LECTURE 19

Last time

- Finish testing linearity of Boolean functions
  [Blum Luby Rubinfeld]
- Tolerant testing and distance estimation

Today

- Estimating distance to sortedness for 0/1 sequences (equivalently, estimating the length of LIS)

Final project reports are due Thursday, April 24

Approximating Distance to Monotonicity for 0/1 Sequences

Input: Parameter $\varepsilon \in (0,1/2]$ and

a list of n zeros and ones (equivalently, $f: [n] \to \{0,1\}$)

Question: How far is this list to being sorted?

(Equivalently, how far is f from monotone?)

dist$(f, MONO)$ =distance from f to monotone

Dist$(f, MONO) = n \cdot$ dist$(f, MONO)$

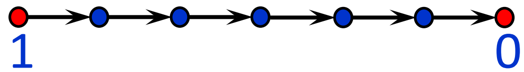
Note: Dist$(f, MONO) = n - |LIS|$,
where LIS is the longest increasing subsequence

Output: An estimate $\hat{\varepsilon}$ such that w.p. $\geq \frac{2}{3}$

$$|\hat{\varepsilon} - \text{dist}(f, MONO)| \leq \varepsilon$$

Today: can answer in $O\left(\frac{1}{\varepsilon^2}\right)$ time [Berman Raskhodnikova Yaroslavtsev]

Distance to Monotonicity over POset Domains

- Let f be a function over a partially ordered domain D.
- Violated pair: 

- The violation graph G_f is a directed graph with vertex set D whose edge set is the set of pairs (x, y) violated by f.
- VC_f is a minimum vertex cover of G_f
- MM_f is a maximum matching in G_f

Characterization of $Dist(f, \mathrm{Mono})$ for $f: D \to \{0,1\}$ [FLNRRS 02]

$$Dist(f, \mathrm{Mono}) = \left|\mathrm{MM}_f\right| = |VC_f|$$

Distance to Monotonicity for 0/1 Sequences

- Let $f : [n] \to \{0,1\}$
- Great notation switch: $g_i = (-1)^{f(i)}$ for $i \in [n]$
- Cumulative sums: $s_0 = 0$ and $s_i = s_{i-1} + g_i$ for $i \in [n]$
- Final sum: $s_f = s_n$
- Maximum sum: $m_f = \max_{i=0}^{n} s_i$

dist$(f, Mono)$ for $f : [n] \to \{0,1\}$ [Berman Raskhodnikova Yaroslavtsev]

$$Dist(f, \mathrm{Mono}) = \frac{n - 2m_f + s_f}{2}$$

Distance to Monotonicity: Algorithm

Algorithm (Input: ε, n; query $acess\ to\ f: [n] \to \{0,1\}$

1. Sample a random subset $\boldsymbol{S} \subset [n]$
   where each element is independently included with probability $\frac{s}{n}$
2. Let $\tilde{f} = f_{|S}$
3. Compute $\tilde{\varepsilon} = Dist(\tilde{f}, Mono)/s$
4. **Return** $\tilde{\varepsilon}$

- Let $\varepsilon_f = dist(f, Mono) = Dist(f, Mono)/n$

Theorem

$$\varepsilon_f - \sqrt{2\varepsilon_f/s} \leq \mathbb{E}[\tilde{\varepsilon}] \leq \varepsilon_f$$

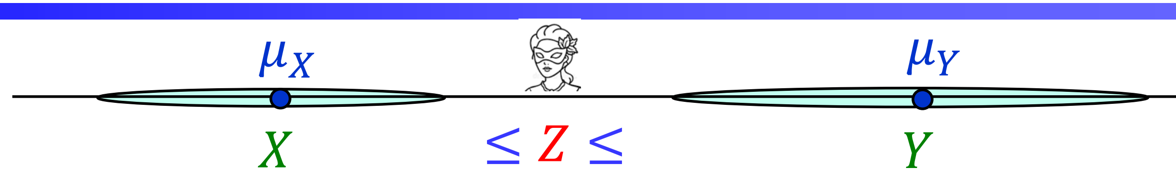
$$\text{Var}[\tilde{\varepsilon}] = O(\varepsilon_f/s)$$

