

## Lecture 11

Lecturer: Michael Kapralov

1  $\Omega(n)$  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*. The following proof is due to [2].

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. For simplifying the proof, we modify Alice's vector to  $u \in \{-1, +1\}^n$ . Also *GAPHAM* problem is defined as, given two vector  $x, y \in \{-1, +1\}^n$ , we want to distinguish whether  $\Delta(x, y) \leq \frac{n}{2} - C\sqrt{n}$  or  $\Delta(x, y) \geq \frac{n}{2} + C\sqrt{n}$ , 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 follows,

- (1) Pick  $N$  i.i.d. vector  $r^1, r^2, \dots, r^N$  where for all  $k \in [N]$ ,  $r^k \sim \text{UNIF}(\{-1, +1\}^n)$
- (2) For each  $k = 1 \dots N$ , 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 vector  $x, y \in \{-1, +1\}^N$  into *GAPHAM* solver. Output  $u_i = -1$  if the *GAPHAM* solver recognizes that  $\Delta(x, y) \geq \frac{n}{2} + C\sqrt{n}$ , otherwise output  $u_i = +1$  if  $\Delta(x, y) \leq \frac{n}{2} - C\sqrt{n}$

Note that,

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

The sketch of this method is to produce a random bit for Alice and Bob without interaction and guarantee that if  $u_j$  is -1, the bit will differ with probability at least  $\frac{1}{2} + \frac{c}{\sqrt{n}}$  and if  $u_j$  is 1, the bit will differ with probability at most  $\frac{1}{2} - \frac{c}{\sqrt{n}}$ . Then repeat this procedure  $N$  times ( $N$  will be specified latter) to make sure that hamming distance either at least  $\frac{n}{2} + C\sqrt{n}$  or at most  $\frac{n}{2} - C\sqrt{n}$  with high probability, which can be proved by Chernoff Bound. We formalize the proof,

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

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

where  $c$  is a positive constant.

**Proof** Assume without loss of generality that  $n$  is odd.  $\langle u, r \rangle = \sum_{j=1}^n u_j r_j = u_i r_i + \sum_{j \neq i}^n u_j r_j$ . Denote  $w = \sum_{j \neq i}^n u_j r_j$ , there are two cases to consider when  $u_i = -1$

- Case 1  $w \neq 0$ , then  $|w| \geq 2$  for  $|w|$  is even. Then we can obtain  $\text{sgn}(\langle u, r \rangle) = \text{sgn}(w)$ , which implies that  $\Pr[\text{sgn}(\langle u, r \rangle) = -1] = \Pr[\text{sgn}(\langle u, r \rangle) = 1] = \frac{1}{2}$ . Thus  $\Pr[\text{sgn}(\langle u, r \rangle) \neq \text{sgn}(\langle e_i, r \rangle)] = \frac{1}{2}$ .
- Case 2  $w = 0$ , then  $\text{sgn}(\langle u, r \rangle) = u_i r_i$ . Thus  $\Pr[\text{sgn}(\langle u, r \rangle) \neq \text{sgn}(\langle e_i, r \rangle)] = 1$ .

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  $c' > 0$ ,  $\Pr[w = 0] \geq \frac{c'}{\sqrt{n}}$  (Another proof is that the distribution of  $w$  is coverage to a Gaussian distribution with variance  $\sqrt{n}$ , thus the pdf of this distribution between  $-\sqrt{n}$  and  $\sqrt{n}$  is  $\Omega(\sqrt{n})$ ). Letting  $c = \frac{c'}{2}$ , we can obtain the following result, when  $u_i = -1$ ,  $\Pr[\text{sgn}(\langle u, r \rangle) \neq \text{sgn}(\langle e_i, r \rangle)] = \Pr[w = 0] + \frac{1}{2}(1 - \Pr[w = 0]) \geq \frac{1}{2} + \frac{c'}{2\sqrt{n}} = \frac{1}{2} + \frac{c}{\sqrt{n}}$ . ■

To boost this probability, we pick  $N$  i.i.d vectors, and denote

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

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

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

**Proof** By Chernoff's 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  $\delta$  is chosen so that  $(1 - \delta) \sum_{k=1}^N \mathbb{E}[Z_k] = \frac{N}{2} + C\sqrt{N}$ . We now lower bound  $\delta$ . Since  $\sum_{k=1}^N \mathbb{E}[Z_k] \geq N/(2 + c/\sqrt{n})$ , 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  $u_i = +1$ ,  $\Pr[\sum_{k=1}^N Z_k > \frac{N}{2} - C\sqrt{N}] \leq 0.1$  ■

## 2 Lower bound for approximating maximum matchings in graph streams

We will prove

**Theorem 3** *Let  $ALG$  be a single pass streaming algorithm that for some constant  $\delta > 0$  outputs a  $(2/3 + \delta)$ -approximation to the maximum matching in an input graph  $G = (V, E)$ ,  $|V| = n$  presented as a stream of edges and succeeds with some constant probability. Then  $ALG$  must use  $n^{1+\Omega(1/\log \log n)} \gg n \log^{O(1)} n$  bits of space.*

