

# Mathematical Programming

## A single cut proximal bundle method for stochastic convex composite optimization

--Manuscript Draft--

<b>Manuscript Number:</b>	
<b>Full Title:</b>	A single cut proximal bundle method for stochastic convex composite optimization
<b>Article Type:</b>	MPA - Full Length Paper
<b>Corresponding Author:</b>	Jiaming Liang Yale University UNITED STATES
<b>Corresponding Author Secondary Information:</b>	
<b>Corresponding Author's Institution:</b>	Yale University
<b>Corresponding Author's Secondary Institution:</b>	
<b>First Author:</b>	Jiaming Liang
<b>First Author Secondary Information:</b>	
<b>Order of Authors:</b>	Jiaming Liang Vincent Guigues Renato Monteiro
<b>Order of Authors Secondary Information:</b>	
<b>Funding Information:</b>	Air Force Research Laboratory (FA9550-22-1-0088)   Prof. Renato Monteiro
<b>Abstract:</b>	In this paper, we consider optimization problems where the objective is the sum of a function given by an expectation and a closed convex composite function. For such problems, we propose a stochastic composite proximal bundle (SCPB) method with optimal complexity. The method does not require estimation of parameters involved in the assumptions on the objective functions. Moreover, to the best of our knowledge, this is the first proximal bundle method for stochastic programming able to deal with continuous distributions. Finally, we present computational results showing that SCPB substantially outperforms the robust stochastic approximation method on all instances considered.

# A single cut proximal bundle method for stochastic convex composite optimization

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

July 18, 2022 (first revision: November 3, 2022)

## Abstract

In this paper, we consider optimization problems where the objective is the sum of a function given by an expectation and a closed convex composite function. For such problems, we propose a stochastic composite proximal bundle (SCPB) method with optimal complexity. The method does not require estimation of parameters involved in the assumptions on the objective functions. Moreover, to the best of our knowledge, this is the first proximal bundle method for stochastic programming able to deal with continuous distributions. Finally, we present computational results showing that SCPB substantially outperforms the robust stochastic approximation method on all instances considered.

**Keywords.** stochastic convex composite optimization, stochastic approximation, proximal bundle method, optimal complexity bound.

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

## 1 Introduction

The main goal of this paper is to propose and study the complexity of a stochastic composite proximal bundle (SCPB) framework to solve the stochastic convex composite optimization (SCCO) problem

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

where

$$f(x) = \mathbb{E}_\xi [F(x, \xi)]. \quad (2)$$

We assume the following conditions hold: i)  $f, h : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  are proper closed convex functions such that  $\text{dom } h \subseteq \text{dom } f$ ; ii) a stochastic first-order oracle, which for every  $x \in \text{dom } h$  and almost every random vector  $\xi$  returns  $s(x, \xi)$  such that  $\mathbb{E}[s(x, \xi)] \in \partial f(x)$ , is available; and iii) for every  $x \in \text{dom } h$ ,  $\mathbb{E}[\|s(x, \xi)\|^2] \leq \bar{M}^2$  for some  $\bar{M} \in \mathbb{R}_+$ .

**Literature Review.** Methods for solving (1) where  $f$  can efficiently be computed exactly (e.g.,  $\xi$  is a discrete random vector with small support) have been discussed for example in [2, 3]

---

<sup>\*</sup>Department of Computer Science, Yale University, New Haven, CT 06511 (email: [jiaming.liang@yale.edu](mailto:jiaming.liang@yale.edu)).

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

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

1  
2  
3  
4  
5  
6  
7  
8 and are usually based on solving a deterministic (but large-scale) reformulation of (1), usually  
9 using decomposition (such as the L-shaped method [23]) possibly combined with regularization as  
10 in [9, 8].

11 Solution methods for problem (1) in which  $\xi$  has a continuous distribution are basically based  
12 on one of the following three ideas: i) a single (usually expensive) approximation of (1) where  
13  $f$  is approximated by a Monte Carlo average  $H_N(x) := \sum_{i=1}^N F(\cdot, \xi_i)/N$  for a large i.i.d. sample  
14  $(\xi_1, \dots, \xi_N)$  of  $\xi$  is constructed at the beginning of the method and is then solved to yield an  
15 approximation solution of (1) (SAA type methods); see for instance [24, 11, 21, 10, 7] and also  
16 Chapter 5 of [22] for their complexity analysis; ii) simple approximations of (1) are constructed at  
17 every iteration based on a small (usually a single) sample and their solutions are used to obtain an  
18 approximation solution of (1) (SA type methods); SA type methods have been originally proposed  
19 in [20] and further extended in [15, 18, 19, 17, 14, 5, 6]; and iii) hybrid type methods which sit in  
20 between SAA and SA type ones in that they use partial Monte Carlo averages  $H_k(\cdot)$  (and their  
21 expensive subgradients) for increasing iteration indices  $k$  [9].

22 **Contributions.** Although cutting plane methodology can be used in the context of SAA  
23 methods to solve single approximations of (1) generated at their outset, such methodology has not  
24 been used in the context of SA type methods. This paper partially addresses this issue by developing  
25 a regularized aggregate cutting plane scheme for solving (1) where some of the most recent (linear  
26 approximation) cuts (in expectation) are combined, i.e., a suitable convex combination of them is  
27 chosen, so that a single aggregated cut (in expectation) is obtained. More specifically, two proximal  
28 (aggregated one cut) bundle variants are proposed which can be viewed as natural extensions of  
29 the one-cut variant developed in [13] (based on the analysis of [12]) for solving the deterministic  
30 version of (1). At every iteration, the SCPB framework solves the prox bundle subproblem

$$31 \quad x = \operatorname{argmin}_{u \in \mathbb{R}^n} \left\{ \Gamma(u) + \frac{1}{2\lambda} \|u - x^c\|^2 \right\} \quad (3)$$

32 where  $\lambda > 0$  is the prox stepsize,  $x^c$  is the current prox-center, and  $\Gamma$  is the current bundle function  
33 in expectation, i.e., it satisfies  $\mathbb{E}[\Gamma(\cdot)] \leq \phi(\cdot)$ . As in the proximal bundle method of [13], it also  
34 performs two types of iterations, i.e., serious or null ones. In the beginning of a serious iteration,  
35 the prox-center is updated to  $x^c \leftarrow x$  and the bundle function  $\Gamma$  is chosen to be the composite  
36 linear approximation of  $F(\cdot, \xi) + h$  at  $x$  given by  $F(x, \xi) + \langle s(x, \xi), \cdot - x \rangle + h(\cdot)$  for some new  
37 independent sample  $\xi$ . In a null iteration, the prox-center remains the same but  $\Gamma$  is set to be a  
38 convex combination of the previous bundle function and the above composite linear approximation  
39 of  $F(\cdot, \xi) + h(\cdot)$  at  $x$ . It is then shown that both variants of SCPB obtain a stochastic iterate  
40  $y \in \mathbb{R}^n$  (determined by some of the above generated  $x$ 's) such that  $\mathbb{E}[\phi(y)] - \phi_* \leq \varepsilon$  where  $\phi_*$  is  
41 as in (1), in  $\mathcal{O}(\varepsilon^{-2})$  iterations/resolvent evaluations. To our knowledge, this is the first SA type  
42 proximal bundle method for SCCO problems where  $\xi$  can have a continuous distribution. Finally,  
43 it is shown that the the robust stochastic approximation (RSA) method of [14] is a special case of  
44 SCPB with a relatively small prox stepsize.

45 **Organization of the paper.** Subsection 1.1 presents basic definitions and notation used  
46 throughout the paper. Section 2 formally describes the assumptions on the SCCO problem (1),  
47 presents the SCPB framework, and two cycles rules for determining the length of the cycles in  
48 SCPB. Section 3 presents various convergence rate bounds for the SCPB variant based on the first  
49 cycle rule and discusses the relationship between SCPB and RSA. Section 4 provides convergence  
50 rate bounds for the SCPB variant based on the second cycle rule. Section 4 collects proofs of the  
51 main results in Sections 3 and 4. Section 6 reports the numerical experiments. Finally, Section 7  
52 presents some concluding remarks and possible extensions.

## 1.1 Basic definitions and notation

Let  $\mathbb{N}_{++}$  denote the set of positive integers. The sets of real numbers, non-negative and positive real numbers are denoted by  $\mathbb{R}$ ,  $\mathbb{R}_+$  and  $\mathbb{R}_{++}$ , respectively. Let  $\mathbb{R}^n$  denote the standard  $n$ -dimensional Euclidean space equipped with inner product and norm denoted by  $\langle \cdot, \cdot \rangle$  and  $\|\cdot\|$ , respectively.

Let  $\psi : \mathbb{R}^n \rightarrow (-\infty, +\infty]$  be given. The effective domain of  $\psi$  is denoted by  $\text{dom } \psi := \{x \in \mathbb{R}^n : \psi(x) < \infty\}$  and  $\psi$  is proper if  $\text{dom } \psi \neq \emptyset$ . Moreover, a proper function  $\psi : \mathbb{R}^n \rightarrow (-\infty, +\infty]$  is  $\mu$ -convex for some  $\mu \geq 0$  if

$$\psi(\alpha z + (1 - \alpha)u) \leq \alpha\psi(z) + (1 - \alpha)\psi(u) - \frac{\alpha(1 - \alpha)\mu}{2}\|z - u\|^2$$

for every  $z, u \in \text{dom } \psi$  and  $\alpha \in [0, 1]$ . The set of all proper lower semicontinuous convex functions  $\psi : \mathbb{R}^n \rightarrow (-\infty, +\infty]$  is denoted by  $\overline{\text{Conv}}(\mathbb{R}^n)$ . For  $\varepsilon \geq 0$ , the  $\varepsilon$ -subdifferential of  $\psi$  at  $z \in \text{dom } \psi$  is denoted by  $\partial_\varepsilon \psi(z) := \{s \in \mathbb{R}^n : \psi(u) \geq \psi(z) + \langle s, u - z \rangle - \varepsilon, \forall u \in \mathbb{R}^n\}$ . The subdifferential of  $\psi$  at  $z \in \text{dom } \psi$ , denoted by  $\partial\psi(z)$ , is by definition the set  $\partial_0\psi(z)$ . We use the notation  $\xi_{[t]} = (\xi_0, \xi_1, \dots, \xi_t)$  for the history of the sampled observations of  $\xi$  up to iteration  $t$ . Define  $\ln_0^+(\cdot) := \max\{0, \ln(\cdot)\}$ . Define the diameter of a set  $X$  to be  $D_X := \sup\{\|x - x'\| : x, x' \in \text{dom } X\}$ .

## 2 Assumptions and the SCPB framework

This section presents the assumptions made on problem (1) and states the SCPB framework for solving it.

### 2.1 Assumptions

Let  $\Xi$  denote the support of random vector  $\xi$  and assume that the following conditions on (1)-(2) are assumed to hold:

(A1)  $f, h \in \overline{\text{Conv}}(\mathbb{R}^n)$  are such that  $\text{dom } f \supset \text{dom } h$ ;

(A2) for almost every  $\xi \in \Xi$ , a functional oracle  $F(\cdot, \xi) : \text{dom } h \rightarrow \mathbb{R}$  and a stochastic subgradient oracle  $s(\cdot, \xi) : \text{dom } h \rightarrow \mathbb{R}^n$  satisfying

$$f(x) = \mathbb{E}[F(x, \xi)], \quad f'(x) := \mathbb{E}[s(x, \xi)] \in \partial f(x)$$

for every  $x \in \text{dom } h$  are available;

(A3)  $\bar{M} := \sup\{\mathbb{E}[\|s(x, \xi)\|^2]^{1/2} : x \in \text{dom } h\} < \infty$ ;

(A4) the set of optimal solutions  $X^*$  of (1)-(2) is nonempty.

We now make some observations about the above conditions. First, as in [14], condition (A2) does not require  $F(\cdot, \xi)$  to be convex. Second, condition (A3) implies that

$$\|f'(x)\| = \|\mathbb{E}[s(x, \xi)]\| \leq \mathbb{E}[\|s(x, \xi)\|] \leq (\mathbb{E}[\|s(x, \xi)\|^2])^{1/2} \leq \bar{M} \quad \forall x \in \text{dom } h. \quad (4)$$

Third, defining for every  $\xi \in \Xi$  and  $x \in \text{dom } h$ ,

$$\Phi(\cdot, \xi) = F(\cdot, \xi) + h(\cdot), \quad \ell(\cdot; x, \xi) = f(x) + \langle s(x, \xi), \cdot - x \rangle + h(\cdot), \quad (5)$$

it follows from (A2), the second identity in (5), and the convexity of  $f$  by (A1), that

$$\mathbb{E}[\Phi(\cdot, \xi)] = \phi(\cdot) \geq f(x) + \langle f'(x), \cdot - x \rangle + h(\cdot) = \mathbb{E}[\ell(\cdot; x, \xi)] \quad (6)$$

where  $\phi(\cdot)$  is as in (1).

## 2.2 The SCPB framework

The SCPB framework is given below. At every iteration  $j \geq 1$ , the SCPB framework samples an independent realization  $\xi_{j-1}$  of  $\xi$ .

