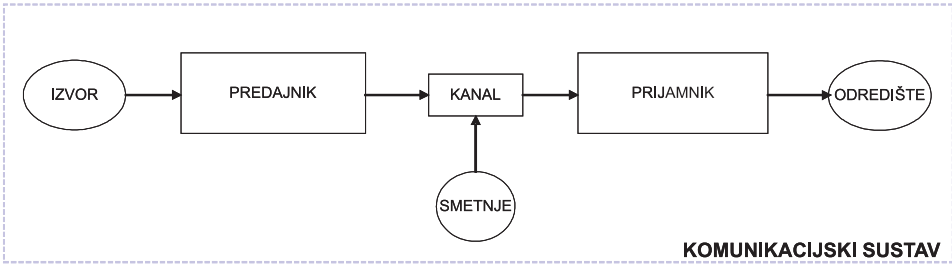
1. Uvod

1.1. Korijeni teorije informacije
1.2. Pregled knjige

Teorija informacije je temeljna matematička teorija koja se bavi problemima komunikacije u smislu prijenosa informacije iz jedne točke (izvor) u drugu (odredište), kao što to prikazuje Slika 1.1. Općenito, prijenos informacije želimo izvršiti:

- što brže,
- što točnije,
- uz što manje utrošene energije,
- usprkos neizbježnim smetnjama u sustavu prijenosa,
- te ponekad uz prikrivanje i zaštitu od zlouporabe.



Slika 1.1: Općeniti shematski prikaz komunikacijskog sustava

Teorija informacije postavlja teoretske osnove za postizanje navedenih ciljeva time što daje definiciju komunikacijskog sustava i mjere za količinu informacije koja protječe kroz taj sustav, te izračunava teoretske granice učinkovitog kodiranja i prijenosa informacije. Konkretnije, jezgreni teoremi daju nam granicu moguće kompresije informacije bez gubitaka, granicu brzine prijenosa informacije kanalom sa ili bez smetnji, te ovisnost moguće brzine prijenosa informacije o odnosu snaga signala i šuma u određenim uvjetima. Teorija informacije se u osnovi bavi prijenosom informacije, no jasno je da kompresija podataka predstavlja važnu osnovu i za pohranu, tj. spremanje podataka.

Premda teorija informacije u svojoj osnovi ne daje praktična rješenja za dostizanje granica koje postavlja, postavljene teoretske osnove su neophodan i ključan korak prema praktičnom rješavanju tih problema. Ova knjiga ne predstavlja samo uvod u jezgrena pitanja teorije informacije, nego i u područja koja na ovim osnovama rješavaju probleme komunikacije na praktične načine, a to su prvenstveno učinkovito kodiranje podataka (kompresija), te zaštitno kodiranje podataka (ispravljanje pogrešaka nastalih zbog smetnji).

Teorija informacije predstavlja neophodnu osnovu modernih informacijskih i komunikacijskih tehnologija. Bez ove bi osnove mnoge današnje tehnologije bile nedostupne, ili daleko manje razvijene, te možemo slobodno reći da je teorija informacije nerazdvojno utkana u život modernog čovjeka i čitavog

društva. Ukoliko ova tvrdnja zvuči pomalo pretjerano, razmotrimo je na sljedeći način. Teorija informacije predstavlja neophodnu osnovu za:

- kodiranje i kompresiju svih vrsta sadržaja, od teksta do govora, glazbe, videa itd.;
- prijenos sadržaja komunikacijskim kanalom u uvjetima smetnji.

Jasno je da su ovo najosnovnije pretpostavke za svaki telefonski razgovor, pristup internetu, prijam satelitske televizije i niz drugih aktivnosti bez kojih bi današnji "moderan život" bio jednostavno nezamisliv. Teorija informacije i tehnologije koje su se razvile na njezinoj osnovi utkane su u društvene promjene čija dalekosežnost i brzina kojom su se dogodile (i kojom se i dalje događaju) vjerojatno nemaju paralelu u povijesti čovječanstva.

1.1. Korijeni teorije informacije

Početkom moderne teorije informacije često se smatra članak Claudea E. Shannona pod naslovom "Matematička teorija komunikacije" iz 1948. g. [45]. Međutim, da bismo shvatili korijene i motivacije koje su dovele do ovog razvoja, potrebno je makar nakratko zaviriti malo dalje u povijest. Teorija informacije vuče korijene iz proučavanja i težnje k poboljšanju elektroničkih komunikacija, konkretno telegrafa, telefona i radiokomunikacija.

Samuel B. Morse je 1838. smislio kôd za prijenos teksta putem telegrafa koji se i danas koristi u nekim vrstama signalizacije. Ovaj poznati kôd predstavlja svako slovo pomoću jedne kodne riječi. Kodna riječ je jednoznačna kombinacija kratkih i dugih signala (strujnih impulsa), što se obično predstavlja točkama (kratki signal) i crtama (dugi signal). Pošto je za slanje dugih signala potrebno više vremena, Morse je uočio da se više isplati dodijeliti kraće kodne riječi onim slovima abecede koja se koriste češće. Tako je, na primjer, najčešće korišteno slovo u engleskom jeziku, slovo E, predstavljeno najkraćom mogućom kodnom riječi, jednom točkom. Kako bi saznao koja slova se najčešće koriste, Morse se snašao na originalan način. Otišao je u tiskaru – tisak se tada radio tako da su se ručno slagala olovna slova jedno do drugog – i prebrojio koliko se olovnih otisaka koristi za svako pojedino slovo abecede. Korištenjem kraćih kodnih riječi za češća slova, Morse je postigao značajnu uštedu (oko 15%) prosječnog vremena potrebnog za slanje poruka. Iz ovog je primjera jasna važnost kodiranja, tj. načina na koji pretvaramo poruke koje želimo poslati u nizove simbola. Ovo je jedno od osnovnih pitanja u teoriji informacije.

Telegrafija je vrlo brzo naišla na ozbiljan problem. Zbog prijenosnih se karakteristika tadašnjih sustava, impulsni strujni signal s predajne strane na prijamnoj strani pretvara u manje ili više razvučeni signal. Dakle, više se

susjednih signala može preklopiti ako vremenski razmaci između njih nisu dovoljno veliki, te primljeni signal postaje nerazumljiv. To znači da postoji granica brzine kojom se poruke mogu slati putem takvog sustava. U slučaju podzemnih ili podmorskih kabela, ta je granica brzine bila toliko niska da je smetala operaterima. Ovaj problem brzine prijenosa informacije putem nekog sustava – komunikacijskog kanala – centralno je pitanje teorije informacije.

