Communication Complexity 14 Nov, 2011 (@ IMSc)

24. Degree/Discrepancy Method

Lecturer: Prahladh Harsha *Scribe: Sajin Koroth*

In today's lecture, we will discuss duality-based methods to prove lower bounds on communication complexity of functions. In particular, we will discuss the degree/discrepancy method due to Sherstov [She09]. The main reference for today's lecture is the survey on application of dual polynomials in communication complexity [She08].

24.1 Duality based lower bounds

To motivate the use of duality in proving lower bounds, let us consider a typical communication complexity problem. For a function f, the communication complexity is cost of the best protocol π solving f. The cost and definition of a protocol solving f is defined according to the type of communication complexity (eg., deterministic, non-deterministic, randomized etc. Hence, it can be casted as a minimization problem,

$$\mathrm{R}(f) = \min_{\pi \text{ solves } f} \mathrm{Cost}(\pi).$$

To prove lower bound on $\mathrm{R}(f)$, one has to lower bound all protocols π "solving" f. But if one could instead write the "dual" of the above minimization program, i.e., a maximization problem over the corresponding dual objects (which we denote with L) with the associated dual cost function (which we denote with ρ), i.e.

$$\mathrm{R}(f) = \max_{l \leftarrow L} \rho(l).$$

Rewriting in the dual form has the following advantage: to show that $\mathrm{R}(f)$ is large it is sufficient to exhibit a specific $l \in L$ say l^* such that $\rho(l^*)$ is large. Such an l^* is called the *dual witness*. It might not be always possible to cast the problem in the primal-dual framework, so one might have to relax the original problem to a convex optimization problem whose dual can then be written. We will see several instances of such relaxations and the corresponding dual formulations in the next few lectures.

Recall that we already used duality (in the form of Yao's min-max theorem) in proving lower bounds for randomized communication complexity.

$$\mathrm{R}_\varepsilon(f) = \max_{\mu} \mathrm{D}^\mu_\varepsilon(f).$$

Here, $\mathrm{R}_\varepsilon(f)$ is the randomized communication complexity for f with error at most ε and $\mathrm{D}^\mu_\varepsilon(f)$ is the distributional communication complexity of f under distribution μ and error ε. The dual witness here is a hard distribution for f, i.e, a distribution μ such that $\mathrm{D}^\mu_\varepsilon(f)$ is large.

How does one come up with a "hard" distribution or a "good" dual witness in general? In today's lecture, we will see one such technique due to Sherstov, that produces hard distributions based on functions with large *threshold degree*.

24.2 Discrepancy bound

We will begin by recalling the discrepancy of a function.

$$\begin{aligned}\mathsf{disc}_\mu(f) &= \max_R \mathsf{disc}_\mu(f;R) = \max_R \left|\Pr_{(x,y)\leftarrow\mu}[f(x,y)=1] - \Pr_{(x,y)\leftarrow\mu}[f(x,y)=0]\right| \\ &= \max_R \left|\sum_{x\in X, y\in Y} (-1)^{f(x,y)}\mu(x,y)\right|,\end{aligned}$$

where the maximum is taken over all rectangles $R = S \times T$. Discrepancy is used to lower bound distributional complexity.

$$\mathrm{D}^\mu_{\frac{1}{2}-\varepsilon}(f) \geq \left(\log\frac{1}{\mathsf{disc}_\mu(f)} - \log\frac{1}{2\varepsilon}\right).$$

This, combined with Yao's minmax lemma yields the following

$$\mathrm{R}_{\frac{1}{2}-\varepsilon}(f) = \max_\mu \ \mathrm{D}^\mu_{\frac{1}{2}-\varepsilon}(f) \geq \max_\mu \ \left(\log\frac{1}{\mathsf{disc}_\mu(f)} - \log\frac{1}{2\varepsilon}\right).$$

Thus, to lower bound $\mathrm{R}(f)$ it suffices to exhibit a distribution (dual witness) μ such that $\mathsf{disc}_\mu(f)$ is very small. To show $\mathsf{disc}_\mu(f)$ is small, one needs to show that for all rectangles R, $\mathsf{disc}_\mu(f;R)$ is small. It will be more convenient to work with a uniform (rectangle independent) bound for discrepancy. The following bound achieves this.

