1

- M. Koucký, IUUK, MS, 3. patro
- kdo jste? – pravděpodobnost?, diskrétní matika?, složitost?
- Domácí úkoly – 4 50% bodů, 3/4 sad
  - ↳ pokud pracujete spolu → Ack
- Informace: 1) "V poledne přijdu na oběd"
  2) "President dnes večer odstoupí s tím, že ho již nebaví být prezidentem."
  3) "Hrubá mast uletěla ve dřevě"
- překvapení & · význam

1) málo překvapení
2) hodně překvapení, má význam
3) hodně překvapení, nemá žádný rozumný význam

→ nás zajímá překvapení, význam neumíme analyzovat

↓

pravděpodobnost

Informace:

1) $I(A)$ klesající s rostoucí prst. A

Př: zavolám balíček 32 karet, vzchu karta je

1) červená A
2) sedma B
3) červená sedma C

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

$I(A \,\&\, B) \geq I(A), I(B)$ $\quad I(B) \geq I(A)$

2) aditivita $I(A \,\&\, B) = I(A) + I(B)$

A, B nezávislé

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

5) $I(A) \geq 0 \quad \forall A$

$\rightarrow \boxed{I(A) = -\log_a P(A)}$ pro nějaké zvolené a

→ alternativní přístup: Kolmogorovská složitost

Opakování pstí:

- jev, prostor psti, náhodná proměnná
- nezávislé jevy
- očekávaná hodnota
- markovova nerovnost

(neurčitost)

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

konvence: $0 \cdot \log 0 = 0$

- $H(X) \geq 0$ Dk: $H(X) = -\sum p(x) \log p(x)$

$\log p(x) < 0$ pro $0 < p(x) < 1$

- $H(x) \leq \log |X|$

$\hookrightarrow \mathrm{supp}(X)$

Př: X on $\{0,1\}^n$

1) $\forall x \in X$; $p(x) = 2^{-n}$

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

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

$H(X) = \frac{1}{2} \cdot \log \left(\frac{1}{2}\right)^{-1} + \frac{1}{2} \cdot \log (2^{n+1} - 2) = \frac{n}{2} + \Theta(1)$

2

Př: $H(p) = -p \log p - (1-p) \log (1-p)$

$$X = \begin{cases} 0 & w/\ p \\ 1 & w/\ 1-p \end{cases}$$

společná entropie: X, Y n. p.

$$H(X,Y) = -\sum_{\substack{x \in X \\ y \in Y}} p(x,y) \log p(x,y)$$

$$= -E_{XY} \log p(X,Y)$$

podmíněná entropie

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

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

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

$$= E_{xy} \log p(Y|X)$$

Př: 1) $H(X|X) = 0$

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

Věta ("chain rule")

Věta: ("chain rule")

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

Dk:

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

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

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

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

∎

Důsledek:

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

$$H(X) - H(X|Y) = H(X,Y) =$$
$$= H(Y) - H(Y|X).$$

Vzájemná informace:

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

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

Př: a) $I(X;X) = H(X) - H(X|X) = H(X)$

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

c) X = fair dice, Y = fair dice, $Z = X+Y \bmod 6$

$I(X:Y) = 0$

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

$I(X,Y:Z) = H(X,Y) - \underbrace{H(X,Y|Z)}_{H(X)} = H(Y)$
$= H(Y)$.

d) $X = \{\, 0^n$ s pstí $\frac{1}{2}$

d) $X = \begin{cases} 0^n & \text{s pst } \frac{1}{2} \\ x \in \{0,1\}^n \setminus \{0^n\} & \text{s pst } \frac{1}{2} \cdot \frac{1}{2^n - 1} \end{cases}$

$Y = \begin{cases} 0 & \text{if } X = 0^n \\ 1 & \text{else} \end{cases}$

$H(X) = \frac{n}{2} + O(1)$

$H(X \mid Y=1) \approx n-1$

$H(X \mid Y=0) = 0$

$H(X \mid Y) \leq \frac{n}{2}$

Př: "Secret sharing scheme".

