COMS E6998-9: Algorithms for Massive Data (Spring'19) March 12, 2019

Lecture Lecture 15: Distribution testing: uniformity, identity

Instructor: *Alex Andoni* Scribes: *Collin Burns, Garrison Grogan*

1 Review

Recall that last time we considered the problem of *uniformity testing*: Given a distribution D over $[n]$, we want to distinguish between D being uniform and D being ϵ-far from uniform. In particular, we saw that $\|D\|_2^2 = \frac{1}{n}$ when D is uniform and $\|D\|_2^2 > \frac{1}{n} + \frac{\epsilon^2}{n}$ if D is ϵ-far from uniform. Given m samples from D, we hope to be able to do distinguish between these cases using a number of samples that is sublinear in n.

We saw before that the algorithm for uniformity testing works as follows: first compute $C := \#$collisions. If $\frac{C}{M} < \frac{1}{n} + \frac{\epsilon^2}{2n}$ output "uniform"; otherwise, output "ϵ-far from uniform".

Throughout this lecture, we let $d = \|D\|_2^2$ for simplicity of notation. We saw last time that $\mathbb{E}[\frac{C}{M}] = d$, where $M := \binom{m}{2}$.

2 Uniformity Testing Continued

We want to argue that $\frac{C}{m}$ concentrates around its mean by bounding its variance and applying Chebyshev.

Claim 1. $Var(\frac{C}{M}) \leq O(\frac{d^{3/2}}{m})$.

Proof. First, note that while C is a sum of indicator random variables, those random variables are not independent, so we must compute the variance explicitly. We have:

$$\mathbb{E}\left[\left(\frac{C}{M}\right)^2\right] = \frac{1}{M^2}\mathbb{E}\left[\left(\sum_{i<j} \chi\left[x_i \;\&\; x_j \text{ collide}\right]\right)^2\right]$$

Using the notation $\chi_{i,j} = \chi\left[x_i \;\&\; x_j \text{ collide}\right]$ and expanding the quadratic yields:

$$= \frac{1}{M^2} \sum_{\substack{i<j,k<l \\ \{i,j\}\cap\{k,l\}=\emptyset}} \mathbb{E}[\chi_{i,j}\chi_{k,l}] + \frac{1}{M^2} \sum_{\substack{i<j,k<l \\ \{i,j\}\cap\{k,l\}\neq\emptyset}} \mathbb{E}[\chi_{i,j}\chi_{k,l}]$$

$$\leq \frac{1}{M^2}M^2d^2 + \frac{O(1)}{M^2}\sum_{i<jk} \mathbb{E}\left[\chi\left[x_i \;\&\; x_j \;\&\; x_k \text{ collide}\right]\right] = d^2 + O\left(\frac{1}{M^2}\right)\sum_{i<jk} Pr\left[x_i \;\&\; x_j \;\&\; x_k \text{ collide}\right]$$

But we have that if i, j, and k are distinct:

$$Pr\left[x_i \;\&\; x_j \;\&\; x_k \text{ collide}\right] = \sum_{z\in[n]} D_z^3 = \|D\|_3^3$$

Moreover, there are $O(m^3)$ such choices over i, j, and k, so we have

$$\mathbb{E}\left[\left(\frac{C}{M}\right)^2\right] \leq d^2 + O\left(\frac{m^3}{M^2}\right)\|D\|_3^3 \leq d^2 + O\left(\frac{1}{m}\right)\|D\|_2^3 = d^2 + O\left(\frac{1}{m}\right)d^{3/2}$$

where we used that $M = \theta(m^2)$. Combining this with $\mathbb{E}[\frac{C}{M}] = d$ implies that:

$$Var\left(\frac{C}{M}\right) = \mathbb{E}\left[\left(\frac{C}{m}\right)^2\right] - d^2 = O\left(\frac{d^{3/2}}{m}\right)$$

as desired. □

Now we will analyze the algorithm. The first case is when D is indeed uniform. Then as we saw, $d = \frac{1}{n}$. By Chebyshev, this implies that with probability 0.9, $\frac{C}{M} \leq d + \sqrt{10 \cdot Var(C/m)} \leq \frac{1}{n} + O\left(\sqrt{\frac{d^{3/2}}{M}}\right)$. We want $O\left(\sqrt{\frac{d^{3/2}}{M}}\right) \leq \frac{\epsilon^2}{2n}$. Since $M = \theta(m^2)$ and $d = \frac{1}{n}$, this is equivalent to $\frac{1}{n^{3/2}m} \leq O(\frac{\epsilon^2}{2n})^2$, for which it suffices that $M = O\left(\frac{\sqrt{n}}{\epsilon^4}\right)$.

