AMS Sampling and Estimating Frequency moments

Lecture 07
February 05, 2019
Frequency Moments

- Stream consists of $e_1, e_2, \ldots, e_m$ where each $e_i$ is an integer in $[n]$. We know $n$ in advance (or an upper bound).
- Given a stream let $f_i$ denote the frequency of $i$ or number of times $i$ is seen in the stream.
- Consider vector $f = (f_1, f_2, \ldots, f_n)$.
- For $k \geq 0$ the $k$'th frequency moment $F_k = \sum_i f_i^k$. We can also consider the $\ell_k$ norm of $f$ which is $(F_k)^{1/k}$.

Example: $n = 5$ and stream is 4, 2, 4, 1, 1, 1, 4, 5

Problem: Estimate $F_k$ from stream using small memory.
A more general estimation problem

- Stream consists of \( e_1, e_2, \ldots, e_m \) where each \( e_i \) is an integer in \([n]\). We know \( n \) in advance (or an upper bound).
- Given a stream let \( f_i \) denote the frequency of \( i \) or number of times \( i \) is seen in the stream.
- Consider vector \( f = (f_1, f_2, \ldots, f_n) \).
- Define a function \( g(\sigma) \) of stream \( \sigma \) to be \( \sum_{i=1}^{m} g_i(f_i) \) where \( g_i : \mathbb{R} \rightarrow \mathbb{R} \) is a real-valued function such that \( g_i(0) = 0 \).
A more general estimation problem

Stream consists of \( e_1, e_2, \ldots, e_m \) where each \( e_i \) is an integer in \([n]\). We know \( n \) in advance (or an upper bound).

Given a stream let \( f_i \) denote the frequency of \( i \) or number of times \( i \) is seen in the stream.

Consider vector \( f = (f_1, f_2, \ldots, f_n) \).

Define a function \( g(\sigma) \) of stream \( \sigma \) to be \( \sum_{i=1}^{m} g_i(f_i) \) where \( g_i : \mathbb{R} \rightarrow \mathbb{R} \) is a real-valued function such that \( g_i(0) = 0 \).

Examples:

- Frequency moments \( F_k \) where for each \( i \), \( g_i(f_i) = h(f_i) \) where \( h(x) = x^k \)
- Entropy of stream: \( g(\sigma) = \sum_i f_i \log(f_i) \)
  (assume \( 0 \log 0 = 0 \))
AMS Sampling

An unbiased statistical estimator for $g(\sigma)$

- Sample $e_J$ uniformly at random from stream of length $m$
- Suppose $e_J = i$ where $i \in [n]$
- Let $R = |\{j \mid J \leq j \leq m, e_j = e_J = i\}|$
- Output $m(g_i(R) - g_i(R - 1))$
AMS Sampling

An unbiased statistical estimator for $g(\sigma)$

- Sample $e_J$ uniformly at random from stream of length $m$
- Suppose $e_J = i$ where $i \in [n]$
- Let $R = \{|j| J \leq j \leq m, e_j = e_J = i\}|$
- Output $m(g_i(R) - g_i(R - 1))$

Can be implemented in streaming setting with reservoir sampling.
AMSEstimate:

\[ s \leftarrow \text{null} \]
\[ m \leftarrow 0 \]
\[ R \leftarrow 0 \]

While (stream is not done)

\[ m \leftarrow m + 1 \]
\[ a_m \text{ is current item} \]
Toss a biased coin that is heads with probability \( \frac{1}{m} \)
If (coin turns up heads)

\[ s \leftarrow a_m \]
\[ R \leftarrow 1 \]

Else If \( a_m == s \)

\[ R \leftarrow R + 1 \]

endWhile

Output \( m(g_s(R) - g_s(R - 1)) \)
Let $Y$ be the output of the algorithm.

**Lemma**

$$E[Y] = g(\sigma) = \sum_{i \in [n]} g_i(f_i).$$
Let $Y$ be the output of the algorithm.

**Lemma**

$$E[Y] = g(\sigma) = \sum_{i \in [n]} g_i(f_i).$$

$$\Pr[e_J = i] = f_i/m$$ since $e_J$ is chosen uniformly from stream.
Let $Y$ be the output of the algorithm.

