Private Hypothesis Testing via Robustness

Audra McMillan

Boston University and Northeastern University

November 11, 2019
**Estimation:** What is the world like?

**Testing:** Is my understanding of the world correct?
**Estimation:** What is the world like?

How many people actually like kale?

**Testing:** Is my understanding of the world correct?
**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?

- How many people actually like kale?
Fundamental Questions in Data Science

**Estimation:** What is the world like?

**Testing:** Is my understanding of the world correct?

- How many people actually like kale?
- Do 0% of people actually like kale?

Deep connections in the non-private world:

- Lower bounds for testing
- Lower bounds for estimation
- Under some circumstances, optimal algorithms for testing
- Optimal algorithms for estimation.
Estimation: What is the world like?

Testing: Is my understanding of the world correct?

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.
Fundamental Questions in Data Science

**Estimation:** What is the world like?

- Estimation algorithms

- Deep connections in the non-private world

**Testing:** Is my understanding of the world correct?

- Testing algorithms

- Lower bounds for testing

- Lower bounds for estimation
Fundamental Questions in Data Science

**Estimation:** What is the world like?

- Estimation algorithms
- Testing algorithms

**Testing:** Is my understanding of the world correct?

- Lower bounds for testing
- Deep connections in the non-private world
- Under some circumstances optimal algorithms for testing
- Optimal algorithms for estimation.
Fundamental Questions in Private Data Science

**Private Estimation:** What is the world like?

**Private Testing:** Is my understanding of the world correct?
Private Estimation:  What is the world like?

How can we translate results?

Private Testing:  Is my understanding of the world correct?
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?

Given a space of distributions $\Delta(\Omega)$, $H_0 \subset \Delta(\Omega)$ null hypothesis and $H_1 \subset \Delta(\Omega)$ alternate hypothesis.

Hypothesis Test

Given $x_1,\ldots,x_n \sim R$, a hypothesis test determines with high probability whether $R \in H_0$ or $R \in H_1$. 
Hypothesis Testing

Do 0% of people actually like kale?

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

- $H_0 \subset \Delta(\Omega)$ and $H_1 \subset \Delta(\Omega)$

null hypothesis

alternate hypothesis

Hypothesis Test

Given $x_1, \ldots, x_n \sim R$, a hypothesis test determines with high probability whether $R \in H_0$ or $R \in H_1$. 
A hypothesis test $T : \Omega^n \rightarrow \{\mathcal{H}_0, \mathcal{H}_1\}$ 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 $T : \Omega^n \rightarrow \{\mathcal{H}_0, \mathcal{H}_1\}$ 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 $T'$, $SC_{\mathcal{H}_0, \mathcal{H}_1}(T) = O(SC_{\mathcal{H}_0, \mathcal{H}_1}(T'))$. 
A $\epsilon$-DP hypothesis test $T : \Omega^n \rightarrow \{\mathcal{H}_0, \mathcal{H}_1\}$ 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 $T : \Omega^n \rightarrow \{\mathcal{H}_0, \mathcal{H}_1\}$ 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 $\epsilon$-DP

A test $T$ has “optimal” sample complexity if for all $\epsilon$-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**

### $\epsilon$-differential privacy

A test $T : \Omega^n \rightarrow \{\mathcal{H}_0, \mathcal{H}_1\}$ is $\epsilon$-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

$\epsilon$-differential privacy

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

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

Other stability notions

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

- $\epsilon$-DP,
- $\epsilon$-Total-variation stability (requires $P[T(x) \in S] \leq P[T(x') \in S] + \epsilon$)
- 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

- 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
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\}

\[ P \quad Q \]

**Identity testing in high dimensions**

\{U_d\} vs. \{product dist \( Q \mid TV(U_d, Q) \geq \alpha\}\}

\[ P \]

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.
Two hypothesis testing problems

**Simple hypothesis testing**

\[ \{P\} \text{ vs. } \{Q\} \]

- **P**
- **Q**

**Identity testing in high dimensions**

\[ \{U_d\} \text{ vs. } \{\text{product dist } Q \mid TV(U_d, Q) \geq \alpha\} \]

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

Simple hypothesis testing

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

\[
P \quad Q
\]

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\}

\[ P \quad Q \]

