Communication complexity of inevitable intersection

Dmitry Gavinsky*

May 21, 2017

Abstract

Many known methods for analysing the communication complexity of unstructured search are based on the hardness of the set disjointness problem. Such techniques may "hide" important aspects of the original problem. Intuitively, search is a much simpler task: while set disjointness is hard even for NP, successful search necessarily results in a short witness, which makes it easy for NP. Accordingly, the possibility to deduce hardness of search problems from that of set disjointness can be viewed as an artefact of specific definitions.

We construct a natural variation of the intersection-search problem, where the input comes from a product distribution, and nevertheless, every pair of input subsets share at least one element. We call this problem *inevitable intersection*, its analysis seems to require a new, more subtle approach – in particular, not relying on the hardness of set disjointness.

To prove a lower bound on the communication complexity of inevitable intersection, we identify such properties of large rectangles that make it impossible to partition the input matrix into a small number of "nearly-monochromatic" rectangles (the analysis is "non-local": it addresses not the existence of a large rectangle of low discrepancy, but the possibility to partition the whole matrix into such rectangles). We believe that this technique provides new insight into the combinatorial structure of "search-like" communication problems.

1 Introduction

Unstructured search is a very basic computational paradigm and its natural example in the context of communication complexity is the *set disjointness problem*, sometimes called the *intersection problem*. Here every player receives a subset of $[n]$ and the goal is to decide whether the intersection of these subsets is empty.[1]

The set disjointness problem is hard in most models of communication and understanding its complexity often comes down to proving a lower bound (e.g., see [BFS86, KS87, Raz92]). Virtually all known lower bound methods are based on the *hardness of witnessing non-intersection of the input sets*: Informally, while it is easy to prove that the given subsets

*Institute of Mathematics, Academy of Sciences, Žitna 25, Praha 1, Czech Republic. Partially funded by the grant P202/12/G061 of GA ČR and by RVO: 67985840. Part of this work was done while visiting the Centre for Quantum Technologies at the National University of Singapore, and was partially supported by the National Research Foundation, Prime Minister's Office, Singapore and the Ministry of Education, Singapore under the Research Centres of Excellence programme and by Grant No. MOE2012-T3-1-009.

[1] In the relational "search version" of the problem the goal is to find an element from the intersection when it is not empty – usually the two versions are nearly-equivalent computationally.

overlap (e.g., by pointing to an element from the intersection), it can be rather difficult to certify that the given subsets are disjoint; formally, disjointness is known to be hard even for the non-deterministic communication class NP. Due to the fact that a communication transcript can be used as a "correctness certificate" of the answer produced by a valid protocol, hardness of witnessing non-intersection implies a lower bound on the communication complexity of the set disjointness problem and a number of its modifications.

Many lower bounds on the complexity of unstructured search that are based on the hardness of witnessing non-intersection can be viewed as addressing the question "*What happens if the set of target elements is empty?*" – these approaches usually identify an "embedded instance" of the set disjointness problem and use it to argue hardness of the problem under consideration. These approaches often provide deep insight into the structure of the analysed problem; nevertheless, we feel that the "origin of hardness" of a search problem can be more subtle than that of disjointness: in particular, while disjointness remains hard even for NP, a successful search inherently produces a short witness, which makes it easy for NP. Accordingly, the possibility to deduce hardness of search problems from that of disjointness may be just an artefact of the definitions.

The most standard model of communication complexity consists of two players – Alice and Bob – who are allowed to exchange messages in order to find the answer. The complexity of a given protocol is defined as the maximum possible number of bits exchanged before producing the answer, when each player's input is of length at most n. The players are allowed to use random bits, the answer must be correct with non-trivial constant probability – say, at least p when the input distribution is assumed to be such that no single answer can be correct with probability more than $p - \Omega(1)$. We denote this model by $\mathcal{R}_p$ or simply by $\mathcal{R}$. The complexity of a problem in $\mathcal{R}$ is the minimum complexity of a valid protocol for solving it. Denote by $\mathcal{R}_{\mu,p}$ the distributional version of $\mathcal{R}_p$, where the input distribution is assumed to be μ.

To model the unstructured search in $\mathcal{R}$, we define a version of the intersection problem that we call *inevitable intersection ($\mathcal{II}$)*, where the input comes from the uniform distribution (i.e., there is no "semantic promise"), and nevertheless every possible pair of input subsets share at least one element:

> Let $A, B \subseteq \{0,1\}^n$ be such that
> $$\forall a \in A, b \in B : \; a \cap b \neq \emptyset,$$
> where a binary string is viewed as the set of its coordinates with value "1". Alice receives $a \in A$, Bob receives $b \in B$ and they have to output some $i \in a \cap b$.

