $\lambda > 0$ is the prox stepsize, recursively computes

$$x_k = \underset{u \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ \ell_{f(\cdot,y_{k-1})}(u; x_{k-1}) + h_1(u) + \frac{1}{2\lambda}\|u - x_{k-1}\|^2 \right\}, \tag{68}$$

$$y_k = \underset{v \in \mathbb{R}^m}{\operatorname{argmin}} \left\{ -\ell_{f(x_{k-1},\cdot)}(v; y_{k-1}) + h_2(v) + \frac{1}{2\lambda}\|v - y_{k-1}\|^2 \right\}. \tag{69}$$

For given tolerance $\bar{\varepsilon} > 0$, letting $\lambda = \bar{\varepsilon}/(32M^2)$, then the iteration-complexity for CS-SPP$(x_0, y_0, \lambda)$ to generate a $\bar{\varepsilon}$-saddle point of (2) is bounded by $\mathcal{O}(M^2 D^2/\bar{\varepsilon}^2)$ (see Theorem C.1).

### 4.1 Inexact proximal point framework for SPP

The generic PPM for solving (64) iteratively solves the proximal subproblem

$$(x_k, y_k) = \underset{x\in\mathbb{R}^n, y\in\mathbb{R}^m}{\operatorname{argmin}} \left\{ \Phi(x,y) + \frac{1}{2\lambda_k}\|x - x_{k-1}\|^2 + \frac{1}{2\lambda_k}\|y - y_{k-1}\|^2 \right\}, \tag{70}$$

which motivates the following proximal point formulation for solving (2)

$$(x_k, y_k) = \underset{x\in\mathbb{R}^n}{\operatorname{argmin}}\, \underset{y\in\mathbb{R}^m}{\operatorname{argmax}} \left\{ \phi(x,y) + \frac{1}{2\lambda_k}\|x - x_{k-1}\|^2 - \frac{1}{2\lambda_k}\|y - y_{k-1}\|^2 \right\}. \tag{71}$$

However, both (70) and (71) are only conceptual PPMs for SPP. In this subsection, we introduce the generic IPPF for solving SPP (2) and show that CS-SPP described previously is a concrete example of IPPF.

---

**Algorithm 5** Inexact Proximal Point Framework for SPP (2)

---

**Initialize:** given initial pair $(x_0, y_0) \in \operatorname{dom} h_1 \times \operatorname{dom} h_2$ and scalar $\sigma \in [0,1]$

**for** $k = 1, 2, \cdots$ **do**

- choose $\lambda_k > 0$, $\varepsilon_k > 0$, and $\delta_k > 0$ and find $(x_k, y_k) \in \operatorname{dom} h_1 \times \operatorname{dom} h_2$ and $(\tilde{x}_k, \tilde{y}_k) \in \operatorname{dom} h_1 \times \operatorname{dom} h_2$ such that

$$\left( \frac{x_{k-1} - x_k}{\lambda_k}, \frac{y_{k-1} - y_k}{\lambda_k} \right) \in \partial_{\varepsilon_k}[\phi(\cdot, y_{k-1}) - \phi(x_{k-1}, \cdot)](\tilde{x}_k, \tilde{y}_k) \tag{72}$$

and

$$\|x_k - \tilde{x}_k\|^2 + \|y_k - \tilde{y}_k\|^2 + 2\lambda_k\varepsilon_k \le \delta_k + \sigma\left( \|\tilde{x}_k - x_{k-1}\|^2 + \|\tilde{y}_k - y_{k-1}\|^2 \right). \tag{73}$$

**end for**

---

**Lemma 4.1.** *For every $k \ge 1$, define $p_k : \mathbb{R}^n \to \mathbb{R}$ and $d_k : \mathbb{R}^m \to \mathbb{R}$ as follows*

$$p_k(\cdot) := f(\cdot, y_{k-1}) + h_1(\cdot), \quad d_k(\cdot) := -f(x_{k-1}, \cdot) + h_2(\cdot). \tag{74}$$

*Then, the inclusion (72) is equivalent to for every $(u,v) \in \operatorname{dom} h_1 \times \operatorname{dom} h_2$,*

$$\begin{aligned} & p_k(u) + d_k(v) - p_k(\tilde{x}_k) - d_k(\tilde{y}_k) \\ \ge & \frac{1}{\lambda_k}\langle x_{k-1} - x_k, u - \tilde{x}_k\rangle + \frac{1}{\lambda_k}\langle y_{k-1} - y_k, v - \tilde{y}_k\rangle - \varepsilon_k. \end{aligned} \tag{75}$$

**Proof**: It follows from (72) and the definition of $\varepsilon$-subdifferential (3) that for every pair $(u,v) \in \operatorname{dom} h_1 \times \operatorname{dom} h_2$,

$$\begin{aligned} & \phi(u, y_{k-1}) - \phi(x_{k-1}, v) - [\phi(\tilde{x}_k, y_{k-1}) - \phi(x_{k-1}, \tilde{y}_k)] \\ \ge & \frac{1}{\lambda_k}\langle x_{k-1} - x_k, u - \tilde{x}_k\rangle + \frac{1}{\lambda_k}\langle y_{k-1} - y_k, v - \tilde{y}_k\rangle - \varepsilon_k. \end{aligned}$$

Observing from the definitions of $p_k$ and $d_k$ in (74) that

$$p_k(u) + d_k(v) - p_k(\tilde{x}_k) - d_k(\tilde{y}_k) = \phi(u, y_{k-1}) - \phi(x_{k-1}, v) - [\phi(\tilde{x}_k, y_{k-1}) - \phi(x_{k-1}, \tilde{y}_k)],$$

which together with the above inequality implies that (75) holds. ∎

We are now ready to present the result showing that CS-SPP is an instance of IPPF with certain parameterizations. The proof is postponed to Subsection A.2.

**Proposition 4.2.** *Given $(x_0, y_0) \in \mathrm{dom}\, h_1 \times \mathrm{dom}\, h_2$, $\delta > 0$, and $\lambda = \sqrt{\delta/8M^2}$, then CS-SPP$(x_0, y_0, \lambda)$ is an instance of IPPF with $\sigma = 1$, $(\lambda_k, \delta_k) = (\lambda, \delta)$ for every $k \geq 1$, $(\tilde{x}_k, \tilde{y}_k) = (x_k, y_k)$ where $x_k$ and $y_k$ are as in (68) and (69), respectively, and $\varepsilon_k = \varepsilon_k^x + \varepsilon_k^y$ where*

$$\varepsilon_k^x = f(x_k, y_{k-1}) - \ell_{f(\cdot, y_{k-1})}(x_k; x_{k-1}), \tag{76}$$

$$\varepsilon_k^y = -f(x_{k-1}, y_k) + \ell_{f(x_{k-1}, \cdot)}(y_k; y_{k-1}). \tag{77}$$

## 4.2 Proximal bundle method for SPP

In this subsection, we describe another instance of IPPF, namely PB-SPP, for solving SPP (2). The inclusion of PB-SPP as an instance of IPPF is presented in Proposition 4.2 below.

We start by stating PB-SPP.

---

**Algorithm 6** Proximal Bundle for SPP (2), PB-SPP$(x_0, y_0, \bar{\varepsilon})$

---

**Initialize:** given $(x_0, y_0) \in \mathrm{dom}\, h_1 \times \mathrm{dom}\, h_2$ and $\bar{\varepsilon} > 0$

**for** $k = 1, 2, \cdots$ **do**

- call oracles $(x_k, \tilde{x}_k) = \mathrm{PDCP}(x_{k-1}, \lambda_k, \bar{\varepsilon}/4)$ and $(y_k, \tilde{y}_k) = \mathrm{PDCP}(y_{k-1}, \lambda_k, \bar{\varepsilon}/4)$ and compute

$$\bar{x}_k = \frac{1}{k} \sum_{i=1}^{k} \tilde{x}_i, \quad \bar{y}_k = \frac{1}{k} \sum_{i=1}^{k} \tilde{y}_i. \tag{78}$$

**end for**

---

Inspired by PPM (71) for solving SPP (2), the $k$-th iteration of PB-SPP aims at approximately solving the decoupled proximal subproblems, i.e.,

$$\min_{x \in \mathbb{R}^n} \left\{ f(x, y_{k-1}) + h_1(x) + \frac{1}{2\lambda_k} \|u - x_{k-1}\|^2 \right\}, \tag{79}$$

$$\min_{y \in \mathbb{R}^m} \left\{ -f(x_{k-1}, y) + h_2(y) + \frac{1}{2\lambda_k} \|v - y_{k-1}\|^2 \right\}. \tag{80}$$

