COMS 6998-3: Sub-Linear Algorithms in Learning and Testing **Spring 2014**

Lecture 8: March 12, 2014

Lecturer: Rocco Servedio *Scriber: Yichi Zhang*

1 Overview

1.1 Last time

- Proper learning for $\mathcal{P}$ implies property testing of $\mathcal{P}$ (generic, but quite inefficient)
- Testing linearity (over $\mathbb{GF}[2]$), i.e. $\mathcal{P} = \{\text{all parities}\}$: (optimal) $O\left(\frac{1}{\epsilon}\right)$-query 1-sided non-adaptive tester.
- Testing monotonicity ($\mathcal{P} = \{\text{all monotone functions}\}$: an efficient $O\left(\frac{n}{\epsilon}\right)$-query 1-sided non-adaptive tester.

1.2 Today

- Finish testing monotonicity ($\mathcal{P} = \{\text{all monotone functions}\}$): an efficient $O\left(\frac{n}{\epsilon}\right)$-query 1-sided non-adaptive algorithm
- Lower bounds:
  - For non-adaptive 1-sided testers, we will show a $\Omega(\sqrt{n})$ lower bound from [FLN$^+$02].
  - Start the proof of $\Omega\left(n^{1/5}\right)$ by Chen–Servedio–Tan for non-adaptive, 2-sided testers, using **Yao's minmax principle** which converts the problem to the problem of lower bound for *deterministic* algorithms (under suitable distribution on inputs).

Relevant Readings:

- E. Fischer and E. Lehman and I. Newman and S. Raskhodnikova and R. Rubinfeld and A. Samorodnitsky: *Monotonicity Testing Over General Poset Domains.* [FLN$^+$02]

- O. Goldreich and S. Goldwasser and E. Lehman and D. Ron and A. Samordinsky: *Testing Monotonicity.* [GGL$^+$00]

2 Testing Monotonicity (contd. from last time)

Recall the set of *violating edges* $V(f) \subseteq E$ can be decomposed as

$$V(f) = V_1(f) \cup V_2(f) \cup \cdots \cup V_n(f)$$

where $V_i(f) \subseteq V(f)$ is the set of *coordinate-i violating edges.*, and we defined the quantity $\eta(f) \stackrel{\text{def}}{=} \frac{|V(f)|}{n2^{n-1}} = \frac{|V(f)|}{n2^{n-1}} = \Pr[\,\textsc{EdgeTester}$ outputs $\mathsf{REJECT}\,]$.

Goal: prove $\eta(f) \geq \frac{\text{dist}(f,\mathcal{M})}{n}$ i.e.

$$n\eta(f) \geq \text{dist}(f, \mathcal{M}). \tag{1}$$

To do so, for any fixed f we will show how to construct a monotone function g such that $\text{dist}(f, g) \leq n\eta(f)$.
Finally, recall the definition of the *shift operator* S_i:

Definition 1 (Shift Operator)**.** *Fix $i \in [n]$. The* shift operator S_i *acts on functions $h\colon \{0,1\}^n \to \{0,1\}$, by sorting $h(x^{i\leftarrow 0}), h(x^{i\leftarrow 1})$: $S_i h$ is a function from $\{0,1\}^n$ to $\{0,1\}$ defined by*

$$S_i h(x^{i\leftarrow 0}) = \min\left(h(x^{i\leftarrow 0}), h(x^{i\leftarrow 1})\right)$$
$$S_i h(x^{i\leftarrow 1}) = \max\left(h(x^{i\leftarrow 0}), h(x^{i\leftarrow 1})\right)$$

In the following, we let $D_i(f) \stackrel{\text{def}}{=} 2\,|V_i(f)|$ be the number of vertices x such that $S_i(f)(x) \neq f(x)$.

Definition 2. *We say $h\colon \{0,1\}^n \to \{0,1\}$ is* i-monotone *if no x has $h(x^{i\leftarrow 0}) = 1$ but $h(x^{i\leftarrow 1}) = 0$, that is if h has no violation in the i^{th} coordinate. For $A \subseteq [n]$, we say h is A-monotone if h is i-monotone for all $i \in A$.*

Claim 3 (2-part claim)**.** *1. If h is A-monotone and $j \notin A$, then $S_j(h)$ is $(A \cup j)$-monotone.*

2. *For every $i,j \in [n]$, we have $D_i(S_j(h)) \leq D_i(h)$ (shifting does not increase violations).*

Before proving this claim, we show how it directly yields our goal:

Proof of Eq. (1) *using Claim 3.* Let $g \stackrel{\text{def}}{=} S_n(S_{n-1}(\cdots S_1(f))\cdots) = S_n \circ S_{n-1} \circ \cdots \circ S_1(f)$. By the Part 1 of the claim, g is monotone (as it is $[n]$-monotone); hence, it is sufficient to prove it is not too far from f – namely, that $n\eta(f) \geq \text{dist}(f,g)$.

Let f_i denote $S_i \circ S_{i-1} \circ \cdots \circ S_1(f)$ (so in particular $f = f_0$ and $g = f_n$). By the triangle inequality,

$$\text{dist}(f,g) \leq \text{dist}(f_0,f_1) + \cdots + \text{dist}(f_{n-1},f_n)$$

Focusing on a fixed term of the sum, for $i \in [n]$

$$\begin{aligned}
\text{dist}(f_{i-1},f_i) = \text{dist}(f_{i-1},S_i(f_{i-1})) &= \frac{D_i(f_{i-1})}{2^n} \\
&= \frac{D_i(S_{i-1}\circ\cdots\circ S_1(f))}{2^n} \\
&\leq \frac{D_i(S_{i-2}\circ\cdots\circ S_1(f))}{2^n} && \text{(Claim 3, Part 2)} \\
&\leq \frac{D_i(f_0)}{2^n} = \frac{|V_i(f)|}{2^{n-1}} && \text{(Repeating the inequality)}
\end{aligned}$$

which, coming back to the sum, gives

$$\text{dist}(f,g) \leq \frac{|V_1(f)| + \cdots + |V_n(f)|}{2^{n-1}} = \frac{|V(f)|}{2^{n-1}} = n\eta(f)$$

as $|V(f)| = |\cup_{i=1}^n V_i(f)| = \sum_{i=1}^n |V_i(f)|$ by disjointness; and finally by definition of $\eta(f)$. □

It remains to prove the claim:

Proof of Claim3. First, observe that (2) ⇒ (1): indeed, assume Part 2 holds, and suppose h is A-monotone. Fix any $j \notin A$. Since $S_j(h)$ is j-monotone by application of the shift operator; we only have to show that $S_j(h)$ is i-monotone as well, for any $i \in A$.

Fix such an $i \in A$: the number of i-edges where $S_j(h)$ violates monotonicity is

$$|V_i(S_j(h))| = \frac{D_i(S_j(h))}{2} \underset{\text{(Part 2)}}{\leq} \frac{D_i(h)}{2} = |V_i(h)| = 0$$

as stated.

