Algorithm Foundations of DaSE

Name: [redacted] Student ID: [redacted] Credits: _____

Read all of the following information before starting the exam:

- Show all work, clearly and in order, if you want to get full credit. I reserve the right to take off points if I cannot see how you arrived at your answer (even if your final answer is correct).
- This test has 8 problems and is worth 100 points. It is your responsibility to make sure that you have all of the pages!
- Good luck!

1. (*10 points*) In n tosses of a coin with head probability $\frac{1}{4}$, let X be # heads, what is the upper bound of $P(X > \frac{n}{2})$ given by following inequalities?

a. Markovs inequality; $P(X>a) < \frac{E(X)}{a}$ $P(X>\frac{n}{2}) < \frac{E(X)}{\frac{n}{2}}$

b. Chebyshevs inequality;

c. Chernoff bound.

2. (*10 points*) Given the input streaming $< b, a, c, d, c, a, f, c, c, a, e >$, and $k = 4$, i.e., four counters.

- Please find the frequent items via using the Misra-Gries algorithm.
- Analyze the pros and cons of the Misra-Gries algorithm for finding the frequent items.

3. (*10 points*) Suppose you are creating an account on Tiktok, you want to enter a cool username, you entered it and got a message, "Username is already taken". You added your birth date along username, still no luck. Now you have added your university roll number also, still got "Username is already taken". It's really frustrating.

a. But have you ever thought how quickly Tiktok check availability of username by searching billions of usernames registered with it. Please provide your solution.

b. Analyze the pros and cons of the solution.

c. Enumerate some applications for your proposed solution.

4. (*10 points*) There is a web crawler that crawls a page every second. If you are asked to implement an API, each call uniformly selects a web page that has been crawled so far.

a. How should you implement the API with $O(1)$ space complexity?

b. And show how to guarantee the selected page evenly distributed from the crawled pages.

5. (*10 points*) Given the following count sketch algorithm:

$C[1...t][1...k] \leftarrow \vec{0}$, where $k = O(\frac{1}{\epsilon^2})$ and $t = O(\log(1/\delta))$;
Choose t random hash functions $h_1, \cdots, h_t : [n] \rightarrow [k]$ (uniformly);
Choose t random hash functions $g_1, \cdots, g_t : [n] \rightarrow \{-1, 1\}$ (uniformly);

Process item (j, c), where $c = 1$:

for i to t do $C[i][h_i(j)] \leftarrow C[i][h_i(j)] + cg_i(j)$;

Output:

On query a, report $\widehat{f_a} = \text{median}_{1 \le i \le t} g_i(a) C[i][h_i(a)]$;

By using Tug-of-War, the algorithm aims at boosting the accuracy of the item frequency estimation, it returns

$$\widehat{f_a} = \text{median}_{1 \le i \le t} g_i(a) C[i][h_i(a)]$$

for a query a. Please show your reason how to determine the value of t, i.e., $O(\log(1/\delta))$.

6. (*10 points*) A certain experiment is believed to be described by a two-state Markov chain with the transition matrix P, where $P = \begin{pmatrix} 0.5 & 0.5 \\ p & 1-p \end{pmatrix}$ and the parameter p is unknown. When the experiment is performed many times, the chain ends in state one approximately 20 percent of the time and in state two approximately 80 percent of the time.

a. Compute a sensible estimate for the unknown parameter p and explain how you found it;

b. Whether is the Markov chain irreducible and aperiodic, or not? Why?

反周期

7. (*20 points*) As shown in the following table, given a universal set U of seven elements, there are three subsets S_1, S_2 and S_3.

a. Compute the Jaccard similarity of each pair of columns.

b. Define a hash function $h_\pi(C)$ = the index of the last (in the permuted order π) row in which column C has value 1, denoted as maxhash. Given a random permutation π, whether is the following statement true or not? And show your reasons.

$$P\big(h_\pi(C_1) = h_\pi(C_2)\big) = Jaccard(C_1, C_2) = \frac{|C_1 \cap C_2|}{|C_1 \cup C_2|}.$$

c. Compute the maxhash signature for each column if we use the following hash functions: $h_1(x) = 3x + 1 \bmod 7$; $h_2(x) = 5x + 2 \bmod 7$; $h_3(x) = 9x + 4 \bmod 7$, where $x \in U$.

d. Compute similarity of each pair of sets via using the maxhash signatures.

Element	S_1	S_2	S_3
0	1	1	1
1	0	0	0
2	1	1	0
3	0	1	1
4	1	0	0
5	0	0	1
6	1	1	1

1 2 4
4 7 13
7 12 22
10 17 31
13 22 40
16 27 49
19 32 58

8. (*20 points*) Let $X_1, X_2, \cdots, X_n$ be i.i.d. Poisson sample (pmf of Poisson is $\frac{\lambda^x}{x!}e^{-\lambda}$)

a. If the sample is drawn from a uni-modal Poisson, please find the maximum likelihood estimation of λ;

b. If sample $X_1, X_2, \cdots, X_n$ is formed by mixing two different Poisson distributions, called Poisson mixture model, please give the likelihood of the sample, i.e., $L(\lambda_1, \lambda_2|\mathbf{x})$, where $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ is the values of the sample;

c. Please find a way to estimate the parameters of the Poisson mixture model, i.e., λ_1, λ_2;