

# Entroopia ja infohulk

Ahto Buldas

## Shannoni entroopia valem

Olgu  $X$  juhuslik suurus väärtuste hulgaga  $S = \{x_1, \dots, x_n\}$ . Seda võib vaadelda katsena, mille tulemusena saadakse väärtus  $x_i \in S$  tõenäosusega  $p_i = \Pr[X = x_i]$ .

Intuitiivselt: juhusliku suuruse  $X$  entroopia  $H[X]$  kui informatsiooni hulk, mida suuruse  $X$  väärtuse teadasaamine (katse sooritamine) meile annab.

Sageli antakse entroopia definitsioonina nn. *Shannoni entroopia* valem:

$$H[X] = - \sum_i p_i \cdot \log p_i , \quad (1)$$

kus summeeritakse üle indeksite  $i$ , nii et  $p_i > 0$ .

Loengu sisu: näidata, et Shannoni valem vastab entroopia intuitiivsele definitsioonile.

## Tõenäosusteooria: sündmused

$\Omega$ -valimiruum, mis sisaldab kõikvõimalikke tulemeid  $\omega \in \Omega$ . Näiteks täringu veeretamisel  $\Omega = \{1, \dots, 6\}$ .

$\mathcal{F} \subseteq 2^\Omega$ - sündmuste hulk. Meie käsitluses põhiliselt  $\mathcal{F} = 2^\Omega$ . Näiteks sündmus  $\{2, 4, 6\}$  tähendab, et visati paarisarv silmi.

Sündmused  $A$  ja  $B$  on teineteist *välistavad* kui  $A \cap B = \emptyset$ .

Tühja sündmust  $\emptyset$  nimetatakse *võimatuks* sündmuseks.

*Juhuslik suurus*  $X$  on funktsioon  $X : \Omega \rightarrow \mathbb{R}$ , mis sõltub tulemist  $\omega \in \Omega$ .

Katse on mõtteline protsess, kus "saadakse teada" juhusliku suuruse  $X$  väärtus  $x$ . Seega peale katset on teada, et  $\omega \in X^{-1}(x)$ .

NB!  $X^{-1}(x) \subseteq \Omega$  võib vaadelda ka sündmusena  $[X = x]$ .

## Tõenäosusteooria: tõenäosus

*Tõenäosus(mõõt)* on funktsioon  $\Pr: \mathcal{F} \rightarrow \mathbb{R}$  järgmiste omadustega:

- Iga sündmuse  $A \in \mathcal{F}$  korral  $0 \leq \Pr[A] \leq 1$ .
- $\Pr[\Omega] = 1$
- Kui  $\{A_i\}_{i \in \mathbb{N}}$  on üksteist välistavate sündmuste pere, siis

$$\Pr[\cup_i A_i] = \sum_i \Pr[A_i] .$$

Kolmikut  $(\Omega, \mathcal{F}, \Pr)$  nimetatakse *tõenäosusruumiks*.

Mõned järeldused omadustest:

$$\begin{aligned} \Pr[\Omega \setminus A] &= 1 - \Pr[A] , \\ \Pr[A \cap B] + \Pr[A \cup B] &= \Pr[A] + \Pr[B] . \end{aligned}$$

## Tõenäosusteooria: tinglik tõenäosus

Kui  $A$  ja  $B$  on sündmused ja  $\Pr[B] \neq 0$ , siis sündmuse  $A$  *tinglikuks tõenäosuseks* eeldades sündmust  $B$  nimetatakse suhet:

$$\Pr[A | B] = \frac{\Pr[A \cap B]}{\Pr[B]} .$$

”Mõõtmisteoreetiline” selgitus: Kui sündmus  $B$  toimub, siis me saame teada, et  $\omega \in B$ . Seega teiseneb kõikvõimalike tulemite hulk  $\Omega' = B \subset \Omega$ .

Uus (lisateabega) maailm peab samuti olema tõenäosusruum ja seega tuleb kõik tõenäosused ”renormeerida”. Sündmuse  $A$  tõenäosus uues tõenäosusruumis on  $\Pr[A | B]$ .

**Bayesi valem:**  $\Pr[A] \neq 0 \neq \Pr[B] \Rightarrow \Pr[A | B] \cdot \Pr[B] = \Pr[B | A] \cdot \Pr[A]$ .

## Tõenäosusteooria: sõltumatus

Sündmusi  $A$  ja  $B$  nimetatakse *sõltumatuteks*, kui

$$\Pr[A \cap B] = \Pr[A] \cdot \Pr[B] .$$

Kui  $\Pr[A] \neq 0 \neq \Pr[B]$ , siis sõltumatus on ekvivalentne seostega:

$$\Pr[A | B] = \Pr[A] \quad \text{ja} \quad \Pr[B | A] = \Pr[B] .$$

Juhuslikke suursi  $X$  ja  $Y$  nimetatakse sõltumatuks kui iga  $x, y \in \mathbb{R}$  korral kehtib:

$$\Pr[X = x, Y = y] = \Pr[X = x] \cdot \Pr[Y = y] .$$

## Tõenäosusteooria: keskväärtus

Olgu  $X$  juhuslik suurus võimalike väärtustega  $x_1, \dots, x_n$ .

Juhusliku suuruse  $X$  **keskväärtuseks**  $\mathbf{E}[X]$  nimetatakse arvu:

$$\mathbf{E}[X] = \sum_{i=1}^n x_i \cdot \Pr[X = x_i] .$$

Keskväärtus  $\mathbf{E}$  on lineaarne operaator:

$$\mathbf{E}[\alpha X + \beta Y] = \alpha \mathbf{E}[X] + \beta \mathbf{E}[Y] .$$

Kui  $X$  ja  $Y$  on sõltumatud, siis

$$\mathbf{E}[X \cdot Y] = \mathbf{E}[X] \cdot \mathbf{E}[Y] .$$

## Juhuslik suurus ja infohulk

Vaatleme esmalt juhtu, kus tõenäosusjaotus on ühtlane, st  $p_i = 1/n$ , iga  $i \in \{1, \dots, n\}$  korral. Kui palju infot annab meile  $X$ -i väärtuse teadasaamine?

Kujutleme mõttelist katset, kus üks katsealune  $A$  (oraakel) teab tegelikku väärtust ette ja teine katsealune  $B$  püüab seda teada saada esitades oraaklile küsimusi, kusjuures iga küsimuse vastus võib olla kas *jah* või *ei*.

Informaatikule kohaselt väljendudes, iga vastus sisaldab ühe *biti* informatsiooni. Katsealusel  $B$  on kaks ekvivalentset viisi suuruse  $X$  väärtuse teadasaamiseks. Ta võib kas teha ise katse või küsida  $X$ -i väärtuse bitt-haaval oraakli  $A$  käest.

Seega võib katses sisalduva infohulga mõõduks võtta küsimuste arvu oraaklile, mis garanteerib suuruse  $X$  väärtuse teadasaamise.

## Infohulk kui keskmine vajalik küsimuste arv

Vajalik küsimuste arv sõltub aga sellest, milline on küsimuste esitamise *strateegia* ja milline on tegelik  $X$ -i väärtus.

Näiteks võib  $B$  küsida järjest (ükshaaval) kõiki väärtusi  $x_1, \dots, x_n$ . See strateegia on edukas niipea, kui saadakse esimene *jah* vastus. Keskmine küsimuste arv on  $n/2$ , kusjuures halvimal juhul läheb vaja  $n - 1$  küsimust.

Entroopia defineerimisel on mõistlik lähtuda strateegiast, mis annab minimaalse vajaliku küsimuste arvu. Selgub, et  $n/2$  ei ole kaugeltki minimaalne (ei keskmise ega halvima juhu mõttes).

Järgnevas näitame, et Shannoni entroopia (1) on keskmise vajaliku küsimuste arvu alamtõke.