Turning to Part 2: rather disappointly, this is a "proof by inspection", as there are actually only 16 cases to consider: only 2 variables are really involved, i and j.
More precisely, without loss of generality, one can take $i = 1$ and $j = 2$; fixing coordinates $x_3, \cdots, x_n \in \{0,1\}^{n-2}$, h becomes a bivariate function $h\colon \{0,1\}^2 \to \{0,1\}$. Hence, it is sufficient to argue that for all $h : \{0,1\}^2 \to \{0,1\}$, $D_1(S_2(h)) \leq D_1(h)$ – which can be done by enumerating all 16 cases. □

Remark 1. *This algorithm was analyzed in 2000; it is known that the analysis is tight, that is that this "edge tester"* needs $\Omega(n)$ *queries: a hard instance would be any dictator function $x \mapsto \bar{x}_i$, anti-monotone.*
In 2013, Chakrabarty and Seshadhri ([CS13]) broke the "linearity barrier" for testing monotonicity by giving a $O\big(n^{7/8}/\epsilon^{3/2}\big)$-query tester[1] which combines the edge tester with a "path tester" (which picks a random path in the hypercube, then queries two points randomly on this path). This has (very) recently been improved to an $n^{5/6}$ dependency, by Chen–Servedio–Tan (2014).

3 $\Omega(\sqrt{n})$ lower bound for non-adaptive 1-sided testers

Theorem 4. *There is an absolute constant $\epsilon_0 > 0$ such that any one-sided non-adaptive ϵ_0-tester for $\mathcal{M}$ must make at least $\frac{\sqrt{n}}{3}$ queries.*

Observation 5. *Suppose $\mathcal{A}$ is such a tester, and say $\mathcal{A}$* reveals a violation of f *if it queries x, y with $x \prec y$ such that $f(x) = 1$, $f(y) = 0$. As it is one-sided, $\mathcal{A}$ can only reject when it is "sure" beyond any doubt; that is, if $\mathcal{A}$ does* not *reveals a violation in an execution, it* must *output* ACCEPT*. Therefore, if A is 1-sided non-adaptive tester for monotonicity, it must be the case that for* every f *with* $\mathrm{dist}(f, \mathcal{M}) > \epsilon_0$, *$\mathcal{A}$ must reveal a violation of f with probability at least $\frac{2}{3}$.*

Definition 6. *For $i \in [n]$, define the* truncated anti-dictator f_i *as*

$$f_i\colon \{0,1\}^n \to \{0,1\}$$
$$x \mapsto \begin{cases} 1 & if \sum_{j=1}^n x_j \geq \frac{n}{2} + \sqrt{n} \\ 0 & if \sum_{j=1}^n x_j < \frac{n}{2} - \sqrt{n} \\ \bar{x}_i & o.w. \end{cases}$$

[1] Note that the quantity of interest in the query complexity is n, so this result is an improvement even though the exponent of ϵ is now $3/2 > 1$. More generally, compared to n the parameter ϵ is seen as a constant, and in property testing $2^{2^{2^{1/\epsilon}}}$ will always be considered better than $\frac{\log^* n}{\epsilon}$.

Fact 7. *There exists an absolute constant $\epsilon_0 > 0$ such that, for every $i \in [n]$,* $\mathrm{dist}(f_i, \mathcal{M}) > \epsilon_0$.

Proof. Indeed, there are at least $c2^n$) (for some suitable constant $c > 0$) many $x \in \{0,1\}^n$ having:

$$\frac{n}{2} - \sqrt{n} < \sum_{i=1}^{n} x_i < \frac{n}{2} + \sqrt{n}$$

(in the "middle slice"). Without loss of generality, we consider the case $i = 1$: we can pair up inputs of the form $z = (1, z_2, \cdots, z_n)$ for which $f_1(z) = 0$ with $z' = (0, z_2, \cdots, z_n)$, for which $f_1(z') = 1$.
Any monotone function g disagrees with f_1 on at least 1 of these two inputs; so any monotone function must disagree with f on at least $\frac{c}{2} \cdot 2^n$ points. □

Lemma 8. *Let $\mathcal{A}$ be any non-adaptive q-query algorithm. Then there exists $i \in [n]$ such that $\mathcal{A}$ reveals a violation on f_i with probability at most $\frac{2q}{\sqrt{n}}$.*

This implies the theorem: any one-sided non-adaptive tester $\mathcal{A}$ with query complexity $q < \frac{\sqrt{n}}{3}$ will reveal a violation on some f_{i^*} with probability $< 2/3$; but it only rejects on such occasions, yet any successful tester should reject f_{i^*} with probability at least $2/3$.

Proof of Lemma 8. Fix $\mathcal{A}$ to be any q-query non-adaptive algorithm, and let Q be the set of q queries it makes. We will show Q reveals violations of f_i for at most $2(q-1)\sqrt{n}$ many $i \in [n]$: this in turn implies that

$$\begin{aligned}\sum_{i=1}^{n} \Pr[\, A \text{ reveals a violation of } f_i \,] &= \sum_{i=1}^{n} \mathbb{E}\left[\mathbf{1}_{\substack{A \text{ reveals a} \\ \text{violation of } f_i}}\right] = \mathbb{E}\left[\sum_{i=1}^{n} \mathbf{1}_{\substack{A \text{ reveals a} \\ \text{violation of } f_i}}\right] \\ &= \mathbb{E}[|\{\, i \in [n] : \ A \text{ reveals a violation of } f_i \,\}|] \\ &\leq 2(q-1)\sqrt{n}\end{aligned}$$

so there exists $i \in [n]$ such that $\Pr[\, \mathcal{A} \text{ reveals a violation of } f_i \,] \leq \frac{2(q-1)}{\sqrt{n}}$.

Q is an arbitrary set of q strings in $\{0,1\}^n$; without loss of generality, one can further assume every string $z \in Q$ has Hamming weight $|z| \in [\frac{n}{2} - \sqrt{n}, \frac{n}{2} + \sqrt{n}]$, as querying any other cannot reveal any violation of f_i. Q reveals violations for f_i only if Q contains 2 comparable strings $u \succ v$ such that $u_i \neq v_i$.

Accordingly, let G_Q be a q-node unirected graph with vertex set $V = Q$ and edge set E containing only comparable pairs: $(u, v) \in E$ iff $u \prec v$ or $v \prec u$.

(1) $|E| \leq \binom{n}{2} \leq q^2$ (pairs of comparable strings); and each pair reveals a violation of at most $2\sqrt{n}$ f_i's (by the Hamming weight assumption: $u, v \in Q$ can differ in at most that many coordinates). Therefore, the total number of i's such that Q can reveal a violation of f_i is at most $2\sqrt{n}\binom{n}{2} \leq 2q^2\sqrt{n}$. *Almost* what we need, but with q^2 instead of q.

(2) A better bound can be achieved by considering a *spanning forest* F_Q of G_Q: F_Q has at most $q-1$ edges. Furthermore, if Q has two comparable strings u, v with $u_i \neq v_i$, u and v will be in the same tree and some edge in the path $u \rightsquigarrow v$ has endpoints with different value on their i^{th} coordinate, and hence presents a violation of f_i. As before, every 2 adjacent vertices in a tree differ by at most $2\sqrt{n}$ coordinates, so the maximum number of i's such that f_i has a violation reveals in F_Q (and thus in G_Q) is $2(q-1)\sqrt{n}$.

□

4 $\tilde{\Omega}(n^{1/5})$ lower bound by Chen–Servedio–Tan for non-adaptive, 2-sided testers

We will now (start to) prove the following lower bound:

Theorem 9. *There exists $\epsilon_0 > 0$ such that any 2-sided non-adaptive tester for $\mathcal{M}$ must make $\tilde{\Omega}(n^{1/5})$ queries.*

To do so, we start by describing a general approach and one of the key tools for property testing lower bounds: "Yao's Minmax Theorem".

