0 Označenia

? Nech Σ je ľubovoľná množina. Hovoríme, že α je *slovo* vytvorené z *abecedy* Σ, t. j. $\alpha \in \Sigma^*$, ak existuje $k \in \mathbb{N}$ a prvky a_1, a_2, …, a_k z Σ také, že $\alpha = \langle a_1, a_2, \ldots, a_k \rangle$, obvykle skrátene $\alpha = a_1 a_2 \ldots a_k$.
Špeciálne pre $k = 0$ ide o tzv. *prázdne slovo*, označujeme ho ε.

? Definujme funkciu $\mathrm{DĺžkaSlova} : \Sigma^* \to \mathbb{N}$ takto: $\mathrm{DĺžkaSlova}(\alpha) = k$, ak $\alpha = \langle a_1, \ldots, a_k \rangle$ pre nejaké a_1, …, a_k z Σ.
Túto funkciu budeme skrátene označovať ds.

– Všimnime si, že $\Sigma^n = \Sigma \times \Sigma \times \cdots \times \Sigma$ je práve množina všetkých slov dĺžky n.
Špeciálne $\Sigma^0 = \{\varepsilon\}$.

? Ak $\alpha = a_1 a_2 \ldots a_m$ a $\beta = b_1 b_2 \ldots b_n$ sú slová z Σ^*, tak ich *konkatenáciou* (*zreťazením*) nazývame slovo $\mathrm{KonkatenáciaSlov}(\alpha, \beta) = \alpha\beta = a_1 a_2 \ldots a_m b_1 b_2 \ldots b_n$.

• Ak Σ je množina všetkých písmen slovenskej abecedy, konkatenáciou slov `myš` a `lienka` je slovo `myšlienka`.

– Zrejme je konkatenácia asociatívna, t. j. platí $(\alpha\beta)\gamma = \alpha(\beta\gamma)$.

– Zrejme tiež platí $\mathrm{ds}(\alpha\beta) = \mathrm{ds}(\alpha) + \mathrm{ds}(\beta)$.

• $\mathrm{ds}(\texttt{myš}) = 3$, $\mathrm{ds}(\texttt{lienka}) = 6$ a $\mathrm{ds}(\texttt{myšlienka}) = 9$.

? Pre slovo α definujme indukciou:

$$\alpha^0 = \varepsilon$$
$$\alpha^{n+1} = \alpha^n \alpha$$

? *Podslovom* slova $a_1 a_2 \ldots a_n$ nazývame ľubovoľné slovo $a_i a_{i+1} \ldots a_{i+(k-1)}$, kde $1 \leq i \leq n$ a $0 \leq k \leq n + 1 - i$. Špeciálne:

– Ak $k = 0$, ide o podslovo ε,
– Ak $i = 1$ a $k = n$, ide o pôvodné slovo.
– Ak $i = 1$, podslovo nazveme *prefix*. Ak navyše $k < n$, tento prefix nazývame *vlastný*.
– Ak $k = n + 1 - i$, podslovo nazveme *sufix*. Ak navyše $i > 1$, tento sufix nazývame *vlastný*.

• Podslovami slova `myšlienka` sú napríklad slová `šli`, `myš`, `lienka`, `myšlienka`, ε, ale nie `šek`. Z toho prefixmi sú iba `myš`, `myšlienka` a ε, vlastným prefixom iba `myš`. Sufixmi sú `lienka`, `myšlienka`, ε, vlastným iba `lienka`.

L Ak má slovo dva prefixy α_1 a α_2, tak nastáva jedna z týchto troch možností:

1 Ak $\mathrm{ds}(\alpha_1) = \mathrm{ds}(\alpha_2)$, tak $\alpha_1 = \alpha_2$.
2 Ak $\mathrm{ds}(\alpha_1) < \mathrm{ds}(\alpha_2)$, tak α_1 je vlastným prefixom α_2.
3 Ak $\mathrm{ds}(\alpha_1) > \mathrm{ds}(\alpha_2)$, tak α_2 je vlastným prefixom α_1.

? *Jazykom nad abecedou* Σ nazveme ľubovoľnú podmnožinu jej slov, t. j. prvkov Σ^*.
Keďže (z definície) platí $\langle x \rangle = x$, jazyk obsahujúci práve jednopísmenné slová je samotná abeceda Σ.

? Ak L_1 a L_2 sú jazyky nad abecedou Σ, tak ich *konkatenáciou* (*zreťazením*) nazývame jazyk $\mathrm{KonkatenáciaJazykov}(L_1, L_2) = L_1 L_2 = \{\alpha_1 \alpha_2 : \alpha_1 \in L_1 \wedge \alpha_2 \in L_2\}$.

– Zrejme je konkatenácia jazykov asociatívna, t. j. platí $(L_1 L_2) L_3 = L_1 (L_2 L_3)$.

? Pre jazyk L definujme indukciou:

$$L^0 = \{\varepsilon\}$$
$$L^{n+1} = L^n L$$

Keďže $\langle x \rangle = x$, špeciálne $L^1 = L$.

? Pre jazyk L definujme $L^* = \bigcup_{i \in \mathbb{N}} L^i$ a $L^+ = \bigcup_{i \in \mathbb{N}^+} L^i$.

1 Základné pojmy

– ...obrázok...:
zdroj → kóder → kanál (včítane prípadného šumu) → dekóder → príjemca
Abecedu zdroja (ale tiež príjemcu) budeme označovať Σ_S, *abecedu kanálu* (alebo *kódovú abecedu*) Σ_C. Pod *správami* rozumieme nejakú množinu $M \subseteq \Sigma_S^+$, pod *kódom* množinu *kódových slov* $C \subseteq \Sigma_C^+$.
Pod *kódovaním* potom rozumieme funkciu $e : M \to C$, pod *dekódovaním* funkciu $d : C^+ \to M^*$.
Snahou je, aby pre každé $m \in M$ platilo $d(e(m)) = m$, inak hovoríme o *šume*.

– Budeme uvažovať prípad, že $M = \Sigma_S$ a že zobrazenie e je prosté. Potom kód bude $C = e[\Sigma_S]$.

? Funkciu e potom môžeme rozšíriť na funkciu $\overline{e} : \Sigma_S^+ \to \Sigma_C^*$ vzťahom
$\overline{e}(a_1 a_2 \dots a_n) = e(a_1)e(a_2)\dots e(a_n)$.

– Od dekódujúcej funkcie d potom budeme navyše požadovať $d(\overline{e}(m)) = m$ pre všetky $m \in \Sigma_S^+$.

– Často tiež budeme uvažovať prípad, že $\Sigma_C = \{0, 1\}$, teda správy budeme kódovať pomocou dvoch znakov, hovoríme potom o *binárnom* kódovaní.

• Nech $\Sigma_S = \{0, 1, 2, 3\}$, nech kódovanie e priraďuje každému číslu jeho binárny zápis, t. j. $e = \{\langle 0, 0\rangle, \langle 1, 1\rangle, \langle 2, 10\rangle, \langle 3, 11\rangle\}$ Alternatívnym zápisom je $0 \mapsto 0$, $1 \mapsto 1$, $2 \mapsto 10$, $3 \mapsto 11$, resp. v grafickej forme *stromu*:
... obrázok...
Potom napr. $\overline{e}(121) = e(1)e(2)e(1) = 1101$.

• V predchádzajúcom príklade, žiaľ, platí tiež $\overline{e}(301) = 1101$ či $\overline{e}(1101) = 1101$. To však znamená, že pre ľubovoľné správne dekódovanie d platí napr. $301 = d(\overline{e}(301)) = d(1101) = d(\overline{e}(1101)) = 1101$, čo je však spor, a teda správne dekódovanie neexistuje.

? Kód C nazývame *rozdeliteľný* (alebo *dekódovateľný*), ak pre jeho ľubovoľné kódové slová $c_1^1, \dots, c_{n^1}^1$ a $c_1^2, \dots, c_{n^2}^2$ také, že $c_1^1 \dots c_{n^1}^1 = c_1^2 \dots c_{n^2}^2$, platí $n^1 = n^2$ a pre všetky $i \in \{1, \dots, n^1\}$ je $c_i^1 = c_i^2$.

• Kód C určený predchádzajúcim kódovaním, t. j. $C = \{0, 1, 10, 11\}$, nie je rozdeliteľný, pretože napr. slovo 1101 má aspoň dva rôzne rozklady: $n^1 = 3$, $c_1^1 = 11$, $c_2^1 = 0$, $c_3^1 = 1$ a $n^2 = 4$, $c_1^2 = 1$, $c_2^2 = 1$, $c_3^2 = 0$, $c_4^2 = 1$.

? V prípade kódovania e určujúceho rozdeliteľný kód bude potom dekódujúca funkcia d korektne definovaná spôsobom:

$$d(\alpha) = \begin{cases} e^{-1}(c_1)e^{-1}(c_2)\dots e^{-1}(c_n), & \text{ak } c_1 c_2 \dots c_n \text{ je (jednoznačný) rozklad } \alpha \text{ na kódové slová,} \\ \varepsilon & \text{inak.} \end{cases}$$

– Pre takto definovanú funkciu d potom platí: $d(c_1 c_2 \dots c_n) = d(c_1)d(c_2)\dots d(c_n)$.

? Kód C sa nazýva *blokový* (alebo *rovnomerný*), ak pre všetky $c_1, c_2 \in C$ platí $\mathrm{ds}(c_1) = \mathrm{ds}(c_2)$.

• Kódovanie:

x	$e(x)$
0	0000
1	0001
2	0010
3	0011
4	0100
5	0101
6	0110
7	0111
8	1000
9	1001

Dekódovanie je jednoduché: Kódové slovo sa rozloží na štvorice znakov, tie sa dekódujú pomocou funkcie e^{-1}.
Napríklad $d(000100101000) = d(0001)d(0010)d(1000) = e^{-1}(0001)e^{-1}(0010)e^{-1}(1000) = 128$.

2 Prefixové kódovania

? Kód C sa nazýva *prefixový*, ak neexistujú $x, y \in C$ také, že x je vlastným prefixom y.

- Vyššie uvedený kód $\{0, 1, 10, 11\}$ nie je prefixový, lebo 1 je vlastným prefixom 10.
- ...automat na dekódovanie prefixového kódu...

V Ak je kód prefixový, tak je rozdeliteľný.

