A Proximal Bundle Type Method for Smooth and Nonsmooth Convex Optimization and Stochastic Programming

Jiaming Liang

School of Industrial and Systems Engineering
Georgia Institute of Technology

Joint work with Renato Monteiro

INFORMS 2021 - October 24, Anaheim, CA

This talk is based on the following papers:

- J. Liang and R. D. C. Monteiro. A unified analysis of a class of proximal bundle methods for solving hybrid convex composite optimization problems. 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. To appear in SIAM Journal on Optimization, available on arXiv:2003.11457, 2020.

The first **optimal** complexity result for a PB type method.

Introduction

Main problem:

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

Main goal:

To present a framework consisting of most proximal bundle methods for hybrid convex composite optimization.

Hybrid (smooth-nonsmooth) convex composite problem

Consider (1), where

(A1) $f, h \in \overline{\mathrm{Conv}}\,(\mathbb{R}^n)$ are such that $\mathrm{dom}\, h \subset \mathrm{dom}\, f$ and a subgradient oracle $f' : \mathrm{dom}\, h \to \mathbb{R}^n$ satisfying $f'(x) \in \partial f(x)$ for every $x \in \mathrm{dom}\, h$ is available;

(A2) the set of optimal solutions X^* of problem (1) is nonempty;

(A3) $\|f'(u) - f'(v)\| \leq 2M_f + L_f\|u - v\|$ for every $u, v \in \mathrm{dom}\, h$;

(A4) h is μ-convex.

In a previous paper [1], we proposed a relaxed proximal bundle (RPB) method that is optimal for convex (and strongly convex) nonsmooth optimization.

In this work, we generalize and improve RPB in the following aspects:

1. hybrid cases;
2. a general framework including 3 bundle update schemes;
3. a unified and much simpler analysis;
4. stronger complexity results;
5. an adaptive PB method.

[1] J. Liang and R. D. C. Monteiro. A proximal bundle variant with optimal iteration-complexity for a large range of prox stepsizes. To appear in SIAM Journal on Optimization, available on arXiv:2003.11457, 2020.

Proximal bundle method

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

Solving the proximal problem

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

can be as difficult as solving $\min\{\phi(u) : u \in \mathbb{R}^n\}$.

Proximal bundle method approximately solves (2) and recursively builds up a model by using a standard cutting-plane approach.

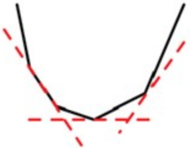
Proximal **bundle** method

The **bundle method** solves a sequence of prox subproblems of the form

$$x_j = \underset{u\in\mathbb{R}^n}{\operatorname{argmin}} \left\{ \Gamma_j^\lambda(u) := f_j(u) + h(u) + \frac{1}{2\lambda}\|u - x_{j-1}^c\|^2 \right\}, \qquad (3)$$

where x_{j-1}^c is the **prox-center**, f_j is the **cutting-plane** model defined as

$$f_j(u) = \max\{f(x) + \langle f'(x), u - x\rangle : x \in C_j\} \quad \forall u \in \mathbb{R}^n.$$



Proximal **bundle** method

The **bundle method** solves a sequence of prox subproblems of the form

$$x_j = \underset{u \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ \Gamma_j^\lambda(u) := f_j(u) + h(u) + \frac{1}{2\lambda} \|u - x_{j-1}^c\|^2 \right\},$$

where x_{j-1}^c is the **prox-center**, f_j is the **cutting-plane** model defined as

$$f_j(u) = \max\{f(x) + \langle f'(x), u - x \rangle : x \in C_j\} \quad \forall u \in \mathbb{R}^n,$$

and decides to perform a **serious** or **null** iteration based on the **descent condition** $\phi\left(x_j\right) \leq (1-\gamma)\phi\left(x_{j-1}^c\right) + \gamma\left(f_j + h\right)\left(x_j\right)$ for some $\gamma \in (0,1)$.

