Problem Set 1

Due: Friday, March 7, 4 p.m. at the ITO

---

1. Recall the randomized min-cut algorithm discussed in the first lecture (see Section 1.4 of the textbook). There may be several different min-cut sets in a graph. Using the analysis of the randomized min-cut algorithm, prove that there can be at most $n(n-1)/2$ distinct min-cut sets.

2. Let $X_1, X_2, \ldots, X_n, \ldots$ be an infinite sequence of independent, identically distributed (i.i.d.) random variables. (For example, each of the X_i's might be the outcome of rolling some die once.) Suppose the X_i's have expectation μ and (finite) standard deviation σ. Use Chebyshev's inequality to prove that, for any fixed $\epsilon > 0$,
$$\lim_{n\to\infty} \mathbf{Pr}\left[\left|\frac{\sum_{i=1}^{n} X_i}{n} - \mu\right| \geq \epsilon\right] = 0.$$

3. Let $a_1, \ldots, a_n$ be a list of n distinct numbers. We say that a_i and a_j are inverted if $i < j$ but $a_i > a_j$. The *Bubblesort* sorting algorithm swaps pairwise adjacent inverted numbers in the list until there are no more inversions, so the list is in sorted order. Suppose that the input to Bubblesort is a random permutation, equally likely to be any of the $n!$ permutations of n distinct numbers. Determine the expected number of inversions that need to be corrected by Bubblesort.

4. The standard proof of the Chernoff bound showing concentration for a sum $X = \sum_{i=1}^{n} X_i$ assumed that the variables X_i are independent. In this problem you are asked to prove a variant of the Chernoff bound that does *not* assume independence.

   Suppose we have a set of $\{0,1\}$–random variables X_i, $i \in [n]$, that satisfy the following "negative correlation" property:

   For any $I \subseteq [n]$, it holds $\mathbf{Pr}[\cap_{i\in I}(X_i = 1)] \leq \prod_{i\in I} \mathbf{Pr}[X_i = 1]$.

   (a) Suppose $\hat{X}_i$ is a random variable with the same distribution as X_i, but the $\hat{X}_i$'s are all independent of each other. Let $\hat{X} = \sum_{i=1}^{n} \hat{X}_i$. Prove that for any $I \subseteq [n]$ it holds
   $$\mathbf{E}\left[\prod_{i\in I} X_i\right] \leq \mathbf{E}\left[\prod_{i\in I} \hat{X}_i\right].$$
   As a consequence of the above, show that $\mathbf{E}[e^{tX}] \leq \mathbf{E}[e^{t\hat{X}}]$ for any $t \geq 0$.

   (b) Read the proof of the Chernoff bound in the textbook (also given in class), and show how to prove the following variant: for $\{0,1\}$–random variables X_i, $i \in [n]$, that satisfy the "negative correlation" property and $\delta > 0$ it holds
   $$\mathbf{Pr}\left[X \geq \mathbf{E}[X] + n\delta\right] \leq e^{-2\delta^2 n}.$$
   Can you see where the proof breaks down if we want to prove the bound on the lower tail? Can you suggest a property similar to negative correlation that suffices to prove the bound on the lower tail?

5. In this problem, we will analyze a simple algorithm to learn an unknown probability distribution from samples.

   A *discrete probability distribution* over the set $[n] = \{1, \ldots, n\}$ can be viewed as a function $p : [n] \to [0, 1]$. The number $p(i)$ represents "the probability the distribution p assigns to point i." Hence, we have that $p(i) \geq 0$ for all $i \in [n]$, and $\sum_{i=1}^{n} p(i) = 1$. For two distributions p, q over $[n]$ the *total variation distance* between p and q is the quantity $\mathrm{d_{TV}}(p, q) := \sum_{i=1}^{n} |p(i) - q(i)|$. ($\mathrm{d_{TV}}(p, q)$ represents a measure of the "closeness" between p and q.)

   In many scenarios we are interested in *learning* an *unknown* probability distribution from *samples*. In more detail, a *learning algorithm* is given access to a *sampling oracle* for p, i.e., a "black-box" with the following property: Every invocation of the oracle (query) yields an output $s \in [n]$ that is a random variable distributed according to p (i.e., $\mathbf{Pr}[s = j] = p(j)$ for all $j \in [n]$) and is independent of all previous outputs. For a given error parameter $0 < \epsilon < 1$, the goal of the learning algorithm is to output a *hypothesis distribution* h over $[n]$ such that with probability at least $2/3$ (over the samples obtained from the oracle) the following condition is satisfied: $\mathrm{d_{TV}}(p, h) \leq \epsilon$.

   Given m independent samples $s_1, \ldots, s_m$, drawn from distribution $p : [n] \to [0, 1]$, the *empirical distribution* $\widehat{p}_m : [n] \to [0, 1]$ is defined as follows: for all $i \in [n]$,

   $$\widehat{p}_m(i) = \frac{|\{j \in [m] \mid s_j = i\}|}{m}.$$

   Consider the following algorithm:

   "Draw m samples from the oracle for p and output the distribution $h = \widehat{p}_m$."

   (a) For $i \in [n]$, let $N_i = |\{j \in [m] \mid s_j = i\}|$ denote the number of samples that "land" on point i. Show that $\mathbf{Var}[N_i] = mp(i)\,(1 - p(i))$.

   (b) Show that $\mathbf{E}\left[|p(i) - \widehat{p}_m(i)|\right] \leq \sqrt{\frac{p(i)}{m}}$. Deduce as a consequence that

   $$\mathbf{E}\left[\mathrm{d_{TV}}(p, \widehat{p}_m)\right] \leq \sqrt{\frac{n}{m}}.$$

   (Hint: Use (a) along with Jensen's inequality.)

   (c) Show that there exists a constant $C > 0$ such that if $m \geq Cn/\epsilon^2$ the above described algorithm satisfies $\mathrm{d_{TV}}(p, h) \leq \epsilon$ with probability at least $9/10$.