CMPUT 675: Algorithms for Streaming and Big Data **Fall 2019**

Lecture 6 (Sep 23, 2019): F_p Estimator and Heavy Hitters

Lecturer: Mohammad R. Salavatipour *Scribe: Haozhou Pang*

Recall that in the previous lecture, we have seen an F_2 estimator via the JL lemma that uses the 2-stable property of the normal distribution. In this lecture, we will extend this idea to construct an F_p estimator by using a p-stable distribution (for $p \in (0, 2]$).

6.1 F_p estimator

Before we introduce the F_p estimator, we need some definitions.

Definition 1 *Let $p > 0$ be a real number. A probability distribution D_p over reals is called p-stable if it has the following property: Suppose $X_1, \dots, X_n \in D_p$, for any real vector $c \in \mathbb{R}^n$, $X = \sum c_i X_i$ has the same distribution as $\bar{c}X$, where $\bar{c} = (\sum c_i^p)^{1/p} = ||c||_p$ and $X \in D_p$.*

It is known that p-stable distribution exists for all $p \in (0, 2]$, for example, the normal distribution is 2-stable and Cauchy distribution is 1-stable. Cauchy distribution is the distribution of the ratio of two standard normal distribution. It has density function $\phi(x) = \frac{1}{\sqrt{2\phi}} e^{-x^2/2}$. However, in general, for any $p > 2$, the p-stable distributions do not have an explicit formula. Also, we can use the Chambers-Mallows-Stuck method to sample from D_p for $p \in (0, 2]$. Sample (θ, r) from $[-\frac{\pi}{2}, \frac{\pi}{2}] \times [0, 1]$ and return $X = \frac{\sin(p\theta)}{(\cos\theta)^{1/p}} \left(\frac{\cos((1-p)\theta)}{\ln(1/r)}\right)^{\frac{1-p}{p}}$. Now if we replace $N(0, 1)$ in the code given for F_2 estimator with D_p where it's a p-stable distribution we can generate a variable X that is distributed according to D_p scalled by $||f||_p$ and this is what we are trying to estimate.

Definition 2 *The median of distribution D is μ if for $X \sim D$, $Pr[X \le \mu] = \frac{1}{2}$. If $\phi(x)$ is the probability density function (PDF) of D, then $\int_{-\infty}^{\mu} \phi(x)dx = \frac{1}{2}$.*

Note that the distribution D_p has a unique median and we denote it by median(D_p). For a distribution D, we let $|D|$ denote the distribution of the absolute value of a random variable drawn from D. One can think of $|D|$ as the negative part of D being folded to the positive part, so if $\phi(x)$ is the density function of D, then the density function of $|D|$ is given by $\psi(x)$, where $\psi(x) = 2\phi(x)$ if $x \ge 0$ and $\psi(x) = 0$ if $x < 0$. The factor 2 arises from the symmetry of the distribution. Then we are ready to state the F_p estimator.

F_p **Estimator**
- $t \leftarrow O(\frac{1}{\epsilon^2} \log \frac{1}{\delta})$
- $x \leftarrow \mathbf{0}$
- Let M be a $t \times n$ matrix where $M_{ij} \sim D_p$
- While there is a token (j, a_j), do
  - for $i = 0$ to t do:
    - $x[i] \leftarrow x[i] + M_{ia_j}$
- return $\frac{median(|x_1|, \dots, |x_t|)}{median(D_p)}$

6.1.1 Analysis of the F_p estimator

For any $p \in (0, 2]$ and $c \in \mathbb{R}$, we use $\phi_{p,c}$ to denote the density function of distribution of $c|X|$ where $X \sim D_p$ and let $\mu_{p,c}$ be the median of this distribution. Then it's easy to verify that $\phi_{p,c} = \frac{1}{c}\phi(\frac{X}{c})$ and $\mu_{p,c} = c \cdot \mu_{p,1}$.

Suppose X_i is the value of x_i at the end of the algorithm. By using the p-stable property we know that $X_i \sim ||f||_p X$, where $X \sim D_p$, so $\frac{|X_i|}{median(|D_p|)}$ has a distribution according to $c|D_p|$ where $c = \frac{||f||_p}{median(|D_p|)}$ and the PDF is $\phi_{p,c}$. Then the median of the distribution (which we try to estimate) is $\mu_{p,c} = c \cdot \mu_{p,1} = ||f||_p$. The algorithm takes t independent samples from the distribution and output the sample median. We use the following lemma to show the sample median gives us good concentration.

Lemma 1 *Let $\epsilon > 0$ and D be a probability distribution over $\mathbb{R}$ with density function ϕ and a unique median $\mu > 0$. Suppose ϕ is absolutely continuous on $[(1-\epsilon)\mu, (1+\epsilon)\mu]$ and let $\phi^* = min\{\phi(x) : x \in [(1-\epsilon)\mu, (1+\epsilon)\mu]\}$. Let $Y = median_{1\le i\le t}(Y_i)$ where Y_i's are independently sampled from D. Then*

$$Pr[|Y - \mu| \ge \epsilon\mu] \le 2e^{-\frac{2}{3}\epsilon^2\mu^2{\phi^*}^2 t}$$

