Optimal Communication Complexity of Chained Index

Janani Sundaresan
Cheriton School of Computer Science, University of Waterloo, Canada

Abstract

We study the CHAIN communication problem introduced by Cormode et al. [ICALP 2019]. For $k \geq 1$, in the $\text{CHAIN}_{n,k}$ problem, there are k string and index pairs (X_i, σ_i) for $i \in [k]$ such that the value at position σ_i in string X_i is the same bit for all k pairs. The input is shared between $k+1$ players as follows. Player 1 has the first string $X_1 \in \{0,1\}^n$, player 2 has the first index $\sigma_1 \in [n]$ and the second string $X_2 \in \{0,1\}^n$, player 3 has the second index $\sigma_2 \in [n]$ along with the third string $X_3 \in \{0,1\}^n$, and so on. Player $k+1$ has the last index $\sigma_k \in [n]$. The communication is one way from each player to the next, starting from player 1 to player 2, then from player 2 to player 3 and so on. Player $k+1$, after receiving the message from player k, has to output a single bit which is the value at position σ_i in X_i for any $i \in [k]$. It is a generalization of the well-studied INDEX problem, which is equivalent to $\text{CHAIN}_{n,2}$.

Cormode et al. proved that the $\text{CHAIN}_{n,k}$ problem requires $\Omega(n/k^2)$ communication, and they used it to prove streaming lower bounds for the approximation of maximum independent sets. Subsequently, Feldman et al. [STOC 2020] used it to prove lower bounds for streaming submodular maximization. However, it is not known whether the $\Omega(n/k^2)$ lower bound used in these works is optimal for the problem, and in fact, it was conjectured by Cormode et al. that $\Omega(n)$ bits are necessary.

We prove the optimal lower bound of $\Omega(n)$ for $\text{CHAIN}_{n,k}$ when $k = o(n/\log n)$ as our main result. This settles the open conjecture of Cormode et al., barring the range of $k = \Omega(n/\log n)$. The main technique is a reduction to a non-standard INDEX problem where the input to the players is such that the answer is biased away from uniform. This biased version of INDEX is analyzed using tools from information theory. As a corollary, we get an improved lower bound for approximation of maximum independent set in vertex arrival streams via a reduction from CHAIN directly.

2012 ACM Subject Classification Theory of computation → Communication complexity

Keywords and phrases communication complexity, index communciation problem

Digital Object Identifier 10.4230/LIPIcs.ITCS.2025.89

Related Version *Full Version*: https://arxiv.org/abs/2404.07026

Funding *Janani Sundaresan*: Supported in part by Sepehr Assadi's Sloan Research Fellowship and startup grant from University of Waterloo.

Acknowledgements The author is thankful to Sepehr Assadi for introducing them to the problem and for insightful discussions on the proof. The author would also like to thank Parth Mittal for useful comments, Christian Konrad for introducing them to the Augmented Chain problem, and the anonymous reviewers of ITCS 2025 for helpful comments and suggestions. The author is very grateful to Mi-Ying Huang, Xinyu Mao, Guangxu Yang and Jiapeng Zhang for an illuminating discussion about the problem. They pointed out an important flaw in an earlier version of this work, and the discussion was instrumental for the new proofs in the current version.

1 Introduction

The INDEX problem is one of the foundational problems in communication complexity. For $n \geq 1$, in the INDEX_n problem, there are two players Alice and Bob. Alice has a string $X \in \{0,1\}^n$ and Bob has an index $\sigma \in [n]$, and Bob has to output the value of X at position σ.

If the communication is one-way from Alice to Bob, it is easy to show that Alice needs to send $\Omega(n)$ bits to get any constant advantage [1, 30]. This problem has been well-studied in multiple settings, and we know tight trade-offs in the two-party communication model for communication complexity [34], information complexity [24], and quantum communication complexity [8, 24].

Among the numerous applications of communication complexity, one that is of interest to us is proving lower bounds for streaming algorithms. INDEX and its variants, in particular, have been quite useful in this context, for example, in [23, 18, 20, 21, 16, 10]. This is by no means an exhaustive list.

In this paper, we study a natural generalization of INDEX, called chained index ($\text{CHAIN}_{n,k}$ for $n, k \geq 1$) introduced by [12]. There are k different instances of INDEX_n, correlated so that they have the same answer. They are "chained" together, where each player holds the index to the previous instance, and also the string for the next instance.

▶ **Definition 1** (Informal). *In $\text{CHAIN}_{n,k}$, there are k instances of INDEX_n, all with the same answer. Players 1 and 2 take on the role of Alice and Bob respectively in the first instance, players 2 and 3 take on the role of Alice and Bob respectively for the second instance, and so on, all the way till players k and $k+1$ for the last instance.*

Communication is one-way from each player to the next in ascending order. The last player has to output the answer. The communication cost is the total number of bits in all the messages sent by the players. See Figure 1 for an illustration.

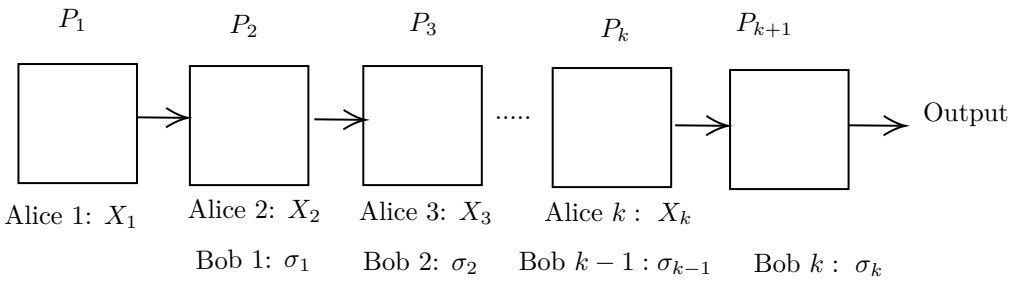

Figure 1 An illustration of the $\text{CHAIN}_{n,k}$ problem with k correlated sub-instances of INDEX_n from Definition 1. The arrows illustrate that the message is from P_i to P_{i+1} for $i \in [k]$.

In [12], a reduction from CHAIN was employed to get a lower bound for approximation of maximum independent sets in vertex arrival streams. Before its introduction, [28] used the problem implicitly to get a lower bound of $(1 - 1/e)$ in the approximation factor for maximum matching in $\tilde{O}(n)$ space in vertex arrival streams.

The problem has been used by the breakthrough result of [19] to study the multi-party communication complexity of submodular maximization. They proved that any randomized p-party protocol which maximizes a monotone submodular function $f : \{0,1\}^N \to \mathbb{R}$, subject to a cardinality constraint of at most p and an approximation factor of at least $(1/2 + \epsilon)$, uses $\Omega(N\epsilon/p^3)$ communication. This also gave a lower bound for streaming submodular maximization. The CHAIN problem was used by [17] also for similar purposes, but subject to stronger matroid constraints. [7] used a reduction from CHAIN to prove lower bounds for interval independent set selection in streams of split intervals.

We do not know tight bounds for the communication complexity of the CHAIN problem, despite finding varied applications of it. There is a trivial protocol of $O(n)$ bits, where any player can send the entire string to the next player who holds the index. Another simple

protocol is for each player to send $O(n/k)$ bits randomly sampled from their strings using public randomness, and with constant probability, in at least one of the k instances, we send the special bit to the player holding the index. However, this still takes $\Omega(n)$ total bits of communication.

In [12], they prove a lower bound of $\Omega(n/k^2)$ for any $k \geq 1$ through a reduction from conservative multi-party pointer jumping problem, introduced by [14]. They state without proof that a stronger lower bound of $\Omega(n/k)$ can be obtained for a restricted range of $k \leq O((n/\log n)^{1/4})$. They posed the following conjecture on the optimal communication lower bound.

▶ **Conjecture 2 ([12]).** *Any protocol that solves* $\textsc{chain}_{n,k}$ *requires* $\Omega(n)$ *bits of communication.*

[19] made some progress on Conjecture 2 by showing that among all the messages sent by the players, there is at least one message with $\Omega(n/k^2)$ bits for every $k \geq 1$. But the original conjecture is still open, and this is the focus of our work.

