A single cut proximal bundle method for stochastic convex composite optimization

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

July 18, 2022 (first revision: November 3, 2022; second revision: April 30, 2023)

Abstract

This paper considers optimization problems where the objective is the sum of a function given by an expectation and a closed convex composite function, and proposes stochastic composite proximal bundle (SCPB) methods for solving it. Complexity guarantees are established for them without requiring knowledge of parameters associated with the problem instance. Moreover, it is shown that they have optimal complexity when these problem parameters are known. 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 (RSA) method in 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 some stochastic composite proximal bundle (SCPB) variants to solve the stochastic convex composite optimization (SCCO) problem

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

where

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

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

Literature Review. Proximal bundle methods for solving the deterministic version of (1), i.e., where an oracle that outputs $f(x)$ for any x is available, have been proposed in [17, 18, 21,

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

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

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

37]. Moreover, convergence (but not complexity) analyses of proximal bundle methods have been developed for example in [7, 26, 30, 34], and their iteration complexities have been derived for example in [2, 5, 6, 15, 19, 20].

We now discuss methods for solving (the stochastic version of) problem (1). Methods for solving (1) when f can be computed exactly (e.g., ξ is a discrete random vector with small support) have been discussed for example in [3, 4] and are usually based on solving a deterministic (but large-scale) reformulation of (1), using decomposition (such as the L-shaped method [33]) possibly combined with regularization as in [12, 13].

Solution methods for problem (1) in which ξ has a continuous distribution are basically based on one of the following three ideas: i) a single (usually expensive) approximation of (1) where 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 $(\xi_1,\ldots,\xi_N)$ of ξ is constructed at the beginning of the method and is then solved to yield an approximate solution of (1) (SAA-type methods); see for instance [11, 14, 16, 31, 35] and also Chapter 5 of [32] for their complexity analysis; ii) simple approximations of (1) are constructed at every iteration based on a small (usually a single) sample and their solutions are used to obtain an approximation solution of (1) (SA-type methods); SA-type methods have been originally proposed in [29] and further extended in [9, 10, 22, 23, 25, 27, 28]; and iii) hybrid type methods which sit in between SAA and SA-type ones in that they use partial Monte Carlo averages $H_k(\cdot)$ (and their expensive subgradients) for increasing iteration indices k [13].

Contributions. Although the cutting plane methodology can be used in the context of SAA methods to solve single approximations of (1) generated at their outset, such methodology has not been used in the context of SA-type methods. This paper partially addresses this issue by developing regularized aggregated cutting plane methods for solving (1) where some of the most recent (linear approximation) cuts (in expectation) are combined, i.e., a suitable convex combination of them is chosen, so that a single aggregated cut (in expectation) is obtained. Two SCPB variants based on the aforementioned aggregated one-cut scheme are proposed which can be viewed as natural extensions of the one-cut variant developed in [20] (based on the analysis of [19]) for solving the deterministic version of (1). More specifically, at every iteration, these SCPB variants solve the prox bundle subproblem

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

where $\lambda > 0$ is the prox stepsize, x^c is the current prox-center, and Γ is the current bundle function in expectation, i.e., it satisfies $\mathbb{E}[\Gamma(\cdot)] \le \phi(\cdot)$. The prox-center remains the same for several consecutive iterations which are referred to as a cycle. In the beginning of a cycle, the prox-center is updated to $x^c \leftarrow x$ and the bundle function Γ is chosen to be the composite linear approximation $F(x,\xi) + \langle s(x,\xi), \cdot - x\rangle + h(\cdot)$ of the function $F(\cdot,\xi) + h(\cdot)$ at x for some new independent sample ξ. For other iterations of the cycle, the prox-center remains the same but Γ is set to be a convex combination of the previous bundle function and the most recent composite linear approximation as constructed above. It is then shown that both SCPB variants obtain a stochastic iterate $y \in \mathbb{R}^n$ (determined by some of the above generated x's) such that $\mathbb{E}[\phi(y)] - \phi_* \le \varepsilon$, where ϕ_* is as in (1), in $\mathcal{O}(\varepsilon^{-2})$ iterations/resolvent evaluations. To our knowledge, these are the first SA-type SCPB methods for solving SCCO problems where ξ can have either a discrete or continuous distribution. Finally, it is shown that the robust stochastic approximation (RSA) method of [22] is a special case of SCPB with a relatively small prox stepsize.

Organization of the paper. Subsection 1.1 presents basic definitions and notation used throughout the paper. Section 2 formally describes the assumptions on the SCCO problem (1), presents the SCPB scheme, and two cycles rules for determining the length of the cycles in SCPB.

Section 3 presents various convergence rate bounds for the SCPB variant based on the first cycle rule and discusses the relationship between SCPB and RSA. Section 4 provides convergence rate bounds for the SCPB variant based on the second cycle rule. Section 5 collects proofs of the main results in Sections 3 and 4. Section 6 reports the numerical experiments. Finally, Section 7 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 $\psi : \mathbb{R}^n \to (-\infty, +\infty]$ be given. The effective domain of ψ is denoted by $\operatorname{dom} \psi := \{x \in \mathbb{R}^n : \psi(x) < \infty\}$ and ψ is proper if $\operatorname{dom} \psi \neq \emptyset$. For $\varepsilon \geq 0$, the *ε-subdifferential* of ψ at $z \in \operatorname{dom} \psi$ 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 ψ at $z \in \operatorname{dom} \psi$, denoted by $\partial \psi(z)$, is by definition the set $\partial_0 \psi(z)$. Moreover, a proper function $\psi : \mathbb{R}^n \to (-\infty, +\infty]$ is μ-strongly 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 $z, u \in \operatorname{dom} \psi$ and $\alpha \in [0,1]$. Note that we say ψ is convex when $\mu = 0$. We use the notation $\xi_{[t]} = (\xi_0, \xi_1, \ldots, \xi_t)$ for the history of the sampled observations of ξ up to iteration t. Define $\ln_0^+(\cdot) := \max\{0, \ln(\cdot)\}$. Define the diameter of a set X to be $D_X := \sup\{\|x - x'\| : x, x' \in \operatorname{dom} X\}$.

2 Assumptions and two SCPB variants

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

2.1 Assumptions

Let Ξ denote the support of random vector ξ and assume that the following conditions on (1) are assumed to hold:

(A1) f and h are proper closed convex functions satisfying $\operatorname{dom} f \supset \operatorname{dom} h$;

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

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

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

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

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

We now make some observations about the above conditions. First, as in [22], condition (A2) does not require $F(\cdot, \xi)$ to be convex. Second, condition (A3) implies that

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

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

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

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

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

where $\phi(\cdot)$ is as in (1). Hence, $\ell(\cdot;x,\xi)$ is a stochastic composite linear approximation of $\phi(\cdot)$ in the sense that its expectation is a true composite linear approximation of $\phi(\cdot)$. (The terminology "composite" refers to the function h which is included in the approximation $\ell(\cdot;x,\xi)$ as is.)

2.2 Description of the two SCPB variants

Before describing the two SCPB variants, we motivate them by interpreting them as inexact implementations of the (theoretical) proximal point method for solving (1).

Their k-th cycle of iterations performs a finite number of iterations to solve the prox subproblem

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

where $\hat{x}_{k-1}$ denotes the prox-center during the cycle. Each iteration within the cycle solves a subproblem of the form

$$x = \underset{u\in\mathbb{R}^n}{\operatorname{argmin}} \left\{\mathcal{A}(u) + h(u) + \frac{1}{2\lambda}\|u - \hat{x}_{k-1}\|^2\right\} \tag{7}$$

where $\mathcal{A}(\cdot)$ is an affine bundle for f in expectation, i.e., an affine function such that $\mathbb{E}[\mathcal{A}(\cdot)] \leq f(\cdot)$. (This type of bundle has been considered in the inexact proximal point approach considered in [20] where it is referred to as a one-cut bundle for f.) The bundle $\mathcal{A}^+$ for the next subproblem in the cycle is then taken to be a linear combination of the current bundle $\mathcal{A}$ and a newly generated stochastic linear approximation of f of the form $F(x,\xi)+\langle s(x,\xi), \cdot - x\rangle$. Moreover, the first iteration of every cycle starts by setting the prox-center to the most recently generated x as in (7) and the bundle to the most recently generated stochastic linear approximation of f at x.

