Boolean Function Monotonicity Testing requires (almost) $\Omega(n^{1/2})$ queries

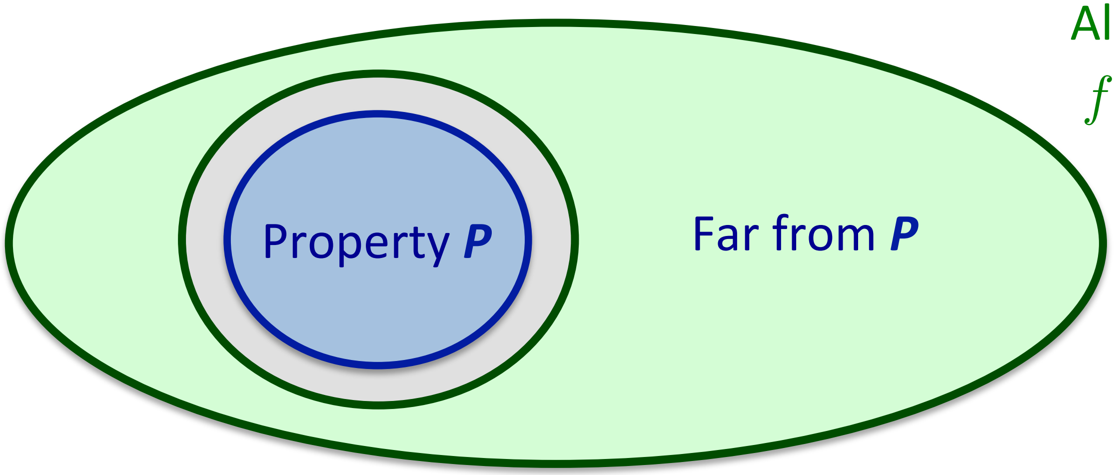
Xi Chen (Columbia), Anindya De (DIMACS and IAS),
Rocco Servedio (Columbia), and
Li-Yang Tan (TTI-Chicago)



STOC 2015
Portland, OR

Property Testing

Simplest question about a Boolean function: Does it have some property ***P***?



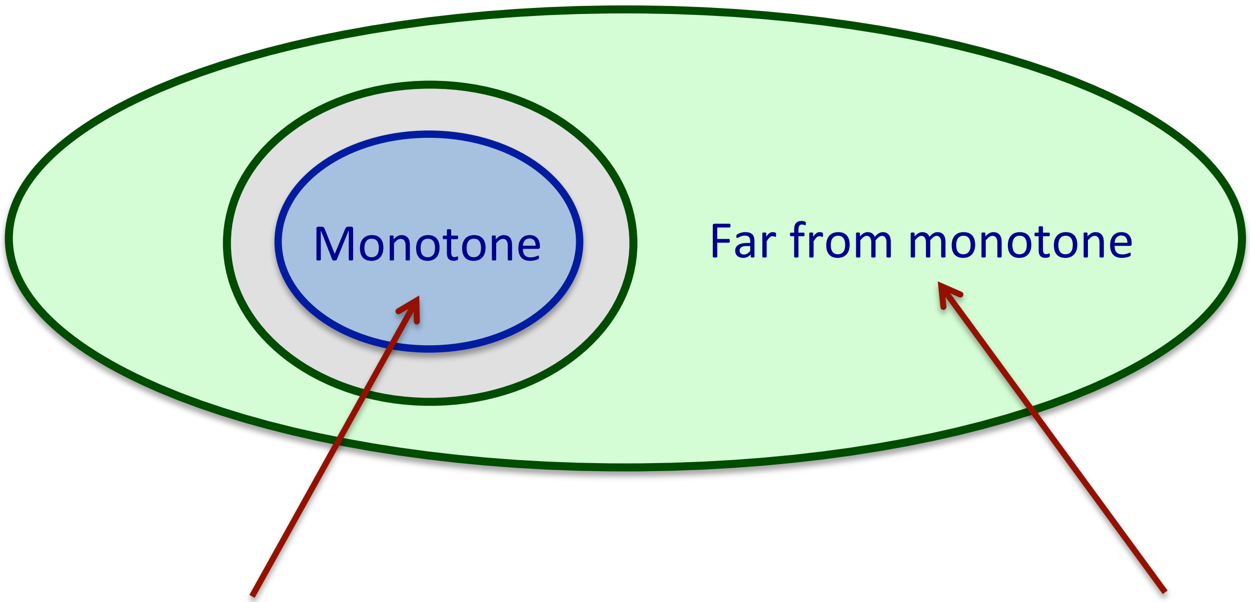
All Boolean functions
$f : \{0,1\}^n \to \{0,1\}$

Query access to unknown *f* on any input *x*

1. If *f* has Property ***P***, accept w.p. > 2/3.
2. If *f* is ***ε*-far** from having Property ***P***, reject w.p. > 2/3.
3. Otherwise: doesn't matter what we do.

