Fakulteta za elektrotehniko
Univerza v Ljubljani

Andrej Vilhar

Teoretične meje digitalnega prenosa

seminarska naloga

Mentor: prof. dr. Sašo Tomažič

Ljubljana, maj 2005

Povzetek

Seminarska naloga je del obveznosti pri predmetu Digitalne komunikacije na podiplomskem študiju. V njej obravnavamo temeljne oziroma teoretične meje prenosa informacij. Osredotočamo se predvsem na osnovne principe in manj na konkretne primere iz sodobne tehnologije, pri čemer podamo tudi izpeljave, po katerih pridemo do končnih rezultatov. Zahtevnejšim matematičnim opisom se poskušamo izogibati, kjer je to za pravilno razumevanje možno.

V prvem poglavju na kratko opišemo zgodovino razvoja informacijske teorije, ki jo je leta 1948 za vedno zaznamoval Claude Shannon. V drugem poglavju podamo strukturo splošnega komunikacijskega sistema, ki služi kot osnovni model pri nadaljnji obravnavi. Nato v tretjem poglavju opišemo izvorno stran komunikacijskega sistema. Definiramo pojma informacije in entropije ter razložimo prvi pomemben teorem, teorem o izvornem kodiranju. V četrtem poglavju se osredotočimo na matematično razlago prenosa informacij. Definiramo kapaciteto kanala in podamo teorem o kanalskem kodiranju. V petem poglavju opisani teorem o informacijski kapaciteti predstavlja povezavo med prej podanimi matematičnimi pojmi in fizikalnimi količinami, ki določajo teoretične meje. Ogledamo si tudi drugačen, grafičen pristop k razlagi teorema o informacijski kapaciteti. Na koncu podamo še izpeljavo t.i. Shannonove meje ter nakažemo, koliko so se ji uspeli približati nekateri realni primeri kodno-modulacijskih postopkov. V zaključku ocenimo, kaj za področje telekomunikacij Shannonova teorija pomeni in ugotovimo, da bo zaradi nedosegljivosti teoretičnih mej to področje iz praktičnega vidika vedno aktualno.

Kazalo

1 UVOD 6
2 SPLOŠNI KOMUNIKACIJSKI SISTEM 9
3 IZVORNA STRAN KOMUNIKACIJSKEGA SISTEMA 10
3.1 Informacija 10
3.2 Entropija 11
3.3 Teorem o izvornem kodiranju 12
3.3.1 Primeri postopkov za izvorno kodiranje 13
3.4 Odvisnost naključnih spremenljivk 15
4 PRENOS INFORMACIJ 18
4.1 Diskretni informacijski kanal 18
4.2 Vzajemna informacija 19
4.3 Kapaciteta kanala 20
4.4 Teorem o kanalskem kodiranju 21
5 FIZIKALNA SLIKA PRENOSA INFORMACIJ 25
5.1 Entropija in vzajemna informacija pri zveznih naključnih spremenljivkah 25
5.2 Gaussov kanal 26
5.3 Teorem o informacijski kapaciteti 27
5.3.1 Kapaciteta kanala v bitih na simbol 28
5.3.2 Nyquistov kriterij 29
5.3.3 Kapaciteta kanala v bitih na sekundo 31
5.4 Kapaciteta kanala z nebelim šumom 33
5.5 Geometrična predstavitev teorema o informacijski kapaciteti 35
5.6 Shannonova meja in primeri kodno-modulacijskih postopkov 39
6 ZAKLJUČEK 42

Kazalo slik

Slika 1: Claude Elwood Shannon (1916-2001) [1] 6
Slika 2: Splošni komunikacijski sistem [3] 9
Slika 3: Diskretni informacijski kanal brez spomina [4] 18
Slika 4: Ilustracija različnih vrst entropij, vzajemne informacije in relacij med njimi [4] ... 20
Slika 5: Shematičen prikaz preslikave sporočil med vhodom in izhodom diskretnega informacijskega kanala [3] 22
Slika 6: Dosegljivo področje negotovosti po sprejemu, glede na entropijo izvora pri podani kapaciteti kanala [3] 23
Slika 7: Gaussov kanal brez spomina 27
Slika 8: Vpliv povečevanja pasovne širine B oziroma moči signala P_S na povečevanje kapacitete kanala 32
Slika 9: Najboljša porazdelitev gostote moči signala pri nebelem šumu [9] 34
Slika 10: Problem pakiranja krogel (*sphere-packing problem*) [10] 38
Slika 11: Spektralna učinkovitost v odvisnosti od razmerja energije bita proti gostoti šumne moči [11] 40
Slika 12: Verjetnost bitne napake v odvisnosti od razmerja energije bita proti gostoti šumne moči [11] 41

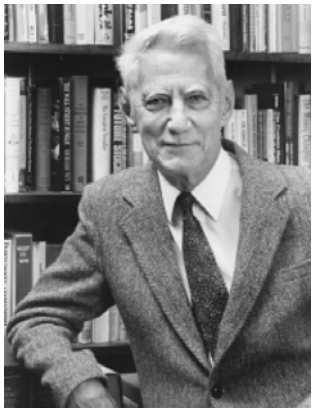
Kazalo tabel

Tabela 1: Povezava med komunikacijskim sistemom in geometrijskimi entitetami 37

1 Uvod

Ljudje se že od nekdaj sporazumevamo s simboli. Ti simboli so lahko bodisi akustični bodisi vizualni. V prazgodovini so uporabljali omejeno godrnjanje in renčanje, kretnje z rokami ter slike na jamskih stenah. Kasneje so se ti simboli izpopolnjevali. Nastal je jezik, širši nabor simbolov, s katerim je bilo mogoče predstaviti več misli in idej. Prišla je pisana beseda, sprva na kamne, kasneje na papirus in papir. Sporazumevanje s simboli se je ohranilo, zamenjal pa se je način, kako jih prenašamo, shranjujemo in nenazadnje kako razumemo njihov pomen.

K razumevanju je kot začetnik informacijske teorije pred več kot 50 leti bistveno pripomogel Claude Elwood Shannon. Leta 1948 je v dveh delih objavil študijo z naslovom *»A Mathematical Theory of Communication«*. Gre za temeljno delo, ki na področju informacijske teorije in teoretičnih mej komunikacij igra vlogo nekakšne »biblije«. Mnogo matematikov in inženirjev je ta študija navdahnila za njihovo nadaljnje delo bodisi na teoretičnem bodisi na praktičnem področju. Še danes razvijalcem komunikacijskih sistemov in različnih principov kodiranja, navkljub številnim novejšim primerom literature s tega področja, to delo predstavlja pomembno osnovo. Navsezadnje teoretične meje, ki jih je Shannon teoretično utemeljil, veljajo za vekomaj in se jim v sedanjosti ter prihodnosti praktično lahko le bolj ali manj približamo.



Slika 1: Claude Elwood Shannon (1916-2001) [1]

Kljub veliki odmevnosti zgoraj opisanega Shannonovega dela in priznavanju njegovih zaslug pa je potrebno omeniti, da pri tem vendarle ni šlo za nenadno odkritje. Njegov članek je bil posledica oziroma logično, pa vendar genialno, nadaljevanje predhodnih raziskav drugih raziskovalcev, kot tudi njegovega lastnega raziskovalnega dela ter nenazadnje znanj, pridobljenih na že obstoječih telekomunikacijskih sistemih. Pred letom 1948 so že poznali telegrafijo, telefonijo,

AM, FM, SSB modulacije, brezžično telegrafijo, televizijo, teleskriptor, PCM, vocoder in razširjeni spekter. V telegrafiji so tako na primer z Morsovo kodo že izkoriščali povprečne pogostosti posameznih črk v abecedi, FM, PCM in razširjeni spekter so nakazovali na možnost zamenjave pasovne širine za razmerje signal/šum, ne da bi pri tem vplivali na učinkovitost prenosa, z vocoderjem pa je bilo moč doseči prenose pri sicer manjši razumljivosti vendar večji izkoriščenosti pasovne širine.

H. Nyquist je že leta 1924 predpostavil, da je hitrost prenosa sorazmerna logaritmu števila nivojev signala glede na časovno enoto. Spraševal se je tudi že o tem, koliko bi bilo moč pridobiti, če bi Morsejevo kodo nadomestili z neko drugo, optimalno kodo. Za komunikacijsko teorijo pa je bila ključnega pomena njegova utemeljitev na podlagi teorema o vzorčenju, da za verodostojno rekonstrukcijo signala zadostuje $2TB$ vzorcev, pri čemer je T čas trajanja signala, B pa njegova pasovna širina. Do podobnih ugotovitev je kasneje prišel tudi D. Gabor. Najhitrejšo oddajo telegrafskih signalov glede na frekvenčno širino sta poleg Nyquista študirala še K. Küpfmüller in V. Kotel'nikov. Leta 1928 je R. Hartley že uporabljal izraze kot so hitrost prenosa, intersimbolna interferenca, kapaciteta prenosa. Ugotovil je, da je prenosna kapaciteta sistema sorazmerna s pasovno širino ter uvedel mero za količino informacije, ki jo je označil s črko H, kot sledi iz enačbe (1).

$$H = n \cdot \log s \tag{1}$$

Pri tem je n število izborov izmed s različnih simbolov. Princip, da je informacija določen izid izmed končne množice možnih izidov, je bil torej definiran že takrat.

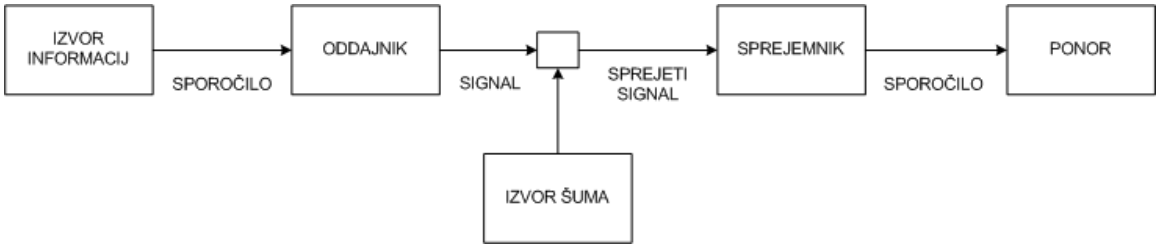
Nyquist in Hartley pa nista kvantitativno določila vpliva šuma na prenos niti nista upoštevala verjetnostnega modeliranja izvorov informacij. Za realizacijo slednjega problema sta verjetno najbolj zaslužna N. Wiener in S. Rice, za upoštevanje prvega pa C. Shannon.

Skratka, pri oblikovanju informacijske in komunikacijske teorije je v tistem času sodelovalo več raziskovalcev. Do leta 1948 je tako kopica znanstvenikov poskušala oblikovati univerzalno teorijo, ki bi upoštevala razmerje oziroma kompromis med hitrostjo prenosa, zanesljivostjo, pasovno širino in razmerjem signal/šum. Med njimi so bili A. Clavier, C. Earp, S. Goldman, J. Laplume, W. Tuller, N. Wiener in seveda tudi C. Shannon, čigar teorija se je konec koncev tudi za vedno obdržala. Poleg zgoraj omenjenih faktorjev, ki so vplivali na nastanek Shannonove teorije, pa je verjetno na njen nastanek pozitivno vplivalo tudi njegovo lastno predhodno raziskovalno delo. Njegovo magistrsko delo, pri katerem je utemeljil uporabo Boolove algebre v preklopnih sistemih je bilo celo nagrajeno z Nobelovo nagrado, v doktoratu se je ukvarjal z genetiko, do ključnih spoznanj o komunikacijskih in informacijskih sistemih pa naj bi Shannon prišel med drugo svetovno vojno, ko je preučeval kriptografijo in o tem spisal takrat še tajno študijo, ki so jo lahko objavili šele nekaj let po vojni.

V seminarski nalogi se opiramo zlasti na njegovo interpretacijo in teoretično utemeljitev mej prenosa informacij. Poudarek je na razumevanju osnovnih principov in manj na konkretnih primerih iz digitalnih komunikacij.

2 Splošni komunikacijski sistem

C. Shannon je pri obravnavi telekomunikacij privzel splošni komunikacijski sistem, kot sledi s slike 2.



Slika 2: Splošni komunikacijski sistem [3]

Cilj takšnega sistema je, da se določeno sporočilo, izbrano izmed več možnih, verodostojno prenese od izvora do ponora informacij. Pri tem je sporočilo lahko praktično karkoli, od zvoka in slike do pisanih črk in številk. Naloga oddajnika je, da prevede izbrano sporočilo v izbrano signalno obliko, ki jo lahko prenašamo preko komunikacijskega kanala. Danes oddajnik razumemo v dveh vlogah, izvornemu kodirniku in kanalskemu kodirniku. Prvi skrbi za predstavitev sporočila v neki standardni obliki (binarni zapis), ki jo nato drugi pretvori v fizikalno količino (npr. napetost), ki ji rečemo signal in je primerna za komunikacijski kanal. Na komunikacijskem kanalu se v vsakem realnem sistemu oddanemu signalu prišteje šum. Tako spremenjenega sprejme sprejemnik, ki izvrši obratno operacijo od oddajnika, sprejetemu signalu pripiše eno od možnih sporočil. Pri tem opisu splošnega komunikacijskega sistema, ki velja v vsakem realnem sistemu se porodijo vprašanja: kaj informacija sploh je, koliko je potrebujemo za predstavitev nekega sporočila in koliko jo lahko zanesljivo prenesemo preko komunikacijskega kanala. Z odgovorom na ta vprašanja postavimo teoretične meje prenosa, ki veljajo za splošen oziroma poljuben komunikacijski sistem.

