Communication Complexity and Applications

Lecturer: Toniann Pitassi

1 Randomized Communication Complexity

1.1 Definitions

A *(private coin) randomized protocol* is a protocol where Alice and Bob have access to random strings r_A and r_B, respectively. These two strings are chosen independently, according to some probability distribution. We can classify randomized protocols by considering different types of error:

- *zero-error protocol* $\mathcal{P}$:
$$\forall x, y \Pr_{r_A, r_B}[\mathcal{P}(x, r_A, y, r_B) = f(x, y)] = 1$$
- *ϵ-error protocol* $\mathcal{P}$:
$$\forall x, y \Pr_{r_A, r_B}[\mathcal{P}(x, r_A, y, r_B) = f(x, y)] \geq 1 - \epsilon$$
- *one-sided ϵ-error protocol* $\mathcal{P}$:
$$\begin{aligned} \forall x, y: \quad & f(x, y) = 0 \Rightarrow \Pr_{r_A, r_B}[\mathcal{P}(x, r_A, y, r_B) = 0] = 1 \\ & f(x, y) = 1 \Rightarrow \Pr_{r_A, r_B}[\mathcal{P}(x, r_A, y, r_B) = 1] \geq 1 - \epsilon \end{aligned}$$

Due to randomization, the number of bits exchanged may differ in different executions of the protocol on the same input (x, y). So, there are two natural choices for measuring the running time of a randomized protocol:

- The *worst case running time $\mathcal{P}$ on input* (x, y) is the maximum number of bits communicated over all choices of the random strings r_A and r_B. The *worst case cost of* $\mathcal{P}$ is the maximum, over all inputs (x, y), of the worst case running time of $\mathcal{P}$ on (x, y).
- The *average case running time $\mathcal{P}$ on input* (x, y) is the expected number of bits communicated over all choices of the random strings r_A and r_B. The *average case cost of* $\mathcal{P}$ is the maximum, over all inputs (x, y), of the average case running time of $\mathcal{P}$ on (x, y).

So, for a function $f : X \times Y \to \{0, 1\}$, we define the following complexity measures. All of these definitions are for private coin protocols.

- $R_0(f)$ is the minimum average case cost of a randomized protocol that computes f with zero error.
- For $0 < \epsilon < \frac{1}{2}$, $R_\epsilon(f)$ is the minimum worst case cost of a randomized protocol that computes f with error ϵ.
- For $0 < \epsilon < 1$, $R^1_\epsilon(f)$ is the minimum worst case cost of a randomized protocol that computes f with one-sided error ϵ.

These lead naturally to the following complexity classes:

- $ZPP^{cc} = \{f \mid R_0(f) \in O(\text{polylog}(n))\}$
- $BPP^{cc} = \{f \mid R_\epsilon(f) \in O(\text{polylog}(n))\}$
- $RP^{cc} = \{f \mid R^1_\epsilon(f) \in O(\text{polylog}(n))\}$

Analogous definitions hold in a *public coin* model, that is, a model where both Alice and Bob see the results of a single series of random coin flips. A randomized protocol in the public coin model can be viewed as a distribution of deterministic protocols, that is, Alice and Bob choose together a string r (according to a probability distribution Π, and independently of x and y) and then follow the deterministic protocol P_r. The *success probability* of a public coin protocol on input (x, y) is the probability of choosing a deterministic protocol, according to the probability distribution Π, that computes $f(x, y)$ correctly. We use the same complexity measures as in the private coin model, but add a superscript 'pub', i.e., $R_0^{pub}(f)$, $R_\epsilon^{pub}(f)$, $R_\epsilon^{1\ pub}(f)$. We have previously seen the following facts:

- $R_\epsilon^{pub}(f) \leq R_\epsilon(f)$
- for every $\delta > 0$ and every $\epsilon > 0$, $R_{\epsilon+\delta}(f) \leq R_\epsilon^{pub}(f) + O(\log n + \log \delta^{-1})$

1.2 Distributional Complexity

Let μ be a probability distribution over $X \times Y$, $X = \{0,1\}^n$, $Y = \{0,1\}^n$. The (μ, ϵ)-*distributional communication complexity* of f, $D_\epsilon^\mu(f)$, is the cost of the best deterministic protocol that gives the correct answer for f on at least a $(1-\epsilon)$ fraction of all inputs in $X \times Y$, weighted by μ.

Theorem 1 $R_\epsilon^{pub}(f) = \max_\mu D_\epsilon^\mu(f)$

Proof First, we show that $R_\epsilon^{pub}(f) \geq \max_\mu D_\epsilon^\mu(f)$. Let $\mathcal{P}$ be a randomized public coin protocol with worst-case cost $R_\epsilon^{pub}(f)$ that computes f with success probability at least $1-\epsilon$ for every input (x, y). Therefore, if Π is the probability distribution of $\mathcal{P}$'s public coin flips,

$$\Pr_{r \in \Pi, (x,y) \in (X \times Y)_\mu} (\mathcal{P}_r(x, y) = f(x, y)) \geq 1 - \epsilon$$

By a counting argument, there exists a fixed choice of public coin flips r' such that

$$\Pr_{(x,y) \in (X \times Y)_\mu} (\mathcal{P}_{r'}(x, y) = f(x, y)) \geq 1 - \epsilon$$

Thus, $\mathcal{P}_{r'}$ is a deterministic protocol that gives the correct answer for f on at least a $1-\epsilon$ fraction of all inputs in $X \times Y$, weighted by μ. So, $R_\epsilon^{pub}(f) \geq \text{cost}(\mathcal{P}_{r'}) \geq \max_\mu D_\epsilon^\mu(f)$.

