Samoobučavajući i adaptivni algoritmi

Računarski upravljački sistemi

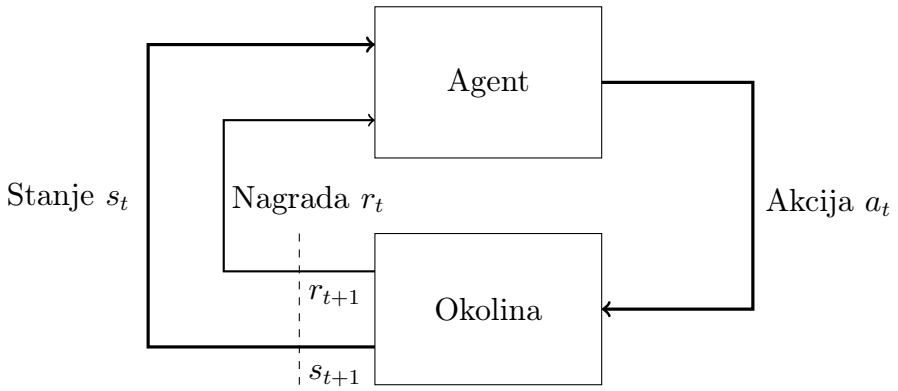
Sadržaj

I. Učenje potkrepljivanjem - *Reinforcement learning*

A. Uvod u učenje potkrepljivanjem

Ideja da (ljudska) bića uče tako što intereaguju sa okolinom je prva ideja koju pomislimo kada govorimo o prirodi učenja. Kroz život, takve interakcije su nesumnjivo ogroman izvor našeg znanja o okolini, kao i o nama. Šta god radili, kakve god akcije preduzimali, svjesni smo činjenica i reakcija okoline. **Učenje iz interakcija** jeste fundamentalna ideja gotovo svih teorija inteligencije i učenja.

Učenje potkrepljivanjem (***Reinforcement learning***) jeste **tehnika učenja** koja za cilj ima takvu **obuku agenta** (onog koji „**uči**") da, u budućnosti, on preduzima **optimalne odluke** u zavisnosti od **stanja** u kojem se on nalazi i „**nagrada**" (povratne informacije) iz okoline koje dobija. Prosto rečeno, agent uči na osnovu stanja i dobijene nagrade šta je najbolje uraditi. Agentu u ovom slučaju nije data mapa akcija koje on mora preduzeti kako bi ostvario najveći uspjeh - naprotiv, na njemu je da pronađe one akcije koje maksimizuju nagradu. Na agentu je, takođe, da prepozna ne samo akcije koje maksimizuju trenutnu nagradu, već da bude svjestan toga da trenutno preuzete akcije utiču na sve buduće nagrade. Ponašanje agenta je tako određeno **skupom akcija** $\mathcal{A}$, a kako ih agent preduzima, on prelazi iz jednog stanja u drugo. Stanja su opisana **skupom stanja** $\mathcal{S}$. Agentu je takođe, pri prelasku između datih stanja, dodjeljena nagrada iz **skupa nagrada** $\mathcal{R}$, koja upravo govori o kvalitetu preduzete akcije.



Slika I-A.1: Interakcija **agent** - **okolina** u Markovljevom procesu odlučivanja

Dvije najbitnije karakteristike *reinforcement learning*-a su **metoda pokušaja i pogrešaka** preduzimanja akcija (neke akcije su bolje od drugih), kao i **buduće nagrade** koje direktno zavise od preuzetih akcija.

Za formalni zapis problema učenja potkrepljivanjem (odnosno dinamike sistema koju posmatramo) koriste se **Markovljevi procesi odlučivanja**, te se na njih primjenjuje **najbolja akcija** (odnosno biranje najboljih poteza - akcija) radi ostvarivanja zadatog cilja (maksimizacije nagrade).

Jedan od izazova *reinforcement learning*-a jeste to što je potrebno postojanje **kompromisa** između **eksploatacije** i **eksploracije**. Kako bi agent maksimizovao nagradu, on mora preduzimati akcije koje je pokušao u prošlosti i koje su davale dobru nagradu, ali kako bi otkrio takve akcije, on mora pokušati one koje nije birao prije, odnosno mora posjedovati eksplorativnu komponentu kako bi verifikovao koje akcije su bolje od drugih. Dakle, agent mora **eksploatisati** ono što je već iskusio, ali mora i **istraživati** kako bi pravio bolje odluke u budućnosti. Jedna karakteristika povlači drugu - agent prosto mora da proba razne akcije i da bira one koje za njega odgovaraju.

