AMS Sampling, Estimating Frequency moments, F_2 Estimation

Lecture 07

September 15, 2020

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 $\mathbf{f} = (f_1, f_2, \ldots, f_n)$
- For $k \geq 0$ the k'th frequency moment $F_k = \sum_i f_i^k$. We can also consider the ℓ_k norm of $\mathbf{f}$ 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 $\mathbf{f} = (f_1, f_2, \ldots, f_n)$
- Define a function $g(\sigma)$ of stream σ to be $\sum_{i=1}^{m} g_i(f_i)$ where $g_i : \mathbb{R} \to \mathbb{R}$ 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 $\mathbf{f} = (f_1, f_2, \ldots, f_n)$
- Define a function $g(\sigma)$ of stream σ 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 $0 \log 0 = 0$)

Part I

AMS Sampling

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 $(g_i(R) - g_i(R-1)) \cdot m$

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 $(g_i(R) - g_i(R-1)) \cdot m$

Can be implemented in streaming setting with reservoir sampling.

Streaming Implementation

```
AMSEstimate:
    s ← null
    m ← 0
    R ← 0
    While (stream is not done)
        m ← m + 1
        a_m is current item
        Toss a biased coin that is heads with probability 1/m
        If (coin turns up heads)
            s ← a_m
            R ← 1
        Else If (a_m == s)
            R ← R + 1
    endWhile
    Output (g_s(R) − g_s(R − 1)) · m
```

Expectation of output

Let Y be the output of the algorithm.

Lemma

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

Expectation of output

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.

Expectation of output

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.

$$\begin{aligned}
\mathbf{E}[Y] &= \sum_{i \in [n]} \Pr[a_J = i]\, \mathbf{E}[Y | a_J = i] \\
&= \sum_{i \in [n]} \frac{f_i}{m} \mathbf{E}[Y | a_J = i] \\
&= \sum_{i \in [n]} \frac{f_i}{m} \sum_{\ell=1}^{f_i} m \frac{1}{f_i} \left(g_i(\ell) - g_i(\ell - 1)\right) \\
&= \sum g_i(f_i).
\end{aligned}$$

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 $Var(Y)$?

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 $Var(Y)$?

Lemma

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

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 $Var(Y)$?

Lemma

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

$\mathbf{E}[Y] = F_k$ and $Var(Y) \leq kn^{1-\frac{1}{k}}F_k^2$. Hence, if we want to use averaging and Cheybyshev we need to average $h = \Omega(\frac{1}{\epsilon^2}kn^{1-\frac{1}{k}})$ parallel runs and space to get a $(1 \pm \epsilon)$ estimate to F_k with constant probability.

Variance calculation

$$
\begin{aligned}
Var[Y] &\leq \mathbf{E}\left[Y^2\right] \\
&\leq \sum_{i\in[n]} \Pr[a_J = i] \sum_{\ell=1}^{f_i} \frac{m^2}{f_i}\left(\ell^k - (\ell-1)^k\right)^2 \\
&\leq \sum_{i\in[n]} \frac{f_i}{m} \sum_{\ell=1}^{f_i} \frac{m^2}{f_i}(\ell^k - (\ell-1)^k)(\ell^k - (\ell-1)^k) \\
&\leq m \sum_{i\in[n]} \sum_{\ell=1}^{f_i} k\ell^{k-1}(\ell^k - (\ell-1)^k) \quad \text{using } x^k - (x-1)^k \leq kx^{k-1} \\
&\leq km \sum_{i\in[n]} f_i^{k-1} f_i^k \\
&\leq kmF_{2k-1} = kF_1 F_{2k-1}.
\end{aligned}
$$

Variance calculation

Claim: For $k \geq 1$, $F_1 F_{2k-1} \leq n^{1-1/k}(F_k)^2$.

Variance calculation

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 $\sum_i x_i/n \leq ((\sum_i x_i^k)/n)^{1/k}$.

$$
\begin{aligned}
F_1 F_{2k-1} &= (\sum_i f_i)(\sum_i f_i^{2k-1}) \leq (\sum_i f_i)(F_\infty)^{k-1}(\sum_i f_i^k) \\
&\leq (\sum_i f_i)(\sum_i f_i^k)^{\frac{k-1}{k}}(\sum_i f_i^k) \\
&\leq n^{1-1/k}(\sum_i f_i^k)^{1/k}(\sum_i f_i^k)^{\frac{k-1}{k}}(\sum_i f_i^k) \\
&= n^{1-1/k}(F_k)^2
\end{aligned}
$$

Worst case is when $f_i = m/n$ for each $i \in [n]$.