Next, we show that $R_\epsilon^{pub}(f) \leq \max_\mu D_\epsilon^\mu(f)$. Let $c = \max_\mu D_\epsilon^\mu(f)$.

1.2.1 Minimax Theorem

We will show this direction of the theorem by an application of Von Neumann's Minimax Theorem. In a two-player, zero-sum game, there are two players, $P1$ and $P2$. $P1$ has a finite set $A = \{a_1, \ldots, a_m\}$ of pure strategies, and $P2$ has a finite set of pure strategies, $B = \{b_1, \ldots, b_n\}$. Each player has a utility for each pair (a_i, b_j) of actions. The utility for $P1$ is denoted by $U_1(a_i, b_j)$ and the utility for $P2$ is denoted by $U_2(a_i, b_j)$. It is a zero-sum game if for all i, j $U_1(a_i, b_j) = -U_2(a_i, b_j)$. In our case, for each (a_i, b_j), one of the players will win and the other one will lose.

Each player can use a mixed strategy by creating a probability mass function and playing each pure strategy with a fixed probability. Let p_i denote the probability that $P1$ plays action a_i and let q_j denote the probability that $P2$ plays action b_j. Since p and q are probabilities, we have that each $p_i, q_j \geq 0$, and the sum of the p_i's is 1, and the sum of the q_j's is 1. A mixed strategy for $P1$ will be denoted by p, and similarly q denotes a mixed strategy for $P2$. For each mixed strategy pair (p, q), the payoff $M(p, q)$ is defined to be

$$\sum_{i=1}^{m} \sum_{j=1}^{n} p_i M(a_i, b_j) q_j.$$

When $P1$ uses pure strategy a_i and $P2$ uses mixed strategy q, then $M(a_i, q) = \sum_{j=1}^{n} M(a_i, b_j) q_j$, and analogously for $M(p, b_j)$. We let P and Q denote the set of all mixed strategies available to player 1 and 2 respectively. Player $P1$'s objective is to select a mixed strategy $p \in P$ soas to maximize $min_q M(p, q)$, and at the same time $P2$'s objective is to select a mixed strategy $q \in Q$ soas to minimize $max_p M(p, q)$.

The Minimax theorem states that for every two-person zero-sum game, there exists an equilibrium strategy. That is there exists a value v such that

$$max_p min_q M(p, q) = min_q max_p M(p, q)$$

In other words, in every two-person zero-sum game with finite strategies, there exists a value v and a mixed strategy for each player such that: (a) given Player 2's strategy, the best payoff for Player 1 is v, and (b) given Player 1's strategy, the best payoff for Player 2 is $-v$.

In our context, we define a two-player zero-sum game as follows:

- $P1$ (the protocol designer): his pure strategies are all c-bit deterministic protocols $\mathcal{P}_\nabla$, one for each choice of coin flips. His mixed strategies are all randomized protocols, P, (each of which is a distribution over the deterministic protocols).
- $P2$ (the adversary): her pure strategies are all inputs (x, y). Her mixed strategies are all distributions μ over the inputs.
- $P1$ has payoff 1 if $\mathcal{P}_r(x, y) = f(x, y)$ and -1 otherwise. That is, the designer ($P1$) wins the game iff this protocol is correct on (x, y), and otherwise $P2$ wins.

We are given as our assumption that for all distributions μ over inputs (x, y), there exists a pure strategy (a protocol) P such that the probability of a win is at least $1 - \epsilon$. This means that $Min_\mu Max_P M(\mu, P) \geq 1 - \epsilon$. (Since for each choice of μ, there is a fixed strategy P_r that achieves payoff $1-\epsilon$, so no matter what μ we choose, the designer will be able to come up with a protocol that wins $1-\epsilon$ of the time. Now by the Minimax theorem, this means that $Max_P Min_\mu M(\mu, P) \geq 1-\epsilon$.

From this it follows that there is a randomized strategy P such that for all fixed (x, y), the payoff is at least $1 - \epsilon$.

Theorem 1 is useful because, for any choice of μ, a lower bound for D_ϵ^μ gives a lower bound on $R_\epsilon^{pub}(f)$.

Definition A distribution μ over $X \times Y$ is a *product distribution* if $\mu(x, y) = \mu_X(x) \cdot \mu_Y(y)$ for some distributions μ_X over X and μ_Y over Y. Let $R^{[\ ]}(f) = \max_\mu D^\mu(f)$, where the maximum is taken over all product distributions μ.

Exercise: Prove that $R_\epsilon^{[\ ]}(DISJ) = O(\sqrt{n} \log n)$. On the other hand, show that $R_\epsilon(DISJ) = \Theta(n)$.

Sherstov showed a separation between product and non-product distributional complexity by proving the existence of a function f such that $R^{[\ ]}(f) = \Theta(1)$ but $R_\epsilon(f) = \Theta(n)$.

2 Discrepancy

We now consider a technique for proving lower bounds for D_ϵ^μ. It consists of finding an upper bound for the size of rectangles in M_f that are "almost" monochromatic. If we can prove that all such rectangles for a given function f are small, then we need a lot of rectangles to "cover" the function.

Definition Let $f : X \times Y \to \{0, 1\}$ be a function, R be any rectangle, and μ be a probability distribution on $X \times Y$.

$$Disc_\mu(R) = |\mu(R \cap f^{-1}(1)) - \mu(R \cap f^{-1}(0)|.$$

The discrepancy of f under μ is the maximum over all possible rectangles:

$$Disc_\mu(f) = max_R Disc_\mu(R).$$

If f has small discrepancy it means (informally) that all large rectangles are roughly balanced.

