The Communication Complexity of Distributed ε-Approximations

Zengfeng Huang[*] Ke Yi [†]

MADALGO, Aarhus University HKUST

Abstract

Data summarization is an effective approach to dealing with the "big data" problem. While data summarization problems traditionally have been studied is the streaming model, the focus is starting to shift to distributed models, as distributed/parallel computation seems to be the only viable way to handle today's massive data sets. In this paper, we study ε-approximations, a classical data summary that, intuitively speaking, preserves approximately the density of the underlying data set over a certain range space. We consider the problem of computing ε-approximations for a data set which is held jointly by k players, and give general communication upper and lower bounds that hold for any range space whose discrepancy is known.

[*] Zengfeng Huang is supported by the Danish National Research Foundation grant DNRF84 through Center for Massive Data Algorithmics (MADALGO).

[†] Ke Yi is supported by HKRGC under grants GRF-621413 and GRF-16211614.

1 Introduction

Data summarization, which makes "big data" small while preserving important properties of the data, is an effective approach in many data analytics tasks. While data summarization problems traditionally have been studied is the streaming model, the focus is starting to shift to distributed models, as distributed/parallel computation seems to be the only viable way to handle today's massive data sets. A standard approach for constructing a summary over a large data set is to break it into k pieces (often tens of thousands), give them to many workers, who compute a summary for each piece (possibly using a streaming algorithm), and then merge them together. This mergeability is trivially enjoyed by all the linear sketches, as well as many non-linear summaries [1]. This standard approach has a communication cost of $k\times$summary size. Thus, an interesting question is if this is the best one can do. It has been shown that for heavy hitters and the quantiles, with randomization, the communication cost can be reduced to $\sqrt{k}\times$summary size [11], while no further improvement is possible for coresets [21], the distinct count [28] and the F_2 problem [27], up to polylog factors.

In this paper, we study another classical data summary, *ε-approximations*, which preserves approximately the density of the underlying data set over a certain range space (a formal definition will be given later). In the centralized setting, ε-approximations have been extensively studied [19, 9], with numerous applications in data analytics, machine learning, and computational geometry. In this work, we are interested in characterizing the communication complexity of computing ε-approximations from large distributed data sets. In a nutshell, we show that the standard approach above is optimal for deterministic algorithms, while the dependency on k can be reduced with randomization, and the exact dependency is related to the discrepancy of the underlying range space.

1.1 Communication models

We study the multi-party communication complexity in the *number-in-hand* model, in which each player gets part of the input and cannot see other players' data. To solve a problem, the players can exchange messages with each other. Based on the way the messages are being exchanged, the following three models have been studied in the literature, with increasing power of communication.

Simultaneous message passing model. In the *simultaneous message passing model* (SMP), there is a coordinator or referee, and each player can only send one message to the coordinator. After receiving all the messages, the coordinator should output the answer. In this model, it is usually assumed that all players know the common parameters like ε, k (the number of players), and n (input size).

Message passing model. The *message passing model* is a fundamental model has been extensively studied in distributed computing and other areas [12, 13]. In this model, every player can send messages to any other players according to some protocol.

Blackboard model. In the *blackboard* model, there is a shared blackboard, and each player will write their messages on it, which can be seen by everyone else. This model corresponds to the situation where there is a broadcast channel in the communication network. This model has the strongest communication power, thus lower bounds proved for the this model also hold for the other two models.

1.2 Range space and ε-approximation

Let X be a set of size n and $\mathcal{S} \subseteq 2^X$ be a family of subsets of X. In the study of ε-approximations, we are interested in set systems $(X, \mathcal{S})$ where X is a set of points in $\mathbb{R}^d$ and $\mathcal{S}$ is induced by some *range space* $\mathcal{R}$ (such as all intervals in $\mathbb{R}^1$ or all halfplanes in $\mathbb{R}^2$), i.e., $\mathcal{S} = \{r \cap X \mid r \in \mathcal{R}\}$. Let $\varepsilon \in [0, 1]$ be a real number. We say that a subset $Y \subseteq X$ is an *ε-approximation* of X w.r.t. the range space $\mathcal{R}$ if we have, for all $r \in \mathcal{R}$,

$$\left| \frac{|Y \cap r|}{|Y|} - \frac{|X \cap r|}{|X|} \right| \leq \varepsilon. \tag{1}$$

This means that the fraction of points of Y lying in r should approximate the fraction of points of X in r with accuracy no worse than ε. Therefore, an ε-approximation is considered as a "density-preserving" sample of X (w.r.t. $\mathcal{R}$), which is a desirable property in many applications.

The study of ε-approximations has spanned mathematics, learning theory, and computational geometry. Classical results [26, 16] show that, if the range space has bounded VC-dimension, a random subset of size $\Theta(\frac{1}{\varepsilon^2})$ is an ε-approximation with constant probability. An interesting fact about ε-approximations is that their size depends only on ε. Nevertheless, keep in mind that ε is an additive error to a fraction, so it should be set quite small; actually, one may consider $\varepsilon = 1/\sqrt{n}$ as the most interesting case, as it makes simple random sampling ineffective and calls for more careful constructions.

Define $\mathsf{approx}(\varepsilon, X, \mathcal{R})$ to be the optimal size of an ε-approximation of X with respect to $\mathcal{R}$, and let $\mathsf{approx}(\varepsilon, \mathcal{R}) = \max_X \mathsf{approx}(\varepsilon, X, \mathcal{R})$. We will omit $\mathcal{R}$ when it is clear from the context. For specific range spaces, the size of ε-approximations can be much smaller than what is achieved by random sampling. For instance, for intervals in $\mathbb{R}^1$, we have (trivially) $\mathsf{approx}(\varepsilon) = \Theta(\frac{1}{\varepsilon})$; for (axis-parallel) boxes in $\mathbb{R}^d$, we have $\mathsf{approx}(\varepsilon) = O(\frac{1}{\varepsilon} \log^{d+1/2} \frac{1}{\varepsilon})$ [19, 15]; for halfspaces in $\mathbb{R}^d$, $\mathsf{approx}(\varepsilon) = \Theta(1/\varepsilon^{\frac{2d}{d+1}})$ [18]. There is a close relationship between $\mathsf{approx}(\varepsilon)$ and the combinatorial discrepancy of the underlying range space. We include a brief introduction to discrepancy below. For a complete treatment on ε-approximations and discrepancy theory, please refer to the books [19] and [9].

1.3 Discrepancy

In this section, we give some preliminaries on discrepancy theory, which will be used in our algorithm, as well as the lower bounds. Let $X = \{x_1, \cdots, x_n\} \subset \mathbb{R}^d$ be a set of n points in $\mathbb{R}^d$ and $\mathcal{R}$ a range space. A coloring is any mapping $\chi : X \to \{-1, +1\}$.

The *(combinatorial) discrepancy* of the set system $(X, \mathcal{R})$ is defined as

$$\mathsf{disc}(X, \mathcal{R}) = \min_{\chi} \max_{r \in \mathcal{R}} |\chi(r)|, \tag{2}$$

where $\chi(r) = \sum_{x \in X \cap r} \chi(x)$. Intuitively, it finds the best coloring such that the two colors in every range of $\mathcal{R}$ are as balanced as possible. For example, if $\mathcal{R}$ is the family of all intervals in $\mathbb{R}^1$, the best coloring is to order the points of X and then color them alternatively, which gives $\mathsf{disc}(X, \mathcal{R}) = 1$. The discrepancy of the range space $\mathcal{R}$ considers the worst case over all point sets X of a given size n, i.e.,

$$\mathsf{disc}(n, \mathcal{R}) = \max_{|X|=n} \mathsf{disc}(X, \mathcal{R}). \tag{3}$$

We will omit $\mathcal{R}$ when it is clear from the context. The discrepancy for many range spaces have been extensively studied. Notable results include: for intervals in $\mathbb{R}^1$, $\mathsf{disc}(n) = 1$; for boxes in $\mathbb{R}^d$, $\mathsf{disc}(n) = O(\log^{d+1/2} n)$ [15]; for halfspaces in $\mathbb{R}^d$, $\mathsf{disc}(n) = \Theta(n^{\frac{d-1}{2d}})$ [18]. The bounds on the size of ε-approximations mentioned above are actually all derived from these discrepancy bounds through a standard procedure [19]. In general, if $\mathsf{disc}(n) = O(\log^\tau n)$ for some constant τ, then $\mathsf{approx}(\varepsilon) = O(\frac{1}{\varepsilon} \log^\tau \frac{1}{\varepsilon})$; if $\mathsf{disc}(n) = O(n^\tau)$, then $\mathsf{approx}(\varepsilon) = O(1/\varepsilon^{1/(1-\tau)})$.

The discrepancy as defined in (2) considers the worst range r of $\mathcal{R}$. The notion of an "average discrepancy" has also been studied, which is often helpful in lower bound proofs. Let m be the number of distinct subsets that can be defined by $\mathcal{R}$ on X. When X is given, we abuse the notation slightly by also using $\mathcal{R} = \{r_1, \cdots, r_m\}$ to denote these subsets of X induced by the ranges of $\mathcal{R}$. We order X and $\mathcal{R}$ arbitrarily, and let A be the incidence matrix of $(X, \mathcal{R})$, which is the $m \times n$ matrix with columns corresponding to the points of X and rows corresponding to $\mathcal{R}$, and has entry $a_{ij} = 1$ iff x_j is in r_i. Interpreting the coloring χ as a vector, the *l_p-discrepancy* of $(X, \mathcal{R})$ is defined as

$$\mathsf{disc}_p(X, \mathcal{R}) = \min_{\chi \in \{-1,+1\}^n} \frac{||A\chi||_p}{m^{1/p}}.$$

Thus $\mathsf{disc}(X, \mathcal{R}) = \mathsf{disc}_\infty(X, \mathcal{R})$. It should be clear that $\mathsf{disc}_1(X, \mathcal{R}) \le \mathsf{disc}_2(X, \mathcal{R}) \le \mathsf{disc}_\infty(X, \mathcal{R})$.

An *α-partial coloring* is any mapping $\chi : X \to \{-1, 0, +1\}$ satisfying $||\chi||_1 \ge \alpha|X|$. We can similarly define the α-partial l_p-discrepancy as

$$\mathsf{disc}_p^\alpha(X, \mathcal{R}) = \min_{\chi \in \{-1,0,+1\}^n, ||\chi||_1 \ge \alpha n} \frac{||A\chi||_p}{m^{1/p}}.$$

The *α-partial l_p-discrepancy* for a range space $\mathcal{R}$ is defined similarly as in (3):

$$\mathsf{disc}_p^\alpha(n, \mathcal{R}) = \max_{|X|=n} \mathsf{disc}_p^\alpha(X, \mathcal{R}).$$

To construct a good full coloring, a general method is to first get a good partial coloring, then recursively apply the method to the points that have not been colored. So the gap between the full coloring discrepancy and partial discrepancy is usually not significant. The following lemma can be easily proved using this idea (proof in Appendix A).

