- Domácí úkoly - 4-5 70% bodů na zápočet
- cvičení pátek 10:40 - nepovinné

Plán - Informace & komprese
- samoopravné kódy
- komunikační složitost

Informace

Scrabble

1x X_{10} 5x E_1

FAX FIX SEX EX

"Přidám slovo obsahující X_{10}" vs.
"Přidám slovo obsahující E_1"

- "Řeknu Vám velkou novinu:
  a) náš prezident odstoupil
  b) v Praze zavedou trolejbusy
  c) v poledne jsem byl na obědě
  d) v poledne jsem nebyl na obědě
  e) hadr snědl roztříštěný koberec"

- překvapení vs význam
  → více překvapení = více informace
  - význam neumíme analyzovat
  → informaci přiřadíme k pravděpodobnostním událostem
  (např. vygenerovaná zpráva)

Informace - chceme

A událost, $I(A)$ informace obsažená v události.

1) $I(A)$ klesá s rostoucí pstí. A

Př: Zamíchám balíček 32 karet, vrchní karta je
1) červená A
2) sedma B
3) červená sedma C

$p(A) \geq p(B) \geq p(C)$
$I(A) \leq I(B) \leq I(C)$

2) $I(A \& B) = I(A) + I(B)$ pro $\forall$ A, B nezávislé
$(p(A \& B) = p(A) \cdot p(B))$

3) $I(A) \geq 0$ pro $\forall A$

$\longrightarrow I(A) = -\log_a P(A)$ pro pevně zvolené $\underline{a} > 1$.

$-\log_a P(A) = \log_a \frac{1}{P(A)}$

- Alternativní přístup – Kolmogorovská složitost
  (uvidíme později:)

Opakování pst:

- pravděpodobnostní prostor Ω ... konečná nebo spočetná množina
  s mírou pravděpodobnosti $p : \Omega \to \mathbb{R}$
  $\forall \omega \in \Omega,\ p(\omega) \geq 0$
  $\sum_{\omega \in \Omega} p(\omega) = 1$

  Př: Ω ... množina možných zpráv, $\{0,1\}^n$

- jev (událost) $A \subseteq \Omega$ $\Pr[A] = \sum_{\omega \in A} p(\omega)$.
- jevy A a B jsou nezávislé:
  $\Pr[A \& B] = \Pr[A] \cdot \Pr[B]$.

- podmíněná pravděpodobnost A na B:
  $$\Pr[A / B] = \frac{\Pr[A \& B]}{\Pr[B]}$$
  Př: A & B jsou nezávislé $\Leftrightarrow \Pr[A/B] = \Pr[A]$.

- náhodná proměnná $X : \Omega \to \mathbb{R}$.
  (veličina)

  Př: náhodná proměnná X definuje různé jevy
  $\forall S \subseteq \mathbb{R} \to$ jev $[X \in S]$.

  dvě náhodné proměnné X, Y jsou nezávislé,
  jestliže $\forall S, S' \subseteq \mathbb{R}$ $[X \in S]$ a $[Y \in S']$ jsou nezávislé

  Př: n.p. $X, Y : \Omega \to \{0,1\}^n$
  1) X vybere náhodný řetězec $\in \{0,1\}^n$
     Y vybere nezávislý náhodný řetězec $\in \{0,1\}^n$

(P1)
$\forall x, y \in \{0,1\}^n$
$$\Pr[X = x \ \& \ Y = y] = \frac{1}{2^{2n}}$$

2) vybereme náhodné řetězce $a, b, c \in \{0,1\}^{n/2}$
položíme $X = ab$ $Y = bc$
$\forall x, y \in \{0,1\}^n$ $\Pr[X = x] = \frac{1}{2^n}$
$\Pr[Y = y] = \frac{1}{2^n}$

(P2)

$$\Pr[X = x \ \& \ Y = y] \neq \frac{1}{2^n}$$

$\Rightarrow X$ a Y jsou závislé

$$\Pr[X = x \ \& \ Y = y] = \begin{cases} 0 & x_{\frac{n}{2} \dots n} \neq y_{1 \dots \frac{n}{2}} \\ \frac{1}{2^{3n/2}} & x_{\frac{n}{2} \dots n} = y_{1 \dots \frac{n}{2}} \end{cases}$$

• střední hodnota: $\mathbb{E}[X] = \sum_{\omega \in \Omega} p(\omega) \cdot X(\omega)$

$$= \sum_{c \in \mathbb{R}} c \cdot \Pr[X=c].$$

• náhodné proměnné X, Y, Z:

$$\mathbb{E}[X+Y+Z] = \mathbb{E}[X] + \mathbb{E}[Y] + \mathbb{E}[Z]$$

"linearita střední hodnoty"

• Markovova nerovnost: pro nezápornou náhodnou proměnnou X a pro $\forall k \in \mathbb{R}$

$$\Pr[X \geq k \cdot \mathbb{E}[X]] \leq \frac{1}{k}.$$

Entropie (neurčitost) X náhodná proměnná

$$H(X) = - \sum_{x} p(x) \cdot \log_2 p(x)$$

konvence $0 \cdot \lg 0 = 0$

→ střední hodnota informace

• $H(X) \geq 0$ Dk: $\forall\, 0 < p < 1 \quad \log_2 p < 0$

$$\Rightarrow H(X) = - \sum_{x} p(x) \cdot \log p(x)$$

• $H(X) \leq \log \underbrace{|X|}_{\hookrightarrow |\mathrm{supp}(X)|}$ (Dk později)

Př: X je náhodná proměnná s hodnotami z $\{0,1\}^n$

1) $\forall x \in \{0,1\}^n, \quad p(x) = 2^{-n} \quad (= \Pr[X=x])$

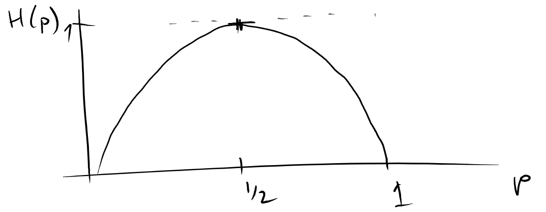
$$H(X) = \sum_{x \in \{0,1\}^n} \frac{1}{2^n} \lg 2^n = n$$

2) $p(0^n) = \frac{1}{2} \qquad \forall x \neq 0^n \quad p(x) = \frac{1}{2^{n+1}-2}$

$$H(X) = \frac{1}{2} \lg 2 + \frac{1}{2} \lg (2^{n+1}-2) = \frac{n}{2} + \Theta(1)$$

Př: $H(p) = p \cdot \lg \frac{1}{p} + (1-p) \lg \frac{1}{1-p}$

$$X = \begin{cases} 0 & \text{s prav. } p \\ 1 & \text{s prav. } 1-p \end{cases}$$



společná entropie X, Y náhodné proměnné

$$H(X,Y) = \sum_{\substack{x \in X \\ y \in Y}} p(x,y) \cdot \lg \frac{1}{p(x,y)}$$

$$= \mathbb{E}_{X,Y} \lg \frac{1}{p(X,Y)}$$

Př: (P1) výše X, Y

$$H(X,Y) = 2n$$

(P2) výše X, Y

$$H(X,Y) = \frac{3}{2}n$$

- $\pi_x, \pi_y : \{0,1\}^n \to \{0,1\}^n$ ... pevně zvolené permutace

$X' = \pi_x(X) \quad Y' = \pi_y(Y) \quad H(X',Y') = \frac{3}{2}n.$

Podmíněná entropie

$$H(Y|X) = \sum_{\substack{x \in X \\ p(x) \neq 0}} p(x) \cdot H(Y \mid X = x)$$

$$= \sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) \lg \frac{1}{p(y|x)}$$

$$= \sum_{x \in X} \sum_{y \in Y} p(x,y) \lg \frac{1}{p(y|x)}$$

$$\text{"} = \mathbb{E}_{X,Y} \lg \frac{1}{p(Y|X)} \text{"}$$

Př: 1) $H(Y|Y) = 0$

2) $H(Y|X) = H(Y)$ pokud X & Y jsou nezávislé

3) (P1) X, Y $\quad H(Y|X) = H(Y) = n$

4) (P2) X, Y $\quad H(Y|X) = \frac{n}{2}$

$H(Y) = n$

Věta ("chain rule")

$$H(X,Y) = H(X) + H(Y|X)$$

Dk: $H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \cdot \lg p(x,y)$

$$= -\sum_{x \in X} \sum_{y \in Y} p(x,y) \cdot \lg p(x) \cdot p(y|x)$$

$$= -\sum_{x \in X} \sum_{y \in Y} p(x,y) \left[\lg p(x) + \lg p(y|x) \right]$$

$$= \underbrace{-\sum_{x \in X} p(x) \lg p(x)}_{H(X)} \underbrace{- \sum_{\substack{x \in X \\ y \in Y}} p(x,y) \lg p(y|x)}_{H(Y|X)}$$

Důsledek: $H(X,Y|Z) = H(X|Z) + H(Y|X,Z)$

$H(X) + H(X|Y) = H(X,Y) = H(Y) + H(X|Y)$

Obecně: náhodné prom. $X_1, X_2, \dots X_k$

$$H(X_1, X_2, \dots X_k) = \sum_{i=1}^{k} H(X_i \mid X_1, X_2, \dots X_{i-1})$$

Dk:

$$= H(X_1) + H(X_2, \dots X_k | X_1)$$
$$= H(X_1) + H(X_2 | X_1) + H(X_3, \dots X_k | X_1, X_2)$$
$$= \dots$$ ∎

Význam entropie

– očekávaný počet bitů při kódování (kompresi) ... uvidíme

...

Vzájemná informace

Př: zprávy v různých novinách o stejné události.
$N_1, N_2, \dots N_k$ k periodik
kolik informace o zprávě N_1 mám, již po přečtení N_2, N_3, atd.