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

$$\tau \bar{\Gamma} + (1-\tau)[\ell_f(\cdot; x) + h] \le \Gamma^+ \tag{4}$$

where

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

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

$$\bar{\Gamma}(x) = \Gamma(x), \quad x = \operatorname*{argmin}_{u \in \mathbb{R}^n} \left\{ \bar{\Gamma}(u) + \frac{1}{2\lambda}\|u - x_0\|^2 \right\}. \tag{5}$$

Examples

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

(E2) **two-cuts scheme:** 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 $A_f^+ = \theta A_f + (1-\theta)\ell_f(\cdot;x^-)$ and

$$\theta \begin{cases} = 1, & \text{if } A_f(x) > \ell_f(x;x^-), \\ = 0, & \text{if } A_f(x) < \ell_f(x;x^-), \\ \in [0,1], & \text{if } A_f(x) = \ell_f(x;x^-). \end{cases}$$

Also set $\bar\Gamma = A_f^+ + h$.

Examples

(E3) **multiple-cuts scheme:** assume $\Gamma = \Gamma(\cdot; C)$ where $C \subset \mathbb{R}^n$ is the current bundle set and $\Gamma(\cdot; C) := \max\{\ell_f(\cdot; c) : c \in C\} + h$, choose the next bundle set C^+ satisfying

$$C(x) \cup \{x\} \subset C^+ \subset C \cup \{x\}, \quad C(x) := \{c \in C : \ell_f(x; c) + h(x) = \Gamma(x)\},$$

and then set $\Gamma^+ = \Gamma(\cdot; C^+)$ and $\bar{\Gamma} = \Gamma(\cdot; C(x))$.

Generic proximal bundle (GPB) framework

0. Let $x_0 \in \operatorname{dom} h$, $\lambda > 0$, $\bar\varepsilon > 0$ and $\tau \in [\bar\tau, 1)$ be given where

$$\bar\tau = \left[1 + \frac{(1+\lambda\bar\mu)\bar\varepsilon}{8\lambda T_{\bar\varepsilon}}\right]^{-1}, \quad T_{\bar\varepsilon} := \left(\bar M_f^2 + \bar\varepsilon \bar L_f\right)^{1/2}, \tag{6}$$

and set $y_0 = x_0$, $t_0 = 0$ and $j = 0$;

1. if $t_j \le \bar\varepsilon/2$, then perform a **serious update**, i.e., set $x_{j+1}^c = x_j$ and find $\Gamma_{j+1} \in \mathcal{C}_\mu(\phi)$ such that $\Gamma_{j+1} \ge \ell_f(\cdot; x_j) + h$; else, perform a **null update**, i.e., set $x_{j+1}^c = x_j^c$ and find $\Gamma_{j+1} \in \mathcal{C}_\phi(\Gamma_j, x_j^c, \lambda, \tau)$;

2. compute

$$x_{j+1} = \operatorname*{argmin}_{u \in \mathbb{R}^n} \left\{ \Gamma^{\lambda}_{j+1}(u) := \Gamma_{j+1}(u) + \frac{1}{2\lambda}\|u - x^c_{j+1}\|^2 \right\}, \qquad (7)$$

choose $y_{j+1} \in \{x_{j+1}, y_j\}$ such that

$$\phi^{\lambda}_{j+1}(y_{j+1}) = \min\left\{\phi^{\lambda}_{j+1}(x_{j+1}), \phi^{\lambda}_{j+1}(y_j)\right\} \qquad (8)$$

where ϕ^{λ}_j is defined as

$$\phi^{\lambda}_j := \phi + \frac{1}{2\lambda}\|\cdot - x^c_j\|^2, \qquad (9)$$

and set

$$m_{j+1} = \Gamma^{\lambda}_{j+1}(x_{j+1}), \quad t_{j+1} = \phi^{\lambda}_{j+1}(y_{j+1}) - m_{j+1}; \qquad (10)$$

3. set $j \leftarrow j+1$ and go to step 1.

