LECTURE 12 UNIVERSAL HASHING 17/12/2020

Recap:
- large universe $U = \{0, 1, \dots, m-1\}$
- small $S \subseteq U$, $|S| \ll |U|$ keys
- FIND, INSERT, DELETE
- hash function: $h : U \to \overbrace{\{0, 1, \dots, n-1\}}^{=V}$
- there is a collision for $x \neq y \in S$, if $h(x) = h(y)$

weakness of a single function hashing: an adversary can select $S = \{x \in U \mid h(x) = i\}$ for some $i \in \{0, 1, \dots, n-1\}$

IDEA: Choose the hash function at random, independently of the keys in S

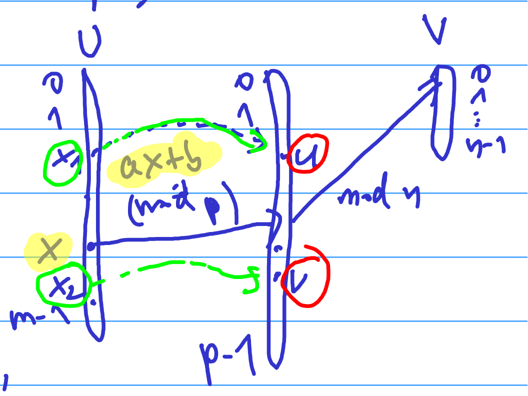
$\Rightarrow$ universal hashing ... in expectation, works well for any S

Def: A family of hash functions $\mathcal{H}$ from U to V is 2-universal if for any $x, y \in U$, $x \neq y$ and a hash function $h \in \mathcal{H}$ chosen uniformly at random, we have $\Pr[h(x) = h(y)] \leq \frac{1}{n}$.

For $U = \{0, \dots, m-1\}$, $V = \{0, \dots, n-1\}$ and a prime $p \geq m$, and integers a, b, define

$$h_{a,b}(x) = ((ax+b) \bmod p) \bmod n$$

and $\mathcal{H} = \{h_{a,b} \mid 1 \leq a \leq p-1, 0 \leq b \leq p\}$



Lemma: $\mathcal{H}$ is 2-universal.

Proof: fix $x_1, x_2 \in U$, $x_1 \neq x_2$

C1: $ax_1 + b \neq ax_2 + b \mod p$

$\Rightarrow$ C2: For each $u, v \in \{0, 1, \dots, p-1\}$, $u \neq v$, there exist an exactly one pair a, b s.t.

$$\left.\begin{aligned} ax_1 + b &= u \mod p \\ ax_2 + b &= v \mod p \end{aligned}\right]$$

Proof: $\begin{pmatrix} x_1 & 1 \\ x_2 & 1 \end{pmatrix}$ ... regular matrix $\rightsquigarrow$ system of lin. equations has a solution goal: count $|\{h \in \mathcal{H} \mid h(x_1) = h(x_2)\}| \le \frac{p(p-1)}{n}$

(*) For each $u \in \{0, \ldots, p-1\}$, there are at most $\lceil \frac{p}{n} \rceil - 1$ values of $v \neq u$ s.t. $u = v \mod n$.

$\Rightarrow$ there are at most $p \cdot \left(\lceil \frac{p}{n} \rceil - 1 \right)$ pairs of u, v, $u \neq v$ for which $u = v \mod n$

For each such pair u, v (i.e., $u \neq v$, $u = v \mod n$) there is a unique function $h_{a,b}$ s.t. x_1, x_2 are in collision under $h_{a,b}$: $h_{a,b}(x_1) = h_{a,b}(x_2)$.

$$p\left(\lceil \tfrac{p}{n} \rceil - 1\right) \le p\left(\frac{p+n-1}{n} - 1\right) = p\,\frac{p-1}{n}$$

$$\Rightarrow \Pr[h_{a,b}(x_1) = h_{a,b}(x_2)] \le \frac{p(p-1)/n}{p(p-1)} = \frac{1}{n}$$

$\blacksquare$

Static dictionary

Def: A hash function $h: U \to V$ is perfect for $S \subseteq U$ if h does not have any collisions among elements in S.

Lemma: Assume that $S \subseteq U$, $|S| = m$, is hashed into V, $|V| = n$, using a hash function h chosen uniformly at random from a 2-universal family. Then for an arbitrary $x \in U$

$$\mathbb{E}[X] \le \begin{cases} \frac{m}{n} & \text{if } x \notin S \\ 1 + \frac{m-1}{n} & x \in S \end{cases}$$

where $X = |\{y \in S \mid h(y) = h(x)\}|$.

Proof: let $S = \{e_1, e_2, \ldots, e_m\}$

$\forall i \quad X_i = 1$ if $h(e_i) = h(x)$; $X = \sum_{i=1}^{m} X_i$

$x \notin S$: $\mathbb{E}[X] = \sum_{i=1}^{m} \mathbb{E}[X_i] \leq m \cdot \frac{1}{n} = \frac{m}{n}$

$x \in S$: wlog $x = e_1$, $Pr[h(e_i) = h(x)] = \begin{cases} 1 & \text{for } i = 1 \\ \frac{1}{n} & i \neq 1 \end{cases}$

$$\mathbb{E}[X] \leq 1 + \frac{m-1}{n}$$ ∎

Lemma: If $h \in \mathcal{H}$ is chosen uniformly at random from a 2-universal family, then for any set $S \subseteq U$ of size m, the probability that

i) h is perfect for S is at least $\frac{1}{2}$, if $n \geq m^2$

ii) h has at most m collisions is at least $\frac{1}{2}$, if $n = m$.

