A Single Cut Proximal Bundle Method for Stochastic Convex Composite Optimization

Jiaming Liang

Yale University

July 27, 2022

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

International Conference on Continuous Optimization 2022

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) h is closed convex and is M_h-Lipschitz continuous, i.e.,

$$|h(x) - h(y)| \le M_h\|x - y\|;$$

(A2) f is closed convex and $\mathrm{dom}\, f \supset \mathrm{dom}\, h$;

(A3) for almost every $\xi \in \Xi$, there exist a functional oracle $F(\cdot,\xi) : \mathrm{dom}\, h \to \mathbb{R}$ and a stochastic subgradient oracle $s(\cdot,\xi) : \mathrm{dom}\, h \to \mathbb{R}^n$ satisfying

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

(A4) for every $x \in \mathrm{dom}\, h$, we have $\mathbb{E}[\|s(x,\xi)\|^2] \le M^2$;

(A5) $\mathrm{dom}\, h$ has a finite diameter $D > 0$;

(A6) the set of optimal solutions X^* is nonempty.

Review of Deterministic PB

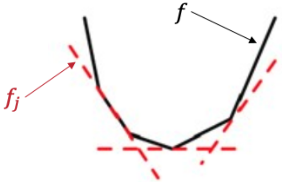
Proximal point method: constructs a sequence of proximal problems.
E.g., Chambolle-Pock for saddle point, ADMM for distributed optimization.

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] $\rightarrow$ $\mathcal{O}(\varepsilon^{-2})$ [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 $\bar{\Gamma} = \Gamma$.

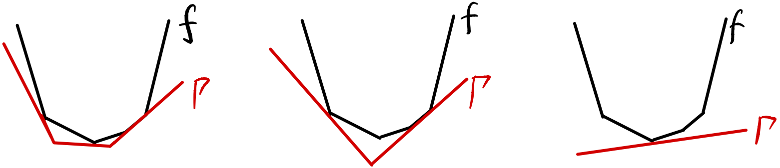
(E2) **two-cuts update:** assume $\Gamma = \max\{A_f, \ell_f(\cdot;x^-)\} + h$ where A_f is an affine function satisfying $A_f \le f$, set

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

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

$f \geq \Gamma$

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



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

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

Stochastic Composite Proximal Bundle Framework

1. Let $\lambda, \theta > 0$, integer $K > 0$, and $x_0 \in \operatorname{dom} h$ be given, and set $x_0^c = x_0$, $j = k = 1$, $j_0 = 0$, and
$$\tau = \frac{\theta K}{\theta K + 1};$$
moreover, take a sample ξ_0 of r.v. ξ;
2. 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
$$x_j = \underset{u \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ \Gamma_j^\lambda(u) := \Gamma_j(u) + \frac{1}{2\lambda}\|u - x_j^c\|^2 \right\},$$
where
$$\Gamma_j(\cdot) := \begin{cases} \ell_\Phi(\cdot; x_{j_{k-1}}, \xi_{j_{k-1}}), & \text{if } j = j_{k-1} + 1, \\ (1-\tau)\ell_\Phi(\cdot; x_{j-1}, \xi_{j-1}) + \tau \Gamma_{j-1}(\cdot), & \text{otherwise,} \end{cases}$$
$$\ell_\Phi(\cdot; x, \xi) := F(x, \xi) + \langle s(x, \xi), \cdot - x \rangle + h(\cdot)$$

Stochastic Composite Proximal Bundle Framework

2. Compute

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

take a sample ξ_j of r.v. ξ and set

$$u_j = (1-\tau)\Phi(x_j,\xi_j) + \begin{cases} \tau\Phi(x_{j-1},\xi_{j-1}), & \text{if } j = j_{k-1}+1, \\ \tau u_{j-1}, & \text{otherwise;} \end{cases}$$

3. Choose an integer $j_k \ge j_{k-1}+2$, and set $\hat{u}_k = u_{j_k}$ and $\hat{y}_k = y_{j_k}$ when the k-th cycle ends;
4. if $k = K$ then **stop** and output

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

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

SCPB vs. other SA methods

Comparison with Dual Averaging (DA) [4]:

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

Comparison with Robust Stochastic Approximation (RSA) [5]:

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

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

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

Remarks on SCPB

- An aggregated single cut
- Estimate function value $\hat{u}^a_K$
- 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 $j_k \geq j_{k-1} + 2$ and $\lambda k \tau^{j_k - i_k} \leq C$;

(B2) the smallest integer $j_k \geq j_{k-1} + 2$ and $\lambda k \tau^{j_k - i_k}[u_{i_k} - \Gamma^{\lambda}_{i_k}(x_{i_k})] \leq C$.

(B1) is deterministic and (B2) is stochastic

Main Results – SCPB 1

SCPB with (B1) satisfies the following statements:

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

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

- Convergence within $\mathcal{C}_k$

$$\mathbb{E}[\text{opt. gap for the prox. problem}] \le \frac{C(2M + M_h)D}{\lambda k} + \frac{2\lambda M^2}{\theta K}.$$

- Convergence of SCPB 1

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

Main Results – SCPB 2

(A7) For every $x \in \operatorname{dom} h$, we have $\mathbb{E}[|F(x,\xi) - f(x)|] \le \sigma$.

