Sublinear Algorithms

LECTURE 20

Last time

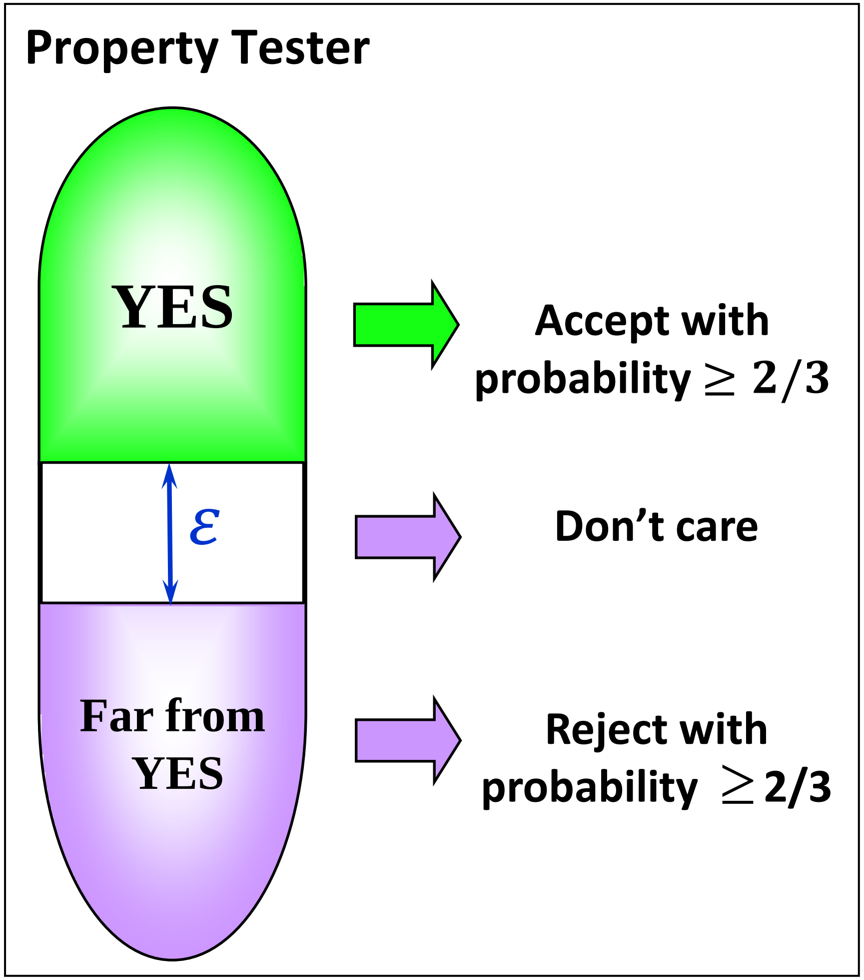
- Approximating the distance to sortedness (length of LIS) of 0/1 sequences

