1. Máme n mincí, z nichž nejvýše jedna je falešná. Pozná se podle toho, že má jinou hmotnost než ostatní mince (ty váží všechny stejně). Mince vážíme na rovnoramenné váze se dvěma miskami. Nalezněte horní mez pro počet mincí n, která umožní nalézt falešnou minci pomocí k vážení a současně rozhodnout, zda je nalezená falešná mince lehčí či těžší. Pokuste se úlohu interpretovat z hlediska teorie informace.

Řešení:

Pěkný výklad problému:

- http://www.geeksforgeeks.org/decision-trees-fake-coin-puzzle/
- http://grokrate.blogspot.cz/2007/05/solving-oddball-problem-with-n-balls.html

Velmi zjednodušeně lze řešení popsat takto. Při každém vážení nastane právě jedna ze tří možností:

(a) první miska je těžší,

(b) první miska je lehčí,

(c) misky jsou v rovnováze.

Pomocí zadaného počtu vážení k tak lze rozlišit nejvýše 3^k možností. Pro n mincí máme dohromady $2n+1$ stavů (jedna z mincí je těžší, jedna z mincí je lehčí, žádná mince není falešná). Pokud chceme splnit zadání, musí nutně platit $2n+1 < 3^k$, nebo-li

$$n < \frac{3^k - 1}{2}.$$

Možná interpretace z hlediska teorie informace: každé vážení přinese právě $\log 3$ bity informace, entropie celé množiny stavů je $\log(2n+1)$. K určení falešné mince je proto třeba realizovat

$$k > \frac{\log(2n+1)}{\log 3}$$

vážení. To dává stejný výsledek.

2. **Entropie a dělení.** Náhodně vybereme přirozené číslo od 1 do 1050 a určíme jeho zbytek po dělení 5. Určete:

(a) entropii vybraného čísla,

(b) entropii jeho modulu,

(c) střední podmíněnou entropii modulu vzhledem k vybranému číslu,

(d) střední podmíněnou entropii vybraného čísla vzhledem k modulu.

Řešení:

(a) $\log 1050$.

(b) Zbytky $0, \ldots, 4$ jsou zřejmě stejně pravděpodobné a proto je entropie modulu $\log 5$.

(c) Pokud známe vybrané číslo, potom zbytek je funkcí tohoto číslo a proto je entropie nulová.

(d) Při znalosti zbytku po dělení 5 máme zřejmě na výběr mezi $\frac{1050}{5}$ stejně pravděpodobnými čísly a tudíž je entropie rovna $\log 210$.

3. **Entropie funkce náhodné veličiny.** Uvažujme konečné abecedy Λ a Γ a zobrazení $g : \Lambda \to \Gamma$. Je-li X náhodná veličina s výběrovým prostorem Λ, položme $Y = g(X)$. Ukažte, že zobrazením X nezvýšíme entropii $H(X)$, přesněji: vždy platí $H(X) \geq H(Y)$ a rovnost nastává právě tehdy, když g je prosté na množině $\{\, x \in \Lambda \mid p_X(x) > 0 \,\}$.

Řešení:

Pro rozdělení veličiny Y platí

$$p_Y(y) = \sum_{\substack{x \in \Lambda \\ g(x)=y}} p_X(x), \quad y \in \Gamma. \tag{1}$$

Entropii $H(X)$ můžeme vyjádřit takto:

$$\begin{aligned} H(X) &= -\sum_{x \in \Lambda} p_X(x) \log p_X(x) = -\sum_{y \in \Gamma} \sum_{\substack{x \in \Lambda \\ g(x)=y}} p_X(x) \log p_X(x) \\ &\geq -\sum_{y \in \Gamma} \sum_{\substack{x \in \Lambda \\ g(x)=y}} p_X(x) \log p_Y(y) = -\sum_{y \in \Gamma} p_Y(y) \log p_Y(y) = H(Y), \end{aligned}$$

kde nerovnost výše plyne z monotonie logaritmu a vztahu (1). Rovnost přitom nastává právě tehdy, když za sumou $\sum_{y \in \Gamma}$ je pouze jeden sčítanec, což je ekvivalentní požadavku, aby funkce g byla prostá s pravděpodobností 1.