Hence, the underlying $f$ in the call to PDCP$(x_{k-1}, \lambda_k, \bar{\varepsilon})$ is $f(\cdot, y_{k-1})$ and the underlying $f$ in the call to PDCP$(y_{k-1}, \lambda_k, \bar{\varepsilon})$ is $-f(x_{k-1}, \cdot)$. Correspondingly, similar to (22), by calling the subroutine PDCP, PB-SPP exactly solves

$$x_k = \underset{u\in\mathbb{R}^n}{\operatorname{argmin}} \left\{ \Gamma_k^x(u) + h_1(u) + \frac{1}{2\lambda_k}\|u - x_{k-1}\|^2 \right\}, \tag{81}$$

$$y_k = \underset{v\in\mathbb{R}^m}{\operatorname{argmin}} \left\{ -\Gamma_k^y(v) + h_2(v) + \frac{1}{2\lambda_k}\|v - y_{k-1}\|^2 \right\}, \tag{82}$$

where $\Gamma_k^x(\cdot)$ and $-\Gamma_k^y(\cdot)$ are the cutting-plane models constructed for $f(\cdot, y_{k-1})$ and $-f(x_{k-1}, \cdot)$, respectively, by GBM (see step 2 of Algorithm 2). Hence, by the construction in GBM (i.e., Algorithm 1) and the convexity of $f(\cdot, y_{k-1})$ and $-f(x_{k-1}, \cdot)$, we have

$$\Gamma_k^x(\cdot) \le f(\cdot, y_{k-1}), \quad -\Gamma_k^y(\cdot) \le -f(x_{k-1}, \cdot). \tag{83}$$

Since GBM is a generic scheme, the models $\Gamma_k^x(\cdot)$ and $-\Gamma_k^y(\cdot)$ can be any one satisfying GBM, e.g., one-cut, two-cuts, and multiple-cuts schemes (i.e., (E1)-(E3)) described in Subsection 3.1 of [17]. As a result, PB-SPP is a template for many possible methods using GBM as their bundle management.

For ease of the convergence analysis of PB-SPP, we define

$$p_k^\lambda(\cdot) := p_k(\cdot) + \frac{1}{2\lambda_k}\|\cdot - x_{k-1}\|^2, \quad d_k^\lambda(\cdot) := d_k(\cdot) + \frac{1}{2\lambda_k}\|\cdot - y_{k-1}\|^2, \tag{84}$$

where $p_k$ and $d_k$ are as in (74), $m_k^x$ and $m_k^y$ as the optimal values of (81) and (82), respectively, and

$$t_k^x = p_k^\lambda(\tilde{x}_k) - m_k^x, \quad t_k^y = d_k^\lambda(\tilde{y}_k) - m_k^x. \tag{85}$$

Following from Proposition 2.3 and a simplification of Lemma 2.8 using (B3), we obtain the convergence rates of $t_k^x$ and $t_k^y$. We omit the proof since it is almost identical to that of Proposition 2.3.

**Proposition 4.3.** *Considering Algorithm 2 with $\tau_j = j/(j+2)$, then for every $j_k \ge 1$, we have*

$$t_k^x \le \frac{4MD}{l_k(l_k+1)} + \frac{16\lambda_k M^2}{l_k+1}, \quad t_k^y \le \frac{4MD}{l_k(l_k+1)} + \frac{16\lambda_k M^2}{l_k+1},$$

*where $l_k$ denotes the length of the $k$-th cycle $\mathcal{C}_k$ (i.e., $l_k = |\mathcal{C}_k| = j_k - i_k + 1$).*

Given Proposition 4.3, PDCP is able to solve (79) and (80) to any desired accuracy. For given tolerance $\bar{\varepsilon} > 0$, the calls to PDCP in Algorithm 6 guarantees

$$t_k^x \le \frac{\bar{\varepsilon}}{4}, \quad t_k^y \le \frac{\bar{\varepsilon}}{4}. \tag{86}$$

Starting from (86), we establish the iteration-complexity for PB-SPP to find a $\bar{\varepsilon}$-saddle-point of SPP (2).

**Lemma 4.4.** *For every $k \geq 1$ and $(u,v) \in \mathrm{dom}\, h_1 \times \mathrm{dom}\, h_2$, we have*

$$p_k(\tilde{x}_k) - p_k(u) \leq \frac{\bar{\varepsilon}}{4} + \frac{1}{2\lambda_k}\|u - x_{k-1}\|^2 - \frac{1}{2\lambda_k}\|u - x_k\|^2 - \frac{1}{2\lambda_k}\|\tilde{x}_k - x_{k-1}\|^2, \tag{87}$$

$$d_k(\tilde{y}_k) - d_k(v) \leq \frac{\bar{\varepsilon}}{4} + \frac{1}{2\lambda_k}\|v - y_{k-1}\|^2 - \frac{1}{2\lambda_k}\|v - y_k\|^2 - \frac{1}{2\lambda_k}\|\tilde{y}_k - y_{k-1}\|^2. \tag{88}$$

**Proof**: We only prove (87) to avoid duplication. Inequality (88) follows similarly. Noting that the objective in (81) is $\lambda_k^{-1}$-strongly convex and using the definition of $m_k^x$, we have for every $u \in \mathbb{R}^n$,

$$\Gamma_k^x(u) + h_1(u) + \frac{1}{2\lambda_k}\|u - x_{k-1}\|^2 \geq m_k^x + \frac{1}{2\lambda_k}\|u - x_k\|^2.$$

It follows from the definition of $p_k$ in (74) and the first inequality in (83) that $p_k(\cdot) \geq (\Gamma_k^x + h_1)(\cdot)$. Hence, we have for every $u \in \mathbb{R}^n$,

$$p_k^\lambda(\tilde{x}_k) - p_k(u) \leq p_k^\lambda(\tilde{x}_k) - m_k^x + \frac{1}{2\lambda_k}\|u - x_{k-1}\|^2 - \frac{1}{2\lambda_k}\|u - x_k\|^2.$$

Therefore, inequality (87) immediately follows from the first inequality in (86). ■

**Lemma 4.5.** *For every $k \geq 1$ and $(u,v) \in \mathrm{dom}\, h_1 \times \mathrm{dom}\, h_2$, we have*

$$\phi(\tilde{x}_k, v) - \phi(u, \tilde{y}_k) \leq \frac{\bar{\varepsilon}}{2} + \frac{1}{2\lambda_k}\|z_{k-1} - w\|^2 - \frac{1}{2\lambda_k}\|z_k - w\|^2 + 4\lambda_k M^2, \tag{89}$$

*where $w = (u,v)$ and $z_k = (x_k, y_k)$.*

**Proof**: It follows from (B2) that for every $u \in \mathrm{dom}\, h_1$,

$$f(u, y_{k-1}) - f(u, \tilde{y}_k) \overset{(60)}{\leq} M\|\tilde{y}_k - y_{k-1}\|, \quad f(\tilde{x}_k, \tilde{y}_k) - f(\tilde{x}_k, y_{k-1}) \overset{(60)}{\leq} M\|\tilde{y}_k - y_{k-1}\|.$$

Noting from (74) that $p_k(\tilde{x}_k) - p_k(u) = f(\tilde{x}_k, y_{k-1}) + h_1(\tilde{x}_k) - f(u, y_{k-1}) - h_1(u)$, using this relation and the above inequality in (87), we have for every $u \in \mathrm{dom}\, h_1$,

$$\begin{aligned} & f(\tilde{x}_k, \tilde{y}_k) + h_1(\tilde{x}_k) - f(u, \tilde{y}_k) - h_1(u) \\ \overset{(87)}{\leq} & \frac{\bar{\varepsilon}}{4} + \frac{1}{2\lambda_k}\|x_{k-1} - u\|^2 - \frac{1}{2\lambda_k}\|x_k - u\|^2 + 2M\|\tilde{y}_k - y_{k-1}\| - \frac{1}{2\lambda_k}\|\tilde{x}_k - x_{k-1}\|^2. \end{aligned}$$

