Electronic Colloquium on Computational Complexity, Report No. 159 (2023)

Communication Lower Bounds for Collision Problems via Density Increment Arguments

Guangxu Yang * guangxuy@usc.edu

Jiapeng Zhang * jiapengz@usc.edu

October 31, 2023

Abstract

Collision problems are important problems in complexity theory and cryptography with diverse applications. Previous fruitful works have mainly focused on query models. Driven by various applications, several works by Bauer, Farshim and Mazaheri (CRYPTO 2018), Itsykson and Riazanov (CCC 2021), Göös and Jain (RANDOM 2022) independently proposed the communication version of collision problems.

In the communication setting, both Alice and Bob receive k uniformly random sets: $S_1, \ldots, S_k$ and $T_1, \ldots, T_k$ with each of size roughly $\sqrt{N}$, where a typical choice of k is in the order of $\sqrt{N}$ for applications. Then Alice and Bob aim to find a pair (x, x') such that $x, x' \in S_i \cap T_j$ for some S_i and T_j. A simple protocol that solves this problem with $\widetilde{O}(N^{1/4})$ communication bits can be the following: Alice sends to Bob a random subset of S_1 of size $N^{1/4}$ and Bob checks if there is a set T_j that has more than two intersections to this subset. All the papers mentioned above believe this bound should be tight up to some log factors.

In this paper, we prove an $\widetilde{\Omega}(N^{1/4})$ randomized communication lower bound, affirming the conjecture above. Previously, only an $\widetilde{\Omega}(N^{1/12})$ was known by a work of Göös and Jain (RANDOM 2022). Our lower bound provides direct applications to cryptography and proof complexity via connections by Bauer, Farshim, and Mazaheri (CRYPTO 2018) and Itsykson and Riazanov (CCC 2021).

Our proof technique could be of independent interest as it is an extension of simulation methods to non-lifted functions. Previously, simulations have been widely applied to lifted functions (a.k.a composed functions), which leads to beautiful query-to-communication lifting theorems. However, many important communication problems are not lifted functions. We believe our methods could give more applications. In particular, it may have applications to communication search problems with many solutions. Note that many existing methods do not apply to this setting.

1 Introduction

Collision problems are important problems in theoretical computer science with wide applications in quantum complexity [Aar02], streaming complexity [Din20, LZ23], cryptography [LZ17, BFM18],

*Thomas Lord Department of Computer Science, University of Southern California. Research supported by NSF CAREER award 2141536.

ISSN 1433-8092

property testing [BHH11], quantum algorithm [MSS07], proof complexity [IR21], distributed computing [GHK13] and approximate counting [AKKT19]. Previous research on collision problems has mainly focused on query models.

However, the communication version of collision problems was not widely studied until several applications have been identified recently. In this paper, we study two communication versions of collision problems, giving direct applications to cryptography and proof complexity.

Cryptography motivations. To analyze the security of cryptographic hash functions, Bauer, Farshim, and Mazaheri [BFM18] formulated and studied the backdoored random-oracle (BRO) model. They showed that, via reductions to lower bounds of communication problems, central cryptographic security properties are achievable by some combiners in the BRO model. Concretely, they formalized the following multi-set double-intersection problem.

Problem 1.1 (Multi-set double-intersection [BFM18])**.** Both Alice and Bob hold $\sqrt{N}$ random sets:

- Alice independently samples sets $S_1, \ldots, S_{\sqrt{N}} \subseteq [N]$ with each element of $[N]$ contained in S_i with probability $1/\sqrt{N}$.
- Similarly, Bob independently samples $T_1, \ldots, T_{\sqrt{N}} \subseteq [N]$ with each element of $[N]$ contained in T_j with probability $1/\sqrt{N}$.

Their goal is to find a pair $s \neq s'$ such that $s, s' \in S_i \cap T_j$ for some i, j.

By a simple calculation, the sampled instances will contain such a solution pair with high probability. Assuming the hardness of the multi-set double-intersection problem, [BFM18] obtained collision-resistant combiners in the BRO model. However, proving such a communication lower bound looks challenging. Hence, they put it as an open problem.

On the other hand, based on the birthday paradox, a simple protocol solving this problem with $\widetilde{O}(N^{1/4})$ communication bits can be:

1. Alice sends to Bob a random subset of S_1 with size of $N^{1/4}$.
2. Bob checks if there is a set T_j that has more than two intersections to this subset.

[BFM18] believed this simple algorithm could be the best attacker of multi-set double intersections. They made the following conjecture.

Conjecture 1.2 ([BFM18])**.** *The randomized communication complexity of Problem 1.1 is $\widetilde{\Omega}(N^{1/4})$.*

Proof complexity motivations. The connections between communication complexity and proof complexity have been extensively studied for many years. To study proof complexity lower bounds for natural formulas, Itsykson and Riazanov [IR21] introduced a communication search problem called the bit-pigeonhole principle problem.

Problem 1.3 (Bit-pigeonhole principle problem [IR21])**.** For $N < M$, the bipartite communication search problem BPHP_N^M is defined below,

- Alice holds $x = (x_1, \ldots, x_M) \in [\sqrt{N}]^M$;

- Bob holds $y = (y_1, \ldots, y_M) \in [\sqrt{N}]^M$;

The goal is to find a pair of distinct coordinates $i, i' \in [M]$ such that $x_i = x_{i'}$ and $y_i = y_{i'}$. We call those collision pairs, or simply collisions.

BPHP_N^M is a total search problem and always has a collision for $N < M$. [IR21] proved that $\Omega(\sqrt{N})$ randomized communication lower bound for BPHP_N^{N+1} via a randomised reduction from set-disjointness. This lower bound implies that any proof system that randomized protocols can efficiently simulate requires exponential size to refute bit-pigeonhole formulas featuring $M = N+1$ pigeons and N holes. Since then, a later result by Göös and Jain [GJ22] showed an $\Omega(N^{1/12})$ lower bound for BPHP_N^{2N}. Both [IR21] and [GJ22] are interested in similar communication lower bounds for the weak pigeonhole principle with arbitrary $M > N$ pigeons and N holes. For $M = 2N$, Göös and Jain [GJ22] proved the following lower bound.

Theorem 1.4 ([GJ22])**.** *The randomized communication complexity of* BPHP_N^{2N} *are* $\Omega(N^{1/12})$.

Built on the birthday paradox again, a simple protocol (see [GJ22]) also solves BPHP_N^{2N} with $\widetilde{O}(N^{1/4})$ communication bits. To this end, a natural question arises:

Conjecture 1.5 ([IR21, GJ22])**.** *The randomized communication complexity of* BPHP_N^{2N} *is* $\widetilde{\Omega}(N^{1/4})$.

1.1 Our Contribution

Our contribution is two-fold. Firstly, we affirm Conjecture 1.2 and Conjecture 1.5, giving several direct applications in both cryptography and proof complexity. On the other hand, our proof technique can be considered as an extension of query-to-communication lifting theorems to general functions without a composed form. Lifting theorem is a nice idea developed in recent years with diverse applications in a lot of areas. However, lifting theorems have mainly focused on applications with lifted functions previously. In this paper, we aim to extend these applications to broader functions. We prove the following two theorems.

Theorem 1.6. *For any randomized communication protocol that solves the multi-set double-intersection problem with constant probability, it must communicate* $\Omega(N^{1/4})$ *bits.*

Theorem 1.7. *For any* $M > N$*, the randomized communication complexity of* BPHP_N^M *is* $\Omega(N^{1/4})$.

Theorem 1.7 holds for any $M > N$, which is an extension of [IR21] ($M = N+1$) and [GJ22] ($M = 2 \cdot N$). Furthermore, this lower bound is tight for $M = (1 + \Omega(1)) \cdot N$ (up to some logarithmic factors), matching the upper bound protocol by [GJ22] [1].

Applications: Our results directly give some applications in cryptography and proof complexity by the connections built by [BFM18, IR21, GJ22].

1. Since we affirm the hardness assumptions by [BFM18], collision-resistance combiners for backdoored random oracles could be obtained directly through the reduction by [BFM18].

[1] Their protocol is for $M = 2 \cdot N$. But it can be extended to any $M = (1 + \Omega(1)) \cdot N$

2. Using the reductions by [IR21], we directly show that any proof system that can be efficiently simulated by randomized protocols (most notably, tree-like Res(⊕) [IS20]) requires exponential size to refute bit-pigeonhole formulas featuring M pigeons and N holes for arbitrary $M > N$, which answer the open problem in [IR21]. Furthermore, our communication lower bound is tight for a large range of M and N.

