

# Samoopravné kódy

Alexandr Kazda

Univerzita Karlova

22. dubna 2020

- $\mathcal{A}$  informační zdroj
- Pro jednoduchost: Zdroj nemá paměť, jednotlivé znaky jsou nezávislé stejně rozdělené náhodné veličiny
- Tedy informační zdroj = pravděpodobnostní rozdělení nad písmeny abecedy  $\Sigma$
- Funkce entropie  $H(\mathcal{A}) = -\sum_{i=1}^k p_i \log p_i$
- T. Kaiser:  $k = 2$  a tedy  $H(p) = p \log(1/p) + (1 - p) \log(1/(1 - p))$
- Informace měří naši míru překvapení zprávami ze zdroje  $\mathcal{A}$

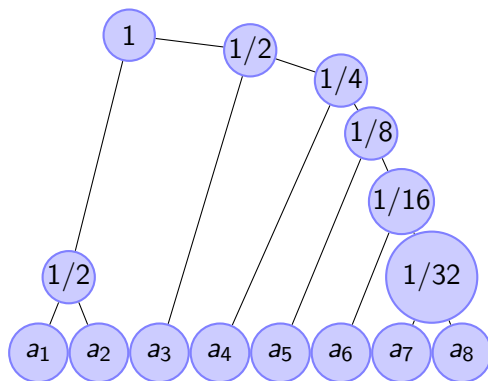
- Značme binární slova  $\{0, 1\}^*$
- Kód = zobrazení  $\Sigma \rightarrow \{0, 1\}^*$
- Kódová slova (obrazy písmen) nemusí být stejně dlouhá. . .
- Vyžadujeme, aby se žádná dvě slova nad  $\Sigma$  nekódovala na stejné binární slova – **jednoznačně dekódovatelný kód**

- Bud'  $\Sigma = \{1, 2, \dots, k\}$ ; kódová slova  $h_1, \dots, h_k \in \{0, 1\}^*$
- **Prefixový kód:** Žádné kódové slovo není prefixem (předponou) jiného
- Necht' pro nějaké  $i_1, \dots, i_n$  a  $j_1, \dots, j_m$  máme

$$h_{i_1} h_{i_2} \cdots h_{i_n} = h_{j_1} h_{j_2} \cdots h_{j_m}$$

- Pak  $h_{i_1}$  je předponou  $h_{j_1}$  nebo naopak
- $\Rightarrow i_1 = j_1$
- Indukcí  $i_2 = j_2, i_3 = j_3 \dots$

# Huffmanovo kódování



- Máme písmena s pravděpodobnostmi

$$p_1 > p_2 > \dots > p_k$$

- Sloučíme  $(k - 1)$ -ní a  $k$ -té písmeno a iterujeme
- Postavíme si tak rozhodovací strom

- Typicky kódová slova různých délek
- Můžeme dekódovat pomocí rozhodovacího stromu
- Kódová slova = cesty z kořene do listů
- Žádné kódové slovo není prefixem druhého – **prefixový kód**
- Očekávaná délka kódového slova vyjde blízko entropie

## Theorem

*Bud'  $\mathcal{A}$  informační zdroj s nezávislými znaky  $\{1, 2, \dots, k\}$  s pravděpodobnostním rozdělením písmen  $(p_1, p_2, \dots, p_k)$ . Potom očekávaná délka kódového slova z Huffmanova kódování  $\mathcal{A}$  leží v intervalu  $[H(\mathcal{A}), H(\mathcal{A}) + 1)$ .*

- Očekávaná délka je

$$p_1 \ell(h_1) + p_2 \ell(h_2) + \dots + p_k \ell(h_k),$$

kde  $\ell(h_1), \dots, \ell(h_k)$  jsou kódová slova Huffmanu.

- Důkaz horního odhadu vynecháme (dělá se to přes to, že Huffman má minimální očekávanou délku mezi všemi prefixovými kódy)
- Dolní odhad viz dále

## Theorem

*Bud'  $C: \Sigma \rightarrow \{0,1\}^*$  jednoznačně dekódovatelný. Bud'  $\ell_1, \dots, \ell_k$  délky kódových slov kódu  $C$ . Potom*

$$\sum_{i=1}^k 2^{-\ell_i} \leq 1.$$

- Intuitivní význam: Délky kódových slov nemohou být libovolně malé
- Intuitivní argument: Vygeneruji si náhodné (nekonečné) binární slovo. Jaká je pravděpodobnost, že se dekóduje na něco začínajícího na 1?  $2^{-\ell_1}$ .
- Jevy „první písmeno se dekóduje na  $i$ “ a „první písmeno se dekóduje na  $j$ “ jsou disjunktní, tedy součet jejich pstí je  $\leq 1$



