6.5240 Sublinear Time Algorithms	October 23, 2024

Lecture 14

Lecturer: Ronitt Rubinfeld — *Scribe: Manya Bansal*

In this class we consider a new model: given a Domain D and the ability to sample elements of this domain using a sampler P, we would like to learn the underlying distribution of the domain. In this model, the size of the domain, i.e. $|D| = n$, is known, and we would like to achieve sublinear sample complexity in n.

For the remainder of the class, we describe a tester for the case when P is uniformly distributed on D, denoted by U_D.

1 Testing Uniformity

We would like to define a tester for P such that:

- If $P = U_D$, then the tester outputs $PASS$.
- If $dist(P, U_D) > \epsilon$, then the tester outputs $FAIL$.

There are several choices for $dist(P, U_D)$, for example we can pick one of the following:

- l_1-distance: $||p - q||_1 = \sum_{i \in D} |p_i - q_i|$.
- l_2-distance: $||p - q||_2 = \sqrt{\sum_{i \in D} (p_i - q_i)^2}$.

During the course of the proof, we will provide an astounding tester with respect to the l_2-distance. In particular, we will estimate the l_2-distance upto a multiplicative factor using only constant number of samples. While this seems very strong, part of the reason why it works out is the fact that the l_2-distance is a "weird" measure of distance. To see why, consider the case when:

$$p = (1, 0, \dots, 0), \qquad q = (\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n}).$$

Then, l_2-distance $= (||p - q||_2)^2 = \left(1 - \frac{1}{n}\right)^2 + (n-1)\left(\frac{1}{n}\right)^2 = 1 + \frac{1}{n^2} - \frac{2}{n} + \frac{n-1}{n} \approx 1$. However, when we define two distributions that can be as far as possible:

$$p = (\frac{2}{n}, \frac{2}{n}, \dots, \frac{2}{n}, 0, 0 \dots, 0), \qquad q = (0, 0 \dots, 0, \frac{2}{n}, \frac{2}{n}, \dots, \frac{2}{n}).$$

i.e. distributions with disjoint support, we get that l_2-distance is $\sqrt{n\left(\frac{2}{n}\right)^2} = \frac{2}{\sqrt{n}}$, which is quite small!

1.1 Naive Algorithm

Algorithm:

- Take m samples from p.
- For all x, estimate $p(x)$ by computing the observed frequency $\hat{p}(x) = \frac{\#\text{times } x \text{ appears in sample}}{m}$
- If the observed frequency is far from unfiorm i.e. $\sum_x \left|\hat{p}(x) - \frac{1}{n}\right| > \epsilon$, $REJECT$, else $ACCEPT$.

The above algorithm will achieve our goals. We want to pick m such that *whp*, for x the error is bound by $\frac{\epsilon}{n}$ to get that:

$$\forall x, |p(x) - \hat{p}(x)| < \frac{\epsilon}{n}$$
$$\implies ||p - p'||_1 = \sum_{x \in D} |p(x) - \hat{p}(x)| < \epsilon.$$

If $p = U_D$, then this clearly gives us the correct answer. To see why it would fail if p is 2ϵ far, we consider the contrapositive. That is, if our test accepts, then p is not 2ϵ far. Notice that if:

$$||p - \hat{p}||_1 < \epsilon.$$

and,

$$||\hat{p} - U_D||_1 < \epsilon.$$

then, by the triangle inequality,

$$||p - U_D||_1 < 2\epsilon.$$

The problem however arises when we start to set m. Since we need to estimate each $p(x)$ within $\frac{\epsilon}{n}$ error, and need to see each sample at least once. Using a coupon collector argument, we will need to consider $\Omega(n)$ samples, which is not sub-linear in n. The hand-written notes also contain a tighter bound for m, but we did not discuss this in class.

1.2 l_2-Distance

We now turn our attention to l_2-distance. Before we proceed, let's establish a useful set of facts:

Fact (*)

$$\begin{aligned} ||p - U_d||_2^2 &= \sum_{x \in D} \left(p_i - \frac{1}{n}\right)^2 \\ &= \sum_i p_i^2 + 2\sum_i p_i \frac{1}{n} - \sum_i \frac{1}{n^2} \\ &= \sum_i p_i^2 + \frac{2}{n}\sum_i p_i - \sum_i \frac{1}{n^2} \\ &= \sum_i p_i^2 + \frac{2}{n} - \frac{1}{n} \\ &= \sum_i p_i^2 - \frac{1}{n}, \end{aligned}$$

where p_i^2 denotes the ***collision probability*** of two elements i.e. we sample the same elements.