Reinforcement learning je dio modernog trenda vještačke inteligencije i mašinskog učenja koji ima za cilj integraciju sa teorijom optimizacije, statističkim računom i ostalim matematičkim oblastima. Pored tehničkih oblasti, učenje potkrepljivanjem redovno intereaguje sa psihologijom i neuronaukama, za cilj imitiranja i razumijevanja čovjekove sposobnosti obzervacije i učenja iz datih.

B. *Sastavni elementi učenja potkrepljivanjem*

Pored **agenta** i **okoline** u kojoj se on nalazi i sa kojom intereaguje, postoje četiri suštinska dijela *reinforcement learning* sistema - **politika** (ili **način**) **odlučivanja**, **nagrada**, **vrijednost stanja** i (opciono) **model okoline**.

Politika odlučivanja (eng. *policy*) je način ponašanja agenta u trenutku vremena, odnosno u određenom stanju. Matematički rečeno, politika odlučivanja je mapiranje stanja u kojem se agent nalazi na akcije koje preduzima kada se nađe u datim stanjima

$$\pi : \mathcal{S} \to \mathcal{A} \tag{I-B.1}$$

opisano rečenicom „*ako sam u stanju* s_1*, preduzimam akciju* a_1" (u deterministčkom slučaju).

Nagrada definiše cilj u kontekstu problema kojeg rješavamo. Pri prelasku između stanja (odnosno, u svakom diskretnom vremenskom trenutku), agent iz okoline dobija signal nagrade, koji govori o tome kako je prethodno preduzeta akcija djelovala na okolinu, te samim tim i na njega. Time se može povući analogija između dobijanja „male" i „velike" nagrade sa osjećanjem „boli" i „zadovoljstva", u biološkom smislu. Iz toga ishodi da je nagrada **mjera kvaliteta** politike odlučivanja, jer iz dobijene nagrade možemo izvući zaključak o tome da li je politika koju agent sprovodi **dobra**, odnosno **optimalna** za dati problem.

Vrijednost stanja jeste kumulativna suma nagrada koje agent može očekivati da dobije, počevši od tog stanja u kojem se trenutno nalazi. Dok nagrade određuju trenutnu dobit, vrijednosti stanja predstavljaju dugoročni plan dobiti kada se uzmu u obzir buduća stanja u kojima se agent može naći. Matematički iskazano, vrijednost stanja je suma nagrada trenutnog i svih budućih stanja (može ih biti beskonačno za neterminirajuće procese) **ponderisana težinskim faktorom** γ

$$v(s) = \begin{cases} \sum_{k=0}^{T} \gamma^k r_k, \text{ za deterministički sistem} \\ \mathbb{E}\left\{\sum_{k=0}^{T} \gamma^k R_k\right\}, \text{ za stohastički sistem} \end{cases} \tag{I-B.2}$$

Razlog uvođenja pondera γ dvojake je prirode. Naime, prvi razlog je matematički - olakšava računanje pomenute sume koja, kao što je rečeno, može sadržati beskonačno mnogo članova, no drugi i bitniji razlog jeste (**ljudska**) **priroda prioritetizovanja trenutnih nad budućim nagradama**. Bić su sklona davati prioritet trenutnom stanju i trenutnim nagradama, dok ih ne zanima šta će se dešavati u budućnosti i kakve će nagrade tada dobijati, te se u tu svrhu uvodi težinski faktor γ koji smanjuje značaj budućih nagrada (ograničava se u intervalu od 0 do 1). Formalna definicija svih gorenavedenih pojmova data je u nastavku teksta.

Naravno, iz prethodnog rečenog nepravilno bi bilo izvući zaključak da su agentu sveohuvatno bitnija trenutna stanja i da će njih maksimalno prioritetisati. Čitav niz donošenja akcija zavisi od vrijednosti stanja i agent potražuje one akcije koje ga dovode u stanja najvećih vrijednosti jer to direktno pravi put akcija koje dovode do ukupne maksimalne nagrade.

II. Markovljev proces odlučivanja

A. *Definicija Markovljevog procesa odlučivanja*

Markovljev proces odlučivanja (*Markov decision process*, **MDP**) je formalizacija sekvencijalog procesa donošenja odluka, matematički idealizovana forma *reinforcement learning* problema.

