CSC265 Fall 2020 Homework Assignment 5

due Tuesday, October 20, 2020

A family of hash functions $\mathcal{H}$ from U to $\{0, \ldots, m-1\}$ is *pairwise independent* if for every two distinct keys $x_1, x_2 \in U$ and for every $y_1, y_2 \in \{0, \ldots, m-1\}$,

$$\Prob_{h \in \mathcal{H}} [h(x_1) = y_1 \text{ and } h(x_2) = y_2] = 1/m^2.$$

Let $\mathcal{H}$ be a pairwise independent family of hash functions from U to $\{0, \ldots, m-1\}$.

1. Prove that $\mathcal{H}$ is universal.

2. Let $u = |U|$ and let $m = u^3$. Prove that $\Prob_{h \in \mathcal{H}} [h \text{ is perfect for } U] > 1 - 1/u$.

3. Let $k \in \{0, \ldots, m-1\}$. For any value $a \in U$, let $X_a : \mathcal{H} \to \{0, 1\}$ be the indicator variable such that $X_a(h) = 1$ if and only if $h(a) < k$.
   Let $S \subseteq U$ and let $Y = \sum\{X_a \mid a \in S\}$. Prove that

   $$\var_{h \in \mathcal{H}} [Y] \leq \E_{h \in \mathcal{H}} [Y] = |S|k/m.$$

   You may use without proof any property of variance given in CLRS section C.3.

4. Consider the following algorithm that takes as input a sequence $a_1, \ldots, a_n$ of n elements from U and is supposed to return an estimate of the number d of distinct elements in the sequence. Here t is a parameter of the algorithm.

   > Let $h \in \mathcal{H}$ be chosen uniformly at random.
   > Determine the set T of the t smallest distinct elements in $\{h(a_i) \mid 1 \leq i \leq n\}$.
   > If there are fewer than t distinct elements in $\{h(a_i) \mid 1 \leq i \leq n\}$,
   > then return the size of this set;
   > else let V be the largest element in T.
   > Return $D = (t - \frac{1}{2})(m-1)/V$.

   Explain how to implement this algorithm so that it takes $O(n \log t)$ time and uses $O(t)$ words of memory, each storing $O(\log m)$ bits.
   Assume that a hash function can be stored in $O(1)$ words of memory and that it can be evaluated on an element of U in $O(1)$ time.

5. Give a brief, intuitive explanation why $\E_{h \in \mathcal{H}} [D]$ is approximately d.