CSE 291G: Computational Statistics Lecturer: Daniel Kane
Winter 2020, January 16 Scribes: Aaron Geelon So, Zhi Wang

Lecture 4: Closeness Testing and Independence Testing

Last time, we discussed the L^2-tester, where given distributions p and q over $[n]$, with $\|q\|_2 \le b$, we can distinguish between (i) $p = q$ and (ii) $\|p - q\|_2 > \varepsilon$ with sample complexity:

$$O\left(\frac{b}{\varepsilon^2} + \frac{1}{\varepsilon}\right) \text{ samples.} \tag{1}$$

By converting an L^1-separation to an L^2-separation, we also saw how we could use the L^2-tester to perform identity testing with L^1-separation of ε in $O(\sqrt{n}/\varepsilon^2)$ samples.

In identity testing, we want to determine whether p is the same as a known distribution q. Today, we will consider a similar problem of *closeness testing*, where q is not known *a priori*. We'll also show how we can use closeness testing to perform *independence testing*, to determine whether a distribution is an independent (product) distribution. Let's quickly recap the technique we saw last time.

1 Flattening technique

Because the number of samples depends on the bound b, we would like to control the L^2-norm of q. This leads to a *flattening* technique, where we transform q into a flatter distribution q' with smaller L^2-norm. In particular, given $q = (q_1, \ldots, q_n)$, we can distribute the probability mass of each bin into a number of sub-bins so that the overall distribution becomes more uniform.

To do this, we split the ith bin into $\lceil q_i \cdot n \rceil$ bins. Then, having expanded the domain from $[n]$ to $[n']$, where $n' = \sum_{i=1}^n \lceil q_i \cdot n \rceil$, we can also transform any distribution p into a new distribution p' on $[n']$ as follows: draw an element from p and assign it to one of its sub-bins uniformly at random. Here were a few important properties of this transformation:

- it preserves L^1-distance, $\|p - q\|_1 = \|p' - q'\|_1$
- we can simulate draws from p' and q' by using draws from p and q
- the domain expands by at most by a constant factor, $n' \le 2n$.

Because q' is flatter, we obtain a better upper bound b, and because this transformation only expands the domain by a constant factor, when we convert the ε-separation in L^1 into an $(\varepsilon/\sqrt{n})$-separation in L^2, the new $(\varepsilon/\sqrt{n'})$-separation is only a constant factor smaller as well.

2 Closeness testing

In *closeness testing*, we aim to distinguish between (i) $p = q$ and (ii) $\|p-q\|_1 > \varepsilon$, where p and q are arbitrary unknown distributions. Because q is unknown, we don't know how to flatten q as we did for identity testing. Instead, we can use samples from q to approximately determine q_i for the heavy bins, before applying the flattening technique again.

Here's a quick sketch of our approach. Draw k i.i.d. samples from q. Let S be this multiset of samples and define $a_i = \{x \in S : x = i\}$ to be the number of samples landing in the ith bin. We then split the ith bin into $(a_i + 1)$ sub-bins. If we define the transformation of p and q into p_S and q_S as above, then we have the following properties:

- it preserves L^1-distance, $\|p - q\|_1 = \|p_S - q_S\|_1$
- we can simulate draws from p_S and q_S by using draws from p and q
- the domain expands to $n' = n + |S|$.

As our goal of this flattening transformation is to reduce the bound of the L^2-norm of q_S, we compute:

$$\|q_S\|_2^2 = \sum_{i=1}^{n} (a_i + 1) \left(\frac{q_i}{a_i + 1} \right)^2 = \sum_{i=1}^{n} \frac{q_i^2}{a_i + 1}. \tag{2}$$

But $\|q_S\|_2$ is random; we would like an upper bound that holds with high probability. To do this, let's bound the expectation and appeal to Markov's inequality (since $\|q_S\|_2$ is a non-negative random variable).

