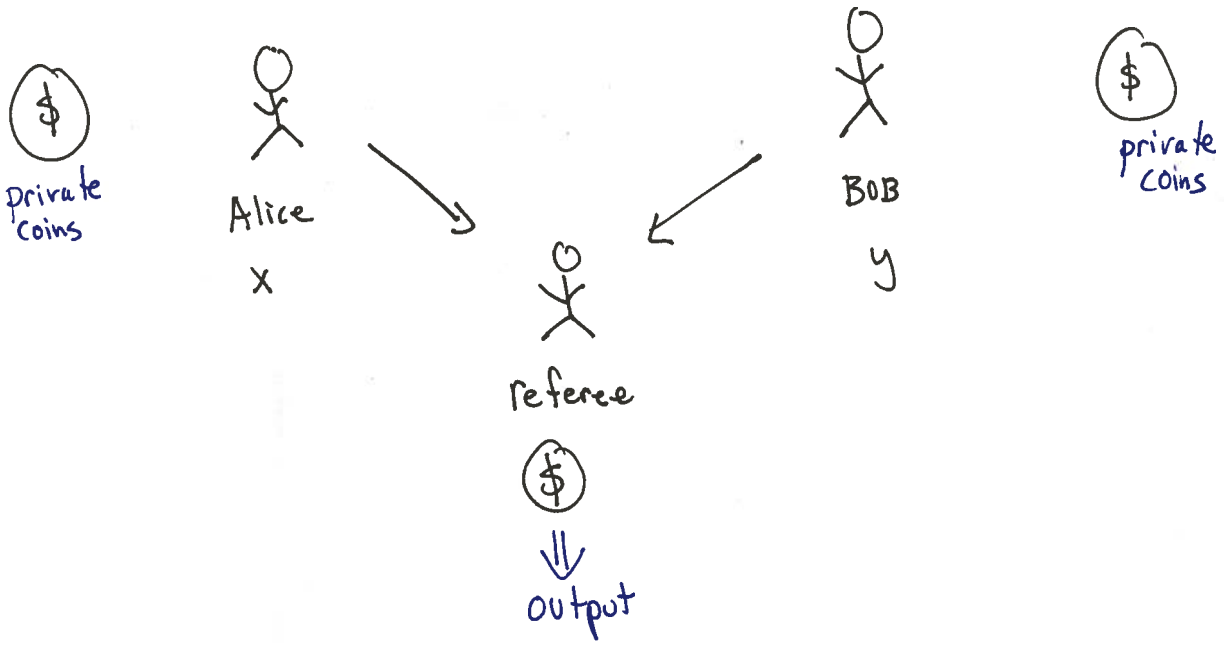
Lecture 21:

More lower bounds for distribution testing problems

- Communication complexity vs. sample complexity
- overview of closeness testing l.b.

Communication Complexity: Simultaneous Message Passing Model (SMP)



note: 1 round

let $f(x,y) = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{o.w} \end{cases}$

Known Thm:

If $|X| = |Y| = K$

must communicate $\Omega(\sqrt{K})$ bits to compute f

Yet another $\tilde{\Omega}(\sqrt{n})$ lower bound for testing uniformity of a distribution:

Useful fact:

$\exists$ error correcting code $C : \{0,1\}^k \to \{0,1\}^n$ s.t.

(1) constant rate: $\frac{k}{n} \geq$ const

(2) relative distance $\delta = \Omega(1)$

(3) each code word (each y in range of C) is "balanced" – i.e., # 1's = # 0's

The reduction:

Alice: given x

- Computes $C(x)$
- $A \leftarrow \{i \mid C(x)_i = 1\}$
- send s uniformly dist samples from A to ref

Bob: given y

- Computes $C(y)$
- $B \leftarrow \{j \mid C(x)_j = 0\}$
- send s unif dist samples from B to ref

Referee:

- receives samples from $A+B$
- constructs s samples of dist $q \equiv \frac{U_A + U_B}{2}$:

  sample i:

  toss coin
  heads: output next sample from A
  tails: " " " B
- feeds to uniformity tester & outputs pass if uniformity test passes

Why does it work?

if $x=y$:

$C(x) = C(y)$

$A+B$ are partition of $[n]$ s.t. $|A| = |B|$

$U_{[n]} = \frac{U_A + U_B}{2} \equiv q$

so uniformity tester accepts q ✓ correct answer for this case

if $x \neq y$:

$C(x) \neq C(y)$ disagree on $\delta = \Omega(1)$ fraction of domain

so $|A \cap B| = \frac{\delta}{2}$ & $|[n] \setminus (A \cup B)| = \frac{\delta}{2}$