Consider a deterministic protocol that partitions the input space into rectangles $R_1, \ldots, R_{2^c}$. And suppose it has success probability $2/3$ with respect to μ. The best thing that the protocol can do if it has to give one output a_i for all inputs in the rectangle R_i is to set a_i to the bit value with the highest weight in that rectangle. This contributes $\mu(R_i \cap f^{-1}(a_i))$ to the success probability and $\mu(R_i \cap f^{-1}(1 - a_i))$ to the failure probability. Thus the overall success probability is $\sum_i \mu(R_i \cap f^{-1}(a_i))$ and the overall error probability is $\sum_i \mu(R_i \cap f^{-1}(1 - a_i))$. Since the difference between these two has to be at least $2/3 - 1/3 = 1/3$, we have

$$
\begin{aligned}
1/3 &\leq \sum_{i=1}^{2^c} \mu(R_i \cap f^{-1}(a_i)) - \sum_{i=1}^{2^c} \mu(R_i \cap f^{-1}(1-a_i)) && (1)\\
&\leq \sum_{i=1}^{2^c} |\mu(R_i \cap f^{-1}(a_i)) - \mu(R_i \cap f^{-1}(1-a_i))| && (2)\\
&= \sum_{i=1}^{2^c} Disc_\mu(R_i) && (3)\\
&\leq 2^c Disc_\mu(f). && (4)
\end{aligned}
$$

This gives a lower bound on communication: $c \geq \log(1/3Disc_\mu(f))$. To get a lower bound for randomized protocols, it suffices to find a distribution μ such that $Disc_\mu(f)$ is small.

We have proved

Theorem 2 *For every distribution μ, $R_\mu(f) \geq \log(1/3Disc_\mu(f))$.*

We now demonstrate how to prove a lower bound for the inner product (IP) function by calculating the discrepancy of IP according to the uniform distribution. Before we prove this result, we will study the communication matrix for the IP function for $n = 3$ to get some intuition. We will actually switch things a little bit and analyze the matrix whose (x, y) entry is $(-1)^{x \cdot y}$. This is just the communication matrix for IP, with 0's replaced by 1's and 1's replaced by -1's. With this switch of basis, The associated IP matrices are the Hadamard matrices. Hadamard matrices are defined to be square matrices where each entry is either $+1$ or -1 and such that all pairs of rows are mutually orthogonal.

The IP matrix, H_n, for $n = 3$ looks like this:

1	1	1	1	1	1	1	1
1	-1	1	-1	1	-1	1	-1
1	1	-1	-1	1	1	-1	-1
1	-1	-1	1	1	-1	-1	1
1	1	1	1	-1	-1	-1	-1
1	-1	1	-1	-1	1	-1	1
1	1	-1	-1	-1	-1	1	1
1	-1	-1	1	-1	1	1	-1

More generally $H_0 = [1]$ and H_n is built from H_{n-1} as follows: the lower right quadrant of H_n is equal to $-H_{n-1}$ and the other three quadrants are equal to H_{n-1}.

The following facts are easy to prove about H_n:

- Every pair of rows is orthogonal, and therefore $H_n^2 = N \cdot I$.
- We can interpret the rows as parity functions
- The matrix is symmetric about the diagonal
- The eigenvectors form an orthonormal basis. (That is $< v_i, v_j >= 0$ for all $i \neq j$, and $v_i^2 = 1$ for all i.)

- The only eigenvalues of H_n are $+/-\sqrt{N}$.

We want to find the eigenvalues of the Hadamard matrices, as claimed in the last bullet point above. Recall these are defined by the following recursive construction:

$$H_0 = [1], \qquad H_{n+1} = \begin{bmatrix} H_n & H_n \\ H_n & -H_n \end{bmatrix}.$$

Lemma 3 *For each n, $H_n^2 = HH^T = 2^n I_{2^n}$.*

Proof The proof is by induction. Since $H_0 = I_1$, the lemma is correct for $n = 0$.

Given that $H_n^2 = 2^n I$, we can calculate H_{n+1}^2 explicitly:

$$\begin{aligned} H_{n+1}^2 &= \begin{bmatrix} H_n & H_n \\ H_n & -H_n \end{bmatrix}^2 \\ &= \begin{bmatrix} H_n^2 + H_n^2 & H_n^2 - H_n^2 \\ H_n^2 - H_n^2 & H_n^2 + H_n^2 \end{bmatrix} = \begin{bmatrix} 2^{n+1} I_{2^n} & 0 \\ 0 & 2^{n+1} I_{2^n} \end{bmatrix} = 2^{n+1} I_{2^{n+1}}. \end{aligned}$$

Corollary 4 *The eigenvalues of H_n are all $\pm 2^{n/2}$.*

Proof

By the above lemma, for all v, $vHH^T = 2^n v$ and therefore 2^n is the only eigenvalue of HH^T. Thus, the only eigenvalues of H are $\pm 2^{n/2}$.

We denote the discrepancy of f (with respect to the uniform distribution) and a rectangle $A \times B$ by $\mathrm{disc}(f, A \times B)$. All our results can be generalized to arbitrary distributions by multiplying each entry of M_f by the probability of the corresponding cell.

Recall that Boolean functions can be considered as taking values in either $\{0, 1\}$ or $\{+1, -1\}$. In this section, we will use the ± 1 convention when describing the matrices and rectangles.