## Küsitlusstrateegiad ja koodid

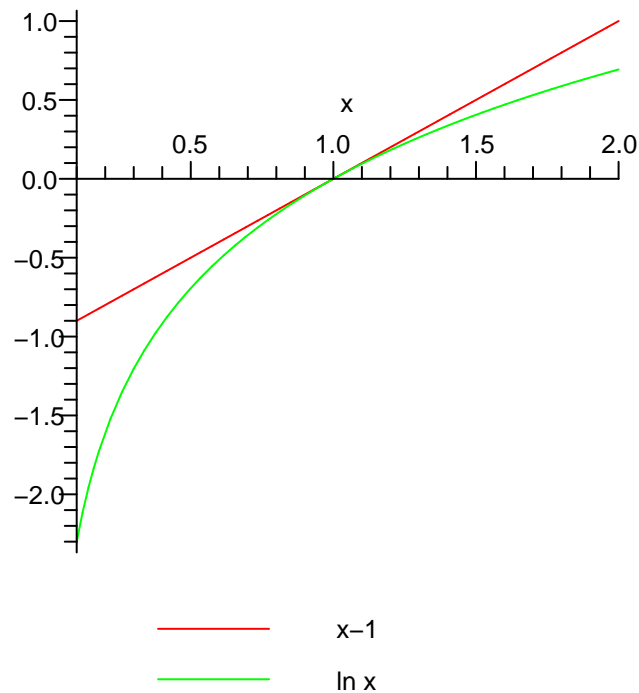
Kõigepealt defineerime küsimuste esitamise strateegia kui teatavat liiki (nn. *prefiksivaba*) koodi.

**Definitsioon 1** *Injektiivset funktsiooni  $D \xrightarrow{f} \{0, 1\}^*$  nimetatakse prefiksivabaks koodiks, kui  $f(y) \neq f(x)z_1, \dots, z_m$  mitte ühegi erinevate elementide paari  $x \neq y$ , ja elementide  $z_1, \dots, z_m \in \{0, 1\}$  korral.*

Prefiksivaba koodi sõnu (kujutisi) võib vaadelda kui jah/ei vastuste komplekte. Igale komplektile koodis vastab kindel hulga  $D$  element.

## Logaritmifunktsiooni omadus

**Lemma 1** Iga  $0 < r \in \mathbb{R}$  korral  $\ln r \leq r - 1$ , kusjuures võrdus kehtib parajasti siis kui  $r = 1$ .



## Kullback-Liebleri võrratus

**Lemma 2 (Kullback-Liebleri võrratus)** *Kui  $X$  on juhuslik suurus väärtuste hulgaga  $D$  ja  $D \xrightarrow{\pi} [0 \dots 1]$  on funktsioon, nii et  $\sum_{x \in D} \pi(x) \leq 1$ , siis*

$$s = \sum_{x \in D} \Pr[X = x] \cdot \ln \frac{\Pr[X = x]}{\pi(x)} \geq 0,$$

*kusjuures võrdus kehtib parajasti siis, kui  $\pi(x) = \Pr[X = x]$  iga  $x$  korral.*

Tõestus.

$$\begin{aligned} s &= - \sum_{x \in D} \Pr[X = x] \cdot \ln \frac{\pi(x)}{\Pr[X = x]} \geq - \sum_{x \in D} \Pr[x] \cdot \left( \frac{\pi(x)}{\Pr[x]} - 1 \right) \\ &= \underbrace{\sum_{x \in D} \Pr[x]}_{=1} - \underbrace{\sum_{x \in D} \pi(x)}_{\leq 1} \geq 0. \end{aligned}$$

Kui  $\pi(x) \neq \Pr[X = x]$  mingi  $x$  korral muutub võrratus mitterangeks.  $\square$

## Krafti võrratus

**Lemma 3 (Krafti võrratus)** Iga prefiksivaba koodi  $D \xrightarrow{f} \{0, 1\}^*$  korral kehtib võrratus:

$$s = \sum_{x \in D} 2^{-\|f(x)\|} \leq 1,$$

kus  $\|f(x)\|$  tähistab koodsõna  $f(x)$  pikkust.

Tõestus. Esitame  $s$  teisel kujul, kus  $c_n$  on kõigi  $n$ -bitiste koodsõnade arv:

$$s = \sum_{n=0}^{\infty} c_n \cdot 2^{-n}.$$

$D$  on lõplik  $\Rightarrow \exists m \geq 0: c_{m+1} = c_{m+2} = \dots = 0$ . Arutleme, kui suur võib maksimaalselt olla  $c_m$ . On selge, et ideaaljuhul võib olla  $c_m = 2^m$ , sest täpselt nii palju on kõikvõimalikke  $m$ -bitiseid koodsõnu. Koodi

prefiksivabadusest tulenevalt ei tohi aga  $m$ -bitised sõnad sisaldada algosana (prefiksina) lühemaid samasse koodi kuuluvaid sõnu. Näiteks kui kood sisaldab tühisõna, siis  $c_m = 0$ , sest tühisõna võib vaadelda mis tahes sõna algosana. Kui kood sisaldab ühebitist sõna '1', siis  $m$ -bitiste koodsõnade hulgas ei tohi esineda 1-ga algavaid sõnu, mida on täpselt  $2^{m-1}$ . Kui lisaks sõnale '1' sisaldab kood ka kahebitist sõna '01', siis ei saa  $m$ -bitiste koodsõnade hulgas olla '01'-ga algavaid sõnu, mida on täpselt  $2^{m-2}$  tükki. Seega kehtib võrratus:

$$c_m \leq 2^m - c_0 2^{m-0} - c_1 2^{m-1} - c_2 2^{m-2} - \dots - c_{m-1} 2^1,$$

mille läbi jagamisel  $2^m$ -ga ja liikmete viimisel vasakule poole, saame

$$s = c_0 2^{-0} + c_1 2^{-1} + c_2 2^{-2} + c_3 2^{-3} + \dots + c_m 2^{-m} \leq 1. \quad (2)$$

□

## Prefiksivaba kood ja Shannoni entroopia

**Teoreem 1** Kui  $X$  on juhuslik suurus võimalike väärtuste hulgaga  $D$ , siis iga prefiksivaba koodi  $D \xrightarrow{f} \{0, 1\}^*$  sõnade keskmine pikkus on ülalt tõkestatud suuruse  $X$  Shannoni entroopiaga, st

$$\mathbf{E}[\|f(X)\|] \geq H[X].$$

Tõestus. Kasutame Kullback-Liebleri ja Krafti võrratusi:

$$\begin{aligned} \mathbf{E}[\|f(X)\|] - H[X] &= \mathbf{E}[\|f(X)\| - H[X]] \\ &= \sum_{x \in D} \Pr[X = x] \cdot \left( \|f(x)\| - \log_2 \frac{1}{\Pr[X = x]} \right) \\ &= \sum_{x \in D} \Pr[X = x] \cdot \log_2 \frac{\Pr[X = x]}{2^{-\|f(x)\|}} \geq 0. \end{aligned}$$

□

## Kombinatoorne entroopia ja Shannoni Entroopia

**Definitsioon 2** *Juhusliku suuruse  $X$  kombinatoorseks entroopiaks nimetatakse suurust*

$$H_{\text{comb}}[X] = \min_f \mathbf{E}[\|f(X)\|],$$

*kus miinimum arvutatakse üle kõigi prefiksivabade koodide  $D \xrightarrow{f} \{0, 1\}^*$ .*

Näitasime juba, et  $H[X] \leq H_{\text{comb}}[X]$ . Nüüd näitame, et kombinatoorne entroopia ei erine palju Shannoni entroopiast, st  $H_{\text{comb}}[X] \leq H[X] + 1$ .

