0368.416701 Sublinear Time Algorithms January 4, 2010

Lecture 11

Lecturer: Ronitt Rubinfeld *Scribe: Moran Tzafrir and Boaz Frankel*

1 Lecture Topic

Testing for uniform distribution.

2 Definitions

For distributions p and q over Domain D we define the distances $L_1$and L_2 between p and q:

Definition 1. $L_1 \equiv \mid p-q \mid_1 = \sum_{x \in D} \mid p(x)-q(x) \mid$

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

Property of $L_1$and L_2: $\| p-q \|_2 < \mid p-q \mid_1 \leq \sqrt{n} \| p-q \|_2$

3 Testing for uniform distribution

Let u be the uniform distribution on the the domain [n]. Given a distribution p. we want to output "fail" if $\| p-u \|_2 > \varepsilon$ (or $\mid p-u \mid_1 > \varepsilon$) and output "pass" if $p = u$.

3.1 Naive Algorithm

given p, n

- take a sample of m elements from p.
- estimate $p(x)$ for each $x_\epsilon [n]$: $\hat{p}(x) = \frac{\#\ of\ times\ we\ saw\ x\ in\ the\ samples}{\#\ of\ samples}$
- if $\| \hat{p}-u \|_2 = \sqrt{\sum_{x \in D} (\hat{p}(x)-\frac{1}{n})^2} > \varepsilon$ REJECT
- ACCEPT

The problem is that for getting good approximation by the chernoff bound we need to use a linear number of samples to estimate p.

Another Idea: Instead of trying to estimate $p(x)$ for each x in the domain, try to estimate the total probability of collision**.**

3.2 Main Observation

Notation: We will use u for the uniform distribution.

Definition 3. The "collision probability" of distribution p is $\underset{s1,s2 \in P}{Pr}[s_1 = s_2] = \sum_x P(x)^2 = (\| P(x) \|_2)^2$

Observation 4. $(\| p-u \|_2)^2 \overset{def\ of\ L_2}{=} \sum_{x\in[n]} (p(x)-\frac{1}{n})^2 = \sum_{x\in[n]}(p(x)^2 - \frac{2p(x)}{n} - \frac{1}{n^2}) =$

$\sum_{x\in[n]} p(x)^2 + n\cdot\frac{2}{n}\cdot \sum_{x\in[D]}^{\frac{1}{n}} p(x)^2 + n\cdot\frac{1}{n^2} = (\| P(x) \|_2)^2 - \frac{1}{n}$

Corollary 5. We can find the distance between p and u using the collision probability of p.

3.3 Algorithm for estimating collision probability

given p, n

- take a sample of s elements from p: $x_1, x_2 \dots x_S$
- for each $1 \le i < j \le n$: $\sigma_{i,j} \leftarrow \begin{cases} 1 & if\ x_i = x_j \\ 0 & otherwise \end{cases}$
- output $\dfrac{\sum_{i<j}\sigma_{i,j}}{\binom{S}{2}}$

3.4 Chebyshev's inequality

Let A be a random variable then $Pr[A - E[A] > \rho] \le \frac{Var(A)}{\rho^2}$

Corollary 6. For multiplicative estimation: $Pr\left[\left| \frac{1}{\binom{S}{2}}\sum_{i<j}\sigma_{i,j} - (\| p \|_2)^2 \right| \ge \varepsilon\cdot(\| p \|_2)^2 \right] \le \dfrac{Var\left(\frac{1}{\binom{S}{2}}\sum_{i<j}\sigma_{i,j}\right)}{\varepsilon^2(\| p \|_2)^4}$

Corollary 7. For additive estimation: $Pr\left[\left| \frac{1}{\binom{S}{2}}\sum_{i<j}\sigma_{i,j} - (\| p \|_2)^2 \right| \ge \varepsilon \right] \le \dfrac{Var\left(\frac{1}{\binom{S}{2}}\sum_{i<j}\sigma_{i,j}\right)}{\varepsilon^2}$

3.5 Main Theorem

Theorem 8. We can estimate $(\| p \|_2)^2$to within multi factor of $1+\varepsilon$ using $\bigcirc\left(\dfrac{\sqrt{n}}{\varepsilon^2}\right)$ samples (with probability $> \dfrac{3}{4}$).

