# Universal Hash Functions

ESO207

Indian Institute of Technology, Kanpur

# Why randomized hashing?

- Given any hash function, an adversary can choose the keys to be hashed so that they all hash to the same slot.
- This would then require $\Theta(n)$ time for the SEARCH operation.
- Any fixed hash function would suffer from a $\Theta(n)$ worst-case time requirement for the SEARCH operation.
- Solution: Have a *family of hash functions* from which a hash function is chosen randomly.

# Randomized choice from family of hash functions

- Suppose hash function is chosen at random from some family of hash function.
- In this case, the adversary would choose a set of keys.
- But since the hash function is chosen randomly, there is a good chance that the chosen hash function would distribute the keys more uniformly than the worst-case choice.

# Universal Hashing: Definition

- Let $\mathcal{H}$ be a finite collection (family) of hash functions each of which map a given universe $U$ of keys into the range $\{0, 1, \ldots, m-1\}$.
- The collection $\mathcal{H}$ is said to be ***universal*** if for each pair of distinct keys $k, l \in U$,
  the number of hash functions $h \in H$ such that $h(k) = h(l)$ is at most $|\mathcal{H}|/m$.
- That is, for any $k, l \in U$ and distinct,

$$\Pr_{h \in \mathcal{H}} \{h(k) = h(l)\} \leq \frac{1}{m}$$

# Universal Hashing

- where, the notation $\Pr_{h \in \mathcal{H}} \{h(k) = h(l)\}$ means that the probability is taken over the random choices of the hash functions in $\mathcal{H}$.
- There is no other source of randomness. Once $h \in \mathcal{H}$ is chosen, the functions INSERT, SEARCH and DELETE all proceed deterministically.

# Universal Hashing: Property

- Let us now see why using a universal family of hash functions gives good average-case behaviour. Recall that $n_i$ denotes the length of the list $T[i]$, that is, the length of the chain at slot $i$.

# Property of Universal Hashing

Property: Suppose that a hash function $h$ is chosen uniformly at random from a universal family of hash functions $\mathcal{H}$ that map a universe $U$ of keys into $\{0, 1, \ldots, m-1\}$. Further, let $h$ be used as the hash function for hashing $n$ keys from $U$ to a hash table $T[0, \ldots, m-1]$ that uses open chaining to handle collisions.

1. If key $k$ is not in the table, then $\mathbb{E}\left[n_{h(k)}\right] \leq \alpha = n/m$.
2. If key $k$ is in the table, then $\mathbb{E}\left[n_{h(k)}\right] \leq 1 + \alpha$.

- Note that the expectation is taken over the choice of $h \in \mathcal{H}$. Let $K$ be the set of keys that are in the table $T$.

# Proof of Property

- For distinct keys $k, l \in K$ and $k \neq l$, define the indicator variable

$$X_{kl} = \begin{cases} 1 & \text{if } h(k) = h(l) \\ 0 & \text{otherwise.} \end{cases}$$

- By definition of universal hashing,

$$\Pr\{X_{kl} = 1\} = \Pr_{h \in \mathcal{H}}\{X_{kl} = 1\} = \Pr_{h \in \mathcal{H}}\{h(k) = h(l)\} \leq 1/m$$

- Hence,

$$\mathbb{E}[X_{kl}] = \Pr\{X_{kl} = 1\} \leq 1/m \ .$$

# Proof Contd.

- Let $I_k = 1$ if $k \in K$ and 0 otherwise. (that is $I_k$ is 1 if $k$ is hashed and is 0 otherwise. ) $I_k$ is a constant, it is not a random variable.
- Then,

$$n_{h(k)} = I_k + \sum_{\substack{l \in K \\ l \neq k}} X_{kl}$$

- Taking expectations, and using linearity of expectation, we have,

$$\begin{aligned} \mathbb{E}\left[n_{h(k)}\right] &= I_k + \sum_{\substack{l \in K \\ l \neq k}} \mathbb{E}\left[X_{kl}\right] \leq I_k + \sum_{\substack{l \in K \\ l \neq k}} \frac{1}{m} \\ &= I_k + \frac{n-1}{m} = \alpha + (I_k - 1/m) \end{aligned}$$

where, $\alpha = n/m$.

- $$\mathbb{E}\left[n_{h(k)}\right] = \alpha + (I_k - 1/m)$$
- Since, $I_k = 1$ if $k \in K$ and $I_k = 0$ otherwise, the statements of the theorem follows.

# Corollary

Using universal hashing and collision resolution by chaining in an initially empty table with $m$ slots, it takes expected time $\Theta(n)$ to handle any sequence of $n$ INSERT, SEARCH and DELETE operations containing $O(m)$ INSERT operations.

# Argument

- Since the number of insertions is $O(m)$, we have $n = O(m)$ and so $\alpha = O(1)$.
- The INSERT and DELETE operations take $O(1)$ time.
- By previous property, each SEARCH operation takes $O(1)$ expected time.
- By linearity of expectation, the expected time for the entire sequence of $n$ operations is $O(n)$.
- Since each operation takes $\Omega(1)$ time, the $\Theta(n)$ bound follows.

