Groszyk Maciej 289761

Sprawozdanie WSI - ćwiczenie 6 Q Learning

Zaimplementuj algorytm Q-learning. Następnie, wykorzystując proste środowisko (np. Taxi-v3), zbadaj wpływ hiperparametrów na działanie algorytmu (np. wpływ strategii eksploracji, współczynnik uczenia).
Pamiętaj, że implementacja musi być wykonana samodzielnie. Brak zrozumienia dostarczonego kodu rozwiązania równoważny jest plagiatowi!

Rozwiązanie:

Zadanie zostało rozwiązane przy pomocy algorytmu q learning z zastosowaniem dwóch strategii eksploracji. Strategia ε-zachłanna i strategia boltzmanna.

Strategia z użyciem epsilona , polega na wybraniu z pewnym prawdopodobieństwem epsilon >0 dowolną akcję losowo według rozkładu równomiernego, a z prawdopodobieństwem 1-epsilon} wybiera się akcję zachłanną (jeśli jest ich wiele, to także losowo). Formalnie można to zapisać następująco:

$$\pi(x,a^*) = \begin{cases} \frac{1-\epsilon}{|\operatorname{Arg\,max}_a Q(x,a)|} + \frac{\epsilon}{|A|} & \text{jeśli } a^* \in \operatorname{Arg\,max}_a Q(x,a), \\ \frac{\epsilon}{|A|} & \text{w przeciwnym przypadku.} \end{cases}$$

Strategia z zachłannym epsilonem posiada wadę, że prawdopodobieństwo losowego zachowania się ucznia nie zależy od tego czego zdołał się nauczyć. Jednym ze sposobów pozbycia się tego problemu jest zastosowanie strategii boltzmanna. Która jest opisana wzorem:

$$\pi(x,a^*) = \frac{\exp(Q(x,a^*)/T)}{\sum_a \exp(Q(x,a)/T)},$$

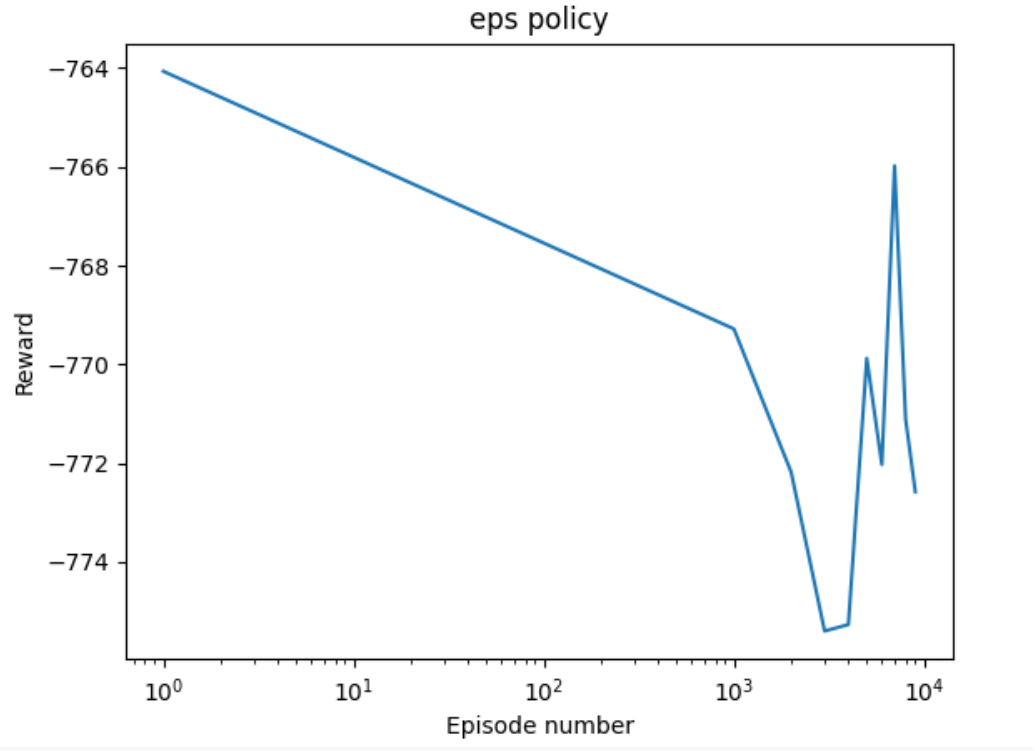
Współczynnik T (temperatura) > 0 reguluje stopień losowości. Przy tej strategii prawdopodobieństwo wyboru akcji niezachłannej jest tym mniejsze, im bardziej akcja zachłanna ma większą Q -wartość od pozostałych.

