# Skriveni Markovljevi modeli

**Subotić, Aron**

**Master's thesis / Diplomski rad**

**2022**

*Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj:* **University of Rijeka / Sveučilište u Rijeci**

*Permanent link / Trajna poveznica:* https://urn.nsk.hr/urn:nbn:hr:196:382474

Fakultet za matematiku, Sveučilište u Rijeci
Diplomski studij Diskretna matematika i primjene

Aron Subotić

# Skriveni Markovljevi modeli

Diplomski rad

Rijeka, 2022.

Fakultet za matematiku, Sveučilište u Rijeci
Diplomski studij Diskretna matematika i primjene

Aron Subotić

# Skriveni Markovljevi modeli

Mentor: doc. dr. sc. Ivana Slamić

Diplomski rad
Rijeka, 2022.

# Sažetak

U radu su prikazana osnovna pitanja i problemi teorije skrivenih Markovljevih modela. Definirani su i ilustrirani na primjerima osnovni pojmovi teorije Markovljevih lanaca. Opisani su osnovni problemi koji se javljaju pri određivanju parametara skrivenog Markovljevog modela te algoritmi koji se koriste za rješavanje tih problema.

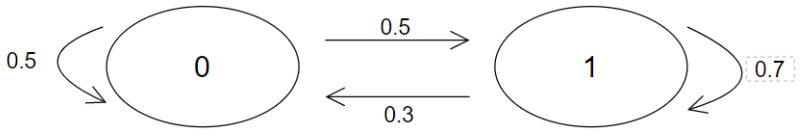
*Ključne riječi*: Markovljevi lanci; stacionarna distribucija; skriveni Markovljevi modeli; Viterbijev algoritam; Baum-Welch algoritam.

# Sadržaj

# 1 Uvod

Razne pojave iz svakodnevnog života, kao što su vremenski uvjeti, kretanja cijena dionica ili temperature, mutacije ili kemijske reakcije odvijaju se na slučajan način. Teorija slučajnih procesa prati tijek tih slučajnih pojava kroz vrijeme. Slučajna varijabla jedan je od temeljnih pojmova u teoriji vjerojatnosti, a slučajni proces možemo shvatiti kao niz slučajnih varijabli. Najjednostavniji je onaj kod kojeg su te varijable nezavisne i jednako distribuirane, kao što je, na primjer, broj koji je pao na kocki ili strana koja je pala na novčiću. *Markovljevi lanci* spadaju u najjednostavnije procese kod kojih se javlja zavisnost među varijablama te u određenom trenutku ishod ovisi o tome što je bilo u prošlom trenutku.

Markovljevi lanci nazvani su po ruskom matematičaru Markovu [1]. Početak razvoja teorije Markovljevih lanaca vezan je uz *slabi zakon velikih brojeva*. Prisjetimo se, taj rezultat kaže da prosjek nezavisnih, jednako distribuiranih varijabli $X_n$ s konačnim očekivanjem $\mu$, konvergira po vjerojatnosti prema očekivanju, tj. vrijedi $\frac{S_n}{n} \xrightarrow{\mathbb{P}} \mu$, kada $n \to \infty$, gdje je $S_n = X_1 + \cdots + X_n$. Markov je početkom dvadesetog stoljeća dokazao slabi zakon velikih brojeva za niz zavisnih slučajnih varijabli kojeg je ilustrirao sljedećim primjerom. Pretpostavimo da imamo dvije posude koje reprezentiramo kao slučajne varijable koje mogu poprimiti vrijednosti 0 ili 1. U jednoj posudi je jednak broj crnih i bijelih loptica i tu posudu ćemo označiti s 0, a u drugoj je 70% crnih loptica i 30 % bijelih loptica i nju ćemo označiti s 1. Na slučajan način izvlačimo loptice; ako je izvučena crna loptica, prelazimo u posudu 1, a ako je izvučena bijela, prelazimo u posudu 0. Nakon izvlačenja, loptica se vraća u istu posudu. Takav sustav možemo grafički prikazati kao na slici 1.




Slika 1: Dijagram Markovljevog lanca za posude

Ovaj jednostavni primjer opisuje osnovnu ideju Markovljevih lanaca. Vrijednost koju slučajna varijabla može poprimiti predstavlja *stanje* u kojem se neki sustav

[1] Andrej Andrejevič Markov (1859.-1922.) - ruski matematičar koji je pridonio razvoju teorije stohastičkih procesa.

može nalaziti. U svakom trenutku sustav može i ne mora promijeniti stanje, a pretpostavljamo da te promjene možemo opisati vjerojatnostima. Uočimo da je u ovom primjeru jednostavno izračunati te vjerojatnosti. Stanja sustava su posude 0 i 1, a u kojem će se stanju naći sustav ovisi o tome koja je loptica izvučena. Kako u posudi 0 imamo jednak broj crnih i bijelih loptica, znači da je vjerojatnost da sustav promijeni stanje (ako se nalazi u stanju 0) jednaka 0.5. Također, u posudi 1 je više crnih loptica pa je veća vjerojatnost da će sustav ostati u stanju 1.

Skriveni Markovljevi modeli proučavaju se još od kraja 1960-ih u svrhu primjene u obradi signala. Procesi u stvarnom svijetu kao što su financijska kretanja ili kretanja temperature proizvode vidljive rezultate koje možemo zvati signalima. Signali mogu biti diskretne prirode ili neprekidne. Također, signali mogu biti stacionarni, tj. ne mijenjaju im se statistička svojstva kroz vrijeme ili nestacionarni. Dobrim modelom možemo simulirati signale i efikasno učiti o njihovom ponašanju u budućnosti. Skriveni Markovljevi modeli dobili su ime po tome što se javlja Markovljev lanac koji je skriven, tj. stanja sustava nisu direktno vidljiva. Razvojem računalne tehnologije skriveni Markovljevi modeli postali su popularni u primjeni u automatskom prepoznavanju govora, strojnom učenju, bioinformatici itd.

Rad je podijeljen na četiri poglavlja. U drugom poglavlju dajemo osnovni pregled pojmova iz teorije vjerojatnosti koji su potrebni u nastavku rada. U trećem poglavlju uvodimo pojmove iz teorije Markovljevih lanaca, a u četvrtom definiramo skrivene Markovljeve lance.

# 2 Osnovni pojmovi teorije vjerojatnosti

U ovom poglavlju dajemo pregled osnovnih pojmova i rezultata teorije vjerojatnosti koje ćemo koristiti u nastavku rada. Kako bismo definirali pojam vjerojatnosti potrebno je najprije definirati $\sigma$-algebru.

**Definicija 2.1.** Familija $\mathcal{F}$ podskupova od $\Omega$ je *$\sigma$-algebra* skupova ako je:

(F1) $\emptyset \in \mathcal{F}$,

(F2) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$,

(F3) $A_i \in \mathcal{F}$ $(i \in \mathbb{N}) \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$ (zatvorenost na prebrojive unije).

**Definicija 2.2.** Neka je $\mathcal{F}$ $\sigma$-algebra na skupu $\Omega \neq \emptyset$. Uređen par $(\Omega, F)$ zovemo *izmjeriv prostor*. Elemente $\sigma$-algebre zovemo *događaji*.

**Definicija 2.3.** Neka je $(\Omega, \mathcal{F})$ izmjeriv prostor. Funkcija $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ je *vjerojatnost* ako vrijedi:

(P1) $\mathbb{P}(A) \geq 0$ za svaki događaj $A \in \mathcal{F}$ (nenegativnost),

(P2) $\mathbb{P}(\Omega)$=1 (normiranost vjerojatnosti),

(P3) Ako su $A_i \in \mathcal{F}$, $i \in \mathbb{N}$ i $A_i \cap A_j = \emptyset$ za $i \neq j$, onda je:

$$\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

($\sigma$-aditivnost ili prebrojiva aditivnost vjerojatnosti)

Uvjete iz prethodne definicije zovemo *aksiomi Kolmogorova*[2].

**Definicija 2.4.** Uređenu trojku $(\Omega, \mathcal{F}, \mathbb{P})$, gdje je $\mathcal{F}$ $\sigma$-algebra na skupu $\Omega$ i $\mathbb{P}$ vjerojatnost na $\mathcal{F}$, zove se *vjerojatnosni prostor*. Broj $\mathbb{P}(A)(A \in \mathcal{F})$ zovemo *vjerojatnost događaja A*.

Ako je $\Omega$ konačan ili prebrojiv skup u tom slučaju vjerojatnosni prostor zovemo *diskretan vjerojatnosni prostor*.

[2]Andrej Nikolaevich Kolmogorov (1903-1987) - ruski matematičar koji je postavio aksiome teoriji vjerojatnosti.

**Definicija 2.5.** Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor i $A, B \in \mathcal{F}$. Kažemo da su događaji $A$ i $B$ *nezavisni* ako vrijedi $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

Definiramo funkciju $P_A$ sa:

$$\mathbb{P}_A(B) = \mathbb{P}(B|A) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)}, \quad A, B \in \mathcal{F}.$$

Može se pokazati da je $\mathbb{P}_A$ vjerojatnost te je nazivamo *uvjetna vjerojatnost*. Broj $\mathbb{P}(B|A)$ zovemo vjerojatnost od $B$ uz uvjet da se dogodio $A$. Uvjetna vjerojatnost je jedan od ključnih pojmova kojeg koristimo za definiranje Markovljevih lanaca.

Najjednostavniji primjer slučajanog pokus je bacanje novčića, a ishodi pokusa su *pismo* ili *glava*. Preciznije, ove ishode nazivamo elementarnim događajima. Ako bismo htjeli analizirati ponašanje takve slučajne pojave, onda je korisno uvesti neku varijablu $X$, definiranu na skupu $\Omega$ koja će poprimiti vrijednost 0 ako je palo pismo, odnosno 1 ako je pala glava. Na taj način, definirali smo jednu slučajnu varijablu.

**Definicija 2.6.** Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor. *Slučajna varijabla* je funkcija $X : \Omega \to \mathbb{R}$ takva da vrijedi

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}, \ \forall B \in \mathcal{B},$$

gdje je $\mathcal{B}$ Borelova $\sigma$-algebra na $\mathbb{R}$.

Prisjetimo se, Borelova $\sigma$-algebra je najmanja $\sigma$-algebra koja sadrži sve otvorene skupove u $\mathbb{R}$. Elemente Borelove $\sigma$-algebre nazivamo *Borelovim skupovima*. Može se pokazati da je Borelova $\sigma$-algebra najmanja $\sigma$-algebra generirana nekom od sljedećih familija: $\{\langle a, b\rangle : a < b\}$, $\{[a, b] : a < b\}$, $\{\langle a, b] : a < b\}$, $\{[a, b\rangle : a < b\}$,$\{\langle a, \infty\rangle : a \in \mathbb{R}\}$, $\{\langle -\infty, a\rangle : a \in \mathbb{R}\}$, $\{[a, \infty\rangle : a \in \mathbb{R}\}$, $\{\langle -\infty, a] : a \in \mathbb{R}\}$.

Slučajne varijable dijelimo na diskrente i neprekidne.

**Definicija 2.7.** Kažemo da je slučajna varijabla $X$ *diskretna* ako postoji prebrojiv skup $\{a_1, a_2, \dots\} \subset \mathbb{R}$ tako da je $\{X = a_i\} \in \mathcal{F}$, za svaki $i$, i $\mathbb{P}(X \in \{a_1, a_2, \dots\}) = 1$.

