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 $1 - 1/n$.

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.

Note: Your algorithm should take ϵ and the failure probability δ as parameters, and return a result that is ϵ-approximate with probability at least $1 - \delta$. The proof of correctness should prove that your result is ϵ-approximate with this probability.

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 using a Hoeffding bound.

Hint: The following one-sided version Hoeffding bound might be useful for this (and other) problems: If $X_1, \ldots X_n$ are i.i.d Bernoulli with $\mathbb{E}(X_i) = p$, we have that:

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

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.

Note: We are not looking for a rigorous proof here. Just an argument is fine.

(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.

Note: We are not looking for a rigorous proof here. Just an argument is fine.

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