Theoretical Computer Science 106 (1992) 385–390
Elsevier

Note

On the distributional complexity of disjointness

A.A. Razborov

Steklov Mathematical Institute, Vavilova, 42, GSP-1, 117966, Moscow, Russia

Communicated by M.S. Paterson
Received October 1990
Revised October 1991

Abstract

Razborov, A.A., On the distributional complexity of disjointness, Theoretical Computer Science 106 (1992) 385–390.

We prove that the distributional communication complexity of the predicate "disjointness" with respect to a very simple measure on inputs is $\Omega(n)$.

1. Introduction

The following concept of the *ε-error probabilistic communication complexity* $C_\varepsilon(A)$ of a binary predicate $A(x, y)$ was introduced by Yao [3]. Assume that two infinitely powerful computers evaluate the predicate $A(x, y)$ in the situation when the first computer possesses x and the second possesses y (x and y are binary strings of length n). They do this by interchanging messages between each other. Both computers are allowed to flip a coin. At the end of the communication for each x and y they must output the correct value of $A(x, y)$ with probability at least $1-\varepsilon$. The complexity is measured by the expected number of communications in the worst case. For more details see [3, 1].

In [4], Yao suggested an approach to estimating $C_\varepsilon(A)$ from below and gave an application of this approach. It is based upon the notion of the ε-error distributional communication complexity $D_\varepsilon(A)$ of a binary predicate $A(x, y)$. This notion, in turn, was generalized in [1] to the concept of the *ε-error distributional communication*

complexity $D_\varepsilon(A \mid \mu)$ *under an arbitrary probabilistic measure* μ *on inputs* ($D_\varepsilon(A)$ is just $D_\varepsilon(A \mid \mu)$ with uniform μ). This concept is somewhat dual to $C_\varepsilon(A)$: now the computers run a *deterministic* protocol and are required to output the correct value of $A(x, y)$ everywhere except for at most ε-fraction (with respect to the measure μ) of inputs. It was proved in [4] for uniform μ and generalized in [1] to arbitrary μ that $C_\varepsilon(A) \geqslant \frac{1}{2} D_{2\varepsilon}(A \mid \mu)$ for any A, μ and $\varepsilon > 0$.

Several authors studied these complexity measures for the predicate "disjointness". Let DIS_n denote this predicate (we will recall its definition below). Babai et al. [1] proved that $D_\varepsilon(\mathrm{DIS}_n \mid \mu) \geqslant \Omega(n^{1/2})$, where μ is some measure on inputs and $\varepsilon > 0$ is sufficiently small. This implies $C_\varepsilon(\mathrm{DIS}_n) \geqslant \Omega(n^{1/2})$ for any $\varepsilon < \frac{1}{2}$. The measure μ in [1] is a *product* measure that is the product of a measure on columns and a measure on rows. In comparison with the lower bound it was also proved in [1] that $D_\varepsilon(\mathrm{DIS}_n \mid \mu) \leqslant O(n^{1/2} \log n)$ for any product measure μ and arbitrary $\varepsilon > 0$. Then Kalyanasundaram and Schnitger [2] established the best possible lower bound $C_\varepsilon(\mathrm{DIS}_n) \geqslant \Omega(n)$ ($\varepsilon < 1/2$) for the ε-error probabilistic communication complexity of "disjointness".

Probably all lower bounds for $C_\varepsilon(A)$ known prior to the paper [2] were actually lower bounds for the distributional complexity $D_\varepsilon(A \mid \mu)$ with some suitable measure μ. But the proof in [2] involves complicated arguments related to the Kolmogorov complexity and this results in the fact that the measure μ implicitly "meant" in the proof depends on the protocol given by "the adversary".

The aim of this note is to show that the "random coupling" arguments of Kalyanasundaram and Schnitger can be carried over to yield the lower bound $D_\varepsilon(\mathrm{DIS}_n \mid \mu) \geqslant \Omega(n)$ for a very simple measure μ described below (this does *not* contradict the result from [1] since our μ is *not* a product measure). The proof involves only classical probabilistic arguments and does not appeal to the Kolmogorov complexity.

