Exercise 8.6

zaterdag 7 oktober 2023 15:07

note: lower bound (proofs) may appear 'at some point'

Suppose a deterministic streaming algorithm, say Alg, for Similar-Items uses at most s bits of storage. Then Alg can be in at most 2^s different states at any point in time. Now, consider the set of streams X defined as follows. Let $\mathbb{S}$ be the set of all subsets of $m-1$ distinct elements from the universe $[n]$ ~~which do not contain two similar items (i.e. items a_i and a_j such that $|a_i - a_j| = 1$)~~, but where all elements are a multiple of 3 and smaller than $n-1$. For each subset $S \in \mathbb{S}$, put a stream σ into X whose elements are exactly the items in S. Note that

$$|X| = |\mathbb{S}| = \binom{n/3}{m-1} \geq \left(\frac{n/3}{m-1}\right)^{m-1} = 2^{(m-1)\log_2 \frac{n/3}{m-1}}$$

we have $\binom{n}{k} \geq \left(\frac{n}{k}\right)^k$

since $a^b = 2^{\log_2(a^b)} = 2^{b \log_2(a)}$

Hence, when $s < (m-1)\log_2 \frac{n/3}{m-1}$, there must be two different streams $\sigma_1 \in X$ and $\sigma_1' \in X$ such that Alg is in exactly the same state after processing σ_1 as it would be after processing σ_1'.

Given that all items in X are distinct, we know that σ_1 and σ_1' do not have similar items. Furthermore, given that each stream in X (including σ_1 and σ_1') only contains elements which are multiples of 3, we can conclude that, for each element a_k which is a multiple of 3, it must be the case that $a_k - 2$, $a_k - 1$, $a_k + 1$ and $a_k + 2$ are not in the stream.
Now, since σ_1 and σ_1' are different streams, each having distinct items, there must be an item $\ell \in [n]$ such that $\ell \in \sigma_1$ and $\ell \notin \sigma_1'$. Now, consider the streams $\kappa_1 = \sigma_1 \circ \langle(\ell+1)\rangle$ and $\kappa_1' = \sigma_1' \circ \langle(\ell+1)\rangle$; these are the streams with the element $\ell + 1$ appended to them. Clearly, streams κ_1 and κ_1' have length m. Furthermore, since ℓ is in stream σ_1, it must be that ℓ is a multiple of 3. This implies that $\ell + 2$ is not in the streams σ_1 and σ_1', nor in any of the streams κ_1 and κ_1'. Thus (given that σ_1 and σ_1' do not contain similar items), κ_1 and κ_1' can only contain a similar item if $\ell + 1$ is similar to some item in the given stream. Since $\ell + 2$ is not in κ_1 or κ_1', it follows that the only item to which $\ell + 1$ can be similar is the item ℓ. But ℓ is only in κ_1 and not in κ_1'. This implies that the algorithm should give the answer Yes for the stream κ_1 and No for the stream κ_1'. But this is impossible, given that the state before processing the item $\ell + 1$ is equal for both streams, and processing the same item from the same starting state cannot lead to different final states. This means that, if Alg uses less than $(m-1)\log_2 \frac{n/3}{m-1} = \Omega\left(m \log_2 \frac{n}{m}\right)$ bits, then it cannot be correct on all inputs. Thus, any deterministic streaming algorithm that solves Similar-Items exactly must use at least $\Omega\left(m \log_2 \frac{n}{m}\right)$ bits.

~~Note: we need the constraint that $m < \frac{n}{3}$ to ensure that there are $\binom{n/3}{m-1}$ streams which do not have a similar item; upon picking more than $n/3$ items, one would have that at least one item needs to be picked twice. *(Usually, one requires that $m < \frac{n}{2}$ to avoid trivially having a similar item, but we note that two distinct elements which are a multiple of 3 cannot be similar; furthermore, an element is not similar to itself in the definition given in the problem. If the problem were to state that elements are similar if $|a_i - a_j| \leq 1$, then the bound $m < \frac{n}{6}$ would be necessary.)*~~

needed to avoid issue with $\log_2 \ldots < 0$, which flips a sign