Individual Assignment 3 E

Casper Kristiansson

December 15, 2023

1 Part 1

We have a stream of numbers where the number **8** appears 10 times, **5** appears 20 times, **3** appears 40 times, **1** appears 20 times, and lastly where **9** appears 10 times. We run the COUNTMIN algorithm on this stream with a $t = 2$ and $k = 4$. We want to calculate the probability when hashing these numbers with a fully-random hash function where $h_i \colon U \to [4]$ that the probability of $\hat{f}(5) > 30$ happening.

As mentioned when we want to query an element we can get the approximate counter by:

$$\hat{f}(x) = \min_{i \in [t]} C[i][h_i(x)]$$

When hashing the numbers, we have 4 different unique hash values. This means that to calculate the probability of $\hat{f}(5) > 30$ happening we want to calculate the different combinations of how the numbers can be combined with the number 5. The exact probability of this happening is **0.21246**.

This can be proved by calculating the combination of numbers that uphold the restriction that $\hat{f}(5) > 30$. These sets are: (5, 3), (5, 1), (5, 8, 3), (5, 8, 1), (5, 8, 9), (5, 3, 1), (5, 3, 9), (5, 1, 9), (5, 8, 3, 1), (5, 8, 3, 9), (5, 3, 1, 9), (5, 8, 3, 1, 9). We then need to calculate the probability of each of these occurrences happening. We have four different types that happen with a probability of:

$$\begin{aligned}
(5, X) &: \left(\frac{1}{4}\right)^1 \times \left(\frac{3}{4}\right)^3 \\
(5, X, X) &: \left(\frac{1}{4}\right)^2 \times \left(\frac{3}{4}\right)^2 \\
(5, X, X, X) &: \left(\frac{1}{4}\right)^3 \times \left(\frac{3}{4}\right)^1 \\
(5, X, X, X, X) &: \left(\frac{1}{4}\right)^4
\end{aligned}$$

This means that we can calculate the probability of the counter for number 5 having a count over 30 for one hashing function as:

$$
\begin{aligned}
&2 \times \left(\left(\frac{1}{4}\right)^1 \times \left(\frac{3}{4}\right)^3 \right) + \\
&6 \times \left(\left(\frac{1}{4}\right)^2 \times \left(\frac{3}{4}\right)^2 \right) + \\
&3 \times \left(\left(\frac{1}{4}\right)^3 \times \left(\frac{3}{4}\right)^1 \right) + \\
&1 \times \left(\left(\frac{1}{4}\right)^4 \right) \\
&= 0.2109375 + 0.2109375 + 0.03515625 + 0.00390625 \\
&= 0.4609375
\end{aligned}
$$

Because we are using two different hash functions we need to take the probability raised to 2. This yields $0.4609375^2 = 0.21246$. This means that the probability of $\hat{f}(5) > 30$ happening is 0.21246. As mentioned in lecture 12 lemma 2 [1] we have a bound for $\hat{f_x} \leq f_x + \frac{2}{k} \cdot (n - f_x)$ with a probability of $1 - \left(\frac{1}{2}^t\right)$. When inserting our variables we get that:

$$
\min_{i \in [t]} C[i][h_i(x)] \leq 20 + \frac{2}{4} \cdot (100 - 20) \text{ with a probability of at least } 1 - \left(\frac{1}{2}\right)^2
$$

.

This states that the probability of $\hat{f}(5) > 60$ is $\leq \left(\frac{1}{2}\right)^2$. Comparing this to our case where the probability of $\hat{f}(5) > 30$ happening is 0.21246 shows us that there is a bit of difference. This is happening due to the distribution of the numbers in the stream. The mentioned lemma 2 bound states that it can guarantee to uphold that constraint while the actual probability is slightly different. The distribution of the numbers will affect the *true exact* probability of $\hat{f_x}$ being larger than a specific number.

2 Part 2

As for the second part where the distribution of the numbers is a bit different where **8** appears **5** times, **7** appears 5 times, **5** appears 20 times, **3** appears 40 times, **1** appears 20 times, and lastly **9** appears 10 times. Same as the first question we are looking for the exact probability of $\hat{f}(5) > 30$ happening. This probability comes to be exactly **0.2490244**.

This can be proven by calculating the probability of all different combinations the numbers might add up to that fulfills the constraint that $\hat{f}(5) > 30$. The possible sets that uphold that constraint are: (5, 3), (5, 1), (5, 8, 3), (5, 8, 1), (5, 8, 9), (5, 7, 3), (5, 7, 1), (5, 7, 9), (5, 3, 1), (5, 3, 9), (5, 1, 9), (5, 8, 7, 3), (5, 8, 7, 1), (5, 8, 7, 9), (5, 8, 3, 1), (5, 8, 3, 9), (5, 8, 1, 9), (5, 7, 3, 1), (5, 7, 3, 9), (5, 7, 1, 9), (5, 3, 1, 9), (5, 8, 7, 3, 1), (5, 8, 7, 3, 9), (5, 8, 7, 1, 9), (5, 8, 3, 1, 9), (5, 7, 3, 1, 9), (5, 8, 7, 3, 1, 9).

The different situations have the following probabilities:

$$
\begin{aligned}
(5, X) &: \left(\frac{1}{4}\right)^1 \times \left(\frac{3}{4}\right)^4 \\
(5, X, X) &: \left(\frac{1}{4}\right)^2 \times \left(\frac{3}{4}\right)^3 \\
(5, X, X, X) &: \left(\frac{1}{4}\right)^3 \times \left(\frac{3}{4}\right)^2 \\
(5, X, X, X, X) &: \left(\frac{1}{4}\right)^4 \times \left(\frac{3}{4}\right)^1 \\
(5, X, X, X, X, X) &: \left(\frac{1}{4}\right)^5
\end{aligned}
$$

This means that we can calculate the probability of the counter for number 5 having a count over 30 for one hashing function as:

$$
\begin{aligned}
&2 \times \left(\left(\frac{1}{4}\right)^1 \times \left(\frac{3}{4}\right)^4\right) + \\
&9 \times \left(\left(\frac{1}{4}\right)^2 \times \left(\frac{3}{4}\right)^3\right) + \\
&10 \times \left(\left(\frac{1}{4}\right)^3 \times \left(\frac{3}{4}\right)^2\right) + \\
&5 \times \left(\left(\frac{1}{4}\right)^4 \times \left(\frac{3}{4}\right)^1\right) + \\
&1 \times \left(\left(\frac{1}{4}\right)^5\right) \\
&= 0.158203125 + 0.2373046875 + 0.087890625 + 0.0146484375 + 0.0009765625 \\
&= 0.4990234375
\end{aligned}
$$

Because we are using two different hash functions we need to take the probability raised to 2. This yields $0.4990234375^2 = 0.2490244$. This means that the probability of $\hat{f}(5) > 30$ happening is 0.2490244.

Comparing this to the answer in the first part of the question we can see that the probability is a bit higher. We see that this probability is close to the bound for $\hat{f}_x \leq f_x + \frac{2}{k} \cdot (n - f_x)$ with a probability of at least $1 - \left(\frac{1}{2}\right)^t$. Probability is higher in part 2 because we see that the frequencies of different numbers have increased, which will lead to higher collisions. It might not always be like this but generally speaking more unique numbers with higher frequencies will increase the chance of collisions. The conclusion that can be drawn is that the distribution of numbers in a stream affects the accuracy of the COUNTMIN algorithm.

References

[1] loana Bercea, "Lecture 12: Estimating frequencies," Novemeber 2023.