INFORMATSIOONITEOORA

Loengukonspekt ja ülesanded

kevad 2018

Jüri Lember

Kirjandus:

1. **T.M. Cover, J.A. Thomas "Elements of information theory"**, Wiley, 1991 ja 2006;
2. Yeung, Raymond W. "A first course of information theory", Kluwer, 2002;
3. Te Sun Han, Kingo Kobayashi "Mathematics of information and coding", AMS, 1994;
4. Csiszar, I., Shields, P. "Information theory and statistics : a tutorial", MA 2004;
5. Mackay, D. "Information theory, inference and learning algorithms", Cambridge 2004; http://www.inference.phy.cam.ac.uk/itprnn/book.pdf
6. McEliece, R. "Information and coding", Cambridge 2004;
7. Gray, R. "Entropy and information theory", Springer 1990;
8. Gray, R. "Entropy and information theory", Springer 1990;
9. Gray, R. "Source coding theory", Kluwer, 1990;
10. Shields, P. "The ergodic theory of discrete sample paths", AMS 1996;
11. Dembo, A., Zeitouni, O. "Large deviation techniques and Applications", Springer 2010.
12. $\cdots$

Konspekt: https://courses.ms.ut.ee/2018

1 Entroopia ja informatsioon

1.1 Entroopia

1.1.1 Definitsioon ja omadused

Vaatleme diskreetset juhuslikku suurust X jaotusega P. Olgu $\mathcal{X} = \{x_1, x_2, \ldots\}$ ülimalt loenduv hulk, mis sisaldab juhusliku suuruse $\mathcal{X}$ võimalikke väärtusi. Tähistame

$$p_i := \mathbf{P}(X = x_i) = P(x_i),$$

s.t. p_i on tõenäosus, et X võtab väärtuse x_i. Jaotus P on üheselt määratud paaridega $\{(x_i, p_i)\}$, sest iga hulga $A \subset \mathcal{X}$ korral

$$P(A) = \mathbf{P}(X \in A) = \sum_{i: x_i \in A} p_i = \sum_{x \in A} P(x).$$

Tihti esitatakse selline jaotus tabelina

$$\begin{array}{c|c|c|c} x_1 & x_2 & x_3 & \ldots \\ \hline p_1 & p_2 & p_3 & \ldots \end{array},$$

kusjuures $x_i \neq x_j$, kui $i \neq j$ ja $p_i \geq 0$. Edaspidi ütleme, et jaotus (tõenäosusmõõt) P on antud hulgal $\mathcal{X}$. Paneme tähele, et $\mathcal{X}$ võib olla suvaline hulk, mitte ilmtingimata reaalarvude alamhulk. Näiteks võib hulk $\mathcal{X}$ olla tähestik, s.t. $\mathcal{X} = \{a, b, \ldots, y\}$. Sellisel juhul on X juhuslik täht. Informatsiooniteoorias nimetataksegi hulka $\mathcal{X}$ tihti *tähestikuks (alphabet)*.

Jaotuse P *kandja (support)* $\mathcal{X}_P$ on tähed, mille korral $P(x) > 0$. Seega

$$\mathcal{X}_P := \{x : P(x) > 0\}.$$

Tuletame meelde, et kui $g : \mathcal{X} \to \mathbb{R}$ on suvaline funktsioon, mis rahuldab tingimust $\sum p_i |g(x_i)| < \infty$, siis

$$Eg(X) = \sum_i p_i g(x_i) = \sum_{x \in \mathcal{X}} P(x) g(x) = \sum_{x \in \mathcal{X}_P} P(x) g(x) \tag{1.1}$$

NB! Alljärgnevas tähistame $\log := \log_2$ ning lepime kokku, et $0 \log 0 = 0$.

Def 1.1 *Juhusliku suuruse X (jaotuse P)* **entroopia (entropy)** *$H(X)$ on*

$$H(X) = -\sum p_i \log p_i = -\sum_{x \in \mathcal{X}} P(x) \log P(x) = -\sum_{x \in \mathcal{X}_P} P(x) \log P(x).$$

Märkused:

- $H(X)$ sõltub vaid juhusliku suuruse X jaotusest P. Seetõttu tähistame entroopiat $H(X)$ ka $H(P)$.

- Seose (1.1) tõttu
$$H(X) = E\big(-\log P(X)\big) = E \log \frac{1}{P(X)}.$$

- Et $-\log p_i \geq 0$, on $\sum -p_i \log p_i$ mittenegatiivsete liikmetega rida. Sellise rea summa on alati defineeritud, kuid võib olla lõpmatu. Seega
$$0 \leq H(X) \leq \infty,$$
kusjuures $H(X) = 0$ parajasti siis, kui X on peaaegu kindlasti konstant.

- Entroopia ei sõltu tähestikust $\mathcal{X}$. Tõepoolest, olgu jaotused P ja Q antud tabelitega
$$P : \begin{array}{c|c|c|c} x_1 & x_2 & x_3 & \dots \\ \hline p_1 & p_2 & p_3 & \dots \end{array} \qquad Q : \begin{array}{c|c|c|c} y_1 & y_2 & y_3 & \dots \\ \hline p_1 & p_2 & p_3 & \dots \end{array}$$
Siis $H(P) = H(Q)$. Et oluline on vaid tõenäosuste vektor $(p_1, p_2, \dots)$, kasutame tihti tähistust
$$H(p_1, p_2, \dots).$$

- Põhimõtteliselt võib entroopia defineerida ka mõne muu logaritmi abil. Logaritmi $\log_b$ abil defineeritud entroopiat tähistame H_b. Seega
$$H_b(X) = -\sum p_i \log_b p_i = -\sum_{x \in \mathcal{X}} P(x) \log_b P(x).$$
Et $\log_b p = \log_b a \log_a p$, siis
$$H_b(X) = (\log_b a) H_a(X),$$
millest $H_b(X) = (\log_b 2) H(X)$ ning $H_e(X) = (\ln 2) H(X)$. Informatsiooniteoorias kasutatakse harilikult kahendlogaritmi abil defineeritud entroopiat. Seda mõõdetakse *bittides*. Naturaallogaritmi kaudu defineeritud entroopiat mõõdetakse *nattides*, kümnendlogaritmi kaudu defineerituid entroopiat mõõdetakse *dittides*.

- Jaotuse P entroopia ei muutu, kui hulka $\mathcal{X}$ laiendada elementidega, mille tõenäosus on 0. Seega, kui $\mathcal{X}'$ on suvaline hulk, mis sisaldab hulka $\mathcal{X}$, siis kehtib
$$H(X) = -\sum_{x \in \mathcal{X}'} P(x) \log P(x). \tag{1.2}$$

Entroopia $H(X)$ mõõdab juhusliku suuruse X "keskmist juhuslikkust". Mida suurem on entroopia, seda "juhuslikum" on X. Konstant ei ole juhuslik, seetõttu on konstandi entroopia 0. Entroopiat võib ka interpreteerida kui informatsioonihulka, mida juhusliku suuruse väärtuse teadasaamine meile annab. Mida "juhuslikum" on X, seda vähem oskame me ära arvata juhusliku suuruse väärtust (juhusliku katse tulemust) ning seda enam informatsiooni selle väärtuse (katse tulemuse) teadasaamine meile annab.

Esmakordselt defineeris entroopia ameerika matemaatik C. Shannon oma 1948.-l aastal ilmunud teedrajavas artiklis "A mathematical theory of communacation". Seetõttu nimetatkse entroopiat tihti ka Shannoni entroopiaks.

Näited:

1 Olgu $\mathcal{X} = \{0, 1\}$, $p = \mathbf{P}(X = 1)$. Seega on X Bernoulli p-jaotusega juhuslik suurus, $X \sim B(1, p)$. Leiame

$$H(X) = -p \log p - (1-p) \log(1-p) =: h(p).$$

Funktsiooni $h(p)$ nimetatakse **binaarseks entroopiafunktsiooniks**. Funktsioon $h(p)$ on nõgus, punkti $\frac{1}{2}$ suhtes sümmeetriline ning saavutab maksimumi juhul, kui $p = \frac{1}{2}$. Siis

$$h(\frac{1}{2}) = -\frac{1}{2} \log \frac{1}{2} - \frac{1}{2} \log \frac{1}{2} = \log 2 = 1.$$

Seega on (nihketa) mündi viske entroopia 1. Teadmine, kas sellise mündi viskel tuli kull või kiri, annab meile täpselt 1 biti informatsiooni (sellest tulenevalt ongi entroopia defineerimisel võetud aluseks kahendlogaritm). Kui kulli tulemise tõenäosus p on väiksem arvust $\frac{1}{2}$, siis on entroopia väiksem kui 1. See ühtib intuitsiooniga: mida väiksem on kulli tulemise tõenäosus, seda "mittejuhuslikum" on X ning seda "kergem" on mündiviske tulemust ära arvata. Sellevõrra vähem informatsiooni mündivise endas kätkeb.

2 Vaatleme jaotusi

$$P: \begin{array}{c|c|c|c|c} a & b & c & d & e \\ \hline \frac{1}{2} & \frac{1}{4} & \frac{1}{8} & \frac{1}{16} & \frac{1}{16} \end{array} \qquad Q: \begin{array}{c|c|c|c} a & b & c & d \\ \hline \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{array}.$$

Leiame

$$H(P) = -\frac{1}{2} \log \frac{1}{2} - \frac{1}{4} \log \frac{1}{4} - \frac{1}{8} \log \frac{1}{8} - \frac{1}{16} \log \frac{1}{16} - \frac{1}{16} \log \frac{1}{16} = \frac{1}{2} + \frac{2}{4} + \frac{3}{8} + \frac{4}{16} + \frac{4}{16} = \frac{15}{8}$$
$$H(Q) = \log 4 = 2.$$

Seega on jaotus P "vähem juhuslik", kuigi tema aatomite arv on suurem.

1.1.2 Entroopia aksiomaatiline definitsioon

On kerge veenduda, et entroopial on nn. *grupeerimisomadus*

$$H(p_1,p_2,p_3,\ldots)=H(\Sigma_{i=1}^k p_i,p_{k+1},p_{k+2},\ldots)+\big(\Sigma_{i=1}^k p_i\big)H\Big(\frac{p_1}{\Sigma_{i=1}^k p_i},\ldots,\frac{p_k}{\Sigma_{i=1}^k p_i}\Big) \quad (1.3)$$

Omaduse (1.3) tõestus on ülesanne 2.

Grupeerimisomadus on teatavas mõttes igati loomulik juhuslikkuse "aditiivsuse" omadus, mistõttu on loogiline eeldada, et iga funktsioon $f(p_1,p_2,\ldots)$, mis mõõdab juhuslikkust, peaks seda omadust rahuldama. Selgub aga, et kui $\mathcal{X}$ on lõplik, siis f mis rahuldab grupeerimisomadust ning in lisaks pidev, sümmeetriline ja normeeritud (igati loomulikud eeldused) saab olla ainult entroopia.

Sõnastame selle väitena. Lõpliku $\mathcal{X}$ korral on iga tõenäosusmõõt vektor $(p_1,\ldots,p_m)$, kus $|\mathcal{X}|=m$, $p_i\ge 0$ ja $\sum_{i=1} p_i=1$. Olgu selliste vektorite hulk $\mathcal{P}^m$, seda hulka nimetatakse ($(m-1)$-dimensionaalseks) *simpleksiks*. Funktsioon $f_m:\mathcal{P}^m\to\mathbb{R}$ on pidev parajasti siis, kui ta on pidev kõikide argumentide järgi. Funktsiooni f_m nimetame sümmeetriliseks, kui $f_m(p_1,\ldots,p_m)$ ei sõltu argumentide järjekorrast.

Väide 1.1 *Olgu iga m korral $f_m:\mathcal{P}^m\to[0,\infty)$ sümmeetrilised funktsioonid, mis rahuldavad järgmisi omadusi (aksioome):*

A1 *f_2 on normaliseeritud, st $f_2(\frac12,\frac12)=1$;*

A2 *f_m on pidev iga $m=2,3,\ldots$ korral;*

A3 *kehtib grupeerimisomadus: iga $1<k<m$ korral*

$$f_m(p_1,p_2,\ldots,p_m)=f_{m-k+1}(\Sigma_{i=1}^k p_i,p_{k+1},\ldots,p_m)+\big(\Sigma_{i=1}^k p_i\big)f_k\Big(\frac{p_1}{\Sigma_{i=1}^k p_i},\ldots,\frac{p_k}{\Sigma_{i=1}^k p_i}\Big).$$

A4 *iga $m<n$ korral $f_m(\frac1m,\ldots,\frac1m)\le f_n(\frac1n,\ldots,\frac1n)$.*

Siis iga m korral

$$f_m(p_1,\ldots,p_m)=-\sum_{i=1}^m p_i\log p_i. \quad (1.4)$$

Tõestus. Olgu iga m korral

$$g(m):=f_m(\frac1m,\ldots,\frac1m).$$

Grupeerimisomadust ja sümmeetriat m korda rakendades saame

$$\begin{aligned} g(mn)&=f_{nm}\Big(\underbrace{\frac{1}{nm},\ldots,\frac{1}{nm}}_{n},\ldots,\underbrace{\frac{1}{nm},\ldots,\frac{1}{nm}}_{n}\Big)\\ &=f_m(\frac1m\ldots,\frac1m)+f_n\big(\frac1n,\ldots,\frac1n\big)=g(m)+g(n).\end{aligned}$$

Seega iga täisarvu n ja k korral $g(n^k) = kg(n)$ ja **A1** tõttu $g(2^k) = kg(2) = k$ ehk

$$g(2^k) = \log(2^k), \quad \forall k.$$

Omadust **A4** kasutades on võimalik näidata, et ülaltoodud võrdus kehtib iga täisarvu n korral, ehk

$$g(n) = \log n, \quad \forall n \in \mathbb{N}.$$

Olgu nüüd m suvaline täisarv ja vaatleme vektorit $(p_1, \ldots, p_m)$, mille kõik komponendid on ratsionaalarvud. Seega leiduvad täisarvud $k_1, \ldots, k_m$ ja ühine nimetaja n nii, et $p_i = \frac{k_i}{n}$, $i = 1, \ldots, m$. Sellisel juhul

$$\begin{aligned} g(n) &= f_n\big(\underbrace{\frac{1}{n}, \ldots, \frac{1}{n}}_{k_1}, \underbrace{\frac{1}{n}, \ldots, \frac{1}{n}}_{k_2}, \ldots, \underbrace{\frac{1}{n}, \ldots, \frac{1}{n}}_{k_m}\big) \\ &= f_m(\frac{k_1}{n}, \ldots, \frac{k_m}{n}) + \sum_{i=1}^{m} \frac{k_i}{n} f_{k_i}(\frac{1}{k_i}, \ldots, \frac{1}{k_i}) \\ &= f_m(p_1, \ldots, p_m) + \sum_{i=1}^{m} \frac{k_i}{n} g(k_i) = f_m(p_1, \ldots, p_m) + \sum_{i=1}^{m} p_i \log(k_i). \end{aligned}$$

Seega

$$f_m(p_1, \ldots, p_m) = \log(n) - \sum_{i=1}^{m} p_i \log(k_i) = -\sum_{i=1}^{m} p_i \log(\frac{k_i}{n}) = -\sum_{i=1}^{m} p_i \log p_i$$

ehk ratsionaalarvuliste argumentide korral (1.4) kehtib. Et f_m on pidev, kehtib (1.4) suvaliste argumentide korral. ■

Märkus: Väide kehtib ka ilma aksioomita **A4**.

1.1.3 Entroopia on rangelt nõgus

Funktsioon $g : \mathbb{R} \to \mathbb{R}$ on *kumer*, kui iga x_1, x_2 ja $\lambda \in [0, 1]$ korral kehtib

$$g(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda g(x_1) + (1 - \lambda)g(x_2).$$

Funktsioon g on *rangelt kumer* kui võrdus kehtib vaid siis, kui $\lambda = 1$ või $\lambda = 0$. Funktsioon g on *nõgus*, kui $-g$ on kumer.

Jaotuste segu. Olgu P_1 ja P_2 kaks hulgal $\mathcal{X}$ antud jaotust. Eeldus, et P_1 ja P_2 on antud ühel ja samal hulgal pole üldisust kitsendav: kui P_1 on antud hulgal $\mathcal{X}_1$ ja P_2 on antud hulgal $\mathcal{X}_2$, siis defineerime $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$. Mõõtude P_1 ja P_2 *segu* on nende kumer kombinatsioon

$$Q = \lambda P_1 + (1 - \lambda)P_2, \quad \lambda \in (0, 1).$$

Kui $X_1 \sim P_1$ ja $X_2 \sim P_2$ ning $Z \sim B(1,\lambda)$, siis järgmine juhuslik suurus on jaotusega Q:

$$Y = \begin{cases} X_1 & \text{kui } Z = 1, \\ X_2 & \text{kui } Z = 0. \end{cases}$$

On selge, et segu Q kätkeb endas nii P_1 kui ka P_2 juhuslikkust. Lisaks on juhuslik komponendi valik (juhuslik suurus Z). Järgnev väide näitab, et $H(Q)$ on suurem kui $\lambda H(P_1) + (1-\lambda)H(P_2)$ ehk entroopia on nõgus.

Väide 1.2 *Entroopia on rangelt nõgus, s.t.*

$$H(Q) \geq \lambda H(P_1) + (1-\lambda)H(P_2),$$

kusjuures võrratus on range välja arvatud juhul, kui $P_1 = P_2$.

Tõestus. Funktsioon $f(y) = -y \log y$ on rangelt nõgus ($y \geq 0$). Seega iga $x \in \mathcal{X}$ korral

$$\begin{aligned} -\lambda P_1(x)\log P_1(x) - (1-\lambda)P_2(x)\log P_2(x) &= \lambda f\big(P_1(x)\big) + (1-\lambda) f\big(P_2(x)\big) \\ &\leq f\Big(\lambda P_1(x) + (1-\lambda)P_2(x)\Big) = -Q(x)\log Q(x). \end{aligned}$$

Summeerides mõlemad pooled üle $\mathcal{X}$, saame

$$\lambda H(P_1) + (1-\lambda)H(P_2) \leq H(Q).$$

Viimane võrratus on range, kui leidub vähemalt üks $x \in \mathcal{X}$ nii, et $P_1(x) \neq P_2(x)$. ■

Näide: Bernoulli p-jaotus $B(1,p)$ on konstantide 1 ja 0 kumer kombinatsioon. Entroopia nõgususest järeldub: $h(\lambda p_1 + (1-\lambda)p_2) \geq \lambda h(p_1) + (1-\lambda)h(p_2)$, st binaarne entroopiafunktsioon on nõgus.

1.1.4 Jenseni võrratus

Edaspidi kasutame tihti Jenseni võrratust. Et Jenseni võrratus käsitleb X keskväärtust, eeldame seejuures, et $\mathcal{X} \subset \mathbb{R}$, st tähed on reaalarvud (vastasel juhul pole EX defineeritud).

Teoreem 1.2 (Jenseni võrratus). *Olgu* $\mathcal{X} \subset \mathbb{R}$, *ja* g *kumer funktsioon, kusjuures* $E|g(X)| < \infty$ *ja* $E|X| < \infty$. *Siis*

$$Eg(X) \geq g(EX). \tag{1.5}$$

Kui g *on rangelt kumer, siis (1.5) on võrdus parajsti siis, kui* $X = EX$ *p.k.*

Tõestus. Tuleta meelde (rangelt) kumera funktsiooni definitisioon. Kumeral funktsioonil g on omadus:

$$\forall y \in \mathbb{R} \quad \exists m(y) \in \mathbb{R}: \quad g(x) - g(y) \geq m(y)(x-y), \quad \forall x \in \mathbb{R}.$$

($m(y) = g'(y)$, kui viimane eksisteerib). Kui g on rangelt kumer, siis on ülaltoodud võrratus võrdus vaid $x = y$ korral.
Olgu $y = EX \in \mathbb{R}$. Iga juhusliku suuruse X väärtuse x_i korral

$$g(x_i) - g(EX) \geq m(EX)(x_i - EX).$$

Seega

$$Eg(X) - g(EX) = \sum \big(g(x_i) - g(EX)\big)p_i \geq m(EX) \sum \big(x_i - EX\big)p_i = m(EX)(EX - EX) = 0$$

ehk

$$Eg(X) \geq g(EX).$$

Näitame nüüd, et rangelt kumera g korral on võrratus võrdus vaid siis, kui $X = EX$ p.k. Olgu

$$Z := \big(g(X) - g(EX)\big) - m(EX)\big(X - EX\big).$$

Juhuslik suurus Z on mittenegatiivne. Seega $EZ = 0$ parajasti siis, kui $Z = 0$ p.k., millest $\big(g(X) - g(EX)\big) = m(EX)\big(X - EX\big)$ p.k.. Rangelt kumera g korral tähendab viimane võrdus, et $X = EX$ p.k. ■

1.2 Ühisentroopia

Olgu X ja Y diskreetsed juhuslikud suurused, mis võtavad väärtusi tähestikel $\mathcal{X}$ ja $\mathcal{Y}$. Seega (X, Y) on diskreetne juhuslik vektor, mille väärtused sisalduvad hulgas

$$\mathcal{X} \times \mathcal{Y} = \{(x, y) : x \in \mathcal{X}, y \in \mathcal{Y}\}.$$

Olgu (X, Y) ühisjaotus P. Seega on P hulgal $\mathcal{X} \times \mathcal{Y}$ antud tõenäosusmõõt. Tähistame

$$p_{ij} := P(x_i, y_j) = \mathbf{P}\big((X, Y) = (x_i, y_j)\big) = \mathbf{P}(X = x_i, Y = y_j).$$

Ühisjaotus esitatakse tihti tabelina

$\mathcal{X}\backslash\mathcal{Y}$	y_1	y_2	$\dots$	y_j	$\dots$	$\sum$
x_1	$P(x_1, y_1) = p_{11}$	$P(x_1, y_2) = p_{12}$	$\dots$	p_{1j}	$\dots$	$\sum_j p_{1j} = P(x_1)$
x_2	$P(x_2, y_1) = p_{21}$	$P(x_1, y_2) = p_{22}$	$\dots$	p_{2j}	$\dots$	$\sum_j p_{2j} = P(x_2)$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
x_i	p_{i1}	p_{i2}	$\dots$	p_{ij}	$\dots$	$\sum_j p_{ij} = P(x_i)$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$\sum$	$\sum_i p_{i1} = P(y_1)$	$\sum_i p_{i2} = P(y_2)$	$\dots$	$\sum_i p_{ij} = P(y_j)$	$\dots$	1

Ülaltoodud tabelis ning ka edaspidi,

$$P(x) := \mathbf{P}(X = x) \quad \text{ja} \quad P(y) := \mathbf{P}(Y = y)$$

tähistavad marginaaltõenäosusi. Pane tähele, et kui mingi paari (x, y) korral $P(x, y) > 0$, siis ka $P(x) > 0$ ja $P(y) > 0$. Kui X ja Y on sõltumatud, siis

$$P(x, y) = P(x)P(y) \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}.$$

Et juhuslikku vektorit (X, Y) võib vaadelda kui diskreetset juhuslikku suurust, avaldub tema entroopia

$$H(X, Y) = -\sum_{ij} p_{ij} \log p_{ij} = -\sum_{(x,y)\in\mathcal{X}\times Y} P(x, y) \log P(x, y) = E\Big(-\log P(X, Y)\Big). \quad (1.6)$$

Def 1.3 *Juhusliku vektori (X, Y) entroopiat (1.6) nimetatakse juhuslike suuruste X ja Y* **ühisentroopiaks (joint entropy)**.

Kui juhuslikud suurused X, Y on sõltumatud, siis

$$\begin{aligned} H(X, Y) &= -\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} P(x, y) \log P(x, y) = -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} P(x)P(y)(\log P(x) + \log P(y)) \\ &= -\sum_{x\in\mathcal{X}} P(x) \log P(x) - \sum_{y\in\mathcal{Y}} P(y) \log P(y) = H(X) + H(Y). \end{aligned}$$

Ülaltoodud argumendi saab esitada ka teisiti. Iga $x \in \mathcal{X}$ ja $y \in \mathcal{Y}$ korral kehtib $\log P(x, y) = \log P(x) + \log P(y)$, millest $\log P(X, Y) = \log P(X) + \log P(Y)$. Keskväärtus on lineaarne, seega

$$\begin{aligned} H(X, Y) &= -E\big(\log P(X, Y)\big) = -E\big(\log P(X) + \log P(Y)\big) \\ &= -E \log P(X) - E \log P(Y) = H(X) + H(Y). \end{aligned}$$

Sõltumatute juhuslike suuruste ühisentroopia on seega komponentide entroopiate summa. See ühtib intuitsiooniga: kui X ja Y on sõltumatud, siis ei anna X väärtuse teadmine mingit informatsiooni Y kohta. See aga tähendab seda, et vektori (X, Y) väärtuse teadasaamine annab niipalju informatsiooni kui mõlematest komponentidest saadava informatsiooni summa.

Analoogiliselt defineeritakse mitme juhusliku suuruse $X_1, \ldots, X_n$ ühisentroopia

$$H(X_1, \ldots, X_n) := -E \log P(X_1, \ldots, X_n).$$

Kui juhuslikud suurused on sõltumatud, siis

$$H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i).$$

1.3 Tinglik entroopia

1.3.1 Definitsioon

Tähistame tinglikud tõenäosused

$$P(x|y) := \mathbf{P}(X = x|Y = y) = \frac{P(x,y)}{P(y)}, \quad P(y|x) := \mathbf{P}(Y = y|X = x) = \frac{P(x,y)}{P(x)}.$$

Tuletame meelde: juhusliku suuruse Y tinglik jaotus tingimusel $X = x$ (eeldusel $P(x) > 0$) on

$$\begin{array}{c|c|c|c} y_1 & y_2 & y_3 & \dots \\ \hline P(y_1|x) & P(y_2|x) & P(y_2|x) & \dots \end{array}.$$

Selle jaotuse entroopia avaldub

$$H(Y|x) :=: H(Y|X = x) := -\sum_{y\in\mathcal{Y}} P(y|x)\log P(y|x).$$

Vaatleme hulgal $\mathcal{X}$ antud funktsiooni $x \mapsto H(Y|x)$. Võttes selle funktsiooni argumendiks juhusliku suuruse X, saame uue juhusliku suuruse (juhusliku suuruse X funktsiooni), mille jaotus on

$$\begin{array}{c|c|c|c} H(Y|x_1) & H(Y|x_2) & H(Y|x_3) & \dots \\ \hline P(x_1) & P(x_2) & P(x_3) & \dots \end{array}.$$

Sellise jaotuse keskväärtus on (tulete meelde $\mathcal{X}_P$ on P kandja – tähed, mille tõenäosus on positiivne)

$$\sum_{x\in\mathcal{X}_P} H(Y|x)P(x).$$

Def 1.4 *Juhusliku suuruse Y* **tinglik entroopia (conditional entropy)** *tingimusel X on*

$$\begin{aligned} H(Y|X) &:= \sum_{x\in\mathcal{X}_P} H(Y|x)P(x) = -\sum_{x\in\mathcal{X}_P} P(x)\sum_{y\in\mathcal{Y}} \log P(y|x)P(y|x) \\ &= -\sum_{x\in\mathcal{X}_P}\sum_{y\in\mathcal{Y}} \log P(y|x)P(x,y) = -E\Big(\log P(Y|X)\Big). \end{aligned}$$

Märkused:

- Kui juhuslikud suurused X ja Y on sõltumatud, siis $P(y|x) = P(y)\ \forall x \in \mathcal{X}, y \in \mathcal{Y}$, millest $H(Y|X) = H(Y)$.
- Üldiselt $H(X|Y)$ ei võrdu $H(Y|X)$. Olgu näiteks X, Y sõltumatud juhuslikud suurused, kusjuures $H(X) \neq H(Y)$. Siis $H(X|Y) = H(X) \neq H(Y) = H(Y|X)$.

- $H(Y|X) = 0$ parajasti siis, kui Y on X funktsioon. Tõepoolest, $H(Y|X) = 0$ parajasti siis, kui $H(Y|X = x) = 0$ iga $x \in \mathcal{X}$ korral. See aga tähendab, et leidub konstant $f(x)$ nii, et $\mathbf{P}(Y = f(x)|X = x) = 1$ ehk $Y = f(X)$. Järelikult kehtib ka $H(X|X) = 0$.

Järgmine väide avab tingliku entroopia olemuse.

Väide 1.3

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y).$$

Tõestus. Iga $(x,y) \in \mathcal{X} \times \mathcal{Y}$ korral nii, et $P(x,y) > 0$ kehtib $P(x,y) = P(x)P(y|x)$, millest

$$\log P(x,y) = \log P(x) + \log P(y|x)$$

Seega

$$H(X,Y) = -E \log P(X,Y) = -E \log P(X) - E \log P(Y|X) = H(X) + H(Y|X).$$

Et $H(X,Y) = H(Y,X)$, siis teine võrdus kehtib ka. ■

1.3.2 Ketireeglid

Olgu X, Y, Z kolm juhuslikku suurust väärtuste hulgaga. Olgu nende kandjad vastavalt $\mathcal{X}, \mathcal{Y}$ ja $\mathcal{Z}$. Analoogiliselt $H(Y|X)$ definitsiooniga defineerime $H(X,Y|Z)$ ja $H(X|Y,Z)$:

$$\begin{aligned}
H(X,Y|Z) &:= -\sum_{z\in\mathcal{Z}} P(z) \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} P(x,y|z) \log P(x,y|z) \\
&= -\sum_{(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}} \log P(x,y|z) P(x,y,z) = -E \log P(X,Y|Z) \\
H(X|Y,Z) &:= -\sum_{(y,z)\in\mathcal{Y}\times\mathcal{Z}} P(y,z) \sum_{x\in\mathcal{X}} P(x|y,z) \log P(x|y,z) \\
&= -\sum_{(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}} \log P(x|y,z) P(x,y,z) = -E \log P(X|Y,Z).
\end{aligned}$$

Nüüd on selge, kuidas suvaliste juhuslike suuruste $X_1, \ldots, X_n$ korral on defineeritud tinglik entroopia

$$H(X_n, X_{n-1}, \ldots, X_j | X_{j-1}, \ldots, X_1).$$

Väide 1.3 üldistub mitmes suunas. Alljärgnev on väite 1.3 tinglik versioon

Väide 1.4

$$H(Y,X|Z) = H(X|Z) + H(Y|X,Z).$$

Tõestus. Iga sellise kolmiku (x, y, z) kus $P(x, y, z) > 0$ korral kehtib

$$P(x, y|z) = P(x|z)P(y|x, z).$$

Nüüd

$$H(X, Y|Z) = -E \log P(X, Y|Z) = -E \log P(X|Z) - E \log P(Y|X, Z) = H(X|Z) + H(Y|X, Z).$$

■

Väitest 1.4 järeldub väide 1.3. Ka järgmine lemma üldistab väidet 1.3.

Lemma 1.1 (Ketireegel) *Olgu $X_1, \ldots, X_n$ juhuslikud suurused. Siis*

$$H(X_1, \ldots, X_n) = H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2) + \cdots + H(X_n|X_1, \ldots, X_{n-1}).$$

Tõestus. Olgu juhuslike suuruste kandjad vastavalt $\mathcal{X}_1, \ldots, \mathcal{X}_n$. Olgu $x_1 \in \mathcal{X}_1, \ldots, x_n \in \mathcal{X}_n$ sellised, et $P(x_1, \ldots, x_n) > 0$. Iga sellise vektori korral kehtib

$$P(x_1, \ldots, x_n) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2) \cdots P(x_n|x_1, \ldots, x_n),$$

millest

$$\begin{aligned} H(X_1, \ldots, X_n) &= -E \log P(X_1, \ldots, X_n) \\ &= -E \log P(X_1) - E \log P(X_2|X_1) - \cdots - E \log P(X_n|X_1, \ldots, X_{n-1}) \\ &= H(X_1) + H(X_2|X_1) + \cdots + H(X_n|X_1, \ldots, X_{n-1}). \end{aligned}$$

■

Kehtib ka ketireegli tinglik versioon.

Lemma 1.2 (Tinglik ketireegel) *Olgu $X_1, \ldots, X_n, Z$ juhuslikud suurused. Siis*

$$H(X_1, \ldots, X_n|Z) = H(X_1|Z) + H(X_2|X_1, Z) + H(X_3|X_1, X_2, Z) + \cdots + H(X_n|X_1, \ldots, X_{n-1}, Z).$$

Tõestus. Olgu juhuslike suuruste $X_1, \ldots, X_n, Z$ kandjad vastavalt $\mathcal{X}_1, \ldots, \mathcal{X}_n$ ja $\mathcal{Z}$. Väide järeldub sellest, et iga $x_i \in \mathcal{X}_i$ ja $z \in \mathcal{Z}$ korral (tingimusel $P(x_1, \ldots, x_n, z) > 0$)

$$P(x_1, \ldots, x_n|z) = P(x_1|z)P(x_2|x_1, z)P(x_3|x_2, x_1, z) \cdots P(x_n|x_1, \ldots, x_{n-1}, z)$$

■

Tinglikust ketireeglist järeldub nii väide 1.4 kui ka ketireegel.

1.4 Kullback-Leibleri kaugus

1.4.1 Definitsioon

Olgu P ja Q kaks jaotust tähestikul $\mathcal{X}$. Tuletame meelde, et need mõõdud esituvad tabelitena

$$P: \begin{array}{c|c|c|c} x_1 & x_2 & x_3 & \dots \\ \hline P(x_1) & P(x_2) & P(x_3) & \dots \end{array} \quad Q: \begin{array}{c|c|c|c} x_1 & x_2 & x_3 & \dots \\ \hline Q(x_1) & Q(x_2) & Q(x_3) & \dots \end{array},$$

kusjuures võib olla, et mõne i korral $Q(x_i) = 0$ või $P(x_i) = 0$.

NB! Lepime kokku, et $0 \log(\frac{0}{q}) = 0$, kui $q \geq 0$, $p \log(\frac{p}{0}) = \infty$, kui $p > 0$.

Def 1.5 *Mõõtude P ja Q* **Kullback-Leibleri kaugus (Kullback-Leibler distance, Kullback-Leibler divergence, relative entropy)** *on*

$$D(P||Q) := \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}. \tag{1.7}$$

Kui $X \sim P$, siis kehtib

$$D(P||Q) = E\Big(\log \frac{P(X)}{Q(X)}\Big).$$

Kui $X \sim P$ ja $Y \sim Q$, siis tähistame ka

$$D(X||Y) := D(P||Q).$$

Märkused:

- $\log \frac{P(x)}{Q(x)}$ ei pruugi olla positiivne. Veendume, et rida (1.7) on sellegipoolest defineeritud. Olgu

$$\mathcal{X}^+ := \Big\{x \in \mathcal{X} : \frac{P(x)}{Q(x)} > 1\Big\}, \quad \mathcal{X}^- := \Big\{x \in \mathcal{X} : \frac{P(x)}{Q(x)} \leq 1\Big\}.$$

  Et

$$\sum_{x \in \mathcal{X}^-} |P(x) \log \frac{P(x)}{Q(x)}| = \sum_{x \in \mathcal{X}^-} P(x) \log \frac{Q(x)}{P(x)} \leq \sum_{x \in \mathcal{X}^-} P(x) \frac{Q(x)}{P(x)} \leq 1.$$

  Seega on rea (1.7) negatiivne osa koonduv. Kui $\sum_{x \in \mathcal{X}^+} P(x) \log \frac{P(x)}{Q(x)} < \infty$, on rida (1.7) koonduv, vastasel juhul on tema summa ∞.

- $D(P||Q)$ nimetatakse küll Kullback-Leibleri kauguseks, kuid ta pole meetrika: kuigi $D(P||Q) \geq 0$, kusjuures $D(P||Q) = 0$ parajasti siis, kui $P = Q$ (tõestus allpool), pole üldiselt $D(P||Q)$ ja $D(Q||P)$ võrdsed (D pole sümmeetriline) ning ei kehti ka kolmurga võrratus (vaata ülesanne 8).

K-L kaugus mõõdab "keskmist üllatust", mille jaotusega P juhuslik suurus meile valmistab, kui eeldame, et tema jaotus on Q. Oletame, et leidub $x' \in \mathcal{X}$ nii, et $Q(x') = 0$, kuid $P(x') > 0$. sellisel juhul

$$\sum_{x\in\mathcal{X}^+} \log\Big(\frac{P(x)}{Q(x)}\Big)P(x) \geq P(x')\log\Big(\frac{P(x')}{Q(x')}\Big) = \infty.$$

Seega on üllatus lõpmatu, kui mingi (meie arvates) võimatu sündmus (x') toimub (vähemalt üks kord). See ühtib intuitsiooniga: võimatu sündmuse toimumist peetakse imeks. Vaatleme aga sellist $x'' \in \mathcal{X}$, et $Q(x'') > 0$, kuid $P(x'') = 0$. sellisel juhul

$$P(x'')\log\Big(\frac{P(x'')}{Q(x'')}\Big) = 0.$$

Selline sündmus kaugust $D(P||Q)$ ei suurenda. Teisisõnu, üllatus ei suurene kui mõni meie meelest positiivse tõenäosusega sündmus x'' toimumata jääb. Ka see ühtib intuitsiooniga: mingi positiivse tõenäosusega sündmuse mittetoimumist üldiselt imeks ei panda. Sellest vaatepunktist lähtudes on K-L kauguse ebasümmeetrilisus igati loogiline.

Näide: Olgu $P = B(1, \frac{1}{2})$, $Q = B(1, q)$. Siis

$$\begin{aligned} D(P||Q) =& \frac{1}{2}\log(\frac{1}{2q}) + \frac{1}{2}\log(\frac{1}{2(1-q)}) = -\frac{1}{2}\log(4q(1-q)) \to \infty, \text{ kui } q \to 0 \\ D(Q||P) =& q\log(2q) + (1-q)\log(2(1-q)) \to 1 \text{ kui } q \to 0. \end{aligned}$$

1.4.2 Gibbsi võrratus ja selle järeldused

Väide 1.5 (Gibbsi võrratus) $D(P||Q) \geq 0$, *kusjuures* $D(P||Q) = 0$ *parajasti siis, kui* $P = Q$.

Tõestus. Kui $D(P||Q) = \infty$, siis väide kehtib triviaalselt. Vaatleme olukorda, kus $D(P||Q) < \infty$, s.t. rida (1.7) on absoluutselt koonduv.
Olgu X jaotusega P juhuslik suurus. Defineerime juhusliku suuruse $Y := \frac{Q(X)}{P(X)}$. Olgu $g(x) := -\log(x)$ rangelt kumer funktsioon. Seega

$$E|g(Y)| = \sum_{x\in\mathcal{X}} |-\log\frac{Q(x)}{P(x)}|P(x) = \sum_{x\in\mathcal{X}} |\log\frac{P(x)}{Q(x)}|P(x) < \infty, \quad E|Y| = \sum_{x\in\mathcal{X}}\frac{Q(x)}{P(x)}P(x) = 1.$$

Jenseni võrratusest järeldub, et

$$D(P||Q) = E\Big(\log\frac{P(X)}{Q(X)}\Big) = E\Big(-\log\frac{Q(X)}{P(X)}\Big) = Eg(Y) \geq g(EY) = -\log(1) = 0,$$

kusjuures $D(P||Q) = 0$ parajasti siis, kui $Y = 1$ p.k. ehk $Q(x) = P(x)$ iga sellise $x \in \mathcal{X}$ korral, et $P(x) > 0$. Sellest järeldub, et $Q(x) = P(x)$ iga $x \in \mathcal{X}$ korral. ■

Gibbsi võrratusest järeldub muuhulgas, et lõpliku tähestiku korral on suurim entroopia ühtlasel jaotusel.

Järeldus 1.1 *Olgu $|\mathcal{X}| < \infty$. Siis iga hulgal $\mathcal{X}$ antud jaotuse P korral $H(P) \leq \log|\mathcal{X}|$, kusjuures võrdus kehtib vaid ühtlase jaotuse korral.*

Tõestus. Olgu U ühtlane jaotus üle $\mathcal{X}$, s.t. $U(x) = |\mathcal{X}|^{-1}$ iga $x \in \mathcal{X}$ korral. Siis

$$D(P||U) = \sum_{x\in\mathcal{X}} P(x)\log\frac{P(x)}{U(x)} = \log|\mathcal{X}| - H(P) \geq 0.$$

■

Väide 1.6 (log-sum võrratus) *Olgu $a_1, a_2, \ldots$ ja $b_1, b_2, \ldots$ mittenegatiivsed arvud, $\sum a_i < \infty$ ja $0 < \sum b_i < \infty$. Siis*

$$\sum a_i \log\frac{a_i}{b_i} \geq \sum a_i \log\frac{\sum a_i}{\sum b_i}, \tag{1.8}$$

kusjuures võrratus on võrdus parajasti siis, kui $\frac{a_i}{b_i} = c \quad \forall i$.

Tõestus. Olgu

$$a_i' = \frac{a_i}{\sum_j a_j}, \quad b_i' = \frac{b_i}{\sum_j b_j}.$$

Seega on $\{a_i'\}$ ja $\{b_i'\}$ tõenäosusjaotused ning väitest 1.5 järeldub

$$0 \leq \sum a_i' \log\frac{a_i'}{b_i'} = \sum \frac{a_i}{\sum_j a_j} \log\frac{\frac{a_i}{\sum_j a_j}}{\frac{b_i}{\sum_j b_j}} = \frac{1}{\sum_j a_j}\Big[\sum a_i \log\frac{a_i}{b_i} - \sum a_i \log\frac{\sum a_j}{\sum b_j}\Big].$$

Et

$$\sum a_i \log\frac{\sum a_j}{\sum b_j} < \infty,$$

siis (1.8) kehtib. Teame, et $D(\{a_i'\}||\{b_i'\}) = 0$ parajasti siis, kui $a_i' = b_i'$, millest

$$\frac{a_i}{b_i} = \frac{\sum_j a_j}{\sum_j b_j} =: c, \quad \forall i.$$

■

Märkus: Log-sum võrratuse tõestus põhineb Gibbsi võrratusel. Samas järeldub viimane otseselt log-sum võrratusest. Seega on need võrratused ekvivalentsed.

Segude K-L kaugus. Olgu P_1, P_2, Q_1, Q_2 hulgal $\mathcal{X}$ antud jaotused. Vaatleme segusi

$$\lambda P_1 + (1-\lambda)P_2 \quad \text{ja } \lambda Q_1 + (1-\lambda)Q_2.$$

Järeldus 1.2

$$D\big(\lambda P_1 + (1-\lambda)P_2||\lambda Q_1 + (1-\lambda)Q_2\big) \leq \lambda D(P_1||Q_1) + (1-\lambda)D(P_2||Q_2). \tag{1.9}$$

Tõestus. Fikseerime $x \in \mathcal{X}$. Log-sum võrratusest järeldub

$$\lambda P_1(x) \log \frac{\lambda P_1(x)}{\lambda Q_1(x)} + (1-\lambda) P_2(x) \log \frac{(1-\lambda) P_2(x)}{(1-\lambda) Q_2(x)}$$
$$\geq \Big(\lambda P_1(x) + (1-\lambda) P_2(x)\Big) \log \frac{\lambda P_1(x) + (1-\lambda) P_2(x)}{\lambda Q_1(x) + (1-\lambda) Q_2(x)}.$$

Summeeri üle hulga $\mathcal{X}$. ■

Võrratust (2.2) võime interpreteerida: K-L kaugus on kumer paaride (P, Q) suhtes. Fikseeritud Q korral järeldub võrratusest (2.2), et funktsioon $P \mapsto D(P||Q)$ on kumer. Samamoodi järeldub, et funktsioon $Q \mapsto D(P||Q)$ on kumer. Veel enam, mõlemad nimetatud funktsioonid on rangelt kumerad (piirkonnas kus nad on lõplikud):

$$D(P||Q) = \sum P(x) \log P(x) - \sum P(x) \log Q(x) = -\sum P(x) \log Q(x) - H(P). \quad (1.10)$$

Funktsioon $P \mapsto \sum P(x) \log Q(x)$ on lineaarne, $P \mapsto H(P)$ aga rangelt nõgus. Seega $P \mapsto D(P||Q)$ on rangelt kumer. Selles mõttes käitub ta kui kaugus.
Seosest (1.10) järeldub ka, et $Q \mapsto D(P||Q)$ on rangelt kumer.

