Private Hypothesis Testing via Robustness

Audra McMillan

Boston University and Northeastern University

November 11, 2019

Fundamental Questions in Data Science

Estimation: What is the world like?

Testing: Is my understanding of the world correct?

Fundamental Questions in Data Science

Estimation: What is the world like?

How many people actually like kale?

Testing: Is my understanding of the world correct?

Fundamental Questions in Data Science

Estimation: What is the world like?

How many people actually like kale?

Do 0% of people actually like kale?

Testing: Is my understanding of the world correct?

Fundamental Questions in Data Science

Estimation: What is the world like?

How many people actually like kale?

Deep connections
in the non-private world

Do 0% of people actually like kale?

Testing: Is my understanding of the world correct?

Fundamental Questions in Data Science

Estimation: What is the world like?

Estimation algorithms
$\Downarrow$
testing algorithms

Deep connections
in the non-private world

Testing: Is my understanding of the world correct?

Fundamental Questions in Data Science

Estimation: What is the world like?

Estimation algorithms
$\Downarrow$
testing algorithms

Lower bounds for testing
$\Downarrow$
lower bounds for estimation

Deep connections
in the non-private world

Testing: Is my understanding of the world correct?

Fundamental Questions in Data Science

Estimation: What is the world like?

Estimation algorithms
$\Downarrow$
testing algorithms

Deep connections
in the non-private world

Lower bounds for testing
$\Downarrow$
lower bounds for estimation

Under some circumstances optimal algorithms for testing
$\Downarrow$
optimal algorithms for estimation.

Testing: Is my understanding of the world correct?

Fundamental Questions in Private Data Science

Private Estimation: What is the world like?

Private Testing: Is my understanding of the world correct?

Fundamental Questions in Private Data Science

Private Estimation: What is the world like?

How can we translate results?

Private Testing: Is my understanding of the world correct?

Fundamental Questions in Private Data Science

Private Estimation: What is the world like?

How can we translate results?

Private Testing: Is my understanding of the world correct?

Hypothesis Testing

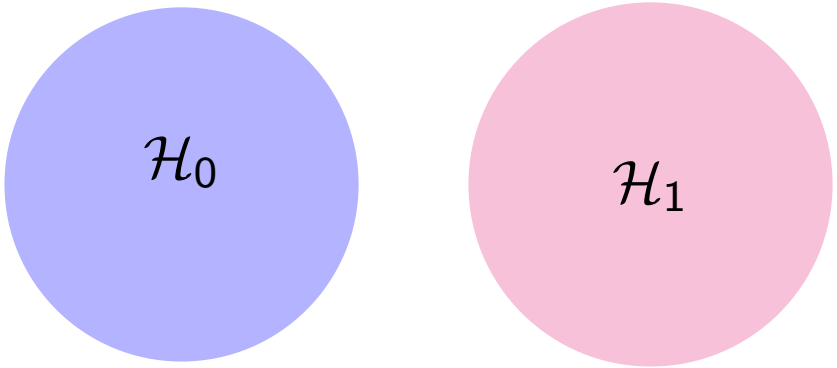
Do 0% of people actually like kale?

Hypothesis Testing

Do 0% of people actually like kale?

Given a space of distributions $\Delta(\Omega)$,

$$\underbrace{\mathcal{H}_0 \subset \Delta(\Omega)}_{\text{null hypothesis}} \quad \text{and} \quad \underbrace{\mathcal{H}_1 \subset \Delta(\Omega)}_{\text{alternate hypothesis}}$$



Hypothesis Test

Given $x_1, \ldots, x_n \sim R$, a hypothesis test determines with high probability whether $R \in \mathcal{H}_0$ or $R \in \mathcal{H}_1$.

Hypothesis Testing

A hypothesis test $T : \Omega^n \to \{\mathcal{H}_0, \mathcal{H}_1\}$ is an algorithm which given $X \sim R^n$ attempts to determine whether $R \in \mathcal{H}_0$ or $R \in \mathcal{H}_1$

The test $T : \Omega^n \to \{\mathcal{H}_0, \mathcal{H}_1\}$ distinguishes between $\mathcal{H}_0$ and $\mathcal{H}_1$ with sample complexity $SC^{\mathcal{H}_0,\mathcal{H}_1}(T)$ if for all $n \geq SC^{\mathcal{H}_0,\mathcal{H}_1}(T)$:

1. $\min_{R \in \mathcal{H}_0} \mathbb{P}_{X \sim R^n}[T(X) = \mathcal{H}_0] \geq 2/3$
2. $\min_{R \in \mathcal{H}_1} \mathbb{P}_{X \sim R^n}[T(X) = \mathcal{H}_1] \geq 2/3$

A test T has "optimal" sample complexity if for all tests T',

$$SC^{\mathcal{H}_0,\mathcal{H}_1}(T) = O(SC^{\mathcal{H}_0,\mathcal{H}_1}(T')).$$

Private Hypothesis Testing

A ϵ-DP hypothesis test $T : \Omega^n \to \{\mathcal{H}_0, \mathcal{H}_1\}$ is an algorithm which given $X \sim R^n$ attempts to determine whether $R \in \mathcal{H}_0$ or $R \in \mathcal{H}_1$ while maintaining the privacy of elements of the database.

The test $T : \Omega^n \to \{\mathcal{H}_0, \mathcal{H}_1\}$ distinguishes between $\mathcal{H}_0$ and $\mathcal{H}_1$ with sample complexity ${SC_\epsilon}^{\mathcal{H}_0,\mathcal{H}_1}(T)$ if for all $n \geq {SC_\epsilon}^{\mathcal{H}_0,\mathcal{H}_1}(T)$:

1. $\min_{R \in \mathcal{H}_0} \mathbb{P}_{X \sim R^n}[T(X) = \mathcal{H}_0] \geq 2/3$
2. $\min_{R \in \mathcal{H}_1} \mathbb{P}_{X \sim R^n}[T(X) = \mathcal{H}_1] \geq 2/3$
3. T is ϵ-DP

A test T has "optimal" sample complexity if for all ϵ-DP tests T',

$$ {SC_\epsilon}^{\mathcal{H}_0,\mathcal{H}_1}(T) = O({SC_\epsilon}^{\mathcal{H}_0,\mathcal{H}_1}(T')). $$

Differential privacy

Desirable: statistical tests stable under small changes in the data.

Reasons:

- **privacy**
- **generalisability under adaptive data analysis**

ϵ-differential privacy

A test $T : \Omega^n \to \{\mathcal{H}_0, \mathcal{H}_1\}$ is ϵ-**differentially private** (DP) if for all databases x and x' that differ on a single element, and all $b \in \{\mathcal{H}_0, \mathcal{H}_1\}$,

$$e^{-\epsilon}\mathbb{P}[T(x') = b] \leq \mathbb{P}[T(x) = b] \leq e^{\epsilon}\mathbb{P}[T(x') = b]$$

Differential privacy

Desirable: statistical tests stable under small changes in the data.

Reasons:

- **privacy**
- **generalisability under adaptive data analysis**

ϵ-differential privacy

A test $T : \Omega^n \to \{\mathcal{H}_0, \mathcal{H}_1\}$ is ϵ-**differentially private** (DP) if for all databases x and x' that differ on a single element, and all $b \in \{\mathcal{H}_0, \mathcal{H}_1\}$,

$$e^{-\epsilon}\mathbb{P}[T(x') = b] \leq \mathbb{P}[T(x) = b] \leq e^{\epsilon}\mathbb{P}[T(x') = b]$$

Other stability notions

Sample complexity is asymptotically the same for (see e,g,, [Acharya, Sun, Zhang '18]):

- ϵ-DP,
- ϵ-Total-variation stability (requires $\mathbb{P}[T(x) \in S] \leq \mathbb{P}[T(x') \in S] + \epsilon$)
- stability notions in-between (KL-stability, "concentrated DP")

Related work

- DP versions of popular statistical tests [Vu, Slavkovic '09, Uhler, Slavkovic, Feinberg '13, Wang, Lee, Kifer '15, Gaboardi, Lim, Rogers, Vadhan '16, Kifers, Rogers '17, Acharya, Sun, Zhang '18, Campbell, Bray, Ritz, Groce '18, Couch, Kazan, Shi, Bray, Groce '18a,b, Swanberg, Globus-Harris, Griffith, Ritz, Groce, Bray '18]
  - goodness-of-fit, closeness, independence
  - focus on small sample sizes

Related work

- DP versions of popular statistical tests [Vu, Slavkovic '09, Uhler, Slavkovic, Feinberg '13, Wang, Lee, Kifer '15, Gaboardi, Lim, Rogers, Vadhan '16, Kifers, Rogers '17, Acharya, Sun, Zhang '18, Campbell, Bray, Ritz, Groce '18, Couch, Kazan, Shi, Bray, Groce '18a,b, Swanberg, Globus-Harris, Griffith, Ritz, Groce, Bray '18]
  - goodness-of-fit, closeness, independence
  - focus on small sample sizes
- Asymptotic sample complexity of private testing [Cai, Daskalakis, Kamath '17, Aliakbarpour, Diakonikolas, Rubinfeld '18, Acharya, Sun, Zhang '18, Acharya, Kamath, Sun, Zhang '18]

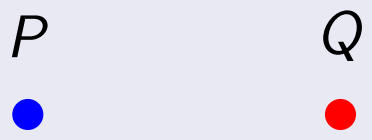
Related work

- DP versions of popular statistical tests [Vu, Slavkovic '09, Uhler, Slavkovic, Feinberg '13, Wang, Lee, Kifer '15, Gaboardi, Lim, Rogers, Vadhan '16, Kifers, Rogers '17, Acharya, Sun, Zhang '18, Campbell, Bray, Ritz, Groce '18, Couch, Kazan, Shi, Bray, Groce '18a,b, Swanberg, Globus-Harris, Griffith, Ritz, Groce, Bray '18]
  - goodness-of-fit, closeness, independence
  - focus on small sample sizes
- Asymptotic sample complexity of private testing [Cai, Daskalakis, Kamath '17, Aliakbarpour, Diakonikolas, Rubinfeld '18, Acharya, Sun, Zhang '18, Acharya, Kamath, Sun, Zhang '18]
- "Local" model (e.g. randomized response) [Duchi, Jordan, Wainwright '13, '18, Sheffet '18, Acharya, Cannone, Freitag, Tyagi '18]
  - Subclass of algorithms where individual data points are randomized

Two hypothesis testing problems

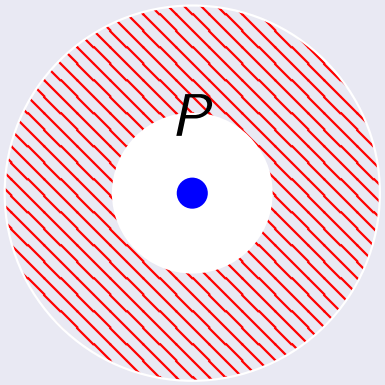
Simple hypothesis testing

$\{P\}$ vs. $\{Q\}$



Identity testing in high dimensions

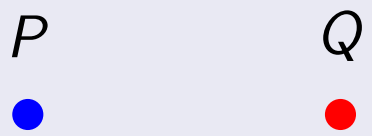
$\{\mathcal{U}_d\}$ vs. $\{\text{product dist } Q \mid TV(\mathcal{U}_d, Q) \geq \alpha\}$



Two hypothesis testing problems

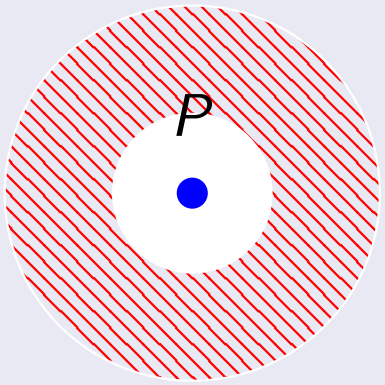
Simple hypothesis testing

$\{P\}$ vs. $\{Q\}$



Identity testing in high dimensions

$\{\mathcal{U}_d\}$ vs. $\{\text{product dist } Q \mid TV(\mathcal{U}_d, Q) \geq \alpha\}$

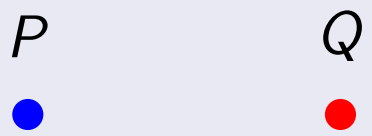


- Foundational problems, well understood in the non-private literature.

Two hypothesis testing problems

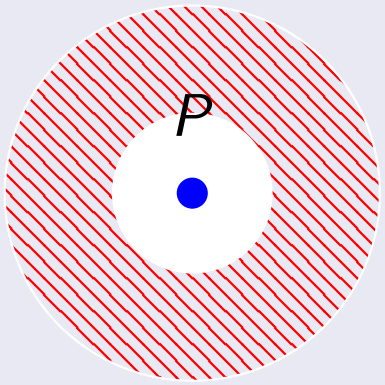
Simple hypothesis testing

$\{P\}$ vs. $\{Q\}$



Identity testing in high dimensions

$\{\mathcal{U}_d\}$ vs. $\{\text{product dist } Q \mid TV(\mathcal{U}_d, Q) \geq \alpha\}$

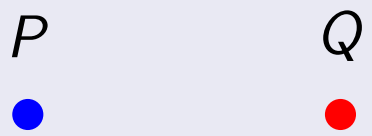


- Foundational problems, well understood in the non-private literature.
- Challenging to solve privately.

Two hypothesis testing problems

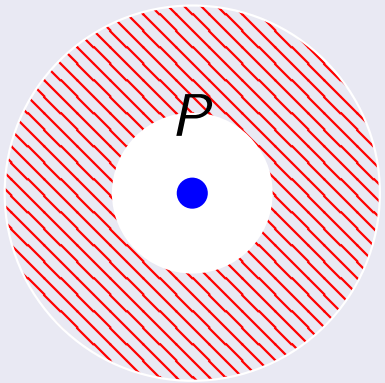
Simple hypothesis testing

$\{P\}$ vs. $\{Q\}$



Identity testing in high dimensions

$\{\mathcal{U}_d\}$ vs. $\{\text{product dist } Q \mid TV(\mathcal{U}_d, Q) \geq \alpha\}$



- Foundational problems, well understood in the non-private literature.
- Challenging to solve privately.

Goals:

- Design private algorithms that adapt to the specific instances.
- Understand dimensionality in private testing

Differential Privacy and Global Sensitivity

$$\text{Global sensitivity} = GS_f = \max_{X,X' \text{ neighbours}} |f(X) - f(X')|$$

$$f(X) + \text{Lap}\left(\frac{GS_f}{\epsilon}\right) \text{ is } \epsilon\text{-DP.}$$

Robustness and reducing sensitivity

Reducing sensitivity is a useful tool for designing DP algorithms

Simple Hypothesis Testing

Joint with Clément Canonne, Gautam Kamath,
Jon Ullman and Adam Smith.

arXiv:1811.11148

Simple Hypothesis Testing

Let P and Q be two distributions on the same domain Ω. In a **simple hypothesis test**,

$$\mathcal{H}_0 = \{P\} \text{ and } \mathcal{H}_1 = \{Q\}.$$

$R = \text{Ber}(0.5)$ or
$R = \text{Ber}(0.49)$

Simple Hypothesis Test

Given $x_1, \ldots, x_n \sim R$, a simple hypothesis test determines with high probability whether $R = P$ or $R = Q$.

Our work: instance-specific sample complexity

First work to give an instance-specific characterisation of sample complexity in the central model.

[DJW13]: instance-specific characterisation for the same problem in the local model.

Contributions of this work

The Optimal Sample Complexity

We characterise the optimal sample complexity for ϵ-DP distinguishing between P and Q, for any distributions P and Q.

An Optimal Efficient* Test

Give a specific efficient* test that achieves this sample complexity.

Classical Solution

Test $$\mathrm{LLR}(X) = \begin{cases} P & \text{if } P^n(X) \geq Q^n(X) \\ Q & \text{if } P^n(X) < Q^n(X) \end{cases}$$

Classical Solution

Test

$$\mathrm{LLR}(X) = \begin{cases} P & \text{if } P^n(X) \geq Q^n(X) \\ Q & \text{if } P^n(X) < Q^n(X) \end{cases}$$

Optimality

[Neyman-Pearson (1933)]
Exactly optimal sample complexity.

