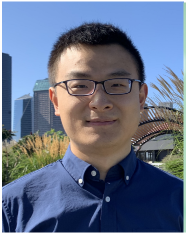
Differential Privacy Meets Robust Statistics

Sewoong Oh

Paul G. Allen School of Computer Science and Engineering
University of Washington

joint work with

Xiyang Liu

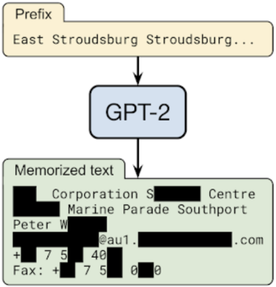


Weihao Kong

Sham Kakade

What can go wrong when training on shared data?

- Increasingly more models are being trained on shared data
- Sensitive information should not be revealed by the trained model
- Membership inference attacks can identify individual's sensitive data used in the training



[*]

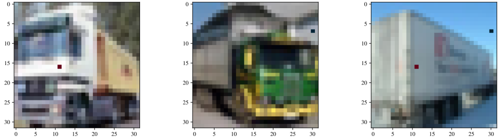
- Potential defense: Differentially Private Stochastic Gradient Descent[†] when computing the average of the gradients in the mini-batch, use differentially private mean estimation

[*] [Carlini et al.,2020]

[†] [Chaudhuri,Monteleoni,Sarwate,2011], [Abadi et al.,2016]

What can go wrong when training on shared data?

- When training on shared data, not all participants are trusted
- Malicious users can easily inject corrupted data
- Data poisoning attacks can create backdoors on the trained model such that any sample with the trigger will be predicts as ‘deer’



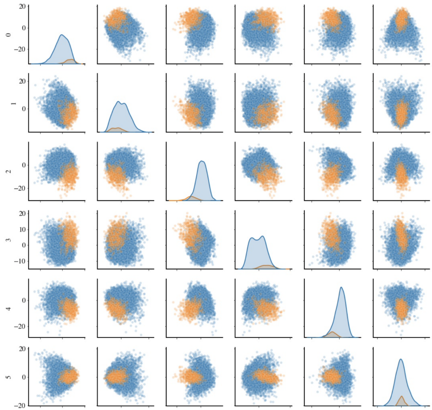
$y_i =$ ‘deer’

- Strong defense: Robust estimation*
- Insight: successful backdoor attacks leave a path of activations in the trained model that are triggered only by the corrupted samples

*[Hayase,Kong,Somani,O.,2021,ICML] inspired by [Tran,Li,Madry,2018]

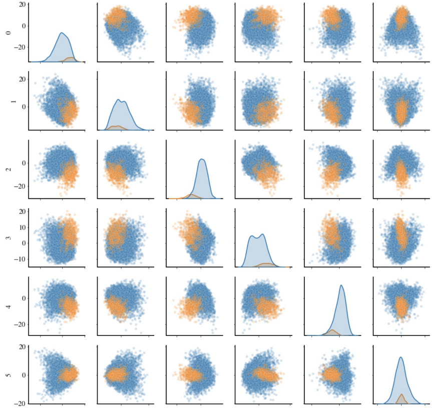
Middle layer of a model trained with corrupted data

- All samples with label 'deer': CLEAN and POISONED
- Top-6 PCA projection of node activations at a middle layer
- Can we separate POISONED from CLEAN?

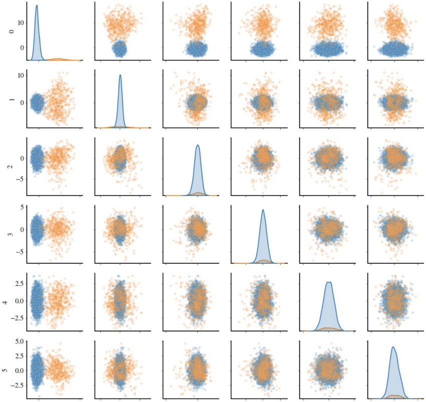


Middle layer of a model trained with corrupted data

- All samples with label 'deer': CLEAN and POISONED
- Top-6 PCA projection of node activations at a middle layer
- Can we separate POISONED from CLEAN?

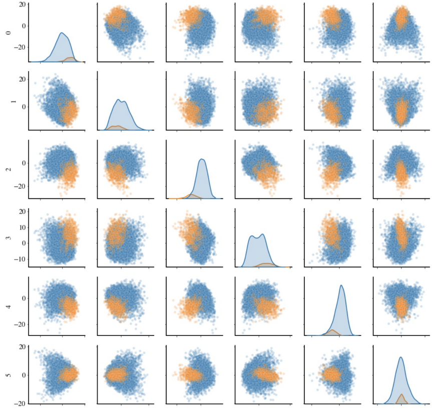


After whitening with
the covariance of CLEAN

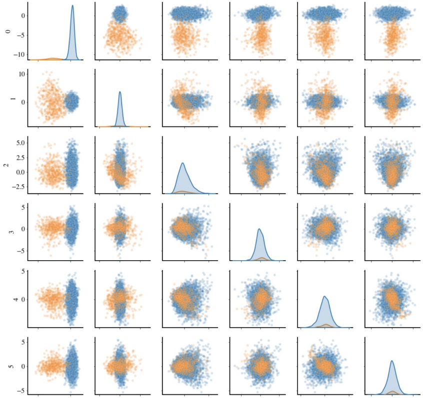


Middle layer of a model trained with corrupted data

- All samples with label 'deer': CLEAN and POISONED
- Top-6 PCA projection of node activations at a middle layer
- Can we separate POISONED from CLEAN?

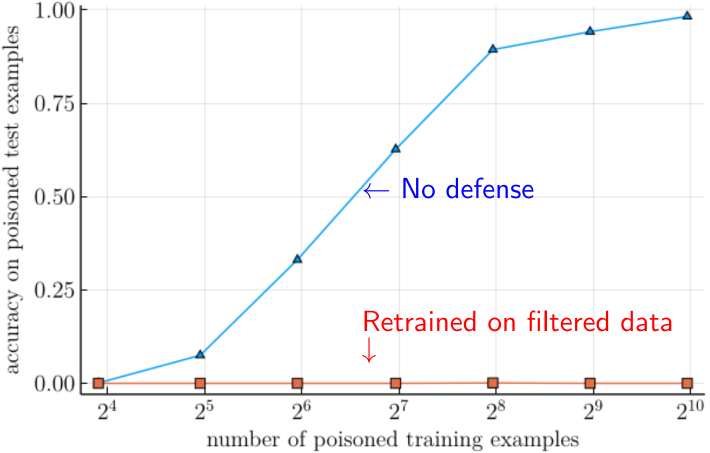


After whitening with
estimated **robust** mean and covariance



SPECTRE: Defense against backdoor attacks

[Hayase,Somani,Kong,O.,2021,ICML]‡



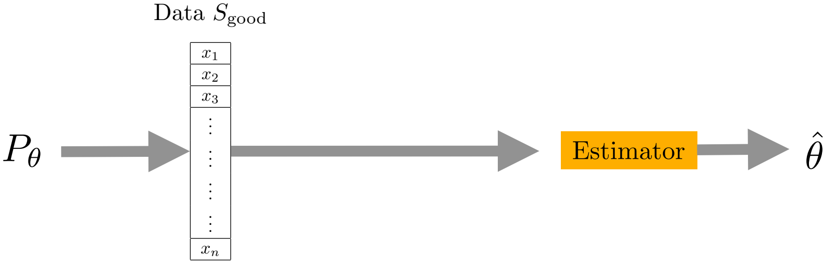
‡ `https://github.com/SewoongLab/backdoor-suite`