---

SCPB

---

0. Let  $\lambda, \theta > 0$ , integer  $K \geq 1$ , and  $x_0 \in \text{dom } h$  be given, and set  $j = k = 1$ ,  $j_0 = 0$ , and

$$\tau = \frac{\theta K}{\theta K + 1}; \quad (7)$$

1. take a sample  $\xi_{j-1}$  of r.v.  $\xi$  independent from the previous samples  $\xi_0, \dots, \xi_{j-2}$  and compute

$$x_j^c = \begin{cases} x_{j_{k-1}}, & \text{if } j = j_{k-1} + 1, \\ x_{j-1}^c, & \text{otherwise,} \end{cases} \quad (8)$$

$$S_j = \begin{cases} s(x_{j_{k-1}}, \xi_{j_{k-1}}), & \text{if } j = j_{k-1} + 1, \\ (1 - \tau)s(x_{j-1}, \xi_{j-1}) + \tau S_{j-1}, & \text{otherwise,} \end{cases} \quad (9)$$

$$x_j = \underset{u \in \mathbb{R}^n}{\text{argmin}} \left\{ h(u) + \langle S_j, u \rangle + \frac{1}{2\lambda} \|u - x_j^c\|^2 \right\}, \quad (10)$$

and

$$y_j = \begin{cases} x_j, & \text{if } j = j_{k-1} + 1, \\ (1 - \tau)x_j + \tau y_{j-1}, & \text{otherwise;} \end{cases} \quad (11)$$

2. if  $j = j_{k-1} + 1$ , then choose an integer  $j_k$  such that

$$j_k \geq j_{k-1} + 1;$$

if  $j < j_k$ , then set  $j \leftarrow j + 1$  and go to Step 1; else, set  $\hat{y}_k = y_{j_k}$ , and go to Step 3;

3. if  $k = K$  then **stop** and output

$$\hat{y}_K^a = \frac{1}{\lceil K/2 \rceil} \sum_{k=\lceil K/2 \rceil+1}^K \hat{y}_k; \quad (12)$$

otherwise, set  $k \leftarrow k + 1$  and  $j \leftarrow j + 1$ , and go to Step 1.

---

We first discuss the roles played by the two index counts  $j$  and  $k$  used by SCPB. First,  $j$  counts the total number of iterations/resolvent evaluations performed by SCPB since it is increased by one every time SCPB returns to Step 1. Second, defining the  $k$ -th cycle as the iteration indices  $j$  lying in

$$\mathcal{C}_k := \{i_k, \dots, j_k\}, \quad \text{where } i_k := j_{k-1} + 1, \quad (13)$$

it immediately follows that  $k$  counts the number of cycles generated by SCPB. Third, Step 1 determines two types of iterations depending on whether  $j = i_k$  (serious iteration) or  $j \in \mathcal{C}_k \setminus \{i_k\}$  (null iteration). Hence, the first iteration of a cycle is a serious one while the others are null ones.

We now make several basic remarks about SCPB. First, every execution of Step 1 involves one resolvent evaluation of  $\partial h$ , i.e., an evaluation of the point-to-point operator  $(I + \alpha\partial h)^{-1}(\cdot)$  for some  $\alpha > 0$ . Second, SCPB generates three sequences of iterates, namely, the sequence of prox centers  $\{x_j^c\}$  computed in (8), the auxiliary sequence  $\{x_j\}$  determined by (10), and the sequence  $\{y_j\}$  given by (11). Third, it follows from (8) that  $x_j^c = x_{j_{k-1}}$  for every  $j \in \mathcal{C}_k$ . Hence, the prox-center  $x_j^c$  remains constant within every cycle and can only change at the beginning of its only serious iteration. Fourth,  $\{\hat{y}_k\}$  is the subsequence of  $\{y_j\}$  consisting of all last iterates of the cycles generated by SCPB. Fifth, the convergence rates described in Theorems 3.1 and 4.1 for SCPB are with respect to the average of the iterates  $\hat{y}_{\lfloor K/2 \rfloor + 1}, \dots, \hat{y}_K$ , namely, the point  $\hat{y}_K^a$  as in (12). Finally, as currently stated, SCPB is not a completely specified algorithm since Step 2 does not describe how to select  $j_k$ .

### 2.3 Two cycle rules for SCPB

We now describe two ways of determining  $j_k$  in step 2 of SCPB, namely:

(B1) for every  $k \geq 1$ ,  $j_k$  is the smallest integer  $\geq i_k$  such that  $\lambda k \tau^{j_k - i_k} \leq R$ .

(B2) for every  $k \geq 1$ ,  $j_k$  is the smallest integer  $\geq i_k + 1$  such that

$$\lambda k \tau^{j_k - i_k} \left( F(x_{i_k}, \xi_{i_k}) - \tilde{\ell}_k(x_{i_k}) - \frac{1}{2\lambda} \|x_{i_k} - x_{i_k}^c\|^2 \right) \leq R \quad (14)$$

where  $i_k$  is as in (13) and

$$\tilde{\ell}_k(\cdot) := F(x_{i_{k-1}}, \xi_{i_{k-1}}) + \langle s(x_{i_{k-1}}, \xi_{i_{k-1}}), \cdot - x_{i_{k-1}} \rangle. \quad (15)$$

We make the following remarks about cycle rules (B1) and (B2). First, the sequence  $\{j_k\}$  determined by the cycle rule (B1) is deterministic, while the one determined by (B2) is stochastic since the sequence  $\{x_{i_k}\}$  used in (14) is stochastic. Second, another difference between the two cycle rules is that (B1) allows  $j_k = i_k$ , while  $j_k$  in (B2) is at least  $i_k + 1$ . In other words, the cycle length for (B1) may be equal to one, but the one for (B2) is at least two.

## 3 Results of the first variant of SCPB

This section presents the main results of SCPB based on cycle rule (B1) under various assumptions and discusses the relationship between SCPB and RSA.

### 3.1 Convergence rate bounds of SCPB with bounded $\text{dom } h$

The following result states a general convergence rate result for SCPB, based on cycle rule (B1), which holds for bounded  $\text{dom } h$  and for any choice of input  $(\lambda, \theta, K)$  in SCPB and constant  $R$  as in (B1). The proof is postponed to Subsection 5.1.

**Theorem 3.1.** *Assume that conditions (A1)-(A4) hold and  $\text{dom } h$  has a finite diameter  $D_h \geq 0$ . Then, for any given  $(\lambda, \theta, K) \in \mathbb{R}_{++}^2 \times \mathbb{N}_{++}$  and  $R > 0$ , SCPB with any input  $(\lambda, \theta, K)$  based on cycle rule (B1) with constant  $R$  satisfies the following statements:*

a) the number of iterations within the  $k$ -th cycle  $\mathcal{C}_k$  (see (13)) is bounded by

$$\left\lceil (\theta K + 1) \ln_0^+ \left( \frac{\lambda k}{R} \right) \right\rceil + 1; \quad (16)$$

1  
2  
3  
4  
5  
6  
7  
8 b) we have

$$\mathbb{E}[\phi(\hat{y}_K^a)] - \phi_* \leq \frac{1}{K} \left( \frac{D_h^2}{\lambda} + \frac{6R \min\{\lambda \bar{M}^2, \bar{M} D_h\}}{\lambda} + \frac{2\lambda \bar{M}^2}{\theta} \right) \quad (17)$$

11 where  $D_h$  is the diameter of  $\text{dom } h$ .

14 We now make some remarks about Theorem 3.1. First, its overall iteration complexity is given by  $K$ , which is its outer iteration complexity, times its inner iteration complexity given in (16). Second, (17) gives a bound on the expected primal gap  $\mathbb{E}[\phi(\hat{y}_K^a)] - \phi_*$  in terms of  $K$ , and hence provides a sufficient condition on how large  $K$  should be chosen for SCPB to generate a desired approximate solution.

20 For any given  $(\lambda, K)$ , the following result describes a convergence rate bound for SCPB with a specific choice of  $(\theta, R)$ .

23 **Corollary 3.2.** *Assume that conditions (A1)-(A4) hold and  $\text{dom } h$  has a finite diameter  $D_h > 0$ . Let a pair  $(\lambda, K)$  be given and consider SCPB based on cycle rule (B1) with input  $(\lambda, \theta, K)$  and constant  $R$  in (B1) satisfying*

$$\theta = \frac{2\lambda^2 M^2}{D^2}, \quad R = \frac{D}{6M} \quad (18)$$

27 where  $(D, M)$  is an estimate for the (usually unknown) pair  $(D_h, \bar{M})$ . Then, the following statements hold:

32 a) we have

$$\mathbb{E}[\phi(\hat{y}_K^a)] - \phi_* \leq \frac{3D^2}{2\lambda K} (\kappa_D + \kappa_M)$$

35 where

$$\kappa_D := \frac{D_h^2}{D^2}, \quad \kappa_M := \frac{\bar{M}^2}{M^2}; \quad (19)$$

38 b) its expected overall iteration complexity (up to a logarithmic term) is

$$\mathcal{O} \left( \frac{\lambda^2 M^2 K^2}{D^2} + K \right). \quad (20)$$

46 **Proof:** a) Using (17), the definitions of  $\kappa_D$  and  $\kappa_M$  in (19), and the definitions of  $\theta$  and  $R$  in (18), we get

$$\begin{aligned} \mathbb{E}[\phi(\hat{y}_K^a)] - \phi_* &\leq \frac{1}{K} \left( \frac{\kappa_D D^2}{\lambda} + \frac{D \min\{\lambda \bar{M}^2, \bar{M} D_h\}}{\lambda M} + \frac{\kappa_M D^2}{\lambda} \right) \\ &\leq \frac{D^2}{\lambda K} (\kappa_D + \kappa_M + \sqrt{\kappa_D \kappa_M}) \\ &\leq \frac{3D^2}{2\lambda K} (\kappa_D + \kappa_M), \end{aligned}$$

56 where in the second inequality we have used  $\min\{\lambda \bar{M}^2, \bar{M} D_h\} \leq \bar{M} D_h$  and the definitions of  $\kappa_D$  and  $\kappa_M$  while in the last inequality we have used the relation  $\sqrt{ab} \leq (a+b)/2$  for every  $a, b \geq 0$ .

58 b) It follows from Theorem 3.1(a) that the overall complexity (up to a logarithmic term) is  $\mathcal{O}(\theta K^2 + K)$ , which in turn is (20) in view of  $\theta$  as in (18). ■

61 For a given tolerance  $\varepsilon > 0$ , we now consider a specialization of the SCPB variant of Corollary 3.2 with  $K$  chosen so that  $x = \hat{y}_K^a$  is a  $\varepsilon$ -solution of (1), i.e., it satisfies  $\mathbb{E}[\phi(x)] - \phi_* \leq \varepsilon$ , and show

that it has optimal overall iteration-complexity for a large range of prox stepsizes. Indeed, setting  $K = \lceil T_\varepsilon \rceil$  where

$$T_\varepsilon := \frac{3D^2}{2\lambda\varepsilon} (\kappa_D + \kappa_M),$$

it follows from the above result that  $\mathbb{E}[\phi(\hat{y}_K^a)] - \phi_* \leq \varepsilon$ . Since  $K \leq T_\varepsilon + 1$ , we conclude from (20) that the expected overall iteration-complexity of SCPB is bounded by

$$\mathcal{O}\left(\frac{\lambda^2 M^2 (T_\varepsilon + 1)^2}{D^2} + T_\varepsilon + 1\right) = \mathcal{O}\left(\frac{M^2 D^2}{\varepsilon^2} [\kappa_D^2 + \kappa_M^2] + \frac{\lambda^2 M^2}{D^2} + \frac{D^2}{\lambda\varepsilon} [\kappa_D + \kappa_M] + 1\right).$$

In particular, if  $D \geq D_h$  and  $M \geq \bar{M}$ , or equivalently,  $\kappa_D \leq 1$  and  $\kappa_M \leq 1$ , then the above complexity reduces to

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

Moreover, under the assumption that the prox stepsize  $\lambda$  lies in the interval  $[\varepsilon/M^2, D^2/\varepsilon]$ , the above complexity bound further reduces to  $\mathcal{O}(M^2 D^2/\varepsilon^2)$ , which is known to be the optimal complexity of finding an  $\varepsilon$ -solution for any instance of (1) such that its corresponding pair  $(D_h, \bar{M})$  satisfies the condition that  $D \geq D_h$  and  $M \geq \bar{M}$  (e.g., see [16]).