Def: náhodné proměnné X, Y
$$I(X:Y) = H(X) - H(X|Y) \dots$$
vzájemná informace X a Y

... o kolik se sníží neurčitost X, když znám Y

Pozorování: $I(X:Y) = I(Y:X)$ ... symetrie informace

Př: 1) $I(X:X) = H(X) - H(X|X) = H(X)$

2) X, Y nezávislé $I(X:Y) = H(X) - H(X|Y) = 0$

3) (P2) $I(X:Y) = H(X) - H(X|Y) = \frac{n}{2}$

4) X, Y nezávislé hody kostkou
$Z = X + Y \pmod 6$
$I(X:Y) = 0 \quad I(X:Z) = 0 \quad I(Y:Z) = 0$
$$I(X,Y:Z) = H(X,Y) - \underbrace{H(X,Y|Z)}_{\underbrace{H(X|Z)}_{H(X)} + \underbrace{H(Y|X,Z)}_{=0}} = H(Y) = H(X)$$

Vlastnosti:

- $I(X:Y) = H(X) + H(Y) - H(X,Y)$

Def: Kullback-Leiblerova vzdálenost
náhodné proměnné X, Y se stejným oborem hodnot
$p(x) = Pr[X = x] \qquad q(x) = Pr[Y = x]$
$$D(X \| Y) = D(p \| q) = \sum_x p(x) \lg \frac{p(x)}{q(x)}$$

konvence: $0 \cdot \lg \frac{0}{0} = 0 \qquad 0 \cdot \lg \frac{0}{q} = 0$

$\qquad\qquad\qquad\qquad\qquad\qquad p \cdot \lg \frac{p}{0} = \infty$

Pozorování: $D(p \| q) = \sum_x p(x) \lg \frac{1}{q(x)} - \underbrace{\sum_x p(x) \lg \frac{1}{p(x)}}_{H(X)}$

uvidíme $D(p \| q) \geq 0$

- možná interpretace: o kolik se prodlouží průměrná délka kódu při distribuci p, když použiji kód pro distribuci q.

Tvrzení: $I(X:Y) = D(p(x,y) \,\|\, p(x) \cdot p(y))$

Dk.: $= \underbrace{\sum_{x \in X} p(x) \lg \frac{1}{p(x)}}_{H(X)} - \underbrace{\sum_{\substack{x \in X \\ y \in Y}} p(x,y) \lg \frac{1}{p(x|y)}}_{H(X|Y)}$

$= \sum_{\substack{x \in X \\ y \in Y}} p(x,y) \lg \frac{1}{p(x)} - \quad -\!\!\text{"}\!\!-$

$= \sum_{x,y} p(x,y) \lg \frac{p(x|y)}{p(x)} \qquad (*)$

$= \sum_{x,y} p(x,y) \lg \frac{p(x|y) \cdot p(y)}{p(x) \cdot p(y)}$ $\;$ "$p(x,y)$" $\qquad \blacksquare$

Tvrzení: $I(X:Y) = \mathbb{E}_{y \in Y}\left[D(X|Y=y \,\|\, X)\right]$.

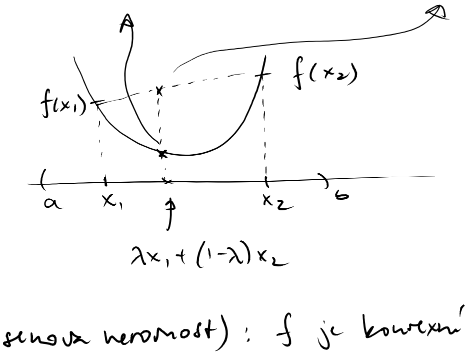
Dk.: plyne z $(*)$ výše.

Michal Koucky at 14. 3. 2016 21:28

- f je konvexní na (a,b) pokud $\forall x_1, x_2 \in (a,b)$

$\forall\, 0 \leq \lambda \leq 1$

$$f(\lambda x_1 + (1-\lambda) x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$



Věta (Jensenova nerovnost): f je konvexní fce na (a,b) a X je náhodná proměnná s hodnotami z (a,b) pak

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$$

Dk.: pro $|X|$ konečné

1) $p(x_1) = \lambda \qquad p(x_2) = 1 - \lambda \qquad X \in \{x_1, x_2\} \subseteq (a,b)$

z definice konvexnosti ✓

2) $p(x_1) = p_1, \dots \quad p(x_n) = p_n$ $\quad (|supp(X)| = n)$

indukce dle n $\qquad p_i' = \frac{p_i}{1-p_n}$

$$\sum_{i=1}^{n} p_i f(x_i) = p_n f(x_n) + (1-p_n) \sum_{i=1}^{n-1} p_i' f(x_i)$$

$$\overset{\text{ind. předp.}}{\geq} p_n f(x_n) + (1-p_n) f\left(\sum_{i=1}^{n-1} p_i' x_i\right)$$

$$\overset{\text{konvexita}}{\geq} f\left(p_n x_n + (1-p_n) \sum_{i=1}^{n-1} p_i' x_i\right)$$

$$= f\left(\sum_{i=1}^{n} p_i x_i\right) \qquad \boxtimes$$

Věta: Nechť $p(x), q(x)$ jsou pstní rozdělení pro $x \in X$.

Pak $D(p \| q) \geq 0$

Dk: $\quad A = \{x ;\ p(x) > 0\}$

$$-D(p\|q) = -\sum_{x \in A} p(x) \lg \frac{p(x)}{q(x)}$$

$$= \sum_{x \in A} p(x) \lg \frac{q(x)}{p(x)}$$

$$\leq \lg \sum_{x \in A} p(x) \frac{q(x)}{p(x)}$$

$$\leq \lg \sum_{x \in X} q(x)$$

$$= \lg 1 = 0 \qquad \boxtimes$$

$\rightarrow \quad D(p\|q) = 0 \iff p = q$

Důsledek: $\quad I(X:Y) = 0$

Dk: $\quad I(X:Y) = D(p(x,y) \,\|\, p(x) \cdot p(y)) \qquad \boxtimes$

Důsledek: $\quad H(X) \leq \lg |X|$ s rovností
pouze pokud X je rovnoměrné rozdělení.

Dk: $\quad$ def $\ u(x) = \frac{1}{|X|} \qquad p$ je rozdělení X

$$D(p\|u) = \sum p(x) \lg \frac{p(x)}{u(x)} = \lg|X| - H(X)$$

$$\underset{0}{\vee\!\!|} \qquad \Rightarrow \quad \lg|X| \geq H(X) \qquad \boxtimes$$

Důsledek: $\quad H(X|Y) \leq H(Y)$ s rovností iff
jsou X a Y nezávislé

Dk: $\quad 0 \leq I(X:Y) = H(X) - H(X|Y) \qquad \boxtimes$

Př: $\quad Pr[X = 0^n] = \frac{1}{2} \qquad \forall x \in \{0,1\}^n \setminus \{0^n\}$
$\qquad Pr[X = x] = \frac{1}{2(2^n-1)}$

$\ldots\ 0 \quad x = 0^n$

$2)\quad X \neq 0^n \qquad H(X) = \frac{n}{2} + O(1)$

$\qquad\qquad\qquad\qquad H(X|Y) \leq \frac{n}{2} + O(1)$

ale: $H(X|Y=1) = n \pm O(1)$

$H(X|Y=0) = 0$

$H(Y) = 1$

- $I(X:Y|Z) = \mathop{\mathbb{E}}_{z \in Z} [I(X:Y|Z=z)]$

Pozorování: $I(X_1, X_2, \dots X_n : Y) = \sum_{i=1}^{n} I(X_i : Y | X_1, \dots X_{i-1})$

Dk: $I(X_1, \dots X_n : Y) = H(X_1, \dots X_n) - H(X_1, X_2 \dots X_n | Y)$

$= \sum_{i=1}^{n} H(X_i | X_1 \dots X_{i-1}) - H(X_i | X_1 \dots X_{i-1}, Y)$

$= \sum_{i=1}^{n} I(X_i : Y | X_1, \dots X_{i-1})$ ∎

Alenka $\qquad\qquad$ Bob

$X \xrightarrow[\text{šum}]{} Y \xrightarrow{g} g(Y) = Z$

$Z \dots$ odhad X

$I(X:Y)$ vs $I(X:Z)$?

Def: X, Y, Z splňují Markovovskou vlastnost

pokud $\forall_{x,y,z}$

$\Pr[Z=z \mid Y=y] = \Pr[Z=z \mid Y=y \ \& \ X=x]$

"$X \to Y \to Z$"

- $p(z|y) = p(z|y,x)$
- $p(x,z|y) = p(x|y) \cdot p(z|y\,x) = p(x|y)\, p(z|y)$

$= p(z|y) \cdot p(x|y\,z)$

$\Rightarrow p(x|y\,z) = p(x|y)$

$\Longrightarrow$ "$X \to Y \to Z$" iff "$Z \to Y \to X$"

... symetrie

Věta: Pokud $X \to Y \to Z$ pak $I(X:Y) \geq I(X:Z)$

Dk: $I(X:Y,Z) = I(X:Y) + I(X:Z|Y)$

$= I(X:Z) + I(X:Y|Z)$

$I(X:Z|Y) = 0$ protože X & Z jsou nezávislé podmíněno na Y

$I(X:Y|Z) \geq 0$ triv.

$\Rightarrow I(X:Z) \leq I(X:Y)$ ∎

$\Rightarrow \quad I(X:Z) \le I(X:Y)$

Kódování

- $C : X \to \Sigma^*$ (abeceda) — chci $\forall x \neq y$ $C(x) \neq C(y)$

  ... kód

- průměrná délka kódu C : $L(C) = \sum_{x \in X} p(x)\, l(x)$
- uzávěr kódu $C^*(x_1, \dots x_k) = C(x_1)\,C(x_2) \dots C(x_k)$
- kód je jednoznačně dekódovatelný, pokud C^* nemá kolizi.
- bezprefixový kód C : $\forall x \neq y$ $C(x)$ není prefix $C(y)$

Př:

10	0
00	10
11	110
110	111
↑ jednoznačně dekódovatelný	↑ bezprefixový

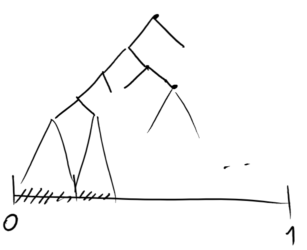
Věta: (Kraftova nerovnost)

Pro bezprefixový kód C s délkami kódu $l_1, l_2, \dots$

platí: $$\sum 2^{-l_i} \le 1$$

(Obecně $\sum |\Sigma|^{-l_i} \le 1$ pro nebinární abecedu.)

Dk:



ke kódovému slovu $a_1 a_2 \dots a_k$ přiřadím interval $[0.a_1 a_2 \dots a_k,\ 0.a_1 a_2 \dots a_k + 2^{-k})$

intervaly přiřazené různým kódovým slovům jsou disjunktní, jejich sjednocení je podmnožina $[0;1)$, tedy celková délka splňuje

$$\sum 2^{-l_i} \le 1 .$$

Podobně pro obecnou abecedu Σ. ∎

- Platí i opačná implikace: pokud mám $l_1, \dots l_n$ t.ž. $\sum 2^{-l_i} \le 1$, pak existuje bezprefixový kód požadovaných délek.

