6.889 Sublinear Time Algorithms March 20, 2019

Lecture 13

Lecturer: Ronitt Rubinfeld *Scribe: Michal Shlapentokh-Rothman*

1 Outline

Today, we will continue our discussion about testing the uniformity of a distribution in sub-linear time. We will focus on the analysis of the algorithm that we went over in the previous lecture.

- Setup and Review from Last Lecture
- Analysis

2 Setup and Review from Last Lecture

Our overall goal is to be able to pass a distribution, D, if it is uniform with probability greater than $\frac{2}{3}$ and reject a distribution D with probability greater than $\frac{2}{3}$, if the $dist(D, Uniform) > \epsilon$. There are two definitions that we consider for distance functions, L_1 and L_2.

Definition 1 *The $L_1(p, q)$ or the L_1 distance between distributions p and q is*

$$L_1(p, q) = ||p - q_1||_1 = \sum_{x \in D} |p(x) - q(x)|$$

Definition 2 *The $L_2(p, q)$ or the L_2 distance between distributions p and q as*

$$L_2(p, q) = ||p - q||_2 = \sqrt{\sum_{x \in D} (p(x) - q(x))^2}$$

It is important to note the relationship between these two definitions:

$$||p - q||_2 \leq ||p - q||_1 \leq \sqrt{n}||p - q||_2$$

where n is the number of points in p and q.

We can also identify an important fact about the *squared* L_2 distance when determining the distance between a distribution p and the uniform distribution U

$$||p - u||_2^2 = \sum_{x \in D} p^2(x) - \frac{1}{n}$$

where we will define $||p||_2^2$ as the collision probability. Please see Lecture 12 notes for a more detailed explanation of this fact.

The last part we will review from last lecture is the algorithm we will use for determining the distance between a distribution and the uniform distribution.

Algorithm 1 Uniformity Testing

```
Take $s$ Samples
$\hat{c} \leftarrow$ estimate of $||p||_2^2$ from sample
if $\hat{c} < \frac{1}{n} + \delta$ then
    Pass
else
    Fail
end if
```

3 Analysis

The rest of the lecture will cover three questions we can ask about the algorithm (described at the end of the last section):

- How well should we estimate $||p||_2^2$?
- What should δ be for the last statement?
- How many samples should we take?

3.1 Estimating Collision Probability

We will start with a naive idea for estimating $||p||_2^2$ or the collision probablility.

Algorithm 2 Naive $||p||_2^2$

```
Take $s$ samples from $p$
for each pair $k$ do
    if $X_k = X_{k+1}$ then
        $\sigma_k = 1$ where $\sigma_k$ is an indicator variable
    else
        $\sigma_k = 0$
    end if
end for
Output $\hat{c} \leftarrow \frac{\sum_{k=1}^{k} \sigma_k}{k}$
```

This algorithm gives us $\Theta(k)$ samples of collision probability from k samples of p. We would like a better query complexity than this.

A better idea is to ‘recycle’ pairs and use **all** pairs in a sample. This way we can get $\Theta(k^2)$ samples of collision probability from k samples.

Algorithm 3 Recycle $||p||_2^2$

```
Take $s$ samples from $p$
for each $1 \leq i \leq j \leq s$ do
    if $X_i = X_j$ then
        $\sigma_{ij} = 1$
    else
        $\sigma_{ij} = 0$
    end if
end for
Output $\hat{C} \leftarrow \frac{\sum_{i<j} \sigma_{ij}}{\binom{S}{2}}$
```

Note that σ_{ij}s are not independent so we cannot use Chernoff to bound our error. Also, note that the expectation of **both** algorithms is correct and equals to the collision probability.

The goal of the rest of our analysis is to show that with enough samples the expectation of our estimate is close enough to the collision probability. Formally, we want to show

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

3.2 Picking δ

Since we are approximating the collision probability, we need to determine how good of an estimate we need to satisfy the algorithm we described in the first section.