Identity testing in high dimensions

\{U_d\} vs. \{product dist \ Q \mid TV(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.
Global sensitivity = \( GS_f = \max_{X, X'} \text{neighbours} |f(X) - f(X')| \)

\[ f(X) + \text{Lap}\left(\frac{GS_f}{\epsilon}\right) \text{ is } \epsilon\text{-DP.} \]
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 $\Omega$. In a simple hypothesis test,

$$H_0 = \{P\} \text{ and } H_1 = \{Q\}.$$ 

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 $\epsilon$-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.
**Test**

\[
\text{LLR}(X) = \begin{cases} 
P & \text{if } P^n(X) \geq Q^n(X) \\
Q & \text{if } P^n(X) < Q^n(X)
\end{cases}
\]
Classical Solution

Test

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

Optimality

[Neyman-Pearson (1933)]

*Exactly* optimal sample complexity.
Classical Solution

Test

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

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**

\[ \text{LLR}(X) = \sum_{i=1}^{n} \log \frac{P(x_i)}{Q(x_i)} \]

**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 | $\text{LLR}(X) = \sum_{i=1}^{n} \log \frac{P(x_i)}{Q(x_i)}$ |
| Test | $\text{LLR}(X) = \begin{cases} P & \text{if } \text{LLR}(X) \geq 0 \\ Q & \text{if } \text{LLR}(X) < 0 \end{cases}$ |
| 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

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

Noisy Clamped Log-likelihood Test

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

**Test Statistic**

\[
c\text{LLR}(X) = \sum_{i=1}^{n} \left\{ \log \frac{P(x_i)}{Q(x_i)} \right\}^\epsilon - \epsilon
\]

**Noisy Clamped Log-likelihood Test**

\[
c\text{cLLR}(X) = \begin{cases} P & \text{if } c\text{LLR} + \text{Lap}(2) \geq 0 \\ Q & \text{otherwise} \end{cases}
\]
The Main Theorem: Optimal Private Sample Complexity

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

\[ SC^P,Q_\epsilon = \Theta \left( \min \left\{ \frac{1}{\epsilon \tau}, \frac{1}{(1 - \tau)H^2(P', Q')} \right\} \right), \]

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

\[ P' = \frac{1}{1 - \tau} \min\{P, e^{\epsilon} Q\} \quad \text{and} \quad Q' = \frac{1}{1 - \tau} \min\{Q, e^{\epsilon} P\} \]
Private identity testing in high dimensions

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

arXiv:1905.11947
$\mathcal{U}_d =$ uniform distribution on $\{-1, 1\}^d$. 

$H_0 = \{\mathcal{U}_d\}$ and $H_1 = \{Q | TV(\mathcal{U}_d, Q) \geq \alpha$ and $Q$ is a product distribution\}.

Requires $2 \Omega(d)$ samples.
Uniformity Testing in High Dimensions

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

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

Requires \(2^{\Omega(d)}\) samples.
$\mathcal{U}_d = \text{uniform distribution on } \{-1,1\}^d.$

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

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

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:

\[
\begin{align*}
\mathcal{U}_d & \quad \text{if } T(X) \leq \frac{1}{4} \alpha^2 \\
TV(\mathcal{U}_d, Q) \geq \alpha & \quad \text{if } T(X) > \frac{1}{4} \alpha^2
\end{align*}
\]

\( \|\mu\|_2 \) is a proxy for \( \|\mathcal{U}_d - R\|_{TV} \)
Non-private sample complexity \cite{Canonne, Diakonikolas, Kane, Stewart '17}

\[
T(U_d) \leq O\left(\sqrt{\frac{d}{n}}\right)
\]

\[
T(Q) \leq O\left(\frac{\sqrt{d}}{n} + \frac{\sqrt{\mathbb{E}_Q[T(X)]}}{n^{1.5}}\right)
\]

Sample Complexity: \[\Theta\left(\frac{\sqrt{d}}{\alpha^2}\right)\]
### Our contributions

#### \( \epsilon \)-DP Uniformity tester

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

#### Non-private

<table>
<thead>
<tr>
<th>Efficiency</th>
<th>( \epsilon )-DP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inefficient</td>
<td>( \epsilon )-DP</td>
</tr>
</tbody>
</table>

- [Canonne, Diakonikolas, Kane, Stewart '17]
- [This work]
- [This work]

