CS369G: Algorithmic Techniques for Big Data	Spring 2015-2016
Lecture 5: Moment estimation via Max-stability	
Prof. Moses Charikar	*Scribes: Jiakun Li*

1 Overview

In this lecture, we will review the sketch for F_p estimation when $0 < p \leq 2$. We will show that this algorithm could be implemented with small space via Nisan's pseudorandom generator [1].

Next, we will present Andoni's algorithm [2] for estimating the $p > 2$ frequency moment. The algorithm approximates an n-dimensional l_p norm with l_∞ of an m-dimensional vector, where $m = O(n^{1-\frac{2}{p}} \cdot \log n)$.

2 Recap for F_p when $0 < p \leq 2$

Recall that in the last lecture, we construct the linear sketch for $0 < p \leq 2$ frequency moment based on p-stable distribution $\mathcal{D}_p$. A distribution $\mathcal{D}_p$ is said to be p-stable if the following property holds: Let $Y_1, \ldots, Y_n$ be independent random variables drawn from $\mathcal{D}_p$, then $\sum_i x_i Y_i$ has the same distribution as $||x||_p Y$, $Y \sim \mathcal{D}_p$. In the last lecture we presented the following algorithm to estimate the p-th frequency moment.

Algorithm 1: F_p estimate where $0 < p \leq 2$

$\mathbf{x} \leftarrow (x_1, ...x_n)$;
$k \leftarrow \Theta(\frac{1}{\epsilon^2} \log \frac{1}{\delta})$;
Let M be a k × n matrix where each $M_{ij} \sim \mathcal{D}_p$;
$\mathbf{y} \leftarrow M\mathbf{x}$;
return $Y \leftarrow \left[\frac{median(|y_1|, |y_2|, ..., |y_k|)}{median(|\mathcal{D}_p|)} \right]$;

Remark: Note that the matrix multiplication could be done in a streaming fashion. We start with all-zero $\mathbf{y}$, and for each x_i take the i^{th} column of M and update $\mathbf{y} \leftarrow \mathbf{y} + \sum_{j=1}^{k} M_{ij} x_i$.

By the p-stability property we see that each $y_i \sim ||x||_p Y$ where $Y \sim \mathcal{D}_p$. The following lemma shows that the median of $|y_i|$'s has good concentration properties.

Lemma 1. *Let $\epsilon > 0$ and $\mathcal{D}_p$ be a p-stable distribution. Let $F(t)$ be the probability density function of $|\mathcal{D}_p|$, μ be the median of $|\mathcal{D}_p|$, and $\alpha = min_{t \in [\mu(1-\epsilon), \mu(1+\epsilon)]} F(t)$. Denote $y = median(|y_1|, |y_2|, ..., |y_k|)$, where y_i are independent random variables drawn from $\mathcal{D}_p$. Then*

$$Pr(y \leq (1-\epsilon)\mu) \leq \frac{\delta}{2}$$

holds when $k = \Theta\left(\frac{1}{\epsilon^2}\log\frac{1}{\delta}\right)$

Proof. Let $F(t)$ be the density function of $|\mathcal{D}_p|$, then F(t) is the density function of $\mathcal{D}_p$ scaled by 2 if $t \geq 0$ and $F(t) = 0$ if $t < 0$. $|y_1|, ..., |y_k| \sim |\mathcal{D}_p|$. The median μ is uniquely defined and it satisfies

$$\int_0^{\mu} F(t)dt = \frac{1}{2}$$

$F(t)$ is continuous on $[(1-\epsilon)\mu, (1+\epsilon)\mu]$.

$$Pr\left(|y_i| \leq \mu(1-\epsilon)\right) = \frac{1}{2} - \int_{\mu(1-\epsilon)}^{\mu} F(t)dt \leq \frac{1}{2} - \alpha\mu\epsilon$$

Let $\gamma = \alpha\mu\epsilon$, L be the number of $|y_i|$'s that fall in the range of $[0, \mu(1-\epsilon)]$.

$$L = |\{i : |y_i| \leq \mu(1-\epsilon)\}|$$

$$E[L] \leq k\left(\frac{1}{2} - \gamma\right) = \frac{k}{2}(1-2\gamma)$$

Since y is the median of $|y_i|$, $y \leq (1-\epsilon)\mu$ only if more than half of $|y_i|$ are low, which is the same as $L > k/2$.

