Strojové učenie II

prednáška 2 – Markovovské rozhodovacie procesy a Bellmanove rovnice

Ing. Ján Magyar, PhD.
Katedra kybernetiky a umelej inteligencie
Technická univerzita v Košiciach
2021/2022 letný semester

Recyklačný robot

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

Markovoská vlastnosť

- „Budúcnosť pri danej prítomnosti nezáleží od minulosti“

$$P[S_{t+1}|S_1, \dots, S_t] = P[S_{t+1}|S_t]$$

- vlastnosť stochastických procesov, modelujúcich sekvenciu stavov
- aktuálny stav obsahuje všetky relevantné informácie pre predikciu budúceho stavu
  - nezáleží na postupnosti, ktorá viedla k súčasnému stavu
  - históriu minulých stavov môžeme ignorovať ak poznáme aktuálny stav

Markovovský proces (reťazec)

- stochastický model sekvencie možných stavov
  - spĺňa Markovoskú vlastnosť

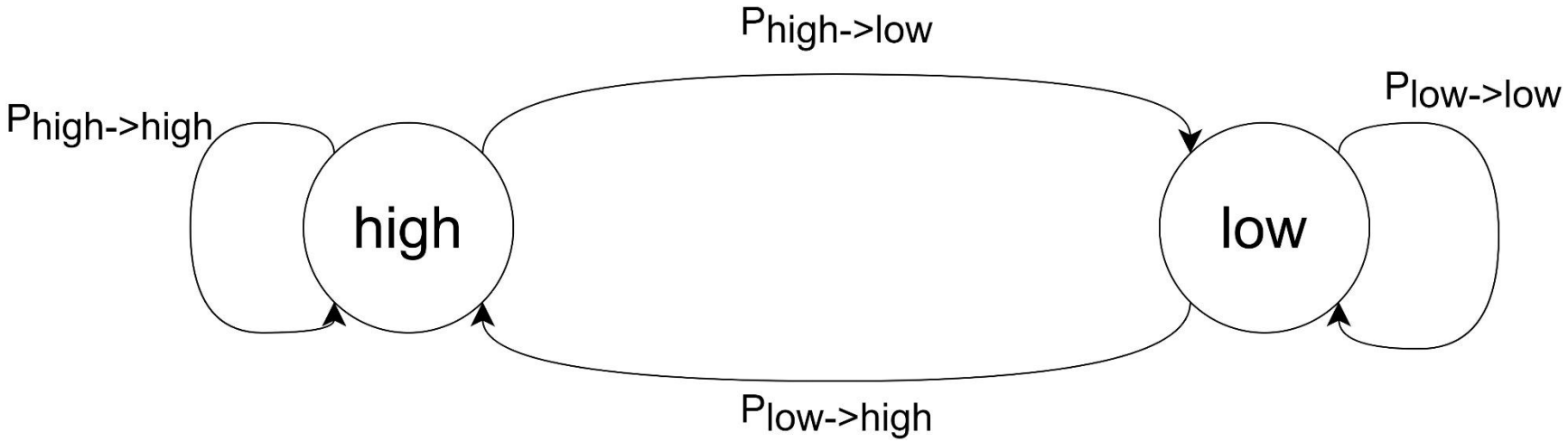
- Markovovský proces je dvojica *(S, P)*, kde:
  - *S* je množina stavov
  - *P* je pravdepodobnostná prechodová matica $P_{ss'} = P[S_{t+1} = s' | S_t = s]$

- prechodová matica definuje pravdepodobnosti prechodov medzi dvojicami stavov

Recyklačný robot ako MP

- možná sekvencia: *high*, *high*, *low*, *high*, *low*, ...

$$P = \begin{bmatrix} P_{high,high} & P_{high,low} \\ P_{low,high} & P_{low,low} \end{bmatrix}$$