**Definicija 2.8.** Kažemo da je slučajna varijabla $X$ *neprekidna* ako postoji nenegativna izmjeriva funkcija $f = f_X : \mathbb{R} \to [0, 1\rangle$ tako da vrijedi

$$\mathbb{P}(a \leq X < b) = \int_a^b f_X(t)\, dt. \tag{1}$$

Funkcija $f_X$ naziva se gustoća slučajne varijable $X$.

U nastavku ćemo trebati i pojam slučajnog vektora:

**Definicija 2.9.** Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor. Funkcija $\mathbf{X} : \Omega \to \mathbb{R}^n$ je $n$-dimenzionalan slučajni vektor ako je

$$\mathbf{X}^{-1}(B) \in \mathcal{F} \text{ za svaki } B \in \mathcal{B}^n,$$

gdje je $\mathcal{B}^n = \sigma\{< a_1, a_2 > \times \cdots \times < a_{n-1}, a_n >: a_i \in \mathbb{R}, a_i < a_{i+1}\}$ Borelova $\sigma$-algebra na $\mathbb{R}^n$

Svaki $n$-dimenzionalan slučajni vektor na $\Omega$ je uređena $n$-torka slučajnih varijabli na $\Omega$. Za neprekidnu slučajnu varijablu, funkcija gustoće zadana je već definicijom. Kod diskretne slučajne varijable $X$ definiramo *funkciju gustoće* sa:

$$f(x) = \mathbb{P}(X = x), x \in \mathbb{R}. \tag{2}$$

Za bilo koju slučajnu varijablu, *funkcija distribucije* definirana je sa:

$$F(x) = \mathbb{P}(X \leq x), x \in \mathbb{R}. \tag{3}$$

Diskretna slučajna varijabla je određena tablicom distribucije

$$\begin{pmatrix} a_1 & a_2 & \dots \\ p_1 & p_2 & \dots \end{pmatrix},$$

gdje su $a_i, i \in \mathbb{N}$ vrijednosti koje $X$ može poprimiti, a $p_i$ vjerojatnost da $X$ poprimi vrijednost $a_i$. Za bilo koji vektor $(p_1, p_2, \dots)$ takav da

$$p_i \geq 0, \forall i \in \mathbb{N} \text{ i } \sum_{i=1}^{n} p_i = 1$$

postoji slučajna varijabla kojoj je to distribucija. Takav vektor ćemo zvati *vjerojatnosna distribucija*. U nastavku definiramo uvjetnu funkciju gustoće

**Definicija 2.10.** Neka je $(X, Y)$ slučajni vektor s gustoćom $f = f_{X,Y}$ i neka je $f_Y$ gustoća od $Y$. Definiramo *uvjetnu funkciju gustoće* od slučajne varijable $X$ za dano $Y = y$ je funkcija $x \mapsto f_{X|Y}(x|y)$ dana sa:

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_y(y)}, x \in \mathbb{R},$$

uz uvjet da je $f_Y(y) > 0$. Analogno definiramo uvjetnu funkciju gustoće od $Y$ uz dani $X = x$.

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)},$$

Ako je $f_X(x) > 0$.

*Napomena:* U diskretnom slučaju vrijedi $f(x|y) = \mathbb{P}(X = x|Y = y)$, ali u neprekindom ne.

**Definicija 2.11.** *Funkcija distribucije* $n$-dimenzionalnog slučajnog vektora $\mathbf{X} = (X_1, \ldots, X_n) : \Omega \to \mathbb{R}^n$ na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$ je funkcija $F : \mathbb{R}^n \to [0, 1]$ definirana na sljedeći način:

$$F(\mathbf{X}) := \mathbb{P}_{\mathbf{X}}(X_1 \leq x_1, \ldots, X_n \leq x_n),$$

gdje je $(x_1, \cdots, x_n) \in \mathbb{R}^n$.

**Definicija 2.12.** *Matematičko očekivanje* slučajne varijable $X$ definira se kao broj

$$\mathbb{E}X = \begin{cases} \sum\limits_{x \in ImX} x \cdot f(x), & \text{ako je X diskretna slučajna varijabla,} \\ \int_{-\infty}^{+\infty} x f(x)\, dx, & \text{ako je X neprekidna slučajna varijabla,} \end{cases}$$

pod pretpostavkom da red, odnosno integral u definiciji apsolutno konvergira.

**Definicija 2.13.** Neka je $X$ slučajna varijabla i neka $\mathbb{E}X$ postoji. *Varijanca* od $X$ definira se sa

$$\mathrm{Var}X = \begin{cases} \mathbb{E}\left[(X - \mathbb{E}(X))^2\right], & \text{ako je X diskretna slučajna varijabla,} \\ \int_{-\infty}^{+\infty} (x - \mathbb{E}X)^2 f(x)\, dx, & \text{ako je X neprekidna slučajna varijabla,} \end{cases}$$

pod pretpostavkom da očekivanje u prvom slučaju postoji, odnosno da integral u drugom slučaju konvergira.

Kako se kod Markovljevih lanaca javlja zavisnost među varijablama, prisjetit ćemo se definicije uvjetnog očekivanja.

**Definicija 2.14.** *Uvjetno očekivanje* slučajne varijable $X$ za zadanu vrijednost $Y = y$:

$$\mathbb{E}\left[X|Y=y\right] = \begin{cases} \sum\limits_{x \in ImX} x \cdot f_{X|Y}(x|y), & \text{ako je X diskretna slučajna varijabla,} \\ \int_{-\infty}^{+\infty} x f_{X|Y}(x|y)\, dx, & \text{ako je X neprekidna slučajna varijabla,} \end{cases}$$

gdje je $f_{X|Y}(x|y)$ uvjetna funkcija gustoće slučajnog vekotra $X$ za zadanu vrijednost $Y = y$.

# 3 Markovljevi lanci

Promotrimo sljedeći primjer. Pretpostavimo da pratimo vlastito zdravlje iz dana u dan i da postoje dva ishoda:

- *bolestan* (stanje 1): u tom slučaju ne možemo otići na posao
- *zdrav* (stanje 2): u tom slučaju možemo otići na posao

Pretpostavimo sada da nas zanimaju neka od sljedećih pitanja: Možemo li sutra na posao? Možemo li za tri dana na posao? Hoćemo li biti na poslu cijeli tjedan? Koliko dana možemo očekivati da ćemo biti bolesni? Htjeli bismo izraditi matematički model koji bi nam dao odgovor na ovakva pitanja. Reprezentiramo dane kao slučajne varijable koje mogu poprimiti dvije vrijednosti bolestan ili zdrav, što možemo označiti sa 1 ili 0. Razumno je pretpostaviti to da je netko zdrav u nekom danu u tjednu, ovisi o tome u kakvom je stanju prethodnih dana, a najviše prethodnog dana. Takvom smo pretpostavkom pretpostavili da je riječ o nizu zavisnih slučajnih varijabli.

## 3.1 Osnovni pojmovi i definicije

Markovljevi lanci su slučajni procesi kod kojih se javlja zavisnost među varijablama, stoga najprije uvodimo definiciju slučajnog procesa.

**Definicija 3.1.** Neka je $S$ skup. *Slučajan proces* s diskretnim vremenom i prostorom stanja $S$ je familija $X = (X_n : n \geq 0)$ slučajnih varijabli definiranih na nekom vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$ s vrijednostima u $S$.

Dakle, za svaki $n \geq 0$, $X_n : \Omega \to S$ je slučajna varijabla. Skup stanja $S$ može biti *diskretan* (konačan ili prebrojiv) ili *općí* (neprebrojiv). U ovom radu ćemo promatrati samo Markovljeve lance s diskretnim skupom stanja. Pod pojmom *konačnodimenzionalne distribucije slučajnog procesa* $(X_n, n \geq 0)$ podrazumijevat ćemo da su to distribucije vektora $(X_1, \ldots, X_k), k \in \mathbb{N}$.

**Definicija 3.2.** Neka je $S$ prebrojiv skup. Slučajni proces $X = (X_n : n \geq 0)$ definiran na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$ s vrijednostima u skupu $S$ je *Markovljev lanac* ako vrijedi

$$\mathbb{P}(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = \mathbb{P}(X_{n+1} = j | X_n = i) \qquad (4)$$

za svaki $n \geq 0$ i za sve $i_0, \ldots, i_{n-1}, i, j \in S$ za koje su obje uvjetne vjerojatnosti dobro definirane.

Svojstvo (4) iz prethodne definicije zove se *Markovljevo svojstvo*. Ako pretpostavimo da se nalazimo u vremenskom trenutku $n$, onda je vrijeme $n+1$ *neposredna budućnost*, dok vremena $0, 1, \ldots, n-1$ predstavljaju *prošlost*. Markovljevo svojstvo možemo interpretirati na sljedeći način. Ponašanje Markovljevog lanca u neposrednoj budućnosti ovisi samo o stanju lanca u sadašnjem trenutku, te nije bitno ponašanje Markovljevog lanca u prošlosti. Prisjetimo se uvodnog primjera - posudica u kojoj ćemo se naći ovisi samo o posudici u kojoj se trenutno nalazimo.

Nas će zanimati samo *homogeni Markovljevi lanci*. To su oni lanci koji ne ovise o vremenu $n \geq 1$, tj. takvi da vrijedi uvjet

$$\mathbb{P}(X_{t_{n+1}} = j | X_{t_n} = i) = \mathbb{P}(X_{t_{n+1}-t_n} = j | X_0 = i).$$

U uvodnom primjeru homogenost je osigurana tako što lopticu nakon izvlačenja vraćamo u istu posudu pa se vjerojatnost prijelaza ne mijenja s vremenom. Uočimo da Markovljevo svojstvo uz vremensku homogenost prelazi u

$$\mathbb{P}(X_{n+1} = j | X_n = i) = \mathbb{P}(X_1 = j | X_0 = i). \qquad (5)$$

U nastavku nam je cilj definirati matricu prijelaza za što će nam trebati pojam stohastičke matrice.

**Definicija 3.3.** Matrica $P = (p_{ij} : i, j \in S)$ naziva se *stohastička matrica* ako je $p_{ij} \geq 0$ za sve $i, j \in S$, te

$$\sum_{j \in S} p_{ij} = 1, \text{ za sve } i \in S.$$

Sada uvedimo pojam matrice prijelaza u koju spremamo vjerojatnosti prijelaza iz stanja u stanje:

**Definicija 3.4.** Neka je $\Pi = (\pi_i : i \in S)$ vjerojatnosna distribucija na S, te neka je $P = (p_{ij} : i, j \in S)$ stohastička matrica. Slučajni proces $X = (X_n : n \geq 0)$ definiran na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$ s prostorom stanja S je *homogen Markovljev* lanac s početnom distribucijom $\Pi$ i matricom prijelaza $P$ ako vrijedi

(i) $\mathbb{P}(X_0 = i) = \pi_i$ za sve $i \in S$, te

(ii) $\mathbb{P}(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = p_{ij}$, za svaki $n \geq 0$ i za sve $i_o, \ldots, i_{n-1}, i, j \in S$.

Broj $p_{ij}$ predstavlja vjerojatnost prijelaza lanca $(X_n, n \geq 0)$ iz stanja $i$ u stanje $j$ u jednom koraku. Matricu $P = (p_{ij} : i, j \in S)$ ćemo zvati *matricom prijelaza* lanca $(X_n, n \geq 0)$. Kako se u prethodnoj definiciji radi o vjerojatnostima, a redak u takvoj matrici predstavlja sve moguće prijelaze, zbroj vrijednosti u svakom retku mora biti jednak 1.