Similarly, using (88), we can prove for every $v \in \mathrm{dom}\, h_2$,

$$\begin{aligned} & -f(\tilde{x}_k, \tilde{y}_k) + h_2(\tilde{y}_k) + f(\tilde{x}_k, v) - h_2(v) \\ \overset{(88)}{\leq} & \frac{\bar{\varepsilon}}{4} + \frac{1}{2\lambda_k}\|y_{k-1} - v\|^2 - \frac{1}{2\lambda_k}\|y_k - v\|^2 + 2M\|\tilde{x}_k - x_{k-1}\| - \frac{1}{2\lambda_k}\|\tilde{y}_k - y_{k-1}\|^2. \end{aligned}$$

Noting that $2Ma - a^2/(2\lambda_k) \le 2\lambda_k M^2$ for $a \in \mathbb{R}$ and summing the above two inequalities, we obtain

$$\begin{aligned}\phi(\tilde{x}_k, v) - \phi(u, \tilde{y}_k) &\overset{(2)}{=} f(\tilde{x}_k, v) + h_1(\tilde{x}_k) - h_2(v) - f(u, \tilde{y}_k) - h_1(u) + h_2(\tilde{y}_k) \\ &\le \frac{\bar{\varepsilon}}{2} + \frac{1}{2\lambda_k}\|z_{k-1} - w\|^2 - \frac{1}{2\lambda_k}\|z_k - w\|^2 + 4\lambda_k M^2,\end{aligned}$$

where the identity is due to the definition of $\phi(\cdot,\cdot)$ in (2). ∎

**Proposition 4.6.** *For every $k \ge 1$, setting $\lambda_k = \lambda_1/\sqrt{k}$ for some $\lambda_1 > 0$, then for every $(u, v) \in \mathrm{dom}\, h_1 \times \mathrm{dom}\, h_2$, we have*

$$\varphi(\bar{x}_k) - \psi(\bar{y}_k) \le \frac{\bar{\varepsilon}}{2} + \frac{8\lambda_1 M^2}{\sqrt{k}} + \frac{D^2}{2\lambda_1\sqrt{k}}, \tag{90}$$

*where $\bar{x}_k$ and $\bar{y}_k$ are as in (78).*

**Proof**: Summing (89) from $k = 1$ to $k$ and using (78) and the convexity of $\phi(\cdot, y)$ and $-\phi(x, \cdot)$, we have for every $(u, v) \in \mathrm{dom}\, h_1 \times \mathrm{dom}\, h_2$,

$$\phi(\bar{x}_k, v) - \phi(u, \bar{y}_k) \le \frac{\bar{\varepsilon}}{2} + \frac{1}{k}\sum_{i=1}^{k}\left[\frac{1}{2\lambda_i}(\|z_{i-1} - w\|^2 - \|z_i - w\|^2) + 4\lambda_i M^2\right]. \tag{91}$$

It follows from the fact that $\lambda_k = \lambda/\sqrt{k}$ and assumption (B3) that

$$\begin{aligned}\frac{1}{k}\sum_{i=1}^{k}\left[\frac{1}{2\lambda_i}(\|z_{i-1} - w\|^2 - \|z_i - w\|^2)\right] &\le \frac{1}{2k}\left[\frac{\|z_0 - w\|^2}{\lambda_1} + \sum_{i=1}^{k-1}\|z_i - w\|^2\left(\frac{1}{\lambda_{i+1}} - \frac{1}{\lambda_i}\right)\right] \\ &\le \frac{D^2}{2k\lambda_k} = \frac{D^2}{2\lambda_1\sqrt{k}}.\end{aligned} \tag{92}$$

Observing that $\sum_{i=1}^{k}(1/\sqrt{i}) \le \int_0^k (1/\sqrt{x})dx = 2\sqrt{k}$, and hence

$$\frac{1}{k}\sum_{i=1}^{k} 4\lambda_i M^2 = \frac{1}{k}\sum_{i=1}^{k}\frac{4\lambda_1 M^2}{\sqrt{i}} \le \frac{8\lambda_1 M^2}{\sqrt{k}}.$$

This observation, (91), and (92) imply that

$$\phi(\bar{x}_k, v) - \phi(u, \bar{y}_k) \le \frac{\bar{\varepsilon}}{2} + \frac{8\lambda_1 M^2}{\sqrt{k}} + \frac{D^2}{2\lambda_1\sqrt{k}}.$$

Maximizing the left-hand side over $(u, v) \in \mathbb{R}^n \times \mathbb{R}^m$ and using (65) yield (90). ∎

We are now ready to establish the iteration-complexity for PB-SPP to find a $\bar{\varepsilon}$-saddle-point.

**Theorem 4.1.** *Given $(x_0, y_0, \bar{\varepsilon}) \in \mathrm{dom}\, h_1 \times \mathrm{dom}\, h_2 \times R_{++}$, letting $\lambda_1 = D/(4M)$, then the iteration-complexity for PB-SPP$(x_0, y_0)$ to find a $\bar{\varepsilon}$-saddle-point $(\bar{x}_k, \bar{y}_k)$ of (2) is $\mathcal{O}((MD/\bar{\varepsilon})^{2.5})$.*

**Proof**: It follows from Proposition 4.6 with $\lambda_1 = D/(4M)$ that

$$\varphi(\bar{x}_k) - \psi(\bar{y}_k) \le \frac{\bar{\varepsilon}}{2} + \frac{4MD}{\sqrt{k}}.$$

Hence, PB-SPP takes $k = 64M^2D^2/\bar{\varepsilon}^2$ iterations to find the $\bar{\varepsilon}$-saddle-point $(\bar{x}_k, \bar{y}_k)$. Using Proposition 4.3, we know to have (86) holds for every cycle $\mathcal{C}_i$, it is sufficient to have

$$l_i = \frac{\sqrt{32MD}}{\sqrt{\bar{\varepsilon}}} + \frac{128\lambda_i M^2}{\bar{\varepsilon}} = \frac{\sqrt{32MD}}{\sqrt{\bar{\varepsilon}}} + \frac{32MD}{\bar{\varepsilon}\sqrt{i}}.$$

As a consequence, the total number of iterations (of proximal mappings of $h_1$ and $h_2$, and of calls to subgradient oracles $f'_x$ and $f'_y$) is

$$\sum_{i=1}^{k} l_i = \frac{\sqrt{32MD}}{\sqrt{\bar{\varepsilon}}}k + \sum_{i=1}^{k} \frac{32MD}{\bar{\varepsilon}\sqrt{i}} \le \frac{256\sqrt{2}M^{2.5}D^{2.5}}{\bar{\varepsilon}^{2.5}} + \frac{512M^2D^2}{\bar{\varepsilon}^2},$$

where we use the facts that$\sum_{i=1}^{k}(1/\sqrt{i}) \le \int_0^k (1/\sqrt{x})dx = 2\sqrt{k}$ and $k = 64M^2D^2/\bar{\varepsilon}^2$. ■

Note that the complexity in Theorem 4.1 holds for any model $\Gamma_k^x$ and $-\Gamma_k^y$ generated by GBM, such as one-cut, two-cuts, and multiple-cuts schemes described in Subsection 3.1 of [17]. However, there is a possiblity that the complexity for PB-SPP using two-cuts and multiple-cuts schemes becomes better as $\mathcal{O}(M^2D^2/\bar{\varepsilon}^2)$. For simplicity, we only present the analysis for the current bound $\mathcal{O}((MD/\bar{\varepsilon})^{2.5})$ and leave the finer bound $\mathcal{O}(M^2D^2/\bar{\varepsilon}^2)$ for future investigation.

Finally, we conclude this subsection by presenting that PB-SPP is an instance of IPPF. The proof is postponed to Subsection A.3.

**Proposition 4.7.** *Given $(x_0, y_0) \in \mathrm{dom}\, h_1 \times \mathrm{dom}\, h_2$, $\bar{\varepsilon} > 0$, then PB-SPP$(x_0, y_0, \bar{\varepsilon})$ is an instance of IPPF with $\sigma = 0$, $\delta_k = \lambda_k\bar{\varepsilon}/2$, and $\varepsilon_k = \varepsilon_k^x + \varepsilon_k^y$ where*

$$\varepsilon_k^x = p_k(\tilde{x}_k) - (\Gamma_k^x + h_1)(x_k) + \frac{1}{\lambda_k}\langle x_{k-1} - x_k, x_k - \tilde{x}_k\rangle, \tag{93}$$

$$\varepsilon_k^y = d_k(\tilde{y}_k) - (-\Gamma_k^y + h_2)(y_k) + \frac{1}{\lambda_k}\langle y_{k-1} - y_k, y_k - \tilde{y}_k\rangle. \tag{94}$$

# 5 Concluding remarks

This paper studies the iteration-complexity of modern PB methods for solving CNCO (1) and SPP (2). It proposes PDPB for solving (1) and provides the iteration-complexity of PDPB in terms of a primal-dual gap. The paper also introduces PB-SPP for solving (2) and establishes the iteration-complexity to find a $\bar{\varepsilon}$-saddle-point. Another interesting feature of the paper is that it investigates the duality between CG and PDCP for solving the proximal subproblem (8). The paper further develops novel variants of both CG and PDCP leveraging the duality.