Kao što smo već vidjeli, onaj koji uči i koji donosi odluke naziva se **agent**. Stvar s kojom intereaguje i koja nije njegov integrisani sastav jeste **okolina**. Sa pomenutom intereaguje u sekvenci vremenskih trenutaka, $t = 0, 1, 2, ...$, donoseći akcije koje utiču na nju i dobijajući kao povratnu informaciju novu situaciju (**stanje**) u sistemu i **nagradu**, kao reakciju na preduzetu akciju. Sekvenca kojom se ovo dešava formira **trajektoriju** dešavanja

$$s_0, a_0, r_0, s_1, a_1, r_1, s_2, a_2, r_2, ... \tag{II-A.1}$$

$$s \in \mathcal{S},\ a \in \mathcal{A},\ r \in \mathcal{R} \tag{II-A.2}$$

Potrebno je razgraničiti dvije vrste MDP-a:

1) **Deterministički Markovljev proces**
   Deterministički Markovljev proces jeste model procesa u kojem su za svako stanje agenta tačno određene akcija koja se preduzima, kao i sledeće stanje, odnosno

$$s^+ = f(s, a) \tag{II-A.3}$$

$$r = h(s, a) \tag{II-A.4}$$

   gdje funkcije f i h određuju **deterministički model**.
   Dakle, unaprijed je moguće odrediti trajektoriju sistema bazirano na trenutnom stanju agenta.
2) **Stohastički Markovljev proces**
   Stohastički Markovljev proces je uopštenje determinističkog. Naime, u zavisnosti od trenutnog stanja i akcije koju agent preduzme, postoji vjerovatnoća da će on završiti u nekom od narednih stanja i dobiti neku nagradu, odnosno

$$p(s^+, r|s, a) = \mathbb{P}\{S^+ = s^+, R = r|S = s, A = a\} \tag{II-A.5}$$

   gdje se vjerovatnoća p naziva **stohastičkim modelom**, **modelom prelaza**, **vjerovatnoća prelaza**, **dinamika sistema**.
   Dakle, nije moguće odrediti tačnu trajektoriju sistema baziranu na trenutnom stanju, već je ona u zavisnosti od raspodjele vjerovatnoća.

B. *Dobitak (Gain), politika odlučivanja, vrijednost stanja, vrijednost akcije*

U procesu interakcije sa okolinom, agent kao povratni signal dobija nagradu. **Ukupna dobijena nagrada** (*gain*) računa se kao suma sekvence svih dobijenih nagrada počevši od trenutnog momenta i nastavljajući se u budućnost, odnosno

$$g = \begin{cases} \sum_{k=0}^{T} \gamma^k r_k, \text{ za deterministički sistem} \\ \mathbb{E}\left\{\sum_{k=0}^{T} \gamma^k R_k\right\}, \text{ za stohastički sistem} \end{cases} \tag{II-B.1}$$

Agent intereaguje sa okolinom donoseći odluke. Pravilo pod kojim se donose odluke naziva se **politika odlučivanja**

$$a = \begin{cases} \pi(s), \text{ za determinističke politike} \\ \pi(a|s) = \mathbb{P}\{A = a|S = s\}, \text{ za stohastičke politike} \end{cases} \tag{II-B.2}$$

i ona očigledno zavisi od stanja u kojem se agent nalazi. Takođe, valjalo bi primjetiti da je moguće primjeniti ili determinističku ili stohastičku politiku, nezavisno od toga da li je sam proces deterministički ili stohastički - **politika odlučivanja ne zavisi od karakteristika modela**.

Ako fiksiramo početno stanje, možemo pričati o **očekivanoj nagradi** u zavisnosti od politike odlučivanja π

$$g_\pi(s_0) = \mathbb{E}_\pi \left\{ \sum_{k=0}^{T} \gamma^k R_k | S_0 = s_0 \right\} \tag{II-B.3}$$

Tada možemo definisati **vrijednost stanja**

$$v_\pi(s) = g_\pi(s) \tag{II-B.4}$$

koje određuje **očekivanu nagradu počevši od stanja** s.

