Lecture 11: Upper bound on Tolerant Testing

Daniel Kane
Scribe: Sankeerth Rao

April 27, 2017

Abstract

In this lecture we finish the upper bound on the number of samples needed to distinguish between $|p - U_n|_1 < \frac{1}{1000}$ vs $|p - U_n|_1 > \frac{1}{10}$.

1 Tolerant testing

Problem: Given the guarantee that either $|p - U_n|_1 < \frac{1}{1000}$ or $|p - U_n|_1 > \frac{1}{10}$, give an algorithm that distinguishes the two cases.

Idea: Take $\frac{cn}{\log n}$ samples and learn p on bins with samples to $\frac{1}{1000}$ error. Note that if a bin i has probability $p_i \geq \frac{\log^2 n}{n}$ then the number of samples that need fall into it is $X_i \sim Poi(c \log n)$. Thus the probability that we don't see a sample from this bin is $\Pr[X_i = 0] = e^{-c \log n} = \frac{1}{n^c}$. Union bounding this over at most $\frac{n}{c \log^2 n}$ such high probability bins we can see that whp we would see atleast one sample from each of the high mass bins and thus we have a good estimate of their probability.

So now we can assume that $p_i \leq \frac{c \log^2 n}{n}$ on the remaining bins for small c. Thus we want to estimate $\sum_i f(p_i)$ for the remaining bins, where $f(x) = |x - \frac{1}{n}|$.

Note that looking at samples and calculating the number of bins with k samples we can estimate $\sum_i e^{-mp_i} \frac{(mp_i)^k}{k!}$. In fact we can estimate linear combinations of these:

$$\sum_k c_k \sum_i e^{-mp_i} \frac{(mp_i)^k}{k!} = \sum_i \Big[\sum_k c_k e^{-mp_i} \frac{(mp_i)^k}{k!} \Big]$$

Thus a natural idea is to find c_k's so that

$$\left| x - \frac{1}{n} \right| \approx \sum_k c_k e^{-mx} \frac{(mx)^k}{k!} \ \textit{Uniformly on} \ \left[0, \frac{c \log^2 n}{n} \right].$$

which is equivalent to finding c_k's to approximate,

$$e^{mx}|x-\frac{1}{n}| \approx \sum_k c_k \frac{(mx)^k}{k!} \ \textit{Uniformly on} \ \Big[0, \frac{c\log^2 n}{n}\Big].$$

We solve this as an approximation problem.

Approximation Theory Note that we can approximate e^{mx} by a Taylor series $e^{mx} \approx 1 + mx + \frac{m^2x^2}{2} + \cdots$. Thus lets first focus only on approximating $\left|x-\frac{1}{n}\right|$ by polynomials. For the ease of calculations we write

$$\left|x-\frac{1}{n}\right| = \Big(x-\frac{1}{n}\Big) + 2\Big(\frac{1}{n}-x\Big)I_{x<\frac{1}{n}}$$

Since $\Big(x-\frac{1}{n}\Big)$ is already a polynomial, we focus on approximating $g(x) = 2\Big(\frac{1}{n}-x\Big)I_{x<\frac{1}{n}}$ plotted below.

First we do the following variable change to change the domain to $[-1,1]$:

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

Now we are ready to make the further variable change $y = \cos(\theta), \theta \in [-\pi, \pi]$ so that we get a periodic function that can be easy to use Fourier analysis on. Thus we need to approximate with polynomials in $\cos\theta$.

$$g(\cos\theta) = \frac{2}{n}\Big(1 - c\sin^2(\theta/2)\log^2 n\Big)I_{x<\frac{1}{n}}$$

Note that $g(\cos\theta)$ is Lipschitz with constant $M = \frac{\sqrt{c}\log n}{n}$. The idea is to use Dirichlet kernels to do this approximation. The following description is copied from wikipedia.