1.4.3 Pinskeri võrratus

Tõenäosusmõõtude omavaheline kaugus. Olgu ühel ja samal tähestikus $\mathcal{X}$ (aga teame, et see eeldus pole kitsendav) antud kaks erinevat tõenäosusmõõtu P ja Q. Kuidas mõõta nende omavahelist kaugust? Tõenäosusteoorias on selleks mitmesuguseid meetrikaid (kaugusi) ja teatavas mõttes mõõdab P ja Q omavahelist kaugust ka K-L kaugus (kuigi ta pole sümmeetriline). Vaadeldes mõõte P ja Q ruumi $\mathbb{R}^{|\mathcal{X}|}$ elementidena (oletame hetkeks, et $|\mathcal{X}| < \infty$) võivad kõne alla tulla kõik ruumis $\mathbb{R}^{|\mathcal{X}|}$ defineeritud kaugused, näiteks eukleidiline kaugus – l_2-meetrika. Selgub, et tõenäosusmõõtude korral on otstarbekas kasutada l_1-meetrikat ja nii defineerimegi P_1 ja P_2 vahelise kauguse järgmiselt:

$$\|P_1 - P_2\| := \sum_{x \in \mathcal{X}} |P_1(x) - P_2(x)|.$$

On lihtne näidata, et defineeritud kaugus on meetrika ning samuti on lihtne näha (ülesanne 9), et

$$\|P_1 - P_2\| = 2 \sup_{B \subseteq \mathcal{X}} |P_1(B) - P_2(B)| = 2|P_1(A) - P_2(A)| \leq 2, \quad (1.11)$$

kus

$$A := \{x \in \mathcal{X} : P_1(x) \geq P_2(x)\}.$$

Seega, kui P_n on tähestikul antud mõõtude jada nii, et $\|P_n - P\| \to 0$, siis iga $B \subset \mathcal{X}$ korral $P_n(B) \to P(B)$, millest loomulikult (aga see tuleneb ju ka vahetult definitsioonist) järeldub, et sellisel juhul iga tähe $x \in \mathcal{X}$ korral $P_n(x) \to P(x)$. Teisest küljest aga on

võimalik näidata (lõpliku tähestiku korral on see ilmne, lõpmatu tähestiku korral järeldub see nn Sheffe lemmast),

$$\|P_n - P\| \to 0 \quad \Leftrightarrow \quad P_n(x) \to P(x), \quad \forall x \in \mathcal{X}.$$

Edaspidi tähistame: $P_n \to P$ tähendab $\|P_n - P\| \to 0$ ja seega $P_n \to P$ parajasti siis, kui $P_n(x) \to P(x)$ iga x korral.

Märkus: Kaugust $\|P_2 - P_2\|$ nimetatakse ingliskeelses kirjanduses *distance of total variation (variational distance)* ja tähistatakse tihti $\|\cdot\|_{TV}$.

Pinskeri võrratus. Pinskeri võrratus väidab muuhulgas, et kui P ja P_n on tähestikul $\mathcal{X}$ defineeritud jaotused nii, et $D(P_n||P) \to 0$ või $D(P||P_n) \to 0$, siis $P_n \to P$.

Teoreem 1.6 (Pinskeri võrratus) *Iga tähestikul $\mathcal{X}$ antud kahe jaotuse P ja Q korral kehtib*

$$D(P||Q) \geq \frac{1}{2\ln 2}\|P - Q\|^2. \tag{1.12}$$

Tõestus. Kõigepealt tõestame võrratuse juhul, kui $|\mathcal{X}| = 2$. Seega olgu $P = (p, 1-p)$ ja $Q = (q, 1-q)$, $\|P - Q\| = 2|p - q|$. Seega on vaja näidata, et

$$g(p,q) := p\log\frac{p}{q} + (1-p)\log\left(\frac{1-p}{1-q}\right) - \frac{4}{2\ln 2}(p-q)^2 \geq 0.$$

Fikseerime p ja võtame tuletise q järgi. Saame (kontrolli!)

$$\frac{dg(p,q)}{dq} = \frac{q-p}{q(1-q)\ln 2} - \frac{4(q-p)}{\ln 2}.$$

Veendu, et kui $0 < q < p$, siis $\frac{dg(p,q)}{dq} < 0$ ehk $q \mapsto g(p,q)$ on kahanev. Et $g(p,p) = 0$, järeldub sellest, et kui $q \leq p$, siis $g(p,q) \geq 0$. Kui $q > p$, siis $1-q < 1-p$ ja tähistades $q := 1-q$, $p := 1-p$ saame jälle, et võrratus kehtib.
Üldise tähestiku korral kasutame log-sum võrratust. Olgu

$$A := \{x \in \mathcal{X} : P(x) \geq Q(x)\}.$$

Defineerime jaotused $\hat{P}$ ja $\hat{Q}$ järgmiselt

$$\hat{P} := (P(A), (1-P(A)), \quad \hat{Q} := (Q(A), (1-Q(A)).$$

Log-sum võrratus:

$$\sum_{x\in A} P(x)\log\frac{P(x)}{Q(x)} \geq P(A)\log\frac{P(A)}{Q(A)}, \quad \sum_{x\in A^c} P(x)\log\frac{P(x)}{Q(x)} \geq ((1-P(A))\log\frac{(1-P(A))}{(1-Q(A))},$$

millest saame, et

$$D(P||Q) \geq D(\hat{P}||\hat{Q}) \geq \frac{4}{2\ln 2}(P(A)-Q(A))^2 = \frac{1}{2\ln 2}\|P-Q\|^2.$$

Siin teine võrratus tulenes sellest, et kahe tähe korral Pinskeri võrratus on juba tõestatud ja viimane võrdus tuleb võrdusest (1.11). ■

Pidevusest. Olles defineerinud tõenäosusmõõtude koondumise on loomulik küsida, kas koondumisest $P_n \to P$ järeldub entroopia koondumine $H(P_n) \to H(P)$, (st kas entroopia on pidev funktsioon) või koondumine $D(P_n||Q) \to D(P||Q)$ või koondumine $D(Q||P_n) \to D(Q||P)$ (st kas K-L kaugus on pidev ühe või teise argumendi järgi).

Entroopia pidevusest. Et $q \mapsto q \log q$ on pidev funktsioon, on lihtne veenduda, et *kui* $|\mathcal{X}| < \infty$, on $P \mapsto H(P)$ pidev funktsioon kõikidel tõenäosusmõõtude hulgal $\mathcal{P}$ (veendu selles!). Tuletame, et pidevus oli ka üks aksioomidest (lõplikumõõtmelise) entroopia defineerimisel. Olukord on aga hoopis teine, kui $|\mathcal{X}| = \infty$. Selgub, et sellisel juhul pole entroopia ühegi mõõdu korral pidev: iga jaotuse P korral leidub jada $P_n \to P$ nii, et $H(P_n) \not\to H(P)$. Väide kehtib ka siis, kui P aatomite hulk on lõplik. Veendume selles. Olgu $|\mathcal{X}| = \infty$, kuid mõõdul P vaid lõplik hulk m aatomeid. Seega olgu

$$P = (p_1, p_2, \ldots, p_m, 0, 0, \ldots).$$

Konstrueerime jaotuste jada P_n järgmiselt:

$$P_n = \big((1 - \frac{1}{n})p_1, \ldots, (1 - \frac{1}{n})p_m, \underbrace{\frac{1}{nM_n}, \ldots \frac{1}{nM_n}}_{M_n}, 0, \ldots\big), \tag{1.13}$$

kus

$$M_n = \lceil 2^{nc} \rceil, \quad c > 0.$$

On kerge veenduda, et et $P_n \to P$ kuid (ülesanne 11)

$$H(P_n) = (1 - \frac{1}{n})H(P) + \frac{1}{n}\log_2 M_n + h(\frac{1}{n}) \to H(P) + c.$$

Vaadeldud näite korral piirjaotusel P on lõplik hulk aatomeid, kuid samasuguse kontranäite saab konstureerida ka siis kui P aatomite arv on lõpmatu ehk kehtib järgmine teoreem.

Teoreem 1.7 (S-W. So ja R. Yeung) *Olgu tähestik $\mathcal{X}$ lõpmatu. Siis iga jaotuse P ja arvu $0 \le c \le \infty$ korral leidub jada P_n nii, et $P_n \to P$, kuid $H(P_n) \to H(P) + c$.*

K-L pidevusest. Vaatleme lühidalt funktsiooni $P \mapsto D(P||Q)$ pidevust. Olgu $|\mathcal{X}| < \infty$. Teame, et $P \mapsto D(P||Q)$ on kumer. Lõplikudimensionaalne kumer funktsioon on pidev piirkonnas kus ta on lõplik. Seega, kui $|\mathcal{X}| < \infty$, $D(P||Q) < \infty$ ja $P_n \to P$ on selline, et $D(P_n||Q) < \infty$ iga n korral, siis kehtib ka koondumine $D(P_n||Q) \to D(P||Q)$. Pane tähele, et ilma lisatingimuseta $D(P_n||Q) < \infty$ ülaltoodud koondumine ei kehti. Kontranäitena vaatleme olukorda, kus $|\mathcal{X}| = 2$, $P = Q = (1, 0)$ ja $P_n = (1 - \frac{1}{n}, \frac{1}{n})$. On selge, et $P_n \to P$, kuid iga n korral $D(P_n||Q) = \infty$.
Lõpliku tähestiku korral on kumer ka funktsioon $Q \mapsto D(P||Q)$ ning sellest järeldub ka selle funktsiooni pidevus.
Juhul, kui $\mathcal{X}$ on lõpmatu, ei järeldu koondumisest $P_n \to P$ koondumine $D(P_n||Q) \to D(P||Q)$. Kontranäide on ülesanne 12.

1.4.4 Tinglik Kullback-Leibleri kaugus

Kullback-Leibleri kaugus mõõdab kahe jaotuse vahelist seost. Tinglik Kullback-Leibleri kaugus mõõdab kahe tingliku jaotuse $P_1(y|x)$ ja $P_2(y|x)$ vahelist seost. Täpsemalt, olgu iga x korral $P_1(y|x)$ ja $P_2(y|x)$ tinglikud jaotused hulgal $\mathcal{Y}$. Seega võime iga sellise x korral, mis rahuldab $P(x) > 0$, defineerida nende jaotuste vahel KL-kauguse

$$D(P_1(y|x)||P_2(y|x)|x) := \sum_{y\in\mathcal{Y}} P_1(y|x)\log\frac{P_1(y|x)}{P_2(y|x)}.$$

Nagu ikka informatsiooniteoorias, keskmistatakse tinglikud karakteristikud üle x-de hulgal $\mathcal{X}$ antud jaotuse $P(x)$.

Def 1.8 *Olgu $P_1(y|x)$ ja $P_2(y|x)$ tingliku jaotused hulgal $\mathcal{Y}$. Hulgal $\mathcal{X}$ antud jaotuse $P(x)$ korral* **tinglik Kullback-Leibleri kaugus (conditional relative entropy)** *on*

$$\begin{aligned} D(P_1(y|x)||P_2(y|x)) &:= \sum_{x\in\mathcal{X}_P} D(P_1(y|x)||P_2(y|x)|x)P(x) = \sum_{x\in\mathcal{X}_P} P(x)\sum_{y\in\mathcal{Y}} P_1(y|x)\log\frac{P_1(y|x)}{P_2(y|x)} \\ &= \sum_{x\in\mathcal{X}_P}\sum_{y\in\mathcal{Y}} P_1(y,x)\log\frac{P_1(y|x)}{P_2(y|x)}, \ \textit{kus}\ P_1(x,y) := P(x)P_1(y|x). \end{aligned}$$

Olgu nüüd X jaotusega P juhuslik suurus; (X,Y_1) ja (X,Y_2) olgu jaotustega $P_1(x,y) = P(x)P_1(y|x)$ ja $P_2(x,y) = P(x)P_2(y|x)$ juhuslikud vektorid, st $P_i(y|x)$ on Y_i tinglik jaotus tingimusel $X = x$, $(i = 1, 2)$. Sellisel juhul

$$D\big(P_1(y|x)||P_2(y|x)\big) = E\log\frac{P_1(Y_1|X)}{P_2(Y_1|X)} =: D(Y_1||Y_2|X) \tag{1.14}$$

Märkused:

1. Tähistusest $D(P_1(y|x)||P_2(y|x))$ ei selgu, milline on jaotus P, üle mille keskmistatakse. Harilikult selgub see kontekstist.

2. Tähistus $D(Y_1||Y_2|X)$ võib olla eksitav. Olgu näiteks (X_1,Y_1) ning (X_2,Y_2) kaks juhuslikku vektorit ühisjaotustega vastavalt $P_1(x,y) = P_1(x)P_1(y|x)$ ja $P_2(x,y) = P_2(x)P_2(y|x)$. Võttes $P(x) = P_1(x)$, saame

$$D\big(P_1(y|x)||P_2(y|x)\big) = E\log\frac{P_1(Y_1|X_1)}{P_2(Y_1|X_1)}. \tag{1.15}$$

   Võrduse (1.15) parem pool on igati korrektne, kuid tähistuse $D(Y_1||Y_2|X_1)$ korral tuleb meeles pidada, et $P_2(x,y)$ pole mitte (X_1,Y_2) vaid (X_2,Y_2) ühisjaotus. Seega $P_2(y|x)$ on juhusliku suuruse Y_2 tinglik jaotus tingimusel X_2 (mis tähistuses ei figureerigi) mitte X_1. Seda tuleb meeles pidada eelkõige KL-kauguse ketireegli (väide 1.9) korral.

Väide 1.7

$$D(P_1(y|x)||P_2(y|x)) \geq 0,$$

kusjuures võrdus kehtib vaid siis kui $P_1(y|x) = P_2(y|x)$ $\forall y \in \mathcal{Y}$ *ja iga* $x \in \mathcal{X}_P$.

Tõestus. Iga $x \in \mathcal{X}$ korral $D(P_1(y|x)||P_2(y|x)|x) \geq 0$, millest järelduvalt

$$D(P_1(y|x)||P_2(y|x)) \geq 0.$$

Oletame, et

$$D(P_1(y|x)||P_2(y|x)) = 0.$$

Siis iga $x \in \mathcal{X}_P$ korral kehtib $D(P_1(y|x)||P_2(y|x)|x) = 0$ ja sellest järeldub väide. ■

Väide 1.8 (Tingimustamine suurendab K-L kaugust)

$$D(P_1(y|x)||P_2(y|x)) \geq D(P_1||P_2),$$

kus $P_i(y) = \sum_x P_i(y|x)P(x)$, *kus* $i = 1, 2$.

Tõestus. Log-sum võrratusest saame, et iga $y \in \mathcal{Y}$ korral

$$\sum_x P_1(y|x)P(x) \log \frac{P_1(y|x)P(x)}{P_2(y|x)P(x)} \geq P_1(y) \log \frac{P_1(y)}{P_2(y)}.$$

Summeeri üle $\mathcal{Y}$. ■

Väide 1.9 (K-L kauguse ketireegel) *Olgu* $(X_1, \ldots, X_n)$ *ja* $(Y_1, \ldots Y_n)$ *juhuslikud vektorid, mis võtavad väärtusi hulgal* $\mathcal{X} \times \cdots \times \mathcal{X}$. *Siis*

$$D\Big((X_1, \ldots, X_n)\Big|\Big|(Y_1, \ldots, Y_n)\Big) =$$
$$D(X_1||Y_1) + D(X_2||Y_2|X_1) + D(X_3||Y_3|X_1, X_2) + \cdots + D(X_n||Y_n|X_1, \ldots, X_{n-1}).$$

Tõestus. Olgu

$$P(x_1, \ldots, x_n) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2) \cdots P(x_n|x_1, \ldots, x_{n-1})$$

vektori $(X_1, \ldots, X_n)$ jaotus ning olgu

$$Q(x_1, \ldots, x_n) = Q(x_1)Q(x_2|x_1) \cdots Q(x_n|x_1, \ldots, x_{n-1})$$

vektori $(Y_1, \ldots, Y_n)$ jaotus. Juhuslike vektorite vaheline K-L kaugus on defineeritud

$$\begin{aligned}
D(X_1, \ldots, X_n||Y_1, \ldots, Y_n) &= E \log \frac{P(X_1, \ldots, X_n)}{Q(X_1, \ldots, X_n)} \\
&= E \log \frac{P(X_1)P(X_2|X_1) \cdots P(X_n|X_1, \ldots, X_{n-1})}{Q(X_1)Q(X_2|X_1) \cdots Q(X_n|X_1, \ldots, X_{n-1})} \\
&= E \log \frac{P(X_1)}{Q(X_1)} + E \log \frac{P(X_2|X_1)}{Q(X_2|X_1)} + \cdots + E \log \frac{P(X_n|X_1, \ldots, X_{n-1})}{Q(X_n|X_1, \ldots, X_{n-1})} \\
&= D(X_1||Y_1) + D(X_2||Y_2|X_1) + \cdots + D(X_n||Y_n|X_1, \ldots, X_{n-1}).
\end{aligned}$$

■

1.5 Vastastikune informatsioon

Olgu (X, Y) juhuslik vektor ühisjaotusega $P(x, y)$, $(x, y) \in \mathcal{X} \times \mathcal{Y}$.

Def 1.9 *Juhuslike suuruste* X, Y **vastastikune informatsioon (mutual information)** *on*

$$I(X;Y) := \sum_{x,y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)} = D\big(P(x,y)||P(x)P(y)\big) = E\Big(\log \frac{P(X,Y)}{P(X)P(Y)}\Big).$$

Vastastikune informatsioon on seega K-L kaugus jaotuse $P(x, y)$ ning korrutismõõdu $P(x)P(y)$ vahel. Teisisõnu, $I(X;Y)$ on K-L kaugus vektori (X, Y) ja samade marginaaljaotusega kuid *sõltumatute* komponentidega vektori vahel.

Märkused:

- Vastastikune informatsioon $I(X;Y)$ ei sõltu mitte ainult juhuslike suuruste X ja Y jaotusest vaid ka nende ühisjaotusest, s.t. vektori (X, Y) jaotusest.
- $0 \le I(X;Y)$.
- Vastastikune informatsioon on sümmeetriline: $I(X;Y) = I(Y;X)$.
- $I(X;Y) = 0$ parajasti siis kui X, Y on sõltumatud.

Vastastikuse informatsiooni olemust aitab mõista järgmine seos:

$$\begin{aligned} I(X;Y) &= E \log \frac{P(X,Y)}{P(X)P(Y)} = E \log \frac{P(X|Y)P(Y)}{P(X)P(Y)} = E \log \frac{P(X|Y)}{P(X)} \\ &= E \log P(X|Y) - E \log P(X) = H(X) - H(X|Y). \end{aligned}$$

Sümmeetria tõttu kehtib

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X). \tag{1.16}$$

Suurus $H(X)$ on juhusliku suuruse X "keskmine juhuslikkus", tema (väärtuse teadasaamisel saadav) informatsioon. Tinglik entroopia $H(X|Y)$ on juhusliku suuruse X entroopia tingimusel, et Y on teada ehk X tinglik "juhuslikkus". On selge, et mida rohkem annab Y informatsiooni X kohta, seda väiksem on $H(X|Y)$. Kui $X = f(Y)$, siis $H(X|Y) = 0$. Kui X ja Y on sõltumatud, siis $H(X|Y) = H(X)$. Mida väiksem on $H(X|Y)$, seda suurem on vahe $H(X) - H(X|Y) = I(X;Y)$. Nüüd on selge, mida $I(X;Y)$ mõõdab: juhusliku suuruse X entroopia kahanemist juhusliku suuruse Y läbi. Valemist (1.16) järeldub, et täpselt sama palju kahaneb $H(Y)$ juhusliku suuruse X läbi. Sellest ka nimetus: vastastikune informatsioon. Kui X ja Y on sõltumatud, siis $I(X;Y) = 0$ - juhuslikud suurused X ka Y ei anna teineteise kohta mingisugust informatsiooni. Paneme tähele, et

$$I(X;X) = H(X) - H(X|X) = H(X),$$

s.t. juhuslik suurus X annab iseenese kohta täpselt $H(X)$ informatsiooni. Inglisekeelses kirjanduses kutsutaksegi entroopiat teinekord *self-information*.

Väide 1.3: $H(X|Y) = H(X,Y) - H(Y)$, millest

$$I(X;Y) = H(X) + H(Y) - H(X,Y). \tag{1.17}$$

Vastastikuse informatsiooni, tingliku entroopia ja entroopia omavahelisi seoseid aitab mõista alljärgnev diagramm.

Teeme veel mõned lihtsad kuid olulised järeldused.

Järeldus 1.3 (tingimustamine vähendab entroopiat) *Juhuslike suuruste X ja Y korral kehtib*

$$H(X|Y) \leq H(X),$$

kusjuures ülaltoodud võrratus on võrdus vaid sõltumatute juhuslike suuruste korral.

Tõestus. $H(X) - H(X|Y) = I(X;Y) \geq 0$. ■

Märkus: Tuleta meelde, et $H(X|Y) = \sum_y H(X|Y=y)P(y)$. Kuigi ülaltoodud summa on väiksem kui $H(X)$, võib mõne $y \in \mathcal{Y}$ korral siiski olla, et $H(X|Y=y) > H(X)$.

Näide:

$\mathcal{Y}\backslash\mathcal{X}$	a	b
u	0	$\frac{3}{4}$
v	$\frac{1}{8}$	$\frac{1}{8}$

Järeldus 1.4 *Juhusliku vektori $(X_1, \ldots, X_n)$ entroopia rahuldab*

$$H(X_1, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i),$$

kusjuures võrratus on võrdus vaid sõltumatute komponentide korral.

Tõestus. Ketireegelist saame

$$H(X_1, \ldots, X_n) = H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2) + \cdots + H(X_n|X_1, \ldots, X_{n-1}).$$

Kasuta eelmist järeldust. ■

1.5.1 Tinglik vastastikune informatsioon

Olgu X, Y, Z juhuslikud suurused, kusjuures Z kandja olgu $\mathcal{Z}$.

Def 1.10 *Juhuslike suuruste* X, Y **vastastikune informatsioon tingimusel** Z **(conditional mutual information)** *on*

$$\begin{aligned}
I(X;Y|Z) :=& H(X|Z) - H(X|Y,Z) = E\log\frac{P(X|Y,Z)}{P(X|Z)} \\
=& E\log\frac{P(X|Y,Z)P(Y|Z)}{P(X|Z)P(Y|Z)} = E\log\frac{P(X,Y|Z)}{P(X|Z)P(Y|Z)} \\
=& \sum_{x,y,z} P(x,y,z)\log\frac{P(x,y|z)}{P(x|z)P(y|z)} \\
=& \sum_{z\in\mathcal{Z}} P(z)\sum_{y,x} P(x,y|z)\log\frac{P(x,y|z)}{P(x|z)P(y|z)} \\
=& \sum_{z\in\mathcal{Z}} D\big(P(x,y|z)||P(x|z)P(y|z)\big)P(z).
\end{aligned}$$

Väide 1.10

$$I(X;Y|Z) \geq 0,$$

kusjuures võrdus kehtib parajasti siis, kui X *ja* Y *on tinglikult sõltumatud, s.t.*

$$P(x,y|z) = P(x|z)P(y|z), \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}, z \in \mathcal{Z}. \tag{1.18}$$

Tõestus. Et iga z korral

$$D\Big(P(x,y|z)||P(x|z)P(y|z)\Big)P(z) \geq 0,$$

siis $I(X;Y|Z) = 0$ parajasti siis, kui iga $z \in \mathcal{Z}$ korral

$$D\Big(P(x,y|z)||P(x|z)P(y|z)\Big) = 0$$

ja sellest järeldub (1.18). ■

Tinglikul vastastikusel informatsioonil on üldiselt samad omadused mis vastastikusel informatsioonil. Kehtib (ülesanne 21)

$$\begin{aligned}
I(X;X|Z) &= H(X|Z) \\
I(X;Y|Z) &= H(Y|Z) - H(Y|X,Z) \\
I(X;Y|Z) &= H(X|Z) + H(Y|Z) - H(X,Y|Z).
\end{aligned}$$

Lisaks kehtib veel (ülesanne 21)

$$I(X;Y|Z) = H(X;Z) + H(Y;Z) - H(X,Y,Z) - H(Z). \tag{1.19}$$

Väide 1.11 (Vastastikuse informatsiooni ketireegel)

$$I(X_1,\dots,X_n;Y) = I(X_1;Y)+I(X_2;Y|X_1)+I(X_3;Y|X_1,X_2)+\cdots+I(X_n;Y|X_1,\dots,X_{n-1}).$$

Tõestus. Kasutame entroopia ketireeglit ja tingliku entroopia ketireeglit.

$$\begin{aligned} I(X_1,\dots,X_n;Y) =& H(X_1,\dots,X_n) - H(X_1,\dots,X_n|Y) \\ =& H(X_1) + H(X_2|X_1) + \cdots + H(X_n|X_1,\dots,X_{n-1}) \\ & - H(X_1|Y) - H(X_2|X_1,Y) - \cdots - H(X_n|X_1,\dots,X_{n-1},Y). \end{aligned}$$

■

Väide 1.12 (Tingliku vastastikuse informatsiooni ketireegel)

$$I(X_1,\dots,X_n;Y|Z) = I(X_1;Y|Z) + I(X_2;Y|X_1,Z) + \cdots + I(X_n;Y|X_1,\dots,X_{n-1},Z).$$

Tõestus. Analoogiline. ■

1.6 Andmetöötlusvõrratus

1.6.1 Lõplik Markovi ahel

Def 1.11 *Juhuslikud suurused $X_1,\dots,X_n$ kandjatega vastavalt $\mathcal{X}_1,\dots,\mathcal{X}_m$ moodustavad* **Markovi ahela** *kui iga $x_i \in \mathcal{X}_i$ ja iga $m = 2,\dots,n-1$ korral*

$$\mathbf{P}(X_{m+1} = x_{m+1}|X_m = x_m,\dots,X_1 = x_1) = \mathbf{P}(X_{m+1} = x_{m+1}|X_m = x_m). \tag{1.20}$$

Seega on $X_1,\dots,X_n$ Markovi ahel parajasti siis, kui iga $x_1,\dots,x_n$ korral

$$P(x_1,\dots,x_n) = \begin{cases} P(x_1,x_2)P(x_3|x_2)\cdots P(x_n|x_{n-1}) & \text{kui } P(x_2)>0,\dots,P(x_n)>0, \\ 0 & \text{muidu.} \end{cases}$$

Asjaolu, et $X_1,\dots,X_n$ on Markovi ahel tähistatakse informatsiooniteoorias tihti:

$$X_1 \to X_2 \to \cdots \to X_n.$$

Seega $X \to Y \to Z$ parajasti siis, kui

$$P(x,y,z) = P(x)P(y|x)P(z|y).$$

Väide 1.13 *Kui $X_1 \to X_2 \to \cdots \to X_n$, siis $X_n \to X_{n-1} \to \cdots \to X_1$.*

Tõestus. $X_1 \to X_2 \to \cdots \to X_n$ parajasti siis kui

$$P(x_1,\dots,x_n)P(x_2)\cdots P(x_{n-1}) = P(x_1,x_2)P(x_2,x_3)\cdots P(x_{n-1},x_n).$$

See on aga sümmeetriline. ■

Väide 1.14 *Markovi ahela iga alamjada on Markovi ahel, s.t. kui* $X_1 \to X_2 \to \cdots \to X_n$, *siis* $X_{n_1} \to X_{n_2} \to \cdots \to X_{n_k}$.

Tõestus. Tuletame meelde tingliku täistõenäosuse valemi: kui $A, B, C_1, C_2, \ldots$ on sündmused ning $C_1, C_2, \ldots$ on täissüsteem (st $C_i \cap C_j = \emptyset$ ja $\mathbf{P}(\cup_i C_i) = 1$), siis

$$\mathbf{P}(A|B) = \sum_i \mathbf{P}(A|B, C_i)\mathbf{P}(C_i|B). \tag{1.21}$$

Fikseerime m ja näitame, et

$$\mathbf{P}(X_{m+2} = x_{m+2}|X_m = x_m, \ldots, X_1 = x_1) = \mathbf{P}(X_{m+2} = x_{m+2}|X_m = x_m)$$

ehk

$$P(x_{m+2}|x_m, \ldots, x_1) = P(x_{m+2}|x_m).$$

Kõigepealt paneme tähele, et valemit (1.21) kasutades saame

$$\begin{aligned} P(x_{m+2}|x_{m+1}, x_m) &= \sum_{x_1,\ldots,x_{m-1}} P(x_{m+2}|x_{m+1}, x_m, x_{m-1}, \ldots, x_1)P(x_{m-1}, \ldots, x_1|x_m, x_{m+1}) \\ &= \sum_{x_1,\ldots,x_{m-1}} P(x_{m+2}|x_{m+1})P(x_{m-1}, \ldots, x_1|x_m, x_{m+1}) = P(x_{m+2}|x_{m+1}). \end{aligned}$$

Analoogiliselt saame, et iga $m_1 < m_2 < \cdots < m_k \leq m$ korral

$$P(x_{m+2}|x_{m+1}, x_{m_k}, x_{m_{k-1}}, \cdots, x_{m_1}) = P(x_{m+2}|x_{m+1}) \tag{1.22}$$

[Seosest (1.22) järeldub $P(x_{m+2}|x_{m+1}, x_m) = P(x_{m+2}|x_{m+1})$ (kuidas?)].
Seega

$$\begin{aligned} P(x_{m+2}, x_{m+1}|x_m, \ldots, x_1) &= P(x_{m+2}|x_{m+1}, x_m, \ldots, x_1)P(x_{m+1}|x_m, \ldots, x_1) \\ &= P(x_{m+2}|x_{m+1}, x_m)P(x_{m+1}|x_m) \\ &= P(x_{m+2}, x_{m+1}|x_m). \end{aligned}$$

Seega

$$\begin{aligned} P(x_{m+2}|x_m, \ldots, x_1) &= \sum_{x_{m+1}} P(x_{m+2}, x_{m+1}|x_m, \ldots, x_1) \\ &= \sum_{x_{m+1}} P(x_{m+2}, x_{m+1}|x_m) = P(x_{m+2}|x_m). \end{aligned}$$

Viimasest võrdusest ja seosest (1.22) järeldub, et $X_1, \ldots, X_m, X_{m+2}, \ldots X_n$ on Markovi ahel. Siit järeldub ülejäänu. ■

Järeldus 1.5 *Kui* $X_1 \to X_2 \to \cdots \to X_n$, *siis iga* $m < n$ *korral*

$$P(x_n, \ldots, x_{m+1}|x_m, \ldots, x_1) = P(x_n, \ldots, x_{m+1}|x_m). \tag{1.23}$$

Tõestus. Tõepoolest, kui $X_1 \to X_2 \to \cdots \to X_n$ on Markovi ahel, siis Väite 1.14 korral on seda ka $X_k \to \cdots \to X_n$ ($k \geq 1$), millest iga $m > k$ korral

$$P(x_m|x_{m-1}, \ldots, x_k) = P(x_m|x_{m-1}) \tag{1.24}$$

Tõestusest saime, et $P(x_{m+2}, x_{m+1}|x_m, \ldots, x_1) = P(x_{m+2}, x_{m+1}|x_m)$. Kasutades seda võrdust saame

$$\begin{aligned} P(x_{m+3}, x_{m+2}, x_{m+1}|x_m, \ldots, x_1) &= P(x_{m+3}|x_{m+2}, x_{m+1}, x_m, \ldots, x_1)P(x_{m+2}, x_{m+1}|x_m, \ldots, x_1) \\ &= P(x_{m+3}|x_{m+2}, x_{m+1}, x_m, \ldots, x_1)P(x_{m+2}, x_{m+1}|x_m) \\ &= P(x_{m+3}|x_{m+2}, x_{m+1}, x_m)P(x_{m+2}, x_{m+1}|x_m) \\ &= P(x_{m+3}, x_{m+2}, x_{m+1}|x_m). \end{aligned}$$

Siin eelviimane võrdus tuleneb seosest (1.24). Edasi jätka induktsiooniga. ■

Väide 1.15 *Juhuslikud suurused $X_1, \ldots, X_n$ on Markovi ahel parajsti siis, kui iga $m = 2, \ldots, n-1$ korral $X_1, \ldots, X_{m-1}$ ja $X_{m+1}, \ldots, X_n$ on antud X_m korral tinglikult sõltumatud.*

Tõestus. Olgu $X_1, \ldots, X_n$ Markovi ahel. Tõestame, et

$$P(x_1, \ldots, x_{m-1}, x_{m+1}, \ldots, x_n|x_m) = P(x_1, \ldots, x_{m-1}|x_m)P(x_{m+1}, \ldots, x_n|x_m). \tag{1.25}$$

Seosest (1.23) saame

$$P(x_1, \ldots, x_n) = P(x_1, \ldots, x_m)P(x_{m+1}, \ldots, x_n|x_1, \ldots, x_m) = P(x_1, \ldots, x_m)P(x_{m+1}, \ldots, x_n|x_m),$$

millest

$$\frac{P(x_1, \ldots, x_n)}{P(x_m)} = \frac{P(x_1, \ldots, x_m)}{P(x_m)}P(x_{m+1}, \ldots, x_n|x_m) = P(x_1, \ldots, x_{m-1}|x_m)P(x_{m+1}, \ldots, x_n|x_m).$$

Kehtigu (1.25). Siis

$$\begin{aligned} P(x_{m+1}, \ldots, x_n|x_1, \ldots, x_m) &= \frac{P(x_1, \ldots, x_n)}{P(x_1, \ldots, x_m)} = \frac{P(x_1, \ldots, x_n)}{P(x_m)P(x_1, \ldots, x_{m-1}|x_m)} \\ &= \frac{P(x_1, \ldots, x_{m-1}, x_{m+1}, \ldots, x_n|x_m)}{P(x_1, \ldots, x_{m-1}|x_m)} = P(x_{m+1}, \ldots, x_n|x_m). \end{aligned}$$

■

Seega $X \to Y \to Z$ parjasti siis, kui antud Y korral on X ja Z tinglikult sõltumatud.

1.6.2 Andmetöötlusvõrratus

Lemma 1.3 (Andmetöötlusvõrratus) *Kui $X \to Y \to Z$, siis*

$$I(X;Y) \geq I(X;Z),$$

kusjuures võrdus kehtib parajasti siis, kui $X \to Z \to Y$.

Tõestus. Et X ja Z on antud Y korral sõltumatud, siis $I(X;Z|Y) = 0$. Seega ketireeglist saame

$$I(X;Y,Z) = I(X;Z) + I(X;Y|Z) = I(X;Y) + I(X;Z|Y) = I(X;Y). \tag{1.26}$$

Et $I(X;Y|Z) \geq 0$, siis $I(X;Z) \leq I(X;Y)$, kusjuures võrdus kehtib parajsti siis, kui $I(X;Y|Z) = 0$ ehk antud Z korral on X ja Y tinglikult sõltumatud ehk $X \to Z \to Y$ on Markovi ahel. ■

Olgu X juhuslik suurus, mille kohta vajame informatsiooni. Juhuslik suurus X on meil teadmata, meie käsutuses on vaid Y (andmed), mis annab X kohta $I(X;Y)$ bitti informatsiooni. Kas aga on võimalik Y töödelda nii, et X kohta saadav informatsioon suureneks? Juhuslikku suurust Y on võimalik töödelda determineeritult, s.t. rakendame talle mingit funktsiooni g. Seega saame uue juhusliku suuruse $g(Y)$. Et aga $X \to Y \to g(Y)$ on Markovi ahel, siis andmetöötlusvõrratusest saame, et $I(X;Y) \geq I(X;g(Y))$ ehk $g(Y)$ ei anna rohkem informatsiooni X kohta, kui Y. Teine võimalus on töödelda Y juhuslikult, s.t. lisada mingi X-st sõltumatu lisajuhuslikkus. Olgu Z andmete Y juhuslikul töötlemisel saadud juhuslik suurus. Et lisajuhuslikkus on X-st sõltumatu, on $X \to Y \to Z$ Markovi ahel ning andmetöötlusvõrratusest järeldub $I(X;Y) \geq (X;Z)$, s.t. ka juhuslik töötlemine ei suurenda informatsiooni. Seega postuleerib andmetöötlusvõrratus väga üldise printsiibi: andemete (juhuslikul või mittejuhuslikul) töötlemisel võib informatsioon vaid kaotsi minna, mitte mingil juhul ei saa aga informatsiooni juurde võita. Kas sellest järeldub igasuguse statistilise andmetöötluse mõttetus?

Järeldus 1.6 *Kui $X \to Y \to Z$, siis*

$$H(X|Z) \geq H(X|Y).$$

Tõestus. Ülesanne 25. ■

Järeldus 1.7 *Kui $X \to Y \to Z$, siis*

$$I(X;Z) \leq I(Y;Z), \quad I(X;Y|Z) \leq I(X;Y).$$

Tõestus. Ülesanne 25. ■

1.6.3 Piisav statistik

Olgu $\{P_\theta\}$ hulgal $\mathcal{X}$ antud tõenäosusjaotuste klass. Statistikas interpreteeritakse hulka $\{P_\theta\}$ kui mudelit, indeksit θ nimetatakse parameetriks. Olgu X juhuslik valim jaotusest P_θ. Juhuslikku valimit X vaatleme kui juhuslikku suurust väärtuste hulgaga $\mathcal{X}^n$. Seega sõltub X jaotus vaid parameetrist θ. Olgu $T(X)$ mingi statistik (valimi funktsioon), mille abil püüame hinnata valimi genereerivat jaotust P_θ ehk siis parameetrit θ. Vaatleme olukorda, kus parameeter θ on juhuslik eeljaotusega π (Bayesi lähenemisviis). Sellisel juhul $\theta \to X \to T(X)$ on Markovi ahel ning andmetöötlusvõrratusest saame, et

$$I(\theta; T(X)) \le I(\theta; X).$$

Kui ülaltoodud võrratus on võrdus, siis on statistik T selline, et $T(X)$ annab parameetri kohta sama palju informatsiooni kui X (sõltumata parameetri eeljaotusest π). Lemmast 1.3 teame, et võrdus kehtib parajasti siis, kui antud $T(X)$ korral on X ja θ sõltumatud ehk $\theta \to T(X) \to X$. Seos $\theta \to T(X) \to X$ kehtib aga parajasti siis, kui iga valimi $x \in \mathcal{X}^n$ korral

$$\mathbf{P}(X = x | T(X) = t, \theta) = \mathbf{P}(X = x | T(X) = t)$$

ehk antud $T(X)$ korral ei sõltu valimi jaotus parameetrist θ. Statistikas nimetatakse selliseid statistikuid *piisavateks*. Seega oleme tõestanud järelduse.

Järeldus 1.8 *Statistik T on piisav parajasti siis, kui iga θ jaotuse korral*

$$I(\theta; T(X)) = I(\theta; X).$$

Näide: Olgu $\{P_\theta\}$ Bernoulli jaotuste hulk. Statistik $T(X) = \sum_{i=1}^n X_i$ on piisav, sest

$$\mathbf{P}(X_1 = x_1, \ldots, X_i = x_i | T(X) = t, \theta) = \begin{cases} 0 & \text{kui } \sum_i x_i \neq t, \\ \frac{1}{C_n^t} & \text{kui } \sum_i x_i = t. \end{cases}$$

Tõepoolest, kui $\sum_i x_i = t$, siis

$$\begin{aligned} \mathbf{P}(X_1 = x_1, \ldots, X_n = x_n | T(X) = t, \theta) &= \frac{\mathbf{P}(X_1 = x_1, \ldots, X_n = x_n, T(X) = t, \theta)}{\mathbf{P}(T(X) = t, \theta)} \\ &= \frac{\theta^t (1-\theta)^{n-t} \pi(\theta)}{\sum_{x_1,\ldots,x_n : \sum_i x_i = t} \theta^t (1-\theta)^{n-t} \pi(\theta)} = \frac{1}{C_n^t}, \end{aligned}$$

sest fikseetud ühtede arvu korral on erinevateks valimiteks täpselt C_n^t võimalust.

1.7 Fano võrratus

Olgu X tundmatu juhuslik suurus ning olgu $\hat{X}$ korreleeritud juhuslik suurus, mida vaatleme kui X hinnangut. Olgu

$$P_e := \mathbf{P}(X \neq \hat{X})$$

hindamisel tehatava vea tõenäosus. Kui $P_e = 0$, siis $X = \hat{X}$ p.k., millest $H(X|\hat{X}) = 0$. Seega on loogiline, et kui P_e on väike, siis $H(X|\hat{X})$ peaks samuti väike olema. Selgub, et lõpliku tähestiku korral see nii ongi.

Teoreem 1.12 (Fano võrratus) *Olgu X ja $\hat{X}$ juhuslikud suurused tähestikul $\mathcal{X}$. Siis*

$$H(X|\hat{X}) \leq h(P_e) + P_e \log(|\mathcal{X}| - 1), \tag{1.27}$$

kus h on binaarne entroopiafunktsioon.

Tõestus. Olgu

$$E = \begin{cases} 1 & \text{kui } \hat{X} \neq X, \\ 0 & \text{kui } \hat{X} = X. \end{cases}$$

Seega

$$E = I_{\{\hat{X} \neq X\}}, \quad E \sim B(1, P_e).$$

Entroopia ketireeglist saame

$$H(E, X|\hat{X}) = H(X|\hat{X}) + H(E|X, \hat{X}) = H(X|\hat{X}), \tag{1.28}$$

sest $H(E|X, \hat{X}) = 0$ (miks?)
Teisest küljest

$$H(E, X|\hat{X}) = H(E|\hat{X}) + H(X|E, \hat{X}) \leq H(E) + H(X|E, \hat{X}) = h(P_e) + H(X|E, \hat{X}).$$

Paneme tähele, et

$$\begin{aligned} H(X|E, \hat{X}) = &\sum_{x \in \mathcal{X}} \mathbf{P}(\hat{X} = x, E = 1) H(X|\hat{X} = x, E = 1) \\ &+ \sum_{x \in \mathcal{X}} \mathbf{P}(\hat{X} = x, E = 0) H(X|\hat{X} = x, E = 0). \end{aligned}$$

Tingimusel $\hat{X} = x$ ja $E = 0$ kehtib $X = x$, siis on $H(X|\hat{X} = x, E = 0) = 0$ ehk

$$H(X|E, \hat{X}) = \sum_{x \in \mathcal{X}} \mathbf{P}(\hat{X} = x, E = 1) H(X|\hat{X} = x, E = 1).$$

Kui $E = 1$ ja $\hat{X} = x$ siis $X \in \mathcal{X} \backslash x$, millest $H(X|\hat{X} = x, E = 1) \leq \log(|\mathcal{X}| - 1)$. Kokkuvõttes

$$H(X|E, \hat{X}) \leq P_e \log(|\mathcal{X}| - 1).$$

Seosest (1.28) saame, et

$$H(X|\hat{X}) \leq P_e \log(|\mathcal{X}| - 1) + h(P_e).$$

■

Järeldus 1.9

$$H(X|\hat{X}) \leq 1 + P_e \log |\mathcal{X}|, \quad \textit{ehk} \quad P_e \geq \frac{H(X|\hat{X}) - 1}{\log |\mathcal{X}|}.$$

Kui $|\mathcal{X}| < \infty$, siis Fano võrratusest järeldub, et kui $P_e \to 0$, siis $H(X|\hat{X}) \to 0$. Kui aga tähestik on lõpmatu, siis Fano võrratus on trivaalne ja ülaltoodud implikatsioon ei pruugi kehtida.

Näide: Olgu $Z \sim B(1,p)$ ning olgu Y mingi selline juhuslik suurus, et $Y > 0$ ja $H(Y) = \infty$. Defineerime juhusliku suuruse X järgmiselt

$$X = \begin{cases} 0 & \text{kui } Z = 0, \\ Y & \text{kui } Z = 1. \end{cases}$$

Olgu $\hat{X} = 0$ p.k. Siis $P_e = \mathbf{P}(X > 0) = \mathbf{P}(X = Y) = \mathbf{P}(Z = 1) = p$. Kuid

$$H(X|\hat{X}) = H(X) \geq H(X|Z) = pH(Y) = \infty.$$

Seega iga $p > 0$ korral $H(X|\hat{X}) = \infty$, mistõttu $H(X|\hat{X}) \not\to 0$, kui $P_e \to 0$.

Millal on Fano võrratus võrdus? Võrratuse tõestusest on näha, et võrdus kehtib parajasti siis, kui iga $x \in \mathcal{X}$ korral

$$H(X|\hat{X} = x, E = 1) = \log(|\mathcal{X}| - 1) \tag{1.29}$$

ning

$$H(E|\hat{X}) = H(E). \tag{1.30}$$

Seos (1.29) tähendab, et vektori X tinglik jaotus tingimusel, et $X \neq \hat{X} = x$ on ühtlane üle ülejäänud tähtede $\mathcal{X}\backslash x$. See aga tähendab, et leidub p_i nii, et iga $x_i \in \mathcal{X}$ korral

$$\mathbf{P}(\hat{X} = x_i, X = x_j) = p_i, \quad \forall j \neq i.$$

Teisisõnu, vektori $(\hat{X}, X)$ ühisjaotuse tabelis

$\hat{X}\backslash X$	x_1	x_2	$\cdots$	x_n
x_1	$\mathbf{P}(\hat{X} = x_1, X = x_1)$	$\mathbf{P}(\hat{X} = x_1, X = x_2)$	$\cdots$	$\mathbf{P}(\hat{X} = x_1, X = x_n)$
x_2	$\mathbf{P}(\hat{X} = x_2, X = x_1)$	$\mathbf{P}(\hat{X} = x_2, X = x_2)$	$\cdots$	$\mathbf{P}(\hat{X} = x_2, X = x_n)$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_n	$\mathbf{P}(\hat{X} = x_n, X = x_1)$	$\cdots$	$\cdots$	$\mathbf{P}(\hat{X} = x_n, X = x_n)$

on igas reas väljaspool peadiagonaali kõik elemendid võrdsed.
Seos (1.30) kehtib, kui iga $x \in \mathcal{X}$ korral $P(X = x|\hat{X} = x) = 1 - P_e$ ehk iga rea peadiagonaali elemendi suhe rea summase on võrdne $1 - P_e$. Selline jaotustabel on näiteks

$\hat{X}\backslash\mathcal{X}$	a	b	a
a	$\frac{3}{10}$	$\frac{1}{10}$	$\frac{1}{10}$
b	$\frac{1}{25}$	$\frac{3}{25}$	$\frac{1}{25}$
c	$\frac{3}{50}$	$\frac{3}{50}$	$\frac{9}{50}$

.

