TEORIJA INFORMACIJA I KODIRANJE

Marko D. Petković

July 7, 2018

Sadržaj

Glava 1

Uvod

1.1 Podatak i informacija

Podaci su registrovane činjenice, oznake ili zapažanja u toku nekog procesa. Podatke prikupljamo i registrujemo da bismo ih mogli čuvati i po potrebi koristiti. Ako se registrovani podatak koristi za preduzimanje akcija i donošenje odluka, onda se on smatra **informacijom**. Reč informacija potiče od latinske reči *in formare* i izvorno je značila stavljanje u određenu formu, odnosno, davanje oblika nečemu.

Primer 1.1.1. Kada smo kod kuće i čujemo obaveštenje da je spoljna temperatura $-10°$C, to je podatak. Međutim, ako se spremamo da izađemo iz kuće i na osnovu toga odlučimo kako da se obučemo, onda se taj podatak može smatrati informacijom.

Ako smo negde u Sibiru, ova informacija je gotovo beznačajna, pošto je tamo uobičajena temperatura od $-10°$C. Situacija je potpuno obrnuta ako je u pitanju neko toplije mesto gde su tako niske temperature prava retkost.

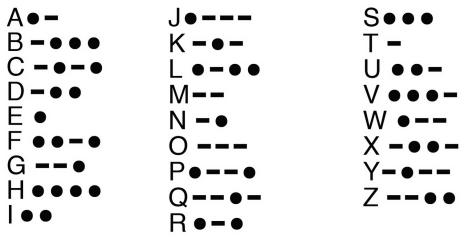
Pod obradom podataka podrazumeva se skup aktivnosti kojima se podaci pretvaraju u informacije, dok je informatika sinonim za automatsku obradu podataka (od francuskih reči *information* i *automatique*).

1.2 Telegrafija i osnovni problemi teorije informacija

Potreba za prenosom određenih informacija sa jednog mesta na drugo stara je koliko i samo čovečanstvo. Prve komunikacije širih razmera nastale su

pronalaskom telegrafa i radio-prenosa. Tada se prvi put javljaju i osnovne ideje teorije informacija.

Samuel B. Morse je 1838. smislio kod za prenos teksta putem telegrafa koji se povremeno i danas koristi. Svaki karakter u ovom kodu predstavljen je nizom dugih i kratkih impulsa. Ovi impulsi su na slici 1.1 redom predstavljeni kao crta i tačka. Ideja je bila da se karakteri koji se češće koriste [1], predstave kraćim kodnim rečima.



Slika 1.1: Morzeov kod. Tačka predstavlja kraći a crta duži signal.

Zbog prenosnih karakteristika tadašnjih (prenosnih) sistema, impulsni signal sa predajne strane postaje manje ili više "razvučen" na prijemnoj strani. Ako je vreme između emitovanja dva signala suviše kratko, oni mogu da se preklope na prijemu, pa samim tim nije moguće razlikovati ih. Dakle postoji maksimalna brzina kojom je moguće slati poruke putem takvog sistema.

Na primeru Morseovog koda ilustrovali smo dva osnovna problema koji nastaju pri prenosu informacija:

- Koji je maksimalni stepen kompresije podataka?
- Koja je maksimalna brzina prenosa podataka kroz određeni medijum?

Odgovor na oba pitanja (tj. teorijske granice) daje teorija informacija. To su **entropija** H i **kapacitet kanala** C. Ove veličine je definisano **Claude E. Shannon**[2] u svom radu:

- C.E. Shannon, *A Mathematical Theory of Communication*, Bell Syst. Tech. J. 27 (1948), 379–423, 623–656.

Pored toga, Shannon je kako se rezultati Boolea mogu primeniti u projektovanju i analizi digitalnih kola sastavljenih od elektromagnetnih releja.

[1] Statistika je rađena za engleski jezik, pošto je kod inicijalno konstruisan za taj jezik.

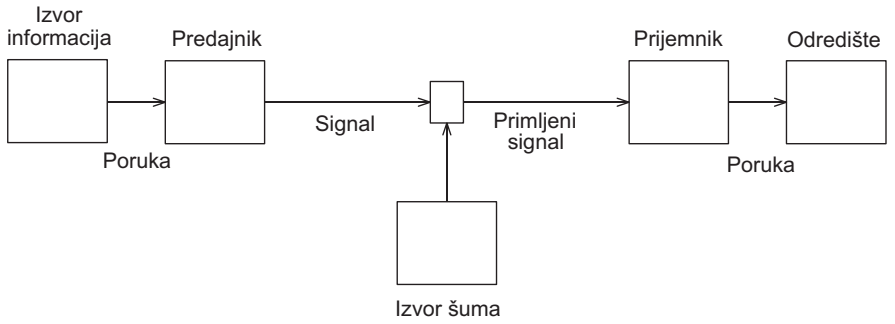
[2] Claude Elwood Shannon (1916-2001), američki naučnik i inženjer. Završio je elektrotehniku i matematiku.

Rezultate do kojih je došao, objavio je 1937. godine u svojoj magistarskoj tezi *A Symbolic Analysis of Relay and Switching Circuits*. Ovi rezultati predstavljuju osnovu projektovanja digitalnih računara i logičkih kola. Tri godine kasnije doktorirao je na Massachusetts Institute of Technology. Autor je i prvog kompjuterskog programa za igranje šaha.

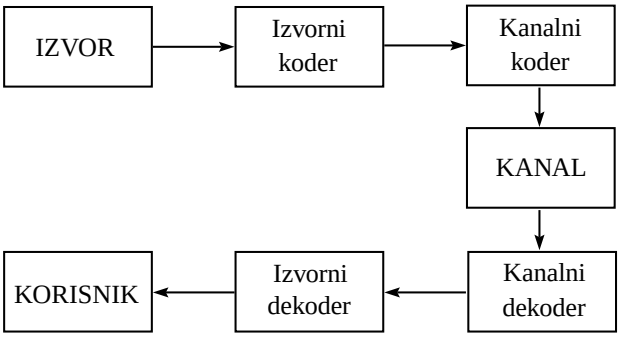
Teorija informacija daje odgovore na neka važna pitanja iz oblasti statističke fizike (termodinamike), računarskih nauka (algoritamska odnosno Kolmogorovljeva složnost), i mnogih drugih.

1.3 Model komunikacionog sistema

Prema Shannonu, opšti model komunikacionog sistema dat je na slici 1.2 dok je detaljnija blok-šema data na slici 1.3.



Slika 1.2: Osnovna blok-šema Shannonovog modela komunikacionog sistema.



Slika 1.3: Detaljnija blok-šema Shannonovog modela komunikacionog sistema.

U nastavku dajemo detaljniji opis blokova sa slike 1.3:

- **Izvor informacija** generiše niz simbola $a_1, a_2, \ldots, a_n, \ldots$ pri čemu je $a_n \in \mathcal{A}$, gde je $\mathcal{A}$ konačan skup mogućih simbola.
- **Koder izvora (statistički koder)** kodira svaki od simbola a_n konačnim nizom simbola skupa $\mathcal{B}$, koji mogu da se šalju kroz kanal, a da pritom srednja dužina generisanog niza bude minimalna moguća.
- **Koder kanala (zaštitni koder)** dodaje poruci izvesnu redundansu, čime će se na prijemu omogućiti detektovanje i ispravljanje eventualnih grešaka, nastalih tokom prenosa informacije kroz komunikacioni kanal.
- **Kanal** je medijum za prenos podataka (bakarna žica, optički kabl, vazduh, vakuum, magnetni medijum itd.).
- **Dekoderi** vrše komplementarne operacije, kako bi se na kraju korisniku predale informacije koje je uputio izvor.

Primer 1.3.1. Potrebno je da informaciju o vremenskoj prognozi u toplom primorskom mestu prenesemo zainteresovanom turisti. Izvor informacija je meteorološka stanica i ona svakog dana (ili sata) emituje podatak o vremenu. Radi jednostavnosti, pretpostavićemo da su mogući vremenski uslovi

$$\mathcal{A} = \{\texttt{Sunce}, \texttt{Oblaci}, \texttt{Kiša}, \texttt{Sneg}\}.$$

Pretpostavimo da kanalom možemo da šaljemo samo binarne podatke, tj. $\mathcal{B} = \{0, 1\}$.

Ukoliko simbole iz $\mathcal{A}$ redom kodiramo sa 00, 01, 10 i 11, tada je srednja dužina poruke (kodne reči) jednaka $l_1 = 2$. Međutim, imajući u vidu da su verovatnoće pojavljivanja svakog od simbola redom

$$0.8,\ 0.1,\ 0.09,\ 0.01,$$

bolje je da kodiranje vršimo na sledeći način: 0, 10, 110, 111 (simbol koji se najčešće javlja, kodiramo najkraćom kodnom reči). Ovim kodiranjem, srednja dužina kodne reči jednaka

$$l_2 = 0.8 \cdot 1 + 0.1 \cdot 2 + 0.09 \cdot 3 + 0.01 \cdot 3 = 1.3 < 2 = l_1.$$

Ukoliko je potrebno da prenesemo 100 simbola, očekivana dužina poruke biće 200 bita za prvo i 130 bita za drugo kodiranje. Prema tome, drugo kodiranje je bolje koristiti

Da bi signal učinili otpornijim na greške, na svaka četiri bita b_1, b_2, b_3 i b_4 prenosićemo i 3 tzv. *parity check* bita:

$$p_1 = b_1 \oplus b_2 \oplus b_4$$
$$p_2 = b_1 \oplus b_3 \oplus b_4$$
$$p_3 = b_2 \oplus b_3 \oplus b_4$$

Samim tim, celokupan ulazni signal delimo na grupe od po 4 bita $b_1b_2b_3b_4$, a šaljemo grupe $p_1p_2b_1p_3b_2b_3b_4$ od po 7 bita. Može se pokazati, da ukoliko je **tačno jedan bit** u prethodnoj sedmorki pogrešan, tada vrednost $\overline{(s_3s_2s_1)}_2$ (broj u binarnom sistemu čije su cifre s_1, s_2 i s_3), gde je

$$s_1 = p_1 \oplus b_1 \oplus b_2 \oplus b_4$$
$$s_2 = p_2 \oplus b_1 \oplus b_3 \oplus b_4$$
$$s_3 = p_3 \oplus b_2 \oplus b_3 \oplus b_4$$

označava indeks bita (s leva na desno, počev od 1) koji je pogrešan i koji treba promeniti.

Radi praktične demostracije ovog tvrđenja (koje ćemo u jednoj od poslednjih sekcija i dokazati), pretpostavimo da želimo da prenesemo bitove $b_1b_2b_3b_4 = 0000$. Tada je i $p_1 = p_2 = p_3 = 0$, odnosno prenosimo $p_1p_2b_1p_3b_2b_3b_4 = 0000000$. Pretpostavimo da je u trećem bitu s leve strane došlo do greške, odnosno da je na prijemu stiglo $p_1p_2b_1p_3b_2b_3b_4 = 0010000$. Tada je $s_1 = 1$, $s_2 = 1$ i $s_3 = 0$, pa je $\overline{(s_3s_2s_1)}_2 = \overline{011}_2 = 3$. Pretpostavimo da je četvrtom bitu došlo do greške, odnosno da je na prijemu stiglo $p_1p_2b_1p_3b_2b_3b_4 = 0001000$. Tada je $s_1 = 0$, $s_2 = 0$ i $s_3 = 1$, pa je $\overline{(s_3s_2s_1)}_2 = \overline{100}_2 = 4$.

Dakle, unošenjem redundanse od 3 bita obezbedili smo da je kod otporan na jednu grešku. Ovaj kod je poznat kao Hammingov $(7, 4)$ kod.

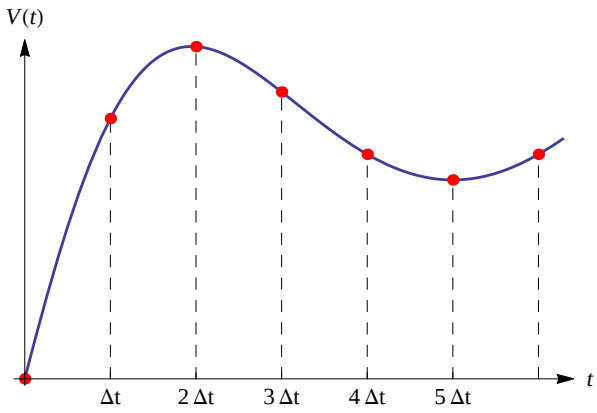
Važno je napomenuti da je neretko potrebno prenositi signal koji je po prirodi kontinualan i opisan funkcijom $V(t)$ za $t \in [0, t_0]$. U tom slučaju, postupa se na sledeći način:

1. Najpre se vrši **uzorkovanje** (semplovanje) signala, pri čemu se dobija niz $V_n = V(n\Delta t)$, gde je Δt fiksirana vrednost.

2. Definiše se konačan broj nivoa $y_0, y_1, \ldots, y_{N-1}$ i za svako V_n odredi se

$$a_n = \underset{k=0,1,\ldots,N-1}{\operatorname{argmin}} |y_k - V_n|.$$

Dakle, odredi se nivo y_k koji je najpribližniji vrednosti V_n. Na ovaj način dobijamo diskretni signal $a_0, a_1, a_2, \ldots, a_n, \ldots$.



Slika 1.4: Postupak uzorkovanja (semplovanja).

Prethodno opisani postupak je zapravo jedna vrsta analogno-digitalne konverzije (A/D konverzije). Važno je napomenuti da, ukoliko je $\Delta t < 1/(2f_{max})$ gde je f_{max} maksimalna frekvencija u spektru signala $V(t)$, tada je moguće na osnovu niza vrednosti $V_0, V_1, \ldots, V_n, \ldots$ u potpunosti rekonstruisati signal $V(t)$. Dakle, tada je operacija u koraku **1.** reverzibilna.

Na primer, ukoliko je u pitanju zvučni signal, tada je $f_{max} = 22$ kHz pa je dovoljno uzeti $\Delta t = 1/f_s$ gde je $f_s = 2f_{max} = 44$ kHz (*frekvencija semplovanja*). To je slučaj u svim muzičkim formatima.

1.4 Dva neobična primera

Primer 1.4.1. Igra pogađanja broja. Igrač 1 zamisli određeni broj od 0 do 99 a igrač 2 treba da pogodi o kom broju se radi postavljajući pitanja oblika:

"Da li zamišljeni broj pripada skupu S?", gde je $S \subset \{0, 1, \ldots, 99\}$.

Potrebno je odrediti:

1. Minimalan (garantovan) broj pitanja koje igrač 1 mora da postavi da bi bio siguran da zna o kom broju se radi.

2. Srednji broj pitanja koje postavlja drugi igrač ukoliko igra "optimalno" i ukoliko su svi izbori prvog igrača podjednako verovatni.

1. Pretpostavimo da je minimalan broj pitanja jednak k. Ukoliko svakom broju pridružimo niz sa elementima 0 i 1, pri čemu 0 označava negativan a 1

pozitivan odgovor na postavljeno pitanje, onda je jasno da različitim brojevima x moraju odgovarati različiti binarni nizovi $x_1x_2\dots x_k$. Pošto je $x_i \in \{0,1\}$, sledi da ovakvih nizova ima bar 2^k, pa samim tim mora da važi $2^k \geq 100$. Minimalni broj k za koji to važi je $k = \lceil \log_2 100 \rceil = 7$. Prema tome, minimalan ganratnovan broj pitanja je $k = 7$.

2. Sada se prirodno postavlja sledeće pitanje. Da li je UVEK potrebno $k = 7$ pokušaja da bi se otkrilo o kom broju se radi? Naravno, ukoliko postavimo pitanje "Da li je zamišljeni broj u skupu $S = \{99\}$" a pritom je igrač 1 zamislio baš $x = 99$, odgovor smo dobili u samo jednom koraku. Ali ako nije u pitanju broj 99, onda smo vrlo "jeftino" istrošili jedan pokušaj, pošto u obzir može doći bilo koji broj od 1 do 98.

Pošto je podjednako verovatno da zamišljeni broj bude bilo koji broj od 0 do 99, možemo primenti sledeću strategiju. Prvo pitanje je:

Da li je zamišljeni broj u skupu $S = \{50, 51, \dots, 99\}$?

Ako je odgovor "NE" znamo da je u pitanju jedan od brojeva $0, 1, \dots, 49$, a u suprotnom je to jedan od brojeva $50, 51, \dots, 99$. Time smo broj mogućnosti prepolovili. Bez umanjenja opštosti, možemo i u drugom slučaju smatrati da pogađamo jedan od brojeva $0, 1, \dots, 49$ (zamolićemo igrača 1 da od svog zamišljenog broja oduzme 50). Ukoliko nastavimo na ovaj način da postavljamo pitanja, broj mogućnosti biće prepolovljen u svakom sledećem potezu. Označimo sa $T(n)$ srednji broj pitanja koje je potrebno postaviti na ovaj način da bi se pogodio jedan od brojeva $0, 1, \dots, n-1$. Ukoliko je $n = 2k$, onda pitanje postavljamo za $S = \{k, k+1, \dots, 2k-1\}$, i tada je

$$T(2k) = T(k) + 1$$

Ako je $n = 2k+1$, onda pitanje postavljamo za $S = \{k+1, k+2, \dots, 2k\}$. Tada u prvom slučaju ("NE") pogađamo jedan od brojeva $0, 1, \dots, k$, a u drugom jedan od brojeva $k+1, k+2, \dots, 2k$ (odnosno $0, 1, \dots, k-1$, ako oduzmemo $k+1$ od zamišljenog broja). Srednji broj pokušaja je sad jednak

$$T(2k+1) = \frac{k+1}{2k+1}T(k+1) + \frac{k}{2k+1}T(k) + 1.$$

Naravno, ukoliko je $n = 1$, onda nemamo šta da pogađamo, pa je $T(n) = 0$. Na ovaj način, dobili smo sledeće rekurentne veze za $T(n)$:

$$T(2k) = T(k) + 1$$
$$T(2k+1) = \frac{k+1}{2k+1}T(k+1) + \frac{k}{2k+1}T(k) + 1$$

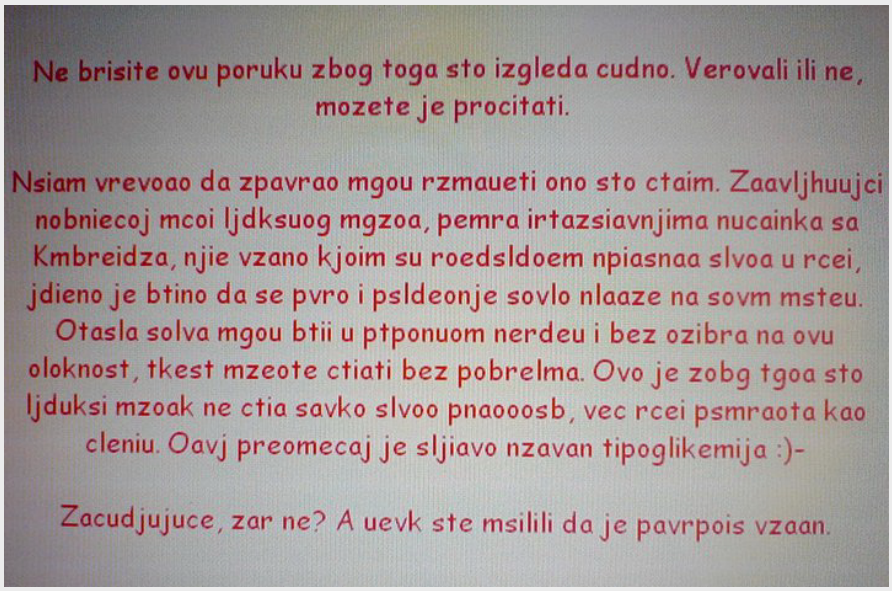
uz startnu vrednost $T(1) = 0$. Lako se dobija da je $T(100) = 168/25 = 6.72$.

Izračunajmo sada srednji broj pokušaja za strategiju opisanu u delu **1.** Ukoliko je $x < 64 + 32 = 96$, tada je jasno da je potrebno tačno 7 pokušaja (za svaku binarnu cifru po jedan). Ako utvrdimo da je $x \geq 96$ (tj. da su prve dve cifre $x_1 = x_2 = 1$), onda nam preostaje još 4 mogućnosti za koje nam je potrebno još 2 pokušaja. Znači, za $x \geq 96$ potrebno nam je 4 pokušaja.

Prema tome, srednji broj pokušaja za strategiju pod **1.** je $(4 \cdot 4 + 96 \cdot 7)/100 = 172/25 = 6.88$.

Vidimo da ukoliko primenjujemo drugu strategiju, ostvarujemo u srednjem određenu uštedu u broju pokušaja odnosno u **količini informacija kojom jednoznačno opisujemo zamišljeni broj**. Kasnije ćemo, korišćenjem tehnika Teorije informacija odnosno izvornog kodiranja, pokazati da je druga strategija ujedno i optimalna, odnosno da srednji broj pitanja ne može biti manji od $T(100) = 6.72$.

Primer 1.4.2. Posmatrajmo sledeći primer neobično napisanog teksta.



Zaključak je da redosled slova u reči (osim prvog i poslednjeg) vrlo malu količinu informacije.

Glava 2

Osnovi teorije verovatnoće

2.1 Algebra događaja i prostor verovatnoća

Definicija 2.1.1. *Neka je dat skup* Ω *i familija* $\mathbb{F}$ *podskupova skupa* Ω. *Familiju* $\mathbb{F}$ *nazivamo* **algebrom događaja** (*ili* σ**-algebrom**) *ako važi:*

$(a1)$ $\Omega \in \mathbb{F}$,

$(a2)$ *Ako je* $A \in \mathbb{F}$ *onda je i* $A^c = \Omega \setminus A \in \mathbb{F}$,

$(a3)$ *Ako je* $A_n \in \mathbb{F}$ *za svako* $n \in N$, *onda je i* $\bigcup_{n\in\mathbb{N}} A_n \in \mathbb{F}$.

U tom slučaju, skup Ω *nazivamo* **skupom elementarnih događaja** *a sve elemente familije* $\mathbb{F}$ *nazivamo* **događajima**.

Iz svojstva $(a2)$ je očigledno i $\emptyset \in \mathbb{F}$ koji nazivamo **nemoguć događaj**, dok je Ω **siguran događaj**. Uređeni par $(\Omega, \mathbb{F})$ se još naziva **prostor elementarnih događaja**. Trivijalno se dokazuje sledeće svojstvo:

$$A_n \in \mathbb{F} \quad \Rightarrow \quad \bigcap_{n\in\mathbb{N}} A_n \in \mathbb{F}.$$

Ukoliko je skup Ω konačan (što će najčešće biti pretpostavka), onda se uslov $(a3)$ zamenjuje sa

$$A, B \in \mathbb{F} \quad \Rightarrow \quad A \cup B \in \mathbb{F}.$$

Umesto $A \cap B$ pisaćemo kraće AB.

Definicija 2.1.2. *Neka je* $(\Omega, \mathbb{F})$ *prostor elementarnih događaja i funkcija* $P : \mathbb{F} \to [0, 1]$ *takva da važi:*

$(p1)$ $P(\emptyset) = 0$ $(P(\Omega) = 1)$,

($p2$) *Ako su* $A_n \in \mathbb{F}$ $(n \in \mathbb{N})$ *disjunktni događaji, onda je*

$$P\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \sum_{n=1}^{+\infty} P(A_n).$$

Tada se funkcija P *naziva* **verovatnoća** *a trojka* $(\Omega, \mathbb{F}, P)$ **prostor verovatnoća**.

I ovde se u slučaju konačnog skupa Ω, uslov ($p2$) zamenjuje uslovom

$$A, B \in \mathbb{F}, A \cap B = \emptyset \quad \Rightarrow \quad P(A) + P(B) = P(A \cup B).$$

Sledeća osnovna svojstva verovatnoće se lako dokazuju

Lema 2.1.1.

1. *Ako su* $A, B \in \mathbb{F}$ *i* $A \subset B$*, onda je* $P(A) \leq P(B)$.
2. $P(A^c) = 1 - P(A)$.
3. $P(A \cup B) = P(A) + P(B) - P(AB)$. *Iz ovoga sledi i* $P(A \cup B) \leq P(A) + P(B)$.

Primer 2.1.1 (SJ. p.11, zad 2)**.** Igrač baca tri kocke i dobija ako zbir bude veći od 10. Naći verovatnoću dobitka.

Primer 2.1.2 (SJ. p.11, zad 2)**.** Kolika je verovatnoća da će, u kup takmičenju sa 2^n ekipa, druga po vrednosti ekipa zauzeti drugo mesto? Smatramo da bolja ekipa uvek pobeđuje lošiju. Naći i graničnu vrednost kad $n \to +\infty$.

2.2 Uslovna verovatnoća

Definicija 2.2.1. (Bayes, 1763) *Neka su* $A, B \in \mathbb{F}$ *takvi da je* $P(B) > 0$. *Veličina*

$$P(A|B) = \frac{P(AB)}{P(B)}$$

naziva se **uslovna verovatnoća** *događaja* A *pod uslovom* B.

Uslovnu verovatnoću možemo posmatrati kao funkciju $P(\,\cdot\,|B) : \mathbb{F} \to [0, 1]$ i kao takva ona zadovoljava svojstva ($p1$) – ($p2$). Samim tim, $P(\,\cdot\,|B)$ predstavlja verovatnoću na $(\Omega, \mathbb{F})$.

Lema 2.2.1. *Za događaje* $A_1, A_2, \ldots, A_n \in \mathbb{F}$ *važi*

$$P(A_1A_2\cdots A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1A_2)\cdots P(A_n|A_1A_2\cdots A_{n-1}).$$

Teorema 2.2.2. (Formula potpune verovatnoće i Bayesova formula)[1] *Ako su* $A, B_1, \ldots, B_n \in \mathbb{F}$ *takvi da je* $B_1 + B_2 + \ldots + B_n = \Omega$ *i da su* B_i $(i = 1, 2, \ldots, n)$ *disjunktni po parovima, onda je*

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \ldots + P(A|B_n)P(B_n)$$

$$P(B_k|A) = \frac{P(A|B_k)P(B_k)}{P(A)}.$$

Definicija 2.2.2. *Događaji* A *i* B *su* **nezavisni** *ako je* $P(AB) = P(A)P(B)$. *Uopšteno, događaji* $A_1, A_2, \ldots, A_n$ *su* **nezavisni** *ako je*

$$P(A_{i_1}A_{i_2}\cdots A_{i_k}) = P(A_{i_1})P(A_{i_2})\cdots P(A_{i_k})$$

za svako $1 \le i_1 < i_2 < \cdots < i_k \le n$ *i* $1 \le k \le n$.

Primer 2.2.1.

2.3 Slučajne promenljive

Pojam slučajne promenljive jedan je od osnovnih pojmova u teoriji verovatnoće. Jednodimenzionalna slučajna promenljiva je funkcija koja svakom mogućem ishodu $\omega \in \Omega$ pridružuje neku njegovu numeričku karakteristiku, tj. broj $X(\omega)$.

Definicija 2.3.1. *Funkcija* $X : \Omega \to \mathbb{R}$ *za koju važi*

$$X^{-1}((-\infty, x)) = \{\omega \mid X(\omega) < x\} \in \mathbb{F}$$

naziva se **slučajna promenljiva** *na prostoru verovatnoća* $(\Omega, \mathbb{F}, P)$. *Funkcija* $F_X : \mathbb{R} \to [0, 1]$ *definisana sa* $F_X(x) = P(\{\omega \mid X(\omega) < x\})$ *naziva se* **funkcija raspodele** *slučajne promenljive* X.

Slučajna promenljiva X indukuje prostor verovatnoće $(\mathbb{R}, \mathcal{B}, P_X)$ gde je $P_X : \mathcal{B} \to [0, 1]$ funkcija (verovatnoća) definisana sa

$$P_X(B) = P(\{X \in B\}) = P(X^{-1}(B)).$$

Elementi algebre događaja $\mathcal{B}$ nazivaju se **Borelovi skupovi**.

Verovatnoće događaja $\{\omega \mid X(\omega)[=, \ge, \le]x\}$ ili $\{\omega \mid X(\omega) \in S\}$ obeležavaćemo kraće sa $P(X = x)$ i $P(X \in S)$. Sličnu notaciju ćemo koristiti i za uslovne verovatnoće.

[1] Ovu formulu je prvi koristio Laplace 1812. godine.

2.3.1 Diskretne slučajne promenljive

Ukoliko je skup vrednosti slučajne promenljive X konačan ili prebrojiv, u pitanju je **diskretna slučajna promenljiva** (**slučajna promenljiva diskretnog tipa**). Ako je $X(\Omega) = \mathcal{X} = \{x_1, x_2, \ldots\}$ onda se najčešće koristi sledeća oznaka

$$X : \begin{pmatrix} x_1 & x_2 & \cdots \\ p_1 & p_2 & \cdots \end{pmatrix}, \quad p_k = P(X = x_k) = P(\{\omega \mid X(\omega) = x_k\}).$$

Funkcija raspodele diskretne slučajne promenljive je

$$F_X(x) = \sum_{x_k < x} p_k.$$

Verovatnoće p_k obeležavaćemo i sa $p_X(x_k)$. Drugim rečima, definišimo funkciju $p : \mathbb{R} \to [0, 1]$ takvu da je

$$p_X(x) = \begin{cases} p_k = p(X = x_k), & x = x_k \\ 0, & \text{u suprotnom} \end{cases}.$$

Funkcija p u potpunosti odreuje diskretnu sluajnu promenljivu. U nastavku emo pisati $p(x)$ umesto $p_X(x)$ kad god je jasno o kojoj se sluajnoj promenljivoj radi.

Definicija 2.3.2. Matematičko očekivanje $\mathbb{E}X$ *diskretne slučajne promenljive* X *definisano je sa*

$$\mathbb{E}X = \sum_k x_k p_X(x_k).$$

Disperzija $\mathbb{D}X$ *slučajne promenljive* X *definisana je sa*

$$\mathbb{D}X = \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \sum_k x_k^2 p_X(x_k) - \left(\sum_k x_k p_X(x_k) \right)^2.$$

Teorema 2.3.1. (Osobine matematičkog očekivanja)

($e1$) $X \geq 0 \quad \Rightarrow \quad \mathbb{E}X \geq 0$,

($e2$) $\mathbb{E}1 = 1$,

($e3$) $\mathbb{E}(cX) = c\mathbb{E}X$,

($e4$) $\mathbb{E}(X + Y) = \mathbb{E}X + \mathbb{E}Y$,

($e5$) $\mathbb{E}(XY) = \mathbb{E}X \cdot \mathbb{E}Y$ *ukoliko su* X *i* Y *nezavisne slučajne promenljive.*

Teorema 2.3.2. (Osobine disperzije)

($d1$) $\mathbb{D}X = 0 \Leftrightarrow X = c$

($d2$) $\mathbb{D}(cX) = c^2\mathbb{D}X$,

($d3$) $\mathbb{D}(X+Y) = \mathbb{D}X + \mathbb{D}Y$ *ukoliko su* X *i* Y *nezavisne slučajne promenljive.*

Poznatije raspodele za slučajne promenljive diskretnog tipa:

Uniformna raspodela

Neka je $n \in N$. Tada je $X : \mathcal{U}(x_1, x_2, \ldots, x_n)$ ako važi

$$p_X(x_k) = \frac{1}{n}, \quad k = 1, 2, \ldots, n.$$

Tada važi i:

$$\mathbb{E}X = \frac{n+1}{2}, \quad \mathbb{D}X = \frac{n^2-1}{12}.$$

Bernoullijeva raspodela

Neka je $A \in \mathbb{F}$, $p = P(A)$ i slučajna promenljiva I_A definisana sa

$$I_A(\omega) = \begin{cases} 1, & \omega \in \mathcal{A} \\ 0, & \omega \notin A \end{cases}.$$

Raspodela slučajne promenljive je

$$I_A : \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}.$$

i naziva se **Bernoullijeva raspodela**. Pritom važi

$$\mathbb{E}I_A = p, \quad \mathbb{D}I_A = p(1-p).$$

Binomna raspodela

Neka je $n \in \mathbb{N}$, $p \in [0, 1]$ i $q = 1 - p$. Tada je $S_n : \mathcal{B}(n, p)$ ako važi:

$$p_{S_n}(k) = \binom{n}{k} p^k q^{n-k}, \quad k = 0, 1, \ldots, n.$$

Tada važi i:

$$\mathbb{E}S_n = np, \qquad \mathbb{D}S_n = npq.$$

Primer slučajne promenljive koja ima binomnu raspodelu je

$$S_n = I_{A_1} + I_{A_2} + \ldots + I_{A_n}$$

gde su $A_1, A_2, \ldots, A_n$ nezavisni događaji takvi da je $P(A_k) = p$ za svako $k = 1, 2, \ldots, n$.

Poissonova raspodela

Neka je $\lambda > 0$ proizvoljan realan broj. Slučajna promenljiva S ima **Poissonovu raspodelu** ako je

$$p_S(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

Ovu raspodelu obeležavamo sa $\mathcal{P}(\lambda)$. Može se pokazati da su očekivanje i disperzija slučajne promenljive S jednaki:

$$\mathbb{E}S = \mathbb{D}S = \lambda.$$

Važi i sledeće tvrđenje.

Teorema 2.3.3. *Ako je $S_n \in \mathcal{B}(n, p_n)$, pri čemu $np_n \to \lambda$ kad $n \to +\infty$, onda je*

$$p_{S_n}(k) = P(S_n = k) \to \frac{\lambda^k}{k!} e^{-\lambda}.$$

2.3.2 Apsolutno neprekidne slučajne promenljive

Slučajna promenljiva X je **apsolutno neprekidna** (**apsolutno neprekidnog tipa**) ukoliko postoji funkcija $p : \mathbb{R} \to \mathbb{R}$ takva da je $p(x) \geq 0$, $\int_{-\infty}^{+\infty} p(x)dx = 1$ i

$$P(X \in [a, b)) = \int_a^b p(x)dx.$$

Funkcija p je **gustina verovatnoće** slučajne promenljive X. Ovu funkciju (p_X) ne treba mešati sa "istoimenom" funkcijom za diskretne slučajne promenljive, pošto imaju potpuno drugačiju ulogu.

Raspodela i funkcija raspodele apsolutno neprekidne slučajne promenljive X dati su sa

$$P_X(B) = \int_B p(x)dx, \quad F(x) = \int_{-\infty}^{x} p(u)du.$$

Matematičko očekivanje i disperzija promenljive X jednaki su

$$\mathbb{E}X = \int_{-\infty}^{+\infty} x p_X(x) dx,$$
$$\mathbb{D}X = \int_{-\infty}^{+\infty} (x - \mathbb{E}X)^2 p_X(x) dx = \mathbb{E}X^2 - (\mathbb{E}X)^2.$$

Matematičko očekivanje i disperzija imaju sva svojstva kao i kod slučajne promenljive diskretnog tipa.

Poznatije raspodele za slučajne promenljive apsolutno neprekidnog tipa:

Uniformna raspodela na intervalu

Neka su $a, b \in \mathbb{R}$ i $a < b$. Tada je $X : \mathcal{U}(a, b)$ ako važi:

$$p_X(x) = \begin{cases} \frac{1}{b-a}, & x \in [a, b) \\ 0, & x \notin [a, b) \end{cases}, \quad F_X(x) = \begin{cases} 0, & x \le a \\ \frac{x-a}{b-a}, & a < x \le b \\ 1, & x > b \end{cases}.$$

Tada važi i

$$\mathbb{E}X = \frac{a+b}{2}, \quad \mathbb{D}X = \frac{(b-a)^2}{12}.$$

Normalna (Gaussova) raspodela

Neka su $m, \sigma \in \mathbb{R}$ i $\sigma > 0$. Tada je $X : \mathcal{N}(m, \sigma)$ ako važi

$$p_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}.$$

Tada je

$$\mathbb{E}X = m, \quad \mathbb{D}X = \sigma^2.$$

Standardna normalna raspodela je $X^* = \frac{X-m}{\sigma} : \mathcal{N}(0, 1)$. Funkcija raspodele promenljive X^* je

$$F_{X^*}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{u^2}{2}} du = \frac{1}{2} \left(1 + \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right)\right), \quad \operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_{-\infty}^{x} e^{-u^2} du.$$

Sa $\operatorname{erf}(x)$ smo označili **funkciju greške** (eng. *error function*). Funkcija raspodele promenljive X data je sa

$$F_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(u-m)^2}{2\sigma^2}} du = \frac{1}{2} \left(1 + \operatorname{erf}\left(\frac{x-m}{\sqrt{2}\sigma}\right)\right).$$

Eksponencijalna raspodela

Neka je $\lambda > 0$ proizvoljan realan broj. Slučajna promenljiva X ima eksponencijalnu raspodelu (u oznaci $X : \mathcal{E}(\lambda)$ ako je

$$p_X(x) = \begin{cases} 0, & x \leq 0 \\ \lambda e^{-\lambda}, & x > 0 \end{cases} \quad \Rightarrow \quad F_X(x) = \begin{cases} 0, & x \leq 0 \\ 1 - e^{-\lambda}, & x > 0 \end{cases}.$$

Tada je

$$\mathbb{E}X = \frac{1}{\lambda}, \quad \mathbb{D}X = \frac{1}{\lambda^2}.$$

2.4 Višedimenzione slučajne promenljive

Kao što jednodimenzionalna slučajna promenljiva predstavlja neku numeričku karakteristiku slučajnog ishoda, tako možemo istovremeno posmatrati n $(n \geq 2)$ numeričkih karakteristika svakog ishoda ω, tj. funkciju koja Ω preslikava u n-dimenzionalni prostor $\mathbb{R}^n$.

Definicija 2.4.1. *Funkcija $X : \Omega \to \mathbb{R}^n$, $X = (X_1, X_2, \ldots, X_n)$ takva da je*

$$\{\omega \mid X_1(\omega) < x_1, X_2(\omega) < x_2, \ldots, X_n(\omega) < x_n\} \in \mathbb{F}$$

*za svako $x_1, x_2, \ldots, x_n \in \mathbb{R}$, naziva se n-***dimenzionalna slučajna promenljiva**.

Funkcija raspodele $F_X(x_1, x_2, \ldots, x_n)$ definiše se na sličan način kao i kod jednodimenzione slučajne promenljive:

$$F_X(x_1, x_2, \ldots, x_n) = P(\{\omega \mid X_1(\omega) < x_1, X_2(\omega) < x_2, \ldots, X_n(\omega) < x_n\}).$$

Nije teško dokazati da su $X_1, X_2, \ldots, X_n$ slučajne promenljive. Raspodela k-te koordinate X_k naziva se **marginalna raspodela**. I ovde razlikujemo diskretne i apsolutno neprekidne slučajne promenljive.

Ukoliko je $X = (X_1, X_2, \ldots, X_n)$ pisaćemo $F_{X_1, X_2, \ldots, X_n}$ umesto $F_{(X_1, X_2, \ldots, X_n)}$.

2.4.1 Diskretne višedimenzione slučajne promenljive

Višedimenziona slučajna promeljiva X je **diskretna** ako je skup vrednosti $X(\Omega) = \{x^{(1)}, x^{(2)}, \ldots\}$ konačan ili prebrojiv. Njihove raspodele su potpuno određene vrednostima

$$p_k = P(X = x^{(k)}) = P(\{\omega \mid X(\omega) = x^{(k)}\}).$$

Raspodela i funkcija raspodele ove slučajne promenljive jednake su

$$P_X(B) = \sum_{x^{(k)} \in B} p_k, \quad F_X(x) = F_X(x_1, x_2, \ldots, x_n) = \sum_{x^{(k)} < x} p_k.$$

I ovde možemo definisati funkciju p_X na sličan način kao u jednodimenzionom slučaju:

$$p_X(x) = \begin{cases} p_k, & x = x^{(k)} \\ 0, & \text{u suprotnom} \end{cases}. \tag{2.1}$$

Primer 2.4.1. Neka je (X, Y) dvodimenziona slučajna promenljiva, pri čemu prva koordinata (X) može da uzima vrednosti $x_1, x_2, \ldots, x_m$ a druga $y_1, y_2, \ldots, y_n$. Tada je raspodela promenljive (X, Y) potpuno određena vrednostima

$$p_{X,Y}(x_i, y_j) = P(X = x_i, Y = y_j), \quad i = 1, 2, \ldots, m, \quad j = 1, 2, \ldots, n.$$

Marginalne raspodele (diskretnih) slučajnih promenljivih X i Y date su sa

$$p_X(x_i) = \sum_{j=1}^{n} p_{X,Y}(x_i, y_j), \quad j = 1, 2, \ldots, n,$$
$$p_Y(y_j) = \sum_{i=1}^{m} p_{X,Y}(x_i, y_j), \quad i = 1, 2, \ldots, m.$$

Polinomna raspodela

Neka je skup vrednosti slučajne promenljive $X : \Omega \to \mathbb{R}^r$ jednak

$$\{(k_1, k_2, \ldots, k_r) \mid k_1 + k_2 + \ldots + k_r = n, \quad k_i \in \mathbb{N}\ (i = 1, 2, \ldots, r)\}$$

i neka važi

$$P_{X_1, X_2, \ldots, X_r}(k_1, k_2, \ldots, k_r) = \frac{n!}{k_1! k_2! \cdots k_r!} p_1^{k_1} p_2^{k_2} \cdots p_r^{k_r}.$$

gde su $p_i > 0$ za $i = 1, 2, \ldots, r$ i $p_1 + p_2 + \ldots + p_r = 1$. Tada kažemo da slučajna promenljiva X ima polinomnu raspodelu $\mathcal{B}(n, p_1, p_2, \ldots, p_r)$.

2.4.2 Apsolutno neprekidne višedimenzione slučajne promenljive

Slično kao u jednodimenzionom slučaju, funkcija raspodele apsolutno neprekidne (višedimenzione) slučajne promenljive ima oblik

$$F_X(x_1, x_2, \ldots, x_n) = \int_{-\infty}^{x_1} du_1 \int_{-\infty}^{x_2} du_2 \cdots \int_{-\infty}^{x_n} du_n p_X(u_1, u_2, \ldots, u_n),$$

gde je $p_X : \mathbb{R}^n \to \mathbb{R}_0^+$ **gustina raspodele**. Dajemo formule za marginalne raspodele u slučaju $n = 2$, koje se lako uopštavaju za proizvoljni slučaj

$$p_{X_1}(x_1) = \int_{-\infty}^{+\infty} p_{X_1,X_2}(x_1, x_2)dx_2, \quad p_{X_2}(x_2) = \int_{-\infty}^{+\infty} p_{X_1,X_2}(x_1, x_2)dx_1.$$

Višedimenzionalna normalna raspodela

U slučaju $n = 2$, gustina raspodele data je sledećim izrazom

$$p_X(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho}} e^{-\frac{1}{2(1-\rho)}\left[\frac{(x_1-m_1)^2}{\sigma_1^2} - 2\rho\frac{(x_1-m_1)(x_2-m_2)}{\sigma_1\sigma_2} + \frac{(x_2-m_2)^2}{\sigma_2^2}\right]}$$

gde su $\sigma_1, \sigma_2 > 0$, $m_1, m_2 \in \mathbb{R}$ i $\rho \in (-1, 1)$. U opštem slučaju je

$$p_X(x) = \frac{1}{\sqrt{2\pi \det C}} e^{-\frac{1}{2}(x-m)^T C(x-m)}$$

gde je $C \in \mathbb{R}^{n\times n}$ simetrična pozitivno-definitna matrica a $m \in \mathbb{R}^{n\times 1}$ vektor srednjih (očekivanih) vrednosti koordinata.

2.4.3 Nezavisnost slučajnih promenljivih

Definicija 2.4.2. *Slučajne promenljive $X_1, X_2, \ldots, X_n$ su* **nezavisne** *ako važi*

$$P(X_1 \in B_1, \ldots, X_n \in B_n) = P(X_1 \in B_1) \cdots P(X_n \in B_n)$$

za proizvoljne skupove $B_1, B_2, \ldots, B_n \in \mathcal{B}$.

Teorema 2.4.1. *Slučajne promenljive $X_1, X_2, \ldots, X_n$ su* **nezavisne** *akko važi*

$$F_X(x_1, x_2, \ldots, x_n) = F_{X_1}(x_1)F_{X_2}(x_2) \cdots F_{X_n}(x_n).$$

Teorema 2.4.2. *Diskretne (apsolutno neprekidne) slučajne promenljive $X_1, X_2, \ldots, X_n$ su nezavisne akko važi*

$$p_X(x_1, x_2, \ldots, x_n) = p_{X_1}(x_1)p_{X_2}(x_2) \cdots p_{X_n}(x_n).^2$$

[2] U diskretnom slučaju, funkcija p_X je definisana izrazom (2.1) a u apsolutno neprekidnom je to gustina raspodele.

2.4.4 Uslovne raspodele slučajnih promenljivih

Definicija 2.4.3. *Ako su $X_1, X_2, \ldots, X_n$ diskretne slučajne promenljive, onda je* **uslovna raspodela** *$p(x_1, \ldots, x_r | x_{r+1}, \ldots, x_n)$, gde su $x_i \in \mathcal{X}_i$ $(i = 1, 2, \ldots, n)$ data sa*

$$\begin{aligned} &p_{X_1,\ldots,X_r|X_{r+1},\ldots,X_n}(x_1, \ldots, x_r | x_{r+1}, \ldots, x_n) \\ &= \frac{p_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n)}{p_{X_{r+1},X_{r+2},\ldots,X_n}(x_{r+1}, x_{r+2}, \ldots, x_n)} \\ &= \frac{P(X_i = x_i \mid i = 1, 2, \ldots, n)}{P(X_i = x_i \mid i = r+1, r+2, \ldots, n)}. \end{aligned}$$

pod uslovom $p_{X_{r+1},X_{r+2},\ldots,X_n}(x_{r+1}, x_{r+2}, \ldots, x_n) \neq 0$. U slučaju dve promenljive X i Y svodi se na

$$p_{Y|X}(y|x) = \frac{p_{X,Y}(x, y)}{p_X(x)} = \frac{P(X = x, Y = y)}{P(X = x)}.$$

pod uslovom $p_X(x) \neq 0$.

Uslovna gustina raspodele $p_{Y|X}(y|x)$ (i analogno višedimensionalni slučaj) definiše se na sličan način i kada su X i Y apsolutno neprekidne slučajne promenljive. Pritom, najpre se uvodi:

$$F_{Y|X}(y|x) = \lim_{\Delta x \to 0} \frac{P(X \in [x, x + \Delta x), Y < y)}{P(X \in [x, x + \Delta x))} = \frac{\int_{-\infty}^{y} p_{X,Y}(x, t)dt}{p_X(x)}$$

pa je onda

$$p_{Y|X}(y|x) = \frac{d}{dy} F_{Y|X}(y|x) = \frac{p_{X,Y}(x, y)}{p_X(x)}.$$

Vidimo da se i ovog puta dobija identičan izraz kao i u diskretnom slučaju, iako su odgovarajuće veličine suštinski različite.

Glava 3

Entropija, uslovna entropija i uzajamna informacija

3.1 Definicija entropije

Neka je dat diskretan izvor informacija (slučajna promenljiva) X koji emituje jednu od poruka iz skupa $\mathcal{X} = \{x_1, x_2, \dots, \}$ redom sa verovatnoćama $p_1, p_2, \dots$. Umesto $p_X(x_i) = p_i$, pisaćemo samo $p(x_i)$. Ukoliko nije drugačije naglašeno, smatraćemo da su sve slučajne promenljive koje razmatramo u ovoj glavi **diskretne**.

Pretpostavimo da je korisnik zainteresovan za poruku $x \in \mathcal{X}$, čija je verovatnoća $p(x)$. Što je ova verovatnoća manja, to je veća neizvesnost u kojoj se nalazi korisnik čekajući da pristigne x. Što je neizvesnost veća, možemo smatrati da odgovarajuća poruka nosi veću količinu informacija [1].

Označimo sa $\mathcal{I}(x)$ **količinu informacija** (**količinu sopstvene informacije**) koju nosi poruka x. Funkcija $\mathcal{I}$ mora da zadovolji sledeće zahteve:

1. $\mathcal{I}(x)$ je neprekidna i opadajuća funkcija verovatnoće $p(x)$.
2. $\mathcal{I}((x, y)) = \mathcal{I}(x) + \mathcal{I}(y)$, ukoliko su X i Y nezavisne slučajne promenljive. Dakle, informacija koju nosi događaj $\{X = x, Y = y\}$ jednaka je zbiru informacija koje nose događaji $\{X = x\}$ i $\{Y = y\}$.

Vidimo da izbor $\mathcal{I}(x) = -\log_b p(x)$ zadovoljava sve prethodne zahteve.

[1] Na primer, činjenica da je u Sibiru temperatura bila $-40°$C nosi manje informacija nego činjenica da je ista temperatura bila u Srbiji.

Zaista, iz nezavisnosti slučajnih promenljivih X i Y sledi

$$\begin{aligned}\mathcal{I}(x,y) &= -\log_b p(x,y) = -\log_b(p(x)p(y)) \\ &= -\log_b p(x) - \log_b p(y) = \mathcal{I}(x) + \mathcal{I}(y).\end{aligned}$$

U sledećem odeljku, pokazaćemo da je ovo i jedini mogući izbor funkcije $\mathcal{I}(x)$ koja zadovoljava prethodna svojstva. Možemo uzeti proizvoljnu bazu logaritma, ali već u narednom primeru objasnićemo zbog čega biramo bazu $b = 2$.

Srednju količinu informacija koju emituje izvor X nazivamo **entropijom** i označavamo sa $H(X)$ (pisaćemo i $H(p)$, gde se podrazumeva da je p zapravo p_X).

Definicija 3.1.1. *Entropija (jednodimenzione ili višedimenzione) slučajne promenljive X, u oznaci $H(X)$, jednaka je matematičkom očekivanju slučajne promenljive $\mathcal{I}(X)$:*

$$H(X) = \mathbb{E}\ \mathcal{I}(X) = \sum_{x\in\mathcal{X}} p(x)\mathcal{I}(x) = -\sum_{x\in\mathcal{X}} p(x)\log_2 p(x)$$

Takođe, definišemo i entropiju u proizvoljnoj bazi b na isti način:

$$H_b(X) = \mathbb{E}\ \mathcal{I}_b(x) = -\sum_{x\in\mathcal{X}} p(x)\log_b p(x).$$

Koristimo i oznake $H(p)$ (gde je p zapravo p_X) kao i $H(p_1, p_2, \ldots, p_n)$ ako je $\mathcal{X} = \{x_1, \ldots, x_n\}$ konačan skup i $p_i = p(x_i)$ za $i = 1, 2, \ldots, n$.

Ukoliko je $p(x) = 0$, smatramo da je $-p(x)\log_2 p(x) = 0$ zato što važi $t\log_2 t \to 0$ kad $t \to 0+$. Pošto je $p(x) \in [0,1]$, važi da je $\log_2 p(x) \leq 0$ odnosno $H(X) \geq 0$.

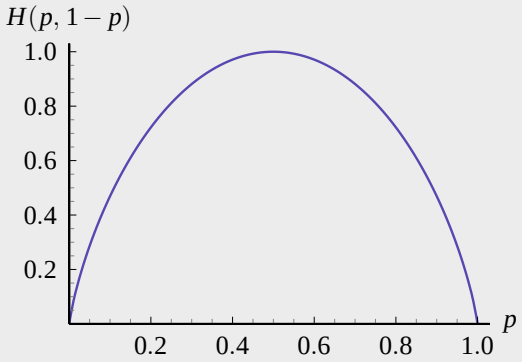
Primer 3.1.1. Neka je

$$X : \begin{pmatrix} 0 & 1 \\ p & 1-p \end{pmatrix}.$$

Tada je

$$H(X) = H(p, 1-p) = -p\log_2 p - (1-p)\log_2(1-p).$$

Kao što se vidi (slika 3.1), za $p = 0, 1$ je neodređenost 0 (tada X skoro sigurno nema slučajni karakter) dok je za $p = 1/2$ neodređenost maksimalna jer su oba ishoda (0 i 1) podjednako verovatna.



Slika 3.1: Grafik funkcije $H(p, 1-p)$ za $p \in [0, 1]$.

Vidimo i da je $H(p, 1-p)$ rastuća funkcija u prvoj polovini segmenta, odnosno za $p \in [0, 1/2]$ a opadajuća u drugoj polovini, odnosno za $p \in [1/2, 1]$. Grafik je simetričan u odnosu na pravu $x = 1/2$, odnosno vrednost npr. funkcije u tački $p = 0.1$ jednaka je vrednosti u tački $p = 0.9$. Ovo je logično, s obzirom da se ova dva primera dobijaju ako vrednosti 0 i 1 zamene mesta.

Označimo sa $h(n) = H(1/n, 1/n, \ldots, 1/n)$. Očigledno je $h(n) = \log_2 n$. Pokazaćemo kasnije da je ovo maksimalna moguća vrednost entropije H, ukoliko slučajna promenljiva X uzima n različitih vrednosti.

Prilikom definisanja entropije, usvojili smo da je osnova logaritma $b = 2$. Sa druge strane, važi $H_b(X) = (\log_2 b)^{-1} H(X)$, odnosno entropija se sa promenom baze menja samo do na multiplikativnu konstantu.

Izbor $b = 2$ proističe iz činjenice da je $h(2) = H(1/2, 1/2) = 1$ (videti sliku 3.1), odnosno da je količina informacije koju sadrži poruka o događaju koji ima samo dva podjednako verovatna ishoda (0 i 1), jednaka 1. Ovako definisana entropija izražava se u jedinicama **bit** (**informacioni bit**) odnosno **shannon**.

Naziv jedinice **(informacioni) bit** motivisan je i činjenicom da ukoliko posmatramo slučajno izabran bit u nekom skupu podataka (memorijski medijum, saobraćaj u računarskoj mreži, itd.), verovatnoca da je taj bit jednak 0 odnosno 1 je 1/2. Dakle, taj bit nosi količinu informacije 1 **bit**.

Pored ove, postoje još dve jedinice:

$$H(X) = -\sum_{i=1}^{n} p_i \log_{10} p_i \quad \text{[hartley]}$$

kao i

$$H(X) = -\sum_{i=1}^{n} p_i \ln p_i \quad [\text{nat}].$$

Ove jedinice koriste se isključivo u nekim specifičnim primenama.

Primer 3.1.2. Posmatrajmo sledeću slučajnu promenljivu

$$X : \begin{pmatrix} a & b & c & d \\ 1/2 & 1/4 & 1/8 & 1/8 \end{pmatrix}$$

i odredimo njenu entropiju:

$$H(X) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{8}\log_2\frac{1}{8} = \frac{7}{4}.$$

Ovo je u isto vreme i minimalan mogući srednji broj pitanja oblika "Da li je $X \in A$", gde je $A \subset \{a, b, c, d\}$, koji nam je potreban da bi utvrdili vrednost X.

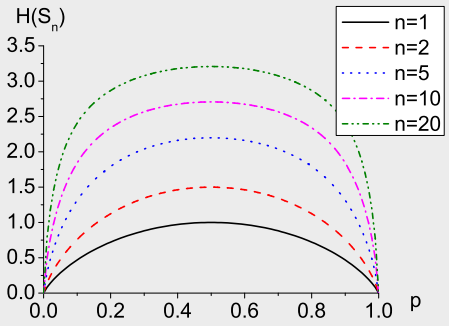
Srednji broj pitanja 7/4 u ovom slučaju daje jednostavna strategija: Najpre pitamo da li je $X = a$? Ako nije, pitamo da li je $X = b$? Ako ni tada nije, onda pitamo da li je $X = c$?

Kasnije ćemo dokazati da je u opštem slučaju, minimalni srednji broj pitanja ovog oblika između $H(X)$ i $H(X) + 1$.

Primer 3.1.3. Odredimo entropiju slučajne promenljive S_n koja ima binomnu $\mathcal{B}(n, p)$ raspodelu. Entropija ove slučajne promenljive jednaka je

$$\begin{aligned}
H(S_n) &= -\sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \log_2 \left[\binom{n}{k} p^k q^{n-k} \right] \\
&= -\sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \left[\log_2 \binom{n}{k} + k \log_2 p + (n-k) \log_2 q \right] \\
&= -\sum_{k=0}^{n} \binom{n}{k} k p^k q^{n-k} \log_2 p - \sum_{k=0}^{n} (n-k) \binom{n}{k} p^k q^{n-k} \log_2 q \\
&\qquad - \sum_{k=0}^{n} \binom{n}{k} \log_2 k \binom{n}{k} p^k q^{n-k} \\
&= -n(p \log_2 p + q \log_2 q) - \sum_{k=0}^{n} \binom{n}{k} \log_2 \binom{n}{k} p^k q^{n-k}.
\end{aligned}$$

Na slici 3.2 dat je grafik zavisnosti $H(S_n)$ u funkciji od p:



Slika 3.2: Grafik entropije $H(S_n)$ u zavisnosti od p

Entropiju na isti način definišemo i za diskretne slučajne promenljive koje mogu da uzmu prebrojivo mnogo vrednosti iz skupa $\mathcal{X} = \{x_1, x_2, \ldots\}$. Razlika je samo u tome što je suma u definicionom izrazu beskonačna, i predstavlja red koji može ili ne mora da bude konvergetan. Samim tim, entropija ovakvih slučajnih promenljivih nije uvek definisana.

Primer 3.1.4. Eksperimentalno je utvrđeno da je na nekom fakultetu kod nekog profesora, stopa padanja na ispitu jednaka p $(0 < p < 1)$. Student polaže ispit sve dok ne položi. Potrebno je odrediti srednju količinu informacija koje nosi slučajna promenljiva X koja označava broj izlazaka na ispit studenta.

Nije teško utvrditi da je $P(X = k) = p^{k-1}q$ gde je $q = 1 - p$. Dakle, raspodela slučajne promenljive X jednaka je

$$X : \begin{pmatrix} 1 & 2 & 3 & \cdots & k & \cdots \\ q & pq & p^2q & \cdots & p^{k-1}q & \cdots \end{pmatrix}$$

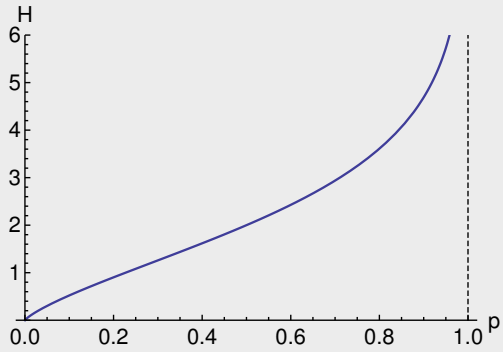
a entropija

$$\begin{aligned} H(X) &= -\sum_{k=1}^{+\infty} p^{k-1}q \log_2(p^{k-1}q) \\ &= -\sum_{k=1}^{+\infty} (k-1)p^{k-1}q \log_2 p - q \log_2 q \sum_{k=1}^{+\infty} p^{k-1} \end{aligned}$$

Sumiranjem prethodnih izraza dobijamo:

$$\begin{aligned} H(X) &= -\frac{pq}{(1-p)^2}\log_2 p - \frac{q}{1-p}\log_2 q \\ &= -\frac{p}{1-p}\log_2 p - \log_2(1-p). \end{aligned}$$

Grafik entropije $H(X)$ u funkciji od p dat je na slici 3.3.



Slika 3.3: Grafik entropije $H(X)$ u zavisnosti od p

Vidimo sa grafika (a nije teško ni dokazati) da $H(X)$ teži beskonačnosti kad $p \to 1$. Dakle, što veća stopa padanja na ispitu (manja prolaznost), to je sam ispit informativniji :)!

Takođe, nije teško utvrditi da je $H(X) = 1$ za $p = 0.227$, tj. da je u slučaju prolaznosti od $q = 1 - p \approx 73\%$, sam ispit daje informaciju od 1 bit. Možda će ovaj podatak biti od koristi kolegama iz Bolonje za sledecu (release) verziju Bolonjske deklaracije.

Entropija poseduje i sledeću važnu osobinu.

Teorema 3.1.1. *Neka je X diskretna slučajna promenljiva koja uzima vrednosti iz skupa $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ sa verovatnoćama $p_1, p_2, \ldots, p_n$. Tada je*

$$\begin{aligned} H(p_1, p_2, \ldots, p_n) =& H(q_1, q_2) + q_1 H\left(\frac{p_1}{q_1}, \frac{p_2}{q_1}, \ldots, \frac{p_r}{q_1}\right) \\ &+ q_2 H\left(\frac{p_{r+1}}{q_2}, \frac{p_{r+2}}{q_2}, \ldots, \frac{p_n}{q_2}\right). \end{aligned} \tag{3.1}$$

za svako $1 \le r \le n$, gde je $q_1 = p_1 + p_2 + \ldots + p_r$ i $q_2 = p_{r+1} + p_{r+2} + \ldots + p_n$.

