Exercise II, Sublinear Algorithms for Big Data Analysis 2024-2025

These exercises are for your own benefit. Feel free to collaborate and share your answers with other students, and solve as many problems as you can. Problems marked (*) are more difficult, but also more rewarding. These problems have been taken from various sources on the Internet, too numerous to cite individually.

1 Show that for large enough integer n there exists a collection C of subsets of $[n] := \{1, 2, 3, \ldots, n\}$ such that

1. $|C| \geq 2^{cn}$ for some absolute constant $c > 0$;
2. every element $S \in C$ is of size $n/4$;
3. for every two $S_1 \neq S_2 \in C$, $|S_1 \cap S_2| \leq n/8$

Solution. We show a probabilistic proof of the existence of such a family C. Create a collection $\mathcal{F}$ with each set constructed as follows: i) divide up $[n]$ into $n/4$ segments with each segment having 4 elements ii) select one element from each segment at random.

For any two sets S_1 and S_2, probability that they chose the same element from any given segment is $1/4$. Indicate this event by I_j for $j \in [n/4]$, then $\mathbf{E}[\sum_{i=1}^{n/4} I_j] = n/16$. We will use the multiplicative form of Chernoff bounds valid for all $\delta > 0$ (refer solution of excercise 1, Problem 3a).

$$\textbf{Prob}\left[\sum_{i=1}^{n/4} I_j > 2(n/16)\right] \leq \left(\frac{e}{4}\right)^{n/16} < 2^{-c_1 n} \text{ for some } c_1 > 0$$

Now let $\mathcal{F}$ contain $2^{(c_1/4)n}$ sets constructed as above. Then the probability that the intersection of at least two of such sets is larger than $n/8$ is upper bounded, by a union bound over all $\binom{2^{c_1/4}}{2} \leq 2^{(c_1/2)n}$ pairs of sets, by $2^{(c_1/2)n} \cdot 2^{-c_1 n} < 1$. Thus, we have that $\mathcal{F}$ satisfies the properties required for C with positive probability, proving existence of the required collection C. □

2 Use the result of **1** to show that any **deterministic** algorithm that achieves a 1 ± 0.1 approximation to the number of distinct elements in a data stream must use $\Omega(n)$ space.

Solution. Suppose towards a contradiction that there exists a deterministic algorithm ALG that uses less than cn bits of space, where $c > 0$ is the constant from Problem **1**, and computes a 1 ± 0.1 approximation to the number of distinct elements in a data stream. Since ALG can only be in strictly fewer than 2^{cn} memory states, there exists a pair of distinct sets $S_1, S_2 \in C$, where C is the collection of sets from Problem **1** such that the memory state of ALG after processing a stream of elements in S_1 (in increasing order, say) is the same as its memory state after processing elements of S_2 (again, in increasing order).

Now consider two streams $\sigma_1 = (S_1, S_1)$ and $\sigma_2 = (S_2, S_1)$. Both streams consist of two phases. In σ_1 elements of S_1 arrive in the first phase (in increasing order) and then elements of S_1 arrive again in the second phase (for concreteness, in increasing order). In σ_2 elements of S_2 arrive in the first phase (in increasing order) and then elements of S_1 arrive again in the second phase (for concreteness, in increasing order). Since the state of ALG's memory after processing the first phase of σ_1 is the same as after processing the first phase of σ_2, ALG outputs the same answer on both of these streams. However, the number of distinct elements in σ_1 is $n/4$ (every element of S_1 arrived twice, and S_1 contains $n/4$ elements), but the number of distinct elements in S_2 is $|S_1| + |S_2| - |S_1 \cap S_2| \geq n/4 + n/4 - n/8 \geq (3/8)n = (3/2) \cdot n/4$. Thus, an algorithm that provides a 1 ± 0.1 approximation to the number of distinct elements must output different answers on σ_1 and σ_2, leading to a contradiction. □