CS7800: Advanced Algorithms

Lecture 19: Randomized IV

- Universal hash functions
- Perfect hashing

Jonathan Ullman

11-18-2022

Hash Tables

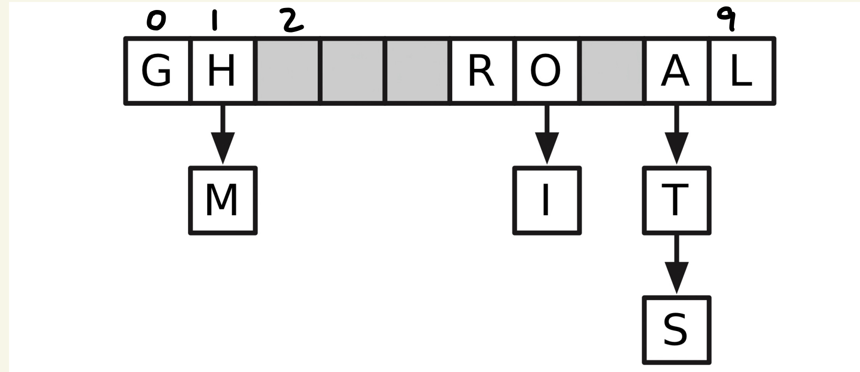
Goal: Store a set of n elements $S \subseteq \underset{\text{universe}}{U}$ so that we can look up whether $x \in U$ is in S

- A dictionary also let's us associate a value with keys x

- A hash table $T[1:m]$ stores the elements
- A hash function $h : U \to \{0, 1, \ldots, m-1\}$ maps elements to slots $x \to T[h(x)]$

Linear Chaining

A method for dealing with hash collisions



load factor n/m

$$\begin{cases} m = 10 \\ n = 10 \end{cases}$$

$U = \{A, B, C, \ldots, Z\}$

$h(G) = 0$

$h(M) = 1$

$h(A) = h(T) = 8$

- Let $\ell(x)$ be the number of elements $y \in S$ such that $h(x) = h(y)$

- Time to lookup x is $O(\ell(x))$

Randomized Hash Functions

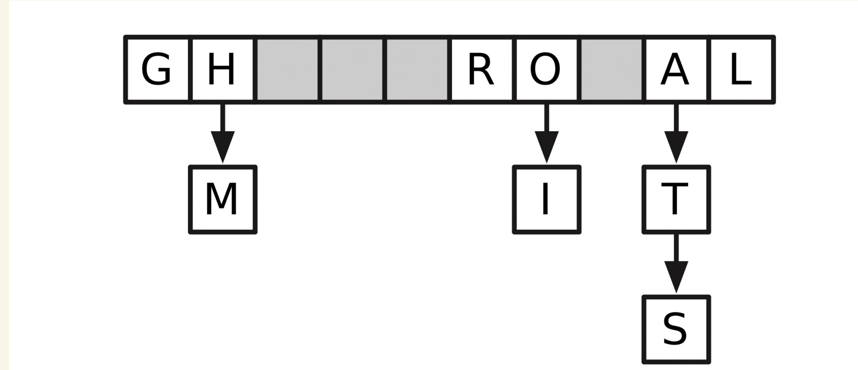
- A <u>hash family</u> $\mathcal{H} = \{ h : \mathcal{U} \to \{0, 1, \ldots, m-1\} \}$
- Choose a hash function h uniformly at random from $\mathcal{H}$.

- If $|\mathcal{H}| = 1$ (h is deterministic) then there is always a set of size $|\mathcal{U}|/m$ that all hash to the same bucket

- A uniformly random function is a "good hash family"

- There simple hash families that "good enough"

Linear Chaining with Ideal Hashing



Property of an ideal hash fn:

For any set of distinct elts $x_1, \dots, x_k$ the values $h(x_1) \dots h(x_k)$ are independent

$\ell(x) = \#$ of elts $y \in S$ s.t. $h(x) = h(y)$

Expected Lookup Time: (fixed S, fixed x, random h)

