Privacy-Preserving Statistical Learning and Testing

Huanyu Zhang August 22, 2019

Microsoft Research, Redmond

- 1. Introduction and Motivation
- 2. Differentially Private Identity Testing
- 3. Differentially Private Property Estimation

Introduction and Motivation

Classical statistical learning and testing problem:

- Distribution learning
 - Estimating the bias of a coin
- Hypothesis testing
 - Testing whether a coin is fair
- Property estimation
 - Estimating the Shannon entropy





The Era of Big Data



2.5 quintillion(2.5×10^{18}) bytes of data are generated everyday¹.

Huge success for ML and statistics, but new challenges.

¹Data Never sleeps 6.0 by Domo, 2018

Modern Challenges

Large domain, small sample

- Distributions over large domains/high dimensions
- Expensive data
- Sample complexity

Privacy

- Samples contain sensitive information
- Perform testing or learning while preserving privacy

Data may contain **sensitive** information.

Medical studies:

- Learn behavior of genetic mutations
- Contains health records or disease history

Navigation:

- Suggests routes based on aggregate positions of individuals
- Position information indicates users' residence

We want to explore privacy-sample complexity tradeoff.

- Sample complexity of non-private algorithm
- Additional cost due to privacy

Question: Is privacy expensive, cheap or even free?

Differential Privacy (DP) [Dwork et al., 2006]

 \hat{f} is ε -DP for any X^n and Y^n , with $d_{Ham}(X^n, Y^n) \leq 1$, for all measurable S,

$$\frac{\Pr\left(\hat{f}(X^n)\in S\right)}{\Pr\left(\hat{f}(Y^n)\in S\right)}\leq e^{\varepsilon}.$$



DP is widely adopted by the industry, e.g., Microsoft, and Google.

From Non-private Algorithm to Private Algorithm

Sensitivity. The *sensitivity* of a non-private estimator *f* is

$$\Delta_{n,f} := \max_{d_{Ham}(X^n, Y^n) \leq 1} \left| f(X^n) - f(Y^n) \right|.$$

Laplace Mechanism [Dwork et al., 2006]:

- Design a non-private estimator with low sensitivity
- Privatize this estimator by adding Laplace noise $X \sim Lap(\Delta_{n,f}/arepsilon)$

This talk will contain the following two works:

- Jayadev Acharya, Ziteng Sun, **Huanyu Zhang**, Differentially Private Testing of Identity and Closeness of Discrete Distributions, Spotlight presentation at NeurIPS 2018.
- Jayadev Acharya, Gautam Kamath, Ziteng Sun, **Huanyu Zhang**, INSPECTRE: Privately Estimating the Unseen, ICML 2018.

Differentially Private Identity Testing

Polish lottery Multilotek

- Choose "uniformly" at random distinct 20 numbers out of 1 to 80.
- Is the lottery fair?



No! Probability of 50 - 59 too small!



The plot credits to "Statistics vs Big Data" by Constantinos Daskalakis.

Identity Testing (IT), Goodness of Fit

- $[k] := \{0, 1, 2, ..., k 1\}$
- q : a known distribution
- Given $X^n := X_1 \dots X_n$ independent samples from **unknown** p
- Is p = q?
- Tester: $\mathcal{A}: [k]^n o \{0,1\}$, which satisfies the following:

With probability at least 2/3,
$$\mathcal{A}(X^n) = \begin{cases} 1, \text{ if } p = q\\ 0, \text{ if } |p - q|_{TV} > \alpha \end{cases}$$

• Sample complexity: Smallest n where such a tester exists

Non-private:

$$S(IT) = \Theta\left(\frac{\sqrt{k}}{\alpha^2}\right)$$
 [Paninski, 2008]

• Lower bound intuition: Birthday Paradox

$$\varepsilon$$
-DP algorithms: $S(IT, \varepsilon) = O\left(\frac{\sqrt{k}}{\alpha^2} + \frac{\sqrt{k \log k}}{\alpha^{3/2}\varepsilon}\right)$ [Cai et al., 2017]

Problem: based on a χ^2 -test, which has **high sensitivity**.