**Teoreem 2 (Shannon 1948)**  $H_{\text{comb}}[X] \leq H[X] + 1$ .

## Shannoni teoreemi tõestus

Tõestus. Olgu  $X$  juhuslik suurus väärtuste hulgaga  $D = \{x_1, \dots, x_N\}$ , kusjuures  $p_i = \Pr[X = x_i] \neq 0$  iga  $i \in \{1, \dots, N\}$ . Olgu elemendid indekseeritud nii, et  $p_1 \geq p_2 \geq \dots p_N$ . Piisab kui näitame, et leidub prefiksivaba kood  $f$ , nii et teoreemi väites olev võrratus kehtib. Vajaliku koodi konstrueerimiseks defineerime suurused

$$\begin{aligned} a_1 &= 0 \\ a_2 &= p_1 \\ a_3 &= p_1 + p_2 \\ a_4 &= p_1 + p_2 + p_3 \\ &\dots \\ a_N &= p_1 + p_2 + p_3 + \dots + p_{N-1} . \end{aligned}$$

Olgu  $m_i$  selline positiivne täisarv, nii et

$$2^{-m_i+1} > p_i \geq 2^{-m_i} . \quad (3)$$

On selge, et  $m_1 \leq m_2 \leq \dots \leq m_N$ . Defineerime  $a_i^*$  kui kahendmurr, mis saadakse arvu  $a_i$  esitusest, millest kustutatakse kõik peale  $m_i$  komakoha, st  $a_i = a_i^* + 2^{-m_i} \cdot \bar{a}_i$ , kus  $\bar{a}_i < 1$ . Defineerime koodi  $f$ , nii et  $f(x_i)$  olgu arvu  $a_i^*$  kümnenndkohtadest koosnev järjend pikkusega  $m_i$ .

Näitame, et defineeritud kood on tõepoolest prefiksivaba. Olgu  $1 \leq i < j \leq N$  ja  $f(x_j) = f(x_i) \| z$ . Et  $m_i \leq m_j$ , siis vastupidi olla ei saa, sest  $\|f(x_i)\| \leq \|f(x_j)\|$ . Siit järeldub, et

$$\begin{aligned} a_i^* + 2^{-m_i} \cdot \bar{a}_i &= a_i = p_1 + \dots + p_{i-1} \\ a_i^* + 2^{-m_i} \cdot \bar{z} &= a_j = p_1 + \dots + p_{j-1} , \end{aligned}$$

kus  $\bar{z} < 1$ . Võrratuse  $a_j > a_i$  tõttu  $0 < \bar{z} - \bar{a}_i < 1$ . Lahutades teisest võrrandist esimese, saame

$$2^{-m_i} > 2^{-m_i}(\bar{z} - \bar{a}_i) = a_j - a_i = p_i + \dots + p_{j-1} \geq 2^{-m_i} ,$$

mis on vastuolu. Järelikult on defineeritud kood prefiksivaba. Võrratustest (3) järeldeb, et  $-\log_2 p_i \leq m_i \leq -\log_2(p_i) + 1$ , millest omakorda saame hinnata koodi keskmist pikkust:

$$\mathbf{E}[\|f(X)\|] = \sum_{i=1}^N p_i \cdot m_i \leq \sum_{i=1}^N p_i \cdot (-\log_2(p_i) + 1) = H[X] + 1 .$$

□

## Optimaalsed prefiksivabad koodid

Prefiksivaba koodi  $f$  nimetatakse *optimaalseks*, kui

$$\mathbf{E}[\|f(X)\|] = H_{\text{comb}}[X] .$$

Kui  $n = 2$ , siis optimaalne kood  $f$  saadakse defineerides  $f(x_1) = 0$  ja  $f(x_2) = 1$ , mis annab keskmiseks koodi pikkuseks

$$\mathbf{E}[\|f(X)\|] = \Pr[X = x_1] \cdot 1 + \Pr[X = x_2] \cdot 1 = 1 .$$

## Optimaalsete koodide omadus I

**Lemma 4** *Kui  $f$  on optimaalne prefiksivaba kood ja  $p_i > p_j$ , siis*

$$\|f(x_i)\| \leq \|f(x_j)\| .$$

Tõestus. Kui  $\|f(x_i)\| > \|f(x_j)\|$ , siis defineerime uue koodi  $f'$ , mis on sarnane koodiga  $f$ , välja arvatud elemendid  $x_i$  ja  $x_j$ , kus  $f'(x_i) = f(x_j)$  ja  $f'(x_j) = f(x_i)$ . Siis oleks

$$\begin{aligned} p_i \cdot \|f'(x_i)\| + p_j \cdot \|f'(x_j)\| &= p_i \cdot \|f(x_j)\| + p_j \cdot \|f(x_i)\| \\ &= p_i \cdot \|f(x_i)\| + p_j \cdot \|f(x_j)\| - \\ &\quad - \underbrace{(p_i - p_j)(\|f(x_i)\| - \|f(x_j)\|)}_{>0} \\ &< p_i \cdot \|f(x_i)\| + p_j \cdot \|f(x_j)\|, \end{aligned}$$

mistõttu ka  $\mathbf{E}[f'(X)] = \sum_i p_i \cdot \|f'(x_i)\| < \sum_i p_i \cdot \|f(x_i)\| = \mathbf{E}[f(X)]$ , mis on vastuolus koodi  $f$  optimaalsusega.  $\square$

## Optimaalsete koodide omadus II

**Lemma 5** *Olgu  $f$  optimaalne prefiksivaba kood juhuslikule suurusele  $X$  väärtuste hulgaga  $D = \{x_1, \dots, x_n\}$  ja tõenäosustega  $p_1, \dots, p_n$ , kusjuures  $p_1 \geq \dots \geq p_n > 0$  ja  $\|f(x_1)\| \leq \dots \leq \|f(x_n)\|$ . Siis  $\|f(x_{n-1})\| = \|f(x_n)\|$  ja leidub  $i \in \{1, \dots, n-1\}$ , nii et koodid  $f(x_i)$  ja  $f(x_n)$  erinevad ainult viimase märgi poolest (näiteks  $f(x_n) = w\|1$  ja  $f(x_i) = w\|0$ ).*

Tõestus. Oletame, et  $\|f(x_{n-1})\| < \|f(x_n)\|$ . Siis  $f(x_n)$  on ainuke pikim kood. Olgu näiteks  $f(x_n) = w\|1$ . On selge, et kood  $w$  ei ole ühegi teise koodi  $f(x_1), \dots, f(x_{n-1})$  algosa, sest see saaks olla võimalik vaid siis, kui  $w \in \{f(x_1), \dots, f(x_{n-1})\}$ , sest  $w$  on vähemalt sama pikk kui mis tahes kood sellest hulgast. Samuti ei ole ükski koodidest  $f(x_1), \dots, f(x_{n-1})$  koodi  $w$  algosa, sest siis oleks ta ühtlasi koodi  $f(x_n)$  algosa. Seega, kui asendada väärtuse  $x_n$  kood  $f(x_n)$  koodiga  $w$ , saaksime koodi, mis oma

efektiivsuselt ületaks koodi  $f$ , mis on aga võimatu koodi  $f$  optimaalsuse tõttu. Seega on koodid  $f(x_{n-1})$  ja  $f(x_n)$  tõepoolest ühepikkused.

Kui  $f(x_n) = w||c$ , kus  $c \in \{0, 1\}$  ja iga  $i \in \{1, \dots, n - 1\}$  erineksid koodid  $f(x_i)$  ja  $f(x_n)$  rohkem kui viimase märgi poolest, siis kood  $w$  ei oleks ühegi koodi  $f(x_1), \dots, f(x_{n-1})$  algosa. Vastasel korral oleks mingi  $i \in \{1, \dots, n - 1\}$  korral  $f(x_i) = w||c'$ , kus  $c' \in \{0, 1\}$ , sest  $\|f(x_i)\| \leq \|w\| + 1 = \|f(x_n)\|$  ja  $w \neq f(x_i)$ , sest vastasel korral  $f(x_i)$  oleks koodi  $f(x_n)$  algosa. Järelikult saab elemendi  $x_n$  koodi  $f(x_n)$  asendada lühema koodiga  $w$ , mis on vastuolus koodi  $f$  optimaalsusega.  $\square$

## Optimaalsete koodide omadus III

Olgu  $x_1, \dots, x_n$  tõenäosused on vastavalt  $(p_1, \dots, p_n)$  (jaotus). Kasutame tähistust  $f: (w_1, \dots, w_n)$  kui  $w_i = f(x_i)$ .

