# A Black-Box Transformation from Robustness to Privacy

**Lydia Zakynthinou**

UC Berkeley

Based on works by Hilal Asi, Jonathan Ullman, Z, Sam Hopkins, Gautam Kamath,  Mahbod Majid, Shyam Narayanan

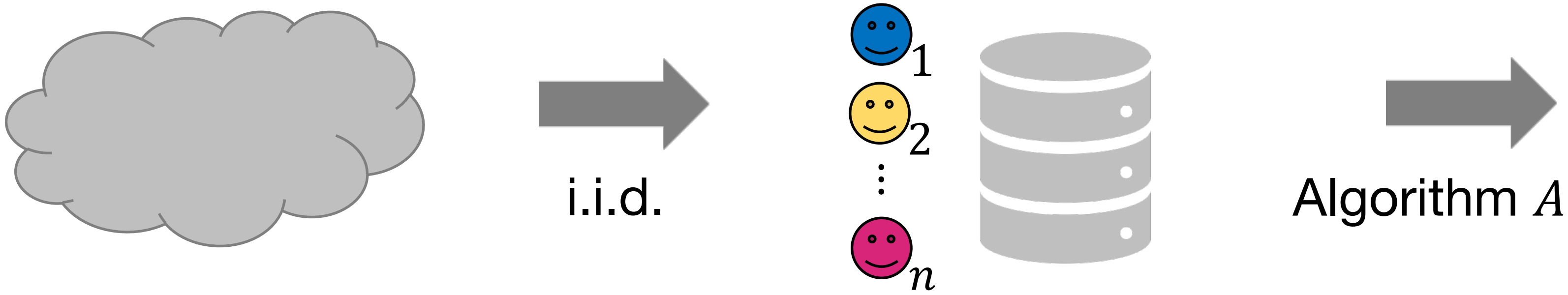
# Outline

- Definitions of Differential Privacy and Robustness
- Prior work (PTR)
- A black-box transformation from robust to DP algorithms
    - Implications
    - Applications
- Summary

# Parameter Estimation

Population $p_{\theta^*} \in \mathcal{P}$,
$\theta^* \in \Theta \subseteq \mathbb{R}^d$

Sample $X = (X_1, \dots, X_n)$




Algorithm $A(X)$

Parameter estimate $\hat{\theta}$

**Accuracy goal** : $\|\hat{\theta} - \theta^*\| \leq \alpha$ w.p. $1 - \beta$

Differential Privacy: Do not leak too much information about the sample $X$.
[Dwork McSherry Nissim Smith 2006]

Robustness: Be accurate even under data corruptions or model misspecification.
[Tukey, Huber '60s]

