Dokumentacja – laboratorium nr 6
Wprowadzenie do Sztucznej Inteligencji

Dominika Wyszyńska 318409

17 stycznia 2024

1 Wstęp

Celem zadania było zaimplementowanie algorytmu Q-learning. Następnie przeprowadzenie eksperymentów, badając wpływ różnych wartości hiperparametrów oraz poznanych strategii eksploracji na działanie algorytmu. Do testów wykorzystano środowisko `Taxi` dostępne w bibliotece Gym. Wybrane strategie:

- Strategia ϵ-zachłanna, z prawdopodobieństwem ϵ wybieramy akcję losową; z prawdopodobieństwem 1-ϵ akcję zachłanną (jeśli jest ich wiele, to losowo wybieramy jedną)
- Strategia oparta na rozkładzie Boltzmanna:

$$\pi(x,a) = \frac{e^{\frac{Q(x,a)}{\tau}}}{\sum_b e^{\frac{Q(x,b)}{\tau}}}$$

  gdzie:

  - $\pi(x,a)$ to prawdopodobieństwo wyboru akcji a w stanie x,
  - $Q(x,a)$ to wartość Q dla pary stanu x i akcji a,
  - τ (temperatura) to parametr kontrolujący "eksploracyjność"strategii Boltzmanna.

2 Opis zaplanowanych eksperymentów

Eksperymenty zostały zaprojektowane w następujący sposób:

a) Badanie wpływu wartości hiperparametru *learning_rate*

b) Badanie wpływu wartości hiperparametru *discount_rate*

c) Badanie wpływu wartości hiperparametru *exploration*

d) Badanie wpływu strategii eksploracji na działanie algorytmu:

- Strategia ϵ-zachłanna
- Strategia oparta na rozkładzie Boltzmanna

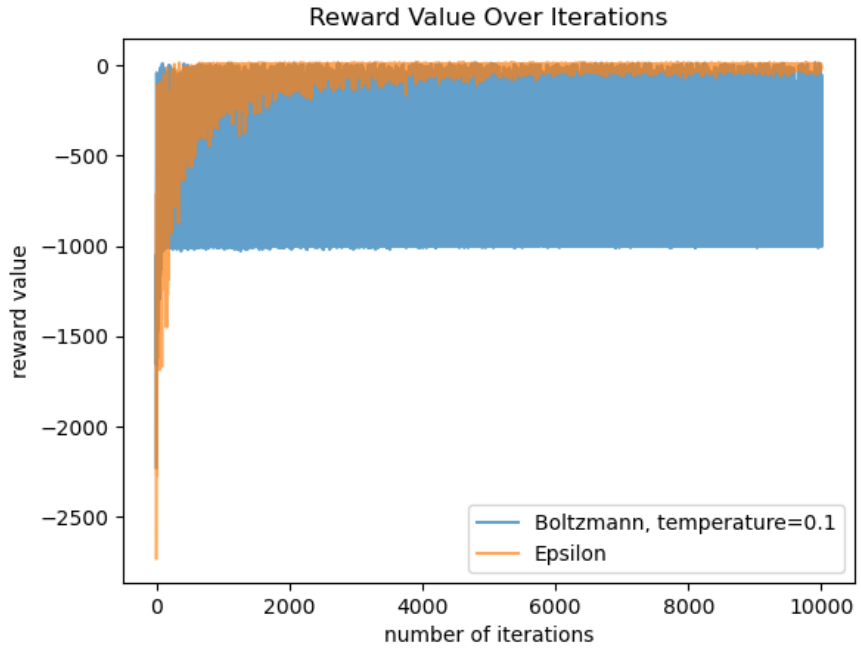
3 Wyniki i opis eksperymentów

Podczas eksperymentów ze zmianą parametrów prezentowane są wyniki dla obu strategii, umożliwiając porównanie skuteczności między nimi w zależności od różnych ustawień.