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

- Kullback-Leiblerova vzdálenost: (Kullback-Leibler divergence)

$$D(p \| q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$$

konvence: $0 \cdot \log \frac{0}{0} = 0$

- $D(p \| q) = -\sum_{x \in X} p(x) \log(q(x)) + \sum_{x \in X} p(x) \log p(x)$

konvence: $0 \cdot \log \frac{0}{0} = 0$, $0 \log \frac{0}{q} = 0$, $p \log \frac{p}{0} = \infty$

- uvidíme $D(p \| q) \geq 0$
- míří, o kolik se prodlouží kód, když použiji nesprávnou distribuci

- $I(X;Y) = D\left(p(x,y) \,\|\, p(x) \cdot p(y)\right)$

$= -\sum_{y \in X} p(x) \log p(x) + \sum_{x,y} p(x,y) \log p(x \mid y)$

$= -\sum_{x,y} p(x,y) \log p(x) + \sum_{x,y} p(x,y) \log p(x \mid y)$

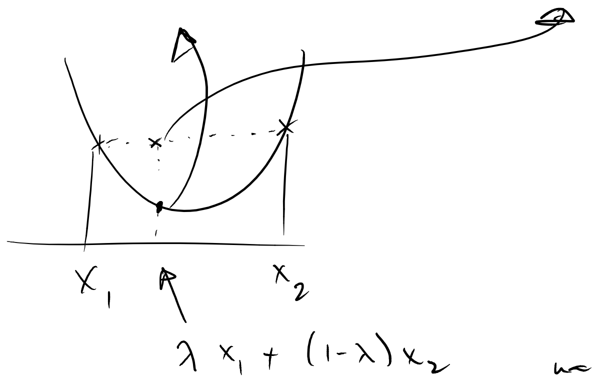
$$= \sum_{x,y} p(x,y) \lg p(x) + \sum_{x,y} p(x,y) \lg p(x|y)$$

$$= \sum_{x,y}' p(x,y) \lg \frac{p(x)}{p(x|y)} = -\sum_{x,y} p(x,y) \lg \frac{p(x)p(y)}{p(x,y)}$$

- fce f je <u>konvexní</u> na (a,b), pokud $\forall x_1, x_2 \in (a,b)$ ③

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

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



- <u>Jensenova nerovnost</u>: f je konvexní fce na (a,b) a X je n.p. s hodnotami z (a,b), pak

$$E[f(X)] \ge f(E[X])$$

<u>Dk</u>: $p(x_1) = \lambda \qquad p(x_2) = (1-\lambda)$

definice konvexnosti $\Rightarrow$ tvrzení

dále indukcí

- $p(x_1) = p_1 \quad p(x_2) = p_2 \quad \dots \quad p(x_n) = p_n$

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

kde $p_i' = \frac{p_i}{(1-p_n)}$

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

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

$$= f\left(\sum_{i=1}^{n} p_n x_n\right)$$

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

<u>Věta</u>: Nechť $\underset{x \in X}{p(x), q(x)}$ jsou pravděpodobnostní rozdělení.

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

<u>Dk</u>: Označ $A = \{x ;\ p(x) > 0\}$. Pak

$$-D(p\|q) = -\sum_{x \in A} p(x) \log \frac{p(x)}{q(x)}$$
$$= \sum_{x \in A} p(x) \log \frac{q(x)}{p(x)}$$
$$\leq \log \sum_{x \in A} p(x) \frac{q(x)}{p(x)}$$
$$\leq \log \sum_{x \in X} q(x)$$
$$= \log 1 = 0 \qquad \blacksquare$$

- Rovnost $D(p\|q) = 0$ nastává pouze pro $p = q$

<u>Důsledek</u>: $I(X:Y) \geq 0$

<u>Dk</u>: $I(X:Y) = D(p(x,y) \,\|\, p(x) \cdot p(y))$ $\blacksquare$

<u>Důsledek</u>: $H(X) \leq \log |X|$, s rovností pouze pokud je X rovnoměrně rozdělená.

