Robust estimation of discrete distributions under local differential privacy

Julien Chhor
ENSAE Paris

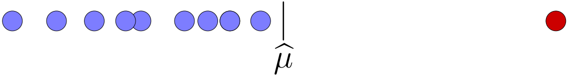
Flore Sentenac
ENSAE Paris

Séminaire de Statistique CREST-CMAP

Why robust estimation?

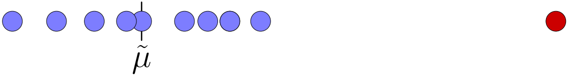
Example: $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. Optimal estimator: $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

What if an adversary replaces one of the X_i's with an outlier?



Contamination strongly impacts $\widehat{\mu}$: the estimator $\widehat{\mu}$ is not robust.

Now consider the empirical median $\widetilde{\mu} \in \mathrm{Med}(X_1, \ldots, X_n)$.



Contamination hardly affects $\widetilde{\mu}$: the estimator $\widetilde{\mu}$ is robust.

Goal: Find estimators that are robust to contamination.

Why Local differential Privacy?

Setting: Let $X_1, \ldots, X_n \overset{iid}{\sim} p$.

The $X_1, \ldots, X_n$ are sensitive: they should not be disclosed to the statistician.

Idea: Add noise to each X_i! If $X_i = x$, draw $Z_i \sim Q(\cdot | X = x)$.

Here, Q denotes some Markov transition kernel.

Goal: (Informal) Ensure that from Z_i, one "cannot recover" X_i.

Local Differential Privacy

Definition: Fix $\alpha \in (0,1)$. A Markov transition kernel $Q : \mathcal{X} \to \mathcal{Z}$ is a (non-interactive) α-*locally differentially private* mechanism if

$$\sup_{B \in \sigma(\mathcal{Z})} \sup_{x, x' \in \mathcal{X}} \frac{Q(B|x)}{Q(B|x')} \leq e^{\alpha}. \quad (*)$$

Intuition: Let $x, x' \in \mathcal{X}$. From the observation $Z_i \sim Q(\cdot|X_i)$, consider

$$H_0 : X_i = x \qquad \text{vs} \qquad H_1 : X_i = x'.$$

The likelihood-ratio test $\mathbb{1}\left\{\frac{Q(Z_i|x')}{Q(Z_i|x)} > 1\right\}$ is minimax optimal.

But under $(*)$, it has Type-I + Type-II error $\in \left[1-\alpha, 1\right]$.

(Random guessing has Type-I + Type-II error $= 1$.)

Our setting

Let $\mathcal{P}_d = \left\{ \left(p_1, \ldots, p_d\right) \in \mathbb{R}_+^d \;\middle|\; \sum_{j=1}^{d} p_j = 1 \right\}$ for $d \geq 3$.

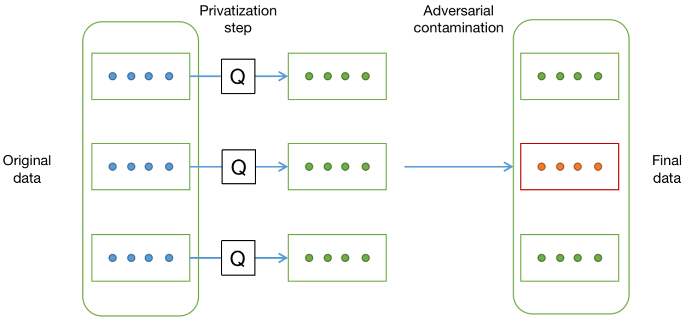
Privacy level $\alpha \in (0, 1)$, Corruption level: $\epsilon \in \left(0, \frac{1}{100}\right)$.

Underlying distribution $p \in \mathcal{P}_d$ to estimate, Q is chosen by the statistician.

1. Collect n iid batches $X^1, \ldots, X^n$ of size k: $X^i = \left[X_1^i, \ldots, X_k^i\right] \overset{iid}{\sim} p^{\otimes k}$.
2. Privatize each X_j^i to define $Y_j^i \sim Q(\,\cdot\,|X_j^i)$.
3. An adversary replaces $n\epsilon$ batches Y^i by arbitrary outliers $\widetilde{Y}^i$.

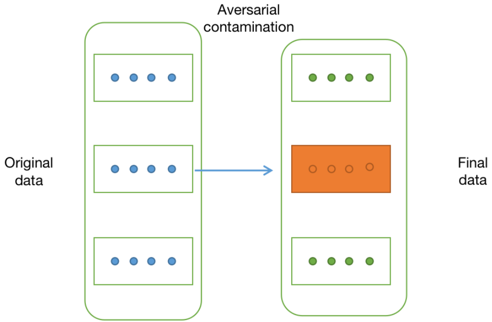
The resulting dataset is denoted as $(Z^1, \ldots, Z^n)$.