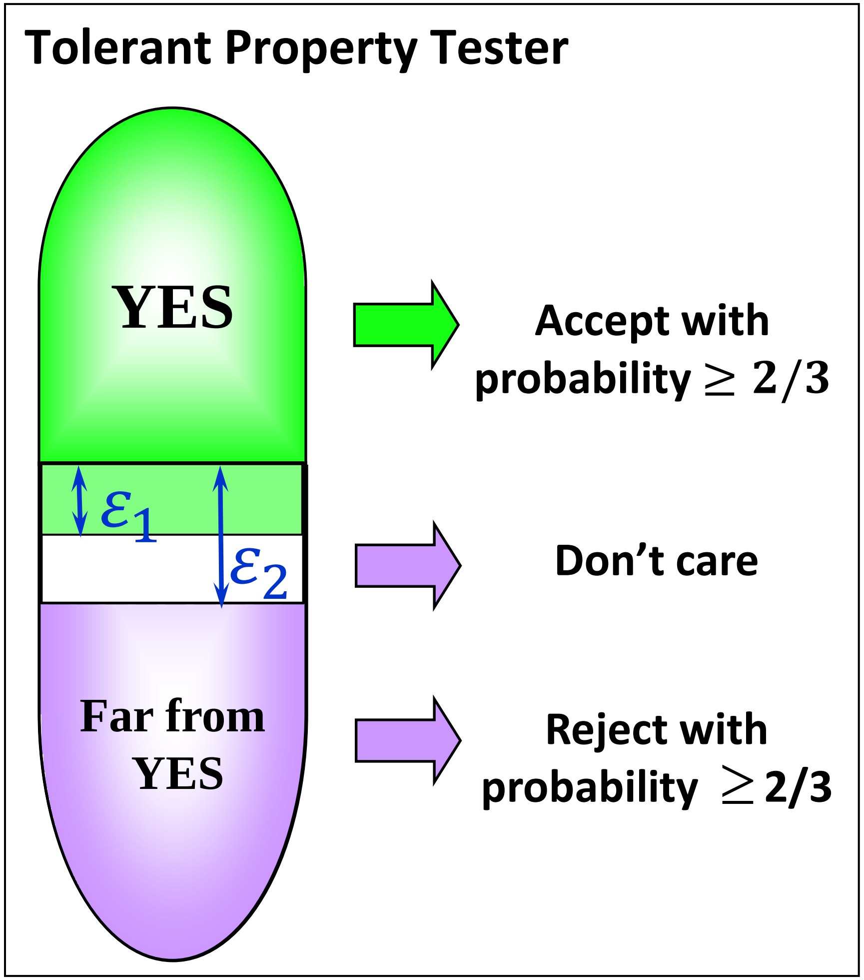
Today

- L_p-testing

Project Reports are due April 24

Testing Models





Two objects are at distance ε = they differ in an ε fraction of places

Equivalent problem: approximating distance to the property.

Why Hamming Distance?

- Nice probabilistic interpretation
  - probability that two functions differ on a random point in the domain
- Natural measure for
  - algebraic properties (linearity, low degree)
  - properties of graphs and other combinatorial objects
- Motivated by applications to probabilistically checkable proofs (PCPs)
- It is equivalent to other natural distances for
  - properties of Boolean functions

Which stocks grew steadily?

Data from
http://finance.google.com

L_p-Testing

for properties of real-valued data

[Berman Raskhodnikova Yaroslavtsev]

Use L_p-metrics to Measure Distances

- Functions $f, g\colon D \to [0,1]$ over (finite) domain D

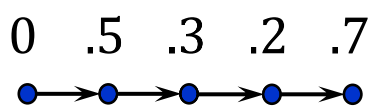
Normalize the values, so they are between 0 and 1

- For $p \geq 1$
$$L_p(f,g) = \left\|f - g\right\|_p = \left(\sum_{x \in D} |f(x) - g(x)|^p\right)^{1/p}$$

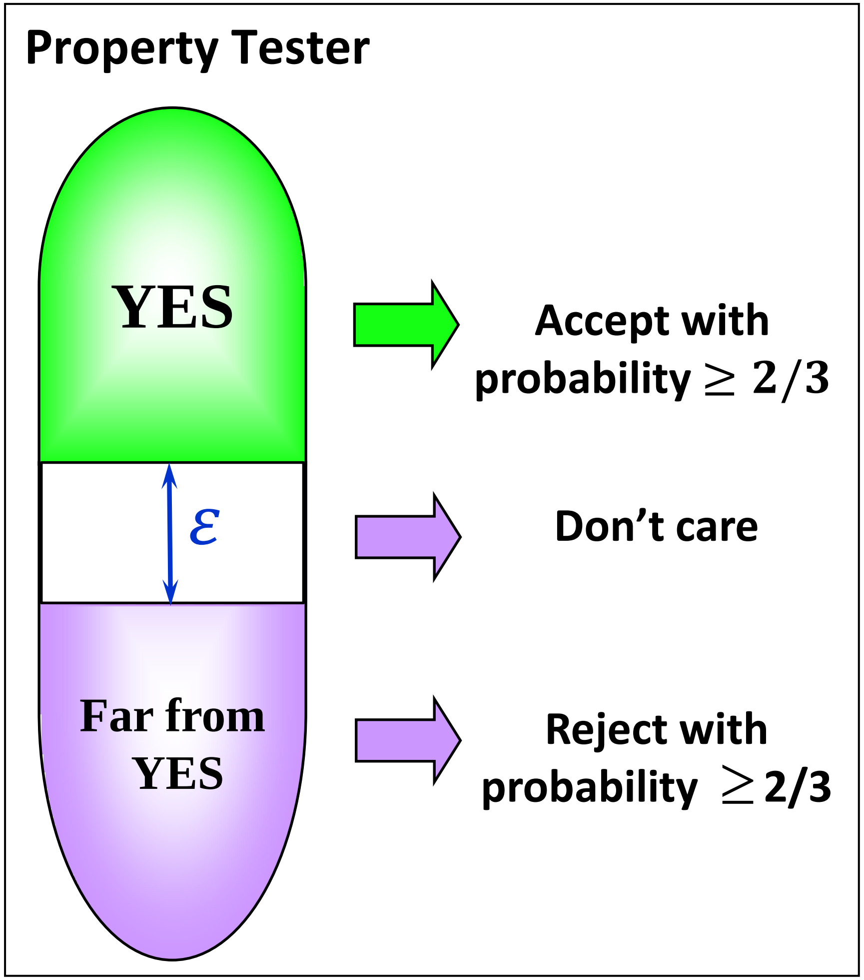
$$L_0(f,g) = \left\|f - g\right\|_0 = |\{x \in D : f(x) \neq g(x)\}|$$