4. Byl nalezen jednoznačně dekódovatelný n-ární kód s délkami kódových slov $(1, 1, 2, 3, 2, 3)$. Jaká je minimální možná hodnota n, tedy počet znaků kódovací abecedy?

Řešení:

Každý jednoznačně dekódovatelný n-ární kód splňuje Kraftovu nerovnost a proto musí platit

$$f(n) := n^{-1} + n^{-1} + n^{-2} + n^{-3} + n^{-2} + n^{-3} \leq 1.$$

Kód nemůže být binární, neboť $f(2) > 1$. Zřejmě

$$f(3) = 2\left(\frac{1}{3} + \frac{1}{9} + \frac{1}{27}\right) = \frac{26}{27} \leq 1$$

a tudíž existuje ternární ($n = 3$) jednoznačně dekódovatelný kód nad abecedou $\{0, 1, 2\}$ se zadanými délkami kódových slov.

5. Navrhněte kódovací algoritmus pro markovský zdroj informace a posuďte jeho efektivitu. Návod: použijte různé kódy v závislosti na aktuálně načteném zdrojovém symbolu.

Řešení:

Uvažujme markovský řetězec s maticí přechodu $\mathbf{P} = (p_{ij})_{i,j\in\Lambda}$, který popisuje náhodné přechody mezi symboly n-prvkové abecedy Λ. Pro každé $i \in \Lambda$, nalezněme Huffmanův kód C_H^i pro zdroj popsaný pravděpodobnostmi $(p_{i1}, \ldots, p_{in})$. Kódování bude probíhat takto:

(a) první znak lze zakódovat kódem s pevnou délkou slova, tedy nejvýše pomocí $\lceil \log n \rceil$ bitů;

(b) libovolný znak $i \in \Lambda$ na pozici m ($m \geq 2$) zakódujeme pomocí kódu C_H^j, je-li na $(m-1)$-pozici znak $j \in \Lambda$.

Efektivitu tohoto kódu budeme vyhodnocovat pomocí rychlosti entropie $H(X_1^\infty)$, která měří entropii na znak markovského zdroje. Při tom lze asymptoticky zanedbat první symbol z části (a), který se v celeém řetězci vyskytne pouze jednou.

Uvažujeme, že $\mathbf{p}_{X_1} = (p_i)_{i\in\Lambda}$ je stacionární rozdělení tohoto řetězce a označme takto vzniklý kód jako C. Jelikož je Huffmanův kód optimální, platí pro každé $i \in \Lambda$

$$H(X_2|X_1 = i) \leq L(C_H^i) < H(X_2|X_1 = i) + 1,$$

což dává součtem přes $i \in \Lambda$

$$H(X_2|X_1) \leq \underbrace{\sum_{i\in\Lambda} p_i \cdot L(C_H^i)}_{L(C)} < H(X_2|X_1) + 1.$$

Jelikož pro markovský řetězec platí $H(X_1^\infty) = H(X_2|X_1)$, redundance kódu C je nejvýše $H(X_1^\infty) + 1$, tedy o jeden bit více než je entropie, což je stejný výsledek jako pro bezpaměťový zdroj.

6. Bezpaměťový zdroj nad abecedou $\Lambda = \{a, b, c\}$ má pravděpodobnosti jednotlivých písmen

$$p(a) = 0.3, \;\; p(b) = 0.6, \;\; p(c) = 0.1.$$

Nad znaky uvažujeme uspořádání $a < b < c$. Použijte algoritmus aritmetického kódování s délkou bloku 5 k zakódování zdrojového řetězce *bacca*. Dále popište způsob dekódování takto nalezeného kódového slova.