(lze přiřazovat od největší 2^{-l_i} do nejmenší:
$l_1 \le l_2 \le l_3 \dots \le l_n$

$$\underbrace{|\longrightarrow)}_{2^{-l_1}}\underbrace{\longrightarrow)}_{2^{-l_2}} \cdots \underbrace{\longrightarrow)}_{2^{-l_n}} \qquad)$$

- optimální bezprefixový kód pro X, chci minimalizovat průměrnou délku slova.

$$C: X \to \{0,1\}^* \qquad L = \mathbb{E}[|C(X)|]$$

$$l(x) = |C(x)|$$

$$L = \sum_x p(x) \cdot l(x)$$

$$L - H(X) = \sum_x p(x)\, l(x) - \sum_x p(x) \frac{1}{p(x)} =$$
$$= \sum_x p(x) \lg \frac{1}{2^{-l(x)}} - \sum_x p(x) \frac{1}{p(x)} =$$
$$= \sum_x p(x) \lg \frac{p(x)}{2^{-l(x)}} = (*)$$

Kraft.

$$c = \sum 2^{-l(x)} \qquad q(x) = \frac{2^{-l(x)}}{c} \qquad c \le 1$$

$$(*) = \sum_x p(x) \lg \frac{p(x)}{q(x) \cdot c} = \sum_x p(x) \lg \frac{p(x)}{q(x)} + \sum_x p(x) \lg \frac{1}{c}$$

$$= \underbrace{D(p\|q)}_{\ge 0} + \underbrace{\lg \frac{1}{c}}_{\ge 0} \ge 0$$

$$\Rightarrow \quad L \ge H(X) \qquad \blacksquare$$

Shannonův kód:

$$l(x) = \left\lceil \lg \frac{1}{p(x)} \right\rceil$$

$$\to \quad \lg \frac{1}{p(x)} \le l(x) \le \left(\lg \frac{1}{p(x)}\right) + 1$$

$$\to \quad H(X) \le L_{\text{shannon}} \le H(X) + 1$$

optimální ±1 bit.

- kódování k symbolů za sebou $X_1, X_2, \dots X_k$ $\quad X_i \sim X$ nezávislé

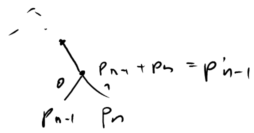
$$H(X_1, X_2, \dots X_k) \le L_{X_1 \dots X_k} \le H(X_1 \dots X_k) + 1$$

$$k\, H(X) \qquad \to \text{průměrně } \tfrac{1}{k} \text{ bitu navíc za každý symbol.}$$

Př: $\quad p(x_1) = 0.9999 \qquad p(x_2) = 0.0001$
$\quad l(x_1) = 1 \qquad l(x_2) = 14 \quad$ (lze lépe :)

Huffmanův kód: opakovaně spojuj dva nejmenší posl. a stav strom od listů.

$p_1 \ge p_2 \ge \dots p_n$



→ dá se ukázat, že dává nejmenší možnou průměrnou délku

Fanův kód: stav strom odshora, minimalizuj rozdíl v poli vlevo a vpravo.

$p_1, p_2, \dots, p_n$ $\qquad \min_k \left| \sum_{i=1}^{k} p_i - \sum_{i=k+1}^{n} p_i \right|$

Fakt: průměrná délka slova $\le H(X) + 2$

Věta: (McMillan) Jednoznačně dekódovatelný kód C s délkami $l(x)$ splňuje

$$\sum 2^{-l(x)} = 1$$

Dk: Uvažme C^k $\qquad k \ge 1$

$$\left(\sum_x 2^{-l(x)}\right)^k = \sum_{x_1} \dots \sum_{x_k} 2^{-l(x_1) - \dots - l(x_k)}$$

$$= \sum_{\bar{x} \in X^k} 2^{-l(\bar{x})} \qquad l_i \le l_{max}$$

$$\sum_{\bar{x} \in X^k} 2^{-l(\bar{x})} \le \sum_{m=1}^{k \cdot l_{max}} c(m) \cdot 2^{-m}$$

$c(m) = \#$ slov v C^k s délkou m

$c(m) \le 2^m$.

$$\Rightarrow \left(\sum_x 2^{-l(x)}\right)^k \le \sum_{m=1}^{k \cdot l_{max}} 2^m \cdot 2^{-m} = k \cdot l_{max}$$

$$\sum_x 2^{-l(x)} \le \underbrace{(k \cdot l_{max})^{1/k}}_{k \to \infty} \to 1 \qquad \blacksquare$$

- generování pstního rozdělení pomocí nezávislých hodů $1/2, 1/2$.

- C libovolný kód $\qquad C : X \to \Sigma^*$

  $L(C)$ vs $H(X)$?

  - $H(X) - 2$; $H(X) - 2 \le L(C)$

Dk: z C lze udělat bezprefixový kód

$$C(x) \longrightarrow \underbrace{e(|C(x)|)}_{\text{zakódovaná délka } C(x) \text{ v binárním vyjádření}} \, 01 \, C(x)$$

a zdvojení každého bitu

$0 \to 00$

$1 \to 11$

$\to$ bezprefixový

$\ell(x) \quad \to \ell'(x) = \ell(x) + 2 \lg \ell(x) + 2$

$$H(X) \le \sum p(x)\,\ell'(x) = L(C) + 2\sum p(x) \lg \ell(x) + 2$$

Jensen $\le L(C) + 2 + 2 \lg \sum p(x)\ell(x)$.

$$H(X) \le L(C) + 2 + 2 \lg L(C) \;.$$

Pokud $H(X) \le L(C)$ není co dokazovat

jinak $2 \lg L(C) \le 2 \lg H(X)$

$\Rightarrow \quad H(X) \le L(C) + 2 + 2 \lg H(X) \quad \boxtimes$

Michal Koucky at 28. 3. 2016 21:07

Kolmogorovská složitost

Q: Co činí konečný řetězec $\{0,1\}^*$ náhodný?

3333333 ... 3 $\to 3^{10}$

3141522 6535 $\to \pi$ ... prvních deset číslic

84354279521 $\to$ náhodný

Odlišují se délkou zápisu. Záleží však na jazyku.

f ... částečná rekurzivní funkce $f: \{0,1\}^* \to \{0,1\}^*$

$x \in \{0,1\}^*$

Df: Kolmogorovská složitost x vzhledem k f:

$$C_f(x) = \min\{\,|p|;\ p \in \{0,1\}^*;\ f(p) = x\,\}$$

Věta: Existuje univerzální částečná rekurzivní fce U t.ž. $\forall$ č.r.f. g $\exists c > 0$ t.ž. $\forall x \in \{0,1\}^*$

$$C_U(x) \le C_g(x) + c$$

Dk:

$\phi_1, \phi_2, \ldots$ enumerace všech částečných rekurzivních fcí

ϕ_i ... i "kód"

uvažme párování $\langle i, z\rangle = 0^{|i|} 1\, i z$

U definujeme programem:

na vstupu $\langle i, z\rangle$

dekóduj i a z,

simuluj ϕ_i na z

pokud se zastaví, vytiskni, co vytiskla ϕ_i na z.

end.

pro nějaké j $\quad \phi_j = g$

$c = 2|j| + 1$

Pokud p je optimální program pro x pro g,
pak $\langle j, p\rangle$ je kód pro x pro U ∎

→ U dává nejmenší složitost ze všech čir. fcí
(což nás konstanta)

→ zafixujme nějaké U $\qquad C \equiv C_U$.

Př: $C(x) \leq |x| + O(1)$
$C(0^n) \leq |n| + O(1)$
↳ n binárně ... $O(\lg n)$ bitů
$C(\pi_{1..n}) \leq |n| + O(1)$

$\forall n\ \exists x \in \{0,1\}^n;\ C(x) \geq |x|$
Dk.: ∃ nejvýše $2^n - 1 = \sum_{i=0}^{n} 2^i$
různých pgm. délky $< n$,
ale je 2^n řetězců délky n ∎

Def: x je Kolmogorovsky náhodný, pokud $C(x) \geq |x|$.

· Podmíněná Kolmogorovská složitost
$x, y \in \{0,1\}^*$ $\qquad f$ čir. f.
$C_f(x|y) = \min\{|p|;\ p \in \{0,1\}^*,\ f(\langle p, y\rangle) = x\}$

Věta: $\exists U;\ \forall$ čir. f. $g\ \exists c\ \forall x, y$
$C_U(x|y) \leq C_g(x|y) + c_g$

→ zafixujeme U ... univerzální
$C(x) \overset{def}{=} C(x \mid \varepsilon)$
↳ prázdný řetězec.

· $C(x, y) \overset{def}{=} C(\langle x, y\rangle)$

Věta: $C(x, y) \leq C(x) + C(y|x) + O(\lg C(x,y))$
Dk: pgm $p \xrightarrow{} C(x)$ pro x
pgm q pro y, když znám x
↳ $C(y|x)$
$O(\lg C(x,y))$ pro separaci p a q ∎

Věta: $C(x, y) \geq C(x) + C(y|x) - O(\lg C(x,y))$
Dk: netriviální, viz standardní učebnice
Kolmogorovské složitosti ∎

Kolmogorovská informace

$I_C(x : y) = C(x) - C(x|y)$

Symetrie informace: $I_C(x:y) = I_C(y:x) + O(\lg C(x,y))$

Př: $\forall n\ \exists x \in \{0,1\}^n$ t.ž. $C(x|n) \geq n$
pokud $C(n) \geq \lg n$ pak

$$I_c(x:n) = C(x) - C(x|n) \le n - n = 0$$
$$I_c(n:x) = C(n) - \underbrace{C(n|x)}_{=O(1)} = \lg n \pm O(1)$$

→ logaritmická ztráta nutná
↳ ∃ pgm pro aproximaci psti

<u>Věta</u>: $X_1, X_2, \dots$ je <u>rekurzivní</u> posloupnost pstních distribucí, X_n na $\{0,1\}^n$. Pak

$$H(X_n) - O(\lg n) \le \mathbb{E}[C(X_n)] \le H(X_n) + O(\lg n)$$

<u>Dk</u>: • $\mathbb{E}[C(X_n)] \le H(X_n) + O(\lg n)$

vezmeme Huffmanův kód pro X_n,
$x \in X_n$ má délku $\ell(x)$
$C(x) \le \ell(x) + O(\lg n)$ pgm pro Huf. kód X_n

$$\mathbb{E}[C(X_n)] \le \mathbb{E}_{x\sim X_n}[\ell(x) + 2\lg n + O(1)] \le H(X) + O(\lg n)$$

• $H(X) \le \mathbb{E}[C(X_n)] + O(\lg n)$

pro každé $x \in X_n$, nechť p_x je jeho nejkratší pgm.
$|p_x| \le n + O(1)$.