Ülaltoodud ühisjaotuse korral $P_e = \frac{2}{5}$, $\log(|\mathcal{X}| - 1) = 1$, millest

$$P_e \log(|\mathcal{X}| - 1) + h(P_e) = \frac{2}{5} + \frac{3}{5}\log\frac{5}{3} + \frac{2}{5}\log\frac{5}{2} = \frac{3}{5}\log\frac{5}{3} + \frac{2}{5}\log 5.$$

Teisest küljest aga

$$H(X|\hat{X} = a) = H(X|\hat{X} = b) = H(X|\hat{X} = c) = \frac{3}{5}\log\frac{5}{3} + \frac{2}{5}\log 5,$$

millest

$$H(X|\hat{X}) = \frac{3}{5}\log\frac{5}{3} + \frac{2}{5}\log 5.$$

Seega on Fano võrratus võrdus.

1.8 Juhusliku protsessi entroopiamäär

Käesolevas alajaotuses vaatleme juhuslikku protsessi $\{X_n\}_{n=1}^{\infty}$.

Def 1.13 *Juhusliku protsessi* $\{X_n\}_{n=1}^{\infty}$ **entroopiamäär (entropy rate)** *on*

$$H_X := \lim_{n\to\infty} \frac{1}{n} H(X_1, \ldots, X_n),$$

kui piirväärtus eksisteerib.

Näited:

- Olgu $\{X_n\}_{n=1}^{\infty}$ i.i.d. juhuslikud suurused jaotusest P, s.t. $X_i \sim P$. Siis

  $$\lim_{n\to\infty} \frac{1}{n} H(X_1, \ldots, X_n) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(X_i) = \lim_{n\to\infty} H(P).$$

  Seega on i.i.d. protsessil entroopiamäär defineeritud, see võrdub jaotuse P entroopiaga.

- Olgu $\{X_n\}_{n=1}^{\infty}$ sõltumatud juhuslikud suurused. Siis

  $$\frac{1}{n} H(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} H(X_i).$$

  Selline rida ei pruugi alati koonduda ja siis pole protsessi entroopiamäär defineeritud.

- Olgu $X_1, X_2, \ldots$ i.i.d. juhuslikud suurused, $X_i \sim P$. Vaatleme juhuslikku ekslemist, $\{S_n\}_{n=0}^{\infty}$, s.t.

  $$S_0 = 0,\ S_1 = X_1,\ S_2 = X_1 + X_2, \ldots, S_n = X_1 + \cdots + X_n.$$

  Juhusliku ekslemise entroopia on $H_S = H(P)$ (ülesanne).

Vaatleme piirväärtust

$$H'_X := \lim_n H(X_n|X_1, \ldots, X_{n-1}),$$

mis muidugi ei pruugi alati eksisteerida. Järgnevas näeme, et statsionaarsete protsesside korral H'_X alati eksisteerib ning see on võrdne protsessi entroopiamääraga H_X. Tuletame meelde statsionaarse protsessi definitsiooni.

Def 1.14 *Juhuslik protsess* $\{X_n\}_{n=1}^{\infty}$ *on* **statsionaarne (stationary)**, *kui iga* $n \geq 1$ *ja iga* $k \geq 1$ *korral on juhuslikud vektorid*

$$(X_1, \ldots, X_n) \text{ ja } (X_{k+1}, \ldots, X_{k+n})$$

ühe ja sama jaotusega.

Kui $\{X_n\}_{n=1}^{\infty}$ on statsionaarne protsess, siis on juhuslikud suurused $X_1, X_2, \ldots$ sama jaotusega, juhuslikud vektorid (X_1, X_2), (X_2, X_3), $\ldots$ on sama jaotusega, juhuslikud vektorid (X_1, X_2, X_3), (X_2, X_3, X_4), $\ldots$ on sama jaotusega, jne.

Väide 1.16 *Kui* $\{X_n\}_{n=1}^{\infty}$ *on statsionaarne protsess, siis* H'_X *on alati defineeritud.*

Tõestus. Et $\{X_n\}_{n=1}^{\infty}$ on statsionaarne, siis iga n korral on juhuslikud vektorid $(X_1, \ldots, X_n)$ ja $(X_2, \ldots, X_{n+1})$ sama jaotusega. Sellest järeldub, et iga n korral

$$H(X_n|X_1, \ldots, X_{n-1}) = H(X_{n+1}|X_2, \ldots, X_n).$$

Seega

$$H(X_{n+1}|X_1, \ldots, X_n) \leq H(X_{n+1}|X_2, \ldots, X_n) = H(X_n|X_1, \ldots, X_{n-1}),$$

millest saame, et $\{H(X_n|X_1, \ldots, X_{n-1})\}$ on mittenegatiivne ja mittekasvav jada ning sellisel jadal on piirväärtus. ■

Järgnevas tõestame, et statsionaarse protsessi entroopiamäär on alatu defineeritud ja see võrdub H'_X. Tõestuses kasutame Cesaro lemmat.

Lemma 1.4 (Cesaro lemma) *Olgu* $\{a_n\}$ *mittenegatiivsete reaalarvude jada, kusjuures* $a_1 > 0$ *ja* $\sum_n a_n = \infty$. *Tähistame* $b_n := \sum_{i=1}^{n} a_i$. *Olgu* $x_n \to x$ *suvaline koonduv jada. Siis*

$$\frac{1}{b_n} \sum_{i=1}^{n} a_i x_i \to x, \quad \text{kui} \quad n \to \infty.$$

Juhul, kui $a_n = 1$, saame

$$\frac{x_1 + \ldots + x_n}{n} \to x.$$

Teoreem 1.15 *Kui* $\{X_n\}_{n=1}^{\infty}$ *on statsionaarne protsess, siis* H_X *on alati defineeritud, kusjuures* $H'_X = H_X$.

Tõestus. Entroopia ketireeglist saame

$$\frac{1}{n}H(X_1,\ldots,X_n) = \frac{1}{n}\sum_{k=1}^{n} H(X_k|X_1,\ldots,X_{k-1}).$$

Et $H(X_k|X_1,\ldots,X_{k-1}) \to H'_X$, siis Cesaro lemmast saame, et

$$\lim_{n\to\infty}\frac{1}{n}H(X_1,\ldots,X_n) = \lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n} H(X_k|X_1,\ldots,X_{k-1}) = H'_X.$$

■

Seega statsionaarse protsessil on entroopiamäär alati defineeritud ning lisaks definitsioonile saab selle leidmiseks kasutada ka seost $H_X = H'_X$. Ülaltoodud näidetest selgus, et ka mittestatsionaarsel protsessil võib leidida entroopiamäär (millised näidetes toodud protsessidest pole statsionaarsed?)

1.8.1 Markovi ahela entroopiamäär

Juhusliku protsessi entroopiamäära leidmine ei pruugi üldiselt olla kerge. Teatud protsesside korral (nagu näiteks i.i.d. protsess), on aga entroopiamäära lihtne leida. Alljärgnevas näeme, et ka satsionaarse Markovi ahela entroopiamäära on lihtne leida. Tuletame meelde (lõpmatu) Markovi ahela definitsiooni. Olgu $\{X_n\}_{n=1}^{\infty}$ juhuslik protsess, kusjuures juhuslikud suurused X_i võtavad väärtusi hulgal $\mathcal{X}$.

Def 1.16 *Juhuslik protsess* $\{X_n\}_{n=1}^{\infty}$ *on* **Markovi ahel**, *kui iga* $x_i \in \mathcal{X}$ *ja iga* $m \geq 1$ *korral kehtib (1.20), s.t.*

$$\mathbf{P}(X_{m+1} = x_{m+1}|X_m = x_m,\ldots,X_1 = x_1) = \mathbf{P}(X_{m+1} = x_{m+1}|X_m = x_m). \tag{1.31}$$

Märkus: Arusaadavalt on võrdus (1.31) defineeritud vaid siis, kui tinglik tõenäosus on defineeritud, s.t. $\mathbf{P}(X_m = x_m,\ldots,X_1 = x_1) > 0$.

Markovi ahelate terminoloogias nimetatakse hulka $\mathcal{X}$ ahela seisundite hulgaks, selle elemente nimetatakse Markovi ahela seisunditeks. Markovi ahel on **homogeene**, kui võrduse (1.31) parem pool ei sõltu m-st. Sellisel juhul iga m ja iga $x_i, x_j \in \mathcal{X}$ korral

$$\mathbf{P}(X_{m+1} = x_j|X_m = x_i) = P(X_2 = x_j|X_1 = x_i) =: P_{ij}.$$

Maatriksit $P = (P_{ij})$ nimetatakse homogeense MA üleminekumaatriksiks. Alljärgnevas vaatlemegi vaid homogeenset Markovi ahelat $\{X_n\}$. Olgu $\pi(i) = \pi(x_i)$ juhusliku suuruse X_1 jaotus (ütleme, et algtõenäosuste vektor). Siis $P(X_2 = x_j) = \sum_i \pi(i)P_{ij}$ ehk X_2 jaotus on $\pi^T P$. Analoogiliselt on X_3 jaotus $\pi^T P^2$ ning X_k jaotus on $\pi^T P^k$. Seega on $\{X_n\}$ jaotus määratud üleminekumaatriksi P ja algtõenäosuste vektoriga π. Markovi ahel on statsionaarne parajasti siis, kui algtõenäosuste vektor π on selline, et $\pi^T P = \pi$

ehk $\pi(j) = \sum_i \pi(i)P_{ij}$ iga j korral. Sellist vektorit nimetatakse statsionaarseks .

Näide: Olgu $|\mathcal{X}| = 2$ ning olgu üleminekumaatriks

$$\begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}.$$

Sellise üleminekumaatriksiga Markovi ahela statsionaarne algtõenäosuste vektor on

$$(\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta}).$$

Teoreem 1.17 *Olgu $\{X_n\}$ statsionaarne Markovi ahel üleminekumaatriksiga (P_{ij}) ja algtõenäosuste vektoriga π. Siis*

$$H_X = H(X_2|X_1) = -\sum_i \pi(i) \sum_j P_{ij} \log P_{ij}.$$

Tõestus. Markovi omadusest saame, et iga n korral $H(X_n|X_{n-1}, \ldots, X_1) = H(X_n|X_{n-1})$. Et ahel on statsionaarne, siis $H(X_n|X_{n-1}) = H(X_2|X_1)$ ja teoreemist 1.15 järeldub

$$H_X = H'_X = \lim_{n\to\infty} H(X_n|X_{n-1}, \ldots, X_1) = \lim_{n\to\infty} H(X_n|X_{n-1}) = H(X_2|X_1).$$

Seos

$$H(X_2|X_1) = -\sum_i \pi(i) \sum_j P_{ij} \log P_{ij}$$

on lihtne ülesanne. ■

1.9 Erinevate algjaotustega Markovi ahelad

Olgu $X_1, X_2, \ldots$ homogeene MA üleminekutõenäosustega $R(x|y)$, (st $R(x|y) = \mathbf{P}(X_n = x|X_{n-1} = y)$) ja algtõenäosustega π (st $\pi(x) = \mathbf{P}(X_1 = x)$). Olgu $X'_1, X'_2, \ldots$ sama üleminekumaatriksi kuid algjaotusega π' MA. Järgnev võrratus näitab, et sõltumata algjaotustest π ja π', juhuslike suuruste X_n ja X_{n+1} jaotused lähenevad teineteisele K-L mõttes.

Väide 1.17 *Iga $n = 1, 2, \ldots$ korral kehtib*

$$D(X_{n+1}||X'_{n+1}) \le D(X_n||X'_n). \tag{1.32}$$

Tõestus. Olgu P_n ja P'_n vastavalt X_n ja X'_n jaotused. Seega (1.32) on

$$D(P_{n+1}||P'_{n+1}) \le D(P_n||P'_n). \tag{1.33}$$

K-L ketireeglist saame

$$\begin{aligned} D\big((X_{n+1}, X_n)||(X'_{n+1}, X'_n)\big) &= D\big(X_{n+1}||X'_{n+1}\big) + D\big(X_n||X'_n|X_{n+1}\big) \\ &= D\big(X_n||X'_n\big) + D\big(X_{n+1}||X'_{n+1}|X_n\big). \end{aligned}$$

Veendu, et $D\big(X_{n+1}||X'_{n+1}|X_n\big) = 0$. Tõepoolest, et

$$\mathbf{P}(X_{n+1} = x|X_n = y) = \mathbf{P}(X'_{n+1} = x|X'_n = y) = R(x|y),$$

siis tähistades

$$P(y) = \mathbf{P}(X_n = y), \quad P(x,y) = \mathbf{P}(X_{n+1} = x, X_n = y), \quad P'(x,y) = \mathbf{P}(X'_{n+1} = x, X'_n = y),$$

saame

$$D\big(X_{n+1}||X'_{n+1}|X_n\big) = \sum_y P(y) \sum_x P(x|y) \log \frac{P(x|y)}{P'(x|y)} = \sum_y P(y) \sum_x P(x|y) \log \frac{R(x|y)}{R(x|y)} = 0.$$

■

Järeldus 1.10 *Kui π' on statsionaarne algjaotus, siis (1.32) on*

$$D(P_{n+1}||\pi') \leq D(P_n||\pi'). \tag{1.34}$$

Seega X_n jaotus P_n läheneb statsionaarsele jaotusele K-L mõttes. Mittenegatiivsete liikmentega mittekahaneval jadal $\{D(P_n||\pi')\}$ on piirväärtus. Juhuslike protsesside teooriast teame, et taandumatu ja mitteperioodilise MA korral $P_n(x) \to \pi'(x)$, $\forall x \in \mathcal{X}$. Kui $\mathcal{X}$ on lõplik, siis sellest järeldub ka koondumine $D(P_n||\pi') \to 0$.

Järeldus 1.11 *Kui statsionaarne algjaotus π' on ühtlane üle lõpliku tähestiku $\mathcal{X}$, siis (1.34) on*

$$H(P_n) \leq H(P_{n+1}) \tag{1.35}$$

Tõestus. Ülesanne 26. ■

Seega ühtlase algjaotuse korral on juhuslike suuruste $X_1, X_2, \ldots$ entroopia mittekahanev.

Näide. Olgu kaardipakis m kaarti: $\{1, \ldots, m\}$. Seega on kaardipakil $m!$ võimalikku seisundit. Kaardipaki segamist võib vaadelda Markovi ahelana. Pole raske veenduda, et sellise Markovi ahela üleminekumaatriks on selline, et ka veergude summa on üks. Seetõttu on statsionaarne jaotus ühtlane. Seega kaardipaki piirjaotus on ühtlane (see ongi segamise mõte, mitteühtlase piirjaotuse korral oleksid mõned kaardid teatud positsioonidel suurema tõenäosusega). Kaardipaki segamine seega suurendab selle entroopiat.

1.10 Ülesanded

1. Olgu mündiviskel kulli saamise tõenäosus p. Münti vistatakse kuni esimese kullini. Olgu X selleks kulunud visete arv. Leida $H(X)$.

2. Tõestada *grupeerimisomadus*

$$H(p_1, p_2, p_3, \ldots) = H(p_1 + p_2, p_3, \ldots) + (p_1 + p_2) H\Big(\frac{p_1}{(p_1 + p_2)}, \frac{p_2}{(p_1 + p_2)}\Big)$$

   ja järeldada sellest (1.3).

3. Leida selline $P(y|x)$ ja $P_1(x)$ ja $P_2(x)$ nii, et $P_1 \neq P_2$, kuid $P_1(y) = P_2(y)$ iga $y \in \mathcal{Y}$ korral.

4. Olgu $g : \mathcal{X} \to \mathcal{X}$ funktsioon. Tõestada, et

$$H(g(X)) \leq H(X), \quad H(g(X)|Y) \leq H(X|Y).$$

5. Leida P nii, et $H(P) = \infty$.

6. Olgu X_1 ja X_2 juhuslikud suurused väärtuste hulgaga vastavalt $\mathcal{X}_1 = \{1, \ldots, m\}$, $\mathcal{X}_2 = \{m+1, \ldots, n\}$. Olgu X segujaotusega, s.t.

$$X = \begin{cases} X_1 & \text{kui } Z = 1, \\ X_2 & \text{kui } Z = 0, \end{cases}$$

kus $Z \sim B(1, p)$. Leida $H(X)$. Veendu, et

$$2^{H(X)} \leq 2^{H(X_1)} + 2^{H(X_2)}.$$

7. Olgu $X \sim P$. Tõestada, et

$$\mathbf{P}\big(P(X) \leq d\big)(\log \frac{1}{d}) \leq H(X).$$

8. Leida jaotused P, Q ja R nii, et

$$D(P\|Q) > D(P\|R) + D(R\|Q).$$

9. Tõestada võrdused (1.11).

10. Olgu P ja Q tähestikul $\mathcal{X}$ antud jaotused. Olgu

$$\chi^2(P||Q) := \sum_{x \in \mathcal{X}} \frac{(P(x) - Q(x))^2}{Q(x)}.$$

Tõesta, et

$$\chi^2(P||Q) \geq (\ln 2) D(P||Q).$$

11. Olgu P_n defineeritud kui (1.13). Näita, et

$$H(P_n) \to H(P) + c.$$

12. Olgu $\mathcal{X}$ lõpmatu,

$$P_n = (1 - \frac{\alpha}{\log n}, \underbrace{\frac{\alpha}{n \log n}, \ldots, \frac{\alpha}{n \log n}}_{n}, 0, \cdots),$$

kus $\alpha > 0$. Veendu, et $P_n \to P$, kus $P = (1, 0, \ldots)$, kuid $H(P_n) \to \alpha$. Olgu nüüd

$$Q = (q_1, q_2, q_3, \ldots),$$

kus $q_i = (1-q)q^{i-1}$. Näita, et $D(P||Q) < \infty$, kuid

$$D(P_n||Q) \to \infty.$$

13. Olgu $X = (X_1, \ldots, X_n)$ binaarsete komponentidega juhuslik vektor. Olgu $R = (R_1, \ldots, R_n)$ vektori X blokipikkuste indikaator. Näiteks, kui $X = (1, 0, 0, 0, 1, 1, 0)$, siis $R = (1, 3, 2, 1)$. Näidata, et

$$0 \leq H(X) - H(R) \leq \min_i H(X_i).$$

14. Olgu X, Y juhuslikud suurused, olgu $Z = X + Y$.
Näita, et $H(Z|X) = H(Y|X)$ ning veendu, et kui X ja Y on sõltumatud, siis $H(X) \leq H(Z)$ ja $H(Y) \leq H(Z)$.
Leida X ja Y nii, et $H(X) > H(Z)$ ja $H(Y) > H(Z)$.
Millal kehtib $H(Z) = H(X) + H(Y)$?

15. Olgu

$$\rho(X, Y) = H(X|Y) + H(Y|X).$$

Tõesta, et ρ on poolmeetrika. Millal $\rho(X, Y) = 0$?
Veendu, et

$$\rho(X, Y) = H(X) + H(Y) - 2I(X; Y) = H(X, Y) - I(X; Y) = 2H(X, Y) - H(X) - H(Y).$$

16. Tõestada, et iga $n \geq 2$ korral

$$H(X_1, \ldots, X_n) \geq \sum_{i=1}^{n} H(X_i | X_j, j \neq i).$$

Veenduda, et

$$\frac{1}{2}[H(X_1, X_2) + H(X_3, X_2) + H(X_1, X_3)] \geq H(X_1, X_2, X_3).$$

17. Olgu X, Y, Z juhuslikud suurused, kusjuures Y ja Z on sõltumatud. Tõesta, et

$$D(X||Y|Z) = -H(X|Z) + D(X||Y) + H(X) \leq H(Z) + D(X||Y).$$

18. Tõesta, et $D\big((X, f(X))||(Y, f(Y))\big) = D(X||Y)$. Järelda sellest, et $D\big(f(X)||f(Y)\big) \leq D(X||Y)$. Veendu, et üldiselt $D\big((X, f(X))||(Y, g(Y))\big) \neq D(X||Y)$.

19. (a) Olgu X_1 ja X_2 sama jaotusega juhuslikud suurused. Olgu

$$\rho(X_1, X_2) := 1 - \frac{H(X_2|X_1)}{H(X_1)}. \tag{1.36}$$

Tõestada, et ρ on sümmeetriline, $\rho \in [0, 1]$. Millal on $\rho = 0$? Millal on $\rho = 1$?

(b) Olgu (X, Y) jaotustabel järgmine $\epsilon \in (0, \frac{1}{4}]$:

$Y \backslash X$	$-n$	-1	1	n
n	0	0	0	ϵ
1	0	$\frac{1}{4} - \epsilon$	$\frac{1}{4}$	0
-1	0	$\frac{1}{4}$	$\frac{1}{4} - \epsilon$	0
$-n$	ϵ	0	0	0

Leida $I(X;Y)$ ning ρ (nagu seoses (1.36)). Leida $\mathrm{cov}(X, Y)$ ja X ning Y korrelatsioonikordaja. Veendu, et kui $n \to \infty$, siis korrelatsioonikordaja piirväärtus on 1 iga $\epsilon > 0$ korral.

(c) Olgu (X, Y) jaotustabel järgmine

$Y \backslash X$	$-n$	-1	1	n
n	0	0	$\frac{1}{4}$	0
1	$\frac{1}{4}$	0	0	0
-1	0	0	0	$\frac{1}{4}$
$-n$	0	$\frac{1}{4}$	0	0

Leida $I(X;Y)$ ning ρ (nagu seoses (1.36)). Leida $\mathrm{cov}(X, Y)$ ja X ning Y korrelatsioonikordaja.

20. Tõestada, et

$$\begin{aligned}
I(X;X|Z) &= H(X|Z) \\
I(X;Y|Z) &= H(Y|Z) - H(Y|X,Z) \\
I(X;Y|Z) &= H(X|Z) + H(Y|Z) - H(X,Y|Z) \\
I(X;Y|Z) &= H(X,Z) + H(Y,Z) - H(X,Y,Z) - H(Z).
\end{aligned}$$

21. Tõestada, et

$$\begin{aligned}
H(X,Y|Z) &\geq H(X|Z) \\
I(X,Y;Z) &\geq I(X;Z) \\
H(X,Y,Z) - H(X,Y) &\leq H(X,Z) - H(X) \\
I(X;Y|Z) &\geq I(Y;Z|X) - I(Y;Z) + I(X;Y).
\end{aligned}$$

Millal kehtivad võrdused?

22. Leida X, Y, Z nii, et

$$I(X;Y|Z) > I(X;Y) = 0$$
$$0 = I(X;Y|Z) < I(X;Y).$$

23. Tõestada, et

$$H(X|g(Y)) \geq H(X|Y).$$

Leida vektor (X, Y) nii, et X ja Y pole sõltumatud, g pole üksühene funktsioon, kuid ülaltoodud võrratus on võrdus.

24. Olgu $X = (X_1, \ldots, X_n)$ binaarsete komponentidega juhuslik vektor, kusjuures X jaotus on järgmine:

$$P(x_1, \ldots, x_n) = \begin{cases} 2^{-(n-1)} & \text{kui } \sum_i x_i \text{ on paarisarv;} \\ 0, & \text{kui } \sum_i x_i \text{ on paaritu arv.} \end{cases}$$

Leida X_i jaotus. Leida (X_i, X_{i+1}) jaotus. Leida

$$I(X_1;X_2), I(X_2;X_3|X_1), I(X_4;X_3|X_1,X_2), \ldots, I(X_n;X_{n-1}|X_1,X_2,\ldots,X_{n-2}).$$

25. Tõestada, et kui $X \to Y \to Z$, siis $H(X|Z) \geq H(X|Y)$, $I(X;Z) \leq I(Y;Z)$ ja $I(X;Y|Z) \leq I(X;Y)$.

26. Olgu $\{P_\theta\}$ Bernoulli jaotuste hulk, $\theta \in \Theta$, kus Θ on mingi ülimalt loenduv hulk, π on parameetri eeljaotus. Olgu X juhuslik valim ja $T(X) = \sum_{i=1}^n X_i$. Leida $H(\theta|T(X))$ ja $H(\theta|X)$. Veenduda, et informatsioonivõrratus on võrdus.

27. Olgu $X_1 \to X_2 \to X_3 \to X_4$. Tõestada, et

$$I(X_1;X_4) \leq I(X_2;X_3).$$

28. Olgu $X_1 \to X_2 \to \cdots \to X_n$. Leida $I(X_1; X_2, X_3, \ldots, X_n)$.

29. Oletame, et $X_1 \to X_2 \to X_3$ on Markovi ahel, kusjuures $|\mathcal{X}_1| = n$, $|\mathcal{X}_2| = k$, $|\mathcal{X}_3| = m$, kusjuures $k < n$ ja $k < m$. Tõestada, et "pudelikael" vähendab vastastikust informatsiooni juhuslike suuruste X_1 ja X_3 vahel, s.t. $I(X_1;X_3) \leq \log k$. Järeldada, et $k = 1$ korral ei saa X_3 kuidagi sõltuda X_3-st.

30. Olgu X juhuslik suurus lõpliku väärtuste hulgaga, s.t. $|\mathcal{X}| = m$. Leida väikseima veatõenäosusega mittejuhuslik hinnang juhuslikule suurusele X. Olgu P_e vea tõenäosus, s.t. $P_e = \mathbf{P}(X \neq \hat{X})$. Millise X jaotuse korral on Fano võrratus võrdus

$$H(X) = P_e \log(|\mathcal{X}| - 1) + h(P_e)?$$

31. Olgu P jaotus väärtuste hulgaga $1, 2, \ldots$. Olgu selle mõõdu keskväärtus μ. Tõestada, et
$$H(P) \leq \mu \log \mu - (\mu - 1) \log(\mu - 1),$$
kusjuures võrratus on võrdus parajasti siis, kui P on geomeetrilise jaotusega. Seega fikseeritud keskväärtuse korral on geomeetriline jaotus suurima entroopiaga.

32. **a)** Tõestada Järeldus 1.11

    b) Olgu $X_0 \to \cdots \to X_n$. Tõestada, et
$$H(X_0|X_1) \leq H(X_0|X_2) \leq H(X_0|X_3) \leq \cdots \leq H(X_0|X_n).$$

33. Olgu $\{X_n\}_{n=1}^{\infty}$ statsionaarne juhuslik protsess. Tõestada, et
$$\frac{H(X_1, \ldots, X_n)}{n} \leq \frac{H(X_1, \ldots, X_{n-1})}{n-1}$$
$$\frac{H(X_1, \ldots, X_n)}{n} \geq H(X_n|X_1, \ldots, X_{n-1}).$$

34. Tõestada, et statsionaarse MA korral
$$H(X_2|X_1) = -\sum_i \pi(i) \sum_j P_{ij} \log P_{ij}.$$

35. Olgu $X_1, X_2, \ldots$ i.i.d. juhuslikud suurused, $X_i \sim P$. Vaatleme juhuslikku ekslemist, $\{S_n\}_{n=0}^{\infty}$, s.t.
$$S_0 = 0,\ S_1 = X_1,\ S_2 = X_1 + X_2, \ldots, S_n = X_1 + \cdots + X_n.$$
Tõestada, et juhusliku ekslemise entroopia on $H_S = H(P)$.

36. Koer liigub juhuslikult täisarvudel: ajahetkel 0 on koer positsioonil 0. Seejärel hakkab ta tõenäosusega 0.5 liikuma paremale ja samasuure tõenäosusega vasakule. Pärast esimest sammu jätkab ta liikumist esialgses suunas tõenäosusega 0.9, tõenäosussega 0.1 vahetab ta suunda jne. Seega on koera tüüpiline trajektoor näiteks
$$(X_0, X_1, \ldots) = (0, -1, -2, -3, -4, -3, -2, -1, 0, 1, 2, 3, \ldots).$$
Leida H_X.

37. Vaatleme juhuslikku ekslemist ringil $(0, 1, \ldots, l)$, s.t. l-le järgneb 0. Olgu
$$S_n = \sum_{i=1}^{n} X_i,$$
kusjuures X_1 on ühtlase jaotusega juhuslik suurus, $X_2, X_3, \ldots$ on i.i.d. juhuslikud suurused $P(X_2 = 1) = P(X_2 = 2) = 0.5$. Leida H_S.

2 Kodeerimine

2.1 Põhimõisted

Vaatleme tähestikku $\mathcal{X}$. Oletame, et informatsiooni edasiandmiseks on meie käsutuses kanal, mille kaudu saab edastada vaid sümboleid etteantud lõplikust *kodeerimistähestikust* $\mathcal{D}$. Kui $D := |\mathcal{D}| < |\mathcal{X}|$ (ja sellist olukorda vaatlemegi), tuleb iga tähestiku $\mathcal{X}$ täht esitada kodeerimistähtede lõpliku stringina - *koodisõnana*. Teisisõnu, tähestik $\mathcal{X}$ tuleb kodeerida. Näiteks kui $\mathcal{D} = \{0, 1\}$, tuleb iga tähestiku $\mathcal{X}$ element kodeerida mingiks bitisõnaks.
Olgu $\mathcal{D}^*$ kõikide kooditähtedest moodustatud lõplike sõnade hulk. Olgu $\mathcal{X}^*$ kõikide tähtedest moodustatud lõplike sõnade hulk. Formaalselt

$$\mathcal{D}^* := \cup_{n=1}^{\infty} \mathcal{D}^n, \quad \mathcal{X}^* := \cup_{n=1}^{\infty} \mathcal{X}^n.$$

Def 2.1 **Kood (code)** *on kujutis*

$$C : \mathcal{X} \to \mathcal{D}^*.$$

Koode on väga palju ning väga erinevate omadustega. Näiteks on kood Morse tähestik, mille korral hulga $\mathcal{X}$ moodustavad tähestik, numbrid ja kirjavahemärgid, kodeerimistähestik $\mathcal{D}$ koosned kolmest elemendist: punkt, kriips ja paus (tegelikult kuulub Morse kodeerimistähestikku ka pikk paus sõnavahedeks, kuid ülalkirjeldatud tähestiku kodeerimiseks pole seda vaja).

Def 2.2 *Kood C on* **ühene (non-singular)**, *kui ta on injektiivne, s.t. $C(x_i) \neq C(x_j)$ iga $x_i \neq x_j \in \mathcal{X}$ korral.*

Ühene kood kodeerib tähestiku üheselt. Sellest üksi ei piisa aga, et üheselt kodeerida mitmest tähest koosnevat sõna $x_1 x_2 \cdots x_n$.

Olgu C kood. Defineerime tema *laiendi*

$$C^* : \mathcal{X}^* \to \mathcal{D}^*, \quad C^*(x_1 \cdots x_n) := C(x_1) \cdots C(x_n).$$

Def 2.3 *Kood C on* **üheselt dekodeeritav (uniquely decodable)**, *kui tema laiend C^* on ühene.*

Üheselt dekodeeritava koodi korral vastab koodisõnale $C(x_1) \cdots C(x_n)$ vaid üks originaalsõna $x_1 \cdots x_n$. Küll aga võib olla nii, et esimese tähe x_1 dekodeerimiseks tuleb lugeda kogu kodeeritud sõna $C(x_1) \cdots C(x_n)$. On aga loomulik eeldada, et kood C on selline, et täht x_1 on dekodeeritud niipea kui see saab loetud (s.t. dekodeerimine toimub "on-line"). Sellisel juhul ei tohi tähe x_1 kood $C(x_1)$ olla ühegi teise tähe koodi algus (vastasel juhul ei teaks me, kas $C(x_1)$ on x_1 kood või järgneb veel midagi ning $C(x_1)$ on vaid osa mingi teise tähe koodist).

Def 2.4 *Kood C on* **prefikskood (prefix-free, instantaneous)**, *kui ei leidu erinevaid tähti x_i ja x_j nii, et tähe x_i kood $C(x_i)$ on tähe x_j koodi $C(x_j)$ algus (prefiks).*

Märkused:

- Prefikskood on üheselt dekodeeritav ja seetõttu ka ühene.
- Termini *prefikskood* asemel oleks ehk loogilisem kasutada terminit *mitteprefikskood*, kuid viimane tundub kohmakas. Inglisekeelses kirjanduses kasutatakse mõlemaid termineid: nii *prefix code* kui ka *prefix-free code*.

Näited:

- Morse tähestikus tähistab iga koodi lõppu paus. Seega on Morse tähestik prefikskood. Ilma pausideta oleks ei oleks Morse tähestik üheselt dekodeeritav.
- Olgu $\mathcal{X} = \{a, b, c, d\}$ ning vaatame kahendkoode C_1, C_2, C_3 ja C_4, millised esitame tabelina

$\mathcal{X}$	C_1	C_2	C_3	C_4
a	0	0	10	0
b	0	010	00	10
c	1	01	11	110
d	0	10	110	111

  Kood C_1 pole ühene. Kood C_2 on küll ühene, kuid pole üheselt dekodeeritav. Näiteks kodeerimissõna 010 võib tähendada nii tähte b kui ka sõnu ad ja ca. Kood C_3 on üheselt dekodeeritav kuid mitte prefikskood. Tõepoolest, saamaks teada, kas jada $1100\ldots0$ kodeerib sõna $cbb\ldots b$ või $dbb\ldots b$, peame lugema üle kõik nullid ning veenduma kas neid on paaris- või paarituarv. Järelikult ei saa me esimest tähte dekodeerida enne kui oleme kogu sõna ära lugenud. See on sellepärast nii, et koodisõna $C(c) = 11$ on koodisõna $C(d) = 110$ prefiks. Kood C_4 on aga prefikskood ning iga tähe saame dekodeerida niipea kui oleme tema koodi lugenud. Dekodeerige "on-line" string 01011111010.

2.2 Krafti võrratus

Prefikskood kui puu. Iga prefikskoodi võib esitada D-ndpuuna, kus igal sõlmel on maksimaalselt D järglast ning igale lehele vastab üks tähestiku $\mathcal{X}$ täht. Koodipuu igale oksale vsatab üks täht kooditähestikust $\mathcal{D}$ ning tee koodipuu juurest leheni ongi lehele vastava tähe kood.

Näide: Olgu $D = 3$. Konstueerige järgmise koodi puu:

a	b	c	d	e	f	g	h
1	2	010	012	02	000	001	002

Olgu C mingi kood. Olgu iga x korral $l(x) := |C(x)|$ tähele x vastava koodisõna pikkus. Ülaltoodud näites $l(a) = l(b) = 1$, $l(c) = l(d) = 3$ jne. Järjestades kõikide koodisõnade pikkused kasvavalt, saame

$$l_1 = l_2 = 1, \quad l_3 = 2, \quad l_4 = l_5 = l_6 = l_7 = l_8 = 3.$$

On selge, et kui C on prefikskood (saab esitada puuna), siis koodisõnade pikkused ei saa olla kuitahes lühikesed. Alljärgnev Krafti võrratus annab kena tõkke: suvalise prefikskoodi koodisõnade pikkused $\{l(x) : x \in \mathcal{X}\}$ on piisavalt pikad rahuldamaks teatud tingimust. Veel enam, nimetatud tingimus on piisav selleks, et leiduks vähemalt üks etteantud pikkustega prefikskood.

Teoreem 2.5 (Krafti võrratus) *Olgu $C : \mathcal{X} \to \mathcal{D}^*$ prefikskood, $l_i = l(x_i)$. Siis*

$$\sum_i D^{-l_i} \leq 1. \tag{2.1}$$

Teistpidi, olgu $\{l_i\}_{i=1}^{|\mathcal{X}|}$ täisarvud. Kui nad rahuldavad võrratust (2.1), siis leidub prefikskood $C : \mathcal{X} \to \mathcal{D}^*$ nii, et $l_i = l(x_i)$ $\forall x_i \in \mathcal{X}$.

Tõestus. Olgu $\mathcal{D} = \{0, \ldots, D-1\}$. Tõestame kõigepealt, et iga prefikskoodi sõnade pikkused rahuldavad Krafti võrratust. Seda on väga lihtne näidata juhul kui tähestik $\mathcal{X}$ on lõplik. Seega vaatame alguses juhtu, kui $|\mathcal{X}| = m < \infty$. Olgu $l^* := \max\{l_1, \ldots, l_m\} < \infty$. Esita kood D-puuna. Koodisõnal (lehel) sügavusel l_i oleks sügavusel l^* täpselt $D^{l^*-l_i}$ järglast. Erinevatele lehtedele kuuluvad (potensiaalsed) järglased sügavusel l^* on lõikumatud. Seega nende summa ei ületa tippude arvu sügavusel l^*. Et sügavusel l^* saab D-puul olla ülimalt D^{l^*} tippu, saame

$$\sum_{i=1}^{m} D^{l^*-l_i} \leq D^{l^*} \quad \Leftrightarrow \quad \sum_{i=1}^{m} D^{-l_i} \leq 1.$$

Tõestame nüüd Krafti võrratuse lõpmatu $\mathcal{X}$ korra. Vaatleme koodisõna $d_1 d_2 \cdots d_{l_i}$. Olgu $0.d_1 d_2 \cdots d_{l_i}$ reaalarv, millele vastav D-ndarv on $0.d_1 d_2 \cdots d_{l_i}$, s.t.

$$0.d_1 d_2 \cdots d_{l_i} = \sum_{j=1}^{l_i} \frac{d_j}{D^j}. \tag{2.2}$$

Vaatleme koodisõnale $d_1 d_2 \cdots d_{l_i}$ vastavat intervalli

$$[0.d_1 d_2 \cdots d_{l_i}, 0.d_1 d_2 \cdots d_{l_i} + D^{-l_i}).$$

Siia intervalli kuuluvad need reaalarvud, millele vastavad D-ndarvud algavad $0.d_1 d_2 \cdots d_{l_i}$. See on intervalli $[0, 1]$ alamintervall, tema pikkus on D^{-l_i}. Et C on prefikskood, on erinevatele koodisõnadele vastavad intervallid lõikumatud, nende intervalli pikkuste summa on seega väiksem või võrdne ühega ehk kehtib (2.1).

Teistpidi: olgu $\{l_i\}_{i=1}^{|\mathcal{X}|}$ tingimust (2.1) rahuldavad täisarvud. Sellisel juhul saab ühikintervalli jagada lõikudeks pikkustega D^{-l_i}. Tõepoolest, reastame arvud l_i nii, et $l_1 \leq l_2 \leq \cdots$. Olgu esimene intervall $[0, D^{-l_1})$, teine $[D^{-l_1}, D^{-l_1} + D^{-l_2})$ jne. Esimese intervalli – , pikkusele l_1 vastava intervalli – saame esitada kujul

$$0.\underbrace{0 \cdots 0}_{l_1},$$

kus koma järel on l_1 nulli. Selle intervalli lõpp-punkti D^{-l_1} esitus D-ndarvuna on

$$0.\underbrace{0\cdots01}_{l_1}.$$

Intervalli $[0.0\ldots0, 0.0\ldots1)$ kuuluvad *parajasti* need D-ndarvud, mille algus on $0.0\ldots0$. Järgmise intervalli – arvule l_2 vastava intervalli $[D^{-l_1}, D^{-l_1}+D^{-l_2})$ – algus- ja lõpp-punkti esitame esitame D-ndarvuna, kus komakohti on l_2 (tuleta meelde, et $l_2 \geq l_1$). Seega teise intervalli alguspunkt on

$$0.\overbrace{\underbrace{0\cdots01}_{l_1}0\cdots0}^{l_2}. \tag{2.3}$$

Sinna intervalli kuuluvad *parajasti* need arvud, mille D-nd esitus algab arvuga (2.3). Järgmise intervalli alguspunkti $D^{-l_1}+D^{-l_2}$ esitame D-ndkujul $0.d_1d_2\cdots d_{l_3}$. Paneme tähele, arvu $D^{-l_1}+D^{-l_2}$ D-ndkujus on (maksimaalselt) l_2 kohta peale koma. Et $l_3 \geq l_2$ tähendab see, et $0.d_1d_2\cdots d_{l_3}$ on sisuliselt arvu $D^{-l_1}+D^{-l_2}$ D-ndkuju ning (vajaduse korral) teatav arv 0-e. Selle intervalli lõpp-punkti saab esitada l_3-kohalise D-ndarvuna. Arvule l_i vastava intervalli algus on $D^{-l_1}+\cdots+D^{-l_{i-1}}$. Selle arvu D-ndkujus on (maksimaalselt) l_{i-1} komakohta. Et $l_i \geq l_{i-1}$, saame (vajaduse korral 0-de lisamisel) selle arvu esitada kujul D-ndkujul (2.2). Arvu $D^{-l_1}+\cdots+D^{-l_i}$ esituseks D-nd kujul läheb samuti vaja maksimaalselt l_i kohta.
Kokkuvõttes: arvule l_i vastava intervalli algus ja lõpp-punkti esitame D-nd kujul, kusjuures komakohti on l_i. Sellest piisab mõlema arvu esitamiseks. Koodi C konstrueerime nii, et arvule l_i (tähele x_i) seame vastavusse koodisõna $d_1d_2\cdots d_{l_i}$, st vastava itervalli alguspunkti komakohad. Seega iga koodisõna kuulub erinevasse intervalli. Intervallid on lõikumatud, mistõttu on saadud kood prefikskood, sest kõik need koodisõnad, millele $d_1d_2\cdots d_{l_i}$ on prefiksiks kuuluvad ühte intervalli. ■

Märkus: Edaspidi tõestame, et sama väide üldistud üheselt dekodeeritavate koodideni (Teoreem 2.11).
Alternatiivse tõestuse teisele implikatsioonile võib leida *Yeung*'i raamatust (Thm 3.1).

Näited:

- Vaatleme veelkord koodi C_4. Siin $l_1=1$, $l_2=2$, $l_3=l_4=3$. Leiame reaalarvud, millele vastavad kahendarvud on $0.d_1d_2\cdots d_{l_i}$. Saame

  $$0.0_2=0, \quad 0.10_2=0.1_2=0.5, \quad 0.110_2=0.11_2=\frac{1}{2}+\frac{1}{4}=0.75, \quad 0.111_2=\frac{1}{2}+\frac{1}{4}+\frac{1}{8}=0.875.$$

  Vastavad intervallid (tõestuse esimene pool) on

  $$[0, 0+\frac{1}{2}), [0.5, 0.5+\frac{1}{4}), [0.75, 0.75+\frac{1}{8}), [0.875, 0.875+\frac{1}{8}).$$

  Antud näite korral on Krafti võrratus võrdus.

- Teistpidi: olgu $\{1,2,3,3\}$ koodisõnade pikkused. Konstrueerime vastavate pikkustega kahendkoodi. Lihtsaim võimalus selleks on konstrueerida vastav kahendpuu. Teoreemi tõestuses kasutatud protseduur oleks aga järgmine.
  Konstrueerime intervallid

$$[0,\frac{1}{2}),[\frac{1}{2},\frac{1}{2}+\frac{1}{4}),[\frac{1}{2}+\frac{1}{4},\frac{1}{2}+\frac{1}{4}+\frac{1}{8}),[\frac{1}{2}+\frac{1}{4}+\frac{1}{8},1).$$

  Vastavad intervallid kahendkujul (komakohti on niipalju kui l_i) on

$$[0.\overbrace{0}^{1}, \quad 0.1), \quad [0.\overbrace{10}^{2}, \quad 0.11), \quad [0.\overbrace{110}^{3}, \quad 0.111), \quad [0.\overbrace{111}^{3}, \quad 1).$$

  Koodisõnad: 0, 10, 110, 111.

- Olgu koodisõnade pikkused $\{2,2,3,3\}$. Intervallid

$$[0,\frac{1}{4}),[\frac{1}{4},\frac{1}{2}),[\frac{1}{2},\frac{1}{2}+\frac{1}{8}),[\frac{1}{2}+\frac{1}{8},\frac{1}{2}+\frac{1}{8}+\frac{1}{8}).$$

  Vastavad intervallid kahendkujul (komakohti on niipalju kui l_i) on

$$[0.00,0.01),[0.01,0.10),[0.100,0.101),[0.101,0.110).$$

  Koodisõnad: 00, 01, 100, 101.

2.3 Keskmine koodipikkus ja entroopia

Vaatleme olukorda, kus tähed on juhuslikud, tähe $x \in \mathcal{X}$ tõenäosus on $P(x)$. Olgu C mingi kood ning $l(x) = |C(x)|$. Jaotusega P juhusliku tähe kodeerimiseks kulub seega keskmiselt

$$L(C) = \sum_x l(x)P(x)$$

kooditähte. Suurust $L(C)$ nimetame koodi C *keskmiseks pikkuseks*.

Näide: Vaatleme koodi C_4. Olgu $P(a) = \frac{1}{2}$, $P(b) = \frac{1}{4}$, $P(c) = P(d) = \frac{1}{8}$. Siis

$$L(C_4) = \frac{1}{2} + \frac{1}{4}2 + \frac{1}{8}3 + \frac{1}{8}3 = \frac{7}{4}.$$

Paneme tähele, et ka $H(P) = \frac{7}{4}$.

Alljärgnevas otsime prefikskoodi, mille keskmine pikkus oleks võimalikult väike, sest sellise koodi korral on (antud jaotusega) juhusliku tähe kodeerimine efektiivne. Sellist koodi (kui see eksisteerib) nimetame *optimaalseks*. Eelnevas nägime, et iga prefikskoodi korral peavad koodisõnade pikkused rahuldama Krafti võrratust ning iga seda võrratust rahuldavate pikkuste hulga korral on võimalik leida etteantud pikkustega prefikskoodi.

