Disjointness Through the Lens of Vapnik–Chervonenkis Dimension: Sparsity and Beyond

Anup Bhattacharya
Indian Statistical Institute, Kolkata, India
bhattacharya.anup@gmail.com

Sourav Chakraborty
Indian Statistical Institute, Kolkata, India
sourav@isical.ac.in

Arijit Ghosh
Indian Statistical Institute, Kolkata, India
arijitiitkgpster@gmail.com

Gopinath Mishra
Indian Statistical Institute, Kolkata, India
gopianjan117@gmail.co

Manaswi Paraashar
Indian Statistical Institute, Kolkata, India
manaswi.isi@gmail.co

Abstract

The disjointness problem – where Alice and Bob are given two subsets of $\{1, \ldots, n\}$ and they have to check if their sets intersect – is a central problem in the world of communication complexity. While both deterministic and randomized communication complexities for this problem are known to be $\Theta(n)$, it is also known that if the sets are assumed to be drawn from some restricted set systems then the communication complexity can be much lower. In this work, we explore how communication complexity measures change with respect to the complexity of the underlying set system. The complexity measure for the set system that we use in this work is the Vapnik–Chervonenkis (VC) dimension. More precisely, on any set system with VC dimension bounded by d, we analyze how large can the deterministic and randomized communication complexities be, as a function of d and n. The d-sparse set disjointness problem, where the sets have size at most d, is one such set system with VC dimension d. The deterministic and the randomized communication complexities of the d-sparse set disjointness problem have been well studied and is known to be $\Theta\left(d \log\left(n/d\right)\right)$ and $\Theta(d)$, respectively, in the multi-round communication setting. In this paper, we address the question of whether the randomized communication complexity is always upper bounded by a function of the VC dimension of the set system, and does there always exist a gap between the deterministic and randomized communication complexity for set systems with small VC dimension.

In this paper, we construct two natural set systems of VC dimension d, motivated from geometry. Using these set systems we show that the deterministic and randomized communication complexity can be $\widetilde{\Theta}\left(d \log\left(n/d\right)\right)$ for set systems of VC dimension d and this matches the deterministic upper bound for all set systems of VC dimension d. We also study the deterministic and randomized communication complexities of the set intersection problem when sets belong to a set system of bounded VC dimension. We show that there exists set systems of VC dimension d such that both deterministic and randomized (one-way and multi-round) complexities for the set intersection problem can be as high as $\Theta\left(d \log\left(n/d\right)\right)$, and this is tight among all set systems of VC dimension d.

2012 ACM Subject Classification Theory of computation → Communication complexity

Keywords and phrases Communication complexity, VC dimension, Sparsity, and Geometric Set System

Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques
(APPROX/RANDOM 2020).
Editors: Jarosław Byrka and Raghu Meka; Article No. 23; pp. 23:1–23:15
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Digital Object Identifier 10.4230/LIPIcs.APPROX/RANDOM.2020.23

Category RANDOM

Funding *Anup Bhattacharya*: Supported by SERB-National Post Doctoral Fellowship, India.
Arijit Ghosh: Supported by Ramanujan Fellowship (No. SB/S2/RJN-064/2015), India.

Acknowledgements The authors would like to thank Sudeshna Kolay and Arijit Bishnu for the many discussions in the early stages of this work.

1 Introduction

Since its introduction by Yao [22], communication complexity occupies a central position in theoretical computer science. A striking feature of communication complexity is its interplay with other diverse areas like analysis, combinatorics, and geometry [9, 17]. Vapnik and Chervonenkis [21] introduced the measure *Vapnik-Chervonenkis dimension* or *VC dimension* for set systems in the context of statistical learning theory. As was the case with communication complexity, VC dimension has found numerous connections and applications in many different areas like approximation algorithms, discrete and combinatorial geometry, computational geometry, discrepancy theory and many other areas [10, 3, 14, 11]. In this work we study both of them under the same lens: of restricted systems and, for the first time, prove that geometric simplicity does not necessarily imply efficient communication complexity.

Lets start with recollecting some definitions from Vapnik–Chervonenkis theory. Let $\mathcal{S}$ be a collection of subsets of a *universe* $\mathcal{U}$. For a subset y of $\mathcal{U}$, we define

$$\mathcal{S}|_y := \{y \cap x \,:\, x \in \mathcal{S}\}\,.$$

We say a subset y of $\mathcal{U}$ is *shattered* by $\mathcal{S}$ if $\mathcal{S}|_y = 2^y$, where 2^y denotes the power set of y. *Vapnik–Chervonenkis (VC) dimension* of $\mathcal{S}$, denoted as VC-dim$(\mathcal{S})$, is the size of the largest subset y of $\mathcal{U}$ shattered by $\mathcal{S}$. VC dimension has been one of the fundamental measures for quantifying complexity of a collection of subsets.

Now let us revisit the world of communication complexity. Let $f : \Omega_1 \times \Omega_2 \to \Omega$. In *communication complexity*, two players Alice and Bob get as inputs $x \in \Omega_1$ and $y \in \Omega_2$ respectively, and the goal for the players is to device a protocol to compute $f(x, y)$ by exchanging as few bits of information between themselves as possible.

The *deterministic communication complexity* $D(f)$ of a function f is the minimum number of bits Alice and Bob will exchange in the worst case to deterministically compute the function f. In the randomized setting, both Alice and Bob share an infinite random source[1] and the goal is to give the correct answer with probability at least 2/3. The randomized communication complexity $R(f)$ of f denotes the minimum number of bits exchanged by the players in the worst case input by the best randomized protocol computing f. In both deterministic and randomized settings, Alice and Bob are allowed to make multiple rounds of interaction. Communication complexity when the number of rounds of interaction is bounded is also often studied. An important special case is when only one round of communication is allowed, that is, only Alice is allowed to send messages to Bob and Bob computes the output. We will denote by $D^{\rightarrow}(f)$ and $R^{\rightarrow}(f)$ the *one way deterministic communication complexity* and *one way randomized communication complexity* respectively, of f.

[1] This is the communication complexity setting with shared random coins. If no random string is shared, it is called the private random coins setting. By [13] we know that the communication complexity in both the setting differs by at most a logarithmic additive factor.

One of the most well studied functions in communication complexity is the disjointness function. Given a universe $\mathcal{U}$ known to both Alice and Bob, the *disjointness function*, $\textsc{Disj}_{\mathcal{U}} : 2^{\mathcal{U}} \times 2^{\mathcal{U}} \to \{0,\, 1\}$, where $2^{\mathcal{U}}$ denotes the power set of $\mathcal{U}$, is defined as follows:

$$\textsc{Disj}_{\mathcal{U}}(x, y) = \begin{cases} 1, & \text{if } x \cap y = \emptyset \\ 0, & \text{o/w} \end{cases} \tag{1}$$

We also define the *intersection function*. Given a universe $\mathcal{U}$ known to both Alice and Bob, the *intersection function*, $\textsc{Int}_{\mathcal{U}} : 2^{\mathcal{U}} \times 2^{\mathcal{U}} \to 2^{\mathcal{U}}$, where $2^{\mathcal{U}}$ denotes the power set of $\mathcal{U}$, is defined as $\textsc{Int}_{\mathcal{U}}(x, y) = x \cap y$. It is easy to see that $\textsc{Int}_{\mathcal{U}}$ is harder function to compute than $\textsc{Disj}_{\mathcal{U}}$. The $\textsc{Disj}_{\mathcal{U}}$ function and its different variants, like $\textsc{Int}_{\mathcal{U}}$, have been one of the most important problems in communication complexity and have found numerous applications in areas like streaming algorithms for proving lower bounds [17, 15]. By abuse of the notation, when $\mathcal{U} = [n]$, where $[n]$ denotes the set $\{1,\, \ldots,\, n\}$, we will denote the functions $\textsc{Disj}_{[n]}$ and $\textsc{Int}_{[n]}$ by $\textsc{Disj}_n$ and $\textsc{Int}_n$ respectively.

Using the standard *rank argument* [9, 15] one can show that $D(\textsc{Disj}_n) = \Theta(n)$. In a breakthrough paper, Kalyanasundaram and Schnitger [8] proved that $R(\textsc{Disj}_n) = \Omega(n)$. Razborov [16] and Bar-Yossef et al. [1] gave alternate proofs for the above result. From the above cited results we can also see the $D(\textsc{Int}_n) = R(\textsc{Int}_n) = \Theta(n)$.

