COL758: Advanced Algorithms	Spring 2019

Lecture 20: April 4

Lecturer: Naveen Garg — *Scribe: Raj Kamal*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

20.1 Important Results

We state some important results which will be used in following sections. We omit the proofs here.

20.1.1 Chebyshev's Inequality

If X is a random variable with mean μ and variance σ^2, then, for any $k > 0$,

$$P\{|X - \mu| \geq k\} \leq \frac{\sigma^2}{k^2}$$

20.1.2 Chernoff Bound

Let $X_1,\ X_2,\ ...,\ X_n$ be independent Poisson trials with $P\{X_i = 1\} = p_i$. If $X = \sum_{i=1}^{n} X_i$ and if $E[X] \leq \mu$, then for any $\eta \in (0, 1]$:

$$P\{X \geq (1+\eta)\mu\} \leq e^{-\frac{\eta^2 \mu}{3}}$$

20.1.3 Mean Value Theorem (MVT)

Let f be a continuous function on $[a,\ b]$ that is differentiable on $(a,\ b)$. Then there exists [at least one] ξ in $(a,\ b)$ such that

$$f'(\xi) = \frac{f(b) - f(a)}{b - a}$$

20.2 Previous Lecture

Following sketching algorithm, called "Count Sketch", was analysed in previous lecture.

Algorithm

Initialize

$C[1...k] \leftarrow \overrightarrow{0}$, where $k := \frac{3}{\epsilon^2}$;
Choose a random hash function $h : [n] \rightarrow [k]$ from a $2-$universal family;
Choose a random hash function $g : [n] \rightarrow \{-1,\ 1\}$ from a $2-$universal family;

Process(j, r)

$C[h(j)] \leftarrow C[h(j)] + r \times g(j)$;

Output

On query a, report $g(a) \times C[h(a)]$;

If we assume that X is a random variable which denotes the value '$g(a) \times C[h(a)]$' returned by the algorithm stated above. Then it has been proved in previous lecture that

$$E[X] = f_a \qquad \textit{and} \qquad Var(X) = \frac{-f_a^2 + \sum_{j\in[n]} f_j^2}{k} \tag{20.1}$$

20.3 The Quality of the Algorithm's Estimate

Let $\overline{f}_a$ and f_a denote estimated and actual frequency respectively of token a. Also let

$$\begin{aligned}(\|f_{-a}\|_2)^2 &= -f_a^2 + \sum_{j\in[n]} f_j^2 \\ (\|f\|_2)^2 &= \sum_{j\in[n]} f_j^2\end{aligned}$$

So equation (20.1) implies $Var(X) = \frac{(\|f_{-a}\|_2)^2}{k}$

Let ϵ be any positive real number. Then Chebyshev's inequality implies

$$P[\ |\bar{f}_a - f_a| \geq \epsilon\sqrt{(\|f_{-a}\|_2)^2}\ ] = P[\ |X - E(X)| \geq \epsilon\sqrt{k\,Var(X)}\ ] \leq \frac{Var(X)}{\epsilon^2 k\,(Var(X))} = \frac{1}{\epsilon^2 k} = \frac{1}{3} \tag{20.2}$$

where we have taken $\epsilon^2 k = 3$.

20.4 Multiple hash functions for better estimate

Algorithm

Initialize

$C[1...t][1...k] \leftarrow \overrightarrow{0}$, where $k := \frac{3}{\epsilon^2}$ and $t := O\left(\log(\frac{1}{\delta})\right)$;

Choose t independent random hash functions $h_1,\ h_2,\ ...,\ h_t : [n] \to [k]$ each from a $2-$universal family;

Choose t independent random hash functions $g_1,\ g_2,\ ...,\ g_t : [n] \to \{-1,\ 1\}$ each from a $2-$universal family;

Process(j, r)

for $i = 1$ **to** t **do** $C[i][h_i(j)] \leftarrow C[i][h_i(j)] + r \times g_i(j)$;

Output

On query a, report $\underset{\mathbf{1\leq i\leq t}}{\textbf{median}}\,(\mathbf{g_i(a)} \times \mathbf{C[i][h_i(a)]})$;