# **Differential Privacy** [Dwork McSherry Nissim Smith 2006]

Sample $X = (X_1, \dots, X_n)$


Algorithm $A_{priv}(X)$

Sample $X' = (X'_1, \dots, X_n)$

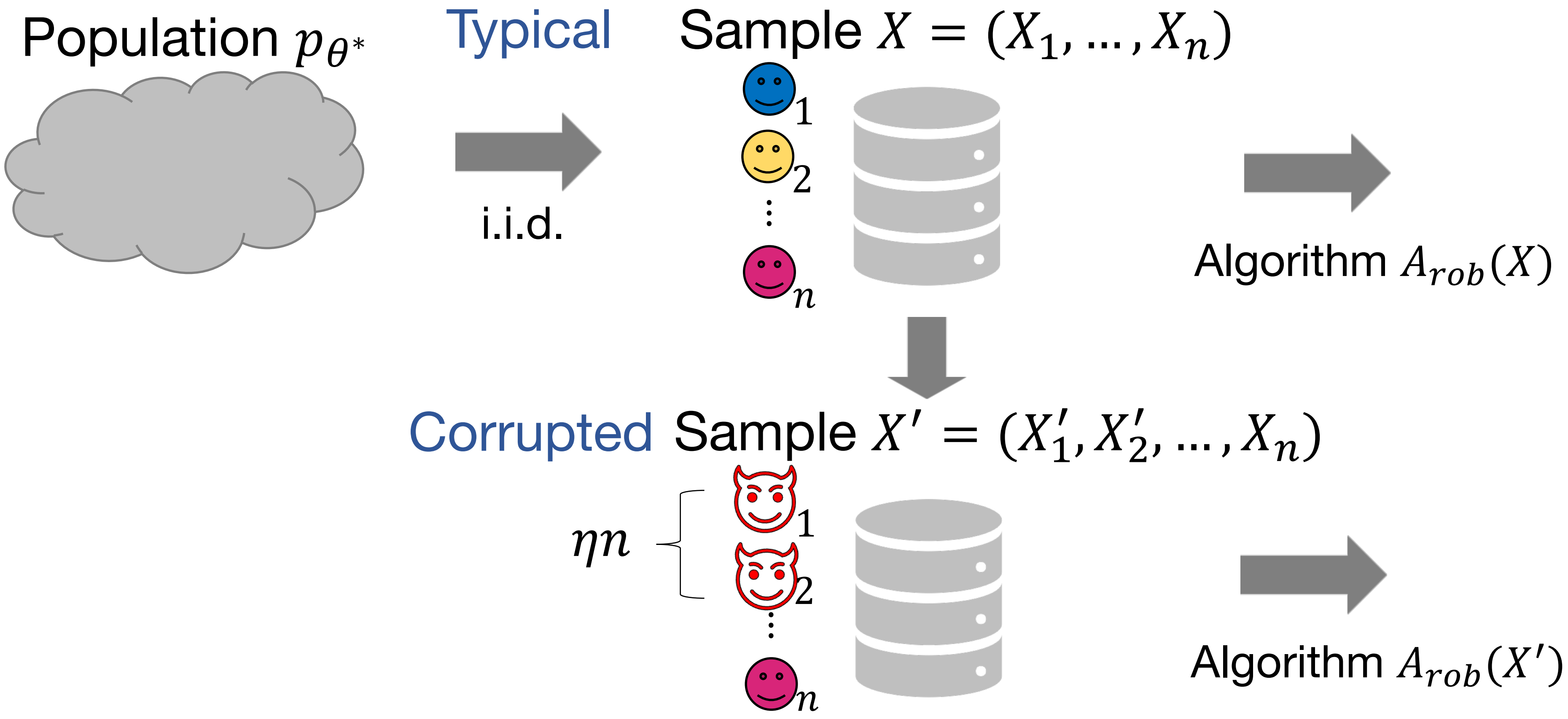

Algorithm $A_{priv}(X')$

| $\delta \neq 0$: "approx" | $\delta = 0$: "pure", $\varepsilon$-DP |
|---|---|

**Def.** Algorithm $A_{priv}\colon \mathcal{X}^n \to \mathcal{W}$ is $(\varepsilon, \delta)$-differentially private (DP) if for all datasets $X, X'$ with $Ham(X, X') = 1$ and all measurable sets $W \subseteq \mathcal{W}$,

$$\Pr\left[A_{priv}(X) \in W\right] \le e^{\varepsilon} \Pr[A_{priv}(X') \in W] + \delta$$

# Robustness




**Def.** Algorithm $A_{rob}: \mathcal{X}^n \to \mathcal{W}$ is $\eta$-robust with accuracy $\alpha(\eta)$ if given $X \sim p_{\theta^*}^n$, with high probability, for all $X'$ differing on at most $\eta n$ points,

$$\|A_{rob}(X') - \theta^*\| \le \alpha(\eta).$$

# History of connection between DP+Robustness

- [Dwork Lei 2009]: Propose-Test-Release (PTR)

- Lots of recent works had given private estimators “inspired” by robust ones [Bun Kamath Steinke Wu 2019], [Kamath Singhal Ullman 2020], [Ramsay Chenouri 2021], [Liu Kong Kakade Oh 2021], [Brown Gaboardi Smith Ullman **Z** 2021], [Liu Kong Oh 2022], [Hopkins Kamath Majid 2022], [Kothari Manurangsi Velingker 2022]

# Very high-level: PTR [Dwork Lei 2009]

Release $f(X) + \text{Laplace}\left(\frac{\Delta_f}{\varepsilon}\right)$.

$f(X)$ good estimator of $\theta$

**Def.**
Global Sensitivity of function $f: \mathcal{X}^n \to \mathbb{R}$ :

$$\Delta_f = \max \{|f(X) - f(X')| \ for\ X, X' \colon Ham(X, X') = 1\}$$

Local Sensitivity of function $f: \mathcal{X}^n \to \mathbb{R}$ on dataset $X$ :

$$\Delta_f(X) = \max \{|f(X) - f(X')| \ for\ X' \colon Ham(X, X') = 1\}$$

But $\Delta_f \geq \Delta_f(X)$…

# PTR: Why does robustness help privacy

Propose local sensitivity bound $B$.

