Lower Bounds on distributions

last time: sketch of lower bound for uniformity testing

Homework: One way of making it formal (not optimal in all parameters)

Today: Another methodology of showing lower bounds

def. Uniformity tester

given samples from p on $[n]$, ε

- if $p = U_n$ output PASS with prob $\geq 3/4$
- if $\|p - U_n\|_1 > \varepsilon$ output FAIL with prob $\geq 3/4$

any constant $> 1/2$ to do better than random guess

Thm Uniformity tester needs $\Omega(\sqrt{n}/\varepsilon^2)$ samples

Proof soon, 1st some observations + basics:

Observation: randomness doesn't help testing algorithms

Pf.: h.w.

Information Theory Basics:

$p(y|x) \equiv P(Y=y|X=x)$

Entropy $\quad H(X) = -\sum_{x \in \text{domain}} p(x) \log p(x)$

Conditional Entropy $\quad H(Y|X) = E_x\left[\sum_{y \text{ st. } p(y)\neq 0} p(y|x) \log \frac{1}{p(y|x)} \right]$

$$= \sum_x p(x) \sum_{y \text{ st. } p(y)\neq 0} p(y|x) \log \frac{1}{p(y|x)}$$

Note:

$H(Y|X) = 0$ iff Y determined by X

$H(Y|X) = H(Y)$ iff Y independent of X

Basic facts:

- $H(x) \geq 0$
- $H(Y|X) \leq H(Y)$
- Chain rule: $H(X,Y) = H(X) + H(Y|X)$

joint entropy of pair (x,y)

Mutual Information:

$$I(X,Y) = H(X) + H(Y) - H(X,Y)$$
$$= H(X) - H(X|Y)$$
$$= H(Y) - H(Y|X)$$

measure of how independent X,Y are or how much X allows you to predict Y

Chain rule: $I(X;(Y,Z)) = I(X;Z) + I(X;Y|Z)$

Main Idea:

define random var X as fair coin flip

X decides whether pick K samples from
- uniform on $[n]$
- uniform on S s.t. $|S| = \frac{n}{2}$ + S chosen randomly

all K from same distribution

we get samples, not X. Can we figure out what X is from samples?

Will show, if K small, $I(X, \text{Samples}) = o(1)$

any improvement over random gives $o(1)$

So what?

Lemma if f any fctn (algorithm) s.t. $\Pr_{X,\text{samples}}[f(\text{samples}) = X] \geq 51\%$

then $I(X;A) \geq 2 \cdot 10^{-4}$

So if $I(X, \text{Samples}) = o(1) \Rightarrow$ no algorithm can solve the testing problem with high enough probability

$a_i \leftarrow$ # times elt i appears in sample

$$I(x, \text{samples}) = I\left(x, \{a_i\}_{i=1}^{n}\right)$$

Lets assume a_i's independent (they are not if k is fixed, but if k chosen as Poisson dist with mean k_0, they are independent)

$$I\left(x, \{a_i\}_{i=1}^{n}\right) \leq \sum_{i=1}^{n} I(x, a_i) \quad \text{by chain rule}$$

drawn identically $\forall i$

$$= n \cdot I(x, a_1) = O\left(\frac{k^2 \varepsilon^4}{n}\right)$$

$$\boxed{\begin{array}{l} \underline{\text{Lemma}} \quad I(x, a_1) = O\left(\dfrac{k^2 \varepsilon^4}{n^2}\right) \\ \text{Proof: calculations} \end{array}}$$

if $k = \Theta\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$

this is $\Theta(1)$

Poissonization

An important way to get rid of dependencies.

why:
if take fixed k # of samples

Pr[see elt i] not independent of Pr[see elt j].

why? if you see elt i, you know 1 sample is not j, so less likely you will see elt j in all k samples (you now only have $k-1$ samples left to "play with").

Poissonization trick:
pick k distributed as Poisson with parameters

def. Poisson dist with parameter λ $(\Psi(\lambda))$:

k occurs with prob $\dfrac{\lambda^k e^{-\lambda}}{k!}$ ← Note: $0! = 1$, $\Psi(0) = 0$

Observe

$$\sum_{k \geq 0} \frac{\lambda^k e^{-\lambda}}{k!} = 1$$

$E[X] = \lambda$ for $X \leftarrow \Psi(\lambda)$

$Var[X] = \lambda$

Poisson Sampling: pick $k \sim \Psi(\lambda)$
take k samples of distribution

Important property of Poisson Sampling:

- # of occurences of elt i is <u>independent</u> of
  " " " " " j (for $i \neq j$)

- # of occurences of elt i $\sim \Psi(k \cdot p_i)$
  E[" " " " " "] $= k \cdot p_i$
  Var[" " " " " "] $= k \cdot p_i$

Why does this give us a lower bound?

Suppose you want to show $\geq S_0$ samples are required for a testing problem,

i.e. $\forall\ \mathcal{A}$ taking S_0 samples, $\mathcal{A}$ correct with probabilty $\geq 2/3$

$\Downarrow$

$\forall\ \mathcal{A}'$ taking $\underbrace{\Psi(c \cdot S_0)}_{c>1}$ samples, $\mathcal{A}'$ correct with prob $\geq 2/3 -$ "tiny"

expectation $c \cdot S_0$
prob # samples $< S_0$ is "tiny"

<u>Contrapositive</u>: if $\mathcal{A}'$ needs $\geq \Psi(c \cdot S_0)$ samples
then $\mathcal{A}$ needs $\geq S_0$ samples

Sketch of l.b. for p, q given by samples $\Leftarrow$ "closeness testing"

