PhD Defense
On the tradeoffs of statistical learning with privacy

Clément Lalanne

École Normale Supérieure de Lyon
Ockham - LIP

clement.lalanne@ens-lyon.fr

2023-10-04

Jury Members

- **Aurélien Bellet**, DR, Inria & Université de Montpellier (Rapporteur)
- **Béatrice Laurent-Bonneau**, PR, INSA Toulouse (Rapporteuse)
- **Élisa Fromont**, PR, Université de Rennes (Examinatrice)
- **Aurélien Garivier**, PR, ENS de Lyon (Directeur de thèse)
- **Rémi Gribonval**, DR, Lyon & ENS de Lyon (Co-directeur de thèse)

Table of Contents

4 Context

Increasing data usage :

- Natural Language Processing (ChatGPT, ...)
- Medical applications
- ...

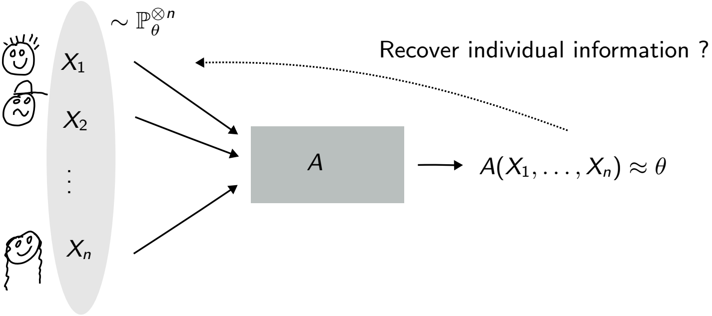
Observation :

Most applications are not interested in the dataset itself, but rather in some quantities defined at the scale of the population.

Threats to privacy :

Incautious use of data can leak personal information.

5 Context

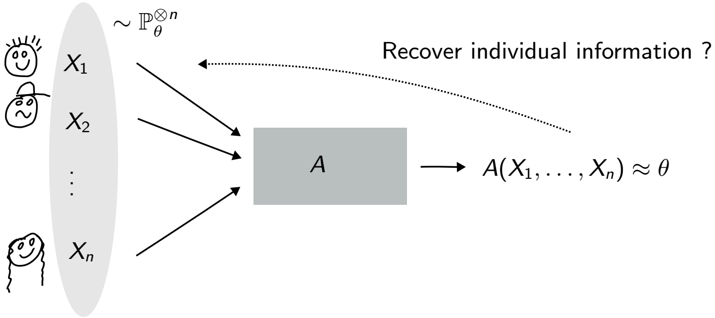


Examples :

- Bernoulli estimation (proportion)
- Complex distributions
- ...

[1] Antoine Gonon et al. *Sparsity in neural networks can improve their privacy*. 2023. arXiv: 2304.10553 [cs.LG].

6 Context



Question :

Is it possible to defend against ***ANY*** *attack ?*

---

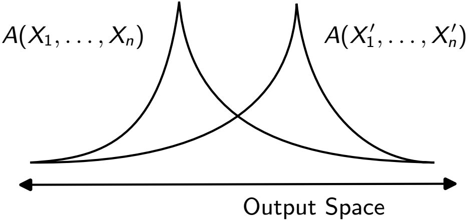
[2]Gonon et al., *Sparsity in neural networks can improve their privacy*.

7 Differential Privacy

Neighboring relation : $\mathbf{X} \sim \mathbf{X}'$ iff $\mathbf{X}$ can be obtained from $\mathbf{X}'$ by changing the data of one individual.

Definition (informal) : A is $\epsilon > 0$-DP if $A(\mathbf{X})$ is ϵ-close to $A(\mathbf{X}')$ for any $\mathbf{X} \sim \mathbf{X}'$.[3]

Impact of ϵ : The smaller ϵ, the more privacy.



Question :

How does privacy affect classical statistical estimation results ?

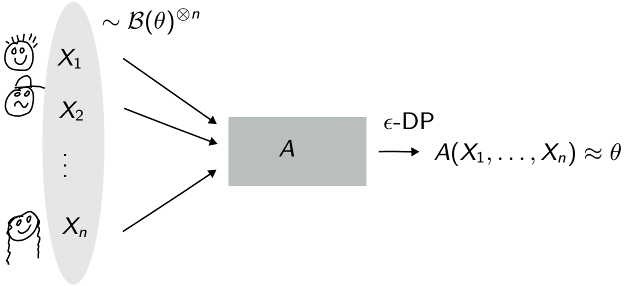
[3] Cynthia Dwork et al. "Calibrating Noise to Sensitivity in Private Data Analysis". In: 2006.

Table of Contents

9 Bernoulli estimation setup

Setup : $\theta \in [0,1]$, $\mathbf{X} = (X_1, \ldots, X_n) \sim \mathcal{B}(\theta)^{\otimes n}$.

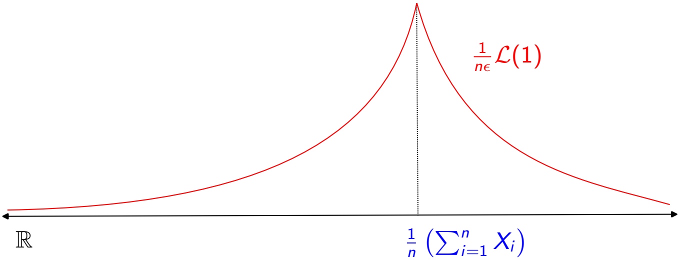
Measure of performance : $\mathbb{E}_{A,\mathbf{X}}\left((A(\mathbf{X}) - \theta))^2\right)$.



10 Private estimator

Laplace mechanism :

$$A(\mathbf{X}) = \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right) + \frac{1}{n\epsilon}\mathcal{L}(1)$$



Laplace mechanism :

$$A(\mathbf{X}) = \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right) + \frac{1}{n\epsilon}\mathcal{L}(1)$$

is ϵ-DP.