Ô Nech kód C je prefixový, ale nie rozdeliteľný, teda existujú slová z Σ_C^* s aspoň dvoma rôznymi rozkladmi na kódové slová. Nech α je najkratšie z nich a nech existujú kódové slová $c_1^1, \ldots, c_{n^1}^1$ a $c_1^2, \ldots, c_{n^2}^2$ také, že $\alpha = c_1^1 \ldots c_{n^1}^1 = c_1^2 \ldots c_{n^2}^2$, ale neplatí, že $n^1 = n^2$ a zároveň pre všetky $i \in \{1, \ldots, n^1\}$ je $c_i^1 = c_i^2$. Slovo α má vlastné prefixy c_1^1 a c_1^2. Podľa lemy ... je buď jedno z nich vlastným prefixom druhého (čo však nemôže nastať, lebo C je prefixový), alebo sa navzájom rovnajú. Platí teda $c_1^1 = c_1^2$, čo však znamená, že slovo β, ktoré vznikne z α odobratím tohto vlastného prefixu, má dva rôzne rozklady na kódové slová: $\beta = c_2^1 \ldots c_{n^1}^1 = c_2^2 \ldots c_{n^2}^2$. To je však spor s predpokladanou minimalitou slova α.

- Opačné tvrdenie neplatí:

  Kód $\{0, 01, 11\}$ nie je prefixový, ale rozložiteľný je. Vznikol „otočením“ prefixového kódu $\{0, 10, 11\}$, je teda „sufixový“. V tako prípade slová rozkladáme odzadu.

  Ak chceme rozložiť slovo 0111...1, musíme ho dočítať až do konca. Ak obsahuje $2k$ jednotiek, tak jediný rozklad je $0(11)^k$, ak $2k+1$ jednotiek, tak jediný rozklad je $01(11)^k$.

– Pripomeňme, že kód je množina slov nad abecedou Σ_C, je to teda jazyk. Kódy teda možno konkatenovať. ...obrázok...

- Nech $C_1 = \{0, 10\}$ a $C_2 = \{0, 10, 11\}$, potom platí:
  - $C_1C_2 = \{00, 010, 011, 100, 1010, 1011\}$.
  - $C_2C_1 = \{00, 100, 110, 010, 1010, 1110\}$.
  - $C_1^2 = \{00, 010, 100, 1010\}$.
  - $C_2^2 = \{00, 010, 011, 100, 110, 1010, 1110, 1011, 1111\}$.

L Ak C_1 a C_2 sú prefixové kódy nad tou istou abecedou, tak platí $|C_1C_2| = |C_1| \cdot |C_2|$.

L Ak C je rozdeliteľný kód, tak platí $|C^{n+1}| = |C^n| \cdot |C|$.

L Ak C je rozdeliteľný kód, tak platí $|C^n| = |C|^n$.

? Pre kód C nad abecedou Σ_C definujme tzv. *mieru úplnosti* ako $\text{múk}(C) = \sum_{c \in C} \frac{1}{|\Sigma_C|^{\text{ds}(c)}}$.

- Nech $C_1 = \{0, 10\}$ a $C_2 = \{0, 10, 11\}$ (pri klasickom $\Sigma_C = \{0, 1\}$), potom platí:
  - $\text{múk}(C_1) = \sum_{c \in C} \frac{1}{2^{\text{ds}(c)}} = \frac{1}{2^{\text{ds}(0)}} + \frac{1}{2^{\text{ds}(10)}} = \frac{1}{2^1} + \frac{1}{2^2} = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}$.
  - $\text{múk}(C_1^2) = \sum_{c \in C} \frac{1}{2^{\text{ds}(c)}} = \frac{1}{2^{\text{ds}(00)}} + \frac{1}{2^{\text{ds}(010)}} + \frac{1}{2^{\text{ds}(100)}} + \frac{1}{2^{\text{ds}(1010)}} = \frac{1}{2^2} + \frac{1}{2^3} + \frac{1}{2^3} + \frac{1}{2^4} = \frac{1}{4} + \frac{1}{8} + \frac{1}{8} + \frac{1}{16} = \frac{9}{16}$.
  - $\text{múk}(C_2) = \frac{1}{2^1} + \frac{1}{2^2} + \frac{1}{2^2} = \frac{1}{2} + \frac{1}{4} + \frac{1}{4} = 1$.
  - $\text{múk}(C_2^2) = \frac{1}{2^2} + \frac{1}{2^3} + \frac{1}{2^3} + \frac{1}{2^3} + \frac{1}{2^3} + \frac{1}{2^4} + \frac{1}{2^4} + \frac{1}{2^4} + \frac{1}{2^4} = \frac{1}{4} + 4 \cdot \frac{1}{8} + 4 \cdot \frac{1}{16} = 1$.

L Ak C_1 a C_2 sú konečné rozdeliteľné kódy a $|C_1C_2| = |C_1| \cdot |C_2|$, tak platí $\text{múk}(C_1C_2) = \text{múk}(C_1)\text{múk}(C_2)$.

L Ak C_1 a C_2 sú konečné prefixové kódy, tak platí $\text{múk}(C_1C_2) = \text{múk}(C_1)\text{múk}(C_2)$.

L Ak C je (konečný) rozdeliteľný kód a $n \in \mathbb{N}^+$, tak $\text{múk}(C^n) = \text{múk}(C)^n$.

V (Kraftova-McMillanova nerovnosť)

Nech $d_1, \ldots, d_n$ sú nenulové prirodzené čísla a Σ_C je kódová abeceda, pre ktorú $r = |\Sigma_C| > 1$.

Potom rozdeliteľný kód C nad touto abecedou taký, že $C = \{c_1, \ldots, c_k\}$, $|C| = k$ a pre všetky $i \in \{1, \ldots, k\}$ platí $\mathrm{ds}(c_i) = d_i$, existuje práve vtedy, keď platí $\sum_{i=1}^{k} \frac{1}{r^{d_i}} \leq 1$. Navyše môžeme predpokladať, že tento kód je prefixový.

Ô→ Nech C je rozdeliteľný kód s uvedenými vlastnosťami, chceme teda dokázať, že $\mathrm{múk}(C) \leq 1$.

Nech $M = \max\{\mathrm{ds}(c_i) : i \in \{1, \ldots, k\}\}$ je maximálna dĺžka slova z C, teda pre každé $i \in \{1, \ldots, k\}$ platí $1 \leq \mathrm{ds}(c_i) \leq M$. Potom zrejme pre ľubovoľné slovo x kódu C^n platí $n \leq \mathrm{ds}(x) \leq nM$.

Označme pre $i \in \{n, n+1, \ldots, nM\}$ premennou m_i^n počet slov dĺžky i z kódu C^n. Keďže kódová abeceda má r znakov, slov dĺžky i je r^i, preto platí $m_i^n \leq r^i$.

Podľa predchádzajúcej lemy potom $\mathrm{múk}(C)^n = \mathrm{múk}(C^n) = \sum_{c \in C} \frac{1}{r^{\mathrm{ds}(c)}} = \sum_{i=n}^{nM} m_i^n \cdot \frac{1}{r^i} \leq \sum_{i=n}^{nM} r_i \cdot \frac{1}{r^i} = \sum_{i=n}^{nM} 1 \leq nM$.

Pre všetky n teda platí $\mathrm{múk}(C)^n \leq nM$, čiže $\frac{1}{n}\mathrm{múk}(C)^n \leq M$. Ale pretože $\lim_{n\to\infty} \frac{1}{n}\alpha^n = \infty$ pre $\alpha > 1$, musí byť $\mathrm{múk}(C) \leq 1$, čo sme chceli dokázať.

Ô← Konštrukcia Shannonovho kódu:

Nech $\Sigma_C = \{z_0, z_1, \ldots, z_{r-1}\}$. Bez ujmy na všeobecnosti predpokladajme, že $d_1 \leq d_2 \leq \cdots \leq d_k$,

Definujme indukciou čísla q_i pre $i \in \{1, \ldots, k\}$:

$$q_1 = 0$$
$$q_{i+1} = q_i + \frac{1}{r^{d_i}}$$

Teda $q_i = \frac{1}{r^{d_1}} + \frac{1}{r^{d_2}} + \cdots + \frac{1}{r^{d_{i-1}}}$, čiže podľa predpokladanej nerovnosti $q_i < \sum_{i=1}^{k} \frac{1}{r^{d_i}} \leq 1$. To teda znamená, že $0 = q_1 < q_2 < \cdots < q_k < 1$.

Zápis čísla q_i v r-árnej sústave má (vzhľadom na usporiadanie $d_1 \leq d_2 \leq \cdots \leq d_k$) najviac d_{i-1} cifier za čiarkou, teda ho môžeme nulami sprava doplniť tak, aby tam mal d_i cifier. Nech teda $q_i = (0, a_1^i a_2^i \ldots a_{d_i}^i)_r$, kde $a_j^i \in \{0, 1, \ldots, r-1\}$ pre všetky $j \in \{1, \ldots, d_i\}$.

Definujme teraz $c_i = z_{a_1^i} z_{a_2^i} \ldots z_{a_{d_i}^i}$, zrejme c_i má dĺžku d_i. Ukážeme, že $C = \{c_1, \ldots, c_k\}$ je prefixový, z čoho podľa vety ... vyplynie, že je rozdeliteľný.

Nech C nie je prefixový, existujú teda rôzne indexy h a i také, že c_i je vlastným prefixom c_h.

To ale znamená, že pre všetky $j \in \{1, \ldots, d_i\}$ platí $z_{a_j^i} = z_{a_j^h}$, čiže $a_j^i = a_j^h$, a teda $q_h = (0, a_1^i a_2^i \ldots a_{d_i}^i a_{d_i+1}^h \ldots a_{d_h}^h)_r = (0, a_1^i a_2^i \ldots a_{d_i}^i)_r + (0, 00 \ldots 0 a_{d_i+1}^h \ldots a_{d_h}^h)_r =$

$= q_i + (0, 00 \ldots 0 a_{d_i+1}^h \ldots a_{d_h}^h)_r$, z čoho jednak $q_h > q_i$ a jednak $q_h < q_i + \frac{1}{r^{d_i}} = q_{i+1}$.

Spolu teda $q_i < q_h < q_{i+1}$, z čoho $i < h < i+1$, čo je spor.

