comput. complex. (2022) 31:9

https://doi.org/10.1007/s00037-022-00225-6

computational complexity

DISJOINTNESS THROUGH THE LENS OF VAPNIK–CHERVONENKIS DIMENSION: SPARSITY AND BEYOND

Anup Bhattacharya, Sourav Chakraborty, Arijit Ghosh, Gopinath Mishra, and Manaswi Paraashar

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 are 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 of the disjointness problem is always upper bounded by a function of the VC dimension of the set system, and does there always exist a gap between

Published online: 05 July 2022

the deterministic and randomized communication complexities of the disjointness problem for set systems with small VC dimension.
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 exist 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 $\Theta\left(d \log\left(n/d\right)\right)$.

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

Subject classification.
94A05 and 52C35

1. Introduction

Since its introduction by Yao (1979), 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 (see, e.g., Kushilevitz & Nisan (1996), Roughgarden (2016), and Rao & Yehudayoff (2020)). Vapnik & Chervonenkis (1971) 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 (see, e.g., Matousek (2009), Chazelle (2001), Pach & Agarwal (2011) and Matousek (2013)). In this work, we study communication complexity under the lens of *restricted systems* and, for the first time, prove that *geometric simplicity* does not necessarily imply *better* communication complexity.

Let us 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 devise 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 (over the inputs) 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 $D^{\to}(f)$ and $R^{\to}(f)$ 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. Newman (1991) proved that the communication complexity in both the settings differ 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

$$\textsc{Disj}_{\mathcal{U}}(x, y) = \begin{cases} 1, & \text{if } x \cap y = \emptyset \\ 0, & \text{otherwise.} \end{cases} \tag{1.1}$$

We also define the *intersection function*. Given a universe $\mathcal{U}$ known to both Alice and Bob, the *intersection function*, $\textsc{Int}_{\mathcal{U}} : 2^{\mathcal{U}} \times 2^{\mathcal{U}} \to 2^{\mathcal{U}}$ is defined as $\textsc{Int}_{\mathcal{U}}(x, y) = x \cap y$. It is easy to see that $\textsc{Int}_{\mathcal{U}}$ is a 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 (see, e.g., Roughgarden (2016) and Rao & Yehudayoff (2020)). 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* (see, e.g., Kushilevitz & Nisan (1996) and Rao & Yehudayoff (2020)), one can show that $D(\textsc{Disj}_n) = \Theta(n)$. In a breakthrough paper, Kalyanasundaram & Schnitger (1992) proved that $R(\textsc{Disj}_n) = \Omega(n)$. Razborov (1992) and Bar-Yossef *et al.* (2004) gave alternate proofs for the above result. From the above cited results, we can also see the $D(\textsc{Int}_n) = R(\textsc{Int}_n) = \Theta(n)$. $R(\textsc{Disj}_n) = R(\textsc{Int}_n) = \Theta(n)$ also follow from a recent result by Braverman *et al.* (2013).

Naturally, one would also like to ask what happens to the deterministic and randomized communication complexities (one way or multiple rounds) 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. We will denote by $\textsc{Disj}_{\mathcal{U}} \mid_{\mathcal{S} \times \mathcal{S}}$ the function $\textsc{Disj}_{\mathcal{U}}$ *restricted* to $\mathcal{S} \times \mathcal{S}$. This problem has 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.

We will denote by d-SPARSEDISJ$_n$ and d-SPARSEINT$_n$, the functions $\text{DISJ}_n \mid_{\mathcal{S}\times\mathcal{S}}$ and $\text{INT}_n \mid_{\mathcal{S}\times\mathcal{S}}$, respectively, where $\mathcal{S}$ is the collections of all subsets of $[n]$ of size at most d. Again, using the rank argument (see, e.g.,Kushilevitz & Nisan (1996) and Rao & Yehudayoff (2020)) one show that, for all $d \leq n$, the deterministic communication complexity of d-SPARSEDISJ$_n$ is $D(d\text{-SPARSEDISJ}_n) = \Omega(d \log(n/d))$. Håstad & Wigderson (2007) and Dasgupta *et al.* (2012) showed that the randomized communication complexity and one way randomized communication complexity of d-SPARSEDISJ$_n$ are $R(d\text{-SPARSEDISJ}_n) = \Theta(d)$ and $R^{\rightarrow}(d\text{-SPARSEDISJ}_n) = \Theta(d \log d)$, respectively. In a follow up work, Saglam & Tardos (2013) gave a randomized communication protocol that uses $O(\log^* d)$ rounds of communication and $O(d)$ bits of communication to compute d-SPARSEDISJ$_n$. More recently, Brody *et al.* (2014) proved that $R^{\rightarrow}(d\text{-SPARSEINT}_n) = \Theta(d \log d)$ and $R(d\text{-SPARSEINT}_n) = \Theta(d)$. These results show that in the d-sparse setting, there is a separation between randomized and deterministic communication complexity of DISJ$_n$ and INT$_n$ functions.

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

PROBLEM 1.2 (DISCRETE LINE DISJ). *Let L be the set of all lines in $\mathbb{R}^2$, and we denote by L^d the collection of all d-size subsets of L. Also, let $G \subset \mathbb{Z}^2$ be a set of n points in $\mathbb{Z}^2$, and $\mathcal{L} \subseteq L^d$. The* DISCRETE LINE DISJ *function on G and $\mathcal{L}$,* $\text{DISJ}_G \mid_{\mathcal{L}\times\mathcal{L}}: \mathcal{L}\times\mathcal{L} \to \{0,1\}$, *is defined as*

$$\text{DISJ}_G \mid_{\mathcal{L}\times\mathcal{L}} (\{\ell_1,\ldots,\ell_d\},\{\ell'_1,\ldots,\ell'_d\}) = \begin{cases} 1, & \text{if } \exists i,j \in [d] \text{ s.t. } \ell_i \cap \ell'_j \cap G \neq \emptyset \\ 0, & \text{otherwise} \end{cases}$$

In other words, $\text{DISJ}_G \mid_{\mathcal{L}\times\mathcal{L}} (\{\ell_1,\ldots,\ell_d\},\{\ell'_1,\ldots,\ell'_d\})$ *is 1 if and only if there exists a line in Alice's set $\{\ell_1,\ldots,\ell_d\}$ that intersects some line in Bob's set $\{\ell'_1,\ldots,\ell'_d\}$ at some point in G.*

