Private Algorithm for Quantile Regression

Tuoyi Zhao[1]
Joint work with Wen-Xin Zhou[2] & Lan Wang[1]

[1]Department of Management Science,
University of Miami
[2]Department of Information and Decision Sciences,
University of Illinois at Chicago

Workshop on Translational Research on Data Heterogeneity
Washington University in St. Louis

April 7, 2024

UNIVERSITY OF MIAMI
MIAMI HERBERT
BUSINESS SCHOOL

Outline

Data Privacy

Data Privacy - Netflix Competition

- Netflix launched a competition in 2006 for predicting customer ratings of movies.
- Netflix released a dataset with over 100 million ratings from more than 480 thousand users. It anonymized user IDs and added noise to the ratings.
- In 2008, researchers from UT Austin de-anonymized some of the Netflix data by comparing rankings and timestamps with public information in the Internet Movie Database, or IMDb.

Data Privacy - Netflix Competition

Robust De-anonymization of Large Datasets
(How to Break Anonymity of the Netflix Prize Dataset)

Arvind Narayanan and Vitaly Shmatikov

The University of Texas at Austin

Data Privacy - Netflix Competition

Robust De-anonymization of Large Datasets
(How to Break Anonymity of the Netflix Prize Dataset)

Arvind Narayanan and Vitaly Shmatikov

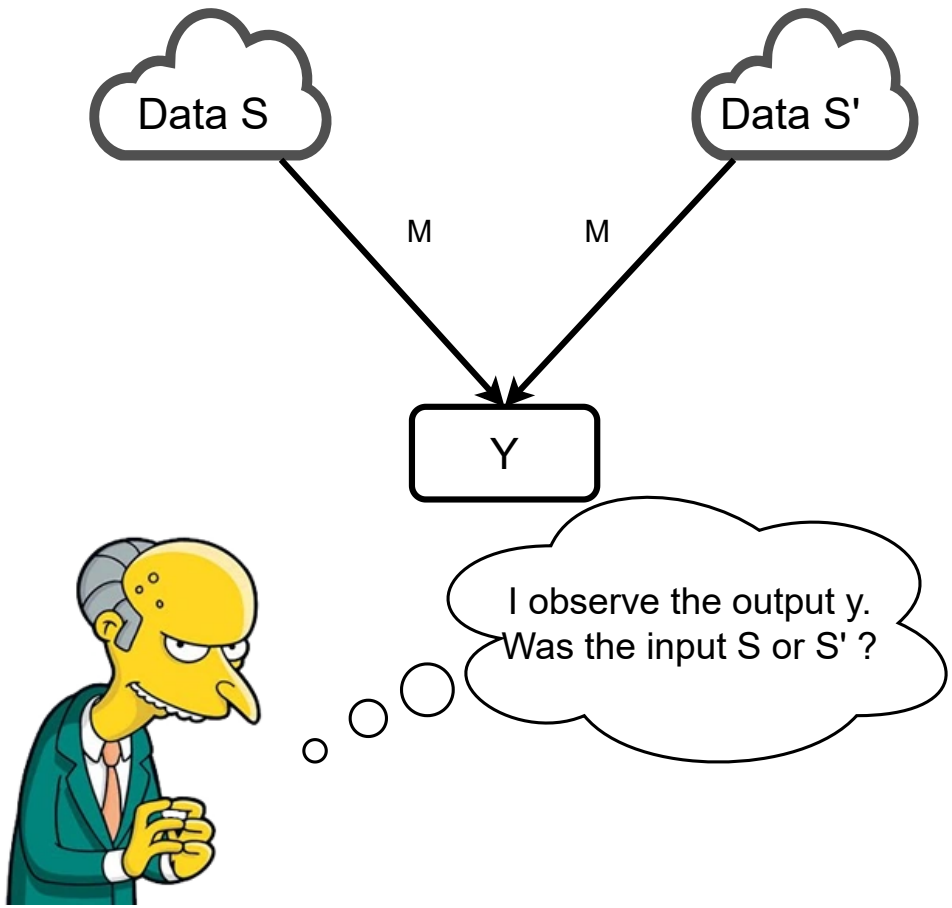
The University of Texas at Austin

- Privacy protection is not an easy task!

