The communication complexity of the inevitable intersection problem

Dmitry Gavinsky*

August 31, 2020

Abstract

Set disjointness (*Disj*) is a central problem in communication complexity. Here Alice and Bob each receive a subset of an n-element universe, and they need to decide whether their inputs intersect or not. The communication complexity of this problem is relatively well understood, and in most models, including – most famously – *interactive randomised communication with bounded error* ($\mathcal{R}$), the problem requires much communication.

In this work we were looking for a variation of *Disj*, as natural and simple as possible, for which the known lower bound methods would fail, and thus a new approach would be required in order to understand its $\mathcal{R}$-complexity. The problem that we have found is a *relational* one: each player receives a subset as input, and the goal is to find an element that belongs to both players. We call it *inevitable intersection* ($\mathcal{II}$).

The following list of its properties seem to let $\mathcal{II}$ resist the old lower bound techniques:

- the domain of $\mathcal{II}$ is $A \times B$, the product of the players' individual input spaces;
- $A \times B$ only contains intersecting pairs of subset;
- the input comes from the uniform distribution over $A \times B$;
- $A \times B$ is chosen in a randomised fashion, both A and B being uniformly-random subsets of $2^{[n]}$ of size $2^{n^{\Theta(1)}}$.

In particular, complexity analysis of $\mathcal{II}$ cannot be based on the hardness of *Disj* (as no pair in $A \times B$ is disjoint); moreover, it cannot be based on any argument based on discrepancy (including corruption, smooth discrepancy and the like), as the problem allows for a cover of $A \times B$ by n perfectly-monochromatic rectangles.

We are using an ad hoc technique to show that $\mathcal{II}$ is *ultimately hard*: it requires $\Omega(\log|A|)$ bits of interactive randomised communication. Besides its ability – apparently unique – to capture the complexity of the inevitable intersection, the new technique can also be applied to other "search-like" problems (including *Disj*), thus providing new insight into their communicational hardness.

*Institute of Mathematics, Academy of Sciences, Žitna 25, Praha 1, Czech Republic. Partially funded by the grant 19-27871X of GA ČR. 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 Singapore National Research Foundation, the Prime Minister's Office and the Ministry of Education under the Research Centres of Excellence programme under grant R 710-000-012-135.

1 Introduction

Unstructured search is a basic computational paradigm and its natural instance in the context of communication complexity is the *intersection problem*. Here each player receives a subset of $[n]$ and the goal is *to decide whether the intersection of these subsets is empty* in the *decision version*, and *to find an element from the intersection when it is not empty* in the *search version* of the problem. The decision version is exactly the same as the well-known *set disjointness problem* (*Disj*). The search version looks more demanding, but for most communication models the decision and the search version are, essentially, equivalent, as the search version can be reduced to a short series of decision instances (e.g., via binary search).

The most common in the context of communication complexity is, probably, the model of *randomised interactive communication*: there are two players, Alice and Bob, who are allowed to use random bits and exchange messages in order to find an answer with respect to the input that is divided between the two of them. 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 has length at most n. 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 $\mathcal{R}$. The complexity of a problem P in $\mathcal{R}$ is the minimum complexity of a valid protocol for solving P, and we write $\mathcal{R}(P)$ to denote it.

No efficient solution for *Disj* exists in most communication models, so understanding its complexity amounts to very little more than proving a tight lower bound. In particular, [KS87] has shown that $\mathcal{R}(Disj) \in \Omega(n)$ (see also [BFS86, Raz92]).

Virtually all known lower bound methods for *Disj* are based on the *hardness of witnessing non-intersection of the input sets*: Informally, while it is easy to demonstrate that the given subsets overlap (e.g., by pointing to an element from the intersection), it can be rather difficult to certify that the given subsets are disjoint.[1] 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 *Disj* and a number of its modifications.

