Assignment - 6   Stream Algorithm

Question 2 :- Estimate the surprise number of a data stream using the method of AMS.

$\boxed{2m - 1}$ m is the number of occurrence of the element of the stream at that timestamp

Stream has 1, 2, ... 10 cycles repeatedly.

timestamp current $\boxed{n = 75}$

Upto 75

$$\begin{tikzcd} 1-5 \arrow[d, "\text{Series}"] & 6-10 \arrow[d, "\text{Series}"] \\ 8 & 7 \end{tikzcd}$$

Range of 1-5 is 5

Range of 6-10 is 5

Surprise number = $(5 \times 8^2) + (5 \times 7^2) = 565$

Let's take $\{24, 44, 65\}$

Our estimate will be ~~the~~ median of three resulting values.

$n(2m-1)$

$24 = 75[2(6) - 1] = 825$

$44 = 75[2(4) - 1] = 525$

$65 = 75[2(2) - 1] = 225$

Since its median has 525 nearer to 565

we choose 44.

Question 2: DGIM Algorithm

End time	100	98	95	92	87	80	65
Size	1	1	2	2	4	8	8

Sliding window length = 40
Current time stamp = 100

Since there are 3 buckets of 1's we combine earlier buckets of 1's by 2 [according to DGIM algorithm]

102	101	100	95	92	87	80	65
1	1	2	2	2	4	8	8

2 2 → since 3 2's are in a rows combine to bucket of 4.

103	102	101	100	95	87	80	65
1	1	1	2	4	4	8	8

104	103	102	100	95	87	80	85
1	1	2	2	4	4	8	8

As 105th stream arrives, again 3 1's appear

105	104	103	102	100	95	87	80	65
1	1	1	2	2	4	4	8	8

105	104	102	100	95	87	80	65
1	2	2	2	4	4	8	8

Combine 2 2's to get 4

105	104	102	93	87	80	65
1	2	4	4	4	8	8

combine 2 4's to get 8

105	104	102	95	80	65
1	2	4	8	8	8

combine 2 8's to get 16

~~So size~~

105	104	102	93	80
1	2	4	8	16

Size of bucket of 105 is 1.

Question 3: Count no. of distinct elements in a stream.

hash function

$\hookrightarrow (3x+7) \% 11$ → modullo

x	$h(x)$	binary
1	10	1010
2	2	0010
3	5	0101
4	8	1000
5	0	0000
6	3	0011
7	6	0110
8	9	1001
9	1	0001
10	4	0100

We consider our option

$\{1,3,6,8\} \Rightarrow \{1+0+0+0\}=1$

$\{2,6,8,10\}=\{1+0+0+2\}=3$

$\{2,6,8,9\}=\{1+0+0+0\}=1$

$\{2,5,7,10\}=\{1+4+1+2\}=8$

$\{2,5,7,10\}$ is violated since has more than 2 0's on right side. So we choose $\{2,6,8,10\}$ since it is closer to our estimate.

Question 4

Users = 10^8

Sample = 10^{10} bytes

User ID's will be hashed to a bucket number from 0 to 999,999.

Threshold (t) ⇒ 100 byte records for all the users whose ID's hash to t or less will be retained and other users records will not be retained.

Considering $n = 10^9$ then threshold t is

$$t = \frac{10^{14}}{10^{19}} - 1 = 10^5 - 1 = 99999 \; //$$

Question 5

Bit array length = 100 → t

set has = 23 members → d

Initially all are 0's and a bit to 1 is set whenever a member of set hashes to it.

fraction of 0's = $e^{-hd/t}$

$= e^{-23/100}$

$\boxed{h = 1}$ → No. of hash functions

fraction of 1's = $1 - e^{-hd/t}$

$= 1 - e^{-23/100}$