Our setting



Privatization
step
Adversarial
contamination
Q
Q
Q
Original
data
Final
data

With contamination only



With contamination only

1. Collect n iid batches $X^1, \ldots, X^n$ of size k: $X^i = \left[X_1^i, \ldots, X_k^i\right] \overset{iid}{\sim} p^{\otimes k}$.
2. An adversary replaces $n\epsilon$ batches X^i by arbitrary outliers $\widetilde{X}_i$.

The resulting dataset is denoted as $(Y_1, \ldots, Y_n)$.

Theorem (Qiao and Valiant, 2017)

There exists $\widehat{p}$ such that w.p. $\geq 1 - O(e^{-d})$,

$$\sup_p TV(p, \widehat{p}) \lesssim \sqrt{\frac{d}{nk}} + \frac{\epsilon}{\sqrt{k}}.$$

There exists a constant $c > 0$ s.t. for all estimator $\widehat{p}$, w.p. $\geq O(e^{-d})$

$$\sup_{p \in \mathcal{P}_d} TV(p, \widehat{p}) \geq c \left\{ \sqrt{\frac{d}{nk}} + \frac{\epsilon}{\sqrt{k}} \right\}.$$

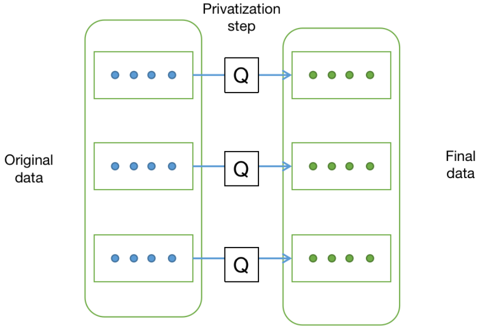
Computational tractability

Theorem (Jain and Orlitsky, 2020)

There exists a polynomial time algorithm $\widehat{p}$ *s.t. w.p.* $\geq 1 - O(e^{-d})$,

$$TV(p, \widehat{p}) \lesssim \sqrt{\frac{d}{nk}} + \frac{\epsilon}{\sqrt{k}} \sqrt{\log(1/\epsilon)}.$$

With privatization only



Randomized response mechanism

- $p = (p_1, \ldots, p_d)$ s.t. $\sum_{j=1}^{d} p_j = 1$, $X_1, \ldots, X_n \overset{iid}{\sim} p$.
- Privacy level α.

The following mechanism is α-LDP and minimax optimal for $\alpha \in (0, 1)$.

RAPPOR mechanism

Input: $X \in [d]$ and $\alpha \in (0, 1)$.

Define $\lambda = \frac{1}{e^{\alpha/2}+1}$.

Output: $Z \in \{0, 1\}^d$ with independent coordinates such that

$$\forall j \in [d] : \quad Z(j) = \begin{cases} \mathbb{1}_{X=j} & \text{with probability } 1 - \lambda, \\ 1 - \mathbb{1}_{X=j} & \text{otherwise.} \end{cases}$$

Randomized response mechanism

We have:

$$\mathbb{E}\left[Z\left(j\right)\right] = \frac{e^{\alpha/2}-1}{e^{\alpha/2}+1}p_j + \frac{1}{1+e^{\alpha/2}}.$$

Define

$$\hat{p}_j := \frac{e^{\alpha/2}+1}{e^{\alpha/2}-1}\left[\frac{1}{n}\sum_{i\in[n]} Z_i - \frac{1}{1+e^{\alpha/2}}\right].$$

If $\alpha \ll 1$, we have:

$$\mathbb{E}\left[\|\hat{p}-p\|_1\right] \approx \frac{d}{\alpha\sqrt{n}}.$$

In comparison:

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} X_i - p\right\|_1\right] \approx \sqrt{\frac{d}{n}}.$$

Effective sample size $\sim \alpha^2 n/d$.

This estimation rate is minimax optimal (up to constants).

Theorem (Duchi, Jordan, and Wainwright, 2014)

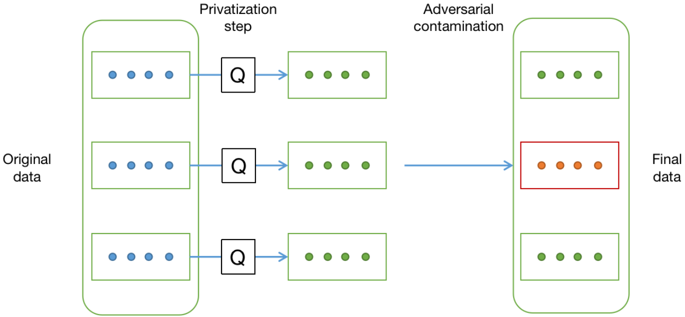
For any α-LDP mechanism Q,

$$\inf_{\hat{p}} \sup_{p} \mathbb{E}\left[\left\|\hat{p} - p\right\|_1\right] \gtrsim \min\left(1, \frac{d}{\alpha\sqrt{n}}\right).$$