Dokaz. Označimo sa D desnu stranu izraza. Tada je:

$$D = H(q_1, q_2) + q_1 H\left(\frac{p_1}{q_1}, \frac{p_2}{q_1}, \ldots, \frac{p_r}{q_1}\right) + q_2 H\left(\frac{p_{r+1}}{q_2}, \frac{p_{r+2}}{q_2}, \ldots, \frac{p_n}{q_2}\right)$$
$$= -q_1 \log_2 q_1 - q_2 \log_2 q_2 - q_1 \sum_{i=1}^{r} \frac{p_i}{q_1} \log_2 \frac{p_i}{q_1} - q_2 \sum_{i=r+1}^{n} \frac{p_i}{q_2} \log_2 \frac{p_i}{q_2}$$

Dalje, pošto je $q_1 = \sum_{i=1}^{r} p_i$ i $q_2 = \sum_{i=r+1}^{n} p_i$, sledi:

$$D = -\sum_{i=1}^{r} p_i \log_2 q_1 - \sum_{i=r+1}^{n} p_i \log_2 \frac{p_i}{q_1} - \sum_{i=1}^{r} p_i \log_2 q_2 - \sum_{i=1}^{r} p_i \log_2 \frac{p_i}{q_2}$$
$$= -\sum_{i=1}^{r} p_i \log_2 p_i - \sum_{i=r+1}^{n} p_i \log_2 p_i = H(p_1, p_2, \ldots, p_n)$$

Ovim je dokaz teoreme završen. □

Prethodna osobina može da se interpretira na sledeći način. Neka je $A = \{x_1, x_2, \ldots, x_r\}$ i $B = \{x_{r+1}, x_{r+2}, \ldots, x_n\}$. Informacija da se realizovao događaj $\{X = x_i\}$ može se saopštiti iz dva dela. Najpre se saopšti skup (A ili B) kom realizovana vrednost pripada, a onda indeks realizovanog elementa (u tom skupu).

Prvo saopštenje je realizacija slučajne promenljive $I_{\{X \in A\}}$ i nosi (srednju) količinu informacije $H(q_1, q_2)$ a dok drugo saopštenje nosi ili $H\left(\frac{p_1}{q_1}, \frac{p_2}{q_1}, \ldots, \frac{p_r}{q_1}\right)$ ili $H\left(\frac{p_{r+1}}{q_2}, \frac{p_{r+2}}{q_2}, \ldots, \frac{p_n}{q_2}\right)$, u zavisnosti od poruke u prvom saopštenju. Zbog toga je $H(X)$ upravo zbir prethodno navedenih količina informacija, kao što je dato u (3.2).

3.2 Aksiomatsko zasnivanje entropije

Pretpostavili smo da količina sopstvene informacije $\mathcal{I}(x)$ koju nosi događaj $\{X = x\}$ zadovoljava sledeće uslove:

1. Neprekidna i opadajuća funkcija verovatnoće $p(x)$;
2. $\mathcal{I}((x, y)) = \mathcal{I}(x) + \mathcal{I}(y)$, ukoliko su X i Y nezavisne slučajne promenljive;

Ovi uslovi su prilično intuitivni, imajući u vidu našu predstavu o tome šta bi mera informativnosti događaja morala da zadovolji. Međutim, ispostavlja se da je logaritamska funkcija jedini mogući izbor. O tome govori naredna teorema.

Teorema 3.2.1. *Jedina funkcija $\mathcal{I}(x)$ koja zadovoljava prethodne uslove je $\mathcal{I}(x) = -\log_b p(x)$ gde je $b > 1$.*

Dokaz. Iz uslova 2 zaključujemo da važi:

$$\begin{aligned}\mathcal{I}((x,y)) &= \mathcal{I}(p(x,y)) = \mathcal{I}(p(x)p(y)) = \mathcal{I}(uv)\\ &= \mathcal{I}(x) + \mathcal{I}(y) = \mathcal{I}(p(x)) + \mathcal{I}(p(y)) = \mathcal{I}(u) + \mathcal{I}(v)\end{aligned}$$

Prema tome, funkcija $\mathcal{I}'$ zadovoljava jednačinu $\mathcal{I}'(uv) = \mathcal{I}'(u) + \mathcal{I}'(v)$ za svako $u, v \in (0, 1]$. Uz to, iz uslova 2 sledi da je $\mathcal{I}'(u)$ neprekidna i opadajuća funkcija. Uvedimo smenu $f(t) = \mathcal{I}'(e^t)$. Tada je

$$f(t_1 + t_2) = \mathcal{I}'(e^{t_1+t_2}) = \mathcal{I}'(e^{t_1}e^{t_2}) = \mathcal{I}'(e^{t_1}) + \mathcal{I}'(e^{t_2}) = f(t_1) + f(t_2)$$

pri čemu je funkcija $f : (-\infty, 0) \to \mathbb{R}$ takođe neprekidna i opadajuća. Primetimo da je $f(0) = f(0+0) = f(0) + f(0)$ odakle je $f(0) = 0$. Funkciju možemo dodefinisati na pozitivnom delu realne prave sa $f(t) = -f(-t)$ za $t > 0$. Direktno se proverava da je uslov $f(t_1 + t_2) = f(t_1) + f(t_2)$ sada ispunjen za svako $t_1, t_2 \in \mathbb{R}$. Funkciju f određujemo u sledeća 3 koraka:

1. Matematičkom indukcijom lako pokazujemo da je

   $$f(t_1 + t_2 + \ldots + t_n) = f(t_1) + f(t_2) + \ldots + f(t_n)$$

   za svako $n \in \mathbb{N}$ i svako $t_1, t_2, \ldots, t_n \in \mathbb{R}$. Specijalno, za $t_1 = t_2 = \ldots = t_n = t$ je $f(n \cdot t) = nf(t)$, dok za $t = 1$ dobijamo $f(n) = nf(1) = cn$, gde je $c = f(1)$.

2. Neka je $t = p/q > 0$ racionalan broj. Iz $f(qt) = qf(t)$ i $f(qt) = f(p) = pf(1)$ sledi

   $$f(t) = \frac{1}{q}f(qt) = \frac{1}{q}f(p) = c\frac{p}{q} = ct$$

   Ukoliko je $t < 0$ racionalan broj, onda je $-t > 0$ takođe racionalan broj i $f(t) = -f(-t) = -ct$.

3. Neka je $t \in \mathbb{R} \setminus \mathbb{Q}$, tj. iracionaln broj. Tada postoji niz racionalnih brojeva t_n takav da $t_n \to t$ kada $n \to +\infty$. Iz neprekidnosti funkcije f sledi da onda $f(t_n) \to f(t)$, a pošto je $f(t_n) = ct_n$ (deo 2) onda sledi da je $f(t_n) = ct_n \to ct = f(t)$.

Prema tome, dokazali smo da je $f(t) = ct$ za neku konstantu $c \in \mathbb{R}$. Neposrednom proverom možemo utvrditi da linearna funkcija $f(t) = ct$ zaista zadovoljava jednačinu za proizvoljno $c \in \mathbb{R}$.

Međutim, funkcija $f(t) = ct$ je opadajuća samo za $c < 0$. Vratimo se sada na funkciju $\mathcal{I}$. Pošto je $f(t) = \mathcal{I}'(e^t)$ onda je $\mathcal{I}'(q) = f(\ln q) = c \ln q$. Neka je $b = e^{-1/c}$, odnosno $c = -(\ln b)^{-1}$. Pošto je $-c > 0$, sledi da je $b = e^{-1/c} > 1$. Tada je $\mathcal{I}'(q) = -(\ln q)/(\ln b) = -\log_b q$, odnosno $\mathcal{I}(x) = -\log_b p(x)$. Ovim je dokaz završen. □

Prethodna teorema nam omogućava da o svojstva 1 i 2 koja zadovoljava količina sopstvene informacije, tretiramo kao o **aksiome**, s obzirom da slede na osnovu naše intuicije, a u isto vreme i formalno određuju veličinu $\mathcal{I}(x)$ do na multiplikativnu konstantu, odnosno bazu logaritma.

Na sličan način je i entropija karakterisana sa ukupno 4 svojstava koja su data u formulaciji naredne teoreme. Ova svojstva takođe mogu da se posmatraju kao **aksiome entropije**. Primetimo da je svojstvo 3 zapravo ono svojstvo koje smo dokazali u prethodnom odeljku (Teorema 3.1.1).

Teorema 3.2.2. *Ukoliko funkcija $H(p_1, p_2, \ldots, p_n)$ i njoj pridružena funkcija $h(n) = H(1/n, 1/n, \ldots, 1/n)$ zadovoljava uslove:*

1. *Ako je $m, n \in \mathbb{N}$ i $m < n$, onda je $h(m) < h(n)$.*
2. *Za svako $m, n \in \mathbb{N}$ važi $h(m \cdot n) = h(m) + h(n)$.*
3. *Neka je $1 \leq r \leq n-1$ i $r \in \mathbb{N}$. Označimo $q_1 = p_1 + p_2 + \ldots + p_r$ i $q_2 = p_{r+1} + p_{r+2} + \ldots + p_n$. Tada je*

$$\begin{aligned} H(p_1, p_2, \ldots, p_n) =& H(q_1, q_2) + q_1 H\left(\frac{p_1}{q_1}, \frac{p_2}{q_1}, \ldots, \frac{p_r}{q_1}\right) \\ &+ q_2 H\left(\frac{p_{r+1}}{q_2}, \frac{p_{r+2}}{q_2}, \ldots, \frac{p_n}{q_2}\right). \end{aligned} \tag{3.2}$$

4. *Funkcija $H(p, 1-p)$ je neprekidna u svakoj tački $p \in (0, 1)$.*

onda je

$$H(p_1, p_2, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log_b p_i, \tag{3.3}$$

gde je $b > 1$ proizvoljna baza logaritma.

Dokaz. Matematičkom indukcijom jednostavno dokazujemo da je

$$h(n_1 n_2 \cdots n_k) = h(n_1) + h(n_2) + \ldots + h(n_k)$$

za svako $k \in \mathbb{N}$ i $n_1, n_2, \ldots, n_k \in \mathbb{N}$. Specijalno, za $n_1 = n_2 = \cdots = n_k = n$ je $h(n^k) = kh(n)$. Zamenom $n = 1$ dobijamo $h(1) = h(1^k) = kh(1)$ odnosno

$h(1) = 0$. Iz svojstva 1, dobijamo da za svako $n > 1$ važi $h(n) > h(1) = 0$ odnosno $h(n) > 0$.

Neka su $n \geq 2$ i r proizvoljni prirodni brojevi. Tada postoji $k \in \mathbb{N}_0$ tako da je

$$n^k \leq 2^r < n^{k+1}.$$

Konkretno, prethodna nejednakost je zadovoljena za $k = \lfloor r \log_n 2 \rfloor$. Pošto je h rastuća funkcija (svojstvo 1) dobijamo da je $h(n^k) \leq h(2^r) < h(n^{k+1})$ odnosno

$$kh(n) \leq rh(2) < (k+1)h(n).$$

Deljenjem sa $rh(n)$, s obzirom da smo pokazali da je $h(n) > 0$, dobijamo

$$\frac{k}{r} \leq \frac{h(2)}{h(n)} < \frac{(k+1)}{r}.$$

Logaritmovanjem i deljenjem sa $r \log n$ dobijamo istu tu nejednakost, samo za logaritamsku funkciju

$$\frac{k}{r} \leq \frac{\log 2}{\log n} < \frac{(k+1)}{r}.$$

Obzirom da se vrednosti $h(2)/h(n)$ i $(\log 2)/(\log n)$ nalaze u istom polusegmentu $[k/r, (k+1)/r)$ dužine $1/r$, njihova razlika mora biti manja od dužine polusegmenta. Drugim rečima:

$$\left| \frac{h(2)}{h(n)} - \frac{\log 2}{\log n} \right| < \frac{1}{r}$$

S obzirom da smo pretpostavili da je r proizvoljan prirodan broj, zaključujemo da razlika u prethodnoj nejednakosti mora biti jednaka nuli. Dakle

$$h(n) = \frac{h(2)}{\log 2} \log n = c \log n$$

gde je $c = h(2)/(\log 2)$. Iz $h(2) > 0$ i $\log 2 > 0$ sledi da je $c > 0$ pa je $c = (\log b)^{-1}$ za $b = e^{1/c}$. Sada je $h(n) = (\log n)/(\log b) = \log_b n$.

Najpre izvodimo izraz za $H(p, 1-p)$ ukoliko je $p = r/s$ racionalan broj. Koristimo svojstvo 3:

$$h(s) = H\underbrace{\left(\frac{1}{s}, \frac{1}{s}, \ldots, \frac{1}{s}\right)}_{s} = H\left(\frac{r}{s}, \frac{s-r}{s}\right) + \frac{r}{s}h(r) + \frac{s-r}{s}h(s-r)$$

odakle je

$$\begin{aligned} H(p,1-p) = H\left(\frac{r}{s},\frac{s-r}{s}\right) &= -\frac{r}{s}h(r) - \frac{s-r}{s}h(s-r) + h(s) \\ &= -\frac{r}{s}\log_b r - \frac{s-r}{s}\log_b(s-r) + \log_b s \\ &= -\frac{r}{s}\log_b \frac{r}{s} - \frac{s-r}{s}\log_b \frac{s-r}{s} \end{aligned}$$

odnosno

$$H(p,1-p) = -p\log_b p - (1-p)\log_b(1-p).$$

Korišćenjem svojstva neprekidnosti funkcije $H(p,1-p)$, na sličan način kao u dokazu teoreme 3.2.1 dokazujemo da prethodna formula važi i za realne brojeve $p \in [0,1]$.

Konačno, dokaz teoreme završavamo matematičkom indukcijom. Već smo dokazali da teorema (odnosno izraz (3.3)) važi za $n = 2$. Pretpostavimo da izraz važi za n, i dokažimo da važi i za $n+1$. Zaista,

$$H(p_1,p_2,\ldots,p_n,p_{n+1}) = H(s_n,p_{n+1}) + s_n H\left(\frac{p_1}{s_n},\frac{p_2}{s_n},\ldots,\frac{p_n}{s_n}\right) + p_{n+1}h(1)$$

gde je $s_n = p_1 + p_2 + \ldots + p_n$. Prema indukcijskoj hipotezi je

$$H(p_1,p_2,\ldots,p_{n+1}) = -s_n\log_b s_n - p_{n+1}\log_b p_{n+1} - s_n\sum_{i=1}^{n}\frac{p_i}{s_n}\log_b\frac{p_i}{s_n}$$

odnosno

$$\begin{aligned} H(p_1,p_2,\ldots,p_{n+1}) &= -\sum_{i=1}^{n} p_i\log_b s_n - \sum_{i=1}^{n} p_i\log_b\frac{p_i}{s_n} - p_{n+1}\log_b p_{n+1} \\ &= -\sum_{i=1}^{n+1} p_i\log_b p_i \end{aligned}$$

Ovim je dokaz teoreme završen. □

3.3 Važna lema i maksimum entropije

Naredna lema će često biti korišćena u narednim dokazima, a kao njenu direktnu posledicu, pokazaćemo da uniformna raspodela maksimizuje entropiju.

Lema 3.3.1. ***(važna lema)*** *Neka je $n \in \mathbb{N}$ i $p_1, p_2, \ldots, p_n$ i $q_1, q_2, \ldots, q_n$ realni brojevi takvi da je $p_i \geq 0$, $q_i \geq 0$, $q_i = 0$ samo ako je $p_i = 0$ i $q_1 + q_2 + \ldots + q_n \leq p_1 + p_2 + \ldots + p_n$. Tada je*

$$-\sum_{i=1}^{n} p_i \log_2 p_i \leq -\sum_{i=1}^{n} p_i \log_2 q_i,$$

gde jednakost važi akko je $p_i = q_i$ za svako $i = 1, 2, \ldots, n$.

Dokaz. [2] Pretpostavimo najpre da je $p_i > 0$ i $q_i > 0$ za svako $i = 1, 2, \ldots, n$. Podsetimo se da za svaki broj $x > 0$ važi $\ln x \leq x - 1$, pošto je $x - 1$ tangenta konkavne funkcije $\ln x$ u tački $x = 1$.

Zamenom $x = q_i/p_i$ i $\ln x = \log_2 x / \log_2 e$ ($\log_2 e > 0$) dobijamo:

$$\log_2 \frac{q_i}{p_i} \leq \log_2 e \left(\frac{q_i}{p_i} - 1 \right)$$

Množenjem prethodne nejednakosti sa $p_i > 0$ i sumiranjem po $i = 1, 2, \ldots, n$ sledi

$$\sum_{i=1}^{n} p_i \log_2 \frac{q_i}{p_i} \leq \log_2 e \left(\sum_{i=1}^{n} q_i - \sum_{i=1}^{n} p_i \right) \leq 0$$

gde poslednja nejednakost važi iz uslova leme. Sada iz $\log_2 q_i/p_i = \log_2 q_i - \log_2 p_i$ sledi

$$0 \geq \sum_{i=1}^{n} p_i \log_2 \frac{q_i}{p_i} = \sum_{i=1}^{n} p_i \log_2 q_i - \sum_{i=1}^{n} p_i \log_2 p_i$$

što je ekvivalentno tvrđenju leme. Jednakost u $\ln x = x - 1$ važi akko je $x = 1$ odakle sledi da je $q_i/p_i = 1$ odnosno $q_i = p_i$ za svako $i = 1, 2, \ldots, n$.

Pretpostavimo sada, bez gubitka opštosti, da je $q_l = q_{l+1} = \ldots = q_n = 0$ kao i da je $q_i > 0$ za $i = 1, 2, \ldots, l-1$. Na sličan način, pretpostavimo da je $p_m = p_{m+1} = \ldots p_n = 0$ a $p_i > 0$ za $i = 1, 2, \ldots, m-1$. Na osnovu uslova leme je $m \leq l$ [3]. Imajući u vidu da je $x \log_2 x = 0$ za $x \to 0+$, važi:

$$-\sum_{i=1}^{n} p_i \log_2 p_i = -\sum_{i=1}^{m-1} p_i \log_2 p_i,$$

$$-\sum_{i=1}^{n} p_i \log_2 q_i = -\sum_{i=1}^{m-1} p_i \log_2 q_i - \sum_{i=m}^{l-1} 0 \cdot \log_2 q_i = -\sum_{i=1}^{m-1} p_i \log_2 q_i,$$

[2] Dokaz i formulacija u [Šešelja, p.36] ne razmatra opštiji slučaj koji dopušta da su neki p_i i q_i jednaki 0, što je značaja za kasnije posledice ove leme.

[3] Odnosno, za svaki $q_i = 0$ je i odgovarajuci $p_i = 0$. Obrnuto u opštem slučaju ne mora da važi. Dakle, $m \leq l$.

Dalje je i

$$\sum_{i=1}^{m-1} p_i = \sum_{i=1}^{n} p_i \geq \sum_{i=1}^{n} q_i = \sum_{i=1}^{l-1} q_i \geq \sum_{i=1}^{m-1} q_i$$

kao i $p_1, p_2, \ldots, p_{m-1} > 0$ i $q_1, q_2, \ldots, q_{m-1} > 0$, čime smo ovaj slučaj sveli na prethodni. Primetimo da jednakost u poslednjoj nejedakosti prethodnog izraza važi akko je $m = l$, pa na osnovu prvog slučaja dobijamo da i ovde jednakost važi akko je $p_i = q_i$ za $i = 1, 2, \ldots, n$. □

Na osnovu prethodne leme lako dobijamo da entropija dostiže svoj maksimum ukoliko su svi ishodi podjednako verovatni.

Posledica 3.3.2. *Entropija $H(p_1, p_2, \ldots, p_n)$ ima maksimalnu vrednost $h(n) = \log_2 n$, tj važi $H(p_1, p_2, \ldots, p_n) \leq \log_2 n$. Jednakost se dostiže akko je $p_i = 1/n$ za svako $i = 1, 2, \ldots, n$, tj. ako su svi ishodi jednako verovatni.*

Dokaz. Primenom prethodne leme za $q_i = 1/n$, $i = 1, 2, \ldots, n$ dobijamo

$$H(p_1, p_2, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log_2 p_i \leq -\sum_{i=1}^{n} p_i \log_2(1/n) = \log_2 n.$$

Jednakost važi akko je $p_i = q_i = 1/n$ za svako $i = 1, 2, \ldots, n$. □

3.4 Entropija višedimenzionalne slučajne promenljive i uslovna entropija

Posmatrajmo dvodimenzionalnu diskretnu slučajnu promenljivu (X, Y). Neka promenljiva X uzima vrednosti iz skupa $\mathcal{X} = \{x_1, x_2, \ldots, x_m\}$ a promenljiva Y vrednosti iz skupa $\mathcal{Y} = \{y_1, y_2, \ldots, y_n\}$. Na osnovu Definicije 3.1.1, entropija promenljive (X, Y) jednaka je

$$H(X, Y) = -\sum_{\substack{x \in \mathcal{X} \\ y \in \mathcal{Y}}} p(x, y) \log_2 p(x, y).$$

Poznato je da su koordinate X i Y dvodimenzionalne slučajne promenljive (X, Y) takođe slučajne promenljive. Naredna lema daje pogodne izraze za entropije ovih slučajnih promenljivih, koje ćemo koristiti kasnije.

Lema 3.4.1. *Entropije slučajnih promenljivih X i Y date su formulama*

$$H(X) = -\sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} p(x,y)\log_2 p(x), \quad H(Y) = -\sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} p(x,y)\log_2 p(y)$$

gde su $p(x)$ i $p(y)$ odgovarajuće marginalne raspodele.

Dokaz. S obzirom da je $p(x) = \sum_{y\in\mathcal{Y}} p(x,y)$ dobijamo da je

$$H(X) = -\sum_{x\in\mathcal{X}} p(x)\log_2 p(x) = -\sum_{x\in\mathcal{X}} \left[\sum_{y\in\mathcal{Y}} p(x,y)\right] \log_2 p(x) = -\sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} p(x,y)\log_2 p(x).$$

Drugi izraz se analogno dokazuje. □

Veza između entropije $H(X,Y)$ i marginalnih entropija $H(X)$ i $H(Y)$ data je sledećom teoremom.

Teorema 3.4.2. *Za slučajne promenljive X i Y važi $H(X,Y) \leq H(X) + H(Y)$, a jednakost važi akko su X i Y nezavisne.*

Dokaz. Neka je $q(x,y) = p(x)p(y)$. Ukoliko je $q(x,y) = 0$ tada je $p(x) = 0$ ili $p(y) = 0$ i u oba slučaja je $p(x,y) = 0$. Sa druge strane je:

$$\begin{aligned}\sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} q(x,y) &= \sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} p(x)p(y) = \left[\sum_{x\in\mathcal{X}} p(x)\right]\left[\sum_{y\in\mathcal{Y}} p(y)\right] \\ &= 1\cdot 1 \leq 1 = \sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} p(x,y).\end{aligned}$$

Primenom Leme 3.3.1 dobijamo da je

$$\begin{aligned}H(X,Y) &= -\sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} p(x,y)\log_2 p(x,y) \leq -\sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} p(x,y)\log_2 q(x,y) \\ &= -\sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} p(x,y)\log_2 p(x)p(y) \\ &= -\sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} p(x,y)\log_2 p(x) - \sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} p(x,y)\log_2 p(y) = H(X) + H(Y)\end{aligned}$$

Jednakost važi ako i samo ako je $p(x,y) = q(x,y) = p(x)p(y)$, odnosno, ako i samo ako su X i Y nezavisne. Ovim je dokaz teoreme završen. □

Prethodnu teoremu možemo interpretirati na sledeći način. Ukoliko su X i Y nezavisne, moramo imati potpunu informaciju i o X i o Y da bi smo imali potpunu informaciju o (X,Y). Pritom su ta dva dela disjunktna. U suprotnom, kada dobijemo potpunu informaciju za X, samim tim dobijamo i "deo" informacije za Y.

Definicija 3.4.1. ***Uslovna sopstvena informacija*** *elementa $x \in \mathcal{X}$ u odnosu na element $y_0 \in \mathcal{Y}$ definisana je sa*

$$\mathcal{I}(x|y_0) = -\log_2 P(\{X = x\}|\{Y = y_0\}) = -\log_2 p(x|y_0).$$

Uslovna entropija *slučajne promenljive X u odnosu na $y_0 \in \mathcal{Y}$ definisana je sa*

$$H(X|y_0) = \sum_{x\in\mathcal{X}} p(x|y_0)\mathcal{I}(x|y_0) = -\sum_{x\in\mathcal{X}} p(x|y_0)\log_2 p(x|y_0).$$

Drugim rečima, to je srednja vrednost za $\mathcal{I}(x|y_0)$ po uslovnoj raspodeli $p(x|y_0)$. Ako sada usrednjimo $H(X|y_0)$ za svako $y_0 \in \mathcal{Y}$ (po marginalnoj raspodeli $p(y_0)$) dobijamo ***uslovnu entropiju slučajne promenljive X u odnosu na Y:***

$$H(X|Y) = \sum_{y\in\mathcal{Y}} p(y)H(X|y).$$

Iz definicionih izraza za $H(X|Y)$ i $H(X|y)$ lako se dobija da važi:

$$\begin{aligned} H(X|Y) &= \sum_{y\in\mathcal{Y}} p(y)H(X|y) = -\sum_{y\in\mathcal{Y}} p(y)\sum_{x\in\mathcal{X}} p(x|y)\log_2 p(x|y) \\ &= -\sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} p(y)p(x|y)\log_2 p(x|y) = -\sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} p(x,y)\log_2 p(x|y). \end{aligned}$$

Prethodni izraz može da se napiše i kao $H(X|Y) = \mathbb{E}\ \mathcal{I}(X|Y)$. Analogno dobijamo da važi

$$H(Y|X) = -\sum_{x\in\mathcal{X},\ y\in\mathcal{Y}} p(x,y)\log_2 p(y|x) = \mathbb{E}\ \mathcal{I}(Y|X)$$

odakle dobijamo da je u opštem slučaju $H(X|Y) \neq H(Y|X)$. Takođe, uslovna entropija se lako generalizuje za slučaj n slučajnih promenljivih:

$$\begin{aligned} &H(X_1, X_2, \ldots, X_r|X_{r+1}, X_{r+2}, \ldots, X_n) \\ &\quad = -\sum_{x_i\in\mathcal{X}_i} p(x_1, x_2, \ldots, x_n)\log_2 p(x_1, x_2, \ldots, x_r|x_{r+1}, x_{r+2}, \ldots, x_n). \end{aligned}$$

Pošto je uslovna raspodela ispod znaka logaritma uvek između 0 i 1, sledi da je uslovna entropija uvek nenegativna, odnosno $H(X_1, \ldots, X_r|X_{r+1}, \ldots, X_n) \geq 0$. Takođe, ukoliko je $p(x_1, \ldots, x_r|x_{r+1}, \ldots, x_n) = 0$ tada je i $p(x_1, \ldots, x_n) = 0$, što znači da je odgovarajući sabirak u sumi oblika $0 \cdot \log_2 0$, koji dogovorom smatramo da je jednak 0.

Teorema 3.4.3. *Za slučajne promenljive X i Y važi $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$.*

Dokaz. Iz $p(x,y) = p(x)p(y|x)$ dobijamo:

$$\begin{aligned} H(X,Y) &= -\sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} p(x,y)\log_2 p(x,y) = -\sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} p(x,y)\log_2 p(x)p(y|x) \\ &= -\sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} p(x,y)\log_2 p(x) - \sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} p(x,y)\log_2 p(y|x) \\ &= H(X) + H(Y|X) \end{aligned}$$

Druga jednakost se dokazuje analogno. □

Sada možemo dati potpuniju interpretaciju od one koju smo dali nakon Teoreme 3.4.2. Potpuna informacija o (X,Y) dobija se na osnovu potpune informacije o X $(H(X))$ i potpune uslovne informacije o Y pod uslovom X $(H(Y|X))$.

Iz definicionog izraza za entropiju i uslovnu entropiju, jasno je da prethodna teorema važi i u slučaju kada su X i Y višedimenzionalne slučajne promenljive.

Iz Teoreme 3.4.2 i Teoreme 3.4.3 dobija se:

Teorema 3.4.4. ***(Shannonova nejednakost)*** *Za slučajne promenljive X i Y važi $H(X|Y) \leq H(X)$. Jednakost važi akko su X i Y nezavisne.*

Dokaz. Tvrđenje sledi direktno iz $H(X,Y) = H(Y) + H(X|Y)$ i $H(X,Y) \leq H(X) + H(Y)$. □

Naredna teorema formalizuje činjenicu da se obradom podataka ne može povećati, već samo smanjiti njihova informativnost. Dakle, primenom (determinističke) funkcije na slučajnu promenljivu X, njena entropija ostaje ista ili se smanjuje.

Teorema 3.4.5. *Neka je $\phi : \mathcal{X} \to \mathcal{Y}$ funkcija koja je "na", odnosno za koju važi da je $\phi(\mathcal{X}) = \mathcal{Y}$, i neka je $Y = \phi(X)$. Tada je $H(Y) \leq H(X)$.*

Dokaz. Pošto je $Y = \phi(X)$, znači da je vrednost slučajne promenljive Y u potpunosti određena vrednošću promenljive X. Drugim rečima, $p(y|x) = 1$ ako je $y = \phi(x)$ a $p(y|x) = 0$ ako je $y \neq \phi(x)$. Iz $p(x, y) = p(x)p(y|x)$ dobijamo da je $p(x, y) = p(x)$ za $y = \phi(x)$ i $p(x, y) = 0$ za $y \neq x$. Dalje računamo uslovnu entropiju $H(Y|X)$ na sledeći način:

$$H(Y|X) = -\sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} p(x,y)\log_2 p(y|x) = -\sum_{y=\phi(x)} p(x)\log_2 1 - \sum_{y\neq\phi(x)} 0\cdot\log_2 0 = 0.$$

Dalje iz $H(X, Y) = H(X) + H(Y|X)$ dobijamo $H(X) = H(X, Y) = H(Y) + H(X|Y) \geq H(Y)$ jer je $H(X|Y) \geq 0$. Ovim je dokaz teoreme završen. □

Ako su X, Y i Z diskretne slučajne promenljive, onda važi:

$$\begin{aligned} p(x, y|z) &= p(x|y, z)p(y|z) \\ &= p(y|x, z)p(x|z) \end{aligned} \tag{3.4}$$

Ovaj izraz se lako dokazuje, imajući u vidu da je

$$p(x|y, z)p(y|z) = \frac{p(x, y, z)}{p(y, z)} \cdot \frac{p(y, z)}{p(z)} = \frac{p(x, y, z)}{p(z)} = p(x, y|z).$$

Množenjem izraza (3.4) sa $p(x, y, z)$ i sumiranjem po svim $x \in \mathcal{X}$, $y \in \mathcal{Y}$ i $z \in \mathcal{Z}$ dobijamo da važi sledeća lema:

Lema 3.4.6. *Za slučajne promenljive X, Y i Z je $H(X, Y|Z) = H(X|Z) + H(Y|X, Z)$.*

Napomenimo da X, Y i Z mogu biti i višedimenzionalne diskretne slučajne promenljive i da tada važe kako Lema 3.4.6 tako i izraz (3.4).

Teorema 3.4.7 predstavlja uopštenje Teoreme 3.4.3 za n promenljivih.

Teorema 3.4.7. *Za slučajne promenljive $X_1, X_2, \ldots, X_n$ i Y važi:*

$$H(X_1, X_2, \ldots, X_n) = \sum_{k=1}^{n} H(X_k|X_1, X_2, \ldots, X_{k-1}),$$

$$H(X_1, X_2, \ldots, X_n|Y) = \sum_{k=1}^{n} H(X_k|X_1, X_2, \ldots, X_{k-1}, Y).$$

*Ovo pravilo je poznato kao **lančano pravilo** (**chain rule**).*

Dokaz. [Cover, p.23] Primenom izraza (3.4) dobijamo:

$$\begin{aligned} &p(x_k, x_{k+1}, \ldots, x_n | x_1, \ldots, x_{k-1}) \\ &\quad = p(x_{k+1}, \ldots, x_n | x_1, \ldots, x_{k-1}, x_k) \cdot p(x_k | x_1, \ldots, x_{k-1}) \end{aligned}$$

Uzastopnom primenom prethodnog izraza dobijamo

$$\begin{aligned} p(x_1, x_2, \ldots, x_n) &= p(x_2, x_3, \ldots, x_n | x_1) p(x_1) \\ &= p(x_3, \ldots, x_n | x_1, x_2) p(x_2 | x_1) p(x_1) \\ &= p(x_4, \ldots, x_n | x_1, x_2, x_3) p(x_3 | x_1, x_2) p(x_2 | x_1) p(x_1) \\ &= \cdots \\ &= \prod_{k=1}^{n} p(x_k | x_1, x_2, \ldots, x_{k-1}) \end{aligned}$$

Logaritmovanjem i množenjem poslednje jednakosti sa $p(x_1, x_2, \ldots, x_n)$, i sumiranjem po svim $x_i \in \mathcal{X}_i$ sledi prva jednakost u teoremi. Druga jednakost se dobija analogno. □

Posledica 3.4.8. *Ako su slučajne promenljive $X_1, X_2, \ldots, X_n$ nezavisne, tada je*

$$H(X_1, X_2, \ldots, X_n) = H(X_1) + H(X_2) + \ldots + H(X_n).$$

Dokaz. Iz nezavisnosti slučajnih promenljivih $X_1, X_2, \ldots, X_n$ sledi

$$p(x_k | x_1, x_2, \ldots, x_{k-1}) = p(x_k)$$

za svako $x_k \in \mathcal{X}_k$ $(k = 1, 2, \ldots, n)$. Odavde sledi jednakost odgovarajućih entropija $H(X_k | X_1, \ldots, X_{k-1}) = H(X_k)$, što zajedno sa Teoremom 3.4.7 daje tvrđenje posledice. □

Primer 3.4.1. Odredimo uslovne entropije za dvodimenzionalnu slučajnu promenljivu, čija je raspodela data u sledećoj tabeli:

		X			
		1	2	3	4
Y	1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$
	2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$
	3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
	4	$\frac{1}{4}$	0	0	0

Najpre računamo marginalne raspodele za X i Y kao zbirove odgovarajućih vrsta odnosno kolona: $p_X = (1/2, 1/4, 1/8, 1/8)$ kao i $p_Y = (1/4, 1/4, 1/4, 1/4)$. Entropije su redom $H(X) = 7/4$ i $H(Y) = 2$.

Sada računamo $p(x|y)$ deljenjem elemenata odgovarajuće vrste tabele sa $p_Y(y)$ i dobijamo:

$$\begin{aligned} H(X|Y) &= p_Y(1)H(X|1) + p_Y(2)H(X|2) + p_Y(3)H(X|3) + p_Y(4)H(X|4) \\ &= \frac{1}{4}H\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right) + \frac{1}{4}H\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{8}, \frac{1}{8}\right) + \frac{1}{4}H\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) \\ &\quad + \frac{1}{4}H(1, 0, 0, 0) = \frac{11}{8} \end{aligned}$$

Na sličan način, računamo $p(y|x)$ deljenjem odgovarajuće kolone sa $p_X(x)$ i dobijamo $H(Y|X) = 13/8$. Vidimo da su vrednosti za $H(X|Y)$ i $H(Y|X)$ različite, ali zato su razlike $H(X) - H(X|Y) = 3/8$ i $H(Y) - H(Y|X) = 3/8$ jednake. Pored toga je i $H(X, Y) = H(X) + H(Y|X) = 27/8$.

Primer 3.4.2. Neka je $S_n = I_{A_1} + I_{A_2} + \ldots + I_{A_n}$ gde su A_k nezavisni događaji takvi da je $P(A_k) = p$ za svako $k = 1, 2, \ldots, n$. U Primeru 3.1.3 izveli smo sledeći izraz za entropiju slučajne promenljive S_n:

$$H(S_n) = -n(p \log_2 p + q \log_2 q) - \sum_{k=0}^{n} \binom{n}{k} \log_2 \binom{n}{k} p^k q^{n-k}.$$

Neka je $I^{(n)} = (I_{A_1}, I_{A_2}, \ldots, I_{A_n})$. Na osnovu Posledice 3.4.8 sledi da je

$$H(I^{(n)}) = H(I_{A_1}) + H(I_{A_2}) + \ldots + H(I_{A_n}) = -n(p \log_2 p + q \log_2 q).$$

Prema tome (na osnovu Teoreme 3.4.3), važi

$$H(I^{(n)}|S_n) = \sum_{k=0}^{n} \binom{n}{k} \log_2 \binom{n}{k} p^k q^{n-k}.$$

Dakle, količinu informacija $H(I^{(n)})$ podelili smo u 2 dela, $H(S_n)$ i $H(I^{(n)}|S_n)$. Neka je $p = q = 0.5$ i $n = 20$. Tada je $H(S_{20}) \approx 3.21$ a $H(I^{(20)}) = 20 \cdot 1 = 20$. Prema tome, približno $3.21/20 \approx 16\%$ informacije o $I^{(20)}$ sadržano je u S_{20}.

3.5 Relativna entropija

U prethodnom odeljku uveli smo pojam entropije kao srednje količine informacija koju nosi jedna realizacija slučajne promenljive X. Definisali smo veličine $H(X|Y)$ i $H(Y|X)$ koje opisuju preostalu količinu informacije za Y koja nije sadržana u X i obratno. Sada ćemo uvesti veličinu koja opisuje zajedničku količinu informacija koje se nalaze u X i Y (uzajamna informacija) kao i relativnu entropiju koja daje neku vrstu "rastojanja" između funkcija raspodele za X i Y.

Relativna entropija $D(p||q)$ ima ulogu "rastojanja" između dve raspodele $p(\cdot)$ i $q(\cdot)$, i predstavlja meru neefikasnosti pretpostavke da je raspodela neke slučajne promenljive $q(\cdot)$, ako je ona $p(\cdot)$. Na primer, ukoliko slučajna promenljiva ima distribuciju $p(\cdot)$, optimalni kod imaće srednju dužinu kodne reči $H(p)$. Međutim, ako pretpostavimo da je raspodela $q(\cdot)$ i konstuišemo optimalni kod za ovu raspodelu, njegova srednja dužina kodne reči (primenjen na promenljivu X) biće $H(p) + D(p||q)$.

Definicija 3.5.1. *Neka su $p(\cdot)$ i $q(\cdot)$ dve raspodele na skupu $\mathcal{X}$ za koje važi da iz $q(x) = 0$ sledi $p(x) = 0$. Relativna entropija ili Kullback-Leibler-ovo rastojanje između ove dve raspodele je*

$$D(p||q) = \sum_{x\in\mathcal{X}} p(x) \log_2 \frac{p(x)}{q(x)}.^4 \tag{3.5}$$

Ako je X slučajna promenljiva sa raspodelom p, onda je

$$D(p||q) = \mathbb{E} \log_2 \frac{p(X)}{q(X)}.$$

Primetimo da $D(p||q)$ nije simetrična funkcija po p i q, kao i da ne zadovoljava nejednakost trougla, pa je (formalno) ne možemo nazvati rastojanje (metrika). I pored toga, korisno je tretirati relativnu entropiju kao neku vrstu "rastojanja" između raspodela.

Teorema 3.5.1. *$D(p||q) \geq 0$. Jednakost važi akko je $p(x) = q(x)$ za svako $x \in \mathcal{X}$.*

Dokaz. Na osnovu Leme 3.3.1 sledi

$$-\sum_{x\in\mathcal{X}} p(x) \log_2 p(x) \leq -\sum_{x\in\mathcal{X}} p(x) \log_2 q(x)$$

[4]Pretpostavljamo da vrednost izraza $0/0$ može biti proizvoljan pozitivan realan broj.

Prebacivanjem sume sa leve na desnu stranu direktno dobijamo tvrđenje leme. □

U nastavku ćemo radi preglednosti (i naglašavanja o kojoj raspodeli se radi) povremeno koristiti oznaku $D(p(x)||q(x))$ umesto $D(p||q)$. Relativna entropija se analogno definiše i za višedimenzionalne raspodele a ovde dajemo izraz u slučaju dvodimenzionalnih raspodela:

$$D(p(x,y)||q(x,y)) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log_2\frac{p(x,y)}{q(x,y)}.$$

Slično se definiše i rastojanje između uslovnih raspodela $p(y|x)$ i $q(y|x)$. Ukoliko fiksiramo element $x \in \mathcal{X}$, tada su $p(\cdot|x)$ i $q(\cdot|x)$ raspodele definisane na skupu $\mathcal{Y}$ i tada je definisana i relativna entropija između njih

$$D(p(\cdot|x)||q(\cdot|x)) = \sum_{y\in\mathcal{Y}} p(y|x)\log_2\frac{p(y|x)}{q(y|x)}$$

Ukoliko usrednjimo relativne entropije $D(p(\cdot|x)||q(\cdot|x))$ za svako $x \in \mathcal{X}$, dobijamo izraz koji predstavlja relativnu entropiju uslovnih raspodela $p(y|x)$ i $q(y|x)$:

$$\begin{aligned} D(p(y|x)||q(y|x)) &= \sum_{x\in\mathcal{X}} p(x)D(p(\cdot|x)||q(\cdot|x)) \\ &= \sum_{x\in\mathcal{X}} p(x)\sum_{y\in\mathcal{Y}} p(y|x)\log_2\frac{p(y|x)}{q(y|x)} \\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log_2\frac{p(y|x)}{q(y|x)} \end{aligned}$$

Prethodni izraz važi i ukoliko su u pitanju višedimenzionalne raspodele.

Takođe važi i odgovarajuće lančano pravilo, dato sledećom teoremom:

Teorema 3.5.2. *Za svake dve diskretne raspodele $p(x,y)$ i $q(x,y)$ važi:*

$$D(p(x,y)||q(x,y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x)).$$

Dokaz. [Cover, p. 25] Direktno iz definicije dobijamo:

$$\begin{aligned} D(p(x,y)||q(x,y)) &= \sum_{\substack{x\in\mathcal{X}\\y\in\mathcal{Y}}} p(x,y)\log_2\frac{p(x,y)}{q(x,y)} = \sum_{\substack{x\in\mathcal{X}\\y\in\mathcal{Y}}} p(x,y)\log_2\frac{p(x)p(y|x)}{q(x)q(y|x)} \\ &= \sum_{\substack{x\in\mathcal{X}\\y\in\mathcal{Y}}} p(x,y)\log_2\frac{p(x)}{q(x)} + \sum_{\substack{x\in\mathcal{X}\\y\in\mathcal{Y}}} p(x,y)\log_2\frac{p(y|x)}{q(y|x)} \\ &= D(p(x)||q(x)) + D(p(y|x)||q(y|x)) \end{aligned}$$

Napomenimo da u prvom sabirku u predposlednjem redu, sumiranje po y marginalizuje $p(x,y)$ u $p(x)$. $\square$

3.6 Uzajamna informacija

Posmatrajmo dve slučajne promenljive X i Y i njihove dve realizacije x i y. Pritom je $\mathcal{I}(y)$ količina informacije koju nosi realizacija y promenljive Y, a $\mathcal{I}(y|x)$ količina informacije koju nosi realizacija y promenljive Y, ako znamo da se prethodno dogodila realizacija x promenljive X. Razlika ove dve vrednosti predstavlja **informaciju o y u x**:

$$\hat{\mathcal{I}}(x,y) = \mathcal{I}(y) - \mathcal{I}(y|x).$$

Pošto je

$$\begin{aligned}\hat{\mathcal{I}}(x,y) &= \mathcal{I}(y) - \mathcal{I}(y|x) = -\log_2 p(y) + \log_2 p(y|x) \\ &= \log_2 \frac{p(x,y)}{p(x)p(y)} = \hat{I}(y,x)\end{aligned}$$

onda se $\hat{\mathcal{I}}(x,y)$ naziva i **uzajamna informacija za x i y**. Ako sada uzmemo srednju vrednost veličine $\hat{\mathcal{I}}(x,y)$ za sve $x \in \mathcal{X}$ i $y \in \mathcal{Y}$ dobijamo **srednju uzajamnu informaciju** za X i Y:

$$I(X,Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\hat{\mathcal{I}}(x,y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log_2\frac{p(x,y)}{p(x)p(y)}.$$

Očigledno je $I(X,Y)$ relativna entropija dvodimenzionalnih raspodela $p(x,y)$ i $p(x)p(y)$, odnosno $I(X,Y) = D(p(x,y)||p(x)p(y))$, odakle sledi da je $I(X,Y) \geq 0$ za svake dve slučajne promenljive X i Y.

Teorema 3.6.1. *Za svake dve slučajne promenljive X i Y važi:*

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

Kao direktnu posledicu prethodnih jednakosti sledi da je $I(X,Y) \geq 0$ i da je $I(X,X) = H(X)$ jer je $H(X|X) = 0$. Definicioni izraz za $I(X,Y)$ kao i prethodna teorema važe i za višedimenzionalne slučajne promenljive.

Uslovna uzajamna informacija za slučajne promenljive X, Y i Z definiše se na sličan način:

$$I(X,Y|Z) = H(X|Z) - H(X|Y,Z)$$

i zadovoljava

$$I(X,Y|Z) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\sum_{z\in\mathcal{Z}} p(x,y,z)\log_2\frac{p(x,y|z)}{p(x|z)p(y|z)}.$$

Promenljive X, Y i Z mogu biti i višedimenzionalne slučajne promenljive. Uslovna uzajamna informacija takođe zadovoljava lančano pravilo:

Teorema 3.6.2. *Za proizvoljnu n-dimenzionalnu slučajnu promenljivu $(X_1, X_2, \ldots, X_n)$, kao i slučajnu promenljivu Y važi:*

$$I((X_1, X_2, \ldots, X_n), Y) = \sum_{k=1}^{n} I(X_k, Y|X_1, X_2, \ldots, X_{k-1}).$$

Dokaz. [Cover, p.24] Na osnovu definicije uzajamne informacije dobijamo

$$I((X_1, X_2, \ldots, X_n), Y) = H(X_1, X_2, \ldots, X_n) - H(X_1, X_2, \ldots, X_n|Y)$$

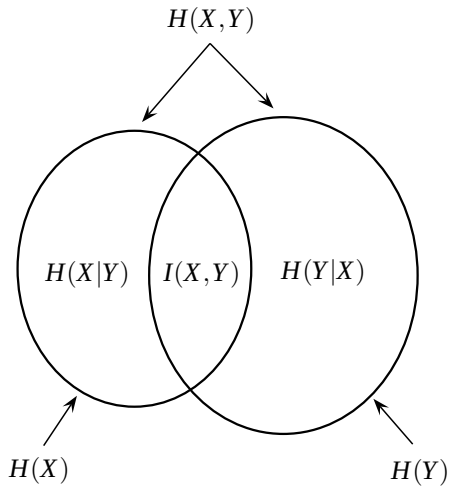
a na osnovu lančanog pravila za entropiju (Teorema 3.4.7) sledi:

$$\begin{aligned} I((X_1, X_2, \ldots, X_n), Y) &= \sum_{k=1}^{n} H(X_k|X_1, \ldots, X_{k-1}) - \sum_{k=1}^{n} H(X_k|X_1, \ldots, X_{k-1}, Y) \\ &= \sum_{k=1}^{n} \Big[H(X_k|\boxed{X_1, \ldots, X_{k-1}}) - H(X_k|\boxed{X_1, \ldots, X_{k-1}}, Y)\Big] \\ &= \sum_{k=1}^{n} I(X_k, Y|\boxed{X_1, \ldots, X_{k-1}}). \end{aligned}$$

Prvi red je direktna primena lančanog pravila, a u trećem primenjujemo izraz za uslovnu uzajamnu informaciju. Pritom, ulogu promenljive Z ima uokvireni niz slučajnih promenljivih $X_1, \ldots, X_{k-1}$, koji predstavlja k-dimenzionalnu slučajnu promenljivu. Ovim je dokaz završen. □

Specijalno, za $n = 2$ važi izraz

$$I((X_1, X_2), Y) = I(X_1, Y) + I(X_2, Y|X_1).$$



Slika 3.4: Odnos između entropije, međusobne informacije i uslovnih entropija.

3.7 Parcijalna uzajamna informacija

Na sličan način možemo definisati i **(parcijalnu) uzajamnu informaciju između promenljive Y i elementa** $x_0 \in \mathcal{X}$ za koji je $p(x_0) > 0$.

$$I(x_0, Y) = \sum_{y \in \mathcal{Y}} p(y|x_0) \log_2 \frac{p(y|x_0)}{p(y)}.$$

Ukoliko je pak $p(x_0) = 0$, možemo uzeti da je $I(x_0, Y) = 0$ po definiciji. Primetimo da je $I(x_0, Y)$ zapravo relativna entropija raspodela $p(y|x_0)$ i $p(y)$, odnosno $I(x_0, Y) = D(p(y|x_0)||p(y))$. Izraz za parcijalnu uzajamnu informaciju $I(X, y_0)$ dobro definisan čak i ako je $p(y) = 0$ za neko y, pošto je tada i $p(y|x_0) = 0$ pa smatramo da je vrednost izraza $0/0$ jednaka 1, kao što je to bio slučaj i kod definicije relativne entropije.

Slično za $y_0 \in \mathcal{Y}$ $(p(y_0) > 0)$ definišemo:

$$I(X, y_0) = \sum_{x \in \mathcal{X}} p(x|y_0) \log \frac{p(x|y_0)}{p(x)} = D(p(x|y_0)||p(x)).$$

Naredna posledica se dokazuje direktno na osnovu definicije uzajamne informacije i svojstva relativne entropije.

Posledica 3.7.1. *Za svako $x_0 \in \mathcal{X}$ i $y_0 \in \mathcal{Y}$ je $I(x_0, Y) \geq 0$ i $I(X, y_0) \geq 0$. Pored toga je:*

$$I(X,Y) = \sum_{x \in \mathcal{X}} p(x) I(x,Y) = \sum_{y \in \mathcal{Y}} p(y) I(X,y).$$

Primer 3.7.1 (Šešelja, primer 2.21., p.47)**.**

3.8 Nejednakost obrade podataka

Često je u praksi potrebno da izvučemo što više informacija o nekoj veličini X pomoću neke druge veličine Y. Nejednakost obrade podataka pokazuje da se količina informacije o X koja je sadržana u Y (tj. $I(X,Y)$) ne može povećati primenom neke transformacije (koja može biti deterministička ili ne) na Y.

Definicija 3.8.1. *Diskretne slučajne promenljive X, Y i Z formiraju **Markovljev lanac** (u oznaci $X \to Y \to Z$), ukoliko uslovna raspodela promenljive Z u odnosu na X i Y, zavisi samo od Y. Drugim rečima, ukoliko važi*

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

Ovo praktično znači da je Y dobijeno na osnovu vrednosti X i nekog slučajnog faktora, dok je Z dobijeno na osnovu Y i nekog drugog slučajnog faktora. Pritom X ne učestvuje direktno u formiranju vrednosti za Z, već samo preko Y.

Neka je $X \to Y \to Z$. Tada važi

$$\begin{aligned} p(x,y,z) &= p(x,y)p(z|x,y) \\ &= p(x)p(y|x)p(z|y). \end{aligned}$$

Slučajne promenljive X i Z su uslovno nezavisne od Y, što sledi iz

$$p(x,z|y) = \frac{p(x,y,z)}{p(y)} = \frac{p(x,y)p(z|y)}{p(y)} = p(x|y)p(z|y).$$

Teorema 3.8.1. ***(Nejednakost obrade podataka)*** *Ako je $X \to Y \to Z$, onda je $I(X,Y) \geq I(X,Z)$ i $I(X,Y) \geq I(X,Y|Z)$.*

Dokaz. [Cover, p.34] Na osnovu lančanog pravila za uzajamnu informaciju (Teorema 3.6.2) sledi

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

Pošto su X i Z uslovno nezavisne od Y, sledi da je $I(X,Z|Y)=0$ pa je

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

Iz prethodnog izraza slede oba dela teoreme. □

Specijalan slučaj Markovljevog lanca je $X\to Y\to g(Y)$, gde je g funkcija definisana na skupu Z. Na osnovu Teoreme 3.8.1 sledi $I(X,Y)\geq I(X,g(Y))$, tj. da se determinističkom obradom vrednosti Y ne može uvećati uzajamna informacija sa X.

Nejednakost $I(X,Y)\geq I(X,Y|Z)$ može da se objasni na sledeći način: Veličina Z može sadržati neke informacije o X, ali je taj deo ujedno sadržan i u Y. Znajući Z, mi ujedno znamo i deo uzajamnih informacija X i Y, pa se ta količina informacija smanjuje ili ostaje ista.

Ukoliko X, Y i Z ne formiraju Markovljev lanac, tada je moguće da je $I(X,Y|Z)>I(X,Y)$.

Primer 3.8.1. Neka su X i Y nezavisne slučajne promenljive sa raspodelom:

$$X,Y:\begin{pmatrix}0 & 1\\ 1/2 & 1/2\end{pmatrix}$$

i neka je $Z=X+Y$. Tada je $I(X,Y)=0$ dok je

$$I(X,Y|Z)=H(X|Z)-H(X|Y,Z).$$

Pritom, pošto je $X=Z-Y$, to je $H(X|Y,Z)=0$ (X je potpuno određena sa Y i Z). Uz to je i

$$H(X|Z)=p_Z(0)H(X|Z=0)+p_Z(1)H(X|Z=1)+p_Z(2)H(X|Z=2)$$

Ukoliko je $Z=0$ ili $Z=2$, onda je sigurno $X=0$ odnosno $X=1$, pa su odgovarajuće uslovne entropije $H(X|Z=0)$ i $H(X|Z=2)$ jednake 0. Primetimo da je

$$p_Z(1)=P(Z=1)=P(X=1,Y=0)+P(X=0,Y=1)=\frac{1}{4}+\frac{1}{4}=\frac{1}{2}$$

kao i

$$p_{X|Z}(0,1)=P(X=0|X+Y=1)=P(X=0|Y=1)=P(X=0)=1/2$$
$$p_{X|Z}(1,1)=P(X=1|X+Y=1)=P(X=1|Y=0)=P(X=1)=1/2$$

Odavde lako računamo preostalu uslovnu entropiju:

$$H(X|Z=1) = -p_{X|Z}(0,1)\log_2 p_{X|Z}(0,1) - p_{X|Z}(1,1)\log_2 p_{X|Z}(1,1) = 1$$

pa je

$$I(X,Y|Z) = H(X|Z) = p_Z(1)H(X|Z=1) = 1/2.$$

Ovim smo dokazali da je $I(X,Y|Z) > I(X,Y)$.

Glava 4

Diskretni izvori informacija

4.1 Definicija i vrste izvora informacija

Definicija 4.1.1. ***Diskretni izvor informacija*** *X je niz diskretnih slučajnih promenljivih $X = (X_n)_{n\in\mathbb{N}}$ definisanih na istom prostoru verovatnoća.*

Pretpostavićemo da je $\mathcal{X}_k = \mathcal{X}_j = \mathcal{X}$ za svako $k, j \in \mathbb{N}$. Drugim rečima, skup vrednosti koje uzimaju slučajne promenljive $X_1, X_2, \ldots$ je isti i ne zavisi od konkretne promenljive X_k. Slučajna promenljiva X_k označava vrednost koju izvor emituje u k-tom vremenskom trenutku. Slučajni niz je okaraktersian nizom raspodela

$$(p(x_1, x_2, \ldots, x_n))_{n\in\mathbb{N}} = (p_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n))_{n\in\mathbb{N}}\,. \tag{4.1}$$

I ovde koristimo skraćenu oznaku u koliko je jasno koje slučajne promenljive treba da stoje u indeksu. Ovaj niz je dovoljan, s obzirom da se proizvoljna raspodela slučajnih promenljivih $X_{i_1}, X_{i_2}, \ldots, X_{i_n}$ dobija marginalizacijom odgovarajuće raspodele za $X_1, X_2, \ldots, X_{i_n}$ (pod pretpostavkom da je $i_1 < i_2 < \cdots < i_n$). Marginalizacija predstavlja sumiranje $p(x_1, x_2, \ldots, x_{i_n})$ po svim $x_j \in \mathcal{X}$ za koje je $j \neq i_k$. U opštem slučaju (kada X_n ne moraju biti diskretne) umesto funkcija $p_{X_1,X_2,\ldots,X_n}$ posmatraju se funkcije raspodele $F_n = F_{X_1,X_2,\ldots,X_n}$ za koje važi naredna teorema.

Teorema 4.1.1. [SJ, p.113] *Niz funkcija raspodela F_n zadovoljava sledeća svojstva:*

(F_1^∞) *F_n je monotono neopadajuća po svim argumentima.*

(F_2^∞) *$F_n(x_1, \ldots, -\infty, \ldots, x_n) = 0$, $F_n(+\infty, \ldots, +\infty) = 1$,*

(F_3^∞) *F_n je neprekidna sa leve strane, po svakom argumentu,*

(F_4^∞) $F_{n+1}(x_1, \ldots, x_n, +\infty) = F_n(x_1, x_2, \ldots, x_n)$.

Važi i obrat, tj. ukoliko niz funkcija $(F_n)_{n\in\mathbb{N}}$ zadovoljava uslove prethodne teoreme, onda postoji niz slučajnih promenljivih $(X_n)_{n\in\mathbb{N}}$ čije su F_n funkcije raspodele (Teorema Daniell-Kolmogorov).

Definicija 4.1.2. *Diskretni izvor je:*

1. ***stacionaran**, ako je $p_{X_1,\ldots,X_n} = p_{X_{k+1},\ldots,X_{k+n}}$ za svako $k, n \in \mathbb{N}$.*
2. ***bez memorije**, ako je $p_{X_n|X_1,\ldots,X_{n-1}}(x_n|x_1, \ldots, x_{n-1}) = p_{X_n}(x_n)$, za svako $x_k \in \mathcal{X}$ $(k = 1, 2, \ldots, n)$ i $n \in \mathbb{N}$. Tada je i $p(x_{i_1}, x_{i_2}, \ldots, x_{i_n}) = p(x_{i_1})p(x_{i_2}) \cdots p(x_{i_n})$.*
3. ***Markovljev**, ako je $p_{X_n|X_1,\ldots,X_{n-1}}(x_n|x_1, \ldots, x_{n-1}) = p_{X_n|X_{n-1}}(x_n|x_{n-1})$, za svako $x_k \in \mathcal{X}$ $(k = 1, 2, \ldots, n)$ i $n \in \mathbb{N}$.*

Prethodne definicije važe i za kontinualne izvore ako se funkcija $p_\bullet$ posmatraju funkcije $F_\bullet$.

4.2 Entropija izvora informacija

Neka je $H_n = H(X_1, X_2, \ldots, X_n)$. Pošto je H_n srednja količina informacije koju nose prvih n simbola koje emituje izvor, onda je srednja količina informacije po simbolu jednaka H_n/n.

Definicija 4.2.1. *Entropija izvora $X = (X_n)_{n\in\mathbb{N}}$ definisana je sa*

$$H(X) = \lim_{n\to+\infty} \frac{H_n}{n}.$$

Ako je X bez memorije i ako sve slučajne promenljive X_k imaju istu raspodelu, onda je $H_n = nH(X_1)$, pa je i $H(X) = H(X_1)$.

Teorema 4.2.1. *Ako je X stacionarni izvor, tada on ima konačnu entropiju.*

Pre nego što damo dokaz ove teoreme, formulisaćemo alternativnu definiciju entropije izvora X:

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

Ovako definisana entropija intuitivno predstavlja srednju dodatnu količinu informacije koju donosi nosi svaka naredna slučajna promenljiva X_k u nizu (odnosno svaki naredni simbol koji emituje izvor), u odnosu na sve prethodne.

Teorema 4.2.2. *Ako je X stacionarni izvor, tada postoji $H'(X)$.*

Dokaz. Neka je $H'_n = H(X_n|X_1, \ldots, X_{n-1})$. Tada je

$$\begin{aligned} H'_{n+1} &= H(X_{n+1}|X_1, X_2, \ldots, X_n) \leq H(X_{n+1}|X_2, \ldots, X_n) \\ &= H(X_n|X_1, \ldots, X_{n-1}) = H'_n. \end{aligned}$$

Prva nejednakost sledi iz svojstva entropije $H(X|Y,Z) \leq H(X|Z)$ a druga iz stacionarnosti izvora, s obzirom da je $p_{X_2,\ldots,X_{n+1}} = p_{X_1,\ldots,X_n}$ i $p_{X_{n+1}|X_2,\ldots,X_n} = p_{X_n|X_1,\ldots,X_{n-1}}$. Drugim rečima, niz $(H'_n)_{n\in\mathbb{N}}$ je opadajući, a pošto je odozdo ograničen nulom, sledi da je konvergentan. □

Pre nego što pokažemo da su entropije $H(X)$ i $H'(X)$ jednake u slučaju stacionarnih izvora, dokažimo sledeću pomoćnu lemu:

Lema 4.2.3. *Ako je niz $(a_n)_{n\in\mathbb{N}}$ konvergentan i ima graničnu vrednost a, tada je i niz aritmetičkih sredina $b_n = (a_1 + a_2 + \ldots + a_n)/n$ konvergentan i ima istu graničnu vrednost.*

Dokaz. Iz definicije granične vrednosti, sledi da za proizvoljno $\epsilon > 0$ postoji n_ϵ takvo da je $|a_n - a| < \epsilon$ za svako $n > n_\epsilon$. Posmatrajmo apsolutnu vrednost razlike:

$$\begin{aligned} |b_n - a| &= \left| \frac{a_1 + a_2 + \ldots + a_n}{n} - a \right| = \left| \frac{a_1 - a + a_2 - a + \ldots + a_n - a}{n} \right| \\ &\leq \frac{|a_1 - a| + |a_2 - a| + \ldots + |a_n - a|}{n} \end{aligned}$$

Podelimo sumu elemenata u brojiocu razlomka na dva dela dobijamo:

$$\begin{aligned} |b_n - a| &\leq \frac{|a_1 - a| + |a_2 - a| + \ldots + |a_{n_\epsilon} - a|}{n} + \frac{|a_{n_\epsilon+1} - a| + \ldots + |a_n - a|}{n} \\ &\leq \frac{A}{n} + \frac{n - n_\epsilon}{n}\epsilon < 2\epsilon \end{aligned}$$

za dovoljno veliko n, odnosno za $n > \max\{n_\epsilon, A/\epsilon\}$ gde je

$$A = |a_1 - a| + |a_2 - a| + \ldots + |a_{n_\epsilon} - a|$$

konstanta koja ne zavisi od n. Ovim smo dokazali da je i niz $(b_n)_{n\in\mathbb{N}}$ konvergentan i da ima istu graničnu vrednost kao $(a_n)_{n\in\mathbb{N}}$. □

Teorema 4.2.4. *Ako je X diskretni izvor informacija takav da je definisana entropija $H'(X)$, tada je definisana i entropija $H(X)$ i važi $H(X) = H'(X)$.*

Dokaz. Na osnovu lančanog pravila:

$$H_n = H(X_1, X_2, \ldots, X_n) = \sum_{k=1}^{n} H(X_k|X_1, \ldots, X_{k-1}) = \sum_{k=1}^{n} H'_k$$

i prethodne leme sledi:

$$\frac{H_n}{n} = \frac{H'_1 + H'_2 + \ldots + H'_n}{n} \to H'(X), \qquad n \to +\infty.$$

Drugim rečima, postoji $H(X)$ i jednako je $H'(X)$. □

Iz prethodne teoreme, i Teoreme 4.2.2 direktno dobijamo da za svaki stacionarni izvor X postoji entropija $H(X)$, odnosno Teoremu 4.2.1.