### 3.2 Relationship between SCPB and the RSA method of [14]

Recall that the RSA method of [14], which is developed under the assumption that  $h$  is the indicator function of a nonempty closed convex set  $X$ , with a given initial point  $x_0 \in X$  and constant prox stepsize  $\lambda > 0$  recursively computes its iteration sequence  $\{x_j\}_{j=1}^N$  according to

$$x_j = \operatorname{argmin}_{u \in X} \left\{ \langle s(x_{j-1}, \xi_{j-1}), u \rangle + \frac{1}{2\lambda} \|u - x_{j-1}\|^2 \right\} \quad \forall j = 1, \dots, N. \quad (21)$$

For  $1 \leq i \leq N$ , letting  $\tilde{x}_i^N$  denote the average of  $\{x_j\}_{j=i}^N$ , i.e.,

$$\tilde{x}_i^N = \frac{1}{N-i+1} \sum_{j=i}^N x_j, \quad (22)$$

it is shown in (2.24) of [14] that if, for some fixed scalar  $\alpha > 0$ ,  $\lambda$  is chosen as

$$\lambda = \frac{\alpha D}{M\sqrt{N}} \quad (23)$$

where  $N$  is the specified number of iterations in (21) and  $D$  (resp.,  $M$ ) is an upper bound on the diameter of  $X$  (resp.,  $\bar{M}$ ), then the ergodic iterate  $\tilde{x}_i^N$  as in (22) satisfies

$$\mathbb{E}[\phi(\tilde{x}_i^N)] - \phi_* \leq \max\{\alpha, \alpha^{-1}\} \frac{DM}{\sqrt{N}} \left( \frac{2N}{N-i+1} + \frac{1}{2} \right).$$

In the case of  $i = \lfloor N/2 \rfloor + 1$ , the above convergence rate bound becomes

$$\mathbb{E}[\phi(\tilde{x}_{\lfloor N/2 \rfloor + 1}^N)] - \phi_* \leq \max\{\alpha, \alpha^{-1}\} \frac{9DM}{2\sqrt{N}}. \quad (24)$$



Hence, for a given tolerance  $\varepsilon > 0$ , (24) implies that the overall iteration complexity of RSA to obtain  $\tilde{x}_{\lfloor N/2 \rfloor + 1}^N$  such that  $\mathbb{E}[\phi(\tilde{x}_{\lfloor N/2 \rfloor + 1}^N)] - \phi_* \leq \varepsilon$  is

$$\mathcal{O}\left(\frac{\max\{\alpha^2, \alpha^{-2}\}M^2D^2}{\varepsilon^2}\right). \quad (25)$$

It turns out that RSA is a special case of SCPB based on cycle rule (B1) with  $R$  given by

$$R = \frac{\alpha D \sqrt{K}}{M}.$$

Indeed, it follows from the above choice of  $R$  and  $\lambda$  as in (23) with  $N$  replaced by  $K$  that

$$\frac{R}{\lambda k} \geq \frac{R}{\lambda K} = 1$$

and hence that  $j_k = i_k$  satisfies (B1). Thus, every cycle only performs one iteration, i.e., its only serious iteration. Moreover, every iteration of this SCPB variant is a serious one and  $K$  is its total number of iterations.

### 3.3 A practical SCPB variant

From a computational point of view, the choice of  $\theta$  in Corollary 3.2 usually results in the quantity  $\theta K$  and hence the inner complexity bound (16), being large. The following result provides a practical variant of SCPB with an alternative choice for  $\theta$  and  $R$  which partially remedies the above drawback by forcing  $\theta K$  to be constant. A nice feature of this variant is that it is able to choose large prox stepsizes without losing the optimality of its overall iteration complexity

**Corollary 3.3.** *Assume that conditions (A1)-(A4) hold and  $\text{dom } h$  has a finite diameter  $D_h > 0$ . Let positive integer  $K$  and constant  $C \geq 1$  be given, and define*

$$\theta = \frac{C}{K}, \quad R = \frac{D}{M}, \quad \lambda = \frac{\sqrt{CD}}{M\sqrt{K}} \quad (26)$$

where  $(D, M)$  is an estimate for the pair  $(D_h, \bar{M})$  such that  $M \geq \bar{M}$  and  $D \geq D_h$ . Then, the following statements about SCPB with input  $(\lambda, \theta, K)$  based on cycle rule (B1) with  $\theta$ ,  $\lambda$ , and  $R$  as above hold:

a) we have

$$\mathbb{E}[\phi(\hat{y}_K^a)] - \phi_* \leq \frac{9DM}{\sqrt{CK}}; \quad (27)$$

b) the number of iterations within the  $k$ -th cycle  $\mathcal{C}_k$  is bounded by

$$\left\lceil (C+1) \ln_0^+ \left( \frac{\sqrt{Ck}}{\sqrt{K}} \right) \right\rceil + 1,$$

and hence, up to a logarithmic term, is  $\mathcal{O}(C)$ ;

c) its expected overall iteration complexity, up to a logarithmic term, is  $\mathcal{O}(CK)$ .

**Proof:** a) Using (17), the definitions of  $\kappa_D$  and  $\kappa_M$  in (19), and the definitions of  $\theta$  and  $R$  in (26), we get

$$\begin{aligned} \mathbb{E}[\phi(\hat{y}_K^a)] - \phi_* &\leq \frac{1}{K} \left( \frac{\kappa_D D^2}{\lambda} + \frac{6D \min\{\lambda \bar{M}^2, \bar{M} D_h\}}{\lambda M} + \frac{2\lambda K \bar{M}^2}{C} \right) \\ &\leq \frac{1}{\lambda K} \left( \kappa_D D^2 + \frac{6D \bar{M} D_h}{M} \right) + \frac{2\lambda M^2 \kappa_M}{C} \\ &= \frac{D^2}{\lambda K} (\kappa_D + 6\sqrt{\kappa_M \kappa_D}) + \frac{2\lambda M^2 \kappa_M}{C}, \end{aligned} \quad (28)$$

where in the second inequality we have used  $\min\{\lambda \bar{M}^2, \bar{M} D_h\} \leq \bar{M} D_h$  and the definition of  $\kappa_M$ . It follows from (28) and the assumptions that  $M \geq \bar{M}$  and  $D \geq D_h$  that

$$\mathbb{E}[\phi(\hat{y}_K^a)] - \phi_* \leq \frac{7D^2}{\lambda K} + \frac{2\lambda M^2}{C}.$$

Finally, the above bound with  $\lambda$  as in (26) implies (27).

b) This statement immediately follows from Theorem 3.1(a) with  $\theta$ ,  $R$ , and  $\lambda$  as in (26).

c) This statement follows from (b) and the fact that SCPB has  $K$  cycles.  $\blacksquare$

The following paragraphs make some remarks about the practical SCPB variant of Corollary 3.3.

First, although  $\theta$  in (26) depends neither on  $M$  nor  $D$ , the choice of  $R$  depends on both of these estimates. A variant of SCPB based on cycle rule (B2) will be analysed in Section 4 where  $\theta$  depends neither on  $M$  nor  $D$ , and  $R$  depends on  $D$  but not  $M$ .

Second, Corollary 3.3 (see its statement (b)) implies that the number of iterations within a cycle of its SCPB variant is bounded (up to a logarithmic term) by the a priori (user specified) constant  $C$ . Thus, the SCPB variant of Corollary 3.3 can be viewed as an extended version of RSA where the number of iterations within a cycle can be larger than one, instead of being equal to one as in RSA (see the discussion in the second paragraph of Subsection 3.2).

Third, if  $K$  is chosen as

$$K = \left\lceil \frac{81D^2 M^2}{C\varepsilon^2} \right\rceil,$$

then Corollary 3.3(a) implies that  $\hat{y}_K^a$  is an  $\varepsilon$ -solution of (1) and Corollary 3.3(c) implies that the overall iteration complexity of SCPB is  $\mathcal{O}(M^2 D^2 / \varepsilon^2)$ . In conclusion, SCPB with the above choice of  $K$  is able to choose large prox stepsizes without losing the optimality of its overall iteration complexity for finding an  $\varepsilon$ -solution of (1).

Fourth, it is interesting to compare the behavior of RSA and the SCPB variant of Corollary 3.3 for find an  $\varepsilon$ -solution of (1) when both choose the same prox stepsize  $\lambda$  as in (26) where  $C \geq 1$  is a possibly large scalar. (For simplicity, we assume as in Subsection 3.2 that  $h$  is the indicator function of a nonempty closed convex set.) Indeed, it follows from (25) with  $\alpha = \sqrt{C}$  that RSA with  $\lambda$  chosen according to (26) has overall iteration complexity given by

$$\mathcal{O} \left( \frac{CD^2 M^2}{\varepsilon^2} \right).$$

Thus, while both RSA and the SCPB variant of Corollary 3.3 converge for any prox stepsize  $\lambda$  as in (26), the overall iteration complexity of RSA becomes worse as  $C$  grows while the one for SCPB does not depend on  $C$  and remains equal to the optimal complexity bound for the class of all instances of (1) such that their corresponding pair  $(D_h, \bar{M})$  satisfies the condition that  $D \geq D_h$  and  $M \geq \bar{M}$ .

Fifth, although the SCPB variant of Corollary 3.3 chooses  $\lambda$  as in (26), our numerical experiments choose a more aggressive prox stepsize, i.e.,

$$\lambda = \beta_1 \frac{\sqrt{C}D}{M\sqrt{K}}$$

where  $\beta_1 = 10$ . It is interesting that SCPB with this aggressive choice of  $\lambda$  substantially outperforms RSA on the (relatively small number of) instances considered in our experiment.

### 3.4 Convergence rate bounds of SCPB with unbounded $\text{dom } h$

Finally, we end this section by presenting convergence guarantees for a slightly modified variant without assuming  $\text{dom } h$  is bounded. The following result is an analogue of Theorem 3.1 under the assumption that  $\text{dom } h$  is not necessarily bounded. We omit its proof since it is similar to that of Theorem 3.1.

We consider a modified variant of SCPB where the output (12) is replaced by

$$\bar{y}_K^a = \frac{1}{K} \sum_{k=1}^K \hat{y}_k.$$

We also consider a modified cycle rule of (B1) as follows:

(B1') for every  $k \geq 1$ ,  $j_k$  is the smallest integer  $\geq i_k$  such that  $\lambda K \tau^{j_k - i_k} \leq R$ .

**Theorem 3.4.** *Assume that conditions (A1)-(A4) hold and let  $d_0$  denote the distance of the initial point  $x_0$  to optimal set  $X^*$ , i.e.,*

$$d_0 := \|x_0 - x_0^*\|, \quad \text{where } x_0^* := \operatorname{argmin} \{\|x_0 - x^*\| : x^* \in X^*\}. \quad (29)$$

*Then, SCPB based on cycle rule (B1') satisfies the following statements:*

a) *the number of iterations within each cycle is bounded by*

$$\left\lceil (\theta K + 1) \ln_0^+ \left( \frac{\lambda K}{R} \right) \right\rceil + 1;$$

b) *we have*

$$\mathbb{E}[\phi(\bar{y}_K^a)] - \phi_* \leq \frac{1}{K} \left( \frac{d_0^2}{2\lambda} + 2R\bar{M}^2 + \frac{2\lambda\bar{M}^2}{\theta} \right).$$

**Corollary 3.5.** *Assume that conditions (A1)-(A4) hold and let  $d_0$  be as in (29). Let pair  $(\lambda, K)$  and constant  $C \geq 1$  be given, and define*

$$\theta = \frac{C}{K}, \quad R = \frac{D^2}{4\lambda M^2},$$

*where  $(D, M)$  is an estimate for the pair  $(d_0, \bar{M})$ . Then, the following statements about SCPB with input  $(\lambda, \theta, K)$  based on cycle rule (B1') with  $R$  as above hold:*

a) *we have*

$$\mathbb{E}[\phi(\bar{y}_K^a)] - \phi_* \leq \frac{D^2}{2\lambda K} (\kappa_{d_0} + \kappa_M) + \frac{2\lambda M^2 \kappa_M}{C}$$

*where  $\kappa_{d_0} = d_0/D$  and  $\kappa_M$  is as in (19);*

1  
2  
3  
4  
5  
6  
7  
8 b) the number of iterations within the  $k$ -th cycle  $\mathcal{C}_k$  is bounded by

$$\left\lceil (C + 1) \ln_0^+ \left( \frac{4\lambda^2 M^2 K}{D^2} \right) \right\rceil + 1;$$

12 and hence, up to a logarithmic term, is  $\mathcal{O}(C)$ ;

13 c) its expected overall iteration complexity, up to a logarithmic term, is  $\mathcal{O}(CK)$ .

## 4 Results of the second variant of SCPB

20 This section provides the main results of the SCPB variant based on cycle rule (B2).

21 The following result is an analogue of Theorem 3.1 and describes the convergence rate bound  
22 for the SCPB variant based on cycle rule (B2) without imposing any condition on its input  $(\lambda, \theta, K)$   
23 and the constant  $R$  in (B2). The proof is postponed to Subsection 5.2.

24 **Theorem 4.1.** *Assume that conditions (A1)-(A4) hold and  $\text{dom } h$  has a finite diameter  $D_h > 0$ .  
25 Then, SCPB based on cycle rule (B2) satisfies the following statements:*

26 a) the expected number of iterations within the  $k$ -th cycle  $\mathcal{C}_k$  (see (13)) is bounded by

$$\left\lceil (\theta K + 1) \ln_0^+ \left( \frac{2\bar{M}^2 \lambda^2 k}{R} \right) \right\rceil + 1; \quad (30)$$

27 b) we have

$$\mathbb{E}[\phi(\hat{y}_K^a)] - \phi_* \leq \frac{1}{K} \left( \frac{3R + D_h^2}{\lambda} + \frac{2\lambda\bar{M}^2}{\theta} + \frac{2\lambda\bar{M}^2}{\theta^2 K} \right).$$

28 Following a similar argument as in the paragraph following Corollary 3.2, it can be shown that  
29 SCPB has optimal iteration complexity (up to a logarithmic term) for finding an  $\varepsilon$ -solution of (1)  
30 for a large range of prox stepsizes.

31 The following result is the analogue of Corollary 3.3 when SCPB is implemented using cycle  
32 rule (B2) instead of (B1). As in Corollary 3.3, it forces the quantity  $\theta K$  to be constant but, in  
33 contrast to the choice of  $R$  of Corollary 3.3, its choice for  $R$  does not depend on an estimate  $M$  for  
34  $\bar{M}$ .