Wszystkie testy zostały przeprowadzone przy stałych parametrach *interval* równym 1000 i *episodes_number* równym 10000. Parametr *interval* odnosi się do co ile epizodów zbierane są dane i generowane wykresy, natomiast *episodes_number* to łączna liczba epizodów, które zostały wykonane w trakcie testów. Ustawienie tych parametrów pozwala na monitorowanie i analizę wyników algorytmu Q-learning na przestrzeni 10000 epizodów, co umożliwia uzyskanie reprezentatywnego obrazu skuteczności algorytmu w danym środowisku. Stałe wartości tych parametrów ułatwiają porównanie wyników między różnymi strategiami eksploracji oraz hiperparametrami, tworząc jednolite warunki testowe dla wszystkich przypadków.

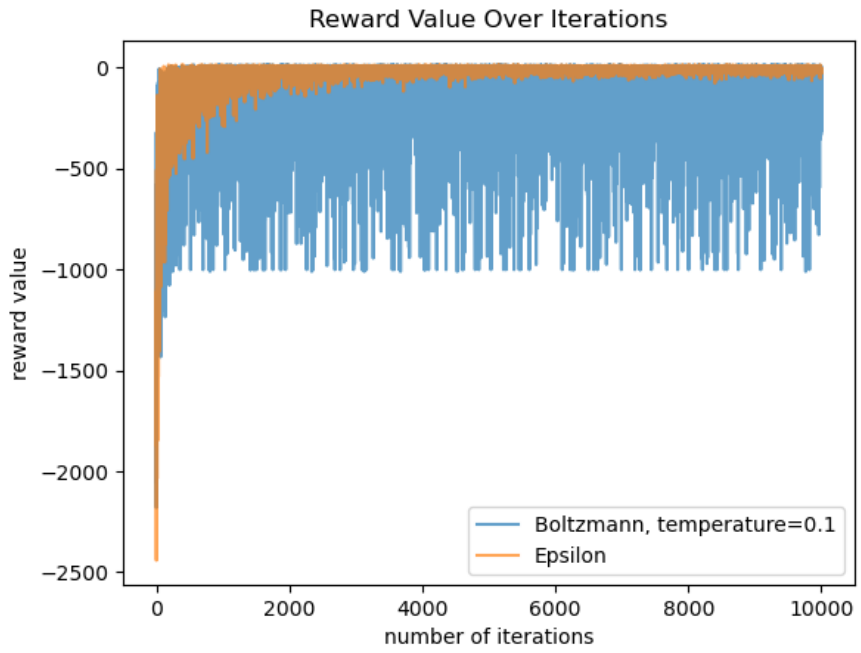
Eksperyment z `discount_rate`

W algorytmie Q-learning, wartość *discount_rate* (często oznaczane jako γ) przyjmuje wartości z zakresu od 0 do 1. Ten parametr wpływa na sposób, w jaki agent ocenia przyszłe nagrody.



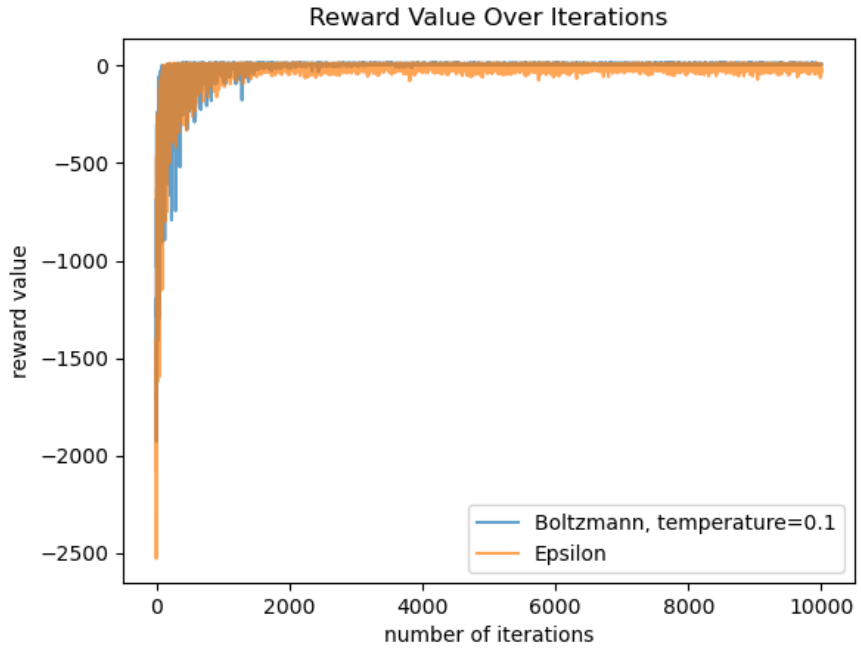
Rysunek 1: Wykres zależności wartości nagrody od iteracji przy zmianie parametru *discount_rate* na 0.2, *learning_rate*=0.1, *exploration*=0.2

