Optimal Private Streaming SCO in ℓ_p-geometry with Applications in High Dimensional Online Decision Making

Yuxuan Han, Zhicong Liang, Zhipeng Liang, Yang Wang, Yuan Yao, Jiheng Zhang

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Stochastic Optimization

Dataset: $D := \{x_1, \dots, x_n\} \subset \mathcal{X}, x_i \sim_{i.i.d.} P$.

Parameter Space: Compact $\mathcal{C} \subset \mathbb{R}^d$

Loss Function: $f(\theta; x) : \mathcal{C} \times \mathcal{X} \to \mathbb{R}$

Population Loss: $F(\theta) := \mathbb{E}_{x \sim P}[f(\theta; x)]$

Population-Level Minimizer: $\theta^\star = \text{argmin}_{\theta \in \mathcal{C}} F(\theta)$

Goal: Control the excess risk

$$\mathcal{R}(\theta) := F(\theta) - F(\theta^\star).$$

Stochastic Convex Optimization (SCO) in $\ell_p, 1 \leq p \leq \infty$ geometry

$\mathcal{C}$ is a ℓ_p-ball.

F is convex and smooth under ℓ_p geometry:

$$\|\nabla F(\theta)\|_q \leq L, \quad \|\nabla F(\theta) - \nabla F(\theta')\|_q{}^1 \leq \beta, \quad \theta, \theta' \in \mathcal{C}.$$

Well studied optimal excess risk [ABRW14]

$$\mathcal{R}(\theta) = \begin{cases} \tilde{\Theta}(\frac{1}{\sqrt{n}}) & 1 < p \leq 2, \\ \tilde{\Theta}(\frac{d^{1/2-1/p}}{\sqrt{n}}) & 2 < p \leq \infty, \end{cases}$$

[1] $1/q + 1/p = 1$

Private SCO in ℓ_p geometry

Well investigated in recent works [AFVT21,BGN21,BFVG19]

$p = 1$

$$\textbf{Non-private: } \mathcal{R}(\theta) = \tilde{\Theta}(\frac{1}{\sqrt{n}}) \implies \textbf{Private: } \mathcal{R}(\theta) = \tilde{\Theta}(\frac{1}{\sqrt{n}} + (\frac{1}{n\varepsilon})^{2/3})$$

$1 < p \leq 2$

$$\textbf{Non-private: } \mathcal{R}(\theta) = \tilde{\Theta}(\frac{1}{\sqrt{n}}) \implies \textbf{Private: } \mathcal{R}(\theta) = \tilde{\Theta}(\frac{1}{\sqrt{n}} + \frac{\sqrt{d}}{n\varepsilon})$$

$2 < p \leq \infty$

$$\textbf{Non-private: } \mathcal{R}(\theta) = \tilde{\Theta}(\frac{d^{1/2-1/p}}{\sqrt{n}}) \implies \textbf{Private: } \mathcal{R}(\theta) = \tilde{O}(\frac{d^{1/2-1/p}}{\sqrt{n}} + \frac{d^{1-1/p}}{n\varepsilon})$$

All above upper bounds are achieved by algorithms with **large-batch** design.

Private SCO in ℓ_p geometry with continual release

1. **Streaming Data Structure:**

$$D = \{x_1, \dots, x_n\} \implies D_0 = \emptyset, D_t = D_{t-1} \cup \{x_t\}, \quad t = 1, \dots, T.$$

2. **Real-time Feedback:**

$\cdots \longrightarrow$ Receive $x_t \longrightarrow D_t = D_{t-1} \cup \{x_t\} \longrightarrow$ Update $\theta_t = \varphi(\theta_{t-1}, D_t) \longrightarrow$ Receive $x_{t+1} \longrightarrow \dots$

3. **Continual Release:**

Keep last iteration output private $\implies$ Keep $(\theta_1, \dots, \theta_T)$ private

Our Contributions

Develop algorithms for private SCO with continual release under ℓ_p.

1. $p = 1$: dimension-independent excess risk $\tilde{O}(\frac{1}{\sqrt{n}\varepsilon})$
2. $1 < p \leq 2$: Optimal excess risk $\tilde{O}(\frac{1}{\sqrt{n}} + \frac{\sqrt{d}}{n\varepsilon})$
3. $2 < p \leq \infty$: SOTA excess risk $\tilde{O}(\frac{d^{1/2-1/p}}{\sqrt{n}} + \frac{d^{1-1/p}}{n\varepsilon})$
4. Fast iteration: $\hat{\theta}_{t+1} \leftarrow \hat{\theta}_t$ with **two** gradient computation.

