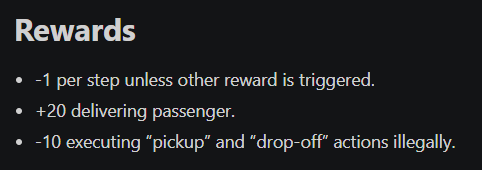
Algorytm Q-Learn

Mateusz Matukiewicz 325195

System Nagradzania modelu (Z dokumentacji środowiska **gym-taxi**):



Wzór do aktualizacji Q-tabeli (**Wikipedia**):

A close-up of a white background

Description automatically generated

**Learning Rate - 0.1**

A graph of a number

Description automatically generated

Jak można zauważyć wartość nagrody dla metody eksploracji epsilon-greedy oscyluje w okolicy 1.5, co oznacza, że algorytm poprawnie nauczył się optymalnych strategii w środowisku i utrzymuje stały poziom nagród na przestrzeni epizodów. Natomiast dla strategii Boltzmann wartość nagrody oscylowała w okolicach -85 co wskazuje na to, że algorytm często otrzymywał dużo kar i ciężko mu było nauczyć się odpowiednich strategii.

A graph with red lines

Description automatically generated

Tutaj możemy zauważyć, że algorytm częściej eksplorował przestrzeń przez co widzimy strome wzrosty i spadki wartości nagrody, ponieważ algorytm częściej decydował się sprawdzać nowe niesprawdzone pola i czasem model dobrze na tym wychodził, a czasem wręcz przeciwnie.

**Learning rate – 0.5 oraz 0.9**

**A graph with numbers and lines

Description automatically generated**

**A graph with red lines and numbers

Description automatically generated**

Dla powyższych współczynników uczenia obie strategie poradziły sobie praktycznie identycznie jak przy współczynniku uczenia 0.1. Z tym, że przy wyższych współczynnikach uczenia osiągnęliśmy wyższe średnie wartości nagrody. Dla strategii Epsilon-greedy wartości oscylujące w okolicach od 2 do 3 a dla strategii Boltzmanna od -45 do -42.