I007 Osnove umjetne inteligencije

Tema: Planiranje uz nepouzdanost

14. 4. 2021.

predavač: Darija Marković

asistent: Darija Marković

1 Planiranje uz nepouzdanost

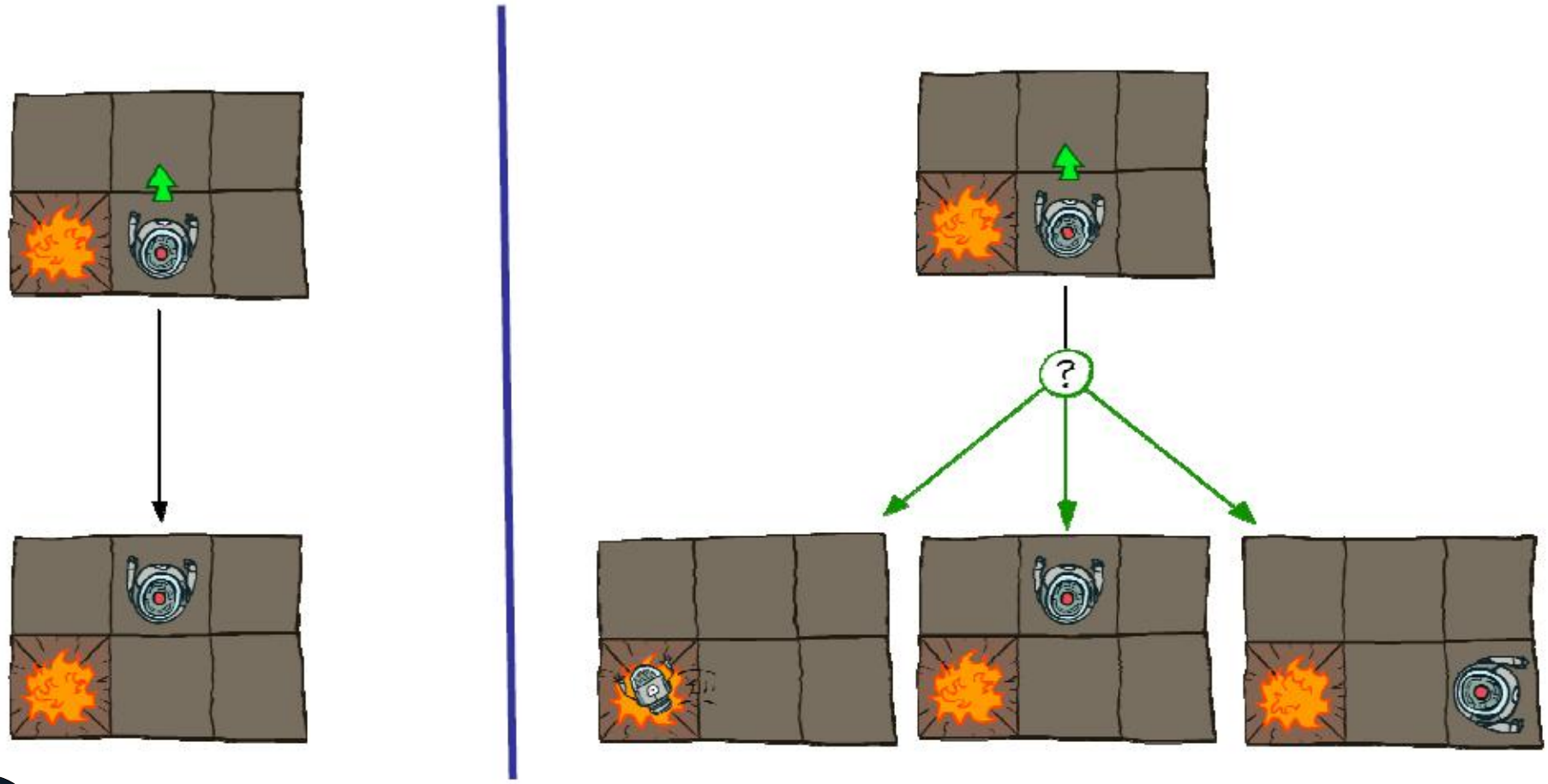
Planiranje uz nepouzdanost

- ukoliko su rezultati akcija stohastički koristit ćemo Markovljeve procese odlučivanja (MPO)

MPO su definirani s:

- skupom stanja s: S
- skupom akcija a: A
- funkcijom prijelaza $T(s, a, s')$
  - vjerojatnostima da a iz s vodi u s', tj. $P(s'|s, a)$
  - također se naziva model ili dinamika
- funkcijom nagrade $R(s, a, s')$
  - ponekad je to samo $R(s)$ ili $R(s')$
- početnim stanjem
- ponekad i završnim stanjem

