

Revisiting the Direct Sum Theorem and Space Lower Bounds in Random Order Streams

Sudipto Guha and Zhiyi Huang*

University of Pennsylvania, Philadelphia PA 19104, USA
{sudipto,hzhiyi}@cis.upenn.edu

Abstract. Estimating frequency moments and L_p distances are well studied problems in the adversarial data stream model and tight space bounds are known for these two problems. There has been growing interest in revisiting these problems in the framework of random-order streams. The best space lower bound known for computing the k^{th} frequency moment in random-order streams is $\Omega(n^{1-2.5/k})$ by Andoni et al., and it is conjectured that the real lower bound shall be $\Omega(n^{1-2/k})$. In this paper, we resolve this conjecture. In our approach, we revisit the direct sum theorem developed by Bar-Yossef et al. in a random-partition private messages model and provide a tight $\Omega(n^{1-2/k}/\ell)$ space lower bound for any ℓ -pass algorithm that approximates the frequency moment in random-order stream model to a constant factor. Finally, we also introduce the notion of space-entropy tradeoffs in random order streams, as a means of studying intermediate models between adversarial and fully random order streams. We show an almost tight space-entropy tradeoff for L_∞ distance and a non-trivial tradeoff for L_p distances.

1 Introduction

The data stream model is a very useful computational model for designing efficient algorithms for massive data sets. In the data stream model, the algorithm can only access the data in a given order and for a limited number of times (passes). Designing sub-linear space algorithms and proving space lower bound for numerous problems have received a lot of attention.

The problem of estimating the Frequency Moments is one of the most studied problems in data stream model. Given an alphabet $\Sigma = \{\sigma_1, \sigma_2, \dots, \sigma_m\}$ of size m and a sequence of n numbers x_1, x_2, \dots, x_n in Σ , y_i is the number of occurrence of σ_i in the sequence, and the k^{th} frequency moment f_k is defined as $f_k = \sum_{i=1}^m y_i^k$.

Usually, it is assumed that the order is given by an adversary and the model is known as adversarially ordered streaming. In this model, there are approximation algorithms for computing the k^{th} frequency moment using only $\tilde{O}(n^{1-2/k})$ space [4,13]. Alon et al. [1] proved the first lower bound of $\Omega(n^{1-5/k})$ for the space required to estimate the k^{th} frequency moment to a constant factor. Bar-Yossef et al. [3] gave an improved lower bound of $\Omega(n^{1-3/k})$ via their direct

* This research was supported in part by NSF award CCF-0644119.

sum theorem. And Chakrabarti et al. [7] showed that any single-pass algorithm required $\Omega(n^{1-2/k})$ space in order to approximate the k^{th} frequency moment, while for algorithms with a constant number of passes require $\Omega(n^{1-2/k}/\log n)$ space. Very recently, Gronemeier [9] improved the lower bound for constant-pass algorithms to $\Omega(n^{1-2/k})$.

A related and almost equally well studied problem in the data stream model is the approximation of L_∞ and L_p distances. Given $x = (x_1, x_2, \dots, x_n) \in [0, \ell]^n$ and $y = (y_1, y_2, \dots, y_n) \in [0, \ell]^n$, the L_p distance between x and y is defined as $L_p(x, y) = (\sum_{i=1}^n (x_i - y_i)^p)^{1/p}$. The L_∞ distance between x and y is $\max_i |x_i - y_i|$. Saks and Sun [15] proved that any two-party one-way protocol that distinguishes $L_\infty(x, y) = 1$ from $L_\infty(x, y) = \ell$ with probability at least $2/3$ uses at least $\Omega(n/\ell^2)$ communication. Later, Bar-Yossef et al. [3] use their direct sum theorem to prove the same space lower bound for general two-party protocols. Matching protocols for this problem are also known. Using a reduction from L_∞ to L_p proposed by Saks and Sun, a space lower bound of $\Omega(n^{1-2/p}/\ell^2)$ holds for L_p , $p > 2$.

In many scenarios, however, an adversarially ordered data stream is not the best model, and recently, random-order data streams has received a lot of attention [12,5,6]. The work which is closest to this paper, by Chakrabarti et al. [5] show that the space complexity of estimating the k^{th} frequency moment is $\Omega(n^{1-3/k})$ and $\Omega(n^{1-3/k}/\log n)$ for single-pass and constant-passes algorithms respectively for the random order stream model. Andoni et al. [2] improve these lower bounds to $\Omega(n^{1-2.5/k}/\log n)$ and conjecture that the lower bound for adversarially ordered streams holds for random-order streams.

Communication complexity [14,16] plays a central role in proofs of most results on space lower bound results. There are two models of communication complexity which are useful in this context. The blackboard model refers to the communication games in which players can broadcast their message to all other players. In the private messages model, only one-to-one communication is allowed. In the literature to date, most lower bound results are based on reductions from various communication complexity problems in the blackboard model. And a key technique is the direct sum theorem developed by Bar-Yossef et al. [3]. In contrast, the private messages model has received less attention so far. The private messages model is more restrictive than the broadcast model, may lead to better space lower bounds; and further, to prove lower bounds in the streaming model, the private message model is more relevant (in fact the order in which the players speak is also preordained). To the best of our knowledge, the only effort on proving space lower bound from communication complexity in private messages model is the work on the longest increasing sequence problem by Gal and Gopalan [8]. We note that direct lower bounds for streaming problems that bypass communication games as in [11] also use ideas which are similar in spirit to the private messages model.

