Electronic Colloquium on Computational Complexity, Report No. 164 (2023)

Communication Complexity of Set-Intersection Problems and Its Applications

Shuo Wang* shuo_wang@sjtu.edu.cn

Guangxu Yang† guangxuy@usc.edu

Jiapeng Zhang† jiapengz@usc.edu

November 5, 2023

Abstract

Set-disjointness is one of the most fundamental and widely studied problems in the area of communication complexity. In this problem, each party i receives a set $S_i \subseteq [N]$. The goal is to determine whether $\bigcap S_i$ is empty via communication between players. The *decision version* simply asks if the common intersection is empty or not, while the *search version* asks players to find an element $a \in \bigcap S_i$ if it exists. Both versions give wide applications in diverse areas.

In this paper, we focus on the communication complexity of the search version under product distributions in the number-in-hand (NIH) model. For the decision version, Babai, Frankl, and Simon (FOCS 86) proposed an $\Omega(\sqrt{N})$ lower bound under product distributions for the two-party setting, and was extended to k-party setting $\Omega(N^{1-1/k}/k^2)$ by Dershowitz, Oshman, and Roth (STOC 21).

For the search version, though it is well-motivated by several applications, lower bounds were less known due to some technical obstacles. In this paper, we study the following natural problem, which was further motivated by Bauer, Farshim, and Mazaheri (CRYPTO 18) from cryptography motivations: there are k players, and each of them receives a (random) set of size $\approx N^{1-\alpha_i}$; the goal is to find a common element. We prove that:

- If each party holds a random set of size $\approx N^{1-\alpha_i}$, the communication complexity lower bound is $\Omega(N^{\sum_i \alpha_i - \max_i\{\alpha_i\}}/k)$.

Furthermore, we show that our lower bounds are indeed almost tight, up to logarithmic factors. Building on this lower bound, we give several applications. The first one involves improved results for the security of several combiners in backdoored random oracles (BRO). The second one is an application in distributed computing via connections established by Huang, Pettie, Zhang, and Zhang (SODA 20).

Our proof is built on a structure-vs-pseudorandomness decomposition inspired by Göös, Pitassi and Watson (FOCS 17), and this technique could be of independent interest since it enables a new method to prove communication lower bounds for search problems with many possible solutions. To the best of our knowledge, existing lower-bound techniques do not apply to this setting.

*Shanghai Jiao Tong University.

†University of Southern California. Research supported by NSF CAREER award 2141536.

ISSN 1433-8092

1 Introduction

The number-in-hand *set-disjointness* problem lies in the heart of communication complexity, whose lower bounds have wide applications in streaming algorithms, data structures, circuit complexity, and related fields (see [CP10, She14] for details). In the setting of k-party number-in-hand *set-disjointness* problem, each player i is assigned with a subset S_i of $[N]$, and the goal of this communication problem is to determine if $\bigcap_{i=1}^{k} S_i = \emptyset$.

In this paper, we consider the *search version* of the *set-disjointness* problem, which is called the *set-intersection* problem. The setting is: each player i is assigned with a subset S_i of $[N]$, and the goal changes to finding an element $a \in \bigcap_{i=1}^{k} S_i$.

We consider the communication complexity under product distribution here, and there are two typical product distributions that have been widely studied before. One is the *fixed-size product distribution*, where each player i receives a uniformly random subset $S_i \subseteq [N]$ with $|S_i| = n_i$. The other one is the *Bernoulli product distribution*, where each player i receives a random set S_i sampled as follows: for each element $a \in [N]$, $a \in S_i$ independently with probability m_i. Babai, Frankl, and Simon [BFS86] first proposed the communication complexity of the *set-disjointness* problem under fixed-size product distribution where $n_i = \sqrt{N}$, and gave an $\Omega(\sqrt{N})$ lower bound. Their proof could also be adapted to the setting of Bernoulli product distribution with $m_i = N^{-1/2}$. Recently, this bound was extended to the k-party setting by a recent paper by Dershowitz, Oshman, and Roth [DOR21]. They showed that when $k \leq \log N/6$, the communication complexity under the Bernoulli product distribution, where $m_i = N^{-1/k}$, is $\Omega(N^{1-1/k}/k^2)$. Both of these decision-version lower bounds gave various applications.

In the context of the *set-intersection* (the search version), lower bounds are less known, though it also provides many applications. One of the main obstacles here is that the size of intersections could be very large, i.e., players only need to find one common element from many possible choices. To the best of our knowledge, many existing methods fail to bypass this barrier. To this end, we now state our main result.

Our results. We consider the same setting as [DOR21]. The difference is that they assumed all players have a similar size set ($\approx N^{1-1/k}$), while we allow the sizes to be asymmetric and also consider a larger range of parameters where the size of intersections could be very large and the set-disjointness problem is easy since the $k$sets are not disjoint with high probability. Concretely, we assume that each player holds a (random) set S_i of size $|S_i| \approx N^{1-\alpha_i}$ with $\sum_i \alpha_i \leq 1$, and prove the following results for *set-intersection*.

Theorem 1.1. *For the k-party set-intersection problem under Bernoulli product distribution, where $m_i = N^{-\alpha_i}$, $\sum_i \alpha_i \leq 1$ and $k \leq 0.1 \cdot \min\{N^{\min_i\{\alpha_i\}/2}, N^{(1-\max_i\{\alpha_i\})/3}\}$[1]:*

1. *the communication complexity lower bound is $\Omega(N^{\sum_i \alpha_i - \max_i\{\alpha_i\}}/k)$ to achieve a constant accuracy;*
2. *there exists a protocol that solves this problem under the distribution mentioned above with a constant accuracy and uses $O(k \log n \cdot N^{\sum_i \alpha_i - \max_i\{\alpha_i\}})$ communication cost.*

Our main theorem suggests that $\widetilde{\Theta}(N^{\sum_i \alpha_i - \max_i\{\alpha_i\}})$ is a tight bound up to logarithmic factors when $k = o(\log N)$. Notice that our bound does not apply to the case $k \geq \Omega(\log N)$ since the

[1] We assume all the distributions considered in this paper satisfy this constraint.

requirement $k \leq 0.1N^{\min_i\{\alpha_i\}/2}$ does not hold for large k. Our bound is similar to [DOR21] when $\alpha_1 = \cdots = \alpha_k = 1/k$, where they prove an lower bound of $\Omega(N^{1-1/k}/k^2)$ for $k \leq \log N/6$.

1.1 Applications

It is well-known that *set-disjointness* is a fundamental problem in communication complexity with wide applications in many areas. Also, our results, lower bounds for the *set-intersection* problem, have applications in different areas. We mainly discuss the applications in two areas.

Applications in cryptography. Collision-resistant Hash functions (CRHFs) are a very important primitive in cryptography. A CRHF is a function $h : \{0,1\}^n \to \{0,1\}^m$ with $m < n$. It usually hopes that m is much smaller than n (the smaller, the better). Hence, there are many collision pairs in the function h. On the other hand, the security part of CRHFs would like to promise that adversaries with bounded computational power can not find such collision pairs.

We expect that our lower bounds could give applications in the analysis of CRHFs. A concrete example could be the applications given by Bauer, Farshim, and Mazaheri [BFM18]. This paper considered the communication complexity for the two-party *set-intersection* problem under Bernoulli product distribution. The main technical lemma in their paper showed an $\Omega(N^{\min\{\alpha_1,\alpha_2\}+\alpha_1+\alpha_2-1})$ lower bound for $\alpha_1 + \alpha_2 \leq 1$, where each player i receives a random set with size $\approx N^{1-\alpha_i}$.

