Estimating Frequency Moments

- We present a better estimator for $F_0(\sigma)$ (# of distinct elements in $\sigma = a_1 \dots a_m$ where $a_i \in [n]$)

- Bar-Yossef et al (BJKST'02) presented three alg's for estimating frequencies, best of which uses $O(\varepsilon^{2} + \log n)$ space.

- Here we see the simpler one using $O(\frac{1}{\varepsilon^2}\log n)$ space & $O(\log n/\varepsilon)$ update per element.

- The idea builds uppon Flajolet-Martin: there we only kept the smallest element produced by our 2-universal hash family $H: [m] \longrightarrow [0,1]$ and we expected that after d distinct elements were seen we would have gaps of roughly size $\frac{1}{d+1}$; using the smallest element we approximate d.

- here instead of tracking the smallest element, we track the $\underline{t}$ smallest elements (this can be done using a heap)

Choose $t = \frac{c}{\varepsilon^2}$ for some fixed c (TBD).

We use a 2-universal hash family $H: [m] \longrightarrow [0,1]$

we expect the # of hash values that are less than $\frac{t}{d}$ should be t; thus we expect the largest of these values be about $\frac{t}{d}$. To get $(1+\varepsilon)$-approx, we choose t charge enough.

BJKST F_0 – estimation

Choose a 2-universal hash family $H: [n] \longrightarrow [N = n^3]$

$t \leftarrow c/\varepsilon^2$

while there is another item a_i in stream σ do

  update the smallest t hash values with $h(a_i)$

let v be the largest of t smallest hash values

return $\tilde{d} = \frac{tN}{v}$ (Since we expect $v \approx \frac{tN}{d}$)

Analysis:

Since we have t hash values, space taken is $O(\frac{1}{\varepsilon^2} \log n)$ and each round of alg. takes $O(\log t)$ time.

Suppose we have d distinct elements, say $b_1 \cdots b_d$.

Then the hash values $h(b_1) \cdots h(b_d)$ are uniformly dist.

So we expect about $\frac{t}{d}$ in $[0, \frac{tN}{d}]$.

Note that since we have a 2-universal hash family from $[n]$ to $[n^3]$, the prob. of collision is $\ll \frac{1}{n}$.

Lemma 1: $\Pr[\tilde{d} > (1+\varepsilon) d] \leq \frac{1}{6}$

Proof: So at least t values in $h(b_1) \cdots h(b_d)$ are smaller than $\frac{tN}{(1+\varepsilon)d} \leq \frac{(1-\frac{\varepsilon}{2})tN}{d}$. Each $h(b_i)$ has probability at most

$\frac{(1-\frac{\varepsilon}{2})t}{d} + \frac{1}{n} < \frac{(1-\frac{\varepsilon}{4})t}{d}$ of being smaller than $\frac{(1-\frac{\varepsilon}{2})tN}{d}$. Thus if

$\frac{(1-\frac{\varepsilon}{2})t}{d} + \frac{1}{N} < \frac{(1-\frac{\varepsilon}{4})t}{d}$ of being smaller than $\frac{(1-\frac{\varepsilon}{2})tN}{d}$. Thus if X_i is indicator for the event "$h(b_i) < \frac{(1-\frac{\varepsilon}{2})tN}{d}$" and $Y = \sum_i X_i$ then since X_i's are pairwise independent and we have t of them: $E[X_i] < \frac{(1-\frac{\varepsilon}{2})t}{d} + \frac{1}{N} \le \frac{(1-\frac{\varepsilon}{4})t}{d}$ and $E[Y] \le (1-\frac{\varepsilon}{4})t$. Using pair-wise ind, $Var[Y] \le (1-\frac{\varepsilon}{4})t$.

The event that $\tilde{d} > (1+\varepsilon)d$ occurs only if $Y > t$.