We need privacy and robustness, simultaneously

- When learning from shared data
  - Differential privacy is crucial in defending against inference attacks
  - Robust estimation is crucial in defending against data poisoning attacks
- Critical components are mean/covariance estimation
  - DP-SGD relies on DP mean estimation
  - Backdoor defense relies on robust mean/covariance estimation
- We provide the first efficient estimators that are provably differentially private and robust against data corruption

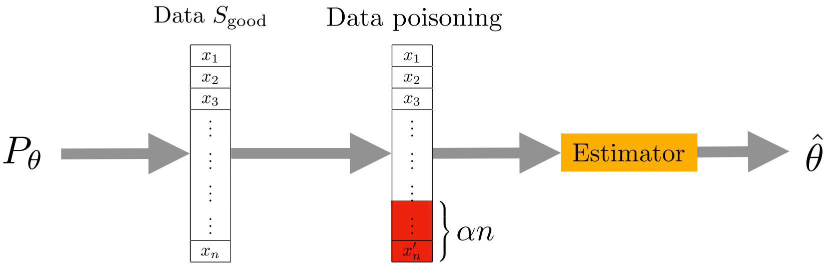
Statistical estimation, robustly and privately

- Statistics



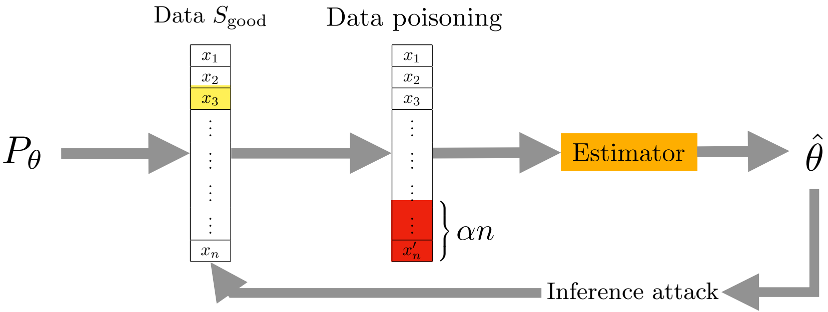
Statistical estimation, robustly and privately

- Statistics$\Rightarrow$ Robust estimation



Statistical estimation, robustly and privately

- Statistics$\Rightarrow$ Robust estimation$\Rightarrow$ Robust and private estimation



- This talk focuses on mean estimation
- Q. What is the extra cost (in the estimation error) we pay for {Robustness, Privacy, and Robustness+Privacy}

Mean estimation

- Estimate the mean μ from n i.i.d. samples
- For this talk,
  we assume sub-Gaussian distribution with identity covariance matrix
- Minimax error rate:

$$\min_{\hat{\mu}\in\mathcal{F}_{S_n}} \max_{P_\mu} \mathbb{E}\left[\, \|\hat{\mu}(S_n) - \mu\| \,\right] \propto \sqrt{\frac{d}{n}}$$

  $\mathcal{F}_{S_n}$ is set of all estimators over n i.i.d. samples in $\mathbb{R}^d$ from P_μ,
  P_μ is maximized over all sub-Gaussian distributions with identity covariance
- In this talk, I will ignore all constant and logarithmic factors

Robust mean estimation

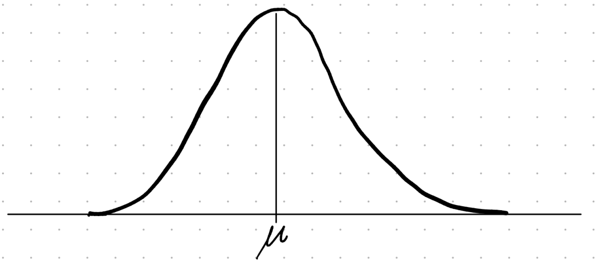
- Threat model
  - Adversarial corruption model:
    $\{x_i\}_{i=1}^n \sim P_\mu$ is drawn, then adversary replaces α-fraction arbitrarily

- Robust mean estimation:
  - Low dimensional:
    [Tukey,1960] [Huber,1964]
  - Computationally intractable methods in high dimension:
    [Donoho,Liu,1988], [ChenGaoRen,2015],[Zhu,Jiao,Steinhardt,2019]
  - Breakthroughs in polynomial time algorithms:
    [Lai,Rao,Vempala,2016],[Diakonikolas,Kamath,Kane,Li,Moitra,Stewart,2019]
  - Linear time algorithms:
    [Cheng,Dianikolas,Ge,2019], [Depersin,Lecué,2019],[Dong,Hopkins,Li,2019]

Robust mean estimation

- Threat model
  - Adversarial corruption model:
    $\{x_i\}_{i=1}^n \sim P_\mu$ is drawn, then adversary replaces α-fraction arbitrarily
- Relatively easy to estimate mean robustly in low-dimensions

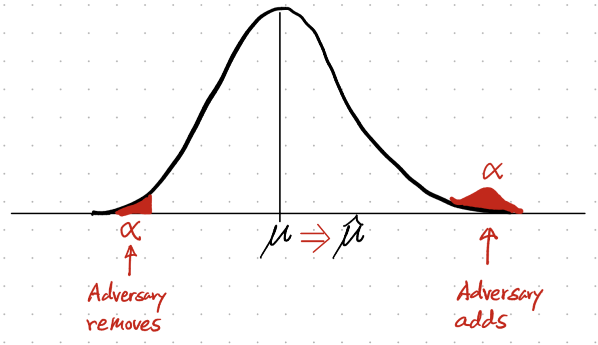
histogram of sub-Gaussian samples in 1D



Robust mean estimation

- Threat model
  - Adversarial corruption model:
    $\{x_i\}_{i=1}^n \sim P_\mu$ is drawn, then adversary replaces α-fraction arbitrarily
- Relatively easy to estimate mean robustly in low-dimensions

histogram of sub-Gaussian samples in 1D

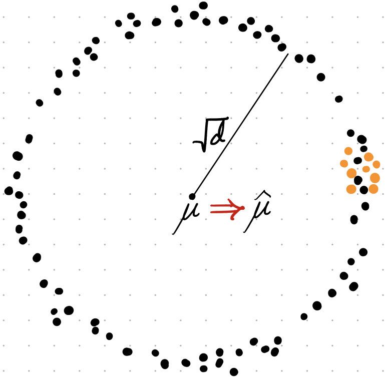


simple outlier detection achieves $|\hat{\mu} - \mu| \leq \alpha$

Robust mean estimation

- Threat model
  - Adversarial corruption model:
    $\{x_i\}_{i=1}^{n} \sim P_\mu$ is drawn, then adversary replaces α-fraction arbitrarily
- Mean estimation becomes challenging in high-dimensions

scatter plot of sub-Gaussian samples in high-dimension



each corrupted sample looks uncorrupted and still $\|\hat{\mu} - \mu\| \geq \alpha\sqrt{d}$

