

# A Note on Space Lower Bound for Samplers

January 5, 2017

We study the space lower bound for maintaining a sampler over a turnstile stream. An  $\ell_p$ -sampler with failure probability at most  $\delta$  is a randomized data structure for maintaining vector  $x \in \mathbb{R}^n$  (initially 0) under a stream of updates in the form of  $(i, \Delta)$  (meaning that  $x_i \leftarrow x_i + \Delta$ ); in the end, with probability at least  $1 - \delta$ , it gives an “ $\ell_p$ -sample” according to  $x$ : namely, item  $i$  is sampled with probability  $\frac{|x_i|^p}{\sum_{j \in [n]} |x_j|^p}$ .

Note that updates are independent of the randomness used in the sampler. That is, for the purpose of proving a lower bound, we assume an oblivious adversary.

To the best of my knowledge, the best space upper bound for  $\ell_0$  sampler is  $O(\log^2 n \log \frac{1}{\delta})$  bits, while the previous best lower bound is  $\Omega(\log^2 n + \log \frac{1}{\delta})$  bits (where  $\Omega(\log^2 n)$  is shown in [JST11]). The bound is tight for constant  $\delta$ , while for example, when  $\delta = \frac{1}{n}$ , the gap is  $\log n$ .

We assume that

$$2^{-n^{c_1}} < \delta < (\log n)^{-c_2}, \quad (1)$$

where  $c_1 = 0.01$  and  $c_2 = 2$ . For other range of  $\delta$  we will study later.

In this note, we show space lower bounds for maintaining a sampler for a *binary* vector. That is, at any time, we are guaranteed that  $x \in \{0, 1\}^n$ . This makes our result strong in the sense that (1) the lower bound applies for any  $p$ ; (2) the lower bound also works for strict turnstile streams.

In the following sections, we give sequentially improved lower bounds. First, we give a lower bound of  $\Omega(\log n \log \frac{1}{\delta})$  bits. Then, we improve it to  $\Omega(\log n \log \frac{1}{\delta} \log \frac{\log n}{\log \log \frac{1}{\delta}})$  bits, and finally  $\Omega(\frac{\log^2 n \log \frac{1}{\delta}}{\log \log n + \log \log \frac{1}{\delta}})$  bits. The lower bounds are based on communication complexity in the public random coin model. Alice wants to send Bob a uniform random set  $A \subseteq [n]$  of size  $m$  (Bob knows  $m$ , but the random source generating  $A$  is independent of the random source accessible to Bob). The one-way communication problem is: Alice sends some message to Bob, and Bob is required to recover  $A$  completely. Since the randomness in  $A$  contains  $\log \binom{n}{m}$  bits of information, any randomized protocol requires at least  $\log \binom{n}{m}$  expected bits.

Now Alice considers to attach (the memory of) a sampler **SAMP** in the message. The sampler uses public random coins as its random source, so that the sampler will behave the same at Alice’s and Bob’s as long as the updates are all the same. Alice will insert all the items in  $A$  into **SAMP** and send **SAMP** to Bob. In addition, Alice will send a subset  $B \subseteq A$  to Bob, so that together with  $B$  and **SAMP**, Bob is able to recover  $A$  with good probability based on some protocol they have agreed on.

Now we turn the previous protocol into a new one without any failure. Let **SUCC** denote the event (or a subset of the event) that Bob successfully recovers  $A$  (note that Alice can simulate Bob, so she knows exactly when **SUCC** happens). If **SUCC** happens Alice will send Bob a message starting with a 1, followed by (the memory of) **SAMP**, then followed by the native encoding of  $B$ ; otherwise, Alice will send a message starting with a 0, followed by the native encoding of  $A$ . We say the native encoding of a set  $S \subseteq [n]$  to be an integer (expressed in binary) in  $[\binom{n}{|S|}]$  together with  $|S|$  (taking  $\log n$  bits). We drop the size of the set if it is known by the receiver.

