Učenie posilňovaním ako Markovovský rozhodovací proces

(Strojové učenie II)

M. Mach

Katedra kybernetiky a umelej inteligencie, FEI, TUKE

január 2021 - február 2024

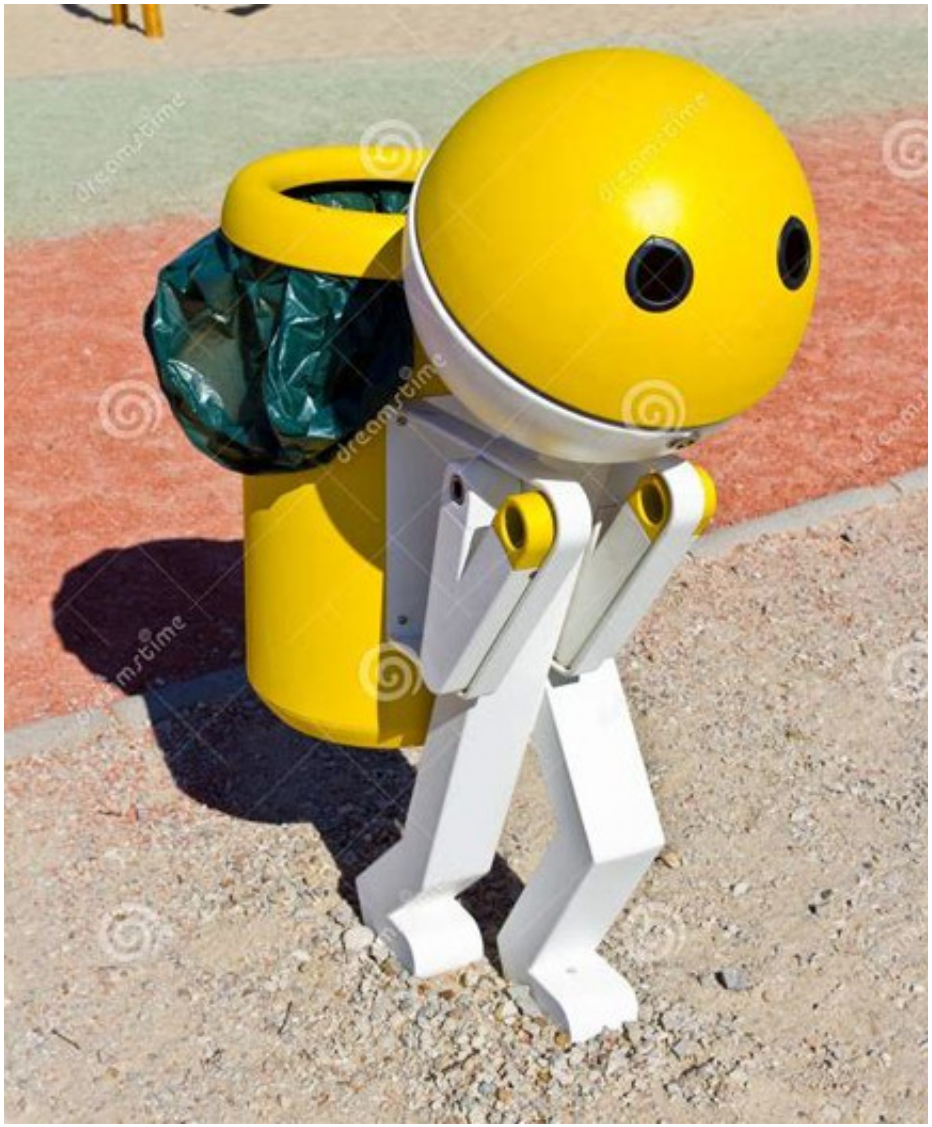
Markovovský rozhodovací proces (MDP)

- Markovovský proces rozšírený o odmeny a akcie
- Akumulácia odmien získavaných počas sekvencie
- Výsledky môžu byť náhodné (mimo kontroly toho, kto prijíma rozhodnutie)
  - deterministický prípad je špecializácia náhodného

MDP je n-tica (S,A,P,R), kde

- S je množina stavov
- A je množina akcií
- P je pravdepodobnostná prechodová matica $P^a_{ss'} = P[S_{t+1} = s' | S_t = s, A_t = a]$
- R je funkcia odmeny $R^a_s = E[R_{t+1} | S_t = s, A_t = a]$

