Multi-Pass Memory Lower Bounds for Learning Problems

Qian Li LIQIAN.ICT@GMAIL.COM
Shenzhen International Center For Industrial And Applied Mathematics, Shenzhen Research Institute of Big Data

Shuo Wang SHUOW@MIT.EDU
Departments of Mathematics, Massachusetts Institute of Technology

Jiapeng Zhang JIAPENGZ@USC.EDU
Thomas Lord Department of Computer Science, University of Southern California

Editors: Nika Haghtalab and Ankur Moitra

Abstract

Space complexity in learning problems has received a lot of attention in recent years. In this direction, Brown, Bun, and Smith (COLT 2022) studied space complexity lower bounds for several natural learning problems under the *one-pass streaming* setting. Assuming that the examples are sampled from $\{0,1\}^d$ and the optimal hypothesis can be encoded using κ bits, they showed learning algorithms with constant error using a near-minimal number of examples, $\tilde{O}(\kappa)$, require $\tilde{\Omega}(d\kappa)$ bits of memory. Moreover, for a general number N of examples, their memory lower bound takes the form $\tilde{\Omega}(d\kappa \cdot \frac{\kappa}{N})$.

However, as mentioned by Brown, Bun, and Smith (COLT 2022), the learning process often involves multiple passes over the data. Hence, it is equally important to study the space complexity in the *multi-pass streaming* setting. The authors conjectured that similar lower bounds should apply but left it as an open problem. In this paper, we resolve this open problem by proving that any L-pass streaming algorithm using N samples requires $\tilde{\Omega}(d\kappa \cdot \frac{\kappa}{NL})$ bits of memory. Intuitively, our lower bound shows that a stream of $L \cdot N$ fresh examples is at least as useful as L passes over N examples.

A key component of our approach is a lower bound on the information complexity of the $\mathsf{Bit\text{-}Bias}(p,q)$ problem in the multi-pass streaming setting, a basic problem that may have independent significance. In the $\mathsf{Bit\text{-}Bias}(p,q)$ problem, one sees a stream of N i.i.d. random bits drawn from either $\mathsf{Bernoulli}(p)$ or $\mathsf{Bernoulli}(q)$, and would like to distinguish the two cases. Our results not only extend the previous lower bound on $\mathsf{Bit\text{-}Bias}(0,1/2)$ by Brown, Bun, and Smith from the one-pass streaming setting to the more general multi-pass setting, but also cover more general values of p and q.

Keywords: Multi-pass streaming algorithm, Agnostic learning, Complexity theory, Information theory

1. Introduction

Streaming algorithms for learning problems have been extensively studied in both theoretical contexts (Alon et al., 1996; Assadi et al., 2020; Ben-Eliezer et al., 2021; Nelson and Yu, 2019; Lovett and Zhang, 2023; Larsen et al., 2015; Mahabadi et al., 2020; Cohen-Addad et al., 2023; Woodruff et al., 2023; Diakonikolas et al., 2023) and practical applications (Vaswani et al., 2017; Brown et al., 2020; Ramesh et al., 2021; Xiao et al., 2024). Notably, recent successful complex machine learning models are typically trained on massive datasets in a streaming manner. The learning algorithms process samples sequentially, handling them one at a time while continuously updating the model as training progresses. Several papers (Vaswani et al., 2017; Brown et al., 2020; Ramesh et al., 2021) have highlighted that the performance of such models is often constrained by memory limitations, which can be as significant as the time complexity during training.

In this context, we focus on the learning process with bounded space. Specifically, we consider a scenario involving a data stream of length N, denoted by $Z = (Z_1 = (X_1, Y_1), Z_2 = (X_2, Y_2), \ldots, Z_N = (X_N, Y_N))$, independently sampled from an unknown distribution P on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Here, $\mathcal{X}$ represents the set of possible samples and $\mathcal{Y}$ denotes the set of possible labels. The learning algorithm processes the data stream in one or multiple passes and is then required to output a hypothesis function $h : \mathcal{X} \to \mathcal{Y}$ from a predefined hypothesis class $\mathcal{H}$, with the goal of minimizing the error of h under P. For discrete-valued functions, the error of hypothesis h is defined as $\mathrm{err}_P(h) := \Pr_{(x,y)\sim P}[h(x) \neq y]$. For real-valued functions (e.g., when $\mathcal{Y} = [0, 1]$), the error is defined as $\mathrm{err}_P(h) := \mathbb{E}_{(x,y)\sim P}[|h(x) - y|]$. Furthermore, the error of a learning algorithm $\mathcal{A}$ with stream length N under P is defined as:

$$\mathrm{err}_{P^{\otimes N}}(\mathcal{A}) := \mathbb{E}_{Z\sim P^{\otimes N}}\left[\mathrm{err}_P\left(\mathcal{A}\left(Z\right)\right)\right].$$

We simplify the notation to $\mathrm{err}_P(\mathcal{A})$ when N is clear from the context. Fix a class $\mathcal{H}$ of hypotheses from $\mathcal{X} \to \mathcal{Y}$. We say that the algorithm $\mathcal{A}$ agnostically learns $\mathcal{H}$ within error ϵ if, for every distribution P on $\mathcal{X} \times \mathcal{Y}$, the algorithm $\mathcal{A}$ satisfies:

$$\mathrm{err}_P(\mathcal{A}) - \inf_{h\in\mathcal{H}} \{\mathrm{err}_P(h)\} \leq \epsilon.$$

Let κ denote the size of the hypothesis class, i.e., $\kappa = \log|\mathcal{H}|$,[1] let $d = \log|\mathcal{X}|$ be the data dimension, and N denote the number of samples (i.e., the length of the data stream). For learning with unbounded space, classical results show that $O(\kappa)$ samples are sufficient to agnostically learn $\mathcal{H}$ within a small constant error. Therefore, there exists a streaming algorithm running on a data stream of length $N = O(\kappa)$ and using $O(\kappa d)$ space, which simply stores the entire stream and attains low expected error.

However, there are relatively fewer results about learning with bounded space. In this direction, Brown et al. (2022) studied *one-pass* space lower bounds for agnostically learning several natural hypothesis classes and proved that the aforementioned naive learning algorithm is optimal. Specifically, they first established a tight one-pass space lower bound for a basic learning problem, referred to as *the core problem* (Theorem 2). Then, by reductions, they obtained tight one-pass space lower bounds for several other natural classes, including:

- *Direct sums of k indicators:* Let $\mathcal{X} = [k] \times \{0, 1\}^{d'}$ (for $d' = d - \log_2 k$ so inputs can be described with d bits) and consider classifiers $h_{i_1,\cdots,i_k}$ specified by k indices in $[d']$, where $h_{i_1,\cdots,i_k}(j, x) = x_{i_j}$ (that is, for each j there is a single bit of x that determines the label).

- *Sparse linear classifiers over degree-2 polynomial features:* Let $\mathcal{X} = \{0, 1\}^d$, and consider classifiers of the form $h(x) = \mathrm{sign}(\langle w, \phi(x)\rangle)$ where $\phi(x)$ denotes a degree-2 monomial in the entries of x (so each entry of $\phi(x)$ equals $x_i x_j$ for two indices $i, j \in [d]$) and $w \in \{0, 1\}^{\binom{d}{\leq 2}}$ has at most k non-zero entries.

[1] For continuous function spaces, κ should be interpreted as the bit length of an appropriate discrete representation of functions in $\mathcal{H}$.

- *Multiclass sparse linear classifiers:* Let $\mathcal{X} = \{0,1\}^d$ and $\mathcal{Y} = [k]$ (so there are k distinct labels). Let $\mathcal{H}$ comprise all functions of the form $h(x) = \arg\max_{j\in[k]}\langle w_j, x\rangle$ where each $w_j \in \{0,1\}^d$ has $O(\log k)$ nonzero entries. The combining function can also be taken to be a "softmax" instead of the exact argmax.
- *Real-valued regression:* Let $\mathcal{X} = \{0,1\}^d$ and $\mathcal{Y} = [0,1]$. Consider functions realizable by a sparse two-layer neural network with a single hidden layer of k ReLU nodes, each of which is connected to at most $O(\log k)$ input nodes. The weights on the wires in the first layer are 0 or 1, and those in the second layer are $\left\{0, \frac{1}{k-1}, \frac{1}{k-1}, \cdots, 1\right\}$.

For each of those settings, $\kappa = \tilde{\Theta}(k \log d)$, and Brown et al. (2022) showed that every one-pass streaming algorithm that learns any function class above requires $\tilde{\Omega}(\kappa^2 d/N)$ bits of space, which matches the aforementioned upper bound $O(\kappa d)$ when $N = O(\kappa)$.

Now, we present the definition of the core problem and the one-pass space lower bound established by Brown et al. (2022).

Definition 1 (The core problem) *The core problem is parameterized by positive integers d, k, and $\rho \le d$. The learning algorithm receives N i.i.d. samples from either the distribution P_0 or P_1, and aims to distinguish between the two distributions. Here, P_0 and P_1 are both defined on $[k] \times \{0,1\}^d$ (a "subpopulation identifier" in $[k]$ and a feature vector in $\{0,1\}^d$):*

- $\underline{P_0}$*: the uniform distribution on $[k] \times \{0,1\}^d$;*
- $\underline{P_1}$*: the uniform mixture of k (randomly picked) sub-distributions. Here, each sub-distribution $j \in [k]$ is associated with a (random) set $I_j \subseteq [d]$ of size at most ρ and a (random) pattern $b_j \in \{0,1\}^{I_j}$. To generate I_j and b_j, we uniformly sample r_j from $\{0,1,\cdots,\rho\}$, sample I_j from all possible sets of size r_j, and then sample b_j from $\{0,1\}^{I_j}$. Once all I_j and b_j are determined, the j-th sub-distribution is defined as follows: to draw a sample, which is an element of $[k] \times \{0,1\}^n$,*
  - *the subpopulation identifier is fixed to j, and the coordinates in I_j are fixed to the corresponding values in b_j;*
  - *all remaining coordinates are independently set to 0 or 1 uniformly at random.*

Theorem 2 (Brown et al. (2022)) *Any one-pass streaming algorithm that solves the core problem with N samples, d dimensions, k components, and at most $\rho = o(d^{1/4})$ fixed features, with a constant advantage*[2] *requires space $\Omega\left(\frac{k^2 d}{N\rho^4}\right) = \Omega\left(dk \cdot \frac{k}{N} \cdot \frac{1}{\rho^4}\right)$.*

In addition, they also proposed a simple one-pass streaming algorithm that matches this lower bound: for $N = \Omega(k\log(d))$ and $1 \le \rho \le d$, the algorithm solves the core problem with a constant advantage while using space $O(dk \cdot \frac{k}{N} \cdot \frac{1}{\rho})$. In particular, if $\rho = \operatorname{poly}\log(d)$, the bounds simplify to $\tilde{\Omega}(k \cdot d/T)$ where $T = N/k$. As N/k is the expected number of samples from each subpopulation, Theorem 2 indicates that: the space required to distinguish a specific subpopulation using T samples scales as d/T, and there is no better way to solve the larger problem than to learn each subpopulation individually.

[2] The precise definition of advantage is given in Section 2. Intuitively, it represents the probability of successfully distinguishing between the two distributions.

Remark 3 *As mentioned by Brown et al. (2022), in addition to guiding the design of practical streaming learning algorithms, the streaming lower bounds suggest an explanation for the empirical success of the widely adopted distillation or pruning strategy: first train a large dense network, and then reduce it to a smaller one through distillation or pruning. This strategy is commonly justified by the notion that optimization is easier in a larger space (Frankle et al., 2021; Bartlett et al., 2021). However, streaming lower bounds offer an alternative perspective: Training algorithms for sparse models sometimes need space proportional to the ambient dimension of the natural encoding, rather than the size of the examples or the final classifier. In other words, the larger parameter vector enables the training process to encode information whose relevance to the problem can only be understood later.*

The space lower bounds in (Brown et al., 2022), including Theorem 2, apply only to one-pass streaming algorithms. However, in practice, machine learning models are typically trained by iterating over the data stream multiple times. The techniques in (Brown et al., 2022) do not extend to the multi-pass setting, and the authors identify the following challenging open problem:

> *Question 1: Do similar lower bounds hold in the multi-pass setting, or does allowing multiple passes substantially improve the efficiency of streaming algorithms for the learning tasks?*

1.1. Our contributions

In this paper, we provide an affirmative answer to Question 1.

