Contents

1 Hash functions for sampling

1.1 Exercise 1

1.1.1 (a)

We are asked to prove $p \leq Pr[h_m(x)/m < p] \leq 1.01p$. We will use various facts to show this. Firstly that $h(x) = h_m(x)/m$ is a Strong Universal Hash Function, secondly as we are give $p \geq 100/m$ this implies that $p/100 \geq 1/m$, thirdly $h_m(x)/m \leq p$ implies $h_m(x) \leq mp$ and finally we will use that for any y, $y \leq \lceil y \rceil < y + 1$.

We observe that

$$\begin{aligned} & Pr[h_m(x)/m < p] \\ &= \sum_{0 \leq k < mp} Pr[h_m(x) = k] \\ &= \sum_{0 \leq k < mp} \frac{1}{m} \\ &= \frac{1}{m} \left|[0, mp)\right| \\ &= \frac{1}{m} \cdot \lceil mp \rceil \\ &= \frac{\lceil mp \rceil}{m} \end{aligned}$$

Thus we conclude

$$p = \frac{pm}{m} \leq Pr[h_m(x)/m < p] = \frac{\lceil pm \rceil}{m} \leq \frac{pm+1}{m} \leq p + \frac{p}{100} = 1.01p$$

1.1.2 (b)

We are asked to bound the probability that two keys share the same hash value $\dfrac{h_m(x)}{m} = \dfrac{h_m(y)}{m}$, given that $A \subset U, m \geq 100|A|^2$.

To prove this we use that $\dfrac{\binom{n}{2}}{2} = \dfrac{\frac{n(n-1)}{2}}{m} = \dfrac{n(n-1)}{2m}$ The probability can be written as

$$\begin{aligned}
Pr[\exists\{x,y\} \in A : \frac{h_m(x)}{m} = \frac{h_m(x)}{m}] \\
&\leq \sum_{\{x,y\}\in A} Pr\left[\frac{h_m(x)}{m} = \frac{h_m(x)}{m}\right] \\
&= \frac{\binom{|A|}{2}}{m} \\
&\leq \frac{|A|(|A|-1)}{2m} \\
&\leq \frac{|A|(|A|-1)}{2 \cdot 100|A|^2} \\
&\leq \frac{|A|(|A|-1)}{200|A|^2}
\end{aligned}$$

Thus the bound for two keys sharing the same hash value is

$$\leq \frac{1}{200}$$

2 Bottom-k sampling

2.1 Frequency Estimation

2.1.1 Exercise 2

2.1.2 Exercise 3

2.2 Similarity Estimation

2.2.1 Exercise 4

3 Bottom-k sampling with strong universality

3.1 A union bound

3.1.1 Exercise 5

3.2 Upper bound with 2-independence

3.2.1 Exercise 6

3.2.2 Exercise 7