Testy:

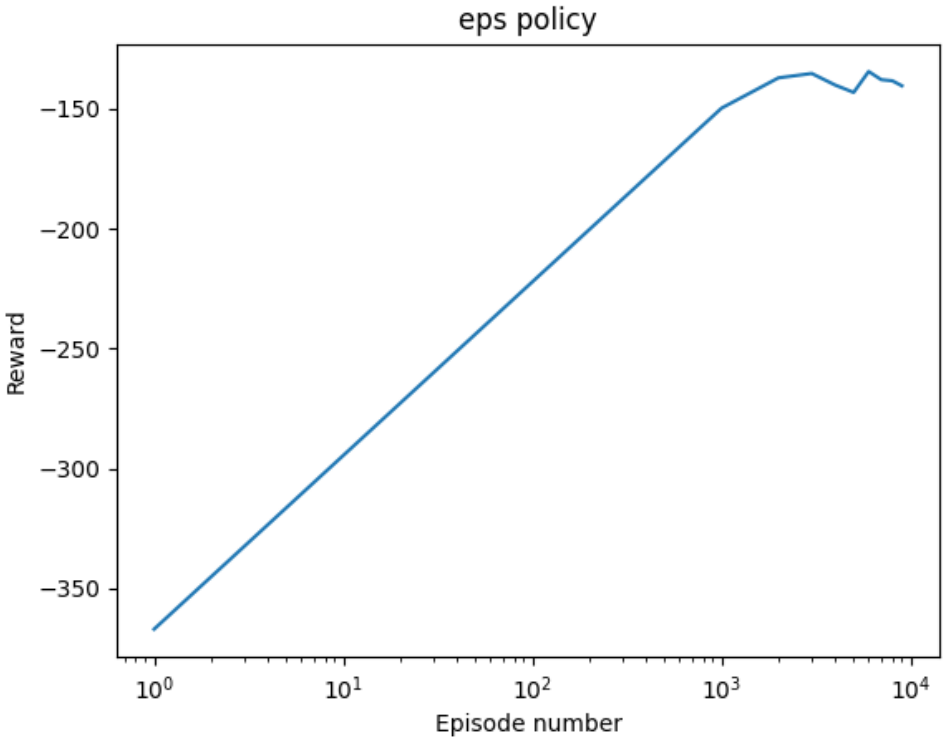
Strategia zachłannego epsilona:

Badanie zachowania algorytmu przy zmianie eps. Pozostałe parametry (discount factor = 1, learning rate = 0.1, episodes = 10000, steps = 1000)