Lemma 1.1. *For any range space $\mathcal{R}$, any constant α and $p \ge 1$, $\mathsf{disc}_p(n) = O(\log n \cdot \mathsf{disc}_p^\alpha(n))$. If $\mathsf{disc}_p^\alpha(n) = \mathrm{poly}(n)$, then $\mathsf{disc}_p(n) = O(\mathsf{disc}_p^\alpha(n))$.*

Finally, we need another version of coloring where we relax $\chi(x)$ to be real numbers instead of $\{-1, +1\}$. We call a point x *heavy* if $|\chi(x)| \ge 1$. A coloring χ is *α-heavy*, if it is a mapping $\chi : X \to \mathbb{R}$, such that there are at least $\alpha|X|$ points in X that are heavy. The *α-generalized l_p-discrepancy* is

$$\mathsf{rdisc}_p^\alpha(X, \mathcal{R}) = \min_{\alpha\text{-heavy } \chi} \frac{||A\chi||_p}{m^{1/p}}.$$

We omit $\mathcal{R}$ when the context is clear; we also omit α when $\alpha = 1$ and omit p when $p = \infty$.

1.4 Our results

In our setting, the point set X is partitioned arbitrarily into k pieces and each player possesses one, and the goal is to compute an ε-approximation (for a certain range space) using minimum communication. Based on known results, there are two straightforward ways to solve this problem. The first one is to draw a random sample of size $O(1/\varepsilon^2)$ over the whole data set, which can be easily implemented with communication cost $O(1/\varepsilon^2)$.[1] Another way is the standard "merging" approach mentioned at the beginning of the paper, which has communication cost $O(k \cdot \mathsf{approx}(\varepsilon))$. It is deterministic if the players' time cost to compute the local ε-approximations is not a concern. The detailed bounds for various range spaces are listed in Table 1, and they remain the best deterministic bounds for these cases.

Our first result is a randomized algorithm in the SMP model to compute an ε-approximation for any range space, which has communication cost better than both methods above. The algorithm works for any range space $\mathcal{R}$, with its communication cost depending on $\mathsf{disc}(n, \mathcal{R})$. The general relationship is described in Theorem 2.1, while we have listed the explicit bounds for some common

[1] We assume $k < 1/\varepsilon^2$ in this paper. When $k > 1/\varepsilon^2$, the bound can be either $O(k)$ or $O(1/\varepsilon^2)$ depending on some subtleties in the model formulation, e.g., whether the players need to be notified when the protocol starts/ends, but in either case, random sampling is already the optimal solution for our problem.

	$\mathsf{approx}(\varepsilon)$	Deterministic		Randomized	
		O	Ω	O	Ω
Intervals in $\mathbb{R}^1$	$1/\varepsilon$	k/ε	k/ε	$\frac{\sqrt{k}}{\varepsilon}\sqrt{\log\frac{1}{\varepsilon}}$	$\sqrt{k}/\varepsilon$
Boxes in $\mathbb{R}^d$	$\frac{1}{\varepsilon}\log^{d+\frac{1}{2}}\frac{1}{\varepsilon}$	$\frac{k}{\varepsilon}\log^{d+\frac{1}{2}}\frac{1}{\varepsilon}$	$\frac{k}{\varepsilon}\log\frac{1}{\varepsilon}$	$\frac{\sqrt{k}}{\varepsilon}\log^{d+1}\frac{1}{\varepsilon}$	$\frac{\sqrt{k}}{\varepsilon}\log\frac{1}{\varepsilon}$ for $d = 2, 3, 4$; $\frac{\sqrt{k}}{\varepsilon}\log^{\frac{d-3}{2}}\frac{1}{\varepsilon}$ for all d
Halfspaces in $\mathbb{R}^d$	$1/\varepsilon^{\frac{2d}{d+1}}$	$k/\varepsilon^{\frac{2d}{d+1}}$	$k/\varepsilon^{\frac{2d}{d+1}}$	$k^{\frac{1}{d+1}}/\varepsilon^{\frac{2d}{d+1}}\cdot\log^{\frac{d}{d+1}}\frac{1}{\varepsilon}$	$k^{\frac{1}{d+1}}/\varepsilon^{\frac{2d}{d+1}}$

Table 1: The k-party communication upper and lower bounds for computing ε-approximations. The upper bounds are in terms of the number of points communicated and hold in the simultaneous message passing model; the lower bounds are in terms of bits and hold in the blackboard model.

range spaces in Table 1. For example, for halfspaces in $\mathbb{R}^d$, the communication cost of our algorithm is $\tilde{O}(k^{\frac{1}{d+1}}/\varepsilon^{\frac{2d}{d+1}}) = \tilde{O}(k^{\frac{1}{d+1}} \cdot \mathsf{approx}(\varepsilon))$. This is obviously better than the deterministic bound $O(k \cdot \mathsf{approx}(\varepsilon))$; it is also better than the random sampling bound $O(1/\varepsilon^2)$ when $k \leq 1/\varepsilon^2$, the parameter range we are interested in.

The bulk of the paper studies lower bound methods, which yield almost tight lower bounds for many interesting range spaces, up to polylog factors. The lower bounds hold in the blackboard model, and actually hold for computing any data structure that allows one to estimate $\frac{|X \cap r|}{|X|}$ for any $r \in \mathbb{R}$ as in (1) within ε accuracy, which may not be a subset of points of X as required by the strict definition of an ε-approximation. The lower bounds use information-theoretical arguments, and the bounds are in terms of bits (note that the upper bounds are in terms of the number of points transmitted). But we note that all the lower bound constructions use points drawn from a d-dimensional grid $[u]^d$ for $u = (k/\varepsilon)^{O(1)}$. In this case, the gap between the upper and lower bounds is thus $\mathrm{polylog}(k/\varepsilon)$ since it takes $d \log u$ bits to represent a point.

We provide both deterministic and randomized lower bounds. We first relate deterministic lower bounds to the partial l_1-discrepancy $\mathsf{disc}_1^{\alpha}(n)$ of the underlying range space (Theorem 3.1). However, since the l_1-discrepancy is still not well understood (there is only the trivial lower bound $\mathsf{disc}_1^{\alpha}(n) = \Omega(1)$), this method yields a rather weak lower bound of $\Omega(k/\varepsilon)$. We then link the deterministic lower bound to the partial generalized l_∞-discrepancy $\mathsf{rdisc}(n)$ (Theorem 4.2), via a new deterministic direct-sum result. This leads to tight deterministic lower bounds, in particular for halfspaces.

For randomized lower bounds, we also provide two connections to discrepancy theory, one to the (partial) l_2-discrepancy (Theorem 3.8) and the other to the (partial) generalized l_∞-discrepancy (Theorem 4.5). Comparing these two relationships with the upper bound in Theorem 2.1, one can notice that the statements are the essentially the same (modulo log factors), except that $\mathsf{disc}(n)$ is replaced by $\mathsf{disc}_2(n)$ and $\mathsf{rdisc}(n)$, respectively. As lower bounds on the l_2-discrepancy and generalized discrepancy have been well studied, this allows us to derive almost tight lower bounds for the range spaces in Table 1.

1.5 Related work

Discrepancy theory, including ε-approximations, is a well-studied topic [19, 9]. There are efficient algorithms to construct ε-approximations for range spaces with bounded VC-dimension in the centralized setting [9]. For most interesting range spaces, optimal ε-approximations now can be computed in polynomial time due to recent breakthrough results on constructive discrepancy minimization [5, 17]. Recently, it has also been studied in the streaming model [24, 4, 1]. However, the communication complexity of computing ε-approximations in a distributed environment is largely an unexplored area. For the simplest range space, intervals in $\mathbb{R}^1$, the ε-approximation problem is equivalent to the

ε*-approximate quantiles* problems. For this problem, we previously gave an algorithm in the simultaneous message passing model with communication cost $O(\frac{\sqrt{k}}{\varepsilon}\log\frac{1}{\varepsilon})$ words [11]; while Woodruff and Zhang [27] proved a randomized lower bound of $\Omega(\frac{\sqrt{k}}{\varepsilon})$ bits. The results presented here match these results for the approximate quantiles problem, but are much more general, as they hold for any range space for which the discrepancy is known.

2 The Upper Bound

In this section we give a general randomized algorithm for any range space $\mathcal{R}$ based on the upper bound of $\mathsf{disc}(n,\mathcal{R})$. We first rewrite the definition of ε-approximations (1) as

$$\left| |Y\cap r|\cdot\frac{|X|}{|Y|}-|X\cap r|\right|\le\varepsilon|X|,$$

which implies that if we give a weight $\frac{|X|}{|Y|}$ to each element in Y, the size of $X\cap r$ can be estimated within an additive error $\varepsilon|X|$ for all $r\in\mathcal{R}$. For a range r, when call $\left||Y\cap r|\cdot\frac{|X|}{|Y|}-|X\cap r|\right|$ the error of Y on r. Then Y is an ε-approximation of X if its error is at most $\varepsilon|X|$ for all $r\in\mathcal{R}$. We allow duplicated points in X and Y; they are treated as different points but sharing the same coordinates.

The Algorithm. Our algorithm works in the simultaneous message passing model. Let X be partitioned into k subsets with player i holding I_i. Let $n_i=|I_i|$ and $n=\sum_i n_i$. We assume that $\mathsf{disc}(n)\le O(\sqrt{n})$, which is a reasonable assumption since this holds for any range space with bounded VC-dimension [19].

Every player i will run the following algorithm to compute a subset of I_i, and sends it to the coordinator. Without loss of generality we assume $|I_i|$ is a power of 2. Each player i reduces its point set I_i iteratively, every time by half. We first color I_i using a coloring χ such that $|\chi(r)|=O(\mathsf{disc}(|I_i|))$ for all r and $|\chi^{-1}(-1)|=|\chi^{-1}(+1)|$.[2] Then randomly pick one of the two classes $\chi^{-1}(-1)$ and $\chi^{-1}(+1)$. Let I_i^1 be the picked subset, and we recursively apply the same procedure on I_i^1. We repeat the procedure for λ times, and let I_i^λ be the final subset. The value of λ will be determined through our analysis.

In the end, the coordinator receives k subsets $I_1^\lambda,\cdots,I_k^\lambda$. We will show that their union forms an ε-approximation of X with at least constant probability. Note that this may not give an ε-approximation of the optimal size, but this can be easily fixed by asking the coordinator to run a centralized algorithm to compute an optimal-size ε-approximation on the union of the k subsets. This would give a (2ε)-approximation of X.

Analysis. The analysis of the algorithm is given in Appendix B, which is based on the following intuition. For $1\le j\le\lambda$, let $\Delta_i^j=2^{j-1}(2|I_i^j\cap r|-|I_i^{j-1}\cap r|)$. Then the final error of using the union of the I_i^λ's to approximate X on r is $\sum_{i=1}^k\sum_{j=1}^\lambda\Delta_i^j$. We observe that the Δ_i^j's are random variables with mean 0 and absolute value bounded by the discrepancy of the coloring. Then, we can apply a tail bound on the probability that their sum does not deviate from 0 too much. Some technicalities have to be taken care of because the Δ_i^j's are not independent, though.