Markovovský proces s odmenou

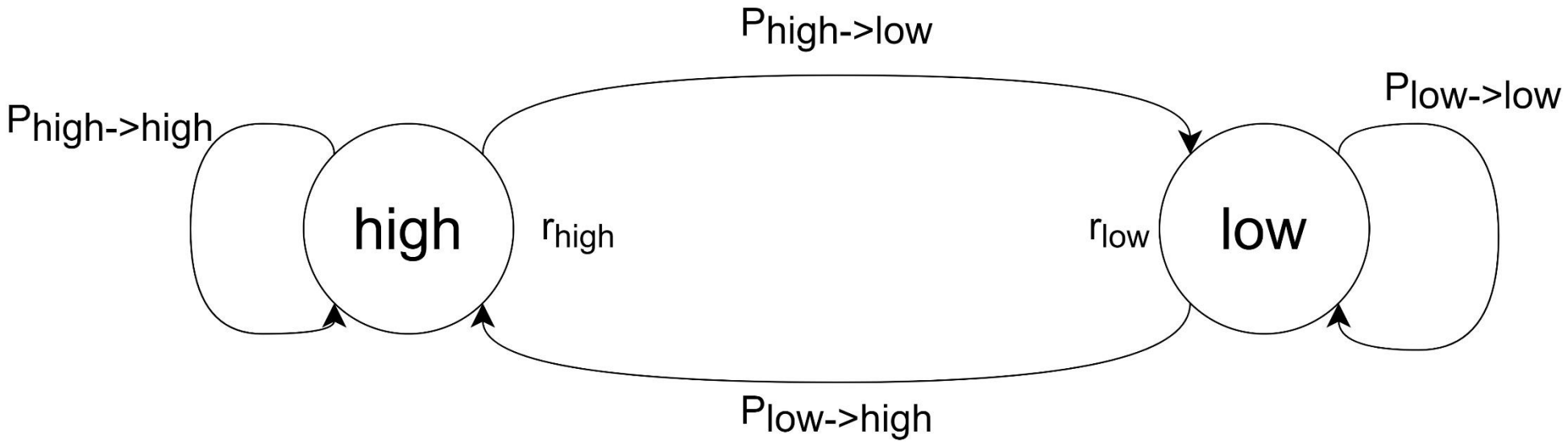
- Markovovský proces rozšírime o hodnoty (odmeny)

- Markovovský proces s odmenou je n-tica *(S, P, R)*, kde:
  - *S* je množina stavov
  - *P* je pravdepodobnostná prechodová matica $P_{ss'} = P[S_{t+1} = s' | S_t = s]$
  - R je funkcia odmeny $R_s = E[R_{t+1} | S_t = s]$

- priradíme hodnotu (odmenu) každému stavu
- akumulácia odmien získavaných počas sekvencie prechádzania stavmi

Recyklačný robot ako MRP

- možná sekvencia: *high*, r_{high}, *high*, r_{high} , *low*, r_{low} , *high*, r_{high} , *low*, r_{low} , *low*, ...
- akumulácia: 0, r_{high} , $2\ r_{high}$, $2\ r_{high} + r_{low}$, ...

