Streaming Lower Bounds for the Needle Problems

Jiapeng Zhang

University of Southern California

October 16, 2023

Joint works with Shachar Lovett; Qian Li and Shuo Wang

Overview

The Needle Problem

The Needle Peoblem

Definition

Let $n > 1$ be a large integer and let $p > 0$ be a parameter.

- **Uniform distribution D_0:** each sample is uniformly sampled from $[n]$.
- **Needle distribution D_1:** First sample a needle $x \in [n]$. Each sample equals x with probability p, and uniformly otherwise.

The Needle Peoblem

Definition

Let $n > 1$ be a large integer and let $p > 0$ be a parameter.

- **Uniform distribution D_0:** each sample is uniformly sampled from $[n]$.
- **Needle distribution D_1:** First sample a needle $x \in [n]$. Each sample equals x with probability p, and uniformly otherwise.

Question:

given a bounded memory of s bits, how many samples t are needed to distinguish these two distributions?

Simple Algorithms

Algorithm 1

Put all samples in the memory, and find the most frequent element.

Simple Algorithms

Algorithm 1

Put all samples in the memory, and find the most frequent element. It needs $\Theta(1/p)$ samples and $\Theta((\log n)/p)$ space.

Simple Algorithms

Algorithm 1

Put all samples in the memory, and find the most frequent element. It needs $\Theta(1/p)$ samples and $\Theta((\log n)/p)$ space.

Algorithm 2

Keep the most recent two samples in the memory, and check the consecutive identical elements.

Simple Algorithms

Algorithm 1

Put all samples in the memory, and find the most frequent element. It needs $\Theta(1/p)$ samples and $\Theta((\log n)/p)$ space.

Algorithm 2

Keep the most recent two samples in the memory, and check the consecutive identical elements. It needs $\Theta(1/p^2)$ samples and $\Theta(\log n)$ space.

Simple Algorithms

Algorithm 1

Put all samples in the memory, and find the most frequent element. It needs $\Theta(1/p)$ samples and $\Theta((\log n)/p)$ space.

Algorithm 2

Keep the most recent two samples in the memory, and check the consecutive identical elements. It needs $\Theta(1/p^2)$ samples and $\Theta(\log n)$ space.

For general space s, we need $t \approx \Theta((\log n)/(s \cdot p^2))$ samples.

Our Results

Theorem (Lovett-Z, Li-Wang-Z)

Any streaming algorithm that distinguishes the needle distribution needs $t = \Omega(1/(s \cdot p^2))$ *samples.*

Our Results

Theorem (Lovett-Z, Li-Wang-Z)

Any streaming algorithm that distinguishes the needle distribution needs $t = \Omega(1/(s \cdot p^2))$ samples. For ℓ-pass streaming algorithm, it needs $t = \Omega(1/(\ell \cdot s \cdot p^2))$ samples.

Frequency Estimation

Corollary

In the random order setting. It requires $\Omega(n^{1-2/k})$ space for a streaming algorithm to approximate k-the frequency moment of a data stream.

Frequency Estimation

Corollary

In the random order setting. It requires $\Omega(n^{1-2/k})$ space for a streaming algorithm to approximate k-the frequency moment of a data stream.

Theorem (Andoni-McGregor-Onak-Panigrahy)

In the random order setting. It requires $\Omega(n^{1-2.5/k})$ space for a streaming algorithm to approximate k-the frequency moment of a data stream.

Lower Bounds of the Needle Problems

Disjointness Problem

Definition

There are k players with each of them holds a (random) set $S_i \subseteq [n]$. It is promised that,

- **Disjoint**: the sets $S_1, \ldots, S_k$ are pairwise disjoint.
- **Unique intersection**: there is an $x \in S_1 \cap \ldots \cap S_k$, and the sets $S_1 \setminus \{x\}, \ldots, S_k \setminus \{x\}$ are pairwise disjoint.

Disjointness Problem

Definition

There are k players with each of them holds a (random) set $S_i \subseteq [n]$. It is promised that,

- **Disjoint**: the sets $S_1, \ldots, S_k$ are pairwise disjoint.
- **Unique intersection**: there is an $x \in S_1 \cap \ldots \cap S_k$, and the sets $S_1 \setminus \{x\}, \ldots, S_k \setminus \{x\}$ are pairwise disjoint.

Theorem

The randomized communication complexity of the disjointness problem is $\Omega((|S_1| + \cdots + |S_k|)/k)$

A Simple Case of the Needle Problem

In expectation, there are $\Theta(p \cdot t)$ needles in the stream.

A Simple Case of the Needle Problem

In expectation, there are $\Theta(p \cdot t)$ needles in the stream.

$$
\begin{aligned}
& X_1, \ldots, X_{i_1}, \ldots, X_{(1/p)} \\
& X_{(1/p)+1}, \ldots, X_{i_2}, \ldots, X_{(2/p)} \\
& \ldots \\
& X_{t+1-(1/p)}, \ldots, X_{i_\ell}, \ldots, X_t
\end{aligned}
$$

The symmetric case.