GPB vs. standard bundle method

- introduce an auxiliary iterate y_j, convergence in $\{y_j\}$
- null/serious iterate decision making based on t_j
- motivation for y_j and t_j:
  define $m_j^* := \min\{\phi_j^\lambda(u) : u \in \mathbb{R}^n\}$, then we have

$$m_j \le m_j^* \le \phi_j^\lambda(y_j),$$

  and hence

$$\phi_j^\lambda(y_j) - m_j^* \le t_j \le \frac{\bar{\varepsilon}}{2}$$

  where $t_j = \phi_j^\lambda(y_j) - m_j$.

RPB can be viewed as GPB with bundle update scheme (E3).

While RPB only deals with the nonsmooth case ($L_f = 0$), GPB extends the analysis to the hybrid case ($L_f \geq 0$).

Complexity for GPB variants

Theorem

Let $x_0 \in \operatorname{dom} h$, $\bar\varepsilon > 0$ and $C > 0$ be given. Then, any variant of GPB with input $(x_0, \lambda, \bar\varepsilon, \tau)$ satisfying

$$\tau = \left[1 + \frac{(1+\lambda\mu)\bar\varepsilon}{8\lambda(M_f^2 + \bar\varepsilon L_f)}\right]^{-1}, \quad \frac{\bar\varepsilon}{C(M_f^2 + \bar\varepsilon L_f)} \le \lambda \le \frac{C d_0^2}{\bar\varepsilon}, \tag{11}$$

has $\bar\varepsilon$-iteration complexity given (up to a logarithmic term) by

$$\mathcal{O}_1\left(\min\left\{\frac{(M_f^2 + \bar\varepsilon L_f)d_0^2}{\bar\varepsilon^2}, \left(\frac{M_f^2 + \bar\varepsilon L_f}{\mu\bar\varepsilon} + 1\right)\log\left(\frac{\mu d_0^2}{\bar\varepsilon} + 1\right)\right\}\right). \tag{12}$$

Complexity for τ-free GPB variants in the nonsmooth case

Theorem

Let $x_0 \in \operatorname{dom} h$, $\bar{\varepsilon} > 0$ and $C > 0$ be given. Then, any variant of the τ-free GPB subclass with input $(x_0, \lambda, \bar{\varepsilon})$ satisfying

$$\frac{\bar{\varepsilon}}{CM_f^2} \leq \lambda \leq \frac{Cd_0^2}{\bar{\varepsilon}}, \tag{13}$$

has $\bar{\varepsilon}$-iteration complexity given (up to a logarithmic term) by

$$\mathcal{O}_1\left(\min\left\{\frac{M_f^2 d_0^2}{\bar{\varepsilon}^2}, \left(\frac{M_f^2}{\mu\bar{\varepsilon}} + 1\right)\log\left(\frac{\mu d_0^2}{\bar{\varepsilon}} + 1\right)\right\}\right). \tag{14}$$

Complexity of RPB in the strongly convex case

Theorem

Let $x_0 \in \mathrm{dom}\, h$, $\bar{\varepsilon} > 0$ *and* $C > 0$ *be given, and assume*

$$\frac{CM_f d_0}{\bar{\varepsilon}} \geq 1, \qquad 0 \leq \mu \leq \frac{CM_f}{d_0}. \tag{15}$$

Then, RPB with input $(x_0, \lambda, \bar{\varepsilon})$ *satisfying*

$$\frac{d_0}{M_f} \leq \lambda \leq \frac{Cd_0^2}{\bar{\varepsilon}} \tag{16}$$

has $\bar{\varepsilon}$*-iteration complexity given by*

$$\mathcal{O}_1\left(\min\left\{\frac{M_f^2 d_0^2}{\bar{\varepsilon}^2}, \left(\frac{M_f^2}{\mu\bar{\varepsilon}} + 1\right)\log\left(\frac{\mu d_0^2}{\bar{\varepsilon}} + 1\right)\right\}\right). \tag{17}$$

Complexity of RPB in the convex case

Theorem

Let $x_0 \in \mathrm{dom}\, h$, $\bar{\varepsilon} > 0$ *and* $C > 0$ *be given, and assume*

$$\frac{CM_f d_0}{\bar{\varepsilon}} \geq 1, \qquad M_h \leq CM_f, \qquad \mu = 0. \tag{18}$$

Then, RPB with input $(x_0, \lambda, \bar{\varepsilon})$ *satisfying*

$$\frac{\bar{\varepsilon}}{CM_f^2} \leq \lambda \leq \frac{Cd_0^2}{\bar{\varepsilon}} \tag{19}$$

has $\bar{\varepsilon}$*-iteration complexity given by* $\mathcal{O}_1(M_f^2 d_0^2/\bar{\varepsilon}^2)$.