First for some notational convenience. It will be nicer to work with $\{\pm\}$-valued functions instead of $\{0,1\}$-valued functions. For any $f : X \times Y \to \{\pm\}$ and a distribution μ on the input space $X \times Y$, for the purposes of notational brevity define $\psi(x,y) \triangleq \psi^f_\mu(x,y) \triangleq f(x,y)\mu(x,y)$. In this notation, discrepancy of a function is given by

$$\mathsf{disc}_\mu(f) = \max_R \left|\sum_{x\in X, y\in Y} \psi(x,y)\right|.$$

Lemma 24.1. *For any function $f : X \times Y \to \{\pm 1\}$*

$$\left(\frac{\mathsf{disc}_\mu(f)}{|X|\cdot|Y|}\right)^2 \leq \mathop{\mathbb{E}}_{x\in X}\left|\mathop{\mathbb{E}}_{y,y'\in Y}\left[\psi^f_\mu(x,y)\psi^f_\mu(x,y')\right]\right|.$$

Proof. Let $R = S \times T$ be the rectangle such that $\mathsf{disc}_\mu(f) = \mathsf{disc}_\mu(f;R)$. Define the two random variables $\alpha : X \to \{\pm 1\}$ and $\beta : Y \to \{\pm 1\}$ as follows:

$$\alpha_x \triangleq \begin{cases} 1 & \text{if } x \in S, \\ \pm 1 & \text{with equal probability if } x \notin S \end{cases}, \quad \beta_y \triangleq \begin{cases} 1 & \text{if } y \in T, \\ \pm 1 & \text{with equal probability if } y \notin T. \end{cases}.$$

We thus have that for $(x,y) \in S \times T$, $\alpha_x\beta_y = 1$, while for $(x,y) \notin S \times T$, $\mathbb{E}_{\alpha,\beta}[\alpha_x\beta_y] = 0$. Hence,

$$\mathsf{disc}_\mu(f) = \mathsf{disc}_\mu(f;R) = \left|\mathop{\mathbb{E}}_{\alpha,\beta}\left[\sum_{x\in X, y\in Y} \alpha_x\beta_y \cdot \psi(x,y)\right]\right|.$$

Hence, there exist functions α, β such that

$$\mathsf{disc}_\mu(f) \quad \leq \quad \left| \sum_{x \in X, y \in Y} \alpha_x \beta_y \cdot \psi(x,y) \right| .$$

Or equivalently,

$$\frac{\mathsf{disc}_\mu(f)}{|X| \cdot |Y|} \quad \leq \quad \left| \mathop{\mathbb{E}}_{(x,y) \in X \times Y} \left[\alpha_x \beta_y \psi(x,y) \right] \right| .$$

Squaring both sides and applying Jensen's inequality we have,

$$\begin{aligned}
\left(\frac{\mathsf{disc}_\mu(f)}{|X| \cdot |Y|} \right)^2 &\leq \left(\mathop{\mathbb{E}}_{x,y} \left[\alpha_x \beta_y \cdot \psi(x,y) \right] \right)^2 \\
&= \left(\mathop{\mathbb{E}}_{x} \left[\alpha_x \cdot \mathop{\mathbb{E}}_{y} \left[\beta_y \cdot \psi(x,y) \right] \right] \right)^2 \\
&\leq \mathop{\mathbb{E}}_{x} \left[\left(\mathop{\mathbb{E}}_{y} \left[\beta_y \cdot \psi(x,y) \right] \right)^2 \right] && [\text{Since } (\mathbb{E}\, X)^2 \leq \mathbb{E}\, X^2 \text{ and } \alpha_x^2 = 1] \\
&= \mathop{\mathbb{E}}_{x} \left[\mathop{\mathbb{E}}_{y,y'} \left[\beta_y \beta_{y'} \cdot \psi(x,y) \psi(x,y') \right] \right] \\
&= \mathop{\mathbb{E}}_{y,y'} \left[\beta_y \beta_{y'} \cdot \mathop{\mathbb{E}}_{x} \left[\psi(x,y) \psi(x,y') \right] \right] \\
&\leq \mathop{\mathbb{E}}_{x} \left[\left| \mathop{\mathbb{E}}_{y,y'} \left[\psi(x,y) \psi(x,y') \right] \right| \right] . && [\text{Since } |\beta_y \beta_{y'}| = 1]
\end{aligned}$$

Thus, proved. □

In later lectures, we will see a generalization of the above bound to higher dimensions.