3 Izvorna stran komunikacijskega sistema

Najprej se v splošnem komunikacijskem sistemu osredotočimo samo na izvor informacij, ki je na sliki 2 uprizorjen s skrajno levim blokom. S tem bomo definirali kaj je pravzaprav tisto, kar v komunikacijskem sistemu prenašamo.

3.1 Informacija

Vzemimo, da imamo na voljo diskreten izvor, ki vsako časovno enoto odda enega izmed možnih simbolov v abecedi. Vsak simbol se pojavi z določeno verjetnostjo. Rečemo, da opazujemo izid diskretne naključne spremenljivke *S*. Verjetnost, da bo nastopil *k*-ti dogodek, oziroma da bo izbran *k*-ti simbol s_k iz abecede, je podana z enačbo (2).

$$P(S = s_k) = p_k \tag{2}$$

Vsota vseh verjetnosti, da bo izbran eden od skupno *K* možnih simbolov, je seveda enaka 1, kot to podaja enačba (3).

$$\sum_{k=1}^{K} p_k = 1 \tag{3}$$

Če poznamo porazdelitev verjetnosti za naključno spremenljivko *S*, lahko z določeno gotovostjo napovemo, kakšen bo naslednji izid. Ko za izid izvemo, smo do določene mere nad izidom presenečeni, odvisno od gotovosti vnaprejšnje napovedi. Bolj kot smo o izidu negotovi oziroma bolj kot smo po izidu presenečeni, večja je njegova informacija. Informacija je torej nekakšno merilo presenečenja. Manjša kot je verjetnost izida, večje bo presenečenje in večja bo informacija. Kot sta to napovedala že Nyquist in Hartley in kot je to po njima privzel tudi Shannon, je za ustrezno merilo informacije potrebno uporabiti logaritemsko funkcijo. Zapisali so, da je v primeru *n* enako verjetnih dogodkov informacija enega enaka logaritmu njihovega števila *n*, kot sledi iz enačbe (4).

$$I = \log n \tag{4}$$

Če verjetnosti niso enake, pa je informacijo o posameznem dogodku potrebno izračunati po natančnejši enačbi (5).

$$I(s_k) = \log\left(\frac{1}{p_k}\right) = -\log p_k \tag{5}$$

Osnova logaritma pri tem pomeni izbiro enote. Če vzamemo osnovo 2, kar je glede na preklopni značaj informacijsko komunikacijskih sistemov tudi najbolj ustrezna izbira, je enota informacije podana v **bit**ih (**bi**nary digi**t**s). Besedo bit je že pred Shannonovim slavnim delom predlagal J. W. Turkey.

Takšna definicija informacije ustreza naslednjim intuitivnim zahtevam:

1. Informacija gotovega dogodka je enaka 0, saj po njegovem izidu ni nikakršnega presenečenja oziroma ga lahko brez možnosti napake napovemo vnaprej (6).

$$I(s_k) = 0 \quad za \quad p_k = 1 \tag{6}$$

2. Posamezen dogodek lahko bodisi vsebuje informacijo, bodisi informacije ne vsebuje. Izguba informacije oziroma negativna informacija ni možna (7).

$$I(s_k) \geq 0 \quad za \quad 0 \leq p_k \leq 1 \tag{7}$$

3. Dogodek, ki je manj verjeten nosi več informacije od tistega, ki je bolj verjeten (8).

$$I(s_k) > I(s_i) \quad za \quad p_k < p_i \tag{8}$$

4. Če sta dva dogodka neodvisna, je vsota informacij posameznih dogodkov enaka skupni informaciji, ki jo dogodka vsebujeta (9).

$$I(s_k s_i) = I(s_k) + I(s_i) \tag{9}$$

3.2 Entropija

Po zgornji definiciji informacije ugotovimo, da v splošnem količina informacije niha, oziroma da je odvisna od vsakokratnega izbora enega od možnih simbolov. Zato po enačbi (10) izračunamo povprečno informacijo izvora s poznano porazdelitvijo verjetnosti naključne spremenljivke.

$$H_S = \sum_{k=1}^{K} p_k I(s_k) = \sum_{k=1}^{K} p_k \log_2 \frac{1}{p_k} \tag{10}$$

Količino H_S imenujemo entropija, po analogiji iz Boltzmannove teorije v termodinamiki. Indeks *S* pri tem pomeni, da se nanaša na abecedo oziroma naključno spremenljivko *S*. Entropija nam torej pove, koliko informacije v povprečju dobimo, če izvemo za enkratni izid dogodka oziroma enkratni izbor simbola iz celotne abecede.

Za entropijo diskretnega izvora brez spomina velja, da je navzgor in navzdol omejena, kot sledi iz enačbe (11).

$$0 \le H_S \le \log_2 K \tag{11}$$

Pri tem spodnjo mejo $H_S = 0$ dosežemo le, če je verjetnost za nek simbol $p_k = 1$, za preostale pa $p_k = 0$.

Da vedno velja $H_S \ge 0$, vidimo iz enačbe (10), saj je pri pogoju $0 \le p_k \le 1$ vsak člen $p_k \log_2 1/p_k$ vedno nenegativno število. Dalje za taisti člen ugotovimo, da je točno enak nič le ob pogoju da je p_k enak bodisi 0 ali 1 in mora torej za njihovo skupno vsoto 0 veljati, da je verjetnost enega simbola 1 ostalih pa 0.

Dokaz zgornje meje sledi iz izraza (12), kjer sta p_k in q_k poljubni verjetnostni porazdelitvi simbolov iz iste abecede.

$$\sum_{k=1}^{K} p_k \log_2 \frac{q_k}{p_k} \tag{12}$$

V tem izrazu upoštevamo neenačbo (13) in dobimo neenačbo (14).

$$\ln x \le x - 1 \quad , \; x \ge 0 \tag{13}$$

$$\sum_{k=1}^{K} p_k \log_2 \frac{q_k}{p_k} \le 0 \tag{14}$$

Po vstavitvi enačbe $q_k = 1/K$ v neenačbo (14) dobimo končno rešitev $H_S \le \log_2 K$.

Največjo povprečno informacijo pri podani končni abecedi simbolov torej dosežemo, če je izbor vsakega izmed simbolov enako verjeten. Pri tem več možnih simbolov seveda pomeni večjo povprečno informacijo.

3.3 Teorem o izvornem kodiranju

Nek niz simbolov oziroma vhodno informacijo, ki jo želimo prenesti, lahko predstavimo z binarnim nizom. Entropija vhodnega niza pri tem natančno določa, koliko je minimalno število bitov, ki jih v povprečju potrebujemo za predstavitev enega simbola. Za dosego tega cilja moramo poznati statistiko izvora in jo

izkoristiti tako, da simbole z manjšo verjetnostjo zapišemo z daljšimi, simbole z večjo verjetnostjo pa s krajšimi kodnimi besedami. Tak princip je bil uporabljen že pri Morsejevi abecedi, kjer je bila npr. črka E, ki je zelo pogosta, predstavljena samo z eno piko, medtem ko je bil Q, ki se pojavi občutno redkeje predstavljen s tremi črticami in eno piko, torej skupno štirimi znaki.

Če imamo na vhodu diskreten izvor s simboli s_k in pripadajočimi verjetnostmi p_k ter vsak simbol s_k predstavimo s kodno besedo dolžine l_k, potem je število bitov, ki jih v povprečju potrebujemo za zapis enega izmed skupno K simbolov, enako $\overline{L}$, podano z enačbo (15).

$$\overline{L} = \sum_{k=1}^{K} p_k l_k \tag{15}$$

Po Shannonovi teoriji velja neenačba (16), kjer je H_S entropija izvora.

$$\overline{L} \geq H_S \tag{16}$$

Teorem o izvornem kodiranju torej pravi, da je najmanjša povprečna dolžina kodne besede L_{min}, ki jo potrebujemo za zapis simbola, natanko enaka entropiji izvora. Ker se v splošnem temu rezultatu lahko le približamo, ne moremo pa ga natanko doseči, lahko efektivnost izvornega kodiranja izrazimo z razmerjem, podanim z enačbo (17).

$$\eta = \frac{L_{\min}}{\overline{L}} = \frac{H_S}{\overline{L}} \tag{17}$$

Bolj kot se η približa enoti, boljša je efektivnost kodiranja in za prenos je potreben manjši bitni pretok. Če imamo opravka z brezizgubnim kodiranjem oziroma t.i. entropijskim kodiranjem, po katerem še lahko pravilno dekodiramo, η ne more biti večji od 1, ki torej predstavlja teoretično mejo. V primeru, da je η manjši od 1, pa imamo opravka z redundanco, to je z odvečno informacijo, ki za enolično predstavitev izvornih simbolov ni potrebna. Cilj izvornega kodiranja je zmanjšati redundanco in s tem povečati njegovo efektivnost.

3.3.1 Primeri postopkov za izvorno kodiranje

Za dosego cilja zmanjšanja redundance oziroma približanju entropiji izvora, obstoji vrsta kodirnih postopkov. V primeru diskretnih izvorov brez spomina, ki smo jih obravnavali do sedaj, so to na primer kodiranje s predpono, Huffmanovo kodiranje, Lempel-Ziv kodiranje. Vsi našteti postopki delujejo po principu, da se pogostejše simbole oziroma vzorce kodira z manj biti in seveda predpostavljajo da je po dekodiranju možna popolna rekonstrukcija prvotnega vhodnega niza.

3.3.1.1 Kodiranje s predpono

Princip kodiranja s predpono (angl. *prefix coding*) je, da kodna beseda kateregakoli simbola ni hkrati tudi predpona kodne besede kateregakoli drugega simbola. Če bi na primer simbol s_k kodirali s kodno besedo z zaporednimi biti m_{k1}, m_{k2}, … m_{kn}, kjer je n dolžina kodne besede, so njene predpone vse kode z zaporedjem bitov m_{k1}, … m_{ki}, kjer je $i \leq n$. Pomembni lastnosti tega postopka sta, da je pri dekodiranju ob vsakokratnem poznavanju cele kodne besede, zakodiran simbol že enolično določen in da je možno doseči povprečno dolžino kodne besede, za katero velja neenačba (18).

$$H_S \leq \overline{L} < H_S + 1 \tag{18}$$

Če je porazdelitev verjetnosti posameznih simbolov s_k takšna, da je verjetnost vsakega simbola enaka 2^{-lk}, kjer je l_k dolžina k-te kodne besede, potem v levi neenakosti neenačbe (18) velja enačaj in je torej povprečna dolžina kodne besede že enaka entropiji. V splošnejšem primeru, kjer porazdelitev verjetnosti ni takšna, pa lahko uvedemo t.i. razširjeno kodo, kjer zakodiramo n simbolov hkrati. Če s H_{Sn} označimo entropijo na ta način novo nastale abecede, kjer vsaka kombinacija n simbolov pomeni nov simbol, je moč pokazati, da velja $H_{Sn} = nH_S$. Označimo sedaj še z $\overline{L}_n$ povprečno dolžino kodne besede tako razširjenega kodiranja in dobimo neenačbo (19), ki jo lahko zapišemo tudi v obliki neenačbe (20).

$$nH_S \leq \overline{L}_n < nH_S + 1 \tag{19}$$

$$H_S \leq \frac{\overline{L}_n}{n} < H_S + \frac{1}{n} \tag{20}$$

Vidimo torej, da se s povečevanjem reda kodiranja n zgornja in spodnja meja enačbe (20) približujeta. Povprečno dolžino kodne besede na posamezen simbol lahko torej poljubno približamo vrednosti entropije izvora, če le dovolj povečamo red kodiranja n. Seveda pa moramo zato plačati tudi ceno – kompleksnost kodirnega sistema raste z naraščanjem n. Ko gre n preko vseh meja, gre 1/n* $\overline{L}_n$ proti H_S, preko vseh meja pa gre tudi kompleksnost sistema.

3.3.1.2 Huffmanovo kodiranje

Huffmanovo kodiranje je pravzaprav poseben primer kodiranja s predpono, kjer kodne besede dodeljujemo simbolom po določenem algoritmu, ki ga imenujemo Huffmanov algoritem. Kratek opis postopka kodiranja po tem algoritmu je sledeč:

- posamezne simbole se uredi po njim pripadajočih verjetnostih,

- najmanj verjetnima simboloma se dodeli vrednosti bitov 0 in 1 ter se ju združi v skupni simbol z vsoto njunih verjetnosti,
- izmenično se ponavlja prva dva koraka, dokler ne preostaneta le še dva simbola, ki se jima dodeli še zadnja bita 0 in 1,
- kodne besede za posamezen simbol se prebere v obratni smeri, kot je bil postopek dodeljevanja bitov.

Postopek je dokaj enostaven in učinkovit, lahko pa ga tudi razširimo z združevanjem simbolov, kot je bilo opisano v zgornjem poglavju, s čimer se s povprečno dolžino kodne besede na simbol še bolj približamo entropiji izvora.