Building on this lower bound, [BFM18] proved one-way security for combiners in 2-BRO models. However, their lower bound only showed hardness when $2\alpha_1 + \alpha_2 \geq 1$ and $\alpha_1 + 2\alpha_2 \geq 1$, and was hard to extend to k-party settings limited by their techniques. They further conjectured that:

1. The lower bound could be further improved.

2. A similar result also holds for k-party settings, which implies one-way security for combiners in k-BRO models.

Once this conjecture has been proved, their results for one-way security for combiners can be automatically improved. In this paper, we address both of them. In the two-player setting, we improve the lower bound from $\Omega(N^{\min\{\alpha_1,\alpha_2\}+\alpha_1+\alpha_2-1})$ as it shown in [BFM18] to $\Omega(N^{\min\{\alpha_1,\alpha_2\}})$, i.e., we remove the extra $(1-\alpha_1-\alpha_2)$ term. For a larger gap of $(1-\alpha_1-\alpha_2)$, our improvement becomes more significant. Notice that $N^{1-\alpha_1-\alpha_2}$ is exactly the expected size of $|S_1 \cap S_2|$, indicating that our improvement is more significant for larger intersections. We also remove the constraints of $2\alpha_1 + \alpha_2 \geq 1$ and $\alpha_1 + 2\alpha_2 \geq 1$. Moreover, our result applies for the k-party setting and is tight up to logarithmic factors.

By the reductions established in [BFM18], our result directly gives the following improvements in cryptography:

1. improved one-way security for three combiners in the BRO model, including concatenation combiners, cascade combiners, and xor combiners;

2. generalization of previous results from the 2-BRO model to the k-BRO model.

Applications in distributed computing: Connections between distributed computing and communication complexity have also been widely studied. However, as Drucker, Kuhn, and Oshman [DKO12] pointed out:

> *Importing problems from the distributed computing world into the communication complexity model raises issues that are not often considered in existing communication complexity lower bounds: search problems, where players are allowed to choose one of many possible outputs; partial knowledge, where each player needs to output only part of the answer; and unicast communication cost.*

Hence, in order to further study the connections between distributed computing and communication complexity, techniques to prove communication lower bounds in new settings are needed. This echoes the cryptography applications in which lower-bound techniques for search problems with many solutions are sought.

We now give a concrete application of our main theorem (Theorem 1.1). A nice paper by Huang, Pettie, Zhang, and Zhang [HPZZ21] studied a similar (two-party) *set-intersection* problem. However, in their setting, Alice and Bob are asked to find the whole set $S_1 \cap S_2$ (enumerating all solutions is somehow similar to decision problems). They observed this *set-intersection* enumeration problem is equivalent to the (local) triangle *enumeration* problem in the CONGEST networks. Building on their lower bounds on *set-intersection* enumeration lower bounds, they proved a lower bound for triangle enumeration in the CONGEST networks. By using the same reduction as in [HPZZ21], our main theorem (Theorem 1.1) implies a lower bound for explicitly finding one triangle in the CONGEST networks. For more backgrounds and motivations about triangle detection and enumeration, we refer to the papers by Dolev, Lenzen, and Peled [DLP12] and Izumi and Le Gall [ILG17].

Further potential applications. Inspired by lifting theorems from communication complexity, Coretti, Dodis, Guo, and Steinberger [CDGS18] employed a pre-sampling technique to prove a number of positive results in the random oracle model. The main idea of this pre-sampling technique is to decompose a high-entropy source into many structured distributions (they called it bit-fixing source). Despite many successes, their deposition still has some limitations. For example, a more recent paper by Dodis, Farshim, Mazaheri, and Tessaro [DFMT20] studied the security in the backdoored random oracle model. However, the decomposition method by [CDGS18] can not solve the question in [DFMT20]. [DFMT20] made the following comment towards this:

> *In order to overcome the bounded adaptivity restriction and prove full indifferentiability, one would require an improved decomposition technique which fixes considerably less points after each leakage. This, at the moment, seems (very) challenging and is left as an open question. In particular, such a result would simultaneously give new proofs of known communication complexity lower bounds for a host of problems, such as set-disjointness and intersection, potentially a proof of the conjecturally hard problem stated in [3], and many others. (We note that improved decomposition techniques can potentially also translate to improved bounds.)*

Since we provide a new decomposition and give an almost tight bound for the *set-intersection* problem, we believe that our decomposition is helpful for the applications in [CDGS18]. The main conceptual difference between our decomposition and pre-sampling is as follows.

- Pre-sampling decomposes a high-entropy source directly.
- We decompose rectangles recursively, i.e., the decomposition of a node in the protocol tree tracks the decomposition information from its ancestors. (See more details in Section 1.2)

Overall, our method enables a new approach to prove communication lower bounds for search problems, even when there are many possible outputs, and it could be applied to the k-party setting without any barrier.

1.2 Proof outline

In this section, we give a brief introduction to our proof for the lower bound and the idea behind it. Instead of considering Bernoulli distributions, we consider the following product distribution to simplify our presentation:

- Each player i independently and uniformly samples $cN^{1-\alpha_i}$ elements from $[N]$ (may have duplicates), where c equals $(1+2/k)$ here.

Thus, each player i receives a vector in $[N]^{cN^{1-\alpha_i}}$ and gets its set $S_i \subseteq [N]$ by removing the duplicate elements in the vector. In general, for any $I \subseteq [cN^{1-\alpha_i}]$ and $\beta \in [N]^I$, we consider β as a subset of $[N]$ in a similar way. We prove the lower bound under this distribution, and then reductions are established in Section 3.3 to prove our main theorem.

It is well known that a deterministic protocol Π partitions the input domain into $2^{|\Pi|}$ rectangles by step-by-step communication. The crucial idea of our proof is to further partition these leaf rectangles in the protocol tree into many structured rectangles defined below.

Definition 1.2 (Structured rectangles). *Assuming $R = X_1 \times X_2 \times \cdots \times X_k$, where each X_i is a subset of $[N]^{cN^{1-\alpha_i}}$, is a rectangle. We say R is a structured rectangle if there exist subsets of coordinates $J_1, J_2, \cdots, J_k$ with $J_i \subseteq [cN^{1-\alpha_i}]$ satisfying that*

- *For each i, there exists a $\beta_i \in [N]^{J_i^c}$ such that $\forall x_i \in X_i, x_i(J_i^c) = \beta_i$. Here, J_i^c is the complement of J_i defined by $J_i^c := [cN^{1-\alpha_i}] - J_i$ and $x_i(J_i^c) \in [N]^{J_i^c}$ is the values of x_i on J_i^c.*
- *For each i, X_i has a high block-wise min-entropy (see definitions in Section 2) on the coordinates J_i.*

The notion of structured rectangle has also been widely used in query-to-communication lifting theorems [GPW17, CFK$^+$19, LMM$^+$22].

In the decomposition, we recursively (starting from the root to the leaves) decompose all rectangles in the protocol tree, i.e., for a node (which is also a rectangle), we decompose it based on the decomposition of its ancestors. This is the key step compared to existing decomposition (presampling techniques) in cryptography, which may lead to new applications. The formal process of this decomposition is referred to Section 3.

After the decomposition process, each leaf has been partitioned into many structured rectangles. For a structured rectangle $R = X_1 \times X_2 \times \cdots \times X_k$ associated with $J_1, \ldots, J_k$ and $\beta_i \in [N]^{J_i^c}$ for $i \in [k]$, we say that:

1. R is *bad* if $\cap_i \beta_i \neq \emptyset$.
2. R is *good* if $\cap_i \beta_i = \emptyset$. We also call good structured rectangles as pseudorandom rectangles.

