TEORIJA INFORMACIJ
IN IZVORNO KODIRANJE

MATEVŽ KUNAVER

TEORIJA INFORMACIJ IN IZVORNO KODIRANJE

TEORIJA INFORMACIJ IN IZVORNO KODIRANJE

Razloženi primeri in naloge

Matevž Kunaver

Univerza *v Ljubljani*
Fakulteta *za elektrotehniko*

Katoložni zapis o publikaciji (CIP) pripravili v Narodni in univerzitetni knjižnici v Ljubljani

COBISS.SI-ID 247040771

ISBN 978-961-243-480-9 (PDF)

URL: https://fides.fe.uni-lj.si/~matevzk/Ucbeniki/TIIK/TIIK_Kunaver.pdf

Naslov: Teorija informacij in izvorno kodiranje : razloženi primeri in naloge

Recenzenta: pro. dr. Árpád Bűrmen, prof. dr. Marko Meža
Založnik: Založba FE, Ljubljana
Izdajatelj: Fakulteta za elektrotehniko, Ljubljana
Urednik: prof. dr. Sašo Tomažič
Kraj in leto izida: Ljubljana, 2025
1. elektronska izdaja

VSEBINA

ZAHVALA

Vsak začetek je težak in pisanje prvega učbenika ni izjema. Za začetek se zahvaljujem Iztoku, ki mi je pokazal, da je to možno in mi hkrati prihranil veliko veliko dela, ker se je on prvi spopadel z latex predlogo za učbenik. Prav tako gre zahvala vsem sodelavcem, ki so mirno prenesli, da sem med pisanjem izginil z radarja in se ukvarjal samo z učbenikom. Nikakor pa ne bi pred seboj imeli tega učbenika, če ne bi bilo mojih domačih, ki so mi vztrajno dvigali moralo in poskrbeli za to, da sem našel nov zagon, ko sem bil že skoraj obupan. Hvala vam!

NAMESTO UVODA

V tem učbeniku se bomo seznanili z osnovnimi pojmi teorije informacij in izvornega kodiranja, ki so ključni za razumevanje prenosa informacij, obdelave signalov in delovanja komunikacijskih sistemov. Postopoma bomo spoznali matematično in inženirsko osnovo, ki jo potrebujemo za kvantitativno obravnavo informacijskih procesov.

Najprej bomo ponovili in razširili znanje iz verjetnostnega računa, naključnih spremenljivk in njihovih porazdelitev, ter spoznali večkratne spremenljivke in odvisnosti med njimi. Na teh temeljih bomo obravnavali statistične lastnosti signalov, vključno z avtokorelacijo in drugimi pomembnimi koncepti, ki opisujejo časovne in frekvenčne značilnosti procesov.

Nato bomo prešli na jedro teorije informacij, kjer bomo spoznali pojme informacije, entropije, kodiranja in vzajemne informacije. Videli bomo, kako informacijo merimo, kako jo učinkovito predstavimo ter kako jo lahko prenašamo z najmanjšo možno izgubo in redundanco. Poseben poudarek bomo namenili Shannonovim temeljnim rezultatom, ki povezujejo kapaciteto komunikacijskega kanala z razmerjem signal–šum in širino pasu.

V zadnjih poglavjih bomo obravnavali kapaciteto splošnega in Gaussovega kanala ter postopek vzorčenja signalov. Na teh primerih bomo spoznali, kako teoretična znanja uporabimo pri zasnovi in analizi realnih telekomunikacijskih sistemov, radijskih prenosov, digitalnih komunikacij in pri obdelavi ter shranjevanju podatkov.

Vsako poglavje vsebuje sklop vaj, ki nam bodo pomagale utrditi razlago in postopoma pridobiti samostojno obvladovanje obravnavanih metod. Na voljo bodo tudi priporočeni dodatni viri, kjer bomo lahko pridobljeno znanje še poglobili s pomočjo klasične in sodobne literature s področja teorije informacij, statistične obdelave signalov in telekomunikacij.

1. POGLAVJE

OSNOVE VERJETNOSTI

V sodobni informacijski teoriji in izvirnem kodiranju je verjetnostna teorija ključna osnova za razumevanje in modeliranje negotovosti, napak pri prenosu in učinkovitosti kodiranja. Informacijski sistemi se redno soočajo z naključnimi procesi — od prenosa bitov prek šumnega kanala, do stiskanja podatkov s spremenljivo dolžino kod. Razumevanje osnovnih pojmov, kot so verjetnostni prostor, dogodki, neodvisnost in pogojna verjetnost, je zato temeljno za nadaljnje študije entropije, kapacitete kanala in kodirnih strategij.

V tem poglavju bomo na primerih pokazali, kako uporabljamo osnovne koncepte klasične verjetnosti pri analizi enostavnih naključnih sistemov, kot so meti kocke, kvizi in telekomunikacijski prenosi.

Pri reševanju nalog v tem poglavju se bomo seznanil z modeliranjem enostavnih sistemov, analizo izidov ter praktičnim računom verjetnosti, ki tvorijo osnovo za kasnejše poglavje o informacijskih ukrepih in kodiranju z izgubo ali brez nje.

Za dodatno branje priporočamo naslednje vire:

- A. Papoulis, S. U. Pillai: *Probability, Random Variables, and Stochastic Processes* [1]
- S. Haykin: *Communication Systems, 5th ed.* [7]

1.1 Naloge

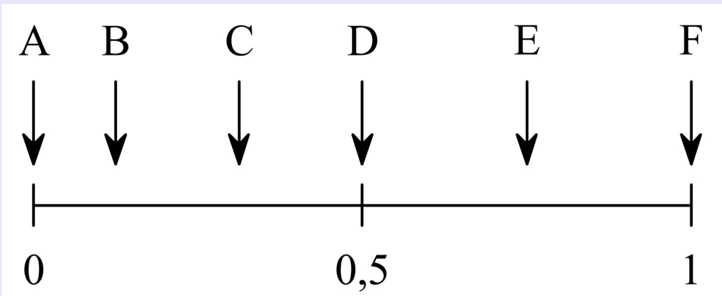
Naloga 1.1: V vreči imamo 15 kroglic različnih barv: 10 modrih in 5 rdečih. V vrečo sežemo z obema rokama in z vsako roko izvlečemo eno kroglico (skupaj izvlečemo dve kroglici) ter si zapomnimo njuni barvi.

- Definiraj naključni dogodek v tem specifičnem primeru.
- Določi vzorčni prostor.
- Izrazi dogodek, da je leva izvlečena kroglica rdeče barve, kot podmnožico vzorčnega prostora.

Rešitev: Naša naloga je torej da najprej opredelimo sistem ki ga opazujemo in nato definiramo posamezne dogodke, ki se lahko zgodijo v njem. Pri tem opozorimo na dejstvo, da ločimo med levo in desno roko, saj je to pomembno za našo analizo.

- Dogodek, ki nas zanima je barva obeh kroglic, ki ju izvlečemo. Pri tem smo pozorni, da sledimo tudi podatku v kateri roki je katera barva.
- Vzorčni prostor je zaloga vseh možnih dogodkov, ki jih v našem sistemu lahko opazimo. Definiramo najprej obe barvi (M za kroglico modre barve in R za rdečo). Dogodek, da smo v levi roki izvlekli rdečo in v desni modro kroglico zabeležimo kot (R, M).
Vse možne dogodke na koncu podamo kot $\{(M, M), (M, R), (R, M), (R, R)\}$
- Podmnožica vzorčnega prostora vsebuje samo tiste dogodke, ki nas zanimajo, torej $\{(R, M), (R, R)\}$

Naloga 1.2: Na sliki so prikazane verjetnosti za dogodke A, B, C, D, E in F.



- Kateri dogodek je gotov?
- Kateri dogodek je najbolj neverjeten, a ne nemogoč?
- Kateri dogodek je nemogoč?
- Kateri dogodki so bolj verjetni kot C?

Rešitev: Čeprav so dogodki A, B, C, D, E in F prikazani skupaj, to še ne pomeni, da pripadajo istemu verjetnostnemu prostoru. To ugotovimo, ker je vsota njihovih verjetnosti večja od 1, kar krši osnovno pravilo verjetnosti. Zato teh dogodkov ne smemo primerjati ali združevati med sabo na enak način, kot bi to storili z dogodki iz istega eksperimenta. Za vsak dogodek lahko tako iz grafične predstavitve (v levo smer se verjetnost manjša, v desno pa veča) ugotovimo v kakšnem odnosu so med seboj.

- Glede na pozicijo na sliki je gotov dogodek F, saj ima 100% verjetnost, da se zgodi.
- Najbolj neverjeten, a ne nemogoč je dogodek B, saj še vedno obstaja majhna možnost, da se zgodi.
- Nemogoč je dogodek A.
- Dogodki, ki so bolj verjetni kot C se na sliki nahajajo desno od njega: D, E in F.

Naloga 1.3: Pošteno igralno kocko vržemo 300 krat.

- Koliko šestic pričakujemo?
- Ali smo lahko presenečeni, če dobimo 55 šestic?

Rešitev: Poštena igralna kocka pomeni, da vsaka cifra pade z enako verjetnostjo $P(X) = \frac{1}{6}$. Vsak met je neodvisen, torej zgodovina metov ne vpliva na naslednjega.

- Verjetnost za posamezno šestico je $P(6) = \frac{1}{6}$, meti so med seboj neodvisni torej lahko pričakujemo $300 \cdot P(6) = 300 \cdot \frac{1}{6} = 50$ šestic.
- Ne.

Naloga 1.4: M&M bonbončki so različnih barv in te različne barve se pojavljajo v različnih razmerjih. Spodnja tabela podaja verjetnosti, da je naključno izbran M&M določene barve, vendar verjetnost za pojavljanje modre barve manjka.

barva	rjava	rdeča	rumena	zelena	oranžna	modra
verjetnost	0,3	0,2	0,2	0,1	0,1	?

a) Kolikšna je manjkajoča verjetnost?

b) Kolikšna je verjetnost da:

- Dobimo rjavega ali rdečega.
- Ne dobimo rumenega.
- Ne dobimo niti oranžnega niti modrega.
- Dobimo rjavega ali rdečega ali rumenega ali zelenega ali oranžnega ali modrega.

Rešitev:

a) Vemo, da mora biti vsota vseh verjetnosti enaka ena. Torej:

$$\begin{aligned} P(moder) &= 1 - \overline{P(moder)} = 1 - \sum P(ostali) \\ &= 1 - (0{,}3 + 0{,}2 + 0{,}2 + 0{,}1 + 0{,}1) = 1 - 0{,}9 = 0{,}1 \end{aligned}$$

Verjetnost, da izvlečemo modri bonbon je torej 0,1.

b) Izračunane verjetnosti:

$$\begin{aligned} P(\text{rjav} \cup \text{rdeč}) &= 0{,}3 + 0{,}2 = 0{,}5 \\ \overline{P(\text{rumen})} &= 1 - P(\text{rumen}) = 1 - 0{,}2 = 0{,}8 \\ 1 - P(\text{oranžen} \cup \text{moder}) &= 1 - (0{,}1 + 0{,}1) = 0{,}8 \\ P(\text{vse}) &= 1 (\text{ker smo pokrili vse možnosti}). \end{aligned}$$

Naloga 1.5: Mečemo pošteno igralno kocko.

- Če je prej padla 5, kakšna je sedaj verjetnost, da sedaj pade 6?
- Trikrat zapored je padla 1. Kolikšna je verjetnost, da sedaj pade 6?
- Kolikšna je verjetnost, da trikrat zapored pade 6?
- Kolikšna je verjetnost, da v treh metih nikoli ne pade 3?
- Kolikšna je verjetnost, da bo v treh metih vsaj enkrat padla 6?

Rešitev:

• Verjetnost, da pade petica ni vezana na prejšnje mete (dogodki so neodvisni) torej je enaka

$$P(5) = \frac{1}{6}$$

• Verjetnost je par tako neodvisna od prejšnjih metov, torej

$$P(6) = \frac{1}{6}$$

• Verjetnost, da trikrat zapored pade 6 je enaka produktu verjetnosti:

$$P(6,6,6) = P(6)P(6)P(6) = \frac{1}{6}\frac{1}{6}\frac{1}{6} = \frac{1}{216}$$

- Verjetnost, da enkrat ne pade 3 $P(\overline{3}) = 1 - P(3) = \frac{5}{6}$. Končna verjetnost je tako

$$P(\overline{333}) = P(\overline{3})P(\overline{3})P(\overline{3}) = \frac{125}{216}$$

- Zadnjo verjetnost lahko izračunamo na dva načina - da upoštevamo vse možnosti ali s pomočjo inverza.

$$\begin{aligned}\text{vse možnosti}: 3(\frac{1}{6}\frac{5}{6}\frac{5}{6}) + 3(\frac{1}{6}\frac{1}{6}\frac{5}{6}) + \frac{1}{6}\frac{1}{6}\frac{1}{6} &= \frac{91}{216}\\ \text{inverz}: P(\text{vsaj ena}) = 1 - P(\text{nobena}) = 1 - \frac{125}{216} &= \frac{91}{216}\end{aligned}$$

Naloga 1.6: Dve osebi trdita, da sta telepata. Naredimo poskus, kjer postavimo vsako osebo v svojo sobo. Osebi A pokažemo niz kart, kjer je vsaka karta z enako verjetnostjo rdeča, modra ali zelena.

Za vsako karto, ki jo pokažemo osebi A, oseba B zabeleži ”telepatsko” barvo.

Kolikšna je verjetnost, da oseba B pravilno ugotovi vsaj eno karto, če jih osebi A pokažemo deset.

Rešitev: Nalogo lažje rešimo, če uporabimo inverzni pristop, torej da najprej izračunamo verjetnost, da oseba B zgreši vse karte. Ko imamo ta podatek znan pa lahko z inverzom dobimo verjetnost za vse ostale možnosti (torej, da ugane eno, dve, tri ali več kart).

$$\begin{aligned}P(\text{prav}) &= \frac{1}{3}\\ P(\text{narobe}) &= \frac{2}{3}\\ P(\text{vse narobe}) &= \big(P(\text{narobe})\big)^{10} = \left(\frac{2}{3}\right)^{10} = 0{,}0173\\ P(\text{vsaj ena}) &= 1 - P(\text{vse narobe}) = 1 - 0{,}0173 = 0{,}9827\end{aligned}$$

Naloga 1.7: Izpit je sestavljen iz vprašanj za obkroževanja s po petimi odgovori. Študent na izpitu ve:

- ker se je za izpit učil, obstaja 75% verjetnost, da pozna pravilen odgovor
- v primeru, da odgovora ne pozna bo ugibal, ker na izpitu ni negativnih točk za napačen odgovor

Kolikšna je torej verjetnost, da bo študent na vprašanje odgovoril pravilno?

Rešitev: Rešitev je kombinacija dveh dogodkov - poznavanje odgovora (A) in ugibanje (B). Pravilni odgovor dobimo v primeru, ko študent pozna odgovor, ker se je učil ali pa vprašanja ne pozna in mu uspe uganiti pravilno izbiro.

$$
\begin{aligned}
P(A) &= 0.75 \\
P(B) &= \frac{1}{5} \\
P(\text{pravilno}) &= P(A) + P(B) \cdot P(\overline{A}) \\
P(\text{pravilno}) &= 0{,}75 + \frac{1}{5} \cdot (1 - 0{,}75) = 0{,}75 + 0{,}05 = 0{,}8
\end{aligned}
$$

Verjetnost, da bo študent na vprašanje odgovoril pravilno je torej $0{,}8$.

Naloga 1.8: Imamo 10 kovancev. Devet jih je "navadnih" (cifra-grb), eden pa je goljufiv (grb-grb).

- Iz žepa vzamemo kovanec. Kolikšna je verjetnost, da je goljufiv?
- Vzamemo kovanec ne da bi ga pogledali ter ga vržemo. Pade grb. Kakšna je verjetnost, da je goljufiv?
- Ko vržemo kovanec pade cifra. Kolikšna je verjetnost, da gre za navadni kovanec?

Rešitev:

- Verjetnost, da vzamemo goljufiv kovanec je $P(G) = \frac{1}{10}$.
- Do verjetnosti, da je kovanec goljufiv pridemo s pomočjo Bayesovega izreka:

$$
\begin{aligned}
P(G|grb) &= \frac{P(grb|G)P(G)}{P(grb|G)P(G) + P(grb|N)P(N)} \\
P(grb|G) &= 1 \quad \text{pri goljufivem kovancu vedno pade grb} \\
P(G) &= \frac{1}{10} \\
P(grb|N) &= \frac{1}{2} \quad \text{pri navadnem kovancu pade grb s 50\% verjetnostjo} \\
P(N) &= \frac{9}{10} \\
P(G|grb) &= 1 \cdot \frac{1}{10} + \frac{1}{2} \cdot \frac{9}{10} = \frac{2}{11}
\end{aligned}
$$

Verjetnost, da je kovanec goljufiv je torej $0{,}1818$.

- Če pade cifra ni možno, da gre za goljufiv kovanec, torej je 100% navadni kovanec.

Naloga 1.9: Mečemo pošteno igralno kocko.

- Kolikšna je verjetnost, da pade liha številka?
- Kolikšna je verjetnost, da pade številka, manjša od 5?
- Ali sta ta dogodka odvisna?
- Ali sta dogodka da pade sodo število in da pade število večje od 3 odvisna?
- Kaj pa dogodka da pade število manjše od 4 in pade število večje od 3?

Rešitev:

- Verjetnost, da pade liha številka je:

$$P(L) = \frac{3}{6} = \frac{1}{2}$$

- Verjetnost, da pade številka manjša od 5 je:

$$P(M) = \frac{4}{6} = \frac{2}{3}$$

- Dogodka sta neodvisna, saj je:

$$P(Liho \cap < 5) = P(Liho)P(< 5)$$

- Dogodka da pade sodo število in da pade število večje od 3 sta neodvisna, saj je:

$$P(Sodo \cap > 3) = P(Sodo)P(> 3)$$

- Dogodka da pade število manjše od 4 in pade število večje od 3 sta odvisna, saj je:

$$P(< 4 \cap > 3) \neq P(< 4)P(> 3)$$

Naloga 1.10: Kupujemo svetilko. Na izbiro imamo tri različne modele. Za vsakega vemo kakšna je verjetnost, da bo svetilka po petih letih še delovala ter kakšna je verjetnost, da bomo ta model dobili.

Model	I	II	III
Verjetnost delovanja po 5 letih	90%	75%	80%
Verjetnost nakupa	20%	30%	50%

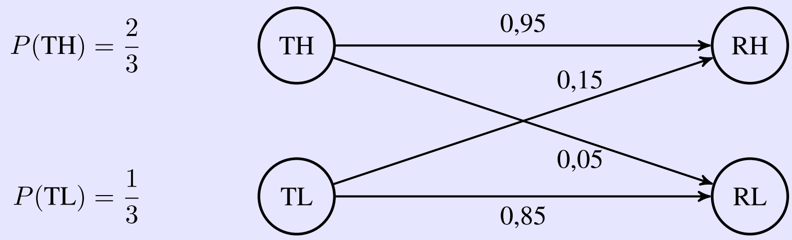
Kolikšna je torej verjetnost, da bo kupljena svetilka po petih letih še delovala.

Rešitev: Verjetnost, da bo svetilka po petih letih še delovala dobimo tako, da izračunamo verjetnost za vsak model posebej in jih seštejemo. Tako dobimo skupno pričakovano verjetnost (tehtano povprečje) delovanja svetilke.

$$\begin{aligned} P(\text{deluje}) &= P(I)P(\text{deluje}|I) + P(II)P(\text{deluje}|II) + P(III)P(\text{deluje}|III) \\ &= 0{,}2 \cdot 0{,}9 + 0{,}3 \cdot 0{,}75 + 0{,}5 \cdot 0{,}8 = 0{,}18 + 0{,}225 + 0{,}4 = 0{,}805 \end{aligned}$$

Verjetnost, da bo kupljena svetilka po petih letih še delovala je torej $0{,}805$.

Naloga 1.11: V binarnem kanalu je visoki nivo H oddajan $\frac{2}{3}$ časa in nizki L preostalo $\frac{1}{3}$. verjetnost za pravilni sprejem H je 0,95, za pravilno sprejet L pa 0,85.



Binarni kanal

Kakšni sta zanesljivosti simbolov?

Rešitev: Zanesljivost simbolov izračunamo tako, da izračunamo verjetnost, da je simbol sprejet pravilno. Torej da je bil sprejet visoki nivo (RH) če je bil oddan visoki nivo (TH) in obratno. Zanesljivosti zato definiramo kot pogojni verjetnosti $P(TH|RH)$ in $P(TL|RL)$.

$$\begin{aligned} P(RH) &= \frac{2}{3} \cdot 0{,}95 + \frac{1}{3} \cdot 0{,}15 = 0{,}686 \\ P(RL) &= \frac{1}{3} \cdot 0{,}85 + \frac{2}{3} \cdot 0{,}05 = 0{,}316 \\ P(TH|RH) &= \frac{P(RH|TH)P(TH)}{P(RH)} = \frac{0{,}95 \cdot \frac{2}{3}}{0{,}6833} = 0{,}927 \\ P(TL|RL) &= \frac{P(RL|TL)P(TL)}{P(RL)} = \frac{0{,}85 \cdot \frac{1}{3}}{0{,}3167} = 0{,}893 \end{aligned}$$

Zanesljivost simbolov je torej $P(TH|RH) = 0{,}927$ in $P(TL|RL) = 0{,}893$.

Naloga 1.12: Na vhodu imamo niz simbolov:

$$X[n]\epsilon\{x_1, x_2\} = \{0,1\};\ P(x_1) = 0{,}4\ P(x_2) = 0{,}6$$

Verjetnost napake pri prenosu po kanalu je $P_n = 0{,}1$.

- Kolikšna je verjetnostna funkcija izhodnega niz $y[n]$
- Kako vpliva kanal na verjetnost izhodnih simbolov?
- Kaj pa če bi bili verjetnosti simbolov enaki?

Rešitev:

- Verjetnostna funkcija izhodnega niza $y[n]$ je odvisna od verjetnosti napake pri prenosu po kanalu.

$$\begin{aligned} P(y_1) &= P(y_1|x_1)P(x_1) + P(y_1|x_2)P(x_2) \\ &= 0{,}9 \cdot 0{,}4 + 0{,}1 \cdot 0{,}6 = 0{,}42 \\ P(y_2) &= P(y_2|x_2)P(x_2) + P(y_2|x_1)P(x_1) \\ &= 0{,}9 \cdot 0{,}6 + 0{,}1 \cdot 0{,}4 = 0{,}58 \end{aligned}$$

- Kanal vpliva na verjetnost izhodnih simbolov, saj se verjetnosti spreminjajo glede na to ali je prišlo do napake ali ne.

- Če bi bile verjetnosti simbolov enake kanal ne bi imel vpliva na verjetnost izhodnih simbolov.

Povzetek ključnih enačb

$P(A) = \frac{ugodniizidi}{vsiizidi}$	Klasična definicija verjetnosti: razmerje med ugodnimi in vsemi možnimi izidi
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$	Verjetnost unije dveh dogodkov
$P(A \cap B) = P(A\|B) \cdot P(B)$	Verjetnost preseka dveh dogodkov preko pogojne verjetnosti
$P[X] = \sum_i x_i \cdot P(x_i)$	Pričakovana vrednost diskretne slučajne spremenljivke
$P[X] = \int_{-\infty}^{\infty} x \cdot p(x)\, dx$	Pričakovana vrednost zvezne slučajne spremenljivke
$P(A \cap B) = P(A) \cdot P(B)$	Če sta A in B neodvisna dogodka
$P(\bar{A}) = 1 - P(A)$	Verjetnost komplementarnega dogodka
$P(A\|B) = \dfrac{P(A \cap B)}{P(B)}$	Pogojna verjetnost

2. POGLAVJE

NAKLJUČNE SPREMENLJIVKE

V tem poglavju bomo na konkretnih nalogah prikazali uporabo naključnih spremenljivk za modeliranje in analizo meritev, napak in izidov v informacijskih sistemih. Naključna spremenljivka nam omogoča, da kvantitativno opišemo izide naključnih poskusov — naj bodo to meritve napetosti v komunikacijskem kanalu, število napak pri prenosu ali rezultat kviza.

Razumeli bomo razlike med *diskretnimi* in *zveznimi* naključnimi spremenljivkami ter načine, kako opisujemo njihove verjetnostne porazdelitve: s funkcijo verjetnosti $P(x)$ za diskretne in z gostoto verjetnosti $p(x)$ za zvezne.

Nadalje bomo definirali pričakovano vrednost naključne spremenljivke X z oznako $P(X)$ in preko nje razvili ključne koncepte, kot so *varianca*, *standardni odklon* in *funkcija porazdelitve*. Posebej nas bodo zanimale lastnosti linearnih transformacij naključnih spremenljivk, simetrije, enakomernih in normalnih porazdelitev ter uporaba teh znanj pri analizi in detekciji signalov.

Naučili se bomo tudi, kako z uporabo tabel (ali izračunov) ocenimo verjetnosti za standardno normalno porazdeljene spremenljivke, kar je pogosto v aplikacijah, kjer predpostavljamo prisotnost šuma (npr. digitalni komunikacijski sistemi).

Poglavje bo ključno za nadaljnje razumevanje entropije, kanalov, ter kodiranja z napako ali brez nje. Osvojena znanja bomo uporabljali pri ocenjevanju zanesljivosti prenosa, dimenzioniranju kanalov in pri analizi porazdelitev, ki vplivajo na učinkovitost kodirnih algoritmov.

Za dodatno branje priporočamo naslednje vire:

- A. Papoulis, S. U. Pillai: *Probability, Random Variables, and Stochastic Processes* [1]
- S. Tomažič: *Osnove telekomunikacij I* [5]

2.1 Naloge

Naloga 2.1: Zgodovina orkanov v ZDA je podana v tabeli.

x_i	0	1	2	3	4	5	6	7
$P_x[i]$	0,05	0,15	0,22	0,26	0,14	0,08	0,07	0,03

- Kolikšno je pričakovano število orkanov?
- Kakšna je standardna deviacija?

Rešitev: Uporabimo diskretno porazdelitev za oceno pričakovanega števila orkanov ter standardne deviacije. Začnemo z izračunom pričakovane vrednosti:

$$\overline{X} = \sum_{i=0}^{7} i \cdot P_x[i] = 0 \cdot 0{,}05 + 1 \cdot 0{,}15 + 2 \cdot 0{,}22 + 3 \cdot 0{,}26 + 4 \cdot 0{,}14$$
$$+ 5 \cdot 0{,}08 + 6 \cdot 0{,}07 + 7 \cdot 0{,}03 = 2{,}96$$

Pričakovano število orkanov je 2,96.
Standardna deviacija je kvadratni koren variance. Ti izračuni so uporabni, kadar želimo razumeti variabilnost pojavov v okolju — tukaj naravnih nesreč.

$$\sigma_X^2 = D(X) = \sum_{i=0}^{7} P_X[i] \cdot (x_i - \overline{X})^2 = \overline{X^2} - \overline{X}^2$$

$$\overline{X^2} = \sum_{i=0}^{7} P_X[i] \cdot (x_i)^2$$

$$\overline{X^2} = 0{,}05 \cdot 0^2 + 0{,}15 \cdot 1^2 + 0{,}22 \cdot 2^2 + 0{,}26 \cdot 3^2 + 0{,}14 \cdot 4^2$$
$$+ 0{,}08 \cdot 5^2 + 0{,}07 \cdot 6^2 + 0{,}03 \cdot 7^2$$

$$\overline{X^2} = 11{,}6$$

$$\sigma_X^2 = 11{,}6 - 2{,}96^2 = 2{,}8384$$

$$\sigma_X = 1{,}685$$

ali

$$\sigma_X^2 = D(X) = \sum_{i=0}^{7} P_X[i] \cdot (x_i - \overline{X})^2$$

$$\begin{aligned}
\sigma_X^2 &= 0{,}05 \cdot (0 - 2{,}96)^2 + 0{,}15 \cdot (1 - 2{,}96)^2 + 0{,}22 \cdot (2 - 2{,}96)^2 \\
&\quad + 0{,}26 \cdot (3 - 2{,}96)^2 + 0{,}14 \cdot (4 - 2{,}96)^2 + 0{,}08 \cdot (5 - 2{,}96)^2 \\
&\quad + 0{,}07 \cdot (6 - 2{,}96)^2 + 0{,}03 \cdot (7 - 2{,}96)^2 \\
\sigma_X^2 &= 2{,}8384 \\
\sigma_X &= 1{,}685
\end{aligned}$$

Standardna deviacija je $1{,}685$.

Naloga 2.2: Predpostavimo, da je X zvezna naključna spremenljivka z gostoto verjetnosti

$$p(x) = \begin{cases} ce^{(3-x)}, & \text{za } 3 \leq x \leq 6 \\ 0, & \text{drugje} \end{cases}$$

Določi

- Vrednost konstante c
- $P(x \leq 5)$
- $P(x \leq 3)$
- $P(4 \leq x \leq 5.5)$
- $P(4 \leq x \leq 7)$
- $P(x \leq 5 \mid x \geq 4)$

Rešitev: Veljati mora:

$$\int_{-\infty}^{\infty} p_X(x)\, dx = 1$$

Torej integral razbijemo na tri dele:

$$\begin{aligned}
\int_{-\infty}^{3} 0\, dx + \int_{3}^{6} ce^{(3-x)}\, dx + \int_{6}^{\infty} 0\, dx &= -ce^{(3-x)}\Big|_3^6 \\
&= -c\left(e^{3-6} - e^{3-3}\right) = c\left(1 - e^{-3}\right) \\
&= c \cdot 0{,}95
\end{aligned}$$

Iz tega sledi, da mora biti konstanta c enaka:

$$c = \frac{1}{0{,}95} = 1{,}0524$$

Sedaj lahko izračunamo ostale verjetnosti, v vsakem primeru s pomočjo integrala:

Verjetnost $P(x \leq 5)$ je $0{,}91$.

$$\begin{aligned} P(X \leq 5) &= \int_{-\infty}^{5} p_X(x)\,dx = 1{,}05 \cdot \int_{3}^{5} e^{(3-x)}\,dx = -1{,}05 \cdot e^{(3-x)}\Big|_3^5 \\ &= 1{,}05\left(1 - e^{-2}\right) = 0{,}91 \end{aligned}$$

Verjetnost $P(x \leq 3)$ je 0.

$$P(X \leq 3) = \int_{-\infty}^{3} p_X(x)\,dx = 1{,}05 \cdot \int_{3}^{3} e^{(3-x)}\,dx = 0$$

Verjetnost $P(4 \leq x \leq 5.5)$ je $0{,}30$.

$$\begin{aligned} P(4 \leq X \leq 5.5) &= P(X \leq 5.5) - P(X < 4) \\ &= 1{,}05\left(1 - e^{3-5.5}\right) - 1{,}05\left(1 - e^{3-4}\right) \\ &= 0{,}30 \end{aligned}$$

Verjetnost $P(4 \leq x \leq 7)$ je $0{,}335$.

$$\begin{aligned} P(4 \leq X \leq 7) &= P(X \leq 7) - P(X < 4) \\ &= 1{,}05\left(1 - e^{3-6}\right) - 1{,}05\left(1 - e^{3-4}\right) \\ &= 0{,}335 \end{aligned}$$

Verjetnost $P(x \leq 5 \mid x \geq 4)$ je $0{,}73$.

$$\begin{aligned} P(X \leq 5 \mid X \geq 4) &= \frac{P(4 \leq X \leq 5)}{P(X \geq 4)} = \frac{P(X \leq 5) - P(X < 4)}{1 - P(X < 4)} \\ P(X \leq 5 \mid X \geq 4) &= \frac{1{,}05\left(1 - e^{3-5}\right) - 1{,}05\left(1 - e^{3-4}\right)}{1 - 1{,}05\,(1 - e^{3-4})} \\ &= 0{,}73 \end{aligned}$$

Naloga 2.3: X je zvezna naključna spremenljivka s porazdelitvijo verjetnosti:

$$p(x) = \begin{cases} 3x^2, & 0 \leq x \leq 1 \\ 0, & \text{drugje} \end{cases}$$

- Ali je $p(x)$ lahko verjetnostna porazdelitev?
- Določi kumulativno verjetnostno porazdelitev.
- Izračunaj povprečno vrednost spremenljivke.
- Izračunaj varianco naključne spremenljivke.

Rešitev: Če je $p(x)$ verjetnostna porazdelitev, mora veljati:

$$\int_{-\infty}^{\infty} p_X(x)\, dx = 1$$
$$\int_0^1 3x^2\, dx = 3 \cdot \left.\frac{x^3}{3}\right|_0^1 = 1$$

Torej je $p(x)$ verjetnostna porazdelitev.
Kumulativna verjetnostna porazdelitev je definirana kot:

$$\begin{aligned} F_X(x) &= \int_{-\infty}^{x} p_X(t)\, dt \\ x < 0: \quad & F_X(x) = 0 \\ 0 \leq x \leq 1: \quad & F_X(x) = \int_0^x 3t^2\, dt = 3 \cdot \left.\frac{t^3}{3}\right|_0^x = x^3 - 0^3 = x^3 \\ x > 1: \quad & F_X(x) = 1 \end{aligned}$$

Za povprečno vrednost spremenljivke rešimo integral:

$$\begin{aligned} \overline{X} &= \int_{-\infty}^{\infty} x \cdot p_X(x)\, dx = \int_0^1 x \cdot 3x^2\, dx = \int_0^1 3x^3\, dx \\ &= 3 \cdot \left.\frac{x^4}{4}\right|_0^1 = \frac{3}{4} = 0{,}75 \end{aligned}$$

Povprečna vrednost spremenljivke je torej $\overline{X} = 0{,}75$.