Najlepsza wartość nagrody dla strategii epsilon=-6.21 oraz Boltzmanna=-436.71.



Rysunek 2: Wykres zależności wartości nagrody od iteracji przy zmianie parametru *discount_rate* na 0.5, *learning_rate*=0.1, *exploration*=0.2

Najlepsza wartość nagrody dla strategii epsilon=-5.20 oraz Boltzmanna=-127.27.



Rysunek 3: Wykres zależności wartości nagrody od iteracji przy zmianie parametru *discount_rate* na 0.9, *learning_rate*=0.1, *exploration*=0.2

Najlepsza wartość nagrody dla strategii epsilon=-4.44 oraz Boltzmanna=8.09.

Im mniejsza wartość *discount_rate* w algorytmie Q-learning, tym bardziej agent skupia się na krótkoterminowych nagrodach, ignorując w dużej mierze przyszłe korzyści. To podejście sprawia, że agent bardziej koncentruje się na natychmiastowych rezultatach, bez dużej uwagi dla długoterminowych konsekwencji swoich decyzji.

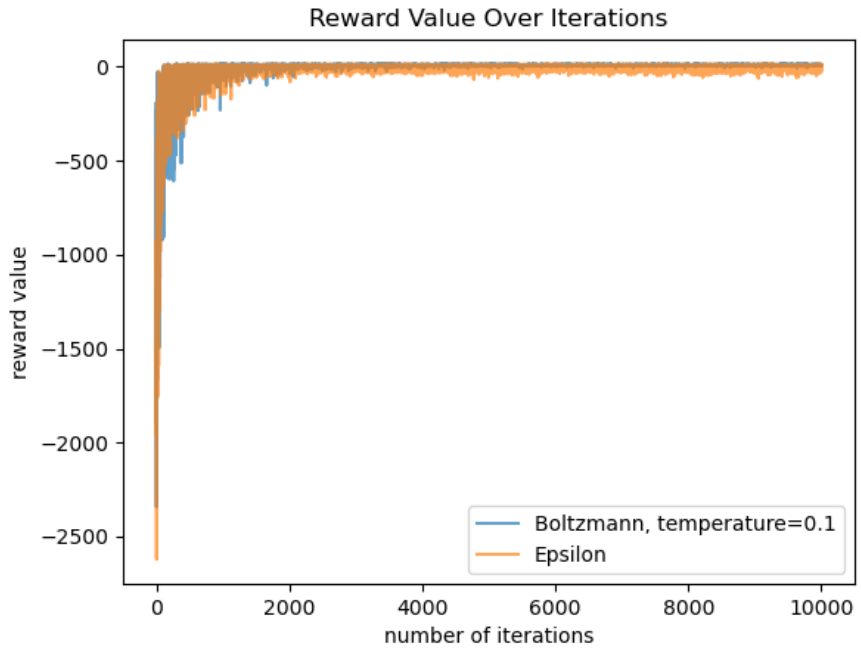
W przeciwieństwie do tego, im większa wartość *discount_rate*, tym bardziej agent przykłada wagę do przyszłych nagród. Działa to jak forma dyskontowania przyszłych korzyści, co sprawia, że agent bardziej dba o długoterminowe efekty swoich działań.