Differential Privacy (DP)

- A pair of data sets $\mathcal{S}$ and $\mathcal{S}'$ are said to be neighboring data sets if they differ in only one data point.
- Given an output from an algorithm, an adversary seeks to determine the source of the data.
- The goal of DP is to enhance algorithms to make it more challenging to identify the original data source while ensuring the output remains usable for its original purpose.

Differential Privacy



Differential Privacy

Definition: (ε, δ)-differential privacy (Dwork et al., 2006)

A randomized algorithm M is (ε, δ)-differentially private if for any neighboring data sets $\mathcal{S}$ and $\mathcal{S}'$, and any event $\mathcal{E}$, we have

$$P(M(\mathcal{S}) \in \mathcal{E}) \leq e^{\epsilon} P(M(\mathcal{S}') \in \mathcal{E}) + \delta,$$

where $\epsilon \geq 0$ and $0 \leq \delta \leq 1$ are constants.

Differential Privacy

One way to understand the concept of differential privacy is via the lens of hypothesis testing for distinguishing two neighboring data sets $\mathcal{S}$ and $\mathcal{S}'$:

$$H_0 : \text{the underlying data set is } \mathcal{S}$$

versus

$$H_1 : \text{the underlying data set is } \mathcal{S}'.$$

- It can be shown that for a test ϕ based on the output of any (ϵ, δ)-differentially private algorithm, its power is bounded by $\min\{e^{\epsilon}\alpha_{\phi} + \delta, 1 - e^{-\epsilon}(1 - \alpha_{\phi} - \delta)\}$.
- If ϵ and δ are both small, then any α-level $(0 < \alpha < 1)$ test will be nearly powerless.

f-differential privacy (Dong et al., 2022)

- It is known (ϵ, δ)-differential privacy suffers from the major drawback that it does not tightly handle composition.
- For any two probability distributions P and Q, the **trade-off function** $T(P, Q) : [0, 1] \to [0, 1]$ is defined as $T(P, Q)(\alpha) = \inf\{\beta_\phi : \alpha_\phi \leq \alpha\}$, where the infimum is taken over all measurable rejection rules to distinguish P and Q.
- A randomized algorithm M is said to be f-**differentially private** if for any neighboring data sets $\mathcal{S}$ and $\mathcal{S}'$

$$T(\mathcal{S}, \mathcal{S}') \geq f$$

for some trade-off function f.

Gaussian Differential Privacy

Definition: μ-GDP (Dong et al., 2022)

A mechanism M is said to satisfy μ-**Gaussian Differential Privacy** (μ-GDP) if it is $G_\mu - DP$. That is,

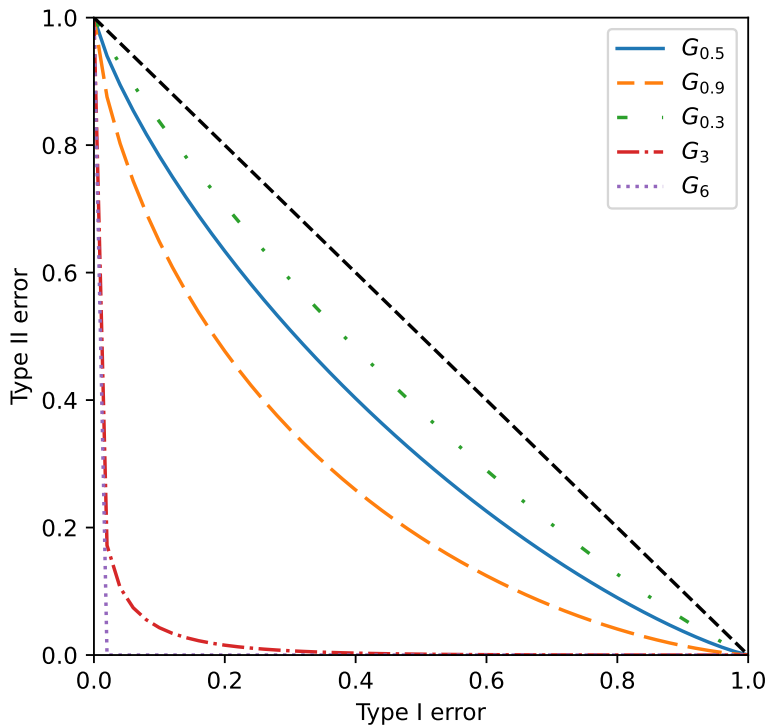
$$T\left(M(S), M\left(S'\right)\right) \geqslant G_\mu$$

for all neighboring datasets S and S', where

$$G_\mu(\alpha) = T(N(0,1), N(\mu,1)) = \Phi\left(\Phi^{-1}(1-\alpha) - \mu\right)(\alpha)$$