The second case is when D is ϵ-far from uniform. We have that with probability 0.9,

$$\frac{C}{M} \geq d - \sqrt{10 \cdot Var(C/M)} \geq d - O\left(\sqrt{\frac{d^{3/2}}{m}}\right) \geq d\left(1 - O\left(\frac{1}{d^{1/4}m^{1/2}}\right)\right)$$

Recall that if D is ϵ-far fro m uniform, $d \geq \frac{1}{n} + \frac{\epsilon^2}{n}$. Moreover, note that if $O\left(\frac{1}{d^{1/4}m^{1/2}}\right) \leq \frac{\epsilon^2}{10}$, then we have:

$$\frac{C}{M} \geq \frac{1}{n}\left(1+\epsilon^2\right)\left(1 - \frac{\epsilon^2}{10}\right) \geq \frac{1}{n} + \frac{9}{10}\frac{\epsilon^2}{n} - \frac{\epsilon^4}{10n} \geq \frac{1}{n} + \frac{\epsilon^2}{2n}$$

where the final inequality holds as long as $\epsilon < \frac{1}{4}$. Moreover, if this holds, then our algorithm correctly outputs "ϵ-far from uniform". To have $O\left(\frac{1}{d^{1/4}m^{1/2}}\right) \leq \frac{\epsilon^2}{10}$, it suffices that $m = O(\frac{1}{\epsilon^4}\frac{1}{\sqrt{d}})$. Since $d \geq \frac{1}{n}$, it thus suffices to have $m = O\left(\sqrt{n}\epsilon^4\right)$. As a final remark, this $\sqrt{n}$ factor is tight, but there are algorithms that improve on the $\frac{1}{\epsilon^4}$ factor.

3 Identity Testing/Closeness to a fixed distribution

Say we fix Q to some distribution on a known $[n]$,

the problem is now given $x_1, ..., x_m \sim D$, we want to distinguish between $D = Q$ and $||D - Q||_1 \geq \epsilon$ i.e D is ϵ-far from Q.

Consider the following idea: We will reduce testing D vs Q to testing $\underbrace{D' \text{ vs } Q'}_{\text{on new domain } S}$,

where Q' is $\approx$ uniform and $||D' - Q'||_1 \approx ||D - Q||_1$

How will we do this? Q may not be uniform so we must change space Q is on. We do this by relabeling Q's values into a new space, such that repeated values on Q' become unique on this new space thus the new distribution is uniform. We do the same scheme for D to get D' on the same space.

For example if there are 5 1s but 2 2s on Q, then we relabel the 1s to $(1,1),(1,2),...$ and the 2s to $(2,1),(2,2)$, thus we get a new uniform distribution in our new domain.

Formally, start by fixing $\delta = c\epsilon$ for small $c > 0$.

Now for each $i \in [n]$ let $\mathscr{S}_i = \lceil \frac{nQ_i}{\delta} \rceil$ i.e how many times we replicate coordinate i. δ acts as a control parameter.

Note, that if $\delta = 1$, then if Q is already uniform, nothing happens.

Our new domain is then the union of all our coordinate changes,

$$S = \bigcup_{i=1}^{n} i \times [\mathscr{S}_i] = \{(1,1),(1,2),...,(1,\mathscr{S}_1),(2,1),...,(2,\mathscr{S}_2),(3,1),...\} \tag{1}$$

and our new distributions are given as

$$Q'_{ij} = \frac{Q_i}{\mathscr{S}_i} \text{ and } D'_{ij} = \frac{D_i}{\mathscr{S}_i}$$

note that D' also depends on Q.