Jedan od načina da se ubrza slanje podataka telegrafom bio je korištenje više razina i smjerova struje. Na primjer, Thomas Edison je 1847. uveo tzv. kvadripleksni telegrafski sustav koji je koristio četiri moguća signala: dvije razine jakosti pozitivne struje (+1 i +3) i dvije negativne (–1 i –3). Ovaj sustav je korišten za slanje dviju poruka istovremeno. Jedna poruka se šalje smjerom struje (+ ili –) a druga razinom jakosti struje (1 ili 3). Ako ovo usporedimo sa slučajem korištenja samo dva simbola (npr. pozitivne i negative struje, bez korištenja različitih razina), vidimo da korištenjem dvostruko više simbola možemo prenositi dvostruko više poruka, dakle dvostruko više korisne informacije. Harry Nyquist [46] je 1924. poopćio ovo zapažanje i pokazao da slanjem k simbola u sekundi, gdje svaki simbol može poprimiti jednu od m različitih vrijednosti, možemo postići brzinu prijenosa $W = k \log_2 m$ [bit/s]. Vidimo da je za spomenuti Edisonov sustav m jednak 4, jer imamo na raspolaganju 4 različite vrijednosti simbola.

U vrijeme Nyquistovog rada bila je uobičajena praksa korištenja iste linije za telefonsku i telegrafsku komunikaciju, omogućena činjenicom da ove dvije vrste signala koriste različite frekvencijske pojaseve. Nyquist je proučavao signale razlažući ih na frekvencijske komponente i došao do zaključaka koji su ugrađeni u teoriju informacije i koji imaju univerzalnu primjenu [47]. Najvažniji Nyquistov zaključak je činjenica da je za prijenos i vjernu rekonstrukciju signala ograničenog frekvencijskog pojasa potreban broj uzoraka u sekundi (frekvencija uzorkovanja) dvostruko veći od najveće frekvencije signala (ovo je kasnije dokazao Shannon; postoje neslaganja oko toga kome pripada najveća zasluga za ovaj teorem, pa se on pojavljuje pod raznim imenima: Nyquistov teorem, Shannonov teorem, teorem uzorkovanja itd.).

Vrlo značajan doprinos razvoju teorije informacije dao je i Ralph Hartley [48]. Hartley je razmišljao o problemu komunikacija u smislu izvora koji šalje poruke prijamniku. Na izvoru se za slanje poruka koriste simboli iz konačnog skupa (npr. slova abecede) te se svaka poruka šalje kao niz takvih simbola. Hartley je definirao veličinu H – informaciju sadržanu u poruci – kao $H = n \log s$, gdje je s broj simbola u abecedi, a n broj simbola upotrijebljenih za slanje poruke (dužina poruke).

Iz ovog vrlo sažetog pregleda naslućuje se da je, počevši od prve uporabe telegrafa, tijekom više od jednog stoljeća razvoja elektroničkih komunikacija, nastao velik skup praktičnih i teorijskih znanja o komunikacijama. Claude E.

Shannon [45] je postavio model komunikacije na matematički formalan način, što mu je omogućilo da postavi dalekosežnu teoriju u koju je ugradio postojeća i dodao nova znanja – teoriju koju danas nazivamo teorija informacije. Shannonove postavke omogućile su i inspirirale lavinu budućih teorijskih i praktičnih dostignuća.

1.2. Pregled knjige

Knjiga je podijeljena u sedam poglavlja. Prvo poglavlje pruža kratak uvod u problematiku teorije informacije i njezine korijene, te daje pregled sadržaja knjige.

Drugo poglavlje pruža osnovno razumijevanje temeljnih pojmova teorije informacije i predstavlja temelj na koji se nadograđuju ostala poglavlja. Definira se komunikacijski sustav, poruka, sadržaj informacije, kapacitet kanala, te informacijske mjere uz objašnjenje njihovih značenja. Uvodno se objašnjava pojam kodiranja i uloge kodiranja u komunikacijskom sustavu. Poglavlje završava potpunim prikazom komunikacijskog sustava (Slika 2.18), s koderima informacije i kanala i objašnjenjem protoka informacije kroz sustav.

Ostala poglavlja knjige nadovezuju se na prikaz komunikacijskog sustava na slici (Slika 2.18) te objašnjavaju pitanja vezana uz pojedine dijelove sustava. Organizacija knjige i redoslijed poglavlja odabrani su na osnovu logičnog slijeda od samog komunikacijskog kanala (u sredini slike) prema rubovima komunikacijskog sustava. Shodno ovakvom objašnjenju, slijede redom poglavlje o komunikacijskim kanalima u kontinuiranom vremenu, zatim poglavlje o zaštitnom kodiranju koje se odvija u koderu kanala, te na kraju tri poglavlja vezana uz kodiranje informacije.

Treće poglavlje bavi se pitanjima komunikacijskih kanala u kontinuiranom vremenu. Kao podloga se objašnjavaju obilježja signala koji se pojavljuju na ulazu i izlazu takvih kanala s posebnim naglaskom na njihovu snagu, odnosno energiju jer su to bitni parametri za određivanje kapaciteta kanala. Objašnjavaju se pitanja u vezi uzorkovanja signala i kvantizacije uzoraka. Nadalje, razmatra se modeliranje kanala linearnim vremenski nepromjenjivim sustavom te određivanje širine prijenosnog pojasa kanala. Napokon, definiraju se informacijske mjere u kanalima u kontinuiranom vremenu i određuje izraz za proračun kapaciteta takvih kanala. Konačno, taj se izraz prilagođuje pojasno ograničenim kanalima.

Četvrto se poglavlje bavi zaštitnim kodiranjem. Zaštitno kodiranje odvija se u koderu kanala a svrha mu je omogućiti otkrivanje i/ili ispravljnje pogrešaka nastalih u prijenosu kodiranih poruka komunikacijskim kanalom. U uvodnom

dijelu poglavlja donose se temeljni pojmovi važni za zaštitno kodiranje – Hammingova udaljenost, najveći broj ostvarivih kodnih riječi, perfektni kodovi i ekvivalencija kodova. Nakon toga slijede prikazi linearnih blok kodova, Hammingovih kodova, cikličkih kodova, BCH kodova, konvolucijskih kodova i konačno kratak prikaz turbo kodova.

Posljednja tri poglavlja bave se pitanjima vezanim uz kodiranje informacije. Osnovna uloga kodera informacije je kompresija. Prezentacija je organizirana tako da se krene od osnovnih pojmova kodiranja i kompresije i osnovnih metoda kodiranja, a na toj osnovi se onda grade složeni algoritmi kodiranja pojedinih medijskih sadržaja.

