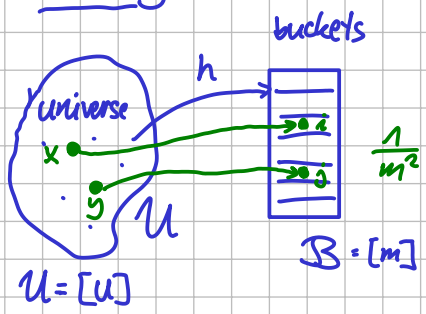
Hashing



choose $h \in \mathcal{H}$ randomly

For hashing with buckets

$\mathbb{E}_{h \in \mathcal{H}}[\text{time complexity}] \in O\left(\frac{n}{m}\right)$

Def: A family $\mathcal{H}$ of functions from $\mathcal{U}$ to $\mathcal{B}$ is **c-universal** for $c > 0$ iff

$\forall x, y \in \mathcal{U},\ x \neq y: \Pr_{h \in \mathcal{H}}[h(x) = h(y)] \leq \frac{c}{m}$.

Def: A family $\mathcal{H}$ of functions from $\mathcal{U}$ to $\mathcal{B}$ is **(k,c)-independent** for $k \in \mathbb{N}^+$, $c > 0$, $|\mathcal{U}| \geq k$ iff

$\forall x_1 \dots x_k \in \mathcal{U}$ all distinct $\forall b_1 \dots b_k \in \mathcal{B}$

$\Pr_{h \in \mathcal{H}}[h(x_1) = b_1 \wedge \dots \wedge h(x_k) = b_k] \leq \frac{c}{m^k}$.

Def: $\mathcal{H}$ is k-independent iff $\exists c > 0$: $\mathcal{H}$ is (k,c)-indep.

- $\mathcal{H}$ is (k,c)-ind. $\Rightarrow$ $(k-1,c)$-indep
- $\mathcal{H}$ is $(2,c)$-ind. $\Rightarrow$ c-universal
- $\mathcal{H}$ is $(1,c)$-ind.

Consider: $\mathcal{C} := \{h_a \mid a \in \mathcal{B}\}$, $h_a(x) := a$ for all x. Then $\mathcal{C}$ is $(1,1)$-indep.

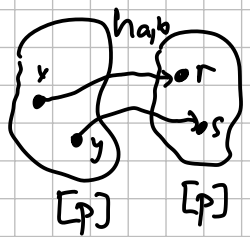
Def: For p prime and $m \leq p$ ($\mathcal{U} = [p]$):

$\mathcal{L} := \{h_{a,b} \mid a, b \in [p]\}$ (linear functions)

$h_{a,b}(x) := ((ax+b) \bmod p) \bmod m$ (calculating in a field $\mathbb{Z}_p$)

Def: $\mathcal{L}' := \{h_{a,b} \mid a, b \in [p], a \neq 0\}$

Theorem: $\mathcal{L}$ is 2-universal. ($\mathcal{L}'$ is 1-universal)



Proof: Let $x, y \in [p]$, $x \neq y$.

First without mod m:

- params $(a,b) \in [p]^2$
- $r := (ax+b) \bmod p$, $s := (ay+b) \bmod p$, $r \neq s$
- $(a,b) \overset{1\text{-}1}{\longleftrightarrow} (r,s)$ — pick (a,b) u.a.r. $\Leftrightarrow$ pick (r,s) u.a.r.

$p(p-1)$ pairs; $r \neq s$: $p(p-1)$ pairs

With mod m:

$\Pr_{a,b}[h_{a,b}(x) = h_{a,b}(y)]$

$= \Pr_{a,b}[((ax+b) \bmod p) \bmod m = ((ay+b) \bmod p) \bmod m]$ (r, s)

$= \Pr_{r,s}[r \equiv s \pmod m] \leq \frac{2}{m}$

Def: (r,s) is a **bad pair** $\Leftrightarrow$ $r \equiv s \pmod m$

Count bad pairs -- for a given r

$\#s: (r,s)$ is BP $\leq \lceil \frac{p}{m} \rceil \leq \frac{p+m-1}{m} \leq \frac{2p}{m}$

Sum over all r: $\#BP \leq \frac{2p^2}{m}$

$\#s \leq \lceil \frac{p}{m} \rceil - 1 \leq \frac{p+m-1}{m} - 1 = \frac{p-1}{m}$

$\#BP \leq \frac{p \cdot (p-1)}{m}$

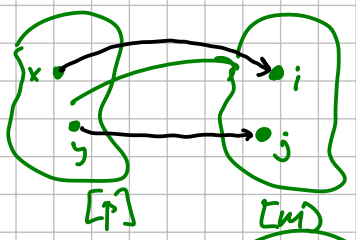
$\#$ all pairs $= p^2 \Rightarrow \Pr[\text{a pair is bad}] = \frac{\#BP}{p^2} \leq \frac{2}{m}$

$\Pr[\text{pair is bad}] \leq \frac{p \cdot (p-1)}{m} \Big/ p(p-1) = \frac{1}{m}$.