In particular, the communication complexity of $\mathcal{II}$ cannot be analysed via asking "*What happens to the non-intersecting input pairs?*". Furthermore, the input matrix can be *fully covered* by a very small number of "monochromatic" rectangles.[2]

Our definition of $\mathcal{II}$ is analogous to the well-known notion of *monotone Karchmer-Wigderson games* due to [RW92]. Nevertheless, considering the following special case of this problem has led us to developing a new lower bound technique.

Question 1. *What is the $\mathcal{R}$-complexity of $\mathcal{II}_{A,B}$, where A and B are uniformly-random subsets of $\binom{[n]}{n^{3/5}}$ of size $2^{\sqrt[5]{n}/5}$?*

[2]Compare it to the case of the set disjointness problem, where the intersecting input pairs are also "easy to cover", unlike the non-intersecting ones: There is no "large rectangle that mostly contains disjoint pairs", and this fact underlies most or all known lower bound proofs.

The main result of this work is a (tight) linear lower bound on the communication complexity of the above problem (cf. Corollary 1). We are not aware of a previously-known technique that would give a non-trivial lower bound here. Besides being able to capture the complexity of $\mathcal{II}_{A,B}$ in $\mathcal{R}$, the technique proposed in this work provides an alternative lower bound proofs for many other cases of "search-like" problems, including set disjointness.

2 Preliminaries

We will often view the elements of $\{0,1\}^n$ as subsets of $[n]$ and vice versa. For $x \in \{0,1\}^n$, let $x(i)$ denote the i'th bit of x, and for $S \subseteq [n]$, let $x(S) \in \{0,1\}^{|S|}$ be the "projection" of x to the coordinates in S.

For a discrete set A, we denote by $\mathcal{U}_A$ the uniform distribution over its elements. Sometimes (e.g., in subscripts) we will write "$X \in A$" instead of "$X \sim \mathcal{U}_A$".

We let log denote the base-2 logarithm; at times, we will write $\exp(\cdot)$ instead of $e^{\cdot}$ to avoid "superscript congestion".

We will use the Chernoff bound in the following form (cf. [DM05]).

Claim 1 (*Chernoff bound*)**.** *Let $X_1, \ldots, X_n$ be mutually independent random variables taking values in $[0,1]$ and $\mathbf{E}[X_i] \equiv \mu$. Then for any $\Delta > 0$,*

$$\mathbf{Pr}\left[\frac{1}{n} \cdot \sum_{i=1}^{n} X_i \geq \mu + \Delta\right] \leq e^{\frac{-n\Delta^2}{2\mu+\Delta}}$$

and

$$\mathbf{Pr}\left[\frac{1}{n} \cdot \sum_{i=1}^{n} X_i \leq \mu - \Delta\right] \leq e^{\frac{-n\Delta^2}{2\mu}}.$$

The following tail bound can be viewed as a variation on Markov's inequality.

Lemma 1. *Let X be a random variable taking values in $[a,b]$, then for any $\Delta > 0$,*

$$\mathbf{Pr}\left[X < \mathbf{E}[X] + \Delta\right],\ \mathbf{Pr}\left[X > \mathbf{E}[X] - \Delta\right] \ \geq \ \min\left\{\frac{\Delta}{b-a}, 1\right\}.$$

Proof. Let $\lambda \stackrel{\text{def}}{=} \mathbf{Pr}[X < \mathbf{E}[X] + \Delta]$, then

$$\mathbf{E}[X] \geq (\mathbf{E}[X] + \Delta) \cdot (1 - \lambda) + a \cdot \lambda,$$

and therefore,

$$\lambda \geq \frac{\Delta}{(\mathbf{E}[X] + \Delta) - a}.$$

If $\mathbf{E}[X] + \Delta \leq b$, then $\lambda \geq \frac{\Delta}{b-a}$; otherwise, $\lambda = 1$ trivially.

The case of $\mathbf{Pr}[X > \mathbf{E}[X] - \Delta]$ is similar. ■

Communication complexity

For a problem $\mathcal{S}$, let $\mathcal{R}_p(\mathcal{S})$ denote its complexity in the model $\mathcal{R}_p$. Unless stated otherwise, we will call $\mathcal{R}_p(\mathcal{S})$ *the communication complexity of* $\mathcal{S}$.

One of the most studied communication complexity problems is set disjointness: Alice receives x and Bob receives y as input, and they have to decide whether the two sets overlap (note that this is a function: for every input pair, there is exactly one correct answer).

Definition 1 (*Set disjointness problem, Disj*)**.** *For* $x, y \subseteq [n]$, *let*

$$\mathit{Disj}(x, y) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } x \cap y = \emptyset \\ 0 & \text{otherwise} \end{cases}.$$

In this work we study the following problem.