Our Contributions. In this paper, we revisit the notion of information cost and information complexity in the framework of private messages model. We prove that the private information cost of a decomposable function is at least

as large as the sum of the private information costs of the primitive functions. Using this direct sum theorem, we prove a tight $\Omega(nm/t^2)$ lower bound for the communication complexity of random-partition multiparty set disjointness. Here n is the number of different items, m is the number of players. The players try to distinguish the case that all items are distinct and the case that there are t identical items. As a corollary of this result, we show that any ℓ -pass algorithm which gives constant factor approximation of the k^{th} frequency moment in random-order stream model requires $\Omega(n^{1-2/k}/\ell)$ space. This result resolves the conjecture by Andoni et al. [2]. It also provides an alternate approach for space lower bound for constant-pass algorithms in adversarially ordered streams.

We then study protocols for L_∞ and the tradeoff of the entropy of the input order and the communication complexity used by the protocol. We show that if the protocol can distinguish $L_\infty = \ell$ and $L_\infty \leq 1$, and $2n \log n - E = \alpha n \log n$, then the $2n$ -party communication complexity is at least $\Omega(n^{2-\alpha(1+\epsilon)}/\ell^2)$ for any constant $\epsilon > 0$. As a corollary, we have $\Omega(n^{1-\alpha(1+\epsilon)}/\ell^2)$ and $\Omega(n^{1-2/p-\alpha(1+\epsilon)}/\ell^2)$ space lower bounds for data stream algorithms which approximates L_∞ and L_p for $p > 2$ respectively. We also prove this tradeoff is essentially tight for L_∞ and give algorithm matching the lower bound.

2 Preliminaries

2.1 Definitions and Notations

Definition 1. Suppose Σ is a finite set. A function $f : \Sigma^T \mapsto \{0, 1\}$ is defined to be decomposable if there exists t , n and functions $g : \{0, 1\}^n \mapsto \{0, 1\}$ and $h : \Sigma^t \mapsto \{0, 1\}$ such that $T = tn$ and the function f is of the form $f(x_1, x_2, \dots, x_T) = g(h(x_1, x_2, \dots, x_t), \dots, h(x_{(n-1)t+1}, \dots, x_T))$. We call function h the primitive function.

We shall consider the following two special cases of decomposable functions in this paper. If h is the AND_t function with t input bits, and g is function OR_n with n input bits, the decomposable function f is denoted as the SET DISJOINTNESS function:

$$\text{SETDISJ}_{n,t} = \text{OR}_n(\text{AND}_t(x_1), \dots, \text{AND}_t(x_n)) \ .$$

If h is the bivariate gap function BIGAP_ℓ such that $\text{BIGAP}_\ell(x, y) = 1$ when $|x - y| = \ell$ and $\text{BIGAP}_\ell(x, y) = 0$ when $|x - y| = 0, 1$, and g is function OR with n input bits, then the decomposable function f is denoted as the GAP DISTANCE function:

$$\text{GAPDIST}_{n,\ell} = \text{OR}_n(\text{BIGAP}_\ell(x_1, x_2), \dots, \text{BIGAP}_\ell(x_{2n-1}, x_{2n})) \ .$$

We use capital letters X , Y , and Z to denote random variables. We use bold-face letters \mathbf{X} and \mathbf{Y} to denote vectors. Moreover, we shall let $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ denote the input vectors of primitive functions and let $\mathbf{X} = \mathbf{X}_1 \times \mathbf{X}_2 \times \dots \times \mathbf{X}_n$ denote the input vector of the decomposable function f . We let ν denote the input distribution of the primitive function and let μ denote the input distribution of the decomposable function. Usually we shall have $\mu = \nu^n$.

We use $[d]$ to denote the set $\{1, 2, \dots, d\}$. We say a distribution μ is *symmetric* if and only if for any permutation π of $[T]$, $\mathbf{X} \sim \pi(\mathbf{X}) \sim \mu$.

Definition 2. A distribution μ is defined as a collapsing distribution if for any input \mathbf{x} drawn from the distribution μ and any $\mathbf{X}_i \in \Sigma^t$, we always have that $f(\mathbf{x}_1, \dots, \mathbf{x}_{i-1}, \mathbf{X}_i, \mathbf{x}_{i+1}, \dots, \mathbf{x}_n) = h(\mathbf{X}_i)$.

We shall use η to denote the distribution of random variable Y_i and let ζ to denote the distribution of random vector \mathbf{Y} . We shall have $\zeta = \eta^n$. We will consider the distribution of random vector \mathbf{X} conditioned on \mathbf{Y} .

Definition 3. \mathbf{Y} is defined to partition \mathbf{X} if the distribution of \mathbf{X} given \mathbf{Y} is a product distribution.

2.2 Communication Games and Various Models

We let \mathcal{P} denote a communication protocol. We shall always use δ to denote the error rate of a protocol. Let Γ denote the set of all protocols and let Γ_δ denote the set of all protocols whose error rate is at most δ . Similarly we shall use Φ and Φ_δ to denote the set of all deterministic protocols and the set of all deterministic protocols with error rate at most δ .

