Karl Bringmann and Vasileios Nakos **Summer 2020**

Sublinear Algorithms, Exercise Sheet 3

www.mpi-inf.mpg.de/departments/algorithms-complexity/teaching/summer20/sublinear-algorithms/

Total Points: **40** Due: 12:00 (noon), Monday, **June 29**, 2020

*You are allowed to collaborate on the exercise sheets, but you have to write down a solution on your own, **using your own words**. Please indicate the names of your collaborators for each exercise you solve. Further, cite all external sources that you use (books, websites, research papers, etc.).*

You need to collect at least 50% of all points on exercise sheets to be admitted to the exam.

Exercise 1 — 10 + 10 points

In this exercise we will focus on a particular pattern of infected persons: Let $n+1$ be a power of two and arrange the universe $[n]$ as a binary tree (see Figure 1). We say that a subset $S \subseteq [n]$ is a *tree group* if S forms a connected subtree rooted at 1 (again, see Figure 1).

Your task is to show that we can achieve improvements for non-uniform group testing, both in terms of measurements and time, if the set of k infected persons is guaranteed to be a tree group. One way of motivating this scenario is that person 1 is the first patient infected by the disease and the tree structure traces the recorded contact persons (the “chain of infections”). In what follows, let $S \subseteq [n]$ be a fixed tree group of k infected persons.

1. Show that in the usual group testing framework, setting $R = O(1)$ suffices to compute a set $T \supseteq S$ of size $O(k)$ with probability 9/10. The running time of your recovery algorithm should be bounded by $O(k)$.

   Hint: For the recovery algorithm you should not use the vanilla Find$(\cdot)$ *procedure we discussed in class, but rather use the tree structure. In order to argue correctness, bound the probability that some healthy person is only contained in infected groups (i.e., is a false positive), and observe that these events are independent for all healthy persons. Then compute the expected size of the connected subtrees containting only false positives.*

2. Show that by setting $R = O(\log k)$, we can exactly recover S with probability $\frac{2}{3}$. The running time of your recovery algorithm should be bounded by $O(k \log k)$.

   Hint: Note that your algorithm cannot be adaptive. In particular, if you decide to compute a set $T \supseteq S$ using part 1, you cannot simply test all $O(k)$ candidates in T.

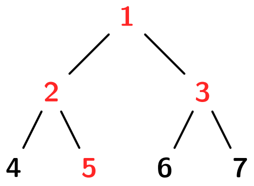

Figure 1. The set [7] visualized as a binary tree. The subset of red numbers $S = \{1, 2, 3, 5\}$ forms a tree group.

Exercise 2 — 10 points

Recall the CountMin algorithm from the lecture, phrased as a measurement problem; the goal is to compute a point-wise approximation to a length-n vector x using few linear measurements:

Measurements: For each $r = 1, \ldots, R$, we execute the following steps:

- Pick a random hash function $h_r : [n] \to [t]$, for $t = \lceil 50\varepsilon^{-1} \rceil$.
- For each group $b = 1, \ldots, t$, perform a linear measurement $y_{r,b} := \sum_{i \in h_r^{-1}(b)} x_i$.

Recovery: Return $\tilde{x}$ with $\tilde{x}_i = \text{median}_r \, y_{r,h_r(i)}$.

In the lecture we proved that for $R = O(\log n)$, $\tilde{x}_i$ is an accurate approximation of x_i with probability $\frac{2}{3}$, that is, $|x_i - \tilde{x}_i| \leq \varepsilon \|x\|_1$ for all coordinates i.

Let $x[-k]$ denote the vector obtained by zeroing out the k largest entries (in absolute value) in x, breaking ties arbitrarily. Prove that in fact that the stronger guarantee $|x_i - \tilde{x}_i| \leq \varepsilon \| \, x[-\varepsilon^{-1}] \, \|_1$ holds, again for $R = O(\log n)$ repetitions and with success probability $\frac{2}{3}$.

Exercise 3 — 10 points

We adapt the CountMin sketch by incorporating random signs:

Measurements: For each $r = 1, \ldots, R$, we execute the following steps:

- Pick a random hash function $h_r : [n] \to [t]$, for $t = \lceil 50\varepsilon^{-1} \rceil$.
- Pick another random function $s_r : [n] \to \{-1, 1\}$.
- For each group $b = 1, \ldots, t$, perform a linear measurement $y_{r,b} := \sum_{i \in h_r^{-1}(b)} s_r(i) \cdot x_i$.

Recovery: Return $\tilde{x}$ with $\tilde{x}_i = \text{median}_r \, s_r(i) \cdot y_{r,h_r(i)}$.

Prove that by setting $R = O(\log n)$, with probability $\frac{2}{3}$ we obtain the following strictly stronger guarantee on the recovery of x: $|x_i - \tilde{x}_i| \leq \varepsilon^{1/2} \| \, x[-\varepsilon^{-1}] \, \|_2$ for all coordinates i.