Test Let $\gamma = \min_{X'}\{Ham(X, X'): \Delta_f(X') > B\}$. If $\gamma + \text{Laplace}\left(\frac{1}{\varepsilon}\right) \leq \frac{\log(1/\delta)}{\varepsilon}$, abort.

Release $\tilde{f}(X) = f(X) + \text{Laplace}\left(\frac{B}{\varepsilon}\right)$.

✓ Propose-Test-Release is $(\varepsilon, \delta)$-DP.

✓ If it passes the test, it has error $\left|\tilde{f}(X) - f(X)\right| \lesssim \frac{B}{\varepsilon}$.

# PTR: Why does robustness help privacy

Propose local sensitivity bound $B$.

Test Let $\gamma = \min_{X'}\{Ham(X, X'): \Delta_f(X') > B\}$. If $\gamma + \mathrm{Laplace}\left(\frac{1}{\varepsilon}\right) \leq \frac{\log(1/\delta)}{\varepsilon}$, abort.

Release $\tilde{f}(X) = f(X) + \mathrm{Laplace}\left(\frac{B}{\varepsilon}\right)$.

Let's apply this to learning the Gaussian mean $\mathcal{N}(\theta^*, 1)$!

- First try: $f(X) = \frac{1}{n}\sum_{i\in[n]} X_i$. Then $\Delta_f(X) = \infty$ and $\gamma = 0$, even for $X \sim \mathcal{N}(\theta^*, 1)^n$.

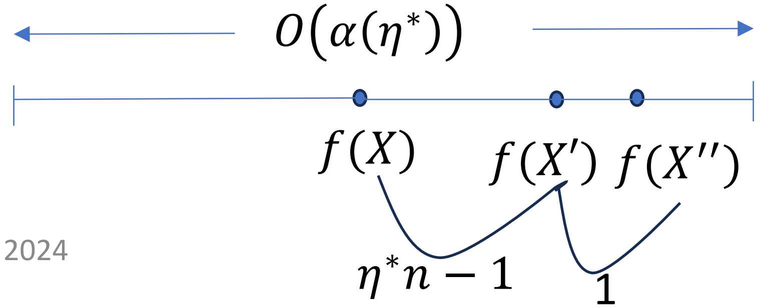
# PTR: Why does robustness help privacy

**Propose** local sensitivity bound $B$.

**Test** Let $\gamma = \min_{X'}\{Ham(X, X') \colon \Delta_f(X') > B\}$. If $\gamma + \text{Laplace}\left(\frac{1}{\varepsilon}\right) \leq \frac{\log(1/\delta)}{\varepsilon}$, abort.

**Release** $\tilde{f}(X) = f(X) + \text{Laplace}\left(\frac{B}{\varepsilon}\right)$.

- Better: Choose $f(X)$ to be an $\eta$-robust estimator of $\theta^*$ with accuracy $\alpha(\eta) = \eta + \frac{1}{\sqrt{n}}$. Set $B = O\big(\alpha(\eta^*)\big)$ for $\eta^* n \approx \frac{\log(1/\delta\beta)}{\varepsilon} + 1$.
  If $X \sim \mathcal{N}(\theta^*, 1)^n$ then whp $\Delta_f(X') \leq O\big(\alpha(\eta^*)\big) = B$ and we will pass the test with overall error $O\left(\frac{1}{\varepsilon^2 n} + \frac{1}{\varepsilon\sqrt{n}}\right)$ 🎉



# History of connection between DP+Robustness

Can we always transform robust estimators to DP ones?

- [Dwork Lei 2009]: Can be used as a black-box transformation from robust to $(\varepsilon, \delta)$-DP but it incurs extra factors.
- [Nissim Raskhodnikova Smith 2007] Smooth sensitivity: Also incurs extra factors.
- [Liu Kong Oh 2022]: Framework which gives statistically optimal estimators for many tasks under $(\varepsilon, \delta)$-DP via generalization of Restricted Exponential Mechanism ([Brown Gaboardi Smith Ullman **Z** 2021] used REM with Tukey depth as a score function) but not black-box.