Frequency moment estimation

AMS-Estimator shows that F_k can be estimated in $O(n^{1-1/k})$ space.

Question: Can one do better?

Frequency moment estimation

AMS-Estimator shows that F_k can be estimated in $O(n^{1-1/k})$ space.

Question: Can one do better?

- For F_2 and $1 \leq k \leq 2$ one can do $O(polylog(n))$ space!
- For $k > 2$ space complexity is $\tilde{O}(n^{1-2/k})$ which is known to be essentially tight.

Thus a phase transition at $k = 2$.

Part II

F_2 Estimation

Estimating F_2

- 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 $\mathbf{f} = (f_1, f_2, \ldots, f_n)$

Question: Estimate $F_2 = \sum_{i=1}^{m} f_i^2$ in small space.

Using generic AMS sampling scheme we can do this in $O(\sqrt{n} \log n)$ space. Can we do it better?

AMS Scheme for F_2

AMS-F_2-Estimate:

```
Let h : [n] → {−1, 1} be chosen from
    a 4-wise independent hash family 𝓗.
z ← 0
While (stream is not empty) do
    a_j is current item
    z ← z + h(a_j)
endWhile
Output z^2
```

AMS Scheme for F_2

AMS-F_2-Estimate:

```
Let h : [n] → {−1, 1} be chosen from
    a 4-wise independent hash family ℋ.
z ← 0
While (stream is not empty) do
    a_j is current item
    z ← z + h(a_j)
endWhile
Output z^2
```

AMS-F_2-Estimate:

```
Let Y_1, Y_2, ..., Y_n be {−1, +1} random variable that are
    4-wise independent
z ← 0
While (stream is not empty) do
    a_j is current item
    z ← z + Y_{a_j}
endWhile
Output z^2
```

Analysis

$Z = \sum_{i=1}^{n} f_i Y_i$ and output is Z^2

Analysis

$Z = \sum_{i=1}^{n} f_i Y_i$ and output is Z^2

- $\mathbf{E}[Y_i] = 0$ and $Var(Y_i) = \mathbf{E}\left[Y_i^2\right] = 1$
- For $i \neq j$, since Y_i and Y_j are pairwise-independent $\mathbf{E}[Y_i Y_j] = 0$.

Analysis

$Z = \sum_{i=1}^{n} f_i Y_i$ and output is Z^2

- $\mathbf{E}[Y_i] = 0$ and $Var(Y_i) = \mathbf{E}\left[Y_i^2\right] = 1$
- For $i \neq j$, since Y_i and Y_j are pairwise-independent $\mathbf{E}[Y_i Y_j] = 0$.

$$Z^2 = \sum_i f_i^2 Y_i^2 + 2 \sum_{i \neq j} f_i f_j Y_i Y_j$$

and hence

$$\mathbf{E}\left[Z^2\right] = \sum_i f_i^2 = F_2.$$

Variance

What is $Var(Z^2)$?

Variance

What is $Var(Z^2)$?

$$E[Z^4] = \sum_{i\in[n]}\sum_{j\in[n]}\sum_{k\in[n]}\sum_{\ell\in[n]} f_i f_j f_k f_\ell E[Y_i Y_j Y_k Y_\ell].$$

Variance

What is $Var(Z^2)$?

$$E[Z^4] = \sum_{i\in[n]} \sum_{j\in[n]} \sum_{k\in[n]} \sum_{\ell\in[n]} f_i f_j f_k f_\ell E[Y_i Y_j Y_k Y_\ell].$$

4-wise independence implies $\mathbf{E}[Y_i Y_j Y_k Y_\ell] = 0$ if there is a number among i, j, k, ℓ that occurs only once. Otherwise 1.