Proof: let $S = \{e_1, \dots, e_m\}$

let $X_{ij} = \begin{cases} 1 & \text{if } h(e_i) = h(e_j) \\ 0 & \text{otherwise} \end{cases}$

$Pr[X_{ij} = 1] = \frac{1}{n}$ $\Leftarrow$ as $\mathcal{H}$ is 2-universal family

$$X = \sum_{1 \leq i < j \leq m} X_{ij}.$$

$$\mathbb{E}[X] = \sum_{1 \leq i < j \leq m} \mathbb{E}[X_{ij}] \leq \binom{m}{2} \frac{1}{n} < \frac{m^2}{2n}$$

By Markov inequality:

$$Pr\left[X \geq \frac{m^2}{n}\right] \leq Pr[X \geq 2\mathbb{E}[X]] \leq \frac{1}{2}$$

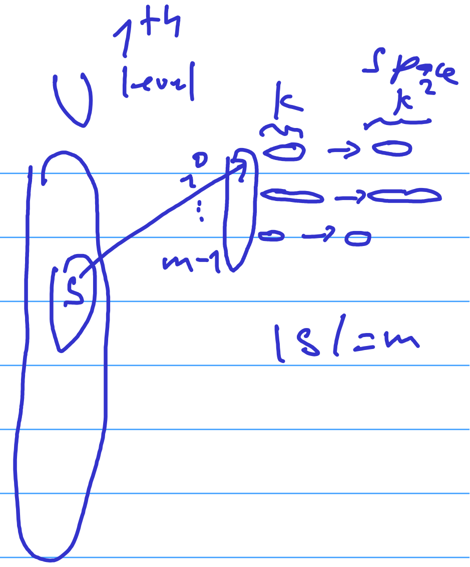
i) if $n \geq m^2$: $\frac{m^2}{n} \leq 1$ <1

$\Rightarrow$ no collision with prob. $\geq \frac{1}{2}$

ii) if $n = m$ $\frac{m^2}{n} = m$

$\Rightarrow$ $\leq m$ collisions with prob $\geq \frac{1}{2}$ ∎

GOAL : PERFECT HASHING FOR $S \subseteq U$



IDEA: two level hashing

first level: map S to $\underbrace{\{0, 1, .., m-1\}}_{n = m \text{ elements}}$

$\Rightarrow$ ensure $\leq m$ collisions

second level: for each $i \in \{0, 1, .., m-1\}$ cell with k elements, map them by a private hash function to k^2 space --- perfect mappings

Theorem: The two-level approach gives a perfect hashing scheme for m items using $O(m)$ space.

Proof: first level - by previous lemma find h with $\leq m$ collisions

second level

let c_i -- # of items in the cell $i \in \{0, 1, --, m-1\}$; $\forall i$

Note: • $\sum_{i=0}^{m-1} c_i = m \quad (= |S|)$

• as # of collisions $\leq \binom{c_i}{2}$ in each cell i

$\Rightarrow$ total number of collisions $\sum_{i=0}^{m-1} \binom{c_i}{2} \leq m$

Total space : $m + \sum_{i=0}^{m-1} c_i^2 = m + 2 \sum_{i=0}^{m-1} \binom{c_i}{2} + \sum_{i=0}^{m-1} c_i \leq$

$\leq m + 2m + m = 4m$ $\blacksquare$

ALGEBRAIC PROBLEMS & RANDOMIZATION

RANDOMIZED VERIFYING OF MATRIX MULTIPLICATION

A, B, C ... $n \times n$ matrices

Task: verify whether $A \cdot B = C$

- Trivial solution: compute $A \cdot B$
  - single matrix multiplication ... $O(n^3)$
  - sophisticated algorithm - $O(n^{2,376})$ ↑ slow
- Goal: a faster algorithm

$$\boxed{\begin{array}{llll}
1. & \text{choose a random vector } \bar{r} = (r_1, \dots, r_n) \in \{0,1\}^n & & \\
2. & \text{compute} \quad B\bar{r} & \dots\ O(n^2) & \\
 & \qquad\qquad A(B\bar{r}) & O(n^2) & \\
 & \qquad\qquad C\bar{r} & O(n^2) & O(n^2) \\
3. & \text{if} \quad A(B\bar{r}) \neq C\bar{r} \quad \text{then return "} AB \neq C \text{"} & & \\
 & \qquad \text{else return "} AB = C \text{"}. & &
\end{array}}$$

Theorem: If $AB \neq C$ and $\bar{r}$ is chosen uniformly at random from $\{0,1\}^n$, then

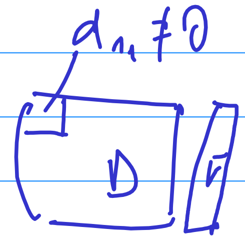
$$\Pr[AB\bar{r} = C\bar{r}] \leq \frac{1}{2}.$$

Proof: let $D = AB - C$

$$AB\bar{r} = C\bar{r} \Rightarrow \underbrace{(AB - C)}_{D}\bar{r} = 0$$

$$D\bar{r} = 0$$

As $D \neq 0$, we assume wlog $d_{11} \neq 0$



$$D\bar{r} = 0 \Rightarrow \underbrace{\sum_{j=1}^{n} d_{1j} \cdot \bar{r}_j = 0}$$

$$\Rightarrow \bar{r}_1 = - \frac{\sum_{j=2}^{n} d_{1j} \bar{r}_j}{d_{11}} \quad \text{fixed}$$

only one choice of $\bar{r}_1 \in \{0,1\}$ (after $\bar{r}_2, \dots, \bar{r}_n$ were fixed) will satisfy

$$\Rightarrow \Pr[D\bar{r} \neq 0] \geq \Pr[d_1\bar{r} \neq 0] \geq \frac{1}{2} \quad \blacksquare$$

principle of deferred decisions