1.1 Our Results

We settle Conjecture 2 almost fully by proving the optimal lower bound of $\Omega(n)$ barring the corner case of when k is too large. As far as we know, this corner case is not a focus for existing reductions from $\textsc{chain}_{n,k}$.

▶ **Theorem 3.** *For any* $n, k \geq 1$*, any protocol for* $\textsc{chain}_{n,k}$ *with probability of success at least 2/3, requires* $\Omega(n - k\log n)$ *total bits of communication.*

Therefore, as long as $k = o(n/\log n)$, we get the optimal $\Omega(n)$ lower bound from Theorem 3.

The proof of Theorem 3 can be found in Section 3. We prove the lower bound in the more general blackboard model of communication, instead of private messages between players (see Section 2.1 for details). The main idea is to analyze $\textsc{index}_n$ where Alice and Bob already have *some prior advantage* in guessing the answer. We elaborate on our techniques in Section 1.2.

As a direct corollary of Theorem 3, we get improvements in streaming lower bounds in [12, 19] through reductions from CHAIN immediately. In particular, we get that any algorithm which α-approximates the size of a maximum independent set in vertex arrival streams requires $\Omega(n^2/\alpha^5 - \log n)$ space, while the previous bound was $\Omega(n^2/\alpha^7)$ in [12]. We present the implications of our result in Section 4.

A further generalization of CHAIN, called Augmented Chain was defined in [15]. Here, instances of Augmented Index are chained together instead. Our lower bound of $\Omega(n - k\log n)$ can be extended to Augmented Chain also, and the details are covered in Section 3.4.

1.2 Our Techniques

In this subsection, we give an overview of the challenges in proving the lower bound and a summary of our techniques. We start by going over the prior techniques.

Prior Techniques

We will briefly talk about the technique used in [19] to prove that there is at least one player who sends $\Omega(n/k^2)$ bits for any $k \geq 1$. We will argue that these techniques can be extended to proving a lower bound of $\Omega(n/k)$ for the total number of bits, but not all the way to $\Omega(n)$ bits.

The first step in proving a lower bound for $\text{CHAIN}_{n,k}$ in [19] is a decorrelation step – the k instances of INDEX_n have the same answer, and they remove this correlation with a hybrid argument. These arguments have been used extensively in the literature (see e.g., [26, 3, 27, 6]). Intuitively, any protocol that tries to solve $\text{CHAIN}_{n,k}$ may attempt to solve "many" of the instances of INDEX_n, albeit each with a "small" advantage over $1/2$, in the hope that the "small" advantages may accrue to get a constant probability of success overall (taking advantage of the fact that all the instances of INDEX_n have the same answer). The hybrid argument is a way to reduce proving a lower bound on the overall problem, to proving a lower bound for k different INDEX_n problems against these low advantages.

Let us assume, for simplicity, that the protocol tries to get an advantage of $\Omega(1/k)$ in each instance of INDEX_n, to get a constant total advantage. We can prove that any protocol that gets an advantage of $\Omega(1/k)$ for INDEX_n uses $\Omega(n/k^2)$ bits of communication using basic tools from information theory [9] (this is quite standard, see e.g., [2] for a direct proof), and this is known to be tight. Therefore, for k instances, we get a lower bound of $\Omega(n/k)$ bits in total. Now, we will argue why this is not the optimal lower bound.

On one hand, for INDEX_n, it is known that for any $\delta \in (0, 1/2)$, there is a protocol that uses $O(n\delta^2)$ bits of communication to get a probability of success $1/2 + \delta$. This means that INDEX_n can be solved with advantage $\Omega(1/k)$ in $O(n/k^2)$ bits; in other words, each "hybrid step" of the previous lower bound argument is optimal. So, then, why can we not get a good protocol for $\text{CHAIN}_{n,k}$ by running the protocol for INDEX_n with $\delta = 1/k$ on all k instances? This protocol would have $O(n/k)$ bits of communication in total. The reason is that the $1/k$ small advantages do not add up as we would like them to. To illustrate this, we will briefly talk about the protocol that gets δ advantage in $O(n\delta^2)$ bits for INDEX_n.

Protocol for INDEX_n

First, we will sketch a protocol for $\delta = 1/\sqrt{n}$ that uses $O(1)$ bits of communication. Let us imagine that the input X is chosen uniformly at random from $\{0,1\}^n$ and the index σ is chosen uniformly at random from $[n]$. Then, Alice finds the majority bit from her string X and sends it to Bob. Bob just outputs the bit sent by Alice. We know, from simple anti-concentration bounds on the binomial distribution, that the number of indices with the majority bit in X is at least $n/2 + c\sqrt{n}$ with constant probability for some appropriate constant c. The protocol succeeds if Bob holds the index σ to a majority bit, and this happens with probability at least $\frac{1}{2} + \frac{c}{\sqrt{n}}$.

The assumption that the input X is chosen uniformly at random from $\{0,1\}^n$ can be removed using public randomness. Alice and Bob collectively sample a random string $A \in \{0,1\}^n$, and Alice changes her input to $X \oplus A$ so that each bit is 0 or 1 with equal probability. Similarly, the assumption that the index σ is chosen uniformly at random from $[n]$ can be removed by Alice and Bob sampling a random permutation of $[n]$. Alice permutes string X according to this permutation, and Bob changes his input to the index that the permutation maps σ to. This is termed as the self-reducibility property of INDEX.

We can also extend the protocol to any δ by partitioning the string X into $n\delta^2$ blocks at random using public randomness and sending the majority bit in each block. Bob knows the block that his input index σ belongs to as public randomness is used.

Challenges for $\text{CHAIN}_{n,k}$

If we use the protocol we described for $\text{CHAIN}_{n,k}$, we are left with k bits from each instance of INDEX_n, which may be the correct answer to the problem with probability $1/2 + \Theta(1/k)$, but, the variance of each of these bits is $1/4 - \Theta(1/k^2)$. We have a protocol for $\text{CHAIN}_{n,k}$,

where, out of the k instances of INDEX_n, it finds the right answer for $\approx k/2 + \Theta(1)$ instances in expectation. However, the standard deviation of the number of right answers is $\approx \sqrt{k}/2$, which is enough to mask the $\Theta(1)$ improvement over $k/2$ we get in expectation. If we use a hybrid argument over the total variation distance, each of the smaller "hybrid steps" is optimal, whereas the overall lower bound is not optimal. We cannot achieve a lower bound stronger than $\Omega(n/k)$.

Our Solution

Instead of keeping track of progress in terms of advantage gained in guessing the answer, we directly keep track of the "information" the message reveals about the answer. Formally, this translates to the change in entropy of the answer, after each successive message. Initially, the players have no information about the answer, and it is uniform over $\{0,1\}$ (the entropy is 1). Any protocol with a large enough probability of success must reduce the entropy of the answer by a large factor (see Fano's inequality in Proposition 8). We prove that after each message, the entropy is reduced only by an additive factor *linear in the length of the message*, by a reduction to the INDEX problem. This will give us a lower bound on the total length of the messages.

For INDEX_n, in any protocol with probability of success at least ε, the entropy of the answer conditioned on the message is at most $H_2(\varepsilon)$ where $H_2(x) = x\log(1/x)+(1-x)\log(1/(1-x))$ is the binary entropy function. In the standard version where the initial entropy is 1, the reduction in entropy is $1 - H_2(\varepsilon)$, and it is known that the protocol requires $\Omega(n(1 - H_2(\varepsilon)))$ communication [24]. This is not sufficient for our application due to the following reason: after the message of $\mathcal{P}_1$, when $\mathcal{P}_2$ and $\mathcal{P}_3$ attempt to solve INDEX_n, they already have some prior advantage that the message of $\mathcal{P}_1$ gives them. Therefore, we need to analyze INDEX_n when the answer is not uniform over $\{0,1\}$.