# Kraft-McMillanova věta, důkaz [dle S. Roman: Introduction to Coding and Information Theory]

- Bud'  $\alpha_\ell$  počet kódových slov délky  $\ell$
- Bud'  $m$  maximální délka kódového slova
- Mějme pro spor

$$\sum_{i=1}^k 2^{-\ell_i} = \sum_{\ell=1}^m \alpha_\ell 2^{-\ell} > 1$$

- Zvolme velké  $u \in \mathbb{N}$  a pojďme si hrát

$$\begin{aligned} \left( \sum_{\ell=1}^m \alpha_\ell 2^{-\ell} \right)^u &= \sum_{\ell_1, \dots, \ell_u} \alpha_{\ell_1} \alpha_{\ell_2} \cdots \alpha_{\ell_u} 2^{-\ell_1 - \ell_2 - \dots - \ell_u} = \\ &= \sum_{k=m}^{um} \sum_{\ell_1 + \dots + \ell_u = k} \alpha_{\ell_1} \alpha_{\ell_2} \cdots \alpha_{\ell_u} 2^{-k} \end{aligned}$$

- Značme

$$N_k = \sum_{\ell_1 + \dots + \ell_u = k} \alpha_{\ell_1} \alpha_{\ell_2} \cdots \alpha_{\ell_u}$$

- Čeho je  $N_k$ ?
- Mezi binárními slovy délky  $k$  zvolme ta, která lze dekodovat na slovo délky  $u$  nad  $\Sigma$
- Takových slov je přesně  $N_k$  (tady potřebuji jednoznačnou dekodovatelnost)
- Jsou to binární slova délky  $k$ , tedy  $N_k \leq 2^k$

# Kraft-McMillanova věta, důkaz III

- Pro spor ať  $\sum_{i=1}^k 2^{-\ell_i} > 1$ .
- Odvodili jsme

$$\begin{aligned}\left(\sum_{\ell=1}^m \alpha_{\ell} 2^{-\ell}\right)^u &= \sum_{\ell_1, \dots, \ell_u} \alpha_{\ell_1} \alpha_{\ell_2} \cdots \alpha_{\ell_u} 2^{-\ell_1 - \ell_2 - \dots - \ell_u} = \\ &= \sum_{k=m}^{um} N_k 2^{-k} \\ &\leq \sum_{k=m}^{um} 1 \leq um\end{aligned}$$

- Co je na tom divného?
- Levá strana roste s rostoucím  $u$  exponenciálně, pravá lineárně
- Tedy pro dost velké  $u$  nebude nerovnost platit

# Entropie jako funkce pravděpodobností

$$H(p_1, \dots, p_k) = \sum_{i=1}^k p_i \log(1/p_i) = - \sum_{i=1}^k p_i \log p_i$$

- Funkce  $f(p) = -p \log p$  je konkávní ( $f''(p) = -1/p < 0$  pro  $p > 0$ )
- $H(p_1, \dots, p_k)$  je součet konkávních funkcí  $\Rightarrow$  konkávní
- Chceme maximalizovat  $H$  na rovině  $p_1 + \dots + p_k = 1$
- Gradient  $H$  má být kolmý na tuto rovinu

$$\nabla H = (\partial H / \partial p_1, \dots, \partial H / \partial p_k) = \lambda(1, 1, \dots, 1)$$

- $\partial H / \partial p_i = -\log p_i - 1 = \lambda$
- Všechna  $p_i$  jsou stejná, tedy  $p_i = 1/k$
- Maximální entropie je  $\log k$

- Pro libovolnou konkávní funkci  $f(x)$  platí, že kdykoli máme  $\alpha_1, \dots, \alpha_k \geq 0$  splňující  $\sum_{i=1}^k \alpha_i = 1$  a  $x_1, \dots, x_k \in \text{dom } f$ , tak

$$\sum_{i=1}^k \alpha_i f(x_i) \leq f\left(\sum_{i=1}^k \alpha_i x_i\right)$$

(a pravá strana je definovaná)

- Funguje to třeba pro entropii...
- ...ale my to časem použijeme pro logaritmus

# Důsledek: Entropie dává dolní odhad na délku kódu

## Theorem

*Bud'  $\mathcal{A}$  informační zdroj s nezávislými znaky  $\{1, 2, \dots, k\}$  s pravděpodobnostním rozdělením písmen  $(p_1, p_2, \dots, p_k)$ . Bud'  $C$  jednoznačně dekódovatelný kód. Pak průměrná délka kódového slova je aspoň entropie zdroje  $\mathcal{A}$ .*