Trade-off Function for GDP

- A lower value of μ corresponds to a stronger protection of privacy.
- For GDP, a value of $\mu = 0.5$ indicates a reasonably private scenario, $\mu = 1$ represents a borderline level of privacy.



Quantile Regression (QR)

- **Conditional Linear QR Model** (Koenker, 2005; Koenker and Bassett Jr, 1978)

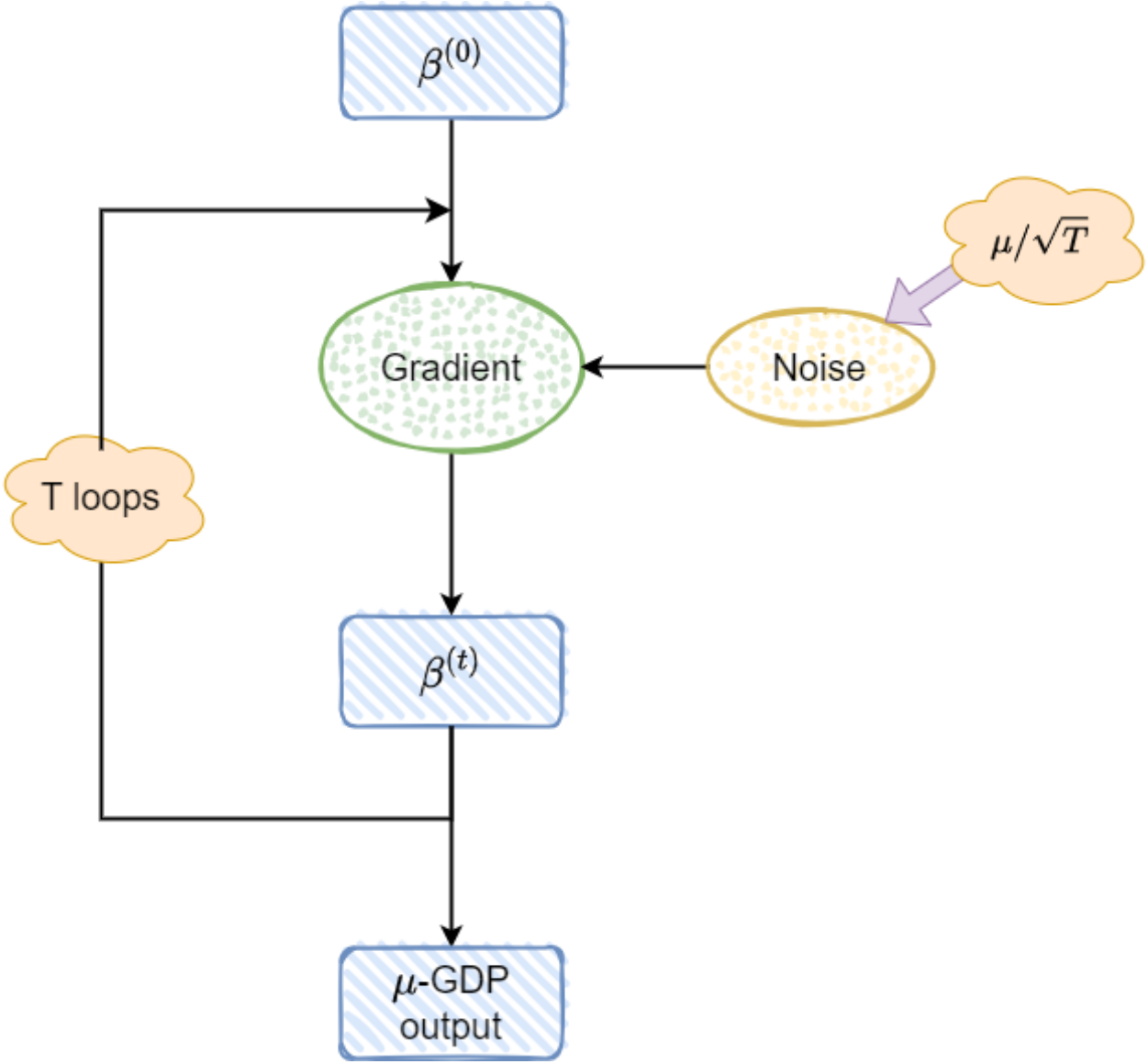
$$F^{-1}_{y|\boldsymbol{x}}(\tau) = \boldsymbol{x}^T \boldsymbol{\beta}^*(\tau).$$

