LECTURE 5

Recall $[n] = \{1, 2, \ldots, n\}$

1

Last time, proved

For every integer $k \geq 1$, every $\lambda > 0$, every $\varepsilon > 0$, there is a randomized algorithm that given $A = (a_1, a_2, \ldots, a_m)$, $a_i \in [n]$,

- makes one pass over A
- computes Y such that
$$\Pr\left[\,|Y - F_k| > \lambda F_k\,\right] \leq \varepsilon$$
- uses $O\left(\dfrac{k \log(1/\varepsilon)}{\lambda^2}\, n^{1-1/k}\, (\log m + \log n)\right)$ bits of memory

Mentioned, but didn't prove, an even better result for F_2: can get rid of $\sqrt{n}$ factor and use only space
$$O(\log m + \log n)$$
(for fixed λ, ε).

Recall $m_i = |\{j \mid a_j = i\}|$

kth frequency moment:

$$F_k = \begin{cases} |\{ i \mid m_i \neq 0\}| & \text{if } k = 0 \\ \max_i m_i & \text{if } k = \infty \\ \|m\|_k^k = \sum_{i=1}^n m_i^k & \text{otherwise} \end{cases}$$

QUESTION: Can't we compute the exact answer with 100% probability?

Both <u>randomization</u> and <u>approximation</u> is needed

Given sequence A, its 1st frequency moment F_1 (ie., its length) can be computed exactly and deterministically in space $\log m$

However, for any $k \neq 1$, $k \geq 0$, in order to achieve close approximation of F_k (say, at most 10% off) with <u>deterministic</u> algorithm must use <u>linear</u> space

<u>Theorem</u> 1 [AMS99] For any integer $k \neq 1$, $k \geq 0$, any deterministic streaming algorithm that computes, given a sequence A, a number Y s.t. $|Y - F_k| \leq 0.1 \cdot F_k$ must use $\Omega(n)$ memory bits.

<u>Proof</u> Let $\mathcal{G}$ be a family of $t = 2^{\Omega(n)}$ subsets of $N = \{1, 2, \dots, n\}$ such that

(1) $\forall G \in \mathcal{G}$ $\quad |G| = n/64$

(2) $\forall G_1, G_2, \in \mathcal{G}$ $G_1 \neq G_2$ $\quad |G_1 \cap G_2| \leq n/128$

Claim such family $\mathcal{G}$ exists. (Accept in faith for now)

Let $EQ_{\mathcal{G}}$ be the set equality function restricted to inputs from $\mathcal{G}$

Same fooling set technique as in last lecture shows $D(EQ_{\mathcal{G}}) = \Omega(n) = \Omega(\log t)$.

Suppose $\exists$ deterministic streaming alg as in statement of Thm. Use alg to build communication protocol too good to be true.

Suppose Alice gets G_1; Bob gets G_2. Let $A(G_1, G_2)$ be datastream with all elements in G_1 followed by G_2.

Alice runs the streaming alg on G_1-part. Then sends memory contents to Bob
$o(n) = o(\log t)$ bits
Bob finishes algorithm and computes estimate of F_k. This estimate will allow Bob to decide whether $G_1 = G_2$ or not, which he announces to Alice. To see this, consider possible cases

	$G_1 = G_2$	$G_1 \neq G_2$
F_0	$= n/64$	$\geq 3n/128$ 50% higher
F_k $k \geq 2$	$= \frac{n}{64} \cdot 2^k$	$\leq \frac{n}{128} 2^k + \frac{n}{64}$ $= 0.75 \cdot \frac{n}{64} \cdot 2^k$ 25% lower

So this gives deterministic protocol for $EQ_{\mathcal{G}}$ with lower cost than $D(EQ_{\mathcal{G}})$ ... Contradiction

To see that there is such a family of sets $\mathcal{G}$, use a counting argument.

Also use known inequalities

$$\left(\frac{n}{s}\right)^s \leq \binom{n}{s} \leq \left(\frac{en}{s}\right)^s$$

[far from tight for the range we are considering, so constants in $\mathcal{G}$ can be optimized (and are better in [AMS99]). But this is good enough for us right now. where cardinality is $n/4$

The detailed argument is left for the next <u>problem set</u>.

