2/19/2025

Pairwise Hash functions or

2-Universal Hash functions

$\mathcal{H} = \{ h : [m] \to [M] \}$

$\forall \; x \neq y \in [m]$, and

any $i, j \in [M]$,

$$\Pr_{h \sim \mathcal{H}} \left[h(x) = i \text{ and } h(y) = j \right] = \frac{1}{M^2}.$$

$\to$ (1) $h \sim \mathcal{H}$, $h(x)$ is uniformly distributed on $[M]$

(2) $h \sim \mathcal{H}$, $h(x)$, $h(y)$ are indep.

$\mathcal{H} = \{ h : [m] \to [M] \}$

Construction: Pick M to be a prime number, $M \geq m$

$\mathcal{H}$ "$= \left\{ h_{a,b}(x) = ax + b \pmod M \right\}_{a,b \in [M]}$

$\hookrightarrow h_{a,b} : [m] \to [M]$

Question: $|\mathcal{H}| = M^2$

Observe: Storing $h_{a,b}$ takes $2 \log M$ bits.

Claim: $\mathcal{H}$ as defined above is a pairwise indep hash family.

Pf. Pick $x \neq y \in [m]$

Let $i, j \in [M]$

WTS:

$$\Pr_{a,b \sim [M]}\left[h_{a,b}(x) = i \wedge h_{a,b}(y) = j \right] = \frac{1}{M^2}.$$

$$\Pr_{a,b \sim [M]}\left[ax + b = i, \ ay + b = j \right]$$

$$\left\{ \begin{array}{l} ax + b = i \\ ay + b = j \end{array} \right\}$$

$a(x - y) = i - j$. Recall $x \neq y$,

$\Rightarrow x - y \in \{-m+1, -m+2, \dots,$

$-1, 1, 2, \ldots, m-1\}$

$\Rightarrow \quad x - y \neq 0 \mod M$

$\hookrightarrow$ prime

$\Rightarrow \quad (x-y)^{-1}$ is well-defined

$[\ (x-y)\cdot(x-y)^{-1} = 1\ ]$

Thus, $\quad a = (i-j)\cdot(x-y)^{-1}.$

$b = i - ax = i - a(i-j)\cdot(x-y)^{-1}.$

$$\therefore \Pr_{a,b\sim[m]}[\ ax+b=i,\ ay+b=j\ ]$$

$$= \frac{1}{m^2}.$$

An application to Streaming Algorithms

Data stream: each a_i is b bits.

$a_1, a_2, a_3, \ldots, a_n$ $\quad$ $m = 2^b$

$\hookrightarrow$ data items

Algorithm has 'limited memory'

S bits of memory.

Trivial: $S = n \cdot b$

$\hookrightarrow$ $S << n b$

$\hookrightarrow$ poly$(\log n), \ldots$

Counting distinct elements

$a_1, a_2, \ldots, a_n$, $a_i = [2^b] = [m]$

Task: with prob. $1-\delta$, output a number in $[(1-\varepsilon)d, (1+\varepsilon)d]$,

trying to estimate where d is the # of distinct elements.

$\rightarrow$ $\mathcal{H} = \{ h : [m] \rightarrow [M] \}$

pairwise indep, $m \le M \le 2m$.

$2\lceil \log M \rceil$ bits.

Alg 1: $z = \perp$; $h \sim \mathcal{H}$

for $i = 1$ to n

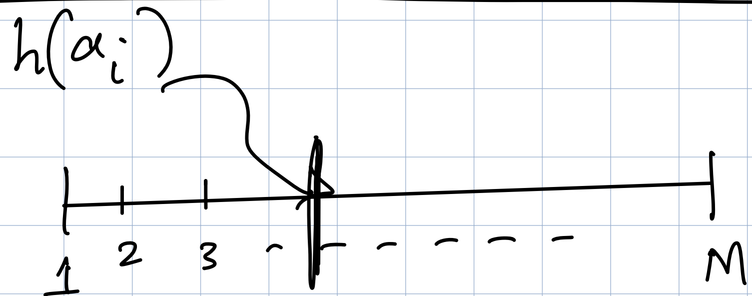
compute $h(a_i)$;

if $h(a_i) < Z$,

update $Z = h(a_i)$

end for

Output: $\dfrac{M}{Z}$



$$Z \sim \frac{M}{d}$$

Space: $\lceil 2 \log M \rceil + \log m \leq 3 \underline{\log} m + O(1)$

For $k = 1, 2, \ldots, m$, define

$\rightarrow \quad X_{ik} = 1$ if $h(a_i) \leq k$

0 o\w

$$Y_k = \sum_{i=1}^{d} X_{ik},$$ → # of distinct elements hashed to atmost k.

w.l.o.g assume $a_1, a_2, \ldots, a_d$ are distinct.

(i) $\mathbb{E}[X_{ik}] = \frac{k}{m}$.

(ii) $\operatorname{Var}(X_{ik}) = \mathbb{E}[X_{ik}^2] - \mathbb{E}[X_{ik}]^2$

$$= \frac{k}{m} - \frac{k^2}{m^2} \leq \frac{k}{m}.$$

(iii) $$\mathbb{E}[Y_k] = \sum_{i=1}^{d} \mathbb{E}[X_{ik}]$$

linearity of expectation

$$= \frac{dK}{M}$$

$$\text{Var}[Y_k] = \text{Var}\left(\sum_{i=1}^{d} X_{ik}\right)$$

$$= \sum_{i=1}^{d} \text{Var}(X_{ik})$$

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

$$= \mathbb{E}\left[\sum_{i=1}^{d} X_{ik}^2 + \sum_{\substack{i_1, i_2, \\ i_1 \neq i_2}} X_{i_1 k} X_{i_2 k}\right]$$

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

$$= \sum_{i=1}^{d} \mathbb{E}[X_{ik}^2] - \sum_{i=1}^{d} \mathbb{E}[X_{ik}]^2$$

$$+ \sum_{\substack{i_1, i_2, \\ i_1 \neq i_2}} \left(\mathbb{E}[X_{i_1 k} X_{i_2 k}] - \mathbb{E}[X_{i_1 k}] \cdot \mathbb{E}[X_{i_2 k}]\right)$$

$\overset{\shortparallel}{0}$

Since $X_{i_1 k}$ and $X_{i_2 k}$ are independent [h is pairwise indep]

$$= \sum_{i=1}^{d} Var(X_{ik}) .$$

$$\boxed{\begin{aligned} \mathbb{E}[Y_k] &= d \cdot \frac{k}{m} \quad \text{and} \\ Var(Y_k) &\leq d \cdot \frac{k}{m} \end{aligned}}$$

Goal:

$\rightarrow$ $\Pr\left[\frac{M}{Z} \notin \left[\frac{d}{6}, 6d\right]\right]$ is small.

① $\Pr\left[\frac{M}{Z} \leq \frac{d}{6}\right]$

$= \Pr\left[Z \geq \frac{6M}{d}\right]$

$= \Pr\left[Y_q = 0\right]$,

where $q = \left\lfloor \frac{6M}{d} \right\rfloor$