CSCI 6220/4030: Substitute quiz

Assigned Monday November 11 2019. Due at beginning of class Monday November 25 2019. If your grade on this quiz is higher than your quiz average so far (viz., after the first 6 quizzes), this grade will replace your quiz average on the first 6 quizzes.

Remember to typeset your submission, and label it with your name. Please start early so you have ample time to see me during office hours. Provide mathematically convincing arguments for the following problems. Ask me if you are unclear whether your arguments are acceptable.

1. [**CSCI 4030 students**] Recall that one way to achieve perfect hashing is to use opoen addressing to hash the items (see §5.8 of Jeff Erichson's notes on hashing). Suppose that we are using an open-addressed hash table of size m to store n items, where $m \geq 2n$. Assume an ideal random hash function. For any i, let X_i denote the number of probes required for the ith insertion into the table, and let $X = \max_i X_i$ denote the length of the longest probe sequence.

   (a) Prove that $\mathbb{P}[X_i > k] \leq \frac{1}{2^k}$ for all i and k.

   (b) Prove that $\mathbb{P}[X_i > 2\log_2 n] \leq \frac{1}{n^2}$ for all i.

   (c) Prove that $\mathbb{P}[X > 2\log_2 n] \leq \frac{1}{n}$.

   (d) Prove that $\mathbb{E}[X] = O(\ln n)$.

2. [**CSCI6220 students**] Remember that one of the uses of randomization is to reduce the storage space of algorithms.

   As an illustration of that fact, we will design an algorithm that, given a streaming view of n arbitary items in some universe U, returns an accurate estimate of the number of unique items d in that stream, using a near optimal amount of storage space.

   For simplicity, we assume that $U = \{0, \ldots, n-1\}$ and that n is known, is greater than 2, and is a power of 2. We will also assume that we have access to an ideal hash function from U to U[1].

   (a) Argue that any algorithm that computes d exactly must use at least $\log_2 n$ bits of storage (recall that d is unknown).

   (b) Let zeros(x) denote the number of zeros that precede the first 1 in the binary expansion of x. For example,

   $$\begin{aligned} \text{zeros}(1) &= \text{zeros}(001_2) = 0 \\ \text{zeros}(2) &= \text{zeros}(010_2) = 1 \\ \text{zeros}(4) &= \text{zeros}(100_2) = 2 \\ \text{zeros}(5) &= \text{zeros}(101_2) = 0. \end{aligned}$$

   Let d random numbers be sampled i.i.d. uniformly at random from U. Show that, for any $z \in \{0, \ldots, \log_2 n\}$, the expected amount of these numbers that satisfy $\{\text{zeros}(x) \geq z\}$ is $\frac{d}{2^z}$.
   HINT: Recall that randomly sampling an integer in $0, \ldots, n-1$ is equivalent to randomly sampling the $\log_2 n$ bits in its binary expansion.

   (c) The above observation tells us that *if* the d unique items making up our stream were selected i.i.d uniformly at random from U, then we could estimate d as 2^z, where z is the largest zeros(a_i) of the items a_i in our stream. This is not generally the case, but we can use an ideal hashing function h to randomize our stream.

[1] This is unrealistic, but makes the analysis simple. One can show that a 2-universal hash function suffices, with more effort; further, there exist such hash functions that can be stored using $O(\log_2 n)$ bits. The algorithm in this problem becomes both practical and space-optimal if such a hashing function is selected.

Consider the following algorithm that reads in a stream of n items a_i and employs h to calculate an estimate $\hat{d}$ of d:

```
$z \leftarrow 0$
**for** $i \leftarrow 1 \dots n$ **do**
    $z \leftarrow \max\{z, \mathrm{zeros}(h(a_i))\}$
$\hat{d} \leftarrow 2^{z+1}$
```

Argue that $\mathbb{P}[\hat{d} \leq d] \leq \mathrm{e}^{-\frac{1}{2}}$ by expressing the event that $z \leq \log_2(d) - 1$ as the event that a sum of d independent indicator r.v.s is less than or equal to 1.

HINT: since $z = \max_{i=1,\dots,d} \mathrm{zeros}(h(a_i))$, it is smaller than a number if and only if all the $\mathrm{zeros}(h(a_i))$ are smaller than that number.

Similarly show that $\mathbb{P}[\hat{d} \geq 4d] \leq \mathrm{e}^{-\frac{1}{6}}$ by expressing the event that $z \geq \log_2(d) + 1$ as the event that a sum of d independent indicator r.v.s is greater than or equal to 1.

Conclude that the estimator $\hat{d}$ is accurate, i.e. satisfies

$$\hat{d} \in [d, 4d]$$

with probability at least $1 - 2\mathrm{e}^{-1/6}$.

(d) The above algorithm uses optimal, $O(\log_2 n)$, storage space and has constant failure probability. Explain how to use $O(\log_2^2 n)$ space to reduce the failure probability to $\frac{1}{n}$. That is, give an explicit algorithm in the above format, and analyze its failure probability.

HINT: maintain multiple independent estimators and return their median.