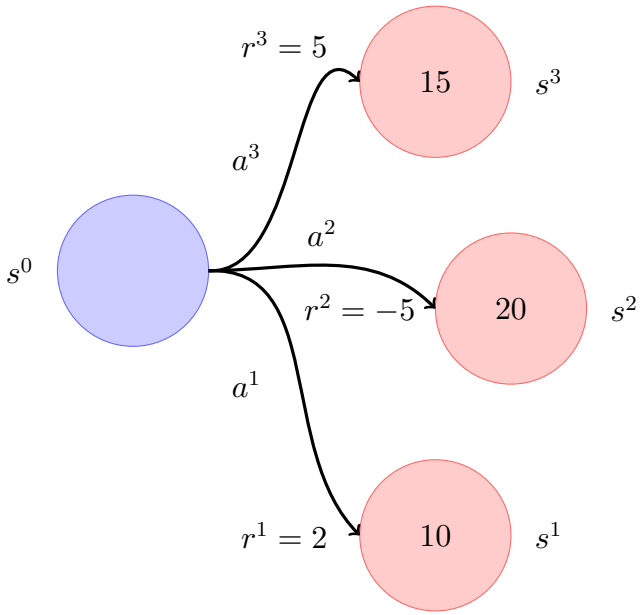
Takođe, definiše se i **vrijednost akcije**

$$q_\pi(s,a) = g_\pi(s|s_0 = s, a_0 = a) = r_0 + \gamma g_\pi(s^+) = r_0 + \gamma q_\pi(s^+, a^+) \tag{II-B.5}$$

koja govori o **nagradi** (označena sa r_0) **dobijenoj kada se krene iz početnog stanja** $s_0 = s$ **i preduzme akcija** $a_0 = a$, te se **nakon toga slijedi politika** π. Dakle, početna akcija a nezavisna je od politike π, no može biti njen dio (bez gubitka opštosti), te zato imamo puno pravo pisati $g_\pi(s^+) = q_\pi(s^+, a^+)$.

Primjer II.1. Određivanje vrijednosti stanja

Posmatrajmo dati sistem odlučivanja



Slika II-B.1: Primjer 1.

Dakle, na raspolaganju su nam date tri akcije, $\mathcal{A} = \{a^1, a^2, a^3\}$. Nagrada koju možemo očekivati da dobijemo počevši u stanju s^0 jeste, po formuli

$$g(s) = \sum_{k=0}^{T} \gamma^k r_k = r_0 + \gamma r_1 + \gamma^2 r_2 + \gamma^3 r_3 ... = r_0 + \gamma(r_1 + \gamma r_2 + \gamma^2 r_3 ...) \tag{II-B.6}$$

Akcija a	**Stanje** s^+	**Vrijednost stanja** $v(s^+)$	**Nagrada** r
a^1	s^1	10	2
a^2	s^2	20	-5
a^3	s^3	15	5

Primjetimo da je red $r_1 + \gamma r_2 + ...$ zapravo **dobijena nagrada**[1] za stanje s^+, te će važiti da je

$$g(s) = r_0 + \gamma g(s^+) \tag{II-B.7}$$

Dobijeni rezultat predstavlja **rekurzivnu formulu za računanje dobijene nagrade**, a kako je, za konkretnu politiku odlučivanja $v_\pi(s) = g_\pi(s)$, slijedi da je

$$v_\pi(s) = r_0 + \gamma v_\pi(s^+) \tag{II-B.8}$$

U zavisnosti od preduzete akcije možemo izračunati vrijednost stanja s^0 (za parametar γ uzeta je vrijednost 0.9)

Akcija a	**Stanje** s^+	**Vrijednost stanja** $v(s^0)$
a^1	s^1	11
a^2	s^2	13
a^3	s^3	18.5

Prethodni primjer dobar je pokazatelj preduzimanja **optimalne politike odlučivanja**, koja se definiše kao **politika koja će donijeti najveću nagradu** od svih mogućih politika $\pi \in \mathcal{P}$

$$v^*_\pi(s) = \max_{\pi\in\mathcal{P}}\{v_\pi(s)\} = \max_{\pi\in\mathcal{P}}\{r_0 + \gamma v^*_\pi(s^+)\} = \max_{\pi\in\mathcal{P}}\{h(s,a) + \gamma v^*_\pi(f(s,a))\} \tag{II-B.9}$$

$$a = \pi(s) \tag{II-B.10}$$

Konkretno za primjer, optimalna politika bi bila akcija a^3, koja bi vrijednost stanja s^0 postavila na 18.5.