Řešení:

Délka bloku $M = 5$. Jelikož je M stejné jako délka kódovaného řetězce, dostaneme pouze jedno kódové slovo. Položme dále $p_0 = 1$, $F_0 = 0$ a

$$f_a = 0, \;\; f_b = p(a) = 0.3, \;\; f_c = p(a) + p(b) = 0.9.$$

Budeme postupně číst jednotlivé znaky řetězce *bacca* a počítat

$$p_k = p_{k-1} \cdot p(u_k), \;\; F_k = F_{k-1} + p_{k-1} \cdot f_{u_k}, \quad k = 1, \ldots, 5.$$

První krok:

$$\begin{aligned} p_1 &= p(b) = 0.6, \\ F_1 &= F_0 + p_0 \cdot f_b = 0.3. \end{aligned}$$

Druhý krok:

$$\begin{aligned} p_2 &= p_1 \cdot p(a) = 0.18, \\ F_2 &= F_1 + p_1 \cdot f_a = 0.3. \end{aligned}$$

Konečně v pátem kroku dostaneme

$$\begin{aligned} p_5 &= p_4 \cdot p(a) = 0.00054, \\ F_5 &= F_4 + p_4 \cdot f_a = 0.4782. \end{aligned}$$

Určíme délku kódového slova:

$$\tilde{l} = \left\lceil \log \frac{1}{p_5} \right\rceil + 1 = 12.$$

Číslo F_5 převedeme do dvojkové soustavy, ořízneme na délku 12 a k výsledku přičteme 2^{-12}:

$$\begin{aligned} (F_5)_2 &= (0.4782)_2 = (0.011110100110)_2, \\ (F_5)_2 + 2^{-12} &= (0.011110100111)_2. \end{aligned}$$

Výsledné binární kódové slovo je 011110100111.

Jak bude probíhat dekódování? Máme k dispozici jen zakódované slovo a pravděpodobnosti zdrojových znaků. Z nich můžeme spočítat horní mez pro délku hypotetického kódového slova:

$$\tilde{l}_{max} = \left\lceil 5 \log \frac{1}{p(c)} \right\rceil + 1 = \lceil 5 \log 10 \rceil + 1 = 18.$$

Jelikož je slovo 011110100111 kratší než $\tilde{l}_{max}$, stačí uvažovat celé zadané kódové slovo. Položme $k = 1$, $p_0 = 1$ a určeme desítkové vyjádření čísla $\tilde{F}_1 = (0.011110100111)_2$: dostaneme

$$\tilde{F}_1 = 0.4782714844.$$

Hledáme největší f_{α_i} takové, že $f_{\alpha_i} \leq \tilde{F}_1$. Zřejmě $\alpha_i = b$, $f_b = 0.3$. První zakódovaný znak byl tedy b a my provedeme update pravděpodobností zbylé části řetězce:

$$\tilde{F}_2 = \frac{\tilde{F}_1 - f_b}{p(b)} = 0.2971191407.$$
$$p_1 = p_0 \cdot p(b) = 0.6.$$

Pravděpodobnost $\tilde{F}_2$ odpovídá zatím nedekódovanému řetězci *acca*. Postupujeme stejně: největší f_{α_i} je nyní $f_a = 0$, což dává:

$$\tilde{F}_3 = \frac{\tilde{F}_2 - f_a}{p(a)} = 0.9903971357.$$
$$p_2 = p_1 \cdot p(a) = 0.18.$$

Takto pokračujeme analogicky až do pátého kroku, kdy dekódujeme poslední symbol a.

7. **Instantní kódy.** Pro informační zdroj nad 6-prvkovou abecedou produkující znaky s pravděpodobnostmi

$$p_1 = 0.1,\ p_2 = p_3 = 0.25,\ p_4 = 0.05,\ p_5 = 0.15,\ p_6 = 0.2$$

nalezněte

(a) 2 různé kódy Shannonovského typu,

(b) Fanův kód,

(c) Shannonův kód,

(d) Huffmanův kód.

Určete jejich střední délky a posuďte efektivitu každého z nich.

Řešení:

(a) Nalezeneme kódový strom, jehož 6 listů je v hloubce $\lceil -\log p_i \rceil$, $i = 1, \ldots, 6$. Tak dostaneme kód Shannonovského typu

$$0011,\ 10,\ 11,\ 00101,\ 011,\ 010.$$

Jiný kód získáme bitovou inverzí výše uvedeného.

(b) Fanův kód: 1110, 00, 01, 1111, 110, 10.

