Variance Reduction and Low Sample Complexity in Stochastic Optimization via Proximal Point Methods

Jiaming Liang

University of Rochester

October 21, 2023

SIAM-NNP Annual Meeting, Newark, NJ

Introduction

Main problem: Stochastic convex composite optimization

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

Black-box model

(A1) f is μ-strongly convex, h is closed convex, and $\mathrm{dom}\, f \supset \mathrm{dom}\, h$;

(A2) for almost every $\xi \in \Xi$, there exist $F(\cdot,\xi) : \mathrm{dom}\, h \to \mathbb{R}$ and $s(\cdot,\xi) : \mathrm{dom}\, h \to \mathbb{R}^n$ satisfying

$$f(x) = \mathbb{E}[F(x,\xi)], \quad \nabla f(x) = \mathbb{E}[s(x,\xi)];$$

(A3) for every $x \in \mathrm{dom}\, h$, $\mathbb{E}[\|s(x,\xi) - \nabla f(x)\|^2] \le \sigma^2$;

(A4) for every $x, y \in \mathrm{dom}\, h$, $\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|$;

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

Motivation

Standard complexity results of stochastic gradient methods

$$x_i = \operatorname{argmin}\left\{\langle s(x_{i-1}, \xi_{i-1}), x\rangle + h(x) + \frac{1}{2\lambda}\|x - x_{i-1}\|^2\right\}, \quad \lambda \le \min\left\{\frac{\varepsilon}{4\sigma^2}, \frac{1}{4L}\right\}$$

To obtain $\mathbb{E}[\phi(x)] - \phi_* \le \varepsilon$,

$$\tilde{\mathcal{O}}\left(\max\left\{\kappa, \frac{\sigma^2}{\mu\varepsilon}\right\}\right)$$

where $\kappa = L/\mu$. To obtain $\mathbb{P}(\phi(x) - \phi_* \le \varepsilon) \ge 1 - p$,

$$\tilde{\mathcal{O}}\left(\max\left\{\kappa, \frac{\sigma^2}{\mu\varepsilon p}\right\}\right).$$

Improving $1/p$ to $\log(1/p)$ with sub-Gaussian assumption,

$$\mathbb{E}\left[\exp\left(\|s(x,\xi) - \nabla f(x)\|^2/\sigma^2\right)\right] \le \exp(1).$$

Preview

Sample complexity $\mathbb{P}(\phi(x) - \phi_* \leq \varepsilon) \geq 1 - p$,

$$\tilde{\mathcal{O}}\left(\max\left\{\frac{L}{\mu}, \frac{\kappa\sigma^2}{\mu\varepsilon}\right\} \log\frac{1}{p}\right)$$

Techniques

- Inexact proximal point method (IPPM)
- Bundle-type stochastic approximation (SA) method
- Probability booster

Inexact Proximal Point Method

Approximately solve the proximal problem

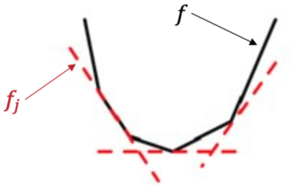
$$\hat{x} := \operatorname{argmin} \left\{ f(x) + h(x) + \frac{1}{2\lambda}\|x - x^c\|^2 \right\}$$

by an iterative process

$$x_j \leftarrow \min \left\{ f_j(x) + h(x) + \frac{1}{2\lambda}\|x - x^c\|^2 \right\}.$$

Recursively build up a cutting-plane model

$$f_j(x) = \max_{0 \le i \le j-1} \{\ell_f(x; x_i) := f(x_i) + \langle f'(x_i), x - x_i \rangle\}$$



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,

$$\mathbb{E}[\Gamma_j(u)] \geq \max_{0 \leq i \leq j-1} \{\mathbb{E}[F(x_i, \xi_i) + \langle s(x_i, \xi_i), u - x_i \rangle]\}.$$