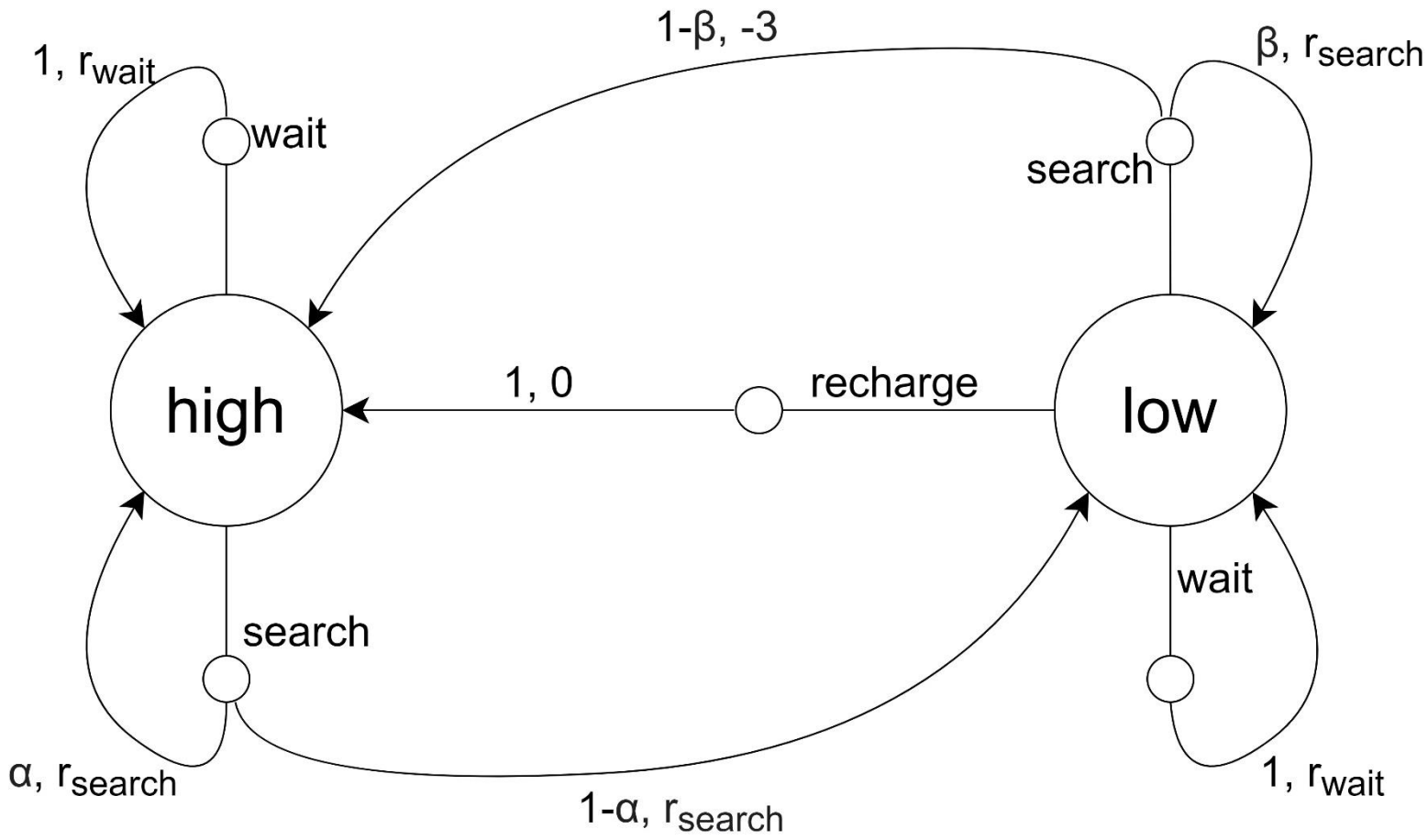


Markovovský rozhodovací proces

- Markovovský proces s odmenou rozšírime o akcie
- modelovanie rozhodovacích procesov v situáciách, kde výsledky sú čiastočne náhodné a čiastočne pod kontrolou toho, kto prijíma rozhodnutie

- Markovovský rozhodovací proces je n-tica *(S, A, P, R)*, kde:
  - *S* je množina stavov
  - *A* je množina akcií
  - *P* je pravdepodobnostná prechodová matica $P_{ss'}^{a} = P[S_{t+1} = s' | S_t = s, A_t = a]$
  - R je funkcia odmeny $R_s^a = E[R_{t+1} | S_t = s, A_t = a]$

Recyklačný robot ako MDP



Zdroj: 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 $a \in A, s \in S$

Dynamika prostredia

- z dynamiky prostredia vieme určiť charakteristiky
  - pravdepodobnosti zmeny stavov

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

  - očakávaná odmena pre dvojicu stav-akcia

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

Odmena

- cieľom je maximalizovať očakávanú kumulatívnu odmenu
- dôraz nie je na bezprostrednú odmenu
- odmena vyjadruje cieľ, ktorý má byť dosiahnutý, teda maximalizovaním odmeny dosiahneme aj cieľ
- odmena komunikuje, čo má byť dosiahnuté, nie ako

Kumulatívna odmena

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

- epizodická úloha
  - T je konečný čas, interakcia s agentom končí
  - sekvencia stavov epizódy končí v koncovom stave (po ňom nasleduje počiatočný stav ďalšej epizódy)
  - 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