Peto poglavlje, "Entropijsko kodiranje", prvo pruža uvod u osnove kompresije i daje okvirnu podjelu metoda kompresije u entropijsko, izvorno i hibridno kodiranje. Kao temelj za razumijevanje samih metoda kodiranja, uvode se karakteristike izvora informacije, vrste kodova i definicija optimalnog kodiranja. Nakon toga se objašnjavaju odabrane osnovne metode entropijskog kodiranja: Shannon-Fanoovo kodiranje, Huffmanovo kodiranje, aritmetičko kodiranje, metode rječnika (LZ77, LZ78, LZW) i metode skraćivanja niza. Postoje i druge metode entropijskog kodiranja, no ove su metode odabrane kao reprezentativne zbog njihove česte uporabe u praktičnim primjenama kompresije podataka. Izuzetak je Shannon-Fanoovo kodiranje koje se uglavnom ne koristi u praksi, ali predstavlja jedan od koraka u razvoju metoda kodiranja, a uvršteno je u sadržaj jer svojom jednostavnošću olakšava razumijevanje osnovnih principa kodiranja.

Posljednja dva poglavlja nisu dio klasične teorije informacije, ali se na njoj temelje. Šesto se poglavlje bavi metodama izvornog kodiranja. To su metode koje, za razliku od entropijskog kodiranja, ne promatraju više izvor informacije samo kao apstraktni izvor niza simbola opisan isključivo statističkim svojstvima, nego u svrhu bolje kompresije koriste poznavanje svih karakteristika izvora, tj. podataka koji se kodiraju, kao i karakteristike ljudskih osjetila koja su obično krajnji korisnik informacije. Ova svojstva koriste se na način da se uklanja nepotrebno ponavljanje podataka (redundancija), kao i podaci koji su za određenu primjenu nepotrebni u smislu da njihov nedostatak neće bitno smanjiti subjektivnu kvalitetu dekodirane informacije. Jasno je da se ovdje, za razliku od entropijskog kodiranja, radi o metodama kodiranja s gubicima, gdje dekodirana poruka nije potpuno jednaka izvornoj poruci nego predstavlja njezinu aproksimaciju. Pritom je važno da ta aproksimacija bude zadovoljavajuća za određenu primjenu. Na primjer, slika se može kodirati s relativno velikim stupnjem kompresije, uvodeći pritom određenu razliku između izvorne i dekodirane slike, a da se pritom ta razlika ne vidi golim okom. Nakon uvodnog izlaganja o analognim medijima u digitalnom komunikacijskom sustavu, te o principima kompresije pri izvornom kodiranju, uvode se redom sljedeće često korištene metode izvornog kodiranja: kvantizacija, poduzorkovanje, trans-

formacijsko kodiranje, diferencijalno (predikcijsko) kodiranje, potpojasno kodiranje i kodiranje zasnovano na modelu.

Posljednje poglavlje knjige bavi se informacijskim svojstvima i principima kodiranja pojedinih medijskih sadržaja – jezika (teksta), zvuka, nepomične slike i videa. Za kodiranje medija (uz iznimku teksta) koriste se gotovo uvijek hibridni koderi koji koriste kombinacije metoda entropijskog i izvornog kodiranja prikazane u dva prethodna poglavlja. Jasno je da cilj ovog poglavlja nije iscrpan prikaz svih metoda kompresije medijskih sadržaja (za daleko dublji i detaljniji prikaz izvrsna je referenca [49]), nego objašnjenje najosnovnijih principa i vrsta metoda kompresije.

2. Osnovni pojmovi teorije informacije

2.1. Opći model komunikacijskog sustava
2.2. Sadržaj informacije
2.3. Kodiranje
2.4. Informacijski opis komunikacijskog sustava
2.5. Prijenos informacije i informacijske mjere
2.6. Kapacitet kanala
2.7. Prijenos informacija komunikacijskim sustavom
2.8. Zadaci

Ovo poglavlje daje podlogu za razumijevanje temeljnih koncepata i pojmova teorije informacije te pruža temelj na koji se logično nadograđuju ostala poglavlja.

2.1. Opći model komunikacijskog sustava

U svakodnevnom životu, pojam "komunikacija" upotrebljavamo u najrazličitijim mogućim kontekstima, od međuljudskih komunikacija do prometnih veza. Da bismo mogli proučavati probleme komunikacije, potrebno je definirati što taj pojam znači u području koje proučavamo, a to su informacijske i komunikacijske tehnologije. Slika 1.1 prikazuje opći model komunikacijskog sustava koji je postavio Claude E. Shannon 1948. g. [45]. Prema Shannonovoj definiciji, temeljni problem komunikacije je točno ili aproksimativno u jednoj točki informacijskog prostora (odredište) reproducirati poruku odabranu na nekoj drugoj točki (izvor). Kako bi ova definicija bila praktično upotrebljiva, Shannon je precizno definirao značenje svih dijelova općeg komunikacijskog sustava (Slika 1.1), i najvažnije, definirao je što točno znači poruka i njezin prijenos.

Definicija

2.1.1. Diskretni komunikacijski sustav

Mi ćemo se zasad zadovoljiti pomalo intuitivnim shvaćanjem općenitog komunikacijskog sustava, te ćemo najosnovnije pojmove potrebne za proučavanje teorije informacije definirati na nešto jednostavnijem slučaju diskretnog komunikacijskog sustava. U diskretnom komunikacijskom sustavu koristimo se diskretnim signalima (vidi poglavlje 3.1), dakle signalima koji (za razliku od kontinuiranih) mogu poprimiti konačan broj diskretnih vrijednosti. Pritom treba primijetiti da većina modernih komunikacijskih sustava koristi upravo diskretne signale, pa stoga takvo sužavanje područja ne sužava značajno praktičnu vrijednost ovih razmatranja. Kontinuiranim komunikacijskim sustavima ćemo se baviti u poglavlju 3.

Zasad ćemo zanemariti detalje predajnika i prijamnika, te promatrati pojednostavljenu sliku na kojoj je lakše uočiti temeljna pitanja, a to su:

- Što je poruka?
- Što znači prenijeti poruku?
- Kako možemo mjeriti količinu informacije u nekoj poruci, te informacije prenesene sustavom?

2.1.2. Poruka

Definicija

Premda je intuitivno prilično jasno što znači riječ poruka, u smislu teorije informacije je ovaj pojam potrebno preciznije definirati. Poruka je niz simbola odabranih iz abecede, gdje je abeceda X konačan skup simbola:

$$X = \{x_1, x_2, ..., x_i, ..., x_n\}.$$

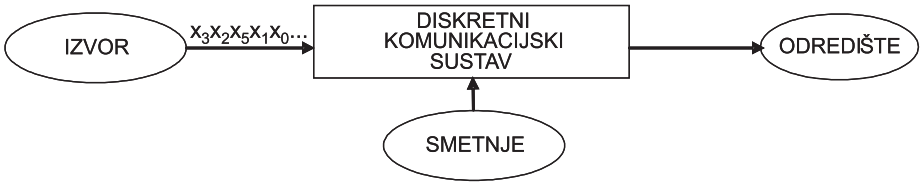
Poruke nastaju na izvoru (Slika 2.1) ponavljanim biranjem simbola iz abecede. Pritom svaki simbol x_i pri N-tom biranju ima neku vjerojatnost $p_N(x_i)$ da bude odabran:

$$x_i \longrightarrow p_N(x_i).$$

Ove vjerojatnosti pojavljivanja simbola određene su karakteristikama izvora.

