CS 49/149: 21st Century Algorithms (Fall 2018): Lecture 14

Date: 30th October, 2018

Topic: Streaming. Estimating the second moment – two algorithms

Scribe: Chao Chen

Disclaimer: These notes have not gone through scrutiny and in all probability contain errors. Please email errors to chao.chen.gr@dartmouth.edu.

1 Recap

1.1 Frequency Moment Estimation

Think of we have a really large array A[1...n] coming in as a stream. Once an element leaves the stream, it is gone. Elements of the array are selected from the set ranging from 1 to m. Thus, A[i]∈{1, 2, ..., m}. And f_j is the number of occurrences of j in the stream for j∈{1, 2, ..., m}.
We want to estimate K^{th} moment:

$$F_k = \sum_{j=1}^{m} f_j^k$$

1.2 Estimation Algorithm

We want to develop an algorithm which returns a random variable **Z** such that

- Unbiased:

$$\mathbb{E}(Z) = \text{"what we want"}$$

- Error:

$$\text{Additive Error:} \quad \mathbf{Pr}[|Z - \mathbb{E}(Z)| \geq \epsilon] \leq \delta \qquad (1)$$
$$\text{Multiplicative Error:} \quad \mathbf{Pr}[Z \notin \mathbb{E}(Z)(1+\epsilon)] \leq \delta \qquad (2)$$

From last class, we know if we want (1), we need $\mathrm{Var}(Z)\frac{1}{\epsilon^2}\ln\frac{1}{\delta}$ samples. If we want (2), we need $\frac{\mathrm{Var}(Z)}{\mathbb{E}(Z)^2}\frac{1}{\epsilon^2}\ln\frac{1}{\delta} \leq \frac{\mathbb{E}(Z^2)}{\mathbb{E}(Z)^2}\frac{1}{\epsilon^2}\ln\frac{1}{\delta}$ samples.

2 Estimate F_2

2.1 Problem

We want to estimate $F_2 = \sum_{j=1}^{m} f_j^2$ under the assumption that $n \approx m$. Consider this two cases:

1. The elements have approximate equal frequency, i.e. for each j, $f_j \approx \frac{n}{m}$:

$$F_2 = m \cdot (\frac{n}{m})^2 = \frac{n^2}{m} \approx n$$

2. Some elements have large frequency while others are small, i.e for some j, f_j is really large (we call this a "surprise factor"), F_2 will be really large.

2.2 Algorithm 1 – Try 1

TRY 1

- Sample a j∈{1, 2, ..., m} u.a.r
- Count/Evaluate f_j
- return $Z = mf_j^2$

Analysis: We need to evaluate the bound of the number of samples Try1 need. Remember that it's bound by $\frac{\mathbb{E}(Z^2)}{\mathbb{E}(Z)^2}\frac{1}{\epsilon^2}\ln\frac{1}{\delta}$.
Calculate $\mathbb{E}(Z)$:

$$\begin{aligned}\mathbb{E}(Z) &= \sum_{j=1}^{m}\mathbf{Pr}[j\ is\ sampled]\cdot mf_j^2\\ &= \sum_{j=1}^{m}\frac{1}{m}\cdot mf_j^2\\ &= \sum_{j=1}^{m}f_j^2\end{aligned}$$

Calculate $\mathbb{E}(Z^2)$:

$$\begin{aligned}\mathbb{E}(Z^2) &= \sum_{j=1}^{m}\mathbf{Pr}[j\ is\ sampled]\cdot (mf_j^2)^2\\ &= \sum_{j=1}^{m}\frac{1}{m}\cdot (mf_j^2)^2\\ &= m\sum_{j=1}^{m}f_j^4\end{aligned}$$

Consider two cases:

- For each j, $j = \frac{n}{m}$:

  $$\frac{\mathbb{E}(Z^2)}{\mathbb{E}(Z)^2} = \frac{m\sum_{j=1}^{m}f_j^4}{(\sum_{j=1}^{m}f_j^2)^2} = \frac{m(\frac{n}{m})^4 m}{((\frac{n}{m})^2 m)^2} = 1$$

  This is good, since the bound of the number of samples is $\frac{1}{\epsilon^2}\ln\frac{1}{\delta}$.

