

Lecture 2

- Element Distinctness
- Testing Probability Distributions
 - uniformity

Element Distinctness

Input $x_1, \dots, x_n, \epsilon$

Question are ~~x_i~~ distinct?

if all x_i distinct, output PASS

if # distinct $x_i < (1-\epsilon)n$ output FAIL ← with prob $> 3/4$

Algorithm

- Take $\frac{1}{\epsilon}$ independent samples
- If there is a duplicate, output "FAIL"
else output "PASS"

Behavior

- If all x_i distinct, will output PASS with prob 1
- What if # distinct $x_i < (1-\epsilon)n$?

example input

$1, 1, 2, 2, 3, 3, \dots, \frac{n}{2}, \frac{n}{2}$ in random order

"Birthday paradox"-like analysis

\Rightarrow need \sqrt{n} samples to see two elements that are the same

Plan:

- pair off duplicates
- argue that $O(\frac{\sqrt{n}}{\epsilon})$ samples likely to hit both members of same pair

New Algorithm:

- $S_1 \leftarrow \sqrt{n}$ samples. If see duplicate, fail + halt
 - $S_2 \leftarrow \frac{c \cdot \sqrt{n}}{\epsilon}$ samples
If see duplicate, fail + halt
else output pass
- same value, different location

New vs Old Algorithm:

both pass distinct lists
if new algorithm fails, then old (with $\frac{c\sqrt{n}}{\epsilon}$ samples) fails with prob $\geq \beta$.

Analysis

first: divide duplicates into disjoint ordered pairs P
then: to show - $O(\frac{\sqrt{n}}{\epsilon})$ samples likely to hit both members of some pair in P
• Phase 1 samples hit first member of lots of pairs in P
• 2nd member of

Pairing the elements in the ^{input} list:

This page has nothing to do with the algorithm or the samples taken by it.

- divide occurrences into \sqrt{n} pairs ^{ordered} } let P be set of pairs
- if odd number, throw out one

Claim if number distinct elts $< (1-\epsilon)n$

then

$$|P| \geq \frac{\epsilon n}{3}$$

examples

1) 7, 1, 2, 1, 3, 4, 2, 1, 2, 4, 4, 3, 5, 3, 6, 4

1, 1, 2, 2, 3, 3, 4, 4, 4, 4, *, *, *

$$|P| = 5$$

2) 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2

$$|P| = 7$$

Pf of Claim

at least ϵn elements have a duplicate

let $I_c \leftarrow \{i \mid x_i = c\}$ set of elts with value c

$D = \bigcup_{c \text{ s.t. } |I_c| \geq 2} I_c$ is set of elts with duplicate

For each c , # paired off elements

is $\begin{cases} |I_c| & \text{if } |I_c| \text{ even} \Rightarrow \text{unpaired} = 0 \\ |I_c| - 1 & \text{if } |I_c| \text{ odd (so } |I_c| \geq 3) \Rightarrow \text{unpaired} = 1 \end{cases}$

$\leq \frac{1}{2} \cdot \epsilon n \leq \frac{1}{2} \epsilon n$

fraction into 1

Let's get back to the algorithm + the samples:

How many ^(ordered) Pairs are hit in first location by S_1 ?

Attempt:

Define $F = \{i \mid i \text{ is first member of an ordered pair + } i \text{ is hit by } S_1\}$

$$\Pr[\text{sample hits first elt of pair}] > \frac{\epsilon}{3}$$

Define indicator variable $y_k = \begin{cases} 1 & \text{if sample } k \text{ hits 1st elt of a pair} \\ 0 & \text{o.w.} \end{cases}$

$$E[y_k] = \Pr[y_k=1] \geq \frac{\epsilon}{3}$$

$$E[\sum_{k \in S_1} y_k] \geq \sqrt{n} \cdot \frac{\epsilon}{3}$$

$$= E[Y]$$

Expected # samples in S_1 that hit 1st location of an ordered pair
 ?

Call this a "sample collision"
 not necessarily equal,
 what if hit the same pair twice?

