

Q-learning

Implementacja od podstaw



Waldemar Kołodziejczyk

Zawodowo:

Machine Learning Engineer ( , )

Background akademicki:

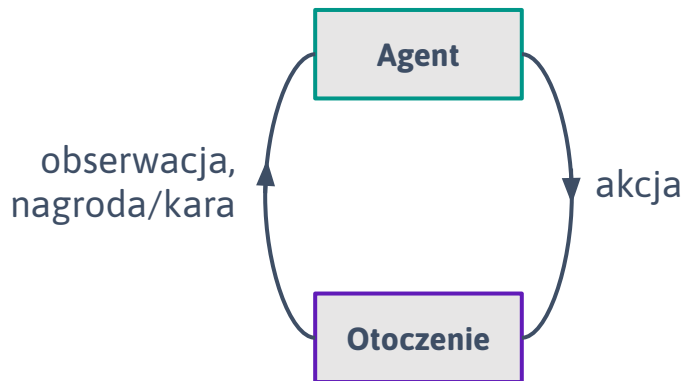
Informatyka, Energetyka (Politechnika Warszawska)

Obszar zainteresowań:

Machine Learning, Reinforcement Learning, algorytmika

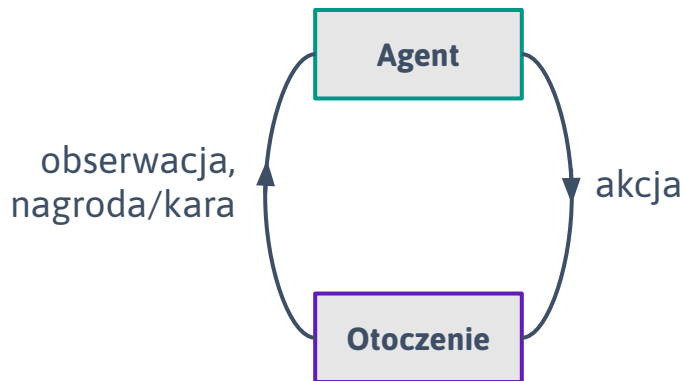


Koncepcja uczenia ze wzmocnieniem



Uczenie ze wzmocnieniem (Reinforcement Learning) - gałąź sztucznej inteligencji, w której **Agent** współoddziałuje z **Otoczeniem** (rzeczywistym lub symulacyjnym) i **na podstawie wielu prób (akcji) odnajduje wzorzec optymalnego zachowania**.

Koncepcja uczenia ze wzmocnieniem



```
import gym

env = gym.make("Taxi-v3")
agent = RLAgent(env)

observation = env.reset()
for _ in range(1000):
    action = agent.act(observation)
    observation, reward, done, info = env.step(action)
    agent.update(action, reward)

    if done:
        observation = env.reset()
env.close()
```

Proces Decyzyjny Markova (MDP)

przestrzeń **akcji** (ang. action space),
ewentualnie \mathbf{A}_s jako zbiór akcji zależny
od obecnego stanu

natychmiastowa **nagroda** lub
oczekiwana natychmiastowa
nagroda po przejściu ze stanu
 \mathbf{s} do \mathbf{s}' po podjęciu akcji \mathbf{a}

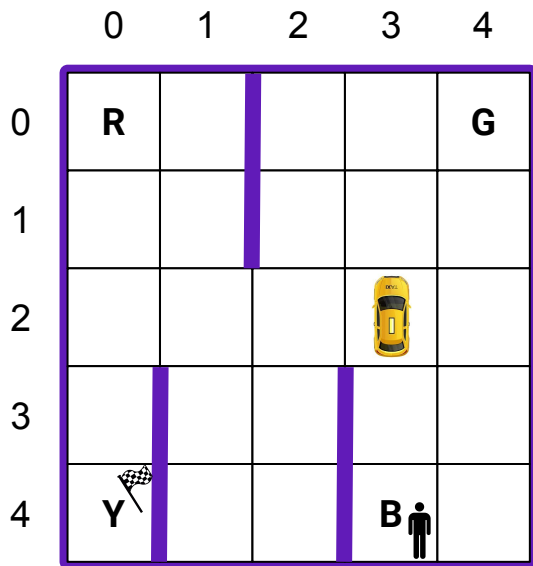
$\mathbf{S}, \mathbf{A}, P_a, R_a$

przestrzeń **stanów**
(ang. state space)

