INFORMATSIOONITEOORA

Loengukonspekt

Kevad 2011

Jüri Lember

Kirjandus:

T.M. Cover, J.A. Thomas "Elements of information theory", Wiley, 1991 ja 2006.

Yeung, Raymond W. "A first course of information theory", Kluwer, 2002.

Mackay, D. "Information theory, inference and learning algorithms", Cambridge 2004

McEliece, R. "Information and coding", Cambridge 2004

Te Sun Han, Kingo Kobayashi "Mathematics of information and coding", AMS, 1994.

Gray, R. "Entropy and information theory", Springer 1990.

Gray, R. "Source coding theory", Kluwer, 1990.

Shields, P. "The ergodic theory of discrete sample paths", AMS 1996.

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)*.

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

$$Eg(X) = \sum p_i g(x_i). \tag{1}$$

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

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

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

Märkused:

- $H(X)$ sõltub vaid juhusliku suuruse X jaotusest P. Seetõttu tähistame entroopiat $H(X)$ ka $H(P)$.
- Seose (1) 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 $H(P) = H(Q)$.
- 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 kehtib
$$H(X) = -\sum_{x\in\mathcal{X}'} P(x)\log P(x), \tag{2}$$
kus $\mathcal{X}'$ on suvaline hulk, mis sisaldab hulka $\mathcal{X}$.

Entroopia $H(X)$ mõõdab juhusliku suuruse X "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 on rangelt nõgus

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 juhuslik suurus

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

on jaotusega Q.
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.

Tuletame meelde Jenseni võrratuse

Teoreem 1.2 (Jenseni võrratus).
Olgu g kumer funktsioon, kusjuures $E|g(X)| < \infty$ ja $E|X| < \infty$. Siis

$$Eg(X) \geq g(EX). \tag{3}$$

Kui g on rangelt kumer, siis (3) 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

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

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. ■

Väide 1.1 *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. Eelpool nägime, et binaarne entroopiafunktsioon on nõgus.

1.2 Ühisentroopia

Olgu X ja Y diskreetsed juhuslikud suurused väärtuste hulgaga vastavalt $\mathcal{X}$ ja $\mathcal{Y}$. Seega (X,Y) on diskreetne juhuslik vektor, mille väärtuste hulk sisaldub 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, $P(x) := \mathbf{P}(X = x)$ ja $P(y) := \mathbf{P}(Y = y)$ tähistavad marginaaltõenäosusi. 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

$$H(X,Y) = -\sum_{ij} p_{ij} \log p_{ij} = -\sum_{(x,y) \in \mathcal{X} \times Y} P(x,y) \log P(x,y) = E\Big(-\log P(X,Y)\Big). \quad (4)$$

Def 1.3 *Juhusliku vektori (X, Y) entroopiat (4) nimetatakse juhuslike suuruste X ja Y* **ühisentroopiaks**.

Kui juhuslikud suurused X, Y on sõltumatud, siis

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

Ülaltoodud argumandi 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

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

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 $X = x$ on

y_1	y_2	y_3	$\ldots$
$P(y_1\|x)$	$P(y_2\|x)$	$P(y_2\|x)$	$\ldots$

.

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

Def 1.4 *Juhusliku suuruse Y* **tinglik entroopia** *tingimusel X on*

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

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 1.2

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

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

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

Seega

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

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

Vaatleme juhusliku vektori (X,Y) ühisjaotust $P(x,y) = P(x)P(y|x)$. Olgu tinglik(ud) jaotus(ed) $P(y|x)$ fikseeritud. Sellisel juhul on vektori (X,Y) jaotus juhusliku suuruse X jaotuse P funktsioon. Järelikult on ka Y jaotus P funktsioon. Millest ka $H(Y)$ on jaotuse P funktsioon. Järgenevas näeme, et see funktsioon on nõgus.

Väide 1.3 *Fikseeritud $P(y|x)$ korral on $P \mapsto H(Y)$ nõgus funktsioon.*

Tõestus. Kehtib $P(y) = \sum_{x\in\mathcal{X}} P(x)P(y|x)$. Olgu $P = \lambda P_1 + (1-\lambda)P_2$. Nüüd

$$\begin{aligned} Q(y) :=& \sum_{x\in\mathcal{X}} \big(\lambda P_1(x) + (1-\lambda)P_2(x)\big)P(y|x) \\ =& \lambda \sum_{x\in\mathcal{X}} P_1(x)P(y|x) + (1-\lambda)\sum_{x\in\mathcal{X}} P_2(x)P(y|x) =: \lambda Q_1(y) + (1-\lambda)Q_2(y). \end{aligned}$$

Funktsioon H on nõgus, seega

$$H(Q) = H(\lambda Q_1 + (1-\lambda)Q_2) \geq \lambda H(Q_1) + (1-\lambda)H(Q_2).$$

■

Märkus: Funktsioon $P \mapsto H(Y)$ ei pruugi olla rangelt nõgus (ülesanne).

Vaatleme nüüd tinglikku entroopiat $H(Y|X)$.

Väide 1.4 *Fikseeritud $P(y|x)$ korral on $P \mapsto H(Y|X)$ lineaarne funktsioon. Fikseeritud $P(x)$ korral on $P(y|x) \mapsto H(Y|X)$ nõgus.*

Tõestus.

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

Fikseeritud $P(y|x)$ korral on $H(Y|X=x)$ fikseeritud ja esimene väide tõestatud.
Iga $x \in \mathcal{X}$ korral on $H(Y|X=x) = \sum_{y\in\mathcal{Y}} -\log P(y|x)\log P(y|x)$ kui $P(y|x)$ funktsioon rangelt nõgus. Seega

$$\begin{aligned} &-\sum_{y\in\mathcal{Y}} \big(\lambda P_1(y|x) + (1-\lambda)P_2(y|x)\big)\log\big(\lambda P_1(y|x) + (1-\lambda)P_2(y|x)\big) \geq \\ &-\lambda\sum_{y\in\mathcal{Y}} P_1(y|x)\log P_1(y|x) - (1-\lambda)\log P_2(y|x)P_2(y|x). \end{aligned}$$

Võrratus kehtib iga x korral. Korrutame läbi $P(x)$-ga ja summeerime üle $\mathcal{X}$. ■

1.3.2 Ketireeglid

Olgu X, Y, Z kolm juhuslikku suurust väärtuste hulgaga $\mathcal{X}, \mathcal{Y}$ 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 1.2 üldistub mitmes suunas. Alljärgnev on väite 1.2 tinglik versioon

Väide 1.5

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

Tõestus. Et $P(x,y|z) = P(x|z)P(y|x,z)$, siis

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

■

Väitest 1.5 järeldub väide 1.2. Ka järgmine lemma üldistab väidet 1.2.

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 väärtuste hulgad vastavalt $\mathcal{X}_1,\ldots,\mathcal{X}_n$. Et iga $x_1 \in \mathcal{X}_1,\ldots,x_n\in\mathcal{X}_n$ korral kehtib

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

siis

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

■

Kehtib ka ketireegli tinglik versioon.

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

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

Tõestus. Olgu juhuslike suuruste $X_1, \dots, X_n, Z$ väärtuste hulgad vastavalt $\mathcal{X}_1, \dots, \mathcal{X}_n$ ja $\mathcal{Z}$. Väide järeldub sellest, et iga $x_i \in \mathcal{X}_i$ ja $z \in \mathcal{Z}$ korral

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

■

Tinglikust ketireeglist järeldub nii väide 1.5 kui ka ketireegel.

1.4 Kullback-Leibleri kaugus

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

$$P: \begin{array}{c|c|c|c} x_1 & x_2 & x_3 & \dots \\ \hline p_1 & p_2 & p_3 & \dots \end{array} \quad Q: \begin{array}{c|c|c|c} x_1 & x_2 & x_3 & \dots \\ \hline q_1 & q_2 & q_3 & \dots \end{array},$$

kusjuures võib olla, et mõne i korral $q_i = 0$ või mõne j korral $p_i = 0$.

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** *on*

$$D(P||Q) := \sum_{x\in\mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}. \tag{5}$$

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 (5) 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 (5) negatiivne osa koonduv. Kui $\sum_{x\in\mathcal{X}^+} P(x) \log \frac{P(x)}{Q(x)} < \infty$, on rida (5) 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).
- Kullback-Leibleri kaugust nimetatakse veel *suhteliseks entroopiaks (relative entropy)* või *divergentsiks (divergence).*

Tõestame, et $D(P||Q) \geq 0$. Selleks kasutame Jenseni võrratust.

Väide 1.6 (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 (5) 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. ■

K-L kaugus mõõdab "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

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

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 $H(P) \le \log|\mathcal{X}|$, 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

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

■

Väide 1.7 (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*

$$\sum a_i \log\frac{a_i}{b_i} \ge \sum a_i \log\frac{\sum a_i}{\sum b_i}, \tag{6}$$

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 1.6 järeldub

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

Et

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

siis (6) 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.

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

$$D\big(\lambda P_1 + (1-\lambda)P_2 || \lambda Q_1 + (1-\lambda)Q_2\big) \leq \lambda D(P_1||Q_1) + (1-\lambda)D(P_2||Q_2). \tag{7}$$

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

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

Summeeri üle hulga $\mathcal{X}$. ■

Võrratust (105) võime interpreteerida: K-L kaugus on kumer paaride (P, Q) suhtes. Fikseeritud Q korral järeldub võrratusest (105), 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):

$$D(P||Q) = \sum P(x) \log P(x) - \sum P(x) \log Q(x) = -\sum P(x) \log Q(x) - H(P). \tag{8}$$

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 (8) järeldub ka, et $Q \mapsto D(P||Q)$ on rangelt kumer.

1.4.1 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 x korral 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 1.6 *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** *on*

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

kus $P_1(x,y) := P(x)P_1(y|x)$.

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

$$D\big(P_1(y|x)||P_2(y|x)\big) = E \log \frac{P_1(Y_1|X)}{P_2(Y_1|X)} =: D(Y_1||Y_2|X) \tag{9}$$

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

$$D\big(P_1(y|x)||P_2(y|x)\big) = E \log \frac{P_1(Y_1|X_1)}{P_2(Y_1|X_1)}. \tag{10}$$

   Võrduse (10) 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 1.10) korral.

Väide 1.8

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

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

Tõestus. Iga $x \in \mathcal{X}$ korral $D(P_1(y|x)||P_2(y|x)|x) \geq 0$, millest järelduvalt $D(P_1(y|x)||P_2(y|x)) \geq 0$. Meil $\mathcal{X}$ on X väärtuste hulk, s.t. $P(x) > 0$ iga $x \in \mathcal{X}$ korral. Oletame, et $D(P_1(y|x)||P_2(y|x)) = 0$. Siis $D(P_1(y|x)||P_2(y|x)|x) = 0$ iga $x \in \mathcal{X}$ korral, millest järeldub väide. ■

Väide 1.9 (Tingimustamine suurendab K-L kaugust)

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

kus $P_i(y) = \sum_y P_i(y|x)P(x)$, *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 1.10 (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 1.7 *Juhuslike suuruste* X, Y **vastastikune informatsioon** *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.
- $0 \leq I(X;Y)$.
- 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

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X). \tag{11}$$

Suurus $H(X)$ on juhusliku suuruse X "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 (11) järeldub, et täpselt sama palju kahaneb $H(Y)$ juhusliku suuruse X läbi. Sellest ka nimetus: vastastikune informatsioon (*mutual information*). 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 1.2: $H(X|Y) = H(X,Y) - H(Y)$, millest

$$I(X;Y) = H(X) + H(Y) - H(X,Y). \tag{12}$$

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. ■

Lemma 1.3
1. *Fikseeritud $P(y|x)$ korral on $P(x) \mapsto I(Y;X)$ nõgus.*
2. *Fikseeritud $P(x)$ korral on $P(y|x) \mapsto I(Y;X)$ kumer.*

Tõestus.
1. $I(X;Y) = H(Y) - H(Y|X)$. Väide 1.3: fikseeritud $P(y|x)$ korral on $P(x) \mapsto H(Y)$ nõgus. Väide 1.4: fikseeritud $P(y|x)$ korral on $P(x) \mapsto H(Y|X)$ lineaarne. Vahe on nõgus.

2. Olgu $P(x)$ fikseeritud, $P_1(y|x)$ ja $P_2(y|x)$ olgu kaks tinglikku jaotust (formaalselt 2 tingliku jaotuse pere). Olgu $P_1(x,y) = P_1(y|x)P(x)$ ja $P_2(x,y) = P_2(y|x)P(x)$. Vastavad marginaalid olgu $P_1(y)$ ja $P_2(y)$.
Vaatleme kumerat kombinatsiooni

$$\lambda P_1(y|x) + (1-\lambda)P_2(y|x).$$

Olgu

$$Q(x,y) := (\lambda P_1(y|x) + (1-\lambda)P_2(y|x))P(x) = \lambda P_1(x,y) + (1-\lambda)P_2(x,y)$$

sellele kombinatsioonile vastav ühisjaotus ning olgu

$$Q(y) := \sum_x Q(x,y) = \sum_x \lambda(P_1(x,y) + (1-\lambda)P_2(x,y)) = \lambda P_1(y) + (1-\lambda)P_2(y).$$

selle ühisjaotuse marginaaljaotus. Paneme tähele, et

$$Q(y)P(x) = \lambda P_1(y)P(x) + (1-\lambda)P_2(y)P(x).$$

Eesmärk on näidata, et

$$D\big(Q(x,y)||Q(y)P(x)\big) \le \lambda D\big(P_1(y,x)||P_1(y)P(x)\big) + (1-\lambda)D\big(P_2(y,x)||P_2(y)P(x)\big).$$

Järeldub järeldusest 1.2 sest

$$D\Big(Q(x,y)\Big\|Q(y)P(x)\Big) = D\Big(\lambda P_1(x,y)+(1-\lambda)P_2(x,y)\Big\|\lambda P_1(y)P(x)+(1-\lambda)P_2(y)P(x)\Big).$$

■

1.5.1 Tinglik vastastikune informatsioon

Olgu X, Y, Z juhuslikud suurused, olgu $\mathcal{Z}$ juhusliku suuruse Z väärtuste hulk.

Def 1.8 *Juhuslike suuruste X, Y vastastikune informatsioon tingimusel Z on*

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

Väide 1.11

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

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

$$P(x,y|z) = P(x|z)P(y|z), \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}, z \in \mathcal{Z}. \tag{13}$$

Tõestus.

$$\begin{aligned} \sum_{x,y,z} P(x,y,z)\log\frac{P(x,y|z)}{P(x|z)P(y|z)} &= \sum_z \Big(\sum_{x,y} P(x,y|z)\log\frac{P(x,y|z)}{P(x|z)P(y|z)}\Big)P(z) \\ &= \sum_z D\Big(P(x,y|z)||P(x|z)P(y|z)\Big)P(z) \ge 0. \end{aligned}$$

Kui võrdus kehtib, siis iga $z \in \mathcal{Z}$ korral (tuletame meelde, et $P(z) > 0$ iga $z \in \mathcal{Z}$ korral)

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

millest järeldub (13). ■

Tinglikul vastastikusel informatsioonil on üldiselt samad omadused mis vastastikusel informatsioonil. Kehtib (ülesanne 14)

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

Lisaks kehtib veel (ülesanne 14)

$$I(X;Y|Z) = H(X;Z) + H(Y;Z) - H(X,Y,Z) - H(Z). \tag{14}$$

Väide 1.12 (Vastastikuse informatsiooni ketireegel)

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

Tõestus. Kasutame entroopia ketireeglit ja tingliku entroopia ketireeglit.

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

■

Väide 1.13 (Tingliku vastastikuse informatsiooni ketireegel)

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

Tõestus. Analoogiline. ■

1.6 Andmetöötlusvõrratus

1.6.1 Lõplik Markovi ahel

Def 1.9 *Juhuslikud suurused $X_1,\ldots,X_n$ väärtuste hulkadega vastavalt $\mathcal{X}_1,\ldots,\mathcal{X}_m$ moodustavad* **Markovi ahela** *kui iga $x_i \in \mathcal{X}_i$ ja iga $m = 2,\ldots,n-1$ korral*

$$\mathbf{P}(X_{m+1} = x_{m+1}|X_m = x_m,\ldots,X_1 = x_1) = \mathbf{P}(X_{m+1} = x_{m+1}|X_m = x_m). \tag{15}$$

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

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

Asjaolu, et $X_1, \ldots, X_n$ 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 1.14 *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

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

See on aga sümmeetriline. ■

Väide 1.15 *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

$$\mathbf{P}(A|B) = \sum_i \mathbf{P}(A|B, C_i)\mathbf{P}(C_i|B). \tag{16}$$

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 (16) 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 $k \geq 3$ korral

$$P(x_{m+k}|x_{m+k-1}, \ldots, x_{m+2}, x_m, x_{m-1} \ldots, x_1) = P(x_{m+k}|x_{m+k-1}) \tag{17}$$

[Seosest (17) järeldub $P(x_{m+2}|x_{m+1},x_m)=P(x_{m+2}|x_{m+1})$ (kuidas?)].
Seega

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

Seega

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

Viimasest võrdusest ja seosest (17) järeldub, et $X_1,\dots,X_m,X_{m+2},\dots X_n$ on Markovi ahel. Siit järeldub ülejäänu. ■

Tõestusest näeme, et kui $X_1\to X_2\to\cdots\to X_n$, siis

$$P(x_n,\dots,x_{m+1}|x_m,\dots,x_1)=P(x_n,\dots,x_{m+1}|x_m). \tag{18}$$

Tõepoolest, kui $X_1\to X_2\to\cdots\to X_n$ on Markovi ahel, siis Väite 1.15 korral on seda ka $X_k\to\cdots\to X_n$ ($k\geq 1$), millest iga $m>k$ korral

$$P(x_m|x_{m-1},\dots,x_k)=P(x_m|x_{m-1}) \tag{19}$$

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

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

Siin eelviimane võrdus tuleneb seosest (19). Edasi jätka induktsiooniga.

Väide 1.16 *Juhuslikud suurused $X_1,\dots,X_n$ on Markovi ahel parajsti siis, kui iga $m=2,\dots,n-1$ korral $X_1,\dots,X_{m-1}$ ja $X_{m+1},\dots,X_n$ on antud X_m korral tinglikult sõltumatud.*

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

$$P(x_1,\dots,x_{m-1},x_{m+1},\dots,x_n|x_m)=P(x_1,\dots,x_{m-1}|x_m)P(x_{m+1},\dots,x_n|x_m). \tag{20}$$

Seosest (18) saame

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

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 (20). 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 1.4 (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

$$I(X;Y,Z) = I(X;Z) + I(X;Y|Z) = I(X;Y) + I(X;Z|Y) = I(X;Y). \tag{21}$$

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 1.5 *Kui $X \to Y \to Z$, siis*

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

Tõestus. Ülesanne. ■

Järeldus 1.6 *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. ■

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

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

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 1.4 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 1.7 *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 1.10 (Fano võrratus) *Olgu X ja $\hat{X}$ juhuslikud suurused tähestikul $\mathcal{X}$. Siis*

$$H(X|\hat{X}) \leq h(P_e) + P_e \log(|\mathcal{X}| - 1), \tag{22}$$

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

$$H(E, X|\hat{X}) = H(X|\hat{X}) + H(E|X, \hat{X}) = H(X|\hat{X}), \tag{23}$$

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 (23) saame, et

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

■

Järeldus 1.8

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

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

$$H(X|\hat{X} = x, E = 1) = \log(|\mathcal{X}| - 1) \tag{24}$$

ning

$$H(E|\hat{X}) = H(E). \tag{25}$$

Seos (24) 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 (25) 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 1.11 *Juhusliku protsessi* $\{X_n\}_{n=1}^{\infty}$ **entroopiamäär** *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 1.12 *Juhuslik protsess* $\{X_n\}_{n=1}^{\infty}$ *on* **statsionaarne**, *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 1.17 *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 1.5 (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 1.13 *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 1.14 *Juhuslik protsess* $\{X_n\}_{n=1}^{\infty}$ *on* **Markovi ahel**, *kui iga* $x_i \in \mathcal{X}$ *ja iga* $m \geq 1$ *korral kehtib (15), s.t.*

$$\mathbf{P}(X_{m+1} = x_{m+1}|X_m = x_m,\ldots,X_1 = x_1) = \mathbf{P}(X_{m+1} = x_{m+1}|X_m = x_m). \qquad (26)$$

Märkus: Arusaadavalt on võrdus (26) 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 (26) 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 1.15 *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 1.13 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 1.18 *Iga $n = 1, 2, \ldots$ korral kehtib*

$$D(X_{n+1}||X'_{n+1}) \le D(X_n||X'_n). \tag{27}$$

Tõestus. Olgu P_n ja P'_n vastavalt X_n ja X'_n jaotused. Seega (27) on

$$D(P_{n+1}||P'_{n+1}) \le D(P_n||P'_n). \tag{28}$$

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 1.9 *Kui π' on statsionaarne algjaotus, siis (27) on*

$$D(P_{n+1}||\pi') \leq D(P_n||\pi'). \tag{29}$$

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 1.10 *Kui statsionaarne algjaotus π' on ühtlane üle lõpliku tähestiku $\mathcal{X}$, siis (29) on*

$$H(P_n) \leq H(P_{n+1}) \tag{30}$$

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*

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

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. 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

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

10. 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 $H(Z) = H(X) + H(Y)$?

11. Olgu

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

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

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

12. 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

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

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

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

14. 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 $D\big((X, f(X))||(Y, g(Y))\big) \neq D(X||Y)$.

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

$$\rho(X_1, X_2) := 1 - \frac{H(X_2|X_1)}{H(X_1)}. \tag{31}$$

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 (31)). 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 (31)). Leida $\mathrm{cov}(X, Y)$ ja X ning Y korrelatsioonikordaja.

16. Tõestada, et

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

17. 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?

18. Leida X, Y, Z nii, et

$$\begin{aligned} I(X;Y|Z) > I(X;Y) = 0 \\ 0 = I(X;Y|Z) < I(X;Y). \end{aligned}$$

19. 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.

20. 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

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

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

22. 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.

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

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

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

25. 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.

26. 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

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

27. Olgu P jaotus väärtuste hulgaga $1, 2, \ldots$. Olgu selle mõõdu keskväärtus μ. Tõestada, et

$$H(P) \leq \mu \log \mu + (1 - \mu) \log(1 - \mu),$$

kusjuures võrratus on võrdus parajasti siis, kui P on geomeetrilise jaotusega. Seega fikseeritud keskväärtuse korral on geomeetriline jaotus suurima entroopiaga.

28. **a)** Tõestada Järeldus 1.10

b) Olgu $X_1 \to \cdots \to X_n$. Tõestada, et

$$H(X_0|X_1) \leq H(X_0|X_2) \leq H(X_0|X_3) \leq \cdots \leq H(X_0|X_n).$$

29. Olgu $\{X_n\}_{n=1}^{\infty}$ statsionaarne juhuslik protsess. Tõestada, et

$$\frac{H(X_1, \ldots, X_n)}{n} \leq \frac{H(X_1, \ldots, X_{n-1})}{n-1}$$
$$\frac{H(X_1, \ldots, X_n)}{n} \geq H(X_n|X_1, \ldots, X_{n-1}).$$

30. Tõestada, et statsionaarse MA korral

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

31. 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 $H_S = H(P)$.

32. 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.

33. 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 *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**, *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**, *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**, *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*. Tihti kasutatakse ka terminit *instantaneous code*. Üheselt dekodeeritav kood on inglisekeelses kirjanduses *uniquely decodable*, ühene kood aga *non-singular*.

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.

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 koodi

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

puu.

2.2 Krafti võrratus

Vaatleme olukorda, kus tähed on juhuslikud, tähe $x \in \mathcal{X}$ tõenäosus on $P(x)$. Seega on tähestikul $\mathcal{X}$ antud mingi tõenäosusjaotus P. Olgu C mingi kood ning olgu $l(x) := |C(x)|$, s.t. $l(x)$ on tähe x koodi pikkus. 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*. Alljärgnevas otsime koodi, mille keskmine pikkus oleks võimalikult väike, sest sellise koodi korral on (antud jaotusega) juhusliku tähe kodeerimine efektiivne. Seejuures on oluline teada, milline on (fikseeritud jaotuse korral) väikseim võimalik keskmine koodipikkus.

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 $H(P) = \frac{7}{4}$.

Otsime minimaalse keskmise pikkusega prefikskoodi. On selge, et keskmine pikkus on seda väiksem, mida lühemad on koodisõnad $C(x)$. Samas on ka selge, et prefikskoodi puhul ei saa koodisõnad olla kuitahes lühikesed. Kui $|\mathcal{X}| \geq 3$, ei saa leida ühest kahendkoodi nii, et $l(x) = 1\ \forall x \in \mathcal{X}$. Alljärgnev teoreem väidab, et 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*

$$\sum_i D^{-l_i} \leq 1. \tag{32}$$

Teistpidi, olgu $\{l_i\}_{i=1}^{|\mathcal{X}|}$ täisarvud. Kui nad rahuldavad võrratust (32), siis leidub prefikskood $C : \mathcal{X} \to \mathcal{D}^*$ nii, et $l_i = l(x_i)\ \forall x_i \in \mathcal{X}$.

Tõestus. Olgu $\mathcal{D} = \{0, \ldots, D-1\}$. Vaatleme koodisõna $d_1d_2\cdots d_{l_i}$. Olgu $0.d_1d_2\cdots d_{l_i}$ reaalarv, millele vastav D-ndarv on $0.d_1d_2\cdots d_{l_i}$, s.t.

$$0.d_1d_2\cdots d_{l_i} = \sum_{j=1}^{l_i} \frac{d_j}{D^j}. \tag{33}$$

Vaatleme koodisõnale $d_1d_2\cdots d_{l_i}$ vastavat intervalli

$$[0.d_1d_2\cdots d_{l_i}, 0.d_1d_2\cdots d_{l_i} + D^{-l_i}).$$

Siia intervalli kuuluvad need reaalarvud, millele vastavad D-ndarvud algavad $0.d_1d_2\cdots d_{l_i}$. 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 (32).

Olgu $\{l_i\}_{i=1}^{|\mathcal{X}|}$ tingimust (32) rahuldavad täisarvud. Sellisel juhul saab ühikintervalli jagada lõikudeks pikkustega D^{-l_i}. Tõepoolest, reastame arvud l_i nii, et $l_1 \le l_2 \le \cdots$. 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 – alguspunkti esitame kujul $0.d_1d_2\ldots d_{l_1} := 0.0\cdots 0$, kus koma järel on l_1 nulli. Selle intervalli lõpp-punkti D^{-l_1} esitus D-ndarvuna on $0.0\ldots 1$, kus peale komakohta on l_1 arvu ($l_1 - 1$ nulli ja üks 1). Intervalli $[0.0\ldots 0, 0.0\ldots 1)$ kuuluvad parajasti need D-ndarvud, mille algus on $0.0\ldots 0$. Järgmise intervalli – arvule l_2 vastava intervalli – alguspunkti esitame D-ndarvuna $0.d_1d_2\cdots d_{l_2}$. Et $l_2 \ge l_1$, on $d_{l_1} = 1$ ja $d_i = 0$, kui $i > l_1$. Selle intervalli lõpp-punkti esitame l_2-kohalise D-ndarvuna. Teise intervalli $[D^{-l_1}, D^{-l_1} + D^{-l_2})$ kuuluvad parajasti need arvud, mille D-nd esitus algab arvuga $0.d_1d_2\cdots d_{l_2}$. 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 $l_3 \ge l_2$ 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 $l_i \ge l_{i-1}$, saame (vajaduse korral 0-de lisamisel) selle arvu esitada kujul D-ndkujul (105). 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: Alternatiivse tõestuse esimesel väitele (iga prefikskood rahuladab tingimust (32)) võib leida *Thomas ja Cover'i* raamatust, samuti *Yeung'i* raamatust. Edaspidi tõestame, et sama väide üldistud üheselt dekodeeritavate koodideni (Teoreem 2.11).
alternatiivse tõestuse teisele implikatsioonile võib leida *Yeung*'i raamatust (Thm 3.1).

Näide: 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

$$0.0_2 = 0, \quad 0.10_2 = 0.1_2 = 0.5, \quad 0.110_2 = 0.11_2 = \frac{1}{2}+\frac{1}{4} = 0.75, \quad 0.111_2 = \frac{1}{2}+\frac{1}{4}+\frac{1}{8} = 0.975.$$

Vastavad intervallid on

$$[0, 0+\frac{1}{2}), [0.5, 0.5+\frac{1}{4}), [0.75, 0.75+\frac{1}{8}), [0.975, 0.975+\frac{1}{8}).$$

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

$$[0.0, 0.1), [0.10, 0.11), [0.110, 0.111), [0.111, 1).$$

Koodisõnad on seega:

$$0 \quad 10 \quad 110 \quad 111.$$

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

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

Kood: 00 01, 100, 101.

2.3 Keskmine pikkus ja entroopia, Shannon-Fano kood

2.3.1 Shannon-Fano kood

Otsime minimaalse keskmise pikkusega prefikskoodi. 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 $C : \mathcal{X} \to \mathcal{D}^*$ kood. 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. ■

Seega, kui jaotus P on selline, et

$$\log_D \frac{1}{P(x)} \in \mathbb{Z}, \quad \forall x \in \mathcal{X}, \tag{34}$$

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 (34) 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 sü-

napikkustele 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 (105)

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

Vastav kood on seega

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

Paraku ei rahulda kõik tõenäosusjaotused seost (34) 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.

$$l(x) = \lceil \log_D \frac{1}{P(x)} \rceil. \tag{35}$$

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 (35).

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*

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

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 on seega näiteks järgmine

$$C(x_1) = 000, \quad C(x_2) = 001, \quad C(x_3) = 010, \quad C(x_4) = 011, \quad C(x_5) = 110. \tag{36}$$

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 $\frac{12}{5} = 2.4$.

2.3.2 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

$$H(P) + D(P\|Q) \leq \sum_x l_Q(x)P(x) < H(P) + D(P\|Q) + 1. \tag{37}$$

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. ■

2.4 Huffmani kood

2.4.1 Huffmani koodi konstrueerimine

Shannon-Fano meetod andis üsna hea keskmise pikkusega prefikskoodi; kui jaotus P rahuldab seost (34), 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 and järgmisel tasemel (eel-eelviimasel tasemel). 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. Saadud kood C on järgmine:

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

- 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 on seega:

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

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

$$P(x_1) \geq P(x_2) \geq \cdots \geq P(x_m). \tag{38}$$

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 iga optimaalse koodi korral leidub järjestus (38) nii, et

$$l(x_1) \leq l(x_2) \leq \cdots \leq l(x_m). \tag{39}$$

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

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. Üldisust kitsendamata eeldame, et $C(x_m)$ on pikim (võrratused (39)). 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 (39) saame, et $l(x_j) = l(x_{m-1}) = l(x_m)$, mistõttu võime optimaalsust rikkumata sõnad $C(x_j)$ ja $C(x_{m-1})$ ära vahetada. ■

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 (39) 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 (ülesanne).

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), 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

$$p_m \geq \sum_{i>m} p_i =: p_m^*. \tag{40}$$

  Kujutagem korraks ette, et tähestikus on lõplik arv (kuid väga palju) tähti. Olgu $p_{m_1}, p_{m_2}, \ldots$ tingimust (40) rahuldavad aatomid. Et p_{m_1} rahuldab tingimust (40), 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

$$p_1, p_2, \ldots, p_{m_1}, p_{m_1}^*. \tag{41}$$

  Et jaotuseni (41) jõutakse suvaliste aatomite p_j, $j > m_1$ korral (kui vaid nende summa on $p_{m_1}^*$), siis võib lõpmatu koodi konstrueerimist alustada jaotusele (41) vastava kahendpuu konstrueerimisest. Edasi asume konstrueerime alampuud, mis väljub sõlmest $p_{m_1}^*$. Selleks vaatleme jaotust, mis on proportsionaalne vektoriga

$$p_{m_1+1}, p_{m_1+2}, \ldots, p_{m_2}, p_{m_2}^*. \tag{42}$$

  Arvud (42) 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 (42) vastava alampuu konstrueerimisest. Edasi alustame sõlmele $p^*_{m_2}$ vastava alampuu konstrueerimist. Selleks vaatleme aatomeid

$$p_{m_2+1}, p_{m_2+2}, \ldots, p_{m_3}, p^*_{m_3}, \tag{43}$$

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 (40) kehtib iga m korral (ülesanne).

