

JY / Matematiikan ja tilastotieteen laitos
MATA2500 Informaatioteoria, Syksy 2019
Harjoitus 4 – Ratkaisuehdotuksia

Tämän viikon demoissa puhutaan koodausteoriasta.

1. Olkoon X TN-jakauma $(\{a, b, c, d, e, f, g, h\}, \{\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{3}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\})$.

- Laske jakauman entropia.
- Määritä $|S_\delta|$ sekä $H_\delta(X)$ kaikilla $0 < \delta < 1$.
- Muodosta jakauman Huffman-koodi, ja laske Huffman-koodin odotuspituus.

Ratkaisu:

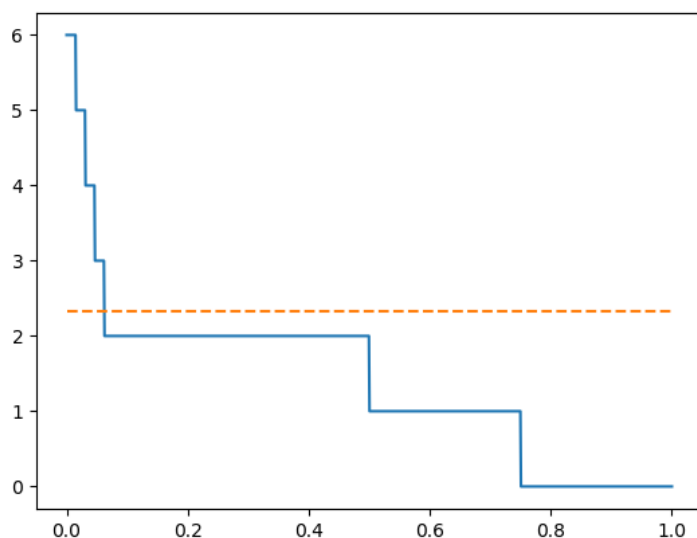
- Tutulla tavalla,

$$H = - \sum_i p_i \log(p_i) = \dots \approx 2.3.$$

- Kirjataan arvot taulukkoon:

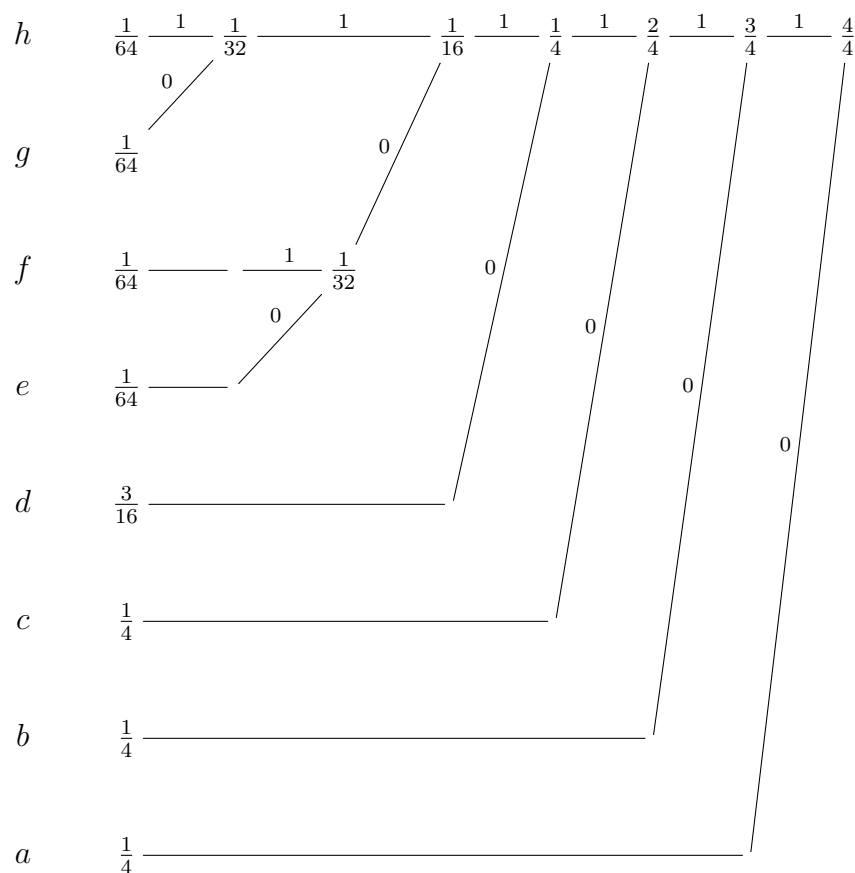
S_δ	$\delta \in$	$ S_\delta $	$H_\delta = \log_2(S_\delta)$
$\{a, b, c, d, e, f, g, h\}$	$(0, \frac{1}{64})$	8	≈ 3
$\{a, b, c, d, e, f, g\}$	$(\frac{1}{64}, \frac{2}{64})$	7	≈ 2.8
$\{a, b, c, d, e, f\}$	$(\frac{2}{64}, \frac{3}{64})$	6	≈ 2.8
$\{a, b, c, d, e\}$	$(\frac{3}{64}, \frac{1}{16})$	5	≈ 2.6
$\{a, b, c, d\}$	$(\frac{1}{16}, \frac{2}{4})$	4	≈ 2.3
$\{a, b, c\}$	$(\frac{1}{4}, \frac{1}{2})$	3	≈ 2
$\{a, b\}$	$(\frac{1}{2}, \frac{3}{4})$	2	≈ 1.6
$\{a\}$	$(\frac{3}{4}, 1)$	1	≈ 1