Problem 1.3 (Discrete Interval Disj). *Let Int be the set of all possible intervals in $\mathbb{R}$ and Int^d denote the collection of all d-size subsets of Int. Let $X \subset \mathbb{Z}$ be a set of n points in $\mathbb{Z}$ and let $\mathcal{I} \subset Int^d$. The* Discrete Interval Disj *function on X and $\mathcal{I}$,* $\text{Disj}_X \mid_{\mathcal{I}\times\mathcal{I}} : \mathcal{I} \times \mathcal{I} \to \{0,1\}$*, is defined as*

$$\begin{aligned}&\text{Disj}_X \mid_{\mathcal{I}\times\mathcal{I}} (\{I_1, \ldots, I_d\}, \{I'_1, \ldots, I'_d\}) \\ &= \begin{cases} 1, & \textit{if } \exists i, j \in [d] \textit{ s.t. } I_i \cap I'_j \cap X \neq \emptyset \\ 0, & \textit{otherwise} \end{cases}\end{aligned}$$

In other words, $\text{Disj}_X \mid_{\mathcal{I}\times\mathcal{I}} (\{I_1, \ldots, I_d\}, \{I'_1, \ldots, I'_d\})$ *is* 1 *if and only if there exists an interval in Alice's set $\{I_1, \ldots, I_d\}$ that intersects some interval in Bob's set $\{I'_1, \ldots, I'_d\}$ at some point in X.*

Note that both the Discrete Line Disj and Discrete Interval Disj functions are generalizations of sparse set disjointness function.[2] Although it may not be obvious at first look, but both Discrete Line Disj and Discrete Interval Disj are disjointness functions restricted to a suitable subset. 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 $\text{Disj}_{\mathcal{U}} \mid_{\mathcal{S}\times\mathcal{S}}$. Here we would like to point out that special cases of Discrete Interval Disj and Discrete Line Disj implies a nontrivial lower bound for $\text{Disj}_{\mathcal{U}} \mid_{\mathcal{S}\times\mathcal{S}}$, and we will discuss these connections shortly. For the time being, we are interested in the following two questions:

Do the randomized communication complexities of Discrete Line Disj *function and* Discrete Interval Disj *function upper bounded by a function of d (independent of n)?*

Observe that an "Yes" answer to the above question implies that these functions also have a separation between their randomized

[2] 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.

and deterministic communication complexities similar to that of the *Sparse Set Disjointness* function (d-SPARSEDISJ$_n$). Unfortunately, the answer to the above question is negative.

THEOREM 1.4. *For* DISCRETE LINE DISJ*: there exists a $G \subset \mathbb{Z}^2$ with n points and $\mathcal{L} \subset L^d$ such that*

$$D(\textsc{Disj}_G\mid_{\mathcal{L}\times\mathcal{L}}) = D^{\rightarrow}(\textsc{Disj}_G\mid_{\mathcal{L}\times\mathcal{L}}) = \Theta\left(d\log\left(n/d\right)\right)$$

and, for the randomized setting, we have

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

THEOREM 1.5. *For* DISCRETE INTERVAL DISJ*: there exists a $X \subset \mathbb{Z}$ with n points and $\mathcal{I} \subset Int^d$ such that*

$$D(\textsc{Disj}_X\mid_{\mathcal{I}\times\mathcal{I}}) = D^{\rightarrow}(\textsc{Disj}_X\mid_{\mathcal{I}\times\mathcal{I}}) = \Theta\left(d\log\left(n/d\right)\right)$$

and, for the randomized setting, we have

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

DISCRETE LINE INT, that is, the intersection finding version of DISCRETE LINE DISJ is defined as follows : the objective is to compute a function $\textsc{Int}_G\mid_{\mathcal{L}\times\mathcal{L}}: \mathcal{L}\times\mathcal{L}\rightarrow G$ that is defined as

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

As we have already mentioned, $R(d\text{-}\textsc{SparseInt}_n) = \Theta(d)$ and $D(d\text{-}\textsc{SparseInt}_n) = \Theta\left(d\log\left(n/d\right)\right)$. We also show that DISCRETE LINE INT does not demonstrate such a separation between its deterministic and randomized communication complexities.

THEOREM 1.6. *For* DISCRETE LINE INT*: there exists a $G \subset \mathbb{Z}^2$ with n points and $\mathcal{L} \subset L^d$ such that*

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

and, for the randomized setting, we have

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

Note that Theorem 3.1, Theorem 2.1 and Theorem 3.2, given in Section 3, Section 2 and Section 3, will directly imply Theorem 1.4, Theorem 1.5 and Theorem 1.6, respectively. Note that the set systems used the proofs of Theorem 3.1, Theorem 2.1 and Theorem 3.2 have VC dimension $\Theta(d)$. For more details, see Section 3.1, Section 2.1 and Section 3.1.

Sauer-Shelah Lemma (see Sauer (1972), Shelah (1972) and Vapnik & Chervonenkis (1971)) states that if $\mathcal{S} \subseteq 2^{[n]}$ and VC-dim$(\mathcal{S})$ $= d$ then

$$|\mathcal{S}| \leq \sum_{i=0}^{d} \binom{n}{i} \leq \left(\frac{en}{d}\right)^d .$$

Thus if VC-dim$(\mathcal{S})$ $= d$, then the Sauer-Shelah Lemma implies that

$$D^{\rightarrow}(\textsc{Int}_n \mid_{\mathcal{S}\times\mathcal{S}}) = O\left(d \log\left(n/d\right)\right).$$

So, $O\left(d \log\left(n/d\right)\right)$ is an upper bound for randomized and deterministic and also for the one-way communication complexities. 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 $\mathcal{S}$ has VC dimension d? Using Theorem 3.1, Theorem 2.1 and Theorem 3.2, given in Section 3, Section 2 and Section 3, we get the following result.

Theorem 1.7 (**Main result**). *Let $1 \leq d \leq n$.*

(i) There exists $\mathcal{S} \subseteq 2^{[n]}$ with VC-dim$(\mathcal{S}) \leq d$ and

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

(ii) There exists $\mathcal{S} \subseteq 2^{[n]}$ with VC-dim$(\mathcal{S}) \leq d$ and

$$R^{\rightarrow}(\textsc{Disj}_n \mid_{\mathcal{S}\times\mathcal{S}}) = \Omega\left(d \log\left(n/d\right)\right).$$

(iii) There exists $\mathcal{S} \subseteq 2^{[n]}$ with VC-dim$(\mathcal{S}) \leq d$ and

$$R(\textsc{Int}_n \mid_{\mathcal{S}\times\mathcal{S}}) = \Omega\left(d \log\left(n/d\right)\right).$$

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 $\mathcal{S} \subset 2^{\mathcal{U}}$ of VC dimension d.

