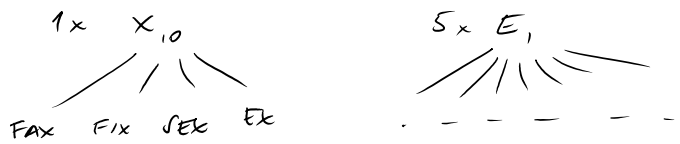
- Domácí úkoly – 4-5 70% bodů na zápočet
- cvičení pátek 10:40 – nepovinné

Plán – Informace & komprese
– samoopravné kódy
– komunikační složitost

Informace

Scrabble



"Přidám sloup obsahující X_{10}" vs.
"Přidám sloup 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 sněděl rozmístěný koberec "

- překvapení vs význam
  → více překvapení = více informace
  - význam neuvažujeme analyzovat
  → informaci přiřazujeme k pravděpodobnostním událostem (např. vygenerované zprávy)

Informace – chceme

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

1) $I(A)$ klesá s rostoucí pstí. A

Př.: Zamíchám balíček 32 karet, vezmu kartu

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 pevné zvolené $a > 1$.

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

- Alternativní přístup – Kolmogorovská složitost
(uvidíme později)

Opakování pojmů:

- pravděpodobnostní prostor Ω ... konečné 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:
$$Pr[A \mid B] = \frac{Pr[A \& B]}{Pr[B]}$$

Př: A & B jsou nezávislé $\Leftrightarrow Pr[A|B] = Pr[A]$.

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

Př: náhodná proměnná X definuje různé jevy
$\forall S \subseteq R \to$ jev $[X \in S]$.
dvě náhodné proměnné X, Y jsou nezávislé, jestliže $\forall S, S' \subseteq R$ $[X \in S]$ 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 $\in \{0,1\}^n$

(PN)

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

2) vybereme náhodné řetězce $a, b, c \in \{0,1\}^{n/2}$
položíme $X = ab$ $Y = bc$

(P2) $\forall x,y \in \{0,1\}^n \quad Pr[X=x] = \frac{1}{2^n}$

$Pr[Y=y] = \frac{1}{2^n}$

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

$\Rightarrow$ 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 $\forall k \in \mathbb{R}$

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

Entropie (nejistota) 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: $\forall \ 0 < p < 1 \quad \log_2 p < 0$

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

- $H(X) \leq \log \underbrace{|X|}_{\hookrightarrow |supp(X)|}$ (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) $p(0^n) = \frac{1}{2} \qquad \forall x \neq 0^n \quad p(x) = \frac{1}{2^{n+1}-2}$

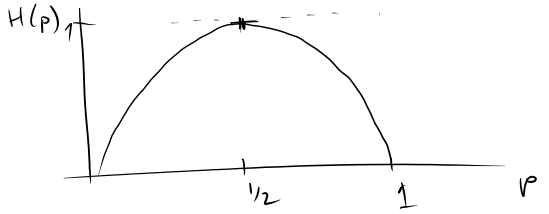
$$H(X) = \frac{1}{2} \lg 2 + \frac{1}{2} \lg (2^{n+1}-2) = \frac{n}{2} + \Theta(1)$$

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

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

$H(p)$ ↑

$X = \{1, 0\}$ s pr. $1-p$



společná entropie X, Y náhodné proměnné

$$H(X,Y) = \sum_{\substack{x \in X \\ y \in Y}} p(x,y) \cdot \log \frac{1}{p(x,y)}$$
$$= \mathop{\mathbb{E}}_{x,y} \log \frac{1}{p(x,y)}$$

Př: (P1) nezávislé X, Y
$H(X,Y) = 2n$

(P2) závislé X, Y
$H(X,Y) = \frac{3}{2}n$

- $\pi_x, \pi_y : \{0,1\}^n \to \{0,1\}^n$ ... pevně zvolené permutace
$X' = \pi_x(X)$ $Y' = \pi_y(Y)$ $H(X',Y') = \frac{3}{2}n$.

podmíněná entropie

$$H(Y|X) = \sum_{\substack{x \in X \\ p(x) \neq 0}} p(x) \cdot H(Y \mid X = x)$$
$$= \sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) \log \frac{1}{p(y|x)}$$
$$= \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{1}{p(y|x)}$$
$$" = \mathop{\mathbb{E}}_{x,y} \log \frac{1}{p(Y|X)} "$$

Př:
1) $H(Y|Y) = 0$
2) $H(Y|X) = H(Y)$ pokud X a Y jsou nezávislé
3) (P1) X, Y $H(Y|X) = H(Y) = n$
4) (P2) X, Y $H(Y|X) = \frac{n}{2}$
$H(Y) = n$

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 proměnné $X_1, X_2, \dots X_k$

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

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 z různých množin o stavu události

$N_1, N_2, \dots N_k$ k periodik

kolik informace o zprávě N_1 nám jistí po přečtení N_2, N_3, atd.

Def: náhodné proměnné X, Y

$I(X:Y) = H(X) - H(X|Y)$ ...

vzájemná informace X a Y

... o kolik se sníží nejistota 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

$Z = X + Y \pmod 6$

$I(X:Y) = 0 \quad I(X:Z) = 0 \quad I(Y:Z) = 0$

$I(X,Y:Z) = H(X,Y) - H(X,Y|Z) = H(Y) = H(X)$

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

$H(X) \qquad = 0$

Vlastnosti:

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

Def: Kullback-Leiblerova vzdálenost

náhodné proměnné X, Y se sdíleným oborem hodnot

$p(x) = Pr[X = x] \qquad q(x) = Pr[Y = x]$

$$D(X \| Y) = D(p \| q) = \sum_x p(x) \lg \frac{p(x)}{q(x)}$$

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

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

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

víme $D(p\|q) \geq 0$

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

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

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

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

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

$= S\left(p(x,y) \,\|\, \underbrace{p(x|y) \cdot p(y)}_{''\,p(x,y)}\right)$ ∎

$$= \sum_{x,y} p(x,y) \log \frac{p(x|y)\cdot p(y)}{p(x)\cdot p(y)} \qquad // \; p(x,y)$$

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

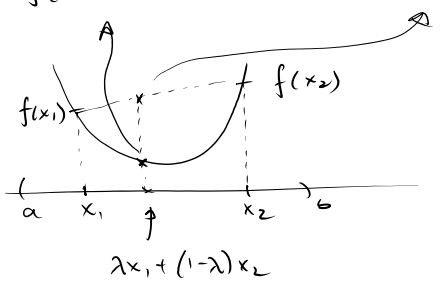
Dk: plyne z (*) výše.

Michal Koucky at 14. 3. 2016 21:28

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

$\forall\, 0 \le \lambda \le 1$

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



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

$$\mathbb{E}[f(x)] \ge f(\mathbb{E}[x])$$

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) $p(x_1) = p_1, \ldots \qquad p(x_n) = p_n$ $\quad (|supp(x)| = n)$

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

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

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

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

$$= f\Big(\sum_{i=1}^{n} p_i x_i\Big) \qquad \blacksquare$$

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

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

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

$$-D(p\|q) = -\sum_{x\in A} p(x) \lg \frac{p(x)}{q(x)}$$
$$= \sum_{x\in A} p(x) \lg \frac{q(x)}{p(x)}$$
$$\le \lg \sum_{x\in A} p(x) \frac{q(x)}{p(x)}$$
$$\le \lg \sum_{x\in X} q(x)$$
$$= \lg 1 = 0$$ ▣

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

Důsledek: $I(X:Y)=0$

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

Důsledek: $H(X) \le \lg |X|$ s rovností, právě pokud X je rovnoměrně rozdělená.

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

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

$0 \quad \Rightarrow \lg|X| \ge H(X)$ ▣

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

Dk: $0 \le I(X:Y) = H(X) - H(X|Y)$ ▣

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

$Y = \begin{cases} 0 & x = 0^n \\ 1 & x \neq 0^n \end{cases}$ $\quad H(X) = \frac{n}{2} + O(1)$

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

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

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

Alenka — Bob

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

Z ... odhad X

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

Def: X, Y, Z splňují Markovovu vlastnost pokud $\forall x,y,z$

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

"$X \to Y \to Z$"

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

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

... symetrie

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

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

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

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

$\Rightarrow I(X:Z) \leq I(X:Y)$ $\blacksquare$

Kódování

- $C : X \to \Sigma^*$ (abeceda) chci $\forall x \neq y$ $C(x) = C(y)$
  ... kód
- průměrná délka kódu C : $L(C) = \sum_{x \in X} p(x)\, l(x)$
- rozšíření kódu $C^*(x_1, \dots x_k) = C(x_1) C(x_2) \dots C(x_k)$

- kód je jednoznačně dekódovatelný, pokud C^* není kódem.

