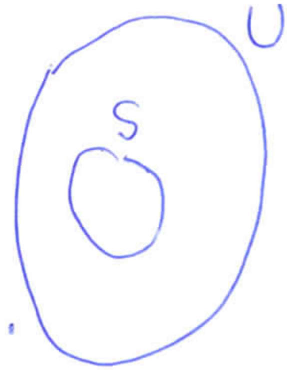
Hashing

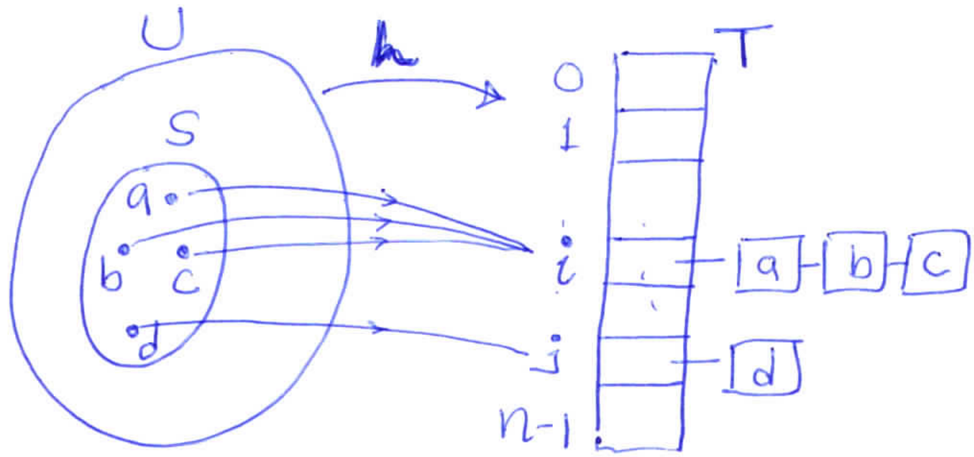


- To maintain $S \subseteq U$, $|S| = n \ll |U|$.
- Search ($x \in S$?), add, delete from S.

Example U = All people in the world.
S = All residents of NYC.

Trivial: - Maintain array of size $|U|$.
- Too much space.

Hash Table



- Maintain an array $T[0], T[1], \ldots, T[n-1]$.
- Pick a "hash function" $h : U \to \{0, 1, \ldots, n-1\}$.
- Store $x \in S$ at location $T[h(x)]$.

Example person → (eye-color, height, nationality).

Collisions - All $x \in S$ s.t. $h(x) = i$ are stored at location $T[i]$ in a list.

- Search(x) takes time $O(k)$ if this list has size k.

Search Given x, search list at $T[h(x)]$.

Add " add to " .

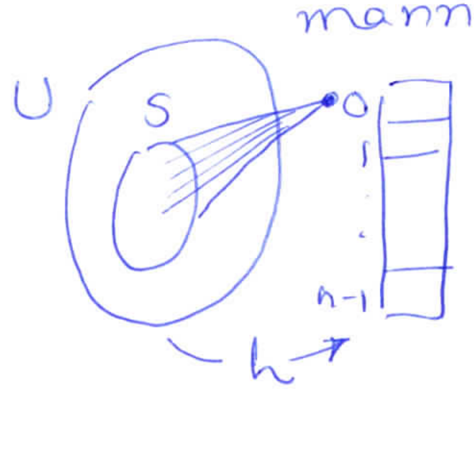
Delete " delete from " .

We desire that - Very few collisions
- Sizes of lists are small.

Note. Randomization is necessary, i.e. we cannot pick the hash function $h : U \to \{0, 1, \ldots, n-1\}$ in fixed, a priori, det. manner.

Because

S coud, adverserially, be such that $\forall x \in S$, $h(x) = 0$.



Hence - Randomization!

- Let $\mathcal{H}$ be a $\frac{\text{family}}{\text{collection}}$ of functions from U to $\{0,1,..,n-1\}$.

  Pick $h \in \mathcal{H}$ (uniformly) at random.

- Show that for any $S \subseteq U$, $|S| = n$, over the choice of $h \in \mathcal{H}$, few collisons, small lists.

Tradeoff If h is $\frac{\text{truly}}{\text{completely}}$ random, (i.e. $\mathcal{H}$ is family of all functions $U \to \{0,1,..,n-1\}$) then the scheme "works". However, then "storing" h takes space proportional to $|U|$.

Hence we desire that

① $h \in \mathcal{H}$ is "random enough"

② - $h \in \mathcal{H}$ has compact representation ("formula")

eq $|\mathcal{H}|$ is "small".