On the other hand,

$$\begin{aligned} &\max_{0 \leq i \leq j-1} \{\mathbb{E}[F(x_i, \xi_i) + \langle s(x_i, \xi_i), u - x_i \rangle]\} \\ = &\max_{0 \leq i \leq j-1} \{f(x_i) + \langle f'(x_i), u - x_i \rangle\} \leq f(u) \end{aligned}$$

So

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

Other Bundle Models

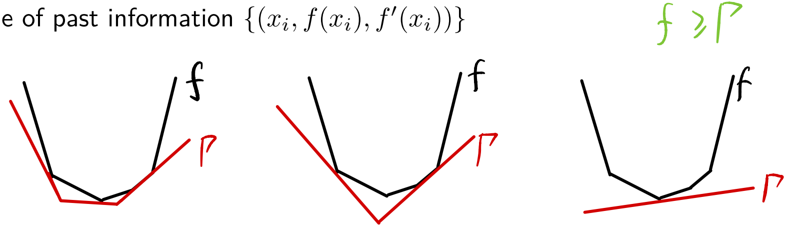
(E1) **single cut update**[1]**:** $\Gamma^+ := \tau\Gamma + (1-\tau)\ell_f(\cdot;x)$.

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

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

where $A_f^+ = \theta A_f + (1-\theta)\ell_f(\cdot;x^-)$.

Bundle of past information $\{(x_i, f(x_i), f'(x_i))\}$



Multiple cuts | Two cuts | One cut

[1] Liang, Guigues, and Monteiro. A single cut proximal bundle method for stochastic convex composite optimization. 2022.

Single Cut Model in the Stochastic Setting

Aggregate all cuts into a single one

$$\Gamma^+(u) = \tau\Gamma(u) + (1-\tau)[F(x,\xi) + \langle s(x,\xi), u - x\rangle].$$

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

$$\mathbb{E}[\Gamma^+(u)] \leq f(u).$$

Proximal Subproblem Solver, PSS(x_0, λ, I)

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

0. Set $i = 1$ and
$$\tau = \frac{I/2 + \lambda L}{1 + I/2 + \lambda L}; \tag{1}$$
1. take an independent sample ξ_{i-1} and compute
$$S_i = \begin{cases} s(x_0, \xi_0), & \text{if } i = 1, \\ (1-\tau)s(x_{i-1}, \xi_{i-1}) + \tau S_{i-1}, & \text{otherwise,} \end{cases} \tag{2}$$
$$x_i = \underset{u \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ h(u) + \langle S_i, u \rangle + \frac{1}{2\lambda} \|u - x_0\|^2 \right\}, \tag{3}$$
$$y_i = \begin{cases} x_i, & \text{if } i = 1, \\ (1-\tau)x_i + \tau y_{i-1}, & \text{otherwise;} \end{cases} \tag{4}$$
2. **if** $i < I + 1$, set $i \leftarrow i + 1$ and go to step 1; **otherwise, stop**.

Output: x_{I+1} and y_{I+1}.

Convergence Guarantee

Define

$$u_i := \begin{cases} F(x_1, \xi_1) + h(x_1) + \frac{1}{2\lambda}\|x_1 - x_0\|^2, & \text{if } i = 1, \\ (1-\tau)\left[\phi(x_i) + \frac{1}{2\lambda}\|x_i - x_0\|^2\right] + \tau u_{i-1}, & \text{otherwise,} \end{cases}$$

$$\Gamma_i(\cdot) := \begin{cases} F(x_0, \xi_0) + \langle s(x_0, \xi_0), \cdot - x_0 \rangle + h(\cdot), & \text{if } i = 1, \\ (1-\tau)\ell(\cdot; x_{i-1}, \xi_{i-1}) + \tau \Gamma_{i-1}(\cdot), & \text{otherwise,} \end{cases}$$

$$t_i := u_i - \left[\Gamma_i(x_i) + \frac{1}{2\lambda}\|x_i - x_0\|^2\right].$$

Then, we have

$$\mathbb{E}[t_{I+1}] \leq \tau^I \left(\sigma D + \frac{LD^2}{2}\right) + \frac{\lambda \sigma^2}{I}.$$

Variance reduction by a factor of I.