We use the notation $\mathbf{1}_A$ for the characteristic vector of A, which contains 1 in positions corresponding to the elements of A, and 0's elsewhere.

2.1 The Eigenvalue Method

The eigenvalue method upper bounds the discrepancy using the maximal eigenvalue of M_f.

Lemma 5 (Eigenvalue Bound) *Let f be a symmetric Boolean function, i.e. $f(x, y) = f(y, x)$. Then*

$$\mathrm{disc}(f, A \times B) \leq 2^{-2n} \lambda_{\max} \sqrt{|A| \cdot |B|},$$

where $n = |x| = |y|$ is the input size, and $\lambda_{\max}$ is the largest eigenvalue of the symmetric matrix M_f.

Proof Since M_f is symmetric, its eigenvectors v_i form an orthonormal basis for $\mathbb{R}^n$. Denote by λ_i the eigenvalue corresponding to v_i, so that $M_f v_i = \lambda_i v_i$.

Expand the characteristic vectors of A and B in this basis:

$$\mathbf{1}_A = \sum \alpha_i v_i, \qquad \mathbf{1}_B = \sum \beta_i v_i$$

Putting these expansions into the definition of discrepancy, we are almost done. Since $2^{2n}\text{disc}(f, A\times B)$ is equal to the absolute value of the difference between the number of 1's and the number of 0's in $A \times B$, we have:

$$\begin{aligned} 2^{2n}\text{disc}(f, A \times B) &= \left|\mathbf{1}_A{}^{\mathrm{T}} M_f \mathbf{1}_B\right| \\ &= \left|\left(\sum \alpha_i v_i\right)^{\mathrm{T}} \left(\sum \beta_i \lambda_i v_i\right)\right| \\ &= \left|\sum \alpha_i \beta_i \lambda_i\right| \leq \lambda_{\max} \left|\sum \alpha_i \beta_i\right|. \end{aligned}$$

Note that $\sum \alpha_i^2 = \|\mathbf{1}_A\|^2 = |A|$ by Parseval's identity. (Parseval's identity relates the values of the Fourier coefficients to the values of the function. Namely, it states that for any function $f : \{0,1\}^n \to R$, the sum of the squares of the Fourier coefficients of f is equal to f^2. Note that in our case we have not normalized. If we had normalized – so that the Fourier coefficients were normalized, then the sum of the squares of the Fouerier coefficients of f would be equal to $E[f^2]$.)

and similarly $\sum \beta_i^2 = |B|$. The lemma follows from an application of Cauchy-Schwarz:

$$\begin{aligned} 2^{2n}\text{disc}(f, A \times B) &\leq \lambda_{\max} \left|\sum \alpha_i \beta_i\right| \\ &\leq \lambda_{\max} \sqrt{\sum \alpha_i^2} \sqrt{\sum \beta_i^2} = \lambda_{\max} \sqrt{|A| \cdot |B|}. \end{aligned}$$

We are now ready to prove Lindsey's Lemma which gives a bound on the disrepancy of the inner product function:

Lemma 6 (Lindsey's Lemma) $2^{2n}\text{disc}(\text{IP}_n, A \times B) \leq \sqrt{2^n |A| \cdot |B|}$.

Here $\text{IP}_n(x, y) = \sum x_i y_i \pmod 2$.

Proof The matrix corresponding to IP_n is H_n. We have shown that $\lambda_{\max}(H_n) = 2^{n/2}$, and so the lemma follows by the Eigenvalue Bound.

We are now ready to prove the following theorem.

Theorem 7 $R^{cc}(IP) = \Omega(n)$

By Lindsey's Lemma, discrepancy is maximized when $|A| = |B| = 2^n$, and this gives $disc(IP_n, A\times B) \leq 2^{3n/2}2^{-2n} = 2^{-n/2}$. Thus $R(IP_n) \geq \log(1/3disc(IP_n)) = log(2^{n/2}/3) = \Omega(n)$.

2.2 The BNS Method

In the previous lecture we've outlined the discrepancy method, which is a method for getting lower bounds on randomized communication complexity given upper bounds on the discrepancy of the matrix M_f corresponding to the function in question. We showed how to bound the discrepancy using the largest eigenvalue of M_f. Today we will first give the BNS lemma which is another way of bounding the discrepancy of M_f.

Let F be a function from $X \times Y$ to $\{-1, 1\}$. Let μ be an arbitrary distribution over $X \times Y$ and let $R = A \times B$, $A \subseteq X$, $B \subseteq Y$ be a combinatorial rectangle. Then

$$\mathrm{disc}_\mu(F, R) = |\sum_{(x,y)\in R} \mu(x, y)F(x, y)|,$$

$$\mathrm{disc}_\mu(F) = max_R \mathrm{disc}_\mu(R),$$

$$\mathrm{disc}(F) = min_\mu \mathrm{disc}_\mu(F).$$

The following theorem has been reproven several times, for example in the original BNS paper, and followup papers by Raz 1995, and in Sherstov's paper "Separating AC0 from depth-2 majority circuits."

Theorem 8 *(BNS Bound) Let $F : X \times Y \to \{-1, 1\}$ and let μ be a distribution over $X \times Y$. Then*

$$\mathrm{disc}_\mu(F)^2 \leq |Y| \sum_{x,x'\in X} |\sum_{y\in Y} \mu(x, y)\mu(x', y)F(x, y)F(x', y)|.$$

Proof Define $\alpha_x = 1$ for all $x \in A$, $\beta_y = 1$ for all $y \in B$ and for all other x, y, let α_x, β_y be independent random variables distributed uniformly over $\{-1, 1\}$.