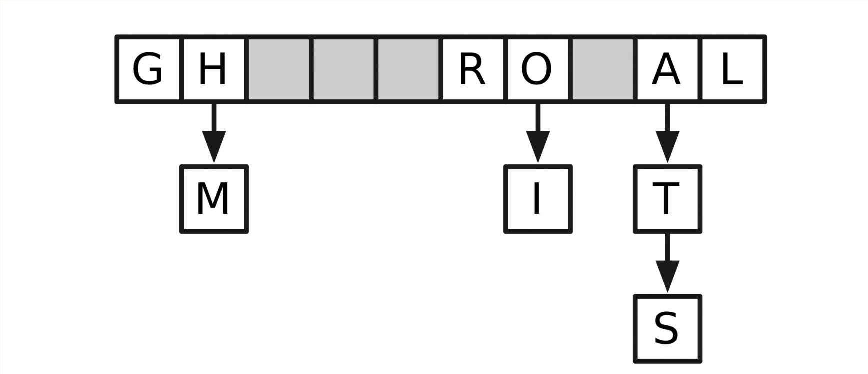
$$C_{x,y} = \begin{cases} 1 & \text{if } h(x) = h(y) \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbb{E}(\ell(x)) = \mathbb{E}\left(\sum_{y \in S} C_{x,y}\right) = \sum_{y \in S} \mathbb{E}(C_{x,y}) = \sum_{y \in S} \mathbb{P}(h(x) = h(y))$$

Assume $x \neq y$

$$= \sum_{y \in S} \frac{1}{m} = \frac{|S|}{m} = \underbrace{\frac{n}{m}}_{\text{load factor}}$$

Linear Chaining with Ideal Hashing



Property of an ideal hash fn:

For any set of distinct elts $x_1, \ldots x_k$

the values $h(x_1) \ldots h(x_k)$ are independent

$\ell(x) = \#$ of elts $y \in S$ s.t. $h(x) = h(y)$

Worst-Case Lookup Time: (S is fixed, h is random)

$\mathbb{E}(\max_{x \in U} \ell(x)) = \mathbb{E}(\text{max items in any bucket})$ ← balls and bins

(if $\frac{m}{n} = 1$) then $= \Theta\left(\frac{\log n}{\log \log n}\right)$

Pseudorandom Hash Families

A <u>hash family</u> $\mathcal{H} = \{ h : \mathcal{U} \to \{0, 1, \dots, m-1\} \}$

- A hash family $\mathcal{H}$ is <u>2-wise uniform</u> if for every $x \neq y \in \mathcal{U}$

$$\mathbb{P}(h(x) = i \text{ and } h(y) = j) = \frac{1}{m^2}$$

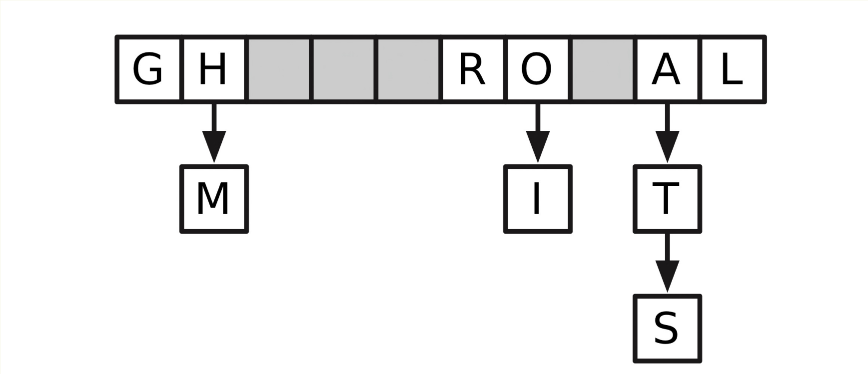
- A hash family $\mathcal{H}$ is <u>2-wise universal</u> if for every $x \neq y \in \mathcal{U}$

$$\mathbb{P}(h(x) = h(y)) \leq \frac{1}{m}$$

c_1, c_2 are coins

$$h_{c_1, c_2}(x) = \begin{cases} c_1 & x = 1 \\ c_2 & x = 2 \\ c_1 \oplus c_2 & x = 3 \end{cases}$$

Linear Chaining with Universal Hashing



If $\mathcal{H}$ is 2-universal then

$$\mathbb{E}(\ell(x)) = \frac{m}{n}$$

Constructing Universal Hash Families

Fix some prime $p > |U|$ and table size m

$$\boxed{\begin{aligned} h_{a,b}(x) &= (ax+b \bmod p) \bmod m \\ \mathcal{H}_{p,m} &= \{ h_{a,b} \text{ for } a \in \mathbb{Z}_p^+, b \in \mathbb{Z}_p \} \end{aligned}}$$