- **QR Estimator:**

$$\widehat{\boldsymbol{\beta}}_\tau = \underset{\boldsymbol{\beta} \in \mathcal{R}^p}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}),$$

where $\rho_\tau(u) = u\{\tau - \mathbb{I}(u < 0)\}$.

Gradient Descent-based Algorithms



Challenges of Differential Privacy for Quantile Regression

- Most of the prior work requires **smoothness** and **convexity** conditions (Avella-Medina et al., 2023; Bassily et al., 2019; Feldman et al., 2020).
- However, the quanitle loss function is **non-smooth** and poses significant challenges to developing a privacy-protection procedure.
- Computational efficiency.

Convolution Smoothing for Quantile Loss

To overcome the non-smooth problem and design an algorithm with privacy protection guarantee, we adopt the following convolution smoothed loss (Fernandes et al., 2021; He et al., 2023; Tan et al., 2022):

$$C_{\varpi}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\left(\rho_{\tau} * K_{\varpi}\right)\left(d_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right),$$

- Closed-form first-order and second-order derivatives.
- Maintain convexity.
- Restricted strong convexity.

Convolution Smoothing for Quantile Loss

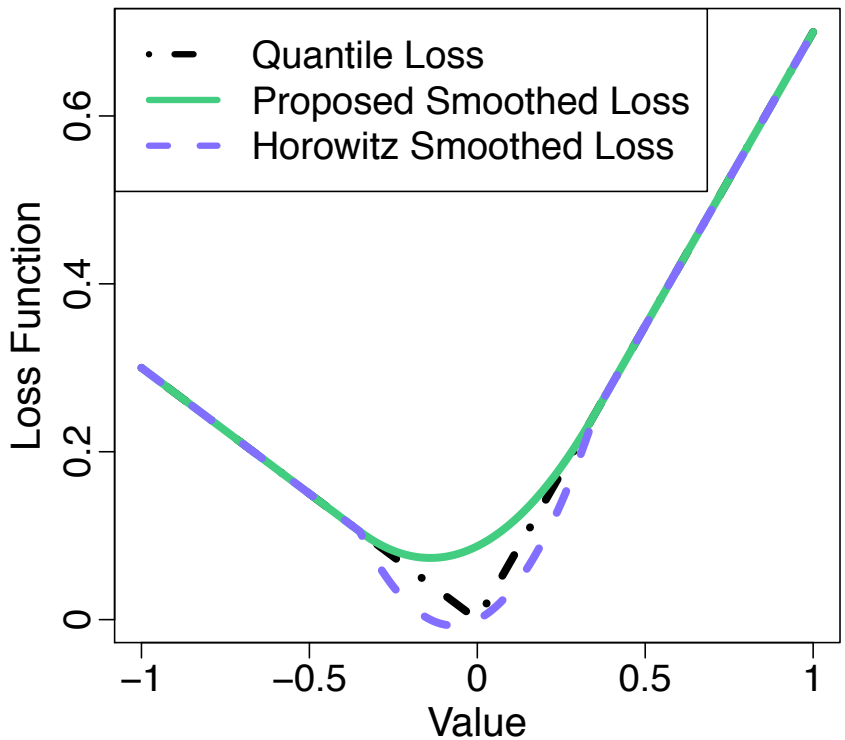


Figure: Comparison of three loss functions

Convolution Smoothing for Quantile Loss

	Convexity	Smoothness
Check loss ρ	✓	×
Convolution smoothed loss	✓	✓
Horowitz smoothed loss	×	✓

A Noisy Clipped Gradient Descent Algorithm for QR

Algorithm 1 Private ERM via Noisy Smoothed Gradient Descent

Input: dataset $\{(d_i, \mathbf{x}_i)\}_{i=1}^n$, probability level $\tau \in (0,1)$, bandwidth $\varpi > 0$, initial value $\boldsymbol{\beta}^{(0)}$, step size $\eta_0 > 0$, noise scale $\sigma > 0$, truncation level $B \geq 1$, number of iterations $T \geq 1$.

