A Single Cut Proximal Bundle Method for Stochastic Convex Composite Optimization

Jiaming Liang

Yale University

October 16, 2022

Joint work with Vincent Guigues (FGV) and Renato Monteiro (Georgia Tech)

INFORMS Annual Meeting 2022, Indianapolis, IN

Introduction

Main problem

$$\phi_* := \min \left\{ \phi(x) := f(x) + h(x) : x \in \mathbb{R}^n \right\}, \quad f(x) = \mathbb{E}_\xi [F(x, \xi)]$$

E.g., two-stage convex stochastic program

$$\min\{f_1(x) + \mathbb{E}[Q(x, \xi)] : x \in X\}$$

where $Q(x, \xi) = \min\{f_2(x, y, \xi) : g_2(x, y, \xi) \le 0, y \in Y\}$.

An instance of the main problem with

$$h(x) = \delta_X(x), \quad F(x, \xi) = f_1(x) + Q(x, \xi).$$

Goal: SA-type algorithm based on the proximal bundle (PB) method

Assumptions

Stochastic convex composite optimization

$$\phi_* := \min\left\{\phi(x) := f(x) + h(x) : x \in \mathbb{R}^n\right\}, \quad f(x) = \mathbb{E}_\xi[F(x,\xi)]$$

Black-box model

(A1) f is closed convex and $\operatorname{dom} f \supset \operatorname{dom} h$;

(A2) for almost every $\xi \in \Xi$, there exist 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

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

(A3) for every $x \in \operatorname{dom} h$, we have $\mathbb{E}[\|s(x,\xi)\|^2] \le M^2$;

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

Review of Deterministic PB

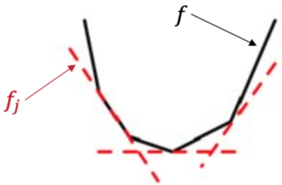
Proximal point method: constructs a sequence of proximal problems.
E.g., Chambolle-Pock for saddle point, ADMM for distributed optimization.

Approximately solve the proximal problem by an iterative process

$$x^+ \leftarrow \min_{u \in \mathbb{R}^n} \left\{ f(u) + \frac{1}{2\lambda} \|u - x^c\|^2 \right\}.$$

Recursively build up a cutting-plane model

$$f_j(u) = \max\{f(x) + \langle f'(x), u - x \rangle : x \in C_j\}, \quad C_{j+1} = C_j \cup \{x_j\}$$



Review of Deterministic PB

Algorithm 1 PB (one cycle)

1. Construct a proximal problem

$$\min_{u \in \mathbb{R}^n} \left\{ f(u) + h(u) + \frac{1}{2\lambda} \|u - x^c\|^2 \right\};$$

2. **If** find an $(\varepsilon/2)$-solution to the current proximal problem, **then** change the prox-center; $\leftarrow$ serious

Otherwise, keep the prox-center, update the cutting-plane model and solve the prox subproblem based on the current model, i.e., $\leftarrow$ null

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

Review of Deterministic PB

- Proximal bundle method $\mathcal{O}(\varepsilon^{-3})$ [1] $\to \mathcal{O}(\varepsilon^{-2})$ [2]
- Lower complexity bound $\Omega(\varepsilon^{-2})$

Proximal bundle method is optimal for black-box model.

[1] Kiwiel, 2000. Efficiency of proximal bundle methods.

[2] Liang and Monteiro, 2020. A proximal bundle variant with optimal iteration-complexity for a large range of prox stepsizes.

Other bundle models

(E1) **one-cut update[3]:** $\Gamma^+ = \Gamma^+_\tau := \tau\Gamma + (1-\tau)[\ell_f(\cdot;x) + h]$ with $\bar{\Gamma} = \Gamma$.