<u>Dk</u>: def. $u(x) = \frac{1}{|X|}$, p je distr. pro X

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

$$0 \leq D(p\|u) \Rightarrow H(X) \leq \log |X| \qquad \blacksquare$$

<u>Důsledek</u>: $H(X|Y) \leq H(X)$, rovnost iff X, Y nezávislé

<u>Dk</u>: $0 \leq I(X:Y) = H(X) - H(X|Y)$ $\blacksquare$

Dk.: $0 \leq I(X:Y) = H(X) - H(X|Y)$ ∎

- X, Y, Z náhodné proměnné, <u>podmíněná informace</u> X o Y, když je dáno Z:

$$I(X:Y|Z) = \mathop{E}_{z \in Z} I(X:Y/Z=z)$$
$$= \sum Pr[Z=z] \cdot H(X/Z=z)$$
$$H(X_1, \ldots \quad X_n) = \sum_z Pr[Z=z] \cdot H(X|Y, Z=z)$$
$$= H(X/Z) - \sum_z Pr[Z=z] \sum_y Pr[Y=y|Z=z] \cdot H(X|Y=y, Z=z)$$
$$= H(X|Z) - \sum_{y,z} P[Z=z \,\&\, Y=y] H(X|Y=y, Z=z)$$
$$= H(X|Z) - H(X/Y, Z)$$

- $H(X_1, X_2, X_3, \ldots X_n) = H(X_1) + H(X_2|X_1) \ldots + H(X_n|X_1 \ldots X_{n-1})$
$$\sum_{i=1}^{n} H(X_i | X_1, X_2, \ldots X_{i-1})$$

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

Dk.: $I(X_1 \ldots X_n : Y) =$

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

∎

Př.: Alenka Bob

$$X \longrightarrow Y \overset{g}{\rightsquigarrow} g(Y) = Z \ \ldots \text{odhad } X$$

$$I(X:Y) \text{ vs } I(X:Z) \ ?$$

X, Y, Z náh. proměnné

$I(X:Y) \geq I(X:Z)$?

X, Y, Z náh. proměnné

Def: X, Y, Z splňují markovovskou vlastnost, pokud
$X \to Y \to Z$

$\forall x, y, z$

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

- $p(z|y) = p(z|y, x)$
- $p(x, z|y) = p(x|y)\, p(z|y, x) = \underbrace{p(x|y)\, p(z|y)}_{p(x,z|y) = p(x|y)p(z|y)}$
  $= p(z|y)\, p(x|y, z)$

$\Rightarrow p(x|y) \cdot p(z|y) = p(z|y) \cdot p(x|y, z) \Rightarrow p(x|y, z) = p(x|y)$
$\Rightarrow (X \to Y \to Z \text{ iff } Z \to Y \to X)$

symetrie

Věta (Data processing inequality): X, Y, Z n. p. splňující markovovskou vlastnost. 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 podmíněně na Y

$I(X:Y|Z) \geq 0$ triviálně

$$\Rightarrow I(X:Y) \geq I(X:Z)$$ ∎

Fanova nerovnost

$$X \to Y \to \hat{X}$$

$P_e = \Pr[X \neq \hat{X}]$ ... pst. chyby

- $H(P_e) + P_e \log |X| \geq H(X|\hat{X}) \geq H(X|Y)$

$p_e = \Pr[\hat{X} \neq X]$ ...

- $H(p_e) + p_e \log|X| \geq H(X|\hat{X}) \geq H(X|Y)$

Důsledek: $1 + p_e \log|X| \geq H(X|Y)$

neboli $p_e \geq \dfrac{H(X|Y) - 1}{\log|X|}$

- $p_e = 0 \Rightarrow H(X|Y) = 0$ ... odpovídá intuici

Dk:

def. n.p. $E = \begin{cases} 1 & \hat{X} \neq X \\ 0 & \hat{X} = X \end{cases}$

$$H(X, E | \hat{X}) = H(X|\hat{X}) + \underbrace{H(E | X, \hat{X})}_{=0}$$

$$= \underbrace{H(E|\hat{X})}_{\leq H(p_e)} + H(X | E, \hat{X})$$

$$H(X | E, \hat{X}) \leq \Pr[E=0]\, H(X|\hat{X}, E=0) + \Pr[E=1]\, H(X|\hat{X}, E=1) \leq 0 + p_e \log|X|$$

$\Rightarrow H(X|\hat{X}) \leq H(p_e) + p_e \log|X|$.