Theorem 2.1. *For any range space $\mathcal{R}$ with bounded VC-dimension, if* $\mathsf{disc}(n)^2$ *is concave and t is any value that satisfies*

$$\frac{t}{\mathsf{disc}(t)}=\Omega\left(\frac{1}{\varepsilon\sqrt{k}}\cdot\sqrt{\log\frac{1}{\varepsilon\delta}}\right),$$

[2] We assume there is an $r\in\mathcal{R}$ that contains the entire X, which is true for all natural geometric range spaces. Then the restriction $|\chi^{-1}(-1)|=|\chi^{-1}(+1)|$ will increase the discrepancy by at most a constant factor.

then there is an algorithm in the simultaneous message passing model that computes an ε-approximation for any input X with respect to $\mathcal{R}$ with probability $1-\delta$, and the communication cost is $O(tk)$.

To apply this theorem, given a range space, we should set t to the smallest value satisfying the above requirement. In case $\mathsf{disc}(t)$ is not yet known, it can be replaced by any upper bound. The reader can easily verify the results in Table 1 by plugging in the known discrepancy upper bounds.

Remarks. Our algorithm actually adopts the same framework as the standard algorithm to compute an ε-approximation [19]. The only difference is that, while we retain $\chi^{-1}(-1)$ or $\chi^{-1}(+1)$ randomly, the standard algorithm just picks one arbitrarily. In the RAM model, making random choices does not help; what we show here through our analysis is that these random choices significantly reduce the communication complexity, and they are necessary as suggested by our deterministic lower bounds. Similar ideas have also been used in designing randomized streaming algorithms for computing ε-approximations [24, 1], but no deterministic lower bounds are known in the streaming model.

3 Lower Bounds via l_p-Discrepancy

In this section, we give deterministic and randomized lower bounds techniques for computing ε-approximations in the blackboard model based on l_p-discrepancy. The deterministic lower bound uses l_1-discrepancy, while the randomized lower bounds use l_2-discrepancy.

3.1 Our techniques

Our lower bounds follow the following common framework. For some parameter t, set Y to be a set of t points with the largest discrepancy, i.e., the discrepancy of $(Y,\mathcal{R})$ matches that of $\mathcal{R}$. The particular form of discrepancy will vary in different lower bounds. It is easy to see that Y should not have duplicated points. Let $\mathcal{R}_{|Y}=\{r_1,\ldots,r_m\}$ be the distinct ranges in $\mathcal{R}$ defined on Y. The input of each player i is a subset $x_i\subseteq Y$, and we write x_i in vector form $x_i\in\{0,1\}^t$ (assuming an arbitrary ordering of the points in Y). Let $x=x_1+\cdots+x_k$ denote the whole input set, for which we need to compute an ε-approximation. Note that x could contain duplicated points, and duplicity is considered. Let $n=|x|$, which is at most tk. Our lower bounds use information-theoretical arguments and hold for computing any data structure that allows one to estimate $|x\cap r|$ for any $r\in\mathcal{R}$ with additive error εn.

Clearly there are totally 2^{tk} possible inputs. Let Π be any deterministic protocol computing an ε-approximation in the blackboard model. It is well-known [14] that Π partitions the set of all possible inputs into a set of combinatorial rectangles $P=\{\rho_1,\rho_2,\cdots,\rho_h\}$. Each rectangle ρ_j is a Cartesian product $\rho_j=B_1\times B_2\times\cdots\times B_k$, where $B_i\subseteq\{0,1\}^t$ is a subset of all possible inputs for player i. For all the inputs in the same rectangle, the transcripts of the protocol are exactly the same, in other words, the protocol cannot distinguish these inputs. In particular, the outputs over a rectangle are the same, thus any correct Π should produce a rectangle partitioning such that all inputs in any rectangle share a common correct output. The communication cost of Π is at least $\log|P|$ bits. A randomized protocol is just a distribution of a set of deterministic protocol.

For deterministic complexity, it is enough to prove that, under certain conditions, any correct deterministic protocol computing an ε-approximation must produce a partitioning P with at least $2^{\Omega(tk)}$ rectangles. Let $\rho=B_1\times B_2\times\cdots\times B_k$ be any combinatorial rectangle in the partition induced by a correct deterministic protocol, the size of which is $|\rho|=|B_1|\cdot|B_2|\cdot\cdots\cdot|B_k|$. We will show $|\rho|\le 2^{tk/2}$, thus the communication cost of the protocol is at least $\log(2^{tk}/|\rho|)=\Omega(tk)$. As all the inputs in a rectangle share the same output, we define the error in a rectangle ρ as

$$\max_{x,x'\in\rho}\max_{r\in\mathcal{R}}\big||x\cap r|-|x'\cap r|\big|.$$

It is easy to see that any correct protocol cannot have a rectangle with error larger than $2\varepsilon n \leq 2\varepsilon tk$, since otherwise no matter what the output is, it cannot be a valid ε-approximation for both $(x, \mathcal{R})$ and $(x', \mathcal{R})$. We will prove that the error in a rectangle with size greater than $2^{tk/2}$ will be too large. To do so, we begin with the observation that a rectangle of size at least $2^{tk/2}$ must have $\Omega(k)$ long sides, where a long side i means $|B_i| \geq 2^t/4$. Then we show that, for a long side B_i, there must be $x_i, x_i' \in B_i$ that are far away in terms of the Hamming distance. The Hamming distance is then related to the α-partial discrepancy of $\mathcal{R}$, which shows that x_i, x_i' will cause a large error for some r. However, the $\Omega(k)$ players corresponding to the long sides may not share the same r, so we cannot add the errors up. This is where the ℓ_1-discrepancy comes to the rescue, which corresponds to the "average" error over all ranges, which can be added up over the $\Omega(k)$ players. We formalize the above intuition and give the full proof of the following theorem in Appendix C.

Theorem 3.1. *For any range space $\mathcal{R}$, any deterministic protocol Π that solves the $\frac{1}{3t} \cdot \mathsf{disc}_1^{1/64}(t)$-approximation problem for $\mathcal{R}$ must communicate at least $\Omega(tk)$ bits in the blackboard model.*

Since for any range space, $\mathsf{disc}_1^{1/64}(t) = \Omega(1)$ is a trivial lower bound, the above theorem yields an $\Omega(k/\varepsilon)$ lower bound for all range spaces by setting $t = \Omega(1/\varepsilon)$. However, as l_1-discrepancy lower bounds are still not well understood, Theorem 3.1 currently does not yield a higher lower bound for specific range spaces.

To prove randomized lower bounds, by Yao's minimax principle, we only need to prove the distributional complexity for some hard input distribution. We follow the common framework as in the deterministic case, i.e., we pick an Y to maximize the discrepancy (the l_2-discrepancy now) of $(Y, \mathcal{R})$ with $|Y| = t$ for some t. The input of each player i is a random subset $x_i \in Y$, and the goal is to compute an ε-approximation for $(x, \mathcal{R})$, where $x = x_1 + \cdots + x_k$. The (still deterministic) algorithm is required to solve the ε-approximation problem with at least constant probability (with respect to the random input).

Our hard distribution is simply the uniform distribution. More precisely, thinking of x_i as a binary vector in $\{0, 1\}^t$, each player independently sets each entry of x_i to 1 or 0 with equal probability. For deterministic complexity, we only need to find two inputs in a large rectangle that are far away, such that they cannot share the same output. However, to lower bound the distributional communication complexity, we need to show that a constant fraction of all inputs in a large rectangle cannot be answered correctly, which is more difficult. Our approach is, instead of analyzing the maximum error, we will analyze the variance of the number of points in each range. We use l_2-discrepancy to lower bound the variance of the range counts, and argue that in a large rectangle, there must exist one range such that the number of points in this range cannot concentrate around its expectation, then at least a constant fraction of the inputs cannot be answered correctly. Similar variance arguments have been used in [27] to prove the information cost of a simpler problem. We next give the details of the proof for randomized lower bounds.

3.2 Randomized lower bound via l_2-discrepancy

As before we set $n = |x|$, $m = |\mathcal{R}_{|Y}|$ and A the incidence matrix of the range space $(Y, \mathcal{R}_{|Y})$. Let us fix a rectangle $\rho = B_1 \times B_2 \times \cdots \times B_k$. Note that conditioned on $x \in \rho$, the distribution of x_i is uniform in B_i, which follows from the fact that the distribution of x is uniform and ρ is a combinatorial rectangle. Let $E_i = \frac{1}{|B_i|} \sum_{x_i \in B_i} Ax_i$ be an m-dimensional vector, which can be viewed as the average of Ax_i over all x_i in the set B_i. Let d_i also be an m-dimensional vector such that $d_{i,j} = \sum_{x_i \in B_i} ((Ax_i)_j - E_{i,j})^2$. Note that $\sum_{x_i \in B_i} ||Ax_i - E_i||_2^2 = \sum_{j=1}^m d_{i,j}$. We have the following lemma.

Lemma 3.2. *For any rectangle $\rho = B_1 \times B_2 \times \cdots \times B_k$ of size at least $2^{tk/2}$, we have*

$$\max_{j\in[m]} \sum_{i=1}^{k} \frac{1}{|B_i|} \cdot d_{i,j} \geq \frac{k}{18} \cdot (\mathsf{disc}_2^{1/64}(Y, \mathcal{R}))^2.$$

We first prove the following technical lemma.

Lemma 3.3. *If $|B_i| \geq 2^{t/4}$, then*

$$\sum_{x\in B_i} ||Ax - E_i||_2^2 \geq \frac{|B_i|m}{6} \cdot (\mathsf{disc}_2^{1/64}(Y, \mathcal{R}))^2.$$

Proof. For any two vectors $x, y \in B_i$, if their Hamming distance $HD(x, y) \geq t/64$, we have

$$||Ax - Ay||_2 = ||A(x - y)||_2 \geq \sqrt{m} \cdot \mathsf{disc}_2^{1/64}(Y, \mathcal{R}),$$

as $x - y$ is a $1/64$-partial coloring. By triangle inequality,

$$||Ax - E_i||_2 + ||Ay - E_i||_2 \geq ||Ax - Ay||_2 \geq \sqrt{m} \cdot \mathsf{disc}_2^{1/64}(Y, \mathcal{R}),$$

which implies $||Ax - E_i||_2^2 + ||Ay - E_i||_2^2 \geq \frac{m}{2} \cdot (\mathsf{disc}_2^{1/64}(Y, \mathcal{R}))^2$.

From the Lemma C.2, we know that, if the size of B_i is at least $2^{t/4}$, then for any vector $x \in B_i$, almost all other vectors in B_i have Hamming distance at least $t/64$ from x. Thus we can easily pick $|B_i|/3$ pairs of (x, y) with $HD(x, y) \geq t/64$, so

$$\sum_{x\in B_i} ||Ax - E||_2^2 \geq \frac{|B_i|}{3} \cdot \frac{m}{2} \cdot (\mathsf{disc}_2^{1/64}(Y, \mathcal{R}))^2 = \frac{|B_i|m}{6} \cdot (\mathsf{disc}_2^{1/64}(Y, \mathcal{R}))^2.$$

□

Proof of Lemma 3.2. By Lemma 3.3, if $|B_i| \geq 2^{t/4}$ we have $\frac{1}{|B_i|} \sum_{j=1}^{m} d_{i,j} \geq \frac{m}{6} \cdot (\mathsf{disc}_2^{1/64}(Y, \mathcal{R}))^2$. By Lemma C.1, there are at least $k/3$ such i's, which implies

$$\sum_{j=1}^{m} \sum_{i=1}^{k} \frac{1}{|B_i|} \cdot d_{i,j} \geq \frac{mk}{18} \cdot (\mathsf{disc}_2^{1/64}(Y, \mathcal{R}))^2.$$

Then the lemma follows using an averaging argument. □

Suppose j achieves the maximum in Lemma 3.2. For any i, let Y_{ij} be a random variable, which is defined as $Y_{ij} = (Ax)_j$ where $x \in_R B_i$, i.e., the number of points in r_j when the underlying set is chosen uniformly from B_i. Put $Y_j = \sum_{i=1}^{k} Y_{ij}$. Since our input distribution is uniform, it easily follows from the combinatorial rectangle property that the random variables $Y_{1j}, Y_{2j}, \cdots, Y_{kj}$ are independent. Thus Lemma 3.2 actually gives a lower bound on $\mathsf{Var}[Y_j]$, the variance of Y_j, since $\mathsf{Var}[Y_{i,j}] = \frac{1}{|B_i|} \cdot d_{i,j}$.

