Exercise 8.2

zaterdag 7 oktober 2023 19:33

The following deterministic streaming algorithm can solve the two missing items problem using a sub-linear number of bits:

Input:
A stream $\langle a_1, \dots, a_{n-2} \rangle$ in the vanilla model, where n is the size of the universe, and all the a_i are distinct.

Initialize:
$sum \leftarrow 0$
$squaredsum \leftarrow 0$

Process(a_i):
1. $sum \leftarrow sum + a_i$
2. $squaredsum \leftarrow squaredsum + a_i^2$

Output:
The missing items are j_1 and j_2 with:

$$D \leftarrow \left(2\left(sum - \left(\frac{1}{2}n(n+1)\right)\right)\right)^2 - 4 \cdot 2 \cdot -\left(\frac{n(n+1)(2n+1)}{6} - squaredsum - \left(\frac{1}{2}n(n+1)\right)^2 - sum^2 + 2\left(\frac{1}{2}n(n+1)\right) \cdot sum\right)$$

$$j_2 \leftarrow \frac{-2\left(sum - \left(\frac{1}{2}n(n+1)\right)\right) + \sqrt{D}}{2 \cdot 2}$$

$$j_1 \leftarrow \frac{1}{2}n(n+1) - sum - j_2$$

Explanation of the equations for j_1 and j_2

Let j_1, j_2 be the two missing items.
Note that we have $\sum_{i=1}^{n} i = \frac{1}{2}n(n+1)$, whereas $sum = \sum_{i=1}^{n}(i) - j_i - j_2$.
Furthermore, $\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$, whereas $squaredsum = \sum_{i=1}^{n}(i^2) - j_1^2 - j_2^2$.

Substituting the value for j_1 obtained from the equation for sum ($j_1 = \sum_{i=1}^{n}(i) - sum - j_2$) into the equation for squaredsum gives us the following:

$$\text{squaredsum} = \sum_{i=1}^{n}(i^2) - j_1^2 - j_2^2$$

$$j_2^2 = \sum_{i=1}^{n}(i^2) - squaredsum - \left(\sum_{i=1}^{n}(i) - sum - j_2\right)^2$$

$$j_2^2 = \frac{n(n+1)(2n+1)}{6} - squaredsum - \left(\frac{1}{2}n(n+1)\right)^2 - sum^2 - j_2^2 + 2\left(\frac{1}{2}n(n+1)\right) \cdot sum + 2\left(\frac{1}{2}n(n+1)\right) \cdot j_2 - 2 \cdot sum \cdot j_2$$

$$2j_2^2 + 2\left(sum - \left(\frac{1}{2}n(n+1)\right)\right) \cdot j_2 = \frac{n(n+1)(2n+1)}{6} - squaredsum - \left(\frac{1}{2}n(n+1)\right)^2 - sum^2 + 2\left(\frac{1}{2}n(n+1)\right) \cdot sum$$

This equation can be solved using the quadratic formula.

$$D = \left(2\left(sum - \left(\frac{1}{2}n(n+1)\right)\right)\right)^2 - 4 \cdot 2 \cdot -\left(\frac{n(n+1)(2n+1)}{6} - squaredsum - \left(\frac{1}{2}n(n+1)\right)^2 - sum^2 + 2\left(\frac{1}{2}n(n+1)\right) \cdot sum\right)$$

$$j_2 = \frac{-2\left(sum - \left(\frac{1}{2}n(n+1)\right)\right) + \sqrt{D}}{2 \cdot 2}$$

Note that we are only interested in the positive square root, given that $j_2 \geq 0$.
Using the value of j_2 obtained from the equations above, we can then obtain j_1:

$$j_1 = \frac{1}{2}n(n+1) - sum - j_2$$

Proof of correctness

Admittedly somewhat informal

We note that only two distinct ***natural*** numbers j_1 and j_2 can satisfy the equations worked out to the left of here. Given that the sum of the items in the stream and the missing items, as well as the sum of squares of these items, are given by constants, we have that the values of j_1 and j_2 must be correct.

Storage requirements analysis

This algorithm, as part of its process phase, only needs to store the numbers sum and $squaredsum$. Since (the bigger of) these numbers cannot grow larger than $n \cdot n^2$, we have that they can each be stored in at most $\log_2 n^3 = 3\log_2 n$ bits, which is clearly sublinear. Furthermore, since the powers of n used in the computation of the values of j_1 and j_2 are constant, the number of bits used to compute the output phase is also sublinear in n. Hence, the total number of bits of storage needed to compute the two missing items is sublinear.