Tehtävä 4

Tutkitaan kahden alkion tasajakaumaa, $X = \{(0,1), (\frac{1}{2}, \frac{1}{2})\}$.
Hahmottele funktion $\delta \mapsto H_\delta(X^N)$ kuvaaja kun $N = 1, 2, 4, 1000$.
Vihje: Normaalijakauma.

Ratkaisu:
Koska X on tasajakauma, kaikilla alkeistapahtumille $x \in \Omega_{X^N}$ pätee $P(x) = (\frac{1}{2})^N$. Olkoon nyt $\delta \in]0,1[$ ja $S \in \mathcal{P}(\Omega_{X^N})$. Tällöin

$$\sum_{x \in S} P(x) = \sum_{i=1}^{|S|} P(x_i) = \frac{|S|}{2^N} \geq 1 - \delta$$

toteutuu jos ja vain jos

$$|S| \geq 2^N(1 - \delta).$$

Tästä saadaan, että $|S_\delta| = \lceil 2^N(1-\delta) \rceil$. Bittisisällöksi tulee suoraan määritelmästä

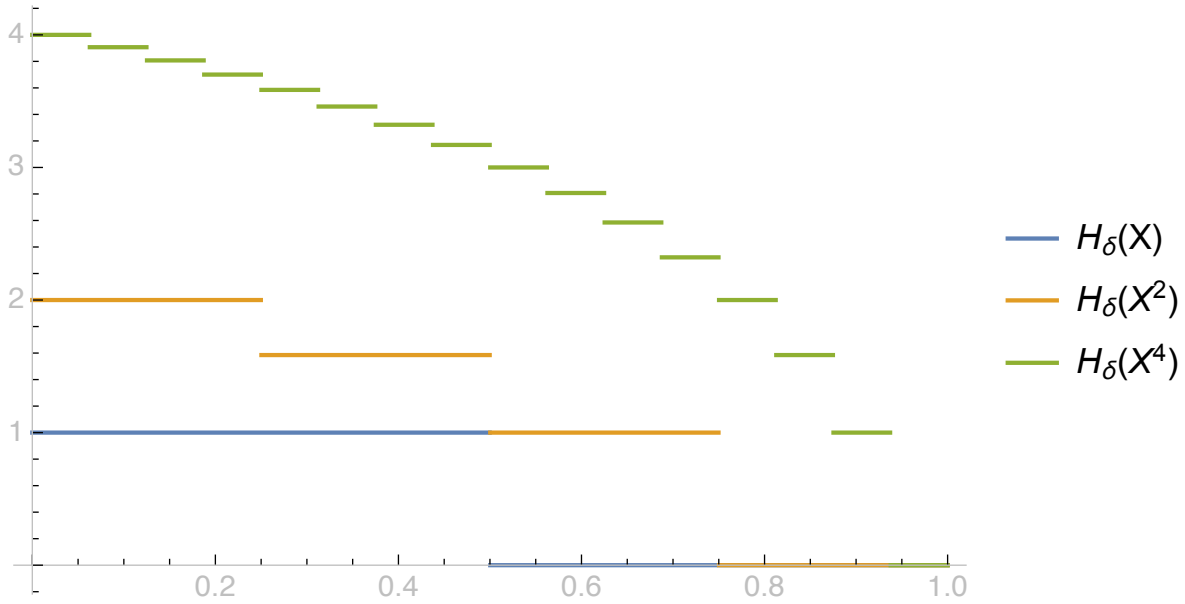
$$H_\delta(X^N) = \log_2(|S_\delta|) = \log_2(\lceil 2^N(1-\delta) \rceil).$$

Valitettavasti kattofunktion $\lceil \cdot \rceil$ käsittely on melko epämiellyttävää. Päättelemällä nähdään kuitenkin, että $|S_\delta|$ riippuu N:stä siten, että $|S_\delta|$ on ensimmäisellä välillä 2^N, toisella välillä $2^N - 1$ ja niin edelleen. Välit puolestaan ovat kaikki $1/N$:n pituisia. Näin ollen $\delta \mapsto |S_\delta|$ on alaspäin tuleva portaikko, jossa portaiden korkeus on verrannollinen 2^N ja portaiden väli on verrannollinen $1/N$. Ottamalla näistä logaritmi saadaan, että $\delta \mapsto H_\delta(X^N)$ on loppua kohden nopeammin laskeva portaikko samoilla väleillä. Tietokoneella piirretyt kuvaajat on kuvassa 1.