For our case of l_2-distance, we will estimate $||p_i||_2^2$ since $||U_D||_2^2$ is already known. To estimate l_2-distance, we follow the algorithm:

Algorithm:

- Take s samples from p.
- For all $\hat{c} \leftarrow$ estimate of $||p||_2^2$ from the sample s.
- If $\hat{c} < \frac{1}{n} + \delta$, then $ACCEPT$, else $REJECT$.

There are three questions that arise:

1. How many samples s should we consider?
2. How will we actually estimate $\hat{c}$?
3. What should δ be?

Question 2

We begin by tackling question 2: how to estimate $\hat{c}$? We follow a strategy called "estimate by recycling".

- Define s samples as $x_1, x_2, \ldots x_s$.
- For each $1 \leq i < j \leq s$, define $\sigma_{ij} \leftarrow$ if $x_i = x_j$ (notice that σ_{ij}'s are identically distributed, though not independent).
- Output $\hat{c} \leftarrow (\sum_{i<j} \sigma_{ij})/\binom{s}{2}$ (i.e. normalize by number of pairs).

Now, we compute the expected value of $\hat{c}$:

$$\mathbb{E}[\hat{c}] = \frac{1}{\binom{s}{2}} \binom{s}{2} \mathbb{E}[\sigma_{ij}] = ||p||_2^2.$$

Question 3

Next, we decide how good our approximation must be i.e. we fix δ. Ultimately, we will use this to decide how many samples we need to take.

We set $\delta = \epsilon^2/2$. To see why this will work, assume that $|\hat{c} - ||p||_2^2| < \epsilon^2/2$. Then, if $p = U_D$:

$$\begin{aligned} \hat{c} &< ||U||_2^2 + \epsilon^2/2 && \text{(From assumption)} \\ &= \frac{1}{n} + \frac{\epsilon^2}{2}. && \text{(Pass!)} \end{aligned}$$

If $p \neq U_D$: i.e. $||p - U_D||_2 > \epsilon$, then $||p - U_D||_2^2 > \epsilon^2$. Using fact $(*)$, we know that:

$$||p||_2^2 = ||p - U_D||_2 + \frac{1}{n} = \epsilon^2 + \frac{1}{n} > \frac{1}{2} + \frac{\epsilon^2}{2}. \quad \text{(Fail!)}$$

So, now all is left to do is to pick the number of samples such that we can make $|\hat{c} - ||p||_2^2| < \epsilon^2/2$ hold.

Question 1

Let's first establish useful facts about the variance of $\hat{c}$. Ultimately, we will plug this into Chebyshev to get a good bound.

First, note that:

$$Var[\hat{c}] = Var\left[\left(\sum_{i<j}\sigma_{ij}\right) / \binom{s}{2}\right] = \frac{1}{\binom{s}{2}^2} Var\left[\left(\sum_{i<j}\sigma_{ij}\right)\right].$$

Lemma 1. *The variance of $\sum_{i<j}\sigma_i j$ is bounded by $O(s^3||p||_2^3)$.*

Corollary 2. *The above lemma immediately implies that the variance of $\hat{c}$ is bounded by $O(||p||_2^3)/s$.*

Proof of Lemma: To make our analysis easier, define $\bar{\sigma}_{ij} = \sigma_{ij} - \mathbb{E}[\sigma_{ij}]$. We use this definition because the $\mathbb{E}[\bar{\sigma}_{ij}] = 0$, which will be a useful fact to exploit. In particular, notice that:

$$\bar{\sigma}_{ij} < \sigma_{ij},$$
$$\mathbb{E}[\bar{\sigma}_{ij}\bar{\sigma}_{kl}] \leq \mathbb{E}[\sigma_{ij}\sigma_{kl}].$$

To begin our analysis, we simply break up the definition of $Var[\sum_{i<j}\sigma_i]$ into multiple cases:

$$\begin{aligned}
Var\left[\sum_{i<j}\sigma_i\right] &= \mathbb{E}\left[\left(\sum_{i<j}\sigma_{ij} - \mathbb{E}\left[\sum_{i<j}\sigma_{ij}\right]\right)^2\right] \\
&= \mathbb{E}\left[\left(\sum_{i<j}\bar{\sigma}_{ij}\right)^2\right] \\
&= \mathbb{E}\left[\underbrace{\sum_{i,j,k,l}\bar{\sigma}_{ij}\bar{\sigma}_{kl}}_{\text{2 unique indices}} + \underbrace{\sum_{i,j,k,l}\bar{\sigma}_{ij}\bar{\sigma}_{kl}}_{\text{3 unique indices}} + \underbrace{\sum_{i,j,k,l}\bar{\sigma}_{ij}\bar{\sigma}_{kl}}_{\text{4 unique indices}}\right].
\end{aligned}$$

We handle each case separately:

