CS 170 HW 13

Due on 2018-04-29, at 9:59 pm

1 Study Group

List the names and SIDs of the members in your study group.

2 One-to-One Functions

Suppose that $\mathcal{H}$ is a pairwise independent hash family that maps the elements $U = \{1, \ldots, n\}$ to $1, \ldots, n^3$. Show that the probability that a randomly chosen hash function h from $\mathcal{H}$ is one-to-one is at least $1 - 1/n$.

Solution: Consider the probability that a random function drawn at uniform from $\mathcal{H}$ is one to one:

$$\begin{aligned}
\Pr_{h\in\mathcal{H}}(h \text{ is one-to-one}) &= \Pr_{h\in\mathcal{H}}(h(x_1) \neq h(x_2) \neq \cdots \neq h(x_n)) \\
&= 1 - \Pr_{h\in\mathcal{H}}(\bigcup_{i\neq j} h(x_i) = h(x_j)) \\
&\overset{\zeta_1}{\geq} 1 - n(n-1)\left[\Pr_{h\in\mathcal{H}}(h(x_i) = h(x_j))\right] \\
&\overset{\zeta_2}{\geq} 1 - n(n-1)\left[\frac{n^3}{n^6}\right] \\
&\geq 1 - \frac{1}{n},
\end{aligned}$$

where ζ_1 follows from a union bound over $n(n-1)$ pairs of (i, j) and ζ_2 follows from the definition of pairwise independent hash family for any x_i, x_j.

3 Approximate Median

Let $S = (x_1, x_2, \ldots, x_m)$ denote a stream of m elements. For simplicity, assume that the elements in the stream are unique. Define the *position* of an element to be $\text{pos}(x) = |\{y \in S | y \leq x\}|$. The ϵ-approximate median is then defined to be an element x such that:

$$\frac{m}{2} - \epsilon m \leq \text{pos}(x) \leq \frac{m}{2} + \epsilon m. \tag{1}$$

Provide an algorithm which returns a ϵ-approximate median with high probability. Provide a 3-part solution providing the algorithm, proof of correctness and the space complexity of the algorithm. Note that the optimal algorithm can solve the problem in space independent of the size of the stream. (Hint: Try to provide a sampling based algorithm and argue that less than $1/2$ of the samples will be from the $1/2 - \epsilon$ and $1/2 + \epsilon$ percentile – use a Hoeffding bound for this argument). The following is a one-sided version of the Hoeffding bound which

might be useful for this and other problems:

$$\Pr\left(\frac{1}{t}\sum_{i=1}^{t} X_i - p \geq \epsilon\right) \leq \exp\left(-2\epsilon^2 t\right).$$

Solution: The algorithm that we will use is as follows: Sample t points from the stream uniformly at random using the Reservoir Sampling algorithm from class and return the median of the sampled t points. Setting $t = \frac{2}{\epsilon^2}\log\left(\frac{2}{\delta}\right)$ ensures that with probability atleast $1-\delta$, the sample median will be ϵ-approximate. Thus the space complexity of our algorithm is $t \cdot \log(|\Sigma|)$ where $|\Sigma|$ is the size of the sample space.

In order to argue about its correctness, let us partition the set S into three parts:

$$\begin{aligned} S_L &:= \{x | \text{pos}(x) \leq \frac{m}{2} - \epsilon m\} \\ S_M &:= \{\{x | \text{pos}(x) \geq \frac{m}{2} - \epsilon m \quad \& \quad \text{pos}(x) \leq \frac{m}{2} + \epsilon m\} \\ S_R &:= \{x | \text{pos}(x) \geq \frac{m}{2} + \epsilon m\}, \end{aligned}$$

where S_L denotes the set of points to the left of the approximate median and S_R denotes the set of points to the right of approximate median. Note that our algorithm will return a correct solution if the number of points from S_R and S_L is less than $t/2$ in the t samples that were collected. We will now show that these are low probability events. Let us start with S_L, the proof for S_R follows similarly.

