Electronic Colloquium on Computational Complexity, Report No. 83 (2015)

The Simultaneous Communication of Disjointness with Applications to Data Streams

Omri Weinstein* David P. Woodruff†

Abstract

We study k-party set disjointness in the simultaneous message-passing model, and show that even if each element $i \in [n]$ is guaranteed to either belong to all k parties or to at most $O(1)$ parties in expectation (and to at most $O(\log n)$ parties with high probability), then $\Omega(n \min(\log 1/\delta, \log k)/k)$ communication is required by any δ-error communication protocol for this problem (assuming $k = \Omega(\log n)$).

We use the strong promise of our lower bound, together with a recent characterization of turnstile streaming algorithms as linear sketches, to obtain new lower bounds for the well-studied problem in data streams of approximating the frequency moments. We obtain a space lower bound of $\Omega(n^{1-2/p}\varepsilon^{-2} \log M \log 1/\delta)$ bits for any algorithm giving a $(1+\varepsilon)$-approximation to the p-th moment $\sum_{i=1}^n |x_i|^p$ of an n-dimensional vector $x \in \{\pm M\}^n$ with probability $1-\delta$, for any $\delta \geq 2^{-o(n^{1/p})}$. Our lower bound improves upon a prior $\Omega(n^{1-2/p}\varepsilon^{-2} \log M)$ lower bound which did not capture the dependence on δ, and our bound is optimal whenever $\varepsilon \leq 1/\text{poly}(\log n)$. This is the first example of a lower bound in data streams which uses a characterization in terms of linear sketches to obtain stronger lower bounds than obtainable via the one-way communication model; indeed, our set disjointness lower bound provably cannot hold in the one-way model.

1 Introduction

Set disjointness is one of the cornerstones of complexity theory. Throughout the years, this communication problem has played a key role in obtaining unconditional lower bounds in many models of computation, including proof complexity, data streams, data structures and algorithmic game theory (see [CP10] and references therein). Many variants of this problem were studied, starting with the standard two-party model (e.g., [KS92]) and recently in several multiparty communication models ([She14, BEO$^+$13, BO15]).

Motivated by streaming applications, we study a promise version of number-in-hand multiparty disjointness in the public-coin *simultaneous message passing* model of communication (SMP). In this setting, there are k players each with a bit string $x^i \in \{0,1\}^n$, $i \in [k] = \{1, 2, \ldots, k\}$, who are promised that their inputs satisfy one of the following cases:

- (NO instance) for all $j \in [n]$, the number of $i \in [k]$ for which $x^i_j = 1$ is distributed as $\text{Bin}(k, 1/k)$, or
- (YES instance) there is a unique $j^* \in [n]$ for which $x^i_{j^*} = 1$ for all $i \in [k]$, and for all $j \neq j^*$, the number of $i \in [n]$ for which $x^i_j = 1$ is distributed as $\text{Bin}(k, 1/k)$.

The players simultaneously send a message $M^i(x^i, R)$ to a referee, where R is a public-coin that the players share. The referee then outputs a function $f(M^1(x^1, R), \ldots, M^k(x^k, R), R)$, which should equal 1 if the inputs form a YES instance, and equal 0 otherwise. Notice that if $X \sim \text{Bin}(k, 1/k)$, then $\Pr[X > \ell] \leq (e/\ell)^\ell$, and so by a union bound for all coordinates j in a NO instance, the number of $i \in [k]$ for which $X^i_j = 1$ is $O(\log n/\log\log n)$. Thus, for $k = \Omega(\log n/\log\log n)$, and in fact $k = n^{\Omega(1)}$ in our context below, NO and YES instances are distinguishable.

Our first contribution. We show an $\Omega(n \min(\log 1/\delta, \log k)/k)$ total communication lower bound for any protocol which succeeds with probability at least $1-\delta$ in solving this promise problem in the public-coin SMP model (Theorem 3.1 below).

*Princeton University, Princeton NJ 08544, USA. Research supported by a Simons Fellowship in Theoretical Computer Science and NSF Award CCF-1215990.

†IBM Research, Almaden.

ISSN 1433-8092

We then show how this result can be used to obtain strong space lower bounds in the turnstile data stream model. In this model, an integer vector x is initialized to 0^n and undergoes a long sequence of additive updates to its coordinates. The t-th update in the stream has the form $x_i \leftarrow x_i + \delta_t$, where δ_t is an arbitrary (positive or negative) integer. At the end of the stream we are promised that $x \in \{-M, -M+1, \ldots, M\}^n$ for some bound M which is typically assumed to be at least n (and which we assume here).

Approximating the frequency moments $F_p = \sum_{i=1}^n |x_i|^p$ is one of the most fundamental problems in data streams, starting with the seminal work of Alon, Matias, and Szegedy [AMS99]. The goal is to output a number $\hat{F}_p \in [(1-\varepsilon)F_p, (1+\varepsilon)F_p]$ with probability at least $1-\delta$ using as little memory in bits as possible. It is known that for $0 < p \leq 2$, $\Theta(\varepsilon^{-1}\log(M)\log 1/\delta)$ bits of space is necessary and sufficient [KNW10, JW13]. Ideas here have been the basis of many other streaming algorithms and lower bounds, with connections to linear algebra [SW11, CDM$^+$13] and information complexity [CKW12, BGPW13].

Perhaps surprisingly, for $p > 2$ a polynomial (in n) amount of space is required [SS02, BYJKS04a, CKS03]. The best known upper bound is due to Ganguly and achieves space

$$O(n^{1-2/p}\epsilon^{-2}\log n \cdot \log(M)\log(1/\delta))/\min(\log n, \epsilon^{4/p-2}))).$$

In the case that $\epsilon \leq 1/\text{poly}(\log n)$, this simplifies to $O(n^{1-2/p}\epsilon^{-2}\log M\log(1/\delta))$. On the other hand, if ϵ is a constant, this simplifies to $O(n^{1-2/p}\log n\log M\log(1/\delta))$. The latter complexity is also achieved by algorithms of [AKO10, And]. The lower bound, on the other hand, for any ε, δ is only $\Omega(n^{1-2/p}\varepsilon^{-2}\log M)$ [LW13]. A natural question is whether there are algorithms using less space and achieving a high success probability, that is, if one can do better than just repeating the constant probability data structure and taking a median of $\Theta(\log 1/\delta)$ independent estimates. While there is some work on tightening the bounds in the context of linear sketches over the reals [ANPW13a, LW13], these lower bounds do not yield lower bounds in the streaming setting; for more discussion on this, see below.

Our Second Contribution. We show for any $\varepsilon \in (0,1)$, any $\delta \geq 2^{-o(n^{1/p})}$, and constant $p > 2$, any algorithm obtaining a $(1+\varepsilon)$-approximation to F_p in the turnstile streaming model requires $\Omega(n^{1-2/p}\varepsilon^{-2}\log M\log(1/\delta))$ bits of space (Theorem 6.1 below). Our lower bound is optimal for any $\epsilon \leq 1/\text{poly}(\log n)$. As argued in [LW13], this is an important regime of parameters. Namely, if $\varepsilon = 1\%$, we have that for, e.g., $n = 2^{32}$, $\varepsilon^{-1} \geq \log n$. Our result is a direct strengthening of the $\Omega(n^{1-2/p}\varepsilon^{-2}\log M)$ lower bound of [LW13] which cannot be made sensitive to the error probability δ. Moreover, even for constant ϵ, our lower bound of $\Omega(n^{1-2/p}\log M\log(1/\delta))$ bits improves prior work by a $\log(1/\delta)$ factor. We note that for constant ε, the upper bounds still have space $O(n^{1-2/k}\log n\log M\log(1/\delta))$ bits, so while we obtain an improvement, there is still a gap in this case.

While the ultimate goal in this line of research is to obtain tight space bounds simultaneously for any $\varepsilon, \delta \in (0,1)$ and $p > 2$, our result is the first to obtain tight bounds simultaneously in ε and δ for a wide range of parameters. Our proof technique is also quite different than previous work, and the first to bypass the limitations of one-way communication complexity. This is necessary since the problem considered in [LW13] has a protocol with information cost $O(n^{1-2/p}\varepsilon^{-2}\log M)$ with 0 error probability, which can be compressed to a protocol with this amount of communication and exponentially small error probability. We give a description of this protocol in Appendix B and explain why it implies the problem considered in [LW13] does not give stronger lower bounds.

Our Techniques. The key ingredient of our result is proving the aforementioned simultaneous communication lower bound on the promise version of k-party set disjointness. To do so, we use the information complexity paradigm, which allows one to reduce the problem, via a direct sum argument, to the δ-error SMP complexity of a primitive problem – the k-party AND function with the aforementioned promise. We lower bound the information complexity of AND under the NO distribution (an independent bit $\sim Ber(1/k)$), by asking how many independent messages (over her private randomness) the player would need to send in order to convince one that her input is 0 or 1. We use the product structure of Hellinger distance, and relate this quantity to the amount of information a single message of the player reveals via the Maximum Likelihood Estimation principle. To obtain our stronger bound of $\Omega(n\log(1/\delta)/k)$ for any $\delta \geq 2^{-o(n^{1/p})}$, we restrict all players to have the same (randomized) message function. This assumption turns out to be possible in our application, as we observe that linear sketches can in fact be simulated by *symmetric* SMP protocols (see below).

A Reduction to Streaming: To lower bound the space complexity of a streaming algorithm we need a way of relating it to the communication cost of a protocol for this disjointness problem. We use a recent result of Li, Nguyen, and Woodruff [LNW14] showing there is a near-optimal streaming algorithm for any problem in the turnstile model which can be implemented by maintaining $A \cdot x$ in the stream, where A is a matrix with poly(n)-bounded integer entries, and A is sampled from a fixed set of $O(n \log m)$ hardwired matrices. In [LNW14] near-optimal meant up to an $O(\log n)$ multiplicative factor in space, which would not suffice here. However, their proof shows if one maintains $A \cdot x \mod q$, where q is a vector of integers one for each coordinate (which depends on A but not on x), then this is optimal up to a *constant* factor (we prove this formally in Section 5). Notice that this need not be optimal for a *specific family of streams*, such as those arising in our communication game, though we use the fact that by results in [LNW14] an algorithm which succeeds with good probability *for any* fixed stream has this form, and therefore we can assume this form in our reduction. This implies a public-coin simultaneous protocol since the players can use the public coin to choose an (A, q) pair, then each communicate $A \cdot x^i \mod q$ to the referee, who can combine these (using linearity) to obtain $A \cdot (\sum_{i=1}^k x^i) \mod q$. This simulation also implies all players have the same message function, even conditioned on the public coin.

We stress that the use of a public-coin simultaneous communication model is essential for our result, as there is an $O(n/k)$ total communication upper bound with exponentially small error probability in the one-way communication model (similar to the multi-round 2-player protocol of Håstad and Wigderson [HW07], which may also be derivable from [BKSV14], we prove this formally in Appendix C).

Given this reduction, one of the player's messages must be $\Omega(n \log(1/\delta)/k^2)$ bits long, which lower bounds the space complexity of the streaming algorithm. By setting $k = \varepsilon n^{1/p}$, and by having the referee add $n^{1/p} e_{j*}$ to the stream, where e_{j*} is the standard unit vector in direction j^*, one can show with probability $1 - \delta$, YES and NO instances differ by a $(1 + \varepsilon)$-factor in $F_p(x)$. This is true even given our relaxed definition of disjointness, in which we allow some coordinates to be as large as $\Theta(\log n / \log \log n)$, provided the average of the k-th powers of these coordinates is $\Theta(1)$.

We are not done though, as we seek an extra $\log M$ factor in the lower bound, and for this we superimpose $\Theta(\log M)$ independent copies of this problem at different scales, in a similar way as done for communication problems in previous work [LW13], and ask the referee to solve a random scaling. There are some technical differences needed to execute this approach in the high $(1 - \delta)$ probability regime.

Related Work: We summarize the previous work on the frequency moments problem in Table 1. A few papers [ANPW13b, PW12, LW13] study the "sketching model" of F_p-estimation in which the underlying vector x is in $\mathbb{R}^n$, rather than in the discrete set $\{-M, -M+1, \ldots, M\}^n$. The goal is to design a distribution over linear maps $A : \mathbb{R}^n \to \mathbb{R}^s$, for some $s \ll n$, so that for any fixed vector $x \in \mathbb{R}^n$, one can $(1+\varepsilon)$-approximate $\|x\|_p^p$ with constant probability by applying an estimation procedure $E : \mathbb{R}^s \to \mathbb{R}$ to Ax. We want the smallest s for a given ε and n. Lower bounds in the sketching model do not imply lower bounds in the turnstile model; this is even true given the recent work [LNW14] characterizing turnstile streaming algorithms as linear sketches. The main issue is that dimension lower bounds in the sketching model are shown for input vectors *over the reals*, while it is conceivable that a linear sketch with fewer dimensions does in fact exist if the input is restricted to be in the integer box $\{-M, -M+1, \ldots, M+1\}^n$. For instance, the inner product of x with the single vector $(1, 1/(M+1), 1/(M+1)^2, \ldots, 1/(M+1)^{n-1})$ is enough to recover x, so a sketching dimension of $s = 1$ suffices. What we are really interested in is a linear sketch with polynomially bounded integer entries, and it is an open question to transport dimension lower bounds in the sketching model to space lower bounds in the turnstile streaming model.