Classical Solution

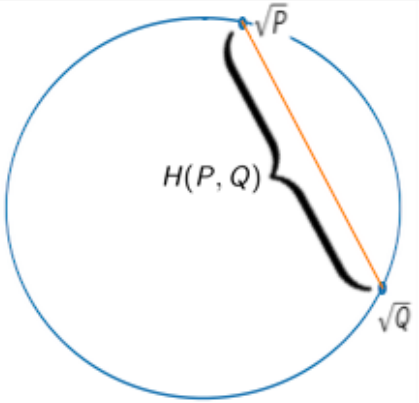
Test $$\mathrm{LLR}(X) = \begin{cases} P & \text{if } P^n(X) \geq Q^n(X) \\ Q & \text{if } P^n(X) < Q^n(X) \end{cases}$$

Optimality [Neyman-Pearson (1933)]
Exactly optimal sample complexity.

Sample Complexity $$SC^{P,Q} = \Theta\left(\frac{1}{H^2(P,Q)}\right)$$

where

$$H^2(P,Q) = \frac{1}{2}\int_\Omega (\sqrt{P(x)} - \sqrt{Q(x)})^2 dx$$

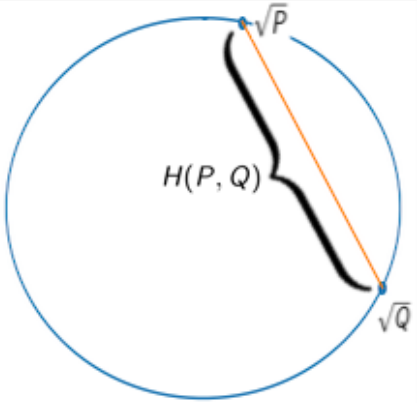


Classical Solution

Test Statistic	$\mathrm{LLR}(X) = \sum_{i=1}^{n} \log \frac{P(x_i)}{Q(x_i)}$
Optimality	[Neyman-Pearson (1933)] *Exactly* optimal sample complexity.
Sample Complexity	$SC^{P,Q} = \Theta\left(\frac{1}{H^2(P,Q)}\right)$

where

$$H^2(P,Q) = \frac{1}{2}\int_{\Omega}(\sqrt{P(x)} - \sqrt{Q(x)})^2 dx$$

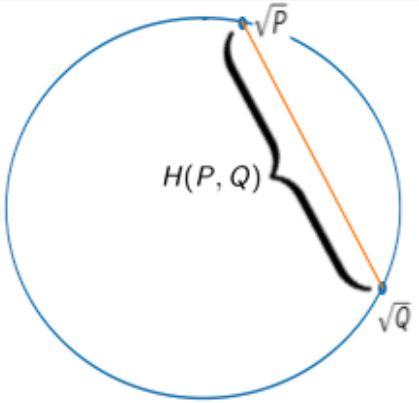


Classical Solution

Test Statistic	$\mathrm{LLR}(X) = \sum_{i=1}^{n} \log \frac{P(x_i)}{Q(x_i)}$
Test	$\mathrm{LLR}(X) = \begin{cases} P & \text{if } \mathrm{LLR}(X) \geq 0 \\ Q & \text{if } \mathrm{LLR}(X) < 0 \end{cases}$
Optimality	[Neyman-Pearson (1933)] *Exactly* optimal sample complexity.
Sample Complexity	$SC^{P,Q} = \Theta\left(\frac{1}{H^2(P,Q)}\right)$

where

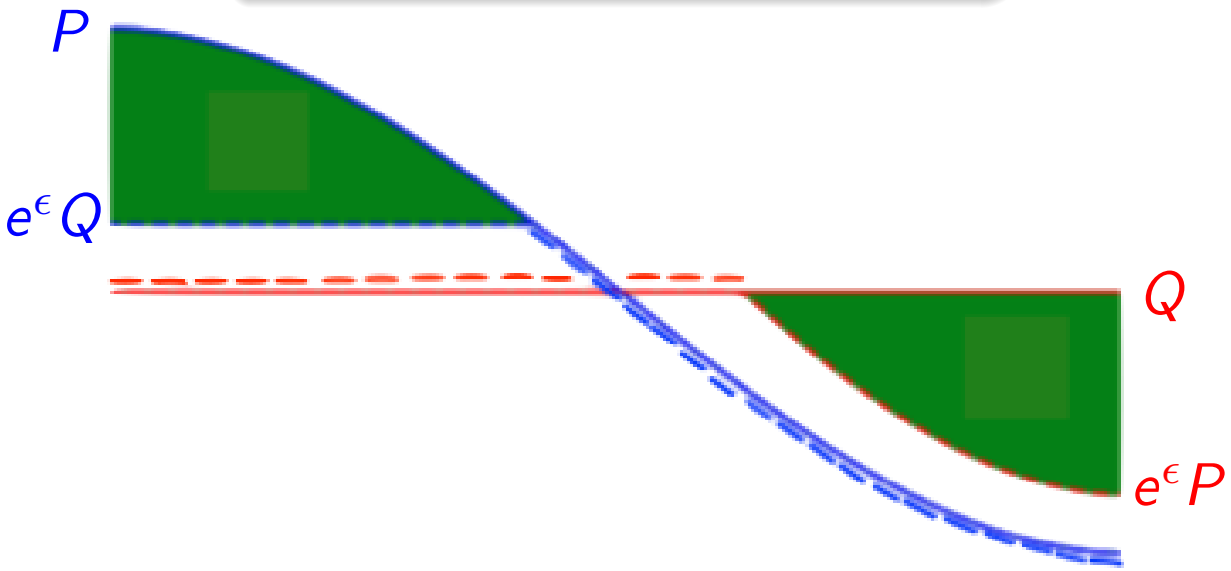
$$H^2(P,Q) = \frac{1}{2}\int_{\Omega} (\sqrt{P(x)} - \sqrt{Q(x)})^2 dx$$



Main Theorem

Test Statistic

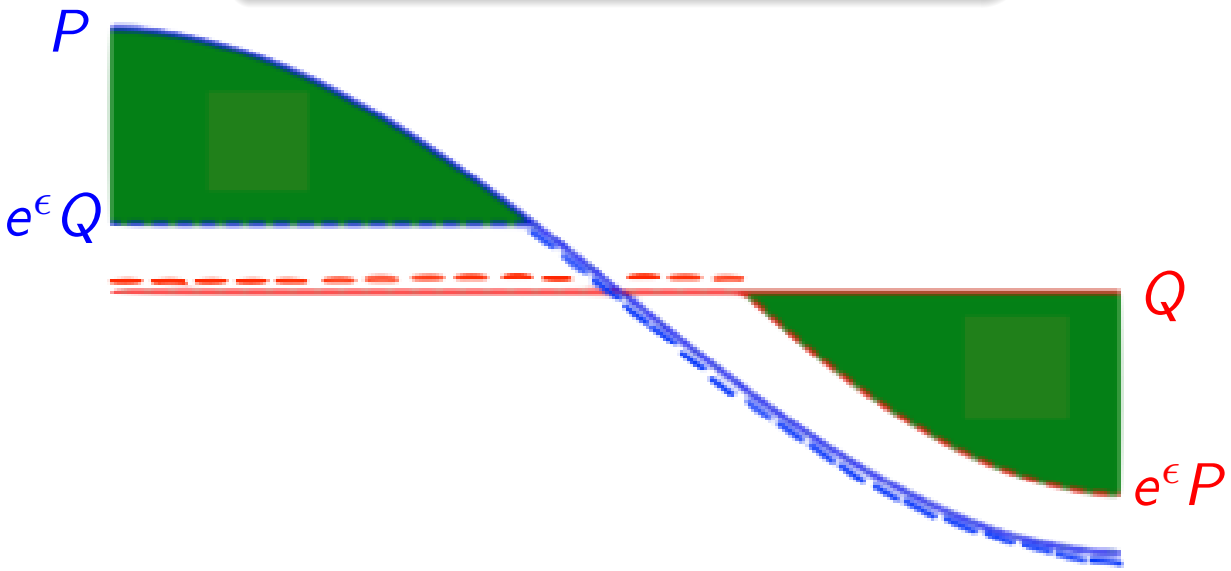
$$\mathrm{cLLR}(X) = \sum_{i=1}^{n} \left[\log \frac{P(x_i)}{Q(x_i)} \right]_{-\epsilon}^{\epsilon}$$



Main Theorem

Test Statistic

$$\mathrm{cLLR}(X) = \sum_{i=1}^{n} \left[\log \frac{P(x_i)}{Q(x_i)} \right]_{-\epsilon}^{\epsilon}$$



Noisy Clamped Log-likelihood Test

$$\mathrm{ncLLR}(X) = \begin{cases} P & \text{if } \mathrm{cLLR} + \mathrm{Lap}(2) \geq 0 \\ Q & \text{otherwise} \end{cases}$$

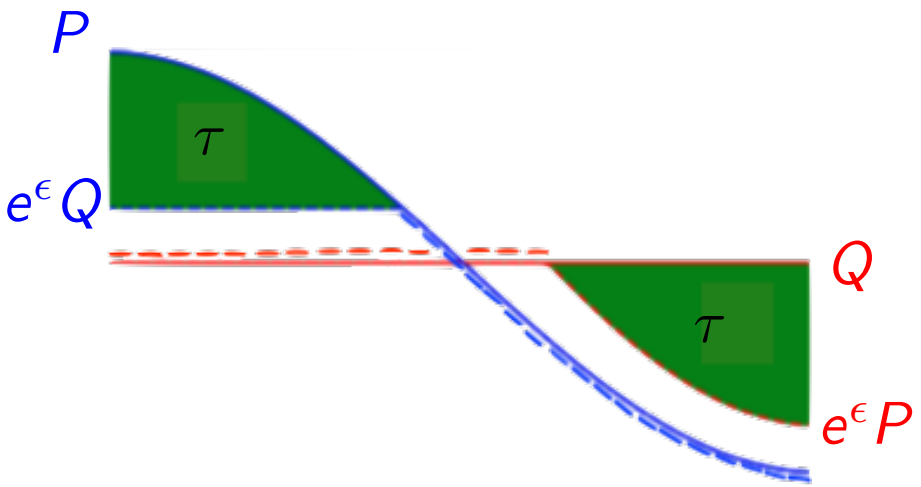
The Main Theorem: Optimal Private Sample Complexity

Theorem

The noisy clamped log-likelihood test, ncLLR, *has sample complexity*

$$SC_{\epsilon}^{P,Q} = \Theta\left(\min\left\{\frac{1}{\epsilon\tau}, \frac{1}{(1-\tau)H^2(P', Q')}\right\}\right),$$

which is minimal (up to constants) among ϵ-DP testing algorithms.



$$P' = \frac{1}{1-\tau}\min\{P, e^{\epsilon'}Q\} \quad \text{and} \quad Q' = \frac{1}{1-\tau}\min\{Q, e^{\epsilon}P\}$$

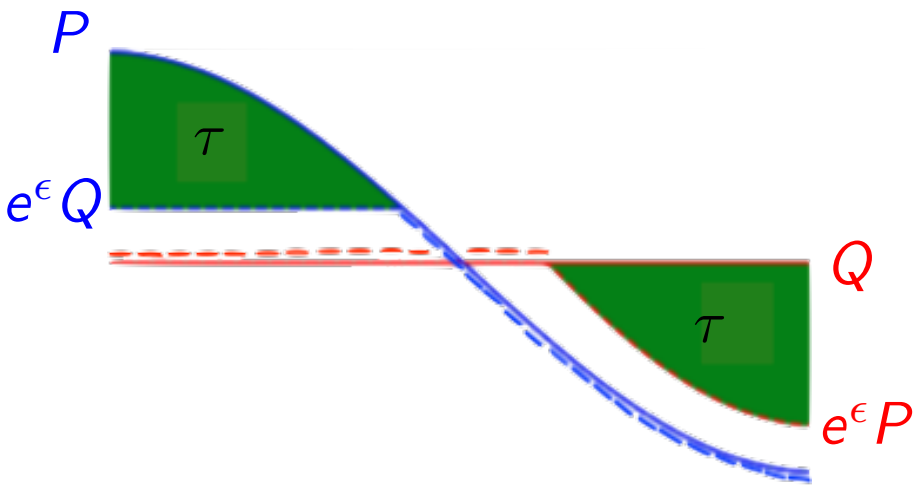
The Main Theorem: Optimal Private Sample Complexity

Theorem

The noisy clamped log-likelihood test, ncLLR, *has sample complexity*

$$SC_{\epsilon}^{P,Q} = \Theta\left(\min\left\{\frac{1}{\epsilon\tau}, \frac{1}{(1-\tau)H^2(P',Q')}\right\}\right),$$

which is minimal (up to constants) among ϵ-DP testing algorithms.



$$SC^{P,Q} = \Theta\left(\frac{1}{H^2(P,Q)}\right)$$

$$P' = \frac{1}{1-\tau}\min\{P, e^{\epsilon'}Q\} \quad \text{and} \quad Q' = \frac{1}{1-\tau}\min\{Q, e^{\epsilon}P\}$$

Private identity testing in high dimensions

Joint work with Clément Cannone, Gautam Kamath,
Jon Ullman and Lydia Zakynthinou

arXiv:1905.11947

Uniformity Testing in High Dimensions

$$\mathcal{U}_d = \text{uniform distribution on } \{-1, 1\}^d.$$

Uniformity Testing in High Dimensions

$$\mathcal{U}_d = \text{uniform distribution on } \{-1, 1\}^d.$$

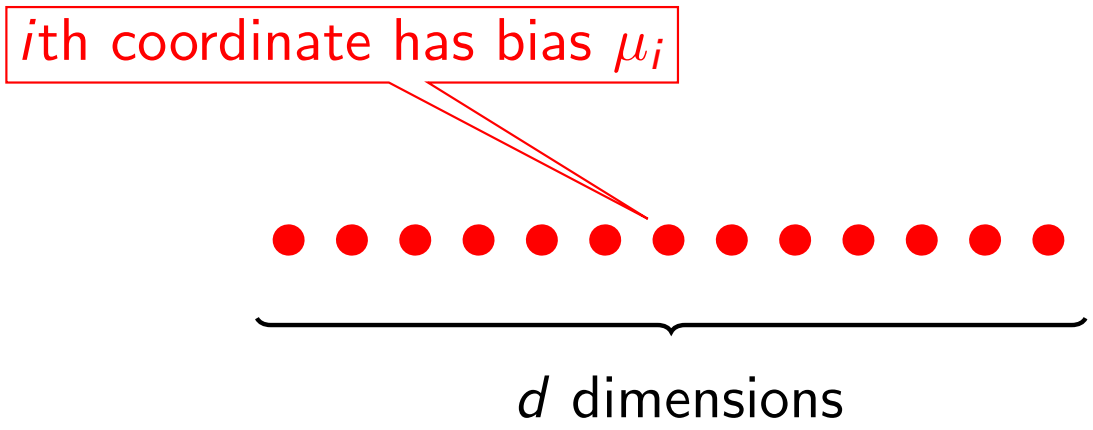
$$\mathcal{H}_0 = \{\mathcal{U}_d\} \text{ and } \mathcal{H}_1 = \{Q \mid TV(\mathcal{U}_d, Q) \geq \alpha \textit{ and } \}$$

Requires $2^{\Omega(d)}$ samples.

