Lower Bounds for property testing algorithms

I. Deterministic lower bounds ⇒ probabilistic lower bounds

a difficulty:

prop testing algs are randomized!
difficult to argue about their behavior

useful lower bnd tool:

Yao's principle:

If there is a probability distribution D on union of "positive" ("YES"/PASS) & "negative" ("NO"/FAIL) elements of domain ("inputs"), such that <u>any</u> <u>deterministic algorithm</u> of query complexity $\le t$ is incorrect with prob $\ge 1/3$ for inputs chosen according to D, then t is a lower bound on randomized query complexity.

So average case deterministic lower bound ⇒ randomized worst case lower bound

(principle works for all types of randomized algorithms) why?

proof omitted

game theoretic view:

Alice selects deterministic alg A } payoff = cost of $A(x)$
Bob selects input x

Von Neuman's minimax $\Rightarrow$ Bob has randomized strategy does as well when A randomized

An example:

note that testing $L_n' = \{ w \mid w = vv^R \}$ is trivial! compare w_i to w_{n-i+1}

$$L_n = \left\{ w \;\middle|\; \begin{array}{l} w \text{ is } n\text{-bit string} \\ w = vv^R uu^R \end{array} \right\}$$

concatenations of palindromes

<u>Thm</u> need $\Omega(\sqrt{n})$ queries to property test L_n

i.e. if A satisfies

$\forall x \in P$, $\Pr[A(x) = \text{PASS}] \geq 2/3$

$\forall x$ ε-far from P, $\Pr[A(x) = \text{FAIL}] \geq 2/3$

then A makes $\Omega(\sqrt{n})$ queries

<u>Pf</u>. Plan: give distribution on inputs that is hard for <u>all</u> algorithms with $o(\sqrt{n})$ queries

Yao $\Rightarrow$ randomized l.b. of $\Omega(\sqrt{n})$

- wlog assume $6 \mid n$
- distribution on negative inputs:

should output FAIL

N = random string of distance $\geq \varepsilon n$ from L_n

Pf of claim 2 (idea)

To show: for every fixed set of $o(\sqrt{n})$ queries, lots of strings in L_n follow that path.

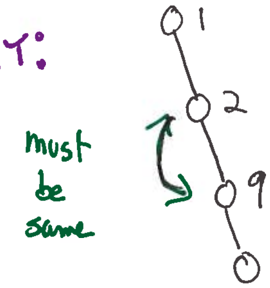
Count # strings that agree with t queries in leaf?

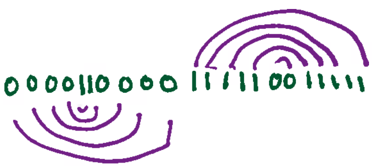
$$= 2^{n-t}$$

Count # strings in L_n that agree with t queries to l?

$$\geq (2^{n-t}) - ?$$

MAIN DIFFICULTY:





Fix $k = 10$

once you see 1, that fixes what you should see at 10

2	9
3	8
4	7
5	6
11	n
12	$n-1$
⋮	

☹ so maybe no string in L_n follows the path?

☺ no! k could be $\frac{n}{6} \dots \frac{n}{3}$

so for each set of queries some k's (but not all) are bad

- distribution on positive inputs:

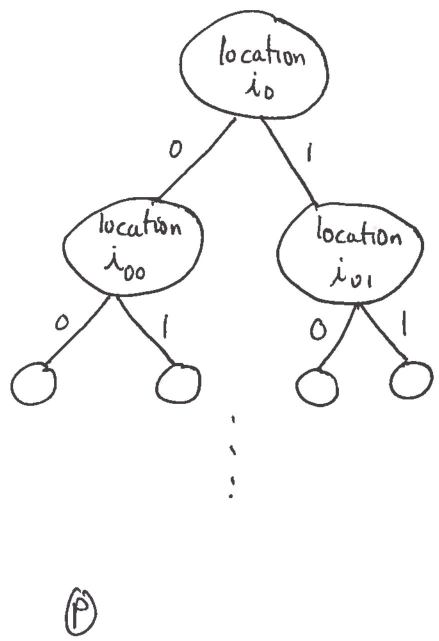
$$P = \begin{cases} 1. & \text{pick } k \in_R [\tfrac{n}{6}H, \tfrac{n}{3}] \\ 2. & \text{pick random } v, u \text{ s.t. } |v| = k,\ |u| = \tfrac{n-k}{2} \\ 3. & \text{output } vv^R uu^R \end{cases}$$

should output PASS

an issue: some strings can be generated via ≥ 1 k

- distribution D = { • flip coin • if H output according to N, else " " " P }

- Assume deterministic algorithm A has behavior above + uses $\leq t = o(\sqrt{n})$ queries



depth t, $\leq 2^t$ root-leaf paths

wlog all leaves have depth t

leaves labelled with A's answer following that path + seeing those bits

note: we can calculate prob of reaching a leaf since we know input distribution

if a input reaches here, hopefully it is a "FAIL" input?

