CSE 522: Sublinear (and Streaming) Algorithms **Spring 2014**

Lecture 6: Frequency Moment Approximation

April 16, 2014

Lecturer: Paul Beame *Scribe: Paul Beame*

1 Approximating F_2

We already defined $F_0 = \sum_{j=1}^{M} f_j^0$, the number of distinct elements in the stream. More generally we can define the k-th frequency moment

$$F_k = \sum_{j=1}^{M} f_j^k = ||f||_k^k.$$

$F_1 = ||f||_1 = n$ is trivial to compute exactly. F_k is an integral value for integer k, which is convenient, but there is no particular reason to restrict this to integer values; we sometimes write F_p for arbitrary non-negative p.

F_2 is particularly interesting and useful to approximate. One reason that we might be interested is in understanding the quality of the output of the COUNT Sketch approximation, where $|\tilde{f}_j - f_j| \leq \varepsilon||f_{-j}||_2 \leq \varepsilon\sqrt{F_2}$. Since F_1 is easy to compute exactly, the errors in the COUNT-MIN and Misra-Gries estimates are easy to compute. The F_2 approximation lets us get good estimates of the (smaller) error for the COUNT Sketch also.

However, there is a more important reason to consider F_2. If we have some relation R with and f is the frequency vector for an attribute in R, then $F_2 = \sum_{j=1}^{M} f_j^2$ is precisely the size of $R \bowtie R$, the join of R with itself on that attribute, which has an entry for each pair of entries in R. The methods we describe here are also useful for estimating the join size $|R \bowtie S|$ for two different relations R and S, as you will show on the first problem set.

The algorithm and ideas we present here is the most remarkable algorithm from the remarkable paper by Alon, Matias, and Szegedy [**?**], which gave algorithms and lower bounds for estimating all frequency moments. This was generalized in work by Alon, Giobbons, Matias, and Szegedy [**?**]. The basic idea is now known as the "tug-of-war" and shows the value of a higher level of independence.

The Basic Tug-of-War

1: Initialize:
2: Choose $h : [M] \to \{-1, 1\}$ independently from a 4-universal family of hash functions
3: $y \leftarrow 0$
4: Process:
5: **for** each i **do**
6: $\quad y \leftarrow y + c_i \cdot h(x_i)$
7: **end for**
8: Output: y^2

The total space of the sketch is $O(\log M + \log n)$.

Analysis Let Y be the value of y at the end of the execution and let $X = Y^2$ be its output. Then

$$Y = \sum_{i=1}^{n} c_i \; h(x_i) = \sum_{j=1}^{M} f_j \; h(j).$$

Therefore

$$\begin{aligned}
\mathbb{E}(Y^2) &= \mathbb{E}((\sum_{j=1}^{M} f_j \; h(j))^2) \\
&= \mathbb{E}(\sum_{i=1}^{M} \sum_{j=1}^{M} f_i \; h(i) \; f_j \; h(j)) \\
&= \mathbb{E}(\sum_{j=1}^{M} f_j^2 \, h(j)^2 + \sum_{i \neq j} f_i \; h(i) \; f_j \; h(j)) \\
&= \sum_{j=1}^{M} f_j^2 + \sum_{i \neq j} f_i \; f_j \; \mathbb{E}(h(i)h(j))
\end{aligned}$$

Now for $i \neq j$, $\mathbb{E}(h(i)h(j)) = \mathbb{E}(h(i))\mathbb{E}(h(j)) = 0 \cdot 0 = 0$ by pairwise independence of h. Therefore

$$\mathbb{E}(Y^2) = \sum_{j=1}^{M} f_j \; h(j))^2) = F_2$$

and hence $X = Y^2$ is an unbiased estimator of F_2. Examining its variance we have

$$Var(X) = Var(Y^2) = \mathbb{E}(Y^4) - \mathbb{E}(Y^2)^2 = \mathbb{E}(Y^4) - F_2^2.$$

Calculating, we have

$$\begin{aligned}\mathbb{E}(Y^4) &= \mathbb{E}((\sum_{j=1}^{M} f_j\ h(j))^4) \\ &= \mathbb{E}(\sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{k=1}^{M}\sum_{\ell=1}^{M} f_i\ f_j\ f_k\ f_\ell\ h(i)h(j)h(k)h(\ell)) \\ &= \sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{k=1}^{M}\sum_{\ell=1}^{M} f_i\ f_j\ f_k\ f_\ell\ \mathbb{E}(h(i)h(j)h(k)h(\ell))\end{aligned}$$

