Učenie posilňovaním

(učenie odmenou a trestom)

Juraj Gazda

iislab.kpi.fei.tuke.sk

Technical University of Košice
Department of Computers and Informatics

Prednášky: Literatúra

Pieter Abbel - predmet: Artificial Intelligence, Berkeley

Sutton, Richard S., and Andrew G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Prednášky: Koncept

Strojové učenie

Učenie s učiteľom

Učenie bez učiteľa

Učenie odmenou a trestom

[KNIHA] **Reinforcement learning**: An introduction
RS Sutton, AG Barto - 1998 - cell.com
The present book is an excellent entry point for someone who wants to understand intuitively the ideas of reinforcement learning and the general connection between its parts. It is not, however, a mathematical 'how-to'book, replete with proofs and pointers to unsolved
Citované 23457-krát Súvisiace články Všetky verzie 48 Citovať Uložiť

Prednášky: Koncept

Strojové učenie

Učenie s učiteľom

Učenie bez učiteľa

Učenie odmenou a trestom

ISI - modelovanie v deterministických prostrediach, definícia stavu
SMAD - Základná báza vedomostí pre reinforcement learning, MDP procesy, Bellman rovnice, SARSA, Q-learning, Actor-Critic prístup, Deep reinforcement learning

- Single-agent vs. multi-agent
  - kooperácia vs kompetetívnosť

- Deterministický vs stochastický priestor
  - Ďalší stav prostredia je plne určený súčasným stavom agenta a jeho akciou

- Statický vs. dynamický priestor
  - Statické prostredie sa nemení, kým agent rozmýšľa čo urobiť :)
  - Semi-dynamické: prostredie sa s časom nemení, ale miera výkonu áno

- Diskrétny vs. spojitý priestor

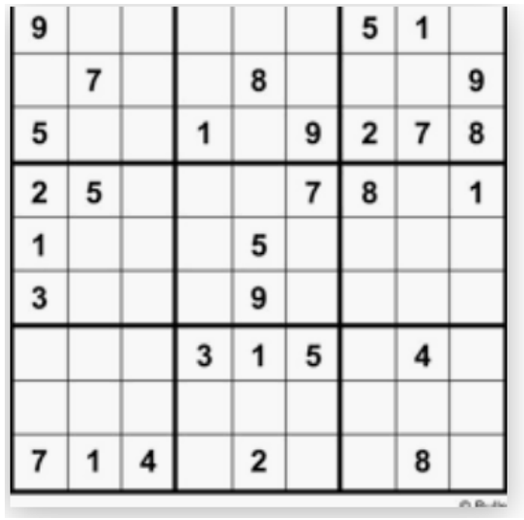
- Úplný vs. čiastočný (Fully observable vs. partially observable)
  - Senzory poskytujú úplný obraz o stavovom prostredí agenta

ISI - 2. ročník

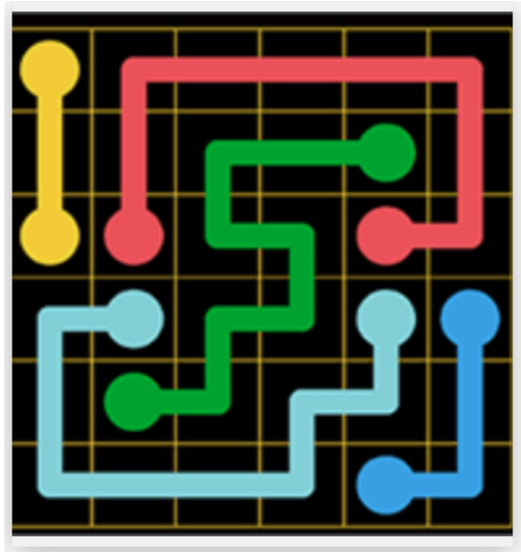
Metódy: vyhľadávanie v strome (DFS, BFS, A*), constraint satisfaction problems

+ lokálne hľadanie: hill climbing, simulované žíhanie

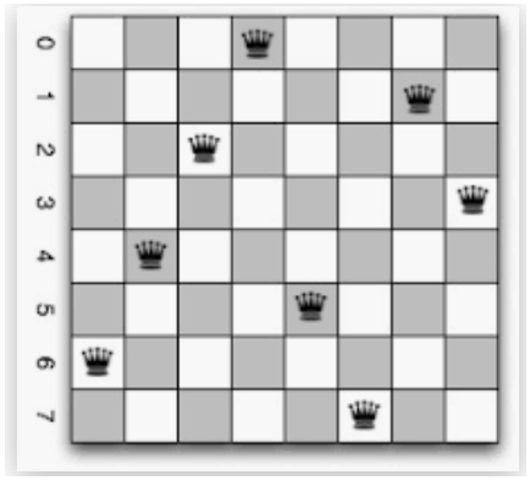
Sudoku



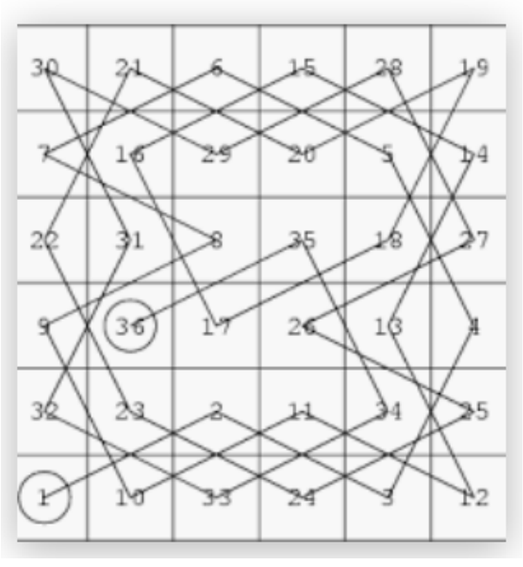
Flow free



N=Queens



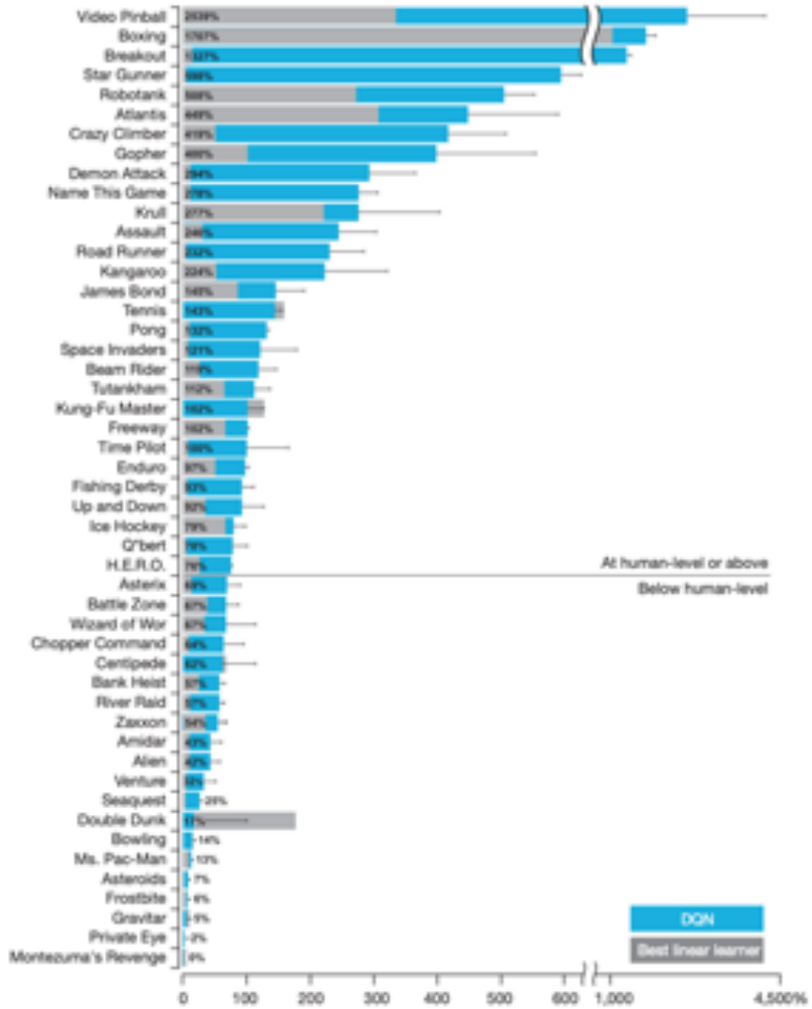
Knight's tour



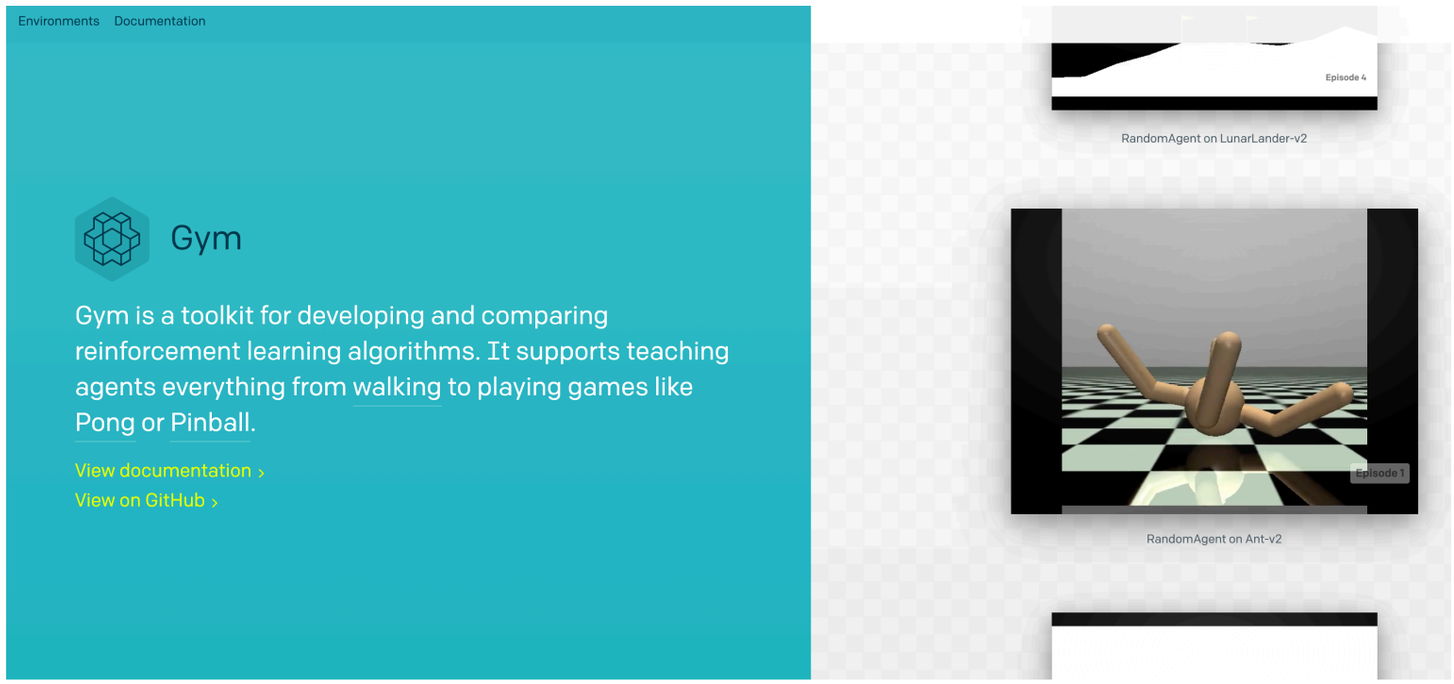
Reinforcement learning

- AlphaGo (Google DeepMind) – 2015,
- 9-dan od Chinese Weiqi Association.
- Využíva vyhľadávanie v strome a hlboké neurónové siete.
  Za účelom trénovania sa využívajú historické hry šampiónov, ako aj počítačové ťahy.
- Hry Atari využitím hlbokého reinforcement learning učenia.