Uniformity Testing in High Dimensions

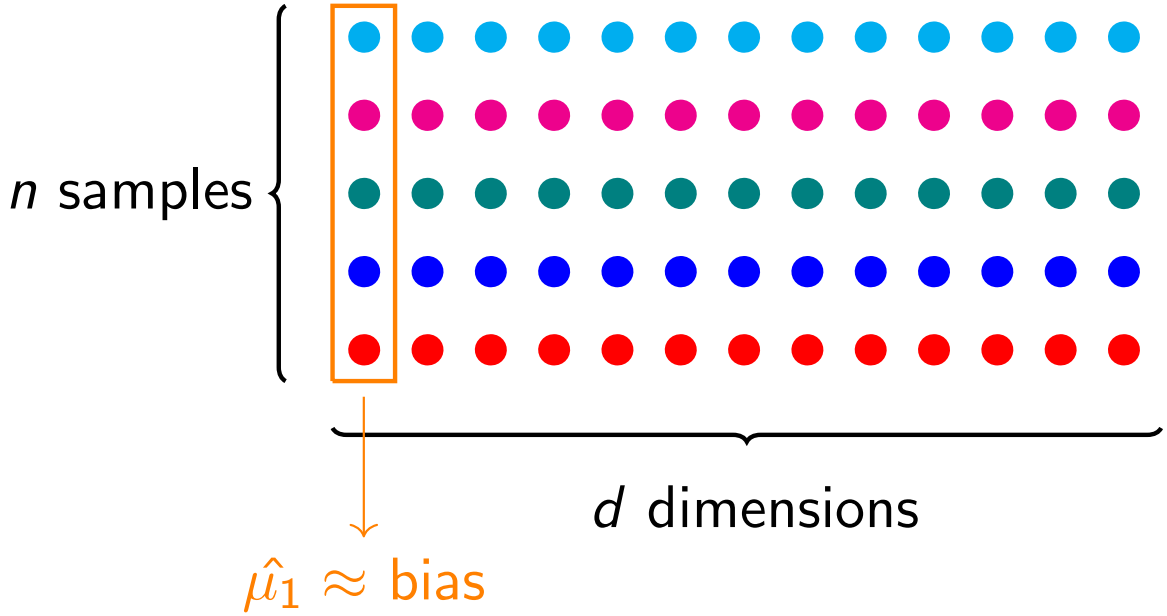
$$\mathcal{U}_d = \text{uniform distribution on } \{-1, 1\}^d.$$

$$\mathcal{H}_0 = \{\mathcal{U}_d\} \text{ and } \mathcal{H}_1 = \{Q \mid TV(\mathcal{U}_d, Q) \geq \alpha \textit{ and } Q \text{ is a product distribution}\}$$



Each coordinate is independent

Non-private solution for uniformity testing [Canonne, Diakonikolas, Kane, Stewart '17]

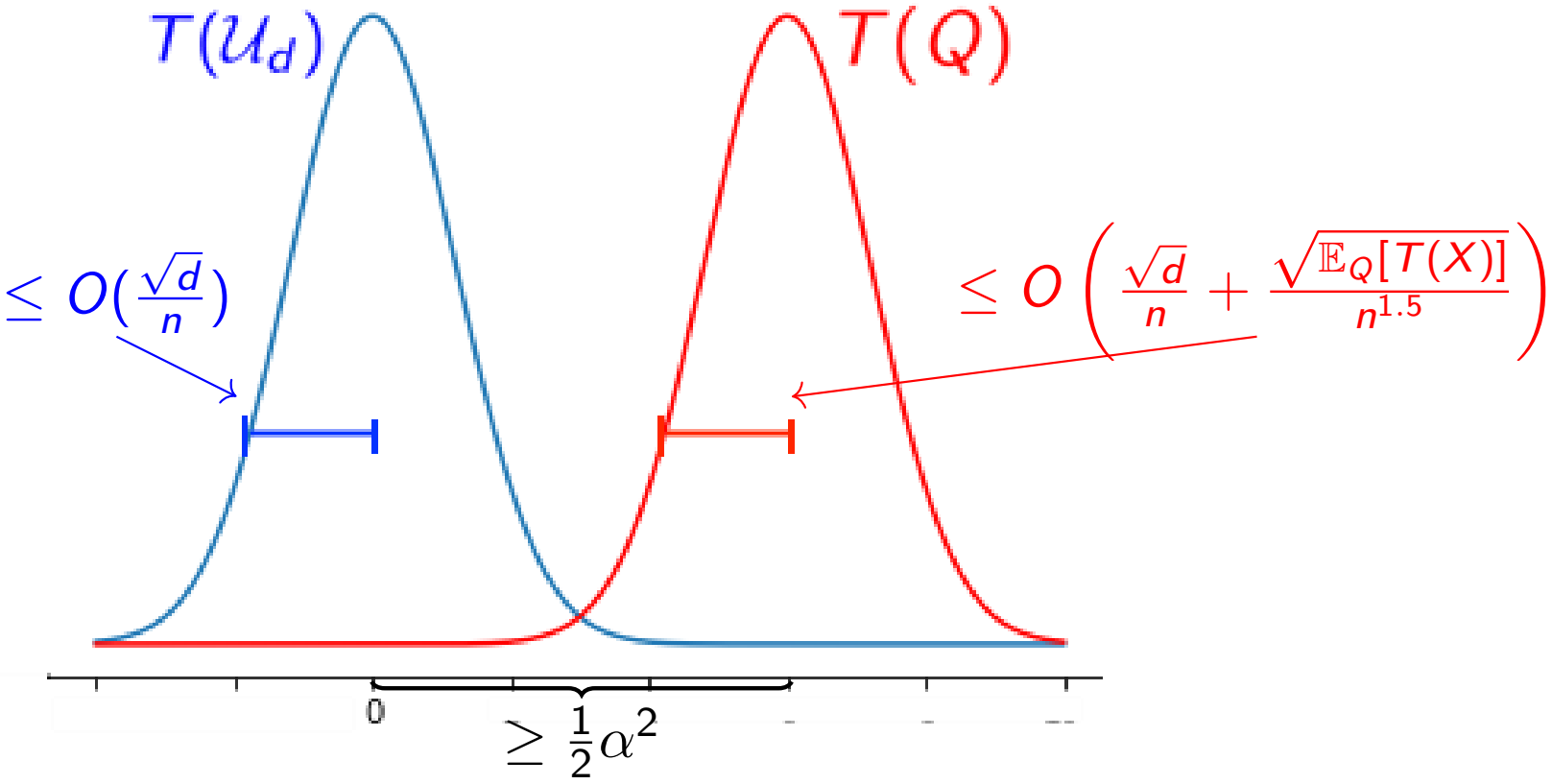


$\|\mu\|_2$ is a proxy for $\|\mathcal{U}_d - R\|_{\mathrm{TV}}$

Test Statistic: $T(X) = \|\hat{\mu}\|_2^2 - \frac{d}{n}$, where $\hat{\mu}_i = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$

Test:
$$\begin{cases} \mathcal{U}_d & \text{if} \quad T(X) \leq \frac{1}{4}\alpha^2 \\ TV(\mathcal{U}_d, Q) \geq \alpha & \text{if} \quad T(X) > \frac{1}{4}\alpha^2 \end{cases}$$

Non-private sample complexity [Canonne, Diakonikolas, Kane, Stewart '17]



Sample Complexity: $\Theta\left(\frac{\sqrt{d}}{\alpha^2}\right)$

Our contributions

ϵ-DP Uniformity tester

Give a test that distinguishes
$\mathcal{H}_0 = \{\mathcal{U}_d\}$ from $\mathcal{H}_1 = \{Q \text{ product distribution } \mid TV(\mathcal{U}_d, Q) \geq \alpha\}$.

Our contributions

ϵ-DP Uniformity tester

Give a test that distinguishes
$\mathcal{H}_0 = \{\mathcal{U}_d\}$ from $\mathcal{H}_1 = \{Q \text{ product distribution } \mid TV(\mathcal{U}_d, Q) \geq \alpha\}$.

ϵ-DP Gaussian mean tester with known covariance

Give a test that distinguishes
$\mathcal{H}_0 = \{\mathcal{N}(\mu, I_d)\}$ from $\mathcal{H}_1 = \{\eta \mid TV(\mathcal{N}(\mu, I_d), \mathcal{N}(\eta, I_d)) \geq \alpha\}$.

Our contributions

ϵ-DP Uniformity tester

Give a test that distinguishes
$\mathcal{H}_0 = \{\mathcal{U}_d\}$ from $\mathcal{H}_1 = \{Q \text{ product distribution } \mid TV(\mathcal{U}_d, Q) \geq \alpha\}$.

ϵ-DP Gaussian mean tester with known covariance

Give a test that distinguishes
$\mathcal{H}_0 = \{\mathcal{N}(\mu, I_d)\}$ from $\mathcal{H}_1 = \{\eta \mid TV(\mathcal{N}(\mu, I_d), \mathcal{N}(\eta, I_d)) \geq \alpha\}$.

	Non-private	**Inefficient ϵ-DP**	**Efficient ϵ-DP**
Testing	[Canonne, Diakonikolas, Kane, Stewart '17] $\Theta\left(\frac{\sqrt{d}}{\alpha^2}\right)$	[This work]	[This work]
Estimation	$\Theta\left(\frac{d}{\alpha^2}\right)$	[Kamath, Li, Singhal, Ullman '19] $\Theta\left(\frac{d}{\alpha^2} + \frac{d}{\alpha\epsilon}\right)$	[Kamath, Li, Singhal, Ullman '19] $\Theta\left(\frac{d}{\alpha^2} + \frac{d}{\alpha\epsilon}\right)$

Our contributions

ϵ-DP Uniformity tester

Give a test that distinguishes
$\mathcal{H}_0 = \{\mathcal{U}_d\}$ from $\mathcal{H}_1 = \{Q \text{ product distribution } \mid TV(\mathcal{U}_d, Q) \geq \alpha\}$.

ϵ-DP Gaussian mean tester with known covariance

Give a test that distinguishes
$\mathcal{H}_0 = \{\mathcal{N}(\mu, I_d)\}$ from $\mathcal{H}_1 = \{\eta \mid TV(\mathcal{N}(\mu, I_d), \mathcal{N}(\eta, I_d)) \geq \alpha\}$.

	Non-private	**Inefficient ϵ-DP**	**Efficient ϵ-DP**
Testing	[Canonne, Diakonikolas, Kane, Stewart '17] $\Theta\left(\frac{\sqrt{d}}{\alpha^2}\right)$	[This work]	[This work] $O\left(\frac{\sqrt{d}}{\alpha^2} + \frac{\sqrt{d}}{\alpha\epsilon}\right)$
Estimation	$\Theta\left(\frac{d}{\alpha^2}\right)$	[Kamath, Li, Singhal, Ullman '19] $\Theta\left(\frac{d}{\alpha^2} + \frac{d}{\alpha\epsilon}\right)$	[Kamath, Li, Singhal, Ullman '19] $\Theta\left(\frac{d}{\alpha^2} + \frac{d}{\alpha\epsilon}\right)$

Our contributions

ϵ-DP Uniformity tester

Give a test that distinguishes
$\mathcal{H}_0 = \{\mathcal{U}_d\}$ from $\mathcal{H}_1 = \{Q \text{ product distribution } \mid TV(\mathcal{U}_d, Q) \geq \alpha\}$.

ϵ-DP Gaussian mean tester with known covariance

Give a test that distinguishes
$\mathcal{H}_0 = \{\mathcal{N}(\mu, I_d)\}$ from $\mathcal{H}_1 = \{\eta \mid TV(\mathcal{N}(\mu, I_d), \mathcal{N}(\eta, I_d)) \geq \alpha\}$.

	Non-private	**Inefficient ϵ-DP**	**Efficient ϵ-DP**
Testing	[Canonne, Diakonikolas, Kane, Stewart '17] $\Theta\left(\frac{\sqrt{d}}{\alpha^2}\right)$	[This work] $O\left(\frac{\sqrt{d}}{\alpha^2} + \frac{\sqrt{d}}{\alpha\sqrt{\epsilon}} + \frac{d^{1/3}}{\alpha^{4/3}\epsilon^{2/3}} + \frac{1}{\alpha\epsilon}\right)$	[This work] $O\left(\frac{\sqrt{d}}{\alpha^2} + \frac{\sqrt{d}}{\alpha\epsilon}\right)$
Estimation	$\Theta\left(\frac{d}{\alpha^2}\right)$	[Kamath, Li, Singhal, Ullman '19] $\Theta\left(\frac{d}{\alpha^2} + \frac{d}{\alpha\epsilon}\right)$	[Kamath, Li, Singhal, Ullman '19] $\Theta\left(\frac{d}{\alpha^2} + \frac{d}{\alpha\epsilon}\right)$

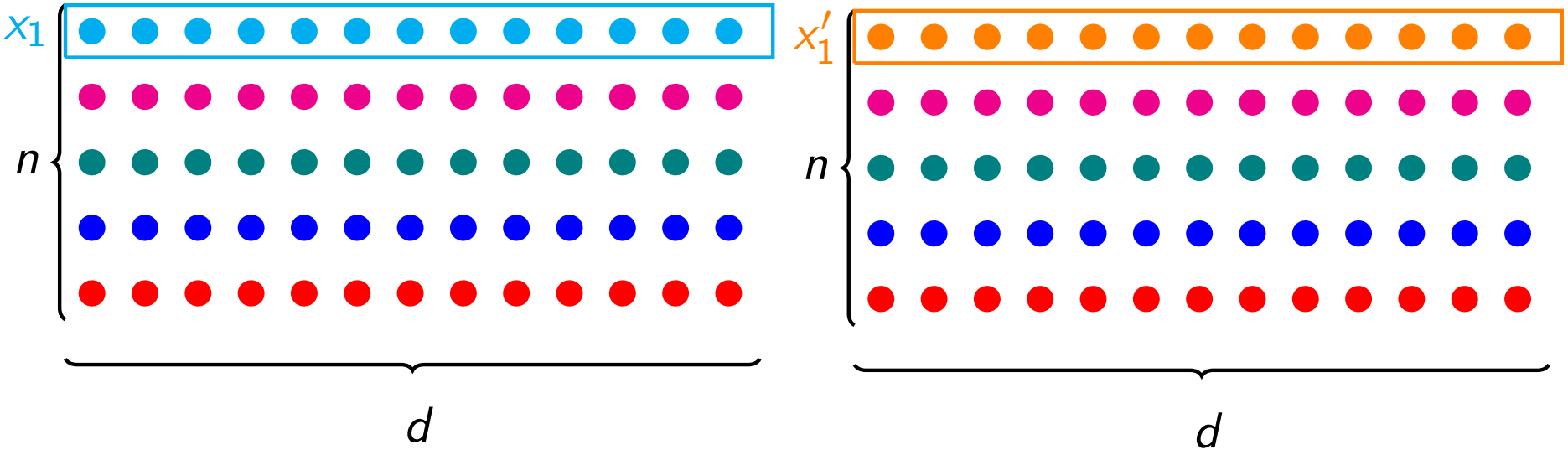
Our contributions

	Non-private	**Inefficient ϵ-DP**	**Efficient ϵ-DP**
Testing	[Canonne, Diakonikolas, Kane, Stewart '17] $\Theta\left(\frac{\sqrt{d}}{\alpha^2}\right)$	[This work] $O\left(\frac{\sqrt{d}}{\alpha^2} + \frac{\sqrt{d}}{\alpha\sqrt{\epsilon}} + \frac{d^{1/3}}{\alpha^{4/3}\epsilon^{2/3}} + \frac{1}{\alpha\epsilon}\right)$	[This work] $O\left(\frac{\sqrt{d}}{\alpha^2} + \frac{\sqrt{d}}{\alpha\epsilon}\right)$
Estimation	$\Theta\left(\frac{d}{\alpha^2}\right)$	[Kamath, Li, Singhal, Ullman '19] $\Theta\left(\frac{d}{\alpha^2} + \frac{d}{\alpha\epsilon}\right)$	[Kamath, Li, Singhal, Ullman '19] $\Theta\left(\frac{d}{\alpha^2} + \frac{d}{\alpha\epsilon}\right)$

Privacy for free?