Expected # ordered pairs hit in 1st location by S_1
 $= E[F]$

will deal with this issue soon

First another issue:

note, y_k 's are independent

$$\text{Chernoff Bnd} \Rightarrow \Pr\left[\sum_{k \in S_1} y_k \geq \frac{E[\sum_{k \in S_1} y_k]}{2}\right]$$

Bad event #1:

$$\sum y_k < \frac{\epsilon \sqrt{n}}{6}$$

$$= \Pr\left[\sum_{k \in S_1} y_k \geq \frac{\epsilon}{6} \sqrt{n}\right]$$

Improvement on attempt:

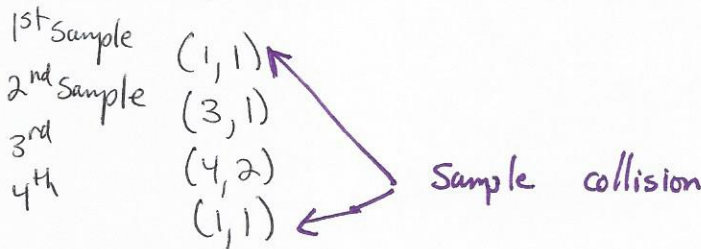
(k, l) -Sample collision: if k^{th} + l^{th} sample hit same input location

$$Z_{kl} = \begin{cases} 1 & \text{if } k^{\text{th}} + l^{\text{th}} \text{ sample are same location in input} \\ 0 & \text{o.w.} \end{cases}$$

ex.

i	1	2	3	4	5	6	7	8	9	10
x_i	1	2	1	2	3	1	2	3	4	5

Sample locations, contents



Sample Collisions $Z = \sum_{k, l} Z_{kl}$

$|F| = \#$ ordered pairs hit in first location

since bad event # didn't happen
 this is $Y \geq \frac{\epsilon \sqrt{n}}{b}$

\geq # samples in S_i that hit 1st member of ordered pair

$- Z$

why? Z is upper bound on # of times that a pair is hit two or more times

Why is Z small?

$$E[Z_{kl}] = \Pr[Z_{kl} = 1] = \frac{1}{n} \quad \text{for } k \neq l$$

$$E[Z] = \binom{\sqrt{n}}{2} \cdot \frac{1}{n} < 1/2$$

$$\text{Markov's } \Rightarrow \Pr[Z > 6] < 1/2$$

Bad event #2

$$Z > 6$$

let's assume it doesn't happen, so $|F| \geq Y$

Lemma with probability $\geq 5/6$
 # ordered pairs

← the prob that neither bad event happens
 hit in 1st locn
 by sample

$$\geq \frac{\epsilon \sqrt{n}}{6} - 6$$

$$\geq \frac{\epsilon \sqrt{n}}{12}$$

Phase 1:

with prob $\geq 5/6$

$$|F| \geq Y - Z$$

$$\geq \frac{\epsilon \sqrt{n}}{6} - 6$$

$$\geq \frac{\epsilon \sqrt{n}}{12}$$

What about Phase 2?

Assume phase 1 "worked", $\geq \frac{\epsilon \sqrt{n}}{12}$ of ordered pairs
 hit in 1st locn by
 the sample

$$\text{i.e. } |F| \geq \frac{\epsilon \sqrt{n}}{12}$$

$$\Pr[\text{one sample hits 2nd member of } i \text{ in } F] \geq \frac{\epsilon \sqrt{n}}{12} \cdot \frac{1}{n} = \frac{\epsilon}{12\sqrt{n}}$$

$$\Pr[\text{no sample in } S_2 \text{ " " " " " some } i \text{ in } F]$$

c. $\sqrt{n} \cdot 12$

for proper

Prob [we don't see both elts of some pair] $\leq \frac{1}{12} + \frac{1}{6} = \frac{1}{4}$

\uparrow failure of phase 2
 \uparrow failure of phase 1
 given phase 1 worked

Thm if elts distinct, alg outputs PASS
 if $\leq (1-\epsilon)n$ elts distinct, $\Pr[\text{alg outputs Fail}] \geq 3/4$

note can improve to $\sqrt{\frac{n}{\epsilon}}$ by setting $|S_1|, |S_2|$ to $\sqrt{\frac{n}{\epsilon}}$

Do we need \sqrt{n} ?

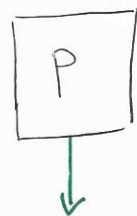
will say more about lower bnds later.

