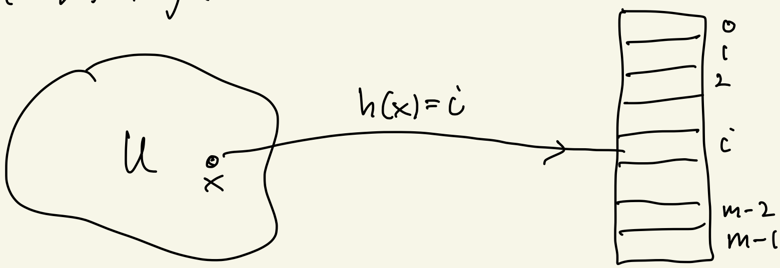
Universal hashing Cormen 11.3.3 + KT 13.6

Recall hashing idea



$|U| >> m$

Any fixed hash function $h: U \to [m]$ is vulnerable to attack by an adversary (can make many elements hash to same value

Solution: choose the hash function randomly and independently of the keys that we wish to hash $\longrightarrow$ Universal hashing

$\boxed{\text{Result: provably good performance on average for all inputs}}$

How do we obtain a good hash function?

Select a random hash function from a carefully designed class of hash functions

Let $\mathcal{H}$ be a finite collection of hash functions such that $h: \mathcal{U} \to [m]_0$ for each $h \in \mathcal{H}$

Then $\mathcal{H}$ is universal if the following holds:

Let $h \in \mathcal{H}$ be randomly chosen. Then
$\forall k, \ell \in \mathcal{U}$ with $k \neq \ell$ $\quad p(h(k) = h(\ell)) \leq \frac{1}{m}$

Recall hashing with chaining:

Let $h: \mathcal{U} \to [m]$ and keep for each $i \in [m]$ a linked list of all elements $x \in \mathcal{U}$ s.t. $h(x) = i$

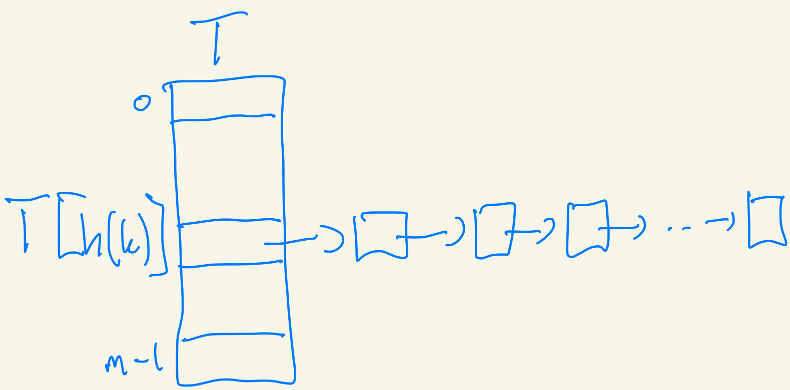
Theorem 11.3 Cormen

Suppose h is chosen randomly from a universal collection $\mathcal{H}$ of hash functions from U to $[m]$.
Assume we have used h to hash a set $S \subseteq U$ with $|S| = n$ using chaining to resolve collisions.
Let $T = T[0], T[1], \ldots T[m-1]$ be the table of linked list we obtain when $T[i]$ is a linked list containing those elements $x \in S$ for which $h(x) = i$
Then the following holds

- If $k \notin S$, then $E(n_{h(k)}) \leq \frac{n}{m} = \alpha$

  when $n_{h(k)}$ is the length of $T[h(k)]$

- If $k \in S$ then $E(n_{h(k)}) \leq \alpha + 1$



Proof:

Note that the expectation is over the choice of $h \in \mathcal{H}$ which is independent of the distribution of the keys in S

$\forall k, \ell \in \mathcal{U}$, $k \neq \ell$ define $X_{k\ell} = \begin{cases} 1 & \text{if } h(k) = h(\ell) \\ 0 & \text{if } h(k) \neq h(\ell) \end{cases}$

As $\mathcal{H}$ is universal $p(h(k) = h(\ell)) \leq \frac{1}{m}$

For fixed $k \in \mathcal{U}$ let $Y_k = \left| \{ \ell \in S \setminus \{k\} \mid h(k) = h(\ell) \} \right|$

so Y_k is the number of keys in $S \setminus \{k\}$ which have the same hash value as k and we have

$$Y_k = \sum_{\substack{\ell \neq k \\ \ell \in S}} X_{k\ell}$$

Now
$$E(Y_k) = E\Big(\sum_{\substack{\ell \neq k \\ \ell \in S}} X_{k\ell}\Big)$$
$$= \sum_{\substack{\ell \neq k \\ \ell \in S}} E(X_{k\ell})$$
$$\leq \sum_{\ell \neq k, \ell \in S} \frac{1}{m}$$

We saw

$$E(Y_k) \leq \sum_{\ell \neq k, \ell \in S} \frac{1}{m}$$

- If $k \notin S$ then $n_{h(k)} = Y_k$ and

  $|\{\ell \in S \mid \ell \neq k\}| = |S| = n$ so

$$E(n_{h(k)}) = E(Y_k) \leq \sum_{\ell \neq k, \ell \in S} \frac{1}{m} = \frac{|S|}{m} = \frac{n}{m} = \alpha$$

- If $k \in S$ then $n_{h(k)} = Y_k + 1$

  and $|\{\ell \in S \mid \ell \neq k\}| = |S| - 1 = n - 1$ so

$$E(n_{h(k)}) = 1 + E(Y_k) \leq 1 + \sum_{\ell \neq k, \ell \in S} \frac{1}{m} = 1 + \frac{n-1}{m} \leq 1 + \alpha$$

$\square$

Corollary 11.4

Using universal hashing + chaining starting from an empty table with m slots it takes expected time $O(n)$ to handle any sequence of INSERT, SEARCH and DELETE operations where $O(m)$ of them are INSERT

P: We insert $O(m)$ elements so $|S| \in O(m)$ implying that $\alpha = \frac{n}{m}$ is $O(1)$
Thus the expected length of each list in the table is $O(1)$ so each operation takes $O(1)$ expected time so $O(n)$ for all operations

How do we construct a universal class of hash functions?