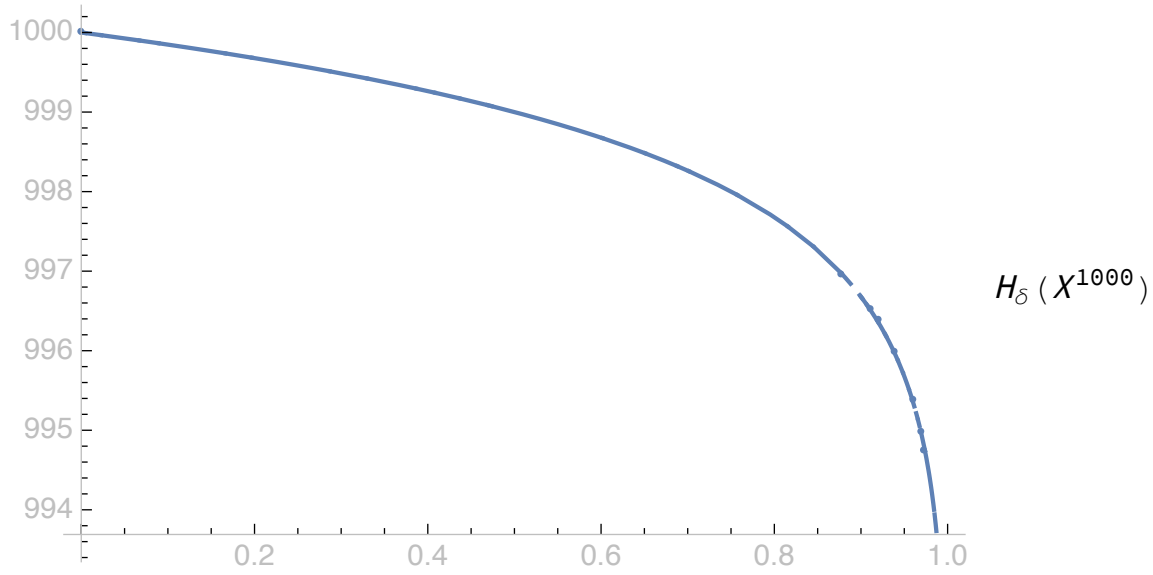
Tarkastellaan vielä tapausta $N = 1000$ lähemmin. Logaritmin takia kuvauksen $\delta \mapsto H_\delta(X^{1000})$ korkeus on enimmillään suunnilleen 1000 ja pienimmillään $\sim 1000 - \log_2(1000) = 990$. Jos siis tarkastellaankin skaalattua versiota $\delta \mapsto \frac{1}{N} H_\delta(X^N)$, niin kuvaaja on lähes vakiokuvaus $\delta \mapsto 1$. Tämä käy järkeen, sillä Shannonin lähdekoodauslauseen nojalla

$$\lim_{N \to \infty} \frac{1}{N} H_\delta(X^N) = H(X)$$

ja tässä tapauksessa $H(X) = 1$. Tämä tilanne on esitetty kuvassa 2.



Kuva 1: Kuvaus $\delta \mapsto H_\delta(X^N)$ kun $N = 1, 2, 4$.



Kuva 2: Kuvaus $\delta \mapsto H_\delta(X^N)$ kun $N = 1000$. Huomattavaa on, että käyrän jyrkkyys johtuu pääasiassa pystyakselin skaalasta. Jos pystyakseli alkaisi nollasta, tuloksena olisi käytännössä suora viiva.