In the meantime, recall example of

$$1, 1, 2, 2, 3, 3, \dots, \frac{n}{2}, \frac{n}{2}$$

Turning to a new model:

Probability distributions - get samples of distribution



Domain D , $|D|=n$

$$p_i = \Pr[p \text{ outputs } i] \leftarrow \text{unknown}$$

Outputs iid samples

← this is all we can learn from

Examples:

Lottery data

Shopping choices

experimental outcomes

⋮

What do we want to know?

is it uniform?

is it high entropy?

large support? (many distinct elements have >0 probability)

is it monotone increasing, k -modal, monotone hazard rate...?

how can we do it?

χ^2 test

plug in estimate

learn distribution, Maximum likelihood estimates

Goal: sample complexity **SUBLINEAR** in n

Testing Uniformity

The goal:

Uniform dist on D

• if $P \equiv U_D$ then tester outputs PASS \leftarrow with prob $\geq 3/4$

• if $\underbrace{\text{dist}(P, U_D)} > \epsilon$ then tester outputs FAIL

which measure of distance?

$l_1, l_2, \text{KL-divergence, Earth mover, Jensen-Shannon}$

$\uparrow \uparrow$
today's focus

Distances

$$l_1\text{-distance} : \|p - q\|_1 = \sum_{i \in D} |p_i - q_i|$$

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

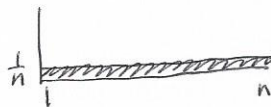
$$\|p - q\|_2 \leq \|p - q\|_1 \leq n^{1/2} \|p - q\|_2$$

examples:

① $p = (1, 0, 0, \dots, 0)$



$q = (\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n})$



l_1 distance:

$$\|p - q\|_1 = \left(\frac{n-1}{n}\right) + (n-1) \cdot \frac{1}{n} \approx 2$$

l_2 distance:

$$\|p - q\|_2^2 = \left(1 - \frac{1}{n}\right)^2 + (n-1) \left(\frac{1}{n}\right)^2 \approx 1$$

②

$p = \left(\frac{2}{n}, \frac{2}{n}, \dots, \frac{2}{n}, 0, 0, \dots, 0\right)$



$q = \left(0, 0, \dots, 0, \frac{2}{n}, \frac{2}{n}, \dots, \frac{2}{n}\right)$



l_1 distance:

$$\|p - q\|_1 = n \cdot \left(\frac{2}{n}\right) = 2$$

$$\|p - q\|_2^2 = n \cdot \left(\frac{2}{n}\right)^2 = \frac{4}{n}$$

$$\|p - q\|_2 = \frac{2}{\sqrt{n}}$$

"Plug-in" Estimate:

Algorithm:

- take m samples from p
- estimate $p(x) \forall x$ via

$$\hat{p}(x) = \frac{\# \text{ times } x \text{ occurs in sample}}{m}$$

- if $\sum_x |\hat{p}(x) - \frac{1}{n}| > \varepsilon$ reject
else accept.

Analysis: (better analyses exist)

pick m st. $\forall x, |\hat{p}(x) - p(x)| < \frac{\varepsilon}{n} \Rightarrow \|\hat{p} - p\|_1 < \varepsilon$

by $\Delta \#$, if $\|\hat{p} - p\|_1 < \varepsilon + \|\hat{p} - u\|_1 < \varepsilon$

then $\|p - u\|_1 < 2\varepsilon$.

how many samples? $\Omega(\frac{n}{\varepsilon})$ maybe even worse ...

Can we do better?

Let's start with L_2 -Distance:

$$\begin{aligned} \|p - u\|_2^2 &= \sum_{i \in [n]} \left(p_i - \frac{1}{n}\right)^2 \\ &= \sum p_i^2 - \underbrace{\frac{2}{n} \sum p_i}_{=1} + \underbrace{\sum \frac{1}{n^2}}_{=\frac{1}{n}} \\ &= \sum p_i^2 - \frac{1}{n} \end{aligned}$$

Collision probability:

$$\Pr_{s, t \in p} [s = t] = \sum_{i \in [n]} p_i^2 \equiv \|p\|_2^2$$

for $p = u$
 $\|u\|_2^2 = \frac{1}{n}$
for any other p
 $\|p\|_2^2 > \frac{1}{n}$

$$\text{so } \|p - u\|_2^2 = \underbrace{\|p\|_2^2}_{\text{we can estimate this}} - \underbrace{\|u\|_2^2}_{\text{we know this}}$$

Algorithm:

1. take s samples from p
2. estimate $\hat{C} = \|p\|_2^2$ from sample
3. if $\hat{C} < \frac{1}{n} + \delta$ pass,
else fail