Translation into IPPM

Our objective in the k-th proximal subproblem is to approximately solve

$$z_k^* := \operatorname{argmin} \left\{ \phi(x) + \frac{1}{2\lambda} \|x - z_{k-1}\|^2 \right\}$$

through

$$z_k = \operatorname{argmin} \left\{ \tilde{\Gamma}_k(x) + \frac{1}{2\lambda} \|x - z_{k-1}\|^2 \right\}.$$

Let $(z_k, w_k) = (x_{I+1}, y_{I+1})$, then the convergence guarantee of PSS translates into

$$\mathbb{E} \left[\phi(w_k) + \frac{1}{2\lambda} \|w_k - z_{k-1}\|^2 - \tilde{\Gamma}_k(z_k) - \frac{1}{2\lambda} \|z_k - z_{k-1}\|^2 \right] \leq \varepsilon$$

where

$$\varepsilon = \tau^I \left(\sigma D + \frac{LD^2}{2} \right) + \frac{\lambda \sigma^2}{I}.$$

Guarantee in Probability

It can be shown that with probability at least $3/4$,

$$\phi(w_k) + \frac{1}{2\lambda}\|w_k - z_{k-1}\|^2 - \phi(z_k^*) - \frac{1}{2\lambda}\|z_k^* - z_{k-1}\|^2 + \frac{1+\lambda\mu}{\lambda(2+\lambda\mu)}\|z_k^* - z_k\|^2 \le 8\varepsilon.$$

Assume $\lambda\mu \ge 3$, then with probability at least $(3/4)^K$, we have

$$\phi(\bar{w}) - \phi^* \le 16\varepsilon,$$

for some $\bar{w} \in \operatorname{dom} h$ and

$$K \ge 2\left(1 + \frac{9}{\lambda\mu}\right)\log\left(\frac{\mu d_0^2}{8\varepsilon} + 1\right).$$

This result is in low probability.

Probability Booster

Consider

$$\hat{z} := \operatorname{argmin}\left\{\phi^\lambda(x) := \phi(x) + \frac{1}{2\lambda}\|x - z\|^2\right\}$$

and suppose the PSS generates (z^j, w^j) such that

$$\mathbb{P}\left(\phi^\lambda(w^j) - \phi^\lambda(\hat{z}) + \frac{1+\lambda\mu}{\lambda(2+\lambda\mu)}\|\hat{z} - z^j\|^2 \le \varepsilon\right) \ge \frac{3}{4}.$$

Calling PSS for n times, the probability booster [2] (based on *Robust Distance Estimate* [3]) improves the result to

$$\mathbb{P}\left(\phi^\lambda(w^j) - \phi^\lambda(\hat{z}) + \frac{1+\lambda\mu}{\lambda(2+\lambda\mu)}\|\hat{z} - z^j\|^2 \le \kappa\varepsilon\right) \ge 1 - 2\exp\left(-\frac{n}{72}\right).$$

[2] Davis, Drusvyatskiy, Xiao, and Zhang. From low probability to high confidence in stochastic convex optimization. JMLR, 2021.

[3] Nemirovski and Yudin. Problem complexity and method efficiency in optimization. 1983.

Robust Distance Estimtor

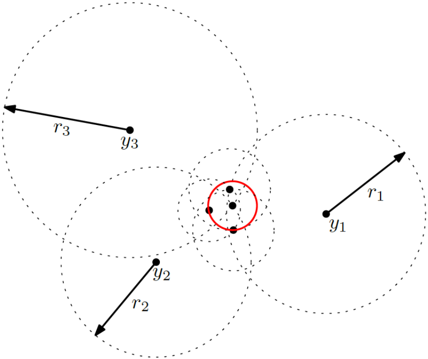


Figure: Robust Distance Esitmater

A clustering technique proposed by Nemirovski and Yudin to improve the low-confidence estimate to a high-confidence guarantee by generating multiple statistically independent points via the low-confidence oracle.