- (paralelne s 2. časťou dôkazu vety)

Nájdime nad klasickou kódovou abecedou $\Sigma_C = \{0, 1\}$ prefixový kód s dĺžkami slov 2, 4, 5, 6, 3, 3, 5, 6:

V prvom rade tieto čísla usporiadajme podľa veľkosti a príslušne ich označme: $d_1 = 2$, $d_2 = 3$, $d_3 = 3$, $d_4 = 4$, $d_5 = 5$, $d_6 = 5$, $d_7 = 6$, $d_8 = 6$, teda počet $k = 8$. Keďže $r = |\Sigma_C| = 2$, pre tieto čísla platí Kraftova-McMillanova nerovnosť:
$\sum_{i=1}^{k} \frac{1}{r^{d_i}} = \sum_{i=1}^{8} \frac{1}{2^{d_i}} = \frac{1}{2^2} + \frac{1}{2^3} + \frac{1}{2^3} + \frac{1}{2^4} + \frac{1}{2^5} + \frac{1}{2^5} + \frac{1}{2^6} + \frac{1}{2^6} = \frac{16+8+8+4+2+2+1+1}{64} = \frac{21}{32} \leq 1$,
takže vyhovujúci prefixový kód určite existuje.

Vypočítajme príslušné q_i a napíšme ich v 2-árnej sústave (s požadovaným počtom miest za čiarkou):

$$q_1 = 0 = (0, 00)_2$$
$$q_2 = 0 + \frac{1}{2^2} = \frac{1}{4} = (0, 010)_2$$
$$q_3 = \frac{1}{4} + \frac{1}{2^3} = \frac{3}{8} = (0, 011)_2$$

$q_4 = \frac{3}{8} + \frac{1}{2^3} = \frac{1}{2} = (0,1000)_2$
$q_5 = \frac{1}{2} + \frac{1}{2^4} = \frac{9}{16} = (0,10010)_2$
$q_6 = \frac{9}{16} + \frac{1}{2^5} = \frac{19}{32} = (0,10011)_2$
$q_7 = \frac{19}{32} + \frac{1}{2^5} = \frac{5}{8} = (0,101000)_2$
$q_8 = \frac{5}{8} + \frac{1}{2^6} = \frac{41}{64} = (0,101001)_2$

A keďže písmená abecedy Σ_C sú $z_0 = 0$ a $z_1 = 1$, hľadané kódové slová sú: $c_1 = 00$, $c_2 = 010$, $c_3 = 011$, $c_4 = 1000$, $c_5 = 10010$, $c_6 = 10011$, $c_7 = 101000$, $c_8 = 101001$.

...strom...

V Pre ľubovoľný rozdeliteľný kód $\{c_1, \ldots, c_n\}$ existuje prefixový kód $\{b_1, \ldots, b_n\}$ taký, že pre všetky $i \in \{1, \ldots, n\}$ je $\mathrm{ds}(b_i) = \mathrm{ds}(c_i)$.

? Kód C nazývame *úplný*, ak pre každé slovo $\alpha \in \Sigma_C \setminus C$ existuje $c \in C$ také, že c je vlastným prefixom α alebo α je prefixom c.

V Rozdeliteľný kód C je úplný práve vtedy, keď je prefixový a platí $\mathrm{múk}(C) = 1$.

Ô← Nech je kód $C = \{c_1, \ldots, c_k\}$ prefixový, nech $\mathrm{múk}(C) = \sum_{i=1}^{k} \frac{1}{r^{\mathrm{ds}(c_i)}} = 1$, ale nech nie je úplný, teda existuje slovo $\alpha \in \Sigma_C \setminus C$, ktoré nie je prefixom žiadneho c_i, ani žiadne c_i nie je prefixom α. To však znamená, že kód $C' = C \cup \{\alpha\}$ je tiež prefixový. Je teda rozdeliteľný, čiže $\mathrm{múk}(C') \leq 1$. Ale $\mathrm{múk}(C') = \sum_{i=1}^{k} \frac{1}{r^{\mathrm{ds}(c_i)}} + \frac{1}{r^{\mathrm{ds}(\alpha)}} = 1 + \frac{1}{r^{\mathrm{ds}(\alpha)}} > 1$, čo je spor. Kód C je teda úplný.

Ô→ Nech $C = \{c_1, \ldots, c_k\}$ je úplný a rozdeliteľný.

Označme $n = \max\{\mathrm{ds}(c_i) : i \in \{1, \ldots, k\}\}$. Z úplnosti kódu C vyplýva, že každé slovo dĺžky $n+1$ (ktoré zrejme nepatrí do C) má vlastný prefix z C (prípad, že by samo bolo vlastným prefixom niektorého slova z C, je vzhľadom na jeho priveľkú dĺžku $n+1$ nemožný).

Nech M je množina slov dĺžky $n+1$ a M_i je množina tých z nich, ktoré majú prefix c_i, platí teda $|M| \leq \sum_{i=1}^{k} |M_i|$.

Zrejme $|M| = r^{n+1}$ a $|M_i| = r^{(n+1)-\mathrm{ds}(c_i)}$ (prvých $\mathrm{ds}(c_i)$ znakov je fixných). Z toho vyplýva $r^{n+1} \leq \sum_{i=1}^{k} r^{(n+1)-\mathrm{ds}(c_i)}$, a teda $1 \leq \sum_{i=1}^{k} r^{-\mathrm{ds}(c_i)} = \sum_{i=1}^{k} \frac{1}{r^{\mathrm{ds}(c_i)}} = \mathrm{múk}(C) \leq 1$ (posledná nerovnosť vyplýva z Kraftovej-McMillanovej vety). Takže $\mathrm{múk}(C) = 1$ a vo vzťahu $|M| \leq \sum_{i=1}^{k} |M_i|$ nastáva rovnosť. To však znamená, že všetky M_i sú disjunktné.

Z toho ale vyplýva, že kód je prefixový: Ak by totiž bolo c_i vlastným prefixom c_j, platilo by $M_j \subseteq M_i$.

3 Cena kódovej postupnosti

? *Kódovou postupnosťou* nazveme prostú konečnú postupnosť slov z $\Sigma_{\mathrm{C}}*$.

? Kódovú postupnosť $\langle c_1, \ldots, c_k \rangle$ nazveme *rozdeliteľnou*, ak je kód $\{c_1, \ldots, c_k\}$ rozdeliteľný.

? Budeme uvažovať len rozdeliteľné kódové postupnosti.

? Kódovú postupnosť $\langle c_1, \ldots, c_k \rangle$ nazveme *prefixovou*, ak je kód $\{c_1, \ldots, c_k\}$ prefixový.

? *Rozdelením pravdepodobností* nazveme konečnú postupnosť $\langle p_1, \ldots, p_k \rangle$, pre ktorú $\sum_{i=1}^{k} p_i = 1$, pričom navyše pre všetky $i \in \{1, \ldots, k\}$ je $p_i > 0$.

– Ak a_1, ..., a_k sú všetky znaky Σ_{S}, pod p_i budeme rozumieť pravdepodobnosť výskytu znaku a_i.

? Nech $P = \langle p_1, \ldots, p_k \rangle$ je rozdelenie pravdepodobností a $C = \langle c_1, \ldots, c_k \rangle$ je kódová postupnosť. Potom *cenou* kódovej postupnosti C za rozdelenia P nazývame $\mathrm{cena}(C, P) = \sum_{i=1}^{k} p_i \mathrm{ds}(c_i)$.

– Ak a_1, ..., a_k sú správy a e je ich kódovanie, tak $\langle e(a_1), \ldots, e(a_k) \rangle$ je zodpovedajúca kódová postupnosť. Jej cena je teda stredná hodnota dĺžky kódového slova kódu $e[\Sigma_{\mathrm{S}}]$ (vzhľadom na dané rozdelenie pravdepodobností). Snahou bude túto cenu minimalizovať.

? *Optimálnou kódovou postupnosťou* pri rozdelení pravdepodobností P nazývame takú kódovú postupnosť C, že pre všetky kódové postupnosti B platí $\mathrm{cena}(C, P) \leq \mathrm{cena}(B, P)$. Jej cenu označíme $\mathrm{cokp}(P)$.

– Uvedomme si, že vzhľadom na konečnosť kódovej abecedy aspoň jedna optimálna kódová postupnosť vždy existuje. Podľa vety ... dokonca môžeme predpokladať, že je prefixová (keďže jej cena závisí len od dĺžky príslušných kódových slov).

? Nech $r \in \mathbb{R}$, $r > 1$ a nech $P = \langle p_1, \ldots, p_k \rangle$ je rozdelenie pravdepodobností. Potom
$\mathrm{H}_r(P) = -\sum_{i=1}^{k} p_i \log_r(p_i) = \sum_{i=1}^{k} p_i \log_r \frac{1}{p_i}$
nazývame *r-entropia*.

L Nech $r_1, r_2 \in \mathbb{R}$, $r_1, r_2 > 1$ a nech $P = \langle p_1, \ldots, p_k \rangle$ je rozdelenie pravdepodobností. Potom
$\mathrm{H}_{r_1}(P) = \frac{\ln r_2}{\ln r_1} \mathrm{H}_{r_2}(P)$.

V Nech $|\Sigma_{\mathrm{C}}| = r$ a nech $P = \langle p_1, \ldots, p_k \rangle$ je rozdelenie pravdepodobností.
Potom $\mathrm{H}_r(P) \leq \mathrm{cokp}(P) < \mathrm{H}_r(P) + 1$.
Rovnosť $\mathrm{H}_r(P) = \mathrm{cokp}(P)$ pritom nastáva práve vtedy, keď pre všetky $i \in \{1, \ldots, k\}$ platí
$p_i \in \{\frac{1}{r^j} : j \in \mathbb{N}\}$.