First **linear time** algorithm achieving optimal rate in $1 < p < 2$.

Frank-Wolfe Algorithm

Initialization: $\theta_1, d_1 = \nabla f(\theta_1; x_1)$.

For $t = 1$ to T:

1. Update d_t.
2. $v_t = \text{argmin}_{v \in \partial \mathcal{C}} \langle v, d_t \rangle$.
3. $\theta_{t+1} = (1 - \eta_t)\theta_t + \eta_t v_t$.

$\eta_t = 1/t$, d_t will be specified later.

Private version:

$$v_t = \text{argmin}_{v \in \partial \mathcal{C}} \langle v, d_t \rangle \implies v_t = \text{ Private Mechanism}\Big(\text{argmin}_{v \in \partial \mathcal{C}} \langle v, d_t \rangle\Big).$$

Main techniques

Recursive variance reduction:[XSZWQ20]

$$\nabla F(\theta_t) \approx f(\theta_t; x_t) \implies \nabla F(\theta_t) \approx d_t,$$

with $d_1 = \nabla f(\theta_1; x_1)$,

$$d_t = \nabla f(\theta_t; x_t) + \frac{t}{1+t}\big(d_{t-1} - \nabla f(\theta_{t-1}; x_t)\big), \quad t \geq 2.$$

Equivalently,

$$d_t = \frac{1}{t+1} \sum_{i=1}^{t} \big((i+1)\nabla f(\theta_i; x_i) - i\nabla f(\theta_{i-1}; x_i)\big).$$

Main techniques

Adaptive Sensitivity:

$$d_t = \frac{1}{t+1}\sum_{s=1}^{t}\Big((s+1)\nabla f(\theta_s; x_s) - s\nabla f(\theta_{s-1}; x_s)\Big).$$

For $D \sim_i D', x_i \neq x_i'$, control $(t+1)[d_t(D') - d_t(D)]$:

$$\underbrace{\Big((i+1)[\nabla f(\theta_i; x_i) - \nabla f(\theta_i; x_i')] - i\nabla[f(\theta_{i-1}; x_i) - \nabla f(\theta_{i-1}; x_i')]\Big)}_{\text{main contribution to sensitivity}, S_i} + \underbrace{\sum_{s<i}\dots}_{\text{un-affected}} + \underbrace{\sum_{s>i}\dots}_{\text{adaptive arguments}}$$

$$\|S_i\|_q \leq 2i\beta\|\theta_i - \theta_{i-1}\|_q + 2L \leq 2(\beta \cdot \mathsf{diam}(\mathcal{C}) + L)$$

Main techniques

With Sensitivity$(d_t) \asymp \frac{1}{t+1}$, we can specify the private mechanisms:

$p = 1$**:**

$$v_t = \text{argmin}_{v \in \partial \mathcal{C}} \langle v, d_t \rangle \implies v_t = \text{ Noisy-Min}_{v \in \partial \mathcal{C}} \langle v, d_t \rangle$$

$$\textbf{utility: } \langle v_t, d_t \rangle < \text{argmin}_v \langle v, \nabla F(\theta_t) \rangle + \tilde{O}(\frac{1}{\sqrt{t}\varepsilon}).$$

$p > 1$**:**

$$v_t = \text{argmin}_{v \in \partial \mathcal{C}} \langle v, d_t \rangle \implies v_t = \text{argmin}_{v \in \partial \mathcal{C}} \langle v, \text{Tree Mechanism}(d_t) \rangle$$

$$\textbf{utility: } \langle v_t, d_t \rangle < \text{argmin}_v \langle v, \nabla F(\theta_t) \rangle + \tilde{O}(\frac{1}{\sqrt{t}} + \frac{\sqrt{d}}{t\varepsilon}).$$

Then ours result can be derived by bringing the utility guarantees into classical routine of Frank-Wolfe analysis.

Conclusion

Contributions:

First rate-optimal private SCO algorithms for streaming data.

First rate-optimal private SCO algorithms with linear complexity for $1 < p < 2$.

Application in high dimensional Bandits.

Possible Directions:

Optimality in $p > 2$?

Optimal rate in linear time without smoothness?

Thanks!