Za varianco spremenljivke uporabimo formulo:

$$D(X) = \int_{-\infty}^{\infty} (x - \overline{X})^2 \cdot p_X(x)\, dx = \int_0^1 (x - 0{,}75)^2 \cdot 3x^2\, dx$$
$$= \int_0^1 \big(x^2 - 1{,}5x + 0{,}5625\big) \cdot 3x^2\, dx$$
$$D(X) = \int_0^1 3x^4\, dx - 1{,}5 \int_0^1 3x^3\, dx + 0{,}5625 \int_0^1 3x^2\, dx$$
$$= 3 \cdot \frac{x^5}{5}\Big|_0^1 - 4{,}5 \cdot \frac{x^4}{4}\Big|_0^1 + 0{,}5625 \cdot 3 \cdot \frac{x^3}{3}\Big|_0^1$$
$$D(X) = \frac{3}{5} - \frac{4{,}5}{4} + 0{,}5625 = 0{,}0375$$

Varianca naključne spremenljivke X je torej $D(X) = 0{,}0375$.

Naloga 2.4: Naključna spremenljivka X s srednjo vrednostjo 2 ima standardno deviacijo 1. Določi srednjo kvadratno vrednost $\overline{X^2}$.

Rešitev: Srednja kvadratna vrednost spremenljivke je $\overline{X^2} = 5$, kar dobimo neposredno iz srednje vrednosti in standardne deviacije po sledeči formuli:

$$\overline{X^2} = \overline{X}^2 + \sigma_X^2 = 2^2 + 1^2 = 5$$

Naloga 2.5: Debelina fotorezista, ki ga nanesemo na silicijeve rezine pri proizvodnji polprevodnikov je enakomerno porazdeljena med 0,202 in 0,215 mikrometri.

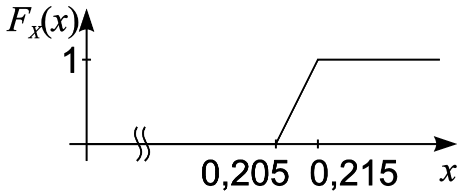
- Določi kumulativno porazdelitev F_x.
- Določi delež rezin, katerih debelina presega 0,2125 mikrometra.
- Katero debelino fotorezistorja presega zgolj 10% vseh rezin?
- Določite povprečno debelino in varianco debeline.

Rešitev: Enakomerna porazdelitev pomeni, da je verjetnost na celem intervalu konstantna. Ker mora biti integral čez celoten interval enak ena, izračunamo konstantno verjetnost kot:

$$p = \frac{1}{\Delta x} = \frac{1}{0{,}01} = 100$$

Kumulativna verjetnost je integral gostote verjetnosti od $-\infty$ do izbrane vrednosti naključne spremenljivke:

$$
\begin{aligned}
x < 0{,}205: &\quad F_X(x) = 0 \\
0{,}205 \le x \le 0{,}215: &\quad F_X(x) = \int_{0{,}205}^{x} p_X(t)\,dt = \int_{0{,}205}^{x} p \cdot dt = 100 \cdot t\Big|_{0{,}205}^{x} \\
&\quad = 100(x - 0{,}205) = 100x - 20{,}5 \\
x > 0{,}215: &\quad F_X(x) = 1
\end{aligned}
$$



Da določimo delež rezin, katerih debelina presega 0,2125 mikrometra, uporabimo kumulativno porazdelitev in inverzno logiko:

$$
\begin{aligned}
P_X(x > 0{,}2125) &= 1 - F_X(0{,}2125) = 1 - (100 \cdot 0{,}2125 - 20{,}5) \\
&= 1 - 0{,}75 = 0{,}25
\end{aligned}
$$

Na podoben način določimo debelino, ki presega zgolj 10% vseh rezin:

$$
\begin{aligned}
P_X(X > x_{10}) = 0{,}1 = 1 - F_X(x_{10}) &\Rightarrow F_X(x_{10}) = 0{,}9 \\
100 \cdot x_{10} - 20{,}5 &= 0{,}9 \\
x_{10} &= \frac{0{,}9 + 20{,}5}{100} = 0{,}214
\end{aligned}
$$

Povprečno debelino in varianco izračunamo z uporabo definicij:

$$
\begin{aligned}
\overline{X} &= \int_{-\infty}^{\infty} x \cdot p_X(x)\,dx = \int_{0{,}205}^{0{,}215} x \cdot \frac{1}{0{,}01}\,dx = 100 \cdot \int_{0{,}205}^{0{,}215} x\,dx \\
&= 100 \cdot \frac{x^2}{2}\Big|_{0{,}205}^{0{,}215} = 50 \cdot (0{,}046225 - 0{,}042025) = 0{,}210 \\
D(X) &= \int_{-\infty}^{\infty} (x - \overline{X})^2\, p_X(x)\,dx = \int_{0{,}205}^{0{,}215} (x - 0{,}21)^2 \cdot 100\,dx \\
&= 100 \cdot \int_{0{,}205}^{0{,}215} (x^2 - 0{,}42x + 0{,}0441)\,dx \\
D(X) &= 100 \cdot \left(\frac{x^3}{3}\Big|_{0{,}205}^{0{,}215} - 0{,}42 \cdot \frac{x^2}{2}\Big|_{0{,}205}^{0{,}215} + 4{,}41 \cdot x\Big|_{0{,}205}^{0{,}215} \right) \\
&= 0{,}04410833 - 0{,}0882 + 0{,}0441 = 8{,}33 \cdot 10^{-6}
\end{aligned}
$$

Naloga 2.6: Naj bo X normalno porazdeljenba zvezna naključna spremenljivka s srednjo vrednostjo 0 in standardno deviacijo 1. Določi:

- $P(x \leq 1{,}4)$
- $P(x \geq -1{,}3)$

Rešitev: Za normalno porazdelitev pogosto ne računamo verjetnosti analitično, temveč s pomočjo tabele. Standardna normalna porazdelitev $\mathcal{N}(0,1)$ omogoča iskanje:

$$P(X \leq x) = \Phi(x)$$

kjer je $\Phi(x)$ vrednost iz tabele. Dodatno si pomagamo s sledečimi pomožnimi enačbami:

$$\begin{aligned} P(X \geq x) &= 1 - P(X \leq x) \\ P(X \leq -x) &= P(X \geq x) \\ P(X \geq -x) &= P(X \leq x) \end{aligned}$$

Za naš primer torej dobimo:

$$\begin{aligned} P(X \leq 1,4) &= \Phi(1,4) = 0,91924 \\ P(X \geq -1,3) &= \Phi(1,3) = 0,90320 \end{aligned}$$

Naloga 2.7: Na digitalni liniji je prisoten šum z normalno porazdelitvijo amplitude, srednjo vrednostjo 0 in standardno deviacijo 0,45V.

Sistem zazna 1, če napetost presega 0,9V. Kolikšna je verjetnost, da sistem zazna 1, če na liniji ni signala?

Rešitev: Pozorni moramo biti na to, da je standardna deviacija različna od 1, kar pomeni, da moramo prilagoditi izračun normalne porazdelitve. Vse tabele v matematičnem priročniku so namreč narejene za standardno normalno porazdelitev, kjer je standardna deviacija 1. Potrebno je torej normalizirati vrednost, da jo lahko potem preberemo iz tabele.
Vrednosti v tabeli nam namreč povejo "kolikokrat moram šum preseči vrednost standardne deviacije". Torej $\Phi(0,4)$ pomeni, da je šum presegel 0,4 standardne deviacije. Povedano drugače - v primerih ko je standardna deviacija različna od 1 uporabimo sledečo enačbo:

$$\Phi\left(\frac{\text{nivo ki me zanima}}{\text{standardna deviacija}}\right)$$

V našem primeru mora torej šum preseči 0,9V, kar je dvakratna vrednost standardne deviacije:

$$x\sigma = 0,9V \quad \Rightarrow \quad x = \frac{0,9V}{\sigma} = \frac{0,9V}{0,45V} = 2$$

$$P_X(X > 0,9) = 1 - P_X(X \leq 0,9) = 1 - P_X\left(\frac{X}{\sigma} \leq \frac{0,9}{\sigma}\right) = 1 - \Phi\left(\frac{0,9}{0,45}\right)$$

$$= 1 - \Phi(2) = 1 - 0,97725$$

$$P_X(X > 0,9) = 0,02275$$

Verjetnost, da sistem zazna 1, če na liniji ni signala, je torej približno 0,02275 ali 2,275%.

Povzetek ključnih enačb

$\bar{X} = \sum x_i \cdot P(x_i)$	Pričakovana vrednost diskretne spremenljivke
$\bar{X} = \int_{-\infty}^{\infty} x \cdot p(x)\,dx$	Pričakovana vrednost zvezne spremenljivke
$\sigma_X^2 = \sum P(x_i) \cdot (x_i - \bar{X})^2$	Varianca – kvadratni odmik od povprečja
$\sigma_X = \sqrt{\sigma_X^2}$	Standardna deviacija – raztros v istih enotah kot X
$P(x \leq A) = \int_{-\infty}^{A} p(x)\,dx$	Verjetnost, da je X manjša ali enaka A (zvezne spremenljivke)
$F(x) = P(X \leq x)$	Kumulativna funkcija porazdelitve

3. POGLAVJE

VEČKRATNE NAKLJUČNE SPREMENLJIVKE

V tem poglavju bomo obravnavali naloge, kjer nastopata dve ali več medsebojno odvisnih naključnih spremenljivk — na primer v prisotnosti šuma ali sočasnem prenosu informacij. Tak pristop je nujen pri analizi sistemov, kjer spremenljivke niso neodvisne, ampak so med seboj statistično povezane — na primer meritev signala v prisotnosti šuma, temperaturni in napetostni profil v sistemu, ali sočasno kodiranje večih informacijskih tokov.

Bistvena pojma, ki ju bomo obravnavali, sta skupna porazdelitev in pogojna porazdelitev. Skupna porazdelitev $P_{XY}(x, y)$ nam pove, s kakšno verjetnostjo se hkrati pojavita vrednosti x in y spremenljivk X in Y. Pogojna porazdelitev $P(Y|X)$ pa modelira verjetnost za vrednosti Y, če je X že znan.

Spoznali bomo tudi koncept neodvisnosti, ki v večrazsežnem prostoru pomeni, da skupna porazdelitev razpade na produkt marginalnih:

$$P_{XY}(x, y) = P_X(x) \cdot P_Y(y)$$

Pri zveznih spremenljivkah bomo delali z gostoto verjetnosti $p(x, y)$, kjer mora integral po celotnem območju enak 1. Pogojne gostote bomo izračunali s pomočjo:

$$p(y \mid x) = \frac{p(x, y)}{p_X(x)}$$

Dodatno bomo analizirali kovarianco in korelacijo, ki podata informacijo o tem, ali se vrednosti spremenljivk gibajo skupaj (in če, kako močno). Kadar je korelacijski koeficient enak 1 ali -1, imamo popolnoma linearno povezanost med spremenljivkami.

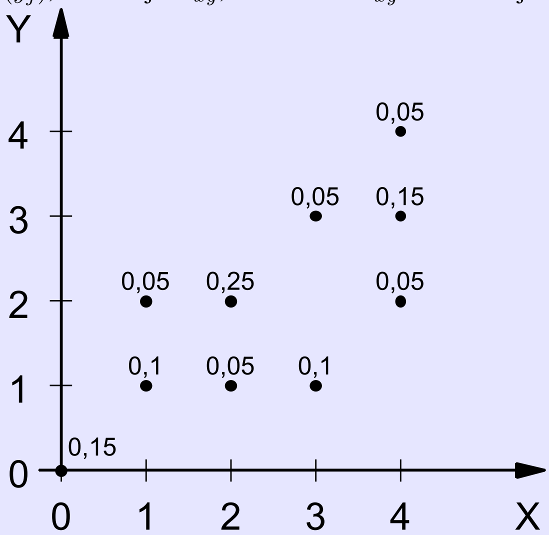
Ti koncepti so temeljni v teoriji informacij, še posebej pri kodiranju povezanih signalov in ocenjevanju napak pri prenosu informacij.

Za dodatno branje priporočamo naslednje vire:

- A. Papoulis, S. U. Pillai: *Probability, Random Variables, and Stochastic Processes* [1]
- N. Pavešič: *Informacija in kodi* [6]

3.1 Naloge

Naloga 3.1: Za dve spremenljivki X in Y določi marginalni verjetnostni funkciji $P_x(x_i), P_y(y_j)$, korelacijo R_{xy}, kovarianco C_{xy} in korelacijski koeficient.



Rešitev: Ker imamo skupno porazdelitev $P_{XY}(x_i, y_j)$, lahko marginalne porazdelitve dobimo s seštevanjem po drugi spremenljivki:

$$P_X(x_i) = \sum_j P_{XY}(x_i, y_j), \quad P_Y(y_j) = \sum_i P_{XY}(x_i, y_j)$$

Korelacija je definirana kot:

$$R_{XY} = \sum_i \sum_j x_i y_j P_{XY}(x_i, y_j)$$

Kovarianca:

$$C_{XY} = R_{XY} - \overline{X} \cdot \overline{Y}$$

Korelacijski koeficient:

$$\hat{\rho}_{XY} = \frac{C_{XY}}{\sigma_X \cdot \sigma_Y}$$

Ta zadnji pove, kako linearno sta spremenljivki povezani. Če je $\hat{\rho}_{XY} = 0$, sta linearno neodvisni.
Marginalni porazdelitvi sta torej:

$$P_X(x_i) = \sum_j P_{XY}(x_i, y_j) \qquad P_Y(y_j) = \sum_i P_{XY}(x_i, y_j)$$

$$\begin{aligned} P_X(0) &= 0,15 \\ P_X(1) &= 0,1 + 0,05 = 0,15 \\ P_X(2) &= 0,05 + 0,25 = 0,3 \\ P_X(3) &= 0,1 + 0,05 = 0,15 \\ P_X(4) &= 0,05 + 0,15 + 0,05 = 0,25 \end{aligned} \qquad \begin{aligned} P_Y(0) &= 0,15 \\ P_Y(1) &= 0,1 + 0,05 + 0,1 = 0,25 \\ P_Y(2) &= 0,05 + 0,25 + 0,05 = 0,35 \\ P_Y(3) &= 0,05 + 0,15 = 0,2 \\ P_Y(4) &= 0,05 \end{aligned}$$

Korelacija je tako:

$$\begin{aligned} R_{XY} &= \sum_i \sum_j x_i y_j P_{XY}(x_i, y_j) \\ R_{XY} &= 0 \cdot 0 \cdot 0,15 \\ &\quad + 1 \cdot (1 \cdot 0,1 + 2 \cdot 0,05) \\ &\quad + 2 \cdot (1 \cdot 0,05 + 2 \cdot 0,25) \\ &\quad + 3 \cdot (1 \cdot 0,1 + 3 \cdot 0,05) \\ &\quad + 4 \cdot (2 \cdot 0,05 + 3 \cdot 0,15 + 4 \cdot 0,05) \\ R_{XY} &= 5,05 \end{aligned}$$

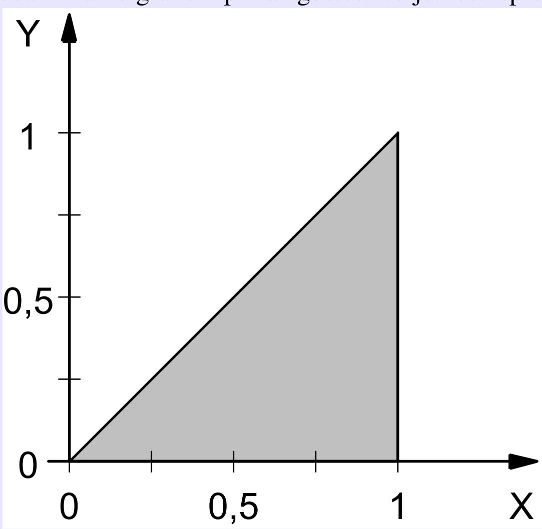
Ko poznamo korelacijo lahko izračunamo kovarianco:

$$\begin{aligned} C_{XY} &= R_{XY} - \overline{X} \cdot \overline{Y} \\ \overline{X} &= \sum_i x_i P_X(x_i), \quad \overline{Y} = \sum_j y_j P_Y(y_j) \\ \overline{X} &= 0 \cdot 0,15 + 1 \cdot 0,15 + 2 \cdot 0,3 + 3 \cdot 0,15 + 4 \cdot 0,25 = 2,2 \\ \overline{Y} &= 0 \cdot 0,15 + 1 \cdot 0,25 + 2 \cdot 0,35 + 3 \cdot 0,2 + 4 \cdot 0,05 = 1,75 \\ C_{XY} &= 5,05 - 2,2 \cdot 1,75 = 1,2 \end{aligned}$$

In na koncu še korelacijski koeficient:

$$\rho_{XY} = \frac{C_{XY}}{\sigma_X \cdot \sigma_Y}$$
$$\sigma_X = \sqrt{\overline{X^2} - \overline{X}^2} = \sqrt{\sum_i x_i^2 P_X(x_i) - \overline{X}^2},$$
$$\sigma_Y = \sqrt{\overline{Y^2} - \overline{Y}^2} = \sqrt{\sum_j y_j^2 P_Y(y_j) - \overline{Y}^2}$$
$$\sigma_X = \sqrt{1^2 \cdot 0,15 + 2^2 \cdot 0,3 + 3^2 \cdot 0,15 + 4^2 \cdot 0,25 - 2,2^2} = 1,3638$$
$$\sigma_Y = \sqrt{1^2 \cdot 0,25 + 2^2 \cdot 0,35 + 3^2 \cdot 0,2 + 4^2 \cdot 0,05 - 1,75^2} = 1,0897$$
$$\rho_{XY} = \frac{1,2}{1,3638 \cdot 1,0897} = 0,807$$

Naloga 3.2: Podan imamo grafični prikaz gostote verjetnosti spremenljivk x in y.



- Določi marginalno porazdelitev komponent.
- Ali sta spremenljivki korelirani?
- Izračunaj gostoto verjetnosti $p_y(y|x)$.

Rešitev: Ker imamo sedaj opravka z zveznimi spremenljivkami, se naše vsote spremenijo v integrale. Najprej moramo določiti gostoto verjetnosti na področju, kjer je različna od 0. Integral po celotnem definicijskem območju (površini x-y) mora biti enak 1. Ker je gostota verjetnosti konstantna, lahko to napravimo grafično (iz slike vidimo, da je površina enaka 1/2):

$$K \cdot \frac{1}{2} = 1 \quad \Rightarrow \quad K = 2$$

ali pa matematično:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{XY}(x,y)\,dx\,dy = 1$$

$$\begin{aligned}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{XY}(x,y)\,dy\,dx &= \int_0^1 \left(\int_0^x K\,dy\right) dx = \int_0^1 (Ky\big|_0^x)dx = \int_0^1 Kx\,dx \\ &= K\frac{x^2}{2}\bigg|_0^1 = K\frac{1}{2}\end{aligned}$$

ali:

$$\begin{aligned}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{XY}(x,y)\,dx\,dy &= \int_0^1 \left(\int_y^1 K\,dx\right) dy = \int_0^1 \left(Kx\big|_y^1\right) dy \\ &= \int_0^1 K(1-y)\,dy = K\left(y\big|_0^1 - \frac{y^2}{2}\bigg|_0^1\right) = K\frac{1}{2}\end{aligned}$$

$$K \cdot \frac{1}{2} = 1 \quad \Rightarrow \quad K = 2$$

Marginalni porazdelitvi sta:

$$\begin{aligned} p_X(x) &= \int_{-\infty}^{\infty} p_{XY}(x,y)\,dy, \quad p_Y(y) = \int_{-\infty}^{\infty} p_{XY}(x,y)\,dx \\ x \le 0: &\quad p_X(x) = 0 \\ 0 < x \le 1: &\quad p_X(x) = \int_0^x K\,dy = Ky\big|_0^x = Kx = 2x \\ x > 1: &\quad p_X(x) = 0 \\ y < 0: &\quad p_Y(y) = 0 \\ 0 \le y \le 1: &\quad p_Y(y) = \int_y^1 K\,dx = Kx\big|_y^1 = K(1-y) = 2(1-y) \\ y > 1: &\quad p_Y(y) = 0 \end{aligned}$$

Naključni spremenljivki sta korelirani, če je kovarianca C_{XY} različna od nič, oziroma, če je korelacija različna od produkta povprečnih vrednosti spremenljivk $\overline{X}$ in $\overline{Y}$.

$$R_{XY} = C_{XY} + \overline{X} \cdot \overline{Y}$$

$$C_{XY} \neq 0 \Rightarrow R_{XY} \neq \overline{X} \cdot \overline{Y}$$

$$R_{XY} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyp_{XY}(x,y)\,dx\,dy$$

$$R_{XY} = \int_0^1 \int_0^x xyK\,dy\,dx = \int_0^1 xK \left(\int_0^x y\,dy \right) dx$$

$$= \int_0^1 xK \left(\frac{y^2}{2}\Big|_0^x \right) dx = \int_0^1 xK\frac{x^2}{2}\,dx = \frac{K}{2}\int_0^1 x^3\,dx$$

$$R_{XY} = \frac{K}{2}\left(\frac{x^4}{4}\Big|_0^1 \right) = \frac{K}{2} \cdot \frac{1}{4} = \frac{1}{4}$$

$$\overline{X} = \int_{-\infty}^{\infty} xp_X(x)\,dx = \int_0^1 x2x\,dx = 2\int_0^1 x^2\,dx = 2\left(\frac{x^3}{3}\Big|_0^1 \right) = \frac{2}{3}$$

$$\overline{Y} = \int_{-\infty}^{\infty} yp_Y(y)\,dy = \int_0^1 y2(1-y)\,dy$$

$$= 2\int_0^1 \left(\frac{y^2}{2} - \frac{y^3}{3} \right)\Big|_0^1 = 2\left(\frac{1}{2} - \frac{1}{3} \right) = \frac{1}{3}$$

$$\overline{X} \cdot \overline{Y} = \frac{2}{3} \cdot \frac{1}{3} = \frac{2}{9} \neq R_{XY} \quad \Rightarrow \quad \text{Naključni spremenljivki sta korelirani.}$$

Korelacijski koeficient $\rho_{XY} = \frac{C_{XY}}{\sigma_X \sigma_Y}$

$$\overline{X^2} = \int_{-\infty}^{\infty} x^2 p_X(x)\,dx = \int_0^1 x^2 Kx\,dx = K\frac{x^4}{4}\Big|_0^1 = \frac{K}{4} = \frac{1}{2}$$

$$\overline{Y^2} = \int_{-\infty}^{\infty} y^2 p_Y(y)\,dy = \int_0^1 y^2 K(1-y)\,dy$$

$$= K\left(\frac{y^3}{3}\Big|_0^1 - \frac{y^4}{4}\Big|_0^1 \right) = K\left(\frac{1}{3} - \frac{1}{4} \right) = K\left(\frac{1}{12} \right) = \frac{1}{6}$$

$$\sigma_X = \sqrt{\overline{X^2} - \overline{X}^2} = \sqrt{\frac{1}{2} - \left(\frac{2}{3} \right)^2} = \sqrt{\frac{1}{2} - \frac{4}{9}} = \frac{1}{\sqrt{18}} = \frac{\sqrt{2}}{6}$$

$$\sigma_Y = \sqrt{\overline{Y^2} - \overline{Y}^2} = \sqrt{\frac{1}{6} - \left(\frac{1}{3} \right)^2} = \sqrt{\frac{1}{6} - \frac{1}{9}} = \frac{1}{\sqrt{18}} = \frac{\sqrt{2}}{6}$$

$$C_{XY} = R_{XY} - \overline{X} \cdot \overline{Y} = \frac{1}{4} - \frac{2}{3} \cdot \frac{1}{3} = \frac{9-8}{36} = \frac{1}{36}$$

$$\rho_{XY} = \frac{C_{XY}}{\sigma_X \sigma_Y} = \frac{\frac{1}{36}}{\frac{\sqrt{2}}{6} \cdot \frac{\sqrt{2}}{6}} = \frac{\frac{1}{36}}{\frac{2}{36}} = \frac{1}{2}$$

Pogojna gostota verjetnosti $p_y(y|x)$

$$p_Y(y|x) = \frac{p_{XY}(x,y)}{p_X(x)}$$

$$x \leq 0 \cup x > 1: \quad p_{XY}(x,y) = 0, \quad p_X(x) = 0 \quad \Rightarrow \quad p_Y(y|x) = \frac{0}{0}$$

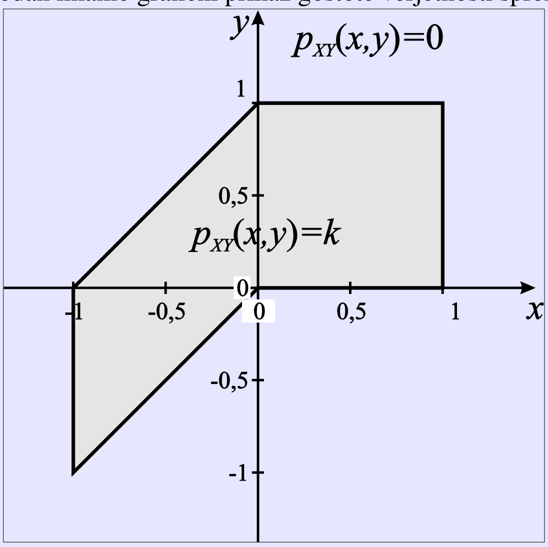
$$0 < x \leq 1:$$

$$p_{XY}(x,y) = \begin{cases} K, & 0 \leq y \leq x \\ 0, & \text{sicer} \end{cases}, \quad p_X(x) = 2x$$

$$p_Y(y|x) = \frac{p_{XY}(x,y)}{p_X(x)} = \begin{cases} \frac{K}{2x}, & 0 \leq y \leq x \\ 0, & \text{sicer} \end{cases}$$

$$p_Y(y|x) = \frac{p_{XY}(x,y)}{p_X(x)} = \begin{cases} \frac{1}{x}, & 0 \leq y \leq x \\ 0, & \text{sicer} \end{cases}$$

Naloga 3.3: Podan imamo grafični prikaz gostote verjetnosti spremenljivk x in y.



- Določi marginalno porazdelitev komponent.
- Izračunaj korelacijo R_{xy}.
- Izračunaj korelacijski koeficient.
- Izračunaj gostoto verjetnosti $p_y(y|x)$.

Rešitev: Najprej moramo določiti gostoto verjetnosti na področju, kjer je različna od 0. Integral po celotnem definicijskem območju (površini x-y) mora biti enak 1. Ker je gostota verjetnosti konstantna, lahko to napravimo grafično (iz slike vidimo, da je površina enaka 2):

$$K \cdot 2 = 1 \quad \Rightarrow \quad K = \frac{1}{2}$$

ali pa z integracijo:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{XY}(x,y)\,dx\,dy = 1$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{XY}(x,y)\,dy\,dx = \int_{-1}^{0}\left(\int_{x}^{x+1} K\,dy\right)dx + \int_{0}^{1}\left(\int_{0}^{1} K\,dy\right)dx$$

$$= \int_{-1}^{0}\left(Ky\big|_{x}^{x+1}\right)dx + \int_{0}^{1}\left(Ky\big|_{0}^{1}\right)dx$$

$$= \int_{-1}^{0}(K(x+1-x))\,dx + \int_{0}^{1}(K)dx$$

$$= K\int_{-1}^{0} dx + K\int_{0}^{1} dx = K(x\big|_{-1}^{0} + x\big|_{0}^{1}) = 2K$$

Marginalni porazdelitvi:

$$p_X(x) = \int_{-\infty}^{\infty} p_{XY}(x,y)\,dy, \quad p_Y(y) = \int_{-\infty}^{\infty} p_{XY}(x,y)\,dx$$

$$\begin{aligned}
x \le -1: &\quad p_X(x) = 0\\
-1 < x \le 0: &\quad p_X(x) = \int_{x}^{x+1} K\,dy = K(x+1-x) = K = \frac{1}{2}\\
0 < x \le 1: &\quad p_X(x) = \int_{0}^{1} K\,dy = Ky\big|_{0}^{1} = K = \frac{1}{2}\\
x > 1: &\quad p_X(x) = 0\\
y < -1: &\quad p_Y(y) = 0\\
-1 \le y < 0: &\quad p_Y(y) = \int_{-1}^{y} K\,dx = Kx\big|_{-1}^{y} = K(y+1) = \frac{1}{2}(y+1)\\
0 \le y \le 1: &\quad p_Y(y) = \int_{y-1}^{1} K\,dx = Kx\big|_{y-1}^{1} = K(1-(y-1)) = \frac{1}{2}(2-y)\\
y > 1: &\quad p_Y(y) = 0
\end{aligned}$$

Naključni spremenljivki sta korelirani, če je kovarianca C_{XY} različna od nič, oziroma, če je korelacija različna od produkta povprečnih vrednosti spremenljivk X in Y.

$$R_{XY} = C_{XY} + \overline{X} \cdot \overline{Y}$$
$$C_{XY} \neq 0 \Rightarrow R_{XY} \neq \overline{X} \cdot \overline{Y}$$
$$R_{XY} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyp_{XY}(x,y)\,dx\,dy$$
$$\begin{aligned}
R_{XY} &= \int_{-1}^{0} \int_{y}^{y+1} xK\,dx\,dy + \int_{0}^{1} \int_{y-1}^{1} xK\,dx\,dy \\
&= K \int_{-1}^{0} y \left(\frac{x^2}{2} \bigg|_{y}^{y+1} \right) dy + K \int_{0}^{1} y \left(\frac{x^2}{2} \bigg|_{y-1}^{1} \right) dy \\
&= \frac{K}{2} \int_{-1}^{0} y \left((y+1)^2 - y^2 \right) dy + \frac{K}{2} \int_{0}^{1} y \left(1 - (y-1)^2 \right) dy \\
&= \frac{K}{2} \int_{-1}^{0} y(2y+1)dy + \frac{K}{2} \int_{0}^{1} y(2y-y^2)dy \\
&= \frac{K}{2} \left(\frac{y^2}{2} + \frac{y}{4} \bigg|_{-1}^{0} \right) + \frac{K}{2} \left(\frac{2y^3}{3} - \frac{y^4}{4} \bigg|_{0}^{1} \right) \\
&= \frac{K}{2} \left(\frac{2}{3} \right) = \frac{K}{3} = \frac{1}{6}
\end{aligned}$$
$$\overline{X} = \int_{-\infty}^{\infty} xp_X(x)\,dx = \frac{1}{2}Kx^2 \bigg|_{-1}^{1} = \frac{K}{2}(1-1) = 0$$
$$\begin{aligned}
\overline{Y} &= \int_{-\infty}^{\infty} yp_Y(y)\,dy = \int_{-1}^{0} yK(y+1)dy + \int_{0}^{1} yK(2-y)dy \\
&= K\left(\frac{1}{3} - \frac{1}{2} \right) + K\left(\frac{1}{3} - \frac{1}{3} \right) = \frac{1}{4}
\end{aligned}$$
$$C_{XY} = R_{XY} - \overline{X} \cdot \overline{Y} = \frac{1}{6} - 0 \cdot \frac{1}{4} = \frac{1}{6}$$

$C_{XY} \neq 0 \quad \Rightarrow \quad$ Naključni spremenljivki sta korelirani.