24.3 Large threshold degree as dual witness

Definition 24.2 (threshold degree)**.** *A function $f : \{\pm 1\}^n \to \{\pm 1\}$ is said to be* sign represented *by a polynomial $p : \{\pm 1\}^n \to \mathbb{R}$ if and only if for all $x \in \{\pm 1\}^n$, $p(x) f(x) > 0$, i.e. the sign of $p(x)$ and $f(x)$ are same for all $x \in \{\pm 1\}^n$.*

For a function $f : \{\pm 1\}^n \to \{\pm 1\}$, the threshold degree, denoted by $\deg_\pm(f)$*, is defined as follows*

$$\deg_\pm(f) \triangleq \min \left\{ d : f \textit{ can be sign represented by a degree } d \textit{ polynomial } \right\} .$$

Note that a trivial upper bound on the threshold degree of any Boolean function $f : \{\pm 1\}^n \to \{\pm 1\}$ is n as any function can be sign represented (in fact exactly represented) by a polynomial of degree n via interpolation.

Which functions have large threshold degree? The parity functions have large threshold degree. More precisely, the functions $\chi_S(x) = \prod_{i \in S} x_i$ for subsets $S \subseteq [n]$ have threshold degree exactly $|S|$. We will now give a dual characterization of functions having large threshold degree.

24.3.1 Dual characterization of threshold degree

We will give a dual characterization using the following theorem about duality.

Theorem 24.3 (Gordon's transportation theorem)**.** *For any matrix $A \in \mathbb{R}^{n\times m}$, **exactly one of** the following holds*

1. *There exists a u, such that $u^T A > 0$*

2. *There exists a $v \geq 0$, v is a non-zero vector, such that $Av = \overline{0}$.*

In the above $v > 0$ means that each coordinate of v is > 0.

Proof. Consider the columns of the matrix A. By Farkas' lemma, either the all zeros vector $\overline{0}$ lies in the convex hull of the column vectors or there is a separating hyperplane separating the convex hull of the column vectors from $\overline{0}$ (but not both). In the former case, we have a non-zero vector $v \geq 0$ such that $Av = \overline{0}$. In the latter case, we have a vector u such that $u^T A > 0$. Thus, proved. □

From this theorem we have the following nice corollary,

Corollary 24.4 (dual characterization of threshold degree)**.** *For a function $f : \{\pm 1\}^n \to \{\pm 1\}$, **exactly one of** the following is true*

1. $\deg_{\pm}(f) \leq d$.

2. *There exists a distribution μ on $\{\pm 1\}^n$ such that $\underset{x\leftarrow\mu}{\mathbb{E}}[f(x)\chi_S(x)] = 0$ for all $S, |S| \leq d$.*

The corollary tells that either f has low threshold degree or there exists a distribution μ such that all the low-order monomials (or parity functions) have zero correlation with the function f under μ. Equivalently, f has zero correlation with functions of d or fewer variables.

Proof. Consider the following matrix,

$$A = [f(x)\chi_S(x)]_{(S,x)}\,.$$

The rows of A are indexed by sets $S \subseteq \{\pm 1\}^n, |S| \leq d$ and columns are indexed by $x \in \{\pm 1\}^n$. Gordon's transportation theorem tells us that either there is a vector $(u_S, |S| \leq d)$ such that $u^T A > 0$ or there exists a non-zero non-negative vector $(v_x, x \in \{\pm 1\}^n)$ such that $Av = \overline{0}$ (but not both). In the former case, set $p(x) = \sum u_S\chi_S(x)$. The fact that $u^T A > 0$ is equivalent to the fact that $p(x)f(x) > 0$ for all $x \in \{\pm 1\}^n$. Hence, f is sign represented by a polynomial of degree at most d, i.e., $\deg_{\pm}(f) \leq d$. In the latter case (i.e., $Av = \overline{0}$), consider the distribution μ on $\{\pm 1\}^n$ defined as the normalized v (i.e., $\mu(x) \triangleq v(x)/\sum_x v(x)$). The fact that $Av = \overline{0}$ is equivalent to the fact that $\mathbb{E}_{x\leftarrow\mu}[f(x)\chi_S(x)] = 0$ for all S such that $|S| \leq d$. Thus, proved. □

24.3.2 Low discrepancy function from large threshold degree function

In this section, we will show how to transform a function with large threshold degree into a function with low discrepancy (and hence large randomized communication complexity).

