Hešovací II

Systém $\mathcal{H}$ fcí $h: U \to [m]$ je **c-univerzální** pro $c \geq 0$, pokud pro lib. $x \neq y$

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

$\mathcal{H}$ je **(k,c)-nezávislý** kde $k \geq 1$, $c \geq 0$, pokud pro lib. $x_1 \dots x_k$ různé, lib. $a_1 \dots a_k$

$$\Pr_{h \in \mathcal{H}}[h(x_1) = a_1 \wedge \dots \wedge h(x_k) = a_k] \leq \frac{c}{m^k}$$

$\mathcal{H}$ je **k-nezávislý**, pokud je (k,c)-nez pro nějaké c. (nezávislé na m)

☺: (k,c)-nez. $\Rightarrow$ $(k-1,c)$-nez. $(k \geq 1)$

$$\Pr[h(x_1) = a_1 \wedge \dots \wedge h(x_{k-1}) = a_{k-1}] = \sum_{i=0}^{m-1} \Pr(h(x_1) = a_1 \wedge \dots \wedge h(x_k) = i]$$

$(2,c)$-nez. $\Rightarrow$ c-univ.

$$\Pr[h(x) = h(y)] = \sum_{i=0}^{m-1} \Pr[h(x) = i \wedge h(y) = i]$$

$(1,c)$-nez. nemusí být c-univ: $\mathcal{H} = \{h_i \mid i \in [m]\}$, $h_i(x) = i$

$\Pr[h(x) = i] = 1/m$, $\Pr[h(x) = h(y)] = 1$

konstrukce univ. systémů

$h_t(x) = \sum_{i=0}^{k-1} t_i x^i$ pro $k \geq 1$, $t \in \mathbb{Z}_p^k$, p prvočíslo $h_t : \mathbb{Z}_p \to \mathbb{Z}_p$ (polynom)

Lemma: $\mathcal{H} = \{ h_t \mid t \in \mathbb{Z}_p^k \}$ je $(k,1)$-univ. pro lib. $k \geq 1$, prvočíslo p.

Důk: $x_1 \dots x_k$ různé, $a_1 \dots a_k \in \mathbb{Z}_p$ (interpolace)

existuje právě jeden polynom nad $\mathbb{Z}_p$ stupně $\leq k-1$ t.ž. $h_t(x_i) = a_i \; \forall i$

$$\Pr_t [h(x_i) = a_i \;\; \forall i] = \frac{1}{p^k}$$ □

Speciálně, $\mathcal{H}_2 = \{ ax+b \mid a,b \in \mathbb{Z}_p \}$ je $(2,1)$-univ.

Lemma (sklížení mod m): Nechť $\mathcal{H}$ je (k,c)-univ. $h : U \to [r]$, $m < r$. Pak

$\mathcal{H} \bmod m = \{ h \bmod m \mid h \in \mathcal{H} \}$ je $(k, 2^k c)$-univ. a $2c$-univ.

$\Downarrow$

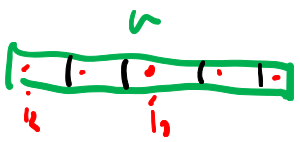
$4c$-univ.

Př: $\mathcal{H}' = \{ h_{a,b} \mid a,b \in \mathbb{Z}_p \} = \mathcal{H}_2 \bmod m$ je $(2,4)$-univ. a 2-univ.

$h_{a,b}(x) = ((ax+b) \bmod p) \bmod m$

Dk: $x_1 \neq x_2$, $h = h' \bmod m$, $h' \in \mathcal{H}$

a) 2c-univ.: $P_r[h(x_1)=h(x_2)] = \sum_{i_1 \equiv i_2} P_r[h'(x_1)=i_1 \wedge h'(x_2)=i_2] \leq \sum_{i_1 \equiv i_2} \frac{c}{r^2}$



$\leq r \lceil r/m \rceil \frac{c}{r^2} \leq \frac{2c}{m}$

$\lceil r/m \rceil \leq \frac{r+m-1}{m} \leq \frac{2m}{m}$

b) (2, 4c)-univ.: $P_r[h(x_1)=j_1 \wedge h(x_2)=j_2] = \sum_{i_1 \equiv j_1,\ i_2 \equiv j_2} P_r[h'(x_1)=i_1 \wedge h'(x_2)=i_2]$

$j_1, j_2 \in [m]$

$\leq \sum_{i_1 \equiv j_1,\ i_2 \equiv j_2} \frac{c}{r^{2k}} \leq \lceil r/m \rceil^{2k} \frac{c}{r^{2k}} \leq \left(\frac{2m}{m}\right)^{2k} \frac{c}{r^{2k}} = \frac{2^k 4c}{m^{2k}}$ □

Lemma: Nechť $\mathcal{H}$ je (k,c)-univ. fce $h: U \to [r]$, $r \geq 2km$. Pak $\mathcal{H} \bmod m$ je $(k, 2c)$-univ. sjednocení.