Theorem 4 *Any (randomized) L-pass streaming algorithm that solves the core problem with N samples, d dimensions, k components and at most $\rho = o(\sqrt{d})$ fixed features, with constant advantage requires at least $\Omega(\frac{k^2 d}{NL\rho^2})$ bits of memory.*

Compared with Theorem 2, our lower bound applies to the more general multi-pass streaming algorithms. Moreover, even in the one-pass scenario, we improve the dependency on ρ from $\frac{1}{\rho^4}$ to $\frac{1}{\rho^2}$. Using similar reductions as those in (Brown et al., 2022), Theorem 4 directly yields multi-pass streaming lower bounds for the other four natural learning problems.

Corollary 5 *Any (randomized) L-pass streaming algorithm that agnostically learns the function classes of direct sums of k indicators, sparse linear classifiers over degree-2 polynomial features, multiclass sparse linear classifiers, or real-valued regression, within a constant error requires $\tilde{\Omega}\left(\kappa d \cdot \frac{\kappa}{NL}\right)$ bits of memory.*

We note that when L is a constant, all our lower bounds are tight (up to logarithmic factors).

The above two theorems show that allowing a constant (or even polylog) number of passes can not substantially improve the efficiency of streaming algorithms for the aforementioned learning tasks.

1.1.1. INFORMATION COMPLEXITY OF BIT-BIAS

The key component of the proof for Theorem 4 is a lower bound for the multi-pass information complexity of the Bit-Bias problem defined below. The lower bound for the Bit-Bias problem may be of independent interest due to its proof technique and potential applications in deriving other streaming lower bounds.

Definition 6 ($\textsf{Bit-Bias}(p, q)$ problem) *Consider a data stream comprising N i.i.d. samples from either $\textsf{Bernoulli}(p)$ or $\textsf{Bernoulli}(q)$. The goal is to distinguish them via the data stream. Without loss of generality, we assume $p < q$ throughout this paper.*

In our proof, we adopt the multi-pass information complexity measure proposed by Braverman et al. (2024). We first recall their notion. Let $\mathcal{A}$ be an L-pass streaming algorithm running on a data stream $X = (X_1, \ldots, X_N)$. For each $i \in [N]$ and $r \in [L]$, we use $M_{i,r}$ to denote the memory state of $\mathcal{A}$ after processing X_i in the r-th pass. Note that $M_{i,r}, \cdots, M_{N,r}$ are random variables depending on X and the private randomness of $\mathcal{A}$. For convenience, we define $M_{0,r}$ as $M_{N,r-1}$.

Definition 7 (Multi-pass information complexity (Braverman et al., 2024)) *For an L-pass streaming algorithm $\mathcal{A}$ running on i.i.d. samples $X = X_1, \ldots, X_N$, we define its multi-pass information complexity as*

$$\begin{aligned}\mathrm{IC}(\mathcal{A}, X) :=& \sum_{r=1}^{L}\sum_{i=1}^{N}\sum_{k=1}^{i} I(M_{i,r}; X_k \mid M_{k-1,\le r}, M_{i,\le r-1}) \\ &+ \sum_{r=1}^{L}\sum_{i=1}^{N}\sum_{k=i+1}^{N} I(M_{i,r}; X_k \mid M_{k-1,\le r-1}, M_{i,\le r-1}),\end{aligned}$$

where $M_{i,\le r}$ denotes $(M_{i,1}, M_{i,2}, \cdots, M_{i,r})$.

Now we introduce our main technical lemma. We say a multi-pass streaming algorithm $\mathcal{A}$ solves the $\textsf{Bit-Bias}(p, q)$ with advantage δ, if

$$\Pr_{X\sim\mathcal{D}_p}[\mathcal{A}(X) = 1] - \Pr_{X\sim\mathcal{D}_q}[\mathcal{A}(X) = 1] \ge \delta.$$

Here, $\mathcal{D}_p$ denotes the distribution of N i.i.d. samples from $\textsf{Bernoulli}(p)$, assuming N is clear from the context. Similar definitions also apply to $\mathcal{D}_q$.

Previously, Brown et al. (2022) used techniques and properties restricted to one-pass streaming algorithms to prove the following result:

Theorem 8 (Brown et al. (2022)) *If a one-pass ($L = 1$) streaming algorithm $\mathcal{A}$ solves the $\textsf{Bit-Bias}(0, 1/2)$ with advantage δ, we have*

$$\mathrm{IC}(\mathcal{A}, X) \ge \Omega\left(\delta^4\right)$$

where $X \sim \mathcal{D}_{1/2}$.

We use different techniques to prove the following more general result:

Theorem 9 *If a multi-pass streaming algorithm $\mathcal{A}$ solves the $\textsf{Bit-Bias}(p, q)$ with advantage δ, we have*

$$\mathrm{IC}(\mathcal{A}, X) \ge \Omega\left(\min(q, 1-q) \cdot \frac{q^2\delta^2}{(q-p)^2}\right),$$

where $X \sim \mathcal{D}_q$. In addition, flipping the input, we get a similar lower bound

$$\mathrm{IC}(\mathcal{A}, X) \ge \Omega\left(\min(p, 1-p) \cdot \frac{(1-p)^2\delta^2}{(q-p)^2}\right),$$

where $X \sim \mathcal{D}_p$.

We improve their results from the following aspects:

1. We extend their findings to multi-pass settings, which is our most significant improvement.
2. We enhance the dependence on δ, leading to stronger lower bounds.
3. Our results cover more general values of p and q, and we establish a stronger lower bound when p is close to q.

The first two improvements allow us to obtain better lower bounds for the core problem, as mentioned earlier. Furthermore, the extension to general p and q makes our approach a more versatile tool for deriving other stream lower bounds through reductions in the future.

1.2. Related works

Multi-pass space lower bounds. How to prove space lower bounds for multi-pass streaming algorithms is always a challenging problem in the area of streaming algorithms. A series of works on space lower bounds (Raz, 2018; Diakonikolas et al., 2019; Braverman et al., 2020, 2021; Brown et al., 2022) are restricted to the one-pass setting due to some technical barriers. They also left the multi-pass space lower bounds as major open problems and future directions. However, known techniques that apply to multi-pass lower bounds are limited. One is to build reductions to communication complexity. However, reductions often fail to give tight bounds. The other one is called *pass elimination* proposed by Guha and McGregor (2008), which is adapted from the *round elimination* technique in communication complexity (see e.g., (Miltersen et al., 1995; Sen, 2003)). These two techniques seem hard to apply to the core problem. A recent advance by Braverman et al. (2024) on multi-pass space lower bounds proposed a new information complexity notion and successfully obtained tight lower bounds via an information theoretical approach for both the needle problem and the coin problem. As we mentioned, they proposed a new information complexity notion for multi-pass streaming algorithms.

Space lower bounds for other learning problems. A closely related line of works focused on space lower bounds for problems defined over finite fields, such as learning parities. In the breakthrough result, Raz (2018) proved that any one-pass streaming algorithm that solves the learning parity functions[3] requires $\Omega(d^2)$ bits of space or $2^{\Omega(d)}$ samples. Some follow-up works by Garg et al. (2018, 2019); Lyu et al. (2023) further generalized and extended this result to the multi-pass setting. However, these lower bounds do not obviously imply bounds for continuous function classes that are normally used in practice.

There is another series of works focusing on streaming lower bounds for statistical problems. Our information complexity notion directly comes from the paper by Braverman et al. (2024), in which they studied both *coin problem* and *needle problem*. In the coin problem, the streaming algorithm $\mathcal{A}$ receives a length N data stream of N independently random coins, and its goal is to determine whether the coins are unbiased or $(1/2 + 1/\sqrt{N}, 1/2 - 1/\sqrt{N})$ biased. Braverman et al. (2020, 2021) proved tight lower bounds for one-pass setting, and Braverman et al. (2024) gave the first constant-pass space lower bounds. In the needle problem, the streaming algorithm $\mathcal{A}$ receives a length-N data stream of N independently random samples from $[n]$, and its goal again is to distinguish between the uniform distribution and the needle distribution (first uniformly sample

[3] A parity function $p_x : \mathbb{F}_2^d \to \mathbb{F}_2$ where $x \in \mathbb{F}_2^d$ is defined as $p_x(y) = \langle x, y \rangle$ for every $y \in \mathbb{F}_2^d$.

a needle $x \in [n]$, then each sample with probability p equals the needle and otherwise is uniformly sampled from $[n]$). Lovett and Zhang (2023) gave a near-optimal lower bound via a reduction, and was improved by Braverman et al. (2024) to a tight lower bound.

In the work of Chen et al. (2022), they studied the space complexity of the continual learning problems. In their setting, the learning algorithm wishes to learn a function from the hypothesis classes $\mathcal{H}$ from a k-stage learning process. In the i-th stage, the learning algorithm can draw as many independent samples as it wants from a planted distribution P_i and update its memory. It is promised that there exists a hypothesis function $h \in \mathcal{H}$ with $\mathrm{err}_{P_i}(h) = 0$ for every $i \in [k]$, and the goal is to find a function h' with a small error under any distribution P_i. They proved lower bounds and upper bounds in the setting mentioned above. Notably, their results reveal a large gap between the power of single-pass and multi-pass algorithms, highlighting the importance of proving space complexity lower bounds in the multi-pass streaming setting.

2. Preliminaries

In this paper, we heavily rely on tools from information theory. Due to the page limit, a basic introduction to information theory is delayed to Appendix A. Readers who are familiar with those tools can skip that part.

2.1. Information complexity notion

In this section, we introduce the main tool adopted in this paper, *a multi-pass information complexity notion* by Braverman et al. (2024).

We start with some notations and definitions. We study L-pass streaming algorithms, denoted by $\mathcal{A}$, with memory size S running on i.i.d. samples $X = X_1, \ldots, X_N$. For an L-pass streaming algorithm $\mathcal{A}$, we use $M_{i,r}$ to denote the memory state of $\mathcal{A}$ after processing X_i in the r-th pass. Notably, we have $|M_{i,r}| \le S$ for every i, r. For convenience, we define $M_{0,r}$ to be $M_{N,r-1}$. Notice that the memory states $M_{1,1}, \cdots, M_{N,r}$ are random variables depending on randomness of the input X and the private randomness of the algorithm $\mathcal{A}$.

We use the notion of multi-pass information complexity introduced by Braverman et al. (2024). For an L-pass streaming algorithm $\mathcal{A}$, we define its multi-pass information complexity as

$$\begin{aligned}\mathrm{IC}(\mathcal{A}, X) := &\sum_{r=1}^{L}\sum_{i=1}^{N}\sum_{k=1}^{i} I(M_{i,r}; X_k \mid M_{k-1,\le r}, M_{i,\le r-1}) \\ &+ \sum_{r=1}^{L}\sum_{i=1}^{N}\sum_{k=i+1}^{N} I(M_{i,r}; X_k \mid M_{k-1,\le r-1}, M_{i,\le r-1}),\end{aligned}$$

where $M_{i,\le r}$ denotes $(M_{i,1}, M_{i,2}, \cdots, M_{i,r})$.

One important property of this notion is the established connection with space complexity of streaming algorithms:

Lemma 10 (Braverman et al. (2024)) *For any L-pass algorithm $\mathcal{A}$ using private randomness with memory size S running on N i.i.d. inputs $X_1, \cdots, X_N$, it holds that*

$$\mathrm{IC}(\mathcal{A}, X) \le 2LSN.$$

Another property also appears in our proof.

Claim 11 (Braverman et al. (2024)) *For any L-pass algorithm $\mathcal{A}$ with semi-private randomness running on N i.i.d. inputs $X = (X_1, \cdots, X_N)$, we have the "diagonal terms" of multi-pass information complexity satisfies that*

$$\sum_{r=1}^{L} \sum_{i=1}^{N} I(M_{i,r}; X_i \mid M_{i-1,\leq r}, M_{i,\leq r-1}) \geq I(M; X).$$

Here, $M := (M_{i,r})_{i,r}$.