**Teoreem 3** *Olgu  $f: (w_1, \dots, w_n)$  jaotuse  $(p_1, \dots, p_n)$  optimaalne prefiksivaba kood. Kui  $p_1 \geq \dots \geq p_n > 0$ ,  $\|w_1\| \leq \dots \leq \|w_n\|$ ,  $w_{n-1} = w|0$  ja  $w_n = w|1$ , siis  $g: (w_1, \dots, w_{n-2}, w)$  on optimaalne prefiksivaba kood jaotusele  $(p_1, \dots, p_{n-2}, p_{n-1} + p_n)$ .*

Tõestus. Kõigepealt näitame, et kood  $g$  on prefiksivaba. Ühelt poolt on selge, et ükski koodsõnadest  $w_1, \dots, w_{n-2}$  ei saa olla sõna  $w$  algosa, sest muidu ei oleks kood  $f$  ise prefiksivaba. Kui  $w$  oleks sõna  $w_i$  algosa, siis  $\|w_i\| \leq \|w\| + 1$  tõttu kas  $w_i = w|0$  või  $w_i = w|1$ , millest aga järelduks, et

$w_i \in \{w_{n-1}, w_n\}$ , mis on vastuolus koodi  $f$  prefiksivabadusega. Oletame, et  $g$  ei ole optimaalne. Siis leiduks kood  $\gamma: (v_1, \dots, v_{n-2}, v)$ , nii et

$$\overline{\ell}_\gamma = \sum_{i=1}^{n-2} p_i \cdot \|v_i\| + (p_{n-1} + p_n) \cdot \|v\| < \sum_{i=1}^{n-2} p_i \cdot \|w_i\| + (p_{n-1} + p_n) \cdot \|w\| = \overline{\ell}_g.$$

Defineerime uue koodi  $\varphi: (v_1, \dots, v_{n-2}, v \| 0, v \| 1)$  jaotusele  $(p_1, \dots, p_n)$

ja näitame, et kood  $\varphi$  on efektiivsem koodist  $f$ . Tõepoolest,

$$\begin{aligned}\bar{\ell}_\varphi &= \sum_{i=1}^{n-2} p_i \cdot \|v_i\| + (p_{n-1} + p_n)(\|v\| + 1) \\ &= \sum_{i=1}^{n-2} p_i \cdot \|v_i\| + (p_{n-1} + p_n)\|v\| + p_{n-1} + p_n \\ &< \sum_{i=1}^{n-2} p_i \cdot \|w_i\| + (p_{n-1} + p_n)\|w\| + p_{n-1} + p_n \\ &= \sum_{i=1}^{n-2} p_i \cdot \|w_i\| + (p_{n-1} + p_n)(\|w\| + 1) = \bar{\ell}_f.\end{aligned}$$

See on aga vastuolus koodi  $f$  optimaalsusega.  $\square$

## Optimaalsete koodide omadus IV

**Teoreem 4** Olgu  $g: (w_1, \dots, w_{n-2}, w)$  optimaalne prefiksivaba kood jaotusele  $(p_1, \dots, p_{n-2}, p_{n-1} + p_n)$ , kus  $p_1 \geq \dots \geq p_n > 0$ . Siis

$$f: (w_1, \dots, w_{n-2}, w||0, w||1)$$

on optimaalne prefiksivaba kood jaotusele  $(p_1, \dots, p_n)$ .

Tõestus. Näitame kõigepealt, et  $f$  on prefiksivaba. On selge, et koodid  $w||0$  ja  $w||1$  ei saa olla koodide  $w_1, \dots, w_{n-2}$  algosad, sest siis oleks ka kood  $w$  ise vastavate koodide algosa, mis aga oleks vastuolus koodi  $g$  prefiksivabadusega. Kui mingi  $i \in \{1, \dots, n-2\}$  korral oleks kood  $w_i$  koodi  $w||0$  algosa, siis järelikult  $w_i = w||0$ , sest muidu oleks  $w_i$  juba sõna  $w$  algosa. Kuid võrdus  $w_i = w||0$  tähendaks, et  $w||0$  oleks koodi

$w_i$  algosa, mille võimatust me aga just põhjendasime. Järelikult on  $f$  prefiksivaba. Kui  $f$  ei oleks optimaalne, siis leiduks optimaalne (NB!) kood  $\varphi: (v_1, \dots, v_{n-2}, v_{n-1}, v_n)$ , nii et  $\bar{\ell}_\varphi < \bar{\ell}_f$ . Lemmale 5 tuginedes võib eeldada üldisust kitsendamata, et  $v_{n-1} = v||0$  ja  $v_n = v||1$ .

Näitame, et kood  $\gamma: (v_1, \dots, v_{n-2}, v)$  jaotusele  $(p_1, \dots, p_{n-2}, p_{n-1} +$

$p_n$ ) on efektiivsem koodist  $g$ . Tõepoolest,

$$\begin{aligned}
 \bar{\ell}_\gamma &= \sum_{i=1}^{n-2} p_i \cdot \|v_i\| + (p_{n-1} + p_n) \cdot \|v\| \\
 &= \sum_{i=1}^{n-2} p_i \cdot \|v_i\| + p_{n-1} \cdot (\|v\| + 1) + p_n \cdot (\|v\| + 1) - p_{n-1} - p_n \\
 &= \bar{\ell}_\varphi - p_{n-1} - p_n \\
 &< \bar{\ell}_f - p_{n-1} - p_n \\
 &= \sum_{i=1}^{n-2} p_i \cdot \|w_i\| + p_{n-1} \cdot (\|w\| + 1) + p_n \cdot (\|w\| + 1) - p_{n-1} - p_n \\
 &= \sum_{i=1}^{n-2} p_i \cdot \|w_i\| + (p_{n-1} + p_n) \cdot \|w\| \\
 &= \bar{\ell}_g.
 \end{aligned}$$

Vastuolu koodi  $g$  optimaalsusega.  $\square$

## Optimaalse koodi leidmine - Huffmani algoritm

*Huffmani algoritmi* võib esitada järgmise kahe sammuna.

(1) Triviaalse jaotuse (1) optimaalne kood on  $(\_)$ , kus  $\_$  tähistab tühisõna.

(2) Kui  $n > 1$ , ja  $(p_1, \dots, p_n)$  on jaotus, nii et  $p_1 \geq \dots \geq p_n > 0$ , siis optimaalne kood saadakse kui:

(2a) protseduuri rekursiivselt rakendades leitakse optimaalne prefiksi-  
vaba kood  $(w_1, \dots, w_{n-2}, w)$  jaotusele  $(p_1, \dots, p_{n-2}, p_{n-1} + p_n)$ ;  
ja

(2b) moodustatakse kood  $(w_1, \dots, w_{n-2}, w||0, w||1)$ .