If S is drawn i.i.d. from q of size k, we claim that in expectation, $\mathbb{E}\left[\|q_S\|_2^2\right] \leq 1/(k+1)$. By Markov,

$$\Pr\left[\|q_S\|_2 \geq \sqrt{\frac{100}{k}}\right] = \Pr\left[\|q_S\|_2^2 \geq \frac{100}{k}\right] \leq \frac{k}{100(k+1)}.$$

In particular, this shows that with 99% probability,

$$\|q_S\|_2 = O\left(\frac{1}{\sqrt{k}}\right). \tag{3}$$

The following is just the technical proof of the claim:

Claim 1. *If S is drawn i.i.d. from q with $|S| = k$, then in expectation, the L^2-norm of q_S is bounded:*

$$\mathbb{E}\left[\|q_S\|_2^2\right] \leq \frac{1}{k+1}.$$

Proof. Assuming i.i.d. draws from q, we take expectations on both sides of Equation 2,

$$\mathbb{E}\left[\|q_S\|_2^2\right] = \sum_{i=1}^{n} q_i^2 \cdot \mathbb{E}\left[\frac{1}{a_i + 1}\right].$$

We can compute this, since $a_i \sim \text{Bin}(q_i, k)$.

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{a_i+1}\right] = \sum_{j=0}^{k} \Pr(a_i = j) \cdot \frac{1}{j+1} &= \int_0^1 \sum_{j=0}^{k} \Pr(a_i = j) \cdot x^j \, dx \\
&= \int_0^1 \sum_{j=0}^{k} \binom{k}{j} q_i^j (1 - q_i)^{k-j} \cdot x^j \, dx \\
&= \int_0^1 \big(q_i x + (1 - q_i)\big)^k \, dx = \frac{1}{k+1} \frac{1}{q_i} \big(1 + q_i(x-1)\big)^{k+1} \Big|_0^1 \leq \frac{1}{q_i} \frac{1}{k+1}.
\end{aligned}$$

Then, when we sum over $i \in [n]$, because q is a distribution, we have:

$$\mathbb{E}\big[\|q_S\|_2^2\big] \leq \sum_{i=1}^{n} \frac{q_i}{k+1} = \frac{1}{k+1}.$$

□

2.1 Sample complexity for closeness testing

Consider the sample complexity bound for L^2-tester given by Equation 1. Suppose that we have a total budget of k samples to discover the heavy bins. From Equation 3, we can let the L^2-bound $\beta = O(1/\sqrt{k})$. Also, given an L^1-separation, we can obtain an L^2-separation lower bound,

$$\|p_S - q_S\|_1 > \varepsilon \quad \Rightarrow \quad \|p_S - q_S\|_2 > \varepsilon/\sqrt{n+k}.$$

Combining these two bounds, we obtain an overall sample complexity bound:

$$O\left(\frac{n+k}{\varepsilon^2\sqrt{k}} + \frac{\sqrt{n+k}}{\varepsilon}\right) \text{ samples.}$$

It follows that any gains we obtain by using more than n samples to flatten the L^2-norm is overtaken by the losses from the increased domain size. In particular, let us use $\min(k, n)$ samples in the flattening step. Thus, we can assume that $\|q_S\|_2 = O\left(1/\sqrt{\min(k,n)}\right)$. There are two domains:

- $k \geq n$. Then, $\|q_S\|_2 = O\left(1/\sqrt{n}\right)$ leads to a sample complexity $O\left(\sqrt{n}/\varepsilon^2\right)$.
- $k \leq n$. Then, $\|q_S\|_2 = O(1/\sqrt{k})$ leads to a sample complexity $O(n/\varepsilon^2\sqrt{k})$. Using a constant fraction of our sample budget for the flattening step, we lower bound $k = \Omega(n/\varepsilon^2\sqrt{k})$, which implies that $\sqrt{k} = \Omega(n^{1/3}/\varepsilon^{2/3})$. Replacing $\sqrt{k}$, we get a sample complexity $O(n^{2/3}/\varepsilon^{4/3})$.

Combining these, we obtain an overall sample complexity of:

$$O\left(\frac{\sqrt{n}}{\varepsilon^2} + \frac{n^{2/3}}{\varepsilon^{4/3}}\right) \text{ samples.} \tag{4}$$

3 Independence testing

In *independence testing*, we'll consider a distribution p over the product space $[n] \times [m]$, where $n \geq m$. We'd like to distinguish between two cases: (i) p is a product distribution (i.e. the coordinates of p are indendent) and (ii) p is ε-far in total variation from any product distribution.

We can define the product distribution $q = p_1 \otimes p_2$ where p_1 and p_2 are the marginal distributions of p. Samples from q can be generated by taking two independent samples $(x_1, y_1), (x_2, y_2) \sim p$ and returning (x_1, y_2). Notice that if p is indeed a product distribution, then $p = q$. But if p is ε-far from any product distribution, then $d_{TV}(p, q) > \varepsilon$. It follows that we can perform independence testing on p by applying closeness testing to p and q. Of course, a direct application of Equation 4 gives an upper bound:

$$O\left(\frac{\sqrt{nm}}{\varepsilon^2} + \frac{(nm)^{2/3}}{\varepsilon^{4/3}}\right) \text{ samples.}$$

But, perhaps we can do better, because we know that q is a product distribution; instead of flattening q directly, we can flatten the marginals p_1 and p_2 first.