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*

$$\sum_x D^{-l(x)} \leq 1. \tag{44}$$

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

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

Seega

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

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 $a(m) \leq D^m$. Seega

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

ehk

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

Võrratuse vasak pool ei sõltu k-st. Järelikult

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

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 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.

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

■

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

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

Järgnev teoreem aga väidab, et optimaalse Shannon-Fano koodi korral (tuletame meelde, et see saab olla vaid siis, kui P rahuldab seost (34)) kehtib võrratus $\mathbf{P}(l'(X) < l(X)) \le \mathbf{P}(l(X) < l'(X))$. 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 (34). Olgu $l(x) = \log_D \frac{1}{P(x)}$ ning olgu $\{l'(x)\}$ mingi üheselt dekodeeritava koodi kodipikkused. Siis*

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

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

$$\text{sign}(a) \le D^a - 1.$$

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

Võrratus on võrdus, kui iga x korral kehtib

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

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

2.7 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 (34) 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 (34) 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 (34)? 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 (34) 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

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

2.8 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:

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

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) $L(C^k) = kL(C)$, millest

$$L_k(C^k) = L(C). \tag{45}$$

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

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

kusjuures parempoolne võrratus võib olla kuitahes täpne.

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

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

millest

$$\frac{H_D(X_1, \ldots, X_k)}{k} \le L_k \le \frac{H_D(X_1, \ldots, X_k)}{k} + \frac{1}{k}. \tag{46}$$

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 (46) saame

$$H_D(P) \le L_k < H_D(P) + \frac{1}{k}. \tag{47}$$

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.

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

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

(Kui $D > 2$, defineerime entroopiamäära analoogiliselt. Meil on D fikseeritud, mistõttu jätame ta tähistusest välja.) Seosest (46) saame, et

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

Seos (2.8) annab entroopiamäärale sisu: H_X on protsessi kodeerimise keskmine pikkus tähe kohta. Kui $X = X_1, X_2, \ldots$ on i.i.d., siis keskmiselt kulub tähe kohta $H_D(P)$ ühikut. 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 on väiksem kui ühte tähte eraldi kodeerides.

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

2.8.1 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 teoreemi 32 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.

Alustame lemmast.

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

$$|E(n)| = \log_D n + o(\log_D n) \tag{48}$$

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. Oletamem 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 $n = m$.

■

Iga naturaalarvude prefiksoodi, mille koodipikkused on kujul (48) nimetatakse **Eliase koodiks**.

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

$$E(12) = u(12)v(12)w(12) = 0001001100, \quad C_k^*(12) = 000100110000101100111.$$

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.

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.

2.9 Ülesanded

1. 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.

2. 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?

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

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

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

5. 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.

6. 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 $L_D(p_1, \ldots, p_m)$.

7. 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.

8. 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.

9. **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 (sõltub üldiselt valitud 4-ndkoodist). Tõestada, et

$$L_2(P) \leq L_T \leq L_2(P) + 1.$$

   d) Näita, et ülaltoodud võrratused võivad olla võrdused.

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

$$\min_{l_1,\ldots,l_m} \sum_{i=1}^{m} u_i l_i$$
$$\text{nii, et } \sum_{i=1}^{m} D^{-l_i} \leq 1.$$

11. 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.

12. 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 $L(C^k) = kL(C)$.

13. 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: vaatle seisunit Y_n. Vali sellele seisundile vastav kood (näiteks, kui $Y_n = 3$, siis C_3) ja kodeeri selle koodiga seisund Y_{n+1}. Siis vali seisundile Y_{n+1} vastav kood ja kodeeri sellega Y_{n+2} jne. Kas leiduvad koodid C_1, C_2, C_3 nii, et kirjeldatud kodeerimisprotseduur annaks keskmiseks pikkuseks tähe kohta H_Y?

14. 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)
$$C(x) = \begin{cases} 00, & \text{kui } x = a; \\ 1, & \text{kui } x = b; \\ 01, & \text{kui } x = c. \end{cases}$$

15. 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 (*almost equipartition property*).

Paneme tähele, et nõrgast suurte arvude seadusest järeldub koondumine

$$-\frac{1}{n} \log P(X_1, \ldots, X_n) = -\frac{1}{n} \sum_{i=1}^{n} \log P(X_i) \xrightarrow{P} -E \log P(X_1) = H. \tag{49}$$

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*

$$2^{-n(H+\epsilon)} \leq P(x_1, \ldots, x_n) \leq 2^{-n(H-\epsilon)}. \tag{50}$$

Tingimust (50) rahuldavaid sõnu nimetame **nõrgalt tüüpilisteks**.

Teoreem 3.2 (Nõrk AEP) *Iga $\epsilon > 0$ korral*

1 *Kui $x^n \in W_\epsilon^n$, siis*

$$2^{-n(H+\epsilon)} \leq P(x^n) \leq 2^{-n(H-\epsilon)}. \tag{51}$$

2 *Piisavalt suure n korral*

$$P(W_\epsilon^n) > 1 - \epsilon. \tag{52}$$

3 *Piisavalt suure n korral*

$$(1-\epsilon)2^{n(H-\epsilon)} \leq |W_\epsilon^n| \leq 2^{n(H+\epsilon)}. \tag{53}$$

Tõestus. Omadus **1** järeldub vahetult definitsioonist (50).

Omadus **2** järeldub vahetult koondumisest (49), sest tõenäosuse järgi koondumis definitsioonist johtuvalt $\forall \epsilon > 0$ korral leidub n_o nii, et

$$\mathbf{P}\Big(\Big| -\frac{1}{n}\sum_{i=1}^{n} \log P(X_i) - H \Big| \le \epsilon \Big) \ge 1-\epsilon, \tag{54}$$

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 tüüpilne 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. Olgu $X_1, \ldots, X_n$ iid juhuslikud suurused lõplikul tähestikul $\mathcal{X}$. 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 nõrgalt tüüpiliste sõnade indeksite kodeerimiseks (kahendkujul) kulub ülimalt $n(H+\epsilon)+1$ bitti. Liidame nendele sõnadele prefiksi 0 (näitab kuulumist nõrgalt tüüpiliste sõnade hulka). Nii kulub iga nõrgalt tüüpilise sõna kodeerimiseks $n(H+\epsilon)+2$ bitti. Et ülejäänud sõnu on vähem kui $2^{n \log |\mathcal{X}|}$, kulub nende sõnade kodeerimiseks $n \log |\mathcal{X}| + 1$ bitti. Lisades prefiksi 1, saame hulka W_ϵ^n mittekuuluvate sõnade koodi. 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

$$\begin{aligned} L &= \sum_{x^n \in \mathcal{X}^n} l(x^n)P(x^n) = \sum_{x^n \in W_\epsilon^n} l(x^n)P(x^n) + \sum_{x^n \notin W_\epsilon^n} l(x^n)P(x^n) \\ &= \sum_{x^n \in W_\epsilon^n} (n(H+\epsilon)+2)P(x^n) + \sum_{x^n \notin W_\epsilon^n} (n \log |\mathcal{X}| + 2)P(x^n) \\ &= P(W_\epsilon^n)(n(H+\epsilon)+2) + (1 - P(W_\epsilon^n))(n \log |\mathcal{X}| + 2) \\ &\leq n(H+\epsilon) + \epsilon(n \log |\mathcal{X}|) + 2 \\ &= n(H+\epsilon'), \end{aligned}$$

kus $\epsilon' = \epsilon + \epsilon \log |\mathcal{X}| + \frac{2}{n}$ ja selle võib (sobiva ϵ ja n valikul) teha kuitahes väikeseks. Oleme tõestanud, et iga $\epsilon > 0$ korral leidub n ja prefikskood $C : \mathcal{X}^n \to \{0,1\}^*$ nii, et

$$H \leq L_n(C) < H + \epsilon. \tag{55}$$

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 (55) 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 $|W_\epsilon^n| \approx 2^{nH}$. Veendume selles.

Lemma 3.1 *Iga $1 > \epsilon > 0$ ja $\delta > 0$ korral leidub n nii suur, et*

$$|B_\epsilon^n| \geq 2^{n(H-\delta)} \tag{56}$$

Tõestus. Valime $\epsilon_1 > 0$ nii väikese, et $\epsilon_1 < \delta$ ja $\epsilon_1 + \epsilon < 1$. Olgu n nii suur, et

$$P(W^n_{\epsilon_1}) > 1 - \epsilon_1. \tag{57}$$

(sellise n olemasolu järeldub Teoreemist 3.2) ning lisaks kehtib

$$\epsilon_1 - \frac{\log(1-(\epsilon+\epsilon_1))}{n} < \delta. \tag{58}$$

Defineerime

$$S := W^n_{\epsilon_1} \cap B^n_\epsilon.$$

Siis

$$1 - (\epsilon_1 + \epsilon) \le P(S) = \sum_{x^n \in S} P(x^n) \le |S| 2^{-n(H-\epsilon_1)} \le |B^n_\epsilon| 2^{-n(H-\epsilon_1)},$$

kus esimene võrratus järeldub B^n_ϵ definitsioonist ja võrratusest (57) ning teine võrratus järeldub $W^n_{\epsilon_1}$ definitsioonist. Seega

$$\log |B^n_\epsilon| \ge \log(1-(\epsilon+\epsilon_1)) + n(H-\epsilon_1) = n\big(\frac{\log(1-(\epsilon+\epsilon_1))}{n} + H - \epsilon_1\big) \ge n(H-\delta).$$

Viimane võrratus tuleb seosest (58). ■

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 $h(0.1) = 468996$, ja võttes $\epsilon = 0.2$, same, et hulka $W^{25}_{0.2}$ kuuluvad klasside $k = 1, 2, 3, 4$ elemendid. Seega

$$P(W^{25}_{0.2}) = 0.199416 + 0.265888 + 0.226497 + 0.138415 = 0.830216 \ge 1 - \epsilon.$$

Samas $|W^{25}_{0.2}| = 25 + 300 + 2300 + 12650 = 15275$, millest

$$\frac{1}{25}\log |W^{25}_{0.2}| \approx 0.556 \in (468996 - 0.2, 468996 + 0.2)$$

Seega $W^{25}_{0.2}$ rahuldab tingimusi (52) ja (53).

Leiame hulga B^{25}_n. 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 $B^{25}_{0.2}$ moodustamine hakkab ülalt kuni klassi mass ületab 0.8. Esimese nelja klassi kogumass on 0.7635908, seega kuuluvad need klassid hulka B^{25}_n. 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

$$|B^{25}_{0.2}| = 1 + 25 + 300 + 2300 + 3325 = 5951$$

ning

$$\frac{1}{25} \log |B^{25}_{0.2}| \approx 0.501.$$

Kuigi hulkadesse $B^{25}_{0.2}$ ja $W^{25}_{0.2}$ 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 $W^{25}_{0.2}$ täielikult, kuid hulka $B^{25}_{0.2}$ 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 ühistüüpilisteks**.

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*

$$P(W_\epsilon^n) > 1-\epsilon. \tag{59}$$

2 *Piisavalt suure n korral*

$$(1-\epsilon)2^{n(H(X,Y)-\epsilon)} \le |W_\epsilon^n| \le 2^{n(H(X,Y)+\epsilon)}. \tag{60}$$

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

$$-\frac{1}{n}\log P(X_1,\ldots,X_n) = -\frac{1}{n}\sum_{i=1}^{n}\log P(X_i) \xrightarrow{P} H(X)$$
$$-\frac{1}{n}\log P(Y_1,\ldots,Y_n) = -\frac{1}{n}\sum_{i=1}^{n}\log P(Y_i) \xrightarrow{P} H(Y)$$
$$-\frac{1}{n}\log P\big((X_1,Y_1),\ldots,(X_n,Y_n)\big) = -\frac{1}{n}\sum_{i=1}^{n}\log P(X_i,Y_i) \xrightarrow{P} H(X,Y)$$

(ülesanne). Ka väite **2** tõestus on analoogiline:

$$1 \geq P(W_\epsilon^n) = \sum_{(x^n,y^n)\in W_\epsilon^n} P(x^n,y^n) \geq |W_\epsilon^n| 2^{-n(H(X,Y)+\epsilon)},$$
$$1-\epsilon \leq P(W_\epsilon^n) \leq \sum_{(x^n,y^n)\in W_\epsilon^n} P(x^n,y^n) \leq |W_\epsilon^n| 2^{-n(H(X,Y)-\epsilon)},$$

millest

$$(1-\epsilon)2^{n(H(X,Y)-\epsilon)} \leq |W_\epsilon^n| \leq 2^{n(H(X,Y)+\epsilon)}.$$

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

$$P(x^n, y^n) = \Big(\frac{1}{80}\Big)\Big(\frac{9}{80}\Big)^2\Big(\frac{63}{80}\Big)^7 = (pq)((1-p)q)^2((1-p)(1-q))^7 = q^3(1-q)^7(1-p)^9 p.$$

$$\begin{aligned}\frac{1}{n}\log P(x^n, y^n) &= \frac{3}{10}\log q + \frac{7}{10}\log(1-q) + \frac{9}{10}\log(1-p) + \frac{1}{10}\log p \\ &= q\log q + \frac{7}{40}\log q - \frac{7}{40}\log(1-q) + (1-q)\log(1-q) + (1-p)\log(1-p) + p\log p \\ &= -h(q) - h(p) + \frac{7}{40}\log(\frac{q}{1-q}).\end{aligned}$$

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

$$-\frac{1}{n}\log P(X_1, \ldots, X_n) \to H_X, \quad \text{p.k.}, \tag{61}$$

kus H_X on X_i entroopia ja seega protsessi entroopiamäär. Sõltumatuse korral järeldub (61) vahetult (tugevast) suurte arvude seadusest. Selgub aga, et koondumine (61) 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 $X_1, X_2 \dots$ on (nõrk) AEP omadus, kui kehtib (61), 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 $X_1, X_2, \dots$ 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
$$2^{-n(A+\epsilon)} \le \frac{Q(x^n)}{P(x^n)} \le 2^{-n(A-\epsilon)};$$

   2 piisavalt suure n korral
$$P(A_\epsilon^n) > 1-\epsilon;$$

   3 piisavalt suure n korral

$$(1-\epsilon)2^{-n(A+\epsilon)} \le Q(A_\epsilon^n) \le 2^{-n(A-\epsilon)}.$$

3. Olgu $X_1, X_2, \dots$ iid juhuslikud suurused, $X_i \sim U[0,1]$ (ühtlane jaotus). Konstrueerime n-tahuka küljepikkustega $X_1, \dots, X_n$, 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 $X_1, X_2, \dots$ lõpliku seisundite hulgaga statsionaarne Markovi ahel üleminekumaatriksiga I (ühikmaatriks). Tõestada koondumine (61).

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 (mäluta) kanal üleminekutõenäosuste maatriks

$$\big(P(y|x)\big)_{x\in\mathcal{X},y\in\mathcal{Y}}. \tag{62}$$

Arv $P(y|x)$ on tõenäosus, et sümboli x sisendamisel kanalisse väljub sealt sümbol y. Vigadeta kanali korral on üleminekumaatriks ühikmaatriks.

Olgu nüüd $P(x)$ mingi jaotus sisendtähestikul $\mathcal{X}$. Seda interpreteerime kui sisendite jaotust. Koos kanaliga (62), 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 (62)* **võimsus** *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. Lemmast 1.3 teame, et see funktsioon on nõgus, mistõttu lokaalne maksimum on ka globaalne ning maksimumi võib leida kumerate optimiseerimismeetoditega.

- Kanali võimsus ei saa ületada logaritmi tähestiku suurusest: $C\le \log\min\{|\mathcal{X}|,|\mathcal{Y}|\}$, sest

  $$C = \max_{P(x)} I(X;Y) \le \max_{P(x)} H(X) \le \log|\mathcal{X}|, \quad C = \max_{P(x)} I(X;Y) \le \max_{P(x)} H(Y) \le \log|\mathcal{Y}|.$$

- 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 maksimaalseslt 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

$$\begin{pmatrix} \text{p} & \text{1-p} & 0 & 0 \\ 0 & 0 & \text{q} & \text{1-q} \end{pmatrix}$$

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 sisendjaoutuse korral. Saadud võimsus ühtib intuitsiooniga – kui vigadeta klaviatuuri korral edastame korraga maksimaalseslt $\log 26$ 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

$$\begin{pmatrix} \text{1-p} & \text{p} \\ \text{p} & \text{1-p} \end{pmatrix}$$

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.

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

$$\begin{pmatrix} \frac{1}{3} & \frac{1}{6} & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & \frac{1}{6} \end{pmatrix}.$$

4.3 Kanaliteoreem

4.3.1 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* $\{1,\ldots,M\}$*;*
- *kodeeriv funktsioon*
$$\mathcal{C}:\{1,\ldots,M\}\to\mathcal{X}^n;$$
- *dekodeeriv funktsioon*
$$g:\mathcal{Y}^n\to\{1,2,\ldots,M\}.$$

Olgu λ_i 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 $\{1,\ldots,M\}$. 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

$$P_e\leq\lambda_{max}.$$

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.

Näide: Olgu $|\mathcal{X}|=2$ ja $\mathcal{C}$ *ühtlane kood*, mis $M=2^n$ korral seab sõnale $i\in\{1,\ldots,M\}$ vastavusse tema kahendesituse. Selle koodi määr on 1. On selge, et kui $|\mathcal{X}|=2$, siis parema määraga koodi konstrueerida pole võimalik.
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):

$$1-\lambda_i=\mathbf{P}(\hat{W}=i|W=i)=\mathbf{P}(Y^n=X^n(i))=(1-p)^n.$$

Kuigi koodil on kõrge määr, pole see antud kanali korral mõistlik.
Binaarse sümmeetrilise kanali korral pakutakse tihti välja nn *kordamiskoodi* (*repetition code*): iga bitt sõna W kahendesituses (ühtlane kood) 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). Iga etteantud $\epsilon > 0$ korral saab valida piisavalt pika m (sõltub M-st) nii, et $\lambda_{max} < \epsilon$. Küll on aga sellise koodi määr $\frac{1}{m}$ (ülesanne). Garanteerimaks, et M kasvades $\lambda_{max} \leq \epsilon$, peab m kasvama, st koodi määr läheneb nullile.

Def 4.4 *Olgu $P(y|x)$ diskreetne mäluta kanal. Arv R on kanali* **saavutatav määr**, *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 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 $R \leq C$.

4.3.2 Esimese väite tõestus

Olgu $R < C$. Näitame, et R on saavutatav määr.

Esimese sammuna fikseerime suvalise $\epsilon > 0$ 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 probleemiks kodeerimine. Teadmata tema struktuuri paistab ainus võimalus kodeerimiseks $n \times 2^{Rn}$ tabelist õige vaste otsimine ning see on ebapraktiline. Samuti on ülalkirjeldatud protseduuri korral ebapraktiline väljundi y^n dekodeerimine, sest selleks tuleb kontrollida paari (x^n, y^n) ühistüüpilisust kõigi 2^{nR} võimaliku sisendi korral.

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*, mida praegu peetakse üheks paljulubavaks praktiliselt rakendatavaks kanali võimsuse saavutamise viisiks. Samuti on populaarsed nn paarsust kontrollivad (*parity check*) veaparanduskoodid (*error-correcting codes*). Viimaste nn prototüüp-kood on nn Hammingi kood, mille tööpõhimõttega põgusalt tutvume.

4.3.3 Hammingi kood

Hammingi kood kuulub binaarse sümmeetrilise kanali tarbeks loodud nn paarsust kontrollivate 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.
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

$$\lambda_i = 1 - ((1-p)^7 + 7p(1-p)^6).$$

Ülesanne: Dekodeerige sõnad

$$1101011, \quad 0110110, \quad 0100111, \quad 1111111.$$

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)

$$t = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} s_1 \\ s_2 \\ s_3 \\ s_4 \end{pmatrix} \tag{63}$$

Olgu seoses (63) olev maatriks P. Siis (63) on $t = Ps$, kus $s = (s_1, s_2, s_3, s_4)^T$. Defineerime 7×4 maatriksi

$$G := \begin{pmatrix} I_4 \\ P \end{pmatrix},$$

kus I_4 on 4×4 ühikmaatriks. Siis iga koodisõna c avaldub

$$c = \begin{pmatrix} s \\ t \end{pmatrix} = Gs = \begin{pmatrix} I_4 \\ P \end{pmatrix} s. \tag{64}$$

Seosest (64) järeldub nüüd koodi lineaarsus. Defineerime nüüd 3×7 maatriksi H järgmiselt

$$H = \begin{pmatrix} P & I_3 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 \end{pmatrix} \tag{65}$$

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

$$Hc = HGs = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}. \tag{66}$$

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 (66). 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

$$c = \arg \min_{i=1,\ldots,16} h(r, c_i), \tag{67}$$

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.

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 (66). 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.

4.3.4 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

$$H(Y^n|X^n) = \sum_{i=1}^{n} H(Y_i|X_i), \tag{68}$$

millest

$$\begin{aligned} I(X^n; Y^n) &= H(Y^n) - H(Y^n|X^n) = H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) \\ &\leq \sum_{i=1}^{n} (H(Y_i) - H(Y_i|X_i)) = \sum_{i=1}^{n} I(X_i; Y_i) \leq nC \end{aligned}$$

■

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

$$I(W; Y^n) \leq I(X^n; Y^n). \tag{69}$$

Arvestades, et

$$I(W; Y^n) = H(W) - H(W|Y^n) = H(W), \tag{70}$$

saame lemmast 4.1 ja andmetöötlusvõrratusest (69)

$$nR = H(W) = I(W; Y^n) \leq I(X^n; Y^n) \leq \sum_{i=1}^{n} I(X_i; Y_i) \leq nC.$$

■

Et $W \to X^n \to Y^n$ on Markovi ahel, siis lisaks andmetöötlusvõrratusele (69) kehtib veel andmetöötlusvõrratus:

$$I(W; Y^n) \leq I(W; X^n). \tag{71}$$

Kui $\lambda_{max} = 0$, siis $H(W|X^n) = 0$. Tõepoolest, veatu infovahetuse korral $H(W) = I(W; Y^n)$ ning seos (71) on kujul: $H(W) = I(W; Y^n) \leq I(W; X^n) = H(W) - H(W|X^n)$. Viimane saab kehtida vaid siis kui $H(W|X^n) = 0$ ja $I(W; Y^n) = I(W; X^n)(= H(W))$. Et $X^n = \mathcal{C}(W)$, siis $H(W|X^n) = H(W|\mathcal{C}(W)) = 0$ tähendab sisuliselt seda, et kood $\mathcal{C}$ on ühene. Sellisel juhul $H(X^n) = H(W)$, millest

$$H(W) = I(W; Y^n) \leq I(X^n; Y^n) = H(X^n) - H(X^n|Y^n) = H(W) - H(X^n|Y^n) \leq H(W).$$

Seega ülaltoodud võrratused on võrdused ja (69) on võrdus:

$$I(W; Y^n) = I(X^n; Y^n). \tag{72}$$

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 (72), mis tuleneb koodi ühesusest. 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

$$\sum_{i=1}^{n} I(X_i; Y_i) \leq nC$$

kehtib siis, kui iga i korral $I(X_i; Y_i) = C$ ehk X_i jaotus on selline, mis saavutab kanali võimsuse.

Seega $(2^{nR}, n)$ kood, mille korral $P_e = 0$ ja $R = C$ peab rahuldama tingimusi:

- $\mathcal{C}$ on (üks)ühene;
- Y_i^n on iid juhuslikud suurused jaotusega

$$P(y) = \sum_x P(y|x)P^*(x), \tag{73}$$

kus $P^*(x)$ saavutab kanali võimsuse.

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: Kui valida sõna $X^n = X_1, \ldots, X_n$ ühtlase jaotusega hulgast $\{1, 3, 5, \ldots, 25\}^n$, siis on $X_1, \ldots, X_n$ iid juhuslikud suurused ning X_i jaotus on ühtlane üle paaritute tähtede $\{1, 3, 5, \ldots, 25\}$. Sellise sisendjaotuse korral on väljund ühtlane üle kõikide tähtede ning $Y_1, \ldots, Y_n$ on iid juhuslikud suurused jaotusega (73). Seega viga $P_e = 0$ ja kanali võimsus on saavutatud, sest hulga $\{1, 3, 5, \ldots, 25\}^n$ võimsus on $13^n = 2^{n \log 13} = 2^{nC}$. Paneme tähele, et saavutatav määr $R = C$.
- 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.

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*

$$H(W|Y^n) \leq 1 + \mathbf{P}(W \neq \hat{W})nR. \tag{74}$$

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 (69) ja lemmast 4.1. Seega

$$R \leq P_e R + \frac{1}{n} + C. \tag{75}$$

Et n kasvades $P_e R + \frac{1}{n} \to 0$, siis $R \leq C$. ■

Märkus: Tõestatud väidet nimetatakse teinekord ka nõrgaks väiteks. Selle väite tugev versioon on: kui leidub $\epsilon > 0$ nii, et $R \geq C + \epsilon$, siis $P_e \to 1$.

Võrratusest (75):

$$P_e \geq 1 - \frac{C}{R} - \frac{1}{nR}.$$

Seega, kui $C < R$, siis $\frac{C}{R} < 1$, millest suure n korral P_e on alt tõkestatud. Sellisel juhul ei saa P_e olla 0 ka väiksese n korral. Järelikult, kui $C < R$, siis $P_e > 0$ iga n korral.

Võrratusest (75) saame veel, et

$$P_e \geq 1 - \frac{C}{R} - \frac{1}{nR} \approx 1 - \frac{C}{R}.$$

Nii saame asümptootilise alumise tõkke veale P_e juhul kui $C < R$. Graafikult

$$R \mapsto 1 - \frac{C}{R}$$

on ilusti näha kui kiiresti see tõke R-i suurenedes kasvab.

4.4 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* $\{1,\ldots,M\}$*;*
- *kodeerivad funktsioonid*

$$\mathcal{C}_i : \{1,\ldots,M\}\times\mathcal{Y}^{i-1}\to\mathcal{X};$$

- *dekodeeriv funktsioon*

$$g:\mathcal{Y}^n\to\{1,2,\ldots,M\}.$$

Tagasisidega infovahetuse kasulikkus tuleb hästi esile näiteks binaarse kadumiskanali korral. Tõepoolest, 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

$$\begin{aligned}
I(W;Y^n)&=H(Y^n)-H(Y^n|W)\\
&=H(Y^n)-H(Y_1|W)-H(Y_2|Y_1,W)-\cdots-H(Y_n|Y_1,\ldots,Y_{n-1},W)\\
&=H(Y^n)-\sum_{i=1}^{n}H(Y_i|Y_1,\ldots,Y_{i-1},W)\\
&=H(Y^n)-\sum_{i=1}^{n}H(Y_i|Y_1,\ldots,Y_{i-1},W,X_i).
\end{aligned}$$

Viimane võrdus kehtib sest $X_i=\mathcal{C}_i(Y_1,\ldots,Y_{i-1},W)$. Et aga Y_i sõltub vaid X_i-st, siis

$$P(y_i|y_1,\ldots,y_{i-1},w,x_i)=P(y_i|x_i)\quad\text{ja}\quad H(Y_i|Y_1,\ldots,Y_{i-1},W,X_i)=H(Y_i|X_i).$$

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 $x_{i+1}, x_{i+1}, \ldots$ annab ka y_i kohta infot.

4.5 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 $V_1, \ldots, V_n$. 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. Et H võib olla palju väiksem kui $\log|\mathcal{V}|$, on teoreem oluline.
Kirjeldatud tulemuse saavutamiseks kasutame *kaheastmelist kodeerimist*: eelkõige kodeerime võimalikult optimaalselt vektori V^n koodisõnadeks hulgast $\{1, \ldots, 2^{nR}\}$ saades nii juhusliku sõna W. Viimase kodeerime ning saadame $(2^{nR}, n)$ koodi abil hinnanguks $\hat{W}$. Kui esimene kodeerimine õnnestub läbi viia nii, et $R < C$ (ja seejuures tehtav võimalik viga on väga väike), on (piisavalt suure n korral) eesmärk saavutatud: vektor V^n saadeti n ülekandega läbi kanali nii, et infovahetuse käigus tehtud viga on etteantud raamides. Olgu g mõlema dekodeerimise kompositsioon, s.t.

$$g : \mathcal{Y}^n \to \mathcal{V}^n.$$

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. Et protsessil on AEP omadus, siis $\forall \epsilon$ korral leidub hulk W_ϵ^n nii, et $P(W_e^n) > 1-\epsilon$ ja $|W_\epsilon^n| \leq 2^{n(H+\epsilon)}$. Indekseerime kõik sõnad hulgast W_ϵ^n. Nii saame sõnastiku, mis koosneb ülimalt $2^{n(H+\epsilon)}$ sõnast. Ainult neid sõnu sisetame kanalisse (tõenäosus, et originaalsõna sinna hulka ei kuulu, on väiksem kui ϵ). Kui $H+\epsilon < C$, siis kanaliteoreemist saame, et saadud sõnad saab edastada kuitahes väikese veaga. Vastuvõtja dekodeerib esmajärjekorras indeksi hulgast W_ϵ^n ja seejärel leiab temale vastava sõna hulgast $\mathcal{V}^n$. On selge, et 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 tõepoolest 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

$$H \leq \frac{H(V_1, \ldots, V_n)}{n} = \frac{1}{n} H(V^n) = \frac{1}{n} H(V^n | \hat{V}^n) + \frac{1}{n} I(V^n; \hat{V}^n).$$

Esimene võrratus kehtib, sest statsionaarsuse tõttu $H(V_n|V_1,\dots,V_{n-1}) \searrow H$ ja

$$\begin{aligned} H(V_1,\dots,V_n) &= H(V_1)+\cdots+H(V_n|V_1,\dots,V_{n-1}) \\ &= H(V_n)+H(V_n|V_{n-1})+\cdots+H(V_n|V_1,\dots,V_{n-1}) \\ &\geq nH(V_n|V_1,\dots,V_{n-1}). \end{aligned}$$

Fano võrratusest saame ($|\mathcal{V}|$ on lõplik)

$$H(V|\hat{V}) \leq 1+\mathbf{P}(\hat{V}\neq V^n)\log|\mathcal{V}|^n = 1+\mathbf{P}(\hat{V}\neq V^n)n\log|\mathcal{V}|.$$

Andmetöötlusvõrratusest ($V^n \to Y^n \to \hat{V}^n$) saame

$$I(V^n;\hat{V}^n) \leq I(V^n;Y^n).$$

Teoreemi 4.7 tõestusest nägime, et

$$I(V^n;Y^n) \leq nC.$$

Seega

$$H \leq \frac{1}{n}+\mathbf{P}(\hat{V}\neq V^n)\log|\mathcal{V}|+C.$$

Kui $P_e \to 0$, siis $H \leq C$; kui $H > C$, siis

$$\mathbf{P}(\hat{V}\neq V^n) \geq \frac{H-C}{\log|\mathcal{V}|}-\frac{1}{n\log|\mathcal{V}|},$$

millest näeme, et kui $H > C$, siis leidub $\epsilon > 0$ nii, et $\mathbf{P}(\hat{V}\neq V^n) > \epsilon$, kui n on piisavalt suur. See aga tähendab, et leidub $\epsilon > 0$ nii, et $\mathbf{P}(\hat{V}\neq V^n) > \epsilon$ iga n korral. ■

Seega üheastmeline (kombineeritud) kodeerimine ja tagasiside ei suurenda infovahetuse efektiivsust: kaheastmeline kodeerimine annab sama hea tulemuse. Kuigi see paistab esmapilgul 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 $\frac{H}{k} < C$.

4.6 Ülesanded

1. Koosnegu sõnastik M sõnast $\{1, \ldots, M\}$. Olgu

   $$g : \{1, \ldots, M\} \to \{0,1\}^{\log M}$$

   kood (sõnade kahendesitus). Et kood on vaja edastada läbi vigadega binaarse kanali, esitatakse iga bitt m kordselt. Nii saame ühe selle kanali koodi – kordamiskoodi (*repetition code*) $\mathcal{C}$. Leida selle koodi määr.