35 **Corollary 4.2.** *Assume that conditions (A1)-(A4) hold and  $\text{dom } h$  has a finite diameter  $D_h > 0$ .  
36 Let positive integers  $K$  and constant  $C \geq 1$  be given, and define*

$$\theta = \frac{C}{K}, \quad R = D^2, \quad \lambda = \frac{\sqrt{C}D}{M\sqrt{K}} \quad (31)$$

37 where  $D$  is an estimate for  $D_h$  and  $M$  is an estimate for  $\bar{M}$  such that  $D \geq D_h$  and  $M \geq \bar{M}$ . Then,  
38 the following statements for SCPB with input  $(\lambda, \theta, K)$  based on cycle rule (B2) with  $R, \theta$ , and  $\lambda$   
39 as above hold:

40 a) we have

$$\mathbb{E}[\phi(\hat{y}_K^a)] - \phi_* \leq \frac{8DM}{\sqrt{CK}}; \quad (32)$$

1  
2  
3  
4  
5  
6  
7  
8 b) the expected number of iterations within the  $k$ -th cycle  $\mathcal{C}_k$  is bounded by

$$\left\lceil (C+1) \ln_0^+ \left( \frac{2Ck}{K} \right) \right\rceil + 1,$$

10  
11  
12 and hence, up to a logarithmic term, is  $\mathcal{O}(C)$ ;

13  
14  
15 c) its expected overall iteration complexity, up to a logarithmic term, is  $\mathcal{O}(CK)$ .

16  
17 **Proof:** a) Using Theorem 4.1(b) with  $\theta$  and  $R$  as in (31) and the assumptions that  $D \geq D_h$  and  
18  $M \geq \bar{M}$ , we have

$$\mathbb{E}[\phi(\hat{y}_K^a)] - \phi_* \leq \frac{4D^2}{\lambda K} + \frac{4\lambda M^2}{C},$$

20  
21 which together with  $\lambda$  in (31) implies (32).

22  
23 b) This statement follows from (30) with  $\theta$ ,  $R$ , and  $\lambda$  as in (31) and the assumption that  $M \geq \bar{M}$ .

24  
25 c) This statement follows from (b) and the fact that SCPB has  $K$  cycles.  $\blacksquare$

## 26 27 28 5 Proofs of main results in Sections 3 and 4

29  
30 This section is devoted to the proofs of Theorems 3.1 and 4.1.

### 31 32 5.1 Proof of Theorem 3.1

33  
34 We recall that for every  $j \geq 0$

$$\xi_{[j]} = (\xi_0, \xi_1, \dots, \xi_j)$$

35  
36 and for  $p \leq q$  positive integers we denote by  $\xi_{[p:q]}$  the portion  $\xi_{[p:q]} = (\xi_p, \xi_{p+1}, \dots, \xi_q)$  of realizations  
37 of the r.v.  $\xi$  over the iterations  $p, p+1, \dots, q$ . For convenience, in what follows we set

$$s_j := s(x_j, \xi_j). \quad (33)$$

38  
39 For every  $k \geq 1$  and  $j \in \mathcal{C}_k$ , define

$$u_j := \begin{cases} \Phi(x_{i_k}, \xi_{i_k}), & \text{if } j = i_k, \\ (1-\tau)\phi(x_j) + \tau u_{j-1}, & \text{otherwise,} \end{cases} \quad (34)$$

40  
41 and

$$\Gamma_j(\cdot) := \begin{cases} \tilde{\ell}_k(\cdot) + h(\cdot), & \text{if } j = i_k, \\ (1-\tau)\ell(\cdot; x_{j-1}, \xi_{j-1}) + \tau\Gamma_{j-1}(\cdot), & \text{otherwise,} \end{cases} \quad (35)$$

42  
43 where  $\ell(\cdot; x, \xi)$  and  $\tilde{\ell}_k(\cdot)$  are as in (5) and (15), respectively. It is easy to see from (9), (10), and  
44 the above definition of  $\Gamma_j$  that

$$x_j = \operatorname{argmin}_{u \in \mathbb{R}^n} \left\{ \Gamma_j^\lambda(u) := \Gamma_j(u) + \frac{1}{2\lambda} \|u - x_j^c\|^2 \right\}. \quad (36)$$

45  
46 The first result below provides some basic relations which are often used in our analysis.

**Lemma 5.1.** *For every  $j \geq 1$ , we have*

$$\mathbb{E}[\Phi(x_j, \xi_j)] = \mathbb{E}[\phi(x_j)], \quad (37)$$

$$\mathbb{E}[\phi(y_j)] \leq \mathbb{E}[u_j], \quad (38)$$

$$\mathbb{E}[\Gamma_j(x)] \leq \phi(x) \quad \forall x \in \text{dom } h. \quad (39)$$

**Proof:** Observe that  $x_j$  is a function of  $\xi_{[j-1]}$  and not of  $\xi_j$ . Hence,  $x_j$  is independent of  $\xi_j$  in view of the fact that  $\xi_j$  is chosen in Step 1 of SCPB to be independent of  $\xi_{[j-1]}$ . Using the relation  $f(x) = \mathbb{E}[F(x, \xi)]$  (see (A2)), it follows that

$$\begin{aligned} \mathbb{E}[\Phi(x_j, \xi_j)] &= \mathbb{E}_{\xi_{[j]}}[F(x_j, \xi_j) + h(x_j)] = \mathbb{E}_{\xi_{[j-1]}}[\mathbb{E}_{\xi_j}[F(x_j, \xi_j) + h(x_j)|\xi_{[j-1]}]] \\ &= \mathbb{E}_{\xi_{[j-1]}}[f(x_j) + h(x_j)] = \mathbb{E}[\phi(x_j)], \end{aligned}$$

which is identity (37). It then suffices to show that, for any given  $k \geq 1$ , (38) and (39) hold for every  $j$  in the  $k$ -th cycle, i.e.,  $j \in \mathcal{C}_k$ . We show this by induction on  $j$  where  $j$  is the iteration count. If  $j = i_k$ , then it follows from (11), (34), and (37) that

$$\mathbb{E}[u_j] \stackrel{(34)}{=} \mathbb{E}[\Phi(x_j, \xi_j)] \stackrel{(37)}{=} \mathbb{E}[\phi(x_j)] \stackrel{(11)}{=} \mathbb{E}[\phi(y_j)],$$

and from (35) with  $j = i_k$ , (15), and assumptions (A1)-(A2) that for every  $x \in \text{dom } h$ ,

$$\mathbb{E}[\Gamma_j(x)] \stackrel{(35)}{=} \mathbb{E}[\tilde{\ell}_k(x) + h(x)] \stackrel{(15),(A2)}{=} f(x_{i_{k-1}}) + \langle f'(x_{i_{k-1}}), x - x_{i_{k-1}} \rangle + h(x) \stackrel{(A1)}{\leq} \phi(x).$$

Let  $j$  be such that  $j > i_k$  and (38) and (39) hold for  $j$ . Then, it follows from (11), (34), the fact that (38) holds for  $j$ , and the convexity of  $\phi$ , that

$$\mathbb{E}[u_{j+1}] \stackrel{(34),(38)}{\geq} (1 - \tau)\mathbb{E}[\phi(x_{j+1})] + \tau\mathbb{E}[\phi(y_j)] \geq \mathbb{E}[\phi((1 - \tau)x_{j+1} + \tau y_j)] \stackrel{(11)}{=} \mathbb{E}[\phi(y_{j+1})],$$

and from (6), (35) and the fact that (39) holds for  $j$ , that

$$\mathbb{E}[\Gamma_{j+1}(x)] \stackrel{(35)}{=} \tau\mathbb{E}[\Gamma_j(x)] + (1 - \tau)\mathbb{E}[\ell(x; x_j, \xi_j)] \stackrel{(6),(39)}{\leq} \tau\phi(x) + (1 - \tau)\phi(x) = \phi(x).$$

We have thus shown that (38) and (39) hold for every  $j \in \mathcal{C}_k$ .  $\blacksquare$

It is worth noting that the proof of (39) is strongly based on the fact that  $\Gamma_j$  is a convex combination of affine functions whose expected values are underneath  $\phi$ . Moreover, this inequality would not necessarily be true if  $\Gamma_j$  were for example the maximum of functions as just described.

The next result provides a useful estimate for the quantity  $\phi(x_j, \xi_j) - \ell(x_j; x_{j-1}, \xi_{j-1})$ .

**Lemma 5.2.** *For every  $j \in \mathcal{C}_k$  such that  $j \geq j_{k-1} + 1$ , we have:*

$$\phi(x_j) - \ell(x_j; x_{j-1}, \xi_{j-1}) \leq (\bar{M} + \|s_{j-1}\|)\|x_j - x_{j-1}\|. \quad (40)$$

**Proof:** Using the definitions of  $\phi$  and  $\ell(\cdot; x, \xi)$  in (1) and (5), respectively, we have

$$\phi(x_j) - \ell(x_j; x_{j-1}, \xi_{j-1}) = f(x_j) - f(x_{j-1}) - \langle s_{j-1}, x_j - x_{j-1} \rangle \leq \langle f'(x_j) - s_{j-1}, x_j - x_{j-1} \rangle$$

where the inequality is due to the convexity of  $f$ . The above inequality, the Cauchy-Schwarz inequality, the triangle inequality and (4) then imply (40).  $\blacksquare$

The technical result below introduces a key quantity, namely, scalar  $t_j$  below, and provides a useful recursive relation for it over the iterations of the  $k$ -th cycle. This recursive relation will then be used in Proposition 5.6 to show that the  $t_j$  at the end of the  $k$ -th cycle, namely  $t_{j_k}$ , is relatively small in expectation.

**Lemma 5.3.** For every  $j \geq 1$ , define

$$t_j := u_j - \Gamma_j^\lambda(x_j), \quad b_{j+1} := \frac{\lambda(\bar{M}^2 + \|s_j\|^2)}{\theta K} \quad (41)$$

where  $\lambda$ ,  $\theta$ , and  $K$  are as in Step 0 of SCPB, and  $s_j$  is as in (33). Then, for every  $j \in \mathcal{C}_k$  such that  $j \geq i_k + 1$ , we have

$$t_j \leq \tau t_{j-1} + (1 - \tau)b_j \quad (42)$$

where  $\tau$  is as in (7), and hence

$$t_j \leq \tau^{j-i_k} t_{i_k} + (1 - \tau) \sum_{i=i_k+1}^j \tau^{j-i} b_i. \quad (43)$$

**Proof:** Let  $j \in \mathcal{C}_k$  with  $j \geq i_k + 1$ . It follows from the definitions of  $\Gamma_j$  and  $\Gamma_j^\lambda$  in (35) and (36), respectively, that

$$\begin{aligned} \Gamma_j^\lambda(x_j) &= (1 - \tau)\ell(x_j; x_{j-1}, \xi_{j-1}) + \tau\Gamma_{j-1}(x_j) + \frac{1}{2\lambda}\|x_j - x_j^c\|^2 \\ &\geq (1 - \tau)\ell(x_j; x_{j-1}, \xi_{j-1}) + \tau \left[ \Gamma_{j-1}(x_j) + \frac{1}{2\lambda}\|x_j - x_{j-1}^c\|^2 \right] \\ &= (1 - \tau)\ell(x_j; x_{j-1}, \xi_{j-1}) + \tau\Gamma_{j-1}^\lambda(x_j) \\ &\geq (1 - \tau)\ell(x_j; x_{j-1}, \xi_{j-1}) + \tau \left[ \Gamma_{j-1}^\lambda(x_{j-1}) + \frac{1}{2\lambda}\|x_j - x_{j-1}\|^2 \right], \end{aligned} \quad (44)$$

where for the first inequality we used the fact that  $\tau < 1$  and  $x_j^c = x_{j-1}^c$  for  $j \in \mathcal{C}_k$  with  $j \geq i_k + 1$  while for the second inequality is due to the facts that  $\Gamma_j^\lambda$  is  $(1/\lambda)$ -strongly convex and  $x_{j-1}$  is the minimizer of  $\Gamma_{j-1}^\lambda$  (see (36)). Using (7), (40) and (44), we have

$$\begin{aligned} \Gamma_j^\lambda(x_j) - \tau\Gamma_{j-1}^\lambda(x_{j-1}) &\stackrel{(7),(44)}{\geq} (1 - \tau) \left[ \ell(x_j; x_{j-1}, \xi_{j-1}) + \frac{\theta K}{2\lambda}\|x_j - x_{j-1}\|^2 \right] \\ &\stackrel{(40)}{\geq} (1 - \tau)\phi(x_j) + (1 - \tau) \left[ \frac{\theta K}{2\lambda}\|x_j - x_{j-1}\|^2 - (\bar{M} + \|s_{j-1}\|)\|x_j - x_{j-1}\| \right] \\ &\geq (1 - \tau)\phi(x_j) - (1 - \tau) \frac{\lambda(\bar{M} + \|s_{j-1}\|)^2}{2\theta K} \end{aligned}$$

where the last inequality is obtained by minimizing its left hand side with respect to  $\|x_j - x_{j-1}\|$ . The above inequality, the fact that  $(\alpha_1 + \alpha_2)^2 \leq 2\alpha_1^2 + 2\alpha_2^2$  for every  $\alpha_1, \alpha_2 \in \mathbb{R}$ , and the definition of  $b_j$  in (41) imply that

$$\Gamma_j^\lambda(x_j) - \tau\Gamma_{j-1}^\lambda(x_{j-1}) \geq (1 - \tau)\phi(x_j) - (1 - \tau) \frac{\lambda(\bar{M}^2 + \|s_{j-1}\|^2)}{\theta K} \stackrel{(41)}{=} (1 - \tau)\phi(x_j) - (1 - \tau)b_j.$$

Rearranging the above inequality and using the definition of  $t_j$  in (41), identity (34), and the fact that  $j \geq i_k + 1$ , we then conclude that

$$\begin{aligned} \Gamma_j^\lambda(x_j) + (1 - \tau)b_j &\geq \tau\Gamma_{j-1}^\lambda(x_{j-1}) + (1 - \tau)\phi(x_j) \\ &\stackrel{(41)}{=} \tau(u_{j-1} - t_{j-1}) + (1 - \tau)\phi(x_j) \stackrel{(34)}{=} u_j - \tau t_{j-1}, \end{aligned}$$

which, in view of the definition of  $t_j$  in (41), implies (42). Inequality (43) follows immediately from (42) and an induction argument.  $\blacksquare$

The following technical result provides some useful bounds on  $b_j$ .