- https://deepmind.com/blog/article/alphastar-mastering-re



Reinforcement learning: Súčasnosť



Reinforcement learning: Súčasnosť

AlphaGo
https://www.youtube.com/watch?v=8tq1C8spV_g
MSc work
https://www.youtube.com/watch?v=6H-FRDOgsCc
Project
https://www.youtube.com/watch?v=V1eYniJ0Rnk

Multi-armed bandit problem

Multi-armed bandit problem

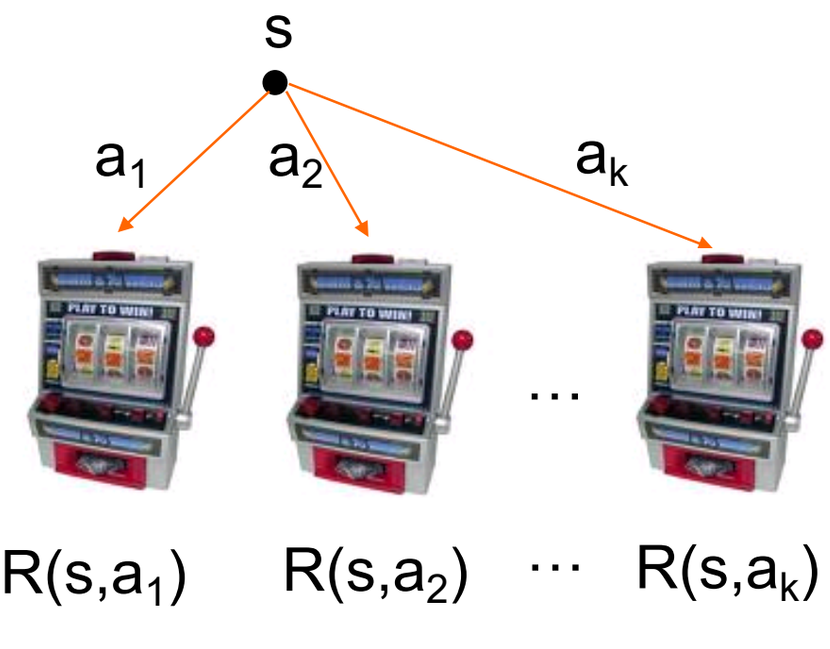
Výherný automat - bandit

Páka automatu - arm

N automatov v kasíne

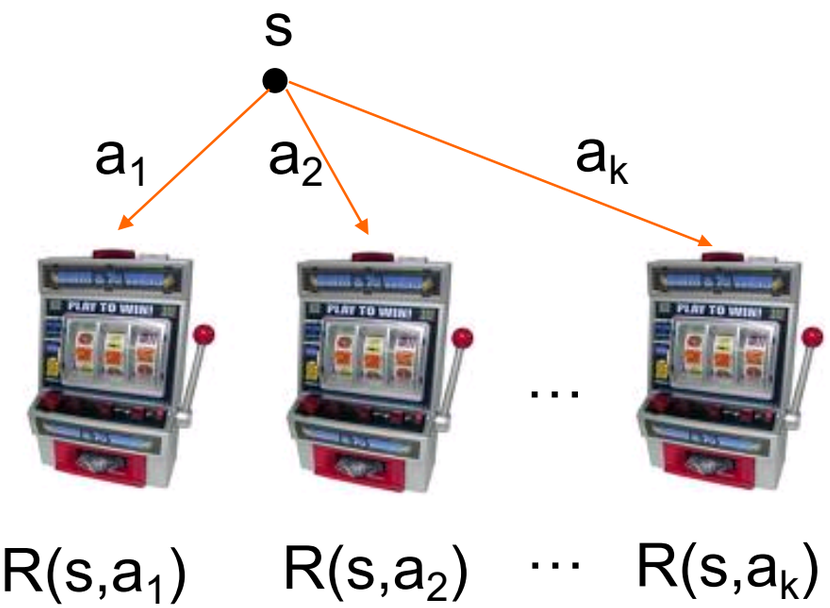
Sutton, R., Barto, A., Reinforcement learning: Introduction (2nd edition), MIT press, 2018, pp: 25-42

Multi-armed bandit problem



- Jedno-stavový problém
- Riešenie: empirické zistenie očakávaných výhier
- t.j. výber výherného automatu a hranie $R(s, a_k, s)$

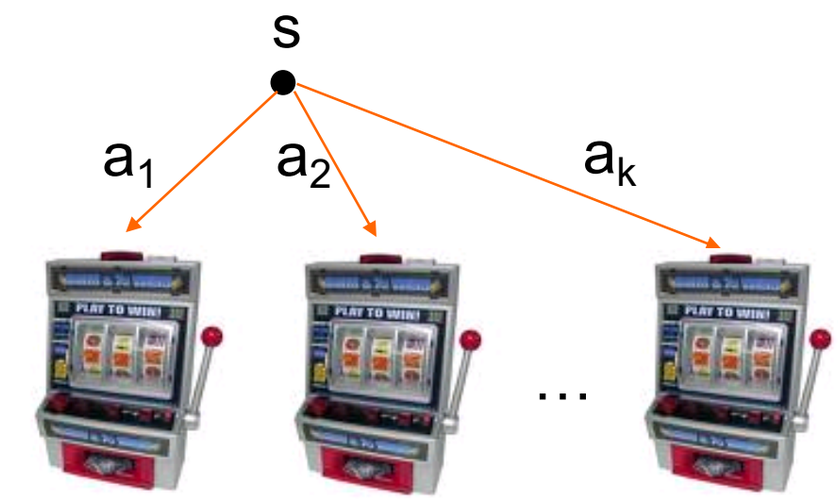
Multi-armed bandit problem



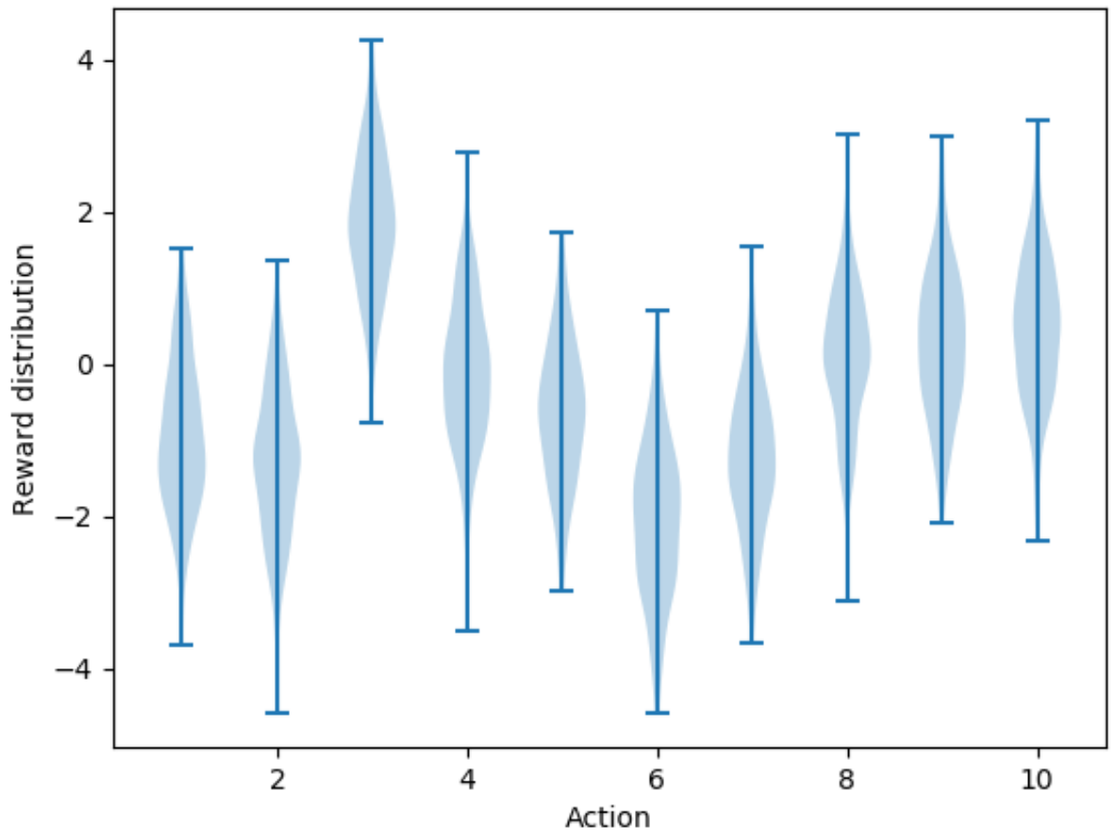
- Predpoklady: množina akcií so stochastickou odmenou *R*
- Trénovanie modelu?
  - Voľba optimálneho automatu *a* (t.j. najväčší očakávaný reward $q_*(a)$, kde $q_*(a) = \mathbb{E}[R_t | A_t = a]$
- Aplikácie:
  - Cieľenie reklamy na koncového zákazníka
  - Evaluácia a výber klinických postupov v medicíne

Vološin, M., Gazda, J., Drotár, P., Bugár, G., & Gazda, V. (2016). Spatial Real-Time Price Competition in the Dynamic Spectrum Access Markets. In Multi-Agent Systems and Agreement Technologies (pp. 217-229), Valencia, Spain, Springer, Cham.

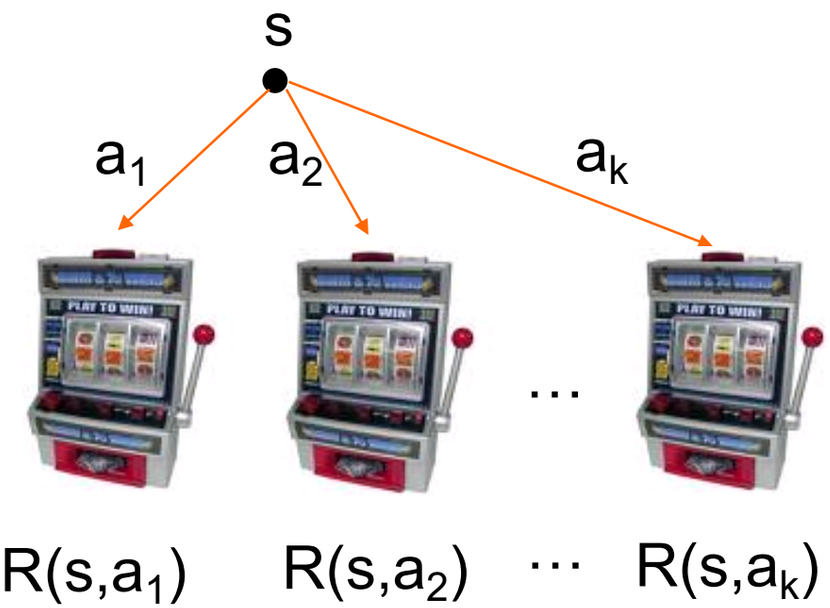
Multi-armed bandit problem



$R(s,a_1)$ $R(s,a_2)$ $\cdots$ $R(s,a_k)$



Multi-armed bandit problem: Riešenie využitím Action value prístupu



$$Q_t(a) \doteq \frac{\text{sum of rewards when } a \text{ taken prior to } t}{\text{number of times } a \text{ taken prior to } t} = \frac{\sum_{i=1}^{t-1} R_i \cdot \mathbb{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i=a}}$$