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

2: Generate standard multivariate normal vector $\mathbf{g}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$;

3: Compute clipped/truncated covariates $\overline{\mathbf{w}}_i = \mathbf{w}_i \min\{1, B/\|\mathbf{w}_i\|_2\}$ for $i = 1, \ldots, n$;

4: Compute $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - (\eta_0/n) \cdot \Sigma^{-1/2} \big[\sum_{i=1}^n \{\bar{K}_\varpi(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}^{(t)} - d_i) - \tau\} \overline{\mathbf{w}}_i + \sigma \mathbf{g}_t \big]$;

5: **end for**

Output: $\boldsymbol{\beta}^{(T)}$.

Theoretical Performance

Theorem (privacy-protection guarantees)

Given a privacy level $\mu > 0$, if $\sigma > 0$ satisfies $\sigma \geq 2\bar{\tau}BT^{1/2}/\mu$ with $T \asymp \log n$, then the final output $\boldsymbol{\beta}^{(T)}$ of algorithm 1 is μ-GDP.

Theoretical Performance

Theorem (finite-sample estimation error bound)

The μ-GDP estimated coefficient $\boldsymbol{\beta}^{(T)}$ obtained from noisy gradient descent algorithm satisfies

$$\|\boldsymbol{\beta}^{(T)} - \boldsymbol{\beta}^*\|_\Sigma \leq C_0 \left(\eta_0 T^{1/2} \frac{p + \log n}{\mu n} + \frac{1}{f_l} \sqrt{\frac{p \log n}{n}} \right)$$

with probability at least $1 - \frac{C_1}{n^2}$.

Theoretical Performance

- Cost of privacy: $O(\frac{p+\log n}{\mu n})$.
- Statistical estimation error: $O(\sqrt{p\log(n)/n})$.
- Our estimation error is near-optimal (up to a $\log n$ term), as for a non-privacy problem, the optimal estimation error is $O(\sqrt{p/n})$.
- Our upper bound of privacy cost also nearly match the lower bound of $\Omega(1/n)$ for private linear regression setting (up to a $\log n$ term) (Cai et al., 2021).

Theoretical Performance

Theorem (finite-sample expected excess risk bound)

$$\mathbb{E}\{\rho_\tau(y - x^T\beta^{(T)}) - \rho_\tau(y - x^T\beta^*)\} \lesssim \log(n)\left\{\frac{p}{n} + \left(\frac{p + \log n}{\mu n}\right)^2\right\}$$

with probability at least $1 - \frac{C_1}{n^2}$.

Theoretical Performance

Here, we want to highlight that our proposed estimator utilize the specific structure of quantile loss and has a faster rate compared to prior general work for non-smooth loss function. For instance,

- Noisy SGD for non-smooth loss (Bassily et al., 2019):

$$O\left(\sqrt{\frac{p}{n}} + \sqrt{\log(1/\delta)}\frac{p}{\epsilon n}\right).$$

- Our algorithm:

$$O\left(\log(n)\left\{\frac{p}{n} + \left(\frac{p+\log n}{\mu n}\right)^2\right\}\right)$$

Synthetic Data

Settings

- Consider the linear model $y = 1.5 + \boldsymbol{x}^T\boldsymbol{\theta}^* + \varepsilon$.
- $\boldsymbol{\theta}^* = (1, -2.5, -1.5, 3)^T$.
- The feature vector $\boldsymbol{z} \in \mathbb{R}^4$ is generated from a centered multivariate normal distribution with covariance matrix $\Sigma = (0.5^{|j-k|})_{1\leq j,k\leq 4}$.
- The observation noise variable ε follows Gaussian mixture distribution $0.9N(0,1) + 0.1N(0,100)$.

