6.897: CSCI8980 Algorithmic Techniques for Big Data	September 19, 2013
Lecture 4	
Dr. Barna Saha	*Scribe: Barna Saha*

Overview

In the previous lecture, we saw that using count-min sketch we can solve a variety of problems related to frequency estimates such as point query, range query, heavy-hitters etc. where the error in the estimate is in terms of the l_1 norm of the stream. Can we obtain frequency estimates where error will be in terms of l_2 norm of the stream ? We will see one such algorithm Count-Sketch as part of an exercise. Today, we consider the following problem given a stream S of size m where elements are coming from domain $[1, n]$ and have unknown frequencies $f_1, f_2, ..., f_n$, what is the second frequency moment F_2 of the stream ? Here F_2 is defined as $F_2 = \sum_{i=1}^{n} f_i^2$. We give an elegant solution based on sketches from [1] that requires logarithmic space and update time.

1 Estimating F_2

Let $\mathcal{H} = \{h : [n] \to \{+1, -1\}\}$ be a family of four-wise independent hash functions (we have seen previously how to construct such families). We initialize t counters $Z_1, Z_2, ...Z_t$ to 0 and maintain $Z_j = Z_j + ah_j(i)$ on arrival of (i, a) for $j = 1, ..., t = \frac{c}{\epsilon^2}$, where c is some constant to be fixed later. We return $Y = \frac{1}{t}\sum_{j=1}^{t} Z_j^2$ as the estimate of F_2 of the stream.

We first show that Y is an unbiased estimator of F_2, that is $\mathsf{E}\big[Y\big] = F_2(S)$. Then we compute $\mathsf{Var}\big[Y\big]$ and apply Chebyshev inequality to bound the deviation of Y from its expectation that is F_2.

Lemma 1. $\mathsf{E}\big[Y\big] = F_2$

Proof.

$$\mathsf{E}\big[Y\big] = \mathsf{E}\big[\frac{1}{t}\sum_{j=1}^{t} Z_j^2\big] = \frac{1}{t}\sum_{i=1}^{t} \mathsf{E}\big[Z_j^2\big].$$

Now, $Z_j = \sum_i h_j(i) f_i$. Hence

$$\begin{aligned}
\mathsf{E}\big[Z_j^2\big] &= \mathsf{E}\big[(\sum_i h_j(i) f_i)^2\big] \\
&= \sum_i \mathsf{E}\big[(h_j(i))^2\big] f_i^2 + 2\sum_{i<k} \mathsf{E}\big[h_j(i)\big]\mathsf{E}\big[h_j(k)\big] f_i f_k \\
&\quad \text{by linearity of expectation and independence of } h_j(i) \text{ and } h_j(k) \\
&= \sum_i f_i^2 \text{ since } \mathsf{E}\big[h_j(i)\big] = 0 \text{ for all } i \text{ and } \mathsf{E}\big[(h_j(i))^2\big] = 1 \text{ for all } i \\
&= F_2
\end{aligned}$$

Therefore,

$$\mathsf{E}\big[Y\big] = \frac{1}{t}\sum_{i=1}^{t}\mathsf{E}\big[Z_j^2\big] = F_2.$$

□

Lemma 2. $\mathsf{Var}\big[Y\big] \leq \frac{4F_2^2}{t}$

Proof.

$$\mathsf{Var}\big[Y\big] = \mathsf{Var}\big[\frac{1}{t}\sum_{j=1}^{t} Z_j^2\big] = \frac{1}{t^2}\sum_{i=1}^{t}\mathsf{Var}\big[Z_j^2\big].$$

In the above we obtained the second inequality by noting that Z_j^2 random variables are all pair-wise independent. We now calculate $\mathsf{Var}\big[Z_j^2\big]$ which is $\mathsf{E}\big[Z_j^4\big] - (\mathsf{E}\big[Z_j^2\big])^2$. Note that

$$\mathsf{E}\big[Z_j^4\big] = \mathsf{E}\big[(\sum_i h_j(i)f_i)^4\big] = \sum_{a\leq b\leq c\leq d} f_a f_b f_c f_d \mathsf{E}\big[f_a f_b f_c f_d\big].$$