Other related work is that of Jayram and Woodruff [JW13] which gives lower bounds in terms of δ for F_p for $p \leq 2$. This regime, as mentioned, is fundamentally different and the communication problems there are based on two-player gap-Hamming and Index problems, which have hard product distributions. In contrast we study multi-player communication problems under non-product distributions.

There is also work on direct sums by Molinaro, Woodruff, and Yaroslavtsev [MWY13], which shows that for some problems, solving all n copies of the problem simultaneously with probability $2/3$, is as hard as solving each copy independently with probability $1 - 1/n$. The techniques in that paper do not seem to apply here, since we are interested in solving an OR rather than all copies, and so the output reveals a lot less information about the inputs. As observed in [BCK$^+$14], there is a quite substantial difference in solving the OR versus all copies of a problem.

As for the communication problem we study, we note that Braverman and Oshman [BO15] recently obtained a tight $\Omega(n \log k + k)$ lower bound on the unbounded-round number-in-hand communication complexity of the k-party set disjointness function. Of course, this lower bound applies in particular to simultaneous protocols and is much stronger than the one proven in this paper $(\Omega(n \cdot \log(1/\delta)/k))$. However, this stronger lower bound holds only for distributions which (vastly) violate the promise required for our streaming application, and therefore their lower bound is useless in our context.

F_p Algorithm	Space Complexity
[IW05]	$O(n^{1-2/p}\epsilon^{-O(1)} \log^{O(1)} n \log(M))$
[BGKS06]	$O(n^{1-2/p}\epsilon^{-2-4/p} \log n \log^2(M))$
[MW10]	$O(n^{1-2/p}\epsilon^{-O(1)} \log^{O(1)} n \log(M))$
[AKO10]	$O(n^{1-2/p}\epsilon^{-2-6/p} \log n \log(M))$
[BO10]	$O(n^{1-2/p}\epsilon^{-2-4/p} \log n \cdot g(p,n) \log(M))$
[And]	$O(n^{1-2/p} \log n \log(M) \epsilon^{-O(1)})$
[Gan11], **Best upper bound**	$O(n^{1-2/p}\epsilon^{-2} \log n \cdot \log(M)/ \min(\log n, \epsilon^{4/p-2})))$
[AMS99]	$\Omega(n^{1-5/p})$
[Woo04]	$\Omega(\epsilon^{-2})$
[BYJKS04a]	$\Omega(n^{1-2/p-\gamma}\epsilon^{-2/p})$, any constant $\gamma > 0$
[CKS03]	$\Omega(n^{1-2/p}\epsilon^{-2/p})$
[WZ12]	$\Omega(n^{1-2/p}\epsilon^{-4/p}/ \log^{O(1)} n)$
[Gan12]	$\Omega(n^{1-2/p}\epsilon^{-2}/ \log n)$
[LW13]	$\Omega(n^{1-2/p}\epsilon^{-2} \log(M))$

Table 1: Results are in bits and for constant $p > 2$. The results are stated for constant probability; all results can be made to achieve $1 - \delta$ success probability by repeating the data structure independently $O(\log 1/\delta)$ times and taking the median of estimates; this blows up the space by a multiplicative $O(\log 1/\delta)$ factor. Here, $g(p, n) = \min_{c \text{ constant}} g_c(n)$, where $g_1(n) = \log n$, $g_c(n) = \log(g_{c-1}(n))/(1 - 2/p)$. We start the upper bound timeline with [IW05], since that is the first work which achieved an exponent of $1 - 2/p$ for n. For earlier works which achieved worse exponents for n, see [AMS99, CK04, Gan04a, Gan04b]. We note that [AMS99] initiated the problem and obtained an $O(n^{1-1/p}\epsilon^{-2} \log(M))$ bound in the insertion-only model (see also [BO12, BKSV14] for work in the insertion model).

Organization. We begin by introducing the formal definitions, complexity measures and information-theoretic tools used in our proofs. In Section 3 we prove the simultaneous communication lower bound on k-party Disjointness (Theorem 3.1). In Sections 5 we describe and analyze the "augmented" multiparty Disjointness communication problem we eventually use for the streaming application. In Section 4 we describe the aforementioned result of [LNW14], asserting that linear sketches have near-optimal space complexity in the turnstile streaming model. Section contains the reduction to frequency moments and the proof of the main streaming lower bound (Theorem 6.1).

2 Preliminaries

For $M \in \mathbb{N}$, we use the shorthands $[M]$ to denote the set $\{1, 2, \ldots, M\}$, and $[-M, M]^n$ to denote the set $\{-M, \ldots, 0, \ldots, M\}^n$. We typically use bold capital letters for random variables, and calligraphic letters to sets (for example, $\mathbf{X} \in \mathcal{X}$ represents a random variable with support $\mathcal{X}$). For a random variable $\mathbf{X}$ and a probability distribution μ, we write $\mathbf{X} \sim \mu$ to denote a random variable distributed according to μ. In particular, we use the notation $\mathbf{X} \sim B(p)$ to denote a Bernoulli-distributed random variable, taking the value 1 with probability p and 0 with probability $1 - p$. We write $\mathbf{X} \perp \mathbf{Y}$ to denote statistical independence between two random variables. All logarithms are in base 2 unless otherwise stated.

We use the following distance measures in our arguments.

Definition 2.1 (Total Variation distance and Hellinger distance)**.** *The* Total Variation *distance between two probability distributions P, Q over the same universe $\mathcal{U}$ is $\Delta(P, Q) := \sup_A |P(A) - Q(A)|$, where A ranges over all measurable events in the probability space.*

The (squared) Hellinger distance *between P and Q is denoted as*

$$h^2(P,Q) = 1 - \sum_{x \in \mathcal{U}} \sqrt{P(x)Q(x)} = \frac{1}{2} \cdot \sum_{x \in \mathcal{U}} \left(\sqrt{P(x)} - \sqrt{Q(x)}\right)^2 .$$

By a slight abuse of notation, we sometimes use the above distance measures with random variables instead of their underlying distributions. For example, if A, B are two random variables in the joint probability space $p(a,b)$, then $\Delta(A,B) = \Delta(p(a), p(b))$, and $h(A,B) = h(p(a), p(b))$.

The following properties and relationships between distance measures will be used throughout the paper. For missing proofs see [BYJKS04b] (Appendix A) and references therein.

Fact 2.2 (Product structure of Hellinger distance). *Let $P := P_1, \ldots, P_t$, $Q := Q_1, \ldots, Q_t$ be two product distributions over the same universe, (i.e., $P(x) = \prod_i P_i(x_i)$, $Q(x) = \prod_i Q_i(x_i)$). Then $h^2(P,Q) = 1 - \prod_{i=1}^t \left(1 - h^2(P_i, Q_i)\right)$.*

Lemma 2.3 (Hellinger vs. Total Variation). *For any two distributions P, Q it holds that*

$$h^2(P,Q) \le \Delta(P,Q) \le h(P,Q) \cdot \sqrt{2 - h^2(P,Q)}.$$

Corollary 2.4. *If $\Delta(P,Q) \ge 1 - \alpha$, then $h^2(P,Q) \ge 1 - 2\sqrt{\alpha}$.*

Proof. Rearranging the RHS inequality of Lemma 2.3 and substituting $x := h^2(P,Q), C := \Delta(P,Q)$, we get the following quadratic equation: $x^2 - 2x + C^2 \le 0$. Solving this equation for x yields

$$h^2(P,Q) = x \ge 1 - \sqrt{1 - C^2} = 1 - \sqrt{(1+C)(1-C)} \ge 1 - 2\sqrt{1-C} \ge 1 - 2\sqrt{\alpha},$$

where the last two inequalities follow since $1 - \alpha \le C = \Delta(P,Q) \le 1$. □

We will need the following fact about the moments of sums of independent random variables (For a proof see [Lat97] Corollary 3).

Lemma 2.5 (Moments of sums of independent random variables). *Let $X_1, X_2, \ldots, X_n$ be independent non-negative random variables, and define $X := \sum_{i=1}^n X_i$, $\Delta_\ell(X) := \left(\sum_i \mathbb{E}[X_i^\ell]\right)^{1/\ell}$. Then for every $m > 1$,*

$$\left(\mathbb{E}[X^\ell]\right)^{1/\ell} \le K \cdot \frac{m}{\log m} \cdot \max\left\{\Delta_2(X), \Delta_m(X)\right\},$$

where $K > 0$ is a universal constant.

Lemma 2.6 (Chebychev inequality for higher moments). *For any $\lambda > 0$ and $m \ge 2$, it holds that* $\Pr\left[|X - \mathbb{E}[X]| > \lambda \cdot \sigma_m(X)\right] \le \frac{1}{\lambda^m}$, *where $\sigma_m(X) := (\mathbb{E}[|X - \mathbb{E}[X]|^m])^{\frac{1}{m}}$.*

Fact 2.7. *If $x \le \varepsilon$, then* $\log(1-x) \ge -\frac{x}{1-\varepsilon}$. *(The proof follows by the known inequality* $\log(1+y) \ge 1 - 1/y$*).*

2.1 Information Theory

We will use basic tools from information theory. For proofs and elaboration on the facts below we refer the reader to an excellent monograph by Cover and Thomas [CT91]. The *entropy* of a random variable $X \sim \mu$ with support $\mathcal{X}$ is defined as $H(X) := \sum_{x \in \mathcal{X}} \mu(x) \log(1/\mu(x))$. A special case used in this paper is the *binary entropy* $H(p) := p\log(1/p) + (1-p)\log(1/(1-p))$, for $p \in (0,1)$. Notice that for a Bernoulli random variable $X \sim B(p)$, $H(X) = H(p)$. The following fact can be proved directly.

Fact 2.8 (Binary entropy). $\forall p \in [0, 1/2]$, $H(p) \le p\log(e/p) \le 2p\log(1/p)$.

The (conditional) *Mutual Information* between two random variables A, B in the joint probability space $\mu(a,b,c)$ is

$$I_\mu(A;B|C) := H(A|C) - H(A|BC) = H(B|A) - H(B|AC).$$

When the distribution μ is clear from the context, we omit the subscript and simply write $I(A;B)$. One of the basic and most useful properties of mutual information is the chain rule:

Fact 2.9 (Chain Rule for Mutual Information)**.** *Let A, B, C, D be jointly distributed random variables. Then*

$$I(AB; C|D) = I(A; C|D) + I(B; C|AD).$$

The following lemma asserts that if a random variable $Y = g(X)$ allows one to reconstruct X with high probability, then Y must "consume" most of the entropy of X:

Lemma 2.10 (Fano's Inequality)**.** *Let X be a random variable chosen from domain $\mathcal{X}$ according to distribution μ_X, and Y be a random variable chosen from domain $\mathcal{Y}$ according to distribution μ_Y . Then for any reconstruction function $g : Y \longrightarrow X$ with error ε_g, it holds that $H(X|Y) \leq H(\varepsilon_g) + \varepsilon_g \log(|\mathcal{X}| - 1)$.*

Lemma 2.11. *Let A, B, C, D be jointly distributed random variables. If A and D are conditionally independent given C, then it holds that $I(A; B|C) \leq I(A; B|CD)$.*

Proof. We apply the chain rule twice. On one hand, we have

$$I(A; BD|C) = I(A; B|C) + I(A; D|CB) \geq I(A; B|C)$$

since mutual information is nonnegative. On the other hand,

$$I(A; BD|C) = I(A; D|C) + I(A; B|CD) = I(A; B|CD)$$

since $I(A; D|C) = 0$ by the independence assumption on A and D. Combining both equations completes the proof. □

2.2 Multiparty Communication and Information Complexity in the SMP Model

We use the framework of communication complexity in the Simultaneous Message-Passing model:

Definition 2.12 (Multiparty SMP Model)**.** *Let P be a k-ary relation with domain $\mathcal{X}^k := \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_k$ and range $\mathcal{Z}$. In the* SMP *communication model, k parties receive inputs $\mathbf{X}_1, \mathbf{X}_2, \ldots \mathbf{X}_k$, jointly distributed according to some prior distribution μ, and are allowed to share a public random tape R. Each of the players simultaneously sends a message $M_j(\mathbf{X}_j, R)$ to an external party called the referee* , and the referee needs to output an answer $v = v(M_1(\mathbf{X}_1, R), \ldots, M_k(\mathbf{X}_k, R), R)$ such that $v = 1$ iff $(\mathbf{X}_1, \ldots \mathbf{X}_k) \in P$.

The *communication cost* of an SMP protocol π is the sum of the (worst-case) lengths of its messages $\|\pi\| := \sum_{j \in [k]} |M_j|$. For a fixed error parameter $\delta > 0$, the *distributional* SMP *communication complexity* of a function f, denoted $\vec{\mathsf{D}}^\delta_\mu(f)$, is the communication complexity of the cheapest deterministic SMP protocol which computes f correctly with error at most δ under input distribution μ.

The *randomized* SMP *communication complexity* of f, denoted $\vec{\mathsf{R}}_\delta(f)$, denotes the communication of the cheapest (public-coin) randomized SMP protocol which computes f correctly with error at most δ on *any* input $x \in \mathcal{X}^k$, under the randomness of the protocol (R).