- bezprefixový kód C: $\forall x \neq y$ $C(x)$ není prefix $C(y)$

Př.:

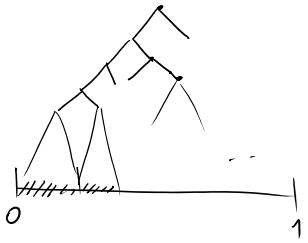
10	0
00	10
11	110
110	111
↑ jednoznačně dekódovatelný	↑ bezprefixový

Věta: (Kraftova nerovnost)
Pro bezprefixový kód C s délkami kódů $l_1, l_2, \dots$ platí:

$$\sum 2^{-l_i} \leq 1$$

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

Dk:



ke kódovému slovu $a_1 a_2 \dots a_k$ přiřadíme interval $[0.a_1a_2\dots a_k,\ 0.a_1a_2\dots a_k + 2^{-k})$

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

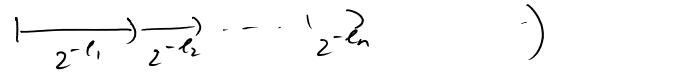
$$\sum 2^{-l_i} \leq 1.$$

Podobně pro obecnou abecedu Σ. ∎

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

(Lze přiřazovat od nejkratší 2^{-l_i} do nejdelší:

$$l_1 \leq l_2 \leq l_3 \dots \leq l_n$$



)

• Optimální bezprefixový kód pro X, chci minimalizovat průměrnou délku slova.

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

$$\ell(x) = |C(x)|$$

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

$$L - H(x) = \sum_x p(x)\ell(x) - \sum_x p(x)\frac{1}{p(x)} =$$

$$= \sum_x p(x) \lg \frac{1}{2^{-\ell(x)}} - \sum_x p(x)\frac{1}{p(x)} =$$

$$= \sum_x p(x) \lg \frac{p(x)}{2^{-\ell(x)}} = (*)$$

kraft.

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

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

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

$$\Rightarrow \quad L \ge H(x)$$ ∎

Shannonův kód:

$$\ell(x) = \lceil \lg \frac{1}{p(x)} \rceil$$

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

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

optimální ± 1 bit.

• kódování k symbolů ze slova $X_1 \sim X$

$X_1, X_2, \dots X_k$ nezávisle

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

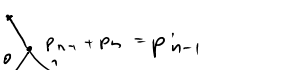
$k \cdot H(X)$ → průměrně $\frac{1}{k}$ bitu navíc za každý symbol.

Př: $p(x_1) = 0.9999$ $\quad p(x_2) = 0.0001$

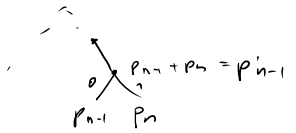
$\ell(x_1) = 1$ $\quad \ell(x_2) = 14$ (ve čísle :)

Huffmanův kód: opakovaně spojuj dva nejmenší pstí a slov stromu od listů.

$p_1 \ge p_2 \ge \dots p_n$



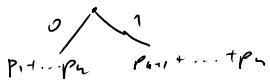
$p_1 \geq p_2 \geq \dots p_n$



→ dá se ukázat, že dobré nejmenší množiny průměrnou délku

Fanův kód: stav stromu odshora, minimalizuj rozdíl v počtu vlevo a vpravo.

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



Fakt: průměrná délka slov $\leq H(X) + 2$

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

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

Dk: Uvažme C^k $k \geq 1$

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

$$= \sum_{\bar{x} \in X^k} 2^{-\ell(\bar{x})} \qquad \ell_i \leq \ell_{max}$$

$$\sum_{\bar{x} \in X^k} 2^{-\ell(\bar{x})} \leq \sum_{m=1}^{k \cdot \ell_{max}} c(m) \cdot 2^{-m}$$

$c(m) = \#$ slov v C^k s délkou m

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

$$\Rightarrow \left(\sum_x 2^{-\ell(x)}\right)^k \leq \sum_{n=1}^{k \cdot \ell_{max}} 2^m \cdot 2^{-m} = k \cdot \ell_{max}$$

$$\sum_x 2^{-\ell(x)} \leq \left(k \cdot \ell_{max}\right)^{1/k}$$

$k \to \infty$ $\to 1$ ∎

- generování pořadí rozdělení pomocí nezávislých bitů $\frac{1}{2}, \frac{1}{2}$.
- C libovolný kód $C : X \to \Sigma^*$

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

  - $H(X) - 2 \lg H(X) - 2 \leq L(C)$

Dk: z C lze udělat bezprefixový kód

$C(x) \longrightarrow e(|C(x)|)\,0\,|\,C(x)$

zakódování délky $C(x)$ v binárním vyjádření a zdvojení každého bitu

$0 \to 00$
$1 \to 11$

$\longrightarrow$ bezprefixový

$l(x) \longrightarrow l'(x) = l(x) + 2\lg l(x) + 2$

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

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

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

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

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

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

Michal Koucky at 28. 3. 2016 21:07

Kolmogorovská složitost

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

3333333 ... 3 $\longrightarrow 3^{10}$
3141522653 5 $\longrightarrow \pi$ ... prvních deset číslic
8435427952 1 $\longrightarrow$ náhodný

Odlišují se délkou zápisu. Záleží však na jazyku.

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

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

Def: Kolmogorovská složitost x vzhledem k f:

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

Věta: Existuje univerzální částečně 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 částečně rekurzivních fcí

ϕ_i ... i "kód"

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

U definujeme programem:
na vstupu $\langle i, z\rangle$
dekóduj i a z.

simuluj ϕ_i na x
pokud se zastaví, vyhlásím, co vyhlásila ϕ_i na x.

konec.

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 $\mathcal{U}$ ☐

$\rightarrow$ $\mathcal{U}$ dává nejmenší složitost ze všech čir. fcí
(což na konstantu)

$\rightarrow$ zafixujme nějaké $\mathcal{U}$ $\quad C \equiv C_{\mathcal{U}}$.

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

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

Def: x je Kolmogorovsky náhodné, pokud $C(x) \ge |x|$.

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

Věta: $\exists \mathcal{U}$; $\forall$ čir. f. g $\exists c$ $\forall x, y$
$C_{\mathcal{U}}(x|y) \le C_g(x|y) + c_g$
$\rightarrow$ zafixujeme $\mathcal{U}$ ... univerzální
$C(x) \overset{\text{def}}{=} C(x|\varepsilon)$
↳ prázdný řetězec.

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

Věta: $C(x,y) \le C(x) + C(y|x) + O(\lg C(x,y))$
Dk: pgm p $\mapsto C(x)$ pro x
pgm q pro y, když znám x
↳ $C(y|x)$
$O(\lg C(x,y))$ pro separaci p a q ☐

Věta: $C(x,y) \ge C(x) + C(y|x) - O(\lg C(x,y))$

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

$$I_C(x:n) = C(x) - C(x|n) \leq n - n = 0$$

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

→ logaritmická ztráta nutná

Věta: $X_1, X_2, \ldots$ je rekurzivní posloupnost prvků distribucí, X_n na $\{0,1\}^n$. Pak (↳ ∃ pgm pro aproximaci pdf)

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

Dk: • $\mathbb{E}[C(X_n)] \leq H(X_n) + O(\lg n)$

vezmeme Huffmanův kód pro X_n, $x \in X_n$ má délku $\ell(x)$

$C(x) \leq \ell(x) + O(\lg n)$ ← pgm pro Huff. kód X_n

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

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

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

$d_x = 0^{\lg n + O(1)} 1 |p_x| p_x$

↳ $\lg n + O(1)$ bitový binární zápis $|p_x|$

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

$$\rightarrow H(X_n) \leq \mathop{\mathbb{E}}_{x \in X_n}[|d_x|] = \mathop{\mathbb{E}}_{x \in X_n}[C(x)] + O(\lg n)$$

Samoopravné kódy

• úvodní přednáška P. Gregor

- podobný
- definice
- Shannonova věta
  + inverzní Shannonova věta

Michal Koucky at 12. 4. 2016 10:43

- Hammingův kód

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

	x_1	x_n
1	1 1	(
:	0 0	1
ℓ	0 0 . .	1

- očíslujme 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.

