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

Lecture 3 (Sep 11, 2019): BJKST Algorithm and AMS $F_k(\sigma)$ Estimator

Lecturer: Mohammad R. Salavatipour *Scribe: Brandon Fuller*

3.1 Bar-yossef, Jayram, Kumar, Sivakumar, Trevisan Algorithm

For a better estimator of $F_0(\sigma)$ (determining the number of unique elements in our stream σ) we look to an algorithm from [BJKST02] that improves upon the *Flajolet-Martin Counter*.

For the *Flajolet-Martin Counter*, we only kept the smallest element produced by our 2-universal hash family $H : [m] \to [0, 1]$ and expected that after d distinct elements were processed we would have gaps of equal spacing roughly equal to $\frac{1}{d+1}$. Using the smallest element we approximated d.

Instead of keeping track of the smallest element, we keep track of the t smallest hash values seen so far (this can be done using a heap). Using a similar analysis as above; for a 2-universal hash family $H : [m] \to [0, 1]$ we expect that the number of hashed values that are less than $\frac{t}{d}$ should be t. Thus we expect largest of these t elements to be $\frac{t}{d}$. To achieve a $(1 + \epsilon)$-approximation we must choose t large enough. Specifically we will consider $t = \frac{c}{\epsilon^2}$ for some constant c.

BJKST Algorithm

1. Choose a 2-universal hash family $H : [m] \to [M = m^3]$
2. $t \leftarrow \frac{c}{\epsilon^2}$
3. While there is another a_i from σ do:
   - Update the smallest t hash values with $h(a_i)$
4. Let v be the largest of the t smallest hash values
5. Return $\tilde{d} = \frac{tM}{v}$ (Since we expect $v \approx \frac{tM}{d}$)

Since the hash function from our 2-universal hash family is mapping m things to m^3 spots we expect that the probability that there are any collisions is at most $\frac{m^2}{m^3} = \frac{1}{m}$. Since m is large, we will assume that there are no collisions.

3.1.1 Analysis

This algorithm uses $O(\frac{1}{\epsilon^2} \log m)$ space (for storing the t smallest elements and $\log m$ bits) which can even be improved to roughly $O(\frac{1}{\epsilon^2} + \log m)$ (see [BJKST02]). We will now prove with the following two lemmas that this is a $(1 + \epsilon, 1 - \delta)$-estimator. For the following proofs, assume our distinct values are $b_1, ..., b_d$.

Lemma 1 $\Pr[\tilde{d} > (1 + \epsilon)d] \leq \frac{16}{c}$.

Proof. Suppose $\frac{tM}{v} = \tilde{d} > (1+\epsilon)d$. So $v < \frac{tM}{(1+\epsilon)d}$ which implies from the definition of v that at least t of the hash values $h(b_1), ..., h(b_d)$ are less than $\frac{tM}{(1+\epsilon)d} \leq \frac{(1-\frac{\epsilon}{2})tM}{d}$ (for small values of ϵ). For each $h(b_i)$, the probability of being smaller than $\frac{(1-\frac{\epsilon}{2})tM}{d}$ is at most $\frac{(1-\frac{\epsilon}{2})t}{d} + \frac{1}{M} < \frac{(1-\frac{\epsilon}{4})t}{d}$, where the $\frac{1}{M}$ comes from scaling.

Let X_i be the 0-1 random variable for $h(b_i) < \frac{(1-\frac{\epsilon}{2})tM}{d}$ and let $Y = \sum_{i=1}^m X_i$. From above, $\mathrm{E}[X_i] \leq \frac{(1-\frac{\epsilon}{4})t}{d}$. Thus from Lemma 1 from Lecture 2 we have that:

$Var[Y] \leq \mathrm{E}[Y] = \sum_{i=1}^m \mathrm{E}[X_i] \leq \sum_{i=1}^d \frac{(1-\frac{\epsilon}{4})t}{d} = (1-\frac{\epsilon}{4})t$