Error :

$$\mathbb{E}_{\mathbb{P}_A, \mathcal{B}(\theta)^{\otimes n}}\left(\left(A(\mathbf{X}) - \theta\right)\right)^2\Big) \leq \frac{1/4}{n} + \frac{2}{n^2\epsilon^2}$$

Two regimes :

- **Low privacy regime :** $\epsilon = \Omega(1/\sqrt{n})$, no significant effect on estimation.
- **High privacy regime :** $\epsilon \ll 1/\sqrt{n}$, the precision can be arbitrarily degraded.

Question :

Is it possible to do better ?

Table of Contents

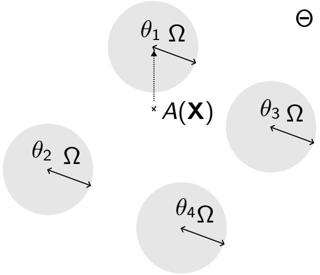
13 Minimax risk and reduction to hypothesis testing

Setup : $\theta \in \Theta$, $\mathbf{X} = (X_1, \dots, X_n) \sim \mathbb{P}_\theta^{\otimes n}$.

Minimax risk :[4][5]

$$\boxed{\mathfrak{M}_n := \inf_A \sup_{\theta\in\Theta} \mathbb{E}_{A,\mathbf{X}\sim\theta}\left(\mathrm{Error}(A(\mathbf{X}),\theta)\right)} \geq \inf_A \boxed{\sup_{i=1,\dots,N} \mathbb{E}_{A,\mathbf{X}\sim\theta_i}\left(\mathrm{Error}(A(\mathbf{X}),\theta_i)\right)}$$

$$\geq \Phi(\Omega) \boxed{\sup_{i=1,\dots,N} \mathbb{P}_{A,\mathbf{X}\sim\theta_i}\left(\hat{i}\,(A(\mathbf{X})) \neq i\right)}$$



[4] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. 2009.

[5] $\mathrm{Error}(\cdot,\cdot) = \Phi(d(\cdot,\cdot))$

14 Without taking privacy into account

Setup : $\theta \in \Theta$, $\mathbf{X} = (X_1, \ldots, X_n) \sim \mathbb{P}_\theta^{\otimes n}$.

Minimax risk :

$$\begin{aligned}\mathfrak{M}_n := \inf_A \sup_{\theta \in \Theta} \mathbb{E}_{A, \mathbf{X} \sim \theta}\left(\mathrm{Error}(A(\mathbf{X}), \theta)\right) &\geq \inf_A \sup_{i=1,\ldots,N} \mathbb{E}_{A, \mathbf{X} \sim \theta_i}\left(\mathrm{Error}(A(\mathbf{X}), \theta_i)\right) \\ &\geq \Phi(\Omega) \sup_{i=1,\ldots,N} \mathbb{P}_{A, \mathbf{X} \sim \theta_i}\left(\hat{i}(A(\mathbf{X})) \neq i\right)\end{aligned}$$

Le Cam's lemma :[6]

$$\boxed{\sup_{i=1,2} \mathbb{P}_{A, \mathbf{X} \sim \theta_i}\left(\hat{i}(A(\mathbf{X})) \neq i\right) \geq \frac{1}{2}\left(1 - \mathrm{TV}\left(\mathbb{P}_{\theta_1}^{\otimes n}, \mathbb{P}_{\theta_2}^{\otimes n}\right)\right)}\,.$$

Fano's lemma :[7] For any dominating measure $\mathbb{Q}$,

$$\boxed{\sup_{i=1,\ldots,N} \mathbb{P}_{A, \mathbf{X} \sim \theta_i}\left(\hat{i}(A(\mathbf{X})) \neq i\right) \geq 1 - \frac{1 + \frac{1}{N}\sum_{i=1}^N \mathrm{KL}\left(\mathbb{P}_{\theta_i}^{\otimes n} \middle\| \mathbb{Q}\right)}{\ln(N)}}\,.$$

[6] $\mathrm{TV}\left(\mathbb{P}_1, \mathbb{P}_2\right) := \sup_S \mathbb{P}_1(S) - \mathbb{P}_2(S)$

[7] $\mathrm{KL}\left(\mathbb{P}_1 \| \mathbb{P}_2\right) := \int \ln\left(\frac{d\mathbb{P}_1}{d\mathbb{P}_2}\right) d\mathbb{P}_1$

Without taking privacy into account

Setup : $\theta \in \Theta$, $\mathbf{X} = (X_1, \ldots, X_n) \sim \mathbb{P}_\theta^{\otimes n}$.

Minimax risk :

$$\mathfrak{M}_n := \inf_{A} \sup_{\theta \in \Theta} \mathbb{E}_{A, \mathbf{X} \sim \theta} \left(\mathrm{Error}(A(\mathbf{X}), \theta) \right) \geq \inf_{A} \sup_{i=1,\ldots,N} \mathbb{E}_{A, \mathbf{X} \sim \theta_i} \left(\mathrm{Error}(A(\mathbf{X}), \theta_i) \right)$$

$$\geq \Phi(\Omega) \sup_{i=1,\ldots,N} \mathbb{P}_{A, \mathbf{X} \sim \theta_i} \left(\hat{i}\,(A(\mathbf{X})) \neq i \right)$$

Question :

Is it possible to obtain similar lower-bounds that take privacy into consideration ?

16 Reduction to a transport problem

Setup : $\theta \in \Theta$, $\mathbf{X} = (X_1, \dots, X_n) \sim \mathbb{P}_\theta^{\otimes n}$.

