Institut für Informatik
Dr. Melanie Schmidt
Clemens Rösner

Cluster Analysis
Winter 2018/19

# Problem Set 4

Please hand in your solutions for this problem set via email (roesner@cs.uni-bonn.de) or personally at Room 2.060 until *Tuesday, 6th of November*.

## Problem 1

Recall the general idea of the $k$-center-via-thresholding algorithm by Hochbaum and Shmoys. Start by guessing the value of the optimal solution. This guess is called a threshold and denoted by $\tau$. For the guessed threshold we then use another routine which, in case $\tau$ is at least as big as opt finds a feasible solution, whose value is at most $2\tau$. In the $k$-center problem this works because we know that the optimal value has to be one of at most $|P|^2$ different values. We now want to take a look at different objectives. Given that we have a routine which can, for a threshold $\tau \geq opt$ and some $\alpha \geq 1$, compute a solution with a value of at most $\alpha\tau$, how could we effectively implement approximation algorithms that use the threshold idea for the following problems

- The $k$-median problem
- The Euclidean $k$-center problem, where we are given a finite set of points $P \subseteq \mathbb{R}^d$ for some $d \in \mathbb{N}$ and are allowed to use any point in $\mathbb{R}^d$ as a center.

## Problem 2

Let us take a look at the streaming model for the $k$-supplier problem. For that case we again assume that we are given elements of our metric space with the additional information if the element is a point or a location. We again assume that the elements appear in a stream and that we have a distance oracle, which can tell us the distance between two points. We are again not able to store all of the information we have seen and more specifically assume that we are only allowed to store $O(k)$ points and locations. Can we find an approximation algorithm for the $k$-supplier problem in the streaming model for the following cases? If not, please explain why it is not possible.

- Points and locations appear arbitrary.
- All locations appear at the beginning.
- All points appear at the beginning.
- The stream starts with $k$ locations, continues with all points and finishes with the rest of the locations.
- We assume that the set of points $P$ is fixed, but unknown and only locations appear on the stream. For each location $\ell$ which appears on the stream we are given the distance $d(\ell, P)$.

## Problem 3

Let us look at a different streaming problem. We are given a stream of points in the Euclidean space and want to randomly select one of the points, such that at any time the expected location of the selected point is the mean location of all points seen so far. Again we are only allowed to choose a point when we see it and are not allowed to store all the points we have seen so far.

- Describe a randomized algorithm which at any time $i$ during the stream has a point $p(i) \in \{p_1, \ldots, p_i\}$ selected and we have $E(p(i)) = \frac{p_1 + \ldots + p_i}{i}$, where $E(p(i))$ denotes the expected value of the location of $p(i)$.