podle data-processing ineq. $I(X:\hat{X}) \leq I(X:Y)$

$\Rightarrow H(X|\hat{X}) \geq H(X|Y)$ ∎

Důsledek: $\forall X, Y$ n.p., $p = \Pr[X \neq Y)$

$$H(p) + p \log|X| \geq H(X|Y)$$

Dk: $\hat{X} = Y$ & use Fano. ∎

Lze vylepšit pokud $\hat{X} \subseteq X$ na

$$H(p_e) + p_e \log(|X| - 1) \geq H(X|Y)$$

Lze vylepšit pokud $\hat{X} \subseteq \mathcal{X}$ ⇝

$$H(p_e) + P_e \lg(|\mathcal{X}|-1) \geq H(X|Y)$$

- Těsné pro Y konst., $\mathcal{X} = \{1, \dots m\}$, $\left(1-p_e, \frac{p_e}{m-1}, \dots \frac{p_e}{m-1}\right)$

$$H(p_e) + p_e \lg(m-1) \geq H(X)$$

$1-p_e \geq \frac{p_e}{m-1}$

- X, X' n.p. i.i.d. $\Pr[X = X'] = \sum_x p^2(x)$

$$\Pr[X = X'] \geq 2^{-H(X)}$$

(rovnost $\iff$ X je uniformní)

Dk: $X \sim p(x)$.

$$2^{\mathbb{E} \lg p(x)} \leq \mathbb{E}\, 2^{\lg p(x)}$$

Jensenova nerovnost

$$2^{-H(X)} = 2^{\sum_x p(x) \lg p(x)} \leq \sum_x p(x)\, 2^{\lg p(x)} = \sum_x p^2(x)$$

"Asymptotic Equipartition property"

$X_1, X_2, \dots$ nezávislé náhodné rozdělené jako X

Tvrzení: $\forall \varepsilon, \delta > 0 \ \exists n_0 \ \forall n \geq n_0$

$$\Pr\left[2^{-n(H(X)+\varepsilon)} \leq p(x_1, x_2, \dots x_n) \leq 2^{-n(H(X)-\varepsilon)}\right] \geq 1-\delta.$$

Připomenutí: • $\mathrm{Var}[X] = \mathbb{E}\left[(X - E(X))^2\right]$

- pokud $X_1, X_2, \dots X_n$ vzájemně nezávislé pak

$$\mathrm{Var}[X_1 + X_2 + \dots + X_n] = \mathrm{Var}(X_1) + \dots + \mathrm{Var}(X_n)$$

- Čebyševova nerovnost: $\Pr[|X - E(X)| \geq a] \leq \frac{\mathrm{Var}[X]}{}$

• Čebyševova ner.: $\Pr[|X - \mathbb{E}(X)| \geq a] \leq \frac{\mathrm{Var}[X]}{a^2}$

Dk: uvažujme $-\log p(x_1), -\log p(x_2), \ldots -\log p(x_n)$ n.p.

$$\mathbb{E}[-\log p(x_i)] = H(X_i) = H(X)$$

$$\Pr\left[\left|\sum_{i=1}^{n} -\log p(X_i) - \mathbb{E}\left[\sum_{i=1}^{n} -\log p(X_i)\right]\right| \geq \varepsilon n H(X)\right] =$$

$$= \Pr\left[\left|\sum_{i=1}^{n} -\log p(X_i) - nH(X)\right| \geq \varepsilon n H(X)\right]$$

$$\leq \frac{n \cdot \mathrm{Var}[-\log p(x)]}{\varepsilon^2 n^2 [H(X)]^2} = \text{const.} \frac{1}{n} \xrightarrow[n\to\infty]{} 0$$

$\rightarrow$ s velkou pstí, vzorek který dostaneme z $X_1, \ldots X_n$ má pst. $\approx n H(X)$.

• $A_\varepsilon^n = \{(x_1, \ldots x_n) \in X_1 \times X_2 \ldots X_n$, t.ž. $2^{-n(H(X)+\varepsilon)} \leq p(x_1, \ldots x_n) \leq 2^{-n(H(X)-\varepsilon)}\}$

$(1-\varepsilon)2^{n(H(X)-\varepsilon)} \leq |A_\varepsilon^n| \leq 2^{n(H(X)+\varepsilon)}$ pro dostatečně velké n.