Corollary 3.4. *For any rectangle $\rho = B_1 \times B_2 \times \cdots \times B_k$ of size at least $2^{tk/2}$, we have*

$$\max_{j\in[m]} \mathrm{Var}[Y_j] \geq \frac{k}{18} \cdot (\mathsf{disc}_2^{1/64}(Y, \mathcal{R}))^2.$$

Then we can use an anti-concentration inequality (Lemma 3.6) to prove that Y_j cannot concentrate around its expectation. Let $w = \max_{r\in\mathcal{R}} |Y \cap r|$. We have the following anti-concentration lemma.

Lemma 3.5. *If the size of the rectangle ρ is at least $2^{tk/2}$ and $\frac{k}{18} \cdot (\mathsf{disc}_2^{1/64}(Y, \mathcal{R}))^2 \geq 40000w^2$, then*

$$\mathsf{Pr}[Y_j \geq E[Y_j] + \sqrt{k} \cdot \mathsf{disc}_2^{1/64}(Y, \mathcal{R})] \geq c, \text{ and}$$

$$\mathsf{Pr}[Y_j \leq E[Y_j] - \sqrt{k} \cdot \mathsf{disc}_2^{1/64}(Y, \mathcal{R})] \geq c,$$

for a sufficiently small constant $c > 0$.

Proof. To prove the lemma, we will use the following anti-concentration inequality (see [20]).

Lemma 3.6. *Let X be as sum of independent random variables, each attaining values in $[0, 1]$, and $\sigma = \sqrt{\mathsf{Var}[X]} \geq 200$. Then for all $t \in [0, \frac{\sigma^2}{100}]$, we have*

$$\mathsf{Pr}[X \geq \mathsf{E}[X] + t] \geq ce^{-t^2/3\sigma^2}$$

for a sufficiently small constant $c > 0$.

To use Lemma 3.6, we first scale each Y_{ij} such that $Y_{ij} \in [0, 1]$. Since $w = \max_{r\in\mathcal{R}} |Y \cap r|$ is an upper bound of Y_{ij}. We set $Y'_{ij} = Y_{ij}/w$, and $Y'_j = Y_j/w$, then $\mathsf{Var}[Y'_j] = \mathsf{Var}[Y_j]/w^2$. We have already shown in Corollary 3.4 that

$$\mathsf{Var}[Y'_j] = \mathsf{Var}[Y_j]/w^2 \geq \frac{k}{18w^2} \cdot (\mathsf{disc}_2^{1/64}(Y, \mathcal{R}))^2.$$

Since Y_j is the sum of independent random variables, which follows from the combinatorial rectangle property and our input distribution, Y'_j is also the sum of independent random variables, each attaining values in $[0, 1]$, and we can directly apply Lemma 3.6. So when $\mathsf{Var}[Y_j]/w^2 \geq 40000$, we have

$$\mathsf{Pr}\left[Y'_j \geq E[Y'_j] + \frac{\sqrt{k}}{w} \cdot \mathsf{disc}_2^{1/64}(Y, \mathcal{R})\right] \geq c$$

for a sufficiently small constant $c \leq 1$, which is equivalent to the first part of the lemma. The second part can be proved similarly. We just replace each Y'_{ij} with $1 - Y'_{ij}$, and Y'_j with $k - Y'_j$. □

Recall that Y is a set of size t such that $\mathsf{disc}_2^{1/64}(Y, \mathcal{R}) = \mathsf{disc}_2^{1/64}(t)$. We are ready to prove our randomized communication lower bound.

Lemma 3.7. *If $w^2 \leq \frac{k}{\beta}(\mathsf{disc}_2^{1/64}(t))^2$ for large enough constant β, any algorithm solving the ε-approximation problem for $\mathcal{R}$ where $\varepsilon \leq \frac{\mathsf{disc}_2^{1/64}(t)}{t\sqrt{k}}$ with probability at least $1 - c/2$ must communicate $\Omega(tk)$ bits, for a sufficiently small constant c.*

Proof. By Yao's minimax principle [29], we only need to prove a lower bound for the distributional complexity, i.e., the communication complexity of any deterministic protocol which will output a correct ε-approximation with probability $1 - c/2$, where the probability is over the input distribution. We know every deterministic protocol partitions all possible inputs into a set of combinatorial rectangles. If there is a rectangle in this partition with size at least $2^{tk/2}$, then a constant fraction of the inputs in this rectangle cannot be answered correctly. This is because the approximation error allowed is at most $\varepsilon kt = \sqrt{k} \cdot \mathsf{disc}_2^{1/64}(t)$, and by Lemma 3.5, the output can not be a ε-approximation for at least c fraction of the inputs in this rectangle. Let σ be the total measure of rectangles of size at least $2^{tk/2}$, then the error probability is at least σc. Hence, $\sigma \leq 1/2$. This implies that the total number of rectangles is at least $2^{tk-1}/2^{tk/2} = 2^{tk/2-1}$, from which it follows that the communication cost of the deterministic protocol is at least $\log 2^{tk/2-1} = \Omega(tk)$ bits. □

The above lower bound can be more effectively used when have a good bound on w. For a general range space, we only have $w \leq t$. So replace the condition $w^2 \leq \frac{k}{\beta}(\mathsf{disc}_2^{1/64}(t))^2$ with $t^2 \leq \frac{k}{\beta}(\mathsf{disc}_2^{1/64}(t))^2$, i.e., $\frac{t}{\mathsf{disc}_2^{1/64}(t)} \leq \sqrt{\frac{k}{\beta}}$. We set t such that $\frac{t}{\mathsf{disc}_2^{1/64}(t)} = \frac{1}{\varepsilon\sqrt{k\beta}}$, which then simplifies the condition to $k \geq 1/\varepsilon$. Note that this means $\varepsilon = \frac{\mathsf{disc}_2^{1/64}(t)}{t\sqrt{k\beta}}$ which satisfies the condition in Lemma 3.7 for $\beta \geq 1$.

Theorem 3.8. *Given any range space $\mathcal{R}$, if t is a value satisfying*

$$\frac{t}{\mathsf{disc}_2^{1/64}(t)} = \frac{1}{\varepsilon\sqrt{k\beta}}$$

for some large constant β, and if $k \geq 1/\varepsilon$, any algorithm solving the ε-approximation problem for $\mathcal{R}$ with probability $1-c$ must communicate $\Omega(kt)$ bits for sufficiently small c.

The annoying assumption $k \geq 1/\varepsilon$ in the above theorem roots from the anti-concentration inequality (Lemma 3.6) having a restriction on the range of the individual random variables. In our case, the Y_{ij}'s can attain any value in $[0, t]$. To get rid of this assumption, we need to restrict each Y_{ij} to a smaller range. Our idea to achieve this is that, for any algorithm, we first modify it by also running a naive deterministic algorithm computing an ε'-approximation with $\varepsilon' = \varepsilon k/2$ along the side. This changes the transcript of the protocol such that in any rectangle, each Y_{ij} is bounded by an interval of size at most εkt. We leave the details of this arguments to Appendix D.

Theorem 3.9. *Given any range space $\mathcal{R}$, if t is a value satisfying*

$$\frac{t}{\mathsf{disc}_2^{1/64}(t)} = \frac{1}{\varepsilon\sqrt{k\beta}}$$

for some large constant β, and if $k < 1/\varepsilon$, any algorithm solving the ε-approximation problem for $\mathcal{R}$ with constant probability must communicate $\Omega(kt - k \cdot \mathsf{approx}(\varepsilon k/2))$ bits.

Applications. For intervals in $\mathbb{R}^1$, we have the trivial bounds $\mathsf{disc}_2^{1/64}(t) = \Omega(1)$ and $\mathsf{approx}(\varepsilon) = O(1/\varepsilon)$. Then the theorems above yield the communication lower bound of $\Omega(\sqrt{k}/\varepsilon)$. This matches the lower bound recently proved in [27], but our proof is elementary, while the proof of [27] needs the machinery of *information complexity* [8, 6].

For boxes in $\mathbb{R}^d$, the highest l_2-discrepancy lower bound known today is $\mathsf{disc}_2(t) = \Omega(\log^{(d-1)/2} t)$ [23], thus $\mathsf{disc}_2^{1/64}(t) = \Omega(\log^{(d-3)/2} t)$ by Lemma 1.1. The current best upper bound on $\mathsf{approx}(\varepsilon)$ is $O(\frac{1}{\varepsilon}\log^{d+1/2}\frac{1}{\varepsilon})$ [19, 15]. So we have the following communication lower bounds.

Corollary 3.10. *Any randomized algorithm solving the ε-approximation problem for d-dimensional boxes must communicate $\Omega(\frac{\sqrt{k}}{\varepsilon} \cdot \log^{(d-3)/2}\frac{1}{\varepsilon})$ bits if $k \geq 1/\varepsilon$. For $k < 1/\varepsilon$, the lower bound is $\Omega(\frac{\sqrt{k}}{\varepsilon} \cdot \log^{(d-3)/2}\frac{1}{\varepsilon} - \frac{1}{\varepsilon} \cdot \log^{d+1/2}\frac{1}{\varepsilon})$.*

Note that the $\frac{1}{\varepsilon} \cdot \log^{d+1/2}\frac{1}{\varepsilon}$ term is insignificant as long as $k \geq \log^{d+4}\frac{1}{\varepsilon}$.

For halfspaces in $\mathbb{R}^d$, we have $\mathsf{disc}_2(t) = \Omega(t^{1/2-1/2d})$ [2], thus $\mathsf{disc}_2^{1/64}(t) = \Omega(t^{1/2-1/2d})$ by Lemma 1.1. Also since $\mathsf{approx}(\varepsilon) = 1/\varepsilon^{2d/(d+1)}$, we get a lower bound of $\Omega(k^{1/(d+1)}/\varepsilon^{2d/(d+1)})$.

4 Lower Bounds via Generalized Discrepancy

In this section, we will give another characterization of both the deterministic and randomized communication complexity which relates them to the generalized l_∞-discrepancy.