Theorem
$$S(IT,\varepsilon) = \Theta\left(\frac{\sqrt{k}}{\alpha^2} + \max\left\{\frac{k^{1/2}}{\alpha\varepsilon^{1/2}}, \frac{k^{1/3}}{\alpha^{4/3}\varepsilon^{2/3}}, \frac{1}{\alpha\varepsilon}\right\}\right).$$

Theorem
$$S(IT,\varepsilon) = \Theta\left(\frac{\sqrt{k}}{\alpha^2} + \max\left\{\frac{k^{1/2}}{\alpha\varepsilon^{1/2}}, \frac{k^{1/3}}{\alpha^{4/3}\varepsilon^{2/3}}, \frac{1}{\alpha\varepsilon}\right\}\right).$$

- When $\varepsilon \to \infty$, $S(IT, \varepsilon) = \Theta\left(\frac{\sqrt{k}}{\alpha^2}\right)$.
- When k is large, $S(IT, \varepsilon) = \Theta\left(\frac{\sqrt{k}}{\alpha^2} + \frac{k^{1/2}}{\alpha\varepsilon^{1/2}}\right)$, which is strictly better than the previous result!

Theorem
$$S(IT,\varepsilon) = \Theta\left(\frac{\sqrt{k}}{\alpha^2} + \max\left\{\frac{k^{1/2}}{\alpha\varepsilon^{1/2}}, \frac{k^{1/3}}{\alpha^{4/3}\varepsilon^{2/3}}, \frac{1}{\alpha\varepsilon}\right\}\right).$$

New algorithms for achieving upper bounds

New methodology to prove lower bounds for hypothesis testing

Uniformity Testing (UT): Identity testing when q is a uniform distribution over [k].

Uniformity Testing (UT): Identity testing when q is a uniform distribution over [k]. [Goldreich, 2016] In the non-private case: Up to constant factors,

S(IT) = S(UT)

Uniformity Testing (UT): Identity testing when q is a uniform distribution over [k]. [Goldreich, 2016] In the non-private case: Up to constant factors,

S(IT) = S(UT)

We proved this also hold for the private case: Up to constant factors,

 $S(IT,\varepsilon) = S(UT,\varepsilon)$

Uniformity Testing (UT): Identity testing when q is a uniform distribution over [k]. [Goldreich, 2016] In the non-private case: Up to constant factors,

S(IT) = S(UT)

We proved this also hold for the private case: Up to constant factors,

$$S(IT,\varepsilon) = S(UT,\varepsilon)$$

It would be sufficient to only consider uniformity testing.

Warm Up - Binary Case (Non-private)

Let q = B(0.5), p = B(b). Test whether b = 0.5 or α away.

Algorithm (hard threshold):

- 1. Let $M_1(X^n)$ be the number of 1's in the samples,
- 2. If $\frac{1}{n} |M_1(X^n) \frac{n}{2}| \le \frac{\alpha}{2}$, output b = 0.5,
- 3. Else, **output** $b \neq 0.5$.

Analysis:

• Expectation Gap:

 $\mathbb{E}_{X^n \sim B(0.5+\alpha)} \left[M_1(X^n) \right] - \mathbb{E}_{X^n \sim B(0.5)} \left[M_1(X^n) \right] \ge \alpha n.$

- Variance of $M_1(X^n)$: Var $(M_1(X^n)) = O(n)$.
- By Chebyshev's inequality, the sample complexity is $O(\frac{1}{\alpha^2})$.

Let q = B(0.5), p = B(b). Test whether b = 0.5 or α away.

Algorithm (soft threshold):

- 1. Let $Z(X^n) = M_1(X^n) \frac{n}{2}$,
- 2. Generate $Y \sim B(\sigma(\varepsilon \cdot (|Z(X^n)| \frac{\alpha n}{2})))$, σ sigmoid function,
- 3. If Y = 0, **output** b = 0.5,
- 4. Else, **output** $b \neq 0.5$.

Lemma

The Algorithm is ε -DP. It has error probability at most 0.1, with $O(\frac{1}{\alpha^2} + \frac{1}{\alpha\varepsilon})$ samples.

Reminder: $Y \sim B(\sigma(\varepsilon \cdot (|M_1(X^n) - \frac{n}{2}| - \frac{\alpha n}{2})))$

Proof idea:

- Privacy: For all $x, \gamma \in \mathbb{R}$, $\exp(-|\gamma|) \le \frac{\sigma(x+\gamma)}{\sigma(x)} \le \exp(|\gamma|)$.
- Sample complexity :
 - 1. Consider the case when b = 0.5,
 - 2. $Z(X^n) = O(\sqrt{n})$ with high probability (**Chebyshev**),
 - 3. Given $n = O\left(\frac{1}{\alpha^2}\right)$, $\frac{\alpha n}{2} |Z(X^n)| = O(\alpha n)$,
 - 4. Given $n = O(\frac{1}{\alpha \varepsilon})$, $\varepsilon(|Z(X^n)| \frac{\alpha n}{2}) < -1000$.
 - 5. Similar argument works for the case when $|b 0.5| > \alpha$.