W skrócie, *discount_rate* kształtuje strategię agenta, determinując, czy bardziej opłaca mu się podejmować decyzje z myślą o natychmiastowej nagrodzie, czy też bardziej przemyślanie uwzględniać potencjalne korzyści w przyszłości.

Eksperyment z `learning_rate`

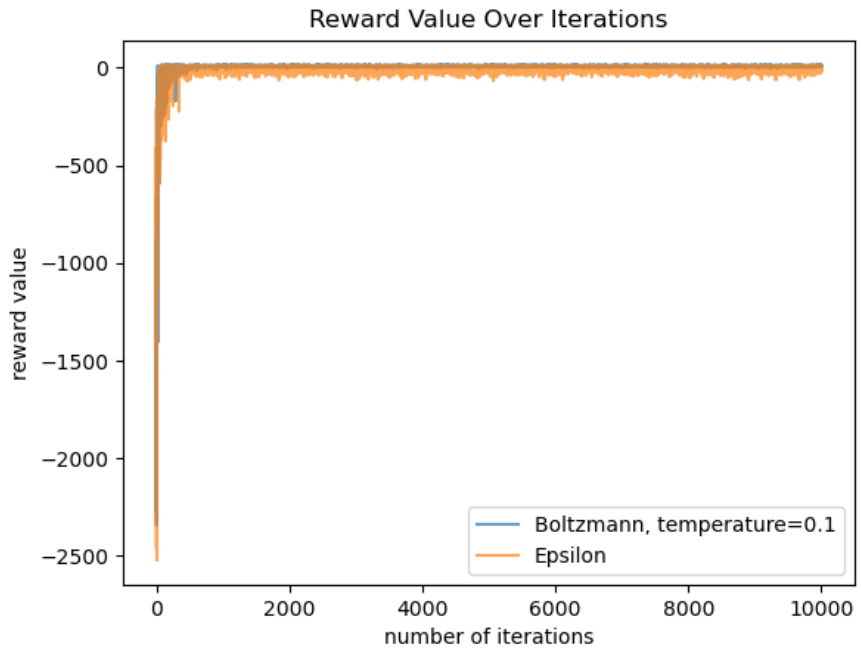
W algorytmie Q-learning, wartość *learning_rate* określa, w jakim stopniu agent uwzględnia nowe informacje (nagrody) podczas aktualizacji swoich oszacowań wartości Q. Wartość *learning_rate* zazwyczaj przyjmuje wartości między 0 a 1. Wartości te reprezentują stopień, w jakim agent uwzględnia nowe informacje w procesie aktualizacji swoich oszacowań wartości Q.
Przeprowadzono eksperyment, aby zobaczyć, jak zmiany w wartości *learning_rate* wpływają na proces uczenia się algorytmu.

Testowano różne wartości *learning_rate*, takie jak 0.1, 0.5, 0.9.



