Differentially Private Stochastic Convex Optimization under a Quantile Loss Function

Du Chen[1], Geoffrey A. Chua[1]

[1]Nanyang Business School, Nanyang Technological University

Introduction

(ε, δ)-Differential Privacy

A randomized algorithm $\mathcal{M} : \mathcal{X}^n \times \mathcal{Y}^n \to \mathcal{Z}$ is (ε, δ)-differential private if, for any pair of neighboring datasets $\mathcal{D} \sim \mathcal{D}'$ that differ in one data point, and for any subset $\mathcal{S} \subseteq \mathcal{Z}$, we have

$$\Pr[\mathcal{M}(\mathcal{D}) \in \mathcal{S}] \le e^{\varepsilon} \cdot \Pr[\mathcal{M}(\mathcal{D}') \in \mathcal{S}] + \delta.$$

Stochastic Convex Optimization (SCO) under a Quantile Loss

The goal is to output a high-quality estimator $\widehat{\boldsymbol{\theta}} := \arg\min_{\boldsymbol{\theta}} \widehat{\mathcal{L}}(\boldsymbol{\theta}; \mathcal{D})$, where $\widehat{\mathcal{L}}(\boldsymbol{\theta}; \mathcal{D}) := \frac{1}{n} \cdot \sum_{i=1}^{n} c(y_i - \langle \boldsymbol{\theta}, \boldsymbol{x} \rangle)$ is the Empiricak Risk Minimization (ERM) problem under a quantile loss.

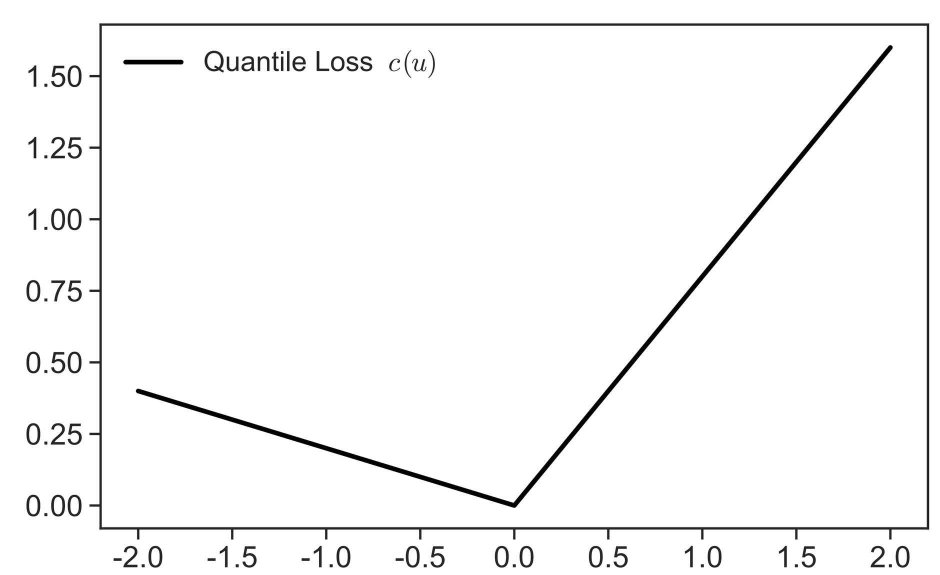

Figure 1. Quantile Loss Function $c(u) := ru^+ + (1-r)(-u)^+$

A quantile loss function allows imposing asymmetric weights on positive or negative values of u, and provides insights into distributional relationships between feature $\boldsymbol{x}$ and dependent variable y.

Research Question & Challenges

We are interested in designing DP algorithms that have provable privacy and performance guarantees for DP-SCO under a quantile loss function. However, the quantile loss is nonsmooth, which will lead to an unstable estimator and prevent gradient-based optimization methods from being efficient.

