CS-448 Sublinear Algorithms for Big Data Analysis April 30, 2025

Lecture 10

Lecturer: Michael Kapralov

In this lecture we prove the lower bound for the INDEX problem, and then show a number of applications.

1 The INDEX problem

Alice has $x \in \{0,1\}^n$ and Bob is given $i \in [n]$. Then, the goal is to compute $f(x,i) = x_i$ on Bob's end with a single message m from Alice. Recall that $R_\delta^{pub,\to}(f)$ stands for the public coin one-way communication complexity of computing a function $f(x,y)$ with error probability at most δ on every input: Alice holds x, Bob holds y, they share a source of random bits and Alice sends a single message to Bob, after which he must output the correct answer with probability at least $1-\delta$ on every fixed pair of inputs.

Claim 1

$$R_\delta^{pub\to}(INDEX) \geq (1 - H_2(\delta))n$$

where $H_2(\delta) = \delta \log_2 \frac{1}{\delta} + (1-\delta)\log_2 \frac{1}{1-\delta}$ *is the binary entropy at* δ*.*

Proof Let X denote the length n vector that Alice holds, and let $X \sim UNIF(\{0,1\})^n$. Let the size of the message that Alice sends be s (we can assume without loss of generality that Alice always sends messages of the same length), and let M be the message. First note that

$$R_\delta^{pub,\to}(M) \geq H(M) \geq I(M;X),$$

The first inequality follows since Alice sends s bits, so $|\text{supp}(M)| \leq 2^s$, and thus $H(M) \leq s$. The second inequality follows from the definition of mutual information and nonnegativity of entropy: $I(M;X) = H(M) - H(M|X) \leq H(M)$.

By correctness of the protocol we know that **for any** x **and** i Bob correctly guesses x_i with probability at least $1-\delta$ (over randomness in Alice's message), i.e., for every i there exists g_i such that $\Pr_M[g_i(M(x)) \neq x_i] \leq \delta$. Letting for any $i \in [n]$ $X_{<i}$ denote the vector $(X_1, \ldots, X_{i-1})$, we get

$$\begin{aligned} I(X;M) &= \sum_{i=1}^{n} I(X_i; M|X_{<i}) \\ &= \sum_{i=1}^{n} H(X_i|X_{<i}) - H(X_i|M, X_{<i}) \\ &\geq \sum_{i=1}^{n} H(X_i) - H(X_i|M) \\ &= \sum_{i=1}^{n} (1 - H(X_i|M)) \\ &\geq (1 - H_2(\delta))n. \end{aligned}$$

We applied the chain rule for mutual information in the first line, used the fact that X_i are i.i.d. and conditioning does not increase entropy in the third line, the fact that X_i are binary in the forth line, and Fano's inequality in the last line.

■

2 MEDIAN lower bound

We define the $MEDIAN$ problem is as follows: given $y_1, \ldots, y_n$, n odd, as a stream, output the exact median of such sequence.

Claim 2 *Any algorithm that solves $MEDIAN$ with $\Omega(1)$ success probability must use $\Omega(n)$ bits of space.*

Proof We reduce $INDEX$ to $MEDIAN$ as follows. Let ALG be an algorithm that solves exact $MEDIAN$. Let $x \in \{0,1\}^n$ be Alice's vector and $i \in [n]$ Bob's index. Alice forms stream $R = (2+x_1, 4+x_2, 6+x_3, \ldots, 2n+x_n)$. Then, Alice sends the state of $ALG(R)$ to Bob (let this memory contents be s bits). Now, Bob feeds 0 ($n-i$ times) and $2n+2$ ($i-1$ times) to ALG in its current memory state. Note that adding 0's and $2n+2$ to the stream centers the stream around the $2i+x_i$. Then, the value of the median in this extended stream will be either $2i$ or $2i+1$ and we recover x_i exactly. Thus, the space complexity of MEDIAN with $\Omega(1)$ success probability is at least $R^{pub,\rightarrow}_{1-\Omega(1)}(INDEX) = \Omega(n)$
■

3 Lower bounds for counting with deletions

We first define the *AUGMENTED-INDEX* problem, prove communication lower bound for it, and then use it to show that any algorithm for constant factor approximate counting with deletions on streams of length n must use $\Omega(\log n)$ bits of space. This is in contrast to the constant factor approximate algorithm for insertion only streams that uses $O(\log \log n)$ space that we developed in the first lecture of the course (Morris' algorithm).

