Sub-linear Algorithms	March 23, 2017

Homework 5

Lecturer: Ronitt Rubinfeld	*Due Date: April 6, 2017*

Homework guidelines: You may work with other students, as long as (1) they have not yet solved the problem, (2) you write down the names of all other students with which you discussed the problem, and (3) you write up the solution on your own. No points will be deducted, no matter how many people you talk to, as long as you are honest. It's ok to look up famous sums and inequalities that help you to solve the problem, but don't look up an entire solution.

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 $|p-q|_1 \leq \epsilon$. 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 $|A - B| \leq \epsilon/3$ and B is independent, then $|A - (\pi_1 A) \times (\pi_2 A)| \leq \epsilon$.
   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 $|X_1 \times Y_1 - X_2 \times Y_2|_1 \leq |X_1 - X_2|_1 + |Y_1 - Y_2|_1$.

   (b) Give an independence tester which makes $\tilde{O}((nm)^{2/3} poly(1/\epsilon))$ queries. (You may use the L1 tester mentioned in class, which uses $\tilde{O}(n^{2/3} poly(1/\epsilon))$ samples, without proving its correctness.)

[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.