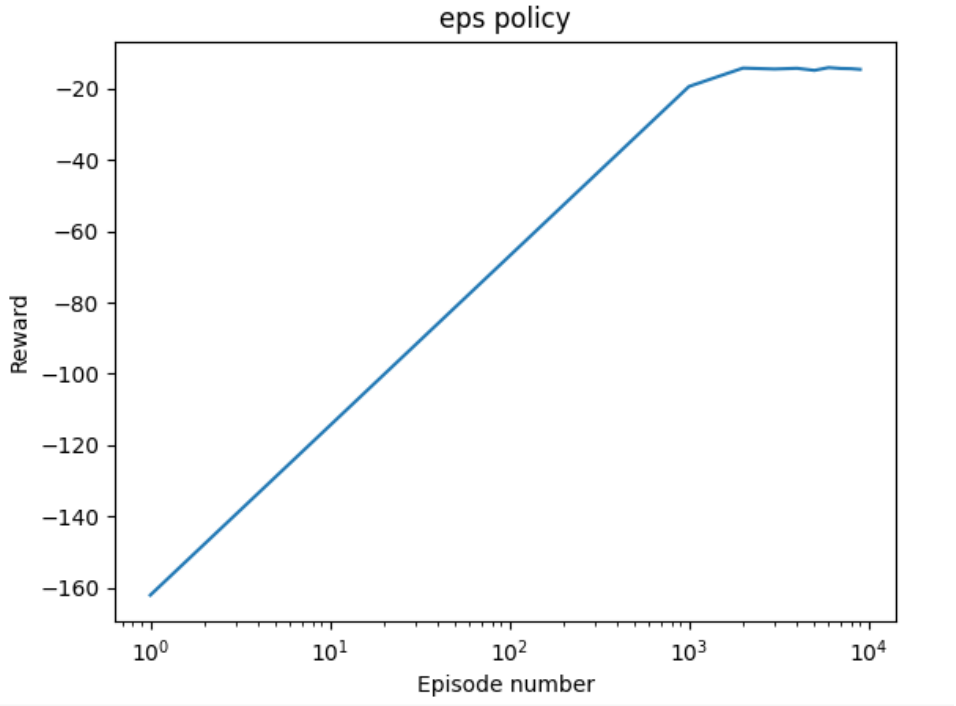
Przy epsilonie = 1 zgodnie z definicją strategii zauważymy 100% losowość:



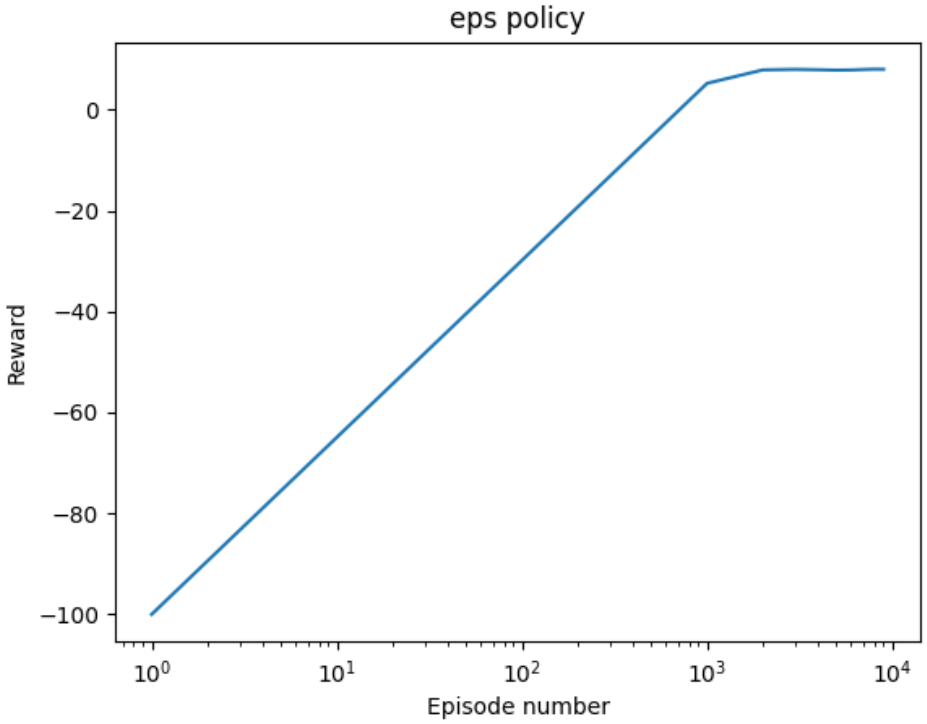
Przy stopniowym zmniejszaniu epsilona dochodzimy do coraz lepszych wyników i jednocześnie wyższych nagród. Poniżej wykres dla eps=0.7



