COS597D: Information Theory in Computer Science	November 14 and 16, 2011
Lecture 15-16	
Lecturer: Mark Braverman	*Scribe: Dan Larkin*

1. Discrepancy

Our goal is to provide tools for proving lower bounds about the distributional communication complexity of a function f. One such tool will be the *discrepancy* of the function with respect to a distribution μ. Recall that for deterministic communication complexity, we used the idea of monochromatic rectangles to prove lower bounds. These rectangles corresponded to leaves in a binary tree representing the communication protocol, and thus we found that a communication protocol which used c bits would have at most 2^c rectangles.

If we allow some mistakes, the rectangles need not be monochromatic. Our goal is to come up with a partition which features large, unbalanced rectangles so that we may bound the error rate with respect to μ, but reduce the depth of the tree. In other words, we want to get some sort of communication advantage from these leaves.

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

$$Disc_\mu(R,f) = \left| \Pr_\mu [f(x,y) = 0 \text{ and } (x,y) \in R] - \Pr_\mu [f(x,y) = 1 \text{ and } (x,y) \in R] \right|$$

Equivalently,

$$Disc_\mu(R,f) = \Pr_\mu [(x,y) \in R] \cdot \left| \Pr_\mu [f(x,y) = 0 | (x,y) \in R] - \Pr_\mu [f(x,y) = 1 | (x,y) \in R] \right|$$

Furthermore we denote the discrepancy of f with respect to μ as follows:

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

Observation 2. *The advantage we get from a leaf is at most the discrepancy of the corresponding rectangle.*

Theorem 3. *For every function $f : X \times Y \mapsto \{0,1\}$, every distribution μ on $X \times Y$, and all $\epsilon > 0$, $D^\mu_{1/2+\epsilon}(f) \geq \log_2 \frac{2\epsilon}{Disc_\mu(f)}$.*

Proof Let Π be a protocol that uses c bits of communication, attempting to compute f, which is correct a fraction $1/2 + \epsilon$ of the time. Then we have a bound on the probability that Π and f disagree. We can then express these probabilities as a sum over all leaves l of Π.

$$2\epsilon \leq \Pr_\mu [\Pi(x,y) = f(x,y)] - \Pr_\mu [\Pi(x,y) \neq f(x,y)]$$

$$= \sum_{\text{leaves } l} \left[\Pr_\mu [\Pi(x,y) = f(x,y) \wedge (x,y) \in l] - \Pr_\mu [\Pi(x,y) \neq f(x,y) \wedge (x,y) \in l] \right]$$

Since $\text{Disc}_\mu(f)$ is defined as the maximum over all leaves of this inner expression, trivially each term in this sum is at most $\text{Disc}_\mu(f)$. Furthermore there are 2^c leaves, so the above expression is at most $2^c \text{Disc}_\mu(f)$.

$$2^c \geq \frac{2\epsilon}{\text{Disc}_\mu(f)} \Rightarrow c \geq \log_2 \frac{2\epsilon}{\text{Disc}_\mu(f)}$$

■

As an aside, suppose that we have a function f and distribution μ such that $\text{Disc}_\mu(f) = 2^{-k}$ for some k. Can we get a randomized protocol with $\epsilon \approx 2^{-k}$ which uses $O(1)$ bits? *Yes.* Take R to be some rectangle such that $\text{Disc}_\mu(R, f) = 2^{-k}$. Such an R must exist or $\text{Disc}_\mu(f)$ would be less than 2^{-k}. We can use a constant number of bits of communication to determine if the given point is in R. If it is, then output the bit that is the more likely value of f on R, otherwise output a bit according to a $B_{1/2}$ distribution.

Let $\mathcal{U}$ be the uniform distribution, and let $\texttt{IP}(x, y) = \langle x, y \rangle$ be the inner product of x and y over $\mathbb{Z}_2$.

Theorem 4. *Suppose f is a random function. Then $D^{\mathcal{U}}_{2/3}(f) \geq \Omega(n)$.*

Proof Idea If you fill the function table of f at random, then any large rectangles you consider are likely to be very balanced. Thus by the discrepancy bound the communication complexity of f would be linear. ■

Theorem 5. $Disc_{\mathcal{U}}(IP) \leq 2^{-n/2} \Rightarrow D^{\mathcal{U}}_{1/2-\epsilon}(IP) = \Omega(n)$

Proof Let H be the Hadamard matrix of size $2^n \times 2^n$. $H(x, y) = 1$ if $\langle x, y \rangle = 0$, and $H(x, y) = -1$ otherwise.