So using chebychev's:

$$\Pr[Y>t] \le \Pr\left[|Y-E[Y]| > \frac{\varepsilon t}{4}\right] \le 16\, Var[Y]/(\varepsilon^2 t^2) \le \frac{16}{\varepsilon^2 t} \le \frac{1}{6}$$

if $t \ge 96/\varepsilon^2$. ∎

Lemma 2: $\Pr[\tilde{d} < (1-\varepsilon)d] \le \frac{1}{6}$

proof: Note that $h(b_1) \dots h(b_d)$ are uniformly dist. in $[1 \dots \underset{N}{n^3}]$. So $\tilde{d} < (1-\varepsilon)$ means $v > \frac{tN}{(1-\varepsilon)d}$ which implies less than t values of $h(\cdot)$ are smaller than $\frac{tN}{(1-\varepsilon)d} \le \frac{(1+\varepsilon)tN}{d}$.

let X_i be the rand. var. indicating "$h(b_i) \le \frac{(1+\varepsilon)tN}{d}$" and $Y = \sum_{i=1}^{d} X_i$. Then since $h(b_i)$ is distributed uniformly in $[N]$, taking into accont rounding error, again we can see $\frac{(1+\frac{\varepsilon}{2})t}{d} \le E[X_i] \le \frac{(1+\frac{3\varepsilon}{2})t}{d}$

and therefore $E[Y] \ge (1+\frac{\varepsilon}{2})t$ and $Var[Y] \le E[Y] \le (1+\frac{3\varepsilon}{2})t$

Now $\tilde{d} < (1-\varepsilon)d$ only if $Y < t$, and so

$$\Pr[Y<t] \le \Pr\left[|Y - E[Y]| \ge \frac{\varepsilon t}{2}\right] \le 4\,Var[Y]/(\varepsilon^2 t^2) \le 12/(\varepsilon^2 t) < \frac{1}{6}$$

$\blacksquare$

So with Prob. $\ge \frac{1}{3}$, $\tilde{d}$ is $(1 \pm \varepsilon)d$.

We can amplify this using the median of means trick to $1-\delta$ using $O(\log \frac{1}{\delta})$ copies of the algorithm

This uses $O(\frac{1}{\varepsilon^2} \log \frac{1}{\delta} \log n)$ space and $O(\log \frac{1}{\varepsilon} \cdot \log n \cdot \log \frac{1}{\delta})$ time per element.

AMS algorithm for F_2 Estimation (linear Sketching)

- Alon/Matias/Szegedy '99 presented a generic algorithm for estimating F_k (k'th moment of frequencies) that has space usage $O(\frac{1}{\varepsilon^2} \cdot \log \frac{1}{\delta} \cdot n^{1-\frac{1}{k}} (\log m + \log n))$ and is (ε, δ)-approximation.
- They also provided a simple and efficiend F_2 estimation
- Estimating $\|x\|_2$ of a data vector x has lots of applications

AMS F_2-estimator

let h be chosen from a family H of 4-universal hash func. from $[n]$ to $\{-1,1\}$

$z \leftarrow 0$

while stream σ is non-empty do

let a_j be the next token

$z \leftarrow z + h(a_j)$

return z^2

one can think of $Y_1 \cdots Y_n$ be $\{+1,-1\}$ 4-wise ind. rand. var's and each round we set $z \leftarrow z + Y_{a_j}$

let Z be the value of z at the end.

Z is pulled in the positive direction by those tokens i with $h(i)=1$ and in the negative direction by those with $h(i)=-1$. So this alg is also called Tug-of-War.

if we simply let $h(i)=Y_i$ for each $i \in [n]$ then $Z = \sum_i f_i \cdot Y_i$. Note $Y_i^2 = 1$ and $E[Y_i] = 0$ and $E[Y_i Y_{i'}] = 0$ for $i \neq i'$ since we have pair-wise ind.

Thus:
$$E[Z^2] = \sum_{i,i' \in [n]} f_i \cdot f_{i'} E[Y_i Y_{i'}]$$
$$= \sum_{i \in [n]} f_i^2 E[Y_i^2] + \sum_{i \neq i'} f_i \cdot f_{i'} E[Y_i \cdot Y_{i'}]$$
$$= \sum f_i^2 = F$$

$$= \sum_{i\in[n]} f_i^2 = F_2$$

So Z^2 is an unbiased estimator for F_2.