Let $\Delta = \frac{\epsilon^2}{2}$. Then we have:

$$|\hat{c} - ||p||_2^2| < \Delta = \frac{\epsilon^2}{2}$$

When $\Delta = \frac{\epsilon^2}{2}$, the algorithm will produce the correct behavior (with enough samples taken):

- If $p = U_{[n]}$ then $\hat{c} \leq ||U_{[n]}||^2 + \Delta = \frac{1}{n} + \frac{\epsilon^2}{2}$, so the test will pass.
- If $||p - U_{[n]}||_2 > \epsilon$ then $||p - U_{[n]}||_2^2 > \epsilon^2$. We will also show that if this happens then $||p||_2^2 = ||p - U_{[n]}||_2^2 + \frac{1}{n} > \epsilon^2 + \frac{1}{n}$ and $\hat{c} > ||p||_2^2 - \Delta \geq \epsilon^2 + \frac{1}{n} - \Delta = \epsilon^2 + \frac{1}{n} - \frac{\epsilon^2}{2} = \frac{\epsilon^2}{2} + \frac{1}{n}$, so the test will fail.

Now we have answered two of our three questions. To estimate the collision probability, we will use the 'Recycle' algorithm and δ should equal $\frac{\epsilon^2}{2}$. The next section will show that the number of samples we need to take can be done in sub-linear queries.

3.3 Determining the Number of Samples

The samples we are drawing are not independent, we will be using Chebyshev's inequality to bound our error:

$$Pr[|\hat{c} - ||p||_2^2| > p] \leq \frac{Var[\hat{c}]}{p^2}$$

From the previous section, we can say the following

$$Var[\hat{c}] = Var[\frac{1}{\binom{s}{2}} \sum_{i<j} \sigma_{ij}] = \frac{1}{\binom{s}{2}^2} Var[\sum_{i<j} \sigma_{ij}]$$

Noting that $Var[aX] = a^2 Var[x]$, we will bound $Var[\sum_{i<j} \sigma_{ij}]$.

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

Proof We will start with a definition:

Definition 4 $\overline{\sigma_{ij}} = \sigma_{ij} - E[\sigma_{ij}] = 0$

The reason why we need this definition is because we will use it as a trick to rewrite $E[\sum \sigma_{ij}]^2$:

$$Var[\sum \overline{\sigma_{ij}}] = E[(\sum \overline{\sigma_{ij}} - E[\sum \overline{\sigma_{ij}}])^2] = E[(\sum \sigma_{ij} - E[\sum \sigma_{ij}])^2] = Var[\sum \sigma_{ij}]$$

So $E[\overline{\sigma_{ij}}] = 0$. We will also use several other facts:

- $E[\overline{\sigma_{ij}}\,\overline{\sigma_{kl}}] \leq E[\sigma_{ij}\sigma_{kl}]$
- $\left(\sum p(x)^3\right)^{\frac{1}{3}} \leq \left(\sum p(x)^2\right)^{\frac{1}{2}}$
- $s^2 \leq 3\binom{s}{2}$
- $\binom{s}{3} \leq \frac{s^3}{6}$

Now, we can say

$$Var[\sum_{i<j} \sigma_{ij}] = E[\left(\sum_{i<j} \sigma_{ij} - E[\sum_{i<j} \sigma_{ij}]\right)^2] = E[(\sum_{i<j} \overline{\sigma_{ij}})^2] =$$

$$\begin{aligned} & E[\sum_{i<j} (\overline{\sigma_{ij}}^2) && (1) \\ & + \sum_{i<j,k<l} \overline{\sigma_{ij}}\,\overline{\sigma_{kl}} && (2) \\ & + \sum_{i<j,k<l} \overline{\sigma_{ij}}\,\overline{\sigma_{il}} && (3) \\ & + \sum_{i<j,k<j} \overline{\sigma_{ij}}\,\overline{\sigma_{kj}} && (4) \\ & + \sum_{i<j,j<l} \overline{\sigma_{ij}}\,\overline{\sigma_{jl}} && (5) \\ & + \sum_{i<j,k<i} \overline{\sigma_{ij}}\,\overline{\sigma_{ki}}] && (6) \end{aligned}$$

We will go through each part of the equation and simplify it using the facts we stated earlier.