3.3.1.3 Lempel-Ziv kodiranje

V dosedanji obravnavi smo predpostavljali diskreten izvor simbolov brez spomina. Verjetnosti simbolov so bile pri tem poznane vnaprej, simboli pa med seboj niso bili odvisni. Pri taki predpostavki je Huffmanovo kodiranje primerno. Pri realnejših, primerih, kjer porazdelitve verjetnosti simbolov niso poznane vnaprej in predvsem, kjer simboli med seboj niso popolnoma neodvisni, pa Huffmanovo kodiranje ni najboljše. Tipičen primer kodirnega postopka, ki opisani lastnosti diskretnega izvora upošteva, je Lempel-Ziv kodiranje.

Osnovni princip po katerem ta postopek deluje je, da kreira lastno kodno knjigo, v katero po vrsti vpisuje najkrajše segmente v izvornem nizu, kateri mu še niso poznani. Te segmente nato opiše s tistimi, ki so mu že poznani in enim dodatnim bitom. Vsem elementom v kodni knjigi dodeli kodno besedo enake dolžine *l*, pri čemer prvih *l-1* bitov pomeni zaporedno mesto, pod katerim se nahaja že poznani del, zadnji bit pa ustreza novemu, še nepoznanemu bitu v izbranem segmentu. Postopek kodiranja in dekodiranja je dokaj enostaven, velikost kodne knjige pa je odvisna od izbire dolžine kodnih besed. V praksi se izbira dolžino 12 in je torej možnih 4096 vnosov posameznih segmentov.

Lempel-Ziv algoritem je danes v množični uporabi pri stiskanju datotek in na primer pri stiskanju teksta prihrani do 55% pomnilniškega prostora, medtem ko znaša prihranek pri Huffmanovem algoritmu le 43%.

3.4 Odvisnost naključnih spremenljivk

Pri zadnjem primeru kodirnega postopka, Lempel-Ziv, smo že nakazali, da v splošnem posamezni simboli diskretnega izvora med seboj niso neodvisni. Videli smo tudi, da je upoštevanje tega dejstva pomembno za učinkovito kodiranje. Za natančnejšo predstavo o medsebojni odvisnosti simbolov pa je potrebno vpeljati nekaj novih pojmov.

Vzemimo, da opazujemo dva dogodka oziroma dve naključni spremenljivki, *X* in *Y*. Verjetnost, da se bo zgodil določen izid posamezne naključne spremenljivke, je definirana z enačbo (21).

$$p(x_i) = P(X = x_i) \ , \quad p(y_j) = P(Y = y_j) \tag{21}$$

Verjetnost skupnega dogodka, to je da se zgodi določen izid prve in določen izid druge naključne spremenljivke pa definirajmo z enačbo (22).

$$p(x_i, y_j) = P(X = x_i, Y = y_j) \tag{22}$$

Entropijo skupnega dogodka oziroma skupno entropijo torej izračunamo po enačbi (23).

$$H_{X,Y} = \sum_{i,j} p(x_i, y_j) \log_2 \frac{1}{p(x_i, y_j)} \tag{23}$$

C. Shannon je v svojem delu [3] pokazal, da velja neenačba (24), kjer dosežemo enakost le, če sta dogodka med seboj neodvisna.

$$H_{X,Y} \leq H_X + H_Y \tag{24}$$

Dogodka, ki sta medseboj odvisna, namreč vsebujeta neko skupno povprečno informacijo, ki jo torej pri seštevanju na desni strani neenačbe (24) upoštevamo dvakrat. Negotovost o izidu nekega dogodka se po poznavanju izida njemu odvisnega dogodka zmanjša. Njeno povprečno vrednost izražamo s pogojno entropijo, podano v enačbi (25).

$$H_{X|Y=y_j} = \sum_{i=1}^{I} p(x_i \mid y_j) \log_2 \left(\frac{1}{p(x_i \mid y_j)} \right) \tag{25}$$

Enačba (25) torej izraža povprečno negotovost o izidu dogodka *X*, če vemo, da se je zgodil *j*-ti dogodek $Y = y_j$. Ker pa je ta vrednost odvisna od izida dogodka *Y*, jo je potrebno za povprečno pogojno entropijo preko vseh možnih izidov dogodka *Y*, povprečiti oziroma utežiti z ustrezno porazdelitvijo verjetnosti. Tako dobimo enačbo (26).

$$H_{X|Y} = \sum_{j=1}^{J} \sum_{i=1}^{I} p(y_j) p(x_i \mid y_j) \log_2 \left(\frac{1}{p(x_i \mid y_j)} \right) = \sum_{j=1}^{J} \sum_{i=1}^{I} p(x_i, y_j) \log_2 \left(\frac{1}{p(x_i \mid y_j)} \right) \tag{26}$$

$H_{X|Y}$ je pogojna entropija, ki predstavlja povprečno mero negotovosti o izidu naključne spremenljivke *X*, pri poznanemu izidu naključne spremenljivke *Y*. Kot

smo že povedali, je povprečna negotovost o izidu naključne spremenljivke *X*, če je izid naključne spremenljivke *Y* poznan, lahko manjša ali kvečjemu enaka povprečni negotovosti o izidu naključne spremenljivke *X* pri nepoznanem izidu naključne spremenljivke *Y*. O tem govori enačba (27).

$$H_{X|Y} \leq H_X \tag{27}$$

Natančno povezavo med vsemi vrstami entropij, ki smo jih do sedaj omenili, pa podaja enačba (28).

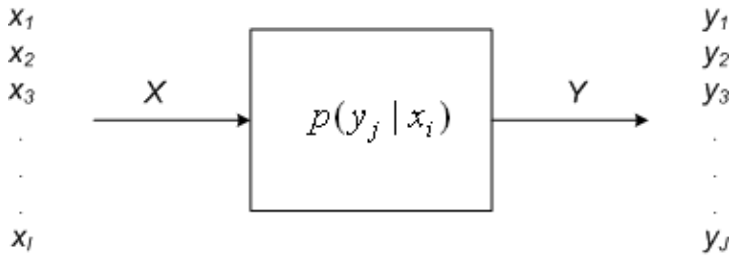
$$H_{X,Y} = H_X + H_{Y|X} \tag{28}$$

Enačba (28) pove, da je negotovost o skupnem dogodku enaka vsoti negotovosti o *X* in negotovosti o *Y*, kjer je *X* poznan. Ilustracija relacij med naštetimi vrstami entropij je podana na sliki 4.

4 Prenos informacij

4.1 Diskretni informacijski kanal

Naključni spremenljivki *X* in *Y* iz prejšnjega poglavja (3.4) naj sedaj predstavljata vhod in izhod diskretnega informacijskega kanala, ki ga uporabimo za prenos informacije. V tem primeru seveda želimo, da sta *X* in *Y* med seboj v čim tesnejši povezavi, torej da je med njima visoka odvisnost. Ker je kanal diskreten, na vhodu sprejema diskretne simbole x_i in na izhodu oddaja ravno tako diskretne simbole y_j. Izhodni simboli so pri tem funkcija vhodnih simbolov in prištetega šuma, ki ga ravno tako obravnavamo kot naključno spremenljivko. Če je vsak izhodni simbol odvisen le od enega pripadajočega vhodnega simbola in šuma, gre za kanal brez spomina. Takšen kanal lahko opišemo z matriko, ki podaja pogojne verjetnosti $p(y_j|x_i)$ vseh kombinacij vhodnih in izhodnih simbolov, torej za vsak *i* in *j*. Pri tem *i* in *j* tečeta od 1 do števila vseh simbolov v abecedi, pri čemer se to število na vhodu in izhodu v splošnem lahko razlikuje. Slika 3 shematično prikazuje diskretni informacijski kanal brez spomina, enačba (29) pa temu kanalu pripadajočo matriko s pogojnimi verjetnostmi, ki določajo s kakšno verjetnostjo se bo zgodila določena pretvorba pri prehodu skozi kanal.



Slika 3: Diskretni informacijski kanal brez spomina [4]

$$\mathbf{P} = \begin{bmatrix} p(y_1 \mid x_1) & p(y_1 \mid x_2) & \cdots & p(y_1 \mid x_I) \\ p(y_2 \mid x_1) & p(y_2 \mid x_2) & \cdots & p(y_2 \mid x_I) \\ \vdots & \vdots & & \vdots \\ p(y_J \mid x_1) & p(y_J \mid x_2) & \cdots & p(y_J \mid x_I) \end{bmatrix} \tag{29}$$

4.2 Vzajemna informacija

Na podlagi do sedaj povedanega se vprašamo, koliko informacije prenesemo preko diskretnega informacijskega kanala, če izvemo, da je na izhodu nastopil določen simbol y_j. Preden smo za y_j izvedeli, je bila povprečna negotovost o simbolu x_i na vhodu enaka H_X. Ko pa smo za y_j izvedeli, se je ta negotovost po neenačbi (27) zmanjšala na $H_{X|Y}$. Očitno smo določeno mero negotovosti po prenosu zmanjšali. Količini, za katero smo prvotno entropijo H_X zmanjšali na pogojno $H_{X|Y}$, pravimo vzajemna ali prenesena informacija. Izračunamo jo kot njuno razliko po enačbi (30). Pri tem seveda želimo, da smo negotovost o simbolu x_i pri prenosu čim bolj zmanjšali in torej želimo, da gre vrednost $I_{X;Y}$ proti H_X in da gre $H_{X|Y}$ proti 0.

$$I_{X;Y} = H_X - H_{X|Y} \tag{30}$$

Pristop, kjer bi kot mero prenesene informacije vzeli na primer število izhodnih simbolov, ki se ujemajo z vhodnimi, bi bil seveda napačen. Kot slikovit primer zakaj je temu tako, je C. Shannon izpostavil primer, pri katerem na vhodu oddajamo niz bitov, ki z enako verjetnostjo ½ zavzamejo vrednost 0 ali 1. Če vzamemo, da je šum na kanalu tako močan, da je izhod popolnoma neodvisen od vhoda, vrednosti 0 in 1 pa še vedno nastopijo z enako verjetnostjo ½, bo v povprečju v polovici primerov vrednost bita na izhodu pravilna. Ker pa je popolnoma neodvisna od vhoda, očitno nismo prenesli nobene informacije.

Naštejmo še nekaj lastnosti prenesene informacije. Prenesena informacija ne more biti negativna in torej velja neenačba (31).

$$I_{X;Y} \geq 0 \tag{31}$$

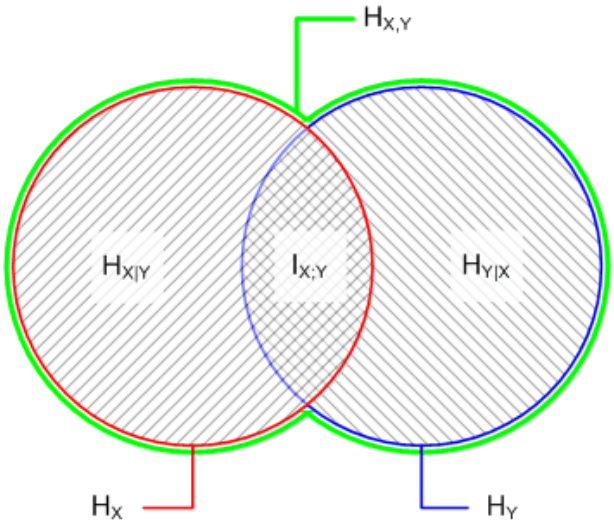
Pokazati je mogoče tudi, da je vzajemna informacija $I_{X;Y}$ enaka $I_{Y;X}$ in jo torej lahko zapišemo po enačbi (32).

$$I_{X;Y} = I_{Y;X} = H_Y - H_{Y|X} \tag{32}$$

Dalje je moč pokazati še veljavnost enačbe (33).

$$I_{X;Y} = H_X + H_Y - H_{X,Y} \tag{33}$$

Vsaka od enačb (30), (32) in (33) predstavlja svojo interpretacijo o tem, kaj vzajemna informacija je. Na sliki 4 je pojem vzajemne informacije in njene relacije do zgoraj opisanih vrst entropij tudi ilustrativno prikazan.