4.1 Our techniques

We obtain our lower bounds through a reduction from t independent instances of a primitive problem, the 1-bit problem, to the ε-approximation problem. Then we prove direct-sum theorems, both deterministic and randomized, for the former. Recall that for a primitive problem which has complexity C, a direct-sum theorem states that the complexity of solving t independent instances of this problem simultaneously is $\Omega(tC)$. However, for deterministic algorithms, the natural requirement, which is that the algorithm should be correct on all instances, turns out to be too strong for the reduction to work. So we have to use a relaxed requirement that the algorithm only needs to solve an arbitrary $1-\alpha$ fraction of the t instances correctly, for some small constant α. For randomized algorithms, we use the usual requirement that the algorithm is correct with constant probability for each of the t instances.

In the *1-bit problem*, each player gets one bit, and their goal is to estimate their sum within additive error Δ, where $\Delta = \Theta(k)$ for deterministic algorithms and $\Delta = \Theta(\sqrt{k})$ for randomized algorithms. It has been shown that the information complexity of the randomized 1-bit problem is $\Omega(k)$ [27]. Then the standard direct-sum arguments using information complexity [6, 7] will give us a direct-sum theorem for the randomized 1-bit problem. There are, however, no effective tools for proving deterministic direct-sum theorems. Our strategy is to first prove a lemma concerning the width of a set of large subsets in the t-dimensional hypercube. The geometric interpretation of the lemma is that, for any ℓ subsets of the t-dimensional hypercube with non-negligible measures, there must be a direction $y \in \{-1, +1\}^t$ such that the sum of the widths of these sets in this direction is $\Omega(\ell\sqrt{t})$. Note that the largest width of any set is $\sqrt{t}$. To show this, the main tool we used is a well-known isoperimetric inequality of the t-dimensional hypercube with the Hamming metric.

Once we have the direct-sum theorem, we use the following idea to build the connection between communication complexity and generalized discrepancy. We fix a range space $(Y, \mathcal{R})$, with $|Y| = t$. Let Π be a deterministic protocol which computes an $\varepsilon/2$-approximation for the range space $(Y, \mathcal{R})$, with $\varepsilon = \frac{\mathsf{rdisc}^{\alpha}(Y,\mathcal{R})}{5t}$. We will show that such a protocol can be used to solve the deterministic direct-sum problem with t independent instances. It seems that this is only possible when $\mathsf{rdisc}^{\alpha}(Y, \mathcal{R}) = O(1)$, since the error allowed in Π is $\varepsilon kt = \frac{k \cdot \mathsf{rdisc}^{\alpha}(Y,\mathcal{R})}{5}$. So we can only set t to be a constant, which would not give us a good lower bound. Our idea is to use discrepancy. We show that if, on the contrary, Π cannot recover a constant fraction of answers within error $O(k)$, then the error will be amplified through discrepancy. More precisely, there will be at least one range $r \in \mathcal{R}$, for which the error made by Π is large than εtk, which contradicts the correctness of Π. The idea for the randomized case is similar.

4.2 Deterministic lower bounds via generalized discrepancy

We first define the deterministic problem more formally. In the deterministic version of the 1-bit problem, each player gets one bit, and their goal is to estimate the sum of their bits within additive error $k/4$. We will use DSE_k to denote this problem. It is not hard to see the communication complexity of DSE_k is $\Omega(k)$. The direct-sum problem $\mathsf{DSE}^t_{k,\alpha}$ is defined as follows. Each player gets an t-bit input x_i. Let $z = \sum_{i=1}^{k} x_i$. The goal is to estimate an arbitrary $(1-\alpha)$ fraction of z_j's within error $k/4$. Note that we do not need to solve all the t copies of the 1-bit problem, but only to solve an arbitrary constant fraction of them. We have the following lower bounds for $\mathsf{DSE}^t_{k,\alpha}$

Lemma 4.1. *For any k, t and small enough constant α, the deterministic communication complexity of $\mathsf{DSE}^t_{k,\alpha}$ is $\Omega(tk)$ bits.*

Before proving the above lemma, we first show how to use this to prove a theorem which relates deterministic communication complexity of the ε-approximation problem to generalized discrepancy.

Theorem 4.2. *Given range space $\mathcal{R}$, constant ε and k, if t is a value satisfying*

$$\frac{t}{\mathsf{rdisc}^{\alpha}(t)} = \frac{1}{5\varepsilon}$$

for some constant d, any deterministic algorithm solving the ε-approximation problem for $\mathcal{R}$ must communicate $\Omega(tk)$ bits.

Proof. As before, we fix a range space $(Y, \mathcal{R})$, where $|Y| = t$ and $\mathcal{R} = \{r_1, \cdots, r_m\}$. The input of each player i is a subset $x_i \subseteq Y$, and we also use $x_i \in \{0,1\}^t$ and $r_\ell \in \{0,1\}^t$ as vectors. Let A be the incidence matrix of this range space. We want to compute an ε-approximation of the multiset $X = x_1 + \cdots + x_k$.

Let Π be a deterministic protocol which computes an $\varepsilon/2$-approximation for the range space $(Y, \mathcal{R})$, with $\varepsilon = \frac{\mathsf{rdisc}^{\alpha}(Y,\mathcal{R})}{5t}$. We will show that such an protocol can be used to solve $\mathsf{DSE}^t_{k,\alpha}$. Let $y_1, \cdots, y_k \in \{0,1\}^t$ be the input of $\mathsf{DSE}^t_{k,\alpha}$. To solve this problem using protocol Π, we first set $x_i = y_i$ for $1 \leq i \leq k$, then run Π on the input $x_1, \cdots, x_k$. We further define z to be a t-dimensional vector such that $z = \sum_i x_i$ and ξ be a m-dimension vector such that $\xi_\ell = \sum_{i=1}^{k} |r_\ell \cap x_i|$ for each $r_\ell \in \mathcal{R}$, i.e., ξ_ℓ is the number of points of X in the range r_ℓ, and it is easy to see $\xi = Az$. After running the protocol we got a vector ξ' such that $||\xi - \xi'||_\infty \leq \varepsilon kt/2$. Given ξ' we find an arbitrary z' such that $||Az' - \xi'||_\infty \leq \varepsilon kt/2$ (there must exists one, as z is valid). By triangle inequality, we have

$$||Az - Az'||_\infty \leq \varepsilon kt \tag{4}$$

We next show that

$$|\{j| \; |z_j - z'_j| > k/4\}| \leq \alpha t. \tag{5}$$

Suppose this is not true, then $4(z - z')/k$ is a α-heavy coloring, and $||4A(z - z')/k||_\infty \geq \mathsf{rdisc}(Y, \mathcal{R})$, which implies that

$$||A(z - z')||_\infty \geq k \cdot \mathsf{rdisc}(Y, \mathcal{R})/4 = 5\varepsilon tk/4 > \varepsilon tk,$$

which is a contradiction to 4. So 5 holds, and z' is a valid answer to $\mathsf{DSE}^t_{k,\alpha}$. Since z' can be computed after running Π without further communication, the communication complexity of Π is at least $\Omega(tk)$. □

Proof of Lemma 4.1. Let $\rho = B_1 \times B_2 \times \cdots \times B_k$ be a combinatorial rectangle and define $|\rho| = |B_1| \cdot |B_2| \cdot \cdots \cdot |B_k|$ as the size of ρ. Let ρ be the largest rectangle in any correct protocol Π. We will show $|\rho| \leq 2^{(1-\alpha)tk}$ for some small enough constant, then the communication cost of the protocol is at least $\log(2^{tk}/|\rho|) = \Omega(tk)$. Let $x, x' \in \rho$ be any two inputs in the rectangle, and we define the t-dimensional vectors z, z' as above, i.e., $z = \sum_{i=1}^{k} x_i$. We will show that in any rectangle ρ of size larger than $2^{(1-\alpha)tk}$, there must be two inputs $x, x' \in \rho$ such that $|\{j \mid |z_j - z'_j| > k/4\}| = \Omega(t)$, which means the rectangle partition induced by any correct deterministic protocol cannot contain such a rectangle. We first prove the following key lemma we will use. The geometric interpretation of the lemma is that, for any ℓ subsets of the t-dimensional hypercube with non-negligible measures, there must be a direction $y \in \{-1, +1\}^t$ such that the sum of the widths of these sets in this direction is $\Omega(\ell\sqrt{t})$. Note that the largest width of any set is $\sqrt{t}$.

Lemma 4.3. *Let $B_1, \cdots, B_\ell$ be a collection of ℓ subsets of the t-dimension hypercube $\{-1,1\}^t$ such that $|B_i| \geq 2^{(1-\beta)t}$ for small enough constant β, then there exists a vector $y \in \{-1,1\}^t$ and a set of vectors $x_i, x'_i \in B_i$ for $1 \leq i \leq \ell$, such that the following holds,*

$$|\sum_{i=1}^{\ell} (x_i - x'_i) \cdot y| \geq \Omega(\ell t), \tag{6}$$

where $x \cdot y$ denotes the inner product of x, y.

Proof. We will first prove that if we uniformly pick a random $y \in \{-1,+1\}^t$, then, for each i, with high probability there exist x_i and x'_i in B_i such that

$$x_i \cdot y - x'_i \cdot y = \Omega(t).$$

The proof is based on the following lemma which is a simple consequence of Talagrand's Inequality [25, 3].

Lemma 4.4. *For any $A \subset \{-1,+1\}^t$, we have*

$$\mathsf{Pr}[A](1 - \mathsf{Pr}[A_d]) \leq e^{-\frac{d^2}{4t}},$$

where A_d is the set of elements in $\{-1,+1\}^t$ with Hamming distance less than d from A, and the probability measure is the uniform measure.

Due to the above lemma, if $\mathsf{Pr}[A] = 2^{-\beta t}$ for some small enough constant, then $1 - \mathsf{Pr}[A_{0.4t}] \leq e^{-\Omega(t)}$. That is for a random y, $\mathsf{Pr}[y \in A_{0.4d}] \geq 1 - e^{-\Omega(t)}$. Since for any $x, y \in \{-1,+1\}^t$, it is obvious that $x \cdot y = t - 2HD(x,y)$, then the probability that there exist $x \in A$ such that $x \cdot y \geq 0.2t$ is at least $1 - e^{-\Omega(t)}$. Similarly, we can prove that the probability that there exists $x \in A$ such that $-x \cdot y \geq 0.2t$ is at least $1 - e^{-\Omega(t)}$, since we can negate all the elements in A and apply the same arguments as above. Now apply the union bound, we have that, for a random y, the probability that there exist x and x' in A such that $x \cdot y - x' \cdot y = \Omega(t)$ is at least $1 - e^{-\Omega(t)} \geq 1/2$.