We can compute $Var[Z^2] = E[Z^4] - E[Z^2]^2 = E[Z^4] - F_2^2$

$$E[Z^4] = \sum_{i\in[n]} \sum_{j\in[n]} \sum_{k\in[n]} \sum_{l\in[n]} f_i \cdot f_j \cdot f_k \cdot f_l \, E[Y_i \cdot Y_j \cdot Y_k \cdot Y_l]$$

if one of the incides appears exactly once, say $i \notin \{j,k,l\}$

then by 4-wise independence $E[Y_i \cdot Y_j \cdot Y_k \cdot Y_l] = \underbrace{E[Y_i]}_{=0} E[Y_j Y_k Y_l]$

So the only potential non-zero case is when all 4 indices are different or two pairs.

$$E[Z^4] = \sum_{i\in[n]} f_i^4 E[Y_i^4] + 6 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} f_i^2 \cdot f_j^2 E[Y_i^2] E[Y_j^2]$$

$$= F_4 + 6 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} f_i^2 \cdot f_j^2$$

Thus $Var[Z^4] = E[Z^4] - F_2^2$

$$= F_4 - F_2^2 + 6 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} f_i^2 f_j^2$$

$$= F_4 - F_2^2 + 3\left(\left(\sum_{i=1}^{n} f_i^2\right)^2 - \sum_{i=1}^{n} f_i^4\right)$$

$$= F_4 - F_2^2 + 3(F_2^2 - F_4)$$

$$\leq 2F_2^2$$

So $E[Z^2] = F_2$ and $Var[Z^2] \leq 2F_2^2 \leq 2E[Z^2]$

Now using the (standard) method of median of

means using $O(1/\varepsilon^2)$ estimators and applying

Now using the (standard) method of median of means using $O(1/\varepsilon^2)$ estimators and applying Chebychev we obtain $(\varepsilon, \frac{1}{2})$-estimator and then median trick with $O(\log \frac{1}{\delta})$ ind. mean estimators we obtain an (ε, δ)-estimator using $O(\frac{1}{\varepsilon^2} \log \frac{1}{\delta})$ parallel copies using $O(\frac{1}{\varepsilon^2} \log \frac{1}{\delta} \log n)$ space.

Linear sketching:

suppose we have two streams σ_1 & σ_2 and an alg. that computes two data structures $Z(\sigma_1)$ & $Z(\sigma_2)$. We call these structures "an sketch" if there is a space efficient algorithm A s.t. for any two streams σ_1 & σ_2, if $\sigma_1 \bullet \sigma_2$ is the concatination of the two then $A(Z(\sigma_1), Z(\sigma_2)) = Z(\sigma_1 \bullet \sigma_2)$

So the algorithm can uses the two sketches to obtain an sketch for the combined stream.

Suppose the values of a stream $\sigma_1 \sigma_2 \dots \sigma_m$ are from $[n]$ and suppose we start with an n-dim vector $x = (0, 0, \dots, 0)$ and each time we see a new token we update x (think of freq. vector f_i)

So each token $a_j = (i_j, \Delta_j)$ where $i_j \in [n]$ and Δ_j is a number (positive or negative) corresponds to an update $x_{i_j} \leftarrow x_{i_j} + \Delta_j$ (typically $\Delta_j = 1$ but could be different) be different)

if $\Delta_j = -1$ we allow token delete or turnstile stream model. If we have requirement that x_j's are always ≥ 0 we have strict turnstile model.

if $\Delta_j > 0$ then we have cash register model.

A linear sketch corresponds to a $k \times n$ projection matrix M and the sketch for vector x is simply $M \cdot x$ Composing linear sketches corresponds to simply adding the sketches: $Mx + Mx' = M(x+x')$

F_2 estimation as a linear sketch:

It is not hard to see that AMS F_2-estimator is a linear sketch.