Korelacijski koeficient $\rho_{XY} = \frac{C_{XY}}{\sigma_X \sigma_Y}$

$$\overline{X^2} = \int_{-\infty}^{\infty} x^2 p_X(x)\, dx = \int_{-1}^{1} x^2 K\, dx = K\frac{x^3}{3}\bigg|_{-1}^{1} = \frac{K}{3}(1+1) = \frac{2}{3} = \frac{1}{3}$$

$$\begin{aligned}
\overline{Y^2} &= \int_{-\infty}^{\infty} y^2 p_Y(y)\, dy = \int_{-1}^{0} y^2 K(y+1)\, dy + \int_{0}^{1} y^2 K(2-y)\, dy \\
&= K\left(\frac{y^4}{4}\bigg|_{-1}^{0} + \frac{y^3}{3}\bigg|_{-1}^{0}\right) + K\left(2\frac{y^3}{3}\bigg|_{0}^{1} - \frac{y^4}{4}\bigg|_{0}^{1}\right) \\
&= K\left(-\frac{1}{4} + \frac{1}{3}\right) + K\left(\frac{2}{3} - \frac{1}{4}\right) = K\left(\frac{1}{2}\right) = \frac{1}{4}
\end{aligned}$$

$$\sigma_X = \sqrt{\overline{X^2} - \overline{X}^2} = \sqrt{\frac{1}{3} - 0} = \frac{1}{\sqrt{3}} = \frac{\sqrt{3}}{3}$$

$$\sigma_Y = \sqrt{\overline{Y^2} - \overline{Y}^2} = \sqrt{\frac{1}{4} - \left(\frac{1}{4}\right)^2} = \sqrt{\frac{3}{16}} = \frac{\sqrt{3}}{4}$$

$$\rho_{XY} = \frac{C_{XY}}{\sigma_X \sigma_Y} = \frac{\frac{1}{6}}{\frac{\sqrt{3}}{3} \cdot \frac{\sqrt{3}}{4}} = \frac{\frac{1}{6}}{\frac{3}{12}} = \frac{4}{6} = \frac{2}{3}$$

Pogojna gostota verjetnosti $p_y(y|x)$

$$p_Y(y \mid x) = \frac{p_{XY}(x,y)}{p_X(x)}$$

$$x \leq -1 \cup x > 1: \quad p_{XY}(x,y) = 0, \quad p_X(x) = 0 \quad \Rightarrow \quad p_Y(y \mid x) = \frac{0}{0}$$

$-1 < x \leq 0:$

$$p_{XY}(x,y) = \begin{cases} K & ; \quad x \leq y \leq x+1 \\ 0 & ; \quad \text{sicer} \end{cases}, \quad p_X(x) = K$$

$$p_Y(y \mid x) = \frac{p_{XY}(x,y)}{p_X(x)} = \begin{cases} \frac{K}{K} & ; \quad x \leq y \leq x+1 \\ \frac{0}{K} & ; \quad \text{sicer} \end{cases}$$

$$p_Y(y \mid x) = \begin{cases} 1 & ; \quad x \leq y \leq x+1 \\ 0 & ; \quad \text{sicer} \end{cases}$$

$0 < x \leq 1:$

$$p_{XY}(x,y) = \begin{cases} K & ; \quad 0 \leq y \leq 1 \\ 0 & ; \quad \text{sicer} \end{cases}, \quad p_X(x) = K$$

$$p_Y(y \mid x) = \frac{p_{XY}(x,y)}{p_X(x)} = \begin{cases} \frac{K}{K} & ; \quad 0 \leq y \leq 1 \\ \frac{0}{K} & ; \quad \text{sicer} \end{cases}$$

$$p_Y(y \mid x) = \begin{cases} 1 & ; \quad 0 \leq y \leq 1 \\ 0 & ; \quad \text{sicer} \end{cases}$$

$P_{XY}(x, y)$	Skupna porazdelitev dveh spremenljivk
$P_X(x) = \sum_y P_{XY}(x, y)$	Marginalna porazdelitev X iz skupne (diskretni primer)
$R_{xy} = \sum_i \sum_j x_i y_j P_{XY}(x_i, y_j)$	Korelacija (diskretno)
$C_{xy} = R_{xy} - \bar{X}\bar{Y}$	Kovarianca (diskretna)
$\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$	Korelacijski koeficient
$p(y \mid x) = \frac{p(x,y)}{p_x(x)}$	Pogojna gostota — za zvezne spremenljivke
$p(x, y) = \frac{1}{\text{ploščina}}$	Enakomerna gostota v zveznem primeru

4. POGLAVJE

AVTOKORELACIJA

V tretjem poglavju smo spoznali **korelacijo** kot orodje za merjenje odnosa med dvema spremenljivkama. Z njeno pomočjo smo ugotavljali, ali sta spremenljivki povezani, in ali morda obstaja takšna povezava, da lahko eno izmed njiju izrazimo z drugo. Takšna redukcija dimenzij ni zgolj matematična eleganca, ampak tudi praktična prednost — omogoča učinkovitejšo predstavitev podatkov in bolj kompaktno kodiranje.

V tem poglavju bomo naredili pomemben korak naprej: spoznali bomo **avtokorelacijo**. Gre za koncept, ki je v osnovi podoben korelaciji, vendar se ne osredotoča na povezavo med dvema različnima spremenljivkama, temveč raziskuje **odnos znotraj ene same spremenljivke skozi čas**.

Zanimalo nas bo:

- ali obstajajo ponavljajoči se vzorci,
- ali se vrednosti v preteklosti ujemajo z vrednostmi v prihodnosti,
- ali obstaja periodičnost, sezonskost ali trendi,
- in ali lahko iz preteklosti napovemo prihodnost.

Avtokorelacija ima ključno vlogo pri analizi **časovnih vrst**, kjer zaporedje podatkov ni naključno, temveč nosi strukturo — red, ki ga lahko odkrijemo in izkoristimo. Zato bomo v tem poglavju naredili odmik od abstraktnih primerov in pokazali praktično uporabo avtokorelacije na realnih časovnih nizih.

To poglavje bo tako predstavljalo pomemben prehod iz teoretičnih temeljev v svet praktične analize podatkov — tam, kjer informacije postanejo uporabne.

Časovne vrste in model ARIMA

Časovna vrsta je zaporedje opazovanj, ki so zbrana skozi čas v enakomernih časovnih intervalih (npr. dnevno, mesečno, letno). V kontekstu elektrotehnike in informatike se časovne vrste pogosto pojavljajo v obliki signalov, meritev, prometnih tokov, obremenitev, števila dogodkov ali ekonomskih indikatorjev.

Sestavni deli časovne vrste Večina realnih časovnih vrst vključuje več naslednjih komponent:

- **Trend** – dolgoročna usmerjenost serije (naraščanje ali upadanje),
- **Sezonskost** – periodično ponavljanje vzorcev (npr. vsakih 12 mesecev),
- **Naključna komponenta (preostanek)** – nestrukturirani ostanki, ki jih ni mogoče pojasniti z modelom,
- **Cikličnost** – dolgoročni neperiodični valovi (npr. gospodarski cikli).

Modeli časovnih vrst Glavni cilj analize časovne vrste je napovedovanje prihodnjih vrednosti na podlagi zgodovinskih podatkov. Eden najpogosteje uporabljenih modelov za to je **ARIMA**.

Kako deluje model ARIMA

Model ARIMA (*Autoregressive Integrated Moving Average*) omogoča modeliranje in napovedovanje časovnih vrst s kombinacijo treh komponent: avtoregresije (AR), integracije (I) in drsečega povprečja (MA). Vsaka komponenta ima svojo vlogo pri razlagi strukture serije:

- **AR** (*autoregressive*) – model predpostavi, da je trenutna vrednost odvisna od preteklih vrednosti časovne vrste.
- **I** (*integrated*) – predstavlja število razlikovanj (diferenc), potrebnih za dosego stacionarnosti.
- **MA** (*moving average*) – modelira odvisnost trenutne vrednosti od preteklih napak napovedi.

Model ARIMA je označen s trojico parametrov (p, d, q), kjer:

- p določa število avtoregresivnih zaostalih vrednosti (lagov),
- d določa število razlikovanj serije,
- q določa število zaostalih napak v modelu.

Avtoregresivni (AR) del Vrednost časovne vrste X_t modeliramo kot linearno kombinacijo preteklih vrednosti:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \varepsilon_t$$

Torej, parameter p določa, **koliko zadnjih vrednosti serije** model upošteva za napoved trenutne vrednosti. Te vrednosti se imenujejo *lag-i*, uteži ϕ_i pa model izračuna med učenjem.

Integrirani (I) del Če serija ni stacionarna, uporabimo razlikovanje:

$$Y_t = X_t - X_{t-1}$$

Za večjo stopnjo trenda ali sezonskosti uporabimo višji nivo odvoda d:

$$Z_t = \nabla^d X_t$$

kjer je ∇ uporabljena kot:

$$\nabla X_t = X_t - X_{t-1}.$$

S tem pretvorimo nestacionarno serijo v stacionarno, ki je primerna za AR in MA modeliranje.

Drseče povprečje (MA) del Model predpostavi, da je trenutna vrednost odvisna tudi od preteklih napak (reziduumov):

$$X_t = \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \ldots + \theta_q\varepsilon_{t-q} + \varepsilon_t$$

Torej q določa, **koliko preteklih napak napovedi** model uporablja. Uporaba napak omogoča kompenzacijo nepojasnjenih nihanj in boljšo prileganje.

Kako model napove novo vrednost Za vsak časovni trenutek t, model izračuna napovedano vrednost $\hat{X}_t$ s kombinacijo:

1. zadnjih p vrednosti serije $(X_{t-1}, \ldots, X_{t-p})$,
2. zadnjih q napak $(\varepsilon_{t-1}, \ldots, \varepsilon_{t-q})$,
3. po potrebi razlikovane serije, če $d > 0$.

Npr. ARIMA(2,1,1) modelira:

$$\Delta X_t = \phi_1 \Delta X_{t-1} + \phi_2 \Delta X_{t-2} + \theta_1 \varepsilon_{t-1} + \varepsilon_t$$

Kar pomeni:

- Predhodno se izračuna razlika $\Delta X_t = X_t - X_{t-1}$,
- Nato se uporabi 2 pretekli razliki $(\Delta X_{t-1}, \Delta X_{t-2})$ in 1 prejšnja napaka za izračun nove vrednosti,
- Napoved se rekonstruira z obratnim razlikovanjem.

Značilnosti modela ARIMA

- Učinkovit za modeliranje nelinearnih trendov in sezonskih pojavov (v kombinaciji z razširitvami, npr. SARIMA),
- Temelji na predpostavki stacionarnosti — podatke pogosto predhodno transformiramo,
- Primeren za enodimenzionalne serije, kjer so vzorci izraženi skozi čas.

ARIMA je robusten model za mnoge vrste podatkov, še posebej ko ni izrazite sezonskosti. Pri sezonskih podatkih se pogosto uporablja njegova razširitev SARIMA.

Primer - analiza števila potnikov v letalskem prometu

Analiza časovnih vrst je eno najmočnejših statističnih orodij, saj nam omogoča, da na podlagi preteklih meritev napovemo prihodnje vrednosti. Uporablja se na različnih področjih — od napovedovanja vremenskih pojavov in finančnih trendov do analize prometa in proizvodnih procesov.

V tej vaji si bomo ogledali primer števila letalskih potnikov na mesečni ravni. Na tem primeru bomo prikazali ključne korake priprave, čiščenja in modeliranja podatkov za napovedovanje prihodnjih vrednosti.

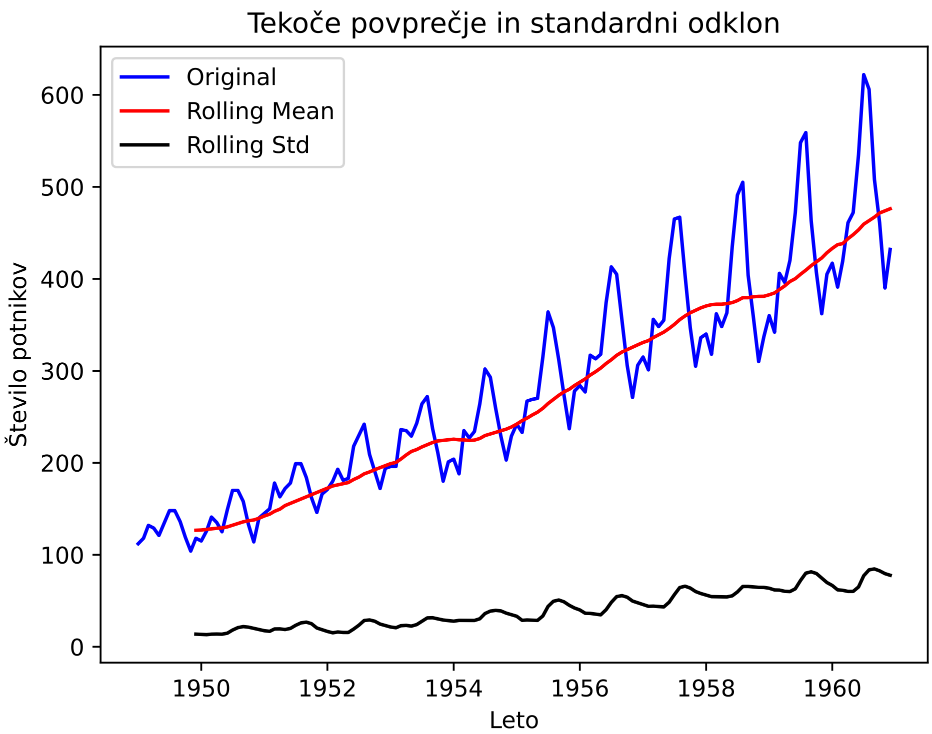
Preden se lotimo celotne procedure razložimo kaj je naš cilj. S pomočjo statistične analize in manipulacije podatkov želimo izdelati model s pomočjo katerega bomo sposobni napovedati naslednje število potnikov. Našo uspešnost bomo vrednotili tako, da bomo dejanske podatke primerjali z našimi napovedmi in pri tem uporabili merilo **koren srednje kvadratne napake**.

Korak 1: Uvoz in vizualizacija podatkov

Začnemo z uvozom podatkov, ki so podani v obliki CSV datoteke. Te podatke shranimo v razpredelnico s časovno osjo kot indeksom, nato pa jih tudi grafično prikažemo. Že ob prvem pogledu lahko opazimo rast števila potnikov ter ponavljajočo se sezonskost.

Opis podatkov

- Časovno obdobje: **januar 1949 – december 1960**
- Število opazovanj: **144 (mesečnih)**
- Frekvenca: **mesečna (12 opazovanj na leto)**
- Spremenljivka: število potnikov, predstavljeno kot celo, pozitivno število



Opazka: Serija očitno ni stacionarna — ima trend in sezonske vzorce, kar pomeni, da je ni mogoče neposredno uporabiti za večino statističnih metod časovne analize. VIdimo predvsem, da povprečje (rdeča) strmo narašča, prav tako pa tekoča standardna deviacija (črna, predstavlja standardno deviacijo zadnjih 12 mesecev) prav tako ni stabilna (ravna).

Korak 2: Preverjanje stacionarnosti

Stacionarnost pomeni, da se povprečje in varianca serije ne spreminjata skozi čas. Če tega pogoja ni, bo napovedovanje manj zanesljivo. Za preverjanje uporabimo Dickey–Fullerjev test.

Test vrne naslednje ključne vrednosti:

- **Testno statistiko**, ki jo uporabimo za primerjavo s kritičnimi vrednostmi,
- **p-vrednost**, ki nam pove, ali lahko zavrnemo hipotezo o nestiacionarnosti (običajno pri p ¡ 0.05),
- **Kritične vrednosti** za 1%, 5% in 10% - če je statistika manjša od kritične vrednosti, lahko zavrnemo hipotezo o nestiacionarnosti. (drugače povedano - če je statistika manjša od kritične vrednosti na 5% nivoju, lahko z 95% gotovostjo trdimo, da je serija stacionarna).

V našem primeru dobimo sledeče rezultate:

```
Results of Dickey-Fuller Test:
```

```
Test Statistic                   0.815369
p-value                          0.991880
#Lags Used                      13.000000
Number of Observations Used    130.000000
Critical Value (1%)             -3.481682
Critical Value (5%)             -2.884042
Critical Value (10%)            -2.578770
dtype: float64
```

Kot vidimo ni testna statistika manjše od nobene kritične vrednosti, kar pomeni da podatki niso stacionarni in jih posledično ne moremo direktno uporabiti v časovni vrsti. Naša naloga je sedaj, da s pomočjo orodij kot so odvajanje, logaritmiranje in drseče povprečje, serijo transformiramo v stacionarno obliko.

Opazka: Uporabiti smemo samo reverzibilne transformacije, da bomo kasneje sposobni naše napovedi vrniti v prvotno območje.

Korak 3: Transformacije za dosego stacionarnosti

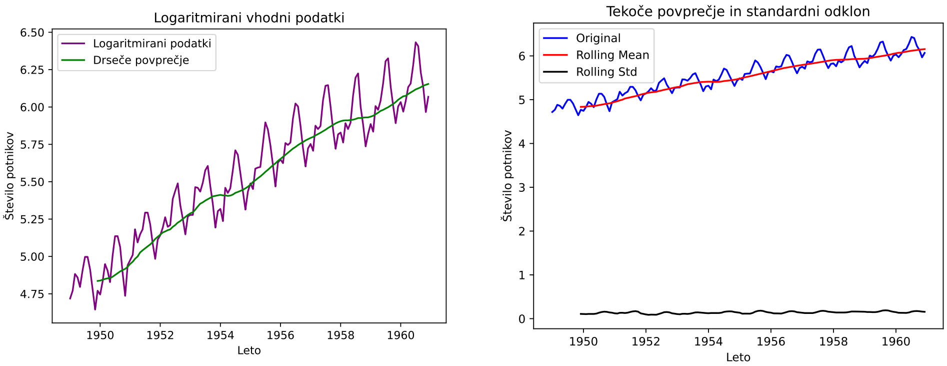
Ker serija vsebuje trend in sezonskost, jo moramo transformirati. Možne strategije vključujejo:

Table 4.1: Pogoste transformacije časovnih vrst za dosego stacionarnosti

Transformacija	**Matematična oblika**	**Namen / učinek**
Logaritemska pretvorba	$y_t = \log(x_t)$	Zmanjšanje variabilnosti, obvladovanje eksponentne rasti
Drseče povprečje	$y_t = \frac{1}{k} \sum_{i=0}^{k-1} x_{t-i}$	Glajenje kratkoročnih nihanj, odstranjevanje šuma
Eksponentno uteženo povprečje (EMA)	$y_t = \alpha x_t + (1 - \alpha) y_{t-1}$	Glajenje podatkov z večjim poudarkom na novejših vrednostih
Diferenciranje	$y_t = x_t - x_{t-1}$	Odstranjevanje trenda, doseganje stacionarnosti
Razstavljanje v komponente	$x_t = T_t + S_t + R_t$	Analiza posameznih komponent: trend, sezona, ostanek

Po vsaki transformaciji (ali kombinaciji večih transformacij) izvedemo Dickey–Fullerjev test, da preverimo, ali je serija postala dovolj stacionarna za izdelavo modela časovne vrste.

Za prvi poskus podatke samo logaritmiramo



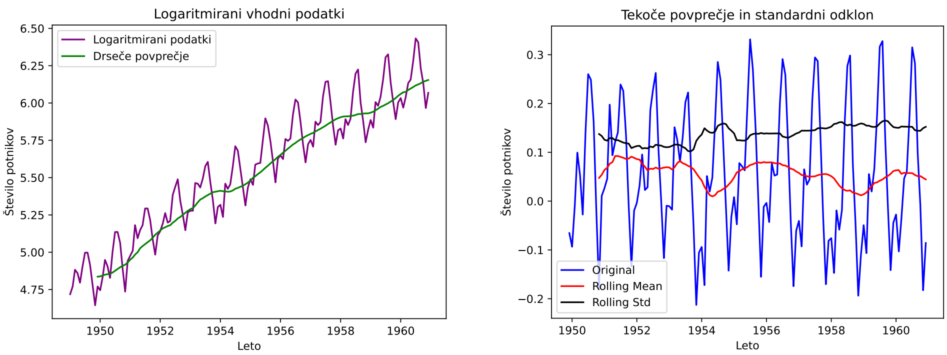
Logaritmirani podatki

Po statističnem testu dobimo sledeče rezultate:

```
Results of Dickey-Fuller Test:
Test Statistic                   -1.717017
p-value                           0.422367
#Lags Used                       13.000000
Number of Observations Used     130.000000
Critical Value (1%)              -3.481682
Critical Value (5%)              -2.884042
Critical Value (10%)             -2.578770
dtype: float64
```

In ugotovimo, da še vedno ni stacionarna. Smo pa bližje kot prej, saj se je sedaj testna statistika vsaj približala kritični vrednosti na 10% nivoju.

Nadaljujemo s kombinacijo logaritmiranja in uporabo (odstranitvijo) 12-mesečnega drsečega povprečja.

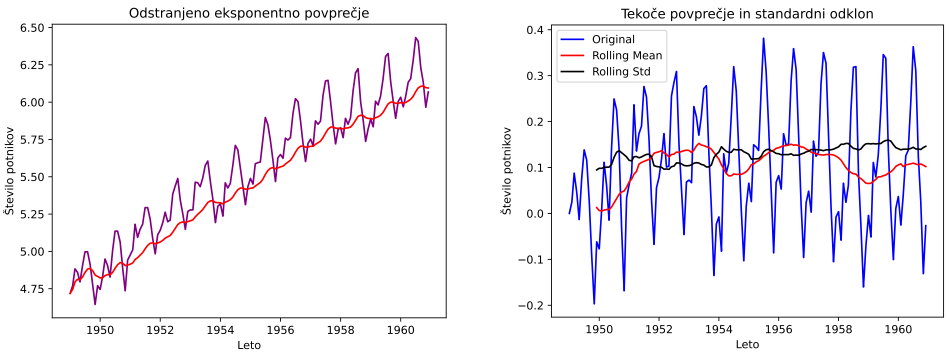


Odstranjeno drseče povprečje

Opazka: Leva slika je ista, saj gre še vedno za logartimirane podatke. Spremeni pa se desna slika, kjer smo sedaj odšteli 12-mesečno drseče povprečje. Po statističnem testu dobimo sledeče rezultate:

```
Results of Dickey-Fuller Test:
Test Statistic                  -3.162908
p-value                          0.022235
#Lags Used                      13.000000
Number of Observations Used    119.000000
Critical Value (1%)             -3.486535
Critical Value (5%)             -2.886151
Critical Value (10%)            -2.579896
dtype: float64
```

To je velik bolje, saj lahko sedaj z 95% gotovostjo trdimo, da je serija stacionarna. Vendar pa še vedno lahko rezultat še izboljšamo, če uporabimo eksponentno uteženo povprečje (EWMA), ki pripisuje večjo težo zadnjim vrednostim.



Odstranjeno eksponentno uteženo povprečje

Po statističnem testu dobimo sledeče rezultate:

```
Results of Dickey-Fuller Test:
Test Statistic                  -3.566092
p-value                          0.006443
#Lags Used                      13.000000
Number of Observations Used    130.000000
Critical Value (1%)             -3.481682
Critical Value (5%)             -2.884042
Critical Value (10%)            -2.578770
dtype: float64
```

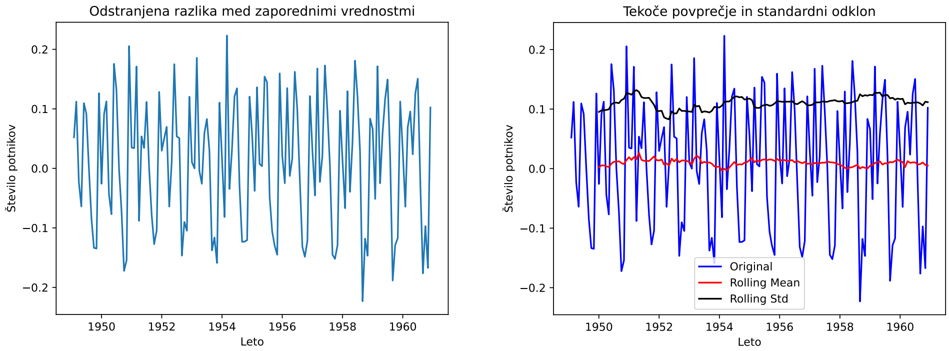
Kar pomeni da je naša serija sedaj stacionarna z 99% gotovostjo. S samimi transformaciji bi sedaj težko še kaj izboljšali doseženi rezultat, zato preidemo na odstranjanje sezonskosti.

Korak 4: Odprava sezonskosti

Naslednji korak je odprava sezonskih vzorcev z **diferenciranjem**:

$$X'(t) = X(t) - X(t-1)$$

Z odštevanjem zamaknjene vrednosti tako skušamo odstraniti sezonske vplive (v resnici je to podobno, kot če bi odvajali naše podatke). Po tem koraku ponovno izvedemo Dickey–Fullerjev test, da preverimo, ali je serija še vedno stacionarna.



Odstranjeno eksponentno uteženo povprečje

Po statističnem testu dobimo sledeče rezultate:

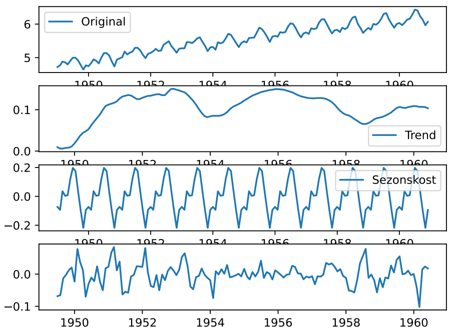
```
Results of Dickey-Fuller Test:
Test Statistic                  -2.649973
p-value                          0.083102
#Lags Used                      14.000000
Number of Observations Used    128.000000
Critical Value (1%)             -3.482501
Critical Value (5%)             -2.884398
Critical Value (10%)            -2.578960
dtype: float64
```

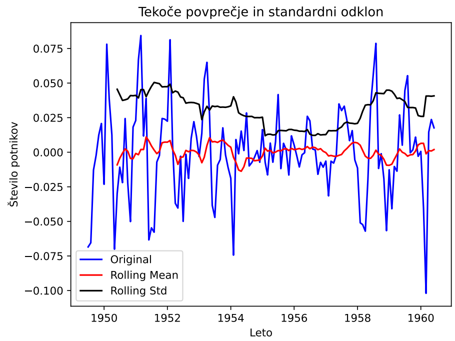
Vidimo, da smo naš rezultat sedaj dejansko poslabšali saj je testna statistika slabša kot prej. To lahko nakazuje na to, da naši podatki nimajo sezonskih vzorcev ALI pa vsebujejo kompleksnejše primere, ki jih z enostavnim diferenciranjem ne moremo odstraniti. Glede na to, da za naše podatke vemo, da predstavljajo letno število potnikov, lahko sklepamo da se naši vzorci pojavljajo na letni ravni (torej npr. sklepamo de veliko več ljudi leti v času počitnic).

V ta namem uporabimo namensko orodje, ki nam je na voljo v okolju Python - **seasonal_decompose**, ki iz podatkov izlušči:

- trendno komponento,
- sezonsko komponento,
- ostanek (rezidual).

Ko imamo enkrat na voljo ostanek (torej edini kos naše serije, ki ga je še potrebno modelirati), ponovno izvedemo Dickey–Fullerjev test, da preverimo, ali je ostanek stacionaren.





Odstranjena sezonskost

Dobimo sledeče rezultate:

```
Results of Dickey-Fuller Test:
Test Statistic                  -6.610539e+00
p-value                          6.393663e-09
#Lags Used                       9.000000e+00
Number of Observations Used      1.220000e+02
Critical Value (1%)             -3.485122e+00
Critical Value (5%)             -2.885538e+00
Critical Value (10%)            -2.579569e+00
dtype: float64
```

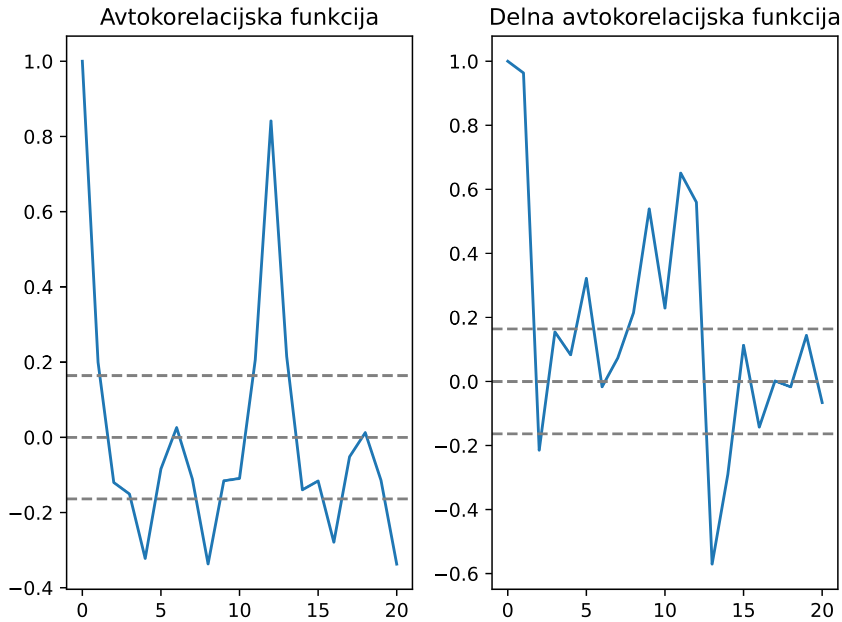
Sedaj lahko z gotovostjo trdimo, da je ostanek stacionaren. To pomeni, da smo uspešno odstranili sezonske vzorce in trendne komponente, kar nam omogoča nadaljnje modeliranje.

Korak 5: Gradnja modela časovne vrste

Zdaj, ko imamo stacionarno serijo, lahko preidemo na modeliranje. Uporabili bomo ARIMA-model, ki je določen s trojico parametrov (p, d, q):

- p: število avtoregresivnih zamikov (ACF),
- d: stopnja diferenčne transformacije,
- q: število zamikov napak (PACF).

Parameter d že poznamo iz prejšnjega koraka (npr. ena diferenčna stopnja pomeni $d = 1$). Za določitev p in q uporabimo avtokorelacijo in delno avtokorelacijo. Poiščemo točko, kjer funkcija prvič pade pod interval zaupanja (črtkasta črta na sliki). V našem primeru dobimo $p = 2$, $q = 2$.



Tako lahko sedaj ovrednotimo tri različne modele:

- AR-model (samo avtoregresija): $(2, 0, 0)$,
- MA-model (samo napake): $(0, 0, 2)$,
- ARIMA-model (kombinirano): $(2, 0, 2)$.

Vsak model bomo vrednotili z vsoto kvadratnih napak (RSS):

Mera napake: vsota kvadratov rezidualov (RSS)

Mera **RSS** (*Residual Sum of Squares*) je ena izmed osnovnih metrik za ocenjevanje kakovosti napovedi v modeliranju časovnih vrst. Predstavlja vsoto kvadratov razlik med dejanskimi vrednostmi in napovedmi.

Enačba:

$$\mathrm{RSS} = \sum_{t=1}^{n} (x_t - \hat{x}_t)^2$$

kjer je:

- x_t — dejanska (opazovana) vrednost v trenutku t,
- $\hat{x}_t$ — napovedana vrednost modela v trenutku t,
- n — skupno število opazovanj.

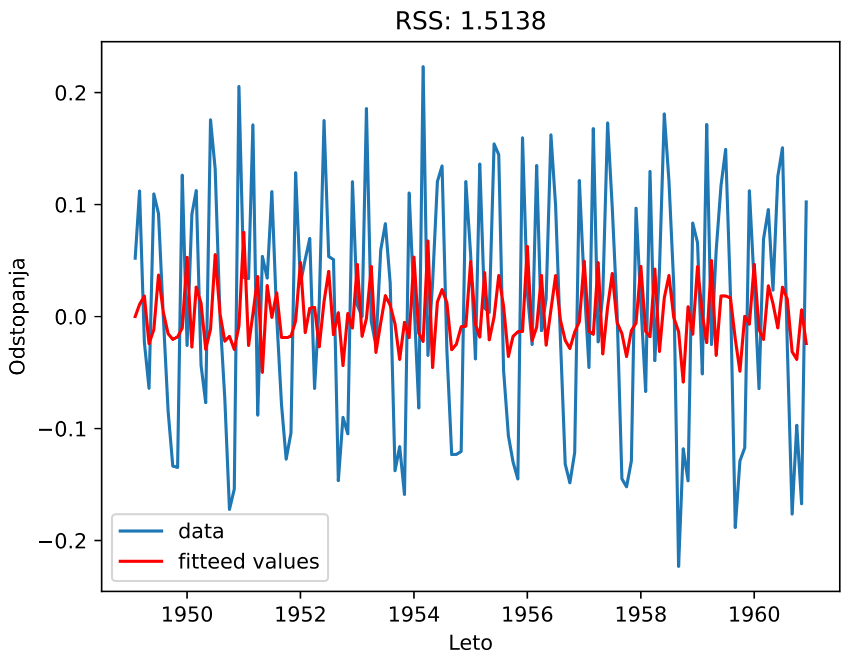
Lastnosti:

- Ni negativna: $\mathrm{RSS} \geq 0$,
- Manjša vrednost RSS pomeni boljše prileganje modela,

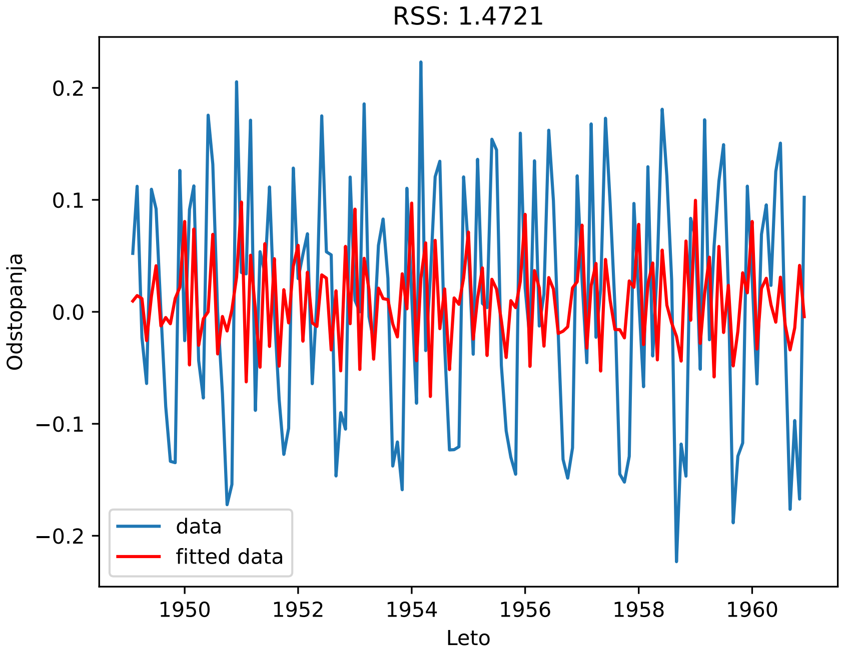
- Zelo občutljiva na velike napake (zaradi kvadriranja),
- Pogosto se uporablja v kombinaciji z RMSE ali AIC.