So by relating the probability of our estimate to the variable Y and using Chebyshev we get the following:

$$
\begin{aligned}
\Pr[\tilde{d} > (1+\epsilon)d] &\leq \Pr[Y > t] \\
&\leq \Pr[|Y - \mathrm{E}[Y]| > \frac{\epsilon t}{4}] \\
&\leq \frac{16Var[Y]}{e^2t^2} \\
&\leq \frac{16(1-\frac{\epsilon}{4})t}{e^2t^2} \\
&= \frac{16-4\epsilon}{c}. \\
&\leq \frac{16}{c}
\end{aligned}
$$

■

Lemma 2 $\Pr[\tilde{d} < (1-\epsilon)d] \leq \frac{1}{c}$.

Proof. We note that for a constant $\alpha > 1$, $\frac{1}{1-\epsilon} \leq (1+\alpha\epsilon)$ for a small enough ϵ.

Suppose now that $\tilde{d} = \frac{tM}{v} < (1-\epsilon)d$. Which implies $v < \frac{tM}{(1-\epsilon)d}$. Again this means that less than t of the hash values $h(b_1), ..., h(b_d)$ are smaller than $\frac{tM}{(1-\epsilon)d} \leq \frac{(1+\frac{3}{2}\epsilon)tM}{d}$. For each $h(b_i)$, the probability of being smaller than $\frac{(1+\frac{3}{2}\epsilon)tM}{d}$ is at most $\frac{(1+\frac{3}{2}\epsilon)t}{d} + \frac{1}{M} < \frac{(1+2\epsilon)t}{d}$.

Let X_i be the 0-1 random variable for $h(b_i) < \frac{(1+2\epsilon)tM}{d}$ and let $Y = \sum_{i=1}^m X_i$. From above, $\mathrm{E}[X_i] \leq \frac{(1+2\epsilon)t}{d}$. Thus from Lemma 1 from Lecture 2 we have that:

$Var[Y] \leq \mathrm{E}[Y] = \sum_{i=1}^m \mathrm{E}[X_i] \leq \sum_{i=1}^d \frac{(1+2\epsilon)t}{d} = (1+2\epsilon)t$.

Similar to the previous lemma we get the following:

$$
\begin{aligned}
\Pr[\tilde{d} < (1-\epsilon)d] &\le \Pr[Y < t] \\
&\le \Pr[|Y - \mathrm{E}[Y]| > 2\epsilon] \\
&\le \frac{Var[Y]}{4e^2t^2} \\
&\le \frac{(1+2\epsilon)t}{4e^2t^2} \\
&= \frac{1+2\epsilon}{4c} \\
&\le \frac{1}{c}
\end{aligned}
$$

■

As an example, if we choose $c = 96$ we get a $(1+\epsilon, 1-\frac{1}{3})$-estimator.

3.2 AMS $F_k(\sigma)$ Estimator

So far we have only looked at $F_0(\sigma)$ and $F_1(\sigma)$ estimators. The first general $F_k(\sigma)$ estimator we will consider is the following algorithm which was presented in [AMS99] alongside their $F_0(\sigma)$ estimator.

AMS $F_k(\sigma)$ Algorithm

1. $m \leftarrow 0$, $r \leftarrow 0$, $a \leftarrow 0$
2. While there is another item do:
   - $m \leftarrow m + 1$
   - $\beta \leftarrow$ random boolean with $\Pr[\beta = 1] = \frac{1}{m}$
   - If $\beta = 1$; $a \leftarrow a_m$, $r \leftarrow 1$
   - Else if $a_m = a$; $r \leftarrow r + 1$
3. Return $m(r^k - (r-1)^k)$

The probability of the j-th item being selected as the last token is exactly equal to $\frac{1}{j} \times \frac{j}{j+1} \times ... \times \frac{m-1}{m} = \frac{1}{m}$. Thus this algorithm will randomly select one of the m items. After the stream has been processed (with, say the J-th item being randomly selected), then $r = |\{j : a_j = a_J, J \le j \le m\}|$ (the number of items in the suffix of the stream past the J-th element that are the same as a_J).