Dk, stejně co na

$P_r[h(x_i)=j_i \ \forall i] \leq \lceil r/m \rceil^k \frac{c}{r^k} \leq \left(\frac{r+m-1}{r}\right)^k \frac{c}{m^k} \leq \frac{c}{m^k}\left(1+\frac{m}{r}\right)^k \leq \frac{c}{m^k} e^{\frac{km}{r}}$

$(1+x) \leq e^x$ (Taylor)

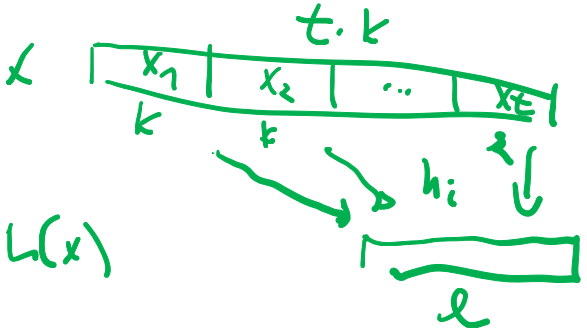
$\leq \frac{c}{m^k} e^{1/2} \leq \frac{2c}{m^k}$ □

$\Rightarrow$ $p_k \bmod m$ je $(k,2)$-nez. systém pro lib. prvočíslo $p \geq 2^{l+m}$, $k \geq 1$

$\{ h_t \bmod m \mid t \in \mathbb{Z}_p^k \}$ $\quad (h_t \bmod m)(x) = \left(\sum_{i=0}^{k-1} t_i x^i \bmod p \right) \bmod m$

Tabulační hešování

idea: pro malé univerzum lze mít zcela náh. fci (k bitů)



$h : [2^{t \cdot k}] \to [2^l]$

$h(x) = \bigoplus_{i=1}^{t} h_i(x_i)$ $\quad$ XOR

kde $h_i : [2^k] \to [2^l]$ jsou zcela náh. fce

- t tabulek s 2^k řádky po l bitech $\Rightarrow$ celkem $t \cdot 2^k \cdot l$ bitů vs $2^{t \cdot k} \cdot l$
- velmi rychlé (implementace), vhodné pro cache
- pro $k=1$ $\quad h(x) = xA \oplus b$ kde A náhodná bin. matice $(t \times l)$
  b náh. bin. vektor dél. l.

Tvrzení: Tabulační hešování (parametrizované tabulkami) tvoří 3-nez. systém ale není 4-nez. (tzl)

Dk: x_1, y_1, z různé $\Rightarrow$ pro něj i b.ú.n.o $x_i \neq y_i$ a $x_i \neq z_i$

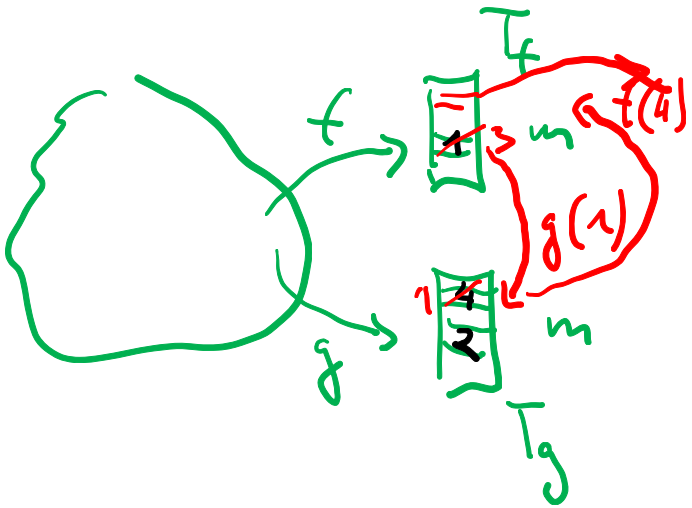
$h_i(x_i)$ nez. $h_i(y_i), h_i(z_i) \Rightarrow h(x)$ nez. na $h(y)$, $h(z)$

$(x_1, x_2), (x_1, \bar{x}_2), (\bar{x}_1, x_2), (\bar{x}_1, \bar{x}_2)$

$h(x_1, x_2) \oplus h(x_1, \bar{x}_2) \oplus h(\bar{x}_1, x_2) \oplus h(\bar{x}_1, \bar{x}_2) = 0 \Rightarrow$ není 4-nez.

$h_1(x_1) \oplus h_2(x_2)$ ☐

Kukačkovské hešování:



2 tabulky T_f, T_g, 2 h.fa f, g, konflikty řešíme přemísťováním mezi tabulkami dokud nenajdeme volné místo nebo přetrváme timeout ~ log n pak vše přehešovat s novými fcemi f, g.

Invariant: každý prvek x je buď $T_f[f(x)]$ nebo $T_g[g(x)]$.

⇒ FIND, DELETE v $O(1)$ nejhorším čase.

Věta: Pro $m \geq (2+\varepsilon)n$ a f, g náhodné (rovnoměrné) z $\lceil c \log n \rceil$-nez. systému. Pak při kukaččím hešování s limitem $\lceil c \log n \rceil$ má INSERT $O(1)$ očekávaný čas (amortizovaně).

(bez důkazu)

Pozn: Ví se, že 6-nez. nestačí pro $O(1)$ očekávaný čas. Ale: tabelační hešování stačí!