CC3: The Good, The Bad, The Ugly

Mihai Pătraşcu
IBM Almaden

April 3, 2009

This is the 3rd post in the thread on communication complexity, following the episode on randomized and distributional complexity.

In today's episode, we will prove our first randomized/distributional lower bound. We will consider a problem that appears quite trivial at first sight – INDEXING, defined as follows:

- Alice receives a bit vector, $x \in \{0,1\}^n$;
- Bob receives an index, $y \in [n]$ (note standard notation: $[n] = \{1, \ldots, n\}$);
- their goal is to output x_y, i.e. the y-th bit of the vector x.

One way to view INDEXING is as the simplest case of lopsided set disjointness, where Bob's set is of size one: Alice receives the set $S = \{i \mid x_i = 1\}$, and Bob receives $T = \{y\}$.

Intuition. What trade-off should we expect between Alice's communication a, and Bob's communication b? Clearly, the problem can be solved by $\big[a = 1, b = \lg n\big]$ and by $\big[a = n, b = 1\big]$.

In between these two extremes, the best use of b seems to be for Bob to send b bits of his index. Alice now knows the index to be in a set of $n/2^b$ possibilities. She can simply send all her bits at these positions, using communication $a = n/2^b$. Finally, Bob announces the answer with one more bit.

We thus showed the upper bound: $a \leq n/2^{b-1}$. It should be fairly intutive that this is also a tight lower bound (up to constants). Indeed, no matter what Bob communicates, his index will be uncertain in a set of $n/2^b$ possibilities (on average). If Alice sends less than $n/2^{b+1}$ bits of information, at least half of the possible index positions will not have a specified answer based on Alice's message. In other words, the protocol fails to determine the answer with constant probability (i.e. makes constant error).

Distributions. Before proving a distributional lower bound, we must find the distribution that makes the problem hard. From the intuition above, it should be clear that the right distributions are uniform, both for Alice's vector and for Bob's index.

Rectangles. We are in the situation of product distributions: the inputs of Alice and Bob are independent. This is a very good situation to be in, if you remember the main take-home lesson from Episode 1: *rectangles*. Remember that some fixed communication transcript is realized in a combinatorial rectangle $A \times B$, where A is a subset of Alice's possible inputs, and B a subset of Bob's possible inputs. The main ideas of a deterministic proof were:

- little communication from Alice means A is large;
- little communication from Bob means that B is large;
- but the rectangle must be monochromatic (the answers must be fixed, since the communication is over);
- if A and B are large, you must have both yes-instances and no-instances.

For a product distribution, we will perform essentially the same analysis, given that the densities $\mu(A)$ and $\mu(B)$ can be measured independently. The only difference will be that the rectangle is not monochromatic, but almost monochromatic. Indeed, if the protocol may make some errors, the answer need not be fixed in the rectangle. But it must happen that one answer (yes or no) is much more frequent — otherwise the error is large.

The first three steps of the analysis are formalized in the following randomized richness lemma, from the classic paper of [Miltersen, Nissan, Safra, Wigderson STOC'95]:

Lemma 1. *(randomized richness) Say Alice and Bob compute a function $f : X \times Y \to \{0,1\}$ on a product distribution over $X \times Y$. Assume that:*

- *f is dense, in the sense that $\mathbb{E}[f(x,y)] \geq \alpha \geq \frac{1}{5}$.*
- *the distributional error is at most ε.*
- *Alice communicates a bits, and Bob b bits.*

Then, there exists a rectangle $A \times B$ ($A \subset X$, $B \subset Y$) satisfying:

- *Alice's side is large: $\mu(A) \geq 2^{-O(a)}$;*
- *Bob's side is large: $\mu(B) \geq 2^{-O(a+b)}$;*
- *the rectangle is almost monochromatic: $\mathbb{E}[f(x,y) \mid x \in A, y \in B] \geq 1 - O(\varepsilon)$.*

Proof. Though the statement is very intuitive, the proof is actually non-obvious. The obvious approach, which fails, would be some induction on the bits of communication: fix more bits of the communication, making sure the rectangle doesn't decrease too much, and the error doesn't increase too much in the remaining rectangle.