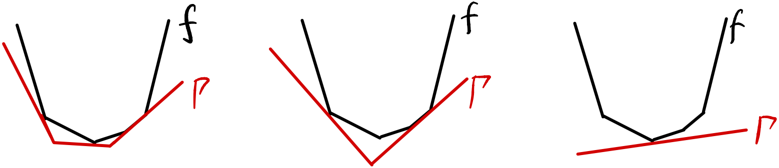
(E2) **two-cuts update:** assume $\Gamma = \max\{A_f, \ell_f(\cdot;x^-)\} + h$ where A_f is an affine function satisfying $A_f \le f$, set

$$\Gamma^+ = \max\{A_f^+, \ell_f(\cdot;x)\} + h$$

where $A_f^+ = \theta A_f + (1-\theta)\ell_f(\cdot;x^-)$. Also set $\bar{\Gamma} = A_f^+ + h$.

Bundle of past information $\{(x, f(x), f'(x)) : x \in C\}$

$f \geq \Gamma$



Multiple cuts Two cuts One cut

[3]Liang and Monteiro, 2021. A unified analysis of a class of proximal bundle methods for solving hybrid convex composite optimization problems.

Cutting-plane Model in the Stochastic Setting

A straightforward fact:

$$\mathbb{E}[\max\{X, Y\}] \geq \max\{\mathbb{E}[X], \mathbb{E}[Y]\}.$$

For a fixed u,

$$\mathbb{E}[\Gamma_j(u)] \geq \max\{\mathbb{E}[F(x,\xi) + \langle s(x,\xi), u - x\rangle] : x \in C_j\}.$$

On the other hand,

$$\begin{aligned}&\max\{\mathbb{E}[F(x,\xi) + \langle s(x,\xi), u - x\rangle] : x \in C_j\}\\ =&\max\{f(x) + \langle f'(x), u - x\rangle : x \in C_j\} \leq f(u)\end{aligned}$$

So

$$\mathbb{E}[\Gamma_j(u)] \ ? \ f(u)$$

A Single Cut Model

Aggregate all cuts into a single one

$$\Gamma_j(u) = \tau\Gamma_{j-1}(u) + (1-\tau)[F(x_{j-1}, \xi_{j-1}) + \langle s(x_{j-1}, \xi_{j-1}), u - x_{j-1}\rangle].$$

Since

$$\mathbb{E}[F(x,\xi) + \langle s(x,\xi), u - x\rangle] = f(x) + \langle f'(x), u - x\rangle \leq f(u),$$

we have by induction

$$\mathbb{E}[\Gamma_j(u)] \leq f(u).$$

1. Let $\lambda, \theta > 0$, integer $K \geq 1$, and $x_0 \in \operatorname{dom} h$ be given, and set $x_0^c = x_0$, $j = k = 1$, $j_0 = 0$, and
$$\tau = \frac{\theta K}{\theta K + 1};$$
2. Take an independent sample ξ_{j-1} of r.v. ξ, set
$$x_j^c = \begin{cases} x_{j_{k-1}}, & \text{if } j = j_{k-1} + 1, \\ x_{j-1}^c, & \text{otherwise,} \end{cases}$$
and compute
$$x_j = \underset{u \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ h(u) + \langle S_j, u \rangle + \frac{1}{2\lambda} \|u - x_j^c\|^2 \right\},$$
where
$$S_j := \begin{cases} s(x_{j_{k-1}}, \xi_{j_{k-1}}), & \text{if } j = j_{k-1} + 1, \\ (1-\tau) s(x_{j-1}, \xi_{j-1}) + \tau S_{j-1}, & \text{otherwise,} \end{cases}$$

Stochastic Composite Proximal Bundle Framework

2. Compute

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

3. Choose an integer $j_k \ge j_{k-1}+1$, and set $\hat{y}_k = y_{j_k}$ when the k-th cycle ends;
4. if $k = K$ then **stop** and output

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

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

Remarks on SCPB

- An aggregated single cut
- No termination criterion for a cycle

Define a cycle

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

Two ways of setting j_k:

(B1) the smallest integer $j_k \geq i_k$ and $\lambda k \tau^{j_k - i_k} \leq C$;

(B2) the smallest integer $j_k \geq i_k + 1$ and $\lambda k \tau^{j_k - i_k} t_{i_k} \leq C$.