[Asi Ullman **Z** 2023], [Hopkins Kamath Majid Narayanan 2023]:
A black-box transformation from any robust to a DP algorithm with optimal rates for several canonical tasks.

# Outline

- Definitions of Differential Privacy and Robustness
- Prior Work (PTR)
- A black-box transformation from robust to DP algorithms
  - Implications
  - Applications
- Summary

# A black-box transformation : Robust→Private

**Theorem** [Asi Ullman **Z** 2023]
Let $\varepsilon, \eta_0, \beta \in (0,1)$, $n \in \mathbb{N}$, distribution $p_{\theta^*}$ for $\theta^* \in \Theta \subseteq \mathcal{B}^d_{||\cdot||}(R)$. Let $A_{rob}: \mathcal{X}^n \to \Theta$ be an $\eta$-robust estimator of $\theta^*$ with accuracy $\alpha(\eta)$ wp $1-\beta$. Let $\eta^* \geq \eta_0$ such that

$$\eta^* \approx \frac{d \log(R/\alpha(\eta_0)) + \log(1/\beta)}{\varepsilon n}$$

Then there exists an $\varepsilon$-DP estimator $A_{priv}$ of $\theta^*$ with accuracy $O\big(\alpha(\eta^*)\big)$ wp $1 - \mathrm{O}(\beta)$.

**Theorem** [Hopkins Kamath Majid Narayanan 2023]
Let $\varepsilon, \eta_0, \beta \in (0,1)$, $n \in \mathbb{N}$, distribution $p_{\theta^*}$ for $\theta^* \in \Theta \subseteq \mathcal{B}^d_{||\cdot||}(R)$. Let $A_{rob}: \mathcal{X}^n \to \Theta$ be an $\eta$-robust estimator of $\theta^*$ with accuracy $\alpha(\eta)$ wp $1-\beta$. Then there exists an $\varepsilon$-DP estimator $A_{priv}$ of $\theta$ with accuracy $O\big(\alpha(\eta_0)\big)$ wp $1 - \mathrm{O}(\beta)$ as long as

$$n \gtrsim \max_{\eta^* \in [\eta_0, 1]} \frac{d \log \frac{2\alpha(\eta^*)}{\alpha(\eta_0)} + \log \frac{1}{\beta}}{\eta^* \varepsilon}$$

# A black-box transformation : Robust→Private

**Theorem** [Asi Ullman **Z** 2023]
Let $\varepsilon, \eta_0, \beta \in (0,1)$, $n \in \mathbb{N}$, distribution $p_{\theta^*}$ for $\theta^* \in \Theta \subseteq \mathcal{B}^d_{||\cdot||}(R)$. Let $A_{rob}: \mathcal{X}^n \to \Theta$ be an $\eta$-robust estimator of $\theta^*$ with accuracy $\alpha(\eta)$ wp $1-\beta$. Let $\eta^* \geq \eta_0$ such that

$$\eta^* \approx \frac{d \log(R/\alpha(\eta_0)) + \log(1/\beta)}{\varepsilon n}$$

Then there exists an $\varepsilon$-DP estimator $A_{priv}$ of $\theta^*$ with accuracy $O\big(\alpha(\eta^*)\big)$ wp $1 - O(\beta)$.

**Theorem** [Hopkins Kamath Majid Narayanan 2023]
Let $\varepsilon, \eta_0, \beta \in (0,1)$, $n \in \mathbb{N}$, distribution $p_{\theta^*}$ for $\theta^* \in \Theta \subseteq \mathcal{B}^d_{||\cdot||}(R)$. Let $A_{rob}: \mathcal{X}^n \to \Theta$ be an $\eta$-robust estimator of $\theta^*$ with accuracy $\alpha(\eta)$ wp $1-\beta$. Then there exists an $\varepsilon$-DP estimator $A_{priv}$ of $\theta$ with accuracy $O\big(\alpha(\eta_0)\big)$ wp $1 - O(\beta)$ as l

$$n \gtrsim \max_{\eta^* \in [\eta_0, 1]} \frac{d \log \frac{2\alpha(\eta^*)}{\alpha(\eta_0)} + \log \frac{1}{\beta}}{\eta^* \varepsilon}$$

$$n \geq \frac{d + \log \frac{1}{\beta}}{\eta_0 \varepsilon} + \frac{d \log(R/\alpha(\eta_0))}{\varepsilon}$$

$$\eta_0 = \alpha, \alpha(\eta) \begin{cases} = \alpha + \eta, & \eta < 1/2 \\ R, & o.w. \end{cases}$$

$$\Rightarrow n \geq \frac{d + \log \frac{1}{\beta}}{\alpha \varepsilon} + \frac{d \log R}{\varepsilon}$$