Idea: Privatizing the statistic used by [Diakonikolas et al., 2017].

Let M_x be the number of samples of x,

$$S(X^n):=\frac{1}{2}\cdot\sum_{x=1}^k\left|\frac{M_x(X^n)}{n}-\frac{1}{k}\right|.$$

- Sample optimal in the non-private case.
- This statistic also has a small sensitivity!

 $S(X^n)$ has the following two properties:

• Expectation gap [Diakonikolas et al., 2017]:

let
$$\mu(p) = \mathbb{E}_{X^n \sim p} [S(X^n)]$$
, if $d_{TV}(u[k], p) > \alpha$,

$$\mu(p) - \mu(u[k]) \ge c\alpha^2 \min\left\{\frac{n^2}{k^2}, \sqrt{\frac{n}{k}}, \frac{1}{\alpha}\right\}.$$

• Small sensitivity:

 $\forall X^n$, Y^n with $d_{Ham}(X^n, Y^n) \leq 1$, we have:

$$|S(X^n) - S(Y^n)| \le \min\left(\frac{1}{n}, \frac{1}{k}\right).$$

Algorithm 1: Private Uniformity Testing **Input:** ε , α , i.i.d. samples X^n from pLet $Z(X^n)$ be defined as follows:

$$Z(X^n) := \begin{cases} k \left(S(X^n) - \mu(u[k]) - \frac{1}{2}c\alpha^2 \cdot \frac{n^2}{k^2} \right), & \text{when } n \leq k, \\ n(S(X^n) - \mu(u[k]) - \frac{1}{2}c\alpha^2 \cdot \sqrt{\frac{n}{k}} \right), & \text{when } k < n \leq \frac{k}{\alpha^2}, \\ n(S(X^n) - \mu(u[k]) - \frac{1}{2}c\alpha), & \text{when } n \geq \frac{k}{\alpha^2}. \end{cases}$$

Generate $Y \sim B(\sigma(\varepsilon \cdot Z(X^n)))$, σ is the sigmoid function. if Y = 0, return p = u[k], else return $p \neq u[k]$

Similar analysis also works here!

Lemma

Suppose there is a coupling between p and q over \mathcal{X}^n (not necessarily i.i.d.), such that $\mathbb{E}[d_{Ham}(X^n, Y^n)] \leq D$.

Then, any ε -differentially private hypothesis testing algorithm satisfies

$$\varepsilon = \Omega\left(\frac{1}{D}\right).$$

For any distribution p_1 and p_2 over \mathcal{X} with $d_{TV}(p_1, p_2) = \alpha$, if we draw *n* samples i.i.d., there exists coupling with **expected Hamming distance** $O(\alpha n)$. Then we have $n = \Omega(\frac{1}{\alpha \epsilon})$.

If we take $p_1 = B(0.5)$ and $p_2 = B(0.5 + \alpha)$, we get the exact lower bound for binary case.

Problem: This bound doesn't contain any dependency on *k*!

Lemma

Suppose there is a coupling between p and q over \mathcal{X}^n (not necessarily i.i.d.), such that $\mathbb{E}[d_{Ham}(X^n, Y^n)] \leq D$.

Then, any ε -differentially private hypothesis testing algorithm satisfies

$$\varepsilon = \Omega\left(\frac{1}{D}\right).$$

Use LeCam's two-point method.

Construct two hypotheses and a coupling between them with small expected Hamming distance.

Lower Bound - Proof Sketch

- Design the following hypothesis testing problem, q: draw n i.i.d. samples from u[k].
 - p: a mixture of distributions:
 - 1. generate the set of $2^{k/2}$ distributions, where for each $\mathbf{z} \in \{\pm 1\}^{k/2}$,

$$p_{\mathsf{z}}(2i-1) = \frac{1+\mathsf{z}_i \cdot 2\alpha}{k}$$
, and $p_{\mathsf{z}}(2i) = \frac{1-\mathsf{z}_i \cdot 2\alpha}{k}$.

- uniformly pick up one distribution, and generate n i.i.d. samples according to it.
- Bound the coupling distance of uniform to mixture,

$$\mathbb{E}\left[d_{Ham}(X^n,Y^n)\right] \leq C \cdot \alpha^2 \min\left\{\frac{n^2}{k},\frac{n^{3/2}}{k^{1/2}}\right\}.$$

• Prove a lower bound by our coupling theorem.