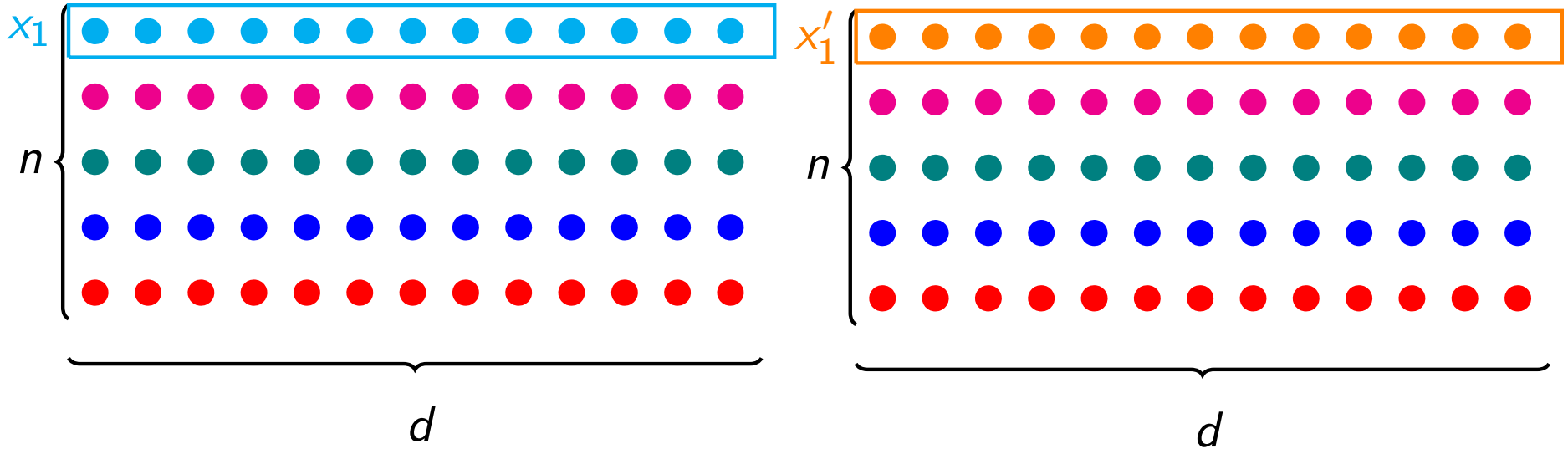
- For the efficient algorithm, when $\epsilon = \Omega(\alpha)$.
- For the inefficient algorithm, when $\epsilon = \Omega(\alpha^2 + \frac{\alpha}{d^{1/4}})$.

First Attempt: Global Sensitivity



$$T(X) = \|\hat{\mu}\|_2^2 - \frac{d}{n}, \qquad T(X) - T(X') = 2\left(\left\langle \frac{1}{n}x_1, \hat{\mu}\right\rangle - \left\langle \frac{1}{n}x_1', \hat{\mu}'\right\rangle\right)$$

First Attempt: Global Sensitivity



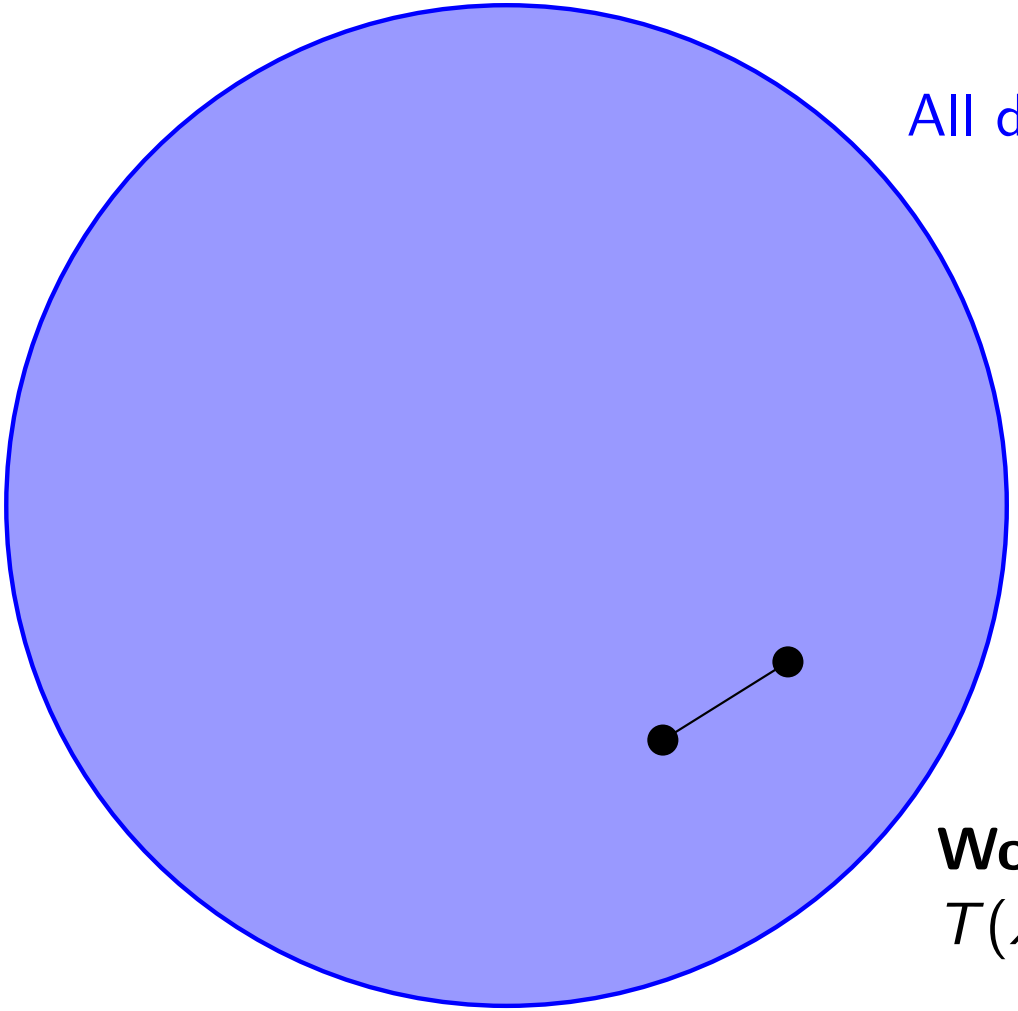
$$T(X) = \|\hat{\mu}\|_2^2 - \frac{d}{n}, \qquad T(X) - T(X') = 2\left(\left\langle \frac{1}{n}x_1, \hat{\mu}\right\rangle - \left\langle \frac{1}{n}x_1', \hat{\mu}'\right\rangle\right)$$

Worst case

The global sensitivity of $T = \Theta(\frac{d}{n})$, $\implies$ sample complexity $\frac{d}{\alpha^2\epsilon}$.

$$X_1 \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \qquad X_1 \begin{bmatrix} -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Typical Sensitivity

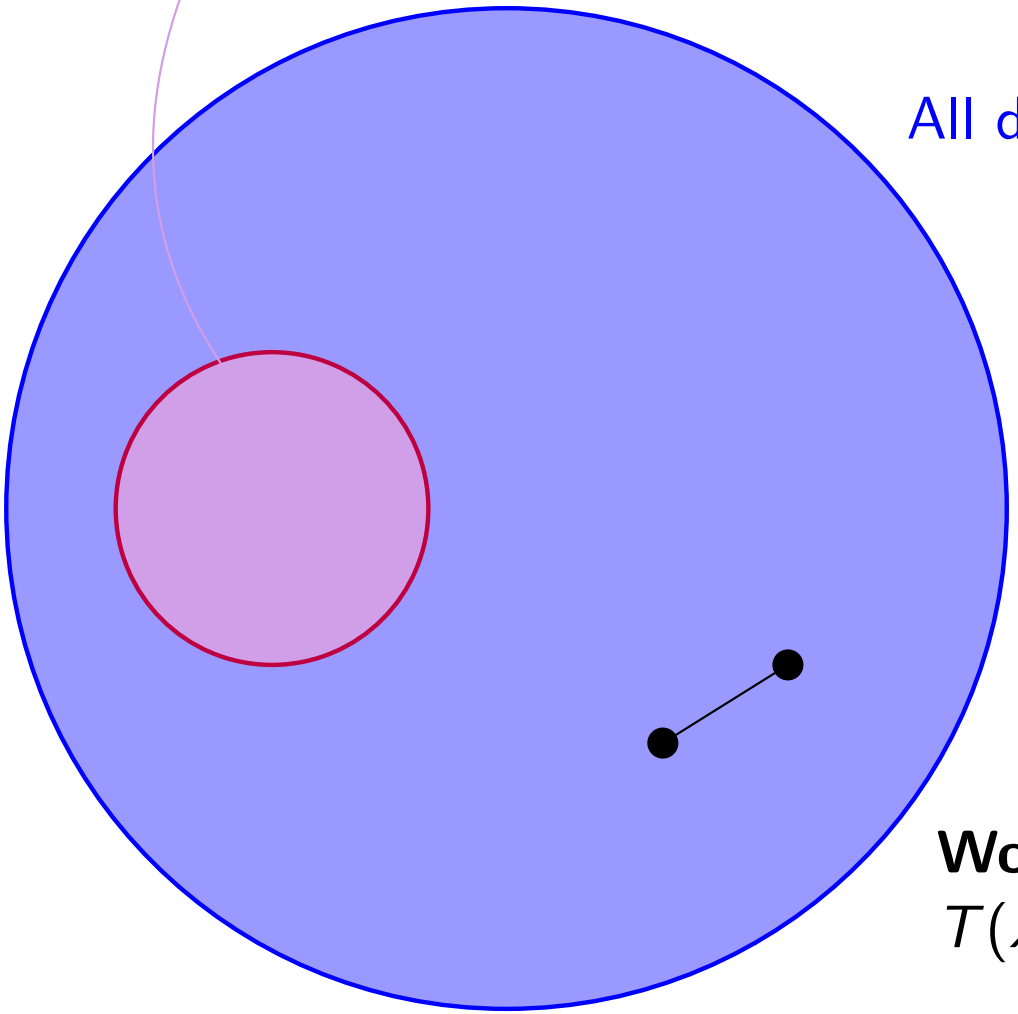


All datasets

Worst Case:
$T(X) - T(X') = \Theta\left(\frac{d}{n}\right)$

Typical Sensitivity

$X \sim \mathcal{U}_d$ lies in purple region w.h.p.



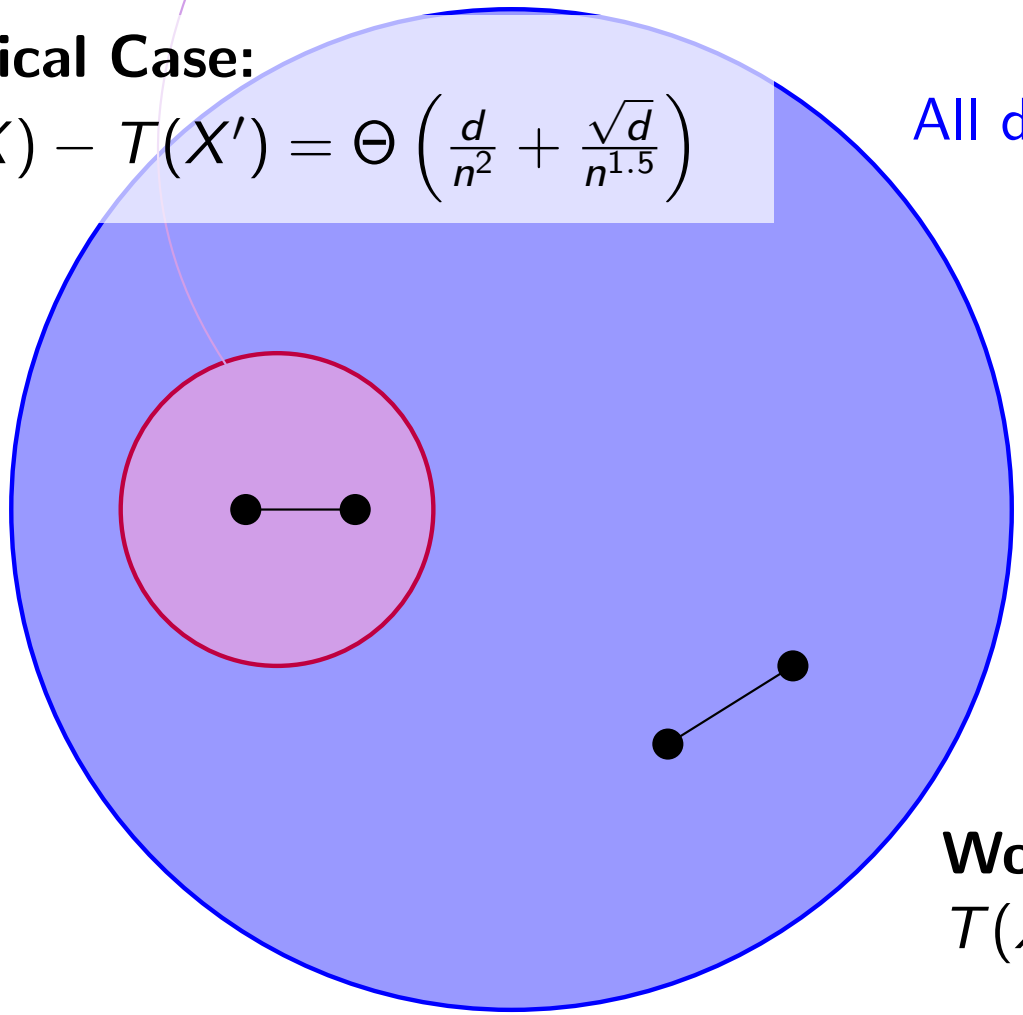
All datasets

Worst Case:
$T(X) - T(X') = \Theta\left(\frac{d}{n}\right)$

Typical Sensitivity

$X \sim \mathcal{U}_d$ lies in purple region w.h.p.