Takva definicija, premda na prvi pogled jednostavna, pokriva začuđujuće velik i raznovrstan skup mogućih poruka koje koristimo u svakodnevnom životu. Na primjer, ova rečenica je jedna takva poruka, sastavljena od simbola – slova hrvatske abecede. Govor, kada se prenosi digitalnim telefonskim sustavom, u nekim je fazama prijenosa predstavljen kao poruka sastavljena od simbola – uzorkovanih vrijednosti amplitude govornog signala. Isto vrijedi za glazbu koju slušamo s CD-a, sliku koju primamo digitalnim satelitskim prijamnikom itd. U ovim je primjerima očito da vjerojatnost odabira sljedećeg simbola jako ovisi o prethodnim simbolima u poruci. Na primjer, ako su prva četiri simbola tekstualne poruke "jeda", prilična je vjerojatnost da će idući simbol biti "n", a vrlo mala vjerojatnost da idući simbol bude npr. "o". Zasad ćemo promatrati jednostavniji slučaj, gdje je vjerojatnost pojavljivanja svakog simbola neovisna o prethodnim simbolima u poruci i ne mijenja se:

$$x_i \longrightarrow p(x_i).$$



Slika 2.1: Poruka u komunikacijskom sustavu

Klasični primjeri ovog slučaja su generiranje poruke bacanjem novčića ili kocke. Svako bacanje novčića odabir je jednog od dvaju mogućih simbola – pisma ili glave. Svako bacanje kocke odabir je jednog od 6 mogućih simbola. U oba je slučaja rezultat svakog bacanja neovisan o svim prethodnim bacanjima. Vjerojatnosti mogu biti ravnomjerno ili neravnomjerno raspodijeljene (npr. kocka s dva ista broja). Premda ovi primjeri ne djeluju naročito praktično,

izvrsno služe za uvod u teoriju informacije. Krenuvši od njih, postupno ćemo stići do zanimljivih praktičnih primjena kao što je npr. kodiranje i prijenos video slike.

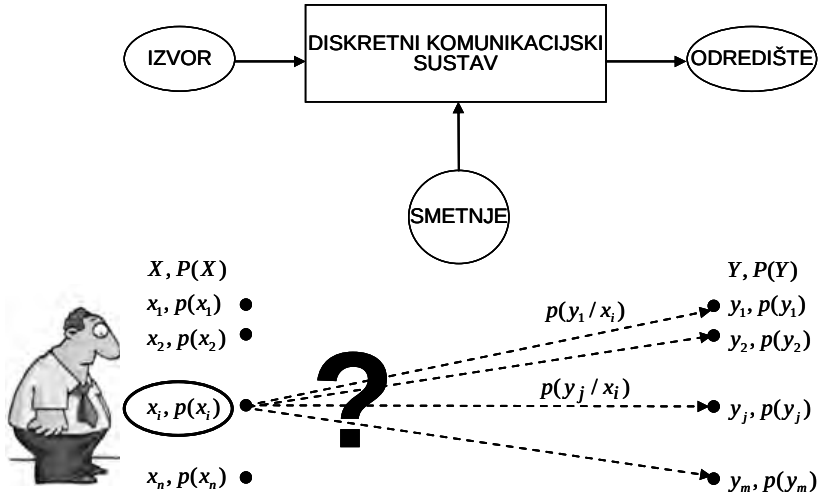
Važno je uočiti da vjerojatnosti pojavljivanja simbola proizlaze iz semantičkih karakteristika izvora koje mogu biti vrlo složene. Na primjer, ako je izvor čovjek koji piše jednu rečenicu, tada vjerojatnosti pojavljivanja pojedinih znakova proizlaze iz složenih faktora koji uključuju jezik i gramatiku. U primjeru govora, odlučujuću ulogu igraju karakteristike ljudskih govornih organa. Teorija informacije se ne bavi semantikom izvora, nego statistički aproksimira izvor vjerojatnostima pojave simbola. To omogućava vrlo općenito razmatranje. Semantiku izvora uzimaju u obzir metode izvornog kodiranja, o kojima će biti više riječi u poglavlju 6.

2.1.3. Prijenos poruke

Pošto smo definirali poruku kao niz simbola, sada možemo na precizniji i formalniji način shvatiti temeljni problem komunikacije – poruku odabranu na izvoru točno ili aproksimativno reproducirati na odredištu. Poruka – niz simbola – odabrana na izvoru kroz mehanizam komunikacije uzrokuje pojavu druge poruke – niza simbola – na odredištu. U idealnom bi slučaju niz simbola koji se pojavi na odredištu bio jednak nizu simbola na izvoru. Međutim, zbog smetnji u komunikaciji to nije slučaj. U općenitom slučaju, na odredištu se pojavljuje neka druga poruka. Iz te druge poruke možemo s određenom sigurnošću zaključiti koja poruka je bila odabrana na izvoru, ali ta sigurnost nije 100%, osim u idealnom slučaju bez smetnji.

Poruku koja se pojavljuje na odredištu moramo dakle promatrati općenito kao niz simbola y_j iz skupa Y, gdje je Y abeceda od m elementarnih simbola koji se mogu pojaviti na odredištu. Radi jednostavnosti, promotrimo prijenos samo jednog simbola iz poruke. Problem komunikacije se time svodi na sljedeće: na izvoru je odabran simbol x_i; djelovanjem mehanizama komunikacijskog sustava na odredištu se pojavio simbol y_j. Na odredištu želimo iz pojave simbola y_j zaključiti koji je simbol x_i odabran na izvoru.

Postavimo se na izvor kao što je predočeno na slici (Slika 2.2). Neka je odabran simbol x_i, te poslan prema odredištu. Na odredištu se može pojaviti bilo koji od simbola $y_1,..., y_m$ i mi kao promatrači na izvoru ne znamo koji se od njih pojavio. Međutim, ako su nam poznata statistička svojstva komunikacijskog sustava, znamo vjerojatnosti pojavljivanja bilo kojeg simbola y_j čiji je uzrok x_i. To su uvjetne vjerojatnosti $p(y_j|x_i)$, dakle vjerojatnosti da se na odredištu pojavi simbol y_j ako je na izvoru odabran simbol x_i.



