CS 15-851: Algorithms for Big Data Spring 2024

Lecture 9 — Mar 21

Prof. David Woodruff *Scribe: Dolev Artzi*

1 Aspects of 1-Way Communication of *INDEX*

Given:

- Alice has $x \in \{0,1\}^n$
- Bob has $i \in [n]$
- Alice sends a randomized message M to Bob
-

$$\begin{aligned} I(M;X|R) &= \sum_i I(M;X_i \mid X_{<i},R) \qquad (1)\\ &\geq \sum_i I(M;X_i \mid R) \qquad \text{(expand summand } (*))\\ &= n - \sum_i H(X_i \mid M,R) \qquad (2)\end{aligned}$$

From Fano's inequality, we know $H(X_i \mid M,R) \leq H(\delta)$. If Bob can guess X_i with probability, then

$$CC_\delta(\mathbf{INDEX}) \geq I(M;X \mid R) \geq n(1-H(\delta))$$

1.1 Dealing With Partial Correctness Guarantees

If we have the following correctness guarantee, for example, over uniformly random X, i for every input pair, we can we can average over input pairs and still have that guarantee:

$$P[\text{Bob outputs } X_i] \geq \frac{9}{10}$$

The same lower bound applies even if the protocol is only correct on average over x and i drawn independently from a uniform distribution. How does the proof change? Instead of Fano's holding for all i, it should hold for a large fraction of i, or a *typical* i. Say, for 4/5ths of the i, the failure probability should be at most 2δ. The intuition is that we have a *small failure probability for a large fractions of the* ***i*****s**. Apply this proof to those ***i***s for which we have the correctness guarantee. The key point is that for $\Omega(n)$ different i, we can still apply the bound.

1.2 Removing Randomness

Note that wlog, we can have have the players act deterministically, by viewing the randomized algorithm as a *distribution over deterministic algorithms*, i.e. there must be some fixing of the randomness that achieves the guarantee. Fixing the randomness, you can prove lower bounds for the deterministic protocol, then argue that the same lower bounds hold once you introduce randomness.

2 Distributional Communication Complexity

- $(X, y) \sim \mu$
- **μ-distributional complexity** $D_\mu(f)$: the minimum communication cost of a protocol which outputs $f(X,Y)$ wp. $\geq 2/3$ for $(X,Y) \sim \mu$
- Yao's Minimax principle: $R(f) = \max_\mu D_\mu(f)$

Theorem 1 ([KNR99])**.** *Indexing is Universal for Product Distribution.*

- Communication matrix A_f of a boolean function $f : X \times Y \to \{0,1\}$ has (x,y)-th entry equal to $f(x,y)$. Alice has X, Bob has Y. Rows correspond to inputs X, col. to inputs Y in the joint (communication) matrix
- $$\max_{\text{product } \mu} D_\mu(f) = \theta(VC \text{ dimension of } A_f)$$
- Product μ means $\exists \mu_1, \mu_2$ such that $\mu(x,y) = \mu_1(x)\mu_2(y)$. VC dimension means "richness". $f(x,y) \in \{0,1\}$. VC dimension d means I can choose $2d$ columns where all row block patterns are represented to "see the whole hypercube". Embed the **INDEX** problem on that hypercube.
- **In INDEX**: In a 2^n by n matrix where every single n pattern occurs, VC-dim is n in that case. Now, we have a $2^d \times d$ communication matrix. Embed product (uniform for x, i) dist. into the communication matrix, and the **Theorem** says that gives you the *optimal lower bound you can get for product distributions.*

3 Indexing with Low Error

- Index problem with 1/3 error probability and 0 error probability both have $\Omega(n)$ communication
- Sometimes want lower bounds in terms of error probability

3.1 Indexing on Large Alphabets

- Alice has $x \in \{0,1\}^{n/\delta}$ with $wt(x) = n$, Bob has $i \in [n/\delta]$
- Bob wants to decide if $x_i = 1$ with error probability δ