Variance

What is $Var(Z^2)$?

$$E[Z^4] = \sum_{i\in[n]}\sum_{j\in[n]}\sum_{k\in[n]}\sum_{\ell\in[n]} f_i f_j f_k f_\ell E[Y_i Y_j Y_k Y_\ell].$$

4-wise independence implies $\mathbf{E}[Y_i Y_j Y_k Y_\ell] = 0$ if there is a number among i, j, k, ℓ that occurs only once. Otherwise 1.

$$\begin{aligned} E[Z^4] &= \sum_{i\in[n]}\sum_{j\in[n]}\sum_{k\in[n]}\sum_{\ell\in[n]} f_i f_j f_k f_\ell E[Y_i Y_j Y_k Y_\ell] \\ &= \sum_{i\in[n]} f_i^4 + 6\sum_{i=1}^{n}\sum_{j=i+1}^{n} f_i^2 f_j^2. \end{aligned}$$

Variance

$$
\begin{aligned}
Var(Z^2) &= \mathbf{E}\left[Z^4\right] - (\mathbf{E}\left[Z^2\right])^2 \\
&= F_4 - F_2^2 + 6\sum_{i=1}^{n}\sum_{j=i+1}^{n} f_i^2 f_j^2 \\
&= F_4 - (F_4 + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} f_i^2 f_j^2) + 6\sum_{i=1}^{n}\sum_{j=i+1}^{n} f_i^2 f_j^2 \\
&= 4\sum_{i=1}^{n}\sum_{j=i+1}^{n} f_i^2 f_j^2 \\
&\leq 2F_2^2.
\end{aligned}
$$

Averaging and median trick again

Output is Z^2: and $\mathbf{E}\left[Z^2\right] = F_2$ and $Var(Z^4) \le 2F_2^2$

- Reduce variance by averaging $8/\epsilon^2$ independent estimates. Let Y be the averaged estimator.
- Apply Chebyshev to average estimator. $\Pr[|Y - F_2| \ge \epsilon F_2] \le 1/4$.
- Reduce error probability to δ by independently doing $O(\log(1/\delta))$ estimators above.
- Total space $O(\log(1/\delta)\frac{1}{\epsilon^2}\log n)$

Geometric Interpretation

Observation: The estimation algorithm works even when f_i's can be negative. What does this mean?

Geometric Interpretation

Observation: The estimation algorithm works even when f_i's can be negative. What does this mean?

Richer model:

- Want to estimate a function of a vector $x \in \mathbb{R}^n$ which is initially assume to be the all $\mathbf{0}$'s vector. (previously we were thinking of the frequency vector f)
- Each element e_j of a stream is a tuple (i_j, Δ_j) where $i_j \in [n]$ and $\Delta_i \in \mathbb{R}$ is a real-value: this updates x_{i_j} to $x_{i_j} + \Delta_j$. (Δ_j can be positive or negative)

Algorithm revisited

```
AMS-ℓ2-Estimate:
    Let Y1, Y2, ..., Yn be {-1, +1} random variable that are
        4-wise independent
    z ← 0
    While (stream is not empty) do
        aj = (ij, Δj) is current update
        z ← z + Δj Y_{ij}
    endWhile
    Output z^2
```

Algorithm revisited

```
AMS-ℓ2-Estimate:
    Let Y_1, Y_2, ..., Y_n be {-1, +1} random variable that are
        4-wise independent
    z ← 0
    While (stream is not empty) do
        a_j = (i_j, Δ_j) is current update
        z ← z + Δ_j Y_{i_j}
    endWhile
    Output z^2
```

Claim: Output estimates $||x||_2^2$ where x is the vector at end of stream of updates.

Analysis

$Z = \sum_{i=1}^{n} x_i Y_i$ and output is Z^2

Analysis

$Z = \sum_{i=1}^{n} x_i Y_i$ and output is Z^2

- $\mathbf{E}[Y_i] = 0$ and $Var(Y_i) = \mathbf{E}\left[Y_i^2\right] = 1$
- For $i \neq j$, since Y_i and Y_j are pairwise-independent $\mathbf{E}[Y_i Y_j] = 0$.

$$Z^2 = \sum_i x_i^2 Y_i^2 + 2 \sum_{i \neq j} x_i x_j Y_i Y_j$$

and hence

$$\mathbf{E}\left[Z^2\right] = \sum_i x_i^2 = ||x||_2^2.$$

Analysis

$Z = \sum_{i=1}^{n} x_i Y_i$ and output is Z^2

- $\mathbf{E}[Y_i] = 0$ and $Var(Y_i) = \mathbf{E}\left[Y_i^2\right] = 1$
- For $i \neq j$, since Y_i and Y_j are pairwise-independent $\mathbf{E}[Y_i Y_j] = 0$.

$$Z^2 = \sum_i x_i^2 Y_i^2 + 2 \sum_{i \neq j} x_i x_j Y_i Y_j$$

and hence

$$\mathbf{E}\left[Z^2\right] = \sum_i x_i^2 = ||x||_2^2.$$

And as before one can show that $Var(Z^2) \leq 2(\mathbf{E}\left[Z^2\right])^2$.