Consider M as the rand. $t \times n$ matrix with entries in $\{+1,-1\}$ with $t = 6/\varepsilon^2$; we can think of entry M_{ij} as $h_i(j)$ where h_i is a 4-wise ind. hash func. for the i'th iteration. It is not hard to see that if x is the freqency vector then $M.x$ is $(Z_1, \ldots Z_t)$ where Z_i is from t'th indepent run of basic AMS. (in the final version we choose $t = O(\frac{1}{\varepsilon^2} \log \frac{1}{\delta})$ and the algorithm post-process $M.x$ to output the estimate.

In this case, after applying Chebychev's inequality:

$$\Pr\left[\left|\frac{1}{t}\sum_{i=1}^{t} Z_i^2 - F_2\right| \geq \varepsilon F_2\right] \leq \frac{1}{3}$$

So we have $M.x$ as a linear sketch where produces this vector $z = (z_1 \dots z_t)$ as estimator for F_2.

$$E\left[\sum_{i=1}^{t} z_i^2\right] = t\,E[z_i^2] = t F_2$$

So if we consider $\frac{1}{\sqrt{t}} M.x$ then using Chebychev's:

$$\left\|\frac{M.x}{\sqrt{t}}\right\|_2 = \frac{1}{\sqrt{t}}\|x\|_2 \in \left[\sqrt{1-\varepsilon}\,F_2, \sqrt{1+\varepsilon}\,F_2\right]$$

- Note that the analysis still works even if x has neg. values. We can estimate $\|f_\sigma - f_{\sigma'}\|_2$ for two streams σ, σ' this way. Similarly if x, x' are two n-dimensional signals representing two time series then the ℓ_2-norm of their difference can be estimated by one pass over them even when the signals are given by a seq. of updates
- We can interpret this as follows. The (random) matrix $\frac{M}{\sqrt{t}}$ (that has values in $\{\sqrt{t}, -\sqrt{t}\}$) performs a dimension reduction, simplifying an n-dimension vector f to a t-dimension vector (sketch) with $t = O(1/\varepsilon^2)$ while preserving ℓ_2-norm within a $(1 \pm \varepsilon)$

This is guaranteed with prob. $\geq \frac{2}{3}$ which we can boost.

Johnson-Lindenstraus Dimension reduction

AMS F_2 sketching that does dimension reduction.

The well-known JL lemma says that any n points in d-dimensional Euclidean space $\mathbb{R}^d$ can be prodjected into $O(\ln n / \varepsilon^2)$-dimension while preserving all pairwise Euclidean distances approximately:

Theorem (JL reduction): Let $\varepsilon \in (0, \frac{1}{2})$ and $x_1 \dots x_n$ be a set of n points in $\mathbb{R}^d$. Then there is a be a set of n points in $\mathbb{R}^d$. Then there is a linear map $f: \mathbb{R}^d \to \mathbb{R}^k$ where $k\,O(\alpha \ln n / \varepsilon^2)$ s.t. for all $1 \le i < j \le n$:

$$\Pr\left[(1-\varepsilon)\|x_i - x_j\|_2 \le \|f(x_i) - f(x_j)\|_2 \le (1+\varepsilon)\|x_i - x_j\|_2\right] \ge 1 - \frac{1}{n^\alpha}$$

The mapping in fact is a $k \times d$ matrix M and

$f(x_i) = M \cdot x_i / \sqrt{k}$

Each entry of M, M_{ij}, is picked independently from the sandard normal dist. $N(0,1)$ (a.k.a. Gaussian).

then $f(x) = \frac{1}{\sqrt{k}} M x$

This lemma relies on properties normal dist.