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

- diskontný faktor $\gamma \in < 0,1 >$
  - $\gamma = 0$ – uvažovanie iba okamžitej odmeny
  - čím je väčšia γ, tým dlhšiu dobu berie agent do úvahy
- $\sum^{\infty}$ má konečnú hodnotu, ak odmeny sú ohraničené ($R_i \leq R$):

$$G_t = \sum_{k=t+1}^{\infty} \gamma^{k-t-1} R_k \leq R \sum_{k=0}^{\infty} \gamma^k = R \frac{1}{1-\gamma}$$

Výhody použitia γ

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots$$
$$G_t = R_{t+1} + \gamma(R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} + \cdots)$$
$$G_t = R_{t+1} + \gamma G_{t+1}$$

- matematicky vhodný spôsob vyjadrenia
- vyhnutie sa nekonečne veľkej kumulatívnej odmene
- lepšia charakterizácia úloh v prípade
  - neurčitosti odmeny v budúcnosti
  - že bezprostredná odmena je zaujímavejšia ako vzdialená
  - že ľudské chovanie preferuje bezprostrednú odmenu

Voľba akcií – politika

- politika ja spôsob, ktorým agent volí svoje akcie
- formálne je to distribúcia pravdepodobnosti nad akciami v určitom stave:

$$\pi(a|s) = P[A_t = a|S_t = s]$$

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

Hodnotová funkcia stavu

$$V_\pi(s) = E_\pi[G_t|S_t = s] = E_\pi[R_{t+1} + \gamma G_{t+1}|S_t = s]$$

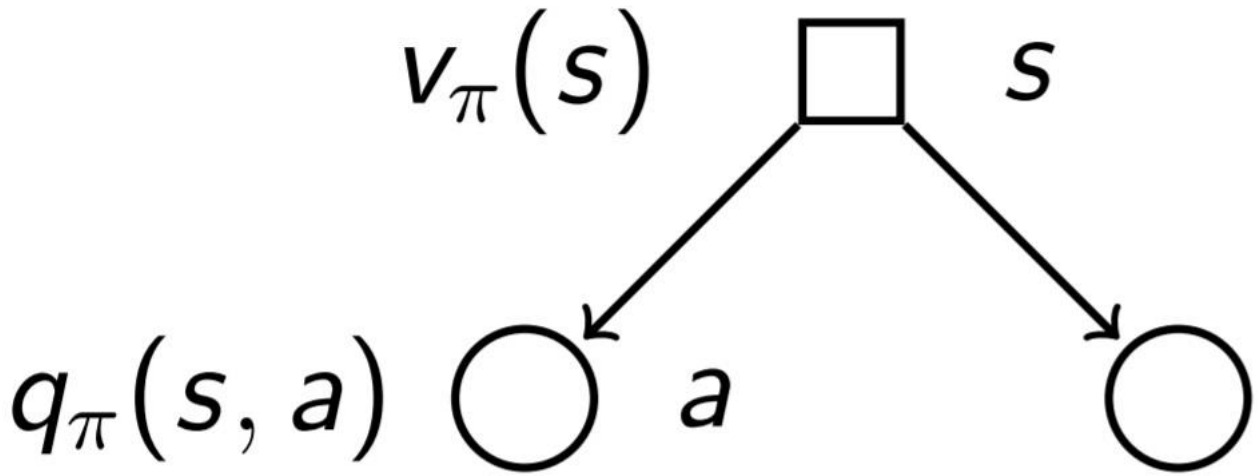
- ako výhodné je byť v danom stave
- očakávaná kumulatívna odmena sekvencie začínajúcej v danom stave pri politike π

