CS5330: Assignment for Week 7

Due: Tuesday, 24th Mar 2020.

Here are solution sketches to the Week 7 problems. If anything is unclear, please talk to me or your TA.

1. Chebyshev's inequality shows that when n items are hashed into n bins using a hash function from a 2-universal family, the maximum load is at most $1 + \sqrt{2n}$ with probability at least $1/2$. Generalize this argument to k-universal hash functions. That is, find a value such that the probability that the maximum load is larger than that value is at most $1/2$.

   We first count the number of k-wise collisions. For any $S \subseteq [n]$ of size k, the probability that all elements in S hash to the same value is $\leq \frac{1}{n^{k-1}}$ by k-universality. Hence, the expected number of k-wise collisions is $\leq \binom{n}{k}\frac{1}{n^{k-1}} \leq \frac{e^k n}{k^k}$ using the inequality $\binom{n}{k} \leq \left(\frac{en}{k}\right)^k$. Now, if L is the maximum load, the number of k-wise collisions is at least $\binom{L}{k}$. Therefore, $\Pr\left[\binom{L}{k} \geq \frac{2e^k n}{k^k}\right] \leq \frac{1}{2}$. Since $\binom{L}{k} \geq \left(\frac{L}{k}\right)^k$, with probability at least $1/2$, $L^k \leq 2e^k n$ or $L \leq e(2n)^{1/k}$. [This is roughly what I was looking for, but your argument could be a little different.]

2. Suppose $M = \{0,1\}^m$ and $N = \{0,1\}^n$. Let $\mathcal{M} = \{0,1\}^{(m+1)\times n}$ denote the space of Boolean matrices with $m+1$ rows and n columns. For any $x \in M$, let $x^{(1)}$ denote the $(m+1)$-bit vector obtained by appending a 1 to the end of x. For $A \in \mathcal{M}$, define $h_A(x) = x^{(1)}A \pmod 2$. Show that $H = \{h_A : A \in \mathcal{M}\}$ is a 2-universal hash family. Is it also strongly 2-universal?

   H is both 2-universal and strongly 2-universal. Let's first argue universality. Take any two distinct $x, y \in \{0,1\}^m$, and suppose $h_A(x) = h_A(y)$. In particular, $\langle x^{(1)} - y^{(1)}, A_j\rangle = 0 \pmod 2$ for any $j \in [n]$ where A_j is the j'th column of A. Since $x^{(1)} - y^{(1)} \neq 0$, this is a non-trivial linear constraint on A_j, so you can argue that $\Pr_{A_j \sim \{0,1\}^{m+1}}[\langle x^{(1)} - y^{(1)}, A_j\rangle = 0 \pmod 2] = \frac{1}{2}$ for every j. Since the A_j's are independent, it follows that $h_A(x) = h_A(y)$ with probability $\frac{1}{2^n} = \frac{1}{|N|}$.

   For strong universality, fix any $u, v \in \{0,1\}^n$. We need to argue that $\Pr[h_A(x) = u, h_A(y) = v] = \frac{1}{2^{2n}}$ for any $x \neq y$. As above, let A_j be the j'th column of A. It's clear that $\Pr[\langle x^{(1)}, A_j\rangle = u_j] = \frac{1}{2}$

as $x^{(1)} \neq 0$ (I am assuming $\mod 2$ everywhere). Also because $x^{(1)}$ and $y^{(1)}$ are linearly independent, $\Pr[\langle y^{(1)}, A_j\rangle = v_j \mid \langle x^{(1)}, A_j\rangle = u_j] = \Pr[\langle y^{(1)}, A_j\rangle = v_j] = \frac{1}{2}$. Hence, $\Pr[\langle x^{(1)}, A_j\rangle = u_j, \langle y^{(1)}, A_j\rangle = v_j] = \frac{1}{4}$. Using the independence of the A_j's proves our claim.

Here, I assumed the useful fact that if $\alpha, \beta \in \{0,1\}^n$ are linearly independent, and if x is drawn uniformly from $\{0,1\}^n$, the random variables $\langle \alpha, x\rangle \pmod 2$ and $\langle \beta, x\rangle \pmod 2$ are independent as random variables. Prove this.

3. For any hash function $h : M \to N$, say it is ϵ-*good* for two sets $A \subseteq M$ and $B \subseteq N$ if for x drawn uniformly from M:

$$\left|\Pr[x \in A, h(x) \in B] - \frac{|A|}{|M|}\frac{|B|}{|N|}\right| \leq \epsilon$$

Suppose h is drawn uniformly from a strongly 2-universal hash family $\mathcal{H}$. Show that for any $\epsilon > 0, A \subseteq M, B \subseteq N$, the probability that h is not ϵ-good for A and B is at most:

$$\frac{|A|/|M| \cdot |B|/|N|}{\epsilon^2 |M|}.$$

Let G_x be the indicator that $h(x) \in B$. By strong 2-universality, $\Pr[G_x = 1] = \frac{|B|}{|N|}$. The key observation is that additionally, G_x and G_y are independent because of strong 2-universality. Hence, if $G = \sum_{x \in A} G_x$, $\mathbb{E}[G] = \sum_{x\in A} \Pr[G_x = 1] = \frac{|A| \cdot |B|}{|N|}$, and $\mathsf{Var}[G] = \sum_{x \in A} \mathsf{Var}[G_x] \leq \sum_{x\in A} \mathbb{E}[G_x^2] = \sum_{x \in A} \mathbb{E}[G_x] = \frac{|A|\cdot|B|}{|N|}$. Using Chebyshev's inequality, $\Pr[|G - |A| \cdot |B|/|N|| > \varepsilon |M|] \leq \left(\frac{|A|\cdot|B|}{|N|}\right)/(\varepsilon^2 M^2) = \frac{|A|/|M| \cdot |B|/|N|}{\epsilon^2 |M|}$.