Given the importance and the natural appeal of both the problem *Disj* and the model $\mathcal{R}$, in this work we have been looking for as simple as possible variation of the former, whose complexity analysis in the latter would lead to a new approach: While the known ones undoubtedly provide deep insight on the core of hardness of *Disj*, there seems to remain something mysterious about it, as witnessed, for example, by the fact that we know nothing about the complexity of *Disj* in the model of AM in communication complexity, which allows both randomness (like $\mathcal{R}$) and non-determinism (like $\mathcal{N}$, the "communication complexity NP").[2]

A communication problem that we call *inevitable intersection* ($\mathcal{II}$) is an instance of the intersection search problem: both Alice and Bob receive a subset as input, and their goal is to find an element that belongs to both players. Formally, $\mathcal{II}_{A,B}$ is the following problem:

Let $A, B \subseteq \{0,1\}^n$ be such that

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

[1] This can be formally expressed in terms of *non-deterministic* communication complexity $\mathcal{N}$, which captures the complexity of *certifying satisfying instances* of the target Boolean function (viewed as a predicate). The $\mathcal{N}$-complexity is $\Theta(n)$ for disjointness and $\Theta(\log n)$ for non-disjointness.

[2] The bound $\mathcal{N}(Disj) \in \Omega(n)$ follows, essentially, from the same argument as the one used to establish $\mathcal{R}(Disj) \in \Omega(n)$. However, the same approach (as well as all others that we are aware of) falls astonishingly short of providing any insight on the complexity of *Disj* in AM, a model that can be viewed as the natural closure, where the strength of both $\mathcal{N}$ and $\mathcal{R}$ is present. Resolving the communication complexity of *Disj* in AM is one of the most important open problems in the field.

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 $i \in a \cap b$.

We will analyse the (expected) $\mathcal{R}$-complexity of $\mathcal{II}_{A,B}$ when A and B are uniformly-random subsets of $\left\{x \in \{0,1\}^n \middle| |x| = n^{3/5}\right\}$ of size $2^{\sqrt[5]{n}/5}$. [3]

The definition of $\mathcal{II}$ is analogous to the well-known monotone case of Karchmer-Wigderson games [KW88, RW92], where lower bounds are known for some instances. Those techniques don't seem applicable to the randomised case of $\mathcal{II}_{A,B}$, as defined above. The special properties of $\mathcal{II}_{A,B}$ that, as we believe, have allowed it to frustrate the previous lower-bound methods are the following:

- the input space $A \times B$ has a product structure, and still it only contains intersecting pairs of subsets;
- $A \times B$ is chosen in a randomised fashion, both A and B being uniformly-random subsets of $2^{[n]}$ of the required size.

As mentioned above, virtually all known hardness arguments for *Disj* are based on the observation that it is hard for a communication protocol to certify that the input subsets are disjoint. More formally, an efficient communication protocol in any of those communication models where *Disj* is known to be hard necessarily implies existence of a relatively large nearly-monochromatic (with respect to the target problem) combinatorial rectangle in the space of pairs of input values; accordingly, showing that no such rectangle is possible proves hardness of the problem for these models. In the case of *Disj* there exist large rectangles that are biased towards intersecting input pairs; however, it has been shown that no large rectangle can be sufficiently biased towards non-intersecting pairs (which is the formal way to say that being disjoint is hard to witness), which implies that *Disj* has no efficient protocol.

In the case of $\mathcal{II}_{A,B}$, on the other hand, not only do large nearly-monochromatic rectangles exist, but in fact the whole input space $A \times B$ can be covered by n perfectly-monochromatic rectangles: e.g., for any $i \in [n]$ the set of all pairs intersecting on i form a rectangle, and the union of all such rectangles equals $A \times B$. This fact alone makes it impossible to deduce the hardness of $\mathcal{II}$ from an argument that is based on discrepancy (that is, from reasoning about the possibility of sufficient bias in large rectangles).

The *product structure* of the support $A \times B$ prevents the possibility to "secretly" add to it disjoint pairs and to base a lower bound argument on the contradiction between, on the one hand, the possibility to use a good $\mathcal{II}$-protocol for detecting disjoint pairs (where the protocol would not find a common element) and, on the other hand, the hardness of deciding whether the input pair is disjoint (to be shown via standard techniques). Adding a non-intersecting pair to $A \times B$ would require extending the marginal support of at least one of the players; even if the input encoding scheme would allowing such an extension, a very short protocol could easily detect an input pair outside $A \times B$ and produce a special response (say, declare an error), thus frustrating the intended hardness reduction. At the same time, the *lack of structure* in the families A and B themselves seems to make it impossible to embed hard instances of *Disj* into instances of $\mathcal{II}_{A,B}$.

