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 any deterministic algorithm of query complexity $\leq t$ is incorrect with prob $\geq 1/3$ 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
Bob selects input x
} payoff = cost of $A(x)$

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

An example:

$$L_n = \{ w \mid w \text{ is } n\text{-bit string}, \; w = vv^R uu^R \}$$

concatenations of palindromes

Thm 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

Pf

- wlog assume $6 \mid n$
- distribution on negative inputs! ← should output FAIL

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

• distribution on positive inputs:

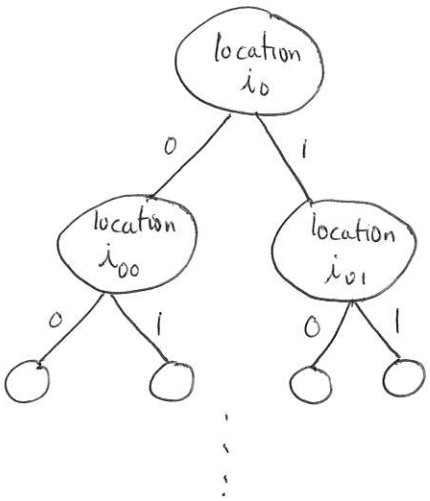
should output PASS

$P = \begin{cases} 1. & \text{pick } k \in_R \left[\frac{n}{6}+1, \frac{n}{3}\right] \\ 2. & \text{pick random } V, U \text{ s.t. } |V|=k,\ |U| = \frac{n-k}{2} \end{cases}$

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

(P) (N) (P)

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

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

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

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

$F_P \leftarrow \{ \ell \text{ s.t. } A \text{ outputs PASS} \}$

$F_F \leftarrow \{ \ell \text{ s.t. } A \text{ outputs FAIL} \}$

Total error of A on D

$$= \sum_{\ell \in F_P} \Pr_{w \in D}[w \in E^-(\ell)] + \sum_{\ell \in F_F} \Pr_{w \in D}[w \in E^+(\ell)]$$

- $\ell \in F_P$: reach PASS leaf
- $w \in E^-(\ell)$: should FAIL
- $\ell \in F_F$: reach FAIL leaf
- $w \in E^+(\ell)$: should PASS

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

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

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

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

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

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

↘↗ but each leaf only has one label!

Putting them together to prove full theorem

error of A on D

$$= \sum_{\ell \in F_p} \Pr_{w \in D}[w \in E^-(\ell)] + \sum_{\ell \in F_n} \Pr_{w \in D}[w \in E^+(\ell)]$$

$$\geq \sum_{\ell \in F_p} \left(\tfrac{1}{2} - o(1)\right) 2^{-t} + \sum_{\ell \in F_n} \left(\tfrac{1}{2} - o(1)\right) 2^{-t}$$

$$\geq \tfrac{1}{2} - o(1) \quad \leftarrow \text{since } |F_p| + |F_n| = 2^t$$

∎

Pf of Claim 1:

Idea N is close to U + U ends up uniformly distributed at each leaf

$$|L_n| \leq 2^{n/2} \cdot \frac{n}{2}$$

↑ choice of u, v ↑ choice of i

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

so $E^-(\ell) \geq 2^{n-t} - 2^{\frac{n}{2} + 2\varepsilon \log\left(\frac{1}{\varepsilon}\right) n} = \left(1 - o(1)\right) 2^{n-t}$

2^{n-t} = # strings that follow path to leaf

$2^{\frac{n}{2} + 2\varepsilon \log\left(\frac{1}{\varepsilon}\right) n}$ = # words at dist $\leq \varepsilon$

assume $\varepsilon \ll 1/8$

t is $o(n)$

So 1st term swamps 2nd term

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

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

∎

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.

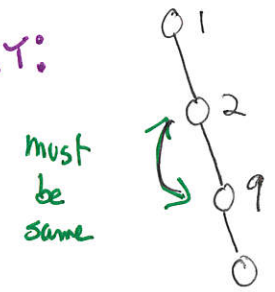
Count # strings that agree with t queries in leaf?

$= 2^{n-t}$

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

$= (2^{n-t}) - ?$

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 $\frac{n}{6} \ldots \frac{n}{3}$

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

Finish this on homework!

