CS880: Advanced Learning Theory **Fall 2019**

Lecture 7: Robust ℓ_2-Estimator & ℓ_1-Tester with Unequal Sized Samples

Lecturer: Ilias Diakonikolas *Scribes: Rojin Rezvan*

1 Introduction

In this lecture, we give the optimal algorithm for tolerant ℓ_2-equivalence testing and present another application of the flattening technique for the problem of ℓ_1-equivalence testing with unequal sized samples. This kind of tester comes in handy when taking samples from one of the distributions is more expensive than the other.

For distributions p and q over $[n]$ with $\max\{\|p\|_2, \|q\|_2\} \leq b$, our ℓ_2-tester distinguishes the case that $\|p-q\|_2 \leq \epsilon$ from the case that $\|p-q\|_2 \geq 2\epsilon$, with probability at least $\frac{3}{4}$ with $O\Big(\frac{\sqrt{b}}{\epsilon^2}\Big)$ samples.

Note that tolerant testing under the ℓ_2-norm has no dependence on the domain size. On the other hand, as will see later in the course, tolerant testing under the ℓ_1-norm requires an almost linear number of samples. Roughly speaking, the reason for the latter is that the ℓ_1-norm cannot be approximated by a low-degree polynomial in the parameters.

2 ℓ_2-Distance Estimator

In this section, we will show the following:

Theorem 1. *Let $p, q : [n] \to [0,1]$ be two unknown distributions to which we have sample access. If $b = \max\{\|p\|_2, \|q\|_2\}$, we can distinguish between $\|p-q\|_2 \leq \epsilon$ versus $\|p-q\|_2 \geq 2\epsilon$ with $O(\frac{\sqrt{b}}{\epsilon^2})$ samples.*

Proof. We build an estimator for $\|p-q|_2$. Let m be the sample size. Draw $Poi(m)$ samples independently from p and q. Let X_i and Y_i be the number of samples from p and q respectively. Now we define:

$$Z = \sum_{i=1}^{n} (X_i - Y_i)^2 - X_i - Y_i \ .$$

Note that taking Poisson samples will not affect the complexity of sample size because the Poisson distribution is concentrated around its mean. If $Z_i = (X_i - Y_i)^2 - X_i - Y_i$, then we have $Z = \sum_{i=1}^{n} Z_i$. We have:

$$\mathbf{E}[Z_i] = \mathbf{E}\left[X_i^2\right] + \mathbf{E}\left[Y_i^2\right] - 2\mathbf{E}[X_i]\,\mathbf{E}[Y_i] - \mathbf{E}[X_i] - \mathbf{E}[Y_i] \tag{1}$$

As $X_i \sim Poi(mp_i)$ and $Y_i \sim Poi(mq_i)$, we have:

$$\mathbf{E}[X_i] = mp_i, \mathbf{E}[Y_i] = mq_i, \quad \mathbf{E}\left[X_i^2\right] = \mathbf{Var}[X_i] + \mathbf{E}[X_i]^2 = mp_i + (mp_i)^2$$

So we get:

$$\begin{aligned} \mathbf{E}[Z_i] &= mp_i(1 + mp_i) + mq_i(1 + mq_i) - 2mp_imq_i - mp_i - mq_i \\ &= m^2p_i^2 + m^2q_i^2 - 2m^2p_iq_i \\ &= m^2(p_i - q_i)^2 \\ \mathbf{E}[Z] &= \sum_{i=1}^{n} \mathbf{E}[Z_i] \\ &= m^2\|p - q\|_2 \end{aligned} \tag{2}$$

With some computations, we can get the value of $\mathbf{Var}[Z]$:

$$\mathbf{Var}[Z] = \sum_{i=1}^{n} 4m^3(p_i - q_i)^2(p_i + q_i) + 2m^2(p_i + q_i)^2$$

Now by Cauchy-Shwarz inequality, since $\sum_{i=1}^{n}(p_i + q_i)^2 \leq 4b$, we have:

$$\sum_{i=1}^{n}(p_i - q_i)^2(p_i + q_i) \leq \sqrt{\sum_{i=1}^{n}(p_i - q_i)^4 \sum_{i=1}^{n}(p_i + q_i)^2} \leq 2\|p - q\|_4^2\sqrt{b}$$

Thus:

$$\mathbf{Var}[Z] \leq 8m^3\sqrt{b}\|p - q\|_4^2 + 8m^2b$$

Lemma 2. *If $m \geq x + y$, then $m^2 \geq mx + y^2$, for any $x, y \geq 0$.*

Proof. We have $m \geq x + y$ so $(m - x)^2 \geq y^2$. Moreover, as $m \geq x + y$ and $y \geq 0$ then $m \geq x$ and $mx \geq x^2$. So $m^2 \geq y^2 - x^2 + 2mx$ and so $m^2 \geq y^2 + mx$. □