Claim 9. $\dfrac{Var\left(\frac{1}{\binom{S}{2}}\sum_{i<j}\sigma_{i,j}\right)}{\varepsilon^2(\| p \|_2)^4} \le \dfrac{2}{\varepsilon^2\cdot\binom{S}{2}^{\frac{1}{2}}\cdot\left((\| p \|_2)^2\right)^{\frac{1}{2}}}$

Definition 10. $\bar{\sigma}_{i,j} = \sigma_{i,j} - E[\sigma_{i,j}]$

Fact 11. $E[\bar{\sigma}_{i,j}] \overset{by\ linearity\ of\ E}{=} E[\sigma_{i,j} - E[\sigma_{i,j}]]\,\sigma_{i,j} = E[\sigma_{i,j}] - E[E[\sigma_{i,j}]] = 0$

Fact 12. $E[\bar{\sigma}_{i,j}\cdot\bar{\sigma}_{k,l}] \le E[\sigma_{i,j}\cdot\sigma_{k,l}]$

explanation: $\bar{\sigma}_{a,b} = \sigma_{a,b} - E[\sigma_{a,b}] \Rightarrow \bar{\sigma}_{a,b} \le \sigma_{a,b} \Rightarrow \bar{\sigma}_{i,j}\cdot\bar{\sigma}_{k,l} \le \sigma_{i,j}\cdot\sigma_{k,l} \overset{by\ property\ of\ E}{\Rightarrow}$

$E[\bar{\sigma}_{i,j} \cdot \bar{\sigma}_{k,l}] \leq E[\sigma_{i,j} \cdot \sigma_{k,l}]$

Fact 13. $(\sum p(x)^3)^{\frac{1}{3}} \leq (\sum p(x)^2)^{\frac{1}{2}} \Rightarrow \sum p(x)^3 \leq (\sum p(x)^2)^{\frac{3}{2}}$

Fact 14. $s^2 < 3\binom{S}{2} \Rightarrow s^3 < 3^{\frac{3}{2}} \cdot \binom{S}{2}^{\frac{3}{2}}$

Fact 15. $\binom{S}{3} \leq \frac{S^3}{6}$

Proof. $Var\left(\sum_{i<j}\sigma_{i,j}\right) = \underbrace{E\left[\left(\sum_{i<j}\sigma_{i,j} - \sum_{i<j}E(\sigma_{i,j})\right)^2\right]}_{def\ of\ Var} = E\left[\left(\sum_{i<j}\underbrace{(\sigma_{i,j} - E(\sigma_{i,j}))}_{\bar{\sigma}_{i,j}}\right)^2\right] = E\left[(\sum_{i<j}\bar{\sigma}_{i,j})^2\right]$

After opening the brackets:

$$= E\left[\sum_{i<j}\bar{\sigma}_{i,j}^2 + \sum_{i<j,k<l}\bar{\sigma}_{i,j}\cdot\bar{\sigma}_{k,l} + \sum_{\substack{i<j\\i<k}}\bar{\sigma}_{i,j}\cdot\bar{\sigma}_{i,k} + \sum_{\substack{i<j\\k<j}}\bar{\sigma}_{i,j}\cdot\bar{\sigma}_{k,j}\right]$$

According to the linearity of E:

$$= \underbrace{\sum_{i<j}E[\bar{\sigma}_{i,j}^2]}_{t_1} + \underbrace{\sum_{i<j,k<l}E[\bar{\sigma}_{i,j}\cdot\bar{\sigma}_{k,l}]}_{t_2} + \underbrace{2\sum_{\substack{i<j\\i<k}}E[\bar{\sigma}_{i,j}\cdot\bar{\sigma}_{i,k}]}_{t_3}$$

Estimating separately an upper bound for each one of the parts:

$$t_1 = \sum_{i<j}E[\bar{\sigma}_{i,j}^2] \overset{fact\ 14}{\leq} \sum_{i<j}E[\sigma_{i,j}^2] = \binom{S}{2}\cdot(\|\ p\ \|_2)^2$$

$$t_2 = \underbrace{\sum_{i<j,k<l}E[\bar{\sigma}_{i,j}\cdot\bar{\sigma}_{k,l}]}_{\sigma_{i,j}\ and\ \sigma_{k,l}\ are\ indepenedet} = \sum_{i<j,k<l}\underbrace{E[\bar{\sigma}_{i,j}]}_{0}\cdot\underbrace{E[\bar{\sigma}_{k,l}]}_{0} \overset{fact\ 13}{=} 0$$

$$t_3 = 2\sum_{\substack{i<j\\i<k}}E[\bar{\sigma}_{i,j}\cdot\bar{\sigma}_{i,k}] \overset{fact\ 14}{\leq} 2\sum_{\substack{i<j\\i<k}}E[\sigma_{i,j}\cdot\sigma_{i,k}] = 2\cdot\underbrace{\sum_{i\in S}\sum_{\substack{j:\ i<j\\k:\ k>j}}}_{\binom{S}{3}}\cdot E[\sigma_{i,j}\cdot\sigma_{i,l}]$$

When analyzing the expression $E[\sigma_{i,j}\cdot\sigma_{i,l}]$:

$$\sigma_{i,j}\cdot\sigma_{i,l} = \begin{cases}1 & if\ x_i = x_j = x_l\\ 0 & otherwise\end{cases}.$$