(c) Uvažujeme permutaci π prvků z $\{1, \ldots, 6\}$ takovou, že

$$\pi(2) = 1, \pi(3) = 2, \pi(6) = 3, \pi(5) = 4, \pi(1) = 5, \pi(4) = 6.$$

Binární kód slova $C(i)$ určíme pomocí binárního rozvoje čísla F_i, kde

$$F_i := \begin{cases} 0 & i = 1, \\ \sum_{j=1}^{i-1} p_{\pi^{-1}(j)} & i = 2, \ldots, 6, \end{cases}$$

přičemž uvažujeme vždy prvních $\ell_i := \lceil -\log p_{\pi^{-1}(i)} \rceil$ pozic z daného rozvoje.

(d) Huffmanův kód může vypadat např. takto: 0010, 01, 00, 0011, 000, 01.

8. **Blokové kódování**. Bezpaměťový zdroj nad abecedou $\Lambda = \{1, 2, 3, 4, 5, 6\}$ generuje řetězec začínající znaky

321463324234

Zakódujte tuto část řetězce pomocí optimálního instantního kódu o délce vstupního bloku 2 a posuďte efektivitu takového kódování ve srovnání s optimálním kódováním jednotlivých symbolů abecedy Λ.

Řešení:

Zdroj je bezpaměťový, pravděpodobnosti jednotlivých symbolů odhadneme na základě relativních četností z pozorované sekvence:

$$p(1) = \tfrac{1}{12}, p(2) = \tfrac{3}{12}, p(3) = \tfrac{4}{12}, p(4) = \tfrac{3}{12}, p(6) = \tfrac{1}{12}.$$

Optimální kódování jednotlivých symbolů je Huffmanovo, jeho střední kódová délka je 2.167. Entropie zdroje je 2.126.

V případě Huffmanova kódování o délce bloku 2 musíme nejprve určit pravděpodobnosti všech dvojic znaků z množiny $\{1,2,3,4,6\} \times \{1,2,3,4,6\}$:

$$p(11) = p(1)p(1) = \tfrac{1}{12^2}, p(12) = p(1)p(2) = \tfrac{2}{12^2} \quad \text{atd.}$$

Pro nový zdroj určíme blokový Huffmanův kód C^2, jeho střední délku $\frac{L(C^2)}{2}$ porovnáme s entropií.

9. **Optimální konstrukce otázek.** Alice vybere náhodně přirozené číslo od 1 do 5 s pravděpodobnostmi

$$p_1 = 0.25,\ p_2 = 0.25,\ p_3 = 0.2,\ p_4 = p_5 = 0.15.$$

Úkolem Boba je uhodnout zvolené číslo pomocí sekvence binárních otázek (odpověď ano/ne). Jak má Bob postupovat, aby byl počet otázek minimální? Úlohu řešte za předpokladu, že pravděpodobnosti $p_1, \ldots, p_5$ jsou pro Boba

(a) známé,

(b) neznámé.

Řešení:

Za předpokladu (a) úlohu převedeme na konstrukci Huffmanova kódu pro zdroj s uvedenými pravděpodobnostmi. Binární Huffmanův kód vypadá např. takto:

$$C(1) = 00,\ C(2) = 10,\ C(3) = 11,\ C(4) = 010,\ C(5) = 011.$$

Otázky zkonstruujeme podle způsobu procházení Huffmanova kódového stromu od kořene. Začneme otázkou „Je neznámé číslo 2 nebo 3?“, neboť ta umožní odlišit první bit kódového slova. V případě kladné odpovědi se ptáme „Je neznámé číslo 3?“, v případě záporné „Je neznámé číslo 4 nebo 5?“ atd. Střední hodnota počtu otázek, které musíme položit, je rovna střední délce Huffmanova kódu: $L(C) = 2.3$.

V situaci (b) bude Bob (pesimisticky) předpokládat, že Alice volí všechna čísla se stejnou pravděpodobností. Tím dosáhne konzervativního odhadu počtu nutných otázek. Postup je stejný jako v (a). Nalezneme tak Huffmanův kód pro zdroj s pravděpodobnostmi rovnými $\frac{1}{5}$:

$$C'(1) = 00,\ C'(2) = 10,\ C'(3) = 11,\ C'(4) = 010,\ C'(5) = 011.$$

Oba kódy C i C' jsou stejné, jejich střední délka je však jiná, neboť $L(C') = 2.4$. Tedy v průměru potřebuje Bob v situaci (b) více otázek.

