0368.416701 Sublinear Time Algorithms November 10, 2014

Lecture 3

Lecturer: Ronitt Rubinfeld *Scribe: Or Zamir, Orr Fischer, Amir Gilad*

0.1 Notions

- A deterministic algorithm $\mathcal{A}$ is called *symmetrical* if $\mathcal{A}(x) = \mathcal{A}(\pi(x))$ for any permutation π.
- A randomized PASS/FAIL algorithm $\mathcal{A}$ is called *symmetrical* if $Pr[\ \mathcal{A}\ \ passes\ \ x\ ] = Pr[\ \mathcal{A}\ \ passes\ \ \pi(x)\ ]$ for any permutation π.
- We denote by U_D the uniform distribution over a domain D.

0.2 Today's lecture

The main subject of today's lecture is testing properties of distributions. More specifically,

- Examples of cases where uniformity testing was used.
- Formal definition of our goal (A uniformity tester).
- Uniformity tester for ℓ_2 distance.
- Modification of the previous uniformity tester for ℓ_1 distance.
- Comparing unknown distributions.

1 Examples of uniformity testing

Uniformity testing was used in order to estimate whether or not the lottery results of the past years (of some different lottery types) are actually uniform.

1.1 New Jersey "Pick 3" Lottery

In that lottery you pick three digits (from $\{0, ..., 9\}$) and if they were the same as the digits that were randomly selected by the lottery you win. If it was uniform the probability of winning was $\frac{1}{10^3} = \frac{1}{1000}$. The *Chi-squared test* on actual data of about twenty years gave low confidence in uniformity despite the fact that the data was probably uniform, the reason for that is probably the relatively low number of samples. That gives us a motivation for looking for *sub-linear uniformity testing algorithm.*

1.2 Multilotek

On that lottery in order to randomly choose the winning numbers, balls with numbers written on them were pulled out of a machine. It turned out that because of size or shape differences, some balls had a lower probability of being selected. The uniformity of such machine could have been estimated using a *uniformity testing algorithm.*

2 Our Goal

Our goal is to construct for any given $\epsilon > 0$ a *uniformity testing algorithm* $\mathcal{A}$ that gets as an input a black-box that can take samples out of an unknown distribution p over the domain $D = [n] = \{1, ..., n\}$ and

1. If $p = U_{[n]}$ then $\mathcal{A}(p) = PASS$ with probability $\geq \frac{3}{4}$.
2. If $distance\left(p, U_{[n]}\right) > \epsilon$ then $\mathcal{A}(p) = FAIL$ with probability $\geq \frac{3}{4}$.
3. Otherwise, either $PASS$ or $FAIL$ may be returned.

The distance functions we will consider are:

- ℓ_2 distance: $distance(p, q) = ||p - q||_2 = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$.
- ℓ_1 distance: $distance(p, q) = ||p - q||_1 = \sum_{i=1}^{n} |p_i - q_i|$.

We note that it is known that $||p - q||_2 \leq ||p - q||_1 \leq \sqrt{n} \cdot ||p - q||_2$.

3 Uniformity testing in ℓ_2

3.1 Collision Probability

We have already noticed that

$$||p - U_{[n]}||_2^2 = \sum_{i=1}^{n} (p_i - 1/n)^2 = \sum_{i=1}^{n} (p_i^2 - 2 \cdot \frac{1}{n} \cdot p_i + 1/n^2) = \sum_{i=1}^{n} p_i^2 - 2 \cdot \frac{1}{n} \sum_{i=1}^{n} p_i + n \cdot \frac{1}{n^2} = \sum_{i=1}^{n} p_i^2 - \frac{1}{n}$$

As the last term may be written as $||p||_2^2 - ||U_{[n]}||_2^2$, it follows that for $p = U = U_{[n]}$, $||p||_2^2 = n \cdot (\frac{1}{n})^2 = \frac{1}{n}$, and for any other distribution $p \neq U$, $||p||_2^2 > \frac{1}{n}$.

The expression $\sum_{i=1}^{n} p_i^2$ is exactly the *collision probability*, which is the probability that two independent samples s_1, s_2 from p would collide, the easy proof follows

$$Pr\left[s_1 = s_2\right] = \sum_{i=1}^{n} Pr\left[s_1 = i\right] \cdot Pr\left[s_2 = i\right] = \sum_{i=1}^{n} p_i^2$$

Thus, in order to estimate $||p - U_{[n]}||_2^2$, it is enough to estimate the *collision probability*. That leads to the following general algorithm outline.