Both SCPB variants are based on the SCPB scheme described below. As stated below, the scheme is not a completely specified algorithm since its step 2 does not describe how to select the index j_k. Two rules for doing so, and hence the complete description of the two aforementioned SCPB variants, are then given following the statement of the scheme.

At every iteration $j \geq 1$, the SCPB scheme samples an independent realization ξ_{j-1} of ξ.

SCPB

Input: Scalars $\lambda, \theta > 0$, integer $K \geq 1$, and initial point $x_0 \in \operatorname{dom} h$.

0. Set $j = k = 1$, $j_0 = 0$, and
$$\tau = \frac{\theta K}{\theta K + 1}; \tag{8}$$

1. take a sample ξ_{j-1} of r.v. ξ independent from the previous samples $\xi_0, \ldots, \xi_{j-2}$ and compute
$$x_j^c = \begin{cases} x_{j_{k-1}}, & \text{if } j = j_{k-1}+1, \\ x_{j-1}^c, & \text{otherwise,} \end{cases} \tag{9}$$

$$S_j = \begin{cases} s(x_{j_{k-1}}, \xi_{j_{k-1}}), & \text{if } j = j_{k-1}+1, \\ (1-\tau)s(x_{j-1}, \xi_{j-1}) + \tau S_{j-1}, & \text{otherwise}, \end{cases} \tag{10}$$

$$x_j = \underset{u\in\mathbb{R}^n}{\operatorname{argmin}} \left\{ h(u) + \langle S_j, u\rangle + \frac{1}{2\lambda}\|u - x_j^c\|^2 \right\}, \tag{11}$$

and

$$y_j = \begin{cases} x_j, & \text{if } j = j_{k-1}+1, \\ (1-\tau)x_j + \tau y_{j-1}, & \text{otherwise;} \end{cases} \tag{12}$$

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

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

   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 set $k \leftarrow k+1$ and $j \leftarrow j+1$, and go to step 1; otherwise, compute

$$\hat{y}_K^a = \frac{1}{\lceil K/2 \rceil} \sum_{k=\lfloor K/2\rfloor+1}^{K} \hat{y}_k \tag{13}$$

   and **stop**.

Output: $\hat{y}_K^a$.

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

$$\mathcal{C}_k := \{i_k, \ldots, j_k\}, \quad \text{where} \quad i_k := j_{k-1}+1, \tag{14}$$

it immediately follows that k counts the number of cycles generated by SCPB. Third, step 1 determines two types of iterations depending on whether $j = j_k$ (serious iteration) or $j \in \mathcal{C}_k \setminus \{j_k\}$ (null iteration). Hence, the last 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 ∂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 (9), the sequence $\{x_j\}$ determined by (11), and the sequence $\{y_j\}$ given by (12). Third, it follows from (9) that $x_j^c = x_{j_{k-1}}$ for every $j \in \mathcal{C}_k$. Hence, the prox-center x_j^c remains constant between consecutive iterations within a cycle and (possibly) changes only at the beginning of the first iteration of the following cycle. Fourth, $\{\hat{y}_k\}$ is the subsequence of $\{y_j\}$ consisting of all the last cycle iterates y_{j_k} generated by SCPB. Fifth, the convergence rates for the two specific variants of the SCPB scheme described below are with respect to the average of the iterates $\hat{y}_{\lfloor K/2\rfloor+1}, \ldots, \hat{y}_K$, namely, the point $\hat{y}_K^a$ as in (13) (see Theorems 3.1 and 4.1 below).

As already mentioned in the second paragraph preceding the description of SCPB, the scheme is not completely specified since its step 2 does not describe how to select j_k. We now describe two cycle rules for doing so which depend on a pre-specified parameter $R > 0$, namely:

(B1) for every $k \ge 1$, let j_k be the smallest integer $\ge i_k$ such that $\lambda k \tau^{j_k - i_k} \le R$;

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

$$\lambda k \tau^{j_k - i_k} \left(F(x_{i_k}, \xi_{i_k}) - \tilde{\ell}_k(x_{i_k}) - \frac{1}{2\lambda}\|x_{i_k} - x^c_{i_k}\|^2 \right) \leq R \tag{15}$$

where i_k is as in (14) and

$$\tilde{\ell}_k(\cdot) := F(x_{i_k - 1}, \xi_{i_k - 1}) + \langle s(x_{i_k-1}, \xi_{i_k-1}), \cdot - x_{i_k-1} \rangle. \tag{16}$$

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 (15) 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. Third, the length of cycle $\mathcal{C}_k$ for both rules above depends on the cycle index k. Hence, even though (B1) is deterministic, the length of the cycles generated by it changes with k.

Throughout our presentation, SCPB based on cycle rule (B1) (resp., (B2)) is referred to as SCPB1 (resp., SCPB2).

3 Complexity results for SCPB1

This section presents the main complexity results for SCPB1 under various assumptions and discusses the relationship between SCPB1 and RSA.

3.1 Convergence rate bounds of SCPB1 with bounded $\operatorname{dom} h$

The following result states a general convergence rate result for SCPB1 that holds for bounded $\operatorname{dom} h$ and for any choice of input (λ, θ, K) in SCPB1 and constant R as in (B1). The proof is postponed to Subsection 5.2.

Theorem 3.1. *Assume that conditions (A1)-(A4) hold and* $\operatorname{dom} h$ *has a finite diameter* $D_h \geq 0$. *Then, for any given* $(\lambda, \theta, K) \in \mathbb{R}^2_{++} \times \mathbb{N}_{++}$ *and* $R > 0$, *SCPB1 with any input* (λ, θ, K) *and constant* R *in (B1) satisfies the following statements:*

a) the number of iterations within the k-th cycle $\mathcal{C}_k$ *(see* (14)*) is bounded by*

$$\left\lceil (\theta K + 1) \ln_0^+ \left(\frac{\lambda k}{R} \right) \right\rceil + 1; \tag{17}$$

b) we have

$$\mathbb{E}[\phi(\hat{y}^a_K)] - \phi_* \leq \frac{1}{K} \left(\frac{D_h^2}{\lambda} + \frac{6R \min\{\lambda \bar{M}^2, \bar{M} D_h\}}{\lambda} + \frac{2\lambda \bar{M}^2}{\theta} \right) \tag{18}$$

where D_h *is the diameter of* $\operatorname{dom} h$.

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 (17). Second, (18) gives a bound on the expected primal gap $\mathbb{E}[\phi(\hat{y}^a_K)] - \phi_*$ in terms of K, and hence provides a sufficient condition on how large K should be chosen for SCPB1 to generate a desired approximate solution.

Even though Theorem 3.1 holds for the general case in which $\operatorname{dom} h$ is unbounded, all its corollaries stated in this subsection and Subsection 3.3 assume that $\operatorname{dom} h$ is bounded. For any given (λ, K), the following result describes a convergence rate bound for SCPB1 with a specific choice of (θ, R).

Corollary 3.2. *Assume that conditions (A1)-(A4) hold and* $\operatorname{dom} h$ *has a finite diameter* $D_h > 0$. *Let a pair* (λ, K) *be given and consider SCPB1 with input* (λ, θ, K) *and* R *in (B1) given by*

$$\theta = \frac{2\lambda^2 M^2}{D^2}, \quad R = \frac{D}{6M} \tag{19}$$

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

a) we have

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

where

$$\kappa_D := \frac{D_h^2}{D^2}, \quad \kappa_M := \frac{\bar{M}^2}{M^2}; \tag{20}$$

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

$$\mathcal{O}\left(\frac{\lambda^2 M^2 K^2}{D^2} + K\right). \tag{21}$$

Proof: a) Using (18), the definitions of κ_D and κ_M in (20), and the definitions of θ and R in (19), we get

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

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 κ_D and κ_M while in the last inequality we have used the relation $\sqrt{ab} \leq (a+b)/2$ for every $a, b \geq 0$.