Definition 2 (*Inevitable intersection problem,* $\mathcal{II}_{A,B}$)**.** *For* $A, B \subseteq \{0,1\}^n$ *such that*

$$\forall a \in A, b \in B: \ a \cap b \neq \emptyset,$$

let

$$\mathcal{II}_{A,B} \stackrel{\text{def}}{=} \left\{(a, b, i) \in A \times B \times [n] \middle| i \in a \cap b\right\}.$$

Informally, when Alice receives a and Bob receives b as their input to $\mathcal{II}_{A,B}$, a correct answer is any $i \in a \cap b$ (a correct answer does not have to be unique, so this is a relational problem). Note that $\mathcal{II}_{A,B}$ can be viewed as a search version of *Disj* with an additional constraint that $\forall a \in A, b \in B : a \cap b \neq \emptyset$. The most important for us is the "syntactic nature" of this constraint: an instance of $\mathcal{II}_{A,B}$ is defined by the choice of $A, B \subseteq \{0,1\}^n$ for every $n \in \mathbb{N}$, and only those instances are valid where "$a \cap b \neq \emptyset$" is a *tautology.*

3 Our argument

A *deterministic* 2-party communication protocol of length c defines a partition of the input matrix into at most 2^c "same-answer" rectangles (the protocol is able to distinguish only between input pairs coming from different rectangles). At the same time, if there exists an efficient *randomised* protocol, then there must exist a reasonably-accurate deterministic protocol for every input distribution μ. A "typical" rectangle defined by such a protocol must be relatively large (otherwise the union of all rectangles would be too small to cover the whole input matrix) and nearly-monochromatic (otherwise the protocol would not be sufficiently accurate).

In the case of *Disj*, one can find such input distribution μ that no large (with respect to μ) rectangle would consist mostly of non-intersecting input pairs, and at the same time, the probability of a pair of sets $(X, Y) \sim \mu$ to not intersect would be close to $1/2$. The above reasoning implies that if a short randomised protocol for *Disj* were possible, the non-intersecting input pairs that are often produced by μ would have "no rectangle to go", thus contradicting the assumption and leading to the desired lower bound on the randomised communication complexity of *Disj*.

In the case of $\mathcal{II}_{A,B}$, non-emptiness of $a \cap b$ holds for every possible input pair $(a,b) \in A \times B$, so one cannot meaningfully ask "*Where do non-intersecting input pairs go?*": If $a \notin A$ or $b \notin B$, at least one of the players would immediately notice the promise violation.

To analyse the communication complexity of $\mathcal{II}_{A,B}$, we will use the following approach. Consider a deterministic protocol of complexity c that solves $\mathcal{II}_{A,B}$ with respect to the uniform (over $A \times B$) input distribution $\mathcal{U}$ with error at most $1/2$. This protocol corresponds to a partition of $A \times B$ into at most 2^c rectangles that are "labelled" by protocol's answers, such that $(X,Y) \sim \mathcal{U}$ belongs to a rectangle labelled by some $i \in X \cap Y$ with probability at least $1/2$.

We would like to get a lower bond of the form $c \in n^{\Omega(1)}$ – that is, we want to show that a *partition* of $A \times B$ with properties as described above must have size $\exp\left(n^{\Omega(1)}\right)$. Note that there always exists a *cover* of $A \times B$ by n perfectly-monochromatic rectangles:

$$r_i \stackrel{\text{def}}{=} \left\{(a,b) \in A \times B \middle| a(i) = b(i) = 1\right\}, \tag{1}$$

where the label of r_i is "i". So, we are looking for a "property of large rectangles" that would obstruct combining them into a partition of $A \times B$, but not into a cover of it.

Let us consider a partition R of $A \times B$ into rectangles. For all $i \in [n]$ and $r \in R$, let

$$p(r,i) \stackrel{\text{def}}{=} \mathop{\mathbf{Pr}}_{(X,Y)\sim\mathcal{U}}\left[X(i) = Y(i) = 1 \middle| (X,Y) \in r\right].$$

For a "typical" $r_0 \in R$ that is labelled by "i_0", we expect $p(r_0, i_0)$ to be high, but what about the rest of i-s? We will see (cf. Lemma 2) that if r_0 is large enough, then, informally speaking, $p(r_0, i)$ cannot be too different from the global (unconditional) expectation

$$p_i \stackrel{\text{def}}{=} \mathop{\mathbf{Pr}}_{\mathcal{U}}\left[X(i) = Y(i) = 1\right]$$

for too many values of $i \in [n]$ (this is the case for the rectangles defined in (1), for instance). Intuitively, a rectangle "pays" in terms of the entropy of its uniformly-random element for making some of its bits "biased" (i.e., making $p(r_0, i)$ significantly different from p_i).