(B1) is deterministic and (B2) is stochastic

Main Results – SCPB1

Assume that conditions (A1)-(A4) hold and $\mathrm{dom}\, h$ has a finite diameter $D_h \geq 0$.

SCPB with (B1) satisfies the following statements:

- Number of iterations within $\mathcal{C}_k$, or number of null steps

$$|\mathcal{C}_k| \leq \left\lceil (\theta K + 1) \ln \left(\frac{\lambda k}{C} + 1 \right) \right\rceil + 1.$$

- Convergence of SCPB1

$$\mathbb{E}[\phi(\hat{y}_K^a)] - \phi_* \leq \frac{1}{K} \left(\frac{D^2}{\lambda} + \frac{6C \min\{\lambda M^2, MD\}}{\lambda} + \frac{2\lambda M^2}{\theta} \right).$$

A Practical Variant of SCPB1

Let pair (λ, K) and constant $m \geq 1$ be given, and define

$$\theta = \frac{m}{K}, \quad C = \frac{D}{6M},$$

SCPB based on (B1) satisfies the following statements:

- Number of iterations within $\mathcal{C}_k$, or number of null steps

$$|\mathcal{C}_k| \leq \left\lceil (m+1) \ln\left(\frac{\lambda k}{C} + 1\right) \right\rceil + 1.$$

- Convergence of SCPB1

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

- Its expected overall iteration complexity is $\tilde{\mathcal{O}}(mK)$.

Robust Stochastic Approximation (RSA) [4]

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

- Convergence of RSA

$$\mathbb{E}[\phi(x_K^a)] - \phi_* \le \frac{2D^2}{\lambda K} + 2\lambda M^2, \quad x_N^a = \frac{1}{\lceil N/2 \rceil} \sum_{j=\lfloor N/2 \rfloor + 1}^{N} x_j.$$

Taking $\lambda = \frac{\sqrt{m} D}{M \sqrt{K}}$, given $\varepsilon > 0$, to obtain $x \in \operatorname{dom} h$ such that $\mathbb{E}[\phi(x)] - \phi_* \le \varepsilon$,

- RSA has iteration complexity $\mathcal{O}\left(\frac{mM^2D^2}{\varepsilon^2}\right)$;
- SCPB1 has iteration complexity $\tilde{\mathcal{O}}\left(\frac{M^2D^2}{\varepsilon^2}\right)$.

[4] Nemirovski, Juditsky, Lan and Shapiro, 2009. Robust stochastic approximation approach to stochastic programming.

Relationship between SCPB1 and RSA

Recall (B1) the smallest integer $j_k \geq i_k$ and $\lambda k \tau^{j_k - i_k} \leq C$.
Choosing

$$C = \frac{\alpha D \sqrt{K}}{M}, \quad \lambda = \frac{\alpha D}{M\sqrt{K}},$$

then (B1) is satisfied with $j_k = i_k$, since

$$\frac{C}{\lambda k} \geq \frac{C}{\lambda K} = 1 = \tau^{j_k - i_k}.$$

In summary,

- RSA performs one iteration per cycle
- RSA $\rightarrow$ SCPB1 is analogous to Subgradient method $\rightarrow$ PB
- RSA is restricted to small stepsizes, while SCPB1 can use large ones
- SCPB1 implicitly reduces the variance and the sample complexity by m

Main Results – SCPB2

Recall (B2) the smallest integer $j_k \geq i_k + 1$ and $\lambda k \tau^{j_k - i_k} t_{i_k} \leq C$.

Assume that conditions (A1)-(A4) hold and $\operatorname{dom} h$ has a finite diameter $D_h \geq 0$.

