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

Definition: Uniformity tester

given samples from \( p \) on \([n]\), \( \epsilon \)

- if \( p = \mathcal{U}_n \) output PASS with prob \( \geq \frac{3}{4} \)
- if \( \|p - \mathcal{U}_n\|_1 \geq \epsilon \) output FAIL with prob \( \geq \frac{3}{4} \)

This uniformity tester needs \( \Omega(n^{3/2}) \) samples

Proof: soon; 1st some observations + basics:

Observation: randomness doesn't help testing algorithms

Proof: h.w.

Information Theory Basics:

Entropy
\[ H(x) = -\sum_{x \in \text{domain}} p(x) \log p(x) \]

Conditional Entropy
\[ H(Y|X) = \mathbb{E}_x \left[ \sum_{y \text{ s.t. } p(y|\neg x) = 0} p(y|x) \log \frac{1}{p(y|x)} \right] \]
\[ = \sum_x p(x) \sum_{y \text{ s.t. } p(y|\neg x) = 0} p(y|x) \log \frac{1}{p(y|x)} \]

Note:

- \( H(Y|X) = 0 \) iff \( Y \) determined by \( X \)
- \( H(Y|X) = H(Y) \) iff \( Y \) independent of \( X \)
Basic facts:

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

Mutual information:

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

$$= H(X) - H(X|Y)$$

$$= H(Y) - H(Y|X)$$

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 $[n]$

$\uparrow$

all $K$

from same distribution

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

So what?

Lemma: if for any func (algorithm) s.t. $Pr[f(\text{samples})=x] \geq 51\%$

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

So if $I(X;\text{samples}) = o(1)$ \Rightarrow no algorithm can solve the testing problem
Let's assume \( a_i \)'s are independent (they are not if \( K \) is fixed, but if \( K \) chosen as Poisson dist with mean \( K_0 \), they are independent).

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

\[
eq n \cdot I(x, a_1) = O\left( \frac{K^2 \varepsilon^4}{n^2} \right)
\]

**Lemma**

\[
I(x, a_1) = O\left( \frac{K^2 \varepsilon^4}{n^2} \right)
\]

*Proof: Calculations*

\[
\text{if } K = O\left( \frac{\sqrt{n}}{\varepsilon^2} \right)
\]

This is \( O(1) \)
Poissonization

An important way to get rid of dependencies.

Why:
if take fixed K # of samples

\[ \text{Pr[see elt } j \text{]} \text{ not independent of Pr[see elt } j \text{]}. \]

why? if you see elt \( j \), you know \( j \) samples
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 parameter \( \lambda \).

def Poisson dist with parameter \( \lambda \) \( (\Psi(\lambda)) \):

\[ \text{K occurs with prob } \frac{\lambda^k e^{-\lambda}}{k!} \]

\[ \sum_{k=0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} = 1 \]

\[ E[X] = \lambda \quad \text{for } X \sim \Psi(\lambda) \]

\[ \text{Var}[X] = \lambda \]

Poisson sampling:
pick \( K \sim \Psi(\lambda) \)
take \( K \) samples of distribution

\[ \Psi(0) = 0 \]

\[ 0! = 1 \]
Important property of Poisson Sampling:

- \# of occurrences of elt i is independent of \# of occurrences of elt j (for i \neq j)

- \# of occurrences of elt i \sim \Psi(K \cdot p_i)

\[ \mathbb{E}[\Psi(K \cdot p_i)] = K \cdot p_i \]

\[ \text{Var}[\Psi(K \cdot p_i)] = K \cdot p_i \]

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. If A taking \( s_0 \) samples, A correct with probability \( \geq 2/3 \).

\[ \forall A' \text{ taking } \Psi(c \cdot s_0) \text{ samples, } A' \text{ correct with prob } \geq 2/3 - \text{"tiny"} \]

Contrapositive: if A' needs \( \geq \Psi(c \cdot s_0) \) samples

Then A needs \( \geq s_0 \) samples
Sketch of I.b. for $p, q$ given by samples "closeness testing"

Thin closeness testing requires $\Omega(n^{2/3})$ samples

Proof idea:

$P_0 = \frac{n^{2/3}}{\frac{2}{n}}$ heavy elements

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

$Q_0 = \frac{n^{2/3}}{\frac{2}{n}}$ heavy elements

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

Positive pairs $\delta, dist = 0 \Rightarrow (\Pi(p_0), \Pi(q_0)) \forall \Pi$ $\Rightarrow (\Pi(p_0), \Pi(q_0)) \forall \Pi \leq \delta, dist = 1$

Negative pairs

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

$\Pi(p_0), \Pi(p)$ applies same relabeling to both

Main idea: Only Collision Statistics matter!

for positive pairs have collisions in both heavy + light elts

for negative pairs have collisions only 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
- Probability see any heavy element $3X$ is small happens, but not too often
- Probability see any small elt $3X$ is tiny, $4X$ is tiny unlikely to happen

So, what collision statistics could we have?

How many elts in domain appear $p_i$ times, $q_i$ times in $p_i q_i$?

<table>
<thead>
<tr>
<th>p</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>q</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
</tbody>
</table>

#domain elts

- Will happen less in pos pairs
- Will happen more in pos pairs than in neg pairs
- Unlikely - can ignore

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

$m = \# \text{ samples}$

$H = \# \text{ heavy collisions}$

$L = \# \text{ light collisions (1 from each dist)}$

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

\[
E[\# \text{ collisions in neg pair}] = E[H] = \frac{m^2}{2n^{2/3}}
\]
Need to show something a bit stronger - can't distinguish the random variables!

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

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

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

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

\( (m^2) \) pairs, each collides with prob \( \frac{1}{n} \)

\[ L_1 \text{ distance small} \]

\[ \text{almost same distribution} \]

\[ \text{hard to distinguish!} \]

how do we show \( L_1 \) dist is small?

if they were Gaussian,

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

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

\[ \Leftrightarrow \ m \leq n^{2/3} \]