Hashing (Cont. from Slides)

Universal Hash functions:

Due to Carter & Wegman (1979)

Defn: family $H$ maps $U \rightarrow [m]$ is universal if for any $x \neq y \in U$

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

$h \in H$
Construction:

\[ |U| = 2^u \]
\[ |M| = 2^m \]

Let \( A \in \text{random binary entries}^{m \times u} \)

For any \( x \in U \) (\( u \)-length binary vector)

\[ h(x) := A x \mod 2 \]

Q: How many hash functions in the family? \( 2^m \)
Thm: \( \ldots \) is universal

Proof: \( h(x) = h(y) \) for \( x \neq y \)

\[ A x = A y \]

\[ A(x - y) = 0 \]

\[ A z = 0 \quad \text{for} \quad z \neq 0 \]

\[ \Rightarrow x \neq y \]

We want to show \( P(Az = 0) \leq \frac{1}{M} \) for any \( z \neq 0 \)

Let \( z_i x \neq 0 \quad \exists i \quad \text{such that} \quad z \neq 0 \)

\[ A z = \sum A_j z_j \]

\[ \uparrow \text{column of } A \]
\[ A \mathbf{2} = 0 \]
\[ \sum A_j z_j = 0 \]
\[ A_{i,x} = -\sum_{j \neq i}^m A_j z_j \]

fixed vector of size \( m \)

\( m \) length binary vector (random)

Prob. of above: \[ \left(\frac{1}{2}\right)^m = \frac{1}{2^m} = \frac{1}{m} \]
Application: Hash table

Handling Collisions:
Approach 1: closed addressing
aka separate chaining.

Look up time \( \leq \text{length of the list} \leq \text{number of collisions} \)
\[ C_x = \text{num. of elements mapped to the same value where } x \text{ is mapped to.} \]

\[ L_x = \text{len. of linked list containing } x \]

\[ L_x = C_x + 1 \]

Q: What is \( E[L_x] \)?

\[ E[L_x] = 1 + E[C_x] = 1 + \frac{(N-1)}{M} \]

If we choose \( M \geq N \)

\[ E[L_x] \leq 2 \]

Look up time constant in expectation.
\[ C = \text{total number of collisions} \]

1. \( E[c] \) ?
   \[ \leq \binom{N}{2} \frac{1}{M} \]

Suppose \( M \geq N^2 \) \( \implies E[c] \leq \frac{1}{2} \)

Prob. [there exists a collision] = ?

\[ \frac{1}{2} \]

Constant time look up (even in worst case).

But we need \( M \geq N^2 \)

Can we get this with \( O(N) \)