Slika 4: Ilustracija različnih vrst entropij, vzajemne informacije in relacij med njimi [4]

4.3 Kapaciteta kanala

Za izračun vzajemne informacije $I_{X;Y}$ je potrebno v enačbo (30) vstaviti enačbi za H_X (po definiciji entropije, enačba (10)) in $H_{X|Y}$ (enačba (26)). Po preureditvi in upoštevanju enačb (34) in (35), dobimo enačbo (36).

$$p(x_i, y_j) = p(y_j \mid x_i)p(x_i) \tag{34}$$

$$p(y_j) = \sum_{i=1}^{I} p(y_j \mid x_i)p(x_i) \tag{35}$$

$$I_{X;Y} = \sum_{j=1}^{J}\sum_{i=1}^{I} p(x_i)p(y_j \mid x_i)\log_2\left(\frac{p(y_j \mid x_i)}{\sum_{i=1}^{I} p(y_j \mid x_i)p(x_i)}\right) \tag{36}$$

Iz tako izpeljane enačbe (36) je razvidno, da je prenesena informacija odvisna od porazdelitve verjetnosti naključne spremenljivke $p(x_i)$ na vhodu in od pogojnih oziroma prehodnih verjetnosti $p(y_j|x_i)$, ki jih določajo lasnosti kanala. Za učinkovit prenos torej ni dovolj, da ima kanal dobre karakteristike, pač pa je pomembna tudi ustrezna prilagoditev izvora na kanal. Če imamo vnaprej podane lastnosti kanala, na njih ne moremo vplivati. Lahko pa maksimiziramo preneseno informacijo tako, da izmed neskončne množice porazdelitev verjetnosti na vhodu izberemo tisto, ki dá najboljši rezultat. Ko smo pri danem kanalu na ta način

dosegli največjo preneseno informacijo, rečemo, da smo dosegli njegovo kapaciteto. Iz tega sledi definicija po enačbi (37).

$$C = \max_{p(x_i)} I_{X;Y} \tag{37}$$

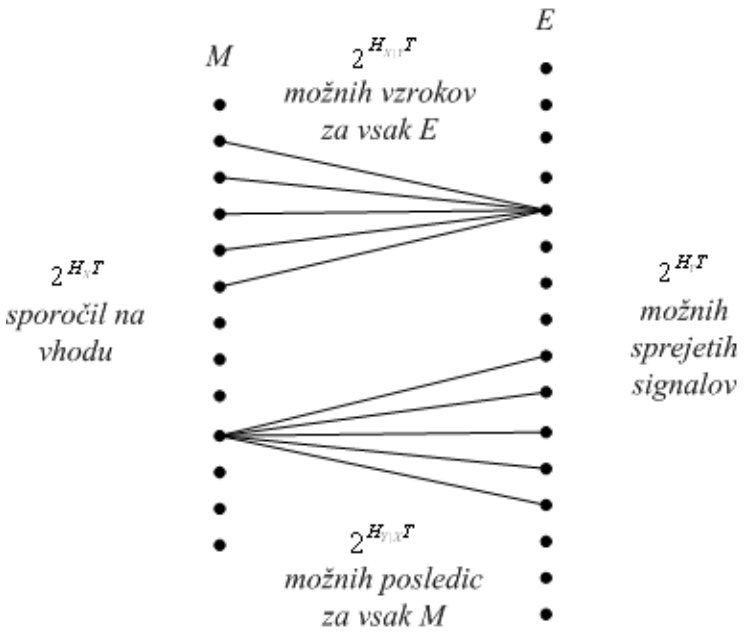
Tako izražena kapaciteta kanala je merjena v bitih na simbol in predstavlja teoretično mejo, ki jo je po tej enačbi v splošnem seveda težko izračunati, saj maksimizacija poteka preko neskončne množice možnih porazdelitev verjetnosti.

4.4 Teorem o kanalskem kodiranju

V prejšnjem poglavju smo definirali, kaj kapaciteta kanala je, sedaj pa bomo opisali kakšne omejitve predstavlja pri kanalskem kodiranju oziroma kakšen pomen ima. Zaradi šuma na kanalu v splošnem pri prenosu pride do napak. Intuitivno bi lahko sklepali, da z dodajanjem redundance zmanjšamo možnost napak oziroma povečamo zmožnost rekonstrukcije originalnega signala. Če bi redundanco povečevali v neskončnost bi s tem napake popolnoma izločili, vendar s tem hkrati tudi izničili preneseno informacijo. Teorem o kanalskem kodiranju, ki ga je postavil C. Shannon, tej intuitivni predpostavki nasprotuje in postavi natančne meje, ki hkrati dajejo tudi eksakten pomen pojmu kapacitete kanala.

Teorem pravi, da v primeru diskretnega kanala s kapaciteto *C*, in diskretnega izvora z entropijo *H*, za katero velja $H \leq C$, obstoji kodirni postopek, s katerim lahko prenašamo izvorni niz preko kanala s poljubno majhno verjetnostjo napake. Drugi del tega teorema pravi, da v primeru entropije diskretnega izvora, ki presega kapaciteto kanala ($H > C$), ni možen prenos s poljubno majhno verjetnosjo napake, pač pa negotovost pri rekostrukciji na sprejemni strani znaša najmanj $H - C$.

Za dokaz zgornjega teorema je Shannon vzel primer, ki je shematično prikazan na sliki 5.



Slika 5: Shematičen prikaz preslikave sporočil med vhodom in izhodom diskretnega informacijskega kanala [3]

Vzemimo, da smo našli izvor, ki po enačbi (37) maksimizira preneseno informacijo na kapaciteto kanala *C*. Na vhodu in izhodu kanala imamo sekvence ali sporočila, ki trajajo nek dolg čas *T*. Če H_X in H_Y predstavljata entropiji na časovno enoto, potem je teh sekvenc na vhodu $2^{H_X T}$, na izhodu pa $2^{H_Y T}$. Za vsako sekvenco na izhodu je lahko na vhodu $2^{H_{X|Y}T}$ vzrokov oziroma sekvenc, ki so to izhodno sekvenco povzročile, saj $H_{X|Y}$ predstavlja povprečno negotovost o tem, kateri simbol je bil izbran na vhodu, če izhodni simbol že poznamo. Podobno velja tudi, da je za vsako vhodno sekvenco lahko $2^{H_{Y|X}T}$ možnih posledic oziroma izhodnih sekvenc. Sedaj pa vzemimo drug izvor, ki oddaja podatke z bitno hitrostjo, za katero velja $R < C$. Tak izvor torej v času *T* lahko odda 2^{RT} različnih sekvenc. Te sekvence želimo izbrati tako, da bo glede na prvotno shemo z izvorom z entropijo H_X frekvenca napak zaradi napačne interpretacije na izhodu čim manjša. Vendar se v prvi fazi ne bomo omejevali na en specifičen primer, pač pa bomo opazovali povprečje vseh možnih sekvenc, ki ustrezajo bitni hitrosti *R*. Verjetnost, da ena od možnih sekvenc pri idealnem izvoru predstavlja sporočilo tudi v izbranem izvoru znaša $2^{T(R-H_X)}$. Če nobena od možnih vhodnih sekvenc v zgornji pahljači na sliki 5, razen tiste, ki dejansko ustreza izhodni sekvenci, ne bi predstavljala sporočila, potem ne bi bilo možne napačne interpretacije na izhodu in torej ne bi bilo napak. Verjetnost, da je temu tako, podaja enačba (38).

$$P = \left[1 - 2^{T(R-H_X)}\right]^{2^{TH_{X|Y}}} \tag{38}$$

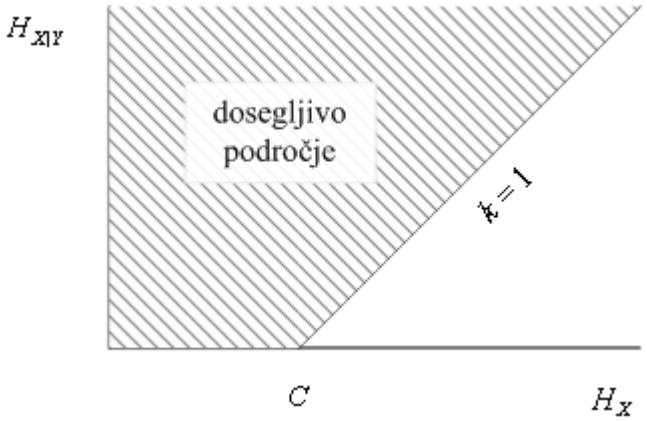
Ker velja $R < C$, mora biti $R + \eta = H_X - H_{X|Y}$. Odtod dobimo $R - H_X = -H_{X|Y} - \eta$, kar vstavimo v enačbo (38) in dobimo enačbo (39).

$$P = \left[1 - 2^{-TH_{X|Y} - T\eta}\right]^{2^{TH_{X|Y}}} \tag{39}$$

Če gre T preko vseh meja ($T \to \infty$), potem gre enačba (39) proti izrazu $1 - 2^{T\eta}$, ta pa gre proti 1. Verjetnost, da bo v pahljači na vhodu samo sporočilo, ki dejansko tudi ustreza izhodni sekvenci, torej zavzame vrednost 1. Kot rečeno smo opazovali povprečje, torej verjetnost za vse možne kombinacije sekvenc. Če to velja za povprečje, zagotovo obstoji vsaj en specifičen primer, kjer je možnost napak še manjša (ali pa je za vse primere enaka).

Dokaz drugega dela teorema je krajši in je posledica definicije kapacitete kanala. Če imamo izvor z entropijo $H_X = C + a$, potem ne moremo doseči $H_{X|Y} = a - \varepsilon$, saj bi to pomenilo, da je $H_X - H_{X|Y} = C + \varepsilon$, kar nasprotuje definiciji kapacitete kanala, ki pravi da je C maksimum razlike $H_X - H_{X|Y}$. $H_{X|Y}$ je torej najmanj a, ta pa je natanko enak razliki $H_X - C$.

Relacijo med kapaciteto kanala C, entropijo izvora H_X in pogojno entropijo oziroma negotovostjo po sprejemu $H_{X|Y}$ nazorno prikazuje tudi slika 6.



Slika 6: Dosegljivo področje negotovosti po sprejemu, glede na entropijo izvora pri podani kapaciteti kanala [3]

Poenostavljeno povedano torej teorem o kanalskem kodiranju pravi, da pri prenosu z bitno hitrostjo $R < C$ v teoriji sicer lahko dosežemo popolnoma zanesljiv prenos brez napak, vendar moramo pri tem uporabiti neskončno kompliciranо kodiranje, z neskončno dolgimi kodami. Bolj kot se R približuje C, težje je to doseči. Če R preseže C, bo do napak nujno prišlo, ne glede na tip kodiranja. V praksi ta meja predstavlja dobro merilo oziroma okvir, katerega se morajo načrtovalci pri izbiri kodiranja držati. Ker časa za prenos ni neskončno mnogo in prevelikih zakasnitev ne moremo tolerirati, se uvaja na primer bločne kode, pri katerih blok traja dovolj dolgo. Teorem sam torej ne govori o tem,

kakšno kodiranje je primerno, pač pa samo kje je meja, ki je ne moremo preseči. Ker pa je možnosti za uspešno približanje meji neskončno, bodo raziskave na tem področju zagotovo vedno aktualne.

5 Fizikalna slika prenosa informacij

Do sedaj smo obravnavali osnovne principe informacijske teorije in prenosa informacij za diskretne naključne spremenljivke in jih matematično opisali. V nadaljevanju pa bomo pogledali realnejše sisteme, ki upoštevajo tudi zvezne naključne spremenljivke in do sedaj razložene principe povezali s fizikalnimi količinami kot so čas, pasovna širina, moč signala in moč šuma.

5.1 Entropija in vzajemna informacija pri zveznih naključnih spremenljivkah

Vzemimo zvezno naključno spremenljivko *X*, s podano gostoto verjetnosti $p_X(x)$. Po analogiji z definicijo entropije za diskretne naključne spremenljivke (enačba (10)), za zvezne naključne spremenljivke uvedemo definicijo po enačbi (40), kjer vsota preide v integral.

$$h_X = \int_{-\infty}^{\infty} p_X(x) \log_2 \left(\frac{1}{p_X(x)} \right) dx \tag{40}$$

Količino h_X imenujemo relativna entropija, saj ne predstavlja absolutnega merila povprečne negotovosti, kot je bilo to v primeru diskretnih naključnih spremenljivk. Izkaže se namreč, da je njena vrednost odvisna od izbire koordinatnega sistema in jo torej lahko obravnavamo kot merilo povprečne negotovosti glede na neko referenco. Ne glede na spremenljivo naravo je ta količina pomembna, saj pri obravnavi prenesene informacije in kapacitete računamo razliko dveh entropij z enakima referencama, ki se torej odštejeta in ne vplivata na končni rezultat. Zaradi relativnega značaja je relativna entropija za razliko od absolutne lahko tudi negativna, kar pa v skladu z zgoraj povedanim še vedno ne vpliva na končni rezultat, računana kapaciteta bo kljub temu nenegativna. Nenazadnje pri zveznih naključnih spremenljivkah mere negotovosti ne moremo izražati drugače kot z relativno entropijo, saj bi bila absolutna entropija neskončna. To je namreč posledica dejstva, da je pri zveznih naključnih spremenljivkah možen izbor neskončno različnih vrednosti in je torej povprečna negotovost neskončna.

Naslednja pomembna lastnost relativne entropije, ki nas zanima je, v kakšnih mejah se lahko giblje, kakšna je njena maksimalna in minimalna vrednost pri podanih omejitvah. Izkaže se, da pri izbranem koordinatnem sistemu, srednji vrednosti porazdelitve gostote verjetnosti 0 in končni varianci σ^2 relativna entropija lahko zavzame vrednosti, kot jih podaja neenačba (41).

$$0 \le h_X \le \tfrac{1}{2}\log_2(2\pi e\sigma^2) \tag{41}$$

Maksimalno vrednost, z enakostjo na desni strani zgornje neenačbe, relativna entropija doseže samo v primeru, če je porazdelitev naključne spremenljivke *X* Gaussova. Z drugimi besedami, Gaussova porazdelitev dá največjo relativno entropijo zvezne naključne spremenljivke. To velja tudi v primeru, ko srednja vrednost ni 0. Zaradi te lastnosti Gaussove porazdelitve, pa tudi zaradi podobnosti iz narave, se kot konservativni model kanala pri prenosu uporablja Gaussov kanal, ki ga bomo opisali v nadaljevanju.