By our assumption, it is satisfied that $\mathsf{Pr}[B_i] \geq 2^{-\beta t}$ for all i, so for at least half of the y's in $\{-1,+1\}^t$, we can find $x_i, x'_i \in B_i$ such that $x_i \cdot y - x'_i \cdot y = \Omega(t)$ for every i. It is implied that

$$\sum_i \sum_y \max_{x_i, x'_i \in B_i} (x_i - x'_i) \cdot y \geq 2^{t-1} \ell \Omega(t)$$

Then by an averaging argument, there must exist y such that

$$\sum_i \max_{x_i, x'_i \in B_i} (x_i - x'_i) \cdot y \geq \Omega(\ell t),$$

and the lemma follows. □

Let us focus on a rectangle $\rho = B_1 \times B_2 \times \cdots \times B_k$ of size $2^{(1-\beta/2)tk}$. Using a similar averaging argument as in the proof of Lemma C.1, we can easily show that the number of i's such that $|B_i| \geq 2^{(1-\beta)t}$ is at least βk for any small enough constant. W.l.o.g, we assume $|B_i| \geq 2^{(1-\beta)t}$ for $1 \leq i \leq \beta k$. For each element x_i in B_i, we transform it into an element $\hat{x_i}$ of $\{-1,+1\}^t$ by replacing each 0 with -1. Now using Lemma 4.1, we have

$$\sum_{i=1}^{k} (\hat{x_i} - \hat{x'_i}) \cdot y \geq \Omega(kt),$$

for some $y \in \{-1,+1\}^t$ and some $x_i, x'_i \in B_i$ (for $i > \beta k$ we just set $x_i = x'_i$). It is observed that $x_i - x'_i = (\hat{x_i} - \hat{x'_i})/2$, thus

$$\sum_{i=1}^{k} (x_i - x'_i) \cdot y \geq \Omega(kt). \tag{7}$$

Now we got two inputs $x, x' \in \rho$, where $x = x_1 x_2 \cdots x_k$ and $x' = x'_1 x'_2 \cdots x'_k$. Define $z = \sum_i x_i$ and $z' = \sum_i x'_i$. By 7, we have $(z - z') \cdot y \geq \Omega(kt)$. Since $y \in \{-1, +1\}$, it follows that

$$||z - z'||_1 \geq (z - z') \cdot y \geq \Omega(kt).$$

The absolute value of each entry of the vector $z - z'$ is at most k, then by an averaging argument, we have that $|\{j \mid |z_j - z'_j| > k/4\}| \geq \alpha t$, for small enough α. So z, z' cannot be in the same rectangle induced by a correct protocol. Since we proved that there are such a pair of inputs in any rectangle ρ of size at least $2^{(1-\beta/2)tk}$, we conclude that any correct protocol cannot have a rectangle of this size, then it follows that the communication complexity is at least $\log(2^{tk}/2^{(1-\beta/2)tk}) = \beta tk/2 = \Omega(tk)$. □

It was known [22] that, for boxes in $\mathbb{R}^2$, $\mathsf{rdisc}^{\alpha}_{\infty}(t) = \Omega(\log t)$ for any constant $0 < \alpha \leq 1$. Plugging this result into Theorem 4.2 yields a better lower bound of $\Omega(k \log(1/\varepsilon)/\varepsilon)$. The lower bound proved in [2] for halfspaces actually holds for the generalized discrepancy for any constant α, i.e. $\mathsf{rdisc}^{\alpha}_{\infty}(t) = \Omega(t^{1/2-1/2d})$ for d-dimensional halfspaces. Applying Theorem 4.2, we got optimal deterministic lower bounds for halfspaces, which is $\Omega(\frac{k}{\varepsilon^{2d/(d+1)}})$.

4.3 Randomized lower bound via generalized discrepancy

Applying similar ideas as in the deterministic case together with information complexity arguments, we get the following analogous result for randomized protocols, the proof of which is given in Appendix E.

Theorem 4.5. *Given range space $\mathcal{R}$, if t is a value satisfying*

$$\frac{t}{\mathsf{rdisc}^{1/6}(t)} = \frac{1}{2\varepsilon\sqrt{k}},$$

any algorithm solving the ε-approximation problem for $\mathcal{R}$ with constant probability must communicate $\Omega(kt)$ bits.

Roth [22] showed that, for boxes in $\mathbb{R}^2$, $\mathsf{rdisc}^{\alpha}_{\infty}(n) = \Omega(\log n)$ for any constant $0 < \alpha \leq 1$. Plugging this result into Theorem 4.5 yields a better lower bound for $d = 2, 3, 4$ than the one derived from Theorem 3.8, as listed in Table 1.

The error we considered so far is the maximum error (l_∞ norm) of the ε-approximation over all ranges. Another advantage of the generalized discrepancy lower bound is that it easily adapts to other forms of errors, by just changing the form of the discrepancy accordingly. Further details are given in Appendix F.

References

[1] P. K. Agarwal, G. Cormode, Z. Huang, J. M. Phillips, Z. Wei, and K. Yi. Mergeable summaries. *ACM Trans. Database Syst.*, 38(4):26:1–26:28, Dec. 2013.

[2] R. Alexander. Geometric methods in the study of irregularities of distribution. *Combinatorica*, 10(2):115–136, 1990.

[3] N. Alon and J. Spencer. *The Probabilistic Method.* Wiley-Interscience, 2000.

[4] A. Bagchi, A. Chaudhary, D. Eppstein, and M. T. Goodrich. Deterministic sampling and range counting in geometric data streams. *ACM Transactions on Algorithms (TALG)*, 3(2):16, 2007.

[5] N. Bansal. Constructive algorithms for discrepancy minimization. In *Proc. IEEE Symposium on Foundations of Computer Science*, pages 3–10. IEEE, 2010.

[6] Z. Bar-Yossef, T. Jayram, R. Kumar, and D. Sivakumar. An information statistics approach to data stream and communication complexity. *Journal of Computer and System Sciences*, 68:702–732, 2004.

[7] M. Braverman. Interactive information complexity. In *Proc. ACM Symposium on Theory of Computing*, pages 505–524. ACM, 2012.

[8] A. Chakrabarti, Y. Shi, A. Wirth, and A. Yao. Informational complexity and the direct sum problem for simultaneous message complexity. In *Proc. IEEE Symposium on Foundations of Computer Science*, pages 270–278. IEEE, 2001.

[9] B. Chazelle. *The Discrepancy Method: Randomness and Complexity*. Cambridge University Press, 2000.

[10] T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2006.

[11] Z. Huang, L. Wang, K. Yi, and Y. Liu. Sampling based algorithms for quantile computation in sensor networks. In *Proc. ACM SIGMOD International Conference on Management of Data*, 2011.

[12] R. Karp, C. Schindelhauer, S. Shenker, and B. Vocking. Randomized rumor spreading. In *Proc. IEEE Symposium on Foundations of Computer Science*, pages 565–574. IEEE, 2000.

[13] D. Kempe, A. Dobra, and J. Gehrke. Gossip-based computation of aggregate information. In *Proc. IEEE Symposium on Foundations of Computer Science*, pages 482–491. IEEE, 2003.

[14] E. Kushilevitz and N. Nisan. *Communication complexity*. Cambridge university press, 1997.

[15] K. G. Larsen. On range searching in the group model and combinatorial discrepancy. In *Proc. IEEE Symposium on Foundations of Computer Science*, pages 542–549. IEEE, 2011.

[16] Y. Li, P. M. Long, and A. Srinivasan. Improved bounds on the sample complexity of learning. *Journal of Computer and System Sciences*, 62(3):516–527, 2001.

[17] S. Lovett and R. Meka. Constructive discrepancy minimization by walking on the edges. In *Proc. IEEE Symposium on Foundations of Computer Science*, pages 61–67. IEEE, 2012.

[18] J. Matoušek. Tight upper bounds for the discrepancy of half-spaces. *Discrete & Computational Geometry*, 13(1):593–601, 1995.

[19] J. Matoušek. *Geometric Discrepancy: An Illustrated Guide*, volume 18. Springer, 1999.

[20] J. Matoušek and J. Vondrák. The probabilistic method. *Lecture Notes, Department of Applied Mathematics, Charles University, Prague*, 2001.

[21] J. M. Phillips, E. Verbin, and Q. Zhang. Lower bounds for number-in-hand multiparty communication complexity, made easy. In *Proc. ACM-SIAM Symposium on Discrete Algorithms*, 2011.

[22] K. Roth. On a theorem of Beck. *Glasgow Mathematical Journal*, 27(1):195–201, 1985.

[23] K. F. Roth. On irregularities of distribution. *Mathematika*, 1(02):73–79, 1954.

[24] S. Suri, C. Tóth, and Y. Zhou. Range counting over multidimensional data streams. *Discrete and Computational Geometry*, 36(4):633–655, 2006.

[25] M. Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l'Institut des Hautes Etudes Scientifiques*, 81(1):73–205, 1995.

[26] 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*, 16:264–280, 1971.

[27] D. P. Woodruff and Q. Zhang. Tight bounds for distributed functional monitoring. In *Proc. ACM Symposium on Theory of Computing*, pages 941–960. ACM, 2012.

[28] D. P. Woodruff and Q. Zhang. An optimal lower bound for distinct elements in the message passing model. In *Proc. ACM-SIAM Symposium on Discrete Algorithms*, 2014.

[29] A. C. Yao. Probabilistic computations: Toward a unified measure of complexity. In *Proc. IEEE Symposium on Foundations of Computer Science*, pages 222–227. IEEE, 1977.

A Proof of Lemma 1.1

Proof. We first construct a partial coloring χ^1 with discrepancy $\mathsf{disc}_p^\alpha(n)$. Now there are only $(1-\alpha)n$ points that have not been colored. We then color these points using a partial coloring χ^2 with discrepancy $\mathsf{disc}_p^\alpha((1-\alpha)n)$, and after this there are only $(1-\alpha)^2 n$ points left uncolored. We repeat this construction until the number of uncolored points becomes trivially small, say smaller than a suitable constant. Let ℓ be the number of iterations, then $\ell = O(\log n)$ since α is a constant. Finally we set $\chi = \sum_{i=1}^{\ell} \chi^i$ to be our full coloring. Note that χ is a valid coloring as each point is colored only once. The discrepancy of χ can be bounded as follows.

$$m^{1/p}\mathsf{disc}(n) \le ||A\chi||_p \le \sum_{i=1}^{\ell} ||A\chi^i||_p \le \ell m^{1/p} \cdot \mathsf{disc}_p^\alpha(n) = m^{1/p} \cdot O(\log n \cdot \mathsf{disc}_p^\alpha(n)).$$

When $\mathsf{disc}_p^\alpha(n) = \mathrm{poly}(n)$, then

$$\sum_{i=1}^{\ell} ||A\chi^i||_p \le \sum_{i=1}^{\ell} \mathsf{disc}_p^\alpha((1-\alpha)^i n) = O(\mathsf{disc}_p^\alpha(n)).$$

□

B Analysis of the Algorithm

It should be clear that the total size of the k subsets is $n/2^\lambda$, so the weight of each point is 2^λ. We will prove that, with at least constant probability, for every $r \in \mathcal{R}$, we have

$$\left| \sum_{i=1}^{k} (2^\lambda |I_i^\lambda \cap r| - |I_i \cap r|) \right| \le \varepsilon n.$$

Fix some range $r \in \mathcal{R}$. For $1 \le j \le \lambda$, let $\Delta_i^j = 2^{j-1}(2|I_i^j \cap r| - |I_i^{j-1} \cap r|)$. Then the final error of using I_i^λ to approximate I_i on r is

$$2^\lambda |I_i^\lambda \cap r| - |I_i \cap r| = \sum_{j=1}^{\lambda} \Delta_i^j. \tag{8}$$

Sum up (8) over i, the total error is $\Delta = \sum_i \sum_j \Delta_i^j$.