**Lemma 5.4.** For every  $\ell \geq 0$  and  $j \geq \ell + 2$ , we have

$$\mathbb{E}[b_j | \xi_{[\ell]}] \leq \frac{2\lambda\bar{M}^2}{\theta K}, \quad \mathbb{E}[b_j] \leq \frac{2\lambda\bar{M}^2}{\theta K}. \quad (45)$$

**Proof:** We first show that for every  $j \geq 1$  and  $\ell \leq j - 1$ ,

$$\mathbb{E}[\|s_j\|^2 | \xi_{[\ell]}] \leq \bar{M}^2. \quad (46)$$

Fix  $j \geq 1$ . Since  $x_j$  becomes deterministic when  $\xi_{[j-1]}$  is given, it follows from (A3) with  $x = x_j$  and the definition of  $s_j$  in (33) that

$$\mathbb{E}_{\xi_j}[\|s_j\|^2 | \xi_{[j-1]}] \leq \bar{M}^2.$$

Now, if  $\ell \leq j - 2$ , then the above relations together with the law of total expectation imply that and

$$\mathbb{E}[\|s_j\|^2 | \xi_{[\ell]}] = \mathbb{E}_{\xi_{[\ell+1:j]}}[\|s_j\|^2 | \xi_{[\ell]}] = \mathbb{E}_{\xi_{[\ell+1:j-1]}}[\mathbb{E}_{\xi_j}[\|s_j\|^2 | \xi_{[j-1]}]] \leq \bar{M}^2.$$

We have thus shown that (46) holds for any  $\ell \leq j - 1$ .

The first inequality in (45) then follows from the definition of  $b_j$  in (41). The second inequality in (45) follows from the first one and the law of total expectation.  $\blacksquare$

The next technical result provides a bound on the initial  $t_j$  for the  $k$ -th cycle, namely  $t_{i_k}$ , in expectation.

**Lemma 5.5.** For every  $k \geq 1$ , we have  $\mathbb{E}[t_{i_k}] \leq 2 \min\{\lambda\bar{M}^2, \bar{M}D_h\}$  where  $i_k$  and  $t_j$  are as in (13) and (41), respectively.

**Proof:** Let

$$\Delta_j = \Phi(x_j, \xi_j) - \phi(x_j) = F(x_j, \xi_j) - f(x_j). \quad (47)$$

Using the definitions of  $t_j$  and  $\Gamma_j^\lambda$  in (41) and (36), respectively, (34) with  $j = i_k = j_{k-1} + 1$  (see (13)), we have

$$\begin{aligned} t_{i_k} &\stackrel{(41)}{=} u_{i_k} - \Gamma_{i_k}^\lambda(x_{i_k}) \\ &\stackrel{(34),(35)}{=} \Phi(x_{i_k}, \xi_{i_k}) - [F(x_{j_{k-1}}, \xi_{j_{k-1}}) + \langle s_{j_{k-1}}, x_{i_k} - x_{j_{k-1}} \rangle + h(x_{i_k})] - \frac{1}{2\lambda} \|x_{i_k} - x_{j_{k-1}}\|^2 \\ &= \Delta_{i_k} - \Delta_{j_{k-1}} + \phi(x_{i_k}) - \ell(x_{i_k}; x_{j_{k-1}}, \xi_{j_{k-1}}) - \frac{1}{2\lambda} \|x_{i_k} - x_{j_{k-1}}\|^2 \\ &\leq \Delta_{i_k} - \Delta_{j_{k-1}} + (\bar{M} + \|s_{j_{k-1}}\|) \|x_{i_k} - x_{j_{k-1}}\| - \frac{1}{2\lambda} \|x_{i_k} - x_{j_{k-1}}\|^2 \end{aligned} \quad (48)$$

where the inequality is due to Lemma 5.2. Maximizing the right hand side of the last inequality above with respect to  $\|x_{i_k} - x_{j_{k-1}}\|$  and using the relation  $(a + b)^2 \leq 2a^2 + 2b^2$  for every  $a, b \in \mathbb{R}$ , we obtain

$$t_{i_k} \leq \Delta_{i_k} - \Delta_{j_{k-1}} + \frac{\lambda}{2} (\bar{M} + \|s_{j_{k-1}}\|)^2 \leq \Delta_{i_k} - \Delta_{j_{k-1}} + \lambda (\bar{M}^2 + \|s_{j_{k-1}}\|^2). \quad (49)$$

Moreover, (48) and the fact that  $\|x_{i_k} - x_{j_{k-1}}\| \leq D_h$  also imply that

$$t_{i_k} \leq \Delta_{i_k} - \Delta_{j_{k-1}} + (\bar{M} + \|s_{j_{k-1}}\|) D_h. \quad (50)$$



1  
2  
3  
4  
5  
6  
7  
8 It follows from (33), (47), and conditions (A2) and (A3) that

$$9 \quad \mathbb{E}[\Delta_{i_k}] = 0, \quad \mathbb{E}[\Delta_{j_{k-1}}] = 0, \quad \mathbb{E}[\|s_{j_{k-1}}\|^2] \leq \bar{M}^2.$$

10  
11 Hence, the lemma follows by taking expectations of (49) and (50) and using the above three  
12 relations.  $\blacksquare$

13 All the results developed above hold regardless of the way  $j_k$  is chosen in Step 2. On the other  
14 hand, the results below all depend on cycle rule (B1).

15 We are now ready to derive a bound on the last  $t_j$  for the  $k$ -th cycle in expectation.

16  
17 **Proposition 5.6.** *In addition to conditions (A1)-(A4), assume also that (B1) holds. Then, for  
18 every  $k \geq 1$ , we have*

$$19 \quad \mathbb{E}[t_{j_k}] \leq \frac{2R \min\{\lambda \bar{M}^2, \bar{M} D_h\}}{\lambda k} + \frac{2\lambda \bar{M}^2}{\theta K} \quad (51)$$

20 where  $t_j$  is as in (41).

21  
22 **Proof:** Fix  $k \geq 1$ . It follows from cycle rule (B1) and inequality (43) with  $j = j_k$  that

$$23 \quad t_{j_k} \leq \tau^{j_k - i_k} t_{i_k} + (1 - \tau) \sum_{i=i_k+1}^{j_k} \tau^{j_k - i} b_i.$$

24  
25 In view of (B1) and (13), it follows that  $j_k$  and  $i_k$  are both deterministic. Hence, taking expectation  
26 of the above inequality and using the last inequality in (45), cycle rule (B1) and Lemma 5.5, we  
27 conclude that

$$28 \quad \begin{aligned} 29 \quad \mathbb{E}[t_{j_k}] &\leq \tau^{j_k - i_k} \mathbb{E}[t_{i_k}] + (1 - \tau) \sum_{i=i_k+1}^{j_k} \tau^{j_k - i} \mathbb{E}[b_i] \\ 30 &\leq \frac{2R \min\{\lambda \bar{M}^2, \bar{M} D_h\}}{\lambda k} + (1 - \tau) \frac{2\lambda \bar{M}^2}{\theta K} \sum_{i=i_k+1}^{j_k} \tau^{j_k - i}, \end{aligned}$$

31 and hence that (51) holds.  $\blacksquare$

32 In the remaining part of this subsection, we analyze the behavior of the “outer” sequence of  
33 iterations  $\{\hat{y}_k\} = \{y_{j_k}\} \subset \mathbb{R}^n$  generated in Step 2 of SCPB. For this purpose, define

$$34 \quad \hat{\Gamma}_k := \Gamma_{j_k} \quad \forall k \geq 1 \quad (52)$$

35 and

$$36 \quad \hat{x}_k := x_{j_k}, \quad \hat{u}_k := u_{j_k}. \quad (53)$$

37 In what follows, we make some remarks about the above “outer” sequences which follow as  
38 immediate consequences of the results developed above. In view of the above definitions, relation  
39 (36) with  $j = j_k$ , and the way the prox centers  $x_j^c$  are updated in (8), we have that

$$40 \quad \hat{x}_k = \operatorname{argmin}_{x \in \mathbb{R}^n} \left\{ \hat{\Gamma}_k(x) + \frac{1}{2\lambda} \|x - \hat{x}_{k-1}\|^2 \right\} \quad \forall k \geq 1. \quad (54)$$

41 Moreover, it follows from (38) and (39) with  $j = j_k$  that

$$42 \quad \mathbb{E}[\phi(\hat{y}_k)] \leq \mathbb{E}[\hat{u}_k] \quad (55)$$

and

$$\mathbb{E}[\hat{\Gamma}_k(z)] \leq \phi(z) \quad \forall z \in \text{dom } h. \quad (56)$$

The following result describes an important recursive formula for the outer sequence  $\{\hat{y}_k\}$  generated by SCPB.

**Lemma 5.7.** *In addition to conditions (A1)-(A4), assume also that (B1) holds. Then, for every  $k \geq 1$  and  $z \in \text{dom } h$ , we have*

$$\frac{2R \min\{\lambda \bar{M}^2, \bar{M} D_h\}}{\lambda k} + \frac{2\lambda \bar{M}^2}{\theta K} + \frac{1}{2\lambda} [d_{k-1}(z)]^2 - \frac{1}{2\lambda} [d_k(z)]^2 \geq \mathbb{E}[\phi(\hat{y}_k)] - \phi(z)$$

where

$$d_k(z) := (\mathbb{E}[\|\hat{x}_k - z\|^2])^{1/2}. \quad (57)$$

**Proof:** First observe that (52), (53), and the definitions of  $\Gamma_j^\lambda$  and  $t_j$  in (36) and (41), respectively, imply that (51) is equivalent to

$$\mathbb{E} \left[ \hat{u}_k - \hat{\Gamma}_k(\hat{x}_k) - \frac{1}{2\lambda} \|\hat{x}_k - \hat{x}_{k-1}\|^2 \right] \leq \frac{2R \min\{\lambda \bar{M}^2, \bar{M} D_h\}}{\lambda k} + \frac{2\lambda \bar{M}^2}{\theta K}. \quad (58)$$

It follows from (54) and the fact that the objective function of (54) is  $(1/\lambda)$ -strongly convex that for every  $z \in \text{dom } h$ ,

$$\hat{\Gamma}_k(\hat{x}_k) + \frac{1}{2\lambda} \|\hat{x}_k - \hat{x}_{k-1}\|^2 \leq \hat{\Gamma}_k(z) + \frac{1}{2\lambda} \|z - \hat{x}_{k-1}\|^2 - \frac{1}{2\lambda} \|z - \hat{x}_k\|^2,$$

and hence that

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

Taking expectation of the above inequality and using (57) and (58), we conclude that

$$\frac{2R \min\{\lambda \bar{M}^2, \bar{M} D_h\}}{\lambda k} + \frac{2\lambda \bar{M}^2}{\theta K} + \frac{1}{2\lambda} [d_{k-1}(z)]^2 \geq \mathbb{E}[\hat{u}_k] - \mathbb{E}[\hat{\Gamma}_k(z)] + \frac{1}{2\lambda} [d_k(z)]^2$$

which, in view of (55) and (56), immediately implies the conclusion of the lemma.  $\blacksquare$

We are now in a position to prove Theorem 3.1.

**Proof of Theorem 3.1:** a) This statement directly follows from (7), cycle rule (B1), the definition of  $\ln_0^+$ , and the facts that  $|\mathcal{C}_k| = j_k - i_k + 1$  and  $\ln \tau^{-1} \geq 1 - \tau$ .

b) Using the definition of  $\hat{y}_K^a$  in (12), Lemma 5.7 with  $z = x^* \in X^*$ , and the facts that  $\lceil K/2 \rceil \geq K/2$  and

$$\sum_{k=\lceil K/2 \rceil+1}^K \frac{1}{k} \leq \int_{\lceil K/2 \rceil}^K \frac{1}{x} dx = \ln \frac{K}{\lceil K/2 \rceil} \leq \ln \frac{K}{K/4} = \ln 4 \leq \frac{3}{2} \quad \forall K \geq 2,$$

we then conclude that for every  $K \geq 2$ ,

$$\begin{aligned} \mathbb{E}[\phi(\hat{y}_K^a)] - \phi_* &\leq \frac{1}{\lceil K/2 \rceil} \sum_{k=\lceil K/2 \rceil+1}^K (\mathbb{E}[\phi(\hat{y}_k)] - \phi_*) \\ &\leq \frac{1}{\lceil K/2 \rceil} \sum_{k=\lceil K/2 \rceil+1}^K \left( \frac{2R \min\{\lambda \bar{M}^2, \bar{M} D_h\}}{\lambda k} + \frac{2\lambda \bar{M}^2}{\theta K} + \frac{1}{2\lambda} [d_{k-1}(x^*)]^2 - \frac{1}{2\lambda} [d_k(x^*)]^2 \right) \\ &\leq \frac{6R \min\{\lambda \bar{M}^2, \bar{M} D_h\}}{\lambda K} + \frac{2\lambda \bar{M}^2}{\theta K} + \frac{[d_{\lceil K/2 \rceil}(x^*)]^2}{\lambda K}. \end{aligned} \quad (59)$$

It is also easy to see from Lemma 5.7 that (59) holds for  $K = 1$ . Then (17) follows from (59) and the fact that  $d_{\lfloor K/2 \rfloor}(x^*) \leq D_h$ .  $\blacksquare$

## 5.2 Proof of Theorem 4.1

The following proposition derives a bound on the last  $t_j$  for the  $k$ -th cycle in expectation. It is an analogue of Proposition 5.6 under the conditions (A1)-(A4) and cycle rule (B2).