(Interesting note concerning collision probability: $\sum_{i=1}^{n} p_i^2 \leq p_{\max} \cdot \sum_{i=1}^{n} p_i = p_{\max}$)

3.2 Algorithm

1. Take $s(n)$ samples from p. *(How many?)*
2. Let $\hat{c} \leftarrow$ estimate of $||p||_2^2$ (the collision probability) from the samples. *(How?)*
3. If $\hat{c} < \frac{1}{n} + \delta$ then $PASS$, else $FAIL$. *(Which δ should we use?)*

3.3 How well can we estimate $||p||_2^2$?

If $||p - U||_2^2 > \epsilon^2$ then by previous arguments $||p||_2^2 > \frac{1}{n} + \epsilon^2$.
Let $\delta < \frac{\epsilon^2}{2}$ (e.g $\delta = \frac{\epsilon^2}{4}$) and pick $s(n)$ such that $|\hat{c} - ||p||_2^2| < \delta$ with high probability ($\geq \frac{3}{4}$). Then

- If $p = U$ then $||p||_2^2 = \frac{1}{n}$ and $|\hat{c} - \frac{1}{n}| < \delta$, thus, $\hat{c} < \frac{1}{n} + \delta$ with high probability.
- If $||p - U_{[n]}||_2^2 \geq \epsilon^2$ then $\hat{c} \geq \frac{1}{n} + \epsilon^2 - \delta = \frac{1}{n} + \frac{3}{4}\epsilon^2 > \frac{1}{n} + \delta$ with high probability.

All that is left to do is to show how can we choose such $s(n)$.

3.4 How well can we estimate $||p - U_{[n]}||_1$?

- If $||p-U||_1 = 0 \iff ||p-U_{[n]}||_2^2 = 0$ then $||p||_2^2 = \frac{1}{n}$ and we need to PASS.
- If $||p-U||_1 > \epsilon$ then $||p-U_{[n]}||_2 > \frac{\epsilon}{\sqrt{n}} \Rightarrow ||p-U_{[n]}||_2^2 > \frac{\epsilon^2}{n}$ so $||p||_2^2 > \frac{1}{n} + \frac{\epsilon^2}{n}$. We need a better estimation to handle this case ($\delta < \frac{\epsilon^2}{2n}$)

3.5 Estimation via recycling

- Take s samples from p: $x_1, x_2, ..., x_s$
- For each $1 \leq i \leq j \leq s$: $\sigma_{ij} = \begin{cases} 1 & \text{if } x_i = x_j \\ 0 & \text{if } x_i \neq x_j \end{cases}$
- Output $\hat{c} \leftarrow \frac{\sum \sigma_{ij}}{\binom{s}{2}}$

We now have $E[\hat{c}] = \frac{1}{\binom{s}{2}} \sum_{i<j} E[\sigma_{ij}] = ||p||_2^2$ where $E[\sigma_{ij}] = Pr[\sigma_{ij} = 1]$

3.6 Reminder: Chebyshev's inequality

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

Now, to find a sufficient sample size s, we'll want to bound $Var(\hat{c})$, beacuse if we do, $Pr[|\hat{c} - ||p||_2^2| > \frac{\epsilon^2}{4}] \leq \frac{16 \cdot Var(\hat{c})}{\epsilon^2}$. but to do that, we'll first need to prove a lemma.

3.7 Bounding the variance of $\hat{c}$

Lemma 1 $Var(\sum_{i,j} \sigma_{ij}) \leq 4(\binom{s}{2}||p||_2^2)^{\frac{3}{2}}$

Proof Denote $\bar{\sigma}_{ij} = \sigma_{ij} - E[\sigma_{ij}]$. We can easily see that each of following facts hold:

1. $E[\bar{\sigma}_{ij}] = 0$
2. $E[\bar{\sigma}_{ij}\bar{\sigma}_{kl}] \leq E[\sigma_{ij}\sigma_{kl}]$ (beacuse $\forall_{a,b} \bar{\sigma}_{ab} \leq \sigma_{ab}$)
3. $(\sum(p(x)^3))^{\frac{1}{3}} \leq (\sum(p(x)^2))^{\frac{1}{2}}$ for a probability vector p.
4. $s \leq \sqrt{3\binom{s}{2}}$ and $\binom{s}{2} \leq \frac{s^3}{6}$
5. For two independant variables X,Y $E[X \cdot Y] = E[X] \cdot E[Y]$.
6. We can see that $(\sigma_{ij} \cdot \sigma_{jk})$ is equivalent to event that x_i, x_j, x_k are all with the same value. So $E[\sigma_{ij} \cdot \sigma_{jk}] = \sum_{x \in D}(p(x))^3$

$$Var(\sum_{i<j}(\sigma_{ij})) = E[(\sum_{i<j}\sigma_{ij} - E[\sum_{i<j}\sigma_{ij}])^2] = E[\sum_{i<j}\bar{\sigma}_{ij}^2] =$$

$$E[\sum_{i<j}\bar{\sigma}_{ij}^2 + \sum_{i\neq k, j\neq l}\bar{\sigma}_{ij}\bar{\sigma}_{kl} + \sum_{j\neq k, i<j,l}\bar{\sigma}_{ij}\bar{\sigma}_{il} + \sum_{i\neq k, i<j<k}\bar{\sigma}_{ij}\bar{\sigma}_{jk}] =$$

$$\sum_{i<j}E[\bar{\sigma}_{ij}^2] + \sum_{i\neq k, j\neq l}E[\bar{\sigma}_{ij}\bar{\sigma}_{kl}] + \sum_{j\neq k, i<j,l}E[\bar{\sigma}_{ij}\bar{\sigma}_{il}] + \sum_{i\neq k, i<j<k}E[\bar{\sigma}_{ij}\bar{\sigma}_{jk}]$$