Proof idea: Let $Z(S) = Dist(\tilde{f}, Mono)$

We'll define random variables $X(S)$ and $\mathrm{Y}(S)$, such that $X(S) \leq Z(S) \leq Y(S)$

$X(S)$ will be in terms of matching MM_f; $Y(S)$ in terms of vertex cover VC_f

Sandwiching a Random Variable



Sandwich Lemma

Let X, Y, Z be discrete random variables with means μ_X, μ_Y, μ_Z, satisfying

$$X \leq Z \leq Y.$$

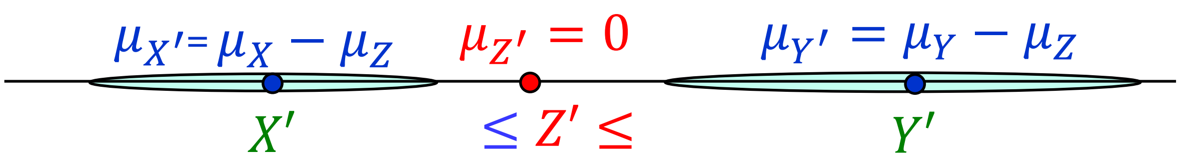
Then $\mathbb{E}[X] \leq \mathbb{E}[Z] \leq \mathbb{E}[Y]$ *and*

By definition of expectation

$$\mathrm{Var}[Z] \leq \mathrm{Var}[X] + \mathrm{Var}[Y] + (\mu_Y - \mu_X)^2.$$

Proof: Shift the variables to make Z unbiased

subtract μ_Z from each to get X', Y', and Z'



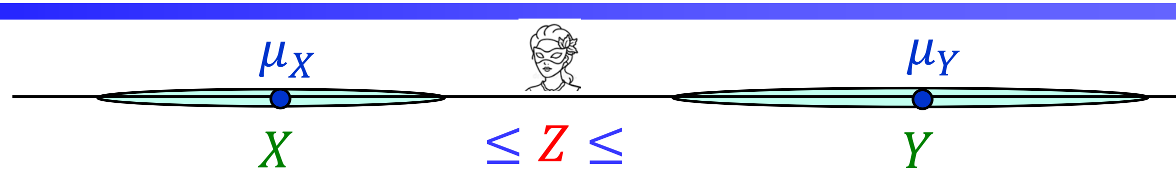
$V^- =$ set of negative values and $V^+ =$ set of positive values in the support of Z'.

$$\mathrm{Var}[Z] = \mathrm{Var}[Z'] = \mathbb{E}[(Z')^2] = \sum_{v \in V^-} \Pr[Z' = v] \cdot v^2 + \sum_{v \in V^+} \Pr[Z' = v] \cdot v^2$$

$$\leq \mathbb{E}[(X')^2] + \mathbb{E}[(Y')^2]$$

because $X' \leq Z' \leq Y'$

Sandwiching a Random Variable



Sandwich Lemma

Let X, Y, Z be discrete random variables with means μ_X, μ_Y, μ_Z, satisfying

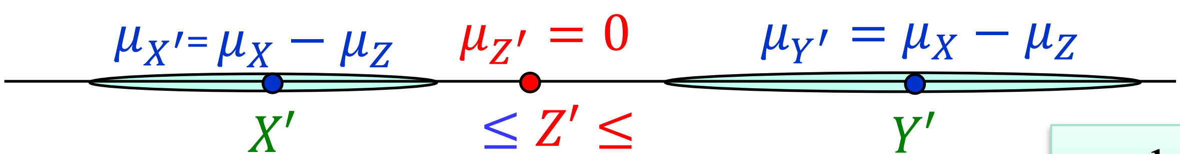
$$X \leq Z \leq Y.$$

Then $\quad \mathbb{E}[X] \leq \mathbb{E}[Z] \leq \mathbb{E}[Y] \quad$ *and*

$$\mathrm{Var}[Z] \leq \mathrm{Var}[X] + \mathrm{Var}[Y] + (\mu_Y - \mu_X)^2.$$

Proof: Shift the variables to make Z unbiased

subtract μ_Z from each to get X', Y', and Z'