Synthetic Data

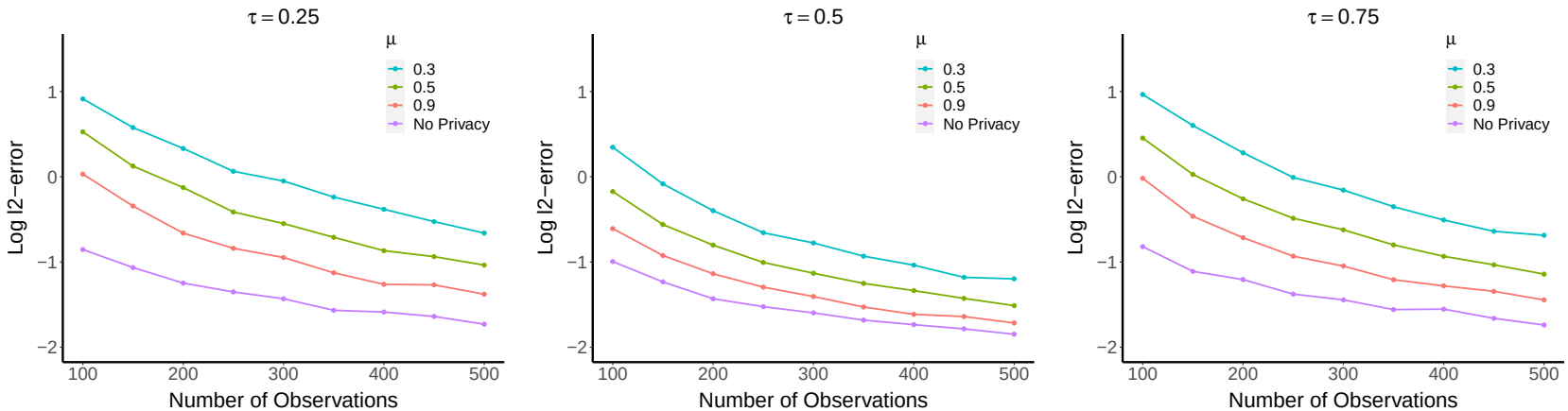


Figure: Plots of logarithmic ℓ_2 estimation error versus the number of observations, averaged over 300 replications, when $\varepsilon \sim 0.9N(0,1) + 0.1N(0,100)$.

Synthetic Data

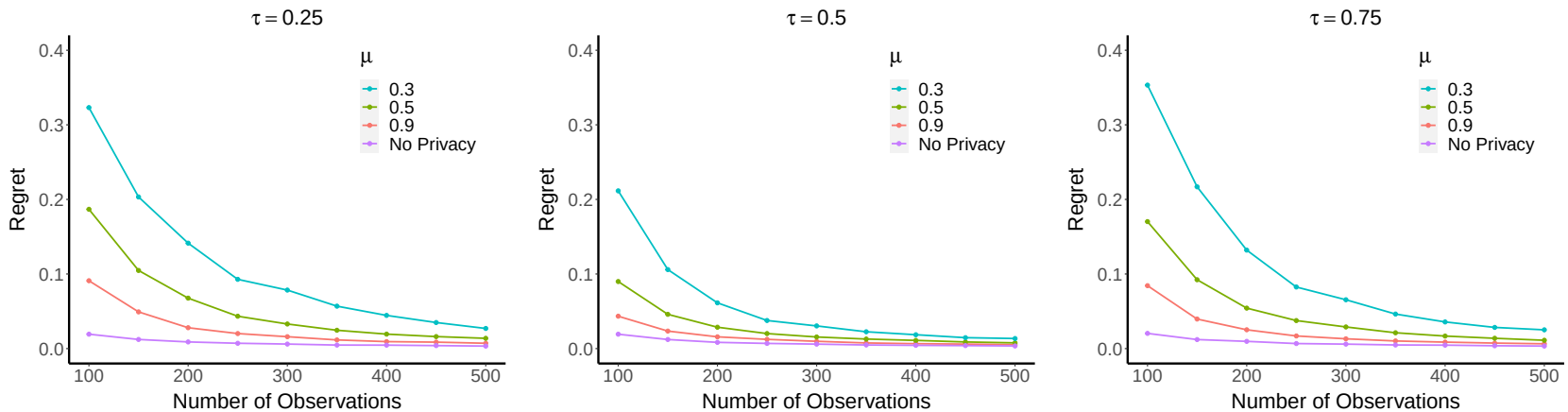


Figure: Plots of regret versus the number of observations, averaged over 300 replications, when $\varepsilon \sim 0.9N(0,1) + 0.1N(0,100)$.

