Streaming Lower Bounds via Communication Complexity

$x, y \in \{0,1\}^n$

$\mathrm{DISJ}(x,y) = \begin{cases} 0 & \text{iff } \langle x,y \rangle = 0 \\ 1 & \end{cases}$

Thm Randomized communication complexity of DISJ is $\Omega(n)$

Suffices to find distrib^n on inputs st. any deterministic protocol w. low communication has large error

INDEX

Given $x \in \{0,1\}^n$ index $i \in [n]$

Compute $\mathrm{INDEX}(x, i) = x_i \in \{0,1\}$

Claim: In order to solve $\mathrm{INDEX}(x,i)$ can solve $\mathrm{DISJ}(x, e_i)$

Thm: Randomized communication complexity of INDEX is $\Omega(n)$

Pf: Yao's lemma: construct a distrib^n D on (x,i) st.

any deterministic protocol w. low error uses $\Omega(n)$ bits.

D: $x \in_R \{0,1\}^n$, $i \in_R [n]$

Any deterministic protocol with $\leq 0.1n$ bits of communication must have error $\geq \frac{1}{8}$

$c = 0.1$

Fix det. one-way protocol P with $\leq cn$ bits of communication

Alice sends only 2^{cn} distinct messages z to Bob

$f : \{0,1\}^n \to \{0,1\}^{cn}$ Alice's f^n mapping input x to output z

Suppose Bob gets z from Alice and his input is i

Bob announces guess for i^{th} bit

Hold z fixed, consider Bob's answers for $i = 1 \ldots n$

Answer vector $a(z) \in \{0,1\}^n$

$\leq 2^{cn}$ messages $z \Rightarrow \leq 2^{cn}$ answer vectors $a(z)$

Fix Alice's input x, resulting in message $z = z(x)$

Protocol is correct for Bob's input i iff $a(z)_i = x_i$

Bob's index i chosen uniformly

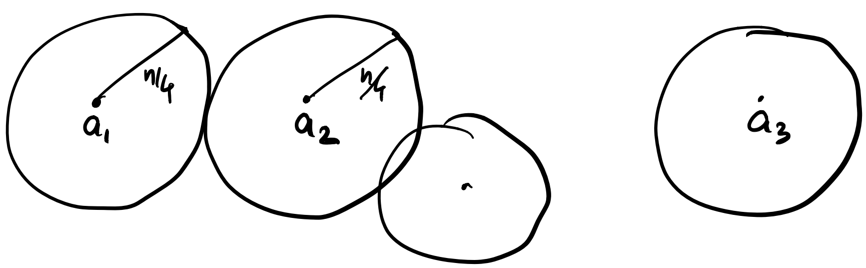
$$\Pr_i[P \text{ is incorrect} \mid x, z] = \frac{d_H(x, a(z))}{n}$$

<u>Goal:</u> w. constant prob. over choice of x this expression is larger than a constant

$A = \{ a(z(x)) : x \in \{0,1\}^n \}$ set of all answer vectors used by protocol P

$|A| \leq 2^{cn}$

Alice's input x is <u>good</u> if $\exists$ answer vector $a \in A$ with $d_H(x,a) < \frac{n}{4}$, <u>bad</u> otherwise



<u>Claim</u>: There are at least 2^{n-1} <u>bad</u> inputs x

Why does this imply theorem?

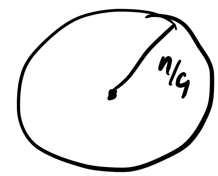
$$\Pr_{(x,y)\in D}[P \text{ is wrong on } (x,y)] = \Pr[x \text{ is good}] \cdot \overbrace{\Pr[P \text{ wrong on } (x,y) \mid x \text{ is good}]}^{\geq 0}$$
$$+ \underbrace{\Pr[x \text{ is bad}]}_{\geq \frac{1}{2}} \cdot \underbrace{\Pr[P \text{ wrong on } (x,y) \mid x \text{ is bad}]}_{\geq 1/4}$$

$$\Pr_{(x,y)\in D}[P \text{ wrong on } (x,y) \mid x \text{ is bad}] = E_x\left[\frac{d_H(x, a(z(x)))}{n} \,\middle|\, x \text{ is bad}\right]$$
$$\geq \left[\min_{a\in A} \frac{d_H(x,a)}{n} \,\middle|\, x \text{ is bad}\right]$$