<u>Thm:</u> $\mathcal{H}_{p,m}$ is a universal hash family

Constructing Universal Hash Families

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

$$\mathcal{H}_{p,m} = \{ h_{a,b} \text{ for } a \in \mathbb{Z}_p^+, b \in \mathbb{Z}_p \}$$

Thm: $\mathcal{H}_{p,m}$ is a universal hash family

Lemma 1: If p is prime and $a \neq 0$ then there is a unique value $a^{-1} \in \{1, \ldots, p-1\}$ such that $a \cdot a^{-1} = 1 \bmod p$

(Division mod p is well defined)

Pf: Suppose that $az = az' \bmod p$ for $z, z' \in \mathbb{Z}_p^+$ $\Rightarrow$ $a(z - z') = 0 \bmod p$

$\Rightarrow$ $z - z'$ is divisible by p

$\Downarrow$

$2 - p \leq z - z' \leq p - 2$ $\Rightarrow$ $z - z' = 0$

$$\boxed{az = h \bmod p \text{ has } \underline{\text{at most}} \text{ one solution}}$$

Must be z s.t.

$$a \cdot z = 1 \bmod p$$

$$\boxed{az = 0 \bmod p \text{ has } \underline{\text{no}} \text{ solutions}}$$

Constructing Universal Hash Families

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

$$\mathcal{H}_{p,m} = \{ h_{a,b} \text{ for } a \in \mathbb{Z}_p^+, b \in \mathbb{Z}_p \}$$

<u>Thm</u>: $\mathcal{H}_{p,m}$ is a universal hash family

<u>Lemma 2</u>: If $x \neq y$ and $r \neq s$ then there is a unique solution (a,b) to the system

$$ax + b = r \bmod p$$
$$ay + b = s \bmod p$$

Constructing Universal Hash Families

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

$$\mathcal{H}_{p,m} = \{ h_{a,b} \text{ for } a \in \mathbb{Z}_p^+, b \in \mathbb{Z}_p \}$$

<u>Thm:</u> $\mathcal{H}_{p,m}$ is a universal hash family

<u>Proof:</u>

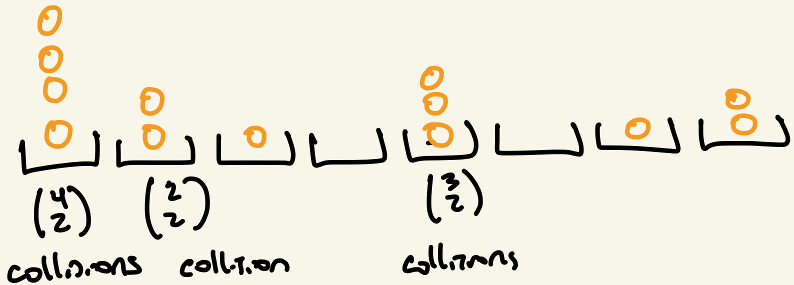
By Lemma 2 $\quad \mathbb{P}\left(ax+b = r \bmod p \text{ and } ay+b = s \bmod p \right) = \frac{1}{p(p-1)}$

$$\mathbb{P}(h_{a,b}(x) = h_{a,b}(y)) = \frac{N}{p(p-1)} \le \frac{p(p-1)}{p(p-1)} \cdot \frac{1}{m}$$

where N is number of $r \neq s \in \mathbb{Z}_p$ such that $r = s \bmod m$

$$N \le \underbrace{p}_{\text{choices of } r} \cdot \frac{\overbrace{p-1}^{\text{choices of } s \text{ for given } r}}{m}$$

Maximum Load for Universal Hashing



if max load $\gtrsim k \approx \sqrt{c}$ then $\#$of collisions $\gtrsim \binom{k}{2} = c$

$$\Longrightarrow \quad \text{max load} \approx \sqrt{\#\text{collisions}}$$

$$\mathbb{E}(\text{max load}) \approx \sqrt{\mathbb{E}(\#\text{collisions})} \approx \sqrt{n + \frac{n^2}{m}}$$

$$\mathbb{E}(\#\text{collisions}) = \sum_{x,y} \mathbb{P}(h(x)=h(y)) = \sum_{x} 1 + \sum_{x \neq y} \frac{1}{m} = n + \frac{n(n-1)}{m}$$

$$\approx n + \frac{n^2}{m}$$