Korak 6: Napoved in ovrednotenje modela

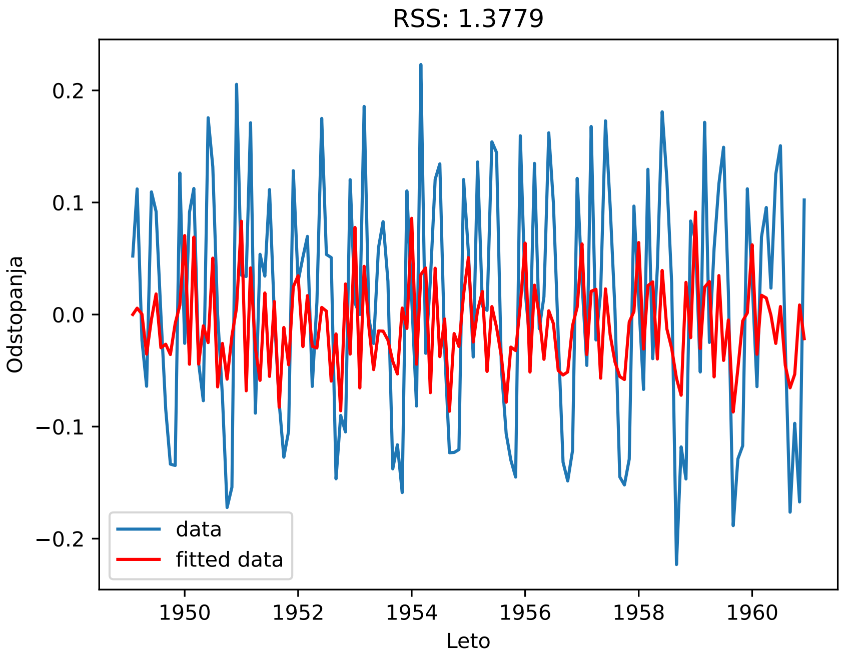
Pri vrednotenju modela se osredotočimo na naš ”ostanek” (rezidual), ki je rezultat odstranitve trenda in sezonskosti. Če namreč želimo, da bo naš model pravilen, mora biti sposoben pravilno napovedati ostanek, saj bomo ostale podatke vrnili z uporabo obratnih procesov (prišteli bomo nazaj trend, sezonskost in po potrebo podatke potencirali).



AR model (2,0,0)



MA model (0,0,2)



ARIMA model (2,0,2)

Pričakovano ugotovimo, da je ARIMA model najboljši od treh saj uporabi vse prednosti modela. Ko imamo tako izbran model se lahko lotimo še celotnega procesa rekonstrukcije podatkov ter končnega vrednotenja modela z uporabo korena srednje kvadratne napake (RMSE).

Korak 7: Rekonstrukcija in vrednotenje

Za ovrednotenje natančnosti modela uporabimo **RMSE** — koren srednje kvadratne napake:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}$$

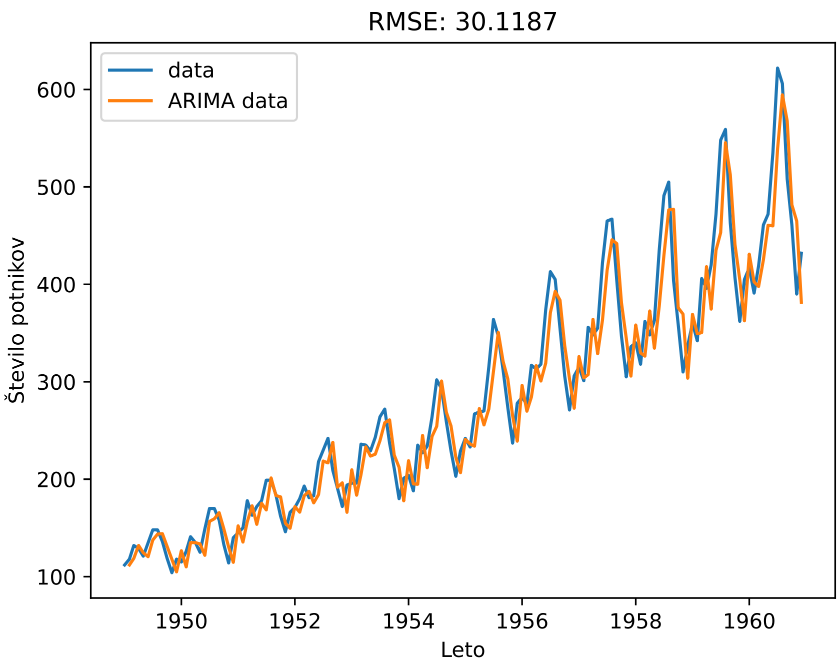
kjer je:

- y_t — dejanska (opazovana) vrednost v trenutku t,
- $\hat{y}_t$ — napovedana vrednost modela v trenutku t,
- n — skupno število opazovanj.

Ob tem se moramo zavedati, da moramo za vsako vrednotenje podatke, dobljene z ARIMA modelom najprej vrniti v prvotno območje.

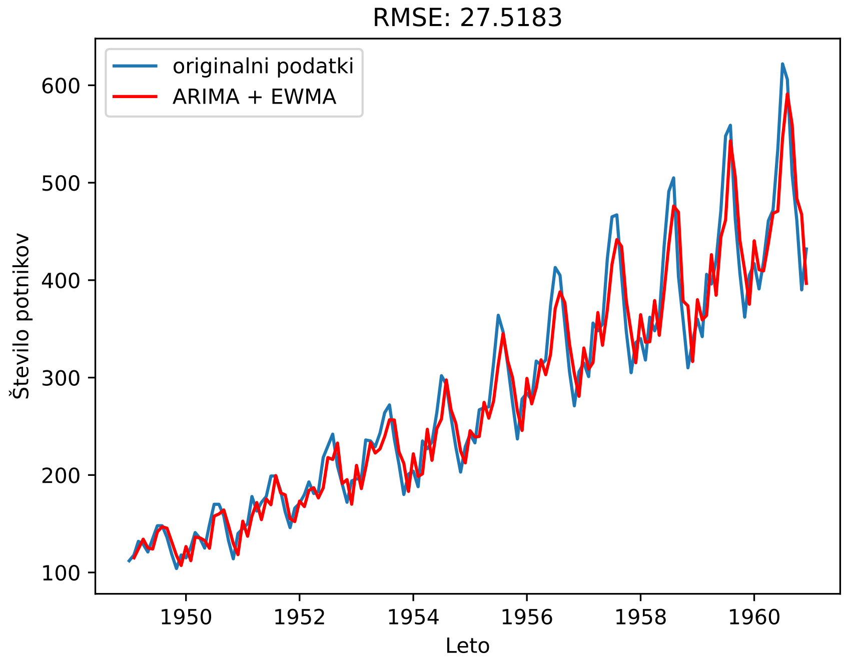
Za prikaz, kako izbira modela in transformacije vpliva na končno natančnost ne bomo ovrednotili samo najboljšega (torej modela z odstranjeno sezonskostjo), temveč tudi ostale modele, ki smo jih ustvarili v prejšnjih korakih.

Prvi model bo tako ARIMA(2,0,2), kjer smo za vhodne podatke uporabili samo logaritmirane podatke brez odstranjene sezonskosti. Za vrednotenje moramo torej vsak napovedani podatek potencirati z osnovo e. Dobljeni rezultat sicer izgleda, kot da sledi izvornim podatkom, vendar je RMSE precej visok (30.11).



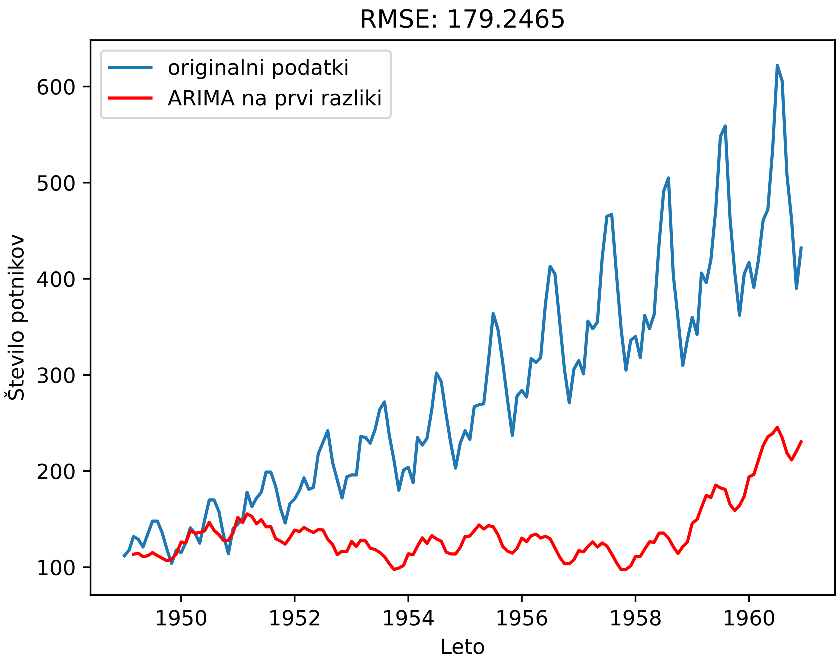
Rekonstrukcija - logaritmiran vhod

Preden smo se lotili odstranjevanja sezonskosti, smo ugotovili, da smo največjo stacionarnost dosegli, če smo odšteli eksponentno drseče povprečje. Če uporabimo te podatke kot vhod za naš model dobimo že malo boljši razultat (RMSE = 27.5), ki nekako predstavlja tudi najboljše, ker lahko dosežemo, če zanemarjamo sezonskost.



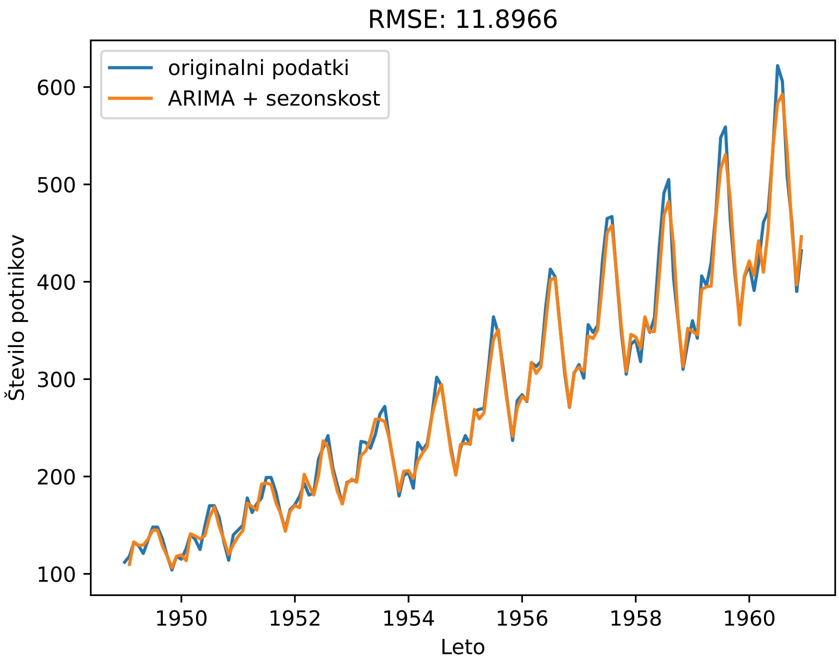
Rekonstrukcija - logaritmiran vhod z odštetim eksponentnim uteženim povprečjem

V prvem poskusu odstranjevanja sezonskosti smo uporabili samo diferenciranje (odštevanje zamaknjene vrednosti) in pokazali, da smo s tem dejansko pokvarili stacionarnost naših podatkov. Pričakovano se zato pokvari tudi rezultat modela, ki sedaj ne vrne uporabnih podatkov (RMSE = 179.2).



Rekonstrukcija - diferenciranje

Za konec uporabimo še naše najoljše vhodne podatke - logaritmirane podatke z odstranjeno sezonskostjo. Že na sliki takoj vidimo, da model sledi izvornim podatkom, mera RMSE (11.89) pa pokaže da smo skoraj trikrat boljši kot pri najboljšem modelu brez sezonskosti.



Rekonstrukcija - odstranjena sezonskost

Sklep

V tem poglavju smo pokazali, kako nam poznavanje lastnosti podatkov omogoča, da jih uporabimo za napovedovanje prihodnjih vrednosti. Tu nam predvsem pomaga avtokoroelacija, ki podatke primerja same s seboj in nam tako omogoča, da najdemo vzorce, ki jih lahko uporabimo za napovedovanje. S pomočjo avtokorelacije in delne avtokorelacije smo prikazali celotni proces priprave vhodnih podatkov, izdelave ARIMA modela ter pravilne rekonstrukcije napovedanih vrednosti.

Predstavljena metoda (poleg same didaktične vrednosti za prikaz uporabnosti avtokorelacije) je pogosto uporabljana tako v financah (napoved vrednosti delnic), telekomunkacijah (napoved obremenitve omrežja), meteorologiji (napoved vremena) in mnogih drugih področjih, kjer so podatki časovno odvisni.

V tej vaji smo predstavili celoten proces analize časovne vrste — od vizualizacije do napovedi. Uporabili smo realne podatke in pokazali, kako pomembna je stacionarnost za uspešno modeliranje. Prav tako smo videli, kako kompleksne strukture (sezonskost, trendi) zahtevajo več korakov čiščenja in predobdelave.

Za tiste, ki se želijo v metodo bolje poglobiti je v **dodatku** na voljo tudi s pomočjo katere smo izvedli celotno poglavje. Za dodatno branje priporočamo naslednje vire:

- S. Haykin: *Communication Systems, 5th ed.* [7]

5. POGLAVJE

INFORMACIJA IN ENTROPIJA

V tem poglavju bomo z reševanjem nalog prikazali, kako merimo količino informacije in izračunamo entropijo v različnih verjetnostnih sistemih. V kontekstu teorije informacij informacija meri količino presenečenja oziroma negotovosti, ki jo povzroči pojav določenega dogodka. Entropija pa predstavlja povprečno informacijo oziroma negotovost, ki je prisotna v danem verjetnostnem sistemu.

Za posamezen dogodek x z verjetnostjo $P(x)$ definiramo informacijo kot:

$$I(x) = \log_2 \frac{1}{P(x)}$$

To pomeni, da redkejši dogodki nosijo večjo informacijo.

Če imamo diskretno naključno spremenljivko X, definiramo njeno entropijo kot povprečno informacijo vseh možnih vrednosti:

$$H(X) = \sum_i P(x_i) \cdot I(x_i) = -\sum_i P(x_i) \log_2 P(x_i)$$

Entropija je največja, ko so vsi dogodki enako verjetni, in znaša $\log_2 n$, kjer je n število možnih vrednosti.

Ko imamo opravka z dvema spremenljivkama, nas zanima tudi skupna entropija:

$$H(X, Y) = -\sum_{x,y} P(x, y) \log_2 P(x, y)$$

ter pogojna entropija:

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

Ta pove, koliko negotovosti ostane v Y, če že poznamo X.

Razumevanje teh konceptov je ključno za kodiranje z minimalno povprečno dolžino, konstrukcijo optimalnih kodov (npr. Huffman) in za ocenjevanje izgub pri prenosu informacij.

Za dodatno branje priporočamo naslednje vire:

- T. M. Cover, J. A. Thomas: *Elements of Information Theory* [2]
- D. J. C. MacKay: *Information Theory, Inference, and Learning Algorithms* [3]
- N. Pavešič: *Informacija in kodi* [6]

5.1 Naloge

Naloga 5.1: Glavni dobitek pri lotu je sedmica - kombinacija 7 števil iz razpona $[1, 39]$. Izračunaj informacijo pravilne kombinacije. Izračunaj tudi informacijo ostalih kombinacij.

Rešitev: Najprej rabimo podatek o vseh možnih kombinacijah, ki jih lahko izberemo

$$N = \binom{39}{7} = \frac{39!}{(39-7)!7!} = 15380937$$

Verjetnost je za vsako kombinacijo enaka.

$$P = \frac{1}{N} = \frac{1}{15380937} = 6{,}5 \cdot 10^{-8}$$

Informacija pravilne kombinacije je torej:

$$I = \log_2 \frac{1}{P} = \log_2 15380937 = 23{,}87 \text{ sh}$$

Naloga 5.2: Diskretna spremenljivka $X, x_i \in \{1, 2, \ldots, 6\}$. Verjetnost simbolov je podana z $P_x(x_i) = P$.

$$P = [0{,}35\ 0{,}35\ 0{,}1\ 0{,}1\ 0{,}05\ 0{,}05]$$

Izračunaj informacijo posamezne vrednosti, njeno povprečno vrednost ter entropijo.

Rešitev: Pri neenakomerni porazdelitvi vsaka vrednost prispeva različno informacijo. Najpogostejše vrednosti imajo manjšo informacijo, redkejše večjo. Entropija

$H(X)$ je povprečna informacija:

$$H(X) = \sum_i P(x_i) \cdot \log_2 \frac{1}{P(x_i)} = \sum_i P_X(x_i) \cdot I_X(x_i)$$

Najprej izračunamo informacijo posameznih vrednosti:

$$I = \log_2 \frac{1}{P}$$
$$I = [1{,}515 \quad 1{,}515 \quad 3{,}322 \quad 3{,}322 \quad 4{,}322 \quad 4{,}322 \quad 4{,}322]$$

Sedaj lahko izračunamo še entropijo:

$$\begin{aligned} H(X) &= 0{,}35 \cdot 1{,}515 + 0{,}35 \cdot 1{,}515 + 0{,}1 \cdot 3{,}322 + 0{,}1 \cdot 3{,}322 \\ &\quad + 0{,}05 \cdot 4{,}322 + 0{,}05 \cdot 4{,}322 \\ H(X) &= 2{,}16 \end{aligned}$$

Naloga 5.3: Naključni spremenljivki $X, Y = 1, 2, 3, 4, 5$ sta podani z vezano verjetnostjo funkcijo, podani v obliki matrike

$$P_{XY}(x_i, y_j); \quad \mathbf{P}_{XY} = \frac{1}{25} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 1 & 0 \\ 2 & 0 & 2 & 1 & 1 \\ 0 & 3 & 1 & 2 & 2 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

- Izračunajte entropiji posameznih spremenljivk $H(X)$ in $H(Y)$. Katera spremenljivka je bolj nedoločena?
- Izračunajte vezano entropijo $H(X, Y)$. Kakšna je relacija med vezano in entropijo posamezne spremenljivke.
- Določite matriko pogojne verjetnosti P_K

$$P_Y(y_j \mid x_i) = \frac{P_{XY}(x_i, y_j)}{P_X(x_i)}$$

$$\mathbf{P}_K = \begin{bmatrix} P_Y(y_1 \mid x_1) & P_Y(y_2 \mid x_1) & P_Y(y_3 \mid x_1) & P_Y(y_4 \mid x_1) & P_Y(y_5 \mid x_1) \\ P_Y(y_1 \mid x_2) & P_Y(y_2 \mid x_2) & P_Y(y_3 \mid x_2) & P_Y(y_4 \mid x_2) & P_Y(y_5 \mid x_2) \\ P_Y(y_1 \mid x_3) & P_Y(y_2 \mid x_3) & P_Y(y_3 \mid x_3) & P_Y(y_4 \mid x_3) & P_Y(y_5 \mid x_3) \\ P_Y(y_1 \mid x_4) & P_Y(y_2 \mid x_4) & P_Y(y_3 \mid x_4) & P_Y(y_4 \mid x_4) & P_Y(y_5 \mid x_4) \\ P_Y(y_1 \mid x_5) & P_Y(y_2 \mid x_5) & P_Y(y_3 \mid x_5) & P_Y(y_4 \mid x_5) & P_Y(y_5 \mid x_5) \end{bmatrix}$$

Rešitev: Iz tabele skupne verjetnosti izračunamo marginalni porazdelitvi:

$$P_X(x_i) = \sum_j P_{XY}(x_i, y_j), \quad P_Y(y_j) = \sum_i P_{XY}(x_i, y_j)$$

Pogojna verjetnost:

$$P(Y = y_j \mid X = x_i) = \frac{P_{XY}(x_i, y_j)}{P_X(x_i)}$$

Entropijo dobimo kot povprečno informacijo:

$$H(X) = \sum_i P_X(x_i) \cdot \log_2 \frac{1}{P_X(x_i)}$$

Začnemo torej z marginalnima verjetnostma $P_X(x)$ in $P_Y(y)$

$$\mathbf{P}_X = \frac{1}{25}\begin{bmatrix} 5 & 4 & 6 & 8 & 2 \end{bmatrix} \quad \mathbf{P}_Y = \frac{1}{25}\begin{bmatrix} 3 & 5 & 7 & 6 & 4 \end{bmatrix}$$

Iz marginalnih verjetnosti izračunamo informacije za posamezne vrednosti x in y:

$$I = -\log_2 P$$
$$\mathbf{I}_X = [2{,}32 \quad 2{,}64 \quad 2{,}06 \quad 1{,}64 \quad 3{,}64] \qquad \mathbf{I}_Y = [3{,}06 \quad 2{,}32 \quad 1{,}84 \quad 2{,}06 \quad 2{,}64]$$

In za X in Y izračunamo povprečno informacijo:

$$H(X) = \sum_i P_X(x_i) \cdot I_X(x_i) \qquad H(Y) = \sum_i P_Y(x_i) \cdot I_Y(x_i)$$
$$H(X) = 2{,}20 \qquad H(Y) = 2{,}26$$

Spremenljivka Y je bolj nedoločena, ker ima večjo entropijo.
Vezano entropijo $H(X, Y)$ izračunamo tako, da izračunamo informacijo za vsak par vrednosti x_i, y_i (vsak element matrike P_{XY}) in nato izračunamo povprečno vrednost informacije:

$$\mathbf{I_{XY}} = \begin{bmatrix} 4{,}64 & 4{,}64 & 4{,}64 & 4{,}64 & 4{,}64 \\ \infty & 4{,}64 & 3{,}64 & 4{,}64 & \infty \\ 3{,}64 & \infty & 3{,}64 & 4{,}64 & 4{,}64 \\ \infty & 3{,}06 & 4{,}64 & 3{,}64 & 3{,}64 \\ \infty & \infty & 4{,}64 & 4{,}64 & \infty \end{bmatrix}$$

$$H(X, Y) = \sum_i \sum_j P_{XY}(x_i, y_j) \cdot I_{XY}(x_i, y_j) = 4{,}05$$

Vezana entropija je večja od entropije posameznih spremenljivk.

Matrika pogojne verjetnosti P_K pa je:

$$\mathbf{P_K} = \begin{bmatrix} 0{,}2 & 0{,}2 & 0{,}2 & 0{,}2 & 0{,}2 \\ 0 & 0{,}25 & 0{,}5 & 0{,}25 & 0 \\ 0{,}333 & 0 & 0{,}333 & 0{,}167 & 0{,}167 \\ 0 & 0{,}375 & 0{,}125 & 0{,}25 & 0{,}25 \\ 0 & 0 & 0{,}5 & 0{,}5 & 0 \end{bmatrix}$$

$I(x) = \log_2 \frac{1}{P(x)} = -\log_2(P(x))$	Informacija posameznega dogodka
$H(X) = \sum_i P(x_i) I(x_i) = -\sum_i P(x_i) \log_2 P(x_i)$	Entropija diskretne spremenljivke
$H(X,Y) = -\sum_{x,y} P(x,y) \log_2 P(x,y)$	Skupna entropija dveh spremenljivk

6. POGLAVJE

VLC KODIRANJE

V tem poglavju bomo obravnavali tehnike kodiranja, pri katerih posamezni simboli nimajo enake dolžine predstavitve, temveč se njihova kodna dolžina prilagaja glede na verjetnost pojavljanja. Cilj takšnega kodiranja je zmanjšanje povprečne dolžine predstavitve zaporedja podatkov, hkrati pa ohraniti nesporočilnost (angl. *prefix-free*) kod.

Te tehnike temeljijo na osnovnem načelu teorije informacije: bolj pogosti simboli naj dobijo krajšo predstavitev, manj pogosti pa daljšo, kar vodi k večji povprečni učinkovitosti prenosa. Takšne metode kodiranja so še posebej uporabne pri stiskanju podatkov, brez izgube informacij.

Osredotočili se bomo na dve razširjeni metodi:

- **Huffmanovo kodiranje**, ki za dano porazdelitev verjetnosti zagotovi optimalno nesporočilno kodo z najkrajšo možno povprečno dolžino.
- **LZW (Lempel–Ziv–Welch)** kodiranje, ki temelji na gradnji slovarja med samim kodiranjem, brez predhodnega poznavanja verjetnosti.

Poglavje bomo začeli z analizo entropije in kodne redundance. Praktične naloge vključujejo generiranje Huffmanovih kod, oceno kodne učinkovitosti ter uporabo LZW kodiranja in dekodiranja.

Pri delu bomo spoznali tudi koncept razširjenih virov – s kombiniranjem osnovnih simbolov v bloke daljše dolžine lahko dodatno znižamo redundanco, če enosimbolna Huffmanova koda ne dosega zahtevane učinkovitosti.

Na koncu bomo s pomočjo tabel povzemali ključne enačbe, ki opisujejo povezavo med entropijo, dolžino kod in redundanco.

Za dodatno branje priporočamo naslednje vire:

- T. M. Cover, J. A. Thomas: *Elements of Information Theory* [2]
- D. J. C. MacKay: *Information Theory, Inference, and Learning Algorithms* [3]
- N. Pavešič: *Informacija in kodi* [6]

6.1 Naloge

Naloga 6.1: Podane so verjetnosti simbolov informacijskega vira $X = \{x_1, x_2, \ldots x_6\}$. Določite binarne kode za posamezne simbole z najmanjšim možnim številom bitov, tako da bodo imele vse kode enako dolžino. Izračunajte entropijo informacijskega vira. Določite Huffmanovo kodo za dani primer. Kolikšni sta absolutna in relativna redundanca informacijskega vira in obeh tipov kodiranja? Kolikšna je učinkovitost obeh tipov kodiranja?

$$P_X = [0{,}4 \quad 0{,}08 \quad 0{,}1 \quad 0{,}22 \quad 0{,}18 \quad 0{,}02]$$

Rešitev: Za oceno uspešnosti našega kodiranja moramo nujno poznati entropijo informacijskega vira:

$$H(X) = -\sum_i P_X[i] \cdot \log_2 P_X[i] = 2{,}19$$

V primeru kodiranje je entropija vira naš ideal - povprečno število bitov, ki jih vir pošilja. Naš cilj je doseči, da bo povprečno število bitov na simbol čim bližje entropiji vira. Tako smo lahko prepričani, da v kanal ne pošiljamo odvečnih bitov.
Če začnemo s kodiranje s fiksno dolžino, za vsak simbol uporabimo enako število bitov in dobimo:

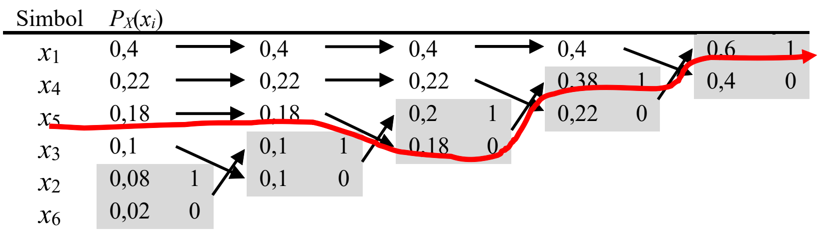
Simbol	Koda
x_1	000
x_2	001
x_3	010
x_4	011
x_5	100
x_6	101

Povprečna dolžina takega kodiranja je:

$$\bar{n} = \sum_i P_X[i] \cdot n_i = 3$$

saj smo za vsak simbol uporabili isto število bitov. Takoj vidimo da je ta številka večja od entropije vira, kar pomeni, da smo v kanal poslali nekaj odvečnih bitov.

Huffmanovo kodiranje:
Simbole razvrstimo v tabelo glede na njihove verjetnosti od simbola z najvišjo verjetnostjo do simbola z najnižjo verjetnostjo. Simboloma z najnižjo verjetnostjo dodelimo logično 0 in 1 ter ju združimo v naslednjem stolpcu. Če je verjetnost združenih simbolov enaka verjetnosti katerega drugega simbola, najprej napišemo verjetnost združenega simbola. Tako nadaljujemo, dokler ne ostaneta le še dve verjetnosti



Kode za posamezne simbole odčitamo iz tabele tako, da sledimo puščicam. Odčitamo jih v obratni smeri. Z rdečo je označen primer za x_5.

Simbol	Koda
x_1	0
x_2	11111
x_3	1110
x_4	10
x_5	110
x_6	11110

Povprečna dolžina kode:

$$\bar{n} = \sum_i P_X[i] \cdot n_i = 0{,}4 \cdot 1 + 0{,}08 \cdot 5 + 0{,}1 \cdot 4 + 0{,}22 \cdot 2 + 0{,}18 \cdot 3 + 0{,}02 \cdot 5$$
$$= 2{,}28$$

Sedaj lahko izračunamo redundanco informacijskega vira, kjer primerjamo entropijo vira z entropijo idealnega vira (kjer so vsi simboli enako verjetni):

$$D_{IV} = H_{\max} - H = \log_2 L - H = \log_2 6 - 2{,}19 = 0{,}39$$

$$R_{IV} = \frac{D_{IV}}{H_{\max}} = \frac{0{,}39}{\log_2 6} = 0{,}15$$

Redundanca kodiranja s kodo konstantne dolžine (Constant Length Code):

$$D_{CLC} = \bar{n} - H = 3 - 2{,}19 = 0{,}81$$

$$R_{CLC} = \frac{D_{CLC}}{\bar{n}} = \frac{0{,}81}{3} = 0{,}27$$

Učinkovitost kodiranja (kako blizu smo entropiji vira) s kodo konstantne dolžine (Constant Length Code):

$$\eta_{CLC} = \frac{H}{\bar{n}} = \frac{2{,}19}{3} = 0{,}73$$

Redundanca kodiranja s kodo spremenljive dolžine (Variable Length Code):

$$D_{VLC} = \bar{n} - H = 2{,}28 - 2{,}19 = 0{,}09$$

$$R_{VLC} = \frac{D_{VLC}}{\bar{n}} = \frac{0{,}09}{2{,}28} = 0{,}04$$

Učinkovitost kodiranja s kodo spremenljive dolžine (Variable Length Code):

$$\eta_{VLC} = \frac{H}{\bar{n}} = \frac{2{,}19}{2{,}28} = 0{,}96$$

Naloga 6.2: Izračunajte entropijo ikone, ki je kodirana s štirimi biti na slikovno točko (0 – črna, 8 – bela). Kolikšno je teoretično najmanjše možno število bitov na slikovno točko, da pri kodiranju še ne pride do izgube informacije? Določite Huffmanovo kodo za dani primer in izračunajte povprečno število bitov na točko. Kolikšna je redundanca dobljene Huffmanove kode? (Da bo vaja krajša, jo izračunajte le za prvih 6×6 točk)

8	8	8	2	1	0	0	0	1	2	7	7	7
8	6	1	1	2	3	3	3	2	1	1	5	7
7	1	2	3	4	4	3	3	3	3	2	1	8
2	1	3	3	3	3	3	2	1	1	3	1	2
1	2	3	2	1	2	3	1	8	0	2	2	1
0	3	3	1	8	0	2	2	0	0	2	3	0
0	4	3	2	1	0	2	3	2	2	3	4	0
0	3	4	3	2	2	3	3	3	3	3	3	0
1	2	4	0	2	3	3	4	2	4	4	2	1
2	1	4	4	1	2	1	1	1	4	3	1	2
7	1	2	3	3	1	2	2	4	3	2	1	8
7	5	1	1	2	3	4	3	2	1	1	6	8
7	7	7	2	1	0	0	0	1	2	8	8	8

Rešitev: Za vsako možno vrednost točke (simbol) $X = \{0, 1, \ldots, 15\}$ določimo verjetnost s štetjem točk, ki imajo to vrednost.

$$N_0 = 3, \quad P_X(0) = \frac{N_0}{N} = \frac{3}{36}$$

$$N_1 = 8, \quad P_X(1) = \frac{8}{36}$$

itd...

