HEŠOVÁNÍ (hashing) (1)

Máme:
- universum $\mathcal{U}$ možných prvků, typicky $\mathcal{U} = [U] := \{0, \dots, U-1\}$
- množinu přihrádek $\mathcal{P} = [m]$
- hešovací funkci $h : \mathcal{U} \to \mathcal{P}$
- reprezentovanou množinu $X \subseteq \mathcal{U}$, $|X| = n$
- v ~~každé~~ i-té přihrádce množinu $P_i := \{x \in X \mid h(x) = i\}$ ... kolize nastane, pokud $|P_i| > 1$.
  reprezentovanou jako seznam

Nepřátelské množiny: Pro libovolnou pevnou funkci h lze najít X: $h \restriction X = \text{const.}$ ($\exists i: P_i = X$)
... stačí, aby $U \geq nm$. $\Rightarrow$ operace stojí $\Omega(n)$ w.c.

Raději zvolíme h náhodně ... pak umíme zaručit dobrou očekávanou složitost.

Cíl: Nastavíme $m \approx n$, v každé přihrádce $\approx$ const. prvků $\Rightarrow$ složitost $\approx$ const.

Možnosti:

(1) Úplně náhodná funkce ... nepraktické (potřebujeme $\Omega(U \log m)$ bitů paměti!)
Ale chová se dobře: $\forall x, y \in \mathcal{U}, x \neq y$: $\Pr_h[h(x) = h(y)] = 1/m$.

(2) Df: Systém funkcí $\mathcal{H}$ z $\mathcal{U}$ do $[m]$ je c-univerzální pro $c \in \mathbb{R}$, pokud
$\forall x, y \in \mathcal{U}, x \neq y$: $\Pr_{h \in \mathcal{H}}[h(x) = h(y)] \leq c/m$.

Chceme: Funkce z $\mathcal{H}$ popsatelná parametry tak, aby šlo
a) efektivně náhodně vybrat $h \in \mathcal{H}$
b) rychle $h(x)$ vyhodnotit (předp. $O(1)$)

Věta: Nechť $\mathcal{H}$ je c-univ. z $\mathcal{U}$ do $[m]$, $X \in \binom{\mathcal{U}}{n}$, $y \in \mathcal{U} \setminus X$.
Potom $\mathbb{E}_{h \in \mathcal{H}}[\underbrace{\#x \in X : h(x) = h(y)}_{A}] \leq \frac{cn}{m}$.

$\Rightarrow$ očekávaná složitost neúspěšného hledání je $O(\frac{n}{m}+1)$
pro úspěšné není horší
(čas úspěš. hledání = čas Insertu = čas neúspěš. hl. v době Insertu)

Dk: Zavedeme indikátory: $A_i = \begin{cases} 0 \\ 1 & \text{pokud } h(x_i) = h(y) \end{cases}$
očíslujeme $X = \{x_1, \dots, x_n\}$

$$\mathbb{E}[A] = \mathbb{E}[\textstyle\sum_i A_i] = \sum_i \underbrace{\mathbb{E}[A_i]}_{= \Pr[A_i = 1] \leq c/m \text{ (z c-univerzality)}} \leq n \cdot c/m = c\frac{n}{m}.$$

Df: Nechť $p \geq U$ je prvočíslo. Pak $\mathcal{L} := \{h_{a,b} \mid a, b \in [p], a \neq 0\}$, kde $h_{a,b}(x) := ((ax+b) \bmod p) \bmod m$.

Věta: Systém $\mathcal{L}$ je 1-univerzální.

Dk: Nechť $x, y \in [U]$, $x \neq y$.
Nejprve analyzujme funkce tvaru $(ax+b) \bmod p$ ... to je výpočet v $\mathbb{Z}_p$.
- Pro dané $(a,b) \in [p]^2$ zavedeme: $r = (ax+b) \bmod p$, $s = (ay+b) \bmod p$
  zatím bez omezení $a \neq 0$
  soustava 2 nezávislých lin. rovnic v tělese $\Rightarrow \exists!$ řešení (a,b) pro každé (r,s)
- Máme tedy bijekci mezi všemi (a,b) a (r,s) (obojí $\in [p]^2$)
  rovnoměrně náhodný výběr jednoho dává r.n. výběr druhého
- Požadavek $a \neq 0$ je ekvivalentní $r \neq s$.

Nyní přidáme modulo m.
Odhadujeme # špatných dvojic (a,b), pro něž $h_{a,b}(x) = h_{a,b}(y)$.
Ty odpovídají dvojicím (r,s) s $r \equiv s \pmod m$.

Pro každé r spočítáme, kolik je s t.ž. $r \equiv s$.