PB$(\{(z^j, w^j)\}_{j=1}^n, q)$

Input: Independent pairs $(z^1, w^1), \ldots, (z^n, w^n)$ generated by the oracle PSS(z, λ, I) and an integer $q \geq 1$.

1. Compute
$$\mathcal{J}_1 := \mathrm{Extract}(\{w^j\}_{j=1}^n, \|\cdot\|_2);$$
2. Compute
$$\mathcal{J}_2 := \mathrm{Extract}(\{z^j\}_{j=1}^n, \|\cdot\|_2);$$
3. Fix arbitrary $j \in \mathcal{J}_1 \cap \mathcal{J}_2$ and set $\bar{w} := w^j$. Use the robust gradient estimator RGE$(\bar{w}, n, q)$ to generate $\widetilde{\nabla} f(\bar{w})$;
4. Define the pseudometric $\rho\left(x, x'\right) := \left| h(x) - h\left(x'\right) + \left\langle \widetilde{\nabla} f(\bar{w}), x - x' \right\rangle \right|$ on dom h and compute
$$\mathcal{J}_3 = \mathrm{Extract}(\{z^j\}_{j=1}^n, \rho).$$

Output: A pair (z^j, w^j) for an arbitrary $j \in \mathcal{J}_1 \cap \mathcal{J}_2 \cap \mathcal{J}_3$.

Improved Probability Result

Proposition

If

$$q \geq \frac{(1+\lambda\mu)\sigma^2}{4L^2\lambda\varepsilon},$$

then with probability at least $1 - 2\exp(-n/72)$, *the pair* (w^j, z^j) *returned by PB satisfies*

$$\phi^\lambda(w^j) - \phi^\lambda(\hat{z}) + \frac{1+\lambda\mu}{\lambda(2+\lambda\mu)}\|z^j - \hat{z}\|^2 \leq 168\varepsilon + 648\kappa\varepsilon.$$

Each PB oracle has an RGE oracle, which takes nq *stochastic gradient samples.*

Inexact Proximal Point Method, IPPM$(z_0, \lambda, n, q, I, K)$

Input: Scalar $\lambda > 0$, integers $n, q, I, K \geq 1$, and initial point $z_0 \in \operatorname{dom} h$.

0. Set $k = 1$;
1. call the oracle PSS(z_{k-1}, λ, I) n times and generate independent pairs $(z_k^1, w_k^1), \ldots, (z_k^n, w_k^n)$;
2. call the oracle PB$(\{(z_k^j, w_k^j)\}_{j=1}^n, q)$ to generate (z_k, w_k);
3. **if** $k < K$, set $k \leftarrow k + 1$ and go to step 1; **otherwise**, **stop**.

Main Result

Assume $\lambda\mu \geq 3$, if $n = \mathcal{O}(\log(1/p))$, then we have

$$\mathbb{P}\left(\phi(\bar{w}) - \phi^* \leq \bar{\varepsilon}\right) \geq 1 - 2K \exp\left(-\frac{n}{72}\right) \geq 1 - p$$

for some $\bar{w} \in \operatorname{dom} h$ and

$$K = \tilde{\mathcal{O}}\left(1 + \frac{1}{\lambda\mu}\right) = \tilde{\mathcal{O}}(1).$$

Inner precision

$$\frac{\bar{\varepsilon}}{\kappa} = \varepsilon = \tau^I\left(\sigma D + \frac{LD^2}{2}\right) + \frac{\lambda\sigma^2}{I},$$

so

$$I = \tilde{\mathcal{O}}\left(\max\left\{\frac{1}{1-\tau}, \frac{\kappa\lambda\sigma^2}{\bar{\varepsilon}}\right\}\right) = \tilde{\mathcal{O}}\left(\max\left\{\kappa, \frac{\kappa\sigma^2}{\mu\bar{\varepsilon}}\right\}\right).$$