Lemma 3: let $X_1 \dots X_k$ be k independent standard normal variables (i.e. $N(0,1)$, normal dist with mean 0 and var 1) and let $Y = a_1 X_1 + \dots a_k X_k$, then Y has a normal dist with mean 0 and variance $a_1^2 + \dots + a_k^2$, i.e. $Y \in N(0, \sqrt{\sum_{i=1}^{k} a_i^2})$

Lemma 4: Let $X_1 \dots X_k$ be independent $N(0,1)$ random variables and let $Y = \frac{1}{k}\sum_{i=1}^{k} X_i^2$. Then for any $\varepsilon \in (0, \frac{1}{2})$, there is a const c s.t.:

$$\Pr\left[(1-\varepsilon)^2 \le Y \le (1+\varepsilon)^2\right] \ge 1 - e^{-c\varepsilon^2 k}$$

In other words, the sum of square of k standard normal dist. is concentrated around its mean.

The proof is similar to chernoff bound and can be found in refrences provided.

To prove JL theorem, by scaling, it is enough to prove it for unit length vectors

Lemma 5: To prove JL theorem it is enough to prove that for any arbitrary unit vector $v \in \mathbb{R}^d$ we have

Lemma 5: To prove JL theorem it is enough to prove that for any arbitrary unit vector $v \in \mathbb{R}^d$ we have

$$\Pr\left[(1-\varepsilon) \le \|f(v)\|_2 \le (1+\varepsilon)\right] \ge 1 - \frac{1}{n^{\alpha+2}} \quad (*)$$

proof: See references.

So we focus on proving $(*)$ only.

Lemma 6: $\mathbb{E}\left[\|f(v)\|_2^2\right] = \|v\|_2^2 = 1$

proof: let $\delta_{jk} = 1$ if $j=k$ and 0 o.w. Then

$$\begin{aligned}
E\left[\|f(v)\|_2^2\right] &= \mathbb{E}\left[\left\|\frac{1}{\sqrt{k}} M.v\right\|_2^2\right] \\
&= \sum_{i=1}^{k} \frac{1}{k} E\left[\left(\sum_{j=1}^{d} M_{ij} v_j\right)^2\right] \\
&= \sum_{i=1}^{k} \frac{1}{k} \sum_{1\le j,k \le d} v_j v_k E\left[M_{ij} M_{ik}\right] \\
&= \sum_{i=1}^{k} \frac{1}{k} \sum_{1\le j,k \le d} v_j v_k \delta_{jk} \qquad \text{since } M_{ij}, M_{ik} \text{ are } \in N(0,1) \\
&= \sum_{i=1}^{k} \frac{1}{k} \sum_{j=1}^{d} v_j^2 \\
&= \|v\|_2^2 \\
&= 1
\end{aligned}$$

∎

Now consider a unit vector $v \in \mathbb{R}^d$

Lemma 7: For $\varepsilon \in (0, \frac{1}{2})$ and $k = \frac{(\alpha+2)\ln n}{c\varepsilon^2}$ where c is the const. in Lemma 4:

$$\Pr\left[1-\varepsilon \le \|f(v)\|_2 \le 1+\varepsilon\right] \ge 1 - \frac{1}{n^{\alpha+2}}$$

Proof: We know $(Mv)_i = \sum_{j=1}^{d} M_{ij} \cdot v_j$. Since M_{ij} are ind. and from $N(0,1)$, by Lemma 3 we have

$(Mv)_i = N\left(0, \sqrt{\sum_{j=1}^{d} v_j^2}\right) = N(0,1)$ since $\|v\|_2^2 = 1$

So we can apply Lemma 4 to $\frac{1}{k}\sum_{i=1}^{k}(Mv)_i^2 = \frac{1}{k}\|Mv\|_2^2$ and conclude that

$$\Pr\left[(1-\varepsilon)^2 \le \frac{\|Mv\|_2^2}{k} \le (1+\varepsilon)^2\right] \ge 1 - e^{-c\varepsilon^2 k} = 1 - e^{-(\alpha+2)\ln n} = 1 - \frac{1}{n^{\alpha+2}}$$

Therefore, $\Pr\left[1-\varepsilon \le \|f(v)\|_2 \le 1+\varepsilon\right] \ge 1 - \frac{1}{n^{\alpha+2}}$

∎