Dla eps=0.3

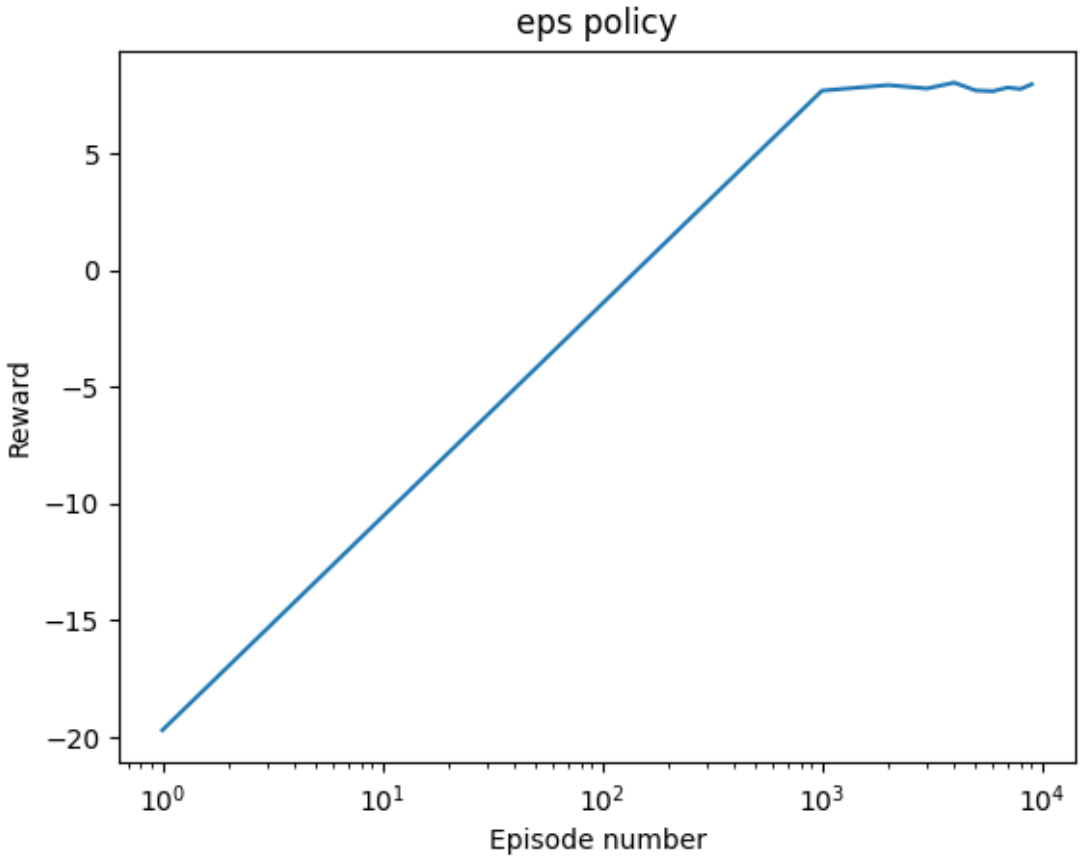


Dla epsilona równego 0 otrzymujemy najlepsze wyniki



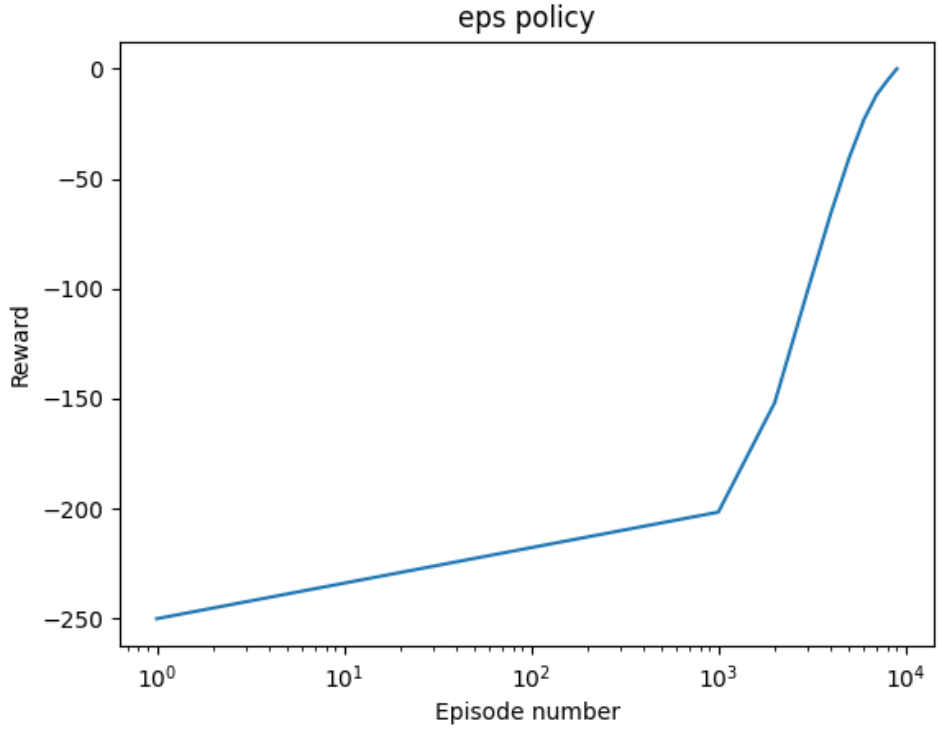
Badanie wpływu zmiany parametru learning rate:

Podczas gwałtownej zmiany learning rate do wartości 0.9 możemy zaobserwować wzrost maksymalnej wartości nagrody. (eps = 0)

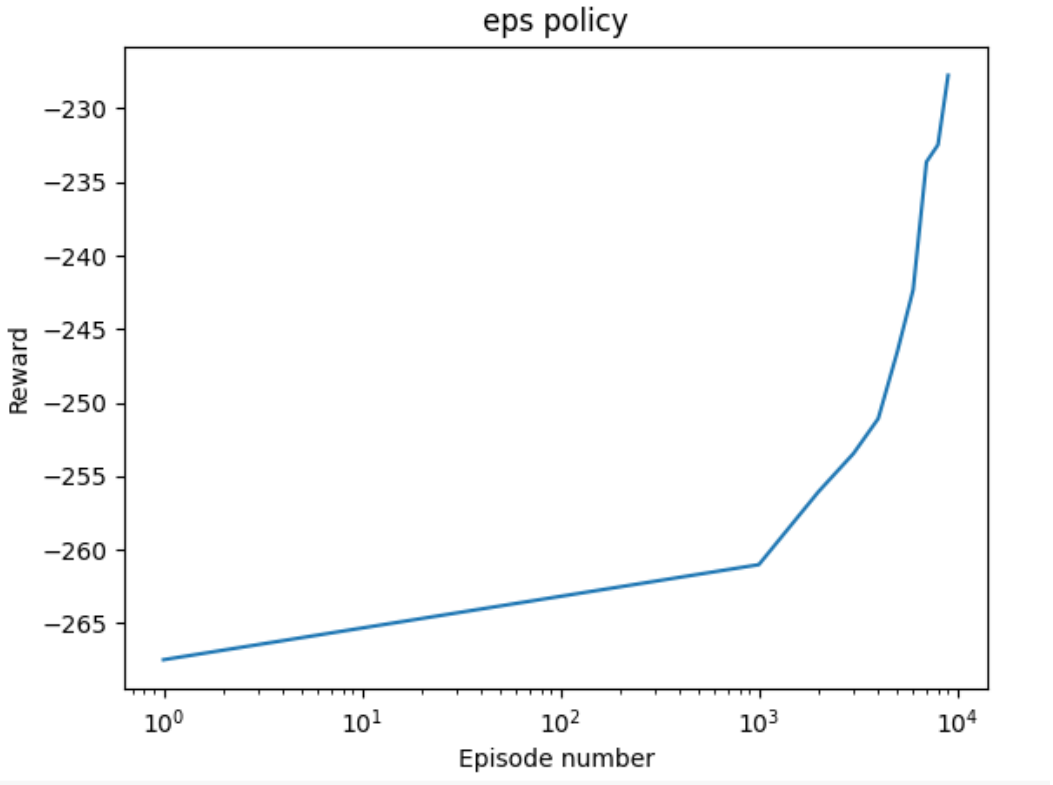


Natomiast dla dużo mniejszych wartości learning rate, algorytm potrzebowałby więcej iteracji aby otrzymał akceptowalne wyniki.