- 1-Way communication is $\Omega(n \log(\frac{1}{\delta}))$ [JW21]
- Better encoding of x: n nonzero locations, $\binom{n/\delta}{n}$ possibilities. $\log_2 \binom{n/\delta}{n} \leq \log_2 (e/\delta)^n \leq O(n \log \frac{1}{\delta})$. This is optimal by [Jayram, W]
- This can be used to get a $\Omega(\log(\frac{1}{\delta}))$ bound for norm estimation. Recall: l_2 estimation: streaming space complexity with $x \in poly(n)$, want $1 \pm \epsilon$ approx, want to succeed wp. $\geq 1 - \delta$ with space complexity (Sx where S is CountSketch) $\frac{1}{\epsilon^2} \log n \log \frac{1}{\delta}$. The log factor comes from **AUGMENTED INDEX**, the $\frac{1}{\epsilon^2}$ from **Gap-Hamming**, and the last part from **Indexing With Large Alphabets**. *The upper bound is their product, the lower bound is their sum.*

4 Beyond Product Distributions

4.1 Non-Product Dists.

- needed for stronger lower bounds on norm estimation, especially for $p > 2$
- estimating $|\cdot|_\infty$ up to a multiplicative factor of B in a stream.

4.2 Gap-∞ Problem

- Alice has $x \in \{0, ..., B\}^n$
- Bob has $y \in \{0, ..., B\}^n$
- Promise: $|x - y|_\infty \leq 1$ or $|x - y|_\infty \geq B$
- Lowerbound: $\Omega(n/B^2)$ [SS02, BJKS04]

4.3 Example

Bob has an index, $y = -e_i$. If we solve Gap-∞ for $B = 2$, can solve **INDEX**. $x - y = x + e_i$: if $x_i = 1$: $|x - y|_\infty = 2$. Else, its 1. To get context for the lower bound, let's think of some possible upper bounds to this problem. l_p-norm estimation is $n^{1-2/p} poly(\log n)$ based on an Exponential rvs-based algorithm from a previous lecture. This allows you to estimate $|x|_p$ (as well as $|x|_p^p$)up to a constant factor. If you output Z s.t. $|x|_p \leq Z \leq C|x|_p$, then $|x|_p^p \leq Z^p \leq C^p |x|_p^p$, where C, p are constant.

Case 1: $|x - y|_\infty \leq 1$ then $|x|_p^p \leq n$

Case 2: $|x - y|_\infty \geq B$ then $|x|_p \geq B^p$, i.e we have separation.

Can distinguish between them if $B^p > 10 \cdot n$. Suppose $B = (10n)^{1/p}$. Estimate the p norm to the p of $x - y$. In one case its at most $10n$, in the other case at most n. The complexity of a p-norm algo is the runtime for the Gap-∞ problem. Send the streaming algorithm. $n^{1-2/p}$ is also $O(\frac{n}{B^2}) poly(\log n)$.

Alice computes Sx and sends it to Bob, who computes Sy, to have $S(x-y)$. When $B > 10\sqrt{n}$, use $p = 2$. when p gets very close to 1 you can't do any better than $\Omega(n)$ anyways, so in this case you can send the whole vector.

References

[BJKS04] Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar, and D. Sivakumar. An information statistics approach to data stream and communication complexity. *J. Comput. Syst. Sci.*, 68(4):702–732, 2004.

[JW21] Rajesh Jayaram and David P. Woodruff. Perfect $\mathrm{l_p}$ sampling in a data stream. *SIAM J. Comput.*, 50(2):382–439, 2021.

[KNR99] Ilan Kremer, Noam Nisan, and Dana Ron. On randomized one-round communication complexity. *Comput. Complex.*, 8(1):21–49, 1999.

[SS02] Michael Saks and Xiaodong Sun. Space lower bounds for distance approximation in the data stream model. In *Proceedings of the Thiry-Fourth Annual ACM Symposium on Theory of Computing*, STOC '02, pages 360–369, 2002.