$\mathbf{N} = [3 \quad 8 \quad 7 \quad 9 \quad 2 \quad 0 \quad 0 \quad 1 \quad 1 \quad 5]$

$$\mathbf{P} = \frac{1}{36}\mathbf{N} = [0{,}083 \quad 0{,}222 \quad 0{,}194 \quad 0{,}250 \quad 0{,}056 \quad 0 \quad 0 \quad 0{,}028 \quad 0{,}028 \quad 0{,}139]$$

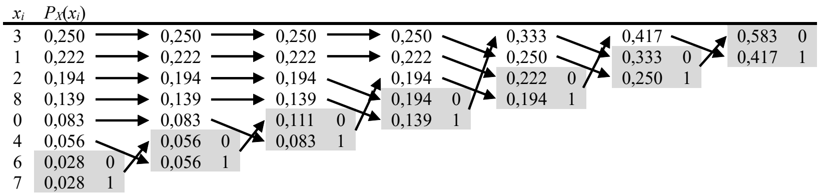
Za vsak simbol izračunamo informacijo

$$\mathbf{I} = -\log_2 \mathbf{P} = [3{,}58 \quad 2{,}17 \quad 2{,}36 \quad 2{,}00 \quad 4{,}17 \quad \infty \quad \infty \quad 5{,}17 \quad 5{,}17 \quad 2{,}85]$$

in nato povprečno informacijo vseh točk

$$H(X) = \sum_i P_X[i] \cdot I_X[i] = 2{,}655$$

To pomeni, da bi bilo dovolj, če bi vsako točko zakodirali v povprečju z 2,66 bita.



x_i	koda	dolžina n_i
0	111	3
1	10	2
2	000	3
3	01	2
4	1101	4
5	0	1
6	11000	5
7	11001	5
8	001	3

Povprečna dolžina kode na slikovno točko:

$$\overline{n} = \sum_i n_i P_X[i] = 0{,}083 \cdot 3 + 0{,}194 \cdot 2 + \ldots + 0{,}139 \cdot 3 = 2{,}694$$

Redundanca:

$$D = \overline{n} - H = 2{,}694 - 2{,}66 = 0{,}040$$

$$R = \frac{D}{\overline{n}} = \frac{0{,}040}{2{,}694} = 0{,}015$$

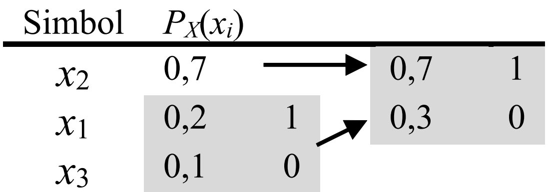
Naloga 6.3: Določite kode simbolov danega informacijskega vira tako, da bo relativna kodirna redundanca manjša od 5 %. Če Huffmanova koda ne zadostuje, razširite informacijski vir s kombiniranjem simbolov v večje bloke.

$$P_X = [0{,}2 \quad 0{,}7 \quad 0{,}1]$$

Rešitev: Entropija vira:

$$H(X) = -\sum_i P_X[i] \cdot \log_2 P_X[i] = 1{,}157$$

Huffmanova koda:



Simbol	Koda
x_1	01
x_2	1
x_3	00

Povprečna dolžina Huffmanove kode:

$$\overline{n} = \sum_i P_X[i] \cdot n_i = 0{,}2 \cdot 2 + 0{,}7 \cdot 1 + 0{,}1 \cdot 2 = 1{,}3$$

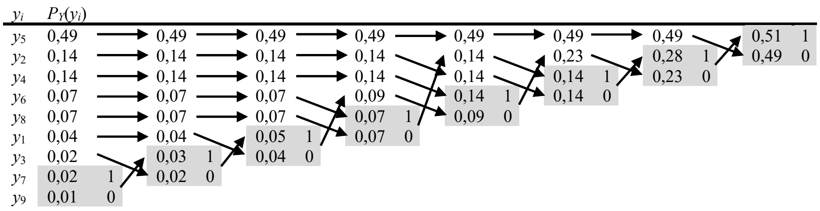
Relativna kodirna redundanca:

$$R = \frac{\overline{n} - H(X)}{\overline{n}} = \frac{1{,}3 - 1{,}157}{1{,}3} = 0{,}11$$

Huffmanova koda ne zadostuje, zato združimo po dva simbola v nov simbol, da dobimo razširjeni vir Y. Verjetnost posameznih simbolov razširjenega vira je odvisna od verjetnosti simbolov primarnega vira. Ker so zaporedni simboli med sabo neodvisni, izračunamo verjetnost novega simbola kot produkt verjetnosti simbolov primarnega vira.

Simbol y_i	Kombinacija	$P_Y[i]$
y_1	$x_1\,x_1$	0,04
y_2	$x_1\,x_2$	0,14
y_3	$x_1\,x_3$	0,02
y_4	$x_2\,x_1$	0,14
y_5	$x_2\,x_2$	0,49
y_6	$x_2\,x_3$	0,07
y_7	$x_3\,x_1$	0,02
y_8	$x_3\,x_2$	0,07
y_9	$x_3\,x_3$	0,01

Huffmanova koda razširjenega vira:



Simbol	Koda
y_1	1000
y_2	110
y_3	10010
y_4	101
y_5	0
y_6	1111
y_7	100111
y_8	1110
y_9	100110

Entropija razširjenega informacijskega vira:

$$H(Y) = -\sum_i P_Y[i] \cdot \log_2 P_Y[i] = 2{,}3136$$

Povprečna dolžina kode razširjenega informacijskega vira:

$$\bar{n} = \sum_i P_Y[i] \cdot n_i = 2{,}33$$

Relativna kodirna redundanca razširjenega vira:

$$R = \frac{\bar{n} - H(Y)}{\bar{n}} = \frac{2{,}33 - 2{,}3136}{2{,}33} = 0{,}0071$$

Naloga 6.4: Podani niz zakodirajte z LZW algoritmom. Dobljeni niz nato ponovno dekodirajte.
Niz: A A B B A B A A A B A A B A A B A B A B
Abeceda: A, B

Rešitev: V začetni kodni knjigi sta samo simbola podane abecede:

0	A
1	B

V kodni knjigi poiščemo najdaljšo kodo, ki se ujema z danim nizom na trenutnem mestu in ustrezen del niza zakodiramo z zaporedno številko kode:

A	A	B	B	A	B	A	A	A	B	A	A	B	A	A	B	A	B	A	B
0																			

V kodno knjigo dodamo novo kodo tako, da trenutni kodi dodamo naslednji znak iz niza:

A	A	B	B	A	B	A	A	A	B	A	A	B	A	A	B	A	B	A	B
0																			

0	A
1	B
2	AA

A	A	B	B	A	B	A	A	A	B	A	A	B	A	A	B	A	B	A	B
0																			

0	A
1	B
2	AA

Koraka ponavljamo, dokler ne pridemo do konca niza:

A	A	B	B	A	B	A	A	A	B	A	A	B	A	A	B	A	B	A	B
0	0																		

0	A
1	B
2	AA
3	AB

A	A	B	B	A	B	A	A	A	B	A	A	B	A	A	B	A	B	A	B
0	0	1																	

0	A
1	B
2	AA
3	AB
4	BB

A	A	B	B	A	B	A	A	A	B	A	A	B	A	A	B	A	B	A	B
0	0	1	1																

0	A
1	B
2	AA
3	AB
4	BB
5	BA

A	A	B	B	AB	A	A	A	B	A	A	B	A	A	B	A	B	A	B
0	0	1	1	3														

0	A
1	B
2	AA
3	AB
4	BB
5	BA
6	ABA

A	A	B	B	AB	AA	A	B	A	A	B	A	A	B	A	B	A	B
0	0	1	1	3	2												

0	A
1	B
2	AA
3	AB
4	BB
5	BA
6	ABA
7	AAA

A	A	B	B	AB	AA	ABA	A	B	A	A	B	A	B	A	B
0	0	1	1	3	2	6									

0	A
1	B
2	AA
3	AB
4	BB
5	BA
6	ABA
7	AAA
8	ABAA

A	A	B	B	AB	AA	ABA	ABAA	B	A	B	A	B
0	0	1	1	3	2	6	8					

0	A
1	B
2	AA
3	AB
4	BB
5	BA
6	ABA
7	AAA
8	ABAA
9	ABAAB

A	A	B	B	AB	AA	ABA	ABAA	BA	B	A	B
0	0	1	1	3	2	6	8	5			

0	A
1	B
2	AA
3	AB
4	BB
5	BA
6	ABA
7	AAA
8	ABAA
9	ABAAB
10	BAB

A	A	B	B	AB	AA	ABA	ABAA	BA	BAB
0	0	1	1	3	2	6	8	5	10

0	A
1	B
2	AA
3	AB
4	BB
5	BA
6	ABA
7	AAA
8	ABAA
9	ABAAB
10	BAB

Zakodiran niz: 0 0 1 1 3 2 6 8 5 10
Začetna kodnja kniga za dekodiranje:

0	A
1	B

S pomočjo kodne knjige dekodiramo prvi znak:

0	0 1 1 3 2 6 8 5 10
A	

S pomočjo kodne knjige dekodiramo drugi znak in dodamo novo kodo v kodno knjigo. Nova koda je sestavljena iz kode, ki je bila dekodirana v prejšnjem koraku in prvega znaka kode, ki je bila odkodirana v tem koraku:

0	0	1 1 3 2 6 8 5 10
A	A	

0	A
1	B
2	AA

Korak 3 ponavljamo, dokler ne pridemo do konca zakodiranega niza:

0	0	1	1 3 2 6 8 5 10
A	A	B	

0	A
1	B
2	AA
3	AB

0	0	1	1	3 2 6 8 5 10
A	A	B	B	

0	A
1	B
2	AA
3	AB
4	BB

0	0	1	1	3	2 6 8 5 10
A	A	B	B	AB	

0	A
1	B
2	AA
3	AB
4	BB
5	BA

0	0	1	1	3	2	6 8 5 10
A	A	B	B	AB	AA	

0	A
1	B
2	AA
3	AB
4	BB
5	BA
6	ABA

0	0	1	1	3	2	6	8 5 10
A	A	B	B	AB	AA	ABA	

0	A
1	B
2	AA
3	AB
4	BB
5	BA
6	ABA
7	AAA

0	0	1	1	3	2	6	8	5 10
A	A	B	B	AB	AA	ABA	???	

0	A
1	B
2	AA
3	AB
4	BB
5	BA
6	ABA
7	AAA
8	???

Pri dekodiranju občasno pride do problema, da za dekodiranje potrebujemo kodo, ki je pravkar v nastajanju. Nova koda bo sestavljena iz prejšnje kode (ABA) in prvega znaka same sebe (prvi zank od ???):

0	0	1	1	3	2	6	8	5	10
A	A	B	B	AB	AA	ABA	ABA?		

0	A
1	B
2	AA
3	AB
4	BB
5	BA
6	ABA
7	AAA
8	ABA?

Ker je zadnji znak kode enak prvemu znaku, lahko tudi to kodo ugotovimo:

0	0	1	1	3	2	6	8	5	10
A	A	B	B	AB	AA	ABA	ABAA		

0	A
1	B
2	AA
3	AB
4	BB
5	BA
6	ABA
7	AAA
8	ABAA

0	0	1	1	3	2	6	8	5	10
A	A	B	B	AB	AA	ABA	ABAA	BA	

0	A
1	B
2	AA
3	AB
4	BB
5	BA
6	ABA
7	AAA
8	ABAA
9	ABAAB

0	A
1	B
2	AA
3	AB
4	BB
5	BA
6	ABA
7	AAA
8	ABAA
9	ABAAB
10	BAB

0	0	1	1	3	2	6	8	5	10
A	A	B	B	AB	AA	ABA	ABAA	BA	BAB

Dekodiran niz: A A B B A B A A A B A A B A A B A B A B

Naloga 6.5: Podani niz zakodirajte z LZW algoritmom. Dobljeni niz nato ponovno dekodirajte.
Niz: ABADBCDBABCDACCABDBD
Abeceda: A, B, C, D

Rešitev:

A	B	A	D	B	C	DB	AB	CD	A	C	C	AB	DB	D
0	1	0	3	1	2	7	4	9	0	2	2	4	7	3

Kodirna knjiga:

0	A
1	B
2	C
3	D
4	AB
5	BA
6	AD
7	DB
8	BC
9	CD
10	DBA
11	ABC
12	CDA
13	AC
14	CC
15	CA
16	ABD
17	DBD

Zakodirani niz: 0 1 0 3 1 2 7 4 9 0 2 2 4 7 3
Dekodiranje:

0	1	0	3	1	2	7	4	9	0	2	2	4	7	3
A	B	A	D	B	C	DB	AB	CD	A	C	C	AB	DB	D

Kodna knjiga:

0	A
1	B
2	C
3	D
4	AB
5	BA
6	AD
7	DB
8	BC
9	CD
10	DBA
11	ABC
12	CDA
13	AC
14	CC
15	CA
16	ABD
17	DBD

Naloga 6.6: S pomočjo aritmetičnega kodiranja zakodirajte niz $X_2, X_3, X_3, X_1, X_{end}$. Verjetnosti posameznih simbolov so $[\frac{4}{9}, \frac{1}{9}, \frac{3}{9}, \frac{1}{9}]$.

Rešitev: Pri aritmetičnem kodiranju vzamemo trenutni interval, ter ga razdelimo na podintervale glede na verjetnosti simbolov. Nato izberemo podinterval, ki ustreza simbolu, ki ga kodiramo, in postopek ponavljamo, dokler ne zakodiramo celotnega niza.

Korak 1, člen: 2, začetni interval: [0.00000, 1.00000]

Simbol	Podinterval
X_1	[0.00000, 0.44444]
X_2	[0.44444, 0.55556] <== izbran
X_3	[0.55556, 0.88889]
X_{end}	[0.88889, 1.00000]

Korak 2, člen: 3, začetni interval: [0.44444, 0.55556]

Simbol	Podinterval
X_1	[0.44444, 0.49383]
X_2	[0.49383, 0.50617]
X_3	[0.50617, 0.54321] <== izbran
X_{end}	[0.54321, 0.55556]

Korak 3, člen: 3, začetni interval: [0.50617, 0.54321]

Simbol	Podinterval
X_1	[0.50617, 0.52263]
X_2	[0.52263, 0.52675]
X_3	[0.52675, 0.53909] <== izbran
X_{end}	[0.53909, 0.54321]

Korak 4, člen: 1, začetni interval: [0.52675, 0.53909]

Simbol	Podinterval
X_1	[0.52675, 0.53224] <== izbran
X_2	[0.53224, 0.53361]
X_3	[0.53361, 0.53772]
X_{end}	[0.53772, 0.53909]

Korak 5, člen: 4, začetni interval: [0.52675, 0.53224]

Simbol	Podinterval
X_1	[0.52675, 0.52919]
X_2	[0.52919, 0.52980]
X_3	[0.52980, 0.53163]
X_{end}	[0.53163, 0.53224] <== izbran

Kodirano število (sredina zadnjega intervala): 0.5319311081

Naloga 6.7: S pomočjo aritmetičnega kodiranja smo dobili 0.5319311081. Ob znanih verjetnosti simbolov $[\frac{4}{9}, \frac{1}{9}, \frac{3}{9}, \frac{1}{9}]$ dekodirajte zaporedje.

Rešitev: V resnici je postopek aritmetičnega dekodiranja praktično enak kodiranju. Vsakič trenutni interval razdelimo na podintervale glede na verjetnosti simbolov, nato pa izberemo podinterval v katerega pade kodirana vrednost. Ta interval ustreza simbolu, ki je bil zakodiran.

Korak 1, vrednost: 0.5319311081, interval: [0.00000, 1.00000]

Simbol	Podinterval
X_1	[0.00000, 0.44444]
X_2	[0.44444, 0.55556]<== izbran
X_3	[0.55556, 0.88889]
X_{end}	[0.88889, 1.00000]

Korak 2, vrednost: 0.5319311081, interval: [0.44444, 0.55556]

Simbol	Podinterval
X_1	[0.44444, 0.49383]
X_2	[0.49383, 0.50617]
X_3	[0.50617, 0.54321] <== izbran
X_{end}	[0.54321, 0.55556]

Korak 3, vrednost: 0.5319311081, interval: [0.50617, 0.54321]

Simbol	Podinterval
X_1	[0.50617, 0.52263]
X_2	[0.52263, 0.52675]
X_3	[0.52675, 0.53909] <== izbran
X_{end}	[0.53909, 0.54321]

Korak 4, vrednost: 0.5319311081, interval: [0.52675, 0.53909]

Simbol	Podinterval
X_1	[0.52675, 0.53224] <== izbran
X_2	[0.53224, 0.53361]
X_3	[0.53361, 0.53772]
X_{end}	[0.53772, 0.53909]

Korak 5, vrednost: 0.5319311081, interval: [0.52675, 0.53224]

Simbol	Podinterval
X_1	[0.52675, 0.52919]
X_2	[0.52919, 0.52980]
X_3	[0.52980, 0.53163]
X_{end}	[0.53163, 0.53224] <== izbran

Prepoznan stop znak (X_{end}) – dekodiranje zaključeno.

Zakodirana sekvenca je bila $X_2, X_3, X_3, X_1, X_{end}$.

$H_{vir} = -\sum_{i=1}^{n} P_x \log_2 P_x$	Entropija informacijskega vira
$H_{MAX} = \log_2(N)$	Maksimalna entropija vira z N simboli
$\bar{n} = \sum_{i=1}^{n} P_x \cdot n_x$	Povprečna dolžina kodnih besed
$R = \frac{\bar{n} - H_{vir}}{\bar{n}}$	Relativna kodirna redundanca
$\eta = \frac{H_{vir}}{\bar{n}}$	Učinkovitost kodiranja

Povzetek postopkov kodiranja

Huffmanovo kodiranje

1. Pridobi verjetnosti (ali relativne frekvence) vseh simbolov.
2. Uredi simbole po naraščajoči verjetnosti.
3. Združi dve najmanj verjetni vrednosti v nov simbol in po potrebi preuredi simbole.

4. Postopek ponavljaj, dokler ne ostaneta samo dva simbola.

5. Dodeli binarne kode tako, da vsak zgornji odcep pomeni 1, spodnji 0.

LZW kodiranje

1. Inicializiraj slovar s posameznimi simboli vhodnega abecednega nabora. Vsak simbol dobi svojo kodo.

2. Preberi zaporedje simbolov in poišči najdaljši niz, ki je že v slovarju.

3. Izdaj kodo za ta niz.

4. Dodaj nov vnos v slovar: trenutni niz + naslednji simbol.

5. Ponovi postopek, dokler ni celoten vhod obdelan.

Opomba: Pri dekodiranju je pomembno, da dekoder uporablja enak slovar kot kodirnik, ki ga gradi sproti in na enak način.

LZW dekodiranje

1. Inicializiraj slovar s posameznimi simboli vhodnega abecednega nabora. Vsak simbol dobi svojo kodo.

2. Preberi prvo kodo iz zaporedja in izpiši ustrezni simbol iz slovarja.

3. Nastavi ta simbol kot začetni dekodirani niz.

4. Za vsako naslednjo kodo v zaporedju:

   - Če je koda v slovarju, pridobi pripadajoči niz.
   - Če kode še ni v slovarju, sestavi nov niz tako, da prejšnjemu dekodiranemu nizu dodaš njegov prvi znak.
   - Izpiši dobljeni niz.
   - Dodaj v slovar nov vnos: prejšnji dekodirani niz, dopolnjen s prvim znakom trenutnega niza.
   - Nastavi trenutni niz kot novi dekodirani niz za naslednji korak.

5. Postopek ponavljaj, dokler ne predelaš vseh kod iz zaporedja.

Aritmetično kodiranje

1. Določi verjetnosti vseh simbolov v zaporedju.

2. Nastavi celotni interval na $[0, 1]$.

3. Intervale razdeli na podintervale za posamezne simbole glede na njihovo verjetnost.

4. Za vsak simbol v zaporedju:

   - Izberi podinterval, ki ustreza simbolu.
   - Celotni interval zožaj na ta podinterval.

   - Ponovi od tretjega koraka dalje, dokler niso uporabljeni vsi simboli.

5. Končni rezultat je poljubno število iz končnega intervala (običajno sredina).
6. Za dekodiranje se število iterativno razširi nazaj v posamezne simbole, glede na znane verjetnosti in intervale.

7. POGLAVJE

VZAJEMNA INFORMACIJA

V tem poglavju bomo raziskali vzajemno informacijo kot kvantitativno mero odvisnosti med dvema naključnima spremenljivkama. Ta pojem je ključen pri analizi informacijskih kanalov, saj nam omogoča ovrednotenje količine informacije, ki jo prejmemo o vhodni spremenljivki, če poznamo izhodno.

Za dve diskretni naključni spremenljivki X in Y je vzajemna informacija definirana kot:

$$I(X;Y) = \sum_{x,y} P(x,y) \log_2 \frac{P(x,y)}{P(x)P(y)}$$

Vzajemna informacija je simetričena in vedno nenegativna. Vzajemna informacija meri, koliko se negotovost o X zmanjša, ko opazujemo Y, kar je formalno:

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

Poleg teoretične osnove bomo v tem poglavju obravnavali:

- primer dogodkov in njihove informacije,
- binarni simetrični kanal z napakami,
- kanal s podano matriko pogojnih verjetnosti,
- modeliranje analognega prenosa preko diskretizacije (Gaussov šum in dekodiranje).

Za dodatno branje priporočamo naslednje vire:

- T. M. Cover, J. A. Thomas: *Elements of Information Theory* [2]
- D. J. C. MacKay: *Information Theory, Inference, and Learning Algorithms* [3]
- N. Pavešič: *Informacija in kodi* [6]

7.1 Naloge

Naloga 7.1: Mečemo pošteno igralno kocko. Dogodek A je, da pade št. 3, dogodek B, da pade liha številka in dogodek C, da pade številka večja od 3. Določite informacijo vsakega dogodka in vzajemno informacijo dogodkov A in B ter vzajemno informacijo dogodkov B in C.

Rešitev:

$$
\begin{aligned}
S &= \{1,2,3,4,5,6\} \\
A &= \{3\}, \quad P(A) = \frac{1}{6} \\
B &= \{1,3,5\}, \quad P(B) = \frac{3}{6} \\
C &= \{4,5,6\}, \quad P(C) = \frac{3}{6} \\
I(A) &= -\log_2 P(A) = -\log_2 \frac{1}{6} = 2{,}58 \\
I(B) &= -\log_2 P(B) = -\log_2 \frac{1}{2} = 1 \\
I(C) &= -\log_2 P(C) = -\log_2 \frac{1}{2} = 1 \\
I(A;B) &= I(A) + I(B) - I(A \cap B) \\
A \cap B &= \{3\} \\
P(A \cap B) &= \frac{1}{6} \\
I(A \cap B) &= -\log_2 P(A \cap B) = -\log_2 \frac{1}{6} = 2{,}58 \\
I(A;B) &= I(A) + I(B) - I(A \cap B) = 2{,}58 + 1 - 2{,}58 = 1 \\
B \cap C &= \{5\}, \quad P(B \cap C) = \frac{1}{6}, \quad I(B \cap C) = 2{,}58 \\
I(B;C) &= I(B) + I(C) - I(B \cap C) = 1 + 1 - 2{,}58 = -0{,}58
\end{aligned}
$$

V tej nalogi smo uporabili izraz:

$I(A;B) = I(A) + I(B) - I(A \cap B)$

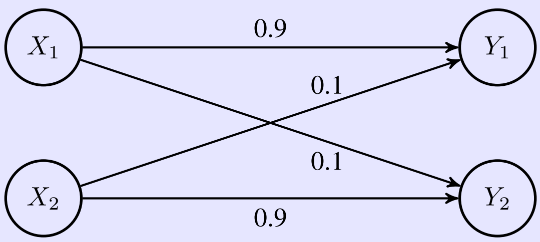
pri čemer je $I(A) = \log_2 \frac{1}{P(A)}$ t. i. "informacija dogodka". Vendar opozorimo, da to **ni formalna definicija vzajemne informacije**, kot jo sicer uporabljamo v teoriji informacij za naključne spremenljivke. Vzajemna informacija med dogodkoma bi bila ustrezno definirana šele, če bi definirali spremenljivki, ki zavzemata vrednosti 1 (če se dogodek zgodi) ali 0 (če se ne zgodi), nato pa izračunali entropiji in skupno entropijo teh spremenljivk:

$I(A;B) = H(A) + H(B) - H(A,B)$

V tem primeru pa smo raje uporabili izraz, ki se intuitivno bere kot "kolikšen presežek informacije bi potrebovali, če bi kodirali dogodka posebej, v primerjavi s skupnim kodiranjem". Zaradi tega lahko pride tudi do negativnih vrednosti izraza $I(A;B)$, kar za pravo vzajemno informacijo nikoli ne velja. Ta pristop je torej zgolj intuitivna ilustracija presežne informacije in **ni enakovreden formalni vzajemni informaciji med naključnima spremenljivkama**. Vzajemna informacija dveh dogodkov (izidov) je lahko pozitivna ali negativna, medtem ko je povprečna vzajemna informacija dveh naključnih spremenljivk vedno pozitivna ali enaka nič.

Naloga 7.2: Verjetnost za napako na simetričnem binarnem kanalu je $0{,}1$. Verjetnost za simbol x_1 vhodne naključne spremenljivke znaša $P_X(x_1 1) = 0{,}3$. Določite:

- informacijo obeh simbolov vhodne in izhodne spremenljivke,
- entropijo vhodne naključne spremenljivke X in izhodne spremenljivke Y,
- vzajemno informacijo vseh kombinacij simbolov vhodne in izhodne naključne spremenljivke,
- povprečno preneseno informacijo,
- pogojni entropiji $H(Y|X)$ in $H(X|Y)$ ter
- vezano entropijo $H(X,Y)$.



Binarni kanal

Rešitev: Informacija simbolov:

$$I(x_1) = -\log_2 0{,}3 = 1{,}74$$

$$I(x_2) = -\log_2 0{,}7 = 0{,}51$$

$$\begin{aligned}
P(y_1) &= P(x_1) \cdot 0{,}9 + P(x_2) \cdot 0{,}1 = 0{,}3 \cdot 0{,}9 + 0{,}7 \cdot 0{,}1 = 0{,}34 \\
P(y_2) &= P(x_1) \cdot 0{,}1 + P(x_2) \cdot 0{,}9 = 0{,}3 \cdot 0{,}1 + 0{,}7 \cdot 0{,}9 = 0{,}66 \\
I(y_1) &= -\log_2 0{,}34 = 1{,}56 \\
I(y_2) &= -\log_2 0{,}66 = 0{,}60
\end{aligned}$$

$$\begin{aligned}
H(X) &= P(x_1) \cdot I(x_1) + P(x_2) \cdot I(x_2) = 0{,}3 \cdot 1{,}74 + 0{,}7 \cdot 0{,}51 = 0{,}88 \\
H(Y) &= P(y_1) \cdot I(y_1) + P(y_2) \cdot I(y_2) = 0{,}34 \cdot 1{,}56 + 0{,}66 \cdot 0{,}60 = 0{,}93
\end{aligned}$$

$$\begin{aligned}
I(x_1; y_1) &= I(x_1) + I(y_1) - I(x_1 \cap y_1) \\
P(x_1 \cap y_1) &= P(y_1 \mid x_1) \cdot P(x_1) = 0{,}9 \cdot 0{,}3 = 0{,}27 \\
I(x_1 \cap y_1) &= -\log_2 0{,}27 = 1{,}89 \\
I(x_1; y_1) &= 1{,}74 + 1{,}56 - 1{,}89 = 1{,}41
\end{aligned}$$

$$\begin{aligned}
P(x_1 \cap y_2) &= P(y_2 \mid x_1) \cdot P(x_1) = 0{,}1 \cdot 0{,}3 = 0{,}03, \quad I(x_1 \cap y_2) = 5{,}06 \\
P(x_2 \cap y_1) &= P(y_1 \mid x_2) \cdot P(x_2) = 0{,}1 \cdot 0{,}7 = 0{,}07, \quad I(x_2 \cap y_1) = 3{,}84 \\
P(x_2 \cap y_2) &= P(y_2 \mid x_2) \cdot P(x_2) = 0{,}9 \cdot 0{,}7 = 0{,}63, \quad I(x_2 \cap y_2) = 0{,}67 \\
I(x_1; y_2) &= 1{,}74 + 0{,}60 - 5{,}06 = -2{,}72 \\
I(x_2; y_1) &= 0{,}51 + 1{,}56 - 3{,}84 = -1{,}77 \\
I(x_2; y_2) &= 0{,}51 + 0{,}60 - 0{,}67 = 0{,}44
\end{aligned}$$

$$\begin{aligned}
I(X; Y) &= \sum_j \sum_k P_{XY}(x_j, y_k) \cdot I(x_j; y_k) \\
I(X; Y) &= P(x_1 \cap y_1) \cdot I(x_1; y_1) + P(x_1 \cap y_2) \cdot I(x_1; y_2) \\
&\quad + P(x_2 \cap y_1) \cdot I(x_2; y_1) + P(x_2 \cap y_2) \cdot I(x_2; y_2) \\
I(X; Y) &= 0{,}27 \cdot 1{,}41 - 0{,}03 \cdot 2{,}72 - 0{,}07 \cdot 1{,}77 + 0{,}63 \cdot 0{,}44 = 0{,}45
\end{aligned}$$

$$\begin{aligned}
H(Y \mid X) &= \sum_j P_X(x_j) \cdot H(Y \mid x_j) \\
H(Y \mid x_1) &= \sum_k P(y_k \mid x_1) \cdot I(y_k \mid x_1) \\
&= -P(y_1 \mid x_1) \cdot \log_2 P(y_1 \mid x_1) - P(y_2 \mid x_1) \cdot \log_2 P(y_2 \mid x_1) \\
H(Y \mid x_1) &= -0{,}9 \cdot \log_2 0{,}9 - 0{,}1 \cdot \log_2 0{,}1 = 0{,}47 \\
H(Y \mid x_2) &= \sum_k P(y_k \mid x_2) \cdot I(y_k \mid x_2) \\
&= -P(y_1 \mid x_2) \cdot \log_2 P(y_1 \mid x_2) - P(y_2 \mid x_2) \cdot \log_2 P(y_2 \mid x_2)
\end{aligned}$$

$$H(Y \mid x_2) = -0{,}1 \cdot \log_2 0{,}1 - 0{,}9 \cdot \log_2 0{,}9 = 0{,}47$$

$$H(Y \mid X) = \sum_j P_X(x_j) \cdot H(Y \mid x_j)$$

$$= P_X(x_1) \cdot H(Y \mid x_1) + P_X(x_2) \cdot H(Y \mid x_2)$$

$$H(Y \mid X) = 0{,}3 \cdot 0{,}47 + 0{,}7 \cdot 0{,}47 = 0{,}47$$

$$H(X \mid Y) = \sum_j P_Y(y_j) \cdot H(X \mid y_j)$$

$$H(X \mid y_1) = \sum_k P(x_k \mid y_1) \cdot I(x_k \mid y_1)$$

$$= -P(x_1 \mid y_1) \cdot \log_2 P(x_1 \mid y_1) - P(x_2 \mid y_1) \cdot \log_2 P(x_2 \mid y_1)$$

$$P(x_1 \cap y_1) = P(y_1 \mid x_1) \cdot P(x_1) = P(x_1 \mid y_1) \cdot P(y_1)$$

$$P(x_1 \mid y_1) = \frac{P(x_1 \cap y_1)}{P(y_1)} = \frac{0{,}27}{0{,}34} = 0{,}79$$

$$P(x_1 \mid y_2) = \frac{0{,}03}{0{,}66} = 0{,}045$$

$$P(x_2 \mid y_1) = \frac{0{,}07}{0{,}34} = 0{,}21$$

$$P(x_2 \mid y_2) = \frac{0{,}63}{0{,}66} = 0{,}95$$

$$H(X \mid y_1) = -0{,}79 \cdot \log_2 0{,}79 - 0{,}21 \cdot \log_2 0{,}21 = 0{,}74$$

$$H(X \mid y_2) = \sum_k P(x_k \mid y_2) \cdot I(x_k \mid y_2)$$

$$= -P(x_1 \mid y_2) \cdot \log_2 P(x_1 \mid y_2) - P(x_2 \mid y_2) \cdot \log_2 P(x_2 \mid y_2)$$

$$H(X \mid y_2) = -0{,}045 \cdot \log_2 0{,}045 - 0{,}95 \cdot \log_2 0{,}095 = 0{,}27$$

$$H(X \mid Y) = \sum_j P_Y(y_j) \cdot H(X \mid y_j) = P_Y(y_1) \cdot H(X \mid y_1) + P_Y(y_2) \cdot H(X \mid y_2)$$

$$H(X \mid Y) = 0{,}34 \cdot 0{,}74 + 0{,}66 \cdot 0{,}27 = 0{,}43$$

$$H(X, Y) = \sum_j \sum_k P_{XY}(x_j, y_k) \cdot I_{XY}(x_j, y_k)$$

$$H(X, Y) = H(X) + H(Y) - I(X;Y) = 0{,}88 + 0{,}93 - 0{,}45 = 1{,}36$$

Opomba: Pri izračunu entropij in vzajemne informacije smo uporabili znane izraze za diskretne naključne spremenljivke:

$$I(X;Y) = H(X) + H(Y) - H(X, Y)$$

Ker imamo podano matriko pogojnih verjetnosti $P(Y|X)$ in porazdelitev $P(X)$, lahko vse ostale verjetnosti (npr. $P(X, Y)$, $P(Y)$) izrazimo preko produktov in vsot. To ustreza modelu sporočilnega kanala, kot ga definira teorija informacij. Kljub temu, da je kanal simetričen, rezultat ni popolnoma enak za vse vrednosti, saj vhodna porazdelitev $P(X)$ ni enakomerna. Ta naloga lepo ponazori, kako se napake v kanalu odražajo kot zmanjšanje vzajemne informacije.