On ka selge, et selliseid koode on mitu (vähemalt $|\mathcal{X}|!$). Kuidas aga valida nende seast väikseima keskmise pikkusega koodi? Intuitiivselt on selge, et keskmine koodipikkus on väike, kui väikese tõenäosusega tähti kodeeritakse pikkade koodisõnadega ning lühikesed koodisõnad hoitakse tähtedele, mille tõenäosus on suur. Ka Morse tähestik on üles ehitatud sarnase printsiibi põhjal. Küll aga on Morse tähestikus sümbol "paus" kasutusel vaid koodisõna lõpu tähistusena, mistõttu seda ei saa kasutada koodisõna keskel, samuti ei saa mitut pausi kasutada kõrvuti. Seega on kooditähestikus olevas kolmest sümbolist ühe kasutamisele seatud ranged kitsendused, mistõttu kindlasti leidub Morse tähestikust väiksema keskmise pikkusega kolmendkood.

Järgnev teoreem annab alumise tõkke antud kõikide prefikskoodide keskmistele pikkustele. Selgub, et ühegi prefikskoodi keskmine pikkus ei saa olla väiksem jaotuse P entroopiast.

Teoreem 2.6 *Olgu $C : \mathcal{X} \to \mathcal{D}^*$ prefikskood. Siis*

$$L(C) \geq H_D(P),$$

kusjuures võrdus kehtib vaid siis, kui $l(x) = -\log_D P(x)$, $\forall x \in \mathcal{X}$.

Tõestus.

$$\begin{aligned} L(C) - H_D(P) &= \sum_x l(x)P(x) - \sum_x P(x) \log_D \frac{1}{P(x)} \\ &= -\sum_x P(x) \log_D D^{-l(x)} + \sum_x P(x) \log_D P(x). \end{aligned}$$

Olgu

$$c := \sum_x D^{-l(x)}, \quad R(x) := \frac{D^{-l(x)}}{c}.$$

Siis

$$L(C) - H_D(P) = \sum_x P(x) \log_D \frac{P(x)}{R(x)} - \log_D c = D(P||R) + \log_D \frac{1}{c} \geq 0,$$

sest $D(P||R) \geq 0$ ning Krafti võrratusest järeldub, et $\log_D \frac{1}{c} \geq 0$.
Ülalolev võrratus on võrdus vaid siis, kui $P = R$ ja $c = 1$. See aga kehtib parajasti siis, kui iga $x \in \mathcal{X}$ korral $P(x) = D^{-l(x)}$. Tarvilik tingimus selleks võrduseks on, et iga $x \in \mathcal{X}$ korral on $-\log_D P(x)$ täisarv. ■

Optimaalsed koodid seost (2.4) rahuldavate jaotuste korral. Eelmisest teoreemist järeldub, et kui jaotus P on selline, et

$$\log_D \frac{1}{P(x)} \in \mathbb{Z}, \quad \forall x \in \mathcal{X}, \tag{2.4}$$

siis on väikseima keskmise pikkusega koodi kerge konstrueerida: võta $l(x) = \log_D \frac{1}{P(x)}$. Nimetatud pikkused rahuldavad Krafti võrratust (võrdusena) ning vastavate pikkustega

koodi võib defineerida näiteks nii nagu Krafti võrratuse tarvilikkuse tõestuses. Selliselt konstrueeritud koodi keskmine pikkus on $H_D(P)$ ning ülaltoodud teoreemist järelduvalt on selline kood optimaalne.

Näide: Jaotus, mis rahuldab seost (2.4) on näiteks

a	b	c	d	e	f	g	h	i
$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{4}$

Pikkused on $\{l(x)\}_{x\in\mathcal{X}} = \{5,5,4,4,4,3,3,2,2\}$. Vastava kahendkoodi konstrueerimiseks on lihtsaim võimalus konstrueerida 5-astmeline kahendpuu ning hakata seda vastavalt sünapikkustele redutseerima. Teine võimalus on formaalselt järgida Krafti võrratuse tõestuses kasutatud skeemi: konstrueerida intervallid

$[0, 2^{-2}), [2^{-2}, 2^{-2}+2^{-2}), [2^{-1}, 2^{-1}+2^{-3}), [2^{-1}+2^{-3}, 2^{-1}+2^{-3}+2^{-3}),$
$[2^{-1}+2^{-2}, 2^{-1}+2^{-2}+2^{-4}), [2^{-1}+2^{-2}+2^{-4}, 2^{-1}+2^{-2}+2^{-3}),$
$[2^{-1}+2^{-2}+2^{-3}, 2^{-1}+2^{-2}+2^{-3}+2^{-4}), [2^{-1}+2^{-2}+2^{-3}+2^{-4}, 2^{-1}+2^{-2}+2^{-3}+2^{-4}+2^{-5})$
$[2^{-1}+2^{-2}+2^{-3}+2^{-4}+2^{-5}, 1).$

Vastavad intervallid kahendkujul (2.2)

$[0.00, 0.01), [0.01, 0.10), [0.100, 0.101), [0.101, 0.110), [0.1100, 0.1101), [0.1101, 0.1110),$
$[0.1110, 0.1111), [0.11110, 0.11111), [0.11111, 1).$

Kood:

a	b	c	d	e	f	g	h	i
11111	11110	1110	1101	1100	101	100	01	00

Shannon-Fano kood. Paraku ei rahulda kõik tõenäosusjaotused seost (2.4) ning selliste jaotuste korral pole ülaltoodud protseduuri võimalik rakendada. Modifitseerime seda aga nii, et arvu $\log_D \frac{1}{P(x)}$ (mis ei pruugi olla täisarv) asemel võtame koodisõna $C(x)$ pikkuseks selle ülemise täisosa s.t.

$$l(x) = \lceil \log_D \frac{1}{P(x)} \rceil. \tag{2.5}$$

On selge, et saadud koodipikkused $\{l(x)\}$ rahuldavad Krafti võrratust ning seetõttu leidub vastavate pikkustega prefikskood C. Kirjeldatud protseduuri abil saadud koodi nimetatakse **Shannon-Fano** koodiks. Teisisõnu on kood C Shannon-Fano kood parajasti siis, kui iga tähe $x \in \mathcal{X}$ korral kehtib (2.5).

Kui palju me aga sellise ümardamise kaudu kaotame keskmises koodipikkuses? Et

$$\lceil \log_D \frac{1}{P(x)} \rceil < \log_D \frac{1}{P(x)} + 1,$$

siis

$$L(C) = \sum_x P(x)\lceil \log_D \frac{1}{P(x)} \rceil < \sum_x P(x) \log_D \frac{1}{P(x)} + 1 = H_D(P) + 1.$$

Seega kehtib järeldus.

Järeldus 2.1 *Alati leidub prefikskood $C : \mathcal{X} \to \mathcal{D}^*$ nii, et*

$$H_D(P) \le L(C) < H_D(P) + 1.$$

Näide: Olgu P ühtlane üle viie tähe, s.t. $P(x_i) = \frac{1}{5}$, $i = 1, \ldots, 5$. Siis

$$l(x) = \log \frac{1}{P(x)} = \log 5 \text{ ja } \lceil \log \frac{1}{P(x)} \rceil = 3.$$

Üks võimalik Shannon-Fano kood:

$$C(x_1) = 000, \quad C(x_2) = 001, \quad C(x_3) = 010, \quad C(x_4) = 011, \quad C(x_5) = 110. \tag{2.6}$$

Sellise koodi keskmine pikkus on 3. Seega kehtib

$$H(P) = \log 5 < L(C) = 3 < \log 10 = H(P) + 1.$$

On aga küllaltki lihtne konstrueerida lühema keskmise pikkusega kahendkoodi koodipikkustega $\{3, 3, 2, 2, 2\}$ (kuidas?). Sellise koodi keskmine pikkus on $\frac{12}{5} = 2.4$.

2.3.1 Valesti hinnatud tõenäosused

Shannon-Fano koodi konstrueerimiseks on vaja teada tähtede tõenäosusjaotust P. Oletame aga, et oleme konstrueerinud Shannon-Fano koodi hoopis jaotuse Q abil, s.t. meie käsutuses olev informatsioon tähtede jaotuse kohta on ebatäpne. On selge, et sellisel juhul on meil üsna vähe lootust saada optimaalsest või sellele suhteliselt lähedast jaotust. Järgnev teoreem väidab, et jaotuse Q põhjal konstrueeritud Shannon-Fano kahendkoodi keskmine pikkuse alumine tõke pole mitte entroopia $H(P)$ vaid $H(P) + D(P\|Q)$, ülemine tõke on pole mitte $H(P) + 1$ vaid $H(P) + D(P\|Q) + 1$. Kui Q ei erine K-L mõttes palju tähtede tegelikust jaotusest P, käitub Q põhjal konstrueeritud Shannon-Fano kahendkoodi keskmine pikkus sarnaselt P põhjal konstrueeritud Shannon-Fano kahendkoodi keskminse pikkusega.

Teoreem 2.7 *Olgu P tähtede tegelik jaotus. Olgu*

$$l_Q(x) := \lceil \log \frac{1}{Q(x)} \rceil.$$

Kehtib

$$H(P) + D(P\|Q) \le \sum_x l_Q(x)P(x) < H(P) + D(P\|Q) + 1. \tag{2.7}$$

Tõestus. Ülemise tõkke leiame järgnevalt

$$\begin{aligned}\sum_x l_Q(x)P(x) &= \sum_x \lceil \log \frac{1}{Q(x)} \rceil P(x) < \sum_x P(x)\Big(\log \frac{1}{Q(x)} + 1\Big) \\ &= \sum_x P(x)\Big(\log \frac{P(x)}{Q(x)} + \log \frac{1}{P(x)} + 1\Big) \\ &= D(P\|Q) + H(P) + 1.\end{aligned}$$

Alumise tõkke leidmine on ülesanne 1. ■

2.4 Huffmani kood

2.4.1 Huffmani koodi konstrueerimine

Shannon-Fano meetod andis üsna hea keskmise pikkusega prefikskoodi; kui jaotus P rahuldab seost (2.4), on Shannon-Fano kood optimaalne. Käesolevas osas kirjeldame aga protseduuri, mis lõpliku tähestiku $\mathcal{X}$ korral alati garanteerib optimaalse koodi. Selle protseduuri abil saadud koode nimetatakse **Huffmani koodideks**.

Näide: Olgu $\mathcal{X} = \{a, b, c, d, e\}$. Jaotus P olgu

a	b	c	d	e
0.35	0.1	0.15	0.2	0.2

Olgu $D = 2$. Tuletame meelde, et iga prefikskood on esitatav puuna, kus lehtedele vastavad tähestiku $\mathcal{X}$ tähed. Seega on kahendkoodi konstrueerimine sisuliselt kahendpuu konstrueerimine. Huffmani protseduur puu leidmiseks on järgnev: leia kaks kõige väiksema tõenäosusega tähte ja ühenda nad eelviimasel tasemel. Antud näite korral ühenda tähed b, c. Summeeri vastavad tõenäosused, antud juhul siis 0.1 ja 0.15 ning vaata vähendatud tähestikku $\{a, \{b, c\}, d, e\}$ tõenäosustega vastavalt $0.35, 0.25, 0.2, 0.2$. Saame n.n. vähendatud jaotuse

a	$\{b, c\}$	d	e
0.35	0.25	0.2	0.2

Nüüd leia järgmised kaks kõige väiksema tõenäosusega tähte, antud juhul d ja e ja ühenda nad uueks täheks. Nii vähendame eelmist jaotust veel ühe tähe võrra ning uus jaotus on järgmine

a	$\{b, c\}$	$\{d, e\}$
0.35	0.25	0.4

Otsi jälle kaks kõige väiksema tõenäosusega tähte ja ühenda need järgmisel tasemel. Saad uue tähestiku $\{a, b, c\}, \{d, e\}$ ja uue jaotuse

$\{a, b, c\}$	$\{d, e\}$
0.6	0.4

Nimetatud tähestikus on vaid kaks tähte, mis ühinevad puu esimesel tasemel. Saad kahendpuu, mille iga hargnemine tähista 0 ja 1-ga. Tee juurest leheni ongi vastava tähe (igale lehele vastab täht) kood. Näiteks saame koodi C, kus

$$C(a) = 00 \quad C(b) = 010 \quad C(c) = 011 \quad C(d) = 10 \quad C(e) = 11.$$

Selle koodi keskmine pikkus $L(C) = 2\frac{3}{4} + 3\frac{1}{4} = \frac{9}{4} = 2.25$. Jaotuse P entroopia on

$$H(P) = -0.35\log(0.35) - 0.1\log(0.1) - 0.15\log(0.15) - 0.4\log(0.4) = 2.202.$$

Kui väikseimate tõenäosustega paar pole ühene, vali Huffmani protseduuris suvaline neist. Lühima pikkusega koodi annab iga valik.

Ülaltoodud näide kirjeldas kahendkoodi (kahendpuu) konstrueerimist Huffmani meetodil. D-ndkoodi konstrueerimine käib põhimõtteliselt sama moodi: igal sammul ühenda D väikseima tõenäosusega tähte ning liida vastavad tõenäosused. Kui selline protseduur jõuab lõpuni $k+1$ sammuga, on konstrueeritud puus $k+1$ sõlme ja $k(D-1)+D$ lehte. Seega peab tähestikus olema $k(D-1)+D$ tähte. Kui see aga nii ei ole, peame tähestikku lisama sobival hulgal (mitte rohkem kui $D-2$) pseudotähti, mille tõenäosus on 0. Selliste tähtede lisamine ei muuda jaotust P, küll aga võimaldab läbi viia Huffmani protseduuri nii, et viimasel saamul ühendatakse D tähte. Paneme tähele, et pseudotähtede mittelisamine ja protseduuri läbiviimine nii, et viimasel sammul ühendatakse vähem kui D tähte võib oluliselt suurendada koodi keskmist pikkust.

Näited:

- Olgu jaotus P ja tähestik $\mathcal{X}$ järgmine

a	b	c	d	e	f
0.25	0.25	0.2	0.1	0.1	0.1

  Olgu $D = 3$. Et $6 \neq 3 + k(3-1)$, siis peame lisama ühe pseudotähe. Uus tabel on järgmine

a	b	c	d	e	f	$*$
0.25	0.25	0.2	0.1	0.1	0.1	0

  Huffmani koodi produtseerime nüüd järgmiselt: esimesel sammul ühendame tähed e, f ja $*$; järgmisel sammul ühendame $\{e, f, *\}$, d ja c; ülejärgmisel sammul ühendame $\{c, d, e, f, *\}$, b ja a.

  Huffmani kood:

  $C(a) = 1,\ C(b) = 2,\ C(c) = 01,\ C(d) = 02,\ C(e) = 000,\ C(f) = 001,\ C(*) = 002.$

- Vaatleme veelkord kõige esimest näidet. Olgu $D = 4$. Et $|\mathcal{X}| = 5$, pole tähtede arv võrdne arvuga $k(D-1)+D$ (mitte ühegi k korral). Lisades 2 pseudotähte, saame $|\mathcal{X}| = 7 = (D-1)+D$. Uus jaotus on

a	b	c	d	e	$*$	$*$
0.35	0.1	0.15	0.2	0.2	0	0

  Esimesel sammul võtame kokku tähed $d, e, *, *$; teisel sammul kõik ülejäänud.

  Huffmani kood:

$$C(a) = 0,\ C(b) = 1,\ C(d) = 2,\ C(e) = 30,\ C(f) = 31,\ C(*) = 32,\ C(*) = 0.$$

Paneme tähele, et Huffmani protseduur on rakendatav vaid lõpliku tähestiku korral, sest kui $|\mathcal{X}| = \infty$, pole võimalik leida väiseimaid tõenäosusi. Järgnevas tõestame, et lõplike $\mathcal{X}$ korral garanteerib Huffmani meetod optimaalse koodi. Eelkõige paneme tähele, et optimaalne kood leidub. Tõepoolest, kui $|\mathcal{X}| < \infty$, siis otsime minimaalse keskmise pikkusega koodi sisuliselt lõplikust koodide hulgast ning seetõttu optimaalne kood leidub (kuid pole üldiselt ühene).

2.4.2 Huffmani koodi optimaalsus

Olgu $\mathcal{X} = \{x_1, \ldots, x_m\}$. Üldisust kitsendamata eeldame, et

$$P(x_1) \geq P(x_2) \geq \cdots \geq P(x_m). \tag{2.8}$$

Teame, et leidub vähemalt üks optimaalne kood. Huffmani koodi optimaalsuse tõestus põhineb optimaalse koodi alljärgnevatel omadustel.
Esimene omadus väidab, et iga optimaalne kood seab väiksema tõenäosusega tähtedele vastavusse pikemad sõnad.

Väide 2.1 *Olgu C optimaalne. Siis $l(x_i) > l(x_j)$ vaid siis, kui $P(x_i) \leq P(x_j)$.*

Tõestus. Oletame vastuväiteliselt, et leiduvad x_i ja x_j nii, et $P(x_i) > P(x_j)$ ja $l(x_i) > l(x_j)$. Vahetades koodis C sõnad $C(x_i)$ ja $C(x_j)$ saame uue koodi C^*. Et aga

$$\begin{aligned} L(C) - L(C^*) &= P(x_i)l(x_i) + P(x_j)l(x_j) - (P(x_i)l(x_j) + P(x_j)l(x_i)) \\ &= (P(x_i) - P(x_j))(l(x_i) - l(x_j)) > 0, \end{aligned}$$

ei saa C olla optimaalne. ■

Vastavalt väitele 2.1 leidub optimaalne koodi nii, et

$$l(x_1) \leq l(x_2) \leq \cdots \leq l(x_m). \tag{2.9}$$

Def 2.8 *Koodisõnad $d', d'' \in \mathcal{D}^*$ on* **vennad** *(siblings), kui nad on ühepikkused ja erinevad üksteisest vaid viimase sümboli poolest.*

Huffmani kahendkoodi optimaalsus. Vaatleme olukorda $D = 2$, s.t. tõestame vaid Huffmani kahendkoodi optimaalsuse. Sellisel juhul on igal koodisõnal vaid üks vend. Järgnev omadus väidab, et leidub optimaalne kood nii, et kahe kõige väiksema tõenäosusega sõna koodid on vennad.

Väide 2.2 *Leidub optimaalne kood C nii, et $C(x_{m-1})$ ja $C(x_m)$ on vennad.*

Tõestus. Olgu C optimaalne kood. Järjestame tähed nii, et kehtivad võrratused (2.8) ja (2.9)). Seega $C(x_m)$ on pikim (võrratused (2.9)). Et $C(x_m)$ on pikim, ei saa koodisõna $C(x_m)$ vend olla ühegi teise koodisõna prefiks. Oletame, et $C(x_m)$ vend pole ühegi tähe kood. Sellisel juhul saaksime aga koodisõna $C(x_m)$ vähendada ühe võrra, mis on vastuolus koodi C optimaalsusega. Seega leidub x_j nii, et $C(x_m)$ ja $C(x_j)$ on vennad. Kui $j = m-1$, siis väide kehtib. Kui $j < m-1$, siis võrratustest (2.9) saame, et $l(x_j) = l(x_{m-1}) = l(x_m)$, mistõttu võime koodisõnad $C(x_j)$ ja $C(x_{m-1})$ ära vahetada. Et $l(x_j) = l(x_m)$, siis selline vahetamine ei muuda keskmist koodipikkust (optimaalsust), võrratusi (2.9) ega ka võrratusi (2.8). ■

Teoreem 2.9 *Huffmani kood on optimaalne kahendkood.*

Tõestus. Väitest 2.2 teame, et leidub optimaalne kahendkood C nii, et $C(x_{m-1})$ ja $C(x_m)$ on vennad. Huffmani koodil on sama omadus. Liigume nüüd mööda koodi C puud edasi, asendades $C(x_{m-1})$ ja $C(x_m)$ nende ühise tüvega. Nii saame uue koodi C', mis vastab redutseeritud (vähendatud) jaotusele, kus x_m ja x_{m-1} on kokku võetud üheks täheks y tõenäosusega $p_m + p_{m-1}$. Kood C' on keskmiselt lühem kui C, nende pikkuste vahe on

$$L(C) - L(C') = lp_m + lp_{m-1} - (p_m + p_{m-1})(l-1) = p_m + p_{m-1},$$

kus $l = l(x_m) = l(x_{m-1})$. Seega ei sõltu koodi pikkuste vahe nende struktuurist, mistõttu C on optimaalne parajasti siis, kui C' on optimaalne. Teisisõnu, iga vähendatud tähestikul antud optimaalsest koodist saame originaaltähestiku optimaalse koodi, lisades y koodile sümboli "0" (ja saades x_{m-1} koodi) ning sümboli "1" (ja saades x_m koodi). Seega oleme optimaalse koodi leidmise probleemi taandanud optimaalse koodi otsimise probleemile vähendatud tähestikul. Väitest 2.2 teame, et vähendatud jaotusel leidub optimaalne kood nii, et kahe väikseima tõenäosusega tähe koodid on vennad. Ühendame need tähed, just nagu Huffmani protseduuris, ning vähendame tähestikku veel ühe tähe võrra. Nüüd otsime optimaalset koodi uuel tähestikul jne. Lõpuks vähendame tähestikku kahe täheni ning sellisel juhul on optimaalne kood ilmne. Seega oleme tõestanud, et Huffmani protseduur annab meile optimaalse kahendkoodi. ■

Analoogiliselt saab tõestada, et Huffmani kood on optimaalne D-ndkood. Skitseerime tõestuse.
Üldisust kitsendamata eeldame, et tähestikus on $D + k(D-1)$ tähte. Kui see nii pole, lisame sobiva arvu pseudotähti. Pseudotähed ei suurenda keskmist koodipikkust, seega optimaalne kood laiendatud tähestikul on optimaalne ka esialgsel tähestikul.

Def 2.10 *Ütleme, et D-ndpuu on täielik, kui igal tema sõlmel on täpselt D alluvat.*

Täielik puu rahuldab Krafti võrratust võrdusena. Täielikul puul on $D + (m-1)(D-1)$ lehte, kus m on sõlmede arv.

Järgnevas paneme tähele, et iga optimaalne koodipuu on täielik, sest

- optimaalse puu igal sõlmel on D alluvat v.a. juhul, kui alampuu pikkus on 1;
- mittetäielikud alampuud saavad olla vaid viimasel tasemel;
- keskmist pikkust suurendamata võib viimasel tasemel olevaid mittetäielikke alampuid muuta nii, et neid jääb maksimaalselt üks;
- kui tähestikus on $J := D + k(D-1)$ tähte (puul on $D + k(D-1)$ lehte), ei saa optimaalsel puul olla vaid ühte mittetäielikku alampuud. Tõepoolest: oletame, et optimaalsel puul on sõlm, millel on vähem kui D järglast. Et puu on optimaalne, saab sellele sõlmele vastava alampuu pikkus olla vaid 1. Olgu selle sõlme järglaste arv a. Et puu on optimaalne, ei saa a olla 1, millest eelöeldu tõttu $2 \leq a \leq D-1$. Elimineerides ainsa mittetäieliku alampuu (ning vaadeldes sõlme uue lehena) saame täieliku puu, millel on $J - a + 1 = D + k(D-1) - a + 1$ lehte. Saadud uus puu on täielik, mistõttu teme lehtede arv peab olema $D + m(D-1)$. See pole aga antud a korral võimalik.

Nüüd on Huffmani D-nd koodi optimaalsuse tõestus analoogiline Huffmani kahendkoodi optimaalsuse tõestusega. Väide 2.1 ja võrratused (2.9) kehtivad suvalise D korral. Arvestades, et leidub alati täielik optimaalne koodipuu, on kerge nähe, et kehtib väite 2.2 analoog: leidub optimaalne D-ndkood C nii, et $C(x_{m-D+1}), C(x_{m-D+2}), \ldots, C(x_m)$ on vennad. Tõepoolest, väikseima tõenäosusega leht peab olema pikima koodisõnaga; et puu on täielik, peavad koodi kuuluma ka kõik tema vennad, optimaalsuse tõttu peavad vendadele vastavad lehed olema võimalikult väikese tõenäosusega.

Teoreemi 2.9 üldistus D-ndkoodidele on nüüd ilmne (veendu!).

Märkused:

- Mitte kõik optimaalsed koodid pole Huffmani koodid, s.t. leidub optimaalseid koode, milliseid pole võimalik konstrueerida Huffmani meetodil.
  Olgu näiteks $\mathcal{X} = \{a, b, c, d, e, f\}$, kõik tähed olgu võrdse tõenäosusega. Vaatleme koode C_1 ja C_2, mis on antud tabelitena

täht \ kood	C_1	C_2
a	11	111
b	101	110
c	100	101
d	011	100
e	010	01
f	00	00

Kood C_2 on Huffmani kood, kui kood C_1 mitte (ülesanne 5), mõlemad on optimaalsed.

- Optimaalse koodi keskmine pikkus ei pruugi alati olla $H_D(P)$. Tõepoolest, eelmises näites on optimaalse (Huffmani) koodi keskmine pikkus $\frac{8}{3}$, mis on rangelt suurem entroopiast $\log 6$. Teame, et Huffmani koodi keskmine pikkus L rahuldab alati võrratusi

$$H_D(P) \leq L < H_D(P) + 1.$$

On kerge veenduda, et antud tõkkeid ei saa parandada. Et alumine tõke võib olla täpne, seda me juba teame. Veendume nüüd, et L võib olla kuitahes lähedal arvule $H_D(P) + 1$. Selleks vaatleme jaotust (k on piisavalt suur)

a	b	c	d
$\frac{1}{k}$	$\frac{1}{k}$	$\frac{1}{k}$	$1 - \frac{3}{k}$

Huffmani kahendkoodi pikkused on $l(a) = l(b) = 3$ $l(c) = 2$ $l(d) = 1$ (kui k on piisavalt suur), millest $L = \frac{8}{k} + 1 - \frac{3}{k} \to 1$, kui $k \to \infty$. Samas entroopia

$$H(P) = \frac{3}{k} \log k - (1 - \frac{3}{k}) \log(1 - \frac{3}{k}) \to 0, \text{ kui } k \to \infty.$$

Seega $H(P) + 1 - L \to 0$, kui $k \to \infty$.
Milline on ülaltoodud jaotuse Shannon-Fano kood?

- Ülaltoodud näidetest võib jääda mulje, otsekui oleks Shannon-Fano koodi sõnapikkused alati pikemad Huffmani (või mõne teise optimaalse koodi sõnapikkustest). Kontranäitena vaatleme jaotust

a	b	c	d
$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{12}$

Huffmani koodisõnade pikkused on vastavalt $(2, 2, 2, 2)$ või $(1, 2, 3, 3)$. Seega leidub Huffmani kood nii, et $l(c) = 3$. Shannon-Fano koodi korral on aga $l(c) = 2$.

- Lõpmatu tähetiku korral Huffmani koodi üldiselt defineerida ei saa, sest selle konstrueerimine alagab altpoolt (kõige väiksema tõenäosusega tähtedest). Teatud tingimustel on Huffmani kahendkoodi võimalik defineerida ka "tükikaupa", st ülalt alla. Kirjeldame üht sellist olukorda. Olgu tõenäosused järjestatud

$$p_1 \geq p_2 \geq \cdots .$$

Oletame, et leidub lõpmata palju aatomeid p_m, mis rahuldavad tingimust

$$p_m \geq \sum_{i>m} p_i =: p_m^*. \tag{2.10}$$

Kujutagem korraks ette, et tähestikus on lõplik arv (kuid väga palju) tähti. Olgu $p_{m_1}, p_{m_2}, \ldots$ tingimust (2.10) rahuldavad aatomid. Et p_{m_1} rahuldab tingimust (2.10),

on selge, et Huffmani protseduuri järgides (lõpliku hulga tähtede korral on see võimalik) ühendatakse kõik aatomid p_j, kus $j > m_1$ enne p_{m_1} (tuletame meelde, et me vaatleme olukorda $D = 2$). Seega, mingil hetkel on protseduur jõudnud jaotuseni

$$p_1, p_2, \ldots, p_{m_1}, p^*_{m_1}. \tag{2.11}$$

Et jaotuseni (2.11) jõutakse suvaliste aatomite p_j, $j > m_1$ korral (kui vaid nende summa on $p^*_{m_1}$), siis võib lõpmatu koodi konstrueerimist alustada jaotusele (2.11) vastava kahendpuu konstrueerimisest. Edasi asume konstrueerime alampuud, mis väljub sõlmest $p^*_{m_1}$. Selleks vaatleme jaotust, mis on proportsionaalne vektoriga

$$p_{m_1+1}, p_{m_1+2}, \ldots, p_{m_2}, p^*_{m_2}. \tag{2.12}$$

Arvud (2.12) ei moodusta tõenäosusjaotust, sest nende summa on $p^*_{m_1}$. Huffmani protseduuri seisukohalt pole kogusumma oluline. Argumenteerides nagu ülalpool, näeme, et sõlmest $p^*_{m_1}$ väljuva alampuu konstrueerimist võime alustada aatomitele (2.12) vastava alampuu konstrueerimisest. Edasi alustame sõlmele $p^*_{m_2}$ vastava alampuu konstrueerimist. Selleks vaatleme aatomeid

$$p_{m_2+1}, p_{m_2+2}, \ldots, p_{m_3}, p^*_{m_3}, \tag{2.13}$$

konstrueerime neile vastava puu jne. On selge, et kirjeldatud protseduur ei sõltu tähtede hulgast ning üldistub seega lõpmatule tähestikule.

Näide: Kui jaotus on geomeetriline parameetriga p, kus $p \geq 0.5$, siis (2.10) kehtib iga m korral (veendu selles!).

2.5 Üheselt dekodeeritavad koodid

Iga prefikskood on üheselt dekodeeritav, vastupidine ei kehti. Et üheselt dekodeeritavate koodide klass on laiem prefikskoodide klassist, on loomulik oletada, et üheselt dekodeeritava koodi sõnapikkused võivad olla "lühemad" kui prefikskoodi sõnadpikkused. Prefikskoodi sõnapikkuste alumise tõkke andis (teatavas mõttes) Krafti võrratus. Järgnev teoreem väidab, et Krafti võrratus kehtib ka üheselt dekodeeritavate koodide korral ehk üheselt dekodeeritavavate koodidide sõnapikkused ei saa tegelikult olla oluliselt "lühemad" prefikskoodide sõnapikkustest. Teisisõnu: üheselt dekodeeritavate koodide klass pole sisuliselt laiem prefikskoodide klassist.

Teoreem 2.11 *Olgu C tähestikul $\mathcal{X}$ antud üheselt dekodeeritav kood, koodipikkustega $\{l(x)\}$. Siis kehtib Krafti võrratus*

$$\sum_x D^{-l(x)} \leq 1. \tag{2.14}$$

Tõestus. Vaatleme erijuhtu, mil $\mathcal{X}$ on lõplik.
Olgu C^k koodi C k-laiend, s.t.

$$C^k : \mathcal{X}^k \to \mathcal{D}^*, \quad C^k(x_1 \cdots x_k) = C(x_1) \cdots C(x_k).$$

$$\begin{aligned}\Big(\sum_x D^{-l(x)}\Big)^k &= \sum_{x_1 \in \mathcal{X}} \sum_{x_2 \in \mathcal{X}} \cdots \sum_{x_k \in \mathcal{X}} D^{-l(x_1)} D^{-l(x_2)} \cdots D^{-l(x_k)} \\ &= \sum_{x_1 x_2 \cdots x_k \in \mathcal{X}^k} D^{-l(x_1)} D^{-l(x_2)} \cdots D^{-l(x_k)} \\ &= \sum_{x^k \in \mathcal{X}^k} D^{-l(x^k)},\end{aligned}$$

kus $x^k := x_1 \cdots x_k$ ja

$$l(x^k) := l(x_1) + \cdots + l(x_k) = |C^k(x^k)|.$$

Olgu $a(m)$ selliste k-sõnade arv, milliseid C^k kodeerib m-sõnaliste koodisõnadega. Formaalselt

$$a(m) = |\{x^k \in \mathcal{X}^k : l(x^k) = m\}|.$$

Kasutame nüüd asjaolu, et $\mathcal{X}$ on lõplik. Olgu

$$l_{max} := \max_{x \in \mathcal{X}} l(x).$$

On selge, et

$$\max_{x^k \in \mathcal{X}^k} l(x^k) = k l_{max}.$$

Seega

$$\Big(\sum_x D^{-l(x)}\Big)^k = \sum_{x^k \in \mathcal{X}^k} D^{-l(x^k)} = \sum_{m=k}^{k l_{max}} a(m) D^{-m}.$$

Nüüd kasutame asjaolu, et C on üheselt dekodeeritav, millest johtuvalt C^k on ühene. Fikseerime m ja vaatleme sõnu hulgast $\{x^k \in \mathcal{X}^k : l(x^k) = m\}$. Pikkusega m koodisõnu on ülimalt D^m. Et C^k on ühene, vastab erinevale koodisõnale erinev x^k, mistõttu $a(m) \le D^m$. Seega

$$\Big(\sum_x D^{-l(x)}\Big)^k = \sum_{m=k}^{k l_{max}} a(m) D^{-m} \le \sum_{m=1}^{k l_{max}} D^m D^{-m} = k l_{max}$$

ehk

$$\sum_x D^{-l(x)} \le \big(k l_{max}\big)^{\frac{1}{k}}.$$

Võrratuse vasak pool ei sõltu k-st. Järelikult

$$\sum_x D^{-l(x)} \le \lim_{k \to \infty} \big(k l_{max}\big)^{\frac{1}{k}} = 1.$$

Lõpmatu $\mathcal{X}$ korral ei lähe ülaltoodud tõestus läbi, sest $l_{max} = \infty$. Vaatleme lõplikku alamtähestikku $\mathcal{X}_m = \{x_1, \ldots, x_m\} \subset \mathcal{X}$. Üheselt dekodeeritava koodi C ahend alamtähestikule $\mathcal{X}_m$ on ikka üheselt dekodeeritav. Alamtähestik on lõplik, seega

$$\sum_{x \in \mathcal{X}_m} D^{-l(x)} \leq 1.$$

Kehtib iga m korral, millest

$$\sum_{x \in \mathcal{X}} D^{-l(x)} = \lim_{m \to \infty} \sum_{x \in \mathcal{X}_m} D^{-l(x)} \leq 1.$$

■

Paneme tähele, et triviaalselt kehtib ka vastupidine väide: kui etteantud koodipikkused rahuldavad Krafti võrratust, siis leidub nende koodipikkustega üheselt dekodeeritav kood. Teame ju, et Krafti võrratuse kehtivuse korral leidub vastavate koodipikkustega prefikskood. Iga prefikskood on aga üheselt dekodeeritav.

Ülaltoodud teoreemist järeldub, et üheselt dekodeeritavad koodide ja prefikskoodide koodipikkuste hulgad langevad kokku. Teisisõnu, igale üheselt dekodeeritavale koodile vastab vähemalt üks samade koodipikkustega prefikskood. See aga tähendab, et igale üheselt dekodeeritavale koodile vastab sama keskmise pikkusega prefikskood ning optimaalne prefikskood on ka optimaalne üheselt dekodeeritav kood. Seega prefikskoodide hulga laiendamine üheselt dekodeeritavate koodideni ei anna keskmise koodipikkuse mõttes mingit efekti. Seetõttu tegeletaksegi informatsiooniteoorias valdavalt prefikskoodidega, sest viimased esituvad puuna.

2.6 Sõnade kodeerimine

Olgu $X_1, \ldots, X_k$ juhuslik vektor tähestikul $\mathcal{X}^k$ (juhuslik sõna). Olgu C tähestiku $\mathcal{X}$ mingi kood. Selle koodi k-laiend C^k kodeerib sõnu $\mathcal{X}^k$. Samas võib hulka $\mathcal{X}^k$ vaadelda omaette tähestikuna ning püüda seda omaette (võimalikult optimaalselt) kodeerida. Kumb on efektiivsem – kas kodeerida optimaalselt tähestik ja laiendada siis seda sõnadele või kodeerida optimaalselt sõnu?
Olgu $C_k : \mathcal{X}^k \to \mathcal{D}^*$ sõnade kood, koodipikkustega $l(x^k)$. Et selle koodi keskmine pikkus kasvab koos k-ga, huvitume koodipikkusest tähe kohta:

$$L_k := \frac{1}{k} L(C_k) = \frac{1}{k} \sum_{x^k \in \mathcal{X}^k} P(x^k) l(x^k) = \frac{1}{k} El(X_1, \ldots, X_k).$$

Sama jaotusega tähed. Uurime kõigepealt tähtede koodi C laiendit C^k. On lihtne veenduda, et kui $X_1, \ldots, X_k$ on sama jaotusega P (kuid mitte ilmtingimata sõltumatud) juhuslikud suurused, siis (ülesanne 14) $L(C^k) = kL(C)$, millest

$$L_k(C^k) = L(C). \tag{2.15}$$

Seega keskmiselt kulub ühe tähe kodeerimiseks ikka $L(C)$ ühikut. Kui C on optimaalne, siis

$$H_D(P) \leq L_k < H_D(P) + 1,$$

kusjuures parempoolne võrratus võib olla kuitahes täpne.

Vaatleme nüüd parimat sõnade koodi. Järeldusest 2.1 saame, et leidub selline kood C_k, et

$$H_D(X_1, \ldots, X_k) \leq L(C_k) < H_D(X_1, \ldots, X_k) + 1,$$

millest

$$\frac{H_D(X_1, \ldots, X_k)}{k} \leq L_k \leq \frac{H_D(X_1, \ldots, X_k)}{k} + \frac{1}{k}. \tag{2.16}$$

Sõltumatud ja sama jaotusega (i.i.d) tähed. Oletame nüüd, et tähed $X_1, \ldots, X_k$ on sõltumatud ja sama jaotusega, $X_i \sim P$. Siis $H_D(X_1, \ldots, X_k) = \sum_{i=1}^{k} H_D(X_i) = kH_D(P)$ ning seosest (2.16) saame

$$H_D(P) \leq L_k < H_D(P) + \frac{1}{k}. \tag{2.17}$$

Seega alati leidub kood, mille korral L_k erineb $H_D(P)$-st ülimalt $\frac{1}{k}$ võrra. Suurendades k-d kui vaja, saame entroopiale $H_D(P)$ kuitahes lähedale. Võrratust (2.17) kutsutakse ka *Shannoni esimeseks teoreemiks (noiseless coding theorem)*. Pane tähele, et selline kood pole üldiselt saadav optimaalse tähtede koodi laiendina.

Statsionaarne protsess. Olgu $X = X_1, X_2, \ldots$ statsionaarne protsess, $X_i \sim P$. Olgu $C_k : \mathcal{X}^k \to \mathcal{D}^*$ optimaalne kood. Tuletame meelde, et statsionaarsel protsessil on alati entroopiamäär

$$H_X = \lim_k \frac{H_D(X_1, \ldots, X_k)}{k} = \lim_k H_D(X_k | X_1, \ldots, X_{k-1}) \leq H(P).$$

(Kui $D > 2$, defineerime entroopiamäära analoogiliselt. Meil on D fikseeritud, mistõttu jätame ta tähistusest välja.) Seosest (2.16) saame, et

$$L^* := \lim_k L_k = \lim_k \frac{H_D(X_1, \ldots, X_k)}{k} = H_X.$$

Seos (2.6) annab entroopiamäärale sisu: H_X on protsessi kodeerimise keskmine pikkus tähe kohta.

Kokkuvõtteks: Kui $X = X_1, X_2, \ldots$ on i.i.d. (väga spetsiifiline statsionaarne protsess), siis parima sõnade koodi ja pikkade sõnade korral keskmiselt kulub ühe tähe kohta $L(P)$ kooditähte. Sellisel juhul võidame sõnakaupa kodeerides vaid seda, et (piisavalt suure k korral) on $H_D(P)$ kuitahes täpselt saavutatav.

Kui $H_X < H_D(P)$, siis keskmine koodipikkus ühe tähe kohta võib olla oluliselt väiksem kui iga tähte eraldi kodeerides.

Näide: Olgu X statsionaarne MA üleminekumaatriksiga I_k (k seisundit). Sellisel juhul $H(P) = \log k$, kuid $L_k = H_X = 0$.

2.6.1 Üheselt dekodeeritava koodi muutmine prefikskoodiks

Igale üheselt dekodeeritavale koodile saab vastavusse seada samade koodipikkustega prefikskoodi. Kui kodeeritavaid tähti (neid on $|\mathcal{X}|$) pole palju, võib ettentud koodipikkustega koodipuu konstrueerimne olla suhteliselt lihtne. Üldiselt võib selleks kasutada Kraffti võrratuse tõestuses kasutatud võtet. Praktikas võib see olla suhteliselt keerukas, iseäranis pikkade sõnade $\mathcal{X}^k$ kodeerimisel. Järgnevas vaatleme, kuidas suvalise üheselt dekodeeritava koodi saab muuta prefikskoodiks sobiva prefiksi lisamisel. Prefiksi lisamine teeb küll koodi pikemaks, kuid seda saab teha nii, et L^* ei muutu, s.t. pikkade sõnade kodeerimisel on vahe tühine.

Eliase delta kood. Alustame lemmast.

Lemma 2.1 (Eliase lemma) *Leidub prefikskood $E : \{1, 2, \ldots\} \to \mathcal{D}^*$ nii, et*

$$|E(n)| = \log_D n + o(\log_D n) \tag{2.18}$$

Tõestus. Iga naturaalarvu kodeerime kolmes osas

$$E(n) = u(n)v(n)w(n),$$

kus $w(n)$ on arvu n D-ndesitus. Seega

$$w(n) = \ulcorner \log_D(n+1) \urcorner.$$

Teine osa $v(n)$ on pikkuse $w(n)$ D-ndesitus ja esimene osa $u(n)$ koosneb nullidest, kusjuures neid nulle on niipalju kui on $v(n)$ pikkus. Seega

$$|u(n)| = |v(n)| = \ulcorner \log_D(1 + \ulcorner \log_D(n+1) \urcorner) \urcorner.$$

Seega

$$|E(n)| = \ulcorner \log_D(n+1) \urcorner + 2\ulcorner \log_D(1 + \ulcorner \log_D(n+1) \urcorner) \urcorner = log_D n + o(\log_D n).$$

Veendume, et $E(n)$ on prefikskood. Oletame et leiduvad n ja m nii, et $E(m)$ on $E(n)$ prefiks, s.t.

$$u(n)v(n)w(n) = u(m)v(m)w(m)w'.$$

Sellisel juhul $u(n) = u(m)$, sest mõlemad koosnevad nullidest ning $v(n)$ ja $v(m)$ esimene sümbol pole 0. Sellise juhul aga $v(n) = v(m)$, sest nende pikkused peavad olema võrdsed. See aga tähendab, et $w(m) = w(n)$ ehk w' on tühi ja $n = m$.

■

Saadud koodi nimetatakse **Eliase (delta) koodiks**.

Näide: Leiame $E(12)$. Numbri 12 kahendkuju on 1100. Seega $w(12) = 1100$. Et $w(12)$ koosneb 4 bitist, saame $v(12) = 100$. Lõpuks $u(12) = 000$. Seega

$$E(12) = u(12)v(12)w(12) = 0001001100.$$

Märkus: Kui $D = 2$, siis Eliase delta koodi saab vähendada kahe biti võrra. Tõepoolest, et iga n korral $|v(n)| \geq 1$, siis ühe nullidest võib "meelde jätta" ja koodi esimene osa on siis $u(n) - 1$. Teiseks, et iga kahendnumber algab ühega, võib ka selle arvu "meelde jätta" ning seega kirjutamata jätta. Seega $w(n)$ on siis arvu n kahendkuju, millest esimene üks on kustutatud. Samas $v(n)$ on ikka terve kahendkuju pikkuse kahendesitus. Seega saadud kood, olgu see E^* on täpselt kahe biti võrra lühem kui $E(n)$. Seega $E^*(12) = 00100100$.

Üheselt dekoteeritavate koodide muutmine prefikskoodideks. Olgu $C_k : \mathcal{X}^k \to \mathcal{D}^*$ sõnade kood, koodipikkustega $\{l(x^k)\}$. Olgu C_k üheselt dekodeeritav. Defineerime koodi C_k *Eliase laiendi*

$$C_k^*(x^k) = E(l(x^k))C^k(x^k).$$

Saadud kood on prefikskood, sest prefiks $E(l(x^k))$ määrab järgneva koodisõna pikkuse. Dekodeerija loeb läbi laiendi $E(l(x^k))$, saab üheselt aru, millal see lõpeb ning kui pikk on järgnev koodisõna. Viimane saab dekodeeritud just siis, kui ta lugemine lõpeb.

Näide: Olgu $D = 2$ ja $C^k(x^k) = 001001100111$. Selle sõna pikkus on 12. Teame, et $E(12) = 0001001100$. Seega

$$C_k^*(x^k) = 000100110000101100111.$$

Kuigi antud näite korral on Eliase laiend peaaegu sama pikk kui koodisõna ise, garanteerib Eliase lemma, et koodisõnade pikkuste kasvamisel (näiteks k kasvamisel) muutub laiendi osa tühiseks.

Koodide kombineerimine. Teine rakendus Eliase laiendile on loenduva hulga koodide kombineerimine üheks koodiks. Oletame, et meil on iga $k \geq 1$ korral defineeritud prefikskood

$$C^k : \mathcal{X}^k \to \mathcal{D}^*.$$

Kasutades Eliase laiendit saame defineerida prefikskoodi

$$C : \mathcal{X}^* \to \mathcal{D}^*, \quad C(x^k) = E(k)C_k(x^k).$$

Seega Eliase laiend määrab ära koodi indeksi, seejärel dekodeeeritakse sõna.