For each leaf ℓ:

$$E^-(\ell) = \{ \overset{\text{inputs}}{w} \in \{0,1\}^n \quad \underbrace{\text{dist}(w, L) \geq \varepsilon n}_{w \text{ should fail}} \quad w \text{ reaches leaf } \ell \}$$

$$E^+(\ell) = \{ \underbrace{\overset{\text{inputs}}{w} \in \{0,1\}^n \cap L}_{w \text{ should Pass}}, \quad w \text{ reaches leaf } \ell \}$$

each leaf ℓ is either passing or failing, not both

Total error of A on D

$$= \sum_{\substack{\ell \\ \text{passing}}} \Pr_{w \in D} [w \in \underset{\substack{\uparrow \\ \text{should FAIL}}}{E^-(\ell)}] + \sum_{\substack{\ell \\ \text{failing}}} \Pr_{w \in D} [w \in \underset{\substack{\uparrow \\ \text{should PASS}}}{E^+(\ell)}]$$

Claim 1 if $t = o(n)$, $\forall \ell$ at depth t

$$\Pr_D [w \in E^-(\ell)] \geq \left(\tfrac{1}{2} - o(1)\right) 2^{-t}$$

(so negative inputs show up at all leaves & should be failed)

but each leaf only has one label!

Claim 2 if $t = o(\sqrt{n})$ $\forall \ell$ at depth t

$$\Pr_D [w \in E^+(\ell)] \geq \left(\tfrac{1}{2} - o(1)\right) 2^{-t}$$

(so positive inputs show up at all leaves & should be passed)

Putting them together to prove full theorem

error of $\mathcal{A}$ on D

$$= \sum_{\ell \text{ passing}} \Pr_{w \in D}[w \in E^-(\ell)] + \sum_{\ell \text{ failing}} \Pr_{w \in D}[w \in E^+(\ell)]$$

$$\geq \sum_{\ell \text{ passing}} \left(\tfrac{1}{2} - o(1)\right) 2^{-t} + \sum_{\ell \text{ failing}} \left(\tfrac{1}{2} - o(1)\right) 2^{-t}$$

$$\geq \tfrac{1}{2} - o(1) \quad \leftarrow \text{since all leaves pass or fail}$$

$\blacksquare$

Pf of Claim 1:

idea: N is close to U

+ U ends up uniformly distributed at each leaf $\Rightarrow \Pr_{w \in U}[w \in E(\ell)] = \frac{2^{n-t}}{2^n} = 2^{-t}$

How much does the distribution change by using N instead of U?

$$|L_n| \leq \underbrace{2^{n/2}}_{\text{choice of } u,v} \cdot \underbrace{\tfrac{n}{2}}_{\text{choice of } i}$$

words at distance $\leq \varepsilon$: $2^{\frac{n}{2}} \cdot \frac{n}{2} \cdot \sum_{i=0}^{\varepsilon n} \binom{n}{i} \leq 2^{\frac{n}{2} + 2\varepsilon \log\left(\frac{1}{\varepsilon}\right) n}$

$$\text{so } E^-(\ell) \geq \underbrace{2^{n-t}}_{\text{\# strings that follow path to leaf}} - \underbrace{2^{\frac{n}{2} + 2\varepsilon \log\left(\frac{1}{\varepsilon}\right) n}}_{\text{\# words at dist } \leq \varepsilon} = (1 - o(1))\, 2^{n-t}$$

assume $\varepsilon \ll 1/8$

t is $o(n)$

So 1st term swamps 2nd term

$$\text{so } \Pr_D[w \in E^-(\ell)] = \tfrac{1}{2} \Pr_N[w \in E^-(\ell)]$$

$$\geq \frac{1}{2} \frac{|E^-(\ell)|}{2^n} \geq \left(\tfrac{1}{2} - o(1)\right) 2^{-t}$$

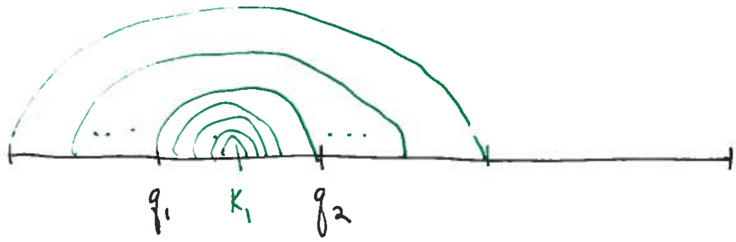
$\blacksquare$

Given leaf ℓ, let $Q_\ell \leftarrow$ <u>indices</u> queried along the way

For each of $\binom{t}{2}$ pairs of queries $q_1, q_2 \in Q_\ell$

at most 2 choices of k for which q_1, q_2

symmetric to k or $n/2 + k$



in this case, only one choice

$\Rightarrow$ # choices of k s.t.

no pair in Q_ℓ symmetric

around k or $n/2+k$

is $\geq \frac{n}{6} - 2\binom{d}{2} = (1-o(1))(n/6)$

For these k, # strings ~~that~~ follow path $= 2^{n/2 - |Q_\ell|}$

$$\text{So } \Pr_p[w \in E^+(\ell)] = \sum_w \sum_k \underbrace{\Pr_D[w|k]}_{2^{-n/2}} \cdot \underbrace{\Pr[\text{choose } k]}_{\frac{6}{n}} \cdot \mathbb{1}_{w \in E^+(\ell)}$$

$$= \frac{1}{\frac{n}{6} \cdot 2^{n/2}} \cdot \left[(1-o(1))\frac{n}{6}\right] \cdot \left[2^{n/2 - |Q_\ell|}\right] = (1-o(1))\, 2^{-|Q_\ell|}$$

$$= (1-o(1))\, 2^{-t}$$

$$\Rightarrow \Pr_D[w \in E^+(\ell)] = \left(\frac{1}{2} - o(1)\right) 2^{-t}$$

∎