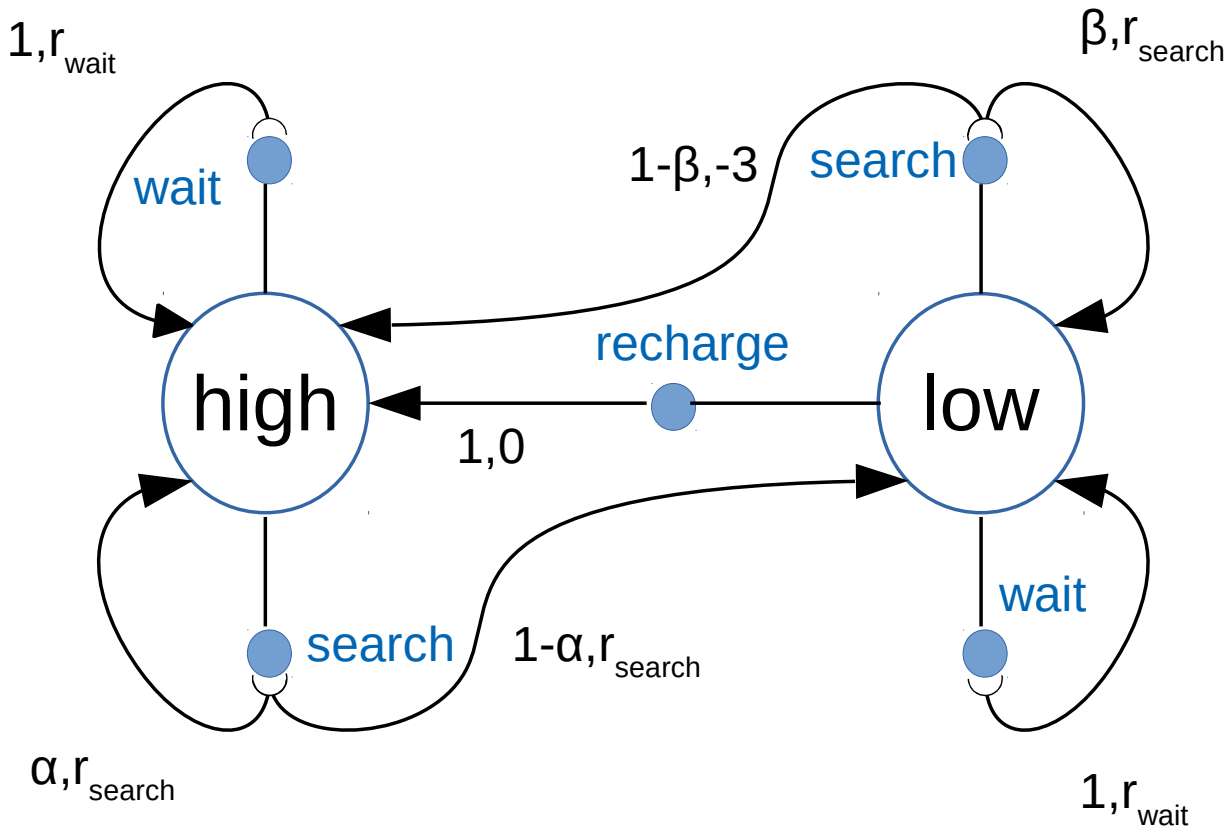
Ilustračný príklad - recyklačný robot



- $S = \{high, low\}$
- $A = \{search, wait, recharge\}$
  - $A(low) = \{search, wait, recharge\}$
  - $A(high) = \{search, wait\}$
- $R = \{r_{wait}, r_{search}, 0, -3\}$
- $P = \{\alpha, \beta, 0, 1\}$

Recyklačný robot ako MDP

$$p(s'|s,a) \qquad r(s,a,s') \rightarrow p(r|s,a,s')$$



©*Sutton-Barto: Reinforcement Learning, 2nd ed., 2018*

MDP pre učenie posilňovaním

- Agent interaguje s prostredím v diskrétnom čase
- Konečný MDP
  - konečný počet stavov
  - konečný počet akcií
  - konečný počet odmien
- Dynamika prostredia je definovaná distribúciou $p(s', r|s, a) = P[S_{t+1} = s', R_{t+1} = r|S_t = s, A_t = a]$
  - pričom platí $\sum_{s' \in S} \sum_{r \in R} p(s', r|s, a) = 1$ pre všetky $a \in A$, $s \in S$
- Deterministický prípad
  - $p : SxRxSxA \rightarrow [0, 1]$
  - výsledky rozhodnutí sú plne pod kontrolou toho, kto prijíma rozhodnutia

Dynamika prostredia

- Z dynamiky prostredia $p(s', r|s, a)$ sa dajú odvodiť ďalšie charakteristiky
  - pravdepodobnosti zmeny stavov ($p(s'|s, a)$, $p(r|s, a)$)

$$P^a_{ss'}: \; p(s'|s,a) = P[S_{t+1} = s'|S_t = s, A_t = a] = \sum_{r \in R} p(s', r|s, a)$$

  - očakávaná odmena pre pár stav-akcia

$$R^a_s: \; r(s,a) = E[R_{t+1}|S_t = s, A_t = a] = \sum_{r \in R} r\, p(r|s,a) = \sum_{r \in R} r \sum_{s' \in S} p(s', r|s, a) = \sum_{r \in R} r \sum_{s' \in S} p(r|s, a, s')p(s'|s, a)$$

Odmena

- Cieľ – maximalizácia očakávanej kumulatívnej odmeny
  - dôraz nie na bezprostrednú odmenu
  - odmena vyjadruje cieľ, ktorý má byť dosiahnutý (maximalizácia odmeny = dosiahnutie cieľa)
  - odmena komunikuje čo má byť dosiahnuté, nie ako

  Ak chceme aby agent
  - niečo dosiahol - tak za to musí byť kladná odmena
  - sa niečomu vyhol - tak za to záporná odmena