prawdopodobieństwo **przejsć**
pomiędzy stanami \mathbf{s} i \mathbf{s}' po
podjęciu akcji \mathbf{a} (ang. transition
probability)

$$P_a(s, s') = Pr(s_{t+1} = s' | s_t = s, a_t = a)$$

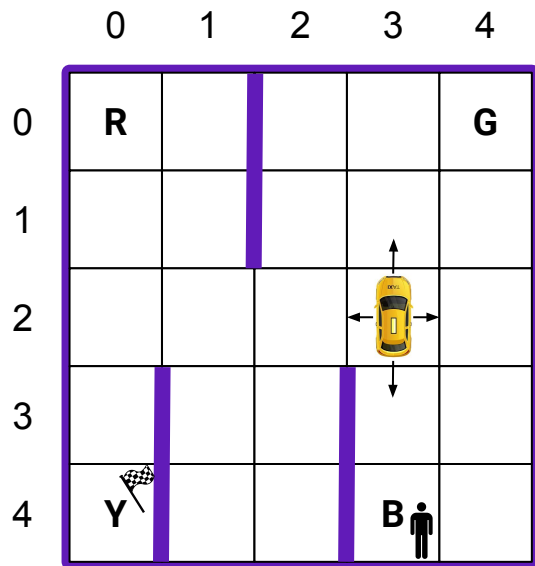
Środowisko symulacyjne



Zasady rozgrywki:

- Otrzymujesz 20 punktów za wysadzenie pasażera w odpowiednim miejscu,
- Tracisz 1 punkt za każdym ruchem, który wykonujesz,
- Tracisz 10 punktów za każde odebranie lub wysadzenie pasażera w nieodpowiednim miejscu.

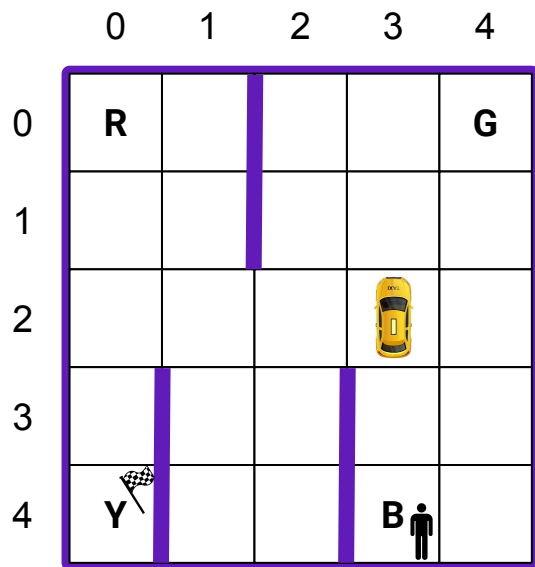
Przestrzeń akcji



Możliwe akcje taksówkarza:

- 0 - ruch w dół
- 1 - ruch w górę
- 2 - ruch w prawo
- 3 - ruch w lewo
- 4 - odebranie pasażera
- 5 - wysadzenie pasażera

Przestrzeń stanów



$$S = 5 \times 5 \times 5 \times 4 = 500$$

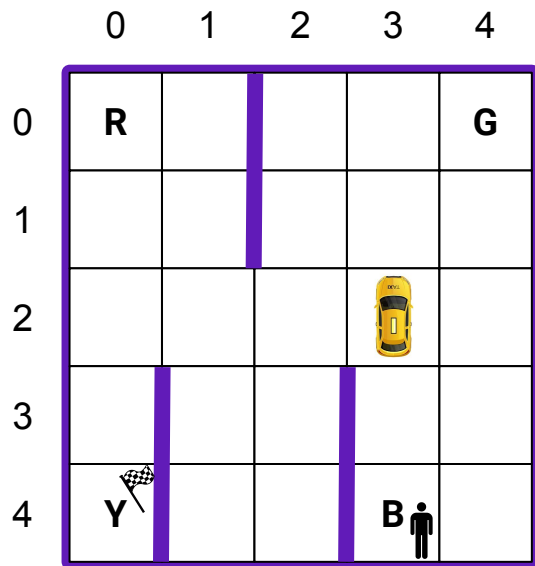
Możliwe lokalizacje pasażera:

0 - R
1 - G
2 - Y
3 - B
4 - taksówka

Możliwe destynacje:

0 - R
1 - G
2 - Y
3 - B

Przestrzeń stanów

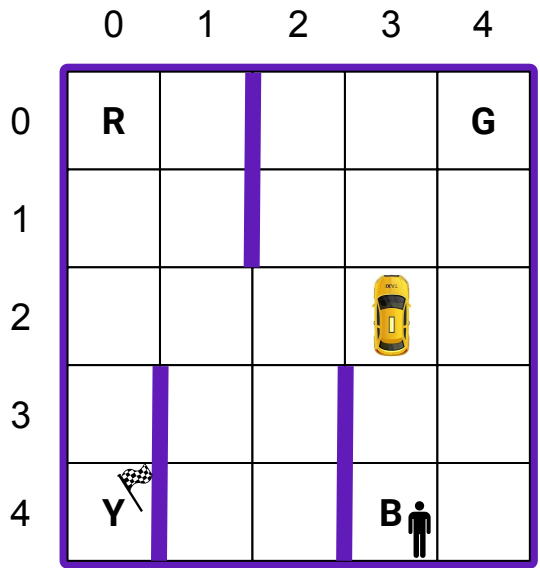


$$S = 5 \times 5 \times 5 \times 4 = 500$$

W rzeczywistości:

$$S = 5 \times 5 \times 5 \times (4 - 1) = 400$$

Przestrzeń stanów



	X	Y	P	D
0	0	0	0	1
1	0	0	0	2
2	0	0	0	3
3	0	0	1	0
...
125	2	3	3	2
...
500	4	4	4	3

Funkcja Q

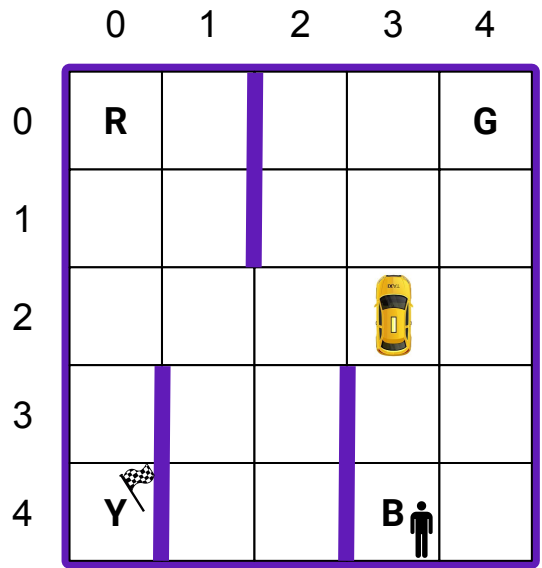
$$Q^{\pi}(s_t, a_t) = E_{\pi} \left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t, a_t \right]$$







Funkcja **Q** zależna od
obecnego stanu \mathbf{s}_t i
podjętej akcji \mathbf{a}_t zgodnie
z przyjętą strategią $\boldsymbol{\pi}$

Oczekiwany zwrot
zdyskontowany względem
przyszłych nagród $\mathbf{r}_{t+...}$ zakładając
przyjętą strategię $\boldsymbol{\pi}$...

