Lower Bounds for property testing algorithms

I. Deterministic lower bounds \( \Rightarrow \) probabilistic lower bounds

a difficulty:

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

useful lower bound tool:

Yao's principle:

If there is a probability distribution \( D \) on union of "positive" and "negative" elements of domain, such that any deterministic algorithm of query complexity \( t \) is incorrect with prob \( \geq \frac{1}{3} \) for inputs chosen according to \( D \), then it is a lower bound on randomized query complexity.

So, average case deterministic lower bound \( \Rightarrow \) 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 \( x \)

\( \text{Von Neumann's minimax } \Rightarrow \text{Bob has randomized strategy} \)

do, as well when \( A \) randomized

**An example:**

\[ \mathcal{L}_n = \{ w | w \text{ is } n \text{-bit string} \} \]

\[ w = v v^R \cup w^R \]

concatenations of palindromes

**Thm**  need \( \Omega(\sqrt{n}) \) queries to properly test \( \mathcal{L}_n \)

i.e., if \( A \) satisfies

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

\[ \forall x \text{ \& far from } \mathcal{P}, \Pr[A(x) = \text{PASS}] \leq \frac{2}{3} \]

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

**Pf.**

Plan: give distribution on inputs that is hard for all algorithms with \( O(\sqrt{n}) \) queries


- Wlog assume \( 6/n \)
- distribution on negative inputs:

\[ N = \text{random string of distance } \geq n \text{ from } \mathcal{L}_n \]

\( \text{Yao } \Rightarrow \text{randomized ub. of } \Omega(\sqrt{n}) \)

\( \text{should output } \text{FAIL} \)
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?

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

Main Difficulty:

so maybe no string infinitely follows the path?

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

so for each set of queries, some $k$'s (but not all) are bad
distribution on positive inputs:

\[ P = \begin{cases} 
1. \text{ pick } k \in \left[ \frac{n}{6}, \frac{n}{3} \right] \\
2. \text{ pick random } V, u \text{ st. } |v| = k \\
3. \text{ output } V^*uu^* 
\end{cases} \]

an issue:
some strings can be generated via \( \geq 1 \) \( k \)

\[ \text{distribution } D = \begin{cases} 
\text{flip coin} \\
\text{if } H \text{ output according to } N \\
\text{else} \quad \quad \quad \quad \quad \quad \quad \text{P} 
\end{cases} \]

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

\[ \text{depth } t, \leq 2^t \text{ root-leaf paths} \]

\[ \text{wlog all leaves have depth } t \]

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

Note: we can calculate probability of reaching a leaf since we know input distribution.

if a input reaches here, hopefully it is a "false" input?
For each leaf \( \ell \):
\[
E^-(\ell) = \sum w \in \{0,1\}^n \quad \text{dist}(w, \ell) = \varepsilon n, \quad w \text{ reaches leaf } \ell.
\]

\( w \text{ should fail} \)

\[
E^+(\ell) = \sum w \in \{0,1\}^n \setminus L, \quad w \text{ reaches leaf } \ell.
\]

\( w \text{ should Pass} \)

each leaf \( \ell \) is either passing or failing, not both

Total error of \( A \) on \( D \)
\[
= \sum_{\ell \text{ passing}} \Pr_D \left[ w \in E^-(\ell) \right] + \sum_{\ell \text{ failing}} \Pr_D \left[ w \in E^+(\ell) \right]
\]

Claim 1: if \( t = o(n) \), \( \forall \ell \) at depth \( t \)
\[
\Pr_D \left[ w \in E^-(\ell) \right] \geq \left( \frac{1}{2} - o(1) \right) 2^{-t}
\]

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

Claim 2: if \( t = o(n) \), \( \forall \ell \) at depth \( t \)
\[
\Pr_D \left[ w \in E^+(\ell) \right] \geq \left( \frac{1}{2} - o(1) \right) 2^{-t}
\]

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

but each leaf only has one label!
Putting them together to prove full theorem

error of \( A \) on \( D \)

\[
\begin{align*}
\mathbb{E} &= \sum_{l \text{ passing}} \Pr_{w \in D} [ w \in E(l) ] + \sum_{l \text{ failing}} \Pr_{w \in D} [ w \in E^+(l) ] \\
&\geq 2^{\left(\frac{1}{2} - o(1)\right)} 2^{-t} + \left(\frac{1}{2} - o(1)\right) 2^{-t} \\
&= \frac{1}{2} - o(1) \quad \leftarrow \text{since all leaves pass or fail}
\end{align*}
\]

Pf of Claim 1:

Idea

\( N \) is close to \( U \)

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

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

\[ |L_n| = 2^n, \frac{n}{2} \]

choice of \( U \)

choice of \( N \)

\# words at distance \( \leq \varepsilon \):

\[ 2^n \cdot \frac{n}{2} \cdot \sum_{i=0}^{\varepsilon n} \binom{n}{i} \leq 2^{n/2 + 2\varepsilon \log(2)n} = (1 - o(1)) 2^n \]

so \( E^-(l) \geq 2^{-n} - 2^{n/2 + 2\varepsilon \log(2)n} = (1 - o(1)) 2^{-n} \)

\# words at dist \( \leq \varepsilon \)

\# words at dist \( \varepsilon \)

\# strings that follow path to leaf

assume \( \varepsilon \ll \frac{1}{n} \)

\( \varepsilon \) is \( o(1) \)

So 1st term swamps 2nd term

so \( \Pr_{D} [ w \in E^{-}(l) ] = \frac{1}{2} \Pr_{N} [ w \in E^{-}(l) ] \)

\[ \geq \frac{1}{2} \frac{|E^{-}(l)|}{2^n} \geq \left(\frac{1}{2} - o(1)\right) 2^{-t} \]
Given list $k$, let $Q_k$ indices queried along the way. For each of $\binom{\frac{n}{2}}{2}$ pairs of queries $q_1, q_2 \in Q_k$ at most 2 choices of $k$ for which $q_1, q_2$ symmetric to $k$ or $\frac{n}{2} + k$.

In this case, only one choice.

$\Rightarrow \# \text{ choices of } k \text{ s.t. }$ 
no pair in $Q_k$ symmetric around $k$ or $\frac{n}{2} + k$

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

For these $k$, $\# \text{ strings that follow path } = 2^{\frac{n}{2} - 1} q_{x1}$

So $Pr_d[w \in E^+(q)] = \sum \sum_{u \in \mathbb{R}} \sum_{k \in \mathbb{R}} \sum_{\frac{m_{u,k}}{2}} \frac{1}{\sum_{\frac{m_{u,k}}{2}} \frac{b}{n}}$ 

$= \frac{1}{n \cdot 2^{n/2}} \left[ (-6(1)) \frac{n}{b} \right] \left[ 2^{\frac{n}{2} - |q_{x1}|} \right] = (1 - o(1)) 2^{-t}$

$\Rightarrow Pr_d[w \in E^{*}(q)] = (\frac{1}{2} - o(1)) 2^{-t}$

$\therefore$