Then

$$\begin{aligned}
\mathrm{disc}_\mu(M) &= |\sum_{(x,y)\in R} \mu(x, y)F(x, y)| && (5)\\
&= |\sum_{(x,y)\in R} \mathbb{E}[\alpha_x\beta_y]\mu(x, y)F(x, y) + \sum_{(x,y)\notin R} \mathbb{E}[\alpha_x\beta_y]\mu(x, y)F(x, y)| && (6)\\
&= |\,\mathbb{E}[\sum_{x,y} \alpha_x\beta_y\mu(x, y)F(x, y)]| && (7)\\
& && (8)
\end{aligned}$$

In particular there is a fixed assignment $\alpha_x, \beta_y \in \{-1, 1\}$ for all x, y such that

$$\mathrm{disc}_\mu(F) \leq |\sum_{x,y} \alpha_x\beta_y\mu(x, y)F(x, y)|.$$

Squaring both sides and applying Cauchy Schwartz gives:

$$\begin{aligned}
\mathrm{disc}_\mu(F)^2 &\leq |Y| \sum_y (\beta_y \sum_x \alpha_x\mu(x, y)F(x, y))^2 && (9)\\
&= |Y| \sum_{x,x'} \alpha_x\alpha_{x'} \sum_y \mu(x, y)\mu(x', y)F(x, y)F(x', y) && (10)\\
&\leq |Y| \sum_{x,x'} |\sum_y \mu(x, y)\mu(x', y)F(x, y)F(x', y)| && (11)\\
& && (12)
\end{aligned}$$

Replacing the sums by expectations we can rewrite the above as:

$$\frac{\mathrm{disc}_\mu(F)^2}{|X|^2 \times |Y|^2} \leq \mathop{\mathbb{E}}_{x,x'} |\mathop{\mathbb{E}}_{y} \mu(x,y)\mu(x',y)F(x,y)F(x',y)|.$$

Alternative Proof. We give an alternative proof of the BNS bound for the uniform distribution.

The definition of discrepancy over the uniform distribution, where $|X| = |Y| = 2^n$ is:

$$\mathrm{disc}(f, A \times B) = \sum_{x \in A, y \in B} F(x,y)/2^{2n}.$$

The discrepancy can be written using expectations as

$$\mathrm{disc}(f, A \times B) = \left| \mathop{\mathbb{E}}_{x,y} \mathbf{1}_A(x)\mathbf{1}_B(y)F(x,y) \right|.$$

We can recast the Cauchy-Schwarz inequality in the form $\mathbb{E}[Z]^2 \leq \mathbb{E}[Z^2]$. Thus we can obtain:

$$\begin{aligned}
\mathrm{disc}(F, A \times B)^2 &= \left(\mathop{\mathbb{E}}_{y} \mathbf{1}_B(y) \mathop{\mathbb{E}}_{x} \mathbf{1}_A(x)F(x,y) \right)^2 \\
&\leq \mathop{\mathbb{E}}_{y} \left(\mathbf{1}_B(y) \mathop{\mathbb{E}}_{x} \mathbf{1}_A(x)F(x,y) \right)^2 \\
&\leq \mathop{\mathbb{E}}_{y} \left(\mathop{\mathbb{E}}_{x} \mathbf{1}_A(x)F(x,y) \right)^2 \\
&= \mathop{\mathbb{E}}_{y} \left(\mathop{\mathbb{E}}_{x,x'} \mathbf{1}_A(x)\mathbf{1}_A(x')F(x,y)F(x',y) \right) \\
&= \mathop{\mathbb{E}}_{x,x'} \mathbf{1}_A(x)\mathbf{1}_A(x') \left(\mathop{\mathbb{E}}_{y} F(x,y)F(x',y) \right) \\
&\leq \mathop{\mathbb{E}}_{x,x'} \left| \mathop{\mathbb{E}}_{y} F(x,y)F(x',y) \right|.
\end{aligned}$$

The bound we get does not depend on the sizes of A and B, and so it is slightly inferior to bounds which do (like Lindsey's lemma). In practice, the difference is usually insignificant (but is the subject of the final question in the first assignment!).

We illustrate the method by proving yet again the upper bound on the discrepancy of the inner product function:

Lemma 9 *We have* $\mathrm{disc}(\mathrm{IP}_n, A \times B) \leq 2^{-n/2}$.

Proof The matrix corresponding to IP_n is H_n. The rows of H_n are orthogonal and so

$$\mathop{\mathbb{E}}_{x} H_n(x,y)H_n(x,z) = \begin{cases} 0 & \text{if } y \neq z, \\ 1 & \text{if } y = z. \end{cases}$$

Using the BNS bound,

$$\mathrm{disc}(\mathrm{IP}_n, A \times B)^2 \leq \mathop{\mathbb{E}}_{y,z} \left| \mathop{\mathbb{E}}_{x} H_n(x,y)H_n(x,z) \right| = \Pr[y = z] = 2^{-n}.$$

3 Degree/Discrepancy Method

The Degree/Discrepancy method, due to Sherstov, is a way to come up with other functions having high randomized communication complexity. The basic idea is to start with some other function (the "base" function) which is difficult under some other complexity measure, and to "lift" it to a function which is difficult in the randomized communication complexity model. Sherstov's main contribution is using polynomial complexity measures to quantify the difficulty of the base function.

3.1 Polynomial Complexity Measures

We will consider several different complexity measures for the base function. All of them try to capture the notion of being hard to approximate by a polynomial over the real numbers.

