Estimating Frequency Moments F_0 and F_2

Anil Maheshwari

anil@scs.carleton.ca
School of Computer Science
Carleton University
Canada

Outline

Frequency Moments

Frequency Moments

Definition

Let $A = (a_1, a_2, \ldots, a_n)$ be a stream, where elements are from universe $U = \{1, \ldots, u\}$. Let $m_i = \#$ of elements in A that are equal to i. The k-th frequency moment $F_k = \sum_{i=1}^{u} m_i^k$, where $0^0 = 0$.

Example: $F_k = \sum_{i=1}^{u} m_i^k$

$A = (3, 2, 4, 7, 2, 2, 3, 2, 2, 1, 4, 2, 2, 2, 1, 1, 2, 3, 2)$ and $m_1 = m_3 = 3$, $m_2 = 10$, $m_4 = 2$, $m_7 = 1$, $m_5 = m_6 = 0$

$F_0 = \sum_{i=1}^{7} m_i^0 = 3^0 + 10^0 + 3^0 + 2^0 + 0^0 + 0^0 + 1^0 = 5$
(# of Distinct Elements in A)

$F_1 = \sum_{i=1}^{7} m_i^1 = 3^1 + 10^1 + 3^1 + 2^1 + 0^1 + 0^1 + 1^1 = 19$
(# of Elements in A)

$F_2 = \sum_{i=1}^{7} m_i^2 = 3^2 + 10^2 + 3^2 + 2^2 + 0^2 + 0^2 + 1^2 = 123$
(Surprise Number)

$\dots$

Streaming Problem

Find frequency moments in a stream

Input: A stream A consisting of n elements from universe $U = \{1, \dots, u\}$.
Output: Estimate Frequency Moments F_k's for different values of k.

Our Task: Estimate F_0 and F_2 using sublinear space

Reference: The space complexity of estimating frequency moments by Noga Alon, Yossi Matias, and Mario Szegedy, Journal of Computer Systems and Science, 1999.

Estimating F_0

Estimating F_0

Computation of F_0

Input: Stream $A = (a_1, a_2, \ldots, a_n)$, where each $a_i \in U = \{1, \ldots, u\}$.

Output: An estimate $\hat{F_0}$ of number of distinct elements F_0 in A such that $Pr\left(\frac{1}{c} \leq \frac{\hat{F_0}}{F_0} \leq c\right) \geq 1 - \frac{2}{c}$ for some constant c using sublinear space.

Algorithm

Algorithm for Estimating F_0

Input: Stream A and a hash function $h : U \to U$

Output: Estimate $\hat{F_0}$

Step 1: Initialize $R := 0$

Step 2: For each elements $a_i \in A$ do:

1. Compute binary representation of $h(a_i)$
2. Let r be the location of the rightmost 1 in the binary representation
3. if $r > R$, $R := r$

Step 3: Return $\hat{F_0} = 2^R$

Space Requirements = $O(\log u)$ bits

Correctness

Observation 1

Let d to be smallest integer such that $2^d \geq u$ (d-bits are sufficient to represent numbers in U)

Observation 1

Pr(rightmost 1 in $h(a_i)$ is at location $\geq r+1) = \frac{1}{2^r}$

Proof: For that to happen the last r bits in $h(a_i)$ must be 0. Since h is a hash function from universal family of hash functions, this happens with probability $(\frac{1}{2})^r$.

□

Observations 2

Observation 2

For $a_i \neq a_j$, $Pr($rightmost 1 in $h(a_i) \geq r+1$ and rightmost 1 in $h(a_j) \geq r+1) = \frac{1}{2^{2r}}$

Proof: $h(a_i)$ and $h(a_j)$ are independent as $a_i \neq a_j$.

