Lecture 9: Tolerant Testing

Daniel Kane
Scribe: Sankeerth Rao

April 24, 2017

Abstract

In this lecture we prove a quasi linear lower bound on the number of samples needed to do tolerant testing for L^1 distance.

1 Tolerant Testing

We have been assuming either

- $p = q$, OR
- p far from q.

What happens if we relax the first assumption? In most natural settings of the problem the distributions need not be exactly equal but are close to each other. Our L^2 tester where we used an unbiased estimator for $|p-q|_2$, already had this property.

L^2 **Tester:** Distinguishes $|p-q|_2 \leq \epsilon/2$ vs $|p-q|_2 \geq \epsilon$ in $O(|q|_2/\epsilon^2)$ samples.

Let us see what happens in the case of $|\cdot|_1$ distance. Lets start with our L^1 identity tester and see what kind of tolerance we achieve from it ?

L^1 **Identity Tester:** We split the ith bin into $\lceil nq_i \rceil$ equal pieces and distinguish $p_S = q_S$ vs $|p_S - q_S|_2 > \frac{\epsilon}{\sqrt{n}}$ based on our L^2 tester. Since the L^2 tester is already tolerant the same analysis will distinguish $|p_S - q_S|_2 < \frac{\epsilon}{2\sqrt{n}}$ vs $|p_S - q_S|_2 > \frac{\epsilon}{\sqrt{n}}$. let us see what is $|p_S - q_S|_2$ explicitly.

$$\begin{aligned} |p_S - q_S|_2^2 &= \sum_i \lceil nq_i \rceil (\frac{p_i - q_i}{\lceil nq_i \rceil})^2. \\ &= \sum_i \frac{(p_i - q_i)^2}{\lceil nq_i \rceil} \leq \frac{1}{n} \sum_i \frac{(p_i - q_i)^2}{q_i} \stackrel{def}{=} \frac{\chi^2(p,q)}{n}. \end{aligned}$$

Thus we are able to distinguish $|p-q|_1 > \epsilon$ vs $\chi(p,q) < \frac{\epsilon}{2}$. In fact we show that $\chi(p,q)$ upper bounds $|p-q|_1$ in the following lemma.

Lemma 1.1.

$$|p-q|_1 \leq \chi(p,q).$$

Proof. We use Cauchy Schwarz to prove this.

$$|p-q|_1^2 = \Big[\sum_i \frac{|p_i-q_i|}{\sqrt{q_i}}\cdot\sqrt{q_i}\Big]^2 \leq \Big[\sum_i \Big(\frac{|p_i-q_i|}{\sqrt{q_i}}\Big)^2\Big]\cdot\Big[\sum_i(\sqrt{q_i})^2\Big] = \sum_i \frac{(p_i-q_i)^2}{q_i} = \chi^2(p,q).$$

□

We have a tester that distinguishes $|p-q|_1 > \epsilon$ vs $\chi(p,q) < \frac{\epsilon}{2}$ but this tester doesn't distinguish between $|p-q|_1 > \epsilon$ vs $|p-q|_1 < \frac{\epsilon}{2}$, because $\chi(p,q)$ is a very weak upper bound on $|p-q|_1$. For instance there could be bins for which q_i is very small and $|p_i-q_i|$ is small - this has a small contribution to $|p-q|_1$ but a huge contribution to $\chi(p,q)$. So when $|p-q|_1 = \frac{\epsilon}{2}$, $\chi(p,q)$ could be very large and thus our tester will fail to distinguish between $|p-q|_1 > \epsilon$ vs $|p-q|_1 < \frac{\epsilon}{2}$.

So we want a tester that distinguishes $|p-q|_1 > \epsilon$ vs $|p-q|_1 < \frac{\epsilon}{2}$. Unfortunately we will show that this is very hard even in the regime of constant ϵ.

Lets focus on the case when $q = U_n$. This kind of tester wants us to approximate $|p-q|_1 = \sum_i |p_i - \frac{1}{n}|$.

Intuition for why L^1 testing is hard: What can a tester do ? The order of the samples doesn't matter by the symmetry argument. The only useful information the tester gets is the counts of the bins. Now if we look at how many bins have k samples. The expected number is given by $\sum_i e^{-mp_i}\frac{(mp_i)^k}{k!}$. For small values of k these give the moments of p upto the exponentials e^{-mp_i}.

