Differentially Private Testing of Identity and Closeness of Discrete Distributions

NeurIPS 2018, Montreal, Canada

Jayadev Acharya, Cornell University

Ziteng Sun, Cornell University

Huanyu Zhang, Cornell University

Hypothesis Testing

- Given data from an unknown statistical source (distribution)

Hypothesis Testing

- Given data from an unknown statistical source (distribution)
- Does the distribution satisfy a postulated hypothesis?

Modern Challenges

Large domain, small samples

- Distributions over large domains/high dimensions

Modern Challenges

Large domain, small samples

- Distributions over large domains/high dimensions
- Expensive data

Modern Challenges

Large domain, small samples

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

Modern Challenges

Large domain, small samples

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

Privacy

- Samples contain sensitive information

Modern Challenges

Large domain, small samples

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

Privacy

- Samples contain sensitive information
- Perform hypothesis testing while **preserving privacy**

Identity Testing (IT), Goodness of Fit

- $[k] := \{0, 1, 2, ..., k-1\}$, a discrete set of size k.

Identity Testing (IT), Goodness of Fit

- $[k] := \{0, 1, 2, ..., k-1\}$, a discrete set of size k.
- q : a **known** distribution over $[k]$.

Identity Testing (IT), Goodness of Fit

- $[k] := \{0, 1, 2, ..., k-1\}$, a discrete set of size k.
- q : a **known** distribution over $[k]$.
- Given $X^n := X_1 \ldots X_n$ independent samples from **unknown** p.

Identity Testing (IT), Goodness of Fit

- $[k] := \{0, 1, 2, ..., k-1\}$, a discrete set of size k.
- q : a **known** distribution over $[k]$.
- Given $X^n := X_1 \ldots X_n$ independent samples from **unknown** p.
- **Is $p = q$?**

Identity Testing (IT), Goodness of Fit

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

With probability at least 2/3,

$$\mathcal{A}(X^n) = \begin{cases} 1, & \text{if } p = q \\ 0, & \text{if } |p - q|_{TV} > \alpha \end{cases}$$

Identity Testing (IT), Goodness of Fit

- $[k] := \{0, 1, 2, ..., k-1\}$, a discrete set of size k.
- q : a **known** distribution over $[k]$.
- Given $X^n := X_1 \dots X_n$ independent samples from **unknown** p.
- **Is** $p = q$**?**
- Tester: $\mathcal{A} : [k]^n \to \{0, 1\}$, 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.

Identity Testing (IT), Goodness of Fit

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

With probability at least 2/3,

$$\mathcal{A}(X^n) = \begin{cases} 1, & \text{if } p = q \\ 0, & \text{if } |p - q|_{TV} > \alpha \end{cases}$$

$$S(IT) = \Theta\left(\sqrt{k}/\alpha^2\right).$$

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

A randomized algorithm $\mathcal{A} : \mathcal{X}^n \to \mathcal{S}$ is ε-differentially private if $\forall S \subset \mathcal{S}$ and $\forall X^n, Y^n$ with $d_H(X^n, Y^n) \leq 1$, we have

$$\Pr\left(\mathcal{A}(X^n) \in S\right) \leq e^{\varepsilon} \cdot \Pr\left(\mathcal{A}(Y^n) \in S\right).$$

Previous Results

Identity Testing:

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

ε-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]

Previous Results

Identity Testing:

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

ε-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]

What is the sample complexity of identity testing?

Our Results

Theorem

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

Our Results

Theorem

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

$$S(IT,\varepsilon) = \begin{cases} \Theta\left(\frac{\sqrt{k}}{\alpha^2} + \frac{k^{1/2}}{\alpha\varepsilon^{1/2}}\right), & \text{if } n \le k \\ \Theta\left(\frac{\sqrt{k}}{\alpha^2} + \frac{k^{1/3}}{\alpha^{4/3}\varepsilon^{2/3}}\right), & \text{if } k < n \le \frac{k}{\alpha^2} \\ \Theta\left(\frac{\sqrt{k}}{\alpha^2} + \frac{1}{\alpha\varepsilon}\right) & \text{if } n \ge \frac{k}{\alpha^2}. \end{cases}$$

Our Results

Theorem

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

$$S(IT,\varepsilon) = \begin{cases} \Theta\left(\frac{\sqrt{k}}{\alpha^2} + \frac{k^{1/2}}{\alpha\varepsilon^{1/2}}\right), & \text{if } n \le k \\ \Theta\left(\frac{\sqrt{k}}{\alpha^2} + \frac{k^{1/3}}{\alpha^{4/3}\varepsilon^{2/3}}\right), & \text{if } k < n \le \frac{k}{\alpha^2} \\ \Theta\left(\frac{\sqrt{k}}{\alpha^2} + \frac{1}{\alpha\varepsilon}\right) & \text{if } n \ge \frac{k}{\alpha^2}. \end{cases}$$

New algorithms for achieving upper bounds

New methodology to prove lower bounds for hypothesis testing

Upper Bound

Privatizing the statistic used by [Diakonikolas et al., 2017], which is sample optimal in the non-private case.

Independent work of [Aliakbarpour et al., 2017] gives a different upper bound.

Lower Bound - Coupling Lemma

Lemma

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

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

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

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

Lower Bound - Coupling Lemma

Lemma

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

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

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

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

Use LeCam's two-point method.

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

The End

Paper available on arxiv:

https://arxiv.org/abs/1707.05128.

See you at the poster session!

Tue Dec 4th 05:00 – 07:00 PM @ Room 210 and 230

AB #151.

Aliakbarpour, M., Diakonikolas, I., and Rubinfeld, R. (2017).
Differentially private identity and closeness testing of discrete distributions.
arXiv preprint arXiv:1707.05497.

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

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 *In Proceedings of the 3rd Theory of Cryptography Conference*.

Paninski, L. (2008).
A coincidence-based test for uniformity given very sparsely sampled discrete data.
IEEE Transactions on Information Theory, 54(10):4750–4755.