In order to do this, suppose we have a budget k to flatten p_1 and p_2. Let's say we use $\min(k_1, n)$ and $\min(k_2, m)$ i.i.d. samples S_1 and S_2 drawn from p_1 and p_2, respectively. Letting $S = S_1 \sqcup S_2$, we define $q_S = (p_1)_{S_1} \otimes (p_2)_{S_2}$. The L^2-norm of a product distribution $\mu \otimes \nu$ is the product of the norms $\|\mu\|\|\nu\|$:

$$\|\mu\nu^\top\|_2^2 = \operatorname{tr}\big((\mu\nu^\top)^\top(\mu\nu^\top)\big) = \operatorname{tr}\big(\nu\mu^\top\mu\nu^\top\big) = \operatorname{tr}\big(\mu^\top\mu\nu^\top\nu\big) = \|\mu\|_2^2\|\nu\|_2^2.$$

So, if follows that we have an upper bound on the L^2-norm of q_S,

$$\|q_S\|_2 = \|p_{1,S_1}\|_2 \times \|p_{2,S_2}\|_2 = O\left(\frac{1}{\sqrt{k_1k_2}}\right).$$

Like before, we should only ever use at most $k_1 = O(n)$ and $k_2 = O(m)$ samples for flattening. If our budget k is sufficiently large, then we may use $k_1 = n$ and $k_2 = m$ samples to flatten, leading to a sample complexity of $O(\sqrt{nm}/\varepsilon^2)$. But if $k \leq n + m$, then we need on the order of $nm/\varepsilon^2\sqrt{k_1 k_2}$ many samples. Again, we could take a constant fraction of samples for the flattening step, so

$$k = \Omega\left(\frac{nm}{\varepsilon^2\sqrt{k_1 k_2}}\right) \geq \Omega\left(\frac{\sqrt{nm}}{\varepsilon^2}\right) \geq \Omega(m),$$

because $n \geq m$. It follows that we have sufficiently many samples to let $k_2 = \Theta(m)$. This means that the overall sample complexity to independence testing is:

$$O\left(\frac{\sqrt{nm}}{\varepsilon^2} + \frac{n^{2/3}m^{1/3}}{\varepsilon^{4/3}}\right) \text{ samples.} \tag{5}$$

In this way, we saved a factor of $m^{1/3}$ in the low-budget domain $k \leq n + m$, and we will see in later weeks that this is in fact optimal.

4 Lower bounds

We now present some lower bound techniques for uniformity testing and closeness testing. We start by providing some intuition and we follow with more rigorous proofs. The lower bounds show that the testers are information-theoretic optimal up to a constant factor.

4.1 Intuition

4.1.1 Uniformity testing

Recall that in uniformity testing, given a distribution p over $[n]$, we aim to distinguish between (i) $p = U_n$ and (ii) $\|p - U_n\|_1 > \varepsilon$. We saw an upper bound of $O(\sqrt{n}/\varepsilon^2)$ samples. Suppose that during testing, we make k i.i.d. draws from p,

$$X_1, X_2, \ldots, X_k \overset{\text{i.i.d.}}{\sim} p.$$

We say that there is a pairwise collision if an element of $[n]$ is sampled twice (i.e. there exists $i \neq j$ such that $X_i = X_j$). Likewise, a 3-wise collision occurs if an element is sampled thrice.

Let's compute the expected number of pairwise collisions:

$$\mathbb{E}\left[\sum_{i<j} \mathbf{1}\{X_i = X_j\}\right] = \sum_{i<j}\sum_{s\in[n]} \Pr[X_i = s]\Pr[X_j = s] = \binom{k}{2}\sum_{s\in[n]} p_s^2 = \binom{k}{2}\|p\|_2^2.$$

Therefore, when p is uniform, the expected number of collisions is around k^2/n. And more generally, the expected number of r-wise collisions for $r \in [k]$ is:

$$\mathbb{E}\left[\sum_{i_1<\cdots<i_r} \mathbf{1}\{X_{i_1} = \cdots = X_{i_r}\}\right] = \binom{k}{r}\|p\|_r^r.$$

In particular, the expected number of 3-wise collisions when p is uniform is k^3/n^2. Thus, by Markov, k needs to be roughly $n^{2/3}$ before we expect to see a single 3-wise collision with constant probability. Assuming that the number of samples required is much smaller than n, or even $n^{2/3}$, then most information about p will come from observing its pairwise collisions.