The term ϵ be denote the relevant approximation parameter (we shall consider either $(1 + \epsilon)$ -approximation or n^ϵ -approximation depending on the problem we study). We use ρ to denote other small values.

Private Messages Model: We shall focus on the communication complexity of various (decomposable) functions in private messages model (with public coins) in this paper. A multiparty communication game in private messages model with m players is as follows. In step 1, the first player sends a message M_1^1 to the second player based merely on her own input. In general, in step $im + j$ such that $i \geq 0$ and $1 \leq j \leq m$ the j^{th} player sends a message M_{i+1}^j to the $(j + 1)^{\text{th}}$ player based on her own input and all messages she received from the $(j - 1)^{\text{th}}$ player. Note that in private messages model, each message is known only by the sender and recipient. This is a major difference from the blackboard model. We shall use $\text{CC}_\delta^P(f)$ to denote the multiparty communication complexity of computing a decomposable function f in private messages model with error rate at most δ .

The transcript of the ℓ^{th} player is the union of all messages sent by player ℓ and is denoted by $\Pi_\ell(\mathbf{X})$. The transcript $\Pi(\mathbf{X})$ is the union of $\Pi_\ell(\mathbf{X})$ for $1 \leq \ell \leq m$. We sometimes abbreviate these notations with Π_ℓ and Π . The communication complexity $\text{CC}_\delta^P = \min_{\mathcal{P} \in \Gamma_\delta} \max_{\mathbf{x} \in \{0,1\}^T} |\Pi(\mathbf{x})|$.

Random Partitioned Communication Games: An allocation is a function $\sigma : [T] \mapsto [m]$. Let $[m]^T$ denote the set of all allocations. In a random partitioned communication game with respect to function f and a distribution Σ on $[m]^T$, an allocation σ is drawn from distribution Σ , and each input bit x_i is given to the $\sigma(i)^{\text{th}}$ player. The players then play a communication game in private messages model to compute the function value of f for the given input. Let $\mathcal{U}_{T,m}$ denote

the uniformly random distribution over $[m]^T$. The special case when $\Sigma = \mathcal{U}_{T,m}$ is of particular interest in proving robust communication complexity and space lower bounds for various functions.

We shall use $\Gamma_{\delta,\Sigma}$ to denote the set of protocols whose error rate is at most δ in the random partitioned communication game with respect to function f and distribution Σ . And the communication complexity in a random partitioned communication game is $\text{CC}_{\delta,\Sigma}^P = \min_{\mathcal{P} \in \Gamma_{\delta,\Sigma}} \max_{x \in \{0,1\}^T} |\Pi|$.

3 Revisiting the Direct Sum Theorem

Now we revisit the definition of information cost and information complexity in the literature of private messages model. A major difference between private messages model and blackboard model is that a player may need to forward information in the messages she received to the other players, while in blackboard model that information is already known by every player.

Therefore, any optimal protocol in blackboard model shall satisfies that $I(\Pi_i; \Pi_j) = 0$ for any $1 \leq i \neq j \leq m$. Thus we shall have that $I(\mathbf{X}; \Pi) = \sum_{i=1}^m I(\mathbf{X}; \Pi_i)$. However, similar statement is not true in private messages model. Based on this observation, we consider the following definition of information cost and information complexity in private messages model.

Definition 4. Suppose \mathcal{P} is a communication protocol and Π is its transcript. The information cost of \mathcal{P} with respect to the input distribution $\mathbf{X} \sim \boldsymbol{\mu}$ is $\text{ICost}_{\boldsymbol{\mu}}(\mathbf{X}; \Pi) = \sum_{i=1}^m I_{\boldsymbol{\mu}}(\mathbf{X}; \Pi_i)$. The δ -error information complexity with respect to function f and input distribution $\mathbf{X} \sim \boldsymbol{\mu}$ is the minimal information cost among all δ -error protocols, that is, $\text{IC}_{\boldsymbol{\mu},\delta}(f) = \min_{\mathcal{P} \in \Gamma_{\delta}} \text{ICost}_{\boldsymbol{\mu}}(\mathbf{X}; \Pi)$.

Similar to the results in blackboard model, we sometimes need to consider the conditional information cost and conditional information complexity, which are defined as follows.

Definition 5. The conditional information cost of a protocol \mathcal{P} with respect to distribution $\mathbf{X} \sim \boldsymbol{\mu}$ and $\mathbf{Y} \sim \boldsymbol{\zeta}$ is $\text{ICost}_{\boldsymbol{\mu},\boldsymbol{\zeta}}(\mathbf{X}; \Pi|\mathbf{Y}) = \sum_{i=1}^m I_{\boldsymbol{\mu},\boldsymbol{\zeta}}(\mathbf{X}; \Pi_i|\mathbf{Y})$. The δ -error conditional information complexity with respect to function f and distribution $\mathbf{X} \sim \boldsymbol{\mu}$ and $\mathbf{Y} \sim \boldsymbol{\zeta}$ is $\text{IC}_{\boldsymbol{\mu},\boldsymbol{\zeta},\delta}(f|\mathbf{Y}) = \min_{\mathcal{P} \in \Gamma_{\delta}} \text{ICost}_{\boldsymbol{\mu}}(\mathbf{X}; \Pi|\mathbf{Y})$.