Now note that if either of the following conditions hold $a < b < c < d$ or exactly three among a, b, c, d are equal, then those terms contribute 0. Hence

$$\begin{aligned}\mathsf{E}\big[Z_j^4\big] = \mathsf{E}\big[(\sum_i h_j(i)f_i)^4\big] &= \sum_i \mathsf{E}\big[(h_j(i))^4\big]f_i^4 + \binom{4}{2}\sum_{i<k}\mathsf{E}\big[(h_j(i))^2\big]\mathsf{E}\big[(h_j(k))^2\big]f_i^2 f_k^2 \\ &= \sum_i \mathsf{E}\big[(h_j(i))^4\big]f_i^4 + 6\sum_{i<k} f_i^2 f_k^2\end{aligned}$$

On the otherhand,

$$(\mathsf{E}\big[Z_j^2\big])^2 = \sum_i \mathsf{E}\big[(h_j(i))^4\big]f_i^4 + 2\sum_{i<k} f_i^2 f_k^2$$

Hence

$$\mathsf{Var}\big[Z_j^2\big] = 4\sum_{i<k} f_i^2 f_k^2 \leq 4\max_i f_i^2 \sum_i f_i^2 \leq 4F_2^2$$

Therefore,

$$\mathsf{Var}\big[Y\big] = \frac{1}{t^2}\sum_{i=1}^{t}\mathsf{Var}\big[Z_j^2\big] \leq \frac{4F_2}{t}.$$

□

Lemma 3. $\mathsf{Pr}\big[|Y - \mathsf{E}\big[Y\big]| > \epsilon F_2\big] \leq \frac{1}{3}$ *where* $t \geq \frac{12}{\epsilon^2}$.

Proof. By Chebyshev Inequality

$$\mathsf{Pr}\big[|Y - \mathsf{E}\big[Y\big]| > \epsilon F_2\big] \leq \frac{\mathsf{Var}\big[Y\big]}{\epsilon^2 F_2^2} \leq \frac{4F_2^2}{t\epsilon^2 F_2^2} \leq \frac{1}{3}$$

□

So, we have an estimate Y for F_2 which guarantees an absolute error at most ϵF_2 with probability at least $\frac{2}{3}$? Can we boost this probability to $(1 - \delta)$ for any $\delta > 0$? To do so, we apply a generic technique, *boosting by median.*

1.1 Boosting by Median

We keep $s = O(\log 1\delta)$ independent estimates Y_1, Y_2, ..., Y_s. We then arrange these values in non-increasing order and return the $\lceil s/2 \rceil$-th estimate, that is the median of $Y_1, Y_2, ..., Y_s$. Let without loss of generality assume, $Y_1 \le Y_2 \le ... \le Y_s$. And for simplicity assume, s is even. First consider the upper tail (the lower tail is similar). If $Y_{s/2} > (1+\epsilon)F_2$ then all of $Y_{s/2+1}, Y_{s/2+2}, ..., Y_s$ must be higher than $F_2(1+\epsilon)$.

Define an indicator random variable X_i, which is 1 if $Y_i > (1+\epsilon)F_2$ and 0 otherwise. From Lemma 3 $\mathsf{Pr}\big[X_i = 1\big] \le \frac{1}{3}$. Hence if we denote by X, the number of estimates that return value more than $(1+\epsilon)F_2$, then $X = \sum_{i=1}^{s} X_i$ and $\mathsf{E}\big[X\big] \le \frac{s}{3}$.

We now apply the Chernoff's bound to obtain

$$\mathsf{Pr}\big[Y_{s/2} > (1+\epsilon)F_2\big] = \mathsf{Pr}\big[X > \frac{s}{2}\big] = \mathsf{Pr}\big[X > \mathsf{E}\big[X\big](1+\frac{1}{3})\big] \le e^{-\frac{s}{3}\frac{1}{9}\frac{1}{3}}$$

Setting $s = C\ln\frac{1}{\delta}$ where C is a large enough constant, the above probability becomes less than $\delta/2$.

Similarly, we have

$$\mathsf{Pr}\big[Y_{s/2} < (1-\epsilon)F_2\big] < \frac{\delta}{2}.$$

Therefore, by union bound

$$\mathsf{Pr}\big[|Y_{s/2} - F_2| > \epsilon F_2\big] < \delta.$$

Finally, we have the following theorem

Theorem 4. *There is a randomized algorithm for estimating F_2 within error $(1 \pm \epsilon)$ with probability at least $(1-\delta)$ that takes space $O(\frac{1}{\epsilon^2}\log\frac{1}{\delta})$ and update time $O(\frac{1}{\epsilon^2}\log\frac{1}{\delta})$.*

References

[1] Noga Alon, Yossi Matias, Mario Szegedy: The Space Complexity of Approximating the Frequency Moments. J. Comput. Syst. Sci. 58(1): 137-147 (1999)