Uvažme libovolný kód $C: X \to \Sigma^*$. $L(C)$ vs $H(X)$?

- $H(X) - 2H(X) - 2 \leq L(C)$
- můžeme udělat z C bezprefixový kód tak, že ke každému slovu přidáme jeho délku zdvojenou a oddělíme 01

nappř. $C(x) = 0011$ ⇝ $\underbrace{1100}0 0\overset{\text{zarážka}}{\widetilde{01}}\,C(x)$

4 binárně a každý bit je zdvojený

Generování náhodných čísel dle zadání distribuce vs entropie

- komprese: je dána distr. objektů, vymyslete nejkratší reprezentaci
    - 01 reprezentace ≈ sekvence hodů mincí.
- obrácená otázka: chceme samplovat z nějaké pstní distribuce, kolik potřebujeme hodů mincí?

Př:

$$X = \begin{cases} a & \text{s pstí } 1/2 \\ b & \text{s pstí } 1/4 \\ c & \text{s pstí } 1/4 \end{cases}$$

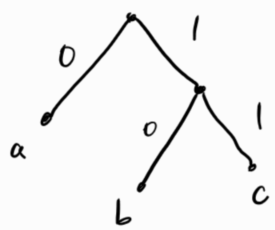
$0 \to a$

$10 \to b$ → průměrně 1.5 bitu

$11 \to c$

$$H(X) = \frac{1}{2}\lg 2 + \frac{1}{4}\lg 4 + \frac{1}{4}\lg 4 = \frac{1}{2} + \frac{1}{2} + \frac{1}{2} = 1.5$$ ... náhoda?

- obrázek



pořadí?

- vnímejme tohle jako algoritmus mapující X na prvky binárního stromu
- vlastnosti
    i) strom by měl být úplný (≈ žádné vrcholy s 1 potomkem)
    ii) pst dosažení listu v hloubce k je 2^{-k}
    iii) očekávaný # hozených bitů ⟺ očekávaná hloubka stromu
- tak nějak čekáme, že náhodnost distribuce ⟺ # bitů ke samplování neboť bit má entropii 1
- pro vrchol v značíme jeho hloubku $h(v)$ a hloubka stromu $h(T)$

Lem: Pro lib úplný strom (tj. # potomků je vždy 0 nebo 2) uvažme pstní rozdělení t.ž. na listy v hloubce k dáme 2^{-k}. Ukažte, že $E[h(T)] = H$(pstní distr. na stromě)

Dk:

· $\mathbb{E}[h(T)] = \sum_{\ell \in listy} h(\ell) \, 2^{-h(\ell)}$

· $H\left(\begin{matrix}\text{distr na} \\ \text{listech}\end{matrix}\right) = -\sum_{\ell \in listy} 2^{-h(\ell)} \lg 2^{-h(\ell)} = \mathbb{E}[\text{hloubka}]$ ← algebra autopilot

□

Věta: Pro lib. alg samplující X, očekávaný # bitů je aspoň entropie X.

↑ tj. $h(T)$ (což je n.v.)

Dk:

· vyvažíme náhodnost lze zaznamenat stromem T jako výše

· označme listy stromu pomocí $\{1, 2, \dots\}$ (pokud je strom nekonečný, Y taky)

· uvažme n.v. Y na listech T t.ž. $\Pr[Y = y] = 2^{-h(y)}$ (h je hloubka)

· lemmatem výše $\mathbb{E}[h(T)] = H(Y)$

· n.v. X je vlastně fcí Y, protože chování algoritmu závisí na náhodných bitech

⇝ $H(X) \leq H(Y)$ □

· pokud psti jsou vždy mocniny dvou, máme rovnost (inspekcí důkazu)

· TODO: horní odhad

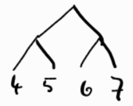
Uvažme n.v.

$$X = \begin{pmatrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 \\ 0.49 & 0.26 & 0.12 & 0.04 & 0.04 & 0.03 & 0.02 \end{pmatrix}$$

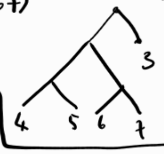
x	1	2	3	(6,7)	4	5
$p(x)$	.49	.26	.12	.05	.04	.04

1	2	3	(4,5)	(6,7)
.49	.26	.12	.08	.05

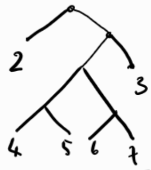
1	2	(4,5,6,7)	3
.49	.26	.13	.12

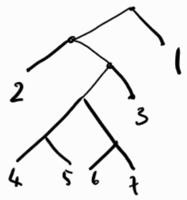


1	2	(34567)
.49	.26	.25



1	234567
.49	.51





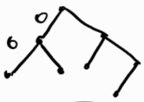
a) Nalezněte Huffmanův kód pro X.