If $\hat{p}$ is estimated through the RAPPOR algorithm and $\alpha \in [0,1]$, then:

$$\inf_{\hat{p}} \sup_{p} \mathbb{E}\left[\left\|\hat{p} - p\right\|_1\right] \lesssim \frac{d}{\alpha\sqrt{n}}.$$

Our setting (reminder)



1. Collect n iid batches $X^1, \ldots, X^n$ of size k: $X^i = \left[X_1^i, \ldots, X_k^i\right] \overset{iid}{\sim} p^{\otimes k}$.
2. Privatize each X_j^i to define $Y_j^i \sim Q(\,\cdot\,|X_j^i)$.
3. An adversary replaces $n\epsilon$ batches Y^i by arbitrary outliers $\widetilde{Y}^i$.

Main theorem

Theorem

- *If $n \geq O(d)$, there is a polynomial time algorithm $\widehat{p}$ such that*

$$\sup_{p \in \mathcal{P}_d} TV(p, \widehat{p}) \lesssim \frac{\epsilon}{\alpha}\sqrt{\frac{d \ln(1/\epsilon)}{k}} + \frac{d}{\alpha\sqrt{nk}}$$

with probability at least $1 - O(e^{-d})$.

- *There exists a constant $c > 0$ s.t. for all estimator $\widehat{p}$, all α-LDP privatization channels Q, w.p. $\geq O(e^{-d})$*

$$\sup_{p \in \mathcal{P}_d} TV(p, \widehat{p}) \geq c \left\{ \frac{\epsilon}{\alpha}\sqrt{\frac{d}{k}} + \frac{d}{\alpha\sqrt{nk}} \right\}.$$

Rates comparison

- n batches of k samples $\to$ nk samples.
- Privacy level α.
- contamination level ϵ.

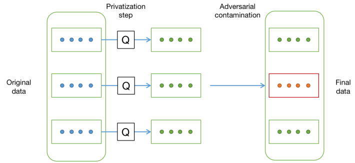
Constraint	Upper bound	Lower bound
Contamination+LDP (Our bound)	$\frac{d}{\alpha\sqrt{nk}} + \frac{\epsilon\sqrt{\log(1/\epsilon)}}{\sqrt{k}}\sqrt{\frac{d}{\alpha^2}}$	$\frac{d}{\alpha\sqrt{nk}} + \frac{\epsilon}{\sqrt{k}}\sqrt{\frac{d}{\alpha^2}}$
LDP only	$\frac{d}{\alpha\sqrt{nk}}$	$\frac{d}{\alpha\sqrt{nk}}$
Contamination only	$\sqrt{\frac{d}{nk}} + \frac{\epsilon\sqrt{\log(1/\epsilon)}}{\sqrt{k}}$	$\sqrt{\frac{d}{nk}} + \frac{\epsilon}{\sqrt{k}}$

Related work

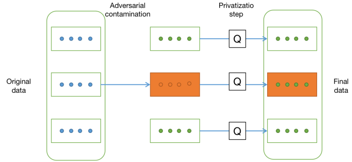
- **acharya2021robust**; Cheu, Smith, and Ullman, 2021: Consider contamination *after privacy* in various settings including discrete distributions. Nearly matching upper and lower bounds.
- Li, Berrett, and Yu, 2022 Consider contamination *before privacy* in various settings.
- Liu et al., 2021: $(X_i)_i$ iid from Subgaussian distribution. The data $(X_i)_i$ are contaminated *before privatization*.

None of them consider batches.

Contamination before vs. after privacy



Estimation error caused by contamination multiplied by $\sqrt{d}/\alpha$.



Estimation error caused by contamination unchanged.

Algorithm

The algorithm proceeds in two mains steps:

1. Privatization step: Using the RAPPOR mechanism.
2. Robust estimation step: estimate the auxiliary quantity

$$q(j) := \mathbb{E}_p\big[Z(j)\big|\, Z \text{ is a good sample}\big] \quad \text{for all } j \in [d].$$

Deduce $\hat{p}$ from $\hat{q}$.

Privatization step

$$q(j) := \mathbb{E}_p\left[Z(j)\middle| Z \text{ is a good sample}\right] \quad \text{for all } j \in [d].$$

One has: $p = \frac{e^\alpha+1}{e^\alpha-1}\left(q - \frac{1}{1+e^\alpha}\mathbb{1}\right)$.