2. The result

We will identify throughout binary predicates and their characteristic 0–1 matrices. Given a predicate $A(x, y)$ ($x \in X$, $y \in Y$), *the* ε*-error distributional complexity* $D_\varepsilon(A \mid \mu)$ *under a probabilistic measure* μ *on inputs* (i.e., on $X \times Y$) is the minimal possible length of a deterministic communication protocol which, given the random input (x, y) according to the measure μ, outputs a_{xy} with probability at least $1 - \varepsilon$ [4, 1]. Fix the notation DIS_n for the so-called *disjointness matrix* DIS_n over $X := Y := \mathscr{P}([n])$ given by $(\mathrm{DIS}_n)_{xy} := 1$ iff $x \cap y = \emptyset$. Let (x_0, y_0) [(x_1, y_1)] be the random input according to the uniform distribution on $\{(x, y) \mid |x| = |y| = \lfloor n/4 \rfloor, |x \cap y| = \emptyset\}$ [$\{(x, y) \mid |x| = |y| = \lfloor n/4 \rfloor, |x \cap y| = 1\}$, respectively]. Let (x, y) be taken with probability $\frac{3}{4}$ as (x_0, y_0) and with probability $\frac{1}{4}$ as (x_1, y_1). Denote by μ the measure corresponding to (x, y). The main result of this note is the following theorem.

Theorem. $D_\varepsilon(\mathrm{DIS}_n \mid \mu) \geqslant \Omega(n)$ *for any sufficiently small* $\varepsilon > 0$.

Proof. Note that (x_ν, y_ν) $(\nu \in \{0, 1\})$ is just the input (x, y) under the condition $|x \cap y| = \nu$. Because of $\mathbf{P}[|x \cap y| = \nu] \geqslant \Omega(1)$ the theorem follows from the following statement (cf. [4, 1]).

Main Lemma. *For any* $\bar{X}, \bar{Y} \subseteq \mathscr{P}([n])$,

$$\mathbf{P}[(x_1, y_1) \in \bar{X} \times \bar{Y}] \geqslant \Omega(\mathbf{P}[(x_0, y_0) \in \bar{X} \times \bar{Y}]) - 2^{-\Omega(n)}.$$

Proof of main lemma. We may assume $n = 4m - 1$ and, therefore, $|x| = |y| = m$. First we need a somewhat exotic way of generating random inputs (x_0, y_0) and (x_1, y_1). Namely, let $t := (z_x, z_y, \{i\})$ be the random partition of $[n]$ into three sets of cardinalities $2m-1$, $2m-1$ and 1, respectively. Let x be the random member of $[z_x \cup \{i\}]^m$, y be the random member of $[z_y \cup \{i\}]^m$ (x and y are assumed to be independent). Let (x_0, y_0) be this (x, y) under the condition $i \notin x$, $i \notin y$ and (x_1, y_1) be (x, y) under the condition $i \in x$, $i \in y$. Note that our construction is invariant under the action of the symmetric group S_n; therefore, we obtain in this way the random inputs corresponding to the *uniform* distributions on $\{(x, y) \mid |x| = |y| = m, |x \cap y| = \emptyset\}$ and $\{(x, y) \mid |x| = |y| = m, |x \cap y| = 1\}$, i.e., exactly (x_0, y_0) and (x_1, y_1) used in the definition of μ (it is also easy to see that (x, y) coincides with our main distribution but we will not need this fact in what follows).

Given a partition $t = (z_x, z_y, \{i\})$, set

$$p_x(t) := \mathbf{P}[x \in \bar{X} \mid (z_x, z_y, \{i\}) = t],$$

$$p_y(t) := \mathbf{P}[y \in \bar{Y} \mid (z_x, z_y, \{i\}) = t],$$

$$p_{x,0}(t) := \mathbf{P}[x_0 \in \bar{X} \mid (z_x, z_y, \{i\}) = t],$$

$$p_{x,1}(t) := \mathbf{P}[x_1 \in \bar{X} \mid (z_x, z_y, \{i\}) = t],$$

$$p_{y,0}(t) := \mathbf{P}[y_0 \in \bar{Y} \mid (z_x, z_y, \{i\}) = t],$$

$$p_{y,1}(t) := \mathbf{P}[y_1 \in \bar{Y} \mid (z_x, z_y, \{i\}) = t].$$

Then

$$\mathbf{P}[(x_0, y_0) \in \bar{X} \times \bar{Y}] = \mathbf{E}[p_{x,0}(t) \cdot p_{y,0}(t)],$$

$$\mathbf{P}[(x_1, y_1) \in \bar{X} \times \bar{Y}] = \mathbf{E}[p_{x,1}(t) \cdot p_{y,1}(t)].$$

We now collect some easy facts about these random variables.