This property of large rectangles is enough to prove a strong lower bound on the cardinality of R (cf. Theorem 1). To understand how, let us consider the following extreme situation: the rectangles in R are either "small" or "large", and for a large $r_0 \in R$ labelled by i_0 it holds that

$$p(r_0, i) \approx \begin{cases} \frac{1}{2} & \text{if } i = i_0 \\ p_i & \text{otherwise} \end{cases}.$$

Let us also assume $p_i \equiv \left(\frac{k}{n}\right)^2$ for some $k \ll n$ (this is almost true almost always if A and B are sufficiently large random subsets of $\binom{[n]}{k}$, which will be the case of interest to us). If, towards contradiction, R is small then a significant fraction of its rectangles must be large – let us again take the extreme and assume that all R's members are large rectangles. Let $i_1 \in [n]$ be such that at least $1/n$-fraction of $A \times B$ belongs to "i_1"-labelled rectangles from R, and let us use the above assumptions to estimate p_{i_1}: On the one hand, with some probability $q \geq 1/n$, a uniformly-random input (X,Y) belongs to a rectangle $r \in R$ labelled by "i_1" – by our assumptions, this event "contributes" roughly $p(r, i_1) \cdot q = \frac{1}{2} \cdot q$ to the probability that

$X(i) = Y(i) = 1$. On the other hand, with probability $1-q$ the input belongs to a rectangle labelled differently, and we have assumed that in that case $X(i_1) = Y(i_1) = 1$ with probability roughly p_{i_1}, so this event "contributes" about $p_{i_1} \cdot (1-q)$. Therefore,

$$p_{i_1} \approx \frac{1}{2} \cdot q + p_{i_1} \cdot (1-q) \implies p_{i_1} \approx \frac{1}{2},$$

which contradicts our assumption that $p_i \equiv \left(\frac{k}{n}\right)^2 \ll 1$.

Our argument can be summarised like this: On the one hand, a large nearly-monochromatic rectangle in R "causes" a noticeable deviation (increase) of the global probability that its "label coordinate" belongs to $X \cap Y$; on the other, large rectangles cannot efficiently "absorb" the deviations caused by other large rectangles – therefore, there must be many small rectangles in R, and the partition itself must be large.

4 The communication complexity of $\mathcal{II}$

We start by proving a lemma that limits "witnessing" against coordinate-wise intersections by a large input rectangle.

Lemma 2. *Let $1 \le k < \frac{n}{2}$ and $1 \le M' \le M \le \frac{1}{2}\binom{n}{k}^{1/2}$, such that*

$$\log\left(\frac{M}{M'}\right) \le \frac{\log M}{3} - 5\log n.$$

Then for

$$\Delta = 51 \cdot \frac{k^{3/2}}{n} \cdot \sqrt{\log\left(\frac{M}{M'}\right) + \log n}$$

it holds that

$$\max_{\substack{A' \subseteq A \\ B' \subseteq B \\ |A'|,|B'| \ge M' \\ T \subseteq [n]}} \left\{ \sum_{i \in T} \left(\left(\frac{k}{n}\right)^2 - \mathop{\mathbf{Pr}}_{(X,Y) \Subset A' \times B'} [X(i) = Y(i) = 1] \right) \right\} < \Delta$$

with probability higher than $1 - \exp(n - M^{1/3})$ when A and B are uniformly-random subsets of $\binom{[n]}{k}$ of size M.

Informally, the lemma states that almost always with respect to A and B, membership of the input pair (X, Y) in a large rectangle $A' \times B' \subseteq A \times B$ cannot significantly decrease the probability that $X(i) = Y(i) = 1$ for many $i \in [n]$ – note that this probability equals $\left(\frac{k}{n}\right)^2$ when $X, Y \Subset \binom{[n]}{k}$. This lemma will be the core technical tool of the lower bound proof for $\mathcal{II}_{A,B}$.

Proof of Lemma 2. Consider some $A' \subseteq A \subseteq \binom{[n]}{k}$ and let $p_i \stackrel{\text{def}}{=} \mathbf{Pr}_{X \Subset A'}[X(i) = 1]$, $\alpha \stackrel{\text{def}}{=} \sum_{i=1}^n \left|p_i - \frac{k}{n}\right|$ and $S \stackrel{\text{def}}{=} \left\{ i \in [n] \middle| p_i < \frac{k}{n} \right\}$. Let us see that if α is big enough, then A' contains a non-negligible fraction of bit strings, whose projection to S has “unnaturally low” Hamming weight. As $\sum p_i = k$ by assumption, $\sum_{i\in S}\left(\frac{k}{n} - p_i\right) = \frac{\alpha}{2}$ and

$$\mathop{\mathbf{E}}_{X \Subset A'}\left[|X(S)|\right] = \frac{k\cdot|S|}{n} - \frac{\alpha}{2}.$$

Therefore by Lemma 1,

$$\mathop{\mathbf{Pr}}_{X \Subset A'}\left[|X(S)| \le \frac{k\cdot|S|}{n} - \frac{\alpha}{4}\right] \ge \frac{\alpha}{4}\Big/\left(\max_{x\in A'}\{|x(S)|\} - \min_{x\in A'}\{|x(S)|\}\right) \ge \frac{\alpha}{4n}.$$

As $A' \subseteq A$, the set A itself must contain enough elements, whose projection to S has low Hamming weight:

$$M' \le |A'| \le \frac{4n}{\alpha}\cdot\left|\left\{a\in A \,\middle|\, |a(S)| \le \frac{k\cdot|S|}{n} - \frac{\alpha}{4}\right\}\right|, \tag{2}$$

and the same holds for B.