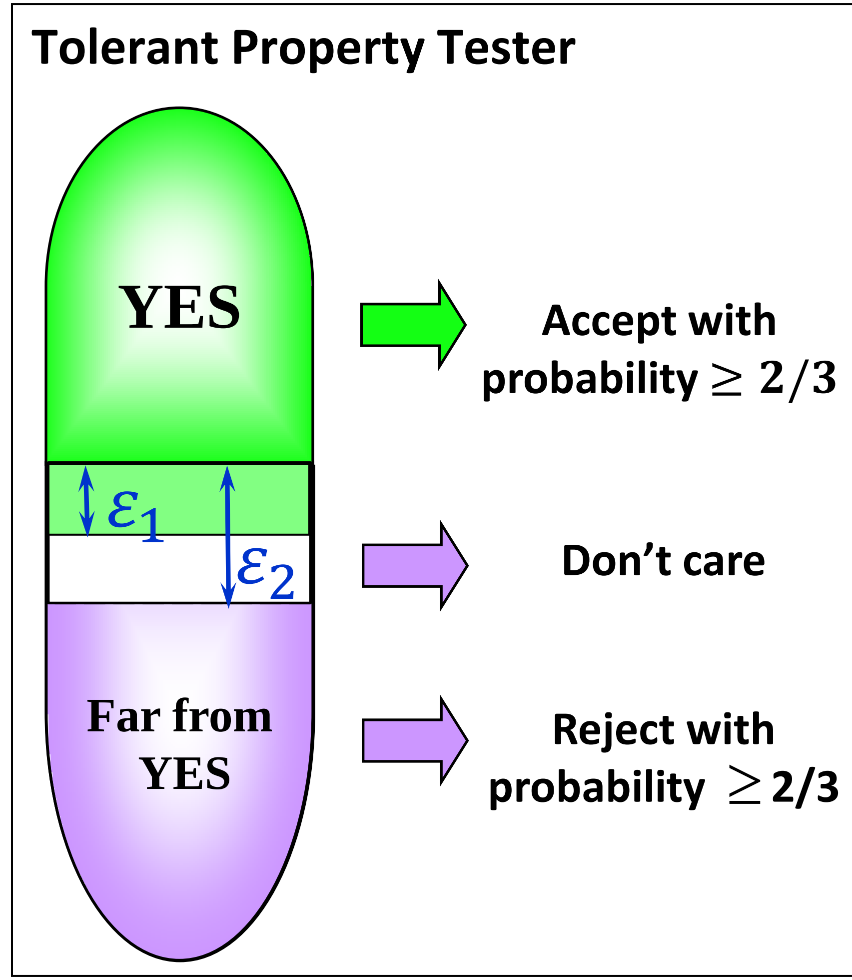
- $d_p(f,g) = \dfrac{\left\|f - g\right\|_p}{\left\|1\right\|_p}$

Example:



L_p-Testing and Tolerant L_p-Testing





Functions $f, g: D \to [0,1]$ are at distance ε if $d_p = \frac{\|f-g\|_p}{\|\mathbf{1}\|_p} = \varepsilon$.

L_p-Testing Model for Real-Valued Data

- Generalizes standard L_0-testing

- For $p > 0$ still have a nice probabilistic interpretation: distance $d_p(f, g) = (\mathbb{E}_{\boldsymbol{x}}[|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{x})|^{\boldsymbol{p}}])^{1/p}$

- Compatible with existing PAC-style learning models (preprocessing for model selection)

- For Boolean functions, $d_0(f, g) = d_p(f, g)^p$.