II Communication Complexity (CC) lower bounds

$\Rightarrow$ Property testing (PT) lower bounds

Idea give reduction from CC problem to PT problem

$\Rightarrow$ L.B. for C.C. problem yields L.B. for P.T. problem

(C.C. ← a lot of great work done in this area)

(P.T. ← so we get this almost for free!!)

Example:

- A hard C.C. problem

SET DISJOINTNESS

Alice

$X \in \{0,1\}^n$

Bob

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

$$Disj(x,y) = \bigvee_{i=1}^{n} (x_i \wedge y_i)$$

do A+B agree on any bit?

Known l.b.: $\Omega(n)$ bits of communication required to solve it.

even if allow many rounds, probabilistic protocols

Sparse Set disjointness: A+B have at most k 1's

needs $\Omega(k)$ bits communication

(even if guaranteed that intersect only once or not at all)

Property testing problem of "unknown" complexity:

k-linearity

def. $f: \{0,1\}^n \to \{0,1\}$ is k-linear if it is a parity of k variables

ie. $f(x) = x_{i_1} \oplus x_{i_2} \oplus \dots \oplus x_{i_k}$ for $S = \{i_1 \dots i_k\}$, $|S| = k$

def PT for k-linearity:

- if f is k-linear, PT passes
- if f is ε-far from k-linear, PT fails

} with prob $\geq 2/3$

ε-far: $\nexists g$ s.t. g is k-linear $\wedge$ $\frac{\#\{x \mid f(x) \neq g(x)\}}{2^n} \leq \varepsilon$

The reduction:

A communication game

Shared Randomness

Alice

set A { n bit vector $\{0,1\}^n$ with exactly k 1's describing k-linear fctn f (ie. f is XOR of bits with indices in A)

Bob

n bit vector $\{0,1\}^n$ with k 1's describing k-linear fctn g } B

Question:

does $h = f \oplus g$

have $2k$- linearity property?

note:

if $A \wedge B = \phi$ then h is $2k$-linear

if $A \wedge B \neq \phi$ then h is j-linear

for $j \leq 2k-2$.

e.g. if $A = \{x_1, x_2\}$ $B = \{x_3, x_4\}$

$A \wedge B = \phi$

$f = x_1 \oplus x_2$ $g = x_3 \oplus x_4$

$h = x_1 \oplus x_2 \oplus x_3 \oplus x_4$ ← 4 linear

if $A = \{x_1, x_2\}$ $B = \{x_2, x_3\}$

$A \wedge B = \{x_2\}$

$f = x_1 \oplus x_2$ $g = x_2 \oplus x_3$

$h = x_1 \oplus \underbrace{x_2 \oplus x_2}_{=1} \oplus x_3$

$= x_1 \oplus x_3$ ← 2 linear

for all x_i in $A \wedge B$,

two variables drop out of h

so h is $(k - 2|A \wedge B|)$-linear

Fact if $h_1 \neq h_2$ are 2 linear fctns (for any k)

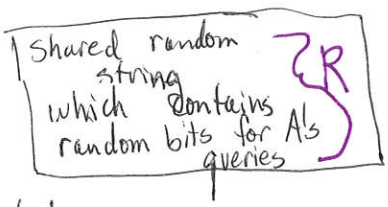
then $\frac{\#x \text{ s.t. } h_1(x) \neq h_2(x)}{2^n} = \frac{1}{2}$

we will prove this in a few weeks

$\Rightarrow$ if $A \cap B \neq \emptyset$, h is $\frac{1}{2}$-far from $2k$-linear

Why is this interesting?

protocol for testing $2k$-linearity of h with q queries $\Rightarrow$ C.C. protocol for set disjointness of A, B



Bob simulates A's run on R. Bob computes x & then g(x)

A runs prop test alg. When needs $h(x) = f(x) \oplus g(x)$:

1) compute $f(x)$
2) ask Bob for $g(x)$
3) output $f(x) \oplus g(x)$ as $h(x)$

what is answer to my next question? $g(x)$

$g(x)$

Note: Alice doesn't need to sent x's

Total communication $= 2q$ bits

$\Rightarrow q = \Omega(k)$

Thm k-linearity testing requires $\Omega(k)$ queries!

Interesting, since linearity testing only needs $O(1)$!