Exercise Set VII

These exercises are for your own benefit. Feel free to collaborate and share your answers with other students. Solve as many problems as you can and ask for help if you get stuck for too long. Problems marked * are more difficult but also more fun :).

These problems are taken from various sources at EPFL and on the Internet, too numerous to cite individually.

1 *(Basic Hashing)* Consider a Hash Family $\mathcal{H}$ where each $h \in \mathcal{H}$ is a function $h : U \to [n]$ that maps the elements U to integers $\{0, 1, \ldots, n-1\}$. We assume throughout this exercise that $|U| = n$ and so we are in the classic balls-and-bin setting (with n balls and n bins). Assume that $\mathcal{H}$ is a pairwise independent hash family, i.e., it satisfies the following:

1. $\Pr_{h\in\mathcal{H}}[h(x) = y \wedge h(x') = y'] = \frac{1}{n^2}$ for all $x \neq x' \in U$ and $y, y' \in [n]$.

1a Let Y be the number of items that hash to value 1, i.e., $Y = |\{x \in U : h(x) = 1\}|$. Prove that $\mathbb{E}_{h\in\mathcal{H}}[Y] = |U|/n = 1$ and $\mathrm{Var}[Y] \leq 1$.

1b Use the solution to the previous subproblem to prove that

$$\Pr[Y \geq 2\sqrt{n} + 1] \leq \frac{1}{4n}.$$

Hint: For the proof use Chebyshev's Inequality: Let Y be a random variable with expectation μ and variance σ^2. Then for any real number $k > 0$,

$$\Pr[|Y - \mu| \geq k\sigma] \leq \frac{1}{k^2}\,.$$

1c Conclude that no hash value (i.e., no bin) will receive more than $2\sqrt{n}$ keys with probability at least $3/4$.

1d Show, using an application of Chernoff bounds, that if h is a uniformly random hash function, then maximum bin load is bounded by $O(\log n/ \log\log n)$ with probability $1-1/n$.

2 *(MinHashing)* Suppose we have a universe U of elements. For $A, B \subseteq U$, the Jaccard distance of A, B is defined as

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

This definition is used in practice to calculate a notion of similarity of documents, webpages, etc. For example, suppose U is the set of English words, and any set A represents a document considered as a bag of words. Note that for any two $A, B \subseteq U$, $0 \leq J(A,B) \leq 1$. If $J(A,B)$ is close to 1, then we can say $A \approx B$.

Let $h : U \to [0,1]$ where for each $i \in U$, $h(i)$ is chosen uniformly and independently at random. For a set $S \subseteq U$, let $h_S := \min_{i \in S} h(i)$. **Show that**

$$\Pr[h_A = h_B] = J(A,B).$$

Now, if we have sets $A_1, A_2, \ldots, A_n$, we can use the above idea to figure out which pair of sets are "close" in time essentially $O(n|U|)$. We can also obtain a good approximation of $J(A,B)$ with high probability by using several independently chosen hash functions. Note that the naive algorithm would take $O(n^2|U|)$ to calculate all pairwise similarities.

3 *(*, Pairwise independent random variables)* Let $y_1, y_2, \ldots, y_n$ be uniform random bits. For each non-empty subset $S \subseteq \{1, 2, \ldots, n\}$, define $X_S = \oplus_{i \in S}\, y_i$. Show that the bits $\{X_S : \emptyset \neq S \subseteq \{1, 2, \ldots, n\}\}$ are pairwise independent.

This shows how to stretch n truly random bits to $2^n - 1$ pairwise independent bits.

Hint: Observe that it is sufficient to prove $\mathbb{E}[X_S] = 1/2$ and $\mathbb{E}[X_S X_T] = 1/4$ to show that they are pairwise independent. Also use the identity $\oplus_{i \in A}\, y_i = \frac{1}{2}\left(1 - \prod_{i \in A}(-1)^{y_i}\right)$.