ALG 4.2

Universal Hash Functions:

<table>
<thead>
<tr>
<th>Hash Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f : A \to B$</td>
</tr>
<tr>
<td>keys</td>
</tr>
<tr>
<td>indices</td>
</tr>
</tbody>
</table>

$f$ has *conflict* at $x,y \in A$ if $x \neq y$ but $f(x) = f(y)$

$\sigma_f(x,y) = \begin{cases} 1 & \text{if } x \neq y \text{ and } f(x) = f(y) \\ 0 & \text{else} \end{cases}$

Auxiliary Reading Selections:
- AHU-Data Section 4.7
- BB Section 8.4.4
If $H$ is a set of hash functions,

$$\sigma_H(x, y) = \sum_{f \in H} \sigma_f(x, y)$$

for set of keys $S$,

$$\sigma_H(x, S) = \sum_{f \in H} \sum_{y \in S} \sigma_f(x, y)$$

Total Conflicts \(\geq b \left[ \left( \frac{a}{b} \right) \left( \frac{a}{b} - 1 \right) \right] \geq \frac{a^2}{b} - a\)
H is a *universal* set of hash functions

\[ \sigma_H(x, y) \leq \frac{|H|}{|B|} \]

for all \( x, y \in A \)

i.e. no pair of keys \( x, y \) are mapped into the same index by more than \( \frac{1}{|B|} \) of all functions in \( H \)

**Proposition 1**

Given any set \( H \) of hash functions, there exist \( x, y \in A \) such that

\[ \sigma_H(x, y) > |H| \left( \frac{1}{|B|} - \frac{1}{|A|} \right) \]

**Proof**

Let \( a = |A|, b = |B| \)

By counting, we can show

\[ \sigma_f(A, A) \geq b\left(\frac{a}{b} - 1\right)^2 \geq \frac{a^2}{b} - a \]
Thus
\[ \sigma_H (A, A) \geq a^2 |H| \left( \frac{1}{b} - \frac{1}{a} \right) \]

By the pigeon hole principle
\[ \exists x, y \in A \text{ s.t. } \sigma_H (x, y) \geq |H| \left( \frac{1}{b} - \frac{1}{a} \right) \]

**note**
in most applications, \(|A| >> |B|\)
and then any universal 2 class has asymptotically a minimum number of conflicts

**Proposition 2:** Let \( x \in A, S \subseteq A \)

For \( f \) chosen randomly from a universal 2 class \( H \) of hash functions, the expected number of collisions is
\[ \sigma_f (x, S) \leq \frac{|S|}{|B|} \]

**proof**
\[ E(\sigma_f (x, S)) = \frac{1}{|H|} \sum_{f \in H} \sigma_f (x, S) \]
\[ = \frac{1}{|H|} \sum_{y \in S} \sigma_H (x, y) \text{ by definition} \]
\[ \leq \frac{1}{|H|} \sum_{y \in S} \frac{|H|}{|B|} \text{ by definition of universal}_2 \]
\[ = \frac{|S|}{|B|} \]
application

associative memory storage of \(|S|\) keys onto \(|B|\) linked lists.

Given key \(x \in A\), store \(x\) in list \(f(x)\)

Proposition 2 implies each list has expected length \(\leq \frac{|S|}{|B|} = O(1)\) if \(|B| \geq |S|\)

Gives \(0(1)\) time for STORE, RETRIEVE, and DELETE operations

Proposition 3

Let \(R\) be a sequence of requests with \(k\) insertion operations into an associative memory.

If \(f\) is chosen at random from set of universal class \(H\), the expected total cost of all \(k\) searches is

\[ \leq |R| \left(1 + \frac{k}{|B|}\right). \]

proof

There are \(|R|\) total search ops, and each takes by Proposition 2 expected time \(\leq 1 + \frac{k}{|B|}\).

note

if \(|B| \geq k\), then expected total time is \(O(|R|)\).
**Bounds on distribution of** $\sigma_f(x,S)$

**Proposition 4** Let $x \in A$, $S \subseteq A$

Let $\mu =$ expected value of $\sigma_f(x,S)$

For $f$ chosen randomly from universal \(H\),

$$\text{Prob} \left( \sigma_f(x,S) > t \cdot \mu \right) < \frac{1}{t}$$

**proof**

Immediate from Markov bound

Improved bounds on probability:

$$\text{prob} \leq \frac{11}{t^4} \text{ for universal hash fns. } H_2, H_3$$

(using 2nd and 4th moments of prob. distribution.)