By Yao's minimax theorem, $\vec{\mathsf{R}}_\delta(f) = \max_\mu \vec{\mathsf{D}}^\delta_\mu(f)$, and therefore it suffices to prove our lower bound for some "hard" distribution μ in the distributional model.

Remark 2.13. *To facilitate our proof techniques, we will sometimes need to give the referee an auxiliary input as well. In the distributional model, this input is jointly distributed with the inputs of the k players. The referee's answer is then a function of the k messages he receives as well as his own input. As a convention, in our definitions we typically ignore this artificial feature of the model, and include it implicitly.*

We also use the notion of (external) Information Cost of a communication protocol. Informally, the external information cost of an SMP protocol is the average amount of (Shannon) information the referee learns about the k player's inputs, by observing the transcript of the messages he receives. More formally, the *external information cost* of a protocol π with respect to inputs $\mathbf{X}_1, \ldots \mathbf{X}_k \sim \mu$ is defined as

$$\mathsf{IC}_\mu(\pi) := I_\mu(\pi; \mathbf{X}_1, \ldots \mathbf{X}_k).$$

Again, when the distribution μ is clear from the context, we omit the subscript and simply write $I(\pi; \mathbf{X}_1, \ldots \mathbf{X}_k)$. The *external information complexity of f* under μ is the least amount of information the players need to disclose to the referee about their inputs under μ, if their SMP protocol is required to solve f *on every input* with high probability:

$$\mathsf{IC}^\delta_\mu(f) := inf_{\pi:\, \forall\ (x_1, \ldots, x_k) \in \mathcal{X}^k\ \ \Pr_\pi[\pi(x_1, \ldots, x_k) \neq f(x_1, \ldots, x_k)] \leq \delta}\ \mathsf{IC}_\mu(\pi).$$

Remark 2.14. *(a) The requirement in the definition above that π is correct everywhere* , i.e., even outside the support of the distribution μ, is crucial: Our lower bounds will rely on analyzing the information cost of protocols under "trivial" distributions, and the only reason these lower bounds will be meaningful (in particular, non-zero) is that these protocols are required to succeed uniformly. (b) We remark that, unlike communication complexity, the usage of private randomness may be crucial to achieve low information cost, and therefore we assume π is randomized even against a fixed prior distribution μ.

Since one bit of communication can never reveal more than one bit of information, the external information cost of a protocol is always upper bounded by its communication:

Fact 2.15. *For any (k-party) communication protocol π and any distribution μ , $\|\pi\| \geq \mathsf{IC}_{\mu}(\pi)$.*

A special class of communication protocols are protocols in which players are restricted to use the same function when sending their messages to the referee. This class will be relevant to our streaming application (Theorem 6.1).

Definition 2.16 (Symmetric SMP protocols)**.** *A k-party* SMP *protocol π is called symmetric* if for any fixed input $\mathbf{X} = x$ and fixing of the public randomness $R = r$,

$$M_1(x, r) = M_2(x, r) = \ldots = M_k(x, r).$$

For a function f, we denote the distributional and randomized communication complexity of f with respect to symmetric SMP protocols by by $\vec{\mathsf{D}}^{\mathsf{SYM},\delta}_{\mu}(f)$ and $\vec{\mathsf{R}}^{\mathsf{SYM}}_{\delta}(f)$. Similarly, we denote by $\mathsf{IC}^{\mathsf{SYM},\delta}_{\mu}(f)$ the (external) information complexity of f with respect to symmetric SMP protocols.

3 Multiparty SMP Complexity of Set-Disjointness

In this section we prove our lower bound on the SMP communication complexity of the k-party Set-Disjointness function. We will prove the following theorem.

Theorem 3.1 (SMP complexity of multiparty Set-Disjointness)**.** *For any $\delta \geq n \cdot 2^{-k}$,*

$$\vec{\mathsf{R}}_{\delta}(\mathsf{Disj}^n_k) \geq \Omega\left(n \cdot \frac{\min\{\log(1/\delta), \log k\}}{k}\right).$$

$$\vec{\mathsf{R}}^{\mathsf{SYM}}_{\delta}(\mathsf{Disj}^n_k) \geq \Omega\left(n \cdot \min\left\{\frac{\log(1/\delta)}{k}, \log k\right\}\right).$$

Recall the k-party Set-Disjointness problem is defined as follows:

Definition 3.2 (Disj^n_k)**.** *Denote by Disj^n_k the multiparty Set-Disjointness problem in which k players each receive an n-dimensional input vector $\mathbf{X}_j = \{\mathbf{X}_{j,i}\}_{i=1}^n$ (where $\mathbf{X}_{j,i} \in \{0, 1\}$). By the end of the protocol, the referee needs to distinguish between the following cases:*

- **(The "NO" case)** $\forall \quad i \in [n], \quad \sum_j \mathbf{X}_{j,i} < k$, or
- **(The "YES" case)** $\exists \quad i \in [n]$ for which $\sum_j \mathbf{X}_{j,i} = k$.

Denote $\mathsf{AND}_k(x_1, x_2, \ldots, x_k) := \bigwedge_{j=1}^k x_j$. Note that

$$\overline{\mathsf{Disj}^n_k}(\mathbf{X}_1, \ldots, \mathbf{X}_k) = \bigvee_{i=1}^n \mathsf{AND}_k\left(\mathbf{X}_{1,i}, \ldots, \mathbf{X}_{k,i}\right).$$

We start by defining a "hard" distribution for Disj^n_k which still satisfies the promise (gap) required for our streaming application. Consider the distribution η on n-bit string inputs, defined by the following process.

The distribution η:

- For each $i \in [n], j \in [k]$ set $\mathbf{X}_{j,i} \sim B(1/k)$, independently at random.
- Pick a uniformly random coordinate $I \in_R [n]$.
- Pick $Z \in_R \{0,1\}$. If $Z = 1$, set all the values $\mathbf{X}_{j,I}$ to 1, for all $j \in [k]$ (If $Z = 0$, keep all coordinates as before.).
- The referee receives the index I (this feature will only be used in Section 6).

Denote by η_0 the distribution of $\eta \mid$ "$Z = 0$", and by μ_0 the projection of η_0 on a single coordinate (this is well defined since the distribution over all coordinates is i.i.d). In particular, notice that $\eta_0 = \mu_0^n$ is a product distribution, and for every $i \in [n]$, $\Pr_{\mu_0}[\mathbf{X}_{i,j} = 1 \text{ for all } j \in [k]] = (1/k)^k$. Thus, by a union bound over all n coordinates and our assumption on δ,

$$\Pr_{\mu_0^n}[\mathsf{Disj}_k^n(\mathbf{X}_1, \ldots, \mathbf{X}_k)] \le n \cdot (1/k)^k \le n \cdot 2^{-k} \le \delta. \tag{1}$$

Remark 3.3. *Notice that the "NO" distribution η_0 contains (w.h.p) coordinates $i \in [n]$ for which $\gg 1$ players (in fact, $\Omega(\log n)$ of them) possess the i'th coordinate. This feature is a by-product of the* product *structure of η_0, which will be crucial to our construction and analysis. To best of our knowledge, this is the first paper to show that distributions with such property (where* disjoint *instances in the support have $\omega(1)$ overlapping items in a coordinate, instead of just 1) are still powerful enough to prove lower bounds on the frequency moments problem.*

3.1 Direct sum and the SMP complexity of AND_k

To prove Theorem 3.1, we first use a direct sum argument, asserting that under product distributions, solving set disjointness is essentially equivalent to solving n copies of the 1-bit AND_k function. The following direct sum argument is well known (See e.g., [BYJKS04a]):

Lemma 3.4 (Direct sum for Disj_k^n). $\forall\, \delta \ge n \cdot 2^{-k}$, $\mathsf{IC}_{\eta_0}^{\delta}(\mathsf{Disj}_k^n) \ge n \cdot \mathsf{IC}_{\mu_0}^{2\delta}(\mathsf{AND}_k)$.

We defer the proof of this claim to Section A of the appendix. With Claim 3.4 in hand, it suffices to prove that any (randomized) SMP protocol solving AND_k with error at most δ, must have a large information cost *under* μ_0 . This is the content of the next theorem, which is one of our central technical contributions.

Theorem 3.5. *For every $\delta > 0$,*

$$\mathsf{IC}_{\mu_0}^{\delta}(\mathsf{AND}_k) \ge \Omega\left(\min\left\{\frac{\log 1/\delta}{k}, \frac{\log k}{k}\right\}\right)$$

$$\mathsf{IC}_{\mu_0}^{\mathsf{SYM},\delta}(\mathsf{AND}_k) \ge \Omega\left(\min\left\{\frac{\log 1/\delta}{k}, \log k\right\}\right).$$

Proof. Let π be a (randomized) SMP protocol which solves $\mathsf{AND}_k(\mathbf{X}_1, \ldots, \mathbf{X}_k)$ for all inputs in $\{0,1\}^k$ with success probability at least $1 - \delta$. For the rest of the analysis, we fix the public randomness of the protocol. Indeed, proving the lower bound for every fixing of the tape suffices as the chain rule for mutual information implies $\mathsf{IC}_{\mu_0}(\pi) = \mathbb{E}_R[\mathsf{IC}_{\mu_0}(\pi_R)]$. For each player $j \in [k]$, let M_j denote the transcript of player j's message, and let $M_0^j := M_j |$"$\mathbf{X}_j = 0$", $M_1^j := M_j |$"$\mathbf{X}_j = 1$" (note that if π is further a symmetric protocol, then M_0^j and M_1^j are the same for every player $j \in [k]$). Since the $\mathbf{X}_j$'s are independent under μ_0, and therefore so are the messages M_j, the chain rule implies that $\mathsf{IC}_{\mu_0}(\pi) = \sum_{j=1}^k I(M_j; \mathbf{X}_j)$. We shall argue that $\sum_{j=1}^k I(M_j; \mathbf{X}_j) \ge \Omega\left(\frac{\log 1/\delta}{k}\ ,\ \frac{\log k}{k}\right)$, and if π is further a symmetric protocol, then $\sum_{j=1}^k I(M_j; \mathbf{X}_j) \ge \Omega\left(\frac{\log 1/\delta}{k}\ ,\ \log k\right)$. To this end, let us denote by

$$h^2(M_1^j, M_0^j) := 1 - z_j$$

the (squared) Hellinger distance between player j's message distributions in both cases. There are two cases: if there is a player j for which $z_j = 0$, then $h^2(M_1^j, M_0^j) = 1$, which means that $I(M_j; \mathbf{X}_j) = H(\mathbf{X}_j) = H(1/k) = \Omega(\log(k)/k)$ and thus $\mathsf{IC}_{\mu_0}^{\delta}(\mathsf{AND}_k) \ge \Omega(\log(k)/k)$. Furthermore, if π is symmetric, then $z_1 = z_2 = \ldots = z_j$, which in this case implies by the same reasoning that $I(M_j; \mathbf{X}_j) = \Omega(\log(k)/k)$ for *all* players $j \in [k]$, and thus $\mathsf{IC}^{\mathsf{SYM},\delta}_{\mu_0}(\mathsf{AND}_k) \geq \Omega(\log k)$, as desired.

We may henceforth assume that all z_j's are non-zero, and the rest of the analysis applies for general (not necessarily symmetric) SMP protocols. To this end, let us introduce one final notation: For a *fixed* input $\mathbf{X}_j$, let $M_j^{\oplus t}$ denote (the concatenation of) t *independent* copies of $M_j|\mathbf{X}_j$ (so $M_j^{\oplus t} = (M_0^j)^t$ whenever $\mathbf{X}_j = 0$ and $M_j^{\oplus t} = (M_1^j)^t$ whenever $\mathbf{X}_j = 1$). By the conditional independence of the t copies of M_j (conditioned on $\mathbf{X}_j$) and the product structure of the Hellinger distance (Fact 2.2 in Section 2), we have that for each $j \in [k]$, the total variation distance between the t-fold message copies in the "YES" and "NO" cases is at least

$$\Delta\Big((M_1^j)^t, (M_0^j)^t\Big) \geq h^2\Big((M_1^j)^t, (M_0^j)^t\Big) = 1 - (z_j)^t, \tag{2}$$

where the first inequality follows from Lemma 2.3. Set $t_j = O(\log k / \log(1/z_j))$ (note that this is well defined as we assumed $z_j \neq 0$). Thus, for each player $j \in [k]$,

$$\Delta\Big((M_1^j)^{t_j}, (M_0^j)^t\Big) \geq 1 - \frac{1}{10k}. \tag{3}$$

Equation (3) implies that the error probability of the MLE predictor[1] for predicting $\mathbf{X}_j$ given $M_j^{\oplus t_j}$ is at most $\varepsilon := 1/(10k)$. Therefore, Fano's inequality (Lemma 2.10) and the data processing inequality together imply that

$$\forall \;\; j \in [k], \quad I(M_j^{\oplus t_j}; \mathbf{X}_j) \geq H(\mathbf{X}_j) - H(\varepsilon) \geq H\left(\frac{1}{k}\right) - H\left(\frac{1}{10k}\right) \geq \Omega\left(\frac{\log k}{k}\right), \tag{4}$$

since $\mathbf{X}_j \sim B(1/k)$ under μ_0, and $H(1/(10k)) \leq \frac{2}{10k}\log(10k) \leq \frac{4}{5}k\log(k)$ by Fact 2.8.