Naturally, one would also like to ask what happens to the deterministic and randomized communication complexity (one way or multiple round) of $\textsc{Disj}_n$, when both Alice and Bob know that their inputs have more structure. In particular what can we say if the inputs are guaranteed to be from a subset of $\mathcal{S} \subseteq 2^{\mathcal{U}}$, where $\mathcal{S}$ is known to both players. Let $\textsc{Disj}_{\mathcal{U}}$ functions *restricted* to $\mathcal{S} \times \mathcal{S}$ be denoted by $\textsc{Disj}_{\mathcal{U}} |_{\mathcal{S}\times\mathcal{S}}$. This problem has also been studied extensively, mostly for certain special classes of subsets $\mathcal{S} \subseteq 2^{\mathcal{U}}$. For example, the *sparse set disjointness* function, where the set $\mathcal{S}$ contains all the subsets of $\mathcal{U}$ of size at most d, is an important special case of these works.

We will denote by d-$\textsc{SparseDisj}_n$ and d-$\textsc{SparseInt}_n$, the functions $\textsc{Disj}_n |_{\mathcal{S}\times\mathcal{S}}$ and $\textsc{Int}_n |_{S\times S}$ respectively, where S is the collections of all subsets of $[n]$ of size at most d. Using the *rank argument* [9, 15], one can again show that, for all $d \leq n$, the deterministic communication complexity of d-$\textsc{SparseDisj}_n$ is $\Omega\left(d \log \frac{n}{d}\right)$. Håstad and Wigderson [6], and Dasgupta et al. [4] showed that the randomized communication complexity and one round randomized communication complexity of d-$\textsc{SparseDisj}_n$ is $R(d\text{-}\textsc{SparseDisj}_n) = \Theta(d)$ and $R^{\to}(d\text{-}\textsc{SparseDisj}_n) = \Theta(d \log d)$ respectively. In a follow up work, Saglam and Tardos [18] proved that with $O(\log^* d)$ rounds of communication and $O(d)$ bits of communication it is possible to compute d-$\textsc{SparseDisj}_n$. More recently, Brody et al. [2] proved that $R^{\to}\left(d\text{-}\textsc{SparseInt}_n\right) = \Theta\left(d \log d\right)$ and $R(d\text{-}\textsc{SparseInt}_n) = \Theta(d)$. These results show that in the d-sparse setting, there is a separation between randomized and deterministic communication complexity of $\textsc{Disj}_n$ and $\textsc{Int}_n$ functions.

One would like to ask what happens to the communication complexity for other restrictions to the disjointness (and intersection) problem. The following are two natural problems, with a geometric flavor, for which one would like to study the communication complexity.

▶ **Problem 1** ($\textsc{Discrete Line Disj}$). Let $G \subset \mathbb{Z}^2$ be a set of n points in $\mathbb{Z}^2$ and L be the set of all lines in $\mathbb{R}^2$. Also, let $\mathcal{L} = L^d$ denote the collection of all d-size subsets of L. The $\textsc{Discrete Line Disj}$ on G and $\mathcal{L}$ is a function, $\textsc{Disj}_G |_{\mathcal{L}\times\mathcal{L}}$: $\mathcal{L} \times \mathcal{L} \to \{0, 1\}$ defined as $\textsc{Disj}_G |_{\mathcal{L}\times\mathcal{L}} (\{\ell_1, \ldots, \ell_d\}, \{\ell'_1, \ldots, \ell'_d\})$ is 1 if and only if there exists a line in Alice's set[2]

[2] We assume that Alice has the set $\{\ell_1, \ldots, \ell_d\}$ and Bob has the set $\{\ell'_1, \ldots, \ell'_d\}$.

that intersects some line in Bob's set at some point in G. Formally,

$$\textsc{Disj}_G\mid_{\mathcal{L}\times\mathcal{L}}(\{\ell_1,\dots,\ell_d\},\{\ell'_1,\dots,\ell'_d\}) = \begin{cases} 1, & \text{if } \exists i,j\in[d] \text{ s.t. } \ell_i\cap\ell'_j\cap G=\emptyset \\ 0, & \text{o/w} \end{cases} \tag{2}$$

▶ **Problem 2** (Discrete Interval Disj)**.** Let $X\subset\mathbb{Z}$ be a set of n points in $\mathbb{Z}$ and Int be the set of all possible intervals. Also, let $\mathcal{I}=Int^d$ denote the collection of all d-size subsets of Int. The Discrete Interval Disj on X and $\mathcal{I}$ is a function, $\textsc{Disj}_X\mid_{\mathcal{I}\times\mathcal{I}}\colon \mathcal{I}\times\mathcal{I}\to\{0,1\}$ defined as $\textsc{Disj}_X\mid_{\mathcal{I}\times\mathcal{I}}(\{I_1,\dots,I_d\},\{I'_1,\dots,I'_d\})$ is 1 if and only if there exists an interval in Alice's set[3] that intersects some interval in Bob's set at some point in X.

$$\textsc{Disj}_X\mid_{\mathcal{I}\times\mathcal{I}}(\{I_1,\dots,I_d\},\{I'_1,\dots,I'_d\}) = \begin{cases} 1, & \text{if } \exists i,j\in[d] \text{ s.t. } I_i\cap I'_j\cap X=\emptyset \\ 0, & \text{o/w} \end{cases} \tag{3}$$

Note that both the Discrete Line Disj and Discrete Interval Disj functions are generalizations of sparse set disjointness function.[4] Although it may not be obvious at first look, but both the Discrete Line Disj function and the Discrete Interval Disj functions are disjointness functions restricted to a suitable subset. In fact, the connection between the *Sparse set disjointness* function (d-$\textsc{SparseDisj}_n$), the Discrete Line Disj function and the Discrete Interval Disj function run deep - all the three subsets of the domain which help to define the functions as restriction of the disjointness function have VC dimension $\Theta(d)$, see Appendix A. Naturally one would like to know, if the fact that the collection of subsets $\mathcal{S}$ has VC dimension d has any implication on the communication complexity of $\textsc{Disj}_{\mathcal{U}}\mid_{\mathcal{S}\times\mathcal{S}}$. For example, is the randomized communication complexity of Discrete Line Disj function and the Discrete Interval Disj function upper bounded by a function of d (independent of n)? And, do the Discrete Line Disj function and the Discrete Interval Disj function also have a separation between their randomized and deterministic communication complexities similar to that of the *Sparse set disjointness* function (d-$\textsc{SparseDisj}_n$)? We show that these are not necessarily the cases.

▶ **Theorem 3.** *For Discrete Interval Disj: there exists a $X\subset\mathbb{Z}$ with n points such that*

$$D(\textsc{Disj}_X\mid_{\mathcal{I}\times\mathcal{I}}) = D^{\to}(\textsc{Disj}_X\mid_{\mathcal{I}\times\mathcal{I}}) = R^{\to}(\textsc{Disj}_X\mid_{\mathcal{I}\times\mathcal{I}}) = \Theta\left(d\log\frac{n}{d}\right).$$

▶ **Theorem 4.** *For Discrete Line Disj: there exists a $G\subset\mathbb{Z}^2$ with n points such that $D(\textsc{Disj}_G\mid_{\mathcal{L}\times\mathcal{L}}) = D^{\to}(\textsc{Disj}_G\mid_{\mathcal{L}\times\mathcal{L}}) = \Theta\left(d\log\frac{n}{d}\right)$ and, for the randomized setting,*

$$R(\textsc{Disj}_G\mid_{\mathcal{L}\times\mathcal{L}}) = \Omega\left(d\frac{\log(n/d)}{\log\log(n/d)}\right)$$

Discrete Line Int , that is, the intersection finding version of Discrete Line Disj is defined as follows : the objective is to compute a function $\textsc{Int}_G\mid_{\mathcal{L}\times\mathcal{L}}\colon \mathcal{L}\times\mathcal{L}\to G$ that is defined as

$$\textsc{Int}_G\mid_{\mathcal{L}\times\mathcal{L}}(\{\ell_1,\dots,\ell_d\},\{\ell'_1,\dots,\ell'_d\}) = \bigcup_{i,j\in[d]}\left(\ell_i\cap\ell'_j\cap G\right).^5$$

[3] We assume that Alice has the set $\{I_1,\dots,I_d\}$ and Bob has the set $\{I'_1,\dots,I'_d\}$.

[4] Take n integer points on the x-axis. For Discrete Line Disj setting, restrict only to lines orthogonal to x-axis. For Discrete Interval Disj setting, take n integer points on $\mathbb{Z}$ and only restrict to intervals containing one integer point. Both of these restriction will give the disjointness problem in the d-sparse setting.

As we have already mentioned, Brody et al. [2] proved that $R(d\text{-}\textsc{SparseInt}_n) = \Theta(d)$, whereas $D(d\text{-}\textsc{SparseInt}_n) = \Theta\left(d \log \frac{n}{d}\right)$. We show that Discrete Line Int does not demonstrate such a separation between the deterministic and randomized communication complexity.