Then, our proof consists of the following two parts.

- If the communication complexity of Π is small, the total size of bad structured rectangle is small compared to the size of the input domain (formalized by Lemma 3.2);

- On the other hand, we show that players can not find a common intersection from pseudo-random rectangles (formalized by Lemma 3.3).

Combining the two parts, we are able to prove the main theorem. We refer the detailed proofs to Section 3.

Comparison to previous proofs. Similar questions have been widely studied in several recent papers [BFM18, HPZZ21, DOR21, OR23]. All of these papers used standard known techniques in communication complexity such as information complexity.

These papers achieved tight bounds for *set-disjointness* (decision version), or set-intersection enumeration (finding whole intersections). However, all of their bounds for search problems are sub-optimal whenever the size of the set intersection (the solution space for the search problem) is large. By contrast, our new method is inspired by lifting theorems [GPW17, CFK$^+$19, LMM$^+$22]. On the other hand, unlike the previous lifting theorems, we do not require the communication function to have a composed form (with a gadget).

2 Preliminaries

To begin with, we formally define the product distributions adopted in this paper. For fixed parameters: k is the number of parties, N is the size of the domain, and $\alpha_i \in (0, 1)$ are parameters indicating the size of each player's set. We consider the following three types of hardness distributions in this paper (two of them have appeared in Section 1):

1. Each player i independently and uniformly samples $cN^{1-\alpha_i}$ elements from $[N]$ (may have duplicates), where c equals $(1 + 2/k)$.

2. Each player i independently and uniformly samples $c_i N^{1-\alpha_i}$ distinct elements from $[N]$, where $1 - 1/k \leq c_i \leq 1 + 1/k$.

3. Each player i independently samples its set S_i with that every element $a \in [N]$ is contained in S_i with probability $N^{-\alpha_i}$.

We assume that $\sum_i \alpha_i \leq 1$, otherwise the existence of intersections can be not guaranteed. Furthermore, if $\sum_i \alpha_i \leq 1 - C$ holds for some constant $C > 0$, the common intersection of all players could be very large ($\approx N^C$).

The hardness distribution 3 is the Bernoulli product distribution with wide applications. Previous papers have mainly focused on this distribution. We prove the lower bound under distribution 1, and use two simple reductions to get the lower bound results for the hardness distribution 2 and 3. We refer the two reductions to Section 3.3. In what follows, our discussion mainly focuses on distribution 1.

For a distribution D and a communication protocol Π, we define the *accuracy* of Π on D by:

$$\mathrm{Acc}_\Pi(D) := \Pr_{S_1,\cdots,S_k \sim D}\left[\Pi(S_1, \cdots, S_k) \in \bigcap_{i=1}^{k} S_i\right].$$

For simplicity in notations, we define this accuracy notion, which does not take the cases when sets are disjoint into consideration, differently from [BFM18] in which they also consider the accuracy of distinguishing disjoint cases, namely they define

$$\mathrm{Acc}'_{\Pi}(D) := \Pr_{S_1,\cdots,S_k\sim D}\left[\Pi(S_1,\cdots,S_k)\in\bigcap_{i=1}^{k}S_i \text{ or } \Pi(S_1,\cdots,S_k)=\bigcap_{i=1}^{k}S_i=\emptyset\right].$$

Since we aim to establish lower bounds for those protocols achieving $\mathrm{Acc}_{\Pi}(D)=\Omega(1)$, we only consider the range of $\alpha_1,\ldots,\alpha_k$ with[2]

$$\Pr_{S_1,\cdots,S_k\sim D}\left[\bigcap_{i=1}^{k}S_i\neq\emptyset\right]>1/2.$$

In this paper, our lower bound result shows that achieving $\mathrm{Acc}_{\Pi}(D)>\epsilon$, where epsilon is a constant less than $1/2$, requires large amounts of communication. This also implies a non-trivial hardness result to achieve $\mathrm{Acc}'_{\Pi}(D)>\epsilon+1/2$ since the disjoint cases could contribute at most $1/2$ to $\mathrm{Acc}'_{\Pi}(D)$ when $\Pr_{S_1,\cdots,S_k\sim D}\left[\bigcap_{i=1}^{k}S_i\neq\emptyset\right]>1/2$ holds. Hence, our results also imply hardness results under the [BFM18] setting.

Next, we introduce some useful notions in communication complexity. In a k-party communication problem, where each party holds an input x_i from a domain Δ_i, a rectangle is defined by $R:=X_1\times X_2\times\cdots\times X_k$ ($X_i\subseteq\Delta_i$).

For a subset $X_i\subseteq\Delta_i$, we denote $\boldsymbol{X}_i$ as the uniform distribution on X_i. In the set-intersection problem (particularly hard distribution 1), we consider the cases that each input is in $\Delta_i=[N]^{M_i}$ where $M_i=cN^{1-\alpha_i}$, and an instance $x_i\in[N]^{M_i}$ can be transformed into a subset of $[N]$ by removing duplicate elements. Also, for two instances $x_i\in[N]^{M_i}, x_j\in[N]^{M_j}$, we define $x_i\cap x_j$ by the intersection of the two subsets of $[N]$ deduced from x_i and x_j.

For a set of coordinates $J_i\subseteq[M_i]$, we use $\boldsymbol{X}_i(J_i)$ to denote marginal distribution of $\boldsymbol{X}_i$ on J_i. For an instance $x_i\in[N]^{M_i}$ and a set of coordinates $J_i\subseteq[M_i]$, define $x_i(J_i)$ to be an instance in $[N]^{J_i}$ by projecting x_i on J_i.

A useful concept adopted in this paper is the dense notion used in lifting theorems [GPW17, CDGS18, CFK+19, LMM+22].

Definition 2.1 (Min-entropy). *For a random variable $\boldsymbol{X}$ taking value on Δ, its min-entropy is defined as follows:*

$$H_\infty(\boldsymbol{X})=\min_{x\in\Delta}\left(\log\frac{1}{\Pr[\boldsymbol{X}=x]}\right).$$

Definition 2.2 (Density function). *We define the one-side density function for a random variable $\boldsymbol{X}$ on its support $[N]^J$ as:*

$$\mathcal{D}(\boldsymbol{X}):=|J|\log N-H_\infty(\boldsymbol{X}).$$

Note that $\mathcal{D}(\boldsymbol{X})\geq 0$ always holds by definitions and $\mathcal{D}(\boldsymbol{X})=0$ when $\boldsymbol{X}$ is a uniform distribution.

Definition 2.3 (k-side density function). *For a structured rectangle $R=X_1\times X_2\times\cdots\times X_k$, where each X_i is subset of $[N]^{M_i}$ and associated with a set $J_i\subseteq[M_i]$, we define its k-side density function as:*

$$\mathcal{D}(R)=\mathcal{D}\big(\boldsymbol{X}_1(J_1)\big)+\mathcal{D}\big(\boldsymbol{X}_2(J_2)\big)+\cdots+\mathcal{D}\big(\boldsymbol{X}_k(J_k)\big).$$

[2] $\Pr_{S_1,\cdots,S_k\sim D}\left[\bigcap_{i=1}^{k}S_i\neq\emptyset\right]>1/2$ is guaranteed by the definitions of hardness distribution 3 when $\sum_i\alpha_i\leq 1$.

The density function is also known as the entropy deficiency in lifting theorem papers, and we design the k-side density function in order to extend the two-party results to the k-party setting.