Naloga 7.3: Za brez spominski informacijski kanal je podana matrika pogojnih verjetnosti:

$$\mathbf{P_K} = \begin{bmatrix} 0{,}8 & 0{,}1 & 0{,}05 & 0{,}05 \\ 0{,}03 & 0{,}85 & 0{,}02 & 0{,}1 \\ 0{,}03 & 0{,}05 & 0{,}9 & 0{,}02 \\ 0{,}06 & 0{,}06 & 0{,}08 & 0{,}8 \end{bmatrix}$$

Za vhodno naključno spremenljivko predpostavite enakomerno porazdelitev. Določite:

- entropijo vhodne naključne spremenljivke X in izhodne spremenljivke Y,
- vezano entropijo $H(X,Y)$,
- povprečno preneseno informacijo $I(Y;X)$ ter
- pogojni entropiji $H(Y|X)$ in $H(X|Y)$.

Rešitev: Vhodna naključna spremenljivka X ima štiri simbole z enako verjetnostjo $0,25$:

$$P_X = [0{,}25 \quad 0{,}25 \quad 0{,}25 \quad 0{,}25]$$

Verjetnost hkratnega nastopa določene kombinacije simbolov spremenljivk X in Y lahko določimo iz pogojne verjetnosti $P_{Y|X}$ in verjetnosti P_X:

$$P_{XY}(x_i, y_j) = P_{Y|X}(x_i, y_j) \cdot P_X(x_i)$$

$$P_{XY} = \begin{bmatrix} 0{,}2 & 0{,}025 & 0{,}0125 & 0{,}0125 \\ 0{,}0075 & 0{,}2125 & 0{,}005 & 0{,}025 \\ 0{,}0075 & 0{,}0125 & 0{,}225 & 0{,}005 \\ 0{,}015 & 0{,}015 & 0{,}02 & 0{,}2 \end{bmatrix}$$

Marginalno verjetnost P_Y določimo s seštevanjem posameznih stolpcev vezane verjetnosti P_{XY}:

$$P_Y = [0{,}23 \quad 0{,}265 \quad 0{,}2625 \quad 0{,}2425]$$

Iz marginalnih verjetnosti P_X in P_Y določimo entropiji naključnih spremenljivk X in Y:

$H(X) = 2$
$H(Y) = 1.9976$

Vezano entropijo izračunamo iz vezane verjetnosti P_{XY}:

$$\begin{aligned} H(X,Y) &= \sum_i \sum_j P_{XY}(x_i, y_j) \cdot I_{XY}(x_i, y_j) \\ &= -\sum_i \sum_j P_{XY}(x_i, y_j) \cdot \log_2 \left(P_{XY}(x_i, y_j)\right) = 2{,}8679 \end{aligned}$$

Povprečno preneseno informacijo lahko izračunamo iz entropij posameznih naključnih spremenljivk in vezane entropije.

$$I(X;Y) = H(X) + H(Y) - H(X,Y) = 2 + 1{,}9976 - 2{,}8679 = 1{,}1296$$

Pogojno entropijo lahko izrazimo iz povprečne prenesene informacije in entropije posamezne naključne spremenljivke.

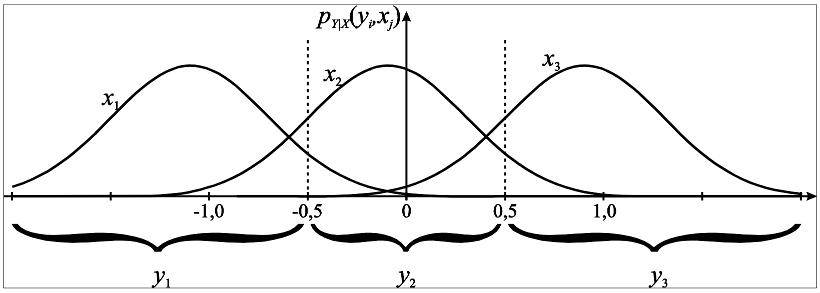
$$I(X;Y) = H(X) - H(X|Y) = I(Y;X) = H(Y) - H(Y|X)$$
$$H(X|Y) = H(X) - I(X;Y) = 2 - 1{,}1296 = 0{,}8704$$
$$H(Y|X) = H(Y) - I(X;Y) = 1{,}9976 - 1{,}1296 = 0{,}8679$$

Naloga 7.4: Tri-nivojski signal z nivoji $-1V$, $0V$ in $1V$ ima dodan beli Gaussov šum z efektivno vrednostjo $0{,}4V$ in povprečno vrednostjo $-0{,}1V$. Linijski dekodirnik interpretira napetosti pod $-0{,}5V$ kot simbol y_1, napetosti med $-0{,}5V$ in $0{,}5V$ kot y_2 in napetosti nad $0{,}5V$ kot y_3. Določite matriko pogojnih verjetnosti P_K informacijskega kanala.

Rešitev: Matrika pogojnih verjetnosti informacijskega kanala določa verjetnosti sprejetih simbolov pri določenih oddanih simbolih:

$$P_K = \begin{bmatrix} P(y_1 \mid x_1) & \cdots & P(y_n \mid x_1) \\ \vdots & \ddots & \vdots \\ P(y_1 \mid x_n) & \cdots & P(y_n \mid x_n) \end{bmatrix}$$

Pogojne verjetnosti za dani primer so ilustrirane na spodnji sliki:



Verjetnosti posameznih izhodnih simbolov pri danem vhodnem simbolu določimo s pomočjo tabele normirane normalne porazdelitve. V ta namen normiramo razdaljo od sredine Gaussove krivulje do meje med simboli. Kjer tabela ne zadostuje, uporabimo približek $Q(x)$.

$$P(y_1 \mid x_1) = \Phi\left(\frac{0{,}6}{0{,}4}\right) = \Phi(1{,}5) = 0{,}93319$$

$$P(y_2 \mid x_1) = \Phi\left(\frac{1{,}6}{0{,}4}\right) - \Phi\left(\frac{0{,}6}{0{,}4}\right) = \Phi(4) - \Phi(1{,}5) = 1 - Q(4) - \Phi(1{,}5)$$

$$Q(x) = \frac{1}{x\sqrt{2\pi}} e^{-x^2/2}, \quad Q(4) = 3{,}35 \cdot 10^{-5}$$

$$P(y_2 \mid x_1) = 1 - 3{,}35 \cdot 10^{-5} - 0{,}93319 = 0{,}06678$$

$$P(y_3 \mid x_1) = F(\infty) - F\left(\frac{1{,}6}{0{,}4}\right) = 1 - F(4) = 1 - 1 + Q(4) = 3{,}35 \cdot 10^{-5}$$

$$P(y_1 \mid x_2) = 1 - \Phi\left(\frac{0{,}4}{0{,}4}\right) = 1 - \Phi(1) = 1 - 0{,}84134 = 0{,}15866$$

$$\begin{aligned} P(y_2 \mid x_2) &= \Phi\left(\frac{0{,}6}{0{,}4}\right) - \left(1 - \Phi\left(\frac{0{,}4}{0{,}4}\right)\right) = \Phi(1{,}5) - 1 + \Phi(1) \\ &= 0{,}93319 - 1 + 0{,}84134 = 0{,}77453 \end{aligned}$$

$$P(y_3 \mid x_2) = 1 - \Phi\left(\frac{0{,}6}{0{,}4}\right) = 1 - \Phi(1{,}5) = 1 - 0{,}93319 = 0{,}06681$$

$$P(y_1 \mid x_3) = 1 - \Phi\left(\frac{1{,}4}{0{,}4}\right) = 1 - \Phi(3{,}5) = 1 - 0{,}99977 = 0{,}00023$$

$$\begin{aligned} P(y_2 \mid x_3) &= \Phi\left(\frac{1{,}4}{0{,}4}\right) - \Phi\left(\frac{0{,}4}{0{,}4}\right) = \Phi(3{,}5) - \Phi(1) \\ &= 0{,}99977 - 0{,}84134 = 0{,}15843 \end{aligned}$$

$$P(y_3 \mid x_3) = \Phi\left(\frac{0{,}4}{0{,}4}\right) = \Phi(1) = 0{,}84134$$

$$P_K = \begin{bmatrix} 0{,}93319 & 0{,}06678 & 0{,}00003 \\ 0{,}15866 & 0{,}77453 & 0{,}06681 \\ 0{,}00023 & 0{,}15843 & 0{,}84134 \end{bmatrix}$$

Opomba: V tej nalogi smo prešli iz zveznega v diskreten model s tem, da smo analogne vrednosti šuma diskretizirali v tri izhodne razrede. To pomeni, da obravnavamo sistem kot kanal z diskretnimi izhodi in zveznimi napakami (Gaussov šum). Za določitev verjetnosti $P(Y|X)$ smo uporabili porazdelitveno funkcijo normalne porazdelitve:

$$P(y_j|x_i) = \int_a^b \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$$

pri čemer so meje a, b odvisne od dekodirnih pragov in trenutne vrednosti x_i. Ta pristop je običajen pri modeliranju kanalov, kjer je vhod simbolen, šum pa analogni — in temelji na numerični aproksimaciji z uporabo Z-tabele ali numerične integracije.

$I(x) = \log_2 \frac{1}{P(x)}$	Informacija posameznega dogodka
$I(A;B) = I(A) + I(B) - I(A \cap B)$	Vzajemna informacija dveh dogodkov
$I(X;Y) = H(X) + H(Y) - H(X,Y)$	Vzajemna informacija dveh spremenljivk
$H(Y\|X) = H(Y) - I(X;Y)$	Pogojna entropija
$P(x \leq A) = \Phi(\frac{A-\mu}{\sigma})$	Verjetnost, da je naključna spremenljivka manjša od A v normalni porazdelitvi z efektivno vrednostjo σ in povprečjem μ
$P(x \geq A) = 1 - P(x \leq A)$	Verjetnost, da je naključna spremenljivka večja od A v normalni porazdelitvi
$P(x \leq -A) = P(x \geq A)$	Verjetnost, da je naključna spremenljivka manjša od $-A$ v normalni porazdelitvi

8. POGLAVJE

KAPACITETA KANALA

V tem poglavju bomo raziskali, kako količinsko opišemo zmožnost komunikacijskega kanala za prenos informacij. Glavni cilj je določiti ***kapaciteto kanala*** — največjo količino informacij, ki jo lahko kanal prenese brez napak.

Obravnavali bomo različne tipe kanalov:

- splošne diskretne kanale z matriko prehodnih verjetnosti,
- binarni simetrični kanal (BSC),
- vpliv napake na prenosno zmogljivost.

Posebno pozornost bomo namenili pojmom:

- entropija izhodne spremenljivke $H(Y)$,
- pogojna entropija $H(Y|X)$,
- vzajemna informacija $I(X;Y)$,
- kapaciteta $C = \max I(X;Y)$.

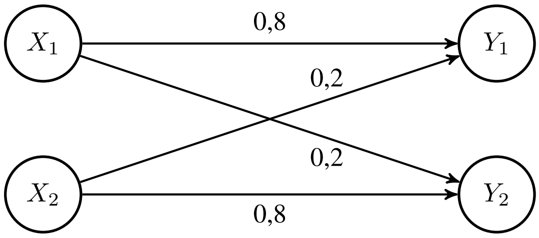
Razumevanje teh konceptov je temeljno za zasnovo učinkovitih kodirnikov in ocenjevanje izgub v prenosnih sistemih. Za dodatno branje priporočamo naslednje vire:

- C. E. Shannon: *A Mathematical Theory of Communication* [4]
- T. M. Cover, J. A. Thomas: *Elements of Information Theory* [2]
- S. Tomažič: *Osnove telekomunikacij I* [5]

8.1 Naloge

Naloga 8.1: Določite kapaciteto simetričnega binarnega kanala z verjetnostjo napake 0,2.

Rešitev: Skica simetričnega binarnega kanala:



Kapaciteta kanala je maksimalna možna vzajemna informacija vhodne in izhodne spremenljivke. Vzajemna informacija je odvisna od kanala in verjetnostne porazdelitve vhodne naključne spremenljivke. Vzajemno informacijo izrazimo kot funkcijo verjetnostne porazdelitve in poiščemo njen maksimum.

$$I(X;Y) = H(X) + H(Y) - H(X,Y) = \sum_i \sum_j P_{XY}(x_i, y_j) \cdot I(x_i; y_j)$$

Povprečno preneseno informacijo lahko izračunamo kot razliko vsote entropij vhodne in izhodne spremenljivke ter vezano entropijo ali kot povprečno vzajemno informacijo posameznih parov vhodnih in izhodnih simbolov. Za iskanje ekstrema je ugodneje, če je izraz za povprečno preneseno informacijo zapisan čim krajše, zato ga raje zapišemo kot povprečno vzajemno informacijo.

$$I(X;Y) = \sum_i \sum_j P_{XY}(x_i, y_j) \cdot \log_2 \left(\frac{P(y_j \mid x_i)}{P(y_j)} \right)$$

$$P_X(x_1) = a$$

$$P_X(x_2) = 1 - a$$

$$P_Y(y_1) = P_X(x_1) \cdot 0{,}8 + P_X(x_2) \cdot 0{,}2 = 0{,}8a + 0{,}2(1-a) = 0{,}2 + 0{,}6a$$

$$P_Y(y_2) = P_X(x_2) \cdot 0{,}8 + P_X(x_1) \cdot 0{,}2 = 0{,}8(1-a) + 0{,}2a = 0{,}8 - 0{,}6a$$

$$P(y_1 \mid x_1) = 0{,}8$$

$$P(y_1 \mid x_2) = 0{,}2$$

$$P(y_2 \mid x_1) = 0{,}2$$

$$P(y_2 \mid x_2) = 0{,}8$$

$$
\begin{aligned}
P_{XY} &= \begin{bmatrix} P(x_1 \cap y_1) & P(x_2 \cap y_1) \\ P(x_1 \cap y_2) & P(x_2 \cap y_2) \end{bmatrix} \\
&= \begin{bmatrix} P(y_1 \mid x_1) \cdot P(x_1) & P(y_1 \mid x_2) \cdot P(x_2) \\ P(y_2 \mid x_1) \cdot P(x_1) & P(y_2 \mid x_2) \cdot P(x_2) \end{bmatrix} \\
P_{XY} &= \begin{bmatrix} 0{,}8 \cdot a & 0{,}2 \cdot (1-a) \\ 0{,}2 \cdot a & 0{,}8 \cdot (1-a) \end{bmatrix}
\end{aligned}
$$

$$
\begin{aligned}
I(X;Y) = 0{,}8a \cdot \log_2\left(\frac{0{,}8}{0{,}2+0{,}6a}\right) + 0{,}2a \cdot \log_2\left(\frac{0{,}2}{0{,}8-0{,}6a}\right) \\
+ 0{,}2(1-a) \cdot \log_2\left(\frac{0{,}2}{0{,}2+0{,}6a}\right) + 0{,}8(1-a) \cdot \log_2\left(\frac{0{,}8}{0{,}8-0{,}6a}\right)
\end{aligned}
$$

Povprečno preneseno informacijo imamo zapisano kot funkcijo porazdelitve verjetnosti vhodne naključne spremenljivke. Njen maksimum poiščemo z odvajanjem. Ker je lažje odvajati naravni logaritem, funkcijo nekoliko predelamo, da na desni strani dobimo le naravne logaritme.

$$
\begin{aligned}
I(X;Y) = 0{,}8a \cdot \frac{\ln\left(\frac{0{,}8}{0{,}2+0{,}6a}\right)}{\ln 2} + 0{,}2a \cdot \frac{\ln\left(\frac{0{,}2}{0{,}8-0{,}6a}\right)}{\ln 2} \\
+0{,}2(1-a) \cdot \frac{\ln\left(\frac{0{,}2}{0{,}2+0{,}6a}\right)}{\ln 2} + 0{,}8(1-a) \cdot \frac{\ln\left(\frac{0{,}8}{0{,}8-0{,}6a}\right)}{\ln 2} \\
\ln 2 \cdot I(X;Y) = 0{,}8a \cdot \ln\left(\frac{0{,}8}{0{,}2+0{,}6a}\right) + 0{,}2a \cdot \ln\left(\frac{0{,}2}{0{,}8-0{,}6a}\right) \\
+0{,}2(1-a) \cdot \ln\left(\frac{0{,}2}{0{,}2+0{,}6a}\right) + 0{,}8(1-a) \cdot \ln\left(\frac{0{,}8}{0{,}8-0{,}6a}\right)
\end{aligned}
$$

$$
\begin{aligned}
\frac{d\left(f(x) \cdot g(x)\right)}{dx} &= \frac{df(x)}{dx} \cdot g(x) + f(x) \cdot \frac{dg(x)}{dx} \\
\frac{d}{dx}\left(\ln \frac{a}{b+cx}\right) &= \frac{d}{dx}\left(\ln a \cdot (b+cx)^{-1}\right) \\
&= \frac{1}{a \cdot (b+cx)^{-1}} \cdot (-1)a(b+cx)^{-2} \cdot c = -\frac{c}{b+cx}
\end{aligned}
$$

$$
\begin{aligned}
\ln 2 \cdot \frac{dI(X;Y)}{da} &= 0{,}8 \cdot \ln \frac{0{,}8}{0{,}2+0{,}6a} + 0{,}8a \cdot \frac{-0{,}6}{0{,}2+0{,}6a} \\
&\quad +0{,}2 \cdot \ln \frac{0{,}2}{0{,}8-0{,}6a} + 0{,}2a \cdot \frac{+0{,}6}{0{,}8-0{,}6a} \\
+0{,}2 \cdot \frac{-0{,}6}{0{,}2+0{,}6a} &- 0{,}2 \cdot \ln \frac{0{,}2}{0{,}2+0{,}6a} - 0{,}2a \cdot \frac{-0{,}6}{0{,}2+0{,}6a} \\
+0{,}8 \cdot \frac{+0{,}6}{0{,}8-0{,}6a} &- 0{,}8 \cdot \ln \frac{0{,}8}{0{,}8-0{,}6a} - 0{,}8a \cdot \frac{+0{,}6}{0{,}8-0{,}6a}
\end{aligned}
$$

$$\begin{aligned}
0 = 0{,}8 \cdot \ln\left(\frac{0{,}8}{0{,}2+0{,}6a} \cdot \frac{0{,}8-0{,}6a}{0{,}8}\right) + 0{,}2 \cdot \ln\left(\frac{0{,}2}{0{,}8-0{,}6a} \cdot \frac{0{,}2+0{,}6a}{0{,}2}\right)\\
+0{,}8 \cdot \frac{+0{,}6}{0{,}8-0{,}6a} - 0{,}8a \cdot \frac{+0{,}6}{0{,}8-0{,}6a} + 0{,}2a \cdot \frac{+0{,}6}{0{,}8-0{,}6a}\\
+0{,}2 \cdot \frac{-0{,}6}{0{,}2+0{,}6a} - 0{,}2a \cdot \frac{-0{,}6}{0{,}2+0{,}6a} + 0{,}8a \cdot \frac{-0{,}6}{0{,}2+0{,}6a}\\
0 = 0{,}8 \cdot \ln\left(\frac{0{,}8-0{,}6a}{0{,}2+0{,}6a}\right) + 0{,}2 \cdot \ln\left(\frac{0{,}2+0{,}6a}{0{,}8-0{,}6a}\right)\\
+\frac{0{,}6}{0{,}8-0{,}6a} \cdot (0{,}8-0{,}8a+0{,}2a)\\
+\frac{-0{,}6}{0{,}2+0{,}6a} \cdot (0{,}2-0{,}2a+0{,}8a)
\end{aligned}$$

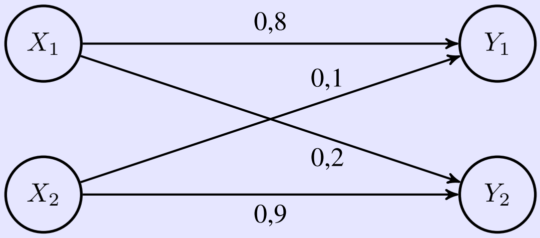
$$\begin{aligned}
0 = \ln\left(\left(\frac{0{,}8-0{,}6a}{0{,}2+0{,}6a}\right)^{0{,}8}\right) + \ln\left(\left(\frac{0{,}2+0{,}6a}{0{,}8-0{,}6a}\right)^{0{,}2}\right)\\
+\frac{0{,}6}{0{,}8-0{,}6a} \cdot (0{,}8-0{,}6a) + \frac{-0{,}6}{0{,}2+0{,}6a} \cdot (0{,}2+0{,}6a)\\
0 = \ln\left(\frac{(0{,}8-0{,}6a)^{0{,}8} \cdot (0{,}2+0{,}6a)^{0{,}2}}{(0{,}2+0{,}6a)^{0{,}8} \cdot (0{,}8-0{,}6a)^{0{,}2}}\right)\\
0 = \ln\left(\left(\frac{0{,}8-0{,}6a}{0{,}2+0{,}6a}\right)^{0{,}6}\right)\\
0 = 0{,}6\ln\left(\frac{0{,}8-0{,}6a}{0{,}2+0{,}6a}\right)
\end{aligned}$$

$$\begin{aligned}
1 &= \frac{0,8-0,6a}{0,2+0,6a}\\
0,2+0,6a &= 0,8-0,6a\\
1,2a &= 0,6\\
a &= 0,5
\end{aligned}$$

Vzajemna informacija je največja, kadar je verjetnost za oba vhodna simbola enaka.

$$\begin{aligned}
C &= I(X;Y)\Big|_{a=0,5}\\
C &= 0,4 \cdot \log_2 \frac{0,8}{0,5} + 0,1 \cdot \log_2 \frac{0,2}{0,5} + 0,1 \cdot \log_2 \frac{0,2}{0,5} + 0,4 \cdot \log_2 \frac{0,8}{0,5}\\
&= 0,8 \cdot \log_2 1,6 + 0,2 \cdot \log_2 0,4\\
C &= 0,28 \text{ bit/sim}
\end{aligned}$$

Naloga 8.2: Določite kapaciteto danega binarnega kanala.



Binarni kanal

Rešitev:

$$I(X;Y) = \sum_i \sum_j P_{XY}(x_i, y_j) \cdot \log_2 \left(\frac{P(y_j|x_i)}{P(y_j)} \right)$$

$$P_X(x_1) = a$$
$$P_X(x_2) = 1 - a$$
$$P_Y(y_1) = P_X(x_1) \cdot 0{,}8 + P_X(x_2) \cdot 0{,}1 = 0{,}8a + 0{,}1(1 - a) = 0{,}1 + 0{,}7a$$
$$P_Y(y_2) = P_X(x_2) \cdot 0{,}9 + P_X(x_1) \cdot 0{,}2 = 0{,}9(1 - a) + 0{,}2a = 0{,}9 - 0{,}7a$$
$$P(y_1|x_1) = 0{,}8$$
$$P(y_1|x_2) = 0{,}1$$
$$P(y_2|x_1) = 0{,}2$$
$$P(y_2|x_2) = 0{,}9$$

$$P_{XY} = \begin{bmatrix} P(x_1 \cap y_1) & P(x_2 \cap y_1) \\ P(x_1 \cap y_2) & P(x_2 \cap y_2) \end{bmatrix} = \begin{bmatrix} P(y_1|x_1) \cdot P(x_1) & P(y_1|x_2) \cdot P(x_2) \\ P(y_2|x_1) \cdot P(x_1) & P(y_2|x_2) \cdot P(x_2) \end{bmatrix}$$

$$P_{XY} = \begin{bmatrix} 0{,}8 \cdot a & 0{,}1 \cdot (1 - a) \\ 0{,}2 \cdot a & 0{,}9 \cdot (1 - a) \end{bmatrix}$$

$$I(X;Y) = 0{,}8a \cdot \log_2 \frac{0{,}8}{0{,}1 + 0{,}7a} + 0{,}2a \cdot \log_2 \frac{0{,}2}{0{,}9 - 0{,}7a} + 0{,}1(1 - a) \cdot \log_2 \frac{0{,}1}{0{,}1 + 0{,}7a} + 0{,}9(1 - a) \cdot \log_2 \frac{0{,}9}{0{,}9 - 0{,}7a}$$

$$\ln 2 \cdot I(X;Y) = 0{,}8a \cdot \ln \frac{0{,}8}{0{,}1 + 0{,}7a} + 0{,}2a \cdot \ln \frac{0{,}2}{0{,}9 - 0{,}7a} + 0{,}1(1 - a) \cdot \ln \frac{0{,}1}{0{,}1 + 0{,}7a} + 0{,}9(1 - a) \cdot \ln \frac{0{,}9}{0{,}9 - 0{,}7a}$$

$$\frac{d\,(f(x) \cdot g(x))}{dx} = \frac{df(x)}{dx} \cdot g(x) + f(x) \cdot \frac{dg(x)}{dx}$$

$$\frac{d}{dx}\left(\ln \frac{a}{b + cx} \right) = \frac{-c}{b + cx}$$

$$\begin{aligned}
\ln 2\cdot\frac{d}{da}I(X;Y) =&0{,}8\cdot\ln\frac{0{,}8}{0{,}1+0{,}7a}+0{,}8a\cdot\frac{-0{,}7}{0{,}1+0{,}7a}\\
&+0{,}2\cdot\ln\frac{0{,}2}{0{,}9-0{,}7a}+0{,}2a\cdot\frac{0{,}7}{0{,}9-0{,}7a}\\
&+0{,}1\cdot\frac{-0{,}7}{0{,}1+0{,}7a}-0{,}1\cdot\ln\frac{0{,}1}{0{,}1+0{,}7a}-0{,}1a\cdot\frac{-0{,}7}{0{,}1+0{,}7a}\\
&+0{,}9\cdot\frac{0{,}7}{0{,}9-0{,}7a}-0{,}9\cdot\ln\frac{0{,}9}{0{,}9-0{,}7a}-0{,}9a\cdot\frac{0{,}7}{0{,}9-0{,}7a}
\end{aligned}$$

$$\begin{aligned}
\ln 2\cdot\frac{d}{da}I(X;Y) =&0{,}8\cdot\ln\frac{0{,}8}{0{,}1+0{,}7a}+0{,}2\cdot\ln\frac{0{,}2}{0{,}9-0{,}7a}\\
&-0{,}1\cdot\ln\frac{0{,}1}{0{,}1+0{,}7a}-0{,}9\cdot\ln\frac{0{,}9}{0{,}9-0{,}7a}\\
&+0{,}2a\cdot\frac{0{,}7}{0{,}9-0{,}7a}-0{,}9a\cdot\frac{0{,}7}{0{,}9-0{,}7a}-0{,}8a\cdot\frac{0{,}7}{0{,}1+0{,}7a}\\
&+0{,}1a\cdot\frac{0{,}7}{0{,}1+0{,}7a}-0{,}1\cdot\frac{0{,}7}{0{,}1+0{,}7a}+0{,}9\cdot\frac{0{,}7}{0{,}9-0{,}7a}
\end{aligned}$$

$$\begin{aligned}
\ln 2\cdot\frac{d}{da}I(X;Y) =&\ln\frac{0{,}8^{0{,}8}}{(0{,}1+0{,}7a)^{0{,}8}}+\ln\frac{0{,}2^{0{,}2}}{(0{,}9-0{,}7a)^{0{,}2}}-\ln\frac{0{,}1^{0{,}1}}{(0{,}1+0{,}7a)^{0{,}1}}\\
&-\ln\frac{0{,}9^{0{,}9}}{(0{,}9-0{,}7a)^{0{,}9}}-0{,}7a\cdot\frac{0{,}7}{0{,}9-0{,}7a}+0{,}9\cdot\frac{0{,}7}{0{,}9-0{,}7a}\\
&-0{,}7a\cdot\frac{0{,}7}{0{,}1+0{,}7a}-0{,}1\cdot\frac{0{,}7}{0{,}1+0{,}7a}
\end{aligned}$$

$$\begin{aligned}
\ln 2\cdot\frac{d}{da}I(X;Y) =&\ln\Bigg(\frac{0{,}8^{0{,}8}}{(0{,}1+0{,}7a)^{0{,}8}}\cdot\frac{0{,}2^{0{,}2}}{(0{,}9-0{,}7a)^{0{,}2}}\\
&\cdot\frac{(0{,}1+0{,}7a)^{0{,}1}}{0{,}1^{0{,}1}}\cdot\frac{(0{,}9-0{,}7a)^{0{,}9}}{0{,}9^{0{,}9}}\Bigg)\\
&+(0{,}9-0{,}7a)\cdot\frac{0{,}7}{0{,}9-0{,}7a}-(0{,}1+0{,}7a)\cdot\frac{0{,}7}{0{,}1+0{,}7a}
\end{aligned}$$

$$\ln 2\cdot\frac{d}{da}I(X;Y)=\ln\left(\frac{(0{,}9-0{,}7a)^{0{,}7}\cdot 0{,}2^{0{,}2}\cdot 0{,}8^{0{,}8}}{(0{,}1+0{,}7a)^{0{,}7}\cdot 0{,}1^{0{,}1}\cdot 0{,}9^{0{,}9}}\right)+0{,}7-0{,}7$$

$$
\begin{aligned}
0 &= \ln\left(\frac{(0{,}9-0{,}7a)^{0{,}7}\cdot 0{,}2^{0{,}2}\cdot 0{,}8^{0{,}8}}{(0{,}1+0{,}7a)^{0{,}7}\cdot 0{,}1^{0{,}1}\cdot 0{,}9^{0{,}9}}\right)\\
&= \ln\left(\left(\frac{0{,}9-0{,}7a}{0{,}1+0{,}7a}\right)^{0{,}7}\right)+\ln\left(\frac{0{,}2^{0{,}2}\cdot 0{,}8^{0{,}8}}{0{,}1^{0{,}1}\cdot 0{,}9^{0{,}9}}\right)\\
&\ln\left(\frac{0{,}1^{0{,}1}\cdot 0{,}9^{0{,}9}}{0{,}2^{0{,}2}\cdot 0{,}8^{0{,}8}}\right) = \ln\left(\left(\frac{0{,}9-0{,}7a}{0{,}1+0{,}7a}\right)^{0{,}7}\right)\\
&\frac{0{,}1^{0{,}1}\cdot 0{,}9^{0{,}9}}{0{,}2^{0{,}2}\cdot 0{,}8^{0{,}8}} = \left(\frac{0{,}9-0{,}7a}{0{,}1+0{,}7a}\right)^{\frac{7}{10}}\\
&\left(\frac{0{,}1^{0{,}1}\cdot 0{,}9^{0{,}9}}{0{,}2^{0{,}2}\cdot 0{,}8^{0{,}8}}\right)^{\frac{10}{7}} = \frac{0{,}9-0{,}7a}{0{,}1+0{,}7a} \quad = K = 1{,}285\\
&0{,}1K+0{,}7Ka = 0{,}9-0{,}7a\\
a &= \frac{0{,}9-0{,}1K}{0{,}7+0{,}7K} = 0{,}482
\end{aligned}
$$

Kapaciteta kanala bo največja, kadar bo verjetnost za simbol x_1 0,482 in verjetnost za simbol x_2 0,518.

$$
\begin{aligned}
C &= I(X;Y)\big|_{a=0{,}482}\\
C &= 0{,}8\cdot 0{,}482\cdot\log_2\frac{0{,}8}{0{,}1+0{,}7\cdot 0{,}482}+0{,}2\cdot 0{,}482\cdot\log_2\frac{0{,}2}{0{,}9-0{,}7\cdot 0{,}482}\\
&+0{,}1(1-0{,}482)\cdot\log_2\frac{0{,}1}{0{,}1+0{,}7\cdot 0{,}482}+0{,}9(1-0{,}482)\cdot\log_2\frac{0{,}9}{0{,}9-0{,}7\cdot 0{,}482}\\
C &= 0{,}398
\end{aligned}
$$

Naloga 8.3: Izračunajte diferencialno entropijo zveznega informacijskega vira z enakomerno verjetnostno porazdelitvijo signala med -1 in 1.