Efficient algorithm: Filtering [Diakonikolas et al.,2017]

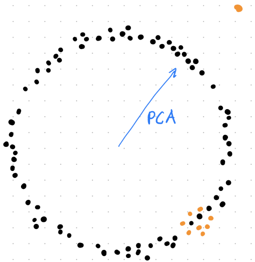
Geometric Lemma [Dong,Hopkins,Li,2019]

Given n i.i.d. samples from a sub-Gaussian distribution with identity covariance matrix, if at most αn samples are corrupted, then, w.h.p.

$$\|\mu_{\mathrm{emp}}(S) - \mu\| \ \leq \ \sqrt{\frac{d}{n}} + \alpha + \sqrt{\alpha \|\mathrm{Cov}(S) - \mathbf{I}\|}$$

- While $\|\mathrm{Cov}(S) - \mathbf{I}\| > c\,\alpha$
  - $v \leftarrow \arg\max_{v:\|v\|=1} v^T \mathrm{Cov}(S) v$
  - $S \leftarrow$ 1D-Filter$(\{\langle v, x_i - \mu_{\mathrm{emp}}(S)\rangle^2\}_{i\in S})$

- Each step guarantees that
  - at least one sample is removed
  - more corrupted samples removed than clean samples in expectation



Efficient algorithm: Filtering [Diakonikolas et al.,2017]

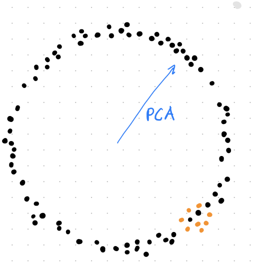
Geometric Lemma [Dong,Hopkins,Li,2019]

Given n i.i.d. samples from a sub-Gaussian distribution with identity covariance matrix, if at most αn samples are corrupted, then, w.h.p.

$$\|\mu_{\mathrm{emp}}(S) - \mu\| \ \le \ \sqrt{\frac{d}{n}} + \alpha + \sqrt{\alpha \|\mathrm{Cov}(S) - \mathbf{I}\|}$$

- While $\|\mathrm{Cov}(S) - \mathbf{I}\| > c\,\alpha$
  - $v \leftarrow \arg\max_{v:\|v\|=1} v^T \mathrm{Cov}(S) v$
  - $S \leftarrow$ 1D-Filter$(\{\langle v, x_i - \mu_{\mathrm{emp}}(S)\rangle^2\}_{i\in S})$

- Each step guarantees that
  - at least one sample is removed
  - more corrupted samples removed than clean samples in expectation



Efficient algorithm: Filtering [Diakonikolas et al.,2017]

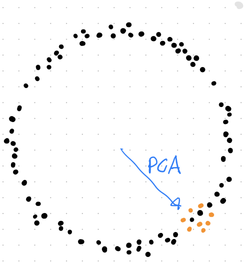
Geometric Lemma [Dong,Hopkins,Li,2019]

Given n i.i.d. samples from a sub-Gaussian distribution with identity covariance matrix, if at most αn samples are corrupted, then, w.h.p.

$$\|\mu_{\text{emp}}(S) - \mu\| \le \sqrt{\frac{d}{n}} + \alpha + \sqrt{\alpha \|\mathrm{Cov}(S) - \mathbf{I}\|}$$

- While $\|\mathrm{Cov}(S) - \mathbf{I}\| > c\,\alpha$
  - $v \leftarrow \arg \max_{v:\|v\|=1} v^T \mathrm{Cov}(S) v$
  - $S \leftarrow$ 1D-Filter$(\{\langle v, x_i - \mu_{\text{emp}}(S)\rangle^2\}_{i\in S})$

- Each step guarantees that
  - at least one sample is removed
  - more corrupted samples removed than clean samples in expectation



Efficient algorithm: Filtering [Diakonikolas et al.,2017]

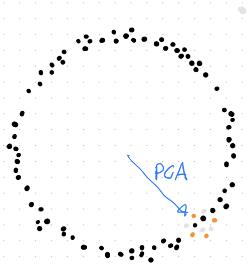
Geometric Lemma [Dong,Hopkins,Li,2019]

Given n i.i.d. samples from a sub-Gaussian distribution with identity covariance matrix, if at most αn samples are corrupted, then, w.h.p.

$$\|\mu_{\text{emp}}(S) - \mu\| \ \le \ \sqrt{\frac{d}{n}} + \alpha + \sqrt{\alpha \|\mathrm{Cov}(S) - \mathbf{I}\|}$$

- While $\|\mathrm{Cov}(S) - \mathbf{I}\| > c\,\alpha$
  - $v \leftarrow \arg\max_{v:\|v\|=1} v^T \mathrm{Cov}(S) v$
  - $S \leftarrow$ 1D-Filter$(\{\langle v, x_i - \mu_{\text{emp}}(S)\rangle^2\}_{i \in S})$

- Each step guarantees that
  - at least one sample is removed
  - more corrupted samples removed than clean samples in expectation



Robust mean estimation

- Minimax error rate under α-corruption

$$\min_{\hat{\mu}} \max_{P_\mu} \mathbb{E}\big[\,\|\hat{\mu}(S_{n,\alpha}) - \mu\|\,\big] \;\propto\; \underbrace{\sqrt{\frac{d}{n}}}_{\text{no corruption}} + \underbrace{\alpha}_{\alpha\text{-corruption}}$$

achieved by filtering algorithm of [Diakonikolas et al.,2017]
information-theoretic lower bound from [Chen,Gao,Ren,2015]

Minimax error rate for mean estimation under sub-Gaussian distributions with identity covariance

	Error $\|\hat{\mu} - \mu\|$
no corruption or privacy	$\sqrt{\frac{d}{n}}$
α-corruption	$\sqrt{\frac{d}{n}} + \alpha$ [Diakonikolas et al.,2017]
(ε, δ)-DP	
α-corruption and (ε, δ)-DP	

