Hashing:

We briefly encountered a similar problem when dealing with balls and bins.

If n balls were placed in n bins uniformly at random, no bin contains more than $\frac{3\ln(n)}{\ln\ln(n)}$ balls w.p $\geq 1-\frac{1}{n}$.

Number of empty bins was concentrated around its expect-ation of $\frac{n}{e}$.

1. Space + time to be $O(n)$
2. Additions ~ $O(1)$ time
3. Query ~ $O(1)$ time.

Perfect hashing:

Given a set S of n keys from the universe U, build a lookup table of size $O(n)$ s.t a membership query can be answered in $O(1)$ time.

we call it "perfect hashing for S".

Defn: A set of hash functions H is called a weak universal family if for all $x, y \in U$, $x \neq y$. $H: U \to [m]$.

$$\Pr_{h \in H}[h(x) = h(y)] = \frac{O(1)}{m}.$$

Fix an element x.

Expected chain length: $\mathbb{E}_{h \sim H}\left[\# \text{ of } y \underset{x \neq y}{\text{ s.t }} h(x) = h(y)\right] + 1$

$$= 1 + \sum_{y} \Pr_{h \sim H}[h(x) = h(y)]$$

$$= 1 + (n-1) \cdot O\left(\frac{1}{m}\right) \quad \leq 1 + \frac{1}{m} \cdot (n-1)$$

We want $1 + \frac{n-1}{m} = O(1)$.

m and n. $m \simeq O(n)$.

Fredman-Komlos-Szemeredi hashing:

[2001].

- → Expected $O(n)$ time
- → Worst case $O(n)$ space
- → $O(1)$ worst case query time.

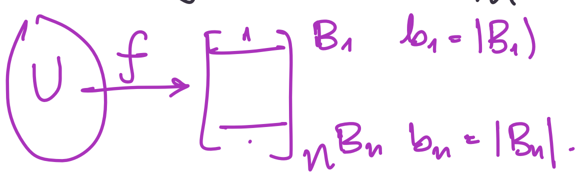
Suppose $m = \Omega(n^2)$; then if we pick h randomly from H

then
$$\mathbb{E}[\text{no. of collisions}] = \sum_{\substack{x,y \in S \\ x \neq y}} \Pr[h(x) = h(y)]$$
$$= \binom{n}{2} \cdot \frac{O(1)}{m}$$
$$\leq O(1).$$

$\Rightarrow$ $O(1)$ trials needed to obtain collision free hashing.

Suppose $m \sim n$ then

$$\mathbb{E}[\text{no. of collisions}] = \binom{n}{2} \cdot \frac{O(1)}{m} = O(n).$$



$m \sim n$

$U = \{0, \dots, |U|-1\}$

FKS gave a 2-step hashing algo.

- Find a hashing fn $f : U \to [n]$ that partitions S into buckets $B_1, \dots, B_n$.

  $|B_i| = b_i \Rightarrow \sum b_i = n.$

← Obtain this from weak univ class of hash functions.

- For each bucket B_i : find a function $g_i : U \to [O(b_i^2)]$

Say $|B_i| = b_i$.

For each element $x \in S$, let C_x be the no. of its collisions

$$\checkmark \ \mathbb{E}_h[C_x] = \frac{|S|-1}{n} = \frac{n-1}{n} < 1. \ \checkmark$$

$$= \sum_{y \neq x} 1 \cdot \Pr_h[y \neq x : h(x) = h(y)] \leq \frac{n-1}{n} < 1$$

1. Select a random function from H.
2. Compute a hash table with chaining, so insertion takes $O(1)$ time
3. Compute an auxillary array B_2 s.t $|B_2(i)| = O(b_i^2)$
4. If $\sum_{i=1}^{n} b_i^2 > \beta n$ then go to step 1. Else "record" f.

$\sum b_i^2 \leq \beta n$.

Let t be the no. of iterations. We want to bound the expected no. of iterations.

Claim: If $\beta \geq 4$, $\mathbb{E}[t] \leq 2$.

Proof: Total no. of collisions (C_s) is as follows

$$C_s = \sum_{i=1} \left| \{(x,y) \mid x, y \in B_i , x \neq y\} \right|$$

$$= \sum_{i=1}^{n} b_i (b_i - 1) \qquad \sum_{i=1}^{n} b_i = n.$$

$$= \sum_{i=1}^{n} (b_i^2 - b_i) = \left(\sum_{i=1}^{n} b_i^2 \right) - n$$

≤ 1

$$\mathbb{E}_h[C_s] = \mathbb{E}_h\left[\sum_{i=1}^{n} b_i^2\right] - n \qquad \mathbb{E}[C_s] = \sum_{x \in S} \mathbb{E}[C_x]$$

We know that $\mathbb{E}[C_s] < n$. $\qquad \leq \sum_{x\in S} 1 = n$.

$\Rightarrow \mathbb{E}\left[\sum b_i^2\right] < 2n$.

Using Markov's ineq: $\Pr\left[\sum_{i=1}^{n} b_i^2 > 4n\right] < \frac{1}{2}$.

$b_i \rightsquigarrow O(b_i^2)$. $\qquad$ w.p $\geq \frac{1}{2}$ $\quad \underbrace{\sum b_i^2 \leq \overset{\beta}{4n}}$.

- Select $g_i : U \to [\alpha b_i^2]$ from H.
- If for some $x_i \in B_i$; there is a collision, pick a new g_i.