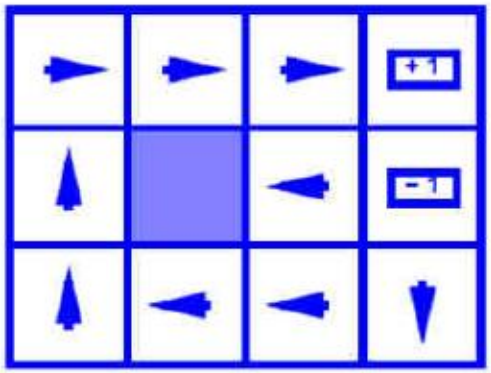
Planiranje uz nepouzdanost



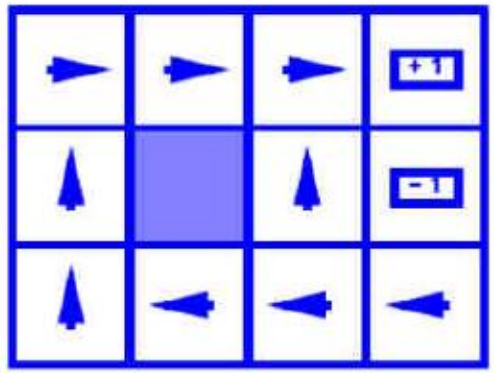
Planiranje uz nepouzdanost

- pretpostavka je da ishodi akcije ovise samo o trenutnom stanju, a ne i o prošlim stanjima
- u deterministčkim okruženjima tražili smo niz akcija iz početnog do ciljnog stanja
- kod MPO tražimo optimalnu strategiju (politiku) $\pi^* : S \to A$
  - strategija π za svakom stanju pridružuje akciju
  - optimalna strategija je ona koja maksimizira očekivanu dobit, ukoliko ju pratimo
  - eksplicitna strategija definira refleksnog agenta