Minimax risk :

$$\mathfrak{M}_n := \inf_A \sup_{\theta \in \Theta} \mathbb{E}_{A, \mathbf{X} \sim \theta} \left(\mathrm{Error}(A(\mathbf{X}), \theta) \right) \geq \inf_A \sup_{i=1,\dots,N} \mathbb{E}_{A, \mathbf{X} \sim \theta_i} \left(\mathrm{Error}(A(\mathbf{X}), \theta_i) \right)$$

$$\geq \Phi(\Omega) \sup_{i=1,\dots,N} \mathbb{P}_{A, \mathbf{X} \sim \theta_i} \left(\hat{i}\left(A(\mathbf{X})\right) \neq i \right)$$

Reduction to a transport problem :[8]

$$\sup_{i=1,\dots,N} \mathbb{P}_{A, \mathbf{X} \sim \theta_i} \left(\hat{i}\left(A(\mathbf{X})\right) \neq i \right) \geq \boxed{\sup_{\mathbb{Q} \in \Pi\left(\mathbb{P}_1^{\otimes n}, \dots, \mathbb{P}_N^{\otimes n}\right)} \int s\left(\mathbf{X}_1, \dots, \mathbf{X}_N\right) d\mathbb{Q}\left(\mathbf{X}_1, \dots, \mathbf{X}_N\right)},$$

where s is a *similarity function* satisfying, for any $\mathbf{X}_1, \dots, \mathbf{X}_N$,

$$\boxed{\frac{1}{N} \sum_{i=1}^{N} \mathbb{P}_A \left(\hat{i}\left(A\left(\mathbf{X}_i\right)\right) \neq i \right) \geq s\left(\mathbf{X}_1, \dots, \mathbf{X}_N\right)}.$$

[8] Clément Lalanne, Aurélien Garivier, and Rémi Gribonval. "On the Statistical Complexity of Estimation and Testing under Privacy Constraints". In: (2023).

17 Building similarity functions

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{P}_A\left(\hat{i}\left(A\left(\mathbf{X}_i\right)\right)\neq i\right)\geq s\left(\mathbf{X}_1,\ldots,\mathbf{X}_N\right).$$

Definition : A is ϵ-DP if $\mathbf{X}\sim\mathbf{X}' \implies \mathbb{P}\big(A(\mathbf{X})\in S\big)\leq e^{\epsilon}\times\mathbb{P}\big(A(\mathbf{X}')\in S\big)$.[9]

Two marginals : $\boxed{\frac{1}{2}\sum_{i=1}^{2}\mathbb{P}_A\left(\hat{i}\left(A\left(\mathbf{X}_i\right)\right)\neq i\right)\geq\frac{1}{2}e^{-\epsilon d_{\mathrm{ham}}(\mathbf{X}_1,\mathbf{X}_2)}}$[10]

Many marginals : $\boxed{\frac{1}{N}\sum_{i=1}^{N}\mathbb{P}_A\left(\hat{i}\left(A\left(\mathbf{X}_i\right)\right)\neq i\right)\geq 1-\frac{1+\frac{\epsilon}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}d_{\mathrm{ham}}\left(\mathbf{X}_i,\mathbf{X}_j\right)}{\ln(N)}}$

9 Dwork et al., "Calibrating Noise to Sensitivity in Private Data Analysis".

10 $d_{\mathrm{ham}}\left((X_1,\ldots,X_n),(X_1',\ldots,X_n')\right):=\sum_{i=1}^{n}\mathbb{1}_{X_i=X_i'}$

18 Back to the transport problem

$$\boxed{\sup_{\mathbb{Q}\in\Pi\left(\mathbb{P}_1^{\otimes n},\dots,\mathbb{P}_N^{\otimes n}\right)} \int s\left(\mathbf{X}_1,\dots,\mathbf{X}_N\right) d\mathbb{Q}\left(\mathbf{X}_1,\dots,\mathbf{X}_N\right)},$$

where s is **non-increasing** *in* $d_{\mathrm{ham}}\left(\mathbf{X}_i,\mathbf{X}_j\right)$ *for any* i,j.

Question :

How to construct a coupling that makes those quantities big ?

19 A good enough coupling

$$\sup_{\mathbb{Q}\in\Pi\left(\mathbb{P}_1^{\otimes n},\dots,\mathbb{P}_N^{\otimes n}\right)} \int s\left(\mathbf{X}_1,\dots,\mathbf{X}_N\right) d\mathbb{Q}\left(\mathbf{X}_1,\dots,\mathbf{X}_N\right),$$

where s is **non-increasing** *in* $d_{\mathrm{ham}}\left(\mathbf{X}_i,\mathbf{X}_j\right)$ *for any* i,j.

Near optimal coupling for equalities : There exists $(X_i)_{i=1,\dots,N}$ of distribution χ, a coupling between $(\mathbb{P}_i)_{i=1,\dots,N}$ such that[11]

$$\forall i,j,\quad \mathrm{TV}\left(\mathbb{P}_i,\mathbb{P}_j\right) \le \boxed{\mathbb{P}(X_i \neq X_j) \le \frac{2\mathrm{TV}\left(\mathbb{P}_i,\mathbb{P}_j\right)}{1+\mathrm{TV}\left(\mathbb{P}_i,\mathbb{P}_j\right)}}.$$

Final coupling : $\boxed{\mathbb{Q}^* = \chi^{\otimes n}}$

[11] Omer Angel and Yinon Spinka. *Pairwise optimal coupling of multiple random variables*. 2021.

20 Final results

Setup : $\theta \in \Theta$, $\mathbf{X} = (X_1, \ldots, X_n) \sim \mathbb{P}_\theta^{\otimes n}$.

Minimax risk :

$$\mathfrak{M}_n := \inf_A \sup_{\theta \in \Theta} \mathbb{E}_{A, \mathbf{X} \sim \theta} \left(\mathrm{Error}(A(\mathbf{X}), \theta) \right) \geq \inf_A \sup_{i=1,\ldots,N} \mathbb{E}_{A, \mathbf{X} \sim \theta_i} \left(\mathrm{Error}(A(\mathbf{X}), \theta_i) \right)$$

$$\geq \Phi(\Omega) \sup_{i=1,\ldots,N} \mathbb{P}_{A, \mathbf{X} \sim \theta_i} \left(\hat{i}(A(\mathbf{X})) \neq i \right)$$

Private Le Cam's lemma :[12]

$$\boxed{\sup_{i=1,2} \mathbb{P}_{A, \mathbf{X} \sim \theta_i} \left(\hat{i}(A(\mathbf{X})) \neq i \right) \geq \frac{1}{2} \left(1 - \left(1 - e^{-\epsilon} \right) \mathrm{TV}\left(\mathbb{P}_{\theta_1}, \mathbb{P}_{\theta_2} \right) \right)^n}.$$