Given the modified definition of information cost and information complexity, we now rephrase the direct sum theorem in the context of private messages model as follows.

Theorem 1 (Direct Sum Theorem). Recall that $f : \{0,1\}^T \mapsto \{0,1\}$ is a decomposable function with primitive function $h : \{0,1\}^t \mapsto \{0,1\}$. Suppose the input distribution $\mathbf{X} \sim \boldsymbol{\mu} = \nu^n$ is a collapsing distribution and random variable $\mathbf{Y} \sim \boldsymbol{\zeta} = \eta^n$ partitions \mathbf{X} . Consider a random partitioned communication game with respect to function f and distribution Σ , then $\text{IC}_{\boldsymbol{\mu},\boldsymbol{\zeta},\delta,\Sigma}(f|\mathbf{Y}) \geq \sum_{i=1}^n \text{IC}_{\nu,\eta,\delta,\Sigma}(h|Y_i)$.

The proof of Theorem 1 is an analogue of the proof by Bar-Yossef et. al. [3], and can be found in the full version [10].

4 Near Optimal Lower Bound for Frequency Moments

In this section, we will prove the following asymptotically optimal space lower bound for computing the k^{th} frequency moments for $k > 2$.

Theorem 2. *Suppose ϵ and δ are small constants. If an algorithm correctly gives a $(1 + \epsilon)$ -approximation of the k^{th} frequency moment of n numbers with probability at least $1 - \delta$ in a random order stream within ℓ passes, then the space it needs is at least $\Omega(n^{1-2/k}/\ell)$.*

We consider the decomposable function $\text{SETDISJ}_{n,t}$. The intuition is the following. Suppose $m = td$ and we shall assume that m is large enough such that if the allocation $\sigma \sim \mathcal{U}_{t,m}$, then with probability $1 - o(1)$ we have $\sigma(i) \neq \sigma(j)$ for all $1 \leq i \neq j \leq t$. Consider a collapsing and symmetric distribution $\mathbf{X} \sim \nu$ partitioned by $Y \sim \eta$, where η is a uniform distribution over $[t]$ and conditioned on $Y_i = j$ we have $\mathbf{X}_i = e^j$ with probability $1/2$ and $\mathbf{X}_i = 0$ with probability $1/2$. From Theorem 1, it suffices to prove lower bound for the primitive function. Recall that the information complexity for AND_t is at least $\text{IC}^B = \Omega(1/t)$ in a blackboard fixed-partition t -player communication game with respect to this input distribution [7,9]. Conditioned on a particular allocation σ , suppose the indexes of the players who get the t bits of the input \mathbf{X}_i are $i_1 < i_2 < \dots < i_t$. We can imagine that these t players play a communication game to compute the function value of AND_t and only the messages these t players receive contribute to the information cost. So the effective information cost is

$$I(\mathbf{X}_i; \Pi_{i_1-1}|Y_i) + I(\mathbf{X}_i; \Pi_{i_2-1}|Y_i) + \dots + I(\mathbf{X}_i; \Pi_{i_t-1}|Y_i) .$$

Now we use the simple fact that the information cost in private messages model is at least as large as the information cost in blackboard model. We get that the above information cost is at least IC^B . Note that for each $1 \leq \ell \leq t$, player $i_\ell + 1, i_\ell + 2, \dots, i_{\ell+1} - 1$ do not have any bit of the input \mathbf{X}_i , we have

$$I(\mathbf{X}_i; \Pi_{i_\ell}|Y_i) \geq I(\mathbf{X}_i; \Pi_{i_{\ell+1}}|Y_i) \geq \dots \geq I(\mathbf{X}_i; \Pi_{i_{\ell+1}-1}|Y_i) .$$

Since the expected distance between i_j and i_{j+1} is d , the next lemma is intuitive.

Lemma 1. *Suppose $\mathbf{X}_i \sim \nu$ is a collapsing symmetric distribution partitioned by $Y_i \sim \eta$, then the information cost of computing the function value of AND_t with small constant error rate δ is at least $\text{IC}(\text{AND}_t|Y_i) = \Omega(d/t)$.*

Now we formally prove this key lemma. Given an allocation $\sigma : [t] \rightarrow [m], m = td$, let $\sigma(\ell)$ be the image of ℓ , and $\pi(\ell)$ be the smallest $\sigma(\ell')$ such that $\ell' \in [t] \setminus \{\ell\}$ and $\sigma(\ell') \geq \sigma(\ell)$ (if $\sigma(\ell) = \max_{\ell' \in [t]} \sigma(\ell')$ then $\pi(\ell) = \min_{\ell' \in [t]} \sigma(\ell') + m$). Let p_j denote the probability that $\pi(\ell) - \sigma(\ell) = j$ when $\sigma \sim \mathcal{U}_{t,m}$. We have $p_j = (t/m)(1 - j/m)^{t-1} = (1 - j/m)^{t-1}/d$. We first prove the following lemmas.