Goal: minimize number of queries

This work: ***P* = monotonicity**



A monotone function is one that satisfies:

$$\forall\, x \preceq y,\ \ f(x) \leq f(y)$$

For all monotone functions *g:*

$$\Pr_{\boldsymbol{x} \in \{0,1\}^n}[f(\boldsymbol{x}) \neq g(\boldsymbol{x})] \geq \varepsilon$$

Well-studied problem:

[GGR98, GGL+98, DGL+99, FLN+02, HK08, BCGM12, RRS+12, BBM12, BRY13, CS13, …]

but still significant gaps in our understanding till recently

Previous work on non-adaptive testers

- Goldreich et al. [FOCS 1998, SICOMP 2000]
  - Introduced problem, gave "edge tester" with $O(n)$ query complexity
- Fischer et al. [STOC 2002]
  - Any tester must make $\Omega(\log n)$ queries
  - Also gave easy $\Omega(n^{1/2})$ lower bound for **one-sided** testers

11 years later…

Chakrabarty-Seshadhri [STOC 2013]

$O(n^{7/8})$-query tester

Chen-Servedio-T. [FOCS 2014]

$\Omega(n^{1/5})$ lower bound,
$O(n^{5/6})$-query tester

Khot-Minzer-Safra 2015:
$O(n^{1/2})$-query tester

This work:
$\Omega(n^{1/2-c})$ lower bound

Precise statement of lower bound

Theorem [Chen-De-Servedio-T. 2015]

For every $c > 0$ there is an $\varepsilon(c) > 0$ such that any non-adaptive algorithm for testing whether f is monotone or $\varepsilon(c)$-far from monotone requires $\Omega(n^{1/2-c})$ many queries.

Outline of this talk

- Sketch of approach in toy setting: 1-query lower bound
- Key ingredient in both [Chen-Servedio-T. 14] and this work:

  Multidimensional Central Limit Theorems

- Going beyond [CST14]: New ideas and ingredients

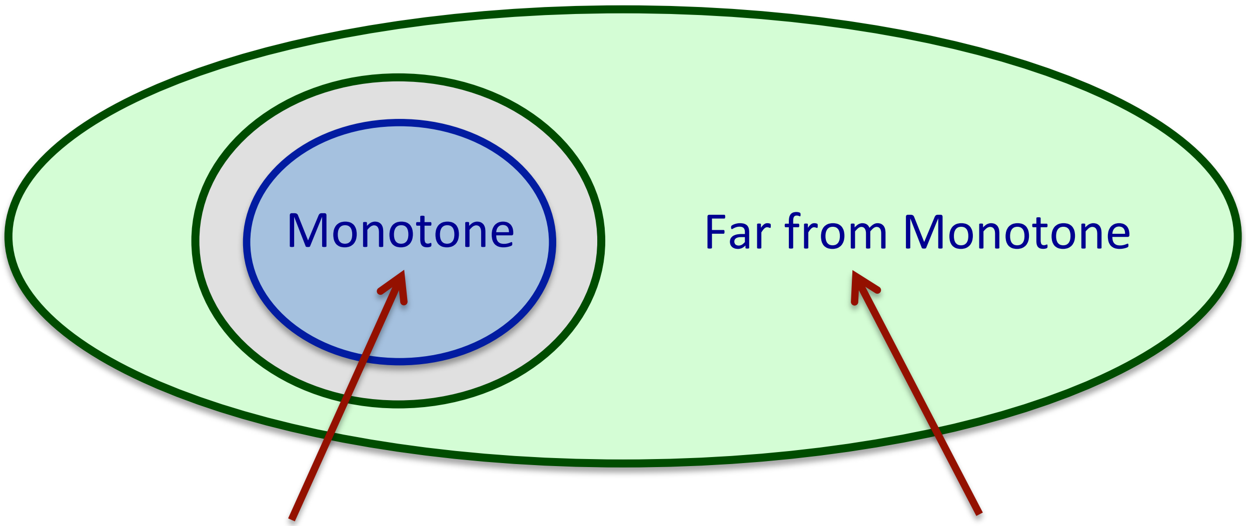
Yao's minimax principle

Lower bound against **randomized** algorithms

implies

Tricky distribution over inputs to **deterministic** algorithms

Yao's principle in our setting



Distribution D_{yes} supported on monotone functions

Distribution D_{no} supported on far from monotone functions

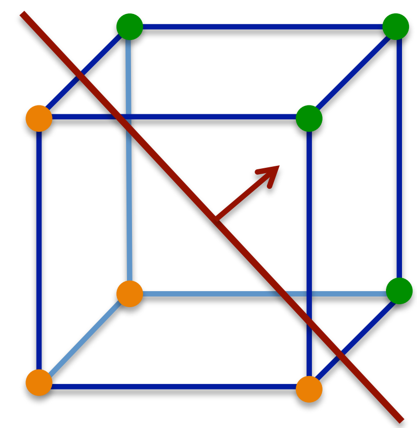
Indistinguishability. For all T = deterministic tester that makes $\boldsymbol{o(n^{1/2})}$ **queries,**

$$\left| \Pr_{\boldsymbol{f}_{yes} \sim \mathcal{D}_{yes}} \left[\mathcal{T} \ accepts \ \boldsymbol{f}_{yes} \right] - \Pr_{\boldsymbol{f}_{no} \sim \mathcal{D}_{no}} \left[\mathcal{T} \ accepts \ \boldsymbol{f}_{no} \right] \right| = o_n(1)$$