Hodnotová funkcia akcie

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

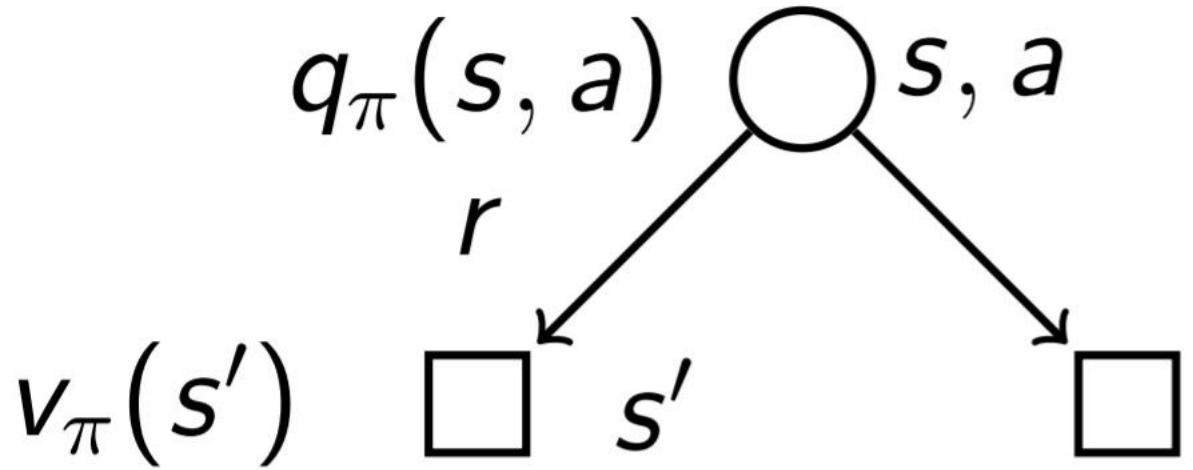
- ako výhodné je v danom stave použiť danú akciu pri politike π
- očakávaná kumulatívna odmena sekvencie začínajúcej v danom stave danou akciou, ak voľba nasledujúcich akcií je podľa politiky π

Bellmanova rovnica očakávania pre v_π



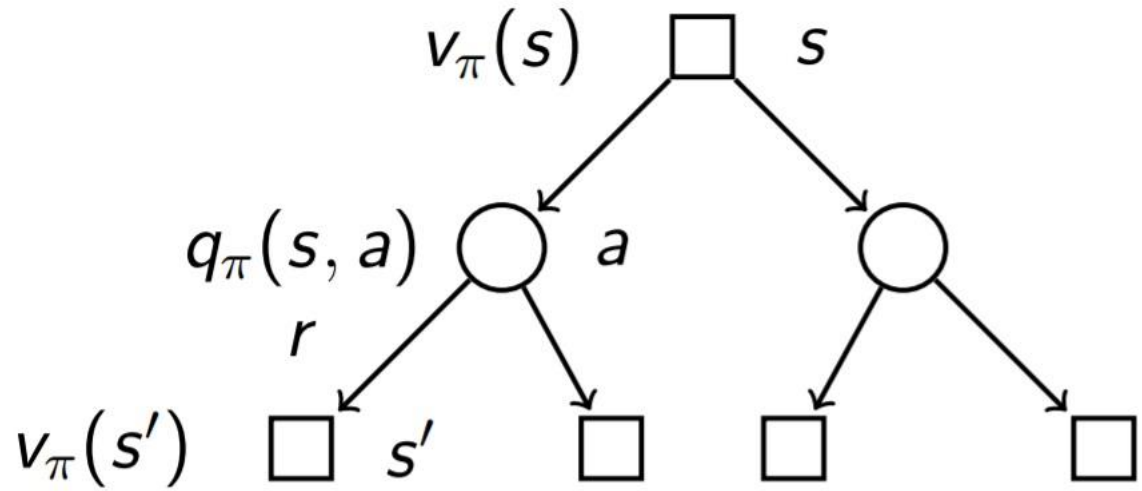
$$v_\pi(s) = \sum_{a \in A} \pi(a|s) q_\pi(s,a)$$

Bellmanova rovnica očakávania pre q_π



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

Bellmanova rovnica očakávania pre v_π (2)