We finally discuss some possible extensions of our methods and analyses. First, we have studied modern PB methods for solving CNCO and SPP in this paper, and we could extend the methods to solving more general nonsmooth problems with convex structures such constrained optimization, equilibrium problems, and variational inequalities. Second, as already noted in the paragraph below Theorem 4.1, it is possible to improve the complexity bound $\mathcal{O}((MD/\bar{\varepsilon})^{2.5})$ of PB-SPP to $\mathcal{O}(M^2D^2/\bar{\varepsilon}^2)$, if we employ only two-cuts and multiple-cuts schemes rather than the GBM. Third, it is interesting to study the duality between PDCP and CG in the context of SPP, which is equivalent to developing a CG method to implement (72) and (73) within IPPF. Fourth, similar to the universal methods proposed in [11], we are also interested in developing universal variants of PB-SPP for SPP (2) under strong convexity assumptions without knowing the problem-dependent parameters a priori. Finally, following the stochastic PB method developed for stochastic CNCO in [15], it is worthwhile to explore stochastic versions of PB-SPP for solving stochastic SPP, particularly those involving decision-dependent distributions.

## References


[1] A. Astorino, A. Frangioni, A. Fuduli, and E. Gorgone. A nonmonotone proximal bundle method with (potentially) continuous step decisions. *SIAM Journal on Optimization*, 23(3):1784–1809, 2013.

[2] F. Bach. Learning with submodular functions: A convex optimization perspective. *Foundations and Trends® in machine learning*, 6(2-3):145–373, 2013.

[3] F. Bach. Duality between subgradient and conditional gradient methods. *SIAM Journal on Optimization*, 25(1):115–129, 2015.

[4] A. Beck. *First-Order Methods in Optimization*, volume 25. SIAM, 2017.

[5] D. P. Bertsekas and H. Yu. A unifying polyhedral approximation framework for convex optimization. *SIAM Journal on Optimization*, 21(1):333–360, 2011.

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

[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] D. Fersztand and X. A. Sun. The proximal bundle algorithm under a frank-wolfe perspective: an improved complexity analysis. *arXiv preprint arXiv:2411.15926*, 2024.

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

[10] E. G. Golshtein. Generalized gradient method for finding saddlepoints. *Matekon*, 10(3):36–52, 1974.

[11] V. Guigues, J. Liang, and R. D. C. Monteiro. Universal subgradient and proximal bundle methods for convex and strongly convex hybrid composite optimization. *arXiv preprint arXiv:2407.10073*, 2024.

[12] K. C. Kiwiel. Efficiency of proximal bundle methods. *Journal of Optimization Theory and Applications*, 104(3):589–603, 2000.

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

[14] C. Lemaréchal. Nonsmooth optimization and descent methods. 1978.

[15] J. Liang, V. Guigues, and R. D. C. Monteiro. A single cut proximal bundle method for stochastic convex composite optimization. *Mathematical programming*, 208(1):173–208, 2024.

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

[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] D. Maistroskii. Gradient methods for finding saddle points. *Matekon*, 13(3):22, 1977.

[19] G. Mazanti, T. Moquet, and L. Pfeiffer. A nonsmooth frank-wolfe algorithm through a dual cutting-plane approach. *arXiv preprint arXiv:2403.18744*, 2024.

[20] R. Mifflin. A modification and an extension of Lemaréchal's algorithm for nonsmooth minimization. In *Nondifferential and variational techniques in optimization*, pages 77–90. Springer, 1982.

[21] A. Nedić and A. Ozdaglar. Subgradient methods for saddle-point problems. *Journal of optimization theory and applications*, 142:205–228, 2009.

[22] Y. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical programming*, 120(1):221–259, 2009.

[23] W. de Oliveira, C. Sagastizábal, and C. Lemaréchal. Convex proximal bundle methods in depth: a unified analysis for inexact oracles. *Mathematical Programming*, 148(1-2):241–277, 2014.

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

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

[26] H. Uzawa. Iterative methods for concave programming. *Studies in linear and nonlinear programming*, 6:154–165, 1958.

[27] W. van Ackooij, V. Berge, W. de Oliveira, and C. Sagastizábal. Probabilistic optimization via approximate p-efficient points and bundle methods. *Computers & Operations Research*, 77:177–193, 2017.

[28] P. Wolfe. A method of conjugate subgradients for minimizing nondifferentiable functions. In *Nondifferentiable optimization*, pages 145–173. Springer, 1975.

[29] I. Y. Zabotin. A subgradient method for finding a saddle point of a convex-concave function. *Issled. Prikl. Mat*, 15(6):12, 1988.

[30] S. Zhou, S. Gupta, and M. Udell. Limited memory kelley's method converges for composite convex and submodular objectives. *Advances in Neural Information Processing Systems*, 31, 2018.

# A Technical results and deferred proofs

This section collects technical results used throughout the paper and deferred proofs from Section 4.

## A.1 Technical results

We present Lemma 13.7 of [4] with slight modification, which is used in the proof of Lemma 3.2.

**Lemma A.1.** *Consider*

$$\min_{x\in\mathbb{R}^n} \{F(x) = f(x) + g(x)\}, \tag{95}$$

*where* $f \in \overline{\text{Conv}}\,(\mathbb{R}^n)$*,* $g \in \overline{\text{Conv}}\,(\mathbb{R}^n)$*, and* $\text{dom}\, g \subset \text{dom}\, f$*. Moreover,* $f$ *is* $L_f$*-smooth over* $\text{dom}\, f$*. Define*

$$S(x) = \max_{p \in \mathbb{R}^n} \{\langle \nabla f(x), x - p\rangle + g(x) - g(p)\}, \quad p(x) = \underset{p \in \mathbb{R}^n}{\text{argmin}}\, \{\langle p, \nabla f(x)\rangle + g(p)\}.$$

*Then, for every* $x \in \text{dom}\, g$ *and* $t \in [0, 1]$*, if* $p(x)$ *exists, we have*

$$F(x + t(p(x) - x)) \le F(x) - tS(x) + \frac{t^2 L_f}{2}\|p(x) - x\|^2. \tag{96}$$

**Lemma A.2.** *Given* $\varepsilon > 0$*, a pair* $(x, y)$ *is an* $\varepsilon$*-saddle-point of* (2) *(i.e., satisfying* (63)*) if and only if the pair satisfies* (66)*.*

**Proof**: It follows from (63) that for every $(u, v) \in \text{dom}\, h_1 \times \text{dom}\, h_2$,

$$\phi(u, y) - \phi(x, v) \ge \phi(x, y) - \phi(x, y) - \varepsilon = -\varepsilon. \tag{97}$$

Hence, (97) holds with $(u, v) = (x(y), y(x))$ where

$$x(y) = \underset{x \in \mathbb{R}^n}{\text{argmin}}\, \phi(x, y), \quad y(x) = \underset{y \in \mathbb{R}^m}{\text{argmax}}\, \phi(x, y),$$

that is

$$\min_{x \in \mathbb{R}^n} \phi(x, y) - \max_{y \in \mathbb{R}^m} \phi(x, y) = \phi(x(y), y) - \phi(x, y(x)) \overset{(97)}{\ge} -\varepsilon.$$

This result, together with (64) and (65), implies that (66) holds. On the other hand, assuming that (66) holds, then for every $(u, v) \in \text{dom}\, h_1 \times \text{dom}\, h_2$, it obviously follows from (65) that

$$\phi(x, v) - \phi(u, y) \overset{(65)}{\le} \varphi(x) - \psi(y) \le \varepsilon,$$

which is (63) in view of (97). ■

**Lemma A.3.** *Given* $\varepsilon > 0$*, a pair* $(x, y)$ *is an* $\varepsilon$*-saddle-point of* (2) *(i.e., satisfying* (63)*) implies* (67)*.*