Our D_{yes} and D_{no} distributions

Both supported on *Linear Threshold Functions* (LTFs) over $\{-1,1\}^n$:

$$f(x) = \text{sign}(w_1 x_1 + \ldots + w_n x_n) \quad \vec{w} \in \mathbb{R}^n$$



D_{yes} : $\boldsymbol{\sigma_i}$ = uniform from {1,3}

D_{no} : $\boldsymbol{\nu_i}$ = −1 with prob 0.1, 7/3 with prob 0.9

Verify: D_{yes} LTFs are monotone, D_{no} LTFs far from monotone w.h.p.

Main Structural Result: **Indistinguishability**

Any deterministic tester that makes few queries cannot tell D_{yes} from D_{no}

Key property: $\mathbb{E}[\boldsymbol{\sigma_i}] = \mathbb{E}[\boldsymbol{\nu_i}]$, $\mathbf{Var}[\boldsymbol{\sigma}_i] = \mathbf{Var}[\boldsymbol{\nu}_i]$.

Indistinguishability: starting small

Claim. For all *T* = deterministic tester that makes **1 query**,

$$\left| \Pr_{\boldsymbol{f}_{yes} \sim \mathcal{D}_{yes}} \left[\mathcal{T} \ accepts \ \boldsymbol{f}_{yes} \right] - \Pr_{\boldsymbol{f}_{no} \sim \mathcal{D}_{no}} \left[\mathcal{T} \ accepts \ \boldsymbol{f}_{no} \right] \right| = o_n(1)$$

Non-trivial proof of a triviality

Claim. Let T = deterministic tester that makes 1 query **z**. Then:

$$\underbrace{\left| \Pr_{\boldsymbol{f}_{yes} \sim \mathcal{D}_{yes}} \left[\mathcal{T} \ accepts \ \boldsymbol{f}_{yes} \right] - \Pr_{\boldsymbol{f}_{no} \sim \mathcal{D}_{no}} \left[\mathcal{T} \ accepts \ \boldsymbol{f}_{no} \right] \right|}_{(*)} = o_n(1)$$

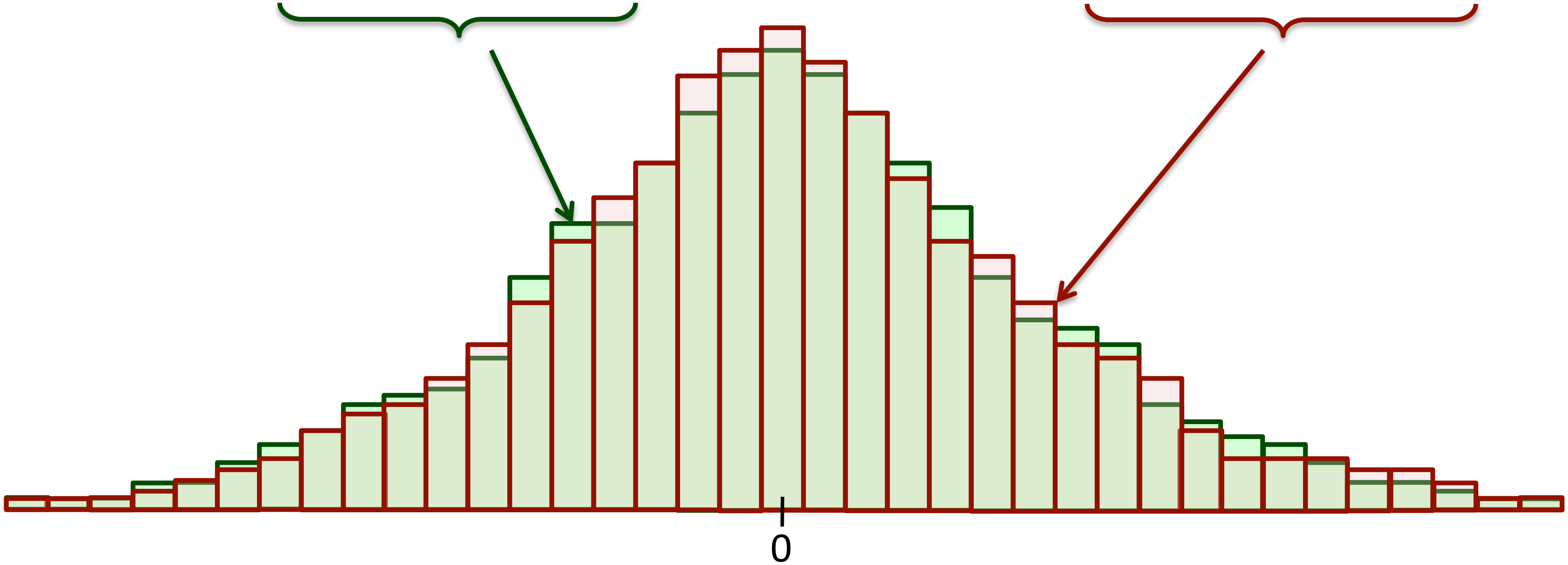
$$(*) \leq d_{\mathrm{TV}}(\mathbf{R}_{yes}, \mathbf{R}_{no})$$