Sličan rezultat dobijamo računajući vrijednosti akcija

Akcija a	**Vrijednost akcije** $q(s^0, a)$
a^1	11
a^2	13
a^3	18.5

te bismo **optimalnu politiku koja donosi najveću vrijednost akcije** računali kao

$$q^*_\pi(s,a) = \max_{\pi\in\mathcal{P}}\{q_\pi(s,a)\} = \max_{\pi\in\mathcal{P}}\{r_0 + q^*_\pi(s^+,a^+)\} = \max_{\pi\in\mathcal{P}}\{h(s,a) + q^*_\pi(f(s,a),a^+)\} \tag{II-B.11}$$

$$a^+ = \pi(s^+) \tag{II-B.12}$$

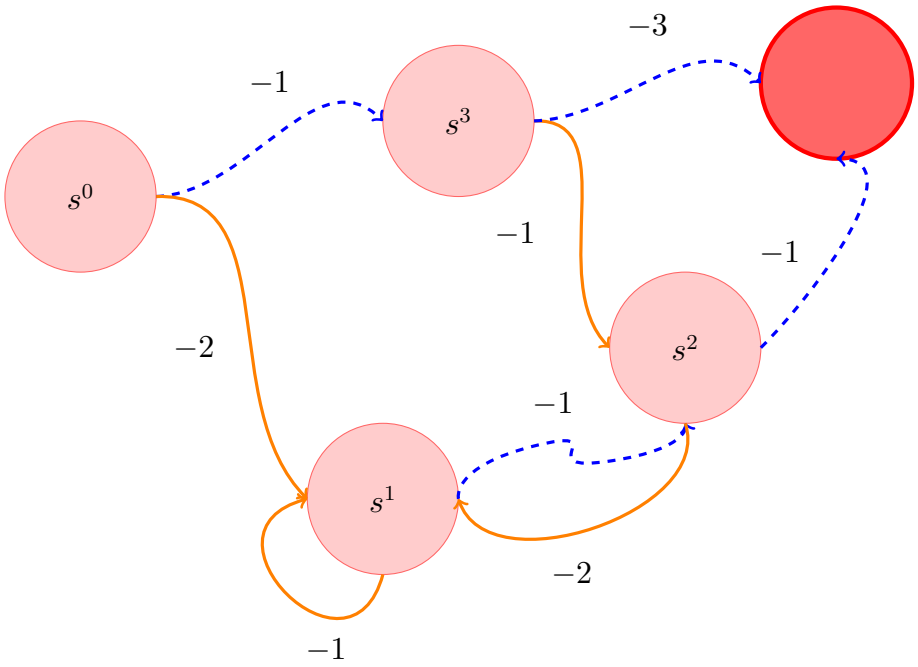
[1]Subskript označava vremenske trenutke, te je za stanje s^+ zapravo početni trenutak $t = 1$ i formula za dobijenu nagradu ostaje validna.

III. Načini računanja vrijednosti stanja i vrijednosti akcija. Optimalna politika odlučivanja

A. *Iterativno računanje vrijednosti stanja i vrijednosti akcija*

Primjer III.1. Evaluacija vrijednosti stanja na osnovu konkretne politike

Pogledajmo sledeći primjer[2]



Slika III-A.1: Primjer 1.

Neka je politika π

$$\pi(s) = \text{"}\rightarrow\text{"} \tag{III-A.1}$$

odnosno „**biraj plavu** (**isprekidanu**) **akciju**" .

Krenuvši unazad, odnosno od terminalnog stanja, vrijednosti svakog od stanja su

$$v^3 = -3 + \gamma v^{\text{term}} \tag{III-A.2}$$
$$v^2 = -1 + \gamma v^{\text{term}} \tag{III-A.3}$$
$$v^1 = -1 + \gamma v^2 \tag{III-A.4}$$
$$v^0 = -1 + \gamma v^3 \tag{III-A.5}$$

Očigledno je da ako izaberemo $v^{\text{term}} = 0$, dobijamo vrijednosti za v^2 i v^3, uz pomoć kojih možemo izračunati vrijednosti preostalih stanja v^0 i v^1. **Algoritam evaluacije vrijednosti stanja** je tako

1) **Izabrati vrijednost** v^{term}
2) **Pronaći i evaluirati sva stanja iz kojih je moguće doći u terminalno**
3) **Pronaći i evaluirati sva stanja iz kojih je moguće do stanja već poznatih vrijednosti**
4) **Ponoviti korak 3.**

[2]Jarko crvenom bojom označeno je **terminalno „stanje"**. Ono zapravo nije suštinski stanje, jer u njemu ne možemo donositi akcije - označava kraj „igre". Obično uzimamo njegovu vrijednost za 0.