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

Bellmanova rovnica očakávania pre v_π (3)

$$v_\pi(s) = E_\pi[G_t|S_t = s] = E_\pi[R_{t+1} + \gamma G_{t+1}|S_t = s]$$

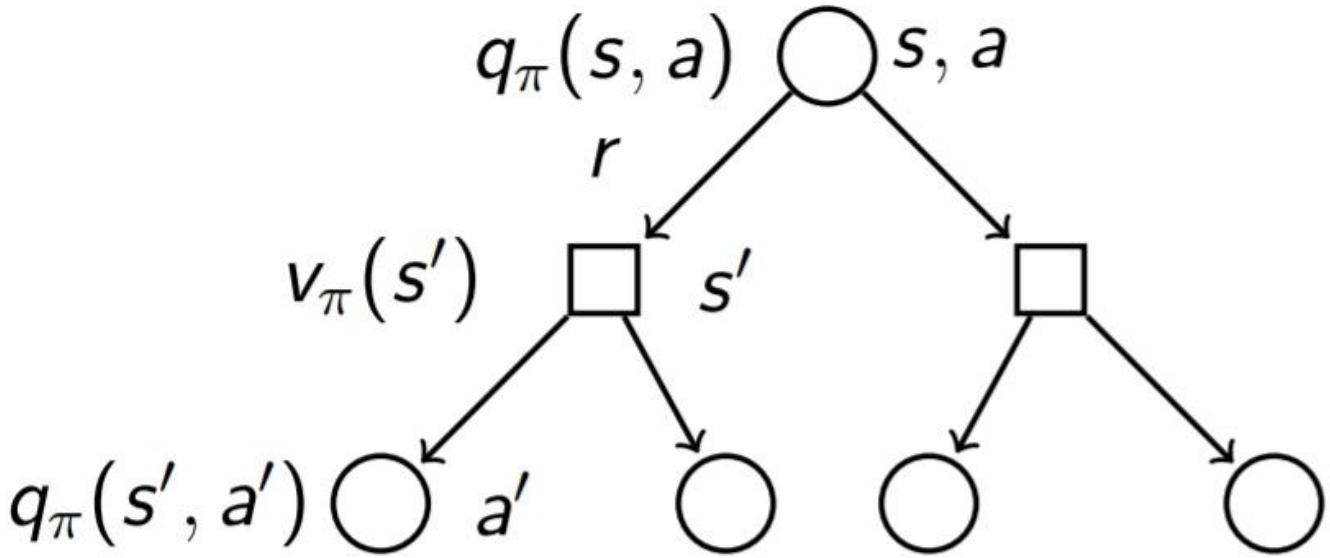
$$v_\pi(s) = \sum_{a\in A} \pi(a|s) \left(r(s,a) + \gamma \sum_{s'\in S} p(s'|s,a) v_\pi(s') \right)$$

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

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

$$v_\pi(s) = E_\pi[R_{t+1} + \gamma v_\pi(S_{t+1})|S_t = s]$$

Bellmanova rovnica očakávania pre q_π (2)



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

Bellmanova rovnica očakávania pre q_π (3)