Def A family of functions $\mathcal{H}$, $U \to \{0,1,..,n-1\}$ is called **2-universal / pairwise independent** if

① $\forall x \in U$, $\forall i \in \{0,1,..,n-1\}$

$$\Pr_{h \in \mathcal{H}} [h(x) = i] = \frac{1}{n}.$$

② $\forall x, y \in U$, $x \neq y$, $\forall i, j \in \{0,1,..,n-1\}$

$$\Pr_{h \in \mathcal{H}} [h(x) = i \wedge h(y) = j] = \frac{1}{n^2}.$$

Note

- ② $\Rightarrow$ ①
- ② $\Rightarrow$ $\forall x, y \in U$, $x \neq y$,

$$\Pr_{h \in \mathcal{H}} [h(x) = h(y)] = \frac{1}{n}.$$

Theorem There is an explicit, concrete, 2-universal family of hash functions $\mathcal{H}$ and all $h \in \mathcal{H}$ are efficiently represented & computed.

Here onwards let $\mathcal{H}$ be a 2-universal family of hash functions $h : U \to \{0,1,\dots,n-1\}$

For $i \in \{0,1,\dots,n-1\}$, let $L(i)$ denote the list of all elements in S hashed to location i.

All probabilities/expectations are over choice of $h \in \mathcal{H}$.

__Lemma__

$$\mathbb{E}[|L(i)|] = 1.$$

__Proof__ For every $a \in S$, let X_a be indicator r.v.

$$X_a = \begin{cases} 1 & \text{if } h(a) = i \\ 0 & \text{otherwise.} \end{cases}$$

$$\mathbb{E}[X_a] = \Pr[h(a) = i] = \frac{1}{n}.$$

$$|L(i)| = \sum_{a \in S} X_a.$$

$$\therefore \mathbb{E}[|L(i)|] = \sum_{a \in S} \mathbb{E}[X_a] = n \cdot \frac{1}{n} = 1.$$

$\blacksquare$

Markov's Inequality

Let X be a non-negative random var and $t \geqslant 1$. Then

$$\Pr[X \geqslant t \cdot \mathbb{E}[X]] \leq \frac{1}{t}.$$

Lemma $\Pr[|L(i)| \geqslant t] \leq \frac{1}{t}$. (Think of $t = 50$).

Proof. $\mathbb{E}[|L(i)|] = 1$.

Markov's inequality.

$\blacksquare$

Chebychev's Inequality

Def. Let X be a r.v. Its variance

$$\mathrm{var}(X) = \mathbb{E}[|X - \mathbb{E}[X]|^2]$$

$$= \mathbb{E}[|X - \mu|^2] \qquad \mu = \mathbb{E}[X].$$

Fact $\mathrm{var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

$$= \mathbb{E}[X^2] - \mu^2.$$

Proof
$$\begin{aligned} var(X) &= \mathbb{E}[|X-\mu|^2] \\ &= \mathbb{E}[X^2 - 2\mu X + \mu^2] \\ &= \mathbb{E}[X^2] - 2\mu\cdot\mathbb{E}[X] + \mu^2 \\ &= \mathbb{E}[X^2] - \mu^2. \end{aligned}$$
∎

Chebychev's Inequality

Let X be a ~~non-negative~~ r.v. Then

$$\Pr[|X-\mu| \geq T] \leq \frac{var(X)}{T^2}.$$

$\mu = \mathbb{E}[X]$.

Proof
$$\begin{aligned} \Pr[|X-\mu| \geq T] &= \Pr[|X-\mu|^2 \geq T^2] \\ &\leq \frac{\mathbb{E}[|X-\mu|^2]}{T^2} \quad \text{Markov.} \\ &= \frac{var(X)}{T^2}. \end{aligned}$$
∎

Corollary. If X is a non-negative r.v. Then for $t \geq 1$

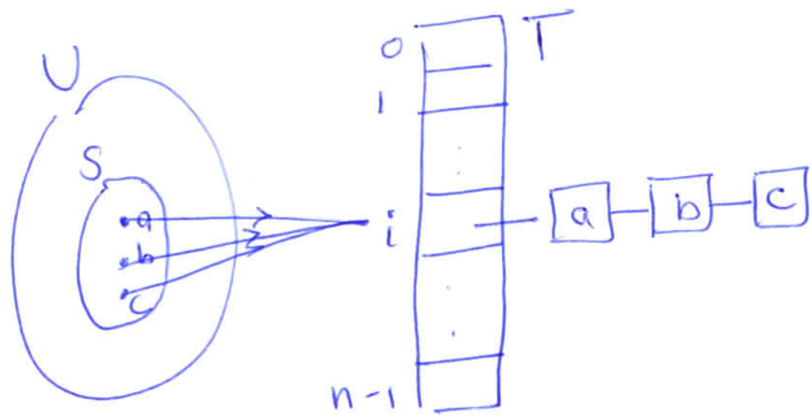
$$\Pr[X \geq t\cdot\mathbb{E}[X]] \leq \frac{var(X)}{(t-1)^2\mu^2}.$$

Proof $\Pr[X \geq t\,\mathbb{E}[X]] \leq \Pr[|X - \mathbb{E}[X]| \geq (t-1)\,\mathbb{E}[X]]$

$$\leq \frac{\mathrm{var}(X)}{(t-1)^2 \cdot \mathbb{E}[X]^2}.$$

$\blacksquare$

Recall



- $h \in \mathcal{H}$ from 2-universal family.
- $L(i) = \{x \in S \mid h(x) = i\}$.
- $\forall\, x, y \in U,\ x \neq y,\ \forall\, i, j \in \{0, 1, \ldots, n-1\}$

$$\Pr[h(x) = i \wedge h(y) = j] = \frac{1}{n^2}.$$

- $\mathbb{E}[|L(i)|] = 1$.