Kada imamo ogroman skup stanja i složeniju politiku odlučivanja, potrebno je primjeniti numeričke iterativne metode za određivanje vrijednosti stanja. Ako matrično zapišemo prethodne jednačine

$$\begin{bmatrix} v^0 \\ v^1 \\ v^2 \\ v^3 \end{bmatrix} = \gamma \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} v^0 \\ v^1 \\ v^2 \\ v^3 \end{bmatrix} + \gamma \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} v^{\text{term}} + \begin{bmatrix} -1 \\ -1 \\ -1 \\ -3 \end{bmatrix} \tag{III-A.6}$$

Direktno rješavanje po matrici $\mathbf{v}$ dobija se formulom

$$\mathbf{v} = (\mathbf{I} - \gamma \mathbf{A})^{-1} \mathbf{r} \tag{III-A.7}$$

ili iterativnom metodom

$$\mathbf{v}^{k+1} = \gamma \mathbf{A} \mathbf{v}^k + \mathbf{r} \tag{III-A.8}$$

gdje je $\mathbf{v}^0$ nasumično inicijalizovano.

*B. **Determinističke i stohastičke jednačine vrijednosti stanja i akcija***

Podsjetimo se rekurzivnih formula za računanje vrijednosti stanja i vrijednost akcija u determinističkom slučaju

$$v_\pi(s) = h(s, \pi(s)) + \gamma v_\pi(f(s, \pi(s))) \tag{III-B.1}$$
$$q_\pi(s, a) = h(s, a) + \gamma q_\pi(f(s, a), \pi(f(s, a))) \tag{III-B.2}$$

U stohastičkom slučaju govorimo o **očekivanoj nagradi**, odnosno

$$g_\pi(s) = \mathbb{E} \left\{ \sum_{k=0}^{T} \gamma^k R_k | S_0 = s \right\} \tag{III-B.3}$$

Razložimo izraz na sledeći način

$$\begin{gathered} g_\pi(s) = \mathbb{E}_\pi \left\{ \sum_{k=0}^{T} \gamma^k R_k | S_0 = s \right\} = \mathbb{E}_\pi \left\{ R_0 + \gamma \sum_{k=1}^{T} \gamma^{k-1} R_k | S_0 = s \right\} = \\ \sum_{a \in \mathcal{A}} \pi(a|s) \mathbb{E}_\pi \left\{ R_0 + \gamma \sum_{k=1}^{T} \gamma^{k-1} R_k | S_0 = s, A_0 = a \right\} = \\ \sum_{a \in \mathcal{A}} \pi(a|s) \sum_{\substack{s^+ \in \mathcal{S} \\ r \in \mathcal{R}}} p(s^+, r|s, a) \mathbb{E}_\pi \left\{ R_0 + \gamma \sum_{k=1}^{T} \gamma^{k-1} R_k | S_0 = s, A_0 = a, S_1 = s^+, R_0 = r \right\} = \\ \sum_{a \in \mathcal{A}} \pi(a|s) \sum_{\substack{s^+ \in \mathcal{S} \\ r \in \mathcal{R}}} p(s^+, r|s, a) \left[r + \gamma \mathbb{E}_\pi \left\{ \sum_{k=1}^{T} \gamma^{k-1} R_k | S_1 = s^+ \right\} \right] \end{gathered} \tag{III-B.4}$$

Poslednja formula može se zapisati kao

$$v_\pi(s) = \sum_{a\in\mathcal{A}} \pi(a|s) \sum_{\substack{s^+\in\mathcal{S}\\ r\in\mathcal{R}}} p(s^+, r|s, a) \left[r + \gamma v_\pi(s^+)\right] \tag{III-B.5}$$

i predstavlja **jednačinu vrijednosti stanja u stohastičkom slučaju**.
Što se tiče vrijednosti akcija

$$q_\pi(s, a) = \mathbb{E}_\pi \left\{ R_0 + \gamma \sum_{k=1}^{T} \gamma^{k-1} R_k | S_0 = s, A_0 = a \right\} \tag{III-B.6}$$