It may not be immediately clear why we choose the specific return value. To understand this choice we will consider the analysis of this algorithm.

3.2.1 Analysis

Clearly we have that $\Pr[J = j] = \frac{f_J}{m}$. To understand the analysis of this algorithm, we will instead use the following equivalent process of selecting a:

1. Pick a random $a \in [d]$.

2. Uniformly at random select one of the occurrences of a from σ.

Let X be the random variable for the output of the algorithm and A and R the random variables for a and r respectively.

$\mathrm{E}[X] = \sum_{j\in[d]} \Pr[A=j]\mathrm{E}[X|A=j] = \sum_{j\in[d]} \frac{f_j}{m}\mathrm{E}[m(R^k-(R-1)^k)|A=j]$

Once we are given that $A=j$, we have that R is equally likely to be any of the values $\{1,...,f_j\}$. If $R=i\in\{1,...,f_j\}$ then $\Pr[R=i|A=j]=\frac{1}{f_j}$ and $X=m(i^k-(i-1)^k)$. Using this we get:

$$\begin{aligned}
\mathrm{E}[X] &= \sum_{j\in[d]} \frac{f_j}{m}\mathrm{E}[m(R^k-(R-1)^k)|A=j] \\
&= \sum_{j\in[d]} \frac{f_j}{m}\sum_{i=1}^{f_j}\frac{1}{f_j}m(i^k-(i-1)^k) \\
&= \sum_{j\in[d]}\sum_{i=1}^{f_j} i^k-(i-1)^k
\end{aligned}$$

Since this is a telescoping sum we finally get $\mathrm{E}[X]=\sum_{j\in[d]} f_j^k$ as desired. We will use similar techniques to compute the variance of X.

$$\begin{aligned}
\mathrm{Var}[X] &\leq \mathrm{E}[X^2] \\
&= \sum_{j\in[d]} \Pr[A=j]\mathrm{E}[X^2|A=j] \\
&= \sum_{j\in[d]} \frac{f_j}{m}\sum_{i=1}^{f_j}\frac{1}{f_j}(m(i^k-(i-1)^k))^2 \\
&= m\sum_{j\in[d]}\sum_{i=1}^{f_j}(i^k-(i-1)^k)^2
\end{aligned}$$

If we consider the polynomial $x^k-(x-1)^k$ we can say, using Mean Value Theorem, that $\exists g(x)\in(x-1,x)$ such that $x^k-(x-1)^k = kg(x)^{k-1}\leq kx^{k-1}$. Thus if we apply this once to the above equation we get

$$
\begin{aligned}
\mathrm{Var}[X] &\leq m \sum_{j\in[d]} \sum_{i=1}^{f_j} (i^k - (i-1)^k)^2 \\
&\leq m \sum_{j\in[d]} \sum_{i=1}^{f_j} k i^{k-1} (i^k - (i-1)^k) \\
&\leq mk \sum_{j\in[d]} f_j^{k-1} \sum_{i=1}^{f_j} (i^k - (i-1)^k) \\
&= mk \sum_{j\in[d]} f_j^{k-1} f_j^k \\
&= mk \sum_{j\in[d]} f_j^{2k-1} \\
&= kF_1 F_{2k-1}
\end{aligned}
$$

So we can see that this algorithm gives us a desirable expected value but the variance can potentially be very large. To finish this bound we will consider the following Lemma.

Lemma 3 $F_1 F_{2k-1} \leq n^{1-\frac{1}{k}} (F_k)^2$.

The proof will be presented in the next lecture.

References

AMS99 N. Alon, Y. Matias, and M. Szegedy, The Space Complexity of Approximating the Frequency Moments. *J. Comput. Syst. Sci.*, 31(2):137-147, 1999.

BJKST02 Z. Bar-Yossef, T. S. Jayram, R. Kumar, D. Sivakumar, L. Trevisan, Counting Distinct Elements in a Data Stream. *Proceedings of the 6th International Workshop on Randomization and Approximation Techniques*, p.1-10, 2002.