Data Stream Algorithms.

Basic definitions.

- Stream: n elements from universe $[m] = \{1, 2, ..., m\}$.
  E.g: Consider $[1000]$
  $$\{x_1, x_2, ..., x_m\} = 3, 5, 7, 100, ...$$
- Goal: Compute a function of stream,
  E.g. Median, number of distinct elements, longest increasing sequence.

- But: — limited working memory, (usually sublinear in n and m, i.e. $O(\log n)$ or $O(\log m)$)
  — access data sequentially.
  — Process quickly.

Why do we care?
  Faster network, cheaper data storage, ...

* Sampling: a general technique to tackling massive amounts of data.

  E.g. we have a large list of all queries made to a search engine, and we want to measure how many queries contain the word "~~iPhone XS~~". Easy! just count them ?!! But we can actually do it faster. $\Rightarrow$ sampling.

Problem: Given a large set of N elements U, $(|U| = N)$, select a subset of elements $\hat{U}$ $(|\hat{U}| \approx n)$ such that from $\hat{U}$ the size of any subset $S \subset U$ can be estimated

· Sampling approach: Pick each element from U independently into set $\hat{U}$ with probability $p = \frac{n}{N}$.

Let the variable X_i be 1 if element i is picked and 0 otherwise.

The number of picked elements is $\sum_{i=1}^{N} X_i$ and its expectation is

$$E\left[\sum_{i=1}^{N} X_i\right] = \sum_{i=1}^{N} E[X_i] = \sum_{i=1}^{N} \frac{n}{N} = n.$$

Let $\hat{S}$ be the set of the intersection of S and $\hat{U}$.

$$\hat{S} = S \cap \hat{U}.$$

Let s_i be 1 if $i \in S$ and 0 otherwise
↑ indicator function.

Let $Z = \frac{N}{n} |\hat{S}|$ be our estimator of $|S|$.

$$E[Z] = E\left[\frac{N}{n}|\hat{S}|\right] = \frac{N}{n} E\left[\sum_{i=1}^{N} X_i s_i\right] = \frac{N}{n} E\left[\sum_{j \in S} X_j\right] = |S|$$

$i \in \hat{S}$ if and only if $X_i s_i = 1$.

∴ $E[Z] = |S|$

question:

Question: How close is Z to $|S|$?
Chernoff's bound would help!

<u>Lemma</u> (Chernoff bound): Let $X_1, ..., X_n$ be independent Bernoulli random variables $\mathbb{P}(X_i = 1) = p_i$ and $\mathbb{P}(X_i = 0) = 1-p_i$. Let $X = \sum_{i=1}^{n} X_i$ and $\mu = \mathbb{E}[X] = \sum_{i=1}^{n} p_i$. Then, for any $\varepsilon > 0$,

$$\mathbb{P}\left(X > (1+\varepsilon)\mu\right) \leq e^{-\mu\varepsilon^2/4}$$

$$\mathbb{P}\left(X < (1-\varepsilon)\mu\right) \leq e^{-\mu\varepsilon^2/2}$$

$$\mathbb{P}\left(|X-\mu| \geq \varepsilon\mu\right) \leq 2e^{-\mu\varepsilon^2/4}.$$

Recall that in our problem, we want to know how go over the elements of S and count how many of them were sampled into $\hat{U}$, and that

$$\frac{n}{N} Z = \sum_{i=1}^{n} X_i \Delta_i = \sum_{j=1}^{\cancel{}} \cancel{X_j} \sum_{j \in S} X_j \Rightarrow Z = \frac{N}{n} \sum_{j \in S} X_j.$$

and $\frac{n}{N}\mathbb{E}[Z] = |S|$

Applying Chernoff's for $\sum_{j \in S} X_j$; $\mathbb{P}\left(\left|\sum_{j \in S} X_j - \frac{n}{N}|S|\right| > (1+\varepsilon \frac{n|S|}{N}\right) \leq 2e^{-\frac{|S| n\varepsilon^2}{4}}$

