CSIT 5500 Advanced Algorithms
2022 Fall Semester
Written Assignment 4
Handed out: November 14, 2022
Due: 23:59 on November 30, 2022
Please submit a soft copy via the canvas system by the due date and time shown above. Late assignments will not be graded.

1. (10 points) This question is about the Misra-Gries algorithm. The elements in the stream come from the range $[1, 9]$. Use $k = 5$. Run the Misra-Gries algorithm on the following stream (in left-to-right order).

$$6, 3, 8, 7, 6, 8, 4, 3, 8, 6, 5, 6, 9, 3, 4, 8, 6, 4, 8, 5, 6, 7, 3, 9, 8, 1, 8, 6, 3, 2, 9, 6, 8, 4, 6, 8, 7, 8, 2$$

   (a) Draw the table A and show its current content (including the elements in the table and their counts) after seeing each element in the stream.

   (b) Draw a table that shows the true frequencies and estimated frequencies (according to the Misra-Gries algorithm) of the values in the range [1,9]. Your table should be updated after seeing each element in the stream.

2. (10 points) This question is about using the count-min sketch to estimate the frequencies of elements in a stream. The elements in the stream come from the range $[0, 9]$. The count-min sketch maintains a two-dimensional array $C[0..k-1][0..\ell-1]$. (For convenience, we let the array indices run over the ranges $[0, k-1]$ and $[0, \ell-1]$ instead of $[1, k]$ and $[1, \ell]$ as given in the lecture notes.)

   Use $k = 2$ and $\ell = 5$. So there are two hash functions h_r for $r \in [0, k-1] = [0, 1]$. Use the following hash functions:

$$\begin{aligned} h_0(x) &= (2x + 3) \bmod 5 \\ h_1(x) &= (3x + 1) \bmod 5 \end{aligned}$$

   That is, given an element x in the stream, we map x to the array entries $C[0, h_0(x)]$ and $C[1, h_1(x)]$. Run the count-min sketch algorithm on the following stream (read from left to right):

$$3, 8, 9, 5, 7, 6, 1, 0, 8, 3, 7, 4, 6, 5, 1, 8, 3, 8, 2, 4, 7, 8, 0, 1, 2, 3, 9, 8, 0, 7, 4, 1, 0, 9, 3, 2, 7, 4, 9, 0$$

   Draw the final array C to show the values of its entries after processing all elements in the stream above. Use the count-min sketch to give the estimates $\hat{f}_a$ of the frequency of a in the stream for all $a \in [0, 9]$. You do not need to show any intermediate step.

3. (10 points) This question is about the q-digest. The elements in the stream come from the range $[1, 8]$. Use $k = 4$. Run the q-digest algorithm on the following stream. **Record the frequencies of all elements in the stream at the leaves of the complete binary tree before running the compression algorithm to identify the q-digest nodes.**

$$6, 5, 4, 6, 3, 7, 2, 5, 4, 8, 1, 3, 4, 3, 6, 1, 8, 2, 5, 4, 6, 7, 8, 1, 4, 6, 3, 7, 5, 4$$

   (a) Draw the final complete binary tree (not only the nodes in the q-digest) and the values stored at the tree nodes. You do not need to give any intermediate step.

(b) Suppose that we sort the elements in the above stream in non-decreasing order. Use your q-digest to estimate the 15th number in this sorted list.

4. (10 points) This question is about developing a randomized algorithm for solving linear programming in $\mathbb{R}^2$. Let $h_i : y_i = a_i x + b_i$, $i \in [1, n]$, be n non-vertical, non-parallel lines. Each line h_i bounds a halfplane h_i^+ bounded by h_i. The halfplane h_i^+ may be $y_i \geq a_i x + b_i$ or $y_i \leq a_i x + b_i$, depending on the input specification. Let $c_1 x + c_2 y + c_3$, where $c_1 \neq 0$ and $c_2 \neq 0$, be the objective function. We seek the point in $(x, y) \in \bigcap_{i=1}^{n} h_i^+$ that maximizes the value of $c_1 x + c_2 y + c_3$. We assume that $\bigcap_{i=1}^{3} h_i^+$ is a bounded triangle. Let $h_4, \cdots, h_n$ be a random permutation of the remaining $n - 3$ input lines other than h_1, h_2 and h_3. For $j \geq 3$, let $H_j = \bigcap_{i=1}^{j} h_i^+$. Let v_j be the point in H_j that maximizes the objective function $c_1 x + c_2 y + c_3$. Throughout this question, you can assume that the intersection point of two lines can be computed in $O(1)$ time.

   (a) (2 points) Prove that for $j \geq 4$, if $v_{j-1} \in h_j^+$, then $v_j = v_{j-1}$; otherwise, $v_j \in h_j$.

   (b) (2 points) Using the fact that $h_4, \cdots, h_n$ is a random permutation, develop a good bound on the probability of $v_j \in h_j$ for $j \geq 4$. Hint: If we consider these $n - 3$ lines without their labels $h_4, \ldots, h_n$, the random permutation acts like assigning the labels $h_4, \ldots, h_n$ to them randomly. So each of them has the same chance of getting the label h_4 and so on.

   (c) (2 points) Let L be an arbitrary line that intersects H_j. Design an algorithm to find the point in $L \cap H_j$ that maximizes the objective function $c_1 x + c_2 y + c_3$. Your algorithm should run in $O(j)$ time. Hint: Intersect L and each of the lines $h_1, \ldots, h_j$.

   (d) (2 points) Part (a) can be viewed as a test of whether $v_j = v_{j-1}$ or whether $v_j \in h_j$. Use parts (a) and (c) to design a randomized algorithm for solving the linear programming problem in $\mathbb{R}^2$. You should be able to reduce a two-dimensional problem to a one-dimensional subproblem.

   (e) (2 points) Analyze your algorithm to show that it runs in $O(n)$ expected time.