Tester sees:

$$\mathbf{R}_{yes} = \mathrm{sign}(\boldsymbol{\sigma_1} z_1 + \cdots + \boldsymbol{\sigma_n} z_n) = \mathbf{1}[\underbrace{\boldsymbol{\sigma_1} z_1 + \cdots + \boldsymbol{\sigma_n} z_n} \geq 0]$$

vs.

$$\mathbf{R}_{no} = \mathrm{sign}(\boldsymbol{\nu_1} z_1 + \cdots + \boldsymbol{\nu_n} z_n) = \mathbf{1}[\underbrace{\boldsymbol{\nu_1} z_1 + \cdots + \boldsymbol{\nu_n} z_n} \geq 0]$$

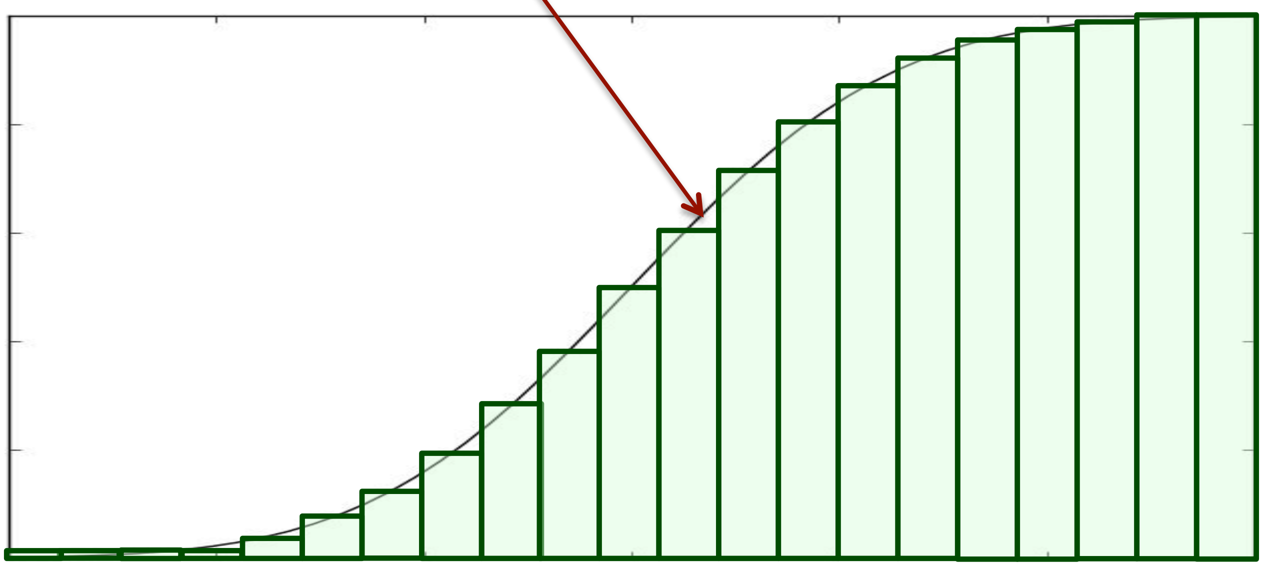


Central Limit Theorems. Sum of many independent "reasonable" random variables converges to Gaussian of same mean and variance.

Main analytic tool (Baby version):