- $f_1 = n$ and for all j∈{2, ..., m}, $f_j = 0$:

  $$\frac{\mathbb{E}(Z^2)}{\mathbb{E}(Z)^2} = \frac{mn^4}{n^4} = m$$

  This is not good, since the bound of number of samples is $m\frac{1}{\epsilon^2}\ln\frac{1}{\delta}$. If m is large, it's going to be large.

2.3 Algorithm 1 – Try 2

<u>TRY 2</u>

- Sample a j∈{1, 2, ..., m} $\propto f_j$ i.e. $\frac{f_j}{n}$ (suppose we know $\frac{f_j}{n}$)
- Count/Evaluate f_j
- return $Z = nf_j$

Analysis: Again, we need to evaluate the bound of the number of samples Try2 need.
Calculate $\mathbb{E}(Z)$:

$$\begin{aligned}\mathbb{E}(Z) &= \sum_{j=1}^{m} \mathbf{Pr}[j\ is\ sampled] \cdot nf_j \\ &= \sum_{j=1}^{m} \frac{f_j}{n} \cdot nf_j \\ &= \sum_{j=1}^{m} f_j^2\end{aligned}$$

Calculate $\mathbb{E}(Z^2)$:

$$\begin{aligned}\mathbb{E}(Z^2) &= \sum_{j=1}^{m} \mathbf{Pr}[j\ is\ sampled] \cdot (nf_j)^2 \\ &= \sum_{j=1}^{m} \frac{f_j}{n} \cdot (nf_j)^2 \\ &= n\sum_{j=1}^{m} f_j^3\end{aligned}$$

Consider four cases:

- For each j, $j = \frac{n}{m}$:
  $$\frac{\mathbb{E}(Z^2)}{\mathbb{E}(Z)^2} = \frac{n\sum_{j=1}^{m} f_j^3}{(\sum_{j=1}^{m} f_j^2)^2} = \frac{n(\frac{n}{m})^3 m}{((\frac{n}{m})^2 m)^2} = 1$$
  This is good, since the bound of the number of samples is $\frac{1}{\epsilon^2} \ln \frac{1}{\delta}$.
- $f_1 = n$ and for all j∈{2, ..., m}, $f_j = 0$:
  $$\frac{\mathbb{E}(Z^2)}{\mathbb{E}(Z)^2} = \frac{n\sum_{j=1}^{m} f_j^3}{(\sum_{j=1}^{m} f_j^2)^2} = \frac{n(n)^3}{((n)^2)^2} = 1$$
  This is good as well.

- $f_1 = \frac{n}{2}$, for all j∈{2, ..., n/2}, $f_j = 1$ and for others $f_j = 0$:

$$\frac{\mathbb{E}(Z^2)}{\mathbb{E}(Z)^2} = \frac{n\sum_{j=1}^{m} f_j^3}{(\sum_{j=1}^{m} f_j^2)^2} = \frac{\frac{n^4}{8} + \frac{n^2}{2}}{(\frac{n^2}{4} + \frac{n}{2})^2} \approx constant$$

  This is good as well.

- $f_1 = \sqrt{n}$, for other j, some $f_j = 1$ and some $f_j = 0$ s.t. $\sum_{j=1}^{m} f_j = n$:

$$\frac{\mathbb{E}(Z^2)}{\mathbb{E}(Z)^2} = \frac{n\sum_{j=1}^{m} f_j^3}{(\sum_{j=1}^{m} f_j^2)^2} \approx \frac{n^{\frac{5}{2}}}{n^2} \approx \sqrt{n}$$

  This is not that good, but it's good enough, since the bound of number of samples is approximately $\sqrt{n}\frac{1}{\epsilon^2}\ln\frac{1}{\delta}$.

Try 2 is an acceptable good approach, but the problem is that we don't know $\frac{f_j}{n}$. So Try 3 will propose an approach that give us $\sqrt{n}$ bounds without the knowledge of $\frac{f_j}{n}$.