10. Pro bezpaměťový zdroj z Příkladu 6 dekódujte posloupnost

$$1110100111111100101$$

které byla zakódována aritmetickým kodérem s délkou bloku $M = 2$.

Řešení:

Spočítejme horní mez pro délku kódových slov:

$$\tilde{l}_{max} = \left\lceil 2 \log \frac{1}{p(c)} \right\rceil + 1 = \lceil 2 \log 10 \rceil + 1 = 8.$$

Vezmeme tedy prvních 8 bitů, položíme $k = 1$, $p_0 = 1$, a hledáme desítkové vyjádření čísla $\tilde{F}_1 = (0.11101001)_2$. Dostaneme

$$\tilde{F}_1 = 0.91015625.$$

Jelikož $f_c = 0.9 < \tilde{F}_1$, první znak je c. Upravíme pravděpodobnosti ve zbylé části řetězce:

$$\tilde{F}_2 = \frac{\tilde{F}_1 - f_c}{p(c)} = 0.1015625.$$
$$p_1 = p_0 \cdot p(c) = 0.1.$$

Položíme $k = 2$. Jelikož $f_a = 0 < \tilde{F}_2$, druhý znak je a. Protože $k = M$, zbývá dopočítat skutečnou délku kódového slova:

$$\tilde{l} = \left\lceil \log \frac{1}{p(ca)} \right\rceil + 1 = 7.$$

Ze sekvence 11101001 délky 8 jsme tak extrahovali kódové slovo 1110100 délky 7. Dále tak pokračujeme dekódováním řetězce 111111100101. Jelikož $\tilde{l}_{max} = 8$, vezmeme slovo 11111110 délky 8 a postupujeme analogicky jako výše. Získáme tak kódové slovo 11111110 identické se vstupním, to odpovídá dvojici cc. V posleddním kroku dostaneme ze vstupního řetězce 0101 kódové slovo, které reprezentuje dvojici ba. Zdrojový řetězec byl tedy

$$caccba.$$

11. Máte k dispozici náhodný generátor bitů, tj. můžete simulovat hodnoty náhodné veličiny X takové, že $p_X(0) = p_X(1) = \frac{1}{2}$. Uvažujme nad abecedou $\Lambda = \{a, b, c\}$ pravděpodobnostní funkce

$$p(a) = \tfrac{1}{2},\ \ p(b) = p(c) = \tfrac{1}{4} \quad \text{a} \quad q(a) = \tfrac{1}{3},\ \ q(b) = \tfrac{1}{6},\ \ q(c) = \tfrac{1}{2}.$$

Popište, jakým způsobem byste simulovali náhodný výběr z rozdělení p a q pomocí opakovaného použití náhodného generátoru bitů. Pokuste se zdola odhadnout střední hodnotu počtu potřebných bitů.

Řešení:

Simulace z p: stačí nalézt Huffmanův kód pro p, např. $C_H(a) = 0$, $C_H(b) = 10$, $C_H(c) = 11$. Algoritmus:

(a) je-li náhodný bit roven 0, pak vypiš a;

(b) je-li náhodný bit roven 1, pak vygeneruj další bit a podle hodnoty vypiš buď b nebo c.

Uvedený postup odpovídá kódovacímu stromu, jehož střední délka je $L(C_H) = H(p) = \frac{3}{2}$. Průměrný počet potřebných bitů je tedy $\frac{3}{2}$.

Simulace z q: pravděpodobnosti nejsou dyadické, dosáhneme jich tedy pouze v dlouhé sérii pokusů. Algoritmus je určen nekonečným rozvojem pravděpodobností: $\frac{1}{3} = \sum_{n=1}^{\infty} 2^{-2n}$, $\frac{1}{6} = \sum_{n=1}^{\infty} 2^{-2n-1}$. Simulaci lze popsat (nekonečným) rozhodovacím stromem, algoritmus vypadá např. takto:

(a) pokud je první bit roven 1, vypiš c;

(b) pokud je první bit roven 0 a generované bitové slovo neobsahuje 11 a končí 00, vypiš a;

(c) pokud je první bit roven 0 a generované bitové slovo neobsahuje 00 a končí 11, vypiš b.

