Advanced Algorithms
Assignment 2

Nikolaj Dybdahl Rathcke (rfq695)
Ola Rønning (vdl761)

October 6, 2015

Hash functions for sampling

Exercise 1 (a)

We want to show that

$$p \leq Pr[h(x) < p] < 1.01p$$

Since p is the sampling probability, we have that $p = \frac{t}{m}$. We also have that $h(x) = h_m(x)/m$, thus we can write

$$\begin{aligned}
\frac{t}{m} &\leq Pr[h(x) < \frac{t}{m}] < 1.01\frac{t}{m} && \Leftrightarrow \\
\frac{t}{m} &\leq Pr[\frac{h_m}{m} < \frac{t}{m}] < 1.01\frac{t}{m} && \Leftrightarrow \\
\frac{t}{m} &\leq Pr[h_m < t] < 1.01\frac{t}{m}
\end{aligned}$$

Since this is a strongly universal hash function $h : U \to [m]$, the probability is at least t/m, which means that the above equation holds.

Exercise 1 (b)

We want to find the complementary probability, that there is no keys $x, y \in A$ that hash to the same value and subtract that from 1. We can find this as

$$\begin{aligned}
P\{h(x) = h(y) \mid \exists x, y \in A\} &= 1 - P\{h(x) \neq h(y) \mid \forall x, y \in A\} \\
&= 1 - \prod_{1 \leq j < |A|} \frac{m - j}{m} \\
&= 1 - \prod_{1 \leq j < |A|} 1 - \frac{j}{m}
\end{aligned}$$

If we let $j = |A|$ for each multiplicand in the product, the probability that there is no pair that hash to the same value become smaller, so it is still an upper bound.

$$\begin{aligned}
P\{h(x) = h(y) \mid \exists x, y \in A\} &= 1 - P\{h(x) \neq h(y) \mid \forall x, y \in A\} \\
&\leq 1 - \prod_{|A|-1} 1 - \frac{|A|}{m} \\
&= 1 - \left(1 - \frac{|A|}{m}\right)^{|A|-1}
\end{aligned}$$

We now let m be the worse case, that is $m = 100|A|^2$, so

$$\begin{aligned} P\{h(x) = h(y) \mid \exists x, y \in A\} &= 1 - P\{h(x) \neq h(y) \mid \forall x, y \in A\} \\ &\leq 1 - \left(1 - \frac{|A|}{100|A|^2}\right)^{|A|-1} \\ &= 1 - \left(1 - \frac{1}{100|A|}\right)^{|A|-1} \end{aligned}$$

So our upper bound becomes $\mathcal{O}\left(1 - \left(1 - \frac{1}{100|A|}\right)^{|A|-1}\right)$.

Bottom-k-sampling

Frequency estimation

Exercise 2

We want to show that

$$E[|C \cap S_h^k(A)|/k] = |C|/|A| \tag{1}$$

The size of $S_h^k(A)$ is equal to k. Since for each element x in A, the probability of x belonging to $S_h^k(A)$ is $k/|A|$, that means that for each element of the random subset C, it also has probability $k/|A|$ of being in $S_h^k(A)$. We can then calculate the expected value of the union between C and $S_h^k(A)$ as it is probability times the number of elements in C:

$$E[|C \cap S_h^k(A)|] = |C|\frac{k}{|A|}$$

Or from equation 1:

$$\begin{aligned} E[|C \cap S_h^k(A)|/k] &= \frac{|C|\frac{k}{|A|}}{k} \\ &= \frac{|C| \cdot k}{|A| \cdot k} \\ &= \frac{|C|}{|A|} \end{aligned}$$

And equation 1 is shown.

Exercise 3 (a)

You would want to be able to find the element with the largest hash value n, compare the incoming key's value m to this, remove the max element if $m < n$ and insert the incoming key.
A fib-heap can extract the maximum $\mathcal{O}(\lg n)$ and find max and insert in $\mathcal{O}(1)$. Furthermore, one could imagine k is a lot smaller than the number of hash values, so it will rarely need to insert and element and extract max, but it will always have to compare with the max element which is done in $\mathcal{O}(1)$ time.