▶ **Theorem 5.** *For* Discrete Line Int *: there exists a $G \subset \mathbb{Z}^2$ with n points such that*

$$D(\textsc{Int}_G\mid_{\mathcal{L}\times\mathcal{L}}) = D^{\rightarrow}(\textsc{Int}_G\mid_{\mathcal{L}\times\mathcal{L}}) = R^{\rightarrow}(\textsc{Int}_G\mid_{\mathcal{L}\times\mathcal{L}}) = R(\textsc{Int}_G\mid_{\mathcal{L}\times\mathcal{L}}) = \Theta\left(d \log \frac{n}{d}\right).$$

The upper bound for all the above three theorems can be obtained from the fact that the corresponding sets have VC dimension $\Theta(d)$, see Appendix A. *Sauer-Shelah Lemma* [19, 20, 21] states that if $\mathcal{S} \subseteq 2^{[n]}$ and VC-dim$(\mathcal{S}) \leq d$ then $\mid \mathcal{S} \mid \leq \left(\frac{en}{d}\right)^d$. Thus if VC-dim$(\mathcal{S}) \leq d$, then the Sauer-Shelah Lemma implies that $D^{\rightarrow}(\textsc{Int}_n\mid_{\mathcal{S}\times\mathcal{S}}) = O\left(d \log \frac{n}{d}\right)$. So, $O\left(d \log \frac{n}{d}\right)$ is a upper bound to the above questions, both for randomized and deterministic and also for the one-way communication. But can the randomized communication complexity of $\textsc{Disj}_{\mathcal{U}}\mid_{\mathcal{S}\times\mathcal{S}}$ and $\textsc{Int}_{\mathcal{U}}\mid_{\mathcal{S}\times\mathcal{S}}$ be even lower when S has VC dimension d? The following result, which is a direct consequence of Theorems 3, 4 and 5, enables us to we answer the question in the negative:

▶ **Theorem 6.** *Let $1 \leq d \leq n$.*

1. *There exists $\mathcal{S} \subseteq 2^{[n]}$ with VC-dim$(\mathcal{S}) \leq d$ and $R^{\rightarrow}(\textsc{Disj}_n\mid_{\mathcal{S}\times\mathcal{S}}) = \Omega\left(d \log \frac{n}{d}\right)$.*
2. *There exists $\mathcal{S} \subseteq 2^{[n]}$ with VC-dim$(\mathcal{S}) \leq d$ and $R(\textsc{Disj}_n\mid_{\mathcal{S}\times\mathcal{S}}) = \Omega\left(d \frac{\log(n/d)}{\log\log(n/d)}\right)$.*
3. *There exists $\mathcal{S} \subseteq 2^{[n]}$ with VC-dim$(\mathcal{S}) \leq d$ and $R(\textsc{Int}_n\mid_{\mathcal{S}\times\mathcal{S}}) = \Omega\left(d \log \frac{n}{d}\right)$.*

The following table compares our result with the previous best known lower bound for $\textsc{Disj}_{\mathcal{U}}\mid_{\mathcal{S}\times\mathcal{S}}$ and $\textsc{Int}_{\mathcal{U}}\mid_{\mathcal{S}\times\mathcal{S}}$ among all sets $S \subset 2^{\mathcal{U}}$ of VC dimension d.

Table 1 The largest communication complexity, for the functions $\textsc{Disj}_n\mid_{\mathcal{S}\times\mathcal{S}}$ and $\textsc{Int}_n\mid_{\mathcal{S}\times\mathcal{S}}$, among all $S \subseteq 2^{[n]}$ of VC dimension d, that was previously known and what we prove in this paper. Tight bounds of $\Omega\left(d \log \frac{n}{d}\right)$ for the largest $D(\textsc{Disj}_n\mid_{\mathcal{S}\times\mathcal{S}})$, $D^{\rightarrow}(\textsc{Disj}_n\mid_{\mathcal{S}\times\mathcal{S}})$, $D(\textsc{Int}_n\mid_{\mathcal{S}\times\mathcal{S}})$ and $D^{\rightarrow}(\textsc{Int}_n\mid_{\mathcal{S}\times\mathcal{S}})$, among all $S \subset 2^{[n]}$ of VC dimension d, follows directly from the fact that if $\mathcal{S}$ is a collection of all subsets of $[n]$ of size at most d then $D(\textsc{Disj}_n\mid_{\mathcal{S}\times\mathcal{S}}) = D(\textsc{Int}_n\mid_{\mathcal{S}\times\mathcal{S}}) = \Omega\left(d \log\left(\frac{n}{d}\right)\right)$, see [9, 15].

Problems	$R(\textsc{Disj}_n\mid_{\mathcal{S}\times\mathcal{S}})$	$R^{\rightarrow}(\textsc{Disj}_n\mid_{\mathcal{S}\times\mathcal{S}})$	$R(\textsc{Int}_n\mid_{\mathcal{S}\times\mathcal{S}})$	$R^{\rightarrow}(\textsc{Int}_n\mid_{\mathcal{S}\times\mathcal{S}})$
Previously Known	$\Omega(d)$ [6]	$\Omega(d \log d)$ [4]	$\Omega(d)$ [2]	$\Omega(d \log d)$ [2]
This Paper	$\Omega\left(d \frac{\log(n/d)}{\log\log(n/d)}\right)$	$\Omega\left(d \log \frac{n}{d}\right)$	$\Omega\left(d \log \frac{n}{d}\right)$	$\Omega\left(d \log \frac{n}{d}\right)$

Notations

We denote the set $\{1, \ldots, n\}$ by $[n]$. For a binary number $\mathbf{x}$, *decimal*$(\mathbf{x})$ denotes the decimal value of $\mathbf{x}$. For two vectors $\mathbf{x}$ and $\mathbf{y}$ in $\{0,1\}^n$, $\mathbf{x} \cap \mathbf{y} = \{i \in [n] : \mathbf{x}_i = \mathbf{y}_i = 1\}$, and $\mathbf{x} \subseteq \mathbf{y}$ when $\mathbf{x}_i \leq \mathbf{y}_i$ for each $i \in [n]$. For a finite set X, 2^X denotes the power set of X. For $x, y \in \mathbb{R}$ with $x < y$, $[x, y]$ denotes the closed interval is the set of all real numbers that lies between x and y.

2 One way communication complexity (Theorems 3 and 6 (1))

In this section we will prove the following result.

▶ **Theorem 7.** *For all $n \geq d$, there exists $X \subset \mathbb{Z}$ with $|X| = n$ and $\mathcal{R} \subseteq 2^X$ with VC-dim$(\mathcal{R}) = 2d$, such that*

$$\mathcal{R} \subseteq \left\{ X \cap \left(\bigcup_{1 \leq j \leq d} I_j \right) \mid \{I_1, \ldots, I_d\} \in \mathcal{I} \right\} \text{ and } R^{\rightarrow}(\textsc{Disj}_X \mid_{\mathcal{R} \times \mathcal{R}}) = \Omega\left(d \log \frac{n}{d}\right).$$

Note that the set $\mathcal{I}$ is defined in Problem 2.

▶ Remark 8. The above result takes care of the proofs of Theorem 3 and Theorem 6 (1).

The *hard* instance, for the proof of the above theorem, is inspired by the *interval* set systems in combinatorial geometry and is constructed in Section 2.1. In Section 2.2, we proof Theorem 7 by using a reduction from AUGMENTED INDEXING, which we denote by $\textsc{AugIndex}_\ell$. Formally the problem $\textsc{AugIndex}_\ell$ is defined as follows: Alice gets a string $\mathbf{x} \in \{0,1\}^\ell$ and Bob gets an index $j \in [\ell]$ and all $x_{j'}$ for $j' < j$. Bob reports x_j as the output.

▶ **Proposition 9.** (Miltersen et al. [12]) $R^{\rightarrow}(\textsc{AugIndex}_\ell) = \Omega(\ell)$.

2.1 Construction of a hard instance

We construct a set $X \subset \mathbb{Z}$ with $|X| = n$ and $\mathcal{R} \subseteq 2^X$ with VC-dim$(\mathcal{R}) = 2d$. Informally, X is the union of the set of points present in the union of d pairwise disjoint intervals, in $\mathbb{Z}$, each containing $\frac{n}{d}$ points. Each set in $\mathcal{R}$ is the union of the set of points in the subintervals anchored either at the left or the right end point of each of the above d intervals. Formally, the description of X and $\mathcal{R}$ are given below along with some of its properties that are desired to show Theorem 7.

