Stream Algorithms

Question 1: We wish to estimate the surprise number (2nd moment) of a data stream, using the method of AMS. It happens that our stream consists of ten different values, which we'll call 1, 2,..., 10, that cycle repeatedly. That is, at timestamps 1 through 10, the element of the stream equals the timestamp, at timestamps 11 through 20, the element is the timestamp minus 10, and so on. It is now timestamp 75, and a 5 has just been read from the stream. As a start, you should calculate the surprise number for this time.

For our estimate of the surprise number, we shall choose three timestamps at random, and estimate the surprise number from each, using the AMS approach (length of the stream times 2*m*-1, where *m* is the number of occurrences of the element of the stream at that timestamp, considering all times from that timestamp on, to the current time). Then, our estimate will be the median of the three resulting values.

You should discover the simple rules that determine the estimate derived from any given timestamp and from any set of three timestamps. Then, take any 4 examples of the set of three "random" timestamps, find out the closest estimate among the 4 examples.

Sol:First, the surprise number is 5*64 + 5*49 = 565. The reason is that the elements 1 to 5 appear 8 times, so they contribute 5*82 , and the elements 6 to 10 appear 7 times,
contributing 5*72. The AMS estimate is a nondecreasing function of the timestamp. Thus, of any three timestamps, the middle one will give the median estimate, and we do not have to calculate all three. At each of the timestamps between 36 and 45, inclusive, the element appearing then appears exactly 4 times, from that time forward. Thus, each of these timestamps generates an estimate of 75*(2*4 - 1) = 525, which is as close to 565 as we can get. Each of the correct answers has a middle timestamp in this range. Similarly, for the timestamps between 26 and 35, the estimate is 75*(2*5 - 1) = 675 and for the timestamps between 46 and 55 the estimate is 75*(2*3 - 1) = 375. Neither of these groups offer as close an estimate, and the timestamps earlier or later offer even worse estimates.

Question 2: Suppose we are using the DGIM algorithm of Section 4.6.2 to estimate the number of 1's in suffixes of a sliding window of length 40. The current timestamp is 100, and we have the following buckets stored:

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

Note: we are showing timestamps as absolute values, rather than modulo the window size, as DGIM would do.

Suppose that at times 101 through 105, 1's appear in the stream. Compute the set of buckets that would exist in the system at time 105. Buckets are represented by pairs (end-time, size).

Sol:Check if there are more than 2 buckets with same size after adding +1 on timeline
2. If there are more than 2 buckets with same size, group them to a new size=sizex2
and with timestamp equals to the latest one.
3. Repeat until there are no more than 2 buckets with same size.
By doing this on the given data, we get finally the set of buckets that exists at time 105 are
{(105,1), (104,2), (102,4), (95, 8), (80, 16)}

Question 3: We wish to use the Flagolet-Martin algorithm of Section 4.4 to count the number of distinct elements in a stream. Suppose that there are ten possible elements, 1, 2,..., 10, that could appear in the stream, but only four of them have actually appeared. To make our estimate of the count of distinct elements, we hash each element to a 4-bit binary number. The element *x* is hashed to 3*x* + 7 (modulo 11). For example, element 8 hashes to 3*8+7 = 31, which is 9 modulo 11 (i.e., the remainder of 31/11 is 9). Thus, the 4-bit string for element 8 is 1001.

A set of four of the elements 1 through 10 could give an estimate that is exact (if the estimate is 4), or too high, or too low. You should figure out under what circumstances a set of four elements falls into each of those categories. Then, take any 4 examples of the set of four elements, find out the exactly correct estimate among 4 examples.

Sol:Given,h(x)=(3x+7) modullo 11

Distinct elements(de)==4 ,if de<4-->low

If de>4---->high

Ex1:3,4,8,10

	h(x)	binary	trailing zeros
3	5	0101	0
4	8	1000	3
8	9	1001	0
10	4	0100	2

Max trailing zeros=3==> 2*3==>8(high)

Ex2:2,3,6,9

	h(x)	binary	trailing zeros
2	2	0010	1
3	5	0101	0
6	3	0011	0
9	1	0001	0

Max trailing zeros=1==>2*1==>2(low)

Ex3:1,6,7,10

	h(x)	binary	trailing zeros
1	10	1010	1
6	3	0011	0
7	6	0110	1
10	4	0100	2

Max trailing zeros=2==>2*2==>4(exact)

Ex4:1,3,9,10

	h(x)	binary	trailing zeros
1	10	1010	1
3	5	0101	0
9	1	0001	0
10	4	0100	2

Max trailing zeros=2==>2*2==>4(exact)

So,exact estimations are:{1,6,7,10} and {1,3,9,10}

Question 4: A certain Web mail service (like gmail, e.g.) has 10^8 users, and wishes to create a sample of data about these users, occupying 10^{10} bytes. Activity at the service can be viewed as a stream of elements, each of which is an email. The element contains the ID of the sender, which must be one of the 10^8 users of the service, and other information, e.g., the recipient(s), and contents of the message. The plan is to pick a subset of the users and collect in the 10^{10} bytes records of length 100 bytes about every email sent by the users in the selected set (and nothing about other users).

The method of Section 4.2.4 will be used. User ID's will be hashed to a bucket number, from 0 to 999,999. At all times, there will be a threshold t such that the 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. You may assume that each user generates emails at exactly the same rate as other users. As a function of n, the number of emails in the stream so far, what should the threshold t be in order that the selected records will not exceed the 10^{10} bytes available to store records?

Sol:Suppose that the fraction of users in the sample is p.
The number of users whose records are stored is 10^8 .
Since each user generates 10^8 emails in the stream when n emails have been seen, then the number of records stored is 10^8 *p*10^(-8) = pn.
Since each record is 100 bytes, we can store 10^10/100 = 10^8 records.
i.e., pn = 10^8 , or p = 10^8 /n.
If the threshold is t, the fraction p of users that will be in the selected set is
(t+1)/1000000 = 10^8 /n.
Therefore, t = (10^14 / n) - 1

Question 5: Suppose we hash the elements of a set S having 23 members, to a bit array of length 100. The array is initially all-0's, and we set a bit to 1 whenever a member of S hashes to it. The hash function is random and uniform in its distribution. What is the expected fraction of 0's in the array after hashing? What is the expected fraction of 1's? You may assume that 100 is large enough that asymptotic limits are reached.

Sol:

Given,Members: 23

Bit array of length: 100; T = 100; hash function = 1 Expected fraction of

0's = $e^{-hd}/t = e^{-23}/100$

Expected fraction of 1's = $1 - e^{-hd}/t = 1 - e^{-23}/100$