$Pr($rightmost 1 in $h(a_i) \geq r+1$ and rightmost 1 in $h(a_j) \geq r+1) = Pr($rightmost 1 in $h(a_i) \geq r+1) \times Pr($rightmost 1 in $h(a_j) \geq r+1) = \frac{1}{2^r} \times \frac{1}{2^r} = \frac{1}{2^{2r}}$

□

Observations 3

Fix $r \in \{1, \ldots, d\}$. $\forall x \in A$, define indicator r.v:

$$I_x^r = \begin{cases} 1, & \text{if the rightmost } 1 \text{ is at location} \geq r+1 \text{ in } h(x) \\ 0, & \text{otherwise} \end{cases}$$

Let $Z^r = \sum I_x^r$ (sum is over **distinct** elements of A)

Observation 3

The following holds:

1. $E[I_x^r] = \frac{1}{2^r}$
2. $Var[I_x^r] = \frac{1}{2^r}\left(1 - \frac{1}{2^r}\right)$
3. $E[Z^r] = \frac{F_0}{2^r}$
4. $Var[Z^r] \leq E[Z^r]$

Observation 3.1

$E[I_x^r] = \frac{1}{2^r}$

Proof: $E[I_x^r] = 1 \times Pr(I_x^r = 1) + 0 \times Pr(I_x^r = 0) = \frac{1}{2^r}$

Note that $Pr(I_x^r = 1)$ corresponds to
Pr(rightmost 1 in $h(x)$ is at location $\geq r + 1) = \frac{1}{2^r}$ by Observation 1.

□

Observation 3.2

$Var[I_x^r] = E[{I_x^r}^2] - E[I_x^r]^2 = \frac{1}{2^r}\left(1 - \frac{1}{2^r}\right)$

Proof: Note that the variance of a random variable X is given by $Var[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2$.

$E[{I_x^r}^2] = 1^2 Pr(I_x^r = 1) = \frac{1}{2^r}$

Then $E[{I_x^r}^2] - E[I_x^r]^2 = \frac{1}{2^r} - (\frac{1}{2^r})^2 = \frac{1}{2^r}\left(1 - \frac{1}{2^r}\right)$

□

Observation 3.3

$E[Z^r] = \frac{F_0}{2^r}$

Proof: Let $A' \subseteq A$ be the set of distinct elements of A.

Note that $F_0 = |A'|$.

By definition $Z^r = \sum_{x \in A'} I_x^r$

Then, $E[Z^r] = E[\sum_{x \in A'} I_x^r] = \sum_{x \in A'} E[I_x^r] = \sum_{x \in A'} \frac{1}{2^r} = \frac{F_0}{2^r}$

$\square$

Observation 3.4

$Var[Z^r] = \frac{F_0}{2^r}\left(1 - \frac{1}{2^r}\right) \leq \frac{F_0}{2^r} = E[Z^r]$

Proof: $Var[Z^r] = Var[\sum_{x \in A'} I_x^r]$

For two independent random variables X and Y, $Var[X + Y] = Var[X] + Var[Y]$.

$Var[Z^r] = Var[\sum_{x \in A'} I_x^r] = \sum_{x \in A'} Var[I_x^r] = F_0 \frac{1}{2^r}\left(1 - \frac{1}{2^r}\right) \leq \frac{F_0}{2^r} = E[Z^r]$

□

Observation 4

Observation 4

If $2^r > cF_0$, $Pr(Z^r > 0) < \frac{1}{c}$

Proof: Recall Markov's inequality for a random variable X, $Pr(X \geq s) \leq \frac{E[X]}{s}$, where $s > 0$ and X takes positive values.

What is the number of distinct elements $x \in A$, whose hash map $h(x)$ has its rightmost 1 in position $\geq r+1$?

$$= Z^r = \sum_{x \in A'} I_x^r$$

What is $Pr(Z^r > 0)$? $\Leftrightarrow$ What is $Pr(Z^r \geq 1)$.

By Markov's inequality: $Pr(Z^r \geq 1) \leq \frac{E[Z^r]}{1} = E[Z^r] = \frac{F_0}{2^r} < \frac{1}{c}$.

$\square$

Chebyshev's Inequality

Chebyshev's Inequality

$Pr(|X - E[X]| \geq \alpha) \leq \frac{Var[X]}{\alpha^2}$

Proof: Recall Markov's inequality for a random variable X, $Pr(X \geq s) \leq \frac{E[X]}{s}$, where $s > 0$ and X takes positive values.

Now

$$\begin{aligned} Pr(|X - E[X]| \geq \alpha) &= Pr((X - E[X])^2 \geq \alpha^2) \\ &\leq \frac{E[(X - E[X])^2]}{\alpha^2} \\ &= \frac{Var[X]}{\alpha^2} \end{aligned}$$

□

Observation 5

If $c2^r < F_0$, $Pr(Z^r = 0) < \frac{1}{c}$

Proof: Recall Chebyshev's inequality $Pr(|X - E[X]| \geq \alpha) \leq \frac{Var[X]}{\alpha^2}$.