2.4 Algorithm 1 – Try 3

TRY 3

- Sample a coordinate r∈ A[1...n] u.a.r
- j = A[r]
- $n_j \equiv$ # of occurrence of j in A[r...n]
- return $Z = (2n_j - 1)n$

Analysis:
Calculate $\mathbb{E}(n_j|j)$:
Given that we sampled j, we are equally likely to sample any of the f_j occuerence.

$$\mathbb{E}(n_j|j) = \frac{f_j + (f_j - 1) + (f_j - 2) + ... + 1}{f_j}$$

$$= \frac{f_j(f_j + 1)}{2f_j}$$

$$= \frac{f_j + 1}{2}$$

Calculate $\mathbb{E}(Z|j)$:

$$\mathbb{E}(Z|j) = (2\mathbb{E}(n_j|j) - 1)n$$

$$= nf_j$$

Calculate $\mathbb{E}(Z)$:

$$\mathbb{E}(Z) = \sum_{j=1}^{m} \mathbf{Pr}[sample\ j] \cdot \mathbb{E}(Z|j)$$
$$= \sum_{j=1}^{m} \frac{f_j}{n} \cdot nf_j$$
$$= \sum_{j=1}^{m} f_j^2$$

Calculate $\mathbb{E}(Z^2)$:
Given that we sampled j, $Z \leq 2nf_j$.

$$\mathbb{E}(Z^2) = \sum_{j=1}^{m} \mathbf{Pr}[sample\ j] \cdot \mathbb{E}(Z^2|j)$$
$$\leq \sum_{j=1}^{m} \frac{f_j}{n} \cdot (2nf_j)^2$$
$$= 4n \sum_{j=1}^{m} f_j^3$$
$$\leq 4\sqrt{n} (\sum_{j=1}^{m} f_j^2)^2$$
$$= 4\sqrt{n}\mathbb{E}(Z)^2$$

So $\frac{\mathbb{E}(Z^2)}{\mathbb{E}(Z)^2} \leq 4\sqrt{n}$, this is good, it give us the bound of the number of samples to be $4\sqrt{n}\frac{1}{\epsilon^2} \ln \frac{1}{\delta}$

2.5 Algorithm 2

ALGORITHM 2

- C = 0
- Sample a g: [m]$\longrightarrow$ {1, -1} from a 4-wise independent hash family.
- When element a arrives:

  C = C + g(a)

- return $Z = C^2$

Analysis:
From the algorithm, we know that

$$C = \sum_{a} f_a \cdot g(a)$$

Calculate $\mathbb{E}(Z)$:

$$\begin{aligned}
\mathbb{E}(Z) &= \mathbb{E}(C^2) \\
&= \mathbb{E}[\sum_{a=1}^{m} f_a \cdot g(a))^2] \\
&= \mathbb{E}[\sum_{a=1}^{m} f_a^2 + (\sum_{a \neq b} f_a f_b g(a) g(b)] \\
&= F_2 + \sum_{a \neq b} f_a f_b \mathbb{E}[g(a) g(b)]
\end{aligned}$$

Since g is from 4-wise independent hash family, the second term is 0.

$$= F_2$$

Calculate $\mathbb{E}(Z^2)$:

$$\begin{aligned}
\mathbb{E}(Z^2) &= \mathbb{E}(C^4) \\
&= \mathbb{E}[\sum_{a=1}^{m} (f_a \cdot g(a))^4] \\
&= \mathbb{E}[\sum_{a=1}^{m} f_a^4 + \sum_{a \neq b} f_a^2 f_b^2 + \sum_{a,b,c,d} f_a f_b f_c f_d g(a) g(b) g(c) g(d)] \\
&= \sum_{a=1}^{m} f_a^4 + \sum_{a \neq b} f_a^2 f_b^2 + \sum_{a,b,c,d} f_a f_b f_c f_d \mathbb{E}[g(a) g(b) g(c) g(d)]
\end{aligned}$$

Since g is from 4-wise independent hash family, the third term is 0.

$$= \sum_{a=1}^{m} f_a^4 + \sum_{a \neq b} f_a^2 f_b^2$$

$$= (\sum_{a=1}^{m} f_j^2)^2$$

$$= F_2^2$$

So $\frac{\mathbb{E}(Z^2)}{\mathbb{E}(Z)^2} = 1$, this is good, since it give us the bound of the number of samples to be $\frac{1}{\epsilon^2} \ln \frac{1}{\delta}$ ✍

Question: *Can algorithm 2 be modified to estimate F_3?*