Lecture 2

- Element Distinctness
- Testing Probability Distributions
  - uniformity

Element Distinctness

Input $x_1 \cdots x_n$, ε

Question ~~are x_i distinct?~~

if all x_i distinct, output PASS ⟷ with prob $> 3/4$

if # distinct $x_i < (1-\varepsilon)n$ output FAIL

Algorithm

- Take ___ 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 $x_i < (1-\varepsilon)n$?

example input

$1, 1, 2, 2, 3, 3, \ldots, \frac{n}{2}, \frac{n}{2}$ 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 $O(\frac{\sqrt{n}}{\varepsilon})$ samples likely to hit both members of same pair

New Algorithm:

- $S_1 \leftarrow \sqrt{n}$ samples. If see duplicate, fail + halt
- $S_2 \leftarrow \frac{c \cdot \sqrt{n}}{\varepsilon}$ 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, (with prob β) then old (with $\frac{c\sqrt{n}}{\varepsilon}$ samples) fails with prob $\geq \beta$.

Analysis

first: divide duplicates into disjoint **ordered** pairs P

then: to show - $O(\frac{\sqrt{n}}{\varepsilon})$ samples likely to hit **both** members of **some** pair in P

- Phase 1 samples hit **first** member of **lots** of pairs in P
- Phase 2 samples hit **2nd** member of **some pair hit in phase 1**

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

Pairing the elements in the input list:

- divide occurences into ordered pairs; let P be set of pairs
- if odd number, throw out one

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

then

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

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, *, *, *

$|P| = 5$

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

$|P| = 7$

3) 1, 1, 1, 2, 2, 2, 3, 3, 3, ..., k, k, k list size $= 3k = \varepsilon n$, $|P| = k = \varepsilon n/3$

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

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

For each c, # paired off elements is

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

fraction unpaired $\leq \frac{1}{3}$

so fraction unpaired is $\leq \max \frac{1}{|I_c|} \leq \frac{1}{3}$

why?
$\frac{|D|}{n} \geq \varepsilon$
$\frac{\text{paired}}{n} \geq \varepsilon \cdot \frac{2}{3}$ ⇒ # pairs $\geq n \cdot \varepsilon \cdot 2/3 / 2 = \varepsilon n/3$

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

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

Define $F = \{ i \mid i$ is first member of an ordered pair in P & i is hit by $S_1 \}$

Attempt:

$\forall k$ $\Pr[$ sample k hits first elt of pair $] > \frac{\varepsilon}{3}$

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

$E[y_k] = \Pr[y_k = 1]$

$\geq \frac{\varepsilon}{3}$

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

$= E[Y]$

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

$\stackrel{?}{=}$

Expected # ordered pairs in P hit in 1st location by S_1

$= E[F]$

not necessarily equal, what if hit the same pair twice?

Call this a "sample collision"

will deal with this issue soon

First another issue:

note, y_k's are independent

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

$= \Pr[\sum_{k \in S_1} y_k < \frac{\varepsilon}{6}\sqrt{n}]$

$< e^{-(\frac{1}{8} \cdot \frac{\varepsilon}{3}\sqrt{n})} << \frac{1}{12}$

Bad event #1: $\sum_{k \in S_1} y_k < \frac{\varepsilon\sqrt{n}}{6}$

Let's assume it doesn't happen

Improvement on attempt:

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

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

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

1st Sample (1,1)
2nd Sample (3,1)
3rd (4,2)
4th (1,1)

Sample collision

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

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

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

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

$-Z$

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

Why is Z small?

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

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

Markov's $\neq \Rightarrow \Pr[Z > 6] < 1/12$

Bad event #2

$Z > 6$

let's assume it doesn't happen, so $|F| \geq Y - 6$

← the prob that neither bad event happens (#1 or #2)

Lemma with probability $\geq 5/6$
ordered pairs hit in 1st locn by sample

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

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

Phase 1:
with prob $\geq 5/6$
$|F| \geq Y - Z$
$\geq \frac{\varepsilon\sqrt{n}}{6} - 6$
$\geq \frac{\varepsilon\sqrt{n}}{12}$

What about Phase 2?

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

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

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

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

Assuming phase 2 takes $\frac{C \cdot 12}{\varepsilon} \cdot \sqrt{n}$ samples $\longrightarrow$

$$\leq \left(1 - \frac{\varepsilon}{12\sqrt{n}}\right)^{C \cdot \frac{\sqrt{n}}{\varepsilon} \cdot 12} \leq e^{-C} \leq \frac{1}{12}$$

for proper choice of C

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

↑ failure of phase 2 given phase 1 worked

↑ failure of phase 1

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

note can improve to $\sqrt{\lceil \frac{n}{\varepsilon} \rceil}$ by setting $|S_1||S_2|$ to $\sqrt{\frac{n}{\varepsilon}}$

Do we need $\sqrt{n}$?

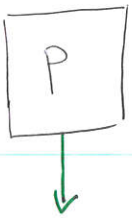
will say more about lower bnds later.

In the meantime, recall example of

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

Turning to a new model:

Probability distributions - get samples of distribution



Domain D, $|D| = n$

$p_i = \Pr[p \text{ outputs } i]$ ← 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 p 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:

- if $P = U_D$ then tester outputs PASS (with prob $\geq 3/4$) ← Uniform dist on D
- if $\text{dist}(P, U_D) > \varepsilon$ then tester outputs FAIL

which measure of distance?

l_1, l_2, KL-divergence, Earth mover, Jensen-Shannon

↕ today's focus

Distances

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

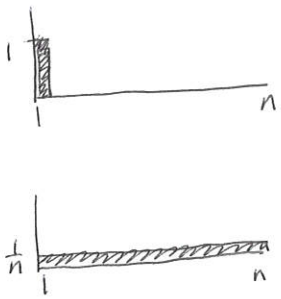
ℓ_2-distance : $\|p-q\|_2 = \sqrt{\sum_{i \in D} (p_i - q_i)^2}$

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

examples:

(1) $p = (1, 0, 0, \ldots 0)$

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



ℓ_1 distance:

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

≈ 2

ℓ_2-distance:

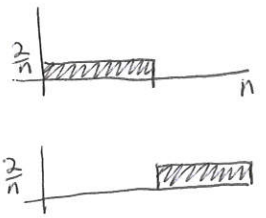
$\|p-q\|_2^2 = \left(1-\frac{1}{n}\right)^2 + (n-1)\left(\frac{1}{n}\right)^2$

≈ 1

(2)

$p = (\frac{2}{n}, \frac{2}{n}, \ldots \frac{2}{n}, 0, 0, \ldots 0)$

$q = (0, 0, \ldots 0, \frac{2}{n}, \frac{2}{n}, \ldots \frac{2}{n})$



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 $\sum_x \left|\hat{p}(x) - \frac{1}{n}\right| > \varepsilon$ reject

else accept.

Analysis: (better analyses exist)

pick m s.t. $\forall x. \left|\hat{p}(x) - p(x)\right| < \frac{\varepsilon}{n} \Rightarrow \|\hat{p} - p\|_1 < \varepsilon$

so, if $p = U_n$ then p passes

by Δ-ineq, if $\|p - p\|_1 < \varepsilon$ & $\|\hat{p} - U\|_1 < \varepsilon$

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

so, if $\|p - U_n\| \geq 2\varepsilon$ this test is likely to fail

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

for each x, need to see it at least once in order to give non zero estimate.

Can we do better?