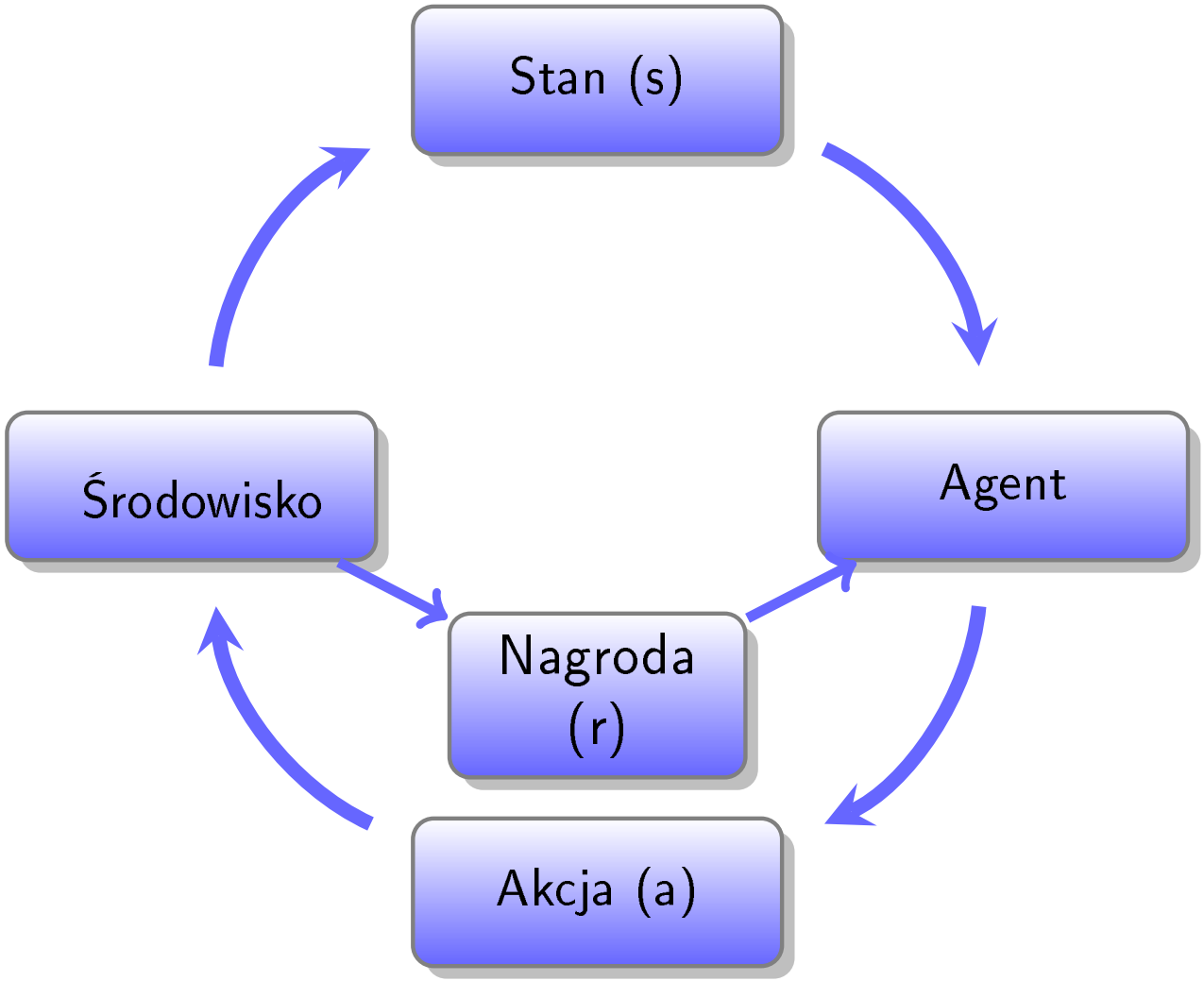
Wskazówki do ćwiczenia 6.

WSI - ćwiczenia 2022Z

Stanisław Pawlak & Jakub Łyskawa

December 22, 2022

Proces decyzyjny Markowa



Uczenie ze wzmocnieniem

Polityka (π) - rozkład prawdopodobieństw akcji w każdym stanie.

Uczenie ze wzmocnieniem - proces szukania polityki maksymalizującej sumę nagród.

Q-Learning

Q-Learning

Uczenie oczekiwanych nagród dla poszczególnych akcji i stanów.
Polityka - w każdym stanie wybór najlepszej akcji.

Q-Learning

Funkcja Q:

$$Q^{\pi}(s_t, a_t) = \mathbb{E}\left(r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + ... | s_t, a_t, \pi\right)$$

$\gamma \in (0, 1)$ - dyskonto - im bardziej odległy krok, tym mniej wpływa

Q-Learning

Funkcja Q:

$$Q^{\pi}(s_t, a_t) = \mathbb{E}\left(r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + ... | s_t, a_t, \pi\right)$$

$\gamma \in (0, 1)$ - dyskonto - im bardziej odległy krok, tym mniej wpływa

Dla polityki w Q-learningu:

$$Q^{\pi}(s_t, a_t) = \mathbb{E}\left(r_t + \gamma \max_a Q^{\pi}(s_{t+1}, a) | s_t, a_t, \pi\right)$$

Uczenie funkcji Q

Funkcja Q jako tablica $|S| \times |A|$

W każdej iteracji minimalizacja $MSE(Q(s_t, a_t), r_t + \gamma \max_a Q(s_{t+1}, a))$

Czyli

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \beta \left(r_t + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t) \right)$$

Uczenie funkcji Q

Funkcja Q jako tablica $|S| \times |A|$

W każdej iteracji minimalizacja $MSE(Q(s_t, a_t), r_t + \gamma \max_a Q(s_{t+1}, a))$

Czyli

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \beta \left(r_t + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t) \right)$$

Jeżeli po akcji był koniec, to oczekiwane przyszłe nagrody =0,
więc 0 zamiast $Q(s_{t+1}, a)$

Eksploracja

Nie można ograniczyć się do decyzji określonych przez politykę - trzeba sprawdzać wyniki innych akcji

Eksploracja

Nie można ograniczyć się do decyzji określonych przez politykę - trzeba sprawdzać wyniki innych akcji

Strategia ϵ-zachłanna

- Z prawdopodobieństwem $1 - \epsilon$ akcja wylosowana spośród najlepszych
- W przeciwnym przypadku, akcja wylosowana spośród wszystkich

Eksploracja

Nie można ograniczyć się do decyzji określonych przez politykę - trzeba sprawdzać wyniki innych akcji

Strategia ϵ-zachłanna

- Z prawdopodobieństwem $1 - \epsilon$ akcja wylosowana spośród najlepszych
- W przeciwnym przypadku, akcja wylosowana spośród wszystkich

Strategia Boltzmannowska

$$P(a|s) = \frac{exp(Q(s,a)/T)}{\sum_{a'} exp(Q(s,a')/T}$$

Q-Learning

$Q \leftarrow [\alpha]_{|S| \times |A|}$
$t \leftarrow 0$
$s_0 \leftarrow$ initial state

while(!end)
$a_t \leftarrow \text{exploration}(Q, s_t)$
$s_{t+1}, r_t \leftarrow \text{env\_step}(a_t)$
$Q(s_t, a_t) \leftarrow Q(s_t, a_t)$
$+\beta \left(r_t + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)\right)$
$t \leftarrow t + 1$

Ewaluacja

W celu ewaluacji należy periodycznie zatrzymać uczenie,
wykonać kilka epizodów bez eksploracji
i zmierzyć otrzymane sumy nagród