Lemma 2. *For any $0 \leq i < j \leq m - 1$,*

$$p_j(p_i + p_{i+1} + \dots + p_{m-1}) \geq p_i(p_j + p_{j+1} + \dots + p_{m-1}) .$$

Proof. Consider the function $p(x) = (1-x)^{t-1}/d$. It is easy to verify that this function is log-concave. Note that $i < j$ and $i \leq i+k < j+k$ for $k \geq 0$, we get that $p_j p_{i+k} \geq p_i p_{j+k}$ and thus $p_{i+k}/p_i \geq p_{j+k}/p_j$. So

$$\frac{p_i + p_{i+1} + \cdots + p_{m-1}}{p_i} \geq \frac{p_i + p_{i+1} + \cdots + p_{i+m-j-1}}{p_i} \geq \frac{p_j + p_{j+1} + \cdots + p_{m-1}}{p_j} . \quad \square$$

Lemma 3. *If $c_1 \geq c_2 \geq \cdots \geq c_{m-1} \geq 0$, then*

$$\sum_{i=1}^{m-1} \sum_{j=i}^{m-1} p_j c_i \geq \sum_{i=1}^{m-1} i p_i \sum_{j=1}^{m-1} p_j c_j .$$

Proof. Note $\sum_{j=1}^{m-1} p_j = 1$. Now,

$$\begin{aligned} & \sum_{i=1}^{m-1} \sum_{j=i}^{m-1} p_j c_i - \sum_{i=1}^{m-1} i p_i \sum_{j=1}^{m-1} p_j c_j = \sum_{j=1}^{m-1} \sum_{i=1}^j p_j c_i - \sum_{i=1}^{m-1} i p_i \sum_{j=1}^{m-1} p_j c_j \\ &= \sum_{i=1}^{m-1} \sum_{j=1}^{m-1} \sum_{\ell=1}^i p_i p_j c_\ell - \sum_{i=1}^{m-1} \sum_{j=1}^{m-1} \sum_{\ell=1}^i p_i p_j c_j \\ &= \sum_{i=1}^{m-1} \sum_{j=1}^{m-1} \sum_{\ell=1}^i p_i p_j (c_\ell - c_j) = \sum_{j=1}^{m-1} \sum_{\ell=1}^{m-1} \sum_{i=\ell}^{m-1} p_i p_j (c_\ell - c_j) \\ &= \sum_{\ell < j} [p_j (p_\ell + \cdots + p_{m-1}) - p_\ell (p_j + \cdots + p_{m-1})] (c_\ell - c_j) \geq 0 . \end{aligned}$$

The last step follows from Lemma 2. \square

Proof (of Lemma 1). Suppose \mathcal{P} is a δ -error protocol and Π is its transcript. Let $1 \leq \ell \leq t$. Let c_j denote the expected communication cost contributed by player $\sigma(\ell) + j - 1$ if $\pi(\ell) - \sigma(\ell) \geq j$, that is, $c_j = I(\mathbf{X}_i; \Pi_{\sigma(\ell)+j-1} | Y_i, \pi(\ell) - \sigma(\ell) \geq j)$. Since we consider private messages model we shall have that $c_1 \geq c_2 \geq \cdots \geq c_{m-1} \geq 0$. By Lemma 3 we get that

$$\sum_{i'=1}^{m-1} c_{i'} \sum_{j=i'}^{m-1} p_j = \sum_{i'=1}^{m-1} \sum_{j=i'}^{m-1} p_j c_{i'} \geq \sum_{i'=1}^{m-1} i' p_{i'} \sum_{j=1}^{m-1} p_j c_j . \quad (1)$$

Note that $\sum_{j=i'}^{m-1} p_j$ is the probability that $\pi(\ell) - \sigma(\ell) \geq i'$. The left-hand side of Equation 1 is the communication cost contributed by players $\sigma(\ell), \sigma(\ell) + 1, \dots, \pi(\ell) - 1$. The first term on the right-hand side $\sum_{i'=1}^{m-1} i' p_{i'}$ is the expected distance between $\sigma(\ell)$ and $\pi(\ell)$, which equals d . The second term on the right-hand side $\sum_{j=1}^{m-1} p_j c_j$ is the information cost contributed by player $\pi(\ell) - 1$. So we have $\sum_{j=\sigma(\ell)}^{\pi(\ell)-1} I(\mathbf{X}_i; \Pi_j | Y_i) \geq d \cdot I(\mathbf{X}_i; \Pi_{\pi(\ell)-1} | Y_i)$. Recall that $\sum_{\ell=1}^t I(\mathbf{X}_i; \Pi_{\pi(\ell)-1} | Y_i) \geq \text{IC}^B = \Omega(1/t)$. We have

$$\text{ICost}(\mathbf{X}_i; \Pi | Y_i) = \sum_{\ell=1}^t \sum_{j=\sigma(\ell)}^{\pi(\ell)-1} I(\mathbf{X}_i; \Pi_j | Y_i) \geq \sum_{\ell=1}^t d \cdot I(\mathbf{X}_i; \Pi_{\pi(\ell)-1} | Y_i) = \Omega\left(\frac{d}{t}\right) .$$

Since the above result is true for any δ -error protocol, we prove Lemma 1. \square

Remark 1. We realize the reduction technique we introduce in Lemma 1, 2, and 3 works for other decomposable functions if we can prove information complexity lower bound for some symmetric collapsing input distribution.

Using the direct sum theorem we have the following corollaries.