Exercise 3 (b)

The probability that we need to insert the incoming key and extract max is k/n where n is the key number $i+1$, which takes $\mathcal{O}(\lg k)$ time. Otherwise we only need constant amount of work to hash x_{i+1} and compare to max. This means it is expected to be processed in $\mathcal{O}(\frac{k}{n}\lg k)$ time.

Similarity estimation

Exercise 4 (a)

We want to show that

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

The left hand side has the k elements with the lowest hash value, while the inner union on the right hand side can have lowest n elements where $k \leq n \leq 2k$.
This means that

$$S_h^k(A \cup B) \subseteq S_h^k(A) \cup S_h^k(B) \tag{2}$$

And that means you can always take the k smallest elements out of n to obtain the same set.

Exercise 4 (b)

We will prove that:

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

by contradiction.
Let $x \in (A \cap B \cap S_h^k(A \cup B))$,

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

That is there exist a element $x \in A \wedge B$ that hashes to one of the k smallest values in $A \cup B$. However, for the inequality to hold there must exist k smaller hashings in either $A \vee B$, but this contradicts that x hashes to one of the k smallest hashings in $A \cup B$, hence the equality must hold.

Exercise 4 (c)

To estimate

$$\frac{\left|S_h^k(A) \cap S_h^k(B) \cap S_h^k(S_h^k(A) \cup S_h^k(B))\right|}{k} \tag{3}$$

we devise a simple algorithm that looks at the k smallest elements and count how many are in both $S_h^k(A)$ and $S_h^k(B)$, we assume $S_h^k(A)$ and $S_h^k(B)$ are sorted in ascending order by hashing value from hashing with h. This algorithm runs in $\mathbb{O}(k)$ time. The pseudo code can be found in Listing 1.

Listing 1: Algorithm to solve unbiased estimator of the Jaccard Similarity, see Formula 3.

```
jaccardEstimator(S_a,S_b) {
  k = S_a.length
  i = 0
  intCardi = 0
  headA = S_a.ptr
  headB = S_b.ptr
  tSize = sizeof(val(headA))
  for(i<k) {
    headValA = val(headA)
    headValB = val(headB)
    if(headValA == headValB) {
      headA += tSize
      headB += tSize
      intCardi += 1
    } else if (headValA < headValB ) {
             headA += tSize
           } else {
             headB += tSize
           }
  }
  return (intCardi/k)
}
```

Bottom-k-sampling with strong universality

A union bound

Exercise 5

If (I) and (II) are false, we can deduct the following:

(I) tells us that all elements from S (which has size k) hash to a value under p.
(II) tells us that at most $(1+b)p|C|$ elements from C hash to a value under p.

This means that $|C \cap S|$ is at most $(1+b)p|C|$. Now we want to see if the following equation holds:

$$|C \cap S| > \frac{1+b}{1-a} fk$$

We can reduce the right hand side:

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

Last reduction is because $p = \frac{k}{n(1-a)}$. But since we know that $|C \cap S| \leq (1+b)p|C|$, that means

$$|C \cap S| \leq (1+b)p|C| < |C \cap S|$$

which can not be true. So it is proved by contradiction that if (I) and (II) are false, then so is (4).