Now, by the chain rule for mutual information (Fact 2.9) we know that

$$I(M_j^{\oplus t_j}; \mathbf{X}_j) = \sum_{s=1}^{t_j} I((M_j)_s; \mathbf{X}_j | (M_j)_{<s}) \leq \sum_{s=1}^{t_j} I((M_j)_s; \mathbf{X}_j), \tag{5}$$

where the last inequality follows from Fact 2.11, as the messages $(M_j)_s$ and $(M_j)_{<s}$ are independent conditioned on $\mathbf{X}_i$ (by construction). Notice that $(M_j)_s \sim M_j$ for all $s \in [t]$, as all the messages are equally distributed conditioned on $\mathbf{X}_j$. Combining equations (4) and (5) therefore implies

$$I(M_j; \mathbf{X}_j) \geq \Omega\left(\frac{\log k}{k \cdot t_j}\right) \geq \Omega\left(\frac{\log(1/z_j)}{k}\right), \tag{6}$$

recalling that $t_j = O(\log k / \log(1/z_j))$. Since (6) holds for any player $j \in [k]$, we have

$$\sum_{j=1}^{k} I(M_j; \mathbf{X}_j) \geq \Omega\left(\frac{1}{k} \cdot \sum_{j=1}^{k} \log\left(\frac{1}{z_j}\right)\right). \tag{7}$$

We finish the proof by showing that

$$\sum_{j=1}^{k} \log\left(\frac{1}{z_j}\right) \geq \Omega(\log(1/\delta)). \tag{8}$$

To this end, we first claim that the correctness of π implies that the total variation distance between the transcript distributions of π on the input 0^k and on the input 1^k must be large (notice that below we crucially use the fact that our information complexity definition requires the protocol to be correct on all inputs, so in particular, a δ-error protocol must distinguish with comparable error, between "YES" and "NO" inputs):

[1] That is, the predictor which given $M_j^{\oplus t} = m$, outputs $Y := argmax_{x \in \{0,1\}} \Pr[(M_x^j)^t = m]$.

Proposition 3.6. $\Delta(\pi(0^k), \pi(1^k)) \geq 1 - 2\delta$.

Proof. Let $\mathcal{Y}$ be the set of transcripts τ for which $\pi(\tau) = \mathsf{AND}_k(1^k) = 1$. By the correctness assumption, $\Pr[\pi(1^k) \in \mathcal{Y}] \geq 1 - \delta$, and $\Pr[\pi(0^k) \in \mathcal{Y}] \leq \delta$, so the above follows by definition of the total variation distance. □

Since μ_0 is a product distribution (the $\mathbf{X}_j$'s are i.i.d), it holds that $\pi(0^k) = \times_{j=1}^k M_0^j$, and $\pi(1^k) = \times_{j=1}^k M_1^j$. Therefore, recalling that $z_j := 1 - h^2(M_0^j, M_1^j)$, the product structure of the Hellinger distance (Fact 2.2) implies

$$1 - \Pi_{j=1}^k z_j = 1 - \Pi_{j=1}^k (1 - h^2(M_0^j, M_1^j)) = h^2(\pi(0^k), \pi(1^k)) \geq 1 - 4\sqrt{\delta} \tag{9}$$

where the last transition follows from the combination of Proposition 3.6 with Corollary 2.4 (taken with $\alpha = 2\delta$). Rearranging (9), we get $\Pi_{j=1}^k z_j \leq 4\sqrt{\delta}$, or equivalently,

$$\sum_{j=1}^k \log\left(\frac{1}{z_j}\right) \geq \frac{1}{2}\log\left(\frac{1}{\delta}\right) - 2 = \Omega\left(\log 1/\delta\right), \tag{10}$$

as desired. Combining equations (8) and (7), we conclude that $\mathsf{IC}_{\mu_0}(\pi) \geq \Omega\left(\frac{\log 1/\delta}{k}\right)$, which completes the proof of Theorem 3.5.

□

Since communication is always lower bounded by information (Fact 2.15), combining Theorem 3.5 and Claim 3.4 directly implies Theorem 3.1:

Corollary 3.7. *For any $\delta \geq n \cdot 2^{-k}$,*

$$\vec{\mathsf{R}}_\delta(\mathsf{Disj}_k^n) \geq \Omega\left(n \cdot \min\left\{\frac{\log 1/\delta}{k}, \frac{\log k}{k}\right\}\right) \quad ,$$

$$\vec{\mathsf{R}}_\delta^{\mathsf{SYM}}(\mathsf{Disj}_k^n) \geq \Omega\left(n \cdot \min\left\{\frac{\log 1/\delta}{k}, \log k\right\}\right).$$

4 The Augmented Disj_k^n Problem

In this section we define the multiparty communication problem which we use as a proxy for our streaming application (Theorem 6.1). This communication problem is constructed using a fairly standard hardness-amplification technique ("augmentation") of the k-party Disjointness problem, in a similar fashion to the work of [LW13] (who used this technique for the L_∞ communication problem).

Definition 4.1 ($\mathsf{Aug\text{-}Disj}(r, k, \delta)$). $\mathsf{Aug\text{-}Disj}(r, k, \delta)$ *is the following k-party communication problem: The players receive r instances of* Disj_k^n*:*

$$(\mathbf{X}_1^1, \ldots, \mathbf{X}_k^1), (\mathbf{X}_1^2, \ldots, \mathbf{X}_k^2), \ldots, (\mathbf{X}_1^r, \ldots, \mathbf{X}_k^r)$$

In addition, the referee receives an index $T \in [r]$ which is unknown to the players, along with the last $(r - T)$ inputs $\{(\mathbf{X}_1^\ell, \ldots, \mathbf{X}_k^\ell)\}_{\ell=T+1}^r$. By the end of the protocol, the referee should output the answer to the T'th instance, i.e, the players need to solve $\mathsf{Disj}_k^n(\mathbf{X}_1^T, \ldots, \mathbf{X}_k^T)$ with probability $1 - \delta$.

For convenience, we henceforth denote $(\mathbf{X}_1^{>t}, \ldots, \mathbf{X}_k^{>t}) := \{(\mathbf{X}_1^\ell, \ldots, \mathbf{X}_k^\ell)\}_{\ell=t+1}^r$. The tuples $\left(\mathbf{X}_1^{<t}, \ldots, \mathbf{X}_k^{<t}\right)$, $(\mathbf{X}_1^{\{-t\}}, \ldots, \mathbf{X}_k^{\{-t\}})$ are defined analogously.

We now define a "hard" distribution ν for $\mathsf{Aug\text{-}Disj}(r, k, \delta)$. To this end, recall the distributions η, η_0 for Disj_k^n from Section 3. The index $T \in [r]$ is chosen independently and uniformly at random. All copies but the T'th copy are independently chosen according to the "NO" distribution for Disj_k^n, i.e., $\left(\mathbf{X}_1^{\{-T\}}, \ldots, \mathbf{X}_k^{\{-T\}}\right) \sim \eta_0^{r-1}$, while $(\mathbf{X}_1^T, \ldots, \mathbf{X}_k^T) \sim \eta$. The next lemma asserts that the the r-augmented Disjointness problem under the distribution ν is r times harder than solving a single instance Disj_k^n under η.

Lemma 4.2 (Direct Sum for $\mathsf{Aug\text{-}Disj}(r,k,\delta)$)**.**

$$\vec{\mathsf{R}}_\delta(\mathsf{Aug\text{-}Disj}(r,k,\delta)) \geq r \cdot \mathsf{IC}^\delta_{\eta_0}(\mathsf{Disj}^n_k).$$

Proof. The proof is essentially the same as that of Claim 3.4, using a standard "embedding" argument: Let Π be a protocol for $\mathsf{Aug\text{-}Disj}(r,k,\delta)$ under ν, such that $\|\Pi\| = \vec{\mathsf{R}}_\delta(\mathsf{Aug\text{-}Disj}(r,k,\delta))$. The k players will use public randomness to sample a random $t \in_R [r]$ along with $(r-t)$ "dummy" inputs $(\mathbf{X}_1^{>t},\ldots,\mathbf{X}_k^{>t})$ where each copy is independently drawn from η_0, and "embed" their inputs $(x_1,\ldots,x_k) \sim \eta$ (to Disj^n_k) to the t'th coordinate of Π, having the referee set $T = t$. Since in the augmented problem player's inputs to each of the r copies are independent, and since η_0 is a product distribution, they can use *private randomness* to "fill in" their inputs to the rest of the coordinates $(\mathbf{X}_1^{<t},\ldots,\mathbf{X}_k^{<t})$ (for a formal argument see the essentially identical proof of Claim 3.4). This process defines a legal input $\{(\mathbf{X}_1^\ell,\ldots,\mathbf{X}_k^\ell)\}_{\ell=1}^r \sim \nu$ for Π, and so the players can now run Π on this input and output its answer. Call this protocol π. By the premise $\left(\mathbf{X}_1^{\{-T\}},\ldots,\mathbf{X}_k^{\{-T\}} \sim \eta_0^{r-1}\right)$, π outputs the correct answer to the t'th copy with probability at least $1-\delta$. Furthermore, we may analyze the information complexity of Π under the distribution η_0^r (notice that $\vec{\mathsf{D}}_{\eta_0^r}(\mathsf{Aug\text{-}Disj}(r,k,\delta)) = 0$ trivially, but we are analyzing the information cost of Π which must be correct with probability $1-\delta$ over all inputs !). Similar to the argument in (22), we have

$$\begin{aligned}
&\mathsf{IC}^\delta_{\eta_0}(\mathsf{Disj}^n_k) \leq \mathsf{IC}_{\eta_0}(\pi) = I_{\eta_0}(\pi; x_1,\ldots x_k)\\
&= \mathbb{E}_{t\in_R[r]}\left[I_{\eta_0}(\Pi;\mathbf{X}_1^t,\ldots\mathbf{X}_k^t)\right]\\
&\leq \mathbb{E}_{t\in_R[r]}\left[I_{\eta_0}(\Pi;\mathbf{X}_1^t,\ldots\mathbf{X}_k^t\ |\mathbf{X}_1^{>t},\ldots,\mathbf{X}_k^{>t})\right] \quad \text{(By Lemma 2.11)}\\
&= \frac{1}{r}\sum_{t=1}^r I_{\eta_0}(\Pi;\mathbf{X}_1^t,\ldots\mathbf{X}_k^t\ |\mathbf{X}_1^{>t},\ldots,\mathbf{X}_k^{>t})\\
&= \frac{1}{r}\cdot I_{\eta_0^r}(\Pi;(\mathbf{X}_1^1,\ldots,\mathbf{X}_k^1),(\mathbf{X}_1^2,\ldots,\mathbf{X}_k^2),\ldots,(\mathbf{X}_1^r,\ldots,\mathbf{X}_k^r))\\
&\leq \frac{\|\Pi\|}{r} = \vec{\mathsf{R}}_\delta(\mathsf{Aug\text{-}Disj}(r,k,\delta)),
\end{aligned}$$

where the last equality follows from the chain rule for mutual information.

□

5 Path-Independent Stream Automata [LNW14]

As mentioned in the introduction, a central fact which facilitates our lower bound is the recent result of [LNW14], asserting that in the turnstile streaming model, linear sketching algorithms achieve optimal space complexity, up to a logarithmic factor. Since we cannot even afford losing a $\log n$ factor in our lower bound, we use the following intermediate result of [LNW14], which shows that oblivious streaming algorithms are optimal up to a *constant* factor. The following exposition largely follows that of [LNW14], from which a number of definitions also occur in the earlier work of [Gan08].

The work of [LNW14] considers problems in which the input is a vector $x \in \mathbb{Z}^n$ represented as a data stream $\sigma = (\sigma_1,\sigma_2,\ldots)$ in which each element σ_i belongs to $\Sigma = \{e_1,\ldots,e_n,-e_i,\ldots,-e_n\}$ (where the e_i's are canonical basis vectors) such that $\sum_i \sigma_i = x$. We write $x = \operatorname{freq}\sigma$.

Definition 5.1 (Deterministic stream automata)**.** *A deterministic stream automaton $\mathcal{A}$ is a deterministic Turing machine that uses two tapes, a one-way (unidirectional) read-only input tape and a (bidirectional) two way work-tape. The input tape contains the input stream σ. After processing its input, the automaton writes an output, denoted by $\phi_{\mathcal{A}}(\sigma)$, on the work-tape.*

A configuration of a stream automaton $\mathcal{A}$ is modeled as a triple (q,h,w), where, q is a state of the finite control, h the current head position of the work-tape and w the content of the work-tape. The set of configurations of a stream automaton $\mathcal{A}$ that are reachable from the initial configuration o on some input stream is denoted by $C(\mathcal{A})$. A stream automaton is a tuple $(n,C,o,\oplus,\phi)$, where n specifies the dimension of the underlying vector, $\oplus : C\times\Sigma\to C$ is the configuration transition function, o is the initial position of the automaton and $\phi : C \to \mathbb{Z}^{p(n)}$ is the output function and $p(n)$ is the dimension of the output. For a stream σ we also write $\phi(o \oplus \sigma)$ as $\phi(\sigma)$ for simplicity.