Theorem: $\mathcal{L}$ is $(2,4)$-independent.

we want this to be $\leq \frac{4}{m^2}$

Proof:



$$Pr_{r,s}\left[\underbrace{r \equiv i \pmod m}_{\text{independent events}} \;\&\; \underbrace{s \equiv j \pmod m}_{\text{independent events}}\right]$$

$$= \underbrace{Pr_{r,s}[r \equiv i \pmod m]}_{\leq 2/m} \cdot \underbrace{Pr_{r,s}[s \equiv j \pmod m]}_{\leq 2/m} \leq \frac{4}{m^2}$$

in each of $\lceil \frac{p}{m} \rceil$ intervals, there is at most one such r

$$\frac{\lceil p/m \rceil}{p} \leq \frac{p + m - 1}{m \cdot p} \leq \frac{2p}{mp} = \frac{2}{m}.$$

($p + m - 1 \leq 2p$)

Construction of k-indep. families for arbitrary k:

Df: For $U = [p]$ and $k \geq 1$:

$\mathcal{H}_k := \{h_{\vec{a}} \mid \vec{a} \in [p]^k\}$ family of functions from $[p]$ to $[p]$

reduce mod m using general lemma later

$h_{\vec{a}}(x) := \sum_{i=0}^{k-1} a_i \cdot x^i$ ← all this mod p

Theorem: $\mathcal{H}_k$ is $(k,1)$-independent

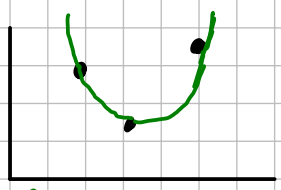
proof sketch:

$\forall i \; h_{\vec{a}}(x_i) = b_i$

↓ unique poly of deg $< k$

↓ unique choice of $\vec{a}$

Given $x_1 \ldots x_k$ items and $b_1 \ldots b_k$ buckets



for k points in the plane $\exists!$ polynomial of deg $< k$ connecting them

$$Pr[\forall i \; h_{\vec{a}}(x_i) = b_i] = \frac{1}{p^k}$$

Good news: this works for arbitrary k

Bad news: time is $\Theta(k)$

Multiply-shift hashing

$a \cdot x$ (a odd)

32-bit machine

$a*x >> (32-\ell)$

$(a*x)\langle 32-\ell : 32\rangle$

Df: $x\langle i:j\rangle$:= bits i to $j-1$ of x

Df: For given w (word size) and ℓ (result size) — in bits

$\mathcal{M} := \{h_a \mid a \in [2^w], a \text{ odd}\}$ from $[2^w] \to [2^\ell]$ ($U \to B$)

$h_a(x) := (ax)\langle w-\ell : w\rangle$

$x\langle 0:1\rangle$ = least-sig. bit = $x \bmod 2$

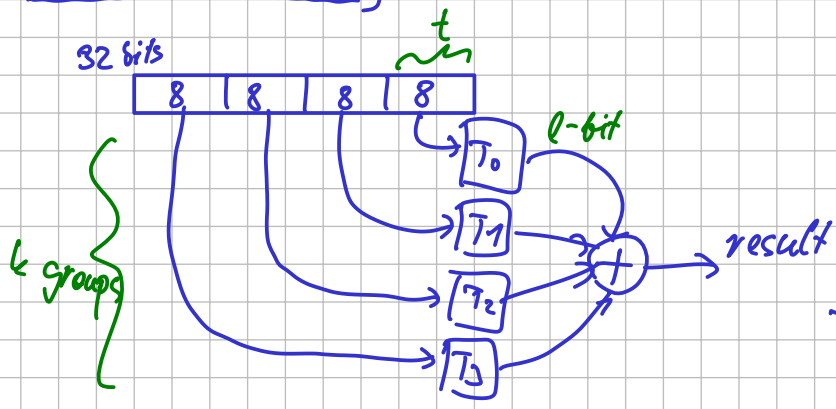
$x\langle 0:8\rangle$ = lowest 8 bits..

$$x\langle i:j\rangle = \lfloor x/2^i \rfloor \bmod 2^{j-i}$$

Claim: $\mathcal{M}$ is 2-universal.

→ ∃ version of mult-shift which is 2-indep.

Tabulation hashing



k groups per t bits

$[2^{kt}] \longrightarrow [2^\ell]$

tables $T_0, \dots, T_{k-1}$

each $T_i : [2^t] \to [2^\ell]$

Then

$$h_{T_0 \dots T_{k-1}}(x) := \bigoplus_{0 \le i < k} T_i\left[x\langle it : (i+1)t\rangle\right]$$

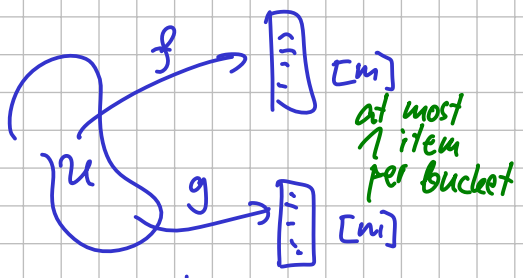
Time for evaluating h is $\Theta(k)$.

Memory needed: $\Theta(k \cdot 2^t \cdot \ell)$

Claim: 3-independent, but not 4-independent.

Cuckoo hashing

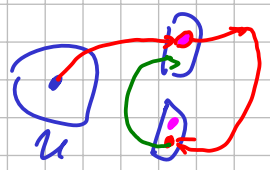
① 2 tables, 2 functions



x is stored either at $B_1[f(x)]$ or at $B_2[g(x)]$

Find is $O(1)$ in worst case!

Insert:

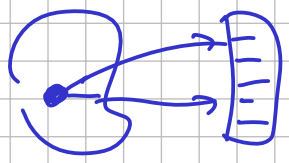


keep kicking out items & moving to the other location until

the location is empty

timeout ↓ rehash with new f, g

② 1 table, 2 functions



Theorem: Let $\varepsilon > 0$ be a fixed constant,

$m \ge (2+\varepsilon) \cdot n$,

insertion timeout $\lceil 6 \cdot \log m \rceil$

f, g uniformly at random from $\lceil 6 \log m \rceil$-indep. family.

Then $\mathbb{E}_{f,g}[\text{time for insert}] \in O(1)$.

depends on ε

known:

6-indep. is insufficient

But tabulation works!