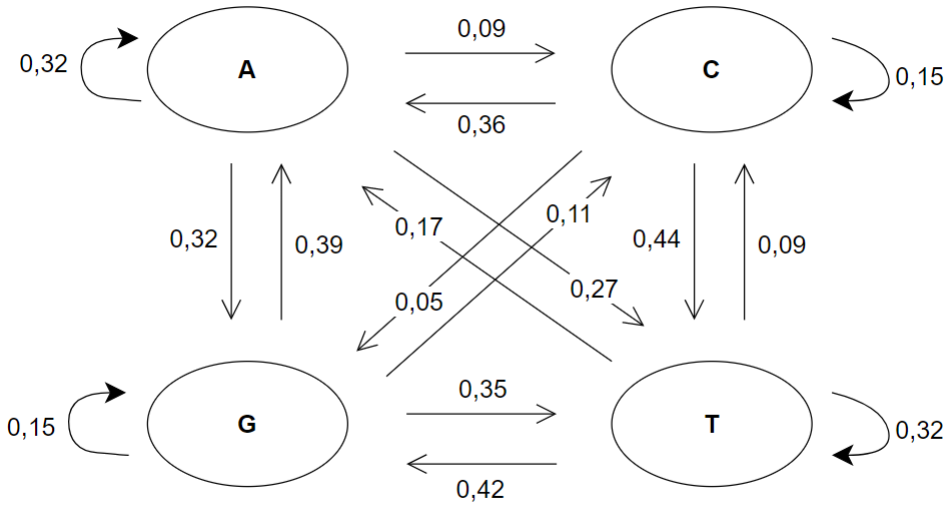
**Primjer 3.1.** Markov je primijenio proces koji danas znamo kao Markovljev lanac u analizi poezije Aleksandra Puškina "Evgenij Onjegin". Markov je proučavao niz od 20000 slova i otkrio je da je vjerojatnost da nakon samoglasnika dolazi samoglasnik $p_1 = 0.128$, da je vjerojatnost da nakon suglasnika sljedi samoglasnik $p_2 = 0.663$.itd.

Pretpostavimo općenito da želimo analizirati neki tekst na engleskom jeziku. Najjednostavniji model bio bi gdje se slova javljaju nezavisno i slučajno. Kako engleska abeceda sadrži 26 slova, možemo pretpostaviti da se svako slovo javlja s vjerojatnošću $\frac{1}{26}$. Naravno, to nije baš najrealističnija pretpostavka jer se rijetko javljaju slova BQ,QG, itd. Dakle, možemo pretpostaviti da pojavljivanje slova ovisi o prethodnom slovu. Ako bismo htjeli koristiti Markovljeve lance, broj stanja bio bi 26, odnosno matrica prijelaza bila bi dimenzije $26 \times 26$. Promotrimo stoga sličan primjer, kod kojeg će ta dimenzija biti manja. Svaki se lanac DNK sastoji od građevnih jedinica nukleotidi koji su određeni jednom od četiri dušične baze: adenin (A), citozin (C), gvanin (G) i timin (T). Stoga na najelementarnijem nivou DNK niz možemo shvatiti kao niz slova A,C,T,G te, u svrhu statističke analize možemo jednu poziciju u nizu shvatiti kao rezultat slučajne varijable koja poprima vrijednosti u skupu A,C,T,G. Promotrimo sljedeći segment DNK niza:

GATCATTGATATGTTGCTAGAACTATGAGTGTTAAAGGTGCTTGT
GGTGAGTTATCAGACAGAAACGCAGAAGATGTTATTGGAAGCTTG
AGGAAAAGTGATCCTGGATTTACAGTGCCAAGAATTGGCCTGTAT
TGTGTTCTCAATGTTTTTGAGGAAGGTAGAAACTGTAAGTGATGA

Markovljevi lanci često se koriste u analizi genetičkih nizova. Ako pretpostavimo da je ovaj niz rezultat opisanog Markovljevog procesa, onda za ovaj primjer matricu prijelaza možemo odrediti na sljedeći način. Uočimo, u 53 javljanja slova A, 17 puta je nakon A slijedilo A (32.1%), 5 puta je slijedilo slovo C (9.4%), 17 puta slovo G (32.1%) i 14 puta slovo T (26.4%). Prebrojavanje ponovimo za slvoa C,G i T te tako procijenimo vjerojatnosti prijelaza i formiramo matricu prijelaza:

$$\mathrm{P}= \begin{bmatrix} 0.32 & 0.09 & 0.32 & 0.27 \\ 0.36 & 0.15 & 0.05 & 0.44 \\ 0.39 & 0.11 & 0.15 & 0.35 \\ 0.17 & 0.09 & 0.42 & 0.32 \end{bmatrix}.$$




Slika 2: Dijagram stanja Markovljevog lanca za nukleotide u DNK-u

## 3.2 Chapman-Kolmogorovljeva jednakost

Pomoću matrice prijelaza možemo odrediti vjerojatnost da se u $n$ vremenskih trenutaka nađemo u pojedinom stanju. Za uvodni primjer to bi značilo da bismo pomoću

matrice prijelaza primjerice mogli odrediti kolika je vjerojatnost da se nakon pet izvlačenja loptica nalazimo u stanju 0. O tome govori sljedeći teorem.

**Teorem 3.1.** Neka je $(X_n, n \geq 0)$ Markovljev lanac. Tada za sve $n \geq 0$ i za sva stanja $i_0, \ldots, i_{n-1}, i_n$ vrijedi

$$\mathbb{P}(X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n) = \pi_{i_0} p_{i_0 i_1} \ldots p_{i_{n-1} i_n}. \tag{6}$$

*Dokaz.* Indukcijom iz $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B|A)$ slijedi $\mathbb{P}(A_0 \cap A_1 \cap \cdots \cap A_n) = \mathbb{P}(A_0)\mathbb{P}(A_1|A_0)\mathbb{P}(A_2|A_0 \cap A_1) \ldots \mathbb{P}(A_n| \bigcap_{i=0}^{n-1} A_i)$. Iz te formule slijedi

$$\begin{aligned}
&\mathbb{P}(X_0 = i_0, X_1 = i_1, \ldots, X_{n-1} = i_{n-1}, X_n = i_n) \\
&\quad = \mathbb{P}(X_0 = i_0)\mathbb{P}(X_1 = i_1|X_0 = i_0) \ldots \mathbb{P}(X_n = i_n|X_0 = i_0, \ldots, X_{n-1} = i_{n-1}) \\
&\quad = \pi_{i_0} p_{i_0 i_1} \ldots p_{i_{n-1} i_n}.
\end{aligned}$$

□

Sljedeći teorem nam daje obrat prethodnog teorema.

**Teorem 3.2.** Pretpostavimo da je $X = (X_n : n \geq 0)$ slučajni proces s konačnodimenzionalnim distribucijama danim formulom (2) iz prethodnog teorema, gdje je $\Pi$ neka vjerojatnosna distribucija na $S$, a $P$ neka matrica prijelaza na $S$. Tada je $(X_n, n \geq 0)$ Markovljev lanac.

*Dokaz.* Kako bismo dokazali ovaj teorem, treba pokazati da vrijede tvrdnje (1) i (2) iz definicije 3.4. Uzimanjem $n = 0$ odmah slijedi da je $\lambda$ početna distribucija. Sada dokazujemo formulu (2).

Pretpostavimo da je $\mathbb{P}(X_0 = i_0, \ldots, X_n = i) > 0$. Tada je

$$\begin{aligned}
\mathbb{P}(X_{n+1} = j|X_n = i, N_{n-1} = i_{n-1}, \ldots, X_0 = i_0) &= \frac{\mathbb{P}(X_{n+1} = j, X_n = i, \ldots, X_0 = i_0)}{\mathbb{P}(X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0)} \\
&= \frac{\lambda_{i_0} p_{i_0 i_1} \ldots p_{i_{n-1} i} p_{ij}}{\lambda_{i_0} p_{i_0 i_1} \ldots p_{i_{n-1} i}} \\
&= p_{ij}.
\end{aligned}$$

□

Pretpostavimo da je matrica prijelaza nekog lanca $(X_n, n \geq 0)$ $P$. Pogledajmo sada matricu $P^2$. Njene elemente ćemo označiti s $p_{ij}^2$, $i, j \in S$. Znamo da je

$$p_{ij}^2 = \sum_{k \in S} p_{ik} p_{kj}.$$

Broj $p_{ij}^2$ je zapravo vjerojatnost da je lanac $(X_n, n \geq 0)$ došao iz stanja $i$ u stanje $j$ u dva koraka. Analogno, elementi matrice $P^{n+1}$ su dani s

$$p_{ij}^{n+1} = \sum_{k \in S} p_{ik}^n p_{kj} = \sum_{k \in S} p_{ik} p_{kj}^n = \sum_{k_1 \in S} \sum_{k2 \in S} \cdots \sum_{k_n \in S} p_{ik_1} p_{k_1 k_2} \ldots p_{k_n j}.$$

Dakle, opet zaključujemo da broj $p_{ij}^{n+1}$ predstavlja vjerojatnost prelaska lanca $(X_n, n \geq 0)$ iz stanja $i$ u stanje $j$ u $(n+1)$ koraka. Iz gornjih razmatranja dolazimo do *Chapman*[3]*-Kolmogorovljeve* jednakosti:

$$p_{ij}^{(m+n)} = \sum_{k \in S} p_{ik}^m p_{kj}^n, \text{ za sve } m, n \in \mathbb{N}_0.$$

**Primjer 3.2.** Pretpostavimo da jesenska prognoza u Rijeci, iz dana u dan, može modelirati Markovljev proces. Skup stanja je $S = \{0, 1, 2\}$, gdje 0 predstavlja kišu, 1 vjetar ili oblačno i 2 sunčano vrijeme. Vjerojatnosti da određeni dan bude kiša, ako je prethodni dan bilo sunce, dana je matricom prijelaza:

$$\mathrm{P} = \begin{bmatrix} 0.5 & 0.2 & 0.3 \\ 0.2 & 0.2 & 0.6 \\ 0.4 & 0.1 & 0.5 \end{bmatrix}.$$

Pretpostavimo da je danas petak i pada kiša. Kolika je vjerojatnost da će u nedjelju biti sunčano?

Ovo pitanje ćemo riješiti pomoću Chapman-Kolmogorovljeve jednakosti. Kako bismo došli iz stanja 0 do stanja 2 potrebno je proći međustanjem, tj. potrebno je uzeti u obzir vrijeme u subotu. Dakle, promatramo tri niza: kiša-kiša-sunčano ili kiša-oblačno-sunčano ili kiša-sunčano-sunčano. Kako znamo da je početno stanje kiša izračunamo vjerojatnost prvog niza tako da pomnožimo vjerojatnost prijelaza iz kiše u kišu te iz kiše u sunčano. Analgno napravimo za druga dva niza. Rješenje problema je zbroj vjerojatnosti prethodna tri niza.

