CSE 291 - Lecture 11

Vaishakh Ravindrakumar

May 14, 2017

In this write-up, we consider the problem of tolerant uniformity testing and use the idea of polynomial approximation to obtain an upper bound on its sample complexity.

1 Definition

In tolerant uniformity testing, given samples from a distribution p, we'd like to distinguish between $|p - U_n| < \frac{1}{1000}$ vs $|p - U_n| > 1/10$.

2 The Big Picture

Our algorithm for testing involves drawing $\frac{c.n}{log(n)}$ samples from p, using the emipirical probability for the 'heavy' bins ($p_i \geq \frac{log^2(n)}{n}$) and polynomial approximation of the total loss contributed by the 'light' bins - $\sum_{i:p_i<\frac{log^2(n)}{n}} |p_i - \frac{1}{n}|$.

Now if our sampling is Poisson($\frac{c.n}{log(n)}$), with high probability, the 'heavy' samples will occur $\Omega(c.log(n))$ times, thus giving us enough samples to learn the combined 'heavy' mass $\sum_{i:p_i\geq\frac{log^2(n)}{n}} |p_i - \frac{1}{n}|$ to within $\frac{1}{1000}$. We then discard the 'heavy' bin samples and henceforth consider only the 'light' bins.

Here we first note that since our parameter of interest - $\sum_i |p_i - \frac{1}{n}|$ is symmetric in p_i, a sufficient statistic is the number of bins with a certain number of samples. Let B_k denote the number of bins with k samples. This allows us to estimate $\sum_i e^{-mp_i} \frac{(mp_i)^k}{k!}$.

Thus if we can approximate the function $f(x) \triangleq |x - \frac{1}{n}| \approx \sum_k c_k e^{-mx} \frac{(mx)^k}{k!}$ uniformly on $[0, \frac{c.log^2 n}{n}]$, we have

$$\sum_i |p_i - \frac{1}{n}| \approx \sum_i \sum_k c_k e^{-mp_i} \frac{(mp_i)^k}{k!} = \sum_k c_k \sum_i e^{-mp_i} \frac{(mp_i)^k}{k!}$$

where the latter can be estimated from B_k for each k.

3 Approximating $|x - \frac{1}{n}|$

We first write $f(x) = |x - \frac{1}{n}| = (x - \frac{1}{n}) + 2(\frac{1}{n} - x)\mathbb{1}_{x<\frac{1}{n}}$. Since $(x - \frac{1}{n})$ is already a polynomial, we consider approximating only $g(x) = 2(\frac{1}{n} - x)\mathbb{1}_{x<\frac{1}{n}}$.

Further, since our domain of interest is $[0, \frac{c.log^2(n)}{n}]$ but we'd like to use Chebyshev polynomials for approximation, we do a change of variables $y = 1 - 2x\frac{n}{c.log^2(n)}$ so that the new domain is $[-1, 1]$. Then, consider

$$g(cos(\theta)) = \frac{2}{n}(1 - c.sin^2(\theta/2)log^2(n))\mathbb{1}_{x(y=cos(\theta))<\frac{1}{n}}.$$

Since $g(cos(\theta))$ is Lipschitz with constant $M = \frac{\sqrt{c}.log(n)}{n}$, we may use Dirchlet kernels to do the approximation. The k^{th} Dirchlet Kernel D_k is defined as

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

Further, the convolution of any 2π periodic function of θ, $r(\theta)$ with the D_k gives rise to a summation of the Fourier coefficients of of $r(\theta)$. This is because

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

where $\hat{r}(l) = \frac{1}{2\pi}\int_{-\pi}^{\pi} r(\phi)e^{-il\phi}d\phi$.

We then use the Dirchlet kernel to approximate $g(.)$. Define

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

As shown above, we then have $g(cos(\theta)) * D_k(\theta) = \sum_{l=-k}^{k} \widehat{g(cos(\theta))}_l e^{il\theta}$ where $\widehat{g(cos(\theta))}_l$ are the Fourier coefficients of $g(cos(\theta))$. Since $g(cos(\theta))$ is an even function of θ, the Fourier co-efficients are also even. Thus

$$s(cos(\theta)) = \sum_{l=-k}^{k} \widehat{g(cos(\theta))}_l e^{il\theta} = \sum_{l=0}^{k} 2\widehat{g(cos(\theta))}_l cos(l\theta) = \sum_{l=0}^{k} 2\widehat{g(cos(\theta))}_l T_l(cos(\theta))$$

where $T_l(cos(\theta))$ is the l^{th} Chebyshev polynomial. It may thus be seen that $s(.)$ is indeed a k-degree polynomial in $cos(\theta)$.