The ground set X

Let $m = \frac{n}{d} - 2$. Without loss of generality we can assume that $m = 2^k$, where $k \in \mathbb{N}$. Let $J_0 = \{0, \ldots, m+1\}$ be the set of $m+2$ consecutive integers that starts from the origin and ends at $m+1$. Similarly, let J_p be the set of $m+1$ consecutive integers that starts at $p \in \mathbb{Z}$ and ends at $p+m+1$. Let $p_1, \ldots, p_d$ be d points in $\mathbb{Z}$ such that the sets $J_{p_1}, \ldots, J_{p_d}$ are pairwise disjoint. Let the *ground* set X be $\bigcup\limits_{i=1}^{d} J_{p_i}$. Note that $X \subset \mathbb{Z}$ and $|X| = (m+2)d = n$.

The subsets of X in $\mathcal{R}$

$\mathcal{R} \subseteq 2^X$ contains two types of sets $\mathcal{R}_0$ and $\mathcal{R}_{m+1}$, where

- Take any d intervals $R_1, \ldots, R_d$ of integer lengths such that, for all $i \in [d]$, length of R_i is at most $m+1$, $R_i \subseteq [p_i, p_i + m + 1]$, and R_i starts at p_i. Note that R_i does not intersect with any $X \setminus J_{p_i}$. The set $A = \bigcup\limits_{i=1}^{d} (R_i \cap X)$ is an element in $\mathcal{R}_0$. We say that A is *generated* by $R_1, \ldots, R_d$.
- Take any d intervals $R'_1, \ldots, R'_d$ of integer lengths such that, for all $i \in [d]$, length of R'_i is at most $m+1$, $R'_i \subseteq [p_i, p_i + m + 1]$ and R'_i ends at $p_i + m + 1$. Note that R'_i does not intersect with any $X \setminus J_{p_i}$. The set $B = \bigcup\limits_{i=1}^{d} (R'_i \cap X)$ is an element in $\mathcal{R}_{m+1}$. We say that B is *generated* by $R'_1, \ldots, R'_d$.

The following claim bounds the VC dimension of $\mathcal{R}$, constructed as above.

▷ Claim 10. For $X \subset \mathbb{Z}$ with $|X| = n$ and $\mathcal{R} \subset 2^X$ as described above, VC-dim$(\mathcal{R}) = 2d$,

Proof. The proof follows from the fact that any subset of of X containing $2d + 1$ points will contain at least three points from some J_{p_i}, where $i \in [d]$. These points in J_{p_i} can not be shattered by the sets in $\mathcal{R}$. Also, observe that there exists $2d$ points, with two from each J_{p_j}, that can be shattered by the sets in $\mathcal{R}$. ◁

Now, we give a claim about X and $\mathcal{R}$ constructed above that will be required for our proof of Theorem 7.

▷ Claim 11. Let $A \in \mathcal{R}_0$ and $B \in \mathcal{R}_{m+1}$ be such that A is generated by $R_1, \ldots, R_d$ and B is generated by $R'_1, \ldots, R'_d$. Then A and B intersects if and only if there exists an $i \in [d]$ such that R_i intersects R'_i at a point in J_{p_i}.

The proof of Claim 11 follows directly from our construction of $X \subset \mathbb{Z}$ and $\mathcal{R} \subseteq 2^X$, as $J_{p_1}, \ldots, J_{p_d}$ are pairwise disjoint.

2.2 Reduction from AugIndex$_{d \log m}$ to Disj$_X$ $|_{\mathcal{R}\times\mathcal{R}}$

Before presenting the reduction we recall the definitions of AugIndex$_{d \log m}$ and Disj$_X|_{\mathcal{R}\times\mathcal{R}}$. In AugIndex$_{d \log m}$, Alice gets $\mathbf{x} \in \{0,1\}^{d \log m}$ and Bob gets an index j and $x_{j'}$ for each $j' < j$. The objective of Bob is to report x_j as the output. In Disj$_X|_{\mathcal{R}\times\mathcal{R}}$, Alice gets $A \in \mathcal{R}_0$ and Bob gets $B \in \mathcal{R}_{m+1}$. The objective of Bob is to determine whether $A \cap B = \emptyset$. Note that $X, \mathcal{R}, \mathcal{R}_0$ and $\mathcal{R}_{m+1}$ are as discussed in the Section 2.1.

Let $\mathcal{P}$ be an one-way protocol that solves Disj$_X|_{\mathcal{R}\times\mathcal{R}}$ with $o\left(d \log \frac{n}{d}\right) = o(d \log m)$ bits of communication. Now, we consider the following protocol $\mathcal{P}'$ for AugIndex$_{d \log m}$ that has the same one way communication cost as that of Disj$_X|_{\mathcal{R}\times\mathcal{R}}$. Then we will be done with the proof of Theorem 7.

Protocol $\mathcal{P}'$ for AugIndex$_{d \log m}$ problem

Step-1 Let $\mathbf{x} \in \{0,1\}^{d \log m}$ be the input of Alice. Bob gets an index $j \in [d \log m]$ and bits $x_{j'}$ for each $j' < j$.

Step-2 Alice will form d strings $\mathbf{a}_1, \ldots, \mathbf{a}_d \in \{0,1\}^{\log m}$ by partitioning the string $\mathbf{x}$ into d parts such that $\mathbf{a}_i = x_{(i-1)\log m+1} \cdots x_{i \log m}$, where $i \in [d]$. Bob first forms a string $\mathbf{y} \in \{0,1\}^{d \log m}$, where $y_{j'} = x_{j'}$ for each $j' < j$, $y_j = 1$, and $y_{j'} = 0$ for each $j' > j$. Then Bob finds $\mathbf{b}_1, \ldots, \mathbf{b}_d \in \{0,1\}^{\log m}$ by partitioning the string $\mathbf{y}$ in to d parts such that $\mathbf{b}_i = y_{(i-1)\log m+1} \cdots y_{i \log m}$, where $i \in [d]$.

Step-3 For each $i \in [d]$, let R_i and R'_i be the intervals that starts at p_i and ends at $p_i + m + 1$, respectively, where $R_i = [p_i, m + p_i - decimal(\mathbf{a}_i)]$ and $R'_i = [p_i + m + 1 - decimal(\mathbf{b}\ _i), p_i{+}m{+}1]$. Alice finds the set $A \in \mathcal{R}_0$ generated by $R_1, \ldots, R_d$ and Bob finds the set $B \in \mathcal{R}_{m+1}$ generated by $R'_1, \ldots, R'_d$, i.e., $A = \bigcup\limits_{i\in[d]} (R_i \cap X)$ and $B = \bigcup\limits_{i\in[d]} (R'_i \cap X)$.

Step-4 Alice and Bob solves Disj$_X$ $|_{\mathcal{R}\times\mathcal{R}}$ on inputs A and B, and report $x_j = 0$ if and only if Disj$_X$ $|_{\mathcal{R}\times\mathcal{R}}$ $(A, B) = 0$. Note that x_j is the output of AugIndex$_{d \log m}$ problem.

The following observation follows from the description of the protocol $\mathcal{P}'$ and from the construction of $X \subset \mathbb{Z}$ and $\mathcal{R} \subseteq 2^X$.

▶ **Observation 12.** Let $i^* \in [d]$ such that $j \in \{(i^*-1)\log m + 1, i^* \log m\}$. Then
(i) $R_i \cap R'_i = \emptyset$ for all $i \neq i^*$.
(ii) $R_{i^*} \cap R'_{i^*} = \emptyset$ if and only if $decimal(\mathbf{b}_{i^*}) \leq decimal(\mathbf{a}_{i^*})$.
(iii) $decimal(\mathbf{b}_{i^*}) \leq decimal(\mathbf{a}_{i^*})$ if and only if $x_j = 0$.
We will use the above observation to show the correctness of the protocol $\mathcal{P}'$.

First consider the case $\text{DISJ}_X \mid_{\mathcal{R}\times\mathcal{R}} (A, B) = 0$. Then, by Claim 11, there exists an $i \in [d]$ such that R_i and R'_i intersects at a point in J_{p_i}. From Observation 12 (i), we can say $R_{i*} \cap R'_{i*} \neq \emptyset$. Combining $R_{i*} \cap R'_{i*} \neq \emptyset$ with Observations 12 (ii) and (iii), we have $x_j = 0$. Hence, $\text{DISJ}_X \mid_{\mathcal{R}\times\mathcal{R}} (A, B) = 0$ implies $x_j = 0$. The converse part, i.e., $x_j = 0$ implies $\text{DISJ}_X \mid_{\mathcal{R}\times\mathcal{R}} (A, B) = 0$, can be shown in the similar fashion.