The set of configurations of an automaton $\mathcal{A}$ that is reachable from the origin o for some input stream σ with $\|\operatorname{freq}\sigma\|_\infty \leq m$ is denoted by $C(\mathcal{A}, m)$. The space of the automaton $\mathcal{A}$ with stream parameter m is defined as $S(A, m) = \log |C(\mathcal{A}, m)|$. An algorithm is said to be a correct randomized algorithm with error probability δ if for any fixed stream σ with $\|\operatorname{freq}\sigma\|_\infty \leq m$, with probability at least $1 - \delta$ the algorithm outputs the correct answer to a relation P for the underlying vector x represented by σ. Note that the streaming algorithm should be correct even if for a substream σ' of σ we have $\|\operatorname{freq}\sigma'\|_\infty > m$, provided that $\|\operatorname{freq}\sigma\|_\infty \leq m$. In this case we say $\mathcal{A}$ solves P on $\mathbb{Z}^n_{|m|}$.

Definition 5.2 (Path-independent stream automata). *A stream automaton $\mathcal{A}$ is said to be* path independent *(*PIA *) if for each configuration s and input stream σ, $s \oplus \sigma$ is dependent only on* $\operatorname{freq}\sigma$ *and s.*

Suppose that $\mathcal{A}$ is a path independent automaton. We can define a function $+ : \mathbb{Z}^n \times C \to C$ as $x + a = a \oplus \sigma$, where $\operatorname{freq}\sigma = x$. Since $\mathcal{A}$ is a path independent automaton, the function $+$ is well-defined. In [Gan08] it is proved that

Theorem 5.3. *Suppose that $\mathcal{A}$ is a path independent automaton with initial configuration o. Let $M = \{x \in \mathbb{Z}^n : x + o = 0 + o\}$, then M is a submodule of $\mathbb{Z}^n$, and the mapping $x + M \mapsto x + o$ is a set isomorphism between $\mathbb{Z}^n/M$ and the set of reachable configurations $\{x + o : x \in \mathbb{Z}^n\}$.*

Definition 5.4 (Randomized stream automata). *A randomized stream automaton is a deterministic stream automaton with one additional tape for the random bits. The random bit string R is initialized on the random bit tape before any input record is read; thereafter the random bit string is used in a two way read-only manner. The rest of the execution proceeds as in a deterministic stream automaton.*

A randomized stream automaton $\mathcal{A}$ is said to be *path-independent* if for each randomness R the deterministic instance $\mathcal{A}_R$ is path-independent. The space complexity of $\mathcal{A}$ is defined to be

$$S(\mathcal{A}, m) = \max_R \{|R| + S(\mathcal{A}_R, m)\}.$$

Theorem 5.5 ([LNW14] Theorems 9 and 10). *Suppose that a randomized algorithm $\mathcal{A}$ solves a relation P on any stream σ with probability at least $1-\delta$. There exists a randomized path-independent automaton (*PIA *) $\mathcal{B}$ which solves P on $\mathbb{Z}^n_{|m|}$ with probability at least $1 - 7\delta$ such that $S(\mathcal{B}, m) \leq S(\mathcal{A}, m) + O(\log n + \log\log m + \log \frac{1}{\delta})$. Further, the number of random bits used by the algorithm is is $O(\log 1/\delta + \log n + \log\log m)$.*

Here we record the corollary of Theorem 5.5 that will be used in the proof of our main result (Theorem 6.1). To this end, we will need the following (refined) restatement of the SMP communication model used in our paper:

Definition 5.6. *Let $P(x_1, \dots, x_k)$ be a k-ary relation. In the public-coin* SMP *communication model, k players receive inputs $x_1, \dots, x_k \in \mathbb{Z}^n$ respectively, such that $x := \sum_j x_j \in Z^n_{|m|}$ (for some $m \in \mathbb{N}$). The players share a public random tape R of $O(\log 1/\delta + \log n + \log\log m)$ uniformly random bits. Each of the players simultaneously sends a message $M_j(x_j, R)$ to an external party called the referee* , and the referee outputs an answer $v = v(M_1(x_1, R), \dots, M_k(x_k, R), R)$, such that $\Pr_R[v = P(x_1, \dots, x_k)] \geq 1 - \delta$. Recall that the *symmetric* SMP communication complexity of P is $\vec{\mathsf{R}}^{\mathsf{SYM}}_\delta(P) := \min_{\pi \,:\, \pi \text{ is symmetric and } \delta\text{-solves } P} \sum_{j=1}^k |M_j(x_j, R)|$ where $|\cdot|$ denotes the worst-case length of the messages, over all choices of $x^1, \dots, x^s$ and R.

Corollary 5.7. *Let $P(x_1, \dots, x_k)$ be a relation such that $\vec{\mathsf{R}}^{\mathsf{SYM}}_\delta(P) = c$. Let $\mathcal{A}$ be a space-optimal streaming algorithm in the turnstile model from which the output of $\mathcal{A}$ on an input stream σ with underlying vector x, can be used to solve P with probability at least $1 - \delta$. Then the space complexity of $\mathcal{A}$ is at least c/k.*

Proof. By Theorem 5.5, we can assume that $\mathcal{A}$ is a randomized path-independent automaton using $O(\log 1/\delta + \log n + \log\log m)$ random bits. The players in the public-coin simultaneous model of communication can therefore use the public coin R to agree upon a deterministic path-independent automaton $\mathcal{B}$. Each player can run $\mathcal{B}$ on his/her local input vector x_j, and transmit the state of $\mathcal{B}$ to the referee. Notice that each player uses the same function to compute his message, and therefore this SMP protocol is also symmetric. By Theorem 5.3, the referee can associate these states with elements of the quotient group Z^n/M, where M is determined from the description of $\mathcal{B}$ (which is in turn determined by R), and perform arithmetic in Z^n/M to add up the states to obtain the result of the execution of $\mathcal{B}$ on the concatenation of streams $\sigma^1, \ldots, \sigma^k$, where σ^j is a stream generating x_j. It follows that $\mathcal{B}$ will be executed on σ with underlying vector x, and by hypothesis can be used to solve P with probability at least $1 - \delta$. As k times the space complexity of $\mathcal{B}$ is the communication cost, the corollary follows. $\square$

6 Frequency Moments

Let $x \in \mathbb{R}^n$ represent a data stream in turnstile streaming model. We say that an algorithm solves the (p, ε, δ)-Norm problem if its output v satisfies $v \in (1 \pm \varepsilon)\|x\|_p^p$ with probability at least $1 - \delta$. Our main result is as follows:

Theorem 6.1. *For any constant $p > 2$, there exists an absolute constant $\alpha > 1$ such that for any $\varepsilon > n^{-\Omega(1)}$ and $\delta \geq 2^{-o(n^{1/p})}$, any randomized streaming algorithm that solves the (p, ε, δ)-Norm problem for $x \in [-M, M]^n$ where $M = \Omega(n^{\alpha/p})$, requires $\Omega\left(\varepsilon^{-2} \cdot n^{1-2/p}(\log M) \log 1/\delta\right)$ bits of space. In particular, the space complexity of any ε-approximate high-probability streaming algorithm (i.e, where $\delta = O(1/n)$) is at least $\Omega\left(\varepsilon^{-2} \cdot n^{1-2/p}(\log n)(\log M)\right)$.*

Proof. Let s be the space complexity of a space-optimal streaming algorithm that solves the (p, ε, δ)-Norm. By Theorem 5.5, there is a *path-independent* streaming algorithm $\mathcal{A}$ which solves $(p, \varepsilon, 7\delta)$-$\mathsf{Norm}$ using $s' = s + O(\log n + \log\log M + \log \frac{1}{\delta})$ bits of space. We will show that $\mathcal{A}$ can be used to (produce a symmetric SMP protocol) solving $\mathsf{Aug\text{-}Disj}(r, k, 18\delta)$ (on $\mathbb{Z}_M^n$), for $r = (1 - 1/\alpha)\log_{10} M, k = \Theta(\varepsilon \cdot n^{1/p})$, under the hard distribution ν. Corollary 5.7 then implies

$$s' \geq \frac{\vec{\mathsf{R}}_{\delta}^{\mathsf{SYM}}(\mathsf{Aug\text{-}Disj}(r, k, 18\delta))}{k} \geq \Omega\left(\frac{rn}{k} \cdot \min\left\{\frac{\log 1/\delta}{k}, \log k\right\}\right) =$$
$$\Omega\left(\min\left\{\frac{n^{1-\frac{2}{p}}(\log M)\log 1/\delta}{\varepsilon^2} \quad , \quad \frac{n^{1-\frac{1}{p}}\log n(\log M)}{\varepsilon}\right\}\right),$$

where the second inequality follows from Lemma 4.2 and Corollary 3.7 (as in our regime of parameters $\delta \geq n \cdot 2^{-k} = 2^{-\Omega(n^{1/p})}$), and the last transition follows by substituting the values of the parameters k, M and noting that $\log k = \Theta(\log n)$ in our regime. Since $s' = s + O(\log n + \log\log M + \log \frac{1}{\delta})$, the last equation implies that so long as $\delta \geq 2^{-o(n^{1/p})}$,

$$s \geq \Omega\left(\frac{n^{1-\frac{2}{p}}(\log M)\log 1/\delta}{\varepsilon^2}\right) \quad , \quad \text{as claimed.}$$

It therefore remains to prove that $\mathcal{A}$ can be used to solve $\mathsf{Aug\text{-}Disj}(r, k, 18\delta)$ under the input distribution ν. To this end, recall that in the $\mathsf{Aug\text{-}Disj}(r, k, \delta)$ problem under the distribution ν, k players receive r instances each, where all instances but a single random instance $t \in_R [r]$ are independently distributed according to η_0, while $(\mathbf{X}_1^t, \ldots, \mathbf{X}_k^t) \sim \eta$. The referee receives t along with $(\mathbf{X}_1^{>t}, \ldots, \mathbf{X}_k^{>t})$ and needs to solve $\mathsf{Disj}_k^n(\mathbf{X}_1^t, \ldots, \mathbf{X}_k^t)$, i.e., distinguish between the "NO" case and the "YES" case in definition 3.2. Recall that for every instance $(\mathbf{X}_1^\ell, \ldots, \mathbf{X}_k^\ell)$, the referee also receives the "spiked" coordinate $I^\ell \in [n]$ of this instance (see the definition of η, η_0 in Subsection 3.1). The players will use the PIA algorithm $\mathcal{A}$ to design the SMP protocol π for $\mathsf{Aug\text{-}Disj}$ described in Figure 1.

We now turn to analyze the correctness of π. For the rest of this analysis, we fix the value of the "special" coordinate $T = t$. Notice that the value v the referee computes in π corresponds to the p-norm of the stream (with underlying frequency vector) $z := (\mathbf{Y}_1^{\leq t}, \ldots, \mathbf{Y}_k^{\leq t}, \mathbf{C}^{\leq t})$. Furthermore, $\|v\|_\infty \leq \left(\sum_{\ell=1}^t \sum_{j=1}^k \mathbf{Y}_j^\ell\right) + C \leq \sum_{\ell=1}^r \sum_{j=1}^k 10^{\ell-1} + C \leq 10^r \cdot k + C \leq O(10^r \cdot n^{1/p}) \leq M$ for a sufficiently small constant $\alpha > 1$, by our assumption $k \leq O(n^{1/p}), C \leq O(10^r \cdot n^{1/p})$, and our assumption that $M = \Omega(n^{\alpha/p})$. Therefore, the correctness of the streaming algorithm $\mathcal{A}$ guarantees that the output v of the referee satisfies

$$\Pr\left[v \notin (1 \pm \varepsilon)\|z\|_p^p\right] \leq 7\delta. \tag{11}$$

Define $L_i := \sum_{j\in[k]} \sum_{\ell\in[t]} \mathbf{Y}_{j,i}^\ell + \mathbb{1}_I \cdot C$, where $\mathbb{1}_I$ is the indicator random variable for the event $I = i$. In this notation, $\|z\|_p^p = \sum_{i=1}^n (L_i)^p$. Recall that for any $i \neq I$, both in the "NO" and "YES" distributions, $\mathbf{X}_{j,i}^\ell \sim B(1/k)$ independently of each other, and in particular, these L_i's are independent random variables. We will need the following concentration bounds on the contribution of the L_i's:

The SMP protocol π

Input : $(\mathbf{X}_1^1, \dots, \mathbf{X}_k^1), (\mathbf{X}_1^2, \dots, \mathbf{X}_k^2), \dots, (\mathbf{X}_1^r, \dots, \mathbf{X}_k^r)$.