Note that these Δ_i^j's may not be independent. But conditioned on the event $I_i^{j-1} = S$ for any set S, it is easy to see that $\mathsf{E}\left[2|I_i^j \cap r| \mid I_i^{j-1} = S\right] = |S \cap r|$, so $\mathsf{E}[\Delta_i^j \mid I_i^{j-1} = S] = 0$. Thus if we order the Δ_i^j's as $\Delta_1^1, \ldots, \Delta_1^\lambda, \Delta_2^1, \ldots, \Delta_2^\lambda, \ldots, \Delta_k^1, \ldots, \Delta_k^\lambda$, and let Y_ℓ be the partial sum of first ℓ terms of this sequence, then the sequence $Y_0, Y_1, \ldots, Y_{k\lambda}$ form a martingale.

Next, we show that $|\Delta| = |Y_{k\lambda}| > \varepsilon n$ with small probability.

Lemma B.1 (Azuma-Hoeffding). *Let $Y_0, \cdots, Y_m$ be a martingale such that $|Y_\ell - Y_{\ell-1}| \le c_\ell$. Then, for all $\beta > 0$,*

$$\mathsf{Pr}[|Y_m - Y_0| \ge \beta] \le 2e^{-\beta^2/(2\sum_{\ell=1}^{m} c_\ell^2)}.$$

To use the Azuma-Hoeffding inequality, we need an upper bound c_ℓ on each $|Y_\ell - Y_{\ell-1}| = |\Delta_i^j| = 2^{j-1}\left|2|I_i^j \cap r| - |I_i^{j-1} \cap r|\right|$ for some i, j. Recall that I_i^j is the subset of I_i^{j-1} with the same color, so $\left|2|I_i^j \cap r| - |I_i^{j-1} \cap r|\right|$ is exactly the discrepancy of the coloring on r. By the definition of discrepancy, we have $|\Delta_i^j| \le 2^{j-1}\mathsf{disc}(|I_i^{j-1}|)$. Let $c_{ij} = 2^{j-1}\mathsf{disc}(|I_i^{j-1}|)$, and let $D = \sum_i \sum_j c_{ij}^2$. For each $i \in [k]$, let $D_i = \sum_{j=1}^{\lambda} c_{ij}^2$. We have $D = \sum_i D_i$ and

$$D_i = \sum_{j=1}^{\lambda} c_{ij}^2 \le \sum_{j=1}^{\lambda}\left(2^{j-1} \cdot \mathsf{disc}(|I_i^{j-1}|)\right)^2 = \sum_{j=1}^{\lambda}\left(2^{j-1} \cdot \mathsf{disc}(n_i/2^{j-1})\right)^2.$$

Since we assume $\mathsf{disc}(n) \le O(\sqrt{n})$, the above sum is bounded by the last term:

$$D_i \le 2\left(2^{\lambda-1} \cdot \mathsf{disc}(n_i/2^{\lambda-1})\right)^2.$$

We further assume that $(\mathsf{disc}(n))^2$ is a concave function, which is true for all natural range spaces. Thus by Jenson's inequality, we have

$$D \le 2^{2\lambda-1} \cdot k\left(\mathsf{disc}(n/(k2^{\lambda-1}))\right)^2.$$

Then by the Azuma-Hoeffding inequality,

$$\mathsf{Pr}[|Y_{k\lambda}| \ge \varepsilon n] \le 2e^{-(\varepsilon n)^2/(2D)} \le 2\exp\left(-\frac{(\varepsilon n)^2}{2^{2\lambda-1} \cdot k\left(\mathsf{disc}(n/(k2^{\lambda-1}))\right)^2}\right).$$

Set $t = \frac{n}{k2^{\lambda-1}}$. To make this probability smaller than some δ', we only need to set t such that

$$\frac{t}{\mathsf{disc}(t)} \ge \Omega\left(\frac{1}{\varepsilon\sqrt{k}} \cdot \sqrt{\log\frac{1}{\delta'}}\right).$$

The number of iterations, λ, is then determined by the value of t. The communication cost is simply the total size of the k subsets $\sum_{i=1}^{k} n_i/2^{\lambda} = tk$.

The above analysis is for any fixed $r \in \mathcal{R}$, while there are infinitely many ranges in $\mathcal{R}$ on which we need to ensure the accuracy of the estimates. But it is well known that, among all the ranges, only $\mathrm{poly}(1/\varepsilon)$ of them are different enough that one needs to consider, if the range space has bounded VC-dimension [26]. Thus, it is sufficient to set $\delta' = \mathrm{poly}(\varepsilon) \cdot \delta$ and apply a union bound.

C Proof of the Deterministic Lower Bound via ℓ_1-Discrepancy

Lemma C.1. *Let $\rho = B_1 \times B_2 \cdots \times B_k$ be a rectangle of size no less than $2^{tk/2}$, and let $U = \{i : |B_i| \ge 2^{t/4}\}$, then $|U| \ge k/3$.*

Proof. U is the same as $\{i : \log|B_i| \ge t/4\}$. By our assumption, we have

$$\sum_{i=1}^{k} \log|B_i| = \log|\rho| \ge tk/2.$$

Because for each i, we have $\log|B_i| \le t$, by an averaging argument, we conclude that $|U| \ge k/3$. □

Now we consider a player i in U, that is $|B_i| \geq 2^{t/4}$. It can be shown that there are two vectors x, y in B_i with Hamming distance no less than $t/64$.

Lemma C.2. *If $|B_i| \geq 2^{t/4}$, then for any $x \in B_i$, it hold for almost all vectors $y \in B_i$ that $HD(x, y) \geq t/64$, where $HD(x, y)$ is the hamming distance between x and y.*

Proof. We will use the following bound for the sum of the first z binomial coefficients.

$$\sum_{i=0}^{z} \binom{n}{i} \leq \left(\frac{en}{z}\right)^z. \tag{9}$$

The total number of different y's with $HD(x, y) \leq t/64$ is

$$\sum_{i=0}^{t/64} \binom{t}{i} \leq \left(\frac{en}{t/64}\right)^{t/64} < 2^{t/8}.$$

The lemma follows because $|B_i| \geq 2^{t/4} > 2^{t/8}$. □

Let A be the incidence matrix of $(Y, \mathcal{R}_{|Y})$ with $|Y| = t$, $|\mathcal{R}_{|Y}| = m$.

Lemma C.3. *Let $\rho = B_1 \times B_2 \cdots \times B_k$ be any rectangle, then the approximation error in the rectangle ρ is at least*

$$\frac{1}{m} \cdot \sum_{i=1}^{k} ||A(x_i - y_i)||_1,$$

for any $x_i, y_i \in B_i$, $i = 1, 2, \cdots, k$.

Proof. Define $d_i = A(x_i - y_i)$. The jth entry of d_i is $d_{i,j} = |x_i \cap r_j| - |y_i \cap r_j|$. Let $E_j = \sum_{i=1}^{k} |d_{i,j}|$, then

$$\max_{1 \leq j \leq m} E_j \geq \frac{1}{m} \cdot \sum_{j=1}^{m} E_j = \frac{1}{m} \cdot \sum_{j=1}^{m} \sum_{i=1}^{k} |d_{i,j}| = \frac{1}{m} \cdot \sum_{i=1}^{k} ||A(x_i - y_i)||_1.$$

W.l.o.g., we assume $E_1 = \max_i E_i$, and we next show that the approximation error of ρ is at least E_1. It is sufficient to find two inputs $(z_1, \cdots, z_k), (z'_1, \cdots, z'_k)$ with $z_i, z'_i \in B_i$ for every i, such that

$$\left| \sum_{i=1}^{k} |z_i \cap r_1| - \sum_{i=1}^{k} |z'_i \cap r_1| \right| \geq E_1. \tag{10}$$

We set

$$z_i = \begin{cases} x_i, & \text{if} \quad (Ax_i)_1 \geq (Ay_i)_1; \\ y_i, & \text{otherwise,} \end{cases}$$

and set z'_i in the opposite way. It is easy to check that these two inputs satisfy (10), then the lemma follows. □

From Lemma C.2, we know for each $i \in U$, we can find a pair $x_i, y_i \in B_i$ with $HD(x_i, y_i) \geq t/64$. As a result, for each $i \in U$, we can view $(x_i - y_i) \in \{-1, 0, +1\}^t$ as a $1/64$-partial coloring for Y, then it directly follows from Lemma C.3 that

Corollary C.4. *Let $\rho = B_1 \times B_2 \cdots \times B_k$ be any rectangle of size no less than $2^{tk/2}$, then the approximation error in ρ is at least*

$$\frac{k}{3} \cdot \mathsf{disc}_1^{1/64}(Y, \mathcal{R}).$$

Proof.

$$\begin{aligned}
\frac{1}{m} \cdot \sum_{i=1}^{k} ||A(x_i - y_i)||_1 &\geq \frac{1}{m} \cdot \sum_{i \in U} ||A(x_i - y_i)||_1 \\
&\geq \sum_{i \in U} \mathsf{disc}_1^{1/64}(Y, \mathcal{R}) \\
&\geq \frac{k}{3} \cdot \mathsf{disc}_1^{1/64}(Y, \mathcal{R})
\end{aligned}$$

□

Since the total number of points in any input is $n \leq tk$, a valid $\frac{1}{3t}\mathsf{disc}_1^{1/64}(Y, \mathcal{R})$-approximation only allows error less than $\frac{k}{3} \cdot \mathsf{disc}_1^{1/64}(Y, \mathcal{R})$. So the above lemma shows that the partitioning $|P|$ produced by any correct protocol computing a $\frac{1}{3t}\mathsf{disc}_1^{1/64}(Y, \mathcal{R})$-approximation cannot contain a rectangle of size more than $2^{tk/2}$, which implies that the communication cost is at least $\log |P| = \Omega(tk)$ bits. As we pick Y to maximize the discrepancy of $(Y, \mathcal{R})$, so $\mathsf{disc}_1^{1/64}(Y, \mathcal{R}) = \mathsf{disc}_1^{1/64}(t)$, then we have proved our deterministic lower bound.

