6.5240 Sublinear Time Algorithms	October 26, 2022

Lecture 13

Lecturer: Ronitt Rubinfeld — *Scribe: Kai Zheng*

In this class we will give an algorithm for uniformity testing. For distributions p, q over a domain D, define the ℓ_1 and ℓ_2 distances as follows.

Definition 1 *The ℓ_1 and ℓ_2 distances are given by,*

- $\ell_1(p,q) = \sum_{x\in D} |p(x) - q(x)|$
- $\ell_2(p,q) = \sqrt{\sum_{x\in D}(p(x) - q(x))^2}$.

We also use $||p||_2$ to denote the ℓ_2-norm which is given by,

- $||p||_2 = \sqrt{\sum_{x\in D} p(x)^2}$.

Let U denote the uniform distribution over D, i.e., $U(x) = \frac{1}{|D|}$ for all $x \in D$. Given sample access to a distribution p, the goal of uniformity testing is to:

- If $p = U$, pass with probability at least 2/3.
- If $\text{dist}(p, U) > \epsilon$, fail with probability at least 2/3.

We will give algorithms for dist as both ℓ_1 and ℓ_2. We start with ℓ_2. The algorithm is as follows.

1. Take $s = \Omega(\epsilon^{-4})$ samples from p, $x_1, \ldots, x_s$
2. Set $\hat{c} \leftarrow$ to be the estimate of $||p||_2^2$ (described next).
3. If $\hat{c} < \frac{1}{n} + \frac{\epsilon^2}{2}$, pass. Otherwise fail.

The idea is that $\hat{c}$ will be an estimate of the collision probability of p, which should be close to $1/n$ if p is close to uniform. To get the estimate of $||p||_2^2$, we do the following,

1. For all i, j, set $\sigma_{ij=1}$ if $x_i = x_j$ and 0 otherwise.
2. Set $\hat{c} \leftarrow \frac{\sum_{i<j}\sigma_{ij}}{\binom{s}{2}}$.

We record some straightforward facts that will be helpful for our analysis.

Lemma 1 *The following are true.*

1. $||p - U||_2^2 = \sum_{i\in D} p(i)^2 - \frac{1}{n}$.
2. $\mathrm{E}[\hat{c}] = ||p||_2^2 = \mathrm{E}[\sigma_{ij}]$.
3. $\mathrm{Var}[\hat{c}] = \frac{\mathrm{Var}(\sum_{i<j}\sigma_{ij})}{\binom{s}{2}}$.
4. $\left(\sum_{x\in D} p(x)^3\right)^{1/3} \le \left(\sum_{x\in D} p(x)\right)^{1/2}$.
5. $s^2 \le 3\binom{s}{2}$.
6. $\binom{s}{3} \le \frac{s^3}{6}$

From this lemma, we see that if $|\hat{c} - ||p||_2^2| < \frac{\epsilon^2}{2}$, then the algorithm outputs the right answer. Indeed, by the first point, if $p = U$, then $||p||_2^2 = \frac{1}{n}$ and we would get $\hat{c} < \frac{1}{n} + \frac{\epsilon^2}{2}$ resulting in pass as desired. Otherwise, if p is ϵ-far from uniform then the first point implies that $||p||_2^2 \geq \frac{1}{n} + \epsilon^2$ and thus $\hat{c} \geq < \frac{1}{n} + \frac{\epsilon^2}{2}$ resulting in reject as desired. To complete our analysis, we will show that $|\hat{c} - ||p||_2^2| < \frac{\epsilon^2}{2}$ with probability at least 2/3 over the random samples $x_1, \ldots, x_s$. To this end, we bound the variance of $\hat{c}$ and use Chebyshev's.

Lemma 2 *We have,*

$$\mathrm{Var}\left[\sum_{i<j} \sigma_{ij}\right] \leq 4\left(\binom{s}{2} ||p||_2^2\right)^{3/2}.$$

Proof First let $\overline{\sigma_{ij}} = \sigma_{ij} - \mathrm{E}[\sigma_{ij}]$. Then, $\mathrm{E}[\overline{\sigma_{ij}}] = 0$. Moreover, we have that

$$\mathrm{E}[\overline{\sigma_{ij}}\ \overline{\sigma_{kl}}] = \mathrm{E}[\sigma_{ij}\sigma_{kl}] - \mathrm{E}[\sigma_{ij}]^2 \leq \mathrm{E}[\sigma_{ij}\sigma_{kl}]. \tag{1}$$

We decompose the variance as,

$$\mathrm{Var}\left(\sum_{i<j} \sigma_{ij}\right) = \mathrm{E}\left[\sum_{i<j} \overline{\sigma_{ij}}^2 + \sum_{i<j,k<\ell,\text{all distinct}} \overline{\sigma_{ij}}\ \overline{\sigma_{k\ell}} + \sum_{i,j,k,\ell\ 3\text{ distinct}} \overline{\sigma_{ij}}\ \overline{\sigma_{k\ell}}\right],$$

and bound each term separately. For the first term,

$$\mathrm{E}\left[\sum_{i<j} \overline{\sigma_{ij}}^2\right] \leq \binom{s}{2} ||p||_2^2,$$

using part 1 of Lemma 1 and (1).