The one-way communication complexity of protocol $\mathcal{P}'$ for $\text{AUGINDEX}_{d\log m}$ is the same as that of $\mathcal{P}$ for $\text{DISJ}_X|_{\mathcal{R}\times\mathcal{R}}$, that is, $o(d \log m)$. However, this is impossible as the one-way communication complexity of AUGMENTED INDEXING, over $d \log m$ bits, is $\Omega(d \log m) = \Omega\left(d \log \frac{n}{d}\right)$ bits. This completes the proof of Theorem 7.

3 Two way communication complexity (Theorems 4, 5, 6 (2) and 6 (3))

In this section, we prove the following theorems.

▶ **Theorem 13.** *For all $n \geq d$, there exists a $G \subset \mathbb{Z}^2$ with $|G| = n$ and $\mathcal{T} \subseteq 2^G$ with VC-dim$(\mathcal{T}) = 2d$, such that*

$$\mathcal{T} \subseteq \left\{ G \cap \left(\bigcup_{1\leq j\leq d} \ell_j \right) \mid \{\ell_1, \ldots, \ell_d\} \in \mathcal{L} \right\} \text{ and } R(\text{DISJ}_G \mid_{\mathcal{T}\times\mathcal{T}}) = \Omega\left(d \frac{\log(n/d)}{\log\log(n/d)} \right).$$

The set $\mathcal{L}$ is as defined in Problem 1.

▶ **Theorem 14.** *For all $n \geq d$, there exists a $G \subset \mathbb{Z}^2$ with $|G| = n$ and $\mathcal{T} \subseteq 2^G$ with VC-dim$(\mathcal{T}) = 2d$, such that*

$$\mathcal{T} \subseteq \left\{ G \cap \left(\bigcup_{1\leq j\leq d} \ell_j \right) \mid \{\ell_1, \ldots, \ell_d\} \in \mathcal{L} \right\} \text{ and } R(\text{INT}_G \mid_{\mathcal{T}\times\mathcal{T}}) = \Omega\left(d \log \frac{n}{d} \right).$$

The set $\mathcal{L}$ is as defined in problem 1.

▶ Remark 15. Theorem 13 takes care of Theorem 4 and 6(2). Theorem 14 takes care of Theorem 5 and 6(3).

Note that the same set system will be used for the proofs of the above theorems. The *hard* instance, for the proof of the above theorems, is inspired by *point line incidence* set systems in computational geometry and is constructed in Section 3.1. We prove Theorems 13 and 14 in Sections 3.2 and 3.3, respectively, using reductions.

3.1 The hard instance for the proofs of Theorems 13 and 14

In this subsection, we give the description of $G \subset \mathbb{Z}^2$ with $|G| = n$ and $\mathcal{T} \subseteq 2^G$, with VC-Dim$(\mathcal{T}) = 2d$. The same G and $\mathcal{T}$ will be our *hard* instance for the proofs of Theorems 13 and 14. In this subsection, without loss of generality, we can assume that d divides n and n/d is a perfect square.

Informally, G is the set of points present in the union of d many pairwise disjoint square grids each containing $\frac{n}{d}$ points and the grids are taken in such a way that any straight line of non-negative can intersects with at most one grid. Also, each set in $\mathcal{T}$ is the union of the set of points present in d many lines of non-negative slope such that one line intersects with exactly one grid. Moreover, all of the d lines have slopes either zero or positive. Formally, the description of G and $\mathcal{T}$ are given below along with some of its properties that are desired to show Theorems 13 and 14.

The ground set G

Let $m = \sqrt{\frac{n}{d}}$, and $G_{(0,0)} = \left\{(x,y) \in \mathbb{Z}^2 : 0 \leq x, y \leq m-1\right\}$ be the grid of size $m \times m$ anchored at the origin $(0,0)$. For any $p, q \in \mathbb{Z}$, the $m \times m$ grid anchored at (p,q) will be denoted by $G_{(p,q)}$, i.e., $G_{(p,q)} = \left\{(i+p, j+q) : (i,j) \in G_{(0,0)}\right\}$. For $d \in \mathbb{N}$, consider $G_{(p_1,q_1)}, \ldots, G_{(p_d,q_d)}$ satisfying the following property:

PROPERTY *For any i, $j \in [d]$, with $i \neq j$, let L_1 and L_2 be lines of non-negative slopes that pass through at least two points of $G_{(p_i,q_i)}$ and $G_{(p_j,q_j)}$, respectively. Then L_1 and L_2 does not intersect at any point* inside $\bigcup_{\ell=1}^{d} G_{(p_\ell,q_\ell)}$.

Observe that there exists $G_{(p_1,q_1)}, \ldots, G_{(p_d,q_d)}$ satisfying PROPERTY. We will take $G = \bigcup_{\ell=1}^{d} G_{(p_\ell,q_\ell)}$ as the ground set. Without loss of generality, we can assume that $(p_1,q_1) = (0,0)$. Note that $G \subset \mathbb{Z}^2$ and $|G| = dm^2 = n$.

The subsets of G in $\mathcal{T}$

$\mathcal{T}$ contains two types of subsets $\mathcal{T}_1$ and $\mathcal{T}_2$ of G, and they are generated by the following ways:

- Take any d lines $L_1, \ldots, L_d$ of non negative slope such that, $\forall i \in [d]$, L_i passes through $(p_i,q_i) \in G_{(p_i,q_i)}$ and (at least) another point in $G_{(p_i,q_i)}$. Note that L_i does not contain any point from $G \setminus G_{(p_i,q_i)}$. The set $A = \bigcup_{i=1}^{d} \left(L_i \cap G_{(p_i,q_i)}\right)$ is in $\mathcal{T}_1$, and we say A is *generated* by the lines $L_1, \ldots, L_d$.
- Take any d vertical lines $L'_1, \ldots, L'_d$ such that, $\forall i \in [d]$, L'_i contains at least one point from $G_{(p_i,q_i)}$. Note that L'_i does not contain any point from $G \setminus G_{(p_i,q_i)}$. The set $B = \bigcup_{i=1}^{d}(L'_i \cap G_{(p_i,q_i)})$ is in $\mathcal{T}_2$, and we say B is generated by the lines $L'_1, \ldots, L'_d$.

The following claim bounds the VC dimension of $\mathcal{T}$, which as described above.

▷ Claim 16. For $G \subset \mathbb{Z}^2$ and $\mathcal{T} \subseteq 2^G$ as described above, VC-dim$(\mathcal{T}) = 2d$.

Proof. The proof follows from the fact that any subset of X containing $2d+1$ points will contain at least three points from some $G_{(p_j,q_j)}, j \in [d]$. These points in $G_{(p_j,q_j)}$ can not be shattered by the sets in $\mathcal{T}$. Also, observe that there exists $2d$ points two from each $G_{(p_j,q_j)}$ that can be shattered by the sets in $\mathcal{T}$. ◁

Now, we give two claims about G and $\mathcal{T}$, constructed above, that follow directly from our construction of $G \subset \mathbb{Z}^2$ and $\mathcal{T} \subseteq 2^G$.

▷ Claim 17. Let $A \in \mathcal{T}_1$ and $B \in \mathcal{T}_2$ such that A is generated by lines $L_1, \ldots, L_d$ and A is generated by lines $L'_1, \ldots, L'_d$. Then A and B intersect if and only if there exists $i \in [d]$ such that L_i and L'_i intersect at a point in $G_{(p_i,q_i)}$.

▷ Claim 18. Let $A \in \mathcal{T}_1$ and $B \in \mathcal{T}_2$ such that A is generated by lines $L_1, \ldots, L_d$ and B is generated by lines $L'_1, \ldots, L'_d$. Also let $|A \cap B| = d$. Then for each $i \in [d]$, L_i and L'_i intersect at a point in $G_{(p_i,q_i)}$. Moreover, A (B) can be determined if we know B (A) and $A \cap B$.

The above claims will be used in the proofs of Theorems 13 and 14.

3.2 Proof of Theorem 13

Let us consider a problem in communication complexity denoted by $\text{OR-DISJ}^t_{\{0,1\}^\ell}$ that will be used in our proof. In $\text{OR-DISJ}^t_{\{0,1\}^\ell}$, Alice gets t strings $\mathbf{x}_1, \ldots, \mathbf{x}_t \in \{0,1\}^\ell$ and Bob also gets t strings $\mathbf{y}_1, \ldots, \mathbf{y}_t \in \{0,1\}^\ell$. The objective is to compute $\bigvee\limits_{i=1}^{t} \text{DISJ}_{\{0,1\}^\ell}(\mathbf{x}_i, \mathbf{y}_i)$. Note that $\text{DISJ}_{\{0,1\}^\ell}(\mathbf{x}_i, \mathbf{y}_i)$ is a binary variable that takes value 1 if and only if $\mathbf{x}_i \cap \mathbf{y}_i = \emptyset$.