Kumulatívna odmena

$$G_t = R_{t+1} + R_{t+2} + R_{t+3} + \ldots + R_T$$

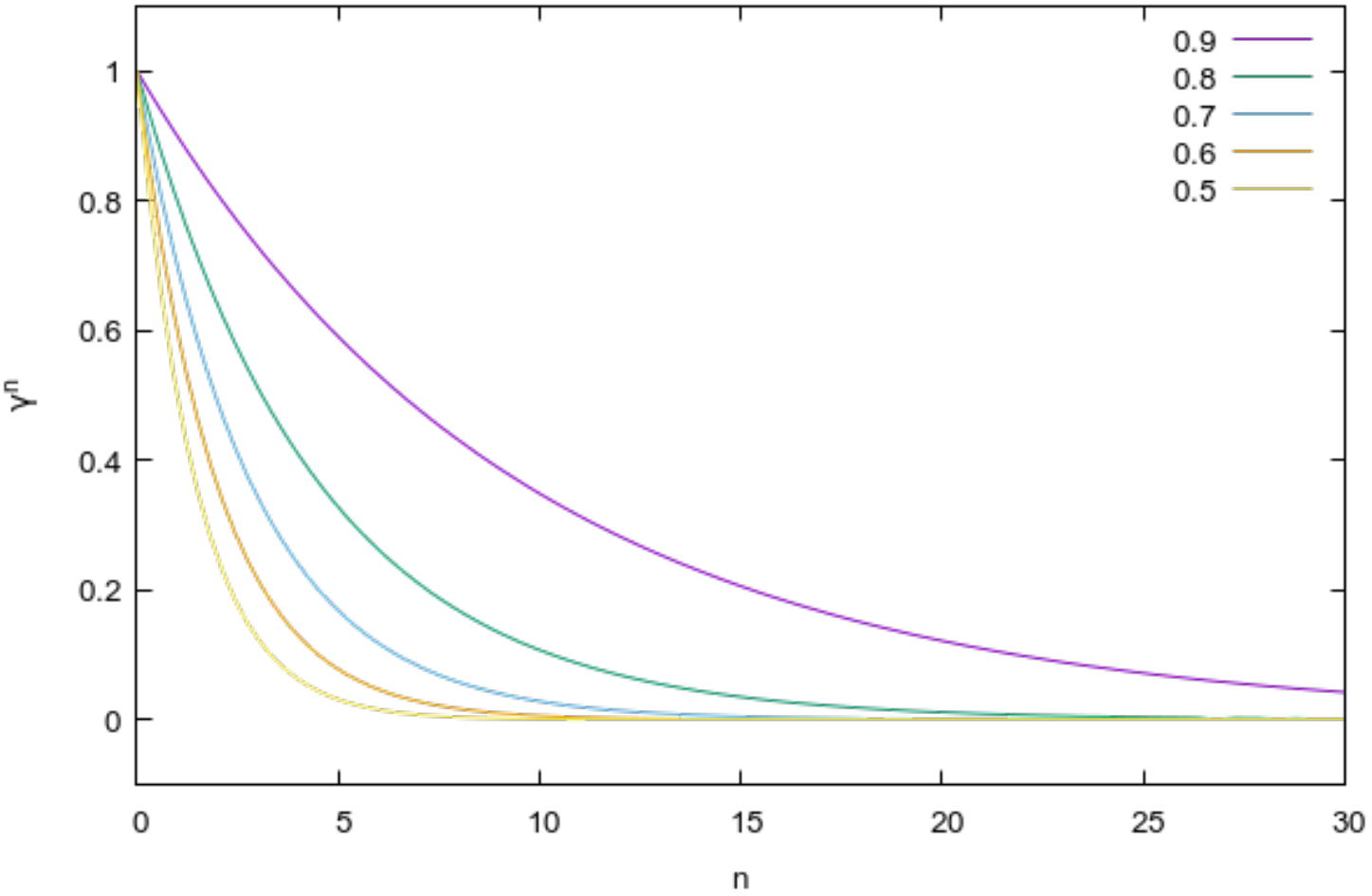
- Epizodická úloha
  - T je konečný čas, interakcia agenta s prostredím končí
  - sekvencia stavov epizódy končí v koncovom stave (po ňom nasleduje počiatočný stav novej epizódy)
  - epizódy sú navzájom nezávislé
  - T je náhodná premenná s rôznymi hodnotami v rôznych epizódach
  - G_t je konečná hodnota
- Kontinuálna úloha
  - $T = \infty$
  - G_t môže ale nemusí konvergovať ku konečnej hodnote

Unifikácia odmeny

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots = \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$$

- Unifikácia epizodických ($T < \infty$) aj kontinuálnych ($T = \infty$, $0 \leq \gamma < 1$) úloh
- Diskontný faktor $\gamma \in < 0, 1 >$
  - $\gamma = 0$ - uvažovanie iba bezprostrednej odmeny
  - čím je γ väčšia, tým agent je viac "ďaleko vidiaci"
  - γ vyjadruje súčasnú cenu budúcich odmien
- $\sum^{\infty}$ je konečná ak odmeny sú ohraničené ($R_i \leq R$)
  $G_t = \sum_{k=t+1}^{\infty} \gamma^{k-t-1} R_k \leq R \sum_{k=0}^{\infty} \gamma^k = R \frac{1}{1-\gamma}$