Now fix $B' \subseteq B \subseteq \binom{[n]}{k}$ and let $q_i \overset{\text{def}}{=} \mathbf{Pr}_{Y\Subset B'}\left[Y(i) = 1\right]$ and $\beta \overset{\text{def}}{=} \sum_{i=1}^{n}\left|q_i - \frac{k}{n}\right|$. Note that the value of

$$\sum_{i\in T}\left(\left(\frac{k}{n}\right)^2 - \mathop{\mathbf{Pr}}_{(X,Y)\Subset A'\times B'}\left[X(i) = Y(i) = 1\right]\right)$$

is maximised by $T = \left\{i\in[n] \,\middle|\, \mathbf{Pr}_{(X,Y)\Subset A'\times B'}\left[X(i) = Y(i) = 1\right] < \left(\frac{k}{n}\right)^2\right\}$, so we fix T this way without loss of generality.

For all $i \in T$ it holds that

$$\begin{aligned}
\mathbf{Pr}\left[X(i) = Y(i) = 1\right] = p_i\cdot q_i &= \left(\frac{k}{n} + p_i - \frac{k}{n}\right)\left(\frac{k}{n} + q_i - \frac{k}{n}\right) \\
&\ge \left(\frac{k}{n}\right)^2 - \frac{k}{n}\cdot\left(\left|p_i - \frac{k}{n}\right| + \left|q_i - \frac{k}{n}\right|\right) - \left|p_i - \frac{k}{n}\right|\cdot\left|q_i - \frac{k}{n}\right| \\
&\ge \left(\frac{k}{n}\right)^2 - \frac{2k}{n}\cdot\left(\left|p_i - \frac{k}{n}\right| + \left|q_i - \frac{k}{n}\right|\right),
\end{aligned}$$

where the last inequality follows from $i\in T \implies p_i < \frac{k}{n}$ or $q_i < \frac{k}{n}$. Accordingly,

$$\sum_{i\in T}\left(\left(\frac{k}{n}\right)^2 - \mathop{\mathbf{Pr}}_{A'\times B'}\left[X(i) = Y(i) = 1\right]\right) \le \frac{2k}{n}\cdot\sum_{i\in T}\left(\left|p_i - \frac{k}{n}\right| + \left|q_i - \frac{k}{n}\right|\right) \le \frac{2k}{n}\cdot(\alpha+\beta).$$

Therefore, if

$$\max_{\substack{A'\subseteq A\\ B'\subseteq B\\ |A'|,|B'|\ge M'\\ T\subseteq[n]}}\left\{\sum_{i\in T}\left(\left(\frac{k}{n}\right)^2 - \mathop{\mathbf{Pr}}_{(X,Y)\Subset A'\times B'}\left[X(i) = Y(i) = 1\right]\right)\right\} \ge \Delta$$

then $\alpha \ge \frac{n\Delta}{4k}$ or $\beta \ge \frac{n\Delta}{4k}$.

Let us see what happens if α is non-negligible. From (2), for some $S \subseteq [n]$:

$$M' \leq \frac{4n}{\alpha} \cdot \left| \left\{ a \in A \middle| |a(S)| \leq \frac{k \cdot |S|}{n} - \frac{\alpha}{4} \right\} \right|,$$

which can be reformulated as

$$\frac{M'}{M} \leq \frac{4n}{\alpha} \cdot \mathop{\mathbf{Pr}}_{X \Subset A} \left[|X(S)| \leq \frac{k \cdot |S|}{n} - \frac{\alpha}{4} \right]. \tag{3}$$

Let $e_S \stackrel{\text{def}}{=} \left[|X(S)| \leq \frac{k \cdot |S|}{n} - \frac{\alpha}{4} \right]$. For a fixed S, this event depends only on the value taken by X. First we analyse the probability of e_S under $X \Subset \binom{[n]}{k}$. To do that (with accuracy sufficient for our needs), we note that in a sequence of n independent Bernoulli trials with individual success probability $\frac{k}{n}$ (next denoted by $\mathcal{B}_{k/n}^{\otimes n}$), exactly k "successes" are observed with probability at least $\frac{1}{n}$; moreover, the corresponding conditional distribution is coordinate-wise symmetric. Accordingly,

$$\mathop{\mathbf{Pr}}_{X \Subset \binom{[n]}{k}} [e_S] \leq n \cdot \mathop{\mathbf{Pr}}_{X \sim \mathcal{B}_{k/n}^{\otimes n}} [e_S] \leq n \cdot e^{\frac{-n^2 \alpha^2}{32k|S|^2}} \leq \exp\left(\ln n - \frac{\alpha^2}{32k} \right), \tag{4}$$

where the second inequality follows from the Chernoff bound (Claim 1), and the last one uses $|S| \leq n$.