2.2. Randomness in streaming algorithms

When we refer to randomized multi-pass streaming algorithms with *private randomness* $\mathcal{A}$, by default we mean the multi-pass streaming algorithm whose randomness R can be partitioned into $(R_{i,r})_{i,r}$ such that $\mathcal{A}$ can only access to $R_{i,r}$ when it proceeds X_i in the r-th pass. In this paper, we consider a stronger computational model called randomized multi-pass streaming algorithms with *semi-private* randomness to make our analysis easier, where the randomness used when proceeding the same X_i in different passes are shared. Formally, it means that the randomness R can be partitioned into $(R_i)_i$ such that $\mathcal{A}$ could access R_i when it proceeds X_i in any pass. Note that the notion of randomized multi-pass streaming algorithms with semi-private randomness is a stronger model than multi-pass streaming algorithms with private randomness, so the lower bounds for multi-pass streaming algorithms with semi-private randomness (e.g., Theorems 19, 27, 28) automatically imply lower bounds for multi-pass algorithms with private randomness. *Thus, in the remaining pages, streaming algorithms refer to streaming algorithms with semi-private randomness for simplicity.* We introduce the special type of randomness since it would appear in reductions. For example, as we will see in Section C.1.1, the algorithm $\mathcal{A}'$ constructed in the reduction from Task B to Task A is associated with semi-private randomness.

3. Proof Sketch

In this section, we outline the proof of our main result (Theorem 4), which establishes a space lower bound for the core problem in the multi-pass streaming setting. Building on the one-pass framework presented in (Brown et al., 2022), we employ a multi-stage reduction that transforms the core problem into the $\mathsf{Bit\text{-}Bias}(0, 1/2)$ problem introduced earlier.

A central component of our argument is Theorem 9, which provides an information complexity lower bound for $\mathsf{Bit\text{-}Bias}(p, q)$. Our novel approach to this problem is the key reason we can extend lower bounds to the multi-pass streaming setting rather than being restricted to the one-pass setting. Moreover, the lower bound itself may be of independent interest; the complete proof is presented in Section B. We will also highlight the idea behind this proof, and most details of the reductions from the core problem to the $\mathsf{Bit\text{-}Bias}$ problem can be found in Appendix C.

We now present the formal definitions of all intermediate problems used in our reductions.

3.1. Task Definition

Definition 12 (Meta-Distributions) *Parameters: three positive integers $k, d, \rho \leq d$.*

Structure distribution $\mathcal{P}$. Given $\mathcal{J} = \{j_1, \cdots, j_r\} \subseteq [d]$ and $\mathcal{B} = (b_1, \cdots, b_r) \in \{0,1\}^{\mathcal{J}}$, $\mathcal{P}_{(\mathcal{J},\mathcal{B})} \in \Delta(\{0,1\}^d)$ is a distribution over $\{0,1\}^d$ defined as follows. To sample $x \in \{0,1\}^d$ from $\mathcal{P} = \mathcal{P}_{(\mathcal{J},\mathcal{B})}$, set $x_{j_i} \leftarrow b_i$ for each $j_i \in \mathcal{J}$. For $j \notin \mathcal{J}$, set $x_j \in \{0,1\}$ uniformly and independently.

Subpopulation meta-distribution $\mathcal{Q}_{d,\rho}$. We sample a distribution $\mathcal{P}_{(\mathcal{J},\mathcal{B})} \in \Delta(\{0,1\}^d)$ from the structured subpopulation meta-distribution $\mathcal{Q}_{d,\rho} \in \Delta(\Delta(\{0,1\}^d))$ as follows:

1. *Draw $r \in \{0, \cdots, \rho\}$ uniformly.*
2. *Draw uniformly $\mathcal{J} = \{j_1, \cdots, j_r\} \subseteq [d]$ without replacement and $\mathcal{B} = (b_1, \cdots, b_r) \in \{0,1\}^{\mathcal{J}}$.*

Mixture of structure distributions $\mathcal{P}_{\mathrm{mix}}$. Given k structured distributions $\mathcal{P}_{(\mathcal{J}_1,\mathcal{B}_1)}, \cdots, \mathcal{P}_{(\mathcal{J}_k,\mathcal{B}_k)}$, define $\mathcal{P}_{\mathrm{mix}} \in \Delta([k] \times \{0,1\}^d)$ simply as a uniform mixture of the k distributions of the form $\delta_j \otimes \mathcal{P}_{(\mathcal{J}_j,\mathcal{B}_j)}$, where δ_j denotes the point mass on j.

Population meta-distribution $\mathcal{Q}_{k,d,\rho}$. We sample a distribution $\mathcal{P}_{\mathrm{mix}} \in \Delta([k] \times \{0,1\}^d)$ from the structured population meta-distribution $\mathcal{Q}_{k,d,\rho} \in \Delta(\Delta([k] \times \{0,1\}^d))$ as follows: for $j = 1, \cdots, k$, draw $\mathcal{P}_{(\mathcal{J}_j,\mathcal{B}_j)} \sim \mathcal{Q}_{d,\rho}$. Then $\mathcal{P}_{\mathrm{mix}}$ is set to be the uniform mixture of $\delta_j \otimes \mathcal{P}_{(\mathcal{J}_j,\mathcal{B}_j)}$.

Now we are ready to give the formal definitions of the core problem and three intermediate problems used in the reductions.

Definition 13 (Task Definitions) *N, k, d and $\rho \leq d$ are positive integer parameters here.*

- *(Task C, Core Problem). First sample a distribution $\mathcal{P}_{\mathrm{mix}}$ from the meta-distribution $\mathcal{Q}_{k,d,\rho}$, and the multi-pass streaming algorithm $\mathcal{A}$ receives a stream of N i.i.d. samples from $\mathcal{P}_{\mathrm{mix}}$. Finally, $\mathcal{A}$ outputs a (possibly randomized) function $m : [k] \times \{0,1\}^d \to \{0,1\}$. The advantage of $\mathcal{A}$ is defined by*

$$\mathop{\mathbb{E}}_{\substack{\mathcal{P}_{\mathrm{mix}} \sim \mathcal{Q}_{k,d,\rho} \\ X=(X_1,\cdots,X_N)\sim_{iid}\mathcal{P}_{\mathrm{mix}} \\ m \leftarrow \mathcal{A}(X)}} \left[\Pr_{(j,y)\sim\mathcal{P}_{\mathrm{mix}}}[m(j,y)=1] - \Pr_{(j,y)\sim\mathcal{U}}[m(j,y)=1] \right].$$

- *(Task B, Single Subpopulation). Sample a distribution $\mathcal{P}$ from the meta-distribution $\mathcal{Q}_{d,\rho}$. The multi-pass streaming algorithm $\mathcal{A}$ receives a stream of T i.i.d. samples from $\mathcal{P}$, then outputs 0 or 1. The advantage of $\mathcal{A}$ is defined by*

$$\Pr_{\substack{\mathcal{P}\sim\mathcal{Q}_{d,\rho} \\ X=(X_1,\cdots,X_T)\sim_{iid}\mathcal{P}}} [\mathcal{A}(X)=1] - \Pr_{X=(X_1,\cdots,X_T)\sim_{iid}\mathcal{U}} [\mathcal{A}(X)=1].$$

- *(Task B', Single Subpopulation). Sample a distribution $\mathcal{P}$ from the meta-distribution $\mathcal{Q}_{d,\rho}$. The multi-pass streaming algorithm $\mathcal{A}$ receives a stream of T i.i.d. samples from $\mathcal{P}$, then outputs a (possibly randomized) function $m : \{0,1\}^d \to \{0,1\}$. The advantage of $\mathcal{A}$ is defined by*

$$\mathop{\mathbb{E}}_{\substack{\mathcal{P}\sim\mathcal{Q}_{d,\rho} \\ X=(X_1,\cdots,X_T)\sim_{iid}\mathcal{P} \\ m\leftarrow\mathcal{A}(X)}} \left[\Pr_{y\sim\mathcal{P}}[m(y)=1] - \Pr_{y\sim\mathcal{U}}[m(y)=1] \right].$$

Intuitively, in Task B' the algorithm receives T structured inputs and must distinguish a structure test example from a uniform one, while in Task B the algorithm must distinguish whether its inputs are all structured or all uniform.

- *(Task A,* Bit-Bias*(0,1/2)). A multi-pass streaming algorithm $\mathcal{A}$ receives a stream of N bits, and outputs 0 or 1. The advantage of $\mathcal{A}$ is*

$$\Pr_{X=(X_1,\cdots,X_T)=0^T}[\mathcal{A}(X)=1] - \Pr_{X=(X_1,\cdots,X_T)\sim_{iid}\mathcal{U}}[\mathcal{A}(X)=1].$$

All the advantages defined above are assumed to be non-negative. This is without loss of generality, since otherwise we can simply flip the output of $\mathcal{A}$ or the function m.

As we discussed before, we generalize the Bit-Bias$(0,1/2)$ problem to Bit-Bias(p,q) for any $p<q$.

Definition 14 (Bit-Bias(p,q)) *A multi-pass streaming algorithm $\mathcal{A}$ receives a stream of N bits, and outputs 0 or 1. The advantage of $\mathcal{A}$ is*

$$\Pr_{X=(X_1,\cdots,X_T)\sim\mathcal{D}_p}[\mathcal{A}(X)=1] - \Pr_{X=(X_1,\cdots,X_T)\sim\mathcal{D}_q}[\mathcal{A}(X)=1].$$

3.2. Information Complexity Lower Bound for Bit-Bias(p,q)

In this subsection, we provide an overview of the proof for obtaining information complexity lower bound for the Bit-Bias(p,q) problem, which is also the main technical contribution of this paper. Details of the proofs are provided in Appendix B. First, we recall our theorem for Bit-Bias(p,q) problem:

Theorem 9 *If a multi-pass streaming algorithm $\mathcal{A}$ solves the Bit-Bias(p,q) with advantage δ, we have*

$$\mathrm{IC}(\mathcal{A},X)\geq\Omega\left(\min(q,1-q)\cdot\frac{q^2\delta^2}{(q-p)^2}\right),$$

where $X\sim\mathcal{D}_q$. In addition, flipping the input, we get a similar lower bound

$$\mathrm{IC}(\mathcal{A},X)\geq\Omega\left(\min(p,1-p)\cdot\frac{(1-p)^2\delta^2}{(q-p)^2}\right),$$

where $X\sim\mathcal{D}_p$.

Now we explain the main idea of proving Bit-Bias(p,q) lower bounds. It consists of two steps.

1. First, we prove a lower bound on the information cost of the communication version of the Bit-Bias$(0,q)$ problem.
2. Then, based on the lower bound of Bit-Bias$(0,q)$, we prove the multi-pass information complexity lower bound for the general Bit-Bias(p,q) problem using a decomposition and reduction argument.

The communication version of $\textsf{Bit-Bias}(0, q)$ problem. Unlike the Theorem 9, which shows the streaming information complexity of $\textsf{Bit-Bias}(0, q)$, we consider $\textsf{Bit-Bias}(0, q)$ as an N-party communication problem in the blackboard model: there are N players, each holding a bit $X_i \in \{0, 1\}$, and aiming to distinguish whether X_i's are from $\textsf{Bernoulli}(0)$ or $\textsf{Bernoulli}(q)$; a communication protocol proceeds in rounds, and in each round, one of the player writes a message on the blackboard seen by all the players.

We show an information complexity lower bound to distinguish these two distributions even in the communication setting. The theorem is formalized as follows.

Theorem 15 *Suppose Π is a protocol solving the $\textsf{Bit-Bias}(0, q)$ problem with advantage δ. Then its information cost under $\mathcal{D}_q$ is at least $\frac{1}{2}\min(q, 1-q)\cdot\delta^2$, i.e.,*

$$I(\Pi(X); X) \geq \frac{1}{2}\min(q, 1-q)\cdot\delta^2.$$

Here, $\Pi(X)$ denotes the transcript induced by running Π on input X.

Our proof of Theorem 15 is inspired by Jayram (2009)'s argument to lower bound the information cost of the AND function. However, there is an essential difference between our setting and Jayram's such that Jayram's method cannot be directly applied: the information cost is defined under the distribution $\mathcal{D}_q$ in our setting, whereas it is under $X \in_R \{0^T, e_i\}$ in Jayram's setting, where $e_i \in \{0, 1\}^T$ denotes the unit vector with the i-th coordinate equaling 1 and other coordinates equaling 0. Our proof adapts and generalizes Jayram's techniques. Specifically, the main tool of our proof is a generalized cut-and-paste property from the rectangle property of communication protocols (see Lemma 21 in the appendix). Based on the property, by extending Jayram's argument, together with some observations in information theory, we managed to show $I(\Pi(X); X) \geq \min(q, 1-q)\cdot\delta^2$ for the $\textsf{Bit-Bias}(0, q)$ problem when $X \sim \mathcal{D}_q$.