Primer 4.2.1. U okviru Teoreme 4.2.4 pokazali smo da ukoliko postoji (je definisana) entropija $H'(X)$, onda postoji i $H(X)$ i važi jednakost $H(X) = H'(X)$. Obrat ovog tvrđenja ne važi, tj. iz postojanja entropije $H(X)$ ne mora da sledi postojanje entropije $H'(X)$.

Da bi to pokazali, posmatrajmo sledeci primer. Neka je niz $(X_n)_{n\in\mathbb{N}}$ definisan sa:

$$X_{2k-1} : \begin{pmatrix} 0 & 1 \\ 1/2 & 1/2 \end{pmatrix}, \quad X_{2k} = X_{2k-1}, \qquad k \in \mathbb{N}.$$

i neka su pritom neparni elementi niza međusobno nezavisni. Drugim rečima, posmatramo sledeći niz:

$$X_1,\ X_1,\ X_3,\ X_3,\ X_5,\ X_5,\ \ldots$$

gde su $X_1, X_3, \ldots$ nezavisna bacanja savršenog novčića. Najpre računamo entropiju $H_n = H(X_1, X_2, \ldots, X_n)$. Očigledno je ova entropija jednaka entropiji neparnih članova niza, pošto su parni jednaki neparnima i ne donose nikakvu informaciju:

$$\begin{aligned} H_{2k-1} &= H(X_1, X_2, \ldots, X_{2k-1}) \\ &= \overbrace{H(X_2, X_4, \ldots, X_{2k-2}|X_1, X_3, \ldots, X_{2k-1})}^{=0} + H(X_1, X_3, \ldots, X_{2k-1}) \\ &= H(X_1, X_3, \ldots, X_{2k-1}) = H(X_1) + H(X_3) + \ldots + H(X_{2k-1}) = k. \end{aligned}$$

Na sličan način je i

$$H_{2k} = H(X_1, X_2, \ldots, X_{2k}) = H(X_1, X_3, \ldots, X_{2k-1}) = k$$

pa je

$$\frac{H_n}{n} = \begin{cases} \dfrac{k}{2k} & n = 2k \\ \dfrac{k}{2k-1} & n = 2k-1 \end{cases} \to \frac{1}{2} = H(X), \qquad (n \to +\infty)$$

s obzirom da obe grane funkcije teže ka $1/2$. Izračunajmo sada alternativnu entropiju. Pošto su $X_1, X_3, \ldots, X_{2k-1}$ nezavisne promenljive, onda je:

$$H'_{2k-1} = H(X_{2k-1}|X_1, \ldots, X_{2k-2}) = H(X_{2k-1}) = 1$$

a pošto je $X_{2k} = X_{2k-1}$ onda je

$$H'_{2k} = H(X_{2k}|X_1, \ldots, X_{2k-1}) = H(X_{2k-1}|X_1, \ldots, X_{2k-1}) = 0.$$

Prema tome, niz $(H'_n)_{n\in\mathbb{N}}$ jednak je $(1, 0, 1, 0, \ldots)$, odnosno nije konvergentan. Prema tome, entropija $H(X)$ postoji i jednaka je $1/2$, dok entropija $H'(X)$ ne postoji.

Naravno, izvor $X = (X_n)_{n\in\mathbb{N}}$ **nije stacionaran**, s obzirom na da je

$$p_{X_1,X_2}(t,s) = \begin{cases} 1/2 & t = s \\ 0 & t \neq s \end{cases}, \qquad p_{X_2,X_3}(t,s) = p_{X_2}(t)p_{X_3}(s) = 1/4$$

pa ne važi da je $p_{X_1,X_2} = p_{X_2,X_3}$.

Na osnovu prethodnog primera zaključujemo da je postojanje entropije $H'(X)$ stroži uslov od postojanja entropije $H(X)$. Međutim, alternativna definicija entropije $H'(X)$ nema mnogo smisla ukoliko izvor nije stacionaran. Naime, izvor razmatran u prethodnom primeru, zaista "*u proseku*" daje $1/2$ bita po simbolu, s obzirom da je svaki simbol dupliran. Entropija $H(X)$ nam upravo daje srednji broj bita (odnosno srednju količinu informacija) po simbolu koji emituje izvor.

Primer 4.2.2. Posmatrajmo tekst napisan na engleskom (ili bilo kom drugom jeziku) kao izvor informacija. Radi jednostavnosti, pretpostavimo da nema razlike između velikih i malih slova kao i da nema interpunkcijskih znakova. Ukupno imamo 26 slova i razmak, što čini ukupno 27 simbola. Posmatrajmo sledeća dva slučaja:

- Tekst je slučajno generisan (npr. nasumičnim kucanjem po tastaturi). Tada se svaki naredni simbol generiše nezavisno od prethodnih i to sa

podjednakim verovatnoćama, pa je entropija izvora $H(X) = \log_2 27 \approx 4.7$.

- Ukoliko posmatramo smislen tekst (knjiga, novine, web-sajt, itd.) onda simboli nisu nezavisni niti su uniformno raspodeljeni. Shannon je procenio da je entropija u ovom slučaju jednaka $H(X) \approx 1.3$. Da bi ovo ilustrovali, posmatrajmo "nastanak" reči `INFORMATION`:

  `I; IN; INF; INFO; INFOR; INFORM; INFORMA; ... INFORMATION`

  Očigledno je da je neizvesnost sve manja svakim narednim slovom, pa je zato i entropija primetno manja nego u prvom slučaju.

4.3 Markovljevi izvori

Neka je X (diskretan) Markovljev izvor. Tada, po definiciji, uslovna raspodela $p(x_n|x_1, \ldots, x_{n-1})$ jednaka je $p(x_n|x_{n-1})$. Drugim rečima, vrednost slučajne promenljive X_n direktno zavisi samo od vrednosti promenljive X_{n-1} ali ne i od prethodnih. Zavisnost promenljive X_n od $X_1, X_2, \ldots, X_{n-2}$ je posredna, preko promenljive X_n.

Lema 4.3.1. *Markovljev izvor* $(X_n)_{n\in\mathbb{N}}$ *poseduje sledeća svojstva:*

1. $p(x_1, x_2, \ldots, x_n) = p(x_n|x_{n-1})p(x_{n-1}|x_{n-2}) \cdots p(x_2|x_1)p(x_1)$, *za svako* $n \in \mathbb{N}$.

2. $p(x_k, x_{k+1}, \ldots, x_{k+n}) = p(x_{k+n}|x_{k+n-1}) \cdots p(x_{k+1}|x_k)p(x_k)$ *za svako* $k, n \in \mathbb{N}$.

3. $p(x_n|x_{i_1}, \ldots, x_{i_k}) = p(x_n|x_{i_k})$ *za svako* $1 \leq i_1 < i_2 < \cdots < i_k$.

Dokaz. Svojstvo **1** sledi direktno iz definicionog svojstva Markovljevih izvora:

$$p(x_1, x_2, \ldots, x_n) = p(x_n|x_1, \ldots, x_{n-1})p(x_{n-1}|x_1, \ldots, x_{n-2}) \cdots p(x_2|x_1)p(x_1)$$
$$p(x_n|x_{n-1})p(x_{n-1}|x_{n-2}) \cdots p(x_2|x_1)p(x_1).$$

Analogno se dokazuje i svojsvto **2**. Svojstvo **3** pokazujemo matematičkom indukcijom po n. Za $n = 2$ jedina moguća raspodela ovog tipa je $p(x_2|x_1)$ za koju tvrđenje trivijalno važi. Pretpostavimo da tvrđenje važi za sve brojeve manje od n. Razmatramo 2 slučaja.

Neka je najpre $i_k = n-1$. Neka su $j_1 < j_2 < \cdots < j_{n-k-1}$ indeksi koji nisu među $i_1, i_2, \ldots, i_k$, odnosno

$$\{j_1, j_2, \ldots, j_{n-1-k}\} = \{1, 2, \ldots, n\} \setminus \{i_1, i_2, \ldots, i_k\}.$$

Tada je:

$$\begin{aligned}
&p(x_n|x_{i_1}, \ldots, x_{i_k}) \\
&= \sum_{x_{j_1},\ldots,x_{j_{n-k-1}} \in \mathcal{X}} p(x_n|x_{i_1}, \ldots, x_{i_k}, x_{j_1}, \ldots, x_{j_{n-k-1}}) p(x_{j_1}, \ldots, x_{j_{n-k-1}}|x_{i_1}, \ldots, x_{i_k}) \\
&= \sum_{x_{j_1},\ldots,x_{j_{n-k-1}} \in \mathcal{X}} p(x_n|x_1, \ldots, x_{n-1}) p(x_{j_1}, \ldots, x_{j_{n-k-1}}|x_{i_1}, \ldots, x_{i_k}) \\
&= \sum_{x_{j_1},\ldots,x_{j_{n-k-1}} \in \mathcal{X}} p(x_n|x_{n-1}) p(x_{j_1}, \ldots, x_{j_{n-k-1}}|x_{i_1}, \ldots, x_{i_k}) \\
&= p(x_n|x_{n-1}) \sum_{x_{j_1},\ldots,x_{j_{n-k-1}} \in \mathcal{X}} p(x_{j_1}, \ldots, x_{j_{n-k-1}}|x_{i_1}, \ldots, x_{i_k}) = p(x_n|x_{n-1})
\end{aligned}$$

Pritom, u poslednjem redu mogli smo da izvučemo $p(x_n|x_{n-1})$ ispred sume, s obzirom da je $i_{k-1} = n-1$ pa samim tim nijedan j_l (odnosno j_{n-k-1}, s obzirom da je maksimalan) nije jednak $n-1$.

Pretpostavimo sada da je $i_k < n-1$. Tada je

$$p(x_n|x_{i_1}, \ldots, x_{i_k}) = \sum_{x_{n-1} \in \mathcal{X}} p(x_n|x_{i_1}, \ldots, x_{i_k}, x_{n-1}) p(x_{n-1}|x_{i_1}, \ldots, x_{i_k})$$

Na osnovu prvog dela dokaza je $p(x_n|x_{i_1}, \ldots, x_{i_k}, x_{n-1}) = p(x_n|x_{n-1})$ a na osnovu indukcijske hipoteze je $p(x_{n-1}|x_{i_1}, \ldots, x_{i_k}) = p(x_{n-1}|x_{i_k})$ pa zamenom dobijamo

$$p(x_n|x_{i_1}, \ldots, x_{i_k}) = \sum_{x_{n-1} \in \mathcal{X}} p(x_n|x_{n-1}) p(x_{n-1}|x_{i_k}) = p(x_n|x_{i_k}).$$

Ovim je dokaz svojstva **3** završen. □

Iz prethodne leme sledi da je Markovljev izvor u potpunosti određen nizom raspodela $p(x_{n+1}|x_n)$ kao i raspodelom prvog elementa niza $p(x_1)$.

Neka je $\mathcal{X} = \{a_1, a_2, \ldots, a_m\}$. Tada je raspodela $p(x_{n+1}|x_n)$ (odnosno $p_{X_{n+1}|X_n}$) definisana matricom

$$\mathbf{\Pi}(n) = [p_{X_{n+1}|X_n}(a_j|a_i)]_{1 \le i,j \le m} = [P(X_{n+1} = a_j|X_n = a_i)]_{1 \le i,j \le m}.$$

Primetimo da je zbir elemenata u svakoj vrsti matrice $\mathbf{\Pi}(n)$ jednak

$$\sum_{j=1}^{m} p_{X_{n+1}|X_n}(a_j|a_i) = 1,$$

odnosno da je $\mathbf{\Pi}(n)$ **stohastička matrica** [1].

Markovljeve izvore bi trebalo posmatrati na sledeći način. Skup $\mathcal{X}$ nazivamo **skupom stanja**, a njegove elemente **stanjima**. Slučajna promenljiva X_n opisuje **stanje izvora u n-tom trenutku** dok je matrica $\mathbf{\Pi}(n)$ **tranziciona matrica** (**matrica prelaza**).

Ako je sistem u n-tom trenutku bio u nekom stanju a_i, tada verovatnoća da će u $n+1$-vom trenutku biti u stanju a_j zavisi **isključivo od stanja** a_i i jednaka je $P(X_{n+1} = a_j|X_n = a_i)$. Pritom važi:

$$p_{X_{n+1}}(a_j) = \sum_{i=1}^{m} p_{X_{n+1}|X_n}(a_j|a_i)p_{X_n}(a_i). \tag{4.2}$$

Dakle sistem "skakuće" iz stanja u stanje, pri čemu verovatnoće prelaza zavise isključivo od stanja u kome se trenutno nalazi.

Ako sa $\mathbf{p}_n$ označimo vektor

$$\mathbf{p}_n = [p_{X_n}(a_1)\ p_{X_n}(a_2)\ \cdots\ p_{X_n}(a_m)],$$

onda izraz (4.2) postaje

$$\mathbf{p}_{n+1} = \mathbf{p}_n\mathbf{\Pi}(n) = \mathbf{p}_1\mathbf{\Pi}(1)\mathbf{\Pi}(2)\cdots\mathbf{\Pi}(n). \tag{4.3}$$

4.3.1 Vremenski nezavisni Markovljevi izvori

Markovljev izvor je **vremenski nezavisan** (ili **homogen**) ukoliko su sve matrice $\mathbf{\Pi}(n)$ jednake, tj. ne zavise od trenutka (indeksa) n. Tada umesto $\mathbf{\Pi}(n)$ koristimo oznaku $\mathbf{\Pi}$ a umesto $P(X_{n+1} = a_j|X_n = a_i) = p_{X_{n+1}|X_n}(a_j|a_i)$ pišemo jednostavno $p_{ij} = p(a_j|a_i)$. Izraz (4.3) postaje

$$\mathbf{p}_{n+1} = \mathbf{p}_1\mathbf{\Pi}^n.$$

Nadalje ćemo podrazumevati (ukoliko drugačije ne bude bilo naglašeno) da je Markovljev izvor X vremenski nezavisan.

[1] Ukoliko isto važi i za kolone, onda je matrica **dvostruko stohastička**.

Primetimo da iz $\mathbf{p}_{t+n} = \mathbf{p}_1\mathbf{\Pi}^{t+n-1}$ i $\mathbf{p}_t = \mathbf{p}_1\mathbf{\Pi}^{t-1}$ sledi $\mathbf{p}_{t+n} = \mathbf{p}_t\mathbf{\Pi}^n$, odnosno da verovatnoća

$$P(X_{t+n} = a_j | X_t = a_i) = [\mathbf{\Pi}^n]_{ij}$$

ne zavisi od vrednosti t. Ukoliko označimo $\mathbf{\Pi}^n = [p_{ij}(n)]$, lako dobijamo da važi

$$p_{ij}(n) = \sum_{k=1}^{m} p_{ik}(s)p_{kj}(n-s).$$

Prethodne jednačine su poznate kao **jednačine Chapman-Kolmogorova**.

Stanja vremenski nezavisnog Markovljevog izvora možemo podeliti u sledeće kategorije:

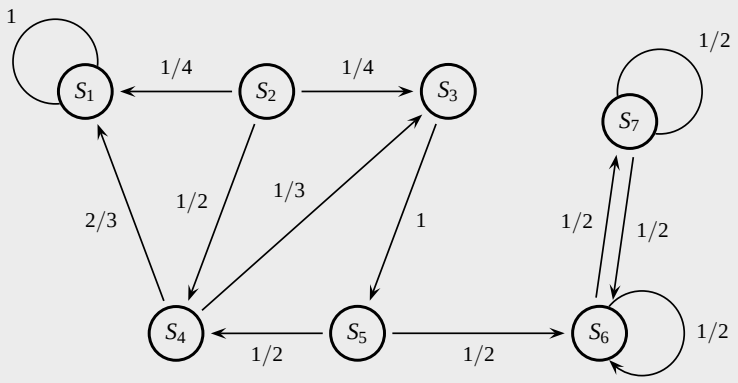
1. **Povratna (esencijalna, bitna, rekurentna) stanja.** Stanje a_i je povratno, ako se izvor sa verovatnoćom 1 vraća u to stanje. U suprotnom, to je **tranzijentno** (**nepovratno**) stanje.

2. **Periodična stanja.** Stanje a_i je periodično, ako postoji ceo broj $d > 1$ takav da se (posle napuštanja stanja a_i) izvor može vratiti u to stanje (sa verovatnoćom većom od 0) samo nakon broja koraka n koji je deljiv sa d. Drugim rečima, potrebno je da postoji ceo broj $d > 1$ takav da iz $p_{ii}(n) > 0$ sledi da d deli n (za svako $n \in \mathbb{N}$).

3. **Apsorbujuća stanja.** Stanje a_i je apsorbujuće, ako izvor, po dolasku u to stanje, ne može više da ga napusti. Drugim rečima, ako je $p_{ii} = 1$.

Teorema 4.3.2. *Stanje a_i je povratno akko je red $\sum_{n=0}^{+\infty} p_{ii}(n)$ divergentan.*

Primer 4.3.1. Markovljev izvor se često prikazuje u obliku grafa na slici 4.1. Stanja $S_1, S_2, \ldots, S_7$ predstavljena su čvorovima grafa, a ivice predstavljaju verovatnoće prelaza, različite od 0.



Slika 4.1: Dijagram stanja jednog Markovljevog izvora.

Vidimo sa slike da su tranzijentna stanja S_2, S_3, S_4 i S_5, rekurentna S_1, S_6 i S_7, periodična S_3, S_4 i S_5, dok je stanje S_1 apsorbujuće.

4.3.2 Stacionarni Markovljevi izvori

Podsetimo se da za stacionarni izvor informacija važi sledeća jednakost

$$p_{X_1,\ldots,X_n} = p_{X_{k+1},\ldots,X_{k+n}}$$

za svako $k, n \in \mathbb{N}$. Sledeće uslove, koje zadovoljava svaki stacionarni izvor informacija dobijamo kao direktnu posledicu ove jednakosti:

1. $p_{X_{k+1}|X_k} = p_{X_2|X_1}$;

2. $p_{X_k} = p_{X_1}$, odnosno da svi elementi niza imaju istu raspodelu. Specijalno je $p_{X_2} = p_{X_1}$.

Primetimo da u slučaju Markovljevog izvora, svojstvo **1** predstavlja vremensku nezavisnost. U tom slučaju, svojstvo **2** ekvivalentno je izrazu $\mathbf{p}_k = \mathbf{p}_1$ odnosno $\mathbf{p}_1\mathbf{\Pi}^{k-1} = \mathbf{p}_1$. Specijalno, za $k = 2$ dobijamo $\mathbf{p}_1\mathbf{\Pi} = \mathbf{p}_1$. Pritom ovde važi i obrnuto tvrđenje, ako je $\mathbf{p}_1$ takva da je $\mathbf{p}_1\mathbf{\Pi} = \mathbf{p}_1$, onda je i $\mathbf{p}_1\mathbf{\Pi}^{k-1} = \mathbf{p}_1$ odnosno $\mathbf{p}_k = \mathbf{p}_1$. Ovaj uslov je, uz uslov vremenske nezavisnosti, dovoljan uslov stacionarnosti Markovljevog izvora. O tome govori sledeća teorema.

Teorema 4.3.3. *Markovljev izvor X je stacionaran ako i samo ako je vremenski nezavisan i vektor raspodele $\mathbf{p}_1$ prvog elementa niza X_1 zadovoljava svojstvo $\mathbf{p}_1\mathbf{\Pi} = \mathbf{p}_1$.*

Dokaz. Već smo pokazali da su ovi uslovi potrebni, odnosno da ih zadovoljava svaki stacionarni izvor informacija, pa samim tim i stacionarni Markovljev izvor.

Pokazujemo da su uslovi dovoljni. Neka je $(X_n)_{n\in\mathbb{N}}$ Markovljev izvor za koji važi da je vremenski nezavisan i da je $\mathbf{p}_1\mathbf{\Pi} = \mathbf{p}_1$. Matematičkom indukcijom dokazujemo da je $\mathbf{p}_n = \mathbf{p}_1$ za svako $n \in \mathbb{N}$. Za $n = 1$ ova nejednakost je trivijalna, a ako pretpostavimo da važi za $n-1$ dobijamo:

$$\mathbf{p}_n = \mathbf{p}_1\mathbf{\Pi}^{n-1} = \mathbf{p}_1\mathbf{\Pi}^{n-2}\mathbf{\Pi} = \mathbf{p}_{n-1}\mathbf{\Pi} = \mathbf{p}_1\mathbf{\Pi} = \mathbf{p}_1$$

čime je dokaz indukcijom završen. Samim tim je $p_{X_1}(t) = p_{X_n}(t) = p(t)$ za svako $n \in \mathbb{N}$ i $t \in \mathcal{X}$. Iz svojstva vremenske nezavisnosti dobijamo da je $p_{X_{n+1}|X_n}(t,s) = p_{X_2|X_1}(t,s) = p(t,s)$ za svako $n \in \mathbb{N}$ i $t, s \in \mathcal{X}$. Dalje sledi

$$\begin{aligned}
&p_{X_{k+1},X_{k+2},\ldots,X_{k+n}}(t_1,t_2,\ldots,t_n)\\
&\qquad = p_{X_{k+n}|X_{k+n-1}}(t_n|t_{n-1})\cdots p_{X_{k+1}|X_k}(t_{k+1}|t_k)p_{X_k}(t_k)\\
&\qquad = p(t_n|t_{n-1})\cdots p(t_2|t_1)p(t_1)\\
&\qquad = p_{X_n|X_{n-1}}(t_n|t_{n-1})\cdots p_{X_2|X_1}(t_2|t_1)p_{X_1}(t_1)\\
&\qquad = p_{X_1,X_2,\ldots,X_n}(t_1,t_2,\ldots,t_n)
\end{aligned}$$

za svako $t_1, t_2, \ldots, t_n \in \mathcal{X}$ i $k, n \in \mathbb{N}$. Ovim smo dokazali da je $p_{X_1,\ldots,X_n} = p_{X_{k+1},\ldots,X_{k+n}}$ odnosno da je Markovljev izvor X stacionaran. □

Raspodelu p_{X_1} za koju je $\mathbf{p}_1\mathbf{\Pi} = \mathbf{p}_1$ nazivamo **stacionarnom raspodelom** vremenski nezavisnog Markovljevog izvora X. U nastavku ćemo vektor $\mathbf{p}_1$ ove raspodele označavati sa $\boldsymbol{\mu} = [\mu_1\ \mu_2\ \cdots\ \mu_m]$. Pošto je u pitanju raspodela verovatnoće, zbir komponenti vektora $\boldsymbol{\mu}$ mora biti jednak 1. Samim tim, $\boldsymbol{\mu}$ predstavlja jedinstveno rešenje sistema linearnih jednačina

$$\begin{aligned}
(p_{11}-1)\mu_1 + \quad p_{21}\mu_2 + \quad &\ldots \quad + \quad p_{m1}\mu_m = 0\\
p_{12}\mu_1 + (p_{22}-1)\mu_2 + \quad &\ldots \quad + \quad p_{m2}\mu_m = 0\\
&\qquad\qquad\qquad\qquad\qquad\qquad\ \vdots\\
p_{1m}\mu_1 + \quad p_{2m}\mu_2 + \quad &\ldots \quad + (p_{mm}-1)\mu_m = 0\\
\mu_1 + \quad \mu_2 + \quad &\ldots \quad + \quad \mu_m = 1
\end{aligned}$$

gde je $\mathbf{\Pi} = [p_{ij}]_{1\le i,j\le m}$. Primetimo da ovaj sistem ima $n+1$ jednačina, ali su samo n od njih nezavisne.

Pored toga, u određenim slučajevima, stacionarna raspodela može da se dobije i na drugi način, o čemu govori naredna teorema.

Teorema 4.3.4. ***(Ergodična teorema za homogene lance Markova sa konačno mnogo stanja)*** *Ako je za neko $n_0 \in \mathbb{N}$, svaki element matrice prelaza $\mathbf{\Pi}^{n_0}$ strogo pozitivan (tj. $p_{ij}(n_0) > 0$ za svako $1 \leq i, j \leq m$, tada za svako $j = 1, 2, \ldots, m$) važi*

$$\lim_{n \to +\infty} p_{ij}(n) = \mu_j$$

gde je $\mathbf{p}_1 = \boldsymbol{\mu} = [\mu_1 \ \mu_2 \ \cdots \ \mu_m]$ stacionarna raspodela.

Inače, **ergodičnost** je svojstvo slučajnog niza (slučajnog procesa) da je određene parametre moguće dobiti na osnovu samo jedne realizacije istog.

4.4 Entropija stacionarnog Markovljevog izvora

Izraz za entropiju stacionarnog izvora informacija možemo da primenimo i na stacionarni Markovljev izvor. S obzirom da je Markovljev izvor definisan stacionarnom raspodelom $\mathbf{p}_1$ i matricom prelaza $\mathbf{\Pi}$, entropija je takođe funkcija ova dva parametra.

Teorema 4.4.1. *Entropija stacionarnog Markovljevog izvora jednaka je*

$$H(X) = H(X_2|X_1) = -\sum_{i=1}^{m} \sum_{j=1}^{m} \mu_i p_{ij} \log_2 p_{ij}.$$

gde su $\mathbf{\Pi} = [p_{ij}]_{1 \leq i,j \leq n}$ matrica prelaza a $\mathbf{p}_1 = \boldsymbol{\mu} = [\mu_1 \ \mu_2 \ \cdots \ \mu_n]$ stacionarna raspodela.

Dokaz. Pošto je X Markovljev izvor, važi:

$$\begin{aligned}
&H(X_n|X_1, \ldots, X_{n-1}) \\
&\quad = \sum_{x_1, \ldots, x_n \in \mathcal{X}} p(x_1, \ldots, x_n) \log_2 p(x_n|x_1, \ldots, x_{n-1}) \\
&\quad = \sum_{x_1, \ldots, x_n \in \mathcal{X}} p(x_n|x_1, \ldots, x_{n-1}) p(x_1, \ldots, x_{n-1}) \log_2 p(x_n|x_1, \ldots, x_{n-1}) \\
&\quad = \sum_{x_{n-1}, x_n \in \mathcal{X}} p(x_n|x_{n-1}) \left[\sum_{x_1, \ldots, x_{n-2} \in \mathcal{X}} p(x_1, \ldots, x_{n-1}) \right] \log_2 p(x_n|x_{n-1}) \\
&\quad = \sum_{x_{n-1}, x_n \in \mathcal{X}} p(x_n|x_{n-1}) p(x_{n-1}) \log_2 p(x_n|x_{n-1}) \\
&\quad = H(X_n|X_{n-1}).
\end{aligned}$$

Sa druge strane, iz stacionarnosti Markovljevog izvora X znamo da je $p_{X_n|X_{n-1}} = p_{X_2|X_1}$ i $p_{X_n} = p_{X_1}$. Odavde je

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

pa je i

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

Druga formula sledi direktno iz izraza za uslovnu entropiju. □

Primer 4.4.1. Neka je dat homogeni lanac markova sa matricom prelaza:

$$\mathbf{\Pi} = \begin{bmatrix} 0.3 & 0.2 & 0.5 & 0 \\ 0 & 0.1 & 0.6 & 0.3 \\ 0 & 0 & 0.2 & 0.8 \\ 0.9 & 0.1 & 0 & 0 \end{bmatrix}$$

Najpre određujemo stacionarnu raspodelu rešavanjem sistema jednačina:

$$\begin{aligned} \boldsymbol{\mu}\mathbf{\Pi} &= \boldsymbol{\mu} \\ \mu_1 + \mu_2 + \mu_3 + \mu_4 &= 1 \end{aligned}$$

Gornji sistem ima ukupno 5 jednačina, ali među prvih 4, samo 3 su linearno nezavisne. Prema tome, možemo da izostavimo na primer četvrtu jednačinu. Tako dobijamo sledeći sistem 4 jednačine sa 4 nepoznate:

$$\begin{aligned} (0.3-1)\mu_1 + 0.9\mu_4 &= 0 \\ 0.2\mu_1 + (0.1-1)\mu_2 + 0.1\mu_4 &= 0 \\ 0.2\mu_1 + (0.1-1)\mu_2 + 0.1\mu_4 &= 0 \\ 0.5\mu_1 + 0.6\mu_2 + (0.2-1)\mu_3 &= 0 \\ \mu_1 + \mu_2 + \mu_3 + \mu_4 &= 1 \end{aligned}$$

čije je rešenje:

$$\boldsymbol{\mu} = (0.3398, 0.1049, 0.2910, 0.2643)$$

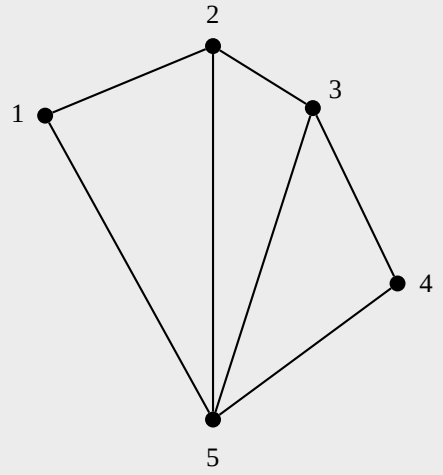
Sistem možemo rešavati npr. Gaussovim metodom, a možemo i primetiti da je $\boldsymbol{\mu}$ sopstveni vektor matrice $\mathbf{\Pi}^T$ koji odgovara sopstvenoj vrednosti $\lambda = 1$ a čiji je zbir komponenti takođe jednak 1. Ukoliko primetimo da $\mathbf{\Pi}^2$ nema elemenata koji su jednaki nuli, možemo primetiti i ergodičnu teoremu (Teorema 4.3.4),

tako što ćemo izračunati npr. $\mathbf{\Pi}^{32}$ (Kako ovo uraditi sa najmanjim brojem operacija?). Na taj način dobijamo:

$$\mathbf{\Pi}^{16} = \begin{bmatrix} 0.3389 & 0.1047 & 0.2910 & 0.2653 \\ 0.3400 & 0.1046 & 0.2903 & 0.2651 \\ 0.3415 & 0.1050 & 0.2903 & 0.2633 \\ 0.3389 & 0.1052 & 0.2922 & 0.2638 \end{bmatrix}, \quad \mathbf{\Pi}^{32} = \begin{bmatrix} 0.3398 & 0.1049 & 0.2910 & 0.2643 \\ 0.3398 & 0.1049 & 0.2910 & 0.2643 \\ 0.3398 & 0.1049 & 0.2910 & 0.2643 \\ 0.3398 & 0.1049 & 0.2910 & 0.2643 \end{bmatrix}.$$

Nakon što smo odredili stacionarnu raspodelu $\boldsymbol{\mu}$, primenom Teoreme 4.4.1 računamo entropiju: $H(X) = 0.975$.

Primer 4.4.2. (Slučajna šetnja na grafu) Kao primer lanca Markova, razmotrimo slučajnu šetnju (*random walk*) na **neorijentisanom** težinskom povezanom grafu. Označimo čvorove grafa brojevima $1, 2, \ldots, n$ a težinu ivice (i, j) sa $W_{ij} \geq 0$ (ukoliko ne postoji ivica, onda je $W_{ij} = 0$). Pretpostavimo da je graf *simetričan* tj. da je $W_{ij} = W_{ji}$ za svako $1 \leq i, j \leq n$.



Slika 4.2: Slučajna šetnja na grafu.

Slučajna šetnja se obavlja na sledeći način: ako smo u n-tom trenutku u čvoru i, verovatnoća prelaza u čvor j je proporcionalna težini W_{ij}. Vidimo da je u pitanju lanac Markova za koji je

$$p_{ij} = \frac{W_{ij}}{W_i}, \quad W_i = \sum_{j=1}^{n} W_{ij}.$$

Stacionarna raspodela u ovom slučaju ima jako jednostavan oblik, koji ćemo pretpostaviti a onda i dokazati. Ispostavlja se da je μ_i proporcionalno ukupnoj težini svih ivica W_i koje izlaze iz čvora i. Iz teorije grafova znamo da je

$$\sum_{i=1}^{n} W_i = 2W, \qquad W = \sum_{i<j} W_{ij},$$

pa je onda

$$\mu_i = \frac{W_i}{2W}.$$

Da je $\boldsymbol{\mu}$ zaista stacionarna distribucija, pokazuje se direktnom proverom jednakosti $\boldsymbol{\mu\Pi} = \boldsymbol{\mu}$:

$$\sum_{i=1}^{n} \mu_i p_{ij} = \sum_{i=1}^{n} \frac{W_i}{2W} \frac{W_{ij}}{W_i} = \sum_{i=1}^{n} \frac{W_{ij}}{2W} = \frac{W_j}{2W} = \mu_j.$$

Primetimo da ova raspodela ima interesantno svojstvo lokalnosti, tj. da verovatnoća μ_i zavisi samo od ivica koje polaze iz čvora i, kao i da se ne menja ukoliko se promene težine ostalih ivica, pod uslovom da ukupna težina svih ivica ostane konstantna.

Odredimo sada entropiju $H(X)$ slučajne šetnje na grafu. Na osnovu Teoreme 4.4.1 sledi

$$\begin{aligned}
H(X) = H(X_2|X_1) &= -\sum_{i=1}^{n} \mu_i \sum_{j=1}^{n} p_{ij} \log_2 p_{ij} \\
&= -\sum_{i=1}^{n} \frac{W_i}{2W} \sum_{j=1}^{n} \frac{W_{ij}}{W_i} \log_2 \frac{W_{ij}}{W_i} \\
&= -\sum_{i=1}^{n} \sum_{j=1}^{n} \frac{W_{ij}}{2W} \log_2 \frac{W_{ij}}{W_i} \\
&= -\sum_{i=1}^{n} \sum_{j=1}^{n} \frac{W_{ij}}{2W} \left(\log_2 \frac{W_{ij}}{2W} - \log_2 \frac{W_i}{2W} \right) \\
&= -\sum_{i=1}^{n} \sum_{j=1}^{n} \frac{W_{ij}}{2W} \log_2 \frac{W_{ij}}{2W} + \sum_{i=1}^{n} \frac{\sum_{j=1}^{n} W_{ij}}{2W} \log_2 \frac{W_i}{2W} \\
&= H\left(\ldots, \frac{W_{ij}}{2W}, \ldots \right) - H(\mu_1, \mu_2, \ldots, \mu_n).
\end{aligned}$$

Dakle, entropija izvora $H(X)$ jednaka je razlici "entropije ivica" i entropije stacionarne raspodele $\boldsymbol{\mu}$. Pretpostavimo sada da sve ivice imaju jednaku težinu,

tj. da graf nije težinski. Tada je

$$H(X) = \log_2(2e) - H\left(\frac{d_1}{2e}, \frac{d_2}{2e}, \ldots, \frac{d_n}{2e}\right) \tag{4.4}$$

gde je d_i broj ivica koje izlaze iz čvora i (stepen čvora i) a e ukupan broj ivica. Prema tome, entropija slučajne šetnje na netežinskom grafu zavisi isključivo od ukupnog broja ivica i entropije stacionarne distribucije.

Primer 4.4.3. Posmatrajmo sada slučajnu šetnju na putu dužine n (čvorovi su $1, 2, \ldots, n+1$ a ivice $\{i, i+1\}$ za $i = 1, 2, \ldots, n$). Pošto je

$$\mu_1 = \frac{1}{2n}, \quad \mu_2 = \frac{1}{n}, \quad \ldots \quad \mu_n = \frac{1}{n}, \quad \mu_{n+1} = \frac{1}{2n}.$$

onda je na osnovu izraza (4.4):

$$\begin{aligned} H(X) &= \log_2(2n) + \frac{1}{2n}\log_2\frac{1}{2n} + (n-1)\frac{1}{n}\log_2\frac{1}{n} + \frac{1}{2n}\log_2\frac{1}{2n} \\ &= \log_2 n + 1 - \frac{1}{n}(\log_2 n + 1) - \frac{n-1}{n}\log_2 n \\ &= 1 - \frac{1}{n}. \end{aligned}$$

Očigledno da $H(X) \to 1$ kad $n \to +\infty$, pošto tada nestaju "efekti krajeva".

4.5 Za dalje čitanje

- Drugi zakon termodinamike i povećanje entropije (Cover, sect. 4.4, p.81–84).
- Funkcije Markovljevih izvora, hidden Markov model (Cover, sect. 4.5, p.84–87).
- Više o Markovljevim izvorima (lancima Markova) [Ivković, p.108–125].

Glava 5

Izvorno kodiranje

Pojmovi *kod* i *kodiranje* često su u upotrebi u svakodnevnom životu. Da bi mogli da formulišemo kriterijum optimalnosti i precizno da kažemo koji kod je najbolji (optimalan), najpre moramo pojmove koda i kodiranja formalno da definišemo.

U ovoj glavi proučavamo kodove koji se koriste za kodiranje simbola koje nam daje izvor informacija. Cilj je odrediti takav kod koji omogućava jednostavno kodiranje odnosno dekodiranje i koji ujedno ima minimalnu srednju dužinu kodne reči. Drugim rečima, za smeštanje niza kodiranih simbola potrebno je (u srednjem) minimalna količina memorije.

5.1 Osnovni pojmovi i definicije

Najpre formalno definišemo pojmove *reči* i *alfabeta* koji su nam takođe poznati iz svakodnevnog života.

Definicija 5.1.1. *Ako je A konačan skup, onda je*

$$A^+ = A \cup A^2 \cup \cdots \cup A^n \cup \cdots = \bigcup_{n=1}^{+\infty} A^n$$

*skup **reči** nad **alfabetom** A.*

Skup A^+ sastoji se od nizova $w = a_1 a_2 \cdots a_n$ gde je $n \in \mathbb{N}$ i $a_i \in A$ za $i = 1, 2, \ldots, n$. Broj n je **dužina** reči a.

Definicija 5.1.2. *Neka su dati konačni skupovi A i B koje redom zovemo **alfabet izvora** i **alfabet koda**. **Kodiranje** je svaka $1-1$ funkcija $f : A' \to B^+$ gde je $A' \subseteq A^+$.*

Skup $V = f(A') \subset B^+$ je **kod** a njegovi elementi su **kodne reči** odnosno **kodne zamene** odgovarajućih reči iz A'.

Kodiranje je **alfabetno** ako je $A' = A$. U nastavku ćemo se baviti samo alfabetnim kodiranjem. U tom slučaju, **kodiranje neke poruke** $m = m_1 m_2 \cdots m_s \in A^+$ predstavlja zamenu svakog simbola (slova) m_i odgovarajućom kodnom reči $f(m_i)$. Tako se dobija kodirana poruka

$$w = f(m) = f(m_1)f(m_2)\cdots f(m_s).$$

Dekodiranje je obrnuti proces. Tada je za datu kodiranu poruku $w \in B^+$ potrebno odrediti originalnu poruku $m \in A^+$ takvu da je $f(m) = w$, odnosno da se kodiranjem poruke m dobija poruka w.

Primer 5.1.1. Neka je $A = \{0, 1, \ldots, 9\}$ i $B = \{0, 1\}$. Slede 2 primera kodiranja cifara iz A rečima skupa B:

$a)$

1	$\mapsto$	10	
2	$\mapsto$	110	
3	$\mapsto$	1110	
$\vdots$			
9	$\mapsto$	1111111110	
0	$\mapsto$	11111111110	

$b)$

1	$\mapsto$	0001	6	$\mapsto$	0110
2	$\mapsto$	0010	7	$\mapsto$	0111
3	$\mapsto$	0011	8	$\mapsto$	1000
4	$\mapsto$	0100	9	$\mapsto$	1001
5	$\mapsto$	0101	0	$\mapsto$	0000

Prvi kod je godinama korišćen u telefoniji (pulsno biranje) a drugi predstavlja standardni BCD kod. Postupak kodiranja poruke $m = 533015$ sastoji se u zameni svake cifre odgovarajućom kodnom reči. Tako se primenom prvog odnosno drugog koda dobijaju kodirane poruke:

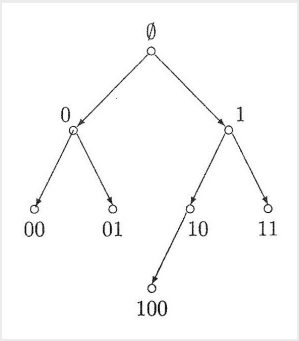
$$w_1 = f_1(m) = 111110\ 1110\ 1110\ 11111111110\ 10\ 111110$$
$$w_2 = f_2(m) = 0101\ 0011\ 0011\ 0000\ 0001\ 0101$$

Dekodirajmo sada poruku $w' = 10110111011110$, tj. odredimo originalni niz cifara m' koji kodiranjem prvim kodom daje poruku w'. Brojanjem uzastopnih jedinica nije teško utvrditi da je $m' = 1234$ i da ne postoji drugo rešenje ovog problema. Kodovi poput ovog, kod kojih je proces dekodiranja jednoznačan, nazivaju se **jednoznačno dekodabilni kodovi** i o njima će biti reči malo kasnije.

Neka je $V \subset B^+$ konačan skup reči nekog alfabeta B i neka je $\vec{V}$ skup svih prefiksa reči iz V. Elementima skupa $\vec{V}$ može se pridružiti stablo na sledeći

način: postoji ivica (x, y) između elemenata $x, y \in V$ ako je $y = x\alpha$ gde je $\alpha \in B$ proizvoljno slovo.

Primer 5.1.2. Neka je $B = \{0, 1\}$ i $V = \{00, 01, 10, 11, 100\}$. Tada je $\vec{V} = \{\emptyset, 0, 1, 01, 00, 10, 11, 100\}$ a odgovarajuće drvo prikazano je na sledećoj slici.

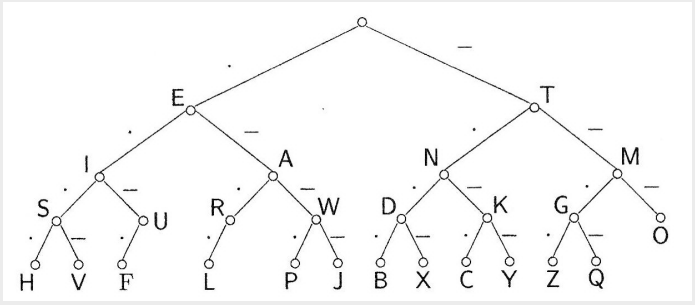


Slika 5.1: Stablo koje odgovara skupu prefiksa $\vec{V}$.

Kodiranje $f : A \to B^+$ je **sa fiksiranom dužinom kodnih reči**, ako je $f(A) \subseteq B^n$ za neko $n \in \mathbb{N}$. U prevodu, kodna reč svakog ima dužinu n. Odgovarajući kod $f(A)$ naziva se **blok kod**. U opštem slučaju, ako nisu sve kodne reči iste dužine, kodiranje je **sa promenljivom dužinom kodne reči**.

Svakom kodu $f(A)$ možemo pridružiti **kodno stablo** kao stablo prefiksa za skup $f(A)$.

Primer 5.1.3. Stablo koje odgovara Morseovom kodu (vidi sliku 1.1) dato je na sledećoj slici.



Slika 5.2: Stablo koje odgovara Morseovom kodu.

Već smo napomenuli da proces **dekodiranja** poruke $w \in B^+$ predstavlja obrnuti proces od procesa kodiranja. Zadatak primaoca je da primljenu poruku (reč) w predstavi u obliku

$$x = v_1 v_2 \cdots v_n$$

gde su $v_1, v_2, \ldots, v_n \in V = f(A)$ kodne reči[1]. Ukoliko se kodne reči mogu odabrati na jedinstven način, kažemo da je dekodiranje **jednoznačno** odnosno da kod V **omogućuje jednoznačno dekodiranje**. Drugim rečima, dekodiranje je jednoznačno ako iz

$$v_1 v_2 \cdots v_n = w_1 w_2 \cdots w_m$$

sledi $m = n$ i $v_i = w_i$ za svako $i = 1, 2, \ldots, m = n$.

S obzirom da kodiranje f predstavlja $1-1$ funkciju (po definiciji)[2], tada nije teško zaključiti da je svaki blok kod (kod sa fiksnom dužinom kodnih reči) jednoznačno dekodabilan. Ovo naravno ne mora da bude slučaj za kodove sa promenljivom dužinom kodnih reči, o čemu će biti više reči kasnije.

5.2 Prefiksni kod

Najvažnija klasa (izvornih) kodova su **prefiksni kodovi**. Kasnije ćemo pokazati da je ova klasa ujedno i dovoljna klasa ukoliko se jedino dužine kodnih reči prethodno fiksiraju.

Definicija 5.2.1. *Kod V ima **svojstvo prefiksa** (je **prefiksni**) ako nijedna kodna reč iz V nije prefiks neke druge kodne reči iz V*[3].

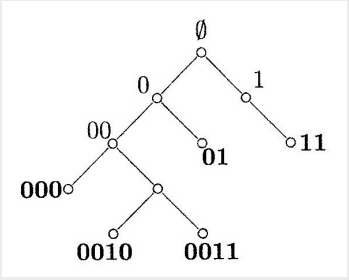
Posmatrajmo kodno stablo za prefiksni kod V. Nije teško zaključiti da se kodne reči nalaze isključivo u listovima stabla.

Primer 5.2.1. Neka je $V = \{01, 11, 000, 0011, 0010\}$. Ovaj kod je prefiksni i na slici je prikazano njegovo kodno stablo.

[1] Pri prenosu reči $w \in B^+$ nekim komunikacionim medijumom može doći do jedne ili više grešaka. Tada je potrebno (sa što manje promena) transformisati poruku w, tako da dekodiranje postane moguće. O dekodiranju pod ovim uslovima biće reči kasnije.

[2] Negde u literaturi, npr. u [Cover], razmatraju se i kodiranja/kodovi kod kojih f nije $1-1$ funkcija. Takvi kodovi nazivaju se **singularni kodovi**, i uglavnom imaju samo teorijski značaj.

[3] Pravilniji naziv za ovu klasu kodova bio bi **prefiks-slobodan** (eng. *prefix-free*), pošto kod ne sadrži prave prefikse svojih kodnih reči. I pored toga, naziv **prefiksni kod** se uglavnom sreće u literaturi.



Slika 5.3: Kodno stablo prefiksnog koda V.

Prefiksni kod se još naziva i **trenutni kod**. To je zato što omogućava dekodiranje nakon samo jednog prolaska kroz poruku. Postupak dekodiranja je jednostavan: čim se formira jedna kodna reč, prelazi se na formiranje druge.

Teorema 5.2.1. *Prefiksni kod omogućava jednoznačno dekodiranje.*

Dokaz. Pretpostavimo suprotno i neka je x najkraća reč iz alfabeta B koja može da se formira spajanjem kodnih reči iz V na dva različita načina:

$$x = v_1 v_2 \cdots v_k = w_1 w_2 \cdots w_m$$

gde su $v_1, v_2, \ldots, v_k \in V$ i $w_1, w_2, \ldots, w_m \in V$ kodne reči. Pretpostavimo da je $|v_1| > |w_1|$ tj. da je reč v_1 duža od reči w_1. Tada je očigledno reč w_1 prefiks reči v_1, što je nemoguće, s obzirom da je kod V prefiksni. Na isti način pokazujemo da ne može biti $|v_1| < |w_1|$ pa zaključujemo da su v_1 i w_1 iste dužine. Dalje je $v_1 = w_1$ kao i

$$v_2 \cdots v_k = w_2 \cdots w_m = x'$$

pa zaključujemo da je i reč x' formirana na dva načina. Ovo je nemoguće s obzirom na pretpostavku da je x najkraća takva reč. □

Primer 5.2.2. Posmatrajmo kodove $V_1 = \{0, 10, 110\}$ i $V_2 = \{10, 101, 001\}$, pri čemu je alfabet izvora u oba slučaja $A = \{a, b, c\}$. Može se pokazati da su oba koda jednoznačno dekodabilna. Međutim, u slučaju koda V_1 (koji je trenutni) dekodiranje je mnogo jednostavnije.

Pretpostavimo da se koristi kod V_1 i da je stigla poruka $x_1 = 10110$. Analizom poruke simbol po simbol dobijamo:

$$1$$
$$10$$
$$10|1$$
$$10|11$$
$$10|110$$

Ako isti algoritam primenimo za poruku $x_2 = 101001$ kodiranu pomoću V_2 dobijamo:

$$1$$
$$10$$
$$10|1$$
$$10|10$$
$$10|10|0$$
$$10|10|\mathbf{01}$$

Očigledno da je dekodiranje neuspešno izvršeno, pošto 01 nije kodna reč. potrebno je primiti celu poruku da bi se izvršilo dekodiranje. Ispravno dekodiranje je naravno 101|001.

Primer 5.2.3. Posmatrajmo sada kod $V = \{01, 10, 100\}$. Neka je poruka za dekodiranje $x_1 = 1001010\cdots10$. Ispravno dekodiranje ove poruke je $x_1 = 100|10|10|\cdots|10$. Međutim, ukoliko pristigne još jedna jedinica, potrebno je da dekodiramo $x_2 = x_1 1 = 1001010\cdots101$, gde je ispravno dekodiranje $x_2 = 10|01|01|\cdots|01|01$. Vidimo da je i ovde (u najgorem slučaju) potrebna cela poruka da bi izvršili dekodiranje, a ona može biti **proizvoljno dugačka**. Ni ovaj kod nije prefiksni, ali omogućava jednoznačno dekodiranje (videti naredni odeljak).

5.3 Kodovi koji omogućuju jednoznačno dekodiranje

Prefiksni kodovi nisu jedini koji omogućavaju jednoznačno dekodiranje. Na primer, kod $V = \{0, 01, 011\}$ nije prefiksni, ali omogućava jednoznačno dekodiranje. Ovo se lako dokazuje, ako se uoči činjenica da se kodne reči ovog koda dobijaju okretanjem reči prefiksnog koda $V' = \{0, 10, 110\}$.

Da bismo formulisali kriterijum jednoznačnog dekodiranja, definišemo niz skupova S_n na sledeći način:

$$S_1 = \{x \in B^+ \mid vx \in V, \text{ za neko } v \in V\}$$

kao i

$$\begin{aligned} S_n &= S'_n \cup S''_n \\ S'_n &= \{x \in B^+ \mid vx \in S_{n-1}, \text{ za neko } v \in V\} \\ S''_n &= \{x \in B^+ \mid s_{n-1}x \in V, \text{ za neko } s_{n-1} \in S_{n-1}\} \end{aligned}$$

Primetimo da su svi elementi skupova S_n sufiksi kodnih reči iz V. Drugim rečima $S_n \subseteq \overleftarrow{V}$ gde je $\overleftarrow{V}$ skup sufiksa reči iz V. Pored toga, neka je $U_1 = S_1$ i $U_{n+1} = S_{n+1} \cup U_n$ za $n \in \mathbb{N}$ (odnosno $U_n = S_1 \cup S_2 \cup \cdots \cup S_n$). Tada važi i $U_n \subseteq \overleftarrow{V}$ za svako $n \in \mathbb{N}$.

Lema 5.3.1. *Postoji prirodan broj $k \in \mathbb{N}$ takav da je $U_{k+1} = U_k$. Tada je i $U_n = U_k$ za svako $n > k$.*

Neka je $U = U_k$, gde je k prirodan broj iz prethodne leme. Narednom teoremom formulišemo kriterijum (tj. algoritam za proveru) jednoznačne dekodabilnosti koda V.

Teorema 5.3.2. *Kod $V = \{v_1, v_2, \ldots, v_a\}$ nad alfabetom B ($V \subset B^+$) omogućuje jednoznačno dekodiranje akko je*

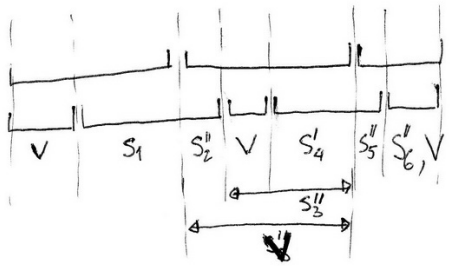
$$U \cap V = \emptyset,$$

tj. akko nijedan sufiks iz U nije istovremeno i kodna reč.

Dokaz. [Šešelja, p.76-77] Napisati precizniji dokaz i algoritam. □

Na slici 5.4 ilustrovan je jedan mogući scenario kada kod nije jednoznačno dekodabilan.

U specijalnom slučaju, kada je V prefiksni kod, on omogućuje jednoznačno dekodiranje, a za njega je $S_1 = \emptyset$. Otuda i $U = \emptyset$, odnosno $U \cap V = \emptyset$. Dokaz prethodne teoreme nam ujedno daje i (efikasan) algoritam provere jednoznačne dekodabilnosti koda. Ukoliko kod nije jednoznačno dekodabilan, primenom gore opisanog metoda može da se konstruiše minimalna reč x koja može na dva različita načina da se predstavi kao kombinacija kodnih reči.



Slika 5.4: Ilustracija Teoreme 5.3.2.

Primer 5.3.1 (Šešelja, primer 3.14, p.77)**.** Neka je $V = \{10, 101, 001\}$. Na osnovu izraza za S_n dobijamo $S_1 = \{0, 1, 01\}$ kao i:

- $S_2' = \emptyset$, $S_2'' = \{01, 0\}$ pa je $S_2 = \{01, 0\}$
- $S_3' = \emptyset$, $S_3'' = \{01\}$ pa je $S_2 = \{01\}$
- $S_4' = \emptyset$, $S_4'' = \emptyset$ pa je $S_4 = \emptyset$, odnosno $U_4 = S_4 \cup U_3 = U_3 = U$.

Prema tome $U = S_1 \cup S_2 \cup S_3 = \{0, 1, 01\}$. Pošto je $U \cap V = \emptyset$ sledi da je je V jednoznačno dekodabilan.

5.4 Kraftova nejednakost

Ova nejednakost predstavlja glavno svojstvo prefiksnih (a kao što ćemo u nastavku sekcije videti) i uopšte, jednoznačno dekodabilnih kodova.

Teorema 5.4.1. ***(Kraftova nejednakost)*** *Ukoliko je kod* $V = \{v_1, v_2, \ldots, v_a\}$ *prefiksni, pri čemu je* $|v_i| = n_i$ $(i = 1, 2, \ldots, a)$*, onda važi*

$$\sum_{i=1}^{a} b^{-n_i} \leq 1.$$

gde je $b = |B|$*. Važi i obrnuto, ukoliko za brojeve* $n_1, n_2, \ldots, n_a$ *važi prethodna nejednakost, onda postoji prefiksni kod* $V = \{v_1, v_2, \ldots, v_a\}$ *za koji važi* $|v_i| = n_i$ $(i = 1, 2, \ldots, a)$*.*

Dokaz.
($\Rightarrow$:) Pretpostavimo da je $n_1 \leq n_2 \leq \cdots \leq n_a$ i neka je $n = n_a$. Formirajmo skupove B_{v_i} koji se sastoje od svih reči dužine n čiji je prefiks reč v_i. Dakle,

$$B_{v_i} = \{x \in B^n \mid x = v_i w, \quad w \in B^{n-n_i}\}.$$

Ukoliko neka reč x pripada skupovima B_{v_i} i B_{v_j} onda je $x = v_i w = v_j w'$, pa sledi da je v_i prefiks reči v_j, što je nemoguće. Prema tome, skupovi B_{v_i} su disjuntni. Skup B_{v_i} ima b^{n-n_i} reči, što ukupno čini

$$b^{n-n_1} + b^{n-n_2} + \ldots + b^{n-n_a}$$

reči. Sa druge strane, broj reči od n slova iz skupa B jednak je b^n, pa je

$$b^{n-n_1} + b^{n-n_2} + \ldots + b^{n-n_a} \leq b^n$$

odakle skraćivanjem sa b^n dobijamo Kraftovu nejednakost.

($\Leftarrow$:) Neka su $n_1, n_2, \ldots, n_a$ brojevi koji zadovoljavaju Kraftovu nejednakost. Pokazaćemo da postoji prefiksni kod $V = \{v_1, v_2, \ldots, v_a\}$ takav da je $|v_i| = n_i$ za svako $i = 1, 2, \ldots, a$. Pretpostavimo da je $n_1 \leq n_2 \leq \cdots \leq n_a$. Neka je

$$q_1 = 0, \qquad q_i = \sum_{k=1}^{i-1} b^{-n_k}, \quad (i = 2, 3, \ldots, a)$$

Očigledno je $q_1 < q_2 < \cdots < q_a < 1$ (poslednja nejednakost $q_a < 1$ sledi direktno iz Kraftove nejednakosti). Predstavimo brojeve q_i u sistemu sa osnovom b na sledeći način:

$$q_i = \left(\overline{0.C_1^{(i)} C_2^{(i)} \cdots C_{n_i}^{(i)}}\right)_b = \sum_{j=1}^{n_i} C_j^{(i)} b^{-j}.$$

Bez gubitka opštosti, možemo pretpostaviti da je $B = \{0, 1, \ldots, b-1\}$. Sada kodne reči v_i definišemo na sledeći način:

$$v_i = C_1^{(i)} C_2^{(i)} \cdots C_{n_i}^{(i)},$$

odnosno kao niz prvih n_i cifara broja q_i iza tačke, u sistemu sa osnovom b. Pošto je $q_i = \sum_{k=1}^{i-1} b^{-n_i}$, sledi da se sve nenula cifre broja q_i nalaze među prvih n_i cifara. Pošto je očigledno $|v_i| = n_i$ za svako $i = 1, 2, \ldots, a$, potrebno je samo pokazati da je kod prefiksni, tj. da ne postoje kodne reči v_i i v_j koje su jedna prefiks drugoj.

Pretpostavimo suprotno, da je $i < j$ i da je v_i prefiks od v_j. Tada je $C_i^{(k)} = C_j^{(k)}$ za svako $k = 1, 2, \ldots, n_i$ pa je

$$q_j - q_i = \left(0.\underbrace{00\cdots0}_{n_i} C_{n_i+1}^{(j)} C_{n_i+2}^{(j)} \cdots C_{n_j}^{(j)}\right)_b < b^{-n_i}.$$

Sa druge strane,

$$q_j - q_i = \sum_{k=i}^{j-1} b^{-n_k} \geq b^{-n_i}$$

što je kontradikcija sa malopre dokazanom nejednakošću. Ovim je dokaz završen. □

Primer 5.4.1. Neka je $V = \{00, 01, 100, 1010, 1011\}$ prefiksni kod čije su dužine kodni reči redom jednake 2,2,3,4,4. Kraftova nejednakost je zadovoljena, pošto je

$$\frac{1}{2^2} + \frac{1}{2^2} + \frac{1}{2^3} + \frac{1}{2^4} + \frac{1}{2^4} = \frac{3}{4} < 1.$$

Obrnuto, polazeći od brojeva $n_1 = n_2 = 2$, $n_3 = 3$, $n_4 = n_5 = 4$ može se konstruisati binarni prefiksni kod čije su to dužine kodnih reči. Zaista, ako je

$$\begin{aligned} q_1 &= 0 \\ q_2 &= 2^{-2} \\ q_3 &= 2^{-2} + 2^{-2} \\ q_4 &= 2^{-2} + 2^{-2} + 2^{-3} \\ q_5 &= 2^{-2} + 2^{-2} + 2^{-3} + 2^{-4} \end{aligned}$$

onda je $q_1 = (0.00)_2$, $q_2 = (0.01)_2$, $q_3 = (0.100)_2$, $q_4 = (0.1010)_2$, $q_5 = (0.1011)_2$. Izdvajanjem cifara desno od zareza dobijaju se upravo kodne reči koda V.

I pored toga što postoje jednoznačno dekodabilni kodovi koji nisu prefiksni, oni daju podjednake rezultate u praksi kao i prefiksni kodovi. To sledi iz činjenice da Kraftovu nejednakost (Teorema 5.4.1) zadovoljavaju ne samo prefiksni, već svi jednoznačno dekodabilni kodovi.