Střední délka takového "kodovacího" stromu je

$$L(C) = \sum_{n=1}^{\infty} n 2^{-n} = 2,$$

přičemž entropie je pouze $H(q) \doteq 1.46$. To znamená, že k simulaci jednoho znaku potřebujeme v průměru 2 bity.

12. Pro bezpaměťový zdroj nad abecedou $\{a, b, c, d\}$ s pravděpodobnostmi

$$p(a) = 0.1, \;\; p(b) = 0.4, \;\; p(c) = p(d) = 0.25,$$

nalezněte Tunstallův kód o délce kódového slova $L = 3$ a posuďte jeho efektivitu.

Řešení:

Nejprve určíme počet neterminálních uzlů budovaného stromu:

$$N = \left\lfloor \frac{2^3 - 1}{4 - 1} \right\rfloor = 2.$$

Dostaneme tedy $n = 1 + 3N = 7$ zpráv. Tunstallova množina zpráv má tento tvar:

$$\{a, ba, bb, bc, bd, c, d\}.$$

Tunstallův kód vznikne např. tímto přiřazením:

$$a \mapsto 000, ba \mapsto 001, bb \mapsto 010, bc \mapsto 100, bd \mapsto 011, c \mapsto 101, d \mapsto 110.$$

Střední délka zpráv je $E(M) = 1 + 0.4 = 1.4$, entropie zdroje je $H(U) = 1.86$. Tedy

$$\frac{L}{E(M)} = \frac{3}{1.4} \doteq 2.14 \geq H(U) = 1.86.$$

13. Markovský řetězec má matici přechodu

$$\mathbf{P} = \begin{pmatrix} 1/16 & 15/16 \\ 1/2 & 1/2 \end{pmatrix}$$

a počáteční rozdělení $\mathbf{p}_{X_1}$ je rovno vektoru $(\frac{8}{23}, \frac{15}{23})$. Určete, zda v této situaci existuje rychlost entropie a v kladném případě určete její hodnotu.

Řešení:

Markovský řetězec je ireducibilní aperiodický a počáteční rozdělení $\mathbf{p}_{X_1} = (\frac{8}{23}, \frac{15}{23})$ je rozdělením stacionárním, neboť platí $\mathbf{p}_{X_1}\mathbf{P} = \mathbf{p}_{X_1}$. Proto existuje rychlost entropie $H(X_1^\infty)$ tohoto řetězce a nalezneme ji takto:

$$H(X_1^\infty) = H(X_2|X_1) = \tfrac{8}{23} \cdot \underbrace{H(1/16, 15/16)}_{\text{entropie 1. řádku } \mathbf{P}} + \tfrac{15}{23} \cdot \underbrace{H(1/2, 1/2)}_{\text{entropie 2. řádku } \mathbf{P}} \doteq 0.7695.$$

14. Experimentálně (např. v MATLABu) zdůvodněte, proč je markovský řetězec z Příkladu 13 ireducibilní aperiodický. Dále přesně vysvětlete, proč je řetězec s maticí

$$\mathbf{Q} = \begin{pmatrix} 0 & 2/3 & 0 & 1/3 \\ 1/3 & 0 & 2/3 & 0 \\ 0 & 1/3 & 0 & 2/3 \\ 2/3 & 0 & 1/3 & 0 \end{pmatrix}$$

ireducibilní, ovšem ne aperiodický. Tento fakt ověřte opět pomocí experimentů. Lze něco tvrdit o limitě

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbf{Q}^k \ ?$$

Řešení:

Pomocí vhodného nástroje snadno určíme hodnoty matice $\mathbf{P}^n$ pro $n \to \infty$. Zjistíme tak, že pro velmi velké n se matice $\mathbf{P}^n$ blíží matici

$$\begin{pmatrix} 2/3 & 1/3 \\ 2/3 & 1/3 \end{pmatrix},$$

která má v každém řádku stacionární rozdělení. To není náhoda: markovský řetězec s maticí $\mathbf{P}$ a stacionárním rozdělením $\mathbf{p}$ je ireducibilní aperiodický právě tehdy, když platí

$$\lim_{n\to\infty} \mathbf{P}^n = \begin{pmatrix} \mathbf{p} \\ \dots \\ \mathbf{p} \end{pmatrix}.$$