- Pokud $[p]$ rozdělíme na m-tice (posl. neúplná), najdeme v 1 m-tici nejvýše 1 takové s.
- $\#s \leq \lceil p/m \rceil \underbrace{-1}_{r=s \text{ jsme zakázali}} \leq \frac{p+m-1}{m} - 1 = \frac{p+m-m-1}{m} = \frac{p-1}{m}$.

Tedy #špatných dvojic $\leq p \cdot \frac{p-1}{m}$, všech dvojic je $p(p-1) \Rightarrow \Pr[\text{dvojice špatná}] \leq 1/m$. ∎

! c-univerzalita (i pro $c=1$) je hodně daleko od skutečné náhodnosti } příklady později

~~např. $\forall h \in \mathcal{H}\ h(0)=0 \Rightarrow h(1)=1$~~

③ **Df:** Systém funkcí $\mathcal{H}$ z $\mathcal{U}$ do $[m]$ je <u>k-nezávislý</u> $\equiv$
$\forall x_1 \dots x_k \in \mathcal{U}$ navzájem různá, $\forall a_1 \dots a_k \in [m]$ $\Pr_{h\in\mathcal{H}}\left[\bigwedge_{i=1}^{k} h(x_i) = a_i\right] \in \mathcal{O}\left(\frac{1}{m^k}\right)$.

<u>Silněji:</u> je (k,c)-nezávislý $\equiv \Pr[\dots] \leq c/m^k$.

- $\mathcal{H}$ je k-nezávislý $\Rightarrow$ $\mathcal{H}$ je $(k-1)$-nezávislý (pro totéž c)
- $\mathcal{H}$ je $(2,c)$-nezávislý $\Rightarrow$ $\mathcal{H}$ je c-univerzální
- 1-nezávislost je příliš slabá, splňuje ji třeba systém konstantních funkcí.

} Dk: cvičení

Df: $\mathcal{L}' := \{ h_{a,b} \mid a,b \in [p]^2 \}$, kde $h_{a,b}(x) = ((ax+b) \bmod p) \bmod m$.

Věta: $\mathcal{L}'$ je $(2,4)$-nezávislý.

Dk: Opět využijeme bijekci mezi (a,b) a (r,s).
Chceme dokázat $\Pr_{r,s}[\underbrace{r \equiv i \ \&\ s \equiv j}] \leq 4/m^2$. (vše mod m)

to jsou nezávislé jevy $\Rightarrow$ stačí $\Pr_r[r \equiv i] \leq 2/m$

$\hookrightarrow$ čísel $r \in [p]$ kongruentních s i je nejvýše $\lceil p/m \rceil \leq \frac{p+m-1}{m} \overset{m \leq p}{\leq} \frac{2p}{m}$

takže $\Pr_r[r \equiv i] \leq \frac{2p}{m}/p = \frac{2}{m}$. ∎

Věta: $\mathcal{L}'$ není 3-nezávislý.

Dk: Zvolíme x,y,z, tak, aby $x+y=2z$, x,y,z různá. Dále $m=p$, takže vše v $\mathbb{Z}_p$.

Chceme, aby platilo (pro dané i,j,k)
$ax+b \equiv i$ ①
$ay+b \equiv j$ ②
$az+b \equiv k$ ③

Z toho: $2k \equiv 2az+2b \equiv a \cdot 2z + 2b \equiv a(x+y)+2b \equiv (ax+b)+(ay+b) \equiv i+j$.

Tedy kdykoli ①–③ platí, je $2k \equiv i+j$.

Proto $\Pr_h[h(x)=i \ \&\ h(y)=j \ \&\ h(z)=k] = \underbrace{\Pr_h[h(z)=k \mid h(x)=i \ \&\ h(y)=j]}_{1} \cdot \underbrace{\Pr_h[h(x)=i \ \&\ h(y)=j]}_{\mathcal{O}(\frac{1}{m^2}),\ \text{neboť } \mathcal{L}' \text{ je 2-nezáv.}}$

pro libovolné i,j a $k=(i+j)\cdot 2^{-1}$

$$= \mathcal{O}(1/m^2) \notin \mathcal{O}(1/m^3).$$

$\boxed{\text{SKLÁDÁNÍ FUNKCÍ}}$ důkaz univerzality $\mathcal{L}$ / 2-nezávislosti $\mathcal{L}'$ fungoval tak, že jsme nejprve dokázali 2-nez. v tělese $\mathbb{Z}_p$ (tedy u funkcí $[p] \to [p]$) a pak nahlédli, že se modulením m moc nepokazí.