For a random variable X, $Pr(X = 0) \leq Pr(|X - E[X]| \geq E[X])$, as the event $|X - E[X]| \geq E[X]$ includes $X \leq 0$ and $X \geq 2E[X]$.

Now, $Pr(Z^r = 0) \leq Pr(|Z^r - E[Z^r]| \geq E[Z^r])$.

$$\begin{aligned} Pr(Z^r = 0) &\leq Pr(|Z^r - E[Z^r]| \geq E[Z^r]) \\ &\leq \frac{Var[Z^r]}{E[Z^r]^2} \\ &\leq \frac{E[Z^r]}{E[Z^r]^2} \\ &= \frac{1}{E[Z^r]} = \frac{2^r}{F_0} < \frac{1}{c} \end{aligned}$$

□

Observation 6

Claim

Set $\hat{F}_0 = 2^R$. We have $Pr\left(\frac{1}{c} \leq \frac{\hat{F}_0}{F_0} \leq c\right) \geq 1 - \frac{2}{c}$

Proof We have that
Observation 4: if $2^r > cF_0$, $Pr(Z^r > 0) < \frac{1}{c}$
Observation 5, if $c2^r < F_0$, $Pr(Z^r = 0) < \frac{1}{c}$

When do we produce a wrong answer?
Case 1: $\hat{F}_0 = 2^R > cF_0$, but this happens with $Pr(Z^R > 0) < \frac{1}{c}$
Case 2: $c2^R = c\hat{F}_0 < F_0$, but this happens with $Pr(Z^R = 0) < \frac{1}{c}$

Therefore, with probability $\leq \frac{2}{c}$, we produce a wrong answer.

$\implies$ with probability $\geq 1 - \frac{2}{c}$, we produce the right answer, i.e.,

$Pr\left(\frac{1}{c} \leq \frac{\hat{F}_0}{F_0} \leq c\right) \geq 1 - \frac{2}{c}$

$\square$

Further Improvements

Improving success probability

Execute the algorithm s times in parallel
(with independent hash functions)
Let R to the median value among these runs
Return $\hat{F_0} = 2^R$

Note: Algorithm uses $O(s \log u)$ bits.

Claim

For $c > 4$, there exists $s = O(\log \frac{1}{\epsilon})$, $\epsilon > 0$, such that $Pr(\frac{1}{c} \leq \frac{\hat{F_0}}{F_0} \leq c) \geq 1 - \epsilon$.

Technique: Median + Chernoff Bounds

Improving success probability (contd.)

i-th Run of the Algorithm:

Step 1: Initialize $R_i := 0$

Step 2: For each elements $a_i \in A$ do:

1. Compute binary representation of $h(a_i)$
2. Let r be the location of the rightmost 1 in the binary representation
3. if $r > R_i$, $R_i := r$

Step 3: Return R_i

Let $R = \text{Median}(R_1, R_2, \ldots, R_s)$

Indicator Random Variables

Define $X_1, \ldots, X_s$ be indicator random variables:

$$X_i = \begin{cases} 0, & \text{if success, i.e. } \frac{1}{c} \leq \frac{2^{R_i}}{F_0} \leq c \\ 1, & \text{otherwise} \end{cases}$$

1. $E[X_i] = Pr(X_i = 1) \leq \frac{2}{c} = \beta < \frac{1}{2}$ (Since $c > 4$)
2. Let $X = \sum_{i=1}^{s} X_i =$ Number of failures in s runs
3. $E[X] \leq s\beta < \frac{s}{2}$
4. If $X < \frac{s}{2}$, then $\frac{1}{c} \leq \frac{2^R}{F_0} \leq c$
   ($R =$ Median$(R_1, R_2, \ldots, R_s)$)

Chernoff Bounds

If r.v. X is sum of independent identical indicator r.v. and $0 < \delta < 1$,
$Pr(X \geq (1+\delta)E[X]) \leq e^{-\frac{\delta^2 E[X]}{3}}$