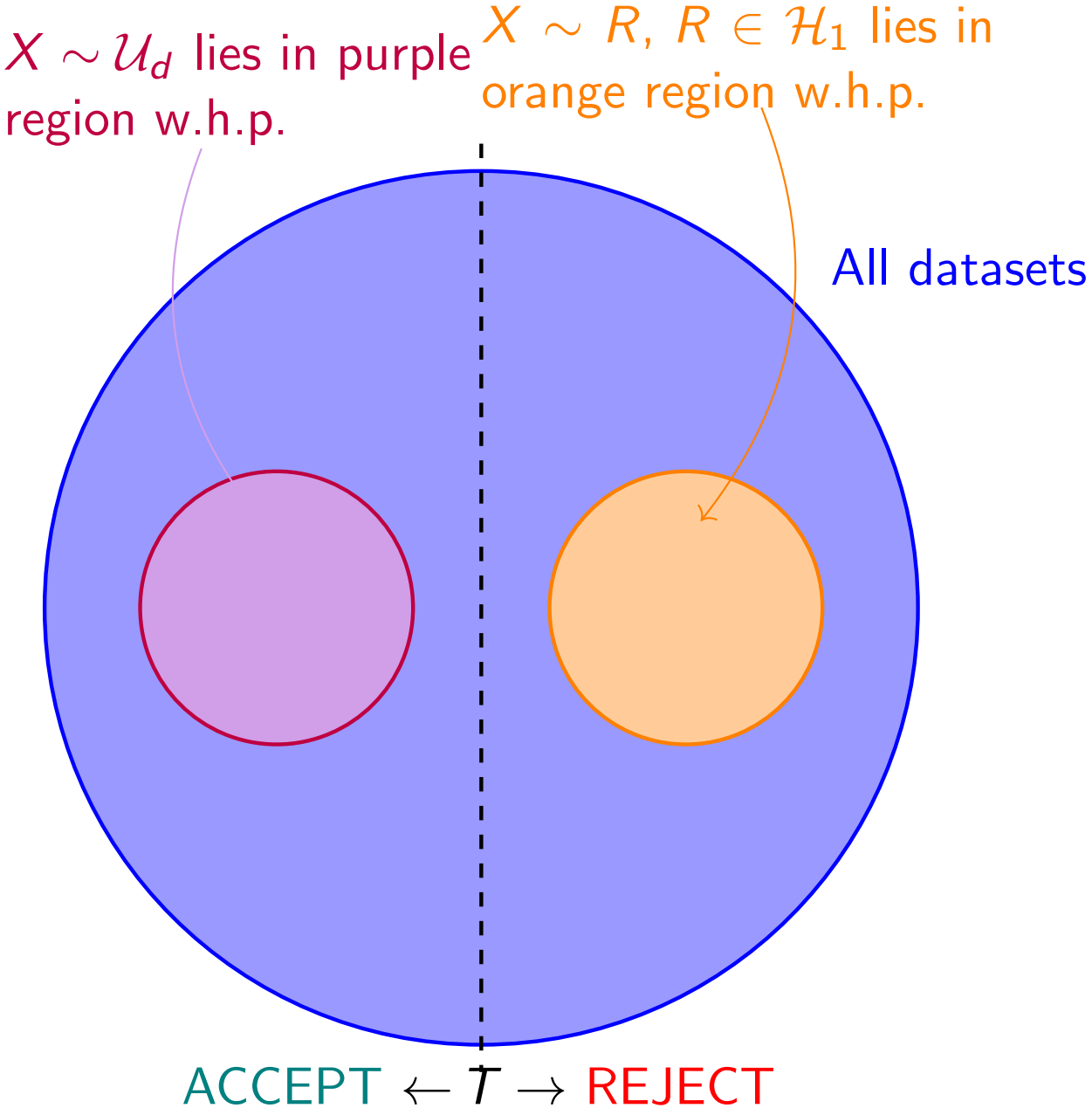
Typical Case:
$T(X) - T(X') = \Theta\left(\frac{d}{n^2} + \frac{\sqrt{d}}{n^{1.5}}\right)$



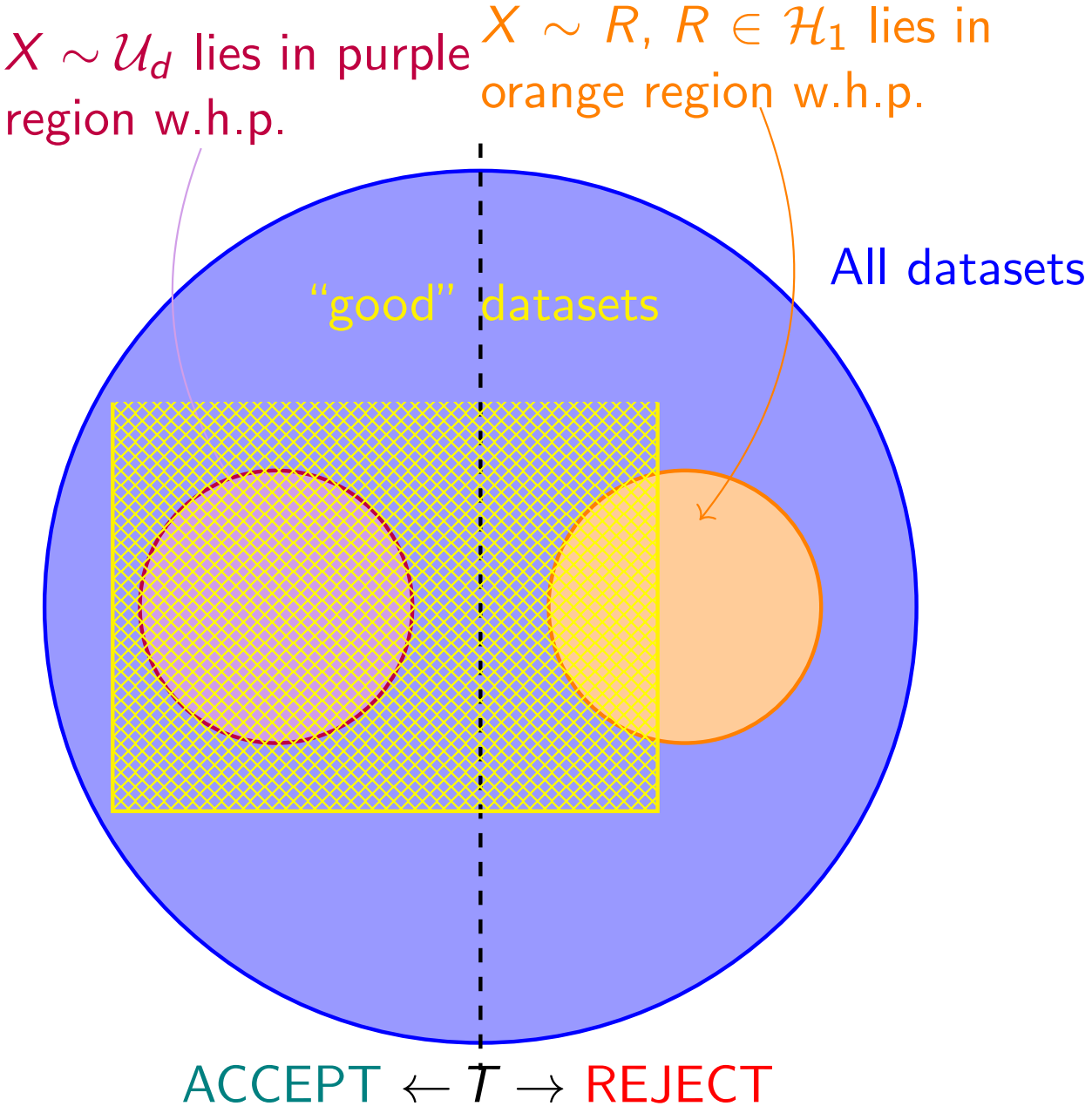
All datasets

Worst Case:
$T(X) - T(X') = \Theta\left(\frac{d}{n}\right)$

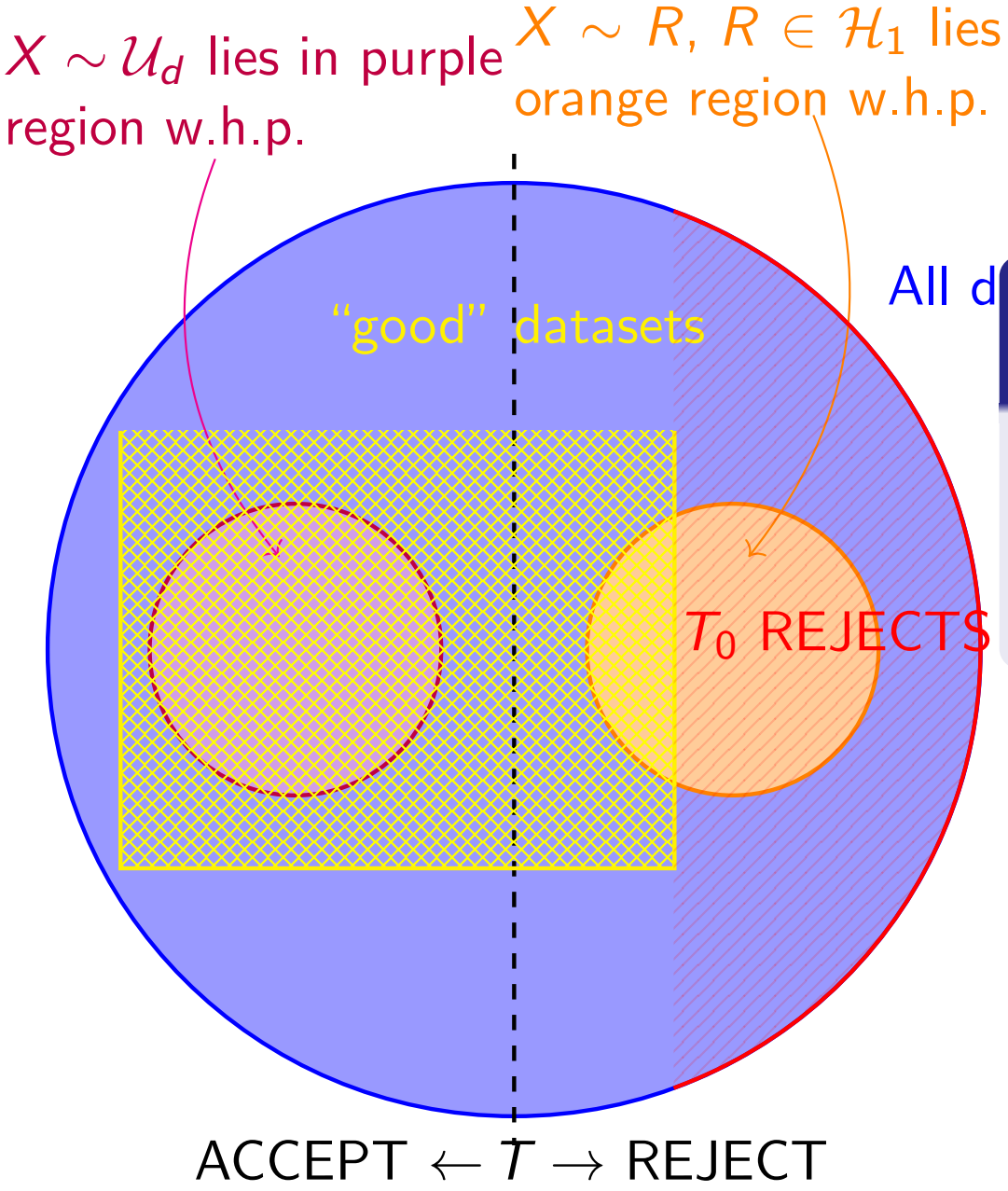
Key Idea: Filtering and Lipschitz Extensions



Key Idea: Filtering and Lipschitz Extensions



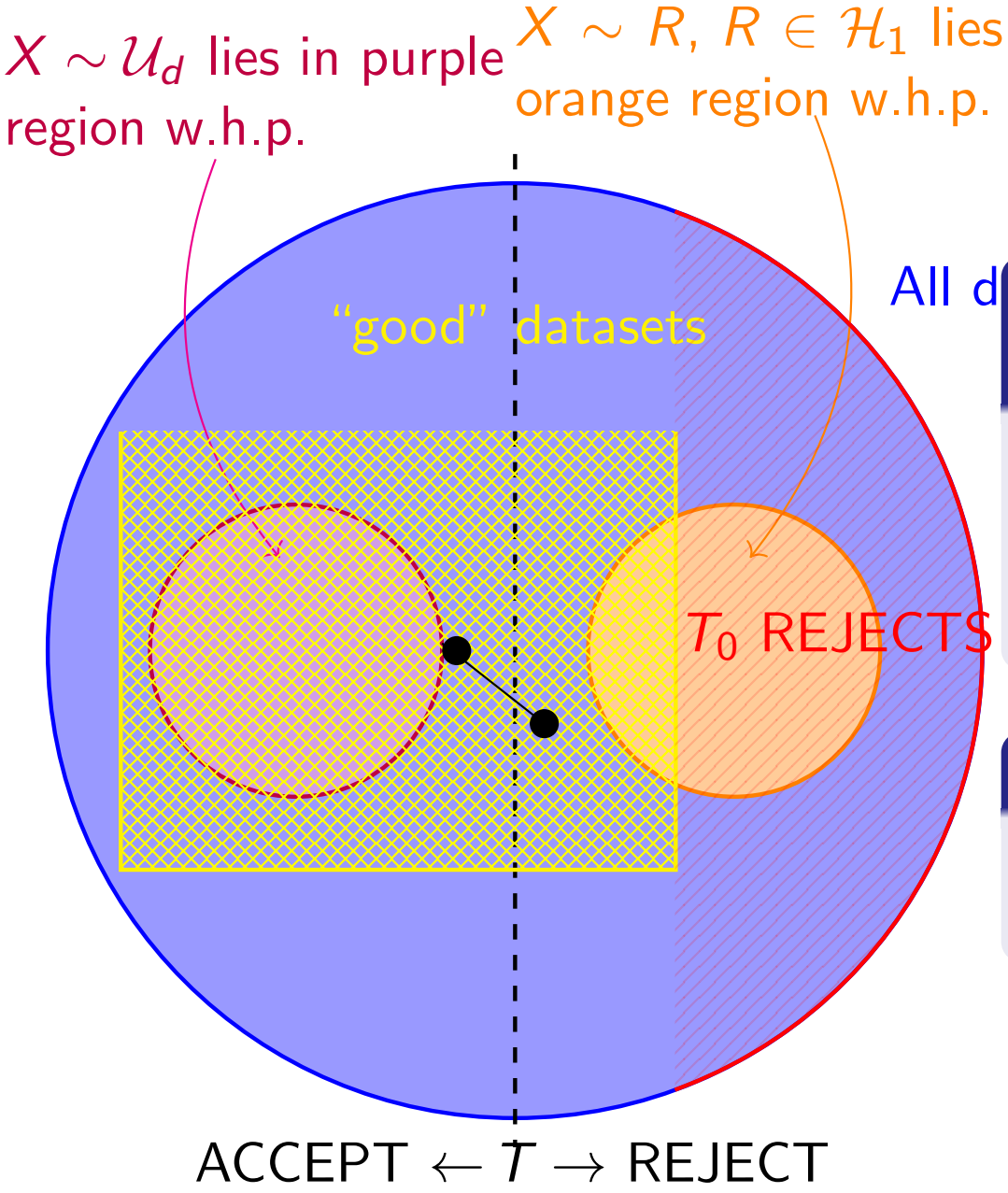
Key Idea: Filtering and Lipschitz Extensions



Conditions on yellow region

- $\exists$ insensitive test T_0.

Key Idea: Filtering and Lipschitz Extensions



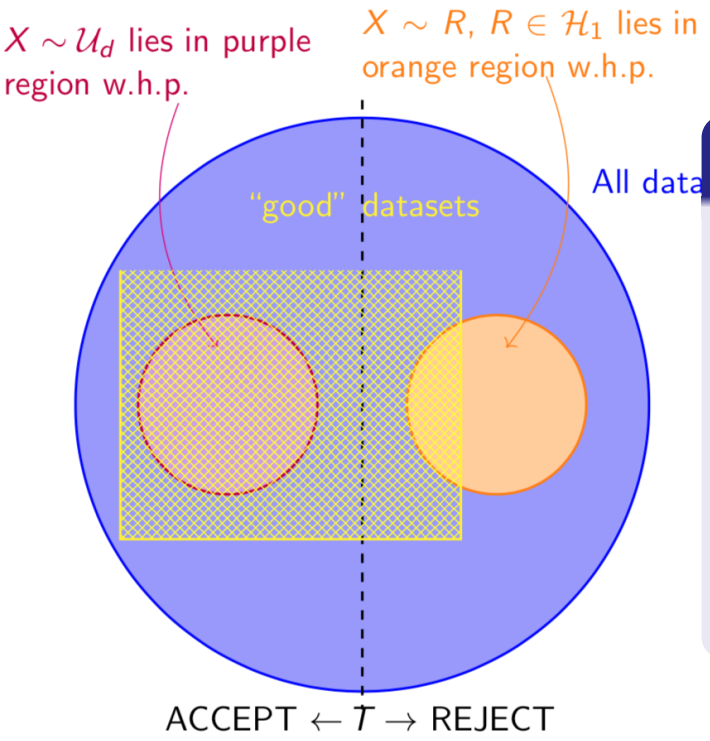
Conditions on yellow region

- $\exists$ insensitive test T_0.
- T = low sensitivity, λ, inside yellow region

Step 1: Filtering

Use T_0 to reject *obviously* non-uniform distributions.