Consider a Boolean function $f(x_1, \ldots, x_q)$. We will assume that the inputs and outputs are the usual 0/1 (rather than ± 1). This function can be represented as a real polynomial by following the following steps:

1. Present f as a logical formula, e.g. conjunctive normal form.
2. Convert the formula to a polynomial using the following rules:

$$\begin{aligned} \neg(x) &= 1 - x, \\ x \wedge y &= xy, \\ x \vee y &= x + y - xy. \end{aligned}$$

3. Use the identity $x^2 = x$ to reduce any repeated variables in the monomials.

The result is some polynomial whose degree is at most q, if f is a q-CNF formula.

This prompts the following definition:

Definition The *degree* (also *polynomial degree*) of a function f, written $\deg(f)$, is the minimal degree of a real polynomial P such that $f(x_1, \ldots, x_q) = P(x_1, \ldots, x_q)$ on all Boolean inputs.

In general, it is difficult to represent functions exactly by polynomials, and so the fact that a function has high polynomial degree isn't strong enough for our purposes. A rather lenient alternative is the following:

Definition The *sign degree* (sometimes *polynomial threshold degree*) of a function f, written $\text{sign-deg}(f)$, is the minimal degree of a real polynomial P such that for all Boolean inputs $x_1, \ldots, x_q$:

- If $f(x_1, \ldots, x_q) = 1$ then $P(x_1, \ldots, x_q) > 0$.
- If $f(x_1, \ldots, x_q) = 0$ then $P(x_1, \ldots, x_q) < 0$.

This definition is so permissive that it is hard to prove lower bounds on the sign degree. Here are two examples of functions for which a lower bound is known:

- The parity function on q inputs has the maximal sign degree q.
- The Minsky-Papert "tribes" function $\bigvee_{i=1}^{m} \bigwedge_{j=1}^{4m^2} x_{ij}$ has sign degree $m = \sqrt[3]{q/4}$.

Lower bounding the sign degree can be difficult simply because a function with high polynomial degree can be sign-represented by a very low degree polynomial. An extreme example is the OR function (the logical inclusive or of all inputs). This function is sign-represented by the linear polynomial $\sum x_i - \frac{1}{2}$, but an exact representation necessitates a degree q polynomial. This prompts the need for some sort of an interpolation between these two extreme definitions.

The following definition generalizes both previous ones:

Definition [ϵ-Approximation Degree] Given a real $0 \leq \epsilon \leq \frac{1}{2}$, the *$\epsilon$-degree* (more officially, *ϵ-approximation degree*) of a function f, written ϵ-deg(f), is the minimal degree of a real polynomial P such that for all Boolean inputs,

$$|f(x_1, \ldots, x_q) - P(x_1, \ldots, x_q)| \leq \epsilon.$$

If $\epsilon = 0$ this reduces to the regular degree, while if $\epsilon = \frac{1}{2}$ then this (almost) reduces to the sign degree. Clearly the ϵ-degree is monotone decreasing in ϵ, and so for general $0 < \epsilon < \frac{1}{2}$ we have

$$0 \leq \text{sign-deg}(f) \leq \epsilon\text{-deg}(f) \leq \deg(f) \leq q.$$

As an example, the OR function, whose sign-degree is 1 and whose polynomial degree is q, has ϵ-degree $O(\sqrt{q})$ for $\epsilon = 1/8$.

Nisan and Szegedy related the ϵ-degree to decision tree complexity, defined as follows:

Definition A *decision tree* for a Boolean function is a binary tree whose inner vertices are labelled by input variables, and whose leaves are labelled by 0/1. The computation outlined by the tree proceeds from the root by querying the labelled variable, taking the left branch if the respective variable is 0, the right branch if it is 1. Upon reaching a leaf, its label is output.

The *decision tree complexity* of a function f, written DTC(f), is the depth of the shallowest decision tree which represents it.

Using the method outlined above for converting a formula into a real polynomial, one sees that the decision tree complexity upper bounds the polynomial degree. In particular, ϵ-deg$(f) \leq$ DTC(f). Nisan and Szegedy proved a matching upper bound:

$$\epsilon\text{-deg}(f) \leq \text{DTC}(f) \leq \epsilon\text{-deg}(f)^8.$$

Formulated differently, we have $\log \epsilon$-deg$(f) = \Theta(\log \text{DTC}(f))$.

4 Discrepancy and Duality of Sign Degree

Theorem 10 (Duality of sign degree) *Let $f : \{-1, 1\}^n$ $d \geq 0$*

Then sign-deg(f) is at least d if and only if there exists a distribution μ over $\{-1, 1\}^n$ such that

$\mathbb{E}_{x \sim \mu}[f(x) \cdot \chi_S(x)] = 0$ $\forall S$, $|S| < d$

That is to say, "f is orthogonal to χ_S for small s", where χ_S is the parity function over the indices in S

Theorem 11 (Duality of approximation degree) *(Sherstov, Shi-Zhu)*
Fix $\varepsilon \geq 0$. Let $f : \{-1,1\}^n \to \{-1,1\}$, $deg_\varepsilon(f) = d \geq 1$.
Then $\exists g : \{-1,1\}^n \to \{-1,1\}$ and a distribution μ over $\{-1,1\}^n$ such that:

$$(1) \quad \mathop{\mathbb{E}}_{x\sim\mu}[g(x)\chi_S(x)] = 0 \quad \forall S \quad |S| \leq d$$

$$(2) \quad corr_\mu(f,g) > \varepsilon \qquad (corr_\mu(f,g) = \mathop{\mathbb{E}}_{x\sim\mu}[f(x)g(x)])$$