Fact 1. $p_x(t) = \frac{1}{2}(p_{x,0}(t) + p_{x,1}(t))$, $p_y(t) = \frac{1}{2}(p_{y,0}(t) + p_{y,1}(t))$.

Proof. The result follows from the observation $\mathbf{P}[i \in x \mid (z_x, z_y, \{i\}) = t] = \mathbf{P}[i \in y \mid (z_x, z_y, \{i\}) = t] = \frac{1}{2}$. □

Fact 2. *$p_x(z_x, z_y, \{i\})$ and $p_{y,0}(z_x, z_y, \{i\})$ depend only on z_y. $p_y(z_x, z_y, \{i\})$ and $p_{x,0}(z_x, z_y, \{i\})$ depend only on z_x.*

Set $\varepsilon := 0.01$. Let us say that t is *x-bad* if

$$p_{x,1}(t) < \tfrac{1}{3} p_{x,0}(t) - 2^{-\varepsilon n} \tag{1}$$

and t is *y-bad* if

$$p_{y,1}(t) < \tfrac{1}{3} p_{y,0}(t) - 2^{-\varepsilon n}.$$

t is *bad* if it is either x-bad or y-bad.

Claim 3. *For any $z_x, z_y \in [n]^{2m-1}$, $\mathbf{P}[(z_x, z_y, \{i\})$ is x-bad $\mid z_y = z_y] < \frac{1}{5}$ and $\mathbf{P}[(z_x, z_y, \{i\})$ is y-bad $\mid z_x = z_x] < \frac{1}{5}$.*

Proof. By symmetry, it is sufficient to prove the first inequality. By Fact 2, having fixed z_y forces $p_x(z_x, z_y, \{i\})$ to be constant. Denote $p_x(z_x, z_y, \{i\})$ by p_x. If $p_x < 2^{-\varepsilon n}$ then, by (1) and $p_{x,0}(t) \leqslant 2p_x(t)$ (see Fact 1), $\mathbf{P}[(z_x, z_y, \{i\})$ is x-bad $\mid z_y = z_y] = 0$ and we are done. So, assume

$$p_x \geqslant 2^{-\varepsilon n}. \tag{2}$$

Denote $\bar{X} \cap [\mathrm{co} - z_y]^m$ by S. Then $p_x = |S| / \binom{2m}{m}$; $p_{x,1}(z_x, z_y, \{i\}) = 2p_x \mathbf{P}[i \in s]$; $p_{x,0}(z_x, z_y, \{i\}) = 2p_x \mathbf{P}[i \notin s]$, where s is the random member of S. So, if $(z_x, z_y, \{i\})$ is x-bad then, by (1),

$$\mathbf{P}[i \in s] \leqslant \tfrac{1}{3} \mathbf{P}[i \notin s], \tag{3}$$

i.e., $\mathbf{P}[i \in s] \leqslant \frac{1}{4}$. On the other hand, $s = (s_1, s_2, \ldots, s_{2m})$, where $s_1, s_2, \ldots, s_{2m}$ are the characteristic functions of events $i_1 \in s$, $i_2 \in s, \ldots, i_{2m} \in s$ ($\{i_1, i_2, \ldots, i_{2m}\} = \mathrm{co} - z_y$). Assume, contrary to the statement of the claim, that $\mathbf{P}[(z_x, z_y, \{i\})$ is x-bad $\mid z_y = z_y] \geqslant \frac{1}{5}$ and, hence, (3) holds for at least $2m/5$ values of $i \in \mathrm{co} - z_y$. Then, counting the entropy, we get

$$m(2 - 4\varepsilon - o(1)) \leqslant H(s) \text{ (by (2))} \leqslant \sum_{i=1}^{2m} H(s_i) \leqslant 8m/5 + 2m/5 \cdot H(1/4) \leqslant 1.93m,$$

a contradiction. □

Let us denote by $\chi_x(t)$ [$\chi_y(t)$, $\chi(t)$] the indicator of the event "t is x-bad" [y-bad, bad].

Claim 4. $\mathbf{E}[p_{x,0}(t) p_{y,0}(t) \chi(t)] \leqslant \frac{4}{5} \mathbf{E}[p_{x,0}(t) p_{y,0}(t)]$.