- Značme  $\ell_1, \dots, \ell_k$  délky kódových slov
- Průměrná délka kódového slova je  $\sum_{i=1}^k p_i \ell_i$
- Chceme  $\sum_{i=1}^k p_i \log(1/p_i) - \sum_{i=1}^k p_i \ell_i \leq 0$

# Entropie dává dolní odhad na délku kódu

- Chceme  $\sum_{i=1}^k p_i \log(1/p_i) - \sum_{i=1}^k p_i \ell_i \leq 0$
- Levá strana je  $\sum_{i=1}^k p_i \log(2^{-\ell_i}/p_i)$
- Jensenova nerovnost pro  $f(x) = \log x$ ,  $\alpha_i = p_i$ ,  $x_i = 2^{-\ell_i}/p_i$

$$\sum_{i=1}^k p_i \log(2^{-\ell_i}/p_i) \leq \log \left( \sum_{i=1}^k p_i 2^{-\ell_i}/p_i \right) = \log \left( \sum_{i=1}^k 2^{-\ell_i} \right)$$

- Kraft-McMillan nám dá  $\sum_{i=1}^k 2^{-\ell_i} \leq 1$

- Dokázali jsme, že bez šumu je průměrná délka (binárního) kódu aspoň entropie zdroje
- Entropie = míra informace
- Dolní mez na míru informace, kterou napěchujeme do  $\{0, 1\}^n$  je  $n$
- Jak to bude, když posíláme zprávy skrz kanál, co dělá chyby? Uvidíme později



- $V(n, r)$  buď objem koule v  $\{0, 1\}^n$  o poloměru  $r$
- $V(n, r) = \sum_{i=0}^r \binom{n}{i}$
- Zhruba  $H(r/n, 1 - r/n) = \frac{1}{n} \log(V(n, r))$

## Theorem

*Bud'  $n \in \mathbb{N}$ . Pro všechna  $0 \leq r \leq n/2$  platí*

$$V(n, r) < 2^{nH(r/n, 1-r/n)}$$

# $V(n, r) < 2^{nH(r/n, 1-r/n)}$ [Drobné překlepy opraveny]

- Proč by něco takového mohlo platit: Pro  $r \ll n$  je zhruba

$$V(n, r) \approx \binom{n}{r} \approx \frac{n^n}{r^r (n-r)^{(n-r)}} = \left(\frac{n}{r}\right)^r \left(\frac{n}{n-r}\right)^{n-r}$$

- Vezmeme logaritmus:

$$\begin{aligned} \log(V(n, r)) &\approx r \log(n/r) + (n-r) \log(n/(n-r)) = \\ &= n \left( \frac{r}{n} \log(n/r) + \frac{n-r}{n} \log(n/(n-r)) \right) \\ &= nH(r/n, 1-r/n) \end{aligned}$$

- Víme něco exaktně? Ano, máme

$$\begin{aligned} nH(r/n, 1-r/n) &= -n \left( \frac{r}{n} \log(r/n) + \left(1 - \frac{r}{n}\right) \log(1-r/n) \right) \\ &= -r \log r/n + (r-n) \log(1-r/n) \end{aligned}$$

$$V(n, r) < 2^{nH(r/n, 1-r/n)}$$

- $nH(r/n, 1-r/n) = -r \log r/n + (r-n) \log(1-r/n)$
- Dosadíme do exponenciální funkce:

$$\begin{aligned} 2^{nH(r/n, 1-r/n)} &= 2^{-r \log(r/n) + (r-n) \log(1-r/n)} = \\ &= \left(\frac{r}{n}\right)^{-r} (1-r/n)^{r-n} = \frac{n^n}{r^r (n-r)^{n-r}} \end{aligned}$$

- To už vypadá skoro jako kombinační číslo...

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

- Protože  $r \leq n/2$ , tak  $r^i (n-r)^{n-i} \geq r^r (n-r)^{n-r}$  pro  $i \leq r$
- Tedy

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

$$V(n, r) < 2^{nH(r/n, 1-r/n)}$$

- Máme

$$2^{nH(r/n, 1-r/n)} = \frac{n^n}{r^r (n-r)^{n-r}}$$

- A zároveň

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

- Po vydělení druhé nerovnosti  $r^r (n-r)^{n-r}$  dosazení do první rovnosti:

$$2^{nH(r/n, 1-r/n)} > \sum_{i=0}^r \binom{n}{i} = V(n, r).$$