Biased Index

We define the biased index problem, parametrized by $\theta \in [-1/2, 1/2]$. Alice and Bob receive input $Y \in \{0,1\}^n$ and $\rho \in [n]$ respectively, such that the value at position ρ in Y (denoted by $Y(\rho)$) is 1 with probability $1/2 + \theta$ and 0 otherwise. Alice sends a message M to Bob and Bob has to output $Y(\rho)$. The initial entropy of the answer is $H_2(1/2 + \theta)$. We prove that entropy of $Y(\rho)$ conditioned on M is smaller by at most $O((|M| + \log n)/n)$ compared to the initial $H_2(1/2 + \theta)$, which is **our main contribution** (see Lemma 14).

We can show that the entropy of input to Alice, i.e. the random string Y, in such a distribution is at least $\Omega(n \cdot H_2(1/2 + \theta))$. Hence, after a message of length s from Alice, the entropy of string Y reduces to $\Omega(n \cdot H_2(1/2 + \theta) - s)$. For a randomly chosen position in Y after conditioning on the message, the entropy is at least $\approx H_2(1/2 + \theta) - s/n$, which gives a lower bound on s.

In the distribution given to Alice and Bob, however, ρ is *not* chosen uniformly at random, and in fact, is correlated with the distribution of Y. Such versions of index where the distributions of Alice and Bob are correlated have been studied before (see Sparse Indexing in Appendix A of [5] and Section 3.3 of [36]). This correlation is the main issue in analyzing biased index with information theoretic tools.

Adapted from the approach in Appendix A of [5], we restrict the randomness in Y to a fixed set of indices of a carefully chosen size (based on θ), and break this correlation. The loss in entropy of Y is not significant enough to hinder us, and the restriction then allows us to use standard information theoretic tools to analyze biased index (see Section 3.3 for more details).

Independent and Concurrent Work

Independently and concurrently of this work, [33] made progress on Conjecture 2. They showed a lower bound of $\Omega(n/k + \sqrt{n})$ for oblivious protocols (where the length of the message sent by each player does not depend on the input), and a lower bound of $\Omega(n/k - k)$ for general protocols. We show a lower bound of $\Omega(n - k \log n)$ for all protocols.[1]

Quantitatively, our lower bounds are a factor of almost k stronger, and are optimal for $k = o(n/\log n)$; moreover, our lower bound also holds for the Augmented Chain problem of [15]. In terms of techniques, however, the two works are entirely disjoint: their proof is based on a new method of analysis through min-entropy and we use information theoretic approaches.

2 Preliminaries

In this section, we will present the required notation and definitions for our proof.

Notation

For any tuple $A = (A_1, A_2, \ldots, A_m)$ of m items, we use $A_{<i}$ to denote the tuple $(A_1, A_2, \ldots, A_{i-1})$ for all $i \in [m]$. We use sans-serif font to denote random variables. For any random variable A, we use $A \sim \mathsf{A}$ to denote any A sampled from the distribution of the random variable A.

For any string $X \in \{0,1\}^n$, we use $X(\sigma)$ to denote the bit at position σ in X for $\sigma \in [n]$. We use $X(< \sigma)$ to denote the string of $\sigma - 1$ bits preceding $X(\sigma)$ in X. We use $X(S)$ for any $S \subseteq [n]$ to denote the bits at positions in set S.

For any $x \in [0,1]$, we use $H_2 : [0,1] \to [0,1]$ to denote the binary entropy function.

$$H_2(x) = -x \log x - (1-x) \log(1-x).$$

We need the following standard approximation of binomial coefficients (see Lemma 7 of Chapter 10 in [32]).

▶ **Fact 4** (c.f. [32]). *For any $p \geq 1$ and any $q \in [p-1]$, we have,*

$$2^{p \cdot H_2(q/p)} \cdot \sqrt{\frac{n}{8\pi q(p-q)}} \leq \binom{p}{q} \leq 2^{p \cdot H_2(q/p)} \cdot \sqrt{\frac{n}{2\pi q(p-q)}}$$

2.1 Communication Complexity Model

We use the standard number-in-hand multi-party model of communication. Only the basic definitions are given in this subsection. More details can be found in textbooks on communication complexity [35, 31].

For any $k \geq 1$, let f be a function from $\mathcal{A}_1 \times \mathcal{A}_2 \times \ldots \times \mathcal{A}_k$ to $\{0,1\}$. There are k players $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_k$ where $\mathcal{P}_i$ gets an input $a_i \in \mathcal{A}_i$ for $i \in [k]$. There is a shared blackboard visible to all the players. The players have access to a shared tape of random bits, along with their own private randomness. In any protocol π for f, the players send a message to the blackboard in increasing order ($\mathcal{P}_1$ sends a message followed by $\mathcal{P}_2$, and so on till

[1] The authors of [33] pointed out an important flaw in the arguments of an earlier version of this work which was posted at around the same time as [33]. This flaw was subsequently fixed in the current version using a global change to the original argument, which now recovers optimal result for $k = o(n/\log n)$.

$\mathcal{P}_k$). The last player $\mathcal{P}_k$, after all the messages are sent, outputs a single bit denoted by $\pi(a_1, a_2, \ldots, a_k)$. Protocol π is said to solve f with probability of success at least $1-\delta$ if, for all $i \in [k]$, for any choice of $a_i \in \mathcal{A}_i$, we have,

$$\Pr[\pi(a_1, a_2, \ldots, a_k) \neq f(a_1, a_2, \ldots, a_k)] \leq \delta.$$

The **communication cost** of a protocol π is defined as the worst case **total communication** of all the players on the blackboard at the end of the protocol.

▶ **Definition 5.** *The* **randomized communication complexity** *of f, with probability of error δ, is defined as the minimum communication cost of any protocol which solves f with probability of success at least $1-\delta$.*

2.2 Information Theoretic Tools

Our proof relies on tools from information theory, and we state the basic definitions and the inequalities we need in this section. Proofs of the statements and more details can be found in Chapter 2 of a textbook on information theory by Cover and Thomas [13].

▶ **Definition 6** (Shannon Entropy)**.** *For any random variable X over support $\mathcal{A}$, the Shannon entropy of X, denoted by $\mathbb{H}(\mathsf{X})$ is defined as,*

$$\mathbb{H}(\mathsf{X}) = \sum_{A \in \mathcal{A}} \Pr[\mathsf{X} = A] \cdot \log(1/\Pr[\mathsf{X} = A]).$$

For any event $\mathcal{E}$ we define $\mathbb{H}(\mathsf{X} \mid \mathcal{E})$ in the same way, as the entropy of distribution of X conditioned on the event $\mathcal{E}$. For any two random variables X and Y, the entropy of X conditioned on Y, denoted by $\mathbb{H}(\mathsf{X} \mid \mathsf{Y})$ is defined as,

$$\mathbb{H}(\mathsf{X} \mid \mathsf{Y}) = \mathop{\mathbb{E}}_{Y \sim \mathsf{Y}} \mathbb{H}(\mathsf{X} \mid \mathsf{Y} = Y).$$

▶ **Fact 7.** *We know the following about entropy and mutual information:*

1. *For any random variable X, the entropy obeys the bound: $0 \leq \mathbb{H}(\mathsf{X}) \leq \log_2(|\mathcal{X}|)$ where $\mathcal{X}$ is the support of X.*
2. *For any two random variables X, Y, $\mathbb{H}(\mathsf{X} \mid \mathsf{Y}) \leq \mathbb{H}(\mathsf{X})$ with equality holding iff $\mathsf{X} \perp \mathsf{Y}$.*
3. *Chain Rule of Entropy: For $m \geq 1$ and any tuple of random variables $\mathsf{X} = (\mathsf{X}_1, \mathsf{X}_2, \ldots, \mathsf{X}_m)$, $\mathbb{H}(\mathsf{X}) = \sum_{i \in [m]} \mathbb{H}(\mathsf{X}_i \mid \mathsf{X}_{<i})$.*
4. *Subadditivity of entropy: For $m \geq 1$ and any tuple of random variables $\mathsf{X} = (\mathsf{X}_1, \mathsf{X}_2, \ldots, \mathsf{X}_m)$, $\mathbb{H}(\mathsf{X}) \leq \sum_{i \in [m]} \mathbb{H}(\mathsf{X}_i)$.*

We also need the following proposition, which relates entropy to the probability of correctness while estimating a random variable.