Theorem 24.5 (degree/discrepancy theorem [She09]). *Let $f : \{\pm 1\}^n \to \{\pm 1\}$ be such that $\deg_{\pm}(f) \geq d$. Then for every $N \geq n$, the function $F : \{\pm 1\}^N \times \binom{[N]}{n} \to \{\pm 1\}$ defined as $F(x, V) \triangleq f(x|_V)$ satisfies*

$$\mathsf{disc}_\lambda(F) \leq \left(\frac{4en^2}{Nd}\right)^{d/2},$$

for some distribution λ on the $\{\pm 1\}^N \times \binom{[N]}{n}$.

In the above $y|_V$ represents the restriction of y to the coordinates in the set V. Formally, if $V = \{v_1 < v_2 < \cdots < v_n\}$, then $y|_V = y_{v_1} y_{v_2} \ldots y_{v_n}$.

Proof. Since $\deg_{\pm}(f) \geq d$ by Corollary 24.4, we are guaranteed a distribution μ such that for all functions g on fewer than d variables $\mathbb{E}_\mu [f(x)g(x)] = 0$. The hard distribution λ is obtained from μ as follows.

$$\lambda(y, V) \triangleq \frac{\mu(y|_V)}{\binom{N}{n} \cdot 2^{N-n}}.$$

An informal description of λ is as follows: Choose V uniformly at random from $\binom{[N]}{n}$. y is chosen according to the following distribution: $y|_V$ is chosen according to the distribution μ while the remaining $N - n$ coordinated of y are chosen independently from the uniform distribution on $\{\pm 1\}$.

Using the discrepancy bound lemma 24.1, we get,

$$\begin{aligned} \left(\frac{\mathsf{disc}_\lambda(F)}{2^N \cdot \binom{N}{n}}\right)^2 &\leq \mathop{\mathbb{E}}_{V,W} \left[\left|\mathop{\mathbb{E}}_{y} \left[\psi_\lambda^F(y, V)\psi_\lambda^F(y, W)\right]\right|\right], \\ \text{i.e., } (\mathsf{disc}_\lambda(F))^2 &\leq 4^N \cdot \binom{N}{n}^2 \cdot \mathop{\mathbb{E}}_{V,W} \left[\left|\mathop{\mathbb{E}}_{y} \left[\psi_\lambda^F(y, V)\psi_\lambda^F(y, W)\right]\right|\right], \end{aligned}$$

where $\psi_\lambda^F(y, V) = F(y, V)\lambda(y, V)$ which upon substitution for F and λ yields $f(y|_V)\frac{\mu(y|_V)}{\binom{N}{n} \cdot 2^{N-n}}$. We can thus, rewrite the final inequality as follows.

$$(\mathsf{disc}_\lambda(F))^2 \leq 4^n \cdot \mathop{\mathbb{E}}_{V,W} \left[\left|\mathop{\mathbb{E}}_{y} \left[f(y|_V)f(y|_W)\mu(y|_V)\mu(y|_W)\right]\right|\right].$$

Define $\Gamma(V, W) \triangleq \mathbb{E}_y \left[\psi_\mu^f(y|_V)\psi_\mu^f(y|_W)\right]$, we then have

$$(\mathsf{disc}_\lambda(F))^2 \leq 4^n \cdot \mathop{\mathbb{E}}_{V,W} \left[|\Gamma(V, W)|\right]. \tag{24.3.1}$$

We now analyze $|\Gamma(V, W)|$. We will show that $\Gamma(V, W)$ is 0 if V and W have small intersection and is not too large when V and W have large intersection. We will then complete the proof by choosing N sufficiently large compared to n, such that with high probability V and W, two random n-sized subsets of $[N]$ will have very small intersection. This would in turn imply that $\mathbb{E}_{V,W}[|\Gamma(V, W)|]$ (and hence discrepancy of F) is small.

