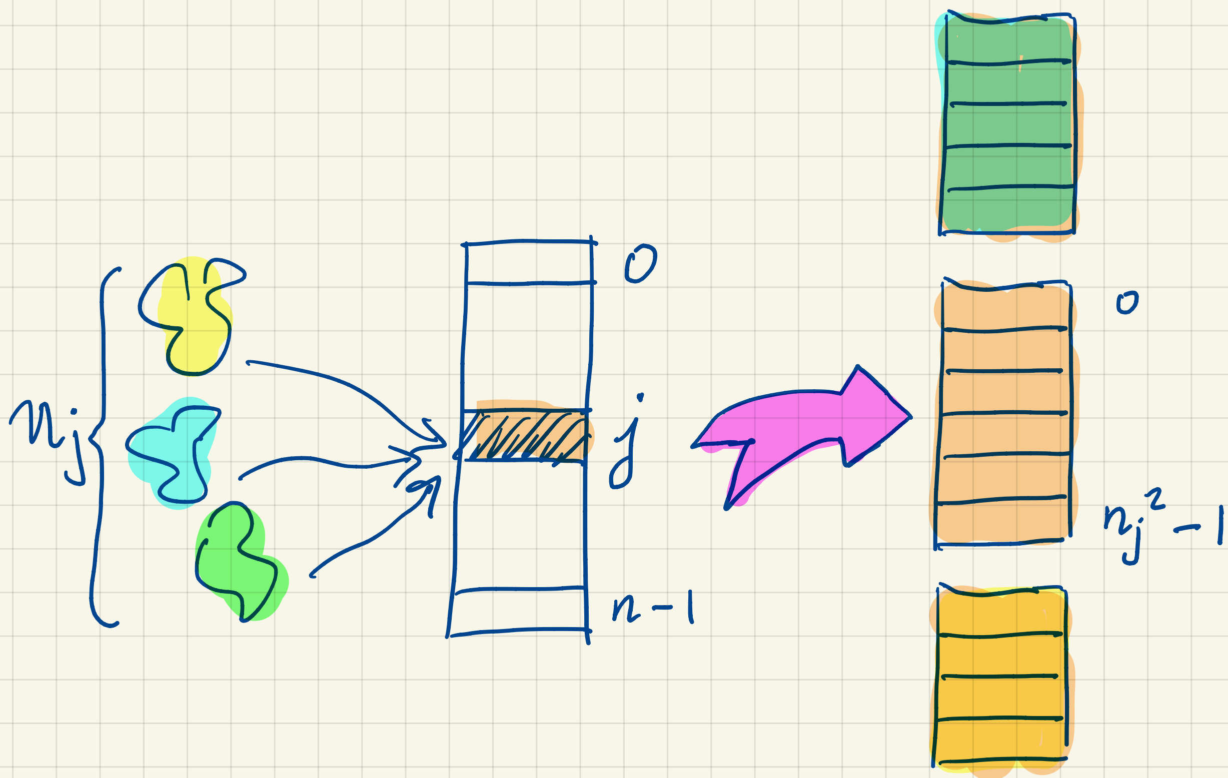
Perfect Hashing

- Static set of n keys
- statis table $\Theta(n)$ size
- no collisions
- deterministic $\Theta(1)$ search time
- $O(n)$ Expected time to construct.

Given n keys, choose $m = n$ and a hash function h from a universal $\mathcal{H}$.

If n_j keys hash to slot j, use for slot j a secondary hash table with n_j^2 entries.



Storage: $E\left[\sum_{j=0}^{n-1} n_j^2\right]$

similar to Bucket sort since

$$n_j^2 = (X_{1j} + X_{2j} + \cdots X_{nj})(X_{1j} + X_{2j} + \cdots + X_{nj})$$

we get

$$E[n_j^2] = \sum_i E[X_{ij}^2] + \sum_{i \neq k} E[X_{ij} X_{kj}]$$

But $E[X_{ij}^2] = P(X_{ij} = 1) = ?$ $(\neq \frac{1}{m})$

and $E[X_{ij} X_{kj}] \neq E[X_{ij}] E[X_{kj}]$ (not independent)

$E[X_{ij} X_{kj}] = P(X_{ij} = X_{kj} = 1) = ?$ $(\neq \frac{1}{m^2})$

But we need to sum over all j anyway

$$E\left[\sum_{j=0}^{n-1} n_j^2\right] = \sum_{j=0}^{n-1}\sum_{i} P(X_{ij}=1) + \sum_{j=0}^{n-1}\sum_{i\neq k} P(X_{ij}=X_{kj}=1)$$

$$= \underbrace{\sum_{i}\underbrace{\sum_{j=0}^{n-1} P(X_{ij}=1)}_{1}}_{n} + \underbrace{\sum_{i\neq k}\underbrace{\sum_{j=0}^{n-1} P(X_{ij}=X_{kj}=1)}_{P(h(key(i))=h(key(k)))\ \leq \frac{1}{m}}}_{\leq n(n-1)/m}$$

$$\leq n + \frac{n(n-1)}{m} = 2n-1 < 2n \quad (m=n)$$

Collisions in secondary table

If we also use a universal hash function, the prob. that any pair collides is $\leq \frac{1}{n_j^2}$

$$E[\#\text{collisions}] \leq \binom{n_j}{2} \frac{1}{n_j^2} < \frac{1}{2}$$

$$\frac{n_j(n_j-1)}{2} \times \frac{1}{n_j^2}$$

Markov Inequality

$$P(X \geq t) \leq \frac{E[x]}{t}$$

Applied to storage: $P(\text{storage} \geq 4n) \leq \frac{2n}{4n} = \frac{1}{2}$

So with few trials of the first hash function we can find a storage $\Theta(n)$

Applied to collisions: $P(\text{collision} \geq 1) \leq \frac{1/2}{1} = \frac{1}{2}$