Tilanteesta voi myös piirtää kuvan:



Kuva 1: Tehtävään 1 liittyvä kuva. Kuvassa näkyy sekä $\delta \mapsto H_\delta$ että katkoviivalla jakauman entropia.

(c) Muodostetaan Huffman-koodi.



Koodiksi siis saadaan:

Symboli:	a	b	c	d	e	f	g	h
Koodi:	0	10	110	1110	111100	111101	111110	111111
Pituus:	1	2	3	4	6	6	6	6

Josta edelleen

$$L(c, X) = 1 \cdot \frac{1}{4} + 2 \cdot \frac{1}{4} + 3 \cdot \frac{1}{4} + 4 \cdot \frac{3}{16} + 4 \left(6 \cdot \frac{1}{64} \right) \approx 2.625.$$

Huomaa, että entropia ja Huffman-koodin odotuspituuksien erotus on alle 1, kuten kuuluukin.

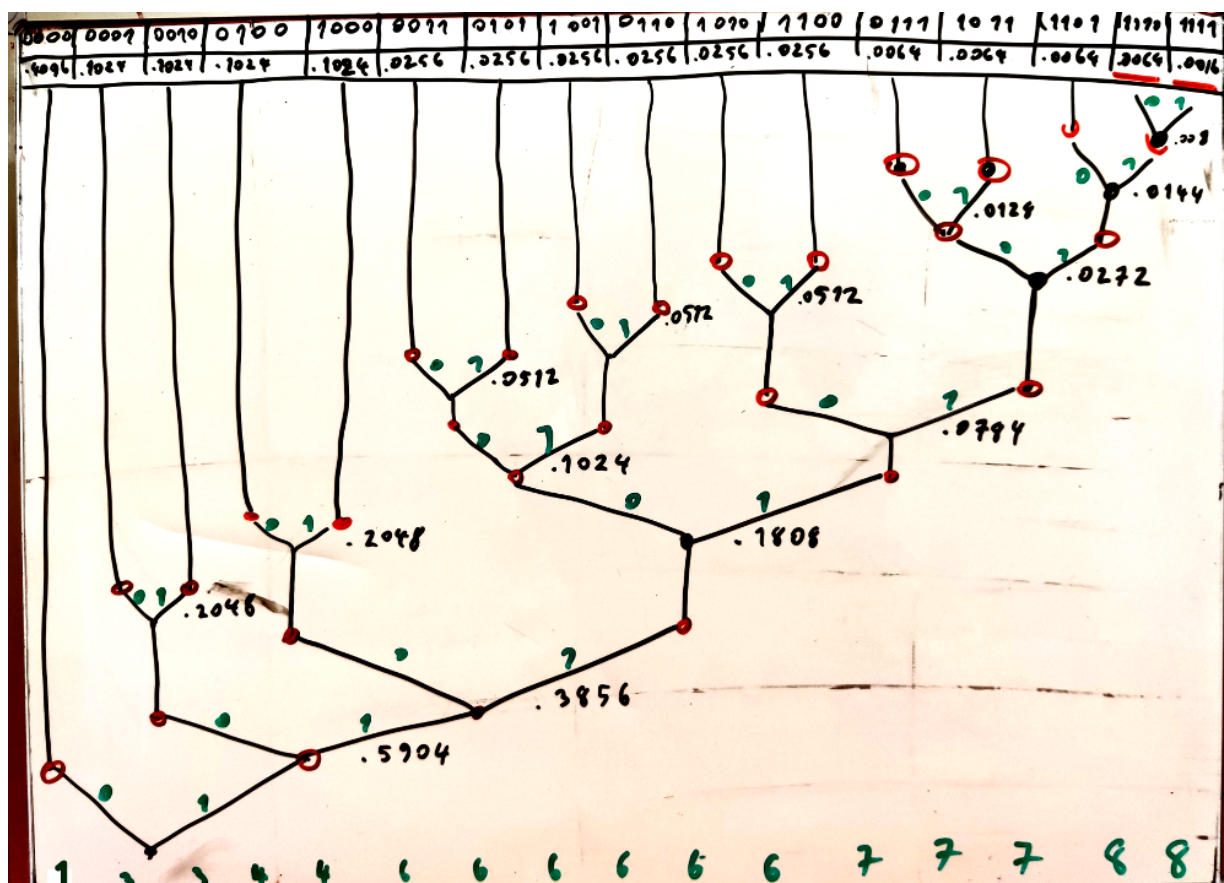
2. Olkoon X TN-jakauma $(\{0, 1\}, \{0.8, 0.2\})$.

Muodosta jakaumista X^2 , X^3 ja X^4 niiden Huffman-koodit. Laske myöskin kunkin jakauman entropia sekä jakaumaa vastaavan Huffman-koodin odotuspituus. *Ratkaisu:*

Huomaa, että Huffman-koodi ei ole täysin yksikäsitteinen, joten ratkaisusi saattaa näyttää hieman erilaiselta. Odotuspituuden tulisi kuitenkin olla ihan sama.

Edelleen huomaa, että $H(X) \approx 0.7219$, ja $H(X^N) = NH(X)$. (Todistimme, että kahdella riippumattomalla jakaumalla $H(X, Y) = H(X) + H(Y)$, ja sama todistus antaa että millä tahansa kokoelmalla riippumattomia jakaumia yhteisjakauman entropia on jakaumien entropioiden summa.)

X^2 :	Syöte	00	01	10	11
	TN	0.64	0.16	0.16	0.04
	Koodi	0	10	110	111



Kuva 2: Tehtävän 2c Huffman-koodin väkerrys, alareunassa pituudet.

	Syöte	000	001	010	100	011	101	110	111
X^3 :	TN	0.512	0.128	0.128	0.128	0.032	0.032	0.032	0.008
	Koodi:	0	100	101	110	11100	11101	11110	11111

X^4 : (Tästä tuli turhan työläs, sorry.)