SCPB with (B2) satisfies the following statements:

- Number of iterations within $\mathcal{C}_k$, or number of null steps

$$|\mathcal{C}_k| \leq \left\lceil (\theta K + 1) \ln \left(\frac{2M^2 \lambda^2 k}{C} + 1 \right) \right\rceil + 1.$$

- Convergence of SCPB2

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

A Practical Variant of SCPB2

Let pair (λ, K) and constant $m \geq 1$ be given, and define

$$\theta = \frac{m}{K}, \quad C = \frac{D^2}{3},$$

SCPB based on (B2) satisfies the following statements:

- Number of iterations within $\mathcal{C}_k$, or number of null steps

$$|\mathcal{C}_k| \leq \left\lceil (m+1) \ln \left(\frac{6M^2\lambda^2 k}{D^2} + 1 \right) \right\rceil + 1.$$

- Convergence of SCPB2

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

- Its expected overall iteration complexity is $\tilde{\mathcal{O}}(mK)$.

Test 1 – Two-stage Stochastic Program

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

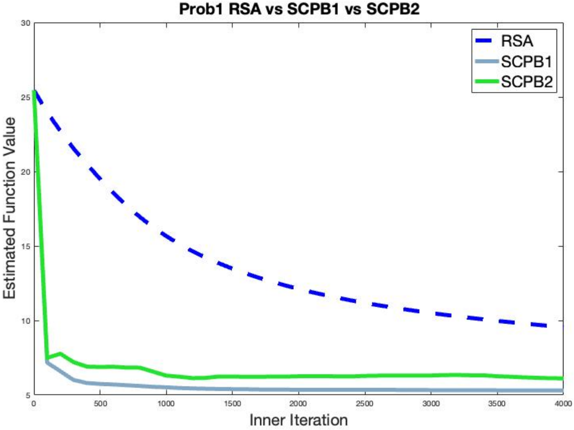
where the second stage recourse function is given by

$$Q(x_1,\xi) = \begin{cases} \min\limits_{x_2 \in \mathbb{R}^n} \; \frac{1}{2} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^T \left(\xi\xi^T + \lambda_0 I_{2n}\right) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \xi^T \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ x_2 \geq 0, \sum\limits_{i=1}^n x_2(i) = 1. \end{cases}$$

Table: $n = 50, \; N = 4000$

Statistics	RSA	SCPB1	SCPB2
λ	7.4×10^{-7}	10^{-3}	10^{-3}
Min Inner	1	9	2
Max Inner	1	52	43
Avg Inner	1	43	5

Test 1 – Two-stage Stochastic Program



Test 2 – Two-stage Stochastic Program

$$\begin{cases} \min \ c^T x_1 + \mathbb{E}[\mathfrak{Q}(x_1, \xi)] \\ x_1 \in \mathbb{R}^n : \|x_1 - x_0\|_2 \leq 1 \end{cases}$$

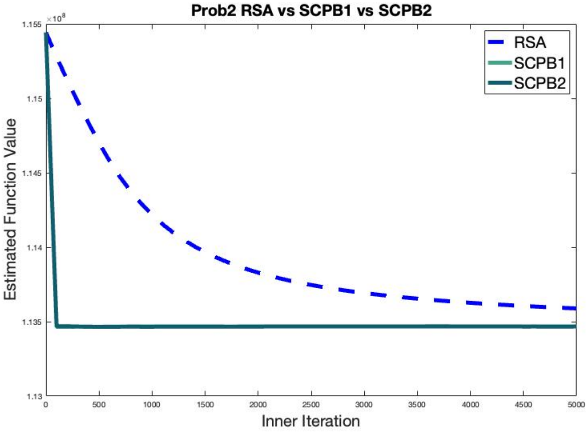
where the second stage recourse function is given by

$$Q(x_1, \xi) = \begin{cases} \min\limits_{x_2 \in \mathbb{R}^n} \ \frac{1}{2} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^T \left(\xi\xi^T + \lambda_0 I_{2n}\right) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \xi^T \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ \|x_2 - y_0\|_2^2 + \|x_1 - x_0\|_2^2 - R^2 \leq 0. \end{cases}$$