Ako raspišemo dalje

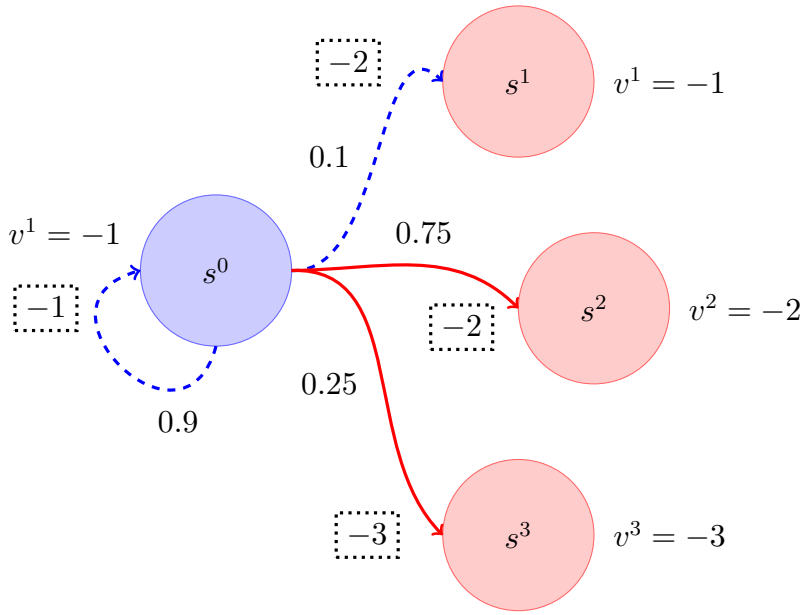
$$\begin{gathered} q_\pi(s, a) = \mathbb{E}_\pi \left\{ R_0 + \gamma \sum_{k=1}^{T} \gamma^{k-1} R_k | S_0 = s, A_0 = a \right\} = \\ \sum_{\substack{s^+\in\mathcal{S}\\ r\in\mathcal{R}}} p(s^+, r|s, a) \mathbb{E}_\pi \left\{ R_0 + \gamma \sum_{k=1}^{T} \gamma^{k-1} R_k | S_0 = s, A_0 = a, S_1 = s^+, R_0 = r \right\} = \\ \sum_{\substack{s^+\in\mathcal{S}\\ r\in\mathcal{R}}} p(s^+, r|s, a) \left[r + \gamma \mathbb{E}_\pi \left\{ \sum_{i=1}^{T} \gamma^{k-1} R_k | S_1 = s^+ \right\} \right] = \\ \sum_{\substack{s^+\in\mathcal{S}\\ r\in\mathcal{R}}} p(s^+, r|s, a) \left[r + \gamma \sum_{a^+\in\mathcal{A}} \pi(a^+|s^+) \mathbb{E}_\pi \left\{ \sum_{k=1}^{T} \gamma^{k-1} R_k | S_1 = s^+, A_1 = a^+ \right\} \right] \end{gathered} \tag{III-B.7}$$

Poslednja formula može se zapisati kao

$$q_\pi(s, a) = \sum_{\substack{s^+\in\mathcal{S}\\ r\in\mathcal{R}}} p(s^+, r|s, a) \left[r + \gamma \sum_{a^+\in\mathcal{A}} \pi(a^+|s^+) q_\pi(s^+, a^+) \right] \tag{III-B.8}$$

i predstavlja **jednačinu vrijednosti stanja u stohastičkom slučaju**.
Prethodne jednačine nazivaju se **Belmanovim jednačinama**.

Primjer III.2. Primjer određivanja vrijednosti stanja i akcija na stohastičkom MDP-u
Razmotrimo sledeći primjer



Slika III-B.1: Primjer 2.

Decimalne vrijednosti na slici iznad označavaju vjerovatnoću završetka u određenim stanjima ako se biraju konkretne akcije, dok su uokvirene vrijednosti nagrade koje agent dobija igranjem određenih akcija. Takođe, poznate su nam vrijednosti stanja s^1, s^2 i s^3.