Rešitev: Po analogiji z entropijo diskretnega informacijskega vira

$$H(X) = -\sum_i P_X(x_i)\log_2 P_X(x_i)$$

zapišemo diferencialno entropijo zveznega informacijskega vira z

$$h(X) = -\int_{-\infty}^{\infty} p_X(x)\log_2 p_X(x)\,dx.$$

Verjetnostna porazdelitev na danem intervalu je dana z

$$p_X(x) = \begin{cases} 0{,}5 & ; \quad -1 < x < 1\\ 0 & ; \quad \textit{sicer}\end{cases}$$

Verjetnostno porazdelitev vstavimo v izraz za izračun diferencialne entropije in dobimo

$$h(X) = -\int_{-1}^{1} 0{,}5 \log_2 0{,}5 \, dx = 0{,}5 \int_{-1}^{1} dx = -0{,}5 \cdot x \Big|_{-1}^{1} = 0{,}5 \cdot 2 = 1.$$

Naloga 8.4: Izračunajte diferencialno entropijo zveznega informacijskega vira z dano verjetnostno porazdelitvijo signala

$$p_x(x) = (1 - |x|)u(1 + x)u(1 - x)$$

Rešitev: Verjetnostno porazdelitev zapišemo po posameznih intervalih

$$p_X(x) = \begin{cases} 1 + x & ; \quad -1 < x < 0 \\ 1 - x & ; \quad 0 \leq x < 1 \\ 0 \quad ; \quad \textit{sicer} \end{cases}$$

in vstavimo v izraz za izračun diferencialne entropije

$$\begin{aligned} h(X) &= -\int_{-\infty}^{\infty} p_X(x) \log_2 p_X(x) \, dx \\ &= -\int_{-1}^{0} (1 + x) \log_2 (1 + x) \, dx - \int_{0}^{1} (1 - x) \log_2 (1 - x) \, dx \\ h(X) &= \frac{1}{\ln 2} \left(-\int_{-1}^{0} (1 + x) \ln(1 + x) \, dx - \int_{0}^{1} (1 - x) \ln(1 - x) \, dx \right) \end{aligned}$$

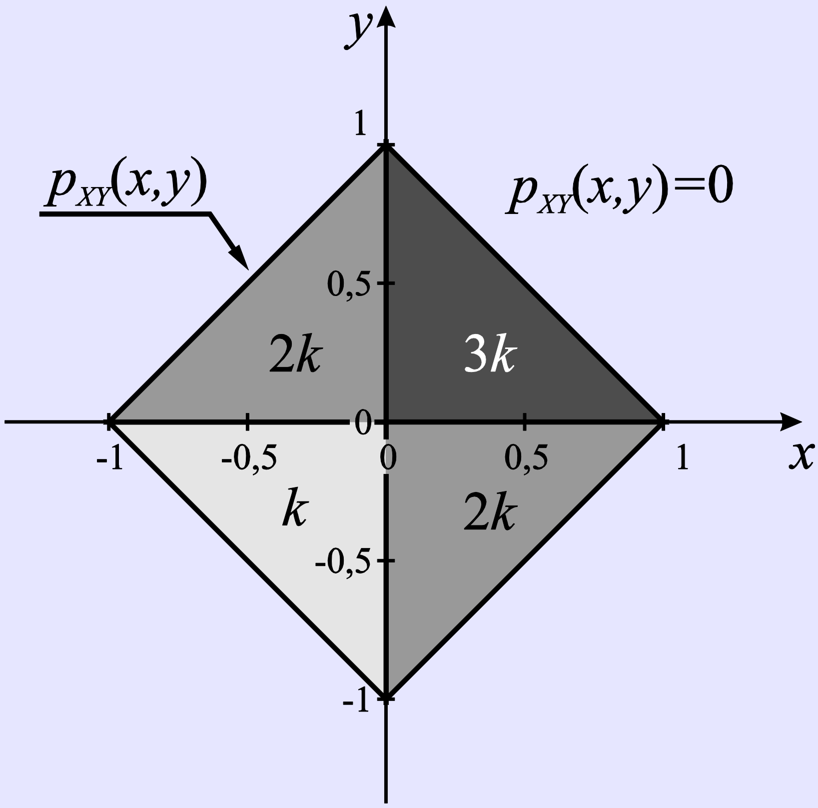
$$\begin{aligned} \int_{-1}^{0} & (1 + x) \ln(1 + x) \, dx = \\ & 1 + x = u \qquad x = -1 : u = 0 \\ & dx = du \qquad x = 0 : u = 1 \\ & = \int_{0}^{1} u \ln u \, du = u^2 \left(\frac{\ln u}{2} - \frac{1}{4} \right) \Big|_{0}^{1} \\ & = \left(\frac{\ln 1}{2} - \frac{1}{4} \right) - 0^2 \left(\frac{\ln 0}{2} - \frac{1}{4} \right) = -\frac{1}{4} \end{aligned}$$

$$\int_0^1 (1-x)\ln(1-x)\,dx =$$

$$\begin{aligned}
&1-x=u \qquad x=0: u=1\\
&-dx=du \qquad x=1: u=0\\
&= -\int_1^0 u\ln u\,du = u^2\left(\frac{\ln u}{2}-\frac{1}{4}\right)\Bigg|_1^0\\
&= 0^2\left(\frac{\ln 0}{2}-\frac{1}{4}\right)-\left(\frac{\ln 1}{2}-\frac{1}{4}\right)=\frac{1}{4}
\end{aligned}$$

$$h(X)=\frac{1}{\ln 2}\left(-\frac{1}{4}+\frac{1}{4}\right)=\frac{1}{2\ln 2}$$

Naloga 8.5: Podana je vezana gostota verjetnosti vhodne spremenljivke X in izhodne spremenljivke Y. Izračunajte entropiji vhodne in izhodne spremenljivke ter povprečno preneseno informacijo.



Rešitev: Velikost konstante k določimo iz volumna pod površino, ki predstavlja gostoto verjetnosti:

$$\iint p_{XY}(x,y)\,dx\,dy = 4k = 1 \quad \Rightarrow \quad k = 0{,}25$$

Za izračun diferencialnih entropij potrebujemo marginalni porazdelitvi verjetnosti:

$$p_X(x) = \int_{-\infty}^{\infty} p_{XY}(x,y)\,dy$$

$$-1 < x < 0:$$

$$\begin{aligned} p_X(x) &= \int_{-1-x}^{0} k\,dy + \int_{0}^{1+x} 2k\,dy = k\,y\Big|_{-1-x}^{0} + 2k\,y\Big|_{0}^{1+x} \\ &= k(1+x) + 2k(1+x) = 3k(1+x) \end{aligned}$$

$$0 \le x < 1:$$

$$\begin{aligned} p_X(x) &= \int_{-1+x}^{0} 2k\,dy + \int_{0}^{1-x} 3k\,dy = 2k\,y\Big|_{-1+x}^{0} + 3k\,y\Big|_{0}^{1-x} \\ &= 2k(1-x) + 3k(1-x) = 5k(1-x) \end{aligned}$$

$$p_X(x) = \begin{cases} 3k(1+x) & ; \quad -1 < x < 0 \\ 5k(1-x) & ; \quad 0 \le x < 1 \\ 0 & ; \quad \textit{sicer} \end{cases}$$

$$p_Y(y) = \int_{-\infty}^{\infty} p_{XY}(x,y)\,dx$$

$$-1 < y < 0:$$

$$\begin{aligned} p_Y(y) &= \int_{-1-y}^{0} k\,dx + \int_{0}^{1+y} 2k\,dx = ky\Big|_{-1-y}^{0} + 2ky\Big|_{0}^{1+y} \\ &= k(1+y) + 2k(1+y) = 3k(1+y) \end{aligned}$$

$$0 \le y < 1:$$

$$\begin{aligned} p_Y(y) &= \int_{-1+y}^{0} 2k\,dy + \int_{0}^{1-y} 3k\,dy = 2ky\Big|_{-1+y}^{0} + 3ky\Big|_{0}^{1-y} \\ &= 2k(1-y) + 3k(1-y) = 5k(1-y) \end{aligned}$$

$$p_Y(y) = \begin{cases} 3k(1+y) & ; -1 < y < 0 \\ 5k(1-y) & ; 0 \le y < 1 \\ 0 & ; \textit{sicer} \end{cases}$$

Diferencialni entropiji:

$$h(X) = -\int_{-\infty}^{\infty} p_X(x)\log_2 p_X(x)\,dx = -\frac{1}{\ln 2}\int_{-\infty}^{\infty} p_X(x)\ln p_X(x)\,dx$$

$$h(X) = -\frac{1}{\ln 2}\int_{-1}^{0} 3k(1+x)\ln 3k(1+x)\,dx - \frac{1}{\ln 2}\int_{0}^{1} 5k(1-x)\ln 5k(1-x)\,dx$$

$$3k(1+x) = u \qquad x=-1 : u = 0 \qquad x = 0 : u = 3k \qquad 3k\,dx = du$$
$$5k(1-x) = v \qquad x = 0 : v = 5k \qquad x = 1 : v = 0 \qquad -5k\,dx = dv$$

$$h(X) = -\frac{1}{\ln 2}\int_{0}^{3k} u\ln u\,\frac{du}{3k} + \frac{1}{\ln 2}\int_{5k}^{0} v\ln v\,\frac{dv}{5k}$$
$$= \frac{1}{\ln 2}\left(-u^2\left(\frac{\ln u}{2} - \frac{1}{4}\right)\Big|_0^{3k} + v^2\left(\frac{\ln v}{2} - \frac{1}{4}\right)\Big|_{5k}^{0}\right)$$
$$h(X) = \frac{1}{\ln 2}\left(-9k^2\left(\frac{\ln 3k}{2} - \frac{1}{4}\right)\right.$$
$$\left. + 0^2\left(\frac{\ln 0}{2} - \frac{1}{4}\right) + 0^2\left(\frac{\ln 0}{2} - \frac{1}{4}\right) - 25k^2\left(\frac{\ln 5k}{2} - \frac{1}{4}\right)\right)$$
$$h(X) = \frac{1}{\ln 2}\left(-9k^2\left(\frac{\ln 3k}{2} - \frac{1}{4}\right) - 25k^2\left(\frac{\ln 5k}{2} - \frac{1}{4}\right)\right) = 0{,}63$$
$$p_Y(y) = p_X(x) \quad \Rightarrow \quad h(Y) = h(X) = 0{,}63$$

Vezana diferencialna entropija:

$$h(X,Y) = -\iint p_{XY}(x,y)\log_2 p_{XY}(x,y)\,dxdy$$
$$h(X,Y) = -\int_{-1}^{0}\int_{-1-x}^{0} k\log_2 k\,dydx - \int_{-1}^{0}\int_{0}^{1+x} 2k\log_2 2k\,dydx$$
$$-\int_{0}^{1}\int_{-1+x}^{0} 2k\log_2 2k\,dydx - \int_{0}^{1}\int_{0}^{1-x} 3k\log_2 3k\,dydx$$

$$h(X,Y) = -\int_{-1}^{0}\int_{-1-x}^{0} 0{,}25\log_2 0{,}25\,dydx - \int_{-1}^{0}\int_{0}^{1+x} 0{,}5\log_2 0{,}5\,dydx$$
$$-\int_{0}^{1}\int_{-1+x}^{0} 0{,}5\log_2 0{,}5\,dydx - \int_{0}^{1}\int_{0}^{1-x} 0{,}75\log_2 0{,}75\,dydx$$

$$\int_{-1}^{0}\int_{-1-x}^{0} dydx = \int_{-1}^{0} y|_{-1-x}^{0}\, dx = \int_{-1}^{0}(1+x)dx = x|_{-1}^{0} + \left.\frac{x^2}{2}\right|_{-1}^{0} = 1 - \frac{1}{2} = 0{,}5$$

$$\int_{-1}^{0}\int_{0}^{1+x} dydx = \int_{-1}^{0} y|_{0}^{1+x}\, dx = \int_{-1}^{0}(1+x)dx = x|_{-1}^{0} + \left.\frac{x^2}{2}\right|_{-1}^{0} = 1 - \frac{1}{2} = 0{,}5$$

$$\int_{0}^{1}\int_{-1+x}^{0} dydx = \int_{0}^{1} y|_{-1+x}^{0}\, dx = \int_{0}^{1}(1-x)dx = x|_{0}^{1} - \left.\frac{x^2}{2}\right|_{0}^{1} = 1 - \frac{1}{2} = 0{,}5$$

$$\int_{0}^{1}\int_{0}^{1-x} dydx = \int_{0}^{1} y|_{0}^{1-x}\, dx = \int_{0}^{1}(1-x)dx = x|_{0}^{1} - \left.\frac{x^2}{2}\right|_{0}^{1} = 1 - \frac{1}{2} = 0{,}5$$

$$h(X,Y) = 0{,}5 \cdot 0{,}5 + 0{,}5 \cdot 0{,}5 + 0{,}5 \cdot 0{,}5 + 0{,}31 \cdot 0{,}5 = 0{,}905$$

Iz diferencialnih entropij marginalnih verjetnostnih porazdelitev in vezane diferencialne entropije izračunamo povprečno preneseno informacijo:

$$I(X;Y) = h(X) + h(Y) - h(X,Y) = 0{,}63 + 0{,}63 - 0{,}91 = 0{,}35$$

$I(x;y) = \sum\sum P_{xy}(x,y) \cdot \log_2 \frac{P(y\|x)}{P(y)}$	Vzajemna informacija med dvema spremenljivkama
$I(X;Y) = H(X) + H(Y) - H(X,Y)$	Vzajemna informacija dveh spremenljivk
$h(X) = -\int_{-\infty}^{\infty} p_X(x) \log_2 p_X(x)\, dx.$	Diferencialna entropija
$h(x,y) = -\int\int p_{xy}(x,y) \log_2 p_{xy}(x,y)\, dx\, dy$	Vezana diferencialna entropija

9. POGLAVJE

KAPACITETA KANALA ZA RAZLIČNE PORAZDELITVE

V tem poglavju bomo raziskali eno najpomembnejših lastnosti komunikacijskih kanalov: njihovo kapaciteto. Osredotočili se bomo predvsem na Gaussov kanal, ki modelira vpliv šuma z normalno porazdelitvijo, kar je v realnih sistemih pogosto zelo dober približek.

Kapaciteta kanala nam pove, kakšna je največja možna hitrost prenosa informacij, pri kateri še vedno lahko dosežemo zanemarljivo napako s primerno kodiranjem. Spoznali bomo tudi diferencialno entropijo kot orodje za kvantifikacijo informacije v neomejenih, zveznih porazdelitvah.

Poglobljeno bomo obravnavali Gaussovo in Laplaceovo porazdelitev, izrazili entropijo signalov in šuma, povezali moč signala z napakami pri prenosu in nazorno izračunali kapacitete za različne realne komunikacijske scenarije (npr. audio kanal, televizijski signal, HD-video).

Za dodatno branje priporočamo naslednje vire:

- C. E. Shannon: *A Mathematical Theory of Communication* [4]
- T. M. Cover, J. A. Thomas: *Elements of Information Theory* [2]
- S. Haykin: *Communication Systems, 5th ed.* [7]

9.1 Naloge

Naloga 9.1: Izračunajte diferencialno entropijo Gaussove naključne spremenljivke z enosmerno vrednostjo 3 in standardno deviacijo 2.

Rešitev:

$$p_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$h(X) = -\int_{-\infty}^{\infty} p_X(x) \log_2 p_X(x)\, dx = -\frac{1}{\ln 2}\int_{-\infty}^{\infty} p_X(x) \ln p_X(x)\, dx$$

$$h(X) = -\frac{1}{\ln 2}\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \ln\left(\frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\right) dx$$

$$h(X) = -\frac{1}{\sigma\sqrt{2\pi}\ln 2}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\left(\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) + \ln\left(e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\right)\right) dx$$

$$h(X) = -\frac{1}{\sigma\sqrt{2\pi}\ln 2}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\left(\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) dx$$

$$h(X) = -\frac{1}{\sigma\sqrt{2\pi}\ln 2}\left(\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right)\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \right.$$
$$\left. - \frac{1}{2}\int_{-\infty}^{\infty}\left(\frac{x-\mu}{\sigma}\right)^2 e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx\right)$$

$$h(X) = -\frac{1}{\ln 2}\left(\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right)\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \right.$$
$$\left. - \frac{1}{2}\int_{-\infty}^{\infty}\left(\frac{x-\mu}{\sigma}\right)^2 p_X(x)\, dx\right)$$

$$h(X) = -\frac{1}{\ln 2}\left(\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right)\int_{-\infty}^{\infty} p_X(x)\, dx - \frac{1}{2\sigma^2}\int_{-\infty}^{\infty}(x-\mu)^2 p_X(x)\, dx\right)$$

$$h(X) = -\frac{1}{\ln 2}\left(\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right)\cdot 1 - \frac{1}{2\sigma^2}\cdot\sigma^2\right)$$

$$h(X) = -\left(\log_2\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2}\frac{\ln e}{\ln 2}\right) = \log_2\left(\sigma\sqrt{2\pi}\right) + \frac{1}{2}\log_2 e = \log_2\left(\sigma\sqrt{2\pi e}\right)$$

$$h(X) = \frac{1}{2}\log_2\left(2\pi e\sigma^2\right) = 3{,}05$$

Naloga 9.2: Izračunajte diferencialno entropijo naključne spremenljivke z Laplaceovo verjetnostno porazdelitvijo, enosmerno vrednostjo 0,1 in standardno deviacijo 0,5.

Rešitev:

$$p_X(x) = \frac{1}{\sigma\sqrt{2}} \cdot e^{-\frac{\sqrt{2}|x-\mu|}{\sigma}}$$

$$h(X) = -\int_{-\infty}^{\infty} p_X(x) \log_2 p_X(x)\, dx = -\frac{1}{\ln 2}\int_{-\infty}^{\infty} p_X(x) \ln p_X(x)\, dx$$

$$h(X) = -\frac{1}{\ln 2}\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2}} \cdot e^{-\frac{\sqrt{2}|x-\mu|}{\sigma}} \cdot \ln\left(\frac{1}{\sigma\sqrt{2}} \cdot e^{-\frac{\sqrt{2}|x-\mu|}{\sigma}}\right) dx$$

$$h(X) = -\frac{1}{\ln 2} \cdot \frac{1}{\sigma\sqrt{2}}\int_{-\infty}^{\infty} e^{-\frac{\sqrt{2}|x-\mu|}{\sigma}} \left(\ln\frac{1}{\sigma\sqrt{2}} + \ln e^{-\frac{\sqrt{2}|x-\mu|}{\sigma}}\right) dx$$

$$h(X) = -\frac{1}{\ln 2} \cdot \frac{1}{\sigma\sqrt{2}}\int_{-\infty}^{\infty} e^{-\frac{\sqrt{2}|x-\mu|}{\sigma}} \left(\ln\frac{1}{\sigma\sqrt{2}} - \frac{\sqrt{2}\,|x-\mu|}{\sigma}\right) dx$$

$$h(X) = -\frac{1}{\ln 2} \cdot \frac{1}{\sigma\sqrt{2}} \ln\frac{1}{\sigma\sqrt{2}}\int_{-\infty}^{\infty} e^{-\frac{\sqrt{2}|x-\mu|}{\sigma}} dx$$
$$+ \frac{1}{\ln 2} \cdot \frac{1}{\sigma\sqrt{2}}\int_{-\infty}^{\infty} \frac{\sqrt{2}\,|x-\mu|}{\sigma} e^{-\frac{\sqrt{2}|x-\mu|}{\sigma}} dx$$

$$t = \frac{x-\mu}{\sigma} \qquad x = -\infty : t = -\infty \qquad x = \infty : t = \infty$$

$$dt = \frac{dx}{\sigma}$$

$$h(X) = -\frac{1}{\ln 2} \ln\frac{1}{\sigma\sqrt{2}}\int_{-\infty}^{\infty} p_X(x) dx + \frac{1}{\ln 2} \cdot \frac{1}{\sigma}\int_{-\infty}^{\infty} |t|\, e^{-\sqrt{2}|t|} \sigma\, dx$$

$$h(X) = -\frac{1}{\ln 2} \ln\frac{1}{\sigma\sqrt{2}} \cdot 1 + \frac{1}{\ln 2}\int_{-\infty}^{\infty} |t|\, e^{-\sqrt{2}|t|} dx$$

$$h(X) = -\frac{1}{\ln 2} \ln\frac{1}{\sigma\sqrt{2}} \cdot 1 + \frac{1}{\ln 2} \cdot 2\int_{0}^{\infty} t e^{-\sqrt{2}t} dx$$
$$= \log_2 \sigma\sqrt{2} + \frac{1}{\ln 2}\left(2 \cdot e^{-\sqrt{2}t}\left(-\frac{t}{\sqrt{2}} - \frac{1}{2}\right)\right)\Bigg|_0^{\infty}$$

$$h(X) = \log_2 \sigma\sqrt{2} + \frac{1}{\ln 2}\left(\lim_{t\to\infty}\left(2 \cdot e^{-\sqrt{2}t}\left(-\frac{t}{\sqrt{2}} - \frac{1}{2}\right)\right)\right.$$
$$\left. -2 \cdot e^{-\sqrt{2}\cdot 0}\left(-\frac{0}{\sqrt{2}} - \frac{1}{2}\right)\right)$$

Člen $e^{\sqrt{2}t}$ raste veliko hitreje kot t, zato celota upada proti 0.

$$h(X) = \log_2 \sigma\sqrt{2} + \frac{1}{\ln 2}\left(0 - 2\left(-\frac{1}{2}\right)\right) = \log_2 \sigma\sqrt{2} + \frac{1}{\ln 2} = \log_2 \sigma\sqrt{2} + \frac{\ln e}{\ln 2}$$

$$h(X) = \log_2 \sigma\sqrt{2} + \log_2 e = \log_2 \sqrt{2}e\sigma$$

$$h(X) = \frac{1}{2}\log_2 2ee\sigma^2 = 0{,}942$$

Naloga 9.3: Naključna spremenljivka Y je definirana kot vsota naključnih spremenljivk X in Z z Gaussovo porazdelitvijo gostote verjetnosti. Izrazite pogojno diferencialno entropijo $h(Y|X)$ z diferencialnima entropijama $h(X)$ in $h(Z)$.

Rešitev: Če seštejemo dve naključni spremenljivki, ima dobljena naključna spremenljivka povprečno vrednost, ki je vsota povprečnih vrednosti originalnih spremenljivk in moč nove naključne spremenljivke je enaka vsoti moči originalnih naključnih spremenljivk.

$$Y = X + Z, \qquad \mu_Y = \mu_X + \mu_Z, \qquad \sigma_Y^2 = \sigma_X^2 + \sigma_Z^2$$

Izraz za pogojno diferencialno entropijo $h(Y|X)$ moramo predelati tako, da ga bomo lahko izrazili s pomočjo diferencialnih entropij $h(X)$ in/ali $h(Z)$, zato si ju tu izpišimo:

$$h(X) = -\int_{-\infty}^{\infty} p_X(x) \log_2 p_X(x)\, dx, \qquad h(Z) = -\int_{-\infty}^{\infty} p_Z(z) \log_2 p_Z(z)\, dz$$

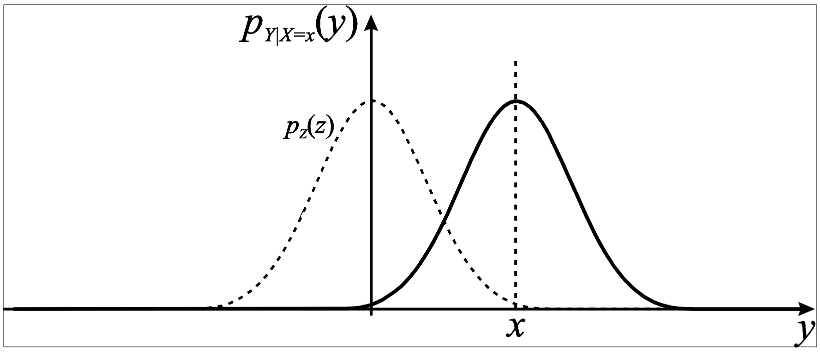
Pogojna diferencialna entropija $h(Y|X)$ je povprečje vseh pogojnih diferencialnih entropij $h(Y|X=x)$, kjer naključno spremenljivko X fiksiramo na eno vrednost x:

$$h(Y \mid X) = \int_{-\infty}^{\infty} p_X(x) \cdot h(Y \mid X = x)\, dx$$

$$h(Y \mid X = x) = -\int_{-\infty}^{\infty} p_{Y|X=x}(y, x) \log_2 p_{Y|X=x}(y, x)\, dx$$

Če je naključna spremenljivka X fiksirana na vrednost x, je naključna spremenljivka Y definirana kot:

$$Y\big|_{X=x} = x + Z, \qquad \mu_Y = x + \mu_Z, \qquad \sigma_Y^2 = \sigma_Z^2.$$



Skratka naključna spremenljivka Y je enaka naključni spremenljivki Z, le da je njena srednja vrednost premaknjena za x. Ker je entropija neodvisna od srednje vrednosti naključne spremenljivke, velja

$$h(Y \mid X = x) = h(Z).$$

Sedaj lahko zapišemo:

$$\begin{aligned} h(Y \mid X) &= \int_{-\infty}^{\infty} p_X(x) \cdot h(Y \mid X = x)\, dx \\ &= \int_{-\infty}^{\infty} p_X(x) \cdot h(Z)\, dx = h(Z) \int_{-\infty}^{\infty} p_X(x)\, dx \end{aligned}$$

$$h(Y \mid X) = h(Z)$$

Naloga 9.4: Izrazite povprečno moč diskretne naključne spremenljivke z L nivoji, povprečno vrednostjo 0 in enakomerno verjetnostno porazdelitvijo. Pri tem predpostavite, da so posamezni diskretni nivoji razmaknjeni za $k\sigma$, kjer σ predstavlja standardno deviacijo prisotnega šuma, in k faktor s katerim bomo na koncu nastavili varno razdaljo med posameznimi nivoji.

Rešitev: Če za L izberete sodo število boste potrebovali izraz

$$\sum_{n=1}^{N} (2n-1)^2 = \frac{N}{3}\left(4N^2 - 1\right)$$

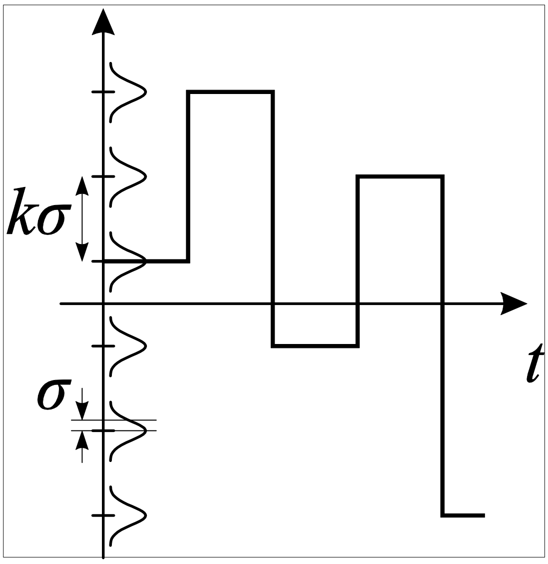
Če za L izberete liho število boste potrebovali izraz

$$\sum_{n=-\frac{N-1}{2}}^{\frac{N-1}{2}} n^2 = \frac{N}{12}\left(N^2 - 1\right)$$

Povprečna moč naključne spremenljivke:

$$S = \overline{X^2} = \sum_i x_i^2 P_X(x_i)$$

Izraziti moramo posamezne nivoje x_i:



$$x_i = ik\sigma - \frac{k\sigma}{2}, \quad i = 1, 2, 3, \ldots, L/2$$

Verjetnost vseh nivojev je 1/L. Upoštevajmo, da je moč pozitivnih in negativnih nivojev enaka:

$$S = 2\sum_{i=1}^{L/2} x_i^2 \cdot \frac{1}{L} = \frac{2}{L}\sum_{i=1}^{L/2}\left(\frac{k\sigma}{2}\right)^2 (2i-1)^2 = \frac{2}{L}\cdot\frac{k^2\sigma^2}{4}\sum_{i=1}^{L/2}(2i-1)^2$$

$$S = \frac{2}{L}\cdot\frac{k^2\sigma^2}{4}\cdot\frac{L/2}{3}\left(4\left(\frac{L}{2}\right)^2 - 1\right) = \frac{k^2\sigma^2}{12}(L^2-1)$$

Entropija naključne spremenljivke z enakomerno verjetnostno porazdelitvijo je $H(X) = \log_2(L)$ Zato iz moči izrazimo L:

$$\frac{12S}{k^2\sigma^2} = L^2 - 1$$

$$L = \sqrt{\frac{12S}{k^2\sigma^2} + 1}$$

In ga vstavimo v izraz za entropijo:

$$H(X) = \log_2\left(\sqrt{\frac{12S}{k^2\sigma^2} + 1}\right) = \frac{1}{2}\log_2\left(\frac{12S}{k^2\sigma^2} + 1\right)$$

Če upoštevamo, da je σ standardna deviacija šuma, je σ^2 moč šuma N:

$$H(X) = \frac{1}{2}\log_2\left(1 + \frac{12}{k^2}\cdot\frac{S}{N}\right)$$

Ta izraz je zelo podoben izrazu za kapaciteto Gaussovega kanala $C = \frac{1}{2}log_2(1+\frac{S}{N})$. Razlikuje se le za člen $\frac{12}{k^2}$. Oba izraza postaneta enaka, če faktor k izberemo tako, da ta člen postane 1:

$k = \sqrt{12} \approx 3,464$

Če za odmik med posameznimi nivoji vhodnega signala izberemo $3,46\sigma$, bo verjetnost, da pride do napake pri prenosu $8,4\%$.

Naloga 9.5: Ali je teoretično možno brez napak prenašati signal s standardno deviacijo $0,1V$ s hitrostjo $10Mbit/s$ preko Gaussovega kanala s pasovno širino $100kHz$ in karakteristično impedanco 50Ω pri temperaturi 20 °C? Izračunajte energijo posameznega bita informacije? Na kolikšno vrednost bi morali nastaviti standardno deviacijo signala, da bi bila kapaciteta kanala enaka želeni hitrost prenosa? Kolikšna bi morala biti pasovna širina kanala, da bi bila kapaciteta kanala enaka želeni hitrost prenosa pri nespremenjeni velikosti? $k = 1,3810J/K$

Rešitev: Moč signala:

$$S = \frac{U_{ef}^2}{R} = \frac{\sigma^2}{R} = \frac{0{,}01\,\mathrm{V}^2}{50\,\Omega} = 0{,}0002\,\mathrm{W}$$

Šumna moč:

$$N = BkT = 10^5\,\mathrm{Hz} \cdot 1{,}38 \cdot 10^{-23}\,\mathrm{J/K} \cdot 293\,\mathrm{K} = 4{,}04 \cdot 10^{-16}\,\mathrm{W}$$

Kapaciteta kanala:

$$C = B \cdot \log_2\left(1 + \frac{S}{N}\right) = 3{,}88 \cdot 10^6\,\mathrm{sh/s}$$

Energija bita:

$$E_b = \frac{S}{C} = \frac{0{,}0002\,\mathrm{W}}{3{,}88 \cdot 10^6\,\mathrm{bit/s}} = 5{,}16 \cdot 10^{-11}\,\mathrm{J}$$

Zahtevana moč signala za mejno kapaciteto kanala:

$$\frac{C}{B} = \log_2\left(1 + \frac{S}{N}\right)$$

$$2^{\frac{C}{B}} = 1 + \frac{S}{N}$$

$$S = \left(2^{\frac{C}{B}} - 1\right) N = \left(2^{\frac{10^7\,\mathrm{bit/s}}{10^5\,\mathrm{Hz}}} - 1\right) \cdot 4{,}04 \cdot 10^{-16}\,\mathrm{W} = 5{,}12 \cdot 10^{14}\,\mathrm{W}$$

$$\sigma = U_{ef} = \sqrt{RS} = \sqrt{5{,}12 \cdot 10^{14}\,\mathrm{W} \cdot 50\,\Omega} = 160\,\mathrm{MV}$$

Naloga 9.6: Telefonski avdio kanal ima pasovno širino 3,4 kHz. Izračunajte kapaciteto kanala za SNR 30 dB. Izračunajte minimalno SNR za prenos 57600 bit/s.

Rešitev: Kapaciteta kanala:

$$\frac{S}{N} = 30\,dB = 10^{\frac{30}{10}} = 1000$$

$$C = B \log_2 \left(1 + \frac{S}{N}\right) = 3400 \cdot \log_2 1001 = 33889\,\mathrm{bit/s}$$

Minimalno razmerje SNR:

$$2^{\frac{C}{B}} = 1 + \frac{S}{N}$$

$$\frac{S}{N} = 2^{\frac{C}{B}} - 1 = 2^{\frac{57600}{3400}} - 1 = 125834 = 51\,dB$$

Naloga 9.7: Telefonski avdio kanal s pasovno širino 3400 Hz in razmerjem med signalom in šumom 20 dB, želimo uporabiti za tekstovni terminal, ki pozna 128 znakov. Predpostavite, da je verjetnostna porazdelitev znakov enakomerna in da je informacijski vir brez-spominski. Izračunajte kapaciteto kanala in maksimalno možno hitrost prenosa znakov, da še ne prihaja do napak.

Rešitev: Kapaciteta kanala:

$$C = B \log_2 \left(1 + \frac{S}{N}\right) = 3400 \log_2 (1 + 100) = 22638$$

Entropija vira:

$$H(X) = \log_2 L = \log_2 128 = 7$$

Maksimalna možna hitrost prenosa znakov:

$$f_D = \frac{C}{H(X)} = \frac{22638 \text{ sh/s}}{7 \text{ sh/sim}} = 3234 \text{ baud}$$

Naloga 9.8: Črno-bela televizijska slika je sestavljena iz $4\text{x}10^5$ točk, od katerih lahko vsaka zasede enega izmed 220-ih odtenkov sivine z enako verjetnostjo. Frekvenca osveževanja slike je 25 Hz. Kolikšna mora biti pasovna širina informacijskega kanala, če predpostavimo razmerje signal/šum 50 dB. (Dejanska pasovna širina kanala za televizijski video signal v standardni ločljivosti znaša 5 MHz in je umeščena v frekvenčni pas širine 8 MHz). Kolikšna bi morala biti kapaciteta kanala za prenos ne-stisnjene barvne slike v HD ločljivosti (1920×1080 točk, povprečno 16 bitov na točko (barvna prostorska ločljivost 4:2:2), 25 slik na sekundo)?