#### Testing

<table>
<thead>
<tr>
<th>Complexity</th>
<th>( \Theta \left( \sqrt{d \alpha^2} \right) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>O(N)</td>
<td>( \Theta \left( \sqrt{d \alpha^2} \right) )</td>
</tr>
<tr>
<td>( \Theta \left( d \alpha^2 \right) )</td>
<td>( \Theta \left( d \alpha^2 + d \alpha \epsilon \right) )</td>
</tr>
</tbody>
</table>

- [Kamath, Li, Singhal, Ullman '19]

#### Estimation

<table>
<thead>
<tr>
<th>Complexity</th>
<th>( \Theta \left( d \alpha^2 \right) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \Theta \left( d \alpha^2 + d \alpha \epsilon \right) )</td>
<td>( \Theta \left( d \alpha^2 + d \alpha \epsilon \right) )</td>
</tr>
</tbody>
</table>

- [Kamath, Li, Singhal, Ullman '19]
Our contributions

<table>
<thead>
<tr>
<th>$\epsilon$-DP Uniformity tester</th>
</tr>
</thead>
<tbody>
<tr>
<td>Give a test that distinguishes $\mathcal{H}_0 = {U_d}$ from $\mathcal{H}_1 = {Q \text{ product distribution} \mid TV(U_d, Q) \geq \alpha}$.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\epsilon$-DP Gaussian mean tester with known covariance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Give a test that distinguishes $\mathcal{H}_0 = {\mathcal{N}(\mu, I_d)}$ from $\mathcal{H}_1 = {\eta \mid TV(\mathcal{N}(\mu, I_d), \mathcal{N}(\eta, I_d)) \geq \alpha}$.</td>
</tr>
</tbody>
</table>
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 \( \mathcal{H}_1 = \{ \eta \mid TV(\mathcal{N}(\mu, I_d), \mathcal{N}(\eta, I_d)) \geq \alpha \} \).

<table>
<thead>
<tr>
<th></th>
<th>Non-private</th>
<th>Inefficient ϵ-DP</th>
<th>Efficient ϵ-DP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Testing</strong></td>
<td>[Canonne, Diakonikolas, Kane, Stewart ’17] ( \Theta \left( \frac{\sqrt{d}}{\alpha^2} \right) )</td>
<td>[This work]</td>
<td>[This work]</td>
</tr>
<tr>
<td><strong>Estimation</strong></td>
<td>( \Theta \left( \frac{d}{\alpha^2} \right) )</td>
<td>( \Theta \left( \frac{d}{\alpha^2} + \frac{d}{\alpha \epsilon} \right) )</td>
<td>( \Theta \left( \frac{d}{\alpha^2} + \frac{d}{\alpha \epsilon} \right) )</td>
</tr>
</tbody>
</table>

[Canonne, Diakonikolas, Kane, Stewart ’17]
[Kamath, Li, Singhal, Ullman ’19]
Our contributions

**ϵ-DP Uniformity tester**
Give a test that distinguishes $\mathcal{H}_0 = \{U_d\}$ from $\mathcal{H}_1 = \{Q \text{ product distribution } | \ T\!V(U_d, Q) \geq \alpha\}$.

**ϵ-DP Gaussian mean tester with known covariance**
Give a test that distinguishes $\mathcal{H}_0 = \{\mathcal{N}(\mu, I_d)\}$ from $\mathcal{H}_1 = \{\eta \ | \ T\!V(\mathcal{N}(\mu, I_d), \mathcal{N}(\eta, I_d)) \geq \alpha\}$.

<table>
<thead>
<tr>
<th></th>
<th>Non-private</th>
<th>Inefficient ϵ-DP</th>
<th>Efficient ϵ-DP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Testing</td>
<td>[Canonne, Diakonikolas, Kane, Stewart '17]</td>
<td>$\Theta\left(\frac{\sqrt{d}}{\alpha^2}\right)$</td>
<td>$[\text{This work}]$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$O\left(\frac{\sqrt{d}}{\alpha^2} + \frac{\sqrt{d}}{\alpha\epsilon}\right)$</td>
</tr>
<tr>
<td>Estimation</td>
<td>$\Theta\left(\frac{d}{\alpha^2}\right)$</td>
<td>$[\text{This work}]$</td>
<td>$\Theta\left(\frac{d}{\alpha^2} + \frac{d}{\alpha\epsilon}\right)$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$[\text{Kamath, Li, Singhal, Ullman '19}]$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\Theta\left(\frac{d}{\alpha^2} + \frac{d}{\alpha\epsilon}\right)$</td>
</tr>
</tbody>
</table>
Our contributions