Then, by a reduction from streaming algorithms to communication protocols and then by Claim 11, we show that $\mathrm{IC}(\mathcal{A}, X) \geq I(M; X) = I(\Pi(X); X) \geq \min(q, 1-q)\cdot\delta^2$. Detailed proofs are contained in Section B.1.

The $\textsf{Bit-Bias}(p, q)$ problem. Our proof is by a decomposition of the $\textsf{Bit-Bias}(p, q)$ problem and a reduction to $\textsf{Bit-Bias}(0, q)$ problem. Precisely, we first decompose $\textsf{Bit-Bias}(p, q)$ into many *local $\textsf{Bit-Bias}(0, q)$ problems* by redefining sampling process of $X_1, \cdots, X_n \sim \mathcal{D}_p$ as follows:

1. sample a set $S \subseteq [n]$ with each element $j \in [n]$ contained in S independently with probability $\frac{q-p}{q}$.
2. for each $j \notin S$, $X_j \sim \textsf{Bernoulli}(q)$;
3. for each $j \in S$, $X_j \sim \textsf{Bernoulli}(0)$ (which means $X_j = 0$).

It is easy to verify that the sampling process is identical to $\mathcal{D}_p$. Thus, solving $\textsf{Bit-Bias}(p, q)$ is equivalent to solving $\textsf{Bit-Bias}(0, q)$ with hidden locations S. Then, we define *local $\mathcal{D}_0$ distribution* $\mathcal{D}_0^S$ as the distribution $\mathcal{D}_p$ condition on that the set sampled in the first step equals S, and define *local $\textsf{Bit-Bias}(0, q)$ problem* as distinguishing between $\mathcal{D}_0^S$ and $\mathcal{D}_q$ within a small error. Since the streaming algorithm $\mathcal{A}$ does not know S, if $\mathcal{A}$ solves the $\textsf{Bit-Bias}(p, q)$ problem, $\mathcal{A}$ has to solve local $\textsf{Bit-Bias}(0, q)$ problems for at least a constant fraction of S at once. Together with the "composability" of the multi-pass information cost, it allows us to derive lower bounds for Bit-Bias(p, q).

To be more specific, suppose $\mathcal{A}$ can solve all the local Bit-Bias$(0, q)$ problems with advantage δ at once. Let $S = \{p_1, p_2, \cdots, p_m\}$, then we have

$$\sum_{r=1}^{L-1}\sum_{i=1}^{m}\sum_{k=1}^{i} I(M_{p_{i+1}-1,r}; X_{p_k} \mid M_{p_k-1,\le r}, M_{p_{i+1}-1,\le r-1}) = \Omega(\min(q, 1-q) \cdot \delta^2).$$

where we define p_{m+1} as p_1. By taking an expectation over S, the L.H.S. of the inequality above is about $\mathrm{IC}(\mathcal{A}, X) \cdot \left(\frac{q-p}{q}\right)^2$ since each term

$$I(M_{i,j}; X_\ell \mid M_{\le i,\ell-1}, M_{\le i-1,j}) \text{ or } I(M_{i,j}; X_\ell \mid M_{\le i-1,\ell-1}, M_{\le i-1,j})$$

in $\mathrm{IC}(\mathcal{A}, X)$ appears in the L.H.S. for an $\left(\frac{q-p}{q}\right)^2$ fraction of S. Consequently, we have $\mathrm{IC}(\mathcal{A}, X) = \left(\frac{q}{q-p}\right)^2 \cdot \Omega(\min(q, 1-q) \cdot \delta^2)$ as desired.

3.3. Results for Other Tasks

With the information complexity lower bound for Bit-Bias$(0, 1/2)$, we obtain information complexity lower bounds (see Theorems 27 and 28) for Task B and B' following the reductions in (Brown et al., 2022). The reduction from Task A to Task B (Bit-Bias) is in Appendix C.1.1, and the reduction from Task B to Task B' is in Appendix C.1.2.

The reduction from the Task B' to Task C is also inspired by Brown et al. (2022), but our analysis makes the argument clearer by adopting a decomposition-and-reduction approach. The intuition is as follows:

- A data stream generated in Task C can be decomposed into k streams of Task B' each located in random positions, and any algorithms solving Task C can also solve most of the local Task B's simultaneously.

Note that this reduction is similar in intuition to the reduction from Bit-Bias$(0, q)$ to Bit-Bias(p, q) discussed earlier.

Organizations of Appendix. In Section B, we prove the information cost lower bound for Bit-Bias. In Section C.1, we establish the lower bounds for Task B and Task B'. In Section C.2, we prove the memory lower bound for the core problem. We finish the whole proof for Theorem 4 in Section C.3.

Acknowledgments

Qian Li's work was supported by Hetao Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone Project (No.HZQSWS-KCCYB-2024016). Jiapeng Zhang is supported by NSF CAREER award 2141536. We thank anonymous COLT reviewers for pointing us to relevant references.

References

Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pages 20–29, 1996.

Sepehr Assadi, Gillat Kol, Raghuvansh R. Saxena, and Huacheng 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, 2020. doi: 10.1109/FOCS46700.2020.00041.

Peter L. Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *Acta Numer.*, 30:87–201, 2021. doi: 10.1017/S0962492921000027. URL https://doi.org/10.1017/S0962492921000027.

Omri Ben-Eliezer, Rajesh Jayaram, David P. Woodruff, and Eylon Yogev. A framework for adversarially robust streaming algorithms. *SIGMOD Rec.*, 50(1):6–13, jun 2021. ISSN 0163-5808. doi: 10.1145/3471485.3471488. URL https://doi.org/10.1145/3471485.3471488.

Mark Braverman, Ankit Garg, Tengyu Ma, Huy L. Nguyen, and David P. Woodruff. Communication lower bounds for statistical estimation problems via a distributed data processing inequality. In *Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing*, STOC '16, page 1011–1020, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450341325. doi: 10.1145/2897518.2897582. URL https://doi.org/10.1145/2897518.2897582.

Mark Braverman, Sumegha Garg, and David P Woodruff. The coin problem with applications to data streams. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science*, pages 318–329. IEEE, 2020.

Mark Braverman, Sumegha Garg, and Or Zamir. Tight space complexity of the coin problem. In *62nd IEEE Annual Symposium on Foundations of Computer Science, FOCS 2021, Denver, CO, USA, February 7-10, 2022*, pages 1068–1079. IEEE, 2021. doi: 10.1109/FOCS52979.2021.00106. URL https://doi.org/10.1109/FOCS52979.2021.00106.

Mark Braverman, Sumegha Garg, Qian Li, Shuo Wang, David P Woodruff, and Jiapeng Zhang. A new information complexity measure for multi-pass streaming with applications. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 1781–1792, 2024.

Gavin Brown, Mark Bun, and Adam D. Smith. Strong memory lower bounds for learning natural models. In Po-Ling Loh and Maxim Raginsky, editors, *Conference on Learning Theory, 2-5 July 2022, London, UK*, volume 178 of *Proceedings of Machine Learning Research*, pages 4989–5029. PMLR, 2022. URL https://proceedings.mlr.press/v178/brown22a.html.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Xi Chen, Christos Papadimitriou, and Binghui Peng. Memory bounds for continual learning. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 519–530. IEEE, 2022.

Vincent Cohen-Addad, David P. Woodruff, and Samson Zhou. Streaming euclidean k-median and k-means with $o(\log n)$ space, 2023. URL https://arxiv.org/abs/2310.02882.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Ilias Diakonikolas, Themis Gouleakis, Daniel M Kane, and Sankeerth Rao. Communication and memory efficient testing of discrete distributions. In *Conference on Learning Theory*, pages 1070–1106. PMLR, 2019.

Ilias Diakonikolas, Daniel Kane, Ankit Pensia, and Thanasis Pittas. Nearly-linear time and streaming algorithms for outlier-robust pca. In *International Conference on Machine Learning*, pages 7886–7921. PMLR, 2023.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M. Roy, and Michael Carbin. Pruning neural networks at initialization: Why are we missing the mark? In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=Ig-VyQc-MLK.

Sumegha Garg, Ran Raz, and Avishay Tal. Extractor-based time-space lower bounds for learning. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 990–1002, 2018.

Sumegha Garg, Ran Raz, and Avishay Tal. Time-space lower bounds for two-pass learning. *Electron. Colloquium Comput. Complex.*, TR19-071, 2019. URL https://eccc.weizmann.ac.il/report/2019/071.

Sudipto Guha and Andrew McGregor. Tight lower bounds for multi-pass stream computation via pass elimination. In Luca Aceto, Ivan Damgård, Leslie Ann Goldberg, Magnús M. Halldórsson, Anna Ingólfsdóttir, and Igor Walukiewicz, editors, *Automata, Languages and Programming*, pages 760–772, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. ISBN 978-3-540-70575-8.

T. S. Jayram. Hellinger strikes back: A note on the multi-party information complexity of AND. In Irit Dinur, Klaus Jansen, Joseph Naor, and José D. P. Rolim, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, 12th International Workshop, APPROX 2009, and 13th International Workshop, RANDOM 2009, Berkeley, CA, USA, August 21-23, 2009. Proceedings*, volume 5687 of *Lecture Notes in Computer Science*, pages 562–573. Springer, 2009. doi: 10.1007/978-3-642-03685-9_42. URL https://doi.org/10.1007/978-3-642-03685-9_42.

Kasper Green Larsen, Jelani Nelson, and Huy L. Nguyên. Time lower bounds for nonadaptive turnstile streaming algorithms. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, STOC '15, page 803–812, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450335362. doi: 10.1145/2746539.2746542. URL https://doi.org/10.1145/2746539.2746542.

Shachar Lovett and Jiapeng Zhang. Streaming lower bounds and asymmetric set-disjointness. In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 871–882. IEEE, 2023.

Xin Lyu, Avishay Tal, Hongxun Wu, and Junzhao Yang. Tight time-space lower bounds for constant-pass learning. In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1195–1202, 2023. doi: 10.1109/FOCS57990.2023.00070.

Sepideh Mahabadi, Ilya Razenshteyn, David P. Woodruff, and Samson Zhou. Non-adaptive adaptive sampling on turnstile streams. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2020, page 1251–1264, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369794. doi: 10.1145/3357713.3384331. URL https://doi.org/10.1145/3357713.3384331.

Peter Bro Miltersen, Noam Nisan, Shmuel Safra, and Avi Wigderson. On data structures and asymmetric communication complexity. In *Proceedings of the Twenty-Seventh Annual ACM Symposium on Theory of Computing*, STOC '95, page 103–111, New York, NY, USA, 1995. Association for Computing Machinery. ISBN 0897917189. doi: 10.1145/225058.225093. URL https://doi.org/10.1145/225058.225093.

Jelani Nelson and Huacheng Yu. Optimal lower bounds for distributed and streaming spanning forest computation. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '19, page 1844–1860, USA, 2019. Society for Industrial and Applied Mathematics.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

Ran Raz. Fast learning requires good memory: A time-space lower bound for parity learning. *J. ACM*, 66(1), dec 2018. ISSN 0004-5411. doi: 10.1145/3186563. URL https://doi.org/10.1145/3186563.

P. Sen. Lower bounds for predecessor searching in the cell probe model. In *18th IEEE Annual Conference on Computational Complexity, 2003. Proceedings.*, pages 73–83, 2003. doi: 10.1109/CCC.2003.1214411.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

David Woodruff, Fred Zhang, and Samson Zhou. On robust streaming for learning with experts: algorithms and lower bounds. *Advances in Neural Information Processing Systems*, 36:79518–79539, 2023.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks, 2024. URL https://arxiv.org/abs/2309.17453.

Appendix A. Basics of Information Theory

Our proof follows an information theoretical approach. Thus, we introduce some background knowledge about information theory and squared Hellinger distance adopted in this paper.

Basics of information theory. Given a random variable Z, $\mathrm{H}(Z)$ denotes the *Shannon entropy* of Z, that is, $\mathrm{H}(Z) = \sum_z \Pr(Z = z)\log(1/\Pr(Z = z))$. We also use $\mathrm{H}(\mathcal{P})$ to denote the entropy of the probability distribution $\mathcal{P}$. For two random variables X, Y, we define the *mutual information* between X and Y by $I(X;Y) := \mathrm{H}(X) - \mathrm{H}(X|Y) = \mathrm{H}(Y) - \mathrm{H}(Y|X)$, where $\mathrm{H}(X|Y) := \mathbb{E}_{y\sim Y}\mathrm{H}(X|Y = y) \le \mathrm{H}(X)$. $I(X;Y|Z)$ represents the mutual information between X and Y conditioned on the random variable Z defined by $I(X;Y|Z) := \mathrm{H}(X|Z) - \mathrm{H}(X|Y,Z)$. Next, we describe some of the important properties of mutual information.