3. Building on connections by [IPU94, IR21], our result implies that every tree-like cutting planes of the weak bit pigeon hole principle BPHP_N^M, $M > N$, has size $2^{\Omega(N^{1/4})}$. It improves the lower bound of $2^{\Omega(N^{1/8})}$ by Hrubeš and Pudlák [HP17]. We note that [HP17]'s lower bound also holds for non-tree-like CPs. Hence, our improvement only applies to tree-like CPs.

4. Besides direct applications, it is also interesting to check if our method gives communication lower bounds for deterministic dag-like protocols. This would provide further applications in proof complexity such as non-tree-like CPs lower bounds [HP17].

1.2 Proof Outline

We now give a high-level description of our proof to Theorem 1.7. In order to prove randomized communication lower bounds, it is sufficient to show that any *deterministic protocol* with a small amount of communication bits can not find collisions under the following distribution.

Both Alice and Bob's inputs are uniformly sampled from $[\sqrt{N}]^M$.

It is well-known that any deterministic communication protocol corresponds to a partition of the input space, denoted by $\mathcal{R}^{\text{leaf}}$. Inspired by simulation methods in lifting theorems, our idea is to further partition the rectangles in $\mathcal{R}^{\text{leaf}}$ into many structures defined below.

Definition 1.8. Let $R = X \times Y \subseteq [\sqrt{N}]^M \times [\sqrt{N}]^M$ be a rectangle, and let $J_1, J_2 \subseteq [M]$. We say that $S := (R, J_1, J_2)$ is a structure if,

- X is fixed on $J_1^c := [M] \setminus J_1$, i.e., for every $i \in J_1^c$, there is an $s_i \in [\sqrt{N}]$ such that $x_i = s_i$ for all $x \in X$.
- Y is fixed on J_2^c, i.e., for every $i \in J_2^c$, there is an $r_i \in [\sqrt{N}]$ such that $y_i = r_i$ for all $y \in Y$.
- For all $i, i' \in J_1^c$, if $s_i = s_{i'}$, then $y_i \neq y_{i'}$ for all $y \in Y$.
- For all $i, i' \in J_2^c$, if $r_i = r_{i'}$, then $x_i \neq x_{i'}$ for all $x \in X$.

We denote $|S| = |R|$ and we say that a rectangle R is a structure if there is a pair (J_1, J_2) such that (R, J_1, J_2) is a structure.

Our proof includes two steps.

- We first show that $R \in \mathcal{R}^{\text{leaf}}$ can be covered by combinations of *pseudorandom* structures.
- Then we show that protocols can not find collision pairs from such pseudorandom structures.

The last two constraints of structures ensure that there is no collision in J_1^c and J_2^c. In order to prevent the protocol from finding collisions from J_1 and J_2, we borrow the dense notion from query-to-communication lifting theorems to capture pseudorandomness. Roughly speaking, we say that X is dense if, for every $I \subseteq J_1$, the marginal distribution X_I has a very high min-entropy. Similarly, we can define it for Y. Building on the dense notion, we prove the following claim.

Claim 1.9 (Informal)**.** *Any protocol can not find a pair of collisions from a dense structure.*

The formal version of this claim is Claim 3.10. Our next step is to show that for a communication protocol with $o(N^{1/4})$ communication bits, the corresponding leaf rectangles can be almost covered by dense structures.

Claim 1.10 (Informal)**.** *Let $\mathcal{R}^{\text{leaf}}$ be a partition associated with a communication protocol. We can further partition each $R \in \mathcal{R}^{\text{leaf}}$ into many smaller rectangles*

$$R = S_1 \cup \cdots \cup S_{\ell_1} \cup B_1 \cup \cdots \cup B_{\ell_2}$$

where $S_1, \ldots, S_{\ell_1}$ are rectangles with the form of pseudorandom structures and $B_1, \ldots, B_{\ell_2}$ could be arbitrary. If the communication complexity of the protocol is small, we then show that the union of all B rectangles is small compared to the input space.

The proof of this claim is inspired by the simulation process in lifting theorems. However, there are two important conceptual differences:

- The purpose of simulations in lifting theorems is to convert a communication protocol to a decision tree. By contrast, our process only decomposes rectangles in $\mathcal{R}^{\text{leaf}}$ into dense structures.
- The simulation in lifting theorems needs a gadget. However, our decomposition only focuses on the structure of rectangles. This enables us to apply simulations for more general non-composed functions.

Connections of BPHP_N^M and multi-set double intersection. For each pair of inputs of BPHP_N^M $x, y \in [\sqrt{N}]^M$, we convert it into a collection of sets

$$\forall i \in [\sqrt{N}], S_i = \{j \in [M] : x_j = i\} \text{ and } T_i = \{j \in [M] : y_j = i\}$$

The definition for BPHP_N^M under uniform distribution can be reformulated as follows:

Problem 1.11 (Restated)**.** For $N < M$,

- Alice samples sets $S_1, \ldots, S_{\sqrt{N}} \subseteq [M]$ with each $s \in [M]$ uniformly assigned to a set S_i.
- Bob samples sets $T_1, \ldots, T_{\sqrt{N}} \subseteq [M]$ with each $s \in [M]$ uniformly assigned to a set S_i.

Their goal is to find a pair $s \neq s'$ such that $s, s' \in S_i \cap T_j$ for some i, j.

Under this interpretation, the only difference between BPHP_N^M and multi-set double intersection is that BPHP_N^M promises that each element is contained in exactly one set but in multi-set double intersection each element is independently sampled for each set. The two distributions are generally similar, and our proof can be applied to multi-set double intersection directly.

Technical contribution and previous barriers. At first glance, the collision problem looks hard to many existing lower bound methods since it is a search problem with many solutions. For random sets S_i and T_j, it has that $|S_i \cap T_j| \geq 2$ with a constant probability. Hence, it expects to have $\Omega(k^2) = \Omega(N)$ pairs of solutions in collision problems. To the best of our knowledge, many existing communication lower bound techniques do not apply to this setting.

To overcome this barrier, Göös and Jain [GJ22] introduced the query-to-communication lifting approach. Concretely, Göös and Jain proposed a new communication problem $\mathsf{Col}_N \circ \mathsf{Ver}^N$, where Col_N is the query version of a collision problem and Ver is a small-size gadget. They proved a BPHP_N^{2N} (Problem 1.3) lower bound via two steps:

1. The communication complexity of $\mathsf{Col}_N \circ \mathsf{Ver}^N$ is $\Omega(N^{1/3})$
2. Builds on $\mathsf{Col}_N \circ \mathsf{Ver}^N$, [GJ22] proves an $\Omega(N^{1/12})$ lower bound for BPHP_N^{2N}via reductions.

Since there is a loss in the reduction [GJ22], the limitation of their framework is an $\Omega(N^{1/6})$ lower bound.

The notion of query-to-communication lifting theorems is a remarkable technique introduced recently [GPW15, GPW17, CFK+19, LMM+22] to prove communication complexity lower bounds with a wide variety of applications in many areas. However, despite many applications, one of the main limitations of lifting theorems is that: it only applies to lifted functions, i.e., a function has the form $f \circ g^n$ where f is a query problem and g is a small-size gadget, such as $\mathsf{Col}_N \circ \mathsf{Ver}^N$. Since collision problems such as BPHP cannot be written as lifted functions directly, lifting theorems can not be applied directly. This is also the main reason that [GJ22] introduced the $\mathsf{Col}_N \circ \mathsf{Ver}^N$ problem and proved BPHP lower bounds through $\mathsf{Col}_N \circ \mathsf{Ver}^N$. However, the reduction caused a loss making the lower bound of BPHP not tight.

By contrast, our proof is not built on reductions. We extend the simulation method (the idea used to prove lifting theorems) into broader functions that do not have the lifted form. Besides the collision problems, we believe that our approach may enable more applications.

2 Preliminary

We first fix some of the notations through this paper.

- A large domain $[N]$ and we set $k = \sqrt{N}$.
- Alice and Bob receive inputs $x \in [k]^M$ and $y \in [k]^M$ respectively. They hope to find a pair (j, j') such that $x_j = x_{j'}$ and $y_j = y_{j'}$.
- We use letters X, Y to denote subsets of $[k]^M$. For a set X, we use the bold font $\boldsymbol{X}$ to denote the uniform distribution on X.
- For a set $J \subseteq [M]$, we use $J^c := [M] \setminus J$ to denote its complement.
- We use $\boldsymbol{X}_J$ to denote the marginal distribution of $\boldsymbol{X}$ on J.