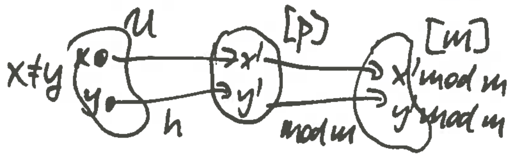
$\Downarrow$

Zkusme o tom dokázat něco obecnějšího

Věta: Nechť $\mathcal{H}$ je $(2,c)$-nezávislý systém funkcí z $\mathcal{U}$ do $[p]$ a $m \le p$.
Potom $\mathcal{H}^* := \{h \bmod m \mid h \in \mathcal{H}\}$ je $2c$-univerzální a $(2,4c)$-nezávislý.

p nemusí být prvočíslo

Dk: $\mathcal{U} \xrightarrow{h\in\mathcal{H}} [p] \xrightarrow{\bmod m} [m]$



① univerzalita:

$$\Pr_h[h(x) \bmod m = h(y) \bmod m]$$
$$= \cancel{\Pr[h(x)=h(y) \vee (h(x)\equiv h(y) \mid h(x)\neq h(y))]}$$
$$\le \underbrace{\Pr[h(x)=h(y)]}_{\le} \cancel{+ \Pr[h(x)\equiv h(y) \mid \dots]}$$

$\le \frac{c}{p^2}$ z 2-nz.

$$= \Pr_h\Big[\bigvee_{\substack{i,j\\ i\equiv j}} h(x)=i \;\&\; h(y)=j\Big] = \sum_{i\equiv j} \Pr[h(x)=i \;\&\; h(y)=j] \le p\cdot\frac{p+m-1}{m}\cdot\frac{c}{p^2} = c\cdot\frac{p+m-1}{pm}$$

$i\equiv j$ ← takových dvojic je $p\cdot\left\lceil\frac{p}{m}\right\rceil \le p\cdot\frac{p+m-1}{m}$

$$= \frac{c}{m}\cdot\frac{p+m-1}{p} \qquad \frac{p+m-1}{p} \le \frac{2p}{p} = 2.$$

② nezávislost: $\Pr_h\Big[\bigvee_{\substack{i',j'\\ i'\equiv i,\ j'\equiv j}} \underbrace{h(x)=i' \;\&\; h(y)=j'}_{\Pr \le c/p^2}\Big] \le \frac{c}{p^2}\cdot\underbrace{\left\lceil\frac{p}{m}\right\rceil^2}_{\le\left(\frac{p+m-1}{m}\right)^2} \le \frac{c}{m^2}\cdot\underbrace{\left(\frac{p+m-1}{p}\right)^2}_{\le 4} \le \frac{4c}{m^2}$

Zobecnění: $\mathcal{H}$ (k,c)-nezávislý $\Rightarrow$ $\mathcal{H}'$ (k,c')-nezávislý pro $c' = c\cdot\underbrace{\left(\frac{p+m-1}{p}\right)^k}_{\le\left(1+\frac{m}{p}\right)^k \le e^{\frac{km}{p}} \le \text{const pro } p=\Omega(km)} \le c\cdot 2^k$

Věta: Nechť $\mathcal{F}$ je c-univerzální z $\mathcal{U}$ do $[r]$
a $\mathcal{G}$ je $(2,d)$-nezávislý z $[r]$ do $[m]$.
Potom $\mathcal{H} := \{f\circ g \mid f\in\mathcal{F},\ g\in\mathcal{G}\}$ je $(2,d')$-nezávislý z $\mathcal{U}$ do $[m]$, pro $d' = \frac{(c+1)d}{m^2}$.

Dk: $\Pr_h[h(x)=i \;\&\; h(y)=j] = \underbrace{\Pr_{f,g}[\underbrace{g(f(x))=i \;\&\; g(f(y))=j}_{A} \;\&\; f(x)=f(y)]}_{\le \underbrace{\Pr[A \mid f(x)=f(y)]}_{\le \frac{d}{m}}\cdot\underbrace{\Pr[f(x)=f(y)]}_{\le\frac{c}{r}\le\frac{c}{m}}} + \underbrace{\Pr_{f,g}[A \;\&\; f(x)\neq f(y)]}_{\le \Pr[A] \le \frac{d}{m^2} \text{ z nezávislosti } \mathcal{G}}$

$$\le \frac{cd}{m^2} + \frac{d}{m^2} = \frac{(c+1)d}{m^2}$$

… šlo by zpřesnit na $\frac{\left(c\cdot\frac{m}{r}+1\right)d}{m^2}$.

POLYNOMY

Df: $\mathcal{P}_k := \{h_{\vec a} \mid \vec a \in \mathbb{Z}_p^k\}$, $h_{\vec a}(x) := \left(\sum_{i=0}^{k-1} a_i x^i\right) \bmod p$ … fce z $[p]$ do $[p]$

Věta: Systém $\mathcal{P}_k$ je $(k,1)$-nezávislý.