$d_x = 0^{\lg n + O(1)}\, 1\, |p_x|\, p_x$
↳ $\lg n + O(1)$ bitový binární zápis $|p_x|$

$\{d_x, x \in X_n\}$ ... bezprefixový kód pro X_n

$$\Rightarrow H(X_n) \le \mathbb{E}_{x\in X_n}[|d_x|] = \mathbb{E}_{x\in X_n}[C(x)] + O(\lg n)$$ ∎

<u>Samoopravné kódy</u>

• úvodní přednáška P. Gregor
- podstata
- definice
- Shannonova věta
+ inverzní Shannonova věta

Michal Koucky at 12. 4. 2016 10:43

• <u>Hammingův kód</u>

$$\begin{array}{c|cccc} & x_1 & & \dots & x_n \\ 1 & 1 & 1 & & 1 \\ \vdots & 0 & 1 & & 1 \\ & 0 & 0 & & 1 \\ \ell & 0 & 0 & \dots & 1 \end{array}$$

$n = 2^\ell - \ell - 1$

• očíslujeme si bity binárními čísly t.ž.

obsahují alespoň dvě jedničky

- pro každý z ℓ řádků spočti paritu bitů x_i, které mají v příslušném řádku binárního indexu 1.

$\to$ ℓ parit $a_1 \dots a_\ell$

kód $x_1 \dots x_n \to x_1 \dots x_n a_1 \dots a_\ell$

$$[2^\ell - 1, 2^\ell - \ell - 1, 3]_2$$

- rozšířený Hammingův kód přidá navíc paritu

$$a_0 = \sum_{i=1}^{n} x_i \mod 2 .$$

$$[2^\ell, 2^\ell - \ell - 1, 4]_2$$

dekódování jedné chyby:

- pokud nastane chyba v části $x_1 \dots x_n$, pak alespoň dvě parity a_i a a_j neodpovídají
  chyba nastala v bitu, jehož binární index je přesně vektor $\{0,1\}^\ell$ ukazující, které parity nesedí
- pokud nastala chyba v části $a_1 \dots a_\ell$
  $\Rightarrow$ nesedí právě jedna parita
  a ta je špatná.

- kód lze reprezentovat maticí

Př.:

$$G = \begin{pmatrix} 1 & & & & 0 & 1 & 1 \\ & 1 & & & 1 & 0 & 1 \\ & & 1 & & 1 & 1 & 0 \\ & & & 1 & 1 & 1 & 1 \end{pmatrix}$$

$\ell = 3$
$[7,4,3]_2$
[Hamming 1950]

$x \longrightarrow xG$
$x \in \{0,1\}^4$

$\to$ lineární kód $[n,k,d]_2$

$G \in \{0,1\}^{k \times n}$ ... generující matice

kontrolní matice $H \in \{0,1\}^{n \times n-k}$

$\forall y \quad yH = 0 \text{ iff } \exists x \ \ xG = y$

$\Rightarrow GH = 0^{k \times n-k}$

$\hookrightarrow$ báze duálního vektorového prostoru (ortogonální doplněk)

$\dots H = n - \dim G$

Kontrolní matice Hammingova kódu:

$$H = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

část matice z G ... binární indexy

jednotková matice

dekódování: $xG = y$ $\qquad y' = y + e_i$

$\hookrightarrow (0,0 \dots 1, 0 0)$ ... i-tá pozice

$$y'H = (y + e_i)H = yH + e_iH = e_iH$$

index bitu s chybou

$\hookrightarrow$ syndrom

- obecně pro lineární kód

$$xG = y \qquad y' = y + e$$

$\hookrightarrow$ chybový vektor

$$y'H = yH + eH = eH$$

pro různé chybové vektory chceme různé syndromy

- pokud chceme opravit d chyb musí platit že pro všechny vektory $e \in \{0,1\}^n \setminus \{0^n\}$ s Ham. vahou $\leq d$, eH jsou různé, tj. součet $\leq 2d$ řádků z H musí být nenulový.

$\rightarrow$ tabulka $eH \rightarrow e$ pro snadné dekódování

pro lineární kód platí: $\qquad [n,k,d]_2$

$$\forall x, y \in C \Rightarrow \begin{matrix} x + y \in C \\ x - y \in C \end{matrix}$$

$$(aG = x, \; bG = y \qquad (a+b)G = x + y)$$

$$\Rightarrow \quad d_H(x,y) = |\{ i : (x-y)_i \neq 0 \}|$$

$$= wt_H(x-y) \; \dots$$ Hammingovská váha $x-y$

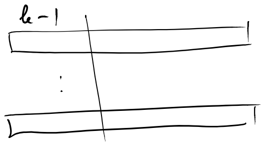
$$d = \min_{x \neq y \in C} d_H(x,y) = \min_{\substack{x \in C \\ x \neq 0}} wt_H(x)$$

$\Rightarrow$ u lineárního kódu se lze zaměřit na minimální váhu nenulových slov pro zjištění minimální vzdálenosti.

Věta (Singletonova): Pro lineární kód C $[n,k,d]_2$ nad tělesem GF_2 platí:

$$n \geq k + d - 1$$

Dk.:



$\exists x, y \in C$ t.ž. $x|_{\{1 \dots k-1\}} = y|_{\{1, \dots k-1\}}$

$d_H(x, y) \le n - (k-1) = n - k + 1$

$d \le n - k + 1$ ∎

Reed - Solomonovy kódy

$GF[q]$, n, k $\qquad n \le q$

$\alpha_1, \alpha_2, \dots \alpha_n \in GF[q]$

$m \in GF[q]^k \longrightarrow p_m(x) = m_0 + m_1 x + \dots m_{k-1} x^{k-1}$

$E(m) = \langle p_m(\alpha_1), p_m(\alpha_2), \dots \; p_m(\alpha_n) \rangle$

- RS kód je lineární:

$$G = \begin{pmatrix} 1 & 1 & & 1 \\ \alpha_0^1 & \alpha_1^1 & \dots & \\ \alpha_0^2 & \alpha_1^2 & & \\ \vdots & \vdots & & \\ \alpha_0^{k-1} & \alpha_1^{k-1} & \dots & \alpha_n^{k-1} \end{pmatrix}$$

$\longrightarrow [n, k, n-k+1]_q$ kód

↑ polynom stupně $k-1$ má nejvýše $k-1$ kořenů, t.j. každé nenulové slovo má Ham. váhu $\ge n-(k-1)$

Př: $k = n-4$ $\qquad [n, k, 5]_q$ kód

každý symbol z GF_q reprezentuji binárně

$\longrightarrow [N, N - 4\log n, 5]_2$ - kód

... binární kód opravující 2 chyby.

$N = n \cdot \log n$.

Michal Koucky at 18. 4. 2016 14:53

[: Konečná tělesa :]

Dekódování Reed - Solomonových kódů

$m = \langle m_0, m_1, \dots m_{k-1} \rangle \qquad \alpha_1, \dots \alpha_n \in \mathbb{F}_q$

$p_m(x) = \sum_{i=1}^{k} m_i x^{i-1}$

$m \longrightarrow \langle p_m(\alpha_1), p_m(\alpha_2), \dots, p_m(\alpha_n) \rangle$ ... kód. zprávy

$\downarrow$

$\langle r_1 \dots r_n \rangle \in \mathbb{F}_q^n$

... přijaté slovo

chceme nalézt polynom stupně $\leq k-1$,
který se nejvíce shoduje s $\langle r_1 \dots r_n \rangle$.

Algoritmus Berlekamp-Welch (1986)

error lokační polynom

$r_1 \dots r_n$ ... víme, že existuje polynom p
stupně $\leq k-1$, který se shoduje s $r_1 \dots r_n$
v alespoň k pozicích $\leq n-k$

chceme nalézt polynom $E(x)$ stupně d_e t. ž.
$\not\equiv 0$

$p(\alpha_i) \neq r_i \Rightarrow E(\alpha_i) = 0 \quad \forall i \in \{1, \dots n\}$

- pokud máme tohle E, dekódování snadné:

Řekneme: $p(\alpha_1) = r_1$ ... $p(\alpha_{n-d_e}) = r_{n-d_e}$

$d_e \leq n-k$.

$$L_{i, \beta_1, \dots \beta_k}(x) = \frac{\prod_{j \neq i} (x - \beta_j)}{\prod_{j \neq i} (\beta_i - \beta_j)} = \begin{cases} 0 & \text{pro } \beta_j \\ 1 & \text{pro } \beta_i \\ * & \text{jinak} \end{cases}$$

... Lagrangeův polynom

$$p(x) = \sum_{i=1}^{k} r_i \cdot L_{i, \alpha_1, \dots \alpha_{n-d_e}}(x)$$

Hledáme $E(x)$:

$\forall i \quad (r_i - p(\alpha_i)) E(\alpha_i) = 0$

$\Rightarrow \quad r_i E(\alpha_i) = p(\alpha_i) E(\alpha_i)$

$0 < d_e \leq \frac{n-k-1}{2} \approx \frac{d}{2}$

$p(x) \cdot E(x)$ je polynom st. $\leq \frac{n-k-1}{2} + k - 1$
$= \frac{n+k-3}{2}$

$$Q(x) = p(x) \cdot E(x) = \sum_{i=0}^{\frac{n+k-3}{2}} c'_i x^i$$

$$E(x) = \sum_{i=0}^{\frac{n-k-1}{2}} c_i x^i$$

$\rightarrow$ soustava $\underline{n}$ lineárních rovnic s $\underline{n}$ neznámými

$$\forall j \in \{1, \dots n\}: \quad \sum_{i=0}^{\dots} c'_i \alpha_j^i = r_j \sum_{i=0}^{\dots} c_i \alpha_j^i \qquad (*)$$

$\rightarrow$ Gaussova eliminace $O(n^3)$
Rychlá Fourierova transformace $O(n \cdot \lg^c n)$
(FFT)

- vezmeme libovolné řešení soustavy $Q(x)$ a $E(x)$ $\not\equiv 0$
- Pokud $E(x)$ nedělí $Q(x)$ $\longrightarrow$ FAIL
(příliš mnoho chyb)
- Spočti $P(x) = \frac{Q(x)}{E(x)}$
- Pokud $d_H(\langle r_1 \dots r_n \rangle, \langle P(\alpha_1), \dots P(\alpha_n) \rangle) > d_e$

⇒ FAIL
(příliš mnoho chyb)

• Výstup: $P(x)$.