Proof: See my notes

An example: Toss a fair coin n-times. Let X be the total number of heads obtained in these n-trials. Evaluate $Pr(X \geq \frac{3}{4}n)$

$$
\begin{aligned}
Pr(X \geq \frac{3}{4}n) &= Pr(X \geq (1+\frac{1}{2})\frac{n}{2}) \\
&= Pr(X \geq (1+\frac{1}{2})E[X]) \\
&\leq e^{-\frac{(\frac{1}{2})^2 E[X]}{3}} \\
&= e^{-\frac{n}{24}}
\end{aligned}
$$

Main Result

Claim

For any $\epsilon > 0$, if $s = O(\log \frac{1}{\epsilon})$, $Pr(X < \frac{s}{2}) \geq 1 - \epsilon$

Proof: We show that $Pr(X \geq \frac{s}{2}) < \epsilon$.

$E[X] = s\beta < \frac{s}{2}$

$$\begin{aligned} Pr(X \geq \frac{s}{2}) &= Pr(X - E[X] \geq \frac{s}{2} - E[X]) \\ &= Pr(X - E[X] \geq \frac{s}{2} - s\beta) \\ &= Pr(X - E[X] \geq \frac{\frac{1}{2} - \beta}{\beta} s\beta) \\ &= Pr(X - E[X] \geq \frac{\frac{1}{2} - \beta}{\beta} E[X]) \\ &= Pr(X \geq \left(1 + \frac{\frac{1}{2} - \beta}{\beta}\right) E[X]) \end{aligned}$$

Proof (contd.)

$$Pr(X \geq \frac{s}{2}) = Pr(X \geq \left(1 + \frac{\frac{1}{2} - \beta}{\beta}\right) E[X])$$
$$\leq e^{-\frac{1}{3}\left(\frac{\frac{1}{2}-\beta}{\beta}\right)^2 E[X]}$$

We want $e^{-\frac{1}{3}\left(\frac{\frac{1}{2}-\beta}{\beta}\right)^2 E[X]} \leq \epsilon$

Substitute $E[X] = s\beta$ and we have

$-\frac{1}{3}\left(\frac{\frac{1}{2}-\beta}{\beta}\right)^2 s\beta \leq \ln \epsilon$

$\Leftrightarrow s \geq \frac{3}{\beta}\left(\frac{\beta}{\frac{1}{2}-\beta}\right)^2 \ln \frac{1}{\epsilon}$

$\implies$ if $s \in O(\ln \frac{1}{\epsilon})$, $Pr(X \geq \frac{s}{2}) < \epsilon$.

$\square$

Estimating F_2

Estimating F_2

Input: Stream A and hash function $h : U \to \{-1, +1\}$

Output: Estimate $\hat{F}_2$ of $F_2 = \sum_{i=1}^{u} m_i^2$

Algorithm (Tug of War)

Step 1: Initialize $Y := 0$.

Step 2: For each element $x \in U$, evaluate $r_x = h(x)$.

Step 3: For each element $a_i \in A$, $Y := Y + r_{a_i}$

Step 4: Return $\hat{F}_2 = Y^2$

Correctness

Observation 1

$E[r_i] = 0$

Proof: $E[r_i] = -1 \times \frac{1}{2} + 1 \times \frac{1}{2} = 0$

□

Observation 2

Observation 2

Let $Y = \sum_{i=1}^{u} r_i m_i$

$E[Y^2] = \sum_{i=1}^{u} m_i^2 = F_2$

Proof:

$$
\begin{aligned}
E[Y^2] &= E[\sum_{i=1}^{u} r_i m_i \sum_{j=1}^{u} r_j m_j] \\
&= E[\sum_{i=1}^{u} r_i^2 m_i^2 + \sum_{i,j:i \neq j} r_i r_j m_i m_j] \\
&= \sum_{i=1}^{u} E[r_i^2 m_i^2] + \sum_{i,j:i \neq j} E[r_i r_j m_i m_j] \\
&= \sum_{i=1}^{u} E[m_i^2] + \sum_{i,j:i \neq j} m_i m_j E[r_i] E[r_j] \\
&= \sum_{i=1}^{u} m_i^2 = F_2
\end{aligned}
$$

Observation 3

$Pr\left(|Y^2 - E[Y^2]| \geq \sqrt{2}cE[Y^2]\right) \leq \frac{1}{c^2}$ for any positive constant c. (I.e., Y^2 approximates $F_2 = E[Y^2]$ within a constant factor with Pr $\geq 1 - \frac{1}{c^2}$)

Proof: Recall Chebyshev's inequality $Pr(|X - E[X]| \geq \alpha) \leq \frac{Var[X]}{\alpha^2}$.