Differential Privacy provably ensures plausible deniability

- Goal: a strong adversary who knows all the other entries in the database except for yours, should not be able to identify whether you participated in that database or not

- Definition*: For two databases S and S' that differ by only one entry, a randomized output to a query is (ε, δ)-differentially private if

$$\mathbb{P}(\text{query\_output}(S) \in A) \;\leq\; e^{\varepsilon}\, \mathbb{P}(\text{query\_output}(S') \in A) + \delta$$

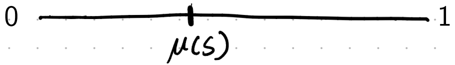
- smaller $\varepsilon, \delta \;\Rightarrow\;$ Testing S or S' fails $\;\Rightarrow\;$ inference attack fails

---

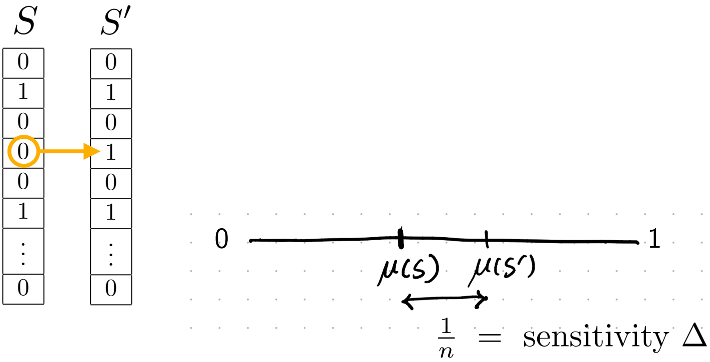
*[Dwork,McSherry,Nissim,Smith,2006]

(ε, δ)-differentially private mean estimation

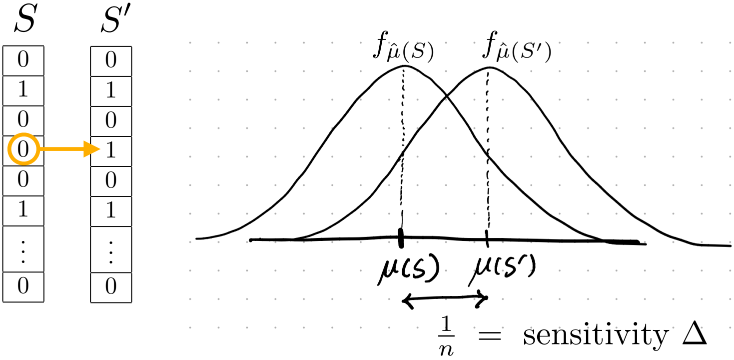
S
0
1
0
0
0
1
$\vdots$
0



(ε, δ)-differentially private mean estimation



(ε, δ)-differentially private mean estimation

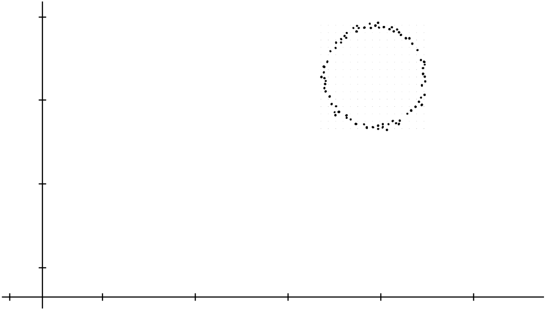


$$\hat{\mu}(S) = \mu(S) + \mathcal{N}\left(0, \left(\frac{\Delta\sqrt{\log 1/\delta}}{\varepsilon}\right)^2\right)$$

- extra error due to (ε, δ)-DP is

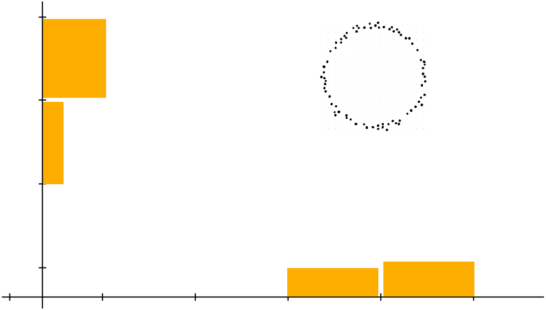
$$|\hat{\mu}(S) - \mu(S)| \simeq \frac{\Delta}{\varepsilon} = \frac{1}{n\,\varepsilon}$$

(ε, δ)-differentially private mean estimation*



*[Karwa,Vadhan,2017], [Kamath,Li,Singhal,Ullman,2019]

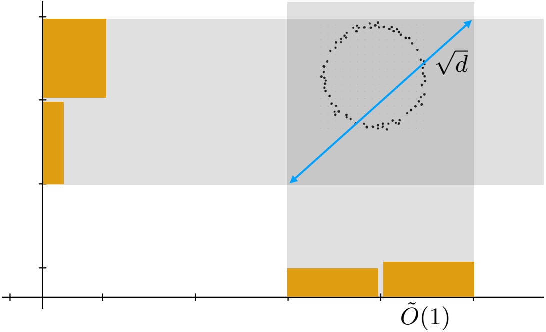
(ε, δ)-differentially private mean estimation*



- step 1. privately find a bounding hypercube

*[Karwa,Vadhan,2017], [Kamath,Li,Singhal,Ullman,2019]

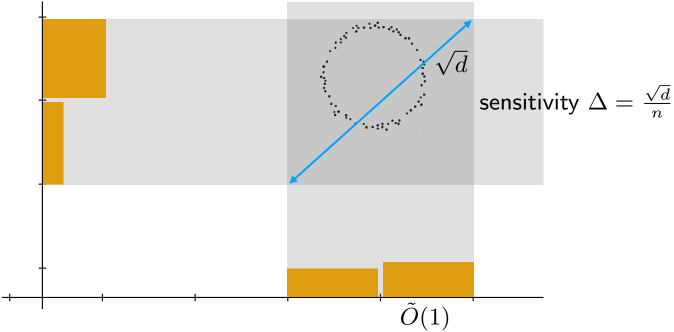
(ε, δ)-differentially private mean estimation*



- step 1. privately find a bounding hypercube

*[Karwa,Vadhan,2017], [Kamath,Li,Singhal,Ullman,2019]

(ε, δ)-differentially private mean estimation*



- step 1. privately find a bounding hypercube
- step 2. add Gaussian noise: $\hat{\mu}(S) = \mu(S) + \mathcal{N}\left(0, \left(\frac{\Delta\sqrt{\log 1/\delta}}{\varepsilon}\right)^2 \mathbf{I}_{d\times d}\right)$
- extra error due to (ε, δ)-DP is

$$\|\hat{\mu}(S) - \mu(S)\| \simeq \frac{\Delta}{\varepsilon}\sqrt{d} = \frac{d}{n\,\varepsilon}$$

