# MCS 549 – Mathematical Foundations of Data Science
Fall 2021
Problem Set 3

Lev Reyzin

**Due**: 12/1/21 at the beginning of class

**Instructions:** Atop your problem set, please write your name and list your collaborators.

## Problems

**1.** Give a streaming algorithm to select symbol $i$ with probability proportional to $a_i^2$, where each $a_i$ is in $\{1, \ldots, m\}$.

**2.** Give an example of a set $H$ of hash functions such that $h(x)$ is equally likely to be any element of $\{0, ..., M-1\}$ but $H$ is not 2-universal. Prove your answer correct.

**3.** For the $k$-median and the $k$-means objectives, prove upper bounds on the ratio between the optimal value when we either require all cluster centers to be data points or allow arbitrary points (sometimes called "Steiner points") to be centers.

**4.** While most clustering problems are NP-Hard, it is possible to formulate clustering objectives which traditional algorithms can solve exactly in polynomial time. Given objects $p_1, \ldots, p_n$, and distances $d(,)$ on them (with $d(p_i, p_i) = 0, d(p_i, p_j) = d(p_j, p_i)$, and $d(p_i, p_j) > 0$ for $i \neq j$), consider the clustering problem of dividing the objects into $k$ sets so as to maximize the minimum distance between any pair of objects in distinct clusters. Give a polynomial time algorithm for solving this version of the clustering problem. (Note that $k$ is part of the input and can also grow with $n$.)

**5.** Fill in the details for the dynamic programming algorithm for clustering $n$ points on the line using $k$ clusters. Let $\text{OPT}(\ell, i)$ be the optimal clustering for points $a_1, \ldots, a_i$ using $\ell \leq k$ clusters for $i \leq n$. As part of your answer, make sure to write this as a function of "smaller" values of $\ell$ and $i$. Use this to derive the complexity of finding $\text{OPT}(k, n)$.