$\Rightarrow$ kódování n-tic

$(y_1, y_2, \ldots y_n) \in A_\varepsilon^n \rightarrow$ index v rámci A_ε^n

kód $\rightarrow 0\,i$

$(y_1 \ldots y_n) \notin A_\varepsilon^n \rightarrow$ index v rámci X^n

kód $\rightarrow 1\,i$

$C(y_1 \ldots y_n) \ldots$ $\to$ indexy v rámci $\sim$

kód $\to$ 1

očekávaná délka kódu:

$$\leq n(H(X)+\varepsilon) + \delta n \log|X| + 2$$

$$\text{pro } \delta = \frac{\varepsilon}{\log|X|}$$

$$\to n(H(X)+2\varepsilon)+2$$

pro libovolné $\varepsilon' > 0$ lze zvolit vhodné n, t.ž. kódujeme s očekávanou délkou kódu

$$n(H(X)+\varepsilon')$$

Př: $\Pr[X=0] = 1/10$ $\Pr[X=1] = 9/10$

$$\Pr[X = 1111\ldots 1] = \left(\frac{9}{10}\right)^n$$

maximální prst.

ale takové tímu nevidíme.

---

kód: $C: X \to \Sigma^*$ $\forall x \neq y \;\; C(x) \neq C(y)$

očekávaná délka: $L(C) = \sum_{x \in X} p(x)\, l(x)$

uzávěr kódu: $C^*(x_1, \ldots x_k) = C(x_1)C(x_2)\ldots C(x_k)$

- kód je jednoznačně dekódovatelný pokud uzávěr kódu nemá kolizi.
- bezprefixový kód C: $\forall x \neq y$ $C(x)$ není prefix $C(y)$

Př: 10 00 0 10 110

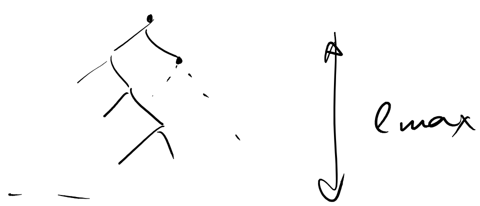
Př.:

10	0
00	10
11	110
110	111
jednoznačně dekódovatelný, ale ne bezprefixový	bezprefixový

• <u>Kraftova nerovnost</u>: kód s délkami $l_1, l_2, l_3, \ldots$ je bezprefixový $\Rightarrow$

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

Dk:



$$\sum 2^{l_{max}-l_i} \le 2^{l_{max}}$$

$$\Rightarrow \sum 2^{-l_i} \le 1$$ ■

• Opačná implikace $\sum 2^{-l_i} \le 1 \Rightarrow \exists$ bezprefixový kód.

• Optimální kód pro danou distribuci X: $l(x) \quad x \in X$

očekávaná délka $L = \sum_{x \in X} p(x)\, l(x)$

$$L - H(X) = \sum p(x) l(x) - \sum p(x) \lg \frac{1}{p(x)} =$$

$$= -\sum p(x) \cdot \lg 2^{-l(x)} + \sum p(x) \lg p(x)$$

$$q(x) = \frac{2^{-l(x)}}{2^{-l(y)}} \qquad c = \sum_y 2^{-l(y)}$$

$$= -\sum p(x) \lg q(x) + \sum p(x) \lg p(x)$$

$$= \underbrace{\sum p(x) \lg \frac{p(x)}{q(x)}}_{D(p\|q) \geq 0} - \underbrace{\lg_2 c}_{\geq 0} \Leftarrow c \in (0,1]$$

$$\geq 0$$

- <u>Shannon-Fano kód</u>: $l(x) = \lceil \lg \frac{1}{p(x)} \rceil$.
  - $H(X) \leq L(C) \leq H(X) + 1$
- opakování – bereme k-tice a ty kódujeme

$$kH(X) = H(X_1, X_2, \ldots X_k) \leq L(C) \leq H(X_1, \ldots X_k) + 1$$

$$\Rightarrow \quad L(\text{"}C_k\text{"}) \leq H(X) + \frac{1}{k}.$$

- <u>Mc Millan</u>: jednoznačně dekódovatelný kód s $l(x)$

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

$\rightarrow$ zobecnění Kraftovy nerovnosti.

