CS 598CSC: Algorithms for Big Data	Lecture date: Sept 9, 2014
Instructor: Chandra Chekuri	Scribe: Chandra Chekuri

1 Sketch for F_p Estimation when $0 < p \leq 2$

We have seen a linear sketching estimate for F_2 estimation that uses $O(\log n)$ space. Indyk [1] obtained a technically sophisticated and interesting sketch for F_k estimation where $0 < p \leq 2$ (note that p can be a real number) which uses polylog(n) space. Since the details are rather technical we will only give the high-level approach and refer the reader to the paper and related notes for more details. Note that for $p > 2$ there is a lower bound of $\Omega(n^{1-2/p})$ on the space required.

To describe the sketch for $0 < p \leq 2$ we will revisit the F_2 estimate via the JL Lemma approach that uses properties of the normal distribution.

F_2-ESTIMATE:

Let $Y_1, Y_2, \ldots, Y_n$ be sampled independenty from the $\mathcal{N}(0,1)$ distribution
$z \leftarrow 0$
While (stream is not empty) do
  (i_j, Δ_j) is current token
  $z \leftarrow z + \Delta_j \cdot Y_{i_j}$
endWhile
Output z^2

Let $Z = \sum_{i \in [n]} x_i Y_i$ be the random variable that represents the value of z at the end of the stream. The variable Z is a sum of independent normal variables and by the properties of the normal distribution $Z \sim \sqrt{\sum_i x_i^2} \cdot \mathcal{N}(0,1)$. Normal distribution is called 2-stable for this reason. More generally a distribution $\mathcal{D}$ is said to be p-stable if the following property holds: Let $Z_1, Z_2, \ldots, Z_n$ be independent random variables distributed according to $\mathcal{D}$. Then $\sum_i x_i Z_i$ has the same distribution as $\|x\|_p Z$ where $Z \sim \mathcal{D}$. Note that a p-stable distribution will be symmetric around 0.

It is known that p-stable distributions exist for all $p \in (0, 2]$ and not for any $p > 2$. The p-stable distributions do not have, in general, an analytical formula except in some cases. We have already seen that the standard normal distribution is 2-stable. The 1-stable distribution is the Cauchy distribution which is the distribution of the ratio of two independent standard normal random variables. The density function of the Cauchy distribution is known to be $\frac{1}{\pi(1+x^2)}$; note that the Cauchy distribution does not have a finite mean or variance. We use $\mathcal{D}_p$ to denote a p-stable distribution.

Although a general p-stable distribution does not have an analytical formula it is known that one can sample from $\mathcal{D}_p$. Chambers-Mallows-Stuck method is the following:

- Sample θ uniformly from $[-\pi/2, \pi/2]$.
- Sample r uniformly from $[0, 1]$.
- Ouput

$$\frac{\sin(p\theta)}{(\cos\theta)^{1/p}} \left(\frac{\cos((1-p)\theta)}{\ln(1/r)} \right)^{(1-p)/p}.$$

We need one more definition.

Definition 1 *The median of a distribution $\mathcal{D}$ is θ if for $Y \sim \mathcal{D}$, $\Pr[Y \leq \mu] = 1/2$. If $\phi(x)$ is the probability density function of $\mathcal{D}$ then we have $\int_{-\infty}^{\mu} \phi(x)dx = 1/2$.*

Note that a median may not be uniquely defined for a distribution. The distribution $\mathcal{D}_p$ has a unique median and so we will use the terminology median($\mathcal{D}_p$) to denote this quantity. For a distribution $\mathcal{D}$ we will refer to $|\mathcal{D}|$ the distribution of the absolute value of a random variable drawn from $\mathcal{D}$. If $\phi(x)$ is the density function of $\mathcal{D}$ then the density function of $|\mathcal{D}|$ is given by ψ, where $\psi(x) = 2\phi(x)$ if $x \geq 0$ and $\psi(x) = 0$ if $x < 0$.

F_p-ESTIMATE:

$k \leftarrow \Theta(\frac{1}{\epsilon^2} \log \frac{1}{\delta})$
Let M be a $k \times n$ matrix where each $M_{ij} \sim \mathcal{D}_p$
$\mathbf{y} \leftarrow M\mathbf{x}$
Output $Y \leftarrow \frac{\text{median}(|y_1|,|y_2|,\ldots,|y_k|)}{\text{median}(|\mathcal{D}_p|)}$

By the p-stability property we see that each $y_i \sim \|x\|_p Y$ where $Y \sim \mathcal{D}_p$. First, consider the case that $k = 1$. Then the output $|y_1|/\text{median}(|\mathcal{D}_p|)$ is distributed according to $c|\mathcal{D}_p|$ where $c = \|x\|_p/\text{median}(|\mathcal{D}_p|)$. It is not hard to verify that the median of this distribution is $\|x\|_p$. Thus, the algorithm take k samples from this distribution and ouputs as the estimator the *sample* median. The lemma below shows that the sample median has good concentration properties.

Lemma 1 *Let $\epsilon > 0$ and let $\mathcal{D}$ be a distribution with density function ϕ and a unique median $\mu > 0$. Suppose ϕ is absolutely continuous on $[(1-\epsilon)\mu, (1+\epsilon)\mu]$ and let $\alpha = \min\{\phi(x) \mid x \in [(1-\epsilon)\mu, (1+\epsilon)\mu]$. Let $Y = median(Y_1, Y_2, \ldots, Y_k)$ where $Y_1, \ldots, Y_k$ are independent samples from the distribution $\mathcal{D}$. Then*

$$\Pr[|Y - \mu| \geq \epsilon\mu] \leq 2e^{-\frac{2}{3}\epsilon^2\mu^2\alpha^2 k}.$$

We sketch the proof to upper bound $\Pr[Y \leq (1-\epsilon)\mu]$. The other direction is similar. Note that by the definition of the median, $\Pr[Y_j \leq \mu] = 1/2$. Hence

$$\Pr[Y_j \leq (1-\epsilon)\mu] = 1/2 - \int_{(1-\epsilon)\mu}^{\mu} \phi(x)dx.$$

Let $\gamma = \int_{(1-\epsilon)\mu}^{\mu} \phi(x)dx$. It is easy to see that $\gamma \geq \alpha\epsilon\mu$.