**ϵ-DP Uniformity tester**

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

**ϵ-DP Gaussian mean tester with known covariance**

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

<table>
<thead>
<tr>
<th></th>
<th>Non-private</th>
<th>Inefficient ϵ-DP</th>
<th>Efficient ϵ-DP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Testing</strong></td>
<td>[Canonne, Diakonikolas, Kane, Stewart ’17] $\Theta\left(\frac{\sqrt{d}}{\alpha^2}\right)$</td>
<td>$O\left(\frac{\sqrt{d}}{\alpha^2} + \frac{\sqrt{d}}{\alpha \epsilon} + \frac{d^{1/3}}{\alpha^4/3 \epsilon^{2/3}} + \frac{1}{\alpha \epsilon}\right)$</td>
<td>$O\left(\frac{\sqrt{d}}{\alpha^2} + \frac{\sqrt{d}}{\alpha \epsilon}\right)$</td>
</tr>
<tr>
<td><strong>Estimation</strong></td>
<td>$\Theta\left(\frac{d}{\alpha^2}\right)$</td>
<td>$\Theta\left(\frac{d}{\alpha^2} + \frac{d}{\alpha \epsilon}\right)$</td>
<td>$\Theta\left(\frac{d}{\alpha^2} + \frac{d}{\alpha \epsilon}\right)$</td>
</tr>
</tbody>
</table>

[Canonne, Diakonikolas, Kane, Stewart ’17] [This work] [This work] [Kamath, Li, Singhal, Ullman ’19] [Kamath, Li, Singhal, Ullman ’19]
### Our contributions

<table>
<thead>
<tr>
<th></th>
<th>Non-private</th>
<th>Inefficient $\epsilon$-DP</th>
<th>Efficient $\epsilon$-DP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Testing</strong></td>
<td>$\Theta \left( \frac{\sqrt{d}}{\alpha^2} \right)$</td>
<td>$O \left( \frac{\sqrt{d}}{\alpha^2} + \frac{\sqrt{d}}{\alpha \sqrt{\epsilon}} + \frac{d^{1/3}}{\alpha^4/3 \epsilon^{2/3}} + \frac{1}{\alpha \epsilon} \right)$</td>
<td>$O \left( \frac{\sqrt{d}}{\alpha^2} + \frac{\sqrt{d}}{\alpha \epsilon} \right)$</td>
</tr>
<tr>
<td><strong>Estimation</strong></td>
<td>$\Theta \left( \frac{d}{\alpha^2} \right)$</td>
<td>$\Theta \left( \frac{d}{\alpha^2} + \frac{d}{\alpha \epsilon} \right)$</td>
<td>$\Theta \left( \frac{d}{\alpha^2} + \frac{d}{\alpha \epsilon} \right)$</td>
</tr>
</tbody>
</table>

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

\[ T(X) = \|\hat{\mu}\|^2_2 - \frac{d}{n}, \quad T(X) - T(X') = 2 \left( \left\langle \frac{1}{n} x_1, \hat{\mu} \right\rangle - \left\langle \frac{1}{n} x'_1, \hat{\mu}' \right\rangle \right) \]
First Attempt: Global Sensitivity

\[ T(X) = \|\hat{\mu}\|_2^2 - \frac{d}{n}, \quad T(X) - T(X') = 2 \left( \left\langle \frac{1}{n}x_1, \hat{\mu} \right\rangle - \left\langle \frac{1}{n}x'_1, \hat{\mu}' \right\rangle \right) \]

Worst case

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

Worst Case:

\[ T(X) - T(X') = \Theta \left( \frac{d}{n} \right) \]

All datasets

\( X \sim U \) lies in purple region w.h.p.
Typical Sensitivity

\[ T(X) - T(X') = \Theta \left( \frac{d}{n} \right) \]

\( X \sim U_d \) lies in purple region w.h.p.
Typical Sensitivity

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

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

Worst Case:
\[ T(X) - T(X') = \Theta \left( \frac{d}{n} \right) \]
Key Idea: Filtering and Lipschitz Extensions

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

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

All datasets

ACCEPT $\leftarrow T \rightarrow$ REJECT
Key Idea: Filtering and Lipschitz Extensions

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

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

"good" datasets

ACCEPT \[ \leftarrow T \rightarrow \text{REJECT} \]

All datasets
Key Idea: Filtering and Lipschitz Extensions

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

"good" datasets

Conditions on yellow region
- $\exists$ insensitive test $T_0$.