Plan

1. Relationships between L_p-testing models
2. L_p-testing monotonicity

Relationships between L_p-Testing Models

Relationships Between L_p-Testing Models

$C_p(P,\varepsilon)$ = complexity of L_p-testing property P with distance parameter ε

- e.g., query or time complexity
- for general or restricted (e.g., nonadaptive) tests

For all properties P

- L_1-testing is no harder than Hamming testing

$$C_1(P,\varepsilon) \le C_0(P,\varepsilon)$$

- L_p-testing for $p > 1$ is close in complexity to L_1-testing

$$C_1(P,\varepsilon) \le C_p(P,\varepsilon) \le C_1(P,\varepsilon^p)$$

Relationships Between L_p-Testing Models

$C_p(P, \varepsilon)$ = complexity of L_p-testing property P with distance parameter ε

- e.g., query or time complexity
- for general or restricted (e.g., nonadaptive) tests

For properties of Boolean functions $f : D \to \{0,1\}$

- L_1-testing is equivalent to Hamming testing

$$C_1(P, \varepsilon) = C_0(P, \varepsilon)$$

- L_p-testing for $p > 1$ is equivalent to L_1-testing with appropriate distance parameter

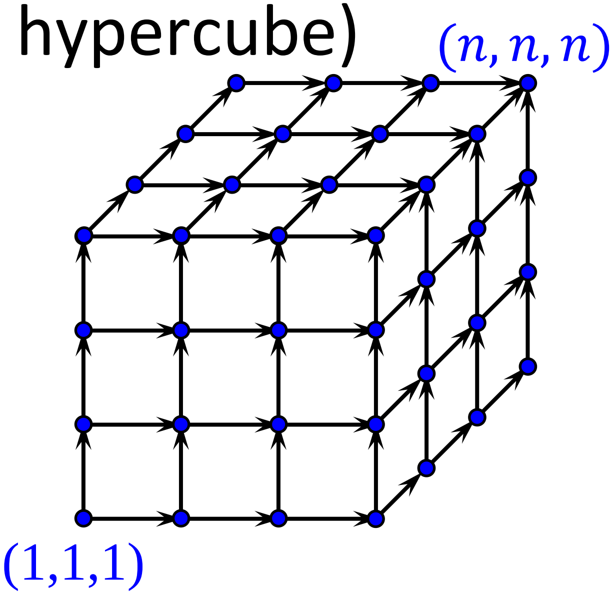
$$C_p(P, \varepsilon) = C_1(P, \varepsilon^p)$$

Property: *Monotonicity of Functions*

Monotonicity

- Domain D=$[n]^d$ (vertices of d-dim hypercube)
- A function $f: D \to \mathbb{R}$ is monotone

  if increasing a coordinate of x does not decrease $f(x)$.



- Special case $d = 1$

  $f: [n] \to \mathbb{R}$ is monotone $\Leftrightarrow f(1), \ldots f(n)$ is sorted.

Monotonicity Testers: Running Time

f	L_0	L_p
$[n] \to [0,1]$	$\Theta\left(\frac{\log n}{\varepsilon}\right)$ [Ergün Kannan Kumar Rubinfeld Viswanathan 00, Fischer 04, Belovs, Chakrabarty Seshadhri]	$\Theta\left(\frac{1}{\varepsilon^p}\right)$
$[n]^d \to [0,1]$	$\Theta\left(\frac{d \cdot \log n}{\varepsilon}\right)$ [Chakrabarty Seshadhri 13]	$O\left(\min\left\{\frac{d}{\varepsilon^p}\log\frac{d}{\varepsilon^p}, \frac{d^{1/2+o(1)}}{\varepsilon^{2p}}\right\}\right)$ $\Omega\left(\frac{1}{\varepsilon^p}\log\frac{1}{\varepsilon^p}\right)$ for $d = 2$ nonadaptive 1-sided error [Berman Raskhodnikova Yaroslavtsev 14, Black Chakrabarty Seshadhri 23]