Problems	Previously Known	This Paper
$R(\text{DISJ}_n \mid_{\mathcal{S}\times\mathcal{S}})$	$\Omega(d)$ Håstad & Wigderson (2007)	$\Omega\left(d\frac{\log(n/d)}{\log\log(n/d)}\right)$
$R^{\to}(\text{DISJ}_n \mid_{\mathcal{S}\times\mathcal{S}})$	$\Omega(d\log d)$ Dasgupta *et al.* (2012)	$\Omega\left(d\log\left(n/d\right)\right)$
$R(\text{INT}_n \mid_{\mathcal{S}\times\mathcal{S}})$	$\Omega(d)$ Brody *et al.* (2014)	$\Omega\left(d\log\left(n/d\right)\right)$
$R^{\to}(\text{INT}_n \mid_{\mathcal{S}\times\mathcal{S}})$	$\Omega(d\log d)$ Brody *et al.* (2014)	$\Omega\left(d\log\left(n/d\right)\right)$

Table (1.1) This table gives the largest communication complexity for the functions $\text{DISJ}_n \mid_{\mathcal{S}\times\mathcal{S}}$ and $\text{INT}_n \mid_{\mathcal{S}\times\mathcal{S}}$, among all $\mathcal{S} \subseteq 2^{[n]}$ of VC dimension d, that was previously known and what we prove in this paper. Note that the lower bound of $\Omega\left(d\log\left(n/d\right)\right)$ for $D(\text{DISJ}_n \mid_{\mathcal{S}\times\mathcal{S}})$, $D^{\to}(\text{DISJ}_n \mid_{\mathcal{S}\times\mathcal{S}})$, $D(\text{INT}_n \mid_{\mathcal{S}\times\mathcal{S}})$, and $D^{\to}(\text{INT}_n \mid_{\mathcal{S}\times\mathcal{S}})$ in the worst case, among all $\mathcal{S} \subset 2^{[n]}$ 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 we have $D\left(\text{DISJ}_n \mid_{\mathcal{S}\times\mathcal{S}}\right) = D(\text{INT}_n \mid_{\mathcal{S}\times\mathcal{S}}) = \Omega\left(d\log\left(n/d\right)\right)$ (see, e.g., Kushilevitz & Nisan (1996) and Rao & Yehudayoff (2020)).

Notations. We denote the set $\{1,\ldots,n\}$ by $[n]$. For any vector $\mathbf{x} \in \{0,1\}^n$, $\text{num}(\mathbf{x})$ denotes the number whose binary representation over n bits is $\mathbf{x}$, that is, $\text{num}(\mathbf{x}) = \sum_{i=1}^{n} 2^{i-1}x_i$ where $\mathbf{x} = (x_1,\ldots,x_n)$. For two vectors $\mathbf{x}$ and $\mathbf{y}$ in $\{0,1\}^n$, $\mathbf{x}\cap\mathbf{y} = \{i \in [n] : x_i = y_i = 1\}$, and $\mathbf{x} \subseteq \mathbf{y}$ when $x_i \le y_i$ 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 $\{z \in \mathbb{R} \mid x \le z \le y\}$.

2. One way communication complexity

In this section, we prove the following result.

THEOREM 2.1. *For all $n \ge d$, there exists $X \subset \mathbb{Z}$ with $|X| = n$*

and $\mathcal{R} \subseteq 2^X$ *with VC-dim*$(\mathcal{R}) = 2d$*, such that*

$$\mathcal{R} \subseteq \left\{ X \cap \left(\bigcup_{1 \leq j \leq d} I_j \right) \;\middle|\; \{I_1, \ldots, I_d\} \in Int^d \right\}$$

and

$$R^{\rightarrow}(\textsc{Disj}_X \mid_{\mathcal{R}\times\mathcal{R}}) = \Omega\left(d \log\left(n/d\right)\right).$$

Note that the set Int^d *is defined in Problem 1.3.*

Remark 2.2. *The above result takes care of the proofs of Theorem 1.5 and Theorem 1.7 (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 prove Theorem 2.1 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 $x_{j'}$ for all $j' < j$. Bob wants to report x_j as the output.

Proposition 2.3 (Miltersen *et al.* 1998). $R^{\rightarrow}(\textsc{AugIndex}_\ell) = \Omega(\ell)$.

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 2.1.

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

Figure (2.1) Let us consider $d = 3, n = 18$ and $m = 4$. J_1, J_2 and J_3 are the intervals of length 4 starting from p_1, p_2 and p_3, respectively. The ground set X is the set of all 18 points present in three intervals.

and ends at $m+1$. Similarly, let J_p be the set of $m+2$ 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

$$X = \bigcup_{i=1}^{d} J_{p_i}.$$

Note that $X \subset \mathbb{Z}$ and $|X| = (m+2)d = n$. See Figure 2.1 for an illustration.

The subsets of X in $\mathcal{R}$: Before defining $\mathcal{R} \subseteq 2^X$, let us define sets $\mathcal{R}_0 \subset 2^X$ and $\mathcal{R}_{m+1} \subset 2^X$.