Ô Najprv dolný odhad:
Nech $C = \langle c_1, \ldots, c_k \rangle$ je ľubovoľná rozdeliteľná kódová postupnosť. Vzhľadom na to, že v definícii ceny ide iba o dĺžky kódových slov, bez ujmy na všeobecnosti predpokladajme, že C je prefixová.
Platí (s využitím vzťahu $\ln x \leq x - 1$ pre všetky $x > 0$):

$$
\begin{aligned}
&\mathrm{H}_r(P) - \mathrm{cena}(C, P) \\
&= -\textstyle\sum_{i=1}^{k} p_i \log_r(p_i) - \sum_{i=1}^{k} p_i \mathrm{ds}(c_i) \\
&= \textstyle\sum_{i=1}^{k} p_i (-\mathrm{ds}(c_i) - \log_r(p_i)) \\
&= \textstyle\sum_{i=1}^{k} p_i \log_r \frac{r^{-\mathrm{ds}(c_i)}}{p_i} \\
&= \textstyle\frac{1}{\ln r} \sum_{i=1}^{k} p_i \ln \frac{r^{-\mathrm{ds}(c_i)}}{p_i} \\
&\leq \textstyle\frac{1}{\ln r} \sum_{i=1}^{k} p_i (\frac{r^{-\mathrm{ds}(c_i)}}{p_i} - 1) \\
&= \textstyle\frac{1}{\ln r} \sum_{i=1}^{k} (r^{-\mathrm{ds}(c_i)} - p_i) \\
&= \textstyle\frac{1}{\ln r} (\sum_{i=1}^{k} r^{-\mathrm{ds}(c_i)} - \sum_{i=1}^{k} p_i) \\
&= \textstyle\frac{1}{\ln r} (\mathrm{múk}(\{c_1, \ldots, c_k\}) - 1) \\
&\leq 0,
\end{aligned}
$$

z čoho vyplýva požadovaná nerovnosť $\mathrm{H}_r(P) \leq \mathrm{cena}(C,P)$.

Rovnosť sa v nej nadobúda práve v prípade, keď pre všetky $i \in \{1,\ldots,k\}$ platí $\ln \frac{r^{-\mathrm{ds}(c_i)}}{p_i} = \frac{r^{-\mathrm{ds}(c_i)}}{p_i} - 1$. Pretože $\ln x = x - 1$ práve v prípade $x = 1$, pre všetky $i \in \{1,\ldots,k\}$ platí $\frac{r^{-\mathrm{ds}(c_i)}}{p_i} = 1$, teda $p_i = r^{-\mathrm{ds}(c_i)} \in \{\frac{1}{r^j} : j \in \mathbb{N}\}$.

Na horný odhad stačí ukázať existenciu kódovania, ktorého cena je menšia než $\mathrm{H}_r(P)+1$: Pre $i \in \{1,\ldots,k\}$ položme $d_i = \lceil \log_r \frac{1}{p_i} \rceil$. Keďže pre tieto čísla platí:

$\sum_{i=1}^k \frac{1}{r^{d_i}} = \sum_{i=1}^k \frac{1}{r^{\lceil \log_r \frac{1}{p_i} \rceil}} \leq \sum_{i=1}^k \frac{1}{r^{\log_r \frac{1}{p_i}}} = \sum_{i=1}^k \frac{1}{\frac{1}{p_i}} = \sum_{i=1}^k p_i = 1,$

podľa Kraftovej-McMillanovej vety existuje prefixový kód $\{c_1,\ldots,c_k\}$, pre ktorý $\mathrm{ds}(c_i) = d_i$. Cena kódovej postupnosti $C = \{c_1,\ldots,c_k\}$ je potom:

$\mathrm{cena}(C,P)$
$= \sum_{i=1}^k p_i \mathrm{ds}(c_i)$
$= \sum_{i=1}^k p_i d_i$
$= \sum_{i=1}^k p_i \lceil \log_r \frac{1}{p_i} \rceil$
$< \sum_{i=1}^k p_i (\log_r \frac{1}{p_i} + 1)$
$= \sum_{i=1}^k p_i \log_r \frac{1}{p_i} + \sum_{i=1}^k p_i$
$= \mathrm{H}_r(P) + 1.$

Z toho vyplýva, že $\mathrm{cokp}(P) \leq \mathrm{cena}(C,P) < \mathrm{H}_r(P) + 1$.

– Konštrukcia Fanovho binárneho kódovania:

Bez ujmy na všeobecnosti predpokladajme, že $p_1 \geq p_2 \geq \cdots \geq p_k > 0$. Navyše predpokladajme, že $\Sigma_{\mathrm{C}} = \{0,1\}$. Nech b_i^j označuje kódové slovo priradené a_i v j. kroku. Definujme indukciou:

1 Pre všetky $i \in \{1,\ldots,k\}$ položme $b_i^0 = \varepsilon$.

2 Pre $j \in \{0,\ldots,k-2\}$ a $i \in \{1,\ldots,k\}$ definujme množinu $M_i^j = \{t : b_i^j = b_t^j\}$.
Ak pre všetky $i \in \{1,\ldots,k\}$ platí $|M_i^j| = 1$, pre všetky $i \in \{1,\ldots,k\}$ položíme $b_i^{j+1} = b_i^j$.
Inak nech v je najmenšie také, že $|M_v^j| > 1$. Nech u je taký index z M_v^j, že $|\sum_{i \in M_v^j \wedge i \leq u} p_t - \sum_{i \in M_v^j \wedge i > u} p_t|$ je minimálne. Ak je takých u viac, vezmime najmenšie. Definujme potom pre všetky $i \in \{1,\ldots,k\}$:

$$b_i^{j+1} = \begin{cases} b_i^j 0, & \text{ak } i \in M_v^j \text{ a } i \leq u \\ b_i^j 1, & \text{ak } i \in M_v^j \text{ a } i > u \\ b_i^j, & \text{ak } i \notin M_v^j \end{cases}$$

Napokon pre všetky $i \in \{1,\ldots,k\}$ položme $c_i = b_i^{k-1}$.

- Nech $p_1 = 0,25$, $p_2 = 0,20$, $p_3 = 0,13$, $p_4 = 0,12$, $p_5 = 0,10$, $p_6 = 0,08$, $p_7 = 0,07$, $p_8 = 0,05$.
Položme $b_1^0 = b_2^0 = b_3^0 = b_4^0 = b_5^0 = b_6^0 = b_7^0 = b_8^0 = \varepsilon$. Ďalej iterujme:

1 Platí $M_1^0 = \{1,2,3,4,5,6,7,8\}$, takže hľadané v je 1.
Skúsme rôzne u:

Ak $u = 1$, tak $p^1 - (p^2 + p^3 + p^4 + p^5 + p^6 + p^7 + p^8) = 0,25 - (0,20 + 0,13 + 0,12 + 0,10 + 0,08 + 0,07 + 0,05) = -0,5$.
Ak $u = 2$, tak $(p^1 + p^2) - (p^3 + p^4 + p^5 + p^6 + p^7 + p^8) = (0,25 + 0,20) - (0,13 + 0,12 + 0,10 + 0,08 + 0,07 + 0,05) = -0,1$.
Ak $u = 3$, tak $(p^1 + p^2 + p^3) - (p^4 + p^5 + p^6 + p^7 + p^8) = (0,25 + 0,20 + 0,13) - (0,12 + 0,10 + 0,08 + 0,07 + 0,05) = +0,16$.

Ďalšie u už nemá význam skúmať, lebo výsledky sú rastúce. Najmenšie (lebo jediné) vyhovujúce u (t. j. to, že uvedený výsledok je (v absolútnej hodnote) najbližší 0) je teda 2.
Platí teda: $b_1^1 = b_2^1 = 0$, $b_3^1 = b_4^1 = b_5^1 = b_6^1 = b_7^1 = b_8^1 = 1$.

2 Platí $M_1^1 = \{1, 2\}$, takže hľadané v je opäť 1.
Vzhľadom na dvojprvkovosť množiny zrejme $u = 1$.
Platí teda: $b_1^2 = 00$, $b_2^2 = 01$, $b_3^2 = b_4^2 = b_5^2 = b_6^2 = b_7^2 = b_8^2 = 1$.

3 Platí $M_1^2 = \{1\}$, $M_2^2 = \{2\}$,$M_3^2 = \{3, 4, 5, 6, 7, 8\}$, takže hľadané v je 3.
Skúsme rôzne u z $M_3^2 = \{3, 4, 5, 6, 7, 8\}$:

Ak $u = 3$, tak $p^3 - (p^4 + p^5 + p^6 + p^7 + p^8) = 0,13 - (0,12 + 0,10 + 0,08 + 0,07 + 0,05) = -0,29$.
Ak $u = 4$, tak $(p^3 + p^4) - (p^5 + p^6 + p^7 + p^8) = (0,13 + 0,12) - (0,10 + 0,08 + 0,07 + 0,05) = -0,05$.
Ak $u = 5$, tak $(p^3 + p^4 + p^5) - (p^6 + p^7 + p^8) = (0,13 + 0,12 + 0,10) - (0,08 + 0,07 + 0,05) = 0,15$.

Ďalšie u už nemá význam skúmať, lebo výsledky sú rastúce. Najmenšie (lebo jediné) vyhovujúce u je teda 4.
Platí teda: $b_1^3 = 00$, $b_2^3 = 01$, $b_3^3 = b_4^3 = 10$, $b_5^3 = b_6^3 = b_7^3 = b_8^3 = 11$.

4 Platí $M_1^3 = \{1\}$, $M_2^3 = \{2\}$,$M_3^3 = \{3, 4\}$, takže hľadané v je 3.
Vzhľadom na dvojprvkovosť množiny zrejme $u = 3$.
Platí teda: $b_1^4 = 00$, $b_2^4 = 01$, $b_3^4 = 100$, $b_4^4 = 101$, $b_5^4 = b_6^4 = b_7^4 = b_8^4 = 11$.

5 Platí $M_1^4 = \{1\}$, $M_2^4 = \{2\}$, $M_3^4 = \{3\}$, $M_4^4 = \{4\}$, $M_5^4 = \{5, 6, 7, 8\}$, takže hľadané v je 5.
Skúsme rôzne u z $M_5^4 = \{5, 6, 7, 8\}$:

Ak $u = 5$, tak $p^5 - (p^6 + p^7 + p^8) = 0,10 - (0,08 + 0,07 + 0,05) = -0,1$.
Ak $u = 6$, tak $(p^5 + p^6) - (p^7 + p^8) = (0,10 + 0,08) - (0,07 + 0,05) = 0,06$.

Ďalšie u už nemá význam skúmať, lebo výsledky sú rastúce. Najmenšie (lebo jediné) vyhovujúce u je teda 6.
Platí teda: $b_1^5 = 00$, $b_2^5 = 01$, $b_3^5 = 100$, $b_4^5 = 101$, $b_5^5 = b_6^5 = 110$, $b_7^5 = b_8^5 = 111$.