Some Intuition when Sparse

- Consider the following two distribution:
 - 1. $p_1 = B(0.5)$,
 - 2. p_2 is a uniform mixture of $B(\frac{1}{2} \alpha)$ and $B(\frac{1}{2} + \alpha)$.
- If we draw (t ≥ 2) samples, d_{TV}(p₁, p₂) ≤ 2tα² and the expected hamming distance is bounded by 2t²α².
- Now we consider the coupling between p and q, for every pair of symbols, roughly appear 2n/k times in total.
- Therefore, the total coupling distance is $\frac{k}{2} \cdot \frac{4n^2\alpha^2}{k^2} = O\left(\frac{n^2\alpha^2}{k}\right)$.

Closeness Testing (CT), Two Sample Test

- $[k] = \{0, 1, 2, \dots, k 1\}$ is a discrete set of size k.
- *p*, *q* two **unknown** distributions over [*k*].
- $X^n = (X_1, X_2, ..., X_n) : n$ independent samples from p.
- $Y^n = (Y_1, Y_2, ..., Y_n) : n$ independent samples from q.
- Tester: $\mathcal{A}: [k]^n imes [k]^n o \{0,1\}$, which satisfies the following:

With probability at least 2/3, $\mathcal{A}(X^n,Y^n) = \begin{cases} 1, \text{ if } p = q\\ 0, \text{ if } |p-q|_{TV} > \alpha \end{cases}$

Closeness Testing (CT), Two Sample Test

- $[k] = \{0, 1, 2, \dots, k-1\}$ is a discrete set of size k.
- p, q two **unknown** distributions over [k].
- $X^n = (X_1, X_2, ..., X_n) : n$ independent samples from p.
- $Y^n = (Y_1, Y_2, ..., Y_n) : n$ independent samples from q.
- Tester: $\mathcal{A}: [k]^n imes [k]^n o \{0,1\}$, which satisfies the following:

With probability at least 2/3,
$$\mathcal{A}(X^n,Y^n) = egin{cases} 1, ext{ if } p=q \ 0, ext{ if } |p-q|_{TV} > lpha \end{cases}$$

$$S(CT) = \Theta\left(k^{2/3}/lpha^{4/3} + \sqrt{k}/lpha^2
ight)$$
 [Chan et al., 2014]

Theorem
$$S(CT,\varepsilon) = O\left(\max\left\{\frac{k^{2/3}}{\alpha^{4/3}} + \frac{\sqrt{k}}{\alpha\sqrt{\varepsilon}}, \frac{\sqrt{k}}{\alpha^2} + \frac{1}{\alpha^2\varepsilon}\right\}\right).$$

- When $\varepsilon \to \infty$, $S(CT, \varepsilon) = O\left(\frac{k^{2/3}}{\alpha^{4/3}} + \frac{\sqrt{k}}{\alpha^2}\right)$.
- When k is large, $S(CT, \varepsilon) = \Theta\left(\frac{k^{2/3}}{\alpha^{4/3}} + \frac{\sqrt{k}}{\alpha\sqrt{\varepsilon}}\right)$.

Conclusion

- We establish a general coupling method to prove lower bounds in DP.
- We derive the optimal sample complexity of DP identity testing for all parameter ranges.
- We also give the sample complexity of DP closeness testing, which is optimal in sparse case.

This work was accepted as spotlight presentation at NeurIPS 2018.

Differentially Private Property Estimation

Property Estimation

- *p*: unknown discrete distribution
- f(p): some property of distribution, e.g. entropy
- α: accuracy
- Input: i.i.d. samples Xⁿ from p
- **Output** $\hat{f} : X^n \to \mathbb{R}$ such that w.p. at least 2/3:

$$\left|\widehat{f}(X^n)-f(p)\right|$$

• Sample complexity: least n to estimate f(p)

Given i.i.d. samples from distribution p, the goals are:

- Accuracy: estimate f(p) up to $\pm \alpha$ with probability $> \frac{2}{3}$
- Privacy: estimator must satisfy ε-DP

Properties of interest:

- **Entropy**, H(p): the Shannon entropy
- **Support Coverage**, *S_m(p)*: expected number of distinct symbols in *m* draws from *p*
- Support Size, S(p): # symbols with non-zero probability

Support Coverage - Motivating Example

• Corbett collected butterflies in Malaya for 1 year.

1	2	3	4	5	6	7	
118	74	44	24	29	22	20	

• Number of seen species = 118 + 74 + 44 + 24 + ...

How many new species can be found next year?

The cost of privacy in private property estimation is often **negligible**.