Analysis

If we assume that X_i is a random variable which takes the value $g_i(a) \times C[i][h_i(a)]$. Then equations (20.1) and (20.2) imply that

$$E[X_i] = f_a\ , \qquad Var(X_i) = \frac{-f_a^2 + \sum_{j\in[n]} f_j^2}{k} = \frac{(\|f_{-a}\|_2)^2}{k} \qquad \forall\ i = 1,\ 2,\ ...,\ t. \tag{20.3}$$

$$P[\ |X_i - f_a| \geq \epsilon\sqrt{(\|f_{-a}\|_2)^2}\ ] = P[\ |X_i - E(X_i)| \geq \epsilon\sqrt{k\,Var(X_i)}\ ] \leq \frac{1}{3} \qquad \forall\ i = 1,\ 2,\ ...,\ t. \tag{20.4}$$

For i=1, 2, ..., t; define random variable W_i by

$$W_i = \begin{cases} 1, & if\ |X_i - f_a| \geq \epsilon\sqrt{(\|f_{-a}\|_2)^2}; \\ 0, & \text{otherwise}; \end{cases}$$

Then equation (20.4) implies that $E(W_i) \leq \frac{1}{3},\ \forall\ i = 1,\ 2,\ ...,\ t$. If we define random variable $W = \sum_{i=1}^{t} W_i$, then $E(W) \leq \frac{t}{3}$. Let Z be the random variable which denotes the value $\underset{\mathbf{1 \leq i \leq t}}{\mathbf{median}}\,(\mathbf{g_i(a)} \times \mathbf{C[i][h_i(a)]})$ returned by the algorithm. Then $|Z - f_a| \geq \epsilon\sqrt{(\|f_{-a}\|_2)^2}$ only if more than $\frac{t}{2}$ random variables out of t random variables X_i ($i = 1,\ 2,\ ...,\ t$) satisfy $|X_i - f_a| \geq \epsilon\sqrt{(\|f_{-a}\|_2)^2}$. Hence $|Z - f_a| \geq \epsilon(\|f_{-a}\|_2)$ only if $W > \frac{t}{2}$. Therefore

$$\begin{aligned} P\{\ |Z - f_a| \geq \epsilon(\|f_{-a}\|_2)\ \} &\leq P\{\ W > \frac{t}{2}\ \} \\ &= P\{W > \left(1 + \frac{1}{2}\right) \times \frac{t}{3}\} \\ &\leq e^{-\frac{\left(\frac{1}{2}\right)^2\left(\frac{t}{3}\right)}{3}} \qquad\qquad \textit{using Chernoff Bound} \\ &= e^{-\frac{t}{36}} = \delta \end{aligned}$$

$\Rightarrow$

$$P\{\ |Z - f_a| \geq \epsilon(\|f_{-a}\|_2)\ \} \leq \delta$$

where we have taken $t = 36 \times \log(\frac{1}{\delta})$ i.e. $t = O\left(\log(\frac{1}{\delta})\right)$.

Space Bound

With a suitable choice of hash family, we can store the hash functions above in $O(t\log(n))$ space. Each of the tk counters in the sketch uses $O(\log(m))$ space. This gives us an overall space bound of $O(t\log(n)+tk\log(m))$, which is

$$O\left(\frac{1}{\epsilon^2}.\log\left(\frac{1}{\delta}\right).(\log(m) + \log(n))\right)$$

20.5 Lemma

Let $n > 0$ be an integer and let $x_1,\ x_2,\ ...,\ x_n \geq 0$ and $k \geq 1$ be reals. Then

$$\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} x_i^{2k-1}\right) \leq n^{1-\frac{1}{k}}\left(\sum_{i=1}^{n} x_i^k\right)^2$$

Proof

Let $v = \underset{i\in[n]}{max}\,(x_i)$. Since $v^k \leq \sum_{i=1}^{n} x_i^k$, so we have

$$v^{k-1} = \left(v^k\right)^{\frac{(k-1)}{k}} \leq \left(\sum_{i=1}^{n} x_i^k\right)^{\frac{(k-1)}{k}} \tag{20.5}$$

Let $f(x) = x^k$, then f is convex function on the set of real numbers for $k \geq 1$. Hence

$$f\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \leq \frac{1}{n}\sum_{i=1}^{n} f(x_i)$$