ACCEPt $\leftarrow T \rightarrow$ REJECT
Key Idea: Filtering and Lipschitz Extensions

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

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

All datasets

“good” datasets

Conditions on yellow region

- \exists \text{ insensitive test } T_0.
- T = \text{ low sensitivity, } \lambda, \text{ inside yellow region}

Step 1: Filtering

Use } T_0 \text{ 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.
- $\text{GS}(\hat{T}) = \text{sensitivity of } T|_{\text{yellow region}}$.
- $\hat{T}(X) = T(X)$ for $X \in \text{yellow region}$.

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

“good” datasets

ACCEPT $\leftarrow T \rightarrow$ REJECT
Key Idea: Lipschitz Extension

There exists a function \( \hat{T} \) such that:
- \( \hat{T} \) is defined on all datasets.
- \( \text{GS}(\hat{T}) = \text{sensitivity of } T \mid \text{yellow region} \).
- \( \hat{T}(X) = T(X) \) for \( X \in \text{yellow region} \).

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

\[ X \sim U_d \text{ lies in purple region w.h.p.} \]
\[ X \sim R, R \in \mathcal{H}_1 \text{ lies in orange region w.h.p.} \]

McShane–Whitney extension theorem

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

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

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.
- $\text{GS}(\hat{T}) = \text{sensitivity of } T|_{\text{yellow region}}$.
- $\hat{T}(X) = T(X)$ for $X \in \text{yellow region}$.

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

All datasets $X \sim U_d$ lies in purple region w.h.p.

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

Step 1: Filtering
Insensitive test, $T_0$

Step 2: Lipschitz Extension
Use test statistic $\hat{T}(X) + \text{Lap}(\lambda \epsilon)$

Sample complexity increases due to use of noisy statistic
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.

Step 1: Filtering

Insensitive test, $T_0$

"good" datasets
The algorithm

\[ X \sim \mathcal{U}_d \text{ lies in purple region w.h.p.} \]
\[ X \sim R, R \in \mathcal{H}_1 \text{ lies in orange region w.h.p.} \]

**Step 1: Filtering**
- Insensitive test, \( T_0 \)

**Step 2: Lipschitz Extension**
- Use test statistic
  \[ \hat{T}(X) + \text{Lap} \left( \frac{\lambda}{\epsilon} \right) \]

ACCEPT \( \leftarrow T \rightarrow \) REJECT
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.

**Step 1: Filtering**

Insensitive test, \( T_0 \)

**Step 2: Lipschitz Extension**

Use test statistic

\[
\hat{T}(X) + \text{Lap}\left(\frac{\lambda}{\epsilon}\right)
\]

Sample complexity increases due to use of noisy statistic

ACCEPT \( \leftrightarrow T \rightarrow \) REJECT

"good" datasets
The algorithm

$X \sim U_d$ lies in purple region w.h.p.

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

**Step 1: Filtering**
Insensitive test, $T_0$

**Step 2: Lipschitz Extension**
Use test statistic

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

Sample complexity increases due to use of noisy statistic

$\text{ACCEPT} \leftarrow T \rightarrow \text{REJECT}$
Uniformity Testing

\[ T(X) = \|\hat{\mu}\|_2^2 - \frac{d}{n}, \quad T(X) - T(X') = 2 \left( \langle \frac{1}{n} x_1, \hat{\mu} \rangle - \langle \frac{1}{n} x'_1, \hat{\mu}' \rangle \right) \]

Worst case

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

<table>
<thead>
<tr>
<th>( X_1 )</th>
<th>1 1 1 1 1 1 1 1 1 1</th>
<th>( X_1 )</th>
<th>-1 1 1 1 1 1 1 1 1 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Uniformity Testing

\[ T(X) = \| \hat{\mu} \|_2^2 - \frac{d}{n}, \quad T(X) - T(X') = 2 \left( \left\langle \frac{1}{n} x_1, \hat{\mu} \right\rangle - \left\langle \frac{1}{n} x'_1, \hat{\mu}' \right\rangle \right) \]