Needle Algorithm to Communication Protocol

Communication protocols

Let $\mathcal{A}$ be an algorithm that distinguishes needles. Recall that each communication player i has a set S_i.

Needle Algorithm to Communication Protocol

Communication protocols

Let $\mathcal{A}$ be an algorithm that distinguishes needles. Recall that each communication player i has a set S_i.

- The first player randomly order S_1, and sends $M_1 := \mathcal{A}(S_1)$ to the second player.

Needle Algorithm to Communication Protocol

Communication protocols

Let $\mathcal{A}$ be an algorithm that distinguishes needles. Recall that each communication player i has a set S_i.

- The first player randomly order S_1, and sends $M_1 := \mathcal{A}(S_1)$ to the second player.
- The second player randomly order S_2, and sends $M_2 := \mathcal{A}(M_1, S_2)$ to the third player.

Needle Algorithm to Communication Protocol

Communication protocols

Let $\mathcal{A}$ be an algorithm that distinguishes needles. Recall that each communication player i has a set S_i.

- The first player randomly order S_1, and sends $M_1 := \mathcal{A}(S_1)$ to the second player.
- The second player randomly order S_2, and sends $M_2 := \mathcal{A}(M_1, S_2)$ to the third player.
- $\ldots$

Needle Algorithm to Communication Protocol

Communication protocols

Let $\mathcal{A}$ be an algorithm that distinguishes needles. Recall that each communication player i has a set S_i.

- The first player randomly order S_1, and sends $M_1 := \mathcal{A}(S_1)$ to the second player.
- The second player randomly order S_2, and sends $M_2 := \mathcal{A}(M_1, S_2)$ to the third player.
- ...
- The last player receives M_{k-1} and outputs $\mathcal{A}(M_{k-1}, S_k)$.

Needle Algorithm to Communication Protocol

Communication protocols

Let $\mathcal{A}$ be an algorithm that distinguishes needles. Recall that each communication player i has a set S_i.

- The first player randomly order S_1, and sends $M_1 := \mathcal{A}(S_1)$ to the second player.
- The second player randomly order S_2, and sends $M_2 := \mathcal{A}(M_1, S_2)$ to the third player.
- ...
- The last player receives M_{k-1} and outputs $\mathcal{A}(M_{k-1}, S_k)$.

The total communication cost is $(k \cdot s)$ bits.

From Communication to Needle Lower Bounds

- The total communication cost is $(k \cdot s)$ bits.

From Communication to Needle Lower Bounds

- The total communication cost is $(k \cdot s)$ bits.
- From the communication lower bounds of disjointness, we have that $(k \cdot s) = \Omega((|S_1| + \cdots + |S_k|)/k = \Omega(t/k)$.

From Communication to Needle Lower Bounds

- The total communication cost is $(k \cdot s)$ bits.
- From the communication lower bounds of disjointness, we have that $(k \cdot s) = \Omega((|S_1| + \cdots + |S_k|)/k = \Omega(t/k)$.
- Recall that $k = t \cdot p$, hence $s \cdot t = \Omega(1/p^2)$

Asymmetric Disjointness

Asymmetric Case

We still assume there are $p \cdot t$ needles.

Asymmetric Case

We still assume there are $p \cdot t$ needles.

$$
\begin{aligned}
&X_1, \ldots, {\color{red}X_{i_1}}, \ldots, {\color{red}X_{i_2}}, \ldots, X_{(1/p)} \\
&X_{(1/p)+1}, \ldots, X_{(2/p)} \\
&\ldots \\
&X_{t+1-(1/p)}, \ldots, {\color{red}X_{i_\ell}}, \ldots, X_t
\end{aligned}
$$

Then the first player would know the answer. In expectation, there are two needles with distance $O(1/(p^2 \cdot t))$.

Asymmetric Disjointness

Definition

There are k players with each of them holds a (random) set $S_i \subseteq [n]$ of size at most s_i. It is promised that either these sets are pairwise disjoint or have a unique intersection

Asymmetric Disjointness

Definition

There are k players with each of them holds a (random) set $S_i \subseteq [n]$ of size at most s_i. It is promised that either these sets are pairwise disjoint or have a unique intersection

Theorem (Lovett-Z)

Let Π be a randomized protocol that solves the asymmetric disjointness. Let c_i be the communication bits by the i-th player. Then we have that,

$$\sum_{i \in [k]} \frac{c_i}{s_i} = \Omega(1).$$

Needle Bounds from Asymmetric Disjointness

Theorem (Lovett-Z)

Any algorithm that distinguishes the needle distribution needs $t = \Omega(1/(s \cdot p^2 \cdot \log n))$ *samples.*

Information Complexity Approaches

The Needle Peoblem

Definition

Let $n > 1$ be a large integer and let $p > 0$ be a parameter.

- **Uniform distribution D_0:** each sample is uniformly sampled from $[n]$.
- **Needle distribution D_1:** Sample a needle x. Each sample equals x with probability p, and uniformly otherwise.

The Needle Peoblem

Definition

Let $n > 1$ be a large integer and let $p > 0$ be a parameter.