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 $C \leq C(K(\epsilon_1)) \wedge C(K(\epsilon_2))$.
   - 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 $I(X^n; Y^n) \leq n - h(\epsilon)$. 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 $(\mathcal{X}, P_1 P_2, \mathcal{X})$ võimsus. Tõestada, et

$$C \leq C_1 \wedge C_2.$$

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 Sagedustüübid ja universaalsed koodid

Käesolevas peatükis vaadeldav nn sagedustüüpide teooria esitasid 80'ndatel Csiszar ja Körner.

5.1 Sagedustüübid

Olgu $\mathcal{X} = \{a_1, \ldots, a_m\}$ lõplik tähestik. Vaatleme jada $x^n \in \mathcal{X}^n$.

Def 5.1 *Jada $x^n = x_1, \ldots, x_n$* **(sagedus)tüüp** *P_{x^n} on selle empiiriline mõõt. Teisisõnu, P_{x^n} on tõenäosusjaotus tähestikul $\mathcal{X}$ ning*

$$P_{x^n}(a) = \frac{1}{n} \sum_{i=1}^{n} I_{x_i}(a).$$

Arv

$$N_{x^n}(a) = \sum_{i=1}^{n} I_{x_i}(a)$$

on tähe a sagedus jadas x^n ja $P_{x^n}(a)$ on selle jada suhteline sagedus jadas x^n.
Jada x^n sagedustüüp on alati selline tõenäosusjaotus, et iga tähe a tõenäosus on kujul $\frac{k}{n}$, kus $k \in \mathbb{Z}^+$. Iga selline tõenäosusjaotus on mingi jada x^n sagedustüüp. Olgu selliste jaotuste hulk $\mathcal{P}_n$.

Def 5.2 *Olgu $\mathcal{P}_n$ kõikide n-elemendiliste jadade sagedustüüpide hulk.*

Näited:

1. Kui $\mathcal{X} = \{0, 1\}$, siis

   $$\mathcal{P}_n = \Big\{(0,1), (\frac{1}{n}, \frac{n-1}{n}), \ldots, (\frac{n-1}{n}, \frac{1}{n}), (1,0)\Big\},$$

   kokku $n+1$ sagedustüüpi.

2. Kui $\mathcal{X} = \{a, b, c\}$, siis

   $$\mathcal{P}_n = \Big\{(0,0,1), (0, \frac{1}{n}, \frac{n-1}{n}), \cdots, (0,1,0), \cdots, (\frac{1}{n}, \frac{n-1}{n}, 0), \ldots, (\frac{n-1}{n}, \frac{1}{n}, 0), (1,0,0)\Big\},$$

   kokku

   $$\frac{(n+2)(n+1)}{2}$$

   sagedustüüpi.

On selge, et hulk $\mathcal{X}^n$ jaguneb ekvivalentsiklassideks sagedustüüpide järgi. Nimetame neid **tüübiklassideks**.

Def 5.3 *Olgu $P \in \mathcal{P}_n$. Hulka*

$$T(P) = \{x^n \in \mathcal{X}^n : P_{x^n} = P\}$$

nimetame P-tüübiklassiks.

Näide: Olgu $\mathcal{X} = \{0, 1, 2\}$, $x^5 = 00210$. Tüüp P_{x^5} on

0	1	2
$\frac{3}{5}$	$\frac{1}{5}$	$\frac{1}{5}$

Tüübiklassi $T(P_{x^5})$ moodustavad jadad

$$T(P_{x^5}) = \{(00012), (00021), (00102), \ldots, (12000)\}.$$

Kokku on selles tüübiklassis

$$|T(P_{x^5})| = \frac{5!}{3!1!1!} = 20$$

elementi.

Kui palju on erinevaid tüübiklasse? Ülemist hinnangut on kerge leida:

$$|\mathcal{P}_n| \leq (n+1)^{|\mathcal{X}|-1}. \tag{76}$$

Tüüp P_{x^n} koosneb $|\mathcal{X}|$ tähest. Iga tähe tõenäosus on üks $n+1$-st arvust

$$\frac{0}{n}, \ldots \frac{n}{n}$$

ning viimase tähe tõenäosus on üheselt määratud eelmistega. Järgnevas kasutame enamasti $(n+1)^{|\mathcal{X}|}$. Täpselt on tüüpe kordustega kombinatsioonide arv $|\mathcal{X}|$ tähest n-kaupa ja see arv on

$$C^{|\mathcal{X}|-1}_{n+|\mathcal{X}|-1}.$$

Oluline, et erinevate tüüpide arv kasvab polünomiaalselt, samal ajal kui $\mathcal{X}^n$ kasvab eksponentsiaalselt. See aga tähendab, et vähemalt ühes tüübiklassis on eksponentsiaalselt palju elemente. Edaspidi tõestame, et tegelikult on peaaegu igas klassis eksponentsiaalselt palju elemente.

Järgnev lemma näitab, et iid juhusliku vektori korral sõltub jada x^n tõenäosus selle tüübist.

Lemma 5.1 *Olgu $X_1, \ldots, X_n$ tähestikul $\mathcal{X}$ antud iid juhuslikud suurused jaotusega Q. Olgu*

$$Q^n(x^n) := \prod_{i=1}^{n} Q(x_i)$$

jada x^n tõenäosus. Kehtib

$$Q^n(x^n) = 2^{-n(H(P_{x^n}) + D(P_{x^n} \| Q))}.$$

Tõestus. Paneme tähele, et iga kahe tähestikul $\mathcal{X}$ antud tõenäosusjaotuse P ja Q korral kehtib

$$\begin{aligned}\sum_{a\in\mathcal{X}} P(a)\log Q(a) &= \sum_{a\in\mathcal{X}} P(a)\log Q(a) - \sum_{a\in\mathcal{X}} P(a)\log P(a) - H(P)\\ &= \sum_{a\in\mathcal{X}} P(a)\log\frac{Q(a)}{P(a)} - H(P) = -(D(P\|Q)+H(P)).\end{aligned}$$

Nüüd

$$\begin{aligned}Q^n(x^n) &= \prod_{i=1}^{n} Q(x_i) = \prod_{a\in\mathcal{X}} Q(a)^{N_{x^n}(a)} = \prod_{a\in\mathcal{X}} Q(a)^{nP_{x^n}(a)}\\ &= \prod_{a\in\mathcal{X}} 2^{nP_{x^n}(a)\log Q(a)} = 2^{-n(H(P_{x^n})+D(P_{x^n}\|Q))}.\end{aligned}$$

■

Olgu jada x^n antud. Olgu $\mathcal{P}$ kõikide tähestikul $\mathcal{X}$ antud tõenäosuste hulk. Millise jaotuse $Q\in\mathcal{P}$ korral on jada x^n tulemise tõenäosus suurim? Lemmast saame, et selliseks jaotuseks on P_{x^n}, sest

$$\arg\max_{Q\in\mathcal{P}} Q^n(x^n) = \arg\min_{Q\in\mathcal{P}}(H(P_{x^n})+D(P_{x^n}\|Q)) = \arg\min_{Q\in\mathcal{P}} D(P_{x^n}\|Q) = P_{x^n}.$$

Järelikult on vektori x^n genereeriva jaotuse suurima tõepära hinnang P_{x^n}, ning suurim tõepära on

$$P_{x^n}(x^n) = 2^{-nH(P_{x^n})}. \tag{77}$$

Teisest küljest, fikseeritud $Q\in\mathcal{P}_n$ korral ei järeldu lemmast , et kõige suurema tõenäosusega väljund on selline jada x^n, mille tüüp on Q. Tõepoolest, fikseeritud allika jaotuse Q korral on kõige suurema tõenäosusega väljund see, mille tüüp on

$$\arg\min_{P\in\mathcal{P}_n}(H(P)+D(P\|Q)).$$

Viimane aga ei pruugi olla Q. Näiteks kui $\mathcal{X}=\{0,1\}$ ja $Q=B(1,q)$, kus $q>0.5$, siis suurima tõenäosusega väljund on vaid ühtedest koosnev vektor, sest sellisel juhul minimiseerib funktsiooni $P\mapsto H(P)+D(P\|Q)$ hoopis jaotus $B(1,1)$. (**Ülesanne.**)

Järgnevas anname alumise ja ülemise hinnangu tüübiklassi võimsusele. Tegelikult on tüübiklassi kuuluvate jadade arvu kerge välja arvutada, sest iga $P\in\mathcal{P}_n$ korral

$$|T(P)| = \frac{n!}{(nP(a_1))!(nP(a_2))!\cdots(nP(a_m))!}.$$

See arv on aga tülikas, mistõttu püüame seda hinnata.

Teoreem 5.4 *Iga $P \in \mathcal{P}_n$ korral*

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{nH(P)} \leq |T(P)| \leq 2^{nH(P)}. \tag{78}$$

Tõestus. Olgu $P \in \mathcal{P}_n$. Seosest (77) saame kergesti ülemise hinnangu tüübiklassi suurusele

$$1 \geq P^n(T(P)) := \sum_{x^n \in T(P)} P^n(x^n) = |T(P)| 2^{-nH(P)},$$

millest

$$T(P) \leq 2^{nH(P)}.$$

Alumise tõkke saamiseks tõestame, et iga $R \in \mathcal{P}_n$ korral

$$P^n(T(P)) \geq P^n(T(R)). \tag{79}$$

Võrratus (79) väidab, et kui $X_1, \ldots X_n$ on iid juhuslikud suurused klassi $\mathcal{P}_n$ kuuluva jaotusega P, siis tüübiklassi $T(P)$ tõenäosus on suurim.

Asume tõestama võrratust (79). Tuletame meelde, et $P, R \in \mathcal{P}_n$.

$$\frac{P^n(T(P))}{P^n(T(R))} = \frac{|T(P)| \prod_{a \in \mathcal{X}} P(a)^{nP(a)}}{|T(R)| \prod_{a \in \mathcal{X}} P(a)^{nR(a)}}.$$

Kehtib

$$\frac{|T(P)|}{|T(R)|} = \frac{n!}{(nP(a_1))!(nP(a_2))! \cdots (nP(a_m))!} \Big/ \frac{n!}{(nR(a_1))!(nR(a_2))! \cdots (nR(a_m))!} = \prod_{a \in \mathcal{X}} \frac{(nR(a))!}{(nP(a))!}.$$

On lihtne veenduda, et iga kahe naturaalarvu n, m korral

$$\frac{m!}{n!} \geq n^{m-n}.$$

Järelikult

$$\frac{(nR(a))!}{(nP(a))!} \geq (nP(a))^{nR(a)-nP(a)}$$

ehk

$$\begin{aligned}
\frac{P^n(T(P))}{P^n(T(R))} &= \prod_{a \in \mathcal{X}} \frac{(nR(a))!}{(nP(a))!} P(a)^{n(P(a)-R(a))} \\
&\geq \prod_{a \in \mathcal{X}} (nP(a))^{n(R(a)-P(a))} P(a)^{n(P(a)-R(a))} = \prod_{a \in \mathcal{X}} n^{n(R(a)-P(a))} \\
&= n^{n(\sum_a R(a) - \sum_a P(a))} = n^{n(1-1)} = 1.
\end{aligned}$$

Nüüd saame arvule $|T(P)|$ alumise tõkke.

$$\begin{aligned}1 &= \sum_{R\in\mathcal{P}_n} P^n(T(R)) \le \sum_{R\in\mathcal{P}_n} \max P^n(T(R)) = \sum_{R\in\mathcal{P}_n} P^n(T(P)) \\ &\le (n+1)^{|\mathcal{X}|} P^n(T(P)) = (n+1)^{|\mathcal{X}|} \sum_{x^n\in T(P)} P^n(x^n) \\ &= (n+1)^{|\mathcal{X}|} \sum_{x^n\in T(P)} 2^{-nH(P)} = (n+1)^{|\mathcal{X}|} |T(P)| 2^{-nH(P)},\end{aligned}$$

millest

$$|T(P)| \ge \frac{2^{nH(P)}}{(n+1)^{|\mathcal{X}|}}.$$

Eelviimane võrdus järeldub seosest (77). ■

Märkus: Arusaadavalt võib seoses (78) olevat alumist tõket täpsustada, asendades hinnangu $(n+1)^{|\mathcal{X}|}$ täpsema hinnangu $(n+1)^{|\mathcal{X}|-1}$ või tüübiklasside täpse arvuga $C_{n+|\mathcal{X}|-1}^{|\mathcal{X}|-1}$. Sama kehtib ka järgnevate tulemiste kohta ning edaspidi me sarnasi märkusi ei tee.

Näide: Olgu $\mathcal{X} = \{a, b\}$ ning olgu $P \in \mathcal{P}_n$. Siis mingi k korral $P(a) = \frac{k}{n}, P(b) = \frac{n-k}{n}$ ja $H(P) = h(\frac{k}{n})$. Kahe tähe korral

$$|T(P)| = \frac{n!}{k!(n-k)!} = C_n^k.$$

Tüübiklasside arv on $(n+1)$. Teoreem 5.4 annab hinnangu

$$\frac{1}{n+1} 2^{nh(\frac{k}{n})} \le C_n^k \le 2^{nh(\frac{k}{n})},$$

mis suure n korral võib olla üsna kasulik. Tõepoolest, kui $k = \alpha n$, siis ülaltoodud võrratustest järeldub, et

$$\log C_n^{\alpha n} \sim nh(\alpha).$$

Teoreem 5.5 *Olgu $X_1, \ldots, X_n$ iid juhuslikud suurused jaotusega Q. Siis iga $P \in \mathcal{P}_n$ korral*

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-nD(P\|Q)} \le Q^n(T(P)) \le 2^{-nD(P\|Q)} \tag{80}$$

Tõestus. Lemmast 5.1

$$Q^n(T(P)) = \sum_{x^n\in T(P)} Q^n(x^n) = \sum_{x^n\in T(P)} 2^{-n(D(P\|Q)+H(P))} = |T(P)| 2^{-n(D(P\|Q)+H(P))}.$$

Hinnangutest (78)

$$Q^n(T(P)) \le 2^{nH(P)} 2^{-n(D(P\|Q)+H(P))} = 2^{-nD(P\|Q)}$$

ning

$$Q^n(T(P)) \geq (n+1)^{-|\mathcal{X}|} 2^{nH(P)} 2^{-n(D(P\|Q)+H(P))} = (n+1)^{-|\mathcal{X}|} 2^{-nD(P\|Q)}.$$

■

Võtame kokku olulisima ülaltõestatust.

1 Hulk $\mathcal{X}^n$ jaguneb tüübiklassideks, neid klasse on vähem kui $(n+1)^{|\mathcal{X}|}$.

2 Tüübiklassi $T(P)$ elementide arv on suurusjärku $2^{nH(P)}$.

3 Vaadeldes hulka $\mathcal{X}^n$ iid juhuslike suuruste $X_1, \ldots, X_n$ võimalike realisatsioonidena ning $X \sim Q$, saame iga $x^n \in T(P)$ $(P \in \mathcal{P}_n)$ korral

$$Q^n(x^n) = 2^{-n(H(P)+D(P\|Q))}.$$

4 Vaadeldes hulka $\mathcal{X}^n$ iid juhuslike suuruste $X_1, \ldots, X_n$ võimalike realisatsioonidena ning $X \sim Q$, saame, et iga tüübiklassi $T(P)$, $P \in \mathcal{P}_n$ tõenäosus on suurusjärku $2^{-nD(P\|Q)}$. Kui $Q \in \mathcal{P}_n$, siis (võrratus (79))

$$Q^n(T(Q)) \geq Q^n(T(P)) \quad \forall P \in \mathcal{P}_n.$$

5.2 Plokk-koodid

Olgu $\mathcal{X}$ tähestik, $\mathcal{D}$ kooditähestik. **Plokk-kood** (mõnikord ka FF kood) kooseneb funktsioonidest:

$$\mathcal{C}_n : \mathcal{X}^n \to \mathcal{D}^m, \quad g_n : \mathcal{D}^m \to \mathcal{X}^n.$$

Def 5.6 *Arvu*

$$\frac{m}{n} \log D$$

nimetatakse koodi $\mathcal{C}_n$ **määraks**.

On selge, et koodi määr on seda väiksem, mida lühemad on koodisõnad. Mõistlik kood ei saa olla liiga väikese määraga.

Hulgas $\mathcal{D}^m$ on 2^{nR} elementi, kus

$$R = \frac{m}{n} \log D. \tag{81}$$

Seega võib plokk-koodi vaadelda kui funktsioone

$$\mathcal{C}_n : \mathcal{X}^n \to \{1, \ldots, 2^{nR}\}, \quad g_n : \{1, \ldots, 2^{nR}\} \to \mathcal{X}^n,$$

kus R on defineeritud kui (81). Koodi määr on sellisel juhul R ning 2^{nR} on täisarv. Sellist plokk-koodi nimetame $(n, 2^{nR})$ koodiks. Arv m on üheselt määratud arvudega n ja R.

Plokk-kood teeb vea, kui

$$g(\mathcal{C}_n(x^n)) \neq x^n),$$

vea tõenäosus on

$$P_e(n) := \mathbf{P}\big(g(\mathcal{C}_n(X^n)) \neq X^n\big),$$

kus $X^n = (X_1, \ldots, X_n)$ on tähestikul $\mathcal{X}$ antud kodeeritav juhuslik vektor.

Ütleme, et määr R on **saavutatav**, kui leidub $(n, \ulcorner 2^{nR} \urcorner)$ koodide jada nii, et $P_e(n) \to 0$.

Seoses nõrga AEP omadusega nägime, et iid vekorit X^n sai kodeerida nii, et keskmine koodipikks biti kohta, L_n, koondub entroopiaks H. Sisuliselt järjestasime kõik nõrgalt tüüpilised sõnad ning kodeerisime neid indeksite järgi. Juhul, kui $D = 2$, siis selleks kulus ligikaudu $n(H + \epsilon)$ bitti. Ülejäänud sõnad kodeerisime pikemalt, kuid neid oli nii vähe, et nad ei mõjutanud keskmist. Selliselt defineeritud kood oli veatu. Lubades aga viga, mille tõenäosus on ülimalt ϵ, võime kodeerida vaid nõrgalt tüüpilisi sõnu (ülejäänute tõenäosus läheneb nullile). Nii saame koodi

$$\mathcal{X}^n \to \{0,1\}^{n(H+\epsilon)},$$

mille määr on $R = H + \epsilon$. Valides ϵ kuitahes väikese, saame iga $R > H$ korral määraga R plokk-koodi, mille veatõenäosus läheneb nullile. Täpselt nii kodeerisime sõnu V^n kanaliga infovahetusel kaheastmelise kodeerimise korral. Formuleerime (veelkord) ülaltoodud arutelu.

Teoreem 5.7 *Olgu $X_1, \ldots, X_n$ esimesed n elementi nõrga AEP omadusega juhuslikust protsessist. Olgu H selle protsessi entroopiamäär. Siis iga $R > H$ on saavutav määr.*

Tõestus. Olgu $0 < 2\epsilon < R - H$. AEP omaduse tõttu leidub n_o nii et $P(W_\epsilon^n) > 1 - \epsilon$, kui $n > n_o$. Valime $n > n_o$ nii, et $\frac{1}{n} < \epsilon$. Nõrgalt tüüpiliste sõnade hulga võimsus oli tõkestatud $|W_\epsilon^n| \leq 2^{n(H+\epsilon)}$. Järjestame need sõnad. Kui $n(H + \epsilon)$ pole täisarv, siis leidub $\epsilon' < \epsilon + \frac{1}{n} < 2\epsilon$ nii, et $n(H + \epsilon')$ on täisarv. Seega saame ülimalt $2^{n(H+\epsilon')}$ sõna. Igale nõrgalt tüüpilisele sõnale same vastavusse tema indeksi, ülejäänud sõnad kodeerime suvaliselt. Nii saame kujutise

$$\mathcal{C}_n : \mathcal{X}^n \to \{1, \ldots, 2^{n(H+\epsilon')}\}. \tag{82}$$

Dekodeeriv funktsioon seab iga nõrgalt tüüpilise sõna indeksile vastavusse originaalsõna, ülejäänud elemendid hulgast (82) dekodeeritakse suvaliselt. Oleme defineerinud plokk-koodi, mille määr on $H + \epsilon' \leq H + 2\epsilon \leq R$. Selline plokk-kood teeb vea vaid mitte-nõrgalt-tüüpilise sõna dekodeerimisel; kui $n > n_o$, on vea tõenäosus väiksem kui ϵ, st $P_e(n) \leq \epsilon$. ■

Märkus: Seosest (81) saame, et tingimus $R > H$ on ekvivalentne

$$\frac{m}{n} > \frac{H}{\log D} = H_D,$$

kus H_D logaritmi D baasil defineeritud entroopiamäär. Mõnikord defieeritaksegi koodi määr kui suhe $\frac{m}{n}$ ning sellisel juhul on määr savutatav, kui ta on suurem kui H_D.

Järgmine teoreem väidab vastupidist: kui $R < H$, siis iga $\lambda > 0$ korral $P_e(n) \geq 1 - \lambda$, kui n on piisavalt suur.

Teoreem 5.8 *Olgu $X_1, \ldots, X_n$ esimesed n elementi nõrga AEP omadusega juhuslikust protsessist, H olgu selle protsessi entroopiamäär. Olgu $\lambda > 0$. Kui $R < H$ siis leidub n_o nii, et suvalise $(n, 2^{Rn})$ koodi korral $P_e > 1 - \lambda$, kui $n > n_o$.*

Tõestus. Olgu $\epsilon > 0$ ja $\delta > 0$ sellised, et

$$0 < \epsilon < \lambda, \quad 0 < \epsilon < \delta < H - R.$$

Et protsessil on AEP omadus, siis iga ϵ korral leidub n_o nii, et $P^n(W_\epsilon^n) > 1 - \epsilon$, kui $n > n_o$. Suurendades arvu n_o kui vaja, saame ta võtta nii suureks, et iga $n > n_o$ korral

$$2^{n(\epsilon - \delta)} < \lambda - \epsilon. \tag{83}$$

Olgu $\mathcal{C}_n$ ja g_n mingi $(n, 2^{nR})$ kood. Defineerime hulga

$$B_n = \{x \in \mathcal{X}^n : g_n(\mathcal{C}_n(x^n)) = x^n\}.$$

On selge, et

$$|B_n| \leq 2^{nR} < 2^{n(H-\delta)}.$$

Tuletades meelde, et iga $x^n \in W_\epsilon^n$ korral

$$P^n(x^n) \leq 2^{-n(H-\epsilon)},$$

saame iga $n > n_o$ korral

$$\begin{aligned} 1 - P_e &= P^n(B_n) = P^n(B_n \cap W_\epsilon^n) + P^n(B_n \cap (W_\epsilon^n)^c) \\ &\leq \epsilon + |B_n \cap W_\epsilon^n| \max_{x^n \in W_\epsilon^n} P^n(x^n) \\ &\leq \epsilon + 2^{n(H-\delta)} 2^{-n(H-\epsilon)} \\ &\leq \epsilon + 2^{n(\epsilon-\delta)} < \lambda, \end{aligned}$$

kus viimane võrratus tuleneb seosest (83). ■

Märkus: Kasutades Fano võrratust, saaksime eelmises peatükis kasutatud argumendi abil tõestada järgmise väite: kui R on saavutatav määr, siis $R \geq H$; juhul kui $R < H$, on $P_e(n)$ alt tõkestatud positiivse konstandiga. Ülaltoodud teoreem on tugevam, sest väidab, et vea tõenäosus $P_e(n)$ läheneb n kasvades ühele.

5.3 Universaalsed plokk-koodid

Teoreem 5.7 väidab, et määr $H < R$ on saavutatav, st leiduvad $(n, 2^{nR})$ koodid nii, et $P_e \to 0$. Nimetatud teoreem põhines nõrgalt tüüpilistel sõnadel: sisuliselt kodeeriti vaid selliseid sõnu, ülejäänute tõenäosus on tühine. Nõrgalt tüüpilised sõnad on need vektorid $x^n \in \mathcal{X}^n$, mille korral $P^n(x^n)$ ei erine palju arvust 2^{-nH}. Otsustamaks, kas x^n on nõrgalt tüüpiline või mitte, tuleb seega teada vektori X^n jaotust P (juhul kui X^n on iid vektor, tuleb seega teada komponendi jaotust). Mida teha, kui P pole teada?
Selgub, et on võimalik konstrueerida nulliks koonduva veaga koode ka siis, kui vektori X^n jaotus pole teada. Selliseid koode nimetatakse **universaalseteks koodideks**. Loomulikult ei saa universaalse $(n, 2^{nR})$ koodi määr olla väiksem kui H, sest teoreem 5.8 kehtib iga koodi korral.

Alljärgnevas eeldame, et $\mathcal{X}$ on lõplik

5.3.1 Funktsioon $F(R, Q)$

Olgu Q tähestikul $\mathcal{X}$ antud tõenäosusmõõt. Defineerime funktsiooni (*reliability function*)

$$F(R, Q) := \min_{P: H(P) \geq R} D(P\|Q).$$

Et $\mathcal{X}$ on lõplik, siis kõikide tõenäosusmõõtude hulk on kinnine kumer hulk ruumis $\mathbb{R}^{|\mathcal{X}|}$ (seda hulka nimetatakse simpleksiks (*simplex*)), funktsioonid

$$P \mapsto H(P) \text{ ning } P \mapsto D(P\|Q)$$

on pidevad, mistõttu on miinimum realiseeruv. Nüüd on selge, et kui $H(Q) < R$, siis $F(R, Q) > 0$, vastasel korral $F(R, Q) = 0$.

Alljärgnevas tõestame, et kui kodeeritav vektor X^n on iid vektor jaotusega Q, siis leidub universaalne $(n, 2^{nR})$ kood nii, et veatõenäosus

$$P_e(n) = \mathbf{P}(g_n(\mathcal{C}_n(X^n)) \neq X^n) \tag{84}$$

rahuldab

$$\liminf_n -\frac{1}{n} \log P_e(n) \geq F(R, Q). \tag{85}$$

Kui $R > H(Q)$, siis $F(R, Q) > 0$. Seosest (85) järeldub sellisel juhul, et iga $0 < \epsilon < F(R, Q)$ leidub n_o nii, et

$$P_e(n) \leq 2^{-n(F(R,Q)-\epsilon)} \to 0, \quad \text{kui } n > n_o.$$

See tulemus üldistab teoreemi 5.7 kahes suunas: esiteks näitab ta asümptootiliselt optimaalsete (nulliks koonduva veaga) *universaalse* koodi olemasolu. Teiseks tõestab ta, et viga koondub nulliks *eksponentsiaalselt*.

Teoreem 5.9 *Olgu $X^n = (X_1, \ldots, X_n)$ iid vektorid jaotusega Q. Iga R korral leidub universaalne $(n, 2^{nR})$ kood nii, et kehtib (85).*

Tõestus. Fikseerime R. Olgu

$$R_n := R - |\mathcal{X}|\frac{\log(n+1)}{n}.$$

Defineerime hulga A:

$$A := \{x^n \in \mathcal{X}^n : H(P_{x^n}) \le R_n\} = \bigcup_{P \in \mathcal{P}_n : H(P) \le R_n} T(P).$$

Hindame selle hulga elementide arvu. Selleks kasutame teoreemi 5.4 ning hinnangut tüübiklasside arvule.

$$\begin{aligned}
|A| &= \sum_{P \in \mathcal{P}_n : H(P) \le R_n} |T(P)| \\
&\le \sum_{P \in \mathcal{P}_n : H(P) \le R_n} 2^{nH(P)} \\
&\le \sum_{P \in \mathcal{P}_n : H(P) \le R_n} 2^{nR_n} \\
&\le (n+1)^{|\mathcal{X}|} 2^{nR_n} \\
&= 2^{n(R_n + |\mathcal{X}|\frac{\log(n+1)}{n})} \\
&= 2^{nR}
\end{aligned}$$

Seega on hulgas A ülimalt 2^{nR} sõna. Järjestame need. Kodeeriv funktsioon $\mathcal{C}_n$, nagu ikka, seab igale hulga A elemendile vastavusse tema järjekorranumbri, ülejäänud sõnad kodeeritakse suvaliselt. Leiame veatõenäosuse. Et $X_1, \ldots, X_n$ on iid jaotusega Q juhuslik vektor, siis

$$\begin{aligned}
P_e(n) = 1 - Q^n(A) &= \sum_{P \in \mathcal{P}_n : H(P) > R_n} Q^n(T(P)) \\
&\le (n+1)^{|\mathcal{X}|} \max_{P \in \mathcal{P}_n : H(P) > R_n} Q^n(T(P)) \\
&\le (n+1)^{|\mathcal{X}|} 2^{-n(\min_{P \in \mathcal{P}_n : H(P) > R_n} D(P\|Q))},
\end{aligned}$$

kus viimane võrratus tuleneb võrratustest (80). Seega

$$-\frac{1}{n}\log P_e(n) \ge -|\mathcal{X}|\frac{\log(n+1)}{n} + \min_{P \in \mathcal{P}_n : H(P) > R_n} D(P\|Q)).$$

Teoreem on tõestatud, kui veendume, et

$$\liminf_n \min_{P \in \mathcal{P}_n : H(P) > R_n} D(P\|Q) \ge F(R, Q).$$

Olgu

$$M_n := \min_{P:H(P)\geq R_n} D(P\|Q).$$

Kehtib

$$\min_{P\in\mathcal{P}_n:H(P)>R_n} D(P\|Q) \geq M_n.$$

Et $R_n \nearrow R$, siis iga jada

$$P_n \in \{P : H(P) \geq R_n\}$$

korral $P_n \to \{P : H(P) \geq R\}$ ($\mathbb{R}^{|\mathcal{X}|}$- meetrikas). Kujutise $P \mapsto D(P\|Q)$ pidevusest saame nüüd, et $M_n \to F(R,Q)$. ■

Tõestuseta märgime, et iga $(n, 2^{nR})$ koodi korral

$$\limsup_n -\frac{1}{n}\log P_e(n) \leq F(R,Q)$$

(vt *Han, Kobayashi*, Thm 3.18). Teoreemi 5.9 tõestuses leitud universaalne kood on seega selline, mille korral

$$\lim_n -\frac{1}{n}\log P_e(n) = F(R,Q)$$

ning paremini (suuremat piirväärtust) ei ole võimalik saada.

Märkus: Teoreemi 5.9 tõestuses konstrueeritud kood on universaalne selles mõttes, et etteantud R korral vektori X^n kodeerimiseks pole vaja teada jaotust Q. Küll aga sõltub kodeerimisprotseduur arvust R, sest sisuliselt kodeeriti vaid neid sõnu mille sagedustüübi entroopia on väiksem kui R. Kui $R > H(Q)$, on selline kodeerimine hea, sest viga koondub eksponentsiaalselt nulliks, vastasel korral aga mitte. Kodeerimaks tundmatu jaotusega genereeritud sõnu nii, et viga läheb kindlasti nulliks, tuleb järelikult valida $R > H(Q)$, mistõttu kirjeldatud kodeerimisprotseduur sõltub jaotusest Q läbi selle entroopia $H(Q)$.

5.3.2 Funktsioon $G(R,Q)$

Järgnevas üldistame teoreemi 5.8. Selleks defineerime

$$G(R,Q) := \min_{P:H(P)\leq R} D(P\|Q).$$

Funktsioon $G(R,Q)$ on positiivne, kui $R < H(Q)$; vastasel juhul $G(R,Q) = 0$. Seega on F positiivne seal kus $G = 0$ ja vastupidi. Ka funktsiooni G nimetatakse *reliability function*. Funktsioon G kirjeldab tõenäosuse

$$P_c(n) := \mathbf{P}\big(g_n(\mathcal{C}_n(X^n)) = X^n\big)$$

asümptootilist käitumist. Järgnev teoreem väidab, et suvalise koodi $(n, 2^{nR})$ koodi korral

$$\liminf_n -\frac{1}{n}\log P_c(n) \geq G(R,Q). \tag{86}$$

Seega, kui $G(R,Q) > 0$, siis iga $0 < \epsilon < G(R,Q)$ korral

$$P_c(n) \le 2^{-n(G(R,Q)-\epsilon)},$$

kui n on piisavalt suur. Saadud tulemus üldistab teoreemi 5.8 väites, et kui $R < H(Q)$, siis ka kõige parema $(n, 2^{nR})$ koodi veatõenäosus läheneb ühele *eksponentsiaalselt*.

