CS880: Advanced Learning Theory **Fall 2019**

Lecture 11: Tolerant Testing in ℓ_1-Distance

Lecturer: Ilias Diakonikolas *Scribes: Yuetian Luo*

1 Introduction to Tolerant Testing

Let $p, q : [n] \to [0, 1]$ be discrete distributions. In previous lectures, we have analyzed the sample complexity of the equivalence testing problem. That is, testing $p = q$ versus $\|p - q\|_1 \geq \epsilon$ with high constant probability. Our goal for this lecture is to see how the complexity of this problem changes if we relax the completeness assumption.

First let us recall that our standard ℓ_2- equivalence tester was tolerant. That is, the same tester can distinguish between the cases that $\|p - q\|_2 \leq \epsilon/2$ and $\|p - q\|_2 \geq \epsilon$ with essentially the same sample size as $\|p - q\|_2 = 0$ versus $\|p - q\|_2 \geq \epsilon$.

The reason is that the test statistic

$$Z = \sum_i \{(X_i - Y_i)^2 - X_i - Y_i\}$$

is an unbiased estimator of $\|p - q\|_2^2$.

First, it is not hard to see that our standard l_1- equivalence tester (obtained via the flattening technique) is tolerant under the χ-squared distance .

When do the flattening, we have that

$$\|p_S - q_S\|_2^2 = \sum_{i=1}^{n} \lceil nq_i \rceil (\frac{p_i - q_i}{\lceil nq_i \rceil})^2 \leq \frac{1}{n} \sum_{i=1}^{n} \frac{(p_i - q_i)^2}{q_i} \ .$$

Suppose $\chi^2(p, q) \leq \frac{\epsilon^2}{4}$, then

$$\|p_S - q_S\|_2^2 \leq \frac{\epsilon^2}{4n} = \frac{\epsilon'^2}{4} \ ,$$

where $\epsilon' = \frac{\epsilon}{\sqrt{n}}$.

So based on the above calculation and the sample complexity of the l_1-tester, we have that the sample complexity of distinguishing between the case that $\chi^2(p, q) \leq \frac{\epsilon^2}{4}$ and $\|p - q\|_1 \geq \epsilon$ is $O(\frac{\sqrt{n}}{\epsilon^2})$.

Although we have that $\|p-q\|_1^2 \leq \chi^2(p, q)$, we can still not replace the null hypothesis "$\chi^2(p, q) \leq \frac{\epsilon^2}{4}$" with "$\|p - q\|_1 \leq \frac{\epsilon}{2}$". The reason is that small ℓ_1 norm small does not imply small χ^2 distance. So we need new tools for tolerant testing w.r.t. the ℓ_1-norm.

2 Tolerant Testing in ℓ_1-norm

Fix ϵ to be a small constant.

Theorem 1. *Distinguishing between $\|p - \mathcal{U}_n\|_1 \leq \epsilon/2$ versus $\|p - \mathcal{U}_n\|_1 \geq \epsilon$ requires $\Theta(\frac{n}{\log n})$ samples.*

Today we will sketch the proof of the lower bound.

The reason for the tolerance of the ℓ_2-tester is because it can be written as a low-degree polynomial. For the l_1-tester, we have $\|p - \mathcal{U}_n\|_1 = \sum_{i=1}^n |p_i - \frac{1}{n}|$, thus our goal is to use low-degree polynomial to approximate the absolute value function.

The reason behind this is that: the only thing a testing algorithm can do in this context is to estimate moments of p_i and q_i.

The reason that moments suffice is that

$$\mathbf{E}[\sharp \text{ domain elements that get } k \text{ samples}] = \sum_{i=1}^{n} e^{-mp_i} \frac{(mp_i)^k}{k!} ,$$

since $X_i \sim Poi(mp_i)$. First, we have $e^{-mp_i} \sim 1$, and when k is small, the expectation is roughly proportional to $\sum_{i=1}^n p_i^k$, which are the k-th parameter moments.

Theorem 2. *Let p be a distribution on domain of size $[n]$. Any tester that can distinguish between $\|p - \mathcal{U}_n\|_1 \leq \frac{1}{10}$ and $\|p - \mathcal{U}_{n/2}\|_1 \leq \frac{1}{10}$ requires $\Omega(\frac{n}{\log n})$ samples. Here $\mathcal{U}_{n/2}$ is the uniform distribution on some subset of size $\frac{n}{2}$, which is a distribution over distributions.*

This is enough for Theorem 1 since $\mathcal{U}_{n/2}$ is far from $\mathcal{U}_n$.

Proof sketch: Let

$$X = \begin{cases} 0 & w.p. 1/2 \\ 1 & w.p. 1/2 \, . \end{cases}$$

If $X = 0$, take $p_i \sim A$. If $X = 1$, take $p_i \sim B$. Note that A, B are the distributions of p_i supported on $[0, 1]$.

We need to carefully choose A, B such that A is close to uniform, i.e., $\mathbf{E}[|A - \frac{1}{n}|] \ll \frac{1}{n}$ and B is close to half to be 0 and half to be $\frac{2}{n}$.

This can be achieved by closeness in earthmover distance (d_{EM}) sense.

Claim: We want

$$A : d_{EM}(A, \delta_{\frac{1}{n}}) \ll \frac{1}{n}$$
$$B : d_{EM}(B, \frac{1}{2}\delta_0 + \frac{1}{2}\delta_{\frac{2}{n}}) \ll \frac{1}{n} ,$$

where $d_{EM}(\mu, \nu) := \inf_{X\sim\mu, Y\sim\nu} \mathbf{E}[|X - Y|]$.

The benefit to get the above distribution is that it has large support so it can match many moments.

Also without loss of generality, we can assume A, B are supported on $[0, \frac{\log^2 n}{n}]$. The reason is that $\alpha \geq \frac{\log n}{n}$, then we can learn large weights with $\frac{n}{\log n}$ samples, so it becomes easy to do the distinction.

We can find distributions A, B satisfying the above such that

$$\mathbf{E}[A^k] = \mathbf{E}[B^k], 0 \leq k \leq \log n .$$

After this has been accomplished, we can use information theory to show the desired lower bound (details are omitted in class).