Since h is 4-wise independent, if any of i, j, k, ℓ occurs only once among the 4 values then we have $\mathbb{E}(h(i)h(j)h(k)h(\ell)) = 0$ since the singleton term factors out and has expectation 0. Therefore the only terms that contribute are when all 4 are equal and when they form 2 pair. There are 3 ways that they can form 2 pair, depending on which of j, k, ℓ the value i is paired with. Moreover, in these cases $\mathbb{E}(h(i)h(j)h(k)h(\ell)) = 1$. Therefore

$$\begin{aligned}\mathbb{E}(Y^4) &= \sum_{j=1}^{M} f_j^4 + 3 \cdot \sum_{i \neq j} f_i^2\ f_j^2 \\ &= \sum_{j=1}^{M} f_j^4 + 3(\sum_{i=1}^{M}\sum_{j=1}^{M} f_i^2\ f_j^2 - \sum_{j=1}^{M} f_j^4) \\ &= F_4 + 3(F_2^2 - F_4) \\ &= 3F_2^2 - 2F_4.\end{aligned}$$

Plugging this in we have

$$Var(X) = 2F_2^2 - 2F_4 \leq 2F_2^2.$$

Unlike the situation with previous approximations, at this point a direct application of Chebyshev's inequality does not give a good error bound for the estimate. Indeed it only yields

$$\mathbb{P}[\ |X - F_2| \geq c|F_2|\ ] \leq \frac{Var(X)}{c^2\mathbb{E}(X)^2} \leq \frac{2}{c^2}$$

which is $\geq 1/2$ unless $c > 2$. In particular it would give no lower bound at all on F_2.

Median of Means Method To remedy this problem Alon, Matias, and Szegedy described the median of means method. The basic idea is to first average the runs of some number of (pairwise) independent copies of the estimate in order to reduce the variance enough to get a failure probability of at most 1/3 for a $(1 \pm \varepsilon)$ error estimate and then apply the median trick to the result. The result is captured in the following lemma.

Lemma 1.1. *There is a constant $c < 25$ such that the following holds. Let X be an unbiased estimator of a real-valued quantity Q. Let $X_{i,j}$ for $i \in [t]$ and $j \in [k]$ be each distributed as X such that the elements of $X_i = (X_{i,1}, \ldots, X_{i,k})$ are pairwise independent for each i and the $X_1, \ldots, X_t$ are fully independent, where*

$$t = \lceil c \, \log_2(1/\delta) \rceil \text{ and } s = \lceil \frac{3 \, Var(X)}{\varepsilon^2 \mathbb{E}(X)^2} \rceil.$$

If $Z = \text{Median}\{\frac{1}{k} \sum_{j=1}^{k} X_{i,j} \mid i \in [t]\}$ then $\mathbb{P}[|Z - Q| \geq \varepsilon Q] \leq \delta$.

(Note that though we state this in mixed form with pairwise independence as well as full independence, it is simplest to apply the lemma with tk fully independent copies of X.)

Proof. For each $i \in [t]$ let

$$Y_i = \frac{1}{k} \sum_{j=1}^{k} X_{i,j}.$$

Since each $X_{i,j}$ is an unbiased estimator of Q, we have

$$\mathbb{E}(Y_i) = Q.$$

Since $k \, Y_i = \sum_{j=1}^{k} X_{i,j}$ and the $X_{i,j}$ are pairwise independent we have

$$Var(k \, Y_i) = \sum_{j=1}^{k} Var(X_{i,j}) = k \, Var(X).$$

But $Var(k \, Y_i) = k^2 \, Var(Y_i)$ so $Var(Y_i) = Var(X)/k$. Now by Chebyshev's inequality we have

$$\begin{aligned} \mathbb{P}[|Y_i - Q| \geq \varepsilon Q] &= \mathbb{P}[|Y_i - \mathbb{E}(Y_i)| \geq \varepsilon \, \mathbb{E}(Y_i)] \\ &\leq \frac{Var(Y_i)}{\varepsilon^2 \, \mathbb{E}(Y_i)^2} \\ &= \frac{Var(X)}{\varepsilon^2 k \, \mathbb{E}(Y_i)^2} \\ &\leq 1/3 \end{aligned}$$

by the choice of k. We can now apply the Median trick. Since $1/2 = (3/2) \cdot (1/3)$, we can use the Chernoff bound with $\delta = 1/2$ to get a failure probability $\leq e^{-\left(\frac{1}{2}\right)^2 \left(\frac{t}{3}\right)/3} = e^{-t/36} \leq \delta$ for $t = 25 \log_2(1/\delta)$. □