$\rightarrow$ ℓ parit $a_1 \dots a_\ell$

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

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

- rozšířený Hammingův kód přidá navíc paritu

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

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

Odhadování 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 $\{0,1\}^\ell$ udávající, kdeod parity nesedí
- pokud nastala chyba v části $a_1 \dots a_\ell$ $\Rightarrow$ nesedí právě jedna parita a to ta správná.

- kód lze specifikovat maticí

Př.:

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

$\ell = 3$

$[7, 4, 3]_2$

Př.:

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

$\ell = 3$

$[7, 4, 3]_2$

[Hamming 1950]

$x \longrightarrow xG$

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

→ lineární kód $[n, k, d]_2$

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

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

$\forall y \quad yH = 0$ iff $\exists x \; xG = y$

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

↳ báze ortogonálního vektorového prostoru (ortogonální doplněk)

$\dim H = n - \dim G$

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: $xG = y$ $\quad y' = y + e_i$ ↳ $(0,0,\ldots,1,\ldots,0,0)$ (i-tá pozice)

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

e_iH: index bitu s chybou

↳ syndrom

- obecně pro lineární kód

$xG = y$ $\quad y' = y + e$ ↳ 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. váhou $\leq d$, eH jsou různé

tj. součet $\leq 2d$ řádků z H musí být nenulový.

→ tabulka $eH \to e$ pro snadné dekódování

pro lineární kód platí: $[n, k, d]_q$

$\forall x, y \in C \Rightarrow x+y \in C$
$x-y \in C$

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

$\Rightarrow d_H(x,y) = |\{i;\ (x-y)_i \neq 0\}|$
$= wt_H(x-y)$ ... Hammingova váha $x-y$

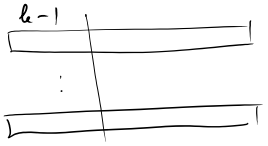
$d = \min_{x \neq y \in C} d_H(x,y) = \min_{\substack{x \in C \\ x \neq 0}} wt_H(x)$

→ u lineárních kódů se lze zaměřit na minimální sám neuložený sbor pro zjištění minimální vzdálenosti.

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

$n \geq k + d - 1$

Dk:



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

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

$d \leq n - k + 1$ ∎

Reed-Solomonovy kódy

$GF[q]$, n, k $\quad n \leq q$

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

$m \in GF[q]^k \longrightarrow p_m(x) = m_0 + m_1 x_1 + \dots m_{k-1} x^{k-1}$

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

- RS kód je lineární:

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

$\rightarrow [n, k, n-k+1]_q$ kód

↑ počet stupňů $k-1$ má nejvýše $k-1$ kořenů, t.j. každé nenulové slovo má Ham. váhu $\geq n-(k-1)$

Př: $k = n-4$ $[n, k, 5]_q$ kód

každý symbol z $\mathbb{F}_q$ reprezentujeme binárně

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

... binární kód opravující 2 chyby.

$N = n \cdot \log n$.

Michal Koucky at 18. 4. 2016 14:53 [1]: Konečná tělesa [1]

Dekódování Reed-Solomonových kódů

$m = (m_0, m_1, \ldots m_{k-1})$ $\alpha_1, \ldots \alpha_n \in \mathbb{F}_q$

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

$m \longrightarrow (p_m(\alpha_1), p_m(\alpha_2), \ldots, p_m(\alpha_n))$

... kód zprávy

$\downarrow$

$(r_1 \ldots r_n) \in \mathbb{F}_q^n$

... přijaté slovo

chceme nalézt polynom stupně $\leq k-1$, který se nejvíce shoduje s $(r_1 \ldots r_n)$.

Algoritmus Berlekamp–Welch (1986)

error lokační polynom

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

chceme nalézt polynom $E(x)$ stupně d_e t.ž. $\neq 0$

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

- pokud máme tělesa E, dekódování snadné:

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

$d_e \leq n-k$.

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

... Lagrangeův polynom

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

Hledání $E(x)$:

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

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

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

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

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

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

→ soustava n lineárních rovnic s n neznámými

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

→ Gaussova eliminace $O(n^3)$

Rychlá Fourierova transformace $O(n \cdot \log n)$ (FFT)

- vezmeme libovolné řešení soustavy $Q(x)$ a $E(x) \neq 0$
- pokud $E(x)$ nedělí $Q(x)$ → 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: $P(x)$.

Tvrzení: 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 stupeň 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 (*) $Q_1(\alpha_j) = r_j E_1(\alpha_j)$ $\quad Q_2(\alpha_j) = r_j E_2(\alpha_j)$

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

$\Rightarrow R(x) \equiv 0$

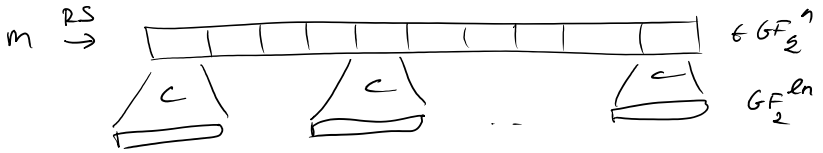
$\Rightarrow Q_1(x) E_2(x) = Q_2(x) E_1(x)$ v n bodech

jelikož jsou to polynomy st. $\le n-2$, mají dvojí rozklad na ireducibilní polynomy

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

- Reed-Solomonovy kódy vyžadují velikost tělesa $\ge n$.
  Zobecnění: Reed-Mullerovy kódy – polynomy ve více proměnných

- Jak z R-S kódů udělat binární kódy?
  → kódovat prvky z GF_2 binárně pomocí samoopravného kódu.



$\ell \approx O(\lg Q)$ $\quad C \dots$ kód $[\ell, \lg Q + 1, d]_2$

pokud RS je $[n, k, D]_Q$ kód, pak výsledný kód je $[n\ell, k \cdot \lg Q, dD]_2$ kód.

Dekódování:
1) dekóduj každý vnitřní symbol
2) dekóduj RS-kód

→ opraví $\frac{dD}{4}$ chyb

- lze opravit až $\frac{dD}{2}$ – Forneyho alg.
- RS kód s min. vzdáleností D umí opravit E chyb a S výmazů, pokud $2E + S < D$.

Forneyho alg. (opraví)
1) dekóduj každý vnitřní symbol $r_i \to r_i'$
   vymaž každou pozici i s pravd. $\frac{2e_i}{d}$, (resp. $\min(\frac{2e_i}{d}, 1)$)
   kde $e_i = d_H(r_i, r_i')$.
2) dekóduj zbylé $r_1', \dots r_n'$ pomocí BW alg.

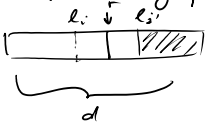
Tvrzení: Pokud E je počet nesprávných symbolů r_i' a S je počet vymazaných symbolů, pak
$\mathbb{E}[2E + S] < D$,

pokud $\sum \ell_i < \frac{d D}{2}$.

Dk: cvičení — symbol i přispěje do součtu hodnotou $\leq \frac{2\ell_i}{d}$

2 případy: a) $\ell_i \leq d/2$ ... b) $\ell_i > d/2$ ...

[Derandomizace — lze vybrat společný threshold r ⌜ všechny prvky:



]

Náhodný lineární kód: (Varshamov)

- výběr náhodné 0-1 matice G velikosti $k \times n$

kódový $x \in \{0,1\}^k \to xG$

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

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

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

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

$$Pr_G[\exists \underset{\neq 0}{x \in \{0,1\}^k};\ xG \in Vol_2(n, d-1)] \leq 2^k \cdot 2^{(H(p)-1)n} < 1$$ ∎

→ 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 ze slov v C, přidej w do C.

Pozorování: algoritmus se zastaví po nejvýše

$$\frac{2^n}{Vol(n,d)}$$ krocích.

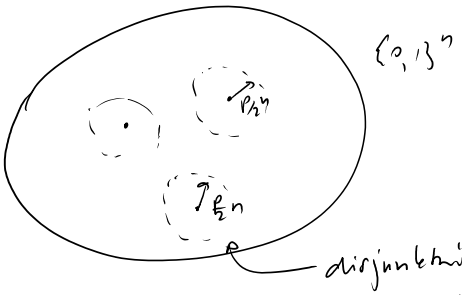
$$\Rightarrow |C| \geq 2^{n - H(\frac{d}{n})n}$$

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

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

Hammingova mez: $[n, k, pn]_2$ kód
má $k \leq \left(1 - H\left(\frac{p}{2}\right)\right) n$.

