ANALÝZA LINEÁRNÍHO PŘIDÁVÁNÍ

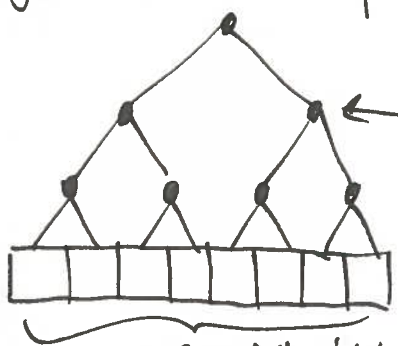
Věta [Knuth 1963]: Pokud $m \geq (1+\varepsilon)\cdot n$, očekávaný # porovnání při neúspěšném hledání je $O(1/\varepsilon^2)$ pro plně náhodnou hešovací fci.

Dokážeme o něco slabší tvrzení: Pokud $m=3n$, $\mathbb{E}[\#\text{porovnání}] \in O(1)$.
(snadno lze upravit na $(1+\varepsilon)$ místo 3)

Bude se hodit Chernoffova nerovnost: Nechť $X_1 - X_n$ jsou ~~vzájemně~~ nezávislé náhodné veličiny s hodnotami $\{0,1\}$, $X = \sum_i X_i$ a $\mu = \mathbb{E}[X]$.
Potom: $\Pr[X > c\cdot\mu] < \left(\frac{e^{c-1}}{c^c}\right)^{\mu}$ pro $c>1$

$\Pr[X < c\cdot\mu] < \left(\frac{e^{c-1}}{c^c}\right)^{\mu}$ pro $c \in (0,1)$.

Uvažujme binární strom postavený nad přihrádkami:



$m=3n$ přihrádek

vnitřní vrchol ve výšce h odpovídá zarovnanému intervalu délky 2^h

tem říkáme bloky délky 2^h

Df: Vrchol je nebezpečný $\equiv$ do jeho intervalu podle hešovací funkce ~~[illegible]~~ více než $\frac{2}{3}\cdot 2^h$ prvků

nepočítáme prvky přesunuté kvůli kolizím

Lemma: $\Pr[\text{daný vrchol ve výšce } h \text{ je nebezpečný}] < (e/4)^{2^h/3}$.

Dk: Nechť I je ~~interval~~ blok délky 2^h pokrytý vrcholem.
Pro prvky $x_1 - x_n$ definujeme indikátory $C_i = \begin{cases} 0 \\ 1 & \text{pokud } h(x_i) \in I \end{cases}$

$C = \sum_i C_i$ je # prvků zahešovaných do intervalu I

$\mathbb{E}[C_i] = 2^h/m \quad \dots \Pr[h(x_i) \in I]$

$\mu = \mathbb{E}[C] = \sum_i \mathbb{E}[C_i] = n\cdot\frac{2^h}{m} \leq 2^h/3$.

Chernoff

$\Pr[\text{vrchol je nebezpečný}] = \Pr[C > \frac{2}{3}\cdot 2^h] = \Pr[C > 2\mu] < \left(\frac{e}{4}\right)^{\mu} = \left(\frac{e}{4}\right)^{2^h/3}$.

Df: Běh je maximální interval přihrádek zaplněných prvky. ← tady už počítáme i přesunuté
(indexujeme mod m)

Lemma: Nechť R je běh délky alespoň $2^{\ell+2}$. Pak aspoň jeden z prvních 4 bloků velikosti 2^ℓ ~~[illegible]~~ protínajících R je nebezpečný.

Dk: Nechť $B_0 - B_3$ jsou zmíněné bloky.
B_0 zasahuje do R aspoň 1 prvkem, $B_1 - B_3$ jsou celé uvnitř R.
Nechť $L := (\bigcup_i B_i) \cap R \quad |L| \geq 3\cdot 2^{\ell} + 1$

Navíc všechny prvky ~~zahešované~~ uložené do L jsou opravdu ~~uložené~~ zahešované do L ← jinak by šlo R rozšířit doleva, tedy by to nebyl běh

$\Rightarrow$ do $\bigcup_i B_i$ se zahešovalo aspoň $|L|$ prvků

... a stačí použít princip holubníku: kdyby žádný B_i nebyl nebezpečný, bylo by v $\bigcup_i B_i$ max. $\frac{8}{3} 2^l$ prvků, což je $< |L|$ ↯

<u>Lemma:</u> Pokud běh obsahující daný prvek x má délku $\in [2^{l+2}, 2^{l+3})$, pak aspoň 1 z ~~následujících~~ těchto 12 bloků délky 2^l je nebezpečný: blok obsahující $h(x)$, 8 předchozích bloků a 3 následující.

<u>Dk:</u> Jelikož běh dané délky obsahuje max. 9 bloků délky 2^l, leží jeho začátek nejvýše 8 bloků před $h(x)$, takže nebezpeč. blok z předch. lemmatu leží mezi určenými 12 bloky.

<u>Nyní počítejme:</u>

$\Pr[\text{běh obsahující } x \text{ má délku} \in [2^{l+2}, 2^{l+3})]$

$\leq \Pr[\text{některý z 12 bloků z lemmatu je nebezpečný}]$

$\leq 12 \cdot \Pr[\text{daný blok velikosti } 2^l \text{ je nebezpečný}]$

$\leq 12 \cdot (e/4)^{2^l/3}$

↓

$$\mathbb{E}[\underbrace{\text{délka běhu obsahujícího } x}_{L}] \leq \sum_{l \geq 0}^{1+2+3} 2^{l+3} \cdot \Pr[L \in [2^{l+2}, 2^{l+3})]$$

$$\leq c + \sum_{l \geq 0} 2^{l+3} \cdot \left(\left(e/4\right)^{1/3}\right)^{2^l}$$

$$\leq c + c' \cdot \sum_{l \geq 0} 2^l \cdot q^{2^l} \quad q \in (0,1)$$

$$\leq c + c' \sum_{i \geq 0} i \cdot q^i \quad \text{konverguje pro každé } q \in (0,1)$$

$$\in O(1).$$

<u>Poznámky:</u>

- pro 2-nezávislost se můžeme pokusit použít Markovovu nerovnost
  → vyjde $\Pr[\text{blok nebezp.}] \leq 1/2$, což je příliš slabé ... dostaneme $\mathbb{E}[L] \in O(n)$.
- pro 3-nezávislost použije Čebyševova nerovnost (1 stupeň volnosti sebere fixace x, zbylé 2 stačí na odhad rozptylu) → $\Pr[\ldots] \leq 9 \cdot 2^{-l}$ ... dostaneme $\mathbb{E}[L] \in O(\log n)$.
- pro 5-nezávislost použijeme analogickou nerovnost pro momenty 4. řádu, z ní dostaneme $\Pr[\ldots] \leq 36 \cdot 2^{-2l}$ ... a to stačí na $\mathbb{E}[L] \in O(1)$.