2.7 Optimaalse koodi tõenäosuslik käitumine*

Optimaalne kood on lühima keskmise pikkusega. Olgu C optimaalne kood ja C' mingi teine kood; nende koodisõnade pikkused olgu vastavalt $\{l(x)\}$ ja $\{l'(x)\}$. Nagu üleelmises osas toodud näidetest nägime, ei pruugi optimaalse koodi kõikide sõnade pikkused olla lühemad teiste sõnade pikkustest: võib leiduda $x \in \mathcal{X}$ nii, et $l'(x) < l(x)$. Kui tihti seda aga juhtub ehk kui suur on selliste tähtede tõenäosus? Kui X on juhuslik täht, siis viimane tõenäosus avaldub $\mathbf{P}(l'(X) < l(X))$.
Optimaalsed koodid on Huffmani koodid, nende pikkustega manipuleerimine pole lihtne. seetõttu uurime tõenäosust $\mathbf{P}(l'(X) < l(X))$ juhul, kui l on Shannon-Fano kood.
Esimene teoreem annab ülemise tõkke tõenäosusele, et $l'(X) \leq l(X) - c$.

Teoreem 2.12 *Olgu $\{l(x)\}$ Shannon-Fano koodipikkused, $\{l'(x)\}$ olgu üheselt dekodeeritava koodi kodipikkused. Siis*

$$\mathbf{P}(l'(X) \leq l(X) - c) \leq D^{1-c}.$$

Tõestus.

$$\begin{aligned}
\mathbf{P}\Big(l'(X) \leq l(X) - c\Big) &= \mathbf{P}\Big(l'(X) \leq \lceil \log_D \frac{1}{P(X)} \rceil - c\Big) \\
&\leq \mathbf{P}\Big(l'(X) \leq \log_D \frac{1}{P(X)} - c + 1\Big) \\
&= \mathbf{P}\Big(l'(X) + c - 1 \leq -\log_D P(X)\Big) \\
&= \mathbf{P}\Big(P(X) \leq D^{-l'(X)-c+1}\Big) \\
&= \sum_{x:P(x)\leq D^{-l'(x)-c+1}} P(x) \\
&\leq \sum_{x:P(x)\leq D^{-l'(x)-c+1}} D^{-l'(x)-c+1} \\
&\leq \sum_{x} D^{-l'(x)-c+1} \\
&\leq D^{-c+1} \sum_{x} D^{-l'(x)} \\
&\leq D^{1-c}.
\end{aligned}$$

■

Optimaalne Shannon-Fano kood. Ülaltoodud teoreem ei anna mingit tõket tõenäosusele $\mathbf{P}(l'(X) < l(X))$, sest teoreemist järeldub vaid triviaalne tõke:

$$\mathbf{P}(l'(X) < l(X)) = \mathbf{P}(l'(X) \leq l(X) - 1) \leq D^{1-1} = 1.$$

Järgnev teoreem aga väidab, et optimaalse Shannon-Fano koodi korral (tuletame meelde, et see saab olla vaid siis, kui P rahuldab seost (2.4)) kehtib võrratus $\mathbf{P}(l'(X) < l(X)) \leq \mathbf{P}(l(X) < l'(X))$. Seega juhuslikult valitud tähe korral on suurima tõenäosusega optimaalse Shannon-Fano kahendkoodi koodisõna pikkus lühem kui teise üheselt dekodeeritava koodi koodisõna pikkus.

Teoreem 2.13 *Rahuldagu P seost (2.4). Olgu $l(x) = \log_D \frac{1}{P(x)}$ ning olgu $\{l'(x)\}$ mingi üheselt dekodeeritava koodi kodipikkused. Siis*

$$\mathbf{P}(l'(X) < l(X)) \leq \mathbf{P}(l(X) < l'(X)),$$

kusjuures võrdus kehtib vaid siis, kui $l'(x) = l(x)$ iga x korral.

Tõestus. Olgu

$$\operatorname{sign}(a) = \begin{cases} 1 & \text{kui } a > 0, \\ 0 & \text{kui } a = 0, \\ -1 & \text{kui } a < 0. \end{cases}$$

Kui $a \in \mathbb{Z}$, siis

$$\operatorname{sign}(a) \leq D^a - 1.$$

$$\begin{aligned} \mathbf{P}(l'(X) < l(X)) - \mathbf{P}(l(X) < l'(X)) &= E\operatorname{sign}(l(X) - l'(X)) \\ &\leq E\big(D^{l(X)-l'(X)} - 1\big) \\ &= \sum_x P(x)(D^{l(x)-l'(x)} - 1) \\ &= \sum_x D^{-l(x)}(D^{l(x)-l'(x)} - 1) \\ &= \sum_x (D^{-l'(x)} - D^{-l(x)}) \\ &= \sum_x D^{-l'(x)} - 1 \\ &\leq 1 - 1. \end{aligned}$$

Võrratus on võrdus, kui iga x korral kehtib

$$\operatorname{sign}(l(x) - l'(x)) = D^{l(x)-l'(x)} - 1.$$

See aga saab olla vaid siis, kui iga x korral $l(x) = l'(x)$. ■

2.8 Diskreetse juhusliku suuruse genereerimine*

Olgu P lõplikul tähestikul $\mathcal{X}$ antud diskreetne jaotus. Seame endale eesmärgiks sellise jaotusega juhusliku suuruse genereerimise mündivistega. Teisisõnu, olgu $Z_1, Z_2, \ldots$ sõltumatud Bernoulli 1/2-jaotusega juhuslikud suurused. Olgu A algoritm, mis juhuslike suuruste $Z_1, Z_2, \ldots, Z_T$ abil tekitab jaotusega P juhusliku suuruse, s.t. $A(Z_1, \ldots, Z_T) \sim P$.

Siin T on juhuslik suurus, mille võimalikud väärtused on mittenegatiivsed täisarvud, kusjuures see, kas $T = n$ või mitte, sõltub juhuslikest suurusetst $Z_1, \ldots, Z_n$ (T on peatumishetk).

Näide: Olgu P järgmine

a	b	c
$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$

Algoritm A võiks olla järgmine

$$A(Z_1, \ldots, Z_T) = \begin{cases} a & \text{kui } Z_1 = 0, \\ b & \text{kui } Z_1 = 1, Z_2 = 1, \\ c & \text{kui } Z_1 = 1, Z_2 = 0. \end{cases}$$

Seega

$$T = \begin{cases} 1 & \text{kui } Z_1 = 0, \\ 2 & \text{mujal.} \end{cases}$$

Muidugi on näites toodud jaotust võimalik tekitada mitmeti. Meid huvitab keskmiselt lühim algoritm, s.t. algoritm, mis kasutab keskmiselt kõige vähem mündiviskeid. Teisisõnu, otsime algoritmi, mille korral algoritmi *keskmine pikkus* ET oleks minimaalne. Ülaltoodud näite korral on $ET = 1.5 = H(P)$.

Paneme tähele, et iga algoritmi võib esitada täieliku kahendpuuna. Puu lehtedel on tähestiku $\mathcal{X}$ tähed, kusjuures erinevatel lehtedel võib olla sama täht. Selliselt konstrueeritud puul võib olla lõpmatu arv lehti. Kui leht on k-ndal tasemel, siis selle lehe tõenäosus on 2^{-k}. Algoritmi keskmine pikkus on selle puu keskmine pikkus.

Olgu A ülalkirjeldatud puu (algoritm). Vaatleme kõiki puu lehti (sõltumata nendel olevast tähest), olgu nende hulk $\mathcal{Y}$. Igal lehel on tõenäosus 2^{-k}, kus k on selle lehe sügavus. Nii saame jaotuse Q. Selle jaotuse entroopia on puu keskmine pikkus ET, sest

$$ET = \sum_{y \in \mathcal{Y}} k(y) 2^{-k(y)} = \sum_{y} -\log 2^{-k(y)} 2^{-k(y)} = H(Q).$$

Nüüd on lihtne tõestada seos algoritmi keskmise pikkuse ja juhusliku suuruse X entroopia vahel.

Teoreem 2.14 *Ükski jaotust P genereeriva algoritmi keskmine pikkus pole suurem kui $H(P)$, s.t.*

$$ET \geq H(P).$$

Tõestus. Olgu A algoritm, mis genereerib X. Olgu Q algoritmile A vastava puu lehtede jaotus, $Y \sim Q$. Teame, et $H(Y) = E(T)$. Et aga algoritm on esitatav puuna, kehtib $X = f(Y)$. Seega $ET = H(Y) \geq H(X) = H(P)$. ■

Ülaltoodud teoreem pole eriti üllatav: et $H(P)$ on jaotuseses sisaldav informatsioon, on üsna loomulik, et $H(P)$ seab alumise piiri selle jaotuse tekitamiseks vajaminevate mündivisete arvule.

Ülaltoodust on ka selge, et seost (2.4) rahuldava P korral leidub algoritm, mille keskmine pikkus on $H(P)$. Tõepoolest, olgu C jaotusele P vastav Shannon-Fano kood. Sellele koodile vastav puu on täielik ning kui seda puud kasutada juhusliku suuruse genereerimiseks, saame, et tähe x tõenäosus on $2^{-k(x)}$, kus $k(x)$ on tähe x sügavus. Et $k(x) = l(x) = \log \frac{1}{P(x)}$, saame, et $2^{-k(x)} = P(x)$. Seega võib seda puud seda võib kasutada X genereerimiseks. Algoritmi keskmine pikkus koodi keskmine pikkus, mis võrdub entroopiaga $H(P)$.

Seega on seost (2.4) rahuldava jaotuse P optimaalne genereerimine sisuliselt ekvivalentne optimaalse kahendkoodi leidmisega. Kas selline ilus seos kodeerimise ja genereerimise vahel kehtib ka juhul, kui P ei rahulda seost (2.4)? Teisisõnu, kas ka üldisel juhul on jaotust P tekitav optimaalne algoritm sisuliselt sama, mis jaotust P kodeeriv minimaalne algoritm. On lihtne veenduda, et üldiselt pole nii, sest iga optimaalne koodipuu (nt. Huffmani puu) tekitab (kui seda kasutada juhuslikkuse genereerimisel) vaid seost (2.4) rahuldava jaotuse. Tekitamaks suvalist jaotust, toimime järgmiselt: aatomi $P(x)$ tekitamiseks leiame suurima arvu 2^{-k_1} nii, et $2^{-k_1} \leq P(x)$ ning seame ühele sügavusel k_1 olevatest lehtedest vastavusse x. Seejärel leiame suurima 2^{-k_2} nii, et $2^{-k_2} \leq P(x) - 2^{-k_1}$ ning seame ülele sügavusel k_2 olevatest lehtedest x jne. Sisuliselt leiame aatomi $P(x)$ kahendesituse:

$$P(x) = \sum_{i \geq 1} 2^{-k_i(x)}.$$

Nüüd konstrueerime kahendpuu, kus sügavusel $k_i(x)$ olevale lehele same vastavusse tähe x. Et $\sum_x P(x) = 1$, siis on sellise puu konstrueerimine alati võimalik ning see on täispuu.

Näited:

- Olgu

a	b	c	d
$\frac{9}{16}$	$\frac{5}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
0.1001_2	0.0101_2	0.0001_2	0.0001_2

  Vastav puu (algoritm) on järgmine.

- Olgu

a	b
$\frac{2}{3}$	$\frac{1}{3}$
$0.1010101\cdots_2$	$0.0101010\cdots_2$

  vastav puu (algoritm) on järgmine.

Saab näidata, et selline algoritm on minimaalse keskmise pikkusega, kusjuures

$$H(X) \leq ET < H(X) + 2.$$

2.9 Ülesanded

1. Tõesta alumine tõke teoreemis 2.7

2. Olgu P

a	b	c	d	e	f	g	h
0.25	0.05	0.1	0.13	0.2	0.12	0.08	0.07

   Konstrueerida optimaalne kahend- ja kolmendkood, leida nende keskmine pikkus.

3. Olgu koodipikkused $1, 1, 2, 2, 3, 3, 3$.

   a) Kas leidub selliste koodipikkustega kahendkood? Kui vastus on jaatav, siis konstrueerida vastavate koodipikkustega kahendkood. Kas leidub jaotus P, mille jaoks konstrueeritud kood on optimaalne?

   b) Kas leidub selliste koodipikkustega kolmendkood? Kui vastus on jaatav, siis konstrueerida vastavate koodipikkustega kolmendkood. Kas leidub jaotus P, mille jaoks konstrueeritud kood on optimaalne?

   c) Kas leidub selliste koodipikkustega neljandkood? Kui vastus on jaatav, siis konstrueerida vastavate koodipikkustega kood. Kas leidub jaotus P, mille jaoks konstrueeritud kood on optimaalne?

4. Kas C saab olla Huffmani kood, kui tema sõnad on

   - $\{0, 10, 11\}$
   - $\{00, 01, 10, 110\}$
   - $\{10, 01, 00, \}$?

5. Olgu$\mathcal{X} = \{a, b, c, d, e, f\}$, kõik tähed olgu võrdse tõenäosusega. Vaatleme koode C_1 ja C_2, mis on antud tabelitena

täht \ kood	C_1	C_2
a	11	111
b	101	110
c	100	101
d	011	100
e	010	01
f	00	00

   Veendu, et C_2 on Huffmani kood, kuid kood C_1 mitte, mõlemad on optimaalsed.

6. Kood on sufikskood, kui ükski koodisõna pole mingi teise koodisõna sufiks. Kas sufikskood on üheselt dekodeeritav?

7. Olgu
$$l_1 \leq l_2 \leq \cdots \leq l_m$$
täisarvud. Iga $1 \leq k \leq m$ korral valitakse binaarne koodisõna pikkusega l_k kõikide pikkusega l_k võimalike koodisõnade seast ühtlase jaotusega. Nii saadakse juhuslik kood C. Olgu $\mathcal{P}$ prefikskoodide hulk. Tõestada, et
$$\mathbf{P}(C \in \mathcal{P}) = \prod_{k=1}^{m} \Big(1 - \sum_{j=1}^{k-1} 2^{-l_j}\Big)^{+}.$$
Tõestada, et $\mathbf{P}(C \in \mathcal{P}) > 0$ parajasti siis, kui $l_1 \leq l_2 \leq \cdots \leq l_m$ rahuldavad Krafti võrratust.

8. Olgu $L_D(p_1, \ldots, p_m)$ jaotusele $(p_1, \ldots, p_m)$ vastava optimaalse D-koodi keskmine pikkus. Veendu, et kuigi optimaalne kood pole tõenäosuste $(p_1, \ldots, p_m)$ pidev funktsioon, on seda $L_D(p_1, \ldots, p_m)$.

9. Näita, et kui $L_D(p_1, \ldots, p_m) = H_D(p_1, \ldots, p_m)$, siis $m = D + k(D-1)$, kus k on mittenegatiivne täisarv.

10. Olgu $q < \frac{2}{3}$. Olgu $p \in [0, 1]$ selline, et
$$L_2(1-q, \frac{q}{2}, \frac{q}{2},) = H_2(1-p, \frac{p}{2}, \frac{p}{2}).$$
Leida seos p ja q vahel.

11. **a)** Leida $L_2(0.5, 0.25, 0.1, 0.05, 0.05, 0.05)$, ja $L_4(0.5, 0.25, 0.1, 0.05, 0.05, 0.05)$.

    b) Vaatleme 2-ndkoodi, mis on saadud 4-ndkoodist järgmiselt: iga 4-ndkoodi täht, olgu need $\{\alpha, \beta, \gamma, \delta\}$, kodeeritakse pikkusega 2 kahendsõnaks järgmiselt:
    $$\alpha \mapsto 00, \beta \mapsto 01, \gamma \mapsto 10, \delta \mapsto 11.$$
    Nimetagem seda protsessi "topeldamiseks". Leida jaotuse $(0.5, 0.25, 0.1, 0.05, 0.05, 0.05)$ Huffmani 4-ndkoodi topeldamisel saadud kahendkood. Mis on selle keskmine pikkus?

    c) Olgu $L_T(P)$ jaotuse P Huffmani 4-ndkoodi topeldamisel saadud 2-ndkoodi keskmine pikkus. Tõestada, et
    $$L_2(P) \leq L_T \leq L_2(P) + 1.$$

    d) Näita, et ülaltoodud võrratused võivad olla võrdused.

12. Olgu $u_1, u_2, \ldots, u_m$ mittenegatiivsed arvud. Leida järgmise optimeerimisülesande lahend:

$$\min_{l_1,\ldots,l_m} \sum_{i=1}^{m} u_i l_i$$
$$\text{nii, et } \sum_{i=1}^{m} D^{-l_i} \leq 1.$$

13. Olgu jaotuse P aatomid järjestatud $P(x_1) > P(x_2) \geq P(x_3) \geq \cdots \geq P(x_m)$. Leiduvad arvud a ja b nii, et

    - kui $P(x_1) > a$, siis iga Huffmani kahendkoodi korral tähe x_1 koodipikkus on 1;
    - kui $P(x_1) < b$, siis iga Huffmani kahendkoodi korral tähe x_1 koodipikkus on vähemalt 2.

    Leida minimaalne a ja maksimaalne b.

14. Olgu $X_1, \ldots, X_n$ sama jaotusega juhuslikud suurused tähestikul $\mathcal{X}$. Olgu C tähestiku $\mathcal{X}$ mingi kood, C^k olgu C laiend sõnadele $\mathcal{X}^k$. Tõestada, et $L(C^k) = kL(C)$.

15. Olgu Y statsionaarne Markovi ahel üleminekumaatriksiga

$$\begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

    Leida selle protsessi entroopiamäär H_Y. Olgu C_1 C_2 ja C_3 seisundite koodid. Vaatleme järgmist kodeerimisprotseduuri: Y_1 kodeeeri koodiga C_1. Edaspidi kodeeri järgmiselt: Y_2 kodeeri koodiga, mis vastab seisundile Y_1 (st kui $Y_1 = 1$, siis vali C_1, kui $Y_1 = 2$, siis vali C_2 jne), Y_3 kodeeri koodiga mis vastab seisundile Y_2 jne. Kas leiduvad koodid C_1, C_2, C_3 nii, et kirjeldatud kodeerimisprotseduuri korral

$$\lim_n \frac{El(X_1, \ldots, X_n)}{n} = H_Y \quad ?$$

16. Olgu P

a	b	c
0.5	0.25	0.25

    Olgu $X_1, X_2, \ldots$ jaotusega P iid juhuslikud suurused. Olgu C tähestikul $\{a, b, c\}$ antud kood. Vaatleme protsessi

$$Z = Z_1 Z_2 Z_3, \ldots = C(X_1)C(X_2)\ldots$$

    Kas Z on üldiselt statsionaarne protsess?
    Leida Z entroopiamäär, kui kood C on järgmine:

(a)

$$C(x) = \begin{cases} 0, & \text{kui } x = a; \\ 10, & \text{kui } x = b; \\ 11, & \text{kui } x = c. \end{cases}$$

(b)

$$C(x) = \begin{cases} 00, & \text{kui } x = a; \\ 10, & \text{kui } x = b; \\ 01, & \text{kui } x = c. \end{cases}$$

(c)

$$C(x) = \begin{cases} 00, & \text{kui } x = a; \\ 1, & \text{kui } x = b; \\ 01, & \text{kui } x = c. \end{cases}$$

17. Olgu $P(x_1) \geq P(x_2) \geq P(x_3) \geq \cdots P(x_m)$. Defineerime

$$F(x_i) := \sum_{k=1}^{i-1} P(x_k).$$

Tähe x_i kood olgu $F(x_i)$ kahendesitus, millest on võetud $l(x_i) = \lceil -\log P(x_i) \rceil$ komakohta. Tõestada, et saadud kood on prefikskood ning et selle koodi keskmine pikkus l_i rahuldab võrratusi $H(P) \leq L < H(P) + 1$. Ülaldefineeritud koodi nimetatakse ka *Shannoni* koodiks.

3 AEP omadus

3.1 Nõrgalt tüüpilised sõnad

Olgu $X_1, X_2, \ldots$ iid juhuslikud suurused (tähestikul $\mathcal{X}$), $X_i \sim P$. Eeldame

$$H := H(P) < \infty.$$

Olgu $X_1, \ldots, X_n$ esimesed n juhuslikku suurust ülaltoodud jadast. Selle juhusliku vektori väärtuste hulk on $\mathcal{X}^n$, iga võimaliku väärtuse tõenäosus on

$$P(x_1, \ldots, x_n) = P(x_1) \cdots P(x_n).$$

Uurime vektori $X_1, \ldots, X_n$ juhusliku väärtuse tõenäosust $P(X_1, \ldots, X_n)$. Olgu $x^* \in \mathcal{X}$ maksimaalse tõenäosusega täht. Kuigi suurima tõenäosusega võtab vektor $X_1, \ldots, X_n$ väärtuse

$$P^n(x^*) = 2^{n \log P(x^*)},$$

selgub, et suure n korral $P(X_1, \ldots, X_n)$ suure tõenäosusega lähedane arvule 2^{-nH}. Viimane võib olla aga oluliselt väiksem maksimaalsest tõenäosusest $2^{n \log P(x^*)}$ Seda asjaolu võib interpreteerida: suure n korral on praktiliselt kõik realisatsioonid võrdtõenäolised. Sõltumatute ja sama jaotusega juhuslike suuruste jada seda omadust nimetame AEP omaduseks (*asymptotic equipartition property*).

Paneme tähele, et nõrgast suurte arvude seadusest järeldub koondumine

$$-\frac{1}{n} \log P(X_1, \ldots, X_n) = -\frac{1}{n} \sum_{i=1}^{n} \log P(X_i) \xrightarrow{P} -E \log P(X_1) = H. \tag{3.1}$$

Tähistame $x^n := x_1, \ldots, x_n$.

Def 3.1 *Hulga W_ϵ^n moodustavad kõik vektorid (sõnad) $x^n \in \mathcal{X}^n$, mis rahuldavad tingimust*

$$2^{-n(H+\epsilon)} \leq P(x_1, \ldots, x_n) \leq 2^{-n(H-\epsilon)}. \tag{3.2}$$

Tingimust (3.2) rahuldavaid sõnu nimetame **nõrgalt ϵ-tüüpilisteks (weakly ϵ-typical)**.

Teoreem 3.2 (Nõrk AEP) *Iga $\epsilon > 0$ korral*

1 *Kui $x^n \in W_\epsilon^n$, siis*

$$2^{-n(H+\epsilon)} \leq P(x^n) \leq 2^{-n(H-\epsilon)}. \tag{3.3}$$

2 *Piisavalt suure n korral*

$$P(W_\epsilon^n) > 1 - \epsilon. \tag{3.4}$$

3 *Piisavalt suure n korral*

$$(1-\epsilon)2^{n(H-\epsilon)} \leq |W_\epsilon^n| \leq 2^{n(H+\epsilon)}. \tag{3.5}$$

Tõestus. Omadus **1** järeldub vahetult definitsioonist (3.2).

Omadus **2** järeldub vahetult koondumisest (3.1), sest tõenäosuse järgi koondumis definitsioonist johtuvalt $\forall \epsilon > 0$ korral leidub n_o nii, et

$$\mathbf{P}\Big(\big| -\frac{1}{n}\sum_{i=1}^{n} \log P(X_i) - H\big| \le \epsilon\Big) \ge 1-\epsilon, \tag{3.6}$$

kui $n > n_o$.

Et nõrgalt tüüpilise sõna tõenäosus on vähemalt $2^{-n(H+\epsilon)}$, siis

$$1 \ge P(W_\epsilon^n) = \sum_{x^n \in W_\epsilon^n} P(x^n) \ge |W_\epsilon^n| 2^{-n(H+\epsilon)},$$

millest

$$|W_\epsilon^n| \le 2^{n(H+\epsilon)}.$$

Paneme tähele, et saadud tõke kehtib iga n korral. Teisest küljest, et suure n korral $P(W_\epsilon^n) > 1-\epsilon$ ning iga nõrgalt tüüpilise sõna tõenäosus on ülimalt $2^{-n(H-\epsilon)}$, siis

$$1-\epsilon \le P(W_\epsilon^n) = \sum_{x^n \in W_\epsilon^n} P(x^n) \le |W_\epsilon^n| 2^{-n(H-\epsilon)},$$

millest

$$|W_\epsilon^n| \ge (1-\epsilon)2^{n(H-\epsilon)}.$$

■

Seega on suure n korral nõrgalt tüüpiliste sõnade hulga W_ϵ^n mõõt praktiliselt üks. Tõenäosus, et iid. juhusliku vektori $X_1, \ldots, X_n$ väärtus pole nõrgalt tüüpiline on väga väike. Kõikide nõrgalt tüüpiliste sõnade tõenäosus on umbes 2^{-nH} ehk kõik nõrgalt tüüpilised sõnad on sisuliselt võrdtõenäosused. Samas on (suure n korral) nõrgalt tüüpiliste sõnade osakaal kõikide pikkusega n sõnade seas väga väike. Tõepoolest, olgu $H < \log|\mathcal{X}| < \infty$. Siis nõrgalt tüüpiliste sõnade osakaal läheb nulliks, sest (piisavalt väikese ϵ korral)

$$\frac{|W_\epsilon^n|}{|\mathcal{X}|^n} \le \frac{2^{n(H+\epsilon)}}{2^{n\log|\mathcal{X}|}} = 2^{n(H+\epsilon-\log|\mathcal{X}|)} \to 0.$$

Nõrk AEP omadus annab järjekordse interpretatsiooni entroopiale.

Näide: Olgu $X_1, \ldots, X_n$ iid Bernoulli p-jaotusega. Siis

$$P(x_1, \ldots, x_n) = p^k(1-p)^{n-k}, \quad k = \sum_{i=1}^{n} x_i.$$

Seega

$$-\frac{1}{n}\log P(x_1, \ldots, x_n) = -\frac{k}{n}\log p - \frac{n-k}{n}\log(1-p),$$

millest järelduvalt on $x_1, \ldots, x_n$ nõrgalt tüüpiline, kui ühtede proportsioon on peaaegu p.

3.1.1 Nõrk AEP ja kodeerimine

Nõrga AEP omaduse abil on lihtne näha, et suure n korral on iid vektorit $X_1, \ldots, X_n$ tõepoolest võimalik kodeerida nii, et keskmine koodipikkus tähe kohta on võrdne entroopiaga. Vaatleme olukorda $D = 2$, suurema D korral on kodeerimine analoogiline.

Olgu $X_1, \ldots, X_n$ iid juhuslikud suurused lõplikul tähestikul $\mathcal{X}$. Fikseerime $\epsilon > 0$ ja jagame kõikvõimalike sõnade hulga $\mathcal{X}^n$ kaheks: nõrgalt tüüpilised sõnad W_ϵ^n ning ülejäänud. Järjestame mõlemad hulgad ning kodeerime nende indekseid. Et $|W_\epsilon^n| \leq 2^{n(H+\epsilon)}$, siis kõigi nõrgalt tüüpiliste sõnade indeksite kodeerimist binaarsteks koodisõnadeks on võimalik teha nii, et kood on ühene ja iga koodisõna pikkus on $\lceil n(H+\epsilon)\rceil \leq n(H+\epsilon)+1$. Liidame nendele sõnadele prefiksi 0, mis näitab kuulumist nõrgalt tüüpiliste sõnade hulka. Seega

$$l(x^n) = \lceil n(H+\epsilon)\rceil + 1 \leq n(H+\epsilon) + 2, \quad \forall\, x^n \in W_\epsilon^n.$$

Ülejäänud sõnad kodeerime samuti võrdse pikkusega koodisõnadeks. Iga hulga $\mathcal{X}^n$ elemendi (neid on ju $2^{n\log|\mathcal{X}|}$) saab kodeerida nii, et kood on ühene ja koodisõna pikkus on $\lceil n\log|\mathcal{X}|\rceil \leq n\log|\mathcal{X}| + 1$. Kasutamegi seda lihtsat koodi ja liidame koodisõnadele prefiksi 1, mis näitab kuulumist hulka $\mathcal{X}^n \setminus W_\epsilon^n$. Seega

$$l(x^n) = \lceil n\log|\mathcal{X}|\rceil + 1 \leq n\log|\mathcal{X}| + 2, \quad \forall\, x^n \notin W_\epsilon^n.$$

Saadud kood on prefikskood, sest esimene bitt näitab järgneva koodi pikkuse. Loomulikult pole kirjeldatud kood optimaalne, sest hulka W_ϵ^n mittekuuluvaid sõnu kodeerisime väga mõtlematult.
Leiame saadud koodi keskmise pikkuse

$$\begin{aligned} L &= \sum_{x^n\in\mathcal{X}^n} l(x^n)P(x^n) = \sum_{x^n\in W_\epsilon^n} l(x^n)P(x^n) + \sum_{x^n\notin W_\epsilon^n} l(x^n)P(x^n) \\ &\leq \sum_{x^n\in W_\epsilon^n} (n(H+\epsilon)+2)P(x^n) + \sum_{x^n\notin W_\epsilon^n} (n\log|\mathcal{X}|+2)P(x^n) \\ &= P(W_\epsilon^n)(n(H+\epsilon)+2) + (1-P(W_\epsilon^n))(n\log|\mathcal{X}|+2). \end{aligned}$$

Seega, kui n on piisavalt suur, siis Teoreemi 3.2 väite **2** tõttu $P(W_\epsilon^n) \leq \epsilon$ nii, et

$$L \leq n(H+\epsilon) + \epsilon(n\log|\mathcal{X}|) + 2 = n(H+\epsilon'),$$

kus $\epsilon' = \epsilon + \epsilon\log|\mathcal{X}| + \frac{2}{n}$ ja selle võib (sobiva ϵ ja n valikul) teha kuitahes väikeseks.

Kokkuvõtteks: Oleme tõestanud, et iga $\epsilon > 0$ korral leidub n ja AEP omadusel põhinev prefikskood $C : \mathcal{X}^n \to \{0,1\}^*$ nii, et

$$H \leq L_n(C) < H + \epsilon. \tag{3.7}$$

3.1.2 Suurima tõepäraga hulk

Eelmises peatükis kirjeldatud lihtne meetod kodeerimiseks keskmise pikkusega nH sai võimalikuks tänu sellel, et suure n korral leidus hulk W_ϵ^n nii, et tema tõenäosus on kuitahes suur, kuid elementide arv võrreldes kõikide sõnade arvuga väike (juhul, kui $H < \log|\mathcal{X}|$). Samas ei kuulu hulka W_ϵ^n üldjuhul kõige suurema tõepäraga sõnad, mistõttu W_ϵ^n pole väikseim (sõnade arvu mõttes) hulk, mille tõenäosus on vähemalt $1-\epsilon$. Olgu B_ϵ^n väikseim hulk mis rahuldab tingimust $P(B_\epsilon^n) \geq 1-\epsilon$. Seega kui eelmises peatükis kirjeldatud koodis hulga W_ϵ^n asemel võtta hulk B_ϵ^n, väheneb keskmine koodipikkus. Kas ka oluliselt? Võrratustest (3.7) on selge, et väga oluliselt keskmine koodimikkus väheneda ei saa. See tuleneb asjaolust, et kuigi $|W_\epsilon^n| \geq |B_\epsilon^n|$ ning enamikul juhtudest on see võrratus range, on nende hulkade elementide arv sama suurusjärku st $|B_\epsilon^n| \approx 2^{nH}$. Veendume selles.

Lemma 3.1 *Iga $1 > \epsilon > 0$ ja $\delta > 0$ korral leidub n nii suur, et*

$$|B_\epsilon^n| \geq 2^{n(H-\delta)} \tag{3.8}$$

Tõestus. Valime $\epsilon_1 > 0$ nii väikese, et $\epsilon_1 < \delta$ ja $\epsilon_1 + \epsilon < 1$. Olgu n nii suur, et

$$P(W_{\epsilon^1}^n) > 1 - \epsilon_1. \tag{3.9}$$

(sellise n olemasolu järeldub Teoreemist 3.2) ning lisaks kehtib

$$\epsilon_1 - \frac{\log(1-(\epsilon+\epsilon_1))}{n} < \delta. \tag{3.10}$$

Defineerime

$$S := W_{\epsilon_1}^n \cap B_\epsilon^n.$$

Siis

$$1-(\epsilon_1+\epsilon) \leq P(S) = \sum_{x^n \in S} P(x^n) \leq |S| 2^{-n(H-\epsilon_1)} \leq |B_\epsilon^n| 2^{-n(H-\epsilon_1)},$$

kus esimene võrratus järeldub B_ϵ^n definitsioonist ja võrratusest (3.9) ning teine võrratus järeldub $W_{\epsilon_1}^n$ definitsioonist. Seega

$$\log|B_\epsilon^n| \geq \log(1-(\epsilon+\epsilon_1)) + n(H-\epsilon_1) = n\Big(\frac{\log(1-(\epsilon+\epsilon_1))}{n} + H - \epsilon_1\Big) \geq n(H-\delta).$$

Viimane võrratus tuleb seosest (3.10). ■

3.1.3 Näide

Olgu $X_1, \ldots, X_{25}$ iid $B(1, 0.1)$ jaotusega juhuslikud suurused. Seega võimalikke vektoreid x^n on 2^{25}. Alljärgnevas tabelis on kõik vektorid x^n jaotatud klassidesse ühtede arvu k järgi. Ühte klassi kuuluvad vektorid on võrdse tõenäosusega. Teises veerus on klassi kuuluvate vektorite arv ja kolmandas veerus on klassi kuuluvate vektorite tõenäosuste

summa: klassi tõenäosus. Neljandas veerus on suurus $\frac{1}{n}\log P(x^n)$, kus $P(x^n)$ on klassi kuuluva *ühe* vektori tõenäosus (mitte klassi tõenäosus).

Arvestades, et $h(0.1) = 0.468996$, ja võttes $\epsilon = 0.2$, same, et hulka $W_{0.2}^{25}$ kuuluvad klasside $k = 1, 2, 3, 4$ elemendid. Seega

$$P(W_{0.2}^{25}) = 0.199416 + 0.265888 + 0.226497 + 0.138415 = 0.830216 \geq 1 - \epsilon.$$

Samas $|W_{0.2}^{25}| = 25 + 300 + 2300 + 12650 = 15275$, millest

$$\frac{1}{25}\log|W_{0.2}^{25}| \approx 0.556 \in (468996 - 0.2, 468996 + 0.2)$$

Seega $W_{0.2}^{25}$ rahuldab tingimusi (3.4) ja (3.5).

Leiame hulga B_n^{25}. Antud näite korral vektorite tõenäosused kahanevad ülalt alla: kõige suurema tõenäosusega vektor koosneb nullidest ja moodustab esimese klassi (selle tõenäosus on 0.0717898); vektorid, milles on vaid 1 null on tõenäosuse järgi teisel kohal, sellise vektori tõenäosus on $0.199416/25 = 0.00797$ jne. Seega hulga $B_{0.2}^{25}$ moodustamine hakkab ülalt kuni klassi mass ületab 0.8. Esimese nelja klassi kogumass on 0.7635908, seega kuuluvad need klassid hulka B_n^{25}. Lisaks peame veel võtma elemente viiendast klassist ($k = 4$). Selle klassi elementide tõenäosus on $\frac{0.138415}{12650} = 0.0000109419$. Seega tuleb sellest klassist võtta

$$\left\lceil \frac{0.8 - 0.7635908}{0.0000109419} \right\rceil = 3328$$

elementi. Seega

$$|B_{0.2}^{25}| = 1 + 25 + 300 + 2300 + 3325 = 5951$$

ning

$$\frac{1}{25}\log|B_{0.2}^{25}| \approx 0.501.$$

Kuigi hulkadesse $B_{0.2}^{25}$ ja $W_{0.2}^{25}$ kuuluvad klassid sisuliselt on samad (esimene klass koosneb vaid ühest elmendist ega oma seega suurt tähtsust), tuleneb võimsuste vahe sellest, et klass $k = 4$ kuulus hulka $W_{0.2}^{25}$ täielikult, kuid hulka $B_{0.2}^{25}$ vaid osaliselt.

k	$\binom{n}{k}$	$\binom{n}{k}p^k(1-p)^{n-k} = \binom{n}{k}P(x^n)$	$-\frac{1}{n}\log P(x^n)$
0	1	0.0717898	0.152003
1	25	0.199416	**0.2788**
2	300	0.265888	**0.405597**
3	2300	0.226497	**0.532394**
4	12650	0.138415	**0.659191**
5	53130	0.0645937	0.785988
6	177100	0.0239236	0.912785
7	480700	0.00721505	1.03958
8	1081575	0.00180376	1.16638
9	2042975	0.000378567	1.29318
10	3268760	0.0000673009	1.41997
11	4457400	0.0000101971	1.54677
12	5200300	1.32185×10^{-6}	1.67357
13	5200300	1.46872×10^{-7}	1.80036
14	4457400	1.39878×10^{-8}	1.92716
15	3268760	≈ 0	2.05396
16	2042975	≈ 0	2.18076
17	1081575	≈ 0	2.30755
18	480700	≈ 0	2.43435
19	177100	≈ 0	2.56115
20	53130	≈ 0	2.68794
21	12650	≈ 0	2.81474
22	2300	≈ 0	2.94154
23	300	≈ 0	3.06833
24	25	≈ 0	3.19513
25	1	≈ 0	3.32193

3.2 Nõrgalt ühistüüpilised sõnad

Olgu $P(x,y)$ jaotus hulgal $\mathcal{X}\times\mathcal{Y}$, $(X,Y)\sim P$. Vaatleme iid juhuslikke vektoreid $(X_1,Y_1),\ldots,(X_n,Y_n)$. Siis iga $(x^n,y^n)\in\mathcal{X}^n\times\mathcal{Y}^n$ korral

$$P(x^n,y^n)=\prod_{i=1}^{n}P(x_i,y_i).$$

Def 3.3 *Hulga W_ϵ^n moodustavad kõik sõnapaarid $(x^n,y^n)\in\mathcal{X}^n\times\mathcal{Y}^n$, mis rahuldavad tingimusi*

- $2^{-n(H(X)+\epsilon)}\le P(x^n)\le 2^{-n(H(X)-\epsilon)}$
- $2^{-n(H(Y)+\epsilon)}\le P(y^n)\le 2^{-n(H(Y)-\epsilon)}$
- $2^{-n(H(X,Y)+\epsilon)}\le P(x^n,y^n)\le 2^{-n(H(X,Y)-\epsilon)}$.

Neid tingimusi rahuldavid sõnu nimetatakse **nõrgalt ϵ- ühistüüpilisteks (jointly ϵ-typical)**.

Seega on paar (x^n, y^n) nõrgalt ühistüüpiline, kui nii x^n ja y^n on nõrgalt tüüpilised ning sõnapaari (x^n, y^n) ühistõenäosus on ligikaudu $2^{-nH(X,Y)}$.

Olgu P_x ja P_y mõõdu P marginaaljaotused. Siis $P_x \times P_y$ on samade marginaalidega sõltumatute komponentidega vektori jaotus. Tähistame

$$P_x \times P_y(x^n, y^n) := \prod_{i=1}^{n} P_x \times P_y(x_i, y_i) = \prod_{i=1}^{n} P_x(x_i) P_y(y_i).$$

Tõestame nüüd teoreemi 3.2 kahemõõtmelise versiooni.

Teoreem 3.4 *Iga $\epsilon > 0$ korral*

1 *Piisavalt suure n korral*

$$P(W_\epsilon^n) > 1 - \epsilon. \tag{3.11}$$

2 *Piisavalt suure n korral*

$$(1-\epsilon)2^{n(H(X,Y)-\epsilon)} \le |W_\epsilon^n| \le 2^{n(H(X,Y)+\epsilon)}. \tag{3.12}$$

3 *Piisavalt suure n korral*

$$(1-\epsilon)2^{-n(I(X;Y)+3\epsilon)} \le P_x \times P_y(W_\epsilon^n) \le 2^{-n(I(X;Y)-3\epsilon)}.$$

Tõestus. Tõestus on analoogiline teoreemi 3.2 tõestusega. Väide **1** järeldub sellest, et

$$\begin{aligned}
-\frac{1}{n}\log P(X_1, \ldots, X_n) &= -\frac{1}{n}\sum_{i=1}^{n} \log P(X_i) \xrightarrow{P} H(X) \\
-\frac{1}{n}\log P(Y_1, \ldots, Y_n) &= -\frac{1}{n}\sum_{i=1}^{n} \log P(Y_i) \xrightarrow{P} H(Y) \\
-\frac{1}{n}\log P\big((X_1, Y_1), \ldots, (X_n, Y_n)\big) &= -\frac{1}{n}\sum_{i=1}^{n} \log P(X_i, Y_i) \xrightarrow{P} H(X, Y)
\end{aligned}$$

(ülesanne 1). Ka väite **2** tõestus on analoogiline:

$$\begin{aligned}
1 \ge P(W_\epsilon^n) &= \sum_{(x^n, y^n) \in W_\epsilon^n} P(x^n, y^n) \ge |W_\epsilon^n| 2^{-n(H(X,Y)+\epsilon)}, \\
1 - \epsilon \le P(W_\epsilon^n) &= \sum_{(x^n, y^n) \in W_\epsilon^n} P(x^n, y^n) \le |W_\epsilon^n| 2^{-n(H(X,Y)-\epsilon)},
\end{aligned}$$

millest

$$(1-\epsilon)2^{n(H(X,Y)-\epsilon)} \le |W_\epsilon^n| \le 2^{n(H(X,Y)+\epsilon)}.$$

Korrutismõõdu korral

$$\begin{aligned}
P_x \times P_y(W_\epsilon^n) &= \sum_{(x^n,y^n)\in W_\epsilon^n} P(x^n)P(y^n) \\
&\leq \sum_{(x^n,y^n)\in W_\epsilon^n} 2^{-n(H(X)-\epsilon)}2^{-n(H(Y)-\epsilon)} \\
&\leq 2^{n(H(X,Y)+\epsilon)}2^{-n(H(X)-\epsilon)}2^{-n(H(Y)-\epsilon)} \\
&= 2^{-n(I(X;Y)-3\epsilon)} \\
P_x \times P_y(W_\epsilon^n) &\geq (1-\epsilon)2^{n(H(X,Y)-\epsilon)}2^{-n(H(X)+\epsilon)}2^{-n(H(Y)+\epsilon)} \\
&= (1-\epsilon)2^{-n(I(X;Y)+3\epsilon)}.
\end{aligned}$$

■

Teoreemi 3.4 esimese kahe väite interpretatsioon jääb samaks: nõrgalt ühistüüpiliste sõnade hulga tõenäosus on ligikaudu üks, kõik nõrgalt ühistüüpilised sõnad on praktiliselt võrdtõenäolised ja nende arv on ligikaudu $2^{nH(X,Y)}$. Tarvilik tingimus sõnapaari (x^n, y^n) (nõrgalt) ühistüüpilisuseks on kummagi sõna (nõrk) tüüpilisus. Paare, kus mõlemad sõnad on (nõrgalt) tüüpilised on ligikaudu $2^{nH(X)}2^{nH(Y)}$ Paneme aga tähele, et üldiselt $2^{nH(X,Y)} < 2^{nH(X)}2^{nH(Y)}$. Seega on selliste paaride seas on vaid väike osa ühistüüpilisi paare. Fikseeritud esimese sõna x^n korral on ühistüüpiliste paaride (x^n, y^n) arv keskmiselt $2^{n(H(X,Y)-H(X))} = 2^{nH(Y|X)}$. Valides teise (nõrgalt tüüpilise) sõna y^n juhuslikult üle kõigi tüüpiliste sõnade (ühtlase jaotusega), saame, et selline juhuslik sõltumatute komponentidega sõnapaar on ühistüüpiline (ligikaudse) tõenäosusega $2^{nH(Y|X)-nH(Y)} = 2^{-nI(X;Y)}$. See ongi sisuliselt teoreemi kolmas väide: kui paar (x^n, y^n) on valitud juhuslikult (vastavalt antud marginaaljaotustele), kusjuures sõna y^n ei sõltu sõnast x^n, on see paar ühistüüpiline tõenäosusega $2^{-nI(X;Y)}$. Mida suurem on vastastikune informatsioon, seda väiksem on nimetatud tõenäosus ning seda raskem on juhuslikult kokku saada ühistüüpilist paari.

Näide: Olgu $\mathcal{X} = \mathcal{Y} = \{0, 1\}$ ning olgu jaotustabel

$\mathcal{X}\backslash\mathcal{Y}$	1	0
1	$\frac{7}{80}$	$\frac{1}{80}$
0	$\frac{9}{80}$	$\frac{63}{80}$

Seega $X \sim B(1, 0.1)$, $Y \sim B(1, 0.2)$. Ühisentroopia

$$H(X,Y) = H(X) + H(Y|X) = h(\frac{1}{10}) + h(\frac{7}{8}).$$

Sõnad $x^n = 1000000000$ ja $y^n = 0110000000$ on mõlemad nõrgalt tüüpilised (suvalise ϵ korral) ehk

$$x^n \in W_\epsilon^{10}, \quad y^n \in W_\epsilon^{10}.$$

Tähistame $p = \frac{1}{10}, q = \frac{1}{8}$ ja leiame

$$P(x^n, y^n) = \left(\frac{1}{80}\right)\left(\frac{9}{80}\right)^2\left(\frac{63}{80}\right)^7 = (pq)((1-p)q)^2((1-p)(1-q))^7 = q^3(1-q)^7(1-p)^9 p.$$

$$\begin{aligned}\frac{1}{n}\log P(x^n, y^n) &= \frac{3}{10}\log q + \frac{7}{10}\log(1-q) + \frac{9}{10}\log(1-p) + \frac{1}{10}\log p \\ &= q\log q + \frac{7}{40}\log q - \frac{7}{40}\log(1-q) + (1-q)\log(1-q) + (1-p)\log(1-p) + p\log p \\ &= -h(q) - h(p) + \frac{7}{40}\log(\frac{q}{1-q}).\end{aligned}$$

Järelikult

$$-\frac{1}{n}\log P(x^n, y^n) - H(X,Y) = \frac{7}{40}\log(7),$$

mistõttu

$$(x^n, y^n) \notin W_\epsilon^{10},$$

kui $\epsilon < \frac{7}{40}\log(7)$.

