CSCI B609 – Foundations of Data Science **Fall 2017**

Homework 3: November 06

Name: YOUR NAME HERE *Due: Monday, November 20, 11:59pm EST*

Problem 3.1 (Approximate median) *You are given a stream of m elements $x_1, \ldots, x_m$ where $x_i \in \{1, \ldots n\}$ where m is an odd integer. For each element x_i in the stream $rank(x_i) = \{j | x_j \leq x_i\}$. Recall the median M is defined as the element x_i in the stream such that $rank(x_i) = m/2 + 1$. Consider an algorithm that takes a uniformly random sample of size t from the stream and computes M', the median value on this random sample.*

For $t = o(m)$ does the above algorithm give a 10%-approximate value of the true median with probability $\geq 2/3$, i.e. is it true that:

$$M - \frac{n}{10} \leq M' \leq M + \frac{n}{10} \text{ ?}$$

Explain your answer – if the algoirthm works provide the analysis. If it doesn't give a counterexample.

Problem 3.2 (Lower bound on F_k) *Recall the definition of a frequency vector f whose entries f_i correspond to the number of occurrences of an element i in the stream. The k-th frequency moment F_k is defined as $F_k = \sum_{i=1}^{n} f_i^k$. Show that for all integer $k \geq 1$ it holds that:*

$$F_k \geq n \left(\frac{m}{n}\right)^k,$$

where m is the length of the stream.

Hint: *Consider the worst case when $f_1 = f_2 = \cdots = f_n = \frac{m}{n}$. Use convexity of x^k to finish the proof.*

Problem 3.3 (Upper bound on the AMS estimator) *Recall the same definitions of f_i and F_k given in the previous problem. Recall that in the AMS estimator we sample an index j between 1 and m and define $r = |\{k \geq j : x_k = x_j\}|$. The AMS estimator is given as $X = m(r^k - (r-1)^k)$. Show that:*

$$X \leq mkf_\star^{k-1}, \text{ where } f_\star = max_{i=1}^{n} f_i.$$

Problem 3.4 (Exercise 7.9, modified) *Let p be a prime. A set of hash functions $H = \{h | \{0, 1, \ldots, p-1\} \to \{0, 1, \ldots, p-1\}\}$ is 3-universal if for all u, v, w, x, y, z in $\{0, 1, \ldots, p-1\}$ where x, y, z are distinct:*

$$\Pr_h[h(x) = u] = 1/p,$$

$$\Pr_h[h(x) = u, h(y) = v, h(z) = w] = \frac{1}{p^3}.$$

1. *Is the set $\{h_{ab}(x) = ax + b \bmod p | 0 \leq a, b < p\}$ of hash functions universal for a uniformly random choice of $a, b \sim \{0, 1, \ldots, p-1\}$?*
2. *Give a 3-universal set of hash functions.*