Lecture 15 & 16 : ε-net(contd.), ε-approximation and Discrepancy

Lecturer: Kasturi Varadarajan *Scribe: T. Alsulaiman*

Let $\sigma =< a_1, a_2, ..., a_m >$ be a stream; each a_i is a pair (j, c), where $j \in [n]$ and c is an integer–meaning of a_i is: update $f_j \leftarrow f_j + c$, where $i \in [1..m]$.

Algorithm 1 Sketch Algorithm

1. **Initialize**: $C[0..k] \leftarrow [0..0]$ //count vector
2. Choose random hash function $h : [n] \rightarrow [k]$ from a 2-universal process
3. Choose random hash function $g : [n] \rightarrow \{-1, +1\}$ from a 2-universal process
4. **Process** $a_i = (j, c')$
   $C[h(j)] \leftarrow C[h(j)] + c' * g(j)$
5. **Output**: on query a, report
   $\hat{f}_a = g(a) * C[h(a)]$

3.0.1 Analysis

Let e_j be the k-vector with 1 in $h(j)$ co-ordinate, and 0 otherwise. For stream σ,

$$\sigma \rightarrow f = (f_0, f_1, ..., f_{n-1}) \rightarrow C[\sigma]$$

$$\sigma \rightarrow f_0 g(0) e_0 + f_1 g(1) e_1 + ... + f_{n-1} g(n-1) e_{n-1}$$

$$\sigma \rightarrow [|M|] \begin{pmatrix} f_0 \\ f_1 \\ . \\ . \\ . \\ f_{n-1} \end{pmatrix}$$

Definition 3.1 *Fix $\sigma \rightarrow C[\sigma]$. C is a* sketch *if, given 2 streams σ_1 and σ_2, the concatenation of the two streams $C[\sigma_1.\sigma_2]$ can be obtained from $C[\sigma_1]$ and $C[\sigma_2]$*

If $C[\sigma_1] = M * f^{\sigma_1}$, $C[\sigma_2] = M * f^{\sigma_2}$,

$$C[\sigma_1.\sigma_2] = M * f^{\sigma_1.\sigma_2} = M * (f^{\sigma_1} + f^{\sigma_2}) = C[\sigma_1] + C[\sigma_2]$$

Fix $a \in [n]$. Let $X = \hat{f}_a$. Define random variableY_j,

$$Y_j = \begin{cases} 1 & \text{if } h(j) = h(a); \text{ //a and j maps to the same bin in C[ ]} \\ 0 & \text{otherwise.} \end{cases}$$

$$\Rightarrow X = g(a) \sum f_j g(j) Y_j = f_a + \sum_{j \in [n] \setminus \{a\}} f_j g(a) g(j) Y_j$$

Now we compute the expected value of X, then the variance.

$$\begin{aligned}
E[X] &= f_a + \textstyle\sum_{j\in[n]\setminus\{a\}} f_j E[g(a)g(j)Y_j] \\
&= f_a + \textstyle\sum_{j\in[n]\setminus\{a\}} f_j E[g(a)g(j)]E[Y_j] \quad //g() \text{ and } h() \text{ are independent} \\
&= f_a + \textstyle\sum_{j\in[n]\setminus\{a\}} f_j E[g(a)]E[g(j)]E[Y_j] \quad //\text{by pairwise independence} \\
\text{note that } E[g(a)] = E[g(j)] = 0 & \\
&= f_a
\end{aligned}$$

Now we compute the variance.

$$\begin{aligned}
Var[x] &= 0 + Var[\textstyle\sum_{j\in[n]\setminus\{a\}} f_j g(a) g(j) Y_j] \\
&= E[\ (\ \textstyle\sum_{j\in[n]\setminus\{a\}} f_j g(a)g(j)Y_j\ )^2\ ] - E[\ \textstyle\sum_{j\in[n]\{a\}} f_j g(a)g(j)Y_j\ ]^2 \\
&= E[\ (\ \textstyle\sum_{j\in[n]\setminus\{a\}} f_j g(a)g(j)Y_j\ )^2\ ] - 0 \\
&= E[\ \textstyle\sum_{j\in[n]\setminus\{a\}} f_j^2 g(a)^2 g(j)^2 Y_j^2\ +\ \sum_{i,j\in[n]\setminus\{a\}, i\neq j} f_i f_j g(a)^2 g(i) g(j) Y_i Y_j\ ] \\
\text{note that } g(i)^2 = (+1)^2 = (-1)^2 = 1 & \\
&= E[\ \textstyle\sum_{j\in[n]\setminus\{a\}} f_j^2 Y_j^2\ + \sum_{i,j\in[n]\setminus\{a\}, i\neq j} f_i f_j g(i) g(j) Y_i Y_j\ ] \\
&= \textstyle\sum_{j\in[n]\setminus\{a\}} E[\ f_j^2 Y_j^2\ ]\ +\ \sum_{i,j\in[n]\setminus\{a\}, i\neq j} f_i f_j\ E[\ g(i)g(j)Y_iY_j\ ] \\
&= \textstyle\sum_{j\in[n]\setminus\{a\}} f_j^2 E[\ Y_j^2\ ]\ +\ \sum_{i,j\in[n]\setminus\{a\}, i\neq j} f_i f_j\ E[\ g(i)g(j)\ ]\ E[\ Y_iY_j\ ] \\
&= \textstyle\sum_{j\in[n]\setminus\{a\}} f_j^2 E[\ Y_j^2\ ]\ +\ \sum_{i,j\in[n]\setminus\{a\}, i\neq j} f_i f_j\ E[\ g(i)\ ]\ E[\ g(j)\ ]\ E[\ Y_iY_j\ ] \\
&= \textstyle\sum_{j\in[n]\setminus\{a\}} f_j^2 E[\ Y_j^2\ ]\ +\ 0 \\
&= \textstyle\sum_{j\in[n]\setminus\{a\}} f_j^2 E[\ Y_j^2\ ] \qquad //Y_j = 0 \text{ or } 1; Y_j^2 = Y_j \\
&= \textstyle\sum_{j\in[n]\setminus\{a\}} f_j^2 E[\ Y_j\ ] \\
&= \tfrac{1}{k} \textstyle\sum_{j\in[n]\setminus\{a\}} f_j^2 \qquad //Pr[h(j) = h(a)] = \tfrac{1}{k} \\
&= \tfrac{1}{k}(\ ||f||_2^2\ -\ f_a^2\ )
\end{aligned}$$