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 (21) in view of θ as in (19). ■

We now argue that the overall iteration (and sample) complexity of the SCPB1 variant of Corollary 3.2 for finding an ε-solution x of (1), i.e., one that satisfies $\mathbb{E}[\phi(x)] - \phi_* \leq \varepsilon$, is optimal for a large range of prox stepsizes. Indeed, setting $K = \lceil T_\varepsilon \rceil$ where

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

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 (21) that the expected overall iteration complexity of SCPB1 is bounded by

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

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

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

Moreover, under the assumption that the prox stepsize λ lies in the interval $[\varepsilon/M^2, D^2/\varepsilon]$, the above complexity bound further reduces to $\mathcal{O}(M^2D^2/\varepsilon^2)$, which is known to be the optimal complexity of finding an ε-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 [24]).

3.2 Relationship between SCPB1 and the RSA method of [22]

We argue in this subsection that RSA can be viewed as a special instance of SCPB1 where every cycle $\mathcal{C}_k$ contains only one index (or equivalently, an instance for which every iteration is serious).

Recall that the RSA method of [22], which is developed under the assumption that h is the indicator function of a nonempty compact 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

$$x_j = \operatorname*{argmin}_{u\in X} \left\{\langle s(x_{j-1}, \xi_{j-1}), u\rangle + \frac{1}{2\lambda}\|u - x_{j-1}\|^2\right\} \qquad \forall j = 1, \ldots, N. \tag{22}$$

For the purpose of reviewing the iteration complexity of RSA, assume that D is an upper bound on the diameter of X and M is an upper bound on $\bar{M}$. For $1 \leq i \leq N$, let $\tilde{x}_i^N$ denote the average of the iterates $\{x_j\}_{j=i}^N$, i.e.,

$$\tilde{x}_i^N = \frac{1}{N - i + 1}\sum_{j=i}^N x_j. \tag{23}$$

It is shown in equation (2.24) of [22] that if the stepsize $\lambda > 0$ is chosen as

$$\lambda = \frac{\alpha D}{M\sqrt{N}} \tag{24}$$

for some fixed scalar $\alpha > 0$, then the ergodic iterate $\tilde{x}_i^N$ with $i = \lfloor N/2\rfloor + 1$ satisfies

$$\mathbb{E}[\phi(\tilde{x}^N_{\lfloor N/2\rfloor+1})] - \phi_* \leq \max\{\alpha, \alpha^{-1}\}\frac{9DM}{2\sqrt{N}}. \tag{25}$$

Hence, for a given tolerance $\varepsilon > 0$, the smallest N satisfying $\mathbb{E}[\phi(\tilde{x}^N_{\lfloor N/2\rfloor+1})] - \phi_* \leq \varepsilon$ has the property that

$$N = \mathcal{O}\left(\frac{\max\{\alpha^2, \alpha^{-2}\}M^2D^2}{\varepsilon^2}\right). \tag{26}$$

It turns out that RSA is a special case of SCPB1 with R in (B1) given by

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

Indeed, it follows from the above choice of R and λ as in (24) 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 SCPB1 variant is a serious one and K is its total number of iterations.

3.3 A practical SCPB1 variant

From a computational point of view, the choice of θ in Corollary 3.2 usually results in the quantity θK, and hence the inner complexity bound (17), being large. The following result provides a practical variant of SCPB1 with an alternative choice for θ and R which partially remedies the above drawback by forcing θK to be constant. A nice feature of this variant is that it is able to choose large prox stepsizes without loosing the optimality of its overall iteration complexity.

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

$$\theta = \frac{C}{K}, \quad R = \frac{D}{M}, \quad \lambda = \frac{\sqrt{C}D}{M\sqrt{K}} \tag{27}$$

where (D, M) *is an estimate for the pair* $(D_h, \bar{M})$. *Then, the following statements about SCPB1 with input* (λ, θ, K) *and* R *as above hold:*

a) we have

$$\mathbb{E}[\phi(\hat{y}_K^a)] - \phi_* \leq \frac{(4\kappa_D + 5\kappa_M)DM}{\sqrt{CK}} \tag{28}$$

where κ_D *and* κ_M *are as in* (20)*;*

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

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

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 (18), the definitions of κ_D and κ_M in (20), and the definitions of θ and R in (27), we get

$$\begin{aligned}
\mathbb{E}[\phi(\hat{y}_K^a)] - \phi_* &\leq \frac{1}{K}\left(\frac{\kappa_D D^2}{\lambda} + \frac{6D\min\{\lambda\bar{M}^2, \bar{M}D_h\}}{\lambda M} + \frac{2\lambda K\bar{M}^2}{C}\right) \\
&\leq \frac{1}{\lambda K}\left(\kappa_D D^2 + \frac{6D\bar{M}D_h}{M}\right) + \frac{2\lambda M^2\kappa_M}{C} \\
&= \frac{D^2}{\lambda K}(\kappa_D + 6\sqrt{\kappa_M\kappa_D}) + \frac{2\lambda M^2\kappa_M}{C},
\end{aligned} \tag{29}$$

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 κ_M. It follows from (29) and the fact that $\sqrt{\kappa_M\kappa_D} \leq (\kappa_M + \kappa_D)/2$ that

$$\mathbb{E}[\phi(\hat{y}_K^a)] - \phi_* \leq \frac{(4\kappa_D + 3\kappa_M)D^2}{\lambda K} + \frac{2\kappa_M\lambda M^2}{C}.$$

Finally, the above bound with λ as in (27) implies (28).

b) This statement immediately follows from Theorem 3.1(a) with θ, R, and λ as in (27).

c) This statement follows from (b) and the fact that SCPB1 has K cycles. ■

We now make two remarks about the practical SCPB1 variant of Corollary 3.3. First, although θ in (27) depends neither on M nor D, the choice of R depends on both of these estimates. On the other hand, SCPB2 will be analyzed in Section 4 where θ 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 SCPB1 is bounded (up to a logarithmic term) by the a priori (user specified) constant C. Thus, the SCPB1 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).

In the remaining part of this subsection, we compare the performance of RSA and SCPB1 when both use the prox stepsize λ as in (27) for some relatively large scalar $C \geq 1$. For this discussion, we assume that their performance measure is the overall iteration complexity (or sample complexity) for finding an ε-solution of (1). For simplicity, we assume as in Subsection 3.2 that h is the indicator function of a nonempty closed convex set and that the estimation pair (D, M) satisfies $D \geq D_h$ and $M \geq \bar{M}$, or equivalently, $\kappa_D \leq 1$ and $\kappa_M \leq 1$.

We first consider the performance (see the previous paragraph) of SCPB1. It follows from Corollary 3.3(a) that there exists $K = \mathcal{O}(D^2M^2/(C\varepsilon^2))$ such that $\hat{y}_K^a$ is an ε-solution of (1). Hence, it follows from Corollary 3.3(c) that the performance of SCPB1 is $\mathcal{O}(M^2D^2/\varepsilon^2)$. In conclusion, SCPB1 with the above choice of K is able to choose a prox stepsize λ as in (27) with a large constant C while preserving its optimal performance. We now consider the performance of RSA. It follows from (24) and (26) with $\alpha = \sqrt{C}$ that the performance of RSA is $\mathcal{O}\left(CD^2M^2/\varepsilon^2\right)$. In conclusion, while both RSA and SCPB1 with prox stepsize λ as in (27) have their own performance guarantee, the one for RSA becomes worse than that of SCPB1 as C becomes large.

Finally, although the SCPB1 variant of Corollary 3.3 chooses λ as in (27), our numerical experiments uses 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 SCPB1 with this aggressive choice of λ substantially outperforms RSA on the (relatively small number of) instances considered in our experiment.

4 Complexity results for SCPB2

This section provides the main complexity results for SCPB2.

The following result is an analogue of Theorem 3.1 and describes the convergence rate bound for the SCPB2 without imposing any condition on its input (λ, θ, K) and the constant R in (B2). The proof is postponed to Subsection 5.3.