Next we claim that the probability of e_S is unlikely to differ significantly under $X \Subset \binom{[n]}{k}$ and under $X \Subset A$ when A is a uniformly-random subset of $\binom{[n]}{k}$ of size M. Let

$$e'_S \stackrel{\text{def}}{=} \left[\mathop{\mathbf{Pr}}_{X \Subset A} [e_S] \geq \mathop{\mathbf{Pr}}_{X \Subset \binom{[n]}{k}} [e_S] + \delta \right]$$

for some $\delta < 1$ to be fixed later. For a fixed S, this event depends only on the content of A (which we now view as a random object).

If instead of choosing A as a subset of size M, we would M times select a uniformly-random element of $\binom{[n]}{k}$ and "add" it to A – possibly with repetitions – then by the assumption about M, no repetition would occur with probability more than $1/2$; conditional on that, the process would indeed generate a uniformly-random subset of size M. Let $Y = (Y_i)_{i=1}^{M}$, where Y_i-s are independent Bernoulli variables that take value "1" with probability $\mathbf{Pr}_{X \Subset \binom{[n]}{k}} [e_S]$, then

$$\mathop{\mathbf{Pr}}_{A \Subset \binom{\binom{[n]}{k}}{M}} [e'_S] \leq 2 \, \mathbf{Pr} \left[\frac{|Y|}{M} \geq \mathop{\mathbf{Pr}}_{X \Subset \binom{[n]}{k}} [e_S] + \delta \right] \leq 2 \cdot e^{\frac{-M\delta^2}{3}},$$

where the second inequality follows from the Chernoff bound (Claim 1). By the union bound and since $n \geq 3$,

$$\mathop{\mathbf{Pr}}_{A \Subset \binom{\binom{[n]}{k}}{M}} \left[\bigvee_S e'_S \right] \leq 2^{n+1} \cdot e^{\frac{-M\delta^2}{3}} < \exp\left(n - \frac{M\delta^2}{3} \right). \tag{5}$$

Now let $\delta \stackrel{\mathrm{def}}{=} \sqrt{3} \cdot M^{-1/3}$. Combining (3), (4) and (5), we conclude that if $\alpha \geq \frac{n\Delta}{4k}$, then

$$\frac{M'}{M} < \frac{4n}{\alpha} \cdot \left(\exp\left(\ln n - \frac{\alpha^2}{32k} \right) + \delta \right) \leq 8n^2 \cdot e^{-\frac{n^2\Delta^2}{512k^3}} + 14n \cdot M^{-1/3} \tag{6}$$

holds with probability greater than $1 - \exp\left(n - M^{1/3}\right)$ with respect to a uniformly-random $A \subseteq \binom{[n]}{k}$ of size M. By symmetry, the same is true if $\beta \geq \frac{n\Delta}{4k}$, and therefore true unconditionally. From (6) we conclude that

$$\frac{M'}{M} < 16n^2 \cdot e^{-\frac{n^2\Delta^2}{512k^3}}$$

or

$$\frac{M'}{M} < 28n \cdot M^{-1/3}.$$

The latter possibility would contradict the lemma assumptions, and the former implies

$$\Delta^2 < \left(\log\left(\frac{M}{M'} \right) + \log n \right) \cdot \frac{2560 \cdot k^3}{n^2}.$$

The result follows. $\blacksquare_{\textit{Lemma 2}}$

We are ready to implement the lower bound method that has been presented in Section 3.

Theorem 1. *Let $1 \leq k < \frac{n}{2}$ and $n^8 \leq M \leq \frac{1}{2} \cdot \binom{n}{k}^{1/2}$. If A and B are uniformly-random subsets of $\binom{[n]}{k}$ of size M, then*

$$\mathcal{R}_{1/2}(\mathcal{II}_{A,B}) \geq \mathcal{R}_{\mathcal{U}_{A\times B},1/2}(\mathcal{II}_{A,B}) > \min\left\{ \frac{\log M}{3} - 8\log n,\ \frac{n^2}{93636 \cdot k^3} - 4\log n \right\}$$

holds with probability higher than $1 - \exp(n - M^{1/3} + 1) - \mathbf{Pr}\left[\exists a \in A, b \in B\ : a \cap b = \emptyset\right]$.