Teorema 5.4.2. ***(McMillanova teorema)*** *Neka je $V = \{v_1, v_2, \ldots, v_a\}$ kod koji omogućuje jednoznačno dekodiranje nad alfabetom B, $b = |B|$ i neka je $|v_k| = n_k$ za svako $k = 1, 2, \ldots, a$. Tada važi Kraftova nejednakost, tj.*

$$\sum_{i=1}^{n} b^{-n_i} \leq 1.$$

Dokaz. Neka je k proizvoljan prirodan broj. Tada je

$$\left(\sum_{i=1}^{a} b^{-n_i}\right)^k = \sum_{(i_1,i_2,\ldots,i_k)\in\{1,2,\ldots,a\}^k} b^{-(n_{i_1}+n_{i_2}+\ldots+n_{i_k})}.$$

gde se zbir na desnoj strani odnosi na sve k-torke nad skupom $\{1, 2, \ldots, a\}$ (kada se izmnože svi zbirovi u zagradama). Svakoj n-torci $(i_1, i_2, \ldots, i_k)$ odgovara reč $x = v_{i_1}v_{i_2}\cdots v_{i_k}$ dužine $j = n_{i_1} + n_{i_2} + \ldots + n_{i_k}$. Neka je m_j broj n-torki koje daju reč x dužine j. Tada je

$$\left(\sum_{i=1}^{a} b^{-n_i}\right)^k = \sum_{j=1}^{kn} m_j b^{-j}$$

Sve ovako dobijene reči su međusobno različite. Zaista, ukoliko bi dvema n-torkama $(i_1, i_2, \ldots, i_k)$ i $(i'_1, i'_2, \ldots, i'_k)$ odgovarale iste reči:

$$x = v_{i_1}v_{i_2}\cdots v_{i_k} = v_{i'_1}v_{i'_2}\cdots v_{i'_k}$$

što je u suprotnosti sa pretpostavkom da je kod jednoznačno dekodabilan[4]. Prema tome, m_j je broj reči dužine j koje se na ovaj način dobijaju. Pošto je b^j maksimalan broj reči dužine j, to je

$$m_j \leq b^j$$

odakle dobijamo

$$\left(\sum_{i=1}^{a} b^{-n_i}\right)^k \leq \sum_{j=1}^{kn} b^j \cdot b^{-j} = kn$$

odnosno

$$\sum_{i=1}^{a} b^{-n_i} \leq \sqrt[k]{kn} \longrightarrow 1 \qquad (k \to +\infty).$$

Ovim je Kraftova nejednakost dokazana. □

Prema tome, ukoliko neki kod W omogućuje jednoznačno dekodiranje, tada postoji prefiksni kod V sa istim dužinama kodnih reči koji takođe omogućuje jednoznačno dekodiranje. Pošto je proces dekodiranja neuporedivo jednostavniji kod prefiksnih kodova, ove kodove treba primenjivati u praksi kad god je to moguće.

[4]Dovoljno je i da pretpostavimo nešto slabiji uslov, da ne postoji reč x koja može da se dobije spajanjem **istog broja** (k) kodnih reči.

5.5 Optimalni kodovi

Pretpostavimo da je alfabet A skup vrednosti koje uzima neka diskretna slučajna promenljiva X, tj. da je $A = \mathcal{X} = X(\Omega)$. Neka je $A = \{\alpha_1, \alpha_2, \ldots, \alpha_a\}$ i $p(\alpha_i) = p_i$ za svako $i = 1, 2, \ldots, n$.

Definicija 5.5.1. *Neka je $V = f(A) = \{v_1, v_2, \ldots, v_a\}$ kod definisan nad alfabetom $A = \mathcal{X} = X(\Omega) = \{\alpha_1, \alpha_2, \ldots, \alpha_a\}$, tako da je $f(\alpha_i) = v_i$ za $i = 1, 2, \ldots, a$. Neka je $n_i = |v_i|$ dužina i-te kodne reči a $p_i = p(\alpha_i)$ verovatnoća pojavljivanja simbola α_i. Veličina*

$$\bar{n}_V := \sum_{i=1}^{a} p_i n_i = \mathbb{E}|f(X)|$$

naziva se ***srednja (prosečna) dužina kodnih reči*** *koda V. Iako $\bar{n}_V$ suštinski zavisi i od koda V i od raspodele p, zavisnost od raspodele nećemo navoditi, s obzirom da će uvek biti jasno o kojoj raspodeli se radi.*

Očigledno je da $\bar{n}_V$ zavisi i od redosleda kodnih reči $v_1, v_2, \ldots, v_a$. Zbog toga ćemo u nastavku podrazumevati da je kod V zapravo a-torka elemenata $V = (v_1, v_2, \ldots, v_a)$, gde v_i odgovara simbolu α_i, čija je verovatnoća pojavljivanja p_i.

Ako kodiramo vrednost slučajne promenljive X pomoću koda V, onda bi srednja dužina odgovarajućih kodnih reči ($f(X)$) bila jednaka $\bar{n}_V$. Ako posmatramo poruku $x = x_1 x_2 \cdots x_N$ koja se sastoji od N nezavisnih realizacija slučajne promenljive X, tada bi "očekivana" dužina reči $w = f(x)$ dobijene kodiranjem poruke x kodom V, bila

$$|f(x_1)| + \ldots + |f(x_N)| \rightsquigarrow N \cdot \bar{n}_V \ ^{5}.$$

Primer 5.5.1 (Šešelja, primer 3.18, p. 79)**.** Neka je $A = \{\alpha_1, \alpha_2, \alpha_3, \alpha_4\}$. U sledećoj tabeli data su dva koda V i W, kao i odgovarajuće dužine kodnih reči za svaki simbol:

α_i	p_i	v_i	n_i	w_i	n'_i
1	0.45	00	2	0	1
2	0.20	01	2	10	2
3	0.20	10	2	110	2
4	0.15	11	2	111	3

[5] Preciznije, ako je $(X_n)_{n\in\mathbb{N}}$ niz nezavisnih diskretnih slučajnih promenljivih koje imaju istu raspodelu p, tada je $(|f(X_1)| + \ldots + |f(X_N)|)/(N \cdot \bar{n}_V) \to 1$, pri čemu je konvergencija skoro izvesna (Strogi zakon velikih brojeva, Teorema 7.2.2).

Srednje dužine kodnih reči su:

$$\bar{n}_V = 0.45 \cdot 2 + 0.2 \cdot 2 + 0.2 \cdot 2 + 0.15 \cdot 2 = 2$$
$$\bar{n}_W = 0.45 \cdot 1 + 0.2 \cdot 2 + 0.2 \cdot 2 + 0.15 \cdot 3 = 1.9$$

Prema tome, kod W je bolji pošto ima manju srednju dužinu kodnih reči. Posmatrajmo sada jednu poruku m sastavljenu od $N = 20$ realizacija slučajne promenljive X:

$$m = 1\ 1\ 4\ 4\ 2\ 1\ 1\ 1\ 1\ 2\ 3\ 1\ 2\ 1\ 1\ 4\ 2\ 1\ 2\ 3$$

Ako ovu poruku kodiramo kodovima V i W dobićemo sledeće kodirane poruke (reči):

$$w_V = 00\ 00\ 11\ 11\ 01\ 00\ 00\ 00\ 00\ 01\ 10\ 00\ 01\ 00\ 00\ 11\ 01\ 00\ 01\ 10$$
$$w_W = 0\ 0\ 111\ 111\ 10\ 0\ 0\ 0\ 0\ 10\ 110\ 0\ 10\ 0\ 0\ 111\ 10\ 0\ 10\ 110$$

Dužina prve kodirane poruke je $|w_V| = 40$ a druge $|w_W| = 35$. Dakle, u ovom slučaju kod W daje bolji rezultat za ukupno 5 bita. Postoje poruke za koje je kod V bolji od koda W (npr. ako je $m = 3\ 4\ 3\ 4\ \cdots\ 3\ 4$), ali je malo verovatno da će se takva poruka dobiti kao niz realizacija slučajne promenljive X.

Iz primera se lako uočava da različiti kodovi daju različite srednje dužine kodnih reči. Pritom je kod bolji ukoliko je srednja dužina kodnih reči manja. Prema tome, za dati izvor X možemo definisati sledeću veličinu

$$n_* := \inf_V \bar{n}_V$$

gde se infimum uzima po svim jednoznačno dekodabilnim kodovima (na osnovu McMillanove teoreme, dovoljno je posmatrati prefiksne kodove). Ukoliko je $n_* = \bar{n}_V$, za neki kod V, onda je taj kod **optimalan**. Sledeća teorema daje važnu donju granicu za $\bar{n}_V$ proizvoljnog koda V, a time u isto vreme i za n_*.

Teorema 5.5.1. *Neka je data diskretna slučajna promenljiva X čiji je skup vrednosti A, i alfabet koda B ($b = |B|$). Tada za svaki kod V koji omogućuje jednoznačno dekodiranje važi*

$$\bar{n}_V \geq \frac{H(X)}{\log_2 b} = H_b(X) \tag{5.1}$$

gde je $H(X)$ entropija slučajne promenljive X. Jednakost važi ako i samo ako je $p_i = q_i = b^{-n_i}$ za svako $i = 1, 2, \ldots, a$.

Dokaz. [Šešelja, p.80] Neka je $q_i = b^{-n_i}$ gde je $n_i = |v_i|$ dužina i-te kodne reči, za $i = 1, 2, \ldots, a$. Pošto je kod V jednoznačno dekodabilan, na osnovu McMillanove teoreme (Teorema 5.4.2) sledi

$$\sum_{i=1}^{a} q_i = \sum_{i=1}^{a} b^{-n_i} \leq 1 = \sum_{i=1}^{a} p_i,$$

a pošto je očigledno $q_i \neq 0$ za svako $i = 1, 2, \ldots, a$, na osnovu važne leme (Lema 3.3.1) sledi

$$\begin{aligned} H_b(X) &= -\sum_{i=1}^{a} p_i \log_b p_i \leq -\sum_{i=1}^{a} p_i \log_b q_i = -\sum_{i=1}^{a} p_i \log_b b^{-n_i} \\ &= \sum_{i=1}^{a} n_i p_i = \bar{n}_V. \end{aligned}$$

Jednakost u važnoj lemi je zadovoljena ako i samo ako je $p_i = q_i = b^{-n_i}$ za svako $i = 1, 2, \ldots, a$. □

U specijalnom slučaju $b = 2$ (binarno kodiranje) dobijamo $\bar{n}_V \geq H(X)$. Pokazali smo da u (5.1) važi jednakost akko je $b^{-n_i} = p_i$ odnosno ako je $n_i = -\log_b p_i$ za svako $i = 1, 2, \ldots, n$. Iako se granica dostiže samo za specijalne raspodele (za koje je $-\log_b p_i$ ceo broj), u nastavku ćemo pokazati da se ovoj granici može prići "relativno blizu".

Teorema 5.5.2. *Neka je data diskretna slučajna promenljiva X čiji je skup vrednosti A, i alfabet koda B ($b = |B|$). Tada postoji prefiksni kod V takav da je*

$$\bar{n}_V < \frac{H(X)}{\log_2 b} + 1. \tag{5.2}$$

Dokaz. [Šešelja, p.81-82] Činjenica da jednakost u prethodnoj teoremi važi samo za raspodele kod kojih je $-\log_b p_i$ ceo broj, sugeriše da konstruišemo kod koji ima srednje dužine kodnih reči $n_i = \lceil -\log_b p_i \rceil$, gde je $\lceil x \rceil$ najmanji ceo broj veći ili jednak x. Pritom očigledno važi $x \leq \lceil x \rceil < x + 1$, pa je $-\log_b p_i \leq n_i$ odnosno $p_i \geq b^{-n_i}$. Brojevi $n_1, n_2, \ldots, n_a$ zadovoljavaju Kraftovu nejednakost. Zaista,

$$\sum_{i=1}^{a} b^{-n_i} \leq \sum_{i=1}^{a} p_i = 1,$$

pa na osnovu Teoreme 5.4.1 sledi da postoji prefiksni kod V takav da je n_i dužina kodne reči v_i za $i = 1, 2, \ldots, n$. Srednju dužinu kodne reči $\bar{n}_V$ ovog koda možemo ograničiti na sledeći način

$$\bar{n}_V = \sum_{i=1}^{a} n_i p_i < \sum_{i=1}^{a} (-\log_b p_i + 1) p_i = -\sum_{i=1}^{a} p_i \log_b p_i + \sum_{i=1}^{a} p_i = \frac{H(X)}{\log_2 b} + 1$$

s obzirom da je $\log_b x = (\log_2 x)/(\log_2 b)$. Prema tome, ovako konstruisan kod V zadovoljava uslove teoreme, cime je dokaz završen. □

Napomena: U prethodnom dokazu implicitno smo pretpostavili da je $p_i \neq 0$ za svako $i = 1, 2, \ldots, a$. Ukoliko je, na primer $p_{s+1} = \cdots = p_a = 0$, možemo postupiti na sledeći način. Posmatrajmo raspodelu p'_i, $i = 1, 2, \ldots, a$ za koju je $p'_i = p_i - \epsilon$ kao i $p'_{s+1} = \cdots = p'_a = \epsilon \cdot s/(n-s)$, i neka je $n_i = \lceil -\log_b p'_i \rceil$ za svako $i = 1, 2, \ldots, a$. S obzirom da je $p'_1 + p'_2 + \ldots + p'_n = 1$, ovako definisani brojevi $n_1, n_2, \ldots, n_a$ zadovoljavaju Kraftovu nejedankost. Za dovoljno malo ϵ je $n_i < -\log_b p_i + 1$ za $i = 1, 2, \ldots, s$. Preciznije, dovoljno je uzeti bilo koje ϵ tako da važi

$$0 < \epsilon < \min\{p_i(1 - b^{-t_i}) \mid i = 1, 2, \ldots, s\}, \quad t_i = -\log_b p_i + 1 - \lceil -\log_b p_i \rceil.$$

Nadalje dokaz ide kao u Teoremi 5.5.2.

Direktno iz prethodne dve teoreme dobijamo sledeću posledicu koja daje granice za minimalnu srednju dužine kodne reči.

Posledica 5.5.3. *Za svaku diskretnu slučajnu promenljivu X je*

$$\frac{H(X)}{\log_2 b} \leq n_* < \frac{H(X)}{\log_2 b} + 1$$

Videli smo da (prefiksni) kod koji zadovoljava gornju granicu (5.2) za $\bar{n}_V$, ima kodne reči dužine $n_i = \lceil -\log_b p_i \rceil$ $(i = 1, 2, \ldots, a)$. Odgovarajući kod, koji se na ovaj na čin dobija (tj. primenom metoda opisanog u dokazu Teoreme 5.4.1 - Kraftova nejednakost), naziva se **Shannonov kod**. Vrednost sa desne strane (5.2) je upravo posledica činjenice da $\log_b p_i$ nije ceo broj za svako $i = 1, 2, \ldots, a$.

Primer 5.5.2. Neka je $A = X(\Omega) = \{a, b\}$ i neka je $p(a) = 0.001$ a $p(b) = 0.999$. Tada je optimalni binarni kod za ovaj izvor $V = \{0, 1\}$ $(f(a) = 0$ i

$f(b) = 1$) za koji važi $\bar{n}_V = 1$. U isto vreme je $H(X) \approx 0.0079$, pa vidimo da je $\bar{n}_V = 1$ jako blizu granice $H(X) + 1 \approx 1.0079$.

Film: `(IC 4.5) An issue with Huffman coding` (Potražiti film na Youtube-u.)

Ova razlika može da se ublaži ukoliko imamo na raspolaganju više odmeraka slučajne promenljive X koje trebamo da kodiramo. Drugim rečima, ukoliko Teoreme 5.5.1 i 5.5.2 primenimo na k-dimenzionalnu slučajnu promenljivu $(X_1, X_2, \ldots, X_k)$ dobićemo

$$\frac{H(X_1, X_2, \ldots, X_k)}{\log_2 b} \leq \bar{n}_{V_k} < \frac{H(X_1, X_2, \ldots, X_k)}{\log_2 b} + 1$$

Ovde je $\bar{n}_{V_n}$ srednja dužina kodne reči za odgovarajući kod V_k, konstruisan primenom postupka opisanog u dokazu Teoreme 5.5.2, za slučajnu promenljivu $(X_1, X_2, \ldots, X_k)$. Ukoliko prethodni izraz podelimo sa k dobijamo:

$$\frac{H(X_1, X_2, \ldots, X_k)}{k \log_2 b} \leq \frac{\bar{n}_{V_k}}{k} < \frac{H(X_1, X_2, \ldots, X_k)}{k \log_2 b} + \frac{1}{k}. \tag{5.3}$$

Primetimo da je $\bar{n}_{V_k}/k$ **srednja dužina kodne reči po simbolu** (kodiramo k simbola odjednom).

Pretpostavimo sada da je $X = (X_k)_{k \in \mathbb{N}}$ stacionarni slučajni niz (izvor informacija). Tada iz (5.3) i definicije entropije slučajnog procesa sledi

$$\lim_{k \to +\infty} \frac{\bar{n}_{V_k}}{k} = \frac{1}{\log_2 b} \lim_{k \to +\infty} \frac{H(X_1, X_2, \ldots, X_k)}{k} = \frac{H(X)}{\log_2 b}.$$

Ovim smo dokazali sledeću teoremu, koja je poznata kao Shannon-Fano stav.

Teorema 5.5.4. ***(Shannon-Fano stav)*** [6] *Minimalna srednja dužina kodnih reči po simbolu $n_{*,k}$ k-dimenzionalne slučajne promenljive zadovoljava nejednakost*

$$\frac{H(X_1, X_2, \ldots, X_k)}{k \log_2 b} \leq n_{*,k} < \frac{H(X_1, X_2, \ldots, X_k)}{k \log_2 b} + \frac{1}{k}. \tag{5.4}$$

Ukoliko je $X = (X_k)_{k \in \mathbb{N}_0}$ stacionarni izvor informacija, tada je

$$\lim_{k \to +\infty} n_{*,k} = \frac{H(X)}{\log_2 b}.$$

[6] Poznata je i kao *Prva Shannonova teorema* ili *Teorema o kodiranju bez šumova.*

Prema tome, tehnikom grupisanja simbola, moguće je **dostići teorijsku donju granicu** za srednju dužinu kodne reči, određenu entropijom slučajne promenljive ili proizvoljnog stacionarnog izvora informacija X. Ovo je jedan od najvažnijih rezultata u teoriji informacija. Naredni primer ilustruje ovu tehniku.

Primer 5.5.3 (Šešelja, primer 3.2, str. 82)**.** Neka su alfabeti izvora odnosno koda $A = \{\alpha_1, \alpha_2, \alpha_3\}$ i $B = \{0, 1\}$ i neka su verovatnoće simbola date sa $p(\alpha_1) = 0.7$, $p(\alpha_2) = 0.2$ i $p(\alpha_3) = 0.1$.

Jedan binarni prefiksni kod za ovu slučajnu promenljivu je $V = (0, 10, 11)$ koji ima srednju dužinu kodne reči

$$\bar{n}_V = 0.7 \cdot 1 + 0.2 \cdot 2 + 0.1 \cdot 2 = 1.3.$$

Iako je ovaj kod optimalan (videćemo kasnije kako se to pokazuje), ova vrednost je primetno veća od vrednosti entropije:

$$H(X) = -0.7 \log_2 0.7 - 0.2 \log_2 0.2 - 0.1 \log_2 0.1 = 1.1568.$$

Ali ako kodiramo po dva simbola istovremeno kodom prikazanom u sledećoj tabeli:

x_1x_2	$p(x_1, x_2)$	$f(x_1x_2)$	$\lvert f(x_1x_2)\rvert$
$\alpha_1\alpha_1$	0.49	0	1
$\alpha_1\alpha_2$	0.14	100	3
$\alpha_1\alpha_3$	0.07	1010	4
$\alpha_2\alpha_1$	0.14	110	3
$\alpha_2\alpha_2$	0.04	1011	4
$\alpha_2\alpha_3$	0.02	11110	5
$\alpha_3\alpha_1$	0.07	1110	4
$\alpha_3\alpha_2$	0.02	111110	6
$\alpha_3\alpha_3$	0.01	111111	6

Srednja dužina kodne reči za ovaj kod je $\bar{n}_{V_2} = 2.33$, odnosno $\bar{n}_{V_2}/2 = 1.165$ po simbolu, što je znatno bliže entropiji nego vrednost $\bar{n}_V$ dobijena kodiranjem samo jednog simbola. Napomenimo i to da su oba koda dobijena primenom Huffmanovog algoritma, i kao što ćemo u nastavku videti, predstavljaju optimalne kodove za odgovarajuće raspodele.

Videli smo da su kod Shannonovog koda, dužine kodnih reči jednake $n_i = \lceil -\log_b p_i \rceil$. Vrednosti verovatnoća p_i često nisu unapred poznate, već je potrebno proceniti njihove vrednosti. Na taj način dobijamo procenjene vrednosti q_i,

$i = 1, 2, \ldots, a$ verovatnoća svakog simbola, pa dužine kodnih reči formiramo na osnovu njih, kao $n_i = \lceil -\log_b p_i \rceil$. Ove vrednosti takođe zadovoljavaju Kraftovu nejednakost (ukoliko je $q_1 + q_2 + \ldots + q_a = 1$), pa sledi da postoji odgovarajući prefiksni kod V. Međutim, on ima nešto lošije granice nego kod dobijen pomoću tačnih vrednosti p_i. I gornja i donja granica pomerene su za vrednost relativne entropije $D(p||q)$ podeljene izrazom $\log_2 b$.

Teorema 5.5.5. ***(Pogrešan kod)*** *Ukoliko je $p_i \neq 0$ i $q_i \neq 0$ za svako $i = 1, 2, \ldots, a$, srednja dužina kodne reči $\bar{n}_V$ koda V za koji je $n_i = |v_i| = \lceil -\log_2 q_i \rceil$ zadovoljava nejednakost*

$$\frac{H(p) + D(p||q)}{\log_2 b} \leq \bar{n}_V < \frac{H(p) + D(p||q)}{\log_2 b} + 1.$$

Dokaz. [Cover, Theorem 5.4.3, p. 115] Iz $-\log_b q_i \leq n_i < -\log_b q_i + 1$ sledi

$$-\sum_{i=1}^{a} p_i \log_b q_i \leq \sum_{i=1}^{a} n_i p_i < -\sum_{i=1}^{a} p_i \log_b q_i + 1$$

pa iz

$$\begin{aligned} -\sum_{i=1}^{a} p_i \log_b q_i &= -\sum_{i=1}^{a} p_i \log_b \left(p_i \cdot \frac{q_i}{p_i} \right) \\ &= -\sum_{i=1}^{a} p_i \log_b p_i - \sum_{i=1}^{a} p_i \log_b \frac{q_i}{p_i} = \frac{H(p) + D(p||q)}{\log_2 b} \end{aligned}$$

dobijamo tvrdjenje teoreme. □

5.6 Potrebni uslovi za optimalnost koda

Iako je konstrukcijom Shannonovog koda, uz korišćenje tehnike grupisanja simbola moguće prići donjoj granici za srednju dužinu kodne reči po simbolu, ovaj kod nije u opštem slučaju optimalan. U nastavku ćemo razmotriti jedan algoritam za konstrukciju optimalnog koda za datu raspodelu. Pre same formulacije ovog algoritma, dokazujemo naredna 3 svojstva koja predstavljaju potrebne uslove za optimalnost nekog koda.

Teorema 5.6.1. *Neka je V optimalni kod za slučajnu promenljivu X. Tada iz $p_i > p_j$ sledi i $n_i \leq n_j$.*

Dokaz. [Šešelja, p.84] Pretpostavimo suprotno. Neka se kod V sastoji od sledećih kodnih reči:

$$v_1, \ldots, v_i, \ldots, v_j, \ldots, v_a$$

pri čemu je $n_i < n_j$ a $p_i > p_j$. Ukoliko rečima v_i i v_j zamenimo mesta dobićemo novi kod V' koji se sastoji (redom) od kodnih reči [7]:

$$v_1, \ldots, v_j, \ldots, v_i, \ldots, v_a$$

Pritom je

$$\bar{n}_V - \bar{n}_{V'} = n_i p_i + n_j p_j - n_j p_i - n_i p_j = (p_i - p_j)(n_j - n_i) > 0$$

što je u kontradikciji sa pretpostavkom da je kod V optimalan. □

U nastavku ćemo pretpostaviti da su svi kodovi koje konstruišemo **binarni**, odnosno da je $b = 2$ i $B = \{0, 1\}$.

Teorema 5.6.2. *Neka je V optimalni **binarni** prefiksni kod za slučajnu promenljivu X i neka je*

$$p_1 \geq p_2 \geq \cdots \geq p_{a-1} \geq p_a.$$

Tada su kodne reči v_{a-1} i v_a iste dužine.

Dokaz. [Šešelja, p.85] Pretpostavimo da je $|v_a| > |v_{a-1}|$. Konstruišimo novi kod V' kod koga je reč $|v_a|$ zamenjena drugom reči v'_a koja se dobija iz v_a kao prefiks dužine $|v_{a-1}|$. Proizvoljna reč v_i ne može biti prefiks reči v'_a zato što bi onda bila i prefiks reči v_a. Samim tim, kod V' je prefiksni i $\bar{n}_{V'} = \bar{n}_V - p_a(|v_a| - |v_{a-1}|) < \bar{n}_V$, što je kontradikcija sa pretpostavkom da je V optimalni kod. □

Teorema 5.6.3. *Za datu slučajnu promenljivu X postoji **binarni** (prefiksni) optimalni kod V takav da se kodne reči v_{a-1} i v_a razlikuju samo u poslednjem slovu.*

Dokaz. [Šešelja, p.85] Na osnovu prethodnog tvrđenja je $|v_a| = |v_{a-1}|$. Neka je, bez gubitka opštosti, $v_{a-1} = w0$. Ukoliko reč $w1$ već postoji u kodu V, stavljamo je na poslednje mesto. U suprotnom, reč v_a zamenimo sa $w1$. Na taj način dobijamo kod V' koji je prefiksni (ako je v_i prefiks od $w1$ onda je ili $v_i = w1$ ili je v_i prefiks od w pa i od $v_{a-1} = w0$) a koji ima istu srednju dužinu kodne reči kao i kod V. □

[7] Formalno gledano, sam kod ($V' = f'(A) = f(A) = V = \{v_1, v_2, \ldots, v_a\}$) je isti u oba slučaja, a ono što se razlikuje su funkcije kodiranja f i f'. Zbog toga ne navodimo skupovne oznake elemenata, ali imamo u vidu da se srednja dužina kodne reči definiše za kodiranje, pa su $\bar{n}_V$ i $\bar{n}_{V'}$ različiti.

5.7 Shannon-Fano kod

Pretpostavimo da je X izvor bez memorije. Izvor X je potpuno opisan skupom vrednosti $A = X_n(\Omega) = \{\alpha_1, \alpha_2, \ldots, \alpha_a\}$ kao i raspodelom verovatnoće $p_i = p(\alpha_i) = p_{X_n}(\alpha_i)$ (za svako $n \in \mathbb{N}$ i $i = 1, 2, \ldots, a$). Zato ćemo u nastavku ove sekcije pod pojmom **raspodela** podrazumevati uređen par (A, p) [8].

Opišimo sada postupak za konstrukciju jednog prefiksnog koda koji zadovoljava sva tri uslova iz prethodnog odeljka. S obzirom da su to potrebni ali ne i dovoljni islovi, dobijeni kod nije obavezno optimalan.

Algoritam 5.7.1. *(Shannon-Fano kod) Neka je $A = \{\alpha_1, \alpha_2, \ldots, \alpha_a\}$ skup simbola alfabeta izvora (skup vrednosti slučajne promenljive X) i neka su $p_1, p_2, \ldots, p_a$ odgovarajuće verovatnoće $p_i = p(\alpha_i)$.*

1. *Ukoliko je $a = 2$, prvom slovu (α_1) pridružiti simbol 0 a drugom (α_2) simbol 1. U suprotnom nastaviti.*
2. *Odrediti i tako da je apsolutna razlika suma $p_1 + p_2 + \ldots + p_i$ i $p_{i+1} + p_{i+2} + \ldots + p_a$ minimalna.*
3. *Slovima $\alpha_1, \alpha_2, \ldots, \alpha_i$ pridružimo simbol 0 a slovima $\alpha_{i+1}, \alpha_{i+2}, \ldots, \alpha_a$ simbol 1.*
4. *Primenimo isti postupak na svaki od novodobijenih skupova $\{\alpha_1, \alpha_2, \ldots, \alpha_i\}$ i $\{\alpha_{i+1}, \alpha_{i+2}, \ldots, \alpha_n\}$.*

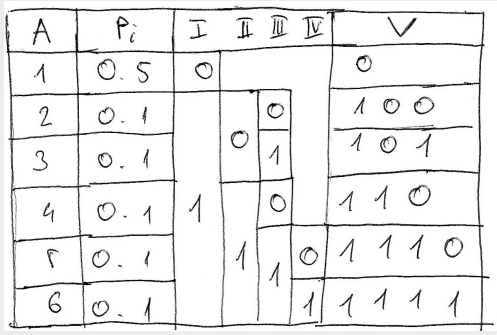
Nije teško pokazati da ovako konstruisan kod zadovoljava uslove prethodne tri teoreme.

Primer 5.7.1. Neka slučajna promenljiva X odgovara "nepravilnoj kockici" i ima raspodelu

$$X : \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 0.5 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \end{pmatrix}.$$

Primenom Shannon-Fano algoritma dobijamo:

[8] Kod Šešelje u knjizi, ovaj pojam se naziva **izvor informacija**.



U prvom koraku, vidimo da je za $i = 1$ razlika $p_1 - (p_2 + p_3 + p_4 + p_5 + p_6)$ minimalna, i jednaka 0. Prema tome, simbole delimo na sledeće grupe: $\{1\}$ i $\{2, 3, 4, 5, 6\}$. Prva grupa se sastoji samo od jednog simbola, a drugu nastavljamo da delimo. Minimalna apsolutna razlika je $|(p_2 + p_3) - (p_4 + p_5 + p_6)| = 0.1$, pa su nove grupe $\{2, 3\}$ i $\{4, 5, 6\}$. Novodobijene grupe delimo na sličan način u naredna dva koraka.

Odgovarajući kod je $V = \{0, 100, 101, 110, 1110, 1111\}$ čija je srednja dužina kodne reči jednaka

$$\bar{n}_V = 0.5 \cdot 1 + 0.1 \cdot 3 + 0.1 \cdot 3 + 0.1 \cdot 3 + 0.1 \cdot 4 + 0.1 \cdot 4 = 2.5.$$

Lako uočavamo kod zadovoljava sva tri potrebna uslova koja smo dokazali u prethodnom odeljku. Kasnije ćemo videti da je ovaj kod takođe i optimalan.

Primer 5.7.2. Posmatrajmo slučajnu promenljivu sa sledećom raspodelom:

$$X : \begin{pmatrix} \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \alpha_5 \\ 0.35 & 0.17 & 0.17 & 0.16 & 0.15 \end{pmatrix}$$

i konstruišimo Shannon-Fano kod. Konstrukcija je prikazana u sledećoj tabeli:

α_i	p_i	I	II	III	v_i	$\lvert v_i \rvert$
α_1	0.35	0	0		00	2
α_2	0.17		1		01	2
α_3	0.17	1	0		10	2
α_4	0.16		1	0	110	3
α_5	0.15			1	111	3

Srednja dužina kodne reči ovog koda je

$$\bar{n}_V = (0.35 + 0.17 + 0.17) \cdot 2 + (0.16 + 0.15) \cdot 3 = 2.31$$

Iako je ova vrednost blizu entropije izvora ($H(X) = 2.23$), dobijeni kod **nije optimalan**. Optimalni kod za ovu slučajnu promenljivu biće konstruisan u Primeru 5.8.1 (naredni odeljak).

Primer 5.7.3 (Šešelja, Primer 3.27, p. 85–86)**.**

Primer 5.7.4 (Šešelja, Primer 3.28, p. 87)**.**

5.8 Huffmanov algoritam za konstrukciju optimalnog koda

Konstrukcija Shannon-Fano koda zasniva se na principu deljenja (particionisanja) skupa simbola $A = \{\alpha_1, \alpha_2, \ldots, \alpha_a\}$ metodom "odozgo nadole". Najpre ceo skup delimo na dva skupa, pa onda svaki od ta dva skupa, po potrebi, dalje delimo. Videli smo da ovakav pristup ne daje uvek optimalan kod.

Sada ćemo razmotriti obrnuti pristup ("odozdo nagore"), gde ćemo u svakom koraku spajati po dva skupa simbola, sve dok ne dobijemo ceo skup A. Razmotrimo najpre jedan primer [9].

Primer 5.8.1. Posmatrajmo ponovo slučajnu promenljivu X sa raspodelom kao u Primeru 5.7.2:

$$X : \begin{pmatrix} \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \alpha_5 \\ 0.35 & 0.17 & 0.17 & 0.16 & 0.15 \end{pmatrix}$$

Simboli α_4 i α_5 imaju najmanje verovatnoće pojavljivanja. Ova dva simbola spajamo u novi simbol α_{45} sa verovatnoćom pojavljivanja $0.16 + 0.15 = 0.31$. Sada imamo sledeću raspodelu

$$X^{(1)} : \begin{pmatrix} \alpha_1 & \boxed{\alpha_{45}} & \alpha_2 & \alpha_3 \\ 0.35 & 0.31 & 0.17 & 0.17 \end{pmatrix}$$

Novi simbol α_{45} pišemo između simbola α_1 i α_2, kako bi verovatnoće pojavljivanja ostale u nerastućem poretku. Isti postupak primenjujemo dalje na

[9] Detaljna vizuelizacija Huffmanovog algoritma primenjenog na raspodelu iz ovog primera data je u filmu pod naslovom: `"(IC 4.1) Huffman coding - introduction and example"` koji se nalazi na `Youtube` (`www.youtube.com`) servisu.

novodobijenu raspodelu, sve dok ne dobijemo raspodelu sa tačno dva simbola:

$$X^{(2)}:\begin{pmatrix} \alpha_1 & \boxed{\alpha_{23}} & \alpha_{45} \\ 0.35 & 0.34 & 0.31 \end{pmatrix} \quad \longrightarrow \quad X^{(3)}:\begin{pmatrix} \boxed{\alpha_{2345}} & \alpha_1 \\ 0.65 & 0.35 \end{pmatrix}$$

Optimalan kod za poslednju slučajnu promenljivu (odnosno raspodelu) $X^{(3)}$ je očigledno $v_{2345} = 0$ i $v_1 = 1$. Pošto je $X^{(3)}$ dobijena iz $X^{(2)}$ spajanjem simbola α_{23} i α_{45}, kodne reči ova dva simbola formiraćemo tako što ćemo na kodnu reč v_{2345} dodati 0 odnosno 1. Tako dobijamo

$$v_{23} = v_{2345}0 = 10, \qquad v_{45} = v_{2345}1 = 11.$$

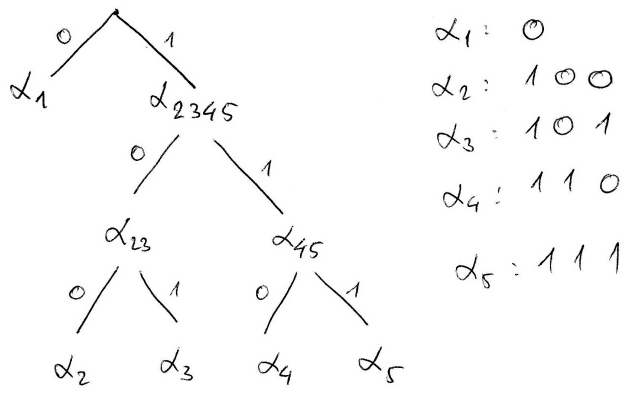
Postupak nastavljamo, tako što u svakom koraku formiramo po dve nove kodne reči za simbole koji su u tom koraku bili spojeni. Spajanje simbola možemo da predstavimo stablom, kao na slici 5.5.

Kodna reč za svaki simbol formira se kao (jedinstveni) put od korena stabla do odgovarajućeg simbola. Tako je na primer, kodna reč za simbol α_3 jednaka $v_5 = 101$.

Srednja dužina kodne reči ovog koda je

$$\bar{n}_V = 0.35 \cdot 1 + (0.17 + 0.17 + 0.16 + 0.15) \cdot 3 = 2.3$$

što je (doduše neznatno) manje nego u slučaju Shannon-Fano koda, datog u Primeru 5.7.2. Prema tome, Shannon-Fano kod definitivno nije optimalan u ovom slučaju, a u nastavku ćemo dokazati da je ovako konstruisan kod optimalan.



Slika 5.5: Kodno stablo za Huffmanov kod.

Postupak opisan u prethodnom primeru naziva se **Huffmanov algoritam**, dok se odgovarajući kod koji se tom prilikom dobija, naziva **Huffmanov kod**. Da bi dokazali da je Huffmanov kod optimalan, najpre dokazujemo da operacijom spajanja (odnosno razdvajanja) simbola ne narušava optimalnost koda. Neka je raspodela (A, p) dobijena spajanjem dva najmanje verovatna simbola raspodele (A', p'). Dokazujemo da ukoliko je V optimalan kod za (A, p), onda je kod V' koji se dobija kao u Primeru 5.8.1 optimalan za (A', p').

Teorema 5.8.1. *Neka je binarni kod* $V = \{v_1, v_2, \ldots, v_n\}$ *optimalan za raspodelu* (A, p). *Ako je* $p_j = q_0 + q_1$, *pri čemu je*

$$p_1 \geq p_2 \geq \cdots \geq p_{j-1} \geq p_{j+1} \geq \cdots \geq p_{a-1} \geq p_a \geq q_0 \geq q_1$$

onda je kod

$$V' = \{v_1, v_2, \ldots, v_{j-1}, v_{j+1}, \ldots, v_a, v_j 0, v_j 1\}$$

optimalan za raspodelu (A', p') *gde je*

$$A' = \{\alpha_1, \ldots, \alpha_{j-1}, \alpha_{j+1}, \ldots, \alpha_a, \alpha_{j0}, \alpha_{j1}\}$$

i $p'(\alpha_i) = p_i$ $(i \neq j)$, $p'(\alpha_{j0}) = q_0$, $p'(\alpha_{j1}) = q_1$.

Dokaz. [Šešelja, p. 88–89] Pošto je kod V prefiksan, sledi da se reči $v_j 0$ i $v_j 1$ ne pojavljuju u kodu V. Iz istog razloga, ove reči nisu prefiks nijedne druge kodne reči v_i $(i \neq j)$ iz V, niti je v_i njihov prefiks. Sledi da je kod V' dobro definisan i da je prefiksni kod. Srednja dužina kodne reči ovog koda jednaka je:

$$\bar{n}_{V'} = \sum_{i \neq j} n_i p_i + (n_j + 1)(q_0 + q_1) = \sum_{i=1}^{a} n_i p_i + p_j = \bar{n}_V + p_j.$$

Pretpostavimo da je W' optimalan kod za raspodelu (A', p'). Označimo kodne reči ovog koda na sledeći način:

$$W' = \{w_1, w_2, \ldots, w_{j-1}, w_{j+1}, \ldots, w_a, w_{a,0}, w_{a,1}\}$$

Bez gubitka opštosti, na osnovu Teoreme 5.6.3, možemo da pretpostavimo da se poslednje dve kodne reči $w_{a,0}$ i $w_{a,1}$ koda W' razlikuju samo u poslednjoj cifri. Drugim rečima, da je $w_{a,0} = w0$ a $w_{a,1} = w1$. Tada je kod

$$W = \{w_1, w_2, \ldots, w_{j-1}, w, w_{j+1}, \ldots, w_a\}$$

prefiksni [10] i važi

$$\bar{n}_{W'} = \bar{n}_W + p_j.$$

[10] Maksimalna dužina kodne reči u W' je $|w_{a,0}| = |w_{a,1}| = |w| + 1$. Odatle sledi da ukoliko je w prefiks neke kodne reči w_i, onda je $w_i = w0$ ili $w_i = w1$. Ovo je nemoguće, jer u kodu W' ne postoje dve jednake kodne reči (u ovom slučaju bi to bile w_i i jedna od $w_{a,0}$ ili $w_{a,1}$).

Pošto je V optimalan kod, sledi da je $\bar{n}_V \leq \bar{n}_W$ odnosno da je

$$\bar{n}_{V'} = \bar{n}_V + p_j \leq \bar{n}_W + p_j = \bar{n}_{W'}.$$

Suprotna nejednakost sledi iz optimalnosti koda W', pa dobijamo da je $\bar{n}_{V'} = \bar{n}_{W'}$, odnosno da je V' takođe optimalan kod za raspodelu (A', p'). □

Transformacija raspodela u prethodnoj teoremi može se predstaviti sledećom šemom:

$$\begin{pmatrix} \alpha_1 & \cdots & \alpha_j & \cdots & \alpha_a \\ p_1 & \cdots & p_j & \cdots & p_a \end{pmatrix} \to \begin{pmatrix} \alpha_1 & \cdots & \alpha_{j-1} & \alpha_{j+1} & \cdots & \alpha_a & \alpha_{j0} & \alpha_{j1} \\ p_1 & \cdots & p_{j-1} & p_{j+1} & \cdots & p_a & q_0 & q_1 \end{pmatrix}$$

Opisaćemo postupak konstrukcije optimalnog (binarnog) koda Huffmanovim algoritmom.

Algoritam 5.8.1. *(Huffmanov algoritam za konstrukciju optimalnog koda)*

1. *Ako je* $|A| = 2$*, konstruisati kod* $V = \{0, 1\}$*. U suprotnom, nastaviti.*
2. *Sortirati slova* $\alpha_1, \alpha_2, \ldots, \alpha_a$ *tako da je* $p_1 \geq p_2 \geq \cdots \geq p_a$*.*
3. *Spojiti slova* α_{a-1} *i* α_a *u jedno* $\alpha_{a-1,a}$ *i formirati novu raspodelu* (A', p') *tako da je* $A' = \{\alpha_1, \alpha_2, \ldots, \alpha_{a-2}, \alpha_{a-1,a}\}$ *kao i* $p'(\alpha_i) = p_i$*,* $p'(\alpha_{a-1,a}) = p_{a-1} + p_a$*.*
4. *Primeniti algoritam rekurzivno na raspodelu* (A', p')*. Neka je kod* $V' = \{v_1, v_2, \ldots, v_{a-2}, v_{a,a-1}\}$ *dobijen na taj način.*
5. *Konstruisati kod* $V = \{v_1, v_2, \ldots, v_{a-2}, v_{a,a-1}0, v_{a,a-1}1\}$*.*

Huffmanov algoritam konstruiše niz raspodela

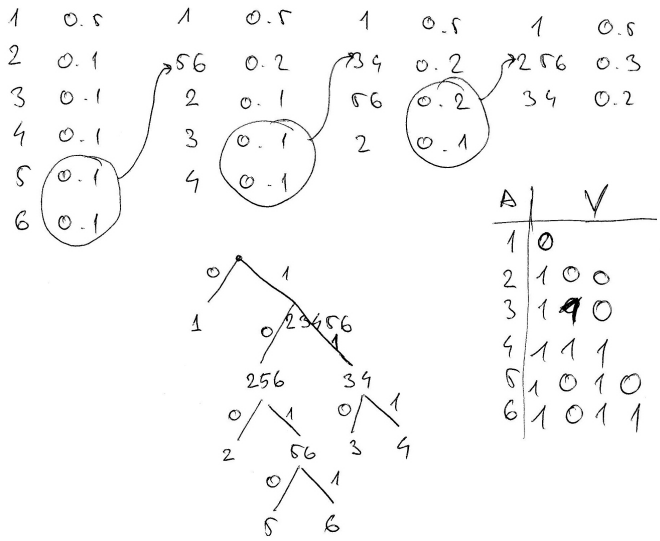
$$(A, p), (A^{(1)}, p^{(1)}), (A^{(2)}, p^{(2)}), \ldots, (A^{(a-2)}, p^{(a-2)})$$

i odgovarajući niz kodova

$$V, V^{(1)}, V^{(2)}, \ldots, V^{(a-2)}.$$

Pritom je $|A^{(i)}| = a - i$. Pošto je kod $V^{(a-2)} = \{0, 1\}$ očigledno optimalan, primenom Teoreme 5.8.1 sledi da su svi kodovi $V^{(i)}$ optimalni za odgovarajuće raspodele $(A^{(i)}, p^{(i)})$ $(i = 1, 2, \ldots, a-2)$, kao i da je kod V optimalan za raspodelu (A, p).

Primer 5.8.2 (Šešelja, Primer 3.30, p. 89-90)**.**



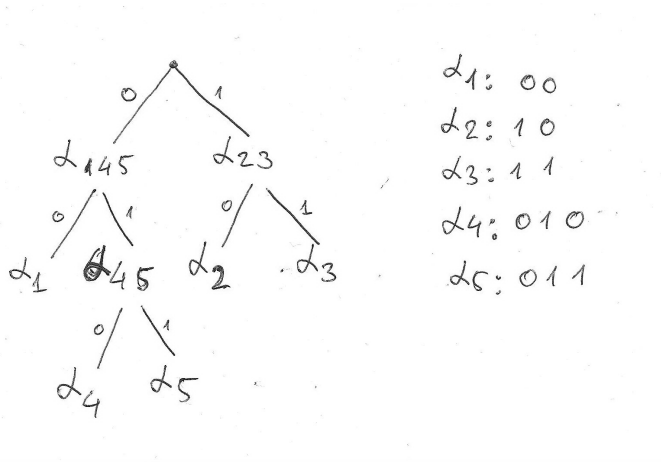
Slika 5.6: Primer konstruisanja optimalnog koda Huffmanovim algoritmom.

Primer 5.8.3. Odredimo sada optimalni kod za nepravilnu kocku iz primera 5.7.1 primenom Huffmanovog algoritma (slika 5.7). Vidimo da smo dobili isti kod kao i primenom Shannon-Fano algoritma.

Primer 5.8.4. Konstruišimo sada Huffmanov i Shannon-Fano kod za sledeću slučajnu promenljivu:

$$X : \begin{pmatrix} \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \alpha_5 \\ 0.35 & 0.2 & 0.2 & 0.15 & 0.10 \end{pmatrix}$$

Vidimo da su verovatnoće pojavljivanja simbola bliske onima iz Primera 5.7.2 i (odnosno Primera 5.8.1). Međutim, u ovom slučaju, dobija se kod dat na slici **??**, koji je u isto vreme i Shannon-Fano kod.



Slika 5.7: Primer konstruisanja optimalnog koda Huffmanovim algoritmom.

Primer 5.8.5. Pogledati film: `(IC 4.2) Huffman coding - more examples` na `www.youtube.com`.

5.9 Univerzalni kodovi: LZ77 i LZ78 algoritam

Videti [Cover, sect. 13.4, p. 440–443] i prezentaciju `Kompresija podataka - Da li su bitovi stišljivi`.

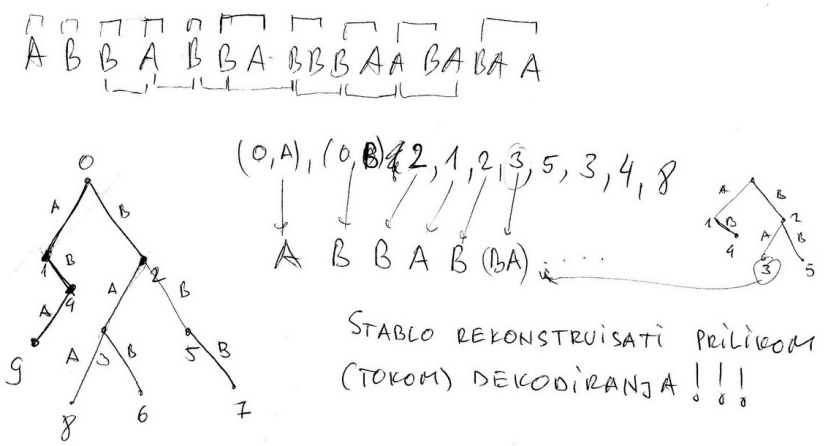
Što se tiče LZW algoritma (slika 5.8), tu je ideja da se umesto npr. $(3, A)$, zapiše samo 3, a da se ovo A tretira kao početak sledećeg elementa. Naravno, novi čvor stabla se kreira kao i kod LZ78. Stablo se rekonstruiše tokom procesa dekodiranja.

Film: `Lempel-Ziv-Welch Compression Algorithm - Tutorial`

5.10 Dodatak

5.10.1 Huffmanov algoritam za proizvoljnu bazu

Postupak je potpuno isti i u slučaju $b \geq 3$, samo što se b najmanje verovatnih simbola kombinuju u svakom koraku. Pritom je potrebno da u poslednjem koraku ostane tačno b simbola.



Slika 5.8: Ilustracija LZW algoritma

Primetimo da se ovde u svakom koraku broj simbola smanji za $b-1$. Ako je poslednjem koraku imamo b simbola, to znači da je $a-b$ simbola ukupno združeno u prethodnim koracima. Dakle, $a-b$ mora da bude deljivo sa $b-1$. Ukoliko to nije ispunjeno, dodaju se fiktivna slova, koja se pojavljuju sa verovatnoćom 0.

Na sličan način mogu da se dokažu i druga dva potrebna uslova za optimalnost (Teorema 5.6.2 i Teorema 5.6.3) ukoliko je $b \geq 3$ i $a-b$ je deljivo sa $b-1$.

Za $b > 2$ važi sledeća generalizacija Teoreme 5.6.2.

Teorema 5.10.1. *Neka je V optimalni prefiksni kod za slučajnu promenljivu X za koji je zbir dužina kodnih reči $n_1 + n_2 + \ldots + n_a$ minimalan. Tada su kodne reči $v_{a-d}, v_{a-d+1}, \ldots, v_a$ (ukupno $d+1$) iste dužine gde je*

$$d = \begin{cases} b-1, & b-1 \mid a-1 \\ (a-1) \bmod (b-1), & \textit{inače} \end{cases}$$

Dokaz. Označimo sa m_k broj kodnih reči dužine k i neka je n dužina najduže kodne reči (v_a). Potrebno je dokazati da je $m_n \geq d+1$. To je očigledno ispunjeno za $m_n \geq b$, pa ćemo stoga pretpostaviti da je $m_n \leq b-1$.

Ako je $m_n = 1$, onda poslednju kodnu reč možemo skratiti za jedan simbol (kao u Teoremi 5.6.2) tako da se dobije optimalniji kod (ili kod sa istim $\bar{n}_V$

koji ima manji zbir dužina kodnih reči) koji je ujedno prefiksni. Prema tome, važi $m_n > 1$.

Neka je:

$$S_1 = \sum_{k=1}^{n-1} m_k b^{-k} = \left(\overline{0.m'_1 m'_2 \cdots m'_n}\right)_b$$

gde su m'_j cifre u reprezentaciji broja S_1 u sistemu sa osnovom b. Na osnovu Kraftove nejednakosti je $S_1 + m_n b^{-n} \leq 1 < 1$, pa broj S_1 ima cifre u sistemu sa osnovom b samo desno od tačke.

Pretpostavimo da postoji indeks j takav da je $m'_j < b - 1$. Tada je

$$S_1 + b^{-j} + (b-1)b^{-n} = \left(\overline{0.m'_1 m'_2 \cdots (m'_j + 1) \cdots m'_n (b-1)}\right)_b < 1$$

pa je $S_1 < 1 - b^{-j} - (b-1)b^{-n}$. Sada ćemo konstruisati nov kod $V^{(1)}$ tako što ćemo jednu kodnu reč dužine n zameniti kodnom reči dužine j. Za novi kod važi $m_k^{(1)} = m_k$ za $k \neq j, n$ kao i $m_j^{(1)} = m_j + 1$ i $m_n^{(1)} = m_n - 1$ ($m_n \geq 1$ jer postoji kodna reč dužine n). Postojanje koda $V^{(1)}$ dokazujemo na osnovu Kraftove nejedakosti:

$$\begin{aligned} S^{(1)} &= S + b^{-j} - b^{-n} = S_1 + (m_n - 1)b^{-n} + b^{-j} \\ &< 1 - b^{-j} - (b-1)b^{-n} + (m_n - 1)b^{-n} + b^{-j} \\ &= 1 - (b - m_n)b^{-n} < 1 \end{aligned}$$

jer je (po pretpostavci) $m_n < b$. Novi kod $V^{(1)}$ ima manju srednju dužinu kodne reči od koda V za $\bar{n}_V - \bar{n}_{V^{(1)}} = p_a(n - j)$. Ukoliko je $p_a = 0$, onda ima manji ukupan zbir dužina za $n - j$. Ovo je kontradikcija sa pretpostavkom optimalnosti koda V (kao i minimalnog zbira svih kodnih reči, među optimalnim kodovima).

Prema tome važi $m'_1 = m'_2 = \cdots = m'_{n-1} = b - 1$. Tada je $S_1 = 1 - b^{-(n-1)}$ pa je $b^{n-1}S_1 = b^{n-1} - 1 \equiv_{b-1} 0$. Sa druge strane, važi

$$b^{n-1}S_1 = m_1 b^{n-2} + m_2 b^{n-3} + \ldots + m_{n-1} \equiv_{b-1} m_1 + m_2 + \ldots + m_{n-1}$$

odnosno $m_1 + m_2 + \ldots + m_{n-1} \equiv_{b-1} 0$. Iz $a = m_1 + m_2 + \ldots + m_{n-1} + m_n$ sledi da je $m_n \equiv_{b-1} a$, a pošto je $0 \leq m_n - 1 \leq b - 2$ ($1 \leq m_n \leq b - 1$ po pretpostavci) sledi da je $m_n = (a-1) \bmod (b-1) + 1$. Primetimo sada da ako $b - 1$ deli $a - 1$, onda je $m_n = 1$, što je kontradikcija. Dakle, u ovom slučaju mora biti $m_n \geq b$. Ovim je dokaz završen. □

Primer 5.10.1 (Šešelja, Primer 3.31, p. 91)**.**

Film: `(IC 4.3) B-ary Huffman codes`

5.10.2 Primena Huffmanovog metoda za težinsku minimizaciju

Film: `(IC 4.4) Weighted minimization with Huffman coding`

5.10.3 Adaptivni Huffmanov algoritam

Film: `FGK (Adaptive huffman coding)`

Film: `Adaptive Huffman Encoding`

5.10.4 Optimalnost Shannonovog koda po drugim kriterijumima

Podsetimo se da smo Shannonov kod definisali kao prefiksni kod čije su dužine kodnih reči jednake $n_i = \lceil -\log p_i \rceil$ za svako $i = 1, 2, \ldots, a$. U dokazu Teoreme 5.5.2 pokazali smo da $n_1, n_2, \ldots, n_a$ zadovoljavaju Kraftovu nejednakost, pa na osnovu Teoreme 5.4.1 možemo da konstruišemo prefiksni kod V_S koji ima baš ove srednje dužine kodnih reči. Takođe, u dokazu iste teoreme, pokazali smo da je

$$H(X) \leq \bar{n}_{V_S} \leq H(X) + 1$$

gde je X izvor bez memorije. Pritom, ukoliko je $\log_2 p_i$ ceo broj, za svako $i = 1, 2, \ldots, a$, ovaj kod je optimalan i važi $\bar{n}_{V_S} = H(X)$.

Činjenica je da nijedan kod ne može biti optimalan za svaku (konačnu) sekvencu simbola ulaznog alfabeta. Kriterijum koji smo do sad koristili bio je srednja dužinu kodne reči $\bar{n}_V$. Sada ćemo pokazati da se Shannonov kod podjednako dobro ponaša iako promenimo kriterijum upoređivanja.

Sledeća teorema tvrdi da verovatnoća da Shannonov kod V bude lošiji od nekog drugog koda V' za c bita, opada (bar) eksponencijalno sa c.

Teorema 5.10.2. *Neka je f kodiranje koje odgovara Shannonovom kodu a f' nekom drugom kodu. Neka su $l(x) = |f(x)|$ i $l'(x) = |f'(x)|$, funkcije koje predstavljaju dužine kodnih reči za kodiranje f odnosno f'. Tada važi*

$$P(l(X) \geq l'(X) + c) \leq \frac{1}{2^{c-1}}.$$

Dokaz. [Cover, p. 131] □

Teorema 5.10.3. *Pod uslovima prethodne teoreme, ukoliko su $\log_2 p_i$ celi brojevi za svako $i = 1, 2, \ldots, a$ važi i:*

$$P(l(X) < l'(X)) \geq P(l(X) > l'(X)).$$

Pritom, jednakost važi ako i samo ako je $l'(x) = l(x)$ za svako $x \in A$.

Dokaz. [Cover, p. 132] □

Posledica 5.10.4. *Ukoliko bar jedna vrednost $\log_2 p_i$ nije ceo broj, onda je*

$$\mathbb{E}\operatorname{sgn}(l(X) - l'(X) - 1) \leq 0.$$

5.11 Za dalje čitanje

- Generisanje proizvoljne distribucije bacanjem novčića. (Cover, sect. 5.11, p. 134)
- Aritmetičko kodiranje. (IC, 5.1-5.14)

Glava 6

Komunikacijski kanali

6.1 Matematički model kanala

Kada kažemo da osoba ili objekat A komunicira sa osobom ili objektom B, podrazumevamo da A i B razmenjuju neke poruke (u jednom ili oba smera). Ta razmena poruka vrši se preko medijuma koji se naziva *komunikacijski kanal*. Na sam transfer mogu da utiču različiti faktori, čijim delovanjem poslata informacija može da se izgubi ili promeni. Sve ove faktore nazivamo *smetnjama u kanalu* ili *šumom*.

S obzirom da postoje (fizički) različite vrste kanala (optički, bežični, satelitski, itd.) definisaćemo opšti matematički model kanala koji može da se prilagodi postojećim tipovima realnih kanala.

Definicija 6.1.1. ***Diskretni (stacionarni) komunikacijski kanal bez memorije*** $K = (\mathcal{X}, p(\cdot|\cdot), \mathcal{Y})$ *definisan je ulaznim alfabetom* $\mathcal{X} = \{x_1, x_2, \ldots, x_a\}$, *izlaznim alfabetom* $\mathcal{Y} = \{y_1, y_2, \ldots, y_b\}$ *i matricom*

$$\Pi = [p(y_j|x_i)]_{1\leq i\leq a, 1\leq j\leq b}$$

gde je $p(y_j|x_i)$ *uslovna verovatnoća da je primljen simbol* y_j, *ukoliko je poslat simbol* x_i. *Pisaćemo i* $K = (\mathcal{X}, p(y|x), \mathcal{Y})$ *da bi naglasili na koje slučajne promenljive se odnosi uslovna raspodela* $p(y|x)$.

Ova definicija podrazumeva da izlaz iz kanala y_n zavisi samo od odgovarajućeg ulaza x_n, a ne zavisi od prethodnih ulaza $x_1, x_2, \ldots, x_{n-1}$ niti od prethodnih izlaza $y_1, y_2, \ldots, y_{n-1}$. Moguće je posmatrati i najopštiji slučaj, gde je dat niz raspodela $p(y_n|x_1, \ldots, x_{n-1}, y_1, \ldots, y_{n-1})$, ali mi ćemo da se zadržimo na diskretnim stacionarnim kanalima bez memorije, za koje ćemo u nastavku jednostavno koristiti termin **kanal**.

Neka je $X = (X_n)_{n\in\mathbb{N}}$ izvor informacija koji emituje simbole (predstavlja niz simbola koje šalje pošiljalac) iz skupa $\mathcal{X}$. Izvor X, i matrica Π jednoznačno opisuju niz slučajnih promenljivih $Y = (Y_n)_{n\in\mathbb{N}}$ koji predstavlja simbole koji se dobijaju na izlazu (tj. koje dobija primalac).

Pretpostavimo da je i izvor X stacionarni izvor bez memorije. Tada sve slučajne promenljive X_n imaju istu raspodelu, pa umesto niza možemo posmatrati samo jednu promenljivu X. Isto važi i za Y.

Raspodela slučajne promenljive Y određena je izrazom

$$p(y_j) = \sum_{i=1}^{a} p(y_j|x_i)p(x_i).$$

6.2 Kapacitet kanala

Neformalno rečeno, kapacitet nekog kanala je maksimalna količina podataka koju je moguće preneti kroz kanal pri jednom slanju. Ako slučajna promenljiva X predstavlja simbol koji saljemo a slučajna promenljiva Y simbol koji primamo, tada je upravo međusobna informacija $I(X, Y)$ ove dve promenljive mera za količinu informacija koja "prolazi" kroz kanal. Kapacitet kanala je onda jednak maksimalnoj mogućoj međusobnoj informaciji ove dve slučajne promenljive.

Definicija 6.2.1. *Kapacitet kanala K jednak je maksimalnoj međusobnoj informaciji $I(X, Y)$, gde se maksimum uzima po svim raspodelama slučajne promenljive X. Drugim rečima*

$$C = \max_{p(x)} I(X, Y). \tag{6.1}$$

Nije teško videti da je raspodela $p(x)$ određena a-dimenzionalnim vektorom $(p(x_1), p(x_2), \ldots, p(x_a))$, čije su sve komponente nenegativne i čiji je zbir jednak 1. Prema tome, ovaj skup je zatvoren i ograničen. Pošto je $I(X, Y)$ neprekidna funkcija, sledi da se maksimum u izrazu (6.1) dostiže.

Primetimo da na osnovu Shannonove nejednakosti sledi da je $I(X, Y) \geq 0$ odnosno $C \geq 0$. Jednakost se dostiže ukoliko je $I(X, Y) = 0$ za svaku raspodelu slučajne promenljive X, odnosno $H(Y) = H(Y|X)$. Ovo važi ukoliko je

$$p(y_j) = p(y_j|x_i), \qquad i = 1, 2, \ldots, a, \quad j = 1, 2, \ldots, b.$$

Drugim rečima, moraju sve vrste matrice Π da budu jednake. Ovakav kanal se naziva **beskorisnim** zato što istim nije moguće preneti nikakvu informaciju.

