REU 2006 · Discrete Math · Lecture 14

Instructor: László Babai
Scribe: Megan Guichard
Editor: Duru Türkoğlu

August 2, 2006. Last updated August 6, 2006 at 5:30pm.
NOT PROOF-READ

14 Communication Complexity (continued)

14.1 Lower Bound Analysis (continued)

Last time we asked a question about the communication complexity of Boolean functions. Today we will consider a related result. As before, let $f : \{0,1\}^{2n} \to \{0,1\}$ be a boolean function. Let M_f be the 2^n by 2^n matrix with entries $f(\underline{x}, \underline{y})$ for $(\underline{x}, \underline{y})$ vectors in $\{0,1\}^n$.

Now suppose M_f can be partitioned into 2^s homogeneous rectangles. (A rectangle is given by taking any subset of the rows and any subset of the columns. Homogeneous means that all entries in that rectangle are the same, i.e. all are zeros or all are ones.)

Claim 1. $C(f) = O(s^2)$, where $C(f)$ is the communication complexity.
Hint: Notice there are 2^s rectangles, so you need s bits to name them. s^2 communication bits come in s groups of s each.

Proof. Suppose Alice says, "My row intersects this rectangle," i.e. the row set of the rectangle includes Alice's row. If Bob's column also intersects the same rectangle, then you know what the value of f is. Indeed, the intersection of Alice's row and Bob's column is in the same rectangle.

There are 2^s rectangles; we want to have just one. To get to just one in s rounds, we need to reduce by a factor of two each time.

We want to show that either Alice or Bob can announce a rectangle which eliminates half the rectangles. Suppose Bob announces that his column intersects a rectangle, $S \times T$, where $S \subset \{\text{rows}\}$ and $T \subset \{\text{columns}\}$. So all rectangles $S' \times T'$ for which $T' \cap T = \emptyset$ are eliminated. Since Bob knows the layout of the rectangles, he can see whether there exists a rectangle which he can announce which would eliminate half the rectangles.

Notice that the intersection of two rectangles $S_1 \times T_1$ and $S_2 \times T_2$ is $(S_1 \cap S_2) \times (T_1 \cap T_2)$, which is nonempty if and only if $S_1 \cap S_2 \neq \emptyset$ **and** $T_1 \cap T_2 \neq \emptyset$. Our rectangles are disjoint. So any two distinct rectangles must have at least one of $S_1 \cap S_2$ and $T_1 \cap T_2$ empty. Having this information at hand we can claim the following:

Claim 2. Either Alice or Bob can announce a rectangle that eliminates at least half of the remaining rectangles.

Proof. Let $R = S \times T$ be the rectangle containing (X, Y). Then either at least half the remaining rectangles are horizontally disjoint from R, or at least half are vertically disjoint (or both). □

This gives us the desired procedure. On a turn, Alice either announces a rectangle or passes. If she passes, then Bob announces a rectangle. Suppose without loss of generality Bob has announced a rectangle. Then all but the rectangles intersecting the columns of his rectangle are eliminated. We can think of this as transforming M_f into a rectangular matrix which has had the columns outside of Bob's rectangle removed. This matrix is divided into at most 2^{s-1} homogeneous rectangles. Repeat this s times; then we know which rectangle our entry lived in, and so we are done. □

This is an $O(s^2)$ algorithm. Recall that we saw last time that the communication complexity is at least the log of the number of partitions.

Lovász defines the communication complexity as follows: the communication is complete when one player knows the answer, and the other player knows this fact about the first player. So for instance, the communication complexity of the function which computes the combined parity of $\underline{x}$ and $\underline{y}$ has communication complexity 1.

Using the above definition, equality testing requires exactly n bits of communication, and we have also shown that randomization improves the communication by an exponential speedup, resulting in randomized complexity of logarithmic order.

One intriguing question would be: "Does there exist functions of boolean inputs, which even using randomization require linear order complexity, say $\frac{1}{2}n$?" Below are some candidates:

1. What is the communication complexity of the function which computes the parity of the number of bits where $\underline{x}$ and $\underline{y}$ agree?

2. Something that might be good for answering this question is the inner product mod 2:
   $IP_2(x, y) = \sum x_i y_i \pmod 2$, where x, y are strings of 1s and -1s.

3. Let p be an n-bit prime, $0 \leq x \leq p-1$, and $0 \leq y \leq p-1$. What is the complexity of the function which computes $\left(\frac{x+y}{p}\right)$? (Technically, we need to change the Legendre symbol so that it only takes on two values; for example, say that zeros get changed into ones.)

14.2 Distributional Complexity

The notion of complexity we have been discussing is known as worst-case analysis of random protocols. It focuses on what the worst complexity for a given pair of inputs is, for general (possibly nondeterministic) protocols.