Claim: If $\alpha \geq 2$ then $\mathbb{E}[t_i] = O(b_i^2)$.

Proof: C_x = no. of collisions of x in B_i

$$\mathbb{E}_{g_i}[C_x] < \frac{b_i}{\alpha b_i^2} = \frac{1}{\alpha b_i}. \qquad \Pr_{g_i \sim H}\left[g_i(x) = g_i(y)\right] \leq \frac{1}{\alpha b_i^2} \quad (x \neq y)$$

From Markov's ineq.

$$\Pr_{g_i}[C_x \geq 1] \leq \mathbb{E}_{g_i}[C_x] < \frac{1}{\alpha b_i}.$$

B_i

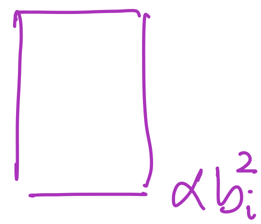
Over all elements in B_i

$$\Pr_{g_i}[\exists x \text{ s.t } C_x \geq 1] \leq \sum_x \Pr_{g_i}[C_x \geq 1] < b_i \cdot \frac{1}{\alpha b_i} = \frac{1}{\alpha}.$$

If $\alpha = 2$ then w.p $\geq \frac{1}{2}$, no collisions happen.

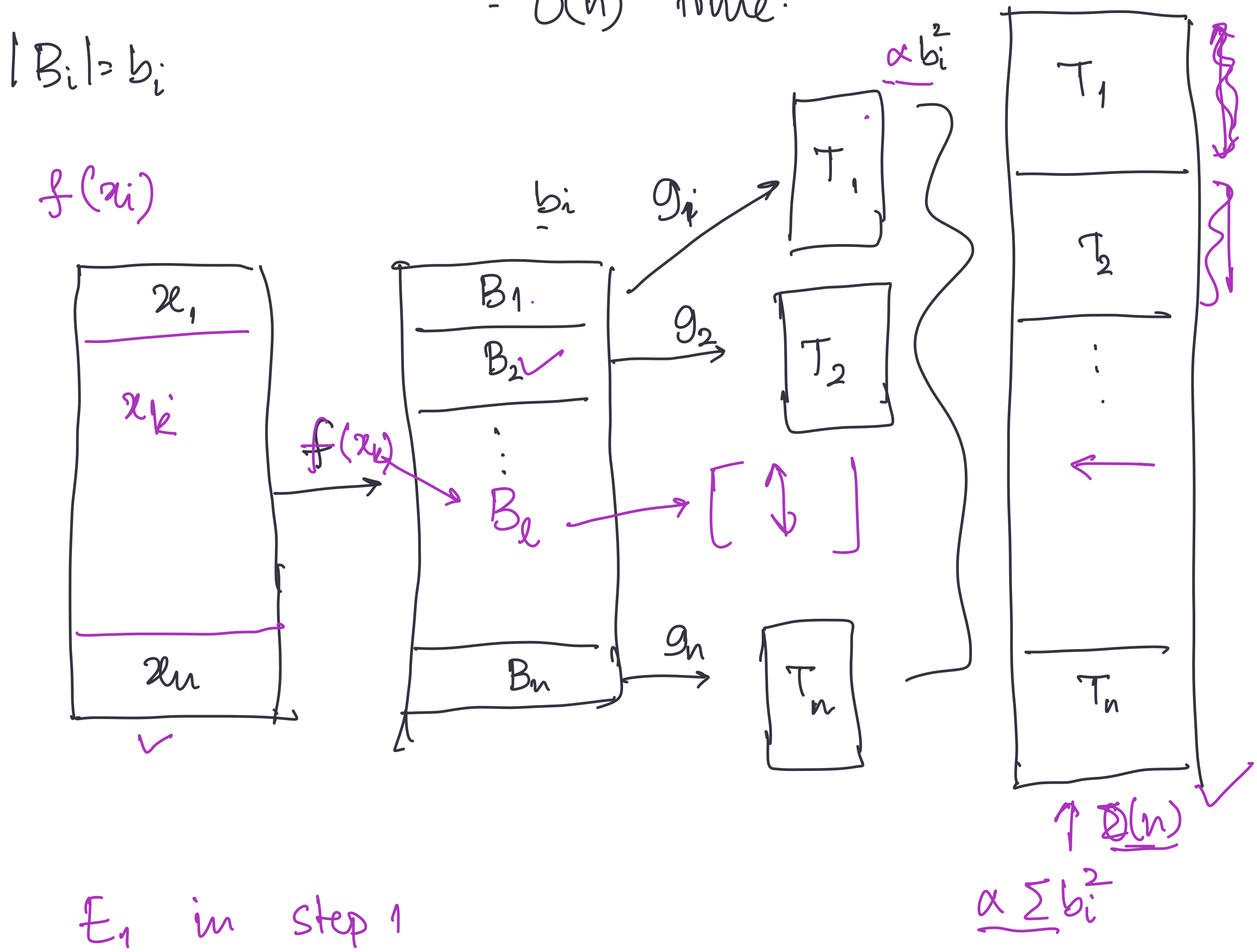
$\Rightarrow$ At most 2 trials to get g_i w/ no collisions.

$\Rightarrow$ Expected time $= O(b_i^2)$ for this step



Total expected running time: $O(n) + \sum_i O(b_i^2) + \underbrace{O(n)}_{\text{preprocessing}}$

$= O(n)$ time.



E_1 in step 1

E_2 in step 2.

Total error $\leq E_1 + E_2 < 1$.