Teoreem 5.10 *Olgu $X^n = (X_1, \ldots, X_n)$ iid vektorid jaotusega Q. Iga R ja iga $(n, 2^{nR})$ koodi korral kehtib (86).*

Tõestus. Tuletame meelde, et $\mathcal{X}^n$ jaguneb tüübiklassideks, $\mathcal{X}^n = \cup T(P)$. Defineerime

$$P_c(P) := Q^n\{x^n \in T(P) : g_n(\mathcal{C}_n(x^n)) = x^n\}.$$

Seega

$$P_c(n) = \sum_{P \in \mathcal{P}_n} P_c(T(P)).$$

Iga tüübiklassi $T(P)$ korral hindame tema osa $P_c(P)$. Õigesti dekodeeritakse maksimaalselt 2^{nR} sõna. Seega

$$|\{x^n \in T(P) : g_n(\mathcal{C}_n(x^n)) = x^n\}| \le 2^{nR} \wedge |T(P)|,$$

millest

$$P_c(P) \le \Big(\frac{2^{nR}}{|T(P)|} \wedge 1\Big)|T(P)|Q^n(x^n),$$

kus $Q^n(x^n)$ on tüübiklassi P kuuluva vektori tõenäosus.
Arvestades, et

$$\log Q^n(x^n) = -n(H(P) + D(P\|Q))$$

ning

$$nH(P) - |\mathcal{X}| \log(n+1) \le \log |T(P)| \le nH(P),$$

saame

$$\log\Big(\frac{2^{nR}}{|T(P)|} \wedge 1\Big) + \log|T(P)| + \log Q(x^n) \le \Big(nR + |\mathcal{X}|\log(n+1) - nH(P)\Big)^+ - nD(P\|Q).$$

Kokkuvõttes,

$$\log P_c(P) \le \max_{P \in \mathcal{P}_n} P_c(P) = \max_{P \in \mathcal{P}_n}\Big(\big(nR + |\mathcal{X}|\log(n+1) - nH(P)\big)^+ - nD(P\|Q)\Big).$$

Et on ülimalt $(n+1)^{|\mathcal{X}|}$ tüüpe

$$\log P_c(n) \le |\mathcal{X}|\log(n+1) + \max_{P \in \mathcal{P}_n}\Big(\big(nR + |\mathcal{X}|\log(n+1) - nH(P)\big)^+ - nD(P\|Q)\Big),$$

millest

$$\liminf_n -\frac{1}{n}\log P_c(n) \geq \liminf_n \min_{P\in\mathcal{P}_n}\Big(\big(H(P)-R\big)^+ + D(P\|Q)\Big) \geq \min_P\Big(\big(H(P)-R\big)^+ + D(P\|Q)\Big).$$

Paneme tähele,

$$\min_{P:H(P)\geq R}\Big((H(P)-R)^+ + D(P\|Q)\Big) = \min_{P:H(P)\geq R}\Big(H(P)-R+D(P\|Q)\Big)$$

ning

$$H(P)+D(P\|Q) = -\sum_x P(x)\log Q(x).$$

Viimane on P suhtes lineaarne funktsioon, mis savutab oma miinimumi hulga

$$\{P : H(P)\geq R\}$$

rajal, kus $H(P)=R$. Seega

$$\min_{P:H(P)\geq R}\Big(H(P)-R+D(P\|Q)\Big) = \min_{P:H(P)=R} D(P\|Q).$$

Kokkuvõttes

$$\min_P\Big(\big(H(P)-R\big)^+ + D(P\|Q)\Big) = \min_{P\leq R} D(P\|Q) = G(R,Q).$$

■

Kehtib ka vastupidine: leidub $(n, 2^{Rn})$ kood nii, et

$$\limsup_n -\frac{1}{n}\log P_c(n) \leq G(R,Q). \tag{87}$$

Sellise koodi korral

$$\lim_n -\frac{1}{n}\log P_c(n) = G(R,Q)$$

ning see kood on teatavas mõttes parim, sest väikese R korral kasvab selle koodi veatõenäosus väiksema võimaliku kuurusega. Selline kood on täpselt sama, mis garanteerib väikseima veatõenäosuse (teoreem 5.9). Veendume selles.

Teoreem 5.11 *Olgu $X^n = (X_1,\ldots,X_n)$ iid vektorid jaotusega Q. Iga R korral leidub universaalne $(n, 2^{nR})$ kood nii, et kehtib (87).*

Tõestus. Fikseerime R. Nagu teoreemi 5.9 tõestuses, olgu

$$R_n := R - |\mathcal{X}|\frac{\log(n+1)}{n}$$

ning hulk A:

$$A := \{x^n \in \mathcal{X}^n : H(P_{x^n}) \le R_n\} = \bigcup_{P \in \mathcal{P}_n : H(P) \le R_n} T(P).$$

Olgu $P_n \in \mathcal{P}_n$ selline, et $H(P_n) \le R_n$ ning

$$D(P_n||Q) = \min_{P \in \mathcal{P}_n : H(P) \le R_n} D(P||Q).$$

Seega $A \supset T(P_n)$ ja seosest (80):

$$P_c(n) \ge Q^n(A) \ge Q^n(T(P_n)) \ge (n+1)^{-|\mathcal{X}|} 2^{-nD(P_n||Q)}.$$

Kodeeriv $\mathcal{C}_n$, nagu teoreemis 5.9, kodeerib vaid hulka A kuuluvaid ning veatu kodeerimise tõenäosus $P_c(n)$ on vähemalt hulga A tõenäosus. Järelikult

$$-\frac{1}{n} \log P_c(n) \le \min_{P \in \mathcal{P}_n : H(P) \le R_n} D(P||Q) + |\mathcal{X}| \frac{\log(n+1)}{n}$$

ja teoreem järeldub sellest, et

$$\lim_n \min_{P \in \mathcal{P}_n : H(P) \le R_n} D(P||Q) = G(R, Q).$$

Viimane koondumine kehtib, sest $\cup_n \mathcal{P}_n$ on kõikjal tihe hulk. ■

5.4 Sanovi teoreem

5.4.1 Suurte hälvete seadused

Sagedustüüpide abil on lihtne tõestada tõenäosusteoorias väga olulisi nn *suurte hälvete seadusi*. Olgu $X_1, X_2, \ldots$ sõltumatud ja sama jaotusega keskväärtusega m juhuslikud suurused. Suurte arvude seadusest järeldub, et

$$\mathbf{P}\Big(|\frac{X_1 + \cdots + X_n}{n} - m| > \epsilon\Big) \to 0.$$

Oma kõige klassikalisemal kujul väidab suurte hälvete seadus, et nimetatud koondumine on (teatud tingimustel) eksponentsiaalne. Täpsemalt, leidub positiivne konstant $c(\delta)$ nii, et

$$\frac{1}{n} \log \mathbf{P}\Big(\frac{X_1 + \cdots + X_n}{n} > \delta + m\Big) \to -c(\delta),$$

millest (suure n korral)

$$\mathbf{P}\Big(\frac{X_1 + \cdots + X_n}{n} - m > \delta\Big) \approx 2^{-c(\delta)n}. \tag{88}$$

Järgnevas, nagu ikka, vaatleme jaotusi lõplikul tähestikul $\mathcal{X}$. Olgu $\mathcal{P}$ nende jaotuste hulk, $\mathcal{P}_n$ on endiselt kõigi sagedustüüpide hulk. Suvalise hulga $E \subset \mathcal{P}$ korral tähistame

$$d(E, Q) := \inf_{P \in E} D(P||Q).$$

Seega $d(E,Q)$ on jaotuse Q kaugus hulgast E K-L mõttes. Eelmises peatükis käsitletud funktsioonid $F(R,Q)$ ja $G(R,Q)$ avalduvad funktsioonina d, kus hulgaks E on vastavalt hulgad $\{P \in \mathcal{P} : H(P) \geq R\}$ ning $\{P \in \mathcal{P} : H(P) \leq R\}$.
Samuti tähistame

$$Q^n(E) := \sum_{P \in E \cap \mathcal{P}_n} Q^n(T(P)). \tag{89}$$

Teoreem 5.12 (Sanovi teoreem) *Olgu $X_1, \ldots, \mathcal{X}_n$ tähestikul $\mathcal{X}$ antud iid jaotusega Q juhuslikud suurused. Olgu $E \subset \mathcal{P}$. Siis*

$$Q^n(E) \leq (n+1)^{|\mathcal{X}|} 2^{-nd(E,Q)}. \tag{90}$$

Kui hulk E on oma sisemuse sulund, siis

$$\frac{1}{n} \log Q^n(E) \to -d(E,Q). \tag{91}$$

Tõestus. Tõestus põhineb võrratustel (80) ja hinnangul $|\mathcal{P}_n| \leq (n+1)^{|\mathcal{X}|}$.

$$\begin{aligned} Q^n(E) &= \sum_{P \subset E \cap \mathcal{P}_n} Q^n(T(P)) \\ &\leq \sum_{P \subset E \cap \mathcal{P}_n} 2^{-D(P||Q)} \\ &\leq \sum_{P \subset E \cap \mathcal{P}_n} \max_{P \subset E \cap \mathcal{P}_n} 2^{-D(P||Q)} \\ &\leq (n+1)^{|\mathcal{X}|} 2^{-d(E \cap \mathcal{P}_n, Q)} \\ &\leq (n+1)^{|\mathcal{X}|} 2^{-d(E,Q)}, \end{aligned}$$

sest

$$d(E \cap \mathcal{P}_n, Q) = \inf_{P \in E \cap \mathcal{P}_n} D(P||Q) \geq \inf_{P \in E} D(P||Q) = d(E,Q).$$

Teise väite tõestuseks paneme tähele, et $\cup_n \mathcal{P}_n$ on kõikjal tihe hulgas $\mathcal{P}$. Vastavalt eeldusele on hulgal E mittetühi sisemus. Sellest järeldub, et hulk

$$E \cap (\cup_n \mathcal{P}_n)$$

on kõikjal tihe hulgas E ehk iga $P \in E$ korral leidub $P_n \in (\cup_n \mathcal{P}_n)$ nii, et $P_n \to P$ ($\mathbb{R}^{|\mathcal{X}|}$ meetrikas). Üldisust kitsendamata võib valida selle jada nii, et $P_n \in \mathcal{P}_n$ iga piisavalt suure n korral. Et E on kompaktne (kinnine), leidub $P^* \in E$ nii, et $D(P^*||Q) = d(E,Q)$. Seega leidub $P_n \in \mathcal{P}_n \cap E$ nii, et $P_n \to P^*$, millest ka $D(P_n||Q) \to D(P^*||Q) = d(E,Q)$. Nüüd

$$Q^n(E) \geq Q^n(T(P_n)) \geq (n+1)^{-|\mathcal{X}|} 2^{-D(P_n||Q)},$$

millest

$$\liminf_n \frac{1}{n} \log Q^n(E) \geq \liminf_n \Big(-\frac{|\mathcal{X}| \log(n+1)}{n} - D(P_n||Q) \Big) = -D(P^*||Q) = -d(E,Q).$$

■

Kuidas ülaltoodud teoreemist järelduvad suurte hälvete seadused? Olgu $g : \mathcal{X} \to \mathbb{R}$ mingi funktsioon. Suurte hälvete seadused käsitlevad tõenäosusi

$$Q^n\Big(x^n \in \mathcal{X}^n : \frac{1}{n}\sum_{j=1}^{n} g(x_j) \geq c\Big). \tag{92}$$

Kui g on samasusfunktsioon, $c = (m+\delta)$ ning $X_i \sim Q$, on (92) võrdne tõenäosusega (88). Sanovi teoreemi rakendused põhinevad asjaolul, et vektori x^n kuulumine hulka

$$\Big\{x^n \in \mathcal{X}^n : \frac{1}{n}\sum_{j=1}^{n} g(x_j) \geq c\Big\}$$

sõltub vaid tema tüübist. Tõepoolest, et

$$\frac{1}{n}\sum_{j=1}^{n} g(x_j) = \sum_{a\in\mathcal{X}} g(a)P_{x^n}(a),$$

siis

$$\frac{1}{n}\sum_{j=1}^{n} g(x_j) \geq c \Leftrightarrow \sum_{a\in\mathcal{X}} g(a)P_{x^n}(a) \geq c \Leftrightarrow P_{x^n} \in E \cap \mathcal{P}_n,$$

kus

$$E := \{P : \sum_{a\in\mathcal{X}} g(a)P(a) \geq c\}. \tag{93}$$

Seega tõenäosus (92) on

$$\sum_{P\in E\cap\mathcal{P}_n} Q^n(T(P)) = Q^n(E),$$

kus E on (93).

5.4.2 Sanovi teoreemi rakendamine

Sanovi teoreemi rakendamiseks suurte hälvete seadustes on oluline konstant $d(E, Q)$, kus E on kujul (93) olev hulk. See hulk on kinnine, mistõttu leidub $P^* \in E$ nii, et

$$D(P^*||Q) = d(E, Q) = \min_{P\in E} D(P||Q).$$

Teades jaotust P^* on kerge arvutada konstanti $d(E, Q)$. Järgnev lemma annab jaotusele P^* üldise kuju.

Lemma 5.2 *Olgu P^* hulgal olev jaotus $\mathcal{X}$, mis avaldub kujul*

$$P^*(a) = \frac{Q(a)2^{\lambda g(a)}}{\sum_{a\in\mathcal{X}} Q(a)2^{\lambda g(a)}}, \tag{94}$$

kus λ on selline, et

$$\sum_{a\in\mathcal{X}} g(a)P^*(a) = c. \tag{95}$$

Siis

$$D(P^*||Q) = \min_{P\in E} D(P||Q) = d(E,Q),$$

kus E on (93).

Tõestus. Olgu $P \in E$ suvaline jaotus. Näitame, et

$$D(P||Q) \geq D(P^*||Q) + D(P||P^*) \geq D(P^*||Q). \tag{96}$$

Olgu $P \in E$, st $\sum_{a\in\mathcal{X}} g(a)P(a) \geq c$. Leiame

$$\begin{aligned} D(P||Q) &= \sum_a P(a)\log\frac{P(a)}{Q(a)} \\ &= \sum_a P(a)\log\frac{P(a)}{P^*(a)} + \sum_a P(a)\log\frac{P^*(a)}{Q(a)} \\ &= D(P||P^*) + \sum_a P(a)\log\frac{P^*(a)}{Q(a)}. \end{aligned}$$

Hindame

$$\sum_a P(a)\log\frac{P^*(a)}{Q(a)}.$$

Olgu

$$C := \sum_a Q(a)2^{\lambda g(a)},$$

millest

$$\frac{P^*(a)}{Q(a)} = \frac{2^{\lambda g(a)}}{C}.$$

Siis

$$\begin{aligned} \sum_a P(a)\log\frac{P^*(a)}{Q(a)} &= \sum_a \Big(\lambda g(a) - \log C\Big)P(a) \\ &= \lambda\sum_a g(a)P(a) - \log C \\ &\geq \lambda c - \log C \\ &= \sum_a \Big(\lambda g(a) - \log C\Big)P^*(a) \\ &= \sum_a P^*(a)\log\frac{P^*(a)}{Q(a)} \\ &= D(P^*||Q), \end{aligned}$$

kus võrratus tuleneb sellest, et

$$\sum_a g(a)P(a) \geq c.$$

■

Märkus: Seos (96) on sisuliselt Pythagorase teoreem ja näitab, et K-L kaugus käitub nagu eukelidilise kauguse ruut.

Näide 1 (mündivisked): Olgu $\mathcal{X} = \{0, 1\}$, $Q = B(1, q)$. Hindame tõenäosust

$$\mathbf{P}\big(\frac{S_n}{n} \geq c\big),$$

kus $S_n = X_1 + \cdots + X_n$ ja $q < c < 1$. Hulk

$$E = \{P : \sum_a P(a)a = P(1) \geq c\} = \{B(1, p) : p \geq c\}.$$

Seega ainus tingimust (102) rahuldav jaotus on $P^* = B(1, c)$ ning

$$D(P^*||Q) = c \log \frac{c}{q} + (1 - c) \log \frac{1 - c}{1 - q}.$$

Näiteks $q = \frac{1}{2}$, siis $D(P^*||Q) = 1 - h(c)$. Kui $c = 0.7$, siis, $1 - h(c) = 1 - h(0.7) = 0.119$. Seega, kui münti visatakse 1000 korda, siis tõenäosus, et kirjade arv on enam kui 700, on ligikaudu

$$2^{-(1000(1-h(0.7))} = 2^{-119}.$$

Range võrratuse saame tõkkest (90):

$$\mathbf{P}\big(S_n \geq 700\big) \leq (1001)2^{-119} < 2^{10-119} = 2^{-109},$$

kus S_n on kirjade arv 1000 mündiviske korral.

Näide 2 (täringuvisked): Kui suur on tõenäosus, et n täringuviskel saadud keskmine on vähemalt 4?
Hulk E on järgmine

$$E = \{P : \sum_{i=1}^{6} iP(i) \geq 4\},$$

jaotus Q on ühtlane üle tähestiku $\{1, 2, 3, 4, 5, 6\}$. Meid huvitab $Q^n(E)$. Sanovi teoreem:

$$\frac{1}{n} \log Q^n(E) \to -D(P^*||Q),$$

kus

$$P^*(i) \propto 2^{\lambda i}, \quad \sum_{i=1}^{6} iP^*(i) = 4.$$

Numbriliselt lahendades saame $\lambda = 0.2519$, millest P^* on järgmine

1	2	3	4	5	6
0.1031	0.1227	0.1467	0.174	0.2072	0.2468

Nüüd $D(P^*||Q) = \sum_{i=1}^{6} P^*(i) \log 6P^*(i) = \log 6 - H(P^*) = 0.0624$. Seega

$$\frac{1}{n} \log Q^n(E) \to -0.0624.$$

Kui $n = 10000$, siis

$$Q^n(E) \approx 2^{-624}.$$

Range võrratus

$$Q^n(E) \leq (n+1)^5 2^{-nD(P^*||Q)} \leq 2^{67-624}.$$

Vaatleme, millise tõkke annab toodud näidetele tõenäosusteooriast tuntud Höffdingi võrratus.

Teoreem 5.13 (Höffdingi võrratus)
Olgu $X_1, \ldots, X_n$ sõltumatud tõkestatud juhuslikud suurused, kusjuures $a_i \leq X_i \leq b_i$ p.k. Siis iga $c > 0$ korral kehtivad võrratused

$$\mathbf{P}\big(S_n - ES_n \geq c\big) \leq \exp\Big[-\frac{2c^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\Big]. \tag{97}$$

Näide 1 (mündivisked): Olgu $\mathcal{X} = \{0, 1\}$, $Q = B(1, 0.5)$. Sellisel juhul $a_i = 0, b_i = 1$ ja (97) on kujul

$$\mathbf{P}\big(S_n - ES_n \geq c\big) \leq \exp[-\frac{2c^2}{n}]. \tag{98}$$

Seega tõenäosus, et 1000 mündiviskest tuleb vähemalt 700 kirja on ülalt tõkestatud järgmiselt

$$\mathbf{P}\big(S_{1000} \geq 700\big) = \mathbf{P}\big(S_{1000} - 500 \geq 200\big) \leq e^{-\frac{80000}{1000}} = e^{-80} = 2^{-\frac{80}{\ln 2}} \leq 2^{-115}.$$

Võrdleme: Sanovi teoreem andis parimaks eksponendiks -119. Seega selle näite korral Höffdingi võrratus annab praktiliselt parima võimaliku hinnangu. Paneme tähele, et Höffdingi avõrratuse abil saame range hinnangu $\mathbf{P}\big(S_{1000} \geq 700\big) \leq 2^{-115}$, mis on isegi parem, kui sanovi teoreemi kasutades saadud range hinnang $\mathbf{P}\big(S_{1000} \geq 700\big) \leq 2^{-109}$.

Näide 2 (täringuvisked): Siin $X_1, \ldots, X_n$ on iid ühtlase jaotusega. Nüüd $a_i = 1, b_i = 6$ ning

$$\mathbf{P}\big(S_n \geq (3.5 + \epsilon)n\big) \leq e^{\frac{-2n\epsilon^2}{25}}.$$

Seega tõenäosust, et 10000 täringuviske keskmine on vähemalt 4 saame ülalt hinnata järgmiselt

$$\mathbf{P}\big(\frac{S_n}{n} \geq 3.5 + 0.5\big) \leq e^{\frac{-5000}{25}} = e^{-200} = 2^{\frac{-200}{\ln 2}} \approx 2^{-288}.$$

Meenutagem, et Sanovi teoreem andis parimaks eksponendiks -624. Seega Höffdingi võrratuse abil saadud hinnang on kaugel optimaalsusest. Sanovi teoreemist saame hinnanguks $2^{67-624} << 2^{-288}$.

5.4.3 Lihthüpoteeside kontroll ja Sanovi teoreem

Olgu $X_1, \ldots, X_n$ iid jaotusega Q juhuslikud suurused. Vaatleme *lihthüpoteese*:

$$H_0 : Q = P_0$$
$$H_1 : Q = P_1$$

Olgu $x^n = x_1, \ldots, x_n$ valim ja $g : \mathcal{X}^n \to \{0, 1\}$ test hüpoteeside kontrollimiseks: hüpoteesi H_i võtame vastu parajasti siis, kui $g(x^n) = i$. Seega test g on funktsioon kujul I_{A_n}, kus $A_n \subset \mathcal{X}^n$ on H_1 vastuvõtmise piirkond. Teisisõnu: kui $x^n \in A_n$, võtame vastu hüpoteesi H_1, kui $x^n \in A_n^c$, võtame vastu hüpoteesi H_0.

Hüpoteeside kontrollimisel tehtud vead on kahte liiki:

$$\alpha := \mathbf{P}(g(X_1, \ldots, X_n) = 1 | H_0) = P_0^n(A_n), \quad \beta := \mathbf{P}(g(X_1, \ldots, X_n) = 0 | H_1) = P_1^n(A_n^c).$$

Viga α nimetatakse tihti esimest liiki veaks ja viga β teist liiki veaks, lihthüpoteeside korral on nad enamsti võrdväärsed.

Üldiselt püüame leida hulka A_n nii, et mõleamd veatõenäosused oleksid minimaalsed. samas ühe minimeerimine toob enesega kaasa teise suurenemise. Näiteks kui mõõtude P_0 ja P_1 kandjad langevad kokku, siis $\alpha = 0$ parajasti siis kui $\beta = 1$. Järgnev Neyman-Pearsoni lemma väidab, et optimaalse testi annab *tõepärasuhete statistiks*.

Teoreem 5.14 *(***Neyman-Pearsoni lemma:***) Olgu $T > 0$,*

$$A_n(T) := \left\{ x^n : \frac{P_1(x^n)}{P_0(x^n)} > T \right\} \tag{99}$$

ning

$$\alpha^* = P_0^n(A_n(T)), \quad \beta^* = P_0^n(A_n^c(T)).$$

Olgu $B \subset \mathcal{X}^n$ mingi teine test veatõenäosustega $\alpha = P_0^n(B)$, $\beta = P_0^n(B^c)$. Siis kehtib võrratus

$$(\beta - \beta^*) + T(\alpha - \alpha^*) \geq 0. \tag{100}$$

Tõestus. Iga x^n korral

$$\big(I_{A_n^c}(x^n) - I_{B^c}(x^n)\big)\big(TP_0(x^n) - P_1(x^n)\big) \geq 0.$$

Summeerides üle x^n, saame (veendu)

$$T\big(P_0^n(A_n^c) - P_0^n(B^c)\big) - P_1^n(A_n^c) + P_1(B^c) = T\big((1 - \alpha^*) - (1 - \alpha)\big) - \beta^* + \beta \geq 0.$$

Viimasest järeldub teoreemi väide. ■

Neyman-pearsoni lemmast järeldub, et kui leidub B mille esimest liiki viga on väiksem

kui $\alpha*$, st $\alpha < \alpha^*$, siis selle testi teist liiki viga peab ilmtingimata olema suurem kui β^*. Loomulikult kehtib ka vastupidine, kui $\beta < \beta^*$, siis $\alpha > \alpha^*$.
Tihti on esimest liiki veale antud ülemine piir p. Neyman-Pearsoni lemma väidab, et tõepärasuhete statistik, mille korral $\alpha^* = p$ minimiseerib teist liiki viga üle kõikide selliste testide mille korral esimest liiki viga on maksimaalselt p. Statistikas öeldakse et selline test on *võimsaim*.

Järgnevas veendume, et tõepärasuhete test pole sisuliselt midagi muud, kui tüübile P_{x^n} K-L mõttes lähima jaotuse valik. Selleks vaatleme *logaritmilist tõepärasuhet*:

$$\begin{aligned} L(x^n) =& \log \frac{P_1^n(x^n)}{P_0^n(x^n)} = \log \frac{P_1^n(x_1,\dots,x_n)}{P_0^n(x_1,\dots,x_n)} = \sum_{i=1}^n \log \frac{P_1(x_i)}{P_0(x_i)} \\ =& \sum_{a\in\mathcal{X}} \log\Big(\frac{P_1(a)}{P_o(a)}\Big) P_{x^n}(a) n = n \sum_{a\in\mathcal{X}} P_{x^n}(a) \log\Big(\frac{P_1(a)}{P_0(a)} \frac{P_{x^n}(a)}{P_{x^n}(a)}\Big) \\ =& n \sum_{a\in\mathcal{X}} P_{x^n}(a) \Big(\log \frac{P_{x^n}(a)}{P_0(a)} - \log \frac{P_{x^n}(a)}{P_1(a)} \Big) = n\big(D(P_{x^n}||P_0) - D(P_{x^n}||P_1)\big). \end{aligned}$$

Järelikult

$$\begin{aligned} \frac{P_1^n(x^n)}{P_0^n(x^n)} > T \;\Leftrightarrow\; \log \frac{P_1^n(x^n)}{P_0^n(x^n)} &= n\big(D(P_{x^n}||P_0) - D(P_{x^n}||P_1)\big) > \log T \\ \Leftrightarrow\; D(P_{x^n}||P_1) &< D(P_{x^n}||P_0) - \frac{\log T}{n}. \end{aligned}$$

Juhul, kui $T = 1$, saame

$$\begin{aligned} H_0 \quad &\Leftrightarrow \quad D(P_{x^n}||P_0) \le D(P_{x^n}||P_1) \\ H_1 \quad &\Leftrightarrow \quad D(P_{x^n}||P_1) < D(P_{x^n}||P_0). \end{aligned}$$

Uurime logaritmilise tõepärasuhete statistiku kasutamisel tehtavate vigade asümptootikat. Olgu need vead

$$\alpha_n(T) := P_0^n(A_n(T)), \quad \beta_n(T) := P_1^n(A_n^c(T)),$$

kus

$$A_n(T) = \{x^n : \frac{1}{n} L(x^n) = D(P_{x^n}||P_0) - D(P_{x^n}||P_1) > \frac{\log T}{n}\}.$$

Edaspidi vaatleme T fikseerituna ja jätame tähistustrest välja. Suurte arvude seadusest on kerge järeldada, et kui $X_1,\dots,X_n$ on iid juhuslikud suurused jaotusega P_1 või P_0, siis mõlemad vead lähevad nulliks, st kehtivad järgmised koondumised (vaata ülesanne 1):

$$\alpha_n \to 0, \quad \beta_n \to 0. \tag{101}$$

Sanovi teoreemi abil on võimalik hinnata aga vigade nulliks koondumise kiirust. See on võimalik seetõttu, et iga valimi x^n kuulumine hulka A_n sõltub vaid vektori x^n tüübist

P_{x^n}. Et kasutada Sanovi teoreemi, peame leidma tõenäosusmõõtude hulga E_n $|\mathcal{X}|$-dimensionaalses simpleksis $\mathcal{P}$ nii, et

$$P_{x^n} \in E_n \quad \Leftrightarrow \quad x^n \in A_n.$$

Antud juhul

$$E_n = \{P \in \mathcal{P} : \sum_a P(a) \log \frac{P_1(a)}{P_0(a)} = D(P||P_0) - D(P||P_1) > \frac{\log T}{n}\}.$$

Paneme tähele, et kui n on piisavalt suur, siis $P_1 \in E_n$ ja $P_0 \in E_n^c$. Hulga E_n raja on

$$\partial E_n = \Big\{P \in \mathcal{P} : \sum_a P(a) \log \frac{P_1(a)}{P_0(a)} = \frac{\log T}{n}\Big\}.$$

Et P_1 ja P_0 on fikseeritud, moodustab raja hüpertasand kujul

$$H = \{(p_1, \ldots, p_{|\mathcal{X}|}) \in \mathbb{R}^{|\mathcal{X}|} : \sum_i p_i r_i = c\}.$$

Sellest järeldub, et nii hulgad E_n kui ka E_n^c on kumerad.
Seega, kasutades definitsiooni (89)

$$\alpha_n = P_0^n(A_n(T)) = P_0^n(E_n), \quad \beta_n = P_1^n(A_n^c(T)) = P_1^n(E_n^c).$$

Olgu P_1' mõõdu P_1 parim lähend K-L mõttes hulgast E_n^c ning olgu P_0' mõõdu P_0 parim lähend K-L mõttes hulgast $\overline{E_n}$. Seega

$$D(P_1'||P_1) = \min_{P \in E_n^c} D(P||P_1) = d(E^c, P_1), \quad D(P_0'||P_0) = \min_{P \in \overline{E}_n} D(P_0'||P_0) = d(\overline{E}_n, P_0).$$

Sanovi teoreem (ülemine hinnang kehtib suvaliste hulkade korral):

$$\alpha_n = P_0^n(E_n) \le (n+1)^{|\mathcal{X}|} 2^{-D(P_0'||P_0)}, \quad \beta_n = P_1^n(E_n^c) \le (n+1)^{|\mathcal{X}|} 2^{-D(P_1'||P_1)}.$$

Arvestades hulkade E_n ja E_n^c ilusat kuju, on Lagrange'i kordajate abil võimalik näidata, et

$$P_0' = P_1' \propto P_0^\lambda P_1^{1-\lambda} =: P_\lambda,$$

kus λ on selline, et

$$D(P_\lambda||P_0) - D(P_\lambda||P_1) = \frac{\log T}{n}. \tag{102}$$

Teisisõnu, iga $a \in \mathcal{X}$ korral

$$P_0'(a) = P_1'(a) = \frac{P_0(a)^\lambda P_1(a)^{1-\lambda}}{C(\lambda)},$$

kus

$$C(\lambda) = \sum_{a \in \mathcal{X}} P_0(a)^{\lambda} P_1(a)^{1-\lambda} \tag{103}$$

on normaliseeriv konstant ja λ on valitud nii, et kehtib (102).

Erijuhul $T = 1$ E_n ei sõltu arvust n, st

$$E_n = E := \big\{P \in \mathcal{P} : D(P||P_0) > D(P||P_1)\big\}.$$

Sellisel juhul P_λ ei sõltu valimi mahust n ning λ, olgu see λ^*, on selline, et

$$D(P_{\lambda^*}||P_0) = D(P_{\lambda^*}||P_1) =: D.$$

Et E ja E^c on mõlemad kumerad, kehtib ka Sanovi teoreemi alumine hinnang, i.e.

$$\frac{1}{n} \log \alpha_n \to -D, \quad \frac{1}{n} \log \beta_n \to -D. \tag{104}$$

Millise testi korral on väikesem esimest ja teist liiku vigade summa $\alpha_n + \beta_n$? Neyman-Pearsoni lemmast järeldub, et selline tast on tõepärasuhete test, kus $T = 1$ (ülesanne 2). Koondumistest (104) saame, et sellisel juhul suure n korral kehtivad

$$\alpha_n = P_0^n(A_n(1)) \approx 2^{-Dn}, \quad \beta_n = P_1^n(A_n^c(1)) \approx 2^{-Dn}.$$

Seega sümmeetrilisel juhul mõlemad vead kahanevad ühe ja sama kiirusega 2^{-Dn}.
Saab näidata, et D avaldub järgmiselt:

$$D = -\min_{0 \le \lambda \le 1} \log C(\lambda), \tag{105}$$

kus $C(\lambda)$ on defineeritud seosega (103) (Ülesanne 3). Arvu D, nimetatakse *Tŝernovi informatsiooniks*. Seega Tŝernovi informatsioon D määrab kiiruse, millega koondub esimest ja teist liiku vea keskmine tõepärasuhete testil konstandiga 1:

$$-\frac{1}{n} \log \frac{\alpha_n + \beta_n}{2} \to D.$$

Teisisõnu, suure n korral

$$\frac{\alpha_n + \beta_n}{2} \approx 2^{-Dn}. \tag{106}$$

Seos (106) järeldus Sanovi teoreemist. Sanovi teoreem on asümptootiliselt täpne, konkreetse n korral võib ta anda liiga jämeda hinnangu. Pole aga raske näidata, et valemis (106) võib aga märgi $\approx$ asendada märgiga $\le$. Alljärgnevas teemegi seda. Veel enam, keskmise $\frac{\alpha_n+\beta_n}{2}$ asemel vaatleme testi, mis minimiseerib (üldisema) kaalutud keskmise $\pi\alpha_n + (1-\pi)\beta_n$. Selline viga tekib juhul, kui π ja $1-\pi$ on nn. *eelmõõt*, mis postuleerib meie eelarvamuse hüpoteeside õigsusest. Sellise eelarvamuse olemasolu tuntakse Bayesi lähenemisena.