$\therefore \mathbb{P}\left(\left|\frac{N}{n}\sum_{j \in S} X_j - |S|\right| > \varepsilon|S|\right) \leq 2\exp\left(-\frac{|S| n\varepsilon^2}{4N}\right)$

$\begin{cases} \mathbb{P}\left(Z > (1+\varepsilon)|S|\right) \leq e^{-|S| n\varepsilon^2/4N} \\ \mathbb{P}\left(Z < (1-\varepsilon)|S|\right) \leq e^{-|S| n\varepsilon^2/2N} \end{cases}$

union bound
$\Rightarrow$

$$\mathbb{P}\left(\left|Z - |S|\right| > \varepsilon|S|\right) \leq 2e^{-|S| n\varepsilon^2/4N}$$

what does it mean?

For example, if $|S|$ is of the size $10^{-5}N$. and we want to have a 10% accuracy with probability at least 0.99, we must keep a sample of roughly 10^8 elements, regardless of N.
                                                                      big number but
                    think of N a really big number, it's
            still small!

* Frequency moments of Data stream:

Given a data stream $a_1, a_2, \ldots, a_n$ of length n, where each $a_j \in \{1, 2, \ldots, m\} =: [m]$. The frequency of $i \in [m]$ in the stream is $f_i = |\{ j \mid a_j = i \}|$.

The vector $\vec{f} = (f_1, f_2, \ldots, f_m)$ is called the frequency vector.

For $p \geq 0$, The pth frequency moment of the input is defined as follows:

number of distinct symbols occuring in the stream

$$F_p = \begin{cases} |\{ i \mid f_i \neq 0 \}| & \text{if } p = 0 \\ \max_i f_i & \text{if } p = \infty \\ \|f\|_p^p = \sum_{i=1}^{m} f_i^p & \text{otherwise} \end{cases}$$

- For $p = 1$, the first frequency moment is just n, the length of the string.
- For $p = 2$, the second frequency moment is useful in computing the variance of the stream:

$$\frac{1}{m} \sum_{i=1}^{m} \left(f_i - \frac{n}{m} \right)^2 = \frac{1}{m} \sum_{i=1}^{m} \left(f_i^2 - 2 f_i \cdot \frac{n}{m} + \frac{n^2}{m^2} \right)$$

$$= \left(\frac{1}{m} \sum_{i=1}^{m} f_i^2 \right) - \frac{n^2}{m^2}.$$

- For $p = \infty$, F_∞ is the frequency of the most frequent element.

* **The uniform distribution:**

A r.v. X assumes values in the interval $[a, b]$ such that all subintervals of equal length have equal probability, we say that X has the **uniform distribution** over $[a, b]$.

The probability distribution function of X is.

$$F(x) = \begin{cases} 0 & \text{if } x \leq a \\ \dfrac{x-a}{b-a} & \text{if } a \leq x \leq b \\ 1 & \text{if } x \geq b. \end{cases}$$

and its density function is

$$f(x) = \begin{cases} 0 & \text{if } x < a. \\ \dfrac{1}{b-a} & \text{if } a \leq x \leq b. \\ 0 & \text{if } x > b. \end{cases}$$

$$E[X] = \int_a^b \frac{x}{b-a}\, dx = \frac{b^2 - a^2}{2(b-a)} = \frac{b+a}{2}.$$

$$E[X^2] \overset{(\text{Exercise!})}{=} \cdots = \frac{b^2 + ab + a^2}{3}$$

$$Var[X] = ?$$

Lemma: Let $X_1, X_2, \ldots, X_k$ be independent uniform random variables over $[0,1]$. Let $Y = \min(X_1, X_2, \ldots, X_k)$.
Then $E[Y] = \dfrac{1}{k+1}$.

$$\mathbb{P}(Y \geq y) = \mathbb{P}\left(\min(X_1, \dots, X_k) \geq y\right)$$

$$= \mathbb{P}\left(\{X_1 \geq y\} \cap \{X_2 \geq y\} \cap \dots \cap \{X_k \geq y\}\right)$$