by definition of variance

$$\mathrm{Var}[Z] \leq \mathbb{E}[(X')^2] + \mathbb{E}[(Y')^2] = \mathrm{Var}[X'] + {\mu_{X'}}^2 + \mathrm{Var}[Y'] + {\mu_{Y'}}^2$$

$$= \mathrm{Var}[X] + (\mu_Z - \mu_Y)^2 + \mathrm{Var}[Y] + (\mu_Y - \mu_Z)^2$$

$$\leq \mathrm{Var}[X] + \mathrm{Var}[Y] + (\mu_X - \mu_Y)^2$$

$a^2 + b^2 \leq (a+b)^2$ for nonnegative a, b

Distance to Monotonicity: Algorithm

Algorithm (Input: ε, n; query *acess to* $f: [n] \to \{0,1\}$**)**

1. Sample a random subset $S \subset [n]$
   where each element is included w.p. s/n independently
2. Let $\tilde{f} = f_{|S}$
3. Compute $\tilde{\varepsilon} = Dist(\tilde{f}, Mono)/s$
4. **Return** $\tilde{\varepsilon}$

- Let $\varepsilon_f = dist(f, Mono) = Dist(f, Mono)/n$

Theorem

$$\varepsilon_f - \sqrt{2\varepsilon_f/s} \leq \mathbb{E}[\tilde{\varepsilon}] \leq \varepsilon_f$$

$$\text{Var}[\tilde{\varepsilon}] = O(\varepsilon_f/s)$$

Proof idea: Let $Z(S) = Dist(\tilde{f}, Mono)$

We'll define random variables $X(S)$ and $\mathrm{Y}(S)$, such that $X(S) \leq Z(S) \leq Y(S)$

$X(S)$ will be in terms of matching MM_f; $Y(S)$ in terms of vertex cover VC_f

Upper Bound on $Z(S)$

- Define $Y(S) = |VC_f \cap S|$

Upper Bound Lemma

(a) $Z(S) \leq Y(S)$, (b) $\mathbb{E}[Y(S)] = \varepsilon_f \cdot s$ and $\mathrm{Var}[Y(S)] \leq \varepsilon_f \cdot s$

Proof: (a) $Z(S) = Dist\left(\tilde{f}, Mono\right) = \left|\mathrm{VC}_{\tilde{f}}\right|$

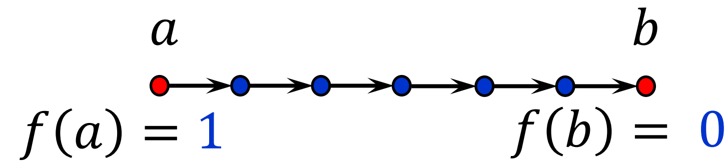
- Each pair violated by $\tilde{f}$ is also violated by f
- $VC_f \cap S$ is a vertex cover (not necessarily minimum) of $G_{\tilde{f}}$

$$Z(S) = Dist\left(\tilde{f}, Mono\right) = \left|\mathrm{VC}_{\tilde{f}}\right| \leq \left|VC_f \cap S\right| = Y(S)$$

(b) Recall that $\left|VC_f\right| = \varepsilon_f \cdot n$

- Each element of VC_f appears in S independently w.p. s/n
- $Y(S)$ is binomial with mean $\left|VC_f\right| \cdot \frac{s}{n} = \varepsilon_f \cdot s$ and variance $\left|VC_f\right| \cdot \frac{s}{n}\left(1 - \frac{s}{n}\right) \leq \varepsilon_f \cdot s$

Lower Bound on $Z(S)$



- Let $\ell = |MM_f| = \varepsilon_f \cdot n$
- MM_f consists of ℓ pairs of the form (a, b)
- Let $a_1 < a_2 < \cdots < a_\ell$ be the lower endpoints of pairs in MM_f
- Let $b_1 < b_2 < \cdots < b_\ell$ be the upper endpoints of pairs in MM_f
- Then $a_i < b_i$ for all $i \in [\ell]$
- Guaranteed edges are pairs of the form (a_i, b_j) where $i \leq j$