Dk:



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

- slabý Gilbert-Varshamov kód
$[n, (1-H(p))n, pn]_2$ kód
s Shannonovým kódem opraví $\frac{p}{2}n$ chyb
$[n, (1-H(p))n, ?]_2$ opravující
pn chyb s velkou pstí

- Dekódování se seznamem (list decoding)
  - pro $[n, k, d]_2$ kód a přijaté slovo $y \in \{0,1\}^n$ hledáme všechna slova do vzdálenosti d' od y.
  $\frac{d-1}{2} \leq d' \leq d$

→ seznam $L \subseteq C$ $\forall y' \in L$, $d_H(y, y') \leq d'$.

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-tý x_i, aniž bych musel dekódovat celé x.
Zároveň ani nechci číst celé y.

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

zároveň ale neobsahuje žádné y.

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

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

$x \in \{0,1\}^n \qquad C(x) = y_{000\ldots} \cdots y_{111\ldots} \qquad y_a \quad a \in \{0,1\}^k$

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

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

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

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

$e_i = (0,0,\ldots,\underset{i\text{-tá pozice}}{1},0\ldots0)$

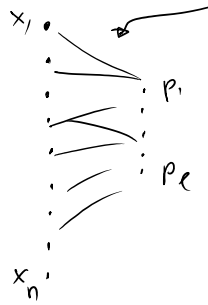
výstup: $y_a \oplus y_{a \oplus e_i}$ $\quad \leq \frac{1}{6}n$ $\quad \neq C(x)$

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

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

- Opakováním a vzhledem většinového výsledku lze pravděpodobnost chyby snížit

- kombinatorické konstrukce kódů



vhodný malý graf (expandér)

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

p_i ... paritní sousedi – pro kódové slovo = 0.

- vlastnosti zdědí na vlastnosti grafu mezi $\bar{x}$ a $\bar{p}$.
- chyby v kódovém slově přepnou většinu paritu
- iterativně lze chyby odstranit.

Michal Koucký 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 f.

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

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

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

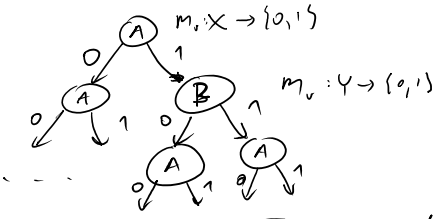
$GT(x,y) = [x \leq y]$ $\quad x \stackrel{?}{\leq} y$

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

A 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,
  ∀ A-ový uzel u je přirozená nějaká fce
  $m_u : X \to \{0,1\}$, která říká Alence, co má poslat v závislosti na jejím vstupu
  ∀ B-čkový 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)

- $D(f) \leq n+1$. Dk.: Alenka pošle svůj vstup x, Bob pošle zpět $f(x,y)$

Př.: 1) $x, y \subseteq \{1, \dots, n\}$ min

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

$D(f) \leq 2 \log n$

2) medián

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

$f(x,y) = \text{medián } x \cup y$

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

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

$D(f) \leq O(\lg n)$ těžší

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

Kombinatorický obdélník

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

$A \subseteq X \quad B \subseteq Y \quad A \times B$ .. kombinatorický obdélník

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

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

Dk: "$\Leftarrow$" $A = \{x : \exists y (x,y) \in R\}$ "$\Rightarrow$" triv.

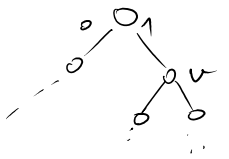
$B = \{y : \exists x (x,y) \in R\}$

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

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

$\overset{\text{předpoklad}}{\Rightarrow} (x,y) \in R$ ▨

Protokol P:



$R_v = \{(x,y) : \text{na vstupu } x, y$ Alenka & Bob dojdou do $v\}$

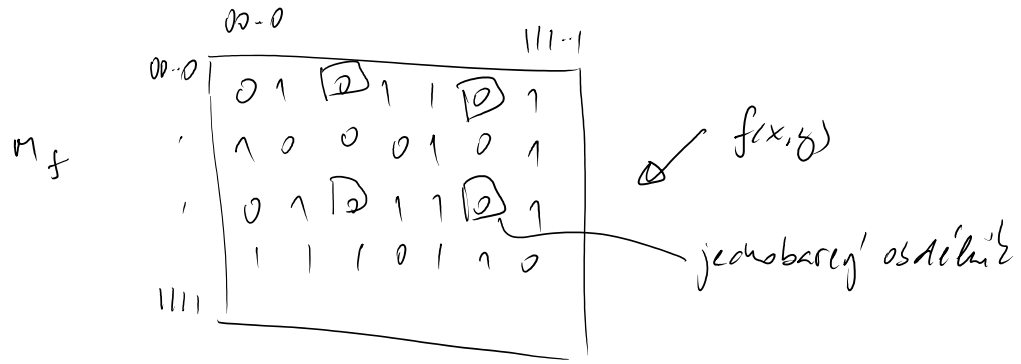
- $\forall v, R_v$ je kombinatorický obdélník.

Dk: 1) indukcí podle hloubky v.

nebo 2) "cut-and-paste" argument ▨

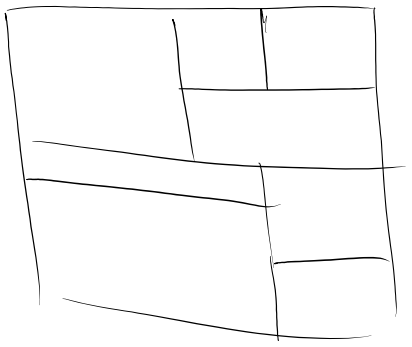
- Pokud P počítá f, pak pro každý list ℓ v P, vstupy v R_ℓ mají stejnou hodnotu $f(x,y)$.

→ jednobarevný obdélník

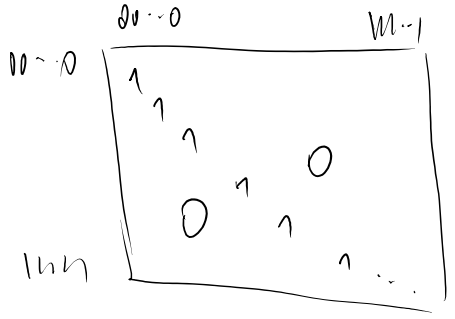


→ každý protokol dává pokrytí matice f jednobarevnými obdélníky

- 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

→ $D(EQ) \geq \log 2^n + 1 > n$

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

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

ustupy $(x, \bar{x})$ musí být v různých jednobarevných obdélníčcích

$\Rightarrow \geq 2^{n+1}$ obdélníčků

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

- 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 $(M_\ell) \leq 1$

$\Rightarrow$ hodnost $M_f \leq$ počet listů P

$=$

r

$\log r \leq D(f)$ ✓

Př: 1) hodnost $(M_{EQ}) = 2^n$

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

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

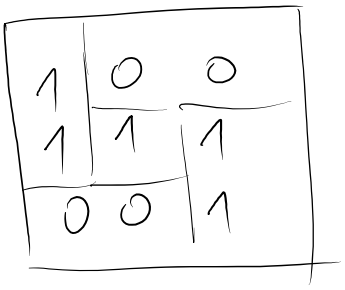
$$(M_{IP})^2 = \begin{array}{|cc|} \hline 0 & \\ 2^{n-1} & 2^{n-2} \\ 2^{n-2} & 2^{n-1} \\ \hline \end{array}$$

Michal Koucký at 10. 5. 2016 11:37

Příklady pokrytí matice M_f obdélníky

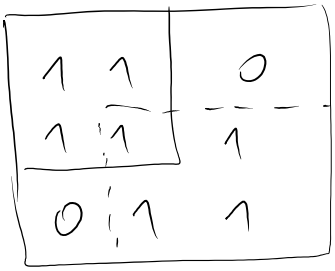
Př:

Př: 1)



∃ pokrytí 5 obdélníky, ale neexistuje protokol pokrývající právě 5 obdélníky.

2)



∃ překrývající se pokrytí 4 obdélníky

→ různé míry pokrytí

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

$C^D(f)$ ... nejmenší možný počet obdélníků pokrývajících nepřekrývajících se matic M_f

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

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

→ nedeterministický komunikační protokol:
P ... všemocný dokazovatel, který zná x a y
a chce ze

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

m ... důkaz, že $f(x,y)=z$

A — $x \in \{0,1\}^n$

B — $y \in \{0,1\}^n$

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

Př: $EQ(x,y)=0$ ... m je jakýkoliv bit, kde $x \neq y$

$|m| = \log_2 n$

$EQ(x,y)=1$ ... m je celý řetězec x.