Definition 2.1 (Dense)**.** Let $\boldsymbol{D}$ be a random variable on $[k]^M$. We say that $\boldsymbol{D}$ is γ-dense on J if for every subset $I \subseteq J$ it holds that

$$H_\infty(\boldsymbol{D}_I) \geq \gamma \cdot |I| \cdot \log k$$

This notation often appears in the query-to-communication lifting theorem [GPW17, CFK+19]. However, in this paper, we set $\gamma = 1 - \frac{1}{\log k}$ which is different from previous methods where they usually set $\gamma = 0.9$.

Definition 2.2. For a distribution $\boldsymbol{X}$ and a set $J \subseteq [M]$, we define its density-loss by,

$$\mathcal{D}_\infty(\boldsymbol{X}, J) = \log(k^{|J|}) - H_\infty(\boldsymbol{X}_J) = |J| \cdot \log k - H_\infty(\boldsymbol{X}_J)$$

For a tuple $(X \times Y, J_1, J_2)$, we define its density-loss by

$$\mathcal{D}_\infty(X \times Y, J_1, J_2) = \mathcal{D}_\infty(\boldsymbol{X}, J_1) + \mathcal{D}_\infty(\boldsymbol{Y}, J_2)$$

We note that the density-loss is non-negative and $\mathcal{D}_\infty(\boldsymbol{X}, J) = 0$ if and only if $\boldsymbol{X}_J$ is uniform.

3 Proof of the Main Theorem

We now prove Theorem 1.7. We first recall the setting. In this problem, Alice and Bob receive (uniform sampled) inputs $x, y \in [k]^M$ and they want to find a pair of distinct coordinates $i, i' \in [M]$ such that $x_i = x_{i'}$ and $y_i = y_{i'}$. We aim to prove an $\Omega(N^{1/4})$ lower bound for this problem. As we briefly mentioned in Section 1.2, a crux in our proof is to decompose each leaf rectangle into dense structures (Definition 1.8).

3.1 Decomposition Process

In this section, we discuss the decomposition process, which is the crucial step in our proof. We need to partition each rectangle $R \in \mathcal{R}^{\text{leaf}}$ into a combination of dense structures and some error rectangles that may have collisions. We first introduce the following density-restoring partition lemma which is similar to the density-restoring partition in [GPW17, CFK+19].

Lemma 3.1 (Density-restoring partition)**.** *Let $S = (X \times Y, J_1, J_2)$ be a structure. If $\boldsymbol{Y}$ is further γ-dense on J_2, then there is a partition of $X \times Y$,*

$$X \times Y = X^1 \times Y^1 \cup X^1 \times Y^1_{\text{error}} \cup \cdots \cup X^t \times Y^t \cup X^t \times Y^t_{\text{error}}$$

such that every X^i is associated with a set $I_i \subseteq J_1$ and $p_{\geq i} := \frac{|\bigcup_{j \geq i} X^j|}{|X|}$ satisfies the following properties:

1. *For every i, $S^i := (X^i \times Y^i, J_1 \setminus I_i, J_2)$ is a structure*
2. *The tuple $B^i := (X^i \times Y^i_{\text{error}}, J_1 \setminus I_i, J_2)$ could be arbitrary*
3. *For every i, $\boldsymbol{X}^i$ is γ-dense on $J_1 \setminus I_i$*
4. *$\mathcal{D}_\infty(\boldsymbol{X}^i, J_1 \setminus I_i) \leq \mathcal{D}_\infty(\boldsymbol{X}, J_1) - |I_i| + \log \frac{1}{p_{\geq i}}$.*
5. *For every i,$(Y^i, Y^i_{\text{error}})$ is a partition of Y such that $|Y^i_{\text{error}}|/|Y| \leq 2 \cdot (|I_i \cup J_1^c|^2 - |J_1^c|^2)/k$.*

Similarly, If $\boldsymbol{X}$ is γ-dense on J_1, analogous conclusions hold for the partition of R with the roles of $\boldsymbol{X}$ and $\boldsymbol{Y}$ interchanged.

The proof of Lemma 3.1 builds on two steps. We first apply the density-restoring partition lemma from [GPW17], decomposing $X \times Y$ to

$$X \times Y = X^1 \times Y \cup \cdots \cup X^t \times Y$$

such that for every i, $\boldsymbol{X}^i$ is γ-dense on $J_1 \setminus I_i$ and $\mathcal{D}_\infty(\boldsymbol{X}^i, J_1 \setminus I_i) \le \mathcal{D}_\infty(\boldsymbol{X}, J_1) - |I_i| + \log \frac{1}{p_{\ge i}}$. However, these tuples $(X^1 \times Y, J_1 \setminus I_1, J_2), \ldots, (X^t \times Y, J_1 \setminus I_t, J_2)$ are not necessarily being structures. We then further decompose Y into $(Y^i, Y^i_{\text{error}})$ by moving the collision part to Y^i_{error}. We defer the formal proof to Section A.

Recursive decomposition. Building on Lemma 3.1, we now describe our decomposition process. We first introduce some notations. Let Π be a fixed protocol tree.

- We use $\mathcal{R}^j$ to denote the rectangles which associated with nodes in j-th depth of the protocol tree. Note that $\mathcal{R}^0 = \{[k]^M \times [k]^M\}$ contains only the root and $\mathcal{R}^{\text{leaf}}$ contains all leaves.
- We recursively (from the root to the leaves) decompose each rectangle associated with a node in the tree. For each R, we decompose it as two parts: $\mathcal{S}(R) = \{(S, J_1, J_2)\}$ and $\mathcal{B}(R) = \{(B, J_1, J_2)\}$, where each (S, J_1, J_2) is a structure and (B, J_1, J_2) could be arbitrary. We also denote $\mathcal{L}(R) = \mathcal{S}(R) \cup \mathcal{B}(R)$.
- For j, let $\mathcal{S}^j := \bigcup_{R \in \mathcal{R}^j} \mathcal{S}(R)$, $\mathcal{B}^j := \bigcup_{R \in \mathcal{R}^j} \mathcal{B}(R)$ and $\mathcal{L}^j := \bigcup_{R \in \mathcal{R}^j} \mathcal{L}(R)$ be the union of all decomposition in the rectangles in the depth-j [2] respectively.

Let us explain our recursive decomposition process.

1. For the root, we simply let $\mathcal{S}([k]^M \times [k]^M) = \{([k]^M \times [k]^M, \emptyset, \emptyset)\}$ and $\mathcal{B}([k]^M \times [k]^M) = \{\}$.
2. For internal nodes, let R be a rectangle with known decomposition $\mathcal{S}(R)$ and $\mathcal{B}(R)$, and let R^0 and R^1 be the children of R. In order to obtain $\mathcal{S}(R^0)$, we simply apply the density-restoring partition lemma (Lemma 3.1) on $(S \cap R^0, J_1, J_2)$ for each $(S, J_1, J_2) \in \mathcal{S}(R)$.

We formalize this process as Algorithm 1 below.

```
Algorithm 1: Decomposition Algorithm (when Alice is speaking)
  Input: A rectangle $R = X \times Y$ and its decomposition $\mathcal{S}(R)$ and $\mathcal{B}(R)$
  Output: Output $\mathcal{S}(R^0), \mathcal{B}(R^0)$ and $\mathcal{S}(R^1), \mathcal{B}(R^1)$, where $R^0, R^1$ are children of $R$ in the tree
  Initialize $\mathcal{S}(R^0), \mathcal{B}(R^0), \mathcal{S}(R^1), \mathcal{B}(R^1) \leftarrow \emptyset$.
  for each $(S, J_1, J_2) \in \mathcal{S}(R)$ do
      For each $b \in \{0, 1\}$, we decompose $(S \cap R^b)$ as $S^{b,1} \cup B^{b,1} \cup \cdots \cup S^{b,t} \cup B^{b,t}$ (Lemma 3.1)
      Update $\mathcal{S}(R^b) \leftarrow \mathcal{S}(R^b) \cup \{S^{b,1}, ..., S^{b,1}\}$ and $\mathcal{B}(R^b) \leftarrow \mathcal{B}(R^b) \cup \{B^{b,1}, ..., B^{b,t}\}$.
  for each $(B, J_1, J_2) \in \mathcal{B}(R)$ do
      For each $b \in \{0, 1\}$, update $\mathcal{B}(R^b) \leftarrow \mathcal{B}(R^b) \cup \{(B \cap R^b, J_1, J_2)\}$.
```

Following the discussion above, an important step in our analysis is to upper bound the size $\mathcal{B}^{\text{leaf}}$. As it has shown in the density-resorting lemma, whenever we put a tuple (B, J_1, J_2) into $\mathcal{B}^j$ for some j, the size of B can be upper bounded by $(|J_1^c|^2 + |J_2^c|^2)/k$. Hence, we first upper the number of fixed coordinates in our analysis.