<u>Tvrzení:</u> Pokud $(Q_1(x), E_1(x)) \neq (Q_2(x), E_2(x))$
splňují (*) a $E_1(x), E_2(x) \not\equiv 0$ pak

$$\frac{Q_1(x)}{E_1(x)} = \frac{Q_2(x)}{E_2(x)}$$

$Q_1(x)\,E_2(x)$ a $Q_2(x)\,E_1(x)$ mají stupně

nejvýše $\frac{n+k-3}{2} + \frac{n-k-1}{2} = n-2$.

definuj $R(x) = Q_1(x)\,E_2(x) - Q_2(x)\,E_1(x)$

z (*) $Q_1(\alpha_j) = r_j\,E_1(\alpha_j)$ $\quad Q_2(\alpha_j) = r_j\,E_2(j)$

$\Rightarrow \forall j \in \{1..n\}\ R(\alpha_j) = 0.$ (st. $R \leq n-2$ ale n kořenů)

$\Rightarrow R(x) \equiv 0$

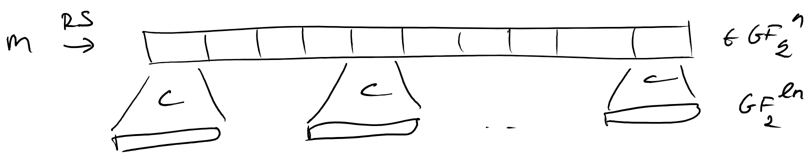
$\Rightarrow Q_1(x)\,E_2(x) = Q_2(x)\,E_1(x)$ v n bodech
jelikož jsou to polynomy st. $\leq n-2$, mají
stejný rozklad na ireducibilní polynomy

$$\Rightarrow \frac{Q_1(x)}{E_1(x)} = \frac{Q_2(x)}{E_2(x)} \qquad \blacksquare$$

• Reed-Solomonovy kódy vyžadují velikost tělesa $\geq n$.

<u>Zobecnění:</u> Reed-Mullerovy kódy – polynomy ve více proměnných

• Jak z R-S kódů udělat binární kódy?
→ kódovat prvky z GF_q binárně pomocí samoopravného kódu.



$l \approx O(\lg q)$ $\quad c \ldots$ kód $[l, \lg q + 1, d]_2$

pokud RS je $[n, k, D]_q$ kód, pak výsledný
kód je $[nl, k \cdot \lg q, dD]_2$ kód.

<u>Dekódování:</u>
1) dekóduj každý vnitřní symbol
2) dekóduj RS-kód

→ opraví $\frac{dD}{4}$ chyb

• lze opravit až $\frac{dD}{2}$ – <u>Forneyho alg.</u>

• RS kód s min. vzdáleností D umí opravit
E chyb a S vymazání, pokud $2E + S < D$.

<u>Forneyho alg.</u>
1) dekóduj ~oprav každý vnitřní symbol $r_i \to r_i'$ ...))

vymaž každou pozici i s pstí $\frac{2e_i}{d}$, (resp. $\min\left(\frac{2e_i}{d}, 1\right)$)

kde $e_i = d_H(r_i, r_i')$.

2) dekóduj zbylé $r_1', \dots r_n'$ pomocí BW alg

Tvrzení: Pokud E je počet nesprávných symbolů r_i' a S je počet smazaných symbolů, pak

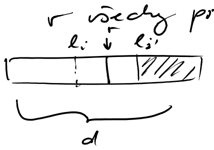
$$\mathbb{E}[2E + S] < D,$$

pokud $\sum e_i < \frac{dD}{2}$.

Dk: cvičení ■ → symbol i přispěje do střední hodnoty $\le \frac{2e_i}{d}$

2 případy: a) $e_i \le d/2$ ... b) $e_i > d/2$ ...

[Derandomizace — lze vybrat společný threshold r v všech pozicích:



]

Náhodný lineární kód: (Varshamov)

- vyber náhodnou 0-1 matici G velikosti $k \times n$

kóduj $x \in \{0,1\}^k \to xG$

Tvrzení: Pokud d je takové, že $2^{k-1} \le \frac{2^n}{Vol_2(n, d-1)}$, pak s velkou pstí $C = \{xG\}$ je kód s minimální vzdáleností d.

- $d = pn$ $\qquad k < n - H(p)n$

Dk: pevné x; xG je náhodný vektor v $\{0,1\}^n$

$$Pr_{G \in \{0,1\}^{k \times n}}[xG \in Vol_2(n, d-1)] \le \frac{Vol_2(n, d-1)}{2^n} \le 2^{(H(p)-1)n}$$

$$Pr_G[\exists x \in \{0,1\}^k, x \neq 0 : xG \in Vol_2(n, d-1)] \le 2^k \cdot 2^{(H(p)-1)n} < 1 \quad ■$$

→ náhodný lineární kód je dobrý $[n, n(1-H(p)), pn]_2$.

(lze použít pro vnitřní kód RS]

prvočísla: (výběr) 251, 257
1021, 1031
4093, 4099

Michal Koucky at 25. 4. 2016 22:01

Gilbertova konstrukce $[n, k, d]_2$

- $C = \emptyset$
- dokud existuje slovo w v $\{0,1\}^n$, které není ve vzdálenosti $< d$ od některého slova v C, přidej w do C.

Pozorování: algoritmus se zastaví po nejdříve

$$\frac{2^n}{Vol(n,d)} \text{ krocích.}$$

$$\Rightarrow |C| \ge 2^{n - H\left(\frac{d}{n}\right)n}$$

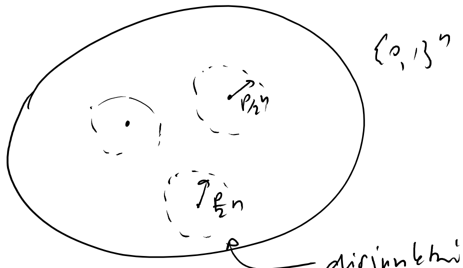
$\to \quad [n, (1-H(p))n, pn]_2$ kód.

- Varshamův kód je lepší v tom, že je lineární, jinak parametry mají stejné.

Hammingova mez: $[n, k, pn]_2$ kód

má $k \le \left(1 - H\left(\frac{p}{2}\right)\right) n$.

Dk:



disjunktní koule
o objemu $Vol(n, \frac{p}{2}n)$
$\geq 2^{H(\frac{p}{2})n}/n$

∎

- srovnej Gilbert-Varshamovy kód

$[n, (1-H(p))n, pn]_2$ kód
opraví $\frac{p}{2}n$ chyb

s Shannonovým kódem

$[n, (1-H(p))n, ?]_2$ opravujícím

pn chyb s velkou pstí

- Dekódování se seznamem (list decoding)

- pro $[n,k,d]_2$ kód a přijaté slovo

$y \in \{0,1\}^n$ hledám všechna slova

do vzdálenosti d' od y.

$\frac{d-1}{2} \le d' \le d$

$\to$ seznam $L \subseteq C \qquad \forall y' \in L, \; d_H(y, y') \le d'$.

pokud d' není příliš velké, seznam L

není příliš velký (je polynomiální v n)

Lokální dekódování

$x \xrightarrow{C} y \rightsquigarrow y'$

chci zjistit souřadnici i tj. x_i, aniž

bych musel dekódovat celé x.

Zároveň ani nechci číst celé y.

Př: Hadamardovy kódy $\quad [2^k, k, \frac{1}{2}\cdot 2^k]_2$

$C: \{0,1\}^k \to \{0,1\}^n \qquad n = 2^k$

$x \in \{0,1\}^n$ $\quad C(x) = y_{0\ldots0} \cdots y_{1\ldots1} \qquad y_a \quad a \in \{0,1\}^k$

$$y_a = \sum_{i=1}^{k} x_i \cdot a_i \pmod 2$$

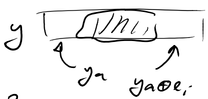
$\forall x, x'$, $x \neq x'$ $\quad d_H(x, x') = \frac{n}{2}$

$y \in \{0,1\}^n$, $i \in \{1, \ldots, k\}$, $d_H(y, C) \leq \frac{n}{6}$

dekódování: náhodně zvol $a \in \{0,1\}^k$

$e_i = (0, 0, \ldots, 1, 0 \ldots 0)$ i-tá pozice

výstup: $y_a \oplus y_{a \oplus e_i}$ $\qquad \leq \frac{1}{6} n$ $\quad \neq C(x)$



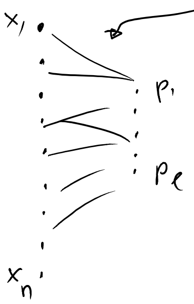
- Pokud $d_H(y, C(x)) \leq \frac{n}{6}$ pak

$$\Pr_{a \in \{0,1\}^n} [y_a \oplus y_{a \oplus e_i} = x_i] \geq 2/3$$

- Opakováním a vzetím většinového výsledku lze prav. chyby snížit

---

- kombinatorické konstrukce kódů



vhodný zvolený graf (expandér)

$[n, n-k, \ldots]_2$

p_i ... parita sousedů – pro kódová slova $= 0$.

- vlastnosti závisí na vlastnostech grafu mezi $\overline{x}$ a $\overline{p}$.
- chyby v kódovém slově přepnou některé parity
- iterativně lze chyby odstranit.

Michal Koucky at 2. 5. 2016 21:22

Komunikační složitost

Alenka	Bob
$x \in \{0,1\}^n$	$y \in \{0,1\}^n$

Alenka a Bob chtějí spočítat nějakou fci

$$f : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$$

na vstupu x, y , tedy $f(x,y)$.

Př.: $EQ(x,y) = [x = y]$ $\qquad x \stackrel{?}{=} y$

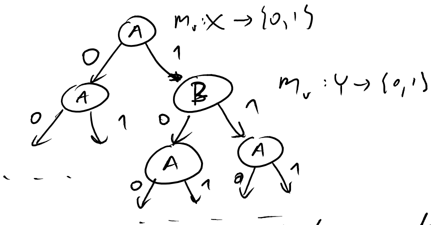
$GT(x,y) = [x \leq y]$ $\qquad x \stackrel{?}{\leq} y$

$IP(x,y) = \sum_{i=1}^{n} x_i y_i \bmod 2$

A & B si vyměňují zprávy o svém vstupu, aby spočítali f.

- Zajímá nás počet bitů potřebných pro spočítání $f(x,y)$. (oba se mají dozvědět výsledek)

- protokol – obecný postup komunikace, lze reprezentovat stromem



- každý uzel je přiřazen jednomu hráči,
  $\forall$ Alicin uzel $\underline{v}$ je přiřazena nějaká fce $m_v : X \to \{0,1\}$, která říká Alici, co má poslat v závislosti na jejím vstupu
  $\forall$ Bobův uzel obdobně.