Contributions

1. We adopt **convolution smoothing** to address the nonsmoothness issue. For DP-SCO under a quantile loss, convolution smoothing **outperforms** existing methods such as Moreau Envelope.
2. We find that with convolution-smoothed functions, both Gradient Perturbation and Objective Perturbation can, under mild assumptions, achieve **optimal excess generalization risks**
$$\mathcal{L}(\widehat{\boldsymbol{\theta}}_h^{\pi}; \mathbb{P}) - \min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}; \mathbb{P}) \le \mathcal{O}\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d \ln(1/\delta)}}{n\varepsilon}\right), \quad \forall \mathbb{P}.$$
3. We derive an upper bound on Objective Perturbation estimator's error:
$$\mathbb{E}_{\mathrm{OP}}\left[\left\|\widehat{\boldsymbol{\theta}}_h^{\mathrm{OP}} - \boldsymbol{\theta}^*\right\|_2\right] \lesssim \frac{1}{\rho_1 \underline{f}} \cdot \left(\sqrt{\frac{d}{n}} + \sqrt{\frac{d \ln(1/\delta)}{n\varepsilon}}\right).$$

DP Algorithms

Main Idea: Convolution Smoothing then Solve Private ERM

Our approach relies on convolution smoothing:

$$\text{(convolution smoothing)} \quad c_h(u) := (c * K_h)(u) = \int_{-\infty}^{\infty} c(v) K_h(u - v)\, dv,$$

where $K_h(\cdot) := K(\cdot/h)/h$ is an adjusted kernel function parameterized by bandwidth $h > 0$, and $K(\cdot)$ is a kernel function. Intuitively, the value $c_h(u)$ is a weighted average over u's neighbors, and the weights are given by the adjusted kernel function $K_h(\cdot)$ so that a closer neighbor has a higher weight.

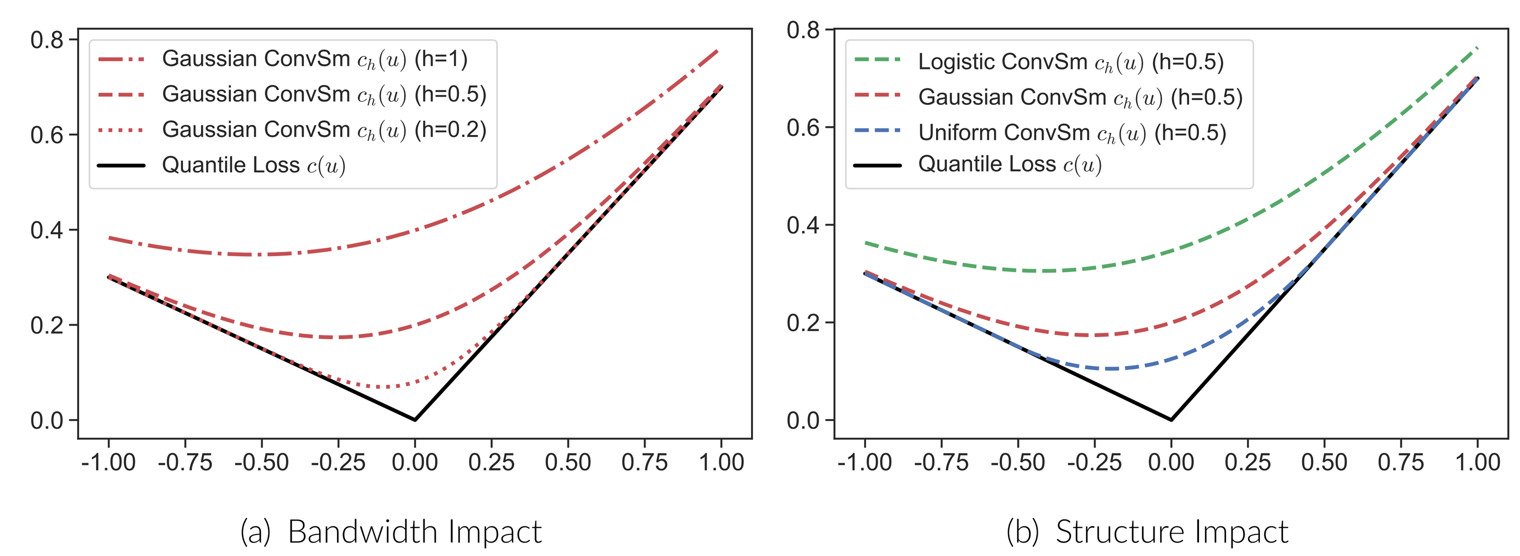

Figure 2. Convolution Smoothing. $r = 0.7$

Algo1: Gradient Perturbation (DP-SGD): We follow classic DP-SGD:

$$\widehat{\boldsymbol{\theta}}_{h,t+1} \leftarrow \widehat{\boldsymbol{\theta}}_{h,t} - \eta \cdot (\nabla \ell_h(\widehat{\boldsymbol{\theta}}_{h,t}; \boldsymbol{x}_{(t)}, y_{(t)}) + \boldsymbol{w}_t),$$

where $\boldsymbol{w}_t \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$ and $\sigma^2 \asymp \ln(1/\delta)/\varepsilon^2$