Private Fano's lemma :[13] For any dominating measure $\mathbb{Q}$,

$$\boxed{\sup_{i=1,\ldots,N} \mathbb{P}_{A, \mathbf{X} \sim \theta_i} \left(\hat{i}(A(\mathbf{X})) \neq i \right) \geq 1 - \frac{1 + \frac{n\epsilon}{N^2} \sum_{i,j=1}^{N} \frac{2\mathrm{TV}\left(\mathbb{P}_{\theta_i}, \mathbb{P}_{\theta_j}\right)}{1 + \mathrm{TV}\left(\mathbb{P}_{\theta_i}, \mathbb{P}_{\theta_j}\right)}}{\ln(N)}}.$$

[12] $\mathrm{TV}\left(\mathbb{P}_1, \mathbb{P}_2\right) := \sup_S \mathbb{P}_1(S) - \mathbb{P}_2(S)$

[13] $\mathrm{KL}\left(\mathbb{P}_1 \| \mathbb{P}_2\right) := \int \ln\left(\frac{d\mathbb{P}_1}{d\mathbb{P}_2}\right) d\mathbb{P}_1$

21 Illustration

$$\inf_{A}\sup_{\theta\in\Theta}\mathbb{E}_{A,\mathbf{X}\sim\theta}\left(\mathrm{Error}(A(\mathbf{X}),\theta)\right) \geq \Phi(\Omega)\left(1-\frac{1+\frac{n\epsilon}{N^2}\sum_{i,j=1}^{N}\frac{2\mathrm{TV}\left(\mathbb{P}_{\theta_i},\mathbb{P}_{\theta_j}\right)}{1+\mathrm{TV}\left(\mathbb{P}_{\theta},\mathbb{P}_{\theta_j}\right)}}{\ln(N)}\right)$$

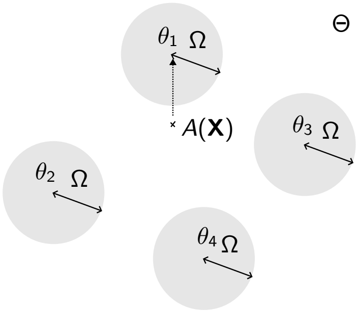
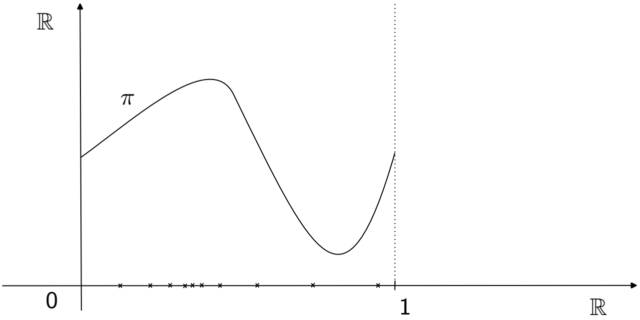


Table of Contents

23 Density Estimation Problem

Setup : $\mathbf{X} = (X_1, \ldots, X_n) \sim \mathbb{P}_\pi^{\otimes n}$, π a **density of probability** with respect to Lebesgue's measure on $[0, 1]$.

Measure of performance : $\mathbb{E}\left(\|A(X) - \pi\|_{L_2}^2\right)$.



24 L_2 approximations and projection estimators

Reference Fourier basis :

$$\begin{aligned}\phi_1(x) &= 1\\ \phi_{2k}(x) &= \sqrt{2}\sin(2\pi kx) \quad k\geq 1\\ \phi_{2k+1}(x) &= \sqrt{2}\cos(2\pi kx) \quad k \geq 1\ .\end{aligned}$$

L_2 **approximation :**

$$\sum_{i=1}^{N}\theta_i\phi_i \xrightarrow[N\to+\infty]{L^2} \pi \quad \text{where} \quad \theta_i := \int_{[0,1]} \pi\ \phi_i\ .$$

Projection estimator :[14]

$$\hat{\pi}^{\text{proj}}(\mathbf{X}) = \sum_{i=1}^{N}\hat{\theta}_i\phi_i \quad \text{where} \quad \hat{\theta}_i := \frac{1}{n}\sum_{j=1}^{n}\phi_i(X_j)\ .$$

Question : *How do we add privacy ?*

[14] Tsybakov, *Introduction to Nonparametric Estimation*.

Private projection estimator :[15]

$$\boxed{\hat{\pi}^{\text{proj}}(\mathbf{X}) = \sum_{i=1}^{N} \left(\hat{\theta}_i + C_{\epsilon,N}\mathcal{L}(1)\right)\phi_i \quad \text{where} \quad \hat{\theta}_i := \frac{1}{n}\sum_{j=1}^{n}\phi_i(X_j)}\,.$$

$\hat{\pi}^{\text{proj}}$ is ϵ-DP.

Question :

What is the utility (error) of this estimator ?

[15] Larry A. Wasserman and Shuheng Zhou. "A Statistical Framework for Differential Privacy". In: (2010).

26 Sobolev spaces and approximation speed

$$\hat{\pi}^{\text{proj}}(\mathbf{X}) = \sum_{i=1}^{N} \left(\hat{\theta}_i + C_{\epsilon,N}\mathcal{L}(1) \right) \phi_i \quad \text{where} \quad \hat{\theta}_i := \frac{1}{n} \sum_{j=1}^{n} \phi_i(X_j) \; .$$

Sobolev spaces :

$$\Theta_{L,\beta}^{\text{PSob}} := \left\{ \pi \;\middle|\; \int_{[0,1]} \left(\pi^{(\beta)} \right)^2 \leq L^2 \text{ and a few other minor hypotheses} \right\} . \tag{1}$$

Approximation speed :[16] When $\pi \in \Theta_{L,\beta}^{\text{PSob}}$,

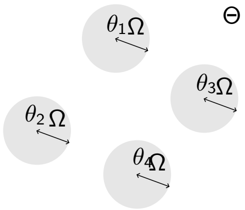
$$\boxed{\mathbb{E}\left(\|\hat{\pi}^{\text{proj}}(\mathbf{X}) - \pi\|_{L_2}^2 \right) \leq C_{L,\beta} \max\left\{ n^{-\frac{2\beta}{2\beta+1}}, (n\epsilon)^{-\frac{2\beta}{\beta+3/2}} \right\}}$$

Question :

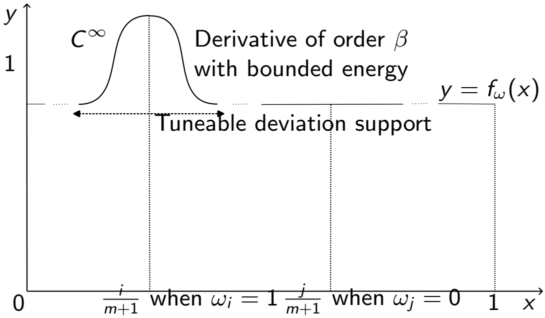
What about lower-bounds ?

