CSIT 5500 Advanced Algorithms
2020 Spring Semester
Written Assignment 5
Handed out: May 6, 2020
Due: 21:00 on May 19, 2020
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 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 $\ell = 3$. So there are two hash functions h_r for $r \in [0, k-1] = [0, 1]$. Use the following hash functions:

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

   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):

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

   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.

2. (10 points) This question is about the q-digest. The elements in the stream come from the range $[1, 8]$. Use $k = 3$. 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.**

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

   (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 20th number in this sorted list.

3. (10 points) Let $P = \{p_1, p_2, \ldots, p_n\}$ be $n \geq 3$ input points in $\mathbb{R}^2$. We assume that no three points in P are collinear. Let $f(x, y) = x^2 + y^2$.

   We want to find a point q in the convex hull of P that minimizes $f(x, y)$. Specialize the Frank-Wolfe algorithm to the problem above. Give the detailed specification of the resulting specialized algorithm. Suggest a stopping condition so that the your final solution is a $(1+\varepsilon)$-approximation for a given $\varepsilon \in (0, 1)$, that is, you obtain a point (x_0, y_0) such that $f(x_0, y_0)$ is at most $1 + \varepsilon$ times the optimum. Explain why your approximation satisfies this quality guarantee. Make sure that your solution satisfies the following requirements in addition to the requirements mentioned above.

- Derive explicit formulae for ∇f and $\langle \nabla f(x_t), z \rangle$.
- Describe explicitly how to compute $\operatorname{argmin}_{z \in D} \langle \nabla f(x_t), z \rangle$.
- Derive an explicit formula for $-\langle \nabla f(x_t), z_t - x_t \rangle$ and give an explicit stopping condition.
- Describe explicitly how to compute $\operatorname{argmin}_{\rho \in [0,1]} f(x_t + \rho(z_t - x_t))$.
- Derive the explicit value of the Lipschitz constant L and relates it to guaranteeing the $1 + \varepsilon$ approximation ratio.
- Make sure that every step of your algorithm can be implemented efficiently and explain how.