**Proposition 5.8.** *In addition to conditions (A1)-(A4), assume also that cycle rule (B2) is used. For every  $k \geq 1$ , we have*

$$\mathbb{E}[t_{j_k}] \leq \frac{R}{\lambda k} + \frac{2\lambda\bar{M}^2}{\theta K} + \frac{2\lambda\bar{M}^2}{\theta^2 K^2}. \quad (60)$$

**Proof:** Using (43) with  $j = j_k$ , (14), recalling that  $j_k \geq i_k + 1$  for cycle rule (B2), and that  $\tau \in (0, 1)$  in view of (7), we conclude that

$$t_{j_k} - (1 - \tau) \sum_{i=i_k+2}^{j_k} \tau^{j_k-i} b_i \stackrel{(43)}{\leq} \tau^{j_k-i_k} t_{i_k} + (1 - \tau) \tau^{j_k-i_k-1} b_{i_k+1} \leq \frac{R}{\lambda k} + (1 - \tau) b_{i_k+1}.$$

Noting that  $j_k$  becomes deterministic once  $\xi_{[i_k]}$  is given, taking expectation of the above inequality conditioned on  $\xi_{[i_k]}$ , rearranging the terms, and using the first inequality in (45), we have

$$\begin{aligned} \mathbb{E}[t_{j_k} | \xi_{[i_k]}] - \frac{R}{\lambda k} - (1 - \tau) \mathbb{E}[b_{i_k+1} | \xi_{[i_k]}] &\leq (1 - \tau) \sum_{i=i_k+2}^{j_k} \tau^{j_k-i} \mathbb{E}[b_i | \xi_{[i_k]}] \\ &\stackrel{(45)}{\leq} (1 - \tau) \left( \sum_{i=i_k+2}^{j_k} \tau^{j_k-i} \right) \frac{2\lambda\bar{M}^2}{\theta K} \leq \frac{2\lambda\bar{M}^2}{\theta K}. \end{aligned}$$

Taking expectation of the above inequality with respect to  $\xi_{[i_k]}$ , rearranging the terms, and using the second inequality (45) and the fact that  $1 - \tau \leq 1/(\theta K)$  by (7), we conclude that

$$\begin{aligned} \mathbb{E}[t_{j_k}] &\leq \frac{R}{\lambda k} + (1 - \tau) \mathbb{E}[b_{i_k+1}] + \frac{2\lambda\bar{M}^2}{\theta K} \\ &\leq \frac{R}{\lambda k} + \frac{1}{\theta K} \frac{2\lambda\bar{M}^2}{\theta K} + \frac{2\lambda\bar{M}^2}{\theta K}, \end{aligned}$$

and hence that (60) holds.  $\blacksquare$

The following lemma is an analogue of Lemma 5.7 under the conditions (A1)-(A4) and cycle rule (B2).

**Lemma 5.9.** *In addition to conditions (A1)-(A4), assume also that cycle rule (B2) is used. Then, for every  $z \in \text{dom } h$  and  $k \geq 1$ , we have*

$$\frac{R}{\lambda k} + \frac{2\lambda\bar{M}^2}{\theta K} + \frac{2\lambda\bar{M}^2}{\theta^2 K^2} + \frac{1}{2\lambda} [d_{k-1}(z)]^2 - \frac{1}{2\lambda} [d_k(z)]^2 \geq \mathbb{E}[\phi(\hat{y}_k)] - \phi(z)$$

where  $d_k(z)$  is as in (57).

**Proof:** First observe that the definitions of  $\Gamma_j^\lambda$  and  $t_j$  in (36) and (41), respectively, imply that (60) is equivalent to

$$\mathbb{E} \left[ \hat{u}_k - \hat{\Gamma}_k(\hat{x}_k) - \frac{1}{2\lambda} \|\hat{x}_k - \hat{x}_{k-1}\|^2 \right] \leq \frac{R}{\lambda k} + \frac{2\lambda\bar{M}^2}{\theta K} + \frac{2\lambda\bar{M}^2}{\theta^2 K^2}. \quad (61)$$

The remaining part of the proof is now similar to that of Lemma 5.7 except that (61) is used in place of (58). ■

We are now ready to prove Theorem 4.1.

**Proof of Theorem 4.1:** a) Using (7), (14), the definition of  $\ln_0^+$ , and the facts that  $|\mathcal{C}_k| = j_k - i_k + 1$  and  $\ln \tau^{-1} \geq 1 - \tau$ , we have

$$|\mathcal{C}_k| \leq \frac{1}{1 - \tau} \ln_0^+ \left( \frac{t_{i_k} \lambda k}{R} \right) + 1 = (\theta K + 1) \ln_0^+ \left( \frac{t_{i_k} \lambda k}{R} \right) + 1.$$

Taking expectation of the above inequality, and using the Jensen's inequality and the fact that  $\ln x$  is a concave function, we then conclude that

$$\begin{aligned} \mathbb{E}[|\mathcal{C}_k|] &\leq (\theta K + 1) \mathbb{E} \left[ \ln_0^+ \left( \frac{t_{i_k} \lambda k}{R} \right) \right] + 1 \leq (\theta K + 1) \ln_0^+ \left( \frac{\mathbb{E}[t_{i_k}] \lambda k}{R} \right) + 1 \\ &\leq (\theta K + 1) \ln_0^+ \left( \frac{2\bar{M}^2 \lambda^2 k}{R} \right) + 1, \end{aligned}$$

where the last inequality is due to Lemma 5.5.

b) This statement follows from the same argument as in the proof of Theorem 3.1(b) except that Lemma 5.9 is used in place of Lemma 5.7. ■

## 6 Numerical experiments

In this section, we report the results of numerical experiments where we compare the performance of RSA and our two variants of SCPB on three stochastic programming problems, namely: a stochastic utility problem given in Section 4.2 of [14] and the two two-stage nonlinear stochastic programs considered in the numerical experiments of [6]. These three problems are of form (1)-(2) with  $h$  the indicator function of a convex compact set  $X$  with diameter  $D_X$ . Therefore, the problems can be written as

$$\min\{f(x) := \mathbb{E}[F(x, \xi)] : x \in X\}. \quad (62)$$

The implementations were coded in MATLAB, using Mosek optimization library [1] to solve sub-problems, and run on a laptop with Intel i7, 1.80 GHz processor. However, when needed (for all methods), we implemented the (efficient) algorithm from [25] to project onto the simplex.

**Parameters for Robust Stochastic Approximation.** Robust Stochastic Approximation, denoted by E-SA (Euclidean Stochastic Approximation) in what follows, is described in Section 2.2 of [14] (as explained in [14], in terms of Section 2.3 of [14], this is mirror descent robust SA with Euclidean setup). In the notation of [14], for E-SA run for  $N$  iterations, we output  $\tilde{x}_1^N = \frac{1}{N} \sum_{i=1}^N x_i$

(this is  $\tilde{x}_i^N$  given by (22) with  $i = 1$  and corresponds to the usual output of RSA) where  $x_i$  is computed at iteration  $i$  taking the constant steps given in (2.23) of [14] by

$$\gamma_t = \frac{\theta D_X}{M\sqrt{N}}$$

where  $D_X$  is the diameter of the feasible set  $X$  in (62).<sup>1</sup> As in [14], we take  $\theta = 0.1$  which was calibrated in [14] using an instance of the stochastic utility problem. For each problem, the value of  $M$  is estimated as in [14] taking the maximum of  $\|s(\cdot, \cdot)\|$  over 10,000 calls to the stochastic oracle at randomly generated feasible solutions.

*Remark:* In [14], E-SA generates approximately  $\log_2(N)$  candidate solutions  $\tilde{x}_i^N = \frac{1}{N-i+1} \sum_{k=i}^N x_k$  with  $N - i + 1 = \min[2^k, N]$ ,  $k = 0, 1, \dots, \log_2(N)$  and an additional sample was used to estimate the objective at these candidate solutions in order to choose the best of these candidates. In [14], the computational effort required by this postprocessing is not reflected in the experiments. However, we believe that for a fair comparison of E-SA using this set of candidate solutions and SCPB, this computational effort should be taken into account and without this additional computational bulk, SCPB is already faster than E-SA in our experiments. ■

**Parameters for the first variant of SCPB.** Our first variant of SCPB, denoted by SCPB 1 in what follows, is based on cycle rule (B1) and uses parameters  $\theta$ ,  $\tau$ ,  $R$ , and  $\lambda$  given by

$$\theta = \frac{C}{K}, \tau = \frac{\theta K}{\theta K + 1}, R = \frac{D_X}{M}, \lambda = \beta_1 \frac{\sqrt{C} D_X}{M\sqrt{K}}$$

where constant  $C = 9$  and constant  $\beta_1$  was calibrated with the stochastic utility problem, see below. We take  $\beta_1 = 10$  in all our experiments. Constant  $M$  was estimated as for RSA taking the maximum of  $\|s(\cdot, \cdot)\|$  over 10,000 calls to the stochastic oracle at randomly generated feasible solutions.

**Parameters for the second variant of SCPB.** Our second variant of SCPB, denoted by SCPB 2 in what follows, uses cycle rule (B2) with parameters  $\theta$ ,  $\tau$ ,  $R$ , and  $\lambda$  given by

$$\theta = \frac{C}{K}, \tau = \frac{\theta K}{\theta K + 1}, R = D_X^2, \lambda = \beta_2 \frac{\sqrt{C} D_X}{M\sqrt{K}}$$

where constant  $C = 9$  and constant  $\beta_2$  was calibrated with the stochastic utility problem, see below. We take  $\beta_2 = 10$  in all our experiments. Constant  $M$  was again estimated as for RSA taking the maximum of  $\|s(\cdot, \cdot)\|$  over 10,000 calls to the stochastic oracle at randomly generated feasible solutions.

**Notation in the tables.** In what follows, we denote by

- $n$  the design dimension of an instance;
- $N$  the sample size used to run the methods; this is also the number of iterations of E-SA;
- $K$  the number of SCPB outer iterations;

---

<sup>1</sup>Parameter  $M$  is denoted by  $M_*$  in [14].

- Obj: the empirical mean

$$\hat{F}_T(x) := \frac{1}{T} \sum_{i=1}^T F(x, \xi_i) \quad (63)$$

of  $F$  at  $x$  based on a sample  $\xi_1, \dots, \xi_T$  of  $\xi$  of size  $T$ , which provides an estimation of  $f(x)$ . The empirical means are computed with  $x$  being the final iterate output by the algorithm and  $T = 10^4$ ;

- CPU is the CPU time in seconds.

## 6.1 A stochastic utility problem

Our first set of experiments was carried out with the stochastic utility problem given by

$$\min_{x \in X} \mathbb{E} \left[ \phi \left( \sum_{i=1}^n \left( \frac{i}{n} + \xi_i \right) x_i \right) \right]$$

where

$$X = \left\{ x \in \mathbb{R}^n : \sum_{i=1}^n x(i) = 1, x \geq 0 \right\}, \quad (64)$$

$\xi_i \sim \mathcal{N}(0, 1)$  are independent and  $\phi(t) = \max(v_1 + s_1 t, \dots, v_m + s_m t)$  is piecewise convex with 10 breakpoints, all located on  $[0, 1]^2$ . We consider four instances:  $L_1, L_2, L_3$ , and  $L_4$  which have respectively problem dimension  $n = 500, 1000, 2000$ , and  $5000$ .

**Calibration of  $\beta_1$  and  $\beta_2$ .** We run SCPB 1 and SCPB 2 with 7 values of  $\beta_1$  and  $\beta_2$  on four instances of the stochastic utility problem for  $K = 1000$  outer iterations (i.e., cycles). For this experiment, the values of  $\beta_1, \beta_2$ , the corresponding values of stepsize  $\lambda$ , and the optimal values computed by SCPB 1 and SCPB 2 are reported in Table 1. We found out that  $\beta_1 = 10$  slightly outperforms other choices of  $\beta_1$  for SCPB 1. Surprisingly, SCPB 2 was not affected by changes in  $\beta_2$  and all tested values allowed us to obtain with similar CPU times a good approximate optimal value. This value  $\beta_1 = 10$  and the same value  $\beta_2 = 10$  will be chosen for all runs of SCPB and all the problem instances (the stepsizes in [14] were calibrated similarly, on the basis of an instance of the stochastic utility problem).

---

<sup>2</sup>Although the same problem class and a similar procedure to build  $\phi$  was used in the experiments of Section 4.2 in [14], we could not find in this reference the precise choices of  $v_k, s_k$  and the optimal values of our instances differ from the optimal values of the instances in [14]. Also, contrary to [14], we use the same function  $\phi$  for all instances. The instances differ for the problem dimension  $n$ .

