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:
$Pr[h_m(x)/m = h_m(y)/m] \leq b$

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

2 Bottom-k sampling

Exercise 2

let X= $\|C \cap S_h^k(A)|$
$X \sim$ Bin $(|C|, \frac{k}{|A|})$
$E(X) = \frac{|C|k}{|A|}$
$\frac{|C|k}{|A|}/k = \frac{|C|}{|A|}$

Exercise 3

a)

Hash table data structure should be used to store the *bottom_k* sample values. In order to tackle collisions when two keys are hashed to the same value some sort of list should be used in the bucket part of hash table.

b)

To process a new key x_{i+1} in this kind of data structure should run in constant time.

Exercise 4

a

$x \in S_h^k(A \cup B) \implies$ (x is one of the the k keys x A with the smallest hash values) or (x is one of the the k keys x B with the smallest hash values) and (x

is one of the the k keys $x \in A \cup B$ with the smallest hash values) Thus it follows that $x \in S_h^k(S_h^k(A) \cup S_h^k(B)) \implies S_h^k(A \cup B) \subseteq S_h^k(S_h^k(A) \cup S_h^k(B))$. Similarly if $x \in S_h^k(S_h^k(A) \cup S_h^k(B)) \implies$ (x is in A or x is in B) and (x is one of the the k keys $x \in A \cup B$ with the smallest hash values). Thus $x \in S_h^k(S_h^k(A) \cup S_h^k(B)) \implies x \in x \in S_h^k(A \cup B) \implies S_h^k(S_h^k(A) \cup S_h^k(B)) \subseteq S_h^k(A \cup B)$. Thus we can conclude

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

b

$$\begin{aligned}
x \in B \wedge x \in S_h^k(A \cup B) &\implies x \in S_h^k(B) \wedge x \in S_h^k(A \cup B) \\
x \in A \wedge x_h^k(A \cup B) &\implies x \in S_h^k(A) \wedge x_h^k(A \cup B) \\
x \in A \wedge x \in B \wedge x_h^k(A \cup B) &\implies x \in S_h^k(A) \wedge x \in S_h^k(A) \wedge x \in S_h^k(A \cup B) \\
\implies A \cap B \cap S_h^k(A \cup B) &\subseteq S_h^k(A) \cap S_h^k(A) \cap S_h^k(A \cup B) \\
x \in B \wedge x \in S_h^k(A \cup B) &\impliedby x \in S_h^k(B) \wedge x \in S_h^k(A \cup B) \\
x \in A \wedge x_h^k(A \cup B) &\impliedby x \in S_h^k(A) \wedge x_h^k(A \cup B) \\
x \in A \wedge x \in B \wedge x_h^k(A \cup B) &\impliedby x \in S_h^k(A) \wedge x \in S_h^k(A) \wedge x \in S_h^k(A \cup B) \\
\implies S_h^k(A) \cap S_h^k(A) \cap S_h^k(A \cup B) &\subseteq A \cap B \cap S_h^k(A \cup B)
\end{aligned}$$

Thus

$$S_h^k(A) \cap S_h^k(A) \cap S_h^k(A \cup B) = A \cap B \cap S_h^k(A \cup B) \tag{1}$$

3 Bottom-k sampling with strong universality

Exercise 5

i) The number of elements that hash below p is greater of equal to K.
ii) The number of elements from C that hash below p is less of equal to $(1+b)p|C|$

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

Exercise 6

$$
\begin{aligned}
Pr(X_A < k) &= Pr(X_A < \mu_A(1 - r\sqrt{k})) \\
&= Pr(\mu_A - X_A > r\mu_A/\sqrt{k}) \\
&\leq Pr(\mu_A - X_A > r\mu_A) \\
&\leq Pr(\mu_A - X_A > r\sqrt{\mu_A}) \\
&\leq (Pr(|\mu_A - X_A| > r\sqrt{\mu_A}) \\
&\leq \frac{1}{r^2}
\end{aligned}
$$

Exercise 7

$$
\begin{aligned}
Pr(X_C > (1+b)\mu_C) &= Pr(X_C - \mu_C > b\mu_C) \\
&= Pr(X_C - \mu_C > \frac{r\mu_C}{\sqrt{fk}}) \\
&\leq Pr(X_C - \mu_C > \frac{r\mu_C}{\sqrt{\mu_C}}) \\
&= Pr(X_C - \mu_C > \sqrt{\mu_C}r) \\
&\leq Pr(|X_C - \mu_C| > r\mu_C) \\
&\leq \frac{1}{r^2}
\end{aligned}
$$