- listy jsou ohodnoceny výstupní hodnotou fce $f(x,y)$.

délka komunikace = hloubka stromu protokolu.
= cena protokolu

$D(f)$ = minimální délka protokolu pro f.
(minimum přes protokoly počítající f)

- $D(f) \le n+1$. <u>Dk:</u> Alice pošle svůj vstup x, Bob pošle zpět $f(x,y)$

<u>Př:</u> 1) $x, y \subseteq \{1,\dots,n\}$ min

$f(x,y) = \min x \cup y$

$D(f) \le 2 \lg n$

2) medián

$x, y \subseteq \{1,\dots,n\}$

$f(x,y) = \text{medián } x \cup y$

$D(f) \le n + \lg_2 n$ triv

$D(f) \le O(\lg^2 n)$

$D(f) \le O(\lg n)$ těžké!

$D(f) \ge \Omega(\lg n)$ ... triv

Kombinatorický obdélník

$X = \{0,1\}^n \qquad Y = \{0,1\}^n$

$A \subseteq X \qquad B \subseteq Y \qquad A \times B$ … kombinatorický obdélník

- $A \times B \subseteq X \times Y$
- $R \subseteq X \times Y$ je kombinatorický obdélník

  $\iff \forall (x,y), (x',y') \in R: (x,y') \in R.$

Dk: "⇐" $A = \{x;\ \exists y\, (x,y) \in R\}$ "⇒" triv.

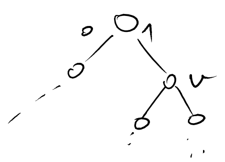
$B = \{y;\ \exists x\, (x,y) \in R\}$

- $R \subseteq A \times B$ triv
- $A \times B \subseteq R$: $(x,y) \in A \times B \Rightarrow \exists x', y'$

  $(x,y') \in R \,\&\, (x',y) \in R$

  $\underset{\text{předpoklad}}{\Longrightarrow} (x,y) \in R$ ∎

Protokol P:



$R_v = \{(x,y);$ na vstupu x,y Alice & Bob dojdou do $v\}$

- $\forall v, R_v$ je kombinatorický obdélník.

Dk: 1) indukcí podle hloubky v.

nebo 2) "cut-and-paste" argument ∎

- Pokud P počítá f, pak pro každý list l v P, vstupy v R_l mají stejnou hodnotu $f(x,y)$.

  → jednobarevný obdélník

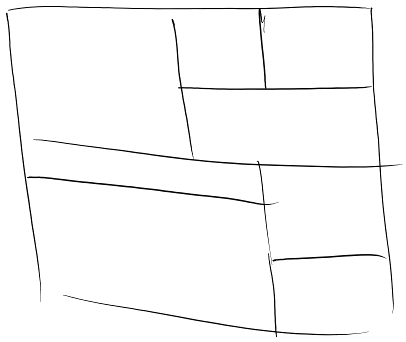
$$M_f \quad \begin{array}{c|ccccccc|} & 00\ldots0 & & & & & & 11\ldots1 \\ \hline 00\ldots0 & 0 & 1 & \boxed{0} & 1 & 1 & \boxed{0} & 1 \\ \vdots & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ \vdots & 0 & 1 & \boxed{0} & 1 & 1 & \boxed{0} & 1 \\ & 1 & 1 & 1 & 0 & 1 & 1 & 0 \\ 11\ldots1 & & & & & & & \\ \hline \end{array}$$

$f(x,y)$

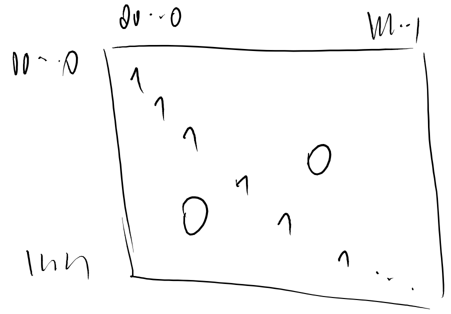
jednobarevný obdélník

→ každý protokol dává pokrytí matice f jednobarevnými obdélníky

• Pokud fce f vyžaduje na pokrytí t jednobarevných obdélníků, pak $D(f) \geq \log_2 t$.



Př: 1) $EQ(x,y) = [x \stackrel{?}{=} y]$



$\geq 2^n+1$ obdélníků – žádné dvě jedničky nemohou být ve stejném obdélníku.
jednobarevném

$\Rightarrow D(EQ) \geq \log 2^n+1 > n$

$D(EQ) \leq n+1 \Rightarrow D(EQ) = n+1$. □

2) $DISJ(x,y) = [\exists i: x_i = y_i = 1?]$

dvojice $(x, \overline{x})$ musí být v různých jednobarevných obdélnících

$\Rightarrow \geq 2^n+1$ obdélníků

$D(DISJ) \geq n+1 \Rightarrow D(DISJ) = n+1$ □

• Pokud hodnost matice M_f je alespoň r, pak $D(f) \geq \log r$.

Dk: vezmeme si protokol P pro f.

$$M_f = \sum_{\substack{\ell \text{ list } P \\ \text{pro hodnotu } 1}}' M_\ell$$

$$M_\ell(x,y) = \begin{cases} 1 & (x,y) \in R_\ell \\ 0 & \text{jinak} \end{cases}$$

hodnost $(M_\ell) \le 1$

$\Rightarrow$ hodnost $M_f \le$ počet listů P $\underset{r}{\underset{\parallel}{}}$

$\log r \le D(f)$ $\boxtimes$

Př: 1) hodnost $(M_{EQ}) = 2^n$

2) hodnost $(M_{IP}) \ge 2^n - 1$

$\Rightarrow D(IP) \ge n.$

$$(M_{IP})^2 = \begin{pmatrix} & 0 & \\ 2^{n-1} & \cdots & 2^{n-2} \\ & \ddots & \\ 2^{n-2} & \cdots & 2^{n-1} \end{pmatrix}$$

Michal Koucky at 10. 5. 2016 11:37

Příklady pokrytí matic M_f obdélníky

Př: 1)

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

$\exists$ pokrytí 5 obdélníky, ale neexistuje protokol pokrývající právě 5 obdélníky.

2)

$$\begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix}$$

$\exists$ překrývající se pokrytí 4 obdélníky

→ různé míry pohledu

$C^P(f)$ ... nejmenší možný počet obdélníků daných protokolem pro f

$C^D(f)$ ... nejmenší možný počet nepřekrývajících se obdélníků pokrývající matici M_f

$C(f)$ ... nejmenší možný počet obd. pokrývající M_f

• Pozorování: $C(f) \le C^D(f) \le C^P(f) \le 2^{D(f)}$

→ nedeterministický komunikační protokol:

P ... všemocný dokazovatel, který zná x & y

m ... důkaz, že $f(x,y) = z$

$A \xleftarrow{m} P \xrightarrow{m} B$

A: $x \in \{0,1\}^n$

B: $y \in \{0,1\}^n$

Alenka a Bob si vymění 1 bit každý, zda souhlasí s důkazem.

Př: $EQ(x,y) = 0$ ... m je index bitu, kde se liší

$|m| = \log_2 n$

$EQ(x,y) = 1$ ... m je celý řetězec x.

$|m| = n$ bitů

• každá zpráva $\underline{m}$ od P definuje z-barevný obdélník

počet možných zpráv = # pokrývajících obdélníků

...

• $N(f) = \log_2 C(f)$ ... nedeterministická složitost f

$N^1(f) = \log_2 C^1(f)$

$N^0(f) = \log_2 C^0(f)$

$C^z(f)$ ... nejmenší počet z-barevných obdélníků

Př: $N^0(EQ) = \lg n$ $\quad N^1(EQ) = n$

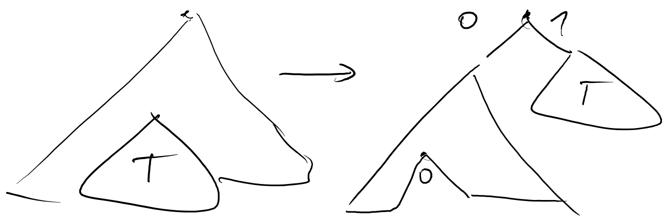
$N^0(DISJ) = n$ $\quad N^1(DISJ) = n$

C vs D

Lema: $\lg_2 C^P(f) \leq D(f) \leq 2 \lg_{3/2} C^P(f)$

Dk: 1. "$\leq$" triv

2. "$\leq$" ... v protokolu nalezni podstrom s $\frac{1}{3} \leq \quad \leq \frac{2}{3}$ listů a přesuň na vrch



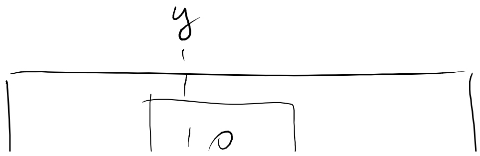
→ vyvážení protokolu ∎

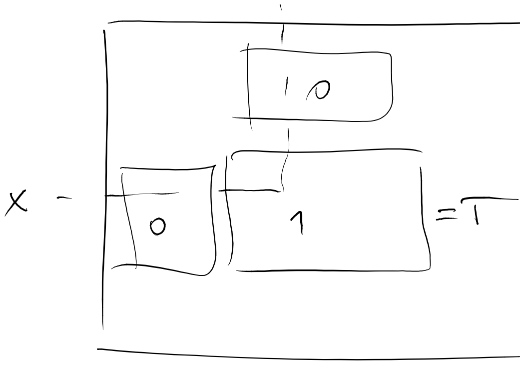
Otázka: $D(f) = O(\log C^D(f))$?

Věta: $D(f) = O(N^0(f) \cdot N^1(f))$.

Dk: idea:

řekněme, že $f(x,y) = 1$





· každý 0-obdélník může protínat T pouze buď v řádcích, nebo sloupcích

Protokol: A & B udržují množinu "živých" 0-obd.

1) Pokud všechny 0-obd. mrtvé, A vyhlásí $f(x,y)=1$
2) A se podívá, zda $\exists$ 1-obd., který obsahuje sloupec y a ve sloupcích protíná nejvýš polovinu živých 0-obd. Pokud ano, pošle číslo 1-obd. Bobovi. Jinak mluví Bob
3) B se podívá, zda $\exists$ 1-obd. obsahující řádek x, který v řádcích protíná $\leq 1/2$ živých 0-obd. Pokud ano, pošle jeho číslo Alence. Jinak skončí s tím, že $f(x,y)=0$.