# A Universal Hash Family

- Let $U$ be the finite universe of the keys that we will assume is the set $\{0, 1, 2, \ldots, |U| - 1\}$.
- Let $p$ be a prime number such that $p \geq |U|$.
- For $a \in \{1, 2, \ldots, p - 1\}$ (called $\mathbb{Z}_p^*$) and $b \in \{0, 1, \ldots, p - 1\}$ (called $\mathbb{Z}_p$), define hash function

$$h_{a,b}(k) = ((ak + b) \mod p) \mod m \ .$$

- For each value of $a, b$ $h_{a,b} : U \to \{0, 1, \ldots, m - 1\}$. Define the family (collection) of hash functions

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

# Universal Hash family

- $$h_{a,b}(k) = ((ak + b) \mod p) \mod m$$
- $\mathcal{H}_{p,m} = \{h_{a,b} \mid a \in \mathbb{Z}_p^*, b \in \mathbb{Z}_p\}$.
- In algebra, $\mathbb{Z}_p^* = \{1, 2, \ldots, p-1\}$ is referred to as the *multiplicative group modulo prime p* and $\mathbb{Z}_p = \{0, 1, \ldots, p-1\}$ as the *prime field* of size $p$.
- Since there are $p-1$ choices for $a$ and $p$ choices for $b$, the collection $\mathcal{H}$ has $p(p-1)$ functions.

**Theorem:**
The class $\mathcal{H}_{pm}$ of hash functions is universal.

# Proof of universality

- First consider the simple case when $m$ is equal to $p$. (Recall $\mathbb{Z}_p = \{0, \dots, p-1\}$, $\mathbb{Z}_p^* = \{1, 2, \dots, p-1\}$.
- Now, hash functions have the simpler form

$$h_{a,b}(k) = ak + b \mod p$$

- Following definition, we have to calculate the number of hash functions $h_{a,b}$ from $\mathcal{H}$ such that

$$h_{a,b}(k) = h_{a,b}(l)$$

  for any $k, l \in U = \mathbb{Z}_p$, $k \neq l$.
- This is equivalent to

$$ak + b = al + b \mod p$$

# Proof of Universality

- $ak + b = al + b \mod p$ gives by transposing,

$$a(k - l) = 0 \mod p$$

- Note: transposition is valid, since,
  - if $ak + b = al + b \mod p$, then, $p$ divides $(ak + b - (al + b))$ or that $p$ divides $ak - al$.
  - Hence, $p$ divides $a(k - l)$, or, $a(k - l) = 0 \mod p$.

# Proof of universality

- $h_{a,b}(k) = h_{a,b}(l)$ is equivalent to $a(k-l) = 0 \mod p$.
- $p$ divides $a(k-l)$. So $p$ being prime divides either $a$ or $k-l$.
- $a \in \{1, 2, \ldots, p-1\}$ and so $p$ does not divide $a$.
- $k, l \in \{0, \ldots, p-1\}$ and are distinct. So, $p$ does not divide $k-l$.
- So $a(k-l) = 0 \mod p$ has no solution.

# Proof: part I

- The number of hash functions from $\mathcal{H}$ such that $h_{a,b}(k) = h_{a,b}(l)$, for $k \neq l$ is 0.
- This of course satisfies the property of universal hashing, since universality needs that for any $k \neq l$,

$$\left|\{(a,b) \mid h_{a,b}(k) = h_{a,b}(l)\}\right| \leq \frac{|\mathcal{H}|}{p} \quad .$$

# Proof of Universality: General Case

- Let us now consider the general case when $m < p$. The hash functions are of the form

$$h_{a,b}(k) = (ak + b \mod p) \mod m$$
$$a \in \{1, \ldots, p-1\}, b \in \{0, 1, \ldots, p-1\}$$

- Let

$$r = ak + b \mod p$$
$$s = al + b \mod p$$

  Then, $r \neq s$, by the argument above.

- We would like to know the number of solutions to $h(k) = h(l)$, that is, the number of solutions to

$$(r - s) \mod m = 0 \ .$$

# Proof: general case

- Fix $r$. The solutions to $(r - s) \mod m = 0$, except for $r = s$ (which is disallowed) are

$$s = r + m, r + 2m, \ldots, r + \left\lfloor \frac{(p - r)}{m} \right\rfloor m$$

and

$$r - m, r - 2m, \ldots, r - \left\lfloor \frac{r}{m} \right\rfloor m \ .$$

- Thus the number of solutions to $r = s \mod m$ is at most

$$\frac{p}{m} - 1 = \frac{p - m}{m}$$

for each fixed $r$.

# Proof of Universality

- There are $p$ possible choices for $r$, namely, $r = 0, 1, 2, \ldots, p-1$.
- Hence, the number of hash functions for which $h(k) = h(l)$ is the number of pairs $(r, s)$ such that $r = s \mod m$, which is at most

$$\frac{p(p-m)}{m} \leq \frac{p(p-1)}{m} = \frac{|\mathcal{H}|}{m} \ .$$

Thus the family is universal.