Dirichlet Kernels The kth Dirichlet Kernel D_k is defined as follows:

$$D_k(\theta) = \sum_{d=-k}^{d=k} e^{id\theta} = 1 + 2\sum_{d=1}^{k}\cos(k\theta) = \frac{\sin((k+1/2)\theta)}{\sin(\theta/2)}.$$

The convolution of $D_k(\theta)$ with any function r of period 2π is the kth-degree Fourier series approximation to r, i.e., we have

$$(D_k * r)(\theta) = \frac{1}{2\pi}\int_{-\pi}^{\pi} r(\theta-\phi)D_k(\phi)\,d\phi = \sum_{d=-k}^{k}\hat{r}(d)e^{id\theta}$$

where

$$\hat{r}(d) = \frac{1}{2\pi}\int_{-\pi}^{\pi} r(\phi)e^{-id\phi}\,d\phi$$

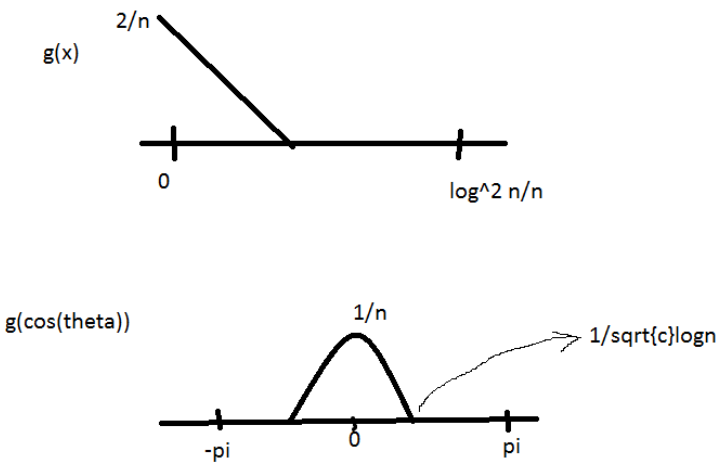

Figure 1: Plots of $g(x), g(\cos\theta)$

Let us use these Dirichlet kernels to come up with polynomials that approximate g. Define

$$s(\cos\theta) = g(\cos\theta) * D_k(\theta).$$

We first verify that this is indeed a polynomial in $\cos\theta$. Firstly using the property of D_k defined above we have

$$(D_k * g\cos)(\theta) = \sum_{d=-k}^{k} \widehat{g\cos}(d)e^{id\theta}.$$

Now since cos is an even function we have $\widehat{g\cos}(d) = \widehat{g\cos}(-d)$. We have

$$s(\cos\theta) = \sum_{d=-k}^{k} \widehat{g\cos}(d)e^{id\theta} = \sum_{d=0}^{k} \widehat{g\cos}(d)2\cos(d\theta) = \sum_{d=0}^{k} 2\widehat{g\cos}(d)T_d(\cos\theta)$$

where T_d is the degree-d Chebyshev polynomial. Thus s is indeed a polynomial of degree-k in $\cos\theta$.

The Dirichlet $D_k(\theta)$ kernel is enveloped by $\frac{1}{\theta}$ but since we need a sharper decay, we raise it to the power of 4.

$$h(\theta) \stackrel{def}{=} D_k^4(\theta).$$

Note that h satisfies the following properties.

- $|h|_1 = \frac{1}{2\pi}\int_{-\pi}^{\pi} h(\theta)\,d\theta = \Theta(k^3)$.