Let Z_i denote a random variable such that $Z_i = 1$ if the i^{th} sample is from S_L and $Z_i = 0$ otherwise. We are interested in showing that $\Pr(\sum_{i=1}^{t} Z_i > t/2)$ is small. Note that $\mathbb{E}[Z_i] = 1/2 - \epsilon$ since the probability that a uniformly sampled element is in S_L is $1/2 - \epsilon$. Using the Hoeffding bound from class, we have:

$$\begin{aligned} \Pr\left(\sum_{i=1}^{t} Z_i > t/2\right) &= \Pr\left(\frac{1}{t}\sum_{i=1}^{t} Z_i - \left[\frac{1}{2} - \epsilon\right] > \epsilon\right) \\ &\leq \exp\left(-2\epsilon^2 t\right), \end{aligned}$$

and therefore setting the value of $t = \frac{2}{\epsilon^2}\log\left(\frac{2}{\delta}\right)$ gives us that with probability atleast $1 - \delta/2$, the number of elements from S_L will be less than $t/2$. We can argue similarly for S_R and combining those two completes our proof.

4 Count-Median-Sketch

Consider the *Count-Min-Sketch* algorithm from class for determining heavy hitters in a stream. Suppose that we change the algorithm to consider $\text{median}_{i=1,\ldots,l} M[i, h_i(x)]$ instead of the min function.

(a) Give an argument that the estimator for heavy hitters is still correct.

(b) If we allow the stream to both insert and delete elements, show that the median based algorithm works to find heavy hitters. Each item x in the stream now comes with a an element $\Delta \in \{+1, -1\}$ indicating whether it is an addition operation or deletion. Note that the number of deletions could be more than the number of additions.

(c) Would *Count-Min-Sketch* work for the setup above?

Solution:

(a) As pointed out in the notes, $\mathbb{E}(M[i, h_i(a)]) \leq f_a + \frac{n}{B}$, where n is the length of the stream and B is the size of the dictionary to which the hash functions h_i map. Using Markov inequality, we get that:

$$\Pr\left[M[i, h_i(a)] \geq f_a + \frac{3n}{B}\right] \leq \frac{1}{3}. \tag{2}$$

For $i = \{1, \ldots, l\}$, let Z_i be a Bernoulli variable such that $Z_i = 1$ if $M[i, h_i(a)] \geq f_a + \frac{3n}{B}$. For the median estimator to work, we need to ensure that $\Pr[\sum_i Z_i > l/2]$ is small. Using a Hoeffding bound, we get:

$$\begin{aligned}
\Pr\left(\sum_i Z_i > l/2]\right) &= \Pr\left(\frac{1}{l}\sum_i Z_i - \mathbb{E}(Z_i) > \frac{1}{2} - \mathbb{E}(Z_i)\right) \\
&\leq \Pr\left(\frac{1}{l}\sum_i Z_i - \mathbb{E}(Z_i) > \frac{1}{2} - \frac{1}{3}\right) \\
&\leq \Pr\left(\frac{1}{l}\sum_i Z_i - \mathbb{E}(Z_i) > \frac{1}{6}\right) \leq \exp\left(-\frac{l}{18}\right),
\end{aligned}$$

where the first inequality follows from Equation (3). Setting $B = 30$ and $l = 36 \cdot \ln n$ gives us that with probability atleast $1 - 1/n^2$, the estimate $\text{median}_{i=1,\ldots,l} M[i, h_i(x)]$ is within f_a and $f_a + 0.1n$.

Intuition. As above, we begin by using Markov's inequality to get:

$$\Pr\left[M[i, h_i(a)] \geq f_a + \frac{3n}{B}\right] \leq \frac{1}{3}. \tag{3}$$

For the median estimator to fail, we would require at least half of $M[i, h_i(a)$ to be greater than $f_a + \frac{3n}{B}$ since this would mean that our estimate of its frquency is higher than $f_a + \frac{3n}{B}$. However, observe that the probability that any specific hash function fails is $1/3$. So overall, one would expect the total number of failures in l hash functions to be around $l/3 < l/2$. This statement indeed would hold with high probability and hence the median would output the correct answer.

(b) In this deletion setup, the *Count-Median-Sketch* algorithm would still perform well. The idea is that for any heavy hitter item a, for the median algorithm to fail, more than half of the l hash functions need to have collisions with other items for which the number of deletions is more than the number of insertions. The hash functions are designed to have a low probability of collision and hence, having more than $l/2$ collisions is a very low-probability event.

(c) No, there is a high chance for *Count-Min-Sketch* to fail for the setup above. Consider any item a which is a heavy hitter. If the minimum entry of the hash table corresponding to this item clashes with another item b and if the number of deletions of b are more than insertions by a margin of $0.2n$, the algorithm will not output a as a heavy hitter.