Contents

1 Hash functions for sampling **2**
1.1 Exercise 1 . . . 2
1.1.1 (a) . . . 2
1.1.2 (b) . . . 2

2 Bottom-k sampling **4**
2.1 Frequency Estimation . . . 4
2.1.1 Exercise 2 . . . 4
2.1.2 Exercise 3 (a) . . . 4
2.1.3 Exercise 3 (b) . . . 5
2.2 Similarity Estimation . . . 5
2.2.1 Exercise 4 . . . 5

3 Bottom-k sampling with strong universality **6**
3.1 A union bound . . . 6
3.1.1 Exercise 5 . . . 6
3.2 Upper bound with 2-independence . . . 6
3.2.1 Exercise 6 . . . 6
3.2.2 Exercise 7 . . . 6

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 $\frac{h_m(x)}{m} = \frac{h_m(y)}{m}$, given that $A \subset U, m \geq 100|A|^2$.

To prove this we use that $\frac{\binom{n}{2}}{2} = \frac{\frac{n(n-1)}{2}}{m} = \frac{n(n-1)}{2m}$ 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

We are asked to show that $E\left[|C \cap S_h^k(A)|/k\right] = |C|/|A|$.

We are told that $S_h^k(A)$ is a uniformly random subset of A and C is a subset of A which is independent from $S_h^k(A)$, knowing these we can say

$$Pr\left[x \in S_h^k(A)\right] = p = \frac{k}{|A|} \tag{1}$$

$$Pr\left[x \in C\right] = Pr(C) = \frac{|C|}{|A|} \tag{2}$$

$$\begin{aligned}
&E\left[|C \cap S_h^k(A)|/k\right] \\
&= \frac{1}{k} \cdot E\left[|C \cap S_h^k(A)|\right] \\
&= \frac{1}{k} \sum_{a \in A} E\left[a \in C \wedge a \in S_h^k(A)\right] \\
&= \frac{1}{k} \sum_{a \in A} Pr\left[a \in C \wedge a \in S_h^k(A)\right] \\
&= \frac{1}{k} \sum_{a \in A} \left(Pr\left[a \in C\right] \cdot Pr\left[a \in S_h^k(A)\right]\right) && \text{Independence} \\
&= \frac{1}{k} \cdot \sum_{a \in A} \left(\frac{|C|}{|A|} \cdot \frac{k}{|A|}\right) && \text{From 1 and 2} \\
&= \frac{1}{k} \cdot \frac{|C|}{|A|} \cdot \frac{k}{|A|} \cdot \sum_{a \in A} 1 && \text{Elements in summation independent of } a \in A \\
&= \frac{1}{k} \cdot \frac{|C|}{|A|} \cdot k \\
&= \frac{|C|}{|A|}
\end{aligned}$$

2.1.2 Exercise 3 (a)

When implementing the hashes we would consider what the greater need is, whether the need is to store or retrieve. The reason this is relevant is that, if the primary need is to retrieve, then we would opt to use a binary tree given that the worst case search time is $O\log_2(h)$ for a balanced tree, whereas it's $O(h)$ for an unbalanced tree, where h is the height. For inserting the running time is the same as for searching, under the same conditions.
If the primary interest is to insert and store we would instead use a linked list or a stack as insertion into those data structure is $O(1)$.

2.1.3 Exercise 3 (b)

As mentioned, processing and inserting the key x_{i+1} would take at worst $O(h)$ using binary trees, whereas it would take $O(1)$ for both stacks and lists.

2.2 Similarity Estimation

2.2.1 Exercise 4

We will use the definition of S_h^k to prove $4(a)$ analytically. $S_h^k(A) = \{$the k keys $x \in A$ with the smallest hash values$\}$

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

References