Algo2: Objective Perturbation (OP): let $\boldsymbol{b} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$ and $\sigma^2 \asymp (\ln(1/\delta) + \varepsilon)/\varepsilon^2$, then

$$\widehat{\boldsymbol{\theta}}_h^{\mathrm{OP}} \leftarrow \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \widehat{\mathcal{L}}_h(\boldsymbol{\theta}; \mathcal{D}) + \lambda \|\boldsymbol{\theta}\|_2^2 + \frac{\langle \boldsymbol{b}, \boldsymbol{\theta} \rangle}{n}$$

Comparison between Convolution Smoothing and Moreau Envelope:

$$\text{(Moreau Envelope)} \quad c_\beta(u) := \inf_v \{c(v) + \tfrac{\beta}{2}\|u - v\|_2^2\}.$$

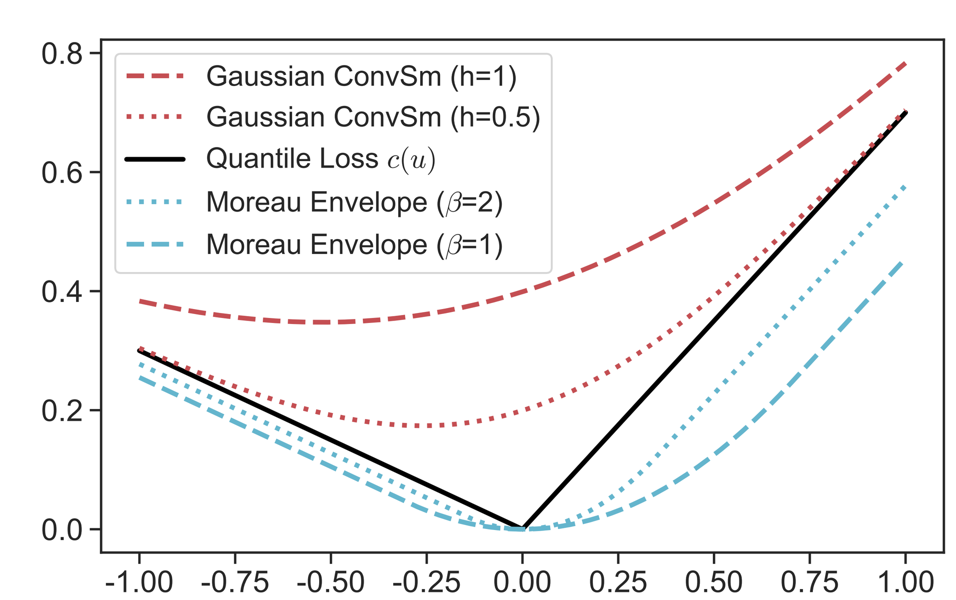

Figure 3. Convolution Smoothing v.s. Moreau Envelope. $r = 0.7$

	Convolution Smoothing (ours)	Moreau Envelope
Flexibility	kernel $K(\cdot)$ & bandwidth h	smoothness param β
Approx. from	above	below
Tolerate outliers?	✓	×

Table 1. Comparison between Convolution Smoothing and Moreau Envelope

Theoretical Results

Both algorithms are (ε, δ)-DP.

Optimal Excess Generalization Risk

By setting proper algorithms' parameters, we can achieve optimal excess generalization risk:

$$\mathcal{L}(\widehat{\boldsymbol{\theta}}_h^{\pi}; \mathbb{P}) - \min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}; \mathbb{P}) \le \mathcal{O}\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d \ln(1/\delta)}}{n\varepsilon}\right), \quad \forall \mathbb{P},$$

where π can be DP-SGD or OP. When running OP, DP parameter should satisfy $\varepsilon^4 + d \ln(1/\delta)\varepsilon^2 \ge \Omega(1/n)$ to ensure optimal rates.