Theorem 4.1. *Assume that conditions (A1)-(A4) hold and* $\operatorname{dom} h$ *has a finite diameter* $D_h > 0$. *Then, SCPB2 satisfies the following statements:*

a) the expected number of iterations within the k-th cycle $\mathcal{C}_k$ (see (14)*) is bounded by*

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

b) we have

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

Following a similar argument as in the paragraph following Corollary 3.2, it can be shown that SCPB2 has optimal iteration complexity (up to a logarithmic term) for finding an ε-solution of (1) for a large range of prox stepsizes.

The following result is the analogue of Corollary 3.3 when SCPB2 is implemented using cycle rule (B2) instead of (B1). As in Corollary 3.3, it forces the quantity θK to be constant but, in contrast to the choice of R of Corollary 3.3, its choice for R does not depend on an estimate M for $\bar{M}$.

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

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

where D *is an estimate for* D_h *and* M *is an estimate for* $\bar{M}$. *Then, the following statements for SCPB2 with input* (λ, θ, K) *based on cycle rule (B2) with* R, θ, *and* λ *as above hold:*

a) we have

$$\mathbb{E}[\phi(\hat{y}_K^a)] - \phi_* \leq \frac{(3 + \kappa_D + 4\kappa_M)DM}{\sqrt{CK}} \tag{32}$$

where κ_D *and* κ_M *are as in* (20)*;*

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

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

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 Theorem 4.1(b) with θ and R as in (31) and the definitions of κ_D and κ_M in (20), we have

$$\mathbb{E}[\phi(\hat{y}_K^a)] - \phi_* \leq \frac{(3+\kappa_D)D^2}{\lambda K} + \frac{4\kappa_M \lambda M^2}{C},$$

which together with λ in (31) implies (32).

b) This statement follows from (30) with θ, R, and λ as in (31) and the definition of κ_M in (20).

c) This statement follows from (b) and the fact that SCPB2 has K cycles. ■

5 Proofs of main results in Sections 3 and 4

This section contains three subsections. The first one presents some technical results that apply to the SCPB scheme regardless of how the index j_k is chosen in step 2. The second and third ones are then devoted to the proofs of Theorems 3.1 and 4.1, respectively.

5.1 Proofs of some technical results

We assume (A1)-(A4) hold throughout this subsection. Recall that for every $j \ge 0$

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

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

$$s_j := s(x_j, \xi_j). \tag{33}$$

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

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

and

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

where $\Phi(\cdot, \xi)$ and $\ell(\cdot; x, \xi)$ are as in (5) and and $\tilde{\ell}_k(\cdot)$ is as in (16). It is easy to see from (10), (11), and the above definition of Γ_j that

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

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

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

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

$$\mathbb{E}[\phi(y_j)] \le \mathbb{E}[u_j], \tag{38}$$

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

Proof: Observe that x_j is a function of $\xi_{[j-1]}$ and not of ξ_j. Hence, x_j is independent of ξ_j in view of the fact that ξ_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 $k \ge 1$, (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 (12), (34), and (37) that

$$\mathbb{E}[u_j] \overset{(34)}{=} \mathbb{E}[\Phi(x_j, \xi_j)] \overset{(37)}{=} \mathbb{E}[\phi(x_j)] \overset{(12)}{=} \mathbb{E}[\phi(y_j)],$$

and from (35) with $j = i_k$, (16), and assumptions (A1)-(A2) that for every $x \in \operatorname{dom} h$,

$$\mathbb{E}[\Gamma_j(x)] \overset{(35)}{=} \mathbb{E}[\tilde{\ell}_k(x) + h(x)] \overset{(16),(A2)}{=} f(x_{i_k-1}) + \langle f'(x_{i_k-1}), x - x_{i_k-1} \rangle + h(x) \overset{(A1)}{\le} \phi(x).$$

Let j be such that $j > i_k$ and (38) and (39) hold for j. Then, it follows from (12), (34), the fact that (38) holds for j, and the convexity of ϕ, that

$$\mathbb{E}[u_{j+1}] \overset{(34),(38)}{\geq} (1-\tau)\mathbb{E}[\phi(x_{j+1})] + \tau\mathbb{E}[\phi(y_j)] \geq \mathbb{E}[\phi((1-\tau)x_{j+1} + \tau y_j)] \overset{(12)}{=} \mathbb{E}[\phi(y_{j+1})],$$

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

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

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

It is worth noting that the proof of (39) is strongly based on the fact that Γ_j is a convex combination of affine functions whose expected values are underneath ϕ. Moreover, this inequality would not necessarily be true if Γ_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 $j \geq i_k$, we have:*

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

Proof: Using the definitions of ϕ 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). ■

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*

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

where λ, θ, 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*

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

where τ is as in (8)*, and hence*

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

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

$$\begin{aligned}
\Gamma^\lambda_j(x_j) &= (1-\tau)\ell(x_j; x_{j-1}, \xi_{j-1}) + \tau\Gamma_{j-1}(x_j) + \frac{1}{2\lambda}\|x_j - x^c_j\|^2 \\
&\geq (1-\tau)\ell(x_j; x_{j-1}, \xi_{j-1}) + \tau\left[\Gamma_{j-1}(x_j) + \frac{1}{2\lambda}\|x_j - x^c_{j-1}\|^2\right] \\
&= (1-\tau)\ell(x_j; x_{j-1}, \xi_{j-1}) + \tau\Gamma^\lambda_{j-1}(x_j) \\
&\geq (1-\tau)\ell(x_j; x_{j-1}, \xi_{j-1}) + \tau\left[\Gamma^\lambda_{j-1}(x_{j-1}) + \frac{1}{2\lambda}\|x_j - x_{j-1}\|^2\right],
\end{aligned} \tag{44}$$

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 Γ_j^λ is $(1/\lambda)$-strongly convex and x_{j-1} is the minimizer of Γ_{j-1}^λ (see (36)). Using (8), (40) and (44), we have

$$\begin{aligned}\Gamma_j^\lambda(x_j) - \tau\Gamma_{j-1}^\lambda(x_{j-1}) &\overset{(8),(44)}{\geq} (1-\tau)\left[\ell(x_j; x_{j-1}, \xi_{j-1}) + \frac{\theta K}{2\lambda}\|x_j - x_{j-1}\|^2\right] \\ &\overset{(40)}{\geq} (1-\tau)\phi(x_j) + (1-\tau)\left[\frac{\theta K}{2\lambda}\|x_j - x_{j-1}\|^2 - (\bar{M} + \|s_{j-1}\|)\|x_j - x_{j-1}\|\right] \\ &\geq (1-\tau)\phi(x_j) - (1-\tau)\frac{\lambda(\bar{M} + \|s_{j-1}\|)^2}{2\theta K}\end{aligned}$$

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

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

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

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

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

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*

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

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

$$\mathbb{E}[\|s_j\|^2 \,|\, \xi_{[\ell]}] \leq \bar{M}^2. \tag{46}$$

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. ■

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* (14) *and* (41)*, respectively.*

Proof: Let

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

Using the definitions of t_j and Γ_j^λ in (41) and (36), respectively, (34) with $j = i_k = j_{k-1} + 1$ (see (14)), we have

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

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

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

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

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

It follows from (33), (47), and conditions (A2) and (A3) that

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

Hence, the lemma follows by taking expectations of (49) and (50) and using the above three relations. ■

We emphasize that all results developed in this subsection hold regardless of the way j_k is chosen in step 2. On the other hand, the results in the following two subsections strongly use the fact that j_k is chosen according to either (B1) or (B2).

5.2 Proof of Theorem 3.1

This subsection is devoted to the proof of Theorem 3.1. The following result derives a bound in expectation for t_{j_k} when j_k is chosen according to cycle rule (B1).

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

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

where t_j is as in (41).

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

$$t_{j_k} \leq \tau^{j_k - i_k} t_{i_k} + (1-\tau) \sum_{i=i_k+1}^{j_k} \tau^{j_k - i} b_i. \tag{52}$$

In view of (B1) and (14), it follows that j_k and i_k are both deterministic. Hence, taking expectation of the above inequality and using the last inequality in (45), cycle rule (B1), and Lemma 5.5, we conclude that

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

and hence that (51) holds. ■

It is worth noting that rule (B1) plays an important role in showing that the expectation of the first term on the right-hand side of (52) is $\mathcal{O}(1/k)$. On the other hand, the proof of the $\mathcal{O}(1/K)$ bound for the expectation of the second term on the right-hand side of (52) does not depend on rule (B1) but on the fact that the expectation of b_j is small, namely, $\mathcal{O}(1/K)$ (see (45) and its definition in (41)). In conclusion, rule (B1) provides a way to estimate the magnitude of $\mathbb{E}[t_{j_k}]$, a quantity which by itself is intractable to compute exactly.