So what is really easy to do is to approximate the moments of p or low degree polynomials of p.

$$\sum_i f(p_i) \ f \textit{ low degree polynomial.}$$

For instance if we are doing L^2 testing we need to estimate $\sum_i \left(p_i - \frac{1}{n}\right)^2$ which is exactly in the above form and thus is easy to estimate. However this is not the case with $|\cdot|_1$ norm because $|x - \frac{1}{n}|$ is not a polynomial. Thus we can't expect estimating $|\cdot|_1$ to be easy. We could always learn the distribution in $\frac{n}{\epsilon^2}$ samples and use that to get an ϵ approximation to $|p-q|_1$. However that requires a linear number of samples.

In fact we will see that it really does take $\frac{n}{\log n}$ samples when ϵ is a constant.

Research Problem: Find out the correct dependence of the number of samples on n and ϵ in the ϵ non constant regime.

Theorem 1.2. *p is a distribution on $[n]$. Any tester that distinguishes between*

$$|p - U_n|_1 < \frac{1}{10} \text{ vs } |p - U_{\frac{n}{2}}|_1 < \frac{1}{10}$$

requires $\Omega(\frac{n}{\log n})$ samples. $U_{\frac{n}{2}}$ is the uniform distribution over some subset of $\frac{n}{2}$ bins.

Note Any good tolerant tester should be able to do this because U_n and $U_{\frac{n}{2}}$ are very far apart. This was also used to show that estimating entropy is hard because $H(U_{\frac{n}{2}}) << H(U_n)$. This shows that testing things that aren't polynomials is going to be difficult.

Proof. We use the adversary method to prove this: Take X a random bit. Once X is fixed each p_i is picked independently according to some distribution.

$$X = 0 \quad p_i \overset{iid}{\sim} A = \text{ some distibution on } [0,1].$$
$$X = 1 \quad p_i \overset{iid}{\sim} B.$$

We should pick A and B carefully. Lets figure out what does our adversary need to do with A, B. When $X = 0$ we want $|p - U_n|_1$ to be small. Thus we need $\mathbb{E}[|A - \frac{1}{n}|] << \frac{1}{n}$. This ensures that in each bin we don't add too much mass to $|p - U_n|_1$, so that when we sum over all the bins we have $|p - U_n|_1 < \frac{1}{10}$. When $X = 0$ we want B to pick 0 half the time and be close to $\frac{1}{n}$ half the time. In order to quantify this for B, we use the Earth mover metric.

Definition 1.3.

$$d_{EM}(r, s) = \inf_{\substack{X \sim r \\ Y \sim s}} \mathbb{E}[|X - Y|].$$

Intuition: Since the marginals of X and Y are fixed we are taking the infimum over all possible correlations that X and Y could have. Say when $r = s$ then we can precisely match X and Y, that is if X takes the real t with probability r_t then we correlate Y to take t with $s_t = r_t$ and thus get $d_{EM} = 0$. Now lets consider the case when $s = \delta_0$. Then the infimum would be $\mathbb{E}_r|X| = \sum_t |t| r_t$. Note that $|t|r_t$ says that we need to move the mass r_t through a distance of $|t|$ to move the probability mass r_t to 0. So in general it captures the least amount of distance times the probability mass that is being moved so that we can make r, s to be the same distributions.

So we want A is close to $\delta_{\frac{1}{n}}$ in EMM, $d_{EM}(A, \delta_{1/n}) << \frac{1}{n}$ and $d_{EM}(B, \frac{1}{2}\delta_0 + \frac{1}{2}\delta_{\frac{2}{n}}) << \frac{1}{n}$. These choices ensure that $X = 0 \implies |p - U_n|_1 < \frac{1}{10}$ and $X = 1 \implies |p - U_{\frac{n}{2}}| < \frac{1}{10}$.

Now we need to figure out what other constraints need to be put on A and B so that we can actually fool the tester.

We need to put the following two constraints for this:

(i) If p_i is large then since the number of samples that fall in bin i are distributed$\sim Poi(mp_i)$, there will be more samples from bin i and thus p_i can be approximated well.

To be precise lets see how far off the empirical estimate of p_i would be. Let X_i denote the number of samples from bin i. Then

$$\Pr\left(\left|\frac{X_i}{m} - p_i\right| > c\sqrt{p_i/m}\right) \leq \frac{p_i/m}{c^2 p_i/m}$$