4.1 Yao's Principle (easy direction)

Consider a decision problem (here, Property Testing) over a (finite) set X of possible inputs (in our case, $X = \mathcal{P} \cup \{ f : \text{dist}(f, \mathcal{P}) > \epsilon \}$, and the inputs are functions), and a randomized non-adaptive decision algorithm $\mathcal{A}$ that makes q queries to its input f. Such an algorithm is equivalent to a probability distribution $\mu = \mu_A$ over *deterministic* q-query decision algorithms. Letting Y be the set of all such determistic algorithms, we consider the $X \times Y$ matrix M with Boolean entries, and

- rows indexed by functions $f \in X$;

- columns indexed by algorithms $y \in Y$ (or, equivalently, by sets Q of queries, possibly with repetitions)

such that $M(f,y) = \begin{cases} 1 & \text{if } y \text{ is right on input } f \\ 0 & \text{o.w.} \end{cases}$.

Our randomized algorithm $\mathcal{A}$ is thus equivalent to a distribution μ over columns (i.e., over Y), non-negative function with $\sum_{y \in Y} \mu(y) = 1$. Similarly, a distribution λ over inputs ($f \in X$) satisfies $\sum_{f \in X} \lambda(f) = 1$.

For $\mathcal{A}$ to be a successful q-query property testing algorithm, it must be such that for every row $f \in X$:

$$\Pr[\,\mathcal{A} \text{ outputs right answer on } f\,] \geq 2/3$$

that is $\Pr_{y \sim \mu}[\,M(f,y) = 1\,] \geq 2/3$.

Suppose there is a distribution λ over X such that *every* $y \in Y$ has:

$$\Pr_{f \sim \lambda}[\,M(f,y) = 1\,] < 2/3.$$

Then, for *any* distribution μ over Y:

$$\Pr_{\substack{f \sim \lambda \\ y \sim \mu}}[\,M(f,y) = 1\,] < 2/3$$

so it cannot be the case that for every $f \in X$, $\Pr_{y \sim \mu}[\,M(f,y) = 1\,] \geq 2/3$.
and in particular $\mathcal{A}$ (which is fully characterized by μ) is not a legit tester – since there exists some f with $\Pr[\,A \text{ right on } f\,] < 2/3$.

This is what Yao's Principle states (at least, what its "easy direction" does): one can reduce the problem of dealing with *randomized* (non-adaptive) algorithms over arbitrary inputs to the one of *deterministic* algorithms over a ("suitably difficult") *distribution* over inputs:

Theorem 10 (Yao's Minmax Principle, easy direction)**.** *Suppose there is a distribution λ over functions (legitimate inputs: $f \in \mathcal{P} \cup \{\, h : \operatorname{dist}(h,\mathcal{P}) > \epsilon \,\}$) such that any q-query* deterministic *algorithm is correct with probability $< 2/3$ when $f \sim \lambda$.*
Then, given any *(non-adaptive) q-query* randomized *algorithm $\mathcal{A}$, there exists $f_{\mathcal{A}} \in X$, such that*

$$\Pr[\,\mathcal{A} \textit{ is correct on } f_{\mathcal{A}}\,] < 2/3$$

Hence, any non-adaptive property testing algorithm for $\mathcal{P}$ must make at least $q+1$ queries.

Goal: find hard distribution over functions, for *deterministic* algorithms.

More precisely, to get a grip on what being a *hard* distribution is, recall the notion of distance between probability distributions we introduced at the beginning of the course:

Definition 11. *Suppose D_1, D_2 are both probability distributions over a finite set Ω; their* total variation distance *is defined*[2] *as*

$$\mathrm{d_{TV}}(D_1, D_2) \stackrel{\text{def}}{=} \max_{S \subseteq \Omega} (D_1(S) - D_2(S)) = \frac{1}{2} \sum_{\omega \in \Omega} |D_1(\omega) - D_2(\omega)| \in [0, 1]$$

This will come in handy to prove our lower bounds, as (very hazily) two sequences of queries/answers whose distribution are very close are impossible to distinguish with high probability:

Exercise 12 (Homework problem)**.** Let D_1, D_2 be probability distributions over a finite set Ω, and fix $\mathcal{A}$ to be any algorithm (deterministic or randomized) which, on input an element $\omega \in \Omega$, either outputs ACCEPT or REJECT. Prove that

$$\left| \Pr_{\omega \sim D_1} [\, \mathcal{A} \text{ outputs } \mathsf{ACCEPT} \,] - \Pr_{\omega \sim D_1} [\, \mathcal{A} \text{ outputs } \mathsf{ACCEPT} \,] \right| \leq \mathrm{d_{TV}}(D_1, D_2)$$

References

[CS13] Deeparnab Chakrabarty and C. Seshadhri. A o(n) monotonicity tester for boolean functions over the hypercube. In *Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing*, STOC '13, pages 411–418, New York, NY, USA, 2013. ACM.

[FLN+02] E. Fischer, E. Lehman, I. Newman, S. Raskhodnikova, R. Rubinfeld, and A. Samorodnitsky. Monotonicity testing over general poset domains. In *STOC*, pages 474–483, 2002.

[GGL+00] O. Goldreich, S. Goldwasser, E. Lehman, D. Ron, and A. Samordinsky. Testing monotonicity. *Combinatorica*, 20(3):301–337, 2000.

[2] The second equality is known as Scheffé's lemma.

COMS 6998-3: Sub-Linear Algorithms in Learning and Testing **Spring 2014**

Lecture 9: March 26, 2014

Lecturer: Rocco Servedio *Scriber: Keith Nichols*

1 Overview

1.1 Last Time

- Finished analysis of $O\left(\frac{n}{\epsilon}\right)$-query algorithm for monotonicity.
- Showed an $\Omega(\sqrt{n})$ lower bound for one-sided non-adaptive monotonicity testers.
- Stated and proved (one direction of) *Yao's Principle*: Suppose there exists a distribution $\mathcal{D}$ over functions $f\colon \{-1,1\}^n \to \{-1,1\}$ (the inputs to the property testing problem) such that any q-query *deterministic* algorithm gives the right answer with probability at most c. Then, given any q-query non-adaptive *randomized* testing algorithm $\mathcal{A}$, there exists some function $f_{\mathcal{A}}$ such that:

$$\Pr[\,\mathcal{A} \text{ outputs correct answer on} f_{\mathcal{A}}\,] \leq c.$$

1.2 Today: lower bound for two-sided non-adaptive monotonicity testers.

We will use Yao's Principle to show the following lower bound:

Theorem 1 (Chen–Servedio–Tan '14)**.** *Any 2-sided non-adaptive property tester for monotonicity, to ϵ_0-test, needs $\tilde{\Omega}\left(n^{1/5}\right)$ queries (where $\epsilon_0 > 0$ is an absolute constant).*

2 $\tilde{\Omega}\left(n^{1/5}\right)$ lower bound: proving Theorem 1

2.1 Preliminaries

Recall the definition of total variation distance between two distributions over the same set Ω:

$$\mathrm{d_{TV}}(\mathcal{D}_1, \mathcal{D}_2) = \frac{1}{2}\sum_x |\mathcal{D}_1(x) - \mathcal{D}_2(x)|\,.$$

As homework problem from last time, we have the lemma[1] below, which relates the probability of distinguishing between samples from two distributions to their total variation distance:

Lemma 2 (HW problem)**.** *Let $\mathcal{D}_1$, $\mathcal{D}_2$ be two distributions over some set Ω, and $\mathcal{A}$ be any algorithm (possibly randomized) that takes $x \in \Omega$ as input and outputs* Yes *or* No*. Then* HW Problem

$$\left| \Pr_{x \sim \mathcal{D}_1} [\, \mathcal{A}(x) = \mathsf{Yes} \,] - \Pr_{x \sim \mathcal{D}_2} [\, \mathcal{A}(x) = \mathsf{Yes} \,] \right| \leq \mathrm{d_{TV}}(\mathcal{D}_1, \mathcal{D}_2)$$

where the probabilities are also taken over the possible randomness of $\mathcal{A}$.