The elegant idea is to use the deterministic richness lemma. Let F be the output of the protocol (what Alice and Bob answer). We know that f and F coincide on $1 - \varepsilon$ of the inputs. By definition, the protocol computes F deterministically with no error (duh!). It is also clear that F is rich, because it is dense $\mathbb{E}[F] \geq \alpha - \varepsilon$.

So apply the deterministic richness lemma from Episode 1. We get a large rectangle of F in which the answer is one. But how do we know that f is mostly one in the rectangle? It is true that F and f differ on only ε of the inputs, but that ε might include this entire rectangle that we chose! (Note: the rectangle size is $\sim 2^{-a} \times 2^{-b}$, so much much smaller than some constant ε. It could all be filled with errors.)

We were too quick to apply richness on F. Now define G, a cleaned-up version of F. Consider some transcript of the protocol, leading to a rectangle $A \times B$. If the answer is zero, let $G = 0$ on $A \times B$. If the answer is one, but the error inside this rectangle is more than 10ε, also let $G = 0$. Otherwise, let $G = 1$ on the rectangle.

How much of the truth table gets zeroed out because of excessive error (above 10ε)? Well, the overall average error is ε, so we can apply a Markov bound to it: the mass of all rectangles in which the error exceeds 10ε is at most $\frac{1}{10}$.

Thus, G is also fairly dense: $\mathbb{E}[G] \geq \mathbb{E}[F] - \frac{1}{10} \geq \alpha - \frac{1}{10} - \varepsilon \geq \frac{1}{10} - \varepsilon$. Thus, G is rich, and we can find a big rectangle in which it is identically one. But in that rectangle, $\mathbb{E}[f] \geq 1 - 10\varepsilon$ by construction of G. □

The good, the bad, the ugly. It remains to prove that in any large rectangle, the fraction of zeros must be non-negligible (it may not be almost all ones). This part is, of course, problem specific, and we shall do it here for INDEXING.

Unfortunately, these proofs are somewhat technical. They typically apply a number of "pruning" steps on the rectangle. An example of pruning on the space of rectangles was seen above: we zeroed out all rectangles that had more than 10ε error. In these proofs, we throw out rows and columns we don't like for various reasons. One usually makes many funny definitions, talking about "good rows", "bad rows", "ugly columns", etc.

While looking at such proofs, it is important to remember the information-theoretical intuition behind them. After you understand why the statement is true (handwaving about the information of this and the entropy of that), you can deal with these technicalities on the night before the STOC/FOCS deadline.

Here is how the proof proceeds for INDEXING:

Lemma 2. *Consider any $A \subset \{0,1\}^n$ and $B \subset [n]$ such that $|A| \geq 2^{n+1}/2^{|B|/2}$. Then $\mathbb{E}_{A,B}[f(x,y)] \geq 0.95$.*

Proof. Assume for contradiction that $Pr_{A,B}[f(x,y) = 0] \leq \frac{1}{20}$. Define an ugly row as a row $x \in A$ such that $\Pr_B[f(x,y) = 0] > \frac{1}{10}$. At most half of the rows are ugly by the Markov bound. Discard all ugly rows, obtaining $A' \subset A$, with $|A'| \geq |A|/2$.

For every remaining $x \in A'$, we have $x_y = 0$ for at least $0.9|B|$ indices from B (this is the definition of f). Call these good indices.

There are $\binom{|B|}{0.9|B|} = \binom{|B|}{0.1|B|}$ choices for the good indices. For every set of good indices, there are at most $2^{n-0.9|B|}$ vectors x which are zero on the good indices. Thus:

$$|A'| \;\leq\; \binom{|B|}{0.1|B|} \cdot 2^{n-0.9|B|} \;\leq\; 2^n \cdot 2^{O(0.1|B| \lg 10)}/2^{0.9|B|} \;=\; 2^n/2^{0.9|B| - O(0.1|B| \lg 10)}$$

This is at most $2^n/2^{0.5|B|}$, at least for a sufficiently small value of $\frac{1}{10}$ (I never care to remember the constants in the binomial inequalities). We have reached a contradiction with the lemma's guarantee that A is large. □

Combine this with the richness lemma with an error $\varepsilon = 0.05$. We have $|A| \approx 2^{n-a}$ and $|B| \approx n/2^b$, so it must be the case that $a = \Omega(|B|) \geq n/2^{O(b)}$. We have obtained a tight lower bound.