1. First, we bound the case where there are only 2 unique indices. We need to compute,

$$\begin{aligned}
\mathbb{E}[\underbrace{\sum_{i,j,k,l}\bar{\sigma}_{ij}\bar{\sigma}_{kl}}_{\text{2 unique indices}}] \leq \mathbb{E}[(\underbrace{\sum_{i,j,k,l}\sigma_{ij}\sigma_{kl})}_{\text{2 unique indices}}] \\
&= \sum_{i<j}\mathbb{E}[(\sigma_{ij})^2] && \text{(Linearity of expectation)} \\
&= \binom{s}{2}||p||_2^2. && (\sigma_{ij}^2 = \sigma_{ij})
\end{aligned}$$

2. Next, we bound the case where there are only 4 unique indices. Since all the indices are distinct, we can exploit independence to factor expectation.

$$\mathbb{E}[\underbrace{\sum_{i,j,k,l} \bar{\sigma}_{ij}\bar{\sigma}_{kl}}_{\text{4 unique indices}}] = \underbrace{\sum_{i,j,k,l} \mathbb{E}[\bar{\sigma}_{ij}]\mathbb{E}[\bar{\sigma}_{kl}]}_{\text{4 unique indices}} = 0.$$

3. Finally, we bound the case where there are 3 unique indices.

$$\begin{aligned}
\underbrace{\sum_{i,j,k,l} \bar{\sigma}_{ij}\bar{\sigma}_{kl}}_{\text{3 unique indices}} &\leq \underbrace{\sum_{i,j,k,l} \sigma_{ij}\sigma_{kl}}_{\text{3 unique indices}} \\
&= \sum_{a,b,c \text{ distinct}} \mathbb{P}[X_a = X_b = X_c] \\
&\leq 6\binom{s}{3}\sum_x p(x)^3 \\
&\leq cs^3\left(\left(\sum_x p(x)^2\right)^{3/2}\right) \quad \text{Using the fact } \sum_x p(x)^3 \leq \left(\sum_x p(x)^2\right)^{3/2} \\
&= O(s^3||p||_2^3).
\end{aligned}$$

Question 2

Finally we turn to the question of number of samples s. We need estimate $||p||_2^2$ within $\epsilon^2/2$. To do so, we will utilize Chebyshev:

$$\begin{aligned}
\mathbb{P}[|\hat{c} - ||p||_2^2| > \epsilon^2/2] &\leq \frac{Var[\hat{c}]}{(\epsilon^2/2)^2} \\
&= \frac{C||p||_2^3}{\epsilon^4 s}.
\end{aligned}$$

We need to pick s big enough to kill the $\frac{C}{\epsilon^4}$ factor since $||p||_2^2 \leq 1$. That is, $s = \Omega(1/\epsilon^4)$. Notably, s is not a functon of n.

Estimating l_1-distance

Using the algorithm described above, we can now show that it is possible to estimate l_1-distance in $O(\sqrt{n}/\epsilon^4)$ samples.

To see why this is correct, notice that:

$$\begin{aligned}
||p - U_d||_1 = 0 &\iff ||p - U_d||_2 = 0 \\
&\iff ||p||_2^2 = \frac{1}{n}.
\end{aligned}$$

and,

$$\begin{aligned}
||p - U_d||_1 > \epsilon &\implies ||p - U_d||_2 > \frac{\epsilon}{\sqrt{n}} \\
&\implies ||p - U_D||_2^2 > \frac{\epsilon^2}{n} \\
&\implies ||p||_2^2 > \frac{1 + \epsilon^2}{n}.
\end{aligned}$$

If we get a multiplicative estimate $\hat{c}$ of $||p||_2^2$ within $\gamma = \epsilon^2/4$, then when $||p - U_d||_1 > \epsilon$, $\hat{c} \geq (1-\gamma)||p_2||^2 \geq \left(1 - \frac{\epsilon^2}{4}\right)\left(\frac{1}{n} + \epsilon^2\right) = \frac{1}{n} + \frac{3\epsilon^2}{4n} - \frac{\epsilon^4}{2n}$, which is sufficiently separated from the other case when $\hat{c} \leq (1 + \frac{\epsilon^2}{4})n$. So, we only need:

$$\begin{aligned}\mathbb{P}\left[|\hat{c} - ||p||_2^2| > \gamma||p||_2^2\right] &\leq \frac{Var[\hat{c}]}{\gamma^2||p||_2^4} \\ &= \frac{C||p||_2^3/s}{||p||_2^4(\epsilon^4/16)} \\ &= \frac{C}{||p||_2(\epsilon^4)s}.\end{aligned}$$

It is always the case that $||p||_2 > \frac{1}{\sqrt{n}}$, so picking $s = \Omega(\sqrt{n}/\epsilon^4)$ suffices.