Všechny stavy v řetězci s maticí $\mathbf{Q}$ jsou trvalé a vzájemně dosažitelné, proto je tento řetězec ireducibilní. Není aperiodický, neboť návraty do libovolného stavu mohou nastat jen po sudém počtu kroků: každý stav má tak periodu 2. Protože je tento markovský řetězec ireducibilní, tvoří ergodický proces a platí pouze

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbf{Q}^k = \begin{pmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{pmatrix},$$

přičemž vektor $(1/4, 1/4, 1/4, 1/4)$ je stacionární rozdělení.

15. Pomocí jednoduché verze binárního LZ-kódování zakódujte řetězec

$$babacbcccbcac.$$

Řešení:

LZ-parsování daného řetězce délky $n = 13$ vypadá takto:

$$\underline{b}\,\underline{a}\,\underline{ba}\,\underline{c}\,\underline{bc}\,\underline{cc}\,\underline{bca}\,a$$

Pro kódování jednotlivých symbolů a, b, c je nutno vyhradit 2 bity, kodér může vypadat takto: $g(a) = 00$, $g(b) = 01$, $g(c) = 10$. Pro kódování ukazatelů je potřeba vyhradit $\lceil \log n \rceil$=4 bity. Dostaneme tak následující kód,

$$0g(b)\,0g(a)\,1\mathbf{0000}0g(a)\,0g(c)\,1\mathbf{0000}0g(c)\,1\mathbf{0011}0g(c)\,1\mathbf{0100}0g(a)\,1\mathbf{0001}$$

v němž jsou tučná bitová slova ukazatele na prefixy.

16. **Alternativní míra závislosti dvou veličin.** Nechť X_1 a X_2 jsou diskrétní náhodné veličiny se stejným pravděpodobnostním rozdělením na konečné množině Λ. Předpokládejme $H(X_1) \neq 0$ a položme

$$\rho = 1 - \frac{H(X_2|X_1)}{H(X_1)}.$$

(a) Ukažte, že platí $\rho = \frac{I(X_1;X_2)}{H(X_1)}$.

(b) Ukažte, že platí $0 \leq \rho \leq 1$.

(c) Kdy platí $\rho = 0$?

(d) Kdy platí $\rho = 1$?

Řešení:

(a) Platí

$$\rho = \frac{H(X_1) - H(X_2|X_1)}{H(X_1)} = \frac{H(X_2) - H(X_2|X_1)}{H(X_1)} = \frac{I(X_1;X_2)}{H(X_1)}.$$

(b) Jelikož $0 \leq H(X_2|X_1) \leq H(X_2) = H(X_1)$, dostáváme

$$0 \leq \frac{H(X_2|X_1)}{H(X_1)} \leq 1,$$

což znamená $0 \leq \rho \leq 1$.
(c) $\rho = 0 \Leftrightarrow I(X_1;X_2) = 0 \Leftrightarrow X_1$ a X_2 jsou nezávislé
(d) $\rho = 1 \Leftrightarrow H(X_2|X_1) = 0 \Leftrightarrow$ existuje funkce $f : \Lambda \to \Lambda$ taková, že $p_{X_2|X_1}(f(x_1)|x_1) = 1$, pro každé $x_1 \in \Lambda$ splňující $p_{X_1}(x_1) > 0$.

17. **Dvojně stochastické matice.** Matice přechodu $\mathbf{P} = (p_{ij})$ Markovova řetězce se stavovým prostorem $\Lambda = \{1, \ldots, n\}$ je *dvojně stochastická*, pokud platí $\sum_{i\in\Lambda} p_{ij} = 1$, pro každé $j \in \Lambda$. Ukažte, že matice $\mathbf{P}$ je dvojně stochastická právě tehdy, když rovnoměrné rozdělení na Λ je stacionárním rozdělením Markovova řetězce.