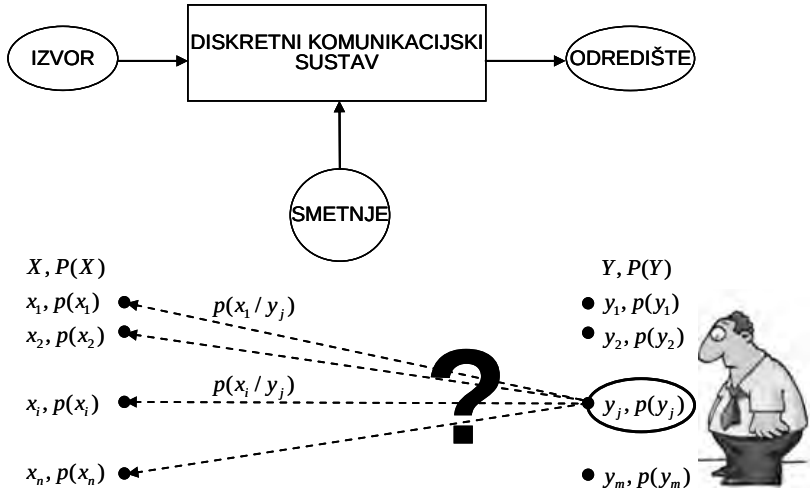
Slika 2.2: Prijenos poruke: pogled s izvora

Vidimo da je u stvarnom slučaju poželjno da je $n = m$ i da su sve vjerojatnosti $p(y_i|x_i)$ što bliže vrijednosti 1, a sve druge vjerojatnosti što bliže vrijednosti 0.

Očito je da su statistička svojstva komunikacijskog sustava određena načinom izvedbe, tj. mehanizmom prijenosa. Takvih mehanizama može biti bezbroj – od signalizacije zastavicama između dva broda do prijenosa signala optičkim nitima. Za bilo koji od tih sustava moguće je odrediti ili procijeniti statistička svojstva prijenosa. Unutar teorije informacije pretpostavit ćemo da su nam ta statistička svojstva poznata i nećemo ulaziti u fizičke mehanizme prijenosa.

Postavimo se sada na odredište (Slika 2.3), i to u trenutku kada je na izvoru odabran simbol x_i, ali malo prije nego što smo na odredištu primili simbol y_j. U ovom trenutku ne znamo još ništa o tome koja je poruka odabrana na izvoru osim vjerojatnosti izbora simbola $p(x_i)$. Nakon prijenosa se na odredištu pojavio simbol y_j. Kao promatrači na odredištu, mi i dalje ne znamo koji je simbol odabran na izvoru. Ukoliko promatramo pojave simbola na odredištu same za sebe, one su neovisne i opisane skupom vjerojatnosti $p(y_j)$. Međutim, poznajemo statistička svojstva sustava, te znamo uvjetne vjerojatnosti $p(x_i|y_j)$, tj. vjerojatnosti da je na izvoru odabran simbol x_i ako znamo da se na odredištu pojavio simbol y_j. Dakle, premda nakon primanja simbola y_j i dalje ne znamo točno koji je simbol bio odabran na izvoru, za svaki od simbola $x_1,..., x_n$ znamo vjerojatnost da je upravo taj simbol bio odabran. Primijetimo da u stvarnom sustavu očekujemo da ta vjerojatnost za jedan od simbola – onaj koji je stvarno bio odabran – bude blizu 1.

Kao promatrači na odredištu, vidimo da se pojavom simbola y_j naše znanje o tome koji je simbol odabran na izvoru bitno povećalo, tj. naša nesigurnost je smanjena – primili smo informaciju.



Slika 2.3: Prijenos poruke: pogled s odredišta

2.2. Sadržaj informacije

Kako bismo mogli promatrati prijenos informacije na sustavan način, potrebna nam je mjera za sadržaj (količinu) informacije u nekoj poruci. Točnije, zanima nas koliko informacije u prosjeku sadrži svaki simbol poruke. Ovu mjeru nazivamo srednji sadržaj informacije a dana je definicijom koja je jedna od temeljnih definicija u teoriji informacije. Prije same definicije, razmotrimo što bi trebala biti mjera za srednji sadržaj informacije. Pošto nas zanima prijenos informacije, onda je to mjera koja izražava koliko informacije možemo prenijeti nekom porukom.

Primjer: Sadržaj informacije pri bacanju novčića

Kako bismo stekli osjećaj za značenje sadržaja informacije, zamislimo jednostavan primjer (Slika 2.4). Na izvoru se stvara poruka bacanjem novčića, i tu poruku – o tome je li palo pismo ili glava – vidi promatrač. Kao promatrač, prije bacanja znamo da može pasti pismo ili glava s jednakom vjerojatnošću, ali ne znamo hoće li pasti pismo ili glava. Nakon bacanja, naša nesigurnost je razriješena i sada znamo točno da je palo npr. pismo – primili smo informaciju. Točnije, primili smo jedan **bit**, jednu osnovnu jedinicu informacije. Vidimo da jedan bit možemo shvatiti kao količinu informacije koja je dovoljna za rješavanje jedne elementarne neodređenosti, jednog izbora između dvije jednako vjerojatne mogućnosti, jednog pitanja na koje odgovor može biti da ili ne. U ovom primjeru, srednji sadržaj informacije je jedan bit po simbolu, tj. svaki simbol (pismo ili glava) nosi jedan bit informacije.

Primjer

Slika 2.4: Sadržaj (količina) informacije: primjer bacanja novčića

Kao protuprimjer, zamislimo da je na novčiću s obje strane pismo, te da kao promatrač unaprijed znamo tu činjenicu. Dakle, znamo da će uvijek pasti pismo i nemamo nikakvu nesigurnost po tom pitanju. Poruke koje opažamo – pismo, pismo, ..., opet pismo – ne donose nam nikakvu novu informaciju. Dakle, sadržaj informacije ovakve poruke je 0 bita.

Zamislimo da imamo novčić koji u prosjeku 70% bacanja daje pismo, a 30% bacanja glavu. Koliki je sada srednji sadržaj informacije koji primamo kao promatrač? Da bismo odgovorili na ovo pitanje, i kasnija složenija pitanja, potrebno je prijeći s ovih intuitivnih razmatranja na definiciju sadržaja informacije.

2.2.1. Entropija

Definicija

Diskretna slučajna varijabla je varijabla odabrana iz skupa od n mogućih vrijednosti, gdje je za svaku vrijednost poznata vjerojatnost odabira. Entropija diskretne slučajne varijable je definirana kao:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i) \ \left[\text{bit/simbol}\right], \tag{2.1}$$

gdje je X diskretna slučajna varijabla koja poprima vrijednosti iz skupa $\{x_1,..., x_i,...x_n\}$, a $p(x_i)$ su vjerojatnosti pojavljivanja vrijednosti x_i, $1 \le i \le n$.

