Sublinear Time and Space Algorithms 2020B – Lecture 14 Communication Complexity and Streaming Lower Bounds*

Robert Krauthgamer

1 Indexing

Problem definition: Alice has input $x \in \{0,1\}^n$ and Bob has as input an index $i \in [n]$. Their goal is to output $INDEX(x,i) = x_i$.

This function would be easy if Bob could send his (short) input to Alice. But we shall consider one-way communication from Alice to Bob, and her input is much longer.

Theorem 1 [Kremer, Nisan, and Ron, 1999]: The randomized one-way communication complexity of indexing is $\Omega(n)$, even with shared randomness.

It's therefore a "canonical" problem for reductions (in this model).

Proof by [Jayram, Kumar and Sivakumar, 2008]: Was seen in class (using an error correcting code and some averaging arguments).

2 Streaming Lower Bounds: Exact ℓ_0

Theorem 2: Every streaming algorithm for computing ℓ_0 exactly in $\mathbb{R}^n$, even a randomized one with error probability $1/6$, requires storage of $\Omega(n)$ bits.

Remark: This is true even for insertions-only streams.

Proof: Was seen in class, by reduction from the indexing problem.

Remark: Notice that our proof works even if random coins are not counted in the storage of the streaming algorithm (because we rely on a communication lower bound with public coins).

Exer: Show a similar lower bound for exact ℓ_1.

Hint: You obviously must use a stream with deletions.

---

*These notes summarize the material covered in class, usually skipping proofs, details, examples and so forth, and possibly adding some remarks, or pointers. The exercises are for self-practice and need not be handed in. In the interest of brevity, most references and credits were omitted.

Exer: Prove that every streaming algorithm for graph connectivity on n vertices (i.e., deciding whether a stream of edge-insertions gives a connected graph), even a randomized one with error probability 1/3, requires storage of $\Omega(n)$ bits.

3 Gap Hamming Distance (GHD)

Problem definition: Alice and Bob's inputs are $x, y \in \{0,1\}^n$, respectively, and their goal is to determine whether the hamming distance between x, y is $\leq \frac{n}{2} - \sqrt{n}$ or $\geq \frac{n}{2} + \sqrt{n}$.

Theorem 3 [Woodruff, 2004]: The randomized one-way communication complexity of GHD is $\Omega(n)$, even with shared randomness.

We skipped the proof of this theorem (For those interested, look for a proof by reduction from Indexing by [Jayram, Kumar and Sivakumar, 2008]).

We mention in passing a stronger result, where the number of rounds is unbounded.

Theorem [Chakrabarti and Regev, 2011]: The communication complexity (with unbounded number of rounds) of GHD is $\Omega(n)$, even with shared randomness.

4 Streaming Lower Bounds: Approximate ℓ_0

Theorem 4: Every streaming algorithm that $(1+\varepsilon)$-approximates ℓ_0 in $\mathbb{R}^n$ for $1/\sqrt{n} \leq \varepsilon < 1$, even a randomized one with error probability 1/6, requires storage of $\Omega(1/\varepsilon^2)$ bits.

Remark: For smaller $0 < \varepsilon < 1/\sqrt{n}$, the required storage is $\Omega(n)$; to see this, observe that an algorithm for such "smaller" ε "solves" $\varepsilon = 1/\sqrt{n}$ which is covered by the above theorem.

We skipped the proof of this theorem (for those interested, it is by reduction from GHD).

5 Current Research Directions

We concluded with a brief mention of research topics related to the course.

Streaming matrices: Different update models, different problems

Streaming (and sampling) edit distance: Different models of the input

Distributed monitoring: Continuously maintain an approximation to data residing in k sites with little communication

Fast algorithms: in classic sense, like near-linear time