Tehtävä 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 etuliitevapailla koodeilla c joilla $\sum_i 2^{-\ell_i} = 1$ 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 alkiolle 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 $\ell(c(c)) \geq \ell(c(b))$. Katso mitä koodin c odotuspituudelle käy kun vaihdat koodit $c(b)$ ja $c(x)$ keskenään.)

Ratkaisu:

Tehdään vastaoletus, että on olemassa jokin optimaalinen, etuliitevapaa ja täydellinen koodi c, joka ei ole Huffmanin koodi. Tarkastellaan Huffmanin algoritmin vaihetta, jossa yhdistetään kaksi vähiten todennäköistä symbolia $a \in X$ ja $b \in X$. Koodaus c antaa tällöin näille antiteesin nojalla eripituiset koodit. Voidaan olettaa yleisyyttä menettämättä olettaa, että $\ell(c(a)) < \ell(c(b))$. Koska alkioiden a ja b todennäköisyydet ovat pienimmät ja c on täydellinen koodi, on olemassa jokin alkio $x \in X$ siten, että $P(x) > P(a)$ ja $\ell(c(x)) \geq \ell(c(b))$. Muodostetaan sitten uusi koodi c', jossa on vaihdettu koodit $c(a)$ ja $c(x)$ keskenään. Tällöin

$$\begin{aligned}
L(c', X) - L(c, X) &= P(x)\ell(c(a)) + P(a)\ell(c(x)) - P(x)\ell(c(x)) - P(a)\ell(c(a)) \\
&= P(x)[\ell(c(a)) - \ell(c(x))] + P(a)[\ell(c(x)) - \ell(c(a))] \\
&= \underbrace{[P(x) - P(a)]}_{>0}\underbrace{[\ell(c(a)) - \ell(c(x))]}_{<0} \\
&< 0.
\end{aligned}$$

Löydettiin siis koodi c', jonka odotuspituus $L(c', X)$ on pienempi kuin koodin c. Antiteesin mukaan c on optimaalinen koodi, joten tämä on ristiriita. Siispä optimaalisen koodin saa vain yhdistämällä kaksi vähiten todennäköistä alkiota eli Huffmanin koodi on optimaalinen.

Tehtävä 6

Olkoon X TN-jakauma missä $\Omega = \{a_1, \ldots, a_n\}$, $p_j = \frac{1}{n}$ 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

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

(a) Osoita, että $f^+ = 2 - (2^{l^+})/n$ ja $L(c, X) = l^+ - 1 + f^+$.

(b) Merkitään edelleen $\Delta L \coloneqq L(c, X) - H(X)$. 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:

(a) Huffmanin koodissa peräkkäisten koodien pituudet eroavat toisistaan korkeintaan yhden bitin verran. Jos $n = 2^b$ jollakin $b \in \mathbb{N}$, niin tällöin $\ell(x) = b$ kaikilla $x \in \Omega$. Muodostetaan nyt uusi joukko $\Omega' = \Omega \cup \{\omega\}$ ja tarkastellaan tämän uuden joukon koodien pituuksia. Uudelle symbolille $\ell(\omega) = b+1$. Lisäksi yksi vanha symboli a joudutaan korvaamaan uudella koodilla, jolloin $\ell(a) = b+1$. Tällöin $b+1$ pituisia koodeja on 2. Toistamalla tätä prosessia ja lisäämällä uusia symboleja nähdään, että $b+1$ pituisia koodeja on $2(n - 2^b)$. Koska $b + 1 = \lceil \log_2(n) \rceil$, niin

$$nf^+ = 2(n - 2^b) = 2n - 2^{b+1} = 2n - 2^{l^+},$$

jolloin jakamalla yhtälö puolittain n:llä saadaan haluttu tulos $f^+ = 2 - 2^{l^+}/n$.