Claim 6. $HH^T = 2^n I$

Proof

$$HH^T(x, y) = \sum_{z \in \{0,1\}^n} H(x, z) \cdot H(z, y)$$

Then consider the diagonal, all pairs (x, y) such that $x = y$. Clearly the product will be 1 for all z, and all the diagonal entries will indeed be 2^n. The off-diagonal entries then, will correspond to pairs such that $x \neq y$. For each such pair, consider one index i such that $x_i \neq y_i$. Then partition the z vectors such that each pair $(z, \overline{z})$ differ only in coordinate i. Then $H(x, z) \cdot H(z, y) + H(x, \overline{z}) \cdot H(\overline{z}, y) = 0$ for any such pair. Therefore the off-diagonal entries all benefit from gratuitous cancellation and are identically 0 as claimed. ■

Now consider each $S \times T \subseteq \{0, 1\}^n \times \{0, 1\}^n$:

$$\text{Disc}_\mu(S \times T, \texttt{IP}) = \frac{\left|\sum_{x \in S, y \in T} H(x, y)\right|}{2^{2n}} = \frac{\left|1_S^T H 1_T\right|}{2^{2n}} \leq \frac{\left|\sqrt{|S|} \cdot 2^{n/2} \cdot \sqrt{|T|}\right|}{2^{2n}} \leq \frac{\left|2^{n/2} \cdot 2^{n/2} \cdot 2^{n/2}\right|}{2^{2n}} = 2^{-n/2}$$

■

2. Direct Sums

It is an interesting question to consider not just solving one instance of a problem, but several instances simultaneously. Can we potentially save some communication costs if we solve these in parallel rather than serially? Clearly we cannot "lose" anything, but can we make the inequality strict for some functions?

$$cost(f(x), f(y)) \stackrel{?}{<} cost(f(x)) + cost(f(y))$$

This question is inspired by such savings in circuit computation. For example, multiplying a fixed $n \times n$ matrix A by a vector v will typically require a circuit of size $\gtrsim n^2$. At the same time, multiplying A by n input vectors v_i can be performed in less than the naïve n^3 time [Strassen, 1969]. We will discuss protocols for the easier variant of solving *each* copy with error ϵ rather than simultaneously solving *all* copies with error ϵ.

Let us consider then EQ, the problem of deciding if a pair of strings are equal. We saw previously that $\mathrm{R}^{pub}_{\epsilon}(\texttt{EQ}) = \log 1/\epsilon$. Intuitively, the complexity of checking equality of k pairs of strings, $\texttt{EQ}^k$, is at most k times the complexity of checking for a single pair. We can actually do much better.

Theorem 7. $R^{pub}_{\epsilon}\left(\texttt{EQ}^k\right) = O(k + \log 1/\epsilon)$ *[Feder, Kushilevitz, Naor, Nisan, 1991]*

Proof Idea Recall that the normal randomized protocol for solving a single copy of EQ involved sending hashes of the input. This protocol will be an adaptation which groups the input pairs for hashing. The protocol will proceed in $\log k$ rounds as follows. At round i, split the inputs into blocks of size 2^i. Send $O(1)$ hashes to look for inequality in each block. If a difference is found, then search the block to find the offending pair and remove it. Continue with the remaining pairs. After all such rounds have completed, run an extra $\log 1/\epsilon$ additional hashes. The main idea is that in each round, in expectation at least half the errors will be removed. The splitting must occur randomly to separate the errors into opposite groups. When a block reports an error, the search can be done with cost logarithmic in the block size. Essentially, there is a very long tail on the expected sum of communication cost, but the tail is very light, and the sum is essentially constant. ■

Consider the problem of sending a random message M across a noiseless channel. How many bits on average does it take to send M? Let e be the encoding of M, and $l(e)$ be the length of the encoding. Then, by some simple encoding (such as Huffman), we can achieve

$$H(M) \leq \mathrm{E}\left[l(e)\right] \leq H(M) + 1$$

Now, how many bits does it take to send k independent copies?

$$k \cdot H(M) \leq \mathrm{E}\left[l(e^k)\right] \leq k \cdot H(M) + 1 \qquad \lim_{k\to\infty} \frac{\mathrm{E}\left[l(e^k)\right]}{k} = H(M)$$

By combining inequalities, we also see that

$$\mathrm{E}\left[l(e^k)\right] \geq k\left(\mathrm{E}\left[l(e)\right] - 1\right)$$

which is the best that we can hope for.