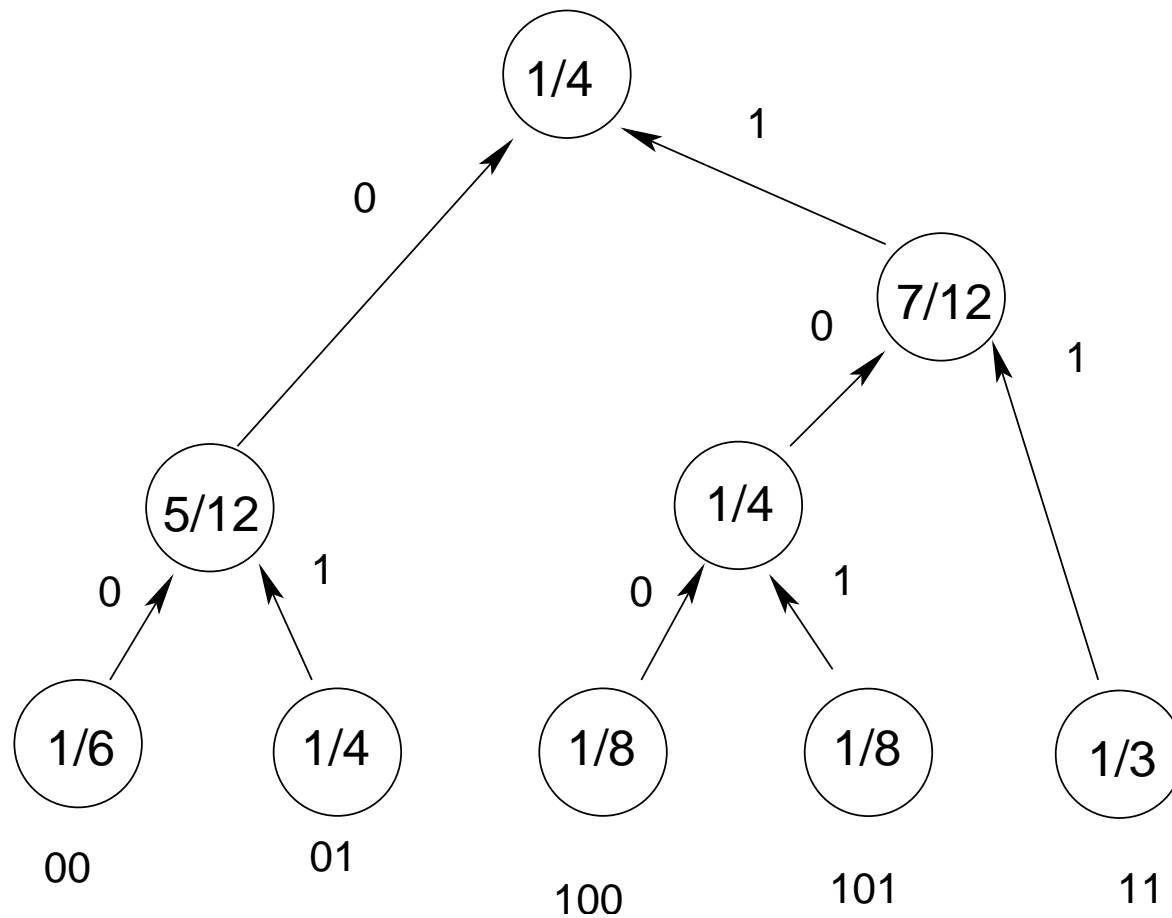
## Huffmani puu

Huffmani algoritmi võib vaadelda kui puu ehitamist “alt üles” meetodil:

- (0) Puu lehtedeks võetakse esialgsed tõenäosused  $p_1, \dots, p_n$ .
- (1) Järjestatakse tõenäosused suuruse järjekorras.
- (2) Võetakse kaks kõige väiksemat tõenäosust (st  $p_{n-1}$  ja  $p_n$ ) ja moodustatakse nendest uus tipp, mis on tõenäosustele  $p_{n-1}$  ja  $p_n$  vastavate tip-pude “ühine vanem”, ning millele vastav tõenäosus on  $p_{n-1} + p_n$ .
- (3) Alustatakse protseduuri uuesti alates sammust (0), lähtudes tõenäosustest  $p_1, \dots, p_{n-2}, p_{n-1} + p_n$ .

Protsess lõpeb, kui alles jääb üksainus tipp tõenäosusega 1.

Seejärel antakse igast tipust väljuvatele koodid 1 ja 0.



## Shannoni entroopia ülemtõke

**Teoreem 5** *Kui  $X$  on juhuslik suurus väärtuste hulgaga  $D = \{x_1, \dots, x_n\}$ , siis kehtib võrratus  $H[X] \leq \log_2 n$ , kusjuures võrdus kehtib parajasti siis, kui iga  $x \in D$  korral  $\Pr[X = x] = 1/n$ .*

Tõestus. Tõestame, et vahe  $\log_2 n - H[X] \geq 0$ . Kasutame Kullback-Liebleri võrratust. Olgu  $p_i = \Pr[X = x_i]$ .

$$\begin{aligned}\log_2 n - H[X] &= \sum_{i=1}^n p_i \cdot \log_2 n + \sum_{i=1}^n p_i \cdot \log_2 p_i \\ &= \sum_{i=1}^n p_i \cdot \log_2 (n \cdot p_i) = \sum_{i=1}^n p_i \cdot \log_2 \frac{p_i}{\frac{1}{n}} \geq 0.\end{aligned}$$

Kullback-Liebleri võrratusest tuleneb ka, et võrdus kehtib ainult siis, kui  $p_i = \frac{1}{n}$ . Eeldades, et  $p_i = \frac{1}{n}$ , saame  $H[X] = \sum_i p_i \log_2 \frac{1}{p_i} = \log_2 n$ , millest järeldub et tõestatud maksimum tõe poolest ka saavutatakse.  $\square$

## Liitentroopia

Juhuslike suuruste  $X$  ja  $Y$  paar  $(X, Y)$  on samuti juhuslik suurus. Tähistame  $H[X, Y] = H[(X, Y)]$ .

**Teoreem 6** *Olgu  $X$  ja  $Y$  kaks juhuslikku suurust. Siis  $H[X, Y] \leq H[X] + H[Y]$ , kusjuures võrdus kehtib vaid siis, kui  $X$  ja  $Y$  on sõltumatud.*

Tõestus. Olgu suuruste  $X$  ja  $Y$  väärtste hulkadega vastavalt  $D_X = \{x_1, \dots, x_n\}$  ja  $D_Y = \{y_1, \dots, y_m\}$ . Olgu  $p_i = \Pr[X = x_i]$ ,  $q_j = \Pr[Y = y_j]$  ja  $r_{ij} = \Pr[X = x_i \text{ ja } Y = y_j]$ . Kasutades võrdusi  $p_i = \sum_j r_{ij}$  ja  $q_j = \sum_i r_{ij}$  hindame suurust  $e = H[X] + H[Y] - H[X, Y]$  ja näitame, et

$$e = H[X] + H[Y] - H[X, Y] \geq 0 .$$

Tõestuskäik:

$$\begin{aligned} e &= \sum_{i=1}^n p_i \log_2 \frac{1}{p_i} + \sum_{j=1}^m q_j \log_2 \frac{1}{q_j} + \sum_{i=1}^n \sum_{j=1}^m r_{ij} \log_2 r_{ij} \\ &= \sum_{i=1}^n \sum_{j=1}^m r_{ij} \log_2 \frac{1}{p_i} + \sum_{j=1}^m \sum_{i=1}^n r_{ij} \log_2 \frac{1}{q_j} + \sum_{i=1}^n \sum_{j=1}^m r_{ij} \log_2 r_{ij} \\ &= \sum_{i=1}^n \sum_{j=1}^m r_{ij} \log_2 \frac{r_{ij}}{p_i q_j} \geq 0, \end{aligned}$$

sest  $\sum_i \sum_j p_i q_j = (\sum_i p_i) \cdot (\sum_j q_j) = 1$  ning võrratus järeldeb seetõttu otseselt Kullback-Liebleri võrratusest. Samuti järeldeb otseselt, et võrdus kehtib parasjagu siis, kui  $r_{ij} = p_i \cdot q_j$ , millest tuleneb suuruste  $X$  ja  $Y$  sõltumatus.  $\square$

## Tinglik entroopia

Olgu  $X \in D_X$  ja  $Y \in D_Y$  juhuslikud suurused. Olgu  $y \in D_Y$   $Y$  mingi fikseeritud väärtus. Defineerime juhusliku suuruse  $X | y$  väärtuste piirkonnaga  $D_X$ , kus  $x \in D_X$  tõenäosus on  $p(x | y) = \Pr[X = x | Y = y]$ . Suuruse  $X | y$  entroopia

$$H[X | y] = \sum_{x \in D_X} p(x | y) \log_2 \frac{1}{p(x | y)}$$

tähendab intuiitiivselt informatsioonihulka, mille me saame suuruse  $X$  tegeliku väärtuse teadasaamisel, eeldusel, et me teame juba, et  $Y = y$ .

Suuruse  $H[X | y]$  keskväärtust tähistame

$$\begin{aligned} H[X | Y] &= \sum_{y \in D_Y} \Pr[Y = y] \cdot H[X | y] \\ &= \sum_{y \in D_Y} \sum_{x \in D_X} p(y)p(x | y) \log_2 \frac{1}{p(x | y)} \\ &= - \sum_{x,y} p(x, y) \log_2 p(x | y). \end{aligned}$$

ja nimetame suuruse  $X$  **tinglikuks entroopiaks** suuruse  $Y$  suhtes. Intuitiivselt tähendab  $H[X | Y]$  informatsiooni hulka, mis annaks suuruse  $X$  teadasaamine, eeldusel, et suuruse  $Y$  tegelik väärtus on juba teada.