Vplyv diskontného faktora



Výhoda použitia γ

$$
\begin{aligned}
G_t &= R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \dots \\
&= R_{t+1} + \gamma (R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} + \dots) \\
&= R_{t+1} + \gamma G_{t+1}
\end{aligned}
$$

- Matematicky vhodný spôsob vyjadrenia
- Vyhnutie sa nekonečne veľkej kumulatívnej odmene
- Lepšia charakterizácia úloh v prípade, že
  - neurčitosť odmeny v budúcnosti
  - bezprostredná odmena je zaujímavejšia ako vzdialená
  - ľudské chovanie preferuje bezprostrednejšiu odmenu

CIT
Centrum inteligentných technológií

Voľba akcií (politika)

- Politika je spôsob, akým agent volí svoje akcie
- Formálne politika je distribúcia pravdepodobnosti nad akciami v určitom stave

$$\pi(a|s) = P[A_t = a|S_t = s] \quad kde \quad \sum_a \pi(a|s) = 1$$

- Zmyslom učenia posilňovaním je špecifikovať vhodnú politiku agenta na základe skúseností s pôsobením agenta v prostredí
- Politika plne definuje chovanie agenta
  - závisí iba na aktuálnom stave (MDP)
  - je stacionárna (časovo invariantná)

Hodnotové funkcie - stavy

- Hodnotová funkcia stavu (State-value function)

$$v_\pi(s) = E_\pi[G_t|S_t = s]$$

  - ako výhodné je byť v danom stave
  - očakávaná kumulatívna odmena sekvencie začínajúcej v danom stave, ak voľba akcií je podľa politiky π

$$\begin{aligned}
v_\pi(s) &= E_\pi[G_t|S_t = s] \\
&= E_\pi[R_{t+1} + \gamma G_{t+1}|S_t = s] \\
&= E_\pi[R_{t+1}|S_t = s] + \gamma E_\pi[G_{t+1}|S_t = s] \\
&= E_\pi[R_{t+1}|S_t = s] + \gamma E_\pi[E_\pi[G_{t+1}|S_{t+1} = s']|S_t = s] \\
&= E_\pi[R_{t+1}|S_t = s] + \gamma E_\pi[v_\pi(S_{t+1})|S_t = s] \\
&= E_\pi[R_{t+1} + \gamma v_\pi(S_{t+1})|S_t = s]
\end{aligned}$$

  - Bellmanova rovnica očakávania

Hodnotové funkcie - akcie

- Hodnotová funkcia akcie (Action-value function)

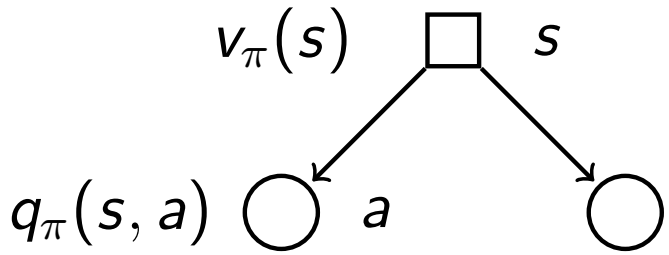
$$q_\pi(s,a) = E_\pi[G_t|S_t = s, A_t = a]$$

  - ako výhodné je v danom stave použiť danú akciu pri politike π
  - očakávaná kumulatívna odmena sekvencie začínajúcej v danom stave danou akciou, ak voľba nasledujúcich akcií je podľa politiky π

$$\begin{aligned} q_\pi(s,a) &= E_\pi[G_t|S_t = s, A_t = a] \\ &= E_\pi[R_{t+1} + \gamma G_{t+1}|S_t = s, A_t = a] \\ &= \dots \\ &= E_\pi[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1})|S_t = s, A_t = a] \end{aligned}$$

  - Bellmanova rovnica očakávania

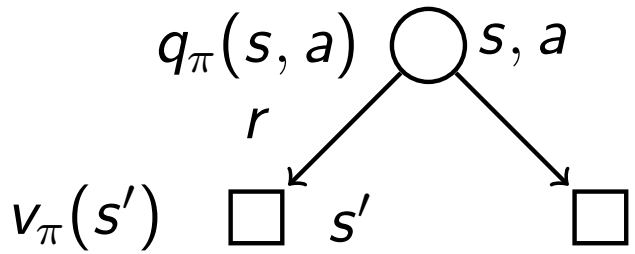
Bellmanova rovnica očakávania pre v_π



- Potrebné zohľadniť všetky agentove možnosti
  - určiť priemernú akciu pomocou strednej hodnoty