Definition 2.4. *A random variable $\boldsymbol{X}$ on $[N]^J$ is called $(1-\delta)$-dense if for every subset $I \subseteq [J]$,*

$$H_\infty(\boldsymbol{X}(I)) \geq (1-\delta) \cdot \log N \cdot |I|.$$

The definition of $(1-\delta)$-dense measures the pseudorandomness of a random variable. In our proof, a typical choice of δ is $\frac{1}{10k \log N}$ [3]

The following lemma tells us that a random variable could be decomposed by a combination of random variables with dense properties by fixing some positions:

Lemma 2.5 (Density-restoring partition [GPW17])**.** *Let X be a subset of $[N]^M$ and J be a subset of $[M]$, and there exists an $\beta \in N^{J^c}$ such that $\forall x \in X, x(J^c) = \beta$. Then, there exists a partition of X:*

$$X := X^1 \cup X^2 \cup \cdots \cup X^r$$

such that every X^i is associated with a set $I_i \subseteq J$ and a value $\tau_i \in [N]^{I_i}$. Then, they satisfy the following properties:

1. *$\forall x \in X^i, x(I_i) = \tau_i$;*
2. *$\boldsymbol{X}^i(J - I_i)$ is $(1-\delta)$-dense;*
3. *$D\Big(\boldsymbol{X}^i(J - I_i)\Big) \leq D\Big(\boldsymbol{X}(J)\Big) - \delta |I_i| \log N + \gamma_i$.*

Here, we define $\gamma_i := \log(|X|/|\cup_{j \geq i} X^j|)$.

For dense random variables, we also have the following useful lemma.

Lemma 2.6. *If $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_\ell$ are $\ell < k$ independent $\big(1 - \frac{1}{10k \log N}\big)$-dense random variables and each $\boldsymbol{X}_i$ takes value from $[N]^{J_i}$ with $|J_i| \leq c \cdot N^{1-\alpha_i}$, where c is a constant and $N^{\alpha_i} = \omega(k)$, then for any element $a \in [N]$, it holds*

$$\Pr\left[a \in \bigcap_{i=1}^{\ell} \boldsymbol{X}_i\right] \leq \frac{ec^\ell}{N^{\sum_i \alpha_i}},$$

here $e \approx 2.7$ denotes the Euler's number.

3 Lower bounds for the product distributions

In this section, we prove the communication lower bound for the hardness distribution 1. Then, in Section 3.3, we use reductions to obtain lower bounds for hardness distributions 2 and 3. Formally, we prove that:

Theorem 3.1. *If a communication protocol Π solves k-party set-intersection problem under the hardness distribution 1 with accuracy bigger than 0.1, the communication complexity $\mathsf{CC}(\Pi)$ is*

$$\Omega\bigg(\frac{N^{\sum_i \alpha_i - \max_i\{\alpha_i\}}}{k}\bigg).$$

[3] $\delta = 0.9$ in previous structure-vs-pseudorandomness decomposition [GPW17, CDGS18, CFK$^+$19, LMM$^+$22].

3.1 The decomposition and sampling process

The key idea of this proof, as we introduce in Section 1, is to decompose rectangles (nodes[4]) of the protocol tree into structured rectangles and analyze the accuracy of the protocol could achieve in those decomposed structured rectangles. We design a *decomposition and sampling process* in this section to

- decompose the rectangles of the protocol tree into structured rectangles;
- sample a decomposed rectangle with respect to its size.

We define the root rectangle of the protocol tree to be R^{root}, which contains all valid inputs. R^{root} is also a structured rectangle by definitions. We start from R^{root} and begin our decomposition and sampling process, which uses a random walk on the protocol tree from the root R^{root} to a leaf, and do the decomposition along the path. See Algorithm 1 for the formal decomposition process.

Algorithm 1: The decomposition and sampling process

```
Input: A rectangle $R^{\text{root}} = X_1 \times X_2 \times \cdots \times X_k$, where each $X_i$ equals $[N]^{cN^{1-\alpha_i}}$.
Output: A rectangle $R^{\text{cur}} = X_1^{\text{cur}} \times X_2^{\text{cur}} \times \cdots \times X_k^{\text{cur}}$, and $k$ sets $J_1, J_2, \cdots, J_k$.
for each $i$, $J_i \leftarrow [cN^{1-\alpha_i}]$;
$R^{\text{cur}} \leftarrow R^{\text{root}}$;
while $R^{cur}$ is not in a leaf level [a] do
    without loss of generality, we assume it is player $i$'s turn to speak;
    $X_i^{\text{cur}}$ is partitioned by: $X_i^{\text{cur}} = X^0 \cup X^1$, and $R^{\text{cur}}$ is thus partitioned by: $R^{\text{cur}} = R^0 \cup R^1$;
    toss a $(\frac{|X^0|}{|X_i^{\text{cur}}|}, \frac{|X^1|}{|X_i^{\text{cur}}|})$ biased coin $c$;
    if $c = 0$:
        $X_i^{\text{cur}} \leftarrow X^0$;
        $R^{\text{cur}} \leftarrow R^0$;
    if $c = 1$:
        $X_i^{\text{cur}} \leftarrow X^1$;
        $R^{\text{cur}} \leftarrow R^1$;
    if $\boldsymbol{X}_i^{\text{cur}}(J_i)$ is $(1 - \frac{1}{10k \log n})$-dense:
        continue;
    else:
        decompose $X_i^{\text{cur}}$ by Lemma 2.5 with $J = J_i$, get $X^1, \cdots, X^r, I_1, \cdots, I_r$;
        $R^{\text{cur}}$ is thus decomposed by $R^{\text{cur}} = R^1 \cup \cdots \cup R^r$;
        sample a random element $\boldsymbol{j} \in [r]$: $\boldsymbol{j}$ w.p. $|X^j|/|X_i^{\text{cur}}|$ equals $j$ for each $j$;
        $X_i^{\text{cur}} \leftarrow X^{\boldsymbol{j}}$, $J_i \leftarrow J_i \backslash I_{\boldsymbol{j}}$;
        $R^{\text{cur}} \leftarrow X_1 \times X_2 \times \cdots \times X_k$;
end
```

[a] R^{cur} is not in a leaf level means R^{cur} is not a sub-rectangle of any leaf rectangle of the protocol tree.

We use R^{cur} to denote the current rectangle of the decomposition and sampling process. It begins with $R^{\text{cur}} = R^{\text{root}}$, and at each step R^{cur} is partitioned into two subrectangles R^0, R^1 by the protocol. Then, we replace R^{cur} with R^0 or R^1 with probability $|R^0|/|R^{\text{cur}}|$ or $|R^1|/|R^{\text{cur}}|$

[4] Note that a node of the protocol tree is a rectangle.

(which also equals to $|X^0|/|X_i^{\mathrm{cur}}|$ or $|X^1|/|X_i^{\mathrm{cur}}|$ as we defined in Algorithm 1), and reach a new rectangle. After reaching the new rectangle, the structured property of R^{cur} may get destructed, and our decomposition works here to maintain the structured property. We use the density-restoring partition (Lemma 2.5) to further decompose the current rectangle R^{cur} into r subrectangles $R^{\mathrm{cur}} = R^1 \cup R^2 \cup \cdots \cup R^r$, and each R^j is a structured rectangle. Again, we choose R^j to be our next rectangle with probability $|R^j|/|R^{\mathrm{cur}}|$, and do the process above recursively until reaching a leaf rectangle. As shown in the decomposition and sampling process, we eventually sample a structured rectangle in the leaf level with respect to its size.

Note that at some point of the random walk, the current rectangle R^{cur} may not exist on the protocol tree since we do the density-restoring partition to further decompose the rectangles. However, every R^{cur} that potentially appears in the random walk must be fully contained in a rectangle of the protocol tree. Thus, the protocol Π also partitions R^{cur} into two sub-rectangles if R^{cur} is not in the leaf level of the protocol tree.

