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 $\text{Var}[Y] \leq 1$.

Solution: Let Y_x be the random indicator variable that takes value 1 if $h(x) = 1$ and 0 otherwise. Then $Y = \sum_{x\in U} Y_x$

$$\mathbb{E}[Y] = \sum_{x\in U} \mathbb{E}[Y_x] = \frac{|U|}{n} = 1\,. \qquad \text{(Recall that pairwise independent imply that } \Pr[h(x) = y] = 1/n \text{ for all } x \in U, y \in [n].)$$

$$\begin{aligned}
\text{Var}[Y] &= \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 \\
&= \sum_{x_1,x_2\in U} \mathbb{E}[Y_{x_1}Y_{x_2}] - \sum_{x_1,x_2\in U} \mathbb{E}[Y_{x_1}]\mathbb{E}[Y_{x_2}] \\
&\leq \sum_{x\in U} (\mathbb{E}[Y_x] - \mathbb{E}[Y_x]^2) \qquad \text{(using the two properties of } \mathcal{H}) \\
&= n \cdot (1/n - 1/n^2) = (1 - 1/n)\,.
\end{aligned}$$

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 $k > 0$,

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

Solution: It is straightforward from applying Chebyshev's Inequality with $k = 2\sqrt{n}$.

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

Solution: This follows by a union bound over all n bins.

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 $1-1/n$.

Solution: Fix a bin j. For $i \in [n]$ let X_i denote the indicator of ball i going to bin j. The X_i's are independent Bernoulli$(1/n)$ random variables, so by the Chernoff bound shown in class $X = \sum_i X_i$ satisfies

$$\Pr\left[X > 1+\delta\right] \le \frac{e^\delta}{(1+\delta)^{1+\delta}},$$

since the expectation of X is 1. Setting $\delta = C \ln n / \ln\ln n$ for a large constant C, we get

$$\frac{e^\delta}{(1+\delta)^{1+\delta}} \le \exp\left(-\frac{1}{2}(1+\delta)\ln(1+\delta)\right) = \exp\left(-\Omega(C\log n)\right) \le 1/n^2.$$

A union bound over all n bins gives the result.

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 $A, B \subseteq U$, $0 \le J(A,B) \le 1$. 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.

Solution: First, let us simplify the situation a little by noticing that with probability 1, all elements $h(i)$ for $i \in U$ are different. This is because $\Pr[h(i) = h(j)] = 0$ for $i \ne j$ (recall that each $h(i)$ is uniform on the interval $[0,1]$).

Given this, let us see where $\min_{i\in A\cup B} h(i)$ is attained:

- if it is attained in $A\cap B$, then $h_A = h_B = h_{A\cup B} = h_{A\cap B}$,
- otherwise, say it is attained in $A \setminus B$: then $h_A < h_B$.

Therefore the event $h_A = h_B$ is (almost everywhere) equal to $h_{A\cup B} = h_{A\cap B}$. Furthermore, notice that for any set $S \subseteq U$ and any $i \in S$ we have $\Pr[h(i) = h_S] = 1/|S|$ due to symmetry. Therefore

$$\Pr[h_A = h_B] = \Pr[h_{A\cap B} = h_{A\cup B}] = \sum_{i\in A\cap B} \Pr[h(i) = h_{A\cup B}] = |A\cap B| \cdot \frac{1}{|A\cup B|} = J(A,B).$$

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* $\oplus_{i\in A}\, y_i = \frac{1}{2}\left(1 - \prod_{i \in A}(-1)^{y_i}\right)$.

Solution: Recall the definition of pairwise independence: for any non-empty S and T such that $S \neq T$ and two bits b_S and b_T, we have

$$\Pr[X_S = b_S \wedge X_T = b_T] = 1/4\,.$$

We now first argue that $\mathbb{E}[X_S] = 1/2, \mathbb{E}[X_T] = 1/2$ and $\mathbb{E}[X_S X_T] = 1/4$ implies that they are pairwise independent. We have

$$\begin{aligned}
&\Pr[X_S = 1 \wedge X_T = 1] = \mathbb{E}[X_S X_T] = 1/4\,,\\
&\Pr[X_S = 1 \wedge X_T = 0] = \mathbb{E}[X_S] - \mathbb{E}[X_S X_T] = 1/4\,,\\
&\Pr[X_S = 0 \wedge X_T = 1] = \mathbb{E}[X_T] - \mathbb{E}[X_S X_T] = 1/4\,,\\
&\Pr[X_S = 0 \wedge X_T = 0] = \text{“remaining probability”} = 1 - 3 \cdot 1/4 = 1/4\,.
\end{aligned}$$

We thus complete the proof by showing that $\mathbb{E}[X_S] = \mathbb{E}[X_T] = 1/2$ and $\mathbb{E}[X_S X_T] = 1/4$. In both calculations we use the identity $\oplus_{i\in A}\, y_i = \frac{1}{2}\left(1 - \prod_{i \in A}(-1)^{y_i}\right)$. For the former,

$$\mathbb{E}[X_S] = \mathbb{E}[\oplus_{i\in S}\, y_i] = \mathbb{E}\left[\frac{1}{2}\left(1 - \prod_{i\in S}(-1)^{y_i}\right)\right] = \frac{1}{2}\left(1 - \prod_{i \in S}\mathbb{E}[(-1)^{y_i}]\right) = \frac{1}{2}\,.$$

The second to last equality is due to the independence of the random bits y_i and the last equality follows because y_i is an uniform random bit. The same calculation also shows that $\mathbb{E}[X_T] = 1/2$.

For the latter,

$$\begin{aligned}
\mathbb{E}[X_S X_T] &= \mathbb{E}[\oplus_{i\in S}\, y_i \cdot \oplus_{i \in T}\, y_i]\\
&= \mathbb{E}\left[\frac{1}{2}\left(1 - \prod_{i\in S}(-1)^{y_i}\right)\cdot \frac{1}{2}\left(1 - \prod_{i\in T}(-1)^{y_i}\right)\right]\\
&= \frac{1}{4}\left(1 - \mathbb{E}\left[\prod_{i\in S}(-1)^{y_i}\right] - \mathbb{E}\left[\prod_{i\in T}(-1)^{y_i}\right] + \mathbb{E}\left[\prod_{i\in S}(-1)^{y_i}\prod_{i\in T}(-1)^{y_i}\right]\right)\\
&= \frac{1}{4}\left(1 + \mathbb{E}\left[\prod_{i\in S}(-1)^{y_i}\prod_{i\in T}(-1)^{y_i}\right]\right) \qquad \text{(by independence of } y_i\text{s)}\\
&= \frac{1}{4}\left(1 + \mathbb{E}\left[\prod_{i\in S\Delta T}(-1)^{y_i}\right]\right) \qquad \text{(recall } S\Delta T = S \setminus T \cup T \setminus S\text{)}\\
&= \frac{1}{4} \qquad (S\Delta T \neq \emptyset \text{ and again using independence of } y_i\text{s.}
\end{aligned}$$