$\Rightarrow$

$$\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^k \leq \frac{1}{n}\sum_{i=1}^{n} x_i^k \qquad \textit{since } f(x) = x^k$$

$\Rightarrow$

$$\frac{1}{n}\sum_{i=1}^{n} x_i \leq \left(\frac{1}{n}\sum_{i=1}^{n} x_i^k\right)^{\frac{1}{k}}$$

$\Rightarrow$

$$\sum_{i=1}^{n} x_i \leq n\left(\frac{1}{n}\right)^{\frac{1}{k}} \left(\sum_{i=1}^{n} x_i^k\right)^{\frac{1}{k}}$$

$\Rightarrow$

$$\sum_{i=1}^{n} x_i \leq n^{1-\frac{1}{k}} \left(\sum_{i=1}^{n} x_i^k\right)^{\frac{1}{k}} \tag{20.6}$$

Hence we have

$$\begin{aligned}
\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} x_i^{2k-1}\right) &\leq \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} x_i^{k-1}x_i^k\right) \\
&\leq \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} v^{k-1}x_i^k\right) && \textit{because} \quad v = \max_{i\in[n]}(x_i) \\
&= \left(\sum_{i=1}^{n} x_i\right) v^{k-1} \left(\sum_{i=1}^{n} x_i^k\right) \\
&\leq \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} x_i^k\right)^{\frac{(k-1)}{k}} \left(\sum_{i=1}^{n} x_i^k\right) && \textit{using } (20.5) \\
&\leq n^{1-\frac{1}{k}} \left(\sum_{i=1}^{n} x_i^k\right)^{\frac{1}{k}} \left(\sum_{i=1}^{n} x_i^k\right)^{\frac{(k-1)}{k}} \left(\sum_{i=1}^{n} x_i^k\right) && \textit{using } (20.6) \\
&= n^{1-\frac{1}{k}} \left(\sum_{i=1}^{n} x_i^k\right)^2
\end{aligned}$$

which completes the proof.

20.6 Higher moments of frequency

For $k > 0$, the kth moment F_k of frequency is defined as $F_k = \sum_{j=1}^{n} f_j^k$

Hence $F_1 = \sum_{j=1}^{n} f_j =$ length of the stream $= m$.

The 0th moment F_0 of frequency is defined as $F_0 = \sum_{\substack{j=1 \\ f_j>0}}^{n} f_j^0 =$ number of distinct tokens in the stream.

20.7 Algorithm and quality of its estimate

Algorithm

Pick a random element in the stream. If this is token a, then count number of occurrences of token a beyond this point. Let this count be r. Return $m\{r^k - (r-1)^k\}$

Analysis

Since algorithm picks the token randomly so r is the value of some random variable Y (say). So Y is a random variable which denotes the total number of remaining occurrences (start counting from where the token is picked and go towards the end of the stream) of picked token in the stream. Also let X be a random variable which denotes the value $m\{r^k - (r-1)^k\}$ returned by the algorithm. Then $X = m\{r^k - (r-1)^k\}$ if and only if $Y = r$. Also let A be a random variable which denotes the picked token. Then we have

$$P\{X = m\{r^k - (r-1)^k\} \mid A = j\} = \mathbf{P\{Y = r \mid A = j\}} = \mathbf{\frac{1}{f_j}} \qquad \textit{where} \;\; r = 1,\ 2,\ 3,\ ...,\ f_j$$

since f_j is the frequency of token j and one of these f_j occurrences of token j is picked. We also have

$$\mathbf{P\{A = j\}} = \mathbf{\frac{f_j}{m}}$$

since token j occurs f_j times in the stream of length m.