1. Set $\gamma \longleftarrow 4ep, \;\; C \longleftarrow 10^t \cdot \gamma \cdot n^{1/p}, \;\; \rho \longleftarrow (1+\varepsilon)(\mathcal{E}_p + (1+7\varepsilon)C^p)$ ($\mathcal{E}_p$ is defined below).
2. Each player $j \in [k]$ locally defines $\mathbf{Y}_j^\ell := 10^{\ell-1} \cdot \mathbf{X}_j^\ell \;\; \forall \ell \in [r]$, and generates the stream
$$\sigma_j := \mathbf{Y}_j^1, \dots, \mathbf{Y}_j^r$$
according to his input, and sends the referee the message $\mathcal{A}(\sigma_j)$.
3. The referee locally computes $A^{>t} := \sum_{j=1}^k \mathcal{A}(\mathbf{Y}_j^{t+1}, \mathbf{Y}_j^{t+2}, \dots, \mathbf{Y}_j^r)$ where the addition is over the quotient ring $\mathbb{Z}^n/M$ (and M is the module representing the kernel of the automaton[a] $\mathcal{A}$).
Notice that he can do so as he has t and $(\mathbf{X}_1^{>t}, \dots, \mathbf{X}_k^{>t})$.
4. The referee adds the value C to the "spiked" coordinate I^ℓ of the ℓ-th instance, for each $\ell \in [r]$. Let $\mathbf{C} := \mathbf{C}^1, \dots, \mathbf{C}^r$ denote the underlying stream representing this vector (notice that he can do so since he receives the "spiked" coordinate I^ℓ of each instance).
5. The referee adds up the messages he receive from the players, over the quotient ring $\mathbb{Z}^n/M$, and outputs 1 ("YES") iff
$v := \left(\sum_{j=1}^k \mathcal{A}(\sigma_j)\right) + \mathcal{A}(\mathbf{C}^{\le t}) - A^{>t} > \rho$.

[a] See Section 5 for the formal definitions and statement.

Figure 1: An SMP protocol for $\mathsf{Aug\text{-}Disj}(r, k, 18\delta)$ using the PIA $\mathcal{A}$

Claim 6.2 (Concentration bounds)**.** *It holds that:*

- $\mathbb{E}_{\eta_0^r}[\sum_{i \ne I}(L_i)^p] \le 2n \cdot 10^{tp}(2ep)^p$.
- *Let K be the universal constant from Lemma 2.5. Then for every $m \in \mathbb{N}$,*
$$\sigma_m\left(\sum_{i \ne I}(L_i)^p\right) := \left(\mathbb{E}\left[\left|\sum_{i \ne I}(L_i)^p - \mathbb{E}\left[\sum_{i \ne I}(L_i)^p\right]\right|^m\right]\right)^{\frac{1}{m}} \le n^{1/m} \cdot (4Kemp)^p \cdot 10^{tp}.$$
- *For every $m \le o(n^{1/p})$,* $\Pr_{\eta_0}[(L_I)^p \ge (1+7\varepsilon)C^p] \le \delta$.

We defer the proof of this technical claim to the end of this argument. In the following denote $\mathcal{E}_p := \mathbb{E}_{\eta_0}\left[\sum_{i \ne I}(L_i)^p\right]$ (note that δ can be computed by the referee as it is public knowledge). Applying the generalized Chebychev's inequality (Lemma 2.6) with $m = \log 1/\delta$ and $\lambda = 2$, the first two propositions of Claim 6.2 guarantee that

$$\Pr\left[\left|\sum_{i \ne I}(L_i)^p - \mathcal{E}_p\right| > \varepsilon \cdot C^p\right] = \Pr\left[\left|\sum_{i \ne I}(L_i)^p - \mathbb{E}\left[\sum_{i \ne I}(L_i)^p\right]\right| > \varepsilon \cdot C^p\right]$$
$$\le \Pr\left[\left|\sum_{i \ne I}(L_i)^p - \mathbb{E}\left[\sum_{i \ne I}(L_i)^p\right]\right| > 2 \cdot \sigma_m\left(\sum_{i \ne I}(L_i)^p\right)\right] \le 2^{-m} = 2^{-\log 1/\delta} < \delta, \tag{12}$$

where the first inequality is by the second proposition of Claim 6.2 (which implies that $2 \cdot \sigma_m\left(\sum_{i \ne I}(L_i)^p\right) \le \varepsilon C^p$ whenever $\varepsilon \ge \Omega(n^{1/m-1}) = n^{-\Omega(1)}$), and the second inequality holds by our assumption that $\delta > 2^{-o(n^{1/p})}$.

Combining (12) with the third proposition of Claim 6.2 implies

$$\Pr_{\eta_0}\left[\|z\|_p^p > \mathcal{E}_p + C^p(1+7\varepsilon)\right] = \Pr_{\eta_0}\left[\|z\|_p^p > (\mathcal{E}_p + \varepsilon C^p) + C^p(1+7\varepsilon)\right] \leq 2\delta.$$

Hence, by definition of the "threshold" $\rho := (1+\varepsilon)[\mathcal{E}_p + (1+7\varepsilon)C^p]$, we conclude by (11) that in the "NO" case,

$$\Pr_{\eta_0}[v > \rho] \leq \Pr\left[\|z\|_p^p > \mathcal{E}_p + (1+7\varepsilon)C^p\right] \leq 9\delta. \tag{13}$$

On the other hand, in the "YES" case, the coordinate I is such that $\mathbf{X}_{j,I}^{\ell} = 1$ for all $j \in [k], \ell \in [r]$. Setting $k = 128e\varepsilon \cdot n^{1/p} = \Theta(\varepsilon \cdot n^{1/p})$, the contribution of this coordinate to the p-norm of z is

$$\begin{aligned}
&\left(C + \sum_{\ell=1}^{t} 10^{\ell-1} \cdot k\right)^p \geq \left(10^t \cdot \gamma \cdot n^{1/p} + 10^t \cdot 128e\varepsilon \cdot n^{1/p}\right)^p = \\
&= n \cdot 10^{tp} \cdot \gamma^p \left(1 + 128\varepsilon/\gamma\right)^p \geq n \cdot 10^{tp} \cdot \gamma^p \cdot e^{\frac{128e\varepsilon p}{2\gamma}} \quad \text{(since } 128e\varepsilon/\gamma < 1/2\text{)} \\
&= n \cdot 10^{tp} \cdot \gamma^p \cdot e^{\frac{128e\varepsilon p}{8ep}} \geq n \cdot 10^{tp} \cdot \gamma^p \cdot (1+16\varepsilon) = (1+16\varepsilon)C^p.
\end{aligned}$$

Furthermore, (12) ensures that, except with probability δ, the contribution of all the rest coordinates ($i \neq I$) is at least $\mathcal{E}_p - \varepsilon C^p$, and thus in the "YES" case,

$$\Pr\left[\|z\|_p^p > \mathcal{E}_p + (1+15\varepsilon)C^p\right] = \Pr\left[\|z\|_p^p > (\mathcal{E}_p - \varepsilon C^p) + C^p(1+16\varepsilon)\right] \geq 1 - 2\delta$$

Finally, (11) implies that under the "YES" distribution,

$$\begin{aligned}
&\Pr[v > \rho] \\
&\geq \Pr\left[\|z\|_p^p > (1+\varepsilon)\rho\right] - 7\delta = \Pr\left[v > (1+\varepsilon)^2 \cdot (\mathcal{E}_p + (1+7\varepsilon)C^p)\right] - 7\delta \\
&\geq \Pr\left[\|z\|_p^p > \mathcal{E}_p + 3\varepsilon\mathcal{E}_p + (1+3\varepsilon)(1+7\varepsilon)C^p\right] - 7\delta \\
&\geq \Pr\left[\|z\|_p^p > \mathcal{E}_p + 3\varepsilon C^p + (1+11\varepsilon)C^p\right] - 7\delta \qquad \text{(since } \mathcal{E}_p \leq C^p\text{)} \\
&= \Pr\left[\|z\|_p^p > \mathcal{E}_p + (1+14\varepsilon)C^p\right] \geq 1 - 9\delta
\end{aligned} \tag{14}$$

We conclude from equations (13) and (14) that setting the threshold ρ guarantees that π is correct with probability at least $1 - 18\delta$, which completes the reduction and the proof of Theorem 6.1.

It remains to prove Claim 6.2.

Proof of Claim 6.2. **First proposition:** Since all coordinates except coordinate I are identically distributed, we can write $\mathbb{E}_{\eta_0}[\sum_{i \neq I}(L_i)^p] = (n-1) \cdot \mathbb{E}[(L_t)^p] + \mathbb{E}[(C + L_t)^p]$, where

$$L_t := \sum_{j \in [k]} \sum_{\ell \in [t]} 10^{\ell-1}\mathbf{X}_j^{\ell} \quad , \quad \text{and } \mathbf{X}_j^{\ell}\text{'s are i.i.d } B(1/k).$$

We first prove the following lemma, which upper bounds the p'th moment of a single coordinate ($i \neq I$) in a "NO" instance. Though it is a spacial case of Lemma 2.5, for completeness we present an elementary (yet slightly weaker) proof that will be sufficient in our applications.

Lemma 6.3. *For every* $p \geq 1$, $\quad \mathbb{E}[(L_t)^p] \leq (2ep)^p \cdot 10^{tp}$.

Proof. We shall show by induction on t, that there exists a function $f(p) \leq (2ep)^p$ for which

$$\mathbb{E}[(L_t)^p] \leq f(p) \cdot 10^{tp}. \tag{15}$$

Indeed, define the function $f(p)$ recursively by the formula: $f(p+1) := (ep)^p + f(p)$. It follows that

$$f(p) = (ep)^p + (e(p-1))^{p-1} + (e(p-2))^{p-2} + \ldots + 1 \leq (2ep)^p,$$

as desired. The proof for the base case ($t = 1$) is very similar to the general case, so we postpone it to the end of the proof. Suppose (15) is true for for all integers up to t. We shall show that

$$\mathbb{E}[(L_{t+1})^p] \le f(p+1) \cdot 10^{(t+1)p}. \tag{16}$$

To this end, we may write $L_{t+1} := \Delta_{t+1} + L_t$, where $\Delta_{t+1} := 10^t \cdot \sum_{j\in[k]} \mathbf{X}_j^{t+1}$. We first bound $\mathbb{E}[(\Delta_{t+1})^p]$:

$$\begin{aligned}
&\mathbb{E}[(\Delta_{t+1})^p] = (10^{tp}) \cdot \mathbb{E}\left[\left(\sum_{j\in[k]} \mathbf{X}_j^{t+1}\right)^p\right] \\
&\le 10^{tp} \cdot \sum_{r=1}^{p} \binom{k}{r} \cdot p^r \cdot \mathbb{E}\left[\prod_{i=1}^{r} \mathbf{X}_{j_i}^{t+1}\right] \quad \text{(By the multinomial formula)} \\
&\le 10^{tp} \cdot \sum_{r=1}^{p} \left(\frac{ek}{r}\right)^r \cdot p^r \cdot \left(\frac{1}{k}\right)^r \quad \text{(By Stirling's approximation and in dependence of } \mathbf{X}_j^{t+1}\text{'s)} \\
&\le 10^{tp} \cdot \sum_{r=1}^{p} \left(\frac{ep}{r}\right)^r \le (ep)^p \cdot 10^{tp}. \qquad (17)
\end{aligned}$$

We therefore have

$$\begin{aligned}
&\mathbb{E}[(L_{t+1})^p] = \mathbb{E}[(\Delta_{t+1} + L_t)^p] \\
&\le 2^{p-1} \cdot (\mathbb{E}[(\Delta_{t+1})^p] + \mathbb{E}[(L_t)^p]) \\
&\le 2^{p-1} \cdot \left(\mathbb{E}[(\Delta_{t+1})^p] + f(p) \cdot 10^{tp}\right) \qquad \text{(by the inductive hypothesis)} \\
&\le 2^{p-1} \cdot 10^{tp} \left[(ep)^p + f(p)\right] \qquad \text{(by (17))} \\
&\le 10^{(t+1)p} \cdot [(2ep)^p + f(p)] \\
&= 10^{(t+1)p} \cdot f(p+1) \ ,
\end{aligned}$$

which finishes the proof of (16). For the base case ($t = 1$), we need to show that $\mathbb{E}[(L_1)^p] := \mathbb{E}[(\sum_{j\in[k]} \mathbf{X}_j^1)^p] \le 10^p \cdot f(p)$. Indeed, repeating essentially the same calculation as in (17), one obtains

$$\begin{aligned}
&\mathbb{E}[(L_1)^p] \le (2ep)^p = 2^p \cdot (ep)^p \le 10^p \cdot (e(p-1))^{p-1} \\
&\le 10^p \cdot \left[(e(p-1))^{p-1} + f(p-1)\right] = 10^p \cdot f(p).
\end{aligned}$$

This finishes the proof of (15), and therefore concludes the proof of Claim 6.3. □

Substituting the value of $C = 10^t \cdot \gamma \cdot n^{1/p}$, we conclude by Lemma 6.3 and (18) that

$$\begin{aligned}
&\mathbb{E}_{\eta_0}\left[\sum_{i\ne I} (L_i)^p\right] = (n-1) \cdot \mathbb{E}[(L_t)^p] + \mathbb{E}[(C + L_t)^p] \\
&\le (n-1) \cdot (2ep)^p \cdot 10^{tp} + 2C^p \\
&= (n-1) \cdot (2ep)^p \cdot 10^{tp} + 2n \cdot \gamma^p \cdot 10^{tp} \\
&\le n \cdot 10^{tp} \left(2\gamma^p + (2ep)^p\right).
\end{aligned}$$

Second proposition: To upper bound the m-th moment of $\sum_{i\ne I}(L_i)^p$, we note that $\sum_{i\ne I}(L_i)^p$ is a sum of independent random variables, and thus Lemma 2.5 implies that

$$\begin{aligned}
&\mathbb{E}\left[\left|\sum_{i\ne I} (L_i)^p\right|^m\right] \le \left(\frac{Km}{\log m}\right)^m \cdot \sum_{i\ne I} \mathbb{E}[(L_i)^{mp}] \\
&\le (n-1) \cdot (3eKm)^m \cdot (2emp)^{mp} \cdot 10^{tmp} \\
&\le n \cdot (4Kemp)^{mp} \cdot 10^{tmp},
\end{aligned}$$

where the second inequality follows again from Lemma 6.3, taken with $p := mp$. The second proposition of Lemma 6.2 now follows by raising both sides of the above inequality to the $1/m$ power.