Now by Chebyshev's inequality, the returned estimate of $\|p - q\|_2$ will be accurate to within $\pm\epsilon$ with probability at least $\frac{3}{4}$ provided

$$\epsilon^2m^2 \geq 2\sqrt{8m^3\sqrt{b}\|p - q\|_4^2 + 8m^2b}$$

Using (2), this holds whenever

$$m \geq 6\frac{\sqrt{b}}{\epsilon^2} + 32\frac{\sqrt{b}\|p-q\|_4^2}{\epsilon^4}$$

□

Remark 1. Note that the statistic $\sum_{i=1}^n (X_i - Y_i)^2$ is close to the number of pairwise collisions. In fact, we get useful information only from elements which have been seen at least twice in the sample. In defining Z, we subtracted X_i and Y_i from each term. The reason for this was two-fold. First, if we didn't make this alteration, we wouldn't get an unbiased estimator for $\|p-q\|_2$. More importantly, the variance of that estimator would have been larger than what we would like. In particular, if we had not subtracted these terms, we would see m^4 in the variance instead of m^3, which would affect the sample complexity of our tester.

Remark 2. The first sample-optimal tester for ℓ_1-equivalence testing was given and analyzed in [1], based on the statistic:

$$\sum_{i=1}^n \frac{(X_i - Y_i)^2 - X_i - Y_i}{X_i + Y_i} \, .$$

Analyzing such a tester is somewhat more cumbersome, as it is a rational function of the X_i, Y_i.

3 ℓ_1-Equivalence Testing with Unequal Sized Samples

The motivation for this testing problem is that in many situations gathering samples from one of the distributions we are dealing with is more expensive. The most extreme case is when we know one of the two distributions exactly.

```
Algorithm   Testing Closeness with Unequal Sample Size
Input:      Sample access to distributions $p$ and $q$ supported on $[n]$ and $\epsilon > 0$
Output:     "YES" with probability at least $\frac{2}{3}$ if $p = q$
            "NO" with probability at least $\frac{2}{3}$ if $\|p - q\|_1 \geq \epsilon$.

            1- Let $k = \min(n, m_1)$.
            2- Define a multiset $S$ by taking $Poi(k)$ samples from $q$.
            3- Use the $\ell_1$-tester from last lecture on $p_S, q_S$ to distinguish between $p_S = q_S$
            versus $\|p_S - q_S\|_1 \geq \epsilon$ using at most $O\Big(\frac{bn}{\epsilon^2}\Big)$ samples.
```

Theorem 3. *Let $p, q : [n] \to [0, 1]$ be two unknown distributions. We are given $m_1 + m_2$ samples from q and m_2 samples from p. One can distinguish between $p = q$ versus $\|p - q\|_1 \geq \epsilon$ with probability greater than $\frac{9}{10}$ with $m_2 = O\Big(\max\{\frac{\sqrt{n}}{\epsilon^2}, \frac{n}{\epsilon^2\sqrt{m_1}}\}\Big)$ using Algorithm 3.*

Proof. This algorithm uses the flattening technique [2]. The main idea is to take samples from the less expensive distribution and use them for flattening both distributions. Here we have assumed q is the less expensive distribution as we have more samples from it.

First, note that with high probability, say $\frac{19}{20}$, we have $|S| = O(n)$. Therefore, the new domain size would be $n + |S| = O(n)$. Furthermore, by a lemma from the previous lecture, it follows that $\mathbf{E}[\|q_S\|_2] \leq O\Big(\frac{1}{\sqrt{k}}\Big)$. Also note that we have $\|p - q\|_1 = \|p_S - q_S\|_1$.

To bound the sample complexity, we have

$$m_2 = O(\frac{n}{\sqrt{m_1}\epsilon^2} + \frac{\sqrt{n}}{\epsilon^2}) \ .$$

This gives us the sample complexity desired, as we consider maximum of the sum we are dealing with order. □

Remark 3. *As m_1 increases, m_2 decreases, which is consistent with out intuition. Furthermore, if we set $m_1 = m_2$, meaning we take twice as samples from q than we do from p, then the sample complexity would be*

$$\max\{\frac{n^{\frac{2}{3}}}{\epsilon^{\frac{4}{3}}}, \frac{n^{\frac{1}{2}}}{\epsilon^2}\} \ ,$$

i.e., the sample complexity of vanilla ℓ_1-equivalence testing.

References

[1] Siu-On Chan, Ilias Diakonikolas, Gregory Valiant, and Paul Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the Twenty-fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '14, pages 1193–1203, Philadelphia, PA, USA, 2014. Society for Industrial and Applied Mathematics.

[2] Ilias Diakonikolas and Daniel M. Kane. A new approach for testing properties of discrete distributions, 2016.