Theorem 1. Sample complexity of support coverage:

$$O\bigg(\frac{m\log(1/\alpha)}{\log m} + \frac{m\log(1/\alpha)}{\log(2+\varepsilon m)}\bigg).$$

Furthermore,

$$C(S_m, \alpha, \varepsilon) = \Omega\left(\frac{m\log(1/\alpha)}{\log m} + \frac{1}{\alpha\varepsilon}\right).$$

Privacy is free unless $\varepsilon < \frac{1}{\sqrt{m}}$. Similar bounds hold for other properties.

Sensitivity. The sensitivity of an estimator f is

$$\Delta_{n,f} := \max_{d_{Ham}(X^n, Y^n) \leq 1} \left| f(X^n) - f(Y^n) \right|.$$

Our algorithms use Laplace Mechanism [Dwork et al., 2006].

- Compute a non-private estimator with **low** sensitivity [Acharya et al., 2017]
- Privatize this estimator by adding Laplace noise $X \sim Lap(\Delta_{n,f}/\varepsilon)$

We borrow the following non-private estimator (SGT) [Orlitsky et al., 2016] with **low sensitivity**:

$$\hat{S}_m(X^n) = \sum_{i=1}^n \Phi_i (1 + (-t)^i \cdot \Pr\left(Z \ge i\right)),$$

where Φ is the profile of X^n , $Z \sim \text{Poi}(r)$ and t = (m - n)/n.

Lemma 1. When $t \ge 1$, the sensitivity of the estimator satisfies

$$\Delta\left(\frac{\hat{S}_m(X^n)}{m}\right) \leq \frac{2}{m} \cdot \left(1 + e^{r(t-1)}\right).$$

Lemma

Suppose there is a coupling between p and q over \mathcal{X}^n , such that

 $\mathbb{E}\left[d_{Ham}(X^n,Y^n)\right] \leq D$

Then, any ε -differentially private hypothesis testing algorithm must satisfy

$$\varepsilon = \Omega\left(\frac{1}{D}\right)$$

Consider the following two distributions:

- u_1 is uniform over $[m(1 + \alpha)]$.
- u_2 is distributed over m + 1 elements $[m] \cup \{ \triangle \}$ where $u_2[i] = \frac{1}{m(1+\alpha)}, \forall i \in [m] \text{ and } u_2[\triangle] = \frac{\alpha}{1+\alpha}.$

We know

$$S_m(u_1) - S_m(u_2) = \Omega(\alpha m).$$

Moreover, their total variation distance is $\frac{\alpha}{1+\alpha}$. So the coupling distance is $\frac{m\alpha}{1+\alpha}$.

Support coverage estimation on synthetic data

- Given n = 10000 samples, then estimate the support coverage at m = n · t, t = 1, 2, ...
- Comparison on performance (RMSE) of private and non-private estimator.



Support coverage estimation on real data

- Comparison on performance (RMSE) of private and non-private estimator
- The dataset: 2000 US Census data, and Hamlet



Conclusion

- Our upper bounds show that the cost of privacy in these settings is often **negligible** compared to the non-private statistical task.
- 2. We derive lower bound for these problems by reducing them into binary hypothesis testing.
- 3. Our methods are realizable in practice, and we demonstrate their effectiveness on several synthetic and real-data examples.

This work was accepted by ICML 2018.

Thank you!

Acharya, J., Das, H., Orlitsky, A., and Suresh, A. T. (2017).
 A unified maximum likelihood approach for estimating symmetric properties of discrete distributions.
 In *ICML*, pages 11–21.

 Cai, B., Daskalakis, C., and Kamath, G. (2017).
 Privit: Private and sample efficient identity testing. In ICML.

Chan, S. O., Diakonikolas, I., Valiant, P., and Valiant, G. (2014).

Optimal algorithms for testing closeness of discrete distributions.

Diakonikolas, I., Gouleakis, T., Peebles, J., and Price, E. (2017).

Sample-optimal identity testing with high probability. *arXiv preprint arXiv:1708.02728.* Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). **Calibrating noise to sensitivity in private data analysis.** In *Proceedings of the 3rd Conference on Theory of Cryptography*, TCC '06, pages 265–284, Berlin, Heidelberg. Springer.

Goldreich, O. (2016).

The uniform distribution is complete with respect to testing identity to a fixed distribution.

In *Electronic Colloquium on Computational Complexity* (ECCC), volume 23.

Orlitsky, A., Suresh, A. T., and Wu, Y. (2016).
 Optimal prediction of the number of unseen species.
 Proceedings of the National Academy of Sciences.



Paninski, L. (2008).

A coincidence-based test for uniformity given very sparsely sampled discrete data.

54(10):4750-4755.