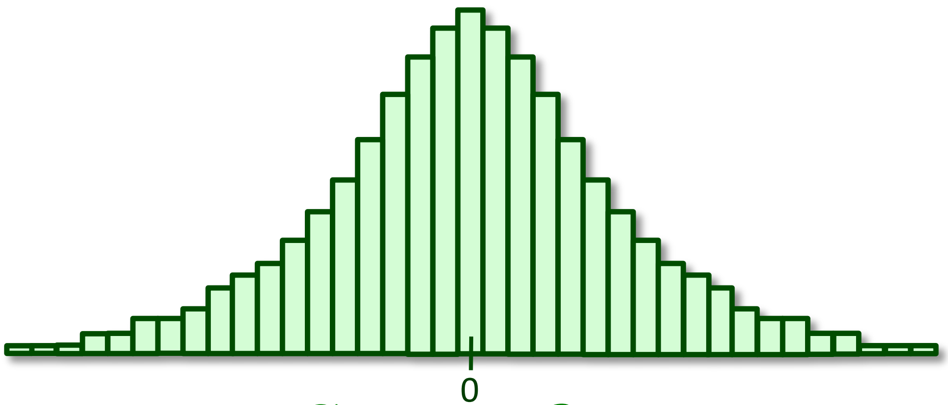
Berry–Esséen CLT. *Let* $\mathbf{S} = \mathbf{X}_1 + \cdots + \mathbf{X}_n$ *where* $\mathbf{X}_1, \ldots, \mathbf{X}_n$ *are independent real-valued random variables satisfying* $|\mathbf{X}_j - \mathbf{E}[\mathbf{X}_j]| \leq \tau$ *with probability* 1 *for all* $j \in [n]$. *Let* $\mathcal{G}$ *be a Gaussian with mean* $\mathbf{E}[\mathbf{S}]$ *and variance* $\mathbf{Var}[\mathbf{S}]$. *Then for all* $\theta \in \mathbb{R}$,

$$\left|\mathbf{Pr}[\mathbf{S} \leq \theta] - \mathbf{Pr}[\mathcal{G} \leq \theta]\right| \leq \frac{O(\tau)}{\mathbf{Var}[\mathbf{S}]^{1/2}}.$$

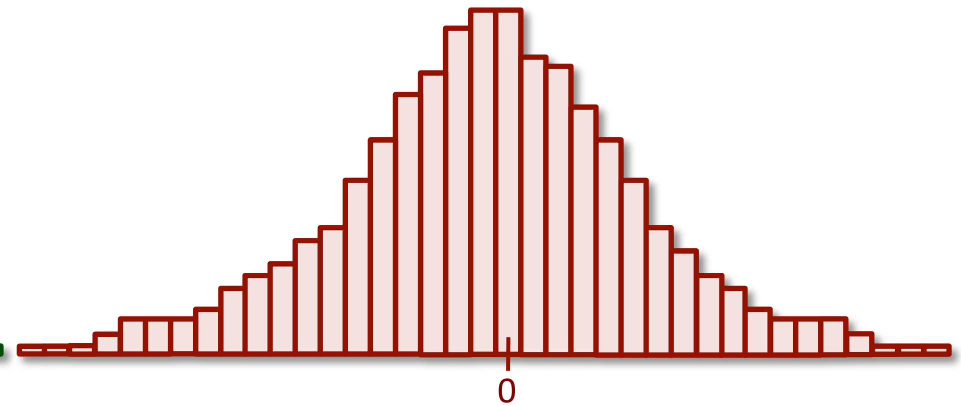


Goal: Upper bound $\left|\mathbf{Pr}[\mathbf{S}_{yes} \geq 0] - \mathbf{Pr}[\mathbf{S}_{no} \geq 0]\right|$

$$\mathbf{S}_{yes} = \boldsymbol{\sigma_1} z_1 + \cdots + \boldsymbol{\sigma_n} z_n \qquad \mathbf{S}_{no} = \boldsymbol{\nu_1} z_1 + \cdots + \boldsymbol{\nu_n} z_n$$



$\mathbf{S}_{yes} \to \mathcal{G}_1$



$\mathbf{S}_{no} \to \mathcal{G}_2$

Recall key property:

$$\left.\begin{aligned} \mathbb{E}[\boldsymbol{\sigma_i}] &= \mathbb{E}[\boldsymbol{\nu_i}] \\ \mathbf{Var}[\boldsymbol{\sigma}_i] &= \mathbf{Var}[\boldsymbol{\nu}_i] \end{aligned}\right\} \longrightarrow \begin{aligned} \mathbb{E}[\mathbf{S}_{yes}] &= \mathbb{E}[\mathbf{S}_{no}] \\ \mathbf{Var}[\mathbf{S}_{yes}] &= \mathbf{Var}[\mathbf{S}_{no}] \\ \mathcal{G}_1 &= \mathcal{G}_2 \end{aligned}$$

We just proved:

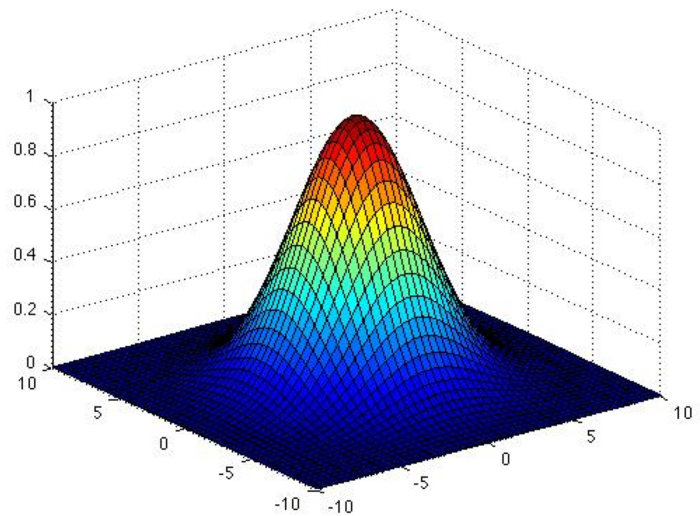
Claim. Let *T* = deterministic tester that makes **1 query**. Then:

$$\left| \underset{\boldsymbol{f}_{yes} \sim \mathcal{D}_{yes}}{\mathbf{Pr}} \left[\mathcal{T} \ accepts \ \boldsymbol{f}_{yes} \right] - \underset{\boldsymbol{f}_{no} \sim \mathcal{D}_{no}}{\mathbf{Pr}} \left[\mathcal{T} \ accepts \ \boldsymbol{f}_{no} \right] \right| = O(n^{-1/2})$$

q queries instead of 1

Main analytic tool (Grown-up version):

Multidimensional CLTs. Sum of many independent "reasonable" ***q*-dimensional** random variables converge to ***q*-dimensional** Gaussian of same mean and covariance.



Main technical work of [Chen-Servedio-T. 14]

Adapting multidimensional CLT for Earth Mover Distance (Valiant-Valiant) to get $\Omega(n^{1/5})$.

[VV]'s proof technique: Stein's method

This work:

Adapt and extend a different multidimensional CLT (Mossel, Gopalan-O'Donnell-Wu-Zuckerman) to get $\Omega(n^{1/2-c})$.

[M-GOWZ]'s proof technique: **Lindeberg's "replacement method"**

Our approach requires **several new ideas** beyond [M-GOWZ].

Three new ideas

1. Random variables that match **arbitrarily many moments** (rather than just two)
2. Careful construction of **mollifiers** in CLT analysis
3. **Pruning** a query set to make it "nice" (main technical work)

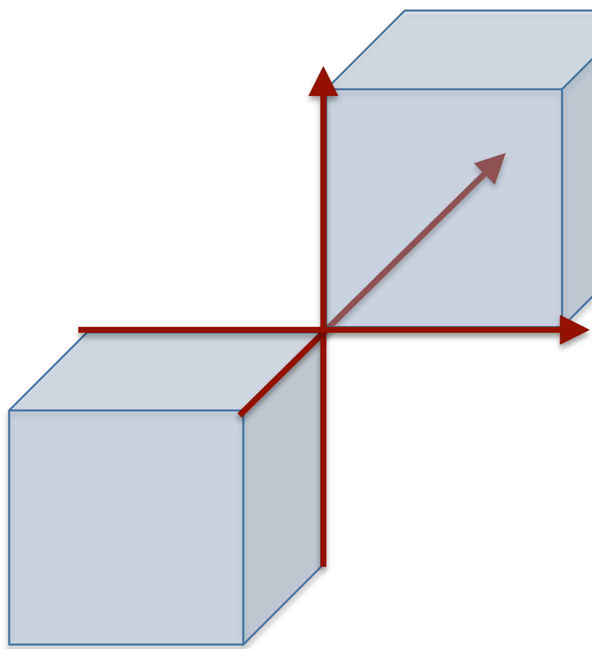
Lindeberg's "replacement method" in one slide

Goal is to bound:

$$\left|\mathbf{E}[\Phi(\boldsymbol{X}_1 + \cdots + \boldsymbol{X}_n)] - \mathbf{E}[\Phi(\boldsymbol{Y}_1 + \cdots + \boldsymbol{Y}_n)]\right|$$

"Mollifier"

In our case, smooth approximation to the indicator of the union of orthants:



Key Ideas:

1. Swap X_i's for Y_i's one by one
2. Bound difference via $\Phi' s$ Taylor expansion

New Idea #1: Matching Higher Moments

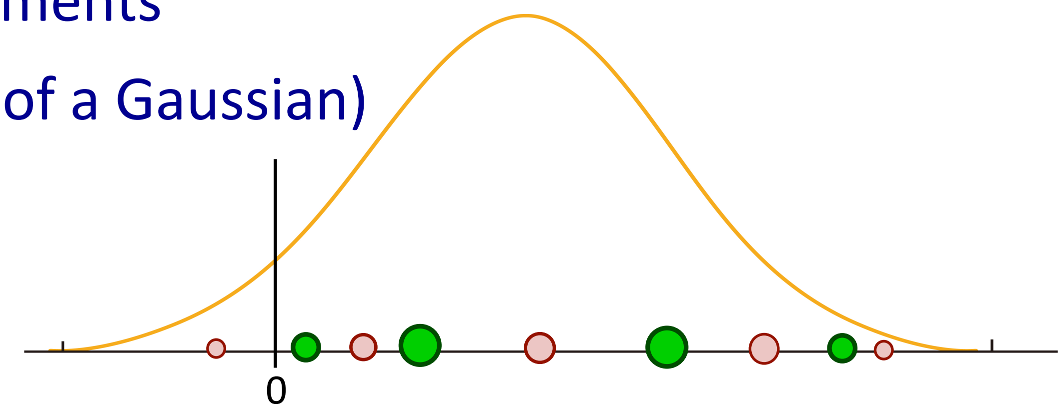
Why? By matching *h* moments: only incur error term of order *h*+1 in Taylor expansion

But first of all, *can* we match higher moments?