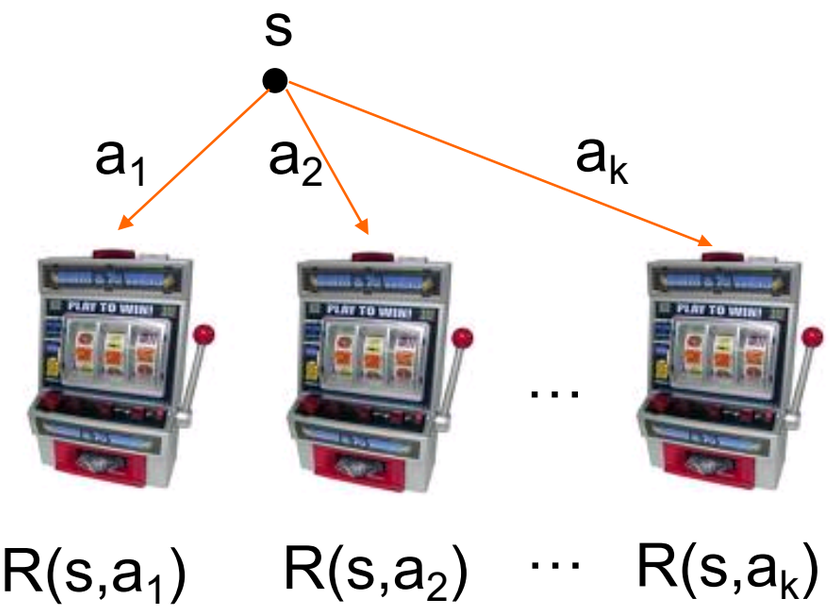
Náš cieľ, odhad charakteristík akcie, ktorá vhodne aproximuje reálne skutočnosti: $Q_t(a) \rightarrow q_*(a)$

Metodológia: výber greedy akcie (exploitation), vs. spoznávanie charakteristík akcií (exploration)

Exploitation fáza: $A_t \doteq \arg\max_a Q_t(a),$

Exploration fáza: Náhodný výber akcie *a*

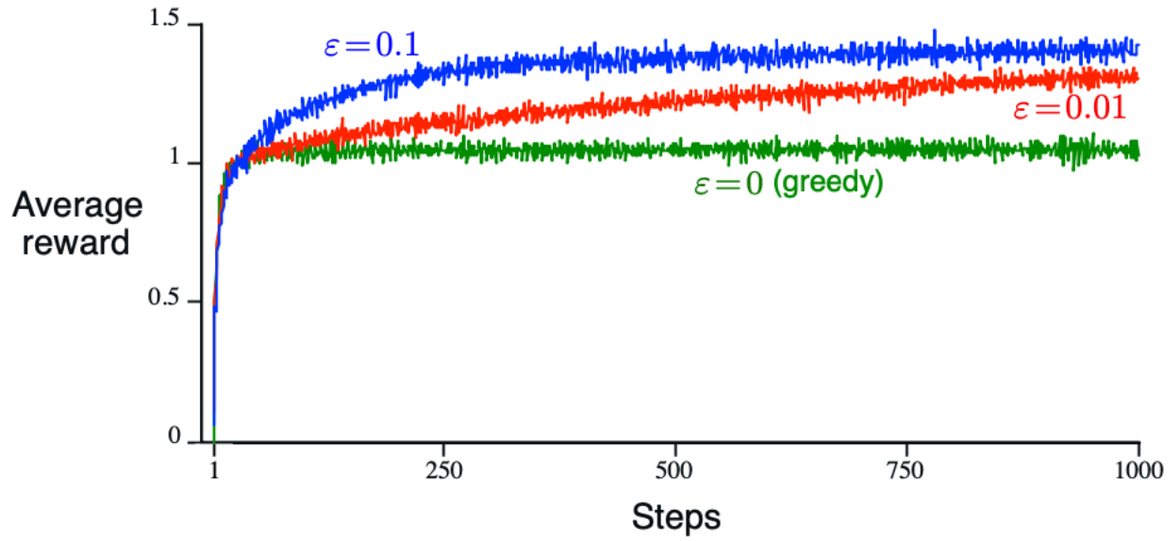
Riešenie: Epsilon-greedy explorácia

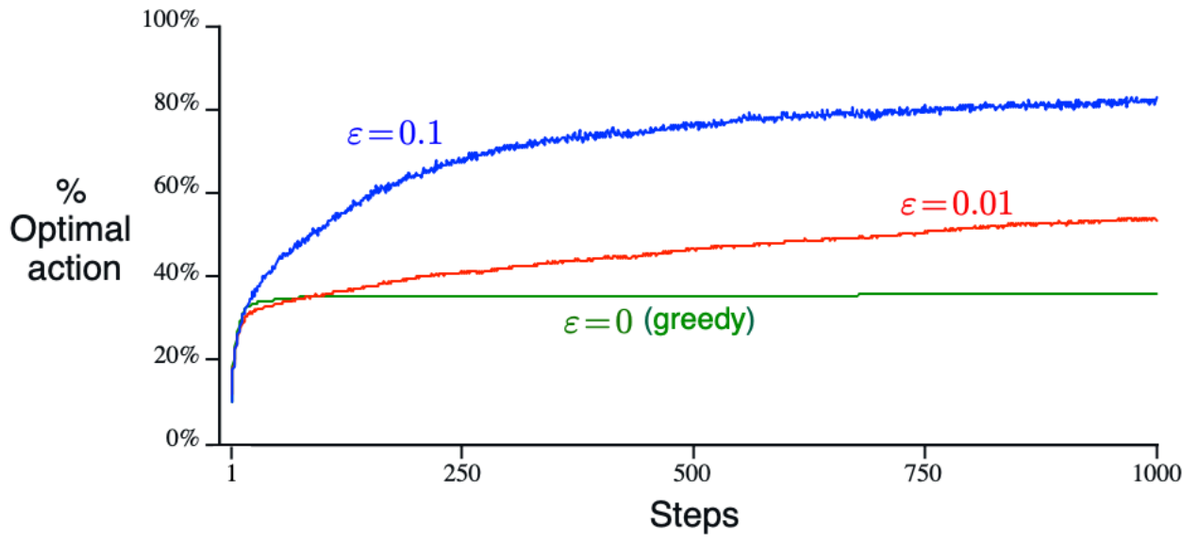


$$Q_t(a) \doteq \frac{\text{sum of rewards when } a \text{ taken prior to } t}{\text{number of times } a \text{ taken prior to } t} = \frac{\sum_{i=1}^{t-1} R_i \cdot \mathbb{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i=a}}$$

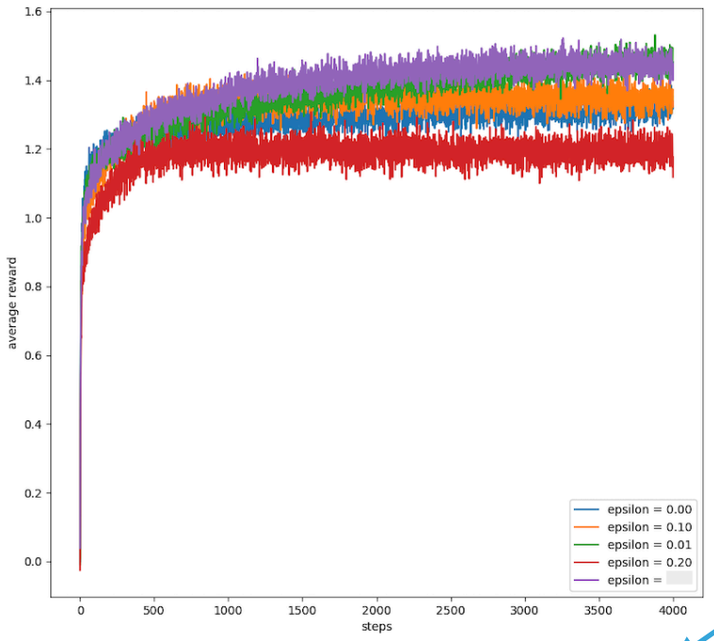
- Správajme sa greedy, avšak s pravdepodobnosťou epsilon (ϵ) hrajme náhodný automat (akciu), t.j.
  - s pravdepodobnosťou ϵ ťahaj náhodnú akciu
  - s pravdepodobnosťou 1- ϵ hraj greedy, $A_t \doteq \underset{a}{\arg\max}\, Q_t(a)$,

Multi-armed bandit problem: Riešenie



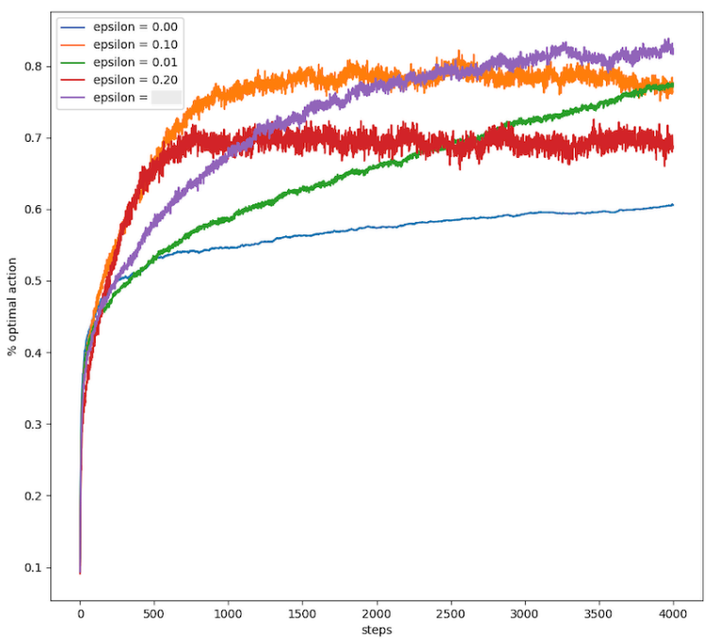


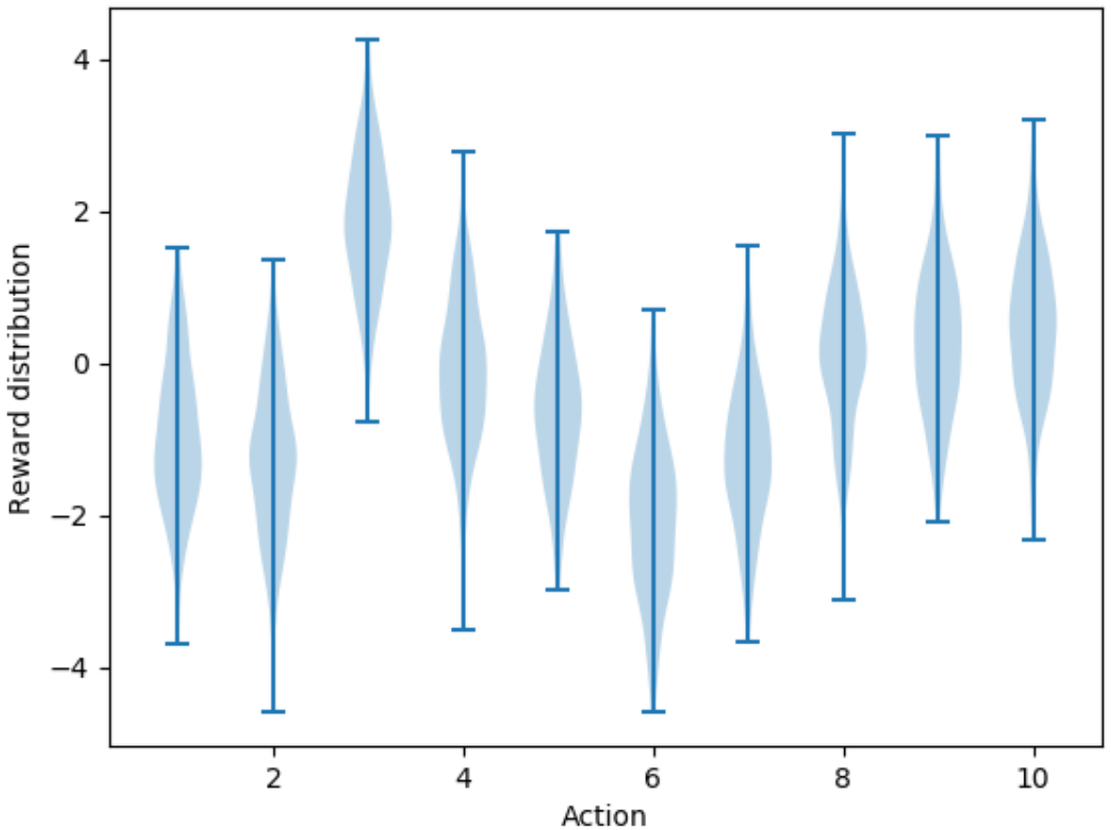
Multi-armed bandit problém: Riešenie, explor. vs. exploitation