*[Karwa,Vadhan,2017], [Kamath,Li,Singhal,Ullman,2019]

Minimax error rate for mean estimation under sub-Gaussian distribution with identity covariance

	Error $\|\hat{\mu} - \mu\|$	
no corruption or privacy	$\sqrt{\frac{d}{n}}$	
α-corruption	$\sqrt{\frac{d}{n}} + \alpha$	[Diakonikolas et al.,2017]
(ε, δ)-DP	$\sqrt{\frac{d}{n}} + \frac{d}{\varepsilon\, n}$	[Kamath,Li,Singhal,Ullman,2019]
α-corruption and (ε, δ)-DP		

Two main challenges in making filtering algorithms private

Algorithm (non-private) robust mean estimation [Diakonikolas et al.,2017]

1: **while** $\|\mathrm{Cov}(S) - \mathbf{I}\| > c\,\alpha$ **do**

2: $\quad v \leftarrow \arg\max_{v:\|v\|=1} v^T \mathrm{Cov}(S) v$

3: $\quad S \leftarrow$ 1D-Filter$(\{\langle v, x_i - \mu_{\mathrm{emp}}(S)\rangle^2\}_{i\in S})$

- First challenge:
  - in the worst case, the filter runs for $O(d)$ iterations
  - this happens if corrupted sample are spread out in orthogonal directions
  - because the filter only checks 1-dimensional subspace at a time
- This is particularly damaging for privacy, as more iterations mean more privacy leakage

Two main challenges in making filtering algorithms private

Algorithm Quantum robust mean estimation [Dong,Hopkins,Li,2019]

1: **while** $\|\mathrm{Cov}(S) - \mathbf{I}\| > c\,\alpha$ **do**

2: $\quad V \leftarrow \frac{1}{\mathrm{Trace}(\exp\{\beta \mathrm{Cov}(S)\})} \exp\{\beta \mathrm{Cov}(S)\}$

3: $\quad S \leftarrow$ 1D-Filter$(\{(x_i - \mu_{\mathrm{emp}}(S))^T V (x_i - \mu_{\mathrm{emp}}(S))\}_{i \in S})$

Two main challenges in making filtering algorithms private

Algorithm Quantum robust mean estimation [Dong,Hopkins,Li,2019]

1: **while** $\|\mathrm{Cov}(S) - \mathbf{I}\| > c\,\alpha$ **do**

2: $\quad V \leftarrow \frac{1}{\mathrm{Trace}(\exp\{\beta \mathrm{Cov}(S)\})} \exp\{\beta \mathrm{Cov}(S)\}$

3: $\quad S \leftarrow$ 1D-Filter$(\{(x_i - \mu_{\mathrm{emp}}(S))^T V (x_i - \mu_{\mathrm{emp}}(S))\}_{i \in S})$

- If $\beta = \infty$, this recovers top PCA and uses only one-dimensional subspace
- If $\beta = 0$, this filters on $\|x_i - \mu_{\mathrm{emp}}(S)\|^2$ treating all directions equally
- For appropriate β, iterations reduce from $O(d)$ to $O((\log d)^2)$

Two main challenges in making filtering algorithms private

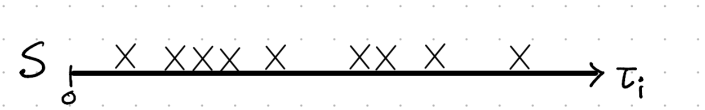
Algorithm Quantum robust mean estimation [Dong,Hopkins,Li,2019]

while $\|\mathrm{Cov}(S) - \mathbf{I}\| > c\,\alpha$ **do**

$\quad V \leftarrow \frac{1}{\mathrm{Trace}(\exp\{\beta \mathrm{Cov}(S)\})} \exp\{\beta \mathrm{Cov}(S)\}$

$\quad S \leftarrow$ 1D-Filter$(\{(x_i - \mu_{\mathrm{emp}}(S))^T V (x_i - \mu_{\mathrm{emp}}(S))\}_{i \in S})$

- Second challenge:
  - 1D-Filter has high sensitivity
  - each sample is independently filtered with probability proportional to $\tau_i \triangleq (x_i - \mu_{\mathrm{emp}}(S))^T V (x_i - \mu_{\mathrm{emp}}(S))$



Two main challenges in making filtering algorithms private

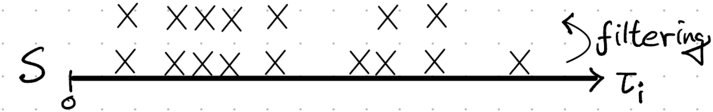
Algorithm Quantum robust mean estimation [Dong,Hopkins,Li,2019]

while $\|\mathrm{Cov}(S) - \mathbf{I}\| > c\,\alpha$ **do**

$\quad V \leftarrow \frac{1}{\mathrm{Trace}(\exp\{\beta \mathrm{Cov}(S)\})} \exp\{\beta \mathrm{Cov}(S)\}$

$\quad S \leftarrow$ 1D-Filter$(\{(x_i - \mu_{\mathrm{emp}}(S))^T V (x_i - \mu_{\mathrm{emp}}(S))\}_{i \in S})$

- Second challenge:
  - 1D-Filter has high sensitivity
  - each sample is independently filtered with probability proportional to $\tau_i \triangleq (x_i - \mu_{\mathrm{emp}}(S))^T V (x_i - \mu_{\mathrm{emp}}(S))$



Two main challenges in making filtering algorithms private

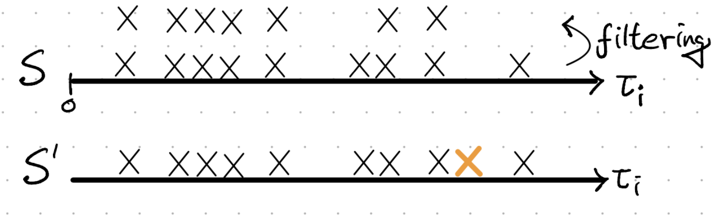
Algorithm Quantum robust mean estimation [Dong,Hopkins,Li,2019]

while $\|\mathrm{Cov}(S) - \mathbf{I}\| > c\,\alpha$ **do**

$\quad V \leftarrow \frac{1}{\mathrm{Trace}(\exp\{\beta \mathrm{Cov}(S)\})} \exp\{\beta \mathrm{Cov}(S)\}$

$\quad S \leftarrow$ 1D-Filter$(\{(x_i - \mu_{\mathrm{emp}}(S))^T V (x_i - \mu_{\mathrm{emp}}(S))\}_{i \in S})$