This means that to distinguish between (i) p is uniform and (ii) p is ε-far from uniform, we should compare the number of pairwise collisions. If p is ε-far from the uniform distribution, then $\|p\|_2^2 > (1 + \varepsilon^2)/n$, and it follows that the number of pairwise collisions is at least $k^2(1 + \varepsilon^2)/n$.

Since pairwise collisions are akin to a Poisson process, the standard deviations are the square roots of the means. Then, in the worst case, we need to distinguish between

$$\frac{k^2}{n} \pm \sqrt{\frac{k^2}{n}} \quad \text{and} \quad \frac{k^2(1+\varepsilon^2)}{n} \pm \sqrt{\frac{k^2(1+\varepsilon^2)}{n}}$$

many pairwise collisions. To tell apart the distributions, the difference in the means needs to be much greater than at least the standard deviations, so k must be sufficiently large to satisfy

$$\frac{k^2\varepsilon^2}{n} \gg \sqrt{\frac{k^2}{n}}.$$

It follows that we need $k \gg \sqrt{n}/\varepsilon^2$ many samples, matching the upper bound.

4.1.2 Closeness testing

In closeness testing, recall that we aim to distinguish between (i) $p = q$ and (ii) $\|p - q\|_1 > \varepsilon$, where p and q are any unknown distributions over $[n]$. If a test draws $k \ll n$ samples from p and q, then most of the signal to distinguish between (i) and (ii) will be based on counting pairwise collisions.

Given k draws from p and q, we can test whether the collisions occur mostly within distribution (draws from p colliding with draws from p, or q with q; this suggests that $p \neq q$) or between distributions (p colliding with q or q colliding with p; this suggests that $p = q$). In particular, let X_i and Y_i be the number of draws from bin i under p and q, respectively. Then, the statistic $X_i^2 + Y_i^2$ is approximately the number of within-distribution collisions in bin i and $X_iY_i + Y_iX_i$ the number of between-distribution collisions.

We saw in the previous lecture how this motivated the following statistic:

$$Z = \sum_{i\in[n]} (X_i^2 + Y_i^2) - (X_iY_i + Y_iX_i) = \sum_{i\in[n]} (X_i - Y_i)^2.$$

In fact, when $X_i \sim \text{Poi}(k \cdot p_i)$ and $Y_i \sim \text{Poi}(k \cdot q_i)$, then a slightly modified statistic has expectation:

$$\sum_{i\in[n]} E\big[(X_i - Y_i)^2 - X_i - Y_i\big] = k^2\|p - q\|_2^2.$$

It follows that using $\Theta(k)$ samples from p and q, we expect a signal on the order of $k^2\varepsilon^2/n$ to help us distinguish between the two cases (i) and (ii).

However, let's consider a hard case for p and q, assuming that $k \ll n$. Suppose that half of the probability mass falls uniformly into the first k bins for both p and q. That is, for all $i, j \in [k]$, we have:

$$p_i = q_j \approx \frac{1}{k}.$$

Then, in either case where the remaining half of the probability mass (i) matches up or (ii) diverges, these first k bins will add significant noise to our test statistic.

In particular, when we draw k times from p and q to perform closeness testing, half of those draws will fall into the first k bins. We already saw that the expected number of pairwise collisions from k draws coming out of a uniform distribution over k elements is around k^2/k. And as a result, we expect to see around $k \pm \sqrt{k}$ collisions in these first k bins occurring both within distribution and between distribution.

Although in expectation, these different types of collisions cancel out, the standard deviation remains on the order of $\sqrt{k}$, contributing to the noise of our test statistic. In this case, our statistic will be:

$$k^2\|p - q\|_2^2 \pm \Omega(\sqrt{k}).$$

Because we need to ensure that the mean of our test statistic for the two cases (i) and (ii) can dominate the noise, we need $k^2\varepsilon^2/n \gg \sqrt{k}$. It follows that the sample complexity is at least $k \gg n^{2/3}/\varepsilon^{4/3}$.