SCPB with (B2) satisfies the following statements:

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

$$\left\lceil (\theta K + 1) \ln \left(\frac{(2M + M_h) D \lambda k}{C} + 1 \right) \right\rceil + 1.$$

- Convergence within $\mathcal{C}_k$

$$\mathbb{E}[\text{opt. gap for the prox. problem}] \le \frac{C}{\lambda k} + \frac{2\sigma + 2\lambda M^2}{\theta K} + \frac{2\lambda M^2}{\theta^2 K^2}.$$

- Convergence of SCPB 2

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

Iteration Complexity – SCPB 1

Let tolerance $\varepsilon > 0$ be given and set the input K of SCPB as

$$K = K_\varepsilon := \left\lfloor \frac{3C(2M + M_h)D + D^2}{\lambda\varepsilon} + \frac{2\lambda M^2}{\theta\varepsilon} \right\rfloor + 1.$$

Then, we have

$$\mathbb{E}[\phi(\hat{y}_K^a)] - \phi_* \le \mathbb{E}[\hat{u}_K^a] - \phi_* \le \varepsilon,$$

and the expected overall iteration complexity of SCPB is

$$\mathcal{O}\left(K_\varepsilon \left[(1 + \theta K_\varepsilon)\log\left(\frac{\lambda K_\varepsilon}{C} + 1\right) + 1\right]\right).$$

Iteration Complexity – SCPB 1

Moreover, if we choose

$$C = \frac{D}{M + M_h}, \quad \theta = \frac{\lambda^2 M^2}{D^2}, \quad K = K_\varepsilon,$$

then the iteration complexity becomes

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

Iteration Complexity – SCPB 2

Let tolerance $\varepsilon > 0$ be given and set the input K of SCPB as

$$K = K_\varepsilon := \left\lfloor \frac{3C + D^2}{\lambda\varepsilon} + \frac{2\lambda M^2 + 2\sigma}{\theta\varepsilon} + \frac{\sqrt{2\lambda}M}{\theta\sqrt{\varepsilon}} \right\rfloor + 1.$$

Then, we have

$$\mathbb{E}[\phi(\hat{y}_K^a)] - \phi_* \leq \mathbb{E}[\hat{u}_K^a] - \phi_* \leq \varepsilon,$$

and the expected overall iteration complexity of SCPB is

$$\mathcal{O}\left(K_\varepsilon \left[(1 + \theta K_\varepsilon) \log\left(\frac{\lambda(M + M_h)DK_\varepsilon}{C} + 1 \right) + 1 \right] \right).$$

Iteration Complexity – SCPB 2

Moreover, if we choose

$$C = D^2, \quad \theta = \frac{\lambda^2 M^2 + \lambda\sigma}{D^2}, \quad K = K_\varepsilon,$$

then the iteration complexity becomes

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

Remarks on Complexity Results

- Assuming sub-Gaussian distribution, i.e.,

$$\mathbb{E}[\exp(\|s(x,\xi)\|^2/M^2)] \le \exp(1),$$

  then we can establish large deviation results.

- Both expected optimality gap and large deviation results are similar to what are for RSA.

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

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

Two-stage Stochastic Program

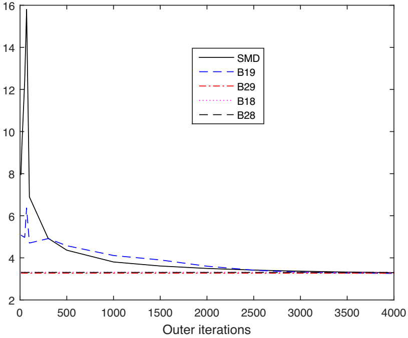
Output	SCPB 1 $\tau = 0.8$	SCPB 1 $\tau = 0.9$	SCPB 2 $\tau = 0.8$	SCPB 2 $\tau = 0.9$	SMD
Outer	10	3000	10	10	4000
Inner	20	17325	20	20	-
Time (s)	0.68	569.5	0.78	0.82	48.5
Obj.	3.32	3.32	3.31	3.27	3.30

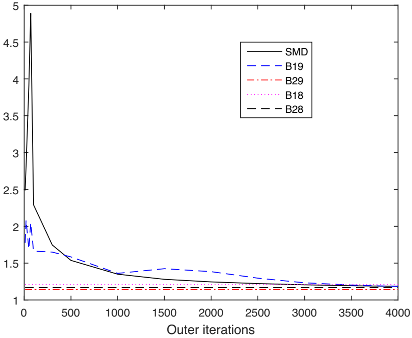
Table: $n = 50$, SMD and four variants of SCPB.

Output	SCPB 1 $\tau = 0.8$	SCPB 1 $\tau = 0.9$	SCPB 2 $\tau = 0.8$	SCPB 2 $\tau = 0.9$	SMD
Outer	10	3000	10	10	4000
Inner	20	17325	20	20	-
Time (s)	3.03	2550	3.57	3.56	234.8
Obj.	1.21	1.23	1.17	1.14	1.18

Table: $n = 100$, SMD and four variants of SCPB.

Two-stage Stochastic Program





Take-away

- A parameter-free single cut proximal bundle method for stochastic programming
- Aggregating all past information by convex combination
- Prescribe $K =$ number of serious steps, no termination criterion
- Establish $\mathcal{O}(1/K)$ rate for expected optimality gap

Reference

- 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!