The *AUGMENTED-INDEX* problem is a version of $INDEX$ where Bob is given, in addition to $i \in [n]$, a prefix of x, i.e. $x_{<i}$. The proof is almost identical to the lower bound for INDEX presented above:

Claim 3 $R_{\delta}^{pub,\to}(\text{AUG-INDEX}) \geq (1 - H_2(\delta))n$

Claim 4 *Any streaming algorithm for* $(1 \pm 1/3)$*-approximate counting with deletions that succeeds with probability at least* $9/10$ *for each fixed input that works in streams of length of length up to* $\text{poly}(n)$ *must use* $\Omega(\log n)$ *bits of space.*

Proof Let Alice have $x \in \{0,1\}^{\log n}$. Also, Bob has $i \in [\log(n)]$ and $x_{>i}$. Alice starts by forming stream R with 10^j events for each j such that $x_j = 1$. Alice sends the memory contents of $ALG(R)$ to Bob. Bob, using $x_{>i}$ that is given to him, runs ALG, starting with memory state communicated by Alice, on 10^j deletions per element in the suffix Bob has. Note that ALG, conditioned on the success event that occurs with probability at least 9/10 by assumption, outputs a $(1 \pm 1/3)$-approximation w to $\sum_{j=0}^{i} 10^j x_j$. If $w \geq \frac{2}{3}10^i$ then Bob concludes that $x_i = 1$, else $x_i = 0$.

We now prove correctness. First note that $\sum_{j \leq i} 10^j x_j = 10^i x_i + \sum_{j=0}^{i-1} 10^j x_j$, where the second term in the sum is upper bounded by $\frac{1}{3}10^i$ (by summing the geometric series). Thus, if $x_i = 1$, $\sum_{j \leq i} 10^j x_j \geq 10^i$, and when $x_i = 0$, $\sum_{j \leq i} 10^j x_j \leq 10^i/3$. Thus, a $1 \pm 1/3$ approximation cannot report a value lower than $(2/3)10^i$ in the former case and higher than $(1 + 1/3) \cdot 10^i/3 < (2/3)10^i$ in the latter case, as required. Thus, Bob guesses x_i correctly with probability at least 9/10, and the lower bound of $\Omega(\log n)$ bits follows. ■

4 Gap Hamming Distance (GAPHAM)

Let Alice and Bob have $x \in \{0,1\}^n$ and $y \in \{0,1\}^n$ respectively. We know that *exactly* one of the following two inequalities is satisfied:

$$\Delta(x,y) \geq \frac{n}{2} + C\sqrt{n}$$
$$\Delta(x,y) \leq \frac{n}{2} - C\sqrt{n},$$

where $\Delta(x,y)$ is the Hamming distance between x and y (e.g. number of entries where they differ). The goal of Alice and Bob is to determine which of the two cases above they are in. We will show

Claim 5 *For any constant* C *one has* $R_{1/10}^{pub,\to}(GAPHAM) = \Omega(n)$.

Note that if x and y are chosen uniformly at random, then each of two cases above occurs with constant probability.

4.1 Lower bound for $(1+\epsilon)$-approximate distinct elements

Claim 6 *Any ALG that outputs a* $1 \pm \epsilon$*-approximation to* $\|x\|_0$ *with* $9/10$ *success probability on every fixed input requires* $\Omega(1/\epsilon^2)$ *space.*

Proof We reduce from GAPHAM on vectors of length m. Let Alice and Bob's inputs be denoted by $x \in \{0,1\}^m$ and $y \in \{0,1\}^m$ respectively.

First note that $2\|x+y\|_0 = \|x\|_0 + \|y\|_0 + \Delta(x,y)$. Indeed, interpreting x and y as indicator vectors for sets in $[m]$, we get that $\|x+y\|_0$ is be the cardinality of the union, $\|x\|_0$ and $\|y\|_0$ the cardinality of each set separately and $\Delta(x,y)$ the cardinality of the symmetric difference, so the identity follows.