While we just gave intuition as to what distributions would be hard to distinguish under this strategy of comparing the types of pairwise collisions, it is important to ensure that the tester doesn't know that the hardness comes from the first k bins (otherwise, they could just ignore these collisions). Instead, these hard distributions must be randomly chosen from ensembles of distributions. For closeness testing, we'll consider the following adversarial setup:

(i) If $p = q$, then for each element $i \in [n]$:

- with probability $\frac{k}{n}$, let $p_i = q_i = \frac{1}{k}$.
- otherwise, let $p_i = q_i = \frac{\varepsilon}{n}$

(ii) If $\|p - q\| > \varepsilon$, then for each element $i \in [n]$:

- with probability $\frac{k}{n}$, let $p_i = q_i = \frac{1}{k}$.
- otherwise, let either $p_i = \frac{2\varepsilon}{n}$ and $q_i = 0$, or $p_i = 0$ and $q_i = \frac{2\varepsilon}{n}$.

Notice that in both cases, there is a $\frac{k}{n}$ probability for each element $i \in [n]$ that $p_i = q_i = \frac{1}{k}$. In expectation, around k of the bins will have a constant fraction of the probability mass, leading to the hardness coming from the noise on the order of $\Omega(\sqrt{k})$. One of the issue here is that p and q defined in this way may not be normalized probability distributions. We will deal with this when we prove the lower bound rigorously.

4.2 Proof sketch of uniformity testing lower bounds

To prove lower bounds for uniformity, we consider an adversary method. Let X be a random bit:

- if $X = 0$, let $p = U_n$
- if $X = 1$, take p from some ensemble $\mathcal{D}$ such that $\|p - U_n\|_1 > \varepsilon$ with high probability.

Our goal is to construct this adversary method in such a way that it is **information-theoretically** impossible for any algorithm to reliably determine the hidden bit X in a given number of samples, that is, to solve the uniformity testing problem.

Notice that it is important to have an adversarial ensemble $\mathcal{D}$. This is because it would only take $O(1/\varepsilon^2)$ samples to distinguish between (i) p is uniform and (ii) p is some fixed distribution q where $d_{TV}(q, U_n) > \varepsilon$. Given such a q, there exists a set $A \subset [n]$ such that $|q(A) - U_n(A)| > \varepsilon$, so we would only need to estimate $p(A)$ up to error $\varepsilon/2$, which uses $O(1/\varepsilon^2)$ samples.

Following this intuition that we want to construct p such that it is close to the uniform distribution, we consider an ensemble $\mathcal{D}$ where $p_i = \frac{1 \pm \varepsilon}{n}$. Let us first Poissonize:

- pick X and p
- take Poi(k) samples from p, and let A_i be the number of samples from bin i
- return $Y = f(A_1, \ldots, A_n) \in \{0, 1\}$ as the output of an algorithm guessing X

We want to show that the mutual information between X and the output of the samples, $A_1, \ldots, A_n$, is small. In particular, we want to show $I(X; A_1, \ldots, A_n) = o(1)$. If so, by the information processing inequality, we must also have $I(X;Y) = o(1)$. Since $I(X;Y) = H(X) - H(X|Y)$, and $H(X) = 1$, if the algorithm is almost always correct, then $H(X|Y)$ is small.[1]

The computation of the mutual information $I(X; A_1, \ldots, A_n)$ is complicated and we would like to avoid it. Therefore, we want the case where A_i's are conditionally independent on X, because if so, then we have,

$$I(X; A_1, \ldots, A_n) \leq \sum_{i=1}^{n} I(X; A_i)$$

and it is much easier to compute each individual $I(X; A_i)$.

Now the question is: Are A_i's conditionally independent on X?

- If $X = 0$, yes. Each bin has probability $1/n$, and because of the Poissonization, the A_i's are independently distributed as Poisson random variables $A_i \sim \text{Poi}(kp_i)$.
- if $X = 1$, no. This is because p_i's are no longer independent: p is a probability distribution, and p_i's need to add up to 1.

The solution is to use non-normalized distribution p, that is, a list of probabilities assigned to each domain element. To sample from p, we have $A_i \sim \text{Poi}(kp_i)$ independently, and we can now take $p_i = \frac{1 \pm \varepsilon}{n}$ independently of each other.

It will take a bit of work to relate this new problem back to the original testing problem. But, at least we now have the desired conditional independence. We will resume in the next lecture.

[1] If we can find an algorithm that has some nontrivial predictive power about what X is, then $I(X;Y) = \Omega(1)$, and this can lead to a proof by contradiction.