Claim $\mathbb{E}[|L(i)|^2] \leq 2$.

Hence $\mathrm{var}(|L(i)|) = \mathbb{E}[|L(i)|^2] - \mathbb{E}[|L(i)|]^2 \leq 1$.

Proof $\forall a \in S$, let X_a be a r.v.

$$X_a = \begin{cases} 1 & \text{if } h(a)=i \\ 0 & \text{otherwise.} \end{cases}$$

$$\therefore |L(i)| = \sum_{a \in S} X_a .$$

$$\therefore \mathbb{E}[|L(i)|^2] = \mathbb{E}\Big[\Big(\sum_{a \in S} X_a\Big)^2\Big]$$

$$= \mathbb{E}\Big[\sum_{a,b \in S} X_a X_b\Big]$$

$$= \sum_{a \in S} \mathbb{E}[X_a^2] + \sum_{\substack{a \neq b \\ a,b \in S}} \mathbb{E}[X_a X_b]$$

$$= \sum_{a \in S} \Pr[X_a = 1] + \sum_{\substack{a \neq b \\ a,b \in S}} \Pr[X_a = 1 \wedge X_b = 1]$$

$$= n \cdot \frac{1}{n} + n(n-1) \cdot \frac{1}{n^2}$$ $\because$ 2-universality

$$\leq 2.$$

$\blacksquare$

Lemma $\Pr[\,|L(i)| \geq t\,] \leq \dfrac{1}{(t-1)^2}$

Proof Applying the corollary,

$$\Pr[\,|L(i)| \geq t\,] = \Pr[\,|L(i)| \geq t \cdot \mathbb{E}[|L(i)|]\,]$$

$$\leq \frac{var(|L(i)|)}{(t-1)^2 \mu^2} \qquad \mu = \mathbb{E}[|L(i)|] = 1.$$

$$\leq \frac{1}{(t-1)^2}. \qquad \blacksquare$$

$\therefore$ For every i, probability that $|L(i)| \geq 50$ is $\leq \dfrac{1}{2000}$.

Note. $\underbrace{\mathbb{E}\left[\sum_{i=0}^{n-1} |L(i)|^2\right]}$ $\leq 2n$.

- Interpretation: Sum over $a \in S$, cost of SEARCH(a).
- $\therefore$ After hashing, average cost of Search(a) is $O(1)$.

Example of 2-Universal Hash family

- Suppose $|S| = |U| = p$. (prime).
- Consider family of hash functions
  $h_{a,b} : U \to \{0,1,\ldots,p-1\}$, $U = \{0,1,\ldots,p-1\}$
- $\mathcal{H} = \{ h_{a,b} \mid a,b \in \{0,1,\ldots,p-1\} \}$ where
  $h_{a,b}(x) = ax + b \pmod p$.
- $|\mathcal{H}| = p^2$.
- 2-universality Fix $x, y \in U = \{0,1,\ldots,p-1\}$
  $x \neq y$.
  $i, j \in \{0,1,\ldots,p-1\}$.

Then $h_{a,b}(x) = i \Rightarrow ax + b = i$

$h_{a,b}(y) = j \Rightarrow ay + b = j$

$$\begin{bmatrix} x & 1 \\ y & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} i \\ j \end{bmatrix}$$

has unique solution (a^*, b^*).

$$\therefore \Pr_{h_{a,b} \in \mathcal{H}}\left[h_{a,b}(x) = i \wedge h_{a,b}(y) = j \right] = \frac{1}{p^2}.$$

- <u>Generalization</u>

  - Let $U = \{0, 1, \dots, p-1\}^k$.

  - $$\mathcal{H} = \left\{ h_{\substack{a_1, \dots, a_k \\ b_1, \dots, b_k}} \;\middle|\; \begin{array}{l} a_1, \dots, a_k \\ b_1, \dots, b_k \end{array} \in \{0, 1, \dots, p-1\} \right\}.$$

  where

$$h_{\substack{a_1, \dots, a_k \\ b_1, \dots, b_k}}(x = (x_1, \dots, x_k)) = \sum_{i=1}^{k} a_i x_i + b_i \pmod{p}.$$

$\blacksquare$