**Proof**: Assuming that $(x, y)$ is an $\varepsilon$-saddle-point, it follows from Lemma A.2 that (66) holds, and hence that for every $(u, v) \in \text{dom}\, h_1 \times \text{dom}\, h_2$,

$$\phi(x, v) - \phi(u, y) \overset{(65)}{\le} \varphi(x) - \psi(y) \le \varepsilon, \tag{98}$$

where the first inequality is due to (65). Taking $(u, v) = (x_*, y)$ in (98) and using the first inequality in (62), we have

$$\phi(x, y) - \phi(x_*, y_*) \overset{(62)}{\le} \phi(x, y) - \phi(x_*, y) \overset{(98)}{\le} \varepsilon.$$

Taking $(u, v) = (x, y_*)$ in (98) and using the second inequality in (62), we have

$$\phi(x_*, y_*) - \phi(x, y) \overset{(62)}{\le} \phi(x, y_*) - \phi(x, y) \overset{(98)}{\le} \varepsilon.$$

Therefore, (67) immediately follows from the above two inequalities. ■

### A.2 Proof of Proposition 4.2

**Proof**: We first show that CS-SPP satisfies (72). It follows from the CS-SPP iterate (68) that

$$\frac{x_{k-1}-x_k}{\lambda} \in \partial[\ell_{f(\cdot,y_{k-1})}(\cdot;x_{k-1})+h_1](x_k).$$

Using the inclusion above, we have for every $u \in \operatorname{dom} h_1$,

$$[\ell_{f(\cdot,y_{k-1})}(\cdot;x_{k-1})+h_1](u) \geq [\ell_{f(\cdot,y_{k-1})}(\cdot;x_{k-1})+h_1](x_k)+\frac{1}{\lambda}\langle x_{k-1}-x_k, u-x_k\rangle.$$

Using the definition of $p_k$ in (74) and the fact that $f(\cdot,y_{k-1})$ is convex, we further obtain

$$p_k(u) \geq p_k(x_k)+\frac{1}{\lambda}\langle x_{k-1}-x_k, u-x_k\rangle - \varepsilon_k^x,$$

where $\varepsilon_k^x$ is as in (76). Similarly, we have for every $v \in \operatorname{dom} h_2$,

$$d_k(v) \geq d_k(y_k)+\frac{1}{\lambda}\langle y_{k-1}-y_k, v-y_k\rangle - \varepsilon_k^y,$$

where $\varepsilon_k^y$ is as in (77). Summing the above two inequalities gives (75) with $\lambda_k=\lambda$, $\varepsilon_k=\varepsilon_k^x+\varepsilon_k^y$ and $(\tilde{x}_k,\tilde{y}_k)=(x_k,y_k)$, and hence (72) holds in view of Lemma 4.1.

We next show that CS-SPP satisfies (73). Indeed, it follows from the definition of $\varepsilon_k^x$ in (76) and the first inequality in (61) that

$$\begin{aligned}2\lambda\varepsilon_k^x-\|x_k-x_{k-1}\|^2 &\overset{(76)}{=} 2\lambda[f(x_k,y_{k-1})-\ell_{f(\cdot,y_{k-1})}(x_k;x_{k-1})]-\|x_k-x_{k-1}\|^2\\ &\overset{(61)}{\leq} 4\lambda M\|x_k-x_{k-1}\|-\|x_k-x_{k-1}\|^2 \leq 4\lambda^2M^2.\end{aligned}$$

Similarly, we have $2\lambda\varepsilon_k^y-\|y_k-y_{k-1}\|^2 \leq 4\lambda^2M^2$. Summing the two inequalities and using the facts that $\lambda=\sqrt{\delta/8M^2}$ and $\varepsilon_k=\varepsilon_k^x+\varepsilon_k^y$, we have

$$2\lambda\varepsilon_k-\|x_k-x_{k-1}\|^2-\|y_k-y_{k-1}\|^2 \leq 8\lambda^2M^2=\delta,$$

which is (73) with $\sigma=1$, $(\lambda_k,\delta_k)=(\lambda,\delta)$, and $(\tilde{x}_k,\tilde{y}_k)=(x_k,y_k)$. ∎

### A.3 Proof of Proposition 4.7

**Proof**: We first show that PB-SPP satisfies (72). It follows from (81) that

$$\frac{x_{k-1}-x_k}{\lambda_k} \in \partial(\Gamma_k^x+h_1)(x_k),$$

which implies that for every $u \in \operatorname{dom} h_1$,

$$(\Gamma_k^x+h_1)(u) \geq (\Gamma_k^x+h_1)(x_k)+\frac{1}{\lambda_k}\langle x_{k-1}-x_k, u-x_k\rangle.$$

Using the first inequality in (83) and the definition of $p_k$ in (74), we have

$$p_k(u) \ge (\Gamma_k^x + h_1)(u) \ge p_k(\tilde{x}_k) + \frac{1}{\lambda_k}\langle x_{k-1} - x_k, u - \tilde{x}_k\rangle - \varepsilon_k^x, \quad \forall u,$$

where $\varepsilon_k^x$ is as in (93). Similarly, we have for every $v \in \operatorname{dom} h_2$,

$$d_k(v) \ge d_k(\tilde{y}_k) + \frac{1}{\lambda_k}\langle y_{k-1} - y_k, v - \tilde{y}_k\rangle - \varepsilon_k^y, \quad \forall v,$$

where $\varepsilon_k^y$ is as in (94). Summing the above two inequalities gives (75) with $\varepsilon_k = \varepsilon_k^x + \varepsilon_k^y$, and hence (72) holds in view of Lemma 4.1.

We next show that PB-SPP satisfies (73). Indeed, it follows from the definitions of $\varepsilon_k^x$ and $\varepsilon_k^y$ in (93) and (94), respectively, that

$$\|x_k - \tilde{x}_k\|^2 + \|y_k - \tilde{y}_k\|^2 + 2\lambda_k\varepsilon_k = \lambda_k\left(p_k^\lambda(\tilde{x}_k) - m_k^x + d_k^\lambda(\tilde{y}_k) - m_k^y\right),$$

where $p_k^\lambda$ and $d_k^\lambda$ are as in (84) and $m_k^x$ and $m_k^y$ as the optimal values of (81) and (82), respectively. In view of (85) and (86), the above relation further implies that

$$\|x_k - \tilde{x}_k\|^2 + \|y_k - \tilde{y}_k\|^2 + 2\lambda_k\varepsilon_k \le \frac{\lambda_k\bar{\varepsilon}}{2},$$

which is (73) with $\sigma = 0$ and $\delta_k = \lambda_k\bar{\varepsilon}/2$. ■

# B Primal-dual subgradient method for CNCO

This section is devoted to the complexity analysis of PDS. The main result is Theorem B.2 below.

Recall the definitions of $d_0$ and $x_0^*$ in (6). Since $x_0^* \in B(x_0, 4d_0)$, which is the ball centered at $x_0$ and with radius $4d_0$, it is easy to see that to solve (1), it suffices to solve

$$\min\left\{\hat{\phi}(x) := f(x) + \hat{h}(x) : x \in \mathbb{R}^n\right\} = \min\left\{\phi(x) : x \in Q\right\}, \tag{99}$$

where $\hat{h} = h + I_Q$ and $I_Q$ is the indicator function of $Q = B(x_0, 4d_0)$. Hence, it is convenient to consider a slightly modified version of PDS$(x_0, \lambda)$ with $h$ replaced by $\hat{h}$ in (7), denoted by MPDS$(x_0, \lambda)$, i.e.,

$$s_k = f'(x_{k-1}), \quad x_k = \underset{u \in \mathbb{R}^n}{\operatorname{argmin}}\left\{\ell_f(u; x_{k-1}) + \hat{h}(u) + \frac{1}{2\lambda}\|u - x_{k-1}\|^2\right\}. \tag{100}$$

It is worth noting that MPDS$(x_0, \lambda)$ is a conceptual method since we do not know $d_0$ and hence $\hat{h}$. We show equivalence between PDS$(x_0, \lambda)$ and MPDS$(x_0, \lambda)$, and only use MPDS$(x_0, \lambda)$ for analyzing the convegence.

