1(a): $p \in [100/m, 1) \Rightarrow p \leq Pr[h(x) < p] < 1.01 \cdot p$

Let $p = \frac{y}{m}$, with $100 \leq y < m$.

$$\frac{y}{m} \leq Pr\left[h(x) < \frac{y}{m}\right] < 1.01\left(\frac{y}{m}\right)$$

Multiply by m/y:

$$1 \leq \frac{m}{y} \cdot Pr\left[h(x) < \frac{y}{m}\right] < 1.01$$

Substitute $h(x)$ for $h_m(x)/m$:

$$1 \leq \frac{m}{y} Pr\left[h_m(x)/m < \frac{y}{m}\right] < 1.01$$

$$1 \leq \frac{m}{y} Pr[h_m(x) < y] < 1.01$$

$Pr[h_m(x) < y] = y/m$ because h_m is strongly universal:

$$1 \leq \frac{m}{y} \cdot \frac{y}{m} < 1.01$$

$$1 \leq 1 < 1.01$$

1(b):

$$\Pr[h(x) = h(y)] = \Pr[h_m(x) = h_m(y)] \quad \text{by def. of } h(x)$$

$$= \sum_{q \in [m]} \Pr[h_m(x) = q \wedge h_m(y) = q]$$

$$= \sum_{q \in [m]} 1/m^2 \quad \text{since strong uni.} \Rightarrow \text{universality}$$

$$= m/m^2 = 1/m$$

$$\leq 1/100|A|^2$$

Ex. 2:

$$I(x) = \begin{cases} 1 & \text{if } x \in C \cap S_h^k(A) \\ 0 & \text{otherwise} \end{cases}$$

$I(x)$ indicator that $x \in C \wedge x \in S_h^k(A)$.

$$E\left[|C \cap S_h^k(A)|\right] = \sum_{x \in C} E[I(x)] \qquad \text{by linearity of expectation}$$

$$= \sum_{x \in C} Pr[x \in C \cap S_h^k(A)]$$

$$= |C| \cdot \frac{k}{|A|}$$

$$E[|C \cap S_h^k(A)|/k] = |C|/|A|$$

Exercise 3

(a)

A binary search tree would be a good choice of data structure to maintain the bottom-k sample. We can have a BST containing at most k nodes which are the hash values for the bottom-k sample. If the tree contains k nodes and a new key arrives, it is inserted and the rightmost node is deleted.

(b)

Assuming the BST is balanced, the running time to process the next key x_{i+1} is $O(\lg(k))$. Otherwise the worst case running time is $O(k)$.

4(a)

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

$$= S_h^k(\{\text{keys } x \in A \mid h(x) \in \text{lowest } k \text{ hash values}\} \cup \{\text{keys } y \in B \mid h(y) \in \text{lowest } k \text{ hash values}\})$$

$$= S_h^k(\{\text{keys } x \in A, y \in B \mid h(x) \in \text{lowest } k \text{ hash values}, h(y) \in \text{lowest } k \text{ hash values}\})$$

$$= S_h^k(\{\text{keys } x \in A \cup B \mid h(x) \in \text{lowest } k \text{ hash values}\})$$

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

■

(b)

It is trivial that

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

since $S_h^k(A) \subseteq A$ and $S_h^k(B) \subseteq B$.

If $x \in A \cap B \cap S_h^k(A \cup B)$ then x is an element in both A and B, and x is in the bottom-k sample of their union. Then x will also be in $S_h^k(A)$:

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

and in $S_h^k(B)$:

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

Clearly

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

since A and B only restricts the left hand side. Since LHS $\subseteq$ RHS and RHS $\subseteq$ LHS:

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

■

4(c) For the expression,

$$|S_h^k(A) \cap S_h^k(B) \cap S_h^k(S_h^k(A) \cup S_h^k(B))|/k$$

given $S_h^k(A)$ and $S_h^k(B)$,
assuming that these are stored in a list,
sorted by hash value,
we can calculate INTERSECTION and UNION
is linear time.
We can also calculate the bottom-k
samples in linear time.
The length of the list can also be
computed in linear time.
Finally, division by k is constant time.

Thus, we can compute the expression
in LINEAR TIME.

Exercise 5:

If (I) is false then we can rewrite $C \cap S$ as

$$C \cap S = C \cap \{x \in A \mid h(x) < p\}$$
$$= \{x \in C \mid h(x) < p\}.$$

This is the same as the elements from C that hash below p, which is the same set as in (II).

We can show that $(1+b)\,p\,|C| = \frac{1+b}{1-a} f k$:

$$(1+b)\,p\,|C| = (1+b)\frac{k}{n(1-a)}|C|$$

$$= \frac{1+b}{1-a}\frac{1}{n} k\,|C|$$

$$= \frac{1+b}{1-a}\frac{|C|}{n} k$$

$$= \frac{1+b}{1-a} f k$$

So (II) is the same as equation (4) and since (II) is false then (4) is also false.

exercise 6.

$P(2) = Pr[X_A < k]$ by the definition of k,

$= Pr[X_A < \mu_A(1 - r/\sqrt{k})]$

$= Pr[X_A < \mu_A - \mu_A \cdot r/\sqrt{k}]$

$= Pr[X_A - \mu_A < -\mu_A r/\sqrt{k}]$

$= Pr[-(X_A - \mu_A) > \mu_A \cdot r/\sqrt{k}]$

$= Pr[|X_A - \mu_A| > \mu_A \cdot r/\sqrt{k}]$

$= Pr[|X_A - \mu_A| > r\sqrt{\mu_A}\sqrt{\mu_A}/\sqrt{k}]$

Also, by the definition of k, $\sqrt{\mu_A}/\sqrt{k} > 1$.

$Pr[|X_A - \mu_A| > r \cdot \sqrt{\mu_A}]$ by Lemma 1, we have

$Pr[|X_A - \mu_A| > r \cdot \sqrt{\mu_A}] \leq 1/r^2$

exercise 7.

$P(11) = Pr[X_C > (1+b)\mu_C]$, by the definition of b, $b = r/\sqrt{fk}$

$= Pr[X_C > (1 + r/\sqrt{fk})\mu_C]$

$= Pr[X_C > \mu_C + \mu_C \cdot r/\sqrt{fk}]$

$= Pr[X_C - \mu_C > \mu_C \cdot r/\sqrt{fk}]$

$= Pr[|X_C - \mu_C| > \mu_C \cdot r/\sqrt{fk}]$

$= Pr[|X_C - \mu_C| > r \cdot \sqrt{\mu_C} \cdot \sqrt{\mu_C}/\sqrt{fk}]$

by the definition of μ_C, $\mu_C > fk$, so $\sqrt{\mu_C}/\sqrt{fk} > 1$.

$Pr[|X_C - \mu_C| > r \cdot \sqrt{\mu_C}]$ by Lemma 1, we have

$Pr[|X_C - \mu_C| > r\sqrt{\mu_C}] \leq 1/r^2$