**Lemma**

$$E[Y] = g(\sigma) = \sum_{i \in [n]} g_i(f_i).$$

$Pr[e_J = i] = f_i/m$ since $e_J$ is chosen uniformly from stream.

$$E[Y] = \sum_{i \in [n]} Pr[a_J = i] E[Y|a_J = i]$$

$$= \sum_{i \in [n]} \frac{f_i}{m} E[Y|a_J = i]$$

$$= \sum_{i \in [n]} \frac{f_i}{m} \left( \sum_{\ell=1}^{f_i} m \frac{1}{f_i} (g_i(\ell) - g_i(\ell - 1)) \right)$$

$$= \sum_{i \in [n]} g_i(f_i).$$
Application to estimating frequency moments

Suppose \( g(\sigma) = F_k \) for some \( k > 1 \). That is \( g_i(x) = x^k \) for each \( i \). What is \( \text{Var}(Y) \)?
Suppose $g(\sigma) = F_k$ for some $k > 1$. That is $g_i(x) = x^k$ for each $i$. What is $\text{Var}(Y)$?

**Lemma**

When $g(x) = x^k$ and $k \geq 1$, $\text{Var}[Y] \leq kF_1 F_{2k-1} \leq kn^{1-\frac{1}{k}} F_k^2$. 
Suppose \( g(\sigma) = F_k \) for some \( k > 1 \). That is \( g_i(x) = x^k \) for each \( i \). What is \( \text{Var}(Y) \)?

**Lemma**

When \( g(x) = x^k \) and \( k \geq 1 \), \( \text{Var}[Y] \leq kF_1F_{2k-1} \leq kn^{1-\frac{1}{k}}F_k^2 \).

\[ \mathbb{E}[Y] = F_k \text{ and } \text{Var}(Y) \leq kn^{1-\frac{1}{k}}F_k^2. \]

Hence, if we want to use averaging and Chebyshev we need to average \( h = \Omega\left(\frac{1}{\epsilon^2}kn^{1-\frac{1}{k}}\right) \) parallel runs and space to get a \((1 \pm \epsilon)\) estimate to \( F_k \) with constant probability.
Application to estimating frequency moments

Suppose \( g(\sigma) = F_k \) for some \( k > 1 \). That is \( g_i(x) = x^k \) for each \( i \). What is \( \text{Var}(Y) \)?

**Lemma**

When \( g(x) = x^k \) and \( k \geq 1 \), \( \text{Var}[Y] \leq kF_1F_{2k-1} \leq kn^{1-\frac{1}{k}}F_k^2 \).

\[
\mathbb{E}[Y] = F_k \quad \text{and} \quad \text{Var}(Y) \leq kn^{1-\frac{1}{k}}F_k^2.
\]

Hence, if we want to use averaging and Chebyshev we need to average \( h = \Omega\left(\frac{1}{\epsilon^2} kn^{1-\frac{1}{k}}\right) \) parallel runs and space to get a \((1 \pm \epsilon)\) estimate to \( F_k \) with constant probability.

Not optimal for frequency moments but shows a general estimating mechanism.
Variance calculation

\[
\text{Var}[Y] \leq \mathbf{E}[Y^2] \\
\leq \sum_{i \in [n]} \Pr[a_J = i] \sum_{\ell=1}^{f_i} \frac{m^2}{f_i} (\ell^k - (\ell - 1)^k)^2 \\
\leq \sum_{i \in [n]} \frac{f_i}{m} \sum_{\ell=1}^{f_i} \frac{m^2}{f_i} (\ell^k - (\ell - 1)^k)(\ell^k - (\ell - 1)^k) \\
\leq m \sum_{i \in [n]} \sum_{\ell=1}^{f_i} k\ell^{k-1}(\ell^k - (\ell - 1)^k) \left((x^k - (x - 1)^k)\right) \\
\leq km \sum_{i \in [n]} f_i^{k-1} f_i^k \\
\leq km F_{2k-1} = kF_1 F_{2k-1}. 
\]
Claim: For $k \geq 1$, $F_1 F_{2k-1} \leq n^{1-1/k} (F_k)^2$. 
Claim: For $k \geq 1$, $F_1 F_{2k-1} \leq n^{1-1/k} (F_k)^2$.

The function $g(x) = x^k$ is convex for $k \geq 1$. Implies $\sum_i x_i / n \leq ((\sum_i x_i^k) / n)^{1/k}$.

\[
F_1 F_{2k-1} = (\sum_i f_i)(\sum_i f_i^{2k-1}) \leq (\sum_i f_i)(F_{\infty})^k (\sum_i f_i^k)
\]
\[
\leq (\sum_i f_i)(\sum_i f_i^k)^{\frac{k-1}{k}} (\sum_i f_i^k)
\]
\[
\leq n^{1-1/k} (\sum_i f_i^k)^{1/k} (\sum_i f_i^k)^{\frac{k-1}{k}} (\sum_i f_i^k)
\]
\[
= n^{1-1/k} (F_k)^2
\]