3. Information Cost

Definition 8. *The information cost of a protocol over a distribution μ of inputs is*

$$IC(\Pi, \mu) = I_\mu(\Pi; Y|X) + I_\mu(\Pi; X|Y)$$

This can be thought of informally as the total of what Alice, who knows X, and Bob, who knows Y, learn from the interaction. We can also define the information cost of a function.

Definition 9.

$$IC(f, \Pi, \epsilon) = \inf_{\Pi: \Pr_\mu[\Pi(x,y) \neq f(x,y)] < \epsilon} IC(\Pi, \mu)$$

Claim 10. $IC(\pi, \mu) \leq E[|\Pi_A|] + E[|\Pi_B|] = E[|\Pi|] \leq |\Pi|$

The final term is just notation for the communication cost of Π. It is clear that you cannot learn more bits of information from the interaction than the number of bits exchanged.

Recall the following formulae, as they will be of use during the proof of the next claim.

$$\mathrm{I}(B; D|AC) = 0 \Rightarrow \mathrm{I}(A; B|C) \geq \mathrm{I}(A; B|CD) \qquad \mathrm{I}(B; D|C) = 0 \Rightarrow \mathrm{I}(A; B|C) \leq \mathrm{I}(A; B|CD)$$

Claim 11. *Private randomness doesn't matter, i.e.* $I(\Pi; X|Y) = I(\Pi; X|YP_B)$.

Proof Suppose that the interaction consists of $2l$ rounds, such that Alice sends bits in the odd rounds, and Bob the even. We will prove the above by induction on l.

$$\Pi = \Pi_1 \Pi_2 \ldots \Pi_{2l-1} \Pi_{2l}$$

The base case of $l = 0$ is trivial since obviously there is no information at all on either side of the equation. So suppose we know the following.

$$\mathrm{I}(\Pi_1 \ldots \Pi_{2l-2}; X|Y) = \mathrm{I}(\Pi_1 \ldots \Pi_{2l-2}; X|YP_B)$$

By applying the chain rule, we find

$$\begin{aligned}\mathrm{I}(\Pi_1 \ldots \Pi_{2l}; X|Y) = &\mathrm{I}(\Pi_1 \ldots \Pi_{2l-2}; X|Y) + \mathrm{I}(\Pi_{2l-1}; X|Y\Pi_1 \ldots \Pi_{2l-2}) + \\ &\mathrm{I}(\Pi_{2l}; X|Y\Pi_1 \ldots \Pi_{2l-1})\end{aligned}$$

The last term we can think of as the shared information between Alice's input and what Bob says conditioned on his prior knowledge. The only dependence on X is based on what he already knows, so this is 0.

$$\begin{aligned}\mathrm{I}(\Pi_1 \ldots \Pi_{2l}; X|YP_B) = &\mathrm{I}(\Pi_1 \ldots \Pi_{2l-2}; X|YP_B) + \mathrm{I}(\Pi_{2l-1}; X|YP_B\Pi_1 \ldots \Pi_{2l-2}) + \\ &\mathrm{I}(\Pi_{2l}; X|YP_B\Pi_1 \ldots \Pi_{2l-1})\end{aligned}$$

By similar reasoning, the last term is again 0. So we are left with the following:

$$\mathrm{I}(\Pi_1 \ldots \Pi_{2l}; X|Y) = \mathrm{I}(\Pi_1 \ldots \Pi_{2l-2}; X|Y) + \mathrm{I}(\Pi_{2l-1}; X|Y\Pi_1 \ldots \Pi_{2l-2})$$

$$\mathrm{I}(\Pi_1 \ldots \Pi_{2l}; X|YP_B) = \mathrm{I}(\Pi_1 \ldots \Pi_{2l-2}; X|YP_B) + \mathrm{I}(\Pi_{2l-1}; X|YP_B\Pi_1 \ldots \Pi_{2l-2})$$

Since, by our inductive argument we know

$$\mathrm{I}(\Pi_1 \ldots \Pi_{2l-2}; X|Y) = \mathrm{I}(\Pi_1 \ldots \Pi_{2l-2}; X|YP_B)$$

It thus suffices to show that

$$\mathrm{I}(\Pi_{2l-1}; X|Y\Pi_1 \ldots \Pi_{2l-2}) = \mathrm{I}(\Pi_{2l-1}; X|YP_B\Pi_1 \ldots \Pi_{2l-2})$$