- Second challenge:
  - 1D-Filter has high sensitivity
  - each sample is independently filtered with probability proportional to $\tau_i \triangleq (x_i - \mu_{\mathrm{emp}}(S))^T V (x_i - \mu_{\mathrm{emp}}(S))$



Two main challenges in making filtering algorithms private

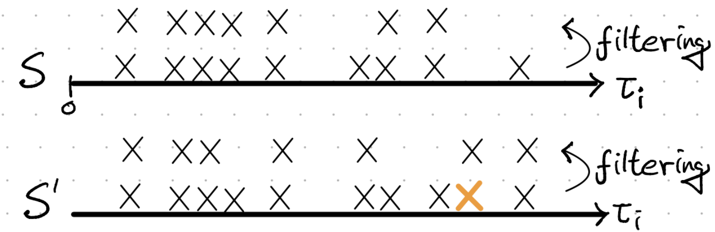
Algorithm Quantum robust mean estimation [Dong,Hopkins,Li,2019]

while $\|\mathrm{Cov}(S) - \mathbf{I}\| > c\,\alpha$ **do**

$\quad V \leftarrow \frac{1}{\mathrm{Trace}(\exp\{\beta \mathrm{Cov}(S)\})} \exp\{\beta \mathrm{Cov}(S)\}$

$\quad S \leftarrow$ 1D-Filter$(\{(x_i - \mu_{\mathrm{emp}}(S))^T V (x_i - \mu_{\mathrm{emp}}(S))\}_{i \in S})$

- Second challenge:
  - 1D-Filter has high sensitivity
  - each sample is independently filtered with probability proportional to $\tau_i \triangleq (x_i - \mu_{\mathrm{emp}}(S))^T V (x_i - \mu_{\mathrm{emp}}(S))$



Two datasets lead to independent filtering, and sensitivity blows up

Two main challenges in making filtering algorithms private

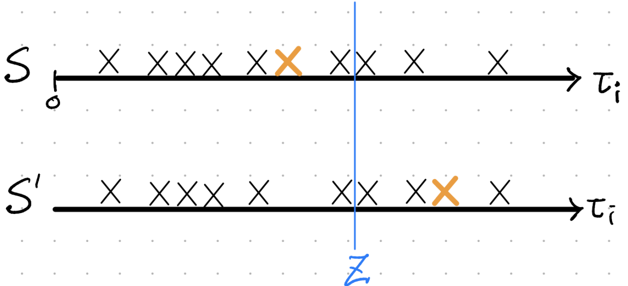
Algorithm Quantum robust mean estimation [Dong,Hopkins,Li,2019]

while $\|\mathrm{Cov}(S) - \mathbf{I}\| > c\,\alpha$ **do**

$\quad V \leftarrow \frac{1}{\mathrm{Trace}(\exp\{\beta \mathrm{Cov}(S)\})} \exp\{\beta \mathrm{Cov}(S)\}$

$\quad S \leftarrow$ 1D-Filter$(\{(x_i - \mu_{\mathrm{emp}}(S))^T V (x_i - \mu_{\mathrm{emp}}(S))\}_{i \in S})$

- Solution:
  - Use a single random threshold $Z \sim \mathrm{Uniform}[0, \rho]$, and filter samples above Z
  - this preserves the sensitivity to be one

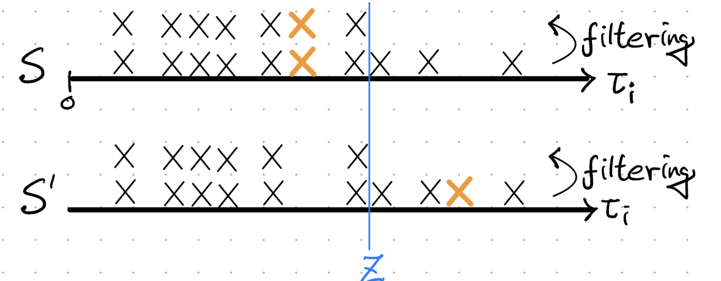


Two main challenges in making filtering algorithms private

Algorithm Quantum robust mean estimation [Dong,Hopkins,Li,2019]

1: **while** $\|\mathrm{Cov}(S) - \mathbf{I}\| > c\,\alpha$ **do**
2: $\quad V \leftarrow \frac{1}{\mathrm{Trace}(\exp\{\beta \mathrm{Cov}(S)\})} \exp\{\beta \mathrm{Cov}(S)\}$
3: $\quad S \leftarrow$ 1D-Filter$(\{(x_i - \mu_{\mathrm{emp}}(S))^T V (x_i - \mu_{\mathrm{emp}}(S))\}_{i \in S})$

- Solution:
  - Use a single random threshold $Z \sim \mathrm{Uniform}[0, \rho]$, and filter samples above Z
  - this preserves the sensitivity to be one



PRIME: PRIvate and robust Mean Estimation

- Run private histogram to get a bounding hypercube
- While $\|\tilde{\Sigma} - \mathbf{I}\| > c\,\alpha$
  - $\tilde{\mu} \leftarrow \mu_{\text{emp}}(S) + \mathcal{N}\left(0, \left(\frac{d^{1/2}\sqrt{\log(1/\delta)}}{n\varepsilon}\right)^2 \mathbf{I}_{d\times d}\right)$
  - $\tilde{\Sigma} \leftarrow \text{Cov}(S) + \mathcal{N}\left(0, \left(\frac{d\sqrt{\log(1/\delta)}}{n\varepsilon}\right)^2 \mathbf{I}_{d^2\times d^2}\right)$
  - $V \leftarrow \frac{1}{\text{Trace}(\exp\{\beta\tilde{\Sigma}\})} \exp\{\beta\tilde{\Sigma}\}$
  - $\rho \leftarrow$ DP-threshold$(\{(x_i - \tilde{\mu})^T V (x_i - \tilde{\mu})\}_{i\in S})$
  - $Z \leftarrow \text{Uniform}[0, \rho]$
  - $S \leftarrow$ 1D-Filter$(\{(x_i - \tilde{\mu})^T V (x_i - \tilde{\mu})\}_{i\in S}, Z)$

Theorem. [Liu,Kong,Kakade,O.,2021,NeurIPS]

PRIME is (ε, δ)-differentially private. For an α-corruption of n i.i.d. samples from a sub-Gaussian distribution with identity covariance matrix, with high probability

$$\|\hat{\mu} - \mu\| \;\lesssim\; \sqrt{\frac{d}{n}} + \alpha + \frac{d^{3/2}}{\varepsilon\, n}.$$

Mean estimation under sub-Gaussian distributions with identity covariance

	Error $\|\hat{\mu} - \mu\|$	
no corruption or privacy	$\sqrt{\frac{d}{n}}$	
α-corruption	$\sqrt{\frac{d}{n}} + \alpha$	[Diakonikolas et al.,2017]
(ε, δ)-DP	$\sqrt{\frac{d}{n}} + \frac{d}{\varepsilon\, n}$	[KamathLiSinghalUllman.,2019]
α-corruption and (ε, δ)-DP	$\sqrt{\frac{d}{n}} + \alpha + \frac{d^{3/2}}{\varepsilon\, n}$ (SVD time)	[LiuKongKakadeO.,2021]

There is a $d^{1/2}$ gap between PRIME and lower bound!