Claim 24.6 (small intersection)**.** *If $|V \cap W| < d$ then $\Gamma(V, W) = 0$*

Proof. Wlog, assume V to be the subset $\{1, \ldots, n\}$. Then,

$$\begin{aligned}
\Gamma(V,W) = &\quad \mathbb{E}_{y_1,\ldots,y_N}\left[\mu(y_1,\ldots,y_n)f(y_1,\ldots,y_n)\cdot\psi_\mu^f(y|_W)\right] \\
= &\mathop{\mathbb{E}}_{y_1\cdots y_n}\left[\mu(y_1,\ldots,y_n)f(y_1,\ldots,y_n)\cdot \mathop{\mathbb{E}}_{y_{n+1}\cdots y_N}\left[\psi_\mu^f(v|_W)\right]\right].
\end{aligned}$$

Since $|V \cap W| < d$, $g(y_1,\ldots,y_n) \triangleq \mathop{\mathbb{E}}_{y_{n+1}\cdots y_N}\left[\psi_\mu^f(v|_W)\right]$ is a function of less than d variables in $y_1,\ldots,y_n$. Hence by Corollary 24.4

$$\begin{aligned}
\Gamma(V,W) &= \mathop{\mathbb{E}}_{y_1\cdots y_n}\left[\mu(y_1,\ldots,y_n)f(y_1,\ldots,y_n)\cdot g(y_1,\ldots,y_n)\right] \\
&= \frac{1}{2^n}\cdot \mathop{\mathbb{E}}_{(y_1,\ldots,y_n)\leftarrow\mu}\left[f(y_1,\ldots,y_n)\cdot g(y_1,\ldots,y_n)\right] = 0.
\end{aligned}$$

□

Claim 24.7 (large intersection)**.** *If $|V \cap W| = k$ and $k \geq d$, then $|\Gamma(V,W)| \leq 2^{k-2n}$.*

Proof. Wlog, let $V = \{1,\ldots,n\}$ and let $W = \{1,\ldots,k\} \cup \{n+1,\ldots,2n-k\}$.

$$\begin{aligned}
\Gamma(V,W) &\leq \mathop{\mathbb{E}}_{y}\left[\mu(y_1,\ldots,y_n)\cdot\mu(y_1,\ldots,y_k,y_{n+1},\ldots,y_{2n-k})\right] \\
&= 2^{k-2n}\cdot\sum_{y_1,\ldots,y_{2n-k}}\mu(y_1,\ldots,y_n)\cdot\mu(y_1,\ldots,y_k,y_{n+1},\ldots,y_{2n-k}) \\
&\leq 2^{k-2n}.
\end{aligned}$$

The last step follows since $\eta(z_1,\ldots,z_{2n}) = \mu(z_1,\ldots,z_n)\cdot\mu(z_{n+1},\ldots,z_{2n})$ is a probability distribution over $\{\pm1\}^{2n}$ and the summation is over 2^{2n-k} distinct elements in $\{\pm1\}^{2n}$. □

We now use the above claims to bound the discrepancy of F.

$$\begin{aligned}
(\mathsf{disc}_\lambda(F))^2 &\leq 4^n\cdot\mathop{\mathbb{E}}_{V,W}\left[|\Gamma(V,W)|\right] && \text{[From (24.3.1)]} \\
&\leq 4^n\cdot\sum_{k=d}^{n}\Pr\left[|V\cap W| = k\right]\cdot 2^{k-2n} \\
&= \sum_{k=d}^{n}\frac{\binom{n}{k}\cdot\binom{N-n}{n-k}}{\binom{N}{n}}\cdot 2^k \\
&\leq \sum_{k=d}^{n}\binom{n}{k}\left(\frac{n}{N}\right)^k 2^k && \text{[By Sterling's approximation]} \\
&\leq \sum_{k=d}^{n}\left(\frac{ne}{k}\right)^k\left(\frac{2n}{N}\right)^k \\
&= \sum_{k=d}^{n}\left(\frac{2n^2e}{Nk}\right)^k \leq \sum_{k=d}^{n}\left(\frac{2n^2e}{Nd}\right)^k.
\end{aligned}$$

If $(2n^2e/Nd) < 1$, then the above sum is a geometric series and can be bounded by $(4n^2e/Nd)^d$. This bound is true even otherwise since discrepancy is at most 1. We thus have,

$$(\mathsf{disc}_\lambda(F))^2 \leq \left(\frac{4n^2e}{Nd}\right)^d.$$

Thus, proved.

