Solutions for Week 7

Bao Jinge A0214306U e0522065@u.nus.edu

1

Suppose we have m items labeled $x_1, x_2, ...x_m$ and n bins. For each $1 \leq i_1, i_2, i_3, ..., i_k$, let $X_{i_1,i_2,...,i_k}$ indicates that ball $x_{i_1}, x_{i_2}, ..., x_{i_k}$ land in same bin.
Let $X = \sum_{1 \leq i_1 < i_2 < ... < i_k \leq m} X_{i_1,i_2,...,i_k}$. By linearity of expectations,

$$\mathbb{E}[X] = \mathbb{E}[\sum_{1 \leq i_1 < i_2 < ... < i_k \leq m} X_{i_1,i_2,...,i_k}] = \sum_{1 \leq i_1 < i_2 < ... < i_k \leq m} \mathbb{E}[X_{i_1,i_2,...,i_k}]$$

Let T denotes the number of collisions, obviously $T = \binom{k}{2} X$

$$T = \binom{k}{2} \sum_{1 \leq i_1 < i_2 < ... < i_k \leq m} \mathbb{E}[X_{i_1,i_2,...,i_k}]$$

Since we are using hash function from a k-universal family,

$$\mathbb{E}[X_{i_1,i_2,...,i_k}] = Pr(h(x_{i_1}) = h(x_{i_2}) = ... = h(x_{i_k})) \leq \frac{1}{n^{k-1}}$$

Hence, expecation of collisions is

$$\mathbb{E}[T] \leq \binom{k}{2}\binom{m}{k}\frac{1}{n^{k-1}}$$

Markov's inequality then yields

$$Pr(T \geq 2\binom{k}{2}\binom{m}{k}\frac{1}{n^{k-1}}) \leq \frac{1}{2}$$

if we uppose that the maximum of items in a bin is Y, then the number of collisions T must be at least $\binom{Y}{2}$. Thus,

$$Pr(\binom{Y}{2} \geq 2\binom{k}{2}\binom{m}{k}\frac{1}{n^{k-1}}) \leq Pr(T \geq 2\binom{k}{2}\binom{m}{k}\frac{1}{n^{k-1}}) \leq \frac{1}{2}$$

Let $m = n$, we get

$$\binom{Y}{2} \geq 2\binom{k}{2}\binom{m}{k}\frac{1}{n^{k-1}}$$

We got

$$Pr(Y \geq 1 + 2\sqrt{\binom{k}{2}\binom{m}{k}\frac{1}{n^{k-1}}}) < \frac{1}{2}$$

As for

$$\binom{m}{k} \leq (\frac{em}{k})^k$$

Consequently with $m = n$,

$$Pr(Y \geq 1 + \sqrt{2n(\frac{e^k}{k^{k-2}}))} \leq \frac{1}{2}$$

To sum up, the maximum load is larger than $1 + \sqrt{2n(\frac{e^k}{k^{k-2}})}$ w.p at most $\frac{1}{2}$.

2

For any distinct $i, j \in M$,

$$\begin{aligned} Pr(h_A(x_1) = h_A(x_2)) &= Pr(x_1^{(1)}A(mod2) = x_2^{(1)}A(mod2)) \\ &= Pr((x_1^{(1)} - x_2^{(1)})A = \vec{0}(mod2))) \end{aligned}$$

$\vec{0}$ in equation above is a row vector. Since x_1 and x_2 are distinct row vector, $x_1^{(1)} - x_2^{(1)}$ can't be a zero vector.

$$h_A(x_1) = x_1^{(1)}A(mod2) = y_1$$
$$h_A(x_2) = x_2^{(1)}A(mod2) = y_2$$

Obviously, $y_1, y_2 \in N$. Suppose

$$\begin{aligned} h_A(x_1) &= h_A(x_2) \\ x_1^{(1)}A(mod2) &= x_2^{(1)}A(mod2) \\ (x_1^{(1)} - x_2^{(1)})A(mod2) &= 0 \end{aligned}$$

Let $z = x_1^{(1)} - x_2^{(1)}$. Since x_1 and x_2 are distinct, $x_1^{(1)}$ and $x_2^{(1)}$ are distinct. Futhermore, z is not a zero vector. However we are sure that $(m+1)$-th coordinate is zero. Without loss of generality, suppose that i^*-th coordinate is not zero, where $1 \leq i^* \leq m$.

$$\begin{cases} \sum_{i=1, i\neq i^*}^m z_i A_{i1} = -A_{i^*1} \\ \sum_{i=1, i\neq i^*}^m z_i A_{i2} = -A_{i^*2} \\ \ldots \\ \sum_{i=1, i\neq i^*}^m z_i A_{in} = -A_{i^*n} \end{cases}$$

As for the first equation, because z is fixed, after we fix elements from A_{11} to A_{1m}, we get

$$Pr(\sum_{i=1, i\neq i^*}^{m} z_i A_{i1} = -A_{i^*1}) \leq \frac{1}{2}$$

Consequently, for all n equations

$$
\begin{aligned}
Pr(\forall x_1, x_2 \in M, x_1 \neq x_2 | h_A(x_1) = h_A(x_2)) = Pr(\sum_{i=1, i\neq i^*}^{m} z_i A_{i1} &= -A_{i^*1} \\
\sum_{i=1, i\neq i^*}^{m} z_i A_{i2} &= -A_{i^*2} \\
&\dots \\
\sum_{i=1, i\neq i^*}^{m} z_i A_{in} = -A_{i^*n}) &\leq \frac{1}{2^n}
\end{aligned}
$$