$\beta_1, \beta_2$	0.01	0.1	1	10	50	150	1000
$\lambda, n = 500$	$1.70 \times 10^{-5}$	$1.70 \times 10^{-4}$	0.0017	0.017	0.09	0.26	1.7
<b>Obj</b> <sub>1</sub> , $n = 500$	14.2795	10.6439	10.1819	10.1811	10.1811	10.1838	10.1937
<b>Obj</b> <sub>2</sub> , $n = 500$	10.1937	10.1937	10.1937	10.1937	10.1937	10.1937	10.1937
$\lambda, n = 10^3$	$1.17 \times 10^{-5}$	$1.17 \times 10^{-4}$	0.0012	0.012	0.0585	0.18	1.17
<b>Obj</b> <sub>1</sub> , $n = 10^3$	14.6307	11.1325	10.0510	10.0504	10.0509	10.0523	10.0710
<b>Obj</b> <sub>2</sub> , $n = 10^3$	10.0710	10.0710	10.0710	10.0710	10.0710	10.0710	10.0710
$\lambda, n = 2 \times 10^3$	$8.36 \times 10^{-6}$	$8.36 \times 10^{-5}$	$8.36 \times 10^{-4}$	0.0084	0.0418	0.1255	0.8364
<b>Obj</b> <sub>1</sub> , $n = 2 \times 10^3$	13.7451	11.0836	10.0365	10.0363	10.0364	10.0375	10.0613
<b>Obj</b> <sub>2</sub> , $n = 2 \times 10^3$	10.0613	10.0613	10.0613	10.0613	10.0613	10.0613	10.0613
$\lambda, n = 5 \times 10^3$	$7.93 \times 10^{-6}$	$7.93 \times 10^{-5}$	$7.93 \times 10^{-4}$	0.0079	0.0397	0.119	0.793
<b>Obj</b> <sub>1</sub> , $n = 5 \times 10^3$	14.0830	11.3370	10.0228	10.0228	10.0231	10.0237	10.0540
<b>Obj</b> <sub>2</sub> , $n = 5 \times 10^3$	10.0540	10.0540	10.0540	10.0540	10.0540	10.0540	10.0540

Table 1: Selecting parameters  $\beta_1$  and  $\beta_2$  of SCPB 1 and SCPB 2. Framework: SCPB,  $K = 1000$  outer iterations, four instances of the stochastic utility problem with  $n = 500, 1000, 2000,$  and  $5000$ . **Obj**<sub>1</sub> (resp. **Obj**<sub>2</sub>) is the approximate optimal value with SCPB 1 (resp. SCPB 2).

We run E-SA, SCPB 1, and SCPB 2 on four instances  $L_1, L_2, L_3,$  and  $L_4$  of the stochastic utility problem with  $n = 500, 1000, 2000,$  and  $5000$ , respectively. For SCPB 1 and SCPB 2, we used  $K = 1000$  outer iterations. The results are reported in Table 2. Several comments are now in order for the results reported in this table.

- For SCPB, approximate solutions can only be computed at the end of every cycle. Namely, at the end of  $L$ -th cycle at iteration  $j_L$  we can compute the approximate solution

$$\frac{1}{\lceil L/2 \rceil} \sum_{\ell=\lfloor L/2 \rfloor+1}^L \hat{y}_\ell = \frac{1}{\lceil L/2 \rceil} \sum_{\ell=\lfloor L/2 \rfloor+1}^L y_{j_\ell}.$$

For a given value of  $N$  in Table 2, the approximate objective value **Obj** we report for E-SA is the empirical mean of  $F(x, \xi)$  at the approximate solution  $\frac{1}{N} \sum_{i=1}^N x_i$  (where  $x_i$ 's are computed along iterations of E-SA) while for SCPB the approximate value **Obj** is the empirical mean of  $F(x, \xi)$  at the approximate solution  $\frac{1}{\lceil L(N)/2 \rceil} \sum_{\ell=\lfloor L(N)/2 \rfloor+1}^{L(N)} \hat{y}_\ell$  where

$$L(N) = \min\{k : j_k \geq N\}$$

(since a cycle may not end at iteration  $N$ ).

- Each iteration of E-SA and SCPB takes a similar amount of time (in both cases we evaluate an inexact prox-operator at some point) and therefore for a given sample size  $N$  the CPU time for E-SA and SCPB is similar.
- For all instances, SCPB computes a good approximate optimal value much quicker than E-SA and the decrease in the objective function value is much slower with E-SA. We also refer to Table 5 which reports the distance between SCPB approximate optimal value and E-SA approximate value as a percentage of SCPB decrease in the objective for several sample sizes. This percentage is above 90% for almost all instances and sample sizes.

-		$L_1 : n = 500$		$L_2 : n = 1000$		$L_3 : n = 2000$		$L_4 : n = 5000$	
ALG.	N	Obj	CPU	Obj	CPU	Obj	CPU	Obj	CPU
E-SA	10	14.6155	0.0012	14.6113	0.109	14.6449	0.001	14.6892	0.05
	50	14.6074	0.0035	14.6039	0.11	14.6322	0.006	14.6813	0.07
	100	14.5982	0.0065	14.5950	0.12	14.6169	0.01	14.6725	0.1
	200	14.5814	0.0125	14.5786	0.13	14.5880	0.03	14.6574	0.2
	1000	14.4682	0.0597	14.4651	0.2	14.3992	0.1	14.5604	0.5
	$10^4$	12.43	0.57	12.34	1.23	12.9656	1.28	12.7410	3.7
SCPB 1	10	12.3639	0.0019	12.8431	0.002	13.9539	0.003	13.7763	0.008
	50	11.4342	0.0048	12.1323	0.005	13.6527	0.01	13.4672	0.02
	100	10.9414	0.0082	11.4249	0.01	13.5986	0.02	13.5346	0.05
	200	10.4741	0.0145	10.7075	0.02	13.5349	0.03	13.4686	0.08
	1000	10.0824	0.0653	10.0574	0.09	13.0370	0.2	12.8376	0.5
SCPB 2	10	12.5790	0.0015	11.2559	0.003	13.5968	0.002	13.7777	0.01
	50	10.1287	0.0045	10.1430	0.01	12.9421	0.008	12.6959	0.05
	100	10.1180	0.0075	10.0799	0.02	12.1317	0.02	11.7614	0.09
	200	10.0993	0.0141	10.0656	0.04	11.3640	0.03	11.3698	0.2
	1000	10.0779	0.0718	10.0569	0.14	11.5681	0.1	11.5572	0.9

Table 2: E-SA versus two variants of SCPB on the stochastic utility problem run with  $K = 1000$  outer iterations.

## 6.2 A first two-stage stochastic program

Our second test problem is the nonlinear two-stage stochastic program

$$\begin{cases} \min c^T x_1 + \mathbb{E}[\Omega(x_1, \xi)] \\ x_1 \in \mathbb{R}^n : x_1 \geq 0, \sum_{i=1}^n x_1(i) = 1 \end{cases} \quad (65)$$

where the second stage recourse function is given by

$$\Omega(x_1, \xi) = \begin{cases} \min_{x_2 \in \mathbb{R}^n} \frac{1}{2} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^T (\xi \xi^T + \gamma_0 I_{2n}) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \xi^T \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ s.t. \quad x_2 \geq 0, \sum_{i=1}^n x_2(i) = 1. \end{cases} \quad (66)$$

Problem (65)-(66) is of form (1)-(2) where  $F(x, \xi) = c^T x + \Omega(x, \xi)$  with  $\Omega$  given by (66) and where  $h$  is the indicator function of set  $X$  where  $X$  given by (64) is the probability simplex. For problem (65) we refer to Lemma 2.1 in [4] for the computation of stochastic subgradients  $s(x, \xi)$ . We take  $\gamma_0 = 2$  and consider a Gaussian random vector in  $\mathbb{R}^{2n}$  for  $\xi$ . We consider two instances of problem (65) with  $n = 50$  and  $n = 100$ . For each instance, the components of  $\xi$  are independent with means and standard deviations randomly generated in respectively intervals  $[5, 25]$  and  $[5, 15]$ . The components of  $c$  are generated randomly in interval  $[1, 3]$ .

We run E-SA, SCPB 1, and SCPB 2 on our two instances  $A_1$  and  $A_2$  with  $n = 50$  and  $n = 100$ . For SCPB 1 and SCPB 2, we used  $K = 1000$  outer iterations. The results are reported in Table 3. The conclusions are similar to the experiments on the stochastic utility problem: SCPB computes a good approximate optimal value much quicker than E-SA and the decrease in the objective function value is much slower with E-SA. We again refer to Table 5 which reports the distance between SCPB



approximate optimal value and E-SA approximate value as a percentage of SCPB decrease in the objective for several sample sizes. This percentage is again above 90% for almost all instances and sample sizes.

-		$A_1 : n = 50$		$A_2 : n = 100$	
ALG.	N	Obj	CPU	Obj	CPU
E-SA	10	24.3477	0.13	7.5134	0.5
	50	24.2378	0.6	7.5018	2.5
	100	24.0816	1.2	7.4868	5.0
	200	23.7947	3.0	7.4566	10.1
	500	22.9185	8.8	7.3790	25.9
	1000	21.5328	24.6	7.2587	55.5
	$2 \times 10^4$	8.5482	377	5.1339	1282
	$10^5$	5.7358	1555.6	3.9193	6147
SCPB 1	10	11.5047	0.2	3.0063	1.3
	50	9.2959	0.6	2.7269	3.2
	100	7.2031	1.5	2.4914	6.9
	200	6.4626	2.9	2.2899	13.0
	500	5.3700	7.5	2.0635	39.2
	1000	5.0582	15.1	1.9609	70.4
SCPB 2	10	8.6325	0.15	3.3113	0.6
	50	7.8378	0.7	2.2478	3.2
	100	7.8602	1.5	2.1929	6.4
	200	6.5839	3.0	2.2913	13.4
	500	6.0361	7.4	1.9974	33.7
	1000	6.1989	14.9	1.8058	65.1

Table 3: E-SA versus two variants of SCPB on the two-stage stochastic program (65)-(66)

We also report the length of SCPB cycle along iterations in the left plot of Figure 1. A few comments are now in order on the length of the cycles with SCPB 1 and SCPB 2:

- We observe that the length of the cycles is much larger with SCPB 1.
- For SCPB 1, sequence  $\{j_k\}$  (and therefore the length of the cycles) can be computed independently of sequence  $\{x_k\}$ , before running SCPB, once constant  $R$  is known. It is worth mentioning that we have an analytic expression for  $j_k$  as a function of  $\lambda$ ,  $R$ ,  $\tau$ , and  $k$ , namely  $j_k - i_k = 0$  if  $R \geq \lambda k$  and

$$j_k - i_k = \left\lceil \frac{\log\left(\frac{R}{\lambda k}\right)}{\log(\tau)} \right\rceil$$

otherwise. Therefore, the cycle length with SCPB 1 is a piecewise constant nondecreasing function of outer iteration  $k$  and the cardinality of the set of consecutive iterations with constant cycle length increases along the cycles.

- For SCPB 2, the length of the cycles is in general small and oscillates between 2 and 20 with an average cycle length of 2.2 for the instance with  $n = 50$  while for instance with  $n = 100$  the cycle length oscillates between 2 and 10 with an average cycle length of 2.1.

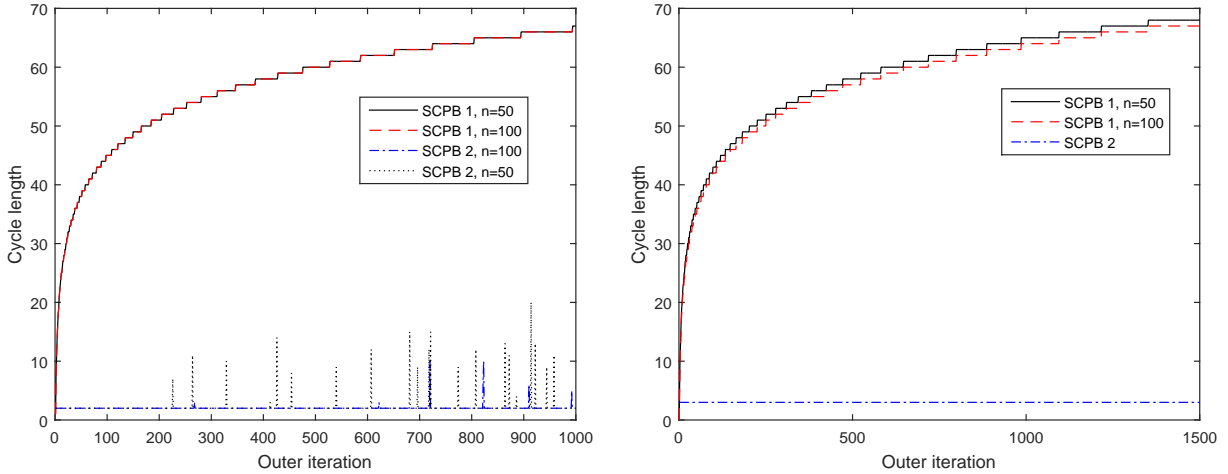


Figure 1: Cycle length for SCPB 1 and SCPB 2 applied to two-stage stochastic program (65)-(66) (left figure) and two-stage stochastic program (67)-(68) (right figure).