Given an estimator $\hat{q}$, one can provide the estimator $\hat{p}$ through

$$\hat{p}_j := \underbrace{\frac{e^\alpha+1}{e^\alpha-1}}_{\asymp 1/\alpha}\left[\hat{q}_j - \frac{1}{1+e^\alpha}\right].$$

Thus, the error on $\hat{p}$ is controlled by the error on $\hat{q}$:,

$$\sum_{j=1}^{n} |\hat{p}_j - p_j| \asymp \frac{1}{\alpha}\sum_{j=1}^{n} |\hat{q}_j - q_j|.$$

Robust estimation step

The algorithm is based on an iterative filtering of the batches.

- We define for a collection of batches B' a contamination rate $\tau_{B'}$.
- For each batch b, we define its corruption score ϵ_b.
- Until the contamination rate is low, batches are eliminated based on the corruption score.

Iterative Filtering Mechanism

Input: Corruption level ϵ, Batch collection B.

Initialize $B' \leftarrow B$

While contamination rate of B', $\tau_{B'} \geq 200$:

$\forall b \in B'$ compute corruption score ε_b

$B^o \leftarrow \{\epsilon n$ Batches with top corruption scores$\}$

Define $\epsilon_{\text{tot}} = \sum_{b \in B^o} \varepsilon_b$

While $\sum_{b \in B^o} \varepsilon_b \geq \epsilon_{\text{tot}}/2$:

Delete a batch from B^o, picking batch b with probability proportional to ε_b

Output: Collection with low contamination rate B'

Under some technical assumptions:

1. If the contamination rate of a collection B' is smaller than a constant, then the empirical mean of the frequencies of each coordinate in B', $\widehat{q}_{B'}$ satisfies:

$$\sup_{S \subseteq [d]} \sum_{j \in S} |\widehat{q}_{B'}(j) - q_j| \lesssim \epsilon \sqrt{\frac{d}{k}}.$$

2. Any collection of "good" batches has a low contamination rate.
3. Each deletion step of the iterative filtering procedure deletes an adversarial batch w.p. at least $3/4$.

Contamination rate

For simplicity, assume we want to estimate the first coordinate q_1.

Define for each batch b and each collection and batches B':

$$\hat{q}_b(1) := \frac{1}{k}\sum_{i=1}^{k} Z_i^b(1) \quad \text{and} \quad \hat{q}_{B'}(1) := \frac{1}{|B'|}\sum_{b\in B'} \hat{q}_b(1).$$

Introduce the following estimates of the second order moment:

$$\widehat{\mathrm{Var}}_1^{B'}(b) := \sum_{b\in B'} \left[\widehat{q}_b(1) - \widehat{q}_{B'}(1)\right]^2,$$

$$\mathrm{Var}_1\left(\widehat{q}_{B'}(1)\right) := \frac{\widehat{q}_{B'}(1)(1-\widehat{q}_{B'}(1))}{k}.$$

The proxy contamination rate is defined through:

$$\tau_{B'} := \frac{1}{\frac{\epsilon d \ln(1/\epsilon)}{k}} \left| \mathrm{Var}_1\left(\widehat{q}_{B'}(1)\right) - \widehat{\mathrm{Var}}_1^{B'}(b) \right|.$$

Corruption scores

The proxy contamination scores are defined as:

$$\epsilon_b := \left[\widehat{q}_b(1) - \widehat{q}_{B'}(1)\right]^2.$$

Lower Bound

For any α-LDP mechanism Q, there exit two probability vectors $p, q \in \mathcal{P}_d$ s.t.:

$$\|p - q\|_1 \gtrsim \frac{\epsilon\sqrt{d}}{\alpha\sqrt{k}} \wedge 1$$

and

$$TV(Qp^{\otimes k}, Qq^{\otimes k}) \leq \epsilon.$$

Thank you !

References

Cheu, Albert, Adam Smith, and Jonathan Ullman (2021). "Manipulation attacks in local differential privacy". In: *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, pp. 883–900.

Duchi, John C., Michael I. Jordan, and Martin J. Wainwright (2014). *Local Privacy, Data Processing Inequalities, and Statistical Minimax Rates*. arXiv: `1302.3203 [math.ST]`.

Jain, Ayush and Alon Orlitsky (2020). *Optimal Robust Learning of Discrete Distributions from Batches*. arXiv: `1911.08532 [cs.LG]`.

Li, Mengchu, Thomas B Berrett, and Yi Yu (2022). "On robustness and local differential privacy". In: *arXiv preprint arXiv:2201.00751*.

Liu, Xiyang et al. (2021). "Robust and differentially private mean estimation". In: *Advances in Neural Information Processing Systems* 34.

Qiao, Mingda and Gregory Valiant (2017). "Learning discrete distributions from untrusted batches". In: *arXiv preprint arXiv:1711.08113*.