$|m| = n$ bitů

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

prvků možných správ = # 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: $N^0(EQ) = \log n$ $\quad N^1(EQ) = n$

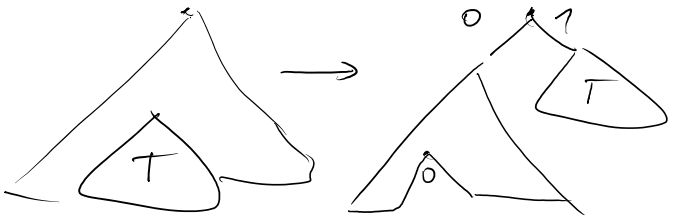
$N^0(DISJ) = n$ $\quad N^1(DISJ) = n$

C vs D

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

Dk: 1. "$\leq$" triv

2. "$\leq$" ... v protokolu nalezni podstrom
$\leq 1/3 \leq \leq 2/3$ listů a přesuň na vrch



→ vyvážený protokol ∎

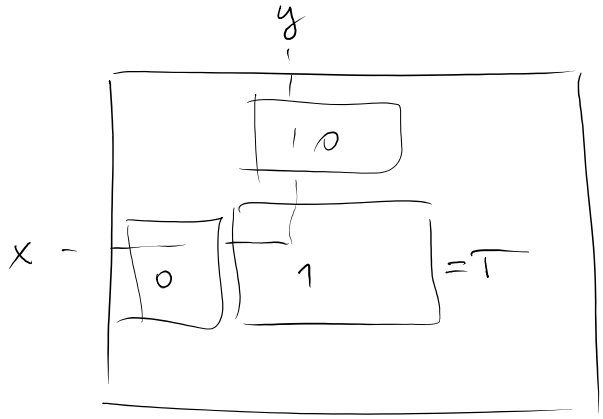
Otázka: $D(f) = O(\lg C^D(f))$?

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

Dk:

idea:

Řekněme, že $f(x,y) = 1$



- každý 0-obdélník může protínat pouze hrany v řádcích, nebo sloupcích

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

1) Pokud všech 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 nejvýše polovinu živých 0-obd. Pokud ano, pošle číslo 1-obd. Bobovi. Jinak nechá Boba.

3) B se podívá, zda $\exists$ 1-obd. obsahující řádek x, který v řádcích protíná $\leq \frac{1}{2}$ 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 ∎

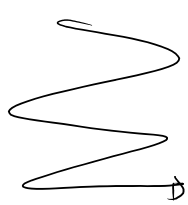
Otázka: $\forall f: D(f) = O(\log \text{hodnost}(M_f))^{O(1)}$?

... "log-rank conjecture"

Pravděpodobnostní protokoly

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

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



· zprávy Alenky můžou záviset též na r_A

· -"- Boba -"- r_B

→ ve stromu protokolu v uzlu v patřícím

Alici máme fci $m_v : \{0,1\}^n \times \{0,1\}^* \to \{0,1\}$
$\quad\quad x \quad\quad r_A \quad\quad m_v$

Bobu —"— $m_v : \{0,1\}^n \times \{0,1\}^* \to \{0,1\}$
$\quad\quad y \quad\quad r_B$

– 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 nezměnilo, neboť každý protokol lze zastavit po $1/\varepsilon$-násobku průměrné délky a to zhorší chybu nejvýše o ε.

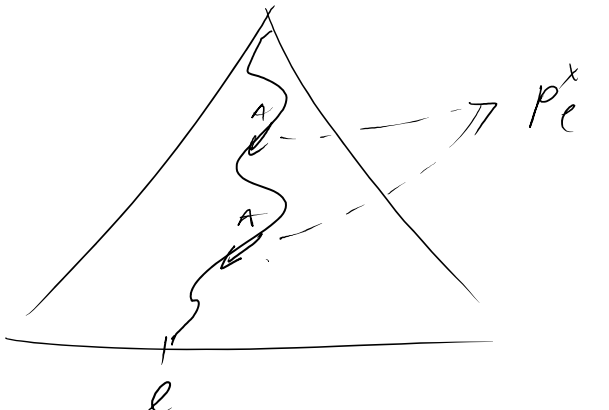
Př: 1) $R(EQ) = O(\lg n)$

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

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



p_e^x je samo součinem pstí v jednotlivých Alenčiných uzlech a obdobně p_e^y.

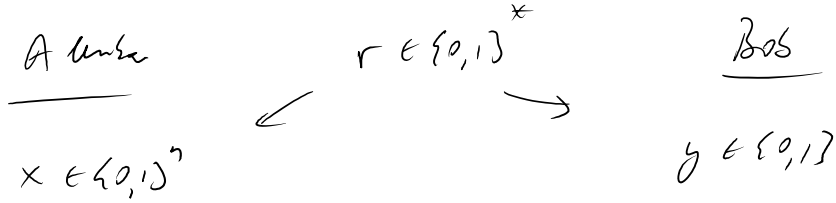
Def. protokol pro f: Alenka spočítá pstí p_e^x pro

vždy listy l a použije tuto informaci Bobovi. Ten spočítá svoje p_l^y a zjistí výsledek pst. jednotlivých výstupů. Vydá odpověď s nejvyšší pstí.

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

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

- protokoly s veřejnými náhodnými bity:



- Alenka i Bob dostanou zadarmo společný náhodný řetězec r.

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

Dk: idea: lze zvolit množinu $R \subseteq \{0,1\}^{O(n)}$ řetězců r t.ž. chyba protokolu na každém vstupu (x,y) se

při použití zadaného náhodného r a náhodného r vygenerovaného z naší množiny R liší právě $0 \leq \frac{1}{n^{O(1)}}$. [Toto plyne z časového už při zvolení si množiny R zadá náhodně.]

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

Př: EQ s veřejnými bity vs soukromými.

$O(1)$ vs. $\Theta(\log n)$

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

$\Omega(\log n)$

Věta: $R(DISJ) \geq \Omega(n)$.

Použití komunikačních složitostí

Data streams

Výpočetní model

Alg.

• Algoritmus má přístup omezenou paměť nedokáže si zapamatovat celý vstup.

• Př: • dána jsou celá čísla z intervalu $[0, n]$

• chci znát:

1) celkový součet — snadno
2) průměr — snadno
3) počet různých prvků — lze s malou pamětí pravděpodobnostní algoritmus (aprox.)
4) kolikrát se vyskytl nejčastější prvek — vyžaduje paměť $\Theta(n)$.

Dk: víme $D(DISJ) \geq n$.

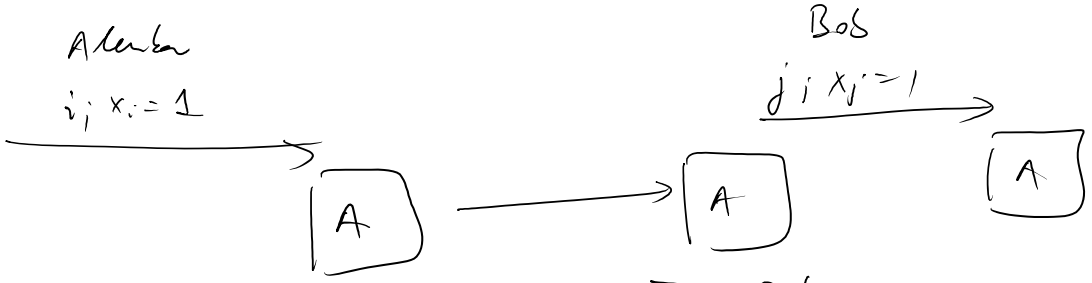
ukážeme, že dobrý algoritmus pro 4) s malou pamětí implikuje příliš efektivní protokol pro DISJ.

• mějme alg. A pro 4). Protokol pro $A \& B$ funguje následujícím způsobem:

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

• Alenka vytvoří posloupnost prvků i t.ž. $x_i = 1$

tuto posloupnost zpracuje alg. A.



poté stav paměti algoritmu A počítá Bobem.
Ten vytvoří posloupnost j; $y_j = 1$ a
zpracuje ji algoritmem.

