This homework is due in class March 20, 2003

Homework 7

1. Recall that a distribution over functions $h : U \to T$ is *2-universal* if for every two different $x, y \in U$ we have $\mathbf{Pr}[h(x) = h(y)] \leq 1/|T|$.

   Consider the following distribution of functions. For each one, prove that it is 2-universal or give a counter-example.

   (a) Let p be a prime and n be an integer, $p \geq n$. Consider the distribution where you pick at random $a \in \{1, \ldots, p-1\}$ and then define $h_a : \{0, \ldots, p-1\} \to \{0, \ldots, n-1\}$ as $h_a(x) = (ax \bmod p) \bmod n$.

   (b) Let p be a prime and n be an integer, $p \geq n$. Consider the distribution where you pick at random $b \in \{0, \ldots, p-1\}$ and then define $h_b : \{0, \ldots, p-1\} \to \{0, \ldots, n-1\}$ as $h_a(x) = (x + b \bmod p) \bmod n$.

   (c) Let m be a multiple of n. Consider the distribution where you pick at random $a \in \{1, \ldots, m-1\}$ and $b \in \{0, \ldots, m-1\}$ and then define $h_b : \{0, \ldots, p-1\} \to \{0, \ldots, n-1\}$ as $h_{a,b}(x) = (ax + b \bmod m) \bmod n$.

2. We say that a distribution over functions $h : U \to T$ is *strongly 2-universal* (or *pairwise independent*) if for every two different $x, y \in U$ and every two values $u, v \in T$ we have $\mathbf{Pr}[h(x) = u \wedge h(y) = v] = 1/|T|^2$.

   Note that if h is picked from a strongly 2-universal family and $x_1, \ldots, x_k$ are different elements of U, then $h(x_1), \ldots, h(x_k)$ are pairwise independent random elements of T.

   Consider the following distribution of functions $h : \{0,1\}^n \to \{0,1\}$. For each one, prove that it is strongly 2-universal or give a counter-example.

   (a) Pick at random $a_0, \ldots, a_n \in \{0,1\}$, and define

   $$h_{a_0,\ldots,a_n}(x_1, \ldots, x_n) = a_0 + a_1 x_1 + \ldots a_n x_n \bmod 2$$

   (b) Pick at random $a_1, \ldots, a_n \in \{0,1\}$, and define

   $$h_{a_0,\ldots,a_n}(x_1, \ldots, x_n) = a_1 x_1 + \ldots a_n x_n \bmod 2$$

3. Consider the following problem: you are given a sequence $a_1, \ldots, a_n$ of elements of some universe U, possibly with repetitions, and you would like to know how many distinct elements are there in the sequence. Also, you would like a *one-pass* algorithm that uses very little memory.[1]

   While it can be shown that there is no memory-efficient one-pass algorithm to solve the problem exactly, we will describe and analyse a probabilistic approximate method.

[1] One-pass memory-efficient algorithms, also called "streaming" algorithms, are a hot topic of current research.

- Pick at random a function $h : U \to \{1, \dots, N\}$ from a *strongly 2-universal* family, where N is a big integer larger than n.
- Let $m = \min\{h(a_1), \dots, h(a_n)\}$
- Output N/m

The idea is that if $a_1, \dots, a_n$ contains k distinct elements, then the sequence $h(a_1), \dots, h(a_n)$ contains k pairwise independent random elements of $1, \dots, N$. Now, it is known that if we pick k random elements from $\{1, \dots, N\}$, then the smallest of the k samples is expected to be around N/k. So, if we let the minimum be m, N/m should be around k. We want to show that this is still true if we only have pairwise independence, and also that there is a reasonably good probability that we do not make very big mistakes.

(a) Suppose that $a_1, \dots, a_n$ contains k distinct elements, and also suppose for simlicity that N is a multiple of $4k$. Show that

$$\mathbf{Pr}[\text{ algorithm's output } > 4k] \leq \frac{1}{4}$$

[Hint: for the algorithm to output more than $4k$ it must be that at least one of $h(a_1), \dots, h(a_n)$ is smaller than $N/4k$. Compute the expected number of $h(a_i)$ that are smaller than $N/4k$ and use Markov's inequality.]

(b) Suppose that $a_1, \dots, a_n$ contains k distinct elements, and also suppose for simlicity that N is a multiple of k. Show that

$$\mathbf{Pr}\left[\text{ algorithm's output } < \frac{k}{4}\right] \leq \frac{1}{4}$$

[Hint: for the algorithm to output less than $k/4$, it must that none of the $h(a_1), \dots, h(a_n)$ is smaller than $4N/k$. Compute the expectation and the variance of the number of $h(a_i)$ that are smaller than $4N/k$ and use Chebyshev's inequality.]

Additional hint: the algorithm will give the same output no matter what is the order of the elements $a_1, \dots, a_n$ so, for the sake of the analysis, one can assume without loss of generality that $a_1, \dots, a_k$ are all distinct, and that there is no new element in $a_{k+1}, \dots, a_n$.