Thus, H is a 2-universal hash family.
To analysis whether H is strongly 2-universal hash family. Suppose we have distinct $x_1, x_2 \in M$ and $y_1, y_2 \in N$. Let

$$
\begin{cases} x_1^{(1)} A &= y_1 \\ x_2^{(1)} A &= y_2 \end{cases}
$$

Since $x_1^{(1)}, x_2^{(1)}$ are non-zero vector, we have

$$
\begin{cases} A &= {x_1^{(1)}}^{-1} y_1 \\ A &= {x_2^{(1)}}^{-1} y_2 \end{cases}
$$

where ${x_1^{(1)}}^{-1}$ is inverse matrix of $x_1^{(1)}$. Consequently,

$$
{x_1^{(1)}}^{-1} y_1 = {x_2^{(1)}}^{-1} y_2
$$

Fix x_1 and x_2, we get

$$
\begin{aligned}
Pr[h(x_1) = y_1 \wedge h(x_2) = y_2] &= Pr[x_1^{(1)} A = y_1 \wedge x_2^{(1)} A = y_2] \\
&= Pr[{x_1^{(1)}}^{-1} y_1 = {x_2^{(1)}}^{-1} y_2] \\
&= \frac{1}{2^n} \frac{1}{2^n}
\end{aligned}
$$

Thus, H is also a strongly 2-universal hash family.

3

Let X_x indicate the event $x \in A, h(x) \in B$. Obviously,

$$
\mathbb{E}[X_x] = Pr(x \in A, h(x) \in B) = Pr(x \in A) Pr(h(x) \in B) = \frac{|A|}{M} \frac{|B|}{N}
$$

Let $X = \sum_{x \in M} X_x$,

$$
\mathbb{E}[X] = \mathbb{E}[\sum_{x \in M} X_x] = \sum_{x \in M} \mathbb{E}[X_x] = |A| \frac{|B|}{N}
$$

To calculate variance of X, we get

$$
\begin{aligned}
Var[X] &= \mathbb{E}[X^2] - \mathbb{E}^2[X] \\
&= \mathbb{E}[\sum_{x\in M} X_x \sum_{x'\in M} X_{x'}] - \mathbb{E}^2[X] \\
&= \mathbb{E}[\sum_{x,x'\in M} X_x X_{x'}] - \mathbb{E}^2[X] \\
&= \mathbb{E}[\sum_{x\in M} X_x^2] + \mathbb{E}[\sum_{x,x'\in M, x\neq x'} X_x X_{x'}] - \mathbb{E}^2[X] \\
&= \sum_{x\in M} \mathbb{E}[X_x^2] + \sum_{x,x'\in M, x\neq x'} \mathbb{E}[X_x X_{x'}] - \mathbb{E}^2[X] \\
&= |A|\frac{|B|}{N} - \frac{|A|^2|B|^2}{N^2} + \sum_{x,x'\in M, x\neq x'} Pr[x \in A, h(x) \in B, x' \in A, h(x') \in B]
\end{aligned}
$$

Since

$$
\begin{aligned}
&Pr[x \in A, h(x) \in B, x' \in A, h(x') \in B] \\
&= Pr[x \in A, h(x) \in B, x' \in A, h(x') \in B | x \in M, h(x) \in N, x' \in M, h(x') \in N] \\
&* Pr[x \in M, h(x) \in N x' \in M, h(x') \in N] \\
&= \frac{|B|}{N}\frac{|B|}{N}\frac{|A|}{M}\frac{|A|}{M}(\frac{1}{N^2}) \\
&= \frac{|A|^2|B|^2}{N^4M^2}
\end{aligned}
$$

where we using the conditional probability and property of strongly 2-universal hash family, the upperbound of variance is

$$
\begin{aligned}
Var[X] &= |A|\frac{|B|}{N} - \frac{|A|^2|B|^2}{N^2} + \sum_{x,x'\in M, x\neq x'} Pr[x \in A, h(x) \in B, x' \in A, h(x') \in B] \\
&\leq |A|\frac{|B|}{N} - \frac{|A|^2|B|^2}{N^2} + \frac{|A|^2|B|^2}{N^4} \\
&\leq |A|\frac{|B|}{N} + (1 - N^2)\frac{|A|^2|B|^2}{N^4} \\
&\leq |A|\frac{|B|}{N}
\end{aligned}
$$

Using Chebyshev's Inequality,

$$
\begin{aligned}
Pr[|X - \mathbb{E}[X]| \geq M\epsilon] &\leq \frac{Var[X]}{M^2\epsilon^2} \\
Pr[|X - |A|\frac{|B|}{N} \geq M\epsilon] &\leq \frac{|A||B|}{NM^2\epsilon^2} \\
Pr[|X - |A|\frac{|B|}{N} \geq M\epsilon] &\leq \frac{|A||B|}{NM^2\epsilon^2} \\
Pr[|X_x - \frac{|A}{M}\frac{|B|}{N} \geq \epsilon] &\leq \frac{|A||B|}{NM^2\epsilon^2} \\
Pr[|Pr(x \in A, h(x) \in B] - \frac{|A}{M}\frac{|B|}{N} \geq \epsilon) &\leq \frac{|A||B|}{NM^2\epsilon^2}
\end{aligned}
$$

Q.E.D.