To apply this lemma, recall that given a deterministic algorithm's set of queries $Q = \{z^{(1)}, \ldots, z^{(q)}\} \subseteq \{-1,1\}^n$, a distribution $\mathcal{D}$ over Boolean functions induces a distribution $\mathcal{D}\Big|_Q$ over $\{-1,1\}^q$: x is drawn from $\mathcal{D}\Big|_Q$ by

- drawing $f \sim \mathcal{D}$;
- outputting $(f(z^{(1)}, \ldots, f(z^{(q)})) \in \{-1,1\}^q$.

With this observation and Yao's principle in hand, we can state and prove a key tool in proving lower bounds in property testing:

Lemma 3 (Key Tool)**.** *Fix any property $\mathcal{P}$ (a set of Boolean functions). Let $\mathcal{D}_{\mathsf{Yes}}$ be a distribution over the Boolean functions that belong to $\mathcal{P}$, and $\mathcal{D}_{\mathsf{No}}$ be a distribution over Boolean functions that all have* $\mathrm{dist}(f, \mathcal{P}) > \epsilon$*.*
Suppose that for all q-query sets Q, one has $\mathrm{d_{TV}}\left(\mathcal{D}_{\mathsf{Yes}}\Big|_Q, \mathcal{D}_{\mathsf{No}}\Big|_Q\right) \leq \frac{1}{4}$. Then any (2-sided) non-adaptive ϵ-tester for $\mathcal{P}$ must use at least $q+1$ queries.

Proof. Let $\mathcal{D}$ be the mixture $\mathcal{D} \stackrel{\text{def}}{=} \frac{1}{2}\mathcal{D}_{\mathsf{Yes}} + \frac{1}{2}\mathcal{D}_{\mathsf{No}}$ (that is, a draw from $\mathcal{D}$ is obtained by tossing a fair coin, and returning accordingly a sample drawn either from $\mathcal{D}_{\mathsf{Yes}}$ or $\mathcal{D}_{\mathsf{No}}$). Fix a q-query deterministic algorithm $\mathcal{A}$. Let

$$p_Y \stackrel{\text{def}}{=} \Pr_{f \sim \mathcal{D}_{\mathsf{Yes}}} [\, \mathcal{A} \text{ accepts on } f \,], \qquad p_N \stackrel{\text{def}}{=} \Pr_{f \sim \mathcal{D}_{\mathsf{No}}} [\, \mathcal{A} \text{ accepts on } f \,]$$

That is, p_Y is the probability that a random "Yes" function is accepted, while p_N is the probability that a random "No" function is accepted. Via the assumption and the

[1] This is sometimes referred to as a "data processing inequality" for the total variation distance.

previous lemma, $|p_Y - p_N| \leq \frac{1}{4}$. However, this means that $\mathcal{A}$ cannot be a succesful tester; as

$$\Pr_{f\sim\mathcal{D}}[\,\mathcal{A} \text{ gives wrong answer}\,] = \frac{1}{2}(1-p_Y) + \frac{1}{2}p_N = \frac{1}{2} + \frac{1}{2}(p_N - p_Y) \geq \frac{3}{8} > \frac{1}{3}$$

So Yao's Principle tells us that any randomized non-adaptive q-query algorithm is wrong on *some* f in support of $\mathcal{D}$ with probability at least $\frac{3}{8}$; but a legit tester can only be wrong on any such f with probability less than $\frac{1}{3}$. □

Exercise 4 (Generalization of Lemma 3)**.** Relax the previous lemma slightly. Prove that the conclusion still holds even under the weaker assumptions HW Problem

$$\Pr_{f\sim\mathcal{D}_{\mathsf{Yes}}}[\,f \in \mathcal{P}\,] \geq \frac{99}{100}, \qquad \Pr_{f\sim\mathcal{D}_{\mathsf{No}}}[\,\mathrm{d_{TV}}(f,\mathcal{P}) > \epsilon\,] \geq \frac{99}{100}.$$

For our lower bound, we need to come up with $\mathcal{D}_{\mathsf{Yes}}$ (resp. $\mathcal{D}_{\mathsf{No}}$) to be over monotone functions (resp. ϵ_0-far from monotone) such that $\forall Q \subseteq \{-1,1\}^n$ with $|Q| = q$, $\mathrm{d_{TV}}\left(\mathcal{D}_{\mathsf{Yes}}\big|_Q, \mathcal{D}_{\mathsf{No}}\big|_Q\right) \leq \frac{1}{4}$.

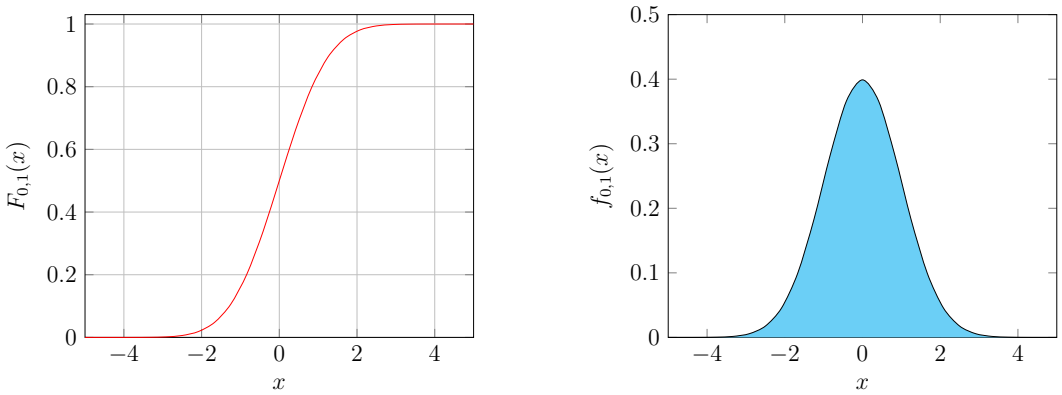
At a high-level, we need to argue that both distributions "look the same". One may thus think of the Central Limit Theorem – *the sum of many independent, "nice" real-valued random variables converges to a Gaussian in distribution (in cumulative distribution function).* For instance, a binomial distribution $\mathrm{Bin}\left(10^6, \frac{1}{2}\right)$ has the same shape ("bell curve") as the corresponding Gaussian distribution $\mathcal{N}\left(\frac{1}{2}, \frac{1}{4}10^6\right)$. For our purpose, however, the convergence guarantees stated by the Central Limit Theorem will not be enough, as they do not give explicit bounds on the rate of convergence; we will use a "quantitative version" of the CLT, the *Berry–Esséen Theorem.*

First, recall the definition a (real-valued) Gaussian random variable:

Definition 5 (One-dimensional Gaussian distribution)**.** *A real-valued random variable is said to be* Gaussian with mean μ and variance σ *if it follows the distribution* $\mathcal{N}(\mu,\sigma)$*, which has probability density function*

$$f_{\mu,\sigma}(x) \stackrel{\text{def}}{=} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad x \in \mathbb{R}$$

Such a random variable has indeed expectation μ and variance σ^2; futhermore, the distribution is fully specified by these two parameters. Extending to higher dimensions, one can define similarly a d-dimensional Gaussian random variable:



(a) Cumulative distribution function (CDF) (b) Probability density function (PDF)

Figure 1: Standard Gaussian $\mathcal{N}(0,1)$.