Corollary 1. *If a protocol \mathcal{P} correctly computes the value of $\text{SETDISJ}_{n,t}$ with probability at least $1 - \delta$ in a random partition communication game, then the total communication complexity is at least*

$$\text{CC}(\text{SETDISJ}_{n,t}) \geq \sum_{i=1}^n \text{IC}(\text{AND}_t | Y_i) = \Omega\left(\frac{nd}{t}\right) = \Omega\left(\frac{nm}{t^2}\right) .$$

Now we can prove Theorem 2 via the a reduction as follows.

Proof (of Theorem 2). Suppose an algorithm gives $(1 + \epsilon)$ -approximation of the k^{th} frequency moment using s bits of space and within ℓ passes. Consider the following ℓ -round protocol which compute the function value of $\text{SETDISJ}_{n,t}$ when $t = (5\epsilon \cdot n)^{1/k}$. Set m to be large¹, $m = \Omega(t^2)$, which rules out collisions with constant probability. Each player shall receive some bits of the input. For each bit of value 1, that indicates some value v in one of the set. And the player take that as probing a number v in the data stream. The first player runs the algorithm on the inputs she receives, then sends the s bits of memory and another $O(\log n)$ bits that indicates the number of 1's she receives to the second player. The second player continues the algorithm on her own inputs, then sends the memory bits and the number of 1's the first two players receive to the third player. And so on and so forth.

Now assume the number of 1's in the input is n' , we get that $n' < n + t < (1 + \epsilon)n$. If the function value of $\text{SETDISJ}_{n,t}$ is 1, then one of the value appears t times in the data stream. So the frequency moment is at least $(n' - t) + t^k = n' - t + 5\epsilon \cdot n \geq n' + 4\epsilon \cdot n > n'(1 + \epsilon)^2$. On the other hand, if the function value of $\text{SETDISJ}_{n,t}$ is 0, then the frequency moment is n' . Therefore, if the last player claims the function value is 1 if the the frequency moment given by the algorithm is at least $(1 + \epsilon)n'$ and claims the function value is 0 otherwise, she will be correct with probability at least $1 - \delta$.

The total communication complexity of this protocol is $O(\ell m(s + \log n))$. Recall that this value is at least $\Omega(nm/t^2)$, we get that

$$s = \Omega\left(\frac{n}{t^2 \ell}\right) = \Omega\left(\frac{n^{1-2/k}}{\ell}\right) . \quad \square$$

¹ Note that the private communication model allows a large number of players, say even one corresponding to each input, which is one of the reasons for getting the improved space lower bounds for streaming algorithms compared to the blackboard model.

5 Entropy–Space Tradeoff for L_∞ and L_p Distances

In this section, we consider the entropy–space tradeoff of finding an n^ϵ -approximation of the L_∞ distance.

We consider the following communication game. The two vectors correspond to $\langle x_1, x_3, \dots, x_{2n-1} \rangle$ and $\langle x_2, x_4, \dots, x_{2n} \rangle$ (we can use any fixed permutation). There are $2n$ players. The input allocation $\sigma : [2n] \mapsto [2n]$ is drawn from a distribution over all permutations of $[2n]$. The entire input x_i is allocated to player $\sigma(i)$. The players then communicate in the private messages model in order to compute the function value of $\text{GAPDIST}_{n,\ell}$.

We shall show the following theorems.

Theorem 3. *Let $\delta > 0$ be a small constant. Let Σ be a distribution of input order with entropy E . Any δ -error n^ϵ -approximation algorithm for L_∞ distances with respect to input order distribution Σ requires space at least*

$$\Omega \left(\frac{n^{1-4\epsilon}}{2^{(2n \log n - E)/(1-2\delta)n}} \right) .$$

Theorem 4. *Theorem 3 is tight, given E there exists an order distribution Σ' with entropy at least E , and a δ -error n^ϵ -approximation algorithm of L_∞ distance with respect to Σ' , using $O \left(\frac{n^{1-4\epsilon}}{2^{(2n \log n - E)/n}} \right)$ space.*

Proof (of Theorem 3). We consider the function $\text{GAPDIST}_{n,\ell}$. Recall that the function BIGAP_ℓ is defined as: $\text{BIGAP}_\ell(x, y) = 1$ when $|x - y| = \ell$ and $\text{BIGAP}_\ell(x, y) = 0$ when $|x - y| = 0, 1$. The decomposable function $\text{GAPDIST}_{n,\ell}$ is defined as $\text{GAPDIST}_{n,\ell} = \text{OR}_n(\text{BIGAP}_\ell(x_1, x_2), \dots, \text{BIGAP}_\ell(x_{2n-1}, x_{2n}))$. If an algorithm can correctly compute the L_∞ distance of two n dimensional vectors up to a n^ϵ factor, then it shall be able to distinguish whether the L_∞ distance is at most 1 or the L_∞ distance is at least $n^{2\epsilon}$. Therefore, the space needed by such an algorithm is as large as the space needed to compute the function value of $\text{GAPDIST}_{n,n^{2\epsilon}}$ with probability at least $1 - \delta$. Hence to prove a space lower bound for computing the L_∞ distances, it suffices to show strong lower bound for the communication complexity of $\text{GAPDIST}_{n,\ell}$.