6.3 Primeri izračunavanja kapaciteta kanala

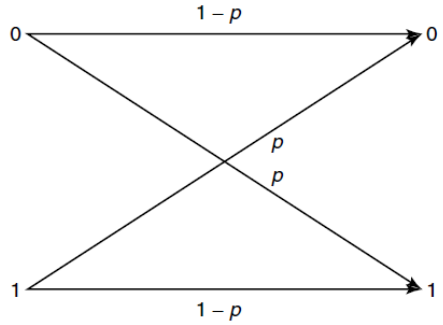
Pokazaćemo na nekoliko konkretnih primera kako se računa kapacitet kanala i šta ova veličina predstavlja u praksi.

6.3.1 Binarni simetrični kanal

Ovaj kanal je definisan dvoelementnim alfabetima $\mathcal{X} = \mathcal{Y} = \{0, 1\}$ i matricom:

$$\Pi = \begin{bmatrix} 1-\alpha & \alpha \\ \alpha & 1-\alpha \end{bmatrix}$$

Komunikacija se sastoji u tome da se na izvoru šalje 0 ili 1, a na odredištu se pritom prima takođe 0 ili 1. Do greške dolazi kada se pošalje 0 a primi 1 ili obrnuto. Pritom, bez poznavanja originalnog niza nije moguće utvrditi da li je došlo do greške ili nije. Kasnije ćemo pokazati da se ovaj kanal može koristiti za komunikaciju kod koje je verovatnoća greške proizvoljno mala.



Slika 6.1: Binarni simetrični kanal.

Izračunajmo sada kapacitet ovog kanala. Srednja uzajamna informacija $I(X, Y)$ može da se izračuna na sledeći način

$$\begin{aligned} I(X,Y) &= H(Y) - H(Y|X) \\ &= H(Y) - p_X(0)H(Y|X=0) - p_X(1)H(Y|X=1) \\ &= H(Y) - p_X(0)H(1-\alpha,\alpha) - p_X(1)H(\alpha,1-\alpha) \\ &= H(Y) - H(\alpha,1-\alpha) \\ &\leq 1 - H(\alpha,1-\alpha) \end{aligned}$$

Poslednja nejednakost važi zato što Y uzima samo dve vrednosti, pa je maksimalna vrednost entropije manja ili jednaka $H(Y) \leq \log_2 2 = 1$. Ova vrednost

je dostižna ukoliko postoji raspodela p_X takva da je $p_Y(0) = p_Y(1) = 1/2$. To očigledno važi baš za $p_X(0) = p_X(1) = 1/2$, odnosno za uniformnu raspodelu ulaznih simbola.

Dakle, kapacitet binarnog simetričnog kanala jednak je

$$C = \max_{p_X} H(Y) - H(\alpha, 1-\alpha) = 1 - H(\alpha, 1-\alpha).$$

Vidimo da za $\alpha = 0, 1$ kanal ima kapacitet $C = 1$ (kanal je **bešumni**), dok je za $\alpha = 1/2$ kapacitet $C = 0$, odnosno kanal je beskorisan.

Zanimljivo je uočiti i da je kanal idealan ukoliko je $\alpha = 1$. Tada je na prvi pogled, verovatnoća greške jednaka 1, tj. kanal uvek prenosi pogrešan podatak. Međutim, onda ćemo jednostavnim komplementiranjem podatka na izlazu dobiti ono što je poslato. Nasuprot tome, kada je $\alpha = 1/2$, na izlazu dobijamo 0 i 1 sa verovatnoćom 1/2, **nezavisno od toga šta je poslato**. Tada nije moguće nikakvu informaciju preneti kanalom.

6.3.2 Kanal sa nepreklapajućim izlazima

Kod ovog kanala, za svaku primljenu vrednost y postoji jedinstvena vrednost x koja je tom prilikom mogla biti poslata. Samim tim, ulaznim simbolima odgovaraju disjunktni skupovi izlaznih simbola koje taj ulazni simbol može da generiše. Drugim rečima, skupovi

$$A(x) = \{y \in \mathcal{Y} \mid p(y|x) \neq 0\}$$

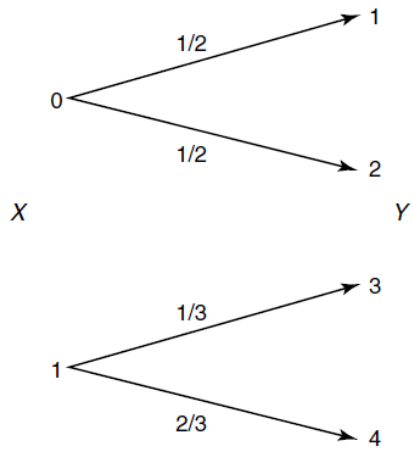
su takvi da je $A(x_1) \cap A(x_2) = \emptyset$ za $x_1 \neq x_2$. Zato je i ovaj kanal **bešumni**, odnosno važi $H(X|Y) = 0$. Odavde dalje sledi:

$$\begin{aligned} C &= \max_{p(x)} I(X,Y) = \max_{p(x)} (H(X) - H(X|Y)) \\ &= \max_{p(x)} H(X) = \log_2 a. \end{aligned}$$

Primer takvog kanala je:

$$\mathcal{X} = \{0,1\}, \quad \mathcal{Y} = \{1,2,3,4\}, \quad \Pi = \begin{bmatrix} 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 1/3 & 2/3 \end{bmatrix}.$$

Na slici 6.2 data je grafička interpretacija ovog kanala. Jasno je da ako na prijemu stigne 1 ili 2, poslat je simbol 0, a ako stigne 3 ili 4, poslat je simbol 1.



Slika 6.2: Kanal sa nepreklapajućim izlazima.

6.3.3 Binarni brišući kanal

Ovaj kanal opisuje situaciju kada primalac ima mehanizam da utvrdi da li je došlo do greške pri prenosu. Kod binarnog simetričnog kanala je to nemoguće, zato što npr. primljeni simbol 0 može da potiče od ispravno poslatog simbola 0, kao i pogrešno poslatog simbola 1.

Ulazni alfabet binarnog brišućeg kanala je $\mathcal{X} = \{0, 1\}$ a izlazni $\mathcal{Y} = \{0, e, 1\}$. Simbol e ovde označava da je došlo do greške. Matrica Π je data sa:

$$\Pi = \begin{bmatrix} 1-\alpha & \alpha & 0 \\ 0 & \alpha & 1-\alpha \end{bmatrix}$$

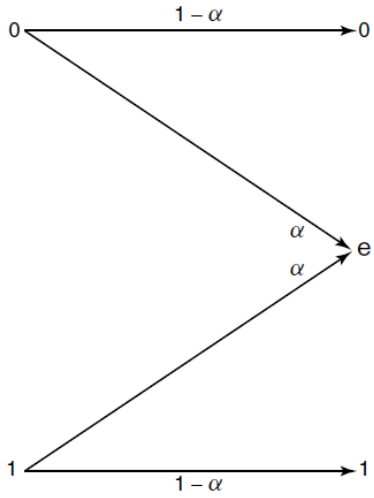
Dakle, do greške dolazi sa verovatnoćama $p(e|0) = p(e|1) = \alpha$ i ona se manifestuje simbolom e. Ukoliko pristigne simbol 0 ili 1, to znači da je slanje bilo ispravno. [1]

Uslovna raspodela za $p_{X|Y}(x|y)$ jednaka je

$$p_{X|Y}(0|1) = p_{X|Y}(1|0) = 0, \quad p_{X|Y}(0|0) = p_{X|Y}(1|1) = 1,$$
$$p_{X|Y}(0|e) = p_X(0), \quad p_{X|Y}(1|e) = p_X(1).$$

Samim tim je $H(X|Y = 0) = H(X|Y = 1) = 0$ a $H(X|Y = e) = H(X)$. Ukoliko su pristigli 0 ili 1, onda sigurno znamo i da su poslati 0 odnosno 1.

[1] Idealan primer binarnog brišućeg kanala u praksi ne postoji. Međutim, ukoliko se koriste zaštitni kodovi, o kojima će biti reči kasnije, onda je sa veoma velikom verovatnoćom moguće utvrditi da li je došlo do greške pri prenosu. Ukoliko jeste, onda najčešće pošiljalac traži ponavljanje poruke.



Slika 6.3: Binarni brišući kanal.

Ukoliko pristigne e, onda nemamo nikakvu (dodatnu) informaciju o tome šta je moglo biti poslato. Samim tim je

$$\begin{aligned} H(X|Y) &= p_Y(0)H(X|Y=0) + p_Y(1)H(X|Y=1) + p_Y(e)H(X|Y=e) \\ &= p_Y(e)H(X|Y=e) = \alpha H(X) \end{aligned}$$

pa je

$$\begin{aligned} C &= \max_{p(x)}(H(X) - H(X|Y)) = \max_{p(x)}(H(X) - \alpha H(X)) \\ &= (1-\alpha)\max_{p(x)} H(X) = 1 - \alpha \end{aligned}$$

Očigledno da je kapacitet maksimalan za $\alpha = 0$ (nema grešaka) a jednak nuli za $\alpha = 1$.

Primer 6.3.1. Pretpostavimo sada da imamo binarni brišući kanal i da pritom pošiljalac ima informaciju da li je došlo do greške u prenosu ili nije. Neka je $x_1x_2\cdots x_m$ ($x_i \in \mathcal{X} = \{0,1\}$) poruka koju pošiljalac želi da pošalje i neka je došlo do greške pri slanju bita x_i. U tom slučaju, pošiljalac ce ponoviti slanje istog bita, i to će nastaviti sve dok se slanje ne izvrši uspešno.

Označimo sa N_i ($i = 1, 2, \ldots, m$) ukupan broj puta kada je poslat bit x_i. Tada je $N = N_1 + N_2 + \ldots + N_m$ ukupan broj poslatih bitova za celu poruku. Očigledno su i N_i ($i = 1, 2, \ldots, m$) i N slučajne promenljive, pa nas ovom prilikom zanimaju njihove prosečne vrednosti (tj. očekivanja). Pošto je

svako slanje nezavisno od prethodnog, i pošto je kanal bez memorije, tada je $\mathbb{E}N_i = \mathbb{E}N_j$ $(i, j = 1, 2, \ldots, m)$ i $\mathbb{E}N = m\mathbb{E}N_1$.

Da bi izračunali $\mathbb{E}N_1$, odredimo verovatnoću događaja da je bit x_1 poslat tačno k puta. Ako je verovatnoća greške jednaka α, tada je

$$p_k = P(N_1 = k) = \alpha^{k-1}(1 - \alpha).$$

Na osnovu ovoga imamo da je očekivanje jednako

$$\mathbb{E}N_1 = \sum_{k=1}^{+\infty} k\alpha^{k-1}(1 - \alpha) = \frac{1}{1 - \alpha}.$$

S obzirom da je $\mathbb{E}N$ bitova poslato a samo m primljeno, zaključujemo da je procenat uspešno prenesenih bitova jednak

$$\frac{m}{\mathbb{E}N} = 1 - \alpha = C.$$

Sa druge strane, ovo je ujedno i procenat "korisnih" bitova u nizu bitova koji su poslati (ostali predstavljaju ponovljene bitove). Prema tome, činjenicu da bitove bez greške prenosimo "nepouzdanim" kanalom, plaćamo unošenjem redundanse u ulazni niz.

Ovim primerom smo pokazali da je $1 - C$ dovoljan procenat redundantnih bitova koji omogućava prenos informacija bez greške kroz binarni brišući kanal sa povratnom spregom. U nastavku ćemo pokazati da je ovo moguće i bez (veštačke) pretpostavke o postojanju kanala povratne sprege, pod uslovom da se umesto prenosa bez greške, posmatra prenos sa proizvoljno malom verovatnoćom greške.

6.4 Simetrični kanali

Jedna od najvažnijih klasa kanala koji se koriste u praksi su **simetrični** kanali. Ideja za uvođenje simetričnih kanala je činjenica da je greška najčešće "na isti način" raspodeljena, koji god ulazni simbol da se pošalje odnosno koji god simbol da se primi.

Definicija 6.4.1. *Kanal je* ***simetričan po ulazu****, ako su sve vrste matrice* Π *obrazovane permutacijama elemenata prve vrste. Kanal je* ***simetričan po izlazu****, ako su sve kolone matrice* Π *obrazovane permutacijama elemenata prve kolone. Kanal je* ***simetričan****, ako je simetričan i po ulazu i po izlazu.*

Primer 6.4.1. Posmatrajmo kanale definisane matricama:

$$\Pi_1 = \begin{bmatrix} 0.5 & 0.3 & 0.2 \\ 0.3 & 0.5 & 0.2 \end{bmatrix}, \quad \Pi_2 = \begin{bmatrix} 0.5 & 0.5 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \Pi_3 = \begin{bmatrix} 0.3 & 0.2 & 0.5 \\ 0.5 & 0.3 & 0.2 \\ 0.2 & 0.5 & 0.3 \end{bmatrix}.$$

Kanal (definisan matricom) Π_1 je simetričan po ulazu, Π_2 simetričan po izlazu, dok je Π_3 simetričan.

Kao što ćemo videti u nastavku, za simetrične kanale moguće je izvesti izraz za kapacitet C u zatvorenom obliku.

Lema 6.4.1. *Ako je kanal simetričan po ulazu, uslovna entropija $H(Y|X)$ ne zavisi od verovatnoća ulaznih signala i jednaka je*

$$H(Y|X) = -\sum_{i=1}^{b} r_i \log_2 r_i$$

gde su $r_1, r_2, \ldots, r_b$ elementi prve vrste matrice kanala.

Dokaz. [Šešelja, p. 60-61] Iz definicionog izraza sledi da je tražena uslovna entropija jednaka

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

gde je

$$H(Y|x) = -\sum_{y \in \mathcal{Y}} p(y|x) \log_2 p(y|x).$$

Vrednosti $p(y|x)$ za $y \in \mathcal{Y}$ formiraju vrstu matrice kanala Π koja odgovara ulaznom simbolu x. Iz simetričnosti kanala po ulazu zaključujemo da $H(Y|x)$ ne zavisi od x (sumiranje se obavlja nad istim skupom brojeva) i da je

$$H(Y|x) = -\sum_{i=1}^{b} r_i \log_2 r_i.$$

Pošto prethodni izraz ne zavisi od x, zaključujemo da je to u isto vreme i izraz za uslovnu entropiju $H(Y|X)$. □

Lema 6.4.2. *Ako su svi ulazni signali jednako verovatni i ako je kanal simetričan po izlazu, onda su i izlazni signali jednako verovatni.*

Dokaz. [Šešelja, p. 61] Iz Bajesove formule i pretpostavke da je $p(x) = 1/a$ za svako $x \in \mathcal{X}$ (uniformna raspodela, $a = |X|$) sledi

$$p(y) = \sum_{x \in \mathcal{X}} p(y|x)p(x) = \frac{1}{a} \sum_{x \in \mathcal{X}} p(y|x).$$

Suma vrednosti $p(y|x)$ po $x \in \mathcal{X}$ predstavlja sumu elemenata u koloni matrice kanala Π koja odgovara elementu $y \in \mathcal{Y}$. Pošto je kanal simetričan po izlazu, sve kolone imaju isti skup elemenata, pa su odgovarajuće sume jednake, kao i verovatnoće $p(y)$. □

Teorema 6.4.3. *Ako je kanal simetričan, njegov kapacitet iznosi*

$$C = \log_2 b + \sum_{i=1}^{b} r_i \log_2 r_i.$$

gde su $r_1, r_2, \ldots, r_b$ *elementi svake vrste matrice* Π.

Dokaz. [Šešelja, p. 61] Podsetimo se da je međusobna informacija slučajnih promenljivih X i Y jednaka $I(X,Y) = H(Y) - H(Y|X)$. Na osnovu Leme 6.4.1 sledi da $H(Y|X)$ ne zavisi od raspodele slučajne promenljive X, već samo od elemenata (svake) vrste kanalne matrice Π. Prema tome, sledi

$$\begin{aligned} C &= \max_{p(x)} I(X,Y) = \max_{p(x)} (H(Y) - H(Y|X)) \\ &= \max_{p(x)} H(Y) + \sum_{i=1}^{b} r_i \log_2 r_i. \end{aligned}$$

Pošto se za uniformu raspodelu $p(x) = 1/a$ postiže (Lema 6.4.2) da je raspodela slučajne promenljive Y takođe uniformna odnosno da $H(Y)$ dostiže svoj maksimum $\log_2 b$, sledi da je $\max_{p(x)} H(Y) = \log_2 b$. Ovim je dokaz završen. □

Za kanale nesimetrične po izlazu ne mora da važi Lema 6.4.2, pa ne mora da postoji raspodela na ulazu koja daje podjednako verovatne izlazne signale. Zato se za kanal simetričan po ulazu kapacitet može proceniti na sledeći način

$$C \leq \log_2 b + \sum_{i=1}^{b} r_i \log_2 r_i.$$

Simetrični kanal čija je matrica Π kvadratna i dimenzija $a \times a$ naziva se **a-arni simetrični kanal**. Najjednostavniji i ujedno i najvažniji slučaj je

takav kanal za $a = 2$, što je zapravo **binarni simetrični kanal**. Direktna generalizacija binarnog simetričnog kanala je kanal definisan matricom:

$$\Pi_a = \begin{bmatrix} 1-(a-1)\alpha & \alpha & \cdots & \alpha \\ \alpha & 1-(a-1)\alpha & & \alpha \\ \vdots & & \ddots & \vdots \\ \alpha & \alpha & & 1-(a-1)\alpha \end{bmatrix}_{a\times a}$$

za svako $0 < \alpha < 1/(a-1)$. Na osnovu Teoreme 6.4.3, kapacitet ovog kanala jednak je

$$C = \log_2 a + (1-(a-1)\alpha)\log_2(1-(a-1)\alpha) + (a-1)\alpha \log_2 \alpha.$$

Ovaj kanal ćemo kasnije koristiti za određivanje performansi nekih klasa zaštitnih kodova.

Glava 7

Zaštitno kodiranje - teorija

7.1 Osnovni pojmovi

Osnovni problem zaštitnog kodiranja je sledeći: potrebno je kodirati ukupno M simbola pomoću kodnih reči alfabeta $\mathcal{X}$ dužine n, tako da se verovatnoća greške pri prenosu kroz kanal $(\mathcal{X}, p(y|x), \mathcal{Y})$ minimizuje.

Definicija 7.1.1. *Kod sa oznakom (M, n) (ili (M, n)-kod) za kanal $(\mathcal{X}, p(y|x), \mathcal{Y})$ sadrži:*

1. *Indeksni skup $\{1, 2, \ldots, M\}$*
2. *Kodiranje $\mathbf{x} : \{1, 2, \ldots, M\} \to \mathcal{X}^n$, koje predstavlja funkciju (algoritam) za dodelu **kodne reči** $\mathbf{x}(k)$ indeksu k. Skup svih kodnih reči naziva se **kodna knjiga** ili samo **kod**.*
3. *Dekodiranje $g : \mathcal{Y}^n \to \{1, 2, \ldots, M\}$ koje predstavlja funkciju na osnovu koje se vrši procena indeksa poslate kodne reči, za svaku moguću pristiglu reč na izlazu iz kanala.*

Pretpostavimo sada da je na ulazu indeks i. Greška pri slanju nastaje, ukoliko je rezultat dekodiranja prispele poruke $\mathbf{y}$ (funkcija g) različit od i. Verovatnoća ovog događaja je

$$\lambda_i = P(g(\mathbf{Y}) \neq i | \mathbf{X} = \mathbf{x}(i)) = \sum_{\mathbf{y},\ g(\mathbf{y}) \neq i} p(\mathbf{y} \mid \mathbf{x}(i)).$$

Ova veličina se naziva i **uslovna verovatnoća greške**. **Maksimalna verovatnoća greške** se prirodno definiše kao

$$\lambda^{(n)} = \max_{i=1,2,\ldots,M} \lambda_i$$

a **srednja verovatnoća greške** kao

$$P_e^{(n)} = \frac{1}{M} \sum_{i=1}^{M} \lambda_i.$$

Primetimo da se kod srednje verovatnoće greške, implicitno pretpostavlja da su indeksi $i = 1, 2, \ldots, M$ jednako verovatni. Drugim rečima, da je slučajna promenljiva W, koja predstavlja indeks kodne reči koja se šalje, uniformno raspodeljena. Očigledno je $P_e^{(n)} \leq \lambda^{(n)}$.

Definicija 7.1.2. ***Kodni količnik*** (***code rate***) *(M, n)-koda definisan je sa*

$$R = \frac{\textit{min. br. bitova za indeks}}{\textit{stvarni. br. bitova za indeks}} = \frac{\log_2 M}{n}$$

Praktično, kodni količnik definiše redundansu koju (svesno) unosimo kodirajući poruke sa n, umesto sa minimalnim mogućim brojem bita $\log_2 M$. Ova redundansa se unosi upravo sa ciljem da bi se smanjila verovatnoća greške.

Primer 7.1.1. Posmatrajmo sada najprostiji primer zaštitnog kodiranja. Na ulazu imamo niz nula i jedinica koje šaljemo binarnim simetričnim kanalom. Da bi smanjili verovatnoću greške, svaki bit šaljemo 3 puta. Dakle, umesto 0, šaljemo 000, a umesto 1 šaljemo 111. Dakle, imamo da je $M = 2$, $n = 3$ kao i

$$\mathcal{X} = \{0, 1\}, \quad \mathcal{Y} = \{0, 1\}, \quad \Pi = \begin{bmatrix} p(0|0) & p(1|0) \\ p(0|1) & p(1|1) \end{bmatrix} = \begin{bmatrix} 1-\alpha & \alpha \\ \alpha & 1-\alpha \end{bmatrix}.$$

Usvajamo kodiranje $\mathbf{x}(0) = 000$ i $\mathbf{x}(1) = 111$. Na izlazu iz kanala, moguće je da stigne bilo koji niz od 3 bita.

Sada je potrebno definisati postupak dekodiranja g. Jasno je da je $g(000) = 0$ i $g(111) = 1$ (u tim slučajevima nije došlo do greške). Ukoliko se na izlazu pojavio neki drugi niz, znači da se pojavila greška bar na jednom bitu. Pritom, ako smo dobili npr. 010, verovatnije je da je došlo do greške samo na drugom bitu (poslato je 0), nego i na prvom i na trećem bitu (poslato je 1). Dakle stavićemo $g(010) = 0$. Na sličan način, za bilo koju drugu moguću reč y na izlazu, prebrojimo koliko ima nula i jedinica, pa ako ima više nula onda je $g(y) = 0$, u suprotnom je $g(y) = 1$. Dakle:

$$\begin{array}{ll} g(000) = 0 & g(111) = 1 \\ g(001) = 0 & g(110) = 1 \\ g(010) = 0 & g(101) = 1 \\ g(100) = 0 & g(011) = 1 \end{array}$$

Nije teško uvideti da je verovatnoća da dođe do greške, pod uslovom da je poslata 0, jednaka:

$$\begin{aligned}\lambda_0 &= P(g(\mathbf{Y}) = 1 \mid \mathbf{X} = x(0)) = P(\mathbf{Y} \in \{011, 101, 110, 111\}\} \mid \{\mathbf{X} = 000) \\ &= p(011|000) + p(101|000) + p(110|000) + p(111|000) \\ &= 3\alpha^2(1-\alpha) + \alpha^3 = 3\alpha^2 - 2\alpha^3.\end{aligned}$$

Isti izraz se dobija i za λ_1, pa je

$$\lambda^{(3)} = \max\{\lambda_0, \lambda_1\} = \alpha^2(3 - 2\alpha).$$

Dakle, dobili smo verovatnoću greške koja je za red veličine bolja od one koju bi dobili bez primene zaštitnog koda ($\lambda^{(1)} = p$). Međutim, cena koju plaćamo je 3 puta manja bitska brzina, zato što je

$$R^{(3)} = \frac{1}{3}.$$

Na sličan način, može da se pokaže da je $\lambda^{(2k+1)} \sim \alpha^k$, ali je pritom $R^{(2k+1)} = 1/(2k+1)$, za svako $k \in \mathbb{N}$.

Dakle, i pored činjenice da metodom opisanim u prethodnom primeru, možemo dobiti proizvoljno malu verovatnoću greške, on je neupotrebljiv zato što kodni količnik (a time i bitska brzina prenosa) teži nuli.

Pitanje koje se prirodno postavlja je da li je moguće (proizvoljno) smanjiti verovatnoću greške, a da pritom kodni količnik ostane isti? Odgovor na ovo pitanje daje druga Shannonova teorema.

Definicija 7.1.3. *Kodni količnik R je **dostižan**, ukoliko postoji niz $(\lceil 2^{nR} \rceil, n)$-kodova takvih da $\lambda^{(n)} \to 0$ kad $n \to +\infty$.*

U nastavku ćemo umesto oznake $(\lceil 2^{nR} \rceil, n)$-kod, često pisati $(2^{nR}, n)$ kod, da bi uprostili notaciju.

Druga Shannonova teorema upravo tvrdi da su sve vrednosti kodnog količnika $0 \leq R < C$ dostižne, gde je C kapacitet kanala.

7.2 Tipične n-torke

Pretpostavimo da su $X_1, X_2, \ldots$ vrednosti koje smo dobili iz nekog izvora informacija, koji opisujemo slučajnom promenljivom X. Formalnije, neka je

$(X_n)_{n\in\mathbb{N}}$ stacionarni izvor informacija bez memorije (sve X_n su međusobno nezavisne i imaju istu raspodelu kao X). Pitamo se koji su to nizovi $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ koji se "često javljaju" (tipični nizovi) kao realizacija $(X_1, X_2, \ldots, X_n)$ a koji su nešto "ređi".

Da bi formalno definisali "tipičnost" i odgovorili prethodno postavljeno pitanje, potrebno je najpre uvesti pojmove (tipove) konvergencije niza slučajnih promenljivih.

Definicija 7.2.1. ***(Konvergencija slučajnih promenljivih)*** *Neka je $(X_n)_{n\in\mathbb{N}}$ niz [1] slučajnih promenljivih. Kažemo da $(X_n)_{n\in\mathbb{N}}$ konvergira ka slučajnoj promenljivoj X:*

1. ***u verovatnoći**, ako za svako $\epsilon > 0$ važi $P(|X_n - X| \geq \epsilon) \to 0$ kad $n \to +\infty$, što pišemo $X_n \xrightarrow{v} X$.*
2. ***srednje-kvadratno**, ako je $\mathbb{E}|X_n - X|^2 \to 0$ kad $n \to +\infty$, što pišemo $X_n \xrightarrow{s.k.} X$.*
3. ***sa verovatnoćom 1** (**skoro izvesno**) ako $P(X_n \to X) = 1$, što pišemo $X_n \xrightarrow{s.i.} X$.*

Jednostavno se pokazuje da ako niz $(X_n)_{n\in\mathbb{N}}$ konvergira u verovatnoći, nejednakost $\geq$ u definicionom izrazu možemo zameniti strogom nejednakošću $>$, odnosno da važi sledeća lema:

Lema 7.2.1. *Niz $(X_n)_{n\in\mathbb{N}}$ slučajnih promenljvih konvergira u verovatnoći ka X ako i samo ako $P(|X_n - X| > \epsilon) \to 0$ kad $n \to +\infty$, za svako $\epsilon > 0$.*

Skoro izvesna i srednje-kvadratna konvergencija povlače konvergenciju u verovatnoći. Ove dve vrste konvergencije su međusobno nezavisne.

Teorema 7.2.2. ***(Drugi (strogi) zakon velikih brojeva Kolmogorova)*** *Neka su $X_1, X_2, \ldots$ nezavisne slučajne promenljive sa istom raspodelom. Tada*

$$\mathbb{E}X_1 = a \quad \Leftrightarrow \quad \frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow{s.i.} a.$$

Posledica 7.2.3. ***(Slabi zakon velikih brojeva)*** *Neka su $X_1, X_2, \ldots$ nezavisne slučajne promenljive sa istom raspodelom i konačnim očekivanjem a. Tada je*

$$\frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow{v} a.$$

[1] Ne nužno diskretnih.

Dokaz prethodne teoreme odnosno posledice može se naći na primer u [SJ, p. 129–130].

Sada možemo formulisati i dokazati teoremu koja predstavlja osnovu formalnog definisanja pojma "tipičnosti" n-torke $\mathbf{x} = (x_1, x_2, \ldots, x_n)$.

Teorema 7.2.4. ***(Asimptotsko ekviparticiono svojstvo)*** *Ako je $X = (X_n)_{n\in\mathbb{N}}$ stacionarni izvor bez memorije (drugim rečima, ako su $X_1, X_2, \ldots$ nezavisne slučajne promenljive sa istom raspodelom $p(x)$), tada je*

$$-\frac{1}{n}\log_2 p(X_1, X_2, \ldots, X_n) \xrightarrow{v} H(X).$$

Dokaz. Pošto je $p(x_1, x_2, \ldots, x_n) = p(x_1)p(x_2)\cdots p(x_n)$ (slučajne promenljive $X_1, X_2, \ldots, X_n$ su nezavisne i sa istom raspodelom), onda je

$$\begin{aligned} -\frac{1}{n}\log_2 p(X_1, X_2, \ldots, X_n) &= -\frac{1}{n}\sum_{i=1}^{n}\log_2 p(X_i) \\ &\to -\mathbb{E}\log_2 p(X) \\ &= -\sum_{x\in\mathcal{X}} p(x)\log_2 p(x) = H(X), \end{aligned}$$

gde konvergencija u verovatnoći sledi na osnovu slabog zakona velikih brojeva. □

Definicija 7.2.2. ***Tipični skup*** *n-torki $A_\epsilon^{(n)}$ u odnosu na slučajnu promenljivu X (ili stacionarni izvor informacija bez memorije $X = (X_n)_{n\in\mathbb{N}}$) je skup svih n-torki $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ takvih da važi*

$$2^{-n(H(X)+\epsilon)} \le p(x_1, x_2, \ldots, x_n) \le 2^{-n(H(X)-\epsilon)}.$$

*Sve n-torke iz skupa $A_\epsilon^{(n)}$ su **tipične** n-torke.*

Teorema 7.2.5. *Pod uslovima prethodne teoreme važi:*

1. *Ako je $(x_1, x_2, \ldots, x_n) \in A_\epsilon^{(n)}$, onda je*

$$\left| -\frac{1}{n}\log_2 p(x_1, x_2, \ldots, x_n) - H(X) \right| \le \epsilon.$$

2. *$P\left((X_1, X_2, \ldots, X_n) \in A_\epsilon^{(n)}\right) > 1 - \epsilon$ za svako n počev od nekog n_0.*

3. $|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$.

4. $|A_\epsilon^{(n)}| \geq (1-\epsilon)2^{n(H(X)-\epsilon)}$, *za svako* n *počev od nekog* n_0.

Dokaz. [Cover, p. 59–60] Deo 1 sledi direktno iz definicije tipične n-torke. Zaista, ako definicionu nejednakost logaritmujemo i podelimo sa n dobijamo

$$H(X) - \epsilon \leq -\frac{1}{n}\log_2 p(\mathbf{x}) \leq H(X) + \epsilon.$$

odakle direktno sledi

$$\left| -\frac{1}{n}\log_2 p(\mathbf{x}) - H(X) \right| \leq \epsilon.$$

Iz asimptotskog ekviparticionog svojstva (Teorema 7.2.4), kao i definicije konvergencije u verovatnoći sledi

$$\begin{aligned} P\left(\mathbf{X} \in A_\epsilon^{(n)}\right) &= P\left(\left| -\frac{1}{n}\log_2 p(\mathbf{X}) - H(X) \right| \leq \epsilon\right) \\ &= 1 - P\left(\left| -\frac{1}{n}\log_2 p(\mathbf{X}) - H(X) \right| > \epsilon\right) > 1 - \epsilon' \end{aligned}$$

za svako $\epsilon' > 0$ i $n \geq n_0$ gde je n_0 prirodni broj koji zavisi od ϵ'. Ako uzmemo upravo $\epsilon' = \epsilon$ (što možemo, zbog proizvoljnosti broja ϵ') dobijamo deo 2 leme. Primetimo da je

$$\begin{aligned} 1 \geq P\left(\left\{\mathbf{X} \in A_\epsilon^{(n)}\right\}\right) &= \sum_{\mathbf{x}\in A_\epsilon^{(n)}} p(\mathbf{x}) \\ \geq \sum_{\mathbf{x}\in A_\epsilon^{(n)}} 2^{-n(H(X)+\epsilon)} &= \left|A_\epsilon^{(n)}\right| 2^{-n(H(X)+\epsilon)}. \end{aligned}$$

Odavde direktno sledi deo 3 leme. Sa druge strane, koristeći prethodno dokazan deo 2 leme kao i suprotnu nejednakost za $p(\mathbf{x})$, dobijamo

$$\begin{aligned} 1 - \epsilon < P\left(\left\{\mathbf{X} \in A_\epsilon^{(n)}\right\}\right) &= \sum_{\mathbf{x}\in A_\epsilon^{(n)}} p(\mathbf{x}) \\ \leq \sum_{\mathbf{x}\in A_\epsilon^{(n)}} 2^{-n(H(X)-\epsilon)} &= \left|A_\epsilon^{(n)}\right| 2^{-n(H(X)-\epsilon)} \end{aligned}$$

za $n \geq n_0$. Ovim smo dokazali i poslednji 4. deo leme. □

Značenje prethodne teoreme je sledeće: Skup $\mathcal{X}^n$ možemo podeliti u dva dela: **tipični** i **netipični**. Neformalnim jezikom rečeno, ova dva skupa možemo opisati na sledeći način:

1. Tipičnih n-torki ima ne više od $2^{n(H(X)+\epsilon)}$ i raspodela $p(x_1, x_2, \ldots, x_n)$ na njima je približno uniformna.

2. Preostale n-torke (kojih ima nezanemarljivo mnogo) nisu mnogo verovatne. Verovatnoća pojavljivanja $\mathbf{x} \notin A_\epsilon^{(n)}$ je manja od ϵ.

7.3 Tipične n-torke i izvorno kodiranje

Na osnovu Teoreme 7.2.5 možemo konstruisati kod V_n za $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, takav da je srednja dužina kodne reči po simbolu ($\bar{n}_{V_n}/n$ proizvoljno bliska entropiji $H(X)$). Konstrukcija se sastoji u sledećem:

1. Razmatramo posebno tipične a posebno netipične n-torke. Prvi bit koda određuje da li je u pitanju tipična ili netipična n-torka.

2. Skup tipičnih n-torki $A_\epsilon^{(n)}$ ima maksimalno $2^{n(H(X)+\epsilon)}$ elemenata (Teorema 7.2.5). Prema tome, sve tipične n-torke možemo kodirati sa maksimalno (uz dodatak prvog bita koji određuje tip n-torke):

$$\lceil n(H(X)+\epsilon)\rceil + 1 \leq n(H(X)+\epsilon) + 2$$

   bita.

3. Broj ostalih (netipičnih) n-torki je očigledno manji od ukupnog broja (tipičnih i netipičnih) n-torki $|\mathcal{X}^n| = |\mathcal{X}|^n$, pa njih možemo kodirati sa maksimalno (ponovo, uz dodatak prvog bita):

$$\lceil \log_2 |\mathcal{X}|^n \rceil + 1 = \lceil n \log_2 |\mathcal{X}| \rceil + 1 \leq n \log_2 |\mathcal{X}| + 2$$

   bita.

Označimo sa $l(\mathbf{x})$ dužinu kodne reči $f(\mathbf{x})$ koja odgovara n-torki $\mathbf{x} \in \mathcal{X}^n$. Srednja dužina kodne reči koda V_n je

$$\begin{aligned}\bar{n}_{V_n} &= \sum_{\mathbf{x}\in\mathcal{X}^n} p(\mathbf{x})l(\mathbf{x}) \\ &= \sum_{\mathbf{x}\in A_\epsilon^{(n)}} p(\mathbf{x})l(\mathbf{x}) + \sum_{\mathbf{x}\in\left[A_\epsilon^{(n)}\right]^C} p(\mathbf{x})l(\mathbf{x})\end{aligned}$$

Ako sada primenimo prethodno izvedene granice za dužinu $l(\mathbf{x})$ u slučaju da je $\mathbf{x}$ tipična odnosno netipična n-torka, dobijamo:

$$\begin{aligned}\bar{n}_{V_n} &\leq \sum_{\mathbf{x}\in A_\epsilon^{(n)}} p(\mathbf{x})(n(H(X)+\epsilon)+2) + \sum_{\mathbf{x}\in\left[A_\epsilon^{(n)}\right]^C} p(\mathbf{x})(n\log_2|\mathcal{X}|+2)\\ &= P\left(\mathbf{X}\in A_\epsilon^{(n)}\right)(n(H(X)+\epsilon)+2) + P\left(\mathbf{X}\notin A_\epsilon^{(n)}\right)(n\log_2|\mathcal{X}|+2)\\ &\leq n(H(X)+\epsilon)+n\epsilon\log_2|\mathcal{X}|+2 = n(H(X)+\epsilon')\end{aligned}$$

gde je $\epsilon' = \epsilon(1+\log_2|\mathcal{X}|)+2/n$. S obzirom da se ϵ' može učiniti proizvoljno malim (za dovoljno malo ϵ i dovoljno veliko n), važi sledeća teorema.

Teorema 7.3.1. *Neka je $\epsilon > 0$ proizvoljan broj i $(X_n)_{n\in\mathbb{N}}$ stacionarni izvor informacija bez memorije. Tada za dovoljno veliko n [2], postoji binarni (prefiksni) kod $V_n = f(\mathcal{X}^n)$ čija je srednja dužina kodne reči*

$$\bar{n}_{V_n} \leq n(H(X)+\epsilon).$$

Primetimo da je prethodna teorema reformulacija Shannon-Fano stava (Teorema 5.5.4).

7.4 Združeni tipični parovi

Definicija 7.4.1. *Združeni tipični skup parova $(\mathbf{x},\mathbf{y}) \in \mathcal{X}^n \times \mathcal{Y}^n$ u odnosu na slučajne promenljive X i Y (ili stacionarni izvor informacija bez memorije $(X,Y) = ((X_n,Y_n))_{n\in\mathbb{N}}$) je*

$$\begin{aligned}\mathbf{A}_\epsilon^{(n)} = \Big\{&(\mathbf{x},\mathbf{y}) \in \mathcal{X}^n \times \mathcal{Y}^n \mid\\ &\left|-\frac{1}{n}\log_2 p(\mathbf{x}) - H(X)\right| \leq \epsilon,\\ &\left|-\frac{1}{n}\log_2 p(\mathbf{y}) - H(Y)\right| \leq \epsilon,\\ &\left|-\frac{1}{n}\log_2 p(\mathbf{x},\mathbf{y}) - H(X,Y)\right| \leq \epsilon\Big\},\end{aligned}$$

gde je

$$p(\mathbf{x},\mathbf{y}) = \prod_{i=1}^{n} p(x_i,y_i).$$

[2] Važi i jači zaključak, da postoji $n_0 \in \mathbb{N}$ takav da tvrđenje važi za svako $n > n_0$.

Teorema 7.4.1. ***(Združeno asimptotsko ekviparticiono svojstvo (AEP))*** *Neka je $((X_n, Y_n))_{n\in\mathbb{N}}$ diskretan izvor informacija bez memorije. Drugim rečima, neka su parovi (X_i, Y_i) $(i = 1, 2, \ldots, n)$ međusobno nezavisni i imaju istu raspodelu $p(x, y)$. Označimo sa $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ i $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$. Neka je (X, Y) proizvoljna dvodimenzionalna slučajna promenljiva koja ima raspodelu $p(x, y)$. Tada je*

1. $P\left((\mathbf{X}, \mathbf{Y}) \in \mathbf{A}_\epsilon^{(n)}\right) \to 1$ *kad* $n \to +\infty$.

2. $|\mathbf{A}_\epsilon^{(n)}| \leq 2^{n(H(X,Y)+\epsilon)}$.

3. *Ako su slučajne promenljive $\bar{\mathbf{X}}$ i $\bar{\mathbf{Y}}$ takve da su nezavisne i da imaju istu raspodelu kao $\mathbf{X}$ i $\mathbf{Y}$ respektivno (drugim rečima da $(\bar{X}_n, \bar{Y}_n)$ ima funkciju raspodele $\bar{p}(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$), onda je*

$$P\left(\left\{(\bar{\mathbf{X}}, \bar{\mathbf{Y}}) \in \mathbf{A}_\epsilon^{(n)}\right\}\right) \leq 2^{-n(I(X,Y)-3\epsilon)}.$$

4. *Za dovoljno veliko n je $|\mathbf{A}_\epsilon^{(n)}| \geq (1-\epsilon)2^{n(H(X,Y)-\epsilon)}$ kao i*

$$P\left(\left\{(\bar{\mathbf{X}}, \bar{\mathbf{Y}}) \in \mathbf{A}_\epsilon^{(n)}\right\}\right) \geq (1-\epsilon)2^{-n(I(X,Y)+3\epsilon)}.$$

Dokaz. [Cover, p. 196–198] Primetimo najpre da je

$$\begin{aligned} P\left((\mathbf{X}, \mathbf{Y}) \in \mathbf{A}_\epsilon^{(n)}\right) &= 1 - P\left((\mathbf{X}, \mathbf{Y}) \notin \mathbf{A}_\epsilon^{(n)}\right) \\ &\geq 1 - P\left(\left|-\frac{1}{n}\log_2 p(\mathbf{x}) - H(X)\right| > \epsilon\right) \\ &\quad - P\left(\left|-\frac{1}{n}\log_2 p(\mathbf{y}) - H(Y)\right| > \epsilon\right) \\ &\quad - P\left(\left|-\frac{1}{n}\log_2 p(\mathbf{x}, \mathbf{y}) - H(X, Y)\right| > \epsilon\right) \end{aligned}$$

Pošto su $X_1, X_2, \ldots, X_n$ nezavisne slučajne promenljive, a isto važi i za $Y_1, Y_2, \ldots, Y_n$ kao i za parove $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, na osnovu asimptotskog ekviparticionog svojstva (Teorema 7.2.4) sledi da su sve tri verovatnoće u prethodnom izrazu manje od $\epsilon'/3$ za svako n počev od nekog n_0 koje zavisi od ϵ'. Tada je

$$P\left((\mathbf{X}, \mathbf{Y}) \in \mathbf{A}_\epsilon^{(n)}\right) > 1 - \epsilon', \qquad n \geq n_0$$

za proizvoljno ϵ', odakle sledi deo 1 tvrđenja. Nejednakosti za broj elemenata skupa $\mathbf{A}_\epsilon^{(n)}$ u delovima 2 i 4 dokazujemo potpuno isto kao i za tipične n-torke:

$$\begin{aligned} 1 \geq P\left((\mathbf{X},\mathbf{Y}) \in \mathbf{A}_\epsilon^{(n)}\right) &= \sum_{(\mathbf{x},\mathbf{y})\in\mathbf{A}_\epsilon^{(n)}} p(\mathbf{x},\mathbf{y}) \\ &\geq \sum_{(\mathbf{x},\mathbf{y})\in\mathbf{A}_\epsilon^{(n)}} 2^{-n(H(X,Y)+\epsilon)} = \left|\mathbf{A}_\epsilon^{(n)}\right| 2^{-n(H(X,Y)+\epsilon)}. \end{aligned}$$

odnosno

$$\begin{aligned} 1-\epsilon < P\left((\mathbf{X},\mathbf{Y}) \in \mathbf{A}_\epsilon^{(n)}\right) &= \sum_{(\mathbf{x},\mathbf{y})\in\mathbf{A}_\epsilon^{(n)}} p(\mathbf{x},\mathbf{y}) \\ &\leq \sum_{(\mathbf{x},\mathbf{y})\in\mathbf{A}_\epsilon^{(n)}} 2^{-n(H(X,Y)-\epsilon)} = \left|\mathbf{A}_\epsilon^{(n)}\right| 2^{-n(H(X,Y)-\epsilon)}. \end{aligned}$$

Dokažimo sada delove 3 i 4. Verovatnoća da vrednost slučajne promenljive $(\bar{\mathbf{X}},\bar{\mathbf{Y}})$ pripada $\mathbf{A}_\epsilon^{(n)}$ jednaka je

$$\begin{aligned} P\left((\bar{\mathbf{X}},\bar{\mathbf{Y}}) \in \mathbf{A}_\epsilon^{(n)}\right) &= \sum_{(\mathbf{x},\mathbf{y})\in\mathbf{A}_\epsilon^{(n)}} \bar{p}(\mathbf{x},\mathbf{y}) = \sum_{(\mathbf{x},\mathbf{y})\in\mathbf{A}_\epsilon^{(n)}} p(\mathbf{x})p(\mathbf{y}) \\ &\leq \sum_{(\mathbf{x},\mathbf{y})\in\mathbf{A}_\epsilon^{(n)}} 2^{-n(H(X)-\epsilon)} \cdot 2^{-n(H(Y)-\epsilon)} \\ &= \left|\mathbf{A}_\epsilon^{(n)}\right| 2^{-n(H(X)+H(Y)-2\epsilon)} \\ &\leq 2^{n(H(X,Y)+\epsilon)} \cdot 2^{-n(H(X)+H(Y)-2\epsilon)} \\ &= 2^{-n(H(X)+H(Y)-H(X,Y)-3\epsilon)} = 2^{-n(I(X,Y)-3\epsilon)}. \end{aligned}$$

Sa druge strane je

$$\begin{aligned} P\left((\bar{\mathbf{X}},\bar{\mathbf{Y}}) \in \mathbf{A}_\epsilon^{(n)}\right) &= \sum_{(\mathbf{x},\mathbf{y})\in\mathbf{A}_\epsilon^{(n)}} p(\mathbf{x})p(\mathbf{y}) \\ &\geq \sum_{(\mathbf{x},\mathbf{y})\in\mathbf{A}_\epsilon^{(n)}} 2^{-n(H(X)+\epsilon)} \cdot 2^{-n(H(Y)+\epsilon)} \\ &= \left|\mathbf{A}_\epsilon^{(n)}\right| 2^{-n(H(X)+H(Y)+2\epsilon)} \\ &\geq (1-\epsilon)2^{n(H(X,Y)-\epsilon)} \cdot 2^{-n(H(X)+H(Y)+2\epsilon)} \\ &= (1-\epsilon)2^{-n(H(X)+H(Y)-H(X,Y)+3\epsilon)} = (1-\epsilon)2^{-n(I(X,Y)+3\epsilon)}. \end{aligned}$$

Ovim je dokaz teoreme završen. □

Iako ima približno $2^{nH(X)}$ i $2^{nH(Y)}$ tipičnih n-torki za izvore X i Y, na osnovu prethodne teoreme zaključujemo da je broj združenih tipičnih parova samo $2^{nH(X,Y)}$. Dakle, nisu svi parovi $(\mathbf{x}, \mathbf{y}) \in \mathcal{X}^n \times \mathcal{Y}^n$ tipičnih n-torki za X i Y ujedno i združeno tipični. Štaviše, verovatnoća da slučajno (uniformno) uzet par $(\mathbf{x}, \mathbf{y})$, gde su $\mathbf{x}$ i $\mathbf{y}$ tipične n-torke, predstavlja združeni tipični par, jednaka je

$$p_z \approx \frac{2^{nH(X,Y)}}{2^{nH(X)}2^{nH(Y)}} = 2^{-nI(X,Y)}.$$

Primer 7.4.1. Neka su X i Y binarne slučajne promenljive ($\mathcal{X} = \mathcal{Y} = \{0, 1\}$) čija je združena funkcija raspodele definisana sa

$$p(0,0) = 0.32, \quad p(0,1) = 0.08, \quad p(1,0) = 0.12, \quad p(1,1) = 0.48.$$

Tada očigledno X i Y imaju raspodele

$$p_X : \begin{pmatrix} 0 & 1 \\ 0.4 & 0.6 \end{pmatrix}, \qquad p_Y : \begin{pmatrix} 0 & 1 \\ 0.44 & 0.56 \end{pmatrix}.$$

pa je $H(X) = 0.971$ i $H(Y) = 0.99$. Sa druge strane, imamo da je združena entropija jednaka

$$\begin{aligned} H(X,Y) &= -p(0,0)\log_2 p(0,0) - p(0,1)\log_2 p(0,1) - p(1,0)\log_2 p(1,0) - p(1,1)\log_2 p(1,1) \\ &= -0.32\log_2 0.32 - 0.08\log_2 0.08 - 0.12\log_2 0.12 - 0.48\log_2 0.48 \\ &= 1.693. \end{aligned}$$

Neka su $\mathbf{x}^{30}$ i $\mathbf{y}^{30}$ (gornji indeksi označavaju dužinu niza) sledeći nizovi:

$$\mathbf{x}^{30} = (0,0,1,1,1,1,1,1,0,1,1,0,0,0,1,0,1,0,1,1,0,1,1,1,1,1,1,1,0,1)$$
$$\mathbf{y}^{30} = (1,0,1,1,1,1,1,0,0,1,0,0,0,0,0,0,1,0,1,1,0,0,1,1,0,1,1,1,1,1)$$

Očigledno je

$$p(\mathbf{x}^{30}) = p(0)p(0)p(1)\cdots p(1)p(0)p(1) = p(0)^{10}p(1)^{20} = 3.834 \cdot 10^{-9}$$
$$p(\mathbf{y}^{30}) = p(1)p(0)p(1)\cdots p(1)p(1)p(1) = p(0)^{13}p(1)^{17} = 1.214 \cdot 10^{-9}$$

pa je

$$\left| -\frac{1}{30}\log_2 p(\mathbf{x}^{30}) - H(X) \right| = 0.039, \quad \left| -\frac{1}{30}\log_2 p(\mathbf{y}^{30}) - H(Y) \right| = 0.002.$$

Ako je $\epsilon = 0.05$ onda sledi da su $\mathbf{x}^{30}$ i $\mathbf{y}^{30}$ tipične 30-torke. Sa druge strane je i

$$\begin{aligned} p(\mathbf{x}^{30}, \mathbf{y}^{30}) &= p(0,1)p(0,0)p(1,1)\cdots p(1,1)p(0,1)p(1,1) \\ &= p(0,0)^8 p(0,1)^2 p(1,0)^5 p(1,1)^{15} = 2.897 \cdot 10^{-16} \end{aligned}$$

pa je

$$\left| -\frac{1}{30}\log_2 p(\mathbf{x}^{30}, \mathbf{y}^{30}) - H(X,Y) \right| = 0.028 < \epsilon$$

odakle zaključujemo da je $(\mathbf{x}^{30}, \mathbf{y}^{30})$ združeni tipični par. Međutim, ukoliko posmatramo 30-torke $\bar{\mathbf{x}}^{30}$ i $\bar{\mathbf{y}}^{30}$ gde je

$$\begin{aligned} \mathbf{x}^{30} = \bar{\mathbf{x}}^{30} &= (0,0,1,1,1,1,1,1,0,1,1,0,0,0,1,0,1,0,1,1,0,1,1,1,1,1,1,1,0,1) \\ \bar{\mathbf{y}}^{30} &= (1,1,1,1,1,0,0,0,1,1,1,1,1,1,0,1,1,0,1,1,0,1,1,1,1,0,1,0,1,0,0) \end{aligned}$$

dobićemo da je

$$\left| -\frac{1}{30}\log_2 p(\bar{\mathbf{y}}^{30}) - H(Y) \right| = 0.037 < \epsilon$$

dok je

$$\left| -\frac{1}{30}\log_2 p(\mathbf{x}^{30}, \mathbf{y}^{30}) - H(X,Y) \right| = 0.628 > \epsilon.$$

Odavde zaključujemo da iako su $\bar{\mathbf{x}}^{30}$ i $\bar{\mathbf{y}}^{30}$ tipični za X i Y, par $(\bar{\mathbf{x}}^{30}, \bar{\mathbf{y}}^{30})$ nije združeno tipičan za X i Y.

Primetimo da se parovi $(1,0)$ i $(0,1)$ pojavljuju čitavih 8 puta u nizu $(\bar{\mathbf{x}}^{30}, \bar{\mathbf{y}}^{30})$, iako su verovatnoće pojavljivanja ovih parova samo $p(0,1) = 0.08$ i $p(1,0) = 0.12$. Sa druge strane, par $(0,0)$ pojavljuje se samo 2 puta, iako je $p(0,0) = 0.32$. Ovo su razlozi zašto ovaj par nije združeno tipičan.

7.5 Druga Shannonova teorema

Sada ćemo dokazati najvažniju teoremu iz teorije kodiranja. Ona tvrdi da je moguće prenositi informaciju sa proizvoljno malom verovatnoćom greške, sve dok unosimo dovoljnu redundansu u procesu slanja, odnosno dok je kodni količnik R manji od kapaciteta kanala C. Ovaj zaključak je na prvi pogled suprotan intuiciji. Ako kanal unosi greške, kako je moguće (skoro) sve ih ispraviti? Svaki pokušaj ispravke je takođe izvor novih grešaka.

Zapravo, teorema pokazuje da postoji kod koji na pogodan način unosi redundansu u podatke, i za dovoljno velike dužine kodne reči čini da oni budu u proizvoljnoj meri otporni na greške. Štaviše, teorema tvrdi da se to (u

srednjem) postiže slučajnim izborom koda. Pošto svi zaključci važe uz pretpostavku da je dužina kodnih reči n dovoljno velika, ova teorema nam nije od pomoći kad je u pitanju konstrukcija konkretnog koda.

Teorema 7.5.1. ***(Druga Shannonova teorema)*** *Za diskretni komunikacijski kanal $(\mathcal{X}, p(\cdot|\cdot), \mathcal{Y})$* ***bez memorije****, dostižni su svi kodni količnici R koji su manji od kapaciteta kanala C. Drugim rečima, ako je $R < C$ tada postoji niz $(2^{nR}, n)$ kodova takav da važi $\lambda^{(n)} \to 0$ kad $n \to +\infty$.*

Pre nego što pređemo na konkretan dokaz Druge Shannonove teoreme, opisaćemo postupak **slučajnog kodiranja (random coding)**. Neka je $p(x)$ data raspodela na skupu $\mathcal{X}$ (kasnije ćemo ovu raspodelu konkretno odrediti i fiksirati). Posmatrajmo slučajno izabrani kod $\mathcal{C}$ generisan raspodelom $p(x)$. Preciznije, neka su kodne reči $\mathbf{X}(1), \mathbf{X}(2), \ldots, \mathbf{X}(2^{nR})$ [3] slučajno izabrane raspodelom

$$p(\mathbf{x}) = \prod_{i=1}^{n} p(\mathbf{x}_i), \qquad \mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$$

Ove kodne reči su redovi $2^{nR} \times n$ matrice:

$$\mathcal{C} = \begin{bmatrix} \mathbf{x}_1(1) & \mathbf{x}_2(1) & \cdots & \mathbf{x}_n(1) \\ \mathbf{x}_1(2) & \mathbf{x}_2(2) & \cdots & \mathbf{x}_n(2) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_1(2^{nR}) & \mathbf{x}_2(2^{nR}) & \cdots & \mathbf{x}_n(2^{nR}) \end{bmatrix}.$$

Svaki element prethodne matrice predstavlja slučajnu promenljivu sa raspodelom $p(x)$, pri čemu su sve te slučajne promenljive međusobno nezavisne. Prema tome, verovatnoća generisanja određenog koda jednaka je

$$P(\mathcal{C}) = \prod_{w=1}^{2^{nR}} \prod_{i=1}^{n} p(\mathbf{x}_i(w)).$$

Postupak kodiranja/dekodiranja je sledeći:

1. Generiše se slučajni kod $\mathcal{C}$ na osnovu raspodele $p(x)$.
2. Kod $\mathcal{C}$ se saopšti i predajnoj i prijemnoj strani.
3. Slučajna promenljiva W koja predstavlja poruku (indeks) koju želimo preneti ima uniformnu raspodelu na skupu $\{1, 2, \ldots, 2^{nR}\}$. Drugim rečima, važi
$$P(W = w) = 1/2^{nR}, \qquad w = 1, 2, \ldots, 2^{nR}.$$

[3] Koristimo velika slova pošto su ovde i kodne reči slučajne promenljive.

4. Ukoliko je generisana poruka w, kodna reč $\mathbf{X}(w)$ se šalje kroz kanal.
5. Neka je $\mathbf{Y}$ slučajna promenljiva koja predstavlja reč koju je prijemnik tom prilikom primio. Uslovna raspodela promenljive $\mathbf{Y}$, pod uslovom $\mathbf{X}(w) = \mathbf{x}(w)$, je

$$P(\mathbf{Y} = \mathbf{y} \mid \mathbf{X}(w) = \mathbf{x}(w)) = \prod_{i=1}^{n} p(\mathbf{y}_i|\mathbf{x}_i(w)).$$

6. Zadatak prijemnika je da na osnovu primljene vrednosti $\mathbf{y}$ slučajne promenljive $\mathbf{Y}$ pretpostavi (zaključi) koja poruka je poslata. Najbolji način za to je **postupak maksimalne verodostojnosti** (*maximum likelihood decoding*). On se sastoji u tome da se, ukoliko je primljena poruka $\mathbf{y}$, nađe maksimum verovatnoće $p(\mathbf{y}|\mathbf{x}(w))$ (po w) i da je onda rezultat dekodera upravo vrednost indeksa $\hat{w}$ koji maksimizuje $p(\mathbf{y}|\mathbf{x}(w))$.

   Iako ovaj dekoder minimizuje verovatnoću greške, radi jednostavnije analize, koristićemo drugačiju konstrukciju. Dekoder (funkcija g) će vratiti indeks $\hat{w}$ takav da je:

   – $(\mathbf{x}(\hat{w}), \mathbf{y})$ združeno tipični par,
   – Ne postoji nijedan drugi indeks $w' \neq \hat{w}$ takav da je $(\mathbf{x}(w'), \mathbf{y})$ združeno tipični par.

   Ukoliko ne postoji takvo $\hat{w}$, smatraćemo da je došlo do greške, tj. uzećemo da je $\hat{w} = 0$ (poruka 0 nikako ne može biti poslata).

Označimo sada sa $\hat{W}$ slučajnu promenljivu koja predstavlja izlaz dekodera. Događaj koji predstavlja grešku (i čiju verovatnoću želimo da odredimo) je očigledno $\mathcal{E} = \{\hat{W} \neq W\}$.

Lema 7.5.2. *U postupku slučajnog kodiranja, za $R < I(X,Y) - 3\epsilon$ je $P(\mathcal{E}) \leq 2\epsilon$ za svako n počev od nekog n_0.*

Dokaz. Da bi pojednostavili pisanje, u dokazu ćemo podrazumevati da "za dovoljno veliko n" znači "za svako n počev od nekog n_0". Na kraju, dokazano tvrđenje važiće za svako n veće od maksimalnog do tada uočenog n_0.

Verovatnoću događaja $\mathcal{E}$ možemo da izračunamo na sledeći način:

$$P(\mathcal{E}) = \sum_{w=1}^{2^{nR}} P(\mathcal{E} \mid W = w) \cdot P(W = w).$$

Pošto je raspodela slučajne promenljive W uniformna, i pošto se sve kodne reči konstruišu na potpuno isti način (slučajnim izborom), zaključujemo da $P(\mathcal{E} \mid W = w)$ ne zavisi od w, tj da je

$$P(\mathcal{E}) = P(\mathcal{E} \mid W = 1).$$

Označimo sa E_i $(i = 1, 2, \ldots, 2^{nR})$ događaj

$$\{(\mathbf{X}(i), \mathbf{Y}) \in \mathbf{A}_\epsilon^{(n)}\}.$$

Drugim rečima, E_i je događaj da su primljena reč $\mathbf{Y}$ i i-ta kodna reč $\mathbf{X}(i)$ združeno tipični par. Na osnovu definicije postupka dekodiranja zaključujemo da do greške dolazi u slučaju da $\mathbf{X}(1)$ i $\mathbf{Y}$ nisu združeno tipične, ili da su $\mathbf{X}(i)$ i $\mathbf{Y}$ združeno tipični za neko $i = 2, 3, \ldots, 2^{nR}$ (bez obzira na to da li su $\mathbf{X}(1)$ i $\mathbf{Y}$ združeno tipični). Dakle, verovatnoća greške je

$$\begin{aligned} P(\mathcal{E} \mid W = 1) &= P(E_1^C \cup E_2 \cup \ldots \cup E_{2^{nR}} \mid W = 1) \\ &\leq P(E_1^c \mid W = 1) + \sum_{i=2}^{2^{nR}} P(E_i \mid W = 1). \end{aligned}$$

Na osnovu združenog AEP-a (Teorema 7.4.1) sledi

$$P(E_1^c \mid W = 1) = 1 - P(E_1 \mid W = 1) = 1 - P\left((\mathbf{X}(1), \mathbf{Y}) \in \mathbf{A}_\epsilon^{(n)}\right) \leq \epsilon$$

za dovoljno veliko n. Na osnovu konstrukcije koda $\mathcal{C}$ imamo da su $\mathbf{X}(1)$ i $\mathbf{X}(i)$ nezavisne slučajne promenljive za $i \neq 1$. Pošto znamo da je raspodela slučajne promenljive $\mathbf{Y}$ pod uslovom $W = 1$ određena raspodelom za $\mathbf{X}(1)$ i uslovnom raspodelom kanala, iz nezavisnosti $\mathbf{X}(1)$ i $\mathbf{X}(i)$ sledi i nezavisnost za $\mathbf{Y}$ i $\mathbf{X}(i)$. Na osnovu Teoreme 7.4.1 je

$$P(E_i \mid W = 1) \leq 2^{-n(I(X,Y)-3\epsilon)}.$$

Ovde smo, kao i u formulaciji Teoreme 7.4.1, sa (X, Y) označili proizvoljnu slučajnu promenljivu koja ima raspodelu $p(x, y) = p(x)p(y|x)$, gde je $p(\cdot|\cdot)$ definisana matricom kanala. Odatle je

$$\begin{aligned} P(\mathcal{E}) = P(\mathcal{E} \mid W = 1) &\leq P(E_1^c \mid W = 1) + \sum_{i=2}^{2^{nR}} P(E_i \mid W = 1) \\ &\leq \epsilon + \sum_{i=2}^{2^{nR}} 2^{-n(I(X,Y)-3\epsilon)} \\ &= \epsilon + (2^{nR} - 1)2^{-n(I(X,Y)-3\epsilon)} \\ &\leq \epsilon + 2^{-n(I(X,Y)-3\epsilon-R)} \\ &\leq 2\epsilon \end{aligned}$$

Poslednja nejednakost važi za dovoljno veliko n uz uslov $R < I(X,Y) - 3\epsilon$. □

Dakle, ukoliko je $R < I(X,Y)$, onda za proizvoljno $0 < \epsilon < (I(X,Y)-R)/3$ i $n \geq n_0$ je verovatnoća greške usrednjena po svim kodovima i svim kodnim rečima $P(\mathcal{E})$ [4] manja od 2ϵ.