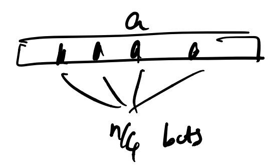
$$\geq \frac{1}{4}$$

$\Rightarrow$ Protocol P has error $\geq \frac{1}{8}$

Proof of Claim: Fix answer vector $a \in A$



inputs x w. Hamming distance $\leq \frac{n}{4}$ from a



$$\underset{\text{dist } 0}{1} + \underset{\text{dist } 1}{\binom{n}{1}} + \underset{\text{dist } 2}{\binom{n}{2}} + \cdots + \underset{\text{dist } \frac{n}{4}}{\binom{n}{n/4}}$$

$$\binom{n}{k} \leq \left(\frac{en}{k}\right)^k \qquad \text{(from Stirling's approx)}$$

$$\sum_{k=0}^{n/4} \binom{n}{k} \leq 2\binom{n}{n/4} \leq 2\left(\frac{en}{n/4}\right)^{n/4} = 2(4e)^{n/4}$$

$$= 2 \cdot 2^{\log_2(4e) \cdot n/4}$$

$$= 2 \cdot 2^{0.861 n}$$

total # good inputs $\leq |A| \cdot 2 \cdot 2^{0.861 n}$

$\leq 2 \cdot 2^{(0.861 + c) n}$ $\qquad c = 0.1$

$\leq 2^{n-1}$ for n sufficiently large $\blacksquare$

So far

INDEX $\longrightarrow$ DISJ $\longrightarrow$ linear lower bounds for streaming (F_∞)

Next: Dependence on approximation parameter ε to compute $(1+\varepsilon)$ approx of freq. moments

Why quadratic dependence on ε?

Goal: Any streaming algo. that computes $(1+\varepsilon)$ approx of F_0, F_2 needs $\Omega(\frac{1}{\varepsilon^2})$ space

Focus on extreme case $(1+\frac{1}{\sqrt{n}})$ approx requires $\Omega(n)$ space

Special case has all ideas needed for $\Omega(\frac{1}{\varepsilon^2})$ bound

Disjointness doesn't work

Suppose we have streaming algo S that gives $(1+\frac{1}{\sqrt{n}})$ approx to F_0

Follow redn for F_∞

Alice's input x, Bob's input y : concatenate xy

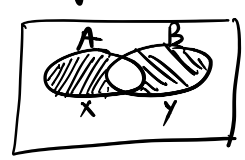
$DISJ(x,y) = 0$	$F_0(xy) = \lvert x\rvert + \lvert y\rvert$	n
$= 1$	$F_0(xy) \leq \lvert x\rvert + \lvert y\rvert - 1$	$n-1$

$(1+\frac{1}{\sqrt{n}})$ approx of F_0 gives additive error of $\sqrt{n}$

Relate to Hamming Distance

$x, y \in \{0,1\}^n$ Universe $\mathcal{U} = \{1, \ldots n\}$

x, y : characteristic vectors of subsets A, B of $\mathcal{U}$



$d_H(x,y) = |A \setminus B| + |B \setminus A|$

$F_0 = |A \cup B|$

$|A \setminus B| = F_0 - |B|$, $|B \setminus A| = F_0 - |A|$

$d_H(x,y) = 2F_0 - |x| - |y|$

Bob knows y, Alice can send $|x|$ to Bob using $\log_2 n$ bits

One way communication protocol that computes F_0 w. comm. c yields one-way protocol that computes $d_H(x,y)$ w. comm. $c + \log_2 n$

$(1+\frac{1}{\sqrt{n}})$ approx to F_0 $\qquad [F_0, \ (1+\frac{1}{\sqrt{n}})F_0]$

yields protocol to estimate $d_H(x,y)$ up to $\frac{2F_0}{\sqrt{n}} \leq 2\sqrt{n}$ additive error with $\log_2 n$ extra communication

<u>Goal</u>: Hamming distance estimation w. additive error has large communication complexity

convert to decision problem

For a parameter t, constant c

$$\text{GAP-HAMMING}(t) = \begin{cases} 1 & \text{if } d_H(x,y) \leq t - c\sqrt{n} \\ 0 & \text{if } d_H(x,y) \geq t + c\sqrt{n} \\ \text{undefined otherwise} \end{cases}$$

Promise problem

GAP-HAMMING(t) reduces to $(1+\frac{c}{\sqrt{n}})$ approx of F_0

How to pick t?

$t = 0$? $\qquad t = c\sqrt{n}$

$$\text{GAP-HAMMING}(c\sqrt{n}) = \begin{cases} 1 & \text{if } d_H(x,y) = 0 \\ 0 & \text{if } d_H(x,y) \geq 2c\sqrt{n} \\ \text{undefined otherwise} \end{cases}$$

We will pick $t = \frac{n}{2}$