Lemma. For every integer *h* there are two random real-valued random variables **u** and **v** satisfying:

1. **u** is supported on *h* values, **all positive** ("yes"/mono LTFs)
2. **v** is supported on *h*+1 values, and **Pr[v < 0] > 0** ("no"/far-from-monotone LTFs)
3. **u** and **v** match first *h* moments

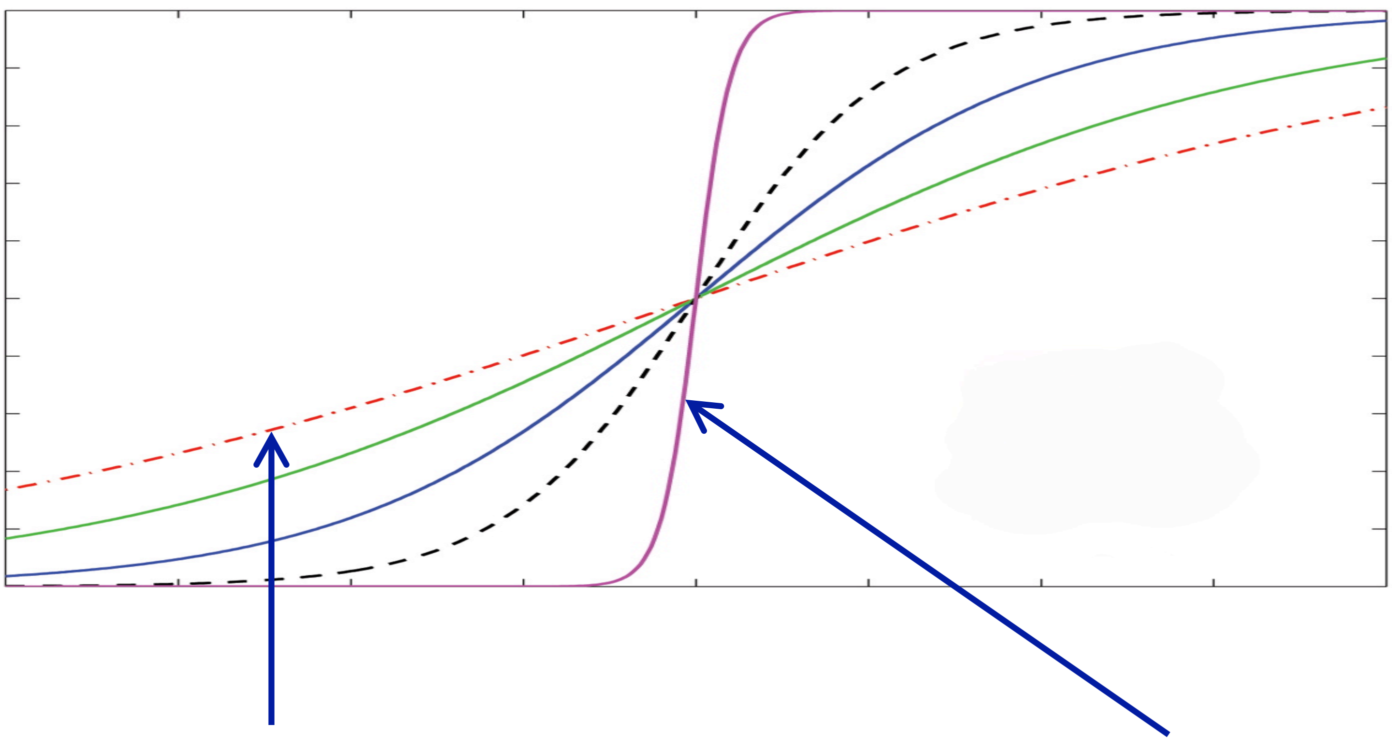
(in fact, they match those of a Gaussian)



Key Ingredient #2: Careful choice of mollifiers

Our mollifier: smooth approximation of indicator of union of orthants

Must carefully control width of "error region" where $0 < \text{mollifier} < 1$



Smooth mollifier (good),
But bad approximation to sign function

Good approximation to sign function,
But high $(h+1)^{st}$ order derivatives (bad)

Using these two ideas, we get $\Omega(n^{1/4})$

Already improves $\Omega(n^{1/5})$ from [Chen-Servedio-T. 14]

To get $\Omega(n^{1/2-c})$, need final new idea …

(main technical work of this paper)

New Idea #3: Pruning the query set

- A delicate CLT analysis yields $\Omega(n^{1/2-c})$ lower bound for "scattered" query sets: no two queries close together.
- Silly but instructive example: our analysis fails for testers that asks **same** query over and over again… but clearly this is equivalent to just 1 query.
- In general, close by queries are likely to take same value, so tester does not "benefit much" from them.