Let $1 + \delta = \frac{1}{1-2\gamma}$.

$$Pr(y \leq (1-\epsilon)\mu) = Pr\left(L > \frac{k}{2}\right) = Pr\left(L > \frac{1}{1-2\gamma}E(L)\right) = Pr(L > (1+\delta)E(L))$$

Using Chernoff bound,

$$Pr(y \leq (1-\epsilon)\mu) \leq e^{\frac{-\delta^2 E(L)}{3}} \leq e^{\frac{-\gamma^2 E(L)}{3}} \leq e^{-\frac{k}{2}\frac{\alpha^2\epsilon^2\mu^2(1-2\alpha\epsilon\mu)}{3}} = e^{-c\epsilon^2 k} \leq \frac{\delta}{2}$$

$$k = O(\frac{1}{\epsilon^2}\log\frac{1}{\delta})$$

□

3 Derandomization of space bounded computation

In the algorithm described above we have to keep the entire matrix M around which is often too expensive for streaming applications. However, given that the algorithm only needs to operate on $S = O(\frac{1}{\epsilon^2}\log 1/\delta)$ bits, one can use pseudorandom generators instead of truly random bits to reduce the required storage.

3.1 Nisan's Pseudorandom Generator

Theorem 2. *Let U_n denote a uniformly random string in $\{0,1\}^n$. There exists $h : \{0,1\}^t \rightarrow \{0,1\}^{SR}$, $t = S\log R$.*

$$Pr(f(U_n)=1) - Pr(f(h(U_m))=1) \leq 2^{-O(S)}$$

for any function f: $\{0,1\}^S \rightarrow \{0,1\}$.

In other words, the distribution of 2^S states generated by a truly random string is indistinguishable from the distribution of a Nisan pseudorandom generator.

The way Nisan works is as follows: Assume we have $h_1, ..., h_{\log n}$, where $h_i : [2^S] \rightarrow [2^S]$ are pair-wise independent hash functions. We choose a random sample $x \in \{0,1\}^S$, place it at the root and repeat the following procedure: on level i, create the left child as the same as its parent p and the right child as $h_i(p)$. Using Nisan, we can take the seed of $S\log R$ bits, expand it to SR bits such that any chunk of S bits can be generated in $S\log R$ time.

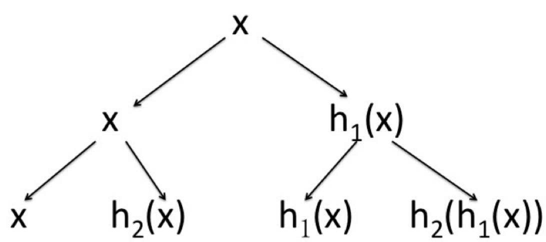

Figure 1: Nisan's pseudorandom generator

4 $p > 2$ Frequency Moments via Max-stability

Andoni proposed an algorithm [2] to estimate F_p when $p > 2$ using space $O(n^{1-\frac{2}{p}}\log n)$. The algorithm consists of two-step mapping. Let $x \in \mathbb{R}^n$ be the input vector. Let u_i's be random variables drawn from an exponential distribution with density e^{-t}, in the first step we scale each x_i by $u_i^{-\frac{1}{p}}$,

$$y_i = \frac{x_i}{u_i^{1/p}}$$

In the second step, we compute $z \in \mathbb{R}^m$ using a random hash function $h : [n] \rightarrow [m]$.

$$z_j = \sum_{i:h(i)=j} \sigma_i \cdot y_i$$

where σ_i are random ± 1. The final estimator is given by $\max_{j\in[m]} |z_j| = \|z\|_\infty$.

4.1 Analysis

We first claim the $\max_i y_i = ||y||_\infty$ is a good estimate on $||x||_p$.

Lemma 3.

$$Pr(||y||_\infty \in [\frac{1}{2}||x||_p, 2||x||_p]) \geq \frac{3}{4}$$

Proof. Let $q = \min\{\frac{u_1}{|x_1|^p}, ..., \frac{u_p}{|x_n|^p}\}$. Given $u_1, u_2, ...u_n$ are i.i.d random variables drawn from the exponential distribution e^{-t},

$$\begin{aligned} P(q > t) &= P(\forall i, u_i > t|x_i|^p) \\ &= \prod_{i=1}^{n} e^{-t|x_i|^p} \\ &= e^{-t|x_i|_p^p} \end{aligned}$$

Therefore,

$$\begin{aligned} P(\frac{1}{2}\|x\|_p \leq \|y\|_\infty < 2\|x\|_p) &= P(\frac{1}{2^p \sum_i |x_i|^p} \leq q \leq \frac{2^p}{\sum_i |x_i|^p}) \\ &= e^{-\frac{1}{2p}} - e^{-2p} \\ &\geq \frac{3}{4} \end{aligned}$$

for $p > 2$. □

In next lecture, we will show that the second step preserves $\|y\|_\infty$ with good probability.

References

[1] Nisan, Noam. ”Pseudorandom generators for space-bounded computation.” Combinatorica 12.4 (1992): 449-461.

[2] Andoni, Alexandr. ”High frequency moment via max stability.” Unpublished manuscript (2012).