Worst case

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

Problems:
- Too biased.
Uniformity Testing

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

Worst case

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

Problems:

- Too biased.
- Coordinates are not independent.
Uniformity Testing

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

Worst case

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

Problems:

- Too biased.
- Coordinates are not independent.
Uniformity Testing

\[ T(X) = \|\hat{\mu}\|_2^2 - \frac{d}{n}, \quad T(X) - T(X') = 2 \left( \left\langle \frac{1}{n} x_1, \hat{\mu} \right\rangle - \left\langle \frac{1}{n} x'_1, \hat{\mu}' \right\rangle \right) \]

Worst case

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

Problems:

- Too biased.
- Coordinates are not independent.
Step 1: Filtering

Reject if any coordinate has too high bias.

\[ \mu \approx \frac{\log(d)}{\sqrt{n}} + \frac{1}{\epsilon n} \]
Uniformity testing: Lipschitz Extension

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

$$\forall i \neq j, \langle x_i, x_j \rangle \text{ small.}$$
Uniformity testing: Lipschitz Extension

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

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

$$T(X) - T(X') = 2 \left( \left\langle \frac{1}{n} x_1, \hat{\mu} \right\rangle - \left\langle \frac{1}{n} x'_1, \hat{\mu}' \right\rangle \right)$$
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 and come from product distributions. 

\(T\) has sensitivity \(4\Delta \ll \frac{d}{n}\) in yellow region.
Uniformity testing: Sample Complexity

\[ \hat{T}(U_d) + \text{Lap}\left(\frac{4\Delta}{\epsilon}\right) \leq \text{old} + \frac{4\Delta}{\epsilon} \]

\[ \hat{T}(Q) + \text{Lap}\left(\frac{4\Delta}{\epsilon}\right) \leq \text{old} + \frac{4\Delta}{\epsilon} \]

Sample Complexity = \( O \left( \frac{d^{1/2}}{\alpha^2} \right) \) + \( \tilde{O} \left( \frac{d^{1/2}}{\alpha \epsilon^{1/2}} + \frac{d^{1/3}}{\alpha^{2/3} \epsilon} + \frac{d^{1/4}}{\alpha \epsilon} \right) \)

- non-private sc
- overhead for privacy
### Inefficient tester

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

<table>
<thead>
<tr>
<th></th>
<th>Non-private</th>
<th>Inefficient $\epsilon$-DP</th>
<th>Efficient $\epsilon$-DP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Testing</strong></td>
<td>$\Theta\left(\frac{\sqrt{d}}{\alpha^2}\right)$</td>
<td>$O\left(\frac{\sqrt{d}}{\alpha^2} + \frac{\sqrt{d}}{\alpha\sqrt{\epsilon}} + \frac{d^{1/3}}{\alpha^{4/3}\epsilon^{2/3}} + \frac{1}{\alpha\epsilon}\right)$</td>
<td>$O\left(\frac{\sqrt{d}}{\alpha^2} + \frac{\sqrt{d}}{\alpha\epsilon}\right)$</td>
</tr>
<tr>
<td>[Canonne, Diakonikolas, Kane, Stewart '17]</td>
<td>[This work]</td>
<td></td>
<td>[This work]</td>
</tr>
<tr>
<td><strong>Estimation</strong></td>
<td>$\Theta\left(\frac{d}{\alpha^2}\right)$</td>
<td>$\Theta\left(\frac{d}{\alpha^2} + \frac{d}{\alpha\epsilon}\right)$</td>
<td>$\Theta\left(\frac{d}{\alpha^2} + \frac{d}{\alpha\epsilon}\right)$</td>
</tr>
<tr>
<td>[Kamath, Li, Singhal, Ullman '19]</td>
<td></td>
<td></td>
<td>[Kamath, Li, Singhal, Ullman '19]</td>
</tr>
</tbody>
</table>
Two hypothesis testing problems

Simple Hypothesis Testing

- $P$
- $Q$

Identity Testing

- $P$

- 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

Privacy Estimation: What is the world like?

What connections are there?

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

Thank you!