Let I_j be the indicator event for $Y_i \leq (1-\epsilon)\mu$; we have $\mathbf{E}[I_j] = \Pr[Y_i \leq (1-\epsilon)\mu] \leq 1/2 - \gamma$. Let $I = \sum_j I_j$; we have $\mathbf{E}[I] = k(1/2 - \gamma)$. Since Y is the median of $Y_1, Y_2, \ldots, Y_k$, $Y \leq (1-\epsilon)\mu$ only if more than $k/2$ of I_j are true which is the same as $\Pr[I > (1+\delta)\mathbf{E}[I]]$ where $1 + \delta = \frac{1}{1-2\gamma}$. Now, via Chernoff bounds, this probability is at most $e^{-\gamma^2 k/3}$ for sufficiently small γ.

We can now apply the lemma to the estimator output by the algorithm. We let ϕ be the distribution of $c|\mathcal{D}_p|$. Recall that the median of this distribution if $\|x\|_p$ and the output of the algorithm is the median of k indepenent samples from this distribution. Thus, from the lemma,

$$\Pr[|Y - \|x\|_p| \geq \epsilon\|x\|_p] \leq 2e^{-\epsilon^2 k\mu^2\alpha^2/3}.$$

Let ϕ' be the distribution of $|\mathcal{D}_{\surd}|$ and μ' be the median of ϕ'. Then it can be seen that $\mu\alpha = \mu'\alpha'$ where $\alpha' = \min\{\phi'(x) \mid (1-\epsilon)\mu' \leq (1+\epsilon)\mu'\}$. Thus $\mu'\alpha'$ depends only on $\mathcal{D}_p$ and ϵ. Letting this be $c_{p,\epsilon}$ we have,

$$\Pr[|Y - \|x\|_p| \geq \epsilon\|x\|_p] \leq 2e^{-\epsilon^2 k c_{p,\epsilon}^2/3} \leq (1-\delta),$$

provided $k = \Omega(c_{p,\epsilon} \cdot \frac{1}{\epsilon^2} \log \frac{1}{\delta})$.

Technical Issues: There are several technical issues that need to be addressed to obtain a proper algorithm from the preceding description. First, the algorithm as described requires one to store the entire matrix M which is too large for streaming applications. Second, the constant k depends on $c_{p,\epsilon}$ which is not explicitly known since $\mathcal{D}_p$ is not well-understood for general p. To obtain a streaming algorithm, the very high-level idea is to derandomize the algorithm via the use of pseudorandom generators for small-space due to Nisan. See [1] for more details.

2 Counting Frequent Items

We have seen various algorithm for estimating various F_p norms for $p \geq 0$. Note that F_0 corresponds to number of distinct elements. In the limit, as $p \to \infty$, ℓ_p norm of a vector $\mathbf{x}$ is the maximum of the absolute values of the entries of $\mathbf{x}$. Thus, we can define the F_∞ norm to corresponds to finding the maximum frequency in $\mathbf{x}$. More generally, we would like to find the frequent items in a stream which are also called "heavy hitters". In general, it is not feasible to estimate the heaviest frequency with limited space if it is too small relative to m.

2.1 Misra-Greis algorithm for frequent items

Suppose we have a stream $\sigma = a_1, a_2, \ldots, a_m$ where $a_j \in [n]$, the simple setting and we want to find all elements in $[n]$ such that $f_i > m/k$. Note that there can be at most k such elements. The simplest case is when $k = 2$ when we want to know whether there is a "majority" element. There is a simple deterministic algorithm that perhaps you have all seen for $k = 2$ in an algorithm class. The algorithm uses an associative array data structure of size k.

MisraGreis(k):

D is an empty associative array
While (stream is not empty) do
  a_j is current item
  If (a_j is in $keys(D)$)
    $D[a_j] \leftarrow D[a_j] + 1$
  Else if ($|keys(A)| < k-1$) then
    $D[a_j] \leftarrow 1$
  Else
    for each $\ell \in keys(D)$ do
      $D[\ell] \leftarrow D[\ell] - 1$
    Remove elements from D whose counter values are 0
  endWhile
  For each $i \in keys(D)$ set $\hat{f}_i = D[i]$
  For each $i \notin keys(D)$ set $\hat{f}_i = 0$

We leave the following as an exercise to the reader.

Lemma 2 *For each $i \in [n]$:*

$$f_i - \frac{m}{k} \leq \hat{f}_i \leq f_i.$$

The lemma implies that if $f_i > m/k$ then $i \in keys(D)$ at the end of the algorithm. Thus one can use a second-pass over the data to compute the exact f_i only for the k itmes in $keys(D)$. This gives an $O(kn)$ time two-pass algorithm for finding all items which have frequency at least m/k.

Bibliographic Notes: For more details on F_p estimation when $0 < p \leq 2$ see the original paper of Indyk [1], notes of Amit Chakrabarti (Chapter 7) and Lecture 4 of Jelani Nelson's course.

References

[1] Piotr Indyk. Stable distributions, pseudorandom generators, embeddings, and data stream computation. *Journal of the ACM (JACM)*, 53(3):307–323, 2006.