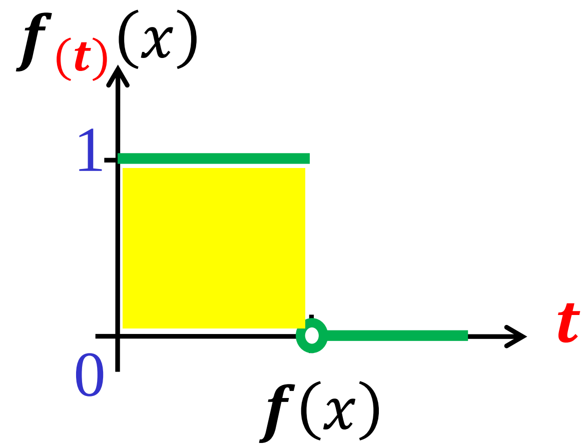
* Hiding some $\log 1/\varepsilon$ dependence

L_1-Testing of Monotonicity

Monotonicity: Reduction to Boolean Functions

Boolean threshold function $\boldsymbol{f}_{(\boldsymbol{t})}: D \to \{0,1\}$

$$\boldsymbol{f}_{(\boldsymbol{t})}(x) = \begin{cases} 1 & \text{if } \boldsymbol{f}(x) \geq t \\ 0 & \text{otherwise} \end{cases}$$

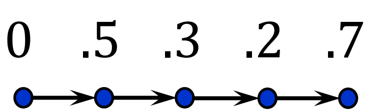


- Decomposition: $f(x) = \int_0^1 \boldsymbol{f}_{(\boldsymbol{t})}(x)\, d\boldsymbol{t}$
- M = class of monotone functions

Characterization Theorem

$$L_1(\boldsymbol{f}, M) = \int_0^1 L_1\left(\boldsymbol{f}_{(\boldsymbol{t})}, M\right) d\boldsymbol{t}$$

Example:



Characterization Theorem: One Direction

$$L_1(\boldsymbol{f}, M) \le \int_0^1 L_1\left(\boldsymbol{f}_{(\boldsymbol{t})}, M\right) d\boldsymbol{t}$$

- $\forall t \in [0,1]$, let g_t=closest monotone (Boolean) function to $\boldsymbol{f}_{(\boldsymbol{t})}$.
- Let $\boldsymbol{g} = \int_0^1 g_t d\boldsymbol{t}$. Then $\boldsymbol{g}$ is monotone, since g_t are monotone.

$$
\begin{aligned}
L_1(\boldsymbol{f}, M) &\le \|\boldsymbol{f} - \boldsymbol{g}\|_1 && \boldsymbol{g} \text{ is monotone} \\
&= \left\| \int_0^1 \boldsymbol{f}_{(\boldsymbol{t})} d\boldsymbol{t} - \int_0^1 g_t d\boldsymbol{t} \right\|_1 && \text{Decomposition \& definition of } \boldsymbol{g} \\
&= \left\| \int_0^1 (\boldsymbol{f}_{(\boldsymbol{t})} - g_t) d\boldsymbol{t} \right\|_1 \\
&\le \int_0^1 \left\| \boldsymbol{f}_{(\boldsymbol{t})} - g_t \right\|_1 d\boldsymbol{t} && \text{Triangle inequality} \\
&= \int_0^1 L_1\left(\boldsymbol{f}_{(\boldsymbol{t})}, M\right) d\boldsymbol{t} && \text{Definition of } g_t
\end{aligned}
$$

Characterization Theorem: the Other Direction

$$L_1(\boldsymbol{f}, M) \geq \int_0^1 L_1\left(\boldsymbol{f}_{(\boldsymbol{t})}, M\right) d\boldsymbol{t}$$

- Let $\boldsymbol{h}$ be closest monotone function to $\boldsymbol{f}$.
- Then $\boldsymbol{h}_{(\boldsymbol{t})}$ is monotone for all $t \in [0,1]$.

$$L_1(\boldsymbol{f}, M) = \|\boldsymbol{f} - \boldsymbol{h}\|_1$$

Definition of $\boldsymbol{h}$

$$= \left\| \int_0^1 \left(\boldsymbol{f}_{(\boldsymbol{t})} - \boldsymbol{h}_{(\boldsymbol{t})}\right) d\boldsymbol{t} \right\|_1$$