Dokaz Druge Shannonove teoreme. [5] Neka je dat kodni količnik $R < C$ i neka je n_0' dovoljno veliko da važi $R' = R + 1/n < C$ za $n \geq n_0'$. Tada postoji raspodela $p(x)$ takva da je $R' < I(X,Y) \leq C$.

Primenom prethodne leme na kodni količnik R' dobijamo

$$P(\mathcal{E}) = \sum_{\mathcal{C}} P(\mathcal{E} \mid \mathcal{C}) P(\mathcal{C}) \leq 2\epsilon$$

za svako $n \geq n_0''$, pa sledi da postoji kod $\mathcal{C}'$ takav da je $P(\mathcal{E} \mid \mathcal{C}') \leq 2\epsilon$. Dalje je

$$P(\mathcal{E} \mid \mathcal{C}') = P_e^{(n)}(\mathcal{C}') = \frac{1}{M'} \sum_{i=1}^{M'} \lambda_i(\mathcal{C}') \leq 2\epsilon$$

za $M' = 2^{nR'}$. Odavde sledi da skup $\mathcal{I}$ svih indeksa i za koje je $\lambda_i(\mathcal{C}') \leq 4\epsilon$, tj.

$$\mathcal{I} = \{i = 1, 2, 3, \ldots, M' \mid \lambda_i(\mathcal{C}') \leq 4\epsilon\},$$

ima bar $M = 2^{nR'-1} = 2^{nR}$ elemenata [6]. Ukoliko konstruišemo novi kod $\mathcal{C}$ koji se sastoji samo iz 2^{nR} kodnih reči $\mathbf{x}(i)$, $i \in \mathcal{I}$, dobićemo da on ima kodni količnik $R = R' - 1/n$ i

$$\lambda_k(\mathcal{C}) \leq \lambda_k(\mathcal{C}') \leq 4\epsilon, \quad (k = 1, 2, \ldots, 2^{nR}) \quad \Rightarrow \quad \lambda^{(n)}(\mathcal{C}) \leq 4\epsilon.$$

Poslednja nejednakost važi jer $g'(\mathbf{y}) = k$ povlači $g(\mathbf{y}) = k$ (drugim rečima, nema greške za $\mathcal{C}'$ sledi nema greške za $\mathcal{C}$). Zaista, $g'(\mathbf{y}) = k$ znači da je $(\mathbf{x}(k), \mathbf{y})$ združeno tipični par, a nijedan od parova $(\mathbf{x}(i), \mathbf{y})$ za $i \neq k$ i $i = 1, 2, \ldots, 2^{nR}, 2^{nR} + 1, \ldots, 2^{nR'}$ nije. Samim tim, tvrđenje ostaje u važnosti kada se umesto šireg skupa indeksa $1, 2, \ldots, 2^{nR'}$ uzme uži $1, 2, \ldots, 2^{nR}$, što povlači $g(\mathbf{y}) = k$.

Dakle, za $R < C$ i dovoljno veliko n, postoji kod $\mathcal{C}$ čiji je kodni količnik jednak R a za koji je verovatnoća greške $\lambda^{(n)}$ proizvoljno mala (tj. manja od 4ϵ). Ovim je dokaz druge Shannonove teoreme završen. □

[4] Dekoder zavisi od ϵ, pa samim tim i $P(\mathcal{E})$.

[5] Blago formalizovan dokaz iz [Cover, p. 200–205]. Zamenjene su uloge R i R'.

[6] Formalno, pokazujemo da važi $|\mathcal{I}| > M'/2$. Ukoliko bi bilo $|\mathcal{I}| \leq M'/2$, tada bi važilo i $|\mathcal{I}^C| \geq M'/2$, odnosno $P_e^{(n)}(\mathcal{C}') > 1/M' \sum_{i \in \mathcal{I}^C} \lambda_i(\mathcal{C}') > (1/M')|\mathcal{I}^C| 4\epsilon \geq 2\epsilon$. Kontradikcija. Dakle, $|\mathcal{I}| > M'/2$, a pošto je $M \geq (M'+1)/2$ to je $|\mathcal{I}| \geq M$.

Iz dokaza Leme 7.5.2 vidimo da je za slučajno generisan kod $\mathcal{C}$ mnogo lakše odrediti verovatnoću greške nego za konkretni kod $\mathcal{C}$. Na kraju dajemo preciznu formulaciju najboljeg mogućeg dekodera g za neki kod $\mathcal{C} = \{\mathbf{x}(1), \ldots, \mathbf{x}(M)\}$.

Teorema 7.5.3. *Za diskretni kanal bez memorije i dati kod $\mathcal{C}$, pod uslovom da je W uniformno raspodeljena, funkcija g koja minimizuje verovatnoću greške data je sa*

$$g(\mathbf{y}) = \hat{w}, \qquad p(\mathbf{y}|\mathbf{x}(\hat{w})) = \max_{w=1,2,\ldots,M} p(\mathbf{y}|\mathbf{x}(w)).$$

Dokaz. Neka je $\mathcal{E}$ događaj da je došlo do greške, odnosno $\mathcal{E} = \{g(\mathbf{Y}) \neq W\}$. Tada je

$$\begin{aligned} P(\mathcal{E}) &= \sum_{g(\mathbf{y})\neq w} P(\mathbf{Y}=\mathbf{y}, W=w) = \sum_{g(\mathbf{y})\neq w} P(\mathbf{Y}=\mathbf{y}|W=w)P(W=w) \\ &= \frac{1}{M}\sum_{g(\mathbf{y})\neq w} p(\mathbf{y}|\mathbf{x}(w)) = \frac{1}{M}\sum_{\mathbf{y}\in\mathcal{Y}}\left[\sum_{w=1}^{M} p(\mathbf{y}|\mathbf{x}(w)) - p(\mathbf{y}|\mathbf{x}(g(\mathbf{y})))\right] \end{aligned}$$

Svaki sabirak po $\mathbf{y}$ u prethodnoj sumi možemo nezavisno da minimizujemo. Vidimo da je njegova vrednost minimalna, ukoliko je $p(\mathbf{y}|\mathbf{x}(g(\mathbf{y})))$ maksimalno, tj ukoliko je

$$p(\mathbf{y}|\mathbf{x}(g(\mathbf{y}))) = \max_{w=1,2,\ldots,M} p(\mathbf{y}|\mathbf{x}(w)).$$

Time je dokaz teoreme završen. □

Dekoder opisan u prethodnoj teoremi naziva se **ML (maximum likelihood) dekoder**. Ukoliko nemamo pretpostavku o uniformnoj raspodeli poruke W, najbolji dekoder se realizuje drugačije.

Teorema 7.5.4. *Za diskretni kanal bez memorije, dati kod $\mathcal{C}$ i poznatu raspodelu p_W poruka W, funkcija g koja minimizuje verovatnoću greške data je sa*

$$g(\mathbf{y}) = \hat{w}, \qquad p(\hat{w}|\mathbf{y}) = \max_{w=1,2,\ldots,M} p(w|\mathbf{y}).$$

Dokaz. Neka je, kao i do sada $\mathcal{E}$ događaj da je došlo do greške. Ovog puta,

razlaganje vršimo po vrednosti slučajnog vektora $\mathbf{Y}$:

$$P(\mathcal{E}) = \sum_{\substack{\mathbf{y},\ w \\ g(\mathbf{y})\neq w}} P(\mathbf{Y}=\mathbf{y}, W=w) = \sum_{\mathbf{y}\in\mathcal{Y}^n} \left[\sum_{\substack{w=1 \\ g(\mathbf{y})\neq w}}^{M} P(W=w|\mathbf{Y}=\mathbf{y}) \right] P(\mathbf{Y}=\mathbf{y})$$
$$= \sum_{\mathbf{y}\in\mathcal{Y}^n} \left[\sum_{w=1}^{M} P(W=w|\mathbf{Y}=\mathbf{y}) - P(W=g(\mathbf{y})|\mathbf{Y}=\mathbf{y}) \right]$$

Na sličan način kao u prethodnoj teoremi zaključujemo da je verovatnoća greške minimalna ukoliko je $P(W=g(\mathbf{y})|\mathbf{Y}=\mathbf{y})$ maksimalno moguće, za svako $\mathbf{y}\in\mathbf{Y}$:

$$g(\mathbf{y})=\hat{w}, \quad P(W=\hat{w}|\mathbf{Y}=\mathbf{y}) = \max_{w=1,2,\ldots,M} P(W=w|\mathbf{Y}=\mathbf{y})$$

Pošto je $P(W=w|\mathbf{Y}=\mathbf{y})$ zapravo $P(\mathbf{X}=\mathbf{x}(w)|\mathbf{Y}=\mathbf{y}) = p(\mathbf{x}(w)|\mathbf{y})$ sledi da je

$$g(\mathbf{y})=\hat{w}, \quad p(\mathbf{x}(\hat{w})|\mathbf{y}) = \max_{w=1,2,\ldots,M} p(\mathbf{x}(w)|\mathbf{y}).$$

Ovim je dokaz završen. □

Dekoder opisan prethodnom teoremom naziva se **MAP (Maximum APosteriori) dekoder** i predstavlja najbolji mogući dekoder u opštem slučaju. Da bi izračunali aposteriornu raspodelu $p(\mathbf{x}(w)|\mathbf{y})$ koristimo sledeći izraz:

$$p(\mathbf{x}(w)|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x}(w))p(\mathbf{x}(w))}{p(\mathbf{y})} = \frac{p(\mathbf{y}|\mathbf{x}(w))p_W(w)}{\sum_{w'=1}^{M} p(\mathbf{y}|\mathbf{x}(w'))p_W(w')}$$

Raspodela $p_W(w)$ je (prema pretpostavci) poznata, dok se $p(\mathbf{y}|\mathbf{x}(w))$ računa na osnovu uslovne raspodele kanala. S obzirom da $p(\mathbf{y})$ ne zavisi od izbora w, dovoljno je naći maksimum:

$$\max_{w=1,2,\ldots,M} p(\mathbf{y}|\mathbf{x}(w))p_W(w).$$

Ukoliko je $p_W(w) = 1/M$ za svako $w = 1, 2, \ldots, M$, maksimizacija $p(\mathbf{x}(w)|\mathbf{y})$ svodi se na maksimizaciju obrnute (apriorne) verovatnoće $p(\mathbf{y}|\mathbf{x}(w))$ po $w = 1, 2, \ldots, M$. Drugim rečima, MAP dekoder se svodi na ML dekoder.

Za dalje čitanje, pogledati:

1. ZERO-ERROR CODES (Cover, p.205)
2. FANO'S INEQUALITY AND THE CONVERSE TO THE CODING THEOREM (Cover, p. 206)
3. EQUALITY IN THE CONVERSE TO THE CHANNEL CODING THEOREM (Cover, p. 208)
4. FEEDBACK CAPACITY (Cover, p. 216)

Glava 8

Zaštitno kodiranje - kodovi

Videli smo da Druga Shannonova teorema tvrdi da je za **slučajno generisan** $(2^{nR}, n)$ **kod** kod koga je $R < C$, verovatnoća greške proizvoljno mala za dovoljno veliku dužinu kodne reči n. Međutim, ona nam ne daje mehanizam kako konstruisati takav kod koji je u isto vreme i praktično upotrebljiv.

Da bi kod bio praktično upotrebljiv, pored male verovatnoće greške, on mora omogućiti i jednostavnu implementaciju kodera i dekodera [1]. Drugim rečima, potrebno je funkcije $w \to \mathbf{x}(w)$ i $g(\mathbf{y})$ realizovati pomoću elementarnih (algebarskih) operacija.

Da bi postigli takvu realizaciju, potrebno je pretpostaviti da su na skupovima $\mathcal{X}$ i $\mathcal{Y}$ definisane odgovarajuće algebarske strukture. To su konačna polja, koje uvodimo u odeljku A.

Pre toga, ograničićemo razmatranje na q-arni simetrični kanal, čija je matrica kanala data sa:

$$\Pi_q = \begin{bmatrix} 1-(q-1)\alpha & \alpha & \cdots & \alpha \\ \alpha & 1-(q-1)\alpha & & \alpha \\ \vdots & & \ddots & \vdots \\ \alpha & \alpha & & 1-(q-1)\alpha \end{bmatrix}_{q\times q}.$$

Podrazumevamo da je $1-(q-1)\alpha > \alpha$ odnosno da je $0 < \alpha < 1/q$. Na osnovu Teoreme 6.4.3, kapacitet ovog kanala jednak je

$$C = \log_2 q - (1-(q-1)\alpha)\log_2(1-(q-1)\alpha) - (q-1)\alpha \log_2 \alpha.$$

[1] U današnje vreme su uobičajene bitske brzine reda veličine 10-100 Gb/s. To znači da je na raspolaganju veoma malo vreme za obradu podataka, pa je potrebno da algoritmi kodiranja i dekodiranja budu najefikasniji mogući.

Za ovaj kanal su skupovi ulaznih i izlaznih simbola jednaki ($\mathcal{X} = \mathcal{Y}$).

Ukoliko nije drugačije naglašeno, pretpostavljamo da se radi o q-arnom simetričnom kanalu. Pored ovog kanala, koristiće se povremeno i binarni brišući kanal, kao i kanal sa adaptivnim Gausovim šumom (AWGN kanal) o kome će kasnije biti više reči.

Na kraju ovog uvodnog dela, napomenimo da se u novije vreme pojavljuju klase kodova koje dostižu Shannonovu granicu, takve da su realizacije kodera i dekodera praktično primenljive. To su najčešće: Reed-Solomonovi kodovi, Turbo kodovi i LDPC kodovi.

8.1 Hammingovo rastojanje i osobine kodova

Ukoliko se radi o q-arnom simetričnom kanalu, optimalni ML dekoder opisan Teoremom 7.5.3 dobija znatno jednostavniju formulaciju. Za ovo nam je potreban pojam **Hammingovog rastojanja** dve reči $\mathbf{x}, \mathbf{y} \in \mathcal{X}^n$:

Definicija 8.1.1. ***Hammingovo rastojanje*** *između reči* $\mathbf{x}, \mathbf{y} \in \mathcal{X}^n$ *jednako je broju pozicija na kojima se ove dve reči razlikuju. Drugim rečima,*

$$d_H(\mathbf{x}, \mathbf{y}) = |\{i \mid x_i \neq y_i, \quad i = 1, 2, \ldots, n\}| .$$

Ukoliko je $\mathcal{X} = \{0, 1, \ldots, q-1\}$*, definiše se i* ***Hammingova težina*** $w_H(\mathbf{x})$ *kao broj ne-nula komponenti vektora* $\mathbf{x}$*, tj.* $w_H(\mathbf{x}) = d_H(\mathbf{x}, \mathbf{0})$.

Nije teško pokazati da je $d_H(\mathbf{x}, \mathbf{y})$ **metrika** na $\mathcal{X}^n$ tj. da važi sledeća lema.

Lema 8.1.1. *Za svako* $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{X}^n$ *važi:*

1. $d_H(\mathbf{x}, \mathbf{y}) = 0$ *akko* $\mathbf{x} = \mathbf{y}$,
2. $d_H(\mathbf{x}, \mathbf{y}) = d_H(\mathbf{y}, \mathbf{x})$,
3. $d_H(\mathbf{x}, \mathbf{z}) \leq d_H(\mathbf{x}, \mathbf{y}) + d_H(\mathbf{y}, \mathbf{z})$.

Dokaz. Svojstva 1 i 2 slede direktno iz definicije. Da bi dokazali svojstvo 3, pretpostavimo da su $\mathcal{I}_1$ i $\mathcal{I}_2$ skupovi indeksa $i = 1, 2, \ldots, n$ na kojima se redom razlikuju $\mathbf{x}$ i $\mathbf{y}$ kao i $\mathbf{y}$ i $\mathbf{z}$. Tada je $x_i = y_i = z_i$ za svako $i \notin \mathcal{I}_1 \cup \mathcal{I}_2$. Ovih elemenata ima bar $n - |\mathcal{I}_1| - |\mathcal{I}_2| = n - d_H(\mathbf{x}, \mathbf{y}) - d_H(\mathbf{y}, \mathbf{z})$. Prema tome,

$$d_H(\mathbf{x}, \mathbf{z}) \leq n - (n - |\mathcal{I}_1| - |\mathcal{I}_2|) = d_H(\mathbf{x}, \mathbf{y}) + d_H(\mathbf{y}, \mathbf{z}).$$

Ovim je dokaz završen. □

Sada možemo formulisati ekvivalentni kriterijum za ML dekodiranje q-arnog simetričnog kanala.

Lema 8.1.2. *ML dekoder za q-arni simetricni kanal dat je sa* $g(\mathbf{y}) = i$ *tako da je* $d_H(\mathbf{x}(i), \mathbf{y})$ *minimalno.*

Dokaz. Neka je $\delta = 1 - (q-1)\alpha$. Podsetimo se da je za q-arni simetrični kanal:

$$p(y|x) = \begin{cases} \delta, & x = y \\ \alpha, & x \neq y \end{cases}$$

kao i

$$\begin{aligned} p(\mathbf{y}|\mathbf{x}) &= p(y_1|x_1)p(y_2|x_2)\cdots p(y_n|x_n) \\ &= \alpha^{d_H(\mathbf{x},\mathbf{y})}\delta^{n-d_H(\mathbf{x},\mathbf{y})} = \delta^n \cdot (\alpha/\delta)^{d_H(\mathbf{x},\mathbf{y})}. \end{aligned}$$

Prethodna jednakost važi jer je kanal bez memorije, kao i zbog činjenice da je $d_H(\mathbf{x}, \mathbf{y})$ vrednosti $p(y_k|x_k)$ jednako α, dok su preostale (njih $n - d_H(\mathbf{x}, \mathbf{y})$) jednake δ.

S obzirom da su α i δ konstante i da je $\alpha/\delta < 1$ (pretpostavka da je $\alpha < 1/q$) vidimo da će $p(\mathbf{y}|\mathbf{x})$ biti maksimalno ukoliko je $d_H(\mathbf{x}, \mathbf{y})$ minimalno. Samim tim je

$$g(\mathbf{y}) = \underset{i=1,2,\ldots,M}{\operatorname{argmax}}\, p(\mathbf{y}|\mathbf{x}(i)) = \underset{i=1,2,\ldots,M}{\operatorname{argmin}}\, d_H(\mathbf{x}(i), \mathbf{y}).$$

Ovim je dokaz završen. □

Primer 8.1.1. Neka je dat kod

$$\mathcal{C} = \{\mathbf{x}(0), \mathbf{x}(1), \mathbf{x}(2), \mathbf{x}(3)\} = \{0000, 0011, 1100, 1111\}$$

i neka je $\mathbf{y} = 1000$. Tada je

$$\begin{aligned} d_H(\mathbf{x}(0), \mathbf{y}) &= d_H(\mathbf{0}000, \mathbf{1}000) = 1, \quad d_H(\mathbf{x}(1), \mathbf{y}) = d_H(\mathbf{00}11, \mathbf{10}00) = 3, \\ d_H(\mathbf{x}(2), \mathbf{y}) &= d_H(1\mathbf{1}00, 1\mathbf{0}00) = 1, \quad d_H(\mathbf{x}(3), \mathbf{y}) = d_H(1\mathbf{111}, 1\mathbf{000}) = 3, \end{aligned}$$

Bitovi koji se razlikuju su boldirani. Dakle, ukoliko smo primili $\mathbf{y} = 1000$, najverovatnije je poslato ili $\mathbf{x}(0)$ ili $\mathbf{x}(2)$. U oba slučaja ($g(\mathbf{y}) = 0$ ili $g(\mathbf{y}) = 2$) dobijamo najbolji mogući dekoder. Implementacija dekodera na način da redom pretražuje kodne reči i računa minimalno Hammingovo rastojanje najčešće nije isplativa, zato što je broj kodnih reči M suviše veliki.

Kada smo uveli pojam rastojanja između dve reči, možemo definisati i "sferu" oko reči na sledeći način:

$$Z_s(\mathbf{x}) = \{\mathbf{y} \in \mathcal{X}^n \mid d_H(\mathbf{x}, \mathbf{y}) \leq s\}.$$

Drugim rečima, elementi sfere $Z_s(\mathbf{x})$ su sve reči $\mathbf{y}$ koje nastaju usled ne više od s grešaka.

Primer 8.1.2. Neka je $\mathbf{x} = 0000$. Tada je

$$Z_1(0000) = \{0000, 0001, 0010, 0100, 1000\}$$
$$Z_2(0000) = Z_1(0000) \cup \{0011, 0101, 1001, 0110, 1010, 1100\}$$

Ukoliko je u pitanju kodna reč $\mathbf{x}(i)$, ova sfera opisuje reči koje mogu biti primljene, ukoliko je u kanalu došlo do ne više od s grešaka. To nam daje povod za sledeće definicije.

Definicija 8.1.2. *Kod **ispravlja** s **grešaka** ako je $g(Z_s(\mathbf{x}(i))) = \{i\}$ za svako $i = 1, 2, \ldots, M$. Kod **detektuje** s **grešaka** ako je $\mathbf{x}(j) \notin Z_s(\mathbf{x}(i))$ za svako $j \neq i$.*

Drugim rečima, kod ispravlja s grešaka, ali se bilo koja n-torka koja može da nastane od $\mathbf{x}(i)$ usled dejstva s ili manje grešaka slika nazad u i. Sa druge strane, kod detektuje grešku, ukoliko primljena n-torka nije kodna reč. Ukoliko se usled s ili manje grešaka, od jedne ne može dobiti druga kodna reč, onda kod detektuje s grešaka.

Definicija 8.1.3. ***Kodno rastojanje koda** $\mathcal{C}$ definisano je sa*

$$d(\mathcal{C}) = \min_{i \neq j} d_H(\mathbf{x}(i), \mathbf{x}(j)).$$

Drugim rečima, to je minimalno rastojanje dve kodne reči u $\mathcal{C}$.

Kodno rastojanje direktno određuje broj grešaka koje kod može da detektuje odnosno ispravi.

Teorema 8.1.3. *Kod omogućuje detektovanje (ispravljanje) s grešaka, ako je $d(\mathcal{C}) > s$ $(d(\mathcal{C}) > 2s)$.*

Dokaz. Ako je $d(\mathcal{C}) > s$ tada ne postoje dve kodne reči na rastojanju manjem ili jednakom s, pa sfera $Z_s(\mathbf{x}(i))$ ne sadrži nijednu drugu kodnu reč, odnosno kod detektuje s grešaka.

Pretpostavimo da je $d(\mathcal{C}) > 2s$ i neka je $\mathbf{y}$ takvo da je $g(\mathbf{y}) = j \neq i$ i da je $d_H(\mathbf{x}(i), \mathbf{y}) \leq s$. Drugim rečima, da je reč $\mathbf{y}$ primljena usled s ili manje grešaka kada je poslata $\mathbf{x}(i)$, a da je pritom (optimalni) dekoder vratio indeks $j \neq i$. Podsetimo se da optimalni dekoder vraća najbližu (po d_H) kodnu reč (Lema 8.1.2), pa je samim tim $d_H(\mathbf{x}(j), \mathbf{y}) \leq d_H(\mathbf{x}(i), \mathbf{y}) \leq s$. Prema tome,

$$d_H(\mathbf{x}(i), \mathbf{x}(j)) \leq d_H(\mathbf{x}(i), \mathbf{y}) + d_H(\mathbf{y}, \mathbf{x}(j)) \leq 2s$$

što je kontradicija sa pretpostavkom da je $d(\mathcal{C}) > 2s$. □

Teorema 8.1.4. ***(Hammingov uslov)*** *Ako kod ispravlja s grešaka, onda je*

$$M \leq \frac{q^n}{\sum_{k=0}^{s} \binom{n}{k}(q-1)^k}$$

Dokaz. Broj reči u sferi $Z_s(\mathbf{x})$ jednak je [2]

$$A = 1 + n(q-1) + \binom{n}{2}(q-1)^2 + \ldots + \binom{n}{s}(q-1)^s$$

Pošto su sve sfere $Z_s(\mathbf{x}(i))$ međusobno disjunktne za $i = 1, 2, \ldots, M$ (kod ispravlja s grešaka) tada je ukupan broj n-torki u ovim sferama jednak $M \cdot A$. Ovaj broj je manji ili jednak od ukupnog broja svih n-torki $M \cdot A \leq q^n$. Odavde dobijamo $M \leq q^n/A$. □

Zanimljivo je uočiti da ukoliko važi jednakost u prethodnoj nejednakosti za M, da je tada ceo prostor reči $\mathcal{X}^n$ prekriven sferama istog poluprečnika s. Na taj način je M kodnih reči optimalno "razmaknuto". Ovakvi kodovi se nazivaju **perfektni kodovi**.

Primer 8.1.3. Naizgled iznenađujuća činjenica je da i kod $\mathcal{C} = \{000, 111\}$ (ponavljajući (*repetition*) kod) spada u grupu perfektnih kodova. Zaista,

$$Z_1(000) = \{000, 001, 010, 110\}, \quad Z_1(111) = \{111, 110, 101, 011\}.$$

Pošto je $d_H(000, 111) = 3$, ovaj kod (na osnovu Teoreme 8.1.3) ispravlja $s = 1$ grešku. Na sličan način, ponavljajući kod dužine $2k + 1$ ispravlja k grešaka. Međutim, i dalje ostaje činjenica da se ovako dobre performanse plaćaju malim brojem kodnih reči ($M = 2$), odnosno kodnim količnikom koji teži 0.

8.2 Linearni blok kodovi

Druga Shannonova teorema (Teorema 7.5.1) tvrdi da za svaki kodni količnik R koji je manji od kapaciteta kanala C, postoji $(2^{nR}, n)$ kod čija je verovatnoća

[2] Ukoliko je došlo do k grešaka, pozicije na kojima je došlo do greške možemo odabrati na $\binom{n}{k}$ načina, a za svaku poziciju možemo odabrati vrednost da se razlikuje od trenutne, na jedan od $q-1$ načina.

greške proizvoljno mala. Štaviše, ukoliko je n dovoljno veliko, ovakve performanse možemo dobiti slučajnim izborom koda.

Međutim, ovi rezultati su nezadovoljavajući sa praktične strane iz (najmanje) 3 razloga:

- Dobri kodovi se u praksi teško konstruišu. Iako teorema tvrdi da slučajno generisan kod ima dobre performanse, to važi samo ako je n dovoljno veliko, što je potpuno neprimenljivo u praksi (još uvek nije moguće raditi sa kodom dužine 10^6 ili čak 10^9).
- Pojedinačni kod se teško analizira. Da bi izračunali verovatnoću greške konkretnog koda, potrebno je da ispitamo sve moguće izlaze iz kanala $\mathbf{y} \in \mathcal{Y}^n$. Broj mogućih izlaza eksponencijalno raste sa povećanjem dužine n.
- Pored male verovatnoće greške, kod primenljiv u praksi mora da ima (računarski) jednostavne procedure za kodiranje i dekodiranje. Dekoderi koje smo do sada pominjali (**ML** i **MAP** dekoder) očigledno ne zadovoljavaju ovaj uslov, pošto zahtevaju pretragu svih $M = 2^{nR}$ kodnih reči, a to nije moguće efikasno sprovesti za iole veće vrednosti broja n.

Prema tome, efikasan kod moramo potražiti na drugi način. Da bi olakšali analizu, kao i konstrukciju kodera i dekodera, potrebno je da nad samim kodom definišemo neku strukturu. Definisanjem strukture omogućavamo bolju kontrolu performansi koda (i razne druge osobine), kao i mnogo efikasniju konstrukciju kodera i dekodera. Pošto su kodne reči nizovi (vektori) simbola iz određenog skupa $\mathcal{X}$, nad samim simbolima uvodimo algebarsku strukturu **polje**, dok sam kod (odnosno skup kodnih reči) posmatramo kao **vektorski prostor**. Ovako dobijeni kodovi nazivaju se **linearni kodovi**.

Pre nego što krenemo sa linearnim blok kodovima, potrebno je uvesti odgovarajuću strukturu na skupu ulaznih odnosno izlaznih simbola kanala $\mathcal{X} = \mathcal{Y}$. U pitanju je struktura **(konačno) polje**. Precizniju definiciju konačnog polja možete videti u odeljku A (dodatak). Za dalji rad, bitno je znati da polja označavamo sa $\mathbb{F}_q$ ili $GF(q)$, kao i da postoje samo za $q = p^m$ gde je p prost broj. Nad poljem su definisane operacije $+$ i $\cdot$ koje se ponašaju **na isti način** kao i za racionalne brojeve (i ovde imamo oduzimanje, deljenje, elemente 0 i 1, i sve ostalo što imamo i kod racionalnih brojeva).

Neka je $p \in \mathbb{N}$ prost broj. **Konačno polje reda** p je skup $\mathbb{F}_p = \{0, 1, \ldots, p-1\}$ sa operacijama $+_p$ i $\cdot_p$ koje su definisane sa

$$x +_p y = (x + y) \bmod p, \qquad x \cdot_p y = (x \cdot y) \bmod p. \tag{8.1}$$

Konstrukcija polja reda $q = p^m$ prilično komplikovana (videti dodatak A), pa ćemo je preskočiti.

8.2.1 Definicija i osnovne osobine

Pre nego što damo formalnu definiciju linearnog blok koda, posmatrajmo primer Hammingovog koda.

Primer 8.2.1. Posmatrajmo binarni simetrični kanal ($\mathcal{X} = \mathcal{Y} = \{0, 1\}$) i neka je verovatnoća greške u kanalu jednaka α. Prenosimo poruku w koja ima vrednost nekog broja $0, 1, \ldots, 15$ ($M = 16$). Umesto broja w, možemo da pretpostavimo da prenosimo njegovu binarnu reprezentaciju $w = (u_1u_2u_3u_4)_2$. Definišimo kod na sledeći način:

$$\mathbf{x}(w) = (u_1, u_2, u_3, u_4, p_1, p_2, p_3)$$

gde je $p_1 = u_1 \oplus u_2 \oplus u_4$, $p_2 = u_2 \oplus u_3 \oplus u_4$ i $p_3 = u_1 \oplus u_3 \oplus u_4$. Na primer,

$$\mathbf{x}(6) = \mathbf{x}((0110)_2) = 0110101, \quad \mathbf{x}(10) = \mathbf{x}((1010)_2) = 1010110.$$

Vidimo da je definicija kodiranja "algebarska" (koristi se operacija $\oplus$ sabiranja po modulu 2). Nije teško dokazati da je zbir (po modulu 2) dve kodne reči $\mathbf{x}(i) \oplus \mathbf{x}(j)$ takođe kodna reč. zaista

$$\mathbf{x}(6) \oplus \mathbf{x}(10) = 1100011 = \mathbf{x}((0110)_2 \oplus (1010)_2) = \mathbf{x}((1100)_2)$$

Kodni količnik ovog koda je $R = (\log_2 M)/n = 4/7$, a pošto je minimalno rastojanje dve kodne reči jednako 3 (dobija se neposrednom proverom), sledi da kod ispravlja jednu grešku odnosno da je verovatnoća greške proporcionalna sa α^2 (verovatnoća pojavljivanja bar dve greške u kanalu).

Ako uporedimo Hammingov kod sa ponavljajućim kodom $\mathbf{x}(0) = 000$ i $\mathbf{x}(1) = 111$, sledi da za isti red veličine verovatnoće greške (u oba slučaja α^2), Hammingov kod daje znatno veći kodni količnik (4/7 naspram 1/3), odnosno znatno manje unešene redundanse.

Na primeru Hammingovog koda demonstrirali smo osnovno svojstvo **binarnih linearnih kodova**, a to je da za svake dve kodne reči $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{C}$ sledi da je njihov zbir $\mathbf{x}_1 \oplus \mathbf{x}_2 \in \mathcal{C}$ takođe kodna reč. Ukoliko kod nije binarni, ovaj uslov se uopštava na sledeći način:

Definicija 8.2.1. *Kod $\mathcal{C}$ je **linearni kod (linearni blok kod)** ukoliko za svake dve kodne reči $\mathbf{x}_1$ i $\mathbf{x}_2$ sledi da je $\alpha\mathbf{x}_1 + \beta\mathbf{x}_2 \in \mathcal{C}$ takođe kodna reč za*

svako $\alpha, \beta \in \mathcal{X}$.

Pritom je operacija + definisana "pokoordinatno" nad vektorima (kodnim rečima) $\mathbf{x}_1$ i $\mathbf{x}_2$, dok su njihove koordinate elementi skupa $\mathcal{X} = \{0, 1, \ldots, q-1\}$ nad kojim uvedena struktura konačnog polja $\mathbb{F}_q$. Prema tome, struktura koja je uvedena na samom kodu $\mathcal{C}$ je **vektorski prostor**!

Vratimo se sada ponovo na Hammingov kod opisan u prethodnom primeru. Operaciju kodiranja možemo da zapišemo na sledeći način:

$$\begin{aligned}\mathbf{x}(w) &= (u_1, u_2, u_3, u_4, u_1+u_2+u_4, u_2+u_3+u_4, u_1+u_3+u_4)\\ &= u_1 \underbrace{(1,0,0,0,1,0,0)}_{\mathbf{g}_1} + u_2 \underbrace{(0,1,0,0,1,1,1)}_{\mathbf{g}_2}\\ &+ u_3 \underbrace{(0,0,1,0,0,1,1)}_{\mathbf{g}_3} + u_4 \underbrace{(0,0,0,1,1,1,1)}_{\mathbf{g}_4}\\ &= u_1\mathbf{g}_1 + u_2\mathbf{g}_2 + u_3\mathbf{g}_3 + u_4\mathbf{g}_4\end{aligned} \tag{8.2}$$

Svaka kodna reč $\mathbf{x}(w)$ može da se zapiše na jedinstven način kao linearna kombinacija kodnih reči $\mathbf{g}_1, \mathbf{g}_2, \mathbf{g}_3$ i $\mathbf{g}_4$. Prema tome, Hammingov kod $\mathcal{C}$ je vektorski prostor **dimenzije** $\dim \mathcal{C} = k = 4$.

Definicija 8.2.2. *Ukoliko je linearni blok kod* $\mathcal{C}$ *dimenzije* $k = \dim \mathcal{C}$*, tada ovaj kod označavamo i kao* (n, k)*-kod.*

Kao što smo malopre zaključili, Hammingov kod iz primera 8.2.1 je $(7, 4)$-kod. Ukoliko je dimenzija nekog koda $\mathcal{C}$ jednaka k, to znači da postoje k linearno nezavisnih vektora (kodnih reči) $\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_k$ takvih da svaku kodnu reč $\mathbf{x} \in \mathcal{C}$ možemo zapisati **na jedinstven način** kao njihovu linearnu kombinaciju:

$$\mathbf{x} = \alpha_1\mathbf{g}_1 + \alpha_2\mathbf{g}_2 + \ldots + \alpha_k\mathbf{g}_k$$

gde su $\alpha_1, \alpha_2, \ldots, \alpha_k \in \mathcal{X} = \mathbb{F}_q$. Postojanje vektora $\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_k$ sledi iz poznate činjenice da svaki vektorski prostor ima bazu, i ovaj izbor ne mora da bude jedinstven. Pošto ima ukupno q^k mogućnosti za koeficijente $\alpha_1, \alpha_2, \ldots, \alpha_k$ (svaki može da uzme proizvoljnu vrednost iz skupa $\{0, 1, \ldots, q-1\}$), sledi da (n, k)-kod ima ukupno $M = q^k$ elemenata, pa je kodni količnik jednak

$$R = \frac{\log_2 M}{n} = \frac{k \log_2 q}{n}.$$

Pretpostavimo da je indeks poruke w broj iz skupa $\{0, 1, \ldots, M-1\}$ (umesto iz skupa $\{1, 2, \ldots, M\}$), i neka je $(u_k u_{k-1} \ldots u_1)_q$ reprezentacija ovog broja u sistemu sa osnovom q. Tada kodiranje $\mathbf{x}(w)$ možemo da posmatramo i kao funkciju koja niz (vektor) $\mathbf{u} = (u_1, u_2, \ldots u_k)$ dužine k prevodi u neki drugi niz $\mathbf{x}(\mathbf{u})$ dužine n.

8.2.2 Generatorska matrica koda

Bez gubitka opštosti[3] možemo da pretpostavimo da je kodiranje dato sličan način kao kod Hammingovog $(7,4)$ koda (izraz 8.2):

$$\mathbf{x}(w) = \mathbf{x}((u_k u_{k-1} \dots u_1)_q) = u_1\mathbf{g}_1 + u_2\mathbf{g}_2 + \dots + u_k\mathbf{g}_k$$

Ovo kodiranje možemo kraće da zapišemo kao $\mathbf{x}(\mathbf{u}) = \mathbf{u}G$, pri čemu je G matrica čije su vrste vektori $\mathbf{g}_1, \mathbf{g}_2, \dots, \mathbf{g}_k$:

$$G = \begin{bmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \\ \vdots \\ \mathbf{g}_k \end{bmatrix} \in \mathcal{X}^{k\times n}.$$

Ova matrica ima format $k \times n$ i naziva se **generatorska (generišuća) matrica** koda $\mathcal{C}$. Vektor $\mathbf{u}$ koji se sastoji od cifara poruke w ubuduće ćemo posmatrati kao vektor vrstu

$$\mathbf{u} = \begin{bmatrix} u_1 & u_2 & \cdots & u_k \end{bmatrix}$$

koji se sastoji od **informacionih simbola** $u_1, u_2, \dots, u_k$. U slučaju $q = 2$, u pitanju su **informacioni bitovi**.

Primer 8.2.2. Ponavljajući kod $\mathcal{C} = \{000, 111\}$ definisan sa $\mathbf{x}(\alpha) = (\alpha, \alpha, \alpha) = \alpha(1,1,1)$ ima generatorsku matricu

$$G_1 = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}.$$

Takođe, proizvoljna matrica formata $k \times n$ koja ima linearno nezavisne vrste predstavlja generišuću matricu nekog (n,k) koda. Na primer, matrica

$$G_2 = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

generiše $(5,3)$ kod čija je funkcija kodiranja data sa

$$\begin{aligned} \mathbf{x}(u_1u_2u_3) = \mathbf{u}G_2 &= \begin{bmatrix} u_1 & u_2 & u_3 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \\ &= \begin{bmatrix} u_1+u_3 & u_1+u_3 & u_1+u_2+u_3 & u_2+u_3 & u_3 \end{bmatrix} \end{aligned}$$

[3] S obzirom da se prenumeracijom indeksa w, odnosno promenom sistema kodiranja $\mathbf{x}(w)$ ne menjaju performanse koda

Naravno, i ovde poistovećujemo vektor vrstu sa uređenom n-torkom (u ovom slučaju petorkom) elemenata iz skupa $\mathcal{X} = \{0, 1, \ldots, q-1\}$ (odnosno $\mathcal{X} = \{0, 1\}$ u konkretnom slučaju). Hammingov $(7, 4)$ kod ima generatorsku matricu

$$G_3 = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

čije su vrste vektori $\mathbf{g}_1, \mathbf{g}_2, \mathbf{g}_3$ i $\mathbf{g}_4$ definisani izrazom (8.2). Odatle trivijalno sledi da je kodiranje dato izrazom

$$\begin{aligned} \mathbf{x}(\mathbf{u}) = \mathbf{u}G_3 &= \begin{bmatrix} u_1 & u_2 & u_3 & u_4 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix} \\ &= \begin{bmatrix} u_1 & u_2 & u_3 & u_4 & u_1+u_2+u_4 & u_2+u_3+u_4 & u_1+u_3+u_4 \end{bmatrix} \end{aligned}$$

Primetimo da su prva četiri elementa kodne reči upravo informacioni bitovi u_1, u_2, u_3 i u_4. Ovakvo kodiranje naziva se **sistematsko kodiranje**.

Svako kodiranje $\mathbf{x}(\mathbf{u})$ koje ima svojstvo da se informacioni simboli $\mathbf{u}$ pojavljuju u kodnoj reči $\mathbf{x}(\mathbf{u})$ naziva se **sistematsko kodiranje**. Kod sistematskog kodiranja, informaciju prenosimo tako što pored informacionih simbola $\mathbf{u}$ dodajemo još nekoliko simbola koji su neka funkcija (linearna ili nelinearna) od $\mathbf{u}$.

Pošto radimo isključivo sa kanalima bez memorije, performanse koda se ne menjaju ukoliko zamenimo redosled simbola u kodnoj reči. Tako da, bez gubitka opštosti, možemo pretpostaviti da su kod sistematskog kodiranja, informacioni simboli $\mathbf{u}$ na početku kodne reči $\mathbf{x}(\mathbf{u})$, odnosno da je

$$\mathbf{x}(\mathbf{u}) = \begin{bmatrix} \mathbf{u} & \mathbf{u}A \end{bmatrix} = \mathbf{u} \begin{bmatrix} I_k & A \end{bmatrix}, \quad A \in \mathcal{X}^{k \times (n-k)}.$$

Generatorska matrica $G = [I_k \; A]$ naziva se **sistematska generatorska matrica** (n, k) koda $\mathcal{C}$.

Svaka generatorska matrica G nekog koda može da se svede na sistematsku matricu postupkom **Gausove eliminacije**. U pitanju je potpuno isti postupak koji se koristi za rešavanje sistema linearnih jednačina, i u kom se matrica

svodi na **stepenastu matricu** sledećeg oblika:

$$\begin{bmatrix} 1 & & * & & * & & * \\ & 1 & * & & * & & * \\ & & & 1 & * & & * \\ & & & & & 1 & * \end{bmatrix}$$

Zvezdica označava proizvoljni (u opštem slučaju nenula) element a prazno polje nula element. Dozvoljene operacije nad matricom su:

E1: Dodavanje (oduzimanje) vrste pomnožene nekom vrednošću drugoj vrsti (od druge vrste), tj. $V_j \pm = \alpha V_i$;

E2: Zamena redosleda vrsta, tj. $V_i \leftrightarrow V_j$.

Ove operacije ne menjaju sam kod (kao skup kodnih reči) već samo operaciju kodiranja. Ukoliko se uvede i operacija zamene kolona ($K_i \leftrightarrow K_j$), onda se menja i redosled simbola u kodnim rečima, ali kao što smo ranije napomenuli, to ne utiče na performanse samog koda. Ovom poslednjom transformacijom, dobija se sistematska matrica oblika $[I_k \ A]$.

U svakom koraku, proverimo da li je i-ti element tekuće kolone j različit od nule, i ako nije, menjamo i-tu vrstu sa k-tom ($k > i$) tako da je $g_{kj} \neq 0$. Ukoliko su svi elementi počev od i-tog jednaki 0, prelazimo na sledeću kolonu. U suprotnom, svakoj vrsti počev od $i+1$-ve dodajemo i-tu pomnoženu odgovarajućim brojem tako da u j-toj koloni ostanu sve nule. Zatim prelazimo na sledeću kolonu a i povećavamo za 1.

Primer 8.2.3. Posmatrajmo ponovo matricu G_2 koda uvedenu u prethodnom primeru i svedimo je na sistematski oblik. Primenom prethodno opisanih transformacija dobijamo:

$$\begin{bmatrix} \boxed{1} & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \xrightarrow{V_3+=V_1} \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & \boxed{1} & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \xrightarrow{V_1+=V_2} \begin{bmatrix} 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & \boxed{1} & 1 \end{bmatrix}$$

$$\xrightarrow[V_1+=V_3]{V_2+=V_3} \begin{bmatrix} 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \xrightarrow[K_3\leftrightarrow K_4]{K_2\leftrightarrow K_3} \begin{bmatrix} 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 \end{bmatrix} = G_2'$$

Ovoj matrici odgovara sledeće kodiranje $\mathbf{x}'(\mathbf{u}) = \mathbf{u}G_2' = [u_1 \ u_2 \ u_3 \ u_1 \ u_1 + u_2 + u_3]$.

Prilikom zamene kolona, poželjno je ažurirati i trenutni redosled kolona originalne matrice. Naime, neka je $\mathbf{col}(k)$ pozicija k-te kolone originalne matrice G_2 u matrici G_2'. Na početku je $\mathbf{col} = (1, 2, 3, 4, 5)$. Prilikom zamene kolona $K_2 \leftrightarrow K_3$, menjamo i vrednosti odgovarajućih indeksa $\mathbf{col}(2)$ i $\mathbf{col}(3)$. Nakon zamene je $\mathbf{col} = (1, \mathbf{3}, \mathbf{2}, 4, 5)$. Na isti način, prilikom zamene $K_3 \leftrightarrow K_4$ menjamo vrednosti $\mathbf{col}(3)$ i $\mathbf{col}(4)$, pa je onda $\mathbf{col} = (1, 3, \mathbf{4}, \mathbf{2}, 5)$. Dobijeni redosled kolona može da posluži pri konstrukciji kontrolne matrice koda.

8.2.3 Kontrolna matrica koda

Posmatrajmo ponovo kod definisan sistematskom generatorskom matricom G_2'. Svaka reč ovog koda je oblika

$$\mathbf{u}G_2' = [u_1 \ u_2 \ u_3 \ u_1 \ u_1 + u_2 + u_3] = [x_1 \ x_2 \ x_3 \ x_4 \ x_5]$$

i zadovoljava jednačine (imajući u vidu da je kod binarni i da je sabiranje isto što i oduzimanje):

$$\begin{aligned} x_1 + x_4 &= 0 \\ x_1 + x_2 + x_3 + x_5 &= 0 \end{aligned}$$

Ovaj sistem može da se napiše u matričnom obliku na sledeći način:

$$\underbrace{\begin{bmatrix} 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 \end{bmatrix}}_{H_2'} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

odnosno $H_2'\mathbf{x}^T = \mathbf{0}^T$. Matrica H_2' naziva se **kontrolna matrica koda**. Naziv potiče iz činjenice pomoću ove matrice (odnosno izraza $H_2'\mathbf{x}^T = \mathbf{0}^T$) direktno proveravamo da li je $\mathbf{x}$ kodna reč ili nije. Koristi se i engleski naziv ***parity check* matrica**.

U opštem slučaju, **kontrolna matrica** koda je svaka matrica $H \in \mathcal{X}^{(n-k)\times n}$ takva da je $\mathbf{x} \in \mathcal{C}$ ako i samo ako je $H\mathbf{x}^T = \mathbf{0}^T$.

Teorema 8.2.1. *Neka je G generatorska matrica koda $\mathcal{C}$ a $\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_k$ vrste generatorske matrice. Tada je $H \in \mathcal{X}^{(n-k)\times n}$ kontrolna matrica koda, ako i samo ako je $H\mathbf{g}_i^T = \mathbf{0}^T$ za svako $i = 1, 2, \ldots, k$. Odnosno, u matričnom obliku, ako i samo ako je $HG^T = \mathbb{O}$.*

Dokaz. S obzirom da su vrste $\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_k$ generatorske matrice G kodne reči, i da svaka druga kodna reč može da se napiše kao linearna kombinacija

$$\mathbf{x} = u_1\mathbf{g}_1 + u_2\mathbf{g}_2 + \ldots + u_k\mathbf{g}_k,$$

zaključujemo da je H kontrolna matrica ako i samo ako važi $H\mathbf{g}_i^T = \mathbf{0}^T$ za svako $i = 1, 2, \ldots, k$. Ovaj uslov može da se napiže u matričnom obliku

$$H\begin{bmatrix}\mathbf{g}_1^T & \mathbf{g}_2^T & \cdots & \mathbf{g}_k^T\end{bmatrix} = HG^T = \mathbb{O},$$

odakle zaključujemo da važi drugi deo teoreme. □

Kontrolna matrica može jednostavno da se odredi iz sistematske generišuće matrice G' koda na sledeći način:

$$G' = \begin{bmatrix}I_k & A\end{bmatrix} \quad \longrightarrow \quad H' = \begin{bmatrix}-A^T & I_{n-k}\end{bmatrix}.$$

Da je H' zaista kontrolna matrica sledi iz identiteta

$$H'G'^T = \begin{bmatrix}-A^T & I_{n-k}\end{bmatrix}\begin{bmatrix}I_k \\ A^T\end{bmatrix} = -A^T + A^T = \mathbb{O}.$$

i Teoreme 8.2.1. Ukoliko matrica G nije sistematska, onda odgovarajuću kontrolnu matricu možemo da odredimo na sledeći način:

1. Odredimo sistematsku matricu $G' = [I_k\ A]$ primenom postupka Gausove eliminacije opisanog u prethodnom pododeljku. Pritom, odredimo i redosled kolona **col** početne matrice G u sistematskoj matrici G'.
2. Formiramo sistematsku kontrolnu matricu $H' = [-A^T\ I_{n-k}]$.
3. Primenimo obrnuti redosled kolona $\mathbf{col}^{-1}$ na matricu H'. Zapravo, formiramo matricu H tako da je na k-ta kolona matrice H' na $\mathbf{col}(k)$-tom mestu u matrici H.

S obzirom da Gausova eliminacija ne menja kod, a da zamena kolona odgovara zameni indeksa kodnih reči $\mathbf{x}$, zaključujemo da je ovako konstruisana matrica H zaista kontrolna matrica koja odgovara originalnom kodu koji je generisan matricom G.

U Primeru 8.2.3 odredili smo redosled kolona $\mathbf{col} = (1, 3, 5, 4, 2)$ za sistematsku matricu G_2'. Na osnovu prethodno opisane procedure, možemo da odredimo kontrolnu matricu koja odgovara originalnoj matrici G_2:

$$H_2' = \begin{array}{c}\begin{matrix}1 & 3 & 4 & 2 & 5\end{matrix} \\ \begin{bmatrix}1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1\end{bmatrix}\end{array} \qquad \longrightarrow \qquad H_2 = \begin{array}{c}\begin{matrix}1 & 2 & 3 & 4 & 5\end{matrix} \\ \begin{bmatrix}1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1\end{bmatrix}\end{array}$$

Kontrolna matrica nije jedinstvena. Zapravo, svaka matrica dobijena primenom elementarnih transformacija **E1** i **E2** na matricu H je takođe kontrolna matrica istog koda.

Primer 8.2.4. Kontrolna matrica koja odgovara ponavljajućem kodu (sa generatorskom matricom $G_1 = [1\ 1\ 1]$) jednaka je

$$H_1 = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}.$$

Ovo je očigledno, s obzirom da su kodne reči (000 i 111) ovog koda okarakterisane sistemom jednačina $x_2 = x_1$ i $x_3 = x_1$. Kontrolna matrica Hammingovog $(7, 4)$ koda data je sa

$$H_3 = \begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}$$

ali je i matrica

$$H_3' = \begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

koja se dobija dodavanjem druge vrste trećoj u matrici H_3, takođe kontrolna matrica ovog koda.

8.2.4 Kodno rastojanje linearnih blok kodova

Podsetimo se da je kodno rastojanje proizvoljnog koda $\mathcal{C}$ definisano kao minimalno Hammingovo rastojanje $d_H(\mathbf{x}, \mathbf{y})$ dve proizvoljne **različite** kodne reči $\mathbf{x}$ i $\mathbf{y}$.

U slučaju linearnih blok kodova, ovu veličinu možemo jednostavnije da izračunamo, a ona je i direktno povezana sa kontrolnom matricom koda H. O tome govore naredna lema i teorema.

Lema 8.2.2. *Kodno rastojanje $d(\mathcal{C})$ kod linearnih blok kodova dato je sa* $d(\mathcal{C}) = \min\limits_{\substack{\mathbf{x}\in\mathcal{C} \\ \mathbf{x}\neq\mathbf{0}}} w_H(\mathbf{x})$.

Dokaz. Označimo sa $m = \min\limits_{\substack{\mathbf{x}\in\mathcal{C} \\ \mathbf{x}\neq\mathbf{0}}} w_H(\mathbf{x})$ i $\delta = d(\mathcal{C}) = \min\limits_{\substack{\mathbf{x},\mathbf{y}\in\mathcal{C} \\ \mathbf{x}\neq\mathbf{y}}} d_H(\mathbf{x}, \mathbf{y})$. Neka su $\mathbf{x}, \mathbf{y} \in \mathcal{C}$ dve proizvoljne različite kodne reči. Tada je

$$d_H(\mathbf{x}, \mathbf{y}) = d_H(\mathbf{x} - \mathbf{y}, \mathbf{0}) = w_H(\mathbf{x} - \mathbf{y}) \geq m$$

jer je $\mathbf{z} = \mathbf{x} - \mathbf{y} \in \mathcal{C}$ kodna reč (kod $\mathcal{C}$ je linearan). Odatle sledi da je i $\delta \geq m$, pošto je δ minimum svih $d_H(\mathbf{x}, \mathbf{y})$ po svim različitim kodnim rečima $\mathbf{x}$ i $\mathbf{y}$. Sa druge strane je

$$w_H(\mathbf{z}) = d_H(\mathbf{z}, \mathbf{0}) \leq \delta$$

pa je $m \leq \delta$, odnosno $m = \delta$ što je i trebalo dokazati. □

Teorema 8.2.3. *Kodno rastojanje $d(\mathcal{C})$ je minimalni broj linearno zavisnih kolona matrice H.*

Dokaz. Neka su $\mathbf{h}_1^T, \mathbf{h}_2^T, \ldots, \mathbf{h}_n^T$ kolone matrice H. Pošto je $H\mathbf{x}^T = \mathbf{0}^T$ ekvivalentno izrazu

$$\mathbf{h}_1^T x_1 + \mathbf{h}_2^T x_2 + \ldots + \mathbf{h}_n^T x_n = \mathbf{0}^T$$

dobijamo da svakoj kodnoj reči $\mathbf{x}$ sa k ne-nula elemenata odgovara skup k linearno zavisnih kolona matrice H. Obrnuto, za svaki skup od k linearno zavisnih kolona, postoji odgovarajuća linearna kombinacija koja je jednaka nuli, i koja se (nulama) može dopuniti do kodne reči $\mathbf{x}$. Na ovaj način smo dokazali da je minimalni broj linearno nezavisnih kolona matrice H jednak minimalnom broju ne-nula elemenata neke kodne reči $\mathbf{x}$. Tvrđenje teoreme dobijamo direktno na osnovu prethodne leme. □

Primer 8.2.5. Videli smo da su kontrolne matrice ponavljajućeg $(3, 1)$ i Hammingovog $(7, 4)$ koda jednake

$$H_1 = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}, \quad H_3 = \begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}$$

Minimalni broj linearno zavisnih kolona je za obe matrice jednak 3 (sve tri kolone za H_1 kao i treća, četvrta i peta za H_3), pa je kodno rastojanje oba koda jednako $d(\mathcal{C}_1) = d(\mathcal{C}_3) = 3$.

8.2.5 Dekodiranje pomoću sindroma

Sada ćemo opisati postupak konstrukcije **ML** dekodera za linearne blok kodove. Podsećanja radi, dokazali smo (Teorema 7.5.3) da je **ML** dekoder najbolji mogući dekoder (dekoder koji daje najmanju verovatnoću greške) ukoliko je raspodela poruka w (odnosno, u ovom slučaju informacionih bitova $\mathbf{u}$) uniformna. Takođe, u slučaju q-arnog simetričnog kanala, pokazali smo da se

ML dekoder svodi na nalaženje kodne reči $\mathbf{x} \in \mathcal{C}$ takve da je Hammingovo rastojanje $d_H(\mathbf{x}, \mathbf{y})$ najmanje, gde je $\mathbf{y}$ reč koju smo primili iz kanala.

Ovaj problem je ekvivalentan nalaženju **korekcionog vektora e** takvog da je $d_H(\mathbf{y} - \mathbf{e}, \mathbf{y})$ najmanje i pritom da je $\mathbf{x} = \mathbf{y} - \mathbf{e} \in \mathcal{C}$ kodna reč. Dakle, potrebno je naći minimum

$$\min_{\mathbf{y}-\mathbf{e}\in\mathcal{C}} d_H(\mathbf{y} - \mathbf{e}, \mathbf{y}) = \min_{\mathbf{y}-\mathbf{e}\in\mathcal{C}} w_H(\mathbf{e})$$

Uslov da je $\mathbf{x} = \mathbf{y} - \mathbf{e}$ kodna reč najlakše ćemo proveriti (tj. obezbediti) pomoću kontrolne matrice koda H. Pošto je $\mathbf{x} \in \mathcal{C}$ ako i samo ako je $H\mathbf{x}^T = \mathbf{0}^T$, sledi da je $\mathbf{y} - \mathbf{e} \in \mathcal{C}$ ako i samo ako je

$$\mathbf{0}^T = H(\mathbf{y} - \mathbf{e})^T = H\mathbf{y}^T - H\mathbf{e}^T$$

odnosno ako i samo je $H\mathbf{y}^T = H\mathbf{e}^T$. Vektor $\mathbf{s}^T = H\mathbf{y}^T$ možemo da izračunamo na osnovu primljene reči $\mathbf{y}$ i on se naziva **sindrom**, dok se vektor $\mathbf{e}$ takav da je $w_H(\mathbf{e})$ minimalno i $H\mathbf{e}^T = \mathbf{s}^T$ naziva **korektor**.

Prema tome, problem dekodiranja svodi se da za datu vrednost sindroma $\mathbf{s}^T = H\mathbf{y}^T$ izračunamo vrednost korektora $\mathbf{e}$ koji ima minimalno ne-nula elemenata ($w_H(\mathbf{e})$ minimalno) i za koji važi $\mathbf{s}^T = H\mathbf{e}^T$. Dakle, među q^k različitih vektora[4] $\mathbf{e}$ koji zadovoljavaju $\mathbf{s}^T = H\mathbf{e}^T$, potrebno je naći onaj koji ima najmanju Hammingovu težinu. Ovaj problem nije moguće efikasno rešiti u opštem slučaju. Ukoliko su n i k relativno mali brojevi, onda možemo unapred da izračunamo korektore za svaki mogući sindrom, i da ih zapamtimo u jedan niz.

Primer 8.2.6. Posmatrajmo ponovo kod $\mathcal{C}_2$ sa kontrolnom matricom:

$$H_2' = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 \end{bmatrix}.$$

Postupak nalaženja korektora svodi se na to da se generišu sve moguće vrednosti vektora $\mathbf{e} \in \mathcal{X}^n$, za svaku da se izračuna Hammingova težina $w_H(\mathbf{e})$ kao i sindrom $\mathbf{s}^T = H_2'\mathbf{e}^T$. Onda se za svaku vrednost sindroma $\mathbf{s}$ uzme vektor $\mathbf{e}$ sa najmanjom Hammingovom težinom:

[4] Nije teško dokazati da ovakvih vektora ima tačno q^k. Ukoliko je $\mathbf{e}_0$ jedan takav vektor, onda je $\mathbf{e} - \mathbf{e}_0 \in \mathcal{C}$ za proizvoljni drugi vektor $\mathbf{e}$. Zaista, $H(\mathbf{e} - \mathbf{e}_0)^T = H\mathbf{e}^T - H\mathbf{e}_0^T = \mathbf{s}^T - \mathbf{s}^T = \mathbf{0}^T$ pa je $\mathbf{e} - \mathbf{e}_0 \in \mathcal{C}$. Dakle, $\mathbf{e} = \mathbf{e}_0 + \mathbf{x}$ gde je $\mathbf{x} \in \mathcal{C}$, pa ovih vektora ima koliko i kodnih reči u kodu $\mathcal{C}$, odnosno q^k.

$\mathbf{e}$	$w_H(\mathbf{e})$	$\mathbf{s}^T = H_2'\mathbf{e}^T$
00000	0	00
00001	1	01
00010	1	10
00100	1	01
01000	1	01
10000	1	11
⋮	⋮	⋮
11111	5	00

$\longrightarrow$

$\mathbf{s}$	$\mathbf{e}_{min}(\mathbf{s})$
00	00000
01	00001
10	00010
11	10000

Da bi uštedeli na vremenu, najbolje je generisati vektore $\mathbf{e}$ sa rastućim Hammingovim težinama (kao u tabeli iznad), i ukoliko se neka vrednost sindroma prvi put pojavi, upisati odgovarajući korektor u tabelu sa desne strane. Generisanje zaustavljamo kada ispunimo desnu tabelu. Ukoliko dva vektora $\mathbf{e}$ iste Hammingove težine daju isti sindrom, možemo uzeti bilo koji za odgovarajući korektor.