$\rightarrow$ maximálně $\log C^0(f)$ kol a každé kolo stojí $O(\log C^1(f))$ bitů $\boxtimes$

Otázka: $\forall f: D(f) = O(\log \text{hodnost}(M_f))^{O(1)}$?

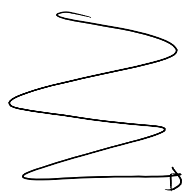
... "log-rank conjecture"

Pravděpodobnostní protokoly

$r_A \in \{0,1\}^*$ Alenka —— Bob $r_B \in \{0,1\}^*$

$x \in \{0,1\}^n$ $y \in \{0,1\}^n$

$r_A \in \{0,1\}^*$ Alenka $x \in \{0,1\}^n$



Bob $y \in \{0,1\}^n$ $r_B \in \{0,1\}^*$

- zprávy Alenky můžou záviset též na r_A
- –||– Boba –||– r_B

→ ve stromu protokolu v uzlu v patřícím Alence máme fci $m_v : \underset{x}{\{0,1\}^n} \times \underset{r_A}{\{0,1\}^*} \to \underset{m_v}{\{0,1\}}$

Bobovi –||– $m_v : \underset{y}{\{0,1\}^n} \times \underset{r_B}{\{0,1\}^*} \to \{0,1\}$

– listy opět obsahují výstupní hodnotu

<u>chceme:</u> protokol <u>P počítá f s chybou</u> $\varepsilon \geq 0$ pokud

$$\forall x,y \quad \Pr[P(x,y) = f(x,y)] \geq 1-\varepsilon$$

- zajímá nás průměrná délka komunikace na daném vstupu x,y. Cena protokolu je nejdelší (nejhorší) průměrná délka na nejhorším x,y.

- $R_\varepsilon(f)$ = minimální cena protokolu P, který počítá f s chybou $\leq \varepsilon$.

$$R(f) := R_{1/3}(f)$$

Všimněte si: kdyby byla cena definována pomocí nejdelší komunikace na x,y, v zásadě –||–

by se moc nezměnila, neboť každý protokol lze zařizovat po $1/\varepsilon$-násobku průměrné délky a to zhorší chybu nejvýše o ε.

Př: 1) $R(EQ) = O(\log n)$

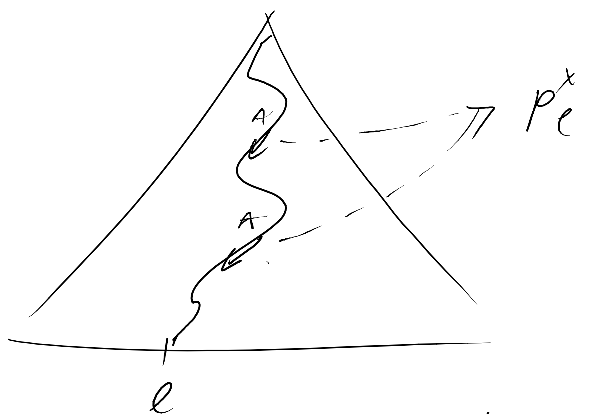
Michal Koucky at 16. 5. 2016 22:29

Lema: $R(f) \geq \Omega(\log D(f))$

Dk: ukážeme, že $D(f) \leq 2^{O(R(f))} \cdot O(R(f))$

- uvažme pstní protokol pro f s chybou $\leq 1/3$ a maximální hloubkou $d = O(R(f))$.
- Protokol má nejvýše 2^d listů.

Na vstupu x a y je pravděpodobnost p_ℓ dosažení listu ℓ dána součinem pravděpodobností p_ℓ^x a p_ℓ^y, kde p_ℓ^x je pravděpodobnost, že Alenka v jejích uzlech na vstupu x jde směrem k ℓ a podobně p_ℓ^y pro Boba.



p_ℓ^x je samo součinem pstí v jednotlivých Alenčiných uzlech a stejné p_ℓ^y.

Det. protokol pro f: Alenka spočítá poči p_ℓ^x pro
všechny listy ℓ a pošle tuto informaci
Bobovi. Ten spočítá svoje p_ℓ^y a zjistí
výsledné poč. jednotlivých výstupů. Vydá
odpověď s nejvyšší poč.

Každou z hodnot p_ℓ^x Alenka pošle s přesností
$d+10$ bitů, tj. zaokrouhlovací chyba
$2^{-(d+10)}$. Celková chyba při výpočtu poč.
jednotlivých odpovědí je tak $\leq 2^d \cdot 2^{-(d+10)}$
$\leq \frac{1}{1000}$

Komunikace vyžaduje $\leq 2^d \cdot (d+10) + 1$ bit

- protokol s veřejnými náhodnými bity:

$$\underline{\text{Alenka}} \quad \xleftarrow{\quad r \in \{0,1\}^* \quad} \quad \xrightarrow{} \quad \underline{\text{Bob}}$$

$x \in \{0,1\}^n$ $\qquad\qquad$ $y \in \{0,1\}^n$

- Alenka i Bob dostanou zadarmo společný náhodný
řetězec r.

Věta: protokol s veřejnými náhodnými bity lze simulovat
se soukromými posláním $O(\log n)$ bitů navíc.
(Dojde k mírnému nárůstu chyby.)

Dk: idea: lze zvolit množinu R $n^{O(1)}$ řetězců r t.ž.

chyba protokolu na každém vstupu (x, y) se
při použití zcela náhodného r a náhodného r
vybraného z naší množiny R liší právě
o $\leq \frac{1}{n^{O(1)}}$. [Toto plyne z Černovovy nerovnosti
při zvolení si množiny R zcela náhodně.]

Simulace protokolu s veřejnými náhodnými bity
pak probíhá tak, že Alenka pomocí svých R_A
vybere náhodný řetězec z R a komunikuje ho Bobovi.
Index tohoto řetězce lze kódovat $\log |R| = O(\log n)$ bity $\blacksquare$

Př: EQ s veřejnými bity vs soukromými.

$$O(1) \quad \text{vs.} \quad \Theta(\log n)$$

$$\uparrow$$

$$R(f) \geq \Omega(\log D(f))$$

$$\Omega(\log n)$$

Věta: $R(\mathrm{DISJ}) \geq \Omega(n)$.

Použití komunikační složitosti

Data streams

Výpočetní model

Alg.

- Algoritmus má pouze omezenou paměť, nedokáže si zapamatovat celý vstup.

- Př: · dána jsou celá čísla z intervalu $[0, n]$
  · chci znát:
    1) celkový součet — snadné
    2) průměr — snadné
    3) počet různých prvků — lze s malou pamětí pravděpodobnostním algoritmem (aprox.)
    4) kolikrát se vyskytl nejčastější prvek — vyžaduje paměť $\Theta(n)$.

Dk: víme $D(DISJ) \geq n$.

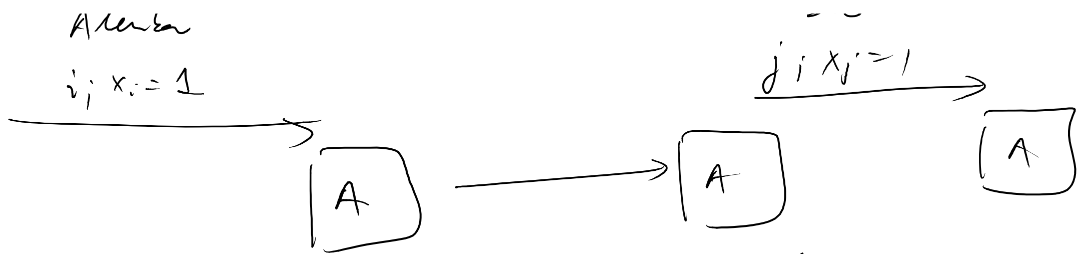
ukážeme, že dobrý algoritmus pro 4) s malou pamětí implikuje příliš efektivní protokol pro DISJ.

- mějme alg. A pro 4). Protokol pro A & B funguje následujícím způsobem:

Alenka $x \in \{0,1\}^n$ — Bob $y \in \{0,1\}^n$

- Alenka vytvoří posloupnost prvků i t.ž. $x_i = 1$; tuto posloupnost zpracuje alg. A.

Alenka $i;\ x_i = 1$ — Bob $j;\ x_j = 1$



pošle stav paměti algoritmu A pošle Bobovi.
Ten vytvoří posloupnost $j;\ y_j = 1$ a
zpracuje ji algoritmem.

Pozorování: $DISJ(x,y) = 1 \iff$ nejčastější
prvek se v datech vytvářejících
Alenkou a Bobem vyskytuje
právě dvakrát.

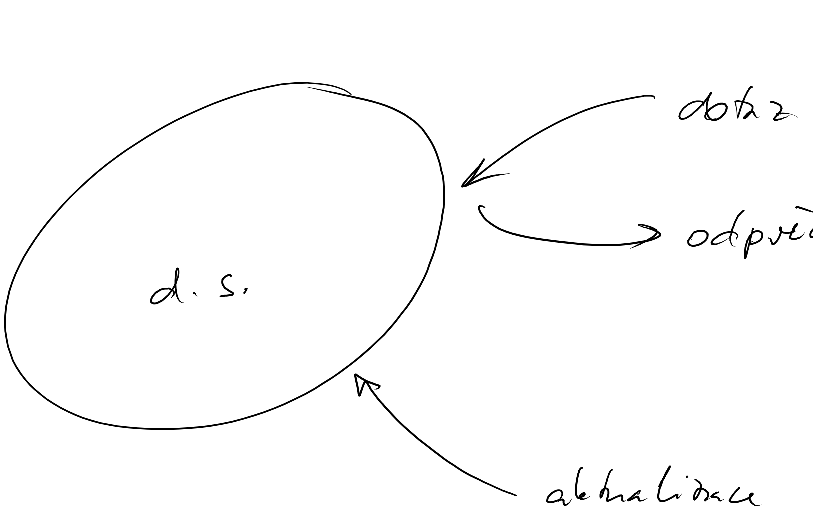
→ objem komunikace mezi Alenkou a Bobem
je velikost paměti použité algoritmem.
(+1 bit na sdělení výsledku Alice)

→ A musí použít paměť $\Omega(n)$ bitů. ∎

Stejné tvrzení s identickým důkazem platí pro
pstní algoritmy

Michal Koucky at 26. 5. 2016 9:39

Datové struktury



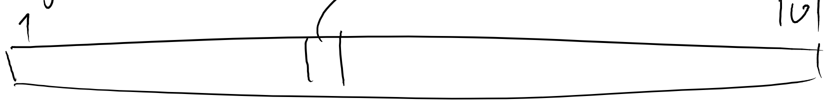
datová struktura uchovává nějaká data a odpovídá o nich dotazy