Estimation Accuracy

Assume privacy parameter $\delta \asymp n^{-w}$ for some $w > 0$. Running OP with proper algorithm parameters yields

$$\mathbb{E}_{\mathrm{OP}}\left[\left\|\widehat{\boldsymbol{\theta}}_h^{\mathrm{OP}} - \boldsymbol{\theta}^*\right\|_2\right] \lesssim \frac{1}{\rho_1 \underline{f}} \cdot \left(\sqrt{\frac{d + \ln(1/\gamma)}{n}} + \sqrt{\frac{d \ln(1/\delta)}{n\varepsilon}}\right),$$

with probability at least $1 - \gamma, \forall \gamma \in (0, 1)$ over the random draw of dataset $\mathcal{D}$, where $\rho_1 := \lambda_{\min}(\boldsymbol{\Sigma}) > 0$ and $\underline{f} > 0$ are parameters of the groundtruth data generating process.

Experiments

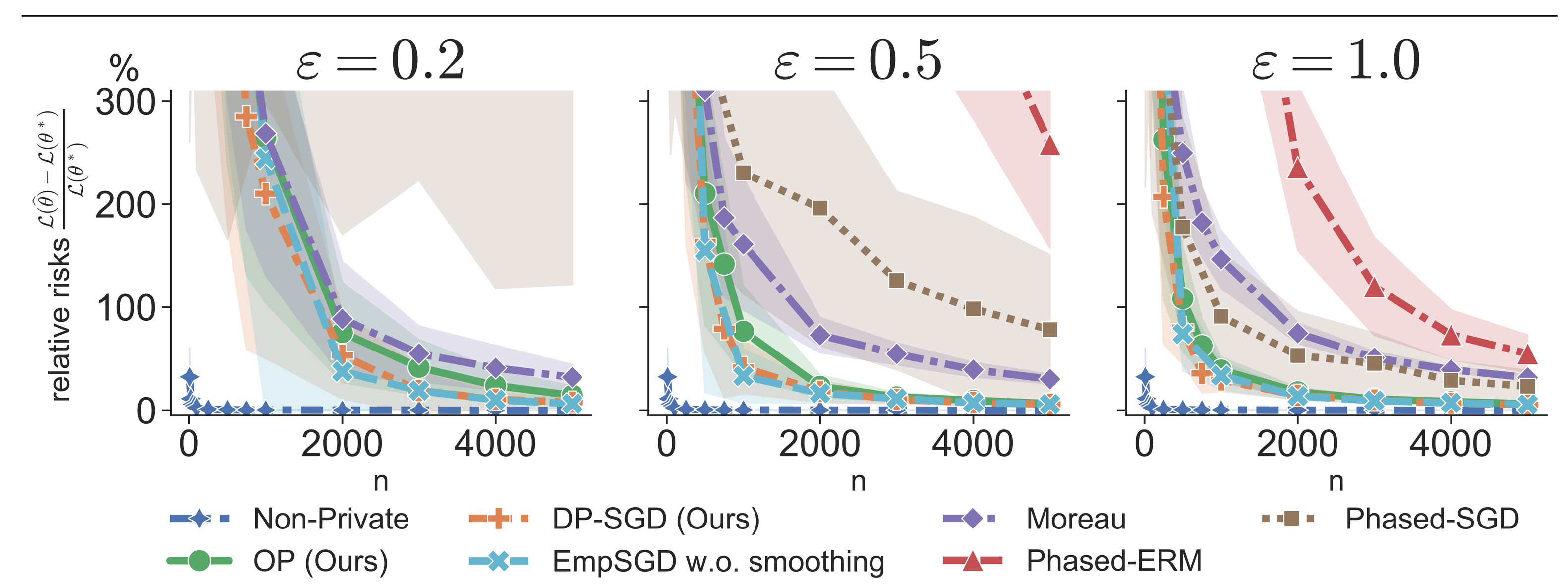

Figure 4. (d=3) Excess generalization risks. Groundtruth $y = 10 + 5x_1 - 2x_2 + \epsilon$, where $(x_1, x_2) \sim \mathcal{N}(\boldsymbol{0}, \begin{pmatrix} 2^2, 0 \\ 0, 3^2 \end{pmatrix})$, and $\epsilon \sim \mathcal{N}(0, 3^2)$. Quantile $r = 0.7$, privacy param $\delta = 10^{-2}$