Kada generišemo tabelu korektora, proces dekodiranja je jednostavan:

1. Izračunamo sindrom $\mathbf{s}^T := H\mathbf{y}^T$;
2. Pročitamo iz tabele korektor $\mathbf{e} := \mathbf{e}_{min}(\mathbf{s})$;
3. Izračunamo kodnu reč $\mathbf{x} := \mathbf{y} - \mathbf{e}$.

Ukoliko je kodiranje $\mathbf{x}(\mathbf{u})$ sistematsko, tada se informacioni vektor $\mathbf{u}$ direktno čita iz kodne reči $\mathbf{x}$. U suprotnom, potrebno je primeniti inverznu transformaciju i na osnovu kodne reči $\mathbf{x}$ izračunati vektor $\mathbf{u}$.

Primer 8.2.7. Pretpostavimo da želimo da pošaljemo informaciju $\mathbf{u} = 100$ kroz binarni simetrični kanal sa verovatnoćom greške $\alpha = p(0|1) = p(1|0) = 0.1$. Koristimo kod $\mathcal{C}_2$ sa sistematskom generatorskom matricom:

$$G_2' = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 \end{bmatrix}.$$

Kodnu reč koja odgovara informacionom vektoru $\mathbf{u}$ dobijamo množenjem

$$\mathbf{x}(\mathbf{u}) = \mathbf{u}G_2' = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 \end{bmatrix}$$

Prema tome, reč $\mathbf{x} = 10011$ šaljemo kroz kanal.

Pretpostavimo sada da je tokom slanja došlo do greške u poslednjem bitu poruke, odnosno da je na prijemnoj strani stigla reč $\mathbf{y} = 10010$. Računamo sindrom:

$$\mathbf{s}^T = H_2'\mathbf{y}^T = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Iz tabele (Primer 8.2.6) korektora čitamo da je za sindrom $\mathbf{s} = 01$ korektor $\mathbf{e} = 00001$ i računamo $\mathbf{x}_{dec} = \mathbf{y} - \mathbf{e} = 10011$. Vidimo da je dekodirana kodna reč ista kao ona koju smo poslali. Pošto je u pitanju sistematsko kodiranje, informacioni bitovi su prva 3 bita dekodirane kodne reči, odnosno $\mathbf{u}_{dec} = 100 = \mathbf{u}$.

Kada bi do greške došlo na 3. poziciji umesto na petoj $\mathbf{y} = 10\mathbf{1}11$, tada bi sindrom ponovo bio jednak $\mathbf{s} = 01$ pa bi dekodirana kodna reč bila $\mathbf{x}_{dec} = 10110$ odnosno na prijemu bi pročitali poruku $\mathbf{u}_{dec} = 101$ koja je različita od poslate poruke $\mathbf{u} = 100$. Vidimo da ovaj kod nije uspeo da ispravi grešku na 3. poziciji, čime potvrđujemo zaključak da on u opštem slučaju ne može da ispravi jednu grešku.

Na kraju, napomenimo još jednom da je generisanje tabele korektora moguće samo za relativno male vrednosti n i k. Navodimo dva glavna razloga za ovu tvrđenje:

1. Da bi generisali tabelu, moramo da ispitamo (u najgorem slučaju) svih q^n mogućih korektora. Iako se ovaj maksimum nikada ne dostiže (imajući u vidu proceduru opisanu u Primeru 8.2.6), postupak je efektivno primenljiv samo za kratke kodove (recimo za $n \leq 30$).
2. Čak i da zanemarimo vreme potrebno za generisanje tabele korektora (ovo se radi samo jednom), količina potrebne memorije da bi se tabela zapamtila eksponencijalno raste sa k i n. Preciznije, pošto je sindrom $\mathbf{s} \in \mathcal{X}^{n-k}$, postoji q^{n-k} različitih vrednosti sindroma, i za svaki od njih, potrebno je zapamtiti korektor $\mathbf{e}$ dužine n. To je ukupno nq^{n-k} simbola koje je potrebno zapamtiti.

U praktičnim primenama koriste se kodovi sa kodnim rečima dužine više stotina, pa i hiljada bitova. Efikasna implementacija dekodera za ove kodove podrazumeva da se, u opštem slučaju, odreknemo najboljeg mogućeg (**ML**)

dekodera. Tada primenjujemo dekoder sa nešto slabijim performansama, ali koji omogućava efikasnu implementaciju koja može da radi u realnom vremenu. Više reči o ovim alternativnim dekoderima biće nešto kasnije, kada budemo razmatrali LDPC kodove.

8.3 Hammingovi kodovi

Podsetimo se da su generišuća i kontrolna matrica Hammingovog $(7, 4)$ koda date sa

$$G_3 = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix} \qquad H_3 = \begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}.$$

S obzirom da matrica H_3 nema dve linearno zavisne kolone, a postoje 3 takve (Primer 8.2.5), Hammingov $(7, 4)$ kod ima kodno rastojanje $d(\mathcal{C}_3) = 3$ odnosno ispravlja jednu grešku.

Ovaj kod je prvi element klase Hammingovih kodova, koji ispravljaju tačno jednu grešku. Pre neko što pokažemo kako se ovi kodovi konstruišu, primetimo da se svih $2^3 - 1 = 7$ ne-nula vektora nalaze među kolonama matrice H_3. Posmatrajmo sad matricu $H_{3,h}$ koja se dobija permutovanjem kolona matrice H_3, tako da je i-ta kolona obrnuta binarna reprezentacija broja i $(i = 1, 2, \ldots, 7)$:

$$H_{3,h} = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

Pretpostavimo sada da je prilikom slanja (npr. kodne reči $\mathbf{x} = 0000000$) došlo do greške u četvrtom bitu, odnosno da je $\mathbf{e} = 0001000$. Tada je sindrom jednak

$$\mathbf{s}^T = H_{3,h}\mathbf{e}^T = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

odnosno $\mathbf{s} = 001$ što predstavlja obrnutu binarnu reprezentaciju pozicije $i = 4 = (100)_2$ na kojoj je došlo do greške. Na potpuno isti način vidimo da

isti zaključak važi za bilo koju poziciju $i = 1, 2, \ldots, 7$. Prema tome, sindrom dekodiranje kod ovog koda je najjednostavnije moguće i ne zahteva formiranje niza korektora. Zapravo, za svaki sindrom $\mathbf{s}$, odgovarajući korektor $\mathbf{e}_{min}(\mathbf{s})$ je broj i, čija je obrnuta binarna reprezentacija upravo $\mathbf{s}$.

Generišuću matricu, koja odgovara matrici $H_{3,h}$ dobijamo tako što odgovarajuću permutaciju kolona (kojom smo $H_{3,h}$ dobili iz H_3) primenimo na kolone matrice G_3. Vidimo da je u pitanju sledeća permutacija $\mathbf{col} = (3, 5, 6, 7, 1, 2, 4)$ pa je onda

$$G_3 = \begin{array}{c} \begin{matrix} 3 & 5 & 6 & 7 & 1 & 2 & 4 \end{matrix} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix} \end{array} \quad \longrightarrow \quad G_{3,h} = \begin{array}{c} \begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{matrix} \\ \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

Odgovarajuće kodiranje je dato sa $\mathbf{x}(\mathbf{u}) = [p_1\ p_2\ u_1\ p_3\ u_2\ u_3\ u_4]$. Iako ovo (formalno) nije sistematsko kodiranje, ono nije ništa manje primenljivo u praksi s obzirom da su informacioni bitovi i dalje sadržani u kodnoj reči.

Na isti način generišu se ostali kodovi iz klase Hammingovih kodova. Matrica $H_{m,h}$ ima m vrsta i $2^m - 1$ kolonu, pri čemu je i-ta kolona obrnuta binarna reprezentacija broja $i = 1, 2, \ldots, 2^m - 1$. Pošto je kontrolna matrica formata $(n - k) \times n$, sledi da je $n = 2^m - 1$ a $n - k = m$ pa je $k = 2^m - m - 1$. Dakle, u pitanju su $(2^m - 1, 2^m - m - 1)$ kodovi.

Očigledno su svake dve kolone kontrolne matrice $H_{m,h}$ linearno nezavisne (odnosno različite, u binarnom slučaju), dok je npr. zbir prve, pretposlednje i poslednje kolone jednak nula vektoru:

$$\begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \mathbf{0}^T.$$

Zaključujemo da je kodno rastojanje ovog koda 3 odnosno da kod ispravlja jednu grešku. Odgovarajuća generišuća matrica dobija se tako što se permutacijom kolona $H_{m,h}$ svede na sistematski oblik, onda odredi sistematska generišuća matrica, a nakon toga primenom obrnute permutacije dobije generišuća matrica $G_{m,h}$.

Sindrom dekodiranje se obavlja na isti način kao kod $(7, 4)$ koda, pošto je i ovde $\mathbf{s}^T = H_{m,h}\mathbf{e}^T$ jednako obrnutoj binarnoj reprezentaciji indeksa na kojem je došlo do greške.

Hammingovi kodovi imaju još jedno interesantno svojstvo. Posmatrajmo sfere $Z_1(\mathbf{x})$ poluprečnika $s = 1$ oko svake kodne reči. Ovo je skup svih reči $\mathbf{y}$ takvih da je $d_H(\mathbf{x}, \mathbf{y}) \leq 1$ odnosno da se $\mathbf{y}$ i $\mathbf{x}$ razlikuju za najviše jedan bit. Podsetimo se, da bi kod ispravljao $s = 1$ grešaka, sfere $Z_1(\mathbf{x})$ moraju biti disjunktne. Sfera $Z_1(\mathbf{x})$ sadrži ukupnop $1 + \binom{n}{1} = n + 1$ reči, a sve sfere zajedno sadrže tačno:

$$\sum_{\mathbf{x}\in\mathcal{C}} |Z_1(\mathbf{x})| = |\mathcal{C}| \cdot |Z_1(\mathbf{x})| = 2^k \cdot (n+1) = 2^{2^m-m-1} \cdot (2^m - 1 + 1)$$
$$= 2^{2^m-m-1+m} = 2^{2^m-1} = 2^n$$

reči. Prema tome, ove sfere pokrivaju ceo skup $\mathcal{X}^n = \{0, 1\}^n$ reči dužine n, odnosno Hammingov kod je **perfektan**.

I pored toga što poseduju svojstvo perfektnosti, Hammingovi kodovi nisu klasa kodova koja dostiže teorijske granice. Štaviše, kodovi se ponašaju suprotno od klase ponavljajućih kodova. Uzmimo kao primer binarni simetrični kanal sa verovatnoćom greške α. S ozbirom da kod ispravlja jednu (i samo jednu) grešku, verovatnoća greške je

$$P(\mathcal{E}) = 1 - (1-\alpha)^n - n\alpha(1-\alpha)^{n-1}.$$

Prema tome, nije teško utvrditi da $P(\mathcal{E}) \to 1$ kada $n \to +\infty$ (odnosno kada $m \to +\infty$, s obzirom da je $n = 2^m - 1$). Sa druge strane je $R = (2^m - m - 1)/(2^m - 1) \to 1$ kada $m \to +\infty$. Ovo je potpuno suprotno ponašanje od ponavljajućeg koda, kod koga su i verovatnoća greške i kodni količnik težili nuli, sa povećanjem dužine kodne reči.

Hammingovi kodovi su osnova za konstrukciju druge klase kodova, poznatijih kao BCH i Reed–Solomonovi kodovi, koji imaju mnogo bolje performanse.

8.4 Ciklični kodovi

Ciklični kodovi predstavljaju podklasu linearnih blok kodova, kod kojih je konstrukcija kodera i dekodera dodatno pojednostavljena. Ovi kodovi imaju široku primenu, a jedno vreme su se isključivo oni i koristili.

8.4.1 Definicija, polinomska reprezentacija i osnovna svojstva

Naziv **ciklični kodovi** potiče direktno iz sledeće definicije, odnosno iz dodatne pretpostavke da je kod zatvoren za operaciju cikličnog pomeranja reči.

Definicija 8.4.1. *Linearni (n,k) kod $\mathcal{C}$ je **cikličnI** kod ukoliko za svaku kodnu reč $\mathbf{c}=(c_0,c_1,\ldots,c_{n-1})$ važi da je reč $\mathbf{c}^R=(c_{n-1},c_0,c_1,\ldots,c_{n-2})\in\mathcal{C}$.*

Iako ima puno cikličkih kodova, u poređenju sa celokupnom klasom linearnih kodova, njihov broj je zanemarljiv.

Primer 8.4.1. Za svaki ceo broj $n\geq 3$, postoje sledeći ciklični kodovi, koje se nazivaju **trivijalni kodovi**:

1. Kod $(n,0)$ koji se sastoji samo od kodne reči $\underbrace{00\ldots0}_{n}$, koji se naziva **neinformativni kod**.
2. Kod $(n,1)$ koji se sastoji od kodnih reči $(b,b,\ldots,b)$ $(b\in\mathbb{F})$ i koji se naziva **ponavljajući kod**.
3. Kod $(n,n-1)$ koji se sastoji od kodnih reči $\mathbf{c}=(c_0,c_1,\ldots,c_{n-1})$ takvih da je $c_0+c_1+\ldots+c_{n-1}=0$, i koji se naziva **kod provere na parnost**.
4. Kod (n,n) koji se sastoji od svih reči dužine n.

Za neke vrednosti n i $\mathbb{F}$ (npr. $n=19$ i polje $GF(2)$) prethodno nabrojani kodovi su jedini ciklični kodovi koji postoje.

Primer 8.4.2. Posmatrajmo linearni $(7,3)$ kod definisan generišućom matricom:

$$G=\begin{bmatrix}1&0&1&1&1&0&0\\0&1&0&1&1&1&0\\0&0&1&0&1&1&1\end{bmatrix}.$$

Ukoliko sa $\mathbf{g}_1,\mathbf{g}_2$ i $\mathbf{g}_3$ označimo prve tri vrste matrice G, sledi da se kod $\mathcal{C}$ sastoji od sledećih kodnih reči:

$$\mathcal{C}=\{\mathbf{0},\mathbf{g}_1,\mathbf{g}_2,\mathbf{g}_3,\mathbf{g}_1+\mathbf{g}_2,\mathbf{g}_2+\mathbf{g}_3,\mathbf{g}_1+\mathbf{g}_3,\mathbf{g}_1+\mathbf{g}_2+\mathbf{g}_3\}$$

Direktno proveravamo da su ciklično pomerene reči $\mathbf{g}_1^R$, $\mathbf{g}_2^R$ i $\mathbf{g}_3^R$ takođe kodne reči, odnosno da je $\mathbf{g}_1^R=\mathbf{g}_2$, $\mathbf{g}_2^R=\mathbf{g}_3$, i $\mathbf{g}_3^R=\mathbf{g}_1+\mathbf{g}_3$. Iz linearnosti operacije cikličnog pomeranja $(\alpha\mathbf{x}^R+\beta\mathbf{y}^R=(\alpha\mathbf{x}+\beta\mathbf{y})^R)$ dobijamo da i za sve ostale kodne reči $\mathbf{c}\in\mathcal{C}$ važi da je ciklično pomerena reč $\mathbf{c}^R$ takođe kodna reč, odnosno:

$$(\mathbf{g}_1+\mathbf{g}_2)^R=\mathbf{g}_2+\mathbf{g}_3,\quad(\mathbf{g}_1+\mathbf{g}_3)^R=\mathbf{g}_1+\mathbf{g}_2+\mathbf{g}_3,$$
$$(\mathbf{g}_2+\mathbf{g}_3)^R=\mathbf{g}_1,\quad(\mathbf{g}_1+\mathbf{g}_2+\mathbf{g}_3)^R=\mathbf{g}_1+\mathbf{g}_2.$$

Dakle, u pitanju je ciklični kod.

Fundamentalne osobine cikličnih kodova dolaze do izražaja tek kada se uvede njihova **polinomna interpretacija**. To je zapravo bila i glavna motivacija da se ova klasa kodova definiše na način opisan u Definiciji 8.4.1. **Generišući polinom** kodne reči $\mathbf{c} = (c_0, c_1, \ldots, c_{n-1}) \in \mathcal{C}$ dat je sa

$$c(x) = c_0 + c_1 x + \ldots + c_{n-1} x^{n-1}.$$

Operacija $\mathbf{c}^R$ cikličkog pomeranja za jedno mesto udesno jednostavno se opisuje pomoću polinoma $c(x)$.

Pre nego što to dokažemo, podsetimo se definicije operacija deljenja i ostatka (div i mod) nad skupom (prstenom) polinoma $\mathbb{F}[x]$, sa koeficijentima iz skupa $\mathbb{F}$.

Lema 8.4.1. *Za svaka dva polinoma $P, M \in \mathbb{F}[x]$ postoje polinomi $Q, R \in \mathbb{F}[x]$ takvi da je*

$$P(x) = Q(x)M(x) + R(x)$$

gde je $\deg R(x) < \deg M(x)$.

Slično kao kod celih brojeva, polinome $Q(x)$ i $R(x)$ iz prethodne leme možemo označiti redom sa $P(x)$ div $M(x)$ i $P(x)$ mod $M(x)$. Ukoliko je $P(x)$ mod $M(x) = 0$ kažemo da polinom $M(x)$ deli polinom $P(x)$ i pišemo $M(x) \mid P(x)$.

Primer 8.4.3. Neka je $\mathbb{F} = \mathbb{R}$, $P(x) = 5x^3 - 1$ i $M(x) = x^2 + 1$. Tada je

$$P(x) = 5xM(x) - 5x - 1$$

pa je $P(x)$ div $M(x) = 5x$ a $P(x)$ mod $M(x) = -5x - 1$.

Primer 8.4.4. Neka je $\mathbb{F} = GF(2)$, $P(x) = x^3 - 1 = x^3 + 1$ i $M(x) = x - 1 = x + 1$. Tada je

$$P(x) = (x^2 + x + 1)M(x) = (x^2 + x + 1)(x + 1)$$

pa je $P(x)$ div $M(x) = x^2 + x + 1$ a $P(x)$ mod $M(x) = 0$.

Lema 8.4.2. *Sledeća svojstva polinomnih operacija važe za proizvoljne polinome $P, Q, M, N \in \mathbb{F}[x]$:*

1. *Ako je* $\deg P(x) < \deg M(x)$*, onda je $P(x)$ mod $M(x) = P(x)$.*
2. *Ako $M(x) \mid P(x)$ onda je $P(x)$ mod $M(x) = 0$*

3. $(P(x)+Q(x))\ mod\ M(x) = P(x)\ mod\ M(x) + Q(x)\ mod\ M(x)$.
4. $(P(x)Q(x))\ mod\ M(x) = (P(x)(Q(x)\ mod\ M(x)))\ mod\ M(x)$.
5. *Ako* $M(x) \mid N(x)$ *onda je* $(P(x)\ mod\ N(x))\ mod\ M(x) = P(x)\ mod\ M(x)$.

Sada možemo dati jednostavan opis operacije pomeranja udesno pomoću generišućeg polinoma.

Teorema 8.4.3. *Generišući polinom reči* $\mathbf{c}^R$ *jednak je*

$$c^R(x) = (xc(x))\ mod\ (x^n - 1).$$

Dokaz. Neka je $\mathbf{c} = (c_0, c_1, \ldots, c_{n-1})$ odnosno $c(x) = c_0 + c_1x + \ldots + c_{n-1}x^{n-1}$. Tada na osnovu Leme 8.4.2 sledi:

$$\begin{aligned}(xc(x)) &\bmod (x^n - 1)\\ &= (c_0x + c_1x^2 + \ldots + c_{n-2}x^{n-1} + c_{n-1}x^n) \bmod (x^n - 1)\\ &= c_0x + c_1x^2 + \ldots + c_{n-2}x^{n-1} + (c_{n-1}x^n) \bmod (x^n - 1)\\ &= c_0x + c_1x^2 + \ldots + c_{n-2}x^{n-1} + c_{n-1} = c^R(x)\end{aligned}$$

jer je $\mathbf{c}^R = (c_{n-1}, c_0, c_1, \ldots, c_{n-2})$. □

U nastavku ćemo većinom koristiti polinom $c(x)$ umesto kodne reči $\mathbf{c}$ (čak ćemo ova dva pojma i poistovećivati, kada to kontekst bude dozvolio). Na ovaj način, ciklični kod $\mathcal{C}$ može da se posmatra i kao skup polinoma

$$\mathcal{C}[x] = \{c(x) \mid \mathbf{c} \in \mathcal{C}\}$$

stepena manjeg ili jednakog $n-1$. Pritom, na osnovu prethodne teoreme važi:

$$c(x) \in \mathcal{C}[x] \quad \Rightarrow \quad (xc(x)) \bmod (x^n - 1) \in \mathcal{C}[x].$$

U nastavku ćemo često poistovećivati kodne reči C sa odgovarajućim polinomima $c(x)$, kao i kod $\mathcal{C}$ sa skupom polinoma $\mathcal{C}[x]$. Takođe ćemo umesto $P(x) \bmod (x^n - 1)$ pisati kraće $[P(x)]_n$.

Teorema 8.4.4. *Ako je* $c(x)$ *kodna reč cikličkog koda* $\mathcal{C}$*, onda je i* $[P(x)c(x)]_n$ *takođe kodna reč.*

Dokaz. Najpre dokazujemo matematičkom indukcijom da je $[x^ic(X)]_n$ kodna reč za svako $i \geq 1$. Za $i = 1$ tvrđenje se svodi na $[xc(x)]_n = xc(x) \bmod (x^n - 1) \in \mathcal{C}[x]$, što smo već dokazali. Pretpostavimo da tvrđenje važi za neko $i-1$ i

dokažimo da važi za i. Neka je $c_1(x) = [x^{i-1}c(x)]_n$ što je kodna reč, na osnovu indukcijske hipoteze. Dalje je:

$$\begin{aligned}\left[x^i c(x)\right]_n &= x^i c(x) \bmod (x^n-1) = x((x^{i-1}c(x)) \bmod (x^n-1)) \bmod (x^n-1)\\ &= (xc_1(x)) \bmod (x^n-1) \in \mathcal{C}[x]\end{aligned}$$

čime je dokaz završen. Napomenimo da je u prethodnom izrazu u poslednjem koraku korišćen identitet 4 iz Leme 8.4.2. □

Ključni pojam vezan za projektovanje i analizu cikličnih kodova je pojam **generišućeg polinoma** $g(x)$ koda $\mathcal{C}$.

Definicija 8.4.2. *Polinom $g(x) \in \mathcal{C}[x] \setminus \{0\}$ minimalnog stepena naziva se **generišući polinom** koda $\mathcal{C}$.*

Primer 8.4.5. Generišući polinom cikličnog $(7,3)$ koda definisanog u primeru 8.4.2, jednak je $g(x) = c_1(x) = 1 + x^2 + x^3 + x^4$.

Lema 8.4.5. *Neka je $\mathcal{C}$ ciklični kod sa generišućim polinomom $g(x)$.*

1. *Ako je $g_1(x)$ drugi generišući polinom, onda je $g_1(x) = \lambda g(x)$, za neko $\lambda \in \mathbb{F} \setminus \{0\}$.*

2. *Ako je $P(x)$ polinom takav da je $[P(x)]_n$ kodna reč, tada $g(x)$ deli $P(x)$.*

Dokaz. Pretpostavimo da $g_1(x)$ nije deljiv polinomom $g(x)$, odnosno da može da se zapiše u obliku $g_1(x) = \lambda g(x) + r(x)$, gde je količnik λ stepena $\deg g_1(x) - \deg g(x) = 0$ a ostatak $r(x) \neq 0$ stepena $\deg r(x) < \deg g(x)$. Pošto su $g_1(x), \lambda g(x) \in \mathcal{C}[x]$ sledi da je $r(x) = g_1(x) - \lambda g(x) \in \mathcal{C}[x]$. Ovo je nemoguće, s obzirom na pretpostavku da je $g(x)$ polinom minimalnog stepena u $\mathcal{C}[x]$. Prema tome, $g_1(x)$ je deljiv polinomom $g(x)$, odnosno važi $g_1(x) = \lambda g(x)$.

Neka je $P(x)$ proizvoljan polinom za koji je $[P(x)]_n$ kodna reč. Pretpostavimo da $P(x)$ nije deljivo sa $g(x)$ odnosno da je $P(x) = q(x)g(x) + r(x)$ gde je $r(x) \neq 0$ i $\deg r(x) < \deg g(x)$. Ako na obe strane jednakosti primenimo operaciju $\bmod (x^n - 1)$ i koristimo svojstva Leme 8.4.2 dobijamo:

$$[P(x)]_n = [q(x)g(x)]_n + [r(x)]_n$$

Pošto je $g(x)$ kodna reč, na osnovu Teoreme 8.4.4 sledi da je $[q(x)g(x)]_n$ kodna reč, pa je i

$$r(x) = [r(x)]_n = [P(x)]_n - [q(x)g(x)]_n \in \mathcal{C}[x].$$

Ovo je nemoguće, opet zato što je $g(x)$ polinom minimalnog stepena u $\mathcal{C}[x]$ i $\deg r(x) < \deg g(x)$. Prema tome, $P(x)$ je deljivo sa $g(x)$, odnosno postoji polinom $q(x)$ takav da je $P(x) = q(x)g(x)$. $\square$

Sada ćemo dokazati glavnu teoremu vezanu za ciklične kodove koja daje vezu između svih cikličnih kodova dužine n i delioca polinoma $x^n - 1$.

Teorema 8.4.6.

1. *Ako je $\mathcal{C}$ ciklični kod nad poljem $\mathbb{F}$, onda je njegov generišući polinom $g(x)$ delioc polinoma $x^n - 1$. Takođe, kodna reč $\mathbf{c}$ je u kodu $\mathcal{C}$ ako i samo ako je $c(x)$ deljivo sa $g(x)$. Ako sa k označimo dimenziju koda $\mathcal{C}$, onda je $k = n - \deg g(x)$.*

2. *Obratno, ako je $g(x)$ proizvoljan delioc polinoma $x^n - 1$ i $k = n - \deg g(x)$, onda postoji (n, k) ciklični kod čiji je generišući polinom jednak $g(x)$.*

Dokaz. S obzirom da je $[x^n - 1]_n = 0 \in \mathcal{C}[x]$, na osnovu prethodne leme zaključujemo da je polinom $x^n - 1$ deljiv polinomom $g(x)$. Takođe, za svaku kodnu reč $c(x)$ važi $[c(x)]_n = c(x) \in \mathcal{C}[x]$ pa opet, na osnovu prethodne leme sledi da je $c(x)$ deljiv polinomom $g(x)$. Obrnuto, ako je $c(x) = u(x)g(x)$ proizvoljni polinom stepena $\deg c(x) < n$, onda je na osnovu Teoreme 8.4.4 $c(x) = [c(x)]_n$ kodna reč. Prema tome, $c(x)$ je kodna reč ako i samo ako je $c(x) = u(x)g(x)$ za neki polinom $u(x)$ stepena manjeg od $k = n - \deg g(x)$ odnosno

$$\begin{aligned} c(x) = u(x)g(x) &= (u_0 + u_1 x + \ldots, u_{k-1}x^{k-1})g(x) \\ &= u_0 g(x) + u_1 x g(x) + \ldots, u_{k-1}x^{k-1}g(x). \end{aligned}$$

Iz linearne nezavisnosti polinoma $g(x), xg(x), \ldots, x^{k-1}g(x)$ (svi su različitog stepena) i prethodnog izraza, sledi da ovi polinomi formiraju bazu prostora $\mathcal{C}[x]$, odnosno da je $\dim \mathcal{C}[x] = k = n - \deg g(x)$.

Pretpostavimo sada da je $g(x)$ proizvoljni delilac polinoma $x^n - 1$ i neka je $\mathcal{C}[x]$ skup polinoma generisan ovim polinomom

$$\mathcal{C}[x] = \{c(x) \mid c(x) = u(x)g(x) = (u_0 + u_1 x + \ldots, u_{k-1}x^{k-1})g(x)\}.$$

Pritom je $k = n - \deg g(x)$. Nije teško uočiti da ako su $c_1(x) = u_1(x)g(x)$ i $c_2(x) = u_2(x)g(x)$ elementi $\mathcal{C}[x]$, da je onda i $\alpha c_1(x) + \beta c_2(x) = (\alpha u_1(x) + \beta u_2(x))g(x) \in \mathcal{C}[x]$. Znači da je $\mathcal{C}[x]$ zatvoren za linearnu kombinaciju, pa predstavlja skup polinoma linearnog koda $\mathcal{C}$. Preostaje još da dokažemo da je

ciklični pomeraj $c^R(x) = [xc(x)]_n$ takođe kodna reč. Polinom $xc(x)$ je najviše n-tog stepena, pa je količnik pri deljenju $xc(x)$ sa $x^n - 1$ konstanta:

$$xc(x) = xu(x)g(x) = \lambda(x^n - 1) + c^R(x).$$

Pošto $g(x)$ deli i levu stranu jednakosti i $x^n - 1$, zaključujemo da deli i ostatak $c^R(x)$ odnosno da $c^R(x) \in \mathcal{C}[x]$. Ovim smo dokazali da je $\mathcal{C}[x]$ ciklični kod dimenzije $k = n - \deg g(x)$. □

Prema tome, cikličnih kodova ima isto onoliko koliko ima i delioca $g(x)$ polinoma $x^n - 1$ u polju $\mathbb{F}$. Svaki kod je određen kao skup polinoma deljivih generišućim polinomom $g(x)$:

$$\mathcal{C}[x] = \{u(x)g(x) \mid \deg u(x) \leq k - 1\}$$

a odgovarajuća dimenzija koda je $k = n - \deg g(x)$.

Primer 8.4.6. Ako izvršimo faktorizaciju polinoma $x^n - 1$ za $n = 3, 4, \ldots, 20$ nad poljem $GF(2)$ dobijamo:

$$\begin{aligned}
x^3 - 1 &= (x+1)\left(x^2 + x + 1\right)\\
x^4 - 1 &= (x+1)^4\\
x^5 - 1 &= (x+1)\left(x^4 + x^3 + x^2 + x + 1\right)\\
x^6 - 1 &= (x+1)^2\left(x^2 + x + 1\right)^2\\
x^7 - 1 &= (x+1)\left(x^3 + x + 1\right)\left(x^3 + x^2 + 1\right)\\
x^8 - 1 &= (x+1)^8\\
x^9 - 1 &= (x+1)\left(x^2 + x + 1\right)\left(x^6 + x^3 + 1\right)\\
x^{10} - 1 &= (x+1)^2\left(x^4 + x^3 + x^2 + x + 1\right)^2\\
x^{11} - 1 &= (x+1)\left(x^{10} + x^9 + x^8 + x^7 + x^6 + x^5 + x^4 + x^3 + x^2 + x + 1\right)\\
x^{12} - 1 &= (x+1)^4\left(x^2 + x + 1\right)^4\\
x^{13} - 1 &= (x+1)\left(x^{12} + x^{11} + x^{10} + x^9 + x^8 + x^7 + x^6 + x^5 + x^4 + x^3 + x^2 + x + 1\right)\\
x^{14} - 1 &= (x+1)^2\left(x^3 + x + 1\right)^2\left(x^3 + x^2 + 1\right)^2\\
x^{15} - 1 &= (x+1)\left(x^2 + x + 1\right)\left(x^4 + x + 1\right)\left(x^4 + x^3 + 1\right)\left(x^4 + x^3 + x^2 + x + 1\right)\\
x^{16} - 1 &= (x+1)^{16}\\
x^{17} - 1 &= (x+1)\left(x^8 + x^5 + x^4 + x^3 + 1\right)\left(x^8 + x^7 + x^6 + x^4 + x^2 + x + 1\right)\\
x^{18} - 1 &= (x+1)^2\left(x^2 + x + 1\right)^2\left(x^6 + x^3 + 1\right)^2\\
x^{19} - 1 &= (x+1)\left(x^{18} + x^{17} + \ldots + x + 1\right)
\end{aligned}$$

$x^{20}-1=(x+1)^4\left(x^4+x^3+x^2+x+1\right)^4$

Vidimo da za dužine $n=5,11,13,19$, polinom x^n-1 ima samo 4 delioca: x^n-1, $x^{n-1}+x^{n-2}+\ldots+1$, $x+1$, 1. Ovi delioci redom odgovaraju trivijalnim kodovima iz Primera 8.4.1 (**neinformativni kod**, **ponavljajući kod**, **kod provere na parnost** i **kod koji se sastoji od svih reči**). Prema tome, za dato n postoje samo trivijalni ciklični kodovi ukoliko polinom $x^{n-1}+x^{n-2}+\ldots+1$ nema netrivijalne delioce (odnosno ako je **ireducibilan**). U suprotnom, ukoliko x^n-1 možemo da zapišemo kao proizvod $x^n-1=p_1(x)^{\alpha_1}p_2(x)^{\alpha_2}\cdots p_k(x)^{\alpha_k}$, gde su $p_1,p_2,\ldots,p_k$ ireducibilni polinomi, broj cikličnih kodova dužine n jednak je broju delioca x^n-1 odnosno $\tau(n)=(\alpha_1+1)(\alpha_2+1)\cdots(\alpha_k+1)$.

8.4.2 Generišuća i kontrolna matrica koda

Polinom $h(x)$ definisan sa

$$h(x)=\frac{x^n-1}{g(x)}$$

naziva se **kontrolni polinom**, ili **polinom provere na parnost**. Značaj polinoma $g(x)$ i $h(x)$ je i u tome što se na osnovu njih lako dobijaju generišuća i kontrolna matrica koda $\mathcal{C}$.

Posledica 8.4.7. *Ako je $\mathcal{C}$ ciklični (n,k) kod sa generišućim i kontrolnim polinomom $g(x)=g_0+g_1x+\ldots+g_rx^r$ i $h(x)=h_0+h_1x+\ldots+h_kx^k$ gde je $k=n-r$, onda su generišuća i kontrolna matrica G_1 i H_1 date sa:*

$$G_1=\begin{bmatrix} g_0 & g_1 & \cdots & \cdots & g_r & 0 & \cdots & \cdots & 0\\ 0 & g_0 & g_1 & \cdots & \cdots & g_r & 0 & \cdots & 0\\ \vdots & & & & & & & & \\ 0 & \cdots & \cdots & 0 & g_0 & g_1 & \cdots & \cdots & g_r \end{bmatrix}=\begin{bmatrix} g(x)\\ xg(x)\\ \vdots\\ x^{k-1}g(x)\end{bmatrix},$$

$$H_1=\begin{bmatrix} h_k & h_{k-1} & \cdots & \cdots & h_0 & 0 & \cdots & \cdots & 0\\ 0 & h_k & h_{k-1} & \cdots & \cdots & h_0 & 0 & \cdots & 0\\ \vdots & & & & & & & & \\ 0 & \cdots & \cdots & 0 & h_k & h_{k-1} & \cdots & \cdots & h_0 \end{bmatrix}=\begin{bmatrix} \tilde{h}(x)\\ x\tilde{h}(x)\\ \vdots\\ x^{r-1}\tilde{h}(x)\end{bmatrix},$$

Svaka vrsta matrica G_1 i H_1 sastoji se od koeficijenata odgovarajućeg polinoma sa desne strane jednakosti. Pritom je

$$\tilde{h}(x)=x^kh(1/x)=h_k+h_{k-1}x+\ldots+h_0x^k$$

recipročni polinom polinoma $h(x)$.

Dokaz. [McEliece, p. 176] Svaka kodna reč $c(x) \in \mathcal{C}[x]$ može da se predstavi kao

$$c(x) = u(x)g(x) = u_0 + u_1 x g(x) + \ldots + u_{k-1}x^{k-1}g(x)$$

gde su $u_0, u_1, \ldots, u_{k-1}$ koeficijenti polinoma $u(x)$ i $k = n - \deg g(x)$. Prethodni izraz možemo zapisati u vektorskom obliku

$$\mathbf{c} = u_0\mathbf{g}_{1,0} + u_1\mathbf{g}_{1,1} + \ldots + u_{k-1}\mathbf{g}_{1,k-1}.$$

gde je $\mathbf{g}_{1,i}$ vektor koeficijenata polinoma $x^i g(x)$ za $i = 0, 1, \ldots, k-1$. Vidimo da su ovi vektori zapravo vrste matrice G_1. Oni su linearno nezavisni, s obzirom da su odgovarajući polinomi $x^i g(x)$ različitog stepena pa su samim tim linearno nezavisni. Prema tome, matrica G_1 je generišuća matrica koda $\mathcal{C}$.

Pokazaćemo da je $H_1\mathbf{g}_{1,j}^T = \mathbf{0}^T$, odnosno da je $\mathbf{h}_{1,i}\mathbf{g}_{1,j}^T = 0$ za svako i, j, gde $\mathbf{h}_{1,i}$ predstavlja i-tu vrstu matrice H_1. S obzirom da su vrste matrica G_1 i H_1 ciklično pomerene prve vrste odgovarajućih matrica, njihov (skalarni) proizvod jednak je

$$\mathbf{h}_{1,i}\mathbf{g}_{1,j}^T = \sum_{l=0}^{k+i-j} g_l h_{k+i-j-l}$$

pri čemu je $g_l = 0$ za $l > n-k$ i $h_l = 0$ za $l > k$. Prethodni izraz je koeficijent uz x^{k+i-j} u proizvodu $g(x)h(x) = x^n - 1$. Pošto je $i = 1, 2, \ldots, n-k$ a $j = 1, 2, \ldots, k$ sledi da je $1 \leq k+i-j \leq n-1$. Koeficijenti uz $x^1, x^2, \ldots, x^{n-1}$ u $g(x)h(x) = x^n - 1$ jednaki su nuli, pa je i $\mathbf{h}_{1,i}\mathbf{g}_{1,j}^T$ za svako i, j. □

Podsetimo se da se operacija kodiranja u slučaju linearnog (n, k) koda svodi na računanje proizvoda $\mathbf{c} = \mathbf{u}G$, gde je G generišuća matrica, a $\mathbf{u}$ reprezentacija indeksa W u bazi b. Ukoliko usvojimo generišuću matricu G_1, u terminima polinoma $c(x)$ ova operacija se svodi na prosto množenje polinoma

$$c(x) = u_0 g(x) + u_1 x g(x) + \ldots + u_k x^{k-1} g(x) = u(x)g(x).$$

Iako se matrice G_1 i H_1 koriste u praksi, mnogo češće se koriste matrice G_2 i H_2 definisane u sledećoj posledici.

Posledica 8.4.8. *Neka je $\mathcal{C}$ ciklični (n, k) kod sa generišućim polinomom $g(x)$ i kontrolnim polinomom $h(x)$ i neka je $g_{2,i}(x) = x^{r+i} - x^{r+i} \bmod g(x)$ za $i = 0, 1, \ldots, k-1$. Onda je $k \times n$ matrica*

$$G_2 = \begin{bmatrix} \mathbf{g}_{2,0} \\ \mathbf{g}_{2,1} \\ \vdots \\ \mathbf{g}_{2,k-1} \end{bmatrix} = [A \; I_k]$$

generišuća za kod $\mathcal{C}$. Slično, ako je $h_{2,j}(x) = x^j \bmod g(x)$ za $j = 0, 1, \ldots, n-1$, onda je $r \times n$ matrica [5]

$$H_2 = \begin{bmatrix} \mathbf{h}_{2,0}^T & \mathbf{h}_{2,1}^T & \cdots & \mathbf{h}_{2,n-1}^T \end{bmatrix} = [I_{n-k} \ -A^T]$$

kontrolna za kod $\mathcal{C}$.

Dokaz. [McEliece, p. 177] Polinomi $g_{2,i}(x)$ $(i = 0, 1, \ldots, k-1)$ su različitih stepena, pa su očigledno linearno nezavisni. Takođe, direktno iz definicije zaključujemo da su deljivi sa $g(x)$, što znači da pripadaju kodu $\mathcal{C}[x]$. Pošto je dimenzija koda $k = \dim\mathcal{C}$, zaključujemo da ovi polinomi formiraju bazu, pa i odgovarajući vektori koeficijenata formiraju generišuću matricu.

Primetimo da su koeficijenti uz $x^{n-1}, x^{n-2}, \ldots, x^r$ jednaki za polinome x^{r+i} i $x^{r+i} - x^{r+i} \bmod g(x)$. Prema tome, poslednjih $k = n - r$ kolona matrice G_2 obrazuju jediničnu matricu, odnosno matrica G_2 ima oblik $G_2 = [A \ I_k]$.

Sa druge strane, koeficijenti polinoma $h_{2,r+i}(x) = x^{r+i} \bmod g(x)$ za $i = 0, 1, \ldots, k-1$ jednaki su negativnim vrednostima odgovarajućih koeficijenata polinoma $g_{2,i}(x) = x^{r+i} - x^{r+i} \bmod g(x) = x^{r+i} - h_{2,r+i}(x)$. Osim toga je $h_{2,i}(x) = x^i \bmod g(x) = x^i$ za $i = 0, 1, \ldots, r-1$. Prema tome, vektori kolone $\mathbf{h}_{2,i}$ za $i < r$ predstavljaju kolone jedinične matrice, a za $i \geq r$ negativne delove (od prvih r komponenata) vektora $g_{2,i-r}$. Time smo dokazali da matrica H_2 ima oblik $H_2 = [I_{n-k} \ -A^T]$, odnosno da predstavlja kontrolnu matricu koda. □

Sistematske matrice G_3 i H_3 dobijamo pomeranjem vrsta matrica G_2 i H_2 za k mesta ulevo:

$$G_3 = [I_k \ A], \quad H_3 = [-A^T \ I_r].$$

Ukoliko usvojimo postupak kodiranja korišćenjem matrice G_2, odnosno $\mathbf{c} = \mathbf{u}G_2$, onda je

$$c(x) = x^r u(x) - x^r u(x) \bmod g(x).$$

Sa druge strane, ako je primljena reč $\mathbf{y} = [y_0 \ y_1 \ \cdots \ y_{n-1}]$, sindrom $\mathbf{s}$ računamo kao $\mathbf{s}^T = H_2\mathbf{y}^T$ odnosno

$$\mathbf{s} = y_0\mathbf{h}_{2,0} + y_1\mathbf{h}_{2,1} + \ldots + y_{n-1}\mathbf{h}_{2,n-1}$$

U terminima polinoma, prethodna jednakost postaje

$$\begin{aligned} s(x) &= y_0x^0 \bmod g(x) + y_1x^1 \bmod g(x) + \ldots + y_{n-1}x^{n-1} \bmod g(x) \\ &= y(x) \bmod g(x). \end{aligned}$$

[5] Ovde $\mathbf{h}_{2,i}^T$ predstavlja **kolonu** matrice H_2.

Primer 8.4.7. Posmatrajmo ponovo kod iz Primera 8.4.2. Videli smo u prethodnom primeru da je generišući polinom za ovaj kod $g(x) = x^4 + x^3 + x^2 + 1$. Odgovarajući kontrolni polinom je $h(x) = (x^7 - 1)/g(x) = x^3 + x^2 + 1$. Kodne reči (tj. odgovarajući polinomi) se lako dobijaju kao umnožci polinoma $g(x)$. Nesistematske matrice G_1 i H_1 jednake su:

$$G_1 = \begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{bmatrix}$$

$$H_1 = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \end{bmatrix}$$

dok su matrice G_2 i H_2 jednake:

$$G_2 = \begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}$$

$$H_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \end{bmatrix}$$

Sistematske matrice G_3 i H_3 dobijamo primenom odgovarajuće permutacije kolona.

Postupak dekodiranja pomoću sindroma i korektora može da se iskoristi i za ciklične kodove. Međutim, kod cikličnih kodova nije potrebno pamtiti ceo vektor korektora već samo poslednju cifru. O tome govori sledeća lema:

Lema 8.4.9. ***(Meggitt dekoder)*** *Ako je* $s(x) = y(x)$ *mod* $g(x)$ *i* $y^R(x) = (xy(x))$ *mod* $(x^n - 1)$ *ciklički pomerena kodna reč, onda je* $s^{(1)}(x) = y^R(x)$ *mod* $g(x) = (xs(x))$ *mod* $g(x)$.

Dokaz. Znamo da je $y^R(x) = xy(x) - y_{n-1}(x^n - 1)$. Iz $s(x) = y(x) \bmod g(x)$, sledi da je $y(x) = q(x)g(x) + r(x)$ za neki polinom $q(x)$. Zamenom dobijamo:

$$\begin{aligned} s^{(1)}(x) = y^R(x) \bmod g(x) &= (xq(x)g(x) + xs(x) - y_{n-1}(x^n - 1)) \bmod g(x) \\ &= xs(x) \bmod g(x), \end{aligned}$$

čime je dokaz leme završen. □

Dakle, ukoliko $y(x)$ odgovara sindromu $s(x)$, onda $y^R(x)$ odgovara sindromu $xs(x) \bmod g(x)$. Važi i obrnuto, ukoliko $y_1(x)$ i $y_2(x)$ odgovaraju različitim sindromima $s_1(x)$ i $s_2(x)$, onda je i $s_1^{(1)}(x) \neq s_2^{(1)}(x)$. Dokaz je jednostavan, s obzirom da je $xs_1(x) \bmod g(x) = xs_2(x) \bmod g(x)$ ekvivalentno sa $g(x) \mid x(s_1(x) - s_2(x))$, a pošto x ne deli $g(x)$ (onda bi delilo $x^n - 1$), sledi da $g(x) \mid s_1(x) - s_2(x)$ odnosno da je $s_1(x) = s_2(x)$ jer su oba polinoma stepena manjeg od r.

Odavde možemo zaključiti da ako je $e(x)$ korektor minimalne težine za $s(x)$, onda je $e^R(x)$ korektor minimalne težine za $s^{(1)}(x)$. Ako je

$$e(x) = e_0 + e_1x + \ldots + e_{n-1}x^{n-1} \quad \Rightarrow \quad e^R(x) = e_{n-1} + e_0x + \ldots + e_{n-2}x^{n-1}$$

tada je dovoljno pamtiti samo e_0 umesto celog polinoma (vektora koeficijenata) $e(x)$. Zaista, e_{n-1} dobijamo kao poslednju cifru korektora ($e^R(x)$) za $s^{(1)}(x) = xs(x) \bmod g(x)$, e_{n-2} za $s^{(2)}(x) = xs^{(1)}(x) \bmod g(x)$, itd.

Ako sa $\mathbf{e}_{min,0}(s(x))$ označimo poslednji simbol korektora minimalne (Hammingove) težine za sindrom $\mathbf{s}$ kome odgovara polinom $s(x)$, algoritam za dekodiranje možemo opisati na sledeći način:

Algoritam 8.4.1. *(Sindrom dekodiranje cikličnih kodova)*

1. $s(x) := y(x)$ *mod* $g(x)$, $e_0 := \mathbf{e}_{min,0}(s(x))$, $s^{(1)}(x) := xs(x)$ *mod* $g(x)$.
2. *Za* $k = 1, 2, \ldots, n-1$ *raditi sledeće:*
   (a) $e_{n-k} := \mathbf{e}_{min,0}(s^{(k)}(x))$
   (b) $s^{(k+1)}(x) = xs^{(k)}(x)$ *mod* $g(x)$
3. $e(x) := e_0 + e_1x + \ldots + e_{n-1}x^{n-1}$

Primer 8.4.8. Prethodni algoritam ilustrovaćemo na jednostavnom primeru cikličnog $(7, 4)$ koda sa generišućim polinomom $g(x) = x^3 + x + 1$. Nije teško izračunati sledeću tablicu sindroma i korektora:

$e(x)$	$s(x) = e(x) \bmod g(x)$	$\mathbf{e}_{min,0}(s(x)) = e_0$
1	1	1
x	x	0
x^2	x^2	0
x^3	$1 + x$	0
x^4	$x + x^2$	0
x^5	$1 + x + x^2$	0
x^6	$1 + x^2$	0

Ukoliko je početni sindrom $s(x) = x^2$, dekodiranje ide na sledeći način: $e_0 = 0$ (iz tablice),

$$s^{(1)}(x) = xs(x) \bmod g(x) = x^3 \bmod (x^3 + x + 1) = x + 1$$

pa je $e_6 = \mathbf{e}_{min,0}(s^{(1)}(x)) = 0$ (iz tablice, treći red). Dalje je:

$$s^{(2)}(x) = xs^{(1)}(x) \bmod g(x) = (x^2 + x) \bmod (x^3 + x + 1) = x^2 + x$$

pa je $e_5 = \mathbf{e}_{min,0}(s^{(2)}(x)) = 0$ (iz tablice, četvrti red). Nastavljajući dalje postupak dobijamo $s^{(3)}(x) = x^2 + x + 1$ odakle je $e_4 = 0$, $s^{(4)}(x) = x^2 + 1$ odakle je $e_3 = 0$, $s^{(5)}(x) = 1$ odakle je $e_2 = 1$, $s^{(6)}(x) = x$ odakle je $e_1 = 0$. Dakle,

$$e(x) = 0 + 0 \cdot x + 1 \cdot x^2 + \ldots + 0 \cdot x^6 = x^2.$$

Prema tome, dobili smo ispravnu vrednost korektora $e(x) = x^2$. Pritom, za potrebe dekodera bilo je dovoljno memorisati samo treću kolonu tablice (indeksiranu drugom), što je ukupno 7 bita, umesto $7 \cdot 7 = 49$ koliko bi bilo potrebno ukoliko bismo pamtili cele korektore (prva kolona).

8.4.3 Specijalni ciklični kodovi

Ciklični kodovi imaju široku primenu u telekomunikacijama, naročito u računarskim komunikacijama. Jedna od najšire primenljivih klasa cikličnih kodova za detektovanje grešaka je klasa **CRC** (Cyclic Redundancy Check) kodova.

Primer 8.4.9. U mrežama koje se baziraju na standardu *Ethernet* (standardi IEEE 802.3 i *Ethernet* II) koristi se **CRC kod** sa oznakom IEEE 802.3 CRC-32 (skraćeno CRC-32) koji ima generišući polinom:

$$g(x) = x^{32}+x^{26}+x^{23}+x^{22}+x^{16}+x^{12}+x^{11}+x^{10}+x^8+x^7+x^5+x^4+x^2+x+1.$$

Pošto je stepen polinoma $g(x)$ jednak 32, dužina zaštitnog dela kodne reči je 32 bita, što tačno staje u 4 okteta okvira predviđenog standardom IEEE 802.3. Ovaj kod se primenjuje i u velikom broju drugih standarda, uključujući: HDLC, ANSI X3.66, ITU-T V.42, Ethernet, Serial ATA, MPEG-2, PKZIP, Gzip, Bzip2, PNG.

U upotrebi je još jedan CRC kod, poznatiji kao **CRC–32C** (Castagnoli),

čiji je generišući polinom jednak:

$$g(x) = x^{32} + x^{28} + x^{27} + x^{26} + x^{25} + x^{23} + x^{22} + x^{20} + x^{19} + x^{18} + x^{14} + x^{13} + x^{11} + x^{10} + x^9 + x^8 + x^6 + 1$$

Ovaj kod ima bolje performanse od CRC–32 koda i našao je primenu u sledećim standardima: iSCSI & SCTP, G.hn payload, SSE4.2.

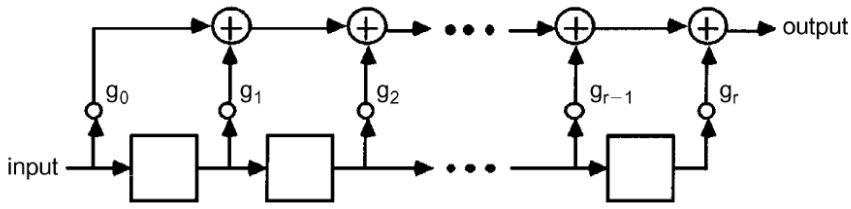
CRC kodovi se označavaju na sledeći način: **CRC**–n–**XXX** gde je n stepen generišućeg polinoma a XXX naziv koda. Ovi kodovi se dele u dve grupe. Jedna grupa kodova (među kojima je i CRC–32) može sa sigurnošću da detektuje jednu ili dve greške u bloku dužine $2^n - 1$. Primera radi, za CRC–32 blok je dužine $2^{32} - 1 \approx 4$GB. Naravno, veći broj grešaka takođe može da se detektuje, što se u praksi i dešava. Generišući polinom druge grupe CRC kodova ima oblik $(1+x)g(x)$ gde je $g(x)$ polinom stepena $n-1$. Ovi polinomi takođe detektuju 1-2 greške, ali ovog puta na približno upola manjem bloku (dužine $2^{n-1} - 1$). Pored toga, detektuju i sve greške sa neparnim brojem pogrešnih bitova.

Kodovi dati u sledećem primeru su važni zbog toga što su **perfektni** ciklični kodovi, tj. sfere poluprečnika $(d(\mathcal{C}) - 1)/2$ u potpunosti prekrivaju skup reči dužine n.

Primer 8.4.10.

Golay $(23, 12)$ **kod:** Perfektni kod (pokazati za vežbu) sa $d = 7$ + ciklični sa gen. polinomom: $g(x) = x^{11} + x^9 + x^7 + x^6 + x^5 + x + 1$. Ispravlja 3 greške!

Golay $(24, 12)$ **kod:** Golay $(23, 12)$ + parity check bit. $d = 8$. Ispravlja 3, detektuje 4 greške i neke kombinacije sa više grešaka.



Slika 8.1: Implementacija kodera cikličkog koda, na osnovu generišućeg polinoma, korišćenjem sabirača i flip-flopova.

8.4.4 Hardverska realizacija

Koderi za ciklične kodove jednostavno se implementiraju pomoću binarnih sabirača i flip-flopova. Blok šema jedne ovakve realizacije prikazana je na slici 8.1.

Na početku se svi flip-flopovi inicijalizuju na 0. Nakon toga se dovede niz

$$u_0, u_1, \ldots, u_{k-1}, \underbrace{0, 0 \ldots, 0}_{r}$$

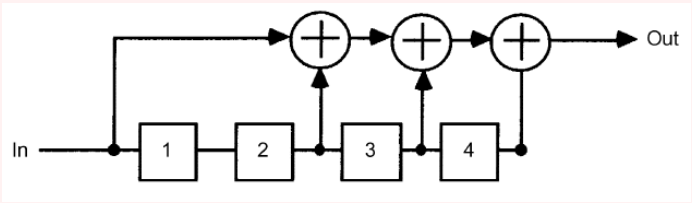
Nije teško pokazati da se na izlazu iz kodera upravo dobija niz

$$\begin{aligned} c_0 &= u_0 g_0 \\ c_1 &= u_0 g_1 + u_1 g_0 \\ c_2 &= u_0 g_2 + u_1 g_1 + u_2 g_0 \\ &\vdots \\ c_j &= u_0 g_j + u_1 g_{j-1} + \ldots + u_j g_0 \\ &\vdots \\ c_{n-1} &= u_{k-1} g_r \end{aligned}$$

koji predstavlja niz koeficijenta polinoma $c(x) = u(x)g(x)$, odnosno kodnu reč $\mathbf{c}$. Slika 8.1 odgovara opštem slučaju, kada kod posmatramo nad proizvoljnim (konačnim) poljem $\mathbb{F}$. Međutim, u opštem slučaju, praktična realizacija odgovarajućih komponenti nije jednostavna. Ako je $\mathbb{F} = GF(2)$, možemo izvršiti sledeću zamenu:

$$\begin{aligned} \text{flip-flop} &\rightarrow D \text{ flip-flop} \\ \text{sabirač} &\rightarrow \text{XOR kolo} \\ \text{množač nulom} &\rightarrow \text{prekid veze} \\ \text{množač jedinicom} &\rightarrow \text{kratak spoj} \end{aligned}$$

Primer 8.4.11. Pošto je generišući polinom za kod iz primera 8.4.2 jednak $g(x) = x^4 + x^3 + x^2 + 1$, odgovarajuća šema kodera prikazana je na slici 8.2.

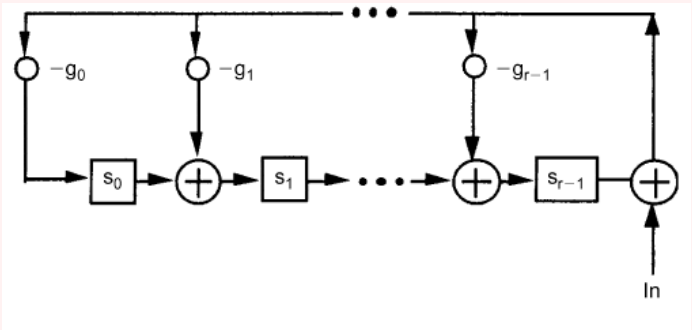


Slika 8.2: Šema kodera za $(7, 4)$ ciklični kod čiji je generišući polinom jednak $g(x) = x^4 + x^3 + x^2 + 1$.

Iako su koderi prikazani na slikama 8.1 i 8.2 najjednostavniji mogući, oni nisu sistematski, tj. informacioni simboli $u_0, u_1, \ldots, u_{k-1}$ se ne pojavljuju nepromenjeni u kodnim rečima. Ipak, moguće je konstruisati sistematski koder na malo komplikovaniji način. Podsetimo se da je operacija kodiranja kod sistematskog kodera svodi na

$$c(x) = x^r u(x) - (x^r u(x)) \bmod g(x).$$

Dakle, najpre je potrebno realizovati " mod $g(x)$" operaciju. To je moguće uraditi pomoću kola prikazanog na slici 8.3.



Slika 8.3: Kolo za računanje operacije " mod $g(x)$".

Lema 8.4.10. *Ako je* $s(x) = s_0 + s_1 x + \ldots + s_{r-1}x^{r-1}$ *polinom stanja u trenutnom a* $s'(x)$ *u sledećem vremenskom trenutku, onda je*

$$s'(x) = (xs(x) + s) \bmod g(x)$$

gde je $s \in \mathbb{F}$ *ulazni simbol.*

Dokaz. [MCEliece, p. 187–188] □

Lema 8.4.11. *Ukoliko stanja flip-flopova inicijalizujemo na 0, za unete koeficijente $a_0, a_1, \ldots$ dobijamo da je polinom stanja nakon t koraka jednak*

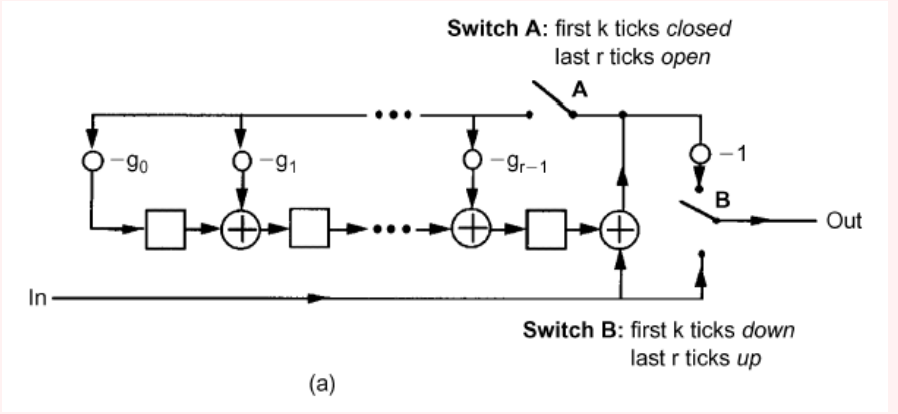
$$s_t(x) = \sum_{j=0}^{t} a_j x^{t-j} \ mod\ g(x).$$

Dokaz. [MCEliece, p. 188–189] □

Prema tome, ako na ulaz dovedemo niz

$$u_{r-1}, u_{r-2}, \ldots, u_0, \underbrace{0, 0, \ldots, 0}_{r}$$

posle k-tog koraka na izlazu (gornji desni ugao) dobijamo redom koeficijente polinoma $(x^r u(x))$ mod $g(x)$. Kompletno kolo za računanje kodne reči (tj. polinoma $c(x)$) prikazano je na slici 8.4.



Slika 8.4: Sistematski koder cikličnog koda.

Prekidač A je zatvoren u prvih k intervala, dok je u narednih r otvoren. Takođe, prekidač B je u prvih k intervala u donjem položaju, dok je u narednih r u gornjem položaju. Ukoliko je niz ulaznih simbola:

$$u_{r-1}, u_{r-2}, \ldots, u_0, \underbrace{0, 0, \ldots, 0}_{r},$$

onda na izlazu dobijamo vektor kodne reči $c_{n-1}, c_{n-2}, \ldots, c_0$ u obrnutom redosledu.

Pošto se sindrom računa kao $s(x) = y(x)$ mod $g(x)$, slično kolo može da se primeni i za konstrukciju dekodera.

8.5 BCH i Reed–Solomonovi kodovi

8.5.1 BCH kodovi

BCH kodovi predstavljaju generalizaciju Hammingovih kodova. To su binarni kodovi koji ispravljaju t grešaka. Iako su ovo u osnovi binarni kodovi, za konstrukciju i analizu (kao i realizaciju dekodera) koristićemo elemente polja $GF(2^m)$. Podsetimo se da elemente polja $GF(2^m)$ možemo sagledati kao m-torke binarnih brojeva nad kojima je definisano pokoordinatno (XOR) sabiranje, kao i operacija množenja u čije detalje realizacije nećemo ulaziti.