Epsilon parameter klesá monotónne k 0, napr. $\epsilon = \epsilon * 0{,}999$

Balans medzi exploráciou a exploitáciou modelu





Inkrementálna implementácia Action Value prístupu

$$Q_n \doteq \frac{R_1 + R_2 + \cdots + R_{n-1}}{n-1}$$

Problém škálovateľnosti, ak $n \to \infty$

Riešenie: inkrementálna implementácia výpočtu Q hodnôt jednotlivých akcií

$$\begin{aligned}
Q_{n+1} &= \frac{1}{n}\sum_{i=1}^{n} R_i \\
&= \frac{1}{n}\left(R_n + \sum_{i=1}^{n-1} R_i\right) \\
&= \frac{1}{n}\left(R_n + (n-1)\frac{1}{n-1}\sum_{i=1}^{n-1} R_i\right) \\
&= \frac{1}{n}\Big(R_n + (n-1)Q_n\Big) \\
&= \frac{1}{n}\Big(R_n + nQ_n - Q_n\Big) \\
&= Q_n + \frac{1}{n}\Big[R_n - Q_n\Big],
\end{aligned}$$

$$NewEstimate \leftarrow OldEstimate \; + \; StepSize \; \Big[Target - OldEstimate\Big]$$

Sutton, R., Barto, A., Reinforcement learning: Introduction (2nd edition), MIT press, 2018, pp: 31

Inkrementálna implementácia Action Value prístupu

A simple bandit algorithm

Initialize, for $a = 1$ to k:
$Q(a) \leftarrow 0$
$N(a) \leftarrow 0$

Loop forever:
$$A \leftarrow \begin{cases} \operatorname{argmax}_a Q(a) & \text{with probability } 1-\varepsilon \quad \text{(breaking ties randomly)} \\ \text{a random action} & \text{with probability } \varepsilon \end{cases}$$
$R \leftarrow bandit(A)$
$N(A) \leftarrow N(A) + 1$
$Q(A) \leftarrow Q(A) + \frac{1}{N(A)}\left[R - Q(A)\right]$

Implementácia Multi-armed bandit využitím Q hodnôt

Gradientný prístup riešenia Multi-armed bandit problému

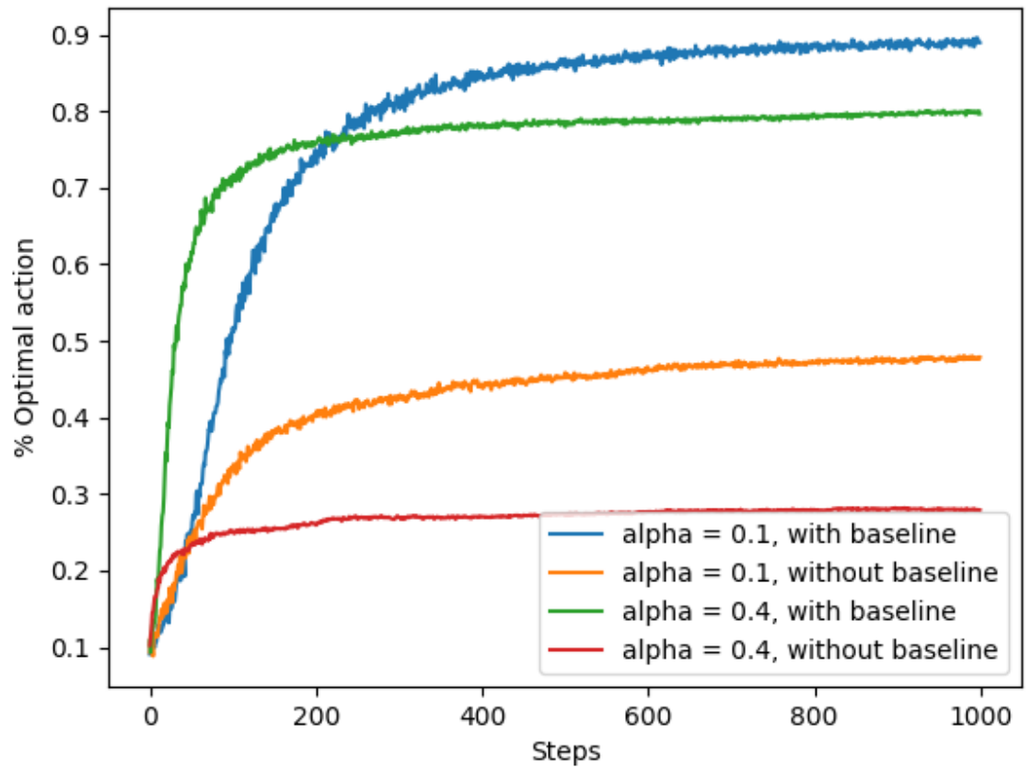
$$\Pr\{A_t = a\} \doteq \frac{e^{H_t(a)}}{\sum_{b=1}^{k} e^{H_t(b)}} \doteq \pi_t(a)$$

$$H_{t+1}(A_t) \doteq H_t(A_t) + \alpha\big(R_t - \bar{R}_t\big)\big(1 - \pi_t(A_t)\big), \qquad \text{and}$$
$$H_{t+1}(a) \doteq H_t(a) - \alpha\big(R_t - \bar{R}_t\big)\pi_t(a), \qquad \text{for all } a \neq A_t,$$

- $\pi_t(a)$ je pravdepodobnosť výberu akcie a v čase t
- $H_t(a)$ je preferencia (logit) agenta pre akciu a*
- $\overline{R_t}$ je priemerná odmena získaná do času t (baseline)

$\mathbb{E}(a_k) = 4$, 10-armed bandit

without baseline, i.e. $\overline{R_t} = 0$

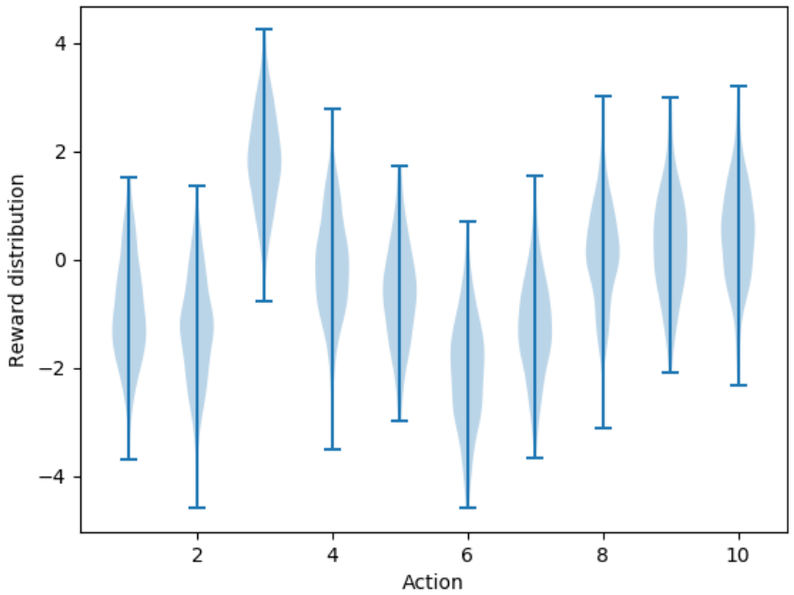


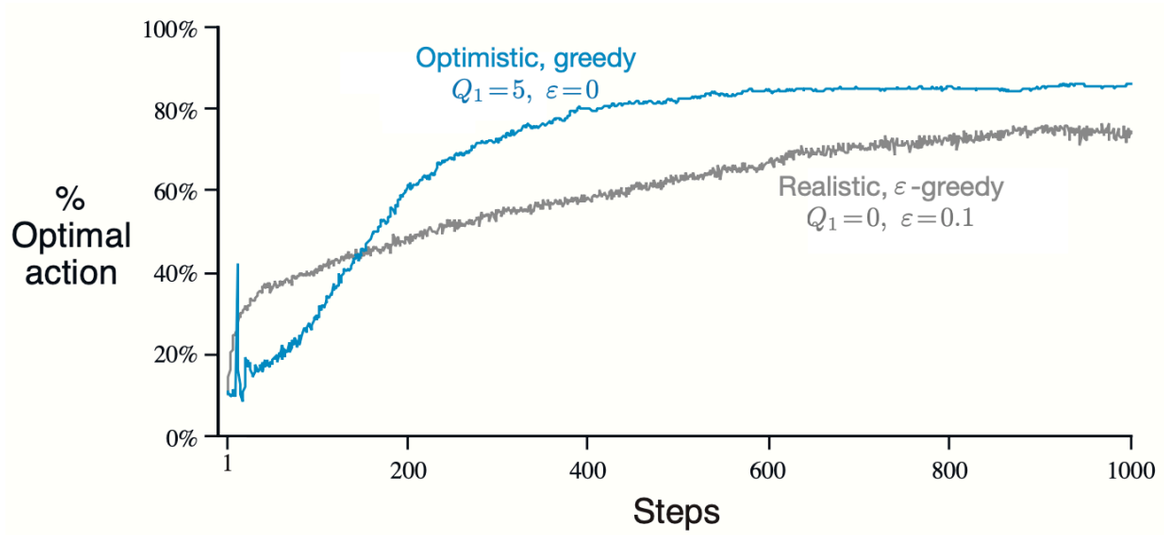
https://en.wikipedia.org/wiki/Softmax_function

Sutton, R., Barto, A., Reinforcement learning: Introduction (2nd edition), MIT press, 2018, pp: 37-41

Alternatíva k explorácii

- Optimistické počiatočné hodnoty
- napr. $Q_0(a) = 5, \forall a$





Základné poznatky

- Explorácia vs využívanie znalostí (exploration vs exploitation) v stochastických prostrediach
- Multi-armed bandit problem: stacionárny vs. nestacionárny problém
- ϵ greedy stratégia, rovnováha medzi získavaním znalosti o modeli, resp využívanie znalostí o modeli s cieľom maximalizovať očakávanú odmenu
- Action-value prístup (vypočet očakávanej hodnoty pre jednotlivé akcie a) (neskôr Q a SARSA learning)
- Gradientný prístup (výpočet preferencie výberu jednotlivých akcií) (neskôr Actor-Critic metódy)

Markovove rozhodovacie problémy

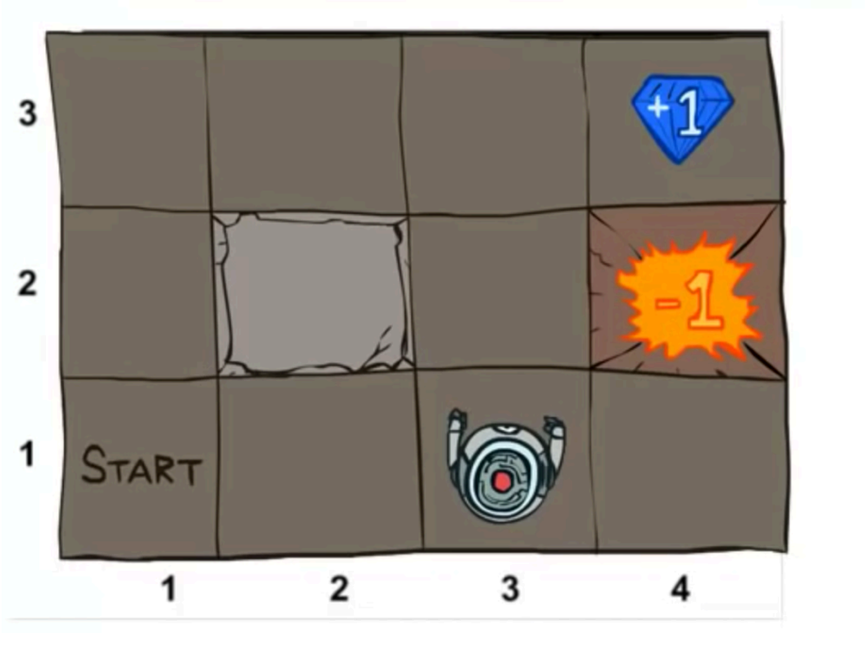
(Markov decision process, MDP)