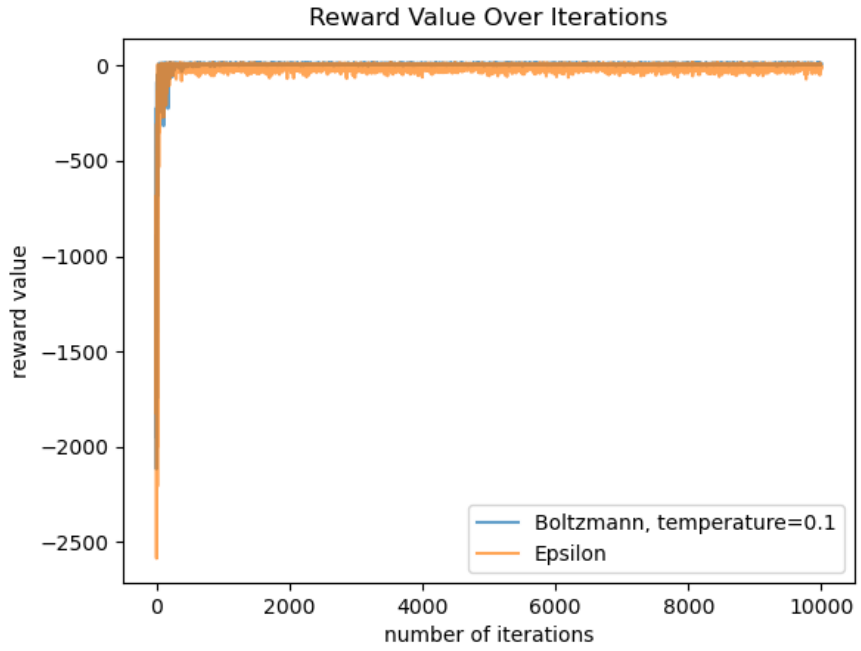
Rysunek 4: Wykres zależności wartości nagrody od iteracji przy zmianie parametru *learning_rate* na 0.1, *discount_rate*=0.8, *exploration*=0.2

Najlepsza wartość nagrody dla strategii epsilon=-4.04 oraz Boltzmanna=7.9.



Rysunek 5: Wykres zależności wartości nagrody od iteracji przy zmianie parametru *learning_rate* na 0.5, *discount_rate*=0.8, *exploration*=0.2

Najlepsza wartość nagrody dla strategii epsilon=-4.10 oraz Boltzmanna=7.93.



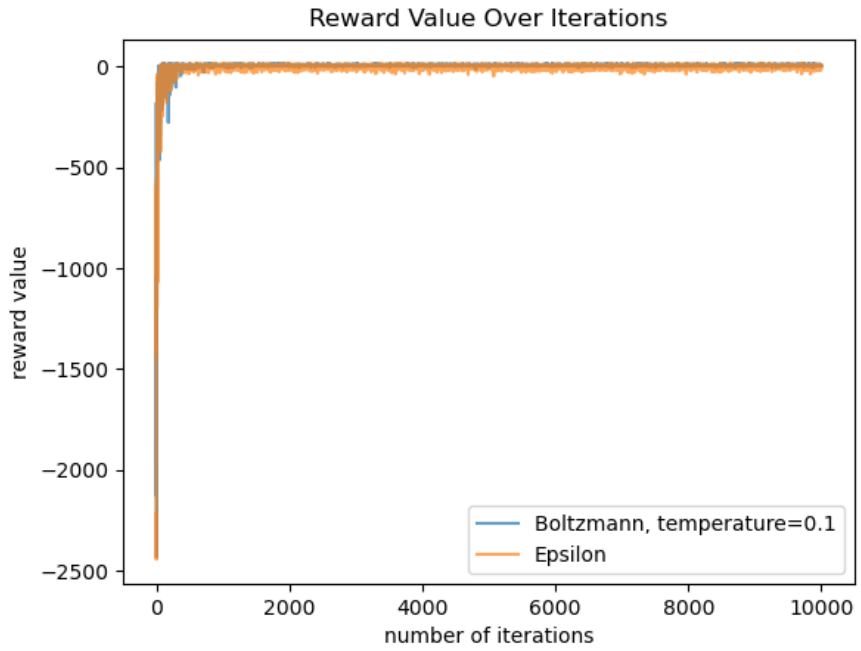
Rysunek 6: Wykres zależności wartości nagrody od iteracji przy zmianie parametru *learning_rate* na 0.9, *discount_rate*=0.8, *exploration*=0.2

Najlepsza wartość nagrody dla strategii epsilon=-4.59 oraz Boltzmanna=7.91.

