Sublinear Time and Space Algorithms 2018B – Lecture 14
More Streaming Lower Bounds*

Robert Krauthgamer

1 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 $\le \frac{n}{2} - \sqrt{n}$ or $\ge \frac{n}{2} + \sqrt{n}$.

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

Proof from [Jayram, Kumar and Sivakumar, 2008]: Was seen in class, by reduction from the Indexing problem.

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.

2 Streaming Lower Bounds: Approximate ℓ_0

Theorem 4: Every streaming algorithm that $(1+\varepsilon)$-approximates ℓ_0 in $\mathbb{R}^n$ for $1/\sqrt{n} \le \varepsilon < 1$, 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 $\Omega(n)$, because any algorithm for such "smaller" ε "solves" $\varepsilon = 1/\sqrt{n}$ which is covered by the above theorem.

Proof: Was seen in class, by reduction from GHD.

Exer: Prove the same bound for insertions-only streams.

Hint: Observe that $2\|x+y\|_0 = \|x\|_0 + \|y\|_0 + \|x-y\|_0$ for all $x, y \in \{0,1\}^n$.

Exer: Show a similar lower bound for $(1+\varepsilon)$-approximation of ℓ_1-norm and ℓ_2-norm.

*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.

3 Set Disjointness and Approximating ℓ_∞-norm

Problem definition: The inputs are $x, y \in \{0,1\}^n$ and the goal is to determine whether the cardinality of $\{i \in [n] : x_i = y_i = 1\}$ is one or zero.

We can view x, y as subsets of $[n]$, and the goal is to decide if the two sets intersect (exactly once) or are disjoint. This is sometimes called the unique intersection property.

Theorem 5 [Kalyanasundaram and Schnitger, 1992] and [Razborov, 1992]: The communication complexity (with unbounded number of rounds) of Set Disjointness in $\{0,1\}^n$ is $\Omega(n)$, even with shared randomness.

Stated without proof.

Corollary 6: Every randomized streaming algorithm that approximates ℓ_∞-norm in $\mathbb{R}^n$ within factor 2.99 requires $\Omega(n)$ bits.

Proof: We sketched in class a lower bound for 1.99-approximation that holds even for insertion-only stream.

Exer: Improve the approximation factor to 2.99, by using negative entries in the input vector (deletions in the stream).

Exer: Extend the above lower bound to p passes over the input.

4 Multiparty Disjointness and ℓ_p-norm

Problem definition: There are t players, with respective inputs $x^{(1)}, \ldots, x^{(t)} \in \{0,1\}^n$ and the goal is to determine whether

- for all $i \neq j$, $\{i \in [n] : x^{(i)} = x^{(j)} = 1\} = \emptyset$; or
- there is $k \in [n]$ such that for all $i \neq j$, $\{i \in [n] : x_i \wedge y_i = 1\} = \{k\}$.

(It may be easier to think of it as set intersection $\left|x^{(i)} \wedge x^{(j)}\right|$.)

We usually consider the model where all messages are written on a blackboard that is seen by all players (equivalently, it is broadcasted to all players without counting it n times).

Theorem 7 [Gronemeier, 2009], following [Bar-Yossef, Jayram, Kumar and Sivakumar, 2002] and [Chakrabarti, Khot and Sun, 2003]: The communication complexity (with unbounded number of rounds) of t-party Set Disjointness in $\{0,1\}^n$ is $\Omega(n/t)$, even with shared randomness.

Stated without proof.

Remarks:
(a) It follows that at least one player has to send $\Omega(n/t^2)$ bits.
(b) The bound holds even in the one-way model, where the messages go first from Player 1 to 2, then from Player 2 to 3, and so forth.

Corollary 8: Every streaming algorithm that 2-approximates the ℓ_p-norm, for $p > 2$, in $\mathbb{R}^n$, requires $\Omega(n^{1-2/p})$ bits of storage.

Remark: Holds even for insertions-only streams.

Proof: Was sketched in class.

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

Massively parallel architectures (e.g., Map-Reduce): Often use techniques from streaming algorithm

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

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