▶ **Proposition 8** (Fano's inequality)**.** *Given a binary random variable X and an estimator random variable Y and a function g such that $g(Y) = X$ with probability at least $1-\delta$ for $\delta < 1/2$,*

$$\mathbb{H}(\mathsf{X} \mid \mathsf{Y}) \leq H_2(\delta).$$

This concludes our preliminaries section.

3 The Lower Bound

In this section, we will prove our lower bound of $\Omega(n)$ on the communication complexity of the $\text{CHAIN}_{n,k}$ problem for $k = o(n/\log n)$. Let us formally define the $\text{CHAIN}_{n,k}$ communication problem first.

▶ **Definition 9.** *The $\text{CHAIN}_{n,k}$ communication problem is defined as follows. Given $k+1$ players $\mathcal{P}_i$ for $i \in [k+1]$ where,*

- *$\mathcal{P}_i$ has a string $X_i \in \{0,1\}^n$ for each $i \in [k]$, and,*
- *$\mathcal{P}_i$ for $1 < i \leq k+1$ has an index $\sigma_{i-1} \in [n]$,*

such that,

$$X_i(\sigma_i) = z,$$

for some bit $z \in \{0,1\}$. The players have a blackboard visible to all the parties. For $i \in [k]$ in ascending order, $\mathcal{P}_i$ sends a single message M_i, after which the index σ_i is revealed to the blackboard at no cost. $\mathcal{P}_{k+1}$ has to output whether z is 0 or 1. Refer to Figure 2 for an illustration.

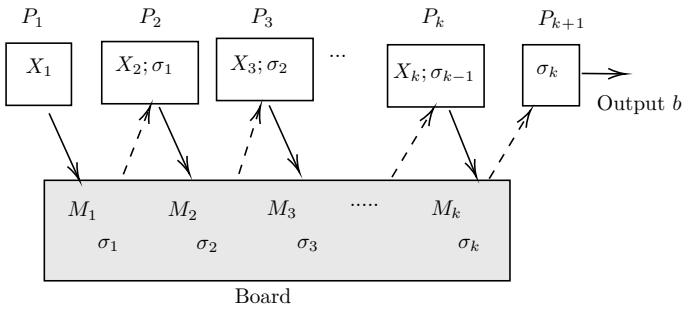

Figure 2 An illustration of the $\text{CHAIN}_{n,k}$ problem from Definition 9. The solid arrows illustrate that player $\mathcal{P}_i$ writes a message M_i to the board. The dashed arrows indicate that $\mathcal{P}_i$ can read the contents of the board. It also shows the order in which the messages are sent by the players and indices are released.

Let us recall the statement of our main result.

▶ **Theorem 3** (restated). *For any $n, k \geq 1$, any protocol for $\text{CHAIN}_{n,k}$ with probability of success at least 2/3, requires $\Omega(n - k\log n)$ total bits of communication.*

We give our hard distribution for $\text{CHAIN}_{n,k}$ in Section 3.1 and give the proof of Theorem 3 in Section 3.2 except for the analysis of biased index, which is given in Section 3.3. Lastly, we extend the arguments to Augmented Chain in Section 3.4.

3.1 Setting Up the Problem

In this subsection, we start by defining the notation for our proof, and we describe the input distributions to $\text{CHAIN}_{n,k}$.

The input hard distribution is as follows. Let $\mathcal{L} \subset \{0,1\}^n$ be the subset of strings where the number of ones is exactly equal to $n/2$.

Distribution $\mathcal{D}$ for $\textsc{chain}_{n,k}$:
1. Pick a bit z uniformly at random from $\{0,1\}$.
2. For each $i \in [k]$, sample (X_i, σ_i) uniformly at random from $\mathcal{L} \times [n]$ and independently conditioned on $X_i(\sigma_i) = z$.

Notation

We use X_i to denote the random variable corresponding to string X_i and $\boldsymbol{\sigma}_i$ to denote the random variable corresponding to index σ_i for $i \in [k]$. To denote the random variable corresponding to the first $i-1$ strings and indices, we use $\mathsf{X}_{<i}$ and $\boldsymbol{\sigma}_{<i}$ respectively.

We use M_i to denote the random variable corresponding to the message M_i sent by $\mathcal{P}_i$ to the blackboard. We use $M = (M_1, M_2, \ldots, M_k)$ to denote the tuple containing the messages of all the players, and M to denote the random variable corresponding to M.

Let π be a deterministic protocol for $\textsc{chain}_{n,k}$ with probability of success at least $2/3$ when the input is distributed according to $\mathcal{D}$. We use Γ to denote the random variable corresponding to the contents of the blackboard (referred to as a **transcript**), and we use γ to also denote transcripts sampled from Γ. We use Γ_i to denote the random variable of the tuple (M^i, σ^i) for $i \in [k]$. We use $\pi(\gamma_{<i}, X_i)$ to denote the output of $\mathcal{P}_i$ when the contents of the blackboard are $\gamma_{<i}$ and the input is X_i for $i \in [k]$.

Let s be the total length of all the messages sent by the players in γ. We assume that the total length of the messages is exactly s by padding. For any random variable A, we use $\mathcal{D}(\mathsf{A})$ to denote the distribution of the random variable A, as the input is distributed according to $\mathcal{D}$. We replace $\mathcal{D}(\mathsf{A} \mid \mathsf{B} = b)$ with $\mathcal{D}(\mathsf{A} \mid b)$ for ease of readability whenever it is clear from context.

Let Z denote the random variable corresponding to bit z. The bit z corresponds to the answer to $\textsc{chain}_{n,k}$. We show the lower bound of $\Omega(n - k \log n)$ for distinguishing between the case when $z = 0$ and $z = 1$ based on the contents of the blackboard.

We need one important observation about the distribution $\mathcal{D}$.

▶ **Observation 10.** *For any $i \in [k]$, random variable $\mathsf{X}_i, \boldsymbol{\sigma}_i$ is independent of $\Gamma_{<i}$ conditioned on Z.*

Proof. For any fixed value of Z, $\mathsf{X}_i, \boldsymbol{\sigma}_i$ are chosen uniformly at random from $\mathcal{L} \times [n]$ such that $\mathsf{X}_i(\boldsymbol{\sigma}_i) = \mathsf{Z}$. This choice is independent of any $\mathsf{X}_j, \boldsymbol{\sigma}_j$ with $j \neq i$, and thus independent of $\Gamma_{<i}$. ◀

We are ready to proceed with the proof of our main theorem.

3.2 Proof of Lower Bound

We start by showing that in any successful protocol, the entropy of the distribution of Z conditioned on the message must be small.

▷ Claim 11 (The transcript reveals information about Z.).

$$\mathbb{H}(\mathsf{Z} \mid \Gamma) \leq 24/25.$$

Proof. We know that $\mathcal{P}_{k+1}$ successfully finds the value of z with probability of success at least $2/3$ using the transcript. Thus, using Fano's inequality in Proposition 8, we have,

$$\mathbb{H}(\mathsf{Z} \mid \Gamma) \leq H_2(2/3) \leq 24/25.$$ ◁

The main part of the proof is that we show a lower bound the entropy of Z conditioned on the transcripts using the entropy of the message.

▶ **Lemma 12.** *For any protocol π,*

$$1 - \frac{12}{n} \cdot (\mathbb{H}(\mathsf{M}) + k \log n) \leq \mathbb{H}(\mathsf{Z} \mid \Gamma).$$

Before we prove Lemma 12, we can easily show that it implies Theorem 3.

Proof of Theorem 3. Combining Lemma 12 with Claim 11, we get,

$$1 - \frac{12}{n} \cdot (\mathbb{H}(\mathsf{M}) + k \log n) \leq \mathbb{H}(\mathsf{Z} \mid \Gamma) \leq \frac{24}{25}.$$

This gives that,

$$\mathbb{H}(\mathsf{M}) \geq \frac{n}{25 \cdot 12} - k \log n.$$

We know from Fact 7-(1) that $\mathbb{H}(\mathsf{M}) \leq \log(2^s) = s$, which proves that the total number of bits $s = \Omega(n - k \log n)$ for any deterministic protocol. By Yao's minimax principle, we get a lower bound of $\Omega(n - k \log n)$ on the randomized communication complexity of $\textsc{chain}_{n,k}$. ◀

The proof of Lemma 12 employs a reduction to the two player $\textsc{index}_n$. However, in these instances of $\textsc{index}_n$, Alice and Bob already have some partial information about the answer. We call this problem the **biased index** problem, and it is defined based on parameter $\theta \in [-1/2, 1/2]$, which is the initial bias known about the answer.