# A black-box transformation : Robust→Private

Via the Inverse-Sensitivity mechanism $M^{\rho}_{Inv}(f\,; X)$ [Johnson Shmatikov 2013], [Asi Duchi 2020]

≡ Exponential mechanism [McSherry Talwar 2007] with the path-length score function

# Exponential Mechanism [McSherry Talwar 2007]

**Def.** Given dataset $X$, score function $score: \Theta \times \mathcal{X}^n \to \mathbb{R}$ with global sensitivity $\max_{\theta} \max_{X,X':Ham(X,X')=1} |score(\theta; X) - score(\theta, X')| \leq 1$, the exponential mechanism returns $\theta$ with probability

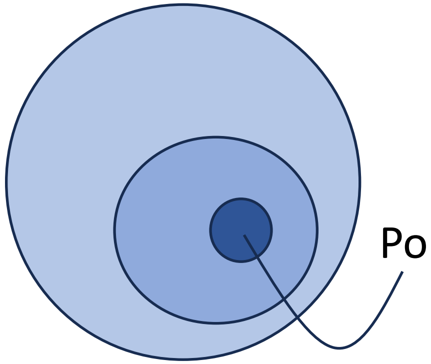
$$\pi_X(\theta) = \frac{e^{-\varepsilon \cdot score(\theta;X)}}{\int_\Theta e^{-\varepsilon \cdot score(\xi;X)} d\xi}$$

0 score: good, high score: bad

- ✓ Satisfies $\varepsilon$-DP.
- ✓ Returns $\theta_{priv}$ with $score(\theta_{priv}; X) \leq K$ with probability at least $1 - e^{-\varepsilon K} \frac{\text{Vol}(\Theta)}{\text{Vol}(\{\theta: score(\theta;X)=0\})}$

Wp $1-\beta$,

$$score(\theta_{priv}; X) \leq \frac{1}{\varepsilon}\left(\log \frac{Vol(\Theta)}{\text{Vol}(\{\theta:\ score(\theta; X) = 0\})} + \log \frac{1}{\beta}\right)$$



Points $\xi \in \Theta$ with low score are sampled whp

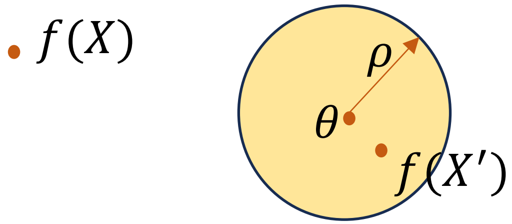
# (Smooth) Inverse Sensitivity Mechanism [Asi Duchi 2020]

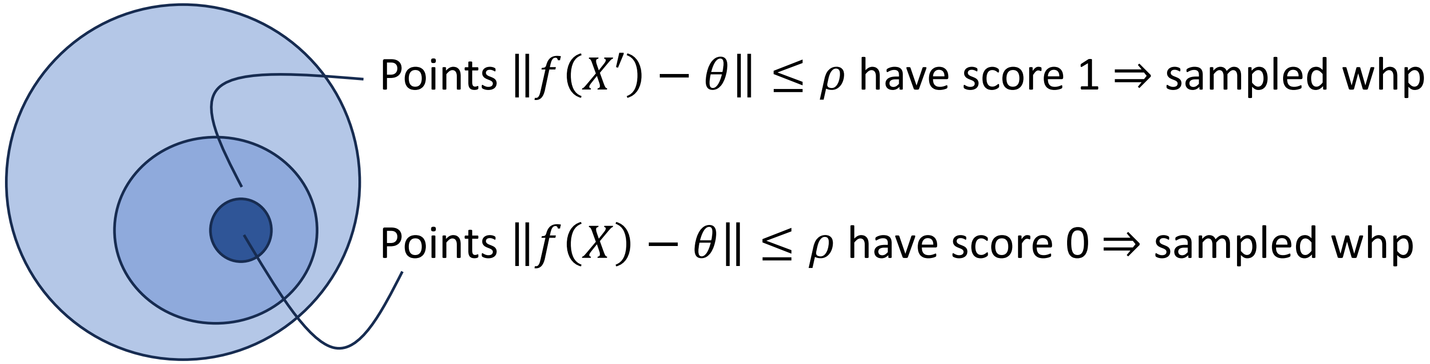
**Def.** Given function $f: \mathcal{X}^n \to \Theta$, dataset $X$, smoothness parameter $\rho$, $M_{Inv}^{\rho}(f; X)$ returns $\theta$ with probability