We will use

**Definition 4** *A bipartite graph  $G = (P, Q, E)$ ,  $|P| = |Q| = n$  is an  $(\epsilon, k, n)$ -Ruzsa-Szemerédi graph if the edge set of  $G$  can be expressed as a union of  $k$  induced matchings of size  $\epsilon n$ , i.e.  $E = \bigcup_{i=1}^k M_i$ , where  $M_i$  is matching between subsets  $A_i \subseteq P$  and  $B_i \subseteq Q$  with  $|A_i| = |B_i| = \epsilon n$ , and the subgraph of  $G$  induced by  $A_i \cup B_i$  is  $M_i$ .*

and

**Lemma 5** [1] *For every  $\delta \in (0, 1)$  there exists an  $(\frac{1}{2} - \delta, k, n)$ -Ruzsa-Szemerédi graph  $G = (P, Q, E)$ ,  $E = \bigcup_{i=1}^k M_i$ , with  $k = n^{1+\Omega_\delta(1/\log \log n)}$ .*

In what follows we prove the lower bound assuming Lemma 5.

**Construction of a hard instance.** Let  $G = (P, Q, E)$  be a  $(1/2 - \delta/10)$ -RS graph, where  $\delta > 0$  is the constant advantage over  $2/3$  approximation that we would like to rule out. Let  $M_i = (A_i, B_i, E_i)$  denote the matchings that form the edges of  $G$ . For each  $i = 1, \dots, k$  let  $X^i \in \{0, 1\}^{M_i}$ , and let  $X = \bigcup_{i=1}^k X^i$ . Let  $X_e = 1$  independently with probability  $1 - \delta/10$  and 0 otherwise. Let  $G'$  contain every edge  $e \in E$  of  $G$  such that  $X_e = 1$ , and let  $M'_i$  denote the corresponding induced matchings. For every  $i \in [k]$  let  $G'_i$  denote the graph obtained from  $G'_i$  by adding two new sets  $S$  and  $T$  together with a perfect matching from  $S$  to  $P \setminus A_i$  and from  $T$  to  $Q \setminus B_i$ .

The following claim follows easily from Chernoff bounds:

**Claim 6** *The graph  $G'_i$  contains a matching of size at least  $(1 - \delta/5)(3/2)n$  for every  $i \in [k]$ ,  $k \leq n$  with probability at least  $1 - e^{-\Omega_\delta(n)}$ .*

Denote the success event from Claim 6 by  $\mathcal{E}_{\text{large-matching}}$ . We also have

**Claim 7** *For every matching  $\widehat{M}$  in  $G'_i$  one has*

$$|\widehat{M}| \leq |P \setminus A_i| + |Q \setminus B_i| + |\widehat{M} \cap M'_i|.$$

**Proof** This follows by the max-flow/min-cut theorem after attaching a source  $s$  with a directed edge to every vertex in  $Q$ , and a sink  $t$  with a directed edge from every vertex in  $P$ , and directing all edges of  $G$  to go from  $Q$  to  $P$ . Indeed, consider the cut with  $\{s\} \cup S \cup (P \setminus A_i) \cup B_i$  on one size and  $\{t\} \cup T \cup (Q \setminus B_i) \cup A_i$  on the other side. There are  $|P \setminus A_i| + |Q \setminus B_i|$  edges that cross the cut and are incident on either  $s$  or  $t$  (these are accounted for by the first two terms on the rhs), and the only edges of  $G$  that cross the cut are the edges that go from  $B_i$

to  $A_i$ . The latter set is exactly the set of edges of  $M'_i$  by the induced property of matchings in Ruzsa-Szemerédi graphs, yielding the  $|\widehat{M} \cap M'_i|$  term.  $\blacksquare$

We now proceed to prove Theorem 3. Let  $\Pi$  denote the state of the memory of a possibly randomized algorithm that on every input with probability at least  $1/3$  outputs a matching  $\widehat{M}$  such that  $\widehat{M} \subseteq E$  and  $|\widehat{M}| \geq (2/3 + \delta)|M_{OPT}|$ , where  $M_{OPT}$  is the maximum matching in the input graph.