Im wyższa wartość *learning_rate*, tym bardziej aktualne informacje mają większy wpływ na dostosowanie oszacowań wartości Q, co przyspiesza proces uczenia się agenta. Z drugiej strony, niższa wartość *learning_rate* sprawia, że agent bardziej ostrożnie uwzględnia nowe informacje, co może prowadzić do bardziej stabilnego, ale wolniejszego procesu uczenia. Wartość *learning_rate* pełni rolę regulacyjną w dostosowywaniu strategii agenta w trakcie interakcji z otoczeniem.

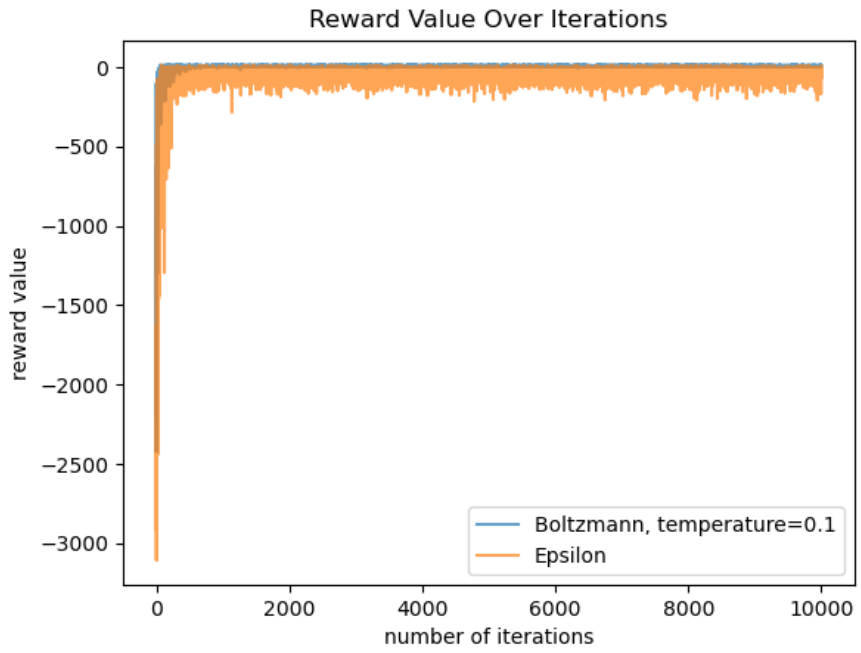
Eksperyment z `exploration`

W kontekście algorytmu Q-learning, strategia eksploracji, kontrolowana przez parametr *exploration*, odnosi się do sposobu, w jaki agent podejmuje decyzje dotyczące eksploracji nowych akcji w przestrzeni stanów, w przeciwieństwie do eksploatacji już znanego, potencjalnie korzystnego zachowania.



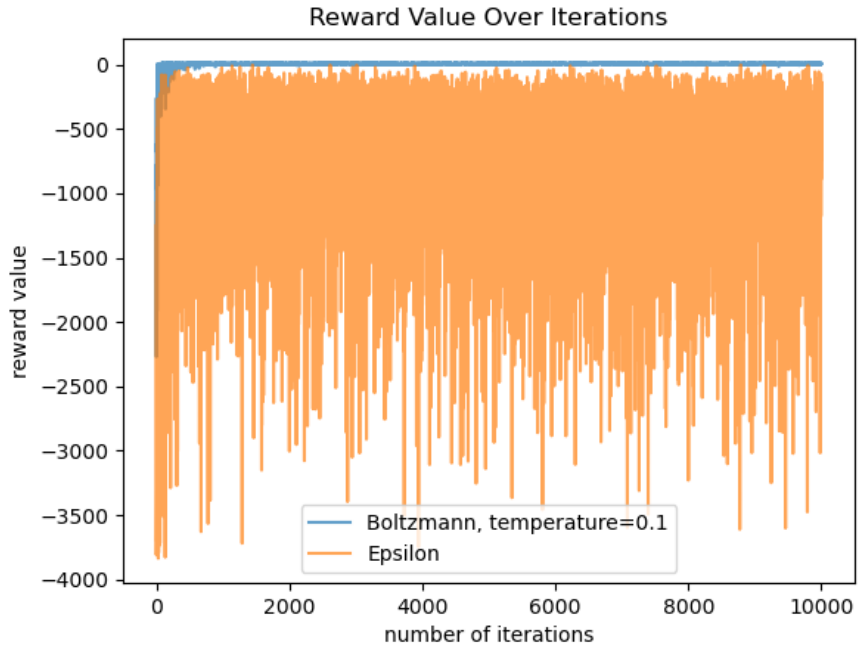
Rysunek 7: Wykres zależności wartości nagrody od iteracji przy zmianie parametru *exploration* na 0.1, *discount_rate*=0.8, *learning_rate*=0.5