[3]Sydney Chapman (1888-1970.) - britanski matematičar i geofizičar.

$$p_{02}^2 = \sum_{i \in S} p_{0i} p_{i2} = 0.5 \cdot 0.3 + 0.2 \cdot 0.6 + 0.3 \cdot 0.5 = 0.42.$$

## 3.3 Klasifikacija stanja Markovljevih lanaca

U ovom poglavlju ćemo opisati kako se stanja Markovljevog lanca klasificiraju s obzirom na relaciju dostižnosti. Postavlja se pitanje koja stanja lanac uopće može posjetiti krenuvši iz nekog zadanog stanja. Započinjemo s definicijom regularnih Markovljevih lanaca.

**Definicija 3.5.** Kažemo da je Markovljev lanac *regularan* ako vrijedi:

$$(\exists n \geq 0) : p_{ij}^{(n)} > 0, \text{ za svaki}(i, j). \tag{7}$$

Primijetimo, ako (7) vrijedi za neki $n \geq 0$, vrijedit će i za svaki $n' \geq n$. Dakle, neovisno o stanju $i$, nakon konačnog broja koraka postoji pozitivna vjerojatnost u regularnom Markovljevom lancu da se proces nalazi u bilo kojem stanju $j$. Kod uvodnog primjera i jedna i druga posudica imaju crne i bijele loptice, što znači da se u bilo kojem trenutku možemo naći u posudici 1 ili u posudici 0. Dakle, to je jedan jednostavan primjer regularnog lanca jer imamo samo dva stanja. Primjer s DNK je također bio primjer regularnog Markovljevog lanca.

**Definicija 3.6.** Za stanja $i, j \in S$ kažemo da je stanje $j$ *dostižno* iz stanja $i$, u oznaci $i \to j$, ako vrijedi

$$\mathbb{P}_i(T_j < \infty) > 0.$$

Vrijednost $T_j$ iz prethodne definicije je definirana kao

$$T_j = \min\{n \geq 0 : X_n = j).$$

Minimum postoji ako sustav može iz stanja $i$ prijeći u stanju $j$. Dakle, to je prvo vrijeme dolaska u stanje $j$. Sljedeća propozicija nam daje kriterij dostižnosti:

**Propozicija 3.1.** Sljedeća svojstva su ekvivalentna:

(i) $i \to j$,

(ii) $p_{ij}^{(n)} > 0$ za neko $n \geq 0$,

(iii) $p_{ii_1}p_{i_1i_2}\dots p_{i_{n-1}j}$ za neka stanja $i_1, \dots, i_{n-1}$.

*Dokaz.* Budući da je $\{X_n = j\} \subsetneq \bigcup_{k=0}^{\infty} \{X_k = j\} = \{T_j < \infty\}$, slijedi

$$p_{ij}^{(n)} = \mathbb{P}_i(X_n = j) \leq \mathbb{P}_i(T_j < \infty) = \mathbb{P}_i(\bigcup_{k=0}^{\infty}\{X_k = j\}) \leq \sum_{k=0}^{\infty} \mathbb{P}_i(X_k = j) = \sum_{k=0} p_{ij}^{(k)}.$$

To dokazuje ekvivalenciju (i) i (ii). Ekvivalencija tvrdnji (ii) i (iii) slijedi iz formule

$$p_{ij}^{(n)} = \sum_{i_1 \in S} \cdots \sum_{i_{n-1} \in S} p_{ii_1}p_{i_1i_2}\dots p_{i_{n-1}j}.$$

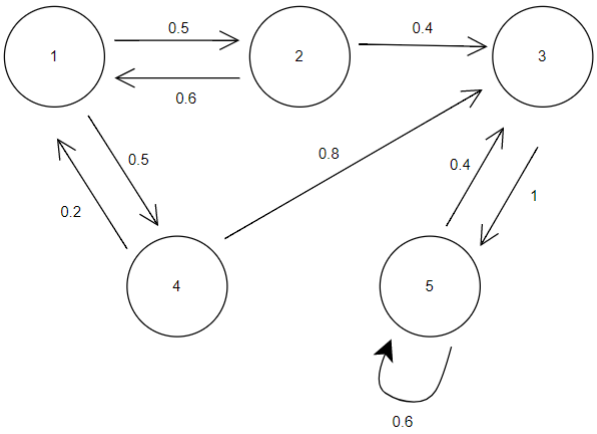
□

Za stanja $i$ i $j$ za koja vrijedi $i \to j$ ne mora vrijediti $j \to i$, međutim ako to vrijedi, za takva stanja ćemo reći da *komuniciraju*, tj.

**Definicija 3.7.** Stanja $i, j \in S$ *komuniciraju*, u oznaci $i \leftrightarrow j$, ako vrijedi $i \to j$ i $j \to i$.

Ako se sustav nalazi u stanju $i$, moguće je da se nakon nekoliko koraka nađe u stanju $j$. Također ako se nalazi u stanju $j$ moguće je da se kasnije nađe u stanju $i$. Može se pokazati da je relacija komuniciranja relacija ekvivalencije na $S \times S$ te inducira particiju skupa $S$ na klase koje ćemo označiti sa $C_i,\ i = 1, \dots, |S|$.

**Primjer 3.3.** Promotrimo idući Markovljev lanac sa skupom stanja $S = \{1, 2, 3, 4, 5\}$ i vjerojatnostima prijelaza prikazanim na slici:




Slika 3: Primjer Markovljevog lanca u kojem se javljaju dvije klase povezanosti

Uočimo da iz stanja 1 možemo doći u stanja 2 i 4. Nadalje, iz stanja 2 možemo u stanje 1, te preko stanja 1 možemo u stanje 4. Također iz stanja 4 možemo u stanje 1 i preko stanja 1 u stanje 2. Dakle, stanja 1,2,4 komuniciraju. U stanje 1 možemo doći iz stanja 1 preko 2 ili preko 4. Slično vrijedi za stanja 2 i 4. Također stanja 3 i 5 komuniciraju. Time smo odredili particiju skupa

$$\mathcal{P}(S) = \{\{1, 2, 4\}, \{3, 5\}\}.$$

**Definicija 3.8.** Markovljev lanac $\{X_n\}_{n\in\mathbb{N}_0}$ je *ireducibilan* ako se prostor stanja $S$ sastoji od jedne klase komuniciranja, tj. za sve $i, j \in S$ vrijedi $i \leftrightarrow j$.

U prethodnom primjeru lanac bi bio ireducibilan kada bi postojao prijelaz iz 3 u 1,2 ili 4 ili prijelaz iz 5 u 1,2 ili 4. Ireducibilni Markovljevi lanci na prvu izgledaju jednaki regularnim Markovljevim lancima. U idućoj propoziciji ćemo opisati njihov odnos.

**Propozicija 3.2.** Svi regularni Markovljevi lanci su ireducibilni.

*Dokaz.* Dokaz slijedi iz definicije regularnog Markovljevog lanca: $(\exists\, n \geq 0)$: $p_{ij}^{(n)} > 0$, za svaki $(i, j)$. Slijedi da možemo doći do stanja $j$ iz bilo kojeg stanja $i$, iz čega zaključujemo da je Markovljev lanac ireducibilan. □

Međutim obrat ove propozicije ne vrijedi, što ćemo pokazati na jednom kratkom primjeru.

**Primjer 3.4.** Neka je zadana sljedeća matrica prijelaza nekog lanca:

$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Ovakav lanac je očito ireducibilan, ali nije regularan. Uočimo: $P^n = \mathcal{I}$ ako je $n$ paran i $P^n = P$ ako je $n$ neprana. Dakle, ne postoji $n$ za kojeg su svi elementi u matrici veći od 0.

**Definicija 3.9.** Bilo koji podskup stanja $C \subset S$ takav da ($\forall i \in C, j \notin C$: $i \not\to j$) nazivamo *zatvoren skup*.

Dakle, bilo koje stanje iz $C$ ne može doseći stanje izvan $C$, bez obzira o broju koraka.

**Primjer 3.5.** Neka je $X$ proces sa skupom stanja $S = \{1, 2, 3\}$ i neka je matrica prijelaza:

$$\mathrm{P}= \begin{bmatrix} 0.3 & 0.2 & 0.5 \\ 0 & 0.3 & 0.7 \\ 0 & 0.5 & 0.5 \end{bmatrix}.$$

Možemo uočiti da ako se nalazimo u stanju 1 i u sljedećem koraku pređemo u stanje 2 da se više ne možemo vratiti u stanje 1. Također vrijedi i ako pređemo u stanje 3. Dakle, u ovom primjeru $C = \{2, 3\}$ je zatvoren skup.

**Definicija 3.10.** Za stanje $i \in S$ kažemo da je *apsorbirajuće* ako vrijedi $p_{ii} = 1$. Za Markovljev lanac $\{X_n\}_{n\in\mathbb{N}_0}$ kažemo da je *apsorbirajući* ako za svako stanje iz $S$ postoji neko apsorbirajuće stanje u koje možemo doći.

Za apsorbirajuće stanje vrijedi da kada jednom uđemo u njega ne možemo više izaći. Primjetimo da je to stanje koje čini jednočlan zatvoreni skup. U primjeru 3.3 kada bi vjerojatnost prijelaza iz stanja 5 u stanje 4 bilo jednako 0, tj. $p_{54} = 0$ stanje 5 bi bilo apsorbirajuće stanje.

## 3.4 Stacionarna distribucija

Stacionarnost kod slučajnih procesa, grubo rečeno predstavlja činjenicu da se vjerojatnosna svojstva procesa ne mijenjaju kroz vrijeme.

**Definicija 3.11.** Slučajni proces $X = (X_n : n \geq 0)$ definiran na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$ zove se *stacionaran* ako za sve $k \geq 0$ i za sve $n \geq 0$ slučajni vektori $(X_0, X_1, \ldots, X_k)$ i $(X_n, X_{n+1}, \ldots, X_{n+k})$ imaju istu distribuciju (u odnosu na vjerojatnost $\mathbb{P}$).

**Definicija 3.12.** Neka je $X = (X_n : n \geq 0)$ Markovljev lanac s skupom stanja $S$ i prijelaznom matricom $P$. Vjerojatnosna distribucija $\pi = (\pi_i : i \in S)$ na $S$ je *stacionarna distribucija* Markovljevog lanca $X = (X_n : n \geq 0)$ ako vrijedi

$$\pi = \pi P,$$

odnosno po komponentama

$$\pi_j = \sum_{k \in S} \pi_k p_{kj}, \text{ za sve } j \in S.$$

Sljedeća propozicija daje kriterij stacionarnosti za konačne skupove stanja. (postoji i općenitiji rezultat, ali ga u radu ne navodimo).

**Propozicija 3.3.** Neka je $S$ konačan skup stanja, te pretpostavimo da za neki $i \in S$,

$$\lim_{n \to \infty} p_{ij}^{(n)} = \pi_j \text{ za sve } j \in S.$$

Tada je $\pi = (\pi_j : j \in S)$ stacionarna distribucija.