Note that the theorem statement can be strengthened as follows: Instead of requiring that $a \cap b \neq \emptyset$ for every possible $a \in A$ and $b \in B$, we could let a uniformly-random pair from $A \times B$ have non-empty intersection with sufficiently high probability $1 - \delta$ and allow protocol error strictly higher than δ (say, looking at $\mathcal{R}_{1/4+\delta}(\mathcal{II}_{A,B})$). Since in this case a valid protocol would be "allowed" to err whenever $a \cap b = \emptyset$, all the challenges in proving a good lower bound that this work aims to address (as discussed in Sections 1 and 3) would still be present. The reason why we impose the restriction that $a \cap b \neq \emptyset$ for every possible input pair is aesthetic: we have been trying to emphasise the *syntactic* nature of the guarantee that the intersection was non-empty.

The above theorem can be applied like this:

Corollary 1. *Let $k = n^{3/5}$ and $M = 2^{\sqrt[5]{n}/5}$, then*

$$\mathcal{R}_{1/2}(\mathcal{II}_{A,B}) \geq \mathcal{R}_{\mathcal{U}_{A\times B},1/2}(\mathcal{II}_{A,B}) \geq \in \Omega\left(\sqrt[5]{n}\right)$$

holds with probability $1 - 2^{-\Omega\left(\sqrt[5]{n}\right)}$ when A and B are uniformly-random M-subsets of $\binom{[n]}{k}$.

The above lower bound is linear in the input size, which is $\log M$. Accordingly, it is tight and $\mathcal{R}_{1/2}(\mathcal{II}_{A,B}) \in \Theta(\sqrt[5]{n})$ almost always (i.e., for almost all A and B).

Proof. Note that

$$\begin{aligned}
\mathbf{Pr}\left[\exists a \in A, b \in B \ : a \cap b = \emptyset\right] &\le M^2 \cdot \mathop{\mathbf{Pr}}_{X,Y \Subset \binom{[n]}{k}} \left[X \cap Y = \emptyset\right] \\
&= M^2 \frac{\binom{n-k}{k}}{\binom{n}{k}} \le M^2 \left(\frac{n-k}{n}\right)^k \\
&\le M^2 \cdot \exp\left(-\frac{k^2}{n}\right) \le 2^{-\sqrt[5]{n}/10}
\end{aligned}$$

and apply Theorem 1. ■*Corollary 1*

Proof of Theorem 1. Let μ be the uniform distribution over $A \times B$ where $|A| = |B| = M$, and assume that $\mathcal{P}$ is a deterministic protocol of complexity c that solves $\mathcal{II}_{A,B}$ with error at most $1/2$ with respect to μ, conditional on $a \cap b \neq \emptyset$ for every $(a,b) \in A \times B$.[3] We will keep track of the events

$$\forall i \in [n]: \ \mathfrak{e}_i \stackrel{\text{def}}{=} [X(i) = Y(i) = 1].$$

Informally, we will say that a "typical" answer "i" is output by $\mathcal{P}$ with probability at least $1/n$ and conditional on the answer "i", the probability of $\mathfrak{e}_i$ is at least $1/2$. That is, "i"-labelled rectangles of $\mathcal{P}$ "boost" the probability of $\mathfrak{e}_i$ by roughly $1/n$, which must be "compensated" by lowering the conditional probability of $\mathfrak{e}_i$ in the rest of the rectangles of $\mathcal{P}$, and Lemma 2 implies that for that to happen, a "typical" rectangle must be rather small.

As $\mathcal{P}$ partitions $A \times B$ into 2^c rectangles, at least a $(1 - n^{-3})$-fraction of the input pairs from $A \times B$ belong to a rectangle with both sides of size at least $M' \stackrel{\text{def}}{=} \frac{M}{n^3 \cdot 2^c}$. Denote by R_+ the set of all such rectangles, by R_- the rest of $\mathcal{P}$'s rectangles and let $R = R_+ \cup R_-$. For every $r \in R$, let $\ell(r)$ be the "label" of the rectangle, i.e., the answer returned by $\mathcal{P}$ when $(X,Y) \in r$.