We shall consider the following input distribution of $\text{GAPDIST}_{n,\ell}$. For each $1 \leq i \leq n$, $Y_i \sim \eta$ is randomly drawn from $[2\ell]$. Conditioned on $Y_i = 2j + 1$, $0 \leq j < \ell$, $X_{2i-1} = j$ and X_{2i} is uniformly distributed in $\{j, j + 1\}$. Conditioned on $Y_i = 2j$, $1 \leq j \leq \ell$, X_{2i-1} is uniformly distributed in $\{j, j - 1\}$ and $X_{2i} = j$. It is clear that $\mathbf{X} \sim \mu = \nu^n$ is a collapsing distribution since we always have the value of each primitive function is $\text{BIGAP}_\ell = 0$. Bar-Yossef et. al. [3] shows the following lower bound for the primitive function BIGAP_ℓ in the literature of blackboard model:

Lemma 4 (Lemma 8.2 in [3]). *Suppose $0 < \delta < 1/4$ is a constant, the two-party communication complexity of computing the function value of BIGAP_ℓ with probability $1 - \delta$ is $\text{IC}^B = \Omega(1/\ell^2)$.*

Now we consider the information complexity lower bound for the i^{th} primitive function BIGAP_ℓ in the private messages model. Suppose player u and player v receive the input X_{2i-1} and X_{2i} . Effectively these two players play a communication game to compute the primitive function and Π_{u-1} and Π_{v-1} are the effective transcripts. So from Lemma 4 we get that $I(X_{2i-1}, X_{2i}; \Pi_{u-1}) + I(X_{2i-1}, X_{2i}; \Pi_{v-1}) = \Omega(1/\ell^2)$. Moreover, we shall have $I(X_{2i-1}, X_{2i}; \Pi_u) \geq I(X_{2i-1}, X_{2i}; \Pi_{u+1}) \geq \dots \geq I(X_{2i-1}, X_{2i}; \Pi_{v-1})$ as well as $I(X_{2i-1}, X_{2i}; \Pi_v) \geq I(X_{2i-1}, X_{2i}; \Pi_{v+1}) \geq \dots \geq I(X_{2i-1}, X_{2i}; \Pi_{u-1})$. So the information cost in private messages model is at least $\Omega(\min\{|u-v|, n-|u-v|\}/\ell^2)$. If we can prove with some constant probability the value of $\min\{|u-v|, n-|u-v|\}$ is large and the protocol correctly gets the function value of BIGAP_ℓ , then we shall have a lower bound for the primitive function.

Suppose E_i is the entropy the allocation distribution for the i^{th} primitive function. We let d' denote the value $n/2^{(2n \log 2n - E_i)/(1-2\delta)}$ for the sake of convenience. We shall prove by contradiction that with probability at least 2δ , $\min\{|u-v|, n-|u-v|\} \geq d'$.

Suppose not. Note that the total number of different allocations for a primitive function $\sigma_i : [2] \mapsto [2n]$ is $2n(2n-1)$, and the number of different allocations such that $\min\{|u-v|, n-|u-v|\} \geq d'$ is $2n(2n-2d'+1)$.

Hence if the probability of getting an allocation $\sigma_i \sim \Sigma_i$ satisfying $\min\{|u-v|, n-|u-v|\} \geq d'$ is at most 2δ , then the entropy of distribution is

$$\begin{aligned} E_i &< 2\delta \log 2n(2n-2d'+1) + (1-2\delta) \log (2n(2n-1) - 2n(2n-2d'+1)) \\ &< 2 \log 2n + (1-2\delta) \log \left(\frac{d'}{n} \right) . \end{aligned}$$

Thus we have $d > n/2^{(2 \log 2n - E_i)/(1-2\delta)}$, a contradiction. Therefore, the information cost for the i^{th} primitive function is at least

$$\Omega(d'/\ell^2) = \Omega \left(\frac{n}{\ell 2^{(2 \log 2n - E_i)/(1-2\delta)}} \right) .$$

Note that we shall have $\sum_{i=1}^n E_i \geq E$ and the function 2^x is convex. Using Theorem 1 and Jensen's inequality we get that

$$\text{IC}(\text{GAPDIST}_{n,\ell} | \mathbf{Y}) = \sum_{i=1}^n \text{IC}(\text{BIGAP}_\ell | Y_i) \geq \Omega \left(\frac{n^2}{\ell 2^{(2n \log n - E)/(1-2\delta)/n}} \right) .$$

Therefore, to compute the function value of $\text{GAPDIST}_{n,n^{2\epsilon}}$ or to compute the L_∞ distance of two n -dimensional vectors up to a n^ϵ factor, we shall need the memory space to be

$$\Omega \left(\frac{n^{1-4\epsilon}}{2^{(2n \log n - E)/(1-2\delta)n}} \right) . \quad \square$$

Proof (of Theorem 4). We let d denote the value $c \cdot n/2^{(2n \log n - E)/n}$ for the sake of convenience, where c is a large constant, then we shall have $\log d =$

$\log c + E/n - n \log n$. Consider the distribution of allocations σ generated by the following algorithm:

- 1: Pick a random permutation π of $[n]$.
- 2: Let $\sigma(2j - 1) = 2\pi(j) - 1$ for $1 \leq j \leq n$.
- 3: **for all** $1 \leq i \leq n/d$ **do**
- 4: Pick a random permutation π_i of $[d]$
- 5: Let $\sigma(2d \cdot i + 2j) = 2\pi(d \cdot i + \pi_i(j))$ for $1 \leq j \leq d$.
- 6: **end for**

This allocation distribution is a uniform distribution over $n!(d!)^{n/d}$ different allocations. So the entropy is $n \log n + (n/d) \cdot d \log d + O(n) > E$ for large c . Here we use the following simple corollary of Stirling's approximation for factorials.