Fact 16 (Chain Rule) *Given random variables X, Y, Z, it holds:*

$$I(XY;Z) = I(X;Z) + I(Y;Z|X).$$

Fact 17 *If $X_1, \cdots, X_n$ are mutually independent random variables, then*

$$I(Y;X_1,\cdots,X_n) \ge \sum_{i=1}^{n} I(Y;X_i).$$

Squared Hellinger Distance. Let p and q be probability distributions over the same space Ω. The *squared Helliger distance* between p and q is defined as $H^2(p,q) := 1 - \sum_{\omega\in\Omega}\sqrt{p_\omega q_\omega}$. The following two basic properties about Hellinger distance appears, and their proof can be found in Cover (1999).

- *Mutual information to Hellinger distance:* Suppose Π, U are possibly dependent variables and U is a random uniform bit. Then
$$I(\Pi;U) \ge \frac{1}{2}H^2(\Pi \mid U = 1, \Pi \mid U = 0). \tag{1}$$
- *Squared Hellinger distance is lower bounded by squared total variation distance:* $H^2(p,q) \ge D_{TV}^2(p,q)$. Here, $D_{TV}(p,q) = \frac{1}{2}\sum_{\omega\in\Omega}|p_\omega - q_\omega|$ denotes the total variation distance.
- *Joint convexity:* Let p, p', q, q' be four distributions on the same support. Then for any $0 \le \lambda \le 1$,
$$\lambda \cdot H^2(p,q) + (1-\lambda)\cdot H^2(p',q') \ge H^2(\lambda p + (1-\lambda)p', \lambda q + (1-\lambda)q').$$

Appendix B. Information Complexity Lower Bound for $\textsf{Bit-Bias}(p,q)$

First, we show an information complexity lower bound for a special case, the $\textsf{Bia-Bias}(0,q)$ problem. To be more specific, we prove the following lemma:

Lemma 18 *If a multi-pass streaming algorithm $\mathcal{A}$ solves the $\textsf{Bit-Bias}(0,q)$ with advantage δ, then we have*

$$I(M;X) \ge \frac{1}{2}\min(q, 1-q)\cdot\delta^2$$

where $X := (X_1,\cdots,X_T) \sim \mathcal{D}_q$ and $M := (M_{i,r})_{i,r}$.

As a direct corollary, we have the lower bound for $\textsf{Bit-Bias}(0, 1/2)$:

Corollary 19 *If a multi-pass streaming algorithm $\mathcal{A}$ solves the $\textsf{Bit-Bias}(0, 1/2)$ problem (Task A) with advantage δ, then we have*

$$I(M; X) \geq \frac{\delta^2}{4}.$$

where $X := (X_1, \cdots, X_T) \sim \mathcal{D}_{1/2}$ and $M := (M_{i,r})_{i,r}$. $\mathcal{D}_{1/2}$ denotes the uniform distribution on $\{0, 1\}^T$.

Remark 20 *By Claim 11, we obtain a lower bound for the multi-pass information complexity of $\textsf{Bit-Bias}(0, 1/2)$: $\mathrm{IC}(\mathcal{A}, X) \geq I(M; X) \geq \delta^2/4$ when $X \sim \mathcal{D}_{1/2}$.*

B.1. Proofs for $\textsf{Bit-Bias}(0, q)$

The basic idea is to see $\textsf{Bit-Bias}(0, q)$ as a T-player communication problem and calculate the information cost using properties of communication protocols. We start with some definitions and notations.

Bit-Bias$(0, q)$ as a multi-party communication problem. There are T players, each receives a single bit $X_i \in \{0, 1\}$ as its input, and their goal is to distinguish whether $X \sim \mathcal{D}_q$ or $X = 0^T$. Here, we consider the communication protocol Π with private randomness in the blackboard model[4]. Given an input $X \in \{0, 1\}^T$ and private coins R, let $\Pi(X)$ denote the transcript induced by running Π on input X. Notice that $\Pi(X)$ might be a random variable here, and the randomness comes from the private randomness of the protocol Π and the randomness of the input X. The advantage of a protocol Π on solving $\textsf{Bit-Bias}(0, q)$ is defined by

$$\Pr_{X=0^T}[\Pi(X) \text{ outputs } 1] - \Pr_{X\sim\mathcal{D}_q}[\Pi(X) \text{ outputs } 1].$$

Information Cost. Suppose Π is a randomized protocol with private randomness for the $\textsf{Bit-Bias}(0, q)$ problem. The *information cost* of Π under $\mathcal{D}_q$ is defined as $I(\Pi(X); X)$, where $X \sim \mathcal{D}_q$.

The following theorem provides a lower bound for the information cost of the $\textsf{Bit-Bias}(0, q)$ problem in the communication setting.

Theorem 15 *Suppose Π is a protocol solving the $\textsf{Bit-Bias}(0, q)$ problem with advantage δ. Then its information cost under $\mathcal{D}_q$ is at least $\frac{1}{2}\min(q, 1-q) \cdot \delta^2$, i.e.,*

$$I(\Pi(X); X) \geq \frac{1}{2}\min(q, 1-q) \cdot \delta^2.$$

Here, $\Pi(X)$ denotes the transcript induced by running Π on input X.

As we mentioned earlier. Our proof adopts further extensions and modifications of Jayram (2009)'s techniques, and also some important observations from information theory. The main tool of our proof is the generalized cut-and-paste property from the rectangle property of communication protocols. We show that it holds even under an "expectation" operator (see Lemma 21).

[4] In each round, one of the players writes a message on the blackboard seen by all the players.

B.1.1. GENERALIZED CUT-AND-PASTE PROPERTY

We start with some notations and definitions.

Notations. We borrow the notion of $\pi(\cdot)$ defined by Jayram (2009). For a fixed communication protocol Π and an input vector $x \in \{0,1\}^T$, let $\pi(x)$ denote the probability distribution over the transcripts induced by running Π on input x, i.e., $\pi(x)$ is the distribution of $\Pi(x)$. Let $\pi(x)_\tau$ denote the probability that the transcript equals τ. If we see $\pi(x)$ as a vector, we have $\|\pi(x)\|_1 = \sum_\tau \pi(x)_\tau = 1$.

For convenience, we extend the definition of $\pi(\cdot)$ to a random transcript over random inputs. Here we use $\boldsymbol{u}$ to denote a vector in $[0,1]^T$. For $\boldsymbol{u} \in [0,1]^T$, define $\pi(\boldsymbol{u})_\tau = \mathbb{E}_X[\pi(X)_\tau]$ where X is a random variable over $\{0,1\}^T$ sampled as follows: for every $i \in [n]$, sample $X_i \in \{0,1\}$ independently with $\mathbb{E}[X_i] = \boldsymbol{u}_i$. In particular, $\pi(1/2, 1/2, \cdots, 1/2)$ is the uniform mixture of all $x \in \{0,1\}, \pi(x)$. It is easy to check that $\|\pi(\boldsymbol{u})\|_1 = 1$, so $\pi(\boldsymbol{u})$ is also a probability distribution.

The following is a generalization of the cut-and-paste lemma used in Jayram (2009), and is also mentioned in Braverman et al. (2016). For completeness, we include its proof here.

Lemma 21 (Cut-and-Paste) *Let u' and v' denote the strings obtained by performing some cut-and-paste on $u, v \in [0,1]^T$. In other words, for each $1 \le i \le T$, either (a) $u'_i = u_i$ and $v'_i = v_i$ or (b) $u'_i = v_i$ and $v'_i = u_i$. Then*

$$H^2(\pi(u), \pi(v)) = H^2(\pi(u'), \pi(v')),$$

where H^2 stands for squared Hellinger distance.

Proof The proof is mainly out of the rectangle property of communication protocols. First, recall the definitions of Hellinger distance: $H^2(p,q) = \frac{1}{2}\sum_i(\sqrt{p_i} - \sqrt{q_i})^2 = 1 - \sum_i \sqrt{p_i q_i}$. Thus, it suffices to show that for each i, it holds that

$$\sqrt{\Pr_{\boldsymbol{i}\sim\pi(u)}[\boldsymbol{i} = i] \Pr_{\boldsymbol{i}\sim\pi(v)}[\boldsymbol{i} = i]} = \sqrt{\Pr_{\boldsymbol{i}\sim\pi(u')}[\boldsymbol{i} = i] \Pr_{\boldsymbol{i}\sim\pi(v')}[\boldsymbol{i} = i]}.$$

From the property of communication protocols, the probability $\Pr_{\boldsymbol{i}\sim\pi(u)}[\boldsymbol{i} = i]$ can be decomposed by:

$$\Pr_{\boldsymbol{i}\sim\pi(u)}[\boldsymbol{i} = i] = \prod_{j=1}^{T} \alpha_j(u_j, i),$$

where u_j denotes the j-th coordinate of the vector u and α_j is a function only depending on u_j and i. Similarly we could decompose $\Pr_{\boldsymbol{i}\sim\pi(v)}[\boldsymbol{i} = i], \Pr_{\boldsymbol{i}\sim\pi(u')}[\boldsymbol{i} = i], \Pr_{\boldsymbol{i}\sim\pi(v')}[\boldsymbol{i} = i]$. Thus,

$$\begin{aligned}\Pr_{\boldsymbol{i}\sim\pi(u)}[\boldsymbol{i} = i] \cdot \Pr_{\boldsymbol{i}\sim\pi(v)}[\boldsymbol{i} = i] &= \prod_{j}^{T} \alpha_j(u_j, i) \cdot \alpha_j(v_j, i) \\ &= \prod_{j}^{T} \alpha_j(u'_j, i) \cdot \alpha_j(v'_j, i) \\ &= \Pr_{\boldsymbol{i}\sim\pi(u')}[\boldsymbol{i} = i] \cdot \Pr_{\boldsymbol{i}\sim\pi(v')}[\boldsymbol{i} = i].\end{aligned}$$

The equality holds since either (a) $u_i' = u_i$ and $v_i' = v_i$ or (b) $u_i' = v_i$ and $v_i' = u_i$ holds. ■

In this paper, we will use the generalized cut-and-paste lemma for $u_i, v_i \in \{0, q\}$. For simplicity of notations, for $I \subseteq [T]$, we use $\pi(I)$ to denote $\pi(u)$ where $u_i = q$ if $i \in I$, otherwise $u_i = 0$. In particular, $\pi([T]) = \pi(q, \cdots, q)$ and $\pi(\emptyset) = \pi(0^N)$. By the new cut-and-paste lemma, we propose a variant of Theorem 7 in (Jayram, 2009) as follows:

Lemma 22 *Suppose $I_1, I_2, \cdots, I_s$ are a pairwise disjoint collection of s subsets of $[T]$, Let $I := I_1 \cup \cdots \cup I_s$. Then*

$$\sum_{i=1}^{s} H^2(\pi(I), \pi(I \setminus I_i)) \geq \frac{1}{4} H^2(\pi(I), \pi(\emptyset)).$$

Based on Lemma 21, the lemma above can be proved by a similar argument of Theorem 7 in (Jayram, 2009).

B.1.2. PROOF OF THEOREM 15

First, we recall the statement of Theorem 15.

Theorem 15 *Suppose Π is a protocol solving the* Bit-Bias$(0, q)$ *problem with advantage δ. Then its information cost under $\mathcal{D}_q$ is at least $\frac{1}{2}\min(q, 1-q) \cdot \delta^2$, i.e.,*

$$I(\Pi(X); X) \geq \frac{1}{2}\min(q, 1-q) \cdot \delta^2.$$

Here, $\Pi(X)$ denotes the transcript induced by running Π on input X.

Proof Suppose $X \sim \mathcal{D}_q$. We have the following observation:

$$I\left(\Pi(X); X\right) \geq \sum_{i=1}^{T} I(\Pi(X); X_i) \geq \min(q, 1-q) \sum_{i=1}^{T} H^2\left(\pi\left(\boldsymbol{q} + \frac{e_i}{2}\right); \pi\left(\boldsymbol{q} - \frac{e_i}{2}\right)\right).$$

Here, the first inequality comes from Fact 17. Moreover, we know that

$$I(\Pi(X); X_i) = H(\Pi(X)) - q \cdot H(\Pi(X)|X_i = 1) - (1-q) \cdot H(\Pi(X)|X_i = 0).$$

Define two distributions here:

- π_0: the distribution of $\Pi(X) \mid X_i = 0$ when $X \sim \mathcal{D}_q$;
- π_1: the distribution of $\Pi(X) \mid X_i = 1$ when $X \sim \mathcal{D}_q$.