Definition 6 (d-dimensional Gaussian distribution)**.** *Fix a vector $\mu \in \mathbb{R}^d$ and a symmetric non-negative definite matrix $\Sigma \in \mathbb{R}^{d\times d}$. A random variable taking values in $\mathbb{R}^d$ is said to be* Gaussian with mean μ and covariance Σ *if it follows the distribution $\mathcal{N}(\mu,\Sigma)$, which has probability density function*

$$f_{\mu,\Sigma}(x) \stackrel{\text{def}}{=} \frac{1}{\sqrt{(2\pi)^k \det \Sigma}} e^{-\frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)}, \qquad x \in \mathbb{R}^d$$

As in the univariate case, μ and Σ uniquely define the distribution; further, one has that for $X \sim \mathcal{N}(\mu,\Sigma)$,

$$\Sigma_{i,j} = \mathrm{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mathbb{E}X_i)(X_j - \mathbb{E}X_j)]\,, \qquad i,j \in [d].$$

Theorem 7 (Berry–Esséen[2])**.** *Let $S \stackrel{\text{def}}{=} X_1 + \ldots + X_n$ be the sum of n independent (real-valued) random variables $X_1,\ldots,X_n$ satisfying*

$$\Pr[\,|X_i - \mathbb{E}[X_i]| \le \tau\,] = 1.$$

that is every X_i is almost surely bounded. For $i \in [n]$, define $\mu_i \stackrel{\text{def}}{=} \mathbb{E}[X_i]$ and $\sigma_i \stackrel{\text{def}}{=} \sqrt{\mathrm{Var}\, X_i}$, so that $\mathbb{E}S = \sum_{i=1}^n \mu_i$ and $\mathrm{Var}\, S = \sum_{i=1}^n \sigma_i^2$ (the last equality by independence). Finally, let G be a $\mathcal{N}\left(\sum_{i=1}^n \mu_i, \sqrt{\sum_{i=1}^n \sigma_i^2}\right)$ Gaussian variable, matching the first two moments of S. Then, for all $\theta \in \mathbb{R}$,

$$|\Pr[\,S \le \theta\,] - \Pr[\,G \le \theta\,]| \le \frac{O(\tau)}{\sqrt{\sum_{i=1}^n \sigma_i^2}}.$$

In other terms[3]*, letting F_S (resp. F_G) denote the CDF of S (resp. G), one has $\|F_S - F_G\|_\infty \leq \frac{O(\tau)}{\sqrt{\sum_{i=1}^n \sigma_i^2}}$.*

Remark 1. *The constant hidden in the $O(\cdot)$ notation is actually very reasonable – one can take it to be equal to 1.*

Application: baby step towards the lower bound. Fix any string $z \in \{-1,1\}^n$, and for $i \in [n]$ let the (independent) random variables γ_i be defined as

$$\gamma_i \stackrel{\text{def}}{=} \begin{cases} +1 & \text{w.p. } \frac{1}{2} \\ -1 & \text{w.p. } \frac{1}{2} \end{cases}$$

Letting $X_i \stackrel{\text{def}}{=} \gamma_i z_i$, we have $\mu_i = \mathbb{E}X_i = 0$, $\sigma_i = \operatorname{Var} X_i = 1$; and can take $\tau = 1$ to apply the Berry–Esséen theorem to $X \stackrel{\text{def}}{=} X_1 + \ldots + X_n$. This allows us to conclude that

$$\forall \theta \in \mathbb{R}, \qquad |\Pr[\, X \leq \theta\,] - \Pr[\, G \leq \theta\,]| \leq \frac{O(1)}{\sqrt{n}}$$

for $G \sim \mathcal{N}(0, \sqrt{n})$.

Now, consider a slightly different distribution than the λ_i's: for the same $z \in \{-1,1\}^n$, define the independent random variables ν_i by

$$\nu_i \stackrel{\text{def}}{=} \begin{cases} \frac{1}{3} & \text{w.p. } \frac{9}{10} \\ -3 & \text{w.p. } \frac{1}{10} \end{cases}$$

and let $Y_i \stackrel{\text{def}}{=} \nu_i z_i$ for $i \in [n]$, $Y \stackrel{\text{def}}{=} Y_1 + \cdots + Y_n$. By our choice of parameters,

$$\mathbb{E}Y_i = \left(\frac{1}{10} \cdot (-3) + \frac{9}{10} \cdot \frac{1}{3}\right) z_i = 0 = \mathbb{E}X_i$$
$$\operatorname{Var} Y_i = \mathbb{E}\left[Y_i^2\right] = \frac{1}{10} \cdot 9 + \frac{9}{10} \cdot \frac{1}{9} = 1 = \operatorname{Var} X_i$$

So $\mathbb{E}[Y] = \mathbb{E}[Y] = 0$ and $\operatorname{Var} Y = \operatorname{Var} X = n$; by the Berry–Esséen theorem (with τ set to 3, and G as before)

$$\forall \theta \in \mathbb{R}, \qquad |\Pr[\, Y \leq \theta\,] - \Pr[\, G \leq \theta\,]| \leq \frac{O(1)}{\sqrt{n}}$$

[3]This quantity $\|F_S - F_G\|_\infty$ is also referred to as the *Kolmogorov distance* between S and G.

[3]There exist other versions of this theorem, with weaker assumptions or phrased in terms of the third moments of the X_i's; we only state here one tailored to our needs.

and by the triangle inequality

$$\forall \theta \in \mathbb{R}, \qquad |\Pr[\,X \le \theta\,] - \Pr[\,Y \le \theta\,]| \le \frac{O(1)}{\sqrt{n}} \tag{1}$$

We can now define $\mathcal{D}_{\mathsf{Yes}}$ and $\mathcal{D}_{\mathsf{No}}$ based on this (that is, based on respectively a random draw of $\lambda, \nu \in \mathbb{R}^n$ distributed as above): a function $f_\lambda \sim \mathcal{D}_{\mathsf{Yes}}$ is given by

$$\forall z \in \{-1,1\}^n, \qquad f_\lambda(z) \stackrel{\text{def}}{=} \text{sign}(\lambda_1 z_1 + \ldots \lambda_n z_n).$$

and similarly for $f_\nu \sim \mathcal{D}_{\mathsf{No}}$:

$$\forall z \in \{-1,1\}^n, \qquad f_\nu(z) \stackrel{\text{def}}{=} \text{sign}(\nu_1 z_1 + \ldots \nu_n z_n)$$

With the notations above, $X \le 0$ if and only if $f_\gamma(z) = -1$ and $Y \le 0$ if and only if $f_\nu(z) = -1$. This implies that for any fixed *single* query z,

$$\mathrm{d_{TV}}\left(\mathcal{D}_{\mathsf{Yes}}\big|_{\{z\}}, \mathcal{D}_{\mathsf{No}}\big|_{\{z\}}\right) = \frac{1}{2}\left(|\Pr[\,X \le 0\,] - \Pr[\,Y \le 0\,]| + |\Pr[\,X > 0\,] - \Pr[\,Y > 0\,]|\right) \le \frac{O(1)}{\sqrt{n}}.$$

This *almost* looks like what we were aiming at – so why aren't we done? There are two problems with what we did above:

1. This only deals the case $q = 1$; that is, would provide a lower bound against *one-query* algorithms.
   Fix: we will use a *multidimensional* version of the Berry–Esséen Theorem for the sums of q-dimensional independent random variables (converging to a multi-dimensional Gaussian).

2. f_γ, f_ν are not monotone (indeed, both the γ_i's and ν_i's can be negative).
   Fix: shift everything by 2:

   - $\gamma_i \in \{1, 3\}$: f_γ is monotone;
   - $\nu_i \in \{-1, \frac{7}{3}\}$: f_ν will be far from monotone with high probability (will show this).