6 Platí $M_1^5 = \{1\}$, $M_2^5 = \{2\}$, $M_3^5 = \{3\}$, $M_4^5 = \{4\}$, $M_5^5 = \{5, 6\}$, takže hľadané v je 5.
Vzhľadom na dvojprvkovosť množiny zrejme $u = 5$.
Platí teda: $b_1^6 = 00$, $b_2^6 = 01$, $b_3^6 = 100$, $b_4^6 = 101$, $b_5^6 = 1100$, $b_6^6 = 1101$, $b_7^6 = b_8^6 = 111$.

6 Platí $M_1^6 = \{1\}$, $M_2^6 = \{2\}$, $M_3^6 = \{3\}$, $M_4^6 = \{4\}$, $M_5^6 = \{5\}$, $M_6^6 = \{6\}$, $M_7^6 = \{7, 8\}$, takže hľadané v je 7.
Vzhľadom na dvojprvkovosť množiny zrejme $u = 7$.
Platí teda: $b_1^7 = 00$, $b_2^7 = 01$, $b_3^7 = 100$, $b_4^7 = 101$, $b_5^7 = 1100$, $b_6^7 = 1101$, $b_7^7 = 1110$, $b_8^7 = 1111$.

Dostávame teda $c_1 = 00$, $c_2 = 01$, $c_3 = 100$, $c_4 = 101$, $c_5 = 1100$, $c_6 = 1101$, $c_7 = 1110$, $c_8 = 1111$. Cena kódovej postupnosti $C = \langle c_1, \ldots, c_k \rangle$ je potom $\mathrm{cena}(C, P) = \sum_{i=1}^{k} p_i \mathrm{ds}(c_i) \doteq 2,85$. Pritom 2-entropia je $\mathrm{H}_2(P) = -\sum_{i=1}^{k} p_i \log_2(p_i) \doteq 2,822$.

4 Konštrukcia optimálnej kódovej postupnosti

? Nech $C = \langle c_1, \ldots, c_k \rangle$ je prefixová kódová postupnosť. Označme $\alpha_{i,j}$ najväčší spoločný (vlastný) prefix slov c_i a c_j a nech $m_i = \max\{\mathrm{ds}(\alpha_{i,j}) : j \in \{1, \ldots, k\} \setminus \{i\}\}$.

Nech b_i je prefix c_i dĺžky m_i+1. Potom kódovú postupnosť $\langle b_1, \ldots, b_k \rangle$ nazveme *minimalizácia* kódovej postupnosti C.

L Minimalizácia prefixovej kódovej postupnosti je tiež prefixová kódová postupnosť.

? Kódovú postupnosť, ktorá sa zhoduje so svojou minimalizáciou, nazveme *minimalizovaná*.

L Nech $P = \langle p_1, \ldots, p_k \rangle$ je ľubovoľné rozdelenie pravdepodobností také, že kde $p_1 \geq \cdots \geq p_k > 0$, nech $C = \langle c_1, \ldots, c_k \rangle$ je kódová postupnosť a $B = \langle b_1, \ldots, b_k \rangle$ je jej minimalizácia. Potom $\mathrm{cena}(B, P) \leq \mathrm{cena}(C, P)$.

L Každá optimálna kódová postupnosť k danému rozdeleniu pravdepodobnosti je minimalizovaná.

? Nech $C = \langle c_1, \ldots, c_k \rangle$ je kódová postupnosť a nech j_1 a j_2 sú indexy z $\{1, \ldots, k\}$. Potom kódovú postupnosť $\mathrm{výmena}(C, j_1, j_2) = \langle b_1, \ldots, b_k \rangle$ definujeme takto:

$$b_i = \begin{cases} c_{j_2}, & \text{ak } i = j_1, \\ c_{j_1}, & \text{ak } i = j_2, \\ c_i, & \text{inak.} \end{cases}$$

– Uvedomme si, že $\{b_1, \ldots, b_k\} = \{c_1, \ldots, c_k\}$, takže ak bola pôvodná kódová postupnosť $\langle c_1, \ldots, c_k \rangle$ prefixová a/alebo minimalizovaná, aj $\langle b_1, \ldots, b_k \rangle$ bude taká.

L Nech $P = \langle p_1, \ldots, p_k \rangle$ je ľubovoľné rozdelenie pravdepodobností také, že kde $p_1 \geq \cdots \geq p_k > 0$, nech $C = \langle c_1, \ldots, c_k \rangle$ je kódová postupnosť.

Nech j_1 a j_2 sú indexy z $\{1, \ldots, k\}$ také, že $j_1 \geq j_2$ (z čoho $p_{j_1} \geq p_{j_2}$), ale $\mathrm{ds}(c_{j_1}) \geq \mathrm{ds}(c_{j_2})$. Potom platí $\mathrm{cena}(\mathrm{výmena}(C, j_1, j_2), P) \leq \mathrm{cena}(C, P)$.

Ô Označme $B = \mathrm{výmena}(C, j_1, j_2) = \langle b_1, \ldots, b_k \rangle$. Potom:

$$\begin{aligned}
&\mathrm{cena}(C, P) \\
&= \textstyle\sum_{i=1}^{k} p_i \mathrm{ds}(c_i) \\
&= \textstyle\sum_{i=1}^{k} p_i \mathrm{ds}(b_i) + p_{j_1}(\mathrm{ds}(c_{j_1}) - \mathrm{ds}(b_{j_1})) + p_{j_2}(\mathrm{ds}(c_{j_2}) - \mathrm{ds}(b_{j_2})) \\
&= \mathrm{cena}(B, P) + p_{j_1}(\mathrm{ds}(c_{j_1}) - \mathrm{ds}(c_{j_2})) + p_{j_2}(\mathrm{ds}(c_{j_2}) - \mathrm{ds}(c_{j_1})) \\
&= \mathrm{cena}(B, P) + (p_{j_1} - p_{j_2})(\mathrm{ds}(c_{j_1}) - \mathrm{ds}(c_{j_2})) \\
&\geq \mathrm{cena}(B, P).
\end{aligned}$$

? Nech $C = \langle c_1, \ldots, c_k \rangle$ je kódová postupnosť. Definujme indukciou pre $i \in \{0, \ldots, k\}$ kódové postupnosti $B^i = \langle b_1^i, \ldots, b_k^i \rangle$ takto:

1 $B^0 = C$.

2 Pre každé $t \in \{1, \ldots, k\}$ definujme množinu „porúch usporiadania“ $\mathrm{P}_t^i = \{j > t : \mathrm{ds}(b_t^i) > \mathrm{ds}(b_j^i)\}$. Ak sú všetky tieto množiny prázdne, položíme $B^{i+1} = B^i$, inak nech
$j_1^i = \min\{t \in \{1, \ldots, k\} : \mathrm{P}_t^i \neq \emptyset\}$ (t. j. index najskoršej poruchy),
$m^i = \min\{\mathrm{ds}(b_j^i) : j \in \mathrm{P}_{j_1^i}^i\}$ (dĺžka najkratšieho slova s indexom väčším než j_1^i, ktoré je kratšie než $b_{j_1^i}^i$)
a $j_2^i = \min\{j \in \mathrm{P}_{j_1^i}^i : \mathrm{ds}(b_j^i) = m^i\}$ (jeho index (ak je ich viac, tak najmenší)).
Potom definujme $B^{i+1} = \mathrm{výmena}(B^i, j_1^i, j_2^i)$.

Potom kódovú postupnosť B^k nazveme *dĺžkovo usporiadaná verzia* kódovej postupnosti B a označíme ju $\mathrm{usporiadanie}(B)$.

L Pre každé $i \in \{1, \ldots, k\}$ pre $B^i = \langle b_1^i, \ldots, b_k^i \rangle$ z predchádzajúcej definície platí $\mathrm{ds}(b_1^i) \leq \cdots \leq \mathrm{ds}(b_i^i)$.

? Kódovú postupnosť nazveme *(dĺžkovo) usporiadaná*, ak sa zhoduje so svojou dĺžkovo usporiadanou verziou.

L Nech $P = \langle p_1, \ldots, p_k \rangle$ je ľubovoľné rozdelenie pravdepodobností také, že kde $p_1 \geq \cdots \geq p_k > 0$, nech $C = \langle c_1, \ldots, c_k \rangle$ je kódová postupnosť. Potom platí $\mathrm{cena}(\mathrm{usporiadanie}(C), P) \leq \mathrm{cena}(C, P)$.

– Uvedomme si, že ak je C prefixová a/alebo minimalizovaná a/alebo optimálna, aj $\mathrm{usporiadanie}(C)$ je taká.

L Nech $\Sigma_{\mathrm{C}} = \{z_0, z_1\}$. Pre ľubovoľné rozdelenie pravdepodobností $P = \langle p_1, \ldots, p_k \rangle$, kde $p_1 \geq \cdots \geq p_k > 0$, existuje optimálna minimalizovaná prefixová usporiadaná kódová postupnosť, ktorej posledné dva členy majú maximálnu a rovnakú dĺžku a líšia sa len v poslednom znaku.

Ô Nech B je optimálna kódová postupnosť, podľa vety ... môžeme bez ujmy na všeobecnosti predpokladať, že je prefixová, minimalizovaná a dĺžkovo usporiadaná.

Nech $b_k = \alpha z$ pre nejaké $\alpha \in \Sigma_{\mathrm{C}}^*$ a $z \in \Sigma_{\mathrm{C}}$. Keďže B je minimalizovaná, musí pre niektoré $i \in \{1, \ldots, k-1\}$ platiť $b_i = \alpha\beta$ pre nejaké $\beta \in \Sigma_{\mathrm{C}}^*$. Vzhľadom na prefixovosť B musí byť $\mathrm{ds}(\beta) \geq 1$, vzhľadom na usporiadanosť B musí byť $\mathrm{ds}(\alpha) + 1 = \mathrm{ds}(\alpha z) = \mathrm{ds}(b_k) \geq \mathrm{ds}(b_i) = \mathrm{ds}(\alpha\beta) = \mathrm{ds}(\alpha) + \mathrm{ds}(\beta)$ z čoho $\mathrm{ds}(\beta) \leq 1$. Takže $\mathrm{ds}(\beta) = 1$, t. j. $\beta = w$ pre nejaké $w \neq z$. Takže b_i a b_k majú rovnakú (a maximálnu) dĺžku a líšia sa len v poslednom znaku.