Key Idea: Lipschitz Extension

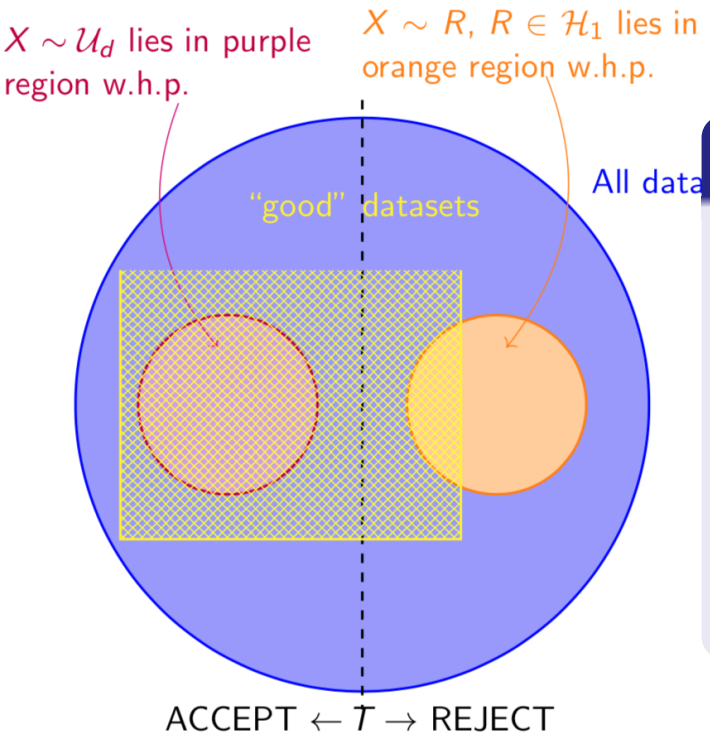


McShane–Whitney extension theorem

There exists a function $\hat{T}$ such that:

- $\hat{T}$ is defined on all datasets.
- $\mathrm{GS}(\hat{T})$ = sensitivity of $T|_{\text{yellow region}}$.
- $\hat{T}(X) = T(X)$ for $X \in$ yellow region.

Key Idea: Lipschitz Extension



McShane–Whitney extension theorem

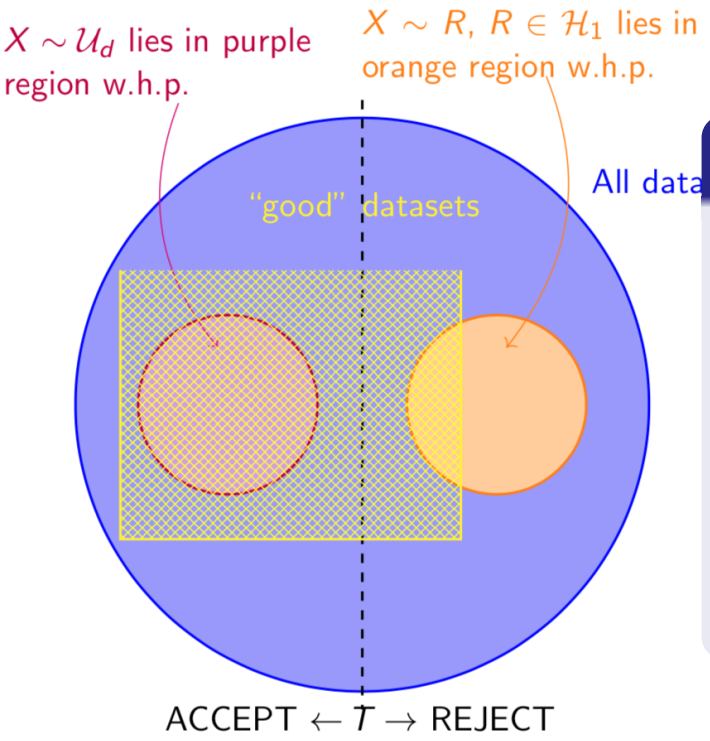
There exists a function $\hat{T}$ such that:

- $\hat{T}$ is defined on all datasets.
- $\mathrm{GS}(\hat{T})$ = sensitivity of $T|_{\text{yellow region}}$.
- $\hat{T}(X) = T(X)$ for $X \in$ yellow region.

Lipschitz extensions to reduce sensitivity in DP:

- introduced in [Blocki, Blum, Datta, Sheffet '13, Kasiviswanathan, Nissim, Raskhodnikova, Smith '13]
  - efficient extensions for graph statistics like edge density.

Key Idea: Lipschitz Extension



McShane–Whitney extension theorem

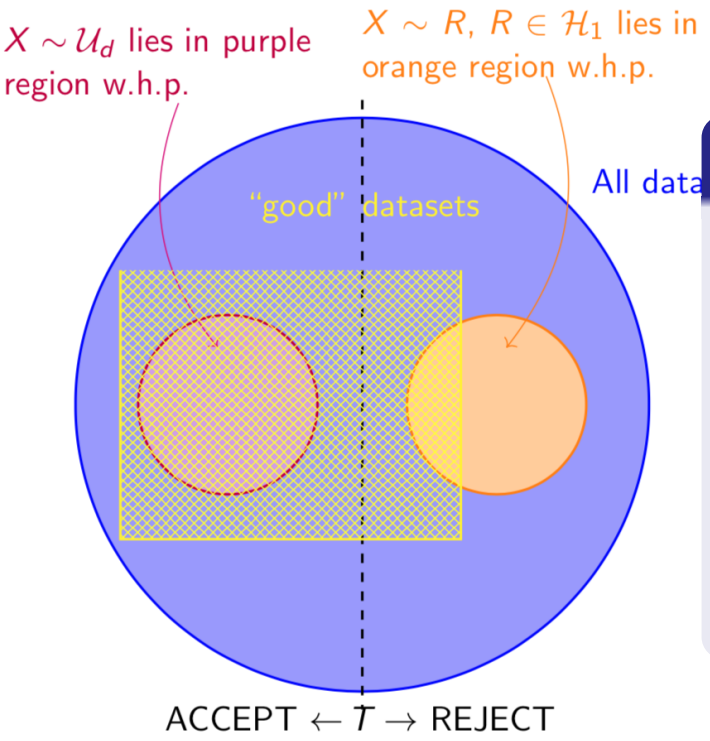
There exists a function $\hat{T}$ such that:

- $\hat{T}$ is defined on all datasets.
- $\mathrm{GS}(\hat{T})$ = sensitivity of $T|_{\text{yellow region}}$.
- $\hat{T}(X) = T(X)$ for $X \in$ yellow region.

Lipschitz extensions to reduce sensitivity in DP:

- introduced in [Blocki, Blum, Datta, Sheffet '13, Kasiviswanathan, Nissim, Raskhodnikova, Smith '13]
  - efficient extensions for graph statistics like edge density.
- for richer classes of graph statistics [Raskhodnikova, Smith '16, Borgs, Chayes, Smith, Zadik '18, Sealfon, Ullman '19]

Key Idea: Lipschitz Extension



McShane–Whitney extension theorem

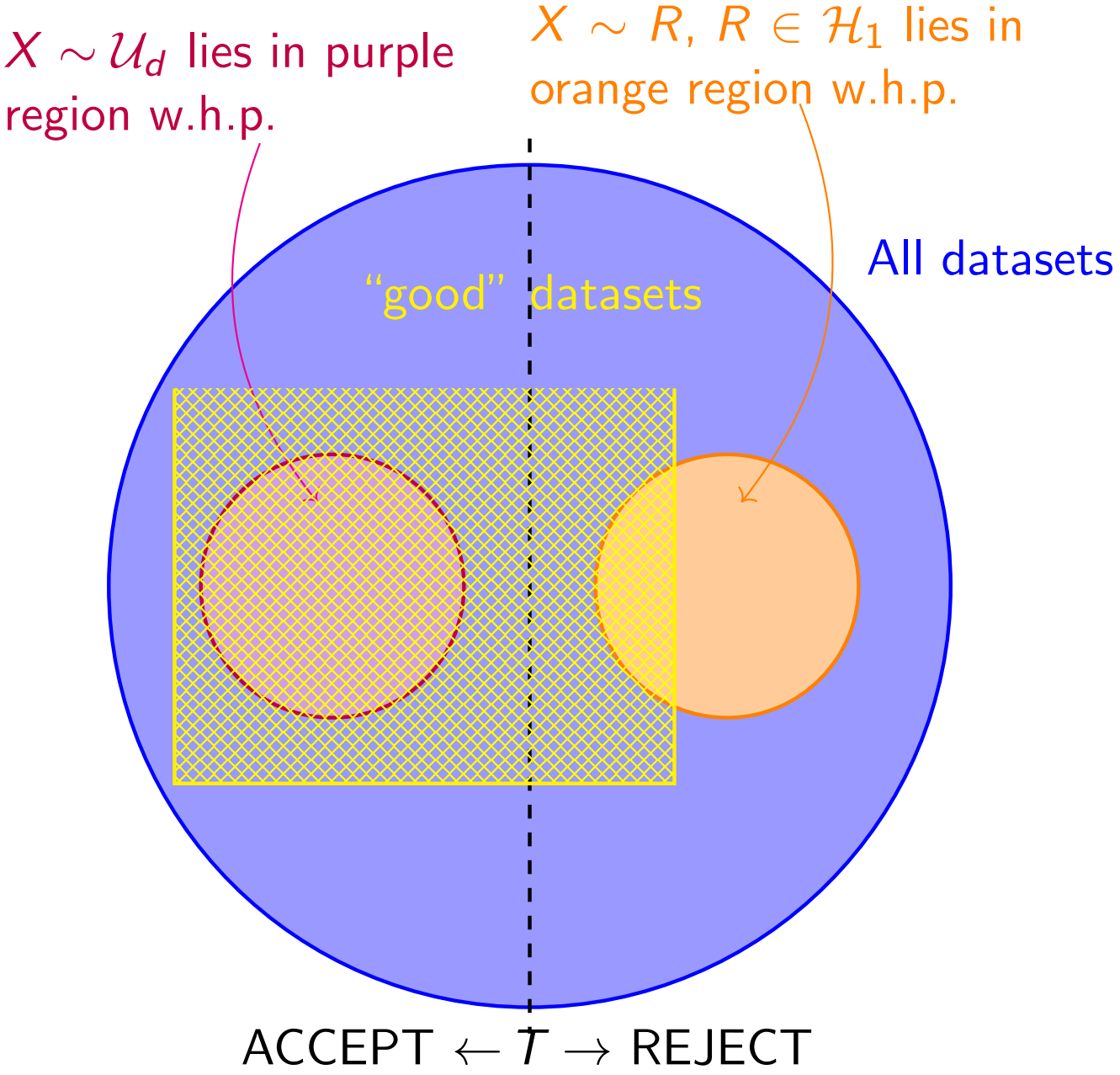
There exists a function $\hat{T}$ such that:

- $\hat{T}$ is defined on all datasets.
- $\mathrm{GS}(\hat{T})$ = sensitivity of $T|_{\text{yellow region}}$.
- $\hat{T}(X) = T(X)$ for $X \in$ yellow region.

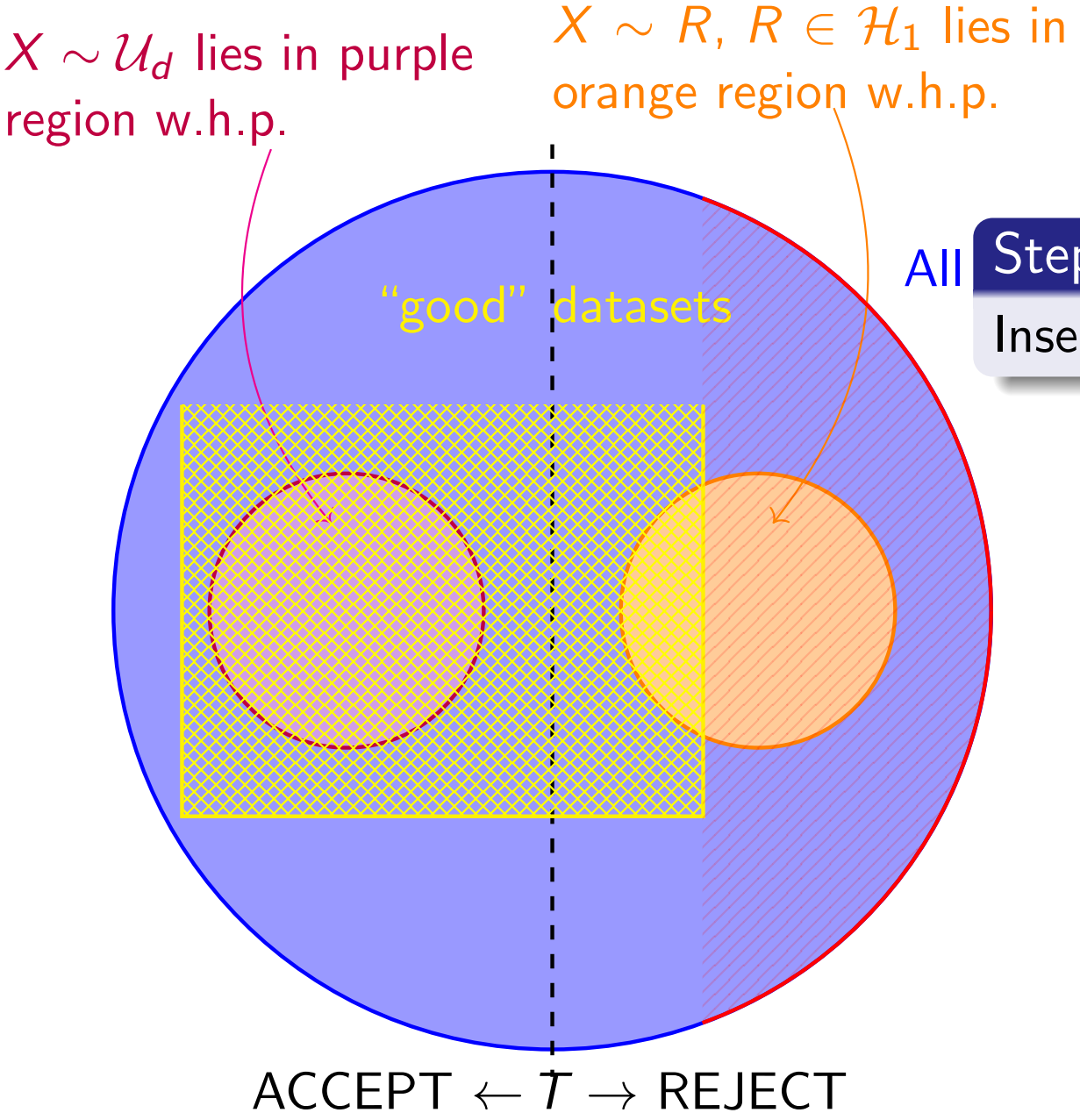
Lipschitz extensions to reduce sensitivity in DP:

- introduced in [Blocki, Blum, Datta, Sheffet '13, Kasiviswanathan, Nissim, Raskhodnikova, Smith '13]
  - efficient extensions for graph statistics like edge density.
- for richer classes of graph statistics [Raskhodnikova, Smith '16, Borgs, Chayes, Smith, Zadik '18, Sealfon, Ullman '19]
- efficient extensions for median and trimmed mean [Cummings, Durfee '18]

The algorithm



The algorithm



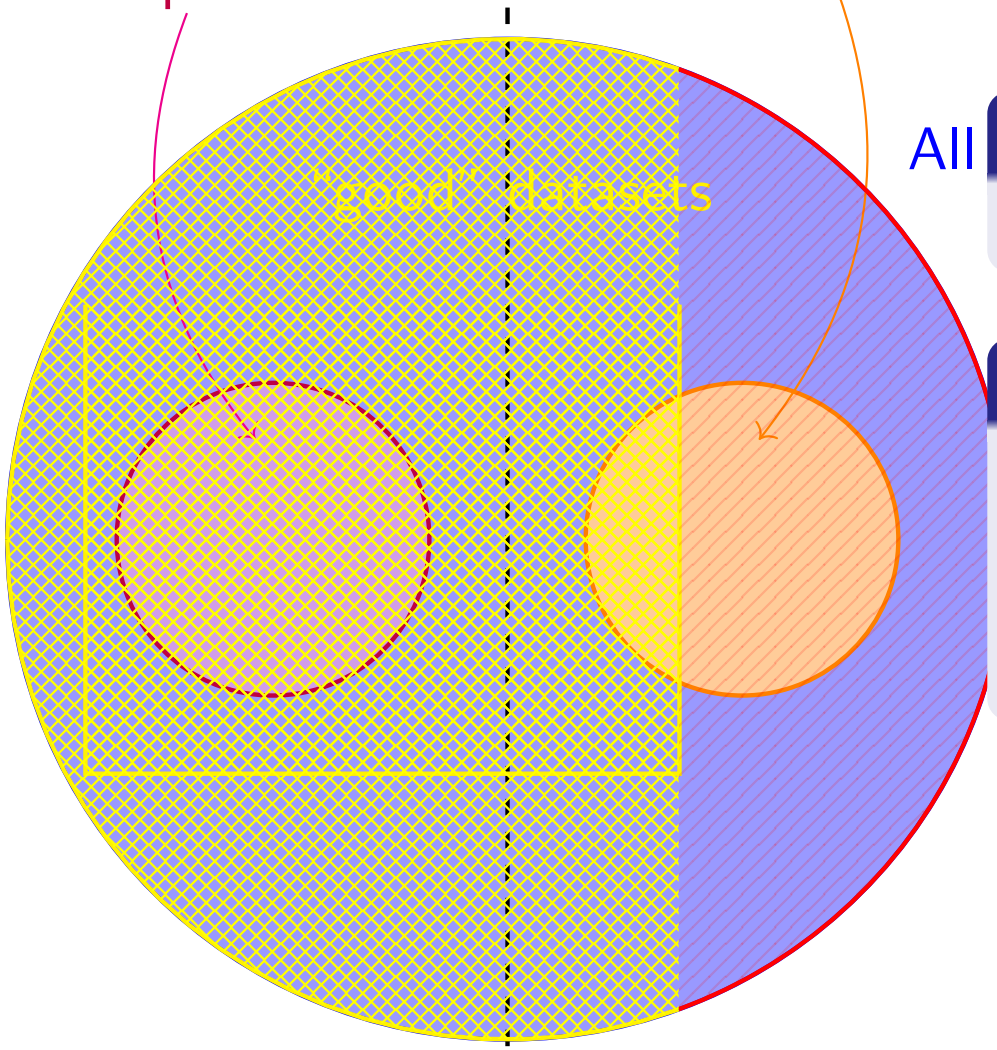
Step 1: Filtering

Insensitive test, T_0

The algorithm

$X \sim \mathcal{U}_d$ lies in purple region w.h.p.

