Exercise 9.5

zaterdag 14 oktober 2023 16:19

First pass

Input: a stream of length m in the vanilla model.

Initialize:
for $j = 1$ to $k\sqrt{m}$ do
$pos_j \leftarrow random(1, m)$
end for

Note: k will be defined in the proof.

Process(a_i):
for $j = 1$ to $k\sqrt{m}$ do
if $pos_j = i$ then
$sample_j = a_i$
end for

Output:
The samples need to be kept.

Second pass

Input: the same stream of length m in the vanilla model.

Initialize:
for $j = 1$ to $k\sqrt{m}$ do
$rank_j \leftarrow 0$
end for

Process(a_i):
for $j = 1$ to $k\sqrt{m}$ do
if $sample_j < a_i$ then
$rank_j \leftarrow rank_j + 1$
end for

Output:
If any of the items have rank $\left\lfloor \frac{m+1}{2} \right\rfloor$ or $\left\lceil \frac{m+1}{2} \right\rceil$, return them as the median; otherwise, the ranks of each sampled position need to be kept.

Third pass

Input: the same stream of length m in the vanilla model.

Initialize:
Select the sampled items which are closest to the median, from each side.
That is, let $rank\_low \leftarrow \max_{(1 \le j \le k\sqrt{m}) \wedge \left(\text{rank}_j \le \left\lfloor \frac{m+1}{2} \right\rfloor\right)} rank_j$. If no such item exists, let $rank\_low \leftarrow 0$.
And let $rank\_high \leftarrow \min_{(1 \le j \le k\sqrt{m}) \wedge \left(\text{rank}_j \ge \left\lceil \frac{m+1}{2} \right\rceil\right)} rank_j$. If no such item exists, let $rank\_high \leftarrow m$.
Let low be the (unique) item with rank $rank\_low$. (If $rank\_low = 0$ because no item existed, set $low \leftarrow 0$.)
Let $high$ be the (unique) item with rank $rank\_high$. (If $rank\_high = m$ because no item existed, set $high \leftarrow n$.)
Let S be a list

Process(a_i):
If $low < a_i < high$
Append a_i to S

Output:
Sort the list S using a sorting algorithm (preferably one which uses a low amount of storage).
Go over the items in S, and pick one of the items with rank $\left\lfloor \frac{m+1}{2} \right\rfloor$ or $\left\lceil \frac{m+1}{2} \right\rceil$ (note: there can be either one or two such items). For this, it is useful to note that the first item in S will have rank $rank\_low + 1$.
Return the picked item.

Proof of correctness

Arguably somewhat informal
The answer returned here is correct because of the following: in the third pass, we simply go over all elements between the two elements which are closest (from each side) to the median. Then, we return as median an item which has appropriate rank (i.e. $\left\lfloor \frac{m+1}{2} \right\rfloor$ or $\left\lceil \frac{m+1}{2} \right\rceil$), which can be accurately determine because the second pass has given us the rank of the items closest to the median.

Proof of storage use

The first pass, as described on the right, needs to sample approximately $4.32\sqrt{m}\log n = O(\sqrt{m}\log n)$ samples; the random numbers which need to be stored only add a constant factor to this term.
The second pass stores one number (a rank), which is at most $\log n$ bits. This is clearly doable in the allotted amount of storage.
The third pass stores, under the situation for which the probability was bounded in pass one, at most $2\sqrt{m}$ items in the list S. This also fits in the allowed amount of storage.

Hence, the total algorithm uses at most $O(\sqrt{m}\log n)$ bits of storage with probability at least 0.95. *(Note that the algorithm always comes to an exact solution; however, in the worst case (with extremely low probability) it might need to store all items in the third pass.)*

Proof of probability of storage

If both the items low and $high$ have a rank at most $\sqrt{m}$ from the median, then the number of items which need to be stored in the third pass is at most $2\sqrt{m} = O(\sqrt{m})$ (and storing such items can be done in $O(\sqrt{m}\log(n))$ bits). It remains to be shown that the probability of the items being picked in such a way can be made at least 0.95.