<u>Dk</u>: uvažme C^k

$$\left(\sum_x 2^{-l(x)}\right)^k = \sum_{x_1} \sum_{x_2} \ldots \sum_{x_k} 2^{-l(x_1) - l(x_2) - \ldots - l(x_k)}$$

$$= \sum_{\bar{x} \in X^k} 2^{-l(\bar{x})}$$

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

$c(m) = \#$ slov s délkou m

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

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

$$\sum_x 2^{-l(x)} \leq (k\, l_{max})^{1/k} \xrightarrow[k \to \infty]{} 1 \qquad \blacksquare$$

- <u>Huffmanův kód</u>

$$p_1 \geq p_2 \geq \ldots \geq p_{n-1} \geq p_n$$

$\to p' = (p_1, p_2, \dots, p_{n-1} + p_n)$ ⇝ repeat

- optimální bezprefixový kód má bez újmy na obecnosti
  1) s pravd. klesá délka slov
  2) dvě nejdelší slova jsou stejně dlouhá
  3) ∃ dvě nejdelší slova, která se liší jen v posledním bitě

Dk: triv.

optimalita Huffmana

$p = (p_1, p_2, \dots, p_n)$

$p' = (p_1, p_2, \dots, p_{n-1} + p_n)$

opt. kód pro p & p', jejich délka $\ell^*(p)$ a $\ell^*(p')$
$C^*(p)$ $C^*(p')$

z $C^*(p')$ udělej kód pro p rozšířením slova pro $p_{n-1} + p_n$ $\to \ell(p)$ (C')

z $C^*(p)$ udělej kód pro p' uříznutím posledního bitu dvou nejdelších slov lišících se v posledním bitu. $\to \ell(p')$

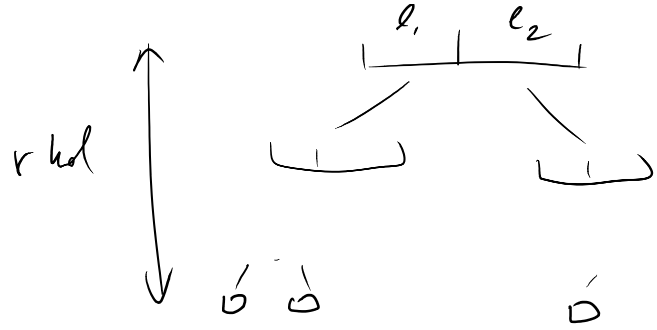
$$\ell(p') = \ell^*(p) - p_{n-1} - p_n$$
$$\ell(p) = \ell^*(p') + p_{n-1} + p_n$$

$$\Rightarrow \quad \underbrace{(\ell(p') - \ell^*(p'))}_{\geq 0 \text{ opt.}} + \underbrace{(\ell(p) - \ell^*(p))}_{\geq 0 \text{ opt.}} = 0$$

$\Rightarrow$ $= 0$ $= 0$

$\Rightarrow$ "$C^*(p')$ optimální $\Rightarrow$ C' optimální"

- Dělení seznamu v k kolech



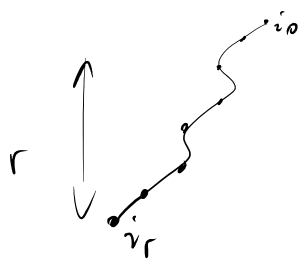
$\ell = \ell_1 + \ell_2$

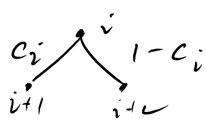
cena $\min(\ell_1, \ell_2)$

Jaká je celková cena dělení seznamu délky N v r kolech

$B(N,r)$

- $B(N,r) \geq \dfrac{N \log N}{4 \log\left(\frac{4r}{\log N}\right)}$





$c_i = \dfrac{N_{i+1}}{N}$

$c_i \in [0, \frac{1}{2}]$

$$\prod_{i=1}^{r} c_{i_j} = \frac{1}{N}$$

$$\rightarrow \quad \left(\frac{1}{N}\right)^N = \prod c_i^{c_i N_i} (1-c_i)^{(1-c_i) N_i}$$

$$N \log N = \sum_i N_i H(c_i)$$

- $\sum_i N_i = rN$

  i je vnitřní uzel

- $c \overset{def}{=} \displaystyle\sum_{i \text{ vnitřní}} \frac{c_i N_i}{rN}$ průměrná cena za prvek při dělení

- $B(N,r) = c \cdot rN$

- H je konkávní $\Rightarrow$

$$\sum_i \frac{N_i}{rN} H(c_i) \leq H\left(\sum \frac{c_i N_i}{rN}\right) = H(c).$$