Decomposition

$\boldsymbol{f}(x) \geq \boldsymbol{h}(x)$
$\Leftrightarrow$
$\boldsymbol{f}_{(\boldsymbol{t})} \geq \boldsymbol{h}_{(\boldsymbol{t})}$
$\forall t \in [0,1]$

$$= \sum_{x:\boldsymbol{f}(x) \geq \boldsymbol{h}(x)} \int_0^1 \left(\boldsymbol{f}_{(\boldsymbol{t})} - \boldsymbol{h}_{(\boldsymbol{t})}\right) d\boldsymbol{t} + \sum_{x:\boldsymbol{f}(x) < \boldsymbol{h}(x)} \int_0^1 \left(\boldsymbol{h}_{(\boldsymbol{t})} - \boldsymbol{f}_{(\boldsymbol{t})}\right) d\boldsymbol{t}$$

$$= \int_0^1 \left(\sum_{x:\boldsymbol{f}(x) \geq \boldsymbol{h}(x)} \left(\boldsymbol{f}_{(\boldsymbol{t})} - \boldsymbol{h}_{(\boldsymbol{t})}\right) + \sum_{x:\boldsymbol{f}(x) < \boldsymbol{h}(x)} \left(\boldsymbol{h}_{(\boldsymbol{t})} - \boldsymbol{f}_{(\boldsymbol{t})}\right) \right) d\boldsymbol{t}$$

$$= \int_0^1 \left\|\boldsymbol{f}_{(\boldsymbol{t})} - \boldsymbol{h}_{(\boldsymbol{t})}\right\|_1 d\boldsymbol{t}$$

all terms are nonnegative

$$\geq \int_0^1 L_1\left(\boldsymbol{f}_{(\boldsymbol{t})}, M\right) d\boldsymbol{t}$$

$\boldsymbol{h}_{(\boldsymbol{t})}$ is monotone

Monotonicity: Using Characterization Theorem

Characterization Theorem

$$d_1(\boldsymbol{f}, M) = \int_0^1 d_1\left(\boldsymbol{f}_{(\boldsymbol{t})}, M\right) d\boldsymbol{t}$$

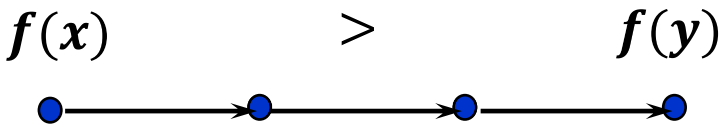
We can use Characterization Theorem
to get monotonicity L_1-testers
and tolerant testers from
standard property testers for Boolean functions.

L_1-Testers from Testers for Boolean Ranges

A nonadaptive, 1-sided error L_0-test for monotonicity of $f: D \to \{0,1\}$ is also an L_1-test for monotonicity of $f: D \to [0,1]$.

Proof:

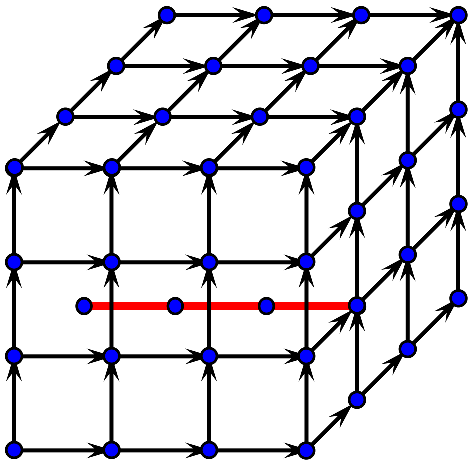
- A violation (x, y):



- A nonadaptive, 1-sided error test queries a random set $Q \subseteq D$ and rejects iff Q contains a violation.
- If $f: D \to [0,1]$ is monotone, Q will not contain a violation.
- If $d_1(f, M) \geq \varepsilon$ then $\exists \boldsymbol{t^*}: d_0\left(\boldsymbol{f}_{(\boldsymbol{t^*})}, M\right) \geq \boldsymbol{\varepsilon}$
- W.p. $\geq 2/3$, set Q contains a violation (x, y) for $\boldsymbol{f}_{(\boldsymbol{t^*})}$

$$\boldsymbol{f}_{(\boldsymbol{t^*})}(x) = 1, \boldsymbol{f}_{(\boldsymbol{t^*})}(y) = 0$$
$$\Downarrow$$
$$\boldsymbol{f}(x) > \boldsymbol{f}(y)$$

L_0-Testing Monotonicity of $f: [n]^d \to \{0,1\}$



Idea: 1. Pick axis-parallel lines ℓ.

2. Sample points from each ℓ,

and check for violations of $f_{|\ell}$.

[DGLRRS 99]

- Testing sortedness: If $f: [n] \to \{0,1\}$ is ε-far from sorted then $O\left(\frac{1}{\varepsilon}\right)$ samples are sufficient to find a violation w/ const. prob.
- Dimension reduction: For $f: [n]^d \to \{0,1\}$

$$\mathbb{E}\left[d_0\left(f_{|\ell}, M\right)\right] \geq \frac{d_0(f, M)}{2d}.$$

How many lines should we sample?

