Problem Set 1

Sublinear Algorithms

Due Tuesday, September 23

1. 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?

2. Let X_1 and X_2 be subgaussian random variables with parameters σ_1 and σ_2 respectively.
   (a) Show that $X_1 + X_2$ is subgaussian with parameter $2\sqrt{\sigma_1^2 + \sigma_2^2}$, 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.

3. Recall that the various algorithms for distinct elements take $\mathrm{poly}(\log n, 1/\epsilon, \log(1/\delta))$ samples to achieve a multiplicative $1 \pm \epsilon$ approximation to the number of distinct elements in the stream with probability $1 - \delta$. Is this dependence on ϵ and δ necessary?
   (a) Show that any streaming algorithm achieving $\epsilon = 0$ and $\delta = 1/10$ must take $\Omega(n)$ space.
   (b) Show that any streaming algorithm achieving $\epsilon = 1/10$ and $\delta = 0$ must take $\Omega(n)$ space.
   (c) (Optional) Show that the dependence must be at least $\mathrm{poly}(1/\epsilon + \log(1/\delta))$.

4. 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 estimates as $\|Ax\|_2^2$. With at least 3/4 probability, we had

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

   (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.

5. Recall the algorithm described in class for testing whether a distribution is uniform: count the fraction of collisions A in the samples $x_1, \ldots, x_m$, and determine whether it is above or below $1/n + \epsilon^2/(2n)$. We showed using Chebyshev's inequality that after $O(\sqrt{n}/\epsilon^4)$ samples, the observed value of A would probably be within a $1 + \epsilon^2/2$ multiplicative factor of its expected value. We also showed that this suffices for the tester to distinguish uniform distributions from those ϵ-far from uniform.

   In this problem, we try to determine whether the dependence is tight. In particular, we know that Pearson's chi-squared test takes $\Theta(n/\epsilon^2)$ samples, so we would like to get a tester with something like a $\Theta(\sqrt{n}/\epsilon^2)$ dependence.

   (a) Show a probability distribution that is ϵ-far from uniform and for which $\Omega(\sqrt{n}/\epsilon^4)$ samples are required for A to be typically a $1+\epsilon^2$ multiplicative approximation of its expected value.

   (b) How many samples does the tester actually take on this distribution to distinguish it from uniform?

   (c) (Optional) If the answer to (b) was less than $\Omega(\sqrt{n}/\epsilon^4)$, find a distribution for which the tester requires $\Omega(\sqrt{n}/\epsilon^4)$ samples or prove that none exists.