Third proposition: We first upper bound the expected contribution of the I'th coordinate under η_0:

$$\begin{aligned}\mathbb{E}_{\eta_0}\left[(L_I)^p\right] = \mathbb{E}[(C+L_t)^p] = C^p \cdot \sum_{r=0}^{p} \binom{p}{r} \cdot \mathbb{E}\left[\left(\frac{L_t}{C}\right)^r\right] \leq C^p \cdot \sum_{r=0}^{p} \left(\frac{ep}{r}\right)^r \cdot \frac{(2er)^r \cdot 10^{tr}}{C^r} \\ = C^p \cdot \sum_{r=0}^{p} \left(\frac{2e^2 \cdot p \cdot 10^t}{C}\right)^r \leq C^p \sum_{r=0}^{\infty} \varepsilon^{-r} \leq \frac{1}{1-\varepsilon} \cdot C^p \leq (1+2\varepsilon)C^p, \qquad (18)\end{aligned}$$

where the third transition follows from Lemma 6.3 (applied p times with $p = r$), and the second before last transition follows since $\frac{2e^2 \cdot p \cdot 10^t}{C} = \frac{2e^2 \cdot p \cdot 10^t}{10^t \cdot \gamma \cdot n^{1/p}} \ll \varepsilon$ for large enough n.

Next, we upper bound the m-th moment of L_I^p. Similar to the calculations in (18), we have

$$\begin{aligned}&\sigma_m((L_I^p))^m := \mathbb{E}_{\eta_0}\left[|(L_t+C)^p - \mathbb{E}[(L_t+C)^p]|^m\right] \leq \\ &\leq \mathbb{E}_{\eta_0}\left[|(L_t+C)^p - C^p|^m\right] \leq \mathbb{E}_{\eta_0}\left[C^{pm} \cdot \left(\sum_{r=0}^{p}\left(\frac{ep}{r \cdot C}\right)^r \cdot L_t^r - 1\right)^m\right] \\ &= \mathbb{E}_{\eta_0}\left[C^{pm} \cdot \left(\sum_{r=1}^{p}\left(\frac{ep}{r \cdot C}\right)^r \cdot L_t^r\right)^m\right] \leq \mathbb{E}_{\eta_0}\left[C^{pm} \cdot p^m \cdot \sum_{r=1}^{p}\left(\frac{ep}{r \cdot C}\right)^{rm} \cdot L_t^{rm}\right] \qquad (19) \\ &\leq C^{pm} \cdot \sum_{r=1}^{p}\left(\frac{ep^2}{r \cdot C}\right)^{rm} \cdot \mathbb{E}_{\eta_0}[L_t^{rm}] \leq C^{pm} \cdot \sum_{r=1}^{p}\left(\frac{ep^2}{r \cdot C}\right)^{rm} \cdot (2eprm)^{rm} \cdot 10^{trm} \qquad \text{(By Lemma 6.3)} \\ &= C^{pm} \cdot \sum_{r=1}^{p}\left(\frac{10^t \cdot 2e^2p^3m}{C}\right)^{rm} \leq C^{pm} \cdot \frac{\varepsilon^m}{1-\varepsilon^m} \leq (2\varepsilon C^p)^m, \qquad (20)\end{aligned}$$

where (19) follows from Jensen's inequality ($(\sum_{i=1}^n a_i)^m \leq n^m \cdot \sum_{i=1}^n a_i^m$), and the second before last transition again follows since $\frac{10^t \cdot 2e^2p^3m}{C} = \frac{10^t \cdot 2e^2p^3m}{10^t \cdot \gamma \cdot n^{1/p}} \ll \varepsilon$ by the premise $m = o(n^{1/p})$.

Given the assumption $\delta > 2^{-o(n^{1/p})}$, we can now apply Lemma 2.6 with $m = \log 1/\delta$ $(\leq o(n^{1/p})))$, $\lambda = 2$, to conclude that

$$\begin{aligned}&\Pr_{\eta_0}[(L_I)^p \geq (1+7\varepsilon)C^p] \leq \Pr_{\eta_0}\left[|(L_I)^p - \mathbb{E}[(L_I)^p]| > 4\varepsilon C^p\right] \\ &\leq \Pr_{\eta_0}\left[\, |(L_I)^p - \mathbb{E}[(L_I)^p]| > 2 \cdot \sigma_m((L_I)^p)\,\right] \leq 2^{-m} = \delta,\end{aligned}$$

where the first and second transitions follow from (18) and (20) respectively. □

□

References

[AKO10] Alexandr Andoni, Robert Krauthgamer, and Krzysztof Onak. Streaming algorithms from precision sampling. *CoRR*, abs/1011.1263, 2010.

[AMS99] Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *JCSS*, 58(1):137–147, 1999.

[And] Alexandr Andoni. High frequency moment via max stability. Available at http://web.mit.edu/andoni/www/papers/fkStable.pdf.

[ANPW13a] Alexandr Andoni, Huy L. Nguyên, Yury Polyanskiy, and Yihong Wu. Tight lower bound for linear sketches of moments. In *ICALP (1)*, pages 25–32, 2013.

[ANPW13b] Alexandr Andoni, Huy Le Nguyen, Yury Polyanskiy, and Yihong Wu. Tight lower bound for linear sketches of moments. In *ICALP*, 2013.

[BCK$^+$14] Joshua Brody, Amit Chakrabarti, Ranganath Kondapally, David P. Woodruff, and Grigory Yaroslavtsev. Certifying equality with limited interaction. In *RANDOM*, pages 545–581, 2014.

[BEO$^+$13] Mark Braverman, Faith Ellen, Rotem Oshman, Toniann Pitassi, and Vinod Vaikuntanathan. A tight bound for set disjointness in the message-passing model. In *FOCS*, pages 668–677, 2013.

[BGKS06] Lakshminath Bhuvanagiri, Sumit Ganguly, Deepanjan Kesh, and Chandan Saha. Simpler algorithm for estimating frequency moments of data streams. In *SODA*, pages 708–713, 2006.

[BGPW13] Mark Braverman, Ankit Garg, Denis Pankratov, and Omri Weinstein. Information lower bounds via self-reducibility. In *CSR*, 2013.

[BKSV14] Vladimir Braverman, Jonathan Katzman, Charles Seidell, and Gregory Vorsanger. An optimal algorithm for large frequency moments using o(nˆ(1-2/k)) bits. In *APPROX/RANDOM*, 2014.

[BO10] Vladimir Braverman and Rafail Ostrovsky. Recursive sketching for frequency moments. *CoRR*, abs/1011.2571, 2010.

[BO12] Vladimir Braverman and Rafail Ostrovsky. Approximating large frequency moments with pick-and-drop sampling. *CoRR*, abs/1212.0202, 2012.

[BO15] Mark Braverman and Rotem Oshman. The communication complexity of number-in-hand set disjointness with no promise. *Electronic Colloquium on Computational Complexity (ECCC)*, 22:2, 2015.

[BR11] Mark Braverman and Anup Rao. Information equals amortized communication. In *FOCS*, 2011.

[BYJKS04a] Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar, and D. Sivakumar. An information statistics approach to data stream and communication complexity. *Journal of Computer and System Sciences*, 68(4):702–732, 2004.

[BYJKS04b] Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar, and D. Sivakumar. An information statistics approach to data stream and communication complexity. *Journal of Computer and System Sciences*, 68(4):702–732, 2004.

[CDM$^+$13] Kenneth L. Clarkson, Petros Drineas, Malik Magdon-Ismail, Michael W. Mahoney, Xiangrui Meng, and David P. Woodruff. The fast cauchy transform and faster robust linear regression. In *SODA*, 2013.

[CK04] Don Coppersmith and Ravi Kumar. An improved data stream algorithm for frequency moments. In *SODA*, 2004.

[CKS03] Amit Chakrabarti, Subhash Khot, and Xiaodong Sun. Near-optimal lower bounds on the multi-party communication complexity of set disjointness. In *CCC*, pages 107–117, 2003.

[CKW12] Amit Chakrabarti, Ranganath Kondapally, and Zhenghui Wang. Information complexity versus corruption and applications to orthogonality and gap-hamming. In *APPROX-RANDOM*, pages 483–494, 2012.

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

[CT91] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley series in telecommunications. J. Wiley and Sons, New York, 1991.

[Gan04a] Sumit Ganguly. Estimating frequency moments of data streams using random linear combinations. In *RANDOM*, 2004.

[Gan04b] Sumit Ganguly. A hybrid algorithm for estimating frequency moments of data streams, 2004. Manuscript.

[Gan08] Sumit Ganguly. Lower bounds on frequency estimation of data streams (extended abstract). In *CSR*, pages 204–215, 2008.

[Gan11] Sumit Ganguly. Polynomial estimators for high frequency moments. *CoRR*, abs/1104.4552, 2011.

[Gan12] Sumit Ganguly. A lower bound for estimating high moments of a data stream. *CoRR*, abs/1201.0253, 2012.

[HW07] Johan Håstad and Avi Wigderson. The randomized communication complexity of set disjointness. *Theory of Computing*, 3(1):211–219, 2007.

[IW05] P. Indyk and D. Woodruff. Optimal approximations of the frequency moments of data streams. In *STOC*. ACM, 2005.

[JW13] T. S. Jayram and David P. Woodruff. Optimal bounds for johnson-lindenstrauss transforms and streaming problems with subconstant error. *ACM Transactions on Algorithms*, 9(3):26, 2013.

[KNW10] Daniel M. Kane, Jelani Nelson, and David P. Woodruff. On the exact space complexity of sketching and streaming small norms. In *SODA*, pages 1161–1178, 2010.

[KS92] Bala Kalyanasundaram and Georg Schnitger. The probabilistic communication complexity of set intersection. *SIAM Journal on Discrete Mathematics*, 5(4):545–557, November 1992.

[Lat97] Rafal Latala. Estimation of moments of sums of independent real random variables. *The Annals of Probability*, 25(3):1502–1513, 07 1997.

[Lin06] J. Lin. Divergence measures based on the shannon entropy. *IEEE Trans. Inf. Theor.*, 37(1):145–151, September 2006.

[LNW14] Yi Li, Huy L. Nguyen, and David P. Woodruff. Turnstile streaming algorithms might as well be linear sketches. In *STOC*, pages 174–183, 2014.

[LW13] Yi Li and David P. Woodruff. A tight lower bound for high frequency moment estimation with small error. In *RANDOM*, pages 623–638, 2013.

[MW10] Morteza Monemizadeh and David P. Woodruff. 1-pass relative-error l_p-sampling with applications. In *SODA*, 2010.

[MWY13] Marco Molinaro, David Woodruff, and Grigory Yaroslavtsev. Beating the direct sum theorem in communication complexity with implications for sketching. In *SODA*, 2013.

[PW12] Eric Price and David P. Woodruff. Applications of the shannon-hartley theorem to data streams and sparse recovery. In *ISIT*, 2012.

[She14] Alexander A. Sherstov. Communication lower bounds using directional derivatives. *J. ACM*, 61(6):34, 2014.

[SS02] Michael Saks and Xiaodong Sun. Space lower bounds for distance approximation in the data stream model. In *STOC*, 2002.

[SW11] Christian Sohler and David P. Woodruff. Subspace embeddings for the l_1-norm with applications. In *STOC*, pages 755–764, 2011.

[Woo04] David P. Woodruff. Optimal space lower bounds for all frequency moments. In *SODA*, pages 167–175, 2004.

[WZ12] David P. Woodruff and Qin Zhang. Tight bounds for distributed functional monitoring. In *STOC*, pages 941–960, 2012.

A Direct Sum for Set-Disjointness

Proof of Lemma 3.4. Let Π be a δ-error protocol for Disj_k^n, and denote

$$\mathsf{IC}_{\mu_0^n}(\Pi) = I_{\mu_0^n}(\Pi; \mathbf{X}_1, \ldots \mathbf{X}_k) := I. \tag{21}$$

We will use the n-fold protocol Π to generate a 2δ-error protocol π for AND_k (under any input $(x_{,} \ldots, x_k) \in \{0,1\}^k$), whose information cost under μ_0 is at most I/n, completing the proof (since $\|\Pi\| \geq I$). The protocol $\pi(x_1, \ldots, x_k)$ is defined as follows:

- The players use public randomness to sample a random coordinate $i \in_R [n]$.
- The players "embed" their inputs to the i'th coordinate of Π, and sample the rest of the coordinates *privately*[2] according to μ_0: Each player $j \in [k]$ sets

$$\mathbf{X}_j := (X_{j,1}, \ldots, X_{j,i-1}, x_j, X_{j,i+1}, \ldots, X_{j,n}), \qquad \text{where } X_{j,t} \sim B(1/k) \text{ for all } t \neq i.$$

- The players run $\Pi(\mathbf{X}_1, \ldots, \mathbf{X}_k)$ and output its answer.