We first establish the complexity of the primal-dual convergence of MPDS$(x_0, \lambda)$ for solving (99), and then we argue that MPDS$(x_0, \lambda)$ and PDS$(x_0, \lambda)$ generate the same primal and dual sequences $\{x_k\}$ and $\{s_k\}$ before convergence (see Lemma B.3). Therefore, we also give the complexity of PDS$(x_0, \lambda)$ for solving (99).

The following lemma is the starting point of the primal-dual convergence analysis.

**Lemma B.1.** *Given $x_0 \in \mathbb{R}^n$, for every $k \ge 1$ and $u \in \mathrm{dom}\, \hat{h}$, the sequence $\{x_k\}$ generated by MPDS$(x_0, \lambda)$ satisfies*

$$\hat{\phi}(x_k) - \ell_f(u; x_{k-1}) - \hat{h}(u) \le 2\lambda M^2 + \frac{1}{2\lambda}\|u - x_{k-1}\|^2 - \frac{1}{2\lambda}\|u - x_k\|^2. \tag{101}$$

**Proof**: Noticing that the objective function in (100) is $\lambda^{-1}$-strongly convex, it then follows from Theorem 5.25(b) of [4] that for every $u \in \mathrm{dom}\, \hat{h}$,

$$\ell_f(u; x_{k-1}) + \hat{h}(u) + \frac{1}{2\lambda}\|u - x_{k-1}\|^2 \ge m_k + \frac{1}{2\lambda}\|u - x_k\|^2, \tag{102}$$

where $m_k = \ell_f(x_k; x_{k-1}) + \hat{h}(x_k) + \|x_k - x_{k-1}\|^2/(2\lambda)$. Using (5) with $(x, y) = (x_k, x_{k-1})$, we have

$$\hat{\phi}(x_k) - m_k = f(x_k) - \ell_f(x_k; x_{k-1}) \overset{(5)}{\le} 2M\|x_k - x_{k-1}\| - \frac{1}{2\lambda}\|x_k - x_{k-1}\|^2 \le 2\lambda M^2,$$

where the last inequality is due to Young's inequality $a^2 + b^2 \ge 2ab$. Hence, (101) follows from combining the above inequality and (102). ■

The next result presents the primal-dual convergence rate of MPDS$(x_0, \lambda)$.

**Lemma B.2.** *For every $k \ge 1$, define*

$$\bar{x}_k = \frac{1}{k}\sum_{i=1}^{k} x_i, \quad \bar{s}_k = \frac{1}{k}\sum_{i=1}^{k} s_i. \tag{103}$$

*Then, we have for every $k \ge 1$, the primal-dual gap of (99) is bounded as follows,*

$$\hat{\phi}(\bar{x}_k) + f^*(\bar{s}_k) + \hat{h}^*(-\bar{s}_k) \le 2\lambda M^2 + \frac{8d_0^2}{\lambda k}. \tag{104}$$

**Proof**: We first note that $\ell_f(\cdot; x_{k-1}) \le f$ and hence $(\ell_f(\cdot; x_{k-1}))^* \ge f^*$. Using this inequality and the fact that $\nabla \ell_f(u; x_{k-1}) = s_k$ for every $u \in \mathbb{R}^n$, we have

$$\ell_f(u; x_{k-1}) = -[\ell_f(\cdot; x_{k-1})]^*(s_k) + \langle s_k, u\rangle \le -f^*(s_k) + \langle s_k, u\rangle.$$

It thus follows from Lemma B.1 that for every $u \in \mathrm{dom}\, \hat{h}$,

$$\hat{\phi}(x_k) + f^*(s_k) - \langle s_k, u\rangle - \hat{h}(u) \overset{(101)}{\le} 2\lambda M^2 + \frac{1}{2\lambda}\|u - x_{k-1}\|^2 - \frac{1}{2\lambda}\|u - x_k\|^2.$$

Summing the above inequality from $k = 1$ to $k$ and using convexity of $\hat{\phi}$ and $f^*$, we obtain for every $u \in \operatorname{dom} \hat{h}$,

$$\hat{\phi}(\bar{x}_k) + f^*(\bar{s}_k) + \langle -\bar{s}_k, u \rangle - \hat{h}(u) \le 2\lambda M^2 + \frac{1}{2\lambda k}\|u - x_0\|^2,$$

where $\bar{x}_k$ and $\bar{s}_k$ are as in (103). Maximizing over $u \in \operatorname{dom} \hat{h}$ on both sides of the above inequality, we have

$$\hat{\phi}(\bar{x}_k) + f^*(\bar{s}_k) + \hat{h}^*(-\bar{s}_k) \le 2\lambda M^2 + \frac{\max\{\|u - x_0\|^2 : u \in \operatorname{dom} \hat{h}\}}{2\lambda k}.$$

Therefore, (104) follows by using the fact that $\operatorname{dom} \hat{h} \subset Q = B(x_0, 4d_0)$. ■

The following theorem provides the complexity of MPDS$(x_0, \lambda)$ for solving (99).

**Theorem B.1.** *Given $(x_0, \bar{\varepsilon}) \in \mathbb{R}^n \times \mathbb{R}_{++}$, letting $\lambda = \bar{\varepsilon}/(16M^2)$, then the number of iterations for MPDS$(x_0, \lambda)$ to generate a primal-dual pair $(\bar{x}_k, \bar{s}_k)$ as in (103) such that $\hat{\phi}(\bar{x}_k) + f^*(\bar{s}_k) + \hat{h}^*(-\bar{s}_k) \le \bar{\varepsilon}$ is at most $256M^2d_0^2/\bar{\varepsilon}^2$.*

**Proof**: It follows from Lemma B.2 with $\lambda = \bar{\varepsilon}/(16M^2)$ and $k = 16d_0^2/(\lambda\bar{\varepsilon})$ that

$$\hat{\phi}(\bar{x}_k) + f^*(\bar{s}_k) + \hat{h}^*(-\bar{s}_k) \le \frac{\bar{\varepsilon}}{8} + \frac{\bar{\varepsilon}}{2} < \bar{\varepsilon}.$$

Therefore, the conclusion of the theorem immediately follows from plugging the choice of $\lambda$ into $k$. ■

The next lemma gives the boundedness of $\{x_k\}$ generated by PDS$(x_0, \lambda)$ and shows that $\{x_k\} \subset Q = B(x_0, 4d_0)$. This result is important since it reveals the equivalence between PDS and MPDS, which is useful in Theorem B.2 below.

**Lemma B.3.** *For every $k \le 256M^2d_0^2/\bar{\varepsilon}^2$, the sequence $\{x_k\}$ generated by PDS$(x_0, \lambda)$ with $\lambda = \bar{\varepsilon}/(16M^2)$ satisfies $x_k \in Q$.*

**Proof**: Following an argument similar to the proof of Lemma B.1, we can prove for every $u \in \operatorname{dom} h$,

$$\phi(x_k) - \ell_f(u; x_{k-1}) - h(u) \le 2\lambda M^2 - \frac{1}{2\lambda}\|u - x_k\|^2 + \frac{1}{2\lambda}\|u - x_{k-1}\|^2,$$

which together with the fact that $\ell_f(\cdot; x_k) \le f$ implies that

$$\phi(x_k) - \phi(u) \le 2\lambda M^2 - \frac{1}{2\lambda}\|u - x_k\|^2 + \frac{1}{2\lambda}\|u - x_{k-1}\|^2.$$

Taking $u = x_0^*$ and using the fact that $\phi(x_k) \ge \phi_* = \phi(x_0^*)$, we obtain

$$\|x_k - x_0^*\|^2 \le 4\lambda^2M^2 + \|x_{k-1} - x_0^*\|^2.$$

Summing the above inequality, we show that for every $k \geq 1$, $\{x_k\}$ generated by PDS$(x_0, \lambda)$ satisfies

$$\|x_k - x_0^*\|^2 \leq d_0^2 + 4\lambda^2 M^2 k. \tag{105}$$

Using the triangle inequality and the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$, we have

$$\|x_k - x_0\| \leq \|x_k - x_0^*\| + \|x_0 - x_0^*\| \overset{(105)}{\leq} 2d_0 + 2\lambda M \sqrt{k}.$$

It thus follows from the assumptions on $k$ and $\lambda$ that

$$\|x_k - x_0\| \leq 2d_0 + \frac{\bar{\varepsilon}}{8M} \frac{16Md_0}{\bar{\varepsilon}} = 4d_0,$$

and hence that $x_k \in Q = B(x_0, 4d_0)$. ■

Finally, using the complexity of MPDS$(x_0, \lambda)$ for solving (99) (i.e., Theorem B.1), we are ready to establish that of PDS$(x_0, \lambda)$.