Pred uvedbo pojma vzajemne informacije za dve zvezni naključni spremenljivki definirajmo še relativno pogojno entropijo, kot jo podaja enačba (42).

$$h_{X|Y} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{X,Y}(x,y)\log_2\left(\frac{1}{p_{X|Y}(x\,|\,y)}\right)dxdy \tag{42}$$

Pri tem je $p_{X,Y}(x,y)$ gostota skupne verjetnosti naključnih spremenljivk *X* in *Y*, $p_{X|Y}(x\,|\,y)$ pa gostota pogojne verjetnosti naključne spremenljivke *X*, kjer je poznan *Y* = *y*. Vzajemna informacija zveznih naključnih spremenljivk *X* in *Y* je torej po analogiji z enačbama (30) in (36) podana z enačbo (43).

$$I_{X;Y} = h_X - h_{X|Y} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{X,Y}(x,y)\log_2\left(\frac{p_{X|Y}(x\,|\,y)}{p_X(x)}\right)dxdy \tag{43}$$

Kot za diskretne naključne spremenljivke tudi pri zveznih veljata lastnosti vzajemne informacije po enačbi (44) in neenačbi (45).

$$I_{X;Y} = I_{Y;X} \tag{44}$$

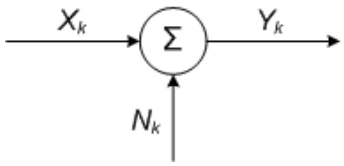
$$I_{X;Y} \ge 0 \tag{45}$$

5.2 Gaussov kanal

Kadar imamo opravka z informacijskim kanalom, na katerem se vhodnim vrednostim prišteje aditiven bel Gaussov šum, govorimo o Gaussovem kanalu. Aditiven bel Gaussov šum (*AWGN – Additive White Gaussian Noise*) ima raven, konstanten spekter in so torej posamezni vzorci tega šuma med seboj neodvisni, saj avtokorelacijska funkcija (kot Fourierov par gostote spektra) zavzame neničelno vrednost le pri ničelnem premiku po abscisni osi. Ker je aditiven, se šum vhodnemu signalu prišteva in ker je Gaussov, za amplitude, ki jih zavzame,

velja Gaussova porazdelitev gostote verjetnosti. Če za vhodno zvezno naključno spremenljivko vzamemo *X* in izhodno *Y*, potem za posamezne vzorce, ki jih označimo z zaporednim indeksom *k* velja enačba (46), kjer *N* predstavlja zvezno naključno spremenljivko aditivnega belega Gaussovega šuma. Gaussov kanal brez spomina torej lahko shematično prikažemo s sliko 7.

$$Y_k = X_k + N_k \tag{46}$$



Slika 7: Gaussov kanal brez spomina

Šum pri prehodu signala skozi kanal vpliva na izhodno vrednost in s tem na gostoto pogojne verjetnosti $p_{Y|X}(y \mid x)$ ter na relativno pogojno entropijo $h_{Y|X}$. Še več, ker predpostavimo, da zvezni naključni spremenljivki *X* in *N* med seboj nista odvisni, velja, da je relativna entropija šuma kar enaka relativni pogojni entropiji pri prehodu skozi kanal in torej velja enačba (47).

$$h_{Y|X} = h_N \tag{47}$$

Za preneseno informacijo preko Gaussovega kanala lahko po do sedaj povedanem zapišemo enačbo (48).

$$I_{X;Y} = h_Y - h_{Y|X} = h_Y - h_N \tag{48}$$

Ker je porazdelitev amplitude šuma Gaussova, je relativna entropija h_N za dano varianco po neenačbi (41) maksimalna in je torej Gaussov kanal s stališča količine prenesene informacije najneugodnejši kanal. Če torej vzamemo najneugodnejši primer kanala in zanj poiščemo kapaciteto, meja ki jo pri tem postavimo ne more biti preveč optimistična, kvečjemu je lahko preveč konservativna. Poleg tega v naravi pogosto nastopa ravno Gaussov kanal oziroma njegov približek. Zato se pri teoretski obravnavi mej prenosa zelo pogosto obravnava ravno Gaussov kanal, za ostale primere pa se Gaussov kanal vzame kot referenco.

5.3 Teorem o informacijski kapaciteti

Najznamenitejši Shannonov teorem povezuje kapaciteto kanala s pasovno širino in močjo oddanega signala ter šuma. Najprej bomo izpeljali kapaciteto kanala v

bitih na simbol, nato pa preko Nyquistovega kriterija dobili še kapaciteto kanala v bitih na sekundo.

5.3.1 Kapaciteta kanala v bitih na simbol

Po definiciji (37) je kapaciteta maksimalna prenesena informacija pri podanem kanalu in izvoru z najugodnejšo porazdelitvijo verjetnosti. Velja enačba (49).

$$C = \max_{p_X(x)} I_{X;Y} = \max_{p_X(x)}(h_Y - h_{Y|X}) = \max_{p_X(x)}(h_Y - h_N) \tag{49}$$

Ker smo predpostavili Gaussov kanal, je relativna entropija h_N maksimalna in z lastnostmi kanala vnaprej podana. Iščemo torej največjo relativno entropijo h_Y. Rekli bi lahko, da iščemo najugodnejši primer prenosa informacije preko najneugodnejšega kanala. Relativna entropija h_Y bo največja, če bo tudi gostota verjetnosti izhodne naključne spremenljivke Y Gaussova. Ta pa bo glede na enačbo (46) Gaussova le, če bo tudi gostota verjetnosti vhodne naključne spremenljivke X Gaussova, saj je vsota Gaussovih porazdelitev zopet Gaussova.

Varianca verjetnostne porazdelitve amplitude signala predstavlja njegovo povprečno moč. Če torej označimo povprečno moč šuma s P_N, bo njegova relativna entropija v skladu z neenačbo (41) podana z enačbo (50).

$$h_N = \tfrac{1}{2}\log_2(2\pi e P_N) \tag{50}$$

Če naprej označimo povprečno moč signala s P_S, bo povprečna moč sprejetega signala enaka vsoti P_N + P_S in bo torej njegova relativna entropija podana z enačbo (51).

$$h_Y = \tfrac{1}{2}\log_2\left[2\pi e(P_N + P_S)\right] \tag{51}$$

Ker smo za enačbo (49) ugotovili, da zavzame največjo vrednost pri vhodni naključni spremenljivki X z Gaussovo porazdelitvijo (vhodni signal ima lastnosti šuma), lahko pri upoštevanju te predpostavke zapišemo enačbo (52).

$$C = I_{X;Y} = h_Y - h_N \qquad , \ kjer\ ima\ X\ Gaussovo\ porazdelitev \tag{52}$$

Po vstavitvi enačb (50) in (51) v enačbo (52) in preureditvi, dobimo končno rešitev za kapaciteto kanala v bitih na simbol, podano z enačbo (53).

$$C = \frac{1}{2}\log_2\left(1 + \frac{P_S}{P_N}\right) \tag{53}$$

Do istega rezultata bi lahko prišli tudi po bolj analitični poti. V tem primeru bi vzeli enačbo za preneseno informacijo (43) in vstavili funkcije Gaussovih porazdelitev verjetnosti po enačbah (54), (55) in (56) ter upoštevali zvezo po enačbi (57).

$$p_X(x) = \frac{1}{\sqrt{2\pi P_S}} e^{-(x^2/2P_S)} \tag{54}$$

$$p_Y(y) = \frac{1}{\sqrt{2\pi (P_S + P_N)}} e^{-(y^2/2(P_S+P_N))} \tag{55}$$

$$p_{X|Y}(x \mid y) = \frac{1}{\sqrt{2\pi P_N}} e^{-((y-x)^2/2P_N)} \tag{56}$$

$$p_{X,Y}(x,y) = p_{X|Y}(x \mid y) p_Y(y) \tag{57}$$

5.3.2 Nyquistov kriterij

Harry Nyquist je k teoretičnim osnovam mej prenosa informacij prispeval pomemben delež. Ugotovil je, da za enolično predstavitev katerega koli časovnega signala, ki je omejen s pasovno širino *B* zadostuje, če ga vzorčimo s frekvenco *2B*. Vzorci so torej med seboj oddaljeni za čas *1/2B* in jih je skupaj *2TB*, pri čemer je *T* čas trajanja signala. Toliko vzorcev lahko popolnoma nadomesti katerokoli funkcijo, ki je frekvenčno omejena s pasovno širino *B* in časovno omejena s časom *T*, ne da bi pri tem prišlo do izgube informacije ali popačitve prvotne funkcije pri pravilni rekonstrukciji. Tej ugotovitvi pravimo tudi vzorčni teorem in ga lahko zapišemo z enačbo (58), kjer z F_S označimo vzorčevalno frekvenco, z *B* pa pasovno širino vzorčenega signala.

$$F_S \geq 2B \tag{58}$$

Če smo zgornjemu pogoju zadostili, signal lahko rekonstruiramo z vsoto sinc funkcij. V primeru, da v enačbi (58) velja enakost, signal rekonstruiramo po enačbi (59), kjer so x_n *n*-te vzorčne točke, za katere velja $x_n = x(n/(2B))$.

$$f(t) = \sum_{n=-\infty}^{\infty} x_n \frac{\sin\left(\pi\left(2Bt - n\right)\right)}{\pi\left(2Bt - n\right)} \tag{59}$$

Na problem števila vzorcev v časovni enoti smo sedaj gledali s stališča izvora. Sedaj pa na taisti problem poglejmo še s stališča prenosa preko kanala oziroma predvsem s stališča dekodiranja na sprejemni strani. Zopet vzemimo kanal, ki je omejen s pasovno širino *B*. Če prenašamo čez ta kanal impulze, s ploščino

enako vrednosti posameznega vzorca, se pri prenosu impulzi preoblikujejo, kot to določa sistemska funkcija prenosnega sistema. Na sprejemni strani nato ob vnaprej določenih trenutkih odčitamo vrednosti, ki ustrezajo oblikovanim vrednostim vhodnih impulzov. Pri tem predstavlja problem predvsem razteg signala po časovni osi, saj je lahko odziv na vhodni impulz neničeln ne samo ob ustreznem vzorčnem trenutku na sprejemni strani, ampak tudi ob vzorčnih trenutkih, ki sledijo. Temu pojavu pravimo intersimbolna interferenca (ISI). ISI lahko povzroča napake pri sprejemu. Da bi se ji izognili, mora odziv na vhodni impulz ob vseh vzorčnih trenutkih zavzeti vrednost 0, razen ob ustreznem časovnem trenutku, kjer mora ustrezati neki vrednosti po kodnem sistemu. Tej zahtevi ustreza sistemska funkcija sinc. Če bi vzorčili ob trenutkih, razmaknjenih za čas T_S, bi morala biti ustrezna sinc funkcija oblike, kakršno podaja enačba (60), kjer je A glede na kodno shemo ustrezno izbrana konstanta.

$$h(t) = A\frac{\sin(\pi / T_S)}{\pi / T_S} \tag{60}$$

Če bi $h(t)$ nato vzorčili ob trenutkih, razmaknjenih za T_S in nad dobljenim rezultatom izvedli Fourierovo transformacijo, bi v frekvenčnem prostoru dobili konstantno vrednost A. Po pravilih vzorčenja pa je ta vrednost dobljena s seštevanjem za ω_S premaknjenih in z $1/T_S$ pomnoženih transformirank nevzorčenega signala, kjer je $\omega_S = 2\pi/T_S$. Če torej želimo prenos brez intersimbolne interference, mora biti tako dobljena prekrita prevajalna funkcija v frekvenčnem prostoru konstantna. Prenosna funkcija z mejno frekvenco $\omega_S/2$, je torej frekvenčno najožja funkcija, pri kateri še dosežemo prenos brez ISI. Če bi namreč vzeli še ožjo frekvenčno mejo $\omega_S'/2 < \omega_S/2$, bi v pasu od $\omega_S'/2$ do $\omega_S/2 + (\omega_S/2 - \omega_S'/2)$ dobili vrednost 0 in torej celotna prekrita prevajalna funkcija ne bi več mogla biti konstantna. Od tu lahko torej postavimo pravilo po enačbi (61), kjer B pomeni pasovno širino kanala, F_S pa frekvenco vzorčenja na sprejemni strani.

$$2\pi B \geq 2\pi / 2T_S = 2\pi F_S / 2 \quad \Rightarrow \quad 2B \geq F_S \tag{61}$$

Pokazali smo torej, da moramo za verodostojno predstavitev signala s pasovno širino B, vzorčiti signal s frekvenco najmanj $2B$, medtem ko pri prenosu preko kanala s pasovno širino B lahko prenašamo največ s simbolno hitrostjo $2B$. Če pogledamo neenačbi (58) in (61) ter privzamemo, da sta pasovni širini signala in kanala enaki, kar je tudi najbolj idealen primer, mora veljati enačba (62), ki z enakostjo definira mejo simbolne hitrosti.

$$F_S = 2B \tag{62}$$

Tako dobljeno mejo lahko torej argumentiramo na dva načina. Pri obeh privzamemo prenosni kanal pasovne širine B. Prvi način pravi, da preko

takšnega kanala lahko prenašamo tudi signal največ taiste pasovne širine *B,* ki je dalje lahko predstavljen z neodvisnimi vrednostmi, podajanimi s hitrostjo najmanj *2B*. Če bi bila hitrost manjša, bi izgubili informacijo, če bi bila večja, pa bi imeli opravka z redundanco. Optimum je torej natanko *2B*. Drugi način pa pravi, da pri prenosu s simbolno hitrostjo, ki je večja od 2*B*, pride do intersimbolne interference in torej ni več možno popolnoma verodostojno dekodiranje na sprejemni strani. Pri manjši hitrosti to sicer je možno, vendar s tem nismo izkoristili kanala. Optimum je torej zopet *2B* in rezultat upoštevamo pri nadaljnji obravnavi.