For the second term,

$$\mathrm{E}\left[\sum_{i<j} \overline{\sigma_{ij}}\ \overline{\sigma_{k\ell}}\right] = 0,$$

by independence and the fact that $\mathrm{E}[\overline{\sigma_{ij}}] = 0$.

For the third term, we can have $i < j$, and $k < \ell$ with 3 distinct in several ways. We could have, $i = k$, $j = \ell$, $j = k$, or $i = \ell$. However, it is not hard to see that the same bound will hold for each, so we simply give a bound for the sum over $i < j$, $k < \ell$ such that $i = k$.

$$\begin{aligned}
\mathrm{E}\left[\sum_{i<j,i<\ell} \overline{\sigma_{ij}}\ \overline{\sigma_{k\ell}}\right] &\leq \mathrm{E}\left[\sum_{i<j,i<\ell} \sigma_{ij}\sigma_{i\ell}\right] \\
&\leq \sum_{i,j,\ell\text{ distinct}} \mathrm{E}[1_{x_i=x_j=x_\ell}] \\
&\leq \binom{s}{3} \sum_{x\in D} p(x)^3 \\
&\leq \frac{s^3}{6}\left(\sum_{x\in D} p(x)^2\right)^{3/2} \\
&\leq \frac{\sqrt{3}}{2}\binom{s}{2}^{3/2} \left(||p||_2^2\right)^{3/2}.
\end{aligned}$$

where we use the fourth part of Lemma 1 in the first line, the sixth part to get the fourth line, and the fifth part to get the last line. As the same bound holds for the other cases with 3 distinct out of i, j, k, ℓ, we get an overall bound of

$$\mathrm{Var}\left(\sum_{i<j} \sigma_{ij}\right) \leq \binom{s}{2}||p||_2^2 + 4 \cdot \frac{\sqrt{3}}{2}\binom{s}{2}^{3/2}\left(||p||_2^2\right)^{3/2} \leq 4\left(\binom{s}{2}||p||_2^2\right)^{3/2}.$$

■

We now apply Chebyshev's to get the following.

Lemma 3

$$\Pr_{x_i' s}[|\hat{c} - ||p||_2^2| > \epsilon^2/2] < \frac{1}{3}.$$

Proof Applying Chebyshev's yields,

$$\begin{aligned}
\Pr_{x_i' s}[|\hat{c} - ||p||_2^2 > \epsilon^2/2] &\leq \frac{\mathrm{Var}(\hat{c})}{(\epsilon^2/2)^2} \\
&\leq \frac{k\binom{s}{2}^{3/2}(||p||_2^2)^{3/2}}{\binom{s}{2}^2 \epsilon^4} \\
&= O\left(\frac{1}{s\epsilon^4}\right) < 1/3,
\end{aligned}$$

where k is some constant in $s = \Omega(\epsilon^{-4})$, chosen so that the last inequality holds. Note that the first line uses fact 3 of Lemma 1 to go from $\mathrm{Var}(\sum \sigma_{ij})$ to $\mathrm{Var}(\hat{c})$. ■

As discussed, this shows the correctness of the algorithm. We now describe how to a similar algorithm for ℓ_1 distance. Notice that $\ell_1(p, U) = 0$ is equivalent to $\ell_2(p, U) = 0$ and $||p||_2^2 = \frac{1}{n}$. On the other hand, if $\ell_1(p, U) > \epsilon$, then $\ell_2(p, U) > \frac{\epsilon}{\sqrt{n}}$ and thus $||p||_2^2 > \frac{1}{n} + \frac{\epsilon^2}{n}$. Therefore we need to estimate $||p||_2^2$ to an within an additive error of $\epsilon^2/(2n)$ and pass if and only if $\hat{c} < \frac{1}{n} + \frac{\epsilon^2}{2n}$. Given the bound on $||p||_2^2$ in the ϵ-far case, this additive error can also be achieved by a multiplicative error of $1 \pm \epsilon^2/3$. To accomplish this we run the same algorithm with $s = \Omega(\sqrt{n}\epsilon^{-4})$. Then by Chebyshev's

$$\begin{aligned}
\Pr_{x_i' s}\left[|\hat{c} - ||p||_2^2| \leq (\epsilon^2/3)||p||_2^2\right] &\leq \frac{\mathrm{Var}(\hat{c})}{\epsilon^4||p||^2)2/9} \\
&\leq \frac{k'}{\epsilon^4||p||_2 s} \\
&\leq \frac{k'\sqrt{n}}{\epsilon^4 s} \\
&\leq \frac{1}{3}
\end{aligned}$$

where we use the fact that $||p||_2 > 1/\sqrt{n}$ to get the second line, and choose k' appropriately to make obtain the last line.