Syöte	0000	0001	0010	0100	1000	0011	0101	1001	0110	1010	1100	0111	1011	1101	1110	1111
TN	0.4096	0.1024	0.1024	0.1024	0.1024	0.0256	0.0256	0.0256	0.0256	0.0256	0.0256	0.0064	0.0064	0.0064	0.0064	0.0016
ℓ	1	3	3	4	4	6	6	6	6	6	6	7	7	7	8	8
Koodit	0	100	101	1100	1101	111000	111001	111010	111011	111100	111101	1111100	1111101	1111110	11111110	11111111

	X	X^2	X^3	X^4
H	0.72	1.44	2.17	2.89
L	1	1.56	2.18	2.96

3. Tutki seuraavista koodeista ovatko ne yksikäsitteisesti purettavissa taikka etuliitevapaita.

- (a) $\{a \mapsto 00, b \mapsto 11, c \mapsto 0101, d \mapsto 111, e \mapsto 1010, f \mapsto 100100, g \mapsto 0110\}$
 (b) $\{x \mapsto 00, y \mapsto 11, z \mapsto 1101, w \mapsto 0010\}$

Osoita, että etuliitevapaa koodi on aina yksikäsitteisesti purettavissa. *Ratkaisu:*

- (a) Ei ole etuliitevapaa sillä koodi $c(b)$ on koodin $c(d)$ etuliite. Myöskään ei ole yksikäsitteisesti purettavissa, sillä $c(bd) = 11111 = c(db)$.
 (b) Ei ole etuliitevapaa sillä koodi $c(x)$ on koodin $c(w)$ etuliite. On yksikäsitteisesti purettavissa seuraavalla algoritmilla:

- i. Huomaa, että kaikissa mahdollisissa viesteissä on parillinen määrä symboleita.
- ii. Jaa viesti $s_1 s_2 \cdots s_{2n}$ kahden mittaiseen blokkeihin, merkitään $B_1 = s_1 s_2$, $B_2 = s_3 s_4$, $B_k = s_{2k-1} s_{2k}$.
- iii. Jokainen blokki B_j joka on muotoa 01 tai 10 on välttämättä joko koodin $c(z)$ tai $c(w)$ loppuosa, ja sen vasemmalla puolella on oltava blokki muotoa 11 tai 00. Puretaan nämä blokkiparit merkeiksi z ja w , ja huomataan että kaikki loput blokit ovat joko $c(x)$ tai $c(y)$.

Huom! Jos haluaa näyttää, että koodi on etuliitevapaa tai yksikäsitteisesti purettavissa, *ei riitä* tarkistaa Kraftin-MacMillanin epäyhtälöä. Kyseinen ey. sanoo ai-noastaan että on olemassa jokin hyvä koodi jolla on samat koodipituudet, eikä, että kyseinen koodi on hyvä. (Voit aina ottaa hyvän koodin, ja vaihtaa kaikki nollat yksiköiksi. Koodisanojen pituus ei muutu, mutta Kraftin-MacMillanin epäyhtälö pätee edelleen.)

Todistetaan väite näyttämällä, että jos koodi ei ole yksikäsitteisesti purettavissa, niin se ei ole myöskään etuliitevapaa. Olkoon siis $c: \Omega \rightarrow \mathcal{M}_{0,1}$ koodi, joka ei ole yksikäsitteisesti purettavissa. Tällöin on olemassa merkkijonot $S, T \in \mathcal{M}_\Omega$ joille pätee $C(S) = C(T)$. Merkitään $S = s_1 s_2 \cdots s_k$, $T = t_1 t_2 \cdots t_n$, ja $C(S) = C(T) = b_1 b_2 \cdots b_m$.

Väitämme ensin, että sanat S ja T voidaan valita siten, että $s_1 \neq t_1$. Voidaan tehdä (mm.) seuraavalla kahdella eri tavalla.

Vähemmän formaalisti Jos väite ei päde, eli $s_1 = t_1$, voidaan nämä merkit poistaa merkkijonojen S ja T alusta. Saadaan merkkijonot S' ja T' joilla edelleen on ominaisuus $C(S') = C(T')$. Jos ominaisuus pätee näillä uusilla merkkijonoilla, niin olemme valmiit, ja jos ei, niin voimme toistaa prosessia. Merkkijonot ovat äärellisen mittaisia, joten prosessi välttämättä pysähtyy äärellisen monen askeleen jälkeen.