Synthetic Data

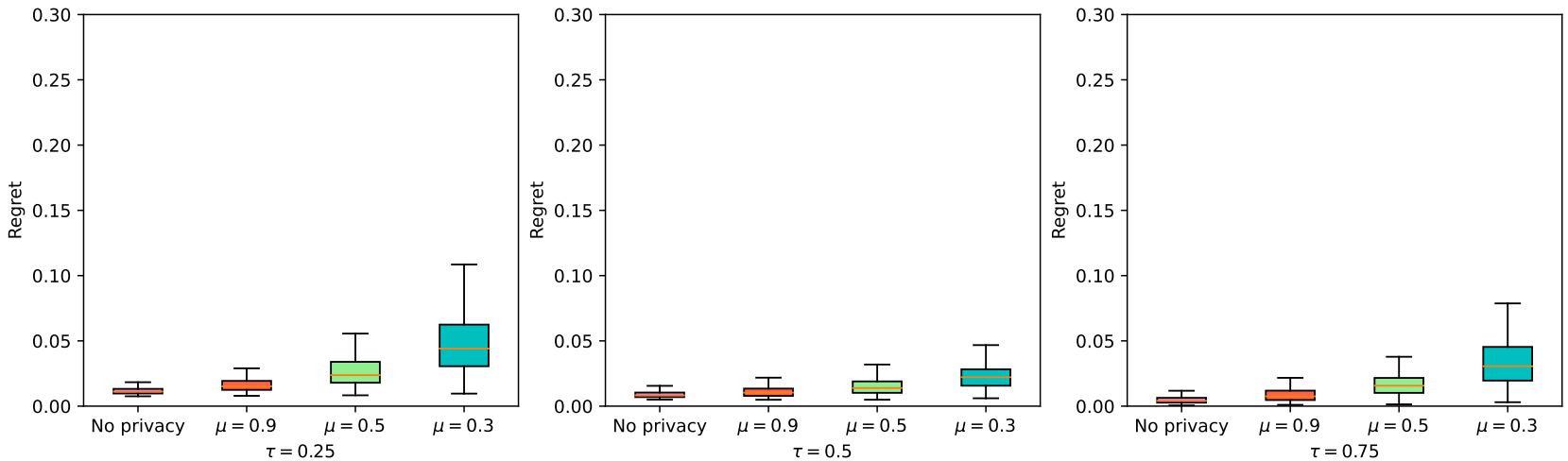


Figure: Box-plots of regret with different privacy level over 300 replications when sample size equals 400.

Conclusions

- We develop a clipped noisy gradient descent algorithm based on convolution smoothing for quantile regression.
- We derive finite-sample high-probability bounds for optimal parameter estimation and the excess population loss.
- By exploring the specific structure of QR, we achieve a faster rate compared to the results obtained by applying existing techniques developed for general non-smooth convex loss.
- Acknowledgment: NSF FRGMS-1952373.

References I

Marco Avella-Medina, Casey Bradshaw, and Po-Ling Loh. Differentially private inference via noisy optimization. *arXiv preprint arXiv:2103.11003*, 2023.

Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

T Tony Cai, Yichen Wang, and Linjun Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics*, 49(5):2825–2850, 2021.

Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. *Journal of the Royal Statistical Society: Series B (Methodological)*, 84(1):3–37, 2022.

References II

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Shai Halevi and Tal Rabin, editors, *Theory of Cryptography*, pages 265–284, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 439–449, 2020.

Marcelo Fernandes, Emmanuel Guerre, and Eduardo Horta. Smoothing quantile regressions. *Journal of Business & Economic Statistics*, 39(1): 338–357, 2021.

Xuming He, Xiaoou Pan, Kean Ming Tan, and Wen-Xin Zhou. Smoothed quantile regression with large-scale inference. *Journal of Econometrics*, 232(2):367–388, 2023.

Roger Koenker. *Quantile regression*. Cambridge University Press, 2005.

References III

Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica*, pages 33–50, 1978.

Kean Ming Tan, Lan Wang, and Wen-Xin Zhou. High-dimensional quantile regression: Convolution smoothing and concave regularization. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):205–233, 2022.