[2] We use the notations $\mathcal{S}^{\text{leaf}}$, $\mathcal{B}^{\text{leaf}}$,$\mathcal{L}^{\text{leaf}}$ for leaf rectangles.

3.2 Upper Bound the Number of Fixed Coordinates

In this subsection, we show that the average size of fixed coordinated for leaf rectangles in $\mathcal{R}^{\text{leaf}}$ is $O(\Pi)$. Firstly, we formalize the definition of the average size of fixed coordinates.

Definition 3.2. Let R be a rectangle with a decomposition into a set of tuples $\mathcal{L}$ (includes both $\mathcal{S}$ and $\mathcal{B}$). We define its average fixing size as

$$E(R; \mathcal{L}) := \sum_{(L, J_1, J_2) \in \mathcal{L}} \frac{|L|}{|R|} \cdot (|J_1^c| + |J_2^c|).$$

For nodes in depth-j of the tree, we define its average fixing size by,

$$E^j = \sum_{R \in \mathcal{R}^j} \frac{|R|}{k^{2M}} \cdot E(R, \mathcal{L}(R)).$$

The main lemma in this section is the following upper bound for E^{leaf}.

Lemma 3.3. *Given a protocol* Π, *then* $E^{\text{leaf}} = O(|\Pi|)$.

Our proof of this lemma is inspired by query-to-communication lifting theorems again. We use the following density function (aka potential function).

Definition 3.4. Let R be a rectangle with a decomposition into a set of tuples $\mathcal{L}$. We define the density function by

$$D(R; \mathcal{L}) := \sum_{(L, J_1, J_2) \in \mathcal{L}} \frac{|L|}{|R|} \cdot \mathcal{D}_\infty(L, J_1, J_2).$$

For nodes in depth-j of the tree, we define its density function by,

$$D^j = \sum_{R \in \mathcal{R}^j} \frac{|R|}{k^{2M}} \cdot D(R, \mathcal{L}(R)).$$

We adopt the density increment arguments [CFK+19, GPW17, YZ22, HMYZ23] in our proof. Concretely, we show that each communication bit increases the density function by at most $O(1)$, and each fixing of a coordinate decreases the density function by at least $\Omega(1)$. The following claim is a direct corollary of the density-restoring lemma (Lemma 3.1).

Claim 3.5. *Let* (R, J_1, J_2) *be a structure, and let the following decomposition obtained by Lemma 3.1.*

$$R = X^1 \times Y^1 \cup X^1 \times Y^1_{\text{error}} \cup \cdots \cup X^t \times Y^t \cup X^t \times Y^t_{\text{error}}.$$

Let $S^1, B^1, \ldots, S^t, B^t$ *be the corresponding tuples from Lemma 3.1. Then we have that*

$$\sum_i \left(\frac{|S^i|}{|R|} \cdot \mathcal{D}_\infty(S^i, J_1 \setminus I_i, J_2) + \frac{|B^i|}{|R|} \cdot \mathcal{D}_\infty(B^i, J_1 \setminus I_i, J_2) \right) \le \mathcal{D}_\infty(R, J_1, J_1) - \sum_i \frac{|X^i|}{|X|} \cdot |I_i| + 2.$$

We note that the lat term $\sum_i \frac{|X^i|}{|X|} \cdot |I_i| + 2$ is the density gain from fixing. We defer the detailed proof of Claim 3.5 in Section B. Now we are ready to Lemma 3.3.

Proof of Lemma 3.3. In order to prove this lemma, it is sufficient to prove that, for all $j > 0$,

$$E^j \leq 3 \cdot j - D^j.$$

Recall that $D^j \geq 0$ for all $j \geq 0$, the above inequality then implies that $E^{\text{leaf}} \leq 3 \cdot |\Pi|$. We prove the statement by induction. In the roof, it is clear that $E^0 = D^0 = 0$. Now we assume that $E^j \leq 3 \cdot j - D^j$ and aim to show that

$$E^{j+1} \leq 3 \cdot (j+1) - D^{j+1}.$$

For any rectangle $R = X \times Y \in \mathcal{R}^j$, we analyze the decomposition process in Algorithm 1.

- For each tuple $(L, J_1, J_2) \in \mathcal{L}(R)$ (either from $\mathcal{S}(R)$ or $\mathcal{B}(R)$), the decomposition algorithm (Algorithm 1) first breaks it into $(L \cap R^0, J_1, J_2)$ and $(L \cap R^1, J_1, J_2)$. Let $\boldsymbol{b}$ be a Bernoulli random variable with $\Pr[\boldsymbol{b} = b] = \frac{|R^b \cap L|}{|L|}$, we then have that

$$\begin{aligned}\sum_{b \in \{0,1\}} \Pr[\boldsymbol{b} = b] \cdot \mathcal{D}_\infty(L \cap R^b, J_1, J_2) &= \mathcal{D}_\infty(L, J_1, J_2) + \sum_{b \in \{0,1\}} \Pr[\boldsymbol{b} = b] \cdot \log \frac{1}{\Pr[\boldsymbol{b} = b]} \\ &= \mathcal{D}_\infty(L, J_1, J_2) + \mathrm{H}(\boldsymbol{b}) \leq \mathcal{D}_\infty(L, J_1, J_2) + 1.\end{aligned} \tag{1}$$

  This inequality shows that the partition step increases the density function by at most 1.

- In Step 3 and Step 5, Algorithm 1 further decomposes (by Lemma 3.1) $S \cap R^0$ and $S \cap R^1$ for those structures $(S, J_1, J_2) \in \mathcal{S}(R)$. For $b \in \{0, 1\}$, let $S^{b,1} \cup B^{b,1} \cup \cdots \cup S^{b,t} \cup B^{b,t}$ be the decomposed rectangles and let $I_1^b, \ldots, I_t^b$ be the associated sets of newly fixed coordinates in the decomposition. By Claim 3.5, we have that,

$$\sum_i \left(\frac{|S^{b,i}|}{|S \cap R^b|} \mathcal{D}_\infty(S^{b,i}, J_1 \setminus I_i^b, J_2) + \frac{|B^{b,i}|}{|S \cap R^b|} \mathcal{D}_\infty(B^{b,i}, J_1 \setminus I_i^b, J_2) \right) \leq \mathcal{D}_\infty(S \cap R^b, J_1, J_2) - \Gamma(S \cap R^b) + 2 \tag{2}$$

  Here $\Gamma(S \cap R^b) := \sum_i \left(\frac{|S^{b,i}| + |B^{b,i}|}{|S \cap R^b|} \cdot |I_i^b| \right)$ is the average size of newly fixed coordinates.

  On the other hand, we have that $\Gamma(B \cap R^0) = 0$ for all $B \in \mathcal{B}(R)$ since the decomposition does not fix new coordinates for those tuples. By the definition of $E(R, \mathcal{L}(R))$, we also have that,

$$\sum_{b \in \{0,1\}} \frac{|R^b|}{|R|} \cdot E(R^b, \mathcal{L}(R^b)) = E(R, \mathcal{L}(R)) + \sum_{b \in \{0,1\}} \frac{|R^b|}{|R|} \cdot \sum_{L \in \mathcal{L}(R)} \frac{|L \cap R^b|}{|R^b|} \cdot \Gamma(L \cap R^b).$$

By combining the Inequalities (1) and (2) and the definition of $D(R, \mathcal{L}(R))$, we have that

$$\begin{aligned}\sum_{b \in \{0,1\}} \frac{|R^b|}{|R|} \cdot D(R^b; \mathcal{L}(R^b)) &\leq D(R; \mathcal{L}(R)) - \sum_{b \in \{0,1\}} \frac{|R^b|}{|R|} \cdot \sum_{L \in \mathcal{L}(R)} \frac{|L \cap R^b|}{|R^b|} \cdot \Gamma(L \cap R^b) + 3 \\ &= D(R; \mathcal{L}(R)) - \sum_{b \in \{0,1\}} \frac{|R^b|}{|R|} \cdot E(R^b, \mathcal{L}(R^b)) + E(R, \mathcal{L}(R)) + 3.\end{aligned} \tag{3}$$