If x denotes value taken by random variable X, then we have

$$\begin{aligned}
E(X \mid \textit{token } j \textit{ is picked}) &= \sum_x x \times P(X = x \mid \textit{token } j \textit{ is picked}) \\
&= \sum_x x \times P(X = x \mid A = j) \\
&= \sum_{r=1}^{f_j} m\{r^k - (r-1)^k\} \times P\{X = m\{r^k - (r-1)^k\} \mid A = j\} \\
&= \sum_{r=1}^{f_j} m\{r^k - (r-1)^k\} \times \mathbf{P\{Y = r \mid A = j\}} \\
&= \sum_{r=1}^{f_j} m\{r^k - (r-1)^k\} \times \mathbf{\frac{1}{f_j}} \\
&= \frac{m}{f_j} \times f_j^k \\
&= m \times f_j^{k-1}
\end{aligned}$$

So we have

$$\begin{aligned}
E(X) &= \sum_{j=1}^{n} E(X|A = j) \times P(A = j) \\
&= \sum_{j=1}^{n} E(X \mid \textit{token } j \textit{ is picked}) \times \mathbf{P(A = j)} \\
&= \sum_{j=1}^{n} m \times f_j^{k-1} \times \mathbf{\frac{f_j}{m}} \\
&= \sum_{j=1}^{n} f_j^k \\
&= F_k
\end{aligned}$$

We also have

$$
\begin{aligned}
Var(X) &\le E(X^2) \\
&= \sum_{j=1}^{n} \mathbf{P}(\mathbf{A} = \mathbf{j}) \times E(X^2 \mid A = j) \\
&= \sum_{j=1}^{n} \left(\frac{\mathbf{f_j}}{\mathbf{m}} \times \sum_{x} x^2 \times P(X = x \mid A = j) \right) \\
&= \sum_{j=1}^{n} \left(\frac{f_j}{m} \times \sum_{r=1}^{f_j} m^2 \{r^k - (r-1)^k\}^2 \times \mathbf{P}(\mathbf{Y} = \mathbf{r} \mid \mathbf{A} = \mathbf{j}) \right) \\
&= \sum_{j=1}^{n} \left(\frac{f_j}{m} \times \sum_{r=1}^{f_j} m^2 \{r^k - (r-1)^k\}^2 \times \frac{\mathbf{1}}{\mathbf{f_j}} \right) \\
&= m \times \sum_{j=1}^{n} \left(\sum_{r=1}^{f_j} \{r^k - (r-1)^k\}^2 \right) \\
&= m \times \sum_{j=1}^{n} \left(\sum_{r=1}^{f_j} \{r^k - (r-1)^k\} \times \{r^k - (r-1)^k\} \right) \\
&= m \times \sum_{j=1}^{n} \left(\sum_{r=1}^{f_j} k \times \xi^{k-1} \times \{r^k - (r-1)^k\} \right) && \textbf{using MVT with } \mathbf{f(x) = x^k} \textbf{ and } \xi \in (\mathbf{r-1},\ \mathbf{r}) \\
&< m \times \sum_{j=1}^{n} \left(\sum_{r=1}^{f_j} k \times r^{k-1} \times \{r^k - (r-1)^k\} \right) && \textbf{because } \xi < \mathbf{r} \\
&\le m \times \sum_{j=1}^{n} \left(\sum_{r=1}^{f_j} k \times f_j^{k-1} \times \{r^k - (r-1)^k\} \right) && \textbf{because } \mathbf{r} \le \mathbf{f_j} \\
&= m \times k \times \sum_{j=1}^{n} f_j^{k-1} \left(\sum_{r=1}^{f_j} \{r^k - (r-1)^k\} \right) \\
&= m \times k \times \sum_{j=1}^{n} f_j^{k-1} f_j^k \\
&= k \times \left(\sum_{j=1}^{n} f_j \right) \times \left(\sum_{j=1}^{n} f_j^{2k-1} \right) && because\ \ m = length\ of\ the\ stream = \left(\sum_{j=1}^{n} f_j \right) \\
&\le k \times n^{1-\frac{1}{k}} \times \left(\sum_{j=1}^{n} f_j^k \right)^2 && Lemma\ 20.5 \\
&= kn^{1-\frac{1}{k}} \left(F_k\right)^2
\end{aligned}
$$

Hence we have proved that $E(X) = F_k$ and $Var(X) < kn^{1-\frac{1}{k}} \left(F_k\right)^2$. Therefore Chebyshev's Inequality implies that

$$Pr[\ |\overline{F}_k - F_k| \ge tF_k\ ] = Pr[\ |X - E(X)| \ge tE(X)\ ] \le \frac{Var(X)}{t^2\,(E(X))^2} < \frac{kn^{1-\frac{1}{k}} F_k^2}{t^2 F_k^2} = \frac{1}{2}$$

where we have chosen t such that $t^2 = 2kn^{1-\frac{1}{k}}$

20.8 Median of Means

Algorithm

Suppose using the algorithm given in previous section, we find st estimates X_{ij} for $i = 1,\ 2,\ ...,\ t$ and $j = 1,\ 2,\ ..., s$. Return $\underset{\mathbf{1\leq i\leq t}}{\mathbf{median}}\left(\frac{\sum_{\mathbf{j=1}}^{\mathbf{s}}\mathbf{X_{ij}}}{\mathbf{s}}\right)$