To argue about the correctness of π, notice that since all coordinates $[n] - \{i\}$ are sampled according to μ_0, (1) implies that except with probability δ,

$$\overline{\mathsf{Disj}_k^n}(\mathbf{X}_1, \ldots, \mathbf{X}_k) = \bigvee_{i=1}^{n} \mathsf{AND}_k(\mathbf{X}_{1,i}, \ldots, \mathbf{X}_{k,i}) = \mathsf{AND}_k(x_1, \ldots, x_k),$$

and since Π itself is correct with probability $1-\delta$, we get that π is correct with probability at least $1-2\delta$, as claimed. To analyze the information cost of π under μ_0, we write

$$\begin{aligned}
\mathsf{IC}_{\mu_0}(\pi) &= \mathbb{E}_i \left[I_{\mu_0}(\Pi; \mathbf{X}_{1,i}, \ldots \mathbf{X}_{k,i}) \right] \\
&\leq \frac{1}{n} \sum_{i=1}^{n} I_{\mu_0}(\Pi; \mathbf{X}_{1,i}, \ldots \mathbf{X}_{k,i} \mid \mathbf{X}_{1,\leq i}, \mathbf{X}_{2,\leq i}, \ldots, \mathbf{X}_{k,\leq i}) \\
&= \frac{1}{n} \cdot I_{\mu_0^n}(\Pi; \mathbf{X}_1, \ldots \mathbf{X}_k) = \frac{I}{n},
\end{aligned} \tag{22}$$

where the inequality follows from Lemma 2.11, since for any $t \neq i$ and $j, \ell \in [k]$, $\mathbf{X}_\ell^t \perp \mathbf{X}_j^i$, and the last equality follows from the chain rule for mutual information.

□

B Tightness of [LW13] Communication Problem

As noted in the introduction, Li and Woodruff [LW13] prove their lower bound via a reduction to a 2-party 1-way communication problem : In the Gap-$L_\infty[B]$ problem, Alice and Bob are each given an n-dimensional vector $x, y \in [B]^n$ with the promise that either $\forall\ i \in [n]\ \ |x_i - y_i| \leq 1$ ("NO") or $\exists\ i \in [n]\ \ |x_i - y_i| = B$ ("YES"), and they need to determine which case it is using a 1-way protocol between Alice and Bob. To facilitate a direct-sum based lower bound, [LW13] analyzed the information complexity of this problem under the following distribution μ supported on "NO" instances: For each coordinate $i \in [n]$, set $x \in_R [B]$, and set $y \in [x, x+1]$ independently at random conditioned on x ($y = B$ if $x = B$).

We now sketch a proof showing that there is an $\exp(-\Omega(n/B^2))$-error randomized 1-way protocol for Gap-$L_\infty[B]$, whose communication cost under the distribution μ is $O(n/B^2)$. This in turn implies that the lower bound of [LW13] cannot be improved in terms of its error parameter, at least using the direct-sum technique (which must analyze the information/communication complexity under the "NO" distribution in order to prove communication lower bounds for the composite problem).

We actually present a 0-*error* protocol for Gap-$L_\infty[B]$ whose (internal) information cost is $O(n/B^2)$, and then use the compression result of [BR11] to compress this protocol (at the price of a tiny error probability). Indeed, consider the protocol τ in which, Bob sends Alice the following vector $v_B \in \{0,1,2\}^n$:

[2] Note that this is indeed possible since μ_0 is a product distribution over the coordinates in $[k]$.

- For each coordinate $i \in [n]$, independently perform the following:
- if $y_i = 0$, Bob sends Alice 0.
- if $y_i = B/2$, Bob sends Alice 1.
- if $y_i = B$, Bob sends Alice 2.
- For all $0 < j < B/2$: If $y_i = j$, Bob sends 0 with probability $\cos(\pi j/B)$, and 1 with probability $1 - \cos(\pi j/B) = \sin(\pi j/B)$ (i.e., Alice sends 0 with probability "evenly" spread on the unit circle).
- For all $B/2 < j < B$: If $y_i = j$ Bob sends 1 with probability $\cos(\pi j/B)$, and 2 with probability $1 - \cos(\pi j/B) = \sin(\pi j/B)$.
- Alice generates a vector $v_A \in \{0,1,2\}^n$ in the exact same process according to her input $x = x_1, \dots, x_n$. When she receives v_B from Bob, she declares "YES" iff there is some $i \in [n]$ for which $(v_A(i), v_B(i)) = (0,2)$ or $(v_A(i), v_B(i)) = (2,0)$. Otherwise, she declares "NO".

The protocol has 0 error since in a "NO" instance, there is no distribution on transcripts of τ on a j and a $j+1$ with a 0 in support of distribution for j and 2 in support of distribution for $j+1$. In a "YES" instance clearly the protocol is correct with zero error. To analyze the information cost of τ, we will use the well know connection between mutual information and KL divergence

$$I(A;B|C) = \mathbb{E}_{b,c}\left[\mathsf{D}\left(A|b,c \| A|c\right)\right],$$

wehre $\mathsf{D}\left(A|b,c\|A|c\right) := \mathbb{E}_a\left[\log\frac{p(a|bc)}{p(a|c)}\right]$. We will also need the following proposition in our analysis:

Proposition B.1 (Jensen-Shannon divergence vs. Hellinger distance, [Lin06]). *For two distributions μ, ν, define the Jensen-Shannon divergence as $JS(\mu,\nu) := \frac{1}{2}\left(\mathsf{D}\left(\mu\|\frac{\mu+\nu}{2}\right) + \mathsf{D}\left(\nu\|\frac{\mu+\nu}{2}\right)\right)$. Then*

$$h^2(\mu,\nu) \le JS(\mu,\nu) \le \ln 2 \cdot h^2(\mu,\nu).$$

We are now ready to show that $I(\tau;Y|X) \le O(n/B^2)$. To see this, let T denote the transcript of Bob's message in τ and let T_i denote Bob's transcript of the i'th coordinate. Recall that by definition, $(T_i|y_i = j) \sim B(j^2/B^2)$ for all $0 < j < B$ (indeed, we will not need to distinguish between the case where the support of this binary random variable is $(0,1)$ or $(1,2)$ since for information purposes these distributions are the same). Therefore,

$$\begin{aligned}
&I(T_i;Y_i|X_i) = \mathbb{E}_{(x_i,y_i)\sim\mu}\mathsf{D}\left(T_i|x_i,y_i\|T_i|x_i\right)\\
&= \frac{1}{B}\cdot\sum_{j=1}^{B/2}\frac{1}{2}\Big(\mathsf{D}\left(T_i|x_i=j,y_i=j\|T_i|x_i=j\right) + \mathsf{D}\left(T_i|x_i=j,y_i=j+1\|T_i|x_i=j\right)\Big)\\
&= \frac{1}{B}\cdot\sum_{j=1}^{B/2} JS(T_i|x_i=j,y_i=j\ ,\ T_i|x_i=j,y_i=j+1) \quad \text{(By the definition in B.1)}\\
&\le \frac{\ln 2}{B}\cdot\sum_{j=1}^{B/2} h^2(T_i|x_i=j,y_i=j\ ,\ T_i|x_i=j,y_i=j+1) \quad \text{(By proposition B.1)}
\end{aligned} \tag{23}$$

Define the vector $u_j := (\cos(\pi j/B), \sin(\pi j/B)) \in \mathbb{R}^2$. By definition of the protocol and the Hellinger distance,

$$\begin{aligned}
&h^2(T_i|x_i=j,y_i=j\ ,\ T_i|x_i=j,y_i=j+1) = \|\sqrt{u_j} - \sqrt{u_{j+1}}\|^2 = 1 + 1 - 2\langle u_j, u_{j+1}\rangle\\
&= 2 - 2\|u_j\|\|u_{j+1}\|\cos(\pi/2B) \qquad \text{(since the angle between } u_j, u_{j+1} \text{ is } \pi/2B)\\
&= 2 - 2(\sqrt{1-(\pi/2B)^2}) \qquad \text{(since } u_j, u_{j+1} \text{ are unit vectors and } \sin(x) \le x)\\
&\le 2 - 2(1-(\pi/2B)^2) \quad \text{(since } \sqrt{1-x} \ge 1-x \text{ for } x \in [0,1])\\
&\le \left(\frac{\pi}{2B}\right)^2 = \Theta\left(\frac{1}{B^2}\right).
\end{aligned}$$

Thus, by (23) we have that

$$I(T_i; Y_i|X_i) \le \frac{\ln 2}{B} \cdot \sum_{j=1}^{B/2} O(1/B^2) = O(1/B^2).$$

By independence of the coordinates and the chain rule, we have $I(T; Y|X) \le O(n/B^2)$, as claimed.

To compress his (1-round) message T, Bob can now use a 1-round version of the correlated-smapling scheme of [BR11] (Theorem 4.1, setting $P := T|Y$ and $Q := T|X$). This simulation asserts that Bob can send Alice a message T' (using shared randomness) such that Alice produces from T' a correct sample $v \sim T|Y$ except with error $\delta = 2^{-\Omega(n/B^2)}$, and $\mathbb{E}[|T'|] \le 2 \cdot I(T; Y|X) + \log(1/\delta) \le O(n/B^2)$. To turn this expected communication into a worst-case guarantee, observe that under the distribution μ, $T = T_1 T_2 \dots T_n$ is a collection of i.i.d random variables, and so a standard Chernoff bound implies that the log-ratio of the distributions $T|Y$ and $T|X$ for any $T = t$, is sharply concentrated around $I := I(T; Y|X) = O(n/B^2)$ (in other words, $\Pr_{x,y,t}\left[\frac{\Pr[(T|y)=t]}{\Pr[T|x=t]} > 2^{O(I)}\right] \le \exp(-\Omega(I)) \le \exp(-\Omega(n/B^2))$. Therefore, the simulation of [BR11] would succeed in a single round with probability $1 - \exp(-\Omega(n/B^2))$, provided that we set $\delta = 2^{-\Theta(n/B^2)}$, $C_t = 2^{O(I)}$ in Step 4 of their compression scheme. In this case, Theorem 4.1 in [BR11] guarantees that the simulation requires only $O(I)$ bits of communication, and by a union bound, the error of this protocol under μ is at most $\delta + 2^{-\Omega(I)} = 2^{-\Omega(n/B^2)} = \exp(-\Omega(n/B^2))$, which completes the proof.

C 1-way Upper Bound for k-Party Disjointness

Here we sketch a proof showing that, in contrast to our SMP lower bound, the k-party communication complexity of Set-Disjointness in the 1-way model under the distribution η (defined in Section 3) is upper bounded by $O(n/k)$ (and therefore resorting to the weaker SMP model is necessary for gaining the extra $\min\{\log k, \log 1/\delta\}$ factor in our lower bound).

The proof is very similar to the 2-party public-coin protocol for sparse Set-Disjointness of Håstad and Wigderson [HW07]: **Stage 1 :** The first player can interpret the public coin as a sequence $S_1, S_2, S_3 \dots$ of random subsets of $[n]$ where each item is independently included with probability 1/2. Player 1 starts by sending player 2 the index t_1 of the first subset S_{t_1} containing his set x_1. Note that since under η_0 $X_{j,i} \sim B(1/k)$, $\mathbb{E}[|x_1|] = n/k$, and therefore the probability that a random subset S_t contains x_1 is $2^{-O(n/k)}$, and therefore after scanning $2^{O(n/k)}$ random subsets, one of them contains the set x_1 with probability $1 - \exp(-n/k)$, in which case $t_1 \le O(n/k)$.

When player 2 receives the index t_1, he sets $x_2' := x_2 \cap S_{t_1}$. Since S_{t_1} is random (conditioned on containing x_1), $\mathbb{E}[x_2 \cap S_{t_1}] = \frac{n}{2k}$ (and the intersection size is concentrated around $n/2k$ by a Chernoff bound). Player 2 then sends player 3 the first index $t_2 (> t_1)$ for which $x_2' \subset S_{t_2}$, using $n/(2k)$ bits (in expectation). Player 3 then sets $x_3' := x_3 \cap S_{t_2}$ and continues as before. The players continue in this fashion until player $j_0 := O(\log\log n)$. Notice that with probability $1 - \exp(-n/k)$, $|x_{j_0}'| = O\left(\frac{n}{k \cdot 2^{-j_0+1}}\right) = O\left(\frac{n}{k \cdot \log^2(n)}\right)$.

Stage 2 : The next players $j_0, \dots j_0 + O(\log n)$ now sequentially communicate their *entire* remaining sets x_j' from one to another, using $O\left(\log^2(n) \cdot \frac{n}{k \cdot \log^2(n)}\right) = O\left(\frac{n}{k}\right)$ bits, where x_j' is the set of elements of player j which belong to x_{j-i}' (note that this stage is deterministic).

Not that the protocol is correct whenever the stage 1 succeeds (that is, assuming $|x_{j_0}'| = O\left(\frac{n}{k \cdot \log^2(n)}\right)$), and the communication bounds in stage 1 are satisfied. which (by a union bound) happens with probability $1 - \exp(n/k)$. Furthermore, in this event the total communication of this protocol is upper bounded by

$$\left(\sum_{j=1}^{O(\log\log(n))} \frac{n}{k \cdot 2^{j-1}}\right) + O\left(\frac{n}{k}\right) \le O\left(\frac{n}{k}\right), \quad \text{as desired.}$$