b) Jaká je prům. délka kód. slova vs entropie? 2.02 2.01

Které z násl. kódů nemůžou být Huffmanův kód pro lib. distr?

a) $\{0, 10, 11\}$ ano, pro $(\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$

b) $\{00, 01, 10, 110\}$



ne, tohle má jednoho potomka

c) $\{01, 10\}$

ne, 1 potomek

$X_i := \{1, 0\}$

Máme n výrobků a každý z nich je funkční/vadný s šancí $p_i > \frac{1}{2}$. Pomocí ano-ne dotazů chceme zjistit které z nich jsou vadné. Lib. ano-ne otázka je povolená.

a) Najděte vhodný dolní odhad na prům. # potřebných dotazů.

b) Dejme tomu, že používáme nějakou opt. sekv. dotazů co do prům. délky. Stane se nám, že jsme ve světě, kdy používáme zrovna tu nejdelší posl. dotazů. Slovně popište, na co se ptá poslední dotaz. Mezi jakými světy rozlišuje?

c) Určete horní odhad na # vyžadovaných dotazů, použijte Huffmanův kód a prokažte ho odstavcem

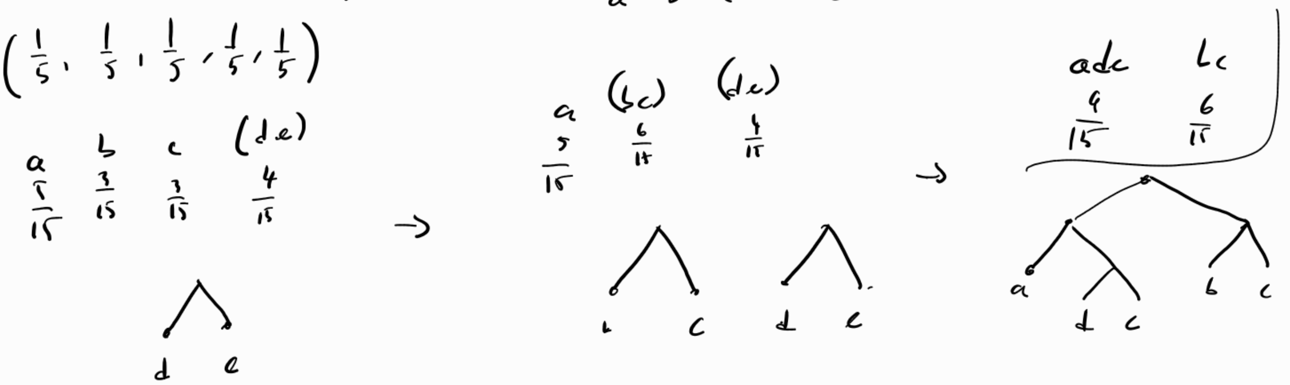
a) X_i : je i-tý item rozbitý? $H(X_1, X_2, \ldots, X_n) \overset{\text{nezávislost}}{=} \sum H(X_i) = \sum H(p_i)$

b)

- při konstr. Huffmana kódu spojujeme od nejnižších pstí nahoru
- takže posl. dotaz se ptá, jestli je posl. item rozbitý

c) $\left[\sum H(p_i)\right]+1$ (= Huffman)

Uvažme Huffmanův kód pro pst. distr. $\left(\frac{1}{3}, \frac{1}{5}, \frac{1}{5}, \frac{2}{15}, \frac{2}{15}\right)$ (a, b, c, d, e). Ukažte, že je opt. pro $\left(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}\right)$



- uvažme lib. kód C
- $\mathbb{E}[L(C)] = \sum_{x \in \{a,b,c,d,e\}} p(x) \cdot \text{délka\_kódu}(x) = \sum \frac{1}{5} \cdot \text{délka} = \frac{\sum \text{délka}}{5}$
- $\sum$ délka musí být celé čísl.
- $\mathbb{E}[\text{Huffman nahoře}] = 2 \cdot \frac{3}{5} + 3 \cdot \frac{2}{5} = \frac{12}{5}$
- lepší kód musí být v čitateli ≤ 11
- ale to je míň než $H[(1/5, 1/5, 1/5, 1/5, 1/5)]$

Source coding thm říká, že „nelze bezztrátově komprimovat X pod prům. délku slova $H(X)$ a že opt. kód má prům. délku $\leq H(X)+1$. Najděte X t.ž. opt. kód má prům. délku slova libovolně blízko k $H(X)+1$.

- stačí vzít $X = \begin{cases} 1 & \text{s pstí } 1-\varepsilon \\ 0 & \varepsilon \end{cases}$ pro $\varepsilon \to 0$
- $H(X) \to 0$, ale aspoň 1 bit je potřeba