Neka su $\alpha_0, \alpha_1, \ldots, \alpha_{n-1} \in GF(2^m)$ različiti elementi. Binarni kod sa kontrolnom matricom (gde se svaki α_i predstavlja kao vektor-kolona dužine m):

$$H = \begin{bmatrix} \alpha_0 & \alpha_1 & \cdots & \alpha_{n-1} \\ \alpha_0^3 & \alpha_1^3 & \cdots & \alpha_{n-1}^3 \\ \vdots & & & \\ \alpha_0^{2t-1} & \alpha_1^{2t-1} & \cdots & \alpha_{n-1}^{2t-1} \end{bmatrix}$$

dimenzije $k = n - mt$ naziva se **BCH kod**. Važi sledeća teorema:

Teorema 8.5.1. *BCH kod ispravlja ispravlja t grešaka.*

Ukoliko odaberemo da $n \mid 2^m - 1$ kao i $\alpha_i = \alpha^i$, $i = 0, 1, \ldots, n-1$ (i pritom da su svi ovi elementi različiti), kod postaje **ciklični** gde je $g(x)$ **minimalni polinom nad** $GF(2)$ takav da je $g(\alpha^{2i+1}) = 0$ za svako $i = 0, 1, \ldots, t-1$.

Primer 8.5.1. Za $t = 3$, $n = 15$ i $m = 4$, kod sa polinomom:

$$g(x) = x^{10} + x^8 + x^5 + x^4 + x^2 + x + 1$$

je BCH kod dimenzije $n - mt = 15 - 12 = 3$. Napomenimo da, pošto zahtevamo da je polinom $g(x)$ binarni (tj. nad $GF(2)$), onda ne možemo garantovati da je on stepena $2t$ (pošto zahtevamo da ima $2t$) nula. Upravo je to ovde slučaj, tj. $\deg g(x) = 10 > 6 = 2t$.

Prednost ove klase kodova nad do sada proučenim kodovima je činjenica da ovde postoji **efektivna procedura** za dekodiranje koja ne zahteva nikakvo dodatno memorisanje. Procedura je bazirana na Euklidovom algoritmu i aritmetici (polinoma) nad $GF(2^m)$.

8.5.2 Reed–Solomonovi kodovi

Za konstrukciju Reed–Solomonovih (RS) kodova koristi se slična ideja kao BCH kodova, sa tom razlikom da se polje $GF(2^m)$ sada koristi i za kodiranje i za dekodiranje. Drugim rečima, ovo su kodovi nad $GF(2^m)$.

Reed–Solomonov (RS) kod je ciklični $(n, n-r)$-kod nad $GF(2^m)$ sa generišućim polinomom $g(x) = \prod_{j=1}^{r}(x - \alpha^j)$. Pritom je $\alpha \in GF(2^m)$ takav da su $\alpha, \alpha^2, \ldots, \alpha^r$ različiti.

Ovi kodovi imaju osobinu da im je kodno rastojanje $(\mathcal{C})$ maksimalno od svih kodova dimenzije $(n, n-r)$.

Teorema 8.5.2. ***(The Singleton bound)*** *Za svaki (n,k)-kod $\mathcal{C}$ nad $\mathbb{F}$ važi $d(\mathcal{C}) \leq n - k + 1$. Ova granica se dostiže za RS kodove, tj. važi $d(\mathcal{C}) = n - (n-r) + 1 = r + 1$.*

Kodovi koji dostižu granicu datu u prethodnoj teoremi nazivaju se **MDS (Maximum-distance separable) kodovi**.

Realizacija kodera je ovde ista kao i za sve ciklične kodove. Drugim rečima, mogu da se primene ista kola data u pododeljku 8.4.4, pri čemu sada svi elementi realizuju operacije nad poljem $GF(2^m)$. Postoje i brže realizacije kodera bazirane na FFT (Fast Fourier Transform) metodu.

Postupak dekodiranja se i ovde može efikasno realizovati. Kao i kod BCH kodova, procedura je bazirana na Euklidovom algoritmu.

Primer 8.5.2. Jedan od najčešće korišćenih RS kodova je $RS(255, 223)$. Ovaj kod se dobija za $m = 3$ $(GF(8))$ i $r = n - k = 32$. U sistematskoj realizaciji, kodna reč ovog koda sadrži 233 informativnih i 32 kontrolna simbola. Svaki simbol se sastoji od 3 bita. Pošto je kodno rastojanje (Teorema 8.5.2) jednako $d(\mathcal{C}) = n-k+1 = 33$, ovaj kod ispravlja $t = (n-k)/2 = 16$ pogrešnih simbola a može da detektuje $d(\mathcal{C}) - 1 = 32$ pogrešna simbola.

RS kodovi, kao kodovi nad $GF(2^m)$, pogodni su za ispravljanje grupisanih binarnih grešaka (**burst–error correcting**).

Primer 8.5.3. Posmatrajmo kod $RS(7,3)$ nad $GF(8)$. Neka je α element ovog polja takav da je $GF(8) = \{1, \alpha, \ldots, \alpha^7\}$ (takav element postoji). Jedna kodna reč ovog koda data je sa

$$\mathbf{c} = (\alpha^3, \alpha, \alpha, 1, 0, \alpha^3, 1) \in RS(7,3,GF(8)) \quad \leftrightarrow \quad \mathbf{c} = [011\ 010\ 010\ 001\ 000\ 011\ 011]$$

Pretpostavimo da se u kanalu pojavila greška, ali takva da su pogrešni bitovi grupisani na sledeći način:

$$\mathbf{e} = [000\ 000\ 0\overbrace{\mathbf{11\ 101}}^{\text{burst}}\ 000\ 000\ 000].$$

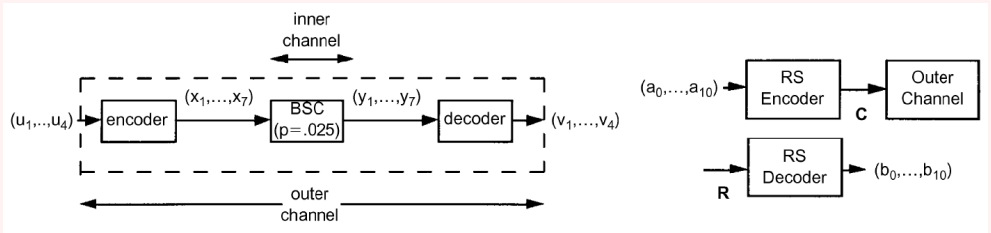
Greške ovog tipa nazivaju se **spojene greške** (**burst–error**). Primljeni vektor jednak je:

$$\mathbf{y} = (\alpha^3, \alpha, \mathbf{1}, \alpha^{\mathbf{2}}, 0, \alpha^3, 1) \in GF(8) \quad \leftrightarrow \quad R = [011\ 010\ 0\mathbf{01}\ 1\mathbf{00}\ 000\ 011\ 011]$$

Iako imamo 4 greške u binarnoj varijanti, one su grupisane i formiraju samo 2 greške u $GF(8)$ reprezentaciji (greške su boldirane), što ovaj kod može da ispravi.

Svojstvo da efikasno ispravljaju spojene greške, omogućilo je ovim kodovima primenu za skladištenje podataka na različitim medijumima, uključujući CD, DVD, Blu-Ray, HDD, SSD, itd. Čak su i moduli u svemirskoj letilici Voyager opremljeni ovim kodom.

Druga velika primena RS kodova je za **nadovezano kodiranje** (**concatenated coding**). Ideja je da se nekoliko kodnih reči koje daje kod čija je dužina kodne reči manja, dodatno zaštiti kodom veće dužine kodne reči. Na slici 8.5 dat je primer jednog takvog kodiranja. Pre nego što se primeni Hammingov $(7,4)$ kod (ili neki drugi kraći kod), najpre se veći blokovi reči zaštite RS kodom. Ukoliko dođe do grešaka u kanalu i pogrešnog dekodiranja kraćeg koda, te greške javiće se kao vezane greše (cele kodne reči biće pogrešne) za RS kod, koje će isti otkloniti.



Slika 8.5: Nadovezano kodiranje.

Ovakve kombinacije daju rezultate **blizu Shannonove granice**. Kao unutrašnji kod najčešće se koriste konvolucioni kodovi o kojima će kasnije biti reči.

Nadovezane greške moguće je "razbiti" pomoću **interlivera** (**interleaver**), koje ćemo kasnije opisati.

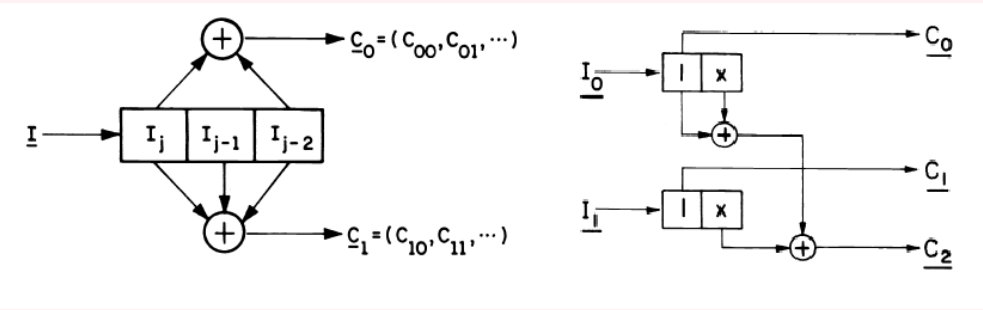
8.6 Konvolucioni kodovi

Ovo je klasa kodova koji su veoma slični opisanim linearnim kodovima, sa tom razlikom što su G, $\mathbf{u}$ i $\mathbf{c}$ polinomi. Zapravo, za razliku od linearnih blok kodova, ovde smatramo da je ulazna poruka $\mathbf{u}(x)$ prilično dugačka. Sledi primer dve matrice koje predstavljaju konvolucione kodove:

$$G_1 = \begin{bmatrix} x^2+1 & x^2+x+1 \end{bmatrix}, \quad G_2 = \begin{bmatrix} 1 & 0 & x+1 \\ 0 & 1 & x \end{bmatrix}$$

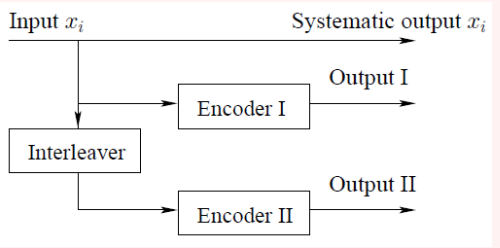
Kodiranje se obavlja slično kao kod linearnih blok kodova: $\mathbf{c}(x) = \mathbf{u}(x) \cdot G(x)$ dok se za dekodiranje koristi Viterbijev algoritam o kome će biti reči u narednom odeljku.

Ovi kodovi se efikasno realizuju pomoću pomeračkih registara (slika 8.6):



Slika 8.6: Realizacija konvolucionih kodova pomoću pomeračkih registara.

Kombinacijom od 2 (najčešće) konvoluciona koda, uz pomoć interlivera, mogu da se dostignu performanse bliske Shannonovoj granici. Ovakvi kodovi su poznati kao **turbo kodovi** ili **PCCC (Parallel Concatenated Convolutional Codes)**. Šema kodera data je na slici 8.7.



Slika 8.7: Blok šema kodera za turbo kodove.

Interliver (*interleaver*) je blok čiji je zadatak da izmeša ulazni niz bitova i tako razbije eventualne grupisane greške. Jednostavan primer mešanja se dobija kada se ulazni niz bitova upisuje u tabelu po redovima a čita po kolonama (*row–column interleaver*. Malo složenija varijanta je kada čitanje ide periodično po dijagonalama (*helical interleaver*). Početak je iz levog donjeg polja, kretanje je udesno i nagore, a kada se dođe do gornje ili desne ivice, kretanje se nastavlja sa donjeg ili levog kraja. Na slici 8.8 data su ova dva interlivera.

Input				
x_1	x_2	x_3	x_4	x_5
x_6	x_7	x_8	x_9	x_{10}
x_{11}	x_{12}	x_{13}	x_{14}	x_{15}

Row-column interleaver output														
x_1	x_6	x_{11}	x_2	x_7	x_{12}	x_3	x_8	x_{13}	x_4	x_9	x_{14}	x_5	x_{10}	x_{15}

Helical interleaver output														
x_{11}	x_7	x_3	x_{14}	x_{10}	x_1	x_{12}	x_8	x_4	x_{15}	x_6	x_2	x_{13}	x_9	x_5

Slika 8.8: Blok šema kodera za turbo kodove.

Ovi kodovi se primenjuju npr. u 3G mobilnoj telefoniji, satelitskim komunikacijama, itd.

8.7 AWGN kanal i dekodiranje linearnih blok kodova

Do sada smo pretpostavljali da su i ulaz i izlaz iz kanala diskretni. Pošto su svi komunikacioni medijumi u osnovi kontinualni, niz ulaznih simbola je najpre potrebno prevesti u vrednosti nekog kontinualnog signala (napona, jačine svetlosti, frekvencije, faze, itd.), zatim poslati kroz kanal, i na kraju dobiti odgovarajuće diskretne vrednosti na osnovu izlaza.

Ukoliko odmah nakon prijema izvršimo diskretizaciju, dobijamo niz simbola na izlazu, na koji dalje primenjujemo odgovarajući dekoder. Ovaj dekoder se naziva **hard dekoder** jer radi sa već zaokruženim (diskretnim) vrednostima.

Sa druge strane, moguće je raditi dekodiranje već na osnovu primljenih kontinualnih vrednosti. Ovakav tip dekodiranja naziva se **soft dekodiranje** i često daje mnogo preciznije procene poslate vrednosti od hard dekodiranja. Pokazaćemo ovo svojstvo na jednom primeru:

Primer 8.7.1. Pretpostavimo da šaljemo vrednost $W = 0, 1$ putem nekog medijuma (kabl, radio link, itd.) na sledeći način: Ukoliko je $W = 0$ postavljamo vrednost signala na ulazu $X = A$ a ukoliko je $W = 1$ postavljamo je na $X = -A$. Vrednost A može biti bilo koji pozitivan broj. Tokom prenosa signala sa ulaza na izlaz medijuma dolazi do smetnji. Pretpostavimo da ove

smetnje mogu da se predstave kao slučajna promenljiva Z koja ima normalnu raspodelu sa parametrom σ^2. Drugim rečima, pretpostavka je da zbog smetnji (šuma) na izlazu primamo:

$$Y = X + Z, \qquad p_Z(z) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{z^2}{2\sigma^2}}, \quad p(y|x) = p_Z(y - x).$$

Šum Z može da potiče i od uticaja drugih signala, nesavršenosti izrade prenosnog medijuma, itd. Pošto trenutno nemamo nijednu drugu informaciju na raspolaganju, ukoliko želimo da procenimo da li je poslata poruka $W = 0$ ili $W = 1$, to možemo da uradimo na sledeći način:

$$\hat{W} = \begin{cases} 0, & Y \geq 0 \\ 1, & Y < 0 \end{cases}$$

Međutim, nisu sve procene koje obavimo na ovaj način "dovoljno dobre", što ćemo pokazati na konkretnim vrednostima.

Pretpostavimo da je npr. $\sigma^2 = A = 1$, da smo poslali $w = 1$ ($x = -1$) i primili vrednost $y = -2$. Tada je

$$P(Y = -2|W = 1) = P(Y = -2|X = -1) = p_Z(-1) \approx 0.242$$
$$P(Y = -2|W = 0) = P(Y = -2|X = 1) = p_Z(-3) \approx 0.004$$

Dakle, prva verovatnoća je skoro 55 puta veća od druge. Sa druge strane, ukoliko je primljena vrednost npr. $y = -0.1$ onda je

$$P(Y = -0.1|W = 1) = P(Y = -0.1|X = -1) = p_Z(0.9) \approx 0.266$$
$$P(Y = -0.1|W = 0) = P(Y = -0.1|X = 1) = p_Z(-1.1) \approx 0.218$$

tj. razlika je drastično manja. Očigledno je da bi u oba slučaja dekodirali $\hat{W} = 1$, ali u prvom je ta odluka mnogo "jasnija".

Kanal u kome je izlazni signal jednak

$$Y = X + Z, \qquad Z : \mathcal{N}(0, \sigma^2)$$

naziva se **AWGN (Additive White Gaussian Noise) kanal**. Detaljniji prikaz ovog kanala dat je u Dodatku B. Ukoliko dodatno pretpostavimo da se i kodiranje odvija na način opisan u prethodnom primeru ($X = A$ ukoliko šaljemo 0 i $X = -A$ ukoliko šaljemo 1) dobijamo **BI-AWGN** kanal. Ovaj

kanal se javlja, između ostalog, i kada se koristi **BPSK** digitalna modulacija, što je takođe detaljnije opisano u Dodatku B.

Razmotrimo sada detaljnije postupak dekodiranja u slučaju **BI-AWGN** kanala. Na sličan način, dokazuje se da je, za slučaj uniformne raspodele ulaznog indeksa W, ML (Maximum Likelihood) dekoder optimalan [6].

Dakle, za primljen niz $\mathbf{y} = (y_1, y_2, \ldots, y_n) \in \mathbb{R}^n$ potrebno je pronaći kodnu reč $\mathbf{x} \in \mathcal{C}$ koda takvu da je $p(\mathbf{y}|\mathbf{x})$ maksimalno (imajmo u vidu da je ovo sada **gustina raspodele** a ne raspodela, kao u diskretnom slučaju). S obzirom da je

$$\begin{aligned} p(\mathbf{y}|\mathbf{x}) &= \prod_{i=1}^{n} p(y_i|x_i) = \prod_{i=1}^{n} p_Z(y_i - (-1)^{x_i} A) \\ &= \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - (-1)^{x_i} A)^2} \end{aligned}$$

cilj nam je da odredimo kodnu reč $\mathbf{x} \in \mathcal{C}$ tako da se minimizuje sledeće "rastojanje":

$$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} d(x_i, y_i), \qquad d(x_i, y_i) = (y_i - (-1)^{x_i} A)^2.$$

U nastavku, opisaćemo metod baziran na dinamičkom programiranju za određivanje kodne reči $\mathbf{x}$ takve da je $d(\mathbf{x}, \mathbf{y})$ minimalno. Ovaj metod je poznat pod nazivom **Viterbijev algoritam**.

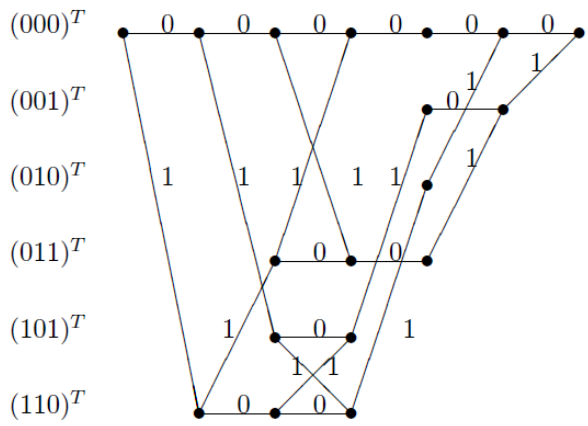
Podsetimo se da je $\mathbf{x} \in \mathcal{C}$ akko je $H\mathbf{x}^T = 0$ gde je H kontrolna (*parity–check*) matrica. Definišimo niz stanja $\mathbf{s}_0, \mathbf{s}_1, \ldots, \mathbf{s}_n$ na sledeći način:

$$\mathbf{s}_0 = 0, \quad \mathbf{s}_l = \mathbf{s}_{l-1} + x_l \mathbf{h}_{\bullet l}, \quad l = 1, 2, \ldots, n.$$

Pritom smo sa $\mathbf{h}_{\bullet l}$ označili l-tu kolonu matrice H. Očigledno je $\mathbf{s}_n$ sindrom, tj. $\mathbf{x} \in \mathcal{C}$ akko $\mathbf{s}_n = 0$. Kod $\mathcal{C}$ može da se predstavi tzv. **trelis dijagramom**. Čvorovi ovog dijagrama podeljeni su u nivoe, i u svakom nivou k postoji čvor za svaku moguću vrednost $\mathbf{s}_l$. Grana između nekog čvora $\mathbf{s}_l$ i $\mathbf{s}_{l-1}$ postoji, ukoliko je $\mathbf{s}_l = \mathbf{s}_{l-1} + x_l \mathbf{h}_{\bullet l}$ za neko $\mathbf{x} \in \mathcal{C}$.

Na slici 8.9 dat je primer trelis dijagrama za kod definisan kontrolnom matricom H.

[6] Dokaz je sličan dokazu Teoreme 7.5.3.



$$H = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix} = [\mathbf{h}_{\bullet 1}\ \mathbf{h}_{\bullet 2}\ \mathbf{h}_{\bullet 3}\ \mathbf{h}_{\bullet 4}\ \mathbf{h}_{\bullet 5}\ \mathbf{h}_{\bullet 6}]$$

Nivoi idu s leva na desno $l = 0, 1, 2, \ldots$ a u svakom nivou su sve moguće vrednosti $\mathbf{s}_l$ (dobijene za svako $\mathbf{x} \in \mathcal{C}$).

Slika 8.9: Primer trelis dijagrama

Vidimo da svaki put u trelisu od $\mathbf{s}_0 = 0$ do $\mathbf{s}_n = 0$ određuje jednu kodnu reč. Dakle, potrebno je naći takav put za koji je $d(\mathbf{x}, \mathbf{y})$ minimalno. Neka je $g_l(\mathbf{s})$ minimalan takav put do $\mathbf{s}$ na nivou l. Tada važi sledeća rekurentna veza:

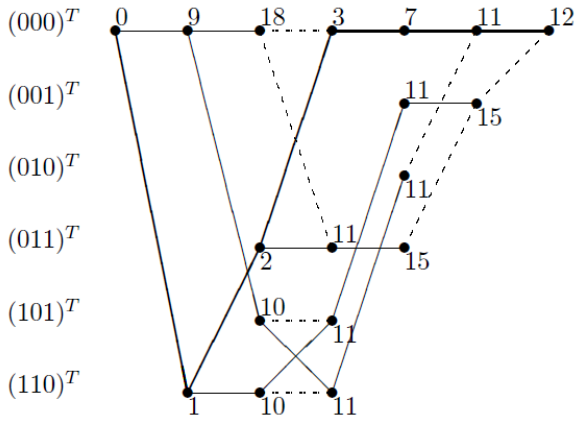
$$g_l(\mathbf{s}) = \min\{g_{l-1}(\mathbf{s}') + d(x_l, y_l) \mid \mathbf{s} = \mathbf{s}' + \mathbf{h}_{\bullet l} x_l, \quad x_l = 0, 1\}$$

kojom računamo $g_l(\mathbf{s})$ za svako $\mathbf{s}$ na trelisu i $l = 1, 2, \ldots, n$. Podsetimo se da je $d(x_l, y_l) = (y_l - (-1)^{x_l} A)^2$.

Primer 8.7.2. Pretpostavimo da je dat kod iz prethodnog primera, da je $A = 1$ i da smo primili $\mathbf{y} = (-2, -2, -2, -1, -1, 0)$. Na slici 8.10 su date izračunate vrednosti minimalne dužine puta $g_l(\mathbf{s})$.

Pune linije odgovaraju vrednostima $x_l = 0$ a isprekidane vrednostima $x_l = 1$. Podebljana linija predstavlja minimalni put od $\mathbf{s}_0 = 0$ do $\mathbf{s}_l = 0$, što ujedno predstavlja i dekodiranu kodnu reč $\mathbf{x} = (1, 1, 1, 0, 0, 0)$.

Napomenimo da Viterbijev algoritam može da se korisiti i u slučaju diskretnih kanala za $d(\mathbf{x}, \mathbf{y}) = d_H(\mathbf{x}, \mathbf{y})$. Uz pomoć Viterbijevog algoritma možemo efikasnije konstruisati tabelu minimalnih korektora za svaki sindrom. Složenost ove konstrukcije je, u najgorem slučaju $\mathcal{O}(q^k \cdot n)$ (treba proći kroz sve moguće vrednosti stanja $\mathbf{s}_l$ (kojih ima, u najgorem slučaju, q^{n-k}) i sve nivoe $l = 0, 1, \ldots, n$). Ovo je primetno manje od $\mathcal{O}(q^n)$, koliko je potrebno da se prođe kroz sve moguće reči $\mathbf{y} \in \{0, 1\}^n$ koje mogu da se jave na prijemnoj strani.



Slika 8.10: Izračunate vrednosti minimalnih dužina puteva za svaki čvor trelis dijagrama.

8.8 LDPC kodovi

Ove kodove je otkrio ih Gallager 1962. godine u svom doktoratu, gde je dao i algoritme za kodiranje i dekodiranje. Međutim, zbog nemogućnosti praktične realizacije ostali su nezapaženi sve do 1996. godine kada su ih MacKay i Neal ponovo otkrili.

To su linearni blok kodovi sa retkom (sparse) kontrolnom matricom H. Regularni LDPC kodovi zadovoljavaju svojstvo da svaki red odnosno kolona matrice H sadrži w_r odnosno w_c jedinica. Matrica H je **retka** ako važi $w_r \ll m$ i $w_c \ll n$.

8.8.1 Konstrukcija

Gallager je u svom radu opisao sledeći metod za konstrukciju ovih kodova. Prvih $(n-k)/w_c$ redova dobijaju se tako što se najpre upiše w_r jedinica u prvi red (kolone sa indeksima od 1 do w_r), zatim se pređe u sledeći red i opet upiše w_r jedinica (kolone sa indeksima od w_r+1 do $2w_r$), itd. Ostalih w_c-1 grupa redova dobijaju se permutacijom kolona prve grupe redova.

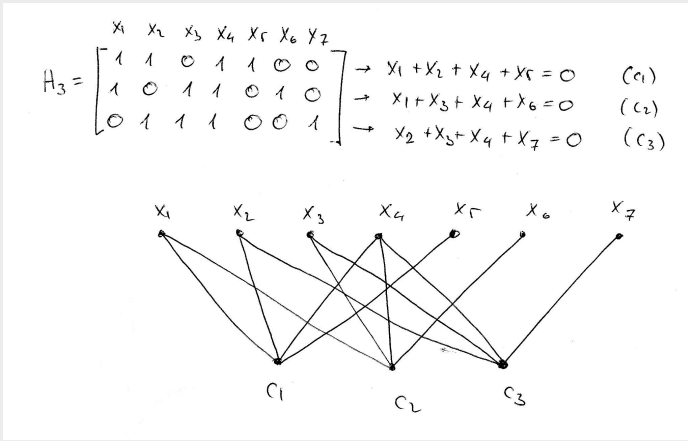
Primer 8.8.1. Dat je primer konstrukcije za $n = 12$, $n - k = 9$, $w_r = 4$,

$w_c = 3$:

$$H = \left[\begin{array}{cccccccccccc} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ \hline 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ \hline 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \end{array}\right]$$

Za predstavljanje LDPC kodova pogodno je koristiti **Tannerov graf**. Ovo je bipartitan graf sa skupovima čvorova $\{x_1, x_2, \ldots, x_n\}$ kao i $\{c_1, c_2, \ldots, c_m\}$ gde je $m = n - k$. Čvorovi x_j odgovaraju elementima kodne reči $\mathbf{x}$, dok čvorovi c_i odgovaraju vrstama matrice H, odnosno linearnim jednačinama koje zadovoljava svaka kodna reč. Veza između čvorova x_j i c_i postoji ako x_j učestvuje u i-toj jednačini, odnosno ako je $h_{ij} = 1$. Tannerov graf je direktno određen kontrolnom matricom koda H.

Primer 8.8.2. Tannerov graf za Hammingov $(7, 4)$ kod dat je na sledećoj slici



Tannerov graf možemo posmatrati za proizvoljan linearni blok kod. Za LDPC kodove, ovaj graf je "redak", odnosno ima mali broj ivica. Simbolom $\mathcal{N}(x_j)$ odnosno $\mathcal{N}(c_i)$ označavamo skupove suseda čvorova x_j odnosno c_i u Tannerovom grafu koda. Tako je npr.

$$\mathcal{N}(x_3) = \{c_2, c_3\}, \qquad \mathcal{N}(c_1) = \{x_1, x_2, x_4, x_5\}.$$

Podsetimo se da je $\mathbf{x}$ kodna reč ako i samo ako je $H\mathbf{x}^T = \mathbf{0}^T$, odnosno ako su sve jednačine $c_1, c_2, \ldots, c_m$ zadovoljene. Ovaj uslov drugačije možemo da zapišemo na sledeći način:

$$\sum_{j \in \mathcal{N}(c_i)} x_j = 0, \qquad i = 1, 2, \ldots, m.$$

8.8.2 Gallager A/B algoritmi za dekodiranje

Pretpostavimo da smo na izlazu kanala primili vektor $\mathbf{y} \in \{0,1\}^n$. Podsetimo se da se proces dekodiranja **ML dekoderom** svodi na nalaženje kodne reči $\mathbf{x} \in \mathcal{C}$ takve da je Hammingovo rastojanje $d_H(\mathbf{x}, \mathbf{y})$ minimalno.

ML dekoder zahteva pretragu po celom skupu kodnih reči, ili bar konstrukciju tabele sindrom-korektor. I jedan i drugi način su algoritamski zahtevni i neprimenljivi za veće vrednosti $m = n - k$ i n. Zato ćemo sada razmotriti aproksimativne dekodere koji **ne pronalaze uvek najbolje (ML) rešenje**, ali su zato algoritamski mnogo manje zahtevni (broj operacija, potrebna memorija, itd.). Svi takvi algoritmi su iterativni i bazirani su na **razmeni poruka** između čvorova x_i i c_j.

Neka je početna vrednost svih informacionih čvorova $\mathbf{x}^{(0)} = \mathbf{y}$, odnosno vrednost primljena iz kanala. Prvi algoritam koji ćemo razmotriti sastoji se iz sledećih koraka:

1. Svaki čvor x_j pošalje svim svojim susedima $c_i \in \mathcal{N}(x_j)$ svoju vrednost u k-tom koraku $x_j^{(k)}$.

2. Svaki čvor c_i izračuna sume $\omega_{i\to j}^{(k)} = \sum_{x_l \in \mathcal{N}(c_i)\setminus\{x_j\}} x_l^{(k)}$ pristiglih poruka iz svih susednih čvorova sem jednog $x_j \in \mathcal{N}(c_i)$. Ove sume su zapravo potrebne vrednosti čvora x_j tako da je jednačina c_i zadovoljena, pod pretpostavkom da svi ostali susedi $x_l \in \mathcal{N}(c_i)$ imaju vrednost $x_l = x_l^{(k)}$. Vrednost $\omega_{i\to j}^{(k)}$ šalje se čvoru x_j. Primetimo da $\omega_{i\to j}^{(k)}$ možemo da izračunamo efikasnije i na sledeći način:

$$\omega_{i\to j}^{(k)} = \omega_i^{(k)} - x_j^{(k)}, \qquad \omega_i^{(k)} = \sum_{x_l \in \mathcal{N}(c_i)} x_l^{(k)}$$

   odnosno kao sumu vrednosti svih suseda minus vrednost tog kom se poruka šalje.

3. Svaki čvor x_j primi sve vrednosti $\omega_{i\to j}^{(k)}$ od svojih suseda $c_i \in \mathcal{N}(x_j)$. Neka je $d_{j,0}^{(k)}$ broj pristiglih nula a $d_{j,1}^{(k)}$ broj pristiglih jedinica na ovaj način.

Čvor računa svoju vrednost u $k+1$-voj iteraciji "većinskim glasanjem":

$$x_j^{(k+1)} = \begin{cases} 0, & d_{j,0}^{(k)} \geq th_0 \cdot n_j \\ 1, & d_{j,1}^{(k)} \geq th_1 \cdot n_j \\ y_j, & \text{u suprotnom} \end{cases}$$

Ovde je $n_j = |\mathcal{N}(x_j)|$ broj suseda čvora x_j a th_0 i th_1 su dati brojevi između 0 i 1. Ukoliko je pristiglo dovoljno nula ili jedinica, nova vrednost čvora je 0 ili 1, u suprotnom čvor vraća vrednost dobijenu iz kanala.

Ovaj algoritam redom generiše reči $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(k)}, \ldots$ i zaustavlja se kada su sve jednačine $c_1, c_2, \ldots, c_m$ zadovoljene. Ako se to dogodi u k-tom koraku, onda je $\mathbf{x}^{(k)} = \mathbf{x}^{(k+1)} = \cdots$. Ukoliko algoritam ne dostigne stacionarno stanje posle određenog broja koraka, postupak se obustavlja i konstatuje se da je došlo do greške prilikom dekodiranja. Druga varijanta je da se vrate dobijene vrednosti informacionih bitova, ali tada nemamo nikakve garancije koliko ima tačnih a koliko pogrešnih bitova.

Ovaj algoritam je formulisao još Gallager u svom radu, pri čemu je **A** varijanta algoritma podrazumevala da je $th_0 = th_1 = 1$. Odnosno, da se promena vrednosti obavlja isključivo ako sve jednačine "tvrde" da x_j ima određenu vrednost (0 ili 1). Za opšte vrednosti th_0 i th_1 u pitanju je **B** varijanta algoritma.

Primer 8.8.3. Posmatrajmo ponovo Hammingov $(7, 4)$ kod i pretpostavimo da smo na prijemu dobili reč $\mathbf{y} = 1000000$. Pretpostavimo da su pragovi odlučivanja $th_0 = th_1 = 0.5$. Nije teško uočiti da ćemo odmah nakon prve iteracije dobiti $\mathbf{x}^{(1)} = 0000000$, što jeste kodna reč.

8.8.3 Algoritmi za dekodiranje za BSC i BEC kanale

[LDPC tutorial, sect. 1–3.1, p. 1–6]

8.8.4 Algoritam razmene poruka

Gallager A/B algoritmi su primenljivi za dekodiranje na BSC kanalu, odnosno za hard dekodiranje. Dekoder koji ćemo sada opisati je soft dekoder i primenljiv je na široki spektar kako diskretnih tako i kontinualnih kanala.

Princip je sličan Gallager A/B algoritmima, samo što poruke imaju drugačiji sadržaj.

Pretpostavimo za početak da se vrednosti čvorova x_j biraju slučajnim izborom. Označimo odgovarajuću slučajnu promenljivu sa X_j i pretpostavimo da su one **nezavisne**, pri čemu je $X_j = 1$ sa verovatnoćom p_j a $X_j = 0$ sa verovatnoćom $q_j = 1 - p_j$.

Pretpostavimo i da je kod takav da ne postoje dva čvora c_i i c_k takva da imaju bar dva zajednička suseda, odnosno da je $|\mathcal{N}(c_i) \cap \mathcal{N}(c_k)| \geq 2$.

Cilj je da izračunamo verovatnoću da je $p'_j = P(X_j = 1 \mid H\mathbf{X}^T = \mathbf{0}^T)$. Ovo je zapravo verovatnoća događaja $X_j = 1$ pod uslovom da je $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ kodna reč.

Lema 8.8.1. *Verovatnoća da je i-ta jednačina zadovoljena (odnosno da je $\mathbf{h}_i\mathbf{X} = 0$) jednaka je*

$$P(Par_i) = \frac{1}{2} + \frac{1}{2} \prod_{j \in \mathcal{N}(c_i)} (q_j - p_j)$$

Ovaj događaj označen je sa Par_i.

Dokaz. Neka je $n_i = |\mathcal{N}(c_i)|$ i

$$A(t) = \prod_{j \in \mathcal{N}(c_i)} (q_j + tp_j) = \sum_{k=0}^{n_i} S_k t^k.$$

Koeficijent S_k jednak je zbiru svih koeficijenata uz t^k u razvoju proizvoda $A(t)$. Svaki od sabiraka predstavlja jedan način da se izabere k čvorova $\{x_j \mid j \in \mathcal{N}(c_j)\}$ koji će imati vrednost 1, dok će ostali imati vrednost 0. Verovatnoća ovog izbora je proizvod vrednosti p_j za čvorove koji imaju vrednost 1, a q_j za one koji imaju vrednost 0, i upravo je jednak vrednosti odgovarajućeg sabirka u razvoju. Samim tim, koeficijent S_k predstavlja verovatnoću da će među vrednostima X_j, $j \in \mathcal{N}(c_i)$ biti tačno k jedinica.

Pošto je

$$A(t) + A(-t) = 2 \sum_{k=0}^{\lfloor n_i \rfloor} S_{2k} t^{2k}$$

onda je

$$\frac{1}{2}(A(1) + A(-1)) = \sum_{k=0}^{\lfloor n_i \rfloor} S_{2k}$$

verovatnoća da će među vrednostima X_j, $j \in \mathcal{N}(c_i)$ biti paran broj jedinica, odnosno da će jednačina i biti zadovoljena. Prema tome

$$P(Par_i) = \frac{1}{2}(A(1) + A(-1)) = \frac{1}{2}\left(1 + \prod_{j \in \mathcal{N}(c_i)} (q_j - p_j)\right).$$

Ovim je dokaz završen. □

Iz prethodne leme, nije teško zaključiti da je verovatnoća da je $X_j = 0$ ukoliko je i-ta jednačina zadovoljena jednaka

$$q'_{j,i} = P(X_j = 0 \mid Par_i) = \frac{P(X_j = 0)P(Par_i \mid X_j = 0)}{P(Par_i)}$$
$$= \frac{q_i \cdot \left(\frac{1}{2} + \frac{1}{2} \prod_{k \in \mathcal{N}(c_i) \setminus \{X_j\}} (q_k - p_k)\right)}{P(Par_i)}$$

Na sličan način je i

$$p'_{j,i} = P(X_j = 1 \mid Par_i) = \frac{P(X_j = 1)P(Par_i \mid X_j = 1)}{P(Par_i)}$$
$$= \frac{p_i \cdot \left(\frac{1}{2} - \frac{1}{2} \prod_{k \in \mathcal{N}(c_i) \setminus \{x_j\}} (q_k - p_k)\right)}{P(Par_i)}$$

Posmatrajmo sada odnose $l_j = \ln(q_j/p_j)$ i $l'_{j,i} = \ln(q'_{j,i}/p'_{j,i})$. Iz prethodna dva izraza dobijamo da je

$$l'_{j,i} = l_j + \tau_{ij}, \qquad \tau_{ij} = \ln \frac{1 + \prod_{k \in \mathcal{N}(c_i) \setminus \{x_j\}} (q_k - p_k)}{1 - \prod_{k \in \mathcal{N}(c_i) \setminus \{x_j\}} (q_k - p_k)}.$$

Pošto je $l_k = \ln(q_k/p_k)$ i $p_k = 1 - q_k$, sledi da je

$$q_k = \frac{e^{l_k}}{1 + e^{l_k}}, \quad p_k = 1 - q_k = \frac{1}{e^{l_k} + 1}$$

odnosno

$$q_k - p_k = \frac{e^{l_k} - 1}{e^{l_k} + 1} = \frac{e^{l_k/2} - e^{-l_k/2}}{e^{l_k/2} + e^{-l_k/2}} = \tanh(l_k/2).$$

Sa druge strane, iz $\operatorname{arctanh}(x) = \ln((1+x)/(1-x))$, dobijamo

$$\tau_{ij} = \operatorname{arctanh}\left(\prod_{k \in \mathcal{N}(c_i) \setminus \{x_j\}} \tanh(l_k/2)\right). \tag{8.3}$$

Izračunajmo sada verovatnoću $q'_j = P(X_j = 0 \mid H\mathbf{X} = \mathbf{0}^T) = P(X_j = 0 \mid Par_1, Par_2, \ldots, Par_m)$. Iz Bajesove formule sledi

$$q'_j = \frac{q_j P(Par_1, Par_2, \ldots, Par_m \mid X_j = 0)}{P(Par_1, Par_2, \ldots, Par_m)}.$$

Pošto ne postoje dva čvora c_j i c_k sa bar dva zajednička suseda, sledi da su skupovi $\mathcal{N}(c_i) \setminus \{x_j\}$ i $i = 1, 2, \ldots, m$ disjunktni, pa je

$$P(Par_1, Par_2, \ldots, Par_m \mid X_j = 0) = \prod_{i \in \mathcal{N}(x_j)} P(Par_i \mid X_j = 0).$$

Na sličan način računamo i $p'_j = P(X_j = 1 \mid H\mathbf{X} = \mathbf{0}^T)$ pa je

$$l'_j = \ln(q'_j/p'_j) = \ln(q_j/p_j) + \sum_{i \in \mathcal{N}(x_j)} \frac{P(Par_i \mid X_j = 0)}{P(Par_i \mid X_j = 1)}$$

odnosno

$$l'_j = l_j + \sum_{i \in \mathcal{N}(x_j)} \tau_{ij}. \tag{8.4}$$

Izrazi (8.3) i (8.4) daju rekurentnu vezu za iterativni metod:

$$l_j^{(k+1)} = l_j^{(k)} + \sum_{i \in \mathcal{N}(x_j)} \tau_{ij}^{(k)}, \qquad \tau_{ij}^{(k)} = \operatorname{arctanh}\left(\prod_{t \in \mathcal{N}(c_i) \setminus \{x_j\}} \tanh(l_t^{(k)}/2)\right). \tag{8.5}$$

Vratimo se sada na problem dekodiranja. Ukoliko iterativni metod (8.5) inicijalizujemo aposteriornim verovatnoćama iz kanala $p_j = p(0|y_j)$ i $q_j = p(1|y_j)$, odnosno

$$l_j^{(0)} = \ln(p(1|y_j)/p(0|y_j)), \quad j = 1, 2, \ldots, n$$

nakon dovoljno iteracija dobijamo aproksimaciju logaritamske verodostojnosti $l_j^{(k)}$ za aposteriorne verovatnoće:

$$l_j^{apo} = \ln(q_j^{apo}/p_j^{apo}), \quad p_j^{apo} = P(X_i = 1 \mid \mathbf{Y} = \mathbf{y},\ H\mathbf{X}^T = \mathbf{0}^T), \quad q_j^{apo} = 1 - p_j^{apo}.$$

Dekodiranje sada možemo da obavimo na osnovu vrednosti l_j^{apo}. Ako je $q_j^{apo} > p_j^{apo}$ onda je $l_j^{apo} > 0$ pa je $\hat{x}_i = 1$ a u suprotnom je $\hat{x}_i = 0$. Dakle,

$$\hat{x}_i = \begin{cases} 1, & l_j^{apo} > 0 \\ 0, & l_j^{apo} \leq 0 \end{cases}.$$

Ovaj metod određuje tačne vrednosti aposteriorne log-verodostojnosti ukoliko je Tannerov graf koda stablo. U suprotnom, ove vrednosti su približne i to tačnije, ukoliko je minimalni ciklus grafa (*girth*) veći.

8.9 Za dalje čitanje

- Burst–error correction [McEliece, sect. 8.4, p. 199]
- BCH codes [McEliece, sect. 9.2–9.5]
- Reed–Solomon codes [McEliece, sect. 9.6]
- Convolutional codes [McEliece, sect. 10]
- LDPC codes (MacKay–Neal construction, encoding, soft decoding, decoding for BEC) [...]
- Određivanje performansi koda Monte–Carlo metodom.
- Digitalne modulacije (ASK, FSK, BPSK, QPSK, M-PSK, QAM, ...)
- Kvantizacija (skalarni, vektorski kvantizeri)

Dodatak A

Konačna polja

Za razmatranje široke klase kodova koji se koriste u praksi, potrebno je uvesti određene algebarske operacije nad ulaznim i izlaznim alfabetom. Najčešća pretpostavka je da je $\mathcal{X} = \mathcal{Y} = \mathbb{F}$ i da $\mathbb{F}$ ima strukturu **polja**.

U ovom odeljku navodimo samo najosnovnije svojstva konačnih polja koja su nam potrebna za dalji rad.

Definicija A.0.1. *Struktura $(\mathbb{F}, +, \cdot)$ je **polje** ukoliko važe sledeći izrazi za svako $x, y, z \in \mathbb{F}$ i za određene elemente $0, 1 \in \mathbb{F}$:*

1. $x + (y + z) = (x + y) + z$
2. $x + 0 = 0 + x = x$
3. *Postoji jedinstveni element* $-x \in \mathbb{F}$ *takav da je* $x + (-x) = (-x) + x = 0$
4. $x + y = y + x$
5. $x \cdot (y \cdot z) = (x \cdot y) \cdot z$
6. $x \cdot 1 = 1 \cdot x = x$
7. *Za* $x \neq 0$, *postoji jedinstveni element* $x^{-1} \in \mathbb{F}$ *takav da je* $x \cdot x^{-1} = x^{-1} \cdot x = 1$
8. $x \cdot y = y \cdot x$
9. $(x + y) \cdot z = x \cdot z + y \cdot z$
10. $x \cdot (y + z) = x \cdot y + x \cdot z$

*Ako je $\mathbb{F}$ konačan skup, onda je u pitanju **konačno polje**.*

Neka je $p \in \mathbb{N}$ prost broj. **Konačno polje reda** p je skup $\mathbb{F}_p = \{0, 1, \ldots, p-1\}$ sa operacijama $+_p$ i $\cdot_p$ koje su definisane sa

$$x +_p y = (x + y) \bmod p, \qquad x \cdot_p y = (x \cdot y) \bmod p. \tag{A.1}$$

Drugim rečima, nad elementima skupa $\mathbb{F}_p$ sve operacije primenjujemo po modulu p. Može se pokazati da je polje $\mathbb{F}_p$ **jedinstveno do na izomorfizam**, tj. da se operacije u svakom drugom polju $\mathbb{F}$ sa p elemenata (**red** označava broj elemenata) obavljaju na isti način kao i u $\mathbb{F}_p$ [1]. Polje $\mathbb{F}_p$ označavamo i sa $GF(p)$ [2]. Direktnom proverom potvrđujemo da važe uslovi 1–10 iz definicije polja.

Primer A.0.1. Posmatrajmo polja $GF(2) = \{0, 1\}$ i $GF(3) = \{0, 1, 2\}$. Ovo su ujedno i polja sa prostim redom koja su najčešće u upotrebi. Date su tablice operacija $+_2$, $\cdot_2$, $+_3$ i $\cdot_3$:

$+_2$	0	1
0	0	1
1	1	0

$\cdot_2$	0	1
0	0	0
1	0	1

$+_3$	0	1	2
0	0	1	2
1	1	2	0
2	2	0	1

$\cdot_3$	0	1	2
0	0	0	0
1	0	1	2
2	0	2	1

Primetimo da je $+_2$ zapravo operacija `XOR` nad bitovima dok je $\cdot_2$ operacija `AND`.

Konačna polja su važna za konstrukciju kodova zato što je nad njima moguće izvoditi operacije na gotovo isti način kao i nad racionalnim brojevima (koji takođe formiraju strukturu polja). Jedina razlika je sto su osnovne operacije ($+$, $-$, $\cdot$ i $/$) drugačije definisane nego inače.

Nadalje ćemo operacije u polju $\mathbb{F}_p$, kao i u bilo kom polju $\mathbb{F}$, označavati standardno sa $+$ i $\cdot$. Pritom, vodićemo računa da uvek naglasimo o kojoj operaciji se radi.

Primer A.0.2. Za svaki element $x \in \mathbb{F}_p$, element $-x$ je zapravo $p - x$. To se lako pokazuje, zato što je $(p - x) +_p x = (p - x + x) \bmod p = 0$. Sa druge strane, element $x = a^{-1}$ je rešenje jednačine $a \cdot_p x = 1$ odnosno $ax \equiv_p 1$. Za dato a i p, ovu jednačinu rešavamo tako što probamo sve mogućnosti za $x \in \mathbb{F}_p$. Posmatrajmo, na primer, element $a = 2$ u polju $\mathbb{F}_5 = GF(5)$. Tada je $x = a^{-1} = 3$ jer je $ax = 3 \cdot 2 \equiv_5 1$.

Jedno interesantno svojstvo polja $GF(p)$ dato je sledećom lemom koju nećemo dokazivati.

[1] Formalno govoreći, ako je $(\mathbb{F}, +', \cdot')$ drugo polje, postoji funkcija $f : \mathbb{F} \to F_p$ koja je bijekcija i za koju važi $f(x +' y) = f(x) +_p f(y)$ kao i $f(x \cdot' y) = f(x) \cdot_p f(y)$. Dakle, sabrati elemente $x, y \in \mathbb{F}$ u $\mathbb{F}$, svodi se na sabiranje po modulu p elemenata $f(x)$ i $f(y)$.

[2] "GF" je skraćenica od "Galois Field", a naziv potiče od Evarista Galoisa, koji je ova polja otkrio.

Lema A.0.1. *U polju $GF(p)$ je $x^p + y^p = (x + y)^p$ za svako $x, y \in GF(p)$.*

Pored polja $\mathbb{F}_p = GF(p)$, moguće je uvesti i polja reda p^m. Važi i obratno, da je svako konačno polje reda p^m za neki prost broj p i prirodan broj m. Skup elemenata polja reda p^m dat je sa

$$GF(p^m) = \{(a_1, a_2, \ldots, a_m) \mid a_i \in GF(p), \quad i = 1, 2, \ldots, m\} = GF(p)^m.$$

Operacija sabiranja + definiše se "pokoordinatno" na sledeći način:

$$a + b = (a_1 + b_1, a_2 + b_2, \ldots, a_m + b_m)$$

dok se operacija množenja definiše malo komplikovanije. Da bi definisali množenje dve m-torke, posmatrajmo ih u obliku polinoma:

$$a(x) = a_1 + a_2x + \ldots + a_mx^{m-1}, \quad b(x) = b_1 + b_2x + \ldots + b_mx^{m-1}.$$

Neka je $f(x)$ polinom stepena m sa koeficijentima u $GF(p)$ takav da ne može da se napiše kao proizvod dva polinoma (stepena većeg od 0) sa koeficijentima takođe u $GF(p)$. Ovakvi polinomi $f(x)$ nazivaju se **ireducibilni polinomi**. Proizvod $c = a \cdot b$ tada definišemo kao

$$c(x) = (a(x) \cdot b(x)) \bmod f(x).$$

Nakon ovakve definicije, prirodno se postavljaju sledeća dva pitanja:

1. Da li postoji (bar jedan) ireducibilni polinom $f(x)$?
2. Da li dva različita polinoma daju isto polje?

Može se pokazati da je odgovor na prvo pitanje potvrdan, tj. da za svako m postoji ireducibilni polinom $f(x)$ stepena m sa koeficijentima iz $GF(p)$. Samim tim, za svako $m \in \mathbb{N}$ postoji konačno polje $GF(p^m)$.

Iako je odgovor na drugo pitanje negativan, pokazuje se da su tako dobijena polja izomorfna. Samim tim, ova polja imaju istu strukturu, pa ih zato poistovećujemo.

Ovom prilikom nećemo se baviti problemom nalaženja (jednog) ireducibilnog polinoma nad $GF(p)$, kao ni uopšte daljim svojstvima konačnih polja. Za konstrukciju kodova je najčešće dovoljno da odgovarajući izvorni i odredišni alfabeti imaju strukturu polja.

Primer A.0.3. Polinom $f(x) = x^3+x+1$ je ireducibilan u $GF(2)$, pa pomoću njega možemo definisati množenje u $GF(2^3)$. Npr, neka je $a = 101$ a $b = 110$. Tada je $a(x) = x^2 + 1$ a $b(x) = x + 1$ pa je

$$\begin{aligned} c(x) &= (a(x)b(x)) \bmod f(x) = (x^2+1)(x+1) \\ &= (x^3 + x^2 + x + 1) \bmod (x^3 + x + 1) = x^2 \end{aligned}$$

Na osnovu ovoga zaključujemo da je $c = a \cdot b = 010$. Pokoordinatnim sabiranjem dobijamo da je $a + b = 011$.

Iako su operacije nad konačnim poljima $GF(p^m)$ (pa i nad $GF(p)$) komplikovane u opštem slučaju, razvijene su specijalne komponente koje ove operacije realizuju (naročito za $GF(2^m)$ pošto ova polja imaju direktnu primenu u realizaciji kompleksnih zaštitnih kodova). Takođe, postoje gotove programske biblioteke (npr. [9]) koje realizuju aritmetiku u konačnim poljima.

U nastavku se nećemo baviti realizacijom operacija, kao i daljim (mnogobrojnim) svojstvima konačnih polja, već ćemo samo podrazumevati da nad alfabetom koda i izvora postoje operacije + i $\cdot$ koje formiraju polje, odnosno ponašaju se kao u slučaju racionalnih brojeva. Za detaljnije upoznavanje sa konačnim poljima, čitalac može da pogleda npr. [11].

Dodatak B

AWGN kanal i digitalne modulacije

B.1 Definicija i kapacitet AWGN kanala

Pored do sada pominjanih (i proučenih) diskretnih kanala, često je u upotrebi i jedan kontinualan kanal, poznat pod nazivom **AWGN (Adaptive White Gaussian Noise) kanal**. Ukoliko je X podatak koji želimo da prenesemo (X može biti, u opštem slučaju, realan broj) onda je izlaz iz kanala dat sa:

$$Y = X + Z, \qquad Z : \mathcal{N}(0, \sigma^2).$$

Dakle, šum u kanalu manifestuje se kao zbir poslate vrednosti X i slučajno generisane vrednosti Z koja ima normalnu raspodelu. Dakle:

$$p_Z(z) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{z^2}{2\sigma^2}}, \quad p(y|x) = p_Z(y - x).$$

Snagu signala [1] označavamo sa $P = \mathbb{E}X^2$ a snagu šuma sa $N = \mathbb{E}Z^2 = \sigma^2$.

Kapacitet kontinualnog kanala definiše se na sličan način kao i za diskretni kanal:

$$C = \max_{p(x),\ \mathbb{E}X^2 \leq P} I(X, Y)$$

[1] Naziv *snaga* potiče od činjenice da je npr. električna snaga na otporniku na kome je napon U jednaka $P = U^2/R$, tj. proporcionalna sa U^2.

uz dodatno ograničenje da je snaga ulaznog signala manja ili jednaka P [2]. Ukoliko su X i Y kontinualne slučajne promenljive, tada je

$$I(X,Y) = h(Y) - h(Y|X) = h(Y) - h(Z)$$

gde su sa $h(X)$ i $h(Y|X)$ označene diferencijalna entropija kao i uslovna diferencijalna entropija apsolutno neprekidne slučajne promenljive. S obzirom da je $h(Z) = 0.5 \log_2(2\pi eN)$,

$$\mathbb{E}Y^2 = \mathbb{E}X^2 + \mathbb{E}Z^2 \leq (P+N)$$

kao i $h(Y) \leq 0.5 \log_2(2\pi e(P+N))$ sledi da je

$$C \leq \frac{1}{2} \log_2 \left(1 + \frac{P}{N}\right).$$

Odnos P/N se često naziva i odnos **signal–šum** i predstavlja u decibelima na sledeći način:

$$SNR = 10 \log_{10}(P/N) \text{ dB}.$$

B.2 BPSK digitalna modulacija

Jedan od najjednostavnijih načina da se (fizički) prenese jedan bit (niz bitova) putem nekog komunikacionog mediuma je BPSK (Binary Phase Shift Keying) modulacija. Pritom, medijum može biti bilo koji, samo je važno da se kroz isti prirodno (putem određenih fizičkih procesa) može izvršiti prenos neke (kontinualne) fizičke veličine. Na primer, par bakarnih žica (parice) nam omogućava da razliku potencijala (napon) sa jednog kraja očitamo na drugom, optičko vlakno omogućava da se svetlosni talas emitovan na jednom kraju vlakna prihvati na drugom, itd.

Pretpostavimo da je $(x_n)_{n\in\mathbb{N}}$ niz bitova koje bi trebalo preneti putem nekog medijuma. Signal koji se prostire kroz medijum možemo definisati na sledeći način:

$$s_n(t) = A\cos(\omega_0 t + \phi_n), \quad \phi_n = \begin{cases} 0 & x_n = 0 \\ \pi & x_n = 1 \end{cases}$$

[2] Ovo ograničenje je prirodno, jer ukoliko je vama dovoljno jak ulaz, on će u svakom slučaju nadvladati šum i ostvarićete skoro idealan prenos. Međutim, u praksi uvek postoji ograničenje maksimalne snage koju možete emitovati (usled fizičkih svojstava uređaja).

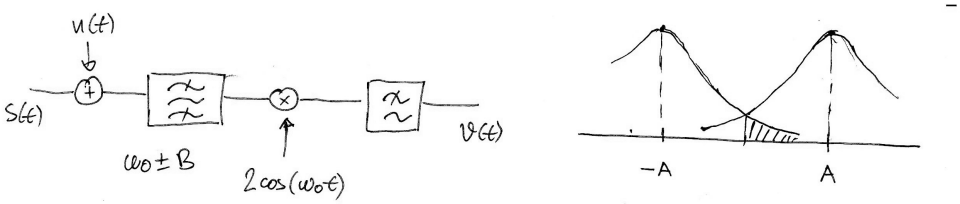
Ovde je $\omega_0 = 2\pi f_0$ gde je f_0 poznata frekvencija (frekvencija nosača), A amplituda signala a ϕ_n faza koja direktno zavisi od vrednosti bita koju prenosimo. Nije teško utvrditi da je

$$s_n(t) = \begin{cases} A\cos(\omega_0 t), & x_n = 0 \\ -A\cos(\omega_0 t), & x_n = 1 \end{cases}$$

Signal prenosimo tako što vremensku osu podelimo u segmente dužine T_b i onda u n-tom segmentu prenosimo bit x_n, odnosno emitujemo signal $s_n(t)$. Dakle:

$$s(t) = \begin{cases} s_0(t), & 0 \le t < T_b \\ s_1(t), & T_b \le t < 2T_b \\ \vdots & \\ s_n(t), & nT_b \le t < (n+1)T_b \\ \vdots & \end{cases}$$

Prilikom prenosa kroz kanal javlja se šum za koji ćemo pretpostaviti da je beli Gaussov šum.



Slika B.1: Šema prijemnika BPSK signala.

Na slici B.1 data je blok šema prijemnika za BPSK signal. Na izlazu (u vremenskom segmentu $[nT_b, (n+1)T_b]$) dobijamo:

$$v_n(t) = \begin{cases} A + Z, & x_n = 0 \\ -A + Z, & x_n = 1 \end{cases}$$

gde je $Z : \mathcal{N}(0, \sigma^2)$ šum. Pritom je $\sigma^2 = N_0 B$ gde je N_0 **spektralna snaga šuma** dok je B propusni opseg filtra. Očigledno je za $v_n(t) > 0$ veća verovatnoća da je poslata nula nego jedinica, dok je za $v_n(t) < 0$ obrnuto.

Upravo tako je realizovan odlučivač (dekoder). Verovatnoća greške jednaka je:

$$P_E = P(v_n(t_0) > 0 | x_n = 1) = P(-A + Z > 0) = \int_0^{+\infty} p_Z(z + A) dz$$
$$= \int_0^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(z+A)^2}{2\sigma^2}} dz = \frac{1}{2}\mathrm{erfc}\,(\rho), \quad \rho = \frac{A}{\sigma\sqrt{2}}.$$

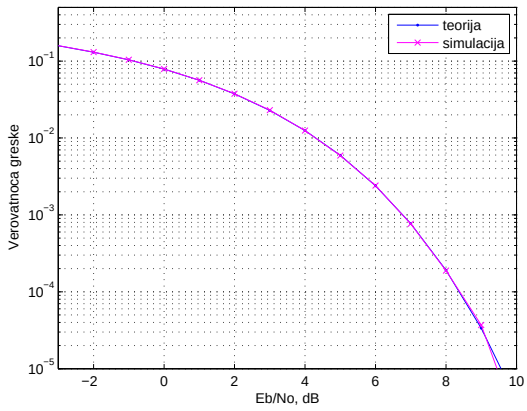
Snaga koja je potrebna da se prenese jedan bit (snaga koju ima signal $s_n(t)$) jednaka je $A^2/2$. Samim tim, ukupna energija po bitu je $E_b = A^2 T_b/2$. Sa druge strane, snaga šuma jednaka je $N = \sigma^2 = N_0 B$. S obzirom da je 99% spektra BPSK signala sadržano u opsegu širine $2/T_b$, možemo uzeti da je $B = 1/T_b$. Tada je

$$\frac{A}{\sigma\sqrt{2}} = \frac{\sqrt{2E_b/T_b}}{N_0 B\sqrt{2}} = \frac{E_b}{N_0}.$$

Prema tome, greška BPSK signala jednaka je:

$$P_E = \frac{1}{2}\mathrm{erfc}\left(\sqrt{\frac{E_b}{N_0}}\right).$$

Na slici B.2 data je zavisnost verovatnoće greške od E_b/N_0 izražene u decibelima ($10\log_{10}(E_b/N_0)$).



Slika B.2: Zavisnost verovatnoće greške BPSK signala od E_b/N_0 [dB]

Literatura

[1] T.M. Cover, J.A. Thomas, *Elements of Information Theory, 2nd ed.*, John Wiley & Sons, 2006.

[2] D. Drajić, P. Ivaniš, *Uvod u teoriju informacija i kodovanje*, Akademska misao, Beograd, 2009.

[3] Z. Ivković, *Teorija verovatnoća sa matematičkom statistikom*, Naučna knjiga, Beograd, 1989.

[4] R.J. McEliece, *The theory of information and coding, 2nd ed.*, Cambridge University Press, 2004.

[5] S. Janković, *Uvod u verovatnoću*, Prirodno-matematički fakultet u Nišu, 2009.

[6] B. Šešelja, *Teorija informacije i kodiranja*, Prirodno-Matematički fakultet u Novom Sadu, 2005.

[7] B. Šešelja, A. Tepavčević, ...

[8] `http://www.youtube.com/playlist?list=PLE125425EC837021F`

[9] Fast Galois Field Arithmetic Library in C/C++,

[10] A.W. Knapp, *Basic algebra*, Birkhauser, 2006.
`http://web.eecs.utk.edu/ plank/plank/papers/CS-07-593/`.

[11] Introduction to Finite Fields, `http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-451-principles-of-digital-communication-ii-spring-2005/lecture-notes/chap7.pdf`.

[12] B.M.J. Leiner, *LDPC Codes a brief Tutorial*,
`www.bernh.net/media/download/papers/ldpc.pdf`.