Our way of making the complexity of *Disj* difficult to analyse has one aspect in common with the problem of analysing the AM-complexity of *Disj*: Namely, both questions escape the reach of the current techniques, partially, due to the fact that *a hardness statement cannot follow from reasoning about the discrepancy of large rectangles.* In the case of AM the model is such that an efficient protocol does not seem to guarantee existence of a large biased rectangle even in the case of

[3] The constants in the definition have been chosen to allow as simple as possible tight lower-bound analysis, based on the new technique (cf. Corollary 1). The approach seems to be adjustable to give tight results for any possible choice of parameters (though we haven't verified the details), but the required changes are rather technical and do not seem to provide enough additional insight to justify the more involved argument.

total functions, and for promise functions or relations an efficient AM protocol is known [Kla11] to be possible in some cases, where every rectangle is either exponentially small, or has error inverse-exponentially close to $1/2$. In our case the problem $\mathcal{II}$ (obviously, a variation on *Disj*) has been defined in such a way that it admits very large monochromatic rectangles, a few of which are enough to cover the whole input space.

We prove that $\mathcal{II}_{A,B}$, as defined above, is *ultimately hard*: it requires $\Omega(\sqrt[5]{n}) = \Omega(\log|A|)$ bits of interactive randomised communication (cf. Corollary 1). We are not aware of a previously-known technique that would give a non-trivial lower bound for $\mathcal{II}_{A,B}$. 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 proof for many other cases of "search-like" problems (including the set disjointness), thus giving new insight on the aspects of their combinatorial structure that are responsible for their communication hardness.

2 Preliminaries

We will write $[n]$ to denote the set $\{1, \ldots, n\} \subset \mathbb{N}$. Let $x(i)$ address the i'th bit of x for $x \in \{0,1\}^n$ and $i \in [n]$. Similarly, for $S \subseteq [n]$, let $x(S)$ denote the $|S|$-bit string, consisting of (naturally-ordered) bits of x, whose indices are in S.

Let $|x|$ denote the Hamming weight of $x \in \{0,1\}^n$. At times we will assume the trivial isomorphism between the n-bit strings and the subsets of $[n]$. In particular, the notation $\binom{[n]}{k}$ will stand for $\big\{x \in \{0,1\}^n \big| |x| = k\big\}$, and $x \cap y$ will address the set $\big\{i \in [n] \big| x_i = y_i = 1\big\}$.

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 "$X \in\!\!_R A$" 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 $\Delta > 0$,*

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

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

We will write $\mathcal{R}_p$ to denote the model of randomised interactive communication with worst-case probability of correct answer at least p. When p is obvious from the context or irrelevant, we drop the subscript $_p$. We will denote by $\mathcal{R}_{\mu,p}$ the distributional version of $\mathcal{R}_p$, where the input distribution is assumed to be μ. For a communication problem $\mathcal{S}$, we will write $\mathcal{R}(\mathcal{S})$, $\mathcal{R}_p(\mathcal{S})$ or $\mathcal{R}_{\mu,p}(\mathcal{S})$ to denote its complexity in the corresponding model.

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*

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

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

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

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 an input distribution μ such 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 the answers of the protocol, 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 bound 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

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

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 we assume by contradiction that R is small, then a significant fraction of its rectangles must be large; let us again take the extreme case and assume that all R's rectangles are large. 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. While in order to prove that *Disj* is communicationally hard one usually argues that in a short protocol *non-intersecting input pairs would have "nowhere to go"*, we will argue that in a hypothetical short protocol for $\mathcal{II}_{A,B}$ there would be *excessive probabilities* of those intersection coordinates that the protocol typically outputs.

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

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

Then for

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

it holds that

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

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 $X, Y \in \binom{[n]}{k}$. 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 $p_i \stackrel{\text{def}}{=} \mathbf{Pr}_{X \in A'}[X(i) = 1]$, $\alpha \stackrel{\text{def}}{=} \sum_{i=1}^{n} \left| p_i - \frac{k}{n} \right|$ 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

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

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

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

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

For all $i \in T$ it holds that

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

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

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

Therefore, if

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

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 $S \subseteq [n]$:

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

which can be reformulated as

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

Let $\mathscr{e}_S \stackrel{\text{def}}{=} \left[|X(S)| \le \frac{k \cdot |S|}{n} - \frac{\alpha}{4}\right]$. For a fixed S, this event depends only on the value taken by X. First we analyse the probability of $\mathscr{e}_S$ 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.[4] Accordingly,

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

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

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

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

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 select M times 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

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

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

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

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

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

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 $\beta \ge \frac{n\Delta}{4k}$, 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. ■*Lemma 2*

[4] That is, the corresponding marginal distribution is the same at each coordinate.

