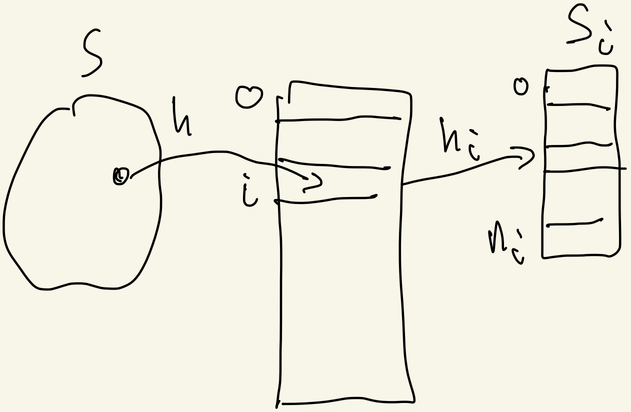
Perfect hashing Cormen 11.5

Goal: achieve excellent worst case behavior when the set of keys is static: once in the table the set of keys never change (ex CD / DVD)

Want:

- O(1) memory access in worst case when searching
- small memory use!

Solution: use two levels of hashing where both use universal hashing



then $x, y \in S$ with $h(x) = h(y) = i$ are hashed to S_i using h_i

1. at level one we use a carefully chosen hash function $h \in \mathcal{H}$ where $\mathcal{H}$ is universal

2. Instead of using a linked list for collisions elements at slot j $(h(x)=j)$ we use a secondary hashtable S_j together with associated hash function h_j to avoid collisions at level 2. The size of S_j will be n_j^2 when $n_j = |\{x \in S \mid h(x)=j\}|$

- at level 1 we use $h \in \mathcal{H}_{pm}$ when $p > |S|$ $(S \subseteq \{0,1,2,\ldots p-1\})$
- keys with $h(x)=j$ are rehashed into second table S_j of size m_j using some $h_j \in \mathcal{H}_{pm_j}$

- We first show how to ensure 0 collisions at level 2
- Then we show that the expected memory use is $O(n)$ $\quad n = |S|$

Theorem 11.9 Suppose we store n distinct keys in a hash table of size $m = n^2$ using a random $h \in \mathcal{H}$, where $\mathcal{H}$ is universal. Then the probability that there are no collisions is at least $1/2$

P: There are $\binom{n}{2}$ possible collisions

let $Z_{k\ell} = \begin{cases} 1 & \text{if } h(k) = h(\ell) \\ 0 & \text{else} \end{cases}$

$P(Z_{k\ell} = 1) \leq \frac{1}{m} = \frac{1}{n^2}$ as h is universal

when $k \neq \ell$

Then $Z = \sum_{\substack{k,\ell \in S \\ k \neq \ell}} Z_{k\ell}$ is the # collisions

$$E(Z) = E\Big(\sum_{\substack{k,\ell\in S\\ k\neq \ell}} Z_{k\ell}\Big) = \sum_{\substack{k,\ell\in S\\ k\neq \ell}} E(Z_{k\ell}) \le \sum_{\substack{k,\ell\in S\\ k\neq \ell}} \frac{1}{n^2} = \frac{\binom{n}{2}}{n^2} < \frac{1}{2}$$

By Markov's inequality we get

$$p(Z \ge 1) \le \frac{E(Z)}{1} = E(Z) < \frac{1}{2}$$

So $p(Z=0) \ge \frac{1}{2}$ as claimed $\square$

By repeating the choice of h' until there are no collisions wrt that choice, we can obtain a collision free hashing of the n keys.

The expected # of repetitions (choosing $h' \in \mathcal{H}$) is at most $\frac{1}{p(Z=0)} \le \frac{1}{1/2} = 2$

Problem: If n is large a table of size n^2 is too large

Solution: use this idea only at the second level

- At level 1 we use a table of size $m=n$ only
- Let $h \in \mathcal{H}$ be the hash function we use at level 1
- Let $n_j = |\{x \mid h(x)=j\}|$ and let S_j, $j \in [m]_0$ be a table with n_j^2 entries and h_j a collision free hash function mapping $\{x \mid h(x)=j\}$ to S_j

At level 1, when we have taken $m=n$, we use $O(n)$ space to store:

- the primary hash table
- the numbers $m_j = n_j^2$ $j \in [m]_0$
- $a_j \in \mathbb{Z}_p^{\times}$, $b_j \in \mathbb{Z}_p$ which define the second level hashfunction h_j to be used on $\{x \mid h(x)=j\}$ according to the definition of $h_{ab}(x)$ in Cormen.

Theorem 11.10

Suppose we store n keys in a hash table of size $m = n$ using universal hashing and let n_j, $j \in \{0, 1, 2, \ldots m-1\}$ be the number of keys hashed to j ($h(x) = j$).

Then $E\left(\sum_{j=0}^{m-1} n_j^2\right) < 2n$ (n_j is a random variable depending on choice of h)

Proof: Recall that $\forall a \in \mathbb{Z}^+$ $\underbrace{a + 2\binom{a}{2}} = a + a(a-1) = \underbrace{a^2}$

$$E\left(\sum_{j=0}^{m-1} n_j^2\right) = E\left(\sum_{j=0}^{m-1} n_j + 2\binom{n_j}{2}\right)$$

$$= E\left(\sum_{j=0}^{m-1} n_j\right) + 2E\left(\sum_{j=0}^{m-1} \binom{n_j}{2}\right)$$

$$= E(n) + 2E(r)$$

$$= n + 2E(r)$$

when r is the total # collisions when we use $h \in \mathcal{H}$

By the universal hashing property $E(r) \le \binom{n}{2} \cdot \frac{1}{m} = \binom{n}{2} \cdot \frac{1}{n} = \frac{n-1}{2}$

Hence $E\left(\sum_{j=0}^{m-1} n_j^2\right) \le n + 2 \cdot \frac{n-1}{2} < 2n$ $\square$.

Corollary 11.1.1

With the hashing scheme chosen (level 1: $m=n$, level 2: $m_j = n_j^2$, $j \in \{0,1,\dots n-1\}$)
The expected total storage for the secondary hash tables is less than $2n$

Proof

$$E\left(\sum_{j=0}^{m-1} m_j\right) = E\left(\sum_{j=0}^{m-1} n_j^2\right) < 2n$$ by theorem 11.10 □.

Corollary 11.12

Using a hashing scheme as above the probability that we need more than $4n$ total storage for second level tables is less than $\frac{1}{2}$

Proof By Markov's inequality

$$P\left(\sum_{j=0}^{m-1} m_j \geq 4n\right) \leq \frac{E\left(\sum_{j=0}^{m-1} m_j\right)}{4n} < \frac{2n}{4n} = \frac{1}{2}$$

Conclusion:

Using a few trials to find a good $h \in \mathcal{H}$ when $\mathcal{H}$ is universal we can quickly obtain a scheme
(h at level 1, $h_1, h_2, \dots, h_{m-1}$ at level 2)
which uses a reasonable amount of space