To solve GAPHAM, Alice sends the memory contents of $ALG(x)$ (s bits) together with $\|x\|_0$ ($\log_2 m$ bits) to Bob. Bob finishes the run of ALG on $x+y$, and thus computes (with constant success probability, as assumed in the claim) $\widehat{L}$ s.t. $|2\|x+y\|_0 - \widehat{L}| \le \epsilon\|x+y\|_0 \le \epsilon m$. Noting that

$$\left|\Delta(x,y) - (2\widehat{L} - ||x||_0 - ||y||_0)\right| = 2\left|||x+y||_0 - \widehat{L}\right| \le 2\epsilon m < \sqrt{m}$$

as long as $m < 1/(2\epsilon^2)$. We thus get that Alice and Bob can distinguish between $\Delta(x,y) \ge m/2 + C\sqrt{m}$ and $\Delta(x,y) \le m/2 - C\sqrt{m}$ with probability at least $9/10$ on every fixed input using space $s + \log_2 m$. We thus get by Claim 5 that $s + \log_2 m = \Omega(m)$, and hence $s = \Omega(m) = \Omega(1/\epsilon^2)$, as required. ■

5 $\Omega(n)$ communication lower bound for GAPHAM

So far, we have seen how to prove the memory lower bound for INDEX problem and reduce GAPHAM to F_0. However to obtain $\Omega(\frac{1}{\epsilon^2})$ space lower bound for F_0, one missing part is to show the reduction from INDEX to GAPHAM, implying an $\Omega(n)$ lower bound for GAPHAM.

Recall the INDEX problem, Alice has a vector $u \in \{0,1\}^n$ and Bob is given a index $i \in [n]$. The goal is to computer u_i on Bob's side after receiving a single message m from Alice. We will think of Alice's input in the INDEX problem as a vector $u \in \{-1,+1\}^n$. Also GAPHAM problem is defined as, given two vector $x, y \in \{-1,+1\}^n$, we want to distinguish whether $\Delta(x,y) \le \frac{n}{2} - C\sqrt{n}$ or $\Delta(x,y) \ge \frac{n}{2} + C\sqrt{n}$, where $\Delta(x,y)$ is the hamming distance between x and y. Now we show how to derive a algorithm for INDEX problem given a protocol for GAPHAM problem. Our plan is described as fellows,

(1) Pick N i.i.d. vectors $r^1, r^2, \ldots, r^N$ where for all $k \in [N]$, $r^k \sim \text{UNIF}(\{-1,+1\}^n)$

(2) For each $k = 1 \ldots N$, let $x_k = \text{sgn}(\langle u, r^k\rangle)$ and $y_k = \text{sgn}(\langle e_i, r^k\rangle)$, where e_i is the standard 0-1 basis vector corresponding to Bob's input.

(3) Feed vectors $x, y \in \{-1,+1\}^N$ into GAPHAM solver. Output $u_i = -1$ if the GAPHAM solver says $\Delta(x,y) \ge \frac{N}{2} + C\sqrt{N}$, otherwise output $u_i = +1$.

Note that,

$$\Delta(x,y) = |\{k \in [n] : \text{sgn}(\langle u, r^k\rangle) \ne \text{sgn}(\langle e_i, r^k\rangle)\}|$$

We start with

Claim 7 *If* $r \sim \mathrm{UNIF}(\{-1,+1\}^n)$*, then*

$$\Pr[\mathrm{sgn}(\langle u, r\rangle) \neq \mathrm{sgn}(\langle e_i, r\rangle)] = \begin{cases} \geq \frac{1}{2} + \frac{c}{\sqrt{n}}, & \text{if } u_i = -1 \\ \leq \frac{1}{2} - \frac{c}{\sqrt{n}}, & \text{if } u_i = 1 \end{cases}$$

where c *is a positive constant.*

Proof Assume without loss of generality that n is odd, and write

$$\langle u, r\rangle = \sum_{j=1}^{n} u_j r_j = u_i r_i + \sum_{j\neq i}^{n} u_j r_j = u_i r_i + w,$$

where $w = \sum_{j\neq i}^{n} u_j r_j$. Suppose that $u_i = -1$. We consider two cases:

$w \neq 0$**.** Then $|w| \geq 2$ for $|w|$ is even. Then $\mathrm{sgn}(\langle u, r\rangle) = \mathrm{sgn}(w)$, which implies

$$\Pr[\mathrm{sgn}(\langle u, r\rangle) = -1] = \Pr[\mathrm{sgn}(\langle u, r\rangle) = 1] = \frac{1}{2}.$$

Thus $\Pr[\mathrm{sgn}(\langle u, r\rangle) \neq \mathrm{sgn}(\langle e_i, r\rangle)] = \frac{1}{2}$.