Pozorování! $DISJ(x,y) = 1 \iff$ nejčastější
pouze se v dotaz vybírají
Alenka a Bobem vyskytuje
právě dvakrát.

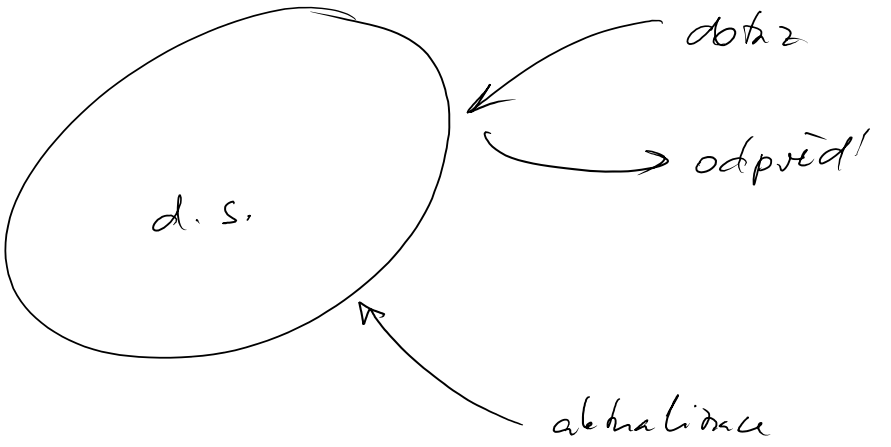
→ objem komunikace mezi Alenkou a Bobem
je velikost paměti použité algoritmu.
(+1 bit na sdělení výsledku Alence)

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

Stejný trik s identickým důkazem platí pro
jiné algoritmy

Michal Koucký at 26. 5. 2016 9:39

Datové struktury

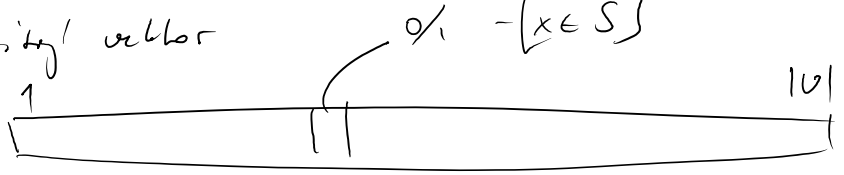


datová struktura uchovává nějaká data a zodpídá
on ních dotazy

Př: d.s. pro množinu $S \subseteq U$ $\quad U = \{1, \dots, |U|\}$.
dotazy typu $[x \overset{?}{\in} S]$ pro každé $x \in U$.

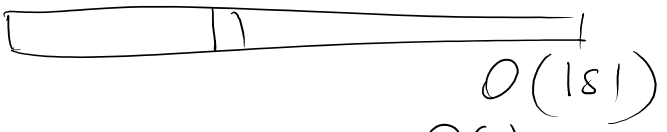
možné implementace

1) bitový vektor



dotaz — čas $O(1)$
paměť $|U|$ špatné, pokud $|S| \ll |U|$

2) hašovací tabulka



čas na dotaz ... $O(1)$
paměť $O(|S|)$

Pozn: není $O(1)$ jako $O(1)$
↑ ↑
1-bit $O(\lg n)$-bitů
1) 2)

- Chtěli datovou strukturu pro lineární prostory

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

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

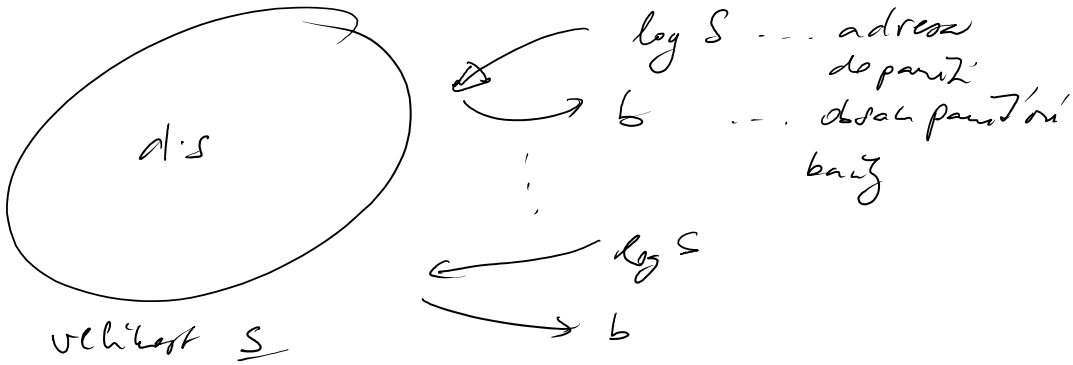
řešení: triviální – pak odpovídá na všechny možné dotazy → prostor 2^n bitů
dotaz na dotaz $O(1)$

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

báze V

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

měřítko: # 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:

Alice $y \in GF_2^n$ → Bob $V \subseteq GF_2^n$

$y \in GF_2^n$ $\quad\rightsquigarrow\quad$ $V \subseteq GF_2^n$ V podprostor

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

Alenka komunikuje a bitů, Bob komunikuje b bitů. Kolik musí být a & b?

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

důkaz níže.

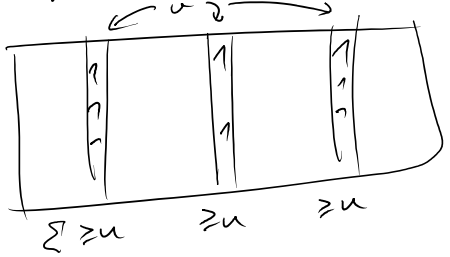
Zpět k datové struktuře:

datová struktura pro $[y \stackrel{?}{\in} V]$, kde na každý dotaz potřebujeme nejvýše t přístupů do paměti dává komunikační protokol pro $[y \stackrel{?}{\in} V]$ 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.

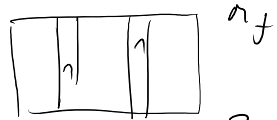


Tvrzení: Pokud f je fce s (u,v)-těžkou maticí M_f a protokolem pro f, kde Alenka posílá a bitů a Bob b, pak M_f obsahuje jednobarevný $u \times v$

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) $a+b=0$



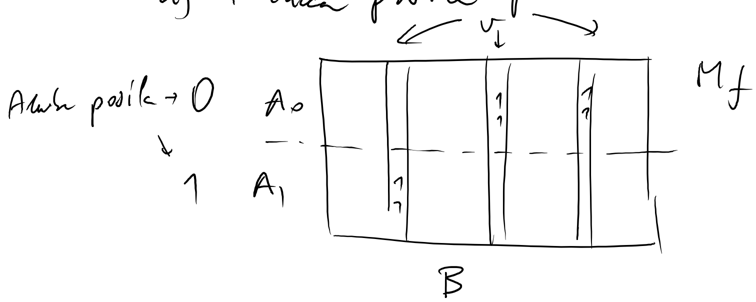
zjevně $M_f \equiv 1$ protože A a B nepotřebují komunikovat

$\Rightarrow M_f \geq u \times v$ ✓

2) $a+b-1$ ✓ $\Rightarrow a+b$

dva případy

a) Alenka posílá první bit

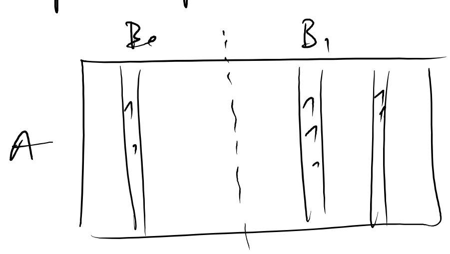


buď podmatice $A_0 \times B$ nebo $A_1 \times B$ je $\left(\frac{u}{2}, \frac{v}{2}\right)$-těžká

$\underset{\text{indukce}}{\Rightarrow}$ 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

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

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

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

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

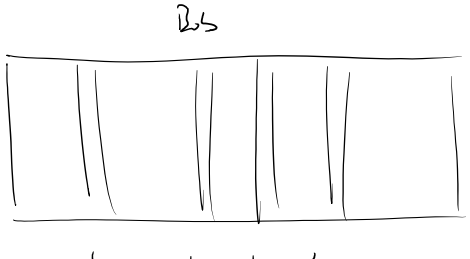
1) & Tom! $\Longrightarrow$ pokud by existoval protokol, kde Alenka posílá $n/6$ bitů a Bob $\frac{n^2}{12} - n/6$, pak by Malin $[y \overset{?}{\in} V]$ obsahovala 1-obd. velikosti $\frac{2^{n/2}}{2^{n/6}} \times \frac{2^{n^2/4}}{2^{n/6 + n^2/12 - n/6}} =$

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

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

Zbývá dokázat 1) a 2)

1)



$V + \bar{z}$. $\dim V = \frac{n}{2}$ obsahují $2^{n/2}$ různých vektorů z GF_2^n, tedy jsou $2^{n/2}$-tého!