In the case of the basic tug of war we have $X = Y^2$, $\mathbb{E}(X) = Q = F_2$, and $Var(X) \leq 2F_2^2 = 2\mathbb{E}(X)^2$, so we choose

$$k = \lceil \frac{3 \, Var(X)}{\varepsilon^2 \, \mathbb{E}(X)^2} \rceil = \lceil 6/\varepsilon^2 \rceil.$$

The Tug-of-War Sketch

Initialize:
$k \leftarrow \lceil 6/\varepsilon^2 \rceil$
$t \leftarrow \lceil c \log_2(1/\delta) \rceil$
Choose tk hash functions $h_{s,j} : [M] \to \{-1, 1\}$ independently from a 4-universal family of hash functions for $s \in [t]$, and $j \in [k]$
$y \leftarrow$ a $t \times k$ array of integers initially 0.
Process:
for each i **do**
    for $s = 1$ to t **do**
        for $j = 1$ to k **do**
            $y[s, j] \leftarrow y[s, j] + c_i \cdot h_{s,j}(x_i)$
        end for
    end for
end for
Output: Median$\{\frac{1}{k} \sum_{j=1}^{k} y^2[s, j] \mid s \in [t]\}$

By the lemma this produces a $1 \pm \varepsilon$ approximation to F_2 with probability at least $1 - \delta$. Its total space is $O(\frac{1}{\varepsilon^2} \log(1/\delta)(\log M + \log n))$. Because of the linearity of the calculation of y, it is not hard to see that we can also use it to estimate the ℓ_2 (Euclidean) difference between the vectors described a pair of interleaved streams; we simply flip the c_i for one of the two streams.

A Geometric Interpretation of the Tug-of-War Consider the part of the sketch, before the application of the median trick, when we have averaged $k = \lceil 6/\varepsilon^2 \rceil$ (pairwise) independent copies of the basic sketch. Before averaging the sketch corresponds to a matrix-vector product given by a $k \times M$ matrix $B = B_{h_1,\ldots,h_k}$ consisting of ± 1 values where the (i, j)-th entry is $h_i(j)$,

$$\begin{bmatrix} -1 & +1 & -1 & -1 & +1 & +1 & -1 & -1 & \cdots & \cdots & -1 & +1 & +1 & -1 \\ +1 & +1 & +1 & -1 & -1 & -1 & -1 & +1 & \cdots & \cdots & +1 & -1 & +1 & -1 \\ +1 & -1 & -1 & +1 & +1 & -1 & +1 & -1 & \cdots & \cdots & +1 & -1 & -1 & +1 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdots & \cdots & \cdot & \cdot & \cdot & \cdot \\ -1 & -1 & +1 & +1 & -1 & +1 & +1 & +1 & \cdots & \cdots & -1 & +1 & -1 & +1 \end{bmatrix},$$

whose columns are 4-wise independent and rows are pairwise independent, The estimate produced by the average for this part of the algorithm is $\frac{1}{k} \sum_{i=1}^{s} y_i^2$ where $y = B \cdot f$. The error estimate given by the variance reduction is

$$\mathbb{P}[|\frac{1}{k} \sum_{i=1}^{k} y_i^2 - F_2| > \varepsilon F_2] \leq \frac{1}{3}.$$

Expanding this in terms of the ℓ_2 norm this says that with probability at least $2/3$,

$$(1 - \varepsilon)||f||_2^2 \leq \frac{1}{k} \sum_{i=1}^{k} y_i^2 = \frac{1}{k}||y||_2^2 = \frac{1}{k}||B \cdot f||_2^2 \leq (1 + \varepsilon)||f||_2^2.$$

Taking square roots, with probabilty at least $2/3$, we have

$$\sqrt{1-\varepsilon}||f||_2 \leq \frac{1}{\sqrt{k}}||B \cdot f||_2 \leq \sqrt{1+\varepsilon}||f||_2.$$

Now $\sqrt{1-\varepsilon}$ is larger than $1-\varepsilon$ and roughly $1-\varepsilon/2$; similarly $\sqrt{1+\varepsilon}$ is smaller than $1+\varepsilon$ and roughly $1-\varepsilon/2$. It follows that with probability at least $2/3$, the matrix $B/\sqrt{k}$ maps M dimensions to only a constant number of dimensions k and approximately preserves the 2-norm, which is a convenient way to think about what is going in this part of the sketch. We will see other methods with stronger guarantees.