Technical University of Košice
Department of Computers and Informatics

Grid world

Problém labyrintu:

- Agent sa pohybuje v bludisku
- Steny labyrintu blokujú pohyb agenta



Akcie agenta:

80% akcie vedie deterministicky na plánované miesto (sever)

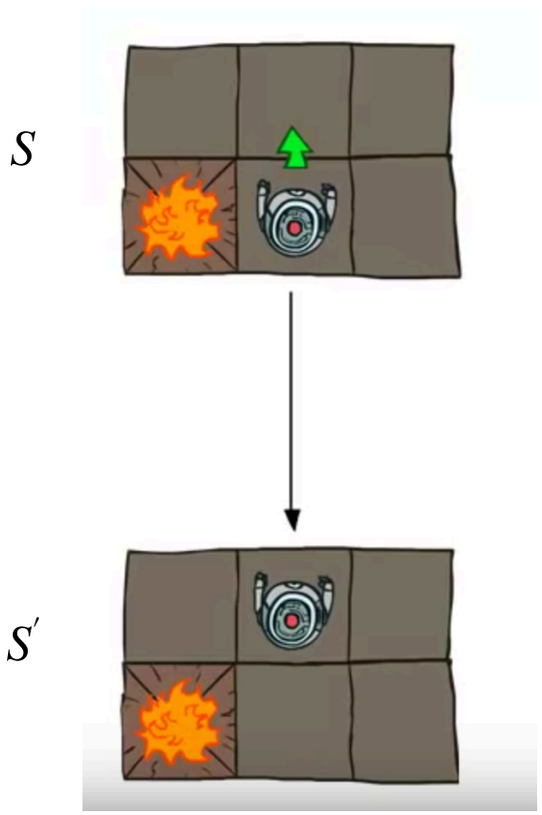
10% akcii vedie agenta na východ a západ

V prípade že agent narazí na miesto, zastane

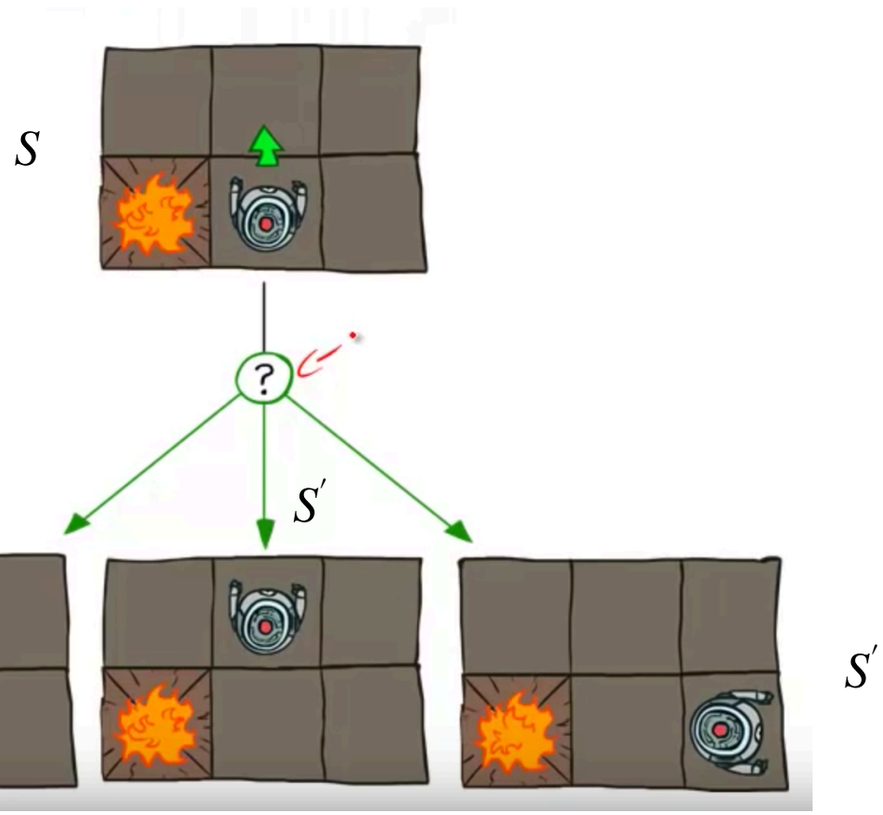
Odmena agenta vs úžitok agenta (reward vs utility (return))

Grid world

Deterministický stavový priestor

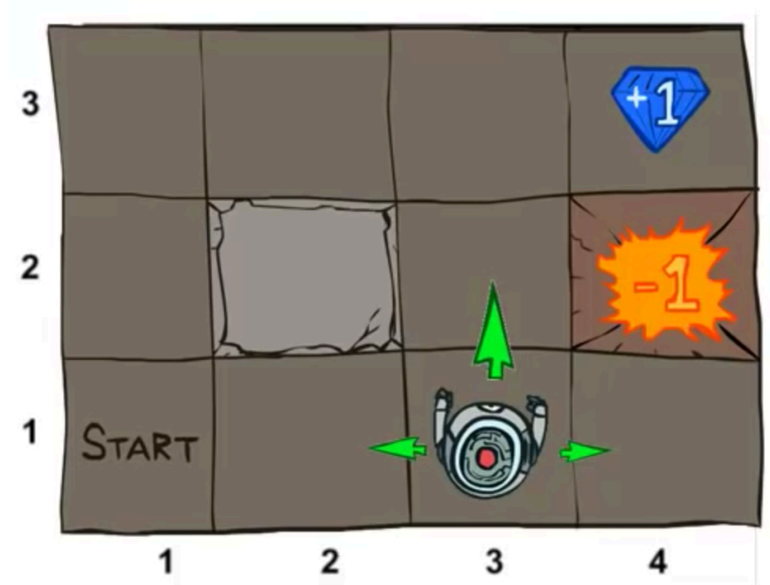


Stochastický stavový priestor

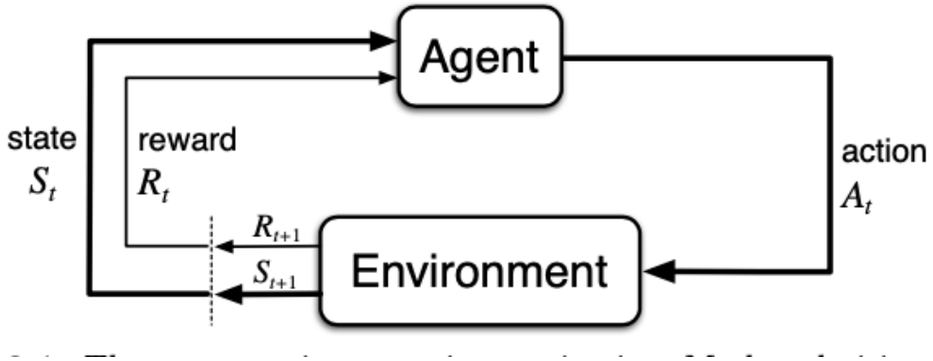


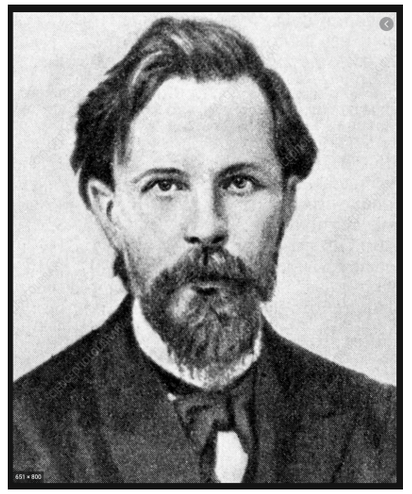
Grid world

- MDP je charakterizovaný:
  - Množinou stavov $s \in S$
  - Množinou akcií $a \in A$
  - Prechodovou pravdepodobnosťou $T(s, a, s')$ resp. podmienenou p. $P(s' \mid s, a)$
  - Odmenou $R(s, a, s')$
  - Štartovací stav
  - Konečný stav



MDP





$$S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, \ldots$$

Markovov problém (vysvetlenie): agent sa nachádza v čase *t* v stave *s*, pričom budúce stavy *s'* a navštívené stavy *s''* sú na sebe nezávislé.

$$P(S_{t+1} = s' | S_t = s_t, A_t = a_t, S_{t-1} = s_{t-1}, A_{t-1}, \ldots S_0 = s_0)$$

$$=$$

$$P(S_{t+1} = s' | S_t = s_t, A_t = a_t)$$

Analógia s plánovaním (ISI): potomok v strome je jednoznačne určený predchádzajúcim stavom a akciou.

Príklad

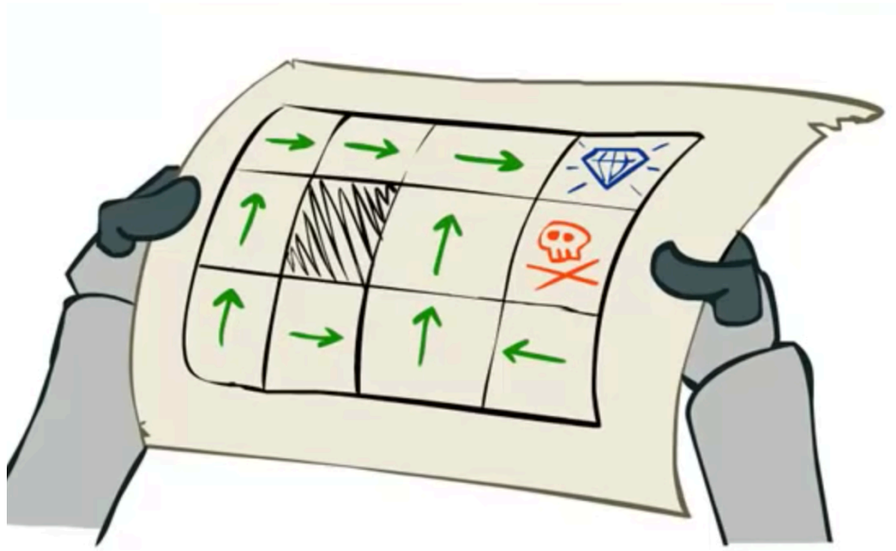
Strategický profil

- V deterministických prostrediach (cieľ):

  sekvencia akcií z štartovacieho stavu S do cieľového stavu G

- Pre MDP, je snahou nájsť optimálny str.profil $\pi^* : S \rightarrow A$:
  - Strategický profil π určuje akciu pre každý stav *s* v MDP
  - Optimálny strategický profil π^* maximalizuje očakávaný