Bayesi lähenemine: $X_1, \ldots, X_n$ olgu iid juhuslikud suurused jaotusega Q, kus

$$\mathbf{P}(Q = P_0) = \pi, \quad \mathbf{P}(Q = P_1) = 1 - \pi.$$

Eesmärk on Q hindamine. Test olgu $g_n = I_{A_n}$ ning seejuures tehtav viga

$$\begin{aligned} P_e &= \mathbf{P}(g_n(X_1, \ldots, X_n) \neq Q) \\ &= \mathbf{P}(g_n(X_1, \ldots, X_n) = 1 | Q = 0)\pi + \mathbf{P}(g_n(X_1, \ldots, X_n) = 0 | Q = 1)(1 - \pi) \\ &= P_0^n(A_n)\pi + P_1^n(A_n^c)(1 - \pi). \end{aligned}$$

Pole raske veenduda, et parim test A_n, mis minimimiseerib summa

$$P_0^n(A_n)\pi + P_1^n(A_n^c)(1 - \pi)$$

on tõepärasuhete statistik konstandiga $T = \frac{\pi}{1-\pi}$ (ülesanne 2). Vaatleme sellise testi esimest ja teist tüüpi vigade kaalutud keskmist.

Väide 5.1 *Olgu*

$$\alpha_n = P_0^n\Big(A_n\big(\frac{\pi}{1-\pi}\big)\Big), \quad \beta_n = P_1^n\Big(A_n^c\big(\frac{\pi}{1-\pi}\big)\Big).$$

Siis

$$\pi\alpha_n + (1 - \pi)\beta_n \leq 2^{-Dn}, \tag{107}$$

kus D on Tŝernovi informatsioon, st

$$D = -\min_{0 \leq \lambda \leq 1} \log\Big(\sum_{a \in \mathcal{X}} P_0(a)^\lambda P_1(a)^{1-\lambda}\Big).$$

Tõestus. Seos (107) on tõestatud, kui näitame, et

$$\pi\alpha_n + (1 - \pi)\beta_n \leq \min_{0 \leq \lambda \leq 1} \Big(\sum_{a \in \mathcal{X}} P_0(a)^\lambda P_1(a)^{1-\lambda}\Big)^n. \tag{108}$$

Et $x^n \in A_n\big(\frac{\pi}{1-\pi}\big)$ parajasti siis, kui $\pi P_0^n(x^n) < (1 - \pi)P_1^n(x^n)$, kehtib

$$\begin{aligned} \pi\alpha_n + (1 - \pi)\beta_n &= \pi P_0^n\Big(A_n\big(\frac{\pi}{1-\pi}\big)\Big) + (1 - \pi)\Big(A_n^c\big(\frac{\pi}{1-\pi}\big)\Big) \\ &= \sum_{x^n} \min\{\pi P_0^n(x^n), (1 - \pi)P_1^n(x^n)\}. \end{aligned}$$

Iga $a, b \geq 0$ ja $\lambda \in (0,1)$ korral kehtib $\min\{a,b\} \leq a^\lambda b^{1-\lambda}$, millest iga $\lambda \in (0,1)$ korral kehtib

$$\begin{aligned}
\sum_{x^n} \min\{\pi P_0^n(x^n), (1-\pi)P_1^n(x^n)\} &\leq \sum_{x^n} \big(\pi P_0^n(x^n)\big)^\lambda \big((1-\pi)P_1^n(x^n)\big)^{1-\lambda} \\
&\leq \sum_{x^n} \big(P_0^n(x^n)\big)^\lambda \big(P_1^n(x^n)\big)^{1-\lambda} \\
&= \sum_{x^n} \prod_{i=1}^n P_0^\lambda(x_i) \prod_{i=1}^n P_1^{1-\lambda}(x_i) \\
&= \sum_{x^n} \prod_{i=1}^n P_0^\lambda(x_i) P_1^{1-\lambda}(x_i) \\
&= \Big(\sum_{a \in \mathcal{X}} P_0^\lambda(a) P_1^{1-\lambda}(a)\Big)^n.
\end{aligned}$$

■

5.5 Tugevalt tüüpilised sõnad

Vaatleme ikka olukorda, kus $X_1, \ldots, X_n$ on iid valim jaotusest P, mis on antud lõplikul tähestikul $\mathcal{X}$. Tuletame meelde, et $x^n \in \mathcal{X}^n$ on nõrgalt ϵ-tüüpiline, kui

$$2^{-n(H(P)+\epsilon)} \leq P^n(x^n) \leq 2^{-n(H(P)-\epsilon)}$$

mis on ekvivalentne tungimusega

$$|\frac{1}{n} \log P^n(x^n) + H(P)| \leq \epsilon. \tag{109}$$

Nõrgalt ϵ-tüüpilisi sõnade hulka tähistame W_ϵ^n. Nõrk AEP-teoreem väitis järgmist:

1. kui $x^n \in W_\epsilon^n$, siis kehtib (109);
2. iga $\delta > 0$ korral $P^n(W_\epsilon^n) \geq 1 - \delta$, kui n on piisavalt suur;
3. iga $\delta > 0$ korral
$$(1-\delta)2^{n(H(P)-\epsilon)} \leq |W_\epsilon^n| \leq 2^{n(H(P)+\epsilon)},$$
kui n on piisavalt suur.

Põhimõte:

- x^n on nõrgalt tüüpiline, kui $P^n(x^n) \approx 2^{-nH(P)}$;
- x^n on tugevalt tüüpiline, kui $P_{x^n}(a) \approx P(a)$ iga $a \in \mathcal{X}$.

Definitsioonid on kergelt erinevad.

Def 5.15 *Sõna $x^n \in \mathcal{X}^n$ on tugevalt ϵ-tüüpiline, kui kehtib:*

1. $P_{x^n}(a) = 0$, *kui* $P(a) = 0$;
2. $\sum_{a\in\mathcal{X}} |P_{x^n}(a) - P(a)| \le \epsilon$.

Def 5.16 *Sõna $x^n \in \mathcal{X}^n$ on tugevalt ϵ-tüüpiline, kui kehtib:*

1. $P_{x^n}(a) = 0$ *kui* $P(a) = 0$;
2. $|P_{x^n}(a) - P(a)| \le \frac{\epsilon}{|\mathcal{X}|}$, *kui* $P(a) > 0$.

Def 5.17 *Sõna $x^n \in \mathcal{X}^n$ on tugevalt ϵ-tüüpiline, kui iga $a \in \mathcal{X}$ korral kehtib*

$$|P_{x^n}(a) - P(a)| \le \frac{\epsilon P(a)}{|\mathcal{X}|}. \tag{110}$$

Seosest (110) järeldub,

- $P_{x^n}(a) = 0$ kui $P(a) = 0$;
- kui $\epsilon < |\mathcal{X}|$, siis $P(a) > 0$ tähendab seda, et $N_{x^n}(a) \ge 1$.

On selge, et Def. 5.17 järeldub Def. 5.16 ja sellest omakorda järeldub Def. 5.15. Seega viimane definitsioon on kõige tugevam. Teisest küljest võttes definitsioonis 5.15 ϵ rolli

$$\epsilon' = \min_{a:P(a)>0} \min \frac{\epsilon P(a)}{|\mathcal{X}|},$$

saame definitsiooni 5.17. Seega sisuliselt on kõik kolm definitsiooni ekvivalentsed.

Alljärgnevas kasutame definitsiooni 5.17 ja defineerime tugevalt ϵ-tüüpiliste sõnade hulga:

$$T_\epsilon^n := \{x^n \in \mathcal{X}^n : |P_{x^n}(a) - P(a)| \le \frac{\epsilon P(a)}{|\mathcal{X}|}, \quad \forall a \in \mathcal{X}\}.$$

Järgnev teoreem näitab hulgal T_ϵ^n on suure n korral samasugused omadused kui hulgal W_ϵ^n, st ta mõõt on ligikaudu 1, sõnad on ligikaudu võrdtõenäosused ja sinna kuuluvate sõnade arv on ligikaudu $2^{nH(p)}$.

Teoreem 5.18 *(Tugev AEP)*

1. *kui $x^n \in T_\epsilon^n$, siis kehtib (109);*
2. *iga $\delta > 0$ korral $P^n(T_\epsilon^n) \ge 1 - \delta$, kui n on piisavalt suur;*
3. *iga $\delta > 0$ korral*
$$(1 - \delta)2^{n(H(P)-\epsilon)} \le |T_\epsilon^n| \le 2^{n(H(P)+\epsilon)},$$
kui n on piisavalt suur.

Tõestus. Et

$$P^n(x^n) = \prod_{a\in\mathcal{X}} P(a)^{N_{x^n}(a)},$$

siis

$$\frac{1}{n}\log P^n(x^n) = \sum_{a\in\mathcal{X}} P_{x^n}(a)\log P(a),$$

millest

$$\begin{aligned}|\frac{1}{n}\log P^n(x^n) + H(P)| &= |\sum_{a\in\mathcal{X}} \log P(a)(P_{x^n}(a) - P(a))| = |\sum_a (P(a) - P_{x^n}(a))(-\log P(a))| \\ &\le \sum_a |(P(a) - P_{x^n}(a)|(-\log P(a)) \le \sum_a \frac{\epsilon}{|\mathcal{X}|} P(a)(-\log P(a)) \\ &= \frac{\epsilon}{|\mathcal{X}|} H(P) \le \epsilon.\end{aligned}$$

Teise väite tõestus põhineb suurte arvude seadusel. Kõigepealt paneme tähele, et kui $P(a) = 0$, siis iga n korral.

$$\mathbf{P}\big(|P_{x^n}(a) - P(a)| > \epsilon\big) = \mathbf{P}\big(|P_{x^n}(a)| > \epsilon\big) = 0. \tag{111}$$

Kui $P(a) > 0$, kasutame NSAS: iga $\epsilon > 0$ korral kehtib

$$\mathbf{P}\big(|P_{x^n}(a) - P(a)| > \epsilon\big) \to 0. \tag{112}$$

Seostest (111) ja (112) saame, et iga a, $\epsilon > 0$ ja $\delta > 0$ korral leidub $N(\epsilon, a, \delta)$ nii, et

$$\mathbf{P}\big(|P_{x^n}(a) - P(a)| > \frac{\epsilon P(a)}{|\mathcal{X}|}\big) \le \frac{\delta}{|\mathcal{X}|}.$$

Seega, kui $n > \max_a N(\epsilon, a, \delta)$, siis ülaltoodust järeldub, et

$$\mathbf{P}\big(\exists a: \quad |P_{x^n}(a) - P(a)| > \frac{\epsilon P(a)}{|\mathcal{X}|}\big) = \mathbf{P}\big(\cup_a \{|P_{x^n}(a) - P(a)| > \frac{\epsilon P(a)}{|\mathcal{X}|}\}\big) \le \delta,$$

millest

$$\mathbf{P}\big(|P_{x^n}(a) - P(a)| \le \frac{\epsilon P(a)}{|\mathcal{X}|}, \quad \forall a \in \mathcal{X}\big) = \mathbf{P}\big(\cap_a \{|P_{x^n}(a) - P(a)| \le \frac{\epsilon P(a)}{|\mathcal{X}|}\}\big) \ge 1 - \delta.$$

Et

$$\mathbf{P}\big(|P_{x^n}(a) - P(a)| \le \frac{\epsilon P(a)}{|\mathcal{X}|}, \quad \forall a \in \mathcal{X}\big) = P^n(T^n_\epsilon),$$

on väide 2 tõestatud.
Kolmanda väite tõestus on analoogiline nõrga AEP omaduse vastava väite tõestusega. ■

Seega iga tugevalt ϵ tüüpiline sõna on ka nõrgalt ϵ tüüpiline ehk $T_\epsilon^n \subset W_\epsilon^n$. Vastupidine ei kehti.

Kontranäide: Olgu $\mathcal{X} = \{a, b, c\}$, $P(a) = 0.5, P(b) = P(c) = 0.25$. Olgu $n = 100$. Vaatame sõna x, kus $N_x(a) = 50$ ja $N_x(b) = 50$. Seega x tüüp on $(0.5, 0.5, 0)$ ja kui $\epsilon < 3$, pole see sõna tugevalt ϵ-tüüpiline. Kuid tema tõenäosus on

$$P^n(x^n) = 0.5^{50}0.25^{50} = 0.5^{50}0.25^{25}0.25^{25},$$

millest

$$\frac{1}{n}\log P^n(x^n) = 0.5\log 0.5 + 0.25\log 0.25 + 0.25\log 0.25 = -H(P).$$

Seega $x^n \in W_\epsilon^n$ iga ϵ korral.

5.6 Ülesanded

1. Olgu $X_1, \ldots, X_n$ iid jaotusega Q. Tõestada, et ledub konstant γ nii, et

   $$\frac{1}{n}L(X_1, \ldots, X_n) \to \gamma,$$

   kus $L(X_1, \ldots, X_n)$ on logaritmiline tõepärasuhe. Avaldada γ juhul, kui $Q = P_1$ ja $Q = P_0$. Järeldada, et kehtivad koondumised (101).

2. Olgu $\pi \in (0, 1)$ Leida test A_n, mille korral oleks väikseim esimest ja teist tüüpi vea kaalutud keskmine $\pi\alpha_n + (1 - \pi)\beta_n$.

3. Tõestada (105).
   Näpunäide:

   a) Tõesta, et $\lambda \mapsto C(\lambda)$ on kumer;

   b) Tõesta, et

   $$\lambda^* = \arg\min_\lambda C(\lambda) \quad \Leftrightarrow \quad D(P_{\lambda^*}||P_1) = D(P_{\lambda^*}||P_0)$$

   c) Veendu, et

   $$D(P_{\lambda^*}||P_0) = -\log C(\lambda^*).$$

4. Suurima tõepära hinnang Olgu $\mathcal{P}$ tõenäosusjaotuste hulk tähestikul $\mathcal{X}$ – *mudel*. Olgu x^n valim ja $\hat{P} \in \mathcal{P}$ olgu suurima tõepära hinnang, st

   $$\hat{P} = \arg\max_{P\in\mathcal{P}} P^n(x^n).$$

   a) Tõesta, et $\hat{P}$ on tüübile P_{x^n} K-L mõttes lähim jaotus hulgast $\mathcal{P}$:

   $$\hat{P} = \arg\min_{P\in\mathcal{P}} D(P_{x^n}||P).$$

b) Olgu

$$l_n(P) = \frac{1}{n} \sum_{i=1}^{n} \log P(x_i)$$

logaritmiline tõepärafunktsioon. Tõesta, et kui $X_1, \ldots, X_n$ on iid juhuslikud suurused jaotusega Q, siis leidub funktsioon $P \mapsto l(P)$ nii, et iga jaotuse P korral $l_n(P) \to l(P)$ p.k. Funktsiooni $l(P)$ nimetatakse *tõepärakontrastiks*. Milline jaotus maksimiseerib $l(P)$ üle $\mathcal{P}$?

6 Ergoodilise teooria elemendid

6.1 Põhimõisted

6.1.1 Kolmogorovi esitus

Olgu $X = \{X_n\}_{n=1}^{\infty}$ juhuslik protsess tähestikul $\mathcal{X}$ (s.t. juhusliku suuruse X_n väärtused kuuluvad hulka $\mathcal{X}$ iga n korral).
Olgu $\mathcal{X}^\infty$ kõikide hulga $\mathcal{X}$ elementidest moodustatud jadade hulk. Juhusliku protsessi $\mathcal{X}$ trajektoor on alati hulga $\mathcal{X}^\infty$ element. Olgu

$$[a_m^n] = \{x^\infty \in \mathcal{X}^\infty : x_i = a_i, i = m, \dots, n\}.$$

Hulka $[a_m^n]$ nimetatakse *silindriks*. Olgu $\sigma(\mathcal{X}^\infty)$ silindrite poolt moodustatud σ-algebra hulgal $\mathcal{X}^\infty$. Protsess X seab igale silindrile $[a_m^n]$ vastavusse arvu

$$P(a_m^n) := \mathbf{P}(X_i = a_i, i = m, \dots, n).$$

Saab näidata, et kujutis P laieneb üheselt mõõduks σ-algebral $\sigma(\mathcal{X}^\infty)$ (see juhuslike protsesside teoorias olulist tulemust tuntakse kui Kolmogorovi olemasoluteoreemi). Protsess $\mathcal{X}$ defineerib seega ruumil $(\mathcal{X}^\infty, \sigma(\mathcal{X}^\infty))$ tõenäosusmõõdu P. Mõõtu P interpreteerime kui juhusliku protsessi X jaotust.

Teisest küljest, olgu ruumil $(\mathcal{X}^\infty, \sigma(\mathcal{X}^\infty))$ antud mingi tõenäosusmõõt P. Veendume, et leidub selle jaotusega juhuslik protsess. Vaatleme *koordinaatfunktsioone*

$$Z_n : \mathcal{X}^\infty \to \mathcal{X}, \quad Z_n(x^\infty) = x_n.$$

On selge, et Z_n on mõõtuv (miks?) ehk iga n korral on Z_n juhuslik suurus. Järelikult on $Z = \{Z_n\}$ juhuslik protsess. On selge, et protsessi Z jaotus on P, sest iga silindri $[a_m^n]$ korral

$$\mathbf{P}(Z_i = a_i, i = m, \dots, n) = P(a_m^n).$$

Kokkuvõttes: iga juhusliku protsessi X korral leidub ruumil $(\mathcal{X}^\infty, \sigma(\mathcal{X}^\infty))$ mõõt P nii, et koordinaatfunktsioonide kaudu esitatud protsessil Z on sama jaotus. Protsessi Z nimetatakse tihti protsessi X *kanooniliseks esituseks* või *Kolmogorovi esituseks*.

6.1.2 Statsionaarsus

Defineerime *Bernoulli nihke*

$$T : \mathcal{X}^\infty \to \mathcal{X}^\infty, \quad (Tx)_n = x_{n+1}, \quad n \geq 1.$$

Seega T nihutab jada elemente ühe võrra ettepoole, st

$$T(x_1, x_2, \dots) = x_2, x_3, \dots.$$

Pamene tähele, et iga silindri originaal on silinder:

$$T^{-1}[a_m^n] = [b_{m+1}^{n+1}], \quad b_{i+1} = a_i.$$

Et silindrid moodustavad $(\sigma(X^\infty))$, on T mõõtuv ehk $T^{-1}(B) \in \sigma(X^\infty)$ iga $B \in \sigma(X^\infty)$ korral.
Olgu P ruumil $(\mathcal{X}^\infty, \sigma(\mathcal{X}^\infty))$ antud tõenäosusmõõt. Kujutis T on *mõõtu säilitav* ehk mõõt P on T-invariantne, kui

$$P(T^{-1}B) = P(B), \quad \forall B \in \sigma(\mathcal{X}^\infty).$$

Tuletame meelde, et juhuslik protsess X on statsionaarne kui iga $n \geq 1$ ja $k \geq 1$ korral on juhuslikud vektorid

$$(X_1, \ldots, X_k) \quad ja \quad (X_{n+1}, \ldots, X_{n+k})$$

sama jaotusega. Veendume, et protsess on statsionaarne parajasti siis, kui tema jaotus on mõõtu säilitav.
Olgu X statsionaarne protsess. Veendume, et tema jaotus P on Bernoulli nihke suhtes invariantne. Olgu $[a_m^n]$ silinder. Statsionaarsusest järeldub

$$P(a_m^n) = \mathbf{P}(X_m = a_m, \ldots, X_n = a_n) = \mathbf{P}(X_{m+1} = a_m, \ldots, X_{n+1} = a_n) = P(b_{m+1}^{n+1}),$$

kus $b_{i+1} = a_i$. Et $[b_{m+1}^{n+1}] = T^{-1}[a_m^n]$ ja X on statsionaarne, siis $P([a_m^n]) = P(T^{-1}[a_m^n])$. Seega on kujutis T silindrite hulgal mõõtu säilitav. Dynkini π-λ teoreemist järeldub nüüd, et kujutus T on mõõtu säilitav ka laiemal hulgal $\sigma(\mathcal{X}^\infty)$) (tõepoolest, silindrilised hulgad moodustavad π-süsteemi; mõõtu säilitavate hulkade klass aga λ-süsteemi. See λ-süsteem sisaldab silindrilisi hulki, seega ka nende tekitatud σ-algebrat.) Järelikult P on T suhtes invariantne.
Teistpidi, olgu X protsess, mille jaotus P on Bernoulli nihke suhtes invariantne. Veendume, et koordinaatprotsess Z on statsionaarne. Siis on statsionaarne ka X (miks?) Kui P on T suhtes invariantne, siis iga $a_1, \ldots, a_k \in \mathcal{X}$ korral

$$P(Z_1 = a_1, \ldots, Z_k = a_k) = P(a_1^k) = P(T^{-1}[a_1^k]) = P(Z_2 = a_1, \ldots, Z_{1+k} = a_k),$$

millest järeldub, et $(Z_1, \ldots, Z_k)$ ja $(Z_2, \ldots, Z_{k+1})$ on sama jaotusega.

6.1.3 Ergoodilise teooria mudel

Nägime, et protsessi X statsionaarsus on ekvivalentne tema jaotuse P invariantsusega Bernoulli nihke T suhtes. Ergoodilises teoorias vaadeldakse abstrakset tõenäosusruumi $(\Omega, \mathcal{F}, \mathbf{P})$ ja sellel antud mõõtuvat kujutist $T : \Omega \to \Omega$, mis säilitab mõõtu, st $\mathbf{P}(T^{-1}(A)) = \mathbf{P}(A)$ iga $A \in \mathcal{F}$ korral. Olgu

$$X : \Omega \to \mathcal{X}$$

mõõtuv funktsioon. Juhusliku suuruse X ja kujutise T abil defineerime juhusliku protsessi

$$X_n := X \circ T^{n-1}, \quad n = 1, 2, \ldots. \tag{113}$$

Ülesanne: Tõestada, et protsess (113) on statsionaarne. Selleks näita, et iga k, n ning $a_1, \ldots, a_k$ korral

$$P(X_1 = a_1, \ldots, X_k = a_k) = \mathbf{P}\{\omega : X(\omega) = a_1, X(T\omega) = a_2, \ldots, X(T^{k-1}\omega) = a_k\} =$$
$$\mathbf{P}\{\omega : X(T^n\omega) = a_1, X(T(T^n\omega)) = a_2, \ldots, X(T^{k-1}(T^n\omega)) = a_k\} = P(X_{n+1} = a_1, \ldots, X_{n+k} = a_k).$$

Võttes $(\Omega, \mathcal{F})$ rolli $(\mathcal{X}^\infty, \sigma(\mathcal{X}^\infty))$ ning T rolli Bernoulli nihke, saame statsionaarse protsessi mudeli. Võttes $X = Z_1$ (esimene koordinaat), saame protsessist (113) koordinaatprotsessi Z. Seega kirjeldatud mudel kätkeb endas ka statsionaarse protsessi kanoonilise esituse. Tihti on hulgal Ω antud mingi ülimalt loenduv tükeldus $\mathcal{P}_a$, $a \in \mathcal{A}$ (a on tüki $\mathcal{P}_a$ "nimi") ning funktsioon X annab selle tüki nime kuhu ω parajasti kuulub, st $X(\omega) = a$ parajasti siis, kui $\omega \in \mathcal{P}_a$. Näiteks statsionaarse protsessi kanoonilise esituses on jadade hulk $\mathcal{X}^\infty$ tükeldatud esimese tähe järgi, iga tüki nimi ongi vastav täht.

Ergoodilise teooria juured on füüsikas. Kujutist T võib interpreteerida kui dünaamikat (näiteks aega) ning teooria keskendub eelkõige trajekooride $\omega, T\omega, T^2\omega, \ldots$ uurimisele (ergoodilises teoorias nimetatakse trajektoori orbiidiks). Funktsiooni X võib interpreteerida kui mõõtmist või eksperimendi tulemust. Näiteks ei pruugi me alati täpselt idenfitseerida ω asukohta, küll aga võib hulgal Ω olla antud tükeldus nii, et saame alati täpselt teada, millisesse tükki ω kuulub.

Et ergoodiline teooria tegeleb orbiitidega $\omega, T\omega, T^2\omega, \ldots$, on erilisel kohal T- invariantsed hulgad. Hulk $B \in \mathcal{F}$ on T-*invariantne*, kui $TB \subset B$. Kui $T\omega$ satub invariantsesse hulka, siis ta jääb sinna.

Def 6.1 *Olgu* $(\Omega, \mathcal{F}, \mathbf{P})$ *tõenäosusruum ning* $T : \Omega \to \Omega$ *mõõtu säilitav kujutis. Kujutis* T *on* **ergoodiline** *kui iga* T *invarantse hulga mõõt on 0 või 1.*

Järgnevas teoreemis B on mõõtuv hulk, $f : \Omega \to \mathbb{R}$ on mõõtuv funktsioon.

Teoreem 6.2 *Olgu* $(\Omega, \mathcal{F}, \mathbf{P})$ *tõenäosusruum ning* $T : \Omega \to \Omega$ *mõõtu säilitav kujutis. Siis järgmised väited on ekvivalentsed:*

1 T *on ergoodiline;*

2 *kui* $T^{-1}B = B$, *siis* $\mathbf{P}(B) = 0$ *või* $\mathbf{P}(B) = 1$;

3 *kui* $\mathbf{P}(T^{-1}B \Delta B) = 0$, *siis* $\mathbf{P}(B) = 0$ *või* $\mathbf{P}(B) = 1$;

4 *kui* $B \subseteq T^{-1}B$, *siis* $\mathbf{P}(B) = 0$ *või* $\mathbf{P}(B) = 1$;

5 *kui* $T^{-1}B \subseteq B$, *siis* $\mathbf{P}(B) = 0$ *või* $\mathbf{P}(B) = 1$;

6 *kui $f = f \circ T$ p.k., siis f on konstantne p.k.*

Tõestus.

1 $\Rightarrow$ **2** Olgu $B \in \mathcal{F}$ selline, et $T^{-1}B = B$. Et iga hulga C korral $TT^{-1}C \subseteq C$, saame $TB \subseteq B$, millest $\mathbf{P}(B) = 0$ või $\mathbf{P}(B) = 1$.

2 $\Rightarrow$ **3** Ülesanne.
(Olgu $\mathbf{P}(T^{-1}B\Delta B) = 0$. Veendu, et $\mathbf{P}(T^{-n}B\Delta B) = 0$ iga n korral. Seejärel veendu, et $(\cup_n T^{-n}B)\Delta B \subset \cup_n (T^{-n}B\Delta B)$. Seejärel veendu, et $\mathbf{P}(\limsup_n T^{-n}B) = \mathbf{P}(B)$ ning järelda, et $\mathbf{P}(B)$ on kas 0 või 1).

3 $\Rightarrow$ **4** Ülesanne.

4 $\Rightarrow$ **1** Olgu $B \in \mathcal{F}$ invariantne hulk. Siis $B \subseteq T^{-1}B$, millest $\mathbf{P}(B) = 0$ või $\mathbf{P}(B) = 1$.

3 $\Rightarrow$ **5** $\Rightarrow$ **2** Ülesanne.

3 $\Rightarrow$ **6** Ülesanne.
(Kui f pole konstant p.k., siis leidub Boreli hulk $C \subset \mathbb{R}$ nii, et

$$1 > \mathbf{P}(\omega : f(\omega) \in C) > 0.$$

Võta $A = f^{-1}(C)$ ja veendu, et $\mathbf{P}(A\Delta T^{-1}A) = 0$.)

6 $\Rightarrow$ **2** Ülesanne.

■

Def 6.3 *Statsionaarne juhuslik protsess $X = \{X_n\}_{n=1}^{\infty}$ on ergoodiline, kui Bernoulli nihe on ergoodiline protsessi X Kolmogorovi esituse suhtes.*

Märkus: Ergoodilisus on kujutise T ja mõõdu $\mathbf{P}$ omadus. Kui T on fikseeritud, siis räägime ergoodilisest mõõdust. juhuslike protsesside korral on alati fikseeritud ruum $(\mathcal{X}^{\infty}, \sigma(\mathcal{X}^{\infty})$ ja Bernoulli nihe T. Protsessi ergoodilisus sõltub vaid jaotusest P. Sellisel juhul räägime ergoodilisest või mitteergoodilisest jaotusest P. Formaalselt: ütleme, et P on ergoodiline, kui Bernoulli nihe on ergoodiline P suhtes.

Märkus: Tihti nimetatakse invariantseks hulgaks sellist hulka B, mis rahuldab $T^{-1}B = B$; ergoodiline kujutis defineeritakse läbi omaduse **2**. Teinekord nimetatakse invariantseks hulgaks sellist, mille korral $\mathbf{P}(T^{-1}B\Delta B) = 0$ ergoodiline kujutis defineeritakse läbi omaduse **3**. Teoreemist 6.2 järeldub, et sisuliselt invariantsuse definitsioonid ekvivalentsed. Meie definitsioon on ehk intuitiivselt mõistatavam, tema puudus on see, et nii defineeritud invariantsed hulgad ei pruugi moodustada σ-algebrat.

Väide 6.1 *Kui tõenäosusruumil $(\Omega, \mathcal{F}, \mathbf{P})$ antud mõõtu säilitab kujutis T pole ergoodiline, siis leidub hulk B nii, et $T^{-1}(B) = B$ ja $1 > \mathbf{P}(B) > 0$.*

Tõestus. Kui T pole ergoodiline, siis leidub A nii, et $1 > \mathbf{P}(A) > 0$ ja $A \subset T^{-1}A$. Defineerime $T^0A := A$ ja $B := \cup_{n=0}^{\infty} T^{-n}A$. Seega $A \cup (\cup_{n=1}^{\infty} T^{-n}A)$. Et aga $A \subset T^{-1}A$, siis $A \cup (\cup_{n=1}^{\infty} T^{-n}A) = \cup_{n=1}^{\infty} T^{-n}A$. Nüüd,

$$T^{-1}(B) = T^{-1}(\cup_{n=0}^{\infty} T^{-n}A) = \cup_{n=1}^{\infty} T^{-n}A = B.$$

Veendu, et $\mathbf{P}(B) = \mathbf{P}(A)$. ■

Seega, kui T pole ergoodiline, siis saab ruumi Ω jagada kaheks: B ja B^c nii, et mõlemad tükid on invariantsed ja mõlema tüki mõõt kuulub hulka $(0, 1)$. Iga ω kuulub ühte neist tükkidest ja jääb sinna igaveseks.

6.1.4 Unustav kujutis

Olgu $(\Omega, \mathcal{F}, \mathbf{P})$ tõenäosusruum ja sellel antud mõõtu säilitav kujutis T.

Def 6.4 *Kujutis T on* **unustav** *(mixing), kui iga $A, B \in \mathcal{F}$ korral*

$$\lim_{n\to\infty} \mathbf{P}(T^{-n}A \cap B) = \mathbf{P}(A)\mathbf{P}(B). \tag{114}$$

Def 6.5 *Kujutis T on* **nõrgalt unustav** *(weakly mixing), kui iga $A, B \in \mathcal{F}$ korral*

$$\frac{1}{n}\sum_{k=1}^{n} |\mathbf{P}(T^{-k}A \cap B) - \mathbf{P}(A)\mathbf{P}(B)| \to 0. \tag{115}$$

Nõrgalt unustav kujutis rahuldab tingimust

$$\lim_{n} \frac{1}{n}\sum_{k=1}^{n} \mathbf{P}(T^{-k}A \cap B) = \mathbf{P}(A)\mathbf{P}(B), \quad \forall A, B \in \mathcal{F}. \tag{116}$$

Ülesanne: Tõestada, et unustav kujutis on nõrgalt unustav ja nõrgalt unustav kujutis on ergoodiline.