Koodin odotuspituuden määritelmästä saadaan nyt edellisten havaintojen avulla

$$\begin{aligned}
L(c, X) &= \sum_{x \in \Omega} P(x)\ell(x) \\
&= \frac{1}{n} \underbrace{\left(2n - 2^{l^+}\right)}_{l^+ \text{ pitkät koodit}} l^+ + \frac{1}{n} \underbrace{\left(n - \left(2n - 2^{l^+}\right)\right)}_{l^+ - 1 \text{ pitkät koodit}} (l^+ - 1) \\
&= 2l^+ - \frac{2^{l^+} l^+}{n} + \frac{2^{l^+} l^+}{n} - \frac{2^{l^+}}{n} - l^+ + 1 \\
&= l^+ + 2 - \frac{2^{l^+}}{n} - 1 \\
&= l^+ - 1 + f^+
\end{aligned}$$

kuten haluttiinkin.

(b) Tasajakauman entropia on vakio

$$H(X) = -\sum_{x\in\Omega} P(x)\log_2(P(x)) = \sum_{i=1}^{n} \frac{1}{n}\log_2(n) = \log_2(n).$$

Kirjoitetaan sitten n muotoon $n = 2^m + d$ joillakin $m, d \in \mathbb{N}$, missä $2^m \leq d < 2^{m+1}$. Todetaan lisäksi, että $l^+ = \lceil \log_2(n) \rceil = \lfloor \log_2(n) \rfloor + 1$. Nyt a-kohdan perusteella

$$L(c, X) = l^+ - 1 + f^+ = \lceil \log_2(n) \rceil - 1 + f^+ = \lfloor \log_2(n) \rfloor + f^+.$$

Tällöin $\lfloor \log_2(n) \rfloor = \lfloor \log_2(2^m + d) \rfloor = m$, joten a-kohdan mukaan

$$\begin{aligned} L(c, X) &= \lfloor \log_2(n) \rfloor + 2 - \frac{2^{l^+}}{n} \\ &= m + 2 - \frac{2^{m+1}}{n}. \end{aligned}$$

Nämä yhdistämällä saadaan, että

$$\begin{aligned} \Delta L &= L(c, X) - H(X) \\ &= m + 2 - \frac{2^{m+1}}{n} - \log_2(n). \end{aligned}$$

Korvataan $n \in \mathbb{N}$ jatkuvalla muuttujalla $x \in \mathbb{R}$, jolloin derivoimalla ΔL:n lauseke x:n suhteen saadaan

$$\frac{\partial \Delta L}{\partial x} = \frac{2^{m+1}}{x^2} - \frac{1}{x \ln 2}.$$

Derivaatta $\partial_x \Delta L$ saavuttaa nollakohdan kun

$$x_0 = 2^{m+1} \ln 2.$$

Lisäksi $\partial_{xx} \Delta L(x_0) < 0$, joten x_0 on maksimikohta. Siispä pätee

$$\begin{aligned} \Delta L &\leq \Delta L(x_0) \\ &= m + 2 - \frac{2^{m+1}}{2^{m+1} \ln 2} - \log_2\left(2^{m+1} \ln 2\right) \\ &= m + 2 - \frac{1}{\ln 2} - m - 1 - \log_2(\ln 2) \\ &= 1 - \frac{\ln(\ln 2)}{\ln 2} - \frac{1}{\ln 2} \approx 0.0861. \end{aligned}$$

ja saatiin väite, sillä $\mathbb{N} \subset \mathbb{R}$.

(c) Kun $n = 11$, niin $l^+ = \lceil \log_2(11) \rceil = 4$, jolloin saadaan a- ja b-kohtien avulla

$$\begin{aligned} \Delta L &= l^+ + 1 - \frac{2^{l^+}}{n} - \log_2(n) \\ &= 4 + 1 - \frac{2^4}{11} - \log_2(11) \\ &= \frac{39}{11} - \log_2(11) \approx 0.08602 < 0.0861. \end{aligned}$$