▶ **Proposition 19** (Jayram et al. [7]). $R\left(\mathit{OR\text{-}DISJ}^t_{\{0,1\}^\ell}\right) = \Omega(\ell t)$.

Note that Proposition 19 directly implies the following result.

▶ **Proposition 20.** $R\left(\mathit{OR\text{-}DISJ}^t_{\{0,1\}^\ell} |_{S_\ell \times S_\ell}\right) = \Omega(\ell t)$, *where* $S_\ell = \{0,1\}^\ell \setminus \{0^\ell\}$.

Let $k \in \mathbb{N}$ be the largest integer such that first k consecutive primes $p_1, \ldots, p_k$ satisfy the following inequalty:

$$\Pi_{i=1}^k p_i \leq \sqrt{\frac{n}{d}}. \tag{4}$$

Using the fact that $\Pi_{i=1}^k p_i = e^{(1+o(1))k \log k}$, we get $k = \Theta\left(\frac{\log(n/d)}{\log\log(n/d)}\right)$.

We prove the theorem by a reduction from $\text{OR-DISJ}^d_{\{0,1\}^k} |_{S_k \times S_k}$ to $\text{DISJ}_G |_{\mathcal{T} \times \mathcal{T}}$, where $S_k := \{0,1\}^k \setminus \{0^k\}$.

Note that $G \subset \mathbb{Z}^2$ with $|G| = n$, and $\mathcal{T} \subseteq 2^G$, with VC-dim$(\mathcal{T}) = 2d$, are the same as that we constructed in Section 3.1. To reach a contradiction, assume that there exists a two way protocol $\mathcal{P}$ that solves $\text{DISJ}_G \ |_{\mathcal{T} \times \mathcal{T}}$ with communication cost of $o\left(d \frac{\log m}{\log\log m}\right) = o\left(d \frac{\log(n/d)}{\log\log(n/d)}\right)$ bits. Now, we give protocol $\mathcal{P}'$ that solves $\text{OR-DISJ}^d_{\{0,1\}^k} |_{S_k \times S_k}$, as described below.

Protocol $\mathcal{P}'$ for Or-Disj$^d_{\{0,1\}^k} |_{S_k \times S_k}$

Step-1 Let $A = (\mathbf{x}_1, \ldots, \mathbf{x}_d) \in [S_k]^d$ [6] and $B = (\mathbf{y}_1, \ldots, \mathbf{y}_d) \in [S_k]^d$ be the inputs of Alice and Bob for $\text{OR-DISJ}^d_{\{0,1\}^k} |_{S_k \times S_k}$. Recall that $S_k = \{0,1\}^k \setminus \{0^k\}$. Bob finds $\bar{B} = (\bar{\mathbf{y}}_1, \ldots, \bar{\mathbf{y}}_d) \in \left[\{0,1\}^k\right]^d$, where $\bar{\mathbf{y}}_i$ is obtained by complementing each bit of $\mathbf{y}_i$.

Step-2 Both Alice and Bob privately determine first k prime numbers $p_1, \ldots, p_k$ without any communication.

Step-3 Let $\Phi : \{0,1\}^k \to \{0,1\}^{\lceil \log(\sqrt{\frac{n}{d}}) \rceil}$ be the function such that $\phi(\mathbf{x})$ is the bit representation of the number $\prod_{i=1}^k p_i^{x_i}$, where $\mathbf{x} = (x_1, \ldots, x_k) \in \{0,1\}^k$. Alice finds $A' = (\mathbf{a}_1, \ldots, \mathbf{a}_d) \in \left[\{0,1\}^{\lceil \log(\sqrt{\frac{n}{d}}) \rceil}\right]^d$ and Bob finds $B' = (\mathbf{b_1}, \ldots, \mathbf{b_1} \in \left[\{0,1\}^{\lceil \log(\sqrt{\frac{n}{d}}) \rceil}\right]^d$ privately without any communication, where $\mathbf{a_i} = \phi(\mathbf{x_i})$ and $\mathbf{b_i} = \phi(\bar{\mathbf{y}}_i)$ for each $i \in [d]$.

[6] For a set W, $[W]^d = W \times \ldots \times W$ (d times).

Step-4 For each $i \in [d]$, let L_i and L'_i be the lines having equation $y - q_i = \frac{decimal(\mathbf{a}_i)-1}{decimal(\mathbf{a}_i)}(x - p_i)$ and $x - p_i = decimal(\mathbf{b}_i)$ respectively. Alice finds $A'' \in \mathcal{T}$ that is generated by the lines $L_1, \ldots, L_d$, and Bob finds $B'' \in \mathcal{T}$ which is generated by the lines $L'_1, \ldots, L'_d$, i.e., $A'' = \bigcup_{i \in [d]} (L_i \cap G_{(p_i,q_i)})$ and $B'' = \bigcup_{i \in [d]} (L'_i \cap G_{(p_i,q_i)})$.

Step-5 Then Alice and Bob solve $\textsc{Disj}_G \mid_{\mathcal{T}\times\mathcal{T}} (A'', B'')$, and report $\bigvee_{i=1}^{d} \textsc{Disj}_{\{0,1\}^k}(\mathbf{x}_i, \mathbf{y}_i) = 1$ if and only if $\textsc{Disj}_G \mid_{\mathcal{T}\times\mathcal{T}} (A'', B'') = 0$.

Now we argue for the correctness of the protocol $\mathcal{P}'$. Let $\textsc{Disj}_G \mid_{\mathcal{T}\times\mathcal{T}} (A'', B'') = 0$, that is, $A'' \cap B'' \neq \emptyset$. By Claim 17 and from the description of $\mathcal{P}'$, there exists $i \in [d]$ such that the lines $L_i : y - q_i = \frac{decimal(\mathbf{a}_i)-1}{decimal(\mathbf{a}_i)}(x - p_i)$ and $L'_i : x - p_i = decimal(\mathbf{b_i})$ intersect at a point in $G_{(p_i,q_i)}$, that is, the lines $y = \frac{decimal(\mathbf{a}_i)-1}{decimal(\mathbf{a}_i)}x$ and $x = decimal(\mathbf{b}_i)$ intersect at a point in $G_{(0,0)}$. Now, we can say that, there exists $i \in [d]$ such that $decimal(\mathbf{a}_i)$ divides $decimal(\mathbf{b_i})$, equivalently, $\phi(\mathbf{x}_i)$ divides $\phi(\bar{\mathbf{y}}_i)$. This implies $\mathbf{x}_i$ is a subset of $\bar{\mathbf{y}}_i$ (or $\mathbf{x}_i \cap \mathbf{y_i} = \emptyset$) for some $i \in [d]$. Hence, $\bigvee_{i=1}^{d} \textsc{Disj}_{\{0,1\}^k}(\mathbf{x}_i, \mathbf{y}_i) = 1$. The converse part, that is, $\bigvee_{i=1}^{d} \textsc{Disj}_{\{0,1\}^k}(\mathbf{x}_i, \mathbf{y}_i) = 1$ implies $\textsc{Disj}_G \mid_{\mathcal{T}\times\mathcal{T}} (A'', B'') = 0$ can be shown in the similar fashion.

Observe that the communication cost of protocol $\mathcal{P}'$ for $\textsc{Or-Disj}^d_{\{0,1\}^k} \mid_{S\times S}$ is same as that of protocol $\mathcal{P}$ for $\textsc{Disj}_G \mid_{\mathcal{T}\times\mathcal{T}}$, which is $o\left(d\frac{\log m}{\log\log m}\right) = o\left(d\frac{\log(n/d)}{\log\log(n/d)}\right) = o(dk)$ as $m = \sqrt{\frac{n}{d}}$ and $k = \Theta\left(\frac{\log(n/d)}{\log\log(n/d)}\right)$. This contradicts Proposition 20 which says that $R\left(\textsc{Or-Disj}^d_{\{0,1\}^k} \mid_{S\times S}\right) = \Omega(dk)$.

3.3 Proof of Theorem 14

With out loss of generality, we also assume that d divides n and, more over, n/d is a perfect square.