Now we take the average on all rectangles in $\mathcal{R}^j$,

$$\begin{aligned}
D^{j+1} &= \sum_{R\in\mathcal{R}^j} \frac{|R|}{k^{2M}} \cdot \sum_{b\in\{0,1\}} \frac{|R^b|}{|R|} \cdot D(R^b; \mathcal{L}(R^b)) \\
&\le \sum_{R\in\mathcal{R}^j} \frac{|R|}{k^{2M}} \cdot \left(D(R; \mathcal{L}(R)) - \sum_{b\in\{0,1\}} \frac{|R^b|}{|R|} \cdot E(R^b, \mathcal{L}(R^b)) + E(R, \mathcal{L}(R)) + 3 \right) \\
&= D^j - E^{j+1} + E^j + 3 \\
&= (D^j + E^j + 3) - E^{j+1} \\
&\le (3 \cdot j + 3) - E^{j+1} \\
&= 3 \cdot (j+1) - E^{j+1}.
\end{aligned}$$

This finishes the proof. □

3.3 Upper Bound the Success Probability of the Protocol

Now we upper bound the success probability of the protocol, i.e., we show that for any communication protocol Π with $o(N^{1/4})$ communication bits, it has that

$$\Pr_{(x,y)} \left[(i,j) \leftarrow \Pi(x,y), (x_i = x_j) \wedge (y_i = y_j) \right] = o(1).$$

Recall that the decomposition process (Algorithm 1) partition rectangles of $\mathcal{R}^{\text{leaf}}$ into $\mathcal{S}^{\text{leaf}} \cup \mathcal{B}^{\text{leaf}}$. The tuples in $\mathcal{S}^{\text{leaf}}$ are dense structures (pseudorandom part), we upper bound the success probability in $\mathcal{S}^{\text{leaf}}$ by Claim 3.10. On the other hand, for those tuples in $\mathcal{B}^{\text{leaf}}$, we simply upper bound its total size, i.e., for communication protocols with $o(N^{1/4})$ communication bits, we show that

$$\sum_{(B,J_1,J_1)\in\mathcal{B}^{\text{leaf}}} \frac{|B|}{k^{2M}} = o(1)$$

To analyze the total size of rectangles in $\mathcal{B}^{\text{leaf}}$, two key points are:

- As it has shown in the density-resorting lemma, whenever we put a tuple (B, J_1, J_2) into $\mathcal{B}^j$ for some j, the size of B can be upper bounded by $(|J_1^c|^2 + |J_2^c|^2)/k$.
- The previous section showed that the average size of fixed sets $(|J_1^c| + |J_2^c|) = O(|\Pi|)$.

However, notice that the second point does not simply imply that $\mathrm{E}[|J_1^c|^2 + |J_2^c|^2] = O(|\Pi|^2)$, so we need more careful analysis. Now, instead of only upper bounding the size of $\mathcal{B}^{\text{leaf}}$, we also upper bound the size of

$$C(R) := \{(S, J_1, J_2) \in \mathcal{S}(R) : |J_1^c| \ge \sqrt{k}/4 \text{ or } |J_2^c| \ge \sqrt{k}/4\}.$$

We define the following modified average quadratic of fixing size to help with our analysis.

Definition 3.6. Let R be a rectangle with a decomposition into a set of tuples $\mathcal{L}$. We define its modified average quadratic of fixing size as

$$Q(R;\mathcal{L}) := \sum_{(L,J_1,J_2)\in\mathcal{L}:|J_1^c|,|J_2^c|<\sqrt{k}/4} \frac{|L|}{|R|}\cdot\left(\frac{2\cdot|J_1^c|^2+2\cdot|J_2^c|^2}{k}\right) + \sum_{(L,J_1,J_2)\in\mathcal{L}:|J_1^c| \text{ or } |J_2^c|\geq\sqrt{k}/4} \frac{|L|}{|R|}\cdot 2.$$

For nodes in the depth-j, we similarly define its total modified average quadratic of fixing size by,

$$Q^j = \sum_{R\in\mathcal{R}^j} \frac{|R|}{k^{2M}}\cdot Q(R,\mathcal{L}(R)).$$

Now we upper bound the size of $\mathcal{B}^{\text{leaf}}\cup\mathcal{C}^{\text{leaf}}$ by the modified average quadratic of fixing size Q^{leaf}. For each j, we denote

$$P^j := \sum_{(L,J_1,J_2)\in\mathcal{B}^j\cup\mathcal{C}^j} \left(\frac{|L|}{k^{2M}}\right).$$

Lemma 3.7. *Given a protocol* Π, $P^{\text{leaf}}\leq Q^{\text{leaf}}$.

The proof of this lemma is based on an induction. In fact, we show that for every j, $P^j\leq Q^j$.

The idea is straightforward. For those tuples $(L,J_1,J_2)\in\mathcal{B}^j\cup\mathcal{C}^j$ with $J_1^c\geq\sqrt{k}/4$ or $J_2^c\geq\sqrt{k}/4$, we upper bound it by the second half of Q^j; For those with both $J_1^c\leq\sqrt{k}/4$ and $J_2^c\leq\sqrt{k}/4$, we upper bound it by $\frac{|L|}{k^{2M}}\cdot\left(\frac{2\cdot|J_1^c|^2+2\cdot|J_2^c|^2}{k}\right)$ according to the density-restoring lemma. We defer the details to Section C.

Now we have the last two steps to finish the proof of Theorem 1.7:

1. If the communication complexity of Π is $o(N^{1/4})$, then $Q^{\text{leaf}}=o(1)$.
2. On the other hand, if Π can find collisions with probability $\Omega(1)$, then $P^{\text{leaf}}=\Omega(1)$.

Formally, we prove the following lemmas.

Lemma 3.8. *If the communication complexity of* Π *is* $o(N^{1/4})$*, then* $Q^{\text{leaf}}=o(1)$.

The proof of this lemma is based on Lemma 3.3 and an average argument.

Proof. By Lemma 3.3,

$$\sum_{(L,J_1,J_2)\in\mathcal{L}^{\text{leaf}}} (|J_1^c|+|J_2^c|)\cdot|L|/k^{2M} = O(|\Pi|)$$

If $|\Pi|=o(N^{1/4})$, then by an average argument, we have that $(|J_1^c|+|J_2^c|)=o(N^{1/4})$ for $(1-o(1))$ fraction of tuples $(L,J_1,J_2)\in\mathcal{L}^{\text{leaf}}$. For those tuples with $(|J_1^c|+|J_2^c|)=o(N^{1/4})$, it contributes only

$$\left(\frac{2\cdot|J_1^c|^2+2\cdot|J_2^c|^2}{k}\right)=o(1)$$

to Q^{leaf}. On the other hand, for the remaining $o(1)$ fraction of tuples, it can also only contribute $o(1)$ to Q^{leaf} as well. □

Lemma 3.9. *For any protocol Π. If it finds a collision with probability $\Omega(1)$, then $P^{\text{leaf}} = \Omega(1)$.*

Proof. For $R \in \mathcal{R}^{\text{leaf}}$, the probability that Alice and Bob find a collision pair is upper bounded by,

$$\max_{i,j\in[M], i\neq j} \Pr_{(x,y)\sim \boldsymbol{R}}[(x_i = x_j) \wedge (y_i = y_j)].$$

Since we decompose R into tuples $\mathcal{S}(R) \cup \mathcal{B}(R)$, we have that

$$\Pr_{(x,y)\sim \boldsymbol{R}}[(x_i = x_j)\wedge(y_i = y_j)] = \sum_S \frac{|S|}{|R|}\cdot \Pr_{(x,y)\sim \boldsymbol{S}}[(x_i = x_j)\wedge(y_i = y_j)] + \sum_B \frac{|B|}{|R|}\cdot \Pr_{(x,y)\sim \boldsymbol{B}}[(x_i = x_j)\wedge(y_i = y_j)].$$

For each dense structure $S \in \mathcal{S}(R)$, we use the following claim (see proof in Section D) to upper bound $\Pr_{(x,y)\sim \boldsymbol{S}}[(x_i = x_j) \wedge (y_i = y_j)]$.