*Dokaz.* Vrijedi

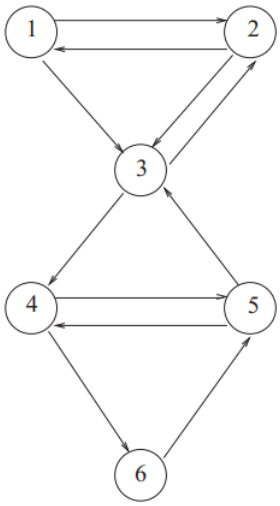
$$\sum_{j \in S} \pi_j = \sum_{j \in S} \lim_{n \to \infty} p_{ij}^{(n)} = \lim_{n \to \infty} \sum_{j \in S} p_{ij}^{(n)} = 1,$$

pa je $\pi$ vjerojatnosna distribucija. Zamjena limesa i sume je opravdana jer je $S$ konačan. Nadalje,

$$\pi_j = \lim_{n \to \infty} p_{ij}^{(n)} = \lim_{n \to \infty} p_{ij}^{(n+1)} = \lim_{n \to \infty} \sum_{k \in S} p_{ik}^{(n)} p_{kj} = \sum_{k \in S} \pi_k p_{kj}.$$

Dakle, $\pi$ je stacionarna distribucija. □

**Primjer 3.6.** Jedna od primjena gdje je prirodno koristiti Markovljev lanac je za pretraživanje weba. *PageRank* algoritam kojeg su osmislili Sergey Brin i Larry Page koristi se za rankiranje internetskih stranica po važnosti samo na temelju poveznica (linkova). Važnost stranice mjerena je vjerojatnošću da će osoba slučajnim klikanjem poveznica stići do određene stranice. PageRank je zapravo stacionarna distribucija Markovljevog lanca koji za skup stanja ima stranice, a broj poveznica na drugu stranicu određuje vjerojatnost prijelaza na tu stranicu. Uzmimo za primjer jedan pojednostavljeni Web sa šest stranica i poveznicama kao na slici 4




Slika 4: Markovljev lanac za Web sa šest stranica preuzeto iz [2]

Matrica prijelaza je sljedeća:

$$\mathrm{P}=\begin{bmatrix} 0 & 0.5 & 0.5 & 0 & 0 & 0 \\ 0.5 & 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

Kako stranica 1 ima dvije poveznice: jednu koja vodi na stranicu 2 i drugu koja vodi na stranicu 3, slučajnim klikanjem prelazimo na stranicu 2, odnosno 3, s vjerojatnošću 0.5. Slično vrijedi za ostale stranice. Kada bismo imali zatvoreni skup stanja, PageRank algoritam bi ocijenio važnost stranica koje nisu u zatvorenom skupu s 0. Naime, jednom kada bi osoba slučajnim klikanjem ušla u stranicu koja

se nalazi u zatvorenom skupu, više ne bi mogla izaći. Stoga uvodimo *teleportaciju* s vjerojatnošću $1-\alpha$. Pod teleportacijom mislimo da osoba koja pretražuje web u jednom trenutku neće pratiti poveznicu nego će se naći na nekoj drugoj stranici (koja nije direktno povezana s trenutnom stranicom). Uzimajući u obzir teleportaciju imamo:

$$G = \alpha P + (1-\alpha)u, 1 \leq \alpha \leq 1, \tag{8}$$

gdje je $u$ vektor duljine $n$ s vrijednostima $\frac{1}{n}$. U našem slučaju vektor $u$ je duljine 6 s vrijednostima $\frac{1}{6}$. Neka je $\alpha = 0.85$. Izračunamo novu matricu prijelaza na temelju 8:

$$G = \begin{bmatrix} 0.025 & 0.45 & 0.45 & 0.025 & 0.025 & 0.025 \\ 0.45 & 0.025 & 0.45 & 0.025 & 0.025 & 0.025 \\ 0.025 & 0.45 & 0.025 & 0.45 & 0.025 & 0.025 \\ 0.025 & 0.025 & 0.025 & 0.025 & 0.45 & 0.45 \\ 0.025 & 0.025 & 0.45 & 0.45 & 0.025 & 0.025 \\ 0.025 & 0.025 & 0.025 & 0.025 & 0.875 & 0.025 \end{bmatrix}.$$

Sada nas zanima postoji li distribucija $\pi = [\pi_1, \pi_2, \pi_3, \pi_4, \pi_5, \pi_6]$ tako da se u bilo kojem ne mijenjaju vjerojatnosti u $\pi$.

Prema definiciji stacionarne distribucije mora vrijediti.

$$\pi = \pi P \text{ i } \sum \pi = 1.$$

Rješavanjem dobijemo:

$$\pi = (0.092, 0.158, 0.220, 0.208, 0.209, 0.113).$$

Dakle, stranice bi bile rangirane od najbitnije do nebitne $\{3, 5, 4, 2, 6, 1\}$.

# 4 Skriveni Markovljevi modeli

Skriveni Markovljev model je Markovljev proces koji je podijeljen u dvije komponente: vidljivu komponentu i nevidljivu, tj. "skrivenu" komponentu. Točnije, skriveni Markovljev model je Markovljev proces $(X_k, Y_k)_{k\geq 0}$ na prostoru stanja $S \times V$. Pretpostavljamo da imamo način kojim bismo promatrali $Y_k$, ali ne i $X_k$. Vrijednosti koje može poprimiti $Y_k$ ćemo zvati *opažaji* i skup $V$ je *skup opažaja*. Vrijednosti koje može poprimiti $X_k$ ćemo zvati *stanja*, a skup $S$ je *skup stanja*. U posljednjih nekoliko desetljeća skriveni Markovljevi modeli su se primjenjivali za modeliranje u *automatskom prepoznavanju govora*[4] (vidjeti npr. [4]). Danas, skrivene Markovljeve modele možemo pronaći u raznim područjima kao što su prepoznavanje uzoraka, obrada signala, telekomunikacija, bioinformatika i slično. S jedne strane, skriveni Markovljevi lanci prirodno opisuju situacije u kojima je stohastički proces promatran kroz mjerenja s određenom greškom. Na primjer u teoriji komunikacije, $X_k$ možemo promatrati kao slučajni signal koji je poslan komunikacijskim kanalom. Kako u komunikacijskom kanalu dolazi do greški, $Y_k$ je oštećena verzija izvornog signala. Onaj tko je primio poruku htio bi rekonstruirati primljeni signal što bolje. S druge strane, $Y_k$ može biti proces koji nas zanima, a $X_k$ je skriveni proces koji utječe na $Y_k$. Na primjer, može nas zanimati cijena dionice $Y_k$, gdje je $X_k$ skriveni ekonomski faktor koji utječe na kretanje cijene dionice.

Iako je $(X_k, Y_k)_{k>0}$ Markovljev proces, opažajna komponenta $(Y_k), k > 0$ neće biti Markovljev lanac. Skrivenim Markovljevim lancima možemo modelirati procese koje nemaju Markovljevo svojstvo.

## 4.1 Osnovni elementi modela

U ovom modelu, opažanje $Y_t$ u trenutku $t$ (koji može označavat vrijeme ili poziciju u nizu opažaja) je posljedica stohastičkog procesa, ali stanje $X_t$ tog procesa ne može se direktno promatrati. Pretpostavljamo da taj skriveni proces zadovoljava Markovljevo svojstvo.

Kako bismo u potpunosti definirali skirvene Markovljeve lance, potrebno je definirati sljedećih pet elemenata:

[4]Prepoznavanje govora ili govor-u-tekst sposobnost je stroja ili programa da identificira riječi izgovorene naglas i pretvori ih u čitljiv tekst.

1. *Broj različitih skrivenih stanja Markovljevog lanca* N. Skup stanja ćemo označiti sa $S = \{s_1, s_2, \ldots, s_N\}$, a stanje u trenutku $t$ reprezentirat ćemo slučajnom varijablom $X_t$.

2. *Broj različitih opažaja* M. Skup opažaja ćemo označitit s $V = \{v_1, v_2, \ldots, v_M\}$. Opažaj u trenutku $t$ može poprimiti bilo koju vrijednost iz skupa $V$ te ćemo ga reprezentirati slučajnom varijablom $Y_t$.

3. *Matrica prijelaza* $P$ koja je $N \times N$ matrica s elementima $p_{ij}$:

$$p_{ij} = \mathbb{P}(X_t = s_j | X_{t-1} = s_i), 1 \leq i, j \leq N.$$

   Dakle, $p_{ij}$ je vjerojatnost prijelaza iz stanja $s_i$ u stanje $s_j$.

4. *Emisijske vjerojatnosti* koje zapisujemo u emisijsku matricu $\Psi$ dimenzija $M \times N$. Elementi matrice su

$$\psi_{ij} = \mathbb{P}(Y_t = v_i | X_t = s_j), 1 \leq i \leq M, 1 \leq j \leq N.$$

   Dakle, $\psi_{ij}$ predstavlja vjerojatnost da se u trenutku $t$ nalazimo u stanju $s_j$ i da se emitiralo opažanje $v_i$.

5. *Distribucija početnog stanja*, $\Pi$ je vektor duljine N s elementima

$$\pi_i = \mathbb{P}(X_1 = s_i), 1 \leq i \leq N.$$

   Dakle, $\pi_i$ predstavlja vjerojatnost da se u početnom trenutku nalazimo u stanju $s_i$.

Skrivene Markovljeve modele određene s gore opisanim parametrima ćemo kraće označavati sa:

$$\lambda(P, \Psi, \Pi). \tag{9}$$

**Primjer 4.1.** Pretpostavimo da netko baca dva novčića. Jedan je pošten (označit ćemo ga s 0), tj. jednaka je vjerojatnost da će pasti pismo ili glava. Drugi novčić je pristran (označit ćemo ga s 1) i veća je vjerojatnost da će pasti glava. Nama su poznati samo rezultati bacanja, a ne koji je novčić bačen niti koja je vjerojatnost da vidimo glavu u stanju 1:

P P G P G P G G P G G G G G G

P G G G P P G P G P G P G P P

G G G G G G P G P G G P P G G

Pretpostavljamo da se u pozadini opažanja skriva Markovljev lanac. Skup svih stanja su nam poznata, tj. znamo da se radi o dva novčića od kojih je jedan pošten, a drugi pristran, ali ne znamo vjerojatnosti prijelaza niti emisijske vjerojatnosti. Potrebne vjerojatnosti možemo približno odrediti na sljedeći način. Radi ilustracije osnovnih pojmova, pretpostavimo da je na početku radi o poštenom novčiću i to označimo s 0. Nakon toga ćemo pretpostaviti da se radi o pristranom novčiću. To označimo na sljedeći način:

P P G P G P G G P G G G G G G

0 0 0 0 0 0 0 0 0 1 1 1 1 1 1

P G G G P P G P G P G P G P P

1 1 1 1 1 0 0 0 0 0 0 0 0 0 0

G G G G P G P G P G G P P G G

1 1 1 1 1 1 1 1 1 1 1 0 0 0 0

Nakon toga procijenimo kolika je vjerojatnost da sustav mijenja stanje. Prebrojimo koliko puta smo označili da je bačen pošten novčić i da je nakon toga opet bačen pošten i vidimo da je od 22 puta 20 puta bio bačen pošten novčić, a dva puta je sustav prešao na pristran novčić. Isto napravimo za pristran novčić. Nadalje, kod pristranog novčića u 22 bacanja 17 puta je pala glava, a pismo 5 puta. Na temelju ovih procjena možemo napraviti skriveni Markovljev model. Matrica prijelaza tog lanca izgleda ovako:

$$P = \begin{bmatrix} 0.91 & 0.09 \\ 0.09 & 0.91 \end{bmatrix}$$

gdje je $p_{11}$ vjerojatnost da je sustav u stanju 0 ako je prethodno stanje bilo 0, $p_{12}$ je vjerojatnost da je sustav u stanju 1 ako je prethodno bio u stanju 0 itd. Nadalje,

