Algorithms, Summer 2019 at CIS

Homework 2: **Due: 7/22/19 before class**

1. Consider the following data stream algorithm for estimating counts of items. We choose a universal hash function $h : \Sigma \to \{0, 1, 2, \ldots, r-1\}$ and a 2-universal hash function $s : \Sigma \to \{-1, 1\}$. We initialize an array A of length r to be all zeros. When we see the stream item a_i, we set $A[h(a_i)] = A[h(a_i)] + s(a_i)$. At the end of the stream, for an element $e \in \Sigma$, suppose we output $\text{est}(e) = s(e) \cdot A[h(e)]$.

   (a) Show that the expected value $\mathbf{E}[\text{est}(e)] = \text{count}(e)$, where $\text{count}(e)$ is the number of occurrences of e in the stream.

   (b) Recall the variance $\mathbf{Var}[X]$ of a random variable X is $\mathbf{E}[(X - \mathbf{E}[X])^2]$, which is also equal to $\mathbf{E}[X^2] - \mathbf{E}^2[X]$. The smaller the variance is, the more likely your random variable is to be close to its expected value. Show $\mathbf{Var}[\text{est}(e)] \leq \frac{1}{r} \sum_{e' \in \Sigma} \text{count}^2(e')$. It might help to use that for independent random variables X and Y, it holds that $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$.

2. Suppose Alice has a bit string $x \in \{0, 1\}^n$ and Bob has a bit string $y \in \{0, 1\}^n$. They are promised that there is exactly one index $i \in \{1, 2, 3, \ldots, n\}$ for which $x_i \neq y_i$, and for all $j \neq i$, $x_j = y_j$. Alice sends a single message to Bob, and Bob needs to figure out what the value of i is with probability at least $4/5$. Show that there is a way to do this where Alice's message length is only $O(\log n)$ bits long.

3. Suppose we have a communication game between Alice, who has x, and Bob, who has y, and where Alice sends a single, possibly randomized, message $M(x)$ to Bob. They would like to solve a problem $f(x, y)$ with probability 1. A correct protocol for f is one where Bob can output $f(x, y)$ given $M(x)$ and y. The *randomized communication complexity* $R(f)$ is the minimum, over all correct protocols for f, of Alice's expected message length, maximized over all inputs.

   Instead of the randomized communication complexity we can look at the *distributional communication complexity* $D_\mu(f)$. Now Alice and Bob do not have random choices, and must behave deterministically. However, x and y are drawn from a distribution μ on the inputs, and $D_\mu(f)$ is the minimum, over all correct protocols for f, of Alice's expected message length, where the expectation is taken over a random input pair (x, y) drawn from μ.

   Prove that $R(f) \geq D_\mu(f)$ for any input distribution μ.