Consider the following substitutions:

$$A = X \qquad B = \Pi_{2l-1} \qquad C = Y\Pi_1 \dots \Pi_{2l-2} \qquad D = P_B \qquad \mathrm{I}(A;B|C) \stackrel{?}{=} \mathrm{I}(A;B|CD)$$

By the formulae referenced prior to this proof, we need only show

$$\mathrm{I}(B;D|AC) = \mathrm{I}(\Pi_{2l-1};P_B|\Pi_1 \dots \Pi_{2l-1}YX) = 0 \qquad \mathrm{I}(B;D|C) = \mathrm{I}(\Pi_{2l-1};P_B|\Pi_1 \dots \Pi_{2l-1}Y) = 0$$

Both statements are about the shared information between Bob's random string and what Alice says. Everything Alice knows about P_B is included in the previous interaction. Since in both cases we condition on the interaction, there will be no extra shared information. ■

The definition of information cost presented here is relatively recent. More traditionally, there was a notion of what we will call *external* information, $\mathrm{IC}_{ext} = \mathrm{I}(\Pi; XY)$, which corresponds to what an outside observer learns from the communication.

Claim 12. $IC_{ext}(\Pi, \mu) \geq IC(\Pi, \mu)$

This relation is easy to get wrong at first, and typically is the opposite of what a cryptographer would want. In an information theoretical sense, someone listening learns more than those talking. For an intuitive explanation, say that Alice and Bob are an old married couple and know everything about each other. They don't really say anything new to each other when they talk, so the normal information cost is close to 0. On the other hand, the outside observer may not know everything about them, so they can still stand to learn something.

Proof

$$\Pi = \Pi_1\Pi_2 \dots \Pi_{2l-1}\Pi_{2l}$$

Let us assume as our inductive hypothesis that

$$\mathrm{I}(\Pi_1 \dots \Pi_{2l-1}; XY) \geq \mathrm{I}(X; \Pi_1 \dots \Pi_{2l-1}|Y) + \mathrm{I}(Y; \Pi_1 \dots \Pi_{2l-1}|X)$$

Then we proceed to prove the claim for $2l$. Assume w.l.o.g. that Alice speaks in the $2l$-th round.

$$\begin{aligned}
\mathrm{I}(\Pi_1 \dots \Pi_{2l}; XY) &= \mathrm{I}(\Pi_1 \dots \Pi_{2l-1}; XY) + \mathrm{I}(\Pi_{2l}; XY|\Pi_1 \dots \Pi_{2l-1}) \\
&\geq \mathrm{I}(\Pi_1 \dots \Pi_{2l-1}; X|Y) + \mathrm{I}(\Pi_1 \dots \Pi_{2l-1}; Y|X) + \\
&\quad \mathrm{I}(\Pi_{2l}; X|\Pi_1 \dots \Pi_{2l-1}) + \mathrm{I}(\Pi_{2l}; Y|\Pi_1 \dots \Pi_{2l-1}X) \\
&= \mathrm{I}(\Pi_1 \dots \Pi_{2l}; Y|X) + \mathrm{I}(\Pi_1 \dots \Pi_{2l-1}; X|Y) + \mathrm{I}(\Pi_{2l}; X|\Pi_1 \dots \Pi_{2l-1}) \\
&\geq \mathrm{I}(\Pi; Y|X) + \mathrm{I}(\Pi_1 \dots \Pi_{2l-1}; X|Y) + \mathrm{I}(\Pi_{2l}; X|\Pi_1 \dots \Pi_{2l-1}Y) \\
&= \mathrm{I}(\Pi; Y|X) + \mathrm{I}(\Pi; X|Y)
\end{aligned}$$

The last inequality follows from substituting

$$A = X \qquad B = \Pi_{2l} \qquad C = \Pi_1 \dots \Pi_{2l-1} \qquad D = Y,$$

and observing that the since Π_{2l} is a message sent by Alice,

$$\mathrm{I}(\Pi_{2l}; Y|\Pi_1 \dots \Pi_{2l-1}X) = 0.$$

■

We note that Claim 12 becomes an equality if μ is a product distribution, i.e. if it can be written as $\mu(x, y) = \mu_X(x) \cdot \mu_Y(y)$. In this case the inputs to Alice and Bob are independent, and there is no difference between what an observer learns and what the participants learn.