Key Reduction:

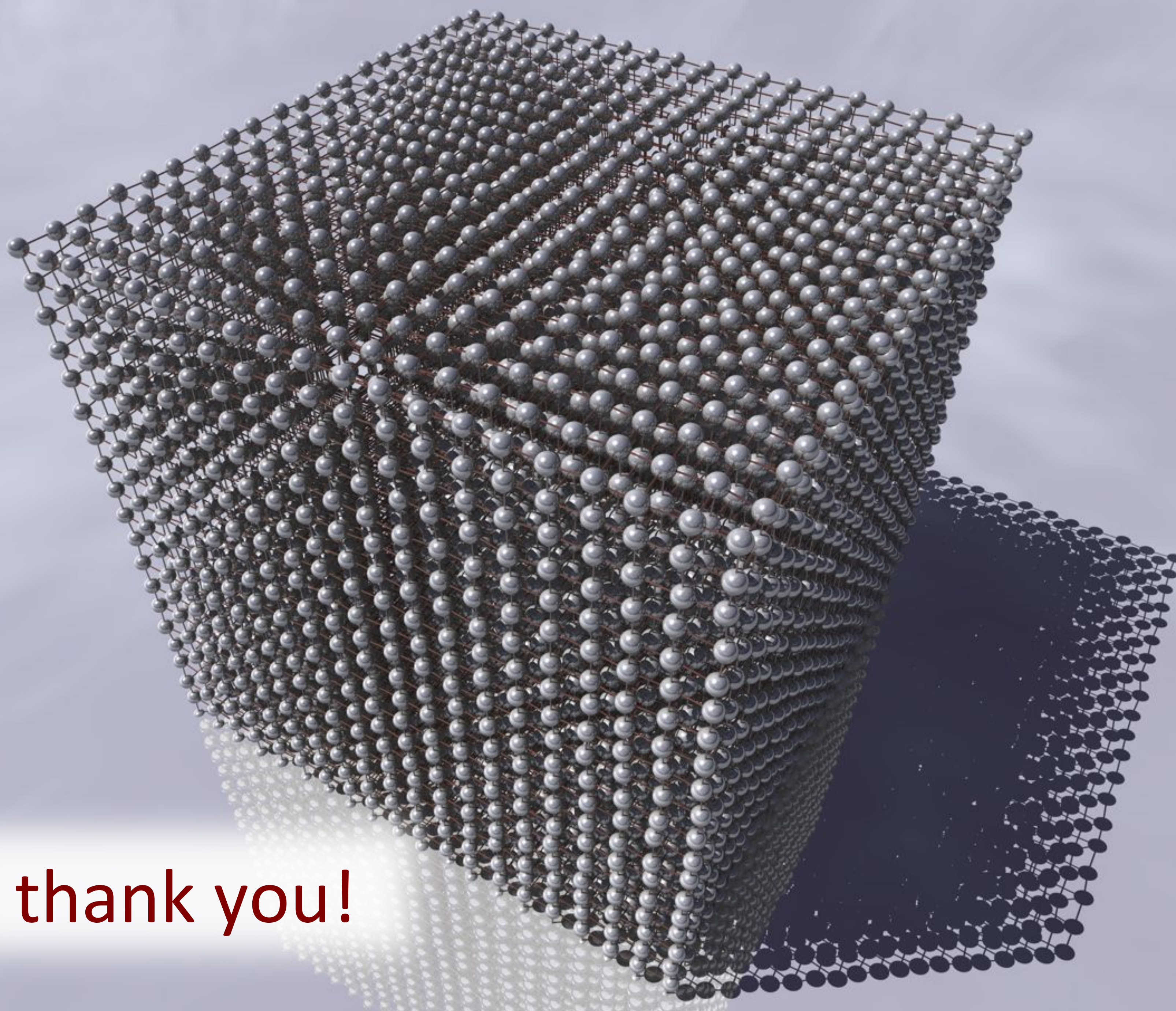
Every set Q of $O(n^{1/2-c})$ queries can be "pruned" to become Q' where

1. Q' is "scattered"
2. Lower bound against Q' yields lower bound against Q

Recap: Our main lower bound

Theorem.

For every $c > 0$ there is an $\varepsilon(c) > 0$ such that any non-adaptive algorithm for testing whether f is monotone or $\varepsilon(c)$-far from monotone requires $\Omega(n^{1/2-c})$ many queries.



thank you!