We now compute the error probability. By Chebyshev's inequality,

$$Pr[\ |\ X - E[X]\ |\ \geq\ \epsilon\sqrt{(\ ||f||_2^2\ -\ f_a^2\ )}\ ] \leq \frac{Var[X]}{\epsilon^2(\ ||f||_2^2\ -\ f_a^2\ )} \leq\ \frac{1}{k\epsilon^2}$$

if $k \geq \frac{3}{\epsilon^2}$,

$$Pr[\ |\ X - E[X]\ |\ \geq\ \epsilon\sqrt{(\ ||f||_2^2\ -\ f_a^2\ )}\ ] \leq \frac{1}{3}$$

Also,

$$Pr[\ |\hat{f}_a\ -\ f_a|\ \geq \epsilon \sum_{j\in[n]} f_j\ ]\ \leq Pr[\ |\ X - E[X]\ |\ \geq\ \epsilon\sqrt{(\ ||f||_2^2\ -\ f_a^2\ )}\ ] \leq \frac{1}{3}$$

3.1 The Tug-of-War Sketch

Problem: We have a stream $a_1, a_2, ..., a_m$, where each a_i has the form (j, c), where $j \in [n]$ and c is an integer. The frequency of element j in the stream is calculated when (j, c) appears in the stream as follows:

$$f_j \leftarrow f_j + c$$

Estimate:

$$F_2 \ = \ \sum_{j \in [n]} f_j^2 \ = \ ||f||_2^2$$

where $f = (f_0, f_1, ..., f_n - 1)$ is the frequency vector of elements appearing in the stream.

The above formula can be generalized for $k \geq 0$ as follows:

$$F_k \ = \ \sum_{j \in [n]} f_j^k$$

Algorithm 2 Tug-of-War Sketch Algorithm

1. **Initialize**:
   $x \leftarrow 0$
   Choose random hash function $h : [n] \rightarrow \{-1, +1\}$ from a 4-universal process
3. **Process** $a_i = (j, c)$
   $x \leftarrow x \ + \ h(j) * c$
5. **Output**: x^2

3.1.1 Analysis

Let X denote x at the end of the stream. Let $Y_j = h(j)$. So, $X \ = \ \sum_{j \in [n]} f_j \ Y_j$.

$$E[X^2] \ = \ \sum_{j \in [n]} f_j^2 \ E[Y_j^2] \ + \sum_{i,j \in [n], i \neq j} f_i^2 \ f_j^2 \ E[Y_i Y_j]$$

note that $E[Y_j^2] = 1$, and by pairwise independence $E[Y_i Y_j] = 0$, hence,

$$E[X^2] \ = \ \sum_{j \in [n]} f_j^2 \ + \ 0 \ = \ F_2$$

$$\Rightarrow var[X^2] \ \leq \ 2F_2^2$$

To reduce the error gap, do:

- Run t parallel, independent copies of $Tug-of-War$ sketch algorithm.
- Return Z, which is the average of the outputs of the t copies.

For Z, $E[Z] \ = \ F_2$, which leads to $var[Z] \ \leq \ \frac{2F_2^2}{t}$.

$$\Rightarrow Pr[|Z - F_2| \ \geq \epsilon F_2] \ \leq \ \frac{var[Z]}{(\epsilon F_2)^2}$$

$$Pr[|Z - F_2| \ \geq \epsilon F_2] \ \leq \frac{2F_2^2}{t\epsilon F_2^2} \ = \ \frac{2}{t\epsilon^2}$$

for $t \ \geq \frac{6}{\epsilon^2}$,

$$Pr[\ |Z - F_2| \ \geq \ \epsilon F_2 \ ] \ \leq 1/3$$

For t copies of the algorithm, with 5 items for example,

$$t \ * \underbrace{\begin{pmatrix} 1, & 1, & -1, & 1, & -1 \\ . & & & & \\ . & & & & \\ . & & & & \\ . & & & & \end{pmatrix}}_{M} * \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \end{pmatrix}$$

$$\Rightarrow Z \ = \ \frac{||Mf||_2^2}{t}$$

where

$$\Rightarrow Z \ = \ \frac{||Mf||_2^2}{t} \ \in \ [(1-\epsilon) \ F_2, \ (1+\epsilon) \ F_2 \ ]$$

by taking square root,

$$\frac{||Mf||_2}{\sqrt{t}} \ \in \ [\sqrt{(1-\epsilon)} \ ||f||_2, \ \sqrt{(1+\epsilon)} \ ||f||_2 \ ]$$

Note: The above operation is called *dimension reduction.* JohnsonLindenstrauss lemma states that a small set of points in a high-dimensional space can be embedded into a space of much lower dimension in such a way that distances between the points are nearly preserved. When $t = \frac{\log n}{\epsilon^2}$, the distance is preserved with high probability.