Lets bound each of the four parts of the last equation seperatly.

i $E[\sum_{i<j}\bar{\sigma}_{ij}^2] \leq_{(2)} E[\sum \sigma_{ij}^2] = \binom{s}{2}\cdot ||p||_2^2$

ii $E[\sum_{i<j,k<l}\bar{\sigma}_{ij}\bar{\sigma}_{kl} =_{(5)} \sum_{i<j,k<l} E[\bar{\sigma}_{ij}]E[\bar{\sigma}_{kl}] =_{(1)} 0$

iii $E[\sum \bar{\sigma}_{ij}\bar{\sigma}_{il}] \leq \sum_{i<j<l} E[\sigma_{ij}\sigma_{il}] =_{(6)} \binom{s}{3}\sum(p(x)^3) \leq_{(3)} 2\cdot\frac{s^3}{6}\sum_{x\in D}(p(x)^2)^{\frac{3}{2}} \leq_{(4)} \sqrt{3}\binom{s}{2}^{\frac{3}{2}}(||p||_2^2)^{\frac{3}{2}}$

iv Identical to (iii).

This gives us that $Var(\sum_{i,j}\sigma_{ij}) \leq \binom{s}{2}||p||_2^2 + 0 + 2\cdot\sqrt{3}\binom{s}{2}^{\frac{3}{2}}(||p||_2^2)^{\frac{3}{2}} \leq 4(\binom{s}{2}||p||_2^2)^{\frac{3}{2}}$
■

Fact 2 $\forall_{\alpha\in R} Var(aX) = a^2 Var(X)$

So $Var(\hat{c}) = Var(\frac{1}{\binom{s}{2}}\cdot\sum\sigma_{ij}) = \frac{1}{\binom{s}{2}^2}\cdot Var(\sum\sigma_{ij})$
Now lets bound the distance between $\hat{c}$ and $||p||_2^2$.
$E[\hat{c}] = \frac{1}{\binom{s}{2}}\sum_{i<j}E[\sigma_{ij}] = \frac{1}{\binom{s}{2}}\sum_{i<j}Pr[\sigma_{ij}=1] = \frac{\binom{s}{2}}{\binom{s}{2}}||p||_2^2 = ||p||_2^2$
$Pr[|\hat{c} - ||p||_2^2| > \frac{\epsilon^2}{4}] \leq_{chebishev} \frac{Var[\hat{c}]}{\epsilon^4}\cdot 4^2 = 64\frac{(\binom{s}{2}||p||_2^2)^{\frac{3}{2}}}{\binom{s}{2}^2\epsilon^4} = O(\frac{1}{\epsilon^4\cdot s})$ (as $||p||_2^3 \leq 1$).
So for $Pr[|\hat{c} - ||p||_2^2| \leq \frac{1}{4}$ we'll need to pick $s \geq \Omega(\frac{1}{\epsilon^4})$.

Conclusion 3 *To test uniformity in l_2 norm, it is sufficient to take sample size $s \geq \Omega(\frac{1}{\epsilon^4})$.*

From this we can get result for l_1 norm, using the $||x||_1 \leq \sqrt{n}||x||_2$ but it will cost us a $\sqrt{n}$ factor.

Corollary 4 *To test uniformity in l_1 norm, it is sufficient to take sample size $s \geq \Omega(\frac{1}{\epsilon^4}\sqrt{n})$.*

4 Difference in distributions

We now have two distributions, p and q and we want our algorithm to behave like this:

- If $p = q$ output PASS.
- If $||p-q||_2^2 > \epsilon^2$ or $||p-q||_1 > \epsilon$ output FAIL.

Again, notice: $||p-q||_2^2 = \sum_i p_i^2 - 2\sum_i p_i q_i + \sum_i q_i^2$ so the variance bound depends on the maximum probability element of q.

4.1 Filter Distribution

1. Learn $\mathcal{B}$ = domain elements with probability $> b$. (We need $\mathcal{O}(\frac{1}{b}\log(\frac{1}{b}))$ samples)
2. Filter the samples:
   - If $\in \mathcal{B}$, use the naive method to estimate $\sum_{i\in B}|p_i - q_i|$ ($\mathcal{O}(\frac{1}{b})$ samples)
   - If $\notin \mathcal{B}$, use the collisions method to distinguish $p_{\overline{\mathcal{B}}} = q_{\overline{\mathcal{B}}}$ from $|p_{\overline{\mathcal{B}}} - q_{\overline{\mathcal{B}}}| > \epsilon$ (small elements, depends on b)

Picking $b = \Theta(\frac{1}{n^{\frac{2}{3}}})$ optimizes the result.