$\boxed{f(a_i) = 1}$

$\boxed{f(b_i) = 0}$

- Let $\widetilde{MM}(S)$ denote a maximum matching that consists of guaranteed edges
- Define $X(S) = |\widetilde{MM}(S)|$

Lower Bound Lemma

(a) $X(S) \leq Z(S)$, (b) $\mathbb{E}[X(S)] \geq \varepsilon_f \cdot s - \sqrt{2\varepsilon_f \cdot s}$ and $\mathrm{Var}[X(S)] = O(\varepsilon_f \cdot s)$

Proof of Lower Bound Lemma: Random Walk

- Recall: $X(S) = |\widetilde{MM}(S)|$
- Let $X'(S) = |V(MM_f) \cap S|$
- $U(S) =$ number of elements of $V(MM_f) \cap S$ left unmatched by $\widetilde{MM}(S)$
- Then $X(S) = \frac{X'(S) - U(S)}{2}$
- $X'(S)$ is binomial with mean $2\varepsilon_f \cdot s$ and variance $\leq 2\varepsilon_f \cdot s$
- To understand $U(S)$ define a random walk that at step $i \in [\ell]$ moves by

$$g_i = \begin{cases} 1 & \text{if } \{a_i, b_i\} \cap S = \{b_i\} \\ -1 & \text{if } \{a_i, b_i\} \cap S = \{a_i\} \\ 0 & \text{otherwise} \end{cases}$$

- Define $M(S) =$ the maximum value reached by the walk
- Define $m(S) =$ the absolute value of the minimum of the ``opposite order'' walk that makes moves $g_\ell, \ldots, g_1$

Claim

$U(S) \leq M(S) + m(S)$

Proof idea: Construct a matching of guaranteed edges that only leaves $M(S) + m(S)$ elements of $V(MM_f) \cap S$ unmatched

Analyzing $M(S)$ and $m(S)$

- By symmetry $M(S)$ and $m(S)$ have the same distribution
- Define $p(S)$ = the final position reached by the walk

Claim 2

$\Pr[M(S) \geq z] \leq \Pr[|p(S)| \geq z]$ for all $z \in [\ell]$

Proof: Define p_i to be the position after i steps.

- Let E_i be the event that $p_i = z$ and $p_j < z$ for all $j \in [i-1]$.
- The event ``$M(S) \geq z$'' is a disjoint union of the events E_i for $i \in [\ell]$.
- By symmetry, $\Pr[\, p(S) \geq z \mid E_i \,] \geq \frac{1}{2}$ for all $i \in [\ell]$.
- Thus, $\Pr[|p(S)| \geq z] = \Pr[p(S) \geq z] + \Pr[p(S) \leq -z]$

$$= 2 \Pr[p(S) \geq z]$$

$$= 2 \sum_{i \in [\ell]} \Pr[p(S) \geq z \mid E_i] \cdot \Pr[E_i]$$

$$\geq \sum_{i \in [\ell]} \Pr[E_i] \quad = \Pr[M(S) \geq z]$$

by symmetry

by law of total probability

Analyzing the Expectation of $U(S)$ and $X(S)$

Claim

$U(S) \le M(S) + m(S)$

Claim 2

$\Pr[M(S) \ge z] \le \Pr[|p(S)| \ge z]$ for all $z \in [\ell]$

$$\mathbb{E}[U] \le \mathbb{E}[M(S) + m(S)] \le 2\mathbb{E}[|p(S)|] \le 2\sqrt{2\varepsilon_f \cdot s}$$

- Recall: $p(S)$ is the sum of $\ell = \varepsilon_f \cdot n$ independent zero-mean variables g_i that take values in $\{-1,0,1\}$ and $\Pr[g_i = 1] = \Pr[g_i = -1] \le \frac{s}{n}$

$$\mathbb{E}[g_i^2] = \Pr[g_i^2 = 1] \le \frac{2s}{n}$$

$$\mathbb{E}^2[|p(S|] \le \mathbb{E}[(p(S)^2]$$

by Jensen's inequality

$$= \mathbb{E}\left[\left(\sum_{i \in [\ell]} g_i\right)^2\right] = \mathbb{E}\left[\sum_{i \in [\ell]} g_i^2\right] \le \ell \cdot \frac{2s}{n} = 2\varepsilon_f \cdot s$$