First, consider the problem $\textsc{Learn}_G \mid_{\mathcal{T}\times\mathcal{T}}$, where the objective of Alice and Bob is to learn each other's set. Note that $G \subset \mathbb{Z}^2$ with $|G| = n$ and $\mathcal{T} \subseteq 2^G$ with VC-Dim$(\mathcal{T}) = 2d$ are same as that constructed in Section 3.1. In $\textsc{Learn}_G \mid_{\mathcal{T}\times\mathcal{T}}$, Alice and Bob get two sets A and B, respectively, from $\mathcal{T}$ with a promise $|A \cap B| = d$. The objective of Alice (Bob) is to learn B (A). Observe that $R(\textsc{Learn}_G \mid_{\mathcal{T}\times\mathcal{T}}) = \Omega(d \log n)$ as there are $\Omega(m^d) = \Omega\left(\left(\sqrt{\frac{n}{d}}\right)^d\right)$ many candidate sets for the inputs of Alice and Bob. We prove the theorem by a reduction from $\textsc{Learn}_G \mid_{\mathcal{T}\times\mathcal{T}}$ to $\textsc{Int}_G \mid_{\mathcal{T}\times\mathcal{T}}$.

Let by contradiction consider a protocol $\mathcal{P}$ that solves $\textsc{Int}_G \mid_{\mathcal{T}\times\mathcal{T}}$ by using $o(d \log n)$ bits of communication. To solve $\textsc{Learn}_G \mid_{\mathcal{T}\times\mathcal{T}}$, Alice and Bob first run a protocol $\mathcal{P}$ and finds $A \cap B$. Now by Claim 17, it is possible for Alice (Bob) to determine B (A) by combining A (B) along with $A \cap B$, with out any communication with Bob (Alice). Now, we have a protocol $\mathcal{P}'$ that solves $\textsc{Learn}_G \mid_{\mathcal{T}\times\mathcal{T}}$ with $o(d \log n)$ bits of communication. However, this is impossible as $R(\textsc{Learn}_G \mid_{\mathcal{T}\times\mathcal{T}}) = \Omega(d \log n)$. Hence, we are done with the proof of Theorem 14.

4 Conclusion and Discussion

In this paper, we studied $\textsc{Disj}_n \mid_{\mathcal{S}\times\mathcal{S}}$ and $\textsc{Int}_n \mid_{\mathcal{S}\times\mathcal{S}}$ when $\mathcal{S}$ is a subset of $2^{[n]}$ and VC-dim$(\mathcal{S}) \leq d$. One of the main contributions of our work is the result (Theorem 6) showing that unlike in the case of d-$\textsc{SparseDisj}_n$ and d-$\textsc{SparseDisj}_n$ functions, there is no separation between randomized and deterministic communication complexity of $\textsc{Disj}_n \mid_{\mathcal{S}\times\mathcal{S}}$ and

$\textsc{Int}_n\ |_{\mathcal{S}\times\mathcal{S}}$ functions when VC-dim$(S) \leq d$. Note that we have settled both the one-way and two-way (randomized) communication complexities of $\textsc{Int}_n\ |_{\mathcal{S}\times\mathcal{S}}$ when VC-dim$(\mathcal{S}) \leq d$ (Theorem 6 (1) and (3)). In the context of $\textsc{Disj}_n\ |_{\mathcal{S}\times\mathcal{S}}$, we have settled the one-way (randomized) communication complexity. The two-way communication complexity for $\textsc{Disj}_n\ |_{\mathcal{S}\times\mathcal{S}}$ is tight up to factor $\log\log\frac{n}{d}$ (See Theorem 6 (2)). However, we believe that the factor of $\log\log\frac{n}{d}$ should not be present in the statement of Theorem 6 (2).

▶ **Conjecture 21.** *There exists $\mathcal{S} \subseteq 2^{[n]}$ with VC-dim$(\mathcal{S}) \leq d$ and $R(\textsc{Disj}_n\ |_{\mathcal{S}\times\mathcal{S}}) = \Omega\left(d\log\frac{n}{d}\right)$.*

Recall $G \subset \mathbb{Z}^2$ with $|G| = n$ and $\mathcal{T} \subseteq 2^G$ with VC-Dim$(\mathcal{T}) = 2d$ construction from Section 3.1, that served as the hard instance for the proof of Theorem 13 and Theorem 14. The same G and $\mathcal{T}$ can not be the hard instance for the proof of Conjecture 21 because of the following result.

▶ **Theorem 22.** *Let us consider $G \subset \mathbb{Z}^2$ with $|G| = n$ and $\mathcal{T} \subseteq 2^G$ with VC-Dim$(\mathcal{T}) = 2d$ as defined in Section 3.1. Also, recall the definition of $\mathcal{T}_1$ and $\mathcal{T}_2$. There exists a randomized communication protocol that can, $\forall A \in \mathcal{T}_1$ and $\forall B \in \mathcal{T}_2$, can compute $\textsc{Disj}_G\ |_{\mathcal{T}\times\mathcal{T}}\ (A, B)$, with probability at least $2/3$, and uses $O\left(\frac{d\log d\log\frac{n}{d}}{\log\log\frac{n}{d}} \cdot \log\log\log\frac{n}{d}\right)$ bits of communication.*

We use the following observation to prove the above theorem.

▶ **Observation 23.** Let us consider the communication problem $\textsc{Gcd}_k(a, b)$, where Alice and Bob get a and b respectively from $\{1, \ldots, k\}$, and the objective is for both the players to compute $\gcd(a, d)$. Then there exists a randomized protocol, with success probability at least $1 - \delta$, for $\textsc{Gcd}_k$ that uses $O\left(\frac{\log k}{\log\log k} \cdot \log\log\log k \cdot \log\frac{1}{\delta}\right)$ bits of communication.

Proof. We will give a protocol P for the case when $\delta = 1/3$ that uses $O\left(\frac{\log k}{\log\log k} \cdot \log\log\log k\right)$ bits of communication. By repeating $O\left(\log\frac{1}{\delta}\right)$ times protocol $\mathcal{P}$ and reporting the majority of the outcomes as the output, we will get the correct answer with probability at least $1 - \delta$. Both Alice and Bob generate all the prime numbers $p_1, \ldots, p_t$ between 1 and k. From the Prime Number Theorem, we known that $t = \Theta\left(\frac{k}{\log k}\right)$. Alice and Bob separately, construct the sets S_a and S_b that contain the prime numbers that divides a and b respectively. Note that $|S_a|$ and $|S_b|$ is bounded by $O\left(\frac{\log k}{\log\log k}\right)$.[7] Alice and Bob compute $S_a \cap S_b$ by solving *Sparse Set Intersection* problem on input S_a and S_b using $O\left(\frac{\log k}{\log\log k}\right)$ bits of communication [2]. For $p \in S_a \cap S_b$, let $\alpha_{p,a}$ and $\alpha_{p,b}$ denote the exponent of p in a and b, respectively. Observe that

$$\gcd(a, b) = \prod_{p\in S_a\cap S_b} p^{\min\{\alpha_{p,a},\alpha_{p,b}\}}.$$

For each $p \in S_a$, Alice sends $\alpha_{p,a}$ to Bob. Number of bits of communication required to send the exponents of all the primes in $S_a \cap S_b$, is

[7] The product of first t prime numbers is $e^{(1+o(1))t\log t}$.

$$
\begin{aligned}
|S_a \cap S_b| + \sum_{p \in S_a \cap S_b} \log(\alpha_{p,a}) &\leq O\left(\frac{\log k}{\log\log k}\right) + |S_a \cap S_b| \log \left(\frac{\sum\limits_{p \in S_a \cap S_b} \alpha_{p,a}}{|S_a \cap S_b|} \right) \\
&\leq O\left(\frac{\log k}{\log\log k}\right) + |S_a \cap S_b| \log \left(\frac{\log k}{|S_a \cap S_b|} \right) \\
&\leq O\left(\frac{\log k}{\log\log k} \cdot \log\log\log k\right)
\end{aligned}
$$

In the above inequalities, we used the facts that $|S_a \cap S_b| = O\left(\frac{\log k}{\log\log k}\right)$, $\sum\limits_{p \in S_a \cap S_b} \alpha_{p,a} \leq \log k$ and $\log x$ is a concave function. After getting the exponents $\alpha_{p,a}$ of the primes $p \in S_a \cap S_b$ from Alice, Bob also sends the exponents $\alpha_{p,b}$ of the primes $p \in S_a \cap S_b$ to Alice using $O\left(\frac{\log k}{\log\log k} \log\log\log k\right)$ bits of communication to Alice. Since both Alice and Bob now know the set $S_a \cap S_b$, and the exponents $\alpha_{p,a}$ and $\alpha_{p,b}$ for all $p \in S_a \cap S_b$, both of them can compute $\gcd(a, b)$. Total number of bits communicated in this protocol is $O\left(\frac{\log k}{\log\log k} \log\log\log k\right)$. ◀

We will now give the proof of Theorem 22.