### 6.3 A second two-stage stochastic program

Our third test problem is the nonlinear two-stage stochastic program

$$\begin{cases} \min c^T x_1 + \mathbb{E}[\mathfrak{Q}(x_1, \xi)] \\ x_1 \in \mathbb{R}^n : \|x_1 - x_0\|_2 \leq 100 \end{cases} \quad (67)$$

where cost-to-go function  $\mathfrak{Q}(x_1, \xi)$  has nonlinear objective and constraint coupling functions and is given by

$$\mathfrak{Q}(x_1, \xi) = \begin{cases} \min_{x_2 \in \mathbb{R}^n} \frac{1}{2} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^T (\xi \xi^T + \gamma_0 I_{2n}) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \xi^T \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ s.t. \quad \|x_2 - y_0\|_2^2 + \|x_1 - x_0\|_2^2 - R^2 \leq 0. \end{cases} \quad (68)$$

Problem (67)-(68) is of form (1)-(2) where  $F(x, \xi) = c^T x + \mathfrak{Q}(x, \xi)$  with  $\mathfrak{Q}$  given by (68) and where  $h$  is the indicator function of set

$$X = \{x \in \mathbb{R}^n : \|x - x_0\|_2 \leq 100\}.$$

For problem (67), we again refer to Lemma 2.1 in [4] for the computation of stochastic subgradients  $s(x, \xi)$ . We take  $\gamma_0 = 2$  and consider for  $\xi$  a Gaussian random vector in  $\mathbb{R}^{2n}$  with the components of  $\xi$  independent with means and standard deviations randomly generated in respectively intervals  $[-5, 5]$  and  $[0, 10]$ . The components of  $c$  are generated randomly in interval  $[-1, 1]$  and we take  $R = 200$ ,  $x_0(i) = 10$  and  $y_0(i) = 1$  for  $i = 1, \dots, n$ .

We run E-SA, SCPB 1, and SCPB 2 on two instances  $B_1$  and  $B_2$  with  $n = 50$  and  $n = 100$ . For SCPB 1 and SCPB 2, we used  $K = 1500$  outer iterations. The results are reported in Tables 4 and 5. The conclusions are similar to the experiments on the stochastic utility problem: it still takes much longer for E-SA to compute a solution with given accuracy. We also report in the right plot of Figure 1 the evolution of the length of the cycles along outer iterations. The behavior of the length of these cycles is similar to what was observed for the previous problem (65)-(66). For SCPB 2 the length of the cycles is still small on all iterations but surprisingly the length of all cycles of SCPB 2 was 3 for both instances.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

-		$B_1 : n = 50$		$B_2 : n = 100$	
ALG.	N	Obj	CPU	Obj	CPU
E-SA	10	15182	0.2	18571	0.6
	50	15108	0.9	18497	4.0
	100	15017	0.9	18405	7.4
	200	14836	1.8	18222	13.9
	1000	13481	3.4	16830	63.8
	$10^5$	99.8	16.2	177.5	6275
SCPB 1	10	2981.5	0.2	4914	0.8
	50	761.2	0.8	1574	2.8
	100	288.1	1.7	679	5.7
	200	64.8	2.9	191	10.2
	1000	-4.38	14.3	-7.87	55.5
SCPB 2	10	1400.9	0.13	2766	0.5
	50	0.45	0.5	17.5	2.1
	100	-4.38	1.0	-7.88	4.1
	200	-4.38	1.9	-7.95	8.1
	1000	-4.38	9.0	-7.95	42.2

Table 4: E-SA versus two variants of SCPB on the two-stage stochastic program (67), (68)

### 6.4 Summarizing performance indicators

The computational results reported in Subsections 6.1-6.3 show that SCPB computes a good approximate solution quicker than E-SA. To properly quantify the speed-up over a fixed number of iterations  $N$ , we compute the quantity

$$100 \frac{\text{Obj}(\text{E-SA}) - \text{Obj}(\text{SCPB } i)}{\hat{F}_T(x_0) - \text{Obj}(\text{SCPB } i)} \tag{69}$$

associated with SCPB  $i$ , where  $\hat{F}_T(x_0)$  is the empirical mean (see (63)) of  $F$  with  $T = 10^4$  and  $x$  equal to the initial point  $x_0$ , and  $\text{Obj}(\text{E-SA})$ ,  $\text{Obj}(\text{SCPB } 1)$ , and  $\text{Obj}(\text{SCPB } 2)$  are the empirical means of  $F$  with  $T = 10^4$  and  $x$  equal to the final iterates output by E-SA, SCPB 1, and SCPB 2, respectively. We see that after  $N = 1000$  iterations this percentage is above 90% for most instances, which clearly shows that both variants of SCPB are faster than E-SA on all instances considered in our computational experiments.

Sample size $N$	10	50	100	200	1000
$L_1$ , SCPB 1	95.2	96.3	96.5	96.5	94.4
$L_1$ , SCPB 2	94.7	97.3	97.1	96.8	94.4
$L_2$ , SCPB 1	95.5	96.5	97.0	97.1	95.1
$L_2$ , SCPB 2	97.6	98.0	97.9	97.5	95.1
$L_3$ , SCPB 1	95.0	95.2	94.1	91.9	82.8
$L_3$ , SCPB 2	96.6	97.2	97.5	97.2	90.9
$L_4$ , SCPB 1	92.1	93.3	92.3	91.5	77.7
$L_4$ , SCPB 2	92.1	95.8	96.8	96.7	93.5
$A_1$ , SCPB 1	99.5	98.8	98.1	96.6	85.1
$A_1$ , SCPB 2	99.6	99.0	98.0	96.5	84.2
$A_2$ , SCPB 1	99.8	99.6	99.3	98.8	95.0
$A_2$ , SCPB 2	99.8	99.6	99.3	98.8	95.4
$B_1$ , SCPB 1	99.8	99.4	98.8	97.6	88.7
$B_1$ , SCPB 2	99.9	99.4	98.8	97.6	88.7
$B_2$ , SCPB 1	99.9	99.4	99.0	98.0	90.5
$B_2$ , SCPB 2	99.9	99.5	99.0	98.0	90.5

Table 5: Percentages (69) for SCPB 1 and SCPB 2

## 7 Concluding remarks

This paper proposes a single-cut stochastic composite proximal bundle framework, called SCPB, for solving SCCO problem (1)-(2) where at each iteration a problem of form (3) is solved. Two variants of this framework, which differ in the way their cycle lengths are determined, are analyzed in Sections 3 and 4, respectively. More specifically, it is shown that both variants of SCPB with proper chosen parameters has optimal iteration complexity (up to a logarithmic term) for finding an  $\varepsilon$ -solution of (1) for a large range of prox stepsizes. Practical variants of SCPB which keep their cycle lengths bounded are also proposed and numerical experiments demonstrating their excellent performance against the RSA method of [14] on the instances considered in this paper are also reported.

**Comparison with other methods:** First, we have shown in Subsection 3.2 that RSA is a special case of SCPB based on cycle rule (B1) which performs only one iteration per cycle. Second, it is worth noting that SCPB has slight similarity with the stochastic dual averaging (SDA) method discussed in [17, 26] since both methods explore the idea of aggregating cuts into a single one. However, there are essential differences between the two methods, namely: 1) while SCPB updates the prox-centers whenever a serious iteration occurs, SDA uses a fixed prox-center, and hence only performs null iterations; and 2) SDA uses variable stepsizes which have to grow sufficiently large, while SCPB uses constant prox stepsizes. In summary, viewed from the viewpoint of this paper, SDA is closest to the special case of SCPB with a single cycle and a sufficiently large prox stepsize; the difference between the latter two methods is that SDA allows the prox stepsizes within its single cycle to gradually become sufficiently large.

In summary, while RSA (resp., SDA) performs only serious (resp., null) iterations, SCPB performs a balanced mix of serious and null iterations. Hence, it is reasonable to conclude that SCPB lies in between RSA and SDA.

**Extensions.** We finally discuss some possible extensions of our analysis in this paper. A first

1  
2  
3  
4  
5  
6  
7  
8 question is how to extend SCPB and the corresponding complexity analysis if instead of condition  
9 (A3) we use the assumption that for every  $u, v \in \text{dom } h$ , we have

$$\|f'(u) - f'(v)\| \leq 2M + L\|u - v\|,$$

13 which is called a uniform  $(M, L)$ -condition in [13]. A second natural question is how to extend  
14 SCPB and its complexity analysis when either the prox stepsize  $\lambda$  and parameter  $\theta$  are allowed to  
15 change with the iteration count  $k$ . Recalling that the prox stepsize is the only ingredient of the  
16 second variant of SCPB that depends on an estimate  $M$  of  $\bar{M}$ , a third natural question is whether  
17 it is possible to develop a SCPB variant which adaptively chooses a (variable) prox stepsize without  
18 the need of knowing  $M$ . Finally, SCPB is able to solve two-stage convex stochastic programs with  
19 continuous distributions, under the assumption that the second-stage subproblems can be exactly  
20 solved (e.g., see Subsections 6.2 and 6.3). It would be interesting to extend it to the setting of  
21 *multistage* stochastic convex problems with continuous distributions.  
22  
23  
24  
25

## 26 References

- 27  
28 [1] E. D. Andersen and K.D. Andersen. *The MOSEK optimization toolbox for MATLAB manual.*  
29 *Version 9.2*, 2019. <https://www.mosek.com/documentation/>.  
30  
31 [2] J. Birge and F. Louveaux. *Introduction to Stochastic Programming.* Springer-Verlag, New  
32 York, 1997.  
33  
34 [3] J.R. Birge and F.V. Louveaux. A multicut algorithm for two-stage stochastic linear programs.  
35 *European Journal of Operational Research*, 34:384–392, 1988.  
36  
37 [4] V. Guigues. Convergence analysis of sampling-based decomposition methods for risk-averse  
38 multistage stochastic convex programs. *SIAM Journal on Optimization*, 26:2468–2494, 2016.  
39  
40 [5] V. Guigues. Multistep stochastic mirror descent for risk-averse convex stochastic programs  
41 based on extended polyhedral risk measures. *Mathematical Programming*, 163:169–212, 2017.  
42  
43 [6] V. Guigues. Inexact Stochastic Mirror Descent for two-stage nonlinear stochastic programs.  
44 *Mathematical Programming*, 187:533–577, 2021.  
45  
46 [7] V. Guigues, A. Juditsky, and A. Nemirovski. Non-asymptotic confidence bounds for the opti-  
47 mal value of a stochastic program. *Optimization Methods & Software*, 32:1033–1058, 2017.  
48  
49 [8] V. Guigues, W. Tekaya, and M. Lejeune. Regularized decomposition methods for deterministic  
50 and stochastic convex optimization and application to portfolio selection with direct transac-  
51 tion and market impact costs. *Optimization & Engineering*, 21:1133–1165, 2020.  
52  
53 [9] J.L. Higle and S. Sen. *Stochastic Decomposition.* Kluwer, Dordrecht, 1996.  
54  
55 [10] A.J. King and R.T. Rockafellar. Asymptotic theory for solutions in statistical estimation and  
56 stochastic programming. *Math. Oper. Res.*, 18:148–162, 1993.  
57  
58 [11] A. J. Kleywegt, A. Shapiro, and T. Homem-de Mello. The sample average approximation  
59 method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12(2):479–502,  
60 2002.  
61  
62  
63  
64  
65

- 1  
2  
3  
4  
5  
6  
7  
8 [12] J. Liang and R. D. C. Monteiro. A proximal bundle variant with optimal iteration-complexity  
9 for a large range of prox stepsizes. *SIAM Journal on Optimization*, 31(4):2955–2986, 2021.  
10  
11 [13] J. Liang and R. D. C. Monteiro. A unified analysis of a class of proximal bundle methods for  
12 solving hybrid convex composite optimization problems. *Available on arXiv:2110.01084*, 2021.  
13  
14 [14] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation ap-  
15 proach to stochastic programming. *SIAM J. Optim.*, 19:1574–1609, 2009.  
16  
17 [15] A. Nemirovski and D. Yudin. On Cezari’s convergence of the steepest descent method for  
18 approximating saddle point of convex-concave functions. *Soviet Math. Dokl.*, 19, 1978.  
19  
20 [16] A. Nemirovski and D.B. Yudin. *Problem complexity and method efficiency in optimization*.  
21 Wiley, 1983.  
22  
23 [17] Y. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical program-*  
24 *ming*, 120(1):221–259, 2009.  
25  
26 [18] B.T. Polyak. New stochastic approximation type procedures. *Automat. i Telemekh (English*  
27 *translation: Automation and Remote Control)*, 7:98–107, 1990.  
28  
29 [19] B.T. Polyak and A. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM*  
30 *J. Contr. and Optim.*, 30:838–855, 1992.  
31  
32 [20] H. Robbins and S. Monroe. A stochastic approximation method. *Annals of Math. Stat.*,  
33 22:400–407, 1951.  
34  
35 [21] A. Shapiro. Asymptotic analysis of stochastic programs. *Ann. Oper. Res.*, 30:169–186, 1991.  
36  
37 [22] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming: Modeling*  
38 *and Theory*. SIAM, Philadelphia, 2009.  
39  
40 [23] R.M. Van Slyke and R.J.-B. Wets. L-shaped linear programs with applications to optimal  
41 control and stochastic programming. *SIAM Journal of Applied Mathematics*, 17:638–663,  
42 1969.  
43  
44 [24] B. Verweij, S. Ahmed, A. J. Kleywegt, G. Nemhauser, and A. Shapiro. The sample aver-  
45 age approximation method applied to stochastic routing problems: a computational study.  
46 *Computational Optimization and Applications*, 24(2-3):289–333, 2003.  
47  
48 [25] W. Wang and M. A. Carreira-Perpinán. Projection onto the probability simplex: An efficient  
49 algorithm with a simple proof, and an application. *Available on arXiv:1309.1541*, 2013.  
50  
51 [26] L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization.  
52 *Journal of Machine Learning Research*, 11(88):2543–2596, 2010.  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65