Proof (Duality of sign degree) This is an instance of the "Gordon Transposition Lemma"
Let A be a matrix of dimension $m \times n$. Then $\exists \vec{u}$ s.t. $\vec{u}^T A > 0$ iff $\exists \vec{v} > 0$ s.t. $A\vec{v} = 0$

We want a polynomial f' which sign-approximates f. We look for coefficients α_s, $|S| < d$ to produce $f' = \sum_S \alpha_s \chi_s$

Fix ρ. If $f(\rho) = 1$ $\sum_S \alpha_s\chi_s > 0$, and if $f(\rho) = -1$ $\sum_S \alpha_s\chi_s < 0$. So, $\sum \alpha_s\chi_s f(\rho) > 0$, that is to say, they match in sign.

We construct a matrix with columns representing values for ρ and rows representing values for s, that is, subsets of $1..n$ of size $\leq d$. For each value we fill in $\chi_s(\rho)f(\rho)$. Then the rows of our matrix are the values for α_s, which is $\vec{u}^T$ in the above lemma, and $\vec{v}$ is a distribution over our columns.

Using duality of sign degree we can prove 2-party communication complexity lower bounds. The outline of the argument is as follows.

(1) We start with a base function $f : \{-1,1\}^n$ with large sign degree d. For example, $f(z) = \bigvee_{i=1}^m \bigwedge_{j=1}^{4m^2} z_{ij}$ has sign-degree m, or the parity function, with sign degree n.

(2) Use the pattern matrix method to "lift" f to obtain a 2-player communication complexity problem $F = f \circ g^n$, where $g : [b] \times \{0,1\}^b \to \{0,1\}$ is the index gadget defined as $g(a,b) = b|_a$. That is, the value of g on input a, b is the bit in b that is pointed to by a.

That is, Alice has input $x_1, \ldots, x_n$ where each x_i consists of a string of $\log b$ bits (that we will view as a pointer to one of the b coordinates); Bob has input $y_1, \ldots, y_n$ where each y_i is a string of b bits, and the value of $F = f \circ g^n$ on (x, y) is obtained by computing the function f on the n bits pointed to by Alice (one per block).

(3) By duality of sign degree, there exists a distribution μ over $\{-1,1\}^n$ such that f is orthogonal to all χ_S, $|S| < d$, with respect to μ. Extend μ to a distribution λ over the domain of F in the natural way. Then by orthogonality, the BNS Lemma will imply small discrepancy (discrepancy less than 2^{-d}) for F with respect to λ.

Using the above plan, we will prove the following theorem:

Theorem 12 *(Sherstov) Let f be boolean over $z_1..z_n$ with sign degree $\geq d$.*
Then $\mathrm{disc}(F)(2en/bd)^d$.

Proof

Extending μ to λ: λ is a distribution on $X \times Y$ induced by μ. To obtain λ we pick $x \in X$ uniformly at random. We choose $y|_x$ according to μ. Then we set the rest of the bits of y uniformly at random. So we have:

$$\lambda(x,y) = b^{-n}\mu(y|_x)2^{-(b-1)n}.$$

By the BNS lemma,

$$\frac{\mathrm{disc}_\lambda(F)^2}{|X|^2 \times |Y|^2} \leq \mathop{\mathbb{E}}_{x,x'} |\mathop{\mathbb{E}}_{y}[f(y|_x)f(y|_{x'})\lambda(x,y)\lambda(x',y)]|$$

Rewriting in terms of μ, since $|X| = b^n$ and $|Y| = 2^{bn}$ we get

$$\mathrm{disc}_\lambda(F)^2 \leq 4^n \mathop{\mathbb{E}}_{x,x'} |\mathop{\mathbb{E}}_{y}[f(y|_x)f(y|_{x'})\mu(y|_x)\mu(y|_{x'})]|.$$

Let $\Gamma(x,x')$ denote $\mathbb{E}_y[f(y|_x)f(y|_{x'})\mu(y|_x)\mu(y|_{x'})]$.

Claim 1 When $|x \cap x'| \leq d-1$ then $\Gamma(x,x') = 0$.

Claim 2 When $|x \cap x'| = i$, $|\Gamma(x,x')| \leq 2^{i-2n}$.

By these claims,

$$\mathrm{disc}_\lambda(F)^2 \leq \sum_{k=d}^{n} 2^k Pr\left[|x \cap x'| = k\right],$$

$$Pr\left[|x \cap x'| = k\right] = \binom{n}{k}(1/b)^k(1-1/b)^{n-k} \leq (en/k)^k(1/b)^k(1-1/b)^k$$

(The above inequality uses $\binom{n}{k} \leq (en/k)^k$.)

Therefore,

$$\begin{aligned}
\mathrm{disc}_\lambda(F)^2 &\leq \sum_{k=d}^{n} 2^k(en/k)^k(1/b)^k(1-1/b)^{n-k} && (13)\\
&= \sum_{k=d}^{n}(2en/bk)^k(1-1/b)^{n-k} && (14)\\
&\leq (2en/bd)^d && (15)\\
& && (16)
\end{aligned}$$

For b sufficiciently large, this is at most 2^{-d}.

Proof of Claim 1 The basic idea here will be that by orthogonality, the expectation is zero.