Vzhľadom na dĺžkovú usporiadanosť B platí $\mathrm{ds}(b_i) = \mathrm{ds}(b_{i+1}) = \cdots = \mathrm{ds}(b_{k-1}) = \mathrm{ds}(b_k)$, teda aj kódová postupnosť $C = \mathrm{výmena}(B, i, k-1)$ je dĺžkovo usporiadaná. Naviac je prefixová, minimalizovaná a jej posledné dva členy sa líšia len v poslednom písmene. Je to teda hľadaná kódová postupnosť.

? Nech z_0 a z_1 sú rôzne a jediné prvky Σ_{C} a nech $j \in \{1, \ldots, k\}$. Nech $C = \langle c_1, \ldots, c_k \rangle$ je kódová postupnosť. Potom označme $\mathrm{expanzia}(C, j)$ kódovú postupnosť $\langle c_1, \ldots, c_{j-1}, c_{j+1}, \ldots, c_k, c_j z_0, c_j z_1 \rangle$.

L Ak $C = \langle c_1, \ldots, c_k \rangle$ je prefixová, tak aj $\mathrm{expanzia}(C, j)$ je prefixová.

L Nech $\Sigma_{\mathrm{C}} = \{z_0, z_1\}$ a $j \in \{1, \ldots, k\}$.

Nech $P = \langle p_1, \ldots, p_k \rangle$ a $P' = \langle p_1, \ldots, p_{j-1}, p_{j+1}, \ldots, p_k, q_0, q_1 \rangle$, kde $p_1 \geq \cdots \geq p_k > 0$, $p_j = q_0 + q_1$ a $p_k \geq q_0 \geq q_1$.

Nech $C = \langle c_1, \ldots, c_k \rangle$ je kódová postupnosť, Potom $\mathrm{cena}(\mathrm{expanzia}(C, j), P') = \mathrm{cena}(C, P) + p_j$.

Ô Označme $C' = \mathrm{expanzia}(C, j)$. Potom platí:

$\mathrm{cena}(C', P')$
$= \sum_{i \in \{1,\ldots,j-1\} \cup \{j+1,\ldots,k\}} p_i \mathrm{ds}(c_i) + q_0 \mathrm{ds}(c_j z_0) + q_1 \mathrm{ds}(c_j z_1)$
$= \sum_{i \in \{1,\ldots,j-1\} \cup \{j+1,\ldots,k\}} p_i \mathrm{ds}(c_i) + q_0(\mathrm{ds}(c_j) + 1) + q_1(\mathrm{ds}(c_j) + 1)$
$= \sum_{i \in \{1,\ldots,j-1\} \cup \{j+1,\ldots,k\}} p_i \mathrm{ds}(c_i) + (q_0 + q_1)\mathrm{ds}(c_j) + (q_0 + q_1)$
$= \sum_{i \in \{1,\ldots,j-1\} \cup \{j+1,\ldots,k\}} p_i \mathrm{ds}(c_i) + p_j \mathrm{ds}(c_j) + p_j$
$= \sum_{i=1}^{k} p_i \mathrm{ds}(c_i) + p_j.$
$= \mathrm{cena}(C, P) + p_j.$

L Nech ...

Potom ak $\mathrm{cena}(C_1, P) = \mathrm{cena}(C_2, P)$, tak $\mathrm{cena}(\mathrm{expanzia}(C_1, j), P') = \mathrm{cena}(\mathrm{expanzia}(C_2, j), P')$.

? Nech $\Sigma_{\mathrm{C}} = \{z_0, z_1\}$ a $j \in \{1, \ldots, k\}$. Nech $C' = \langle c_1, \ldots, c_{j-1}, c_{j+1}, \ldots, c_k, c_j z_0, c_j z_1 \rangle$. Potom označme $\mathrm{redukcia}(C', j)$ kódovú postupnosť $\langle c_1, \ldots, c_{j-1}, c_j, c_{j+1}, \ldots, c_k \rangle$.

L Nech ... Potom $\mathrm{redukcia}(\mathrm{expanzia}(C, j), j) = C$.

L Nech ... Potom $\mathrm{expanzia}(\mathrm{redukcia}(C', j), j) = C'$.

L Nech ... Ak C' je prefixová, tak aj $\mathrm{redukcia}(C', j)$ je tiež prefixová.

L Nech ... Potom ak $\mathrm{cena}(C_1', P') = \mathrm{cena}(C_2', P')$, tak
$\mathrm{cena}(\mathrm{redukcia}(C_1', j), P) = \mathrm{cena}(\mathrm{redukcia}(C_2', j), P)$.

V Nech $\Sigma_{\mathrm{C}} = \{z_0, z_1\}$ a $j \in \{1, \ldots, k\}$.
Nech $P = \langle p_1, \ldots, p_k \rangle$ a $P' = \langle p_1, \ldots, p_{j-1}, p_{j+1}, \ldots, p_k, q_0, q_1 \rangle$, kde $p_1 \geq \cdots \geq p_k > 0$, $p_j = q_0 + q_1$ a $p_k \geq q_0 \geq q_1$.
Potom C je optimálna prefixová kódová postupnosť pre P práve vtedy, keď $\mathrm{expanzia}(C, j)$ je optimálna prefixová kódová postupnosť pre P'.

→ Podľa lemy ... existuje pre P' optimálna prefixová kódová postupnosť
$B' = \langle b_1, \ldots, b_{j-1}, b_{j+1}, \ldots, b_k, b_j z_0, b_j z_1 \rangle$.
Z optimality kódovej postupnosti C máme $\mathrm{cena}(C, P) \leq \mathrm{cena}(\mathrm{redukcia}(B', j), P)$.
Z toho podľa lemy ... dostávame
$\mathrm{cena}(\mathrm{expanzia}(C, j), P') \leq \mathrm{cena}(\mathrm{expanzia}(\mathrm{redukcia}(B', j), j), P') = \mathrm{cena}(B', P')$,
takže aj kódová postupnosť $\mathrm{expanzia}(C, j)$ musí byť optimálna.

← Nech B je ľubovoľná kódová postupnosť pre P, bez ujmy na všeobecnosti prefixová.
Potom $\mathrm{expanzia}(B, j)$ je prefixová kódová postupnosť pre P', takže z optimality kódovej postupnosti $\mathrm{expanzia}(C, j)$ dostávame $\mathrm{cena}(\mathrm{expanzia}(C, j), P') \leq \mathrm{cena}(\mathrm{expanzia}(B, j), P')$.
Z toho podľa lemy ... dostávame
$\mathrm{cena}(C, P) = \mathrm{cena}(\mathrm{redukcia}(\mathrm{expanzia}(C, j), j), P) \leq \mathrm{cena}(\mathrm{redukcia}(\mathrm{expanzia}(B, j), j), P) = \mathrm{cena}(B, P)$,
takže kódová postupnosť C je optimálna.

L Ak $P = \langle p_1, \ldots, p_r \rangle$ a $\Sigma_{\mathrm{C}} = \{z_0, \ldots, z_{r-1}\}$, tak $\langle z_0, \ldots, z_{r-1} \rangle$ je optimálna kódová postupnosť (a jej cena je 1).

– ... konštrukcia Huffmannovho optimálneho binárneho kódovania...

– ... konštrukcia Huffmannovho r-árneho kódovania...

– Ak $\Sigma_{\mathrm{S}} = \{\mathtt{a}, \mathtt{b}\}$, $P = \langle 0.9, 0.1 \rangle$ a $\Sigma_{\mathrm{C}} = \{\mathtt{0}, \mathtt{1}\}$, tak cena optimálnej kódovej postupnosti (napr.) $\langle \mathtt{0}, \mathtt{1} \rangle$ je 1, 2-entropia je však $\mathrm{H}_2(P) = 0.9 \log_2 \frac{1}{0.9} + 0.1 \log_2 \frac{1}{0.1} \doteq 0.469$. Odchýlka (redundancia) je teda pomerne veľká.
Ako sa entropii priblížiť ešte viac? Kódovať budeme celé n-tice naraz, teda budeme akoby pracovať s $\Sigma'_{\mathrm{S}} = \{\mathtt{aa}, \mathtt{ba}, \mathtt{ab}, \mathtt{bb}\} = \Sigma_{\mathrm{S}}^2$, teda s $P' = \{0.9 \cdot 0.9, 0.9 \cdot 0.1, 0.1 \cdot 0.9, 0.1 \cdot 0.1\}$. Optimálna kódová postupnosť je potom $C' = \langle \mathtt{0}, \mathtt{10}, \mathtt{110}, \mathtt{111} \rangle$, s cenou $\mathrm{cena}(C') = 0.81 \cdot 1 + 0.09 \cdot 2 + 0.09 \cdot 3 + 0.01 \cdot 3 = 1.29$. Fakticky sa však kódované texty z Σ_{S}^2 dvakrát skrátili, teda „cena“ tohto kódovania je $\frac{1.29}{2} = 0.645$, čo je podstatné zlepšenie.
A ak budeme pracovať s Σ_{S}^n pre čoraz väčšie n, „cena“ sa bude zlepšovať ešte viac. Daňou za to je však nárast počtu kódovaných znakov (pôvodných n-tíc znakov) na $|\Sigma_{\mathrm{S}}|^n$.

5 Aritmetické kódovanie

– Myšlienku kódovať nielen samostatné znaky zo Σ_S, ale celé ich postupnosti používame aj v tzv. *aritmetickom kódovaní*:

Každému slovu $w = a_1 a_2 \dots a_n$, t. j. postupnosti indexov $i(w) = \langle i_1, \dots, i_n \rangle$ priradíme istý (polouzavretý) interval $\mathrm{A}(i(w))$ s dĺžkou úmernou pravdepodobnosti výskytu tohto slova. Potom z tohto intervalu vyberieme jedného reprezentanta $\mathrm{rep}(\mathrm{A}(i(w)))$, ktorý je *dyadickým číslom*, t. j. racionálnym číslom tvaru $\frac{x}{2^t}$. Cifry binárneho zápisu tohto čísla budú tvoriť zodpovedajúce slovo $e(w)$.