## Liitentroopia ja tinglik entroopia

**Teoreem 7**  $H[X, Y] = H[Y] + H[X | Y]$ .

Tõestus.

$$\begin{aligned} H[X, Y] &= - \sum_{x,y} p(x, y) \log_2 p(x, y) = - \sum_{x,y} p(x, y) \log_2 p(y)p(x | y) \\ &= - \sum_{x,y} p(x, y) [\log_2 p(y) + \log_2 p(x | y)] \\ &= - \sum_{x,y} p(x, y) \log_2 p(y) - \sum_{x,y} p(x, y) \log_2 p(x | y) \\ &= - \sum_y \left( \underbrace{\sum_x p(x, y)}_{=p(y)} \right) \cdot \log_2 p(y) + H[X | Y] = H[Y] + H[X | Y] . \end{aligned}$$

□

## Infohulk

Suurust  $I[X; Y] = H[X] - H[X | Y]$  nimetatakse *infohulgaks* suuruses  $Y$  suuruse  $X$  kohta.

**Teoreem 8** *Infohulk on sümmeetriline, st.  $I[X; Y] = I[Y; X]$ , ja mittenegatiivne, st  $I[X; Y] \geq 0$ , kusjuures  $I[X; Y] = 0$  parajasti siis, kui  $X$  ja  $Y$  on sõltumatud juhuslikud suurused.*

Tõestus. Võrdustest  $H[Y] + H[X | Y] = H[X, Y] = H[X] + H[Y | X]$  tuleneb, et  $I[X; Y] = I[Y; X]$ . Mittenegatiivsus tuleneb seoste ahelast:

$$\begin{aligned} I[X; Y] &= H[X] - H[X | Y] = H[X] + H[Y] - (H[Y] + H[X | Y]) \\ &= H[X] + H[Y] - H[X, Y] \geq 0, \end{aligned}$$

kusjuures võrdus kehtib parajasti siis, kui  $X$  ja  $Y$  on sõltumatud.  $\square$

## Entroopia aksiomaatika

Näitasime entroopia kombinatoorse definitsiooni (kombinatoorse entroopia) seotust Shannoni entroopiaga. Nüüd näitame, et Shannoni entroopia avaldiseni võib jõuda üldistest kaalutlustest lähtudes. Näitame, et eeldades entroopialt kui infohulga mõõdult teatud loomulikke omadusi, saame tõestada, et seljuhul peab entroopia olema arvutatav Shannoni entroopia avaldisega.

Vaatleme juhusliku suuruse  $X$  entroopiat kui funktsiooni  $H$ , mille argumentiks (sisendiks) on suuruse  $X$  võimalike väärtuste tõenäosustest moodustatud jada  $p_1, \dots, p_i, \dots$ , st iga positiivsetest reaalarvudest koosnev jada, mis rahuldab tingimust  $\sum_i p_i = 1$ . Vaatleme komplekti kaheksast omadusest, millest igaühe kohta tõestame, et Shannoni entroopia seda omadust rahuldab. Lõpuks näitame, et kui mingi funktsioon  $H$  rahuldab toodud kaheksat omadust, siis langeb ta kordaja täpsusega kokku Shannoni entroopiaga, st  $H(X) = \lambda \cdot H[X]$ .

## Entroopia põhiomadused

- **OM1:**  $H(p_1, \dots, p_n)$  on iga fikseeritud  $n$  korral maksimaalne parajasti siis, kui  $p_1 = \dots = p_n = 1/n$ .
- **OM2:** Hulga  $\{1, \dots, n\}$  iga permutatsiooni  $\sigma$  korral  $H(p_1, \dots, p_n) = H(p_{\sigma(1)}, \dots, p_{\sigma(n)})$ .
- **OM3:**  $H(p_1, \dots, p_n) \geq 0$  ja võrdus kehtib parajasti siis, kui  $p_i = 1$  mingi  $i \in \{1, \dots, n\}$  korral.
- **OM4:**  $H(p_1, \dots, p_n, 0) = H(p_1, \dots, p_n)$ .
- **OM5:**  $H(\frac{1}{n}, \dots, \frac{1}{n}) < H(\frac{1}{n+1}, \dots, \frac{1}{n+1})$ .
- **OM6:**  $H(p_1, \dots, p_n)$  on pidev funktsioon.
- **OM7:**  $H(\frac{1}{mn}, \dots, \frac{1}{mn}) = H(\frac{1}{n}, \dots, \frac{1}{n}) + H(\frac{1}{m}, \dots, \frac{1}{m})$  suvaliste positiivsete täisarvude  $m$  ja  $n$  korral.

## "Totalisaatori" omadus

- **OM8:** Olgu  $p = p_1 + \dots + p_m$  ja  $q = q_1 + \dots + q_n$ , kus  $p + q = 1$  ja nii  $p_i$  kui  $q_j$  on mittenegatiivsed reaalarvud. Siis

$$H(p_1, \dots, p_m, q_1, \dots, q_n) = H(p, q) + p \cdot H\left(\frac{p_1}{p}, \dots, \frac{p_m}{p}\right) + q \cdot H\left(\frac{q_1}{q}, \dots, \frac{q_n}{q}\right).$$

Võidujooksus osalevad  $m$  musta ja  $n$  valget hobust. Mustade hobuste võitmise tõenäosused on  $p_1, \dots, p_m$  ning valgete tõenäosused  $q_1, \dots, q_n$ .

Olgu  $X$  juhuslik suurus, mille tegelikuks väärtuseks on võitev hobune (ei ole vahet kas must või valge).

Olgu  $Y$  juhuslik suurus, millel on kaks võimalikku väärtust: võidab must ja võidab valge.

## Entroopia mõiste ühesus

**Teoreem 9** Kui funktsioon  $H$  rahuldab omadusi 1-8, siis

$$H(p_1, \dots, p_n) = \lambda \cdot H[X],$$

kus  $X$  on juhuslik suurus, mille väärtuste tõenäosused on  $p_1, \dots, p_n$ .

Tõestus. Olgu  $H$  funktsioon, millel on kõik omadused 1-8. Tähistame  $g(n) = H(\frac{1}{n}, \dots, \frac{1}{n})$ , st funktsioon  $g$  on defineeritud iga positiivse naturaalarvu  $n \in \mathbb{N}$  korral. Omadusest 7 järelduvalt  $g(n^k) = g(n) + g(n^{k-1})$ , millest järeldub seos

$$g(n^k) = k \cdot g(n), \quad (4)$$

mis kehtib kõigi positiivsete naturaalarvude  $n, k \in \mathbb{N}$  korral. Olgu nüüd  $r, s, n \in \mathbb{N}$  suvalised positiivsed naturaalarvud. On selge, et leidub  $m \in \mathbb{N}$ ,

nii et

$$r^m \leq s^n \leq r^{m+1}. \quad (5)$$

Omadusest 5 tulenevalt  $g(r^m) \leq g(s^n) \leq g(r^{m+1})$ , millest võrduse (4) põhjal saame

$$m \cdot g(r) \leq n \cdot g(n) \leq (m + 1) \cdot g(r).$$

Samal ajal, rakendades naturaalogaritmi võrratuse (5) liikmetele, saame võrratused

$$m \cdot \ln r \leq n \cdot \ln n \leq (m + 1) \cdot \ln r.$$

Teisendades neid kahte sarnast võrratuste ahelat, saame süsteemi

$$\begin{cases} \frac{m}{n} \leq \frac{g(n)}{g(r)} \leq \frac{m}{n} + \frac{1}{n} \\ \frac{m}{n} \leq \frac{\ln n}{\ln r} \leq \frac{m}{n} + \frac{1}{n}, \end{cases}$$

millest järeldub, et  $\left| \frac{g(s)}{g(r)} - \frac{\ln s}{\ln r} \right| \leq \frac{1}{n}$ , iga positiivse  $n \in \mathbb{N}$  korral. Siit järeldub, et  $\frac{g(s)}{g(r)} = \frac{\ln s}{\ln r}$  ja  $\frac{g(s)}{\ln s} = \frac{g(r)}{\ln r} = c = \text{const}$ , st iga positiivse naturaalarvu  $s$  korral  $g(s) = c \cdot \ln s = \lambda \cdot \log_2 s$ . Olgu  $p = \frac{t}{n}$  mingi positiivne ratsionaalarv, kus  $t, n \in \mathbb{Q}$ . Omandusest 8 järelduvalt:

$$g(n) = H\left(\frac{1}{n}, \dots, \frac{1}{n}\right) = H\left(\frac{t}{n}, \frac{n-t}{n}\right) + \frac{t}{n}g(t) + \frac{n-t}{n}g(n-t),$$

millest tulenevalt

$$\begin{aligned} H(p, 1-p) &= H\left(\frac{t}{n}, \frac{n-t}{n}\right) = g(n) - \frac{t}{n}g(t) - \frac{n-t}{n}g(n-t) \\ &= \lambda \log_2 n - \lambda \frac{t}{n} \log_2 t - \lambda \frac{n-t}{n} \log_2(n-t) \\ &= -\lambda \left[ -\frac{t}{n} \log_2 n - \frac{n-t}{n} \log_2 n + \frac{t}{n} \log_2 t + \frac{n-t}{n} \log_2(n-t) \right] \\ &= -\lambda \left[ \frac{t}{n} \log_2 \frac{t}{n} + \frac{n-t}{n} \log_2 \frac{n-t}{n} \right] \\ &= -\lambda p \log_2 p - \lambda(1-p) \log_2(1-p). \end{aligned}$$

See võrdus kehtib iga ratsionaalarvu  $p \in [0, 1]$  korral. Funktsiooni  $H$  pidevuse (Omadus 6) tõttu kehtib võrdus ka iga reaalarvu  $r \in [0, 1]$  korral. Tõestuseks, et  $H(p_1, \dots, p_n) = -\lambda \sum_{i=1}^n p_i \log_2 p_i$  suvaliste reaalarvude  $p_1 + \dots + p_n = 1$  korral, kasutame induktsiooni  $n$  järgi. Oleme juba tõestanud, et väide kehtib  $n = 2$  korral. Oletame, et ta kehtib  $n - 1$  korral. Defineerime  $p = p_1 + \dots + p_{n-1}$  ja  $q = p_n$ . Kasutame Omadust 8 ja

induktsiooni eeldust:

$$\begin{aligned}
 H(p_1, \dots, p_n) &= H(p, q) + p \cdot H\left(\frac{p_1}{p}, \dots, \frac{p_{n-1}}{p}\right) + q \cdot H(1) \\
 &= -\lambda p \log_2 p - \lambda q \log_2 q - \lambda p \sum_{i=1}^{n-1} \frac{p_i}{p} \log_2 \frac{p_i}{p} \\
 &= -\lambda p \log_2 p - \lambda p_n \log_2 p_n - \lambda \sum_{i=1}^{n-1} p_i (\log_2 p_i - \log_2 p) \\
 &= -\lambda p \log_2 p - \lambda p_n \log_2 p_n - \lambda \sum_{i=1}^{n-1} p_i \log_2 p_i + \lambda \log_2 p \cdot \underbrace{\sum_{i=1}^{n-1} p_i}_{=p} \\
 &= -\lambda \sum_{i=1}^n p_i \log_2 p_i = \lambda \cdot H[X],
 \end{aligned}$$

Kus  $X$  on juhuslik suurus, mille võimalike väärtuste tõenäosused on  $p_1, \dots, p_n$ .

□

## Krüptosüsteemi tõenäosuslik mudel

Avateksti  $X$ , võtit  $Z$  ja krüptogrammi  $Y$  käsitletakse juhuslike suurustena, mille jaotusi saab hinnata vastane, kellel on juurdepääs krüptogrammidele  $Y$ .

Suurused on seotud funktsionaalse seosega:

$$Y = E_Z(X) ,$$

kus  $X \in \mathbf{X}$ ,  $Y \in \mathbf{Y}$  ja  $Z \in \mathbf{Z}$ .

Olgu  $p(x) = \Pr_X[X = x]$ .

Eeldame, et  $X$  ja  $Z$  on *sõltumatud* juhuslikud suurused.

## Väljundjaotus kui funktsioon sisendjaotusest

Kõigepealt anname valemi tingimusliku tõenäosuse  $p(y | x) = \Pr[Y = y | X = x]$  arvutamiseks. Tähistame:

$$\mathbf{Z}(x, y) = \{z \in \mathbf{Z} : E_z(x) = y\} ,$$

st  $\mathbf{Z}(x, y) \subseteq \mathbf{Z}$  on kõigi selliste võtmete hulk, mille abil avatekst  $x$  krüpteeritakse avatekstiks  $y$ . Tõenäosus  $p(x, y)$  avaldub seljuhul nii:

$$p(y | x) = \Pr_Z[Z \in \mathbf{Z}(x, y)] = \sum_{z \in \mathbf{Z}(x, y)} p(z). \quad (6)$$

Tõenäosus  $p(y) = \Pr[Y = y]$  on arvutatav täistõenäosuse valemi järgi:

$$p(y) = \sum_{x \in \mathbf{X}} p(y | x) \cdot p(x) = \sum_{x \in \mathbf{X}} \sum_{z \in \mathbf{Z}(x, y)} p(z) \cdot p(x). \quad (7)$$

Kasutades Bayesi valemit, saab arvutada ka duaalse tingimusliku tõenäosuse, mis iseloomustab (vastase) teavet avateksti  $x$  kohta, eeldusel, et krüptogramm  $y$  on teada:

$$p(x | y) = \frac{p(x) \cdot p(y | x)}{p(y)}. \quad (8)$$

## Täieliku salastuse definitsioon

Loomulik on defineerida krüptosüsteemi turvalisus tingimusena, et krüptogramm  $Y$  (ja selle statistilised omadused) ei anna mingisugust informatsiooni avateksti kohta, st  $I(Y; X) = 0$ . Kasutades seost  $I(Y; X) = H[X] - H[X | Y]$ , saab sama tingimuse avaldada entroopia kaudu järgmiselt:

$$H[X | Y] = H[X], \quad (9)$$

mis, nagu eelnevalt tõestatud, on samaväärne tingimusega, et  $X$  ja  $Y$  on sõltumatud juhuslikud suurused. Seega, kasutades juhuslike suuruste sõltumatuse definitsiooni ja Bayesi valemit (8), saame et tingimus (9) on samaväärne mõlemaga järgmistest tingimustest

$$\begin{aligned} \forall x \in \mathbf{X}, \forall y \in \mathbf{Y}: \quad & p(x) = p(x | y), \\ \forall x \in \mathbf{X}, \forall y \in \mathbf{Y}: \quad & p(y) = p(y | x). \end{aligned}$$

## Nihkešiffer on täielikult salastav

**Teoreem 10** Nihkešiffer  $y = E_z(x) = x + z \pmod{p}$  on täielikult salastav.

Tõestus. Näitame, et  $p(y) = p(y | x)$ . Et  $|\mathbf{Z}(x, y)| = 1$ , siis iga  $x, y$  korral on võrrandil  $x + z \equiv y \pmod{p}$  täpselt üks lahend  $z$ . Valemist (7):

$$\begin{aligned} p(y) &= \sum_{x \in \mathbf{X}} \sum_{z \in \mathbf{Z}(x, y)} p(z) \cdot p(x) = \frac{1}{p} \sum_{x \in \mathbf{X}} \sum_{z \in \mathbf{Z}(x, y)} p(x) \\ &= \frac{1}{p} \sum_{x \in \mathbf{X}} |\mathbf{Z}(x, y)| p(x) = \frac{1}{p} \sum_{x \in \mathbf{X}} p(x) = \frac{1}{p}. \end{aligned}$$

Teiselt poolt, vastavalt valemile (6),

$$p(y | x) = \sum_{z \in \mathbf{Z}(x, y)} p(z) = \sum_{z \in \mathbf{Z}(x, y)} \frac{1}{p} = \frac{|\mathbf{Z}(x, y)|}{p} = \frac{1}{p},$$

millest järeldub suuruste  $X$  ja  $Y$  sõltumatus ja nihkešifri turvalisus.  $\square$

## Täieliku salastuse “hind”

Nagu nägime, leidub šifreid, mis tagavad täieliku salastuse, st on turvalised selles mõttes, et krüptogramm ei sisalda mingit informatsiooni avateksti kohta, eeldusel, et võti  $Z$  ei ole teada.