Claim 3.10. *Let $S = (X \times Y, J_1, J_2) \in \mathcal{S}^{\text{leaf}}(R)$ be a structure. If either of $\boldsymbol{X}$ is γ-dense on J_1 or $\boldsymbol{Y}$ is γ-dense on J_2, then for any distinct pair $i, j \in [M]$,*

$$\Pr_{(x,y)\sim \boldsymbol{S}}[(x_i = x_j) \wedge (y_i = y_j)] \leq \frac{4}{k}.$$

For those $B \in \mathcal{B}(R)$, we simply upper bound $\Pr_{(x,y)\sim \boldsymbol{B}}[(x_i = x_j) \wedge (y_i = y_j)]$ by 1. Hence,

$$\max_{i,j\in[M], i\neq j} \Pr_{(x,y)\sim \boldsymbol{R}}[(x_i = x_j) \wedge (y_i = y_j)] \leq \sum_{S\in\mathcal{S}(R)} \frac{|S|}{|R|}\cdot\frac{4}{k} + \sum_{B\in\mathcal{B}(R)} \frac{|B|}{|R|} = o(1) + \sum_{B\in\mathcal{B}(R)} \frac{|B|}{|R|}.$$

If the protocol tree Π finds a collision with probability $\Omega(1)$, we must have that

$$\sum_{R\in\mathcal{R}^{\text{leaf}}} \sum_{B\in\mathcal{B}(R)} \frac{|R|}{k^{2M}} \cdot \frac{|B|}{|R|} = \Omega(1).$$

which implies that

$$P^{\text{leaf}} = \sum_{(L,J_1,J_2)\in\mathcal{B}^{\text{leaf}}\cup\mathcal{C}^{\text{leaf}}} \left(\frac{|L|}{k^{2M}}\right) \geq \sum_{B\in\mathcal{B}^{\text{leaf}}} \left(\frac{|B|}{k^{2M}}\right) = \sum_{R\in\mathcal{R}^{\text{leaf}}} \sum_{B\in\mathcal{B}(R)} \frac{|R|}{k^{2M}}\cdot\frac{|B|}{|R|} = \Omega(1).$$

□

Now the proof of Theorem 1.7 simply follows by combining the above lemmas.

References

[Aar02] Scott Aaronson. Quantum lower bound for the collision problem. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 635–642, 2002. 1

[AKKT19] Scott Aaronson, Robin Kothari, William Kretschmer, and Justin Thaler. Quantum lower bounds for approximate counting via laurent polynomials. *arXiv preprint arXiv:1904.08914*, 2019. 2

[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. 1, 2, 3

[BHH11] Sergey Bravyi, Aram W Harrow, and Avinatan Hassidim. Quantum algorithms for testing properties of distributions. *IEEE Transactions on Information Theory*, 57(6):3971–3981, 2011. 2

[CFK+19] Arkadev Chattopadhyay, Yuval Filmus, Sajin Koroth, Or Meir, and Toniann Pitassi. Query-to-communication lifting for bpp using inner product. In *46th International Colloquium on Automata, Languages, and Programming (ICALP 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019. 6, 7, 9

[Din20] Itai Dinur. On the streaming indistinguishability of a random permutation and a random function. In *Advances in Cryptology–EUROCRYPT 2020: 39th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Zagreb, Croatia, May 10–14, 2020, Proceedings, Part II 30*, pages 433–460. Springer, 2020. 1

[GHK13] Mohsen Ghaffari, Bernhard Haeupler, and Majid Khabbazian. Randomized broadcast in radio networks with collision detection. In *Proceedings of the 2013 ACM symposium on Principles of distributed computing*, pages 325–334, 2013. 2

[GJ22] Mika Göös and Siddhartha Jain. Communication complexity of collision. *arXiv preprint arXiv:2208.00029*, 2022. 3, 6

[GPW15] Mika Göös, Toniann Pitassi, and Thomas Watson. Deterministic communication vs. partition number. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 1077–1088. IEEE, 2015. 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. IEEE, 2017. 6, 7, 8, 9, 16

[HMYZ23] Mi-Ying (Miryam) Huang, Xinyu Mao, Guangxu Yang, and Jiapeng Zhang. Communication lower bounds of key-agreement protocols via density increment arguments. Cryptology ePrint Archive, Paper 2023/1349, 2023. `https://eprint.iacr.org/2023/1349`. 9

[HP17] Pavel Hrubeš and Pavel Pudlák. Random formulas, monotone circuits, and interpolation. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 121–131. IEEE, 2017. 4

[IPU94] Russell Impagliazzo, Toniann Pitassi, and Alasdair Urquhart. Upper and lower bounds for tree-like cutting planes proofs. In *Proceedings Ninth Annual IEEE Symposium on Logic in Computer Science*, pages 220–228. IEEE, 1994. 4

[IR21] Dmitry Itsykson and Artur Riazanov. Proof complexity of natural formulas via communication arguments. In *36th Computational Complexity Conference (CCC 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2021. 2, 3, 4

[IS20] Dmitry Itsykson and Dmitry Sokolov. Resolution over linear equations modulo two. *Annals of Pure and Applied Logic*, 171(1):102722, 2020. 4

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

[LZ17] Shachar Lovett and Jiapeng Zhang. On the impossibility of entropy reversal, and its application to zero-knowledge proofs. In *Theory of Cryptography: 15th International Conference, TCC 2017, Baltimore, MD, USA, November 12-15, 2017, Proceedings, Part I 15*, pages 31–55. Springer, 2017. 1

[LZ23] Shachar Lovett and Jiapeng Zhang. Streaming lower bounds and asymmetric set-disjointness. *arXiv preprint arXiv:2301.05658*, 2023. 1

[MSS07] Frédéric Magniez, Miklos Santha, and Mario Szegedy. Quantum algorithms for the triangle problem. *SIAM Journal on Computing*, 37(2):413–424, 2007. 2

[YZ22] Guangxu Yang and Jiapeng Zhang. Simulation methods in communication complexity, revisited. In *Electron. Colloquium Comput. Complex., TR22-019*, 2022. 9

A Proof of Lemma 3.1

Lemma A.1 (Density-restoring partition)**.** *Let $S = (X \times Y, J_1, J_2)$ be a structure. If $\mathbf{Y}$ is further γ-dense on J_2, then there is a partition of $X \times Y$,*

$$X \times Y = X^1 \times Y^1 \cup X^1 \times Y^1_{\text{error}} \cup \cdots \cup X^t \times Y^t \cup X^t \times Y^t_{\text{error}}$$

such that every X^i is associated with a set $I_i \subseteq J_1$ and $p_{\geq i} := \frac{|\bigcup_{j \geq i} X^j|}{|X|}$ satisfies the following properties:

1. *For every i, $S^i := (X^i \times Y^i, J_1 \setminus I_i, J_2)$ is a structure*
2. *The tuple $B^i := (X^i \times Y^i_{\text{error}}, J_1 \setminus I_i, J_2)$ could be arbitrary*
3. *For every i, $\mathbf{X}^i$ is γ-dense on $J_1 \setminus I_i$*
4. $\mathcal{D}_\infty(\mathbf{X}^i, J_1 \setminus I_i) \leq \mathcal{D}_\infty(\mathbf{X}, J_1) - |I_i| + \log \frac{1}{p_{\geq i}}$.
5. *For every i,$(Y^i, Y^i_{\text{error}})$ is a partition of Y such that $|Y^i_{\text{error}}|/|Y| \leq 2 \cdot (|I_i \cup J_1^c|^2 - |J_1^c|^2)/k$.*

Similarly, If $\mathbf{X}$ is γ-dense on J_1, analogous conclusions hold for the partition of R with the roles of $\mathbf{X}$ and $\mathbf{Y}$ interchanged.

Proof. Since $(X \times Y, J_1, J_2)$ is a structure, $X_{J_1^c}$ is a fixed value and we denote it by $s := X_{J_1^c}$. We first apply the following density-restoring partition process on S.

Algorithm 2: Density-restoring partition process

```
Input: A rectangle $S = (X \times Y, J_1, J_2)$ be a structure and $\boldsymbol{Y}$ is $\gamma$-dense on $J_2$.
Output: A decomposition of $X \times Y = S^1 \cup B^1 \cup \cdots \cup S^t \cup B^t$.
Initialize $t \leftarrow 0$.
while $X$ is not nonempty do
    Let $I_t \subseteq J_1$ and $z^t \in [k]^{I_t}$ be the largest set (possibly $I_t = \emptyset$) such that ,
      $\Pr[\boldsymbol{X}_{I_t} = z^t] > 2^{-\gamma \cdot |I_t| \cdot \log k}$.
    Update $t \leftarrow t + 1$.
    Let $X^t = \{x \in X : x_{I_t} = z^t\}$ and $s^t = (s, z^t)$.
    Let $Y^t = \{y \in Y : y_i \neq y_j$ for all $(i,j) \in I_t \cup J_1^c$ with $s_i^t = s_j^t\}$ and $Y_{\text{error}}^t = Y \setminus Y^t$.
    Let $S^t = (X^t \times Y^t, J_1 \setminus I_t, J_2)$ and $B^t = (X^t \times Y_{\text{error}}^t, J_1 \setminus I_t, J_2)$.
    Update $X \leftarrow X \setminus X^t$.
```