Now, $Pr\left(|Y^2 - E[Y^2]| \geq \sqrt{2}cE[Y^2]\right) \leq \frac{Var[Y^2]}{(\sqrt{2}cE[Y^2])^2}$.

$Var[Y^2] = E[Y^4] - E[Y^2]^2$

$$\begin{aligned} E[Y^4] &= E\left[\sum_{i=1}^{u} r_i m_i \sum_{j=1}^{u} r_j m_j \sum_{k=1}^{u} r_k m_k \sum_{l=1}^{u} r_l m_l\right] \\ &= \sum_{i=1}^{u} E[r_i^4 m_i^4] + 6 \sum_{1 \leq i < j \leq u} E[r_i^2 r_j^2 m_i^2 m_j^2] \\ &= \sum_{i=1}^{u} m_i^4 + 6 \sum_{1 \leq i < j \leq u} m_i^2 m_j^2 \end{aligned}$$

Observation 4 contd.

$$
\begin{aligned}
Var[Y^2] &= E[Y^4] - E[Y^2]^2 \\
&= \sum_{i=1}^{u} m_i^4 + 6 \sum_{1\le i<j\le u} m_i^2 m_j^2 - \left(\sum_{i=1}^{u} m_i^2\right)^2 \\
&= 4 \sum_{1\le i<j\le u} m_i^2 m_j^2 \\
&\le 2F_2^2
\end{aligned}
$$

Now, $\frac{Var[Y^2]}{(\sqrt{2}cE[Y^2])^2} = \frac{2F_2^2}{(\sqrt{2}cE[Y^2])^2} = \frac{2F_2^2}{2c^2F_2^2} = \frac{1}{c^2}$

Thus, $Pr\left(|Y^2 - E[Y^2]| \ge \sqrt{2}cE[Y^2]\right) \le \frac{Var[Y^2]}{(\sqrt{2}cE[Y^2])^2} = \frac{1}{c^2}$

□

Improving Variance

Improving the Variance

Execute the algorithm k times (using independent hash functions) resulting in $Y_1^2, Y_2^2, \ldots, Y_k^2$.

Output $\bar{Y}^2 = \frac{1}{k} \sum_{i=1}^{k} Y_i^2$

Observations:

1. $E[\bar{Y}^2] = E[Y^2] = F_2$
2. $Var[\bar{Y}^2] = \frac{1}{k} Var[Y^2]$
   (Note: $Var[cX] = c^2 Var[X]$)
3. $Pr\left(|\bar{Y}^2 - E[\bar{Y}^2]| \geq \sqrt{\frac{2}{k}} c E[\bar{Y}^2]\right) \leq \frac{1}{c^2}$
4. Set $k = O(\frac{1}{\epsilon^2})$, we have
   $Pr\left(|\bar{Y}^2 - E[\bar{Y}^2]| \geq \epsilon c E[\bar{Y}^2]\right) \leq \frac{1}{c^2}$

Complexity

Space Complexity

Algorithm (Tug of War)

Step 1: Initialize $Y := 0$.

Step 2: For each element $x \in U$, evaluate $r_x = h(x)$.

Step 3: For each element $a_i \in A$, $Y := Y + r_{a_i}$

Step 4: Return $\hat{F}_2 = Y^2$

- Need to store Y and $(r_1, r_2, \ldots, r_u)$.
  Y requires $O(\log n)$ bits.
- We needed r_i's to be 2-wise and 4-wise independent hash functions.
- 4-wise independent functions can be maintained using $O(\log u)$ bits.
- Total space required is $O(\log n + \log u)$.

References

1. The space complexity of estimating frequency moments by Noga Alon, Yossi Matias, and Mario Szegedy, Journal of Computer Systems and Science, 1999.
2. Probabilistic Counting by Philippe Flajolet and G. Nigel Martin, 24th Annual Symposium on Foundations of Computer Science, 1983.
3. Notes on Algorithm Design by A.M
4. Several Lecture Notes (Tim Roughgarden, Ankush Moitra, Lap Chi Lau, Yufei Tao, John Augustine,...)