$$v_\pi(s) = \sum_{a \in A} \pi(a|s) q_\pi(s,a)$$

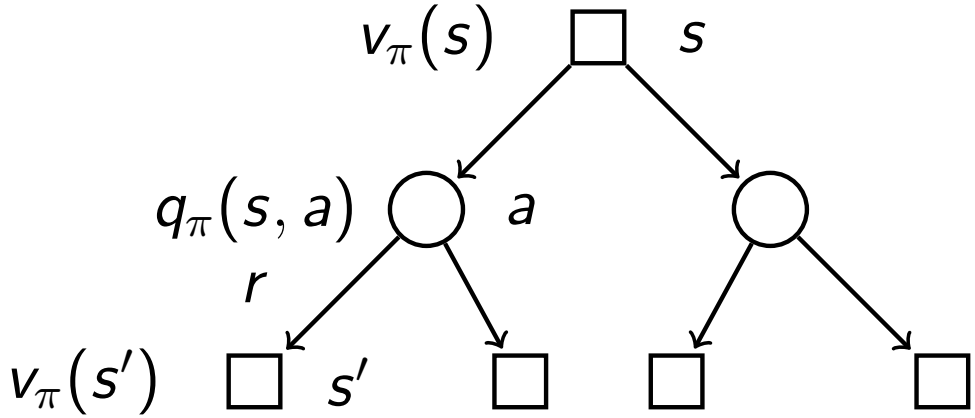
Bellmanova rovnica očakávania pre q_π



- Potrebné zohľadniť všetky reakcie prostredia
  - určiť priemernú reakciu pomocou strednej hodnoty

$$q_\pi(s,a) = r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) v_\pi(s')$$

Bellmanova rovnica očakávania pre v_π (2)



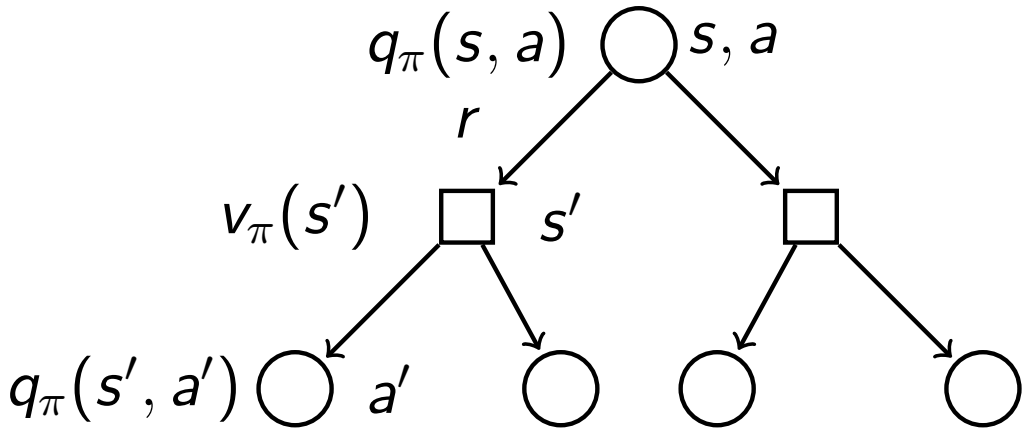
$$
\begin{aligned}
v_\pi(s) &= \sum_{a \in A} \pi(a|s) q_\pi(s,a) \\
&= \sum_{a \in A} \pi(a|s) \left(r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) v_\pi(s') \right)
\end{aligned}
$$

- Spriemerňuje cez všetky možnosti, pričom možnosti váži ich pravdepodobnosťami

Rôzne tvary BR očakávania pre v_π

$$
\begin{aligned}
v_\pi(s) &= \sum_{a\in A}\pi(a|s)\left(r(s,a)+\gamma\sum_{s'\in S}p(s'|s,a)v_\pi(s')\right)\\
v_\pi(s) &= \sum_{a\in A}\pi(a|s)\left(r(s,a)\sum_{s'\in S}p(s'|s,a)+\gamma\sum_{s'\in S}p(s'|s,a)v_\pi(s')\right)\\
&= \sum_{a\in A}\pi(a|s)\sum_{s'\in S}p(s'|s,a)\left(r(s,a)+\gamma v_\pi(s')\right)\\
v_\pi(s) &= \sum_{a\in A}\pi(a|s)\left(\sum_{r\in R}r\sum_{s'\in S}p(s',r|s,a)+\right.\\
&\qquad \left.\gamma\sum_{s'\in S}\sum_{r\in R}p(s',r|s,a)v_\pi(s')\right)\\
&= \sum_{a\in A}\pi(a|s)\sum_{s'\in S}\sum_{r\in R}p(s',r|s,a)\left(r+\gamma v_\pi(s')\right)
\end{aligned}
$$

Bellmanova rovnica očakávania pre q_π (2)

