Assignment 1

Albertas Kazakevicius (zvd518),
Philip Lassen (vgh804),
Nicolas Ringsmose Larsen (vgn209)

October 2018

1 Hash functions for sampling

Exercise 1

Given:
$h_m : U \to [m]$ Strongly universal,
$h(x) = h_m(x)/m$
$p \in [0, 1)$

a)

Show that if: $p \geq 100/m$

then $p \leq Pr[\frac{h_m(x)}{m} \leq p] \leq 1.01p$

First assume that mp $\in \mathbb{N}$.

$Pr[h_m(x) \leq mp] = Pr[h_m(x) \in \{0, 1, 2...mp - 1\} = \frac{mp}{p} = p$

Then when $mp \notin \mathbb{N}$:

$Pr[h_m(x) < mp] = Pr[h_m(x) \in \{0, 1, 2...\lfloor mp \rfloor\}] = \frac{\lfloor mp \rfloor + 1}{m}$

$\frac{mp - 1 + 1}{m} = p \leq \frac{\lfloor mp \rfloor + 1}{m} \leq \frac{mp + 1}{m} = p + \frac{1}{m} \leq p + \frac{p}{100}$

b)

If: $A \subseteq U$ and $m \geq 100|A|^2$
Then find upper and bound b) of:
$Pr[h_m(x)/m = h_m(y)/m] \leq b$

We make a union bound over all events of the form $h(x) = h(y)$ where $x \neq y$. There are $\binom{A}{2} = A(A-1)/2$ of them and they each occur with probability $1/m$ thus

$$Pr(h_m(x) = h_m(y)) \leq \frac{|A|!}{2!(A-2)!} \cdot \frac{1}{m} = \frac{A(A-1)}{2m}$$
$$\binom{|A|}{2} = \frac{|A|}{2!(A-2)} \cdot \frac{1}{m} \leq \frac{A(A-1)}{AA(2 \cdot 100)} \leq \frac{1}{200}$$

2 Bottom-k sampling

Exercise 2

The probability of an event being sample is $k/|A|$. There are C elements being sampled and so from the linearity of Expectation we get $E[|C \cap S_h^k(A)|] = (C) * (k/|A|)$. Thus it follows that $E[|C \cap S_h^k(A)|/k] = (C) * (k/|A|)/k = C/|A|$

Exercise 3

a)

Binary max heap data structure should be used to store *bottom_k* sample values.

b)

to process a new key x_{i+1} in this data structure should run in O(1) time and the collisions (if there is an attempt to add the same value) could be checked while insert is made. The insert would be made in O(log n) time and the max element would be changed.

Exercise 4

a

Let $x \in S_h^k(A \cup B)$ and assume that $x \in A$.
Then, since x is among the k smallest element of $(A \cup B)$ it must also be in the k smallest elements of A. Thus,

$$x \in S_h^k(A) \cup S_h^k(B)$$

Since this union contains the k smallest elements of both A and B, and since we already know that $x \in S_h^k(A \cup B)$ and $S_h^k(A \cup B) \subseteq S_h^k(A) \cup S_h^k(B)$, it must also be true that

$$x \in S_h^k(S_h^k(A) \cup S_h^k(B))$$

Similarly, let $x \in S_h^k(S_h^k(A) \cup S_h^k(B))$.
This implies that x is in A or x is in B and also, x is one of the k keys with smallest values in $A \cup B$. Thus

$$x \in S_h^k(S_h^k(A) \cup S_h^k(B)) \implies x \in S_h^k(A \cup B)$$

From this, it follows that

$$S_h^k(S_h^k(A) \cup S_h^k(B)) \subseteq S_h^k(A \cup B)$$

And thus:

$$S_h^k(A \cup B) = S_h^k(S_h^k(A) \cup S_h^k(B))$$

b

Let

$$\begin{aligned} LHS &= A \cap B \cap S_h^k(A \cup B) \\ RHS &= S_h^k(A) \cap S_h^k(B) \cap S_h^k(A \cup B) \end{aligned}$$

Assuming that x is in the k smallest elements of A and the k smallest elements of B and x is amongst the k smallest elements of the union of A and B. Then x is clearly in A and x is in B and x is in the k smallest elements of the union of A and B. Thus RHS is a subset of LHS. Conversely if x is in the k smallest elements of the union of A and B that are also in A and in B then x is going to be an element of the k smallest elements of A and an element of the k smallest elements of B. Thus it follows that the LHS is a subset of the RHS. Since we have shown both side of the inclusion we can conclude that the LHS and RHS are equivalent

3 Bottom-k sampling with strong universality

Exercise 5

If (i) is false than the number of elements that hash below p is greater or equal to k. Similarly if (ii) is false than the number of elements from C that hash below p is less than or equal to $(1+b)p|C|$. Since we know at least k elements hash below p, and at most $(1+b)p|C|$ elements hash below p, thus it follows that $|C \cap S| \leq (1+b)p|C|$.
Let

$$\begin{aligned} n &= |A| \\ f &= \frac{|C|}{|A|} \end{aligned}$$

Then using are conclusions from above

$$\begin{aligned} |C \cap S| &\leq (1+b)p|C| \\ &\leq \frac{(1+b)k}{n(1-a)}|C| \end{aligned}$$

if $n = |A|$

$$\begin{aligned} |C \cap S| &\leq (1+b)p|C| = \frac{(1+b)k}{n(1-a)}|C| \\ |C \cap S| &\leq \frac{(1+b)k|C|}{(1-a)|A|} \\ |C \cap S| &\leq \frac{1+b}{1-a}kp \end{aligned}$$

Exercise 6

$$\begin{aligned}
Pr(X_A < k) &= Pr(X_A < \mu_A(1 - r\sqrt{k})) \\
&= Pr(\mu_A - X_A > r\mu_A/\sqrt{k})
\end{aligned}$$

We know that $k = (1 - a)\mu_A \le \mu_A$. Thus

$$\begin{aligned}
Pr(\mu_A - X_A > r\mu_A/\sqrt{k}) &= Pr(\mu_A - X_A > r\mu_A/\sqrt{(1-a)\mu_A}) \\
&\le Pr(\mu_A - X_A > r\mu_A/\sqrt{\mu_A}) \\
&= Pr(\mu_A - X_A > r\sqrt{\mu_A}) \\
&\le Pr(|\mu_A - X_A| > r\sqrt{\mu_A})
\end{aligned}$$

Then using Lemma 1 We get that

$$Pr(|X_A - \mu_A| > r\mu_A) \le \frac{1}{r^2}$$

Thus we can conclude that

$$Pr(X_A < k) \le \frac{1}{r^2}$$

Exercise 7

$$\begin{aligned}
Pr(X_C > (1 + b)\mu_C) &= Pr(X_C - \mu_C > b\mu_C) \\
&= Pr(X_C - \mu_C > \frac{r\mu_C}{\sqrt{fk}}) \\
&\le Pr(X_C - \mu_C > \frac{r\mu_C}{\sqrt{\mu_C}}) \\
&= Pr(X_C - \mu_C > r\sqrt{\mu_C}) \quad \le Pr(|X_C - \mu_C| > r\sqrt{\mu_C})
\end{aligned}$$

Then using Lemma 1 We get that

$$Pr(|X_C - \mu_C| > r\sqrt{\mu_C}) \le \frac{1}{r^2}$$

Thus we can conclude that

$$Pr(X_C > (1 + b)\mu_C) \le \frac{1}{r^2}$$