3.3 Nõrga AEP omadusega protsessid

Nõrk AEP omadus (teoreemid 3.2 ja 3.4) põhinevad sõltutmatute sama jaotusega juhuslike suuruste (iid protsessi) $X = \{X_n\}_{n=1}^{\infty}$ omadusel

$$-\frac{1}{n}\log P(X_1, \ldots, X_n) \to H_X, \quad \text{p.k.}, \tag{3.13}$$

kus H_X on X_i entroopia ja seega protsessi entroopiamäär. Sõltumatuse korral järeldub (3.13) vahetult (tugevast) suurte arvude seadusest. Selgub aga, et koondumine (3.13) ei kehti mitte ainult iid protsesside korral vaid ka mitmete teiste statsionaarsete protsesside korral (tuleta meelde, et statsionaarsel protsessil on alati defineeritud entroopiamäär). Sellisel juhul, arusaadavalt, kehtivad ka teoreemi 3.2 kõik väited.

Def 3.5 *Protsessil $X_1, X_2 \ldots$ on* **AEP omadus (AEP property)** *, kui kehtib (3.13), kus H_X on protsessi entroopiamäär.*

Nõrga AEP omadusega on kõik *ergoodilised* protsessid. Näiteks lahutamatu Markovi ahel.

3.4 Ülesanded

1. Tõestada teoreemi 3.4 väide **1**.

2. Olgu $X_1, X_2, \ldots$ iid juhuslikud tähed jaotusega P. Olgu Q mingi teine tähestikul $\mathcal{X}$ antud jaotus. Vaatleme tõepärasuhet

$$\frac{Q(X_1)\cdots Q(X_n)}{P(X_1)\cdots P(X_n)}.$$

   Tõestada, et leidub hulk $A_\epsilon^n \subset \mathcal{X}^n$ ja konstant A nii, et

1 Kui $x^n \in A_\epsilon^n$, siis

$$2^{-n(A+\epsilon)} \leq \frac{Q(x^n)}{P(x^n)} \leq 2^{-n(A-\epsilon)};$$

2 piisavalt suure n korral

$$P(A_\epsilon^n) > 1 - \epsilon;$$

3 piisavalt suure n korral

$$(1-\epsilon)2^{-n(A+\epsilon)} \leq Q(A_\epsilon^n) \leq 2^{-n(A-\epsilon)}.$$

3. Olgu $X_1, X_2, \ldots$ iid juhuslikud suurused, $X_i \sim U[0,1]$ (ühtlane jaotus). Konstrueerime n-tahuka küljepikkustega $X_1, \ldots, X_n$, selle tahuka ruumala on $V_n = \prod_{i=1}^n X_i$. Sama ruumalaga n-kuubi küljepikkus on ${V_n}^{\frac{1}{n}}$. Leida $E({V_n}^{\frac{1}{n}})$, $\lim_n E({V_n}^{\frac{1}{n}})$ ja $\lim_n {V_n}^{\frac{1}{n}}$ (tõenäosuse järgi) ning võrdluseks leia $(EV_n)^{\frac{1}{n}}$ ja $\lim_n (EV_n)^{\frac{1}{n}}$.

4. Olgu $X_1, X_2, \ldots$ lõpliku seisundite hulgaga statsionaarne Markovi ahel üleminekumaatriksiga I (ühikmaatriks). Tõestada koondumine (3.13).

4 Infovahetus läbi kanali

Käsitleme informatsiooni edastamist läbi diskreetse (näiteks binaarse) infokanali. Selleks kodeerime edastatava teksti (binaarse infokanali korral kahendkoodi abil) ja sisestame saadud koodi bitikaupa kanalisse. Vastuvõtja dekodeerib saadud jada. Selline süsteem ei tekita mingeid probleeme kui kanal töötab vigadeta, s.t. iga sisestatud sümbol väljub iseendana. Paraku pole see alati nii – sisestatud sümbolid võivad kanalis teatud tõenäosusega muutuda või kaduda. Sellisel juhul ei pruugi vastuvõetud tekst olla identne saadetuga ning informatsioon läheb kaotsi. Alljärgnevas uurime, kuids ülalkirjeldatud vigase kanali abil informatsiooni võimalikult täpselt vahetada.

4.1 Diskreetne kanal

Olgu $\mathcal{X}$ mingi lõplik tähestik. Seda interpreteerime kui *sisendtähestikku*. Olgu $\mathcal{Y}$ mingi teine lõplik tähestik, mida interpreteerime kui *väljundtähestikku*. Meie käsitluses on diskreetne kanal üleminekutõenäosuste maatriks

$$\big(P(y|x)\big)_{x\in\mathcal{X},y\in\mathcal{Y}}. \tag{4.1}$$

Arv $P(y|x)$ on tõenäosus, et sümboli x – *sisend*– sisendamisel kanalisse väljub sealt sümbol y – *väljund*. Selline kanal on **diskreetne kanal (discrete channel)**. Kanal on **mäluta (memoryless)**, kui väljund sõltub ainult sisendist, kuid mitte eelnevatest sisenditest või väljundistest. Vigadeta kanali korral on üleminekumaatriks ühikmaatriks.

Kanali võimsus. Olgu nüüd $P(x)$ mingi jaotus sisendtähestikul $\mathcal{X}$. Seda interpreteerime kui sisendite jaotust. Koos kanaliga (4.1), saame nüüd mingi ühisjaotuse $P(x,y) = P(x)P(y|x)$ tähestikul $\mathcal{X}\times\mathcal{Y}$. Olgu nüüd $(X,Y)\sim P(x,y)$ antud ühisjaotusega juhuslik vektor. s.t. X on jaotusega $P(x)$ juhuslik sisend ning Y on juhuslik väljund.

Def 4.1 *Kanali (4.1)* **võimsus (capacity)** *on*

$$C = \max_{P(x)} I(X;Y),$$

kus maksimum on võetud üle kõikide võimalike sisendjaotuste hulgal $\mathcal{X}$.

Märkused:

- Funktsioonil $P(x)\mapsto I(X;Y)$ on pidev ning kõikide sisendjaotuste hulk on ruumi $\mathbb{R}^{|\mathcal{X}|}$ kompaktne kumer alamhulk (simpleks). Seega on funktsioonil $P(x)\mapsto I(X;Y)$ maksimum. Saab näidata, et see funktsioon on nõgus, mistõttu lokaalne maksimum on ka globaalne ning maksimumi võib leida kumerate optimiseerimismeetoditega.
- Kanali võimsus rahuldab võrratust: $C\le \log\min\{|\mathcal{X}|,|\mathcal{Y}|\}$, sest
  $C = \max\limits_{P(x)} I(X;Y) \le \max\limits_{P(x)} H(X) \le \log|\mathcal{X}|, \quad C = \max\limits_{P(x)} I(X;Y) \le \max\limits_{P(x)} H(Y) \le \log|\mathcal{Y}|.$
- Kanali võimsust võib interpreteerida kui maksimaalset infohulka, mida ühe edastamise käigus läbi kanali on võimalik saata.

4.2 Näiteid kanalitest

Vigadeta binaarne kanal. Sellise kanali korral $\mathcal{X} = \mathcal{Y} = \{0, 1\}$ ning $P(y|x)$ on ühikmaatriks. Seega iga sisestatud bitt edastatakse muutmatuna. On selge, et ühe edastamise käigus saabki maksimaalselt edastada ühe biti, seega sellise kanali võimsus on 1, mis ühtlasi on ka maksimaalne võimsus, mis binaarsel kanalil võib olla. Formaalselt $I(X;Y) = H(X;X) = H(X)$, millest

$$C = \max_{P(x)} H(X) = 1,$$

kus maksimim saavutatakse $B(1, \frac{1}{2})$ jaotuse korral.

Ebaoluliste vigadega kanal. Selle kanali korral $\mathcal{X} = \{0, 1\}$, $\mathcal{Y} = \{0, 1, 2, 3\}$, üleminekumaatriks on

$$\begin{pmatrix} p & 1-p & 0 & 0 \\ 0 & 0 & q & 1-q \end{pmatrix}$$

Sellises kanalis on küll üksjagu juhuslikkust, kuid erinenevatele sisenditele vastavate väljundite hulgad on lõikumatud. Seega määrab väljund (selle klass) üheselt sisendi ja kanal on vigadeta. Arusaadavalt on selle kanali võimsus samuti 1. Formaalselt

$$C = \max_{P(x)} \big(H(X) - H(X|Y)\big) = \max_{P(x)} H(X) = 1,$$

sest $X = f(Y)$ ja seetõttu $H(X|Y) = 0$.

Vigadega klaviatuur. Siin $\mathcal{X} = \mathcal{Y}$ on tähestik, $|\mathcal{X}| = 26$. Vigase klaviatuuri korral iga $x \in \mathcal{X}$ korral

$$P(x|x) = P(\text{järgmine täht}|x) = 0.5.$$

Seega sellise klaviatuuri korral edastatakse täht vigadeta vaid pooltel juhtudel. Ülejäänud juhtudel edastatakse järgmine täht. Leiame võimsuse

$$C = \max_{P(x)} \big(H(Y) - H(Y|X)\big) = \max_{P(x)} H(Y) - 1 = \log 26 - 1 = \log 13,$$

kusjuures maksimum saavutatakse ühtlase sisendjaotuse korral. Saadud võimsus ühtib intuitsiooniga – kui vigadeta klaviatuuri korral edastame korraga maksimaalselt $\log 26$ bitti informatsiooni, siis vigase klavituuri korral saame vigadeta edastada vaid pooltest tähtedest.

Binaarne sümmeetriline kanal. Siin $\mathcal{X} = \mathcal{Y} = \{0, 1\}$ ja üleminekumaatriks on

$$\begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}$$

Seega sümbol edastetekse täpselt tõenäosusega $1-p$, kuid tõenäosusega p muutub ta teiseks sümboliks. Leiame vastastikuse informatsiooni

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - \sum_x P(x)H(Y|X=x)$$
$$= H(Y) - \sum_x P(x)h(p) = H(Y) - h(p).$$

Seega on $I(X;Y)$ maksimaalne siis, kui Y on ühtlase jaotusega. See saavutatakse ühtlase sisendjaotuse korral ning seega

$$C = \max_{P(x)} I(X;Y) = 1 - h(p).$$

Kui $p = 0$, on kanal vigadeta ning tema võimsus on 1. Kui $p = 0.5$, on X ja Y sõltumatud. Sellisel juhul ei toimu mingisugust infovahetust ning kanali võimsus on arusaadavalt 0.

J. Thomas and T. Cover: "This is the simplest model of a channel with errors; yet it captures most of the complexity of the general problem".

Binaarne kadumiskanal. Sellisel juhul $\mathcal{X} = \{0,1\}$ ja $\mathcal{Y} = \{0,1,e\}$. Sümbolit e interpreteerime kui signaali selle kohta, et sisend on kaduma läinud (vaikus). Kumbki signaal läheb kaduma tõenäosusega p. Üleminekumaatriks on selline, et

$$P(x|x) = 1-p, \quad P(e|x) = p, \quad x = 0,1.$$

Leiame binaarse kadumiskanali võimsuse

$$C = \max_{P(x)}\big(H(Y) - H(Y|X)\big) = \max_{P(x)} H(Y) - h(p).$$

Leidmaks $\max_{P(x)} H(Y)$ defineerime sündmuse $E = \{Y = e\}$. Et $E = f(Y)$, siis

$$H(Y) = H(Y,E) = H(E) + H(Y|E) = h(p) + H(Y|E).$$

Olgu $\pi = P(X=1)$. Siis $P(Y=1|Y\neq e) = \pi$ ja $P(Y=0|Y\neq e) = (1-\pi)$ ja

$$H(Y|E) = H(Y|Y\neq e)P(Y\neq e) = h(\pi)(1-p).$$

Seega

$$C = \max_{P(x)} H(Y|E) = \max_{\pi} h(\pi)(1-p) = 1-p.$$

Sümmeetriline kanal. Selle kanali korral koosnevad üleminekumaatriksi read samadest elementidest. Teisisõnu, maatriksi read on ükseteise permutatsioonid. Samuti on permutatsioonid üleminekumaatriksi veerud. Sümmeetrilised kanalid on näiteks

$$\begin{pmatrix} 0.3 & 0.2 & 0.5 \\ 0.5 & 0.3 & 0.2 \\ 0.2 & 0.5 & 0.3 \end{pmatrix} \quad \begin{pmatrix} 0.2 & 0.2 & 0.3 & 0.3 \\ 0.3 & 0.3 & 0.2 & 0.2 \end{pmatrix}.$$

Sellise kanali võimsust on kerge leida. Olgu rea entroopia H_r. Siis

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - H_r \leq \log|\mathcal{Y}| - H_r,$$

kusjuures võrdus kehtib ühtlase väljundjaotuse korral. Veendume, et ühtlane sisendjaotus garanteerib ühtlase väljundjaotuse. Ühtlase sisendjaotuse korral

$$P(y) = \sum_{x\in\mathcal{X}} P(y|x)P(x) = \frac{1}{|\mathcal{X}|}\sum_{x} P(y|x) = \frac{c}{|\mathcal{X}|},$$

kus c on veeruelementide summa. Saadud arv ei sõltu y-st, mistõttu on välundjaotus ühtlane ja

$$C = \log|\mathcal{Y}| - H_r.$$

Ülaltoodud argument kehtib ka siis, kui üleminekumaatriksi read on üksteise permutatsioonid ja veergude summa on konstantne (kuid veerud ei pruugi olla üksteise permutatsioonid). Selliseid kanalaeid nimetatakse *nõrgalt sümmeetrilisteks*. Nõrgalt sümmeetriline kuid mitte sümmeetriline kanal on näiteks

$$\begin{pmatrix} \frac{1}{3} & \frac{1}{6} & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & \frac{1}{6} \end{pmatrix}.$$

J. Thomas and T. Cover: "In general, there are no closed form solution for the capacity. but for many simple channels it is possible to calculate the capacity using properties like symmetry."

4.3 Kanaliteoreem

Infovahetus läbi kanali. Olgu $\{1, 2, \ldots, M\}$ sõnad. Nende seast valitakse juhuslikult üks. Olgu juhuslik suurus W see juhuslik sõna. Sõna W kodeeritakse n-elemendiliseks koodisõnaks. Olgu

$$\mathcal{C} : \{1, 2, \ldots, M\} \mapsto \mathcal{X}^n$$

kood. Kodeeritud sõna (n-dimensionaalne juhuslik vektor) $X^n := \mathcal{C}(W)$ saadetakse bitikaupa läbi kanali

$$\{P(y|x)\}_{x\in\mathcal{X}, y\in\mathcal{Y}}.$$

Et kanal on mäluta, siis tõenäosus sõna y^n saamiseks sõna x^n sisestamisel on

$$P(y^n|x^n) = \prod_{i=1}^{n} P(y_i|x_i).$$

Saadud sõna, olgu see Y^n, dekodeeritakse. Olgu

$$g : \mathcal{Y}^n \to \{1, 2, \ldots, M\}$$

dekodeeriv funktsioon. Pärest dekodeerimist saame sõna $\hat{W} = g(Y^n)$, mis paraku ei pruugi alati olla esialgne sõna W.

Def 4.2 *Olgu* $\{P(y|x)\}_{x\in\mathcal{X},y\in\mathcal{Y}}$ *diskreetne mäluta kanal. Selle kanali* (M,n) **kood** *koosneb järgmistest komponentidest:*

- *hulk* $\{1,\dots,M\}$ *(sõnade indeksid);*
- *kodeeriv funktsioon*
$$\mathcal{C}:\{1,\dots,M\}\to\mathcal{X}^n;$$
Koodisõnad
$$\{\mathcal{C}(1),\dots,\mathcal{C}(M)\}$$
moodustavad **koodiraamatu**.
- *dekodeeriv funktsioon*
$$g:\mathcal{Y}^n\to\{1,2,\dots,M\}.$$

Veatõenäosused. Olgu λ_i (tinglik) tõenäosus, et (M,n) kood teeb sõna i edastamisel vea. Seega

$$\lambda_i=\mathbf{P}(\hat{W}\neq i|W=i)=\mathbf{P}\big(g(Y^n)\neq i|W=i\big)=\sum_{y^n:\,g(y^n)\neq i}P(y^n|\mathcal{C}(i)).$$

Olgu

$$\lambda_{max}:=\max_i\lambda_i$$

ning olgu P_e vea tegemise tõenäosus juhul, kui sõna valitakse ühtlaselt üle kõikide sõnade hulga $\{1,\dots,M\}$. Seega

$$P_e=\mathbf{P}(\hat{W}\neq W)=\sum_i\mathbf{P}(\hat{W}\neq i|W=i)\mathbf{P}(W=i)=\frac{1}{M}\sum_i\mathbf{P}(\hat{W}\neq i|W=i)=\frac{1}{M}\sum_i\lambda_i.$$

On selge, et

$$P_e\le\lambda_{max}.$$

Def 4.3 (M,n) **koodi määr (rate)** *on*

$$R:=\frac{\log M}{n}.$$

Formaalselt on koodi määr vaid koodi $\mathcal{C}$ omadus (tingimusel et $\mathcal{X}$ on fikseeritud) ja näitab mitu bitti informatsiooni $\mathcal{C}$ korral läbi kanali saadetakse. Praktikas otsime aga koodi $\mathcal{C}$ kanalist sõltuvalt – nii, et viga oleks maksimaalselt väike.

4.4 Näited koodidest binaarse sümmeetrilise kanali korral

4.4.1 Ühtlane kood ja kordamiskood

Ühtlane kood. Olgu $|\mathcal{X}| = 2$ ja $\mathcal{C}$ *ühtlane kood*, mis $M = 2^n$ korral hulga $\{1, \ldots, 2^n\}$ üks-ühesesse vastavusse hulgaga $\mathcal{X}^n$ (tuleta meelde kodeerimist nõrga AEP abil). Selle koodi määr on 1. On selge, et kui $|\mathcal{X}| = 2$, siis parema määraga koodi konstrueerida pole võimalik.
Kui $M = 16$, siis ühtlase koodi koodiraamat on

$$(0000),\ (1000),\ (0100),\ (0010),\ (0001),\ (1100),\ (1010),\ (1001),$$
$$(0110),\ (0101),\ (0011),\ (1110),\ (1101),\ (1011),\ (0111),\ (1111).$$

Kui kanal on vigadeta binaarne kanal, on vaadeldud kood igati mõistlik: tal on maksimaalne määr ja $\lambda_{max} = 0$.
Sama koodi võib ka kasutada binaarse sümmeetrilise kanali korral. Koodi määr on endiselt 1, kuid veatõenäosus kasvab koos n-ga (koos M-ga):

$$1 - \lambda_i = \mathbf{P}(\hat{W} = i | W = i) = \mathbf{P}(Y^n = \mathcal{C}(i)) = (1-p)^n.$$

Kuigi koodil on kõrge määr, pole see antud kanali korral mõistlik.

Kordamiskood. Binaarse sümmeetrilise kanali korral pakutakse tihti välja nn *kordamiskoodi* (*repetition code*): iga bitt ühtlases koodis esitatakse m kordselt. Kui m on piisavalt suur ja $p < 0.5$, siis suurte arvude seaduse tõttu suure tõenäosusega enamik neist jõuab kohale. Seega kordamiskoodi korral edastatakse ühtlase koodi bitid pikkusega m blokkide kaupa, vastuvõtja seab igale blokile vastavusse ühe biti vastavalt sellele, milliseid bitte on vastuvõetud blokis enamus (viikide vältimiseks olgu m paaritu arv). Teinekord tähistatakse sellist koodi R_m. Esialgne ühtlane kood on siis R_1. Näiteks kui $M = 16$, siis koodi R_3 koodiraamat on

$$(000\,000\,000\,000),\ (111\,000\,000\,000),\ (000\,111\,000\,000),\ (000\,000\,111\,000),$$
$$(000\,000\,000\,111),\ (111\,111\,000\,000),\ (111\,000\,111\,000),\ (111\,000\,000\,111),$$
$$(000\,111\,111\,000),\ (000\,111\,000\,111),\ (000\,000\,111\,111),\ (111\,111\,111\,000),$$
$$(111\,111\,000\,111),\ (111\,000\,111\,111),\ (000\,111\,111\,111),\ (111\,111\,111\,111).$$

Leiame veatõenäosuse λ_i. Selleks paneme tähele, et piisab kui leiame ühe ühtlase koodi biti (ühe m-bloki) valesti esitamise tõenäosuse p_b, sest otsitav veatõenäosus on siis

$$1 - \lambda_i = \mathbf{P}(Y^n = \mathcal{C}(i)) = (1-p_b)^4.$$

Üks m-blokk dekodeeritakse vigaselt, kui vähemalt kaks bitti edastatakse valesti. Seega R_3 korral

$$p_b = 3p^2(1-p) + p^3.$$

Kui $p = 0.1$, siis $p_b = 3 \cdot 0.01 \cdot 0.9 + (0.01)^3 = 0.028$ ja iga $i = 1, \ldots, M$ korral

$$\lambda_i = 1 - (1 - 0.028)^4 = 0.107... = \lambda_{max} = P_e.$$

Nägime, et R_3 kahandas ühe biti edastamise veatõenäosuse esialgselt 0.1-lt (R_1 korral) 0.028-ni (R_3 korral). Antud näite korral keskmine viga P_e on nii R_1 kui ka R_3 korral enam-vähem võrdne (0.1 ja 0.107...) kuid tuletame meelde, et R_1 korral saab selle veaga edastada vaid kaks sõna (ühe biti), kuid R_3 korral 16 sõna (neli bitti). Küll aga vähenes koodi määr. Antud juhul on tegemist $(16, 12)$ koodiga ja tema määr seega

$$R = \frac{\log M}{n} = \frac{4}{12} = \frac{1}{3}.$$

On selge, et kui $\log M$ on täisarv, siis koodi R_m määr on alati $\frac{1}{m}$ (miks?).

Suurte arvude seadusest järeldub, et kui $p < 0.5$, siis valides m piisavalt suure, saame tõenäosuse p_b teha kuitahes väikeseks (ülesanne 1). Teisisõnu, iga $\epsilon_0 > 0$ korral leidub $m_0(\epsilon_0)$ nii, et kui $m \geq m_0(\epsilon_0)$, siis koodi R_m korral $p_b \leq \epsilon_0$. Seega saab iga sõnade arvu M korral teha kuitahes väikeseks ka λ_{max}. Tõepoolest, et (olgu $\log M$ täisarv)

$$\lambda_{max} = 1 - (1 - p_b)^{\log M},$$

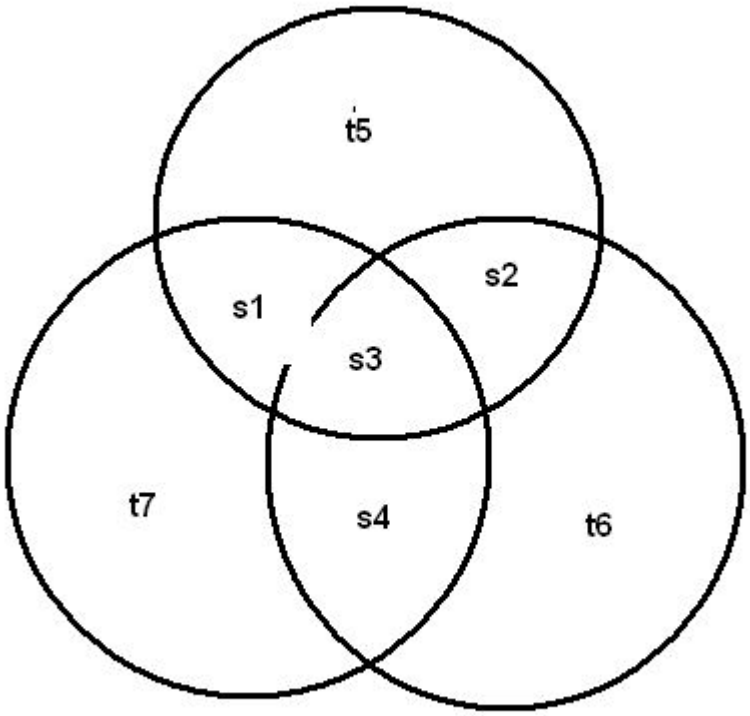
pole raske näha, et $\lambda_{max} \to 0$, kui $p_b \to 0$. Seega saab veatõenäosuse teha kuitahes väikeseks, kuid seejuures peab m olema väga suur ja koodi määr $\frac{1}{m}$ seega väga väike. Saab näidata, et kui $p = 0.1$, siis saavutamaks $p_b = 10^{-15}$, mis on teatav tehniline standard arvuti kettaseadmetel, peaks m olema ligikaudu 61. Sisuliselt tähendab see sõnastikust 61 koopia tegemist.

4.4.2 Hammingi kood

Hammingi kood kuulub binaarse sümmeetrilise kanali tarbeks loodud nn *paarsust kontrollivate (parity check)* koodide hulka. Sellised koodid põhinevad lihtsal asjaolul – kui ülekande käigus muutub ainult üks bitt, muudab see koodisõna ühtede paarsust. Viimast on aga lihtne kontrollida. Lihtne näide sellisest koodist on järgmine: olgu koodisõna pikkus paaritu arv. Liidame sellele ühe biti nii, et ühtede arv koodisõnas oleks paarisarv. Kui ülekande käigus ainult üks bitt (paaritu arv bitte) muutub, muutub ka koodisõnas olevate ühtede paarsus. Nii saab dekodeerija aru, et juhtunud on viga. Kahjuks ei oska ta aga seda viga parandada. Hammingi kood on selline, et ühe biti muutumist saab dekodeerimise käigus korrigeerida ning esialgse sõna seega restaureerida. Kui koodisõna pole liiga pikk ja veatõenäosus liiga suur, on kahe või enama biti muutumise tõenäosus väike võrreldes ühe biti muutumise tõenäosusega.

Hammingi $(7, 4)$ kood: idee. Tutvume Hammingi (16,7)-koodiga (kirjanduses nimetatakse seda $(7, 4)$ koodiks). Selle koodi määr on seega $\frac{4}{7}$ ning ta on mõeldud 16 sõna edastamiseks

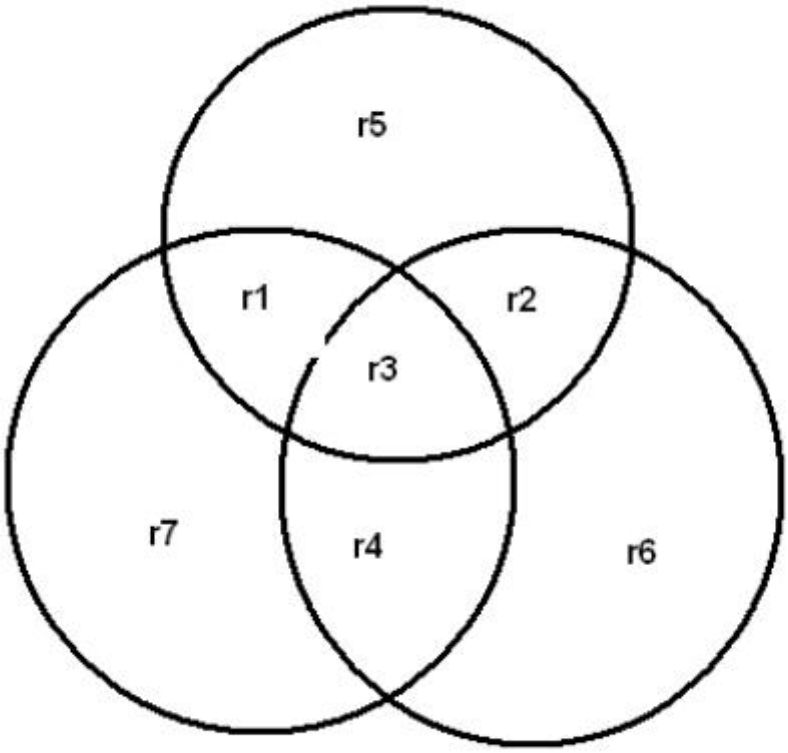
läbi binaarse sümmeetrilise kanali. Kood on järgmine: sõna $W \in \{1, \ldots, 16\}$ kahendesitusele s_1, s_2, s_3, s_4 liidetakse kolm (paarsus)bitti t_5, t_6, t_7 eeskirja alusel, mida on kõige lihtsam selgitada järgmise diagrammi põhjal.



Arvud t_5, t_6, t_7 valitakse nii, et igas ringis oleks ühtesi paarisarv. Nii saadakse järgmised 16 koodisõna (paksult on trükitud bitid $s_1 s_2 s_3 s_4$):

0000000	**0100**110	**1100**011	**1000**101
0001011	**0101**101	**1101**000	**1001**110
0010111	**0110**001	**1110**100	**1010**010
0011100	**0111**010	**1111**111	**1011**001

Dekodeerimine käib analoogiliselt: ülekandel saadud sõna $r_1, r_2, r_3, r_4, r_5, r_6, r_7$ bitid paigutatakse ringidesse samasse positsioonidesse, mis bitid $s_1, s_2, s_3, s_4, t_5, t_6, t_7$. Seega



Nüüd kontrollitakse kõikides ringides olevate ühtede paarsust. Seejuures on 8 võimalust: kas kõigis kolmes ringis on ühtesid paarisarv, ühes kolmest ringis pole see nii, kahes ringis pole see nii, kolmes ringis pole see nii. Kui kõikides ringides on ühtesi paarisarv, loetakse saadud sõna veatuks. Sellisele sõnale vastab üks koodisõna ning see koodisõna on $\hat{W}$. Ülejäänud juhtudel on vähemalt ühes ringis ühtesi paaritu arv. Ütleme, et need ringid on vigased. Hammingi kood on aga konstrueeritud nii, et ükskõik mitu vigast ringi korraga ka ei oleks, ikka saab vaid ühe biti muutmisega ringide paarsused korda seada. Selleks tuleb lihtsalt muuta seda bitti, mis asub kõikide vigaste ringide ühisosas. Näiteks kui vigased on kaks alumist ringi, tuleb muuta bitti r_4; kui vigased on kõik kolm ringi, tuleb muuta bitti r_3 jne. Pärast vea parandamist, on saadud sõna üks 16 koodisõnast ning see koodisõna on $\hat{W}$.
Kui koodisõna edastamisel ei muutunud ükski bitt, siis dekodeerimisel ühtki viga ei parandatud ning $\hat{W} = W$. Kui ülekandel muutus üks bitt, siis muutus mõne ringi paarsus ning antud meetod võimaldab seda viga parandada (muutunud bitt leitakse üles). Ka sellisel juhul $\hat{W} = W$. Kui ülekande käigus muutus kaks või enam bitti, siis sõltumata sellest kui palju ringe on vigased, parandatakse vahetatakse ülimalt üks bitt. Saadud sõna on alati koodisõna, mis aga erineb sisestatust ning $\hat{W} \neq W$. Seega

$$\lambda_{\max} = \lambda_i = 1 - ((1-p)^7 + 7p(1-p)^6).$$

When $p = 0.1$, then $\lambda_{\max} \approx 0.15$. Proovi dekodeerida sõnad

$$1101011, \quad 0110110, \quad 0100111, \quad 1111111.$$

Hammingi kood kui lineaarne kood. Hammingi kood on *lineaarne*: st iga kahe koodisõna summa 2-jäägiklassiringis (st $1+1=0, 0+1=1, 1+0=1, 0+0=0$) on omakorda koodisõna. Nimelt iga koodisõna

$$c^T = (s_1, s_2, s_3, s_4, t_5, t_6, t_7)$$

paarsusvektor $t^T = (t_5, t_6, t_7)$ avaldub korrutisena (jäägiklassiringis)

$$t = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} s_1 \\ s_2 \\ s_3 \\ s_4 \end{pmatrix} \tag{4.2}$$

Olgu seoses (4.2) olev maatriks P. Siis (4.2) on $t = Ps$, kus $s^T = (s_1, s_2, s_3, s_4)$. Defineerime 7×4 maatriksi

$$G := \begin{pmatrix} I_4 \\ P \end{pmatrix},$$

kus I_4 on 4×4 ühikmaatriks. Siis iga koodisõna c avaldub

$$c = \begin{pmatrix} s \\ t \end{pmatrix} = Gs = \begin{pmatrix} I_4 \\ P \end{pmatrix} s. \tag{4.3}$$

Seosest (4.3) järeldub nüüd koodi lineaarsus. Defineerime nüüd 3×7 maatriksi H järgmiselt

$$H = \begin{pmatrix} P & I_3 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 \end{pmatrix}. \tag{4.4}$$

Jäägiklassiringis on $-P = P$ (sest $1+1=0$), mistõttu

$$HG = \begin{pmatrix} -P & I_3 \end{pmatrix} \begin{pmatrix} I_4 \\ P \end{pmatrix} = (-P + P) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Seega korrutades maatriksit H koodisõnaga c (ikka jäägiklassiringis), saame

$$Hc = HGs = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}. \tag{4.5}$$

Maatriksi H veerud on kõik hulga $\{0,1\}^3$ elemendid välja arvatud 0 vektor. Kõik vektorid on erinevad, mistõttu suvalise kahe vektori summa ei saa olla 0. Kui koodivektoris pole ainult nullid, peab tas olema vähemalt 3 ühte sest ühe või kahe ühega ei saaks kehtida (4.5). Samas koodi lineaarsuse tõttu on iga kahe koodivektori vahe koodivektor. Seega erinevad kaks koodivektorit vähemalt kolme biti võrra. Teisisõnu, kahe koodisõna omavaheline kaugus Hammingi mõttes on vähemalt 3. Kui nüüd ühe koodisõna c üks bitt muutub, siis

erineb muudetud vektor, olgu see r, sõnast c täpselt ühe biti võrra (kaugus on 1), kuid kõikidest teistest koodisõnadest vähemalt 2 biti võrra (kaugus vähemalt 2). Seega on c vektor mis minimiseerib üheselt Hammingi kauguse r ja teiste koodisõnade vahel, st

$$c = \arg \min_{i=1,\ldots,16} h(r, c_i), \tag{4.6}$$

kus h on Hammingi kaugus ja $c_1, \ldots, c_{16}$ kõikvõimalikus koodisõnad. Loomulikult pole c leidmiseks vaja leida $h(r, c_i)$ kõikide koodisõnade korral. Seda nägime juba ülalpool (ringide abil) ning selles on kerge veenduda ka maatriksite abil.
Tõepoolest, kui ülekande käigus muutub täpselt üks bitt, siis vastuvõtjani jõuab vektor $r = c + e_i$, kus e_i koosneb nullides välja arvatud i-s positsioon, kus on 1 ($i = 1, \ldots, 7$). korrutades vektorit r maatriksiga H, saame

$$Hr = H(c + e_i) = He_i.$$

Ent He_i on maatriksi i-s veerg. Maatriksi H veerud on üksteisest erinevad. Seega teades veergu He_i, teame positsiooni i ning seega on viga võimalik parandada.

Hammingu koodi suuremate sõnastike jaoks. Nüüd on kerge üldistada kirjeldatud meetodi suurema sõnastiku kodeerimiseks. Oletame, et tahame kodeerida 2^5 koodisõna. Siis on paarsusbitte tarvis vähemalt 4 (miks?). Seega konstrueerime 4×5 maatriksi P, mille veerud on kõik erinevad ja sisaldavad vähemalt 2 ühte. Näiteks

$$P = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Iga algse koodisõna $s^T = (s_1, \ldots, s_5) \in \{0,1\}^5$ korral vektor $t = Ps$ määrab paarsuslaiendi. Näiteks kui $s = (1,0,0,1,1)$, on paarsusbitid $1,0,1,1$ ja nii on koodisõna **10011**1011. Maatriks H on nüüd (P, I_4):

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Maatriksi H veerud on kõik erinevad ja H read on ortogonaalsed kõikide koodisõnadega. Kõik vektorid on erinevad, mistõttu suvalise kahe vektori summa ei saa olla 0. Peale nullvektori igas koodivektoris peab olema vähemalt 3 ühte sest ühe või kahe ühega ei saaks kehtida (4.5). Seega on kõikide koodisõnade Hammingi kaugus vähemalt 3.
Äsja konstrueeritud koodi on (32,9)-kood, tema määr on $\frac{5}{9}$, Tegelikult saab 4 paarsusbiti abil laiendada rohkem sõnu kui 2^5. Tõepoolest, et maatriksi ridu on 4, saab maksimaalne veergude arv maatriksis H olla $2^4 - 1 = 15$. See saab olla maksimaalne (laiendatud) koodisõna pikkus. Originaalkoodisõna pikkus saab olla maksimaalselt $15 - 4 = 11$. Seega

saab nelja paarsusbitiga laiendada maksimaalselt 2^{11} koodisõna. Selle koodi määr on $\frac{11}{15}$. Analoogiliselt saame, et k paarsusbitiga saab laiendada

$$2^{2^k-1-k}$$

sõna, koodi määr on siis $\frac{2^k-1-k}{2^k-1}$. Määr läheneb ühele, kui k kasvab, kuid seejuures kasvab ka veatõenäosus, sest pikkade koodisõnade puhul on suurem tõenäosus, et muutub rohkem kui 1 bitt.

4.5 Kanaliteoreem

Def 4.4 *Olgu $P(y|x)$ diskreetne mäluta kanal. Arv R on kanali* **saavutatav määr (achievable rate)** *, kui leidub selle kanali $(\lceil 2^{nR} \rceil, n)$ koodide jada nii, et nende maksimaalne viga λ_{max} läheneb nullile.*

Kas arv R on saavutatav määr või mitte, on kanali omadus. Kui R on kanali saavutatav määr, siis leidub selline kanali $(\lceil 2^{nR} \rceil, n)$ koodide jada, et maksimaalne viga läheneb nullile. Kui maksimaalne viga läheneb nullile, siis suvalise W jaotuse korral läheneb nullile ka viga $\mathbf{P}(\hat{W} \neq W)$. Seega, kui $R > 0$ on kanali saavutatav määr, siis kuitahes suure sõnade arvu M ja kuitahes väikese $\epsilon > 0$ korral leidub alati mingi n ja mingi $(\lceil 2^{nR} \rceil, n)$ kood nii, et selle koodi maksimaalne viga on väiksem kui ϵ. Seega selle koodi korral võib juhuslikult valitud sõna hulgast $\{1, \ldots, \lceil 2^{nR} \rceil\}$ läbi kanali edastada nii, et vea tõenäosus on väiksem kui ϵ.
Binaarse vigadeta kanali korral on 1 koodi saavutatav määr.

Edaspidi tähistame $\lceil 2^{nR} \rceil$ lihtsalt 2^{nR}.

Järgnev teoreem, nn *Shannoni teine teoreem* on informatsiooniteooria keskne tulemus.

Teoreem 4.5 (Kanaliteoreem) *Olgu C kanali võimsus. Siis iga arv $R < C$ on selle kanali saavutatav määr. Teisisõnu, iga sellise arvu R korral leidub $(2^{nR}, n)$ koodid nii, et $\lambda_{max} \to 0$.*
Teistpidi, kui leidub $(2^{nR}, n)$ kood nii, et $\lambda_{max} \to 0$, siis $R \leq C$.

4.5.1 Esimese väite tõestus

Olgu $R < C$. Näitame, et R on saavutatav määr.

Esimese sammuna fikseerime suvalise $\frac{C-R}{3} > \epsilon > 0$ ning näitame, et leidub kood $\mathcal{C}^*$ nii, et $P_e(\mathcal{C}^*) \leq 2\epsilon$, kus $P_e(\mathcal{C}^*)$ on edastamisel tehtud viga juhul, kui W on ühtlase jaotusega ning kood on $\mathcal{C}^*$. Selleks toimime järgmiselt:

1) Fikseerime sisendjaotuse $P(x)$, mille korral $I(X;Y) = C$. See jaotus, nagu ka kanal $\{P(y|x)\}$ on teada nii vastuvõtjale kui ka sisendajale.

2) Jaotuse $P(x)$ abil genereerime 2^{nR} juhuslikku sõna $x^n(1), \ldots, x^n(2^{nR})$. Saadud 2^{nR} sõna vaatleme hulga

$$\{1, \ldots, 2^{nR}\}$$

koodina:

$$\mathcal{C} : \{1, \ldots, 2^{nR}\} \to \mathcal{X}^n, \quad \mathcal{C}(i) = x^n(i).$$

Olgu

$$X^n(1), \ldots, X^n(2^{nR})$$

sõltumatud juhuslikud sama jaotusega juhuslikud vektorid, kusjuures iga vektor

$$X^n(i) = (X_1(i), \ldots, X_n(i))$$

omakorda koosneb samuti iid komponentidest. Vektor $X^n(i)$ modelleerib koodisõna $x^n(i)$. Seega

$$\mathbf{P}(X^n(i) = x^n(i)) = \prod_{j=1}^{n} P(x_j(i)),$$

kus $x^n(i) = x_1(i), \ldots, x_n(i)$.

Juhuslik iid komponentidega maatriks

$$X := \begin{pmatrix} X_1^n(1) & X_2(1) & \cdots & X_n(1) \\ \cdots & \cdots & \cdots & \cdots \\ X_1(2^{nR}) & X_2(2^{nR}) & \cdots & X_n(2^{nR}) \end{pmatrix}$$

modelleerib juhjuslikku koodi. Iga maatriksi rida on üks koodisõna, tõenäosus koodi $\mathcal{C}$ saamiseks on

$$\mathbf{P}(X = \mathcal{C}) = P(\mathcal{C}) = \prod_{j=1}^{2^{nR}} \prod_{i=1}^{n} P(x_i(j)).$$

3) Saadud kood edastatakse informatsiooni saatjale ning vastuvõtjale.

4) Sõnastikust $\{1, \ldots, 2^{nR}\}$ valime ühtlase jaotusega sõna w. Olgu W juhuslik sõna, s.t.

$$\mathbf{P}(W = w) = 2^{-nR}.$$

5) Valitud sõna w kodeeritakse selle koodi abil ja saadud koodisõna $x^n(w)$ saadetakse läbi kanali.

6) Vastuvõtja saab signaali y^n vastavalt jaotusele

$$P\big(y^n | x^n(w)\big) = \prod_{i}^{n} P\big(y_i | x_i(w)\big).$$

7) Vastuvõtja dekodeerib saadud sõna y^n vastavalt järgmisele eeskirjale

$$g(y^n) = \begin{cases} k & \text{kui } (x^n(k), y^n) \in W_\epsilon^n \text{ ning iga } i \neq k \text{ korral } (x^n(i), y^n) \notin W_\epsilon^n, \\ * & \text{muidu.} \end{cases}$$

Siin $* \notin \mathcal{Y}$, mistõttu see väljund on kindlasti viga. Püüame hinnata ülalkirjeldatud juhuslikul kodeerimisel saadud viga. Selleks hindame keskmist viga üle kõigi juhuslike koodide

$$\sum_{\mathcal{C}} P(\mathcal{C}) P_e(\mathcal{C}) = \sum_{\mathcal{C}} P(\mathcal{C}) \frac{1}{2^{nR}} \sum_{j}^{2^{nR}} \lambda_j(\mathcal{C}) = \frac{1}{2^{nR}} \sum_{j}^{2^{nR}} \sum_{\mathcal{C}} P(\mathcal{C}) \lambda_j(\mathcal{C}),$$

kus

$$\lambda_j(\mathcal{C}) := \mathbf{P}(\hat{W} \neq W | W = j, \mathcal{C})$$

on sõna j edastamisel tehtud viga koodi $\mathcal{C}$ korral. Summa

$$\sum_{\mathcal{C}} P(\mathcal{C}) \lambda_j(\mathcal{C})$$

on tähe j dekodeerimisel tehtud keskmine viga (üle kõikide koodide). Olgu $\mathcal{C}_1$ ja $\mathcal{C}_j$ koodid, kus esimene ja j-s rida on ära vahetatud, muidu samad. On selge, et $P(\mathcal{C}_1) = P(\mathcal{C}_j)$. Sellest järeldub, et

$$\sum_{\mathcal{C}} P(\mathcal{C}) \lambda_j(\mathcal{C}) = \sum_{\mathcal{C}} P(\mathcal{C}) \lambda_1(\mathcal{C})$$

ehk

$$\begin{aligned} \sum_{\mathcal{C}} P(\mathcal{C}) P_e(\mathcal{C}) &= \frac{1}{2^{nR}} \sum_{j}^{2^{nR}} \sum_{\mathcal{C}} P(\mathcal{C}) \lambda_j(\mathcal{C}) = \sum_{\mathcal{C}} P(\mathcal{C}) \lambda_1(\mathcal{C}) \\ &= \sum_{\mathcal{C}} P(\mathcal{C}) \mathbf{P}(\hat{W} \neq W | W = 1, \mathcal{C}) \\ &= \sum_{\mathcal{C}} P(\mathcal{C} | W = 1) \mathbf{P}(\hat{W} \neq W | W = 1, \mathcal{C}) \\ &= \mathbf{P}(\hat{W} \neq W | W = 1), \end{aligned}$$

kus kolmas võrdus järeldub sellest, et sõna- ja koodivalik on sõltumatud, $P(\mathcal{C}|W = 1) = P(\mathcal{C})$. Tuletame meelde, et $\mathbf{P}(\hat{W} \neq W | W = 1, \mathcal{C})$ on esimese sõna edastamisel tehtud viga koodi $\mathcal{C}$ korral, $\mathbf{P}(\hat{W} \neq W | W = 1)$ on aga kogu kirjeldatud juhusliku kodeerimise kaudu esimese sõna edastamisel tehtud viga.