Table: $n = 50$, $N = 5000$

Statistics	RSA	SCPB1	SCPB2
λ	8.9×10^{-10}	10^{-3}	10^{-3}
Min Inner	1	71	54
Max Inner	1	109	89
Avg Inner	1	100	77

Test 2 – Two-stage Stochastic Program



Test 3 – One-stage Stochastic Program

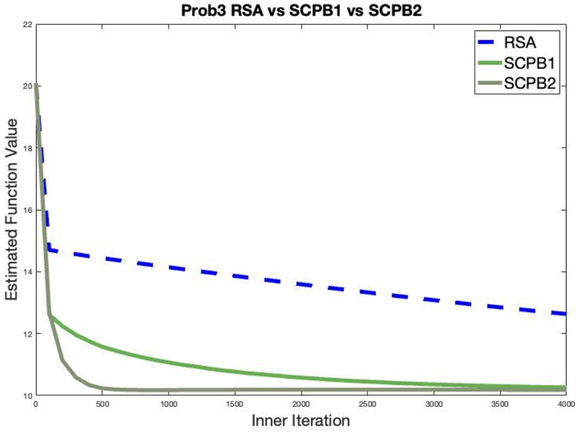
$$\min_{x\in X} \mathbb{E}\left[\phi\left(\sum_{i=1}^{n}(\frac{i}{n}+\xi_i)x_i\right)\right]$$

where X is the unit simplex.

Table: $n = 100,\ N = 4000$

Statistics	RSA	SCPB1	SCPB2
λ	2.8×10^{-5}	10^{-3}	10^{-3}
Min Inner	1	1	2
Max Inner	1	26	6
Avg Inner	1	17	2

Test 3 – One-stage Stochastic Program



Take-away

- A parameter-free single cut proximal bundle method for stochastic programming
- Aggregating all past information by convex combination
- $\mathcal{O}(1/K)$ convergence rate
- Includes RSA as an instance, outperforms RSA in theory and practice
- Variance reduction

References

- J. Liang, V. Guigues and R. D. C. Monteiro. A single cut proximal bundle method for stochastic convex composite optimization. Available on arXiv:2207.09024, 2022.
- J. Liang and R. D. C. Monteiro. A unified analysis of a class of proximal bundle methods for smooth-nonsmooth convex composite optimization. Available on arXiv:2110.01084, 2021.
- 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.

Thank you!

Supplementary Materials

A generic bundle update scheme

Definition

Let $\mathcal{C}_\mu(\phi)$ denote a class of convex functions Γ satisfying $\Gamma \le \phi$ and Γ is μ-convex.

For a given quadruple $(\Gamma, x_0, \lambda, \tau) \in \mathcal{C}_\mu(\phi) \times \mathbb{R}^n \times \mathbb{R}_{++} \times (0,1)$, the generic bundle update scheme returns $\Gamma^+ \in \mathcal{C}_\mu(\phi)$ satisfying

$$\tau\bar{\Gamma} + (1-\tau)[\ell_f(\cdot; x) + h] \le \Gamma^+ \tag{1}$$

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

$$x = \underset{u\in\mathbb{R}^n}{\operatorname{argmin}} \left\{ \Gamma(u) + \frac{1}{2\lambda}\|u - x_0\|^2 \right\}$$

and $\bar{\Gamma} \in \mathcal{C}_\mu(\phi)$ is such that

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

SCPB vs. other SA methods

Comparison with Dual Averaging (DA) [5]:

- DA uses a fixed prox-center throughout the process
- DA uses variable prox stepsizes

Comparison with Robust Stochastic Approximation (RSA) [6]:

- RSA does not use previous cuts
- RSA performs one iteration per cycle
- RSA $\rightarrow$ SCPB is analogous to Subgradient method $\rightarrow$ PB

[5] Nesterov, 2009. Primal-dual subgradient methods for convex problems.

[6] Nemirovski, Juditsky, Lan and Shapiro, 2009. Robust stochastic approximation approach to stochastic programming.