How many points form each line?

General Work Investment Problem [Goldreich 13]

- Algorithm needs to find ``evidence'' (e.g., **a violation**).
- It can select an element from distr. Π (e.g., **a uniform line**).
- Elements e have different quality $q(e) \in [0,1]$

  (e.g., $\boldsymbol{d_0(f_{|\ell}, M)}$).
- Algorithm must invest more work into e with lower $q(e)$ to extract evidence from e (e.g., **need $\boldsymbol{\Theta\left(\frac{1}{q(e)}\right)}$ samples**).
- $\mathbb{E}_{\boldsymbol{e}\leftarrow\boldsymbol{\Pi}}[q(e)] \geq \mu.$

What's a good work investment strategy?

Used in [Levin 85, Goldreich Levin 89], testing connectedness of a graph [Goldreich Ron 97], testing properties of images [R 03], multi-input testing problems [G13]

Work Investment Strategies

- ``Reverse'' Markov Inequality

For a random variable $X \in [0,1]$ with expectation $\mathbb{E}[X] \geq \mu$,

$$\Pr\left[X \geq \frac{\mu}{2}\right] \geq \frac{\mu}{2}.$$

Proof: $\mu \leq \mathbb{E}[X] \leq \Pr\left[X \geq \frac{\mu}{2}\right] \cdot 1 + \Pr\left[X < \frac{\mu}{2}\right] \cdot \frac{\mu}{2}.$

``Reverse'' Markov Strategy:

1. Sample $\Theta\left(\frac{1}{\mu}\right)$ lines.
2. Sample $\Theta\left(\frac{1}{\mu}\right)$ points from each line.

Cost: $\Theta\left(\frac{1}{\mu^2}\right)$ queries.

Work Investment Strategies

Bucketing idea [Levin, Goldreich 13]:

Invest in elements of quality $q(e) \geq \frac{1}{2^i}$ separately.

Bucketing Inequality [Berman R Yaroslavtsev 14]

For a random variable $X \in [0,1]$ with $\mathbb{E}[X] \geq \mu$, let

$$p_i = \Pr\left[X \geq \frac{1}{2^i}\right] \text{ and } k_i = \Theta\left(\frac{1}{2^i \mu}\right) \text{ for all } i \in \left[\log \frac{4}{\mu}\right].$$

Then $\prod_{i=1}^{\log 4/\mu} (1 - p_i)^{k_i} \leq 1/3$.

Bucketing Strategy: For each bucket $i \in \left[\log \frac{4}{\mu}\right]$

1. Sample $k_i = \Theta\left(\frac{1}{2^i \mu}\right)$ lines.
2. Sample $\Theta(2^i)$ points from each line.

Cost: $\Theta\left(\frac{1}{\mu} \log \frac{1}{\mu}\right)$ queries (**for monotonicity, $\mu = \frac{\varepsilon}{2d}$**)

Proof of Bucketing Inequality

Bucketing Inequality [Berman R Yaroslavtsev 14]

For a random variable $X \in [0,1]$ with $\mathbb{E}[X] \geq \mu$, let

$$t = \log \frac{4}{\mu}, \qquad p_i = \Pr\left[X \geq \frac{1}{2^i}\right], \quad \text{and } k_i = \Theta\left(\frac{1}{2^i \mu}\right).$$

Then $\prod_{i=1}^{t} (1 - p_i)^{k_i} \leq \delta$.

Proof: It suffices to prove $\sum_{i \in [t]} \frac{p_i}{2^i} \geq \frac{\mu}{4}$

Proof of Bucketing Inequality (Continued)

Bucketing Inequality [Berman R Yaroslavtsev 14]

For a random variable $X \in [0,1]$ with $\mathbb{E}[X] \geq \mu$, let

$$t = \log\frac{4}{\mu}, \qquad p_i = \Pr\left[X \geq \frac{1}{2^i}\right], \quad \text{and } k_i = \Theta\left(\frac{1}{2^i \mu}\right).$$

Then $\prod_{i=1}^{t}(1-p_i)^{k_i} \leq \delta$.

Proof: It suffices to prove $\sum_{i\in[t]}\frac{p_i}{2^i} \geq \frac{\mu}{4}$