▶ **Definition 13.** *The* **biased index distributional communication problem**, *denoted by* $\textsc{bias-ind}(\theta)$ *for $\theta \in [-1/2, 1/2]$, is defined as follows.*

Sample $W \in \{0, 1\}$ such that $W = 1$ with probability $1/2 + \theta$, and $W = 0$ otherwise. Sample (Y, ρ) uniformly at random from $\mathcal{L} \times [n]$ conditioned on $Y(\rho) = W$. Give string Y to Alice, and index ρ to Bob. Bob has to output $Y(\rho)$ after a single message $M_{\textsc{index}}$ from Alice.

Let $\pi_{\textsc{index}}$ be a deterministic protocol for $\textsc{bias-ind}(\theta)$. Let $\mathsf{W}, \mathsf{Y}, \boldsymbol{\rho}, \mathsf{M}_{\textsc{index}}$ denote the random variables corresponding to W, Y, ρ and $M_{\textsc{index}}$ respectively. Let $\mathcal{D}_\theta$ denote the joint distribution of $\mathsf{W}, \mathsf{Y}, \boldsymbol{\rho}$ and $\mathsf{M}_{\textsc{index}}$ in $\textsc{bias-ind}(\theta)$.

We prove the following lemma about $\textsc{bias-ind}(\theta)$ in Section 3.3.

▶ **Lemma 14** (Biased Index)**.** *For any protocol $\pi_{\textit{INDEX}}$ for $\textsc{bias-ind}(\theta)$,*

$$\mathbb{H}(\mathsf{W} \mid \mathsf{M}_{\textit{INDEX}}, \boldsymbol{\rho}) \geq H_2(1/2 + \theta) - \frac{2}{n} \cdot (\mathbb{H}(\mathsf{M}_{\textit{INDEX}}) + \log n).$$

We can prove Lemma 12 using Lemma 14, but the proof is deferred to the full version.

3.3 Biased Index

In this subsection, we will prove Lemma 14. Let us first recall the input distribution to $\textsc{bias-ind}(\theta)$.

Distribution $\mathcal{D}_\theta$: Sample $W = 1$ with probability $1/2 + \theta$ and set $W = 0$ otherwise. Sample $(Y, \rho) \in \mathcal{L} \times [n]$ conditioned on $Y(\rho) = W$.

In $\mathcal{D}_\theta$, the distribution of Y and $\boldsymbol{\rho}$ are highly correlated. We give an alternate way of sampling Y, ρ so that this correlation is removed partially.

Distribution $\mathcal{D}'_\theta$:
For $\theta \geq 0$:
1. Sample set $T \subset [n]$ of size $b = n/(1+2\theta)$ uniformly at random.
2. Sample a set S of $n/2$ indices from T uniformly at random and set them to 1, and set $[n] \setminus S$ to 0 to get Y.
3. Sample ρ by sampling an index uniformly at random from T.

For $\theta < 0$:
1. Sample set $T \subset [n]$ of size $b = n/(1-2\theta)$ uniformly at random.
2. Sample a set S of $n/2$ indices from T uniformly at random and set them to 0, and set $[n] \setminus S$ to 1 to get Y.
3. Sample ρ by sampling an index uniformly at random from T.

In this section, we assume that $\theta \geq 0$. For the case when $\theta < 0$, the proof follows in the same vein, and is not presented. Let T, S denote the random variables corresponding to set T and set S respectively. We show that distributions $\mathcal{D}_\theta$ and $\mathcal{D}'_\theta$ are in fact identical, and the proofs can be found in the full version.

▷ Claim 15. In distribution $\mathcal{D}_\theta$, for any $(Y, \rho) \in \mathcal{L} \times [n]$,

$$\Pr(\mathsf{Y} = Y, \boldsymbol{\rho} = \rho) = \begin{cases} \frac{1+2\theta}{n \cdot \binom{n}{n/2}} & \text{when } Y(\rho) = 1, \\ \frac{1-2\theta}{n \cdot \binom{n}{n/2}} & \text{when } Y(\rho) = 0. \end{cases}$$

▷ Claim 16. Distribution $\mathcal{D}_\theta$ is the same as $\mathcal{D}'_\theta$.

Using this alternate way of sampling, it is easy to see that random variables Y and $\boldsymbol{\rho}$ are independent of each other conditioned on T. It can also be extended to include random variable $\mathsf{M}_{\text{INDEX}}$, as it is only a function of Y.

▶ **Observation 17.** *In distribution $\mathcal{D}'_\theta$, conditioned on $\mathsf{T} = T$, for any $i \in T$, distribution of random variables $\mathsf{Y}, \mathsf{M}_{INDEX}$ is independent of event $\boldsymbol{\rho} = i$.*

Proof. Conditioned on $\mathsf{T} = T$, string Y is chosen by picking an $n/2$ size set S uniformly at random from T, and setting these indices to 1, and this choice also fixes $\mathsf{M}_{\text{INDEX}}$ as the protocol is deterministic. And index ρ is chosen uniformly at random from T, independently of the choice of S, by definition of $\mathcal{D}'_\theta$. Thus, choice of $\boldsymbol{\rho} = i$ is independent of random variables $\mathsf{Y}, \mathsf{M}_{\text{INDEX}}$. ◀

Next, we will show that even conditioned on T, the entropy of Y remains large.

▷ Claim 18.

$$\mathbb{H}(\mathsf{Y} \mid \mathsf{T}) \geq \frac{n}{(1+2\theta)} \cdot H_2(1/2+\theta) - 2\log n.$$

Proof. We assume that $\theta < 1/2$, as otherwise, the statement is vacuously true. $H_2(1) = 0$ by definition, and entropy is always non-negative by Fact 7-(1).

Conditioned on $\mathsf{T} = T$, we know that Y is fixed by choosing set S uniformly at random. Thus,

$$\mathbb{H}(\mathsf{Y} \mid \mathsf{T}) = \log\left(\binom{b}{n/2}\right) \qquad \text{(by Fact 7-(1))}$$

$$\geq \log\left(2^{bH_2(n/2b)} \cdot \sqrt{\frac{b}{8(n/2)(b-n/2)}}\right) \quad \text{(by Fact 4, and } n/2 < b \text{, as } \theta < 1/2)$$
$$= b \cdot H_2(n/2b) + \frac{1}{2} \cdot \log(\frac{b}{8(n/2)(b-n/2)})$$
$$= \frac{n}{(1+2\theta)} \cdot H_2(1/2+\theta) + \frac{1}{2}\log(\frac{1}{4n \cdot (1/2-\theta)})$$
$$\geq \frac{n}{(1+2\theta)} \cdot H_2(1/2+\theta) + \frac{1}{2}\log(1/4n) \quad \text{(as } 1/2-\theta \leq 1)$$
$$\geq \frac{n}{(1+2\theta)} \cdot H_2(1/2+\theta) - 2\log n. \qquad \lhd$$

We are ready to prove Lemma 14.