**Lemma 1.** *Let  $s$  denote the space (in bits) used by a sampler with failure probability at most  $\delta$ . Let  $s'$  denote the expected number of bits to represent  $B$  conditioned on **SUCC** (if we need to send some extra auxiliary information, we will also count it into  $s'$ ). We have*

$$(1 + s + s') \cdot \mathbb{P}(SUCC) + (1 + \log \binom{n}{m}) \cdot (1 - \mathbb{P}(SUCC)) \geq \log \binom{n}{m}.$$

If  $\mathbb{P}(SUCC) \geq 1/2$ , we have

$$s \geq \log \binom{n}{m} - s' - 2. \quad (2)$$

**Remark 1.** Because the space lower bound in this note is proven via communication complexity under public random coin model, it also applies to non-uniform models of computation such as circuits and branching programs.

## 1 $\Omega(\log n \log \frac{1}{\delta})$ Bits Lower Bound

Let  $m = \frac{1}{2} \log \frac{1}{\delta}$ , namely, Alice wants to send a uniform random set  $A \subseteq [n]$  of size  $\frac{1}{2} \log \frac{1}{\delta}$  to Bob. Let  $A = \{a_1, \dots, a_m\}$  and  $a_1 < \dots < a_m$ .

---

**Algorithm 1** Alice's Encoder.

---

```

1: procedure ENC1(A)
2:   SAMP  $\leftarrow \emptyset$ 
3:   for  $i = 1, 2, \dots, m$  do
4:     Insert  $a_i$  into SAMP
5:   end for
6:   return SAMP
7: end procedure

```

---



---

**Algorithm 2** Bob's Decoder.

---

```

1: procedure DEC1(SAMP)
2:    $S \leftarrow \emptyset$ 
3:   for  $i = 1, 2, \dots, m$  do
4:     Let SAMP $i$  be a copy of SAMP
5:     for  $s \in S$  do
6:       Remove  $s$  from SAMP $i$ 
7:     end for
8:     Obtain a sample  $s_i$  from SAMP $i$ 
9:      $S \leftarrow S \cup \{s_i\}$ 
10:  end for
11:  return  $S$ 
12: end procedure

```

---

**Lemma 2.** For any  $A \subseteq [n]$ , where  $|A| = m = \frac{1}{2} \log \frac{1}{\delta}$ ,  $\mathbb{P}(\text{DEC}_1(\text{ENC}_1(A)) = A) \geq 1/2$ .

*Proof.* Let  $E_S$  denote the event that after removing all the items in  $S$  (in the order from smallest to biggest) from SAMP, it gives a valid sample when queried. We have

$$\mathbb{P}(\text{DEC}_1(\text{ENC}_1(A)) = A) \geq \mathbb{P}\left(\bigcap_{S \subseteq A, S \neq \emptyset} E_S\right) \geq 1 - \sum_{S \subseteq A, S \neq \emptyset} \mathbb{P}(\overline{E_S}) \geq 1 - \delta \cdot 2^{\frac{1}{2} \log \frac{1}{\delta}} \geq 1/2.$$

□

**Lemma 3.**  $s = \Omega(\log n \log \frac{1}{\delta})$  for  $2^{-n^{0.9}} < \delta < \frac{1}{4}$ .

*Proof.* It follows from Formula 2 in Lemma 1, where  $\log \left( \frac{n}{\frac{1}{2} \log \frac{1}{\delta}} \right) = \Omega(\log n \log \frac{1}{\delta})$  and  $\mathbf{s}' = 0$ .  $\square$

**Remark 2.** The following decoder  $DEC'_1$  is similar to  $DEC_1$ , but we will lose a factor of  $\log \log \frac{1}{\delta}$  in the lower bound because by doing so we have to union-bound  $O(m!)$  events instead of  $O(2^m)$  events (so that in turn we have to set  $m$  to be  $\frac{\log \frac{1}{\delta}}{\log \log \frac{1}{\delta}}$  in order to have good success probability).

---

**Algorithm 3** A Worse Decoder.

---

```

1: procedure  $DEC'_1(\text{SAMP})$ 
2:    $S \leftarrow \emptyset$ 
3:   for  $i = 1, 2, \dots, m$  do
4:     Obtain a sample  $s_i$  from  $\text{SAMP}$ 
5:     Remove  $s_i$  from  $\text{SAMP}$ 
6:      $S \leftarrow S \cup \{s_i\}$ 
7:   end for
8:   return  $S$ 
9: end procedure