Where does $\frac{d^{3/2}}{\varepsilon n}$ come from?

- Sample complexity bottleneck: we need to privately compute

$$\tilde{\Sigma} \leftarrow \mathsf{Cov}(S) + W$$

- Best known algorithm adds i.i.d. entry Gaussian matrix $W \in \mathbb{R}^{d \times d}$ with $\mathcal{N}(0, (\frac{d\sqrt{\log 1/\delta}}{\varepsilon n})^2)$ to the covariance matrix
- The spectral norm perturbation is $\|W\|_{\text{spectral}} = O(\frac{d^{3/2}}{\varepsilon n})$

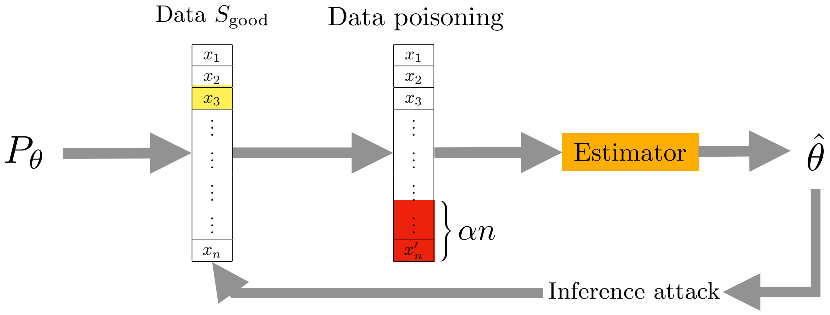
- In general, this cannot be improved as it matches a known lower bound [Dwork,Talwar,Thakurta,Zhang,2014]

Minimax optimal mean estimation

	Error $\|\hat{\mu} - \mu\|$	
no corruption or privacy	$\sqrt{\frac{d}{n}}$	
α-corruption	$\sqrt{\frac{d}{n}} + \alpha$	[Diakonikolas et al.,2017]
(ε, δ)-DP	$\sqrt{\frac{d}{n}} + \frac{d}{\varepsilon\, n}$	[KamathLiSinghalUllman.,2019]
α-corruption and (ε, δ)-DP	$\sqrt{\frac{d}{n}} + \alpha + \frac{d^{3/2}}{\varepsilon\, n}$ (SVD time) $\sqrt{\frac{d}{n}} + \alpha + \frac{d}{\varepsilon\, n}$ (exponential time)	[LiuKongKakadeO.,2021]

There is no extra *statistical* cost in requiring robustness and privacy simultaneously.

High-dimensional Propose-Test-Release



What is the fundamental connection between robust estimators and DP estimators?

High-dimensional Propose-Test-Release

- General framework for solving (inefficiently) statistical estimation problems with (ε, δ)-DP guarantee
- as a byproduct, we get robustness against α-corruption for free
- gives optimal sample complexity for mean estimation, covariance estimation, linear regression, and principal component analysis

HPTR step 1: design the score function

- Problem instance:
  mean estimation with i.i.d. samples from a sub-Gaussian distribution with mean μ and covariance Σ with error metric
  $$\|\Sigma^{-1/2}(\hat{\mu} - \mu)\|$$

HPTR step 1: design the score function

- Problem instance:
  mean estimation with i.i.d. samples from a sub-Gaussian distribution with mean μ and covariance Σ with error metric

$$\|\Sigma^{-1/2}(\hat{\mu} - \mu)\|$$

- Polynomial-time [Kamath,Mouzakis,Singhal,Steinke,Ullman,2021]:
  if $n \geq d^{5/2}/\varepsilon$

$$\|\Sigma^{-1/2}(\hat{\mu} - \mu)\| \ \leq \ \sqrt{\frac{d}{n}} + \frac{d}{\varepsilon n}$$

- Exponential-time [Brown,Gaboardi,Smith,Ullman,Zakynthinou,2021]:

$$\|\Sigma^{-1/2}(\hat{\mu} - \mu)\| \ \leq \ \sqrt{\frac{d}{n}} + \frac{d}{\varepsilon^2 n}$$

- Lower bound [Barber,Duchi,2014]:

$$\min_{\hat{\mu} \in \mathcal{F}_{\varepsilon,\delta}} \ \max_{P_{\mu,\Sigma}} \mathbb{E}\big[\, \|\Sigma^{-1/2}(\hat{\mu} - \mu)\| \,\big] \ \geq \ \sqrt{\frac{d}{n}} + \frac{d}{\varepsilon n}$$

HPTR step 1: design the score function

- Problem instance:
  mean estimation with i.i.d. samples from a sub-Gaussian distribution with mean μ and covariance Σ with error metric

$$\|\Sigma^{-1/2}(\hat{\mu} - \mu)\|$$

HPTR step 1: design the score function

- Problem instance:
  mean estimation with i.i.d. samples from a sub-Gaussian distribution with mean μ and covariance Σ with error metric

$$
\begin{aligned}
\|\Sigma^{-1/2}(\hat{\mu}-\mu)\| &= \max_{\|v\|=1} v^T \Sigma^{-1/2}(\hat{\mu}-\mu) \\
&= \max_{\|v\|=1} \frac{v^T \hat{\mu} - \overbrace{v^T \mu}^{\mu_v}}{\underbrace{\sqrt{v^T \Sigma v}}_{\sigma_v}}
\end{aligned}
$$

HPTR step 1: design the score function

- Problem instance:
  mean estimation with i.i.d. samples from a sub-Gaussian distribution with mean μ and covariance Σ with error metric

$$
\begin{aligned}
\|\Sigma^{-1/2}(\hat{\mu} - \mu)\| &= \max_{\|v\|=1} v^T \Sigma^{-1/2}(\hat{\mu} - \mu) \\
&= \max_{\|v\|=1} \frac{v^T \hat{\mu} - \overbrace{v^T \mu}^{\mu_v}}{\underbrace{\sqrt{v^T \Sigma v}}_{\sigma_v}}
\end{aligned}
$$

- Design empirical loss function:

$$
D_S(\hat{\mu}) = \max_{\|v\|=1} \frac{v^T \hat{\mu} - \mu_v^{\text{robust}}}{\sigma_v^{\text{robust}}}
$$