First of all, let us see that $\mathbf{E}_{(X,Y) \Subset A \times B}\left[|X \cap Y|\right]$ is unlikely to be too different from $\frac{k^2}{n}$:

$$\forall x_0 \in \binom{[n]}{k}: \ \mathop{\mathbf{Pr}}_{|B|=M}\left[\mathop{\mathbf{E}}_{Y \Subset B}\left[|x_0 \cap Y|\right] > \frac{k^2}{n} + \frac{1}{n^2}\right] \le e^{-\frac{M}{n^5}} < \exp\left(-M^{1/3}\right),$$

by the Chernoff bound (Claim 1) and the lemma assumptions. By the union bound,

$$\begin{aligned}
\mathop{\mathbf{Pr}}_{A,B}\left[\mathop{\mathbf{E}}_{(X,Y) \Subset A \times B}\left[|X \cap Y|\right] > \frac{k^2}{n} + \frac{1}{n^2}\right] &\le \mathop{\mathbf{Pr}}_{B}\left[\exists x_0 : \mathop{\mathbf{E}}_{Y \Subset B}\left[|x_0 \cap Y|\right] > \frac{k^2}{n} + \frac{1}{n^2}\right] \\
&< \exp\left(n - M^{1/3}\right).
\end{aligned}$$

For the rest of this proof, assume that $\mathbf{E}\left[|X \cap Y|\right] \le \frac{k^2}{n} + \frac{1}{n^2}$.

[3] In the rest of the proof of Theorem 1 we implicitly assume μ, unless stated otherwise.

Now we come back to the protocol $\mathcal{P}$. By the correctness assumption,

$$\sum_{r\in R}\mu(r)\cdot \mathbf{Pr}\left[e_{\ell(r)}\middle|(X,Y)\in r\right]\geq \frac{1}{2}.$$

On the other hand,

$$\sum_{r\in R}\mu(r)\cdot \sum_{i\in[n]}\mathbf{Pr}\left[e_i\middle|(X,Y)\in r\right]=\mathbf{E}\left[|X\cap Y|\right]\leq \frac{k^2}{n}+\frac{1}{n^2}.$$

Accordingly,

$$\sum_{r\in R}\mu(r)\cdot \sum_{i\neq\ell(r)}\mathbf{Pr}\left[e_i\middle|(X,Y)\in r\right]\leq \frac{k^2}{n}+\frac{1}{n^2}-\frac{1}{2}.$$

Let $\mu(R_+)\stackrel{\text{def}}{=}\sum_{r\in R_+}\mu(r)$, then $\mu(R_+)\geq 1-n^{-3}$ and

$$\sum_{r\in R_+}\frac{\mu(r)}{\mu(R_+)}\cdot \sum_{i\neq\ell(r)}\mathbf{Pr}\left[e_i\middle|(X,Y)\in r\right]\leq \left(\frac{k^2}{n}+\frac{1}{n^2}-\frac{1}{2}\right)\cdot\frac{1}{\mu(R_+)}\leq \frac{k^2}{n}+\frac{1}{n^2}+\frac{2}{n^3}-\frac{1}{2},$$

and therefore for some $r_0\in R_+$,

$$\sum_{i\neq\ell(r_0)}\mathbf{Pr}\left[e_i\middle|(X,Y)\in r_0\right]\leq \frac{k^2}{n}+\frac{1}{n^2}+\frac{2}{n^3}-\frac{1}{2},$$

which can be rewritten as

$$\sum_{i\neq\ell(r_0)}\left(\frac{k^2}{n^2}-\mathop{\mathbf{Pr}}_{(X,Y)\sim r_0}\left[e_i\right]\right)\geq \frac{1}{2}-\frac{1+k^2}{n^2}-\frac{2}{n^3}>\frac{1}{6}.$$

By Lemma 2, with probability at least $1-\exp(n-M^{1/3})$ this implies

$$\frac{k^{3/2}}{n}\cdot\sqrt{\log\left(\frac{M}{M'}\right)+\log n}>\frac{1}{306}$$

or

$$\log\left(\frac{M}{M'}\right)>\frac{\log M}{3}-5\log n,$$

where the former can be rewritten as

$$\log\left(\frac{M}{M'}\right)>\frac{n^2}{93636\cdot k^3}-\log n.$$

The result follows. ■ *Theorem 1*

Acknowledgements

I am grateful to Pavel Pudlák and Ronald de Wolf for many insightful discussions. Critical comments and suggestions that I received from Mika Göös helped me to make the informal discussion less ambiguous.

References

[BFS86] L. Babai, P. Frankl, and J. Simon. Complexity Classes in Communication Complexity Theory. *Proceedings of the 27th Annual Symposium on Foundations of Computer Science*, pages 337–347, 1986.

[DM05] E. Drukh and Y. Mansour. Concentration Bounds for Unigram Language Models. *Journal of Machine Learning Research 6*, pages 1231–1264, 2005.

[KS87] B. Kalyanasundaram and G. Schnitger. The Probabilistic Communication Complexity of Set Intersection. *Proceedings of the 2nd Annual Conference on Structure in Complexity Theory*, pages 41–49, 1987.

[Raz92] A. Razborov. On the Distributional Complexity of Disjointness. *Theoretical Computer Science 106(2)*, pages 385–390, 1992.

[RW92] R. Raz and A. Wigderson. Monotone Circuits for Matching Require Linear Depth. *Journal of the ACM 39*, pages 736–744, 1992.