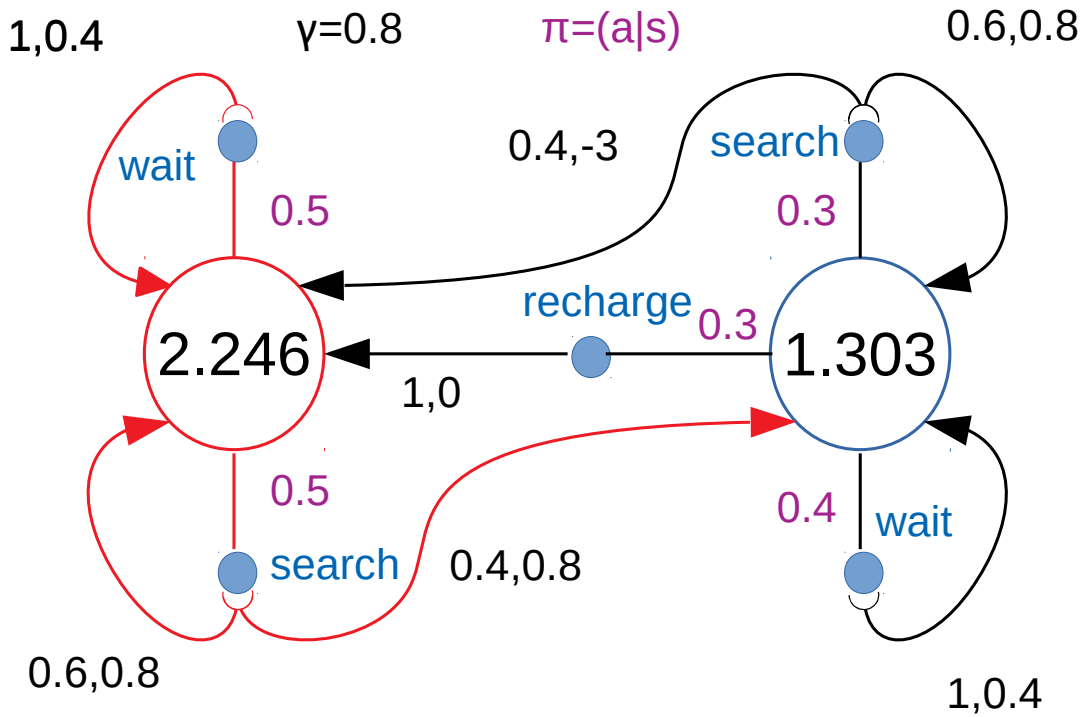


$$
\begin{aligned}
q_\pi(s,a) &= r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) v_\pi(s') \\
&= r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) \sum_{a' \in A} \pi(a'|s') q_\pi(s',a')
\end{aligned}
$$

Rôzne tvary BR očakávania pre q_π

$$
\begin{aligned}
q_\pi(s,a) &= r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) \sum_{a' \in A} \pi(a'|s') q_\pi(s',a') \\
q_\pi(s,a) &= r(s,a) \sum_{s' \in S} p(s'|s,a) + \gamma \sum_{s' \in S} p(s'|s,a) \sum_{a' \in A} \pi(a'|s') q_\pi(s',a') \\
&= \sum_{s' \in S} p(s'|s,a) \left(r(s,a) + \gamma \sum_{a' \in A} \pi(a'|s') q_\pi(s',a') \right) \\
q_\pi(s,a) &= \sum_{r \in R} r \sum_{s' \in S} p(s',r|s,a) + \\
&\quad \gamma \sum_{s' \in S} \sum_{r \in R} p(s',r|s,a) \sum_{a' \in A} \pi(a'|s') q_\pi(s',a') \\
&= \sum_{r \in R} \sum_{s' \in S} p(s',r|s,a) \left(r + \gamma \sum_{a' \in A} \pi(a'|s') q_\pi(s',a') \right)
\end{aligned}
$$

Príklad: Bellmanova rovnica očakávania



$$2.246 = 0.5 * [0.6 * (0.8 + 0.8 * 2.246) + 0.4 * (0.8 + 0.8 * 1.303)] + 0.5 * [1 * (0.4 + 0.8 * 2.246) + 0]$$

$$v_\pi(s) = \sum_{a \in A} \pi(a|s) \sum_{s' \in S} p(s'|s,a) \left(r(s,a) + \gamma v_\pi(s')\right)$$

Optimálny výber akcií

- Cieľom je výber akcií, ktorý maximalizuje G_t
- Hodnotová funkcia stavu umožňuje parciálne usporiadanie výberových politík

$\pi \geq \pi'$ ak $v_\pi(s) \geq v_{\pi'}(s)$ pre všetky $s \in S$

$\pi > \pi'$ ak $v_\pi(s) \geq v_{\pi'}(s)$ a $v_\pi(s_1) > v_{\pi'}(s_1)$ pre všetky $s \in S$ a nejaký $s_1 \in S$

? ak $v_\pi(s_1) < v_{\pi'}(s_1)$ a $v_\pi(s_2) > v_{\pi'}(s_2)$ pre $s_1, s_2 \in S$

- Optimálna polika $\pi*$
  - lepšia alebo rovnako dobrá ako každá iná
  - vždy existuje (môže ich byť viac)