Proof of Theorem 22. Consider the case when $d = 1$. From the description of G and $\mathcal{T}$ in Section 3.1, we can say that $G = G_{(0,0)}$, where $G_{(0,0)} = \{(x, y) \in \mathbb{Z}^2 : 0 \leq x, y \leq \sqrt{n}\}$ [8]. Moreover, each set in $\mathcal{T}_1$ is a set of points present in a straight line of non-negative slope that passes through two points of $G_{(0,0)}$ with one point being $(0, 0)$ and each set in $\mathcal{T}_2$ is a set of points present in a vertical straight line that passes through exactly $\sqrt{n}$ many grid points. Keeping Claims 17 and 18 in mind, we will be done if we can show the existence of a randomized communication protocol for computing the function $\textsc{Disj}_G \mid_{\mathcal{T} \times \mathcal{T}}$, with probability of success at least $1 - \delta$ and number of bits communicated by the protocol being bounded by $O\left(\frac{\log n}{\log\log n} \cdot \log\log\log n \cdot \log \frac{1}{\delta}\right)$, for the special case when $d = 1$. This is because for general d, we will be solving d instances of the above problem, with the number of points in each grid being $\frac{n}{d}$ [9] and setting $\delta = \frac{1}{3d}$ for each of the d instances.

Protocol for $d = 1$

Alice and Bob get A and B from $\mathcal{T}_1$ and $\mathcal{T}_2$, respectively. Let A is generated by the straight line L_A and B is generated by L_B, where L_A is a straight line with non-negative slope and L_B is a vertical line. If L_A is a horizontal one : Alice just sends this information to Bob and then both report that $A \cap B \neq \emptyset$. If L_A is a vertical line : Alice sends this information to Bob and he reports $A \cap B \neq \emptyset$ if and only if L_B passes through origin. Now assume that L_A is neither a horizontal nor a vertical line. Let the equation of L_A be $y = \frac{p}{q}x$, where $1 \leq p, q \leq \sqrt{n}$, and p and q are relatively prime to each other. Also, let equation of Bob's line L_B be $x = r$, where $0 \leq r \leq \sqrt{n}$. Observe that $A \cap B \neq \emptyset$ if and only if L_A and L_B intersects at a point of $G_{(0,0)}$. Moreover, L_A and L_B intersects at a grid point if and only if q divides r and $1 \leq \frac{pr}{q} \leq \sqrt{n}$. So, Alice and Bob run the communication protocol for $\textsc{Gcd}_{\sqrt{n}}(q, r)$ to decide whether $q = \gcd(q, r)$. If $q = \gcd(q, r)$ and $1 \leq \frac{pr}{q} \leq \sqrt{n}$ (again Alice and Bob can decide this using $O(1)$ bits of communications) then $A \cap B \neq \emptyset$, otherwise $A \cap B = \emptyset$. Alice and Bob can decide if $q = \gcd(q, r)$ and $1 \leq \frac{pr}{q} \leq \sqrt{n}$ using $O(1)$ bits of communication.

[8] With out loss of generality assume that $\sqrt{n}$ is an integer

[9] Recall that we have assumed, without loss of generality, that d divides n.

The communication cost of our protocol is dominated by the communication complexity of $\textsc{Gcd}_{\sqrt{n}}(q, r)$, which is equal to $O\left(\frac{\log n}{\log\log n}\log\log\log n\log\frac{1}{\delta}\right)$ by Observation 23. ◀

References

1 Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar, and D. Sivakumar. An Information Statistics Approach to Data Stream and Communication Complexity. *J. Comput. Syst. Sci.*, 68(4):702–732, 2004.

2 Joshua Brody, Amit Chakrabarti, Ranganath Kondapally, David P. Woodruff, and Grigory Yaroslavtsev. Beyond Set Disjointness: The Communication Complexity of Finding the Intersection. In *ACM Symposium on Principles of Distributed Computing, PODC*, pages 106–113, 2014.

3 Bernard Chazelle. *The Discrepancy Method: Randomness and Complexity*. Cambridge University Press, 2001.

4 Anirban Dasgupta, Ravi Kumar, and D. Sivakumar. Sparse and Lopsided Set Disjointness via Information Theory. In *Proceedings of the RANDOM-APPROX*, pages 517–528, 2012.

5 Sariel Har-Peled. *Geometric Approximation Algorithms*. Number 173 in Mathematical Surveys and Monographs. American Mathematical Soc., 2011.

6 Johan Håstad and Avi Wigderson. The Randomized Communication Complexity of Set Disjointness. *Theory of Computing*, 3(1):211–219, 2007.

7 T. S. Jayram, Ravi Kumar, and D. Sivakumar. Two Applications of Information Complexity. In *Proceedings of the 35th Annual ACM Symposium on Theory of Computing, STOC*, pages 673–682, 2003.

8 Bala Kalyanasundaram and Georg Schnitger. The Probabilistic Communication Complexity of Set Intersection. *SIAM J. Discrete Math.*, 5(4):545–557, 1992.

9 Eyal Kushilevitz and Noam Nisan. *Communication Complexity*. Cambridge University Press, USA, 1996.

10 Jiri Matousek. *Geometric Discrepancy: An Illustrated Guide*, volume 18. Springer Science & Business Media, 2009.

11 Jiri Matousek. *Lectures on Discrete Geometry*, volume 212. Springer Science & Business Media, 2013.

12 Peter Bro Miltersen, Noam Nisan, Shmuel Safra, and Avi Wigderson. On Data Structures and Asymmetric Communication Complexity. *J. Comput. Syst. Sci.*, 57(1):37–49, 1998.

13 Ilan Newman. Private vs. common random bits in communication complexity. *Information Processing Letters*, 39(2):67–71, 1991.

14 János Pach and Pankaj K Agarwal. *Combinatorial Geometry*, volume 37. John Wiley & Sons, 2011.

15 Anup Rao and Amir Yehudayoff. *Communication Complexity: and Applications*. Cambridge University Press, 2020.

16 Alexander A. Razborov. On the Distributional Complexity of Disjointness. *Theor. Comput. Sci.*, 106(2):385–390, 1992.

17 Tim Roughgarden. Communication Complexity (for Algorithm Designers). *Foundations and Trends in Theoretical Computer Science*, 11(3-4):217–404, 2016.

18 Mert Saglam and Gábor Tardos. On the Communication Complexity of Sparse Set Disjointness and Exists-Equal Problems. In *Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science, FOCS*, pages 678–687, 2013.

19 N. Sauer. On the Density of Families of Sets. *Journal of Combinatorial Theory, Series A*, 13(1):145–147, 1972.

20 S. Shelah. A Combinatorial Problem, Stability and Order for Models and Theories in Infinitary Languages. *Pacific Journal of Mathematics*, 41:247–261, 1972.

21 V. N. Vapnik and A. Y. Chervonenkis. On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.

22 Andrew Chi-Chih Yao. Some Complexity Questions Related to Distributive Computing (Preliminary Report). In *Proceedings of the 11h Annual ACM Symposium on Theory of Computing, STOC*, pages 209–213, 1979.

A VC dimension, and Problems 1 and 2

VC dimension, and collection of d lines

Let $G \subset \mathbb{Z}^2$ be a set of n points in $\mathbb{Z}^2$. Observe, that the communication functions $\textsc{Disj}_G \mid_{\mathcal{L}\times\mathcal{L}}$ (defined in Problem 1) and $\textsc{Disj}_G \mid_{\mathcal{G}\times\mathcal{G}}$, where

$$\mathcal{G} = \left\{ G \cap \left(\bigcup_{1\le j\le d} \ell_j \right) \mid \{\ell_1, \ldots, \ell_d\} \in \mathcal{L} \right\},$$

are equivalent problems. Note that the set $\mathcal{L}$ is defined in Problem 1. Using standard geometric arguments, see [11, Chap. 10] and [5, Chap. 5], we can show that VC-dim$(\mathcal{G}) = 2d$.

VC dimension, and collection of d intervals

Let $X \subset \mathbb{Z}$ be a set of n points in $\mathbb{Z}$. Observe, that the communication functions $\textsc{Disj}_X \mid_{\mathcal{I}\times\mathcal{I}}$ (defined in Problem 2) and $\textsc{Disj}_X \mid_{\mathcal{F}\times\mathcal{F}}$, where

$$\mathcal{F} = \left\{ X \cap \left(\bigcup_{1\le j\le d} I_j \right) \mid \{I_1, \ldots, I_d\} \in \mathcal{I} \right\},$$

are equivalent problems. Note that the set $\mathcal{I}$ is defined in Problem 2. Using standard geometric arguments, as in the above case, we can show that VC-dim$(\mathcal{F}) = 2d$.