$\mathcal{R}_0 \subset 2^X$**:** Set of d intervals $R_1, \ldots, R_d$ of integer lengths are said to be *left good* if they satisfy the following: for all $i \in [d]$, we have $R_i = [p_i, q_i]$ where $q_i \in \{p_i, p_i+1, \ldots, p_i+m+1\}$. Note that R_i does not intersect with any $X \setminus J_{p_i}$. For a set of left good d-intervals $R_1, \ldots, R_d$, the set $A = \bigcup\limits_{i=1}^{d} (R_i \cap X)$ is said to be *generated* by $R_1, \ldots, R_d$. The set $\mathcal{R}_0 \subset 2^X$ is defined as:

$$\mathcal{R}_0 = \{A \mid A \text{ is generated by left good set of } d\text{-intervals}\}$$

$\mathcal{R}_{m+1} \subset 2^X$**:** Set of d-intervals $R'_1, \ldots, R'_d$ of integer lengths are said to be *right good* if they satisfy the following: for all $i \in [d]$, we have $R'_i = [q_i, p_i+m+1]$ where $q_i \in \{p_i, p_i+1, \ldots, p_i+m+1\}$. Note that R'_i does not intersect with any $X \setminus J_{p_i}$. For a set of right good d-intervals $R'_1, \ldots, R'_d$,

the set $B = \bigcup_{i=1}^{d} (R'_i \cap X)$ is said to be *generated* by $R'_1, \ldots, R'_d$.
The set $\mathcal{R}_{m+1} \subset 2^X$ is defined as:

$$\mathcal{R}_{m+1} = \{B \mid B \text{ is generated by right good set of } d\text{-intervals}\}$$

Finally, $\mathcal{R} = \mathcal{R}_0 \cup \mathcal{R}_{m+1}$.

See Figure 2.2 for an illustration.

(a) A and B *intersect.*

(b) A and B does not *intersect.*

(a) A and B *intersect.*

(b) A and B does not *intersect.*

Figure (2.2) Consider n, d, m and X as in Figure 2.1. A is the set of points in X that are present in the three blue intervals. Similarly, B is the set of points in X that are present in the three red intervals.

The following claim bounds the VC dimension of $\mathcal{R}$.

Claim 2.4. *For $X \subset \mathbb{Z}$ with $|X| = n$ and $\mathcal{R} \subset 2^X$, as described above, we have VC-dim$(\mathcal{R}) = 2d$.*

Proof. The proof follows from the fact that any subset of X containing $2d + 1$ points will contain at least three points from some J_{p_i}, where $i \in [d]$. These points in J_{p_i} cannot 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 $\mathcal{R}$. □

The following claim will be used in the proof of Theorem 2.1.

Claim 2.5. *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 2.5 follows directly from our construction of $X \subset \mathbb{Z}$ and $\mathcal{R} \subseteq 2^X$, and the fact that $J_{p_1}, \ldots, J_{p_d}$ are pairwise disjoint.

2.2. Reduction from $\textsc{AugIndex}_{d \log m}$ to $\textsc{Disj}_X\,|_{\mathcal{R}\times\mathcal{R}}$. Before presenting the reduction, we recall the definitions of $\textsc{AugIndex}_{d\log m}$ and $\textsc{Disj}_X|_{\mathcal{R}\times\mathcal{R}}$. In $\textsc{AugIndex}_{d\log m}$, 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 $\textsc{Disj}_X|_{\mathcal{R}\times\mathcal{R}}$, 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 Section 2.1.

Let $\mathcal{P}$ be a one-way protocol that solves $\textsc{Disj}_X|_{\mathcal{R}\times\mathcal{R}}$ using $o\left(d \log \frac{n}{d}\right) = o(d\log m)$ bits of communication. Now, we consider the following protocol $\mathcal{P}'$ for $\textsc{AugIndex}_{d\log m}$ that has the same one way communication cost as that of $\textsc{Disj}_X|_{\mathcal{R}\times\mathcal{R}}$. Then we will be done with the proof of Theorem 2.1.

Protocol $\mathcal{P}'$ for $\textsc{AugIndex}_{d\log m}$ 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, $\forall i \in [d]$, we have

$$\mathbf{a}_i = x_{(i-1)\log m + 1} \cdots x_{i \log m}.$$

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 $\mathbf{y}$ into d parts such that, $\forall i \in [d]$, we have

$$\mathbf{b}_i = y_{(i-1)\log m + 1} \cdots y_{i \log m}.$$

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

$$R_i = [p_i, m + p_i - \text{num}(\mathbf{a}_i)]$$

and

$$R'_i = [p_i + m + 1 - \text{num}(\mathbf{b}_i), p_i + m + 1].$$

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 $R'_1, \ldots, R'_d$, that is,

$$A = \bigcup_{i \in [d]} (R_i \cap X) \text{ and } B = \bigcup_{i \in [d]} (R'_i \cap X).$$

Step-4 Alice and Bob solves $\text{DISJ}_X \mid_{\mathcal{R} \times \mathcal{R}}$ on inputs A and B, and report $x_j = 0$ if and only if $\text{DISJ}_X \mid_{\mathcal{R} \times \mathcal{R}} (A, B) = 0$. Note that x_j is the output of $\text{AUGINDEX}_{d \log m}$ 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 2.6. *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 num$(\mathbf{b}_{i^*}) \leq$ num$(\mathbf{a}_{i^*})$.

(iii) num$(\mathbf{b}_{i^*}) >$ num$(\mathbf{a}_{i^*})$ 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 2.5, there exists an $i \in [d]$ such that R_i and R'_i intersects at a point in J_{p_i}. From Observation 2.6 (i), we can say $R_{i*} \cap R'_{i^*} \neq \emptyset$. Combining $R_{i*} \cap R'_{i^*} \neq \emptyset$ with Observation 2.6 (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, that is, $x_j = 0$ 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 $\textsc{AugIndex}_{d \log m}$ is the same as that of $\mathcal{P}$ for $\textsc{Disj}_X|_{\mathcal{R}\times\mathcal{R}}$, 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 $\Omega(d \log m) = \Omega\left(d \log\left(n/d\right)\right)$ bits. This completes the proof of Theorem 2.1.

3. Two way communication complexity

In this section, we prove the following theorems.

Theorem 3.1. *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*

$$\mathcal{T} \subseteq \left\{ G \cap \left(\bigcup_{1 \leq j \leq d} \ell_j \right) \;\middle|\; \{\ell_1, \ldots, \ell_d\} \in L^d \right\}$$

and

$$R(\textsc{Disj}_G\,|_{\mathcal{T}\times\mathcal{T}}) = \Omega\left(d \frac{\log(n/d)}{\log\log(n/d)} \right).$$

The set L^d is as defined in Problem 1.2.

Theorem 3.2. *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*

$$\mathcal{T} \subseteq \left\{ G \cap \left(\bigcup_{1 \leq j \leq d} \ell_j \right) \;\middle|\; \{\ell_1, \ldots, \ell_d\} \in L^d \right\}$$

and

$$R(\textsc{Int}_G\,|_{\mathcal{T}\times\mathcal{T}}) = \Omega\left(d \log\left(n/d\right)\right).$$

The set L^d is as defined in Problem 1.2.