```

---

## 2 $\Omega(\log n \log \frac{1}{\delta} \log \frac{\log n}{\log \log \frac{1}{\delta}})$ Bits Lower Bound

In the previous section we have shown how to extract  $\Theta(\log \frac{1}{\delta})$  words of information from a sampler. Our goal is to extract more words. New observation comes from the upper bound for constructing an  $\ell_0$  sampler. Let  $\delta = \frac{1}{n}$ , we have the following sampler algorithm that consumes  $O(\log^3 n)$  bits. The sampler consists of  $\log n$  layers, and on layer  $i \in [\log n]$  it maintains a separate  $\log n$ -sparse recover system for the sub-stream generated by sub-sampling the items from the universe  $[n]$  with probability  $2^{-i}$ . Each sparse recovery system takes  $O(\log^2 n)$  bits. Its correctness comes from the fact that

1. With probability at least  $1 - n^{-c}$  there is some layer  $i$  that contains at least 1 and at most  $\log n$  items whose coordinates are non-zero.
2. Conditioned on the event that on layer  $i$  the number of items whose coordinates are non-zero is between 1 and  $\log n$ , the sparse recovery system on layer  $i$  works with failure probability at most  $n^{-c}$ . In this context, we say the sparse recovery system works if it could recover at least one item.

Intuitively, the previous lower bound only extracts information from one single layer (i.e. layer 0). In this section, we build a framework to extract information from  $\Theta(\log n)$  layers.

Alice wants to send random set  $A$  to Bob where  $|A| = m = n^{1/4}$ . Similar to  $\text{ENC}_1$ , Alice constructs  $\text{SAMP}$  by inserting all the items in  $A$ , and sends it to Bob. Moreover, Alice will send Bob a subset  $B \subseteq A$  computed as follows. Initially  $B = A$ . Alice proceeds in  $R = \log m - \log \log \frac{1}{\delta}$  rounds. Let  $A_r = \{a \in A \mid a < \frac{n}{2^{r-1}}\}$ . Furthermore, let  $A_r^- = A_{r+1}$  and  $A_r^+ = A_r \setminus A_r^-$ . On round  $r$  (for  $r = 1, \dots, R$ ):

1. Alice makes a copy of  $\text{SAMP}$ , denoted by  $\text{SAMP}_r$ , and removes all items in  $A \setminus A_r$  from  $\text{SAMP}_r$ .
2. Similar to  $\text{ENC}'_1$ , obtain  $n_r = \frac{1}{2} \cdot \frac{\log \frac{1}{\delta}}{(\log m) + 2 - r}$  samples from  $\text{SAMP}_r$ , denoted by  $S_r$ .
3. Remove items in  $S_r \cap A_r^+$  from  $B$ .

---

**Algorithm 4** Alice's Encoder.

---

```
1: procedure ENC2( $A$ )
2:    $\text{SAMP} \leftarrow \emptyset$ 
3:   Insert all elements in  $A$  into  $\text{SAMP}$ 
4:    $B \leftarrow A$ 
5:   for  $r = 1, \dots, R$  do
6:     Let  $\text{SAMP}_r$  be a copy of  $\text{SAMP}$ 
7:     Remove all elements in  $A \setminus A_r$  from  $\text{SAMP}_r$ 
8:      $S_r \leftarrow \emptyset$ 
9:     for  $i = 1, \dots, n_r$  do
10:      Obtain a sample  $s$  from  $\text{SAMP}_r$ 
11:      Remove  $s$  from  $\text{SAMP}_r$ 
12:       $S_r \leftarrow S_r \cup \{s\}$ 
13:     end for
14:      $B \leftarrow B \setminus (A_r^+ \cap S_r)$ 
15:   end for
16:   return ( $\text{SAMP}, B$ )
17: end procedure
```

---

---

**Algorithm 5** Bob's Decoder.

---

```
1: procedure DEC2( $\text{SAMP}, B$ )
2:    $A \leftarrow B$ 
3:   for  $r = 1, 2, \dots, R$  do
4:     Let  $\text{SAMP}_r$  be a copy of  $\text{SAMP}$ 
5:     Remove all the items in  $\{a \in A \mid a \geq \frac{n}{2^{r-1}}\}$  from  $\text{SAMP}_r$ 
6:      $S_r \leftarrow \emptyset$ 
7:     for  $i = 1, \dots, n_r$  do
8:      Obtain a sample  $s$  from  $\text{SAMP}_r$ 
9:      Remove  $s$  from  $\text{SAMP}_r$ 
10:       $S_r \leftarrow S_r \cup \{s\}$ 
11:     end for
12:      $A \leftarrow A \cup \{a \in S_r \mid a \geq \frac{n}{2^r}\}$ 
13:   end for
14:   return  $A$ 
15: end procedure
```

---

### 3 Lower Bound for a Promised Problem

Peeling off by chunks (instead of one by one).