Řešení:

Nechť $\mathbf{P}$ je dvojně stochastická a $\mathbf{p} = (\frac{1}{n}, \ldots, \frac{1}{n})$ je rovnoměrné rozdělení na Λ. Potom platí $\mathbf{pP} = (\frac{1}{n}, \ldots, \frac{1}{n})\mathbf{P} = \mathbf{p}$ a $\mathbf{p}$ je tudíž stacionární rozdělení Markovova řetězce. Obráceně, buď stacionární rozdělení $\mathbf{p}$ rovnoměrné, $\mathbf{p} = (p_1, \ldots, p_n) = (\frac{1}{n}, \ldots, \frac{1}{n})$. Potom pro každé $j = 1, \ldots, n$,

$$\tfrac{1}{n} = p_j = \sum_{i=1}^{n} p_i p_{ij} = \tfrac{1}{n} \sum_{i=1}^{n} p_{ij}.$$

Tedy $\mathbf{P}$ je nutně dvojně stochastická.

18. **Entropie geometrického rozdělení.** Opakujme pokus se dvěma možnými výsledky 0 a 1 nezávisle na sobě až do prvního výskytu 1. Pravděpodobnost 1 je $p \in (0, 1)$. Náhodná veličina X označuje číslo pokusu, v němž došlo k prvnímu výskytu 1. Zřejmě tedy platí $p_X(x) = p(1-p)^{x-1}$, $x \in \mathbb{N}$. Určete entropii H_p, která je (přirozeně) definována jako $H_p = -\sum_{x=1}^{\infty} p_X(x) \log p_X(x)$, a stanovte průběh funkce H_p v závislosti na proměnné p.

Řešení:

$$\begin{aligned} H_p &= -\sum_{x=1}^{\infty} (1-p)^{x-1} p\big((x-1)\log(1-p) + \log p\big) \\ &= -p\left(\log(1-p) \sum_{x=1}^{\infty} (x-1)(1-p)^{x-1} + \log p \sum_{x=1}^{\infty} (1-p)^{x-1}\right). \end{aligned}$$

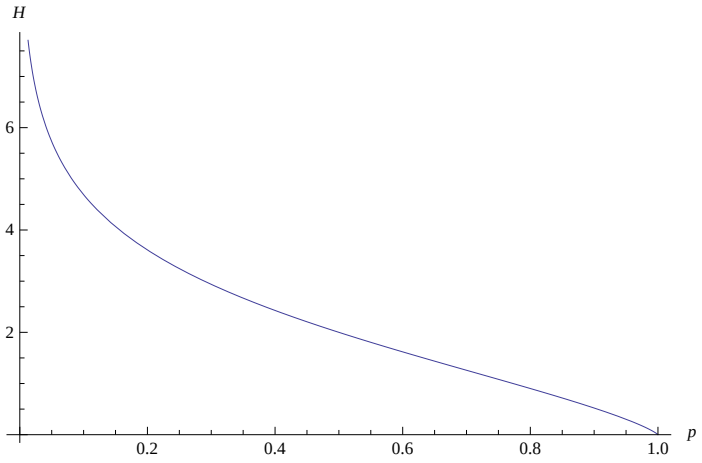
Součet první řady získáme takto:

$$\begin{aligned} \sum_{x=1}^{\infty} (x-1)(1-p)^{x-1} &= (1-p)\left(1 - \left(\sum_{x=2}^{\infty} (1-p)^x\right)'\right) = (1-p)\left(1 - \left(\frac{(1-p)^2}{p}\right)'\right) \\ &= \frac{1-p}{p^2}. \end{aligned}$$

Druhá řada je geometrická. Z toho plyne

$$H_p = -p\left(\frac{1-p}{p^2}\log(1-p) + \frac{1}{p}\log p\right) = -\frac{1-p}{p}\log(1-p) - \log p = \frac{H(p, 1-p)}{p}.$$

Povšimněme si, že platí $\lim_{p \to 1_-} H_p = 0$ a $\lim_{p \to 0_+} H_p = +\infty$. Výpočtem derivace se lze přesvědčit, že funkce H_p je klesající (Obrázek 1).



Obrázek 1: Průběh funkce H_p