$\to \quad N \log N = rN \sum \frac{N_i}{rN} H(c_i) \leq rN H(c)$

Pokud $B(N,r) = crN < \dfrac{N \log N}{4 \log\left(\frac{4r}{\log N}\right)}$

$\to \quad c < \dfrac{\log N}{4r \log\left(\frac{4r}{\log N}\right)} \qquad c = \dfrac{x}{-\log x}$ pro $x = \dfrac{\log N}{4r}$

$\dfrac{\log N}{r} \leq H(c) = H\left(\dfrac{x}{-\log x}\right) < 4x = 4\dfrac{\log N}{r4} = \dfrac{\log N}{r}$

spor

Lemma: $\forall x \in (0, \frac{1}{2}], \quad H\left(\dfrac{x}{\log 1/x}\right) < 4x$

Dk: $\quad y = \dfrac{x}{\log 1/x} \qquad x \in (0, 1/2] \Rightarrow y \in (0, \frac{1}{2}]$

$H(y) = y \log \frac{1}{y} + (1-y) \log \frac{1}{1-y}$

$\forall 0 < y \leq \frac{1}{2} \qquad \log \frac{1}{1-y} \leq 2y \qquad 1-y < 1$

$H(y) \leq y \log \frac{1}{y} - (1-y) 2y \leq y \log \frac{1}{y} + 2y$

$\to H\left(\dfrac{x}{\log 1/x}\right) < x \underbrace{\left(\dfrac{\log\log 1/x}{\log 1/x}\right)}_{\leq 1} + x \underbrace{\left(\dfrac{\log \frac{1}{x}}{\log \frac{1}{x}}\right)}_{\leq 1} + 2x \underbrace{\left(\dfrac{1}{\log 1/x}\right)}_{\leq 1}$

$\leq 4x$ ■

* viz Brief introduction to Kolmogorov Complexity na mojí webové stránce

Symetrie informace

$C(x,y) = C(x) + C(y|x) + O(\log n)$

- f ... částečně rekurzivní funkce $\quad f: \{0,1\}^* \to \{0,1\}^*$

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

- <u>Věta</u>: Existuje <u>univerzální</u> část. rek. fce $\mathcal{U}$ t.ž.

$\forall$ č.r.f. $g \;\; \exists c_g \;\; \forall x \in \{0,1\}^*$

$$C_{\mathcal{U}}(x) \le C_g(x) + c_g$$

<u>Dk</u>: ... $\blacksquare$

$\Rightarrow \mathcal{U}$ dává nejmenší složitost ze všech č.r.fcí

$\rightsquigarrow$ zafixujeme $\mathcal{U}$

- $\forall x \qquad C(x) \le |x| + O(1)$

$\forall n \qquad C(0^n) \le |n| + O(1)$

$\hookrightarrow \sim \log n$

podmíněná Kolmogorovská složitost:

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

<u>Věta</u>: $\exists$ č.r.f. $\mathcal{U}$ t.ž. $\forall$ č.r.f. $g \; \exists c_g \;\; \forall x,y$

$$C_{\mathcal{U}}(x|y) \le C_g(x|y) + c_g$$

$\rightarrow$ zafixujeme $\mathcal{U}$.

$$C(x) \stackrel{def}{=} C(x|\varepsilon)$$

- $\forall n \qquad C(0^n|n) = O(1)$

$\forall n \; \exists x \in \{0,1\}^n \qquad C(x|n) \ge n$ ... Kolmogorovsky náhodné

<u>Dk</u>: počet programů délky $\le n-1$ je $2^n - 1$. $\blacksquare$

- A rekurzivně spočetná množina

$$\forall x \in A,\ |x| = n \qquad C(x|n) \le \log|A^{=n}| + O(1).$$

- $C(x,y) \le C(x) + C(y|x) + O(\log C(x,y))$

Dk: program p pro x
program q pro y když znám x

$O(\log C(x,y))$ extra info na separaci p a q ⊠

- $C(x,y) \ge C(x) + C(y|x) - O(\log C(x,y))$

Dk: sporem. Předpokládejme, že pro každou konstantu k $\exists x,y$ t.ž.