Proof of Lemma 14. We can lower bound the entropy of W conditioned on $\mathsf{M}_{\text{INDEX}}, \boldsymbol{\rho}$ as,

$$\mathbb{H}(\mathsf{W} \mid \mathsf{M}_{\text{INDEX}}, \boldsymbol{\rho}) \geq \mathbb{H}(\mathsf{W} \mid \mathsf{M}_{\text{INDEX}}, \boldsymbol{\rho}, \mathsf{T}) \quad \text{(as conditioning reduces entropy, Fact 7-(2))}$$
$$= \mathop{\mathbb{E}}_{\mathsf{T}=T}\Big[\frac{1}{b} \cdot \sum_{\rho \in T} \mathbb{H}(\mathsf{W} \mid \mathsf{M}_{\text{INDEX}}, \boldsymbol{\rho} = \rho, \mathsf{T} = T)\Big] \quad \text{(as } \boldsymbol{\rho} \text{ is uniform over } T)$$
$$= \mathop{\mathbb{E}}_{\mathsf{T}=T}\Big[\frac{1}{b} \cdot \sum_{\rho \in T} \mathbb{H}(\mathsf{Y}(\rho) \mid \mathsf{M}_{\text{INDEX}}, \boldsymbol{\rho} = \rho, \mathsf{T} = T)\Big]$$
(as $W = Y(\rho)$ by definition of $\mathcal{D}_\theta$)
$$= \mathop{\mathbb{E}}_{\mathsf{T}=T}\Big[\frac{1}{b} \cdot \sum_{\rho \in T} \mathbb{H}(\mathsf{Y}(\rho) \mid \mathsf{M}_{\text{INDEX}}, \mathsf{T} = T)\Big]$$
(as $\mathsf{Y}, \mathsf{M}_{\text{INDEX}} \perp (\boldsymbol{\rho} = \rho) \mid \mathsf{T} = T$, by Observation 17)
$$\geq \mathop{\mathbb{E}}_{\mathsf{T}=T}\Big[\frac{1}{b} \cdot \mathbb{H}(\mathsf{Y}(T) \mid \mathsf{M}_{\text{INDEX}}, \mathsf{T} = T)\Big]$$
(by subadditivity of entropy, Fact 7-(4))
$$= \mathop{\mathbb{E}}_{\mathsf{T}=T}\Big[\frac{1}{b} \cdot \mathbb{H}(\mathsf{Y} \mid \mathsf{M}_{\text{INDEX}}, \mathsf{T} = T)\Big] \quad \text{(as } Y([n] \setminus T) \text{ is fixed to be 0)}$$
$$= \frac{1}{b} \cdot \mathbb{H}(\mathsf{Y} \mid \mathsf{M}_{\text{INDEX}}, \mathsf{T}).$$

Thus it follows that,

$$\mathbb{H}(\mathsf{Y} \mid \mathsf{M}_{\text{INDEX}}, \mathsf{T}) \leq b \cdot \mathbb{H}(\mathsf{W} \mid \mathsf{M}_{\text{INDEX}}, \boldsymbol{\rho}). \tag{1}$$

We also have,

$$\mathbb{H}(\mathsf{Y} \mid \mathsf{T}) = \mathbb{H}(\mathsf{Y}, \mathsf{M}_{\text{INDEX}} \mid \mathsf{T}) \quad \text{(as } M \text{ is fixed by } Y)$$
$$= \mathbb{H}(\mathsf{M}_{\text{INDEX}} \mid \mathsf{T}) + \mathbb{H}(\mathsf{Y} \mid \mathsf{M}_{\text{INDEX}}, \mathsf{T}) \quad \text{(by chain rule of entropy, Fact 7-(3))}$$
$$\leq \mathbb{H}(\mathsf{M}_{\text{INDEX}} \mid \mathsf{T}) + b \cdot \mathbb{H}(\mathsf{W} \mid \mathsf{M}_{\text{INDEX}}, \boldsymbol{\rho}) \quad \text{(by Eq (1))}$$
$$\leq \mathbb{H}(\mathsf{M}_{\text{INDEX}}) + b \cdot \mathbb{H}(\mathsf{W} \mid \mathsf{M}_{\text{INDEX}}, \boldsymbol{\rho}).$$
(as conditioning reduces entropy, by Fact 7-(2))

Combining with Claim 18, we get,

$$\mathbb{H}(\mathsf{M}_{\text{INDEX}}) + b \cdot \mathbb{H}(\mathsf{W} \mid \mathsf{M}_{\text{INDEX}}, \boldsymbol{\rho}) \geq b \cdot H_2(1/2+\theta) - 2\log n \quad \text{(as } b = n/(1+2\theta))$$
$$b \cdot \mathbb{H}(\mathsf{W} \mid \mathsf{M}_{\text{INDEX}}, \boldsymbol{\rho}) \geq b \cdot H_2(1/2+\theta) - (\mathbb{H}(\mathsf{M}_{\text{INDEX}}) + 2\log n).$$
(rearranging the terms)

Dividing both sides by b, we get,

$$\begin{aligned}\mathbb{H}(\mathsf{W} \mid \mathsf{M}_{\text{INDEX}}, \boldsymbol{\rho}) &\geq H_2(1/2+\theta) - \frac{1}{b} \cdot (\mathbb{H}(\mathsf{M}_{\text{INDEX}}) + 2\log n) \\ &\geq H_2(1/2+\theta) - \frac{2}{n} \cdot (\mathbb{H}(\mathsf{M}_{\text{INDEX}}) + 2\log n), \qquad (\text{as } b \geq n/2)\end{aligned}$$

finishing the proof. ◀

3.4 Extension to Augmented Chain

In this subsection, we will extend our lower bound to the Augmented Chain problem introduced by [15]. We begin by defining Augmented Index.

Augmented Index is a close variant of the INDEX problem. Here, in addition to having the index σ, Bob also has the bits $X(<\sigma)$. We know that this generalization also requires $\Omega(n)$ communication when Alice sends a message to Bob [34]. This problem is particularly useful for proving lower bounds for turnstile streams (see e.g., [11, 25, 16], and references therein). Tight information cost trade-off for this variant in the two-way communication model was proved by [9].

The formal definition of Augmented Chain follows.

▶ **Definition 19** (Augmented Chain). *The* AUG-CHAIN$_{n,k}$ *communication problem is defined as follows. Given $k+1$ players $\mathcal{P}_i$ for $i \in [k+1]$ where,*

- *$\mathcal{P}_i$ has a string $X_i \in \{0,1\}^n$ for each $i \in [k]$,*
- *$\mathcal{P}_i$ for $1 < i \leq k+1$ has an index $\sigma_{i-1} \in [n]$, and a string $X_{i-1}(<\sigma_{i-1})$,*

such that,

$$X_i(\sigma_i) = z,$$

for some bit $z \in \{0,1\}$. The players have a blackboard visible to all the parties. For $i \in [k]$ in ascending order, $\mathcal{P}_i$ sends a single message M_i, after which the index σ_i, and the string $X_{i-1}(<\sigma_{i-1})$ are revealed to the blackboard at no cost. $\mathcal{P}_{k+1}$ has to output whether z is 0 or 1. Refer to Figure 3 for an illustration.

For the chained version of Augmented Index, the lower bound that at least one player sends $\Omega(n/k^2)$ bits holds true, and the proof follows with minimal changes. Using this, [15] proved lower bounds for interval independent set selection in turnstile streams with weighted intervals.

We prove the following result about AUG-CHAIN$_{n,k}$.

▶ **Theorem 20.** *For any $n, k \geq 1$, any protocol for* AUG-CHAIN$_{n,k}$ *with probability of success at least 2/3 requires communication $\Omega(n - k\log n)$.*

A proof sketch detailing the changes needed to prove Theorem 20 is given in the full version. Most parts of the proof are similar to the proof of Theorem 3.

4 Applications to Streaming

In this section we give applications of our main result to independent sets in vertex arrival streams and streaming submodular maximization.