Lemma 5. *Suppose $n > 0$ is a positive integer, then*

$$\log(n!) = n \log n + O(n) .$$

For each allocation in this distribution, the first $2d$ numbers are the inputs of d dimensions, and the next $2d$ numbers are the inputs of another d dimensions, and so on and so forth. Therefore, we can divide the original problem into n/d subproblems of computing the L_∞ distance for d dimensional vectors. And the space can be reused for each subproblem. Saks and Sun [15] showed these subproblems can be resolve using only $O(d/n^{4\epsilon})$ space. So we can n^ϵ -approximate the L_∞ distance using $O(d/n^{4\epsilon}) = O(n^{1-4\epsilon}/2^{(2n \log n - E)/n})$ of space. \square

Using a reduction proposed by Saks and Sun [15] we get the following entropy space tradeoff for approximating L_p distances.

Theorem 5. *Let $\delta > 0$ be a small constant and $p > 2$. Let Σ be a distribution of input order with entropy E . Any δ -error n^ϵ -approximation algorithm for L_p distances with respect to input order distribution Σ requires space*

$$\Omega \left(\frac{n^{1-2/p-4\epsilon}}{2^{(2n \log n - E)/(1-2\delta)n}} \right) .$$

References

1. Alon, N., Matias, Y., Szegedy, M.: The Space Complexity of Approximating the Frequency Moments. *Journal of Computer and System Sciences* 58(1), 137–147 (1999)
2. Andoni, A., McGregor, A., Onak, K., Panigrahy, R.: Better Bounds for Frequency Moments in Random-Order Streams. *Arxiv preprint arXiv:0808.2222* (2008)
3. Bar-Yossef, Z., Jayram, T.S., Kumar, R., Sivakumar, D.: An information statistics approach to data stream and communication complexity. *Journal of Computer and System Sciences* 68(4), 702–732 (2004)
4. Bhuvanagiri, L., Ganguly, S., Kesh, D., Saha, C.: Simpler algorithm for estimating frequency moments of data streams. In: *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pp. 708–713. ACM, New York (2006)

5. Chakrabarti, A., Cormode, G., McGregor, A.: Robust lower bounds for communication and stream computation. In: Proceedings of the fortieth annual ACM symposium on Theory of computing, pp. 641–650. ACM Press, New York (2008)
6. Chakrabarti, A., Jayram, T.S., Patrascu, M.: Tight lower bounds for selection in randomly ordered streams. In: Proceedings of the nineteenth annual ACM-SIAM Symposium on Discrete Algorithms, pp. 720–729. Society for Industrial and Applied Mathematics, Philadelphia (2008)
7. Chakrabarti, A., Khot, S., Sun, X.: Near-optimal lower bounds on the multi-party communication complexity of set disjointness. In: IEEE Conference on Computational Complexity, pp. 107–117 (2003)
8. Gal, A., Gopalan, P.: Lower Bounds on Streaming Algorithms for Approximating the Length of the Longest Increasing Subsequence. In: 48th Annual IEEE Symposium on Foundations of Computer Science, 2007. FOCS 2007, pp. 294–304 (2007)
9. Gronemeier, A.: Asymptotically optimal lower bounds on the nlh -multi-party information. In: 26th International Symposium on Theoretical Aspects of Computer, p. 505 (2009)
10. Guha, S., Huang, Z.: Revisiting the direct sum theorem and space lower bounds for random order streams. Technical Report (2009), http://repository.upenn.edu/cis_papers/
11. Guha, S., McGregor, A.: Tight lower bounds for multi-pass stream computation via pass elimination. In: Aceto, L., Damgård, I., Goldberg, L.A., Halldórsson, M.M., Ingólfssdóttir, A., Walukiewicz, I. (eds.) ICALP 2008, Part I. LNCS, vol. 5125, pp. 760–772. Springer, Heidelberg (2008)
12. Guha, S., McGregor, A.: Stream-Order and Order-Statistics: Quantile Estimation in Random-Order Streams. *SIAM Journal of Computing* 38(5), 2044–2059 (2009)
13. Indyk, P., Woodruff, D.: Optimal approximations of the frequency moments of data streams. In: Proceedings of the thirty-seventh annual ACM symposium on Theory of computing, pp. 202–208. ACM, New York (2005)
14. Kushilevitz, E., Nisan, N.: *Communication Complexity*. Cambridge University Press, Cambridge (1996)
15. Saks, M., Sun, X.: Space lower bounds for distance approximation in the data stream model. In: Proceedings of the thirty-fourth annual ACM symposium on Theory of computing, pp. 360–369. ACM, New York (2002)
16. Yao, A.C.: Some complexity questions related to distributed computing. In: Proceedings of the 11th Annual ACM Symposium on Theory of Computing, pp. 209–213 (1979)