$H =$ universal set of hash functions.

$E_1 =$ Expected cost of *random* set of $k$ requests using a *worst case* function $f$ in $H$ *(random input)*

$E_2 =$ Expected cost of *worst case* set of $k$ requests using a *random* function $f$ in $H$ *(randomized algorithm)*
**Prop 5** \( E_1 \geq (1 - \epsilon) E_2 \) where \( \epsilon = \frac{|B|}{|A|} \)

**proof**

Let \( a = |A|, b = |B| \).

Prop 2 implies \( E_2 \leq 1 + \frac{|S|}{b} \)

Suppose \( S \) is chosen randomly. For \( x, y \in S \),

\[
E(\sigma_f(x, y)) = \frac{1}{a} \sigma_f(A, A)
\]

\[
\geq \frac{1}{2} \left[ a^2 \left( \frac{1}{b} - \frac{1}{a} \right) \right] \text{ by Prop 1}
\]

\[
\geq \left( \frac{1}{b} - \frac{1}{a} \right)
\]

So \( E_1 \geq 1 + E(\sigma_f(x, S)) \)

\[
\geq 1 + |S| \left( \frac{1}{b} - \frac{1}{a} \right)
\]

---

**Example of Universal Class**

**Set of Keys** Table

Let \( A = \{0, 1, \ldots, a-1\} \) Set of Keys

B = \( \{0, 1, \ldots, b-1\} \) Table

Let \( p \) be a prime \( \geq a \)

\( Z_p = \{0, 1, \ldots, p-1\} = \text{number field mod } p \)

**Define** \( g : Z_p \rightarrow B \) s.t.

\( g(x) = x \mod b \)

**Define** for \( n, m \in Z_p \) with \( m \neq 0 \),

\( h_{n,m} : A \rightarrow Z_p \) with \( h_{n,m}(x) = (mx+n) \mod p \)

**Define** \( f_{n,m} : A \rightarrow B \) s.t.

\( f_{n,m}(x) = g(h_{m,n}(x)) \)

\( H_1 = \{ f_{m,n} \mid m, n \in Z_p, m \neq 0 \} \)

**Claim:** \( H_1 \) is universal 2
Lemma

for distinct \( x, y \in A \),

\[
\sigma_{H_1}(x,y) = \sigma_g(Z_p, Z_p)
\]

proof

\[
\sigma_g(Z_p, Z_p) = |\{(r, s) \mid r, s \in Z_p, r \neq s, g(r) = g(s)\}|
\]

Observe that the linear equations:

\[
\begin{align*}
xm + n &= r \pmod{p} \\
ym + n &= s \pmod{p}
\end{align*}
\]

have unique solutions in \( Z_p \)

So \((r, s) = (h_{m,n}(x), h_{m,n}(y))\) then

\[
( f_{m,n}(x) = f_{m,n}(y) \text{ if and only if } g(r) = g(s))
\]

\(\sigma_{H_1}(x,y)\) is the number of such pairs in

\[
(r, s) \in \sigma_g(Z_p, Z_p)
\]

Theorem

\( H_1 \) is universal

proof

Let \( n_i = |\{ t \in Z_p \mid g(t) = i\}| \)

By definition of \( g(x) = x \mod b \),

\[
\Rightarrow n_i \leq \frac{p-1}{b} + 1
\]

For any given \( r \), the number of \( s \) where \( s \neq r \) and \( g(r) = g(s) \) is

\[
\sigma_g(r, Z_p) \leq \frac{p-1}{b}.
\]

But there are \( p \) choices of \( r \),

so \( p \cdot \left( \frac{(p-1)}{b} \right) \geq \sigma_g(Z_p, Z_p) \)

\[
= \sigma_{H_1}(x,y) \text{ by Lemma}
\]

(Also note \( \sigma_{H_1}(x,x) = 0 \))