Dk: Danými k body prochází právě 1 polynom stupně menšího než d. (platí v libovolném tělese)

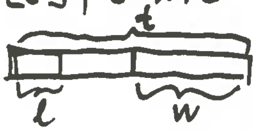
Chceme-li hešovat do $[m]$ pro $m<p$, složíme s mod m, pro $p=\Omega(km)$ je (k, const)-nezávislý.
↑ nezávislá na k

$\boxed{\text{MULTIPLY-SHIFT}}$ – rychlé hešovací fce $ax\langle w-l:w\rangle$

Df: $\mathcal{M} := \{h_a \mid a \in [2^w]\}$ (liché), kde $h_a(x) := \lfloor (ax \bmod 2^w) / 2^{w-l} \rfloor$ … z $[2^w]$ do $[2^l]$

$\mathcal{M}' := \{h'_{a,b} \mid a,b \in [2^t]\}$, kde $h'_{a,b}(x) := \lfloor ((ax+b) \bmod 2^t) / 2^{t-l} \rfloor$ … z $[2^w]$ do $[2^l]$, $t \geq w+l$

$(ax+b)\langle t-l:t\rangle$ $\bmod 2^l$



Věta: $\mathcal{M}$ je 2-univerzální, $\mathcal{M}'$ je 2-nezávislý.
(Dk viz literatura)

lepší notace: $x\langle i:j\rangle$ jsou bity i až $j-1$ čísla x

$\boxed{\text{TABELAČNÍ HEŠOVÁNÍ}}$

Mějme $\mathcal{U} = [2^{dt}]$. Náhodně zvolíme fce (tabulky) $T_0, \dots, T_{d-1} : [2^t] \to [2^t]$. } potřebujeme $dt \cdot 2^t$ bitů paměti

$h(x) := \bigoplus_{i=0}^{d-1} T_i(x \langle ti : t(i+1) \rangle)$

← obecně: čas $O(d)$, prostor $O(\mathcal{U}^{1/d+\varepsilon})$

Věta: Tabelační hešování je 3-nezávislé, ale není 4-nezávislé.
(Dk: cvičení)

$\boxed{\text{HEŠOVÁNÍ VEKTORŮ A ŘETĚZCŮ}}$

Df: $\mathcal{S}$ z $\mathbb{Z}_p^k$ do $\mathbb{Z}_p$: $\mathcal{S} := \{h_{\vec a} \mid \vec a \in \mathbb{Z}_p^k\}$, $h_{\vec a}(\vec x) := \vec a \cdot \vec x$

Věta: $\mathcal{S}$ je 1-~~nezávislý~~ univerzální.

Dk: $\Pr_h[h(\vec x) = h(\vec y)] = \Pr_{\vec a}[\vec a \vec x = \vec a \cdot \vec y] = \Pr_{\vec a}[\vec a \cdot \underbrace{(\vec x - \vec y)}_{\text{nenulový vektor } \vec z,\ \text{BÚNO } z_k \neq 0} = 0] =$

$= \Pr_{\vec a}\left[\sum_{i=1}^{k-1} a_i z_i + a_k z_k = 0\right] = \frac{1}{p}$.

intuice: cokoliv + rovnoměrně náhodné (nezávislé) = rovnoměrně náhodné

pro každou volbu $a_1 \dots a_{k-1}$ $\exists!\, a_k$, s nímž rovnost platí.

Df: $\mathcal{S}'$ z $\mathbb{Z}_p^k$ do $\mathbb{Z}_p$: $\mathcal{S}' := \{h_{\vec a, b} \mid \vec a \in \mathbb{Z}_p^k, b \in \mathbb{Z}_p\}$, $h_{\vec a, b}(\vec x) := \vec a \cdot \vec x + b$

Věta: $\mathcal{S}'$ je (2,1)-~~univerzální~~ nezávislý.
[(2,2)-univerzálnost plyne z obecné věty]

Df: $\mathcal{Q}$ z $\mathbb{Z}_p^k$ do $\mathbb{Z}_p$: $\mathcal{Q} := \{h_a \mid a \in \mathbb{Z}_p\}$, $h_a(\vec x) := \sum_{i=0}^{k-1} x_i \cdot a^i$

Věta: $\mathcal{Q}$ je k-univerzální.

Dk: $\Pr_h[h(\vec x) = h(\vec y)] = \Pr[h(\vec x - \vec y) = 0]$, ale to je polynom stupně $<k$, čili má max. k kořenů

lze používat i na řetězce proměnlivé velikosti (omezené max. k), stačí padding znaky, které se jinde nevyskytují (nejlépe nulami)

↓ složíme s $((bx + c) \bmod p) \bmod m$

$\mathcal{Q}' := \{h'_{a,b,c} \mid a,b,c \in [p]\}$,
$h'_{a,b,c}(\vec x) := \left(\left(b \cdot \sum_{i=0}^{k-1} x_i a^i\right) + c\right) \bmod p\Big) \bmod m$

Věta: $\mathcal{Q}'$ je 2-nezávislý pro $p \geq km$

Dk: Z obecné věty o skládání.