Now we several claims using our new distributions:

Claim 1: For $\delta = c\epsilon$ and small $c > 0$, say $c < 1/10$, $||Q - U_S||_1 < O(\delta) < \epsilon/10$

Proof.

$$Q'_{i,j} = Q_i \frac{1}{\mathscr{S}_i} = Q_i \frac{\delta}{nQ_i} = \delta/n \text{ and}$$

$$|S| = \sum_i \mathscr{S}_i \leq \sum_i \frac{nQ_i}{\delta} + 1 = \frac{n}{\delta}\sum_i^n Q_i + n = n(\frac{1}{\delta} + 1) \implies$$

$$||Q' - U_S||_1 = \sum_{i\in[n]} \sum_{j=1}^{\mathscr{S}_i} |Q'_{ij} - \frac{1}{n(1/\delta+1)}| \leq \sum i \in [n] \sum_{j=1}^{\mathscr{S}_i} \frac{\delta}{n}|1 - \frac{1}{\delta+1}|$$

$$\leq \sum_{i\in[n]} \sum_{j=1}^{\mathscr{S}_i} \frac{\delta}{n}|1 - (1 - \delta * 2\delta^2)| \leq \sum_{i\in[n]} \sum_{j=1}^{\mathscr{S}_i} \frac{\delta}{n}\delta(1+2\delta)$$

$$\leq n(\frac{1}{\delta}+1)\frac{\delta^2(1+2\delta)}{n} = \delta(1+\delta)(1+2\delta) = O(\delta)$$

□

Claim 2: $||D' - Q'|| = ||D - Q||_1$

Proof.

$$||D' - Q'||_1 = \sum\sum |D'_{ij} - Q'_{ij}| = \sum_{i\in[n]} \sum_{j=1}^{\mathscr{S}_i} |\frac{D_i}{\mathscr{S}_i} - \frac{Q_i}{\mathscr{S}_i}| = \sum\sum \frac{1}{\mathscr{S}_i}|D_i - Q_i| = \sum_{i\in[n]} |D_i - Q_i| = ||D - Q||_1$$

□

Thus testing if D is close to $Q \iff$ Testing if D' is uniform over S, say up to $\pm\epsilon/10$ or whatever $c\epsilon$ is used.

Now we need to also show that we can emulate $x'_i \sim D'$ using $x_i \sim D$

Claim 3: We can emulate $x'_i \sim D$ using $x_i \sim D$

Proof.

$$\text{take } x_i \sim D \text{ say } j = x_i \text{ and set}$$

$$x'_i = (j,k), k \in_r \{1,2,...,\mathscr{S}_j\} \implies \Pr[x'_i = (j,k)] = D_j \frac{1}{S_j} = D'_{jk}$$

□

Theorem 2. *We can test identity using* $m = O(\frac{1}{\epsilon^4}\sqrt{|S|}) = O(\frac{1}{\epsilon^4}|S|\sqrt{n})$ *samples*

There are better algorithms for nicer distributions of Q which are not arbitrary, for example when Q is a point distribution or some other fixed distribution (eg, normal distribution). In the first case, $O(1/\epsilon)$ samples are enough! So instance optimal algorithms give a better dependence on m if Q is nice. An example of such tester is the [Valiant - Valiant '14] algorithm, which has the following test:

$\sum_i \frac{(m\hat{D}_i - mQ_i)^2 - m\hat{D}_i}{D_i^{2/3}}$ vs α

where $\hat{D}_i = \frac{\#occ \text{ of } i}{m}$. This can be contrasted to the classic χ^2-test [Pearson' 1900]:

$\sum_i \frac{m\hat{D}_i - mQ_i)^2 - mQ_i}{Q_i}$ vs α

which has much worse complexity.

4 Other Distribution Testing Problems

There are many other distribution testing problems.

• Closeness: Testing where D and Q are the same, or ϵ-far from each other (in ℓ_1 or total variation distance), when *both* D and Q are unknown. $m = \theta_\epsilon(n^{2/3})$ is optimal, which is only a little worse.

• Independence: Given $(x_i, y_i) \sim D$, where $D = (D_1, D_2)$ is over $[n_1] \times [n_2]$, distinguish between when D_1 and D_2 are independent (i.e., $D = D_1 \times D_2$) and when D is ϵ-far form *any* product distribution.

• Robustness: Solving distribution testing problems when a subset of samples are noisy or adversarial. This is a much more recent subarea, with these sorts of problems mainly starting to be proposed in just the last few years.

5 Next Time

We are finished with distribution testing in this clas, and will continue with property testing, and sublinear algorithms more generally. Next time, we'll see *monotonicity testing*: given a string of n integers, determine if that string is monotone (i.e., sorted), or ϵ-far from monotone. Specifically, we want to distinguish between these cases using the smallest number of queries to the list as possible. The notion of distance from monotonicity is the following:

Definition 3. *A string x of length n is ϵ-far from being sorted if it is necessary to delete ϵn items to get a sorted list.*

We will see that $O(1/\epsilon \cdot \log n)$ queries suffices.