The tester would want to estimate with a constant probability and thus corresponds to some constant c. Thus the tester would see an error of $\sqrt{\frac{p_i}{m}}$ between its empirical estimate of p_i and actual p_i. Now say there are a bunch of bins whose probabilities p_is are all close to α. There can be at most $\frac{1}{\alpha}$ such bins. Now the total L^1 error the tester incurs by choosing the empirical estimates for p_i's for these bins (also ensuring that the tester wants to be sure with a constant probability) is $E = \frac{1}{\alpha}\sqrt{\frac{\alpha}{m}} = \frac{1}{\sqrt{\alpha m}}$. Now we know $m \approx \frac{n}{\log n}$(we are aiming for this lower bound), so if $\alpha >> \frac{\log n}{n}$ then $E \to 0$ and the tester would learn these p_i's. We don't want this to happen and thus we constrain the support of A and B to $[0, \frac{\log^2 n}{n}]$. The calculation here gives $\frac{\log n}{n}$, but for technical reasons we can allow coordinates with size as large as $\log^2(n)/n$, because you cannot reliably distinguish between these entries and the smaller ones with only $n/\log(n)$ samples.

(ii) We know that the best the tester could do is compute the moments of p_i. So we want the moments of A, B to be equal. Also since A, B are supported on $[0, \frac{\log^2 n}{n}]$ and since the number of samples in bin i is $\sim Poi(mp_i)$ which is tightly concentrated around $mp_i \leq \log n$, we wont see bins with more than $O(\log n)$ samples. Note that the expected number of bins with at least k samples is given by $\sum_i e^{-mp_i}\frac{(mp_i)^k}{k!}$.Hence the kth moment computation corresponds to looking at the number of bins with k samples. Thus the tester can compute the moments up to $O(\log n)$. Hence we only need to ensure that A and B agree on moments up to $O(\log n)$.

$$\mathbb{E}_{X\sim A}[X^k] = \mathbb{E}_{X\sim B}[X^k] \quad 1 \leq k \leq O(\log n).$$

In fact let $D = A - B$ be a pseudo distribution(need not be positive or normalized). It is supported on $[0, c_2\frac{\log^2 n}{n}]$ and satifies $|D|_1 \leq 2$ and $d_{EM}(D, \delta_{\frac{1}{n}} - \frac{1}{2}\delta_0 - \frac{1}{2}\delta_{\frac{2}{n}})$ is small. Its low order moments are 0,

$$\mathbb{E}_{X\sim D}[X^k] = 0 \quad 1 \leq k \leq c_3 \log n.$$

Our goal is to come up with such a D. Its positive and negative parts give the required A and B. We thus need D to have 0 expectation for any low degree polynomial. There is a clever way to do this. Its the following lemma.

Lemma 1.4. *p is a degree d polynomial with distinct real roots $r_1, \ldots, r_d$. Let $a_i = \frac{1}{p'(r_i)}$, then $\sum_i a_i r_i^k = 0$, for any integer $0 \leq k \leq d-2$.*

Note that the pseudo distribution $\sum_i a_i \delta_{r_i}$ has zero low degree moments. This will be the tool we will use to construct D.

Proof. We use polynomial interpolation to find the unique at most degree $d-1$ polynomial f that satisfies $f(r_i) = y_i \quad i \in [d]$. Then

$$f(x) = \sum_i y_i \prod_{j \neq i} \frac{(x - r_j)}{(r_i - r_j)}.$$

The coefficient of x^{d-1} in $f(x)$ is given by

$$\sum_i y_i \prod_{j \neq i} \frac{1}{r_i - r_j} = \sum_i y_i a_i.$$

Now if we want to prove that $\sum_i a_i r_i^k = 0$, we choose $y_i = r_i^k, \forall i \in [d]$. Note that the unique at most degree $d-1$ polynomial that satisfies $f(r_i) = r_i^k, \forall i \in [d]$ is just $f(x) = x^k$. Since $k \leq d-2$ note that the x^{d-1} coefficient of f is 0. Thus we have

$$\sum_i a_i r_i^k = 0, \quad 0 \leq k \leq d-2.$$

□

We get D by picking a polynomial g and using the above lemma's trick with the roots. What g do we need to pick ?

We need g to have a root at $0, \frac{1}{n}, \frac{2}{n}$ and another reasonable polynomial factor say T. We expect T to have roots in $[0, c_2 \frac{\log^2 n}{n}]$. Furthermore it needs to have nice derivatives which are not huge (Since this would ensure that A, B are close). One natural thing to try are the Chebyshev polynomials.

$$g(x) = x\Big(x - \frac{1}{n}\Big)\Big(x - \frac{2}{n}\Big) T_d\Big(1 - \frac{2x}{c_2(\frac{\log^2 n}{n})}\Big)$$

Chebyshev Polynomials The degree-d Chebyshev polynomial is the unique polynomial that satisfies $T_d(\cos\theta) = \cos(d\theta)$.