Def 6.6 *Statsionaarne juhuslik protsess $X = \{X_n\}_{n=1}^{\infty}$ on unustav, kui Bernoulli nihe on unustav protsessi X Kolmogorovi esituse suhtes.*

Unustav protsess on seega ergoodiline. Kuigi unustavus on tugevam omadus kui ergoodilisus, on seda tihti kergem kontrollida. Koondumine (114) kehtib, kui (114) kehtib iga $\mathcal{F}$ tekitava π-süsteemi korral (seda saab näidata Dynkini π-λ-teoreemi abil). Seega on T unustav, kui (114) kehtib silindrite korral. Olgu $A = [a_{m_1}^{n_1}]$ ja $B = [b_{m_2}^{n_2}]$ kaks silindrit. Tuletame meelde, et

$$T^{-n}[a_{m_1}^{n_1}] = [a_{n+m_1}^{n+n_1}], \quad a_{n+i} = a_i. \tag{117}$$

Tähistame, $x_m^n = (x_m, \ldots, x_n)$. Seega X on unustav parajasti siis, kui

$$\mathbf{P}\Big((X_{m_2}, \ldots, X_{n_2}) = b_{m_2}^{n_2}, (X_{n+m_1}, \ldots, X_{n+n_1}) = a_{m_1}^{n_1}\Big) = P(B \cap T^{-n}A) \to$$

$$\to P(B)P(A) = P(b_{m_2}^{n_2})P(a_{m_1}^{n_1}) = \mathbf{P}\Big((X_{m_2}, \ldots, X_{n_2}) = b_{m_2}^{n_2}\Big)\mathbf{P}\Big((X_{m_1}, \ldots, X_{n_1}) = a_{m_2}^{n_2}\Big).$$

Seega A ja B on asümptootiliselt sõltumatud.

6.1.5 Näiteid ergoodilistest kujutistest

IID protsess. Veendume, et iid protsess X on unustav. Kui $n > n_2$, siis silindrid $T^{-n}[a_{m_1}^{n_1}] = [a_{n+m_1}^{n+n_1}]$ ja $[b_{m_2}^{n_2}]$ on määratud lõikumatute indeksihulkadega. Seega, kui $n > n_2$, siis

$$\begin{aligned} P(B \cap T^{-n}A) &= \mathbf{P}\Big((X_{m_2}, \ldots, X_{n_2}) = b_{m_2}^{n_2}, (X_{n+m_1}, \ldots, X_{n+n_1}) = a_{m_1}^{n_1}\Big) \\ &= \mathbf{P}\Big((X_{m_2}, \ldots, X_{n_2}) = b_{m_2}^{n_2}\Big)\mathbf{P}\Big((X_{m_1}, \ldots, X_{n_1}) = a_{m_1}^{n_1}\Big) \\ &= P(b_{m_2}^{n_2})P(a_{m_1}^{n_1}) = P(B)P(A). \end{aligned}$$

Selle protsessi ergoodilisust on kerge kontrollida ka vahetult definitsioonist lähtudes. Olgu $\{Z_n\}_{n=1}^{\infty}$ iid koordinaatprotsess ruumil $(\mathcal{X}^{\infty}, \sigma(\mathcal{X}^{\infty}))$. Siis $\sigma(\mathcal{X}^{\infty}) = \sigma(Z_1, Z_2, \ldots)$. Olgu $A \in \sigma(\mathcal{X}^{\infty})$ selline, et $T^{-1}A = A$. Intuitiivselt on selga (ja seda on ka võimalik näidata), et $T^{-1}A \in \sigma(Z_2, Z_3, \ldots)$. Seega $A \in \sigma(Z_2, Z_3, \ldots)$. Samuti $A \in \sigma(Z_n, Z_{n+1}, \ldots)$ iga n korral. Seega A kuulub jääk σ-algebrasse. Kolmogorovi 0-1 seaduse järgi on tema mõõt kas 0 või 1.

Markovi ahel. Olgu $\mathcal{X}$ ülimalt loenduv seisundite hulk ja P sellel antud mittelahutuv (*irreducible*) üleminekutõenäosuste maatriks. Definitsiooni järgi on ergoodiline protsess eelkõige statsionaarne, st leidub statsionaarne algjaotus π (see on tõenäosusjaotus mis rahuldab tingimust $\pi P = \pi$). Mittelahutuval üleminekumaatriksil ahelal on statsionaarne algjaotus parjasti siis kui kõik ahela kõik seisundid on positiivselt rekurrentsed (*positive recurrent*). Seega vaatleme sellist ahelat (võib olla perioodiline!). Viimane tingimus on alati täidetud lõpliku seisundite hulga korral. Sellise Markovi ahela korral kehtib alati koondumine

$$\lim_n \frac{1}{n} \sum_{k=1}^{n} P^k = Q, \tag{118}$$

kusjuures maatriksi Q iga rida on π. (Vaata näiteks Rick Durrett "Probability: Theory and Examples" ptk 5 (5.2).) Olgu X kirjeldatud MA. Veendume, et X on ergoodiline. Selleks veendume, et iga $A, B \in \sigma(\mathcal{X}^{\infty})$ korral kehtib (116). Piisab, kui näitame, et (116) kehtib suvaliste silindrite korral (Dynkini $\pi - \lambda$ teoreem). Olgu $B = [b_{m_1}^{n_1}]$ ja $A = [a_{m_2}^{n_2}]$ kaks silindrit. Üldisust kitsendamata võime eeldada, et $m_1 = 1$ (miks?). Seega

$$T^{-k}A = [a_{k+m_2}^{k+n_2}], \quad a_{k+i} = a_i.$$

Seega, kui $k > n_1$, sõltuvad $T^{-k}A$ ja B erinevatest indeksitest. Olgu P protsessi jaotus. Kui $k > n_1$, siis

$$P(B \cap T^{-k}A) = UV_kW,$$

kus

$$
\begin{aligned}
U &= P(B) = P(b_1^{n_1}) = \pi(b_1) \prod_{j=1}^{n_1-1} P_{b_i,b_{i+1}} \\
V_k &= \mathbf{P}(X_{k+m_2} = a_{k+m_2} | X_{n_1} = b_{n_1}) = \mathbf{P}(X_{k+m_2} = a_{m_2} | X_{n_1} = b_{n_1}) = P_{b_{m_1},a_{m_2}}^{k+m_2-n_1} \\
W &= \mathbf{P}(X_{k+m_2+1} = a_{k+m_2+1}, X_{k+m_2+2} = a_{k+m_2+2}, \ldots, X_{k+n_2} = a_{k+n_2} | X_{k+m_2} = a_{k+m_2}) \\
&= \mathbf{P}(X_{m_2+1} = a_{m_2+1}, X_{m_2+2} = a_{m_2+2}, \ldots, X_{n_2} = a_{n_2} | X_{m_2} = a_{m_2}) \\
&= \prod_{i=m_2}^{n_2-1} P_{a_i,a_{i+1}}.
\end{aligned}
$$

Paneme tähele, et U ja W ei sõltu k-st. Seosest (118) saame aga, et

$$\lim_n \frac{1}{n} \sum_{k=1}^{n} W_k = \lim \frac{1}{n} \sum_{k=1}^{n} P_{b_{m_1},a_{m_2}}^{k+m_2-n_1} = \pi(a_{m_2}),$$

millest

$$\lim \frac{1}{n} \sum_{k=1}^{n} P(B \cap T^{-k} A) = \lim \frac{1}{n} \sum_{k=1}^{n} U V_k W = U \pi(a_{m_2}) W = P(b_1^{n_1}) P(a_{m_2}^{n_2}),$$

sest

$$\pi(a_{m_2}) W = P(a_{m_2}^{n_2}).$$

Seega statsionaarne (leidub π) taandumatu MA on ergoodiline.
Veendume, et kehtib ka teistpidine: iga ergoodiline MA on taandumatu. Olgu $X = X_1, X_2, \ldots$ ergoodiline MA üleminekumaatriksiga P ja statsionaarse jaotusega π tähestikul $\mathcal{X}$. On loomulik eeldada, et $\pi(a) > 0$ iga $a \in \mathcal{X}$ korral. Tõepoolest, kui leidub täht b nii, et $\pi(b) = 0$, siis iga n korral $\mathbf{P}(X_n = b) = 0$ (miks?) ehk tõenäosusega 1 ei satu ahel kunagi sinna seisundisse. Seega võime midagi kaotamata selle seisundi tähestikust minema heita.
Kui $\pi(a) > 0$ iga $a \in \mathcal{X}$ korral, on kõik seisundid rekurrentsed ja seisundite hulga $\mathcal{X}$ võib jagada lõikumatuteks alamhulkadeks $\mathcal{X} = \cup_i \mathcal{X}_i$, kus $\mathcal{X}_i$ on kinnised rekurrentsed seisundite klassid (dekompositsiooniteoreem). Kui X_1 satub ühte neist klassidest, näiteks $\mathcal{X}_0$, jääb ahel sinna klassi tõenäosusega 1. Teisisõnu

$$\mathbf{P}\big(X_1 \in \mathcal{X}_0, X_2 \in \mathcal{X}_0, X_3 \in \mathcal{X}_0, \ldots) = \mathbf{P}\big(X_1 \in \mathcal{X}_0).$$

Teisest küljest aga

$$
\begin{aligned}
\mathbf{P}\big(X_1 \in \mathcal{X}_0, X_2 \in \mathcal{X}_0, X_3 \in \mathcal{X}_0, \ldots) &= P(x \in \mathcal{X}^\infty : x_1 \in \mathcal{X}_0, x_2 \in \mathcal{X}_0, \cdots) \\
&= P(A \cap T^{-1} A \cap T^{-2} A \cap \ldots) = P(B),
\end{aligned}
$$

kus

$$A := \{x \in \mathcal{X}^\infty : x_1 \in \mathcal{X}_0\}, \quad B := \cap_{n=0}^{\infty} T^{-n} A.$$

Et $B \subset T^{-1}B$ on B invariantne hulk. Et P on ergoodiline, on $P(B) = 1$ (miks $P(B) \neq 0$?) ja seega on $\mathbf{P}(X_1 \in \mathcal{X}_0) = 1$ ehk $\mathcal{X}_0 = \mathcal{X}$ on ainus kinnine taandumatu klass – ahel on taandumatu.

Järeldus 6.1 *Statsionaarne MA on ergoodiline parajasti siis, kui ta on mittelahutuv.*

Märkus: Klassikalises tõenäosusteoorias nimetatakse MA ergoodiliseks, kui lisaks ergoodilisusele ergoodilise teooria mõttes (lahutamatu üleminekumaatriks ja statsionaarne algjaotus) on ta lisaks mitteperioodiline. Sellisel juhul kehtib koondumise (118) tugevam versioon

$$\lim_n P^n = Q,$$

millest järeldub (ülaltoodud argument), et ahel on unustav.

Pööre. Olgu $\Omega = [0, 1)$ ja

$$T\omega = \omega + \alpha \mod 1.$$

Sisuliselt on T pööre ühikringil, sest ω võib samastada ühikringi punktiga $(\sin \omega 2\pi, \cos \omega 2\pi)$ ning $T\omega$ on siis pööre nurgaga $\alpha 2\pi$. T on üksühene kujutis ning pööratud intervalli pikkus ei muutu. Seega on T invariantne Lebesgue'i mõõdu suhtes. Millal on T ergoodiline?

Ülesanne: Tõestada, et kui α on ratsionaalarv, siis T pole ergoodiline.

Tõestame, et irratsionaalse α korral on pööre ergoodiline kujutis. Selleks tõestame kõigepealt, et suvalise ω korral on trajektoor

$$F := \{T\omega, T^2\omega, T^3\omega, \ldots\}$$

kõikjal tihe hulk. Oletame, et see nii pole. Siis leidub lahtine intervall I nii, et I pikkus on positiivne ja $I \cap F = \emptyset$. Olgu I maksimaalne selles mõttes, et kui J on selline lahtine intervall, et $I \subseteq J$ ning $J \cap F = \emptyset$, siis $J = I$. On selge, et $T^{-n}I \cap F = \emptyset$ (miks?). Et pööre on irratsionaalne, siis $T^{-n}I \neq I$. Et I in maksimaalne, siis $T^{-n}I \cap I = \emptyset$. Seega $T^{-n}I$, $n = 1, 2, \ldots$ on lõikumatute hulkade jada. Et T on pikkust säilitav, ei saa I pikkus olla positiivne. Saime vastuolu.
Olgu nüüd A selline, et $T^{-1}A = A$. Oletame, et $\mathbf{P}(A) > 0$. Veendume, et sellisel juhul $\mathbf{P}(A) = 1$. Olgu $\epsilon > 0$ ning olgu I selline intervall, et $\mathbf{P}(I) < \epsilon$ kuid $\mathbf{P}(A \cap I) \geq (1-\epsilon)\mathbf{P}(I)$. Tänu sellele, et iga punkti trajektoor on kõikjal tihe, leiduvad indeksid $n_1, \ldots, n_k$ nii, et $T^{n_1}I, \ldots, T^{n_k}I$ on lõikumatud ning $\sum_{i=1}^k \mathbf{P}(T^{n_i}I) \geq 1 - 2\epsilon$. Et $T^{-n_i}A = A$ ja $\mathbf{P}(A \cap I) \geq (1 - \epsilon)\mathbf{P}(I)$, saame

$$\begin{aligned}\mathbf{P}(A) &\geq \sum_{i=1}^k \mathbf{P}(T^{n_i}I \cap A) = \sum_{i=1}^k \mathbf{P}(T^{-n_i}(T^{n_i}I \cap A)) = \sum_{i=1}^k \mathbf{P}(I \cap T^{-n_i}A)) \\ &= \sum_{i=1}^k \mathbf{P}(I \cap A) = \sum_{i=1}^k (1-\epsilon)\mathbf{P}(I) \geq \sum_{i=1}^k (1-\epsilon)\mathbf{P}(T^{n_i}I) \geq (1-\epsilon)(1-2\epsilon).\end{aligned}$$

Ülesanne: Kas pööre on unustav kujutis?

6.2 Ergoodilised teoreemid

Meeldetuletus: Alljärgnevas kasutame tihti integraali muutuja vahetust. Tuletame selle meelde. Olgu (S, Σ, μ) mõõduga ruum ja selle kõrval veel teinegi mõõtuv ruum (S', Σ'). Olgu $T : S \to S'$ Σ/Σ'-mõõtuv. Ruumil (S', Σ') vaatleme mõõtu μT^{-1}.
Siis $f \in \mathcal{L}_1(S', \Sigma', \mu T^{-1})$ parajasti siis kui $f \circ T \in \mathcal{L}_1(S, \Sigma, \mu)$ ning kehtib

$$\int_{T^{-1}A'} f(Ts)\mu(ds) = \int_{A'} f(s')\mu T^{-1}(ds'). \tag{119}$$

6.2.1 Birkhoffi ergoodiline teoreem

Ergoodilise protsessi korral kehtib suurte arvude seadus: iga trajektoori aritmeetiline keskmine läheneb keskväärtusele EX_1.
Ergoodilises teoorias sõnastatakse see järgmiselt ($f \circ T^0 = f$).

Teoreem 6.7 (Birkhoff, 1931) *Olgu T tõenäosusruumil $(\Omega, \mathcal{F}, \mathbf{P})$ antud mõõtu säilitav kujutis. Siis iga $f \in L_1(\Omega, \mathcal{F}, \mathbf{P})$ korral leidub T invariantne funktsioon f^* nii, et kehtivad koondumised*

$$\frac{1}{n}\sum_{i=0}^{n-1} f \circ T^i \to f^*, \ p.k. \qquad \frac{1}{n}\sum_{i=0}^{n-1} f \circ T^i \xrightarrow{1} f^*. \tag{120}$$

Kui T on ergoodiline kujutis, siis f^* on $\mathbf{P}$-p.k. konstant. Et koondumine toimub ka ruumis keskmise mõttes, siis koonduvad ka keskväärtused, millest

$$\begin{aligned}\int \frac{1}{n}\sum_{i=0}^{n-1} f(T^i(\omega))\mathbf{P}(d\omega) &= \frac{1}{n}\sum_{i=0}^{n-1}\int f(T^i(\omega))\mathbf{P}(d\omega) = \frac{1}{n}\sum_{i=0}^{n-1}\int f(\omega)\mathbf{P}T^{-i}(d\omega)\\ &= \frac{1}{n}\sum_{i=0}^{n-1}\int f(\omega)\mathbf{P}(d\omega) = \int f(\omega)\mathbf{P}(d\omega).\end{aligned}$$

Olgu

$$\mathcal{G} := \{B \in \mathcal{F} : T^{-1}B = B\} \tag{121}$$

Ülesanne:

1) Veendu, et $\mathcal{G}$ on σ-algebra.

2) Veendu, et Birkhoffi ergooodilisest teoreemist järeldub:

$$\int_A f d\mathbf{P} = \int_A f^* d\mathbf{P}, \quad \forall A \in \mathcal{G}.$$

Järelda, et $f^* = E[f|\mathcal{G}]$.

Järeldus 6.2 *Olgu T tõenäosusruumil $(\Omega, \mathcal{F}, \mathbf{P})$ antud ergoodiline kujutis. Siis iga $f \in L_1(\Omega, \mathcal{F}, \mathbf{P})$ korral kehtivad koondumised*

$$\frac{1}{n}\sum_{i=0}^{n-1} f \circ T^i \to \int f d\mathbf{P}, \ \ p.k. \qquad \frac{1}{n}\sum_{i=0}^{n-1} f \circ T^i \xrightarrow{1} \int f d\mathbf{P}. \tag{122}$$

Esitame saadud tulemused tõenäosusteooria keeles.

Olgu $X = \{X_n\}_{n=1}^{\infty}$ mingil tõenäosusruumil $(\Omega', \mathcal{F}', \mathbf{P}')$ antud statsionaarne protsess. Protsess X on $\mathcal{F}'/\sigma(\mathcal{X}^\infty)$-mõõtuv kujutis

$$X : \Omega' \to \mathcal{X}^\infty,$$

tema jaotus P on mõõt $\mathbf{P}'X^{-1}$. Et X on statsionaarne, on Bernoulli nihe T ruumil$(\mathcal{X}^\infty, \sigma(\mathcal{X}^\infty), P)$ mõõtu säilitav. Olgu $g : \mathcal{X}^\infty \to \mathbb{R}$ mõõtuv funktsioon, $\int_{X^\infty} |g(x)|P(dx) < \infty$. Olgu

$$\mathcal{G} = \{A \in \sigma(\mathcal{X}^\infty) : T^{-1}A = A\}.$$

Birkhoffi ergoodilisest teoreemist saame, et

$$\frac{1}{n}\sum_{i=0}^{n-1} g \circ T^i \to E[g|\mathcal{G}], \quad \text{p.k. ja ruumis } L_1. \tag{123}$$

Peaaegu kindlasti koondumine tähendab, et

$$\begin{aligned}
1 &= P\{x \in \mathcal{X}^\infty : \frac{1}{n}\sum_{i=0}^{n-1} g(T^i x) \to E[g|\mathcal{G}](x)\} \\
&= \mathbf{P}'X^{-1}\{x \in \mathcal{X}^\infty : \frac{1}{n}\sum_{i=0}^{n-1} g(T^i x) \to E[g|\mathcal{G}](x)\} \\
&= \mathbf{P}'\{\omega' \in \Omega' : \frac{1}{n}\sum_{i=0}^{n-1} g(T^i X(\omega')) \to E[g|\mathcal{G}](X(\omega'))\} \\
&= \mathbf{P}'\{\omega' \in \Omega' : \frac{1}{n}\sum_{i=0}^{n-1} g(X_{i+1}(\omega'), X_{i+2}(\omega'), \ldots)) \to E[g|\mathcal{G}](X(\omega'))\}.
\end{aligned}$$

Seega

$$\frac{1}{n}\sum_{i=0}^{n-1} g(X_{i+1}, X_{i+1}, \cdots) \to E[g|\mathcal{G}](X), \quad \mathbf{P}'\text{p.k.}. \tag{124}$$

Olgu

$$\mathcal{G}' = \{X^{-1}B : B \in \mathcal{G}\}. \tag{125}$$

Definitsioonist johtuvalt $A' \in \mathcal{G}'$ parajasti siis, kui leidub $B \in \mathcal{G}$ nii, et $A' = X^{-1}B$ ehk

$$A' = \{\omega' : X(\omega') \in B\} = \{\omega' : X_1(\omega'), X_2(\omega'), \ldots \in B\}.$$

Et $B \in \mathcal{G}$, siis $B = T^{-1}B$, millest $A' = X^{-1}T^{-1}B = (TX)^{-1}B$ ehk

$$A' = \{\omega' : TX(\omega') \in B\} = \{\omega' : X_2(\omega'), X_3(\omega'), \ldots \in B\}.$$

Seega , kui $A' \in \mathcal{G}'$, siis leidub $B \in \mathcal{G}$ nii, et iga $n \geq 1$ korral

$$A' = \{\omega' : X_n(\omega'), X_{n+1}(\omega'), \ldots \in B\}. \tag{126}$$

Teisest küljest, kui leidub $B \in \sigma(\mathcal{X}^\infty)$ nii, et iga n korral kehtib (126), siis

$$A' = X^{-1}B = X^{-1}T^{-1}B = X^{-1}T^{-2}B = \cdots,$$

millest

$$A' = X^{-1}(\limsup_{i \geq 0} T^{-i}B).$$

Et $\limsup_{i\geq 0} T^{-i}B$, siis $A' \in \mathcal{G}'$.

Kokkuvõttes: $A' \in \mathcal{G}'$ parajasti siis, kui leidub $B \in \sigma(\mathcal{X}^\infty)$ nii, et iga n korral kehtib (126).

Pole raske veenduda, et

$$E[g|\mathcal{G}](X) = E[g(X)|\mathcal{G}'] \quad \text{p.}k.. \tag{127}$$

Tõepoolest, Kõigepealt paneme tähele, et funktsioon

$$E[g|\mathcal{G}](X) : \Omega' \to \mathbb{R}$$

on $\mathcal{G}'$-mõõtuv (see on sellepärast, et $E[g|\mathcal{G}]$ on $\mathcal{G}$-mõõtuv ja iga $\mathcal{G}$ hulga originaal on $\mathcal{G}'$ element. Järelikult iga Boreli hulga B korral $E[g|\mathcal{G}](X)^{-1}(B) = X^{-1}(E[g|\mathcal{G}]^{-1}(B)) \in \mathcal{G}'$.) Võrduse (127) kehtivuseks piisab, kui näitame, et

$$\int_{A'} E[g|\mathcal{G}](X)d\mathbf{P}' = \int_{A'} g(X)d\mathbf{P}', \quad \forall A' \in \mathcal{G}'.$$

Ent

$$\int_{A'} g(X)d\mathbf{P}' = \int_{X^{-1}B} g(X)d\mathbf{P}' = \int_B gd\mathbf{P}'X^{-1} = \int_B gdP = \int_B E[g|\mathcal{G}]dP =$$
$$\int_B E[g|\mathcal{G}]d\mathbf{P}'X^{-1} = \int_{X^{-1}B} E[g|\mathcal{G}](X)d\mathbf{P}' = \int_{A'} E[g|\mathcal{G}](X)d\mathbf{P}'.$$

Seega (124) on

$$\frac{1}{n}\sum_{i=0}^{n-1} g(X_{i+1}, X_{i+1}, \cdots) \to E[g(X_1, X_2, \ldots)|\mathcal{G}'], \quad \mathbf{P}'\text{p.k.}. \tag{128}$$

Koondumine ruumis L_1 kandub üle analoogiliselt:

$$\int_{\mathcal{X}^\infty} \Big|\frac{1}{n}\sum_{i=0}^{n-1} g(T^i x) - E[g|\mathcal{G}](x)\Big| P(dx) = \int_{X^\infty} \Big|\frac{1}{n}\sum_{i=0}^{n-1} g(T^i x) - E[g|\mathcal{G}](x)\Big| \mathbf{P}'X^{-1}(dx)$$
$$= \int_{\Omega'} \Big|\frac{1}{n}\sum_{i=0}^{n-1} g(T^i X) - E[g|\mathcal{G}](X)\Big| d\mathbf{P}'.$$

Seega

$$\frac{1}{n}\sum_{i=0}^{n-1} g(X_{i+1}, X_{i+1}, \cdots) \to E[g(X_1, X_2, \ldots)|\mathcal{G}'] \quad \text{ruumis} L_1. \tag{129}$$

Juhul, kui X on ergoodiline, siis

$$E[g|\mathcal{G}] = Eg = \int g dP = \int_{\mathcal{X}^\infty} g(x)P(dx) = \int_{\Omega'} g(X) d\mathbf{P}' = Eg(X_1, X_2, \ldots).$$

Võtame ülaltoodu kokku.

Teoreem 6.8 (Birkhoffi ergoodiline teoreem) *Olgu $X = \{X_n\}_{n=1}^\infty$ statsionaarne protsess, $g : \mathcal{X}^\infty \to \mathbb{R}$ olgu mõõtuv, $E|g(X_1, X_2 \ldots)| < \infty$. Siis*

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i, X_{i+1}, \ldots) \to E[g(X_1, X_2, \ldots)|\mathcal{G}'] \text{ p.k. ja ruumis } L_1, \tag{130}$$

kus $\mathcal{G}'$ on defineeritud kui (125).
Kui X on ergoodiline, siis

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i, X_{i+1}, \ldots) \to E[g(X_1, X_2, \ldots)] \text{ p.k. ja ruumis } L_1, \tag{131}$$

Võttes funktsiooniks g esimese koordinaatfunktsiooni, saame tuntud järelduse.

Järeldus 6.3 *Olgu $X = \{X_n\}_{n=1}^\infty$ statsionaarne protsess, mille korral $E|X_1| < \infty$. Siis*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to E[X_1|\mathcal{G}'] \text{ p.k. ja ruumis } L_1, \tag{132}$$

kus $\mathcal{G}'$ on defineeritud kui (125).
Kui X on ergoodiline, siis

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to E(X_1) \text{ p.k. ja ruumis } L_1. \tag{133}$$

Näited:

SAS Olgu $X_1, X_2, \ldots$ iid jaotusega P. Siis (133) on tugev suurte arvude seadus (tugev SAS).

Markovi ahela SAS Olgu $X_1, X_2, \ldots$ statsionaarne mittelahutuv MA. Olgu π statsionaarne jaotus. Siis iga funktsiooni $g : \mathcal{X} \to \mathbb{R}$ korral

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) \to \sum_x g(x)\pi(x).$$

6.2.2 Subaditiivne ergoodiline teoreem

Praktikas on väga kasulik nn. subaditiivsed ergoodilised teoreemid. Tuletame meelde, et kui $\{a_n\}$ on subaditiivne mittenegatiivsete reaalarvude jada, s.t.

$$a_{n+m} \le a_n + a_m,$$

siis leidub $\lim_n \frac{a_n}{n}$ ja see võrdub $\inf_n \frac{a_n}{n}$.

Teoreem 6.9 (Kingman, 1968) *Olgu T tõenäosusruumil $(\Omega, \mathcal{F}, \mathbf{P})$ antud mõõtu säilitav kujutis ning olgu $\{g_n\}$ ruumi $L_1(\Omega, \mathcal{F}, \mathbf{P})$ elementide jada, mis rahuldab tingimust*

$$g_{n+m} \le g_n + g_m \circ T^n, \quad \textit{iga } m, n \textit{korral.} \tag{134}$$

Siis leidub T invariantne funktsioon $g \ge -\infty$ nii, et

$$\lim_n \frac{g_n}{n} \to g$$

peaaegu kindlasti ja ruumis L_1.

Kui T on ergoodiline, on g loomulikult konstant.

Paneme tähele, et ülaltoodud teoreemist järeldub Birkhoffi ergoodiline teoreem. Tõepoolest, olgu $f \in L_1(\Omega, \mathcal{F}, \mathbf{P})$ ja defineerime

$$g_n := \sum_{i=0}^{n-1} f \circ T^i, \quad n = 1, 2, \ldots.$$

Ülesanne: Tõestada, et $\{g_n\}$ rahuldab seost (134) ja järeldada, et Kingmani subaditiivsest teoreemist järeldub Birkhoffi ergoodiline teoreem.

Tõenäosusteooria keeles on Kingmani subaditiivne ergoodiline teoreem järgmine.

Teoreem 6.10 (Kingman, 1968) *Olgu $X = \{X_n\}$ statsionaarne protsess. Olgu $\{g_n\}$ sellised funktsioonid, et*

$$E|g_n(X_1, X_2, \ldots)| < \infty, \quad \forall n$$

ja

$$g_{n+m}(X_1, X_2, \ldots) \le g_n(X_1, X_2, \ldots) + g_m(X_{n+1}, X_{n+2}, \ldots) \text{ iga } m, n \text{korral.} \tag{135}$$

Siis leidub funktsioon g nii, et

$$\frac{1}{n} g_n(X_1, X_2, \ldots) \to E[g(X_1, X_2, \ldots)|\mathcal{G}'] \text{ p.k. ja ruumis } L_1, \tag{136}$$

kus $\mathcal{G}'$ on defineeritud kui (125).
Kui X on ergoodiline, siis

$$\frac{1}{n} g_n(X_1, X_2, \ldots) \to E[g(X_1, X_2, \ldots)] \text{ p.k. ja ruumis } L_1, \tag{137}$$

Näited:

Juhusliku ekslemise poolt väisatud punktid. Olgu $X_1, X_2, \ldots$ iid integreeruvad juhuslikud suurused. Olgu $S_n = X_1 + \cdots + X_n$ ja vaatleme juhuslikku ekslemist $S = \{S_n\}$. Olgu R_n juhusliku ekslemise poolt esimese n sammu jooksul väisatud punktide hulk. Formaalselt,

$$R_n = |\{S_1, S_2, \ldots, S_n\}|.$$

On selge, et $R_n \le n$, kuid kuidas käitub $\frac{R_n}{n}$? Olgu

$$g_n : \mathcal{X}^\infty \to \mathbb{N}, \quad g_n(x_1, x_2, \ldots) = |x_1, x_1 + x_2, x_1 + x_2 + x_3, \ldots, x_1 + \ldots + x_n|.$$

Nüüd $R_n = g_n(X_1, X_2, \ldots)$. On selge, et

$$g_{m+n}(x_1, x_2, \ldots) \le g_n(x_1, x_2, \ldots) + g_m(x_{n+1}, x_{n+2}, \ldots).$$

Seega kehtib (135). Kingmani subaditiivne ergoodiline teoreem väidab, et leidub konstant γ nii, et

$$\frac{g_n(X_1, X_2, \ldots)}{n} = \frac{R_n}{n} \to \gamma, \quad \text{p.k. ja ruumis } L_1.$$

Saab näidata, et

$$\gamma = \mathbf{P}(S_n \neq 0, \quad \forall n).$$

Pikim ühisjada. Olgu X ja Y kaks sõltumatut ergoodilist protsessi tähestikul $\mathcal{X}$. Olgu juhuslik suurus L_n pikima ühisjada (*longest common subsequence*) pikkus. Näiteks jadade 101010101 ja 110100011 pikimad ühisjadad on näiteks 1101001 või 1101011 ja pikima ühisjada pikkus sellisel juhul on 7. Defineerime 2-dimensionaalse protsessi $Z = (X, Y)$, st $Z_n = (X_n, Y_n)$. Saab näidata, et Z_n on ergoodiline protsess. Olgu $l(x_1, \ldots, x_m; y_1, \ldots, y_m)$ jadade $x_1, \ldots, x_m$ ja $y_1, \ldots, y_m$ pikima ühisjada pikkus. Olgu

$$g_n(z_1, z_2, \ldots) = l(x_1, \ldots, x_n; y_1, \ldots, y_n).$$

On selge et $0 \le g_n(z_1, z_2, \ldots) \le n$. Funktsioonid $\{g_n\}$ on superaditiivsed:

$$\begin{aligned} g_{m+n}(z_1, z_2, \ldots) &= l(x_1, \ldots, x_{m+n}; y_1, \ldots, y_{m+n}) \\ &\ge l(x_1, \ldots, x_n; y_1, \ldots, y_n) + l(x_{n+1}, \ldots, x_{n+m}; y_{n+1}, \ldots, y_{n+m}) \\ &= g_n(z_1, z_2, \ldots) + g_m(z_{n+1}, z_{n+1}, \ldots), \end{aligned}$$

kuid funktsioonid $\{-g_n\}$ on subaditiivsed. Kingmani subaditiivsest ergoodilisest teoreemist järeldub, et

$$\frac{g_n(Z_1, Z_2, \ldots)}{n} = \frac{L_n}{n} \to \gamma \quad \text{p.k. ja ruumis } L_1.$$

Juhul, kui X ja Y on mõlemad $B(1, \frac{1}{2})$ jaotusega iid juhuslike suuruste jadad, siis piirväärtust γ nimetatakse *Chvatal-Sankovi* konstandiks. Selle täpne väärtus pole teada, kuid leitud on päris täpsed tõkked, mis näitavad, et $\gamma \approx 0.81$.