$$q_\pi(s,a) = E_\pi[R_{t+1} + \gamma G_{t+1} | S_t = s, A_t = a]$$

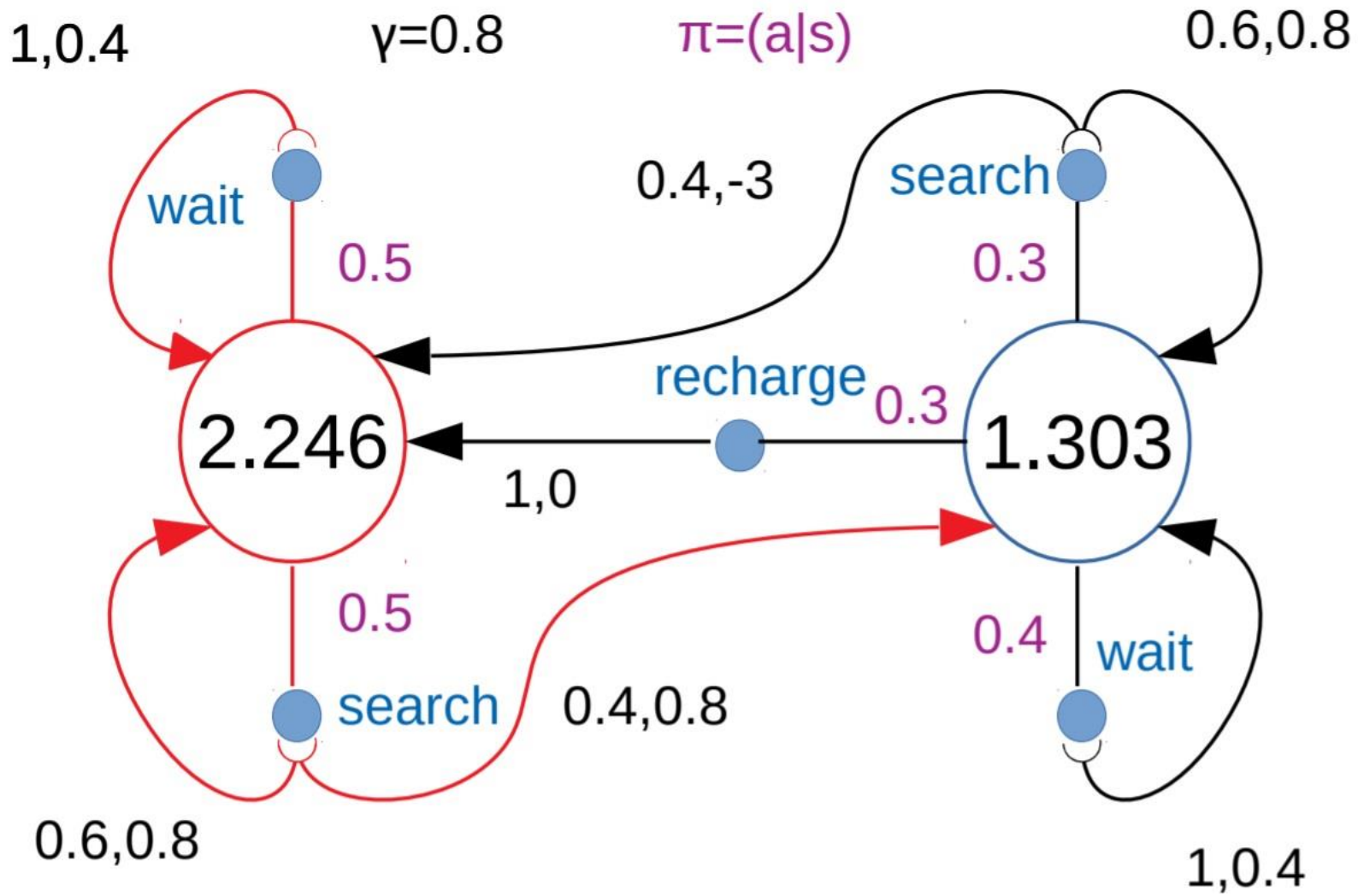
$$q_\pi(s,a) = r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) \sum_{a' \in A} \pi(a'|s') q_\pi(s',a')$$

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

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

$$q_\pi(s,a) = E_\pi[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1}) | S_t = s, A_t = a]$$

Príklad: Bellmanova rovnica očakávania



Optimálny výber akcií

- cieľom je vybrať akcie, ktoré maximalizujú kumulatívnu odmenu
- hodnotová funkcia stavu umožňuje parciálne usporiadanie výberových politík

  $\pi \geq \pi'$ ak platí, že $v_\pi(s) \geq v_{\pi'}(s)$ pre všetky $s \in S$
- optimálna politika π^*
  - lepšia alebo rovnako dobrá ako ostatné politiky
  - vždy existuje
  - môže ich byť viac

Optimálne hodnotové funkcie

- $v_*(s)$, $q_*(s,a)$ – hodnotové funkcie pri použití optimálnej politiky π^*
  - 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 – stavy