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

$$\hat{\Gamma}_k := \Gamma_{j_k} \quad \forall k \geq 1 \tag{53}$$

and

$$\hat{x}_k := x_{j_k}, \quad \hat{u}_k := u_{j_k}. \tag{54}$$

In what follows, we make some remarks about the above "outer" sequences which follow as immediate consequences of the results developed above. In view of the above definitions, relation (36) with $j = j_k$, and the way the prox-centers x_j^c are updated in (9), we have that

$$\hat{x}_k = \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ \hat{\Gamma}_k(x) + \frac{1}{2\lambda} \|x - \hat{x}_{k-1}\|^2 \right\} \quad \forall k \geq 1. \tag{55}$$

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

$$\mathbb{E}[\phi(\hat{y}_k)] \leq \mathbb{E}[\hat{u}_k] \tag{56}$$

and

$$\mathbb{E}[\hat{\Gamma}_k(z)] \leq \phi(z) \quad \forall z \in \operatorname{dom} h. \tag{57}$$

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 $z \in \operatorname{dom} h$, 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

$$d_k(z) := \left(\mathbb{E}[\|\hat{x}_k - z\|^2]\right)^{1/2}. \tag{58}$$

Proof: First observe that (53), (54), and the definitions of Γ_j^λ and t_j in (36) and (41), respectively, imply that (51) is equivalent to

$$\mathbb{E}\left[\hat{u}_k - \hat{\Gamma}_k(\hat{x}_k) - \frac{1}{2\lambda}\|\hat{x}_k - \hat{x}_{k-1}\|^2\right] \le \frac{2R\min\{\lambda\bar{M}^2, \bar{M}D_h\}}{\lambda k} + \frac{2\lambda\bar{M}^2}{\theta K}. \tag{59}$$

It follows from (55) and the fact that the objective function of (55) is $(1/\lambda)$-strongly convex that for every $z \in \operatorname{dom} h$,

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

and hence that

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

Taking expectation of the above inequality and using (58) and (59), we conclude that

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

which, in view of (56) and (57), immediately implies the conclusion of the lemma. ■

We are now in a position to prove Theorem 3.1.

Proof of Theorem 3.1: a) This statement directly follows from (8), cycle rule (B1), the definition of $\ln_0^+$, and the facts that $|\mathcal{C}_k| = j_k - i_k + 1$ and $\ln\tau^{-1} \ge 1 - \tau$.

b) Using the definition of $\hat{y}_K^a$ in (13), Lemma 5.7 with $z = x^* \in X^*$, and the facts that $\lceil K/2 \rceil \ge K/2$ and

$$\sum_{k=\lfloor K/2 \rfloor+1}^{K} \frac{1}{k} \le \int_{\lfloor K/2 \rfloor}^{K} \frac{1}{x}dx = \ln\frac{K}{\lfloor K/2 \rfloor} \le \ln\frac{K}{K/4} = \ln 4 \le \frac{3}{2} \quad \forall K \ge 2,$$

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

$$\begin{aligned}
\mathbb{E}[\phi(\hat{y}_K^a)] - \phi_* &\le \frac{1}{\lceil K/2 \rceil}\sum_{k=\lfloor K/2 \rfloor+1}^{K} (\mathbb{E}[\phi(\hat{y}_k)] - \phi_*) \\
&\le \frac{1}{\lceil K/2 \rceil}\sum_{k=\lfloor K/2 \rfloor+1}^{K} \left(\frac{2R\min\{\lambda\bar{M}^2, \bar{M}D_h\}}{\lambda k} + \frac{2\lambda\bar{M}^2}{\theta K} + \frac{1}{2\lambda}[d_{k-1}(x^*)]^2 - \frac{1}{2\lambda}[d_k(x^*)]^2\right) \\
&\le \frac{6R\min\{\lambda\bar{M}^2, \bar{M}D_h\}}{\lambda K} + \frac{2\lambda\bar{M}^2}{\theta K} + \frac{\left[d_{\lfloor K/2 \rfloor}(x^*)\right]^2}{\lambda K}
\end{aligned} \tag{60}$$

where the first inequality is due to the convexity of ϕ. It is also easy to see from Lemma 5.7 that (60) holds for $K = 1$. Then (18) follows from (60) and the fact that $d_{\lfloor K/2 \rfloor}(x^*) \le D_h$. ■

The above result strongly uses the fact that $\operatorname{dom} h$ is bounded in view of the last inequality of its proof.

We end this subsection by describing a complexity bound for a slightly modified SCPB1 variant which is derived without assuming that $\operatorname{dom} h$ is bounded. We start by describing the two changes one needs to make to SCPB1 in order to obtain the aforementioned variant. First, instead of the point $\hat{y}_K^a$ as in (13), it outputs

$$\bar{y}_K^a = \frac{1}{K}\sum_{k=1}^{K} \hat{y}_k. \tag{61}$$

Second, instead of computing j_k as in (B1), it sets j_k as the smallest integer greater than or equal to i_k such that $\lambda K \tau^{j_k - i_k} \le R$.

Theorem 5.8. *Assume that conditions (A1)-(A4) hold and let d_0 denote the distance of the initial point x_0 to the optimal set X^*, i.e.,*

$$d_0 := \|x_0 - x_0^*\|, \quad \text{where} \quad x_0^* := \operatorname{argmin}\{\|x_0 - x^*\| : x^* \in X^*\}. \tag{62}$$

Then, the aforementioned SCPB1 variant 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) there holds

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

Proof: (a) The proof of (a) is similar to that of Theorem 3.1(a).

(b) First note that the new way of choosing j_k and slightly different arguments than the ones used in the proofs of Proposition 5.6 and Lemma 5.7 imply that for every $z \in \operatorname{dom} h$,

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

Using the fact that ϕ is convex and the definition of $\bar{y}_K^a$ in (61), and summing the above inequality from $k = 1$ to K, we conclude that for every $z \in \operatorname{dom} h$,

$$\begin{aligned}
\mathbb{E}[\phi(\bar{y}_K^a)] - \phi(z) &\le \frac{1}{K}\sum_{k=1}^{K}[\mathbb{E}[\phi(\hat{y}_k)] - \phi(z)] \\
&\le \frac{1}{K}\sum_{k=1}^{K}\left(\frac{2R\bar{M}^2}{K} + \frac{2\lambda\bar{M}^2}{\theta K} + \frac{1}{2\lambda}[d_{k-1}(z)]^2 - \frac{1}{2\lambda}[d_k(z)]^2\right) \\
&\le \frac{2R\bar{M}^2}{K} + \frac{2\lambda\bar{M}^2}{\theta K} + \frac{[d_0(z)]^2}{2\lambda K}.
\end{aligned}$$

The statement now follows from the above inequality with $z = x_0^*$ where x_0^* is as in (62). ■

5.3 Proof of Theorem 4.1

The following result, which is an analogue of Proposition 5.6 with cycle rule (B1) replaced by (B2), derives a bound on t_{j_k} in expectation.

Proposition 5.9. *In addition to conditions (A1)-(A4), assume also that cycle rule (B2) is used. For every $k \geq 1$, we have*

$$\mathbb{E}[t_{j_k}] \leq \frac{R}{\lambda k} + \frac{2\lambda \bar{M}^2}{\theta K} + \frac{2\lambda \bar{M}^2}{\theta^2 K^2}. \tag{63}$$

Proof: Using (43) with $j = j_k$, we conclude that

$$t_{j_k} - (1-\tau) \sum_{i=i_k+2}^{j_k} \tau^{j_k - i} b_i \overset{(43)}{\leq} \tau^{j_k - i_k} t_{i_k} + (1-\tau)\tau^{j_k - i_k - 1} b_{i_k+1} \overset{(B2)}{\leq} \frac{R}{\lambda k} + (1-\tau) b_{i_k+1} \tag{64}$$

where the second inequality is due to cycle rule (B2), the observation that (B2) is equivalent to $\lambda k \tau^{j_k - i_k} t_{i_k} \leq R$, and $\tau \in (0,1)$ in view of (8). 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

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

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 (8), 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 (63) holds. ■

It is worth noting that rule (B2) plays an important role in showing that the first term on the right-hand side of the (64) is $\mathcal{O}(1/k)$.