## 4 $\Omega\left(\frac{\log^2 n \log \frac{1}{\delta}}{(\log \log n + \log \log \frac{1}{\delta})^2}\right)$ Bits Lower Bound

---

**Algorithm 6** Variables Shared by Alice's  $\text{ENC}_4$  and Bob's  $\text{DEC}_4$ .

---

```

1:  $m \leftarrow n^{0.99}$ 
2:  $K \leftarrow \log n \cdot \log \frac{1}{\delta}$ 
3:  $R \leftarrow \frac{\log n}{50 \log K}$ 
4:  $i_1 \leftarrow 0$ 
5:  $\Delta_1 \leftarrow n$ 
6: for  $r = 1, \dots, R - 1$  do
7:   Let  $k_r$  be a uniform sample from  $\{1, \dots, K\}$ 
8:    $i_{r+1} \leftarrow i_r + \frac{k_r - 1}{K} \Delta_r$ 
9:    $\Delta_{r+1} \leftarrow \frac{\Delta_r}{K}$ 
10: end for
11: for  $r = 1, \dots, R$  do
12:    $n_r \leftarrow \frac{1}{2} \cdot \frac{\log \frac{1}{\delta}}{\log K}$ 
13:    $I_r \leftarrow \{i \in \mathbb{N} \mid i_r \leq i < i_r + \Delta_r\}$ 
14:   for  $k = 1, \dots, K$  do
15:      $I_{r,k} \leftarrow \{i \in \mathbb{N} \mid i_r + (k-1) \frac{\Delta_r}{K} \leq i < i_r + k \frac{\Delta_r}{K}\}$ 
16:   end for
17: end for

```

---



---

**Algorithm 7** Alice's Encoder.

---

```

1: procedure  $\text{ENC}_4(A)$ 
2:    $\text{SAMP} \leftarrow \emptyset$ 
3:   Insert all elements in  $A$  into  $\text{SAMP}$ 
4:    $B \leftarrow A$ 
5:   for  $r = 1, \dots, R$  do
6:     Let  $\text{SAMP}_r$  be a copy of  $\text{SAMP}$ 
7:     Remove all elements in  $A \setminus (A \cap I_r)$  from  $\text{SAMP}_r$ 
8:     for  $i = 1, \dots, n_r$  do
9:       Obtain a sample  $s$  from  $\text{SAMP}_r$ 
10:      if  $s$  is not valid then
11:        return "fail"
12:      end if
13:      Find  $k$  such that  $s \in I_{r,k}$ 
14:      if  $k = k_r$  then
15:        return "fail"
16:      end if
17:      Remove all elements in  $A \cap I_{r,k}$  from  $\text{SAMP}_r$ 
18:       $B \leftarrow B \setminus \{s\}$ 
19:    end for
20:  end for
21:  return  $(\text{SAMP}, B)$ 
22: end procedure

```

---

---

**Algorithm 8** Bob's Decoder.

---

```
1: procedure DEC4(SAMP,  $B$ )
2:    $A \leftarrow B$ 
3:   for  $r = 1, 2, \dots, R$  do
4:     Let SAMP $r$  be a copy of SAMP
5:     Remove all the items in  $A \setminus (A \cap I_r)$  from SAMP $r$ 
6:     for  $i = 1, \dots, n_r$  do
7:       Obtain a sample  $s$  from SAMP $r$ 
8:        $A \leftarrow A \cup \{s\}$ 
9:       Find  $k$  such that  $k \in I_{r,k}$ 
10:      Remove items in  $A \cap I_{r,k}$  from SAMP $r$ 
11:    end for
12:  end for
13:  return  $A$ 
14: end procedure
```

---

#### 4.1 Analysis

Note that  $\mathbb{E}(|A \cap I_{r,k}|) = m \cdot K^{-r}$ . By the choice of parameters we have  $m = n^{0.99}$ ,  $K^{-r} \geq K^{-R} = n^{-0.02}$  (note that the range of  $\delta$  is specified by Formula 1). Therefore  $\mathbb{E}(|A \cap I_{r,k}|) \geq n^{0.97}$ , for  $r = 1, \dots, R$  and  $k = 1, \dots, K$ . For any pair of  $(r, k)$ , because of the randomness in  $A$ , we can prove the probability that  $A \cap I_{r,k}$  is empty is exponentially small. By union bound, with high probability, all  $A \cap I_{r,k}$  are not empty. In the following, we will discuss conditioned on that.

