Lecture 10: Tight bounds on Tolerant Testing

Daniel Kane
Scribe: Sankeerth Rao

April 25, 2017

Abstract

In this lecture we finish up the lower bound from the last lecture on Tolerant testing. We start proving a matching upper bound too.

1 Last Time

Tolerant Testing: We constructed a pseudo distribution D such that $d_{EM}(D, -\delta_{\frac{1}{n}} + \frac{1}{2}\delta_0 + \frac{1}{2}\delta_{\frac{2}{n}}) < \frac{1}{100n}$. Now looking at positive part and negative parts of D we get the distributions A, B supported in $[0, c_2\frac{\log^2 n}{n}]$. D has masses of $\approx \pm\frac{1}{c_5 l^5}$ at $\approx c_4\frac{(2l+1)^2}{n}, l \in \mathbb{N}$. The positive contributions to A are given by those l for which $\frac{\sin(\frac{(2l+1)\pi}{2})}{\sin(\frac{(2l+1)\pi}{2d})} > 0$. Lets call such l 'positive l'. Let the negative contribution to B are given by the other $l \in \mathbb{N}$.Lets call these negative l. Now a summation of the mass x distance to $\frac{1}{n}$ for the former l gives $d_{EM}(A, \delta_{\frac{1}{n}}) = o(\frac{1}{n})$, and a similar calculation for latter l gives $d_{EM}(B, \frac{1}{2}\delta_0 + \frac{1}{2}\delta_{\frac{2}{n}}) = o(\frac{1}{n})$. The low order moments of A, B are equal. So $\mathop{\mathbb{E}}_{X\sim A}[X^k] = \mathop{\mathbb{E}}_{X\sim B}[X^k]$ for $k < d = c_3 \log n$. Now we use the information theory arguments to finish the lower bound.

2 Lower Bound

We work with the adversarial method of proof. Let X be a uniform random bit.

$$X = 0 \quad p_i \overset{iid}{\sim} A$$
$$X = 1 \quad p_i \overset{iid}{\sim} B.$$

Let us first check that the distributions we generated are indeed close to $U_n, U_{\frac{n}{2}}$ accordingly.

X = 0 We compute the expected value of $|p - U_n|$ where each p_i is picked from probability distribution A.

$$\mathbb{E}[|p - U_n|] = \mathbb{E}\Big[\sum_i |p_i - \frac{1}{n}|\Big] = \sum_i o(\frac{1}{n}) = o(1).$$

Now we compute the variance.

$$Var(|p - U_n|) = \sum_i Var\Big(|p_i - \frac{1}{n}|\Big)$$

where

$$Var(|p_i - \frac{1}{n}|) \leq \mathbb{E}[(p_i - 1/n)^2] = O(\log^4(n)/n^2)$$

Summing this up we have

$$Var(|p - U_n|) \ll \frac{\log^4(n)}{n}$$

Thus

$$|p - U_n| < o(1) \quad whp.$$

X = 1 Let $C \sim \frac{1}{2}\delta_0 + \frac{1}{2}\delta_{\frac{2}{n}}$. Since $d_{EM}(B, \frac{1}{2}\delta_0 + \frac{1}{2}\delta_{\frac{2}{n}}) = o(\frac{1}{n})$, we can find correlated copies B, C such that $\mathbb{E}[|B - C|] = o(\frac{1}{n})$. Let $q_i \sim C$. That is

$$q_i = \begin{cases} 0 & \text{wp } \frac{1}{2} \\ \frac{2}{n} & \text{wp } \frac{1}{2} \end{cases}$$

Then $q_i = 0$ on $\frac{n}{2} \pm n^{\frac{2}{3}}$ bins whp. So $|q - U_{\frac{n}{2}}|_1 \ll n^{-\frac{1}{3}}$ whp.

Now since $\mathbb{E}[|B - C|] = o(\frac{1}{n})$ we have $\mathbb{E}[|p_i - q_i|] = o(\frac{1}{n})$. Summing over all indices we have $\mathbb{E}[|p - q|_1] = o(1)$. Computing the variance by looking at T^c we can show that $|p - q|_1 = o(1)$ whp. Thus $|p - U_{\frac{n}{2}}|_1 = o(1)$ whp.