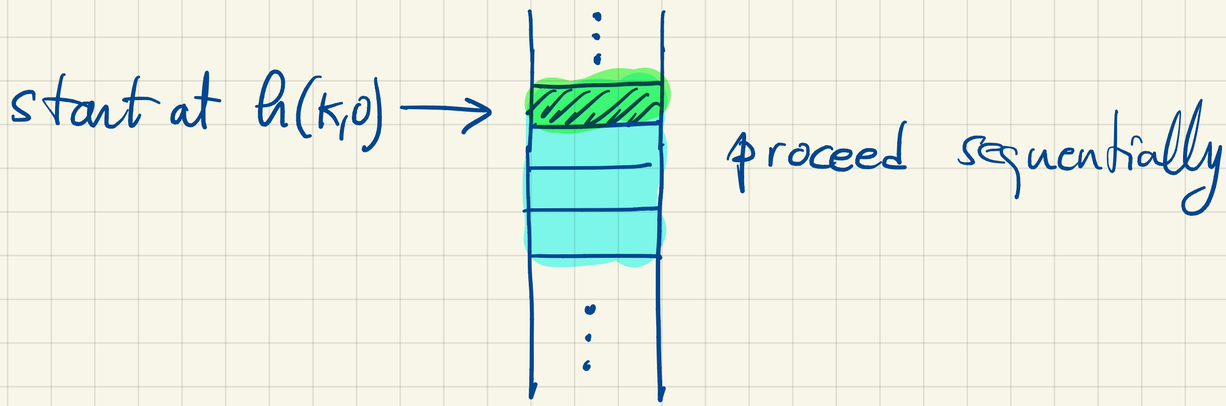
so with few trials of the secondary hash function for slot j, we can find a collision free table

Open addressing

All keys are in the table

<u>Idea:</u> generate $h(k,0), h(k,1), \ldots, h(k,m-1)$ which is a permutation of $(0,1,\ldots,m-1)$ until an empty slot is found.

<u>Linear probing:</u> $h(k,i) = (h'(k)+i) \bmod m$



long runs of occupied slots build up.

Quadratic probing: $h(k,i) = \left(h'(k) + c_1 i + c_2 i^2\right) \bmod m$

- Better clustering behavior
- As with linear, the initial probe determines entire sequence, so only $\Theta(m)$ probe sequences

Double hashing:

$$h(k,i) = [h_1(k) + i h_2(k)] \bmod m$$

- $h_2(k)$ must be relatively prime to m to cover entire table.

  e.g. let m be prime, $h_1(k) = k \bmod m$
  and $h_2(k) = 1 + (k \bmod m-1) \in \{1, \dots, m-1\}$
- We have $\Theta(m^2)$ possible probe sequences.

For analysis, assume each permutation is equally likely given the key. Since $n \leq m$, assume $\alpha < 1$.

Expected # probes until unsuccessful search: (empty slot)

Let $X_i = \begin{cases} 1, & \text{probes } 0 \ldots i \text{ do not find empty slot.} \\ 0 & \text{otherwise} \end{cases}$

$$\# \text{probes} = 1 + \sum_{i=0}^{n-1} X_i$$

$$P(X_i = 1) = \frac{n}{m} \cdot \frac{n-1}{m-1} \cdot \ldots \cdot \frac{n-i}{m-i} \leq \alpha^{i+1}$$

$$n \leq m$$
$$-im \leq -in$$
$$mn - im \leq mn - in$$
$$m(n-i) \leq n(m-i)$$
$$\frac{n-i}{m-i} \leq \frac{n}{m} = \alpha$$

$$E[\#\text{probes}] \leq 1 + \alpha + \alpha^2 + \ldots + \alpha^n \leq \sum_{i=0}^{\infty} \alpha^i = \frac{1}{1-\alpha} \qquad \left(\alpha = 0.9 \Rightarrow \frac{1}{0.1} = 10\right)$$

[Same for insert, search until empty]

What about successful search?

Assume each element is equally likely to be searched for

$$\frac{1}{n}\sum_{i=0}^{n-1}\underbrace{\frac{1}{1-i/m}}_{\text{Expect. \# probes to insert } (i+1)^{st} \text{ elem}} \qquad \left(\text{Before inserting } (i+1)^{st} \text{ elem}, \ \alpha = \frac{i}{m}\right)$$

$$= \frac{m}{n}\sum_{i=0}^{n-1}\frac{1}{m-i} = \frac{m}{n}\left[\frac{1}{m}+\frac{1}{m-1}+\cdots+\frac{1}{m-n+1}\right]$$

$$\approx \frac{m}{n}\left[\ln m - \ln(m-n)\right] = \frac{m}{n}\ln\frac{m}{m-n}$$

$$= \frac{1}{\alpha}\ln\frac{1}{1-\alpha} \qquad \left(\alpha=0.9 \Rightarrow \frac{10}{9}\ln\frac{1}{0.1} < 3\right)$$