Järgnevast lihtsast arutelust selgub, et täieliku turvalisuse saavutamise hind on väga kõrge: kasutatav võti  $Z$  peab olema sama mahukas kui edastatav sõnum  $X$ .

Tuletame meelde, et võtit saab kasutada vaid üheainsa sõnumi krüpteerimiseks, mistõttu võib ka öelda, et võti peab olema sama mahukas kui kõik edastatavad sõnumid kokku.

Põhjenduses kasutatakse entroopia üldisi omadusi.

## Krüptosüsteemi omadused

- *Krüptogrammi taastatavus* – kasutaja, kellel on võti  $Z$ , suudab üheselt taastada krüptogrammile  $Y$  vastava avateksti  $X$ . Ehk: krüptogramm ja võti sisaldavad piisavalt informatsiooni avateksti taastamiseks:

$$H[X | Y, Z] = 0.$$

- *Täielik salastus* – krüptogramm  $Y$  üksi ei sisalda mingit informatsiooni avateksti  $X$  kohta.

$$H[X | Y] = H[X].$$

Neist eeldustest lähtuvalt saame, et

$$H[X] = H[X | Y] \leq H[X, Z | Y] = H[Z] + \underbrace{H[X | Y, Z]}_0 = H[Z] .$$

Seega võtme infosisaldus on vähemalt sama suur kui krüptogrammi infosisaldus, mistõttu on võtme kodeerimiseks vaja vähemalt umbes sama arv bitte kui krüptogrammi kodeerimiseks.

## Võtme korduvkasutus ja selle turvalisus

Teame, et täielikult turvalise šifri saamiseks peab võti olema sama pikk kui avatekst.

Samas, ei järeldu veel otseselt, et võtme korduvkasutus tekitab praktikas olulise turvalisuse kao.

Näiteks kui ühte võtit kasutada kümme korda, siis kui palju infot võtmest sellega vastasele lekitatakse?

Järgnevas näitame, et kui edastatavad sõnumid  $X$  on loomuliku keele tekstid, siis juba paarikümne tähelise sõnumi krüptogramm sisaldab piisava hulga informatsiooni võtme (ja seega ka avateksti) üheseks tuvastamiseks.

## Teoreem võtme tinglikust entroopiast

**Teoreem 11**  $H[Z | Y] = H[Z] + H[X] - H[Y]$ .

Tõestus. Definitsiooni järgi  $H[Z, X, Y] = H[Y | Z, X] + H[Z, X] = H[Z, X]$ , sest  $H[Y | Z, X] = 0$  (kuna  $Y$  on funktsioon  $(Z, X)$ -paarist). Eeldatavasti on  $X$  ja  $Z$  sõltumatud suurused, mistõttu  $H[Z, X] = H[Z] + H[X]$ . Sarnaselt eelnevale arutelule ja eeldusele avateksti ühesest taastatavusest krüptogrammi ja võtme abil ( $H[X | Z, Y] = 0$ ) saame, et  $H[Z, X, Y] = H[Z, Y]$ , mistõttu:

$$\begin{aligned} H[Z | Y] &= H[Z, Y] - H[Y] \\ &= H[Z, X, Y] - H[Y] \\ &= H[Z, X] - H[Y] \\ &= H[Z] + H[X] - H[Y], \end{aligned}$$

mida oligi vaja näidata.  $\square$

## Võtme korduvkasutus ja "valevõtmed"

Oletame, et edastatav sõnum koosneb  $n$  blokidest  $X_1X_2 \dots X_n$ , mis krüpteeritakse blokkideks  $Y_1Y_2 \dots Y_n$ , nii et

$$Y_i = E_Z(X_i),$$

st kõigi blokkide krüpteerimiseks kasutatakse ühte ja sama võtit  $Z$ .

Kui ründaja teab, et  $X_1X_2 \dots X_n$  on loomuliku keele tekst  $X_1, \dots, X_n$ , siis võib ta läbi proovida kõik võtmed  $Z \in \mathcal{Z}$ , mis krüptogrammi  $Y_1Y_2 \dots Y_n$  dešifreerimisel annavad loomuliku keele teksti.

Sobilike kandidaatide hulkas on ka tegelik võti  $Z$ . Ülejäänud kandidaate nimetatakse *valevõtmeteks*.

Intuitsioon: mida vähem on  $n$ -kombinatsioonide seas loomuliku keele sõnu, seda vähem võtmekandidaate tekib ja seda edukam on kirjeldatud rünne.

## Loomuliku keele entroopia

- $\Lambda$  – juhuslik suurus, mille väärtusteks on loomuliku keele tähed tõenäosustega, millega nad esinevad loomuliku keele tekstides.
- $\Lambda^n$  – juhuslik suurus, mille väärtusteks on  $n$ -tähelised loomuliku keele tekstilõigud tekstides esinemise tõenäosusega.

**Definitsioon 3** *Loomuliku keele entroopiaks nimetatakse suurust*

$$H_\Lambda = \lim_{n \rightarrow \infty} \frac{H[\Lambda^n]}{n},$$

*ja liiasuseks suurust*

$$R_\Lambda = 1 - \frac{H_\Lambda}{\log_2 |\mathbf{X}|} = \frac{\log_2 |\mathbf{X}| - H_\Lambda}{\log_2 |\mathbf{X}|}.$$

Inglise keele entroopia:  $1.0 \leq H_\Lambda \leq 1.5$  ja keskmine liiasus  $R_\Lambda \approx 0.75$ .

## Lekkepiir (*unicity distance*)

Olgu  $Y^n$  sisendjaotusele  $\Lambda^n$  vastav väljundjaotus.

Kui  $n$  on piisavalt suur, siis on võti üheselt määratud ja mingi  $n = n_0$  korral  $H[Z | Y^{n_0}] = 0$ . Arvu  $n_0$  nimetatakse *lekkepiiriks*. Teoreemist 11:

$$H[Z] + H[\Lambda^{n_0}] - H[Y^{n_0}] \approx 0,$$

Eeldades, et  $n_0$  on piisavalt suur, saame kasutada lähendit

$$H[\Lambda^{n_0}] \approx n_0 \cdot H_\Lambda = n_0(1 - R_\Lambda) \log_2 |\mathbf{X}| .$$

Eeldades, et  $H[Y^{n_0}] \approx n_0 \log_2 |\mathbf{X}|$ , saame:

$$\begin{aligned} H[Z] + n_0 \cdot H_\Lambda - n_0 \cdot \log_2 |\mathbf{X}| &\approx 0 \\ H[Z] + n_0(1 - R_\Lambda) \log_2 |\mathbf{X}| - n_0 \cdot \log_2 |\mathbf{X}| &\approx 0 \\ H[Z] - n_0 \cdot R_\Lambda \cdot \log_2 |\mathbf{X}| &\approx 0 . \end{aligned}$$

Eeldades, et võti  $Z$  on valitud juhuslikult ja ühtlaselt,  $H[Z] \approx \log_2 |Z|$  ja

$$n_0 \approx \frac{\log_2 |Z|}{R_{\wedge} \log_2 |X|} .$$

Näiteks asendusšifri korral on  $|X| = 26$  ja  $|Z| = 26!$ . Võttes  $R_{\wedge} = 0.75$  saame, et  $n_0 \approx 25$ . See on üsna täpselt kooskõlas praktikaga, et 20 – 30 täheline krüptogramm on suure tõenäosusega üheselt dešifreeritav.

Kirjeldatud ründe läbiviimiseks piisab avateksti liiasusest, mis eristab korrektseid avatekste mittekorrektsetest tekstidest ja võimaldab seega vastasel kõiki võtmeid läbi vaadates selgitada välja võtmekandidaatide hulk, mis väheneb iga kord kui ründaja saab teada uusi krüptogramme. See rünne ei sõltu kasutatavast krüptosüsteemist ja õnnestub niipea, kui avateksti jaotus erineb ühtlasest jaotusest (mis peaaegu alati ongi nii) ja kui võtme entroopia on väiksem avateksti entroopiast.