Remark 3.3. *Theorem 3.1 takes care of Theorem 1.4 and Theorem 1.7 (2). Theorem 3.2 takes care of Theorem 1.6 and Theorem 1.7 (3).*

Note that the same set system will be used for proving both of the above theorems. The *hard* instance used in the proof of the above theorems is inspired by *point line incidence* set systems from combinatorial geometry, see Section 3.1 for the details. We prove Theorem 3.1 and Theorem 3.2 in Section 3.2 and Section 3.3, respectively.

3.1. Set system used in the proofs of Theorem 3.1 and Theorem 3.2. In this subsection, we give the descriptions 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 in the proofs of Theorem 3.1 and Theorem 3.2. 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 n/d points and the grids are taken in such a way that any straight line of non-negative slope 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. Formal details of the constructions of G and $\mathcal{T}$ are given below along with some of their properties.

The ground set G: Let $m = \sqrt{\frac{n}{d}}$, and

$$G_{(0,0)} := \left\{(x, y) \in \mathbb{Z}^2 \mid 0 \leq x, y \leq m - 1\right\}$$

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 $G_{(p,q)}$, that is,

$$G_{(p,q)} := \left\{(i + p, j + q) \mid (i, j) \in G_{(0,0)}\right\}.$$

For $d \in \mathbb{N}$, consider $G_{(p_1,q_1)}, \ldots, G_{(p_d,q_d)}$ satisfying the following property:

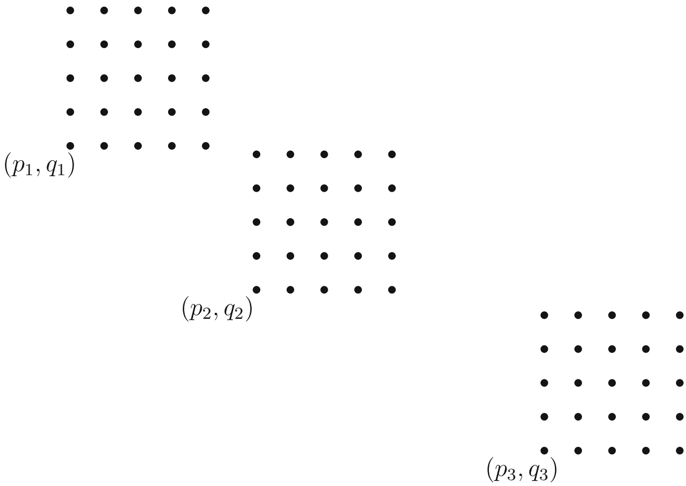

Figure (3.1) Let us take $n = 75, d = 3$ and $m = 5$. The 5×5 grids centered at $(p_1, q_1), (p_2, q_2)$ and (p_3, q_3) are $G_{(p_1,q_1)}$, $G_{(p_2,q_2)}$ and $G_{(p_3,q_3)}$; respectively. The ground set G is the set of all 75 points present in three grids.

<u>Property</u> *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. See Figure 3.1 for an illustration. We will take the ground set G as

$$G := \bigcup_{\ell=1}^{d} G_{(p_\ell,q_\ell)}.$$

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 $B = \bigcup\limits_{i=1}^{d} \left(L'_i \cap G_{(p_i,q_i)}\right)$ is in $\mathcal{T}_2$, and we say B is *generated* by the lines $L'_1, \ldots, L'_d$.

See Figure 3.2 for an illustration.

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

Claim 3.4. *For $G \subset \mathbb{Z}^2$ and $\mathcal{T} \subseteq 2^G$ as described above, we have 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 $G_{(p_j,q_j)}$ cannot be shattered by the sets in $\mathcal{T}$. Also, observe that there exists $2d$ points from G, two from each $G_{(p_j,q_j)}$, that can be shattered by the sets in $\mathcal{T}$. $\square$

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 3.5. *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 $L'_1, \ldots, L'_d$. 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 3.6. *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 Theorem 3.1 and Theorem 3.2.

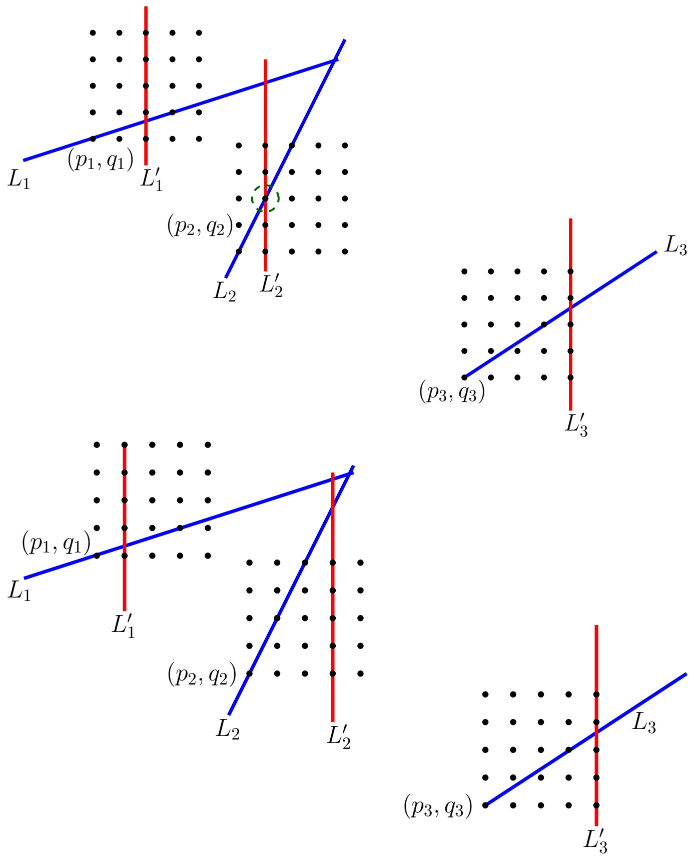

Figure (3.2) Consider n, d, m and G as in Figure 3.1. A is the set of points in G that are present in three blue lines, that is, $L_1 \cup L_2 \cup L_3$. Similarly, B is the set of points in G that are present in three red line $L_1' \cup L_2' \cup L_3'$. First figure shows the instance where A and B *intersect* at a grid point, and the second figure shows an instance where A and B does not *intersect* at a grid point.