Ako koristimo logaritam s bazom 2, entropiju izražavamo u bitovima. Općenito, definicija entropije vrijedi uz bilo koju odabranu bazu logaritma, samo što se tada ne izražava u bitovima nego u drukčijim jedinicama (npr. nit ili nat – baza e, dit – baza 10); u praksi se gotovo uvijek koristi baza 2 i bit. **U ovoj knjizi, ukoliko nije drukčije naznačeno, uvijek koristimo logaritam baze 2.**

Poruka sastavljena od ovakvih simbola sadrži u prosjeku $H(X)$ bita po simbolu.

Entropija i termodinamika

Nije naodmet napomenuti da je Shannon pojam "entropija" preuzeo iz područja termodinamike, te postoji čvrsta analogija između entropije poznate u termodinamici i entropije u teoriji informacije (međutim, to nipošto ne znači da bi za shvaćanje osnova teorije informacije bilo neophodno poznavanje područja termodinamike). Budući da se radi o matematičkoj, općenitoj definiciji, postoje brojni načini poimanja i interpretacije entropije, od kojih će neki biti prikazani u ovoj knjizi. Za početak je važno da entropiju shvatimo kao mjeru za sadržaj informacije, odnosno za neodređenost izvora informacije.

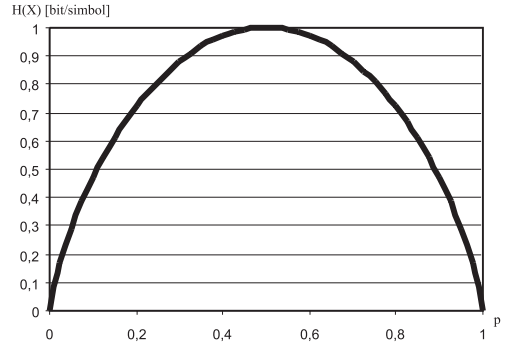
Definicija entropije je izuzetno korisna zbog toga što točno odgovara onome što intuitivno shvaćamo kao sadržaj informacije. To možda nije očito na prvi pogled iz same formule, ali kroz primjere i dodatne diskusije vrlo brzo postaje jasno da entropija posjeduje svojstva koja je čine idealnom mjerom za sadržaj (količinu) informacije.

Primjer: Entropija bacanja novčića

Primjer

Vratimo se nakratko primjeru bacanja novčića (Slika 2.4). U ovom slučaju, imamo dva moguća simbola ($n = 2$), i oba se pojavljuju s vjerojatnošću $p(x_i) = 0{,}5$. Jednostavnim uvrštavanjem u formulu entropije dobivamo $H(X) = 1$ bit/simbol. Dakle, entropija bacanja novčića je 1 bit/simbol, odnosno srednji sadržaj informacije poruke koja se sastoji od uzastopnih rezultata bacanja novčića je 1 bit po simbolu. Za slučaj "nepoštenog" novčića koji uvijek daje pismo, imamo $p(x_1) = 1$, $p(x_2) = 0$, dobivamo očekivano $H(X) = 0$ bit/simbol ($0 \log 0 = 0$, jer vrijedi $x \log x \to 0$ kada $x \to 0$). Za slučaj "nesimetričnog novčića" koji daje 70% pismo, dobivamo $H(X) = 0{,}88$. Dakle, poruka nastala bacanjem ovakvog novčića daje u prosjeku 0,88 bita po simbolu.

Uvrštavanjem svih mogućih vjerojatnosti pojave pisma u formulu entropije, dobivamo graf ovisnosti entropije o toj vjerojatnosti (Slika 2.5).



Slika 2.5: Entropija bacanja novčića *H* u ovisnosti o vjerojatnosti pojavljivanja pisma *p*

Maksimum (1 bit/simbol) je postignut kada je vjerojatnost pisma jednaka vjerojatnosti glave (p = 0,5) – dakle kada je najveća nesigurnost pojave jednog ili drugog. Primijetimo simetriju ovog grafa. Svejedno je pojavljuje li se s većom vjerojatnošću pismo ili glava. Zamjenom njihovih uloga situacija se s informacijskog gledišta ne mijenja.

Primjer: Entropija bacanja novčića – složeniji eksperiment

Primjer

U laboratoriju se izvodi eksperiment bacanja novčića – ako padne pismo, eksperiment se zaustavlja, a ako ne padne pismo, eksperiment se izvodi dalje. Događaji su sljedovi {**g**lava, **p**ismo}. Svaki događaj opisujemo jednim od simbola iz skupa $X=\{x_0, x_1,..., x_i,...x_n\}$, gdje indeks "$i$" označava broj pojava glave prije pojave pisma. Potrebno je odrediti entropiju skupa simbola X.

Rješenje:

događaj i	simbol, x_i	vjerojatnost, $p(x_i)$
p	x_0	0,5
gp	x_1	$0{,}5\cdot 0{,}5=(0{,}5)^2=0{,}25$
ggp	x_2	$0{,}5\cdot 0{,}5\cdot 0{,}5=(0{,}5)^3=0{,}125$
gggp	x_3	$0{,}5\cdot 0{,}5\cdot 0{,}5\cdot 0{,}5=(0{,}5)^4=0{,}0625$
...		

$$H(X)=-\sum_{i=1}^{\infty}p(x_i)\log_2 p(x_i)=-\sum_{i=1}^{\infty}\left(\frac{1}{2}\right)^i\log_2\left(\frac{1}{2}\right)^i=-\sum_{i=1}^{\infty}i\left(\frac{1}{2}\right)^i\log_2\left(\frac{1}{2}\right)$$

$$=\sum_{i=1}^{\infty}i\left(\frac{1}{2}\right)^i=\frac{1}{2}\sum_{i=1}^{\infty}i\left(\frac{1}{2}\right)^{i-1}=\dots$$

Za izračunavanje zbroja u izrazu za entropiju poslužit ćemo se sljedećom jednakošću: $\sum_{i=0}^{\infty}x^i=\frac{1}{1-x}$; uz $|x|<1$.

Deriviranjem prethodnog izraza dobit ćemo:

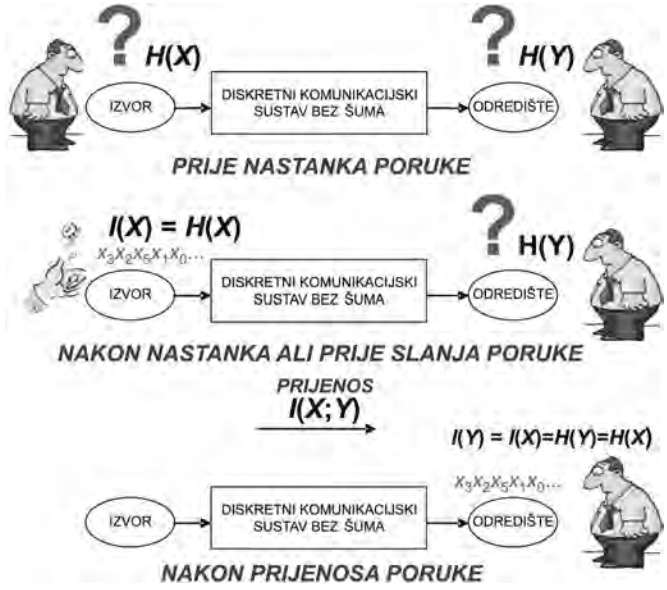
$$\sum_{i=0}^{\infty}ix^{i-1}=\frac{1}{(1-x)^2}=\sum_{i=1}^{\infty}ix^{i-1} \qquad (*)$$