5.3.3 Kapaciteta kanala v bitih na sekundo

Definirali smo že, koliko bitov informacije lahko preko Gaussovega kanala, glede na jakost šuma in signala, prenese posamezni simbol in koliko simbolov lahko prenesemo v časovni enoti glede na omejitev kanala s pasovno širino. Preostane nam le še, da obe ugotovitvi združimo v eno, ki podaja mejo količine informacije, ki jo lahko prenesemo v časovni enoti glede na naštete omejitve. Enačbo (53), ki podaja količino v bitih na simbol je potrebno pomnožiti z desno stranjo enačbe (62), ki podaja količino v simbolih na sekundo. Dobimo rezultat, ki ga podaja enačba (63) v bitih na sekundo.

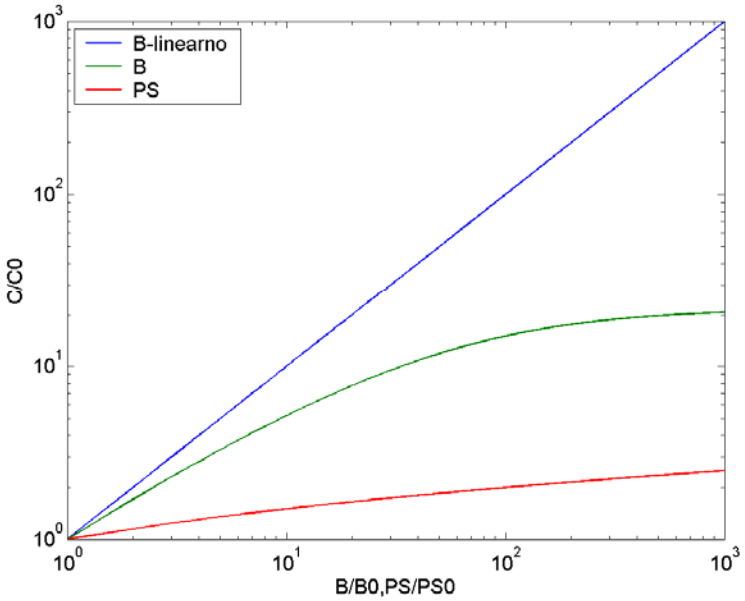
$$C = B\log_2\left(1+\frac{P_S}{P_N}\right) \tag{63}$$

Gre za najslavnejšo Shannonovo enačbo, ki jo je Shannon uporabil v teoremu, v tej nalogi imenovanemu teorem o informacijski kapaciteti. Teorem pravi, da je preko kanala z močjo belega Gaussovega šuma P_N, omejenega s pasovno širino *B,* mogoče prenašati največ z bitno hitrostjo *C*, kjer je P_S povprečna moč oddanega signala. Pri dovolj učinkovitem kodiranju lahko s to hitrostjo prenašamo pri poljubno majhni pogostosti napak. Če je bitna hitrost večja, bo pogostost napak nujno večja od 0, ne glede na izbrani način kodiranja. Teorem torej postavlja temeljno mejo, ki je ni moč preseči in ki se ji lahko približamo le, če se statistične lastnosti oddanega signala približujejo statističnim lastnostim belega Gaussovega šuma.

Lepa lastnost enačbe (63) je, da jasno povezuje vse tri fizikalne količine, ki vplivajo učinkovitost komunikacijskega sistema. Razvidno je, da lahko prenosno hitrost povečujemo bodisi s povečevanjem pasovne širine, bodisi s povečevanjem moči oddanega signala. Na prvi pogled se pri povečevanju pasovne širine kapaciteta povečuje linearno, pri povečevanju moči signala pa logaritmično. Temu žal ni povsem tako, saj pri večji uporabljeni pasovni širini zaobjamemo tudi več šuma. To upoštevamo v enačbi (63) tako, da namesto šumne moči zapišemo produkt pasovne širine kanala *B* in gostote šumne moči p_{N0}. Tako dobimo enačbo (64), ki bolj korektno opisuje povezavo med pasovno širino in kapaciteto kanala.

$$C = B\log_2\left(1+\frac{P_S}{B\cdot p_{N0}}\right) \tag{64}$$

Posledica tega je, da se s povečevanjem pasovne širine kapaciteta kanala ne povečuje linearno, kot to nakazuje modra krivulja v grafu na sliki 8, ampak nekoliko počasneje, kot to prikazuje zelena krivulja. Kljub temu s povečevanjem pasovne širine kapaciteta kanala narašča hitreje kot s povečevanjem moči signala (rdeča krivulja). S tega stališča se v splošnem za povečanje kapacitete kanala bolj izplača povečati pasovno širino kanala kot jakost signala.



Slika 8: Vpliv povečevanja pasovne širine *B* oziroma moči signala P_S na povečevanje kapacitete kanala

Poudariti velja še, da sta obe osi grafa na sliki 8 risani v logaritmičnem merilu in so torej končne razlike v povečanju kapacitete pri tisočkratnem povečanju bodisi pasovne širine bodisi moči signala res velike. Oznake *C0*, *B0* in *PS0* pri tem pomenijo izbrane referenčne vrednosti, uporabljene za normalizacijo. V spodnjem primeru te vrednosti znašajo *C0* = 20,6 kbit/s, *B0* = 3,1 kHz, *PS0* = 1 W. Vrednost p_{N0} iz njih izhaja in je ves čas konstantna. Oznake v legendi grafa (levi zgornji kot) pomenijo katero količino smo spreminjali. Če smo spreminjali *B*, je *PS* držal konstantno vrednost *PS0* in obratno.

5.4 Kapaciteta kanala z nebelim šumom

Do sedaj smo obravnavali kapaciteto kanala v najslabšem primeru, to je primeru, ko je šum bel in ima Gaussovo porazdelitev. V nadaljevanju pa poglejmo še, splošnejši primer, ko se na kanalu signalu prišteva šum, katerega spektralna gostota ni konstanta in torej ni bel. Za začetek predpostavimo, da ima še vedno Gaussovo porazdelitev. V tem primeru lahko izhajamo iz osnovne enačbe (63), pri čemer moramo upoštevati, da le-ta velja le za dovolj majhne frekvenčne odseke. Če so namreč frekvenčni odseki dovolj majhni, gostoti moči signala in šuma zavzameta vrednosti znotraj dovolj majhnih meja in ju v približku lahko obravnavamo kot konstantni. Nato posamezne prispevke k skupni kapaciteti med seboj seštejemo. Od tod dobimo enačbo (65), kjer *k* pomeni zaporeden frekvenčni odsek *Δf*, ki jih je skupno *N* = *B/Δf*.

$$C \cong \sum_{k=1}^{N} C_k = \frac{1}{2} \sum_{k=1}^{N} \Delta f \cdot \log_2 \left(1 + \frac{P_{S,k}}{P_{N,k}} \right) \tag{65}$$

Če se v limiti približamo infinitezimalno majhnim frekvenčnim odsekom, vsota preide v integral po enačbi (66), kjer sta $p_S(f)$ in $p_N(f)$ spektralni gostoti moči signala in šuma.

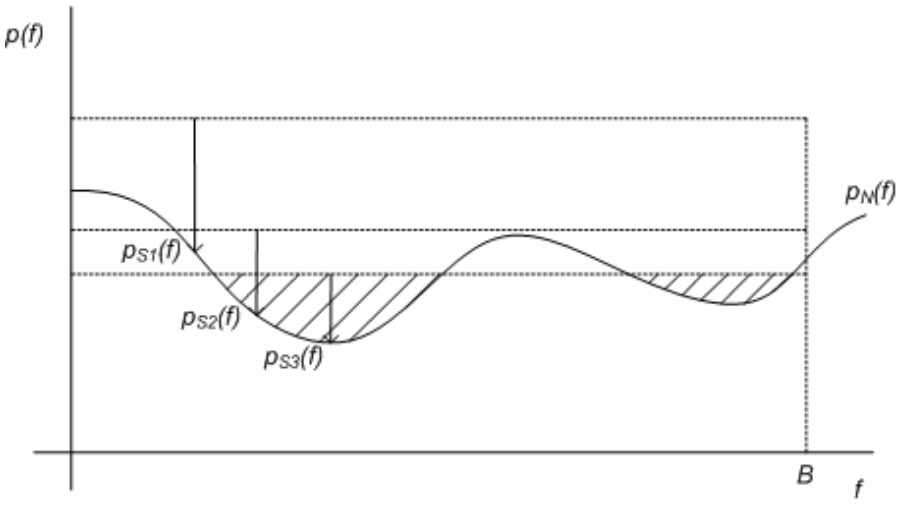
$$C = \frac{1}{2} \int_0^B \log_2 \left(1 + \frac{p_S(f)}{p_N(f)} \right) df \tag{66}$$

Enačba (66) sicer za razliko od enačbe (65) natančno določa kapaciteto kanala, vendar sama po sebi še ne zadošča. Po analogiji iskanja kapacitete v poglavju 4.3 moramo podati tudi najustreznejšo porazdelitev gostote moči izvora, ki ob danih pogojih preneseno informacijo maksimizira. Dani pogoji so v tem primeru funkcija gostote moči šuma $p_N(f)$ in pasovna širina *B* ter skupna povprečna oddajna moč P_S, za katero velja enačba (67).

$$P_S = \int_0^B p_S(f) df \tag{67}$$

Drugače povedano, omejeno oddajno moč moramo glede na dani šum porazdeliti po dani pasovni širini *B* na najboljši način tako, da bo skupna prenesena informacija največja in torej enaka kapaciteti kanala. Izpeljava najugodnejše rešitve ni matematično trivialna in presega namen seminarske naloge, zato podajamo le končno rešitev, ki pa je zelo preprosta. Vsota gostote moči signala in gostote moči šuma mora biti preko celotnega spektra *B* konstantna. To pomeni, da je za maksimalen izkoristek kanala potrebno na frekvenčnih odsekih, kjer je šum šibkejši, oddajati signale z večjo jakostjo in

obratno. Grafično je to prikazano na sliki 9. Tri vzporedne črtkane črte pomenijo izbrano konstantno vsoto. Puščica od črte, ki nakazuje izbrano vsoto, do krivulje gostote šumne moči pomeni gostoto moči signala pri dani frekvenci. Skupna ploščina, ki jo oklepata šumna krivulja spodaj in izbrana črtkana črta zgoraj, predstavlja skupno oddajno moč, ki je konstantna in vnaprej podana. Opazimo lahko, da v primeru, ko ni na voljo dovolj skupne oddajne moči, vsota gostote šumne moči in gostote moči signala ne more biti ves čas konstantna, saj bi to pomenilo na določenih mestih negativno gostoto oddajne moči. V tem primeru velja, da na mestih, kjer je gostota šumne moči večja od mejne vrednosti, ki ravno še zadošča pogoju (67), vsota gostote šumne moči in gostote moči signala ni več konstantna pač pa enaka gostoti šumne moči. Na teh mestih je gostota moči signala enaka nič in posledično ti frekvenčni pasovi niso izkoriščeni. Na sliki 9 sta takšna primera pri gostotah moči $p_{S2}(f)$ in $p_{S3}(f)$.