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

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

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

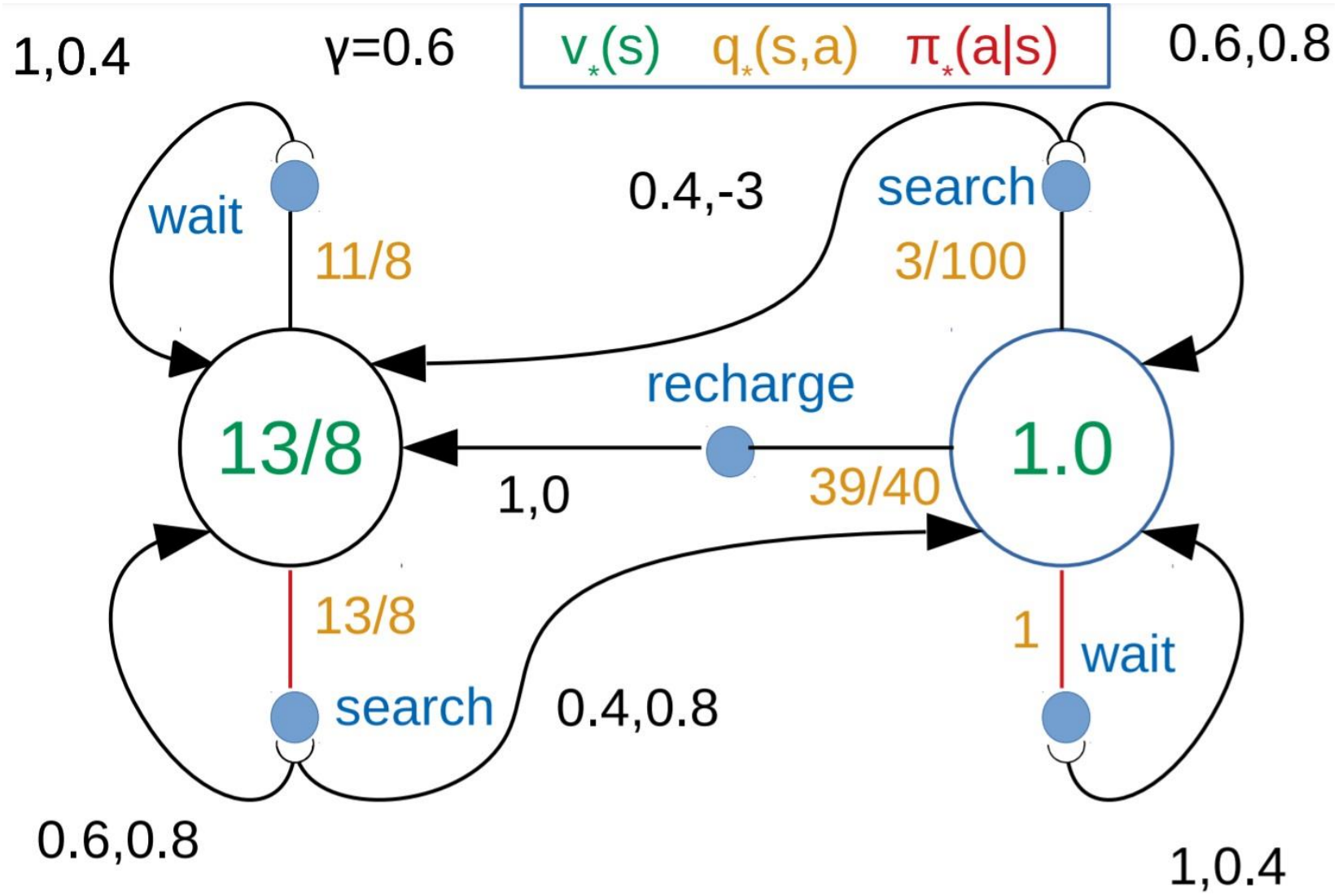
Bellmanova funkcia optimality – akcie

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

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

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

Príklad: Bellmanova funkcia optimality



Nájdenie optimálnej politiky

- ak poznáme $q_*(s,a)$, bezprostredný výber vhodnej akcie:

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

- ak poznáme $v_*(s)$
  - prehľadávanie všetkých akcií prípustných 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
  - hry s plnou informáciou (backgammon asi 10^{20} stavov)
- Bellmanove rovnice optimálnosti nie sú nelineárne
  - vo všeobecnosti neexistuje riešenie v uzavretej forme
- metódy pre aproximatívne riešenie
  - výpočtová náročnosť
  - pamäťová náročnosť