- Recall: $X(S) = [X'(S) - U(S)]/2$ and $\mathbb{E}[X'(S)] = 2\varepsilon_f \cdot s$

$$\mathbb{E}[X(S)] = \frac{\mathbb{E}[X'(S)]}{2} + \frac{\mathbb{E}[U(S)]}{2} \ge \varepsilon_f \cdot s - \sqrt{2\varepsilon_f \cdot s}$$

Analyzing the Variance of $U(S)$ and $X(S)$

Claim

$U(S) \le M(S) + m(S)$

Claim 2

$\Pr[M(S) \ge z] \le \Pr[|p(S)| \ge z]$ for all $z \in [\ell]$

$$\mathbb{E}[U] \le \mathbb{E}[M(S) + m(S)] \le 2\mathbb{E}[|p(S)|] \le 2\sqrt{2\varepsilon_f \cdot s}$$

$$\mathbb{E}[(p(S)^2] = 2\varepsilon_f \cdot s$$

$$\mathbb{E}[X(S)] = \frac{\mathbb{E}[X'(S)]}{2} + \frac{\mathbb{E}[U(S)]}{2} \ge \varepsilon_f \cdot s - \sqrt{2\varepsilon_f \cdot s}$$

- Analyzing the variance of $U(S)$

$$\mathrm{Var}[U(S)] \le \mathbb{E}[(U(S)^2] \le 4 \cdot \mathbb{E}[(p(S)^2] \le 8\varepsilon_f \cdot s$$

- Standard deviation: $\sigma(U(S)) \le \sqrt{8\varepsilon_f \cdot s}$
- Since $X(S) = [X'(S) - U(S)]/2$ \$ and $\sigma(X'(S)) \le \sqrt{2\varepsilon_f \cdot s}$,

$$\sigma(X(S)) \le \frac{1}{2}(\sqrt{8} + \sqrt{2})\sqrt{\varepsilon_f \cdot s} \le \sqrt{4.5\ \varepsilon_f \cdot s}$$

by subadditivity of standard deviation

Completing the Analysis

Lower Bound Lemma

(a) $X(S) \le Z(S)$, (b) $\mathbb{E}[X(S)] \ge \varepsilon_f \cdot s - \sqrt{2\varepsilon_f \cdot s}$ and $\mathrm{Var}[X(S)] \le \sqrt{4.5\, \varepsilon_f / s}$

Upper Bound Lemma

(a) $Z(S) \le Y(S)$, (b) $\mathbb{E}[Y(S)] = \varepsilon_f \cdot s$ and $\mathrm{Var}[Y(S)] \le \varepsilon_f \cdot s$

Theorem

$$\varepsilon_f - \sqrt{2\varepsilon_f / s} \le \mathbb{E}[\tilde{\varepsilon}] \le \varepsilon_f$$

$$\mathrm{Var}[\tilde{\varepsilon}] \le \sqrt{7.5\, \varepsilon_f / s}$$

Distance to Monotonicity: Algorithm

Algorithm (**Input:** ε, n; query $acess\ to\ f: [n] \rightarrow \{0,1\}$)

1. Sample a random subset $\boldsymbol{S} \subset [n]$

   where each element is included w.p. s/n independently
2. Let $\tilde{f} = f_{|S}$
3. Compute $\tilde{\varepsilon} = Dist(\tilde{f}, Mono)/s$
4. **Return** $\tilde{\varepsilon}$

- Let $\varepsilon_f = dist(f, Mono) = Dist(f, Mono)/n$

Theorem

$$\varepsilon_f - \sqrt{2\varepsilon_f/s} \leq \mathbb{E}[\tilde{\varepsilon}] \leq \varepsilon_f$$

$$\text{Var}[\tilde{\varepsilon}] = O(\varepsilon_f/s)$$