Exercise VIII, 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 (Lower bounds for ℓ_0 samplers) Recall from class that an ℓ_0 sampler is a randomized linear sketch $A \in \mathbb{R}^{m \times n}$ together with a decoding algorithm $D : \mathbb{R}^m \to [n] \cup \{\perp\}$ such that for every fixed $x \in \mathbb{R}^n$ one has

$$\textbf{Prob}[\text{decoding } D(A(x)) \text{ fails or } D(A(x)) = \perp] \leq \delta$$

for some $\delta \in (0, 1)$ and for all $j \in \text{supp}(x)$

$$\textbf{Prob}[D(A(x)) = j | \text{ decoding } D(A(x)) \text{ succeeds and } D(A(x)) \neq \perp] = \frac{1}{|\text{supp}(x)|}.$$

In both cases above the probability is over the random string R used to generate A (we write A_R to denote the matrix A generated using random string R). *Note that it is crucial that A is independent of x.*

Also note that the ℓ_0 sampler sketch construction that we saw in class uses a matrix A with integer entries that are polynomially bounded in n. Thus, if $x \in \mathbb{R}^n$ has integer entries bounded in absolute value by $n^{O(1)}$, the product Ax can be stored using $O(m \log n)$ bits. Show that any ℓ_0 sampler that uses a matrix A with polynomially bounded entries and succeeds with probability $1 - \delta$ as above must use $m = \Omega(\log(1/\delta))$ rows for every $\delta > 1/n^{10}$.

1a Show that there exists a family $\mathcal{F}$ of subsets $S \subseteq [n]$ of size $\Omega(\log(1/\delta))$ and a random string R such that **(a)** $\log_2 |\mathcal{F}| = \Omega(\log n \cdot \log(1/\delta))$ and **(b)** every $S \in \mathcal{F}$ can be uniquely recovered from $A_R \mathbf{1}_S$, where $\mathbf{1}_S \in \mathbb{R}^n$ stands for the indicator vector of S:

$$(\mathbf{1}_S)_j = \begin{cases} 1 & \text{if } i \in S \\ 0 & \text{o.w.} \end{cases}$$

Hint: use linearity of the sketch to perform iterative recovery of S, and be very careful when computing success probability

Solution. Let $\mathcal{F} = \{S \subseteq [n] || S| = -c \log \delta\}$ be the set of all subsets of $[n]$ of size $-c \log \delta$ for some constant $0 < c < 1$ to be set later.

$$\begin{aligned}
\log |\mathcal{F}| &= \log \binom{n}{-c \log \delta} \\
&= \log \left(\frac{n!}{(n + c \log \delta)!(-c \log \delta)!} \right) \\
&= \log \left(\frac{n \cdot (n-1) \ldots (n + c \log \delta)}{(-c \log \delta)!} \right) \\
&\geq c \log \frac{1}{\delta} \log \left(\frac{n}{c \log \frac{1}{\delta}} \right) \quad \left(\text{since } \frac{n-i}{k-i} \geq \frac{n}{k} \text{ for } i \leq k \leq n \right) \\
&\geq \Omega(\log n \log \frac{1}{\delta}) \quad (\text{assuming} \log \frac{1}{\delta} \leq n^{9/10})
\end{aligned}$$

Given any S of such a family, use $A_R 1_S$ as a sketch of the subset S where A_R is an instance of the l_0 sampling matrix. To recover the subset, use the l_0 sampling algorithm to recover i such that $1_S(i) = 1$ i.e. $i \in S$. Then remove i from S using the linearity of the sketch i.e. compute $A_R 1_{S \setminus i} = A_R 1_S - A_R 1_i$. Repeat the step $\log \frac{1}{\delta}$ times till all elements are recovered.

The probability that the above procedure will fail is bounded by the probability that the step fails at some step i. However we cannot just use union bound on the $\log \frac{1}{\delta}$ steps since we are using the same randomness R in all the steps. Let $E_R(S)$ indicate the event that the l_0 sampler using randomness R fails to sample an element of S. Let S_i denote the subset of S at step i during out procedure.

$$\mathbf{Prob}(\text{Recovery fails}) = \mathbf{Prob}_R(\cup_{i \in [|S|]} E_R(S_i)) \leq \mathbf{Prob}_R(\cup_{s \subseteq S} E_R(s)) \leq \delta 2^{|S|} = \delta^{1-c}$$

Thus the procedure outlined succeeds with probability $1 - \sqrt{\delta} = O(1)$ by setting $c = 0.5$ and assuming $\delta \leq \frac{1}{2}$. Using Markov's inequality, this means that there exists a string R such that for a constant fraction of sets in $\mathcal{F}$ A_R can recover the set. We will refer to this recoverable subset as $\mathcal{F}$ in the next part. □

1b Conclude that an ℓ_0 sampler with error probability bounded by $\delta > 1/n^{10}$ for every fixed input must use $m = \Omega(\log(1/\delta))$ rows. You may use the result of **1a** even if you did not prove it.

Solution. If all elements of the sketch are polynomially bounded, then an l_0 sampler with m rows gives us a sketch of size $m \log n$. As shown in the previous part, this sketch can be used to recover any set in $\mathcal{F}$. This means that $m \log n \geq \log |\mathcal{F}|$ i.e. $m = \Omega(\log \frac{1}{\delta})$. □