A different concept of complexity is **distributional complexity**. Here the protocol is deterministic, but the inputs are chosen at random. So this is a kind of "average-case" analysis.

There is a generic principle that is capable of using the distributional complexity to give a lower bound on the worst-case complexity.

Let $C_\varepsilon(f)$ be the worst-case complexity of calculating f by a randomized protocol. Then we have the following guarantee:

$$(\forall(x, y))(\Pr(error) \leq \varepsilon) \tag{14.2.1}$$

The $\forall$ quantifier here shows that we are talking about worst-case complexity. On the other hand, distributional complexity reflects average case complexity.

Let $D_\varepsilon(f)$ be

$$\max_{\mu} \min_{P} (\text{cost of } P) \tag{14.2.2}$$

where μ is a distribution over the inputs x, y and we are chosing among those protocols P such that $\Pr_{\mu,P}(\text{error}) \leq \varepsilon$. As before, cost of P is the maximum cost over inputs (x, y).

Theorem 3. $C_\varepsilon(f) \geq D_\varepsilon(f)$.

Proof. Create a matrix with rows corresponding to pairs of inputs, and columns corresponding to deterministic protocols. Put a 0 if the answer given by the protocol is correct, and a 1 if the answer is wrong. Then the average of a column corresponding to a protocol matching $D_\varepsilon(f)$ is $\leq \varepsilon$.

Our strategy, very roughly speaking, is as follows: a deterministic protocol chosen at random is essentially the same as a randomized protocol.

In more detail, suppose we list all deterministic protocols and choose one at random, according to some probability distribution. Let π be a probability distribution on deterministic protocols; π is essentially a randomized protocol.

Returning to our matrix, the average of a row according to π will be at most ε, given that π is a randomized protocol with probability of error at most ε. So the average of the entire table according to π is at most ε. Hence there must be a column whose vertical average is at most ε! This column is a deterministic protocol which will make at most ε error. So there exists a deterministic protocol whose probability of error is at most the probability of error of the best randomized protocol. In other words, $C_\varepsilon(f) \geq D_\varepsilon(f)$. □

In the D_ε concept, the protocol P computes not f but some function g such that

$$\Pr_{\mu}(f(x, y) \neq g(x, y)) \leq \varepsilon$$

Since the protocol computes g, there is a rectangle cover (a partition of M_g into homogeneous rectangles) of M_g. But M_f differs from M_g in at most ε portion of the entries; so the partititon of M_g gives a partition of M_f which is not quite homogeneous. This partition computes f with probability of error at most ε. Hence if we are interested in proving lower bounds on

$D_\varepsilon(f)$, we must somehow account for not just homogeneous partitions but also almost-homogeneous partitions.

One method is as follows. Pick your favorite distribution μ; often, this will be simply the uniform distribution (inputs are picked uniformly at random.) Then prove an upper bound on the **discrepancy** of every rectangle. The discrepancy is the difference between the number of ones and the number of zeros in a rectangle, normalized by the size of the matrix.

If M is an $N \times N$ $(0,1)$-matrix, $R \subset M$ is a $k \times \ell$ rectangle in M, and R has a ones and b zeros (where $a + b = k\ell$), then the discrepancy of R with respect to the uniform distribution is

$$\mathrm{disc}(R) = \frac{|a-b|}{N^2}.$$

Suppose we can prove, for a matrix M_f, an upper bound on the discrepancy; that is, we show $(\forall R \subset M)(\mathrm{disc}(R) \leq \delta)$. Then write $M = \bigcup_{k=1}^{2^s} R_j$ as the disjoint union of 2^s rectangles R_j. Let $s = D_\varepsilon(f)$. Let R_j be a $k_j \times \ell_j$ rectangle. If $\mathrm{disc}(R_j) \leq \delta$, then we have $k_j\ell_j = a_j + b_j$, where $a_j - b_j \leq \delta N^2$. So $2b_j \geq k_j\ell_j - \delta N^2$. Next time we will see how this gives a lower bound on the error on R_j. We will need Hadamard matrices in this task.

Definition 4. A square matrix H is a **Hadamard matrix** if all entries of H are ± 1, and the rows of H are orthogonal (hence $HH^T = n \cdot I$, if H is $n \times n$).

Exercise 5. Construct $2^n \times 2^n$ Hadamard matrices for every n, and relate them to M_f where f is the inner product modulo 2.

Exercise* 6. (Lindsay's Lemma) If R is a $k \times \ell$ rectangle in an $N \times N$ Hadamard matrix H, then

$$\left| \sum_{(i,j) \in R} h_{ij} \right| \leq \sqrt{k \cdot \ell \cdot N}.$$