Optimálne hodnotové funkcie

- $v_*(s)$, $q_*(s, a)$ - hodnotové funkcie pri použití optimálneho spôsobu výberu akcií
  - všetky optimálne politiky produkujú rovnaké funkcie $v_*(s)$ a $q_*(s, a)$
- $v_*(s)$ je maximum hodnotovej funkcie stavu pri uvažovaní všetkých možných politík

$$v_*(s) = \max_\pi v_\pi(s)$$

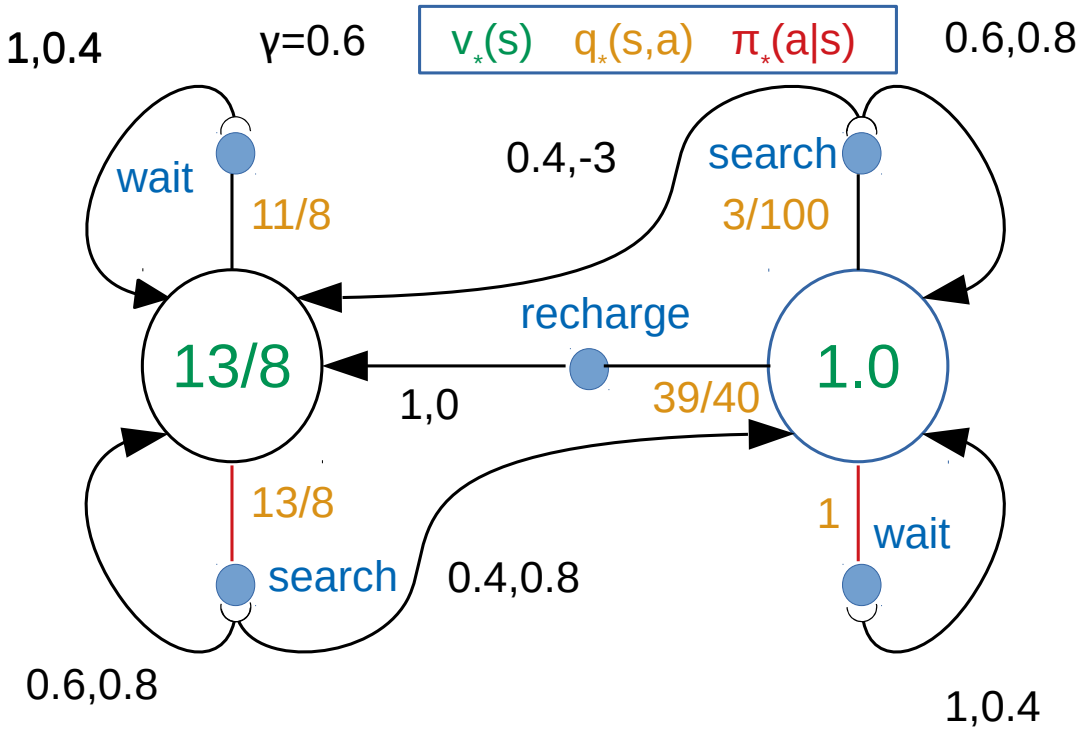
- $q_*(s, a)$ je maximum hodnotovej funkcie akcie pri uvažovaní všetkých možných politík

$$q_*(s, a) = \max_\pi q_\pi(s, a)$$

Bellmanova funkcia optimality

$$
\begin{aligned}
v_*(s) &= \max_a q_*(s,a) \\
&= \max_a \left(r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) v_*(s') \right) \\
&= \max_a \sum_{s' \in S} p(s'|s,a) \left(r(s,a) + \gamma v_*(s') \right) \\
&= \max_a \sum_{s' \in S} \sum_{r \in R} p(s', r|s,a) \left(r + \gamma v_*(s') \right) \\
q_*(s,a) &= r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) v_*(s') \\
&= r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) \max_{a'} q_*(s',a') \\
&= \sum_{s' \in S} p(s'|s,a) \left(r(s,a) + \gamma \max_{a'} q_*(s',a') \right) \\
&= \sum_{r \in R} \sum_{s' \in S} p(s', r|s,a) \left(r + \gamma \max_{a'} q_*(s',a') \right)
\end{aligned}
$$