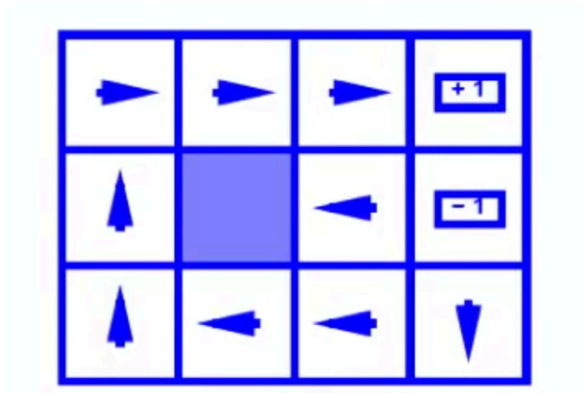
  úžitok (utility, return) agenta



Optimálny strategický profil, kde $R(s, a, s') = -0.03$,

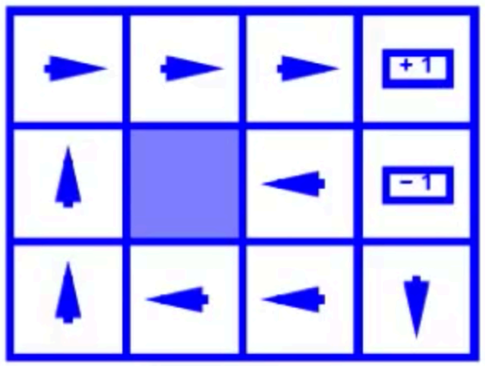
pre všetky stavy s

Optimálny strategický profil

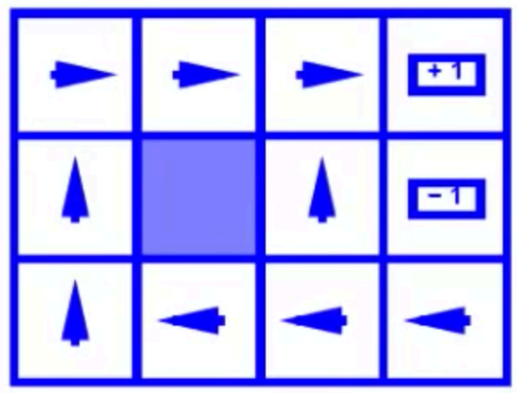


R(s) = -0.01

Optimálny strategický profil

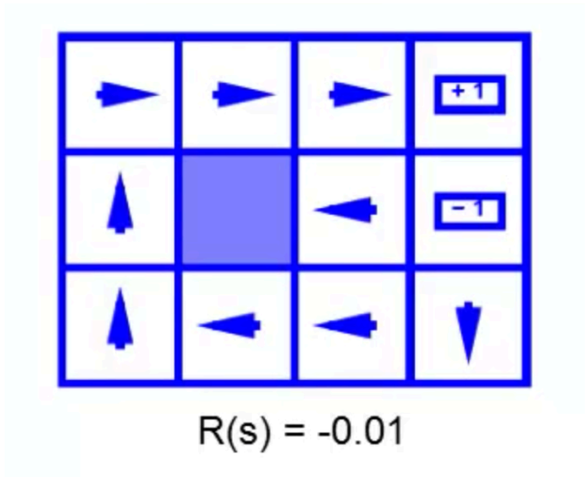


R(s) = -0.01

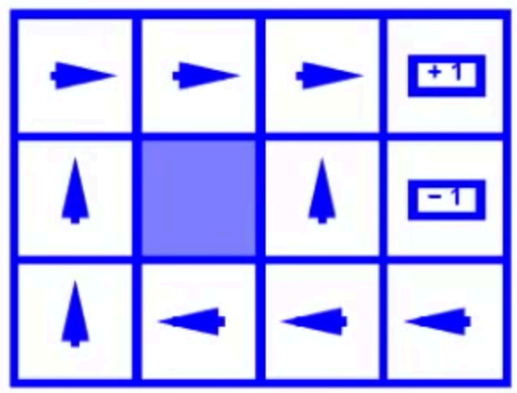


R(s) = -0.03

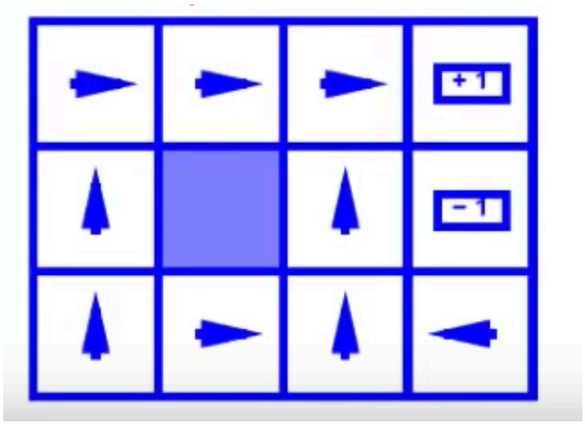
Optimálny strategický profil



R(s) = -0.01

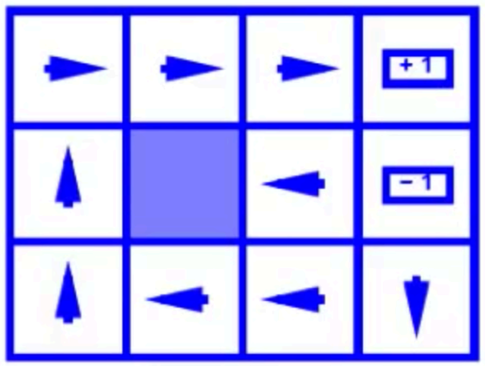


R(s) = -0.03

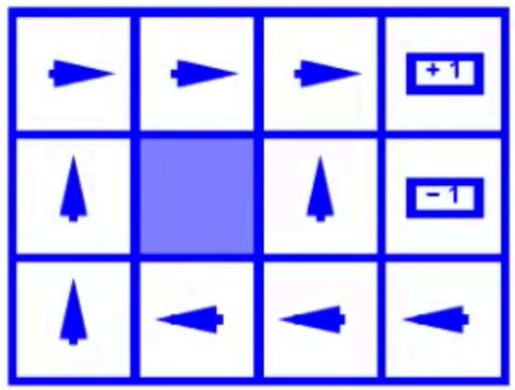


$R(s) = -0.4$

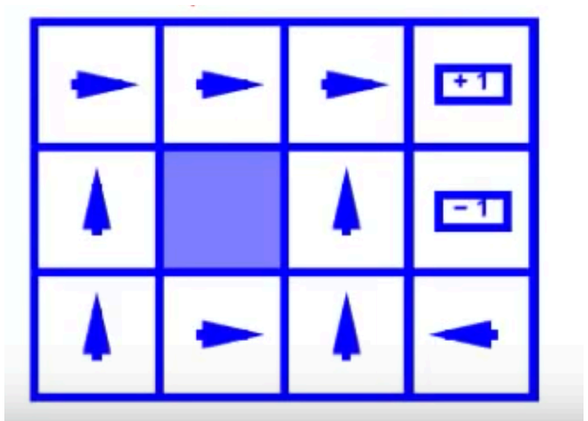
Optimálny strategický profil



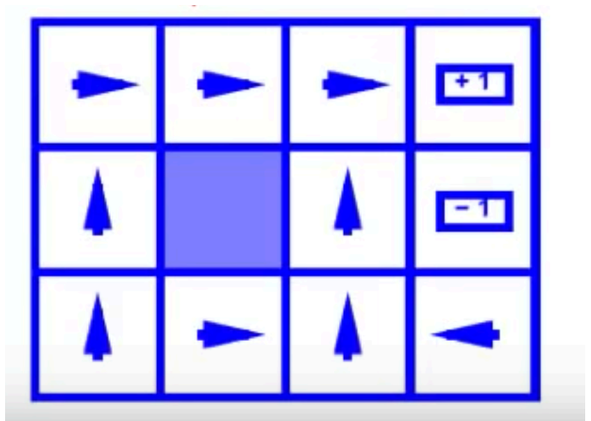
R(s) = -0.01



R(s) = -0.03

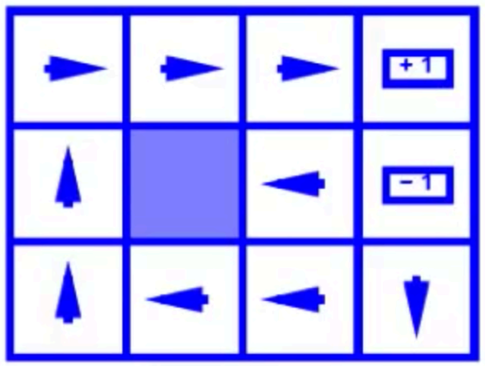


$R(s) = -0.4$

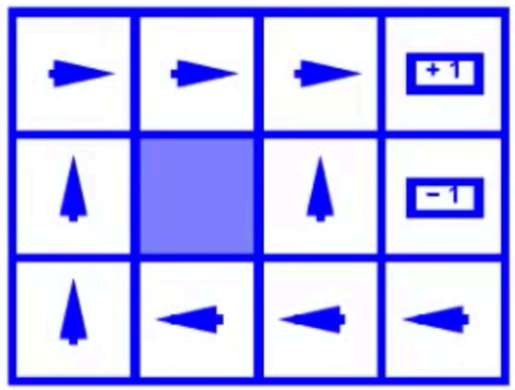


$R(s) = -2$

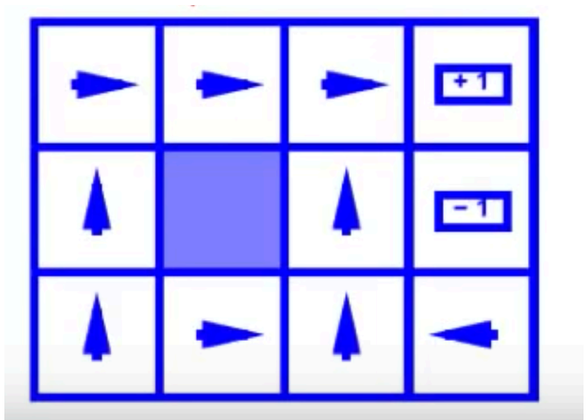
Optimálny strategický profil



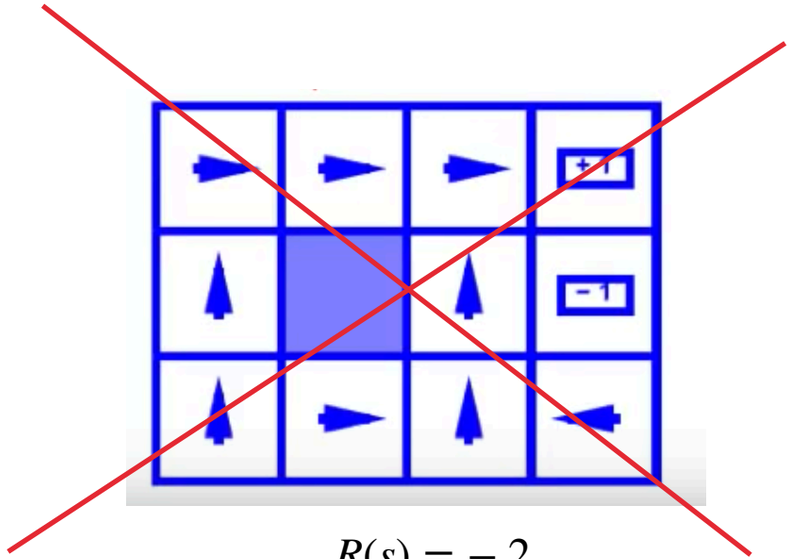
R(s) = -0.01



R(s) = -0.03

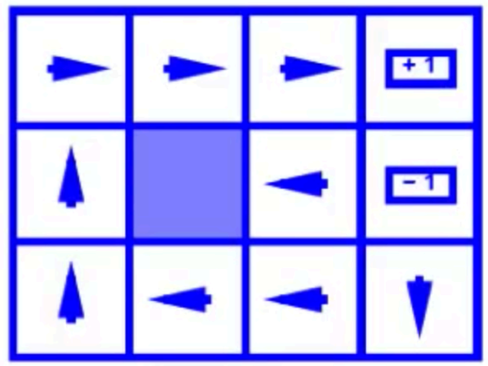


$R(s) = -0.4$

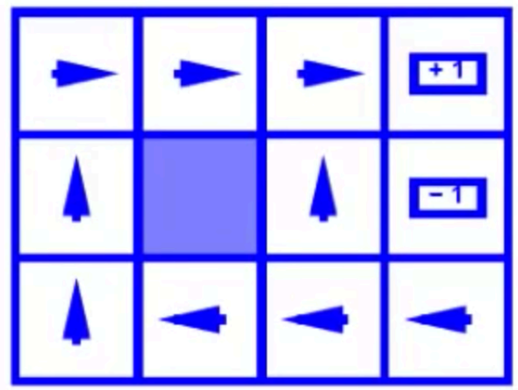


$R(s) = -2$

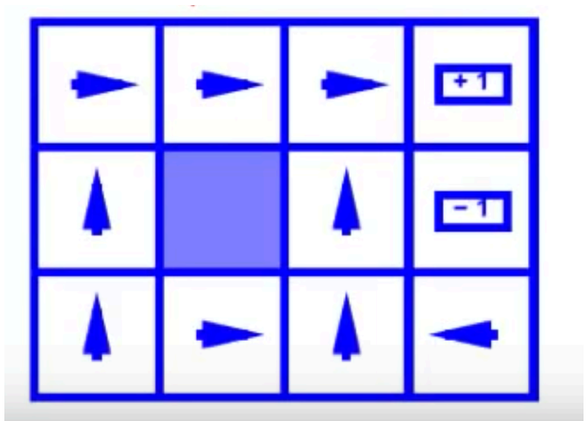
Optimálny strategický profil



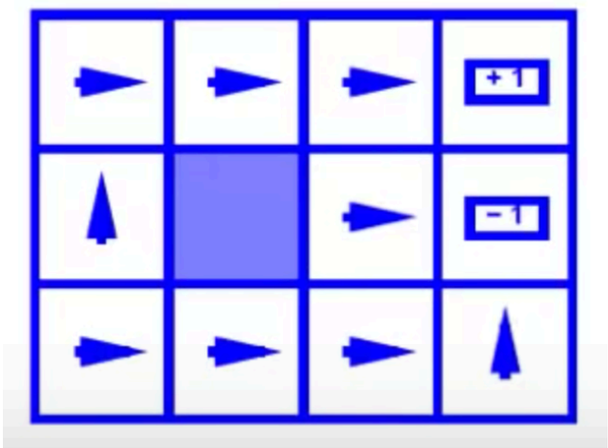
R(s) = -0.01



R(s) = -0.03