Thus we ensured that the distributions we generated are indeed close to $U_n, U_{\frac{n}{2}}$ depending on $X = 0|1$. Now we proceed with the lower bound.

Information Theory Pick X and pick p. Take Poi(m) samples from p where $m = c_6 \frac{n}{\log n}$ where c_6 is very small relative to c_2, c_3. Say we get S_i samples from bin i, $i \in [n]$. We know $S_i \sim Poi(mp_i)$. We need to upper bound the amount of information the bin counts have about X. Since the bins counts are conditionally independent over X, we have

$$I(X; S_1 \dots S_n) \leq nI(X; S_1).$$

If $X = 0$ let α denote the distribution of the number of samples that fall in bin 1, in this case p_1 would have been picked from the distribution A. Similarily when $X = 1$ let β denote the distibution of the number of samples that fall in bin 1,in this case p_1 would have been picked from the distribution B. Then we have the following estimate like in Lecture 6.

$$I(X; S_1) = \sum_k \Theta\Big(\frac{(\alpha_k - \beta_k)^2}{\alpha_k + \beta_k}\Big) \leq \sum_k \Theta(|\alpha_k - \beta_k|) = \Theta(|\alpha - \beta|_1).$$

where

$$\alpha_k = \mathop{\mathbb{E}}_{X\sim A}[e^{-mX}\frac{(mX)^k}{k!}] = \mathbb{E}[e^{-mA}\frac{m^k}{k!}A^k]$$

where we are abusing notation and using A to denote the random variable whose distribution is A. Similarly

$$\beta_k = \mathop{\mathbb{E}}_{X\sim B}[e^{-mX}\frac{(mX)^k}{k!}] = \mathbb{E}[e^{-mB}\frac{m^k}{k!}B^k]$$

Now we know that the moments of A, B upto low orders are equal but we need to get rid of the exponentials in the above expression. So we Taylor expand the exponentials and truncate the higher moments on which A, B disagree and bound the error we incur. We have

$$e^{-mA} = 1 - mA + \frac{(mA)^2}{2} - \frac{(mA)^3}{6} + \cdots \pm \frac{(mA)^t}{t!} + \ \mathit{error}$$

The ratio of consecutive terms is $\frac{mA}{t}$. Since $m = c_6\frac{\log n}{n}$ and A is supported on $\left[0, c_2\frac{\log^2 n}{n}\right]$, we have

$$\frac{mA}{t} < c_6c_2\frac{\log n}{t}$$

If $t \gg 2^a c_6c_2 \log n$ then the tth term of above Taylor series can be bounded by $\frac{1}{2^{ta}}$. Thus when we multiply $e^{-mA}\frac{(mA)^k}{k!}$ we get a polynomial of degree $t+k$ and an error term.

$$e^{-mA}\frac{(mA)^k}{k!} = \frac{(mA)^k}{k!} - \frac{(mA)^{k+1}}{k!} + \frac{(mA)^{k+2}}{2\cdot k!} - \frac{(mA)^{k+3}}{6\cdot k!} + \cdots \pm \frac{(mA)^{t+k}}{t!k!} + 2^{-ta}\frac{(mA)^k}{k!}$$

Note that the error term $2^{-ta}\frac{(mA)^k}{k!} \leq 2^{-ta+O(c_6c_2\log n)}$. Now if we choose $t \geq \frac{4\log n}{a}$ this error would be bounded by n^{-3}. We choose a to be a very large multiple of $\frac{1}{c_3}$ and choose $t = \frac{d}{2} = \frac{c_3\log n}{2}$. Note that for this to satisfy the error needed in the Taylor expansion of e^{-mA}, we need to pick c_6 small enough. So essentially we proved that $e^{-mA}\frac{(mA)^k}{k!}$ is a bunch of low order terms plus an error of atmost n^{-3}. Now we use the fact that A, B agree of these low order moments and the errors are small enough.