$X \sim R$, $R \in \mathcal{H}_1$ lies in orange region w.h.p.



ACCEPT $\leftarrow T \rightarrow$ REJECT

Step 1: Filtering

Insensitive test, T_0

Step 2: Lipschitz Extension

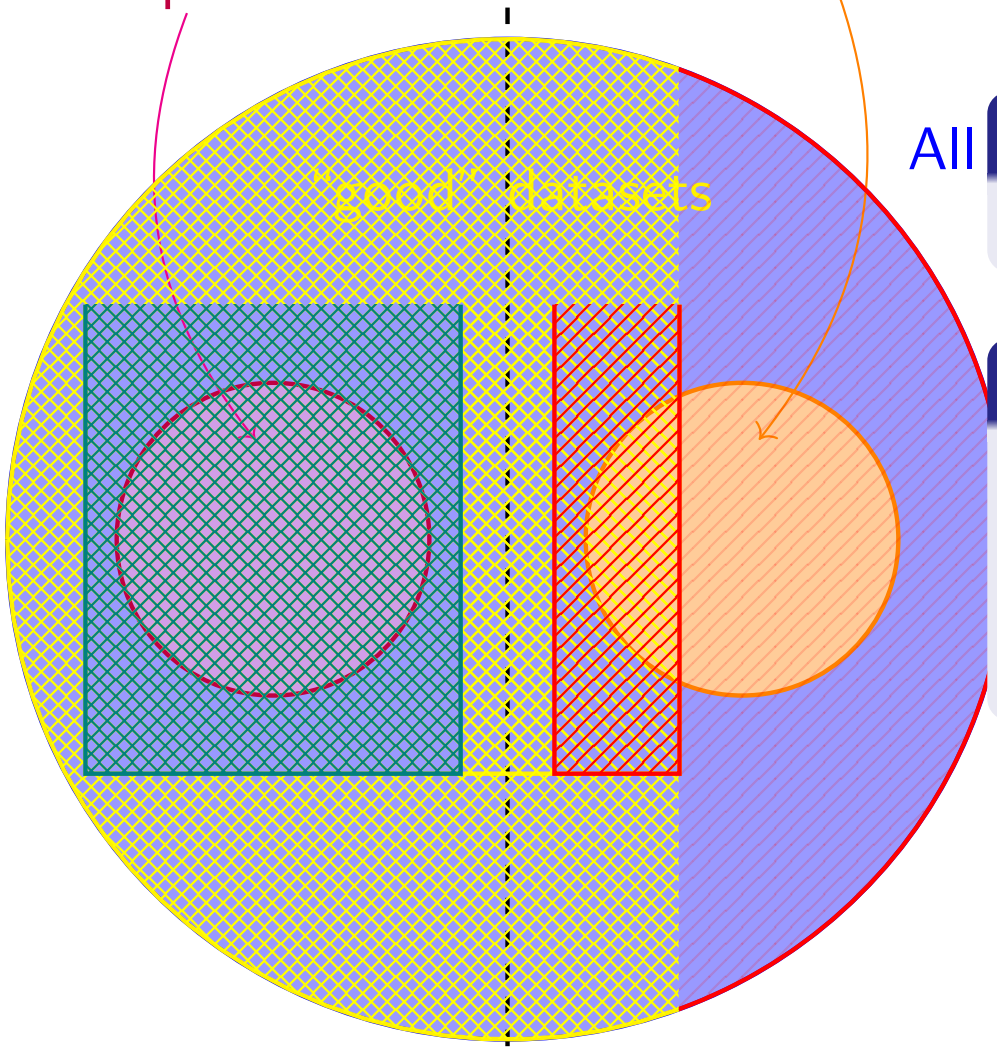
Use test statistic

$$\hat{T}(X) + \mathrm{Lap}\left(\frac{\lambda}{\epsilon}\right)$$

The algorithm

$X \sim \mathcal{U}_d$ lies in purple region w.h.p.

$X \sim R$, $R \in \mathcal{H}_1$ lies in orange region w.h.p.



ACCEPT $\leftarrow T \rightarrow$ REJECT

Step 1: Filtering

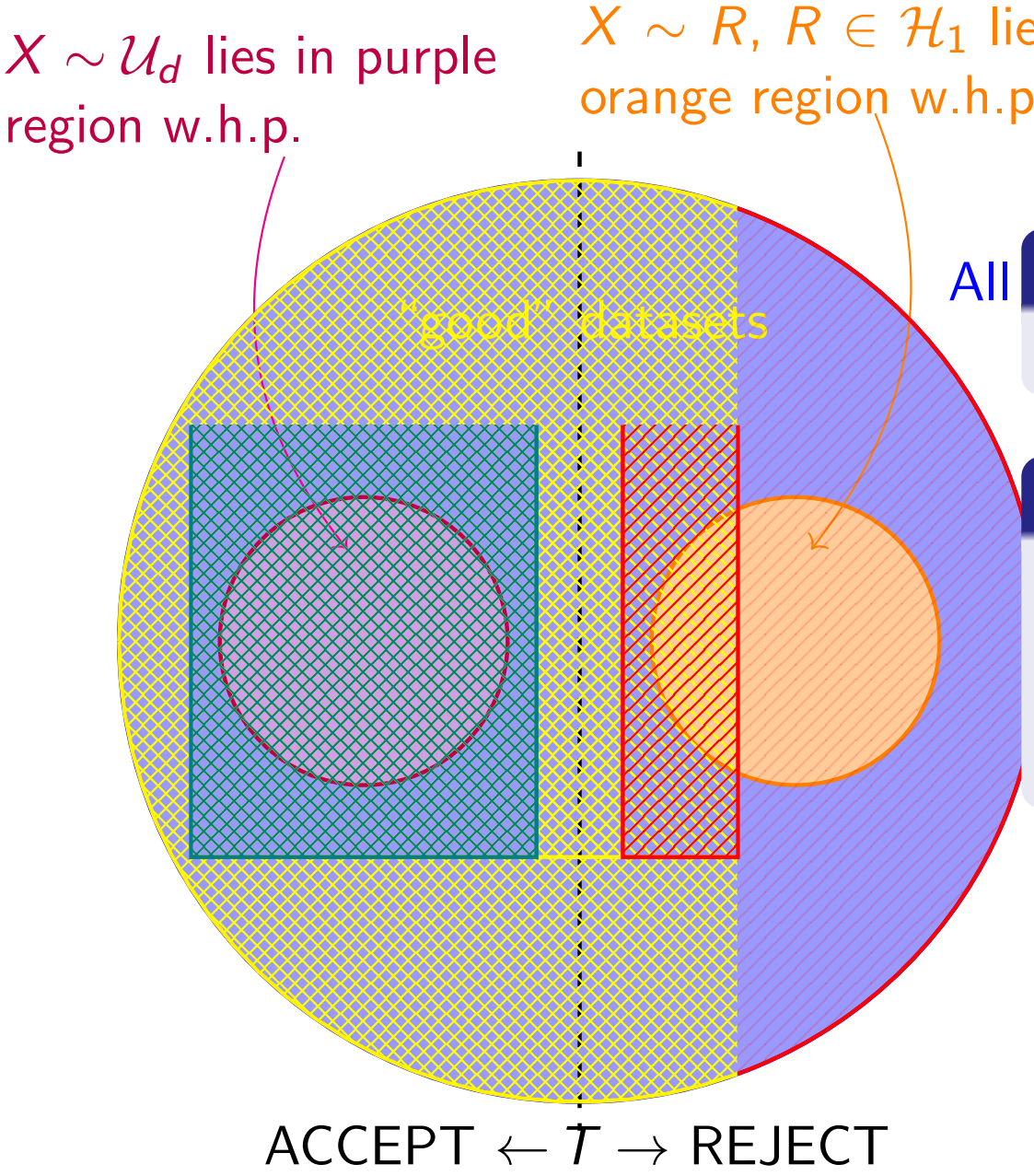
Insensitive test, T_0

Step 2: Lipschitz Extension

Use test statistic

$$\hat{T}(X) + \mathrm{Lap}\left(\frac{\lambda}{\epsilon}\right)$$

The algorithm



Step 1: Filtering

Insensitive test, T_0

Step 2: Lipschitz Extension

Use test statistic

$$\hat{T}(X) + \mathrm{Lap}\left(\frac{\lambda}{\epsilon}\right)$$

Sample complexity increases due to use of noisy statistic

Uniformity Testing

$$T(X) = \|\hat{\mu}\|_2^2 - \frac{d}{n}, \qquad T(X) - T(X') = 2\left(\left\langle \frac{1}{n} x_1, \hat{\mu} \right\rangle - \left\langle \frac{1}{n} x_1', \hat{\mu}' \right\rangle\right)$$

Worst case

The global sensitivity of $T = \Theta(\frac{d}{n})$, and sample complexity $\frac{d}{\alpha^2 \epsilon}$.

$$X_1 \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \qquad X_1 \begin{bmatrix} -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Uniformity Testing

$$T(X) = \|\hat{\mu}\|_2^2 - \frac{d}{n}, \qquad T(X) - T(X') = 2\left(\left\langle \frac{1}{n} x_1, \hat{\mu} \right\rangle - \left\langle \frac{1}{n} x_1', \hat{\mu}' \right\rangle\right)$$

Worst case

The global sensitivity of $T = \Theta(\frac{d}{n})$, and sample complexity $\frac{d}{\alpha^2 \epsilon}$.

$$\begin{array}{cccccccccccc} X_1 & \boxed{1} & \boxed{1} & \boxed{1} & \boxed{1} & \boxed{1} & \boxed{1} & \boxed{1} & \boxed{1} & \boxed{1} & \boxed{1} \\ & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{array} \qquad \begin{array}{cccccccccccc} X_1 & \boxed{-1} & \boxed{-1} & \boxed{-1} & \boxed{-1} & \boxed{-1} & \boxed{-1} & \boxed{-1} & \boxed{-1} & \boxed{-1} & \boxed{-1} \\ & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{array}$$

Problems:

- Too biased.

Uniformity Testing

$$T(X) = \|\hat{\mu}\|_2^2 - \frac{d}{n}, \qquad T(X) - T(X') = 2\left(\left\langle \frac{1}{n} x_1, \hat{\mu} \right\rangle - \left\langle \frac{1}{n} x_1', \hat{\mu}' \right\rangle\right)$$

Worst case

The global sensitivity of $T = \Theta(\frac{d}{n})$, and sample complexity $\frac{d}{\alpha^2 \epsilon}$.

$$\begin{array}{cccccccccccc} X_1 & \boxed{1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1} \\ & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{array} \qquad \begin{array}{cccccccccccc} X_1 & \boxed{-1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1} \\ & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{array}$$

Problems:

- Too biased.
- Coordinates are not independent.

Uniformity Testing

$$T(X) = \|\hat{\mu}\|_2^2 - \frac{d}{n}, \qquad T(X) - T(X') = 2\left(\left\langle \frac{1}{n} x_1, \hat{\mu} \right\rangle - \left\langle \frac{1}{n} x_1', \hat{\mu}' \right\rangle\right)$$

Worst case

The global sensitivity of $T = \Theta(\frac{d}{n})$, and sample complexity $\frac{d}{\alpha^2 \epsilon}$.

$$\begin{array}{c|cccccccccc}
X_1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1
\end{array}
\qquad
\begin{array}{c|cccccccccc}
X_1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\
 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1
\end{array}$$

Problems:

- Too biased.
- Coordinates are not independent.