Note that the output R^{cur} of the process above is a random variable over rectangles. We define $\boldsymbol{R}^{\mathrm{leaf}}$ to be the random variables over decomposed structured rectangles in the leaf level (not leaf rectangles of the protocol tree, but sub-rectangles of those leaves after decomposition) sampled by the process above, and $\boldsymbol{R}^{\mathrm{leaf}}$ is associated with random sets $\boldsymbol{J}_i^{\mathrm{leaf}}$s. For convenience, we define the support of $\boldsymbol{R}^{\mathrm{leaf}}$ to be $\mathcal{R}^{\mathrm{leaf}}$. One may see the two important properties of the decomposition and sampling process:

- Every rectangle $R \in \mathcal{R}^{\mathrm{leaf}}$ is a structured rectangle;
- For a rectangle $R = X_1 \times X_2 \times \cdots \times X_k \in \mathcal{R}^{\mathrm{leaf}}$, we have that

$$\Pr[\boldsymbol{R}^{\mathrm{leaf}} = R] = \prod_i \frac{|X_i|}{cN^{1-\alpha_i}} = \frac{|R|}{c^k N^{k-\sum_i \alpha_i}}.$$

The verification of the two properties is straightforward from the definition of our decomposition and sampling process. The first statement offers the structured property which makes it easier to analysis the rectangles. The second statement tells us that: the probability that $\boldsymbol{R}^{\mathrm{leaf}} = R$ equals the probability that the input lies in R. This is crucial in later bounding the accuracy of Π.

Next, we bound the accuracy of Π. For every structured rectangle $R = X_1 \times X_2 \times \cdots \times X_k \in \mathcal{R}^{\mathrm{leaf}}$ associated with $J_1, J_2, \cdots, J_k$, we define J_i^c as $[cN^{1-\alpha_i}] - J_i$, namely the fixed parts of X_i. Hence, for each X_i, it holds $\forall x \in X_i, x(J_i^c) = \beta_i$ since R is a structured rectangle. We can then divide all the rectangles in $\mathcal{R}^{\mathrm{leaf}}$ into two types:

1. R is a *bad* structured rectangle if $\cap_i \beta_i \neq \emptyset$;
2. R is a *good* structured rectangle if $\cap_i \beta_i = \emptyset$.

Assume R is a bad structured rectangle. Then, there exists a universal common element a[5] such that $a \in \cap_i x_i$ for any possible instance $(x_1, x_2, \cdots, x_k)$ in R. The protocol is thus able to achieve perfect correctness by outputting a when the input lies in R. Hence, we need to show with a low probability that $\boldsymbol{R}^{\mathrm{leaf}}$ is a bad rectangle, namely the probability that input lies in bad rectangles is small. To be more specific, we prove the following lemma:

Lemma 3.2. *If $CC(\Pi) \leq 0.0001 N^{\sum_i \alpha_i - \max_i\{\alpha_i\}}/k$, it holds that $\Pr_{R\sim \boldsymbol{R}^{leaf}}[R \text{ is bad}] \leq 0.05$.*

[5] We can choose any element that lies in $\cap_i \beta_i$ here.

For those good structured rectangles, we show the following facts: for a good structured rectangle, the protocol Π cannot achieve high accuracy since there is no intersection on the fixed parts, while the other parts are dense. Formally, we prove the following lemma:

Lemma 3.3. *For a good structured rectangle $R = X_1 \times X_2 \times \cdots \times X_k$, it holds that for any $a \in [N]$,*

$$\Pr[a \in \cap_i \boldsymbol{X}_i] \le 0.05.$$

Combining the three lemmas above, we can easily prove Theorem 3.1.

Proof of Theorem 3.1. We prove the theorem by showing that communication protocol Π with $\mathrm{CC}(\Pi) \le 0.0001 N^{\sum_i \alpha_i - \max_i\{\alpha_i\}}/k$ can achieve at most 0.1 accuracy.

It is well known that a communication protocol Π partitions the whole input domain into several leaf rectangles and assigns an answer to each leaf rectangle. With our decomposition and sampling process, original leaf rectangles are further decomposed into two types of structured rectangles mentioned above. The accuracy of Π comes from the following two parts:

1. The probability $\Pr[\boldsymbol{R}^{\text{leaf}} \text{ is bad}] = p_1$.

2. The probability that protocol outputs the correct answer in a good structured rectangle is p_2.

From Lemma 3.2 and 3.3, we know that $p_1 \le 0.05, p_2 \le 0.05$. By a union bound, the total accuracy is thus no more than $p_1 + p_2 \le 0.1$ as desired. □

It suffices to prove the two important lemmas above.

3.2 Proofs of technical lemmas

We first prove Lemma 3.2 by the following round-by-round analysis.

Proof of Lemma 3.2. Recall the decomposition process from line 4 to line 12. In each communication round, player i sends one bit, and partitions X_i^{cur} into two parts X^0, X^1. Then, X_i^{cur} is replaced by X^0 (or X^1) with probability $\frac{|X^0|}{|X_i^{\text{cur}}|}$ (or $\frac{|X^1|}{|X_i^{\text{cur}}|}$). In this process, the density function $\mathcal{D}(\boldsymbol{X}_i^{\text{cur}}(J_i))$ would increase since the size of $|X_i^{\text{cur}}|$ decreases. This contributes to the density function with an increment of:

- $\log(\frac{|X_i^{\text{cur}}|}{|X^0|})$ with probability $|X^0|/|X_i^{\text{cur}}|$;

- $\log(\frac{|X_i^{\text{cur}}|}{|X^1|})$ with probability $|X^1|/|X_i^{\text{cur}}|$.

Thus, in expectation, the density function of $R^{\text{cur}} = X_1^{\text{cur}} \times X_2^{\text{cur}} \times \cdots \times X_k^{\text{cur}}$ after partitioning will increase

$$\frac{|X^0|}{|X_i^{\text{cur}}|} \log\left(\frac{|X_i^{\text{cur}}|}{|X^0|}\right) + \frac{|X^1|}{|X_i^{\text{cur}}|} \log\left(\frac{|X_i^{\text{cur}}|}{|X^1|}\right) \le 1, \tag{1}$$

where $|X_i^{\text{cur}}|$ denotes the size of X_i^{cur} before partitioning. Furthermore, if $\boldsymbol{X}_i^{\text{cur}}(J_i)$ is no longer $(1 - \frac{1}{10k \log n})$-dense, we partition X_i^{cur} by Lemma 2.5 and get $X_i^{\text{cur}} = X^1 \cup X^2 \cup \cdots \cup X^r$ and

$I_1 \cup I_2 \cup \cdots \cup I_r$ with $X^j(I_j) = \tau_j$ for all j. We use Lemma 2.5, where we take $\delta = 1/(10k \log N)$, and get:

$$\mathcal{D}(\boldsymbol{X}^j(J_i - I_i)) \leq \mathcal{D}(\boldsymbol{X}_i^{\text{cur}}(J_i)) - \delta |I_j| \log N + \gamma_j = \mathcal{D}(\boldsymbol{X}_i^{\text{cur}}(J_i)) - \frac{|I_j|}{10k} + \gamma_j. \tag{2}$$

Recall that $\gamma_j := \log(|X_i^{\text{cur}}|/|\cup_{p \geq j} X^p|)$ here. In the decomposition process, X_i^{cur} is replaced with X^j with probability $|X^j|/|X_i^{\text{cur}}|$. Hence, taking expectation in one communication round, we have

$$\mathbb{E}[\gamma_j] = \sum_j \frac{|X^j|}{|X_i^{\text{cur}}|} \log(|X_i^{\text{cur}}|/|\cup_{p \geq j} X^p|) \leq \int_0^1 \log \frac{1}{1-x} dx = 1. \tag{3}$$