□

Suppose $f : \{\pm 1\}^n \to \{\pm 1\}$ is a function on n inputs with threshold degree d, then we can generate a function F with exponentially small discrepancy (in terms of d) by setting $N = \lceil 16en^2/d \rceil$.

$$\mathsf{disc}_\lambda(F) \leq \left(\frac{4en^2}{\left\lceil N = \frac{16en^2}{d}\right\rceil d}\right)^{d/2} = 2^{-d}.$$

24.4 Connection to AC^0 and majority circuits

An easy corollary of the above theorem is that the function F corresponding to parity function $\chi_{[n]}$ has large communication complexity since $\deg_\pm(\chi_S) = |S|$. Recall that parity is not in AC^0. In this section, we will show that there are considerably simpler functions which also exhibit large threshold degree.

An interesting depth 2 AC^0 function, arising from learning theory, is the Minksy Papert function MP_m defined on $4n^3$ variables as follows:

$$\mathsf{MP}_m(x) = \bigvee_{i=1}^{m} \bigwedge_{j=1}^{4m^2} x_{i,j}.$$

Note that the Minsky Papert function can be sign represented by a degree m polynomial as follows.

$$\mathsf{MP}_m(\{x_{i,j}\}) = \text{sign}\left(-\frac{1}{2} + \prod_{i=1}^{m}\left(4m^2 + x_{i,1} + x_{i,2} + x_{i,3} + \cdots + x_{i,4m}\right)\right).$$

Hence the threshold degree of MP_m is bounded by m, $\deg_\pm(\mathsf{MP}_m) \leq m$. Minsky and Papert showed that this is in fact tight.

Theorem 24.8 (Minsky Papert Theorem [MP87]). $\deg_\pm(\mathsf{MP}_m) \geq m$.

We can now build the function F_{MP} corresponding to the Minsky-Papert function to obtain a problem in depth 3 AC^0 with exponentially small discrepancy (and hence polynomially large communication complexity). More precisely, plugging $\deg_\pm(\mathsf{MP}_m) = m$ into the degree/discrepancy theorem 24.5, we get (by setting $N = \lceil 16en^2/m \rceil = \lceil 256em^5 \rceil$),

$$\mathsf{disc}_\lambda(F_{\mathsf{MP}_m}) \leq 2^{-m} \leq e^{-\Theta(N^{\frac{1}{5}})}.$$

MP_m is in depth 2 AC_0. It is easy to see that F_{MP_m} is in depth-3 AC^0 as

$$F_{\mathsf{MP}_m}(x, y^{(1)}, \ldots, y^{(n)}) = \mathsf{MP}_m(x|_S)$$

where S is the subset of indexes represented by $y^{(1)}, \ldots, y^{(n)}$, each $y^{(i)} \in \{\pm 1\}^{\log N}$. Note that F_{MP_m} can be written as

$$F(x, y) = \mathsf{MP}_m(\varphi(x, y^{(1)}), \ldots, \varphi(x, y^{(n)}))$$

where $\varphi(x, y^{(i)})$ returns the value x at the index represented by the binary number $y^{(i)}$. Each $\varphi(x, y^{(i)})$ can be computed by a DNF formula having $2^{\log N}$ clauses and each clause having $\log N + 1$ ($\log N$ literals corresponding to values of $y_j^{(i)}$ and one extra literal corresponding to the value of x at the index represented by the binary number $y^{(i)}$) literals, which also can be represented by a CNF formula of similar size. We are interested in CNF representation because then the second layer AND gates of Minsky-Papert function can be collapsed with the AND gates of the CNF formula to obtain a depth 3 AC^0 circuit for $F(x, y)$. Here we have constructed an explicit depth 3 circuit of exponentially small discrepancy.

