CMPSCI 711 SPRING 2012: HOMEWORK 1
DUE 2:05 PM, FEBRUARY 22ND

Rules: You may work in groups of at most four. Each group should submit (emailed or handed in at the start of class) one set of typed solutions (remember to include the names of everyone in the group). Cite all sources although work that uses published papers or material from related classes will receive no credit.

Question 1. *Given a stream S of m distinct values for a value $k \ll m$, suppose you want to find an element x such that $rank_S(x) = k \pm \epsilon k$. Consider drawing t samples (with replacement) from the data stream. Given these samples, how would you choose x? How large would t need to be for your choice to be suitable with probability at least $1-\delta$?*

Question 2. *The empirical entropy of a stream is defined as $H = \sum_{i=1}^{n} \frac{f_i}{m} \ln \frac{m}{f_i}$. Throughout this question, you should assume that $\max f_i \leq m/4$.*

(1) *Prove an upper and lower bound on H given the assumption.*
(2) *For any $r \in [m/4]$, prove an upper and lower bound on $r \ln(m/r) - (r-1)\ln(m/(r-1))$ given the assumption.*
(3) *Design a space-efficient (ϵ, δ) approximation algorithm for H based on AMS sampling. Remember to analyze the space required by the algorithm.*

Question 3. *In ℓ_2-sampling the goal is to return a random value $I \in_R [n]$ such that $\mathbb{P}[I = i] = (1 \pm \epsilon) f_i^2/F_2$. Design an simple, small-space stream algorithm for ℓ_2-sampling that takes $O(\log n)$ passes over the data stream. Hint: You can use an F_2 approximation algorithm as a subroutine.*

Question 4. *Let p be a prime number and let a, b be chosen uniformly at random from $\{0, 1, \ldots, p-1\}$. Define the hash-function $h : \{0, \ldots, p-1\} \to \{0, \ldots, p-1\}$ where $h(i) = ai + b \pmod{p}$. Prove that for any distinct $i, j \in \{0, \ldots, p-1\}$, $h(i)$ and $h(j)$ independent, i.e. for $\alpha, \beta \in \{0, 1, \ldots, p-1\}$*

$$\mathbb{P}[h(i) = \alpha, h(j) = \beta] = \mathbb{P}[h(i) = \alpha]\, \mathbb{P}[h(j) = \beta] \ .$$

Is the hash function 3-wise independent, i.e., for any distinct $i, j \in \{0, \ldots, p-1\}$ and all $\alpha, \beta, \gamma \in \{0, 1, \ldots, p-1\}$, is it true that $\mathbb{P}[h(i) = \alpha, h(j) = \beta, h(k) = \gamma] = \mathbb{P}[h(i) = \alpha]\, \mathbb{P}[h(j) = \beta]\, \mathbb{P}[h(k) = \gamma]$?

Question 5. *Prove that for any $1 \leq i \leq j \leq n$, the interval $[i, i+1, \ldots, j]$ can be partitioned into at most $2\log_2 n$ intervals of the form $[1 + k2^l, 2 + k2^l, \ldots, (k+1)2^l]$ where $k, l \in \mathbb{N}_0$. You may assume n is a power of 2.*

Question 6. *Suppose you may assume that there are at most k values of i such that $f_i > 0$. Adapt the CR-Precis sketch to find all (i, f_i) pairs where $f_i > 0$.*