Slika 9: Najboljša porazdelitev gostote moči signala pri nebelem šumu [9]

Slika 9 šrafirano prikazuje tudi ploščino, ki predstavlja skupno oddajno moč v primeru signala *S3*. Vidimo lahko, da označena ploskev ustreza obliki, ki bi jo zavzela voda, če bi jo natočili v relief takšnega prereza. Zato v angleški literaturi takšni interpretaciji pravijo *»water-filling interpretation«*. To kar mora veljati za gostoto moči signala, da obvelja enačba (66) lahko po zgornjem besednem opisu sedaj zapišemo še z enačbo (68). Pri tem je K konstantna vrednost, izbrana tako, da obvelja pogoj (67), kjer je P_S podan vnaprej.

$$p_S(f) = \begin{cases} K - p_N(f) & , \check{c}e\ p_N(f) < K \\ 0 & , sicer \end{cases} \tag{68}$$

Če bi sedaj upoštevali pravilo po enačbi (68) in opazovali kako na skupno kapaciteto kanala vplivajo različne porazdelitve gostote šumne moči, bi dobili rezultat, da lahko pri neki konstantni oddajni moči, ki jo po frekvenčnem področju

optimalno razporedimo, preko kanala prenesemo najmanj informacije v primeru belega šuma. Dokaz te trditve lahko najdemo v Shannonovih delih, tu pa navajamo le še meje, znotraj katerih se kapaciteta kanala giblje, če šum ni bel. Za poljuben tip šuma je kapaciteta kanala omejena z zgornjo in spodnjo mejo, kot ju navaja neenačba (69). Pri tem pomeni P_N moč šuma, P_{N1} pa entropijsko moč šuma. Entropijska moč je definirana kot količina, pri kateri je vpliv poljubnega šuma na preneseno informacijo tako velik, kot bi bil v primeru belega šuma z isto dejansko močjo. Splošen tip šuma ima dejansko moč večjo od entropijske moči, v primeru belega šuma pa se obe moči izenačita.

$$B\log_2\left(\frac{P_S+P_{N1}}{P_{N1}}\right)\le C\le B\log_2\left(\frac{P_S+P_N}{P_{N1}}\right) \tag{69}$$

Neenačba (69) torej podaja splošnejšo definicijo kapacitete kanala in se reducira na slavno enačbo (63) le v primeru, ko je šum bel. Takrat velja $P_N = P_{N1}$ in se torej obe meji izenačita. Ker je običajno moč signala mnogo večja od šumne moči, pa sta za splošen tip šuma obe meji zelo blizu skupaj. Pokazati je možno tudi, da se kapaciteta C pri naraščanju P_S v limiti približuje zgornji meji. Dalje je moč pokazati še, da bel šum ni najslabši primer le v primeru Gaussove porazdelitve, ampak je v splošnem najslabši primer pri dani skupni šumni moči in pasovni širini, ne glede na verjetnostno porazdelitev.

5.5 Geometrična predstavitev teorema o informacijski kapaciteti

Enačbo (63), ki postavlja temeljne meje prenosa informacij smo v prejšnjih poglavjih izpeljali po analitični poti. Do enakega rezultata pa je možno priti tudi preko geometrične interpretacije komunikacij, ki je verjetno tudi bolj intuitivna. Shannon jo je opisal v članku [9] in kot prednost navedel možnost operiranja z geometrijskimi orodji, področje katerih je bilo takrat seveda bistveno starejše in bolj raziskano. Zaradi zanimive razlage, ki zagotovo prispeva k boljši predstavi o komunikacijskem sistemu, tovrsten pristop k problemu opisujemo tudi v tej nalogi.

V prvi fazi želimo dobiti geometrično predstavitev signalov. V ta namen moramo zopet izhajati iz Nyquistovega teorema o vzorčenju. Ta pravi, da za popolnoma verodostojno predstavitev kateregakoli signala pasovne širine B in trajanja T potrebujemo natanko $2TB$ vzorcev. Na te vzorce lahko gledamo kot na ortogonalne koordinate, ki določajo natanko eno točko v $2TB$ dimenzionalnem prostoru. Vsak signal pasovne širine B in trajanja T torej predstavlja natanko ena točka $2TB$ dimenzionalnega koordinatnega sistema. Po upoštevanju definicije oddaljenosti poljubne točke od koordinatnega izhodišča, ki jo podaja enačba (70) in upoštevanju splošnega zapisa funkcije po enačbi (59), dobimo preko definicije energije signala po enačbi (71), enačbo (72). Ta povezuje razdaljo točke od

koordinatnega izhodišča s pasovno širino signala B, njegovim trajanjem T in povprečno močjo P_S.

$$d = \sqrt{\sum_{n=1}^{2TB} x_n^2} \tag{70}$$

$$E = \int_{-\infty}^{\infty} f(t)^2 \, dt \tag{71}$$

$$d = \sqrt{2TBP_S} \tag{72}$$

Če torej opazujemo le signale, ki imajo povprečno moč manjšo od P_S, le-ti ustrezajo točkam znotraj krogle polmera d, podanega z enačbo (72). Če pa se k izbranemu signalu prišteje šum, se prvotna točka, ki ustreza signalu, v prostoru premakne za razdaljo, sorazmerno z močjo šuma. Ker ne vemo v kateri smeri, šum torej posledično okoli vsake signalne točke ustvari manjši volumen negotovosti o tem, kje se bo signalna točka pojavila po prenosu. Če prihaja na kanalu do linearnega popačenja ali prenosa preko neke sistemske funkcije, se točke v koordinatnem sistemu prestavijo, vendar v fiksnem definiranem smislu, ki ga je za razliko od šumnega vpliva možno popraviti z inverznim postopkom.

Dalje je potrebno v drugi fazi geometrično predstaviti tudi možna sporočila na oddajni strani. To lahko naredimo na podoben način kot smo ga uporabili pri prenosnih signalih. Še posebej načina med seboj sovpadata v primeru prenosa govora. Če prenašamo govor s pasovno širino B_1, ki traja čas T_1, ga lahko predstavimo s točko v $2T_1B_1$ dimenzionalnem prostoru. S tako izbranim koordinatnim sistemom lahko poleg vseh govornih signalov predstavimo tudi vse druge glasovne signale z enako pasovno širino in enakim trajanjem. Poleg tega je človeško uho na fazni potek, če ne gre za stereo zvok, gluho. To pomeni, da bi za tovrsten prenos govora zadostovalo tudi manj dimenzij D. Po analogiji iz prejšnjih poglavij to ustreza zmanjševanju redundance v postopku izvornega kodiranja. Tudi v splošnih primerih izvornih sporočil in ne samo v primeru govora, lahko izvorna sporočila predstavimo kot točko v koordinatnem sistemu z dovolj velikim številom dimenzij.

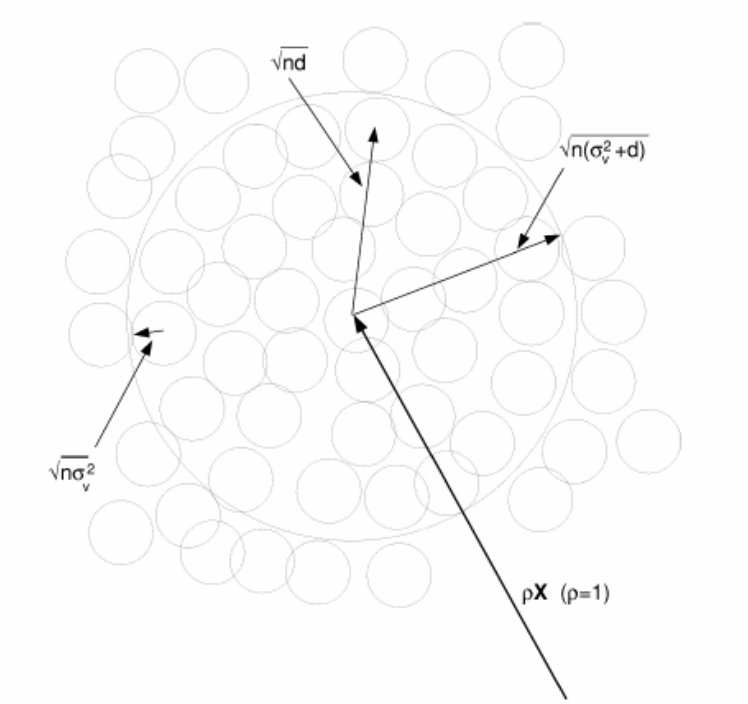
V tretji fazi geometrično predstavimo prenos informacij. Oddajnik ima nalogo da sporočilo pretvori v signal, sprejemnik pa obratno. V geometričnem smislu to pomeni, da je potrebno točko v sporočilnem prostoru dimenzij D na oddajni strani preslikati v točko v signalnem prostoru dimenzij $2TB$ ter obratno na sprejemni strani. Gre torej za preslikavo med koordinatnimi sistemi.

V treh fazah smo torej prikazali povezavo med komunikacijskim sistemom in geometrijo. Nekatere ključne povezave so še enkrat povzete v tabeli 1.

Komunikacijski sistem	Geometrijska entiteta
Skupina možnih prenosnih signalov	*2TB* dimenzionalen prostor
Določen prenosni signal	Točka v tem prostoru
Popačenje med prenosom	Preoblikovanje prostora
Šum	Prostor negotovosti okoli vsake točke v prostoru
Skupina signalov moči, manjše od P_S	Skupina točk v krogli z radijem $(2TBP_S)^{1/2}$
Skupina vseh sporočil	$2T_1B_1$ dimenzionalen prostor
Skupina sporočil, ki jih želimo prenašati	$D < 2T_1B_1$ dimenzionalen prostor
Sporočilo	Točka v tem prostoru
Oddajnik	Preslikava sporočilnega v signalni prostor
Sprejemnik	Preslikava signalnega v sporočilni prostor

Tabela 1: Povezava med komunikacijskim sistemom in geometrijskimi entitetami

V nadaljevanju geometrijske razlage želimo oceniti mejo prenosa informacij, kot jo določajo zgoraj opisane geometrijske entitete in jo povezati z močjo signala, močjo šuma ter pasovno širino kanala. Izpostavili smo že, da se pri prištevanju šuma signalu, originalna signalna točka premakne v neznani smeri za razdaljo, proporcionalno moči šuma. Posledica nepoznane smeri je, da je verjetnost nahajanja nove točke na določenih koordinatah odvisna samo od razdalje od prvotne točke in ne od smeri. To pa pomeni, da je oblika volumna negotovosti, ki se oblikuje okoli originalne točke, okrogla. Za majhne vrednosti produkta *2TB* meje te krogle niso natanko definirane, saj šum lahko zavzame precej različne povprečne amplitude. Ko *T* narašča, pa se povprečna šumna moč približuje neki definirani vrednosti P_N. Pri zelo velikem *T*, bo torej originalna signalna točka z zelo veliko verjetnostjo premaknjena natanko na površino krogle s središčem v originalni signalni točki in radijem $\sqrt{2TBP_N}$. Po drugi strani je povprečna moč sprejetih signalov enaka $P_S + P_N$ in torej po analogiji s prejšnjim primerom pri velikem *T* ležijo sprejete signalne točke na površini krogle z radijem $\sqrt{2TB(P_S + P_N)}$. Na tem mestu si postavimo vprašanje: koliko različnih signalnih točk lahko najdemo, tako da jih bomo po prištetju šuma še vedno lahko med seboj razločili? Odgovor se glasi, da nikakor ne več kot toliko, kolikokrat je volumen krogle z radijem $\sqrt{2TB(P_S + P_N)}$ večji od volumna krogle z radijem $\sqrt{2TBP_N}$. V tem skrajnem primeru smo vse šumne krogle ravno zapakirali v večjo kroglo, ki predstavlja sprejeti signal, pri čemer nismo pustili neizkoriščenega prostora. Če bi bila samo ena več, bi se manjše krogle nujno prekrivale med seboj in razlikovanje vsaj dveh signalnih točk ne bi bilo več možno. Pri izbiranju kodno-modulacijskega sistema gre torej za t.i. problem pakiranja krogel, v angleški literaturi imenovanega »*sphere-packing problem*«, ki je ilustrativno prikazan na sliki 10.



Slika 10: Problem pakiranja krogel (*sphere-packing problem*) [10]

Volumen *n*-dimenzionalne krogle je definiran z enačbo (73).

$$V = \frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2}+1\right)} r^n \tag{73}$$

Če torej v enačbo (73) vstavimo ustrezna izraza za radij večje in manjše krogle ter poiščemo razmerje, dobimo pri dimenziji $n = 2TB$ zgornjo mejo števila signalnih točk, definirano z enačbo (74).

$$M = \left[\sqrt{\frac{P_S + P_N}{P_N}} \right]^{2TB} \tag{74}$$

Če dalje enačbo (74) logaritmiramo in delimo s časom trajanja signala, dobimo rezultat v bitih na sekundo, ki je podan z enačbo (75) in ki se očitno ujema z dobro poznano enačbo (63).

$$C = \frac{\log_2 M}{T} = B \log_2 \frac{P_S + P_N}{P_N} \tag{75}$$

S tem smo dokazali, da ni možno prenašati informacije hitreje kot s hitrostjo *C*, ne da bi pri tem prišlo do napak. Drugega dela teorema o informacijski kapaciteti,

ki pravi, da pod to mejo lahko prenašamo s poljubno majhno napako, preko geometrije tu ne bomo dokazovali. Dokaz je moč poiskati v članku [9].

5.6 Shannonova meja in primeri kodno-modulacijskih postopkov

V prejšnjih poglavjih smo pokazali, da lahko na kapaciteto kanala vplivamo bodisi s povečevanjem pasovne širine, bodisi s povečevanjem moči signala. Če pa želimo primerjati več tako izbranih sistemov med seboj, lahko uvedemo pojem spektralne učinkovitosti *R/B*, kjer *R* pomeni oddajno prenosno hitrost. Spektralna učinkovitost pove, koliko smo izkoristili uporabljeno pasovno širino in je seveda odvisna od razmerja moči signal/šum.

Povprečna moč signala v primeru idealnega prenosa ($R = C$) znaša $P_S = E_b C$, kjer je E_b energija posameznega bita. Če ta izraz vstavimo v izhodiščno enačbo (64), dobimo enačbo (76), ki podaja povezavo med spektralno učinkovitostjo in razmerjem energije bita proti gostoti šumne moči.

$$\frac{C}{B} = \log_2\left(1 + \frac{E_b C}{p_{N0} B}\right) \tag{76}$$

Obratno povezavo pa podaja enačba (77).

