Give problem in Flajolet-Martin (FM) Algorithm to count distinct elements in a stream.

To estimate the number of different elements appearing in a stream, we can hash elements to integers interpreted as binary numbers. 2 raised to the power that is the longest sequence of 0's seen in the hash value of any stream element is an estimate of the number of different elements.  
Eg. Stream: 4, 2, 5 ,9, 1, 6, 3, 7  
Hash function,  h(x) = (ax + b) mod 32  
a) h(x) = 3x + 1 mod 32  
  
a) h(x) = 3x + 7 mod 32  
h(4) = 3(4) + 7 mod 32 = 19 mod 32 = 19 = (10011)  
h(2) = 3(2) + 7 mod 32 = 13 mod 32 = 13 = (01101)  
h(5) = 3(5) + 7 mod 32 = 22 mod 32 = 22 = (10110)  
h(9) = 3(9) + 7 mod 32 = 34 mod 32 = 2 = (00010)  
h(1) = 3(1) + 7 mod 32 = 10 mod 32 = 10 = (01010)  
h(6) = 3(6) + 7 mod 32 = 25 mod 32 = 25  = (11001)  
h(3) = 3(3) + 7 mod 32 = 16 mod 32 = 16  = (10000)  
h(7) = 3(7) + 7 mod 32 = 28 mod 32 = 28  = (11100)  
Trailing zero's {0, 0, 1, 1, 1,  0, 4, 2}  
R = max [Trailing Zero] = 4  
Output = 2R  = 24  = 16

b) h(x) = x + 6 mod 32

h(4) = (4) + 6 mod 32 = 10 mod 32 = 10 = (01010)  
h(2) = (2) + 6 mod 32 = 8 mod 32 = 8  = (01000)  
h(5) = (5) + 6 mod 32 = 11 mod 32 = 11  = (01011)  
h(9) = (9) + 6 mod 32 = 15 mod 32 = 15 = (01111)  
h(1) = (1) + 6 mod 32 = 7 mod 32 = 7 = (00111)  
h(6) = (6) + 6 mod 32 = 12 mod 32 = 12 = (01110)  
h(3) = (3) + 6 mod 32 = 9 mod 32 = 9 = (01001)  
h(7) = (7) + 6 mod 32 = 13 mod 32 = 13 = (01101)  
Trailing zero's {1, 3, 0, 0, 0, 1, 0, 0}  
R = max [Trailing Zero] = 3  
Output = 2R  = 23  = 8

Flajolet-Martin algorithm approximates the number of unique objects in a stream or a database in one pass. If the stream contains n elements with m of them unique, this algorithm runs in O(n) time and needs O(log(m)) memory.

Algorithm:

Create a bit vector (bit array) of sufficient length L, such that 2L>n, the number of elements in the stream. Usually a 64-bit vector is sufficient since 264 is quite large for most purposes.

The i-th bit in this vector/array represents whether we have seen a hash function value whose binary representation ends in 0i. So initialize each bit to 0.

The i-th bit in this vector/array represents whether we have seen a hash function value whose binary representation ends in 0i. So initialize each bit to 0.

The i-th bit in this vector/array represents whether we have seen a hash function value whose binary representation ends in 0i. So initialize each bit to 0.

Once input is exhausted, get the index of the first 0 in the bit array (call this R). By the way, this is just the number of consecutive 1s (i.e. we have seen 0,00,...,0R−1 as the output of the hash function) plus one.

Calculate the number of unique words as 2R/ϕ, where ϕ is 0.77351. A proof for this can be found in the original paper listed in the reference section.

The standard deviation of R is a constant: σ(R)=1.12. (In other words, R can be off by about 1 for 1 - 0.68 = 32% of the observations, off by 2 for about 1 - 0.95 = 5% of the observations, off by 3 for 1 - 0.997 = 0.3% of the observations using the Empirical rule of statistics). This implies that our count can be off by a factor of 2 for 32% of the observations, off by a factory of 4 for 5% of the observations, off by a factor of 8 for 0.3% of the observations and so on.

Example:

S=1,3,2,1,2,3,4,3,1,2,3,1

h(x)=(6x+1) mod 5

Assume |b| = 5

x h(x) Rem Binary r(a)

1 7 2 00010 1

3 19 4 00100 2

2 13 3 00011 0

1 7 2 00010 1

2 13 3 00011 0

3 19 4 00100 2

4 25 0 00000 5

3 19 4 00100 2

1 7 2 00010 1

2 13 3 00011 0

3 19 4 00100 2

1 7 2 00010 1

R = max( r(a) ) = 5

So no. of distinct elements = N=2R=25=32

Flajolet-martin algorithm – Example

•Input streams x=1,3,2,1,2,3,4,3,1,2,3,1

•Hash function

•H(x)=6x + 1 mod 5

H(1)=6(1)+1 mod 5=7 mod 5=2

Example: Input stream =1,3,2,1,2,3,4,3,1,2,3,1

h(x)=6x + 1 mod 5

h(1)=2

h(3)=4

h(2)=3

h(1)=2

h(2)=3

h(3)=4

h(4)=0

h(3)=4

h(1)=2

h(2)=3

h(3)=3

h(1)=2

Flajolet-martin algorithm - Example

Binary representation

h(1)=2=010

h(3)=4=100

h(2)=3=011

h(1)=2=010

h(2)=3=011

h(3)=4 =100

h(4)=0=000

h(3)=4=100

h(1)=2=010

h(2)=3=011

h(3)=3=011

h(1)=2=010

Flajolet-martin algorithm - Example

Trailing zeros

h(1)=2=010=1

h(3)=4=100=2

h(2)=3=011=0

h(1)=2=010=1

h(2)=3=011=0

h(3)=4 =100 =2

h(4)=0=000=0

h(3)=4=100=2

h(1)=2=010=1

h(2)=3=011=0

h(3)= 4 =100 =2

h(1)=2=010=1

Calculating distinct elements

•r=2

•R=2 ^r

•R=2^2=4

•Hence we have 4 distinct elements 1,3,2,4