Proof of Claim 2 This claim follows because μ is a probability distribution. We want to show that if $|x \cap x'| = i$, then $|\Gamma(x,x')| \leq 2^{i-2n}$. For notational convenience we will assume that x and

x' have the same pointers in the first i blocks and different pointers in the remaining blocks. (That is, $x_j = x'_j$ for all $j \leq i$ and $x_j \neq x'_j$ for all $j > i$.) Then we have:

$$|\Gamma(x, x')| \leq \mathop{\mathbb{E}}_{y}[|f(y|_x)\mu(y|_x)f(y|_{x'})\mu(y|_{x'})|]$$

$$|\Gamma(x, x')| \leq \mathop{\mathbb{E}}_{y}[\mu(y|_x)\mu(y_{x'}]$$

$$|\Gamma(x, x')| \leq \mathop{\mathbb{E}}_{y}\mu(y|_x) \cdot max_{\alpha \text{ to } y|_{x\cap x'}} \mathop{\mathbb{E}}_{x'_{i+1},\ldots,x'_n}[\mu(\alpha, x'_{i+1}, \ldots, x'_n]$$

The first expectation above is at most 2^{-n} because μ is a probability distribution, and similarly the second expectation in the last equation is at most $2^{-(n-i)}$ again because μ is a probability distribution.

5 Application to Circuits

In 1989, Allender proved the following theorem, showing that any AC^0 function can be computed by quasipolynomial-size depth-3 majority circuits.

Theorem 13 (Allender) *Any AC^0 function can be computed by a depth-3 majority circuit of quasipolynomial ($O(n^{polylog(n)})$) size.*

An open question was whether or not his result could be improved. In particular, is it possible to improve the depth, showing that every function in AC^0 be computed by depth-2 majority-of-threshold circuits of quasipolynomial size? A corollary to Sherstov's theorem is a negative resolution to this open problem:

Theorem 14 (Sherstov) *$\exists F \in AC^0_3$ (depth 3) whose computation requires majority of exponentially many threshold gates.*

It suffices to show an AC^0 function with exponentially small discrepancy. We start with the AC^0_2 function:

$$f = \bigvee_{i=1}^{m} \bigwedge_{j=1}^{4m^2} e_{ij}$$

We construct F(x,y) where $F(x, y) = f(x|_y)$, that is, f of the bits of x specified by y. F(x,y) is in AC^0_3:

$$F(x, y) = \bigvee_{i=1}^{m} \bigwedge_{j=1}^{4m^2} \bigvee_{\alpha} \left(y_{ij\alpha_1} \wedge y_{ij\alpha_2} \wedge \ldots \wedge y_{ij\alpha_q} \wedge x_{ij\alpha}\right)$$

because we can swap the order of the $\wedge$'s within the brackets with the last $\bigvee$ and then merge them with the middle $\bigwedge$.

By the degree/discrepancy theorem we know that because f requires a high degree polynomial to compute, F(x,y) has low discrepancy. Each threshold gate can be computed by a $O(\log n)$ bit

probabilistic CC protocol with $R_\epsilon^{pub}(f) = O(\log n + \log \frac{1}{\epsilon})$.

Suppose F has (low) discrepancy e^{-N^ε}. Then any randomized protocol requires N^ε bits. Also let $F = MAJ(h_1..h_S)$ where each h_i is a threshold circuit.

The players pick a random $i \in [S]$. They evaluate h_i, using $O(\log n)$ bits and output the result.

The probability of correctness of the threshold-computing protocol is $1 - \frac{1}{4S}$ if we set $\varepsilon' \sim \frac{1}{S}$.

The total cost is $O(\log n) + \log S$ bits. The probability of correctness is $(\frac{1}{2} + \frac{1}{2S}) - \frac{1}{4S} = \frac{1}{2} + \frac{1}{4S}$ on every input.

Since we know that F requires $O(N^\varepsilon)$ bits to compute, S must be exponentially large! And so there is no polynomially-sized majority-of-threshold circuit to compute $F \in AC_3^0$.

6 Extensions of Sherstov

6.1 High approximation degree to high probabilistic communication complexity

First, the above theorem can be generalized to prove lower bounds on 2-party communication complexity of lifted functions where the base function has high ϵ-approximate degree, rather than high sign degree. The idea here is to replace the duality theorem for sign degree by the duality theorm for approximate degree.

We follow the same three steps, showing that if f (the base function) has large approximate degree, then there exists a function g that is highly correlated with f, and a distribution μ such that g is orthogonal to all low degree characters with respect to μ. We then lift g to a two-party communication complexity problem G, and lift μ to a distribution λ over G to show (using orthogonality and BNS) that G has low discrepancy. Finally, since f is highly correlated with g, F is highly correlated with G, and thus it follows that F also has high randomized communication complexity.

6.2 NOF lower bounds

The above ideas can also be extended to prove lower bounds in the NOF model as well. The BNS lemma stated above can be generalized straightforwardly to prove a similar lemma in the NOF case. Its generalization for $k = 3$ looks like this:

$$disc(F)^{2^2} \leq E_{y_1,y_1' \in Y_1} E_{y_2,y_2' \in Y_2} |E_{x \in X} f(x,y_1,y_2) f(x,y_1,y_2') f(x,y_1',y_2) f(x,y_1',y_2')|.$$

More generally for arbitrary k we will have a similar expression, but where the LHS is raised to the power 2^{k-1}. Using this stronger BNS lemma, one can prove a similar general theorem following the basic outline that we presented.

Note that for $k = \log n$ players, the bound becomes trivial. It is a longstanding open problem to prove a NOF communication complexity bound for an explicit function (say in NP) for more than $\log n$ many players.