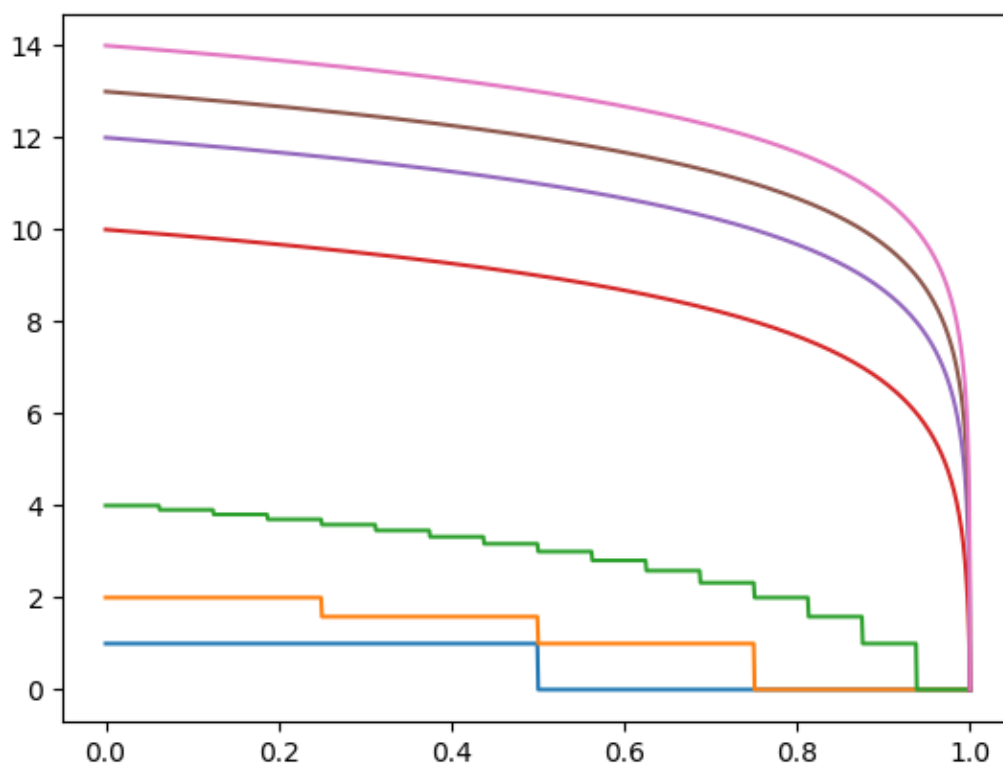
Enemmän formaalisti Huomaa, että jos koodaus vie kaksi eri symbolia a, b samalle merkkijonolle, eli $c(a) = c(b)$ niin koodi ei ole tällöin etuliitevapaa, koska tällöin $c(a) = c(b)\epsilon$. Myöskin, jos koodi vie jonkin symbolin d tyhjälle merkkijonolle (aika outo koodi) niin koodi ei ole etuliitevapaa koska $c(a) = c(d)c(a)$ kaikilla $a \in \Omega \setminus \{d\}$.

Koska koodi c ei ole yksikäsitteisesti purettavissa, löytyy jotkin eri merkkijonot jotka tuottavat saman koodin $b_1 b_2 \cdots b_m$. Katsotaan kaikkien tällaisten merkkijonoparien joukkoa, ja valitaan sieltä merkkijonopari jotka tuottavat tässä joukossa pituudeltaan kaikista lyhimmän koodin. Nyt jos $s_1 = t_1$, niin voimme poistaa koodista $b_1 b_2 \cdots b_m$ merkkiä s_1 vastaavan alkupätkän $c(s_1)$. Koska aiemmin huomasimme, että etuliitevapaassa koodissa mikään merkki ei voi kuvautua tyhjälle merkkijonolle, tämä alkupätkän poisto välttämättä lyhentää output-koodia, mikä on ristiriitä minimaalisen valinnan kanssa.

Nyt kun tiedämme, että $C(S) = C(T)$ ja $s_1 \neq t_1$, niin huomamme että joko $\ell(c(s_1)) \geq \ell(c(t_1))$ jolloin $c(t_1)$ on $c(s_1)$:n etuliite, tai $\ell(c(s_1)) \leq \ell(c(t_1))$ jolloin $c(s_1)$ on $c(t_1)$:n etuliite. Nähdään siis, että ei-yksikäsitteisesti purettavissa oleva koodi ei voi olla etuliitevapaa, mikä todistaa väitteen.