? Nech $P = \langle p_1, \dots, p_k \rangle$ je rozdelenie pravdepodobností, nech $I = [u, v)$ je polouzavretý interval reálnych čísel, pričom $u < v$, a nech $j \in \{1, \dots, k\}$. Definujme *j-podinterval intervalu I podľa rozdelenia pravdepodobností P* takto:

$$\mathrm{podinterval}(I, j, P) = \left[u + (v-u) \sum_{i=1}^{j-1} p_i, \quad u + (v-u) \sum_{i=1}^{j} p_i \right).$$

L $\mathrm{DĺžkaIntervalu}(\mathrm{podinterval}(I, j, P)) = p_j \cdot \mathrm{DĺžkaIntervalu}(I)$.

L I je disjunktné zjednotenie intervalov $\mathrm{podinterval}(I, 1, P)$, …, $\mathrm{podinterval}(I, k, P)$.

? Nech $P = \langle p_1, \dots, p_k \rangle$ je rozdelenie pravdepodobností.
Definujme potom funkciu $\mathrm{A}_P : \{1, \dots, k\}^* \to \mathcal{P}([0, 1))$ takto:

1 $\mathrm{A}_P(\varepsilon) = [0, 1)$.

2 Ak $\alpha \in \{1, \dots, k\}^*$ a $j \in \{1, \dots, k\}$, tak $\mathrm{A}_P(\alpha j) = \mathrm{podinterval}(\mathrm{A}_P(\alpha), j, P)$.

L $\mathrm{DĺžkaIntervalu}(\mathrm{A}_P(\langle i_1, i_2, \dots, i_n \rangle)) = p_{i_1} p_{i_2} \dots p_{i_n}$.

• Ak $P = \langle 0.4, 0.3, 0.2, 0.1 \rangle$, tak (...obrázok...):
$\mathrm{A}_P(\langle 1 \rangle) = [0, 0.4)$, $\mathrm{A}_P(\langle 2 \rangle) = [0.4, 0.7)$, $\mathrm{A}_P(\langle 3 \rangle) = [0.7, 0.9)$, $\mathrm{A}_P(\langle 4 \rangle) = [0.9, 1)$,
$\mathrm{A}_P(\langle 2, 1 \rangle) = [0.4, 0.52)$, $\mathrm{A}_P(\langle 2, 2 \rangle) = [0.52, 0.61)$, $\mathrm{A}_P(\langle 2, 3 \rangle) = [0.61, 0.67)$,
$\mathrm{A}_P(\langle 2, 4 \rangle) = [0.67, 0.7)$,
$\mathrm{A}_P(\langle 2, 1, 1 \rangle) = [0.4, 0.448)$, $\mathrm{A}_P(\langle 2, 1, 2 \rangle) = [0.448, 0.484)$, $\mathrm{A}_P(\langle 2, 1, 3 \rangle) = [0.484, 0.508)$,
$\mathrm{A}_P(\langle 2, 1, 4 \rangle) = [0.508, 0.52)$,
$\mathrm{A}_P(\langle 2, 1, 3, 2 \rangle) = [0.4936, 0.5008)$.

? Nech $I = [u, v)$, pričom $u, v \in [0, 1]$ a $u < v$. Nech t je (jediné) riešenie nerovnice $\frac{1}{2^t} \le v - u < \frac{1}{2^{t-1}}$. Nech $M = \{x \in \mathbb{N} : u \le \frac{x}{2^t} < v\}$, zrejme $1 \le |M| \le 2$.
Ak $|M| = 1$, definujme $\mathrm{rep}(I) = \frac{x}{2^t}$.
Ak $|M| = 2$ a $x_1, x_2 \in M$, zrejme $|x_1 - x_2| = 1$, t. j. majú rôznu paritu. Nech x je párne z nich, potom definujme $\mathrm{rep}(I) = \frac{x}{2^t}$.

L $\mathrm{rep}(I) \in I$.

• Ak $I = [0.484, 0.508)$, tak riešenie nerovnosti $\frac{1}{2^t} \le 0.508 - 0.484 < \frac{1}{2^{t-1}}$ je $t = 3$. Potom riešime nerovnicu $0.484 \le \frac{x}{2^3} < 0.508$, teda $3.872 \le x < 4.064$. Z toho $x = 4$, teda $\mathrm{rep}(I) = \frac{4}{8} = \frac{1}{2}$.

• Ak $I = [0.4936, 0.5008)$, tak riešenie nerovnosti $\frac{1}{2^t} \le 0.5008 - 0.4936 < \frac{1}{2^{t-1}}$ je $t = 8$. Potom riešime nerovnicu $0.4936 \le \frac{x}{2^8} < 0.5008$, teda $126.3616 \le x < 128.2048$. Z toho $M = \{127, 128\}$, teda $x = 128$, a potom $\mathrm{rep}(I) = \frac{128}{2^8} = \frac{1}{2}$.

L (o ekvivalentnom získaní reprezentanta)
Nech $I = [u, v)$, pričom $u, v \in [0, 1]$ a $u < v$. Nech $u = (0.u_1 u_2 \dots)_2$ a $v = (0.v_1 v_2 \dots)_2$, pričom nekonečne veľa u_i je 0, a nech $t \in \mathbb{N}$ je také, že pre všetky $i \in \mathbb{N}^+$, $i \le t$ platí $u_i = v_i = c_i$ a $u_{t+1} = 0$, $v_{t+1} = 1$ (lebo veď $u < v$). Rozlíšme možnosti:

1 Ak $u = (0.c_1 c_2 \dots c_t)_2$, t. j. pre všetky $i > t$ je $u_i = 0$, tak $\mathrm{rep}(I) = u$.

2 Nech $u > (0.c_1c_2\dots c_t)_2$, t. j. pre niektoré $i > t$ je $u_i = 1$, ale nech $v = (0.c_1c_2\dots c_t1)_2$, t. j. pre všetky $i \geq t+2$ je $u_i = 0$. Nech $j > t+1$ je najmenšie také, že $u_j = 0$, potom $\mathrm{rep}(I) = (0.u_1u_2\dots u_{j-1}1)_2$.

3 V ostatných prípadoch $\mathrm{rep}(I) = (0.c_1c_2\dots c_t1)_2$.

? Definujme funkciu cwd, ktorá každému dyadickému číslu z $[0,1)$ priradí konečnú postupnosť z $\{0,1\}^*$ takto:

1 $\mathrm{cwd}(0) = 0$.

2 Nech $r = (0.c_1c_2\dots c_m)_2$ také, že $c_m = 1$, potom $\mathrm{cwd}(r) = c_1c_2\dots c_m$.

? Nech $P = \langle p_1,\dots,p_k\rangle$ je rozdelenie pravdepodobností a $n \in \mathbb{N}^+$, potom *cena aritmetického kódovania* slov dĺžky n je

$$\mathrm{cak}(P,n) = \sum_{\langle i_1,\dots,i_n\rangle\in\{1,\dots,k\}^n} |\mathrm{A}_P(\langle i_1,\dots,i_n\rangle)| \cdot \mathrm{ds}(\mathrm{cwd}(\mathrm{rep}(\mathrm{A}_P(\langle i_1,\dots,i_n\rangle)))),$$

t. j.

$$\mathrm{cak}(P,n) = \sum_{\langle i_1,\dots,i_n\rangle\in\{1,\dots,k\}^n} p_{i_1}\dots p_{i_n} \cdot \mathrm{ds}(\mathrm{cwd}(\mathrm{rep}(\mathrm{A}_P(\langle i_1,\dots,i_n\rangle)))).$$

L Nech $P = \langle p_1,\dots,p_k\rangle$ je rozdelenie pravdepodobností. Potom

$$\sum_{\langle i_1,\dots,i_n\rangle\in\{1,\dots,k\}^n} p_{i_1}\dots p_{i_n} = 1.$$

L Nech $P = \langle p_1,\dots,p_k\rangle$ je rozdelenie pravdepodobností, $t \in \{1,\dots,k\}$ a $j \in \{1,\dots,n\}$. Potom

$$\sum_{\langle i_1,\dots,i_n\rangle\in\{1,\dots,k\}^n, i_j=t} p_{i_1}\dots p_{i_n} = p_t.$$

Ô Pomocou distributívneho zákona dostávame:
$\sum_{\langle i_1,\dots,i_n\rangle\in\{1,\dots,k\}^n, i_j=t} p_{i_1}\dots p_{i_n}$,
$= (\sum_{i\in\{1,\dots,k\}} p_i)\cdots(\sum_{i\in\{1,\dots,k\}} p_i)\cdot p_t\cdot(\sum_{i\in\{1,\dots,k\}} p_i)\cdots(\sum_{i\in\{1,\dots,k\}} p_i)$,
$= 1\cdots 1\cdot p_t\cdot 1\cdots 1$
$= p_t$.

L Nech $P = \langle p_1,\dots,p_k\rangle$ je rozdelenie pravdepodobností, $f : \{1,\dots,k\} \to \mathbb{R}$ a $j \in \{1,\dots,k\}$. Potom

$$\sum_{\langle i_1,\dots,i_n\rangle\in\{1,\dots,k\}^n} p_{i_1}\dots p_{i_n} f(i_j) = \sum_{t=1}^{k} p_t f(t).$$

Ô $\sum_{\langle i_1,\dots,i_n\rangle\in\{1,\dots,k\}^n} p_{i_1}\dots p_{i_n} f(i_j)$,
$= \sum_{t=1}^{k}\sum_{\langle i_1,\dots,i_n\rangle\in\{1,\dots,k\}^n, i_j=t} p_{i_1}\dots p_{i_n} f(t)$,
$= \sum_{t=1}^{k} f(t)\cdot\left(\sum_{\langle i_1,\dots,i_n\rangle\in\{1,\dots,k\}^n, i_j=t} p_{i_1}\dots p_{i_n}\right)$,
$= \sum_{t=1}^{k} f(t)p_t$.

L Nech $P = \langle p_1,\dots,p_k\rangle$ je rozdelenie pravdepodobností a $j \in \{1,\dots,k\}$. Potom

$$-\sum_{\langle i_1,\dots,i_n\rangle\in\{1,\dots,k\}^n} p_{i_1}\dots p_{i_n}\log_2 p_{i_j} = \mathrm{H}_2(P).$$

Ô Špeciálny prípad predchádzajúcej lemy pre $f(t) = -\log_2(p_t)$.