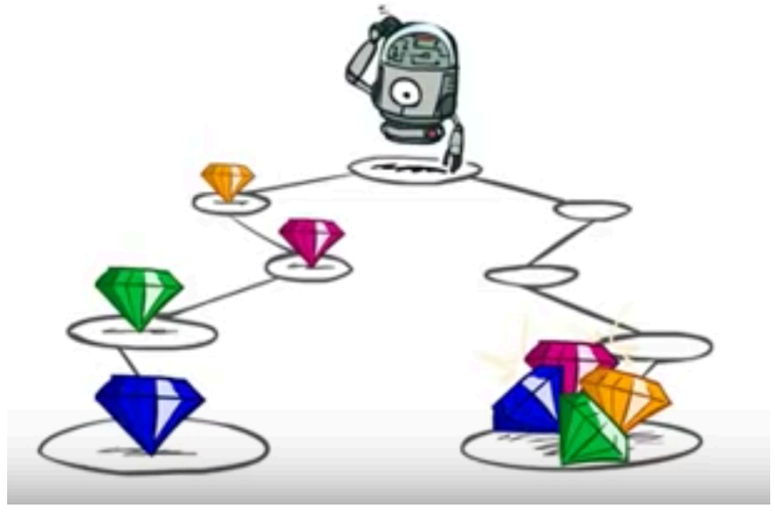
$R(s) = -0.4$



$R(s) = -2$

Postupnosť odmien v MDP

- Preferencie agenta vzhľadom na postupnosť odmien?
- Príklad: [1,2,2] vs [2,3,4]
- Príklad: [0,0,1] vs [1,0,0]



Discounting (diskontovanie)

- Snaha agenta (max úžitok): maximalizovanie hodnoty súčtu odmien získaných v MDP
- V dôsledku stochasticity MDP je vhodné preferovať odmeny získané skôr, ako neskôr
- Riešenie: discounting, hodnoty odmien klesajú exponenciálne s časom (budúcnosťou)

1

Súčasnosť

γ

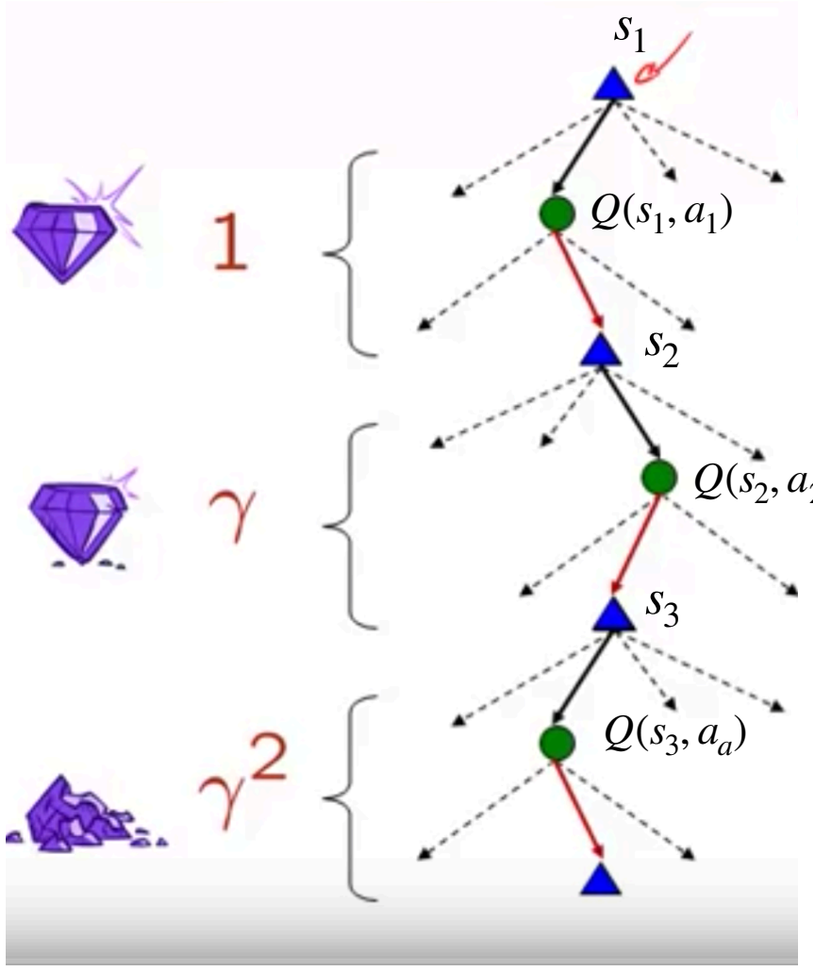
Ďalší krok

γ^2

Ďalšie 2 kroky

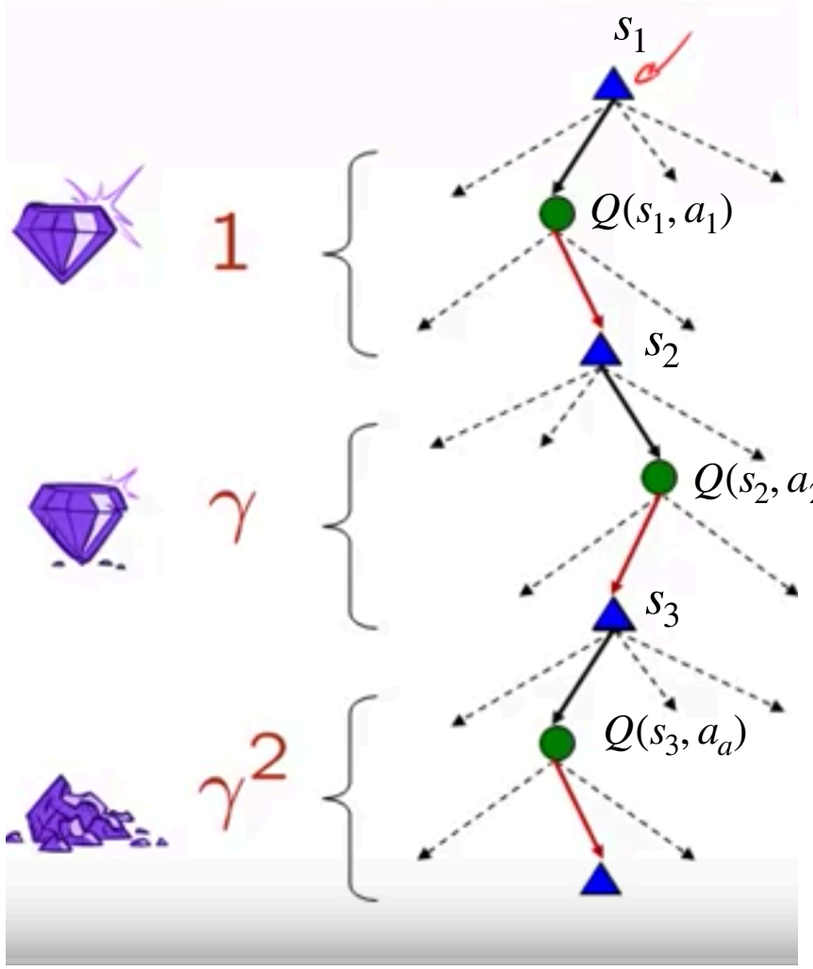
Discounting (diskontovanie)

- Ako diskontovať?
  - V každom vnorení v strome násobime γ
- Prečo diskontovať?
  - Odmeny získané skôr majú väčší úžitok ako tie neskôr
- Diskontovanie umožňuje konvergenciu algoritmov



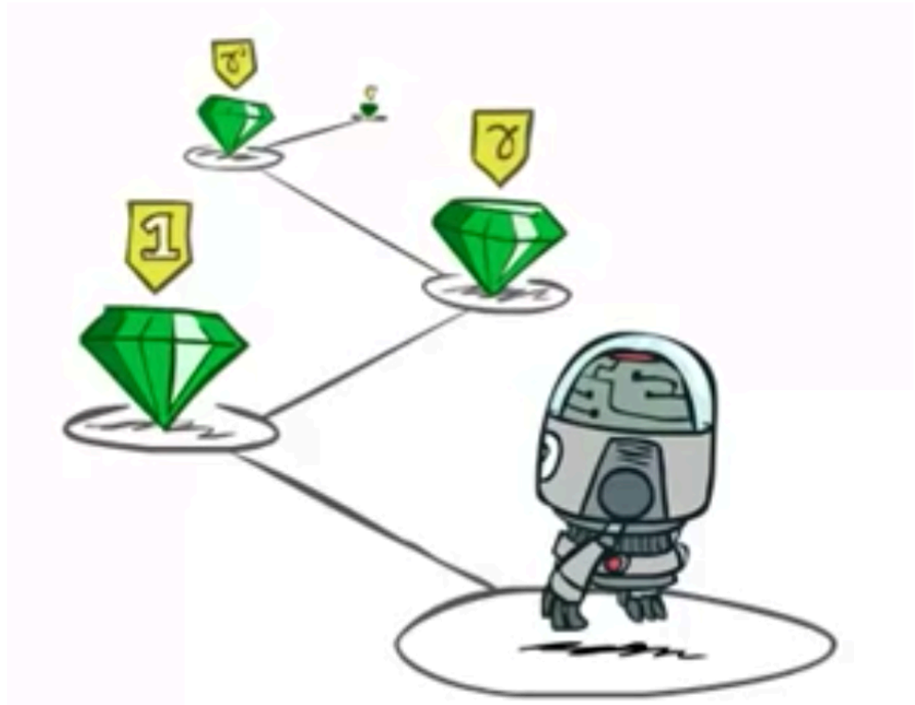
Discounting (diskontovanie)

- Ako diskontovať?
  - V každom vnorení v strome násobime γ
- Prečo diskontovať?
  - Odmeny získané skôr majú väčší úžitok ako tie neskôr
  - Diskontovanie umožňuje konvergenciu algoritmov
- U[1,2,3] vs [3,2,1], $\gamma = 0{,}5$



Sutton, R., Barto, A., Reinforcement learning: Introduction (2nd edition), MIT press, 2018, pp: 53-58

Teória preferencie (stacionarita preferencií)