Then, we define the following function

$$g(q') := H(q'\pi_1 + (1-q')\pi_0) - q' \cdot H(\pi_1) - (1-q') \cdot H(\pi_0).$$

We further know that $g(q')$ is concave when $q' \in [0, 1]$ by calculations. Together with the fact that $g(0) = g(1) = 0$, we have

$$g(q) \geq 2\min(q, 1-q) \cdot g(1/2).$$

Then, by Inequality (1), we know that $g(1/2) \geq 1/2 \cdot H^2\left(\pi\left(\boldsymbol{q}+\frac{e_i}{2}\right);\pi\left(\boldsymbol{q}-\frac{e_i}{2}\right)\right)$, as well as

$$I\left(\Pi(X);X\right) \geq \min(q,1-q)\sum_{i=1}^{T} H^2\left(\pi\left(\boldsymbol{q}+\frac{e_i}{2}\right);\pi\left(\boldsymbol{q}-\frac{e_i}{2}\right)\right).$$

Let e_i denote the vector with a 1 in the i-th position and 0 elsewhere. Let $\boldsymbol{q} = (q,\cdots,q)$ be the all-q vector with N coordinates. We have

$$\begin{aligned}
I\left(\Pi(X);X\right) \geq& \min(q,1-q)\sum_{i=1}^{T} H^2\left(\pi\left(\boldsymbol{q}+\frac{e_i}{2}\right);\pi\left(\boldsymbol{q}-\frac{e_i}{2}\right)\right)\\
=&2\min(q,1-q)\sum_{i=1}^{T}\left[\frac{1}{2}H^2\left(\pi\left(\boldsymbol{q}+\frac{e_i}{2}\right);\pi\left(\boldsymbol{q}-\frac{e_i}{2}\right)\right)+\frac{1}{2}H^2\left(\pi\left(\boldsymbol{q}-\frac{e_i}{2}\right);\pi\left(\boldsymbol{q}-\frac{e_i}{2}\right)\right)\right]\\
\geq&2\min(q,1-q)\sum_{i=1}^{T} H^2\left(\pi\left(\boldsymbol{q}\right);\pi\left(\boldsymbol{q}-\frac{e_i}{2}\right)\right)\\
=&2\min(q,1-q)\sum_{i=1}^{T} H^2\big(\pi([T]);\pi([T]\setminus\{i\})\big) \geq \frac{1}{2}H^2\big(\pi([T]);\pi(\emptyset)\big) \geq \frac{1}{2}D_{TV}^2\big(\pi([T]);\pi(\emptyset)\big)\\
\geq&\frac{1}{2}\min(q,1-q)\cdot\delta^2.
\end{aligned}$$

Here, the second inequality is by the joint convexity of squared Hellinger distance, and the third inequality is by Lemma 22. ■

B.1.3. PROOF OF LEMMA 18

Now, we are ready to prove Lemma 18 by showing the relation between information complexity for communication protocols and streaming algorithms.

Lemma 23 *If a multi-pass streaming algorithm $\mathcal{A}$ solves the* Bit-Bias$(0,q)$ *with advantage δ, then we have*

$$I(M;X) \geq \frac{1}{2}\min(q,1-q)\cdot\delta^2$$

where $X := (X_1,\cdots,X_T) \sim \mathcal{D}_q$ and $M := (M_{i,r})_{i,r}$.

Proof This proof is a standard reduction from streaming algorithms to communication protocols. Suppose $\mathcal{A}$ is an L-pass streaming algorithm (with semi-private randomness) to solve Bit-Bias$(0,q)$ with advantage δ. Then, $\mathcal{A}$ can be used to construct a communication protocol to solve the communication version of Bit-Bias$(0,q)$ as follows: each player i simulates $\mathcal{A}$ by running $\mathcal{A}$ on X_i, and send the internal memory of $\mathcal{A}$ to the next player $i+1$. At the end of each pass, the last player sends the internal memory back to the first player. Finally, the last player in the last pass outputs the answer $\mathcal{A}(X)$. It is easy to see that this communication protocol has the same advantage δ and its transcript $\Pi(X)$ shares the same distribution with $M = (M_{i,r})_{i,r}$. By Theorem 15, we have

$$I(M;X) = I(\Pi(X);X) \geq \frac{1}{2}\min(q,1-q)\cdot\delta^2$$

as desired. ■

B.2. Proofs for $\textsf{Bit-Bias}(p, q)$

Based on the result for the $\textsf{Bit-Bias}(0, q)$ problem, we prove the following information complexity lower bound for the $\textsf{Bit-Bias}(p, q)$ problem.

Theorem 9 *If a multi-pass streaming algorithm $\mathcal{A}$ solves the $\textsf{Bit-Bias}(p, q)$ with advantage δ, we have*

$$\mathrm{IC}(\mathcal{A}, X) \geq \Omega\left(\min(q, 1-q) \cdot \frac{q^2\delta^2}{(q-p)^2}\right),$$

where $X \sim \mathcal{D}_q$. In addition, flipping the input, we get a similar lower bound

$$\mathrm{IC}(\mathcal{A}, X) \geq \Omega\left(\min(p, 1-p) \cdot \frac{(1-p)^2\delta^2}{(q-p)^2}\right),$$

where $X \sim \mathcal{D}_p$.

Proof The proof is inspired by the technique of the needle problem (Braverman et al., 2024). Roughly speaking, our proof is by a decomposition of the $\textsf{Bit-Bias}(p, q)$ problem and a reduction to the $\textsf{Bit-Bias}(0, q)$ problem as mentioned in Section 3. We define the following alternative sampling process for $\mathcal{D}_p$:

1. sample a set $S \subseteq [N]$ with each element $j \in [N]$ contained in S independently with probability $\frac{q-p}{q}$;
2. for each $j \notin S$, $X_j \sim \textsf{Bernoulli}(q)$;
3. for each $j \in S$, $X_j \sim \textsf{Bernoulli}(0)$ (which means $X_j = 0$ with probability 1).

It is easy to verify that the sampling process is identical to $\mathcal{D}_p$ since for each $j \in [N]$,

$$\Pr[X_j = 1] = (1 - \frac{q-p}{q}) \cdot q = p.$$

Here, we use $\mathcal{D}_0^S$ to denote the distribution of $\mathcal{D}_p$ conditioned on the set S is sampled in the first step. Intuitively, the differences between $\mathcal{D}_0^S$ and $\mathcal{D}_q$ is that: $\mathcal{D}_0^S$ locally (on the coordinates in S) equals zero. Then, we know $\mathcal{D}_p$ can be represented as the linear of combination of $\mathcal{D}_0^S$s by the sampling process: $\mathcal{D}_p = \sum_S \Pr[S]\mathcal{D}_0^S$.

We know that

$$\delta = \Pr[\mathcal{A}(\mathcal{D}_p) = 1] - \Pr[\mathcal{A}(\mathcal{D}_q) = 1] = 1 - \Pr[\mathcal{A}(\mathcal{D}_p) = 0] - \Pr[\mathcal{A}(\mathcal{D}_q) = 1].$$

We define $\mathrm{Err}_{\mathcal{A}}(P, Q)$ as $\Pr[\mathcal{A}(P) = 0] + \Pr[\mathcal{A}(Q) = 1]$ for any distribution P, Q.

By Markov inequality, we have

$$\Pr_S[\mathrm{Err}_{\mathcal{A}}(\mathcal{D}_0^S, \mathcal{D}_q) \geq 1 - \frac{\delta}{2}] \leq \frac{1-\delta}{1-\frac{\delta}{2}} = \frac{2-2\delta}{2-\delta}.$$

We say a set S is *good* if and only if $\mathrm{Err}_{\mathcal{A}}(\mathcal{D}_0^S, \mathcal{D}_q) \leq 1-\frac{\delta}{2}$. Then, $\Pr[S \text{ is good}] \geq \frac{\delta}{2}$ by definitions.

For those good S, we show that they contribute a lot to the information complexity by a reduction to $\textsf{Bit-Bias}(0, q)$, which is formalized by the following lemma:

Lemma 24 *If an L-pass streaming algorithm $\mathcal{A}$ could distinguish between $\mathcal{D}_p$ and $\mathcal{D}_0^S$, where $S = (p_1, p_2, \cdots, p_m)$, with $Err_{\mathcal{A}}(\mathcal{D}_p, \mathcal{D}_0^S) \le 1 - \delta$, it holds that*

$$\sum_{r=1}^{L-1} \sum_{i=1}^{m} \sum_{k=1}^{i} I(M_{p_{i+1}-1,r}; X_{p_k} \mid M_{p_k-1,\le r}, M_{p_{i+1}-1,\le r-1}) \ge \frac{1}{2}\min(q, 1-q) \cdot \delta^2.$$

The proof of this lemma appears at the end of this section.

For simplicity of notations, we let

$$\mathrm{IC}_S = \sum_{r=1}^{L-1} \sum_{i=1}^{m} \sum_{k=1}^{i} I(M_{p_{i+1}-1,r}; X_{p_k} \mid M_{p_k-1,\le r}, M_{p_{i+1}-1,\le r-1}).$$

Then, we have

$$\mathbb{E}_S[\mathrm{IC}_S] = \sum_S \Pr[S] \cdot \mathrm{IC}_S \ge \sum_{\text{good } S} \Pr[S] \cdot \mathrm{IC}_S.$$

By the lemma above, we have $\mathrm{IC}_S \ge \min(q, 1-q) \cdot \delta^2/2$ for every good S. Hence, we have the following lower bound for $\mathbb{E}_S[\mathrm{IC}_S]$:

$$\mathbb{E}_S[\mathrm{IC}_S] \ge \Pr[S \text{ is good}] \cdot \min(q, 1-q) \cdot \delta^2/8 \ge \min(q, 1-q) \cdot \delta^3/4.$$

Next, we show the relations between $\mathbb{E}_S[\mathrm{IC}_S]$ and $\mathrm{IC}(\mathcal{A}, X)$. Recall the definition,

$$\begin{aligned}\mathrm{IC}(\mathcal{A}, X) :=& \sum_{r=1}^{L} \sum_{i=1}^{N} \sum_{k=1}^{i} I(M_{i,r}; X_k \mid M_{k-1,\le r}, M_{i,\le r-1}) \\ &+ \sum_{r=1}^{L} \sum_{i=1}^{N} \sum_{k=i+1}^{N} I(M_{i,r}; X_k \mid M_{k-1,\le r-1}, M_{i,\le r-1}),\end{aligned}$$

For random S, each term $I(M_{i,r}; X_k \mid M_{k-1,\le r}, M_{i,\le r-1})$ with $i \ge k$ appears in IC_S with probability exactly $\frac{(q-p)^2}{q^2}$ since it appears if only if both $i+1$ and k are contained by S. Similarly, each term $I(M_{i,r}; X_k \mid M_{k-1,\le r-1}, M_{i,\le r-1})$ with $i < k$ appears in IC_S with probability at most $\frac{(q-p)^2}{q^2}$ since this happens if and only if both $i+1$ and k are contained by S and $i+1$ is the smallest element in S. Finally, we have

$$\mathrm{IC}(\mathcal{A}, X) \ge \mathbb{E}_S[\mathrm{IC}_S] \cdot \frac{q^2}{(q-p)^2} \ge \Omega\left(\min(q, 1-q) \cdot \frac{q^2\delta^2}{(q-p)^2}\right),$$

as desired. ■

Proof [Proof of Lemma 24] *We use a reduction to* ***Bit-Bias***$(0, q)$ *problem in the communication setting.* To be specific, given an algorithm $\mathcal{A}$ with $\mathrm{Err}_{\mathcal{A}}(\mathcal{D}_0^S, \mathcal{D}_q) \le \delta$, where $S = (p_1, \cdots, p_m)$, we construct a communication protocol Π with $\mathrm{Err}_{\Pi}(\mathcal{D}_0, \mathcal{D}_q) \le \delta$ presented as Algorithm B.2.