Introduction to (Linear) Sketching

A *sketch* of a stream σ is a summary data structure $C(\sigma)$ (ideally of small space) such that the sketch of the composition $\sigma_1 \cdot \sigma_2$ of two streams σ_1 and σ_1 can be computed from $C(\sigma_1)$ and $C(\sigma_2)$. The output of the algorithm is some function of the sketch.

Introduction to (Linear) Sketching

A *sketch* of a stream σ is a summary data structure $C(\sigma)$ (ideally of small space) such that the sketch of the composition $\sigma_1 \cdot \sigma_2$ of two streams σ_1 and σ_1 can be computed from $C(\sigma_1)$ and $C(\sigma_2)$. The output of the algorithm is some function of the sketch.

What is the summary of algorithm for F_2 estimation? Is it a sketch?

Introduction to (Linear) Sketching

A *sketch* of a stream σ is a summary data structure $C(\sigma)$ (ideally of small space) such that the sketch of the composition $\sigma_1 \cdot \sigma_2$ of two streams σ_1 and σ_1 can be computed from $C(\sigma_1)$ and $C(\sigma_2)$. The output of the algorithm is some function of the sketch.

What is the summary of algorithm for F_2 estimation? Is it a sketch?

A sketch is a *linear* sketch if $C(\sigma_1 \cdot \sigma_2) = C(\sigma_1) + C(\sigma_2)$.

Introduction to (Linear) Sketching

A *sketch* of a stream σ is a summary data structure $C(\sigma)$ (ideally of small space) such that the sketch of the composition $\sigma_1 \cdot \sigma_2$ of two streams σ_1 and σ_1 can be computed from $C(\sigma_1)$ and $C(\sigma_2)$. The output of the algorithm is some function of the sketch.

What is the summary of algorithm for F_2 estimation? Is it a sketch?

A sketch is a *linear* sketch if $C(\sigma_1 \cdot \sigma_2) = C(\sigma_1) + C(\sigma_2)$.

Is the sketch for F_2 estimation a linear sketch?

F_2 Estimation as Linear Sketching

Recall that we take average of independent estimators and take median to reduce error. Can we view all this as a sketch?

```
AMS-ℓ2-Sketch:
    ℓ = c log(1/δ)/ϵ^2
    Let M be a ℓ × n matrix with entries in {−1, 1} s.t
        (i) rows are independent and
        (ii) in each row entries are 4-wise independent
    z is a ℓ × 1 vector initialized to 0
    While (stream is not empty) do
        a_j = (i_j, Δ_j) is current update
        z ← z + Δ_j M e_{i_j}
    endWhile
    Output vector z as sketch.
```

M is compactly represented via ℓ hash functions, one per row, independently chosen from 4-wise independent hash familty.

An Application to Join Size Estimation

In Databases an important operation is the "join" operation

- A relation/table r of arity k consists of tuples of size k where each tuple element is from some given type. Example: (netid, uin, last name, first name, dob, address) in a student data base
- Given two relations r and s and a common attribute a one often needs to compute their join $r \bowtie s$ over some common attribute that they share
- $r \bowtie s$ can have size quadratic in size of r and s

Question: Estimate size of $r \bowtie s$ without computing it explicitly. Very useful in database query optimization.

An Application to Join Size Estimation

In Databases an important operation is the "join" operation

- A relation/table r of arity k consists of tuples of size k where each tuple element is from some given type. Example: (netid, uin, last name, first name, dob, address) in a student data base
- Given two relations r and s and a common attribute a one often needs to compute their join $r \bowtie s$ over some common attribute that they share
- $r \bowtie s$ can have size quadratic in size of r and s

Question: Estimate size of $r \bowtie s$ without computing it explicitly.
Very useful in database query optimization.
Estimating $r \bowtie r$ over an attribute a is same as F_2 estimation.
Why?

Sketching: a shift in perspective

- Sketching ideas have many powerful applications in theory and practice
- In particular linear sketches are powerful. Allows one to handle negative entries and deletions. Surprisingly linear sketches are feasible in several settings.
- Connected to dimension reduction (JL Lemma), subspace embeddings and other important topics