So randomization is needed.
Also, even given randomization still cannot get exact answer

<u>Thm 2</u> [AMS '99] For any integer k, $k \neq 1$, $k \geq 0$, any randomized algorithm that computes, given A, a number Y s.t. $Y = F_k$ with probability $\geq 1 - \varepsilon$ for some fixed $\varepsilon < 1/2$ must use $\Omega(n)$ memory bits

Proof uses result by [Kalyanasundaram & Schnitger '87] that for any fixed $\varepsilon < 1/2$

$$R_\varepsilon(DISJ_n) = \Theta(n)$$

Identify (a) $x \in \{0,1\}^n$ and
(b) subset $\subseteq N = \{1,2,\dots,n\}$ with characteristic vector x

<u>Proof</u> Suppose $\exists$ streaming alg using $o(n)$ memory bits. Given x, y, let A be stream consisting of members of x followed by members of y. Alice runs algorithm on x and then sends over memory contents plus $|x|$ to Bob

$$o(n) \text{ bits} + O(\log n) \text{ bits}$$

Bob complets algorithm and computes $|x|+|y|$

For $k < \infty$ if $Y = |x| + |y|$ Bob concludes that $x \cap y = \emptyset$ and answers 1, else 0
For $k = \infty$, Bob answers 1 if $Y = 1$
<u>Next problem set</u>: Fill in details why this is correct!

So both randomization & approximation necessary.

But even given this, impossible to do well for F_∞

<u>Theorem 3</u> [AMS '99]

Any randomized streaming alg that outputs, given A, a number Y s.t.

$$\Pr\left[\,|Y - F_\infty| \geq F_\infty/3\right] < \varepsilon \quad \text{for some}$$

fixed $\varepsilon < 1/2$ must use $\Omega(n)$ memory bits.

<u>Proof</u> Same setup as for last theorem. Alice runs alg on her input, then sends snapshot of memory to Bob who completes alg on his input. If $Y < 4/3$ Bob outputs 1 (for $x \cap y = \emptyset$) otherwise 0.

If $x \cap y = \emptyset$, $F_\infty = 1$ and $Y < 4/3\, F_\infty = 4/3$ w prob $1-\varepsilon$. Otherwise $F_\infty \geq 2$ and $Y \geq 2/3\, F_\infty \geq 4/3$ w prob $1-\varepsilon$.

But unless alg uses $\Omega(n)$ memory, this gives a protocol contradicting

$$R_\varepsilon(DISJ_n) = \Theta(n).$$

<u>Note</u> that in fact LB holds for streaming alg with constant # passes over data (since CC lower bound holds in this more general model).

The strongest and most general bound in [AMS99] is as follows ~~next~~.
Won't be able to prove it, but will sketch some ingredients.

THM 4 [AMS 99]

For any fixed integer $k > 5$ and any $\varepsilon < 1/2$, any randomized algorithm that computes, given a sequence A, a number Y such that $\Pr[\,|Y - F_k| > 0.1\, F_k] < \varepsilon$ must use $\Omega(n^{1-5/k})$ memory bits.

Comments

Improved in [BJKS04] to

- "for any real $k > 2$"
- "space essentially $\Omega(n^{1-2/k})$"

Proved by reduction to communication cplx (no prizes for guessing)

But in new model of MULTI-PARTY COMMUNICATION COMPLEXITY

MULTIPARTY COMMUNICATION COMPLEXITY

t parties / players $P_1, \dots, P_t$
t inputs $x_1, \dots, x_t$
Want to compute prespecified function $f(x_1, x_2, \dots, x_t)$