Najlepsza wartość nagrody dla strategii epsilon=2.63 oraz Boltzmanna=7.86.



Rysunek 8: Wykres zależności wartości nagrody od iteracji przy zmianie parametru *exploration* na 0.5, *discount_rate*=0.8, *learning_rate*=0.5

Najlepsza wartość nagrody dla strategii epsilon=-46.26 oraz Boltzmanna=7.91.



Rysunek 9: Wykres zależności wartości nagrody od iteracji przy zmianie parametru *exploration* na 0.9, *discount_rate*=0.8, *learning_rate*=0.5

Najlepsza wartość nagrody dla strategii epsilon=-717.80 oraz Boltzmanna=7.90.

Niska wartość exploration skłania agenta do bardziej skupiania się na sprawdzonych akcjach (eksploatacji), co może prowadzić do unikania nowych, potencjalnie korzystnych rozwiązań. W przeciwnym razie, wysoka wartość exploration sprawia, że agent bardziej intensywnie eksploruje nowe akcje, co pomaga odkryć lepsze strategie w dłuższej perspektywie. W skrócie, niskie exploration faworyzuje eksploatację, podczas gdy wysokie exploration stawiają nacisk na eksplorację.

Porównanie obu strategii eksploracji

Wykresy porównujące obie strategie zostały przedstawione w powyższych eksperymentach. Strategia eksploracyjna w algorytmie Q-learning determinuje, w jaki sposób agent podejmuje decyzje w trakcie uczenia, wpływając bezpośrednio na skuteczność optymalizacji. Oto krótki opis różnic między dwiema popularnymi strategiami:

Rozkład Boltzmanna:

- Równowaga między eksploracją a eksploatacją jest regulowana poprzez temperaturę.
- Stabilne osiąganie pozytywnych nagród z umiarkowaną liczbą iteracji.
- Elastyczność w adaptowaniu się do zmiennych warunków środowiska.

Epsilon:

- Decyzje podejmowane z określonym prawdopodobieństwem, gdzie epsilon kontroluje eksplorację.
- Może prowadzić do większej zmienności wyników, zarówno pod względem nagród, jak i liczby iteracji.
- Mniej elastyczna w długotrwałych zadaniach, ze stałym prawdopodobieństwem eksploracji.

Wybór strategii eksploracyjnej zależy od charakterystyki danego zadania i pożądanych cech uczenia agenta. Stabilność, adaptacyjność, oraz złożoność obliczeniowa są kluczowymi kryteriami do rozważenia przy dokonywaniu tego wyboru.

4 Podsumowanie

Podsumowując, wyniki eksperymentów wskazują na to, że pewne kombinacje hiperparametrów i strategii eksploracji mogą prowadzić do lepszych wyników w uczeniu się algorytmu Q-learning w środowisku `Taxi`. Dostosowywanie tych parametrów może być kluczowe dla osiągnięcia optymalnej wydajności algorytmu w konkretnym środowisku.
Wartość *learning_rate* reguluje tempo aktualizacji oszacowań wartości Q, decydując o balansie między stabilnością a szybkością uczenia. Parametr *discount_rate* wpływa na to, czy agent skupi się bardziej na nagrodach krótkoterminowych czy długoterminowych. Zmieniając wartość *exploration*, można dostosować strategię eksploracji, decydując o intensywności poszukiwania nowych akcji.