$(*)$ $C(y|x) \ge C(x,y) - C(x) + k \log C(x,y)$.

def. $A = \{ \langle u,z\rangle ;\ C(u,z) \le C(x,y)\}$
$A_x = \{ z ;\ C(x,z) \le C(x,y)\}$ ... rek. spočetná

$(**)$ $C(y|x) \le \log |A_x| + 2 \log C(x,y) + O(1)$

$\underset{(*)(**)}{\Rightarrow}$ $\forall k\ \exists x,y$
$|A_x| \ge 2^{\ell}$ $\qquad \ell = C(x,y) - C(x) + k \log C(x,y)$

→ nalezneme příliš krátký pgm pro x:
uvažujme množinu $\mathcal{U}$ t.ž. $\forall u \in \mathcal{U}$
$A_u = \{ z ;\ C(z,u) \le C(x,y)\}$
$|A_u| \ge 2^{\ell}$

1) $x \in \mathcal{U}$, $\mathcal{U}$ je rek. spočetná
2) $\{ \langle u,z\rangle ;\ u \in \mathcal{U}, z \in A_u\} \subseteq A$
3) $|A| \le 2^{C(x,y)+1}$

$\rightarrow \quad |\mathcal{U}| \le \frac{|A|}{2^{\ell}} \le \frac{2^{C(x,y)+1}}{2^{\ell}}$

x lze zkonstruovat z $C(x,y)$, ℓ a indexu v rámci množiny $\mathcal{U}$

$$C(x) \le 2\log C(x,y) + 2\log \ell + \overbrace{C(x,y) - \ell}^{\text{indexu}} + O(1)$$

$\rightarrow$ $C(x) < C(x)$ pro dostatečně velké k.

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

<u>Symetrie informace</u>: $I_C(x:y) = I_C(y:x) + O(\log C(x,y))$

<u>Př.</u>: $\forall n\ \exists x \quad |x| = n \qquad C(x|n) \ge n$

pokud $C(n) \ge \log n$ pak

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

$$I_C(n:x) = C(n) - \underbrace{C(n|x)}_{=0} \ge \log n.$$

$\rightarrow$ logaritmická ztráta nutná

<u>Věta</u>: $H(X_n) - O(\log n) \le E[C(X_n)] \le H(X_n) + O(\log n)$

$X_1, X_2, \dots X_n \dots$ rekurzivní posloupnost pravděpodobnostních distribucí na $\{0,1\}^n$

<u>Dk</u>: • $E[C(X_n)] \le H(X_n)$

Huffmanův kód / Shannon-Fano kód

$$E[\ell(X_n)] \le H(X_n)$$

$$C(X_n) \le \ell(X_n) + 2\log n + O(1)$$

kóduje délku, tj. n

• fix n. X_n

$i = 0, \dots, 2n$ $\quad V_i = \{ x \in \Sigma^n ; \ 2^{i-1} \le \Pr[x = X_n] \le 2^{-i} \}$

$g_i = \Pr[X_n \in V_i]$ $\quad V_{2n} = \{x, \Pr[x = X_n] \le 2^{-2n}\}$

$$|V_i| \geq \frac{g_i}{2^i} \Rightarrow V_i' = \{ x \in V_i, \ C(x) \geq \log \frac{g_i}{2^i} - \log n \}$$

- $\Pr[X_n \in V_i'] \geq g_i (1 - \frac{1}{n})$
- $|V_i'| \geq \frac{|V_i|}{2}$

$$H(X_n) \leq \sum_{i=0}^{2n-1} g_i \cdot i + 1 \quad \text{(ex)} \qquad \text{pro } n \geq 10$$

$$E[C(X_n)] \geq \sum_{i=0}^{2n-1} g_i (1 - \tfrac{1}{n}) \cdot \left(\log \frac{g_i}{2^i} - \log n \right)$$

$$\geq (1 - \tfrac{1}{n}) \left[\sum_{i=0}^{2n-1} g_i \log g_i + \sum_{i=0}^{2n-1} g_i \cdot i - \log n \right]$$

$$\geq (1 - \tfrac{1}{n}) \left[-\log 2n + \sum_{i=0}^{2n-1} g_i \cdot i - \log n \right]$$

$$\geq \sum_{i=0}^{2n-1} g_i \, i - 2 \log n - O(1) \qquad \blacksquare$$