22 Apr 2022 — Streaming Algorithm for Distinct Elements

Input stream $a_1, \ldots, a_n$.

Each a_i belongs to $\{0,1\}^b$ (set of "tokens")

$m = 2^b = \#$ of possible tokens.

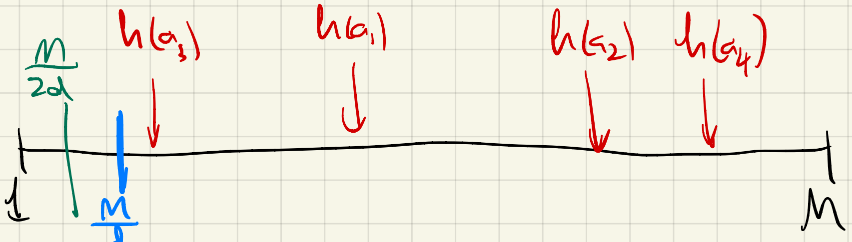
Algorithm with working space $s = \mathrm{poly}(\log n, \log m)$ observes $a_1, \ldots, a_n$ one by one, and afterward must estimate # distinct tokens appearing in the sequence.

GOAL: (ε, δ)-PAC meaning: if true number of distinct tokens is d, then

$$\Pr\left(\text{ALG's output is not in } [(1-\varepsilon)d, (1+\varepsilon)d]\right) < \delta.$$

Use hash function $h : \{0,1\}^b \to [M]$ for some large integer M. (Typically $M \geq 2^b$.)

Suppose for now that $a_1, \ldots, a_d$ are the d distinct tokens that appear in the stream and that $h(a_1), h(a_2), \ldots, h(a_d)$ are independent uniformly random elements of $[M]$.



$$\Pr\left(h(a_1), \ldots, h(a_d) > \frac{M}{d}\right) = \left(1 - \frac{1}{d}\right)^d \approx e^{-1}.$$

$$E[\min\{h(a_1), \ldots, h(a_d)\}] \approx \frac{M}{d}.$$

$$\Pr\left(\min\{h(a_i)\} \leq \frac{M}{2d}\right) \leq \sum_{i=1}^{d} \Pr\left(h(a_i) \leq \frac{M}{2d}\right) = \frac{1}{2}.$$

Alg.

① Initialize $z = M$

② for each a_i in succession

- compute $h(a_i)$
- if $h(a_i) < z$, $z \leftarrow h(a_i)$

// invariant: after t iterations

$z = \min\{h(a_j) \mid 1 \le j \le t\}$

③ output M/z as our estimate of d.

Analysis. For $k \in [M]$ let $X_{ik} = \begin{cases} 1 & \text{if } h(a_i) \le k \\ \emptyset & \text{if } h(a_i) > k \end{cases}$

$$Y_k = \sum_{i=1}^{d} X_{ik}$$

assuming $a_1, \ldots, a_d$ are the distinct token

$=$ # distinct tokens in stream whose hash value is $\le k$.

Fact 1. $\mathbb{E}[X_{ik}] = \frac{k}{M}$

I: $h(a_i)$ is uniform in $[M]$ for all i

Fact 2. $\mathbb{E}[Y_k] = \frac{dk}{M}$

Fact 3. $\mathrm{Var}(Y_k) < \frac{dk}{M}$.

II: $h(a_i)$ and $h(a_j)$ are independent

$$\mathrm{Var}(Y_k) = \mathbb{E}(Y_k^2) - \mathbb{E}(Y_k)^2$$

$$= \sum_{i=1}^{d}\sum_{j=1}^{d} \mathbb{E}[X_{ik}X_{jk}] - \left(\frac{dk}{M}\right)^2$$

$$= \sum_{i=1}^{d} \mathbb{E}[X_{ik}^2] + 2\sum_{1 \le i < j \le d} \mathbb{E}[X_{ik}X_{jk}] - \left(\frac{dk}{M}\right)^2$$

$$= \frac{dk}{M} + 2\binom{d}{2}\left(\frac{k}{M}\right)^2 - \left(\frac{dk}{M}\right)^2$$

$$= \frac{dk}{M} + (d^2-d)\left(\frac{k}{M}\right)^2 - \left(\frac{dk}{M}\right)^2$$

$$= \frac{dk}{M} - \frac{dk^2}{M^2}$$

$$< \frac{dk}{M}.$$

<u>Def.</u> A 2-universal hash function (UHF) is a probability distribution over hash functions that satisfies:

I. $h(a_i)$ is unif. distrib. over its range for all i.

II. $h(a_i)$, $h(a_j)$ independent $\forall i \neq j$.

"pairwise independence"

<u>Example.</u> Say M is prime, and identify $\{0,1\}^b$ with $[m]$, and assume $M \geq m$.

Define h as follows: $h_{ab}(x) = ax+b \pmod M$

Sample a, b as indep't, unif random mod M, set $h = h_{ab}$.

Storing a description of h takes $2\log(M)$ bits to store a, b.

To verify 2-universality, must show

$$\forall a_i \neq a_j \quad \Pr_{a,b}\left(h(a_i) = x,\ h(a_j) = y \right) = \frac{1}{M^2} \quad \forall x, y$$

$h_{ab}(a_i) = x$, $h_{ab}(a_j) = y$ means

$$\begin{pmatrix} a_i & 1 \\ a_j & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} \equiv \begin{pmatrix} x \\ y \end{pmatrix} \pmod{M}$$

The linear system has a unique solution $\pmod{M}$ b/c $\mathbb{Z}/(M)$ is a field and $\det\begin{pmatrix} a_i & 1 \\ a_j & 1 \end{pmatrix} = a_i - a_j \not\equiv 0 \pmod{M}$