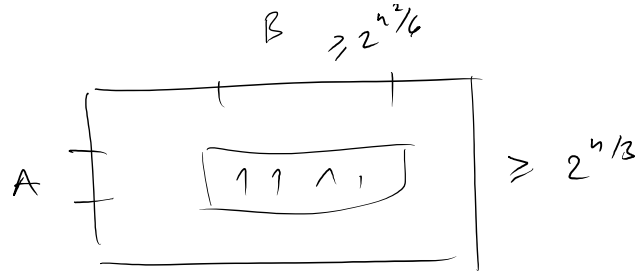
$\#V ; \dim V = \frac{n}{2} \geq 2^{n^2/4}$:

$\#$ bází vel. $\frac{n}{2}$ (lin. nezávislých $\frac{n}{2}$-tic vektorů) $\longrightarrow$ $\frac{\prod_{i=0}^{n/2-1}(2^n - 2^i)}{\prod_{i=0}^{n/2-1}(2^{n/2} - 2^i)} = \prod_{i=0}^{n/2-1} \frac{2^n - 2^i}{2^{n/2} - 2^i} \geq \prod_{i=0}^{n/2-1} 2^{n/2} \geq 2^{n^2/4}$

$\hookrightarrow$ $\#$ 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.

$\forall V \in B: \quad \langle A \rangle \leq V$

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

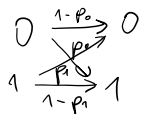
Koucky v 4/12/2017 8:47 PM

Přenos po nespolehlivém kanálu

$X \rightsquigarrow Y$ (šum)

spolehlivý kanál: $0 \to 0$, $1 \to 1$

nespolehlivý kanál:



p_0 ... pst. chyby při vyslání 0
p_1 ... pst. chyby při vyslání 1

Často pro jednoduchost $p_0 = p_1 = p$

bez újmy na obecnosti: $0 \leq p < \frac{1}{2}$

(pro $p = \frac{1}{2}$ nelze přenášet nic, pro $p > \frac{1}{2}$ je potřeba obrátit interpretaci výstupu.)

"binární symetrický kanál"

- Pokud X je n.p. $\in \{0,1\}$, Y je také n.p. $\in \{0,1\}$, zajímá nás

$$I(X:Y) = H(Y) - H(Y|X) = H(Y) - \sum p_x \underbrace{H(Y|p_x)}_{=H(p)}$$
$$= H(Y) - H(p) \leq 1 - H(p).$$

- kapacita kanálu: $C = \max_{\text{n.p. } X} I(X:Y)$

- kódování k bitů $\to n$ bitů

předpoklad chyby nezávislé!

Př.: $k = n = 100$ $\quad p = 0.01$

- kódování k bitů $\to$ n bitů chyby nezávisle

Př: • $k = n = 100$ $p = 0.01$

$$Pr[\text{přenos bez chyby}] = (1-0.01)^{100} \approx 0.37$$

- $k = 100$ $n = 300$

každý bit zopakuji: 3-krát

$0 \to 000$

$1 \to 111$

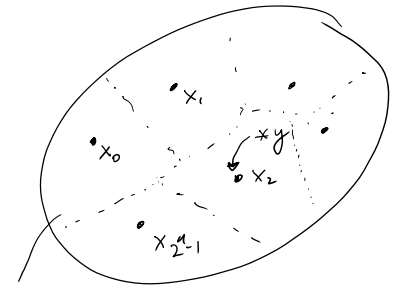
dekóduji: každou trojici zvlášť:

000	111
001	110
010	101
100	011
$\to 0$	$\to 1$

$$Pr[\text{chyba v dané trojici}] = \binom{3}{2} \cdot p^2(1-p) + p^3 = 3p^2 - 2p^3 = 0.000298 = p'$$

$$Pr[\text{zpráva dekódována správně}] = (1-p')^{100} \geq 0.97$$

kód: $C \subseteq \{0,1\}^n$ $|C| = 2^k$



Voronoi diagram

y - dekódujeme na nejbližší kódové slovo b v Hammingově vzdálenosti.

$$\Delta_{Ham}(x,y) = |\{i;\ x_i \neq y_i\}|.$$

rozumné dekódovací pravidlo - mělo by dekódovat na nejpravděpodobnější vysílané zprávu vzhledem k přijatému y. To je nejbližší slovo (Bayesovo pravidlo)

rate $R = \frac{k}{n}$ ... efektivita využití kanálu

Věta: (Shannonova) Uvažme binární symetrický kanál s pravděpodobností chyby $p < \frac{1}{2}$. Nechť $0 < R < 1 - H(p)$. $\forall \varepsilon > 0$ $\forall n$ dostatečně velké $\exists C \subseteq \{0,1\}^n$, $|C| = 2^{Rn}$ a chyba špatného dekódování při přenosu po kanálu s chybou p je $\leq \varepsilon$.

- ukázáno posloupnost kódů $C_1, C_2, \ldots$ $C_n \subseteq \{0,1\}^n$

t.ž. chyba při přenosu $\varepsilon_n \to 0$ pro $n \to \infty$.

Dokonce $\varepsilon_n \approx 2^{-\Theta(n)}$.

Potřebujeme dvě pomocná tvrzení

Věta (Čebyševova nerovnost): Nechť $0 < p < 1$ a $0 < \alpha < 1$. Pak

existuje konstanta $c_{p,\alpha} > 0$ t.ž. pro $\forall n \geq 1$ a posloupnost $X_1, X_2, \dots X_n$ nezávislých n.p. t.ž. $X = \begin{cases} 1 & \text{s pstí } p \\ 0 & \text{s pstí } 1-p \end{cases}$

$$Pr\left[\left| \sum X_i - np \right| \geq \alpha n \right] \leq 2 e^{-c_{p,\alpha} n}.$$

Dk viz. předmětka "Pravděpodobnostní techniky"

Lemma: $\forall n, r \geq 1$, def. $Vol(n,r) = |Ball_{n,r}(0^n)| = \sum_{i=0}^{r} \binom{n}{i}$.

Pak $Vol(n,r) < 2^{n H(\frac{r}{n})}$ pro $r \leq \frac{n}{2}$.

Dk:
$$n^n = (r + (n-r))^n = \sum_{i=0}^{n} \binom{n}{i} r^i (n-r)^{n-i} > \sum_{i=0}^{r} \binom{n}{i} r^r (n-r)^{n-r}$$
pro $r \leq \frac{n}{2}$

$$\Rightarrow \frac{n^n}{r^r (n-r)^{n-r}} > \sum_{i=0}^{r} \binom{n}{i}$$

$$\left(\frac{n}{r}\right)^r \left(\frac{n}{n-r}\right)^{n-r} = 2^{r \lg \left(\frac{n}{r}\right) + (n-r) \lg \left(\frac{n}{n-r}\right)}$$
$$= 2^{n \left[\frac{r}{n} \lg \frac{n}{r} + \frac{n-r}{n} \lg \frac{n}{n-r} \right]}$$
$$= 2^{n H(\frac{r}{n})}$$ ☐

$Ball_{n,r}(x) = \{ x' \in \{0,1\}^n ; o_{Ham}(x,x') \leq r \}$

Dk Shannonovy věty:

1) kód C_n zvolíme rovnoměrně náhodně
$C_n : \{0,1\}^k \to \{0,1\}^n$
t.j. pro každé $m \in \{0,1\}^k$ vybereme nezávisle n-bitové kódové slovo.

2) budeme počítat pst. chybného dekódování při přenosu náhodně zvoleného slova $X \in C_n$.

