MA421 Advanced Algorithms — Lent 2019

Exercises 8

1. Let $\langle x_1, x_2, \ldots \rangle$ be a data stream of unknown length, where each $x_i \in \{1, 2, \ldots, n\}$. Devise an algorithm to produce a single sample s after one pass over the stream, such that the sample is distributed proportionally to its value: that is,

$$\Pr[s = x_i] = \frac{x_i}{\sum_i x_i}.$$

   Argue the correctness of your algorithm.

2. For a set S and $x \in S$, we define $\text{rank}_S(x) = |\{y \in S \mid y \leq x\}|$. Suppose, given a data stream S of m distinct values, you want to find an approximate median: an element x such that $\text{rank}_S(x) = \frac{m}{2} \pm \epsilon m$, for some ϵ such that $0 < \epsilon < 1/4$ (that is, $|\text{rank}_S(x) - m/2| \leq \epsilon m$). Consider drawing t samples (with replacement) from the data stream. Given these samples, how would you choose x? How large would t need to be for your choice of x to be suitable with probability at least $1 - \delta$? What if you wanted x such that $\text{rank}_S(x) = k \pm \epsilon k$, for some $k \ll m$?

3. Let $Z \in \{0, 1\}^{k \times n}$ be a fixed matrix and $f = (f_1, \ldots, f_n)$ denote the frequency vector of a data stream: that is, f_i is the number of times value i appears in the stream. Describe a data stream algorithm to compute the product $Z \cdot f$ exactly, using $O(k \log m)$ bits of memory (in addition to the memory required to hold Z) for a stream of length m.

4. Suppose your data stream algorithm is allowed to make two passes over the data (possibly retaining any memory content in between the passes). Describe and analyse a data stream algorithm to find the exact median of the stream using only $O(n^{2/3} \cdot \ln(1/\delta))$ storage, where n is the known length of the data stream and δ is the error probability of the algorithm.

Solutions to these exercises are to be handed in at the lecture on **12 March 2019**.