X_i's are independent $\longleftarrow$
$$= \prod_{i=1}^{k} \mathbb{P}(X_i \geq y)$$

$$= (1-y)^k$$

$$\therefore \quad \mathbb{P}(Y \leq y) = 1 - (1-y)^k$$
$$F(y) = 1 - (1-y)^k$$

density function of y is $\quad F'(y) = f(y) = k(1-y)^{k-1}$

$$\Rightarrow \quad E[Y] = \int_0^1 ky(1-y)^{k-1} \, dy = y(1-y)^k \Big|_{y=0}^1 + \int_0^1 (1-y)^k \, dy$$

Integration by parts
$\qquad\qquad\qquad\qquad\qquad = 0 \qquad + \int_0^1 (1-y)^k \, dy$

$u = y \qquad dv = k(1-y)^{k-1}$

$du = dy \qquad v = -(1-y)^k$
$$= -\frac{(1-y)^{k+1}}{k+1} \Big|_{y=0}^1$$

$$= \frac{1}{k+1}$$

* <u>Estimating F_0</u> = <u>Counting distinct elements</u>.

[1] Noga Alon, Yossi Matias, and Mario Szegedy.
  The space complexity of approximating the frequency moments. STOC '96.

[2] Edith Cohen.
  Size-estimation framework with applications to transitive closure and reachability. '97.

Idea: use a (hash) function $h: [m] \to [0,1]$. ← universe

We hash each entry a_i of the data as we see it, and keep track of the minimum seen hash value in our memory. Suppose in $a_1, a_2, ..., a_n$, there are k distinct elements $x_1, x_2, ..., x_k$.

Let $Y = \min\left(h(a^{x_1}), h(a^{x_2}), ..., h(a^{x_k})\right)$.

Suppose that the values $h(a_1), ..., h(a_n)$ are independently distributed uniform r.v. over the interval $[0,1]$.

$\Rightarrow$ From the previous lemma:

$$\cancel{E[Y] \geq \frac{1}{n+1}}$$

$$E[Y] = \frac{1}{k+1}$$

Recall that we want to estimate k, so Y may be used to estimate it.

$\hookrightarrow$ Can use Chebyshev's inequality.

$$E[Y^2] = \int_0^1 y^2 \, k(1-y)^{k-1} \, dy$$

$$= .... = \quad ? \quad \leq \quad \frac{2}{(k+1)^2}$$
$$\text{(Exercise!)}$$

$\Rightarrow$ $Var[Y] = E[Y^2] - E[Y]^2 \leq \frac{1}{(k+1)^2} = E[Y]^2$

Chebyshev's inequality:

$$\mathbb{P}\left(|Y - E[Y]| > \varepsilon E[Y]\right) \leq \frac{Var[Y]}{\varepsilon^2 E[Y]^2} \leq \frac{1}{\varepsilon^2}$$

which is a useless bound for small ε.

To improve, we take a mean of estimators.

Consider multiple independent versions $Y_1, Y_2, \ldots, Y_t$ of Y.

Y_1 has b corresponding hash function h_1

Y_2 $\qquad\qquad\qquad\qquad\qquad\qquad h_2$

$\vdots$ $\qquad\qquad\qquad\qquad\qquad\qquad \vdots$

Y_t $\qquad\qquad\qquad\qquad\qquad\qquad h_t$

And let $Z = \dfrac{Y_1 + Y_2 + \ldots + Y_t}{t}$ as our new estimator.