Universal Hashing a la Cormen

- Choose prime $p \geq |U|$ and assume $U \subseteq \{0,1,2,\ldots p-1\}$

  $\mathbb{Z}_p = \{0,1,2,\ldots p-1\}$, $\mathbb{Z}_p^* = \{1,2,\ldots p-1\}$

- p prime $\Rightarrow$ we can solve equations modulo p

- $p \geq |U| > m$ so $p > m$

- For $a \in \mathbb{Z}_p^*$ and $b \in \mathbb{Z}_p$ define

  $h_{ab}(k) = ((ak+b) \bmod p) \bmod m$

  $h_{ab} : \mathbb{Z}_p \to \mathbb{Z}_m$

- Set $\mathcal{H} = \mathcal{H}_{pm} = \{ h_{ab} \mid a \in \mathbb{Z}_p^*, b \in \mathbb{Z}_p \}$

Theorem 11.5

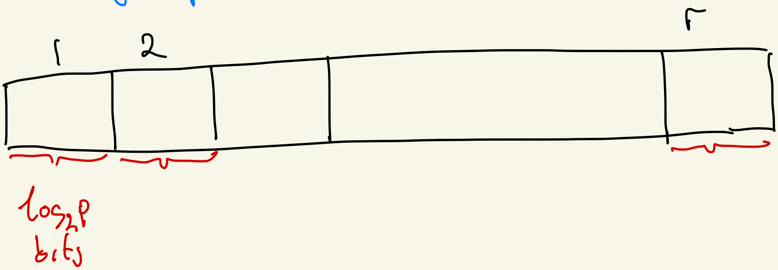
The class $\mathcal{H}_{pm}$ is universal

proof omitted!

Universal hashing a la Kleinberg & Tardos

Identify Universe $\mathcal{U}$ with vectors of the form $(x_1, x_2, \ldots, x_r)$ for some integer r when

$0 \le x_i < p$ for $i = 1, 2 \ldots r$



assume $\mathcal{U} \subseteq \{0, 1, 2, \ldots, N-1\}$

Assume that n is the size of the hash table

(in Cormen we used m as the size of the table)

and that $p \ge n$ p close to n. Then

$$r \simeq \frac{\log N}{\log p} \simeq \frac{\log N}{\log n}$$

let $A = \{(a_1, a_2, \ldots, a_r) \mid a_i \in \{0, 1, 2, \ldots p-1\} \; \forall i \in [r]\}$

for $a \in A$ let

$$h_a(x) = \left(\sum_{i=1}^{r} a_i x_i\right) \bmod p$$

$$\mathcal{H} = \{ h_a \mid a \in A \}$$

Theorem 13.25 $\mathcal{H}$ is a universal family

Proof let $x = (x_1, x_2, \ldots, x_r)$ and $y = (y_1, y_2, \ldots y_r)$ be distinct elements of $\mathcal{U}$.

Need to show that when $a = (a_1, a_2, \ldots a_r) \in A$ is randomly chosen, then $p(h_a(x) = h_a(y)) \leq \frac{1}{p}$

As $x \neq y$ there is a $j \in [r]$ such that

$$x_j \neq y_j$$

Consider the following way of choosing a random $a \in A$:
- first choose all a_i with $i \neq j$
- then choose a_j

We now prove that for every choice of the a_i's with $i \neq j$, the probability that the final choice of a_j will result in $h_a(x) = h_a(y)$ is exactly $\frac{1}{p}$

- Note that $h_a(x) = h_a(y)$

$$\Updownarrow$$

$$\sum_{q=1}^{r} a_q x_q = \sum_{q=1}^{r} a_q y_q \mod p$$

$$\Updownarrow$$

$$\sum_{q=1}^{r} a_q (x_q - y_q) = 0 \mod p$$

$$\Updownarrow$$

$$\sum_{q \neq j} a_q (x_q - y_q) + a_j (x_j - y_j) = 0 \mod p$$

$$\Updownarrow$$

$$\sum_{q \neq j} a_q (x_q - y_q) = a_j (y_j - x_j) \mod p$$

- after fixing a_i for $i \neq j$ we have

$$\sum_{q \neq j} a_q (x_q - y_q) = s \mod p \text{ for some } s \in \{0, 1, 2 \ldots p-1\}$$

Hence $h_a(x) = h_a(y)$ if and only if

$$(\square) \quad a_j(y_j - x_j) = s \mod p$$

as $z = y_j - x_j \neq 0$ since we assumed $x_j \neq y_j$

the equation $(\square)$ has a unique solution

$$a_j = s \cdot (y_j - x_j)^{-1} \mod p \qquad (*)$$

a_j receives a random value in $\{0, 1, 2, \dots p-1\}$ when constructing $a = (a_1, a_2, \dots a_r)$

Hence the probability that $(*)$ holds (and therefore $h_a(x) = h_a(y)$) is $\frac{1}{p}$

$\square$