Without of loss of generality, we add an extra constraint: the streaming algorithm does not change its memory state in the last pass, which does not affect our lower bounds (discussed in Section C.3). Then, we have the following properties from the simulation process:

Algorithm 1 Communication protocol Π for $\textsf{Bit-Bias}(0, q)$

```
Input: a length $m$ data stream $X = (X_1, \cdots, X_m)$;
Output: an answer from $\{0, 1\}$;
Recall: $S = \{p_1, \ldots, p_m\}$  for player $j$ from $1$ to $m$ do
    let $X'_{p_j} = X_j$  independently sample $X'_{p_j+1}, \cdots, X'_{p_{j+1}-1}$ from $\textsf{Bernoulli}(q)$
end
Player $m$ simulates $M_{p_1-1,1} = \mathcal{A}(X'_1, \ldots, X'_{p_1-1})$ and sends $M_{p_1-1,1}$ to Player $1$  for $i$ from $1$ to $L-1$ do
    for $j$ from $1$ to $m$ do
        Player $j$ simulates $M_{p_{j+1}-1,i} = \mathcal{A}(M_{p_j-1,i}, X'_{p_j}, \ldots, X'_{p_{j+1}-1})$  Player $j$ sends $M_{p_{j+1}-1,i}$
          to Player $j+1$ (send to Player $1$ when $j = m$)
    end
end
return the output of Player $m$;
```

- When $X \sim \mathcal{D}_q$, it holds $X' \sim \mathcal{D}_q$[5]; when $X \sim \mathcal{D}_0$, it holds $X' \sim \mathcal{D}_0^S$.
- $\mathrm{Err}_\Pi(\mathcal{D}_0, \mathcal{D}_q) = \mathrm{Err}_\mathcal{A}(\mathcal{D}_0^S, \mathcal{D}_q)$ due to the fact above and the constraints.
- When $X \sim \mathcal{D}_q$ and $X' \sim \mathcal{D}_q$, the information complexity for the streaming algorithm $\mathcal{A}$ and the communication protocol Π are closely related:

$$I\left((M_{p_1-1,1}, M_{p_2-1,1}, \cdots, M_{p_1-1,L}); (X'_{p_1}, \cdots, X'_{p_m})\right) = I\left(\Pi(X); X\right).$$

By Theorem 15, we have the following lower bound:

$$I\left((M_{p_1-1,1}, M_{p_2-1,1}, \cdots, M_{p_1-1,L}); (X'_{p_1}, \cdots, X'_{p_m})\right) = I\left(\Pi(X); X\right) \geq \frac{1}{2}\min(q, 1-q) \cdot \delta^2.$$

Finally, Claim 5.4 in (Braverman et al., 2024) shows that

$$\mathrm{IC}_S \geq I\left((M_{p_1-1,1}, M_{p_2-1,1}, \cdots, M_{p_1-1,L}); (X'_{p_1}, \cdots, X'_{p_m})\right),$$

and as a consequence

$$\mathrm{IC}_S \geq \frac{1}{2}\min(q, 1-q) \cdot \delta^2.$$

■

Appendix C. Reductions from $\textsf{Bit-Bias}(0, 1/2)$ to the Core Problem

C.1. A lower bound for single-subpopulation tasks

In this section, we turn the lower bound for Task A (Corollary 19) into lower bounds for Task B (Theorem 27) and then Task B' (Theorem 28).

[5] Note that the length of the two data streams are different. For simplicity, we use $\mathcal{D}_q$ to denote two distributions respectively.

C.1.1. A LOWER BOUND FOR TASK B

We use the same reduction as in the paper by Brown et al. (2022). To be specific, suppose $\mathcal{A}$ is a randomized streaming algorithm solving Task B with advantage δ. Define

$$\pi_r := \Pr_{\substack{\mathcal{P}\sim\mathcal{Q}_{d,\rho} \\ X=(X_1,\cdots,X_T)\sim_{iid}\mathcal{P}}} \left[\mathcal{A}(X)=1 \mid \text{the size of } \mathcal{J} \text{ is } r\right].$$

We then construct a randomized multi-pass streaming algorithm $\mathcal{A}'$ for Task A as follows (see Appendix C in (Brown et al., 2022) for a formal description):

1. Before the stream coming, $\mathcal{A}'$ draws a random $r \in \{0,\cdots,\rho\}$, a random $\mathcal{J} \subseteq [d]$ of size r, and a random $\mathcal{B} \in \{0,1\}^r$; Furthermore, $\mathcal{A}'$ draw a random $j_0 \in [d]$; $\mathcal{A}'$ stores $(\mathcal{J},\mathcal{B},j_0)$;
2. At each pass, when receiving $X_i \in \{0,1\}$, $\mathcal{A}'$ simulates $\mathcal{A}$ on $Z_i \in \{0,1\}^d$ which is generated as follows:
   - draw $Z_i \in \{0,1\}^d$ uniformly;[6]
   - then overwrite the feature j_0 of Z_i to be X_i;
   - then overwrite the features $j_t \in \mathcal{J}$ to be b_t.
3. Finally, obtaining $\mathcal{A}$'s output $\mathsf{ans} \in \{0,1\}$, $\mathcal{A}'$ output ans if $\pi_{r+1} \geq \pi_r$ and $\neg\mathsf{ans}$ otherwise.

Then, we have the following observations for the construction.

Claim 25 (Brown et al. (2022)) *$\mathcal{A}'$ solving Task A with advantage at least $\frac{\delta}{\rho+1} + \frac{\rho}{d}$.*

Claim 26 *Let $X = (X_1,\cdots,X_T) \sim \mathcal{U}$ and Z_i be defined as above. Let $M = (M_{i,r})_{i,r}$ and $M' = (M'_{i,r})_{i,r}$ denote the memory states of $\mathcal{A}$ and $\mathcal{A}'$ respectively. Then*

$$I(M;Z \mid \mathcal{J},\mathcal{B}) \geq d \cdot I(M';X).$$

Proof Noticing that the memory state of $\mathcal{A}'$ is $M'_{i,r} = (M_{i,r}, j_0, \mathcal{J}, \mathcal{B})$, we have

$$\begin{aligned} I(M';X) =& I(M,j_0,\mathcal{J},\mathcal{B};X) = I(j_0,\mathcal{J},\mathcal{B};X) + I(M;X \mid j_0,\mathcal{J},\mathcal{B}) = I(M;X \mid j_0,\mathcal{J},\mathcal{B}) \\ =& \frac{1}{d}\sum_{j=1}^{d} I(M;X \mid \mathcal{J},\mathcal{B},j_0=j). \end{aligned}$$

Let $Z^j := (Z_1^j,\cdots,Z_T^j)$ and Z_t^j is the j-th bit of Z_t. Next, note that conditioning on $j_0 = j$ and any value of $(\mathcal{J},\mathcal{B})$, we have the Markov chain $M \to Z^j \to X$. So by the data processing inequality, we have

$$I(M';X) \leq \frac{1}{d}\sum_{j=1}^{d} I(M;Z^j \mid \mathcal{J},\mathcal{B},j_0=j).$$

[6] Here, $\mathcal{A}'$ uses the same Z_i in all passes rather than draws an independent Z_i in each pass.

Then, recall that $X = (X_1, \cdots, X_T) \sim \mathcal{U}$. So we can remove the condition that $j_0 = j$, since the distribution of $(M, Z^j, \mathcal{J}, \mathcal{B})$ remains the same whatever value j_0 takes. Thus

$$I(M'; X) \leq \frac{1}{d} \sum_{j=1}^{d} I(M; Z^j \mid \mathcal{J}, \mathcal{B}) \leq \frac{1}{d} I(M; Z \mid \mathcal{J}, \mathcal{B}),$$

where the last inequality is by Fact 17. ■

By Corollary 19, and Claims 25 and 26, we directly get the following lower bound for Task B.

Theorem 27 *For any semi-private random multi-pass algorithm $\mathcal{A}$ solving Task B with advantage δ, we have*

$$I(M; X \mid \mathcal{J}, \mathcal{B}) \geq \frac{d}{2} \left(\frac{\delta}{\rho + 1} - \frac{\rho}{d} \right)^2,$$

where $X = (X_1, \cdots, X_T) \sim_{iid} \mathcal{P}_{(\mathcal{J},\mathcal{B})}$.

C.1.2. A LOWER BOUND FOR TASK B'

The lower bound for Task B can be further turned into a lower bound for Task B'. Recall that Task B' asks the algorithm to distinguish whether a test sample is structured or uniform, while Task B whether the input is structured or uniform. The proof of Theorem 28 is almost the same to that in (Brown et al., 2022).

Theorem 28 *For any semi-private random multi-pass algorithm $\mathcal{A}$ solving Task B' with advantage δ, we have*

$$I(M; X \mid \mathcal{J}, \mathcal{B}) \geq \frac{d}{2} \left(\frac{\delta}{2(\rho + 1)} - \frac{\rho}{d} \right)^2 - 1$$

where $X = (X_1, \cdots, X_T) \sim_{iid} \mathcal{P}_{(\mathcal{J},\mathcal{B})}$.

Theorem 28 directly follows from Theorem 27 and the following lemma.

Lemma 29 *For any semi-private random multi-pass algorithm $\mathcal{A}$ solving Task B' on T samples $X = (X_1, \cdots, X_T)$ with advantage at least δ, there is an semi-private random multi-pass algorithm $\mathcal{A}'$ solving Task B on $T + 1$ samples (X, X_{T+1}) with advantage at least $\delta/2$ and satisfying*

$$I(M'; X, X_{T+1} \mid \mathcal{J}, \mathcal{B}) \leq I(M; X \mid \mathcal{J}, \mathcal{B}) + 1.$$

Here $X_1, \cdots, X_{T+1} \sim_{iid} \mathcal{P}_{(\mathcal{J},\mathcal{B})}$.

Proof The proof is almost the same as that of Lemma 31 in (Brown et al., 2022). Let X_i denote a sample drawn from $\mathcal{P}_{(\mathcal{J},\mathcal{B})}$ and U_i a uniform sample. Define

$p_{x,x} := \Pr[\mathcal{A}(X, X_{T+1}) = 1]$, $p_{x,u} := \Pr[\mathcal{A}(X, U_{T+1}) = 1]$, and $p_{u,u} := \Pr[\mathcal{A}(U, U_{T+1}) = 1]$.

Here $U := (U_1, \cdots, U_T)$. Since $\mathcal{A}$ solves Task B' with advantage at least δ, we have $p_{x,x} - p_{x,u} \geq \delta$. So at least one of the following two cases happens:

Case 1: If $p_{x,x} - p_{u,u} \geq \delta/2$ happens, we construct $\mathcal{A}'$ by simply simulating $\mathcal{A}$ and outputs $\mathcal{A}'s$ output. Obviously, $\mathcal{A}'$ has advantage $p_{x,x} - p_{u,u} \geq \delta/2$ on solving Task B.

Case 2: If $p_{u,u} - p_{x,u} \geq \delta/2$ happens, we construct $\mathcal{A}'$ running on $T + 1$ samples as follows:

1. $\mathcal{A}'$ first simulates $\mathcal{A}$ with ignoring the $(T+1)$-th sample (i.e., when $\mathcal{A}'$ receives the $(T+1)$-th sample in any pass, $\mathcal{A}'$ does nothing and skip it);
2. finally $\mathcal{A}'$ generates a fresh uniform sample U_{T+1} and takes the testing result of $\mathcal{A}$ on U_{T+1} as the output.

In this case, $\mathcal{A}'$ has advantage $p_{u,u} - p_{x,u} \geq \delta/2$ on solving Task B.

In both cases, $\mathcal{A}'$ just runs $\mathcal{A}$ until the finial step, so we have $M'_{i,r} = M_{i,r}$ for any $i \in [T]$ and $r \in [L]$. Besides, $M'_{T+1,r} = M_{T,r}$ if $r \in [L-1]$, since $\mathcal{A}'$ skips the $(T+1)$-th input sample except the last pass. Furthermore, without loss of generality, we can assume the final state of $\mathcal{A}'$ is a single bit (i.e., its answer) and thus the entropy of $M'_{T+1,L}$ is at most 1. Finally, we have

$$\begin{aligned}
&I(M'; X, X_{T+1} \mid \mathcal{J}, \mathcal{B}) \\
=&I(M'_{\leq T, \leq L}; X, X_{T+1} \mid \mathcal{J}, \mathcal{B}) + I(M'_{T+1, \leq L-1}; X, X_{T+1} \mid M'_{\leq T, \leq L}, \mathcal{J}, \mathcal{B}) \\
&+ I(M'_{T+1,L}; X, X_{T+1} \mid M'_{\leq T, \leq L}, M'_{T+1, \leq L-1}, \mathcal{J}, \mathcal{B}) \\
=&I(M; X, X_{T+1} \mid \mathcal{J}, \mathcal{B}) + 0 + I(M'_{T+1,L}; X, X_{T+1} \mid M'_{\leq T, \leq L}, M'_{T+1, \leq L-1}, \mathcal{J}, \mathcal{B}) \\
\leq&I(M; X, X_{T+1} \mid \mathcal{J}, \mathcal{B}) + 1 = I(M; X \mid \mathcal{J}, \mathcal{B}) + 1
\end{aligned}$$

as desired. ■

C.2. A lower bound for the Core problem

In this section, we obtain a lower bound for Task C (Theorem 30) from a reduction to Task B' (Theorem 28). To simplify our analysis, we first add the following restriction on the streaming algorithms.