2.2 The lower bound construction

Up until this point, everything has been a warmup; we are now ready to go into more detail.

$\mathcal{D}_{\mathsf{Yes}}$ and $\mathcal{D}_{\mathsf{No}}$. As we mentioned in the previous section, we need to (re)define the distributions $\mathcal{D}_{\mathsf{Yes}}$ and $\mathcal{D}_{\mathsf{No}}$ (that is, of γ and ν) to solve the second issue:

$\mathcal{D}_{\mathsf{Yes}}$ Draw $f \sim \mathcal{D}_{\mathsf{Yes}}$ by independently drawing, for $i \in [n]$,

$$\gamma_i \stackrel{\text{def}}{=} \begin{cases} +3 & \text{w.p. } \frac{1}{2} \\ +1 & \text{w.p. } \frac{1}{2} \end{cases}$$

and setting $f\colon x \in \{-1,1\}^n \mapsto \text{sign}(\sum_{i=1}^n \gamma_i x_i)$. Any such f is monotone, as the weights are all positive.

$\mathcal{D}_{\mathsf{No}}$ Similarly, draw $f \sim \mathcal{D}_{\mathsf{No}}$ by independently drawing, for $i \in [n]$,

$$\nu_i \stackrel{\text{def}}{=} \begin{cases} +\frac{7}{3} & \text{w.p. } \frac{9}{10} \\ -1 & \text{w.p. } \frac{1}{10} \end{cases}$$

and setting $f\colon x \in \{-1,1\}^n \mapsto \text{sign}(\sum_{i=1}^n \nu_i x_i)$. f is not *always* far from monotone – actually, one of the functions in the support of $\mathcal{D}_{\mathsf{No}}$ (the one with all weights set to 7/3) is even monotone. However, we shall argue that $f \sim \mathcal{D}_{\mathsf{No}}$ is far from monotone with overwhelming probability, and then apply the relaxation of the key tool (HW Problem 4) to conclude.

The theorem will stem from the following two lemmas, that states respectively that (†) No-functions are almost all far from monotone, and (‡) that the two distributions are hard to distinguish:

Lemma 8 (Lemma †)**.** *There exists a universal constant $\epsilon_0 > 0$ such that*

$$\Pr_{f \sim \mathcal{D}_{\mathsf{No}}} [\, \text{dist}(f, \mathcal{M}) > \epsilon_0 \,] \geq 1 - \frac{1}{2^{\Theta(n)}}.$$

(note that this $1 - o(1)$ probability is actually stronger than what the relaxation from Problem 4 requires.)

Lemma 9 (Lemma ‡)**.** *Let $\mathcal{A}$ be any* deterministic *q-query algorithm. Then*

$$\left| \Pr_{f_{\mathsf{Yes}} \sim \mathcal{D}_{\mathsf{Yes}}} [\, \mathcal{A} \text{ accepts} \,] - \Pr_{f_{\mathsf{No}} \sim \mathcal{D}_{\mathsf{No}}} [\, \mathcal{A} \text{ accepts} \,] \right| \leq O\left(\frac{q^{5/4} (\log n)^{1/2}}{n^{1/4}} \right)$$

so that if $q = \tilde{O}\left(n^{1/5}\right)$ the RHS is at most 0.01, which implies with the earlier lemmas and discussion that at least $q+1$ queries are needed for any 2-sided, non-adaptive randomized tester.

Proof of Lemma 8. By an additive Chernoff bound, with probability at least $1 - \frac{1}{2^{\Theta(n)}}$ the random variables ν_i satisfy

$$m \stackrel{\text{def}}{=} |\{\ i \in [n] :\ \nu_i = -1\ \}| \in [0.09n, 0.11n]. \qquad (\star)$$

Say that any linear threshold function for which $(\star)$ holds is *nice*. Fix any nice f in the support of $\mathcal{D}_{\mathsf{No}}$, and rename the variables so that the negative weights correspond to the first variables:

$$f(x) = \text{sign}\left(-(x_1 + \cdots + x_m) + \frac{7}{3}(x_{m+1} + \cdots + x_n)\right), \quad x \in \{-1,1\}^n$$

It is not difficult to show that for this f (remembering that $m = \Theta(n)$), these first variables have high influence – roughly of the same order as for the MAJ function:

Claim 10 (HW Problem)**.** *For $i \in [m]$, $\mathbf{Inf}_i[f] = \Omega\left(\frac{1}{\sqrt{n}}\right)$.* HW Problem

Observe further that f is *unate* (i.e., monotone increasing in some coordinates, and monotone decreasing in the others). Indeed, *any* LTF $g\colon x \mapsto \text{sign}(w \cdot x)$ is unate:

- non-decreasing in coordinate x_i if and only if $w_i \geq 0$;
- non-increasing in coordinate x_i if and only if $w_i \leq 0$.

We saw in previous lectures that, for g monotone, $\hat{g}(i) = \mathbf{Inf}_i[g]$; it turns out the same proof generalizes to unate g, yielding

$$\hat{g}(i) = \pm\mathbf{Inf}_i[g]$$

where the sign depends on whether g is non-decreasing or non-increasing in x_i. Back to our function f, this means that

$$\mathbf{Inf}_i[f] = \begin{cases} +\hat{f}(i) & \text{if } \nu_i = \frac{7}{3} \\ -\hat{f}(i) & \text{if } \nu_i = -1 \end{cases}$$

and thus for all $i \in [m]$ $\hat{f}(i) = -\Omega\left(\frac{1}{\sqrt{n}}\right)$.

Fix any monotone Boolean function g: we will show that $\text{dist}(f, g) \geq \epsilon_0$, for some

choice of $\epsilon_0 > 0$ independent of f and g.

$$\begin{aligned}
4 \cdot \mathrm{dist}(f,g) &= \mathbb{E}_{x \sim \mathcal{U}_{\{-1,1\}^n}}\left[\,(f(x)-g(x))^2\,\right] \underset{\text{(Parseval)}}{=} \sum_{S \subseteq [n]} (\hat{f}(S)-\hat{g}(S))^2 \\
&\geq \sum_{i=1}^{n} (\hat{f}(i)-\hat{g}(i))^2 \geq \sum_{i=1}^{m} (\hat{f}(i)-\hat{g}(i))^2 \underset{(g \text{ mon.})}{=} \sum_{i=1}^{m} \left(-\mathbf{Inf}_i[f]-\mathbf{Inf}_i[g]\right)^2 \\
&= \sum_{i=1}^{m} \left(\mathbf{Inf}_i[f]+\mathbf{Inf}_i[g]\right)^2 \geq \sum_{i=1}^{m} \left(\mathbf{Inf}_i[f]\right)^2 \\
&= \sum_{i=1}^{m} \left(\Omega\left(\frac{1}{\sqrt{n}}\right)\right)^2 = \Omega\left(\frac{m}{n}\right) \\
&= \Omega(1).
\end{aligned}$$

□

Proof (sketch) of Lemma 9. Fix any deterministic, non-adaptive q-query algorithm $\mathcal{A}$; and view its q queries $z^{(1)}, \ldots, z^{(q)} \in \{-1,1\}^n$ as a $q \times n$ matrix $Q \in \{-1,1\}^{q \times n}$, where $z^{(i)}$ corresponds to the i^{th} row of Q.

$$q \left\{ \overbrace{\begin{pmatrix} z_1^{(1)} & z_2^{(1)} & z_3^{(1)} & \cdots & \cdots & \cdots & z_n^{(1)} \\ z_1^{(2)} & z_2^{(2)} & z_3^{(2)} & \cdots & \cdots & \cdots & z_n^{(2)} \\ \vdots & \vdots & \vdots & & \ddots & & \vdots \\ z_1^{(q)} & z_2^{(q)} & z_3^{(q)} & \cdots & \cdots & \cdots & z_n^{(q)} \end{pmatrix}}^{n} \right.$$

Define the "Yes-response vector" R_Y, random variable over $\{-1,1\}^q$, by the process of

(i) drawing $f_{\mathsf{Yes}} \sim \mathcal{D}_{\mathsf{Yes}}$, where $f_{\mathsf{Yes}}(x) = \mathrm{sign}(\gamma_1 x_1 + \cdots + \gamma_n x_n)$;

(ii) setting the i^{th} coordinate of R_Y to $f_{\mathsf{Yes}}(Q_{i,\cdot})$ (f_{Yes} on the i^{th} row of Q, i.e. $z^{(i)}$).

