6.893 Sub-linear Algorithms	November 17, 2014
Homework 2	
Lecturer: Ronitt Rubinfeld	*Due Date: December 1, 2014*

Turn in your solution to each *problem on a separate sheet of paper, with your name on each one.*

1. Let p be a distribution over $[n] \times [m]$. We say that p is *independent* if the induced distributions $\pi_1 p$ and $\pi_2 p$ are independent, i.e., that $p = (\pi_1 p) \times (\pi_2 p)$.[1] Equivalently, p is independent if for all $i \in [n]$ and $j \in [m]$, $p(i,j) = (\pi_1 p)(i) \cdot (\pi_2 p)(j)$.

   We say that p is *ϵ-independent* if there is a distribution q that is independent and $|p-q|_1 \le \epsilon$. Otherwise, we say p is *not ϵ-independent* or is *ϵ-far from being independent.*

   Given access to independent samples of a distribution p over $[n] \times [m]$, an *independence tester* outputs "pass" if p is independent, and "fail" if p is ϵ-far from independent (with error probability at most 1/3).

   (a) Prove the following: Let A, B be distributions over $S \times T$. If $|A - B| \le \epsilon/3$ and B is independent, then $|A - (\pi_1 A) \times (\pi_2 A)| \le \epsilon$.

   Hint: It may help to first prove the following. Let X_1, X_2 be distributions over S and Y_1, Y_2 be distributions over T. Then $|X_1 \times Y_1 - X_2 \times Y_2|_1 \le |X_1 - X_2|_1 + |Y_1 - Y_2|_1$.

   (b) Give an independence tester which uses $\tilde{O}((nm)^{2/3} poly(1/\epsilon))$ samples. You may use (without proof) the claim from the end of lecture 3 which says that there is an algorithm for distinguishing whether two distributions given by samples are identical or ϵ-far in ℓ_1 distance using $\tilde{O}(n^{2/3} poly(1/\epsilon))$ samples.

2. (Don't do this one! already assigned last time.) Suppose an algorithm has the following behavior when given error parameter ϵ and access to samples of a distribution p over a domain $D = \{1, \ldots, n\}$:

   - if p is monotone, then $\mathcal{A}$ outputs "pass" with probability at least 2/3.
   - if for all monotone distributions q over D, $|p - q|_1 > \epsilon$, then $\mathcal{A}$ outputs "fail" with probability at least 2/3

   Show that this algorithm must make $\Omega(\sqrt{n})$ queries.

   Hint: Reduce from the problem of testing uniformity.

3. In class we gave an MST approximation algorithm for graphs in which the weights on each edge were integers in the set $\{1..w\}$. Show that one can get an approximation algorithm when the weights can be any value in the range $[1..w]$ (it is ok to get a slightly worse running time).

[1] For a distribution A over $[n] \times [m]$, and for $i \in \{1, 2\}$, we use $\pi_i A$ to denote the distribution you get from the procedure of choosing an element according to A and then outputting only the value of the the i-th coordinate.

4. Given a graph G of max degree d, and a parameter ϵ, give an algorithm for property testing of connectivity. That is, if G is connected, then the algorithm should pass with probability 1, and if G is ϵ-far from connected (at least $\epsilon \cdot n$ edges must be added to connect G), then the algorithm should fail with probability at least 3/4. You algorithm should be as efficient as possible in terms of n, d, ϵ.

5. Show a lower bound on giving a multiplicative estimate on the MST: Give two distributions over graphs of degree at most d and weights in the range $\{1, \ldots, w\}$ such that

   (a) graphs in one distribution have an MST weight that is at least twice the MST weight of the graphs in the in other distribution

   (b) in order to distinguish the two distributions with constant probability of success, one must make at least $\Omega(w)$ queries

   If you can get the lower bound to have some nontrivial dependence on d and ϵ, even better!

6. The diameter of an unweighted graph is the maximum distance between any pair of nodes. Give a tester for graphs with degree at most d (in the adjacency list model) that have low diameter. The tester should have the following specific behavior:

   (a) Graphs with diameter at most D are always accepted.

   (b) Graphs which are ϵ-far (that is, at least ϵdn edges must be added) from having diameter $4D + 2$ are failed with probability at least 2/3.

   (c) The query complexity of the tester should be $O(1/\epsilon^c)$ for some constant $1 \leq c \leq \infty$.