$$\pi_X(t) = \frac{e^{-\varepsilon \cdot len_f^{\rho}(\theta; X)}}{\int e^{-\varepsilon \cdot len_f^{\rho}(\xi; X)} d\xi},$$

where the score function is the smooth path-length

$$len_f^{\rho}(\theta; X) = \min_{X'} \{Ham(X, X') : \|f(X') - \theta\| \leq \rho\}$$



# (Smooth) Inverse Sensitivity Mechanism [Asi Duchi 2020]

**Def.** Given function $f: \mathcal{X}^n \to \Theta$, dataset $X$, smoothness parameter $\rho$, $M^{\rho}_{Inv}(f; X)$ returns $\theta$ with probability

$$\pi_X(t) = \frac{e^{-\varepsilon \cdot len^{\rho}_f(\theta; X)}}{\int e^{-\varepsilon \cdot len^{\rho}_f(\xi; X)} d\xi},$$

where the score function is the smooth path-length

$$len^{\rho}_f(\theta; X) = \min_{X'} \{Ham(X, X') : \|f(X') - \theta\| \le \rho\}$$

✓ **Theorem** [Asi Duchi 2020]: If $f: \mathcal{X}^n \to \mathcal{B}^d_{\|\cdot\|}(R + \rho)$ then $\forall X \in \mathcal{X}^n$, with probability $1 - \beta$,

$$\left\| M^{\rho}_{Inv}(f; X) - f(X) \right\| \le \omega_f(X; \eta^*) + \rho,$$

where $\omega_f(X; \eta^*) = \sup_{X'} \{\|f(X) - f(X')\| : Ham(X, X') \le \eta^* n\}$ and $\eta^* n \approx \frac{d \log \frac{R}{\rho} + \log \frac{1}{\beta}}{\varepsilon}$.

# A black-box transformation : Robust→Private

[AU**Z**23, HKMN23] Black-Box Transformation:
Sample a random $\theta_{priv} \in \Theta + \mathcal{B}^d_{||\cdot||}(\rho) \subseteq \mathcal{B}^d(R+\rho)$ with probability

$$\pi_X(\theta) \propto e^{-\varepsilon \cdot len_f^{\rho}(\theta;X)}$$

where $f = A_{rob}$, $\rho = \alpha(\eta_0)$. Whp $\|\theta_{priv} - \theta^*\| = O(\alpha(\eta^*))$ for $\eta^* \approx \frac{d\log\frac{R}{\rho} + \log\frac{1}{\beta}}{\varepsilon n}$.

**Proof**.

- By [Asi Duchi 2020] : $\|\theta_{priv} - A_{rob}(X)\| \leq \omega_{A_{rob}}(X;\eta^*) + \alpha(\eta_0)$ for $\eta^* \approx \frac{d\log\frac{R}{\rho} + \log\frac{1}{\beta}}{\varepsilon n}$.
- By robustness: $\omega_{A_{rob}}(X;\eta^*) \leq \sup_{X':Ham(X,X')\leq\eta^* n} \|A_{rob}(X) - A_{rob}(X')\| \leq 2\alpha(\eta^*)$.
- Overall: $\|\theta_{priv} - \theta^*\| \leq \|\theta_{priv} - A_{rob}(X)\| + \|A_{rob}(X) - \theta^*\| \leq 4\alpha(\eta^*)$ for $\eta^* \geq \eta_0$.