Hence \( \sigma_{H_1}(x,y) \leq \frac{|H_1|}{b} \) since \( |H_1| = p(p-1) \)

so \( H_1 \) is universal"
Universal Hash Fns on Long keys

Given class of hash functions $H$, define hash functions $J = \{ h_{f,g} | f, g \in H \}$
where $h_{f,g}(x_1, x_2) = f(x_1) \oplus g(x_2)$

exclusive or

Theorem Suppose $B = \{0, 1, \ldots, b = 1\}$ where $b$ is a power of 2. Suppose this class of fns $A \rightarrow B$

$\exists$ real $r \forall i \in B \forall x_1, y_1 \in A, x_1 \neq y_1$

$\Rightarrow |\{ f \in H | f(x_1) \oplus f(y_1) = i \}| \leq r|H|$

Then $\forall x, y \in (A \times A), x \neq y$

$|\{ h \in J | h(x) \oplus h(y) = i \}| \leq r|H|$

Proof for $x = (x_1, x_2), y = (y_1, y_2) \in A \times A$

$i \in B$ then $|\{ h \in J | h(x) \oplus h(y) = i \}|$

$= |\{ f, g \in H | f(x_1) \oplus g(x_2) \oplus f(y_1) \oplus g(y_2) = i \}|$

$= \sum_{y \in H} |\{ f \in H | f(x_1) \oplus f(y_1) = i \oplus g(x_2) \oplus g(y_2) \}|$

$\leq |\{ f \in H | f(x_1) \oplus f(y_1) = i \}| \leq r|H|$

example $H_1$ with $m = 0$ gives $J$ with $r = \frac{1}{|B|}$ universal!
Universal 2 Hashing \textit{with out} Multiplication

A = set of d digit numbers base \( \alpha \) so, \(|A| = \alpha^d\)

B = set of binary numbers length j

M = arrays of length d \( \cdot \alpha \), with elements in B

\( \forall m \in M \) let \( m(k) \) = kth element of array \( m \)

\( \forall x \in A \) let \( x_k \) = kth digit of \( x \) base \( \alpha \)

\textbf{definition} \( f_m(x) = m(x_1+1) \oplus m(x_1+x_2+2) \oplus \ldots \oplus m\left(\sum_{k=1}^{d} x_k+k\right) \)

\textbf{Theorem}

\( H_2 = \{ f_m \mid m \in M \} \) is universal \( 2 \)

\textbf{proof} for \( x, y \in A \),

let \( f_m(x) = r_1 \oplus r_2 \oplus \ldots \oplus r_s \) rows of \( m \)

\( f_m(y) = r_{s+1} \oplus \ldots \oplus r_t \)

Then \( f_m(x) = f_m(y) \) iff \( r_1 \oplus \ldots \oplus r_t = \overline{0} \)

But if \( x \neq y \Rightarrow \exists k \text{ s.t. } r_k \text{ in only one of } f_m(x), f_m(y) \)

so \( f_m(x) = f_m(y) \) iff \( r_k = \bigoplus_{i \neq k} r_i \)

But there are only \(|B|\) possibilities for row \( r_k \)

so \( x, y \) will collide for \( \frac{1}{|B|} \) of fns \( f_m \in H_2 \)

\textbf{Hence } \( H_2 \) \text{ is universal } 2 \)
Analysis of Hashing
for Uniform Random Hash fn

load factor $\alpha = \frac{\# \text{ of keys hashed}}{\# \text{ of indices in Hash Table}}$

Hashing with Chaining
keep list of conflicts at each index

Conflicts

length is binomial variable

expected length = $\alpha$

Expected Time Cost per hash = $O(1 + \alpha)$

By Chernoff Bounds, with high likelyhood
time cost per hash $\leq O(\alpha \log(\# \text{ keys}))$
Open Address Hashing
(With Uniform Random Hash fn)

Resolve conflicts by applying another hash function

\[ \alpha = \text{load factor} = \text{prob. of occupied hash address} \]

# rehashes as geometric variable

expected hash time = \( \frac{1}{1-\alpha} = 1 + \alpha + \alpha^2 + \ldots \)