By Claim 7 we have

$$|\widehat{M}| \leq |P \setminus A_i| + |Q \setminus B_i| + |\widehat{M} \cap M'_i| \leq \left(\frac{1}{2} + \delta/10\right) 2n + |\widehat{M} \cap M'_i|.$$

Thus, since by Claim 6 the graph  $G'_i$  contains a matching of size at least  $(1 - \delta/5)(3/2)n$  with probability at least  $9/10$  if  $n$  is large enough, it must be that

$$(2/3 + \delta)(1 - \delta/5)(3/2)n \leq \left(\frac{1}{2} + \delta/10\right) 2n + |\widehat{M} \cap M'_i|.$$

This in particular implies that

$$\begin{aligned} |\widehat{M} \cap M'_i| &\geq (2/3 + \delta)(1 - \delta/5)(3/2)n - \left(\frac{1}{2} + \delta/10\right) 2n \\ &\geq [(1 + \delta)(1 - \delta/5) - (1 + \delta/10)]n \\ &\geq [(1 + \delta - \delta/5) - (1 + \delta/10)]n \\ &\geq (\delta/2)n. \end{aligned} \tag{1}$$

Let  $E_i$  be a binary variable that equals 1 if the algorithm is not correct on the graph  $G'_i$  or if the maximum matching size in  $G'_i$  is below  $(1 - \delta/5)(3/2)n$  and 0 otherwise. By Claim 6 and the assumption on correctness of ALG we have

$$\mathbf{Prob}[E_i = 1] \leq 2/3 + e^{-\Omega_\delta(n)} \leq 3/4. \tag{2}$$

We have

$$\begin{aligned} H(X|\Pi) &= \sum_{i=1}^k H(X_i|\Pi, X_{<i}) \\ &\leq \sum_{i=1}^k H(X_i, E_i|\Pi, X_{<i}) \\ &= \sum_{i=1}^k H(E_i|\Pi, X_{<i}) + H(X_i|\Pi, X_{<i}, E_i) \\ &= \sum_{i=1}^k (1 + H(X_i|\Pi, X_{<i}, E_i)) \end{aligned} \tag{3}$$

We now upper bound  $H(X_i|\Pi, X_{<i}, E_i)$ . Note that if  $E_i = 0$ , then by (1) one has

$$|\widehat{M} \cap M'_i| \geq (\delta/2)n. \tag{4}$$

For every  $e \in M_i$  such that  $e \in \widehat{M}$  we know that if  $E_i = 0$  (i.e. the algorithm is

correct) then  $X_e = 1$  (the edge is present in the graph). We thus have

$$\begin{aligned}
& H(X_i|\Pi, X_{<i}, E_i) \\
&= H(X_i|\Pi, X_{<i}, E_i = 1)\mathbf{Prob}[E_i = 1] + H(X_i|\Pi, X_{<i}, E_i = 0)\mathbf{Prob}[E_i = 0] \\
&= H(X^i)\mathbf{Prob}[E_i = 1] + H(X^i|\Pi, X_{<i}, E_i = 0)\mathbf{Prob}[E_i = 0] \\
&\leq H(X^i)\mathbf{Prob}[E_i = 1] + \sum_{e \in M_i} H(X_e^i|\Pi, X_{<i}, E_i = 0)\mathbf{Prob}[E_i = 0] \quad (\text{by subadditivity of entropy}) \\
&\leq H(X^i)\mathbf{Prob}[E_i = 1] + \sum_{e \in M_i \setminus \widehat{M}} H(X_e^i|\Pi, X_{<i}, E_i = 0)\mathbf{Prob}[E_i = 0] \quad (\text{since } X_e^i = 1 \text{ for all } e \in \widehat{M}) \\
&\leq H(X^i)\mathbf{Prob}[E_i = 1] + (|M_i| - (\delta/2)n)H(X_e^i)\mathbf{Prob}[E_i = 0] \quad (\text{by (4) and since conditioning reduces entropy}) \\
&\leq H(X^i)\mathbf{Prob}[E_i = 1] + (1 - \Omega(1))H(X^i)\mathbf{Prob}[E_i = 0] \\
&\leq (1 - \Omega(1))H(X^i) \quad (\text{since } \mathbf{Prob}[E_i] \text{ is larger than a constant by (2)})
\end{aligned}$$

Putting this together with (5), we get

$$\begin{aligned}
H(X|\Pi) &\leq \sum_{i=1}^k (1 + (1 - \Omega(1))H(X_i)) \\
&\leq \sum_{i=1}^k (1 - \Omega(1))H(X_i) \quad (\text{for sufficiently large } n)
\end{aligned} \tag{5}$$

Thus, we get that

$$H(X|\Pi) \leq (1 - \Omega(1))H(X),$$

implying that

$$H(\Pi) \geq I(X; \Pi) = H(X) - H(X|\Pi) = \Omega(1)H(X),$$

and thus message length must be  $n^{1+\Omega(1/\log \log n)}$  bits for any constant  $\delta > 0$ , as required.

## References

- [1] A. Goel, M. Kapralov, and S. Khanna. On the communication and streaming complexity of maximum bipartite matching. *SODA*, 2012.
- [2] Thathachar S Jayram, Ravi Kumar, and D Sivakumar. The one-way communication complexity of hamming distance. *Theory of Computing*, 4(1):129–135, 2008.