[16] Clément Lalanne, Aurélien Garivier, and Rémi Gribonval. "About the Cost of Central Privacy in Density Estimation". In: (2023).

27 Lower-bounds



A packing of densities $f_{\omega_1}, f_{\omega_2}, \ldots$ where for any $\omega \in \{0,1\}^m$,



Lower-bound against ϵ-DP estimators :[17]

$$\boxed{\inf_A \sup_\pi \mathbb{E}\left(\|A(\mathbf{X}) - \pi\|_{L_2}^2\right) \geq C_{L,\beta} \max\left\{n^{-\frac{2\beta}{2\beta+1}}, (n\epsilon)^{-\frac{2\beta}{\beta+1}}\right\}}$$

Best known upper-bound :

$$\mathbb{E}\left(\|\hat{\pi}^{\text{proj}}(\mathbf{X}) - \pi\|_{L_2}^2\right) \leq C_{L,\beta} \max\left\{n^{-\frac{2\beta}{2\beta+1}}, (n\epsilon)^{-\frac{2\beta}{\beta+3/2}}\right\}$$

Question :

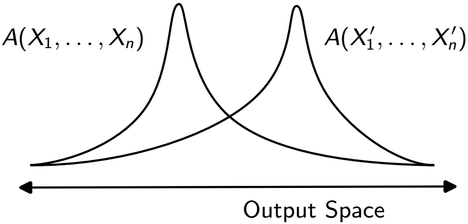
Is it possible to bridge the gap ?

[17] Lalanne, Garivier, and Gribonval, "About the Cost of Central Privacy in Density Estimation".

29 Concentrated Differential Privacy

Definition :[18,19] A is ρ-zCDP if $\mathbf{X} \sim \mathbf{X}' \implies \forall \alpha > 0, \quad \mathrm{D}_\alpha\left(A(\mathbf{X}) \| A(\mathbf{X}')\right) \le \alpha\rho$, where

$$\mathrm{D}_\alpha\left(\mathbb{P} \| \mathbb{Q}\right) := \frac{1}{\alpha - 1} \ln \int \left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right)^{\alpha-1} d\mathbb{Q} \, .$$



[18] Cynthia Dwork and Guy N Rothblum. "Concentrated differential privacy". In: (2016).

[19] Mark Bun and Thomas Steinke. "Concentrated Differential Privacy: Simplifications, Extensions, and Lower Bounds". In: 2016.

Private projection estimator :

$$\boxed{\hat{\pi}^{\text{proj}}(\mathbf{X}) = \sum_{i=1}^{N} \left(\hat{\theta}_i + C_{\rho,N}\mathcal{N}(0,1)\right)\phi_i \quad \text{where} \quad \hat{\theta}_i := \frac{1}{n}\sum_{j=1}^{n}\phi_i(X_j)}\,.$$

$\hat{\pi}^{\text{proj}}$ is ρ-zCDP.

Resulting upper-bound :

$$\boxed{\mathbb{E}\left(\|\hat{\pi}^{\text{proj}}(\mathbf{X}) - \pi\|_{L_2}^2\right) \leq C_{L,\beta} \max\left\{n^{-\frac{2\beta}{2\beta+1}}, (n\sqrt{\rho})^{-\frac{2\beta}{\beta+1}}\right\}}$$

Lower-bound against ρ-zCDP estimators :

$$\boxed{\inf_{A}\sup_{\pi}\mathbb{E}\left(\|A(\mathbf{X}) - \pi\|_{L_2}^2\right) \geq C_{L,\beta} \max\left\{n^{-\frac{2\beta}{2\beta+1}}, (n\sqrt{\rho})^{-\frac{2\beta}{\beta+1}}\right\}}$$

Table of Contents

32 Joint work with

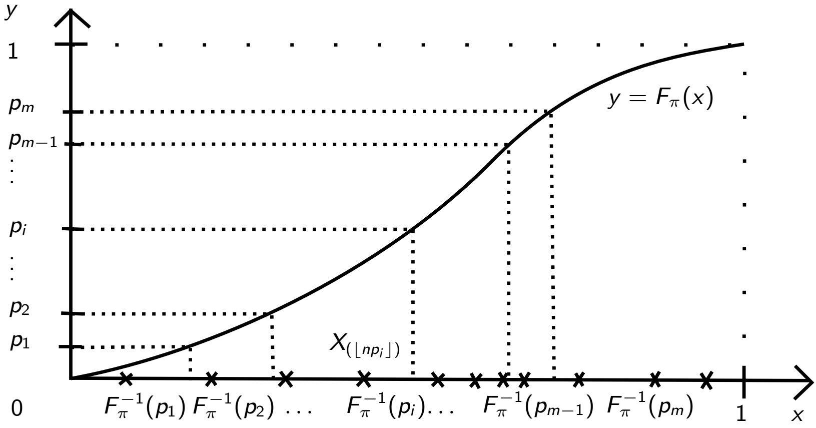
Nicolas Grislain and Clément Gastaud from Sarus Technologies[20]


[20] Clément Lalanne et al. "Private quantiles estimation in the presence of atoms". In: (2023). ISSN: 2049-8772.

