Frequency Moments of Streaming Data

Scribed by Vineet Goyal

January 28, 2003

1 Introduction

We consider the problem of calculating frequency moments of a large data set by making only 1 sequential pass of the entire data using space which is only polylog in the size of the input.

Problem: We are given a stream of integers $a_1, a_2, \ldots, a_m$ such that $a_i \in 1, 2, \ldots, n \; \forall 1 \leq i \leq m$. The problem is to find the number of distinct integers $j \in 1, 2, \ldots, n$ that occur in the stream approximately within a constant factor with high probability.

2 Algorithm

Consider an intuitive scheme of randomly picking a small set of integers from the stream and making a decision based on that. This doesn't work as it is easily fooled in the case when one particular integer occurs very frequently. We need to use an appropriately chosen hash function as we have only space which is polylog in the size of the input. We give the following algorithm.

Input: A stream of integers $a_1, a_2, \ldots, a_m$ where $a_i \in 1, 2, i \ldots, n$.

Step 1: Fix a prime p between n and $2n$.

Step 2: Choose 2 independent uniform random numbers a, b $1 \leq a \leq p - 1,\ 0 \leq b \leq p - 1$.

Step 3: Find $\forall i,\ Z_i = (a_i a + b) \text{mod } p$. Let $Z = min_i Z_i$.

Result: $\frac{p}{Z}$ is the approximate number of distinct integers in the stream.

Let us first understand the intuition behind the algorithm. We try to map each integer a_i from the stream to an integer $Z - i \in [0, p-1]$ such that each Z_i is uniformly random in $[0, p-1]$. Note that with the above hash function 2 distinct integers can't be mapped to the same integer. So we obtain distinct integers Z_i for each distinct integer a_i which are all uniformly random between 0 and $p - 1$.

Suppose $j_1, j_2, \ldots, j_N$ be N distinct $a_i's$ and let $X_t = (aj_t + b) \text{mod } p$.

Claim 1: $X_t's$ are uniformly random and pairwise independent.
Proof:

i) For each fixed $i_0 \in 0, 1, \ldots, p-1$

$$\begin{aligned} Pr(X_t = i_0) &= Pr(aj_t = (i_0 - b) \text{mod } p) \\ &= Pr(a = j_t^{-1}(i_0 - b) \text{mod } p) \\ &= \frac{(p-1)}{p(p-1)} \\ &= \frac{1}{p} \end{aligned}$$

Thus, $X_t's$ are uniformly random between 0 and $p-1$.

ii) $X_t's$ are pairwise independent. Let i_0, i_1 be fixed integers such that $i_0 \neq i_1$.

$$Pr(X_t = i_0, X_{t'} = i_1) = Pr(aj_t + b = i_0(\text{mod } p), aj_{t'} + b = i_1(\text{mod } p)$$

Since, $t \neq t'$ and $i_0 \neq i_1$ we have a unique solution for a, b
Thus, $Pr(X_t = i_0, X_{t'} = i_1) = \frac{1}{p(p-1)}$.
Now we want to show that with high probability $Z_{min} = \frac{p}{N}$ where N is the number of distinct integers in the stream. Let us try to find $Pr(Z \leq s-1)$.

Define

$$Y_t = \begin{cases} 1 & \text{if } X_t \leq s-1 \\ 0 & \text{if } X_t \geq s \end{cases}$$

$$Pr(Z \leq s-1) = Pr(\sum_{t=1}^{N} Y_t \geq 1)$$

$$\begin{aligned} E[Y] &= \sum_{t=1}^{N} E[Y_t] \\ &= \frac{sN}{p} \end{aligned}$$

Since, $E[Y_t] = \frac{s}{p}$.

$$\begin{aligned} E[(\sum_{t=1}^{N} Y_t)^2] &= E[\sum_{t=1}^{N} Y_t^2] + 2E[\sum_{t<t'} Y_t Y_{t'}] \\ &= \frac{sN}{p} + \frac{N(N-1)s^2}{p^2} \end{aligned}$$

Thus,

$$\begin{aligned} Var[Y] &= E[(\sum_{t=1}^{N} Y_t)^2] - (E[(\sum_{t=1}^{N} Y_t)])^2 \\ &= \frac{sN}{p} + \frac{N(N-1)s^2}{p^2} - \frac{s^2N^2}{p^2} \\ &= \frac{sN}{p}(1 - \frac{s}{p}) \end{aligned}$$

Thus, $Var[Y] \leq \frac{sN}{p}$. Now we can bound the value of Z and its probability of occurence in that bound.
Claim 2: $\frac{p}{cN} \leq\ Z <\ \frac{cp}{N}$, with probability atleast $1 - \frac{2}{c}$ for any constant c.
Proof :
We try to find the failiure probabilities i.e. $Pr(Z \geq \frac{cp}{N})$ and $Pr(Z < \frac{p}{cN})$ and then use a union bound to lower bound the success probability.

1. $Pr(Z \geq \frac{cp}{N})$

   Let $s = \frac{cp}{N}$
   $$E[Y] = c \text{ and } Var[Y] = c$$
   $$\begin{aligned} Pr(Y < 1) &= Pr(Y \leq 0) \\ &\leq Pr(|Y - c| \geq c) \\ &\leq \frac{1}{c} \text{ (Using Chebychev's inequality)} \end{aligned}$$
   From this we have,
   $$\begin{aligned} Pr(Z \leq s - 1) &= Pr(Y \geq 1) \\ &= 1 - Pr(Y < 1) \\ &\geq 1 - \frac{1}{c} \end{aligned}$$
   Thus, $Pr(Z \geq \frac{cp}{N}) = 1 - Pr(Z \leq s - 1) \leq \frac{1}{c}$

2. $Pr(Z < \frac{p}{cN})$

   Let $s = \frac{p}{cN}$
   $$E[Y] = \frac{1}{c} \text{ and } Var[Y] = \frac{1}{c}$$
   $$\begin{aligned} Pr(Z \leq s - 1) &= Pr(Y \geq 1) \\ &\leq \frac{1}{c} \text{ (Using Markov's Inequality)} \end{aligned}$$

Thus, with probability atleast $1 - \frac{2}{c}$, we have $\frac{p}{cN} \leq\ Z <\ \frac{cp}{N}$

3 Comparing Two Integers over a Network

Consider the problem of comparing two n-bit integers A and B on two distinct computers with communication of minimum numbers of bits between the two computers. Any deterministic algorithm will clearly require $\Omega(n)$ bits of communication for the comparison.

We give the following randomized algorithm.

Step 1: Pick a random prime p between 1 and k.
Step 2: Computer 1 sends p and A (mod p).
Step 3: Computer 2 checks $A(\text{mod p}) = B(\text{mod p})$

If $A(\text{mod p}) = B(\text{mod p})$, then $p|(A-B)$. If $A \neq B$, then the algorithm will give a wrong result if $p|(A-B)$. So let us calculate the failiure probability.

Let $D =$ Total # primes b/w 1 and k.

$$\begin{aligned}\text{Failiure Prob} &= Pr(p|(A-B)) \\ &= \frac{\text{\# primes b/w 1 and k dividing } (A-B)}{\text{Total \# primes b/w 1 and k}} \\ &\leq \frac{\log_2(A-B)}{D} \\ &= \frac{n}{D}\end{aligned}$$

We want $D \geq cn$ to make the failiure probability $< \frac{1}{c}$.
For this choose $k \approx cn \log n$ and then for communicating both p and $A(\text{mod p})$ we require only $O(\log n)$ bits.