V Nech $P = \langle p_1, \ldots, p_k \rangle$ je rozdelenie pravdepodobností. Potom

$$\mathrm{cak}(P, n) < n\mathrm{H}_2(P) + 1.$$

Ô Všimnime si, že dĺžka kódového slova $\mathrm{cwd}(\mathrm{rep}(\mathrm{A}_P(\langle i_1, \ldots, i_n \rangle)))$ je počet dyadických cifier čísla $\mathrm{rep}(\mathrm{A}_P(\langle i_1, \ldots, i_n \rangle))$ za dyadickou čiarkou. Ak $\mathrm{rep}(\mathrm{A}_P(\langle i_1, \ldots, i_n \rangle)) = \frac{x}{2^t}$, tento počet nepresahuje t.

Ak $I = [u, v)$, tak $\mathrm{DĺžkaIntervalu}(I) = v - u$, a teda ak platí $\frac{1}{2^t} \leq v - u < \frac{1}{2^{t-1}}$, tak $t - 1 < -\log_2(v - u) \leq t$, z čoho $t < 1 - \log_2(v - u) = 1 - \log_2(\mathrm{DĺžkaIntervalu}(I))$.

V našom prípade $I = \mathrm{A}_P(\langle i_1, \ldots, i_n \rangle)$ teda dostávame $\mathrm{ds}(\mathrm{cwd}(\mathrm{rep}(\mathrm{A}_P(\langle i_1, \ldots, i_n \rangle)))) < 1 - \log_2(\mathrm{DĺžkaIntervalu}(\mathrm{A}_P(\langle i_1, \ldots, i_n \rangle))) = 1 - \log_2(p_{i_1} \cdots p_{i_n}) = 1 - \sum_{j=1}^{n} \log_2 p_{i_j}$.

Upravujme:

$\mathrm{cak}(P, n) =$
$= \sum_{\langle i_1, \ldots, i_n \rangle \in \{1, \ldots, k\}^n} p_{i_1} \ldots p_{i_n} \cdot \mathrm{ds}(\mathrm{cwd}(\mathrm{rep}(\mathrm{A}_P(\langle i_1, \ldots, i_n \rangle)))),$
$< \sum_{\langle i_1, \ldots, i_n \rangle \in \{1, \ldots, k\}^n} p_{i_1} \ldots p_{i_n} \cdot (1 - \sum_{j=1}^{n} \log_2 p_{i_j}),$
$= \sum_{\langle i_1, \ldots, i_n \rangle \in \{1, \ldots, k\}^n} p_{i_1} \ldots p_{i_n} - \sum_{\langle i_1, \ldots, i_n \rangle \in \{1, \ldots, k\}^n} p_{i_1} \ldots p_{i_n} \sum_{j=1}^{n} \log_2 p_{i_j},$
$= 1 - \sum_{\langle i_1, \ldots, i_n \rangle \in \{1, \ldots, k\}^n} p_{i_1} \ldots p_{i_n} \sum_{j=1}^{n} \log_2 p_{i_j}$ (podľa lemy ...),
$= 1 - \sum_{j=1}^{n} \sum_{\langle i_1, \ldots, i_n \rangle \in \{1, \ldots, k\}^n} p_{i_1} \ldots p_{i_n} \log_2 p_{i_j},$
$= 1 - \sum_{j=1}^{n} (-\mathrm{H}_2(P)),$
$= 1 + n\mathrm{H}_2(P).$

6 Samoopravné kódy

– Budeme konštruovať rovnomerné (blokové) kódy v binárnej abecede $\Sigma_{\mathrm{C}} = \{0, 1\}$, pričom chápeme prirodzene $0 = 0$ a $1 = 1$.

Budeme predpokladať, že v nich nastanú iba chyby typu, že symbol sa pri prenose šumom zmení na iný symbol, a to na každý s rovnakou pravdepodobnosťou.

– Nemôžeme konštruovať úplné kódy, lebo príjemca by nikdy nedokázal zistiť, či nastala zmena.

Budeme preto konštruovať iba také kódy, aby príjemca (s dostatočne veľkou pravdepodobnosťou) dokázal chybné kódové slovo zachytiť, ba dokonca opraviť.

- *Kód celkovej kontroly parity* obsahujúci len slová s párnym počtom 1 dokáže odhaliť chyby veľkosti 1.

(Napr.) posledný znak môžeme chápať ako kontrolný – je to celková parita ostatných – informačných.

- *Kód dvojrozmernej kontroly parity* (*obdĺžnikový kód*) má kontrolné parity v riadkoch i stĺpcoch a tiež celkovú kontrolu parity.

– ...opravovanie 1 chyby pri tomto kóde...

? Nech $k \in \mathbb{N}^+$. Pod *Hammingovou k-vzdialenosťou* budeme rozumieť funkciu $\mathrm{Ham}_k : \{0,1\}^k \times \{0,1\}^k \to \mathbb{N}$ definovanú vzťahom:

$$\mathrm{Ham}_k(\langle a_1, \ldots, a_k\rangle, \langle b_1, \ldots, b_k\rangle) = \sum_{i=1}^{k} |a_i - b_i|.$$

L Hammingova k-vzdialenosť je metrika.

? Definujme operáciu $\oplus : \{0,1\} \times \{0,1\} \to \{0,1\}$ tak, že $\oplus(0,0) = \oplus(1,1) = 0$ a $\oplus(0,1) = \oplus(1,0) = 1$.

– $\oplus$ je vlastne exkluzívne alebo (tzv. xor).

L $\mathrm{Ham}_k(\langle a_1, \ldots, a_k\rangle, \langle b_1, \ldots, b_k\rangle) = \sum_{i=1}^{k} a_i \oplus b_i$.

? Definujme operáciu $\oplus_k : \{0,1\}^k \times \{0,1\}^k \to \{0,1\}^k$:

$$\oplus_k(\langle a_1, \ldots, a_k\rangle, \langle b_1, \ldots, b_k\rangle) = \langle a_1 \oplus b_1, \ldots, a_k \oplus b_k\rangle.$$

– Index k budeme pri $\oplus_k$ vynechávať.

? Pod *dĺžkou rovnomerného kódu* budeme rozumieť dĺžku ľubovoľného jeho slova.

? Pod *Hammingovou minimálnou vzdialenosťou kódu* C dĺžky k budeme rozumieť číslo $\mathrm{Hvk}(C) = \min\{\mathrm{Ham}_k(c_1, c_2) : c_1, c_2 \in C, c_1 \neq c_2\}$.

– Namiesto o príjemcovi budeme hovoriť o samotnom kóde:

Aby kód vedel odhaliť r chýb, musí platiť $\mathrm{Hvk}(C) \geq r + 1$.

Aby kód vedel opraviť r chýb, musí platiť $\mathrm{Hvk}(C) \geq 2r + 1$.

Ako také kódy zostrojiť?

? Ak $r \in \mathbb{N}$ a $v \in \Sigma_{\mathrm{C}}^k$, definujme *guľu so stredom v a polomerom r* takto:

$$\mathrm{Hg}_k(v, r) = \{u \in \Sigma_{\mathrm{C}}^k : \mathrm{Ham}_k(u, v) \leq r\}.$$

L $|\mathrm{Hg}_k(v, r)| = \sum_{i=0}^{r} \binom{k}{i}$.

– Na to, aby takýto kód C vedel odhaliť r chýb, nesmie žiadna guľa polomeru r so stredom vo vektore z C obsahovať iný vektor z C.

– Na to, aby takýto kód C vedel opraviť r chýb, musia byť každé dve gule polomeru r so stredmi vo vektoroch z C disjunktné.

- *Opakovací kód* $\{0^{2r+1}, 1^{2r+1}\}$ dokáže opraviť chyby veľkosti r.

L Ak pre všetky $u, v \in C$ platí $\mathrm{Hg}_k(u, r) \cap \mathrm{Hg}_k(v, r) = \emptyset$, tak

$$|C| \leq \left\lfloor \frac{2^k}{\sum_{i=0}^{r} \binom{k}{i}} \right\rfloor .$$

? Kód C nazývame *lineárny*, ak je podpriestorom vektorového priestoru $\{0,1\}^k$ (t. j. ak z $u, v \in C$ platí $u \oplus v \in C$).

Ak jeho dimenzia (počet lineárne nezávislých vektorov) je n, hovoríme o *lineárnom* (k, n)*-kóde*.

- Kód celkovej kontroly parity je lineárny $(k, k-1)$-kód.
- Opakovací kód je lineárny $(k, 1)$-kód.

– Každý lineárny (k, n)-kód C má n informačných symbolov a $k - n$ kontrolných. Ak je totiž $b_1, \ldots, b_n$ báza priestoru C, každé kódové slovo v vieme napísať v tvare $v = x_1 b_1 \oplus \cdots \oplus x_n b_n$ pre nejaké $x_1, \ldots, x_n \in \{0,1\}$.

- Báza kódu celkovej kontroly parity je $\{10^{k-2}1, 010^{k-3}1, 0^2 10^{k-4}1, \ldots, 0^{k-2}11\}$, má teda $k-1$ prvkov.
- Báza opakovacieho kódu dĺžky k je jednoprvková: $\{1^k\}$.

? Lineárny $(2^m - 1, 2^m - 1 - m)$-kód pre $m \geq 2$ nazývame Hammingov kód.

- ... konštrukcia Hammingovho kódu pre $m = 4$, t. j. lineárneho $(15, 11)$-kódu...
- ... dekódovanie...

X písomka

1 (4 body)
Vyslovte a dokážte Kraftovu-McMillanovu vetu.

2 (4 body)
Vyslovte a dokážte čo najlepší horný i dolný odhad ceny optimálnej kódovej postupnosti pomocou entropie.

3 (4 body)
Skonštruujte Fanovu kódovú postupnosť pre rozdelenie pravdepodobností $\langle 0.2, 0.4, 0.1, 0.25, 0.05\rangle$.

4 (4 body)
Skonštruujte binárnu Huffmanovu kódovú postupnosť pre rozdelenie pravdepodobností $\langle 0.2, 0.4, 0.1, 0.25, 0.05\rangle$.

5 (4 body)
Ak sú pravdepodobnosti výskytu písmen `B` a `L` po rade $\frac{1}{3}$ a $\frac{2}{3}$, zakódujte aritmetickým kódovaním vetu `BLB L. B. BLBL.` (bez interpunkcie a medzier) s kódmi veľkosti 3.