Proof. We only give the proof to the upper bound $Pr[Y \le (1-\epsilon)\mu] \le e^{-\frac{2}{3}\epsilon^2\mu^2{\phi^*}^2 t}$ The other direction is similar and omitted here. Note that by the definition of median, $Pr[Y_i \le \mu] = \frac{1}{2}$. Let $\Phi(y) = \int_{-\infty}^{y} dx$ be the cumulative density function, then

$$\begin{aligned}
Pr[Y_i \le (1-\epsilon)\mu] &= \frac{1}{2} - \int_{(1-\epsilon)\mu}^{\mu} \phi(x)dx \\
&= \frac{1}{2} - (\Phi(\mu) - \Phi((1-\epsilon)\mu)) \\
&= \frac{1}{2} - \underbrace{\epsilon\mu\phi(\zeta)}_{\gamma} && \text{for some } \zeta \in [(1-\epsilon)\mu, \mu] \\
&\le \frac{1}{2} - \epsilon\mu\phi^* && \text{by the definition of } \phi^*
\end{aligned}$$

Let I_j be the indicator variable for the event $Y_j \le (1-\epsilon)\mu$. Then

$$E[I_j] = Pr[Y_j \le (1-\epsilon)\mu] \le \tfrac{1}{2} - \epsilon\mu\phi^*$$

Let $I = \sum_{j=1}^{t} I_j$, then $E[I] = t \cdot (\frac{1}{2} - \gamma)$. Since Y is the median of $Y_1, \ldots, Y_t$, $Y \le (1-\epsilon)\mu$ requires at least $\frac{t}{2}$ of I_j's being true, which is equivalent to $Pr[I \ge (1+\alpha)E[I]]$. If we choose $(1+\alpha) = \frac{1}{1-2\gamma}$ and apply the Chernoff bounds, then we have

$$Pr[Y \le (1-\epsilon)\mu] \le e^{-\frac{2}{3}\epsilon^2\mu^2\phi(\zeta)^2 t} \le e^{-\frac{2}{3}\epsilon^2\mu^2{\phi^*}^2 t}$$

as required. ■

It remains to apply the lemma to show the concentration of our F_p estimator. Let ϕ be the density function of the distribution of $c|D_p|$, and recall that the median of this distribution $\mu = ||f||_p$. The algorithm returns median of the t independent samples from $c|D_p|$. Therefore by applying the lemma,

$$Pr[|Y - ||f||_p| \geq \epsilon ||f||_p] \leq 2e^{-\frac{2}{3}\epsilon^2\mu^2\phi^{*2}t}$$

Observe that $\mu\phi^*$ only depends on D_p and ϵ, let $\mu\phi^* = c_{p,\epsilon}$ (some constant depending on p and ϵ), given $t = O(\frac{1}{\epsilon^2}\log\frac{1}{\delta})$, thus

$$Pr[|Y - ||f||_p| \geq \epsilon ||f||_p] \leq \delta$$

Remarks: As readers might have noticed, there are several issues that make the F_p estimator as described impractical:

- The algorithm requires space to store the entire matrix M, which is too large for a streaming model.
- The value of t depends on $c_{p,\epsilon}$, which is not explicitly known due to the lack of knowledge on D_p for $p > 2$.
- The algorithm involves calculations on reals, which is expensive and would introduce rounding errors.

To obtain an efficient streaming algorithm, we need to use pseudorandom generators to store a compressed version of M, for more details, see [I06].

6.2 Heavy Hitters

We have seen several algorithms for estimating F_p for $p \geq 0$. Recall that F_0 corresponds to the number of distinct items in the stream and we define F_∞ to be finding the largest frequency in a stream. An interesting question that one may ask is that what if we want to find the frequent items (a.k.a heavy hitters) in a stream?

The problem can be described as given a stream $\sigma = a_1, a_2, \ldots, a_m$ with frequency vector $(f_1, f_2, \ldots, f_m)$, given k, we want to find all values $\{j | f_j > \frac{m}{k}\}$. Note that the number of such items is at most k, and the Majority problem, in which we want to know is there an item that appears more than $\frac{m}{2}$ times in the stream, is a special case when $k = 2$. Misra and Gries [MG82] gave a simple algorithm to solve this problem:

Misra-Gries (82')

let A be an empty list

while stream is not empty **do**

let j be the next token
if $(j \in keys(A))$ **then**
$A[j] \leftarrow A[j] + 1$
else if $|keys(A)| < k - 1$ **then**
$A[j] \leftarrow 1$
else for each $l \in keys(A)$ **do**
$A[l] \leftarrow A[l] - 1$
remove keys with $A[l] = 0$
end while

for each $i \in keys(A)$, set $\hat{f}_i = A[i]$

for each $i \notin keys(A)$, set $\hat{f}_i = 0$

We maintain A as a balanced BST. We have at most k key/value pairs and each pair needs $O(\log n)$ bits, so the total space is $O(k(\log m + \log n))$.

The following theorem is left as an exercise.

Theorem 1 *For each $i \in [n]$: $f_i - \frac{m}{k} \leq \hat{f}_i \leq f_i$.*

The theorem implies that every item that occurs more than $\frac{m}{k}$ times in the stream is guaranteed to appear in the output list, so we can do a second pass to find exact f_i values for the at most k keys in A. The drawback of this algorithm is also obvious, it requires 2 passes on the data instead of 1, and it does not provide a sketch.

References

I06 P. Indyk, Stable distributions, pseudorandom generators, embeddings, and data stream computation. *Journal of the ACM (JACM)*, 21(10):53(3):307323, 2006.

MG82 J. Misra, D. Gries, Finding repeated elements. *Science of Computer Programming*,143-152, 1982.