This is a standard process (see [GPW17]), the only difference is that in Step 6, we partition Y into Y^t and Y_{error}^t, maintaining S^t as a structure. Following the density-restoring partition process, it is clear that $S^i = (X^i \times Y^i, J_1 \setminus I_i, J_2)$ is a structure. The facts that $\boldsymbol{X}^i$ is γ-dense on $J_1 \setminus I_i$ and

$$\mathcal{D}_\infty(\boldsymbol{X}^i, J_1 \setminus I_i) \leq \mathcal{D}_\infty(\boldsymbol{X}, J_1) - |I_i| + \log \frac{1}{p_{\geq i}}$$

hold by the standard proof (see [GPW17]). Now, we prove that, for every t,

$$|Y_{\text{error}}^t|/|Y| \leq 2 \cdot (|I_t \cup J_1^c|^2 - |J_1^c|^2)/k.$$

It is equivalent to show that

$$\Pr_{y \sim Y}\left[\exists i, j \in I_t \cup J_1^c, (y_i = y_j) \wedge (s_i^t = s_j^t)\right] \leq 2 \cdot (|I_t \cup J_1^c|^2 - |J_1^c|^2)/k.$$

Recall that S is a structure, or the event won't happen for $i, j \in J_1^c$, i.e.,

$$\Pr_{y \sim Y}\left[\exists i, j \in J_1^c, (y_i = y_j) \wedge (s_i^t = s_j^t)\right] = 0.$$

For any pair (i, j), if both $i, j \in J_2^c$, we also know there is no collision since S is a structure. On the other hand, if any of i or j is in J_2, by using the fact that $\boldsymbol{Y}$ is dense on J_2,

$$\Pr_{y \sim Y}\left[y_i = y_j\right] \leq \frac{4}{k}.$$

Now we only need to consider those pairs (i, j) such that: at least one of them is in $I_t \setminus J_1^c$ and at least one of them is in J_2. By union bound, we have that

$$\begin{aligned}
&\Pr_{y \sim Y}\left[\exists i, j \in I_t \cup J_1^c, (y_i = y_j) \wedge (s_i^t = s_j^t)\right] \\
\leq &\Pr_{y \sim Y}\left[\exists i, j \in I_t, (y_i = y_j) \wedge (s_i^t = s_j^t)\right] + \Pr_{y \sim Y}\left[\exists i \in I_t, j \in J_1^c, (y_i = y_j) \wedge (s_i^t = s_j^t)\right] \\
\leq &\frac{|I_t|^2}{2} \cdot \frac{4}{k} + |I_t| \cdot |J_1^c| \cdot \frac{4}{k} = \frac{2 \cdot \left(|I_i \cup J_1^c|^2 - |J_1^c|^2\right)}{k}.
\end{aligned}$$

We then finish the proof. □

B Proof of Claim 3.5

Claim B.1. *Let (R, J_1, J_2) be a structure, and let the following decomposition obtained by Lemma 3.1.*

$$R = X^1 \times Y^1 \cup X^1 \times Y^1_{\text{error}} \cup \cdots \cup X^t \times Y^t \cup X^t \times Y^t_{\text{error}}.$$

Let $S^1, B^1, \ldots, S^t, B^t$ be the corresponding tuples from Lemma 3.1. Then we have that

$$\sum_i \left(\frac{|S^i|}{|R|} \cdot \mathcal{D}_\infty(S^i, J_1 \setminus I_i, J_2) + \frac{|B^i|}{|R|} \cdot \mathcal{D}_\infty(B^i, J_1 \setminus I_i, J_2) \right) \le \mathcal{D}_\infty(R, J_1, J_1) - \sum_i \frac{|X^i|}{|X|} \cdot |I_i| + 2.$$

Proof. Let $p_i = \frac{|S^i| + |B^i|}{|R|} = \frac{|X^i|}{|X|}$. By Lemma 3.1, for any $i \in [t]$,

$$\mathcal{D}_\infty(\boldsymbol{X}^i, J_1 \setminus I_i) \le \mathcal{D}_\infty(\boldsymbol{X}, J_1) - |I_i| + \log \frac{1}{p_{\ge i}}.$$

By taking an average on X^i, we have that

$$\sum_i p_i \cdot \mathcal{D}_\infty(\boldsymbol{X}^i, J_1 \setminus I_i) \le \mathcal{D}_\infty(\boldsymbol{X}, J_1) - \sum_i p_i \cdot |I_i| + \sum_i p_i \cdot \log \frac{1}{p_{\ge i}}.$$

Note that $\sum_i p_i \cdot \log \frac{1}{p_{\ge i}} \le \int_0^1 \log \frac{1}{x} dx = 1$, we have

$$\sum_i p_i \cdot \mathcal{D}_\infty(\boldsymbol{X}^i, J_1 \setminus I_i) \le \mathcal{D}_\infty(\boldsymbol{X}, J_1) - \sum_i p_i \cdot |I_i| + 1. \tag{4}$$

On the other hand, $(Y^i, Y^i_{\text{error}})$ is a partition of Y for every i. Let $\boldsymbol{q}_i$ be a Bernoulli random variable with $\Pr[\boldsymbol{q}_i = 1] = \frac{|Y^i|}{|Y|} = \frac{|S^i|}{|S^i \cup B^i|}$. We have that,

$$\begin{aligned} \mathcal{D}_\infty(Y, J_2) &= \Pr[\boldsymbol{q}_i = 1] \cdot \mathcal{D}_\infty(Y^i, J_2) + \Pr[\boldsymbol{q}_i = 0] \cdot \mathcal{D}_\infty(Y^i_{\text{error}}, J_2) - \mathrm{H}(\boldsymbol{q}_i) \\ &\ge \Pr[\boldsymbol{q}_i = 1] \cdot \mathcal{D}_\infty(Y^i, J_2) + \Pr[\boldsymbol{q}_i = 0] \cdot \mathcal{D}_\infty(Y^i_{\text{error}}, J_2) - 1. \end{aligned} \tag{5}$$

By adding inequality (4) and (5), we finish the proof of this claim. □

C Proof of Lemma 3.7

First, we recall the definitions of P^j and Q^j.

Definition C.1. Let R be a rectangle with a decomposition into a set of tuples $\mathcal{L}$. We define its modified average quadratic of fixing size as

$$Q(R; \mathcal{L}) := \sum_{(L, J_1, J_2) \in \mathcal{L} : |J_1^c|, |J_2^c| < \sqrt{k}/4} \frac{|L|}{|R|} \cdot \left(\frac{2 \cdot |J_1^c|^2 + 2 \cdot |J_2^c|^2}{k} \right) + \sum_{(L, J_1, J_2) \in \mathcal{L} : |J_1^c| \text{ or } |J_2^c| \ge \sqrt{k}/4} \frac{|L|}{|R|} \cdot 2.$$

For nodes in the depth-j, we similarly define its total modified average quadratic of fixing size by,

$$Q^j = \sum_{R \in \mathcal{R}^j} \frac{|R|}{k^{2M}} \cdot Q(R, \mathcal{L}(R)).$$

For each j, we denote

$$P^j := \sum_{(L, J_1, J_2) \in \mathcal{B}^j \cup \mathcal{C}^j} \left(\frac{|L|}{k^{2M}} \right).$$

Now we are ready to prove the following lemma.

Lemma C.2. *Given a protocol* Π, $P^{\text{leaf}} \le Q^{\text{leaf}}$.

Proof. We prove that $P^j \le Q^j$ for all $j \ge 0$ by an induction proof. It is clear that $P^0 = Q^0 = 0$ in the roof. Now we assume that $P^j \le Q^j$ and aim to prove $P^{j+1} \le Q^{j+1}$.

For any rectangle $R = R^0 \cup R^1 \in \mathcal{R}^j$, in the $j+1$ interaction, Step 3 and Step 5 in Algorithm 1 decomposes $S \cap R^0$ and $S \cap R^1$ for each $(S, J_1, J_2) \in \mathcal{S}^j(R)$. For $b \in \{0,1\}$, let $S^{b,1} \cup B^{b,1} \cup ... S^{b,t} \cup B^{b,t}$ be the decomposed rectangles and let $I_1^b, ..., I_t^b$ be the associated sets of newly fixing coordinates. Thus, we have $\mathcal{L}(S \cap R^b), \mathcal{B}(S \cap R^b), \mathcal{C}(S \cap R^b)$ just follow the definitions. For $B \in \mathcal{B}^j(R)$ the decomposition does not fix new coordinates for those tuples.