Similarly, define the "No-response vector" R_N over $\{-1,1\}^q$. Via Lemma 2 (the homework problem on total variation distance),

$$(\text{LHS of Lemma 9}) \leq \mathrm{d_{TV}}(R_Y, R_N).$$

(abusing the notation of total variation distance, by identifying the random variables with their distribution.) Hence, our new goal is to show that:

$$\mathrm{d_{TV}}(R_Y, R_N) \underset{?}{\leq} (\text{RHS of Lemma 9}).$$

Multidimensional Berry–Esséen setup. For fixed Q as above, define two random variables $S, T \in \mathbb{R}^q$ as

- $S = Q\gamma$, with $\gamma \sim \mathcal{U}_{\{1,3\}^n}$;
- $T = Q\nu$, with
$$\nu_i = \begin{cases} +\frac{7}{3} & \text{w.p. } \frac{9}{10} \\ -1 & \text{w.p. } \frac{1}{10} \end{cases}$$
for each $i \in [n]$ (independently).

We will also need the following geometric notion:

Definition 11. *An* orthant in $\mathbb{R}^q$ *is the analogue in q-dimensional Euclidean space of a quadrant in the plane $\mathbb{R}^2$; that is, it is a set of the form*

$$\mathcal{O} = \mathcal{O}_1 \times \mathcal{O}_2 \times \cdots \times \mathcal{O}_q$$

where each $\mathcal{O}_i$ is either $\mathbb{R}_+$ or $\mathbb{R}_-$. There are 2^q different orthants in $\mathbb{R}^q$.

The random variable R_Y is fully determined by the orthant S lies in: the i^{th} coordinate of R_Y is the sign of the i^{th} coordinate of S, as $S_i = (Q\gamma)_i = Q_{i,\cdot} \cdot \gamma$. Likewise, R_N is determined by the orthant T lies in. Abusing slightly the notation, we will write $R_Y = \operatorname{sign}(S)$ for $\forall i \in [q]$, $R_{Y,i} = \operatorname{sign}(S_i)$ (and similarly, $R_T = \operatorname{sign}(T)$).

Now, it is enough to show that for any union $\mathcal{O}$ of orthants,

$$|\Pr[\, S \in \mathcal{O}\,] - \Pr[\, T \in \mathcal{O}\,]| \leq O\left(\frac{q^{5/4}(\log n)^{1/2}}{n^{1/4}}\right). \tag{$\diamond$}$$

as this is equivalent to proving that, for any subset $U \subseteq \{-1, 1\}^q$, $|\Pr[\, R_S \in U\,] - \Pr[\, R_T \in U\,]| \leq O\left(\frac{q^{5/4}(\log n)^{1/2}}{n^{1/4}}\right)$ (and the LHS is by definition equal to $\mathrm{d}_{\mathrm{TV}}(R_Y, R_N)$).

Note that for $q = 1$ we get back to the "regular" Berry–Esséen Theorem; for $q > 1$, we will need a "multidimensional Berry–Esséen". The key will be to have random variables with matching means and *covariances* (instead of means and variances for the one-dimensional case).

(Rest of the proof during next lecture.)

□

COMS 6998-3: Sub-Linear Algorithms in Learning and Testing **Spring 2014**

Lecture 10: 03/04/2014

Lecturer: Rocco Servedio *Scribes: Ting-Chu Lin*

1 Administrative

- Sign up for project presentation slot (Doodle)
- Final projects (due May 7)
- New homework problem (due April 30)

2 Overview

2.1 Last Time

Started $\tilde{\Omega}(n^{1/5})$ lower bound non-adaptive monotonicity testers (introducing Yao's principle; Berry–Esséen Theorem, $\mathcal{D}_{\mathsf{Yes}}$, $\mathcal{D}_{\mathsf{No}}$; multidimensional Berry–Esséen Theorem).

2.2 Today

Finish this lower bound (using a multidimensional analogue of the Berry–Esséen Theorem); start testing juntas.

3 Monotonicity testing lower bound: wrapping up

Recall the definitions of $\mathcal{D}_{\mathsf{Yes}}$ and $\mathcal{D}_{\mathsf{No}}$: $f \sim \mathcal{D}_{\mathsf{Yes}}$ is a linear threshold function (LTF) drawn by choosing independently

$$\gamma_i \stackrel{\text{def}}{=} \begin{cases} +3 & \text{w.p. } \frac{1}{2} \\ +1 & \text{w.p. } \frac{1}{2} \end{cases}$$

and setting $f\colon x \in \{-1,1\}^n \mapsto \mathrm{sign}(\gamma \cdot x)$; similarly, for $f \sim \mathcal{D}_{\mathsf{No}}$,

$$\nu_i \stackrel{\text{def}}{=} \begin{cases} +\frac{7}{3} & \text{w.p. } \frac{9}{10} \\ -1 & \text{w.p. } \frac{1}{10} \end{cases}$$

and $f\colon x \in \{-1,1\}^n \mapsto \mathrm{sign}(\nu \cdot x)$. The set of queries Q of any q-query non-adaptive tester will be seen as a $q \times n$ Boolean matrix

$$Q \;=\; q \left\{ \left(\overbrace{\qquad\qquad}^{n} \right. \right. \left. \left. \begin{matrix} \\ \pm 1 \\ \\ \end{matrix} \right) \right.$$

where the i^{th} row $q^{(i)}$ is the i^{th} query string. We also defined the random variables $R_Y, R_N \in \{-1,1\}^q$ by

$$\begin{aligned} (R_Y)_i &= f(q^{(i)}) \qquad \text{for } f \sim \mathcal{D}_{\mathsf{Yes}} \\ (R_N)_i &= f(q^{(i)}) \qquad \text{for } f \sim \mathcal{D}_{\mathsf{No}} \end{aligned}$$

so that by setting $S \stackrel{\text{def}}{=} Q\sigma \in \mathbb{R}^q$ and $T \stackrel{\text{def}}{=} Q\nu \in \mathbb{R}^q$, we get $R_Y = \mathrm{sign}(S)$ and $R_N = \mathrm{sign}(T)$.

Need to show: For S, T as above, for $\mathcal{O}$ any union of orthants in $\mathbb{R}^q$, one has

$$|\Pr[\, S \in \mathcal{O}\,] - \Pr[\, T \in \mathcal{O}\,]| \leq O\left(\frac{q^{5/4}(\log n)^{1/2}}{n^{1/4}}\right) \tag{$\dagger$}$$

as this would imply an $\Omega\left(\frac{n^{1/5}}{\log^{4/5} n}\right)$ lower bound on q for the RHS to be less than 0.01.