Konačno, ako primijenimo (*) na $H(X)$, dobit ćemo:

$$H(X)=\frac{1}{2}\frac{1}{\left(1-1/2\right)^2}=2\ \left[\text{bit/simbol}\right].$$

2.2.2. Entropija, neodređenost i sadržaj informacije u sustavu bez smetnji

Promotrimo ponovo prijenos poruke u komunikacijskom sustavu bez smetnji (Slika 2.6). U trenutku prije nastanka poruke i na izvoru i na odredištu postoji nesigurnost, odnosno neodređenost, jer ne znamo koji će simboli biti odabrani. Mjera ove neodređenosti je entropija izvora $H(X)$. Iz njezine definicije već znamo da ona ovisi isključivo o svojstvima izvora informacije. Promatrana na ovaj način i izražena kao broj bita po simbolu, entropija daje mjeru za prosječnu neodređenost pojave nekog simbola na izvoru, a pritom jedan bit shvaćamo kao jednu elementarnu neodređenost – jedno bacanje "poštenog" novčića.



Slika 2.6: Entropija, neodređenost i sadržaj informacije u sustavu bez smetnji

Nakon nastanka poruke (tj. odabira simbola) neodređenost na izvoru nestaje, a stvoren je sadržaj informacije $I(X)$ upravo jednak dosadašnjoj neodređenosti. Na odredištu, gdje poruka još nije primljena, i dalje imamo neodređenost $H(Y)$. Nakon prijenosa poruke, i ova neodređenost nestaje primanjem sadržaja informacije $I(X)$.

Vidimo da su informacija i neodređenost dvije strane iste medalje. Neodređenost izvora je potencijal za nastajanje informacije u poruci, a informacija u poruci je potencijal za razrješavanje neodređenosti na odredištu. Jedan bit neodređenosti rješavamo jednim bitom informacije.

2.2.3. Svojstva entropije

Entropija ima niz svojstava koja proizlaze iz njezine definicije, a koja je čine pogodnom mjerom za sadržaj informacije, odnosno za neodređenost. Budući da su vjerojatnosti $p(x_i)$ ograničene između 0 i 1, entropija ne može biti negativna, tj. $H(X) \geq 0$.

Intuitivno je jasno da sadržaj informacije ne može biti negativan. Minimalni sadržaj informacije je nula. To je slučaj ilustriran u primjeru s novčićem koji uvijek daje pismo. Općenito, ako se jedan od simbola pojavljuje s vjerojatnošću 1 (iz čega proizlazi da je vjerojatnost pojave svih ostalih simbola 0), entropija je 0: $H(X) = 0 \Leftrightarrow \exists x_i$ takav da je $p(x_i) = 1$.

Entropija postiže maksimum kada su vjerojatnosti simbola jednako raspoređene. Za abecedu od n simbola, vjerojatnost pojavljivanja bilo kojeg simbola x_i je $p(x_i) = 1/n$. U tom slučaju jednostavnim uvrštavanjem $p(x_i)$ u izraz za entropiju (2.1) dobivamo: $H(X) = \log n$.

Može se pokazati da je ova vrijednost najveća moguća vrijednost entropije, tj. za bilo koju razdiobu vjerojatnosti pojavljivanja simbola iz abecede od n simbola vrijedi: $H(X) \leq \log n$.

Dokaz ove tvrdnje može se pronaći u [20] (str. 33). Promatranjem entropije kao mjere neodređenosti, intuitivno je jasno da je neodređenost najveća kada su vjerojatnosti jednako raspoređene po svim simbolima.

Promatrajući definiciju entropije može se postaviti pitanje zbog čega u njoj središnju ulogu igra logaritam. Zašto ne bismo sadržaj informacije mjerili na neki drugi način? Osnovni razlog je u činjenici da prosječni sadržaj informacije poruke koja se sastoji od kombinacija dvaju simbola mora biti jednak zbroju prosječnih sadržaja informacija poruka koje se sastoje od pojedinačnih simbola [22]. Konkretno, uzmimo dva međusobno neovisna skupa simbola – skup X od n simbola x_i i skup Y od m simbola y_j. Uređeni parovi simbola x_i i y_j tvore treći skup XY koji se sastoji od mn simbola oblika $(x_i y_j)$. Gradimo poruke sastavljene od simbola iz svakog od ova tri skupa. U prosjeku, sadržaj informacije kombiniranih simbola mora biti jednak zbroju sadržaja informacije pojedinačnih simbola. Drugim rječima, mjera H za sadržaj informacije mora imati svojstvo:

$$H(XY) = H(X) + H(Y). \tag{2.2}$$

Upravo logaritamska funkcija zadovoljava ovo svojstvo. Pretpostavimo radi jednostavnosti da su vjerojatnosti pojavljivanja jednako raspoređene unutar svakog skupa (dakle $p(x_i) = 1/n$, $p(y_j) = 1/m$, $p(x_i y_j) = 1/mn$). Tada je:

$$H(XY) = \log mn = \log m + \log n = H(X) + H(Y).$$

Potkrijepimo ovo razmatranje jednostavnim primjerom. Proširimo primjer bacanja novčića i zamislimo da se istovremeno bacaju dva novčića, te promatramo niz ovakvih bacanja. Možemo promatrati svaki novčić pojedinačno ili ih promatrati kao par. Na taj način ovakav niz bacanja generira tri poruke: niz rezultata prvog novčića, niz rezultata drugog novčića, i niz združenih rezultata. Jasno je da u prosjeku združeni rezultat bacanja obaju novčića nosi dvostruko više informacije nego rezultat bacanja pojedinog novčića, te očekujemo da mjera za sadržaj informacije to i pokaže. Ovo je pojednostavljeni slučaj izraza (2.2), gdje su pojedinačna bacanja simboli x_i i y_j, a združeni rezultat dvaju bacanja daje kombinirane simbole $x_i y_j$. Mjera za sadržaj informacije mora zadovoljavati ovo svojstvo, a zadovoljava ga isključivo logaritamska funkcija.

Bit i binarna znamenka

Bit, u smislu teorije informacije, je osnovna jedinica za količinu informacije. Količinu informacije od jednog bita možemo zamisliti kao jedan odgovor na binarno pitanje da ili ne, jedno bacanje "poštenog" novčića itd.