$$\mathbb{E}[Z] = \mathbb{E}[Y] = \frac{1}{k+1}.$$

Since $Y_1, \ldots, Y_t$ are independent,

$$\mathbb{E} \quad \operatorname{Var}[Z] = \frac{1}{t^2} \sum_{i=1}^{t} \operatorname{Var}[Y_i] = \frac{\operatorname{Var}[Y]}{t} \leq \frac{\mathbb{E}[Y]^2}{t}$$

$$\therefore \quad \operatorname{Var}[Z] \leq \frac{\mathbb{E}[Z]^2}{t}$$

Applying Chebyshev's ineq:

$$\mathbb{P}\left(|Z - \mathbb{E}[Z]| \geq \varepsilon \, \mathbb{E}[Z] \right) \leq \frac{\operatorname{Var}[Z]}{\varepsilon^2 \, \mathbb{E}[Z]^2} \leq \frac{1}{\varepsilon^2 t}.$$

This means that by increasing t, we can reduce the probability of bad event $\{ |Z - \mathbb{E}[Z]| \geq \varepsilon \, \mathbb{E}[Z] \}$.

Setting $t = \dfrac{10}{\varepsilon^2}$, we can bound the probability of failure by $\dfrac{1}{1} \cdot \dfrac{1}{10}$

* **Estimating the second moment F_2**

Recall $\quad F_2 = \sum_{i=1}^{m} f_i^2$

Goal: Estimate F_2.

Consider a hash function $h: [m] \longrightarrow \{-1, 1\}$.

For each symbol i, $1 \leq i \leq m$, independently set a random variable X_i such that

$$P(X_i = 1) = P(X_i = -1) = 1/2$$

then consider

$$S = \sum_{i=1}^{m} X_i f_i \quad \text{and} \quad V = \left(\sum_{i=1}^{m} X_i f_i \right)^2$$

Fact: $\quad \mathbb{E}[V] = \sum_{i=1}^{m} f_i^2$

pf: Note that

$$\left(\sum_{i=1}^{m} X_i f_i \right)^2 = \sum_{i=1}^{m} X_i^2 f_i^2 + 2 \sum_{i \neq j} X_i X_j f_i f_j$$

$$\Rightarrow \quad \underbrace{\mathbb{E}[V]}_{} = \mathbb{E}\left(\sum_{i=1}^{m} X_i^2 f_i^2 \right) + 2 \mathbb{E}\left(\sum_{i \neq j} X_i X_j f_i f_j \right)$$

$$= \sum_{i=1}^{m} \mathbb{E}[X_i^2 f_i^2] + 2 \sum_{i \neq j} \mathbb{E}[X_i X_j f_i f_j]$$

$\mathbb{E}[X_i X_j f_i f_j]$

$\quad = \mathbb{E}[X_i] \mathbb{E}[X_j] f_i f_j \quad$ since X_i, X_j independent for $i \neq j$.

$\quad = 0$

$$\longleftarrow \quad = \sum_{i=1}^{m} f_i^2 \quad + \quad 0.$$

$$\Rightarrow \quad V \text{ is an estimator of } F_2.$$

We can show that (see [BHK] p. 190)
$$\mathbb{E}[V^2] \leq 3\mathbb{E}[V]^2$$
$$\therefore \leq \mathbb{E}[V]$$
$$\therefore \operatorname{Var}[V] = \mathbb{E}[V^2] - \mathbb{E}[V]^2$$
$$\leq 2\mathbb{E}[V]^2.$$

By Chebyshev's inequality:

$$\mathbb{P}\left(\left|V - \mathbb{E}[V]\right| \geq \varepsilon\mathbb{E}[V]\right) \leq \frac{\operatorname{Var}[V]}{\varepsilon^2\mathbb{E}[V]^2} \leq \frac{2}{\varepsilon^2}.$$

Not a good bound for ε small.

$\Rightarrow$ we can consider multiple independent version of V
$V_1, V_2, \ldots, V_s$. and so Let $Y = \frac{1}{s}\sum_{i=1}^{s} V_i$.

Then $\mathbb{P}\left(\left|Y - F_2\right| \geq eF_2\right) \leq \delta$

if $s \geq \dfrac{2}{\varepsilon^2\delta}$.

Alon–Matias–Szegedy was able to construct
Y and V using $O(\log m)$ space. (See [BHK] p. 190).