Juhuslik vektor $X^n(i)$ on juhusliku koodi i-s sõna, $Y^n(i)$ olgu selle väljund läbi kanali. Defineerime sündmuse

$$E_i = \{(X^n(i), Y^n(1)) \in W_\epsilon^n\}.$$

Esimese sõna kodeerimine on vigane siis, kui toimub sündmus E_1^c või üks sündmustest $E_2, \ldots, E_{2^{nR}}$. Seega

$$\mathbf{P}(\hat{W} \neq W | W = 1) \leq \mathbf{P}(E_1^c \cup E_2 \cup \cdots \cup E_{2^{nR}}) \leq \mathbf{P}(E_1^c) + \sum_{i=2}^{2^{nR}} \mathbf{P}(E_i).$$

Teoreemi 3.4 esimesest väitest järeldub, et piisavalt suure n korral

$$\mathbf{P}(E_1^c) \leq \epsilon.$$

Tuletame meelde, et $X^n(i)$ on iid vektor jaotusest P. See jaotus oli aga selline, et

$$I(X_1(i); Y_1(i)) = C.$$

Vektorid $X^n(i)$ ja $X^n(1)$ on sõltumatud, mistõttu on sõltumatud ka $X^n(i)$ ja $Y^n(1)$. Teoreemi 3.4 viimasest väitest saame, et piisavalt suure n korral

$$\mathbf{P}(E_i) = \mathbf{P}\big((X^n(i), Y^n(1)) \in W_\epsilon^n\big) \leq 2^{-n(I(X_1(i);Y_1(i))-3\epsilon)} = 2^{-n(C-3\epsilon)}, \quad j = 2, \ldots, 2^{nR}.$$

Kokkuvõttes,

$$\mathbf{P}(\hat{W} \neq W | W = 1) \leq \epsilon + \sum_{i=1}^{2^{nR}} 2^{-n(C-3\epsilon)} = \epsilon + 2^{-n(C-R-3\epsilon)} \leq 2\epsilon,$$

kui n on piisavalt suur ja ϵ on nii väike, et $C - R - 3\epsilon > 0$, s.t. $R + 3\epsilon < C$.

Tõestasime, et kuitahes väikese ϵ korral leidub piisavalt suur n nii, et

$$\sum_{\mathcal{C}} P(\mathcal{C}) P_e(\mathcal{C}) \leq 2\epsilon.$$

Et keskmine on väiksem kui 2ϵ, siis peab leidume vähemalt üks kood $\mathcal{C}^*$ nii, et

$$P_e(\mathcal{C}^*) \leq 2\epsilon.$$

Edaspidi võib kasutada seda (mittejuhuslikku) koodi.

Tuletame meelde, et P_e on keskmine viga (üle ühtlase jaotusega sõnavaliku). Seega oleme tõestanud, et koodi $\mathcal{C}^*$ korral on

$$\frac{1}{2^{nR}} \sum_{i}^{2^{nR}} \lambda_i \leq 2\epsilon.$$

Ülaltoodud võrratusest järeldub, et leidub vähemalt 2^{nR-1} indeksit i nii, et $\lambda_i \leq 4\epsilon$. Tõepoolest, kui see nii, ei ole, s.t. leidub vähemalt $2^{nR-1} + 1$ λ_i-d mis on suuremad kui 4ϵ, siis oleks $\sum_i^{2^{nR}} \lambda_i > 2\epsilon$. Jätame koodist $\mathcal{C}^*$ alles pooled koodisõnad, need mille korral $\lambda_i \leq 4\epsilon$. Sellise pooliku koodiga saame kodeerida

$$2^{nR-1} = 2^{n(R-\frac{1}{n})}$$

sõna. See tähendab, et meil on $(2^{n((R-\frac{1}{n})}, n)$ kood nii, et $\lambda_{max} \leq 4\epsilon$. Vahe R ja $R - \frac{1}{n}$ vahel läheneb n kasvamisel nullile. Seega on iga $R < C$ saavutavav määr. ■

Märkused:

- Teoreemi tõestus põhineb sisuliselt järgneval: juhuslikult valitud koodisõna x^n on suure tõenäosusega nõrgalt tüüpiline. Sellise sõna kanali kaudu edastamisel on väljund y^n suure tõenäosusega üks neist $2^{nH(Y|X)}$ vektorist mis on sisendiga koos ühistüüpilised. Ülalkirjeldatud infovahetus töötab hästi, kui erinevatele sisenditele vastavad ühistüüpilised vektorite hulgad on sisuliselt kattumatud. See aga seabki piiri sisendite arvule. Tõepoolest, kui kõikide nõrgalt tüüpiliste väljundite hulk on jagatud lõikumatuteks klassideks, millistes igaühes on umbes $2^{nH(Y|X)}$ elementi ning kui kõiki nõrgalt tüüpilisi väjundeid on umbes $2^{nH(Y)}$, siis peab nende klasside arv olema ligikaudu
$$\frac{2^{nH(Y)}}{2^{nH(Y|X)}} = 2^{nI(X;Y)}.$$
Igale klassile vastab üks sisend. Kokku peab olema ligikaudu $2^{nI(X;Y)}$ sisendit.

- Ülaltoodud tõestus on olemasolu, mitte konstruktsioonitõestus. Tõestus ei anna eeskirja parima koodi $\mathcal{C}^*$ konstrueerimiseks. Põhimõtteliselt võiks küll leida iga võimaliku koodi korral tema maksimaalse vea ning otsida parimat koodi kõikide võimalike koodide seast. Et aga $(2^{nR}, n)$ koodi konstrueerimiseks tuleb läbi vaadata $2^{n2^{nR}}$ võimalikku koodi, langeb see variant ära.
Muidugi võib koodi konstrueerida ka juhuslikult, nii nagu ülaltoodud tõestuses. Suure tõenäosusega (ja suure n korral) see kood töötab hästi. Sellise juhuslikult genereeritud koodi korral on probleem dekodeerimine. Teadmata tema struktuuri paistab ainus võimalus dekodeerimiseks $n \times 2^{Rn}$ tabelist õige vaste otsimine ning see on ebapraktiline.

Töö praktiliselt rakendatava kõrge määraga $(2^{nR}, n)$ koodi leidmiseks on algas sisuliselt juba pärast Shannoni esimese artikli ilmumiset ning kestab siiamaani. Pikka aega ei suudetud selliseid koode leida või nende efektiivsust tõestada. 1993 aastal pakuti välja nn. *turbokood*, mis peaaegu saavutab kanali võimsuse. Samuti saavutavad kanali võimsuse nn *Low Density Parity Check* koodid.

4.5.2 Teise väite tõestus

Lemma 4.1 *Olgu $X^n = \mathcal{C}(W)$ juhuslik koodisõna, $Y^n = (Y_1, \ldots, Y_n)$ selle väljund. Siis*
$$I(X^n; Y^n) \leq nC.$$

Tõestus. Entroopia tinglikust ketireeglist järeldub, et
$$H(Y^n|X^n) = H(Y_1|X^n) + H(Y_2|Y_1, X^n) + \cdots + H(Y_n|Y_1, \ldots, Y_{n-1}, X^n).$$
Vastavalt definitsioonile
$$H(Y_i|Y_1, \ldots, Y_{i-1}, X^n) = -\sum_{y_i, y^{i-1}, x^n} \log P(y_i|y_1, \ldots, y_{i-1}, x_1, \ldots, x_n) P(y_1, \ldots, y_i, x_1, \ldots, x_n).$$

Kanal on mäluta, s.t. iga i korral

$$P(y_i|y_1,\ldots,y_{i-1},x_1,\ldots,x_n) = P(y_i|x_i)$$

ja

$$P(y_1,\ldots,y_i,x_1,\ldots,x_n) = P(y_i|x_i)P(y_1,\ldots,y_{i-1},x_1,\ldots,x_n),$$

millest

$$H(Y_i|Y_1,\ldots,Y_{i-1},X^n) = H(Y_i|X_i).$$

Järelikult

$$H(Y^n|X^n) = \sum_{i=1}^{n} H(Y_i|X_i), \tag{4.7}$$

millest

$$\begin{aligned} I(X^n;Y^n) &= H(Y^n) - H(Y^n|X^n) = H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) \\ &\leq \sum_{i=1}^{n}(H(Y_i) - H(Y_i|X_i)) = \sum_{i=1}^{n} I(X_i;Y_i) \leq nC \end{aligned}$$

■

Veata koodid. Kanaliteoreemi teine väide on sisuliselt järgmine: kui leidub $(2^{nR}, n)$ kood, mille maksimaalne viga on väike, siis $R \leq C$. Tõestuse idee selgitamiseks tõestame esialgu nõrgema väite.

Väide 4.1 *Kui leidub $(2^{nC}, n)$ kood, mille maksimaalne viga on 0, siis $R \leq C$.*

Tõestus. Oletame sellise $(2^{nR}, n)$ koodi olemasolu. Seega leidub dekodeeriv funktsioon g nii, et $g(Y^n) = W$ p.k.. Teisisõnu, $H(W|Y^n) = 0$. Kui juhuslik sõna W on ühtlase jaotusega, siis $H(W) = nR$. Tuletame meelde, et $X^n = \mathcal{C}(W)$ on juhuslik koodisõna. Et

$$W \to X^n \to Y^n$$

on Markovi ahel, siis andmetöötlusvõrratusest järeldub

$$I(W;Y^n) \leq I(X^n;Y^n). \tag{4.8}$$

Arvestades, et

$$I(W;Y^n) = H(W) - H(W|Y^n) = H(W), \tag{4.9}$$

saame lemmast 4.1 ja andmetöötlusvõrratusest (4.8)

$$nR = H(W) = I(W;Y^n) \leq I(X^n;Y^n) \leq \sum_{i=1}^{n} I(X_i;Y_i) \leq nC.$$

■

Milline peab olema $\mathcal{C}$, et poleks vigu ? Kui $\lambda_{max} = 0$, siis $W = g(Y^n)$ nii, et

$$W \to X^n \to Y^n \to W.$$

Nüüd on kerge näha, et

$$I(W;Y^n) = I(W;X^n) = I(X^n;Y^n) = I(W;W) = H(W) = H(X^n) = H(Y^n).$$

Tõepoolest, esimese võrduse saame kui rakendame andmetöötlusvõrratust ahelatele $W \to X^n \to Y^n$ ja $X^n \to Y^n \to W$. Sama võrratust rakendades ahelale $Y^n \to W \to X^n$ koos võrratuse (4.8) ja esimese võrdusega annab teise võrduse. Neljas võrdus on ilmne ja kolmas järeldub neljandast võrratuse (4.9) tõttu. Et $X^n = \mathcal{C}(W)$, siis $H(X^n|W) = 0$, millest $H(W) = I(X^n;W) = H(X^n) - H(X^n|W) = H(X^n)$ ja nii saame viienda võrduse. Viimase võrduse tõestus on analoogiline, sest $W = g(Y^n) \Rightarrow H(W|Y^n) = 0$ ja $H(W) = I(Y^n;W) = H(Y^n) - H(W|Y^n)$.

Võrdusest $I(X^n;W) = H(W) - H(W|X^n) = H(W)$ järeldub, et

$$H(W|X^n) = H(W|\mathcal{C}(W)) = 0$$

st kood $\mathcal{C}$ on ühene.

Oletame nüüd, et koodi $\mathcal{C}$ määr on kanali võimsus C ja $\lambda_{max} = 0$. Siis Väite 4.1 tõestuses olevad võrratused peavad olema võrdused. Neist esimene on

$$I(W;Y^n) = I(X^n;Y^n),$$

mis tuleneb sellest, et $\lambda_{max} = 0$ ja toob enesega kaasa $\mathcal{C}$ ühesuse. Teine võrratus võrdus siis, kui $H(Y^n) = \sum_{i=1}^n H(Y_i)$, mis tähendab, et juhuslikud suurused Y_i on sõltumatud. Kolmas võrdus

$$\sum_{i=1}^{n} I(X_i;Y_i) = nC$$

kehtib siis, kui iga i korral $I(X_i;Y_i) = C$ ehk X_i jaotus on selline, mis saavutab kanali võimsuse.

Kokkuvõtteks: Seega $(2^{nR}, n)$ kood, mille korral $P_e = 0$ ja $R = C$ peab rahuldama tingimusi:

- $\mathcal{C}$ on (üks)ühene;
- ühtlase jaotusega W korral juhuslikud suurused X_i on kanali võimsust saavutava jaotusega $P^*(x)$;
- ühtlase jaotusega W korral juhuslikud suurused Y_i^n on iid juhuslikud suurused jaotusega

$$P(y) = \sum_x P(y|x)P^*(x). \tag{4.10}$$

Siit järeldub, et (peaaegu) samasugused omadused peavad olema $(2^{nR}, n)$ koodil, mille maksimaalne viga on väike.

Näited:

- Vigadega klaviatuur. Sellisel juhul on lihtne saavutada kanali võimsust. Tõepoolest, olgu $M = 13^n$ ja olgu $\mathcal{C}$ ühtlane kood koodiraamatuga $\{1, 3, 5, \ldots, 25\}^n$. Selle koodi määr on $R = (\log M)/n = \log 13 = C$, mis ühtlasi on kanali võimsus. On selge, et sellise koodi korral $\lambda_{max} = 0$. Kas ülaltoodud tingimused on täidetud?
  Kui W on ühtlase jotusega, siis juhuslik koodisõna $X^n = X_1, \ldots, X_n$ on ühtlase jaotusega hulgal $\{1, 3, 5, \ldots, 25\}^n$ ja on lihtne veenduda (aga veenduge!), et siis $X_1, \ldots, X_n$ iid juhuslikud suurused ning X_i jaotus on ühtlane üle paaritute täht-ede $\{1, 3, 5, \ldots, 25\}$. See jaotus (ühtlane üle $\{1, 3, 5, \ldots, 25\}$) on ka kanali võimsust saavutav jaotus P^*. Sellise sisendjaotuse korral on väljund ühtlane üle kõikide täht-ede ning $Y_1, \ldots, Y_n$ on iid juhuslikud suurused jaotusega (4.10).

- Binaarne kadumiskanal. Selle kanali korral ei saa viga P_e olla 0. Samas peaks efektiivne kood ikkagi olema selline, et vektori $Y_1, \ldots, Y_n$ jaotus on lähedane Bernoulli $\frac{1}{2}$ iid jaotusele. Kordamiskoodi korral pole see kindlasti nii.

Fano võrratus taaskord. Väite 4.1 üldistus juhule, kui väikesed vead on lubatud põhineb Fano võrratusel. Esitame Fano võrratuse meile sobival kujul.

Lemma 4.2 (Fano võrratus) *Olgu W juhuslik täht. Siis*

$$H(W|Y^n) \leq 1 + \mathbf{P}(W \neq \hat{W})nR. \tag{4.11}$$

Tõestus. Tuletame meelde Fano võrratuse:

$$H(W|\hat{W}) \leq h(\mathbf{P}(W \neq \hat{W})) + \mathbf{P}(W \neq \hat{W}) \log(2^{nR} - 1) \leq 1 + \mathbf{P}(W \neq \hat{W})nR.$$

Et $\hat{W} = g(Y^n)$, siis (andmetöötlusvõrratus: $I(W; Y^n) \geq I(W, \hat{W})$)

$$H(W|\hat{W}) = H(W|g(Y^n)) \geq H(W|Y^n).$$

■

Teise väite tõestus. Olgu $(2^{nR}, n)$ koodide jada nii, et $\lambda_{max} \to 0$. Näiteme, et $R \leq C$. Et $\lambda_{max} \to 0$, siis

$$P_e = \frac{1}{2^{nR}} \sum_{i=1}^{2^n R} \lambda_i \to 0.$$

Seega piisab, kui näitame, et seosest $P_e \to 0$ järeldub, et $R \leq C$. Arv P_e on tõenäosus $\mathbf{P}(\hat{W} \neq W)$ juhul kui W on ühtlase jaotusega üle tähestiku. Seega tõestuseks piisab, kui

vaatame sellise jaotusega W ning veendume, et $\mathbf{P}(\hat{W} \neq W) = P_e \to 0$ viib seoseni $R \leq C$. Tõestus on põhimõtteliselt sama, mis väitel 4.1, kus näitasime, et

$$nR = H(W) = H(W) - H(W|Y^n) + H(W|Y^n) = I(W;Y^n) + H(W|Y^n) = I(W;Y^n),$$

sest veatu dekodeerimise korral $H(W|Y^n) = 0$. Praegusel juhul $H(W|Y^n) \neq 0$, kuid Fano võrratuse abil saame $H(W|Y^n)$ ülalt hinnata suurusega $1 + P_e nR$. Muu on kõik samamoodi:

$$\begin{aligned} nR = H(W) &= H(W|Y^n) + I(W;Y^n) \leq 1 + P_e nR + I(W;Y^n) \\ &\leq 1 + P_e nR + I(X^n;Y^n) \leq 1 + P_e nR + nC. \end{aligned}$$

Tuletame meelde et kaks viimast võrratust järelduvad andmetöötlusvõrratusest (4.8) ja lemmast 4.1. Seega

$$R \leq P_e R + \frac{1}{n} + C. \tag{4.12}$$

Et n kasvades $P_e R + \frac{1}{n} \to 0$, siis $R \leq C$.

Märkused:

1. Võrratus (4.12): iga n korral

$$P_e \geq 1 - \frac{C}{R} - \frac{1}{nR} \quad \Rightarrow \quad \lim_n P_e \geq \lim_n \left(1 - \frac{C}{R} - \frac{1}{nR}\right) = 1 - \frac{C}{R}.$$

   Seega, kui $C < R$, siis $\frac{C}{R} < 1$, millest järeldub, et leidub $\delta' > 0$ nii, et $P_e > \delta'$, kui n on piisavalt suur. Sellisel juhul ei saa P_e olla 0 ka väiksese n korral (sest kui mingi n korral on $P_e = 0$, siis on see nii ka $2n$ korral ja $3n$ korral jne). Järelikult, kui $C < R$, siis leidub $\delta > 0$ nii, et $P_e > \delta$ iga n korral.

2. Tõestatud väidet nimetatakse teinekord ka nõrgaks väiteks. Saab näidata, et kehtib ka tugev versioon: kui leidub $\epsilon > 0$ nii, et $R \geq C + \epsilon$, siis $P_e \to 1$.

4.6 Tagasisidega infovahetus

Tagasisidega (feedback) infovahetus on järgmine: pärast koodisõna x^n i-nda biti edastamist läbi kanali, saadab vastuvõtja saadud signaali y_i muutusteta saatjale tagasi. Saatja arvestab saadud informatsiooni järgmise biti saatmisel. Seega on sellise kanali korral koodi $\mathcal{C}$ asemel jada $\mathcal{C}_i$, kusjuures $\mathcal{C}_i$ argumendid on täht W ning siiani saadetud bittide tulemused $y_1, \ldots, y_{i-1}$. Nii saadakse väljund y^n, mis dekodeeritakse funktsiooni g abil.

Def 4.6 *Olgu* $\{P(y|x)\}_{x\in\mathcal{X},y\in\mathcal{Y}}$ *diskreetne kanal. Selle kanali* **tagasisidega** (M,n) **kood** *koosneb järgmistest komponentidest:*

- *hulk* $\{1,\dots,M\}$*;*
- *kodeerivad funktsioonid*

$$\mathcal{C}_i : \{1,\dots,M\}\times\mathcal{Y}^{i-1}\to\mathcal{X};$$

- *dekodeeriv funktsioon*

$$g:\mathcal{Y}^n\to\{1,2,\dots,M\}.$$

Tagasisidega infovahetuse kasulikkus tuleb hästi esile näiteks binaarse kadumiskanali korral: sümboli e saamisel edastab saatja eelnevalt saadetud sümboli veelkord kuni see lõpuks kohale jõuab.

Tagasisideta infovahetus on tagasisidega infovahetuse erijuht. Seega iga tagasisideta infovahetuse korral saavutatav määr on saavutatav ka tagasisidega infovahetuse korral. Võiks arvata, et tagasiside korral saab ehk saavutada kõrgemat määra kui C. Üllataval kombel pole see nii: ka tagasisidega infovahetuse korral ei saa saavutada võimsusest C kõrgemat määra.

Teoreem 4.7 *Kui* R *on tagasisidega infovahetuse saavutatav määr, siis* $R\le C$*.*

Tõestus. Argumenteerime analoogiliselt teise väite tõestusega tagasisideta kanali korral. Olgu $(2^{nR},n)$ koodide jada nii, et $\lambda_{max}\to 0$. Näitame, et $R\le C$.
Olgu W ühtlane üle tähestiku. Siis $P_e=\mathbf{P}(\hat{W}\neq W)\to 0$. Fano võrratusest saame

$$nR=H(W)=H(W|Y^n)+I(W;Y^n)\le 1+P_enR+I(W;Y^n).$$

Hindame

$$\begin{aligned}
I(W;Y^n)&=H(Y^n)-H(Y^n|W)\\
&=H(Y^n)-H(Y_1|W)-H(Y_2|Y_1,W)-\cdots-H(Y_n|Y_1,\dots,Y_{n-1},W)\\
&=H(Y^n)-\sum_{i=1}^{n}H(Y_i|Y_1,\dots,Y_{i-1},W)\\
&=H(Y^n)-\sum_{i=1}^{n}H(Y_i|Y_1,\dots,Y_{i-1},W,X_i).
\end{aligned}$$

Viimane võrdus kehtib sest $X_i=\mathcal{C}_i(Y_1,\dots,Y_{i-1},W)$. Et aga Y_i sõltub vaid X_i-st, siis

$$P(y_i|y_1,\dots,y_{i-1},w,x_i)=P(y_i|x_i)\quad\text{ja}\quad H(Y_i|Y_1,\dots,Y_{i-1},W,X_i)=H(Y_i|X_i).$$

Nüüd läheb jälle kõik vanamoodi

$$I(W;Y^n) = H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_1,\ldots,Y_{i-1},W,X_i) = H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i)$$
$$\leq \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) = \sum_{i}^{n} I(X_i,Y_i) \leq nC.$$

Kokkuvõttes $nR \leq P_e nR + 1 + nC$ ehk $R \leq P_e R + \frac{1}{n} + C \to C$. ■

Märkus: Tagasisideta infovahetuse korral kasutasime Lemmat 4.1, mis tugineb võrdusele

$$H(Y^n|X^n) \leq \sum_i H(Y_i|X_i),$$

täpsemalt seosele

$$P(y_i|y_1,\ldots,y_{i-1},x_1,\ldots,x_n) = p(y_i|x_i),$$

mis aga tagasiside korral ei kehti, sest $x_{i+1}, x_{i+2}, \ldots$ annab ka y_i kohta infot.

4.7 Kaheastmeline kodeerimine

Siiani vaatlesime juhusliku sõna W edastamist läbi kanali. Alljärgnevas uurime mõnevõrra reaalsemat probleemi. Olgu meie infoallikas juhuslik protsess $V_1, V_2, \ldots$ (digitaliseeritud kõne, muusika jne), kus iga juhusliku suuruse väärtuste hulk on $\mathcal{V}$. Eesmärk on n ülekandega läbi kanali edastada allika esimesed n sümbolit $V_1, \ldots, V_n$. Kas see on väikese veaga võimalik?

Muidugi võib vektorit $V^n = (V_1, \ldots, V_n)$ vaadelda juhusliku sõnana hulgast $\mathcal{V}^n$ ja rakendada kanaliteoreemi. Viimasest järeldub, et kui $\log|\mathcal{V}| < C$, siis leidub $(|\mathcal{V}|^n, n)$ koodide jada nii, et maksimaalne viga läheneb nullile ehk vektorit V^n võib n ülekande abil edastada kuitahes väikese veaga. Mida aga teha, kui $\log|\mathcal{V}| > C$? Järgnev teoreem väidab, et juhul kui $V_1, V_2, \ldots$ on nõrga AEP omadusega protsess, võib soovitud (n ülekannet, nulliks koonduv viga) infovahetus olla võimalik ka siis, kui $\log|\mathcal{V}| > C$. Piisav tingimus selleks on $H < C$, kus H on protsessi $V_1, V_2, \ldots$ entroopiamäär. Tähestiku suurus $\mathcal{V}$ pole enam oluline.

Teoreem 4.8 *Olgu $V^n = V_1, \ldots, V_n$ esimesed n juhuslikku suurust nõrga AEP omadusega juhuslikust protsessist, H olgu selle protsessi entroopiamäär. Kui $H < C$, siis on vektorit V^n võimalik n ülekandega edastada läbi kanali nii, et $\mathbf{P}(\hat{V}_n \neq V_n) \to 0$.*

Tõestus. Vali $\epsilon > 0$ nii väike, et $H + 2\epsilon < C$. Et protsessil on AEP omadus, siis iga piisavalt suure n korral leidub hulk W_ϵ^n (nõrgalt tüüpilised sõnad) nii, et $P(W_e^n) > 1 - \epsilon$ ja $|W_\epsilon^n| \leq 2^{n(H+\epsilon)}$. Indekseerime kõik sõnad hulgast W_ϵ^n ja nüüd võib hulka W_n^ϵ võib

vaadelda kui sõnastikku, mis koosneb 2^{nR} sõnast, kus $R \leq H + \epsilon < C$. Formaalselt oleme defineerinud funktsiooni (allikakood)

$$f : W_\epsilon^n \to \{1, \ldots, 2^{nR}\},$$

mis igale nõrgalt tüüpilisele sõnale seab vastavusse tema indeksi. Et $R \leq H + \epsilon < C$, siis kanaliteoreemist saame, et saadud sõnad saab edastada kuitahes väikese veaga. Teisisõnu, leidub $(2^{nR}, n)$ kood nii, et $\lambda_{max} \to 0$. Vastuvõtja dekodeerib esmajärjekorras indeksi hulgast W_ϵ^n ja seejärel leiab temale vastava sõna hulgast $\mathcal{V}^n$. Olgu

$$g : \mathcal{Y}^n \to \mathcal{V}^n$$

saadud dekooder. Et $\lambda_{max} \to 0$, siis piisavalt suure n korral iga nõrgalt tüüpilise sõna dekodeerimisel tehtav viga on väiksem kui ϵ. Kokkuvõttes iga piisavalt suure n korral sellisel infovahetusel tekkiva vea tõenäosus rahuldab seoseid

$$\mathbf{P}(\hat{V}^n \neq V^n) \leq \mathbf{P}(V^n \notin W_\epsilon^n) + \mathbf{P}(g(Y^n) \neq V^n | V^n \in W_\epsilon^n) \leq 2\epsilon.$$

■

Ülaltoodud tõestuses kasutasime *kaheastmelist kodeerimist:* esimene aste on allika V^n kodeerimine optimaalselt (kuid kanalist sõltumatult) ligikaudu 2^{nH} koodisõnaks (tuletame meelde, et nõrgalt tüüpilised sõnad annavad suure n korral optimaalse koodi), teine aste on saadud sõnade kodeerimine (esimesest osast sõltumatult) optimaalse infovahetuse käigus, s.t. ka kood $\mathcal{C}$ on teatavas mõttes optimaalne (kuid sõltumatu allikast V^n). Seega allika optimaalne kodeerimine koos optimaalse ning allikast sõltumatu kanali koodiga annab hea tulemuse. Samas võib need kaks sammu ühendada: sõna V^n kodeeritakse otse sõnaks x^n, mis saadetakse kohe kanalisse. Nimetame sellist protseduuri **üheastmeliseks kodeerimiseks (joint source-channel coding)**. Kui infovahetus on tagasisidega, siis üheastmeline kodeerimine tähendab koode $\mathcal{C}_i$ nii, et

$$\mathcal{C}_i : \mathcal{V}^n \times \mathcal{Y}^{i-1} \to \mathcal{X}.$$

Ülalkirjeldatud kaheastmelist kodeerimist võib vaadelda üheastmelise kodeerimise erijuhuna, mistõttu on loomulik küsida, kas üheastmelisel kodeerimisel ei saa äkki paremat tulemust, s.t. kas ei saa äkki n ülekande abil väikese veaga läbi kanali saata sõna V^n ka siis, kui $H > C$? Järgnev teoreem annab esitatud küsimusele eitava vastuse: diskreetse mäluta kanali korral tagab kaheastmeline kodeerimine optimaalse tulemuse (isegi tagasiside korral). Lisaeeldus on $|\mathcal{V}| < \infty$.

Teoreem 4.9 (Separation theorem) *Olgu $V_1, \ldots, V_n$ esimesed n juhuslikku suurust nõrga AEP omadusega statsionaarsest juhuslikust protsessist, H olgu selle protsessi entroopiamäär, $|\mathcal{V}| < \infty$. Olgu $\hat{V}^n$ vektori V^n väljund, mis on saadud tagasisidega infovahetusel n ülekande abil. Kui $H > C$, siis leidub $\epsilon > 0$ nii, et $\mathbf{P}(\hat{V} \neq V) > \epsilon$ iga n korral.*

Tõestus. Olgu

$$\mathcal{C}_i : \mathcal{V}^n \times \mathcal{Y}^{i-1} \to \mathcal{X}, \quad i = 1, \ldots, n$$

(n ülekannet) ja

$$g : \mathcal{Y}^n \to \mathcal{V}^n, \quad \hat{V} = g(Y^n).$$

Statsionaarse juhusliku protsessi korral

$$H \le \frac{H(V_1, \ldots, V_n)}{n} = \frac{1}{n} H(V^n) = \frac{1}{n} H(V^n|\hat{V}^n) + \frac{1}{n} I(V^n; \hat{V}^n).$$

Esimene võrratus kehtib, sest statsionaarsuse tõttu $H(V_n|V_1, \ldots, V_{n-1}) \searrow H$ ja

$$\begin{aligned} H(V_1, \ldots, V_n) &= H(V_1) + \cdots + H(V_n|V_1, \ldots, V_{n-1}) \\ &= H(V_n) + H(V_n|V_{n-1}) + \cdots + H(V_n|V_1, \ldots, V_{n-1}) \\ &\ge nH(V_n|V_1, \ldots, V_{n-1}). \end{aligned}$$

Fano võrratusest saame ($|\mathcal{V}|$ on lõplik)

$$H(V|\hat{V}) \le 1 + \mathbf{P}(\hat{V} \ne V^n) \log |\mathcal{V}|^n = 1 + \mathbf{P}(\hat{V} \ne V^n) n \log |\mathcal{V}|.$$

Andmetöötlusvõrratusest ($V^n \to Y^n \to \hat{V}^n$) saame

$$I(V^n; \hat{V}^n) \le I(V^n; Y^n).$$

Teoreemi 4.7 tõestusest nägime, et

$$I(V^n; Y^n) \le nC.$$

Seega

$$H \le \frac{1}{n} + \mathbf{P}(\hat{V} \ne V^n) \log |\mathcal{V}| + C.$$

Kui $P_e \to 0$, siis $H \le C$; kui $H > C$, siis

$$\mathbf{P}(\hat{V} \ne V^n) \ge \frac{H - C}{\log |\mathcal{V}|} - \frac{1}{n \log |\mathcal{V}|},$$

millest näeme, et kui $H > C$, siis leidub $\epsilon > 0$ nii, et $\mathbf{P}(\hat{V} \ne V^n) > \epsilon$, kui n on piisavalt suur. See aga tähendab, et leidub $\epsilon > 0$ nii, et $\mathbf{P}(\hat{V} \ne V^n) > \epsilon$ iga n korral. ■

Seega üheastmeline (kombineeritud) kodeerimine ja tagasiside ei suurenda infovahetuse efektiivsust: kaheastmeline kodeerimine annab sama hea tulemuse. Kuigi see paistab loomulik, pole see iseenesestmõistetav ning keerulisemate kanalite korral ei pruugi ka kehtida. Seetõttu on teoreemil 4.9 suur tähtsus praktikas, sest ta lubab allika koode ja infovahetust optimiseerida teineteisest sõltumatult. Samuti lubab see teoreem saata erinevaid allikaid läbi sama (kord juba optimiseeritud infovahetusega) kanali. Samuti lubab ta saata sama (kord juba optimaalselt kodeeritud) allikat läbi erinevate kanalite.

Teisest küljest aga tuleb alati meeles pidada, et tõestatud kahe- ja üheastmelise kodeerimise ekvivalentsus on asümptootiline. Lõpliku n korral võib aga üheastmeline kodeerimine ikkagi vähendada vea tõenäosust.

Mida teha, kui $H > C$? Teoreemist 4.9 järeldub, et n ülekandega soovitud tulemust ei saavuta: leidub $\delta > 0$ nii, et n ülekande abil saadud hinnang $\hat{V}^n$ rahuldab seost $\mathbf{P}(\hat{V}^n \neq V^n) > \delta$. Saavutamaks väikest viga, tuleb seega teha rohkem ülekandeid. Tuletame meelde, et kaheastmelise kodeerimise korral on esimese kodeerimise tulemus ligikaudu $M := 2^{nH}$ koodisõna. Kui $H > C$, siis n ülekandega neid koodisõnu nulliks koonduva veaga edastada ei saa. Et aga

$$M = 2^{nH} = 2^{\frac{H}{k}(kn)},$$

siis mingi positiivse täisarvu k (ja piisavalt suure n) korral saab neid M koodisõna edastada kn ülekandega nii, et viga on kuitahes väike. Siin k peab olema selline, et $\frac{H}{k} < C$.

4.8 Ülesanded

1. Vaatleme binaarset sümmeetrilist kanalit, $p < 0.5$. Olgu m paaritu ja olgu $p_b(m)$ kordamiskoodi R_m korral ühe bloki vigase dekodeerimise tõenäosus.

   1 Tõesta, et
   $$p_b(m) = \sum_{k>\frac{m}{2}}^{m} \binom{m}{k} p^k (1-p)^{m-k}.$$

   2 Suurte arvude seadusest järelda, et $\lim_m P_b(m) = 0$.

2. Olgu $\mathcal{X} = \{0, 1\}$. Vaatleme kanalit, kus sisendile X liidetakse sõltumatu juhuslik suurus aZ, kus $Z \sim B(1, 0.5)$. Leida selle kanali võimsus.

3. Olgu $\mathcal{X} = \{0, \ldots, 10\}$. Vaatleme kanalit, kus $Y = X + Z \pmod{11}$, kus X on sisend, Y on väljund ning Z on sõltumatu juhuslikust suurusest X. Juhusliku suuruse Z jaotus olgu

1	2	3
$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

   Leida kanali võimsus. Milline jaotus saavutab võimsuse?

4. Olgu $(\mathcal{X}_1, P_1(y|x), \mathcal{Y}_1)$ ja $(\mathcal{X}_2, P_2(y|x), \mathcal{Y}_2)$ kanalid võimsustega C_1 ja C_2. Defineerime korrutiskanali
   $$(\mathcal{X}_1 \times \mathcal{X}_2, P_1(y_1|x_1)P_2(y_2|x_2), \mathcal{Y}_1 \times \mathcal{Y}_2).$$
   Leida selle kanali võimsus.

5. Olgu $K(\epsilon)$ binaarne sümmeetriline kanal veatõenäosusega ϵ. Olgu $K(\epsilon_1) \to K(\epsilon_2)$ jadaühendus.

- Leida jadaühendusel saadud kanali võimsus C.
- Tõestada, et $C \le C(K(\epsilon_1)) \wedge C(K(\epsilon_2))$.
- Tõestada, et kanali $K(\epsilon)$ n-kordsel jadaühendusel

$$X \to K(\epsilon) \to K(\epsilon) \to \cdots \to K(\epsilon) \to Y(n)$$

saadud kanal on $K(\frac{1}{2}\big(1-(1-2\epsilon)^n\big))$, millest $\lim_n I(X;Y(n)) = 0$.

6. Leida järgmise *Z-kanali* võimsus ja seda saavutav jaotus

$$\begin{pmatrix} 1 & 0 \\ 0.5 & 0.5 \end{pmatrix}$$

Olgu kanal Z-kanal. Vaatleme juhuslikku $(n, 2^{nR})$ koodi, kus iga koodisõna on iid $B(1, \frac{1}{2})$ jaotusega. Millise R korral läheneb üle kõigi võimalike koodide keskmine viga P_e nullile?

7. Vaatleme binaarseid sümmeetrilisi kanaleid $Y_i = X_i + Z_i \pmod 2$, kus $\mathcal{X} = \mathcal{Y} = \{0,1\}$. Olgu $Z^n = Z_1, \ldots, Z_n$ sama jaotusega (kuid mitte sõltumatud) juhuslikud suurused, $Z_i \sim B(1,\epsilon)$, vektor Z^n on sõltumatu juhuslikust vektorist $X^n = X_1, \ldots, X_n$. Seega on n binaarset sümmeetrilist kanalit veatõenäosusega ϵ. Kui aga juhuslikud suurused Z_i pole sõltumatud, on kanalid mäluga.

- Tõestada, et $I(X^n;Y^n) \le n - h(\epsilon)$. Leida X^n ja Z^n jaotus, mis saavutab võrduse.
- Veenduda, et mälu suurendab kanali võimsust ehk

$$\max_{P(x^n)} I(X^n, Y^n) > nC.$$

8. Olgu $(\mathcal{X}, P_1, \mathcal{X})$ ja $(\mathcal{X}, P_2, \mathcal{X})$ kaks kanalit võimsustega vastavalt C_1 ja C_2. Olgu C kanali $(\mathcal{X}, P_1P_2, \mathcal{X})$ võimsus. Tõestada, et

$$C \le C_1 \wedge C_2.$$

9. Olgu $x^n(1), \ldots, x^n(2^{nR})$ koodiraamat. Dekodeeriv funktsioon (suurime tõepära dekooder)g olgu

$$g(y^n) = \arg\max_i P(y^n|x^n(i)) = \arg\max_i \mathbf{P}(Y^n = y^n | W = i).$$

Olgu W jaotus ühtlane.

- Tõestada, et g minimiseerib vea tõenäosuse

$$P_e = \mathbf{P}(g(Y^n) \neq W) = \frac{1}{2^{nR}} \sum_{i=1}^{2^{nR}} P(g(Y^n) \neq i | W = i) = \frac{1}{2^{nR}} \sum_{i=1}^{2^{nR}} \lambda_i$$

üle kõikide dekodeerivate funktsioonide.

- Leida kontranäide selle kohta, et g ei minimiseeri λ_{max} üle kõikide dekodeerivate funktsioonide.

Näpunäide: Näita, et

$$\arg\max_i \mathbf{P}(Y^n = y^n | W = i) = \arg\max_i \mathbf{P}(W = i | Y^n = y^n) =: g^*(y^n).$$

Seejärel veendu, et iga teise dekodeeriva funktsioooni g korral

$$\mathbf{P}(W \neq g^*(y^n) | Y^n = y^n) \leq \mathbf{P}(W \neq g(y^n) | Y^n = y^n), \quad \forall y^n.$$

10. Olgu $K(\epsilon)$ binaarne sümmeetriline kanal, kusjuures $\epsilon < \frac{1}{2}$. Olgu $x^n(1), \ldots, x^n(2^{nR})$ koodiraamat. Iga kahe vektori $x^n, y^n \in \{0,1\}$ korral defineerime *Hammingu kauguse*

$$d(x^n, y^n) = \sum_{i=1}^{n} |x_i - y_i|.$$

Olgu dekodeeriv funktsioon

$$g(y^n) = \arg\min_i d(y^n, x^n(i)).$$

Tõestada, et g on eelmises ülesandes defineeritud suurime tõepära dekooder.

11. Olgu $\mathcal{X} = \mathcal{Y} = \{0,1,2,3,4\}$. Olgu kanal antud üleminekutõenäosuste maatriksiga

$$\frac{1}{2}\begin{pmatrix} 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Leida koodiraamat $x^2(1), \ldots, x^2(5)$ nii, et iga sõna saab edasi anad veatult, st leidub g nii, et $\mathbf{P}(g(Y^2) = i | W = i) = 0$ iga $i = 1, \ldots, 5$ korral.

5 Lempel-Ziv kood

5.1 Asümptootiliselt optimaalsed koodid

Tuleta meelde sõnade kodeerimist (alam-peatükk 2.6). Sellest teame, et kui informatsiooniallikas on statsionaarne protsess, siis leiduvad prefikskoodid

$$\mathcal{C}_n : \mathcal{X}^n \to \{0,1\}^*$$

nii, et keskmised koodipikkused tähe kohta koonduvad protsessi entroopiamääraks (koondumine (2.6)):

$$L_n = \frac{1}{n} El(X_1, \ldots, X_n) \to H_X.$$

Ülaltoodud koondumine kehtib ka siis, kui iga n korral $\mathcal{C}_n$ on optimaalne (Huffmani kood) vektori $(X_1, \ldots, X_n)$ jaoks. Sellisel juhul L_n on vähim võimalik üle kõikide võimalike sõnade koodide, mistõttu suvaliste koodide jada $\mathcal{C}_n$ korral kehtib

$$\liminf_n L_n = \liminf \frac{1}{n} El(X_1, \ldots, X_n) \geq H_X.$$

Kui informatsiooniallikas on lisaks nõrk AEP protsess, siis teame, et leiduvad prefikskoodid – Shannon-Fano koodid – mille korral koodisõnade pikkuste entroopiamääraks koondumine kehtib ka peaaegu kindlasti (veendu selles!):

$$\frac{l(X_1, \ldots, X_n)}{n} \to H_X, \quad \text{p.k..} \tag{5.1}$$

Selgub, et ka seda tõket ei saa parandada, sest kehtib järgmine teoreem.

Teoreem 5.1 *Olgu $\mathcal{C}_n : \mathcal{X}^n \to \{0,1\}^*$ prefikskoodide jada, $X_1, X_2, \ldots$ olgu nõrga AEP omadusega protsess entroopiamääraga H_X. Siis*

$$\liminf_n \frac{1}{n} l(X_1, \ldots, X_n) \geq H_X \quad p.k., \tag{5.2}$$

kus $l(x_1, \ldots, x_n) = |\mathcal{C}_n(x_1, \ldots, x_n)|$.

Teoreemist 5.1 järeldub, et kui jada $\frac{l(X_1,\ldots,X_n)}{n}$ koondub mingiks (mitte ilmtingimata konstantseks) piirväärtuseks p.k., siis see piirväärtus peab olema vähemalt H_X p.k.. Teisisõnu, koondumine (5.1) on (teatavas mõttes) parim võimalik. Siit ka järgmine definitsioon.

Def 5.2 *Koodide $\mathcal{C}_n : \mathcal{X}^n \to \{0,1\}^*$ jada nimetatakse* **asümptootiliselt optimaalseks (asymptotically optimal)**, *kui*

$$\frac{1}{n} l(X_1, \ldots, X_n) \to H_X, \ \ p.k..$$

Seega nõrga AEP omadusega informatsiooniallika korral Shannon-Fano koodide jada on asümptootiliselt optimaalne.

Teoreemi 5.1 tõestus. Tuleta meelde, et

$$x^n := (x_1, \ldots, x_n), \quad X^n = (X_1, \ldots, X_n).$$

Suvalise juhusliku protsessi $X = X_1, X_2, \ldots$ korral tähistame $x = x_1, x_2, \ldots$ (võimalik realisatsioon) ning

$$P(x^n) := \mathbf{P}(X^n = x^n).$$

Teoreemi 5.1 tõestus põhineb järgmisel lemmal.