6.3 Sagedusteoreem

Olgu $x^n \in \mathcal{X}^n$. Tuletame meelde, et iga tähe $a \in \mathcal{X}$ korral

$$N_{x^n}(a), \quad \text{ja } P_{x^n}(a)$$

on tähe a sagedus ja suhteline sagedus jadas x^n. Vaatleme blokki $a_1^k := a^k \in \mathcal{X}^k$ ja defineerime selle bloki (suhtelise) sageduse jadas x^n:

$$N_{x^n}(a^k) := \sum_{i=1}^{n-k+1} I_{a^k}(x_i, \ldots, x_{i+k-1}), \quad P_{x^n}(a^k) := \frac{1}{n-k+1} N_{x^n}(a^k).$$

Näide: Olgu $\mathcal{X} = \{0, 1\}$, $a^3 = 010$, $x^{11} = 00101011010$. Siis

$$N_{x^{11}}(a^3) = 3, \quad P_{x^{11}}(a^3) = \frac{1}{3}.$$

Paneme tähele, kui T on Bernoulli nihe, $x \in \mathcal{X}^\infty$ ja x^n on esimesed n elementi sellest, siis bloki sagedus avaldub

$$N_{x^n}(a^k) := \sum_{i=0}^{n-k} I_{[a_1^k]}(T^i x). \tag{138}$$

Olgu P statsionaarse prtosessi jaotus ruumil $(X^\infty, \sigma(\mathcal{X}^\infty))$. Vaatleme jada $x \in \mathcal{X}^\infty$. Oletame, et leidub piirväärtus

$$P_x(a^k) := \lim_n P_{x^n}(a^k).$$

Kuulugu hulka $\mathcal{T}(P) \subset \mathcal{X}^\infty$ kõik need jadad x, mille korral *iga* bloki a^k suhtelise sageduse piirväärtus on $P(a^k)$. Seega

$$\mathcal{T}(P) := \{x \in \mathcal{X}^\infty : \lim_n P_{x^n}(a^k) = P(a^k),\ \forall k \ge 1,\ \forall a^k \in \mathcal{X}^k\}.$$

Hulga $\mathcal{T}(P)$ elemente nimetame *sagedustüüpilisteks* jadadeks (realisatsioonideks). Birkhoffi ergoodilisest teoreemist järeldub vahetult.

Teoreem 6.11 (Sagedusteoreem) *Olgu P ergoodiline mõõt ruumil $(X^\infty, \sigma(\mathcal{X}^\infty))$. Siis $P(\mathcal{T}(P)) = 1$, st peaaegu iga x kuulub hulka $\mathcal{T}(P)$.*

Tõestus. Olgu $a^k \in \mathcal{X}^k$ fikseeritud blokk. Siis Birkhoffi ergoodilisest teoreemist saame

$$\frac{1}{n-k+1}\sum_{i=0}^{n-k} I_{[a_1^k]}(T^i x) \to \int I_{[a_1^k]} dP = P(a^k).$$

Seosest (138) saame, et

$$P_{x^n}(a^k) \to P(a^k), \quad P - \text{p.k.}. \tag{139}$$

Et blokke a^k on ülimalt loenduv hulk, on teoreem tõestatud. ■

Tõenäosusteooria keeles on saadud tulemus järgmine.

Järeldus 6.4 *Olgu X ergoodiline protsess. Siis*

$$\mathbf{P}\Big(\frac{1}{n-k+1}\sum_{i=1}^{n-k+1} I_{a^k}(X_i, \dots, X_{i+k-1}) \to \mathbf{P}\big((X_1, \dots, X_k) = a^k\big), \quad \forall k \geq 1,\, a^k \in \mathcal{X}^k\Big) = 1.$$

Selgub, et koondumine (138) on piisav ergoodilisuseks.

Teoreem 6.12 *Olgu P statsionaarne (T-invariantne) mõõt ruumil $(X^\infty, \sigma(\mathcal{X}^\infty))$. Kui iga bloki $a^k \in \mathcal{X}^k$ korral leidub p.k. piirväärtus $P_x(a^k)$ ning see $P_x(a^k)$ on p.k. konstant, siis on P ergoodiline.*

Tõestus. Vastavalt eeldusele on P statsionaarne ja iga a^k korral leidub konstant $c(a^k)$ nii, et

$$\frac{1}{n-k+1}\sum_{i=0}^{n-k} I_{[a_1^k]}(T^i x) \to c(a^k), \quad p.k. \tag{140}$$

Birkhoffi ergoodilisest teoreemist järeldub, et leidub T-invariantne funktsioon $c^*(a^k)$ nii, et

$$\frac{1}{n-k+1}\sum_{i=0}^{n-k} I_{[a_1^k]}(T^i x) \to c^*(a^k)$$

ruumis L_1 ja p.k. Seega $c^*(a^k) = c(a^k)$. Et koondumine toimub ka ruumis L_1, siis n kasvades

$$\frac{1}{n-k+1}\sum_{i=0}^{n-k} \int I_{[a^k]}(T^i x) P(dx) = P(a^k) \to c^*(a^k),$$

millest $c(a^k) = P(a^k)$. Olgu $C \in \sigma(\mathcal{X}^\infty)$ suvaline. Koondumisest (140) saame

$$\frac{1}{n-k+1}\sum_{i=0}^{n-k} I_{[a_1^k]}(T^i x) I_C(x) \to P(a^k) I_C(x), \quad \text{p.k.}$$

Domineeritud koondumise teoreemist saame

$$\frac{1}{n-k+1}\sum_{i=0}^{n-k}\int I_{[a_1^k]}(T^i x)I_C(x)P(dx)\to P(a_1^k)P(C).$$

Et aga

$$I_{[a_1^k]}(T^i x)I_C(x)=I_{T^{-i}[a_1^k]}(x)I_C(x),$$

siis

$$\frac{1}{n-k+1}\sum_{i=0}^{n-k}\int I_{[a_1^k]}(T^i x)I_C(x)P(dx)=\frac{1}{n-k+1}\sum_{i=0}^{n-k}P(T^{-i}[a_1^k]\cap C)\to P(a_1^k)P(C).$$

Ülaltoodud argument kehtib iga silindri korral. Seega kehtib ta ka iga $B\in\sigma(\mathcal{X}^\infty)$ korral, millest saame, et iga $B,C\in\sigma(\mathcal{X}^\infty)$ korral

$$\frac{1}{n}\sum_{i=0}^{n-1}P(T^{-i}B\cap C)\to P(B)P(C). \tag{141}$$

Et koondumisest (141) järeldub ergoodilisus, seda ma juba teame. ■

Sama tulemus tõenäosusteooria keeles.

Järeldus 6.5 *Statsionaarne protsess X on ergoodiline parajasti siis, kui iga bloki a^k korral leidub konstant $c(a^k)$ nii, et*

$$\frac{1}{n}\sum_{i=1}^{n}I_{a^k}(X_i,\dots,X_{i+k-1})\to c(a^k),\quad p.k..$$

Sellisel juhul

$$c(a^k)=\mathbf{P}\big((X_1,\dots,X_k)=a^k\big).$$

Sagedusteoreemist järeldub ka järgmine ergoodilisuse kriteerium.

Järeldus 6.6 *Olgu P statsionaarne mõõt ruumil $(X^\infty,\sigma(\mathcal{X}^\infty))$. Järgmised väited on samaväärsed*

1 *P on ergoodiline;*

2 *iga bloki a^k korral kehtib*

$$\frac{1}{n}\sum_{i=0}^{n-1}I_{[a_1^k]}(T^i x)\to P(a^k);$$

3 *iga $A,B\in\sigma(\mathcal{X}^\infty)$ korral*

$$\frac{1}{n}\sum_{i=0}^{n-1}P(T^{-i}B\cap C)\to P(B)P(C).$$

Olgu

$$T_\epsilon^n(k) := \{x^n \in \mathcal{X}^n : |P_{x^n}(a^k) - P(a^k)| < \epsilon, \quad \forall a^k \in \mathcal{X}^n\}.$$

Kui $|\mathcal{X}| < \infty$, siis sagedusteoreemist järeldub, et iga k korral

$$P_n(T_\epsilon^n(k)) \to 1,$$

kus P_n on mõõdu P ahend ruumile $(\mathcal{X}^n, \sigma(\mathcal{X}^n))$.

Juhul, kui $k = 1$, on $T_\epsilon^n(1)$ sisuluselt tugevalt tüüpiliste sõnade hulk T_ϵ^n ning koondumisest $P_n(T_\epsilon^n(1)) \to 1$ järeldub tugev AEP: sobiva ϵ ja lõpliku tähestiku korral rahuldab T_ϵ^n teoreemi 5.18 väiteid.

6.4 Shannon-McMillian-Breimani teoreem

Olgu X iid protsess. Sellisel juhul järeldub tugevast suurte arvude seadusest vahetult, et

$$\lim_n -\frac{1}{n} \log P(X_1, \ldots, X_n) \to H_X, \quad \text{p.k.} \tag{142}$$

kus H_X on protsessi entroopia. Antud juhul (iid protsess), $H_X = H(X_1)$.
Olgu X ergoodiline protsess. Sellisel juhul (142) vahetult Birkhoffi ergoodilisest teoreemist ei järeldu. Selgub aga, et koondumine (142) kehtib. See nn. *Shannon-McMillian-Breimani* teoreem on üks ergoodilise teooria põhitulemustest.

Teoreem 6.13 (Shannon-McMillian-Breimani teoreem) *Olgu P ergoodiline mõõt ruumil $(\mathcal{X}^\infty, \sigma(\mathcal{X}^\infty))$. Siis*

$$\lim_n -\frac{1}{n} \log P(x^n) = \limsup_n -\frac{1}{n} \sum_{x^n \in \mathcal{X}^n} P(x^n) \log P(x^n), \quad P\text{-p.k.}.$$

Et P on ergoodiline ja seetõttu nihke-invariantne (statsionaarne), siis

$$-\limsup_n \frac{1}{n} \sum_{x^n \in \mathcal{X}^n} P(x^n) \log P(x^n) = -\lim_n \frac{1}{n} \sum_{x^n \in \mathcal{X}^n} P(x^n) \log P(x^n)$$

ja piirväärtus on vastava protsessi entroopia. Seega tõenäosusteooria keeles on Shannon-McMillian-Breimani teoreem järgmine.

Teoreem 6.14 (Shannon-McMillian-Breimani teoreem) *Olgu $X = \{X_n\}_{n=1}^\infty$ ergoodiline protsess. Siis kehtib koondumine (142).*

Shannon-Macmillian-Breimani teoreemist järeldub, et ergoodilisel protsessil on nõrk AEP omadus, s.t. hulk

$$W_\epsilon^n = \{x^n \in \mathcal{X}^n : 2^{-n(H_X+\epsilon)} < P(x^n) < 2^{-n(H_X-\epsilon)}\}$$

rahuldab teoreemi 3.2 väiteid.

Ülesanne: Olgu X ergoodiline MA. Tõestada (142).

6.5 Kac'i lemma

Olgu $X_0, X_1, X_2, \ldots$ jaotusega P iid jada tähestikul $\mathcal{X}$. Olgu $a \in \mathcal{X}$ ning T_a esimene a ilmumine jadas $X_1, X_2, \ldots$:

$$T_a = \min\{i \geq 1 : X_i = a\}.$$

Elementaarsest tõenäosusteooriast on teada, et $T_a \sim G(P(a))$ ja

$$ET_a = \frac{1}{P(a)}.$$

Võime kirjutada, et

$$E(T_a|X_0 = a) = \frac{1}{P(a)}. \tag{143}$$

Olgu $X_0, X_1, X_2, \ldots$ ergoodiline MA, $X_0 \sim P$. Seega P on statsionaarne algjaotus ja X on taandumatu. Markovi ahelate teooriast on teada, et sellisel juhul

$$P(a) = \frac{1}{E(T_a|X_0 = a)}$$

ehk kehtib (143) (vt näiteks *Durret* (4.6)). Järgnev nn. Kac'i lemma väidab, et (143) kehtib kõikide ergoodiliste protsesside korral.

Lemma 6.1 *Olgu $X_0, X_1, X_2, \ldots$ ergoodiline protsess. Olgu $a \in \mathcal{X}$ nii, et $P(a) > 0$. Siis kehtib (143).*

Tõestus. Vaatleme kahepoolset protsessi

$$\ldots, X_{-2}, X_{-1}, X_0, X_1, X_2, \ldots$$

Valime a. Defineerime

$$A_{kj} = \{X_{-k} = a, X_l \neq a, -k < l < j, X_j = a\}.$$

Sündmused A_{jk} on lõikumatud ning, et $P(a) > 0$, siis X ergoodilisuse tõttu

$$\mathbf{P}(\cup_{j\geq 1} \cup_{k\geq 0} A_{kj}) = 1.$$

Seega

$$\begin{aligned}
1 &= \sum_{j=1}^{\infty}\sum_{k=0}^{\infty}\mathbf{P}(A_{kj}) = \sum_{j=1}^{\infty}\sum_{k=0}^{\infty}\mathbf{P}(X_{-k}=a)\mathbf{P}\big(X_l\neq a,\ -k<l<j, X_j=a|X_{-k}=a\big)\\
&= \sum_{j=1}^{\infty}\sum_{k=0}^{\infty}\mathbf{P}(X_{-k}=a)P(T_a=j+k|X_0=a)\\
&= \sum_{j=1}^{\infty}\sum_{k=0}^{\infty}P(a)\mathbf{P}(T_a=j+k|X_0=a)\\
&= P(a)\sum_{j=1}^{\infty}\sum_{k=0}^{\infty}\mathbf{P}(T_a=j+k|X_0=a)\\
&= P(a)\sum_{i=1}^{\infty}\mathbf{P}(T_a=i|X_0=a) = P(a)E(T_a|X_0=a).
\end{aligned}$$

Viimane võrdus tuleneb sellest, et paare (j,k), mille summa võrdub i on i tükki. ■

6.6 Statsionaarne kujutis

Olgu lisaks tähestikule $\mathcal{X}$ antud veel tähestik $\mathcal{Y}$. Vaatleme ruume $(\mathcal{X}^\infty, \sigma(\mathcal{X}^\infty))$ ja $(\mathcal{Y}, \sigma(\mathcal{Y}^\infty))$. Olgu T_x ja T_y vastavalt ruumides $\mathcal{X}^\infty$ ja $\mathcal{Y}^\infty$ antud nihked.

Mõõtuvat kujutist

$$F : \mathcal{X}^\infty \to \mathcal{Y}^\infty$$

nimetatakse *statsionaarseks* (*stationary coding*), kui

$$F(T_x x) = T_y F(x).$$

Seega F on stasionaarne, kui iga $x = x_1, x_2, \ldots$ korral $F_1, F_2, \ldots := F(x_1, x_2, \ldots)$ rahuldab tingimust

$$F_2, F_3, \ldots = F(x_2, x_3, \ldots).$$

Olgu nüüd tähestikul $\mathcal{X}$ antud statsionaarne protsess $X_1, X_2, \ldots$. Rakendades protsessile X statsionaarset kujutist F saame juhusliku protsessi $F(X)$:

$$F(X_1, X_2, \ldots) = F_1, F_2, \ldots,$$

mis samuti on statsionaarne. Juhul kui X on ergoodiline, on $F(X)$ samuti ergoodiline protsess. Nendes väidetes on kerge veenduda, kasutades statsionaarse protsessi Kolmogorovi esitust.

Ülesanne: Olgu P ruumil $(\mathcal{X}^\infty, \sigma(\mathcal{X}^\infty))$ mõõt. Olgu F statsionaarne kujutis. Tõestada, et ruumil $(\mathcal{Y}, \sigma(\mathcal{Y}^\infty))$ antud mõõt PF^{-1} on:

- T_y-invariantne, kui P on T_x-invariantne;
- ergoodiline, kui P on ergoodiline.

Kuigi lihtne tõestada, on toodud tulemus tihti kasulik rakendustes. Näitena vaatleme Kaci lemma üldistamist blokkidele. Olgu $X_0, X_1, \ldots$ ergoodiline protsess ning vaatleme blokke

$$(X_n, X_{n+1}, \ldots, X_{n+K}), \quad n = 0, 1, \ldots,$$

kus $K \geq 1$ on fikseeritud täisarv. Olgu $a^{K+1} \in \mathcal{X}^{K+1}$ fikseeritud blokk pikkusega $K+1$ ning uurime peatumishetke

$$T := \min\{n \geq 1 : (X_n, X_{n+1}, \ldots, X_{n+K}) = a^{K+1}\}.$$

Defineerime *blokk-protsessi* $Y = Y_0, Y_1, \ldots$ nii, et

$$Y_n = (X_n, X_{n+1}, \ldots, X_{n+K}).$$

Seega protsess Y võtab väärtusi tähestikul $\mathcal{Y} = \mathcal{X}^{K+1}$. On kerge näha, et leidub statsionaarne kujutis F (milline?) nii, et $Y = F(X)$. Seega Y on ergoodiline protsess ja Kac' lemma rakendub probleemideta:

$$E[T|(X_0, \ldots, X_K) = a^{K+1}) = E(T|Y_0 = a^{K+1}) = \frac{1}{\mathbf{P}(Y_0 = a^{K+1})} = \frac{1}{\mathbf{P}\big((X_0, \ldots, X_K) = a^{K+1}\big)}.$$

7 Lempel-Ziv kood

Peatükis 5 käsitlesime plokk-koode

$$\mathcal{C}_n : \mathcal{X}^n \to \mathcal{D}^m.$$

Selliste koodide korral on koodisõnade pikkus etteantud, mistõttu kodeerimisviga on enamasti vältimatu. Nägime, et kui koodi määr on piisavalt suur ja informatsiooniallikas on iid protsess, saab n valides teha kodeerimisvea kuitahes väikeseks ja seda isegi *universaalselt*.

Peatükist 2 teame, et kui informatsiooniallikas on statsionaarne protsess, siis leiduvad (mitteuniversaalsed) prefikskoodid

$$\mathcal{C}_n : \mathcal{X}^n \to \{0,1\}^*$$

nii, et keskmised koodipikkused tähe kohta koonduvad protsessi entroopiamääraks:

$$L_n = \frac{1}{n} El(X_1, \ldots, X_n) \to H_X.$$

Kui informatsiooniallikas on lisaks ergoodiline, siis Shannon-MacMillian-Breimani teoreemist saame, et leiduvad (mitteuniversaalsed) prefikskoodid – Shannon-Fano koodid – mille korral entroopiamääraks koondumine kehtib ka peaaegu kindlasti:

$$\frac{l(X_1, \ldots, X_n)}{n} \to H_X, \quad \text{p.k..} \tag{144}$$

Käesolevas peatükis vaatleme prefikskoode

$$\mathcal{C}_n : \mathcal{X}^n \to \{0,1\}^*$$

ja uurime jada $l(X_1, \ldots, X_n)$ asümptootilist käitumist. Näitame, et koondumine (144) on teatavas mõttes parim, st iga prefikskoodide jada korral kehtib

$$\liminf_n \frac{l(X_1, \ldots, X_n)}{n} \geq H_X, \quad \text{p.k..}$$

Sellest johtuvalt nimetame koodide jada $\mathcal{C}_n$ on **asümptootiliselt optimaalseks**, kui

$$\frac{1}{n} l(X_1, \ldots, X_n) \to H_X, \text{ p.k..}$$

Teame, et Shannon-Fano kood on asümptootiliselt optimaalne, kuid ta pole universaalne. Käesolevas peatükis käsitleme nn **Lempel-Ziv (LZ)** koode, mis on universaalsed asümptootiliselt optimaalsed koodid. Kõik LZ koodid põhinevad liigendamisel, meie käsitletav kood kannab teinekord nimetust LZ78 või ka puu struktruuriga (*tree-structured*) LZ kood. 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.

7.1 Liigendamine ja kodeerimine

Olgu $\mathcal{X}$ lõplik tähestik. Vektorit $x^n \in \mathcal{X}^n$ käsitleme sisendina. Lempel-Ziv (LZ78) kodeerimine põhineb sisendi x^n **liigendamisel** (*parsing*). Liigendamine on vektori 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 $x^{n_j} = w(1) \cdots w(j)$.

$$\begin{aligned} &\text{kui } x_{n_j+1} \notin \{w(1), \ldots, w(j)\}, \text{ siis } w(j+1) = x_{n_j+1}, \\ &\text{kui } x_{n_j+1} \in \{w(1), \ldots, w(j)\}, \text{ siis } w(j+1) = x_{n_j+1}^{m+1}, \end{aligned}$$

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 $x^{17} = 11001010001000100$, siis liigendus on järgmine:

$$1, 10, 0, 101, 00, 01, 000, 100, 010, 0.$$

Pärast liigendust esitub sisendvektor sõnade jadana:

$$x^n = w(1)w(2) \cdots w(K)v, \tag{145}$$

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. Kodeerides sõnade indeksid ja viimased tähed saame 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 (145) järgmise eeskirja alusel.

7.1.1 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 $n = 17$. Siis $f : \{1,\ldots,17\} \to \{0,1\}^5$. Olgu $f(i)$ arvu $i-1$ kahendesitus. Leiame $C_{17}(11001010001000100)$. Toodud vektori liigendus on meile tuttav:

$$w(1) = 1,\ w(2) = 10,\ w(3) = 0,\ w(4) = 101,\ w(5) = 00,$$
$$w(6) = 01,\ w(7) = 000,\ w(8) = 100,\ w(9) = 010,\ v = 0.$$

Seega $K = 9$ ja $b(1) = 0g(1) = 01$, $b(2) = 1f(1)g(0) = 1000000$, $b(3) = 0g(0) = 00$, $b(4) = 1f(2)g(1) = 1000011$, $b(5) = 1f(3)g(0) = 1000100$, $b(6) = 1f(3)g(1) = 1000101$, $b(7) = 1f(5)g(0) = 1001000$, $b(8) = 1f(2)g(0) = 1000010$, $b(9) = 1f(6)g(0) = 1001010$, $b(10) = 1f(3) = 100010$. Seega

$C_{17}(11001010001000100) = 011000000001000011100010010001011001000100001010010101000010.$

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 $K = 9$, seega võib võtta $f : \{1,\ldots,9\} \to \{0,1\}^4$. Selline f sõltub aga sisendist x^{17} 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.

7.2 Lempel-Ziv teoreem

Olgu

$$\mathcal{C}_n : \mathcal{X}^n \to \{0,1\}^*$$

LZ kahendkood, $l(x^n) := |C(x^n)|$ olgu koodisõna x^n pikkus. Meid huvitab suuruse $\frac{l(x^n)}{n}$ – ühe tähe kodeerimiseks keskmiselt kulunud bittide arv – asümptootiline käitumine (n kasvades). Selle uurimiseks on vaja üht-teist teada sisendist x^n. Nagu ikka, olgu informatsiooniallikas $X := X_1, X_2, \ldots$ juhuslik protsess tähestikul $\mathcal{X}$, x^n on siis juhusliku vektori $X^n = (X_1,\ldots,X_n)$ realisatsioon. Lempel-Zivi teoreem väidab, et kui protsess X on ergoodiline entroopiamääraga H_X, siis

$$\limsup_n \frac{l(X^n)}{n} = H_X, \quad \text{p.k.} \tag{146}$$

Hindame suurust $l(x^n)$. Vaatleme veelkord LZ 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 $B := 1 + \lceil \log |\mathcal{X}| \rceil$. Osa **3)** nõuab ülimalt

$$\lceil \log n \rceil + 1$$

bitti. Tuletame meelde, et liigendusel saadud sõnade arv $K = K(x^n)$ sõltub sisendist x^n. Seega

$$\begin{aligned} l(x^n) &\le A + K(\lceil \log n \rceil + B) + \lceil \log n \rceil + 1 \le (\log n + 1)(K + 1) + KB + A + 1 \\ &= K \log n + \log n + K(B + 1) + A + 2. \end{aligned}$$

Et A ja B on konstandid, on domineeriv liige $K \log n$. LZ teoreemi tõestus seisnebki seose

$$\limsup_n \frac{K(X^n) \log n}{n} = H_X, \quad \text{p.k.} \tag{147}$$

näitamises.

7.2.1 Kombinatoorika

Olgu $t > 1$. Siis

$$\sum_{j=1}^{m} t^j = \frac{t^{m+1} - t}{t - 1},$$

millest saame hinnangud

$$\frac{t}{t-1} m t^m \Big(1 - \frac{1}{m(t-1)}\Big) \le \sum_{j=1}^{m} j t^j \le \frac{t}{t-1} m t^m. \tag{148}$$

Arv $K(x^n)$ on suur siis, kui x^n sisaldab lühikesi sõnu; $K(x^n)$ on kõige suurem, kui x^n sisaldab võimalikult palju ühetähelisi sõnu, võimalikult palju kahetähelisi sõnu jne. Kui $n = \sum_{j=1}^{m} j|\mathcal{X}|^j$, siis maksimaalne liigendamisel saadud sõnade arv on $\sum_{j=1}^{m} |\mathcal{X}|^j$. Kui $m(n)$ on selline, et

$$\sum_{j=1}^{m} j|\mathcal{X}|^j \le n < \sum_{i=1}^{m+1} j|\mathcal{X}|^j,$$

siis

$$K(m(n)) \le \sum_{j=1}^{m} |\mathcal{X}|^j + k_2,$$

kus $k_2(m(n)) < |\mathcal{X}|^{m+1}$. Tähistame $t = |\mathcal{X}|$,

$$k_1(m(n)) = \sum_{j=1}^{m} t^j \le \frac{t^{m+1}}{t-1}, \quad k_2(m) < t^{m+1}.$$

Seega $K(m) \le k_1(m) + k_2(m)$. Sellisel juhul (148) annavad meile

$$\frac{t}{t-1} m t^m \Big(1 - \frac{1}{m(t-1)}\Big) \le \frac{t}{t-1} m t^m \Big(1 - \frac{1}{m(t-1)}\Big) + k_2(m+1) \le \sum_{j=1}^{m} j t^j + k_2(m+1) \le$$

$$n \le \sum_{j=1}^{m} j t^j + (k_2+1)(m+1) \le \sum_{j=1}^{m+1} j t^j \le \frac{t}{t-1}(m+1)t^{m+1}.$$

Sellest saame, et

$$\log n \le \log \frac{t}{t-1} + \log(m+1) + (m+1)\log t$$

ja

$$\begin{aligned}
\frac{(k_1+k_2)\log n}{n} \le & \frac{(k_1+k_2)\Big(\log \frac{t}{t-1} + \log(m+1) + (m+1)\log t\Big)}{n} \\
\le & \frac{(k_1+k_2)\Big(\log \frac{t}{t-1} + \log(m+1) + (m+1)\log t\Big)}{\frac{t}{t-1} m t^m \Big(1 - \frac{1}{m(t-1)}\Big) + k_2(m+1)} \\
\le & \frac{(\frac{t^{m+1}}{t-1} + k_2)\Big(\log \frac{t}{t-1} + \log(m+1) + (m+1)\log t\Big)}{\frac{m t^{m+1}}{t-1} \Big(1 - \frac{1}{m(t-1)}\Big) + k_2(m+1)} \\
= & \frac{(1+\beta(m))\Big(\log \frac{t}{t-1} + \log(m+1) + (m+1)\log t\Big)}{m\Big(1 - \frac{1}{m(t-1)}\Big) + \beta(m)(m+1)} \\
= & \frac{(1+\beta(m))\Big(\alpha(m) + \log t\Big)}{\frac{m}{m+1}\Big(1 - \frac{1}{m(t-1)}\Big) + \beta(m)} \\
= & \Big(\frac{1+\beta(m)}{\gamma(m)+\beta(m)}\Big)\Big(\alpha(m) + \log t\Big) \\
= & \Big(\frac{1+\beta(m)}{\gamma(m)+\beta(m)}\Big)\alpha(m) + \Big(\frac{1+\beta(m)}{\gamma(m)+\beta(m)}\Big)\log t
\end{aligned}$$

kus protsessis $m(n) \to \infty$

$$\beta(m) := \frac{k_2(m)}{\frac{t^{m+1}}{t-1}} \le t-1, \quad \alpha(m) := \frac{\log \frac{t}{t-1} + \log(m+1)}{m+1} \to 0,$$

$$\gamma(m) := \frac{m}{m+1}\Big(1 - \frac{1}{m(t-1)}\Big) \to 1.$$

Seega

$$\Big(\frac{1+\beta(m)}{\gamma(m)+\beta(m)}\Big)\alpha(m) \to 0, \quad \Big(\frac{1+\beta(m)}{\gamma(m)+\beta(m)}\Big)\log t \to \log t,$$

siis

$$\limsup_n \frac{K(n)\log n}{n} \le \log|\mathcal{X}|. \tag{149}$$

7.2.2 Topoloogiline entroopia

Olgu $x \in \mathcal{X}^\infty$. Defineerime

$$\mathcal{U}_k(x) := |\{a^k : x_i^{i+k-1} = a^k \text{ mingi } i \text{ korral }\}|.$$

Seega on $\mathcal{U}_k(x)$ kõigi jadas x sisalduvate erinevate k-elemendiliste blokkide arv. Järelikult

$$0 \le \mathcal{U}_k(x) \le |\mathcal{X}|^k.$$

On selge, et

$$\mathcal{U}_{m+n}(x) \le \mathcal{U}_m(x)\mathcal{U}_n(x),$$

millest

$$\log\mathcal{U}_{m+n}(x) \le \log\mathcal{U}_m(x) + \log\mathcal{U}_n(x), \tag{150}$$

ja subaditiivsuse tõttu leidub piirväärtus

$$h(x) := \lim_n \frac{1}{n}\log\mathcal{U}_n(x), \tag{151}$$

mida nimetatakse **topoloogiliseks entroopiaks**. Topoloogiline entroopia võtab arvesse kõikide jadas x olevate blokkide arvu kuid mitte nende sagedust.

Ülesanne: Olgu $X = X_1, X_2, \ldots$ iid Bernoulli juhuslikud suurused parameetriga $0 < p < 1$. Olgu P protsessi X jaotus ning $x \in \mathcal{T}(P)$, st x on sagedustüüpiline realisatsioon. Leida $h(x)$.