Thus, combining (1), (2) and (3) and taking expectations, we know that after $\mathsf{CC}(\Pi)$ rounds of communication (where each round communicates exact one bit message), it holds:

$$\mathbb{E}_{R \sim \boldsymbol{R}^{\text{leaf}}}[\mathcal{D}(R)] \leq 2 \cdot \mathsf{CC}(\Pi) - \frac{\mathbb{E}_{J_1 \sim \boldsymbol{J}_1^{\text{leaf}}, \cdots, J_k \sim \boldsymbol{J}_k^{\text{leaf}}} \left[\sum_{j=1}^k |J_j^c| \right]}{10k}.$$

Here, the $2 \cdot \mathsf{CC}(\Pi)$ comes from (1) and (3). We know that $\mathbb{E}_{R \sim \boldsymbol{R}^{\text{leaf}}}[\mathcal{D}(R)] \geq 0$ from definitions. Hence, we have

$$\sum_{j=1}^k \mathbb{E}_{J_j \sim \boldsymbol{J}_j^{\text{leaf}}}[|J_j^c|] \leq 20k \cdot \mathsf{CC}(\Pi). \tag{4}$$

We can bound the probability that the bad structured rectangle appears round by round. At each round of communication, if we choose X^j to replace X_i^{cur}, then we will fix $|I_j|$ more positions for X_i^{cur}. We then consider the probability that this new fixed part contributes to forming a bad structured rectangle with future fixed positions.

Let $R^j = X_1^{\text{cur}} \times X_2^{\text{cur}} \times \cdots X^j \cdots \times X_k^{\text{cur}}$, for any $x = (x_1^{\text{cur}}, x_2^{\text{cur}}, \cdots, x^j, \cdots, x_k^{\text{cur}}) \in R^j$, we label it as a error term if $\exists a \in \tau_j$, $a \in \bigcap_{p \neq i} x_p^{\text{cur}}(J_p)$ [6]. By Lemma 2.6, for any $a \in \tau_j$,

$$\Pr[a \in \bigcap_{p \neq i} \boldsymbol{X}_p^{\text{cur}}(J_p)] \leq \frac{ec^{k-1}}{N^{(\sum_{p=1}^k \alpha_p) - \alpha_i}}$$

By a union bound, the probability that error terms appear in R^j is

$$\Pr[\exists a \in \tau_j, a \in \bigcap_{p \neq i} \boldsymbol{X}_p^{\text{cur}}(J_p)] \leq \frac{|I_j| \cdot ec^{k-1}}{N^{(\sum_{p=1}^k \alpha_p) - \alpha_i}}$$

Also, we know that the total number of fixed elements equals $\sum_{i=1}^k |J_i^c|$, which is identical to the summation of $|I_j|$ of every step, thus, the average probability of error terms at the end of the decomposition process is at most

$$\frac{ec^{k-1}}{N^{(\sum_i \alpha_i) - \max_i \{\alpha_i\}}} \cdot \sum_{i=1}^k \mathbb{E}_{J_i \sim \boldsymbol{J}_i^{\text{leaf}}} [|J_i^c|].$$

[6] τ_j is a fixed subset of $[N]$ with size at most $|I_j|$ since X^j is fixed on I_j. Input x may be labeled many times during the decomposition process.

We note that for any $R \in \mathcal{R}^{\text{leaf}}$, if R is bad, then all instances $x \in R$ have been labeled as an error term in the decomposition process, together with (4), we have

$$\Pr_{R\sim \boldsymbol{R}^{\text{leaf}}}[R \text{ is bad}] \le \frac{ec^{k-1}}{N^{(\sum_i \alpha_i)-\max_i\{\alpha_i\}}} \cdot \sum_{i=1}^{k} \mathbb{E}_{J_i\sim \boldsymbol{J}_i^{\text{leaf}}}\left[|J_i^c|\right] \le 0.05.$$

The last inequality holds since $c = (1 + 2/k)$ and $\mathsf{CC}(\Pi) \le 0.0001 N^{\sum_i \alpha_i - \max_i\{\alpha_i\}}/k$. □

Next, we show that in the good structured rectangles, the protocol Π cannot achieve large accuracy in finding the common element. This also comes from the structured properties of the rectangles:

Proof of Lemma 3.3. Notice that we consider the rectangle $R = X_1 \times X_2 \times \cdots \times X_k$ associated with $J_1, J_2, \cdots J_k$ that has no common elements on fixed parts J_i^c. Thus, for any element $a \in [N]$, there exists at least a party i which does not contain a on its fixed part. Thus, we use Lemma 2.6 for $\boldsymbol{X}_i(J_i)$ with $\ell = 1$, and get

$$\Pr[a \in \boldsymbol{X}_i] = \Pr[a \in \boldsymbol{X}_i(J_i)] \le ce/N^{\alpha_i} = o(1).$$

□

3.3 Lower bounds for other hardness distributions

In this section, we first establish a reduction from Bernoulli hardness distribution (hardness distribution 3) to hardness distribution 2 by the following lemma:

Lemma 3.4. *If a communication protocol Π that solves set-intersection under hardness distribution 3 with accuracy ϵ, there exists parameters $c_1, \cdots, c_k$ with each $1 - 1/k \le c_i \le 1 + 1/k$ for hardness distribution 2 so that Π can find set intersection under this distribution with accuracy $\epsilon - 2k\exp(-\frac{N^{1-\max_i\{\alpha_i\}}}{3k^2})$, which is bigger than $\epsilon - 0.01$ when $N^{1-\max_i\{\alpha_i\}} \ge 100k^2 \log k$.*

Proof. We first use Chernoff bound to bound the probability of the size of set S_i of each player i exceeding $(1 + 1/k) \cdot N^{1-\alpha_i}$ or less than $(1 - 1/k) \cdot N^{1-\alpha_i}$ under the hardness distribution 3:

$$\Pr[||S_i| - N^{1-\alpha_i}| > 1/k \cdot N^{1-\alpha_i}] \le 2\exp\left(-\frac{N^{1-\alpha_i}}{3k^2}\right).$$

We use A to denote the event that $\exists i, ||X_i| - N^{1-\alpha_i}| > 1/k \cdot N^{1-\alpha_i}$. Then, by a union bound, we know that:

$$\Pr[A] \le 2k \cdot \exp\left(-\frac{N^{1-\max_i\{\alpha_i\}}}{3k^2}\right).$$

Then, condition on $\neg A$, we have the success probability of Π in finding set intersection under hardness distribution 3 is bigger than $\epsilon - 2k \cdot \exp\left(-\frac{N^{1-\max_i\{\alpha_i\}}}{3k^2}\right)$. Furthermore, condition on $\neg A$, the hardness distribution 3 can be represented by a combination of product distributions:

$$\sum_{c_1,c_2,\cdots,c_k} \sigma(c_1, c_2, \cdots, c_k) D_{c_1,c_2,\cdots,c_k},$$

where $D_{c_1,c_2,\cdots,c_k}$ denotes the hardness distribution 2 with parameters $c_1, c_2, \cdots, c_k$. Then, the lemma follows by an averaging argument. □

It suffices to construct a reduction from hardness distribution 2 to hardness distribution 1.