Proof. Because $\chi(t) \leqslant \chi_x(t) + \chi_y(t)$, it is sufficient to prove that $\mathbf{E}[p_{x,0}(t)p_{y,0}(t)\chi_x(t)] \leqslant \frac{2}{5}\mathbf{E}[p_{x,0}(t)p_{y,0}(t)]$. Let us fix z_y and prove that

$$\mathbf{E}[p_{x,0}(z_x, z_y, \{i\})p_{y,0}(z_x, z_y, \{i\})\chi_x(z_x, z_y, \{i\}) \mid z_y = z_y]$$

$$\leqslant \tfrac{2}{5}\mathbf{E}[p_{x,0}(z_x, z_y, \{i\})p_{y,0}(z_x, z_y, \{i\}) \mid z_y = z_y].$$

Note that $p_{y,0}(z_x, z_y, \{i\})$ and $p_x(z_x, z_y, \{i\})$ are constant under the condition $z_y = z_y$ (see Fact 2). Denote them by $p_{y,0}$ and p_x. It is also clear that $\mathbf{E}[p_{x,0}(z_x, z_y, \{i\}) \mid z_y = z_y] = p_x$ since this expectation is just $\mathbf{P}[x_0 \in X \mid z_y = z_y]$ and x_0 under the only condition $z_y = z_y$ takes all values from $[\mathrm{co} - z_y]^m$ with the same probability $\binom{2m}{m}^{-1}$, i.e., coincides with x under the same condition. Therefore,

$$\mathbf{E}[p_{x,0}(z_x, z_y, \{i\})p_{y,0}(z_x, z_y, \{i\})\chi_x(z_x, z_y, \{i\}) \mid z_y = z_y]$$

$$= p_{y,0}\mathbf{E}[p_{x,0}(z_x, z_y, \{i\})\chi_x(z_x, z_y, \{i\}) \mid z_y = z_y]$$

$$\leqslant 2p_{y,0}p_x\mathbf{E}[\chi_x(z_x, z_y, \{i\}) \mid z_y = z_y] \quad \text{(by Fact 1)}$$

$$\leqslant \tfrac{2}{5}p_{y,0}p_x \quad \text{(by Claim 3)}$$

$$= \tfrac{2}{5}p_{y,0}\mathbf{E}[p_{x,0}(z_x, z_y, \{i\}) \mid z_y = z_y]$$

$$= \tfrac{2}{5}\mathbf{E}[p_{x,0}(z_x, z_y, \{i\})p_{y,0}(z_x, z_y, \{i\}) \mid z_y = z_y]. \qquad \square$$

Proof of main lemma (*conclusion*). Now the proof of the main lemma is completed by the easy computation

$$\mathbf{P}[(x_1, y_1) \in \bar{X} \times \bar{Y}] = \mathbf{E}[p_{x,1}(t) \cdot p_{y,1}(t)] \geqslant \mathbf{E}[p_{x,1}(t) \cdot p_{y,1}(t) \cdot (1 - \chi(t)]$$

$$\geqslant \mathbf{E}[(\tfrac{1}{3}p_{x,0}(t) - 2^{-\varepsilon n}) \cdot (\tfrac{1}{3}p_{y,0}(t) - 2^{-\varepsilon n}) \cdot (1 - \chi(t)] \quad \text{(by (1))}$$

$$\geqslant \Omega(\mathbf{E}[p_{x,0}(t) \cdot p_{y,0}(t) \cdot (1 - \chi(t)]) - 2^{-\Omega(n)}$$

$$\geqslant \Omega(\mathbf{E}[p_{x,0}(t) \cdot p_{y,0}(t)]) - 2^{-\Omega(n)} \quad \text{(by Claim 4)}$$

$$= \Omega(\mathbf{P}[(x_0, y_0) \in \bar{X} \times \bar{Y}]) - 2^{-\Omega(n)}. \qquad \square$$

Acknowledgment

I am thankful to Laci Babai and Nathan Linial for valuable remarks.

References

[1] L. Babai, P. Frankl and J. Simon, Complexity classes in communication complexity theory, in: *Proc. 27th IEEE FOCS* (1986) 337–347.
[2] B. Kalyanasundaram and G. Schnitger, The probabilistic communication complexity of set intersection, in: *Proc. 2nd Ann. Conf. on Structure in Complexity Theory* (1987) 41–49.
[3] A.C. Yao, Some complexity questions related to distributed computing, in: *Proc. 11th ACM STOC* (1979) 209–213.
[4] A.C. Yao, Lower bounds by probabilistic arguments, in: *Proc. 24th IEEE FOCS* (1983) 420–428.