- **Restriction**: $\mathcal{A}$ does not do any operations during the last pass.

We note that this restriction does not affect our main theorem since for any given algorithm solving Task C, we can construct another restricted algorithm that also solves Task C. Please see details in Section C.3.

Theorem 30 *For any randomized multi-pass algorithm $\mathcal{A}$ solving Task C (the Core problem) with advantage δ and satisfies the Restriction, we have*

$$\mathrm{IC}(\mathcal{A}, X \mid \mathcal{J}, \mathcal{B}) \geq \frac{k^2 d}{2} \left(\frac{\delta}{2(\rho+1)} - \frac{\rho}{d} \right)^2 - k^2.$$

Note that the definition of $\mathrm{IC}(\mathcal{A}, X \mid \mathcal{J}, \mathcal{B})$ is by adding the condition $\mathcal{J}, \mathcal{B}$ to each term appearing in $\mathrm{IC}(\mathcal{A}, X)$. Here, $X = (X_1, \cdots, X_N)$ are i.i.d. samples of $\mathcal{P}_{\text{mix}}$ with parameters $\mathcal{J} = (\mathcal{J}_1, \cdots, \mathcal{J}_k)$ and $\mathcal{B} = (\mathcal{B}_1, \cdots, \mathcal{B}_k)$.

The proof of Theorem 30 is inspired by that of Theorem 12 in (Brown et al., 2022), but our analysis makes the argument clearer by adopting a decomposition-and-reduction approach. The intuition is that a stream of Task C can be decomposed into k streams of Task B' each located in random positions, and any algorithm solving Task C has to solve most of the inserted Task B's simultaneously. To explain the decomposition, we consider an alternative sampling process for the stream $X = (X_1, \cdots, X_N) \sim_{iid} \mathcal{P}_{\text{mix}}$:

1. First, pick an uniformly random k-partition $(\mathcal{S}_1, \cdots, \mathcal{S}_k)$ of $[N]$. That is, we throw each $i \in [N]$ into an uniformly random $\mathcal{S}_j$, independent of other elements in $[N]$;
2. Then, for each $j \in [k]$ and each $i \in \mathcal{S}_j$, draw a sample Y_i from $\mathcal{P}_{(\mathcal{J}_j, \mathcal{B}_j)}$, and let $X_i = (j, Y_i)$. Intuitively, we insert a stream of i.i.d. samples from $\mathcal{P}_{(\mathcal{J}_j, \mathcal{B}_j)}$ in the positions $\mathcal{S}_j$.

Given any $j \in [k]$ and any value s_j of $\mathcal{S}_j$, we define

$$\delta(j, s_j) := \mathop{\mathbb{E}}_{\substack{\mathcal{P}_{\text{mix}} \sim \mathcal{Q}_{k,d,\rho} \\ X=(X_1,\cdots,X_N) \sim_{iid} \mathcal{P}_{\text{mix}}|s_j \\ m \leftarrow \mathcal{A}(X)}} \left[\Pr_{y \sim \mathcal{P}_{(\mathcal{J}_j, \mathcal{B}_j)}} [m(j, y) = 1] - \Pr_{y \sim \mathcal{U}} [m(j, y) = 1] \right]$$

to be the advantage of $\mathcal{A}$ conditioned on that the test sample is drawn from $\mathcal{P}_{(\mathcal{J}_j, \mathcal{B}_j)}$ and $\mathcal{S}_j$ takes value of s_j. It is easy to see that

$$\delta = \mathop{\mathbb{E}}_{j, s_j} [\delta(j, s_j)] \,.$$

By reduction to Task B', we have the following lemma. For completeness, its proof is included in Section C.2.1.

Lemma 31 *Given any $j \in [k]$ and any value $s_j = \{p_1, \cdots, p_T\} \subseteq [N]$ of $\mathcal{S}_j$, it holds that*

$$\widetilde{\text{IC}}_{j,s_j} := \sum_{r=1}^{L-1} \sum_{i=1}^{T} I(M_{p_{i+1}-1,r}; X_{p_i} \mid M_{p_i-1,\le r}, M_{p_{i+1}-1,\le r-1}, \mathcal{J}, \mathcal{B}) \ge \frac{d}{2} \left(\frac{\delta(j, s_j)}{2(\rho+1)} - \frac{\rho}{d} \right)^2 - 1. \tag{2}$$

Here we define $M_{p_{T+1},\le r-1}$ as $M_{p_1,\le r}$.

Now we are ready to prove Theorem 30:

Proof The idea is to take expectation over j and s_j of Inequality (2). On the right hand side, we have

$$\mathbb{E}_{j,s_j} \left[\frac{d}{2} \left(\frac{\delta(j, s_j)}{2(\rho+1)} - \frac{\rho}{d} \right)^2 - 1 \right] \ge \frac{d}{2} \left(\frac{\mathbb{E}[\delta(j, s_j)]}{2(\rho+1)} - \frac{\rho}{d} \right)^2 - 1 = \frac{d}{2} \left(\frac{\delta}{2(\rho+1)} - \frac{\rho}{d} \right)^2 - 1$$

by Jensen's inequality. For the left hand side, we claim that

$$\mathbb{E}_{j,s_j} \left[\widetilde{\text{IC}}_{j,s_j} \right] \le \frac{1}{k^2} \cdot \text{IC}(\mathcal{A}, X \mid \mathcal{J}, \mathcal{B}), \tag{3}$$

which together with Lemma 31 implies Theorem 30 immediately. To show Equality (3), we first recall the definition of $\text{IC}(\mathcal{A}, X \mid \mathcal{F}, \mathcal{B})$:

$$\begin{aligned} \text{IC}(\mathcal{A}, X \mid \mathcal{F}, \mathcal{B}) := & \sum_{r=1}^{L} \sum_{i=1}^{T} \sum_{k=1}^{i} I(M_{i,r}; X_k \mid M_{k-1,\le r}, M_{i,\le r-1}, \mathcal{F}, \mathcal{B}) \\ & + \sum_{r=1}^{L} \sum_{i=1}^{T} \sum_{k=i+1}^{T} I(M_{i,r}; X_k \mid M_{k-1,\le r-1}, M_{i,\le r-1}, \mathcal{F}, \mathcal{B}). \end{aligned}$$

For a random (j, s_j), each term $I(M_{i,r}; X_k | M_{k-1,\le r}, M_{i,\le r-1}, \mathcal{F}, \mathcal{B})$ with $i \ge k$ appears in $\widetilde{\text{IC}}_{j,s_j}$ with probability at most $1/k^2$ since this happens only if both $i+1$ and k are chosen by $\mathcal{S}_j$. Similarly, each term $I(M_{i,r}; X_k | M_{k-1,\le r-1}, M_{i,\le r-1})$ with $i < k$ appears in $\widetilde{\text{IC}}_{j,s_j}$ with probability at most $1/k^2$ since this happens only if both $i+1$ and k are chosen by $\mathcal{S}_j$ and $i+1$ is the smallest element in $\mathcal{S}_j$. Thus, we have Equality (3) as claimed. ■

Algorithm 2 $\mathcal{A}'$ for Task B'

```
   Input: a stream $Y = (Y_1, \cdots, Y_T)$;
   Output: a function $m' : \{0,1\}^d \to \{0,1\}$;
10 sample $\mathcal{P}'_{\text{mix}} \sim \mathcal{Q}_{k-1,d,\rho}$;
     draw $(X_1, \cdots, X_{p_1-1})$ from $\mathcal{P}'_{\text{mix}}$ independently;
     execute $\mathcal{A}$ on $(X_1, \cdots, X_{p_1-1})$;
     for $r$ from $1$ to $L-1$ do
11       for $i$ from $1$ to $T$ do
12           receive sample $Y_i \in \{0,1\}^d$;
               set $X_{p_i} \leftarrow (j, Y_i)$;
               if $i = n$ then
13                 draw $\hat{X}_i := (X_{p_n+1}, \cdots, X_N, X_1, \cdots, X_{p_1-1})$ from $\mathcal{P}'_{\text{mix}}$ independently;
14           end
15           else
16               draw $\hat{X}_i := (X_{p_i+1}, \cdots, X_{p_{i+1}-1})$ from $\mathcal{P}'_{\text{mix}}$ independently;
17           end
18           execute $\mathcal{A}$ on $(X_{p_i}, \hat{X}_i)$;
19       end
20 end
21 receive the output $m : [k] \times \{0,1\}^d \to \{0,1\}$ of $\mathcal{A}$;
     return the function $m'(\cdot) := m(j, \cdot)$.
```

C.2.1. PROOF OF LEMMA 31

The proof is by reduction to Task B'. Given $j \in [k]$, $s_j = \{p_1, \cdots, p_T\}$, and an algorithm $\mathcal{A}$ for Task C that satisfies the restriction, we construct a semi-private random $(L-1)$-pass streaming algorithm $\mathcal{A}'$ for Task B' (see Algorithm 1). Line 13 is doable since $\mathcal{A}$ does nothing in the last pass. Obviously, $\mathcal{A}'$ solves Task B' with advantage $\delta(j, s_j)$. Furthermore, to generate inputs for $\mathcal{A}$, $\mathcal{A}'$ stores (j, s_j) and the parameters of $\mathcal{P}'_{\text{mix}}$: $\mathcal{J}_{-j} := (\mathcal{J}_\ell : \ell \in [k] \setminus \{j\})$ and $\mathcal{B}_{-j} := (\mathcal{B}_\ell : \ell \in [k] \setminus \{j\})$. So the memory state of $\mathcal{A}'$ is $M'_{i,r} := (M_{p_{i+1}-1,r}, j, s_j, \mathcal{J}_{-j}, \mathcal{B}_{-j})$. Noting that (j, s_j) are fixed values rather than random variables, we have

$$\begin{aligned}
\widetilde{\text{IC}}_{j,s_j} =& \sum_{r=1}^{L-1} \sum_{i=1}^{T} I(M_{p_{i+1}-1,r}; X_{p_i} \mid M_{p_i-1,\leq r}, M_{p_{i+1}-1,\leq r-1}, \mathcal{J}, \mathcal{B}) \\
=& \sum_{r=1}^{L-1} \sum_{i=1}^{T} I(M_{p_{i+1}-1,r}, j, s_j, \mathcal{J}_{-j}, \mathcal{B}_{-j}; X_{p_i}, j \mid M_{p_i-1,\leq r}, M_{p_{i+1}-1,\leq r-1}, \mathcal{J}, \mathcal{B}) \\
=& \sum_{r=1}^{L-1} \sum_{i=1}^{T} I(M'_{i,r}; Y_i \mid M'_{i-1,\leq r}, M'_{i,\leq r-1}, \mathcal{J}, \mathcal{B}) \\
\geq& I(M'; Y \mid \mathcal{J}, \mathcal{B}) \geq \frac{d}{2}\left(\frac{\delta(j, s_j)}{2(\rho+1)} - \frac{\rho}{d}\right)^2 - 1.
\end{aligned}$$

The last inequality is by Theorem 28, and the second last inequality is by Claim 11.

C.3. Proof of Theorem 4

By Lemma 10 and Theorem 30, we immediately conclude Theorem 4 for any algorithm satisfying the restriction (i.e., the algorithm does not do any operations during the last pass). Finally, this restriction can removed without affect the theorem: given any L-pass streaming algorithm $\mathcal{A}$, we can easily get an $(L+1)$-pass restricted streaming algorithm $\mathcal{A}'$ with the same advantage and memory size as follows:

1. $\mathcal{A}'$ simulate $\mathcal{A}$ in the first L passes.
2. $\mathcal{A}'$ does not do any operation during the $(L+1)$-th pass.

It is easy to verify that if the lower bound $\Omega\left(\frac{k^2 d}{N(L+1)\rho^2}\right)$ holds for $\mathcal{A}'$, it also holds for $\mathcal{A}$.