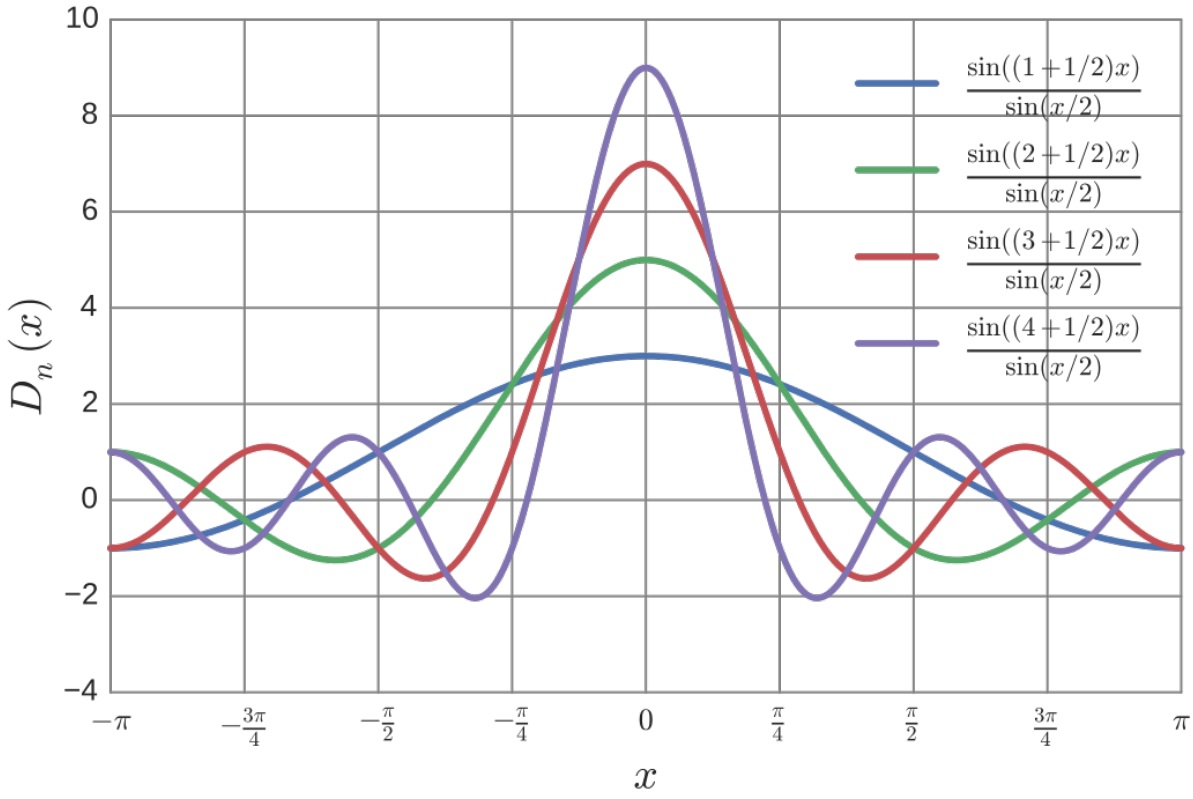

Figure 2: Dirichlet Kernel

- $\int_{-\pi}^{\pi} \theta h(\theta)\, d\theta = O(k^2)$.

Define $p(\cos\theta) = g(\cos\theta) * \frac{h(\theta)}{|h|_1}$, This would still be a polynomial of degree $4k$ in $\cos\theta$ using a similar argument as above. We show that this approximates g well if k is chosen appropriately.

$$\begin{aligned}
|p(\cos\theta) - g(\cos\theta)| &= \Big|\frac{1}{2\pi}\int_{-\pi}^{\pi} g(\cos(\theta-\phi))\frac{h(\phi)}{|h_1|}\,d\phi - g(\cos\theta)\Big| \\
&= \Big|\frac{1}{2\pi}\int_{-\pi}^{\pi} g(\cos(\theta-\phi)) - g(\cos\theta)\frac{h(\phi)}{|h_1|}\,d\phi\Big| \\
&\le \frac{1}{2\pi}\int_{-\pi}^{\pi} |g(\cos(\theta-\phi)) - g(\cos\theta)|\frac{h(\phi)}{|h_1|}\,d\phi \\
&\le \frac{1}{2\pi}\int_{-\pi}^{\pi} M|\phi|\frac{h(\phi)}{|h_1|}\,d\phi \le O\Big(\frac{M}{k}\Big)
\end{aligned}$$

where M is the Lipschitz constant of $g\cos$ and was given by $M = \frac{\sqrt{c}\log n}{n}$. So we pick $k = c^{1/4}\log n$ so that the approximation error is $O(\frac{c^{1/4}}{n}) < \frac{1}{1000n}$.

This shows that the degree $c^{1/4}\log n$ polynomial we constructed in fact approximates $e^{mx}\Big|x - \frac{1}{n}\Big|$ uniformly to error $\frac{1}{1000n}$ on $\Big[0, \frac{c\log^2 n}{n}\Big]$ where $m = \frac{n}{\log n}$. But we would also need a good bound on the coefficients because that would be useful in bounding the variance of the statistic we will eventually come up with.

Note that $\widehat{g\cos}(d) = O(\frac{1}{n})$ because $g \le \frac{2}{n}$ and also the coefficients of $T_k(y)$ are of the size $2^{O(k)}$. When we do a variable change $y \to x : y = 1 - x\frac{n}{c\log^2 n}$, $T_k\Big(1 - x\frac{n}{c\log^2 n}\Big)$ would now have coefficients of size $2^{O(k)}\Big(\frac{n}{c\log^2 n}\Big)^k$. Thus $p(\cos\theta)$ when seen as a function of x would have the coefficient of x^t of the size $\frac{1}{n}\Big(\frac{n}{\log^2 n}\Big)^t O(\frac{1}{c})^k$.