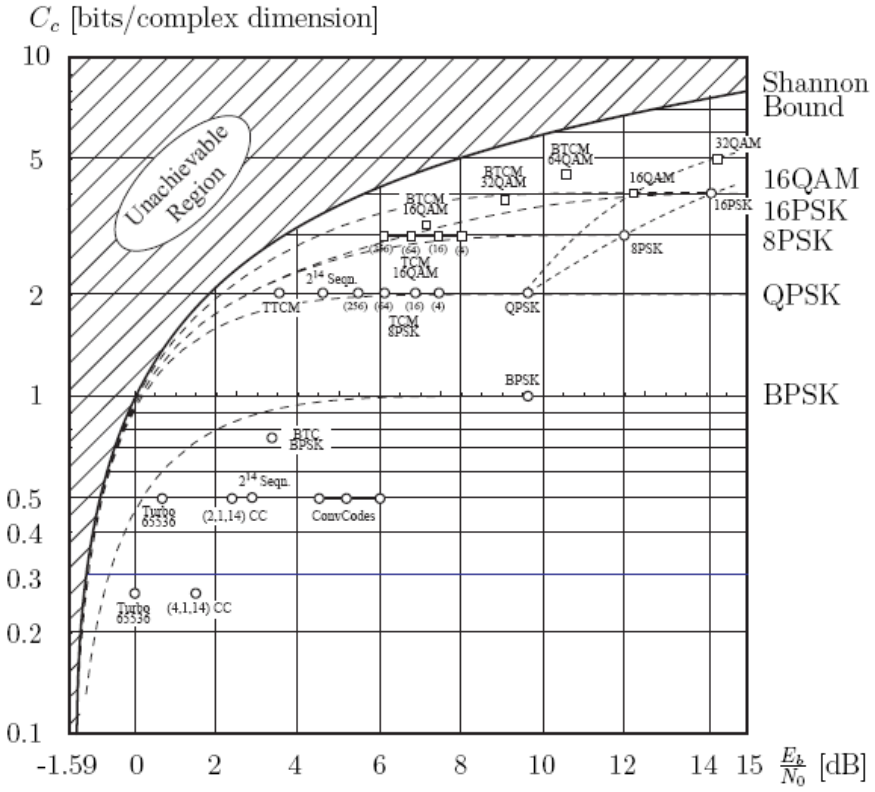
$$\frac{E_b}{p_{N0}} = \frac{2^{C/B} - 1}{C/B} \tag{77}$$

Od tod sledi graf na sliki 11. Odebeljena črta z oznako »*Shannon Bound*« oziroma »Shannonova meja« podaja povezavo, podano z enačbama (76) in (77). Gre torej za črto oziroma mejo, ki jo dosežemo le v primeru idealnega sistema. Nad to črto je nedosegljivo področje, kjer je prenosna hitrost večja od kapacitete in do napak pri prenosu nujno prihaja. Pod črto je prenosna hitrost manjša od kapacitete in lahko torej podatke prenašamo s poljubno majhno verjetnostno napako. V istem grafu so izrisani tudi realni primeri modulacijskih postopkov, ki so seveda pod črto. Bolj kot se posamezen sistem približa meji, boljši je.

Izračunamo lahko še dobro poznano Shannonovo mejo v primeru predpostavke, da imamo na voljo neskončno pasovno širino. Izpeljava sledi iz enačbe (78).

$$\left(\frac{E_b}{p_{N0}}\right)_{\infty} = \lim_{B\to\infty}\left(\frac{2^{C/B} - 1}{C/B}\right) = \ln 2 = 0.693 = -1.59 dB \tag{78}$$

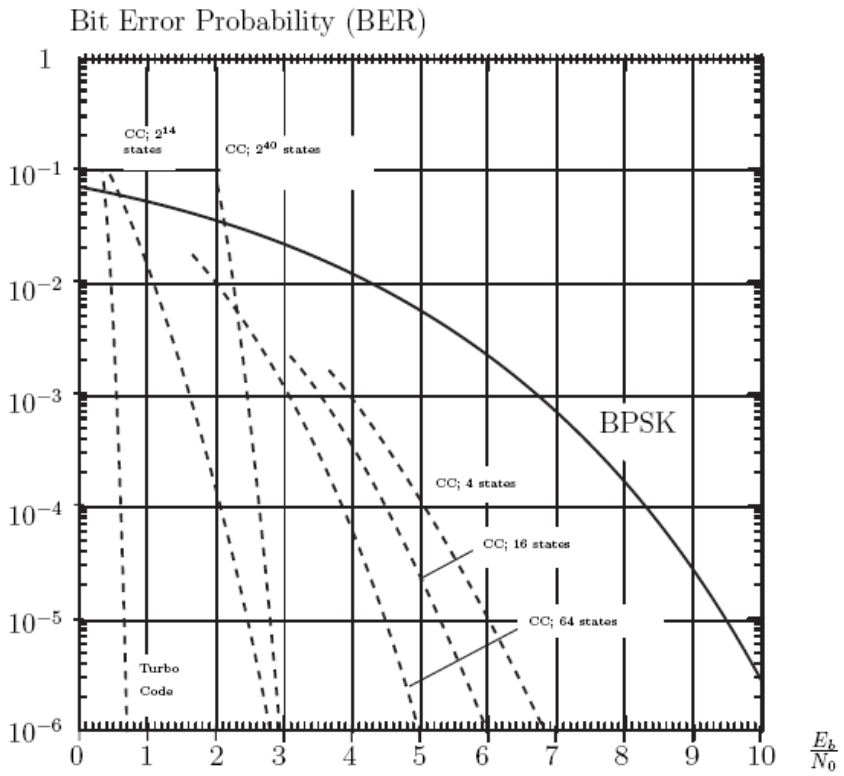
To razmerje energije bita proti gostoti šumne moči je absolutna meja oziroma minimalna vrednost, ki jo moramo doseči, če želimo zanesljiv prenos informacij. Pri manjšem razmerju bo kljub neskončni pasovni širini nujno prihajalo do napak.



Slika 11: Spektralna učinkovitost v odvisnosti od razmerja energije bita proti gostoti šumne moči [11]

Pri realnih sistemih, kjer kodiranje ni neskončno zapleteno in kjer zakasnitev ravno tako ni neskončna, se s približevanjem Shannonovi meji povečuje tudi verjetnost bitnih napak. V grafu na sliki 11 zato na primer pri konstantni spektralni učinkovitosti in pomikanju točke v vodoravni smeri izmenjujemo razmerje E_b/p_{N0} za verjetnost napak, saj ena vrednost pada medtem ko druga raste. Podobno pri pomikanju v navpični smeri držimo konstantno razmerje E_b/p_{N0} in izmenjujemo spektralno učinkovitost za verjetnost bitnih napak.

O prvem primeru, kjer izmenjujemo razmerje E_b/p_{N0} za verjetnost bitnih napak *BER (Bit Error Rate)*, govori tudi graf na sliki 12, kjer so zopet vrisane krivulje za posamezne realne sisteme. Od vrisanih so najučinkovitejše Turbo kode, za katere je značilna strma krivulja, imenovana *Turbo cliff*, ki se Shannonovi meji že zelo približa.



Slika 12: Verjetnost bitne napake v odvisnosti od razmerja energije bita proti gostoti šumne moči [11]

6 Zaključek

Shannon je s svojo informacijsko teorijo postavil temelje, na katerih sloni snovanje komunikacijskih sistemov še danes. Postavil je matematične in fizikalne osnove, brez katerih je dobra predstava o komunikacijskem procesu nemogoča. Začrtal je meje, ki so in bodo ostale za vedno enake. Skratka, njegov dosežek je bil s teoretičnega vidika edinstven. Kljub temu pa Shannonovim delom manjka univerzalen recept, kako doseči postavljene meje. Kako realizirati komunikacijski sistem, ki bi bil v vseh pogledih učinkovit. Zdi se, kot da je inženirje in matematike postavil v neskončen labirint, na koncu katerega čaka lepa in vsem poznana nagrada, ki pa je zaradi kompleksnosti labirinta ni moč doseči. Številnim se je nagradi že uspelo zelo približati, še številnejši pa bi se ji radi še bolj.

Pa pustimo poetično razlago in še enkrat povzemimo tri najpomembnejše teoreme, ki so zaznamovali Shannonovo teorijo in ki smo jih v tej nalogi poimenovali teorem o izvornem kodiranju, teorem o kanalskem kodiranju in teorem o informacijski kapaciteti. Teorem o izvornem kodiranju pravi, da lahko izvorni niz simbolov s pripadajočo entropijo zakodiramo najmanj s povprečnim številom bitov na simbol, ki je enako entropiji. Govori torej o tem, koliko informacije posamezen niz dejansko vsebuje in koliko je posledično potrebujemo za verodostojno predstavitev tega niza. Teorem o kanalskem kodiranju precej preseneča z dejstvom, da za popolnoma zanesljiv prenos ni potrebna neskončna redundanca in s tem ničelni prenos informacije, pač pa zadošča, da je bitna hitrost manjša od kapacitete kanala. Res pa je, da bi v tem primeru morali uporabiti neskončno komplicirano kodiranje z neskončnim trajanjem. Zato se v tem primeru lahko raje odločimo za poljubno majhno verjetnost napake in končno zakasnitev pri prenosu. Če bi bila bitna hitrost večja od kapacitete kanala, pa poljubno majhne napake ne bi bilo več možno doseči. Zadnji teorem, teorem o informacijski kapaciteti, povezuje teorem o kanalskem kodiranju oziroma kapaciteto kanala s fizikalnimi količinami kot so pasovna širina, moč signala in moč šuma. Za verjetnost bitnih napak v povezavi z bitno hitrostjo glede na kapaciteto kanala pravila ostanejo enaka.

Če bi na podlagi zgornjih teoremov in ostalih ugotovitev v seminarski nalogi definirali idealni prenosni sistem, bi zanj ugotovili naslednje:

- informacija, ki jo prenašamo, ne vsebuje redundance,
- hitrost prenosa je enaka kapaciteti kanala,
- frekvenca napak je enaka 0,
- oddani signal ima lastnosti šuma,

- prag, ki določa kdaj bo prišlo do napak je neskončno oster, kar pomeni da bo že pri neskončno majhnem povečanju šumne moči, verjetnost napak bistveno narasla,
- zakasnitve pri prenosu so neskončne.

Z zgornjimi tremi teoremi je torej okvir postavljen, cilj inženirjev pa je, da se mu približajo kolikor se le da. Ker idealnega sistema ni moč doseči (in si ga glede na neskončne zakasnitve niti ne želimo), bodo izboljšave vedno možne in s tem raziskave na tem področju vedno aktualne. Trenutno se Shannonovi meji zelo približujejo Turbo kode, čas pa bo pokazal, katere bodo naslednje, ki bodo še učinkovitejše. Potencial za izboljšave pa je tudi drugod, na primer v raznoliki oddaji in sprejemu, uvajanju kanalov z boljšimi karakteristikami, sofisticirani tehniki vedno višjih frekvenc ter še v mnogih drugih rešitvah. Navsezadnje pa je potrebno pri načrtovanju komunikacijskih sistemov vedno upoštevati tudi njihovo specifiko in iskati optimalno rešitev pri danih razmerah. To pa je tisto, česar Shannon s svojimi deli ni razkril in za kar so zadolženi načrtovalci sodobnih telekomunikacijskih tehnologij.

Literatura

[1] S. W. Golomb, E. Berlekamp, T. M. Cover, R. G. Gallager, J. L. Massey, A. J. Viterbi, Claude Elwood Shannon (1916–2001), Notices of the AMS, vol. 49, no.1, January 2002.

[2] S. Verdu, Fifty Years of Shannon Theory, IEEE Transactions on Information Theory, vol. 44, no. 6, October 1998.

[3] C. E. Shannon, A Mathematical theory of Communication, The Bell System Technical Journal, vol. 27, pp. 379-423, 623-656, July, October, 1948.

[4] S. Haykin, Communication Systems, 4th edition, John Wiley & Sons, Inc., 2001.

[5] E. A. Lee, D. G. Messerschmitt, Digital Communication, 2nd edition, Kluwer Academic Publishers, 1994.

[6] S. Tomažič, Osnove telekomunikacij I, 1. izdaja, Založba FE in FRI, Ljubljana, 2000.

[7] R. G. Gallager, Information Theory and Reliable Communication, John Wiley & Sons, Inc., 1968.

[8] B. P. Lathi, Communication Systems, John Wiley & Sons, Inc., 1968.

[9] C. E. Shannon, Communication in the Presence of Noise, Proceedings of the IEEE, vol. 86, no. 2, pp. 447-457, February, 1998.

[10] R. J. Barron, B. Chen, G. W. Wornell, The Duality Between Information Embedding and Source Coding With Side Information and Some Applications, IEEE Transactions on Information Theory, vol. 49, no. 5, may 2003

[11] www.ee.ualberta.ca/~schlegel/ lecturenotes/ChannelCapacity.pdf