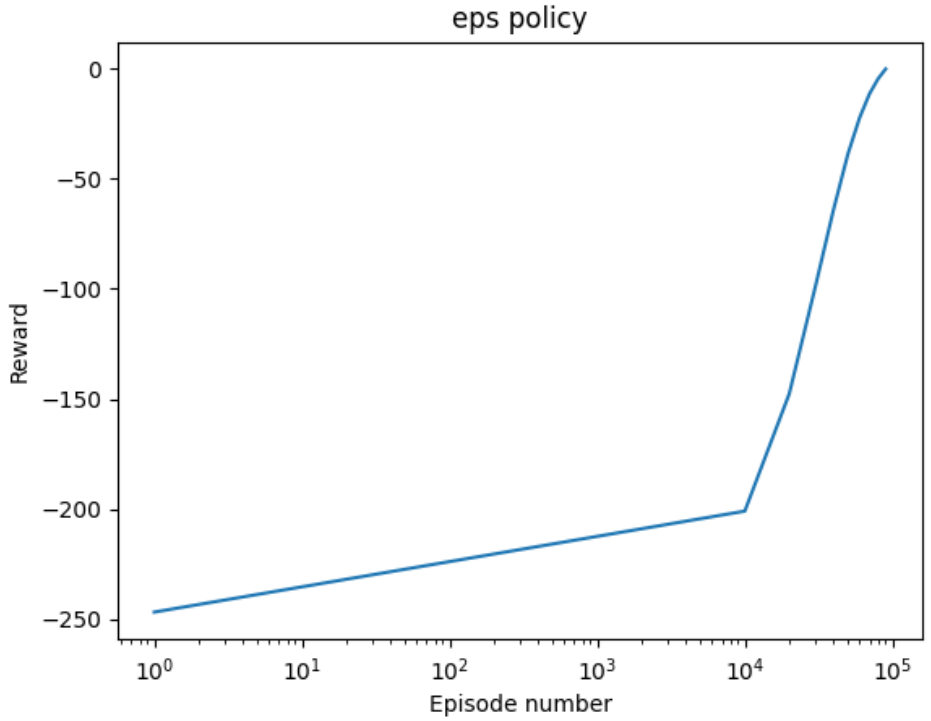
Dla learning rate = 0.01



W przypadku zmniejszenia wartości do 0.001



można zauważyć że w tym przypadku algorytm miał przyjętą za małą ilość epizodów jak na tak mały learning rate. W momencie zwiększenia ilości epizodów algorytm mimo bardzo małej wartości learning rate dochodzi do średniej wartości nagród w okolicy 0.



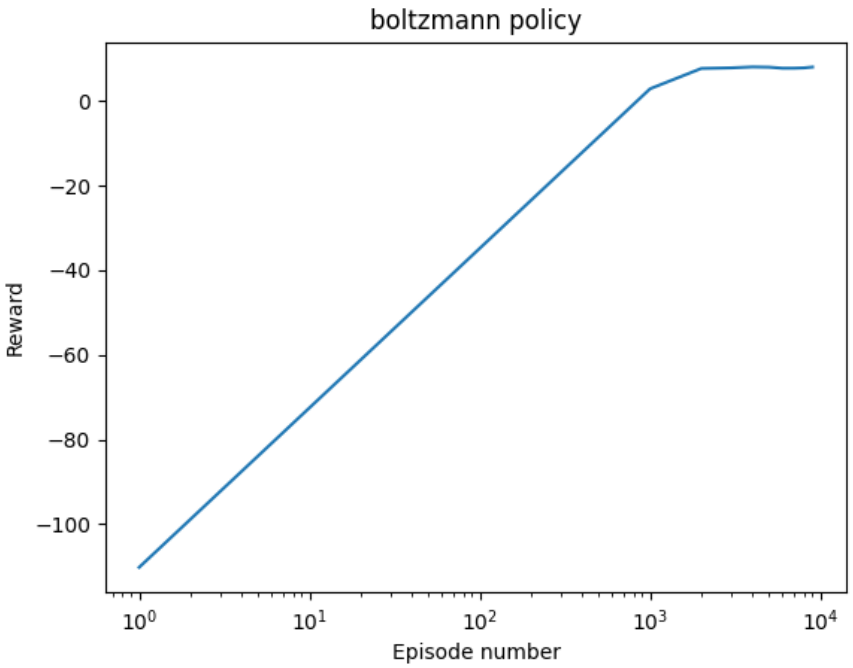
Wnioski: metoda zachłannego epsilona przy dobraniu odpowiednich parametrów spełnia swoje zadanie. Zmiana epsilona działa zgodnie z teorią.Manipulacja parametrami dot. learning rate i ilości epizodów ma znaczący wpływ na wynik końcowy algorytmu. W

przypadku dobrania mniejszego learning rate'u należy zwiększać liczbe epizodów jeżeli chcemy otrzymać podobne wyniki.