$w = 0$ Then $\mathrm{sgn}(\langle u, r\rangle) = u_i r_i$. Thus

$$\Pr[\mathrm{sgn}(\langle u, r\rangle) \neq \mathrm{sgn}(\langle e_i, r\rangle)] = 1.$$

Note that w is the sum of $n-1$ even number uniformly distributed variables in $\{-1,+1\}$. By Stirling's formula, when n is large enough, for some constant $c' > 0$, $\Pr[w = 0] \geq \frac{c'}{\sqrt{n}}$ (also, intuitively the distribution of w coverages to a Gaussian distribution with standard deviation $\approx \sqrt{n}$, thus the pdf of this distribution between $-\sqrt{n}$ and $\sqrt{n}$ is $\Omega(\sqrt{n})$). Thus, when $u_i = -1$, we have

$$\Pr[\mathrm{sgn}(\langle u, r\rangle) \neq \mathrm{sgn}(\langle e_i, r\rangle)] = \Pr[w = 0] + \frac{1}{2}(1 - \Pr[w = 0]) \geq \frac{1}{2} + \frac{c'}{2\sqrt{n}} = \frac{1}{2} + \frac{c}{\sqrt{n}}$$

for a constant $c > 0$. Similarly, when $u_i = +1$, we have

$$\Pr[\mathrm{sgn}(\langle u, r\rangle) \neq \mathrm{sgn}(\langle e_i, r\rangle)] \leq \frac{1}{2} - \frac{c}{\sqrt{n}}.$$

■

For $k = 1, 2, \ldots, N$ let

$$Z_k = \begin{cases} 1, & \text{if } x_k \neq y_k \\ 0, & \text{if } x_k = y_k \end{cases}$$

Then $\Delta(x, y) = \sum_{k=1}^{N} Z_k$ and $\mathbb{E}[Z_k] \geq \frac{1}{2} + \frac{c}{\sqrt{n}}$.

Claim 8 *When* $u_i = -1$, $\Pr[\sum_{k=1}^{N} Z_k < \frac{N}{2} + C\sqrt{N}] < 0.1$

Proof By the Chernoff bound, we have

$$\Pr\left[\sum_{k=1}^{N} Z_k < (1-\delta)\sum_{k=1}^{N}\mathbb{E}[Z_k]\right] \leq \exp(-N\mathbb{E}[Z_k]\delta^2/3) \leq \exp(-N\delta^2/6),$$

where δ is chosen so that $(1-\delta)\sum_{k=1}^{N}\mathbb{E}[Z_k] = \frac{N}{2}+C\sqrt{N}$. We now lower bound δ. Since $\sum_{k=1}^{N}\mathbb{E}[Z_k] \geq N/(1/2+c/\sqrt{n})$, we have

$$\delta \geq 1 - \frac{\frac{N}{2}+C\sqrt{N}}{N(\frac{1}{2}+\frac{c}{\sqrt{n}})} = 1 - \frac{1+\frac{2C}{\sqrt{n}}}{1+\frac{2c}{\sqrt{n}}} = \frac{\frac{2c}{\sqrt{n}}-\frac{2C}{\sqrt{N}}}{1+\frac{2c}{\sqrt{n}}} \geq \frac{\frac{2c}{\sqrt{n}}-\frac{2C}{\sqrt{N}}}{2} = \frac{c}{\sqrt{n}} - \frac{C}{\sqrt{N}}$$

If we choose N so that $\frac{c}{\sqrt{n}} \geq \frac{3C}{2\sqrt{N}}$ (which can be achieved by choosing any $N \geq \frac{9C^2 n}{4c^2}$) and also assume $C > 100$ (this is without loss of generality, as $C > 100$ corresponds to an easier GAPHAM problem), then $\delta \geq \frac{C}{2\sqrt{N}} \geq \frac{50}{\sqrt{N}}$. Thus we can conclude that when $u_i = -1$, $\Pr[\sum_{k=1}^{N} Z_k < \frac{N}{2}+C\sqrt{N}] \leq \exp(-N\delta^2/6) \leq \exp(-\frac{50^2}{N}N/6) \leq 0.1$ Similarly, we can also prove that when $u_i = +1$, $\Pr[\sum_{k=1}^{N} Z_k > \frac{N}{2}-C\sqrt{N}] \leq 0.1$ ■