3.2. Proof of Theorem 3.1. Let us consider a problem in communication complexity denoted by $\textsc{Or-Disj}^t_{\{0,1\}^\ell}$ that will be used in our proof. In $\textsc{Or-Disj}^t_{\{0,1\}^\ell}$, 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

$$\textsc{Or-Disj}^t_{\{0,1\}^\ell}\left(\left(\mathbf{x}_1, \ldots, \mathbf{x}_t\right), \left(\mathbf{y}_1, \ldots, \mathbf{y}_t\right)\right) = \bigvee_{i=1}^{t} \textsc{Disj}_{\{0,1\}^\ell}\left(\mathbf{x}_i, \mathbf{y}_i\right).$$

Note that $\textsc{Disj}_{\{0,1\}^\ell}(\mathbf{x}_i, \mathbf{y}_i)$ is a binary variable that takes value 1 if and only if $\mathbf{x}_i \cap \mathbf{y}_i = \emptyset$.

Proposition 3.7 (Jayram *et al.* 2003). *$R\left(\textsc{Or-Disj}^t_{\{0,1\}^\ell}\right) = \Omega(\ell t)$.*

Note that Proposition 3.7 directly implies the following result.

Proposition 3.8. *$R\left(\textsc{Or-Disj}^t_{\{0,1\}^\ell} |_{S_\ell \times S_\ell}\right) = \Omega(\ell t)$, where $S_\ell = \{0,1\}^\ell \setminus \{0^\ell\}$.*

Let $k \in \mathbb{N}$ be the largest integer such that first k consecutive primes $\pi_1, \ldots, \pi_k$ satisfy the following inequality:

$$\prod_{i=1}^{k} \pi_i \leq \sqrt{\frac{n}{d}}. \tag{3.9}$$

Using the fact that

$$\prod_{i=d}^{k} \pi_i = e^{(1+o(1))k \log k},$$

we get

$$k = \Theta\left(\frac{\log(n/d)}{\log\log(n/d)}\right).$$

We prove the theorem by a reduction from $\textsc{Or-Disj}^d_{\{0,1\}^k} |_{S_k \times S_k}$ to $\textsc{Disj}_G |_{\mathcal{T} \times \mathcal{T}}$, 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 $\textsc{Disj}_G \;|_{\mathcal{T}\times\mathcal{T}}$ with communication cost of

$$o\left(d\frac{\log m}{\log\log m}\right) = o\left(d\frac{\log(n/d)}{\log\log(n/d)}\right).$$

We will now give the details of the protocol $\mathcal{P}'$ that computes the function $\textsc{Or-Disj}^d_{\{0,1\}^k} \;|_{S_k\times S_k}$, and it will use protocol $\mathcal{P}$ as a subroutine.

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

Step-1 Let $A = (\mathbf{x}_1, \ldots, \mathbf{x}_d) \in [S_k]^d$ [3] and $B = (\mathbf{y}_1, \ldots, \mathbf{y}_d) \in [S_k]^d$ be the inputs of Alice and Bob for $\textsc{Or-Disj}^d_{\{0,1\}^k} \;|_{S_k\times S_k}$. Recall that $S_k = \{0,1\}^k \setminus \{0^k\}$. Bob finds $\overline{B} = (\overline{\mathbf{y}}_1, \ldots, \overline{\mathbf{y}}_d) \in \left[\{0,1\}^k\right]^d$, where $\overline{\mathbf{y}}_i$ is obtained by complementing each bit of $\mathbf{y}_i$.

Step-2 Both Alice and Bob privately determine the first k prime numbers $\pi_1, \ldots, \pi_k$ without any communication.

Step-3 Let

$$\phi : \{0,1\}^k \to \{0,1\}^{\left\lceil \log\left(\sqrt{\frac{n}{d}}\right)\right\rceil}$$

be the function such that $\phi(\mathbf{x})$ is the $\left\lceil \log\left(\sqrt{n/d}\right)\right\rceil$ bit representation of the number $\prod_{i=1}^{k} \pi_i^{x_i}$, where $\mathbf{x} = (x_1, \ldots, x_k) \in \{0,1\}^k$. Alice finds

$$A' = (\mathbf{a}_1, \ldots, \mathbf{a}_d) \in \left[\{0,1\}^{\left\lceil \log\left(\sqrt{\frac{n}{d}}\right)\right\rceil}\right]^d$$

and Bob finds

$$B' = (\mathbf{b_1}, \ldots, \mathbf{b_d}) \in \left[\{0,1\}^{\left\lceil \log\left(\sqrt{\frac{n}{d}}\right)\right\rceil}\right]^d$$

privately without any communication. Here $\mathbf{a_i} = \phi(\mathbf{x_i})$ and $\mathbf{b_i} = \phi(\overline{\mathbf{y}}_i)$ for each $i \in [d]$.

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

Step-4 For each $i \in [d]$, let L_i and L_i' be the lines having equation

$$L_i : y - q_i = \frac{\mathrm{num}(\mathbf{a}_i) - 1}{\mathrm{num}(\mathbf{a}_i)}(x - p_i)$$

and

$$L_i' : x - p_i = \mathrm{num}(\mathbf{b}_i).$$

Here p_i's and q_i's are selected to satisfy PROPERTY. 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 $L_1', \ldots, L_d'$, that is,

$$A'' = \bigcup_{i \in [d]} (L_i \cap G_{(p_i,q_i)}) \text{ and } B'' = \bigcup_{i \in [d]} (L_i' \cap G_{(p_i,q_i)}).$$