**Theorem B.2.** *Given $(x_0, \bar{\varepsilon}) \in \mathbb{R}^n \times \mathbb{R}_{++}$, letting $\lambda = \bar{\varepsilon}/(16M^2)$, then the number of iterations for PDS$(x_0, \lambda)$ to generate $(\bar{x}_k, \bar{s}_k)$ such that $\hat{\phi}(\bar{x}_k) + f^*(\bar{s}_k) + \hat{h}^*(-\bar{s}_k) \leq \bar{\varepsilon}$ is at most $256M^2 d_0^2/\bar{\varepsilon}^2$.*

**Proof**: In view of Lemma B.3, for $\lambda = \bar{\varepsilon}/(16M^2)$ and $k \leq 256M^2 d_0^2/\bar{\varepsilon}^2$, the sequence $\{x_k\}$ generated by PDS$(x_0, \lambda)$ is the same as the one generated by MPDS$(x_0, \lambda)$. Hence, sequences $\{s_k\}$ generated by the two methods are also the same, that is, (100) is identical to (7). Therefore, we conclude that the same primal-dual convergence guarantee holds for PDS$(x_0, \lambda)$ as the one for MPDS$(x_0, \lambda)$ in Theorem B.1. ■

# C Composite subgradient method for SPP

This section is devoted to the complexity analysis of CS-SPP. The main result is Theorem C.1 below.

**Lemma C.1.** *For every $k \geq 1$ and $(u, v) \in \mathbb{R}^n \times \mathbb{R}^m$, we have*

$$p_k(x_k) - \ell_{f(\cdot, y_{k-1})}(u; x_{k-1}) - h_1(u) \leq \delta_k^x + \frac{1}{2\lambda}\|x_{k-1} - u\|^2 - \frac{1}{2\lambda}\|x_k - u\|^2, \tag{106}$$

$$d_k(y_k) + \ell_{f(x_{k-1}, \cdot)}(v; y_{k-1}) - h_2(v) \leq \delta_k^y + \frac{1}{2\lambda}\|y_{k-1} - v\|^2 - \frac{1}{2\lambda}\|y_k - v\|^2, \tag{107}$$

*where*

$$\delta_k^x = 2M\|x_k - x_{k-1}\| - \frac{1}{2\lambda}\|x_k - x_{k-1}\|^2, \quad \delta_k^y = 2M\|y_k - y_{k-1}\| - \frac{1}{2\lambda}\|y_k - y_{k-1}\|^2. \tag{108}$$

**Proof**: We only prove (106) to avoid duplication. Inequality (107) follows similarly. Since the objective in (68) is $\lambda^{-1}$-strongly convex, we have for every $u \in \mathbb{R}^n$,

$$\ell_{f(\cdot,y_{k-1})}(u;x_{k-1}) + h_1(u) + \frac{1}{2\lambda}\|u - x_{k-1}\|^2 \geq m_k^x + \frac{1}{2\lambda}\|u - x_k\|^2, \tag{109}$$

where $m_k^x$ denotes the optimal value of (68). Using the definition of $p_k$ in (74), we have

$$p_k(x_k) - m_k^x = f(x_k, y_{k-1}) - \ell_{f(\cdot,y_{k-1})}(x_k; x_{k-1}) - \frac{1}{2\lambda}\|x_k - x_{k-1}\|^2.$$

It thus follows from the first inequality in (61) with $(u,x,y) = (x_k, x_{k-1}, y_{k-1})$ the definition of $\delta_k^x$ in (108) that

$$p_k(x_k) - m_k^x \leq \delta_k^x,$$

which together with (109) implies that (106). ■

For $k \geq 1$, denote

$$s_k = (s_k^x, s_k^y), \quad s_k^x = f'_x(x_{k-1}, y_{k-1}), \quad s_k^y = -f'_y(x_{k-1}, y_{k-1}). \tag{110}$$

We also denote $w = (u,v)$ and $z_k = (x_k, y_k)$ for all $k \geq 0$.

**Lemma C.2.** *For every $(u,v) \in \mathbb{R}^n \times \mathbb{R}^m$ and $k \geq 1$, we have*

$$\begin{aligned} & p_k(x_k) + f(\cdot, y_{k-1})^*(s_k^x) - h_1(u) + d_k(y_k) + [-f(x_{k-1},\cdot)]^*(s_k^y) - h_2(v) - \langle s_k, w\rangle \\ \leq & \delta_k^x + \delta_k^y + \frac{1}{2\lambda}\|z_{k-1} - w\|^2 - \frac{1}{2\lambda}\|z_k - w\|^2. \end{aligned} \tag{111}$$

**Proof**: It follows from the second identity in (110) that for every $u \in \mathbb{R}^n$,

$$\nabla \ell_{f(\cdot,y_{k-1})}(u; x_{k-1}) = s_k^x,$$

which together with Theorem 4.20 of [4] implies that

$$\ell_{f(\cdot,y_{k-1})}(u;x_{k-1}) + [\ell_{f(\cdot,y_{k-1})}(\cdot;x_{k-1})]^*(s_k^x) = \langle u, s_k^x\rangle.$$

Clearly, $\ell_{f(\cdot,y_{k-1})}(\cdot;x_{k-1}) \leq f(\cdot, y_{k-1})$ and hence $[\ell_{f(\cdot,y_{k-1})}(\cdot;x_{k-1})]^* \geq f(\cdot,y_{k-1})^*$. This inequality and the above identity imply that

$$\ell_{f(\cdot,y_{k-1})}(u;x_{k-1}) \leq -f(\cdot,y_{k-1})^*(s_k^x) + \langle s_k^x, u\rangle.$$

It thus follows from (106) that

$$p_k(x_k) + f(\cdot,y_{k-1})^*(s_k^x) - \langle s_k^x, u\rangle - h_1(u) \leq \delta_k^x + \frac{1}{2\lambda}\|x_{k-1} - u\|^2 - \frac{1}{2\lambda}\|x_k - u\|^2.$$

Similarly, we have for every $v \in \mathbb{R}^m$,

$$d_k(y_k) + [-f(x_{k-1},\cdot)]^*(s_k^y) - \langle s_k^y, v\rangle - h_2(v) \leq \delta_k^y + \frac{1}{2\lambda}\|y_{k-1} - v\|^2 - \frac{1}{2\lambda}\|y_k - v\|^2.$$

Finally, summing the above two inequalities and using (110) and the facts that $w = (u,v)$ and $z_k = (x_k, y_k)$, we conclude that (111) holds. ■

**Lemma C.3.** *For every $(u,v) \in \mathbb{R}^n \times \mathbb{R}^m$ and $k \geq 1$, we have*

$$\begin{aligned} & h_1(x_k) + f(\cdot, y_k)^*(s_k^x) - h_1(u) + h_2(y_k) + [-f(x_{k-1}, \cdot)]^*(s_k^y) - h_2(v) - \langle s_k, w\rangle \\ \leq & 16\lambda M^2 + \frac{1}{2\lambda}\|z_{k-1} - w\|^2 - \frac{1}{2\lambda}\|z_k - w\|^2. \end{aligned} \tag{112}$$

**Proof**: Using (74) and (111), we have for every $(u,v) \in \mathbb{R}^n \times \mathbb{R}^m$,

$$\begin{aligned} & h_1(x_k) + f(\cdot, y_{k-1})^*(s_k^x) - h_1(u) + h_2(y_k) + [-f(x_{k-1}, \cdot)]^*(s_k^y) - h_2(v) - \langle g_k, w\rangle \\ \leq & \delta_k^x + \delta_k^y + \frac{1}{2\lambda}\|z_{k-1} - w\|^2 - \frac{1}{2\lambda}\|z_k - w\|^2 + f(x_{k-1}, y_k) - f(x_k, y_{k-1}). \end{aligned} \tag{113}$$

It immediately follows from (60) that

$$\begin{aligned} f(x_{k-1}, y_k) - f(x_k, y_{k-1}) &= f(x_{k-1}, y_k) - f(x_k, y_k) + f(x_k, y_k) - f(x_k, y_{k-1}) \\ &\leq M\|x_k - x_{k-1}\| + M\|y_k - y_{k-1}\|. \end{aligned}$$