Finally, sample complexity (of stochastic gradient oracles) is

$$KIn = \tilde{\mathcal{O}}\left(\max\left\{\kappa, \frac{\kappa\sigma^2}{\mu\bar{\varepsilon}}\right\}\log\frac{1}{p}\right).$$

Conclusion

- A bundle-type SA method for stochastic programming
- A single cut aggregating all past information
- Variance reduction and low sample complexity

Extensions:

- Adaptive $\{\lambda_k\}$ and $\{I_k\}$ to remove the overhead κ
- Universal method without knowing μ and L
- Nesterov's acceleration with restart

Thank you!

Comparison with proxBoost[4]

Algorithm 2: proxBoost(δ, p, T)

Input: $\delta \geq 0$, $p \in (0,1)$, $T \in \mathbb{N}$

Set $\lambda_{-1} = 0$, $\varepsilon_{-1} = \sqrt{\frac{2\delta}{\mu}}$

Generate a point x_0 satisfying $\boxed{\|x_0 - \bar{x}_0\| \leq \varepsilon_{-1}}$ with probability $1 - p$.

for $j = 0, \ldots, T-1$ **do**

Set $\varepsilon_j = \sqrt{\frac{2\delta}{\mu + \lambda_j}}$

Generate a point x_{j+1} satisfying

$$\boxed{\mathbb{P}\left[\|x_{j+1} - \bar{x}_{j+1}\| \leq \varepsilon_j \mid E_j\right] \geq 1 - p,} \tag{13}$$

where E_j denotes the event $E_j := \left\{x_i \in B_{\varepsilon_{i-1}}(\bar{x}_i) \text{ for all } i \in [0, j]\right\}$.

end

Generate a point x_{T+1} satisfying

$$\boxed{\mathbb{P}\left[f^T(x_{T+1}) - \min f^T \leq \delta \mid E_T\right] \geq 1 - p.} \tag{14}$$

Return x_{T+1}

Our proposed algorithm does not rely on other (possibly not implementable) oracles, but we loose a factor of κ in sample complexity.

[4] Davis, Drusvyatskiy, Xiao, and Zhang. From low probability to high confidence in stochastic convex optimization. JMLR, 2021.

Robust Gradient Estimator, RGE(x, n, q)

Input: A point $x \in \operatorname{dom} h$ and integers $n, q \geq 1$.

1. Repeat for $j = 1, \ldots, n$: generate q independent stochastic gradients $s(x, \xi_j^1), \ldots, s(x, \xi_j^q)$, compute
$$\bar{s}_j(x) = \frac{1}{q} \sum_{i=1}^{q} s(x, \xi_j^i);$$
2. Denote $S = \{\bar{s}_1(x), \ldots, \bar{s}_j(x)\}$;
3. Repeat for $j = 1, \ldots, n$: compute $r_j = \min\left\{r \geq 0 : \left|B_r^2\left(\bar{s}_j(x)\right) \cap S\right| > 2n/3\right\}$;
4. Set $j^* = \operatorname{argmin}\{r_j : j \in \{1, \ldots, n\}\}$.

Output: $\bar{s}_{j^*}(x)$.

Extract$(\{z^j\}_{j=1}^n, \rho)$

Input: A set of n points $Z = \{z^1, \ldots, z^n\} \subset \operatorname{dom} h$ and a metric ρ on $\operatorname{dom} h$.

1. Repeat for $j = 1, \ldots, n$: compute $r_j = \min \left\{r \geq 0 : \left|B_r^\rho\left(z^j\right) \cap Z\right| > 2n/3\right\}$.

2. Compute the second tertile $\hat{r} = \operatorname{secondtertile}(r_1, \ldots, r_n)$;

Output: $\mathcal{J} = \{j \in \{1, \ldots, n\} : r_j \leq \hat{r}\}$.