33 Quantiles Estimation Problem

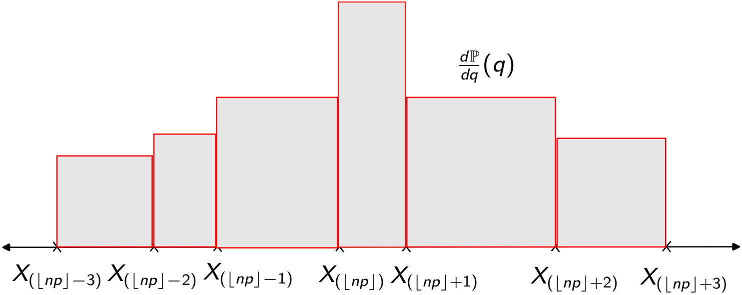
Inputs : samples $\mathbf{X} = (X_1, \ldots, X_n) \overset{\text{i.i.d.}}{\sim} \mathbb{P}_\pi$ $\mathbf{p} = (p_1, \ldots, p_m) \in (0,1)^m$ sorted.

Desired output : Quantile estimator $\mathbf{q} \in [0,1]^m$ of $(F_\pi^{-1}(p_1), \ldots, F_\pi^{-1}(p_m))$.



34 Private Exponential Quantiles

Mechanism :[21] For a single quantile q (associated with p),



Concentration result :[23] When π is away from 0 on a neighborhood of $F_\pi^{-1}(p)$,

$$\mathbb{P}\left(|q - F_\pi^{-1}(p)| > \gamma\right) \leq P(n) \max\left(e^{-C_1 \epsilon n \gamma}, e^{-C_2 \gamma^2 n}\right).$$

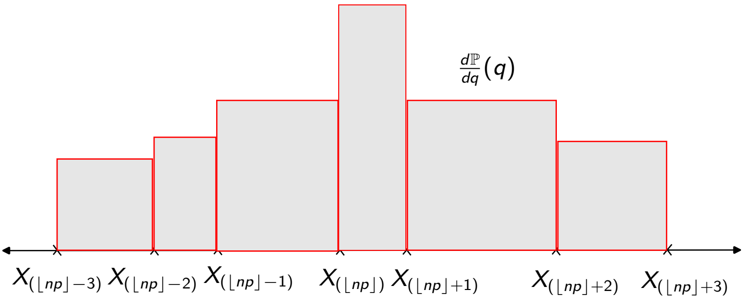
[21] Adam D. Smith. "Privacy-preserving statistical estimation with optimal convergence rates". In: 2011.

[22] $d\mathbb{P}(q) \propto e^{-\frac{\epsilon}{2}\left| |\{i | X_i < q\}| - \lfloor np \rfloor \right|} dq$

[23] Clément Lalanne, Aurélien Garivier, and Rémi Gribonval. "Private Statistical Estimation of Many Quantiles". In: 2023.

Independent Private Quantiles

Idea : Use QExp independently on **p** with simple composition.



Concentration result :[25] When π is away from 0 on a neighborhood of $F_\pi^{-1}(\mathbf{p})$,

$$\mathbb{P}\left(\|\mathbf{q}-F_\pi^{-1}(\mathbf{p})\|_\infty > \gamma\right) \leq P(n,m)\max\left(e^{-C_1\frac{\epsilon n\gamma}{\mathbf{m}}}, e^{-C_2\gamma^2 n}\right).$$

[24] $d\mathbb{P}(q_i) \propto e^{-\frac{\epsilon}{2m}\left||\{i|X_i<q_i\}|-\lfloor np_i\rfloor\right|}dq_i$

[25] Lalanne, Garivier, and Gribonval, "Private Statistical Estimation of Many Quantiles".

 Joint Exponential Private Quantiles

Idea :[26] Leverage structural dependencies. Quantiles are non-decreasing, between q_i and q_j should fall approximately $n|p_i - p_j|$ points. [27]

"Fun" discovery :[28] JointExp $\approx$ Inverse Sensitivity Mechanism[29].

Consistency result : When π is away from 0 on a neighborhood of $F_\pi^{-1}(\mathbf{p})$, JointExp is consistent.

[26] Jennifer Gillenwater, Matthew Joseph, and Alex Kulesza. "Differentially Private Quantiles". In: 2021.

[27] $d\mathbb{P}(\mathbf{q}) \propto e^{-\frac{\epsilon}{2}\sum_{i=1}^{m+1}\left|\delta^{\mathrm{JE}}(i,\mathbf{X},\mathbf{q})\right|} d\mathbf{q}$ where $\delta^{\mathrm{JE}}(i, \mathbf{X}, \mathbf{q}) := n(p_i - p_{i-1}) - \#\left(\mathbf{X} \cap (q_{i-1}, q_i]\right)$

[28] Lalanne et al., "Private quantiles estimation in the presence of atoms".

[29] Hilal Asi and John C. Duchi. "Near Instance-Optimality in Differential Privacy". In: *CoRR* (2020). arXiv: 2005.10630.

37 Recursive Private Quantiles

Idea :[30] Use QExp recursively with a dichotomy on $\mathbf{p}$.

Concentration result :[31] When π is away from 0 on a neighborhood of $F_\pi^{-1}(\mathbf{p})$,

$$\mathbb{P}\left(\|\mathbf{q} - F_\pi^{-1}(\mathbf{p})\|_\infty > \gamma\right) \leq P(n, m) \max\left(e^{-C_1 \frac{\epsilon n \gamma}{(\log_2(\mathbf{m}))^2}}, e^{-C_2 \gamma^2 n}\right).$$

Remark : Almost polylogarithmic degradation in m !

[30] Haim Kaplan, Shachar Schnapp, and Uri Stemmer. "Differentially Private Approximate Quantiles". In: ed. by Kamalika Chaudhuri et al. PMLR, 2022.

[31] Lalanne, Garivier, and Gribonval, "Private Statistical Estimation of Many Quantiles".