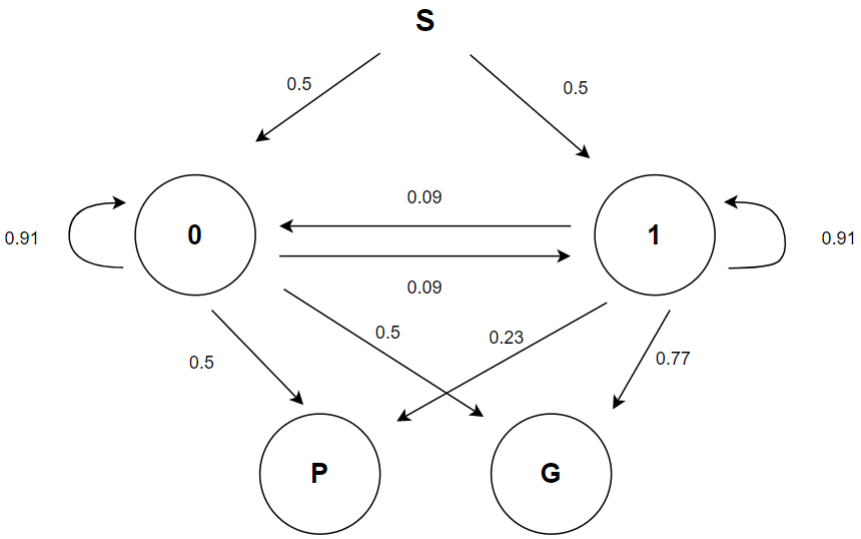
$\psi_{11}$ vjerojatnost da je palo pismo ako smo u stanju 0, $\psi_{12}$ je vjerojatnost da je pala glava ako smo u stanju 0 itd. Emisijska matrica je:

$$\Psi = \begin{bmatrix} 0.5 & 0.5 \\ 0.23 & 0.77 \end{bmatrix}.$$

Za kraj treba odrediti još početne vjerojatnosti. Pretpostavimo da se na početku baca pošteni novčić i ako je pala glava počinjemo s pristranim novčićem, a ako je palo pismo, počinjemo s poštenim novčićem

$$\pi_1 = 0.5, \text{ vjerojatnost da sustav započinje u stanju 0,}$$
$$\pi_2 = 0.5, \text{ vjerojatnost da sustav započinje u stanju 1.}$$

Primjer je prikazan grafički na slici:




Slika 5: Dijagram skrivenog Markovljevog lanca

Ovim primjerom smo procijenili parametre skrivenog Markovljevog modela. U sljedećem poglavlju ćemo, točnije kod problema 3, dati metodu kojom se može optimizirati odabir parametara modela.

## 4.2 Osnovni problemi skrivenih Markovljevih lanaca

Kako bi skriveni Markovljevi modeli bili primjenjivi na probleme iz stvarnog svijeta, potrebno je riješiti sljedeća tri problema.

1. *Problem evaluacije* - za dani niz opažanja $Y_1 = v_1, \cdots, Y_T = v_T$ i model $\lambda = (P, \Psi, \Pi)$, odrediti $\mathbb{P}(V|\lambda)$, odnosno, preciznije efikasno odrediti vjerojatnost da smo dobili takav niz opažanja uz dane parametre modela.

2. *Problem dekodiranja* - želimo odrediti optimalan niz skrivenih stanja za dani niz opažajnih stanja.

3. *Problem učenja* - odrediti model kojim bismo maksimizirali vjerojatnost emitiranja danog niza simbola.

### 4.2.1 Problem evaluacije

Kod ovog problema, zanima nas vjerojatnost da se dobije niz opažanja $V = v_1 v_2 \dots v_T$ za dani model $\lambda$, tj. $\mathbb{P}(Y_1 = v_1, Y_2 = v_2, \dots, Y_T = v_T|\lambda)$ što ćemo kraće pisati kao $\mathbb{P}(V|\lambda)$. Najjednostavniji način je računati vjerojatnost za sve moguće nizove stanja duljine $T$. Fiksirajmo jedan takav niz stanja:

$$S = s_1, s_2, \dots, s_T,$$

gdje je $s_1$ početno stanje. Vjerojatnost da vidimo niz opažaja $V$ za niz stanja $S$, uz pretpostavljenu nezavisnost varijabli opažaja, jednaka je

$$\mathbb{P}(V|S, \lambda) = \prod_{t=1}^{T} \mathbb{P}_\lambda(Y_t = v_t|S_t = s_t) = \prod_{t=1}^{T} \psi_{v_t, s_t}. \tag{10}$$

Vjerojatnost da je nastao niz $S$ možemo zapisati kao:

$$\mathbb{P}_\lambda(S) = \pi_{s_1} p_{12} \dots p_{T-1,T}. \tag{11}$$

Vjerojatnost da se $V$ i $S$ dogode istovremeno je umnožak vjerojatnosti (7) i (8), tj.

$$\mathbb{P}_\lambda(V, S) = \mathbb{P}(V|S, \lambda)\mathbb{P}_\lambda(S). \tag{12}$$

Vjerojatnost niza $V$ uz dani model dobijemo tako da zbrojimo prethodnu jednakost po svim mogućim nizovima stanja duljine $T$. Dakle, imamo:

$$\begin{aligned}\mathbb{P}(V|S,\lambda) &= \sum_{s_1,s_2,\ldots,s_T\in S}\prod_{t=1}^{T}\psi_{v_t,s_t}\cdot\pi_{s_1}\prod_{t=2}^{T}p_{t-1,t}\cdot \\ &= \sum_{s_1,s_2,\ldots,s_T\in S}\pi_{s_1}\cdot\psi_{v_1,s_1}\cdot p_{12}\cdot\psi_{v_2,s_2}\cdot\cdots\cdot p_{T-1,T}\cdot\psi_{v_T,s_T}. \qquad (13)\end{aligned}$$

U trenutku $t = 1$ nalazimo se u stanju $s_1$ s vjerojatnosti $\pi_{s_1}$, te se emitirao simbol $v_1$ s vjerojatnosti $\psi_{v_1,s_1}$. Nakon toga prelazimo u stanje $s_2$ s vjerojatnosti prijelaza $p_{12}$ i emitirao se simbol $v_2$ s vjerojatnosti $\psi_{v_2,s_2}$. Postupak se ponavlja do zadnjeg prijelaza $p_{T-1,T}$ i zadnjeg emisijskog simbola $v_T$. Rješenje ovog problema je relativno jednostavno, no zbog previše računskih operacija nije efikasno. Direktni izračun za $P_\lambda(V|S)$ kao u (13) zahtjeva broj računskih operacija reda $2T\cdot N^T$. U svakom trenutku $t = 1,\cdots,T$ imamo $N$ mogućih stanja i za svaki takav niz stanja imamo $2T$ operacija.

S obzirom da je prethodni račun operacijski zahtjevan, uvodimo algoritma kojim bismo smanjili potreban broj računskih operacija kako bi računanje vjerojatnosti da je nastao niz opažanja uz dani model bilo efikasnije. *Forward-backward* je algoritam dinamičkog programiranja kojeg možemo koristiti u klasificiranju nizova, detekciji anomalija itd. Najprije definiramo *forward varijablu*:

**Definicija 4.1.** Neka su dana opažanja $v_1, v_2, \ldots, v_T \in V$. Definiramo

$$\alpha_t(s_j) := \mathbb{P}(Y_1 = v_1, Y_2 = v_2, \ldots, Y_t = v_t, X_t = s_j|\lambda), \text{ za svaki } t \le T, s_j \in S.$$

Forward varijabla predstavlja vjerojatnost da se model u trenutku $t$ nalazi u stanju $s_j$ i da je model emitirao prvih $t$ elemenata opažajnog niza.

*Forward algoritam* opisujemo sljedećim koracima:

1. Početno stanje:

$$\alpha_1(j) = \pi_{s_j}\psi_{s_jv_t},\ s_j \in S.$$

2. Rekurzija:

$$\alpha_{t+1}(s_j) = \psi_{s_jv_{t+1}}\sum_{s_i\in S}\alpha_t(s_i)p_{ij},\ \ s_j \in S, 1 \le t \le T-1$$

3. Kraj:

$$\mathbb{P}(Y_1 = v_1, Y_2 = v_2, \ldots, Y_T = v_T|\lambda) = \sum_{i=1}^{N} \alpha_T(s_i).$$

Prvi korak je forward varijabla za duljinu $t = 1$. Dakle, to je vjerojatnost da je početno stanje $s_j$ i da se emitirao prvi simbol $v_1$. Kako je $\alpha_t(s_i)$ vjerojatnost da se model u trenutku $t$ nalazi u stanju $s_i$ i da je emitirao opažajni niz $v_1 v_2 \ldots v_t$, produkt $\alpha_t(s_i)p_{ij}$ je onda vjerojatnost da se model u trenutku $t+1$ nalazi u stanju $s_j$ i da se emitirao niz opažanja $v_1 v_2 \ldots v_t$. Sumiranjem produkta po svim mogućim stanjima $s_i, 1 \leq i \leq N$ u trenutku $t$ rezultira vjerojatnosti da se model nalazi u stanju $s_j$ u trenutku $t+1$. Kada smo to napravili i $s_j$ je poznat, lako vidimo da se $\alpha_{t+1}(s_j)$ dobije množenjem sa vjerojatnošću da je stanje $s_j$ emitiralo sljedeće opažanje, $v_{t+1}$. Na kraju korak 3 daje traženu vjerojatnost jer po definiciji $\alpha_t(s_j)$ slijedi

$$\alpha_T(s_j) = \mathbb{P}(Y_1 = v_1, \ldots, Y_T = v_t, X_T = s_j),$$

pa se sumiranjem po svim stanjima $s_j \in S$ u kojima $X_T$ može biti dobije vjerojatnost da se dogodio opažajni niz uz dani model. Opisani forward algoritam je složenosti reda $TN^2$, što je puno manje nego direktno računanje.

**Primjer 4.2.** Dan je opažajni niz $V = \{\text{glava}, \text{glava}, \text{pismo}\}$ i skup stanja $S = \{0, 1\}$. Matrice koje su potrebne kako bismo izračunali forward varijablu za $t = 3$ su matrica prijelaza, emisijska matrica i početna distribucija.

$$\text{P}= \begin{bmatrix} 0.91 & 0.09 \\ 0.09 & 0.91 \end{bmatrix}, \ \Psi = \begin{bmatrix} 0.5 & 0.5 \\ 0.23 & 0.77 \end{bmatrix}, \ \Pi= \begin{bmatrix} 0.5 & 0.5 \end{bmatrix}$$

Najprije računamo za $t = 1$:

$$\alpha_1(0) = \pi_1 \psi_{1g} = 0.5 \cdot 0.5 = 0.25$$
$$\alpha_1(1) = \pi_2 \psi_{2g} = 0.5 \cdot 0.77 = 0.385,$$

zatim za $t = 2$:

$$\alpha_2(0) = \psi_{1g}(\alpha_1(0)p_{11} + \alpha_1(1)p_{12}) = 0.5(0.25 \cdot 0.91 + 0.385 \cdot 0.09) = 0.131$$
$$\alpha_2(1) = 0.77(0.25 \cdot 0.09 + 0.385 \cdot 0.91) = 0.287,$$

te na kraju za $t = 3$:

$$\alpha_3(0) = 0.5(0.131 \cdot 0.91 + 0.287 \cdot 0.09) = 0.073$$
$$\alpha_3(1) = 0.23(0.131 \cdot 0.09 + 0.287 \cdot 0.91) = 0.063.$$

Na kraju treba zbrojiti izračunate $\alpha_3$:

$$\mathbb{P}(Y_1 = g, Y_2 = g, Y_3 = p) = \alpha_3(0) + \alpha_3(1) = 0.073 + 0.063 = 0.136.$$

Za rješenje problema evaluacije dovoljna je forward varijabla, ali sada ćemo uvesti backward algoritam koji će nam biti potreban kod rješavanja problema 3. Na sličan način možemo promatrati backward varijablu.