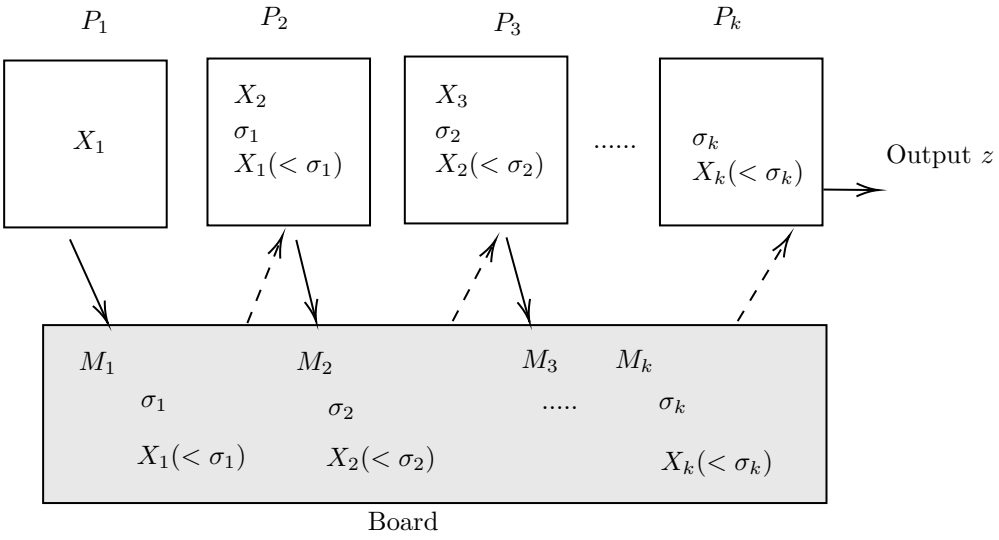

Figure 3 An illustration of the AUG-CHAIN$_{n,k}$ problem from Definition 19. The solid arrows illustrate that player $\mathcal{P}_i$ writes a message M_i to the board. The dashed arrows indicate that $\mathcal{P}_i$ can read the contents of the board. The order in which the messages are sent and indices, strings are released is also shown.

4.1 Independent Sets

In edge arrival streams, for any graph $G = (V, E)$, the vertex set V with n vertices is given, and the edges E arrive in any arbitrary order. We are required to process the graph in limited space.

In vertex arrival streams, for any graph $G = (V, E)$, the edges are grouped by their incident vertices. Vertices from V arrive one by one (in any arbitrary order), and when a vertex arrives, all the edges connecting it to any previously arrived vertices are revealed. This makes the vertex arrival stream a strictly easier model than the edge arrival stream, as the order of edges is restricted.

Indeed, for the maximal independent set problem, we know that finding algorithms in vertex arrival streams is easier; the greedy algorithm produces a maximal independent set in $\widetilde{O}(n)$ space, whereas, in edge arrival streams, any algorithm which finds a maximal independent set requires $\Omega(n^2)$ space [4, 12].

Maximum independent set (MIS), however, is provably hard in both vertex arrival streams and edge arrival streams. It is known that, any algorithm which performs an α-approximation of MIS in edge arrival streams requires $\Omega(n^2/\alpha^2)$ space from [22]. In vertex arrival streams, [12] proved a lower bound of $\Omega(n^2/\alpha^7)$. They also gave the following connection between CHAIN problem and MIS in the proof of Theorem 9 of their paper.

▶ **Proposition 21** (Rephrased from [12]). *For any $\alpha \geq 1$, any algorithm that gives an α-approximation of maximum independent sets in vertex arrival streams for n-vertex graphs using space at most s and probability of success at least 2/3 can be used to solve* CHAIN$_{n^2/64\alpha^4, 2\alpha}$ *with communication at most $2\alpha \cdot s$ bits and success probability at least 2/3.*

Our lower bound Theorem 3, along with Proposition 21 directly gives the following corollary.

▶ **Corollary 22.** *For $\alpha \geq 1$, any α-approximation of maximum independent sets in n-vertex graphs in vertex arrival streams uses $\Omega(n^2/\alpha^5 - \log n)$ space.*

This further reduces the gap between the lower bounds for α-approximation of MIS in vertex arrival streams and edge arrival streams by an α^2 factor.

4.2 Submodular Maximization

In this subsection, we will summarize our slight improvements to lower bounds for streaming submodular maximization.

A function f over ground set V from $f : 2^V \to \mathbb{R}$ is submodular if and only if, for any two sets $A \subset B \subset V$ and for any element $x \in V \setminus B$,

$$f(B \cup \{x\}) - f(B) \leq f(A \cup \{x\}) - f(A).$$

This captures the diminishing returns property of any submodular function.

We are interested in maximization of a monotone submodular function subject to a cardinality constraint. That is, for a given ℓ, we want to find a subset $S \subset V$ with $|S| \leq \ell$ such that for any other set $T \subset V$ with $|T| \leq \ell$, $f(S) \geq f(T)$.

We are given oracle access to function f, however, we do not have access to the entirety of the ground set. The elements of the ground set V arrive one by one, and the algorithm has space s to either store the incoming element or to discard it. The algorithm can query the oracle to f with any subset of the elements currently in storage. We want the storage to be roughly the same as the output size, which is ℓ.

In this model, [29] gave an algorithm which finds a $(1/2 - \epsilon)$-approximation in $O(\ell/\epsilon)$ space. [19] showed that a better approximation was not possible. They proved that any algorithm which gets a $(1/2 + \epsilon)$-approximation uses $\Omega(\epsilon |V| / \ell^3)$. They give the following connection to the CHAIN problem in Theorem 1.3 and Theorem 1.4 of their paper.

▶ **Proposition 23** (Rephrased from [19]). *For any $\epsilon > 0$, there exists a constant ℓ_0 such that for any $\ell \geq \ell_0$, any randomized streaming algorithm which maximizes a monotone submodular function $f : 2^V \to \mathbb{R}$ subject to cardinality constraint of at most ℓ, using space at most s and with approximation factor at least $(1/2 + \epsilon)$ in expectation can be used to solve $\text{CHAIN}_{|V|/\ell,\ell}$ with probability of success at least 2/3 and communication at most $s \cdot O(\ell/\epsilon)$.*

We get an improvement of ℓ factor over the current state-of-art lower bound in [19] as a corollary of Theorem 3.

▶ **Corollary 24.** *For any $\epsilon > 0$, there exists a constant ℓ_0 such that for any $\ell \geq \ell_0$, any streaming algorithm that maximizes a monotone submodular function $f : 2^V \to \mathbb{R}$ subject to a cardinality constraint of at most ℓ, with an approximation factor at least $(1/2 + \epsilon)$, requires $\Omega(|V| \epsilon/\ell^2 - \varepsilon \log(|V|))$ space.*

References

1. Farid Ablayev. Lower bounds for one-way probabilistic communication complexity and their application to space complexity. *Theoretical Computer Science*, 157(2):139–159, 1996. doi:10.1016/0304-3975(95)00157-3.
2. S. Assadi. Lecture notes on sublinear algorithms. https://sepehr.assadi.info/courses/cs514-s20/lec8.pdf, 2020.
3. S. Assadi, G. Kol, R. R. Saxena, and H. Yu. Multi-pass graph streaming lower bounds for cycle counting, max-cut, matching size, and other problems. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 354–364, Los Alamitos, CA, USA, November 2020. IEEE Computer Society. doi:10.1109/FOCS46700.2020.00041.

4 Sepehr Assadi, Yu Chen, and Sanjeev Khanna. Sublinear algorithms for $(\Delta + 1)$ vertex coloring. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019*, pages 767–786, 2019. doi:10.1137/1.9781611975482.48.

5 Sepehr Assadi, Sanjeev Khanna, and Yang Li. Tight bounds for single-pass streaming complexity of the set cover problem. In *Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing*, STOC '16, pages 698–711, New York, NY, USA, 2016. Association for Computing Machinery. doi:10.1145/2897518.2897576.

6 Sepehr Assadi and Janani Sundaresan. (noisy) gap cycle counting strikes back: Random order streaming lower bounds for connected components and beyond. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, STOC 2023, pages 183–195, New York, NY, USA, 2023. Association for Computing Machinery. doi:10.1145/3564246.3585192.

7 Sujoy Bhore, Fabian Klute, and Jelle J. Oostveen. On streaming algorithms for geometric independent set and clique. In Parinya Chalermsook and Bundit Laekhanukit, editors, *Approximation and Online Algorithms*, pages 211–224, Cham, 2022. Springer International Publishing. doi:10.1007/978-3-031-18367-6_11.

8 Harry Buhrman and Ronald Wolf. Communication complexity lower bounds by polynomials. In *Proceedings of the Annual IEEE Conference on Computational Complexity*, pages 120–130, February 2001. doi:10.1109/CCC.2001.933879.

9 Amit Chakrabarti, Graham Cormode, Ranganath Kondapally, and Andrew McGregor. Information cost tradeoffs for augmented index and streaming language recognition. *SIAM Journal on Computing*, 42(1):61–83, 2013. doi:10.1137/100816481.