First, since $P^{j+1} = \sum_{(L,J_1,J_2) \in \mathcal{B}^{j+1} \cup \mathcal{C}^{j+1}} \frac{|L|}{k^{2M}}$, we notice that

$$\begin{aligned}
P^{j+1} &= \sum_{R \in \mathcal{R}^j} \frac{|R|}{k^{2M}} \cdot \sum_{b \in \{0,1\}} \frac{|R^b|}{|R|} \cdot \left(\sum_{L \in \mathcal{C}^j(R) \cup \mathcal{B}^j(R)} \frac{|R^b \cap L|}{|R^b|} + \sum_{S \in \mathcal{S}^j(R) \setminus \mathcal{C}^j(R)} \frac{|S \cap R^b|}{|R^b|} \cdot \sum_{L \in \mathcal{B}(S \cap R^b) \cup \mathcal{C}(S \cap R^b)} \frac{|L|}{|S \cap R^b|} \right) \\
&= P^j + \sum_{R \in \mathcal{R}^j} \frac{|R|}{k^{2M}} \cdot \sum_{b \in \{0,1\}} \frac{|R^b|}{|R|} \cdot \sum_{S \in \mathcal{S}^j(R) \setminus \mathcal{C}^j(R)} \frac{|S \cap R^b|}{|R^b|} \cdot \sum_{L \in \mathcal{B}(S \cap R^b) \cup \mathcal{C}(S \cap R^b)} \frac{|L|}{|S \cap R^b|}.
\end{aligned}$$

Fix $(S, J_1, J_2) \in \mathcal{S}^j(R) \setminus \mathcal{C}^j(R)$ with $|J_1^c| \le \frac{\sqrt{k}}{4}$ and $|J_2^c| \le \frac{\sqrt{k}}{4}$. Let

$$\Delta(S \cap R^b) = \sum_{L \in \mathcal{B}(S \cap R^b) \cup \mathcal{C}(S \cap R^b)} \frac{|L|}{|S \cap R^b|}$$

be the increase in $S \cap R^b$, we aim to upper bound it by $Q(S \cap R^b; \mathcal{L}(S \cap R^b)) - Q(S \cap R^b; \{S \cap R^b\})$.

By Lemma 3.1, for any i, if $|I_i^b \cup J_1^c| \le \frac{\sqrt{k}}{4}$, then

$$\frac{|B^{b,i}|}{|S^{b,i}| + |B^{b,i}|} \le \frac{2 \cdot (|I_i(S) \cup J_1^c|^2 - |J_1^c|^2)}{k}. \tag{6}$$

Otherwise, since $|J_1^c| \le \frac{\sqrt{k}}{4}$, we can bound

$$\frac{|B^{b,i}|}{|S^{b,i}| + |B^{b,i}|} \le 1 \le \frac{9}{8} - \frac{2 \cdot |J_1^c|^2}{k}. \tag{7}$$

Let $C = \{i : |I_i^b \cup J_1^c| \ge \frac{\sqrt{k}}{4}\}$, by inequality (6) and inequality (7), we have

$$\begin{aligned}
\Delta(S \cap R^b) = \sum_{L \in \mathcal{C}(S \cap R^b) \cup \mathcal{B}(R^b)} \cdot \frac{|L|}{|S \cap R^b|} &= \sum_i \frac{|B^{b,i}|}{|S \cap R^b|} + \sum_{i \in C} \frac{|S^{b,i}|}{|S \cap R^b|} = \sum_{i \notin C} \frac{|B^{b,i}|}{|S \cap R^b|} + \sum_{i \in C} \frac{|S^{b,i}| + |B^{b,i}|}{|S \cap R^b|} \\
&\le \sum_{i \notin C} \frac{|S^{b,i}| + |B^{b,i}|}{|S \cap R^b|} \cdot \frac{2 \cdot (|I_i^b \cup J_1^c|^2 - |J_1^c|^2)}{k} + \sum_{i \in C} \frac{|S^{b,i}| + |B^{b,i}|}{|S \cap R^b|} \cdot \left(\frac{9}{8} - \frac{2 \cdot |J_1^c|^2}{k} \right) \\
&\le \sum_{i \notin C(S)} \frac{|S^{b,i}| + |B^{b,i}|}{|S \cap R^b|} \cdot \frac{2 \cdot (|I_i^b \cup J_1^c|^2 + |J_2^c|^2)}{k} + \sum_{i \in C(S)} \frac{|S^{b,i}| + |B^{b,i}|}{|S \cap R^b|} \cdot 2 - \frac{2 \cdot (|J_1^c|^2 + |J_2^c|^2)}{k} \\
&= Q(S \cap R^b; \mathcal{L}(S \cap R^b)) - Q(S \cap R^b; \{S \cap R^b\}),
\end{aligned} \tag{8}$$

where the second inequality is held by the fact that $|J_2^c| \le \frac{\sqrt{k}}{4}$. Moreover, for any $B \in \mathcal{B}^j(R)$, since we don't do decomposition on it, $Q(B \cap R^b; \mathcal{L}(B \cap R^b)) - Q^j(B \cap R^b; \{B \cap R^b\}) = 0$. For any $C \in \mathcal{C}^j(R)$, $Q(C \cap R^b; \mathcal{L}(C \cap R^b)) - Q^j(C \cap R^b; \{C \cap R^b\}) = \sum_{L \in \mathcal{L}(C \cap R^b)} \frac{|L|}{|C \cap R^b|} \cdot 2 - 2 = 0$.

Now we take an average on all rectangles in $\mathcal{R}^j$, by inequality (8), we have

$$\begin{aligned} P^{j+1} - P^j &= \sum_{R \in \mathcal{R}^j} \frac{|R|}{k^{2M}} \cdot \sum_{b \in \{0,1\}} \frac{|R^b|}{|R|} \cdot \sum_{S \in \mathcal{S}^j(R) \setminus \mathcal{C}^j(R)} \frac{|R^b \cap S|}{|R^b|} \cdot \Delta(S \cap R^b) \\ &\le \sum_{R \in \mathcal{R}^j} \frac{|R|}{k^{2M}} \cdot \sum_{b \in \{0,1\}} \frac{|R^b|}{|R|} \cdot \sum_{L \in \mathcal{L}^j(R)} \frac{|L \cap R^b|}{|R^b|} \cdot (Q(L \cap R^b; \mathcal{L}(L \cap R^b)) - Q(L \cap R^b; \{L \cap R^b\})) \\ &= Q^{j+1} - Q^j. \end{aligned}$$

Since $P^j \le Q^j$, $P^{j+1} \le Q^{j+1} - Q^j + P_j \le Q^{j+1}$. This finishes the proof. □

D Proof of Claim 3.10

Claim D.1. *Let $R \in \mathcal{R}^{leaf}$, for $S = (X \times Y, J_1, J_2) \in \mathcal{S}^{\text{leaf}}(R)$ and S is a structure, if either $\boldsymbol{X}$ is γ-dense on J_1 or $\boldsymbol{Y}$ is γ-dense on J_2, then for any distinct pair $i, j \in [M]$,*

$$\Pr_{(x,y) \sim (\boldsymbol{X}, \boldsymbol{Y})} [(x_i = x_j) \wedge (y_i = y_j)] \le \frac{4}{k}.$$

Proof. WLOG, we assume that $\boldsymbol{X}$ is γ-dense on J_1. Since (S, J_1, J_2) is a structure, $\boldsymbol{X}_{J_1^c}$ is fixed and we denote it by $s = X_{J_1^c}$. We consider the two cases.

- **Case 1:** Both $i, j \in J_1^c$, i.e., for all $x \in X$, $x_i = s_i$ and $x_j = s_j$. Since $S = (X \times Y, J_1, J_2)$ is a structure, then either $s_i \ne s_j$, or $s_i = s_j$ and $y_i \ne y_j$ for all $y \in Y$. For both cases, we have that

$$\Pr_{y \sim \boldsymbol{Y}} [(y_i = y_j)] = 0.$$

- **Case 2:** Either $i \in J_1$ or $j \in J_1$. WLOG, we assume that $i \in J_1$. Now we have two sub-cases. If j is also in J_1, then by the fact that $\boldsymbol{X}$ is γ-dense on J_1 (in particular, dense on $\{i, j\}$),

$$\Pr_{x \sim \boldsymbol{X}} [x_i = x_j] \le \frac{k}{k^{2 \cdot \gamma}} = \frac{4}{k}$$

  On the other hand, if $j \in J_1^c$, i.e., $X_j = s_j$, by using the fact that $\boldsymbol{X}$ is γ-dense on J_1 again,

$$\Pr_{x \sim \boldsymbol{X}} [x_i = s_j] \le \frac{1}{k^\gamma} = \frac{2}{k}$$

  For both cases, we have

$$\Pr_{(x,y) \sim (\boldsymbol{X}, \boldsymbol{Y})} [(x_i = x_j) \wedge (y_i = y_j)] \le \Pr_{x \sim \boldsymbol{X}} [x_i = x_j] \le \frac{4}{k}$$

The claim then follows. □