Input models

(a) Number in hand (NIH)
Each P_i knows only x_i

(b) Number on forehead (NOF)
Each P_i knows all x_j except x_i

Communication models

(a) Blackboard or Broadcast
All messages are seen by all

(b) Message passing
P_i sends private message to P_j

(c) One-way
Each P_i sends exactly one private message to P_{i+1}

DEF 5 Unique disjointness problem $UDISJ_n(t)$ ~~$UDISJ(n,t)$~~
Input of each P_i is subset $x_i \subseteq \{1, 2, \dots, n\} = [n]$
t players

(1) Say $(x_1, x_2, \dots, x_t)$ disjoint if $\forall i \neq j \; x_i \cap x_j = \emptyset$
(2) Say $(x_1, \dots, x_t)$ uniquely intersecting if
$$\exists \ell \in [n] \text{ s.t. } \forall i \neq j \; x_i \cap x_j = \{\ell\}$$
Given $(x_1, \dots, x_t)$ guaranteed to be of type 1 or 2, determine which is the case!

Good streaming algorithm for F_k yields good NIH protocol for ~~UDISJ(n, t)~~ $UDISJ_n(t)$

Set $t = n^{1/k}$ Choose $T = \Theta(n^{1-1/k})$ s.t.

$n = (2T-1)t + 1$

Let all sets x_i have size T

Suppose algorithm uses M memory bits

Consider stream $A = x_1, x_2, \dots, x_t$ concatenated.

Each player P_i runs algorithm on own input x_i and then broadcasts content of memory to P_{i+1} (gives protocol for all communication models (a), (b), (c))

Total communication $\leq tM$ bits

Consider output Y of algorithm.

If $Y < 1.1n$, say, decide that $(x_1, \dots, x_t)$ disjoint, else uniquely intersecting.

<u>If disjoint</u>: $F_k = t \cdot T < n$

<u>If intersecting</u> $F_k = t^k + t(T-1)$

$$= n + t(T-1)$$

$$\gtrsim \frac{3}{2} n$$

Hence $tM \geq$ lower bound for ~~UDISJ(n, t)~~ $UDISJ_n(t)$

[AMS99]

Randomized communication cplx in NIH broadcast model

$$\geq \Omega(n/t^4)$$

$$\Rightarrow \text{space } \mathcal{M} \geq \Omega(n/t^5) = \Omega(n^{1-5/k})$$

Holds for CONSTANT # passes

[BJKS04]

One-way LB for $UDISJ_n(t)$ $\Omega(n/t^{1+\gamma})$ for any $\gamma > 0$

$$\Rightarrow \text{space } \mathcal{M} \geq \Omega(n/t^{2+\gamma}) \text{ for any } \gamma > 0$$

$$= \tilde{\Omega}(n/t^2) = \tilde{\Omega}(n^{1-2/k})$$

Holds for ONE pass

Also get

$$\text{space } \mathcal{M} \geq \tilde{\Omega}(n/t^3) = \tilde{\Omega}(n^{1-3/k})$$

for CONSTANT # passes

[Gronemeier 09]

Improved to $\tilde{\Omega}(n/t^2) = \tilde{\Omega}(n^{1-2/k})$

for CONSTANT # passes

[Indyk & Woodruff '05] showed that this lower bound is <u>optimal</u> up to lower order terms

Give algorithm for any real $k > 2$ that uses $\tilde{O}(n^{1-2/k})$ space.

Briefly about ideas in [AMS99]

Equality
Randomized comm cplx
= distributional cplx for worst μ

holds also in multi-player setting

For deterministic protocol, consider multidimensional rectangles (boxes)

To get μ, first randomly partition $[n]$ into $I_1 \dot\cup I_2 \dot\cup \dots \dot\cup I_t \cup \{\ell\}$

$|I_j| = 2T-1$

Pick $A_j \subseteq I_j$, $|A_j| = t$ randomly for $j=1,\dots,t$

~~Wp~~ With prob $1/2$, let

$x_j = A_j$ for all j disjoint

With prob $1/2$, let

$x_j = (I_j \setminus A_j) \cup \{\ell\}$ uniquely intersecting

Prove that if # boxes is too small, then too many inputs of the two distinct types will end up in the same boxes
$\implies$ too large error

[Omitting all the details, of course...]