$\Rightarrow$ q not supported on $\frac{\delta}{2}$ fraction of domain
& "double weight" on " " "

so $\|q - U_{[n]}\|_1 > \delta$

if $X \neq y$ (cont.)

so unif test rejects q ✓ correct answer for this case

sample complexity:
each sample encoded via $O(\log n)$ bits

$\Rightarrow$ comm. complexity is $O(s \log n)$

but we know that comm complexity for $X \stackrel{?}{=} y$ is $\Omega(\sqrt{k})$

$$\Rightarrow s = \Omega\left(\frac{\sqrt{k}}{\log n}\right) \overset{\text{const rate}}{=} \Omega\left(\frac{\sqrt{n}}{\log n}\right) = \tilde{\Omega}(\sqrt{n})$$

So what?

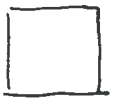
why yet another (weaker) proof of uniformity testing?

methodology works for many other problems in dist testing (via fancier codes, reductions)

Sketch of 1.b. for p, q given by samples $\Leftarrow$ "closeness testing"

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

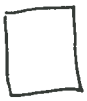
Proof idea:

$p_0 =$ 

$n^{2/3}$ 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 $=0 \Rightarrow (\pi(p_0), \pi(p_0)) \quad \forall \pi$

Negative pairs

$(\pi(p_0), \pi(q_0)) \quad \forall \pi \Leftarrow \ell_1$ dist $= 1$

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

$\pi(p_0), \pi(p_1)$ 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

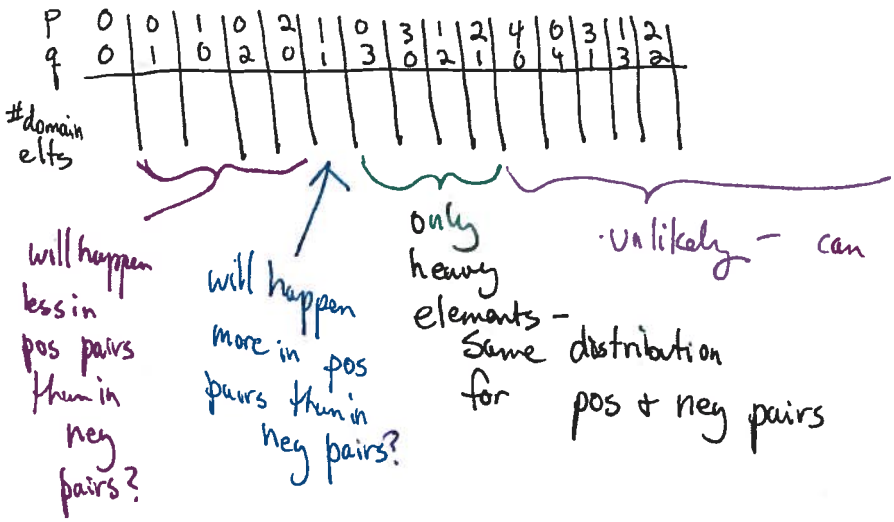
when see a collision, usually can't tell if it was a heavy or light element!

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 ⇐
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

$2m$ = # samples (pretend m heavy, m light from each of p, q) * cheat #1
H = # heavy collisions ← same distribution for pos & neg pairs
L = # light collisions (1 from each dist) ← = 0 when neg pair

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

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

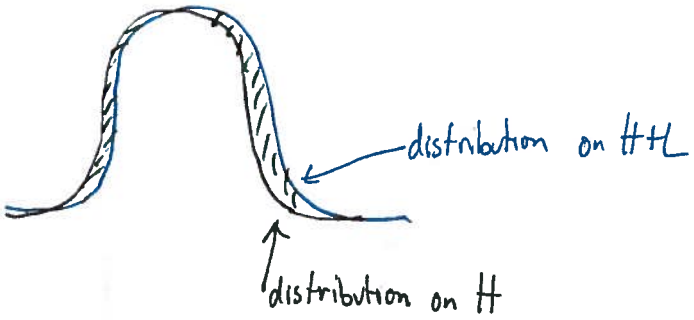
Need to show something a bit stronger - can't distinguish the random variables!

$E[H] = \frac{m^2}{8n^{2/3}}$ $\approx m^2$ pairs, each collides with prob $\frac{1}{2n^{2/3}}$

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

$E[L], Var[L] \approx \frac{m^2}{n}$ $\approx m^2$ pairs, each collides with prob $\frac{2}{n}$

L_1 distance small
⇓
almost same distribution
⇓
hard to distinguish!



how do we show L_1 dist is small?

if they were Gaussian, could show that $\sqrt{Var(H)} \leq E[L]$

⇐ they aren't quite, so it's more difficult.

* cheat #2 ⇑ bigger difficulty!!

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

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