The following result is an analogue of Lemma 5.7 with (B1) replaced by (B2).

Lemma 5.10. *In addition to conditions (A1)-(A4), assume also that cycle rule (B2) is used. Then, for every $z \in \operatorname{dom} h$ 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 (58).

Proof: First observe that the definitions of Γ_j^λ and t_j in (36) and (41), respectively, imply that (63) is equivalent to

$$\mathbb{E}\left[\hat{u}_k - \hat{\Gamma}_k(\hat{x}_k) - \frac{1}{2\lambda}\|\hat{x}_k - \hat{x}_{k-1}\|^2\right] \leq \frac{R}{\lambda k} + \frac{2\lambda \bar{M}^2}{\theta K} + \frac{2\lambda \bar{M}^2}{\theta^2 K^2}. \tag{65}$$

The remaining part of the proof is now similar to that of Lemma 5.7 except that (65) is used in place of (59). ■

We are now ready to prove Theorem 4.1.

Proof of Theorem 4.1: a) Using (8), (15), 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 5.10 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 [22] and the two two-stage nonlinear stochastic programs considered in the numerical experiments of [10]. 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

$$\min\{f(x) := \mathbb{E}[F(x, \xi)] : x \in X\}. \tag{66}$$

The implementations are coded in MATLAB, using Mosek optimization library [1] to generate stochastic oracles $F(x, \xi)$ and $s(x, \xi)$, and run on a laptop with Intel i7, 1.80 GHz processor. For solving subproblem (11), we do not use Mosek but implement algorithms for projection onto X. In particular, we follow [36] to implement an exact algorithm for projection onto the unit 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 [22] (as explained in [22], in terms of Section 2.3 of [22], this is mirror descent robust SA with Euclidean setup). In the notation of [22], 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 (23) 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 [22] by

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

where D_X is the diameter of the feasible set X in (66).[1] As in [22], we take $\theta = 0.1$ which was calibrated in [22] using an instance of the stochastic utility problem. For each problem, the value of

[1] Parameter M is denoted by M_* in [22].

M is estimated as in [22] taking the maximum of $\|s(\cdot,\cdot)\|$ over 10,000 calls to the stochastic oracle at randomly generated feasible solutions.

Remark: In [22], 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 $N-i+1 = \min[2^k, N]$, $k = 0, 1, \ldots, \log_2(N)$ and an additional sample was used to estimate the objective at these candidate solutions in order to choose the best of these candidates. In [22], 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 SCPB1. SCPB1 uses parameters θ, τ, R, and λ 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 β_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 SCPB2. SCPB2 uses parameters θ, τ, R, and λ 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 β_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;
- `Obj` the empirical mean
$$\hat{F}_T(x) := \frac{1}{T}\sum_{i=1}^T F(x, \xi_i) \tag{67}$$
of F at x based on a sample $\xi_1, \ldots, \xi_T$ of ξ 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` 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

$$X=\left\{x \in \mathbb{R}^n : \sum_{i=1}^{n} x_i = 1,\, x \geq 0\right\}, \tag{68}$$

$\xi_i \sim \mathcal{N}(0,1)$ are independent and $\phi(t)=\max(v_1+s_1 t, \ldots, v_m+s_m t)$ is piecewise convex with 10 breakpoints, all located on $[0,1]$[2].

Calibration of β_1 and β_2. We run SCPB1 and SCPB2 with 7 values of β_1 and β_2 on four instances of the stochastic utility problem for $K=1000$ outer iterations (i.e., cycles) and $n=500$, $n=1000$, $n=2000$, and $n=5000$. For this experiment, the values of β_1, β_2, the corresponding values of stepsize λ, and the optimal values computed by SCPB1 and SCPB2 are reported in Table 1. We found out that $\beta_1=10$ slightly outperforms other choices of β_1 for SCPB1. Surprisingly, SCPB2 was not affected by changes in β_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 [22] were calibrated similarly, on the basis of an instance of the stochastic utility problem).

β_1, β_2	0.01	0.1	1	10	50	150	1000
$\lambda, n=500$	1.70×10^{-5}	1.70×10^{-4}	0.0017	0.017	0.09	0.26	1.7
$\texttt{Obj}_1$, $n=500$	14.2795	10.6439	10.1819	10.1811	10.1811	10.1838	10.1937
$\texttt{Obj}_2$, $n=500$	10.1937	10.1937	10.1937	10.1937	10.1937	10.1937	10.1937
$\lambda, n=10^3$	1.17×10^{-5}	1.17×10^{-4}	0.0012	0.012	0.0585	0.18	1.17
$\texttt{Obj}_1$, $n=10^3$	14.6307	11.1325	10.0510	10.0504	10.0509	10.0523	10.0710
$\texttt{Obj}_2$, $n=10^3$	10.0710	10.0710	10.0710	10.0710	10.0710	10.0710	10.0710
$\lambda, n=2\times10^3$	8.36×10^{-6}	8.36×10^{-5}	8.36×10^{-4}	0.0084	0.0418	0.1255	0.8364
$\texttt{Obj}_1$, $n=2\times10^3$	13.7451	11.0836	10.0365	10.0363	10.0364	10.0375	10.0613
$\texttt{Obj}_2$, $n=2\times10^3$	10.0613	10.0613	10.0613	10.0613	10.0613	10.0613	10.0613
$\lambda, n=5\times10^3$	7.93×10^{-6}	7.93×10^{-5}	7.93×10^{-4}	0.0079	0.0397	0.119	0.793
$\texttt{Obj}_1$, $n=5\times10^3$	14.0830	11.3370	10.0228	10.0228	10.0231	10.0237	10.0540
$\texttt{Obj}_2$, $n=5\times10^3$	10.0540	10.0540	10.0540	10.0540	10.0540	10.0540	10.0540

Table 1: Selecting parameters β_1 and β_2 of SCPB1 and SCPB2. Framework: SCPB, $K=1000$ outer iterations, four instances of the stochastic utility problem with $n=500$, 1000, 2000, and 5000. $\texttt{Obj}_1$ (resp., $\texttt{Obj}_2$) is the approximate optimal value with SCPB1 (resp., SCPB2).

We now run E-SA, SCPB1, and SCPB2 on three instances L_1, L_2, and L_3 of the stochastic utility problem with $n=2000$, 5000, and 10^5 respectively. For SCPB1 and SCPB2, we used

[2] Although the same problem class and a similar procedure to build ϕ was used in the experiments of Section 4.2 in [22], 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 [22]. Also, contrary to [22], we use the same function ϕ for all instances. The instances differ for the problem dimension n.

$K = 1000$ 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 7 which reports for L_1 and L_2 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 table confirms the slower convergence of E-SA in these instances.

