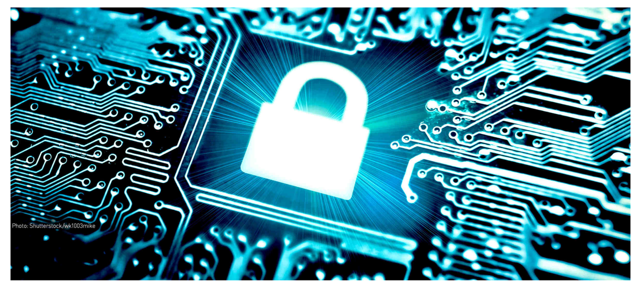
Differentially Private Testing of Identity and Closeness of Discrete Distributions

Jayadev Acharya, Ziteng Sun, Huanyu Zhang

ECE, Cornell University

Objective

Hypothesis testing with privacy constraints.



Problem Formulation

Identity Testing (IT):

- q: a *known* distribution over $[k]$
- X^n : n independent samples from an *unknown* p

Goal: Design $\mathcal{A} : [k]^n \rightarrow \{0,1\}$ such that:

$$p = q \quad \Rightarrow \mathcal{A}(X^n) = 1, \text{ w.p. } \geq 2/3,$$
$$d_{\mathrm{TV}}(p,q) > \alpha \Rightarrow \mathcal{A}(X^n) = 0, \text{ w.p. } \geq 2/3.$$

Differential Privacy (DP): $\mathcal{A}$ is ε-DP if for any X^n and Y^n, with $d_{ham}(X^n, Y^n) \leq 1$, for all measurable S,

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

Private Identity Testing: $\mathcal{A}$ should be ε-DP.

Private Closeness Testing (CT):

- X^n, Y^n : samples from p, and q, both *unknown*
- Is $p = q$, or $d_{\mathrm{TV}}(p,q) > \alpha$?

Previous Work

Previous results: $S(IT,\varepsilon) = O\left(\frac{k^{1/2}}{\alpha^2} + \frac{(k \log k)^{1/2}}{\alpha^{3/2}\varepsilon}\right)$ [1].

Independent work: $S(IT,\varepsilon) = O\left(\frac{k^{1/2}}{\alpha^2} + \frac{k^{1/2}}{\alpha\varepsilon^{1/2}}\right)$, if $n << k$ [2]

Main Results

Theorem 1. Sample complexity of identity testing:

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

we can write it according to the parameter range:

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

We give **tight** bound for all parameter ranges.

Reduction from IT to UT

Uniformity Testing (UT): Identity testing when q is a uniform distribution.

- Non-private: [3] proposes a reduction from IT to UT.
- Differential Privacy: we show that up to constant factors,

$$S(\mathrm{IT},\varepsilon) \quad = \quad S(\mathrm{UT},\varepsilon).$$

Uniformity Testing: Upper Bound

$N(x)$: the number of appearances of x in X_1^n.

Our private tester comes from privatizing and thresholding the following statistic [4]:

$$S(X_1^n) \stackrel{\text{def}}{=} \frac{1}{2} \cdot \sum \left|\frac{N(x)}{m} - \frac{1}{k}\right|.$$

This statistic has following two properties:

- Accuracy: It is optimal in non-private case.
- Small sensitivity: for all values of m, and k, we prove

$$\Delta(S) \quad \leq \quad \min\left\{\frac{1}{k}, \frac{1}{m}\right\}.$$

Privacy Bounds Via Coupling

Theorem 3. Suppose there is a coupling between $X_1^n \sim p$ and $Y_1^n \sim q$, such that

$$\mathbb{E}\left[d_{ham}(X_1^n, Y_1^n)\right] \leq D.$$

Then any ε-DP hypothesis testing algorithm $\mathcal{A}$ on p and q must satisfy $\varepsilon = \Omega\left(\frac{1}{D}\right)$.

Uniformity Testing: Lower Bound

Our proof consists of the following steps:

- Design the following hypothesis testing problem,

  Q_1: uniform distribution over $[k]$.
  Q_2: mixture of $2^{k/2}$ distributions (Paninski construction).
- Bound the coupling distance from uniform to mixture,

$$\mathbb{E}\left[d_{ham}(X_1^n, Y_1^n)\right] \leq C \cdot \alpha^2 \min\left\{\frac{n^2}{k}, \frac{n^{3/2}}{k^{1/2}}\right\}.$$

- Prove a lower bound by our coupling theorem.

Closeness Testing

Theorem 2. Sample complexity of closeness testing: If $\alpha > 1/k^{1/4}$, and $\varepsilon\alpha^2 > 1/k$ $(n < k)$,

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

otherwise,

$$\Omega\left(\frac{k^{1/2}}{\alpha^2} + \frac{k^{1/2}}{\alpha\sqrt{\varepsilon}} + \frac{1}{\alpha\varepsilon}\right) \leq S(CT,\varepsilon) \leq O\left(\frac{k^{1/2}}{\alpha^2} + \frac{1}{\alpha^2\varepsilon}\right).$$

References

[1] B. Cai, C. Daskalakis, and G. Kamath, "Priv'it: Private and sample efficient identity testing," in *ICML*, 2017.

[2] M. Aliakbarpour, I. Diakonikolas, and R. Rubinfeld, "Differentially private identity and closeness testing of discrete distributions," *arXiv preprint arXiv:1707.05497*, 2017.

[3] O. Goldreich, "The uniform distribution is complete with respect to testing identity to a fixed distribution.," in *Electronic Colloquium on Computational Complexity (ECCC)*, vol. 23, p. 1, 2016.

[4] I. Diakonikolas, D. M. Kane, and V. Nikishkin, "Testing identity of structured distributions," in *SODA*, pp. 1841–1854, 2015.