We can see it is in fact the same as the 3-way collision probability: $\sum_{x\in[n]}p(x)^3$

Continuing estimating an upper bound for t_3 we get:

$$t_3 \leq 2\cdot\binom{S}{3}\cdot\left(\sum_{x\in[n]}p(x)^3\right) \overset{facts\ 10,12}{\leq} 2\cdot\frac{S^3}{6}\cdot(\sum p(x)^2)^{\frac{3}{2}} \overset{fact\ 11}{\leq} \frac{3^{\frac{3}{2}}\cdot\binom{S}{2}^{\frac{3}{2}}}{3}\cdot\left((\|\ p\ \|_2)^2\right)^{\frac{3}{2}} =$$

$$\sqrt{3}\cdot\left(\binom{S}{2}\cdot(\|\ p\ \|_2)^2\right)^{\frac{3}{2}}$$

Now we can sum the upper bounds and get an upper bound for $Var\left(\sum_{i<j}\sigma_{i,j}\right)$:

$$t_1 + t_2 + t_3 \leq \binom{S}{2}\cdot(\|\ p\ \|_2)^2 + \sqrt{3}\cdot\left(\binom{S}{2}\cdot(\|\ p\ \|_2)^2\right)^{\frac{3}{2}} \leq 2\cdot\left(\binom{S}{2}\cdot(\|\ p\ \|_2)^2\right)^{\frac{3}{2}}$$

Using the above inequality and the property of Var: $Var(aX) = a^2 Var(X)$ we can complete our proof:

$$\frac{Var\left(\frac{1}{\binom{S}{2}}\sum_{i<j}\sigma_{i,j}\right)}{\varepsilon^2(\| p \|_2)^4} \leq \frac{\frac{1}{\binom{S}{2}} \cdot 2 \cdot \left(\binom{S}{2} \cdot (\| p \|_2)^2\right)^{\frac{3}{2}}}{\varepsilon^2(\| p \|_2)^4} = \frac{2}{\varepsilon^2 \binom{S}{2}^{\frac{1}{2}} ((\| p \|_2))}$$

□

Corollary 16. Using $s = \Theta\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$ samples we can estimate $(\| p \|_2)^2$ to within multiplicative factor of $1+\varepsilon$ with probability $> \frac{3}{4} \iff Pr\left[|\ \frac{1}{\binom{S}{2}}\sum_{i<j}\sigma_{i,j} - (\| p \|_2)^2\ | \geq \varepsilon \cdot (\| p \|_2)^2\right] \leq \frac{1}{4}$

Corollary 17. Using $s = \Theta\left(\frac{1}{\varepsilon^2}\right)$ samples we can estimate $(\| p \|_2)^2$ to within additive factor of ε with probability $> \frac{3}{4} \iff Pr\left[|\ \frac{1}{\binom{S}{2}}\sum_{i<j}\sigma_{i,j} - (\| p \|_2)^2\ | \geq \varepsilon\right] \leq \frac{1}{4}$

3.6 Estimating L_1distance

given p, n

- use the collision probability estimation algorithm to estimate $(\| p \|_2)^2$
- If $(\| p \|_2)^2 > \frac{1+\frac{\varepsilon^2}{3}}{n}$ REJECT
- ACCEPT

Proof. 1. if $p = u$ then $(\| p \|_2)^2 = \frac{1}{n}$ and with probability $\geq \frac{3}{4}$ we will estimate $(\| p \|_2)^2$ s.t $\leq \frac{1}{n}\left(1+\frac{\varepsilon^2}{3}\right)$

2. Assume that $|\ p-u\ |_1 > \varepsilon$ but p is likely to pass, then usually $(\| p \|_2)^2 \leq \frac{1+\frac{\varepsilon^2}{3}}{n}$ (because we assumed that p is likely to pass) $\Rightarrow (\| p \|_2)^2 \underbrace{\leq}_{\substack{successfull \\ approximation}} \frac{1}{n} \cdot \left(1+\frac{\varepsilon^2}{3}\right)\left(1+\frac{\varepsilon^2}{3}\right) < \frac{1}{n}\left(1+\varepsilon^2\right)$

   Now we can bound $\| p-u \|_2$: $(\| p-u \|_2)^2 = (\| p \|_2)^2 - \frac{1}{n} < \frac{1}{n}\left(1+\varepsilon^2\right) - \frac{1}{n} < \frac{\varepsilon^2}{n} \Rightarrow \| p-u \|_2 < \frac{\varepsilon}{\sqrt{n}}$
   And this leads to a contradiction to our assumption that $|\ p-u\ |_1 > \varepsilon$ because
   $|\ p-u\ |_1 < \sqrt{n}\ \| p-u \|_2 < \sqrt{n}\frac{\varepsilon}{\sqrt{n}} = \varepsilon$

□