To this end, let X_j denote an indicator random variable defined as follows:

- $X_j = 1$ if there is the j^{th} sample picked in the first pass has rank at least $\left\lfloor \frac{m+1}{2} \right\rfloor$ and at most $\left\lfloor \frac{m+1}{2} \right\rfloor + k\sqrt{m}$;
- $X_j = 0$ otherwise.

We have that $\Pr[X_j = 1] = \frac{\sqrt{m}}{m} = \frac{1}{\sqrt{m}}$, as the random variable will have value 1 if and only if one of the $\sqrt{m}$ out of m options which fall in the bound are chosen; due to uniform randomness of the choice, each option then has an equal chance of $\frac{\sqrt{m}}{m} = \frac{1}{\sqrt{m}}$ to be chosen.

Now, we have that at least one of the $k\sqrt{m}$ samples is within $\sqrt{m}$ from the median (and at least as large as the median) if and only if $X_j = 1$ for at least one j. Stated differently, the probability of at least one sample complying with our demands is $\Pr[at\ least\ one\ complies] = 1 - \Pr[none\ comply] = 1 - (\Pr[one\ does\ not\ comply])^{\#samples} = 1 - (1 - \Pr[one\ complies])^{\#\ samples} = 1 - \left(1 - \frac{1}{\sqrt{m}}\right)^{k\sqrt{m}}$.
By symmetry, the same reasoning can be applied to obtain the probability that we have one sample at most the median, with rank at most $\sqrt{m}$ below the median.

~~Let us use the simpler version of the probability, as obtained from the Markov inequality.~~

~~Thus, we desire to have $1 - 2\Pr[at\ least\ one\ complies] \ge 0.95$~~

~~$1 - 2k \ge 0.95$~~
~~$1 - 0.95 \ge 2k$~~
~~$0.05 \ge 2k$~~
~~$0.025 \ge k$~~

We desire to have that the probability that at least one item is appropriate (on each side of the median) is at least 0.95. Thus, we need:

$\Pr[at\ least\ one\ item\ at\ most\ the\ median\ has\ rank\ less\ than\ \sqrt{m}\ \text{from the median}]$
$\cdot \Pr[at\ least\ one\ item\ at\ least\ the\ median\ has\ rank\ less\ than\ \sqrt{m}\ \text{from the median}] \ge 0.95$
$\Pr[at\ least\ one\ complies]^2 \ge 0.95$
$\left(1 - \left(1 - \frac{1}{\sqrt{m}}\right)^{k\sqrt{m}}\right)^2 \ge 0.95$
Using the inequality in the exercise, we obtain

$\left(1 - \left(1 - \frac{1}{\sqrt{m}}\right)^{k\sqrt{m}}\right) > 1 - \left(\frac{1}{2}\right)^k$, and hence, we need
$1 - \left(\frac{1}{2}\right)^k \ge 0.95$
$0.05 \ge \left(\frac{1}{2}\right)^k$
$0.05 \ge 2^{-k}$
$\log_2 0.05 \ge -k$
Hence, we take $k \ge -\log_2 0.05 \approx 4.32$
Thus, it is possible with probability at least 0.95 to use at most $4.32\sqrt{m}\log n = O(\sqrt{m}\log n)$ bits of storage to store the items in between low and $high$, and from these items, the median can be exactly determined.

Alternatively, consider the Markov inequality. Define a random variable $X = \sum_j X_j$. Then, $E[X] = E\left[\sum_j X_j\right] = \sum_j E[X_j] = \sum_j \frac{k}{\sqrt{m}} = \sqrt{m} \cdot \frac{k}{\sqrt{m}} = k$.
We obtain that the probability $\Pr[X \ge 1] = \Pr\left[X \ge \frac{1}{k} \cdot E[X]\right] \le \frac{1}{\frac{1}{k}} = k$ by the Markov inequality.

~~Alternatively, use the Chernoff bound. Then, we have that $\Pr[X \ge 1] \ge 1 - \Pr[X > (1+?)E[X]]$~~

Let us use the inequality suggested in the exercise. Then, we obtain $\Pr[at\ least\ one\ complies] = 1 - \left(1 - \frac{1}{\sqrt{m}}\right)^{k\sqrt{m}} > 1 - \left(\frac{1}{2}\right)^k$.