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 $f_1, f_2, \ldots, f_n$, 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 $H = \{ h : [n] \to \{+1, -1\} \}$ be a family of four-wise independent hash functions (we have seen previously how to construct such families). We initialize $t$ counters $Z_1, Z_2, \ldots, Z_t$ to 0 and maintain $Z_j = Z_j + ah_j(i)$ on arrival of $(i, a)$ for $j = 1, \ldots, t = \frac{c}{\epsilon^2}$, 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 $E[Y] = F_2(S)$. Then we compute Var[$Y$] and apply Chebyshev inequality to bound the deviation of $Y$ from its expectation that is $F_2$.

**Lemma 1.** $E[Y] = F_2$

**Proof.**

\[
E[Y] = E\left[ \frac{1}{t} \sum_{j=1}^{t} Z_j^2 \right] = \frac{1}{t} \sum_{i=1}^{t} E[Z_j^2].
\]

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

\[
E[Z_j^2] = E[(\sum_i h_j(i)f_i)^2] = \sum_i E[(h_j(i))^2]f_i^2 + 2\sum_{i<k} E[h_j(i)]E[h_j(k)]f_if_k.
\]

by linearity of expectation and independence of $h_j(i)$ and $h_j(k)$

\[
= \sum_i f_i^2 \text{ since } E[h_j(i)] = 0 \text{ for all } i \text{ and } E[(h_j(i))^2] = 1 \text{ for all } i
\]

\[
= F_2
\]
Therefore,
\[ E[Y] = \frac{1}{t} \sum_{i=1}^{t} E[Z_j^2] = F_2. \]

**Lemma 2.** \( \text{Var}[Y] \leq \frac{4F_2^2}{t} \)

*Proof.*

\[ \text{Var}[Y] = \text{Var}\left[\frac{1}{t} \sum_{j=1}^{t} Z_j^2\right] = \frac{1}{t^2} \sum_{i=1}^{t} \text{Var}[Z_j^2]. \]

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

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

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

\[ E[Z_j^4] = \sum_i E[(h_j(i))^4] f_i^4 + \left(\frac{4}{2}\right) \sum_{i<k} E[(h_j(i))^2] E[(h_j(k))^2] f_i^2 f_k^2 \]

\[ = \sum_i E[(h_j(i))^4] f_i^4 + 6 \sum_{i<k} f_i^2 f_k^2 \]

On the other hand,

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

Hence

\[ \text{Var}[Z_j^2] = 4 \sum_{i<k} f_i^2 f_k^2 \leq 4 \max_i f_i^2 \sum_{i<k} f_k^2 \leq 4F_2^2. \]

Therefore,

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

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

*Proof.* By Chebyshev Inequality

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

So, we have an estimate \( Y \) for \( F_2 \) which guarantees an absolute error at most \( \epsilon F_2 \) 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*. 

2
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, \( Y_1 \leq Y_2 \leq ... \leq Y_s \). 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 \( \Pr[X_i = 1] \leq 1/3 \). 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 \( E[X] \leq s/3 \).

We now apply the Chernoff’s bound to obtain

\[
\Pr[Y_{s/2} > (1 + \epsilon)F_2] = \Pr[X > \frac{s}{2}] = \Pr[X > E[X] \left(1 + \frac{1}{3}\right)] \leq e^{-\frac{s}{3} \frac{1}{9}}
\]

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

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

Therefore, by union bound

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

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 \( O\left(\frac{1}{\epsilon^2} \log \frac{1}{\delta}\right) \) and update time \( O\left(\frac{1}{\epsilon^2} \log \frac{1}{\delta}\right) \).

**References**