Lecturer: Thomas Rothvoss Due date: Friday, Oct 18, 2024, 11pm, on GradeScope

Problem Set 3

CSE 521 - Design and Analysis of Algorithms

Fall 2024

Exercise 1 (10pts)
Suppose we have a universe U of elements. For $A,B \subseteq U$, the *Jaccard distance* of A,B is defined as

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

This definition is used in practice to calculate a notion of similarity of documents, webpages, etc. For example, suppose U is the set of English words, and any set A represents a document considered as a bag of words. Note that for any two $A,B \subseteq U$, $0 \leq J(A,B) \leq 1$. If $J(A,B)$ is close to 1, then we can say $A \approx B$.

Let $h : U \to [0,1]$ where for each $i \in U$, $h(i)$ is chosen uniformly and independently at random from $[0,1]$. For a set $S \subseteq U$, let $h_S := \min_{i \in S} h(i)$. Show that

$$\Pr[h_A = h_B] = J(A,B).$$

Exercise 2 (Optional 0 points!)
Let $X_1,\dots,X_n$ be independent random variables uniformly distributed in $[0,1]$ and let $Y = \min\{X_1,\dots,X_n\}$. Show that $\mathbb{E}[Y] = \frac{1}{n+1}$ and $\mathrm{Var}(Y) \leq \frac{1}{(n+1)^2}$.
Comment. This homework problem is optional. You do not need to solve it for points; only if you are interested. However the claim itself might be useful in Exercise 3 and may be used there without proof.

Exercise 3 (10pts)
Consider the following algorithm for estimating F_0, the number of distinct elements in a sequence $x_1,\dots,x_m$ in the set $\{0,1,\dots,n-1\}$. Let $h : \{0,1,\dots,n-1\} \to [0,1]$ s.t., $h(i)$ is chosen uniformly and independently at random in $[0,1]$ for each i. We start with $Y = 1$. After reading each element x_i in the sequence we let $Y = \min\{Y, h(x_i)\}$.

i) Show that by the end of the stream $\frac{1}{\mathbb{E}[Y]} - 1$ is equal to F_0.

ii) Use the above idea to design a streaming algorithm to estimate the number of distinct elements in the sequence with multiplicative error $1 \pm \varepsilon$. For the analysis you can assume that you have access to k independent hash functions as described above. Show that $k \leq O(1/\varepsilon^2)$ many such hash functions is enough to estimate the number of distinct elements within $1+\varepsilon$ factor with probability at least $9/10$ (where $0 < \varepsilon < 1$).

Hint. There is a fact that may come in handy that you are allowed to use without proof: For any $\alpha > 0$ and any $0 < \beta \leq \frac{1}{2}$ and $0 < \varepsilon \leq \frac{1}{10}$ one has

$$\Big((1-\varepsilon)\alpha \leq \beta \leq (1+\varepsilon)\alpha\Big) \Longrightarrow \Big((1-4\varepsilon)\Big(\frac{1}{\alpha}-1\Big) \leq \frac{1}{\beta} - 1 \leq (1+4\varepsilon)\Big(\frac{1}{\alpha}-1\Big)\Big)$$