FILTERING

Uniformity Testing

$$T(X) = \|\hat{\mu}\|_2^2 - \frac{d}{n}, \qquad T(X) - T(X') = 2\left(\left\langle \frac{1}{n} x_1, \hat{\mu} \right\rangle - \left\langle \frac{1}{n} x_1', \hat{\mu}' \right\rangle\right)$$

Worst case

The global sensitivity of $T = \Theta(\frac{d}{n})$, and sample complexity $\frac{d}{\alpha^2 \epsilon}$.

$$\begin{array}{c|cccccccccc}
X_1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1
\end{array}
\qquad
\begin{array}{c|cccccccccc}
X_1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\
 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1
\end{array}$$

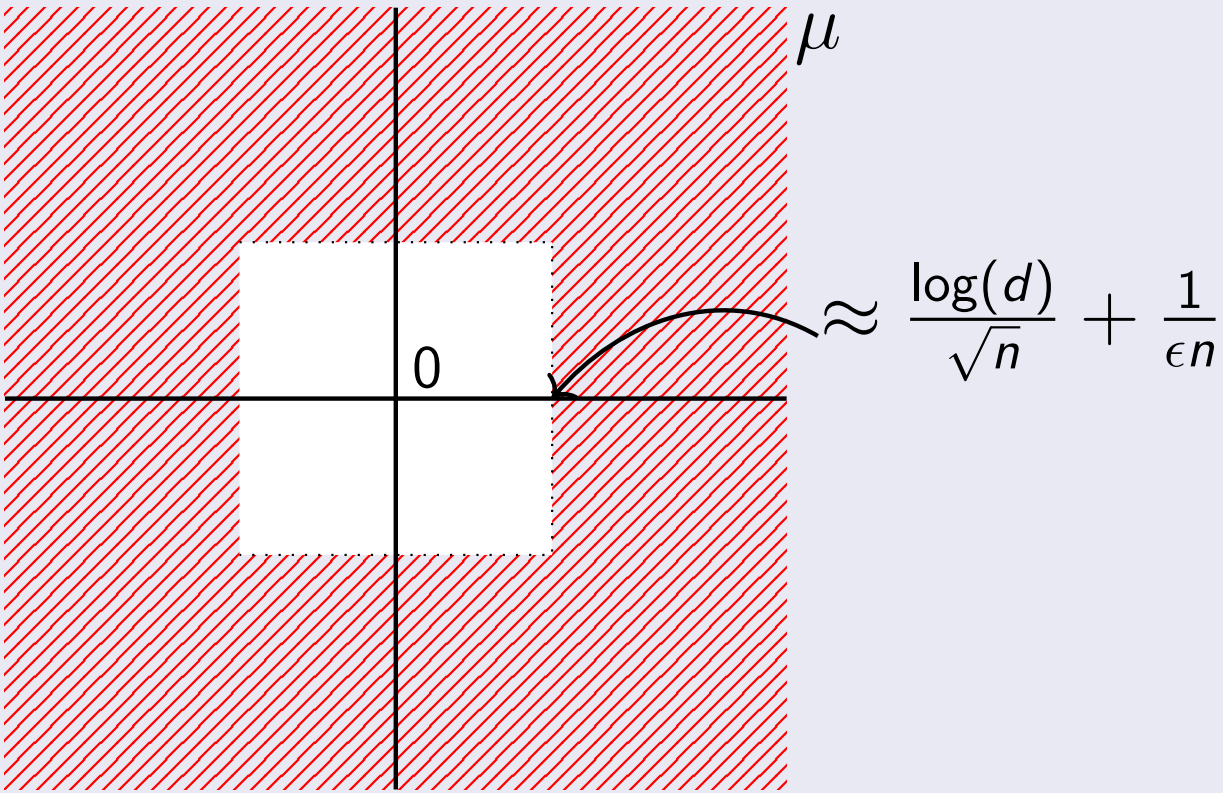
Problems:

- Too biased. FILTERING
- Coordinates are not independent. LIPSCHITZ EXTENSION

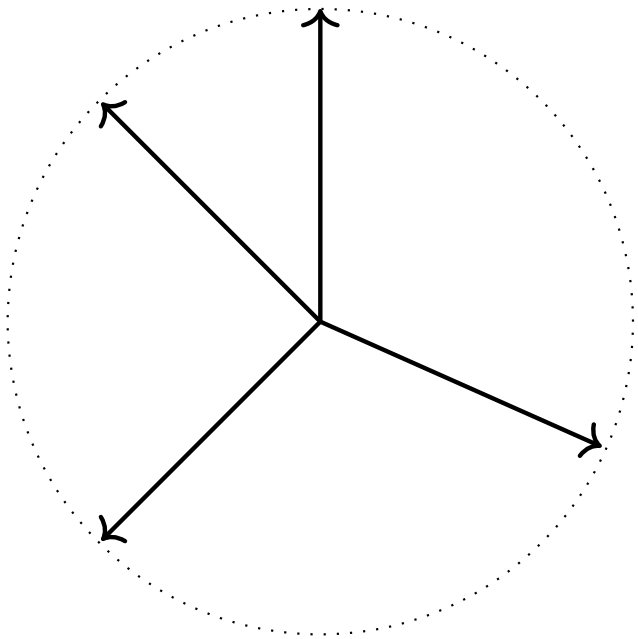
Uniformity Testing: Filtering

Step 1: Filtering

Reject if any coordinate has too high bias.



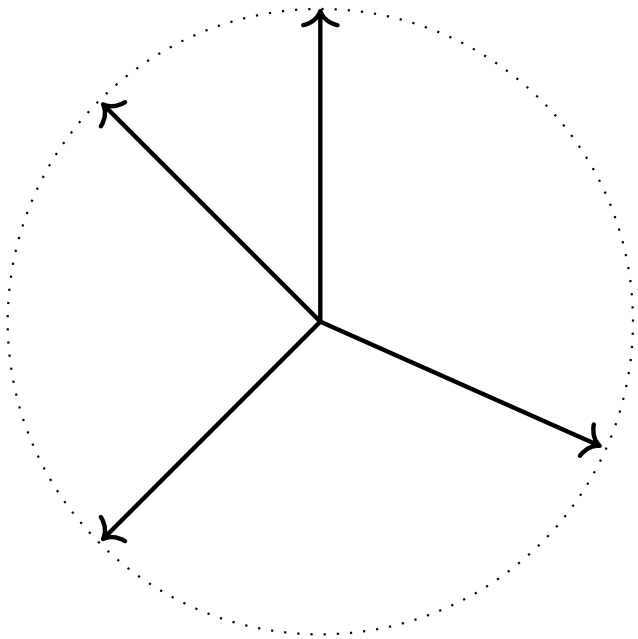
Uniformity testing: Lipschitz Extension



If X drawn from uniform, then the samples should be *independent*.

$$\forall i \neq j, \quad \langle x_i, x_j \rangle \quad \text{small.}$$

Uniformity testing: Lipschitz Extension



If X drawn from uniform, then the samples should be *independent*.

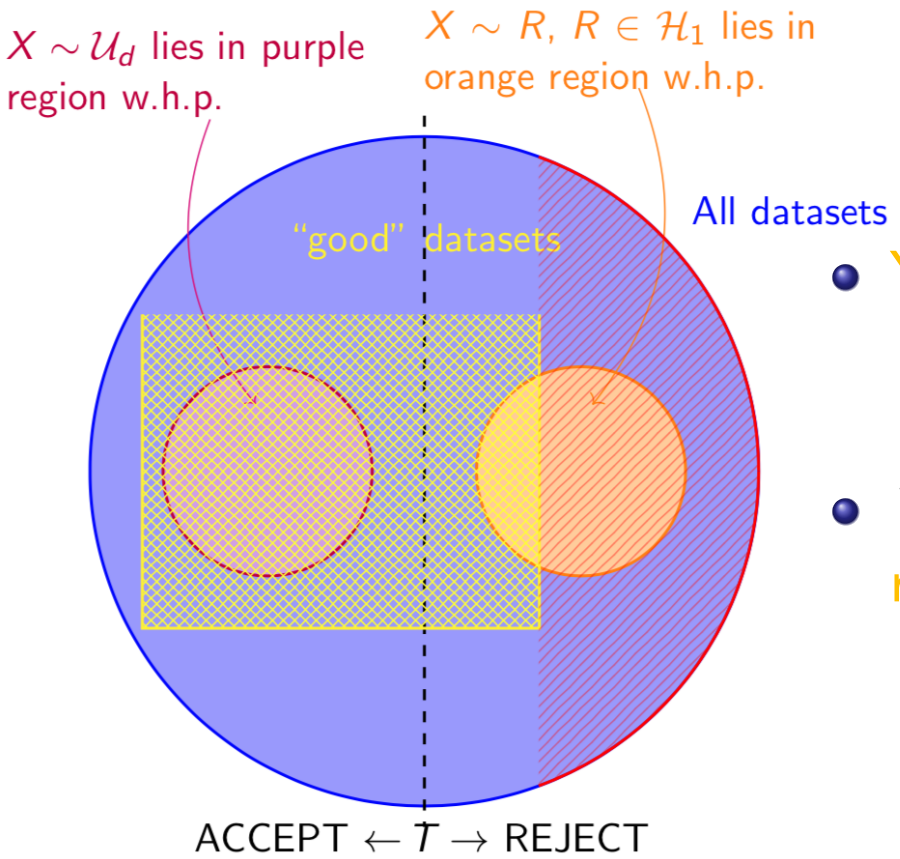
$$\forall i \neq j, \quad \langle x_i, x_j \rangle \quad \text{small.}$$

$$T(X) - T(X') = 2\left(\left\langle \frac{1}{n} x_1, \hat{\mu} \right\rangle - \left\langle \frac{1}{n} x_1', \hat{\mu}' \right\rangle\right)$$

Uniformity testing: Step 2 (Lipschitz Extension)

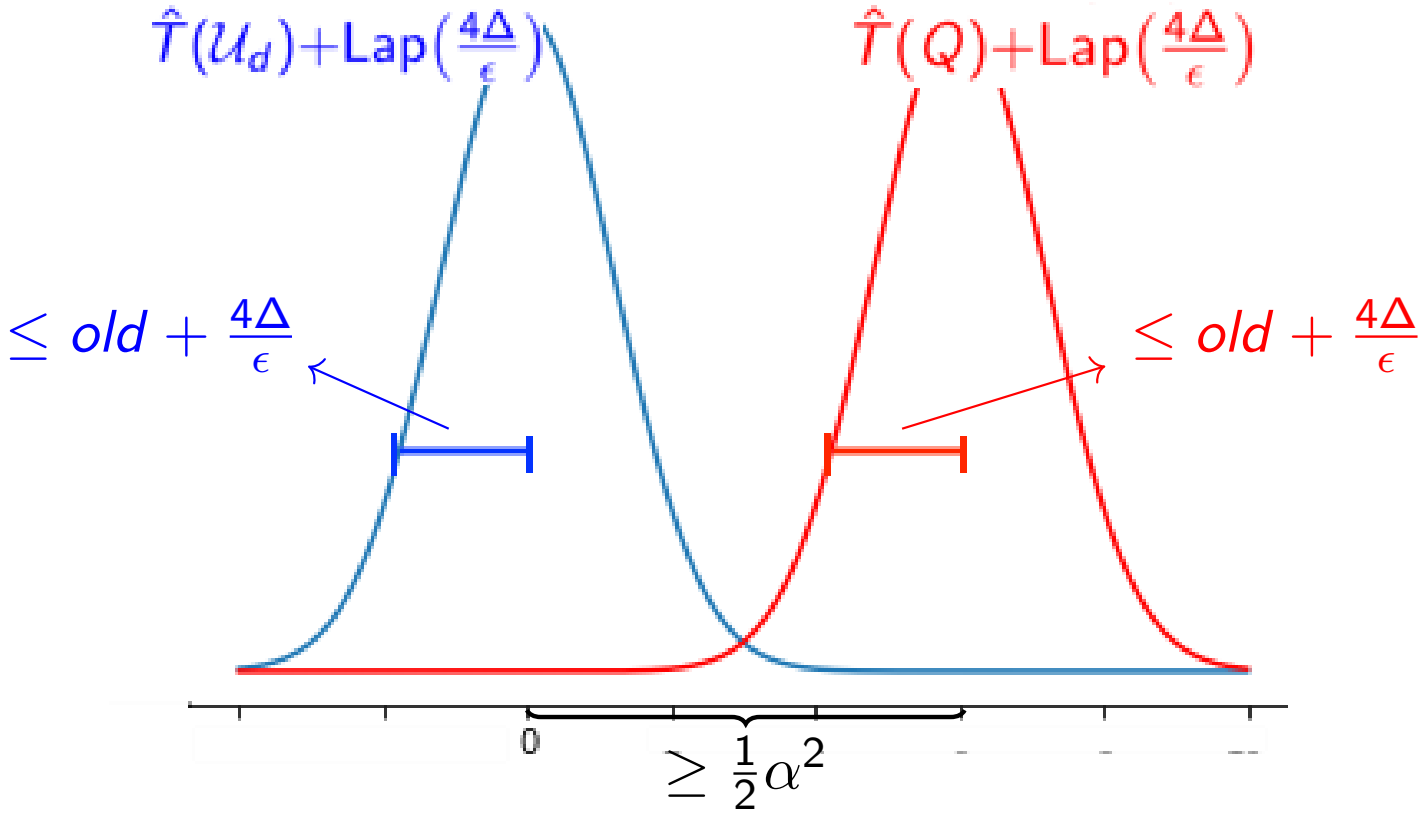
Yellow region $= \{X \mid \forall i, \langle \frac{1}{n} x_i, \hat{\mu} \rangle \leq \Delta\}$, where

$$\Delta = \tilde{O}\left(\frac{d}{n^2} + \frac{d}{n^3\epsilon^2} + \frac{\sqrt{d}}{n^{1.5}} + \frac{\sqrt{d}}{n^2\epsilon}\right) \ll \frac{d}{n}$$



- Yellow region contains datasets that
  - survive filtering
  - come from product distributions.
- T has sensitivity $4\Delta \ll \frac{d}{n}$ in yellow region.

Uniformity testing: Sample Complexity



$$\text{Sample Complexity} = \underbrace{O\left(\frac{d^{1/2}}{\alpha^2}\right)}_{\textit{non-private sc}} + \underbrace{\tilde{O}\left(\frac{d^{1/2}}{\alpha\epsilon^{1/2}} + \frac{d^{1/3}}{\alpha^{2/3}\epsilon} + \frac{d^{1/4}}{\alpha\epsilon}\right)}_{\textit{overhead for privacy}}$$

Conclusion

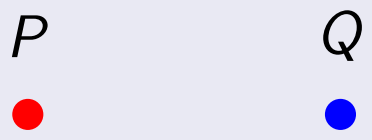
Inefficient tester

Designed an inefficient tester using *filtering* and *Lipschitz extension* for uniformity testing and Gaussian mean testing.

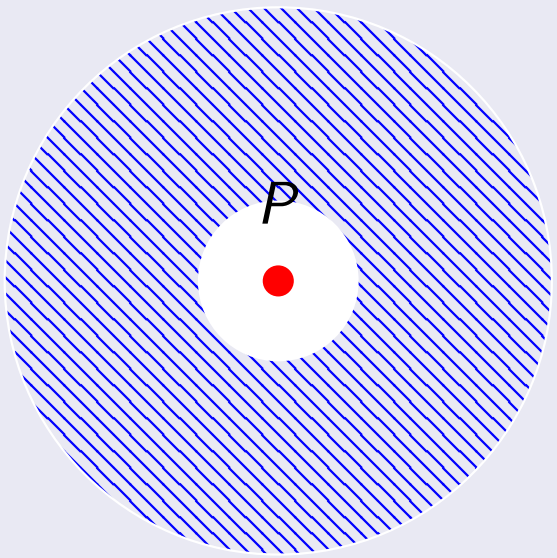
	Non-private	**Inefficient ϵ-DP**	**Efficient ϵ-DP**
Testing	[Canonne, Diakonikolas, Kane, Stewart '17] $\Theta\left(\frac{\sqrt{d}}{\alpha^2}\right)$	[This work] $O\left(\frac{\sqrt{d}}{\alpha^2} + \frac{\sqrt{d}}{\alpha\sqrt{\epsilon}} + \frac{d^{1/3}}{\alpha^{4/3}\epsilon^{2/3}} + \frac{1}{\alpha\epsilon}\right)$	[This work] $O\left(\frac{\sqrt{d}}{\alpha^2} + \frac{\sqrt{d}}{\alpha\epsilon}\right)$
Estimation	$\Theta\left(\frac{d}{\alpha^2}\right)$	[Kamath, Li, Singhal, Ullman '19] $\Theta\left(\frac{d}{\alpha^2} + \frac{d}{\alpha\epsilon}\right)$	[Kamath, Li, Singhal, Ullman '19] $\Theta\left(\frac{d}{\alpha^2} + \frac{d}{\alpha\epsilon}\right)$

Two hypothesis testing problems

Simple Hypothesis Testing



Identity Testing



- Created more robust versions of non-private tests.
- Took advantage of the structure of the problems.
- Achieved *privacy for free* in some parameter regimes.

Some Open Problems

- What is the optimal sample complexity for identity testing of product distributions?

Some Open Problems

- What is the optimal sample complexity for identity testing of product distributions?

- What are good general methods for making statistical analyses differentially private?
  - Our result: transformation of the non-private test

Some Open Problems

- What is the optimal sample complexity for identity testing of product distributions?

- What are good general methods for making statistical analyses differentially private?
  - Our result: transformation of the non-private test

Private Estimation: What is the world like?

What connections are there?

Private Testing: Is my understanding of the world correct?

- What can we learn about private estimation from private testing?
  - In the non-private and local DP settings, nice connections are known.

Some Open Problems

- What is the optimal sample complexity for identity testing of product distributions?

- What are good general methods for making statistical analyses differentially private?
  - Our result: transformation of the non-private test

Private Estimation: What is the world like?

What connections are there?

Private Testing: Is my understanding of the world correct?

- What can we learn about private estimation from private testing?
  - In the non-private and local DP settings, nice connections are known.

Thank you!