- ukážeme, že pst. chybného dekódování $\leq 2^{-cn}$ pro nějakou konstantu c závislející pouze na p a $1 - H(p) - R$.
- pokud průměrná chyba přes volbu C_n je $\leq 2^{-cn}$, pak existuje kód C_n, který dosahuje chyby $\leq 2^{-cn}$ pro dekódování náhodného zvoleného $X \in C_n$.
- zvolíme tento kód. Pro nejvýše 1/2 slov v C_n je pst. špatného dekódování $\geq 2 \cdot 2^{-cn}$, jinak průměr přes náhodné $X \in C_n$ by byl větší než 2^{-cn} $\Rightarrow \exists C_n' \subseteq C_n$ (Markov) $\Rightarrow$
$|C_n'| \geq |C_n|/2$ t.ž. pst. špatného dekódování pro každé $X \in C$ je $\leq 2^{-cn}$.
To je náš kód. (jeho rate $= R - \frac{1}{n} \approx R$)

→ Stačí tedy ukázat, že pro náhodný kód C_n a náhodně vybraném kódové slovo $X \in C_n$, pst. špatného dekódování $\leq 2^{-cn}$.

Zvol $\alpha \in (0,1)$ tž. $H(p+\alpha) + R < 1$

(lze ze spojitosti $H(p)$)

Polož $r = n(p+\alpha)$.

Pro $i = 1, \dots, n$ nechť $E_i = \begin{cases} 1 & \text{nastala chyba} \\ & \text{v } i\text{tém bitu} \\ 0 & \text{jinak} \end{cases}$

$E_1 \dots E_n$ jsou nezávislé $\Pr[E_i = 1] = p$

$\sum E_i$ ... počet chyb.

$\mathbb{E}[\sum E_i] = pn$ ... očekávaný počet chyb.

$\Pr[|\sum E_i - pn| \geq \alpha n] \leq 2 \cdot e^{-c p,\alpha n}$

(Černovova věta)

tedy pst. že nastane víc než $(p+\alpha)n$ chyb je $\leq 2 \cdot e^{-c p,\alpha n}$.

Pokud nastane méně než $(p+\alpha)n$ chyb, pak tvrdíme, že pst. špatného dekódování

$\leq 2^{Rn} \cdot \frac{\mathrm{Vol}(n, (p+\alpha)n)}{2^n} \leq 2^{-\Theta(n)}$.

$\Rightarrow$ v obou případech chyba dekódování $\leq 2^{-\Theta(n)}$

tedy celková pst. chybného dekódování $\leq 2^{-\Theta(n)}$.

pstní protože můžeme postupovat v následujících pořadí:

1) zvol náhodné $m \in \{0,1\}^{Rn}$
2) zvol náhodné $C(m) \in \{0,1\}^n$ (přeměříme kódovací slovo)
3) zvol náhodnou chybu $E_1, \dots E_n$
4) zvol náhodné $C(m') \in \{0,1\}^n$ pro všechna $m' \neq m \in \{0,1\}^{Rn}$.

pokud $\sum E_i \leq (p+\alpha)n$, což je případ který nás zajímá, pak přijatý y a odeslané $x = C(m)$ splňují $\delta_{Ham}(x,y) \leq (p+\alpha)n$.

V takovém případě při dekódování nastane chyba pouze tehdy, pokud v C existuje další slovo ve vzdálenosti $\leq (p+\alpha)n$ od y.

$\Pr[\mathrm{Ball}_{n,r}(y) \cap (C \setminus \{x\}) \neq \emptyset] \leq$

$\leq (2^{Rn} - 1) \cdot \frac{|\mathrm{Ball}_{n,r}(y)|}{2^n} \leq 2^{nH(\frac{r}{n})} \cdot 2^{n(R + H(p+\alpha) - 1)}$

$$\leq \left(2^{Rn}-1\right) \cdot \frac{|\mathrm{Ball}_{n,r}(y)|}{2^n}$$

$$\leq 2^{Rn} \cdot \frac{2^{nH(\frac{r}{n})}}{2^n} = 2^{n(\underbrace{R+H(p+\alpha)-1}_{<0})}$$

$$\leq 2^{-cn}$$ pro nějaké $c>0$ závisející pouze na p, R.

Opačná Shannonova věta říká, že pokud $R > 1-H(p)$, pak každý kód $C_n \subseteq \{0,1\}^n$, $|C_n| \geq 2^{Rn}$, bude dosahovat průměrné chyby dekódování blížící se 1.

Věta: (Shannon) Nechť $0 < p < \frac{1}{2}$, $R > 1-H(p)$ a $\delta \in (0,1)$. Pak pro dostatečně velké n, $\forall C_n \subseteq \{0,1\}^n$, průměrná chyba dekódování $\geq 1-\delta$.

Dk: X ... náhodně zvolené kódové slovo z C_n

E ... chybový vektor $\quad |E|$ .. počet chyb

$D_x = \{ y \in \{0,1\}^n ;\ y \text{ decoduje to } x \}$

$|\cup D_x| \leq 2^n \qquad D_x \cap D_{x'} = \emptyset$ pro $x \neq x'$

nechť $\alpha > 0$ t.č. $\quad R > 1 - H(p) + \alpha \lg \frac{1-p}{p}$

(takové α existuje)

$$\Pr[X+E \text{ se dekóduje správně}] \leq$$

$$< \Pr[\ \text{—}''\text{—} \ \&\ |E| \in [(p-\alpha)n, (p+\alpha)n]]$$

$$+ \Pr[\ \text{—}''\text{—} \ \&\ |E| \notin [(p-\alpha)n, (p+\alpha)n]]$$

$$\leq \Pr[|E| \notin [(p-\alpha)n, (p+\alpha)n]]$$

$$\leq 2^{-c_{p,\alpha} n} \leq \frac{\delta}{2}$$

(čímž volba n dostatečně velké)

$$\leq \frac{\delta}{2} + \Pr[X+E \text{ se dekóduje správně} \ \&\ |E| \in [(p-\alpha)n, (p+\alpha)n]]$$

ukážeme, že $(*) \leq \frac{\delta}{2}$ pro n dostatečně velké

$$(*) \leq \sum_{\substack{x \in C \\ e \in \mathrm{Ball}_{n,pn,\alpha n}(0^n) \\ x+e \in D_x}} \Pr[X=x, E=e]$$

kde $\mathrm{Ball}_{n,pn,\alpha n}(w) = \{ e \in \{0,1\}^n ;\ \Delta_{\mathrm{Ham}}(w,e) \in [(p-\alpha)n, (p+\alpha)n] \}$

$$= \sum_{x \in C} \Pr[X=x] \cdot \sum_{\substack{e \in \mathrm{Ball}_{n,pn,\alpha n}(0^n) \\ x+e \in D_x}} \Pr[E=e]$$

$$\leq \sum_{x \in C} \Pr[X=x] \cdot \sum_{e \in \mathrm{Ball}_{n,pn,\alpha n}(0^n)} \Pr[E=e]$$

$$\leq \sum_{x \in C} \Pr[X=x] \cdot \sum_{k=0}^{(p+\alpha)n} \binom{n}{k} p^k (1-p)^{n-k}$$

$$\leq \sum_{x \in C} \Pr[X=x] \cdot 2^{nH(p+\alpha)} \cdot p^{(p+\alpha)n} (1-p)^{(1-(p+\alpha))n}$$

$$= 2^{nH(p+\alpha)} \cdot p^{(p+\alpha)n} (1-p)^{(1-(p+\alpha))n}$$

$$= 2^{n(H(p+\alpha) + (p+\alpha)\lg p + (1-(p+\alpha))\lg(1-p))}$$

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

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

$$= 2^{n(\underbrace{R - H(p+\alpha) + (p+\alpha)\lg \frac{1-p}{p} + H(p) - R}_{<0})}$$

$$\leq 2^{-cn} \leq \frac{\delta}{2}$$ pro n dostatečně velké

$$\leq \sum_{\substack{x \in C \\ x+e \in D_x \\ e \in \text{shell}_{n,pn,\alpha n}(0^n)}} 2^{-Rn} \cdot p^{(p-\alpha)n} (1-p)^{(1-(p-\alpha))n}$$

$$\leq \left(\bigcup_x |D_x| \right) \cdot 2^{-Rn} \cdot \underbrace{p^{pn} (1-p)^{(1-p)n}}_{2^{-H(p)n}} \cdot \underbrace{p^{-\alpha n} (1-p)^{\alpha n}}_{\left(\frac{1-p}{p}\right)^{\alpha n}}$$

$$\leq 2^{n\left(1 - H(p) + \alpha \lg \frac{1-p}{p}\right)} \cdot 2^{-Rn} \leq 2^{-\varepsilon n} \leq \frac{\delta}{2}$$

pro nějaké $\varepsilon > 0$ (pro n dostatečně velké)