Lemma 3.5. *If there exists a communication protocol Π with communication complexity C which solves set-intersection under hardness distribution 2 with accuracy ϵ, there exists a communication protocol Π' with communication complexity C which solves set-intersection under hardness distribution 1 with accuracy $\epsilon - 0.05$ when $k^2 N^{-\min_i\{\alpha_i\}} \leq \frac{1}{100}$ holds.*

Proof. We construct the communication protocol Π' as follows:

1. For each player i, remove the duplicate elements of its input and get a $S_i \subseteq [N]$.
2. Randomly sample $c_i N^{1-\alpha_i}$ elements from S_i, Π' fail if $|Y_i| < c_i N^{1-\alpha_i}$.
3. Run the communication protocol Π on Y_is to find intersection.

We know that the successful probability of Π' under hardness distribution 1 is bigger than

$$\epsilon - \Pr[\Pi' \text{ fail at step 2}].$$

It suffices to bound $\Pr[\Pi' \text{ fail at step 2}]$. From the union bound, we have:

$$\begin{aligned}\Pr[\Pi' \text{ fail at step 2}] &\leq k \cdot \Pr[|S_i| < c_i N^{1-\alpha_i}] \\ &\leq k \cdot \Pr[\#\text{repeated elements in } S_i > (c - c_i)N^{1-\alpha_i}].\end{aligned}$$

We know that

$$\mathbb{E}[\#\text{repeated elements}] = cN^{1-\alpha_i}\left(1 - (1 - 1/N)^{cN^{1-\alpha_i}-1}\right) \leq c^2 N^{1-2\alpha_i}.$$

From Markov's Inequality, we have

$$\Pr[\#\text{repeated elements} > (c - c_i)N^{1-\alpha_i}] \leq \mathbb{E}[\#\text{repeated elements}]/(c - c_i)N^{1-\alpha_i} \leq kc^2 N^{-\alpha_i}.$$

If $kN^{-\alpha_i} \leq \frac{1}{100k}$ holds, which is guaranteed by the constraints, $\Pr[\Pi' \text{ fail at step 2}] \leq 0.05$ also holds. This concludes the lemma. □

4 Efficient protocols for the hardness distribution

In this section, we first explain an efficient protocol for the hardness distribution 3, where we use D_3 to denote the distribution, showing that our lower bound result is almost tight for this distribution. Also, this protocol can be easily extended to some more general product distributions sharing *"similarities"* with the Bernoulli product distribution. Formally, we prove:

Theorem 4.1. *There is a protocol Π, which solves the hardness distribution 3, with $Acc_\Pi(D_3) \geq 0.1$ and*

$$CC(\Pi) = O(N^{\sum_i \alpha_i - \max_i\{\alpha_i\}} \log N).$$

Furthermore, this protocol can be extended to more general distributions. Let D be any distribution that satisfies the following properties:

1. *each party holds a set of size $\Theta(N^{1-\alpha_i})$;*

2. *the size of intersecting part of all parties is $\Omega(N^{1-\sum_i \alpha_i})$;*

there exists a protocol Π' with $O(k \log N \cdot N^{\sum_i \alpha_i - \max_i\{\alpha_i\}})$ communication cost that achieves $\Omega(1)$ accuracy under D.

Proof. To begin with, we first propose an efficient protocol to solve D_3. Without loss of generality, we assume $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_k$ and each party i gets a subset $S_i \subseteq [N]$. Then, the communication protocol Π proceeds as follows:

1. The first party uniformly and randomly picks $\min\{|S_1|, N^{\sum_i \alpha_i - \max_i\{\alpha_i\}}\}$ elements from S_1 and sends them, denoted by M_1, to the second party.
2. The second party receives the message M_1 from the first one, and sends $M_2 := M_1 \bigcap S_2$ to the third party.
3. The process goes on, and the last party computes $M_{k-1} \bigcap S_k$. If it is not empty, the last party outputs any element in it. Otherwise, the protocol fails.

Then, we bound $\mathrm{Acc}_\Pi(D_3)$ and its communication complexity to show Π is highly efficient. From the definitions, we know that

$$\mathrm{Acc}_\Pi(D_3) = \Pr[M_1 \cap S_2 \cap \cdots \cap S_k \neq \emptyset].$$

Also, we have that

$$\Pr[M_1 \cap S_2 \cap \cdots \cap S_k \neq \emptyset \mid |M_1| = m] = 1 - \left(1 - \frac{1}{N^{\sum_i \alpha_i - \max_i\{\alpha_i\}}}\right)^m \geq \frac{m}{e \cdot N^{\sum_i \alpha_i - \max_i\{\alpha_i\}}}.$$

The last inequality holds since $m \leq N^{\sum_i \alpha_i - \max_i\{\alpha_i\}}$. From Chernoff bound, we know that the probability that $\Pr\left[|M_1| \leq N^{\sum_i \alpha_i - \max_i\{\alpha_i\}}/2\right] \leq e^{-N^{1-\alpha_1}/12} \leq e^{-10k^3}$. The last inequality is from the constraint of $k \leq 0.1 \cdot \min\{N^{\min_i\{\alpha_i\}/2}, N^{(1-\max_i\{\alpha_i\})/3}\}$. Furthermore, when

$$|M_1| \geq N^{\sum_i \alpha_i - \max_i\{\alpha_i\}}/2,$$

it holds that

$$\Pr[M_1 \cap S_2 \cap \cdots \cap S_k \neq \emptyset \mid |M_1| = m] \geq \frac{1}{2e}.$$

Combining the facts above, we have $\mathrm{Acc}_\Pi(D_3) \geq \frac{1}{2e}\left(1 - e^{-10k^3}\right) \geq 0.1$.

On the other hand, we bound the communication complexity by bounding the expected size of $|M_i|$. $\mathbb{E}[|M_1|] \leq N^{\sum_i \alpha_i - \max_i\{\alpha_i\}} \log N$ holds from definitions. Furthermore, we have

$$\mathbb{E}[|M_i|] \leq \mathbb{E}[|M_{i-1}|] \cdot N^{-\alpha_i}.$$

Then, $\mathbb{E}[\sum_i M_i] \leq O(N^{\sum_i \alpha_i - \max_i\{\alpha_i\}} \log n)$ follows by $N^{-\alpha_i} \leq N^{\min_i\{\alpha_i\}} \leq 1/2$. This concludes the first statement.

Next, we slightly change the protocol above to match the second statement. The protocol Π' proceeds as follows:

1. The first party uniformly and randomly picks $\Theta(N^{\sum_i \alpha_i - \max_i\{\alpha_i\}})$ elements from S_1 and sends them, denoted by M_1, to the second party.

2. The second party receives the message M_1 from the first one, and sends $M_2 := M_1 \bigcap S_2$ to the third party.

3. The process goes on, and the last party computes $M_{k-1} \bigcap S_k$. If it is not empty, the last party outputs any element in it.

Obviously, the communication complexity of this protocol Π' is $O(k \log n \cdot N^{\sum_i \alpha_i - \max_i\{\alpha_i\}})$. Also, we know the accuracy is bigger than

$$\Omega\left(1 - (1 - \frac{\Omega(N^{1-\sum_i \alpha_i})}{|S_1|})^{\Theta(N^{\sum_i \alpha_i - \max_i\{\alpha_i\}})}\right) = \Omega(1).$$

□

Thus, our lower bounds show that those trivial protocols are nearly optimal.