One-cut Adaptive Proximal Bundle (1C-APB)

0. Let $x_0 \in \mathrm{dom}\, h$, $\lambda > 0$, $\tau_0 = 0$ and $\bar{\varepsilon} > 0$ be given, and set $y_0 = x_0$, $t_0 = 0$ and $j = 0$;

1. set $\tau = \tau_j$;

2. if $t_j \le \bar{\varepsilon}/2$, then perform a **serious update**, i.e., set $x^c_{j+1} = x_j$ and $\Gamma_{j+1} = \ell_f(\cdot; x_j) + h$; else, perform a **null update**, i.e., set $x^c_{j+1} = x^c_j$ and $\Gamma_{j+1} = \tau\Gamma_j + (1-\tau)[\ell_f(\cdot; x_j) + h]$;

3. compute x_{j+1}, y_{j+1}, m_{j+1} and t_{j+1} as in step 2 of GPB;

4. if $t_j > \bar{\varepsilon}/2$ and $t_{j+1} > \tau t_j + (1-\tau)\bar{\varepsilon}/4$, then set $\tau = (1+\tau)/2$ and go to step 2; else, set $\tau_{j+1} = \tau$ and $j \leftarrow j+1$, and go to step 1.

Complexity of 1C-APB

The general $\bar{\varepsilon}$-iteration complexity for 1C-APB is

$$\tilde{\mathcal{O}}\left(\left[\frac{\lambda_\mu(M_f^2+\bar{\varepsilon}L_f)}{\bar{\varepsilon}}+1\right]\left[\min\left\{\frac{d_0^2}{\lambda\bar{\varepsilon}},\frac{1}{\mu\lambda_\mu}\right\}+1\right]\right).$$

Under same assumptions as in previous theorems of GPB, 1C-APB has the same iteration complexity as GPB.
The total number of times τ is updated in step 4 is at most

$$\left\lceil\log\left(1+\frac{8\lambda_\mu(M_f^2+\bar{\varepsilon}L_f)}{\bar{\varepsilon}}\right)\right\rceil.$$

Extension: Stochastic Proximal Bundle Method

Assume f is deterministic and f has stochastic first-order oracle $s(x;\xi)$ such that

(A5) $\mathbb{E}\left[s\left(x;\xi\right)\right] = f'(x) \in \partial f(x)$ and $\mathbb{E}\left[\left\|s\left(x;\xi\right) - f'(x)\right\|^2\right] \le \sigma^2$.

It is uncertain whether $\mathbb{E}[\Gamma] \le \phi$ if Γ is as in (E2) or (E3), hence the standard PB method is unsuitable to handle stochastic problems.

On the other hand, it is easy to verify that $\mathbb{E}[\Gamma] \le \phi$ if Γ is as in (E1), i.e., one-cut model.

Another extension: sampling from a density $\propto \exp(-f(x))$ where f is a convex and Lipschitz continuous function.
Complexity $\tilde{\mathcal{O}}(d\varepsilon^{-1})$, the best in the high accuracy regime, i.e., $\varepsilon < d^{-1/3}$.

Concluding remarks

- A generic GPB framework for hybrid convex composite optimization
- Including most proximal bundle variants such as RPB, and a novel one based on the one-cut model
- A unified and simple analysis and stronger complexity results
- An adaptive variant that requires no prior knowledge of problem parameters
- A stochastic proximal bundle method based on the one-cut model

THE END

Thanks!