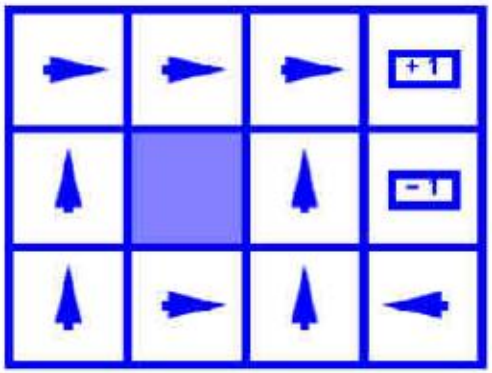
Primjer 3. Optimalna strategija



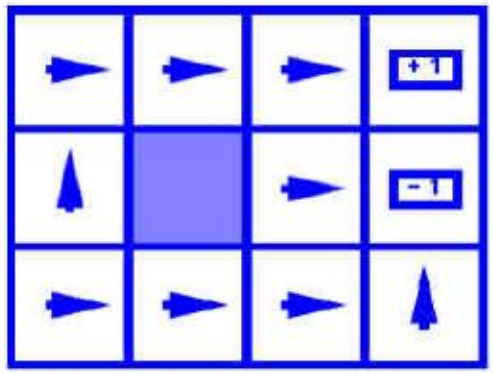
R(s) = -0.01



R(s) = -0.03



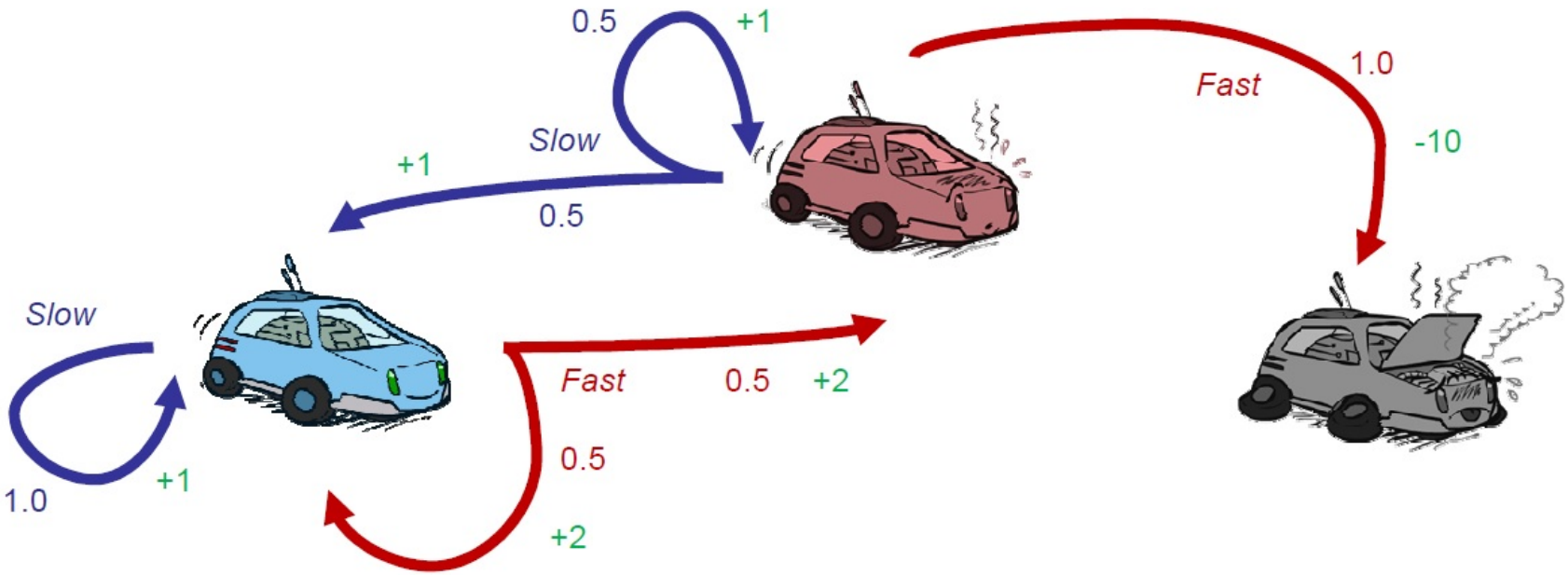
R(s) = -0.4



R(s) = -2.0

Primjer 4. Trkaći automobil

- skup stanja: { Hladan, Topao, Pregrijan }
- skup akcija: { Polako, Brzo}
- nagrada je dvostruka ukoliko se ide brzo



Planiranje uz nepouzdanost

- uobičajeno je preferirati nagrade (dobit) koje se dobiju odmah u odnosu na one koje se dobiju kasnije
- vrlo često se uzima da važnost nagrada opada eksponencijalno
- faktor umanjenja $0 < \gamma \leq 1$
- optimalna vrijednost (dobit, korisnost) stanja s: $V^*(s)$ očekivana dobit ukoliko se počinje u stanju s i djeluje optimalno
- q-vrijednost q-stanja (s, a): $Q^*(s, a)$ očekivana dobit ukoliko se u stanju s napravi akcija a i nakon toga djelujemo optimalno
- optimalna strategija: $\pi^*(s)$ optimalna akcija u stanju s

Planiranje uz nepouzdanost

$$V^*(s) = \max_a Q^*(s,a)$$

$$Q^*(s,a) = \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V^*(s')]$$

$$V^*(s) = \max_a \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V^*(s')]$$

Iteracija vrijednosti

- započinjemo s $V_0(s) = 0$, tj. pretpostavljamo da je očekivana dobit 0
- ako nam je poznat $V_k(s)$, odradimo jedan sloj expectimax

$$V_{k+1}(s) \leftarrow \max_a \sum_{s'} T(s, a, s')[R(s, a, s') + \gamma V_k(s')]$$

Primjer 5. Trkaći automobil 2

Odredite vrijednost $V_2(s)$ za MPO iz Primjera 4.

Iteracija vrijednosti

konvergencija?

- ako je stablo maksimalne dubini M, tada je V_M točna vrijednost optimalne dobiti
- u slučaju ako je $\gamma < 1$: u $k-$tom koraku V_k i V_{k+1} se razlikuju za maksimalno $\gamma^k \max |R|$ pa s povećanjem k vrijednosti konvergiraju

Ocjena (procjena) strategije

- za odabranu strategiju π trebamo odrediti $V^{\pi}(s)$

$$V^{\pi}(s) = \sum_{s'} T(s, \pi(s), s')[R(s, \pi(s), s') + \gamma V^{\pi}(s')]$$

- određivanje vrijednosti $V^{\pi}(s)$ radimo na sljedeći način

$$V_0^{\pi}(s) = 0$$
$$V_{k+1}^{\pi}(s) \leftarrow \sum_{s'} T(s, \pi(s), s')[R(s, \pi(s), s') + \gamma V_k^{\pi}(s')]$$

Izvod strategije

- strategiju vidimo iz q-vrijednosti

$$\pi^*(s) = \arg\max_a Q^*(s,a)$$

- iteracija strategija: za odabranu strategiju π_i odredimo vrijednosti uz pomoć ocjene strategije

$$V_{k+1}^{\pi_i}(s) \leftarrow \sum_{s'} T(s,\pi_i(s),s')[R(s,\pi_i(s),s') + \gamma V_k^{\pi_i}(s')]$$

- nakon toga radimo poboljšanje kako bi dobili bolju strategiju uz pomoć izvoda strategija

$$\pi_{i+1}^*(s) = \arg\max_a \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V^{\pi_i}(s')]$$