Upper bound with 2-independence

Exercise 6

We want to show that we can use Lemma 1 to show that

$$P_{(I)} = Pr\left[X_A < k\right] \leq 1/r^2 \tag{4}$$

Lemma 1 tells us that

$$Pr\left[|X - \mu| \geq r\sqrt{\mu}\right] \leq 1/r^2$$

Since X corresponds to X_A, μ to μ_A, and X_A is always less than μ_A since $X_A < \mu_A(1-a)$ for $0 < a < 1$, we can remove the absolute value tags and write:

$$\begin{aligned}
1/r^2 &\geq Pr\left[-(X_A - \mu_A) \geq r\sqrt{\mu_A}\right] \\
&= Pr\left[X_A - \mu_A \leq -r\sqrt{\mu_A}\right] && \text{(Multiply both sides with } -1\text{)} \\
&= Pr\left[X_A \leq \mu_A - r\sqrt{\mu_A}\right] \\
&= Pr\left[\frac{X_A}{\mu_A} \leq \frac{\mu_A}{\mu_A} - \frac{r\sqrt{\mu_A}}{\mu_A}\right] && \text{(Divide by } \mu_A\text{)} \\
&= Pr\left[\frac{X_A}{\mu_A} \leq \frac{\mu_A}{\mu_A} - \frac{r\sqrt{\frac{k}{1-a}}}{\frac{k}{1-a}}\right] && \text{(Since } \mu_A = \tfrac{k}{1-a}\text{)} \\
&= Pr\left[\frac{X_A}{\mu_A} \leq 1 - \frac{\frac{r\sqrt{k}}{\sqrt{1-a}}}{\frac{k}{1-a}}\right] \\
&= Pr\left[\frac{X_A}{\mu_A} \leq 1 - \frac{r\sqrt{k}}{\sqrt{1-a}\frac{k}{1-a}}\right] \\
&= Pr\left[\frac{X_A}{\mu_A} \leq 1 - \frac{r\sqrt{k}\sqrt{1-a}}{k}\right] \\
&= Pr\left[\frac{X_A}{\mu_A} \leq 1 - \frac{r\sqrt{1-a}}{\sqrt{k}}\right] \\
&= Pr\left[\frac{X_A}{\mu_A} < 1 - \frac{r}{\sqrt{k}}\right] && \text{(Since } 0 < \sqrt{1-a} < 1\text{)} \\
&= Pr\left[X_A < \left(1 - \frac{r}{\sqrt{k}}\right)\mu_A\right] \\
&= Pr\left[X_A < k\right] && \text{(As } k = \left(1 - \tfrac{r}{\sqrt{k}}\right)\mu_A\text{)}
\end{aligned}$$

Note that the inequality becomes strictly less when we remove the term $\sqrt{1-a}$ as this makes the right hand side smaller. We have now proved equation 4.

Exercise 7

We want to show that we can use Lemma 1 to show that

$$P_{(II)} = Pr\left[X_C > (1+b)\mu_C\right] \leq 1/r^2 \tag{5}$$

Lemma 1 tells us that

$$Pr\left[|X - \mu| \geq r\sqrt{\mu}\right] \leq 1/r^2$$

Since X corresponds to X_C, μ to μ_C, and X_C is always larger than μ_C since $X_C > \mu_C(1+b)$ for positive b, we can remove the absolute value tags and write:

$$\begin{aligned}
1/r^2 &\geq Pr\left[X_C - \mu_C \geq r\sqrt{\mu_C}\right] \\
&= Pr\left[X_C \geq r\sqrt{\mu_C} + \mu_C\right] \\
&= Pr\left[\frac{X_C}{\sqrt{\mu_C}^2} \geq \frac{r\sqrt{\mu_C}}{\sqrt{\mu_C}^2} + \frac{\mu_C}{\sqrt{\mu_C}^2}\right] && \text{(Divide by } \sqrt{\mu_C}^2\text{)} \\
&= Pr\left[\frac{X_C}{\mu_C} \geq \frac{r}{\sqrt{\mu_C}} + 1\right] \\
&= Pr\left[X_C \geq \left(1 + \frac{r}{\sqrt{\mu_C}}\right)\mu_C\right] \\
&= Pr\left[X_C > (1+b)\,\mu_C\right] && \text{(Since } b = r/\sqrt{fk}\text{)}
\end{aligned}$$

Last reduction is possible since $fk < \mu_C$ which also makes the inequality strict and thus we have proved equation 5.