Eksempel på entropi.

Teksten 'Then the hen began to eat'
ble med Huffman koding kodet med 75 bits.
Teksten består av 25 tegn, altså 3 bits/tegn.
Hva er entropien for denne teksten?
(egentlig for alfabetet).

Hvis $P_i = P(\alpha_i)$ – sannsynligheten til α_i
er entropien definert som

$$H(P_1, \ldots, P_n) = -\sum_{i=1}^{n} P(\alpha_i) \log_2 P(\alpha_i)$$

$$2^{\log_2 x} = x$$

$$\log_2 x = \frac{\ln x}{\ln 2}$$

For vår tekst får vi

$H(P_1, \ldots, P_n) = 2.93$

Det betyr at vi i beste fall kunne
klare oss $2.93 \cdot 25$ bits, 74 bits

Aritmetisk koding

Ex. Anta at $A = \{0,1\}$, $P(0) = 0.9$, $P(1) = 0.1$

Huffmankoding gir $C(0) = 0$, $C(1) = 1$ så vi trenger 1 bit pr. tegn.

Entropi: $H(P_1, P_2) = -\left(P(0)\log_2 P(0) + P(1)\log_2 P(1)\right)$

$= 0.47$

Idéen bak aritmetisk koding.

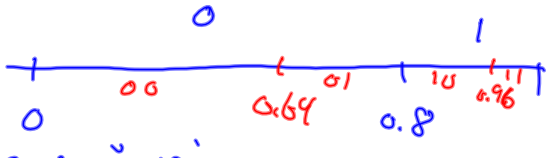
Aritmetisk koding assosierer ulike tekster med ulike delintervall av $[0,1)$. Bredden til delintervallet er lik sannsynligheten til teksten og den aritmetiske koden er ett tall i intervallet, representert i 2-tallsystemet.

Eksempel på aritmetisk koding.

La $X = 00100$, da er $P(0) = 0.8$, $P(1) = 0.2$

Vi begynner med å dele $[0,1]$ i forhold til sannsynlighetene

Siden første tegn er 0 ligger koden i $[0, 0.8)$



For å komme videre må vi dele dette intervallet på samme måte, i $[0, 0.64)$ og $[0.64, 0.8)$. Siden tegn nr. 2 er 0 havner vi i $[0, 0.64)$.

Dette må så deles på samme måte:

i $[0, 0.512)$ og $[0.512, 0.64)$.

Siden tegn 3 er 1 havner vi i $[0.512, 0.64)$.

Dette må deles i $[0.512, 0.6144)$, $[0.6144, 0.64)$

Tegn nr. 4 er 0 så vi havner i $[0.512, 0.6144)$. Dette deles i $[0.512, 0.59392)$ og $[0.59392, 0.6144)$

Tegn 5 er 0 så vi ender til slutt i $[0.512, 0.59392)$.

Vi ønsker et tall i 2-tallsystemet i dette intervallet. Tallet i 2-tallssystemet med minst mulig sifre som ligger i dette intervallet er $9/16$

$\frac{9}{16} = 0.5625 = 0.1001_2 \quad (\frac{1}{2} + \frac{1}{16})$

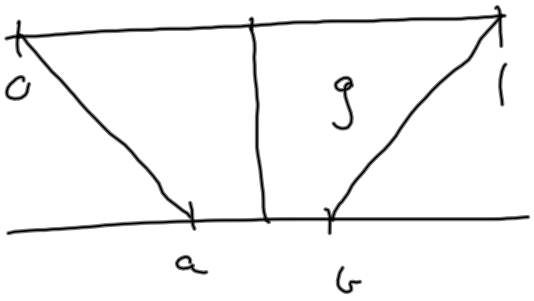
Aritmetisk kode: 1001

Lineære funksjoner:

La $[a,b]$ være gitt med $a<b$.

Funksjonen $g(z) = a + z(b-a)$ avbilder $[0,1]$ på $[a,b]$. Spesielt

$g(0) = a$, $g(\frac{1}{2}) = \frac{a+b}{2}$, $g(1) = b$.



Kumulative sannsynlighetsfordelinger:

La $A = \{\alpha_1, \dots, \alpha_n\}$ være et alfabet med sannsynligheter $P_i = P(\alpha_i)$.

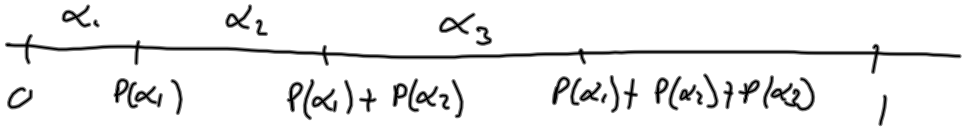
Da definerer vi

$$F(\alpha_j) = \sum_{i=1}^{j} P(\alpha_i), \quad j = 1,2,\dots,n.$$

$$L(\alpha_j) = F(\alpha_j) - P(\alpha_j) = F(\alpha_{j-1}), \quad j = 2,3,\dots,n.$$

Med n symboler får første oppdeling av $[0,1]$

$[0, F(\alpha_1))$, $[F(\alpha_1), F(\alpha_2))$, $[F(\alpha_2), F(\alpha_3))$ - - - -



α_i har intervallet $[L(\alpha_i), F(\alpha_i))$

Generell algoritme:

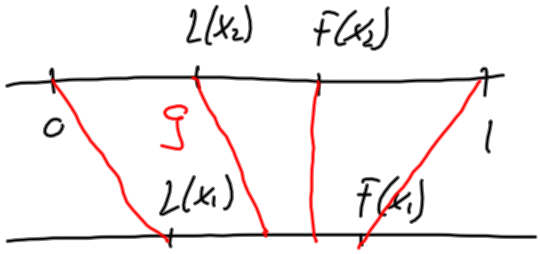
Anta at teksten er $X = \{x_1, x_2, \ldots, x_m\}$

Første tegn er x_1. Dette svarer til intervallet $[L(x_1), F(x_1))$.

Neste symbol er x_2. Nå skal $[L(x_1), F(x_1))$ deles som $[0, 1)$.

Så intervallet $[L(x_2), F(x_2))$ må plasseres riktig i $[L(x_1), F(x_1))$.

Må lage g som sender $[0,1)$ på $[L(x_1), F(x_1))$. Da vil intervallet $[L(x_2), F(x_2))$ automatisk bli sendt til riktig delintervall av $[L(x_1), F(x_1))$ av g



Algoritme:

1. Sett $[a_0, b_0] = [0,1)$.
2. For $k = 1, \ldots, m$
   (a) Definer $q_k = a_{k-1} + z(b_{k-1} - a_{k-1})$
   (b) Sett $[a_k, b_k) = [q_k(L(x_k)), q_k(F(x_k)))$

Den aritmetiske kodeen er midtpunktet til $[a_m, b_m)$ trunkert til

$$\lceil -\log_2[P(x_1)P(x_2)\cdots P(x_m)] \rceil + 1$$

binære siffer. $\lceil\ \rceil$ - tak funksjonen