References

[BFM18] Balthazar Bauer, Pooya Farshim, and Sogol Mazaheri. Combiners for backdoored random oracles. In *Advances in Cryptology–CRYPTO 2018: 38th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 19–23, 2018, Proceedings, Part II 38*, pages 272–302. Springer, 2018. 3, 6, 7

[BFS86] Laszlo Babai, Peter Frankl, and Janos Simon. Complexity classes in communication complexity theory. In *27th Annual Symposium on Foundations of Computer Science (sfcs 1986)*, pages 337–347, 1986. 2

[CDGS18] Sandro Coretti, Yevgeniy Dodis, Siyao Guo, and John Steinberger. Random oracles and non-uniformity. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 227–258. Springer, 2018. 4, 7, 8

[CFK+19] Arkadev Chattopadhyay, Yuval Filmus, Sajin Koroth, Or Meir, and Toniann Pitassi. Query-To-Communication Lifting for BPP Using Inner Product. In Christel Baier, Ioannis Chatzigiannakis, Paola Flocchini, and Stefano Leonardi, editors, *46th International Colloquium on Automata, Languages, and Programming (ICALP 2019)*, volume 132 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 35:1–35:15, Dagstuhl, Germany, 2019. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. 5, 6, 7, 8

[CP10] Arkadev Chattopadhyay and Toniann Pitassi. The story of set disjointness. *SIGACT News*, 41(3):59–85, sep 2010. 2

[DFMT20] Yevgeniy Dodis, Pooya Farshim, Sogol Mazaheri, and Stefano Tessaro. Towards defeating backdoored random oracles: indifferentiability with bounded adaptivity. In *Theory of Cryptography Conference*, pages 241–273. Springer, 2020. 4

[DKO12] Andrew Drucker, Fabian Kuhn, and Rotem Oshman. The communication complexity of distributed task allocation. In *Proceedings of the 2012 ACM symposium on Principles of distributed computing*, pages 67–76, 2012. 4

[DLP12] Danny Dolev, Christoph Lenzen, and Shir Peled. "tri, tri again": finding triangles and small subgraphs in a distributed setting. In *International Symposium on Distributed Computing*, pages 195–209. Springer, 2012. 4

[DOR21] Nachum Dershowitz, Rotem Oshman, and Tal Roth. The communication complexity of multiparty set disjointness under product distributions. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1194–1207, 2021. 2, 3, 6

[GPW17] Mika Göös, Toniann Pitassi, and Thomas Watson. Query-to-communication lifting for bpp. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 132–143, 2017. 5, 6, 7, 8

[HPZZ21] Dawei Huang, Seth Pettie, Yixiang Zhang, and Zhijun Zhang. The communication complexity of set intersection and multiple equality testing. *SIAM Journal on Computing*, 50(2):674–717, 2021. 4, 6

[ILG17] Taisuke Izumi and François Le Gall. Triangle finding and listing in congest networks. In *Proceedings of the ACM Symposium on Principles of Distributed Computing*, pages 381–389, 2017. 4

[LMM$^+$22] Shachar Lovett, Raghu Meka, Ian Mertz, Toniann Pitassi, and Jiapeng Zhang. Lifting with sunflowers. *Leibniz international proceedings in informatics*, 215, 2022. 5, 6, 7, 8

[OR23] Rotem Oshman and Tal Roth. The Communication Complexity of Set Intersection Under Product Distributions. In Kousha Etessami, Uriel Feige, and Gabriele Puppis, editors, *50th International Colloquium on Automata, Languages, and Programming (ICALP 2023)*, volume 261 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 95:1–95:20, Dagstuhl, Germany, 2023. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. 6

[She14] Alexander A. Sherstov. Communication complexity theory: Thirty-five years of set disjointness. In Erzsébet Csuhaj-Varjú, Martin Dietzfelbinger, and Zoltán Ésik, editors, *Mathematical Foundations of Computer Science 2014*, pages 24–43, Berlin, Heidelberg, 2014. Springer Berlin Heidelberg. 2

A Omited Proofs

In this section, we prove lemma 2.6.

Lemma 2.6. *If $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_\ell$ are $\ell < k$ independent $\left(1 - \frac{1}{10k \log N}\right)$-dense random variables and each $\boldsymbol{X}_i$ takes value from $[N]^{J_i}$ with $|J_i| \le c \cdot N^{1-\alpha_i}$, where c is a constant and $N^{\alpha_i} = \omega(k)$, then for any element $a \in [N]$, it holds*

$$\Pr\left[a \in \bigcap_{i=1}^{\ell} \boldsymbol{X}_i\right] \le \frac{ec^\ell}{N^{\sum_i \alpha_i}},$$

here $e \approx 2.7$ denotes the Euler's number.

Proof. We know that all $\boldsymbol{X}_i$'s are independent. Thus, we first bound the probability that $\Pr[a \in \boldsymbol{X}_i]$. Assuming that $J_i = (j_1, j_2, \cdots, j_{|J_i|})$, we then prove for any $p \le |J_i|$

$$\Pr[a \notin \bigcap_{q=1}^{p} \boldsymbol{X}_i(j_q)] \ge \left(1 - \frac{1+1/k}{N}\right)^p$$

by induction. Here, $\boldsymbol{X}_i(j_q)$ denotes the value of $\boldsymbol{X}_i$ on the coordinate j_q.

1. When $p = 1$, we have this inequality directly from the fact that $\boldsymbol{X}_i$ is $(1 - \frac{1}{10k \log N})$-dense.
2. When $p > 1$, we assume this inequality holds for $p - 1$. In that case,

$$\Pr[a \notin \bigcap_{q=1}^{p} \boldsymbol{X}_i(j_q)] = \left(1 - \Pr[a = \boldsymbol{X}_i(j_p) \mid a \notin \bigcap_{q=1}^{p-1} \boldsymbol{X}_i(j_q)]\right) \cdot \Pr[a \notin \bigcap_{q=1}^{p-1} \boldsymbol{X}_i(j_q)].$$

It suffices to show that $\Pr[a = \boldsymbol{X}_i(j_p) \mid x \notin \bigcap_{q=1}^{p-1} \boldsymbol{X}_i(j_q)] \le \frac{1+1/k}{N}$. If we assume

$$\Pr[a = \boldsymbol{X}_i(j_p) \mid a \notin \bigcap_{q=1}^{p-1} \boldsymbol{X}_i(j_q)] > \frac{1+1/k}{N}$$

holds, we have:

$$\begin{aligned}
\Pr[a = \boldsymbol{X}_i(j_p)] &= \Pr\left[a = \boldsymbol{X}_i(j_p) \mid a \notin \bigcap_{q=1}^{p-1} \boldsymbol{X}_i(j_q)\right] \cdot \Pr[a \notin \bigcap_{q=1}^{p-1} \boldsymbol{X}_i(j_q)] \\
&\ge \left(1 - \frac{1+1/k}{N}\right)^{cN^{1-\alpha_i}} \frac{1+1/k}{N} \\
&\ge \left(1 - \frac{2c}{N^{\alpha_i}}\right) \frac{1+1/k}{N} \\
&\ge \frac{1+1/(2k)}{N}.
\end{aligned}$$

Here, the last inequality holds from the fact that $N^{\alpha_i} = \omega(k)$. This contradicts with the fact that $\boldsymbol{X}_i$ is $\left(1 - \frac{1}{(10k \log N)}\right)$-dense. Hence, we know that

$$\Pr[a \notin \boldsymbol{X}_i] \ge (1 - \frac{1+1/k}{N})^{|J_i|} \ge (1 - \frac{1+1/k}{N})^{cN^{1-\alpha_i}} \ge 1 - c \cdot \frac{1+1/k}{N^{\alpha_i}},$$

and

$$\Pr[a \in \cap_i \boldsymbol{X}_i] \le \prod_i \Pr[a \in \boldsymbol{X}_i] \le c^\ell (1 + 1/k)^\ell \cdot \frac{1}{N^{\sum_i \alpha_i}} \le c^\ell \frac{e}{N^{\sum_i \alpha_i}}.$$

□