Now we work on approximating e^{mx}. Using Taylor series we have,

$$e^{mx} = 1 + mx + \frac{(mx)^2}{2} + \cdots + \frac{(mx)^t}{t!} + \ \mathit{error}\ .$$

where the error is $O\Big(\frac{mx}{t}\Big)^t$. Since $mx \le \frac{Cn}{\log n}\frac{c\log^2 n}{n} = c'\log n$, choose $t = \frac{2\log n}{\log\frac{1}{c'}}$ so that the Taylor error is $\ll \frac{1}{n}$.

Thus at this point we have,

$$\Big|x - \frac{1}{n}\Big| = \frac{1}{n}\Big[\sum_{d=0}^{k} a_d\Big(\frac{xn}{\log^2 n}\Big)^d\Big] + \frac{1}{1000n}$$

where $a_d = O\Big(\frac{1}{c}\Big)^k$ where $k = o(\log n)$. Now let b_d be defined as

$$a_d\Big(\frac{n}{\log^2 n}\Big)^d = b_d\frac{m^d}{d!}$$

Since $m = C\frac{\log n}{n}$, we note that $\frac{b_d}{a_d} = \frac{d!}{C\log^d n} < \left(\frac{d}{\log n}\right)^d < 1$ because $d \le k = o(\log n)$. Thus after this substitution the coefficients could only get smaller.

Thus we have

$$\left|x - \frac{1}{n}\right| = \frac{1}{n}\Big[\sum_{d=0}^{k} b_d \frac{(mx)^d}{d!}\Big] + \frac{1}{1000n}$$

$$e^{mx} = \sum_{d=0}^{t} \frac{(mx)^d}{d!} + \frac{1}{1000n}$$

So

$$e^{mx}\left|x - \frac{1}{n}\right| = \sum_{d} \frac{c_d}{d!}(mx)^d + \frac{e^{mx}}{1000n}.$$

where $c_d = \frac{1}{n}\sum_{l+w=d}\binom{d}{w}b_w < \frac{2^{O(k)}}{n}$ where $k = o(\log n)$. Thus choose the constants such that $2^{O(k)} < n^{1/3}$. Then $c_d < \frac{1}{n^{2/3}}$. Thus transposing e^{mx}, we have

$$\left|x - \frac{1}{n}\right| = \sum_{d=0}^{k} c_d\, e^{-mx}\frac{(mx)^d}{d!} + \frac{1}{1000n}.$$

where $c_d < \frac{1}{n^{2/3}}, k = o(\log n)$.

2 Tester

We came up with coefficients c_d such that

$$\left|x - \frac{1}{n}\right| \approx \sum_{d=0}^{k} c_d e^{-mx}\frac{(mx)^d}{d!}$$

where $k = o(\log n)$ and $|c_d| \le \frac{1}{n^{2/3}}$.

We now use this to construct a tester. Take $Poi(m)$ samples from p. Let X_i samples fall in bin i. Then we define our statistic

$$Z = \sum_i c_{X_i}.$$

Then the expected value of Z is

$$\mathbb{E}[Z] = \sum_i \sum_{d=0}^{k} c_d e^{-mx}\frac{(mx)^d}{d!} = \sum_i |p_i - \frac{1}{n}| \pm \frac{1}{1000n} = |p - U_n|_1 \pm \frac{1}{1000}.$$

We also need to bound the variance of Z so that we have a good estimate whp.

$$Var[Z] = \sum_i Var[c_{X_i}] \le \sum_i \mathbb{E}[c_{X_i}^2] \le \sum_i \frac{1}{n^{4/3}} = \frac{1}{n^{1/3}}.$$

Thus using Chebyshev we have whp $|Z - \mathbb{E}[Z]| < \frac{1}{1000}$. Thus we have whp $Z - |p - U_n|_1 < \frac{1}{100}$ which proves the upper bound of $\frac{n}{\log n}$ samples.

Open Problem: Figure out the exact dependence of the number of samples needed on ϵ.