# A black-box transformation : Robust→Private

**Theorem** [Asi Ullman **Z** 2023] [Hopkins Kamath Majid Narayanan 2023]
Let $\varepsilon, \eta_0, \beta \in (0,1)$, $n \in \mathbb{N}$, distribution $p_{\theta^*}$ for $\theta^* \in \Theta \subseteq \mathcal{B}^d_{\|\cdot\|}(R)$. Let $A_{rob}: \mathcal{X}^n \to \Theta'$ be an $\eta$-robust estimator of $\theta^*$ with accuracy $\alpha(\eta)$ wp $1-\beta$. Let $\eta^* \geq \eta_0$ such that

$$\eta^* \approx \frac{d \log(R/\alpha(\eta_0)) + \log(1/\beta)}{\varepsilon n}$$

Then there exists an $\varepsilon$-DP estimator $A_{priv}$ of $\theta^*$ with accuracy $O\big(\alpha(\eta^*)\big)$ wp $1 - O(\beta)$.

Extend to $(\varepsilon, \delta)$-DP using PTR and a *truncated* inverse-sensitivity mechanism.

Replaced by $d + \log\left(\frac{1}{\delta}\right)$

$A_{priv}$: $\varepsilon$-DP estimator

Sample $X = (x_1, \dots, x_n)$

$A_{rob}$ : $\eta$-robust estimator with accuracy $\alpha(\eta)$

$A_{priv}(X)$ has accuracy $\approx \alpha\left(\frac{d \log(R/\alpha(\eta_0))}{\varepsilon n}\right)$

# Outline

- Definitions of Differential Privacy and Robustness
- Prior Work
- A black-box transformation from robust to DP algorithms
  - Implications
  - Applications
- Summary

# Implications [AUZ ‘23]

1. $\varepsilon$-DP and $\frac{\log n}{\varepsilon n}$ - robustness are equivalent for low-dimensional tasks.

**Theorem (informal).** For low-dimensional tasks ($d = O(1)$), under (natural) assumptions ( e.g., the non-private error is $\Omega(1/\mathrm{poly}(n))$ ),

**minimax error $\varepsilon$-DP $\approx$ minimax error $\eta$-robustness for $\eta = \frac{\log n}{\varepsilon n}$.**

Failure probabilities can be different $\propto 1/\mathrm{poly}(n)$ and $R$.

**Idea**: $\alpha^*_{rob}\left(\eta = \frac{\log n}{\varepsilon n}\right) \leq \alpha^*_{priv}(\varepsilon) \leq \alpha^*_{rob}\left(\eta = \frac{d \log n}{\varepsilon n}\right)$

[Dwork Lei 2009] [This work]

# Implications [AUZ '23]

1. $\varepsilon$-DP and $\frac{\log n}{\varepsilon n}$ - robustness are equivalent for low-dimensional tasks.
2. Our transformation is optimal for low-dimensional tasks.

**Theorem (informal).** For low-dimensional tasks ($d = O(1)$), there exists a robust algorithm to instantiate our transformation, such that the resulting private algorithm has **optimal minimax error up to constants**.

What about high-dimensional tasks?

# Applications [HKMN & AUZ '23]

(Near) Optimal private estimators in high dimensions for many statistical tasks, e.g.:

- Gaussian mean estimation,
- Gaussian covariance estimation,
- (Sub)Gaussian PCA [new for $\varepsilon$-DP],
- Gaussian linear regression [new for $\varepsilon$-DP]
- Sparse Gaussian linear regression [new for $\varepsilon$-DP] (via a slightly modified transformation).

Mahbod will fix this next!

A drawback: the transformation is computationally inefficient in general.

# Summary

- We give the first black-box transformation from robust to private estimators.
- We show that $\varepsilon$-privacy and $\frac{\log n}{\varepsilon n}$-robustness are equivalent for low-dim tasks.
- We show that the transformation gives optimal estimators in low-dim.
- And it often gives optimal estimators in high dimensions, including new near-optimal results for PCA and (sparse) linear regression.
- We extend it to $(\varepsilon, \delta)$-DP for $\tau \approx \frac{d+\log(1/\beta\delta)}{\varepsilon n}$, avoiding the dependence on $R$.

# Summary

- In general, the transformation is computationally inefficient.
  - [Asi Duchi 2020] give approximations for special cases (PCA, LR).
  - Using Sum-of-Squares-based techniques (as in [Hopkins Kamath Majid 2022]), [Hopkins Kamath Majid Narayanan 2023] show that if the score function satisfies some properties, then the transformation can be implemented in polynomial time (e.g., for Gaussian estimation).
- The dependence on $d, R$ is optimal in general (via lower bounds on applications). But it may be improved for special cases.
- When does the equivalence result hold for high dimensions?