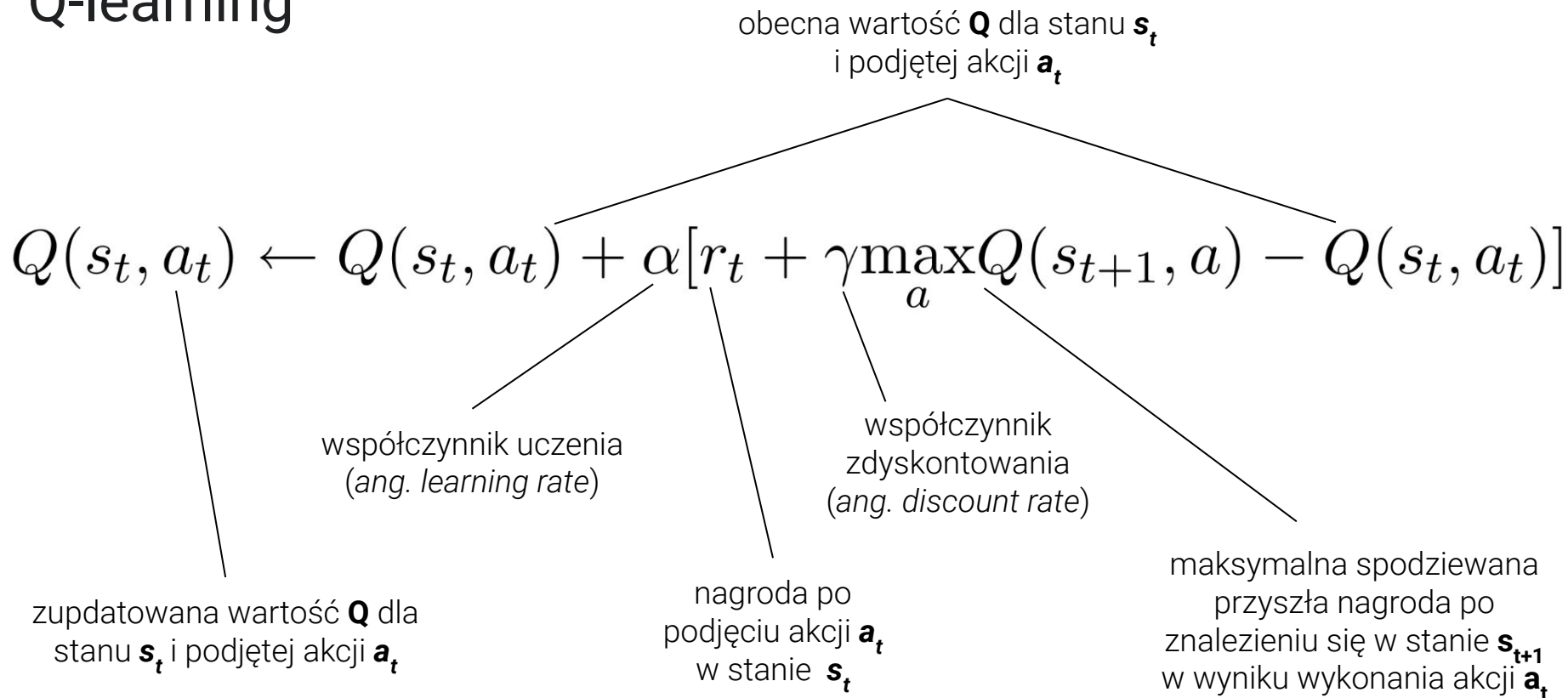
...pod warunkiem
obecnego stanu \mathbf{s}_t
i podjętej akcji \mathbf{a}_t

Tablica wartości Q



						
	A_0	A_1	A_2	A_3	A_4	A_5
S_0	-4.8	-1.1	-0.2	0.5	-8.1	-9.4
...
S_{125}	4.3	0.1	2.1	-0.4	-9.7	-9.3
...
S_{500}	2.6	2.2	-0.1	2.8	-9.2	-7.9

Q-learning



Q-learning

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_t + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)]$$

$$Q(s_t, a_t) \leftarrow (1 - \alpha)Q(s_t, a_t) + \alpha[r_t + \gamma \max_a Q(s_{t+1}, a)]$$

Q-learning - warunki zbieżności gwarantowanej

1. Współczynnik uczenia musi dążyć do zera, ale nie za szybko. Formalnie, **suma ciągu kolejnych wartości współczynnika musi być rozbieżna, a suma ich kwadratów musi być zbieżna do skończonej wartości**. Przykładowym ciągiem spełniającym te własności jest $1/(i+1)$ dla $i \geq 0$.
2. Każda para stan-akcja musi być wypróbowana nieskończoną liczbę razy. Formalnie, należy zagwarantować, że **każda akcja ma niezerowe prawdopodobieństwo wybrania dla dowolnego stanu**. W praktyce, odpowiednia strategia epsilon-greedy spełnia ten warunek.

[Źródło: Q-learning Christopher J. C. H. Watkins & Peter Dayan](#)

Materialy i źródła

1. *"Reinforcement Learning: An Introduction"*, R. S. Sutton, A. G. Barto - [sprawdź](#)
2. *"RL Course by David Silver"*, YouTube, [sprawdź](#)
3. *"Reinforcement Learning Specialization"*, Coursera, [sprawdź](#)
4. *"Practical Reinforcement Learning"*, Coursera, [sprawdź](#)