Syventävät tehtävät

Seuraavat tehtävät palautetaan kirjallisina harjoitusten pitäjälle, kurssin palautuslokeroon tai sähköpostitse. Deadline 03.12.2019 klo 08.30.



Kuva 3: Tehtävän 4 kuvaajia

4. Tutkitaan kahden alkion tasajakaumaa, $X = \{(0, 1), (\frac{1}{2}, \frac{1}{2})\}$.

Hahmottele funktion $\delta \mapsto H_\delta(X^N)$ kuvaajaa kun $N = 1, 2, 4, 1000$.

Vihje: Normaali-jakauma. Ratkaisu:

Oheisessa kuvassa on hahmoteltu kuvaajaa kun $N = 1, 2, 4, 10, 12, 13, 14$. Kuvaaja kun $N = 1000$ olisi näyttää suunnilleen vaakasuoralta viivalta korkeudella 1000. (Muista, että Shannonin lauseen idealla kuvaajan painotettuna $1/N$:llä on suunnilleen vaakasuora viiva alkuperäisen jakauman entropian $H = 1$ korkeudella.)

5. Olkoon X TN-jakauma ja merkitään tämän jakauman Huffmanin koodia c_H . Osoita, että c_H on optimaalinen etuliitevapaiden täydellisten koodien joukossa, eli että kaikilla muilla etuliitevapailta koodilla c joilla $\sum_j 2^{-\ell_j} = 1$ pätee

$$L(c_H, X) \leq L(c, X).$$

(*Vihje: Jos on olemassa koodi jolla on pienempi odotuspituus, niin löytyy optimaalinen koodi jolla on pienempi odotuspituus. Tutki askelta, jossa Huffman asettaa kahdelle vähiten todennäköiselle alkioille a ja b yhtä pitkät koodit. Oleta, että c antaa näille eri mittaiset koodit $\ell(c(a)) < \ell(c(b))$ ja etsi alkio x jolle $\ell(c(x)) \geq \ell(c(b))$. Katso mitä koodin c odotuspituudelle käy kun vaihdat koodit $c(b)$ ja $c(x)$ keskenään.*) *Ratkaisu:*

MacKayn kirja, s. 105, "Solution to exercise 5.16".

6. Olkoon X TN-jakauma missä $\Omega = \{a_1, \dots, a_n\}$, $p_j = \frac{1}{j}$ kaikilla j ja oletetaan että n ei ole muotoa 2^k millään $k \in \mathbb{N}$. Olkoon c optimaalinen koodi (eli Huffmanin koodi) jakaumalle X .

Merkitään $l^+ = \lceil \log_2(n) \rceil$, ja asetetaan

$$f^+ := \frac{\#\{x \in \Omega \mid \ell(x) = l^+\}}{n}.$$

- (a) Osoita, että $f^+ = 2 - (2^{l^+})/n$ ja $L(c, X) = l^+ - 1 + f^+$.
 (b) Merkitään edelleen $\Delta L := L(c, X) - H(X)$. Osoita että

$$\Delta L \leq 1 - \frac{\ln(\ln 2)}{\ln 2} - \frac{1}{\ln(2)} \approx 0.086.$$

(Vihje: a-kohdan nojalla voit kirjoittaa muutoksen ΔL parametrin n muuttujana. Korvaa n jatkuvalla muuttujalla x ja tutki derivaattaa.)

- (c) Laske ΔL tilanteessa, jossa X on yhdentoista alkion tasajakauma ja c sen Huffmanin koodi.

Ratkaisu: Tämä tehtävä oli aika lailla raskaampi kuin mitä sen piti olla, sry!

Malliratkaisu jätetään toistaiseksi väliin aikataulupaineiden johdosta. Pahoittelut tästäkin; näin työläitä tehtäviä ei tule tenttiin.