Step-5 Then Alice and Bob solve $\text{DISJ}_G \mid_{\mathcal{T}\times\mathcal{T}} (A'', B'')$, and report

$$\bigvee_{i=1}^{d} \text{DISJ}_{\{0,1\}^k}(\mathbf{x}_i, \mathbf{y}_i) = 1$$

if and only if

$$\text{DISJ}_G \mid_{\mathcal{T}\times\mathcal{T}} (A'', B'') = 0.$$

Now we argue for the correctness of the protocol $\mathcal{P}'$. Let $\text{DISJ}_G \mid_{\mathcal{T}\times\mathcal{T}} (A'', B'') = 0$, that is, $A'' \cap B'' \neq \emptyset$. By Claim 3.5 and from the description of $\mathcal{P}'$, there exists $i \in [d]$ such that the lines $L_i : y - q_i = \frac{\mathrm{num}(\mathbf{a}_i)-1}{\mathrm{num}(\mathbf{a}_i)}(x - p_i)$ and $L_i' : x - p_i = \mathrm{num}(\mathbf{b_i})$ intersect at a point in $G_{(p_i,q_i)}$, that is, the lines $y = \frac{\mathrm{num}(\mathbf{a}_i)-1}{\mathrm{num}(\mathbf{a}_i)}x$ and $x = \mathrm{num}(\mathbf{b}_i)$ intersect at a point in $G_{(0,0)}$. Now, we can say that, there exists $i \in [d]$ such that $\mathrm{num}(\mathbf{a}_i)$ divides $\mathrm{num}(\mathbf{b_i})$. This implies $\mathbf{x}_i$ is a subset of $\overline{\mathbf{y}}_i$ (or $\mathbf{x}_i \cap \mathbf{y_i} = \emptyset$) for some $i \in [d]$. Hence, $\bigvee_{i=1}^{d} \text{DISJ}_{\{0,1\}^k}(\mathbf{x}_i, \mathbf{y}_i) = 1$. The converse part, that is, $\bigvee_{i=1}^{d} \text{DISJ}_{\{0,1\}^k}(\mathbf{x}_i, \mathbf{y}_i) = 1$ implies $\text{DISJ}_G \mid_{\mathcal{T}\times\mathcal{T}} (A'', B'') = 0$ can be shown in the similar fashion.

Observe that the communication cost of the protocol $\mathcal{P}'$ for $\textsc{Or-Disj}^d_{\{0,1\}^k} \ |_{S_k \times S_k}$ is same as that of the protocol $\mathcal{P}$ for $\textsc{Disj}_G \ |_{\mathcal{T}\times\mathcal{T}}$, is

$$o\left(d\frac{\log m}{\log\log m}\right) = o\left(d\frac{\log(n/d)}{\log\log(n/d)}\right) = o(dk).$$

The above two equalities follows from the facts that $m = \sqrt{\frac{n}{d}}$ and $k = \Theta\left(\frac{\log(n/d)}{\log\log(n/d)}\right)$. This contradicts Proposition 3.8 which says that

$$R\left(\textsc{Or-Disj}^d_{\{0,1\}^k} \ |_{S_k \times S_k}\right) = \Omega(dk).$$

3.3. Proof of Theorem 3.2. Consider the problem $\textsc{Learn}_G \ |_{\mathcal{T}\times\mathcal{T}}$, 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 $\mathcal{T} \subsetneq 2^G$ with VC-dim$(\mathcal{T}) = 2d$ are same as that constructed in Section 3.1. In $\textsc{Learn}_G \ |_{\mathcal{T}\times\mathcal{T}}$, 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 $R(\textsc{Learn}_G \ |_{\mathcal{T}\times\mathcal{T}}) = \Omega(d\log n)$ as there are $\Omega(m^d) = \Omega\left(\left(\sqrt{n/d}\right)^d\right)$ many candidate sets for the inputs of Alice and Bob. We prove the theorem by a reduction from $\textsc{Learn}_G \ |_{\mathcal{T}\times\mathcal{T}}$ to $\textsc{Int}_G \ |_{\mathcal{T}\times\mathcal{T}}$.

Let, by contradiction, us consider a protocol $\mathcal{P}$ that solves $\textsc{Int}_G \ |_{\mathcal{T}\times\mathcal{T}}$ by using $o(d\log n)$ bits of communication. To solve $\textsc{Learn}_G \ |_{\mathcal{T}\times\mathcal{T}}$, Alice and Bob first run a protocol $\mathcal{P}$ and finds $A \cap B$. Now by Claim 3.5, it is possible for Alice (Bob) to determine B (A) by combining A (B) along with $A \cap B$, without any communication with Bob (Alice). Now, we have a protocol $\mathcal{P}'$ that solves $\textsc{Learn}_G \ |_{\mathcal{T}\times\mathcal{T}}$ with $o(d\log n)$ bits of communication. However, this is impossible as $R(\textsc{Learn}_G \ |_{\mathcal{T}\times\mathcal{T}}) = \Omega(d\log n)$. Hence, we are done with the proof of Theorem 3.2.

4. Conclusion and discussion

In this paper, we studied $\textsc{Disj}_n \ |_{\mathcal{S}\times\mathcal{S}}$ and $\textsc{Int}_n \ |_{\mathcal{S}\times\mathcal{S}}$ 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 1.7) showing that unlike in the case of d-SPARSEDISJ$_n$ and d-SPARSEINT$_n$ functions, there is no separation between randomized and deterministic communication complexity of $\text{DISJ}_n \mid_{\mathcal{S}\times\mathcal{S}}$ and $\text{INT}_n \mid_{\mathcal{S}\times\mathcal{S}}$ functions when VC-dim$(S) \leq d$. Note that we have settled both the one-way and two-way (randomized) communication complexities of $\text{INT}_n \mid_{\mathcal{S}\times\mathcal{S}}$ when VC-dim$(\mathcal{S}) \leq d$ (Theorem 1.7 (1) and (3)). In the context of $\text{DISJ}_n \mid_{\mathcal{S}\times\mathcal{S}}$, we have settled the one-way (randomized) communication complexity. The two-way communication complexity for $\text{DISJ}_n \mid_{\mathcal{S}\times\mathcal{S}}$ is tight up to factor $\log\log(n/d)$ (see Theorem 1.7 (2)). However, we believe that the factor of $\log\log(n/d)$ should not be present in the statement of Theorem 1.7 (2).

CONJECTURE 4.1. *There exists $\mathcal{S} \subseteq 2^{[n]}$ with VC-dim$(\mathcal{S}) \leq d$ and $R(\text{DISJ}_n \mid_{\mathcal{S}\times\mathcal{S}}) = \Omega(d\log(n/d))$.*

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 3.1 and Theorem 3.2. The same G and $\mathcal{T}$ cannot be the hard instance for the proof of Conjecture 4.1 because of the following result.

THEOREM 4.2. *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 $\text{DISJ}_G \mid_{\mathcal{T}\times\mathcal{T}}$ (A, B), with probability at least $2/3$, and uses*

$$O\left(\frac{d\log d\log(n/d)}{\log\log(n/d)} \cdot \log\log\log(n/d)\right)$$

bits of communication.

