

Deterministic data streaming

Consider a stream of n items, where items can appear more than once in the stream. The problem is to find the most frequently appearing item in the stream (where ties broken arbitrarily if more than one item satisfies the latter). Suppose that only k items can be stored, one item per memory cell, where the available storage is $k + O(1)$ memory cells. Show that the problem cannot be solved deterministically under the following rules: the algorithm can access only $O(\log^c n)$ bits for each of the k items that it can store, and can read the next item of the stream; you, the adversary, have access to all the stream, and the content of the k items stored by the algorithm, and can decide what is the next item that the algorithm reads (please note that you cannot change the past, namely, the items already read by the algorithm). Hint: it is an adversarial argument based on the k items chosen by the hypothetical deterministic streaming algorithm, and the fact that there can be a tie on $> k$ items till the last minute.

SOLUTION

Any deterministic algorithm can be described by a deterministic finite state machine, in which every state given an input move deterministically to another state. Suppose to have a streaming of n element, $k + 1$ memory cell to store the frequencies, and θ are the bit necessary to store the algorithm state. Therefore we can have 2^θ possible states.

Let's suppose to input $2^\theta + 1$ different stream, where each stream has got a different most frequent element. Therefore, we can notice that the algorithm is going to be in the same state with two different input streaming (i.e. then with two different most frequent element). Therefore, the algorithm is going to answer in two different way, that it's an absurd for a deterministic algorithm. Instead, in the case answer in the same way, one result is going to be wrong.

The stream in input is made by $2^\theta + 1$ items, all different a part of one that is going to be repeated twice, becoming then the most frequent. Obviously, the item repeated twice is going to be different in each input.