10 Lijie Chen, Gillat Kol, Dmitry Paramonov, Raghuvansh R. Saxena, Zhao Song, and Huacheng Yu. Near-optimal two-pass streaming algorithm for sampling random walks over directed graphs. In *International Colloquium on Automata, Languages and Programming*, 2021. URL: https://api.semanticscholar.org/CorpusID:232014583.

11 Kenneth L. Clarkson and David P. Woodruff. Numerical linear algebra in the streaming model. In *Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing*, STOC '09, pages 205–214, New York, NY, USA, 2009. Association for Computing Machinery. doi:10.1145/1536414.1536445.

12 Graham Cormode, Jacques Dark, and Christian Konrad. Independent Sets in Vertex-Arrival Streams. In Christel Baier, Ioannis Chatzigiannakis, Paola Flocchini, and Stefano Leonardi, editors, *46th International Colloquium on Automata, Languages, and Programming (ICALP 2019)*, volume 132 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 45:1–45:14, Dagstuhl, Germany, 2019. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. doi:10.4230/LIPIcs.ICALP.2019.45.

13 Thomas M. Cover and Joy A. Thomas. *Elements of information theory (2. ed.)*. Wiley, 2006.

14 Carsten Damm, Stasys Jukna, and Jirí Sgall. Some bounds on multiparty communication complexity of pointer jumping. In Claude Puech and Rüdiger Reischuk, editors, *STACS 96, 13th Annual Symposium on Theoretical Aspects of Computer Science, Grenoble, France, February 22-24, 1996, Proceedings*, volume 1046 of *Lecture Notes in Computer Science*, pages 643–654. Springer, 1996. doi:10.1007/3-540-60922-9_52.

15 Jacques Dark, Adithya Diddapur, and Christian Konrad. Interval Selection in Data Streams: Weighted Intervals and the Insertion-Deletion Setting. In Patricia Bouyer and Srikanth Srinivasan, editors, *43rd IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS 2023)*, volume 284 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 24:1–24:17, Dagstuhl, Germany, 2023. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. doi:10.4230/LIPIcs.FSTTCS.2023.24.

16 Jacques Dark and Christian Konrad. Optimal lower bounds for matching and vertex cover in dynamic graph streams. In Shubhangi Saraf, editor, *35th Computational Complexity Conference, CCC 2020, July 28-31, 2020, Saarbrücken, Germany (Virtual Conference)*, volume 169 of *LIPIcs*, pages 30:1–30:14. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2020. doi:10.4230/LIPICS.CCC.2020.30.

17 Ashkan Norouzi Fard, Moran Feldman, Ola Svensson, and Rico Zenklusen. Submodular maximization subject to matroid intersection on the fly. In *30th Annual European Symposium on Algorithms, ESA 2022, September 5-9, 2022, Berlin/Potsdam, Germany*, pages 52:1–52:14, 2022. doi:10.4230/LIPICS.ESA.2022.52.

18 Joan Feigenbaum, Sampath Kannan, Andrew McGregor, Siddharth Suri, and Jian Zhang. Graph distances in the data-stream model. *SIAM J. Comput.*, 38(5):1709–1727, 2008. doi:10.1137/070683155.

19 Moran Feldman, Ashkan Norouzi-Fard, Ola Svensson, and Rico Zenklusen. The one-way communication complexity of submodular maximization with applications to streaming and robustness. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2020, pages 1363–1374, New York, NY, USA, 2020. Association for Computing Machinery. doi:10.1145/3357713.3384286.

20 Sudipto Guha and Andrew McGregor. Stream order and order statistics: Quantile estimation in random-order streams. *SIAM J. Comput.*, 38(5):2044–2059, 2009. doi:10.1137/07069328X.

21 Venkatesan Guruswami and Krzysztof Onak. Superlinear lower bounds for multipass graph processing. In *Proceedings of the 28th Conference on Computational Complexity, CCC 2013, K.lo Alto, California, USA, 5-7 June, 2013*, pages 287–298, 2013. doi:10.1109/CCC.2013.37.

22 Magnús M. Halldórsson, Xiaoming Sun, Mario Szegedy, and Chengu Wang. Streaming and communication complexity of clique approximation. In Artur Czumaj, Kurt Mehlhorn, Andrew M. Pitts, and Roger Wattenhofer, editors, *Automata, Languages, and Programming - 39th International Colloquium, ICALP 2012, Warwick, UK, July 9-13, 2012, Proceedings, Part I*, volume 7391 of *Lecture Notes in Computer Science*, pages 449–460. Springer, 2012. doi:10.1007/978-3-642-31594-7_38.

23 P. Indyk and D. Woodruff. Tight lower bounds for the distinct elements problem. In *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.*, pages 283–288, 2003. doi:10.1109/SFCS.2003.1238202.

24 Rahul Jain, Jaikumar Radhakrishnan, and Pranab Sen. A property of quantum relative entropy with an application to privacy in quantum communication. *J. ACM*, 56(6), September 2009. doi:10.1145/1568318.1568323.

25 Daniel M. Kane, Jelani Nelson, and David P. Woodruff. On the exact space complexity of sketching and streaming small norms. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '10, pages 1161–1178, USA, 2010. Society for Industrial and Applied Mathematics. doi:10.1137/1.9781611973075.93.

26 Michael Kapralov, Sanjeev Khanna, and Madhu Sudan. Streaming lower bounds for approximating MAX-CUT. In Piotr Indyk, editor, *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015, San Diego, CA, USA, January 4-6, 2015*, pages 1263–1282. SIAM, 2015. doi:10.1137/1.9781611973730.84.

27 Michael Kapralov, Amulya Musipatla, Jakab Tardos, David P. Woodruff, and Samson Zhou. Noisy Boolean Hidden Matching with Applications. In Mark Braverman, editor, *13th Innovations in Theoretical Computer Science Conference (ITCS 2022)*, volume 215 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 91:1–91:19, Dagstuhl, Germany, 2022. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. doi:10.4230/LIPIcs.ITCS.2022.91.

28 Mikhail Kapralov. Better bounds for matchings in the streaming model. In *ACM-SIAM Symposium on Discrete Algorithms*, 2012. URL: https://api.semanticscholar.org/CorpusID:448251.

29 Ehsan Kazemi, Marko Mitrovic, Morteza Zadimoghaddam, Silvio Lattanzi, and Amin Karbasi. Submodular streaming in all its glory: Tight approximation, minimum memory and low adaptive complexity. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3311–3320. PMLR, 09–15 June 2019. URL: https://proceedings.mlr.press/v97/kazemi19a.html.

30 Ilan Kremer, Noam Nisan, and Dana Ron. On randomized one-round communication complexity. In *Proceedings of the Twenty-Seventh Annual ACM Symposium on Theory of Computing*, STOC '95, pages 596–605, New York, NY, USA, 1995. Association for Computing Machinery. doi:10.1145/225058.225277.

31 Eyal Kushilevitz and Noam Nisan. *Communication complexity*. Cambridge University Press, 1997.

32 F. J. (Florence Jessie) MacWilliams and N. J. A. (Neil James Alexander) Sloane. *The theory of error-correcting codes*. North-Holland mathematical library ; v. 16. North-Holland Pub. Co., Amsterdam, 1978 - 1977.

33 Guangxu Yang Mi-Ying Huang, Xinyu Mao and Jiapeng Zhang. Breaking square-root loss barriers via min-entropy. *Electron. Colloquium Comput. Complex.*, pages TR24–067, 2024. URL: https://eccc.weizmann.ac.il/report/2024/067/.

34 Peter Bro Miltersen, Noam Nisan, Shmuel Safra, and Avi Wigderson. On data structures and asymmetric communication complexity. *Journal of Computer and System Sciences*, 57(1):37–49, 1998. doi:10.1006/jcss.1998.1577.

35 Anup Rao and Amir Yehudayoff. *Communication Complexity: and Applications*. Cambridge University Press, 2020. doi:10.1017/9781108671644.

36 Mert Saglam. Tight bounds for data stream algorithms and communication problems. Master's thesis, Simon Fraser University, 2011.