Př: d.s. pro množinu $S \subseteq U$ $\quad U=\{1,\dots,|U|\}$.

dotazy typu $[x \overset{?}{\in} S]$ pro každé $x \in U$.

možné implementace

1) bitový vektor $\quad$ 0/1 – $[x \in S]$



dotaz ... čas $O(1)$

paměť $|U|$ $\quad$ špatné, pokud $|S| << |U|$

2) hašovací tabulka

$O(|S|)$

čas na dotaz ... $O(1)$

paměť $O(|S|)$

Pozn: není $O(1)$ jako $O(1)$

$\uparrow$ 1-bitů 1) $\qquad$ $\uparrow$ $O(\log n)$-bitů 2)

- Chci datovou strukturu pro lineární prostory

d.s. uchovává $V \subseteq GF_2^n$, V lineární prostor

dotazy: $[y \overset{?}{\in} V]$; $y \in GF_2^n$

řešení: triviální – pole odpovědí na všechny možné dotazy $\rightarrow$ prostor 2^n bitů

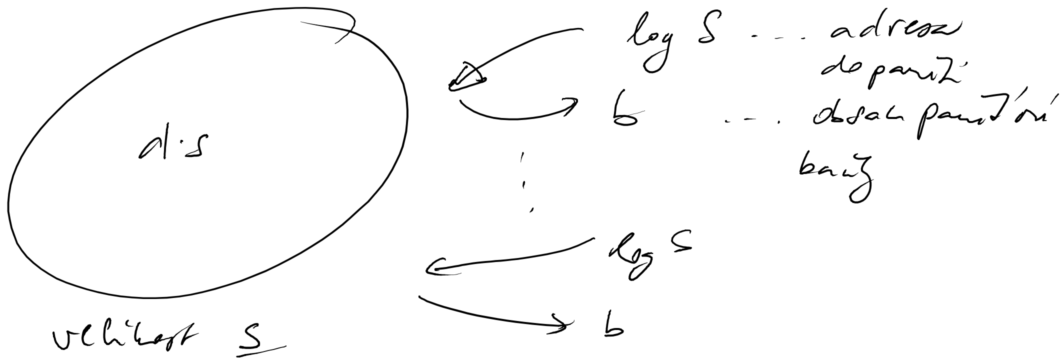
čas na dotaz $O(1)$

- V se dá popsat n vektory, každý potřebuje n bitů $\rightarrow$ stačí n^2 bitů na popis V

báze V

Otázka: $\exists$ d.s., která by měla $n^{O(1)}$ bitů
a „rychle" zodpovídala dotaz $[y \overset{?}{\in} V]$?

měříme: # přístupů do paměti d.s.
při každém přístupu přečteme b bitů



celkem t přístupů

Řekněme, že $b = n$. Kolik přístupů potřebujeme,
když $S = n^{O(1)}$?

odbočka:

Alenka $\qquad\qquad$ Bob

$y \in GF_2^n$ $\qquad$ $V \subseteq GF_2^n$

V podprostor

$[y \overset{?}{\in} V]$

Alenka komunikuje a bitů . Kolik musí být
Bob komunikuje b bitů $\qquad$ a & b ?

Tvrzení: $a \geq n/6$ nebo $b \geq n^2/12 - n/6$

důkaz níže.

zpět k datové struktuře:

datová struktura pro $[y \overset{?}{\in} V]$, kde na každý
dotaz potřebujeme nejvýše t přístupů
. . . . $[y \overset{?}{\in} V]$

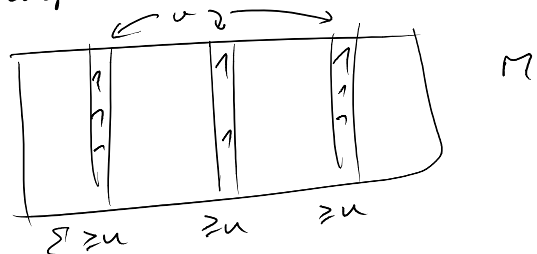
další ... pro ...

dotaz potřebujeme nejvýše t přístupů do paměti dává komunikační protokol pro $[y \in V]$ kde Alenka pošle $t \cdot \log S$ bitů a Bob $t \cdot b$.

(Tvrzení) $\Rightarrow$ d.s. pro $[y \in V]$ s $b = n$ vyžaduje alespoň $\Omega(n / \log s)$ přístupů do paměti na dotaz, tj. $\Omega(n/\log n)$ pro $s = n^{O(1)}$.

Dk.:

matice M je (u,v)-těžká, pokud obsahuje alespoň v sloupečků s alespoň u jedničkami.



Tvrzení: Pokud f je fce s (u,v)-těžkou maticí M_f a protokol pro f, kde Alenka pošle a bitů a Bob b, pak M_f obsahuje jednobarevný 1-obdélník o rozměrech $\geq \frac{u}{2^a} \times \frac{v}{2^{a+b}}$.

Dk: indukce podle $a+b$

1) $a + b = 0$

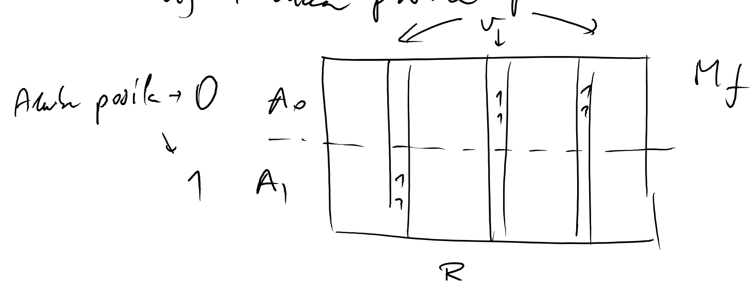
zjevně $M_f \equiv 1$ protože A & B nepotřebují komunikovat

$\Rightarrow M_f \geq u \times v$ ✓

2) $a+b-1$ ✓ $\Longrightarrow$ $a+b$

dva případy

a) Alenka posílá první bit



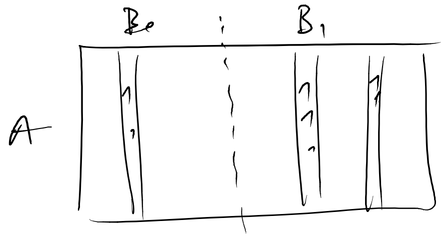
buď podmatice $A_0 \times B$ nebo $A_1 \times B$

je $\left(\frac{u}{2}, \frac{v}{2}\right)$ - těžká

$\underset{\text{indukce}}{\Longrightarrow}$ M_f obsahuje $\frac{\frac{u}{2}}{2^{a-1}} \times \frac{\frac{v}{2}}{2^{(a-1)+b}}$ 1-obd.

$= \frac{u}{2^a} \times \frac{v}{2^{a+b}}$ ✓

b) Bob posílá první bit



buď $A \times B_0$ nebo $A \times B_1$ je $\left(u, \frac{v}{2}\right)$ - těžký

$\underset{\text{indukce}}{\Longrightarrow}$ $\frac{u}{2^a} \times \frac{\frac{v}{2}}{2^{a+(b-1)}} = \frac{u}{2^a} \times \frac{v}{2^{a+b}}$ 1-obd. ☒

• Matice $[y \stackrel{?}{\in} V]$

je 1) $\left(2^{n/2}, 2^{n^2/4}\right)$ - těžká

2) neobsahuje jednobarevný 1-obdélník velikosti $2^{n/3} \times 2^{n^2/6}$.

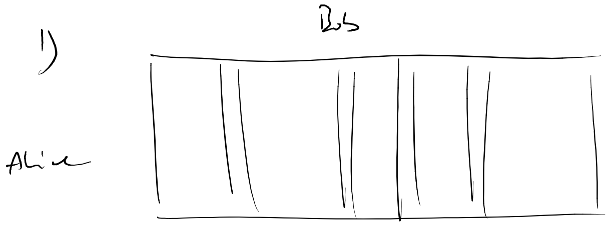
1)&Tvrzení $\Longrightarrow$ pokud by existoval protokol, kde Alice posílá $n/6$ bitů a Bob $\frac{n^2}{12} - n/6$, pak by matice $[y \stackrel{?}{\in} V]$ obsahovala 1-obd. velikosti $\frac{2^{n/2}}{2^{n/6}} \times \frac{2^{n^2/4}}{2^{n/6 + n^2/12 - n/6}} =$

$= 2^{n/3} \times 2^{n^2/6}$ což by byl spor s 2).

Tedy $a \geq n/6$ nebo $b \geq \frac{n^2}{12} - n/6$

Zbývá dokázat 1) a 2)

1)



V t.ž. $\dim V = \frac{n}{2}$ obsahují $2^{n/2}$ různých vektorů z GF_2^n, tedy jsou $2^{n/2}$-těžké

$\# V$; $\dim V = \frac{n}{2}$ $\geq 2^{n^2/4}$:

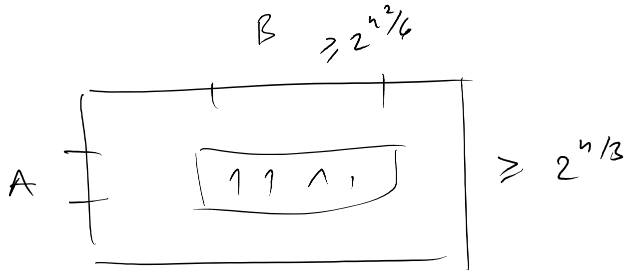
#bází vel. $\frac{n}{2}$ (lin. nezávislých $\frac{n}{2}$-tic vektorů) $\longrightarrow$

$$\frac{\prod_{i=0}^{n/2-1} (2^n - 2^i)}{\prod_{i=0}^{n/2-1} (2^{n/2} - 2^i)} = \prod_{i=0}^{n/2-1} \frac{2^n - 2^i}{2^{n/2} - 2^i} \geq \prod_{i=0}^{n/2-1} 2^{n/2} \geq 2^{n^2/4}$$

#bází prostoru V dim $\frac{n}{2}$

$\Rightarrow \# V$ dim $n/2$ $\geq 2^{n^2/2} / 2^{n^2/4} = 2^{n^2/4}$

2)



A obsahuje alespoň $n/3$ lineárně nezávislých vektorů.

$\forall V \in B$; $\langle A \rangle \subseteq V$

$\Rightarrow$ stačí vybrat dalších $n/6$ vektorů na získání báze V. $\Rightarrow |B| \leq (2^n - 1)^{n/6} < 2^{n^2/6}$ ∎