Theorem 1 (Original (unidimensional) Berry–Esséen Theorem)**.** *Let $X_1, X_2, \ldots, X_n$ be n independent real-valued random variables such that $|X_i - \mathbb{E}[X_i]| \leq \tau$ almost surely (with probability 1); and let $\mathcal{G}$ be Gaussian with mean and variance matching $S \stackrel{\text{def}}{=} \sum_{i=1}^n X_i$. Then*

$$\forall \theta \in \mathbb{R}, \quad |\Pr[\, S \leq \theta\,] - \Pr[\, \mathcal{G} \leq \theta\,]| \leq \frac{O(\tau)}{\sqrt{\mathrm{Var}\, S}}.$$

Theorem 2 (Multidimensional Berry–Esséen Theorem[1]). *Let $S \stackrel{\text{def}}{=} X^{(1)} + \cdots + X^{(n)}$, where the $X^{(i)}$'s are independent random variables in $\mathbb{R}^q$ with $\|X_j^{(i)} - \mathbb{E}\left[X_j^{(i)}\right]\|_\infty \le \tau$ a.s.; and let $\mathcal{G}$ be a q-dimensional Gaussian with mean and covariance matrix matching those of S. Then, for any $\mathcal{O}$ union of orthants and any $r > 0$,*

$$|\Pr[\,S \in \mathcal{O}\,] - \Pr[\,\mathcal{G} \in \mathcal{O}\,]| \le O\left(\frac{\tau q^{3/2} \log n}{r} + \sum_{i=1}^{q} \frac{r+\tau}{\sqrt{\sum_{j=1}^{n} \operatorname{Var} X_i^{(j)}}}\right) \qquad (\ddagger)$$

Proof of (†) *using* (‡). Note that in (†), $S = Q\sigma$ is the sum of the $\sigma_i \cdot (i^{th}$ column of $Q)$'s, which are independent q-dimensional vector-valued random variables (likewise for $T = Q\nu$). $\sigma_i \cdot (i^{th}$ column of $Q)$ is a q-dim independent vector-valued random variables. So

$$S \cong \mathcal{G}_S, \quad T \cong \mathcal{G}_T$$

by our multidimensional Berry–Esséen theorem. But as the means and covariance matrices of these two Gaussians match (because – as we will prove momentarily – $\mathbb{E}S = \mathbb{E}T$ and $\operatorname{Cov} S = \operatorname{Cov} T$), we get $\mathcal{G}_S \equiv \mathcal{G}_T$ and by the triangle inequality

$$\forall \mathcal{O}, \forall r > 0, \quad |\Pr[\,S \in \mathcal{O}\,] - \Pr[\,\mathcal{G} \in \mathcal{O}\,]| \le 2 \cdot (\text{RHS of } (\ddagger))\,.$$

Hence, it only remains to check the expectations and covariance matrices of S and T do match: we have

$$S = Q\sigma = X^{(1)} + \cdots + X^{(n)}, \text{ where } X^{(j)} \stackrel{\text{def}}{=} \sigma_j \cdot Q_{*,j}$$
$$T = Q\nu = Y^{(1)} + \cdots + Y^{(n)}, \text{ where } Y^{(j)} \stackrel{\text{def}}{=} \nu_j \cdot Q_{*,j}$$

($Q_{*,j} \in \mathbb{R}^q$ denoting the j^{th} column of Q); and it is not hard to see that the expectations are equal termwise, i.e. $\mathbb{E}X^{(j)} = \mathbb{E}Y^{(j)}$ for all $j \in [n]$:

$$\mathbb{E}X^{(j)} = \frac{1}{2} \cdot 1 \cdot Q_{*,j} + \frac{1}{2} \cdot 3 \cdot Q_{*,j} = 2Q_{*,j}$$
$$\mathbb{E}Y^{(j)} = \frac{1}{10} \cdot (-1) \cdot Q_{*,j} + \frac{9}{10} \cdot \frac{7}{3} \cdot Q_{*,j} = 2Q_{*,j}$$

so $\mathbb{E}S = \mathbb{E}T$. As for the covariance matrices, as for any random variable $Z \in \mathbb{R}^q$ by definition

$$(\operatorname{Cov} Z)_{k,\ell} = \mathbb{E}[(Z_k - \mathbb{E}Z_k)\,(Z_\ell - \mathbb{E}Z_\ell)] = \mathbb{E}[Z_k Z_\ell] - \mathbb{E}[Z_k] \cdot \mathbb{E}[Z_\ell]$$

[1] From [Chen–Servedio–Tan'14], building upon a Central Limit Theorem for Earthmover distance of [**?**]

for all $k, \ell \in [q]$, one can check that, using the independence of the $X^{(j)}$'s (resp. $Y^{(j)}$'s),

$$\forall j \in [n], \quad (\operatorname{Cov} X^{(j)})_{k,\ell} = (\operatorname{Cov} Y^{(j)})_{k,\ell} = Q_{k,j} Q_{\ell,j}$$

and hence $\operatorname{Cov} X^{(j)} = \operatorname{Cov} Y^{(j)}$; so that (again by independence) $\operatorname{Cov} S = \sum_{j=1}^{n} \operatorname{Cov} X^{(j)} = \sum_{j=1}^{n} \operatorname{Cov} Y^{(j)} = \operatorname{Cov} T$.
This finally results in

$$\forall r > 0, \qquad |\Pr[\, S \in \mathcal{O}\,] - \Pr[\, \mathcal{G} \in \mathcal{O}\,]| \leq O\left(\frac{\tau q^{3/2} \log n}{r} + q \cdot \frac{r+\tau}{\sqrt{n}}\right)$$

(as $\operatorname{Var} X_i^{(j)} = \operatorname{Var} Y_i^{(j)} = 1$) which holds for any r. Taking $r = (qn)^{1/4}\sqrt{\log n}$, the RHS becomes $O\left(\frac{q^{5/4}(\log n)^{1/2}}{n^{1/4}}\right)$. □

4 Testing Juntas

We will now describe and analyze an algorithm for testing juntas (recall that a *k-junta* is a Boolean function with at most k relevant variables).
Let us write $\mathcal{J}_k$ for the class of all k-juntas over $\{-1,1\}^n$ (where k can depend on n); from earlier lectures, we know that one can *learn* $\mathcal{J}_k$ with $2^k \log n$ (membership) queries. As we shall see, however, testing is significantly more query-efficient:

Theorem 3. *There is an $O\big(k \log k + \frac{k}{\epsilon}\big)$-query (one-sided) algorithm for testing $\mathcal{J}_k$.*

Remark 1. *Next time, we will prove an $\Omega(k)$ lower bound for this problem, which shows this theorem is roughly optimal.*

4.1 Setup

Let $S \subseteq [n]$ be a set of variables, and $\bar{S} = [n] \setminus S$. For $x, y \in \{-1,1\}^n$, we write $y_S x_{\bar{S}}$ for the string in $\{-1,1\}^n$ which has for i^{th} coordinate

$$(y_S x_{\bar{S}})_i \stackrel{\text{def}}{=} \begin{cases} x_i & \text{if } i \notin S \\ y_i & \text{if } i \in S \end{cases}.$$

Definition 4. *Given $f\colon \{-1,1\}^n \to \{-1,1\}$ and $S \subseteq [n]$, then $\mathbf{Inf}_f(S)$ is defined as*

$$\mathbf{Inf}_f(S) = 2 \Pr_{x,y \sim \mathcal{U}_{\{-1,1\}^n}} [\, f(y_S x_{\bar{S}}) \neq f(x)\,].$$