Međutim, u širem informatičkom smislu, pa i u svakodnevnom životu, uobičajeno je koristiti naziv "bit" za binarnu znamenku, dakle 0 ili 1 u binarnom zapisu broja.

Suštinsku razliku između bita kao jedinice informacije i bita kao binarne znamenke najlakše je uočiti na najjednostavnijem primjeru, već poznatom bacanju novčića. Recimo da pismo označavamo binarnom znamenkom 1, a glavu binarnom znamenkom 0. Promotrimo opet slučaj novčića koji uvijek daje pismo. Nakon deset bacanja imamo poruku: "1111111111". Očito je da ova poruka ima 10 bitova – binarnih znamenki. Međutim, ako govorimo o bitovima u smislu teorije informacije, ova poruka sadrži 0 bita.

Najčešće je iz konteksta jasno u kojem smislu se koristi riječ bit, te u ovoj knjizi to nećemo posebno napominjati, osim u slučaju kada postoji mogućnost nesporazuma.

2.3. Kodiranje

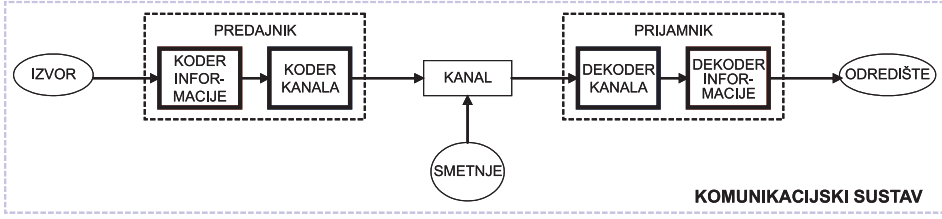
Kodiranje je jedan od najosnovnijih pojmova kako u teoriji, tako i u praksi informacijskih i komunikacijskih tehnologija. Premda većina čitatelja vjerojatno ima prilično dobru ideju o tome što je kodiranje, neophodno je dati jednostavnu i općenitu definiciju. Prisjetimo se da smo poruku u komunikacijskom sustavu definirali kao niz simbola odabranih iz abecede, gdje je abeceda konačan skup elementarnih simbola.

Definicija

Kodiranje je postupak dodjeljivanja kodnih riječi (kôdova) simbolima poruke. Svaka se kodna riječ sastoji od jednog ili više simbola iz neke druge abecede. Dakle, kodiranjem se poruka (niz simbola) pretvara u niz kodnih riječi.

Iz definicije je jasno da kodiranje u konačnici jednostavno pretvara poruku (niz simbola iz neke abecede) u drukčiji oblik poruke – niz simbola iz neke druge abecede. Postavlja se pitanje zbog čega bismo htjeli vršiti ovakvu pretvorbu. Odgovora na ovo pitanje ima onoliko koliko ima i različitih vrsta kodiranja, ali općenito se može reći da je smisao kodiranja pretvorba poruke u oblik koji ima neka bolja svojstva, npr. povoljniji je za prijenos, zaštitu ili pohranu. Na primjer, kompresija je vrsta kodiranja kod koje je kodirana poruka kraća od izvorne poruke; kriptografija je kodiranje kod kojeg kodirana poruka ima određena sigurnosna svojstva; zaštitno kodiranje daje poruci svojstva koja olakšavaju otkrivanje i/ili ispravljanje pogrešaka uzrokovanih smetnjama u prijenosu. U sklopu općenitog komunikacijskog sustava kodiranje se događa unutar predajnika i pritom koder informacije vrši kompresiju, a koder kanala zaštitno kodiranje (Slika 2.7).

U praktičnim primjenama u području informacijskih i komunikacijskih tehnologija, abeceda od koje se grade kodne riječi je gotovo uvijek binarna, dakle svaka kodna riječ je niz od jedne ili više binarnih znamenki, te je i kodirana poruka niz binarnih znamenki.



Slika 2.7: Kodiranje u sklopu komunikacijskog sustava

2.3.1. Kodiranje i entropija

Primjer

Kada se vrši kodiranje s ciljem kompresije podataka, jasno je da mora postojati neka granica do koje se može sažimati bez gubitaka. Ta granica je upravo entropija izvora. U poglavlju 5 (o entropijskom kodiranju) ovo ćemo pokazati na formalniji način, a zasad promotrimo jednostavan primjer.

Primjer: Kodiranje

Promatrajmo izvor koji proizvodi simbole iz skupa $X = \{1, 2, 3, 4\}$ s vjerojatnostima pojavljivanja koje daje Tablica 2.1.

Tablica 2.1: Primjer kodiranja

SIMBOL (x_i)	VJEROJATNOST POJAVLJIVANJA ($p(x_i) = p_i$)	KODNA RIJEČ (C_i)	DULJINA KODNE RIJEČI (l_i)
1	1/2	0	1
2	1/4	10	2
3	1/8	110	3
4	1/8	111	3

Jednostavnim uvrštavanjem vjerojatnosti p_i u izraz za entropiju (2.1) dobivamo da je entropija ovakvog izvora $H(X) = 1,75$ [bit/simbol].

Tablica 2.1 u trećem stupcu daje jedan mogući kôd za ovakav izvor, tj. način pridruživanja kodnih riječi simbolima. Na primjer, niz simbola na izvoru 134213 kodirao bi se kao 0110111100110. Pošto znamo vjerojatnosti pojavljivanja simbola, možemo jednostavno izračunati prosječnu duljinu kodne riječi (uz pretpostavku da su poruke dugačke, tj. da duljina poruke teži beskonačnosti):

$$L = \sum_{i=1}^{n} p_i l_i = 0,5 \cdot 1 + 0,25 \cdot 2 + 0,125 \cdot 3 + 0,125 \cdot 3 = 1,75 \ \left[\text{bit/simbol}\right].$$

Dakle, uz ovaj kôd potrebno je u prosjeku 1,75 bita (binarnih znamenki) po simbolu za kodiranje informacije s ovakvog izvora. Vidimo da je to upravo jednako entropiji izvora. Nemoguće je pronaći neki drugi skup kodnih riječi kojim bi se simboli s izvora mogli jednoznačno prikazati, a čija bi prosječna duljina kodne riječi bila manja od 1,75 bita po simbolu. Na primjer, ukoliko bismo naivno pokušali skratiti kodnu riječ C_4 i koristiti 11 umjesto 111, kôd više ne bi bio jednoznačan jer bi se kodirani niz 110 mogao dekodirati bilo kao simbol "3", bilo kao dva simbola "41".

Općenito, može se dokazati (vidi poglavlje 5.5) da je nemoguće jednoznačno kodirati simbole s nekog izvora uz prosječnu duljinu koda manju od entropije izvora. Drugim riječima, **entropija je granica kompresije bez gubitaka**. Ova temeljna činjenica ukazuje na praktičnu važnost entropije.