1. $E[\sum_{i<j} \overline{\sigma_{ij}^2}] \leq E[\sum \sigma_{ij}^2] = \binom{s}{2}||p||_2^2$. Note that we can make this statement because $E[\sigma_{ij}] = E[\sigma_{ij}^2]$ since σ_{ij} is an indicator variable.
2. We can use our trick to simplify the second term where i, j, k, l are all distinct.

$$E[\sum_{i<j,k<l} \overline{\sigma_{ij}}\,\overline{\sigma_{kl}}] \leq \sum E[\overline{\sigma_{ij}}]E[\overline{\sigma_{kl}}] = 0$$

3. $E[\sum \overline{\sigma_{ij}}\,\overline{\sigma_{il}}] \leq E[\sum_{i,j,l \text{ distinct}} \sigma_{ij}\sigma_{il}] = \sum_{i,j,l \text{ distinct}} Pr[x_i = x_j = x_l]$. The probability that the three indicator variables are equal to each other is the same as the probability of a three way collision, which we can simplify using the facts: $\binom{s}{3} \sum_x p(x)^3 \leq \frac{s^3}{6}(\sum_x p(x)^2)^{\frac{3}{2}} \leq \frac{\sqrt{3}}{2}\binom{s}{2}^{\frac{3}{2}}(||p||_2^2)^{\frac{3}{2}}$
4. Same as 3
5. Same as 3
6. Same as 3

We can put all of this together to say the following

$$Var[\sum_{i<j} \sigma_{ij}] = Var[\sum_{i<j} \overline{\sigma_{ij}}] \leq \binom{s}{2}||p||_2^2 + 0 + 4(\frac{\sqrt{3}}{2})(\binom{s}{2}||p||_2^2)^{\frac{3}{2}} \leq 4[\binom{s}{2}||p||_2^2]^{\frac{3}{2}}$$

■ We can now plug the lemma into Chebyshev with $p = \frac{\epsilon^2}{2}$:

$$Pr[|\hat{c} - ||p||_2^2| > \frac{\epsilon^2}{2}] \leq \frac{Var[\hat{c}]}{\epsilon^4}4 \leq \frac{4[\binom{s}{2}||p||_2^2]^{\frac{3}{2}}}{\binom{s}{2}^2\epsilon^4}4 \leq \frac{32}{\epsilon^4}\frac{1}{s}||p||_2^3$$

Thus, to get the approximating we would like, we can set the number of samples, s to be $\geq \omega\frac{1}{\epsilon^4}$.

3.4 Adjustment for L_1

For the analysis above, we were assuming that the number of samples needed for the L_2 distance was sufficient. However, we need to modify that value to satisfy the L_1 distance as well. If a distribution is uniform, then we can say the following

$$||p-U||_1 = 0 \Leftrightarrow ||p-U||_2^2 = 0 \Leftrightarrow ||p||_2^2 = \frac{1}{n}$$

If a distribution is ϵ-far from uniform, then

$$||p-U||_1 > \epsilon \Rightarrow ||p-U||_2 > \frac{\epsilon}{\sqrt{n}} \Rightarrow ||p-U||_2^2 > \frac{\epsilon^2}{n} \Rightarrow ||p||_2^2 > \frac{1}{n} + \frac{\epsilon^2}{n}$$

This implies that we can have an addistive estimate with error $\leq \frac{\epsilon^2}{2n}$ or multiplicative error $\leq (1 \pm \frac{\epsilon^2}{3})$ which would occur if additive error is $\leq \frac{\epsilon^2}{3n}||p||_2^2$. Now, we need to figure out how to choose the right number of samples such that our additive error is less than or equal to $\frac{\epsilon^2}{3n}||p||_2^2$. We can state the following where k is a constant:

$$Pr[|\hat{c} - ||p||_2^2|] \geq \delta ||p||_2^2] \leq \frac{k||p||_2^3}{s\delta^2(||p||_2^2)^2} \leq \frac{k}{s\delta^2||p||_2} \leq \frac{k\sqrt{n}}{s\delta^2}$$

The statement is true because $||p||_2^2 > \frac{1}{n}$ so $||p||_2 > \frac{1}{\sqrt{n}}$ meaning $\frac{1}{||p||_2} < \sqrt{n}$. Because of the statements above, if we pick the number of samples s to be $>> \frac{\sqrt{n}}{\epsilon^4}$ then we have a small probability of error which is approximately $\frac{k\sqrt{n}}{s\epsilon^4}$.