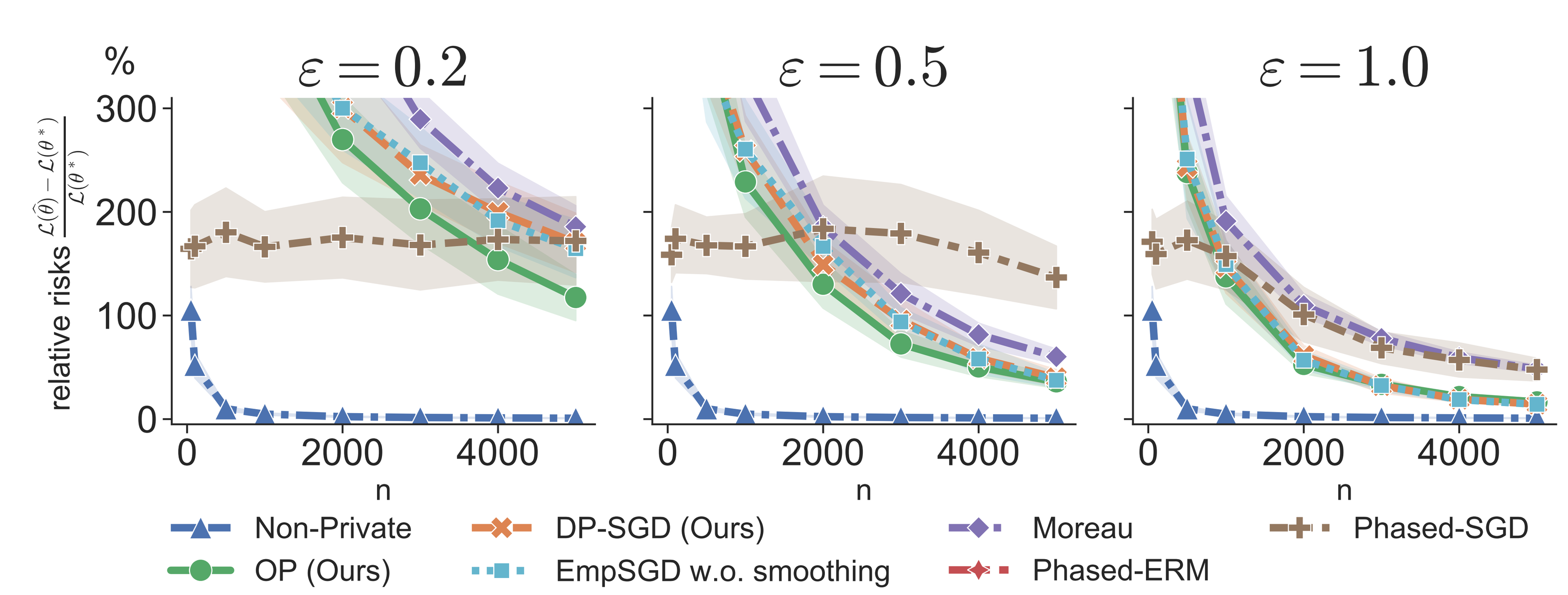

Figure 5. (d=51) Excess generalization risks. Groundtruth $y = 10 + \langle \boldsymbol{\theta}, \boldsymbol{x} \rangle + \epsilon$, $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{x}}, \Sigma_{\boldsymbol{x}})$ with mean $\boldsymbol{\mu}_{\boldsymbol{x}} = [0, \dots, 0] \in \mathbb{R}^{50}$ and covariance matrix $\Sigma_{\boldsymbol{x}} = Diag([\frac{1}{\sqrt{50}}, \dots, \frac{1}{\sqrt{50}}])$; $\boldsymbol{\theta}_{[1:50]} \in \mathbb{R}^{50}$ take values ascendingly from [-2, 5] with even steps, and $\epsilon \sim \mathcal{N}(0, 3^2)$. Quantile $r = 0.7$, privacy param $\delta = 10^{-2}$