Corollary 24.9. *The function F_{MP_m} in depth 3 AC^0 has discrepancy at most $2^{-\Omega(N^{1/5})}$ (wrt. to some explicit distribution) and hence has randomized communication complexity $R_{1/2-\varepsilon}(F_{\mathsf{MP}_m})$ at least $\Omega(N^{1/5}) - O(\log(1/\varepsilon))$.*

We will use the above property about F_{MP_m} to show that it cannot be written as a subexponential sized majority of majority circuit. More precisely, we will show the following (even stronger) statement.

Theorem 24.10 ([She09])**.** *Suppose $F_{\mathsf{MP}_m} = \mathsf{MAJ}(h_1, \ldots, h_s)$, where each h_i is a linear threshold function*[1]*, then s is exponential.*

This is in stark contrast to the following theorem about depth 2 AC^0 computable functions due to Allender.

Theorem 24.11 ([All89])**.** *Any function computable by a depth 2 AC^0 circuit can be written as a quasi-polynomial sized Majority of Majority circuit.*

To prove Theorem 24.10, we first need the following theorem due to Nisan which relates discrepancy to the size of majority of threshold circuits.

Theorem 24.12 (communication complexity of threshold functions [Nis94])**.** *Let $f : \{\pm 1\}^n \to \{\pm 1\}$, be a linear threshold function. Then $\mathrm{R}_\varepsilon^{pub}(f) = O(\log n + \log \frac{1}{\varepsilon})$, for any partition of the variables and any $\varepsilon = \varepsilon(n)$.*

Proof of Theorem 24.10. Let us design a (public coins) randomized protocol for F_{MP_m} assuming $F_{\mathsf{MP}_m} = \mathsf{MAJ}(h_1, \ldots, h_s)$ where each h_i is a linear threshold function. The two parties can randomly pick (using public coins) an $i \in \{1, 2, 3, \ldots, s\}$ and evaluate h_i with probability $1 - \frac{1}{4s}$ using Theorem 24.12. The total communication complexity of this protocol would be $O(\log s + \log N)$ and would predict F correctly with probability at least

[1] A linear threshold function is any function of the form $\mathrm{sign}(w_1x_1 + w_2x_2 + \cdots + w_nx_n - \theta)$ where $w_i, \theta \in \mathbb{R}$. Note that $\mathsf{MAJ}(x_1, \ldots, x_n) = \mathrm{sign}(x_1 + x_2 + \cdots + x_n)$

$\left(\frac{1}{2}+\frac{1}{2s}\right)-\frac{1}{4s}=\frac{1}{2}+\frac{1}{4s}$, on every input (this is because if at least $\frac{1}{2}s+1$ of h_i's will have the value of the majority and hence the probability that value of h_i is the majority is at least $\frac{\frac{1}{2}s+1}{s}=\frac{1}{2}+\frac{1}{s}$ and then one has to subtract the probability that Nisan's protocol errors to get the total error bound). Thus we get

$$\mathrm{R}^{\mathrm{pub}}_{\frac{1}{2}-\frac{1}{4s}}(F_{\mathsf{MP}_m})=O(\log N+\log s).$$

Comparing this with Corollary 24.9 we get that $s=\exp(\Omega(N^{\frac{1}{5}}))$. Hence the theorem. □

References

[All89] Eric Allender. *A note on the power of threshold circuits.* In *Proc. 30th IEEE Symp. on Foundations of Comp. Science (FOCS)*, pages 580–584. 1989. doi:10.1109/SFCS.1989.63538.

[MP87] Marvin Minsky and Seymour Papert. *Perceptrons - an introduction to computational geometry.* MIT Press, 1987.

[Nis94] Noam Nisan. *The communication complexity of threshold gates.* In *Combinatorics, Paul Erdös is Eighty (Vol. 1)*, pages 301–315. Bolyai Math. Soc., 1994.

[She08] Alexander A. Sherstov. *Communication lower bounds using dual polynomials*. Bulletin of the EATCS, 95:59–93, 2008. arXiv:0805.2135.

[She09] ———. *Separating AC^0 from depth-2 majority circuits.* SIAM J. Computing, 38(6):2113–2129, 2009. (Preliminary version in *39th STOC*, 2007). doi:10.1137/08071421X.