Príklad: Bellmanova funkcia optimality



$$13/8 = 0.6 * (0.8 + 0.6 * 13/8) + 0.4 * (0.8 + 0.6 * 1.0)$$

$$1.0 = \max(3/100, 39/40, 1.0)$$

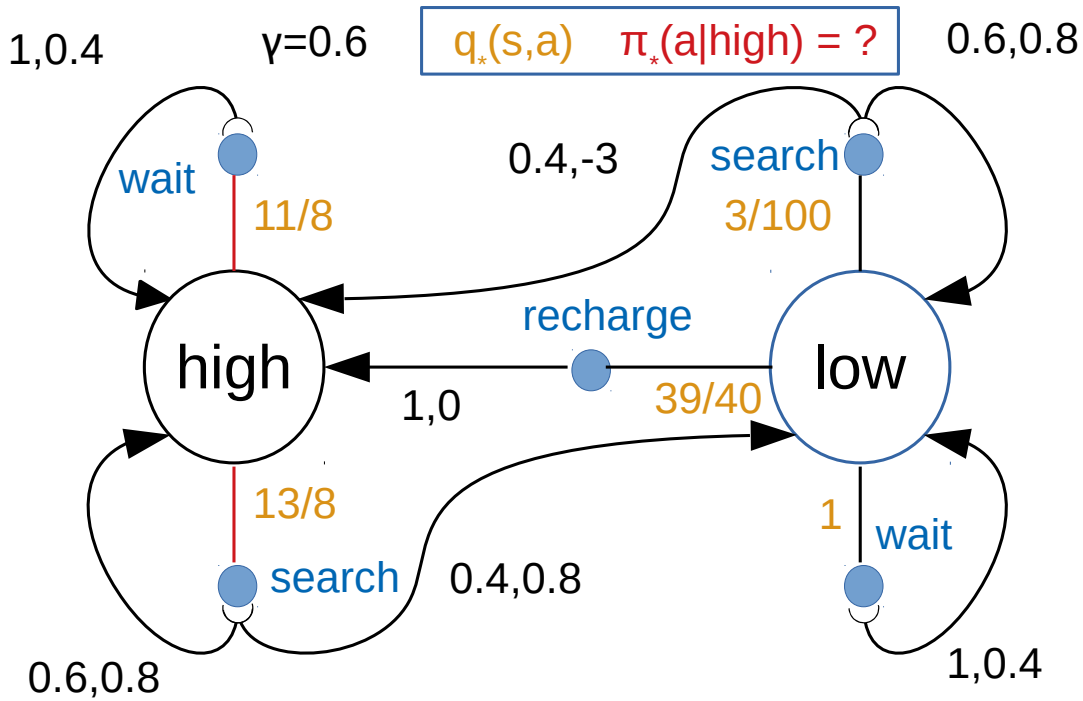
$$q_*(s,a) = \sum_{s' \in S} p(s'|s,a) \left(r(s,a) + \gamma \max_a q_*(s',a') \right)$$

$$v_*(s) = \max_a q_*(s,a)$$

Nájdenie optimálnej politiky pomocou q_*

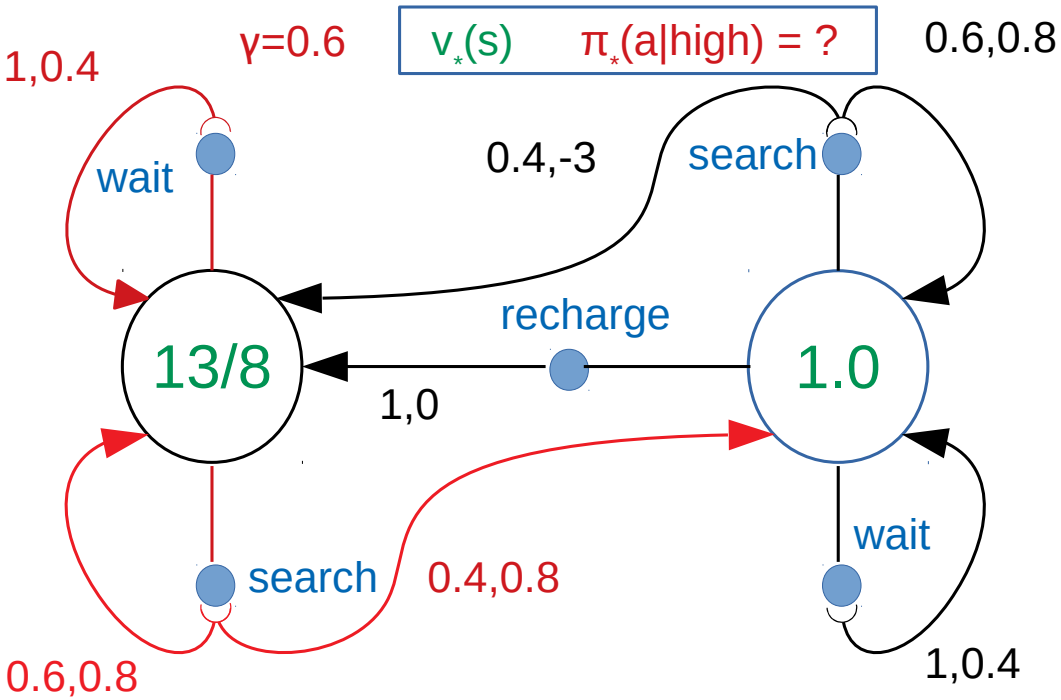
- Známe $q_*(s, a)$
  - bezprostredný výber vhodnej akcie

$$\pi * (a|s) = \begin{cases} 1, & ak\ a = argmax_a\ q(s, a) \\ 0, & inak \end{cases}$$



Nájdenie optimálnej politiky pomocou v_*

- Známe $v_*(s)$
  - prehľadávanie všetkých možností (akcií prístupných v danom stave)
  - hľadanie do hĺbky 1 (obmedzené iba na jeden krok)
  - výber najlepšej možnosti (greedy princíp výberu)



Použiteľnosť explicitného riešenia

- Explicitné riešenie Bellmanových rovníc vyžaduje splnenie podmienok
  1. dynamika prostredia je známa
  2. dostatok výpočtových zdrojov
  3. Markovova vlastnosť
- Často podmienky nie sú splnené
- Bellmanove rovnice optimálnosti sú nelineárne
  - vo všeobecnosti neexistuje riešenie v uzavretej forme
- Metódy pre aproximatívne riešenie
  - výpočtová náročnosť (aproximácia výpočtového procesu)
  - pamäťová náročnosť (aproximácia reprezentácie)