**Definicija 4.2.** Neka su dana opažanja $v_1, v_2, \ldots, v_T \in V$ i neka je

$$\beta_t(s_i) := \mathbb{P}(Y_{t+1} = v_{t+1}, Y_{t+2} = v_{t+2}, \ldots, Y_T = v_T | X_t = s_i), \text{ za svaki } t \leq T, s_i \in S.$$

Varijablu $\beta_t(s_i)$ zovemo *backward varijabla* i ona predstavlja vjerojatnost da je model emitirao zadnjih $T - t$ elemenata opažajnog niza uz uvjet da se u trenutku $t$ lanac nalazi u stanju $s_i$.

Backward algoritam možemo opisujemo sljedećim koracima:

1. Početni uvjet:

$$\beta_T(s_i) = 1,\ 1 \leq i \leq N.$$

2. Rekurzija:

$$\beta_t(s_i) = \sum_{j=1}^{N} \beta_{t+1}(s_j) p_{ij} \psi_{s_j, v_{t+1}},\ \ s_i \in S,\ 1 \leq t \leq T - 1$$

3. Kraj:

$$\mathbb{P}(Y_1 = v_1, Y_2 = v_2, \ldots, Y_T = v_T) = \sum_{j=1}^{\mathcal{N}} \pi_{s_j} \beta_1(s_j) \psi_{s_j, v_1}$$

Kako bismo došli do stanja $s_i$ u trenutku $t$, uzimajući u obzi opažajni niz od trenutka $t + 1$ na dalje, moramo promatrati sva moguća stanja $s_j$ u trenutku $t + 1$. U obzir uzimamo vjerojatnost prijelaza iz stanja $s_i$ u stanje $s_j$ ($p_{ij}$) kao i vjerojatnost da se

u stanju $s_j$ u trenutku $t+1$ emitirao opažaj $v_{t+1}$ ($\psi_{s_j,v_t+1}$). Sumiranjem po svim stanjima $s_j \in S$ dobije se tražena $\beta_t(s_i)$. Kao i kod forward varijable, računanje backward varijable složenosti je reda $TN^2$.

### 4.2.2 Problem dekodiranja

Za razliku od problema evaluacije, problem dekodiranja, tj. pronalazak optimalnog niza stanja za dani opažajni niz, možemo riješiti na više načina. Postoje različiti kriteriji optimalnosti, npr. jedan od njih bi bio da za svaki emisijski simbol izaberemo stanje koja ima najveću vjerojatnost emitiranja tog simbola. Definirajmo varijablu

$$\gamma_t(s_i) = \mathbb{P}_\lambda(X_t = s_i|V),$$

tj. vjerojatnost da smo u trenutku $t$ u stanju $s_i$ uz dani niz opažaja i model $\lambda$. Varijabli $\gamma$ možemo izraziti pomoću forward-backward varijabli

$$\gamma_t(s_i) = \frac{\alpha_t(s_i)\beta_t(s_i)}{\mathbb{P}_\lambda(V)} = \frac{\alpha_t(s_i)\beta_t(s_i)}{\sum_{i=1}^{N}\alpha_t(s_i)\beta_t(s_i)},$$

kako $\alpha_t(s_i)$ odgovara opažajnom nizu $v_1v_2\ldots v_t$ i stanju $s_i$ u trenutku t, dok $\beta_t(s_i)$ odgovara ostatku opažajnog niza $v_{t+1}v_{t+2}\ldots v_T$ uz stanje $s_i$ u trenutku t.
Faktor $\sum_{i=1}^{N}\alpha_t(s_i)\beta_t(s_i)$ normalizira $\gamma_t(s_i)$ kako bi vrijedilo:

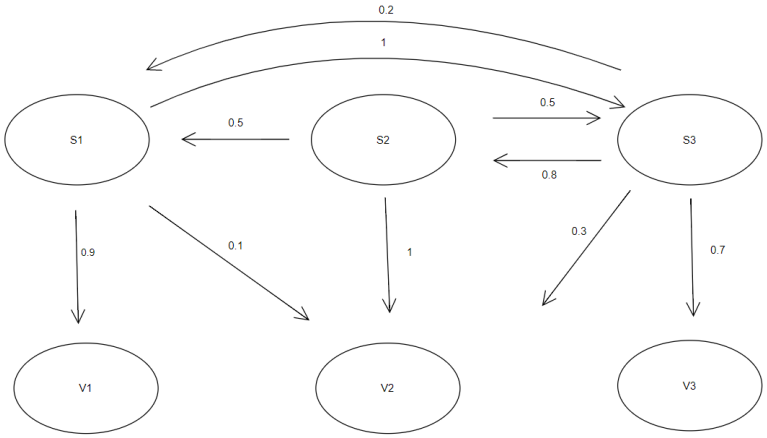
$$\sum_{i=1}^{N}\gamma_t(s_i) = 1,$$

tj. kako bi $\gamma_t(s_i)$ bila vjerojatnosna mjera. Koristeći $\gamma_t(s_i)$ možemo izračunati individualno najbolje stanje za emisijski simbol u trenutku $t$

$$X_t = \max_{1\leq i\leq N}[\gamma_t(s_i)]. \tag{14}$$

Iako gore navedeni postupak maksimizira očekivani broj točnih stanja (tako što bira najvjerojatnije stanje za svaki $t$), ne mora značiti da je tako odabrani niz moguć.

**Primjer 4.3.** Pretpostavimo da imamo skirveni Markovljev model sa skupom stanja $S = \{S_1, S_2, S_3\}$ i skupom emisijskih simbola $V = \{V_1, V_2, V_3\}$. Vjerojatnosti prijelaza kao i emisijske vjerojatnosti su definirane na slici:




Pretpostavimo da se emitirao niz $V_1V_2V_3$. Lako se vidi, kada bismo gledali vjerojatnosti da je pojedino stanje emitiralo pojedini simbol, da je niz stanja $S_1S_2S_3$. No, kada se nalazimo u stanju $S_1$ sa vjerojatnošću 1 prelazimo u stanje $S_3$, tj. $p_{12} = 0$. Dakle, nije moguće da imamo takav niz stanja.

Moguće rješenje nastalog problema je da modificiramo kriterij optimalnosti. Mogli bismo maksimizirati očekivani broj točnih parova stanja $(X_t, X_{t+1})$ ili trojke $(X_t, X_{t+1}, X_{t+2})$ itd.

Viterbijev[5] algoritam je algoritam dinamičkog programiranja kojim tražimo najvjerojatniji niz skriveni stanja koji odgovara nizu opažaja. Najvjerojatniji niz zovemo Viterbijev put. Viterbijev algoritam je u početku služio za ispravljanje grešaka nastalih u komunikacijskom kanalu, a danas se često koristi u automatskom prepoznavanju govora, obradi jezika te bioinformatici. Želimo pronaći najbolji niz stanja, $Q = \{X_1, X_2, \ldots, X_T\}$ za dani niz opažaja $V = v_1v_2 \ldots v_T$. Najprije moramo definirati:

$$\delta_t(s_i) = \max_{X_1, X_2, \ldots, X_{t-1}} \mathbb{P}(X_1, X_2, \ldots, X_t = s_i, Y_1 = v_1, Y_2 = v_2, \ldots, Y_t = v_t).$$

Dakle, definirali smo najveću vjerojatnost duž jednog niza stanja u trenutku $t$, koja se računa za prvih $t$ opažanja. Indukcijom dobivamo:

$$\delta_{t+1}(s_j) = \max_i [\delta_t(s_i)p_{ij}]\dot{\psi}_{s_j, v_k} \tag{15}$$

[5]Andrew James Viterbi (1935.) je američki inženjer elektrotehnike i biznismen koji je suosnivač Qualcomm Inc.

Kako bi dobili traženi niz stanja moramo pratiti argumente koji su maksimizirali (15), za svaki $t$ i za svaki $j$. To zapisujemo u niz $\phi_t(s_j)$.

Sada Viterbijev algoritam možemo opisati koracima:

1. Početno stanje:

$$\delta_1(s_i) = \pi_i \psi_{s_i,v_1},\ 1 \leq i \leq \mathcal{N}$$
$$\phi_i(s_i) = 0.$$

2. Rekurzija:

$$\delta_t(s_j) = \max_i [\delta_{t-1}(s_i)p_{ij}]\dot{\psi}_{s_j,v_t}$$
$$\phi_t(s_j) = arg \max_{1\leq i\leq \mathcal{N}} [\delta_{t-1}(s_i)p_{ij}],$$

   za sve $2 \leq t \leq T$, $1 \leq i \leq \mathcal{N}$

3. Kraj:

$$P^* = \max_{1\leq i\leq \mathcal{N}} [\delta_T(s_i)]$$
$$X_T^* = arg \max_{1\leq i\leq \mathcal{N}} [\delta_T(s_i)].$$

4. Put, tj. niz stanja;

$$X_t^* = \phi_{t+1}(X_{t+1}^*),\ t = T-1, T-2, \ldots, 1.$$

Prethodni algoritam objasnit ćemo na sljedećem primjeru koristeći programski jezik Python.

**Primjer 4.4.** Jedna od primjena Skrivenih Markovljevih lanaca je u označavanju teksta (engl. part of speech tagging, POS). Odnosi se na kategorizaciju riječi u tekstu, što bi značilo da nas zanima vrsta riječ koja se javlja u tekstu, tj. je li riječ koju vidimo imenica, glagol, pridjev itd. U ovom primjeru riječi su opažajni simboli, a vrsta riječi je skriveno stanje. vjerojatnosti prijelaza i emisijske vjerojatnosti možemo procijeniti pomoću već označenih rečenica. Označene rečenice preuzimamo u Python-u iz paketa "nltk" pomoću naredbi: `nltk.download('treebank')`

i `nltk.download('universal_tagset')` iz paketa nltk, kao što je prikazano na slici:

| | NOUN | VERB | ADP | . | DET | PRT | PRON | X | ADJ | CONJ | ADV | NUM |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **NOUN** | 0.263668 | 0.147912 | 0.177451 | 0.240150 | 0.012828 | 0.044548 | 0.004887 | 0.028928 | 0.011301 | 0.042105 | 0.016929 | 0.009294 |
| **VERB** | 0.111922 | 0.167883 | 0.091144 | 0.033983 | 0.130668 | 0.031582 | 0.034906 | 0.221073 | 0.065288 | 0.005356 | 0.083387 | 0.022809 |
| **ADP** | 0.325876 | 0.008566 | 0.017515 | 0.041549 | 0.321529 | 0.001278 | 0.068908 | 0.034646 | 0.103938 | 0.001023 | 0.012401 | 0.062772 |
| **.** | 0.224214 | 0.088454 | 0.088454 | 0.092991 | 0.173129 | 0.001836 | 0.066422 | 0.028189 | 0.041797 | 0.058970 | 0.054110 | 0.081326 |
| **DET** | 0.639354 | 0.038713 | 0.008732 | 0.016883 | 0.005967 | 0.000291 | 0.003493 | 0.044098 | 0.205356 | 0.000582 | 0.012953 | 0.023577 |
| **PRT** | 0.245750 | 0.403787 | 0.021638 | 0.043663 | 0.100850 | 0.002318 | 0.017774 | 0.014683 | 0.084621 | 0.001932 | 0.009660 | 0.053323 |
| **PRON** | 0.211635 | 0.487861 | 0.021072 | 0.038479 | 0.010078 | 0.011910 | 0.006413 | 0.090701 | 0.074668 | 0.005955 | 0.033898 | 0.007329 |
| **X** | 0.057656 | 0.200378 | 0.148393 | 0.164083 | 0.056333 | 0.186011 | 0.056333 | 0.073913 | 0.015690 | 0.009830 | 0.027977 | 0.003403 |
| **ADJ** | 0.704327 | 0.012620 | 0.075120 | 0.062300 | 0.003806 | 0.011418 | 0.000401 | 0.020433 | 0.067308 | 0.016426 | 0.004607 | 0.021234 |
| **CONJ** | 0.352254 | 0.159711 | 0.053422 | 0.034502 | 0.121870 | 0.005565 | 0.060100 | 0.007234 | 0.106288 | 0.000556 | 0.059544 | 0.038954 |
| **ADV** | 0.030763 | 0.350857 | 0.116433 | 0.130062 | 0.068146 | 0.014798 | 0.016745 | 0.024143 | 0.129673 | 0.007399 | 0.079439 | 0.031542 |
| **NUM** | 0.360597 | 0.017070 | 0.033784 | 0.111309 | 0.003201 | 0.026671 | 0.001422 | 0.211238 | 0.032006 | 0.015647 | 0.003201 | 0.183855 |

Slika 6: Matrica prijelaza Primjer 4.4

Nadalje, početne vjerojatnosti prikazane su na slici:

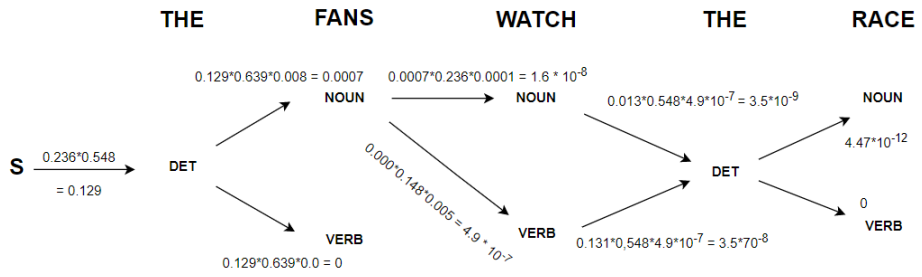
| | 0 |
|---|---|
| **NOUN** | 0.293949 |
| **VERB** | 0.009236 |
| **ADP** | 0.124522 |
| **.** | 0.082166 |
| **DET** | 0.235669 |
| **PRT** | 0.000955 |
| **PRON** | 0.072293 |
| **X** | 0.024204 |
| **ADJ** | 0.039490 |
| **CONJ** | 0.051274 |
| **ADV** | 0.053822 |
| **NUM** | 0.009554 |

Slika 7: Distribucija početnog stanja za Primjer 4.4

Kako je u paketu 10000 riječi, nećemo ubaciti emisijsku matricu. Promotrimo rečenicu: "The fans watch the race.". Radi jednostavnosti pretpostavimo da je u ovom primjeru riječ the "DET", a *fans*,*watch* i *race* mogu biti ili "NOUN" ili "VERB". Dakle, cilj je

$$\max_{P_1,P_2,P_3,P_4,P_5} \mathbb{P}(P_1, P_2, P_3, P_4, P_5, \text{the,fans,watch,the,race})$$

gdje su $P_i$, $i = 1, \ldots, 5$ skrivena stanja. Problem ćemo riješiti pomoću Viterbijevog algoritma.




Slika 8: Grafički prikaz Vitarbijevog algoritma

Tražimo put s najvećom vjerojatnosti, tj. Viterbijev put. Vrijednost s prethodne strelice množimo s vjerojatnosti prijelaza i emisijskom vjerojatnosti. Dakle, množimo vjerojatnost da smo s imenice prešli na imenicu i, ako smo u stanju imenice, vjerojatnost da vidimo riječ watch. Ono što čini algoritam efikasnim je to što kada se dva različita puta sretnu u istom čvoru možemo zanemariti put koji je imao manju vjerojatnost. U ovom primjeru algoritam je dekodirao riječ *the* kao "DET", *fans* kao "NOUN", *watch* kao "VERB" i *race* kao "NOUN".

### 4.2.3 Problem učenja

Treći i najteži problem je problem učenja, tj. kako odrediti metodu kojom bismo podesili parametre skrivenog Markovljevog modela da bismo maksimizirali vjerojatnost opažajnog niza za dani model. Dakle, potrebno je iz dobivenih podataka i početnog modela odrediti optimalne verzije prijelazne matrice $P$, emisijske matrice $\psi$ i početne distribucije $\Pi$. Zapravo, ako imamo bilo kakav niz opažajnih simbola koji nam služi za učenje, ne postoji optimalan način za određivanje parametara modela. Međutim, koristeći Baum[6]-Welch[7] algoritam možemo izabrati $\lambda(P, \psi, \Pi)$ tako da je vjerojatnost $P(V)$ opažajnog niza lokalno najveća. Kako bismo opisali algoritam za učenje potrebno je definirati vjerojatnost da se nalazimo u stanju $s_i$ u trenutku $t$ i da se nalazimo u stanju $s_j$ u trenutku $t+1$ uz dani model i opažajni niz:

$$\xi_t(s_i, s_j) = \mathbb{P}(X_t = s_i, X_{t+1} = s_j | V).$$

[6]Leonard Esau Baum (1931-2017.) Američki matematičar

[7]Lloyd Richard Welch (1927.) Američki matematičar, također poznat po Berlekamp–Welch algoritmu.

Prisjetimo se, $\alpha_t(s_i)$ je vjerojatnost da je emitirano prvih $t$ simbola i da se nalazimo u stanju $s_i$, dok je $\beta_{t+1}(s_j)$ vjerojatnost da se u trenutku $t+1$ nalazimo u stanju $s_j$ i da je emitirano zadnjih $T-t-1$ simbola. Dakle, možemo zapisati

$$\begin{aligned}\xi_t(s_i, s_j) &= \frac{\alpha_t(s_i)p_{ij}\psi_{s_j,v_t}\beta_{t+1}(s_j)}{\mathbb{P}_\lambda(V)} \\ &= \frac{\alpha_t(s_i)p_{ij}\psi_{s_j,v_t}\beta_{t+1}(s_j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(s_i)p_{ij}\psi_{s_j,v_t}\beta_{t+1}(s_j)}.\end{aligned}$$

Kod problema dekodiranja smo definirali $\gamma_t(s_i)$ kao vjerojatnost da se nalazimo u stanju $s_i$ u trenutku t, uz dani opažajni niz i model. Možemo povezati $\gamma_t(s_i)$ i $\xi_t(s_i, s_j)$ sumiranjem po $j$

$$\gamma_t(s_i) = \sum_{j=1}^{N}\xi_t(s_i, s_j).$$

Ako sumiramo $\gamma_t(s_i)$ po $t$, dobijemo nešto što možemo interpretirati kao očekivani broj posjeta stanja $s_i$ ili ekvivalentno, kao očekivani broj prijelaza iz stanja $s_i$. Na sličan način, sumiranjem $\xi_t(s_i, s_j)$ po $t$ (od $t=1$ do $t=T-1$) dobijemo očekivani broj prijelaza iz stanja $s_i$ u stanje $s_j$. Dakle, imamo:

$$\begin{aligned}\sum_{t=1}^{N-1}\gamma_t(s_i) &= \text{očekivani broj prijelaza iz stanja } s_i, \\ \sum_{t=1}^{N-1}\xi_t(s_i, s_j) &= \text{očekivani broj prijelaza iz stanja } s_i \text{ u stanje } s_j.\end{aligned}$$

Koristeći ove veličine možemo dati metodu za ponovno procjenjivanje parametara skrivenog Markovljevog modela.

$$\hat{\pi}_i = \text{očekivani broj posjeta stanja } s_i \text{ u trenutku } t = 1,$$

$$\begin{aligned}\hat{p}_{ij} &= \frac{\text{očekivani broj prijelaza iz stanja } s_i \text{ u stanje } s_j}{\text{očekivani broj prijelaza iz stanja } s_i} \\ &= \frac{\sum_{t=1}^{N-1} \xi_t(s_i, s_j)}{\sum_{t=1}^{N-1} \gamma_t(s_i)},\end{aligned}$$

$$\begin{aligned}\hat{\psi}_{s_j, v_k} &= \frac{\text{očekivani broj posjeta stanja } s_j \text{ i viđanja simbola } v_k}{\text{očekivani broj posjeta stanja } s_j} \\ &= \frac{\sum_{t=1}^{N-1} \gamma_t(s_j), \text{ tako da } Y_t = v_k}{\sum_{t=1}^{N-1} \gamma_t(s_i)}.\end{aligned}$$

# 5 Zaključak

U prvom dijelu rada dan je pregled osnovnih pojmova i rezultata teorije Markovljevih lanaca. Markovljevi lanci su procesi koji imaju, grubo rečeno, svojstvo da ako znamo vrijednost koju je proces preuzeo u određenom trenutku nećemo dobiti nikakve informacije o budućem ponašanju procesa prikupljanjem više znanja o prošlosti. Na toj ideji razvila se metoda koja je važna u primjenama, primjerice u obradi signala ili strojnom učenju, a poznata je pod nazivom skriveni Markovljevi modeli. To je statistički model koji se sastoji od dva procesa, jednog vidljivog i Markovljevog lanca koji je skriven. Model je nastao 1960. godine i uglavnom se koristi u područjima kao što su prepoznavanje uzoraka, obrada signala, telekomunikacije, bioinformatika i u analizi volatilnosti financijskih tržišta. Kroz rad je objašnjeno da se vjerojatnosnim metodama dobije velika složenost, pa su stoga razvijeni različiti algoritmi: forward-backward algoritam, Viterbijev algoritam, Baum-Welch algoritam. Algoritmi daju efikasniji odgovor na tri osnovna problema: problem evaluacije, problem dekodiranja i problem učenja.

# Popis slika

# Literatura


[1] Handel, R. (July 28, 2008.) Hidden Markov Models, Lecture Notes.

[2] Hilgers, P. i Langville, A. (2006) The five greatest applications of Markov chains.

[3] Rabiner, L. i Juang, B. (1986) An Introduction to Hidden Markov Model, IEEE ASSP Magazine 3(1),

[4] Rabiner, L. (1989) A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, SProceedings of the IEEE, 77 (2).

[5] Sarapa, N. (2002.) Teorija vjerojatnosti, Zagreb: Školska knjiga

[6] Vondraček, Z. (2008) Markovljevi lanci, predavanja, Prirodoslovno-matematički fakultet, Zagreb.

[7] `https://www.philol.msu.ru/~lex/khmelev/published/llc/khmelev.html` 19. kolovoza 2022

[8] `http://www.alpha60.de/research/markov/DavidLink_AnExampleOfStatistical_MarkovTrans_2007.pdf.` 1. rujna 2022.