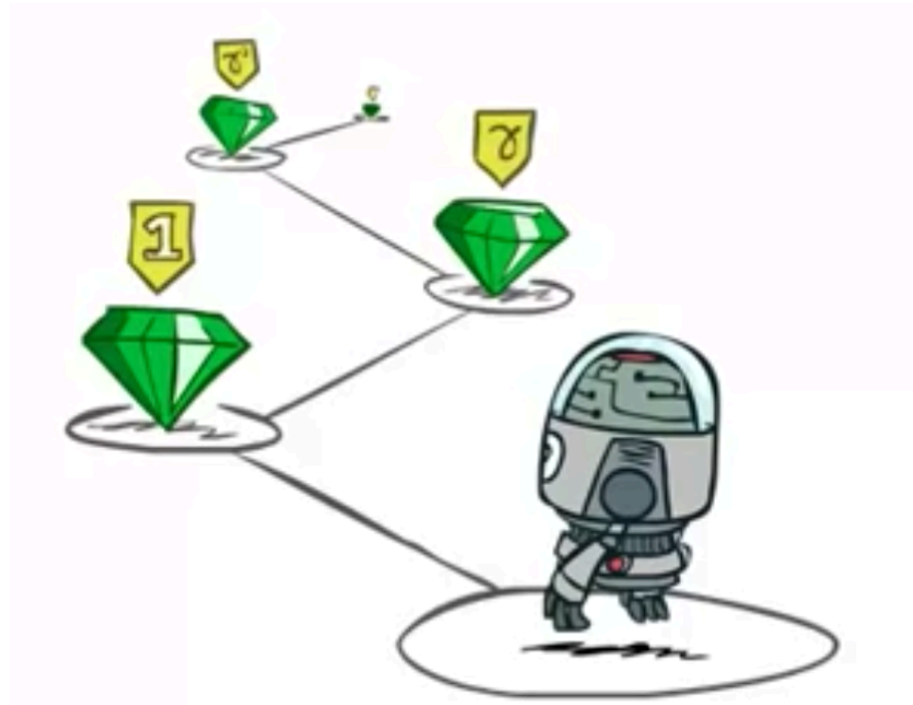
Teória preferencie (stacionarita preferencií)

- Uvažujme stacionárne preferencie agenta, t.j. platí*

$$[a_1, a_2, \ldots] \succ [b_1, b_2, \ldots]$$

$$\Downarrow$$

$$[r, a_1, a_2, \ldots] \succ [r, b_1, b_2, \ldots]$$



Za predpokladu stacionarity preferencií agenta, existujú 2 typy užitkových funkcií agenta:

- Additívny úžitok:

$$U([R_0, R_1, R_2, \ldots]) = R_0 + R_1 + R_2 + \ldots = \sum_{k=0}^{\infty} R_{t+k+1}$$

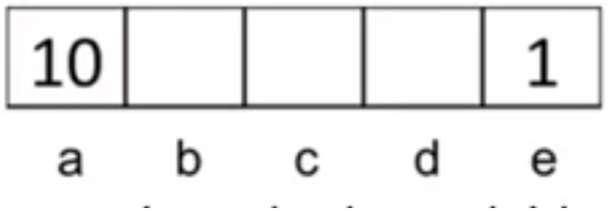
- Diskontovaný úžitok:

$$U([R_0, R_1, R_2, \ldots]) = R_0 + \gamma R_1 + \gamma^2 R_2 + \ldots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

*Drouhin, Nicolas. "Non-stationary additive utility and time consistency." Journal of Mathematical Economics 86 (2020): 1-14.

Kvíz

Uvažujme:



- Akcie: vľavo, vpravo a terminálne stavy (a, e)
- Deterministický stavový priestor

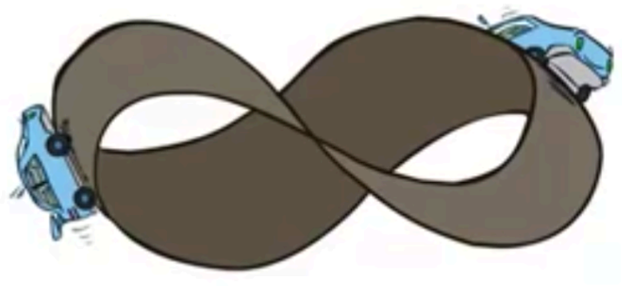
Kvíz č. 1: Pre $\gamma = 1$, načrtnite optimálny strategický profil

Kvíz č. 2: Pre $\gamma = 0.1$, načrtnite optimálny strategický profil

Kvíz č. 3: Pre aké γ, bud mať akcie vľavo a vpravo v stave d rovnaký úžitok

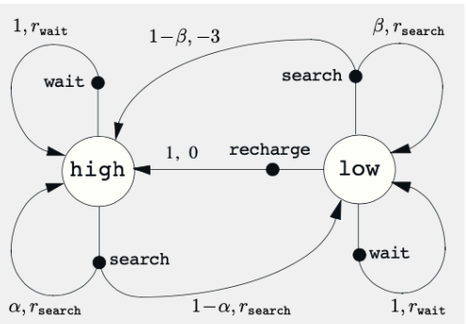
Nekonečný úžitok?

- Formulácia problému: Čo ak je náš MDP nekonečný, t.j. neexistuje terminálny stav
- Riešenie: Konečný horizont (idea depth-limited search)
  - Ukončenie epizódy po T časových jednotkách
  - Vzniká nestacionárny strategický profil



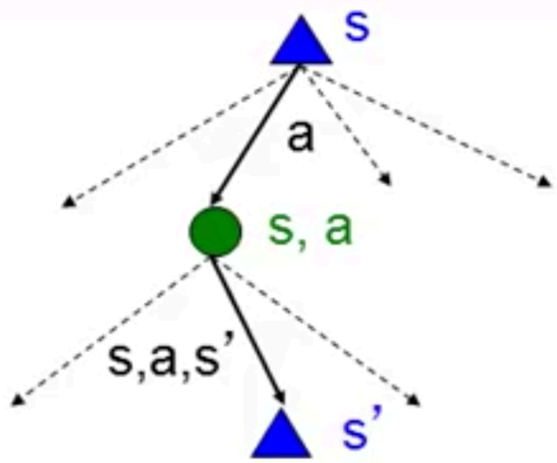
- Diskontovanie: využitie $\gamma \in (0,1)$
  - $U([R_0, R_1, R_2, \ldots) = \sum_{t=0}^{t=\infty} \gamma^t R_t \leq R_{MAX}/(1-\gamma)$

s	a	s'	$p(s' \mid s,a)$	$r(s,a,s')$
high	search	high	α	r_{search}
high	search	low	$1-\alpha$	r_{search}
low	search	high	$1-\beta$	-3
low	search	low	β	r_{search}
high	wait	high	1	r_{wait}
high	wait	low	0	-
low	wait	high	0	-
low	wait	low	1	r_{wait}
low	recharge	high	1	0
low	recharge	low	0	-



Zhrnutie

- Markovov rozhodovací problém:
  - Množina stavov S
  - Začiatočný stav s_0
  - Množina akcií A
  - Prechodové pravdepodobnosti $P(s' \mid s, a)$, resp. $T(s, a, s')$
  - Odmeny $R(s, a, s')$ a diskontovací faktor γ
  - $\pi(s)$: deterministický vs. stochastický



- Prechádzanie MDP:
  - Strategický profil: Výber akcie pre každý stav MDP
  - Úžitok: suma diskontovaných odmien

Využitie v telekomunikáciách

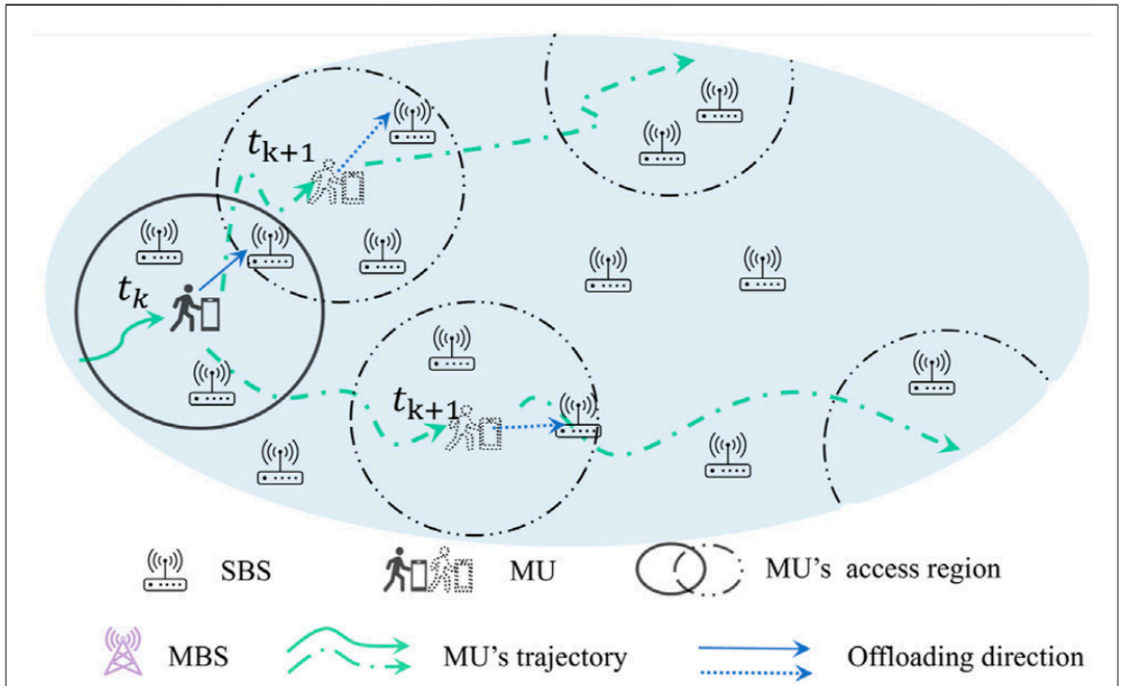
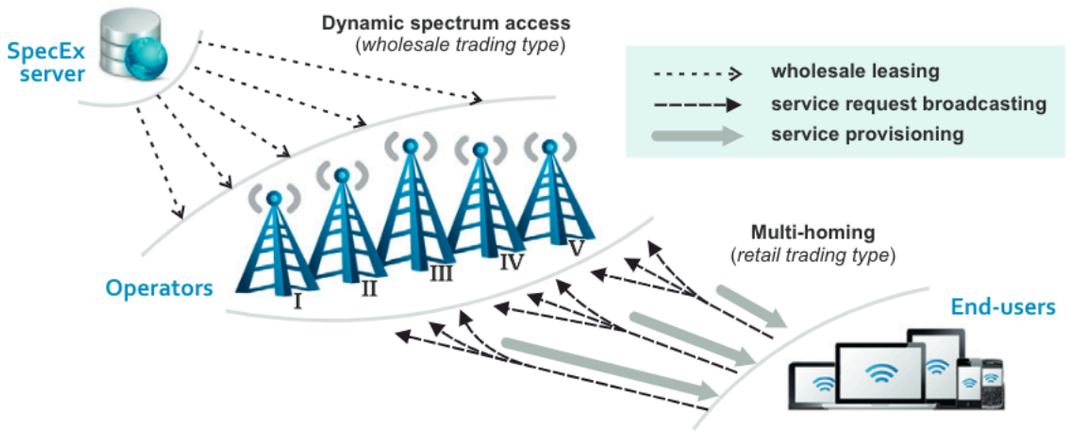


FIGURE 2 | Illustration of user-centric task offloading for MEC in UDN. The two dotted lines represent the uncertain future tracks of the MU.

LIU, S., Cheng, P., Chen, Z., Vucetic, B., & Li, Y. A Tutorial on Bandit Learning and Its Applications in 5G Mobile Edge Computing. *Frontiers in Signal Processing*, 29.



Gazda, J., Bugár, G., Vološin, M., Drotár, P., Horváth, D., & Gazda, V. (2017). Dynamic spectrum leasing and retail pricing using an experimental economy. *Computer Networks*, *121*, 173-184.