Conclusion: We want to evaluate $\sum_k |\alpha_k - \beta_k|$. We break this evaluation into two regimes. Essentially for the regime $k < \frac{d}{2}$ the idea is to carefully bound $|\alpha_k - \beta_k|$ using the analysis above . For the regime of $k > \frac{d}{2}$ we show that α_k, β_k are themselves very small so add negligible mass. Note that in the latter regime the moments don't match anyway because each term's degree is at least $\frac{d}{2}$ and we were only assured moment matching upto d for A, B from the construction of D in the last lecture.

Regime $k < \frac{d}{2}$

$$|\alpha_k - \beta_k| = \overbrace{\mathbb{E}[poly(A) - poly(B)]}^{0} + O(n^{-3})$$

Regime $k > \frac{d}{2}$ We just bound α_k, β_k individually here.

$$\alpha_k \leq O(\frac{(mA)^k}{k!}) \leq O\Big(\frac{m^{\frac{c_2 \log^2 n}{n}}}{k}\Big)^k \leq (\frac{c_6 c_2}{c_3})^k = (\frac{c_6 c_2}{c_3})^{\frac{d}{2}+(k-\frac{d}{2})} \leq \frac{1}{n^2} 2^{-(k-\frac{d}{2})}$$

The last inequality holds because d is proportional to $\log(n)$ and because c_6 (and thus) $c_6 c_2 / c_3$ is taken to be sufficiently small.

Now we sum this over all $k > \frac{d}{2}$ to get

$$\Pr\Big(\alpha > \frac{d}{2}\Big) = O\Big(\frac{1}{n^2}\Big)$$

Adding over the two regimes we have,

$$|\alpha - \beta|_1 < O(\frac{1}{n^2})$$

Thus we have

$$I(X; S_1, \ldots, S_n) = O(\frac{1}{n}) \ll 1 = H(X)$$

Thus there are not enough samples to learn X. This completes the proof of the lower bound. We summarize it in the following theorem.

Theorem 2.1. *Any algorithm that given sample access to p on $[n]$ that reliably distinguishes $|p - U_n|_1 < \frac{1}{100}$ and $|p - U_{\frac{n}{2}}| < \frac{1}{100}$ requires $\Omega(\frac{n}{\log n})$ samples.*

3 Upper Bound

We show that the lower bound is indeed tight by the following algorithm.
Tolerant Uniformity Testing Algorithm $O(\frac{n}{\log n})$ samples.

Idea:

(i) Learn the heavy bins.

(ii) Estimate $\sum_{lightbins} |p_i - \frac{1}{n}|$

For the second step the idea is to approximate $f(x) = |x - \frac{1}{n}|$ by a low degree polynomial $g(x)$ on $][0, \frac{\log^2 n}{n}]$. This works because it is easy to evaluate polynomials of p_i. Essentially

$$\sum_i g(p_i) \approx \sum_k a_k \frac{\#\ bins\ with\ k\ samples}{n}$$

where the number of bins with k samples is just $\sum_i e^{-mp_i} \frac{(mp_i)^k}{k!}$. We need to get rid of the exponentials though.

Note that our algorithm distinguishes more strongly between $|p - U_n|_1 < \frac{1}{1000}$ vs $|p - U_n|_1 > \frac{1}{10}$.

Step 1 Take $m = \frac{cn}{\log n}$ samples and learn p on the bins that got samples. So we look at the sample and make a list of the bins that occured and coalesce all the bins that did not occur into one bin. We just learn p_i these bins that occurred by a standard learning algorithm that takes at most $O(m)$ samples.

Now this approach could fail if an actual high probability bin did not occur in these m samples. Lets show that this happens with a very low probability. Say $p_i > \frac{\log^2 n}{n}$. Then the number of samples from bin i is $\sim Poi(mp_i) = Poi(c \log n)$ which is pretty big. Thus with very high probability we would see one of these samples. In fact we could union bound over all such bins with high p_i's and say that all of these occur in the m samples we picked with very high probability. Thus we essentially learned all the heavy bins with a total error of say $\frac{1}{100}$.

In the next class we will show Step 2.