Lemma 5.1 (Barron) *Olgu $\mathcal{C}_n : \mathcal{X}^n \to \{0,1\}^*$ prefikskoodide jada, X olgu juhuslik protsess. Olgu α_n selline positiivsete numbrite jada, et $\sum_n 2^{-\alpha_n} < \infty$. Siis*

$$\mathbf{P}\Big(l(X^n) + \log P(X_1, \ldots, X_n) \geq -\alpha_n \quad ev.\Big) = 1. \tag{5.3}$$

Tõestus. Paneme tähele, et

$$\begin{aligned} B_n :=& \{x^n : l(x^n) + \log P(x^n) \leq -\alpha_n\} = \{x^n : 2^{l(x^n)+\log P(x^n)} \leq 2^{-\alpha_n}\} \\ =& \{x^n : 2^{l(x^n)} 2^{\log P(x^n)} \leq 2^{-\alpha_n}\} = \{x^n : P(x^n) \leq 2^{-\alpha_n} 2^{-l(x^n)}\}. \end{aligned}$$

Seega Krafti võrratusest järeldub

$$P(B_n) = \sum_{x \in B_n} P(x^n) \leq \sum_{x \in B_n} 2^{-\alpha_n} 2^{-l(x^n)} = 2^{-\alpha_n} \sum_{x^n \in \mathcal{X}^n} 2^{-l(x^n)} \leq 2^{-\alpha_n}.$$

Boreli-Cantelli I lemmast järeldub, et

$$P(\limsup_n B_n) = P\{x : x \in B_n \text{ i.o. }\} = 0 \quad \Rightarrow \quad = P\{x : x \in B_n^c \text{ ev. }\} = P(\liminf_n B_n^c) = 1$$

ehk

$$P\{x : x \in B_n^c \text{ ev. }\} = P\{x : l(x^n) + \log P(x^n) > -\alpha_n \text{ ev. }\} = \mathbf{P}(l(X^n) + \log P(X^n) > -\alpha_n \quad \text{ev.}) = 1.$$

■

Võttes $\alpha_n = 2 \log n = \log n^2$, saame

$$\sum_n 2^{-\alpha_n} = \sum_n n^{-2} < \infty, \quad \frac{\alpha_n}{n} \to 0.$$

Rakendades ülaltoodud lemmat, saame nõrga AEP omaduse tõttu

$$\liminf_n \frac{l(X^n)}{n} \geq \liminf_n \frac{-\log P(X^n) - \alpha_n}{n} = \liminf_n \frac{-\log P(X^n)}{n} = H_X, \quad \text{p.k.}$$

mis on (5.2). ■

Järeldus 5.1 *Olgu $\mathcal{C}_n : \mathcal{X}^n \to \{0,1\}^*$ üheselt dekodeeritavate koodide jada, $X_1, X_2, \ldots$ olgu nõrga AEP omadusega protsess entroopiamääraga H_X. Siis kehtib (5.2)*

Tõestus. Eliase laiendi abil saab suvalise üheselt dekodeeritava koodi muuta prefikskkoodiks. Sõna x^n koodisõna pikkus $l(x^n)$ suureneb $\log l(x^n) + o(\log l(x^n))$ võrra. Seega, kui

$$\liminf_n \frac{l(x^n)}{n} < H_X$$

on ka

$$\liminf_n \frac{l(x^n) + \log l(x^n) + o(\log l(x^n))}{n} < H_X.$$

■

5.2 Universaalsed koodid ning Lempel-Ziv kood

Shannon-Fano kood on küll asümptootiliselt optimaalne, kuid selle konstrueerimiseks on vaja teada iga n korral vektori $X^n = (X_1, \ldots, X_n)$ jaotust. Ka Huffmani koodi konstrueerimiseks on tarvis teada sõna jaotust. Praktikas pole aga jaotus enamasti teada, mistõttu pakuvad huvi koodid, mis oleksid asümptootiliselt optimaalsed iga nõrga AEP omadusega protsessi korral. Koode, mis ei sõltu allika jaotusest nimetatakse **universaalseteks**. Kas aga sellised koodid üldse leiduvad? Esimesed universaalsed asümptootiliselt optimaalsed koodid esitasid aastatel 1977 ja 1978 A. Lempel ja J. Ziv. Seetõttu nimetatakse neid (ja teisi sarnasel põhimõttel töötavaid koode) **Lempel-Ziv (LZ)** koodideks, lühidalt LZ77 ja LZ78. Järgnevas tutvume põgusalt koodiga LZ78. LZ koodid on olemuselt väga lihtsad, mistõttu neid (eriti LZ78 koodi) kasutatakse kompressiooniprogramides (UNIX: "compress", Mac "StuffIt", PC: "arc"). Asümptootilise optimaalsuse tõttu on LZ koodide kasutamine (teatud mõttes) teoreetiliselt õigustatud.

5.2.1 Liigendamine ja kodeerimine

Olgu $\mathcal{X}$ *lõplik* tähestik, vektor (jada) $x^n \in \mathcal{X}^n$ on koodi sisend. Kõik LZ koodid põhinevad sisendjada x^n jagamiseks alamsõnadeks – **liigendamisel (parsing)**. Koodi LZ78 liigendamine seisneb jada x^n jagamine sõnadeks $w(1), w(2), \ldots, w(K)$ nii, et **järgmine sõna on lühim uus sõna**. Seega esimene sõna on alati ühetäheline, teine sõna ülimalt kahetäheline jne. Formaalselt on liigendamiseeskiri järgmine:

a) Esimene sõna on x_1.

b) Olgu $x^{n_j} = w(1) \cdots w(j)$.

$$\text{kui } x_{n_j+1} \notin \{w(1), \ldots, w(j)\}, \text{ siis } w(j+1) = x_{n_j+1},$$
$$\text{kui } x_{n_j+1} \in \{w(1), \ldots, w(j)\}, \text{ siis } w(j+1) = x_{n_j+1}^{m+1},$$

kus $m > n_j$ on väikseim indeks nii, et

$$x_{n_j+1}^m \in \{w(1), \ldots, w(j)\}, \text{ kuid } x_{n_j+1}^{m+1} \notin \{w(1), \ldots, w(j)\}.$$

Näide: Kui $x^{18} = 110010100010001000$, siis liigendus on järgmine:

$$1, 10, 0, 101, 00, 01, 000, 100, 0$$

Pärast liigendust esitub sisendvektor sõnade jadana:

$$x^n = w(1)w(2)\cdots w(K)v, \tag{5.4}$$

kus viimane osa v on on kas tühi hulk või võrdub mingi eelpool oleva sõnaga. Ülaltoodud näites $v = w(3) = 0$.
On selge, et iga liigenduses olev sõna $w(i)$ erineb ühest oma eelkäijast vaid viimase tähe poolest. Seega on iga sõna üheselt määratud eelpoolnimetatud eelkäija ja viimase tähega.

Näide: Ülaltoodud liigenduse võib esitada seega järgmiselt

$$(0,1), (1,0), (0,0), (2,1), (3,0), (3,1), (5,0), (2,0), 0.$$

Siin igas paaris esimene arv näitab eelpoololeva sõna indeksit ja teine arv lisatud bitti. Kui esimene arv on 0, siis järgnev sümbol on uus sõna. Veendu, et kasutades ülaltoodud paare saad rekonstrueerida esialgse jada. Nüüd kodeerime sõnade indeksid ja viimased tähed ning saamegi LZ koodi. Formaalselt käib see järgmiselt: (kahend)kodeerigu

$$f : \{1, \ldots, n\} \to \{0,1\}^{\lceil \log n \rceil}$$

sõnade indekseid. Kodeerigu

$$g : \mathcal{X} \to \{0,1\}^{\lceil \log |\mathcal{X}| \rceil}$$

tähti. Defineerime koodi

$$\mathcal{C}_n : \mathcal{X}^n \to \{0,1\}^*, \quad \mathcal{C}_n(x^n) = b(1)b(2)\cdots b(K)b(K+1),$$

kus sõnad $b(i)$ on saadud liigendusest (5.4) järgmise eeskirja alusel.

5.2.2 LZ algoritm:

1) kui $j \leq K$ ja $|w(j)| = 1$, siis $b(j) = 0g(w(j))$.

2) kui $j \leq K$ ja $i < j$ on selline, et $w(j) = w(i)a$, siis $b(j) = 1f(i)g(a)$.

3) kui $v = \emptyset$, siis $b(K+1) = \emptyset$. Kui $v = w(i)$, siis $b(K+1) = 1f(i)$.

Seega lisasümbol 0 näitab, et järgneb tähe kood; lisasümbol 1 näitab, et järgneb sõna (indeksi) kood ning sellele järgnev kood on tähe kood (või ei järgne midagi).

Näide: Olgu $\mathcal{X} = \{0,1\}$. Siis $g(a) = a$. Olgu $n = 18$. Siis $f : \{1, \ldots, 18\} \to \{0,1\}^5$. Olgu $f(i)$ arvu $i-1$ ühtlane kood, st

$$f(1) = 00000, \quad f(2) = 10000, \quad f(3) = 01000, \quad f(4) = 00100 \quad f(5) = 00010, \quad f(6) = 00001, \cdots .$$

Leiame $\mathcal{C}_{18}(1100101000100010 00)$. Toodud vektori liigendus on meile tuttav:

$$(0,1),(1,0),(0,0),(2,1),(3,0),(3,1),(5,0),(2,0),0.$$

Seega $K = 8$ ja $b(1) = 0g(1) = 01$, $b(2) = 1f(1)g(0) = 1000000$, $b(3) = 0g(0) = 00$, $b(4) = 1f(2)g(1) = 1100001$, $b(5) = 1f(3)g(0) = 1010000$, $b(6) = 1f(3)g(1) = 1010001$, $b(7) = 1f(5)g(0) = 1000100$, $b(8) = 1f(2)g(0) = 1100000$, $b(9) = 1f(3) = 101000$. Seega

$$\mathcal{C}_{18}(1100101000100010 00) = 0110000000011000011010000101000110001001100000101000.$$

Dekodeerija peab teadma numbrite koodi f ja tähtede koogi g. Näites oleva teksti dekodeerimiseks liigendame omakorda kodeeritud teksti

0 1 1 00000 0 0 0 0 1 00001 1 1 00010 0 1 00010 1 1 00010 0 1 00001 0 1 00101 0 1 00010.

Kui liigendusel saadud sõna algab ühega, järgneb sellele viietäheline numbrikood (antud juhul number) ja uue tähe kood, kui uus sõna algab nulliga, järgneb sellele number. Seejärel dekodeerime numrid, vaatame eelmisi sõnu ja dekodeerime teksti.

Nagu näha, ei anna lühikeste sõnade LZ kodeerimine erilist efekti, pigem vastupidi. Paneme tähele, et koodi saab lühendada, kui f kodeerib vaid sõnade $w(i)$ indeksi. Ülaltoodud näites $K = 8$, seega võib võtta $f : \{1,\dots,8\} \to \{0,1\}^4$. Selline f sõltub aga sisendist x^{18} ja nii tuleks kodeerimisel sisend läbida kaks korda: esimene kord liigendada sisend ja määrata sõnade arv, teisel korral aga kodeerida. Ülalesitatud algoritm kodeerib sisendit *on-line*. Asümptootiliselt on erinevad kodeerimisvariandid samad.

Veel üks võimalus kahendteksti kodeerimisel on kasutada numbrite kodeerimisel kahendkoodi, kusjuures koodisõnade pikkus sõltub viidatavate sõnade arvust: kui viidatavate sõnade arv on k, kasutame $\lceil \log k \rceil$ biti pikkusi kahendsõnu. Vaatame tuttavat näidet: vektori 1100101000100010 00 liigendus on endiselt järgmine

$$(0,1),(1,0),(0,0),(2,1),(3,0),(3,1),(5,0),(2,0),0.$$

Esimene täht on alati uus sõna, ja esimesele tähele eelnevat nulli me ei kodeeri, seega paarist (0,1) saab 1. Nüüd on meil kaks eelnevat sõna $\emptyset, \{1\}$ ja neile saab viidata ühe bitiga (0-uus, 1-esimene), seega paar (1,0) jääb muutumatuks. Nüüd on meil 3 eelnevat sõna (uus, esimene ja teine) ning nendele viitamiseks on vaja kaht bitti [00-uus, 01-esimene ja 10-teine), seega paarist (0,0) saab (00,0) ja eale seda on meil 4 eelnevat sõna (uus, esimene, teine ja kolmas) ning ka neile seeb viidata kahe bitiga ehk (2,1) $\mapsto$ (10,0). Edaspidi tuleb numbrite kodeerimiseks kasutada kolmekohalisi kahendarve, seega (3,0) $\mapsto$ (011,0); (3,1) $\mapsto$ (011,1); (5,0) $\mapsto$ (101,0) ja (2,0) $\mapsto$ (010,0). Seega

$$\mathcal{C}(1100101000100010 00) = 1100001000110011110100100.$$

Saame palju lühema sõna. Dekodeerimisel peame samuti arvestama, et koodide pikkused muutuvad. Dekodeerimise liigendus

1 1 0 00 0 10 0 011 0 011 1 101 0 010 0

Siit saame paarid

$$(0,1),(1,0),(0,0),(2,1),(3,0),(3,1),(5,0),(2,0)$$

ning dekodeerime teksti. Paneme tähele, et dekodeerimisel pole enan vaja teada kodeeritava sõna pikkust või liigenduste arvu.

Näited: Olgu sisendssõna 000000000000100000000000. Liigendus

$$(0,0),(1,0),(2,0),(3,0),(2,1),(4,0)(6,0).$$

Pärast numbrite dekodeerimist saame liigenduseks

$$0,(1,0),(10,0),(11,0),(010,1),(100,0),(110,0).$$

Seega kood

$$\mathcal{C}(000000000000100000000000) = 010100110010110001100.$$

Proovige kirjeldatud meetodil dekodeerida sõna

$$00101011101100100100011010101000011$$

(Vastus: 010000100010001010100000 1 .)

5.2.3 Lempel-Ziv teoreem

LZ koodi asümptootilise optimaalsuse näitab järgmine kuulus teoreem. Teoreem eeldab, et sisendprotsess $X = X_2, X_2, \ldots$ on *ergoodiline protsess*. Iga ergoodiline protsess on statsionaarne (st tal on entroopiamäär) ning nõrga AEP omadusega.

Teoreem 5.3 (Lempel-Zivi teoreem) *Kui X on ergoodiline protsess entroopiamääraga H_X, ja $\mathcal{C}_n$ on LZ kood, siis*

$$\limsup_n \frac{l(X_1,\ldots,X_n)}{n} = \limsup_n \frac{K(X_1,\ldots,X_n)\log n}{n} = H_X, \quad \text{p.k.},$$

kus $l(x_1,\ldots,x_n) = |\mathcal{C}_n(x_1,\ldots,x_n)|$.

Teoreemi tõestamiseks tuleb hinnata suurust $l(x^n) = l(x_1,\ldots,x_n)$. Vaatleme veelkord LZ78 algoritmi. Osa **1)** järgi kodeerimiseks kulub

$$|\mathcal{X}|(\lceil \log |\mathcal{X}| \rceil + 1) =: A$$

bitti. Osa **2)** järgi kulub ühe sõna $w(j)$ kodeerimiseks $(\lceil \log n \rceil + 1 + \lceil \log |\mathcal{X}| \rceil)$ bitti. Kokku kulub osa **2)** järgi kodeerimiseks

$$K(\lceil \log n \rceil + 1 + \lceil \log |\mathcal{X}| \rceil) = K(\lceil \log n \rceil + B)$$

bitti (siin $B := 1 + \ulcorner \log |\mathcal{X}| \urcorner$. Osa **3)** nõuab ülimalt

$$\ulcorner \log n \urcorner + 1$$

bitti. Tuletame meelde, et liigendusel saadud sõnade arv $K = K(x^n)$ sõltub sisendist x^n. Seega

$$\begin{aligned} l(x^n) &\leq A + K(\ulcorner \log n \urcorner + B) + \ulcorner \log n \urcorner + 1 \leq (\log n + 1)(K+1) + KB + A + 1 \\ &= K \log n + \log n + K(B+1) + A + 2. \end{aligned}$$

Et A ja B on konstandid, on domineeriv liige $K \log n$. LZ teoreemi tõestus seisnebki seose

$$\limsup_n \frac{K(X^n) \log n}{n} = H_X, \quad \text{p.k.}$$

näitamises.

Märkus: Lempel-Zivi teoreemist järeldub LZ koodi asümptootiline optimaalsus:

$$\frac{1}{n} l(X^n) \to H_X \quad \text{p.k.}$$

Sellest koondumises aga ei järeldu vahetult koondumine

$$L_n = \frac{El(X^n)}{n} \to H_X. \tag{5.5}$$

Lempel-Zivi teoreemi tõestusest selgub aga, et jada $\frac{l(X^n)}{n}$ on p.k. tõkestatud, sest

$$\frac{K(X^n) \log n}{n}$$

on tõkestatud jadaga, mille ülemine piirväärtus on $\log |\mathcal{X}|$. Domineeritud koondumise tereemist saame, et kehtib ka (5.5).

6 Diferentsiaalentroopia ja MaxEnt printsiip

Informatsiooniteooria põhimõisted – entroopia, tinglik entroopia, vastastikune informatsioon, K-L kaugus jt – olid siiani defineeritud vaid diskreetsetel jaotustel. Loomulikult tekib küsimus: kas ja kuidas üldistuvad need mõisted pidevatele (ja kõikidele muudele) tõenäosusjaotustele. Järgnevas tutvustame nende mõistete loomulikku üldistust pidevatele jaotustele. Kuigi üldistused on enesestmõistetavad, puudub neil selline üheselt interpreteeritav tähendus kui diskreetsete jaotuste korral.

6.1 Diferentsiaalentroopia

Olgu X pidev juhuslik suurus jaotusega P ja tihedusega f. Olgu $S = \text{supp}(P)$ jaotuse P *kandja* – väikseim kinnine hulk, mis sisaldab hulka $\{x : f(x) > 0\}$. Olgu $0 \log 0 := 0$.

Def 6.1 *Juhusliku suuruse X (jaotuse P, tiheduse f)* **diferentsiaalentroopia (differential entropy)** *on*

$$h(X) :=: h(P) :=: h(f) := \int -f(x) \log f(x) dx = \int_S -f(x) \log f(x) dx, \tag{6.1}$$

kui see integraal eksisteerib.

Märkused:

- Integraal (6.1) ei pruugi alati olla defineeritud. Sellisel juhul pole ka diferentsiaalentroopia defineeritud.
- Erinevalt entroopiast võib diferentsiaalentroopia olla ka negatiivne. Üldiselt võib diferentsiaalentroopia olla nii $+\infty$ kui ka $-\infty$.
- Ülaltoodust johtuvalt võib diferentsiaalentroopia olla 0 ka siis, kui X pole p.k. konstant. Teisisõnu: sellest, et difenrentsiaalentroopia on 0 ei järeldu, et X on mittejuhuslik.
- Harilikult defineeritakse diferentsiaalentroopia (ning kõik teised alljärgnevad mõisted) naturaallogaritmi abil. Meie jääme kahendlogaritmi juurde.

6.1.1 Näited

Ühtlane jaotus. Olgu $X \sim U(0, a)$. Siis $f(x) = \frac{1}{a} I_{(0,a)}$ ja

$$h(X) = -\int_0^a \frac{1}{a} \log a dx = \log a.$$

Nagu näha, kui $a = 1$, siis $h(X) = 0$ ning

$$\lim_{a \to \infty} h(X) = \infty, \quad \lim_{a \to 0} h(X) = -\infty.$$

Normaaljaotus. Olgu $X \sim \mathcal{N}(0, \sigma^2)$. Siis

$$\begin{aligned}\int_{-\infty}^{\infty} f(x) \ln f(x) dx &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \Big(-\ln \sqrt{2\pi\sigma^2} - \frac{x^2}{2\sigma^2} \Big) dx \\ &= -\ln \sqrt{2\pi\sigma^2} - \int_{-\infty}^{\infty} \frac{x^2}{2\sigma^2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} dx \\ &= -\frac{EX^2}{2\sigma^2} - \ln \sqrt{2\pi\sigma^2} \\ &= -(\frac{1}{2} + \ln \sqrt{2\pi\sigma^2}) = -\frac{1}{2} \ln(e 2\pi\sigma^2).\end{aligned}$$

Seega

$$h_e(X) := -\int_{-\infty}^{\infty} f(x) \ln f(x) dx = \frac{1}{2} \ln(e 2\pi\sigma^2)$$

ning, et $\ln(a) = \ln 2 \log a$, siis

$$-\int_{-\infty}^{\infty} f(x) \log f(x) dx = \frac{1}{\ln 2} h_e(X) = \frac{1}{2} \log(e 2\pi\sigma^2).$$

Eksponentjaotus. Olgu $X \sim E(\lambda)$ s.t

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0.$$

Seega

$$\int_0^{\infty} f(x) \ln f(x) dx = \ln \lambda - \int_0^{\infty} \lambda x f(x) dx = \ln \lambda - 1,$$

millest $h_e(X) = 1 - \ln \lambda$ ja

$$h(X) = \frac{1}{\ln 2} - \log \lambda.$$

Märkus: Ülaltoodud näidetes on $h > -\infty$, kusjuures entroopia läheneb $-\infty$ siis, kui dispersioon läheneb nullile ehk juhuslikud suurused lähenevad (mittejuhuslikule) konstandile. Sellest võib sugeneda lootus, et $h(X) = -\infty$ parajasti siis, kui $X = c$ p.k. See ei ole nii, sest leidub (mittekõdunenud) jaotusi, mille korral differentsiaalentroopia on $-\infty$.

6.2 Pideva juhusliku suuruse kvantiseerimine

Pideva jaotuse **kvantiseerimine (quantization)** on jaotuse lähendamine diskreetse jaotusega (nt histogramm). Esmapilgul võib tunduda, et kvantiseerimisel saadud diskreetse jaotuse entroopia peaks olema "lähedane" vastavale diferentsiaalentroopiale. Arusaadavalt pole see aga nii (kas või juba sellepärast, et difenentsiaalentroopia võib olla ka negatiivne).

Oletame, et tihedusega f antud pideva jaotuse kandja on jaotatud pikkusega Δ intervsallideks. Eeldame (lihtsuse mõttes), et tihedusfunktsioon on igal intervallil

$$I_i := (i\Delta, (i+1)\Delta)$$

pidev. Siis leidub $x_i \in I_i$ nii, et

$$f(x_i)\Delta = \int_{i\Delta}^{(i+1)\Delta} f(x)dx.$$

Defineerime diskreetse jaotuse

$$P(\Delta) = \{x_i, p_i\}, \text{ kus } p_i := \int_{i\Delta}^{(i+1)\Delta} f(x)dx = f(x_i)\Delta.$$

Selle jaotuse entroopia on

$$\begin{aligned} H(P(\Delta)) &= -\sum_i p_i \log p_i \\ &= -\sum_i f(x_i)\Delta \log(f(x_i)\Delta) \\ &= -\sum_i f(x_i)\Delta \log(f(x_i)) - \log(\Delta)\sum_i f(x_i)\Delta \\ &= -\sum_i f(x_i)\Delta \log(f(x_i)) - \log(\Delta), \end{aligned}$$

sest

$$\Delta\sum_i f(x_i) = \sum \int_{i\Delta}^{(i+1)\Delta} f(x)dx = \int f(x)dx = 1.$$

Kui $f(x)\log f(x)$ on Riemanni mõttes integreeruv, siis

$$\lim_{\Delta\to 0} -\sum_i f(x_i)\Delta \log(f(x_i)) = -\int f(x)\log f(x)dx = h(f),$$

millest

$$\lim_{\Delta\to 0} H(P(\Delta)) + \log\Delta = h(f). \tag{6.2}$$

Kui näiteks $\Delta = n^{-1}$, siis suure n korral seosest (6.2) saame

$$H(P(\frac{1}{n})) - \log n \approx h(f).$$

Näide: Olgu $X \sim U(0,1)$, $\Delta = 2^{-n}$. Siis $H(P(\Delta)) = n$ ja $\log\Delta = -n$, millest näeme, et (6.2) kehtib iga n korral võrdusena:

$$H(P(\Delta)) + \log\Delta = 0 = h(f).$$

Kokkuvõtteks: Kvantiseerides saab hinnata pideva jaotuse momente. Näiteks, ülaltoodud kvantiseerimise korral iga Riemaani mõttes integreeruva funktsiooni g korral

$$\sum_i g(x_i)p_i = \sum_i g(x_i)f(x_i)\Delta_i \to \int g(x)f(x)dx,$$

kui $\Delta \to 0$ ja parempoolne integraal eksisteerib. Kuid kvantiseerimist ei saa kasutada entroopia hindamiseks.

6.3 AEP ja diferentsiaalentroopia

Tuletame meelde, et kui $X_1, X_2, \ldots$ on AEP omadusega juhuslik protsess (tähestikul $\mathcal{X}$) entroopiamääraga H, siis iga $\epsilon > 0$ korral leidub $n(\epsilon)$ ja hulk $W_\epsilon^n \subset \mathcal{X}^n$ nii, et $P(W_\epsilon^n) > 1-\epsilon$,

$$(1-\epsilon)2^{n(H-\epsilon)} \le |W_\epsilon^n| \le 2^{n(H+\epsilon)} \tag{6.3}$$

ning iga $x \in W_\epsilon^n$ korral

$$2^{-n(H+\epsilon)} \le P(x^n) \le 2^{-n(H-\epsilon)}.$$

Muuhulgas kehtib ülaltoodud omadus siis, kui $X_1, X_2, \ldots$ on i.i.d. juhuslikud suurused $X_i \sim P$ ja $H = H(P)$.

Olgu nüüd $X_1, X_2, \ldots$ i.i.d. pidevad juhuslikud suurused. Selgub, et AEP omadus kehtib ka nüüd, kuid hulga W_ϵ^n võimsuse asemel on seoses (6.3) tema ruumala ja entroopia asemel on diferentsiaalentroopia.

Def 6.2 *Mõõtuva hulga $A \subset \mathbb{R}$ ruumala on*

$$\mathbf{V}(A) := \int_A dx_1 \cdots dx_n.$$

Teoreem 6.3 *Olgu $X_1, X_2, \ldots$ iid juhuslikud suurused, X_i jaotus on pidev tihedusega f. Olgu $f \log f$ integreeruv ja $\epsilon > 0$. Siis leidub $n(\epsilon)$ nii, et iga $n > n(\epsilon)$ korral leidub hulk $W_\epsilon^n \subset \mathbb{R}^n$ nii, et*

1

$$P(W_\epsilon^n) > 1-\epsilon. \tag{6.4}$$

2

$$(1-\epsilon)2^{n(h-\epsilon)} \le \mathbf{V}(W_\epsilon^n) \le 2^{n(h+\epsilon)}. \tag{6.5}$$

3 *iga $x^n \in W_\epsilon^n$ korral*

$$2^{-n(h+\epsilon)} \le f(x^n) \le 2^{-n(h-\epsilon)}, \tag{6.6}$$

kus $h := h(f)$ ja $f(x^n) = f(x_1, \ldots, x_n) = f(x_1) \cdots f(x_n)$.

Tõestus. Tõestus on analoogiline diskreetse AEP omaduse tõestusega. Olgu

$$W_\epsilon^n := \{x^n \in \mathbb{R}^n : 2^{-n(h+\epsilon)} \le f(x^n) \le 2^{-n(h-\epsilon)}\}.$$

Suurte arvude seadusest järeldub, et

$$-\frac{\log f(X_1,\dots,X_n)}{n} \to -E\Big(\log f(X_1)\Big) = h(f), \quad \text{p.k.},$$

millest järeldub (6.6). Hinnangutest

$$1-\epsilon \le P(W_\epsilon^n) = \int_{W_\epsilon^n} f(x_1,\dots,x_n)dx_1\cdots dx_n \le 1$$

saame (6.5). ■

6.4 Ühisdiferentsiaalentroopia

Juhusliku vektori $(X_1,\dots,X_n)$ (ühis)diferentsiaalentroopia defineeritakse analoogiliselt diskreetse vektori entroopiaga.

Def 6.4 *Olgu $X^n = (X_1,\dots,X_n)$ pidev juhuslik vektor ühistihedusega f. Vektori X^n* **ühisdiferentsiaalentroopiaks (joint differential entropy)** *on*

$$h(X^n) = h(X_1,\dots,X_n) := -\int f(x^n)\log f(x^n)dx^n = -\int f(x_1,\dots,x_n)\log f(x_1,\dots,x_n)dx_1\cdots dx_n,$$

kui integraal eksisteerib.

Näide: Olgu $\phi(x^n)$ mitmemõõtmelise normaaljaotuse $N(\mu,\Sigma)$ tihedusfunktsioon,

$$\phi(x^n) = \frac{1}{(\sqrt{2\pi})^n|\Sigma|^{\frac12}}\exp[-\frac12(x^n-\mu)'\Sigma^{-1}(x^n-\mu)].$$

$$\begin{aligned}-\int_{-\infty}^{\infty}\phi(x^n)\ln\phi(x^n)dx^n &= \int_{-\infty}^{\infty}\frac12(x^n-\mu)'\Sigma^{-1}(x^n-\mu)\phi(x^n)dx^n + \ln[(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac12}]\\ &= \frac12 E\Big((X^n-\mu)'\Sigma^{-1}(X^n-\mu)\Big) + \frac12\ln[(2\pi)^n|\Sigma|],\end{aligned}$$

kus X^n on jaotusega ϕ juhuslik vektor. Et $\mathrm{tr}(AB)=\mathrm{tr}(BA)$, saame

$$(X^n-\mu)'\Sigma^{-1}(X^n-\mu) = \mathrm{tr}\big((X^n-\mu)'\Sigma^{-1}(X^n-\mu)\big) = \mathrm{tr}\big(\Sigma^{-1}(X^n-\mu)(X^n-\mu)'\big),$$

millest

$$\begin{aligned}E(X^n-\mu)'\Sigma^{-1}(X^n-\mu) &= E\mathrm{tr}\big((X^n-\mu)'\Sigma^{-1}(X^n-\mu)\big) = \mathrm{tr}\Big(E\big(\Sigma^{-1}(X^n-\mu)(X^n-\mu)'\big)\Big)\\ &= \mathrm{tr}\big(\Sigma^{-1}E(X^n-\mu)(X^n-\mu)'\big) = \mathrm{tr}(I_n) = n.\end{aligned}$$

Seega

$$-\int_{-\infty}^{\infty}\phi(x^n)\ln\phi(x^n)dx^n=\frac{1}{2}[n+\ln\big((2\pi)^n|\Sigma|\big)]=\frac{1}{2}[\ln e^n+\ln\big((2\pi)^n|\Sigma|\big)]=\frac{1}{2}\ln[(2\pi e)^n|\Sigma|].$$

Seega diferentsiaalentroopia on $\frac{1}{2}\ln[(2\pi e)^n|\Sigma|]$ natti ja

$$\frac{1}{2}\log[(2\pi e)^n|\Sigma|]$$

bitti.

Diferentsiaalentroopia omadused:

- Olgu X^n pidev juhuslik vektor, $\mu\in\mathbb{R}^n$. Siis $h(X^n+\mu)=h(X^n)$
- Olgu pidev juhuslik vektor, A olgu pööratav maatriks. Siis

  $$h(AX^n)=h(X^n)+\log|A|,$$

  kus $|A|$ on A determinandi absoluutväärtus.

Nende omaduste tõestus on ülesanne 2

6.5 Tinglik diferantsiaalentroopia, Kullback-Leibleri kaugus ja vastastikune informatsioon

Tinglik diferentsiaalentroopia. Tuletame meelde, et kui (X,Y) on tihedusega $f(x,y)$ juhuslik vektor, siis

$$f(x|y)=\frac{f(x,y)}{f(y)}$$

on juhusliku suuruse X tinglik tihedus. Siin $f(x)$ ja $f(y)$ on marginaaltihedused.

Def 6.5 *Olgu (X,Y) on tihedusega $f(x,y)$ juhuslik vektor.* **Tinglik diferentsiaalentroopia** *on*

$$h(X|Y)=-\int\int f(x|y)\log f(x|y)dxf(y)dy=-\int\int f(x,y)\log f(x|y)dxdy,$$

kui see integraal eksisteerib.

Analoogiliselt entroopiaga saame

$$\begin{aligned}h(X,Y)&=-\int\int f(x,y)\log f(x,y)dxdy=-\int\int f(x,y)\log\big(\frac{f(x,y)}{f(y)}f(y)\big)dxdy\\&=-\int\int f(x,y)\log f(x|y)dxdy-\int\int f(x,y)\log f(y)dxdy\\&=h(X|Y)+h(Y).\end{aligned}$$

Siit järeldub ketireegel

$$h(X_1,\dots,X_n)=h(X_1)+h(X_2|X_1)+\cdots+h(X_n|X_1,\dots,X_{n-1}).$$

Kullback-Leibleri kaugus.

Def 6.6 *Olgu f, g kaks tõenäosustihedust. Nende* **Kullback-Leibleri kaugus** *on*

$$D(f||g) := \int f(x) \log \frac{f(x)}{g(x)} dx.$$

Märkused:

1. Ülaltoodud definitisioonis, nagu ikka, $0 \log \frac{0}{0} := 0$.
2. Erinevalt diferentsiaalentroopiast on $D(f||g) \leq \infty$ on alati defineeritud (võib olla ∞). Tõestus on sama, mis diskreetsel juhul (kontrolli !)
3. Kui $D(f||g) < \infty$, siis tiheduse g kandja sisaldab f kandjat.

Lemma 6.1 (Gibbsi võrratus) *Iga kahe tiheduse f ja g korral*

$$D(f||g) \geq 0,$$

kusjuures $D(f||g) = 0$ parajasti siis, kui $f = g$ p.k.

Tõestus. Sama, mis diskreetsel juhul (kontrolli !) ■

Näide: Olgu $\{f_\theta : \theta \in \Theta\}$ tõenäosustiheduste pere. Olgu $\theta^* \in \Theta$ fikseeritud parameeter (õige parameeter) ja $X_1, \ldots, X_n$ iid valim jaotusest f_{θ^*} ja vaatame logaritmilist tõepärafunktsiooni

$$l_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \ln f_\theta(X_i).$$

Suurte arvude seaduse põhjal koondub $l_n(\theta)$ p.k. piirväärtuseks

$$\int \ln f_\theta(x) f_{\theta^*}(x) dx =: l(\theta),$$

mida nimetatakse **tõepärakontrastiks (likelihood contrast)**. Gibbsi võrratusest järeldub, et

$$0 \leq \int f_{\theta^*}(x) \ln \left(\frac{f_{\theta^*}(x)}{f_\theta(x)} \right) dx = \int f_{\theta^*}(x) \ln f_{\theta^*}(x) dx - \int f_{\theta^*}(x) \ln f_\theta(x) dx = l(\theta^*) - l(\theta).$$

Seega θ^* maksimiseerib tõepärakontrasti, st $l(\theta^*) \geq l(\theta)$ iga $\theta \in \Theta$ korral. Sellel asjaolul põhineb STP hinnangu mõjusus.

Vastastikune informatsioon.

Def 6.7 *Olgu (X, Y) juhuslik vektor ühistihedusega $f(x, y)$, marginaaltihedustega $f(x)$ ja $f(y)$. Juhuslike suuruste* **vastastikune informatsioon** *on*

$$I(X;Y) := D(f(x,y)||f(x)f(y)) = \int f(x,y) \log \frac{f(x,y)}{f(x)f(y)} dxdy.$$

Võrreldes diskreetse juhuga vastastikuse informatsiooni omadused ei muutu:

- Vastastikune informatsioon $I(X;Y)$ ei sõltu mitte ainult juhuslike suuruste X ja Y jaotusest vaid ka nende ühisjaotusest, s.t. vektori (X, Y) jaotusest.
- $0 \leq I(X;Y)$.
- Vastastikune informatsioon on sümmeetriline: $I(X;Y) = I(Y;X)$.
- $I(X;Y) = 0$ parajasti siis kui $f(x, y) = f(x)f(y)$, st X ja Y on sõltumatud.

Analoogiliselt diskreetse juhuga kehtib (kui $h(X|Y)$ ja $h(Y|X)$ on lõplikud)

$$I(X;Y) = h(X) - h(X|Y) = h(Y) - h(Y|X) \geq 0.$$

Ketireeglist saame

$$h(X_1, \ldots, X_n) \leq \sum_{i=1}^{n} h(X_i).$$

Mitmemõõtmelise normaaljaotuse korral saame ülaltoodud võrratuses nn. *Hadamardi võrratuse*

$$\frac{1}{2} \log[(2\pi e)^n |\Sigma|] \leq \sum_{i=1}^{n} \frac{1}{2} \log[(2\pi e)\sigma_i^2] \quad \Leftrightarrow \quad |\Sigma| \leq \prod_{i=1}^{n} \sigma_i^2. \tag{6.7}$$

6.6 MaxEnt printsiip

Vaatleme järgmist ülesannet: leida tundmatu jaotus P, kui on teada (valimi põhjal hinnatud):

- $\text{supp}(P) = S$ (kandja);
- $\int F_i dP = c_i$, $i = 1, \ldots, k$,

kus F_i on mingisugused funktsioonid (näiteks polünoomid) ja c_i on (harilikult valimi põhjal hinnatud) jaotuse P F_i-momendid.

Üks lähenemine antud ülesandele on *momentide meetod*, kus antud jaotuste hulgast (mudelist) valitakse hinnanguks (ainus) selline, mille F_i-momendid on c_i. Selline lähenemine eeldab aga mudeli olemasolu.

Maksimaalse entroopia printsiip: Kõikide ülaltoodud tingimusi rahuldavate jaotuste hulgast leida selline, mille (diferentsiaal)entroopia on maksimaalne. Sellist jaotust nimetatakse maksimaalse entroopiaga (MaxEnt) jaotuseks.

Juhul kui otsitav (hinnatav) jaotus on pidev (see, kas otsitav jaotus on pidev, diskreetne või midagi muud on harilikult selge ülesande püstitusest), saame järgmise optimiseerimisülesande:

Maksimaalse entroopia ülesanne pidevate jaotuste korral: maksimiseerida

$$h(f) = -\int f(x) \log f(x) dx$$

üle funktsioonide, mis rahuldavad tingimusi:

1) $f(x) \geq 0$, $f(x) = 0 \Leftrightarrow x \notin S$;

2) $\int_S f(x)dx = 1$;

3) $\int_S F_i(x)f(x)dx = c_i$, $i = 1, \ldots, k$.

Järgnev teoreem annab lihtsa eeskirja maksimaalse entroopiaga jaotuse leidmiseks. Tuletame meelde, et iga funktsiooni f ja hulga $S \subset \mathbb{R}$ korral

$$f(x)I_S(x) = \begin{cases} f(x), & \text{kui } x \in S; \\ 0, & \text{mujal.} \end{cases}$$

Teoreem 6.8 *Kui leiduvad konstandid $a_0, a_1, \ldots, a_k$ nii, et funktsioon*

$$f^*(x) = \exp[a_0 + \sum_{i=1}^{k} a_i F_i(x)] I_S(x) \tag{6.8}$$

rahuldab tingimusi **2), 3)**, *siis f^* on ainus (Lebesgue p.k.) maksimaalse entroopiaga tihedusfunktsioon.*

Tõestus. Olgu g suvaline tingimusi **1),2), 3)** rahuldav jaotus. Veendume, et $h_e(g) \leq h_e(f^*)$, kusjuures võrdus kehtib vaid siis, kui $g = f^*$ p.k.. Siis ka $h(g) \leq h(f^*)$ ja võrdus

kehtib vaid siis, kui tihedused on p.k. võrdsed.

$$\begin{aligned}
h_e(g) &= -\int_S g(x)\ln g(x)dx \\
&= -\int_S g(x)\ln\Big(f^*(x)\frac{g(x)}{f^*(x)}\Big)dx \\
&= -D_e(g||f^*) - \int_S g(x)\ln f^*(x)dx \\
&\leq -\int_S g(x)\ln f^*(x)dx \\
&= -\int_S (a_0 + \sum_{i=1}^{k} a_i F_i(x))g(x)dx \\
&= -(a_0 + \sum_{i=1}^{k} a_i c_i) \\
&= -\int_S (a_0 + \sum_{i=1}^{k} a_i F_i(x))f^*(x)dx \\
&= -\int_S f^*(x)\ln f^*(x)dx \\
&= h_e(f^*)
\end{aligned}$$

Võrdus $h_e(f^*) = h_e(g)$ kehtib parajasti siis, kui

$$D_e(g||f^*) = \int_S g(x)\ln\frac{g(x)}{f^*(x)}dx = 0.$$

Aga Gibbsi võrratusest teame, et see on nii vaid siis, kui $g = f^*$ p.k. ■

Märkused:

- Teoreem kehtib ka mitmemõõtmeliste jaotuste korral (sellisel juhul otsime maksimaalse ühisentroopiaga jaotust). Tõestus on sama.
- Kui kandja S on ülimalt loenduv hulk $\mathcal{X}$, otsime diskreetset jaotust. Asendades ülaltoodud tõestuses integreerimise summeerimisega, saame, et teoreem kehtib ka diskreetsete jaotuste korral. Seega diskreetsel juhul MaxEnt jaotus (kui leidub) on

$$P^*(x) = \exp[a_0 + \sum_{i=1}^{k} a_i F_i(x)], \tag{6.9}$$

  kus a_i on valitud nii, et P^* rahuldaks kitsendusi.

6.6.1 Näited

Keskväärtus ja dispersioon: Olgu $S = \mathbb{R}$, $F_1(x) = x$, $c_1 = 0$ ja $F_2(x) = x^2, c_2 = \sigma^2$. Otsime MaxEnt tihedust (üle reaaltelje) keskväärtusega 0 ja disp. σ^2 tiheduste seast. Jaotus (6.8) on kujul

$$\exp[a_0 + a_1 x + a_2 x^2].$$

Normaaljaotuse kuju; MaxEnt jaotus: $\mathcal{N}(0, \sigma^2)$.

Esimest ja teist järku moment: Olgu $S = \mathbb{R}$, $F_1(x) = x$, $c_1 = \mu$ ja $F_2(x) = x^2, c_2 = \alpha$. Jaotus (6.8) on kujul

$$\exp[a_0 + a_1 x + a_2 x^2].$$

Normaaljaotuse kuju; MaxEnt jaotus: $\mathcal{N}(\mu, \alpha - \mu^2)$

Keskväärtus: Olgu $S = \mathbb{R}$, $F_1(x) = x$, $c_1 = \mu$. Otsime MaxEnt tihedust (üle $\mathbb{R}$) keskväärtusega μ. Sellist pole.

Keskväärtus ning mittenegatiivsus: Olgu $S = [0, \infty)$, $F_1(x) = x$, $c_1 = \mu$. Otsime MaxEnt tihedust üle $[0, \infty)$ keskväärtusega μ. Jaotus (6.8):

$$\exp[a_0 + a_1 x] I_{[0,\infty)}.$$

Eksponentjaotuse kuju; MaxEnt jaotus: $E(\mu^{-1})$.

Tõkestatud kandja: Olgu $S = [a, b]$, tingimusi pole. Jaotus (6.8):

$$\exp[a_0] I_{[a,b]}.$$

Ühtlase jaotuse kuju; MaxEnt jaotus :$U(a, b)$.

Keskväärtus ja loenduv kandja: Olgu $S = \{1, 2, \ldots\}$ $F_1(x) = x$, $c_1 = \mu$. Jaotus (6.9):

$$P^*(x) = \exp[a_0 + a_1 x]$$

MaxEnt distribution: Geometric $G(\frac{1}{\mu})$.

Teine moment ja mittenegatiivsus: Olgu $S = [0, \infty)$ ja $F_1(x) = x^2$. Jaotus (6.8):

$$\exp[a_0 + a_1 x^2] I_{[0,\infty)}$$

Kui $c_1 = 1$, siis MaxEnt jaotus on

$$f(x) = \sqrt{\frac{2}{\pi}} \exp[-\frac{x^2}{2}], \quad x \geq 0 \tag{6.10}$$

Lõplik kandja: Olgu $S = \{1, 2, 3, 4, 5, 6\}$, tingimusi pole. Jaotus (6.9):

$$P^*(x) = \exp[a_0].$$

MaxEnt jaotus on ühtlane.

Etteantud segamomendid: Olgu $S = \mathbb{R}^n$, $F_{ij} = x_i x_j$, $c_{ij} = \sigma_{ij}$ $i, j = 1, \ldots, n$. Seega on etteantud segamomendid $EX_iX_j = \sigma_{ij}$. Jaotus (6.8):

$$f(x^n) = \exp[a_0 + \sum_{ij} a_{ij} x_i x_j].$$

Mitmemõõtmelise normaaljaotuse kuju; MaxEnt jaotus on $\mathcal{N}(0, \Sigma)$, kus $\Sigma = (\sigma_{ij})$.

6.7 Ülesanded

1. Tõestada teoreem 6.3.
2. Tõestada ühisdiferentsiaalentroopias omadused:
   - Olgu X^n pidev juhuslik vektor, $\mu \in \mathbb{R}^n$. Siis $h(X^n + \mu) = h(X^n)$
   - Olgu pidev juhuslik vektor, A olgu pööratav maatriks. Siis
   $$h(AX^n) = h(X^n) + \log |A|,$$
   kus $|A|$ on A determinandi absoluutväärtus.
3. Leida $h(f)$, kus $f(x) = \frac{1}{2}\lambda \exp[-\lambda|x|]$ (Laplace'i jaotus ehk kahepoolne eksponentjaotus).
4. Olgu $X \sim U(-\frac{1}{2}, \frac{1}{2})$, $Z \sim U(-\frac{a}{2}, \frac{a}{2})$, $a > 0$, X ja Z on sõltumatud, $Y = X + Z$. Leida $I(X; Y)$.
5. Olgu Π kõikide kõikide ruumil $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$ olevate korrutismõõtude hulk – **korrutismuutkond** (*product manifold*). Olgu (X, Y) juhusklik vektor ühistihedusega $f(x, y)$. Tõestada, et
   $$I(X; Y) = \inf_{g_1(x) \times g_2(y) \in \Pi} D(f(x, y) || g_1(x) \times g_2(y)).$$
   Miinimumi realiseerib vektori (X, Y) marginaaljaotuste korrutis.
6. Vaatleme diskreetsel tähestikul $\mathcal{X}$ antud tõenäosusjaotusi. Olgu $\mathcal{P}$ selliste jaotuste klass, mille korral
   $$\sum_j F_i(x_j) P(x_j) = c_i. \quad i = 1, \ldots, k.$$

Olgu Q suvaline jaotus. Tõestada, et kui leiduvad konstandid a_i, $i = 0, \ldots, k$ nii, et $P^* \in \mathcal{P}$, kus

$$P^*(x_j) = Q(x_j) \exp[a_0 + \sum_{i=1}^{k} a_i F_i(x_j)],$$

siis

$$P^* = \arg\min_{P \in \mathcal{P}} D(P||Q).$$

7. Olgu f_o suvaline tõenäosusjaotus. Tõestada. et leidub tingimus F ja konstant c (mis sõltuvad f_o-st) nii, et f_o on MaxEnt tihedus.