Following from the definition of conjugate functions and (60) again, we have

$$\begin{aligned} f(\cdot, y_{k-1})^*(s_k^x) &= \max_x\{\langle x, s_k^x\rangle - f(x, y_k) + f(x, y_k) - f(x, y_{k-1})\} \\ &\geq \max_x\{\langle x, s_k^x\rangle - f(x, y_k)\} - M\|y_k - y_{k-1}\| \\ &= f(\cdot, y_k)^*(s_k^x) - M\|y_k - y_{k-1}\|. \end{aligned}$$

Similarly, we also have

$$f(x_{k-1}, \cdot)^*(-s_k^y) \leq f(x_k, \cdot)^*(-s_k^y) + M\|x_k - x_{k-1}\|.$$

Plugging the above three inequalities into (113), we obtain for every $(u,v) \in \mathbb{R}^n \times \mathbb{R}^m$,

$$\begin{aligned} & h_1(x_k) + f(\cdot, y_k)^*(s_k^x) - h_1(u) + h_2(y_k) + [-f(x_{k-1}, \cdot)]^*(s_k^y) - h_2(v) - \langle s_k, w\rangle \\ \leq & \delta_k^x + \delta_k^y + \frac{1}{2\lambda}\|z_{k-1} - w\|^2 - \frac{1}{2\lambda}\|z_k - w\|^2 + 2M\|x_k - x_{k-1}\| + 2M\|y_k - y_{k-1}\|. \end{aligned}$$

Noting from the definitions in (108) that

$$\begin{aligned} & \delta_k^x + \delta_k^y + 2M\|x_k - x_{k-1}\| + 2M\|y_k - y_{k-1}\| \\ \overset{(108)}{=} & 4M\|x_k - x_{k-1}\| - \frac{1}{2\lambda}\|x_k - x_{k-1}\|^2 + 4M\|y_k - y_{k-1}\| - \frac{1}{2\lambda}\|y_k - y_{k-1}\|^2 \\ \leq & \ 16\lambda M^2, \end{aligned}$$

we finally conclude that (112) holds. ∎

The following lemma collects technical results revealing relationships about the averages defined in (114) below.

**Lemma C.4.** *Define*

$$\bar{x}_k = \frac{1}{k}\sum_{i=1}^{k} x_i, \quad \bar{y}_k = \frac{1}{k}\sum_{i=1}^{k} y_i, \quad \bar{s}_k^x = \frac{1}{k}\sum_{i=1}^{k} s_i^x, \quad \bar{s}_k^y = \frac{1}{k}\sum_{i=1}^{k} s_i^y. \tag{114}$$

*Then, the following statements hold for every $k \ge 1$:*

*(a)*

$$\frac{1}{k}\sum_{i=1}^{k} f(\cdot, y_i)^*(s_i^x) \ge f(\cdot, \bar{y}_k)^*(\bar{s}_k^x), \quad \frac{1}{k}\sum_{i=1}^{k}[-f(x_i,\cdot)]^*(s_i^y) \ge [-f(\bar{x}_k,\cdot)]^*(\bar{s}_k^y);$$

*(b)*

$$\begin{aligned} \varphi(\bar{x}_k) &\le h_1(\bar{x}_k) + [-f(\bar{x}_k,\cdot)]^*(\bar{s}_k^y) + h_2^*(-\bar{s}_k^y), \\ -\psi(\bar{y}_k) &\le h_2(\bar{y}_k) + f(\cdot,\bar{y}_k)^*(\bar{s}_k^x) + h_1^*(-\bar{s}_k^x). \end{aligned}$$

**Proof**: a) We only prove the first inequality to avoid duplication. The second one follows similarly. It follows from the definition of conjugate functions, (114), concavity of $f(x,\cdot)$, and basic inequalities that

$$\begin{aligned} \frac{1}{k}\sum_{i=1}^{k} f(\cdot, y_i)^*(s_i^x) &= \frac{1}{k}\sum_{i=1}^{k} \max_{x\in\mathbb{R}^n}\{\langle x, s_i^x\rangle - f(x, y_i)\} \\ &\ge \max_{x\in\mathbb{R}^n}\left\{\frac{1}{k}\sum_{i=1}^{k}\langle x, s_i^x\rangle - \frac{1}{k}\sum_{i=1}^{k} f(x,y_i)\right\} \\ &\overset{(114)}{\ge} \max_{x\in\mathbb{R}^n}\left\{\langle x, \bar{s}_k^x\rangle - f(x,\bar{y}_k)\right\} = f(\cdot,\bar{y}_k)^*(\bar{s}_k^x). \end{aligned}$$

b) For simplicity, we only prove the first inequality. The second one follows similarly. It follows from the definition of $\varphi$ in (65), basic inequalities, and the definition of conjugate functions that

$$\begin{aligned} \varphi(\bar{x}_k) &\overset{(65)}{=} \max_{y\in\mathbb{R}^m} \phi(\bar{x}_k, y) = h_1(\bar{x}_k) + \max_{y\in\mathbb{R}^m}\{f(\bar{x}_k, y) - h_2(y)\} \\ &\le h_1(\bar{x}_k) + \max_{y\in\mathbb{R}^m}\{\langle y, \bar{s}_k^y\rangle - (-f(\bar{x}_k, y))\} + \max_{y\in\mathbb{R}^m}\{\langle y, -\bar{s}_k^y\rangle - h_2(y)\} \\ &= h_1(\bar{x}_k) + [-f(\bar{x}_k,\cdot)]^*(\bar{s}_k^y) + h_2^*(-\bar{s}_k^y). \end{aligned}$$

■

**Proposition C.5.** *For every $k \geq 1$, we have*

$$\Phi(\bar{x}_k, \bar{y}_k) = \varphi(\bar{x}_k) - \psi(\bar{y}_k) \leq 16\lambda M^2 + \frac{D^2}{2\lambda k} \tag{115}$$

*where $\Phi(\cdot,\cdot)$ in as in (64).*

**Proof**: Summing (112) from $k = 1$ to $k$, and using Lemma C.4(a), convexity, and (114), we have for every $(u, v) \in \mathbb{R}^n \times \mathbb{R}^m$,

$$\begin{aligned}&h_1(\bar{x}_k) + f(\cdot, \bar{y}_k)^*(\bar{s}_k^x) - \langle \bar{s}_k^x, u\rangle - h_1(u) + h_2(\bar{y}_k) + [-f(\bar{x}_k, \cdot)]^*(\bar{s}_k^y) - \langle \bar{s}_k^y, v\rangle - h_2(v)\\ \leq &16\lambda M^2 + \frac{1}{2\lambda k}\|z_0 - w\|^2.\end{aligned}$$

Maximizing both sides of the above inequality over $(u, v) \in \operatorname{dom} h_1 \times \operatorname{dom} h_2$ yields

$$\begin{aligned}&h_1(\bar{x}_k) + f(\cdot, \bar{y}_k)^*(\bar{s}_k^x) + h_1^*(-\bar{s}_k^x) + h_2(\bar{y}_k) + [-f(\bar{x}_k, \cdot)]^*(\bar{s}_k^y) + h_2^*(-\bar{s}_k^y)\\ \leq &16\lambda M^2 + \frac{1}{2\lambda k}\max\{\|z_0 - w\|^2 : w \in \operatorname{dom} h_1 \times \operatorname{dom} h_2\}.\end{aligned}$$

Finally, (115) follows from Lemma C.4(b), (B3), and the definition of $\Phi(\cdot,\cdot)$ in (64). ∎

**Theorem C.1.** *Given $(x_0, y_0, \bar{\varepsilon}) \in \operatorname{dom} h_1 \times \operatorname{dom} h_2 \times R_{++}$, letting $\lambda = \bar{\varepsilon}/32M^2$, then the number of iterations of CS-SPP$(x_0, y_0, \lambda)$ to find a $\bar{\varepsilon}$-saddle-point $(\bar{x}_k, \bar{y}_k)$ of (2) is at most $128M^2D^2/\bar{\varepsilon}^2$.*

**Proof**: It follows from Proposition C.5 and the choice of $\lambda$ that

$$\Phi(\bar{x}_k, \bar{y}_k) \leq \frac{\bar{\varepsilon}}{2} + \frac{64D^2}{\bar{\varepsilon}k}.$$

Hence, the conclusion of the theorem follows immediately. ∎