Idea :[32]

$$\hat{\pi}^{\text{hist}}(t) := \sum_{b \in \text{bins}} \mathbb{1}_b(t) \frac{1}{nh} \left(\sum_{i=1}^{n} \mathbb{1}_b(X_i) + \frac{2}{\epsilon} \mathcal{L}_b \right) .$$

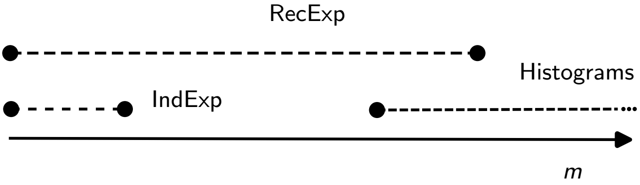
Concentration result :[33] Bins of size h, $\gamma > C_4 h$, π is L-Lipschitz, I is a strict sub-interval

$$\mathbb{P}\left(\| F^{-1}_{\hat{\pi}^{\text{hist}}} - F^{-1}_{\pi} \|_{\infty, I} > \gamma \right)$$
$$\leq \frac{1}{h} e^{-C_1 \gamma h n \epsilon} + \frac{2}{h} e^{-C_2 h^2 (C_3 \gamma - L h)^2 n} .$$

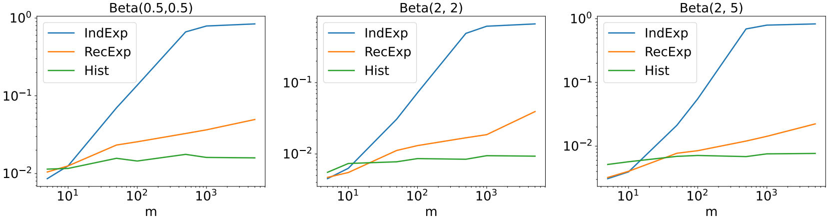
Remark : No degradation in m, but high entry cost.

[32] Wasserman and Zhou, "A Statistical Framework for Differential Privacy".
[33] Lalanne, Garivier, and Gribonval, "Private Statistical Estimation of Many Quantiles".

39 Theoretical choice of algorithm



40 Numerical validation

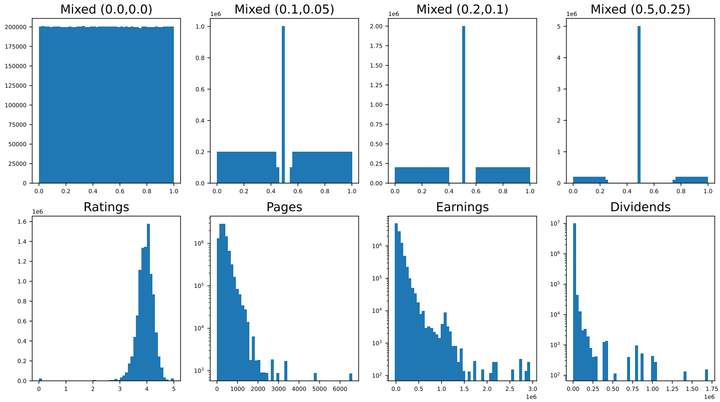


The vertical axis reads the error $\mathbb{E}\left(\|\mathbf{q} - F^{-1}(\mathbf{p})\|_\infty\right)$ where $\mathbf{p} = \left(\frac{1}{4} + \frac{1}{2(m+1)}, \ldots, \frac{1}{4} + \frac{m}{2(m+1)}\right)$ for different values of m, $n = 10000$, $\epsilon = 0.1$, and $\mathbb{E}$ is estimated by Monte-Carlo averaging over 50 runs. The histogram is computed on 200 bins.

41 Dealing with atomic distributions

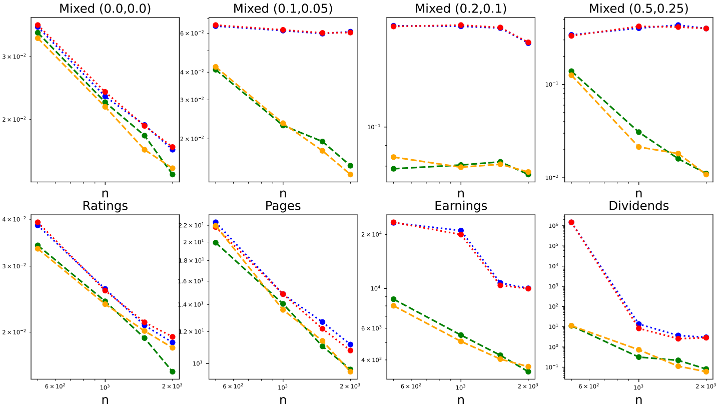
Inconsistency result :[34] When dealing with atomic distributions, all the *Exp mechanisms are inconsistent or have poor performances.

Proposed solution : Smoothing the distribution with noise addition can make those mechanisms consistent and helps the performances.



[34]Lalanne et al., "Private quantiles estimation in the presence of atoms".

42 Dealing with atomic distributions



The vertical axis reads the error $\mathbb{E}\left(\|\hat{\mathbf{q}} - F^{-1}(\mathbf{p})\|_\infty\right)$ where $\mathbf{p} = \left(\frac{1}{m+1}, \ldots, \frac{m}{m+1}\right)$ for $m = 8$, $\epsilon = 1$, $\hat{\mathbf{q}}$ is the private estimator, and $\mathbb{E}$ is estimated by Monte-Carlo averaging over 50 runs.

Table of Contents

Depending on the level of privacy, the effects on the estimation can be either negligible, or they can be made arbitrarily bad.

On a positive note, there exist regimes of increasing privacy at a negligible cost.

The complexity of the statistical problem greatly affects those different regimes.

Main contributions :

- Propose a framework for lower-bounds that builds on recent coupling constructions.
- Give a unified view on the cost of privacy for multiple definitions of privacy.
- Advances on the private density estimation problem with matching or near-matching lower and upper-bounds.
- An in depth analysis of some of the statistical properties of existing mechanisms for the pointwise estimation of the quantile function.