Since we require our Kernel to have a sharp decay, we use $h(\theta) \triangleq D_k^4(\theta)$ in place of $D_k(\theta)$. Then we have that $|h|_1 = \int_{-\pi}^{\pi} h(\theta)d\theta = \Theta(k^3)$ and $\int_{-\pi}^{\pi} \theta h(\theta) = \Theta(k^2)$. Then, on defining $p(cos(\theta)) = g(cos(\theta)) * \frac{h(\theta)}{|h|_1}$, $p(cos(\theta))$ is still a polynomial of degree $4k$ and the approximation error becomes

$$\begin{aligned}
|p(cos(\theta)) - g(cos(\theta))| &= \left|\frac{1}{2\pi}\int_{-\pi}^{\pi} g(cos(\theta - \phi))\frac{h(\phi)}{|h|_1}d\phi - g(cos(\theta))\right| \\
&= \left|\frac{1}{2\pi}\int_{-\pi}^{\pi} \left(g(cos(\theta - \phi)) - g(cos(\theta))\right)\frac{h(\phi)}{|h|_1}d\phi\right| \\
&\leq \frac{1}{2\pi}\int_{-\pi}^{\pi} |g(cos(\theta - \phi)) - g(cos(\theta))|\frac{|h(\phi)|}{|h|_1}d\phi \\
&\leq \frac{1}{2\pi}\int_{-\pi}^{\pi} M|\phi|\frac{|h(\phi)|}{|h|_1}d\phi \\
&\leq O(M/k).
\end{aligned}$$

Here M is the Lipschitz constant of $g(\cos(\theta))$, which is given by $\frac{\sqrt{c}.log(n)}{n}$. Thus, choosing $k = c^{1/4}log(n)$ allows us to bound the approximation error as $O(\frac{c^{1/4}}{n}) < \frac{1}{1000n}$. This shows that our construction approximates $|x - \frac{1}{n}|$ uniformly upto error $\frac{1}{1000n}$ in the interval $[0, \frac{c.log^2(n)}{n}]$.

The coefficients of this polynomial $p(.)$ are of the order of $\widehat{g(cos(\theta))}_l = O(\frac{1}{n})$ where the latter follows since $g(.) = O(\frac{1}{n})$. Now on changing back the variable to x, by substituting $y = 1 - 2x\frac{n}{c.log^2(n)}$ we have that the t^{th} coefficient is of size $O(\frac{1}{n}(\frac{n}{log^2(n)})^t(\frac{1}{c})^k)$.

Now consider e^{mx}. Taylor expansion of the same gives

$$e^{mx} = 1 + mx + \frac{(mx)^2}{2!} + ... + \frac{(mx)^t}{t!} + O(mx/t)^t.$$

Since $mx \leq \frac{Cn}{log(n)}\frac{c.log^2(n)}{n}$, we choose $t = \frac{2log(n)}{log(1/c')}$, so that $O(mx/t)^t \ll \frac{1}{n}$.

We thus have

$$\left|x - \frac{1}{n}\right| = \frac{1}{n}\left[\sum_{l=0}^{k} a_l(\frac{xn}{log^2(n)})^l\right] + \frac{1}{1000n}$$

where $a_l = O(\frac{1}{c})^k$ with $k = o(log(n))$. Defining $b_l\frac{m^l}{l!} = a_l(\frac{n}{log^2(n)})^l$, this becomes

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

Note that the coefficients $b_l \leq a_l$ from the fact that $m = \frac{C.n}{log(n)}$ giving $\frac{b_l}{a_l} = \frac{l!}{C.log^l(n)} < (\frac{l}{log(n)})^l < 1$ since $l \leq k = o(log(n))$. This, along with

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

gives

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

where $c_l = \frac{1}{n}\sum_{w:d+w=l}\binom{l}{w}b_w < \frac{2^{O(k)}}{n}$ where $k = o(log(n))$. From our choice of constants (since $k = o(log(n))$, $2^{O(k)} \leq n^{1/3}$, implying that $c_l < \frac{1}{n^{2/3}}$.

4 Testing

In the previous section, we obtained the coefficients c_l such that

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

with $k = o(log(n))$ and $c_l < \frac{1}{n^{2/3}}$.

To use this to construct a tester, we take Poisson(m) samples from p and let X_i be the number of samples in Bin i. Our test statistic is then $X = \sum_i C_{X_i}$. The expected value of Z is then

$$\mathbb{E}[Z] \approx \sum_i \sum_{l=0}^{k} c_l e^{-mp_i} \frac{(mp_i)^l}{l!} = \sum_i \left(\left| p_i - \frac{1}{n} \right| \pm \frac{1}{1000n} \right) = |p - U_n|_1 \pm \frac{1}{1000n}.$$

Moreover, the variance of Z is given by

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

Thus with high probability $|Z - \mathbb{E}[Z]| \leq \frac{1}{1000}$ and from the triangle inequality, $|Z - |p - U_n|_1| \leq \frac{2}{1000}$, finally giving us our upper bound of $\frac{n}{log(n)}$ samples.