CS 787: Advanced Algorithms	**Homework 2**
Out: 10/21/16	**Due: 11/11/16**

Guidelines

- This homework consists of 7 problems. You are required to solve and turn in all of the problems.
- Some of the problems are difficult, so please get started early. Late submissions do not get any credit.
- Please typeset your solutions.
- Homework may be done in pairs. Please write your names clearly on your homework.

Problems

1. For this problem you will need the following definition.

   Let H be a family of functions mapping a universe U to a set R. This family is called a k-wise independent hash family if for every choice of k elements from U, $x_1, x_2, \cdots, x_k$, and every choice of k elements in R, $r_1, r_2, \cdots, r_k$, over the choice of h drawn uniformly from H, we have

   $$\Pr\left[(h(x_1) = r_1) \wedge (h(x_2) = r_2) \wedge \cdots \wedge (h(x_k) = r_k)\right] = 1/|R|^k.$$

   In the following parts, both the universe U and the range R are defined to be the set of integers modulo a prime p: $\{0, 1, \cdots, p-1\}$.

   (a) Let a, b be any integers in $\{0, 1, \cdots, p-1\}$. For any integer x, let $h_{a,b}(x) = ax + b \pmod p$. Let $H = \{h_{a,b} \quad \forall a, b \in \{0, 1, \cdots, p-1\}\}$. Prove that H is a 2-way independent hash family, but does not satisfy 3-way independence.

   (b) For this problem, let $h_{a,b,c}(x) = ax^2 + bx + c \pmod p$. Let $H = \{h_{a,b,c} \quad \forall a, b, c \in \{0, 1, \cdots, p-1\}\}$. Prove that H is a 3-way independent hash family. How would you construct a 4-way independent hash family? How many bits does one need to store a hash function drawn from your hash family?

2. n balls are thrown into n bins, with each ball landing in a uniformly random bin independently of other throws. Observe that in expectation each bin gets one ball. Find a number k, as a function of n, such that with probability at least $1 - 1/n$, no bin gets more than k balls. Try to make k as small as possible as a function of n.

   You may need to use the following approximation for binomials, derived from Sterling's approximation:

   $$\left(\frac{n}{k}\right)^k \leq \binom{n}{k} \leq \left(\frac{en}{k}\right)^k.$$

3. Consider the following balls and bins process that proceeds in rounds. In the first round, we throw n balls independently and uniformly at random into n bins. At the end of each round, we discard every ball that fell into a bin by itself (that is, had no collisions). The remaining balls are retained for the next round, in which they are again thrown independently and uniformly at random into the n bins. Prove that this process takes $O(\log \log n)$ steps in expectation.

4. Consider the following estimator for keeping count. The estimator X is initialized to be 0. Whenever the count increases, the estimator is incremented with probability $1/2^X$ and otherwise kept the same. Suppose that at the end of the process, the count is C. Compute $E[2^X]$. Use this estimator as a basis for a streaming algorithm that produces an (ϵ, δ)-approximation for the length of the stream. Try to use as little storage space as you can.

5. In class we saw how to maintain a uniformly random sample over a stream of elements. In this question our goal is to sample elements in proportion to their values. Design an algorithm that maintains a sample over a stream $\{a_1, a_2, \cdots\}$, where at any time t the sample is drawn from the first t elements of the stream and is equal to the i element, a_i, with probability proportional to a_i. In other words, if S_t denotes the random sample at time t, we should have for all $i \leq t$,
$$\Pr[S_t = a_i] = \frac{a_i}{\sum_{j \leq t} a_j}.$$

6. Consider a stream $\{a_1, a_2, \cdots, a_m\}$ of m elements drawn from universe U. Let $f_i = |\{j : a_j = i\}|$ denote the frequency of an element $i \in U$. Let $H = \sum_{i \in U} \frac{f_i}{m} \log(\frac{m}{f_i})$ denote the *empirical entropy* of the stream. In this problem, you will develop a streaming algorithm to obtain an (ϵ, δ) approximation for the empirical entropy. We will assume that $f_i \leq m/4$ for every $i \in U$, and will interpret $0 \log(1/0)$ as 0.

   Define $G(r) = r \log\left(\frac{m}{r}\right) - (r-1) \log\left(\frac{m}{r-1}\right)$ for $r \leq m$.

   (a) Prove that $\log(\frac{m}{r}) - 1 \leq G(r) \leq \log(\frac{m}{r})$ for all $r \in \{1, \cdots, m-1\}$.

   (b) Let J be an index drawn uniformly at random from the set $[m]$. Let $r(J)$ denote the number of occurences of the element a_J in the stream after and including itself. That is, $r(J) = |\{j \geq J : a_j = a_J\}|$. Consider the random variable $X = G(r(J))$.

   Compute the expectation of X.

   (c) Let Y be a random variable that is a sum of n independent random variables each taking a value from the range $[0, B]$. State a version of the Chernoff bound for the variable Y.

   (d) Use the estimator from part (b) to develop an (ϵ, δ) approximation to H. Use Chernoff bounds to obtain your result. (I do not recommend trying to compute the variance of your estimator and applying Chebyshev's.)

7. Mention a topic you would like to see discussed in class.