Topoloogilise entroopia abil saab tõket (149) täpsustada. Olgu

$$h_j(x) := \frac{1}{j}\log\mathcal{U}_j(x).$$

Siis on selge, et kui

$$n = \sum_{j=1}^m j\mathcal{U}_j(x) = \sum_{j=1}^m j2^{jh_j(x)},$$

siis

$$K(x^n) \le \sum_{j=1}^m \mathcal{U}_j(x) = \sum_{j=1}^m 2^{jh_j(x)}. \tag{152}$$

Et

$$h(x) = \inf_j h_j(x),$$

siis $h(x) + \epsilon_j$, kus $\epsilon_j \searrow 0$, siis

$$\sum_{j=1}^{m} 2^{jh_j(x)} = \sum_{j=1}^{m} 2^{j(h(x)+\epsilon_j)} = \sum_{j=1}^{m} (2^{h(x)+\epsilon_j})^j = \sum_{j=1}^{m} (t2^{\epsilon_j})^j = \sum_{j=1}^{m} (t\delta_j)^j,$$

kus

$$t = 2^{h(x)} \text{ ja } \delta_j = 2^{\epsilon_j} \to 1.$$

Seega sellise n korral

$$\frac{K(x^n)}{n} \le \frac{\sum_{j=1}^{m} (t\delta_j)^j}{\sum_{j=1}^{m} j(t\delta_j)^j}.$$

Et $\delta_j \to 1$, saab näidata, et

$$m \frac{\sum_{j=1}^{m} (t\delta_j)^j}{\sum_{j=1}^{m} j(t\delta_j)^j} \to 1.$$

Seega võrratusest (152) saame, et vaadeldava n korral

$$\begin{aligned}
\frac{K(x^n)\log n}{n} &\le \frac{\sum_{j=1}^{m} (t\delta_j)^j \log(\sum_{j=1}^{m} j(t\delta_j)^j)}{\sum_{j=1}^{m} j(t\delta_j)^j} \\
&\le \frac{\sum_{j=1}^{m} (t\delta_j)^j \Big(\log(\sum_{j=1}^{j_o} j(t\delta_j)^j) + \log(\sum_{j=j_o+1}^{m} j(t\delta_j)^j)\Big)}{\sum_{j=1}^{m} j(t\delta_j)^j} \\
&\le o(1) + \frac{\sum_{j=1}^{m} (t\delta_j)^j \log(\sum_{j=j_o}^{m} j(t\delta_j)^j)}{\sum_{j=1}^{m} j(t\delta_j)^j} \\
&\le o(1) + \frac{\sum_{j=1}^{m} (t\delta_j)^j \log(m \sum_{j=1}^{m} (t\delta_{j_o})^j)}{\sum_{j=1}^{m} j(t\delta_j)^j} \\
&\le o(1) + \frac{\sum_{j=1}^{m} (t\delta_j)^j \Big(\log m + \log\big(\frac{(t\delta_{j_o})^{m+1}}{(t\delta_{j_o})-1}\big)\Big)}{\sum_{j=1}^{m} j(t\delta_j)^j} \\
&\le o(1) + \frac{\sum_{j=1}^{m} (t\delta_j)^j}{\sum_{j=1}^{m} j(t\delta_j)^j} \Big(\log m + (m+1)\log(t\delta_{j_o}) - \log((t\delta_{j_o})-1)\Big).
\end{aligned}$$

Siin j_o on selline, et $\delta_j \le \delta_{j_o}$, kui $j \ge j_o$. Et

$$\frac{\sum_{j=1}^{m} (t\delta_j)^j}{\sum_{j=1}^{m} j(t\delta_j)^j} \sim \frac{1}{m},$$

siis koondub võrratusteahela parem pool suuruseks

$$\log(t\delta_{j_o}),$$

mille j_o valides saame teha kuitahes lähedaseks suurusele $\log t = h(x)$.
Vaadeldes

$$\sum_{j=1}^{m} j2^{jh_j(x)} \le n < \sum_{j=1}^{m+1} j2^{jh_j(x)},$$

on ülaltoodud idee võimalik formaliseerida (analoogiliselt tõkke (149) tõestusega) järgmiseks lemmaks.

Lemma 7.1 *Iga $x \in \mathcal{X}^\infty$ korral*

$$\limsup_n \frac{K(x^n)\log n}{n} \le h(x) \tag{153}$$

7.2.3 $\bar{d}$-poolmeetrika

Defineerime hulgal $\mathcal{X}^n$ meetrika d_n järgmise eeskirja alusel:

$$d_n(x^n, y^n) = \frac{1}{n}\sum_{i=1}^{n} d(x_i, y_i), \quad \text{kus } d(a,b) = \begin{cases} 0, & \text{kui } a = b; \\ 1, & \text{kui } a \ne b \end{cases}$$

Meetrikat d_n nimetatakse **Hammingu kauguseks** (teinekord ka Hammingu kauguseks tähe kohta).

Defineerime hulgal $\mathcal{X}^\infty$ poolmeetrika

$$\bar{d}(x,y) := \limsup_n d_n(x^n, y^n).$$

Ülesanne: Veenduda, et $\bar{d}$ on poolmeetrika.

Olgu $S \subset \mathcal{X}^n$,

$$d_n(x^n, S) = \inf_{s\in S} d_n(x^n, s), \quad [S]_\delta := \{x^n : d_n(x^n, S) \le \delta\}.$$

Lemma 7.2

$$|[S]_\delta| \le |S| 2^{nh(\delta)} (|\mathcal{X}| - 1)^{n\delta}. \tag{154}$$

Tõestus. Paneme tähele: kui $\delta < 0.5$, siis

$$\sum_{k \le n\delta} C_k^n \le 2^{nh(\delta)}. \tag{155}$$

Tõepoolest, funktsioon

$$q \mapsto q\log\frac{1}{\delta} + (1-q)\log\frac{1}{1-\delta}$$

on kasvav, sest $\delta < 0.5$. Sellest saame, et iga $q \le \delta$ korral

$$-nh(\delta) \le n(q\log\delta + (1-q)\log(1-\delta)).$$

Seega, kui $k \leq n\delta$, st $k = nq$, siis

$$2^{-nh(\delta)} \leq 2^{nq\log\delta + n(1-q)\log(1-\delta)} = \delta^{nq}(1-\delta)^{n(1-q)} = \delta^k(1-\delta)^{n-k},$$

millest

$$2^{-nh(\delta)} \sum_{k\leq n\delta} C_k^n \leq \sum_{k\leq n\delta} C_k^n \delta^k (1-\delta)^{n-k} \leq \sum_k C_k^n \delta^k (1-\delta)^{n-k} = 1.$$

Kuulugu $x^n \in S$. Siis $[x^n]_\delta$ koosneb vektoritest y^n, mis erineb x^n-st ülimalt $n\delta$ elemendi võrra. Neid vektoreid, mis erinevad x^n-st k elemendi võrra on $(|\mathcal{X}|-1)^k \leq (|\mathcal{X}|-1)^{n\delta}$. Hinnangust (154) saame

$$|[x^n]_\delta| \leq 2^{nh(\delta)}(|\mathcal{X}|-1)^{n\delta}$$

ja

$$|[S]_\delta| \leq |S| 2^{nh(\delta)}(|\mathcal{X}|-1)^{n\delta}.$$

■

7.2.4 Täpsustatud tõke

Lemma 7.3 *Olgu $x \in \mathcal{X}^\infty$. Iga $\epsilon > 0$ korral leidub $\delta > 0$ nii, et kui $\bar{d}(x,y) < \delta$, siis*

$$\limsup_n \frac{K(x^n)\log n}{n} \leq h(y) + \epsilon. \tag{156}$$

Tõestus. Olgu $w(1)w(2)\cdots$ jada x liigendus. Olgu y selline, et $\bar{d}(x,y) < \delta$ ning jagame jada y samapikkusteks tükkideks $v(1)v(2)\cdots$ (see ei ole y liigendus). Olgu $l(i) := |w(i)| = |v(i)|$ ning vaatleme sõnade $w(i)$ ja $v(i)$ erinevust, mida mõõdame $d_{l(i)}(w(i), v(i))$. Ütleme, et $w(i)$ ja $v(i)$ on sarnased, kui

$$d_{l(i)}(w(i), v(i)) < \sqrt{\delta},$$

vastasel juhul olgu nad mittesarnased. Olgu $I_1(x)$ mittesarnaste sõnade indeksite hulk. Vaatleme nüüd lõplikku vektorit x^n, liigendame selle ning saadud liigenduse järgi ka y^n. Analoogiliselt defineerime mittesarnaste sõnade indeksite hulga $I_1(x^n)$. Olgu $K_1(x^n)$ hulga $I_1(x^n)$ võimsus, seega mittesarnaste sõnade arv. Veendume, et piisavalt suure n korral

$$\sum_{i\in I_1(x^n)} l(i) \leq n\sqrt{\delta}. \tag{157}$$

Tõepoolest, et $\bar{d}(x,y) < \delta$, siis piisavalt suure n korral

$$d_n(x^n, y^n) < \delta,$$

millest $nd_n(x^n, y^n) < n\delta$ ning

$$n\delta > nd_n(x^n, y^n) = \sum_{i\leq K(x^n)+1} l(i) d_{l(i)}(w(i), v(i)) \geq \sum_{i\in I_1(x^n)} l(i) d_{l(i)}(w(i), v(i)) \geq \sum_{i\in I_1(x^n)} l(i)\sqrt{\delta}.$$

Seega mittesarnased sõnad moodustavad ülimalt $\sqrt{\delta}n$ osa vektorist x^n. Tõkkest (149) saame

$$\limsup_n \frac{K_1(x^n)\log(\sqrt{\delta}n)}{\sqrt{\delta}n} \le \log|\mathcal{X}|$$

ehk

$$\limsup_n \frac{K_1(x^n)\log n}{n} \le \sqrt{\delta}\log|\mathcal{X}|, \tag{158}$$

sest

$$\log(n\sqrt{\delta}) \sim \log(n).$$

Vaatleme nüüd sarnaseid sõnu. Olgu $G_k(x)$ pikkusega k sarnaste sõnade hulk ja $G_k(y)$ olgu neile vastavate sõnade hulk jadast y. Kui $w(i) \in G_k(x)$, siis $d_k(w(i), v(i)) \le \sqrt{\delta}$, millest võrratuse (154) abil saame

$$|G_k(x)| \le |G_k(y)|2^{kh(\delta)}|\mathcal{X}|^{k\delta}. \tag{159}$$

On selge, et

$$|G_k(y)| \le \mathcal{U}_k(y) = 2^{kh_k(y)},$$

kus, nagu ikka,

$$h_k(x) = \frac{\log \mathcal{U}_k(x)}{k}.$$

Võrratuse (159) parem pool on seega

$$|G_k(y)|2^{kh(\delta)}|\mathcal{X}|^{k\delta} \le 2^{k(h_k(y)+h(\delta)+\delta\log|\mathcal{X}|)}.$$

Et $h_k(y) \to h(y)$, siis leidub piisavalt suur k_o nii, et $h_k(y) \le h(y) + \frac{\epsilon}{2}$, kui $k > k_o$. Sellise k korral

$$|G_k(x)| \le 2^{k(h(y)+\frac{\epsilon}{2}+h(\delta)+\delta\log|\mathcal{X}|)}.$$

Võttes nüüd $G_k(x)$ $\mathcal{U}_k(x)$ rolli, saame võrratusest (153), et iga $\epsilon > 0$ korral

$$\limsup_n \frac{K_2(x^n)\log n}{n} \le \limsup_k \frac{\log|G_k(x)|}{k} \le h(y) + \frac{\epsilon}{2} + h(\delta) + \delta\log|\mathcal{X}|),$$

kus $K_2(x)$ on sarnaste sõnade arv. Et

$$h(\delta) + \delta\log|\mathcal{X}| + \sqrt{\delta}\log|\mathcal{X}| \to 0,$$

kui $\delta \to 0$, siis leidub $\delta(\epsilon) > 0$ nii, et

$$\begin{aligned}
\limsup_n \frac{K(x^n)\log n}{\log n} &\le \limsup_n \frac{K_1(x^n)\log n}{n} + \limsup_n \frac{K_2(x^n)\log n}{n} \\
&\le \sqrt{\delta}\log|\mathcal{X}| + \limsup_k \frac{\log G_k(x)}{k} \\
&\le \sqrt{\delta}\log|\mathcal{X}| + h(\delta) + \delta\log|\mathcal{X}| + h(y) + \frac{\epsilon}{2} \\
&\le h(y) + \epsilon,
\end{aligned}$$

■

Lemmast 7.3 järeldub

Järeldus 7.1

$$\limsup_n \frac{K(x^n)\log n}{n} \leq \lim_{\delta\to 0} \inf_{y:\bar{d}(x,y)<\delta} h(y) =: H_Z(x).$$

Ülesanne: Tõestada järeldus.

Märkus: Suurust $H_Z(x)$ nimetatakse ka **Zivi entroopiaks**.

7.2.5 Ehituskivid

Olgu $B_n \subset \mathcal{X}^n$ – ehituskivid. Olgu $M > n$ täisarv ja $\delta > 0$. Ütleme, et jada x^M on $(1-\delta)$-ehitatud kividest B_n kui leiduvad intervalli $[1, M]$ lõikumatud alamintervallid

$$[n_1, m_1], [n_2, m_2], \ldots, [n_I, m_I] \tag{160}$$

nii, et

- $\sum_{i=1}^{I}(n_i - m_i + 1) \geq (1-\delta)M$
- $x_{n_i}^{m_i} \in B_n$.

Seega x^M on (1-)ehitatud kividest B_n, kui ta on täielikult liigendatav hulga B_n elementideks. On selge, et sellisel juhul on jadas $\frac{M}{n}$ ehituskivi ning kividest B_n saab ehitada $|B_n|^{\frac{M}{n}}$ erinevat jada. Järgnev lemma annab tõkke $(1-\delta)$-ehitatud jadade arvule.

Lemma 7.4 (Lemma ehituskividest) *Olgu D_M kõikide kividest B_n $(1-\delta)$ ehitatud jadade hulk. Siis*

$$|D_M| \leq |B_m|^{\frac{M}{n}} 2^{Mh(\delta)} |\mathcal{X}|^{M\delta}.$$

Tõestus. Erinevaid võimalusi valida intervalle (160) nii, et nende kogupikkus kataks $(1-\delta)M$ osa intrervallist $[1, M]$ on tõkestatud arvuga

$$\sum_{j\leq \delta M} C_j^M \leq 2^{Mh(\delta)}.$$

Fikseeritud intervallide korral on ehituskivide valikul ülimalt

$$|B_n|^I \leq |B_n|^{\frac{M}{n}}$$

võimalust. Ning aukude katteks on ülimalt

$$|\mathcal{X}|^{M\delta}$$

võimalust. ■

7.2.6 LZ teoreemi tõestus

Lemma 7.5 *Olgu x ergoodilise protsessi X sagedustüüpiline realisatsioon. Siis iga $\epsilon > 0$ korral leidub $y \in \mathcal{X}^\infty$ nii, et $\bar{d}(x,y) < \epsilon$ ja $h(y) < H_X + \epsilon$*

Tõestus. Olgu P ergoodilise protsessi jaotus. Fikseerime $\epsilon > 0$. Olgu k nii suur, et $P(W_\epsilon^k) > 1 - \epsilon$. Olgu x protsessi sagedustüüpiline realisatsioon. Seega

$$\lim_n \frac{1}{n} \sum_{i=1}^{n} I_{W_\epsilon^k}(x_i, \ldots, x_{i+k-1}) \to P(W_\epsilon^k) > 1 - \epsilon.$$

Sama peab kehtime ka lõikumatute intervallide korral: leidub $s = 0, \ldots k-1$ nii, et

$$\liminf_n \frac{1}{n} \sum_{i=1}^{n} I_{W_\epsilon^k}(x_{s+(i-1)k}, \ldots, x_{s+ik-1}) > 1 - \epsilon.$$

Teisisõnu,

$$x = uw(1)w(2)\cdots$$

nii, et ploki u pikkus on s, $w(i)$ pikkus on k ja

$$\liminf_n \frac{|i \le n : w(i) \in W_\epsilon^k|}{n} > 1 - \epsilon. \tag{161}$$

Defineerime jada

$$y = uv(1)v(2)\cdots,$$

kus

$$v(i) = \begin{cases} w(i), & \text{kui } w(i) \in W; \\ a, & \text{mujal} \end{cases}$$

Siin $a \in \mathcal{X}$ on mingi fikseeritud täht. Tingimus (161) ja y definitsioon garanteerivad, et

$$\bar{d}(x,y) \le \epsilon.$$

Olgu M võrreldes k-ga suur. Siis hulga $\mathcal{U}_M(y)$ elemendid on kujul

$$qv(j)v(j+1)\cdots v(j+m-1)r,$$

kus q ja r pikkus on ülimalt k ja

$$\frac{(M-2k)}{k} \le m \le \frac{M}{k}.$$

Et $v(i) \in W_\epsilon^k \cup \{a\}^k$, siis $\mathcal{U}_M(y)$ on $(1 - \frac{2k}{M})$-ehitatud ja ehituskivide lemmast saame, et

$$\mathcal{U}_M(y) \le |W_\epsilon^k \cup \{a\}^k|^{\frac{M}{k}} 2^{M(h(\delta_M) + \delta_M \log |\mathcal{X}|)},$$

kus

$$\delta_M = \frac{2k}{M}.$$

Et

$$|W_\epsilon^k| \le 2^{k(H_X+\epsilon)},$$

siis

$$|W_\epsilon^k \cup \{a\}|^{\frac{M}{k}} \le 2^{M(H_X+\epsilon)},$$

millest

$$\mathcal{U}_M(y) \le 2^{M(H_X+\epsilon+h(\delta_M)+\delta_M \log|\mathcal{X}|)}$$

ja

$$h_M(y) \le H_X + \epsilon + h(\delta_M) + \delta_M \log|\mathcal{X}|,$$

kusjuures

$$\lim_M (h(\delta_M) + \delta_M \log|\mathcal{X}|) = 0.$$

Seega

$$h(y) = \lim_M h_M(y) \le H_X + \epsilon.$$

■

Järeldus 7.2 *Kui $x \in \mathcal{X}^\infty$ on ergoodilise protsessi X sagedustüüpiline realisatsioon, siis*

$$H_Z(x) \le H_X.$$

Teoreem 7.1 (Lempel-Zivi teoreem) *Kui X on ergoodiline protsess entroopiamääraga H_X, ja $\mathcal{C}_n$ on LZ kood, siis*

$$\limsup_n \frac{l(X^n)}{n} = \limsup_n \frac{K(X^n)\log n}{n} = H_X, \quad \text{p.k..}$$

Tõestus. Järeldusest 7.1 saame, et iga $x \in \mathcal{X}^\infty$ korral

$$\limsup_n \frac{K(x^n)\log n}{n} \le H_Z(x).$$

Järeldusest 7.2 saame, et kui $x \in \mathcal{X}^\infty$ on sagedustüüpiline, siis

$$\limsup_n \frac{K(x^n)\log n}{n} \le H_Z(x) \le H_X.$$

Vastavalt eeldusele on X ergoodiline. Sagedusteoreemi tõttu on sagedustüüpiliste realisatsioonide P-mõõt 1. Seega

$$\limsup_n \frac{K(x^n)\log n}{n} \le H_X, \quad P - \text{p.k.},$$

millest järeldub teoreemi väide. ■

7.3 LZ koodi asümptootiline optimaalsus

Lempel-Ziv teoreem väidab, et kui informatsiooniallikas on ergoodiline protsess, siis

$$\limsup_n \frac{1}{n} l(X^n) \le H_X, \quad \text{p.k.}$$

Selle tulemuse tähtsust aitab mõista järgnev teoreem.

Teoreem 7.2 *Olgu $\mathcal{C}_n : \mathcal{X}^n \to \{0,1\}^*$ prefikskoodide jada, $X_1, X_2, \ldots$ olgu ergoodiline protsess entroopiamääraga H_X. Siis*

$$\liminf_n \frac{1}{n} l(X^n) \ge H_X \quad p.k., \tag{162}$$

kus $l(x^n) = |\mathcal{C}_n(x^n)|$.

LZ teoreemist ja teoreemist 7.2 järeldub LZ koodi asümptootiline optimaalsus:

$$\frac{1}{n} l(X^n) \to H_X \quad \text{p.k.} \tag{163}$$

Muidugi on asümptootiliselt optimaalseid koode teisigi (näiteks?), kuid LZ kood on universaalne, olles seega **universaalne asümptootiliselt optimaalne kood**.

Teoreemi 7.2 tõestus põhineb järgmisel lemmal.

Lemma 7.6 (Barron) *Olgu $\mathcal{C}_n : \mathcal{X}^n \to \{0,1\}^*$ prefikskoodide jada, X olgu juhuslik protsess jaotusega P. Olgu α_n selline positiivsete numbrite jada, et $\sum_n 2^{-\alpha_n} < \infty$. Siis*

$$\mathbf{P}\Big(l(X^n) + \log P(X^n) \ge -\alpha_n \quad ev\Big) = 1. \tag{164}$$

Tõestus. Paneme tähele, et

$$\begin{aligned} B_n :=& \{x : l(x^n) + \log P(x^n) \le -\alpha_n\} = \{x : 2^{l(x^n)+\log P(x^n)} \le 2^{-\alpha_n}\} \\ =& \{x : 2^{l(x^n)} 2^{\log P(x^n)} \le 2^{-\alpha_n}\} = \{x : P(x^n) \le 2^{-\alpha_n} 2^{-l(x^n)}\}. \end{aligned}$$

Seega, Krafti võrratusest järeldub

$$P(B_n) = \sum_{x \in B_n} P(x^n) \le \sum_{x \in B_n} 2^{-\alpha_n} 2^{-l(x^n)} \le 2^{-\alpha_n} \sum_{x^n \in \mathcal{X}^n} 2^{-l(x^n)} \le 2^{-\alpha_n}.$$

Boreli-Cantelli I lemmast järeldub, et

$$P(\limsup_n B_n) = P\{x : x \in B_n \text{ i.o. }\} = 0$$

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 Shannon-McMillian-Breimani teoreemist

$$\liminf_n \frac{l(X^n)}{n} \geq \liminf_n \frac{-\log P(X^n) - \alpha_n}{n} = \liminf_n \frac{-\log P(X^n)}{n} = H_X, \quad \text{p.k.}$$

mis on (162).

Järeldus 7.3 *Olgu $\mathcal{C}_n : \mathcal{X}^n \to \{0,1\}^*$ üheselt dekodeeritavate koodide jada, $X_1, X_2, \ldots$ olgu ergoodiline protsess entroopiamääraga H_X. Siis kehtib (162)*

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

$$\liminf_n \frac{l(x^n) + \log l(x^n) + o(\log l(x^n))}{n} < H_X.$$

■

Märkus: Koondumisest (163) ei järeldu vahetult koondumine

$$E\frac{l(X^n)}{n} \to H_X. \tag{165}$$

Lempel-Zivi teoreemi tõestuses nägime aga, et $\frac{l(X^n)}{n}$ 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 (165). Iga prefikskoodide jada $\mathcal{C}_n : \mathcal{X}^n \to \{0,1\}^*$ korral

$$\liminf_n \frac{El(X^n)}{n} \geq H_X$$

(miks?), seega on ka koondumine (165) (keskmise mõttes) asümptootiliselt optimaalne.

8 Diferentsiaalentroopia

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.

8.1 Diferentsiaalentroopia

Olgu X pidev juhuslik suurus jaotusega P ja tihedusega f. Olgu S jaotuse P kandja (väikseim kinnine hulk, mis sisaldab hulka $\{x : f(x) > 0\}$). Olgu $0 \log 0 := 0$.

Def 8.1 *Juhusliku suuruse X (jaotuse P, tiheduse f)* **diferentsiaalentroopia** *on*

$$h(X) :=: h(P) :=: h(f) := \int -f(x) \log f(x) dx = \int_S -f(x) \log f(x) dx, \tag{166}$$

kui see integraal eksisteerib.

Märkused:

- Integraal (166) ei pruugi alati olla defineeritud. Sellisel juhul pole ka diferentsiaalentroopia defineeritud.
- Erinevalt entroopiast võib diferentsiaalentroopia olla ka negatiivne. Üldiselt võib diskreetse jaotuse 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.

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 $h(X) = 0$, $\lim_{a\to\infty} h(X) = \infty$ ja $\lim_{a\to 0} h(X) = -\infty$.

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 $-\infty$.

8.2 Pideva juhusliku suuruse kvantiseerimine

Pideva jaotuse kvantiseerimine (diskretiseerimine) on levinud võte lähendamaks pidevat jaotust "sarnase" 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 mõõtes), 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

$$\lim_{\Delta\to 0} H(P(\Delta)) + \log\Delta = h(f). \tag{167}$$

Kui näiteks $\Delta = n^{-1}$, siis suure n korral seosest (167) 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

$$H(P(\Delta)) + \log\Delta = 0 = h(f),$$

st (167) kehtib iga n korral võrdusena.

8.3 AEP ja diferentsiaalentroopia

Tuletame meelde, et kui $X_1, X_2, \ldots$ on AEP omadusega juhuslik protsess (tähestikul $\mathcal{X}$), siis iga $\epsilon > 0$ korral leidub $n(\epsilon)$ ja hulk $W_\epsilon^n \subset \mathcal{X}^n$ nii, et $P(W_\epsilon^n) > 1 - \epsilon$,

$$(1-\epsilon)2^{n(H-\epsilon)} \le |W_\epsilon^n| \le 2^{n(H+\epsilon)} \tag{168}$$

ning iga $x \in W_\epsilon^n$ korral

$$2^{-n(H+\epsilon)} \le P(x^n) \le 2^{-n(H-\epsilon)}.$$

Siin $H = H(P)$. AEP omadus kehtib ka iid pidevate jaotuste korral; hulga W_ϵ^n võimsuse asemel on seoses (168) tema ruumala ja entroopia asemel on diferentsiaalentroopia.

Def 8.2 *Mõõtuva hulga $A \subset \mathbb{R}$ ruumala on*

$$\mathbf{V}(A) := \int_A dx_1 \cdots dx_n.$$

Teoreem 8.3 *Olgu $X_1, X_2, \ldots$ iid juhuslikud suurused, X_i jaotus on pidev tihedusega f. Olgu $f \log f$ integreeruv. Siis iga $\epsilon > 0$ korral leidub $n(\epsilon)$ ja hulk $W_\epsilon^n \subset \mathbb{R}^n$ nii, et*

1

$$P(W_\epsilon^n) > 1 - \epsilon. \tag{169}$$

2

$$(1-\epsilon)2^{n(h-\epsilon)} \le \mathbf{V}(W_\epsilon^n) \le 2^{n(h+\epsilon)}. \tag{170}$$

3 *iga $x^n \in W_\epsilon^n$ korral*

$$2^{-n(h+\epsilon)} \le f(x^n) \le 2^{-n(h-\epsilon)}, \tag{171}$$

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

$$-\frac{\log f(X_1, \ldots, X_n)}{n} \to -E\Big(\log f(X_1)\Big) = h(f), \quad \text{p.k.},$$

millest järeldub (171). Hinnangutest

$$1 - \epsilon \le P(W_\epsilon^n) = \int_{W_\epsilon^n} f(x_1, \ldots, x_n) dx_1 \cdots dx_n \le 1$$

saame (170). ■

Ülesanne: Tõestada teoreem 8.3.

8.4 Ühisdiferentsiaalentroopia

Juhusliku vektori $(X_1, \ldots, X_n)$ (ühis)diferentsiaalentroopia defineeritakse analoogiliselt diskreetse vektori entroopiaga.

Def 8.4 *Olgu $X^n = (X_1, \ldots, X_n)$ pidev juhuslik vektor ühistihedusega f. Vektori X^n (ühis)diferentsiaalentroopia on*

$$h(X^n) = h(X_1, \ldots, X_n) := -\int f(x^n) \log f(x^n) dx^n = -\int f(x_1, \ldots, x_n) \log f(x_1, \ldots, x_n) dx_1 \cdots dx_n,$$

kui integraal eksisteerib.

Näide: Olgu $\phi(x^n)$ mitmemõõtmelise normaaljaotuse $N(\mu, \Sigma)$ tihedusfunktsioon,

$$\phi(x^n) = \frac{1}{(\sqrt{2\pi})^n |\Sigma|^{\frac{1}{2}}} \exp[-\frac{1}{2}(x^n - \mu)' \Sigma^{-1} (x^n - \mu)].$$

$$\begin{aligned} -\int_{-\infty}^{\infty} \phi(x^n) \ln \phi(x^n) dx^n &= \int_{-\infty}^{\infty} \frac{1}{2}(x^n - \mu)' \Sigma^{-1} (x^n - \mu) \phi(x^n) dx^n + \ln[(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}] \\ &= \frac{1}{2} E\Big((X^n - \mu)' \Sigma^{-1} (X^n - \mu)\Big) + \frac{1}{2} \ln[(2\pi)^n |\Sigma|], \end{aligned}$$

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.

Ülesanne: Tõestada ülaltoodud omadused.

8.5 Tinglik diferantsiaalentroopia, Kullback-Leibleri kaugus ja vastastikune informatsioon

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 8.5 *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

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

Lepime kokku, et $0\log\frac{0}{0} = 0$.

Def 8.6 *Olgu f, g kaks tõenäosustihedust. Nende Kullback-Leibleri kaugus on*

$$D(f||g) := \int f(x)\log\frac{f(x)}{g(x)}dx.$$

Ülesanne: Tõestada, et $D(f||g) \leq \infty$ on alati defineeritud (võib olla ∞).

Paneme tähele, et kui $D(f||g) < \infty$, siis tiheduse g kandja sisaldab f kandjat.

Kehtib Gibbsi võrratus.

Lemma 8.1

$$D(f||g) \geq 0,$$

kusjuures $D(f||g) = 0$ parajasti siis, kui $f = g$ p.k.

Tõestus. Ülesanne ■

Näide: Olgu $\{f_\theta : \theta \in \Theta\}$ tõenäosustiheduste pere. Olgu $\theta^o \in \Theta$ fikseeritud jaotus (õige jaotus). Defineerime **tõepärakontrasti**

$$\theta \mapsto \int \ln f_\theta(x) f_{\theta^*}(x) dx =: l(\theta),$$

mille saame logaritmilise tõepärafunktsiooni piirväärtusena. Gibbsi võrratusest järeldub, et θ^* maksimiseerib tõepärakontrasti, st $l(\theta^*) \geq l(\theta)$ iga $\theta \in \Theta$ korral. Sellel asjaolul põhineb STP hinnangu mõjusus.

Ülesanne: Olgu f ja g Riemanni mõttes integreeruvad tõenäosustihedused. Jagame reaaltelje intervallideks pikkusega Δ. Olgu $P(\Delta)$ ja $Q(\Delta)$ f ja g kvantiseerimisel (antud tükelduse järgi) saadud diskreetsed tõenäosusjaotused. Tõestada, et protsessis $\Delta \to 0$

$$D(P(\Delta)||Q(\Delta)) \to D(f||g).$$

Def 8.7 *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

$$|\Sigma| \leq \prod_{i=1}^{n} \sigma_i^2. \tag{172}$$

8.6 MaxEnt jaotused

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:

maksimiseerida $h(f)$ 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.

Teoreem 8.8 *Kui leiduvad konstandid $a_0, a_1, \ldots, a_k$ nii, et funktsioon*

$$f^*(x) = \exp[a_0 + \sum_{i=1}^{k} a_i F_i(x)] \tag{173}$$

rahuldab tingimusi **1,2,3**, *siis f^* on ainus 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, otsime diskreetset jaotust. Asendades ülaltoodud tõestuses integreerimise summeerimisega, saame, et teoreem kehtib ka diskreetsete jaotuste korral.

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 (173) 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 (173) 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 (173):

$$\exp[a_0 + a_1 x].$$

Eksponentjaotuse kuju; MaxEnt jaotus: $E(\mu^{-1})$.

tõkestatud kandja: Olgu $S = [a, b]$, tingimusi pole. Jaotus (173):

$$\exp[a_0].$$

Ühtlase jaotuse kuju; MaxEnt jaotus :$U(a, b)$.

lõplik kandja: Olgu $S = \{1, 2, 3, 4, 5, 6\}$, tingimusi pole. Jaotus (173): $\exp[a_0]$. 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 $EX_i X_j = \sigma_{ij}$. Jaotus (173):

$$f(x^n) = \exp[a_0 + \sum_{ij} a_{ij} x_i x_j].$$

Mitmemõõtmelise normaaljaotuse kuju; MaxEnt jaotus on $\mathcal{N}(0, \Sigma)$, kus $\Sigma = (\sigma_{ij})$.

8.7 Ülesanded

1. Tõestada Hadamardi võrratus.

2. Leida $h(f)$, kus $f(x) = \frac{1}{2}\lambda \exp[-\lambda|x|]$ (Laplace'i jaotus ehk kahepoolne eksponentjaotus).

3. 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 $I(X;Y)$.

4. 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.

5. 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

$$P^* = \arg\min_{P\in\mathcal{P}} D(P||Q).$$

6. Olgu f_o suvaline tõenöosusjaotus. Tõestada. et leidub tingimus F ja konstant c (mis sõltuvad f_o-st) nii, et f_o on MaxEnt tihedus.