$F_2$ Estimation and Intro to Sketching

Lecture 08
February 07, 2019
Part I

$F_2$ 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 $f = (f_1, f_2, \ldots, f_n)$

**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 $h : [n] \rightarrow \{-1, 1\}$ be chosen from a 4-wise independent hash family $\mathcal{H}$.

1. $z \leftarrow 0$
2. While (stream is not empty) do
   - $a_j$ is current item
   - $z \leftarrow z + h(a_j)$
3. endwhile
4. Output $z^2$
AMS Scheme for $F_2$

**AMS-$F_2$-Estimate:**

Let $h : [n] \rightarrow \{-1, 1\}$ be chosen from a 4-wise independent hash family $\mathcal{H}$.

$z \leftarrow 0$

While (stream is not empty) do

- $a_j$ is current item
- $z \leftarrow z + h(a_j)$

endWhile

Output $z^2$

---

**AMS-$F_2$-Estimate:**

Let $Y_1, Y_2, \ldots, Y_n$ be $\{-1, +1\}$ random variables that are 4-wise independent.

$z \leftarrow 0$

While (stream is not empty) do

- $a_j$ is current item
- $z \leftarrow z + Y_{a_j}$

endWhile

Output $z^2$
Analysis

\[ Z = \sum_{i=1}^{n} f_i Y_i \text{ and output is } Z^2 \]
Analysis

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

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

\[ Z = \sum_{i=1}^{n} f_i Y_i \] and output is \( Z^2 \)

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

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

and hence

\[ \mathbb{E}[Z^2] = \sum_i f_i^2 = F_2. \]
What is $\text{Var}(Z^2)$?
What is $\text{Var}(Z^2)$?

$$E[Z^4] = \sum_{i \in [n]} \sum_{j \in [n]} \sum_{k \in [n]} \sum_{\ell \in [n]} f_i f_j f_k f_\ell E[Y_i Y_j Y_k Y_\ell].$$
What is $\text{Var}(Z^2)$?

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

4-wise independence implies $E[Y_i Y_j Y_k Y_\ell] = 0$ if there is a number among $i, j, k, \ell$ that occurs only once. Otherwise 1.
What is $\text{Var}(Z^2)$?

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

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

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

$$= \sum_{i \in [n]} f_i^4 + 6 \sum_{i=1}^n \sum_{j=i+1}^n f_i^2 f_j^2.$$
\[
\text{Var}(Z^2) = E[Z^4] - (E[Z^2])^2
\]

\[
= F_4 - F_2^2 + 6 \sum_{i=1}^{n} \sum_{j=i+1}^{n} f_i^2 f_j^2
\]

\[
= F_4 - (F_4 + 2 \sum_{i=1}^{n} \sum_{j=i+1}^{n} f_i^2 f_j^2) + 6 \sum_{i=1}^{n} \sum_{j=i+1}^{n} f_i^2 f_j^2
\]

\[
= 4 \sum_{i=1}^{n} \sum_{j=i+1}^{n} f_i^2 f_j^2
\]

\[
\leq 2F_2^2.
\]
Output is $Z^2$: and $\mathbb{E}[Z^2] = F_2$ and $\text{Var}(Z^4) \leq 2F_2^2$

- Reduce variance by averaging $\frac{8}{\epsilon^2}$ independent estimates. Let $Y$ be the averaged estimator.
- Apply Chebyshev to average estimator. $\Pr[|Y - F_2| \geq \epsilon F_2] \leq 1/4$.
- Reduce error probability to $\delta$ by independently doing $O(\log(1/\delta))$ estimators above.
- Total space $O(\log(1/\delta) \frac{1}{\epsilon^2} \log n)$
**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 0’s vector. (previously we were thinking of the frequency vector $f$)
- Each element $e_j$ of a stream is a tuple $(i_j, \Delta_j)$ where $i_j \in [n]$ and $\Delta_j \in \mathbb{R}$ is a real-value: this updates $x_{ij}$ to $x_{ij} + \Delta_j$. ($\Delta_j$ can be positive or negative)
Algorithm revisited

AMS-$\ell_2$-Estimate:

Let $Y_1, Y_2, \ldots, Y_n$ be $\{-1, +1\}$ random variables that are 4-wise independent.

$z \leftarrow 0$

While (stream is not empty) do

$a_j = (i_j, \Delta_j)$ is current update

$z \leftarrow z + \Delta_j Y_{i_j}$

endWhile

Output $z^2$
Algorithm revisited

AMS-$\ell_2$-Estimate:

Let $Y_1, Y_2, \ldots, Y_n$ be $\{-1, +1\}$ random variables that are 4-wise independent

$z \leftarrow 0$

While (stream is not empty) do

$a_j = (i_j, \Delta_j)$ is current update

$z \leftarrow z + \Delta_j Y_{i_j}$

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

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

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

and hence

$$E[Z^2] = \sum_i x_i^2 = ||x||_2^2.$$
Analysis

\[ Z = \sum_{i=1}^{n} x_i Y_i \] and output is \( Z^2 \)

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

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

and hence

\[ \mathbb{E}[Z^2] = \sum_i x_i^2 = \|x\|_2^2. \]

And as before one can show that \( \text{Var}(Z^2) \leq 2(\mathbb{E}[Z^2])^2 \).
A *sketch* of a stream $\sigma$ 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 $\sigma_1$ and $\sigma_1$ can be computed from $C(\sigma_1)$ and $C(\sigma_2)$. The output of the algorithm is some function of the sketch.
A *sketch* of a stream $\sigma$ 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 $\sigma_1$ and $\sigma_1$ 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* of a stream $\sigma$ 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 $\sigma_1$ and $\sigma_1$ 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 $C(\sigma_1 \cdot \sigma_2) = C(\sigma_1) + C(\sigma_2)$. 
A sketch of a stream $\sigma$ 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 $\sigma_1$ and $\sigma_1$ 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?
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:**

\[ ℓ = c \log(1/δ)/\epsilon^2 \]

Let \( M \) be a \( ℓ \times n \) 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 \( ℓ \times 1 \) vector initialized to 0

While (stream is not empty) do

\( a_j = (i_j, Δ_j) \) is current update

\( z ← z + Δ_j Me_{i_j} \)

endWhile

Output vector \( z \) as sketch.

\( M \) is compactly represented via \( ℓ \) hash functions, one per row, independently chosen from 4-wise independent hash family.
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 database

- 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.