Problem Set 1

Sublinear Algorithms

Due Thursday, September 15

1. Lower bounds for distinct elements in the insertion-only model.

   (a) Using a reduction from the indexing problem, show that any streaming algorithm achieving $\epsilon = 0$ and $\delta = 1/10$ must take $\Omega(k)$ space.

   (b) In the *gap Hamming* problem, Alice and Bob are respectively given vectors $x, y \in \{0,1\}^m$. It is guaranteed that the Hamming distance between x and y is either at least $m/2 + \sqrt{m}$ or at most $m/2 - \sqrt{m}$. Alice and Bob must communicate to figure out which situation they are in, i.e., at the end of the communication Bob should output 1 if $\|x - y\|_1 > m/2 + \sqrt{m}$ and 0 if $\|x - y\|_1 < m/2 - \sqrt{m}$, with probability 3/4. It is known that gap Hamming requires $\Omega(m)$ communication.

   Using a reduction from the gap Hamming problem, show that any streaming algorithm for distinct elements with $\delta = 1/10$ must take $\Omega(1/\epsilon^2)$ space.

2. For a data stream received under the non-strict turnstile model, let $x \in \mathbb{Z}^n$ be the final histogram. You may assume that each entry in x is an $O(\log n)$ bit integer.

   (a) Show how to determine whether x is the all-zero vector, using $\mathrm{poly}(\log(n/\delta))$ space. Bonus: use $O(\log(n/\delta))$ space.

   (b) Give an algorithm that detects if x has a single non-zero entry, and if so finds that location. That is, give an algorithm to compute

   $$\begin{cases} i & \text{if } x = e_i \\ \perp & \text{otherwise} \end{cases}$$

   with probability $1 - \delta$. Hint: first identify whether x has zero, one, or at least two non-zero entries.

3. **[Deferred till problem set 2]** Recall the AMS sketch from class for $\|\cdot\|_2$ estimation: a random $m \times n$ matrix A with entries $A_{ij} \in \{\pm 1/\sqrt{m}\}$ is drawn for $m = O(1/\epsilon^2)$, and $\|x\|_2^2$ is estimated as $\|Ax\|_2^2$. With at least 3/4 probability, we had

   $$(1 - \epsilon)\|x\|_2^2 \le \|Ax\|_2^2 \le (1 + \epsilon)\|x\|_2^2. \tag{1}$$

   (a) Consider the following matrix instead: for each $i \in [n]$, let the ith column of A have a single ± 1 in a random row, and 0s elsewhere. Because this matrix is sparse, it can be maintained under turnstile updates in *constant* time. Show that this A still satisfies (1) with 3/4 probability for $m = O(1/\epsilon^2)$.

(b) Show how to generate A using only $O(\log n)$ bits of randomness.

4. Let $x_1, \ldots, x_n \sim N(0,1)$. Define
$$z = \max_{i \in [n]} x_i.$$
   (a) Prove that $\mathrm{E}[z] = \Theta(\sqrt{\log n})$.
   (b) What if the x_i were instead subgaussian with parameter $\sigma = 1$? What would be the bounds on $\mathrm{E}[z]$ then?

5. Let X_1 and X_2 be mean zero subgaussian random variables with parameters σ_1 and σ_2 respectively.
   (a) Show that $X_1 + X_2$ is subgaussian with parameter $2\max(\sigma_1, \sigma_2)$, regardless of whether X_1 and X_2 are independent.
   (b) If X_1 and X_2 are independent, show that $X_1 X_2$ is subexponential and specify the parameters in terms of σ_1 and σ_2.