<u>Thm</u> Closeness testing requires $\Omega(n^{2/3})$ samples

<u>Proof idea</u>:

$p_0 =$

$n^{2/3}$ heavy elements weight $\frac{1}{2n^{2/3}}$

$\frac{n}{4}$ light elements weight $\frac{2}{n}$

light elements are disjoint

q_0

$n^{2/3}$ heavy elements

$\frac{n}{4}$ light elements

Positive pairs

ℓ_1 dist $= 0 \Rightarrow$ $(\pi(p_0), \pi(p_0))$ $\forall \pi$

Negative pairs

$(\pi(p_0), \pi(q_0))$ $\forall \pi$ $\Leftarrow$ ℓ_1 dist $= 1$

where $\pi(p)$ relabels domain elts randomly

$\pi(p_0), \pi(p_0)$ applies <u>same</u> relabeling to both

<u>Main idea</u>: $\boxed{\text{Only Collision Statistics matter!}}$

for positive pairs have collisions in both heavy + light elts

for negative pairs have collisions <u>only</u> in heavy elts

when see a collision, usually can't tell if it was a heavy or light element!

After $o(n^{2/3})$ samples:

probability see any small element twice really small ⇐ happens, but not too often

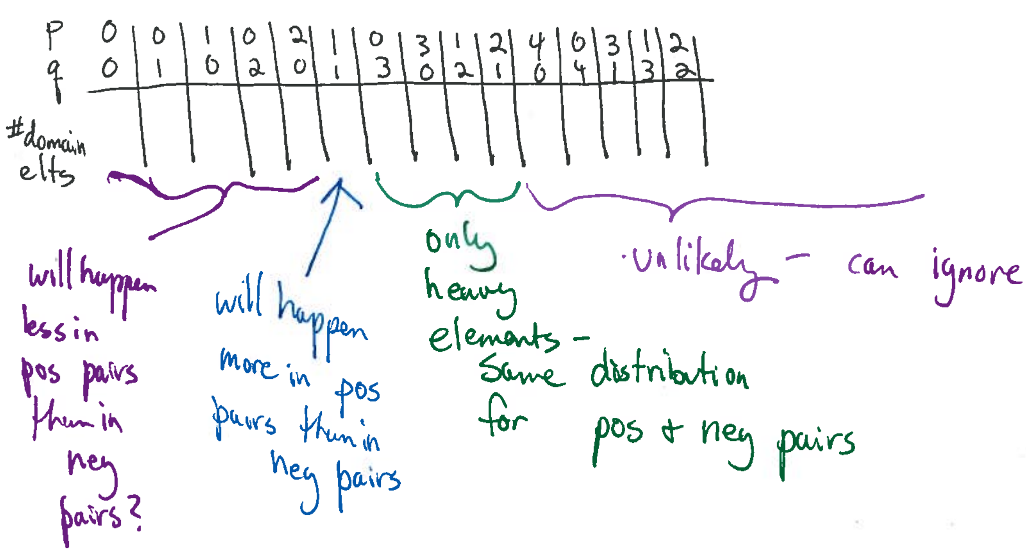
probability see any heavy element 3X is small ⇐ happens, but not too often

probability see any small elt 3X is tiny ⇐ unlikely to happen

heavy " 4X is tiny ⇐ unlikely to happen

So, what collision statistics could we have?

how many elts in domain appear n_p times, n_q times in p, q?



When you see collision, you don't know if it came from heavy or light element

m = # samples

H = # heavy collisions ← same distribution for pos + neg pairs

L = # light collisions (1 from each dist) ← $= 0$ when neg pair

$$E[\# \text{collisions in pos pair}] = E[H] + E[L] = \frac{m^2}{2n^{2/3}} + \frac{m^2}{n} \approx \frac{m^2}{2n^{2/3}}$$

$$E[\# \text{collisions in neg pair}] = E[H] = \frac{m^2}{2n^{2/3}}$$

Need to show something a bit stronger - can't distinguish the random variables!

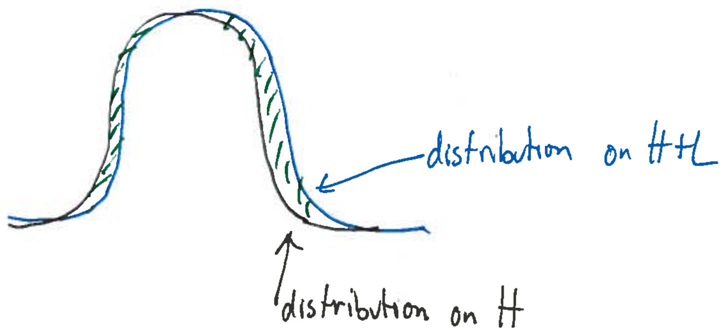
$$E[H] = \frac{m^2}{4n^{2/3}}$$

$\binom{m}{2}$ pairs, each collides with prob $\frac{1}{2n^{2/3}}$

$$\mathrm{Var}[H] \approx \frac{m^2}{n^{2/3}}$$

$$E[L], \mathrm{Var}[L] \approx \frac{m^2}{n}$$

$\binom{m}{2}$ pairs, each collides with prob $\frac{1}{n}$



L_1 distance small

$\Downarrow$

almost same distribution

$\Downarrow$

hard to distinguish!

how do we show L_1 dist is small?

if they were Gaussian,

could show that $\sqrt{\mathrm{Var}(H)} \leq E[L]$

$\Leftarrow$ they aren't quite, so it's more difficult.

$$\Leftrightarrow \frac{m}{n^{1/3}} \leq \frac{m^2}{n}$$

$$\Leftrightarrow m \geq n^{2/3}$$