-		$L_1 : n = 2000$		$L_2 : n = 5000$		$L_3 : n = 10^5$	
ALG.	N	Obj	CPU	Obj	CPU	Obj	CPU
E-SA	10	14.6449	0.001	14.6892	0.05	15.4	0.05
	50	14.6322	0.006	14.6813	0.07	14.7	0.35
	100	14.6169	0.01	14.6725	0.1	14.6	0.74
	200	14.5880	0.03	14.6574	0.2	14.6	1.44
	1000	14.3992	0.1	14.5604	0.5	14.3	17.2
	10^4	12.9656	1.28	12.7410	3.7	14.2	80.3
	10^5	-	-	-	-	13.2	860.1
SCPB1	10	13.9539	0.003	13.7763	0.008	14.7	0.08
	50	13.6527	0.01	13.4672	0.02	14.4	0.39
	100	13.5986	0.02	13.5346	0.05	14.2	0.9
	200	13.5349	0.03	13.4686	0.08	14.3	1.6
	1000	13.0370	0.2	12.8376	1.6	14.2	12.5
	10^4	-	-	-	-	12.7	72.3
SCPB2	10	13.5968	0.002	13.7777	0.01	14.2	0.06
	50	12.9421	0.008	12.6959	0.05	13.2	0.7
	100	12.1317	0.02	11.7614	0.09	12.2	1.5
	200	11.3640	0.03	11.3698	0.2	11.6	3.4
	1000	11.5681	0.1	11.5572	0.9	11.2	25.4

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

$$\left\{ \begin{array}{l} \min \ c^T x_1 + \mathbb{E}[\mathfrak{Q}(x_1, \xi)] \\ x_1 \in \mathbb{R}^n : x_1 \geq 0, \sum_{i=1}^n x_1(i) = 1 \end{array} \right. \tag{69}$$

where the second stage recourse function is given by

$$\mathfrak{Q}(x_1, \xi) = \left\{ \begin{array}{l} \displaystyle\min_{x_2 \in \mathbb{R}^n} \ \frac{1}{2} \left(\begin{array}{c} x_1 \\ x_2 \end{array} \right)^T \Big(\xi \xi^T + \gamma_0 I_{2n} \Big) \left(\begin{array}{c} x_1 \\ x_2 \end{array} \right) + \xi^T \left(\begin{array}{c} x_1 \\ x_2 \end{array} \right) \\ s.t. \quad x_2 \geq 0, \displaystyle\sum_{i=1}^n x_2(i) = 1. \end{array} \right. \tag{70}$$

Problem (69)-(70) is of form (1)-(2) where $F(x, \xi) = c^T x + \mathfrak{Q}(x, \xi)$ with $\mathfrak{Q}$ given by (70) and where h is the indicator function of set X where X given by (68) is the unit simplex. For problem (69) we refer to Lemma 2.1 in [8] 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 ξ. We consider two instances of problem (69) with $n = 50$ and $n = 100$. For each instance, the components of ξ 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, SCPB1, and SCPB2 on our two instances A_1 and A_2 with $n = 50$ and $n = 100$, respectively. For SCPB1 and SCPB2, 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 7 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
	2×10^4	8.5482	377	5.1339	1282
	10^5	5.7358	1555.6	3.9193	6147
SCPB1	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
SCPB2	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 (69)-(70)

Table 4 reports the impact of overestimating M (taking M 10 times the Monte Carlo estimation $\overline{M}_e$) and underestimating M (taking M 10 times smaller than the Monte Carlo estimation $\overline{M}_e$). In this experiment, SCPB is essentially not affected by a bad estimation of M while E-SA converge much slower when M is overestimated. Additionally, Table 5 reports the computational results for all methods applied to a variant of two-stage stochastic program (69)-(70) of size $n = 50$ where the feasible set of the first stage problem is replaced by the larger simplex set $\{x_1 \in \mathbb{R}^n : x_1 \geq 0, \sum_{i=1}^{n} x_1(i) = 100\}$. These results show that the SCPB variants are more efficient that E-SA on this specific instance. Also SCPB is not much affected by an overestimation of the diameter D_X.

-		$M = \overline{M}_e$		$M = 10\overline{M}_e$		$M = 0.1\overline{M}_e$	
ALG.	N	`Obj`	`CPU`	`Obj`	`CPU`	`Obj`	`CPU`
E-SA	2000	9.8	29.7	22.1	32.5	10.5	36.2
SCPB1	2000	5.1	36.3	5.2	33.8	5.1	42.2
SCPB2	2000	5.5	31.2	4.6	33.5	5.6	37.2

Table 4: E-SA versus two variants of SCPB on the two-stage stochastic program (69)-(70) with overestimated and underestimated values of M on an instance with $n = 50$.

-		$D = \overline{D}$		$D = 5\overline{D}$	
ALG.	N	`Obj`	`CPU`	`Obj`	`CPU`
E-SA	2000	1.0338×10^6	33.6	1.0365×10^6	33.8
	10000	8.894×10^5	155.5	1.0055×10^6	160.8
SCPB1	2000	2.894×10^5	35.4	3.029×10^5	35.6
SCPB2	2000	2.889×10^5	34.2	3.003×10^5	30.8

Table 5: E-SA versus SCPB1 and SCPB2 on a variant of the two-stage stochastic program (69)-(70) of size $n = 50$ where the simplex feasible set of the first stage problem is replaced by the larger feasible set $\{x_1 \in \mathbb{R}^n : x_1 \geq 0, \sum_{i=1}^n x_1(i) = 100\}$. Exact value $D = D_X = \overline{D}$ of the diameter used to solve the first instance and overestimated value $D = 5\overline{D} = 5D_X$ used to solve the second instance.

Finally, we 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 SCPB1 and SCPB2:

- We observe that the length of the cycles is much larger with SCPB1.
- For SCPB1, 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 λ, R, τ, 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 SCPB1 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 SCPB2, the length of the cycles is in general small with small variability, with an average cycle length of 2.2 for the instance with $n = 50$ and an average cycle length of 2.1 for the instance with $n = 100$.

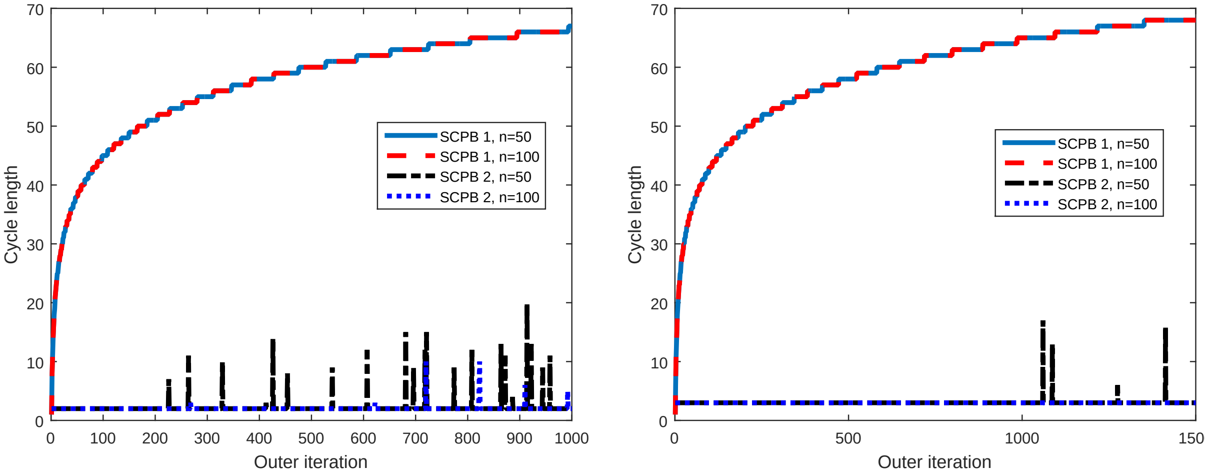

Figure 1: Cycle length for SCPB1 and SCPB2 applied to two-stage stochastic program (69)-(70) (left figure) and two-stage stochastic program (71)-(72) (right figure).

6.3 A second two-stage stochastic program

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

$$\left\{ \begin{array}{l} \min \; c^T x_1 + \mathbb{E}[\mathfrak{Q}(x_1,\xi)] \\ x_1 \in \mathbb{R}^n : \|x_1 - x_0\|_2 \le 100 \end{array} \right. \tag{71}$$

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

$$\mathfrak{Q}(x_1,\xi) = \left\{ \begin{array}{l} \min\limits_{x_2\in\mathbb{R}^n} \; \frac{1}{2}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^T \Big(\xi\xi^T + \gamma_0 I_{2n}\Big)\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \xi^T \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ s.t. \quad \|x_2 - y_0\|_2^2 + \|x_1 - x_0\|_2^2 - R^2 \le 0. \end{array} \right. \tag{72}$$

Problem (71)-(72) is of form (1)-(2) where $F(x,\xi) = c^T x + \mathfrak{Q}(x,\xi)$ with $\mathfrak{Q}$ given by (72) and where h is the indicator function of set

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

For problem (71), we again refer to Lemma 2.1 in [8] for the computation of stochastic subgradients $s(x,\xi)$. We take $\gamma_0 = 2$ and consider for ξ a Gaussian random vector in $\mathbb{R}^{2n}$ with the components of ξ 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 $i = 1,\ldots,n$.