REMARK 4.3. *If $d = 2^{o\left(\frac{\log\log n}{\log\log\log n}\right)}$ then Theorem 4.2 implies existence of a randomized communication protocol that uses $o(d\log(n/d))$ bits of communication.*

We use the following observation to prove Theorem 4.2.

Observation 4.4. *Let us consider the communication problem* $\mathrm{GCD}_k(a,b)$*, where Alice and Bob get* a *and* b *respectively from* $\{1,\dots,k\}$*, and the objective is for both the players to compute* $\gcd(a,b)$*. There exists a randomized protocol, with success probability at least* $1-\delta$*, for* 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 $\pi_1, \dots, \pi_t$ between 1 and k. From the Prime Number Theorem, we know that $t = \Theta\left(\frac{k}{\log k}\right)$ (see, e.g., Chandrasekharan (1968) and Apostol (1976)). 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 $O\left(\frac{\log k}{\log\log k}\right)$ (see, e.g., Theorem 12 of Robin (1983)). 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, see Brody *et al.* (2014). 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

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

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

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

$$\leq O\left(\frac{\log k}{\log\log k}\right) + |S_a \cap S_b| \log\left(\frac{\log k}{|S_a \cap S_b|}\right)$$
$$\leq O\left(\frac{\log k}{\log\log k} \cdot \log\log\log k\right)$$

In the above inequalities, we use the facts that $|S_a \cap S_b| = O\left(\frac{\log k}{\log\log k}\right)$, $\sum_{p \in S_a \cap S_b} \alpha_{p,a} \leq \log k$ 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 4.2.

Proof (Proof of the Theorem 4.2). 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

$$G_{(0,0)} := \{(x, y) \in \mathbb{Z}^2 \ | \ 0 \leq x, y \leq \sqrt{n}\}.[4]$$

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 Claim 3.5 and Claim 3.6 in mind, we will be done if we can show the existence of a randomized communication protocol for computing the function $\textsc{Disj}_G |_{\mathcal{T}\times\mathcal{T}}$, 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 n/d[5] and setting $\delta = \frac{1}{3d}$ for each of the d instances.

[4] Without loss of generality, we assume that $\sqrt{n}$ is an integer.

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

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 intersect at a point of $G_{(0,0)}$. Moreover, L_A and L_B intersect 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.

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 4.4. □

Acknowledgements

Part of this work was done when Anup Bhattacharya was supported by SERB-National Post Doctoral Fellowship, India, and Arijit Ghosh was supported in part by Ramanujan Fellowship (No. SB/S2/RJN-064/2015), India. Gopinath Mishra is supported in part by the Centre for Discrete Mathematics and its Applications (DIMAP) and by EPSRC award EP/V01305X/1. The authors would like to thank Sudeshna Kolay and Arijit Bishnu for the many discussions in the early stages of this work.

References

Tom M. Apostol (1976). *Introduction to Analytic Number Theory.* Springer, New York, NY, 1st edition.

Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar & D. Sivakumar (2004). An Information Statistics Approach to Data Stream and Communication Complexity. *Journal of Computer and System Sciences* **68**(4), 702–732.

Mark Braverman, Ankit Garg, Denis Pankratov & Omri Weinstein (2013). From information to exact communication. In *Symposium on Theory of Computing Conference, STOC*, 151–160.

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

Komaravolu Chandrasekharan (1968). *Introduction to Analytic Number Theory.* Springer, Berlin, Heidelberg, 1st edition.

BERNARD CHAZELLE (2001). *The Discrepancy Method: Randomness and Complexity*. Cambridge University Press.

ANIRBAN DASGUPTA, RAVI KUMAR & D. SIVAKUMAR (2012). Sparse and Lopsided Set Disjointness via Information Theory. In *International Workshop on Randomization and Computation, RANDOM*, 517–528.

JOHAN HÅSTAD & AVI WIGDERSON (2007). The Randomized Communication Complexity of Set Disjointness. *Theory of Computing* **3**(1), 211–219.

T. S. JAYRAM, RAVI KUMAR & D. SIVAKUMAR (2003). Two Applications of Information Complexity. In *ACM Symposium on Theory of Computing, STOC*, 673–682.

BALA KALYANASUNDARAM & GEORG SCHNITGER (1992). The Probabilistic Communication Complexity of Set Intersection. *SIAM Journal on Discrete Mathematics* **5**(4), 545–557.

EYAL KUSHILEVITZ & NOAM NISAN (1996). *Communication Complexity*. Cambridge University Press.

JIRI MATOUSEK (2009). *Geometric Discrepancy: An Illustrated Guide*, volume 18. Springer Science & Business Media.

JIRI MATOUSEK (2013). *Lectures on Discrete Geometry*, volume 212. Springer Science & Business Media.

PETER BRO MILTERSEN, NOAM NISAN, SHMUEL SAFRA & AVI WIGDERSON (1998). On Data Structures and Asymmetric Communication Complexity. *Journal of Computer and System Sciences* **57**(1), 37–49.

ILAN NEWMAN (1991). Private vs. Common Random Bits in Communication Complexity. *Information Processing Letters* **39**(2), 67–71.

JÁNOS PACH & PANKAJ K. AGARWAL (2011). *Combinatorial Geometry*, volume 37. John Wiley & Sons.

ANUP RAO & AMIR YEHUDAYOFF (2020). *Communication Complexity: and Applications*. Cambridge University Press.

ALEXANDER A. RAZBOROV (1992). On the Distributional Complexity of Disjointness. *Theoretical Computer Science* **106**(2), 385–390.

Guy Robin (1983). Estimation de la fonction de Tchebychef θ sur le k-ième nombre premier et grandes valeurs de la fonction $\omega(n)$ nombre de diviseurs premiers de n. *Acta Arithmetica* **42**(4), 367–389.

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

Mert Saglam & Gábor Tardos (2013). On the Communication Complexity of Sparse Set Disjointness and Exists-Equal Problems. In *IEEE Symposium on Foundations of Computer Science, FOCS*, 678–687.

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

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

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

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

Manuscript received 9 April 2021

Anup Bhattacharya
NISER
Bhubaneswar, 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
University of Warwick
Coventry, UK
gopianjan117@gmail.com

Manaswi Paraashar
Aarhus University
Aarhus, Denmark
manaswi.isi@gmail.com