$$y \stackrel{def}{=} \Big(1 - \frac{2x}{c_2(\frac{\log^2 n}{n})}\Big) : [0, c_2 \frac{\log^2 n}{n}] \to [-1, 1].$$

The roots of $T_d(y)$ are $y = \cos\left(\frac{(2m+1)\pi}{2d}\right) \approx 1 - \frac{1}{2}(\frac{(2m+1)\pi}{2d})^2$, where $d = c_3 \log n$. So the first root is at $y = 1 - \frac{\pi^2}{8d^2}$. Let $c_4 = \frac{\pi^2 c_2}{16 c_3^2}$. So the corresponding x would be at approximately $c_4 \frac{1}{n}$, the other roots would be at $c_4 \frac{9}{n}, c_4 \frac{25}{n}, \ldots$... Note that we choose c_2, c_3 such that the first root of T_d as a function of x above is very far right from $\frac{2}{n}$. Thus the three artificially implanted

roots of g are well before the roots of T. In particular if we take c_3 very small then $T_d(y)$ will be $\geq 99/100$ on $[0, \frac{2}{n}]$. Note that,

$$g'(0) \approx \frac{2}{n^2}, \quad g'\Big(\frac{1}{n}\Big) \approx \frac{-1}{n^2}, \quad g'\Big(\frac{2}{n}\Big) \approx \frac{2}{n^2}.$$

Now we need to figure out g' at the roots of T_d. But observe

$$T_d'(\cos\theta) = \frac{d\sin(d\theta)}{\sin\theta}.$$

At the roots of T_d, $\cos(d\theta) = 0$, thus $\sin(d\theta) = \pm 1$, hence T_d' at the roots is $\frac{\pm d}{\sin\theta}$ where θ corresponds to the root. We need to compute g at these roots of T_d.

The roots of g are firstly the three artificial roots $0, \frac{1}{n}, \frac{2}{n}$ and then after a large gap followed by the roots $c_4\frac{1}{n}, c_4\frac{9}{n}, c_4\frac{25}{n}, \ldots$. Now say we evaluate $g'(c_4\frac{2m+1^2}{n})$ then we get

$$\begin{aligned} g'\Big(c_4\frac{(2m+1)^2}{n}\Big) &= \frac{c_4(2m+1)^2}{n}\frac{c_4(2m+1)^2-1}{n}\frac{c_4(2m+1)^2-2}{n}\frac{d}{\sin(\frac{(2m+1)\pi}{2d})}\frac{dy}{dx} \\ &\leq \frac{64c_4^3m^6}{n^3}\frac{2d^2}{(2m+1)\pi}\frac{n}{\log^2 n} \approx \frac{64c_4^3m^5c_3^2}{\pi c_2 n^2} = c_5\frac{m^5}{n^2} \end{aligned}$$

Note that these have the same dependence on n when compared to g' evaluated at the artificial roots but there is a large constant c_5 before which makes the pseudo distribution $d_{EM}(D, \delta_{\frac{1}{n}} - \frac{1}{2}\delta_0 - \frac{1}{2}\delta_{\frac{2}{n}})$ small. We get D from this polynomial we constructed using the lemma above. Since g' evaluated at the roots of g all have n^2 dependence we could scale D to get rid of n^2. Thus the pseudo distribution D we get satisfies

$$D(0) \approx \frac{1}{2} \quad D\Big(\frac{1}{n}\Big) \approx -1 \quad D\Big(\frac{2}{n}\Big) \approx \frac{1}{2}.$$

And at the other roots

$$D\Big(c_4\frac{(2m+1)^2}{n}\Big) \approx \frac{1}{c_5 m^5}.$$

Now lets evaluate

$$d_{EM}(D, \delta_{\frac{1}{n}} - \frac{1}{2}\delta_0 - \frac{1}{2}\delta_{\frac{2}{n}}) = \sum_m \frac{4c_4m^2}{n}\frac{1}{c_5m^5} \xrightarrow{c_4} 0$$

Note that $c_5 >> c_4^3$ so the above term goes to 0 with c_4. Thus we choose the constants such that $d_{EM}(D, \delta_{\frac{1}{n}} - \frac{1}{2}\delta_0 - \frac{1}{2}\delta_{\frac{2}{n}}) < \frac{1}{100n}$. Thus, we should get an average error of $1/(100n)$ per coordinate or a total error of at most $1/100$.

Conclusion: Thus we constructed a D supported on $[0, \frac{c_2\log^2 n}{n}]$ such that $\mathbb{E}_{X\sim D}[X^k] = 0$ for $k \leq d = c_3 \log n$. Next class we will use this D with Information theory arguments to complete the adversarial argument for the lower bound. □