We run E-SA, SCPB1, and SCPB2 on two instances B_1 and B_2 with $n = 50$ and $n = 100$, respectively. For SCPB1 and SCPB2, we used $K = 1500$ outer iterations. The results are reported in Tables 6 and 7. 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 (69)-(70). For SCPB2 the length of the cycles is still small on all iterations and almost constant.

-		$B_1 : n = 50$		$B_2 : n = 100$	
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
SCPB1	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
SCPB2	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 6: E-SA versus two variants of SCPB on the two-stage stochastic program (71), (72)

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

$$100\frac{\texttt{Obj}(\text{E-SA}) - \texttt{Obj}(\text{SCPBi})}{\hat{F}_T(x_0) - \texttt{Obj}(\text{SCPBi})} \tag{73}$$

associated with SCPBi, where $\hat{F}_T(x_0)$ is the empirical mean (see (67)) of F with $T = 10^4$ and x equal to the initial point x_0, and $\texttt{Obj}$(E-SA), $\texttt{Obj}$(SCPB1), and $\texttt{Obj}$(SCPB2) are the empirical means of F with $T = 10^4$ and x equal to the final iterates output by E-SA, SCPB1, and SCPB2, 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.

Sample size N	10	50	100	200	1000
L_1, SCPB1	95.0	95.2	94.1	91.9	82.8
L_1, SCPB2	96.6	97.2	97.5	97.2	90.9
L_2, SCPB1	92.1	93.3	92.3	91.5	77.7
L_2, SCPB2	92.1	95.8	96.8	96.7	93.5
A_1, SCPB1	99.5	98.8	98.1	96.6	85.1
A_1, SCPB2	99.6	99.0	98.0	96.5	84.2
A_2, SCPB1	99.8	99.6	99.3	98.8	95.0
A_2, SCPB2	99.8	99.6	99.3	98.8	95.4
B_1, SCPB1	99.8	99.4	98.8	97.6	88.7
B_1, SCPB2	99.9	99.4	98.8	97.6	88.7
B_2, SCPB1	99.9	99.4	99.0	98.0	90.5
B_2, SCPB2	99.9	99.5	99.0	98.0	90.5

Table 7: Percentages (73) for SCPB1 and SCPB2

7 Concluding remarks

This paper proposes two single-cut stochastic composite proximal bundle variants, called SCPB, for solving SCCO problem (1)-(2) where at each iteration a problem of form (3) is solved. The two SCPB variants, 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 properly chosen parameters have optimal iteration complexity (up to a logarithmic term) for finding an ε-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 [22] 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 SCPB1 which performs only one iteration per cycle. Second, it is worth noting that SCPB has a slight similarity with the stochastic dual averaging (SDA) method discussed in [25, 38] 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, 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 between RSA and SDA.

Extensions. We finally discuss some possible extensions of our analysis in this paper. A first question is how to extend SCPB and the corresponding complexity analysis if instead of condition (A3) we use the assumption that for every $u, v \in \operatorname{dom} h$, we have

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

which is called a uniform (M, L)-condition in [20]. A second natural question is how to extend SCPB and its complexity analysis when either the prox stepsize λ and parameter θ are allowed to change with the iteration count k. Recalling that the prox stepsize is the only ingredient of the second variant of SCPB that depends on an estimate M of $\bar{M}$, a third natural question is whether it is possible to develop a SCPB variant which adaptively chooses a (variable) prox stepsize without the need of knowing M. A fourth question is whether it is possible to establish global convergence rate guarantees for proximal bundle methods based on two-cut or multiple-cut bundle models instead of the single-cut bundle models considered in this paper. Finally, SCPB is able to solve two-stage convex stochastic programs with continuous distributions, under the assumption that the second-stage subproblems can be exactly solved (e.g., see Subsections 6.2 and 6.3). It would be interesting to extend it to the setting of *multistage* stochastic convex problems with continuous distributions.

References

[1] E. D. Andersen and K.D. Andersen. *The MOSEK optimization toolbox for MATLAB manual. Version 9.2*, 2019. `https://www.mosek.com/documentation/`.

[2] 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.

[3] J. Birge and F. Louveaux. *Introduction to Stochastic Programming.* Springer-Verlag, New York, 1997.

[4] J.R. Birge and F.V. Louveaux. A multicut algorithm for two-stage stochastic linear programs. *European Journal of Operational Research*, 34:384–392, 1988.

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

[6] Y. Du and A. Ruszczyński. Rate of convergence of the bundle method. *Journal of Optimization Theory and Applications*, 173(3):908–922, 2017.

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

[8] V. Guigues. Convergence analysis of sampling-based decomposition methods for risk-averse multistage stochastic convex programs. *SIAM Journal on Optimization*, 26:2468–2494, 2016.

[9] V. Guigues. Multistep stochastic mirror descent for risk-averse convex stochastic programs based on extended polyhedral risk measures. *Mathematical Programming*, 163:169–212, 2017.

[10] V. Guigues. Inexact Stochastic Mirror Descent for two-stage nonlinear stochastic programs. *Mathematical Programming*, 187:533–577, 2021.

[11] V. Guigues, A. Juditsky, and A. Nemirovski. Non-asymptotic confidence bounds for the optimal value of a stochastic program. *Optimization Methods & Software*, 32:1033–1058, 2017.

[12] V. Guigues, W. Tekaya, and M. Lejeune. Regularized decomposition methods for deterministic and stochastic convex optimization and application to portfolio selection with direct transaction and market impact costs. *Optimization & Engineering*, 21:1133–1165, 2020.

[13] J.L. Higle and S. Sen. *Stochastic Decomposition.* Kluwer, Dordrecht, 1996.

[14] A.J. King and R.T. Rockafellar. Asymptotic theory for solutions in statistical estimation and stochastic programming. *Math. Oper. Res.*, 18:148–162, 1993.

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

[16] A. J. Kleywegt, A. Shapiro, and T. Homem-de Mello. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12(2):479–502, 2002.

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

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

[19] 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.

[20] J. Liang and R. D. C. Monteiro. A unified analysis of a class of proximal bundle methods for solving hybrid convex composite optimization problems. *To appear in Mathematics of Operations Research, available on arXiv:2110.01084*, 2021.

[21] 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.

[22] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.*, 19:1574–1609, 2009.

[23] A. Nemirovski and D. Yudin. On Cezari's convergence of the steepest descent method for approximating saddle point of convex-concave functions. *Soviet Math. Dokl.*, 19, 1978.

[24] A. Nemirovski and D.B. Yudin. *Problem complexity and method efficiency in optimization.* Wiley, 1983.

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

[26] 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.

[27] B.T. Polyak. New stochastic approximation type procedures. *Automat. i Telemekh (English translation: Automation and Remote Control)*, 7:98–107, 1990.

[28] B.T. Polyak and A. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM J. Contr. and Optim.*, 30:838–855, 1992.

[29] H. Robbins and S. Monroe. A stochastic approximation method. *Annals of Math. Stat.*, 22:400–407, 1951.

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

[31] A. Shapiro. Asymptotic analysis of stochastic programs. *Ann. Oper. Res.*, 30:169–186, 1991.

[32] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory.* SIAM, Philadelphia, 2009.

[33] R.M. Van Slyke and R.J.-B. Wets. L-shaped linear programs with applications to optimal control and stochastic programming. *SIAM Journal of Applied Mathematics*, 17:638–663, 1969.

[34] 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.

[35] B. Verweij, S. Ahmed, A. J. Kleywegt, G. Nemhauser, and A. Shapiro. The sample average approximation method applied to stochastic routing problems: a computational study. *Computational Optimization and Applications*, 24(2-3):289–333, 2003.

[36] W. Wang and M. A. Carreira-Perpinán. Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application. *Available on arXiv:1309.1541*, 2013.

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

[38] L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11(88):2543–2596, 2010.