HPTR step 2: sensitivity analysis

We want to minimize the loss function:

$$D_S(\hat{\mu}) \;=\; \max_{\|v\|=1} \frac{v^T \hat{\mu} - \mu_v^{\text{robust}}}{\sigma_v^{\text{robust}}}$$

- To stochastically minimize this robust empirical loss, we want to sample from (exponential mechanism*)

$$\hat{\mu} \;\sim\; \frac{1}{Z} \exp\Big\{ -\frac{\varepsilon}{2\Delta} D_S(\hat{\mu}) \Big\}$$

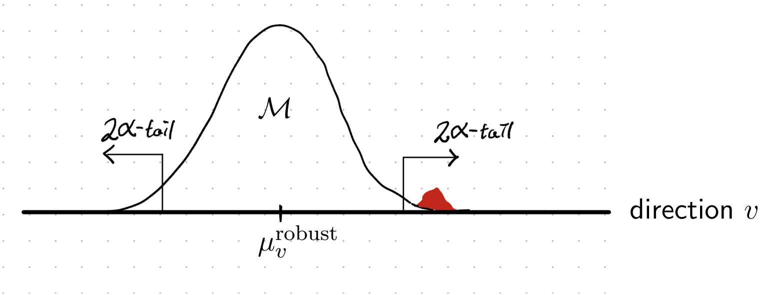
- If Δ is the sensitivity, then this is $(\varepsilon, 0)$-differentially private
- **The sensitivity of $D_S(\hat{\mu})$ dramatically reduces if we use 1-d robust statistics**
- Key ingredient is resilience property

---

* [McSherry,Talwar,2007]

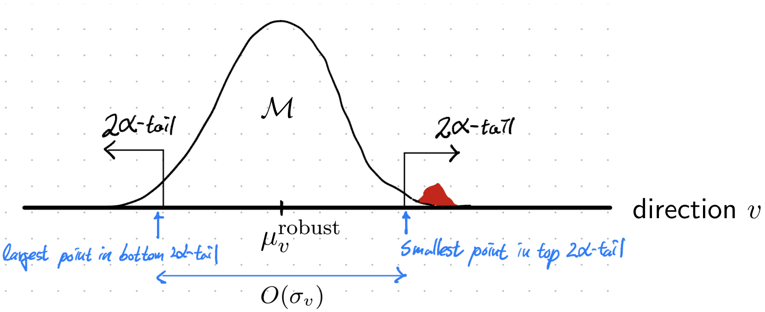
HPTR step 2: sensitivity analysis

- $\mu_v^{\text{robust}} = \frac{1}{|\mathcal{M}|}\sum_{\mathcal{M}} v^T x_i$ has sensitivity $\Delta = \frac{\sigma_v}{n}$



HPTR step 2: sensitivity analysis

- $\mu_v^{\text{robust}} = \frac{1}{|\mathcal{M}|}\sum_{\mathcal{M}} v^T x_i$ has sensitivity $\Delta = \frac{\sigma_v}{n}$



Resilience property for sub-Gaussian [Steinhardt,Charikar,Valiant,2018]

Given n i.i.d. sub-Gaussian samples S with $n \geq d/\alpha^2$, for all $S' \subset S$ of size at least αn,

$$\left| v^T(\mu(S) - \mu(S')) \right| \;\leq\; \sigma_v \;.$$

High-dimensional Propose-Test-Release*

- HPTR(S)

Propose : Propose $\Delta = O(1/n)$ based on the resilience of the distribution

Test : Privately test the sensitivity for all neighboring dataset S'

Release : If S passes the test, release $\hat{\mu}$ sampled from

$$\hat{\mu} \quad \sim \quad \frac{1}{Z} \exp\Big\{ -\frac{\varepsilon}{2\Delta} D_S(\hat{\mu}) \Big\}$$

*inspired by original PTR [Dwork,Lei,2009] and a more advanced PTR [Brown,Gaboardi,Smith,Ullman,Zakynthinou,2021]

Generality of HPTR

- HPTR can be applied to any statistical estimation problem to achieve the near-optimal error rate under (ε, δ)-DP
  - sub-Gaussian mean estimation:
  $$\|\Sigma^{-1/2}(\hat{\mu} - \mu)\| = O\Big(\sqrt{\frac{d}{n}} + \frac{d}{\varepsilon n}\Big)$$
  - k-th moment bounded mean estimation:
  $$\|\Sigma^{-1/2}(\hat{\mu} - \mu)\| = O\Big(\sqrt{\frac{d}{n}} + \Big(\frac{d}{\varepsilon n}\Big)^{1-\frac{1}{k}}\Big)$$
  - sub-Gaussian linear regression:
  $$\|\Sigma^{1/2}(\hat{\beta} - \beta)\| = O\Big(\sqrt{\frac{d}{n}} + \frac{d}{\varepsilon n}\Big)$$
  - Gaussian covariance estimation:
  $$\|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - \mathbf{I}\|_F = O\Big(\sqrt{\frac{d^2}{n}} + \frac{d^2}{\varepsilon n}\Big)$$
  - sub-Gaussian principal component analysis:
  $$1 - \frac{\hat{v}^\top \Sigma \hat{v}}{\|\Sigma\|} = O\Big(\sqrt{\frac{d}{n}} + \frac{d}{\varepsilon n}\Big)$$

Conclusion and open questions

- First half of the talk, we gave the first efficient algorithm that achieves both differential privacy and robustness:

$$\|\hat{\mu} - \mu\| \;\le\; \sqrt{\frac{d}{n}} + \alpha + \frac{d^{1.5}}{\varepsilon\, n}$$

$$\|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - \mathbf{I}\|_F \;\le\; \sqrt{\frac{d^2}{n}} + \alpha + \frac{d^3}{\varepsilon\, n}$$

  - Can we have an efficient algorithm that closes the $d^{1/2}$ gap (for mean)?
  - Can we use it to make DP-SGD robust?
  - Can we use it to make defense against backdoor attacks (such as SPECTRE) also private?
  - Can we design efficient algorithms for other problems:
    - Principal component analysis, linear regression, convex optimization

Conclusion and open questions

- Second half of the talk, we introduced HPTR that achieves optimal error rate on mean estimation, covariance estimation, linear reression, and PCA
  - Characterize fundamental tradeoffs in structured data (sparsity and low-rank)
  - Characterize fundamental tradeoffs in discrete or graph data

- arXiv:2102.09159, Xiyang Liu, Weihao Kong, Sham Kakade, Sewoong Oh "Robust and Differentially Private Mean Estimation"
- arxiv:2111.06578, Xiyang Liu, Weihao Kong, Sewoong Oh "Differential Privacy and Robust Statistics in High Dimensions"
- arXiv:2104.11315, Jonathan Hayase, Weihao Kong, Raghav Somani, S. Oh "SPECTRE: Defending Against Backdoor Attacks Using Robust Covariance Estimation"