Rešitev: Entropija ene slike (ob predpostavki, da slikovne točke niso med sabo odvisne):

$$H(X) = \log_2 L^N = N \log_2 L = 4 \cdot 10^5 \log_2 220 = 3112544 \text{ sh}$$

Potrebna kapaciteta kanala:

$$C = H(X) \cdot f_{\text{SLIK}} = 3112544 \text{ sh} \cdot 25 \text{ s}^{-1} = 77813597 \text{ sh/s}$$

Pasovna širina kanala, potrebna za prenos videa:

$$B = \frac{C}{\log_2\left(1 + \frac{S}{N}\right)} = \frac{77813597 \text{ sh/s}}{\log_2(1 + 100000)} = 4684841 \text{ Hz} \approx 4{,}7 \text{ MHz}$$

Entropija ene slike v HD ločljivosti:

$$H_{HD}(X) = N_{HD} \log_2 L_{HD} = 1920 \cdot 1080 \cdot 16 = 33177600 \text{ sh}$$

Potrebna kapaciteta kanala za video HD ločljivosti:

$$C_{HD} = H_{HD}(X) \cdot f_{\text{HDSLIK}} = 33177600 \text{ sh} \cdot 30 \text{ s}^{-1} = 99532800 \text{ sh/s}$$

Pasovna širina kanala, potrebna za prenos HD videa:

$$B_{HD} = \frac{C_{HD}}{\log_2\left(1 + \frac{S}{N}\right)} = \frac{99532800 \text{ sh/s}}{\log_2(1 + 100000)} = 59924665 \text{ Hz} \approx 60 \text{ MHz}$$

$p_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$	Gaussova porazdelitev
$p_X(x) = \frac{1}{\sigma\sqrt{2}} \cdot e^{-\sqrt{2}\frac{\lvert x-\mu\rvert}{\sigma}}$	Laplaceova porazdelitev
$h(X) = \frac{1}{2}\log_2\left(2\pi e\sigma^2\right)$	Diferencialna entropija Gaussove porazdelitve
$h(X) = \frac{1}{2}\log_2\left(2ee\sigma^2\right)$	Diferencialna entropija Laplaceove porazdelitve
$SNR_{\text{dB}} = 10\log_{10}\left(\frac{S}{N}\right)$	Razmerje signal/šum v decibelih
$C = B \cdot \log_2\left(1 + \frac{S}{N}\right)$	Kapaciteta kanala

10. POGLAVJE

VZORČENJE

V tem poglavju bomo obravnavali osnovne pojme vzorčenja analognih signalov, kot jih naletimo pri digitalizaciji v sodobnih komunikacijskih sistemih. Pri pretvorbi analognega signala v diskretno časovno obliko moramo zagotoviti, da signal vsebuje dovolj informacij, da ga lahko kasneje rekonstruiramo.

Ključni koncept, ki ga bomo podrobneje spoznali, je **Nyquistov vzorčevalni teorem**, ki določa najmanjšo vzorčevalno frekvenco, potrebno za brezizgubno rekonstrukcijo signala. Če vzorčimo z nižjo frekvenco, se lahko zgodi **aliasing** — pojav, pri katerem višje frekvence navidezno postanejo nižje zaradi zrcaljenja spektra.

Na konkretnih primerih bomo prikazali:

- kako frekvence v signalih vplivajo na spekter po vzorčenju,
- kako ugotovimo, ali bo pri dani vzorčevalni frekvenci prišlo do aliasinga,
- kako se spektralne komponente preslikajo po vzorčenju.

Poglavje vključuje več grafičnih primerov, ki prikazujejo zrcaljenje spektra okoli večkratnikov vzorčevalne frekvence in kako se določajo zaznane frekvence.

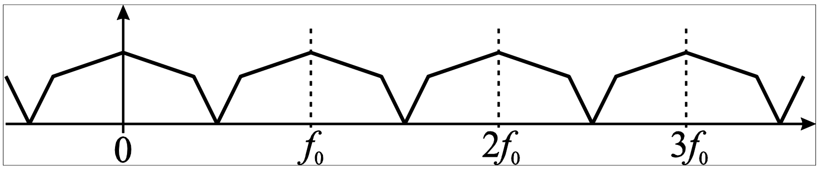
Za dodatno branje priporočamo naslednje vire:

- S. Haykin: *Communication Systems, 5th ed.* [7]
- S. Tomažič: *Osnove telekomunikacij I* [5]

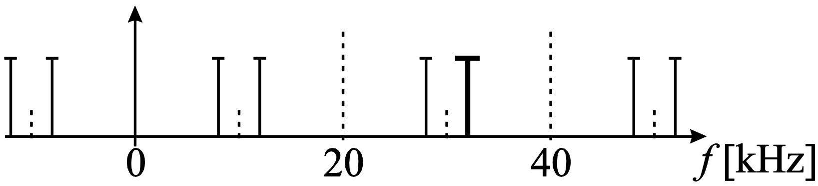
10.1 Naloge

Naloga 10.1: Z vzorčevalno frekvenco $20kHz$ vzorčimo sinusni signal s frekvenco $32kHz$. Kolikšna bo frekvenca najnižje spektralne komponente vzorčenega signala.

Rešitev: Pri vzorčenju se spekter zrcali okoli večkratnikov vzorčevalne frekvence f_0 v območju $\frac{f_0}{2}$ okoli vsakega večkratnika vzorčevalne frekvence.

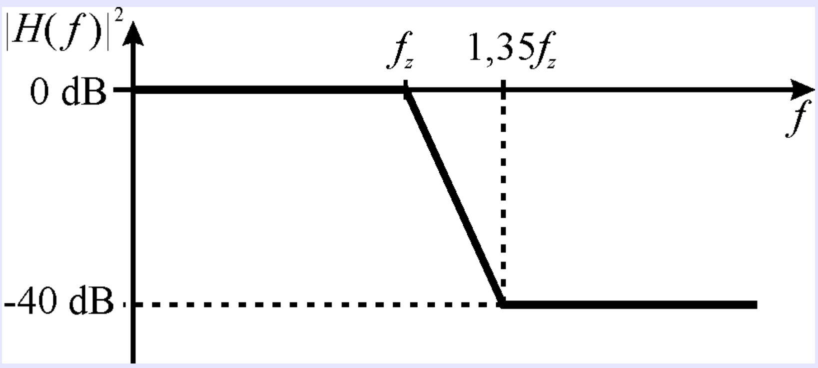


Večkratniki vzorčevalne frekvence so 20, 40, 60 kHz,... V graf vrišemo spektralno komponento merjenega signala (označena z debelo črto). Prezrcalimo jo okoli najbližjega večkratnika vzorčevalne frekvence. Nato spekter okoli tega večkratnika vzorčevalne frekvence skopiramo še okoli vseh ostalih večkratnikov vzorčevalne frekvence. Nato z grafa odčitamo frekvenco najnižje spektralne komponente



Frekvenca najnižje spektralne komponente znaša 8 kHz.

Naloga 10.2: Določite minimalno vzorčevalno frekvenco f_0, da bo velikost tujega spektra v vzorčenem signalu vsaj za $40dB$ manjša od koristnega spektra, ki je v frekvenčnem območju od 0 do $3400Hz$. Za proti-prekrivni filter uporabite nizkoprepustni filter z normiranim potekom na sliki.



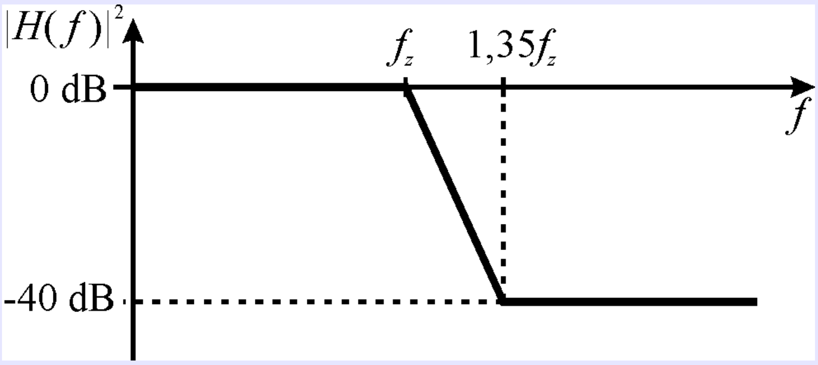
Rešitev: Tuji spekter je tisti del spektra, ki se zaradi vzorčenja preslika v osnovni spekter. V danem primeru je osnovni spekter v območju med 0 Hz in polovico vzorčevalne frekvence. Filter nastavimo tako, da bo zgornja frekvenčna meja filtra na zgornji meji koristnega signala

$$f_z = 3400Hz$$

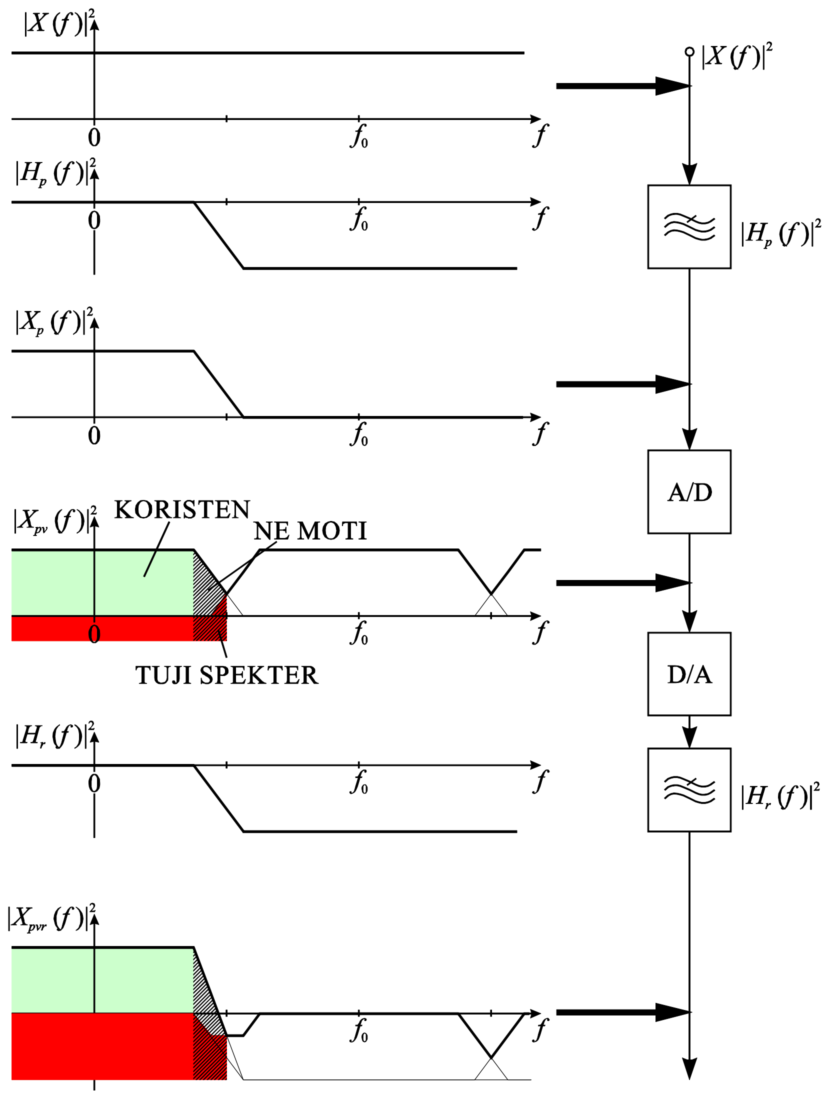
Filter dovolj zaduši spektralne komponente nad $1{,}35 f_z$, da ne motijo, tudi če se preslikajo nazaj v osnovni frekvenčni pas, ato je lahko pri tej frekvenci polovica vzorčevalne frekvence. Vzorčevalna frekvenca je torej

$f_0 = 2 \cdot 1{,}35 f_z = 2 \cdot 1{,}35 \cdot 3400$ Hz $= 9180$ Hz.

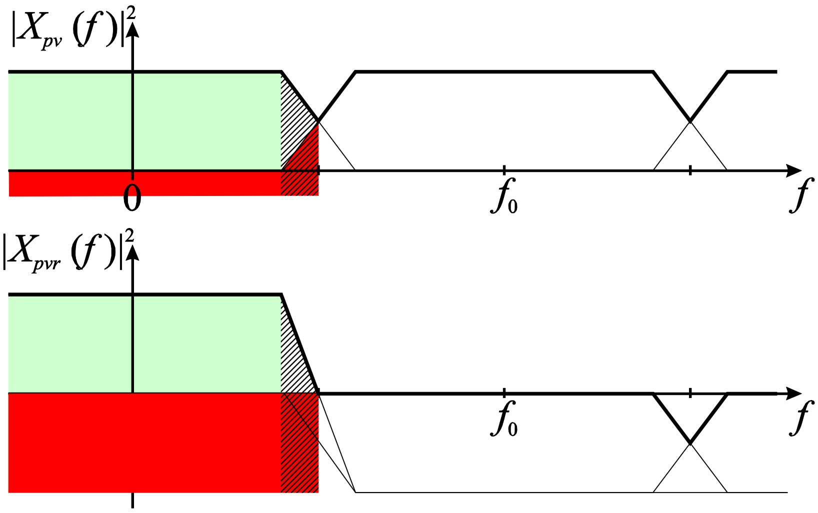
Naloga 10.3: Določite minimalno vzorčno frekvenco f_0 za rekonstrukcijo govornega signala v frekvenčnem območju od 0 do $3400Hz$. Maksimalna velikost tujega spektra sme biti $-40dB$. Za protiprekrivni in rekonstrukcijski filter uporabite nizkoprepustni filter z normiranim potekom na sliki.



Rešitev: Če je poleg proti-prekrivnega filtra znan tudi rekonstrukcijski filter, lahko pri izbiri vzorčevalne frekvence upoštevamo oba. V tem primeru lahko vzorčevalno frekvenco še malo znižamo tako, da proti-prekrivni filter ne zaduši tujega spektra popolnoma, temveč preostanek tujega spektra dovolj zaduši šele rekonstrukcijski filter. Primer je prikazan na sliki:



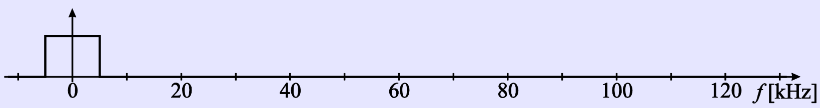
V primeru, ko je frekvenčni potek proti-prekrivnega in rekonstrukcijskega filtra enak in linearen v semi-logaritmičnem merilu, lahko polovico vzorčevalne frekvence nastavimo na sredino med zgornjo mejno frekvenco filtra in frekvenco pri kateri filter doseže zadostno dušenje. Skica:



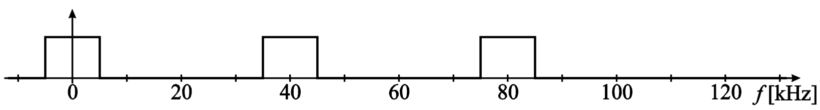
$$\frac{f_0}{2} = \frac{f_z + 1{,}35 \cdot f_z}{2}$$

$$f_0 = 2{,}35 \cdot f_z = 7990 \text{ Hz} \approx 8 \text{ kHz}$$

Naloga 10.4: Podan je dvostranski amplitudni spekter zveznega signala. Zvezni signal vzorčimo s frekvenco $40kHz$. Skicirajte spekter rekonstruiranega signala pred rekonstrukcijskim filtrom, če uporabimo regularno vzorčenje s širino impulza $\tau = \frac{T}{3}$. Določite zgornjo frekvenčno mejo f_z rekonstrukcijskega filtra in njegovo minimalno dušenje pri frekvenci $f_0 - f_z$, da bo tuji spekter zadušen vsaj za $40dB$.

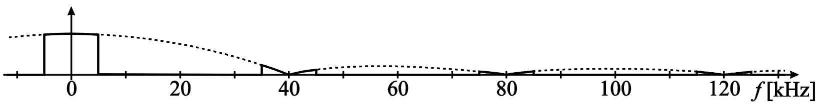


Rešitev: Če bi lahko rekonstruirali signal z idealnim vzorčenjem (z Diracovimi impulzi) bi imel rekonstruirani signal periodičen spekter:

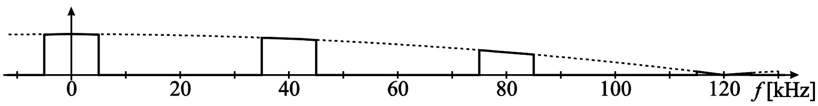


Spekter regularno vzorčenega signala, kjer je čas zadrževanja enak času med vzorci, je enak spektru idealno vzorčenega signala pomnoženim s prevajalno funkcijo zadrževalnega vezja.

$$|H_R(f)| = \frac{\tau}{T} \left| \frac{\sin(\pi \tau f)}{\pi \tau f} \right| .$$



Če je čas zadrževanja krajši od časa med vzorci, se funkcija $|HR(f)|$ ustrezno raztegne in velikost spektra signala se zniža za faktor $\frac{\tau}{T}$. Pred rekonstrukcijo velikost signala običajno ustrezno povečamo za faktor $\frac{T}{\tau}$, tako da velikost spektra ostane enako velika, a se tega dostikrat niti ne omeni:



Zgornja frekvenčna meja rekonstrukcijskega filtra f_z mora biti $5kHz$. Dušenje zadrževalnega vezja pri frekvenci 0 je

$$|H_R\,(0\,\mathrm{kHz})| = \frac{\tau}{T} = \frac{1}{3}$$

Dušenje zadrževalnega vezja pri $f_0 - f_z = 35kHz$ je

$$|H_R\,(35\,\mathrm{kHz})| = \left|\frac{\frac{T}{3}}{T} \cdot \frac{\sin\left(\pi \cdot \frac{T}{3} \cdot 35\,\mathrm{kHz}\right)}{\pi \cdot \frac{T}{3} \cdot 35\,\mathrm{kHz}}\right| = \frac{1}{3}\left|\frac{\sin\left(\pi \cdot \frac{35\,\mathrm{kHz}}{3\cdot 40\,\mathrm{kHz}}\right)}{\pi \cdot \frac{35\,\mathrm{kHz}}{3\cdot 40\,\mathrm{kHz}}}\right| = \frac{1}{3} \cdot 0{,}866$$

Glede na dušenje pri frekvenci 0 je to

$$\left|\frac{H_R(35\,\mathrm{kHz})}{H_R(0\,\mathrm{kHz})}\right| = 0{,}866$$

oziroma

$$\left|\frac{H_R(35\,\mathrm{kHz})}{H_R(0\,\mathrm{kHz})}\right|[dB] = 20 \cdot \log(0{,}866) = -1{,}25\,\mathrm{dB}$$

Skupno dušenje zadrževalnega vezja in rekonstrukcijskega filtra pri frekvenci $f_0 - f_z$ mora biti $40dB$, kar pomeni, da mora biti dušenje rekonstrukcijskega filtra

$$\begin{aligned}|H_r(f_0 - f_z)| &= -40\,\mathrm{dB} - \left|\frac{H_R(35\,\mathrm{kHz})}{H_R(0\,\mathrm{kHz})}\right|[\mathrm{dB}] \\ &= -40\,\mathrm{dB} - (-1{,}25\,\mathrm{dB}) = -38{,}75\,\mathrm{dB}\end{aligned}$$

Naloga 10.5: Spekter zveznega signala je omejen na frekvenčno območje $20{,}6$ do $23{,}6kHz$. Določite najnižjo vzorčevalno frekvenco f_0 tako, da bosta frekvenčna razmika osnovnega spektra do zrcalnih na obeh mejah enaka. Za izbrano frekvenco f_0 narišite tudi simbolično predstavitev spektra idealno vzorčenega signala

Rešitev: Nalogo lahko rešimo z uporabo neenačbe

$$\frac{2f_z}{K} \leq f_0 \leq \frac{2f_s}{K-1}$$

kjer

$$K = \text{int}\left(\frac{f_z}{f_z - f_s}\right) = \text{int}\left(\frac{23{,}6}{23{,}6 - 20{,}6}\right) = 7$$

pove, kolikšno je največje število frekvenčnih pasov dane širine, ki jih lahko umestimo na frekvenčno območje med 0 in f.

$$\begin{aligned} \frac{2f_z}{7} \leq & f_0 \leq \frac{2f_s}{7-1} \\ 6{,}743\,\text{kHz} \leq & f_0 \leq 6{,}867\,\text{kHz} \end{aligned}$$

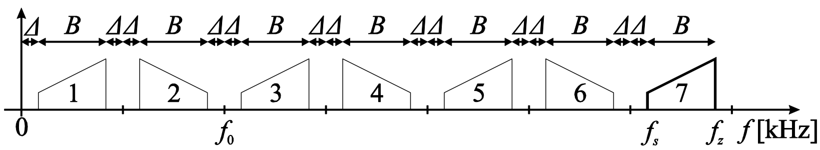
Vzorčevalno frekvenco postavimo v sredino med obe meji

$$f_0 = \frac{6{,}743\,\text{kHz} + 6{,}867\,\text{kHz}}{2} = 6{,}805\,\text{kHz}$$

Nalogo lahko rešimo tudi s pomočjo skice. Tudi v tem primeru izračunamo največje možno število frekvenčnih pasov v območju med 0 in f_z na enak način:

$$K = \text{int}\left(\frac{f_z}{f_z - f_s}\right) = \text{int}\left(\frac{23{,}6}{23{,}6 - 20{,}6}\right) = 7$$

Nato narišemo skico in iz skice zapišemo enačbe:



$B = f_z - f_s = 23{,}6\text{ kHz} - 20{,}6\text{ kHz} = 3\text{ kHz}$

$f_s = 13\Delta + 6B$

$\Delta = \dfrac{f_s - 6B}{13} = \dfrac{20{,}6\text{ kHz} - 6 \cdot 3\text{ kHz}}{13} = 0{,}2\text{ kHz}$

$f_0 = 4\Delta + 2B = 4 \cdot 0{,}2\text{ kHz} + 2 \cdot 3\text{ kHz} = 6{,}8\text{ kHz}$

$\frac{2f_z}{K} \leq f_0 \leq \frac{2f_s}{K-1}$	Nastavek za najnižjo vzorčevalno frekvenco
$K = \text{int}\left(\frac{f_z}{f_z-f_s}\right)$	dodatek k zgornji enačbi
$F_{\text{min}} = \min\limits_{k \in Z} \lvert f - kf_s \rvert$	Najnižja opažena (aliasna) frekvenca po vzorčenju

DODATEK A

ANALIZA ČASOVNIH VRST

Primer ureditve vhodnih podatkov:

```
Month,#Passengers
1949-01,112
1949-02,118
1949-03,132
1949-04,129
1949-05,121
```

S pomočjo spodnje kode, lahko potem podatke uvozimo, analiziramo in izdelamo primeren model časovne vrste.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from statsmodels.tsa.stattools import adfuller, acf, pacf
from statsmodels.tsa.seasonal import seasonal_decompose
from statsmodels.tsa.arima.model import ARIMA
import warnings
warnings.filterwarnings('ignore')

# Uvoz podatkov
df = pd.read_csv('AirPassengers.csv')
```

```
df['Month'] = pd.to_datetime(df['Month'], format='%Y-%m')
df.set_index('Month', inplace=True)
ts = df['#Passengers']

# Funkcija za testiranje stacionarnosti
def test_stationarity(timeseries,figName):

    #Determing rolling statistics
    rolmean = timeseries.rolling(12).mean()
    rolstd = timeseries.rolling(12).std()

    #Plot rolling statistics:
    plt.figure()
    orig = plt.plot(timeseries, color='blue',label='Original')
    mean = plt.plot(rolmean, color='red', label='Rolling Mean')
    std = plt.plot(rolstd, color='black', label = 'Rolling Std')

    plt.legend(loc='best')
    plt.title('Tekoče povprečje in standardni odklon')
    plt.xlabel("Leto")
    plt.ylabel("Število potnikov")
    plt.savefig(figName, format="pdf", bbox_inches="tight")
    plt.show()

    #Perform Dickey-Fuller test:
    print ('Results of Dickey-Fuller Test:')
    dftest = adfuller(timeseries, autolag='AIC')
    dfoutput = pd.Series(dftest[0:4], index=['Test Statistic','p-value',
    '#Lags Used','Number of Observations Used'])
    for key,value in dftest[4].items():
        dfoutput['Critical Value (%s)'%key] = value
    print (dfoutput)


# Naslednji koraki služijo doseganju zadovoljive stopnje stacionarnosti.

# Logaritemska pretvorba
ts_log = np.log(ts)

# Odstranitev trenda: drseče povprečje
moving_avg = ts_log.rolling(window=12).mean()
ts_log_mv_diff = ts_log - moving_avg
ts_log_mv_diff.dropna(inplace=True)

# EWMA
expwighted_avg = ts_log.ewm(halflife=12).mean()
ts_log_ewma_diff = ts_log - expwighted_avg

# Diferenciranje
```

```
ts_log_diff = ts_log.diff().dropna()

# Sezonska razčlenitev
decomposition = seasonal_decompose(ts_log, model='additive')
trend = decomposition.trend
seasonal = decomposition.seasonal
residual = decomposition.resid

# Ko dosežemo stacionarnost s pomočjo ACF in PACF
# določimo parametre ARIMA modela
# ACF in PACF
lag_acf = acf(ts_log_diff, nlags=20)
lag_pacf = pacf(ts_log_diff, nlags=20)

# ARIMA model
model = ARIMA(ts_log, order=(2, 1, 2))
results_ARIMA = model.fit()

# Inverzna transformacija in napoved
predictions_ARIMA_diff = results_ARIMA.fittedvalues
predictions_ARIMA_log = ts_log + predictions_ARIMA_diff
predictions_ARIMA = np.exp(predictions_ARIMA_log)
```

DODATEK B

ARITMETIČNO KODIRANJE

Spodnjo kodo lahko uporabimo za aritmetično kodiranje, opisano v poglavju 6.

```
def aritmeticnoKodiranje(verjetnosti,zaporedje,interval,reverse=False):
    if reverse:  # če želimo, da je stop znak na vrhu
        verjetnosti.reverse()

    new_int = [];
    output = ""
    for j in zaporedje:
        new_int = [];
        for i in range(len(verjetnosti)):
            razpon = (interval[1] - interval[0])/100
            if i == 0:
                temp = [interval[0],interval[0]
                temp += razpon*verjetnosti[i]*100]
                new_int.append(temp)
                next_b = temp[1];
            elif i == (len(verjetnosti)-1):
                temp = [next_b,interval[1]]
                new_int.append(temp)
            else:
                temp = [next_b,next_b + razpon*verjetnosti[i]*100]
```

```
                new_int.append(temp)
                next_b = temp[1]
        if not reverse:
            interval = new_int[j-1]
        else:
            interval = new_int[len(verjetnosti)-j]
        output += f"Trenutni člen je {j}"
        output += f"interval pa je {interval[0]:.5f}"
        output += f"------{interval[1]:.5f}\n"
    print(output)
```

Primer uporabe:

```
verjetnosti = [4/9,1/9,3/9,1/9]
zaporedje = [2,3,3,1,4]
interval = [0.0,1.0]
reverse = False
aritmeticnoKodiranje(verjetnosti,zaporedje,interval,reverse)
```

DODATEK C

LZW (DE)KODIRANJE

Spodnji funckiji uporabimo za LZW kodiranje in dekodiranje. Pri tem smo pozorni, da je potrebno ustrezno inicializirati slovar (ločeno podati začetne simbole).

```
def kodiraj(uncompressed,dictionary):
    dict_size = len(dictionary)
    w = ""
    result = []
    for c in uncompressed:
        wc = w + c
        if wc in dictionary: #preverim če sem kombinacijo že srečal
            w = wc
        else:
            result.append(dictionary[w])
            # Dodam v slovar.
            dictionary[wc] = dict_size
            dict_size += 1
            w = c
    # Dodam kodo v kodirano vsebino
    if w:
        result.append(dictionary[w])
    print("Kodirni slovar:", dictionary)
    return result
```

```
def dekodiraj(compressed,dictionary):
    from io import StringIO
    dict_size = len(dictionary)
    result = StringIO()
    w = dictionary[compressed.pop(0)]
    result.write(w)
    for k in compressed:
        if k in dictionary:
            entry = dictionary[k]
        elif k == dict_size:
            entry = w + w[0]
        else:
            raise ValueError('Napaka pri kodiranju k: %s' % k)
        result.write(entry)

        # Razširim slovar
        dictionary[dict_size] = w + entry[0]
        dict_size += 1

        w = entry
    print("Dekodirni slovar:", dictionary)
    return result.getvalue()
```

Primer uporabe:

```
# Niz, ki ga kodiramo
input_string = "BITIALINEBITI"

# Pripravimo začetna slovarja
char_to_num = {}
num_to_char = {}
count = 0

# Preletimo vhodni niz
for char in input_string:
    if char not in char_to_num:
        # Če znak ni v slovarju, ga dodamo
        char_to_num[char] = count
        num_to_char[count] = char
        count += 1

# Izpis začetnih slovarjev
print("Kodirni slovar:", char_to_num)
print("Dekodirni slovar:", num_to_char)

# Kodiramo
compressed = kodiraj(input_string,char_to_num)
print ("Kodirana vsebina:", compressed)
```

```
# Dekodiramo
decompressed = dekodiraj(compressed,num_to_char)
print ("Dekodirana vsebina:", decompressed)
```

DODATEK D

HUFFMANOVO KODIRANJE

```
# Razred za vozlišče Huffmanovega drevesa
class Vozlisce:
    def __init__(self, verjetnost, simbol, levo=None, desno=None):
        self.verjetnost = verjetnost  # verjetnost simbola
        self.simbol = simbol          # znak (simbol)
        self.levo = levo              # levo poddrevo
        self.desno = desno            # desno poddrevo
        self.koda = ''                # dodeljena binarna koda

# Slovar za shranjevanje kod posameznih simbolov
kode = dict()

def izracunaj_kode(drevo, trenutna_koda=''):
    """
    Rekurzivni sprehod po drevesu za določitev kod posameznih simbolov.
    """
    nova_koda = trenutna_koda + str(drevo.koda)

    if drevo.levo:
        izracunaj_kode(drevo.levo, nova_koda)
    if drevo.desno:
        izracunaj_kode(drevo.desno, nova_koda)
```

```
    if not drevo.levo and not drevo.desno:
        kode[drevo.simbol] = nova_koda

    return kode

def prestej_frekvence(besedilo):
    """
    Prešteje frekvence pojavitve posameznih simbolov.
    """
    frekvence = dict()
    for znak in besedilo:
        frekvence[znak] = frekvence.get(znak, 0) + 1
    return frekvence

def zakodiraj(besedilo, kodiranje):
    """
    Zakodira besedilo s pomočjo slovarja kod.
    """
    return ’’.join([kodiranje[znak] for znak in besedilo])

def izpisi_kompresijo(besedilo, kodiranje):
    """
    Izpiše število bitov pred in po stiskanju.
    """
    pred = len(besedilo) * 8
    po = sum(besedilo.count(simbol) * len(koda) for simbol, koda in kodiranje.it
    print("Bitov pred stiskanjem:", pred)
    print("Bitov po stiskanju:", po)

def huffmanovo_kodiranje(besedilo):
    """
    Glavna funkcija za gradnjo Huffmanovega drevesa in kodiranje podatkov.
    Vrne: zakodirano besedilo in koren drevesa.
    """
    frekvence = prestej_frekvence(besedilo)
    vozlisca = [Vozlisce(frekvence[s], s) for s in frekvence]

    while len(vozlisca) > 1:
        vozlisca = sorted(vozlisca, key=lambda x: x.verjetnost)
        levo = vozlisca.pop(0)
        desno = vozlisca.pop(0)

        levo.koda = ’0’
        desno.koda = ’1’

        zdruzeno = Vozlisce(
            levo.verjetnost + desno.verjetnost,
            levo.simbol + desno.simbol,
```

```
            levo,
            desno
        )
        vozlisca.append(zdruzeno)

    drevo = vozlisca[0]
    kodne_kombinacije = izracunaj_kode(drevo)
    print("Kode simbolov:", kodne_kombinacije)
    izpisi_kompresijo(besedilo, kodne_kombinacije)
    kodirano = zakodiraj(besedilo, kodne_kombinacije)
    return kodirano, drevo

def huffmanovo_dekodiranje(kodirano_besedilo, drevo):
    """
    Dekodira besedilo na osnovi Huffmanovega drevesa.
    """
    izhod = []
    trenutni = drevo

    for bit in kodirano_besedilo:
        trenutni = trenutni.desno if bit == '1' else trenutni.levo

        if not trenutni.levo and not trenutni.desno:
            izhod.append(trenutni.simbol)
            trenutni = drevo

    return ''.join(izhod)

# === TESTNI PRIMER ===
if __name__ == "__main__":
    besedilo = "AAAAAAABCCCCCCDDEEEEE"
    print("Izvirno besedilo:", besedilo)
    kodirano, drevo = huffmanovo_kodiranje(besedilo)
    print("Kodirano besedilo:", kodirano)
    print("Dekodirano besedilo:", huffmanovo_dekodiranje(kodirano, drevo))
```

VIRI

[1] Papoulis, A., Pillai, S. U. (2002). *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill.

[2] Cover, T. M., Thomas, J. A. (2006). *Elements of Information Theory*, 2nd ed., Wiley-Interscience.

[3] MacKay, D. J. C. (2003). *Information Theory, Inference and Learning Algorithms*, Cambridge University Press.

[4] Shannon, C. E. (1948). *A Mathematical Theory of Communication*, Bell System Technical Journal.

[5] Tomažič, S. (2002). *Osnove telekomunikacij I*, Založba FE in FRI.

[6] Pavešič, N. (2010). *Informacija in kodi*, Založba FE in FRI.

[7] Haykin, S. (2009). *Communication Systems*, 5th ed., John Wiley & Sons, Inc., New York.

https://fides.fe.uni-lj.si/~matevzk/Ucbeniki/
TIIK/TIIK_Kunaver.pdf