**Lemma 4.** *With probability at least  $\frac{9}{10}$ , ENC<sub>4</sub> does not return “fail”.*

*Proof.* The probability that ENC<sub>4</sub> returns fail on line 11 is at most  $\sum_{i=1}^R K^{n_i} \cdot \delta = \frac{\delta^{1/2} \log n}{50 \log K} < \frac{1}{10}$ .

The probability that ENC<sub>4</sub> returns fail on line 15 is at most  $\sum_{r=1}^R \frac{n_r}{K} = \frac{1}{100 \log^2 K}$ . □

**Lemma 5.** *If ENC<sub>4</sub> does not return “fail”, we have DEC<sub>4</sub>(ENC<sub>4</sub>( $A$ )) =  $A$ .*

**Lemma 6.** *If ENC<sub>4</sub> does not return “fail”, we have  $|A| - |B| = \Omega(\frac{\log n \log \frac{1}{\delta}}{(\log \log n + \log \log \frac{1}{\delta})^2})$ , where  $B$  is the set ENC<sub>4</sub> outputs.*

*Proof.*  $|A| - |B| = \sum_{r=1}^R n_r = \frac{\log n}{50 \log K} \cdot \frac{1}{2} \cdot \frac{\log \frac{1}{\delta}}{\log K} = \Omega(\frac{\log n \log \frac{1}{\delta}}{(\log \log n + \log \log \frac{1}{\delta})^2})$ . □

**Lemma 7.** *Let  $m = n^{0.99}$ . Let  $X \in \mathbb{N}$  be a random variable, and  $X \leq m$ . Moreover,  $\mathbb{E}(X) \leq m - d$ . We have  $\mathbb{E}(\log \binom{n}{m} - \log \binom{n}{X}) = \Omega(d \log n)$ .*

*Proof.*

$$\begin{aligned} \log \binom{n}{m} - \log \binom{n}{X} &= \log \frac{n!/(m!(n-m)!)}{n!/(X!(n-X)!)} \\ &= \sum_{i=1}^{m-X} \log \frac{n-X-i+1}{m-i+1} \\ &\geq (m-X) \cdot \log \frac{n-X}{m} \\ &\geq (m-X) \cdot \log n^{1/200} \end{aligned}$$

Taking expectation on both sides, we get  $\mathbb{E}(\log \binom{n}{m} - \log \binom{n}{X}) \geq \frac{d}{200} \log n$ . □

**Theorem 1.**  $s = \Omega\left(\frac{\log^2 n \log \frac{1}{\delta}}{(\log \log n + \log \log \frac{1}{\delta})^2}\right)$  for  $2^{-n^{0.01}} < \delta < (\log n)^{-2}$ .

*Proof.* Let SUCC be the conjunction of the following events:

1. All  $A \cap I_{r,k}$  are non-empty ( $r = 1, \dots, R, k = 1, \dots, K$ ).
2. ENC<sub>4</sub> does not return “fail” on input  $A$ .

By Lemma 4, we have  $\mathbb{P}(\text{SUCC}) \geq \frac{1}{2}$ . By Lemma 1, we have  $s \geq \log\binom{n}{m} - s' - 2$ . By definition,  $s' = \log n + \mathbb{E}(\log\binom{n}{|B|} | \text{SUCC})$ . By Lemma 6 and Lemma 7, we get  $\mathbb{E}(\log\binom{n}{|B|} | \text{SUCC}) = \log\binom{n}{m} - \Omega\left(\frac{\log^2 n \log \frac{1}{\delta}}{(\log \log n + \log \log \frac{1}{\delta})^2}\right)$ .  $\square$

## References

- [JST11] Hossein Jowhari, Mert Sağlam, and Gábor Tardos. Tight bounds for lp samplers, finding duplicates in streams, and related problems. In *Proceedings of the thirtieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 49–58. ACM, 2011.