D Dealing with Small k

Here we briefly describe how to deal with the case when $k < 1/\varepsilon$. Let $(Y, \mathcal{R})$ and $(X, \mathcal{R})$ be defined as above. Given any protocol Π computing an ε-approximation of $(X, \mathcal{R})$, we modify it as follows. We first run the protocol Π, and when it finishes, each player deterministically computes an $(\varepsilon k/2)$-approximation of its own set x_i and sends it to the coordinator. We use Π' to denote this new protocol, clearly it is correct if Π is correct. If the cost of Π is ℓ, the cost of Π' is at most $\ell + k \cdot \mathsf{approx}(\varepsilon k/2)$. Suppose there is a rectangle $\rho = B_1 \times B_2 \times \cdots \times B_k$ of size at least $2^{tk/2}$ in Π'. Our new protocol has the property that for any $x_1, x_2 \in B_i$, $|A(x_1 - x_2)|_\infty \leq \varepsilon kt$, since a deterministic local $\varepsilon k/2$-approximation was computed and sent to the coordinator. This means the range of the random variable $Y_{i,j}$ has size at most εkt, i.e., there is a fixed number ϕ_i such that $0 \leq Y_{i,j} - \phi_i \leq \varepsilon kt$. We now set $Y'_{i,j} = \frac{Y_{i,j} - \phi_i}{\varepsilon kt}$, then it is in the range $[0, 1]$, and put $Y'_j = \sum_i Y'_{i,j}$. Observing that $\mathsf{Var}[Y'_{i,j}] = \mathsf{Var}[Y_{i,j}]/(\varepsilon kt)^2$, we can apply the same arguments as in the proof of Lemma 3.5, except now $w = \varepsilon kt$, and get an analogous result as follows: If $\mathsf{Var}[Y_j] = \frac{k}{18} \cdot (\mathsf{disc}_2^{1/64}(Y, \mathcal{R}))^2 \geq 40000(\varepsilon kt)^2$, then

$$\mathsf{Pr}[Y_j \geq E[Y_j] + \sqrt{k} \cdot \mathsf{disc}_2^{1/64}(Y, \mathcal{R})] \geq c$$

and

$$\mathsf{Pr}[Y_j \leq E[Y_j] - \sqrt{k} \cdot \mathsf{disc}_2^{1/64}(Y, \mathcal{R})] \geq c,$$

for some sufficiently small constant $c > 0$. Thus the same result as Lemma 3.7 follows. Now, given ε, k, we set t such that $\frac{t}{\mathsf{disc}_2^{1/64}(t)} = \frac{1}{\varepsilon\sqrt{k\beta}}$. Then in order to satisfy the variance requirement that $\frac{k}{18} \cdot (\mathsf{disc}_2^{1/64}(Y, \mathcal{R}))^2 \geq 40000(\varepsilon kt)^2$, we only need to set β to be a large enough constant

E Proof of Theorem 4.5

Here we introduce some basic definitions from information theory, see [10] for a comprehensive introduction. For any random variables X, Y, Z, we use $H(X|Y)$ to denote the conditional entropy of X given Y, and $I(X, Y|Z) = H(X|Z) - H(X|Y, Z)$ to denote the conditional mutual information between X and Y given Z. Recall the randomized k-party 1-bit problem, which is defined as follows. The input for each player is a bit y_i, and the goal is to determine whether the sum $\sum_{i=1}^{k} y_i$ is greater than $k/2 + \sqrt{k}/2$ or smaller than $k/2 - \sqrt{k}/2$. We need the following result.

Lemma E.1 ([27])**.** *Let P be the transcript of any randomized protocol solving the* 1*-bit problem in the blackboard model with probability* $3/4$*, then we have*

$$I(Y_1, \cdots, Y_k; P|R) = \Omega(k),$$

where the mutual information is measured under uniform input distribution, and R is the public randomness used in P.

As before, we fix a range space $(Y, \mathcal{R})$, where $|Y| = t$ and $\mathcal{R} = \{r_1, \cdots, r_m\}$. The input of each player i is a subset $x_i \subseteq Y$, and we also use $x_i \in \{0,1\}^t$ and $r_\ell \in \{0,1\}^t$ as vectors. We want to compute an ε-approximation of the multiset $x = x_1 + \cdots + x_k$. We further define z to be an t-dimensional vector such that $z_j = \sum_i x_{i,j}$ for each $j \in [t]$.

Let Π be a randomized protocol which computes an $\varepsilon/2$-approximation with probability 0.9, where $\varepsilon \le \frac{\mathsf{rdisc}^{1/6}(Y,\mathcal{R})}{2\sqrt{kt}}$. We will show that, the information cost of Π measured under the uniform distribution is $\Omega(tk)$, which is t times the information cost of the 1-bit problem. To do so, we show that, by running Π, we can recover $\Omega(t)$ z_j's within additive error $\sqrt{k}$, which implies that Π actually solves $\Omega(t)$ independent instances of the 1-bit problem. This is quite counter-intuitive, because the error of Π is $O(\varepsilon tk) = O(\sqrt{k} \cdot \mathsf{rdisc}^{1/6}(t))$, which is too large. It seems that if we want to recover the z_j's within additive error $\sqrt{k}$, we can only set t to be a constant, which would not give us a good lower bound. Our idea is to use discrepancy. We show that if, on the contrary, Π cannot recover a constant fraction of the z_j's within error $\sqrt{k}$, then the error will be amplified through discrepancy. More precisely, there will be at least one range $r \in \mathcal{R}$, for which the error made by Π is large than εtk, which contradicts the correctness of Π. To formalize the above intuition, we first give a reduction that uses Π to solve the 1-bit problem by embedding the input instance to a random position in Y. Then we use the symmetry of our reduction to prove the correctness of this reduction. As the embedding position is random and there are t positions, we can then use information-complexity arguments to prove that the information cost of Π is at least $\Omega(t)$ times that of the 1-bit problem.

Let ξ be an m-dimensional vector such that $\xi_\ell = \sum_{i=1}^{k} |r_\ell \cap x_i|$ for each $r_\ell \in \mathcal{R}$. Given input $y_1, \cdots, y_k$ in the 1-bit problem, which are drawn from uniform distribution ($Y_1, \cdots, Y_k$ are the corresponding random variables). Our protocol for solving the 1-bit problem using Π is as follows.

(1) The players use public randomness to sample an element $j \in [t]$, and each player i sets $x_{i,j} = y_i$. (2) For each $j' \neq j$, each player i sets $x_{i,j'} = 0$ or 1 with equal probability. Note that this only needs private randomness. (3) All players run together the protocol Π on the input they have just constructed, and the coordinator gets an $\varepsilon/2$-approximation. (4) For each $\ell \in [m]$, the coordinator uses the $\varepsilon/2$-approximation computed to recover a value ξ'_ℓ for each ℓ, which is an approximation of ξ_ℓ. (5) The coordinator then computes a t-dimensional vector z' that is compatible with ξ', i.e., $||Az' - \xi'||_\infty \le \varepsilon tk/2$. (6) Answer 1 if $z'_j \ge k/2$, otherwise 0.

Lemma E.2. *If Π is a randomized protocol which computes an $\varepsilon/2$-approximation with probability* 0.9*, where $\varepsilon \le \frac{\mathsf{rdisc}^{1/6}(Y,\mathcal{R})}{2\sqrt{kt}}$, then the above protocol correctly solves the* 1*-bit problem with constant probability, and the information cost of Π is at least $\Omega(tk)$ under uniform distribution.*

Proof. We first show the correctness of the protocol. Suppose the $\varepsilon/2$-approximation computed by Π is correct, which happens with probability 0.9. Then both $||Az-\xi'||_\infty \leq \varepsilon tk/2$ and $||Az'-\xi'||_\infty \leq \varepsilon tk/2$. By triangle inequality, we have

$$||A(z-z')||_\infty = ||Az - Az'||_\infty \leq \varepsilon tk.$$

We claim that for $5/6$ fraction of $j \in [t]$, it holds that $|z_j - z'_j| \leq \sqrt{k}/2$. Suppose it is not true, then $2(z-z')/\sqrt{k}$ is a $1/6$-heavy coloring, and

$$||A(z-z')||_\infty = \frac{\sqrt{k}}{2}||2A(z-z')/\sqrt{k}||_\infty > \frac{\sqrt{k}}{2} \cdot \mathsf{rdisc}^{1/6}(Y,\mathcal{R}) \geq \varepsilon tk,$$

which is a contradiction.

One important property of our reduction is that the input constructed for Π is totally symmetric. In other words, although we sample j first, we can apply the principle of deferred decision, and reveal the value of j after z' is computed. Thus, with probability $5/6$, we have $|z_j - z'_j| \leq \sqrt{k}/2$. Conditioned on this happening, $z_j \geq k/2 + \sqrt{k}/2$ implies $z'_j \geq k/2$, which shows that the output of the protocol is correct. In all, the error probability of the protocol is at most $0.1 + 0.9 \cdot 1/6 = 1/4$.

Next we analysis of the information cost of Π. We need the following property of mutual information.

Proposition 1 (see [10]). Super-additivity of mutual information: *If $X^1, \cdots, X^t$ are conditional independent given Z, then $I(X^1,\cdots,X^t;Y|Z) \geq \sum_{i=1}^t I(X^i;Y|Z)$.*

We use P to denote the above protocol for 1-bit. Let J, Y_i, X_i be the corresponding random variable of j, y_i, x_i. The public randomness used in P is J and R, where R is the public randomness of Π. By Lemma E.1, we have

$$\begin{aligned}
\Omega(k) &= I(Y_1,\cdots,Y_k;P|J,R) \\
&= \sum_{j=1}^{t} \mathsf{Pr}[J=j] \cdot I(Y_1,\cdots,Y_k;P|J=j,R) \\
&= \frac{1}{t} \cdot \sum_{j=1}^{t} I(X_{1,j},\cdots,X_{k,j};\Pi|R) \qquad (11) \\
&\leq \frac{1}{t} \cdot I(X_1,\cdots,X_k;\Pi|R). \qquad (12)
\end{aligned}$$

The equality (11) holds because the joint distributions $(Y_1,\cdots,Y_k,P,R|J=j)$ and $(X_{1,j},\cdots,X_{k,j},\Pi,R)$ are the same by our construction. Inequality (12) holds because the tuples $(X_{1,j},\cdots,X_{k,j})$ for $j \in [q]$ are conditionally independent given R, and we apply the supper-additivity of mutual information. So we have shown the information cost, $I(X_1,\cdots,X_k;\Pi|R)$, of the protocol is $\Omega(kt)$, when the inputs $(X_1,\cdots,X_k)$ are distributed uniformly. By standard arguments, the communication cost is at least the information cost (e.g., see [6]). □

Finally, we set Y of size t satisfying $\mathsf{rdisc}^{1/6}(Y,\mathcal{R}) = \mathsf{rdisc}^{1/6}(t)$, and Theorem 4.5 is proved.

F Handling different error norms

So far we have only studied the l_∞ norm of the error (i.e., maximum error) from an ε-approximation. It is too strong for some applications, and different norms have been considered. The l_p ε-approximation of X with respect to a range space $\mathcal{R}$ is a subset $Y \subset X$ if

$$\left(\frac{1}{m} \cdot \sum_{r \in \mathcal{R}, X \cap r \text{ unique}} \left| \frac{|Y \cap r|}{|Y|} - \frac{|X \cap r|}{|X|} \right|^p \right)^{1/p} \leq \varepsilon.$$

Theorem 4.5 can easily be extended to hold for l_p ε-approximations. In the proof, any norm will work (actually only triangle inequality is needed). Here we state the following result without going through the details again.

Theorem F.1. *Given range space $\mathcal{R}$, if t is a value satisfying*

$$\frac{t}{\mathsf{rdisc}_p^{1/6}(t)} = \frac{1}{2\sqrt{k}\varepsilon},$$

any algorithm solving the l_p ε-approximation problem for $\mathcal{R}$ must communicate $\Omega(kt)$ bits.

The lower bound proved in [2] for halfspaces actually holds for the generalized l_2-discrepancy. Thus, the lower bound listed in Table 1 for computing (l_∞) ε-approximations for halfspaces actually also holds for computing l_2 ε-approximations.

Corollary F.2. *Any randomized algorithm solving the l_2 ε-approximation problem for d-dimensional halfspaces must communicate $\Omega(k^{1/(d+1)}/\varepsilon^{2d/(d+1)})$ bits.*