CS 2429 - Foundations of Communication Complexity

Lecturer: Toniann Pitassi

1 The Discrepancy Method — Cont'd

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.

We denote the discrepancy of f (with respect to the uniform distribution) and a rectangle $A \times B$ by $\text{disc}(f, A \times B)$. 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.

1.1 The BNS Method

The BNS method is another way to bound the discrepancy, and will furnish us with yet another proof of the upper bound on $\text{disc}(\text{IP}_n)$. The method first appeared in a paper by Babai, Nisan and Szegedy.

The method is given by the following lemma:

Lemma 1 (BNS) *The discrepancy of a function $f : X \times Y \to \mathbb{Z}_2$ can be bounded as follows:*

$$\text{disc}(f, A \times B)^2 \leq \mathop{\mathbb{E}}_{y,y'} \left| \mathop{\mathbb{E}}_{x} M_f(x,y) M_f(x,y') \right|,$$

where x, y, y' are chosen independently and uniformly at random, x from X and y, y' from Y.

Proof Recall the definition of discrepancy.

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

The discrepancy can be written using expectations as

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

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

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

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 2 *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,

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

The above theorem can also be proven with respect to an arbitrary distribution λ. The more general theorem is as follows.

Theorem 3 *Let F be a function from $X \times Y$ to $\{-1, 1\}$. Then*

$$\frac{Disc_\lambda(F)^2}{|X|^2 \times |Y|^2} \leq E_{y,y'} |E_x f(x,y) f(x,y') \lambda(x,y) \lambda(x,y')|.$$

2 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.

2.1 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 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$(f) = \Theta(\log \text{DTC}(f))$.

3 Discrepancy and Duality of Sign Degree

Theorem 4 (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

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

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 5 (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:

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

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

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, $f(x) = \bigvee_{i=1}^{m} \bigwedge_{j=1}^{4m^2} x_{ij}$ 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 $F(\bar{x}, \bar{y})$ $|\bar{x}| = N$ and $|\bar{y}| = \log \binom{N}{n}$, $N = O(n^k)$. $F(\bar{x}, \bar{y}) = f(\bar{(x)}|_{\bar{y}})$, which is read "f of $\bar{x}$, restricted to the bits specified by $\bar{y}$"
That is, Alice has N bits (N will be chosen to be polynomial in n), and Bob has $\log \binom{N}{n}$ many bits. We interpret Bob's input as pointing to n locations of Alice's string. They want to compute the function f on these n (consecutive) bits.

(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 6 *(Sherstov) Let f be boolean over $x_1..x_n$ with sign degree $\geq d$.*
Then $disc(F) \leq (\frac{4en^2}{Nd})^{\frac{d}{2}}$ where e has its usual meaning as the base of the natural logarithm.

We set $N = \frac{16en^2}{d}$ so that $disc \leq 2^{-d}$. See Sherstov, Seperating AC^0 from depth-2 majority circuits, and Sherstov, Pattern Matrix Method.

Proof (Proof of Sherstov's theorem) We rename y, y' V and W.

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

$$\lambda(x, V) = 2^{-N+n} \mu(x|_V) / \binom{N}{n}.$$

By the BNS lemma,

$$\frac{disc_\lambda(F)^2}{|X|^2 \times |Y|^2} \leq E_{V,W}|E_x[f(x|_V)f(x|_W)\lambda(x,V)\lambda(x,W)]|$$

Rewriting in terms of μ we get

$$disc_\lambda(F)^2 \leq 4^n E_{V,W}|E_x[f(x|_V)f(x|_W)\mu(x|_V)\mu(x|_W)]|.$$

Let $\Gamma(V,W)$ denote $E_x[f(x|_V)f(x|_W)\mu(x|_V)\mu(x|_W)]$.

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

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

By these claims,

$$disc_\lambda(F)^2 \leq \sum_{k=d}^{n} 2^k Pr\left[|V \cap W| = k\right]$$

$$Pr\left[|V \cap W| = k\right] = \frac{\binom{n}{k}\binom{N-n}{n-k}}{\binom{N}{n}} \leq \left(\frac{en^2}{Nk}\right)^k$$

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

$$disc_\lambda(F)^2 \leq \sum_{k=d}^{n} 2^k \left(\frac{en^2}{Nk}\right)^k = \sum_{k=d}^{n} \left(\frac{2en^2}{Nk}\right)^k \leq \left(\frac{4en^2}{Nd}\right)^d$$

.

Proof of Claim 1 The basic idea here will be that by orthogonality, the expectation is zero. Let V be $x_1...x_n$ (for notational convenience).

$$\Gamma(V,W) = E_x\left[\mu(x_1...x_n)f(x_1...x_n)\mu(x|_W)f(x|_W)\right]$$

$$\Gamma(V,W) = \frac{1}{2^N}\sum_{x_1..x_n} \mu(x_1..x_n)f(x_1..x_n) \sum_{x_{n+1}..x_N} \mu(x|_W)f(x|_W)$$

$$\Gamma(V,W) = \frac{1}{2^N} E_{x_1..x_n \sim \mu} f(x_1..x_n) \left[\sum_{x_{n+1}..x_N} \mu(x|_W)f(x|_W)\right]$$

$\sum_{x_{n+1}..x_N} \mu(x|_W)f(x|_W)$ depends on $\leq d$ bits, so

$$\Gamma(V,W) = 0$$

Proof of Claim 2 We want to show that if $|V \cap W| = i$, then $|\Gamma(V,W)| \leq 2^{i-2}$. Again for notational convenience we will assume that $V = \{1,2,\ldots,n\}$ and $W = \{1,2,\ldots,i\} \cup \{n+1, n+2, \ldots, n+(n-i)\}$. Then we have:

$$|\Gamma(V,W)| \leq E_x[|f(x|_V)\mu(x|_V)f(x|_W)\mu(x|_W)|]$$

$$|\Gamma(V,W)| \leq E_{x_1,\ldots,x_{2n-i}}[\mu(x_1,\ldots,x_n)\mu(x_1,\ldots,x_i,x_{n+1},\ldots,x_{2n-i}]$$

$$|\Gamma(V,W)| \leq E_{x_1,\ldots,x_n}[\mu(x_1,\ldots,x_n)] \cdot max_{x_1,\ldots,x_i} E_{x_{n+1},\ldots,x_{2n-i}}[\mu(x_1,\ldots,x_i,x_{n+1},\ldots,x_{2n-i})]$$

The first quantity above, $E_{x_1,\ldots,x_n}[\mu(x_1,\ldots,x_n]$ 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.

4 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 7 (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 8 (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 $R^{pub}_{\epsilon}(f) = O(\log n + \log \frac{1}{\epsilon})$.

Suppose F has (low) discrepancy $e^{-N^{\varepsilon}}$. 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 $F \in AC^0_3$.

5 Extensions of Sherstov

5.1 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.

5.2 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.