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

Theorem 1. *Let* $1 \leq k \leq \frac{n}{3}$ *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 $1 - \exp(n - M^{1/3} + 1) - \mathbf{Pr}[\exists a \in A, b \in B\ : a \cap b = \emptyset]$.

Note that the statement of the theorem 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 theorem above can be applied to prove the following.

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

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

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

The lower bound above 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

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

and apply Theorem 1. ■*Corollary 1*

Proof of Theorem 1. Let μ be the input distribution of (X, Y) – namely, 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 $(a, b) \in A \times B$.[5] We will keep track of the events

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

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 $\mathscr{e}_i$ is at least $1/2$. That is, "i"-labelled rectangles of $\mathcal{P}$ boost the probability of $\mathscr{e}_i$ by roughly $1/n$, which must be compensated by lowering the conditional

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

probability of 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, let us show that $\mathbf{E}_{(X,Y) \in A \times B}[|X \cap Y|]$ is unlikely to be too different from $\frac{k^2}{n}$. Let ν be the distribution of the *multiset* B, resulting from selecting M times uniformly at random an element of $\binom{[n]}{k}$ and adding it to B (i.e., $|B| \leq M$). Then the inequality below follows from the Chernoff bound (Claim 1) and the fact that the distribution of $Y \Subset B$ that results from $B \sim \nu$ is uniformly-random:

$$\forall x_0 \in \binom{[n]}{k} : \quad \mathop{\mathbf{Pr}}_{B \sim \nu}\left[\mathop{\mathbf{E}}_{Y \Subset B}[|x_0 \cap Y|] > \frac{k^2}{n} + \frac{1}{n^2}\right] \leq e^{-\frac{M}{n^5}}.$$

On the other hand, the probability that $B \sim \nu$ is a *set* is more than

$$1 - \frac{M^2}{\binom{n}{k}} \geq \frac{3}{4},$$

in which case B is a uniformly-random subset of $\binom{[n]}{k}$ of size M. Accordingly,

$$\mathop{\mathbf{Pr}}_{|B|=M}\left[\mathop{\mathbf{E}}_{Y \Subset B}[|x_0 \cap Y|] > \frac{k^2}{n} + \frac{1}{n^2}\right] < \frac{4}{3} \cdot e^{-\frac{M}{n^5}} < \exp\Big(-M^{1/3}\Big),$$

where B is a uniformly-random subset of $\binom{[n]}{k}$ of size M. By the union bound,

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

For the rest of the proof we assume that $\mathbf{E}[|X \cap Y|] \leq \frac{k^2}{n} + \frac{1}{n^2}$.

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

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

On the other hand,

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

Accordingly,

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

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

$$\begin{aligned}\sum_{r \in R_+} \frac{\mu(r)}{\mu(R_+)} \cdot \sum_{i \neq \ell(r)} \mathbf{Pr}\big[e_i \big| (X,Y) \in r\big] &\leq \left(\frac{k^2}{n} + \frac{1}{n^2} - \frac{1}{2}\right) \cdot \frac{1}{\mu(R_+)} \\ &\leq \left(\frac{k^2}{n} + \frac{1}{n^2} - \frac{1}{2}\right) \cdot \left(1 + \frac{2}{n^3}\right) \\ &\leq \frac{k^2}{n} + \frac{1}{n^2} + \frac{2k^2}{n^4} - \frac{1}{2} \\ &\leq \frac{k^2}{n} + \frac{3}{2n^2} - \frac{1}{2}.\end{aligned}$$

Therefore for some $r_0 \in R_+$,

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

which can be rewritten as

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

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 insightful discussions. I have received many valuable comments from Mika Göös and several anonymous reviewers.

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.

[Kla11] H. Klauck. On Arthur Merlin Games in Communication Complexity. *Proceedings of the 26th IEEE Conference on Computational Complexity*, pages 189–199, 2011.

[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.

[KW88] M. Karchmer and A. Wigderson. Monotone Circuits for Connectivity Require Super-Logarithmic depth. *Proceedings of the 20th Symposium on Theory of Computing*, pages 539–550, 1988.

[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.