Analysis

We have $E(X_{ij}) = F_k$ and $Var(X_{ij}) \leq kn^{1-\frac{1}{k}}\left(F_k\right)^2$. Let $X_i = \frac{\sum_{j=1}^{s} X_{ij}}{s}$, then $E(X_i) = F_k$ and $Var(X_i) \leq \frac{kn^{1-\frac{1}{k}}(F_k)^2}{s}$.

$$\begin{aligned} P\{|X_i - F_k| \geq \epsilon F_k\} &= P\{|X_i - E(X_i)| \geq \epsilon F_k\} \\ &\leq \frac{kn^{1-\frac{1}{k}}\left(F_k\right)^2}{s\epsilon^2(F_k)^2} \qquad\qquad using\ Chebyshev's\ Inequality \\ &= \frac{1}{3} \end{aligned}$$

where we have taken $s = \frac{3kn^{1-\frac{1}{k}}}{\epsilon^2}$. Therefore

$$P\{|X_i - F_k| \geq \epsilon F_k\} \leq \frac{1}{3} \qquad\qquad \forall\ i = 1,\ 2,\ ...,\ t. \tag{20.7}$$

For i=1, 2, ..., t; define random variable W_i by

$$W_i = \begin{cases} 1, & if\ |X_i - F_k| \geq \epsilon F_k; \\ 0, & \text{otherwise}; \end{cases}$$

Then equation (20.7) implies that $E(W_i) \leq \frac{1}{3},\ \forall\ i = 1,\ 2,\ ...,\ t$. If we define random variable $W = \sum_{i=1}^{t} W_i$, then $E(W) \leq \frac{t}{3}$. Let Z be the random variable which denotes the value $\underset{\mathbf{1\leq i\leq t}}{\mathbf{median}}\left(\frac{\sum_{\mathbf{j=1}}^{\mathbf{s}}\mathbf{X_{ij}}}{\mathbf{s}}\right)$ returned by the algorithm. Then $|Z - F_k| \geq \epsilon F_k$ only if more than $\frac{t}{2}$ random variables out of t random variables X_i ($i = 1,\ 2,\ ...,\ t$) satisfy $|X_i - F_k| \geq \epsilon F_k$. Hence $|Z - F_k| \geq \epsilon F_k$ only if $W > \frac{t}{2}$. Therefore

$$\begin{aligned} P\{|X_i - F_k| \geq \epsilon F_k\} &\leq P\{\ W > \frac{t}{2}\ \} \\ &= P\{W > \left(1 + \frac{1}{2}\right) \times \frac{t}{3}\} \\ &\leq e^{-\frac{\left(\frac{1}{2}\right)^2\left(\frac{t}{3}\right)}{3}} \qquad\qquad using\ Chernoff\ Bound \\ &= e^{-\frac{t}{36}} = \delta \end{aligned}$$

$\Rightarrow$

$$P\{|X_i - F_k| \geq \epsilon F_k\} \leq \delta$$

where we have taken $t = 36 \times \log(\frac{1}{\delta})$ i.e. $t = O\left(\log(\frac{1}{\delta})\right)$.

Space Bound

$$space \leq st.\left(\log(m) + \log(n)\right) = O\left(\frac{1}{\epsilon^2}.\log\left(\frac{1}{\delta}\right).kn^{\left(1-\frac{1}{k}\right)}\left(\log(m) + \log(n)\right)\right)$$

References

1. Amit Chakrabarti, CS49: Data Stream Algorithms, Lecture Notes, Fall 2011.
2. Kenneth A. Ross, Elementary Analysis: The Theory of Calculus, Springer, 2013.
3. Sheldon M. Ross, Introduction to Probability Models, Academic Press, 2010.