Better similarity functions for different regimes :

- What we used : $\mathrm{KL}\left(A(\mathbf{X}) \| A(\mathbf{Y})\right) \leq \epsilon d_{\mathrm{ham}}\left(\mathbf{X}, \mathbf{Y}\right)$
- State of the art[35] : $\mathrm{KL}\left(A(\mathbf{X}) \| A(\mathbf{Y})\right) \leq \epsilon d_{\mathrm{ham}}\left(\mathbf{X}, \mathbf{Y}\right) \frac{e^{\epsilon d_{\mathrm{ham}}(\mathbf{X},\mathbf{Y})}-1}{e^{\epsilon d_{\mathrm{ham}}(\mathbf{X},\mathbf{Y})}+1}$

Other constraints : Can it be applied to other forms of constraints (quantized, ...) ?

Numerical optimization :

$$\sup_{\mathbb{Q}\in\Pi\left(\mathbb{P}_1^{\otimes n},\ldots,\mathbb{P}_N^{\otimes n}\right)} \int s\left(\mathbf{X}_1,\ldots,\mathbf{X}_N\right) d\mathbb{Q}\left(\mathbf{X}_1,\ldots,\mathbf{X}_N\right)$$

[35] Fengxiang He, Bohan Wang, and Dacheng Tao. "Tighter Generalization Bounds for Iterative Differentially Private Learning Algorithms". In: 2021.

46 Remarks, Open Questions and Future work on Density Estimation

Optimality : What is the true optimal rate of estimation under ϵ-DP ?

$$\max\left\{n^{-\frac{2\beta}{2\beta+1}}, (n\epsilon)^{-\frac{2\beta}{\beta+1}}\right\} \quad \text{VS} \quad \max\left\{n^{-\frac{2\beta}{2\beta+1}}, (n\epsilon)^{-\frac{2\beta}{\beta+3/2}}\right\}$$

Higher dimensions : What happens when the dimensionality increases ?

47 Remarks, Open Questions and Future work on Quantiles Estimation

Lower bounds : Can we obtain meaningful lower-bounds for the problem ?

"Best of both worlds" mechanism : Can we find a mechanism that behaves like RecExp for a moderate amount of quantiles, but then behaves like Histograms when m is large ?

Functional estimation with regularity assumptions : Can we go from a point estimation problem to a functional estimation one ?

From quantiles to density estimation : We used density estimation to perform quantile function estimation, can we do the converse ?

- Angel, Omer and Yinon Spinka. *Pairwise optimal coupling of multiple random variables*. 2021.
- Asi, Hilal and John C. Duchi. "Near Instance-Optimality in Differential Privacy". In: *CoRR* abs/2005.10630 (2020). arXiv: 2005.10630. URL: https://arxiv.org/abs/2005.10630.
- Bun, Mark and Thomas Steinke. "Concentrated Differential Privacy: Simplifications, Extensions, and Lower Bounds". In: *Theory of Cryptography - 14th International Conference, TCC 2016-B, Beijing, China, October 31 - November 3, 2016, Proceedings, Part I*. Vol. 9985. Lecture Notes in Computer Science. 2016, pp. 635–658. DOI: 10.1007/978-3-662-53641-4_24. URL: https://doi.org/10.1007/978-3-662-53641-4%5C_24.
- Dwork, Cynthia and Guy N Rothblum. "Concentrated differential privacy". In: (2016).
- Dwork, Cynthia et al. "Calibrating Noise to Sensitivity in Private Data Analysis". In: *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006, Proceedings*. Vol. 3876. Lecture Notes in Computer Science. 2006, pp. 265–284. DOI: 10.1007/11681878_14. URL: https://doi.org/10.1007/11681878%5C_14.

Gillenwater, Jennifer, Matthew Joseph, and Alex Kulesza. "Differentially Private Quantiles". In: *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*. Vol. 139. Proceedings of Machine Learning Research. 2021, pp. 3713–3722. URL: http://proceedings.mlr.press/v139/gillenwater21a.html.

Gonon, Antoine et al. *Sparsity in neural networks can improve their privacy*. 2023. arXiv: 2304.10553 [cs.LG].

He, Fengxiang, Bohan Wang, and Dacheng Tao. "Tighter Generalization Bounds for Iterative Differentially Private Learning Algorithms". In: *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence, UAI 2021, Virtual Event, 27-30 July 2021*. Vol. 161. Proceedings of Machine Learning Research. 2021, pp. 802–812. URL: https://proceedings.mlr.press/v161/he21a.html.

Kaplan, Haim, Shachar Schnapp, and Uri Stemmer. "Differentially Private Approximate Quantiles". In: *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*. Ed. by Kamalika Chaudhuri et al. Vol. 162. Proceedings of Machine Learning Research. PMLR, 2022, pp. 10751–10761. URL: https://proceedings.mlr.press/v162/kaplan22a.html.

Lalanne, Clément, Aurélien Garivier, and Rémi Gribonval. "About the Cost of Central Privacy in Density Estimation". In: (2023). URL: https://openreview.net/forum?id=uq29MIWvIV.

- Lalanne, Clément, Aurélien Garivier, and Rémi Gribonval. "On the Statistical Complexity of Estimation and Testing under Privacy Constraints". In: (2023). URL: https://hal.science/hal-03794374.
- — ."Private Statistical Estimation of Many Quantiles". In: *ICML 2023 - 40th International Conference on Machine Learning*. 2023. URL: https://hal.science/hal-03986170.
- Lalanne, Clément et al. "Private quantiles estimation in the presence of atoms". In: 12 (2023). ISSN: 2049-8772. DOI: 10.1093/imaiai/iaad030. URL: https://doi.org/10.1093/imaiai/iaad030.
- Smith, Adam D. "Privacy-preserving statistical estimation with optimal convergence rates". In: *Proceedings of the 43rd ACM Symposium on Theory of Computing, STOC 2011, San Jose, CA, USA, 6-8 June 2011*. 2011, pp. 813–822. DOI: 10.1145/1993636.1993743. URL: https://doi.org/10.1145/1993636.1993743.
- Tsybakov, Alexandre B. *Introduction to Nonparametric Estimation*. Springer series in statistics. 2009. ISBN: 978-0-387-79051-0. DOI: 10.1007/b13794. URL: https://doi.org/10.1007/b13794.
- Wasserman, Larry A. and Shuheng Zhou. "A Statistical Framework for Differential Privacy". In: 105 (2010), pp. 375–389.