- **Uniform distribution D_0:** each sample is uniformly sampled from $[n]$.
- **Needle distribution D_1:** Sample a needle x. Each sample equals x with probability p, and uniformly otherwise.
- **Local needle distribution D^S:** Sample a needle x. Each sample in S equals $x \in S$ with probability p, and uniformly otherwise.

Proof Strategy

- We consider $|S| \approx 2 \cdot p \cdot t$

Proof Strategy

- We consider $|S| \approx 2 \cdot p \cdot t$
- A half of elements from S are the needle

Proof Strategy

- We consider $|S| \approx 2 \cdot p \cdot t$
- A half of elements from S are the needle
- $\boldsymbol{D}_1 = \sum_S \alpha_S \cdot \boldsymbol{D}^S$, where $\sum_S \alpha_S = 1$.

Proof Strategy

- We consider $|S| \approx 2 \cdot p \cdot t$
- A half of elements from S are the needle
- $\boldsymbol{D}_1 = \sum_S \alpha_S \cdot \boldsymbol{D}^S$, where $\sum_S \alpha_S = 1$.
- If $\mathcal{A}$ distinguishes $\boldsymbol{D}_0$ and $\boldsymbol{D}_1$, then it distinguishes $\boldsymbol{D}_0$ and $\boldsymbol{D}^S$ for many S.

Proof Strategy

- We consider $|S| \approx 2 \cdot p \cdot t$
- A half of elements from S are the needle
- $\boldsymbol{D}_1 = \sum_S \alpha_S \cdot \boldsymbol{D}^S$, where $\sum_S \alpha_S = 1$.
- If $\mathcal{A}$ distinguishes $\boldsymbol{D}_0$ and $\boldsymbol{D}_1$, then it distinguishes $\boldsymbol{D}_0$ and $\boldsymbol{D}^S$ for many S.
- The information cost of distinguishing $\boldsymbol{D}^S$ and $\boldsymbol{D}^0$ is $\Omega(1)$

Proof Strategy

- We consider $|S| \approx 2 \cdot p \cdot t$
- A half of elements from S are the needle
- $\boldsymbol{D}_1 = \sum_S \alpha_S \cdot \boldsymbol{D}^S$, where $\sum_S \alpha_S = 1$.
- If $\mathcal{A}$ distinguishes $\boldsymbol{D}_0$ and $\boldsymbol{D}_1$, then it distinguishes $\boldsymbol{D}_0$ and $\boldsymbol{D}^S$ for many S.
- The information cost of distinguishing $\boldsymbol{D}^S$ and $\boldsymbol{D}^0$ is $\Omega(1)$
- The information cost of distinguishing $\boldsymbol{D}_1$ and $\boldsymbol{D}^0$ is $\Omega(1/p^2)$

Information Complexity

Definition (Braverman-Garg-Woodruff)

Let $\mathcal{A}$ be a streaming algorithm. We define its information complexity by,

$$\mathrm{IC}(\mathcal{A}, \boldsymbol{D_0}) := \sum_{i=1}^{t} \sum_{k=1}^{i} I(\boldsymbol{M}_i ; \boldsymbol{X}_k \mid \boldsymbol{M}_{k-1})$$

Information Complexity

Definition (Braverman-Garg-Woodruff)

Let $\mathcal{A}$ be a streaming algorithm. We define its information complexity by,

$$\mathrm{IC}(\mathcal{A}, \boldsymbol{D_0}) := \sum_{i=1}^{t} \sum_{k=1}^{i} I(\boldsymbol{M}_i; \boldsymbol{X}_k \mid \boldsymbol{M}_{k-1})$$

Lemma

$IC(\mathcal{A}, \boldsymbol{D_0}) \leq t \cdot s$

Local Information Complexity

Definition

Let $\mathcal{A}$ be a streaming algorithm and let $S = \{p_1, \ldots, p_m\}$ be a set. We define the local information complexity by,

$$\mathrm{IC}^S(\mathcal{A}, \boldsymbol{D_0}) := \sum_{i=1}^{m} \sum_{k=1}^{i} I(\boldsymbol{M}_{p_{i+1}-1}; \boldsymbol{X}_{p_k} \mid \boldsymbol{M}_{p_k-1}).$$

Lemma

- *If $\mathcal{A}$ distinguishes $\boldsymbol{D_0}$ and $\boldsymbol{D}^S$, then $IC^S(\mathcal{A}, \boldsymbol{D_0}) = \Omega(1)$*
- $IC(\mathcal{A}, \boldsymbol{D_0}) \approx \mathbb{E}_S[IC^S(\mathcal{A}, \boldsymbol{D_0})]/p^2$

Tight Needle Lower Bounds

Theorem (Li-Wang-Z)

Any algorithm that distinguishes the needle distribution needs $t = \Omega(1/(s \cdot p^2))$ samples.

Tight Needle Lower Bounds

Theorem (Li-Wang-Z)

Any algorithm that distinguishes the needle distribution needs $t = \Omega(1/(s \cdot p^2))$ samples.

Lower bounds can be extended to the multi-pass setting by a multi-pass information complexity notion.

Thank you!