Neka je politika odlučivanja „**Ako si u stanju s^0, sa 30% biraj akciju '→' (plavu isprekidanu akciju), dok sa 70% biraj akciju '→' (crvenu punu akciju)**". Matematički zapisano

$$\pi(a|s) = \begin{cases} 0.3, \ s = s^0 \to a = \textit{Isprekidana plava} \\ 0.7, \ s = s^0 \to a = \textit{Puna crvena} \end{cases} \tag{III-B.9}$$

Ako zapišemo tablično šta vidimo na slici iznad

Stanje	**Akcija**	**Prelaz u stanje**	$p(s^+, r\|s, a)$
s^0	*Isprekidana plava*	s^0	$p(s^+ = s^0, r = -1\|s = s^0, a = \textit{Ispr. plava}) = 0.9$
	Isprekidana plava	s^1	$p(s^+ = s^1, r = -2\|s = s^0, a = \textit{Ispr. plava}) = 0.1$
	Puna crvena	s^2	$p(s^+ = s^2, r = -2\|s = s^0, a = \textit{Puna crv.}) = 0.75$
	Puna crvena	s^3	$p(s^+ = s^3, r = -3\|s = s^0, a = \textit{Puna crv.}) = 0.25$

Dakle, imamo **stohastičku politiku odlučivanja**, kao i **stohastičku okolinu**, te ćemo primjeniti formulu

$$v_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \sum_{\substack{s^+ \in \mathcal{S} \\ r \in \mathcal{R}}} p(s^+, r|s, a) \left[r + \gamma v_\pi(s^+) \right] \tag{III-B.10}$$

Prvo rješimo unutrašnju sumu. To radimo fiksiranjem po akcijama, odnosno posmatrajući šta se dešava kada izaberemo partikularnu akciju a

$$v^0_{\text{Ispr. pl.}} = 0.9(-1 + \gamma v^0_{\text{Ispr. pl.}}) + 0.1(-2 + \gamma v^1) \tag{III-B.11}$$
$$v^0_{\text{Puna crv.}} = 0.75(-2 + \gamma v^2) + 0.25(-3 + \gamma v^3) \tag{III-B.12}$$

te sabiranjem dobijenih rezultata ponderisanih vjerovatnoćama π

$$v^0 = 0.3v^0_{\text{Ispr. pl.}} + 0.7v^0_{\text{Puna crv.}} \tag{III-B.13}$$

C. *Belmanove jednačine optimalne politike*

Zapisano opet, jednačine optimalnih politika u zavisnosti od vrijednosti stanja i vrijednosti akcija

$$v^*_\pi(s) = \max_{\pi\in\mathcal{P}}\{h(s,a) + \gamma v^*_\pi(f(s,a))\}, \text{ gdje } a = \pi(s) \tag{III-C.1}$$
$$q^*_\pi(s,a) = \max_{\pi\in\mathcal{P}}\{h(s,a) + \gamma q^*_\pi(f(s,a), a^+)\}, \text{ gdje } a^+ = \pi(s^+) = \pi(f(s,a)) \tag{III-C.2}$$

Kako su politike konstruisane od akcija, mi suštinski za svako stanje moramo pronaći onakvu akciju čija primjena dovodi do najveće nagrade, te se **pronalaženja optimalnih politika odlučivanja** svodi na **pronalaženje optimalnih akcija**.

Dakle, to znači da se prethodno napisana formula za evaluaciju vrijednosti stanja može zapisati kao

$$v^*(s) = \max_{a\in\mathcal{A}}\{h(s,a) + \gamma v^*(f(s,a))\} \tag{III-C.3}$$

Što se tiče optimalne vrijednosti akcija, optimalna politika se slijedi tek nakon prvog koraka (prva akcija nije zavisna od politike odlučivanja), te se jednačina optimuma vrijednosti akcija može zapisati kao

$$q^*(s,a) = h(s,a) + \gamma \max_{a^+\in\mathcal{A}}\left\{q^*(f(s,a), a^+)\right\} \tag{III-C.4}$$

Prethodne jednačine predstavljaju **Belmanove jednačine optimalnosti**.

Stohastičke verzije jednačina optimalnosti su

$$v^*(s) = \max_{a\in\mathcal{A}}\left\{\sum_{\substack{s^+\in\mathcal{S}\\ r\in\mathcal{R}}} p(s^+, r|s,a)\left[r + \gamma v^*(s^+)\right]\right\} \tag{III-C.5}$$
$$q^*(s,a) = \sum_{\substack{s^+\in\mathcal{S}\\ r\in\mathcal{R}}} p(s^+, r|s,a)\left[r + \gamma \max_{a^+\in\mathcal{A}}\left\{q^*(s^+, a^+)\right\}\right] \tag{III-C.6}$$