CS60021: Scalable Data Mining 2019
Sample Questions: Streaming Algorithms and Sketches

1. Recall the *Count-Min* Sketch algorithm. Let $x_1, x_2, \ldots, x_n$ be the stream. Let f_j be the number of times element $j \in [m]$ appears in the stream. Let $\mathcal{F}$ be the family of pairwise independent functions $\{h : [m] \to [k]\}$. Let C be the sketch with t hash functions $\{h_1, \ldots, h_t\}$ picked up uniformly at random from $\mathcal{F}$. For $j \in [n] \setminus \{a\}$, let $Y_{i,j}$ be the excess in the counter $C[i][h_i(a)]$. i.e. $Y_{i,j} = C[i][h_i(a)] - f_a = f_a^{(i)} - f_a$, where $f_a^{(i)}$ is an estimate of f_a from the i^{th} hash function. Let $Y_i = \sum_{j=1}^{k} Y_{i,j}$. Our goal is to show that for $\delta, \epsilon > 0$, $\mathbb{P}[min_i Y_i > \epsilon n] \leq \delta$. We will do that step-by-step.

   So, what can you say about $\mathbb{P}[Y_i > \epsilon n]$? (*Hint: Calculate* $\mathbb{E}[Y_i]$ *and use Markov's inequality.*)

   A. $\mathbb{P}[Y_i > \epsilon n] \leq \frac{f_a}{k\epsilon n}$
   B. $\mathbb{P}[Y_i > \epsilon n] \leq \frac{n - f_a}{kn}$
   C. $\mathbb{P}[Y_i > \epsilon n] \leq \frac{n - f_a}{k\epsilon n}$
   D. $\mathbb{P}[Y_i > \epsilon n] \leq \frac{f_a^2}{k\epsilon n}$

   Ans: C.

2. Using the information from Q.8 choose an appropriate option for the question below: What is the value of k, if you want $\mathbb{P}[Y_i > \epsilon n] \leq \frac{1}{2}$?

   A. $k \geq \frac{2f_a}{\epsilon n}$
   B. $k \geq \frac{2(n - f_a)}{n}$
   C. $k \geq \frac{2(n - f_a)}{\epsilon n}$
   D. $k \geq \frac{2f_a^2}{\epsilon n}$

   Ans: C.

3. Using the information from Q.8 and Q.2 choose an appropriate option for the question below: What is the value of t, if you want $\mathbb{P}[min_i Y_i > \epsilon n] \leq \delta$?

   A. $t \leq \frac{(1/\delta)}{2}$
   B. $t \geq \frac{\log\ 2}{\log(1/\delta)}$
   C. $t \leq \frac{\log(1/\delta)}{\log 2}$
   D. $t \geq \log(1/\delta)$

Ans: C.

4. You have a stream of numbers and you want to get the approximate frequency of elements present in your stream. So, you implement Count sketch. But you make a "small" mistake while implementing it. The random function g which was supposed to be defined as $g : [n] \to \{-1, 1\}$, you mistakenly define it as $g : [n] \to \{-1/2, 1/2\}$. Your estimate for a still uses the same formulae as before.

   What do you think will be the expected value of the estimates you get for any element a i.e. $\mathbb{E}[f_a]$?

   A. $\mathbb{E}[f_a] = f_a$
   B. $\mathbb{E}[f_a] = f_a/2$
   C. $\mathbb{E}[f_a] = f_a^2$
   D. $\mathbb{E}[f_a] = f_a/4$

   Ans: D.

5. Suppose you are creating the SpaceSaving sketch with $k = 1$. Which of the following statements is true if you run this sketch over a stream of length m?

   (a) If there is an item with frequency at least $m/3$, then this item will be stored at the end.
   (b) We cannot say anything about which item will be stored at the end.
   (c) If there is an item with frequency $> m/2$ then this item will be stored at the end.

6. You have designed a Count-Min sketch. However, when returning the estimated frequency for a query, you forgot to return the *minimum* of the estimates, instead you returned the *median* (i.e. you return $\text{median}_i A[i, h_i(x)]$). You are running your code over a stream with only positive updates to frequencies. Then which of the following statements are *True*.

   A. Your estimates are hopelessly wrong, and do not satisfy the guarantees proved for Count-Min.
   B. Relax, you are still within the guarantees of Count-Min sketch, but your estimates could be somewhat worse in practice.
   C. Your estimates are strictly better than the ones obtained by Count-Min.
   D. None of the above are true.

7. Suppose instead of taking the median in the above question, you took the maximum of the estimators ($\max_i A[i, h_i(x)]$). Then which of the following statements are *True*.

   A. Now you should be worried, your estimates do not satisfy the guarantees proved for Count-Min.
   B. Relax, you are still within the guarantees of Count-Min sketch, but your estimates could be somewhat worse in practice.
   C. Your estimates are strictly better than the ones obtained by Count-Min.
   D. None of the above are true.

8. Recall the *Count-Min* Sketch algorithm. Let $x_1, x_2, \ldots, x_n$ be the stream. Let f_j be the number of times element $j \in [m]$ appears in the stream. Let $\mathcal{F}$ be the family of pairwise independent functions $\{h : [m] \to [k]\}$. Let C be the sketch with t hash functions $\{h_1, \ldots, h_t\}$ picked up uniformly at random from $\mathcal{F}$. For $j \in [n] \setminus \{a\}$, let $Y_{i,j}$ be the excess in the counter $C[i, h_i(a)]$. i.e.

$$Y_{i,j} = \begin{cases} f_j & \text{if } h_i(a) = h_i(j) \quad \text{(with probability } 1/k\text{)}, \\ 0 & \text{otherwise} \quad \text{(with probability } 1 - (1/k)\text{)} \end{cases}$$

i.e. $Y_{i,j}$ is the increment in f_a because of some other element $j \in [n] \setminus \{a\}$ for i^{th} hash function in the same bucket $h_i(a)$.
Let $Y_i = \sum_{j \in [n] \setminus \{a\}} Y_{i,j}$. Therefore, $\hat{f}_a$, the estimated frequency of a by the sketch, equals $\hat{f}_a = f_a + \underset{i}{min}\ Y_i$. What is the smallest value of t, such that for $\delta, \epsilon > 0$, we can claim $\mathbb{P}[\underset{i}{min}\ Y_i > \epsilon n] \leq \delta$.

(a) $t \leq \frac{(1/\delta)}{2}$
(b) $t \geq \frac{\log\ 2}{\log(1/\delta)}$
(c) $t \leq \frac{\log(1/\delta)}{\log 2}$
(d) $t \geq \log(1/\delta)$

9. You have a stream of numbers and you want to get the approximate frequency of elements present in your stream. So, you implement Count sketch. But you make a mistake while implementing it. While implementing you have implemented the random function g as $g : [n] \to \{\frac{-1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\}$. Your estimate for a is $X_a = g(a)C[h(a)]$.

What do you think will be the expected value of the estimates you get for any element a i.e. $\mathbb{E}[f_a]$?

(a) $\mathbb{E}[f_a] = f_a$
(b) $\mathbb{E}[f_a] = f_a/2$
(c) $\mathbb{E}[f_a] = f_a^2$
(d) $\mathbb{E}[f_a] = f_a/4$