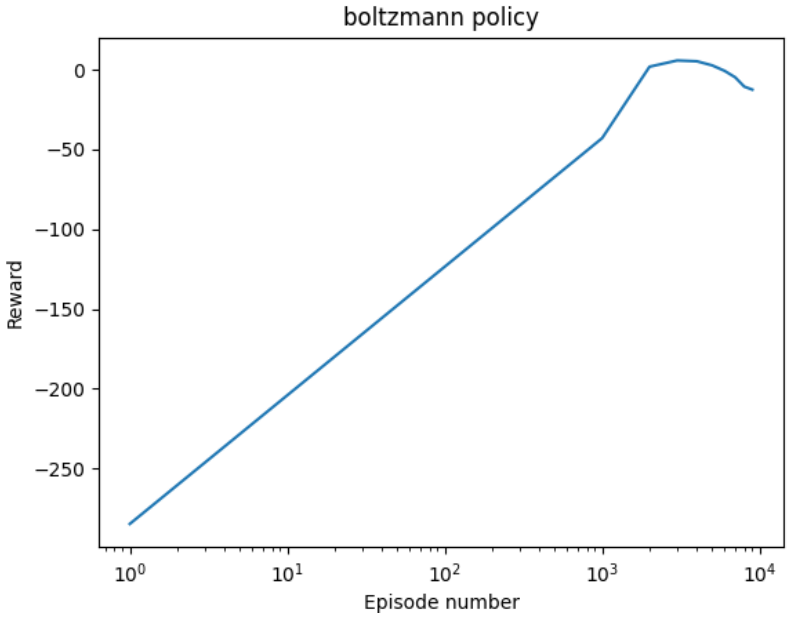
Strategia Boltzmanna

Wpływ parametru temperatury na algorytm:

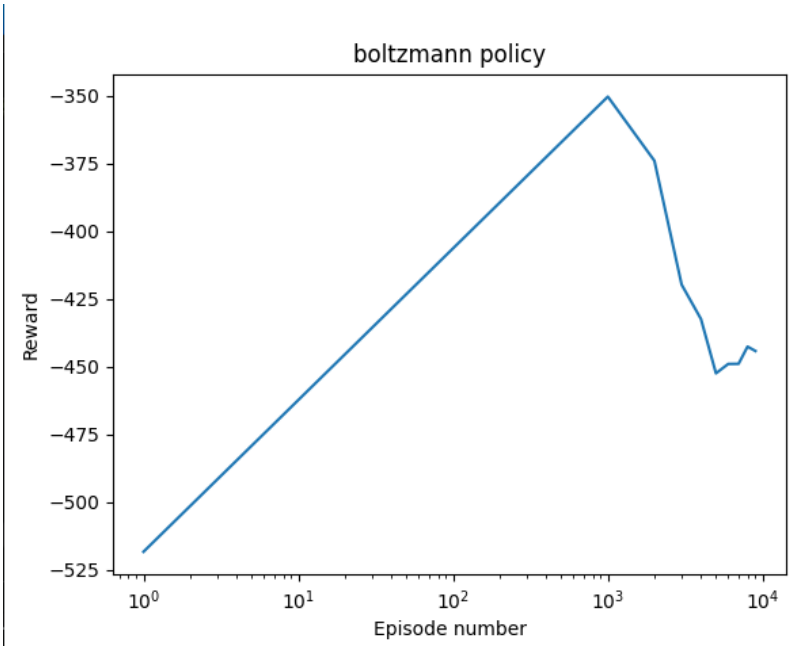
Dla T = 0.1



Dla T = 3

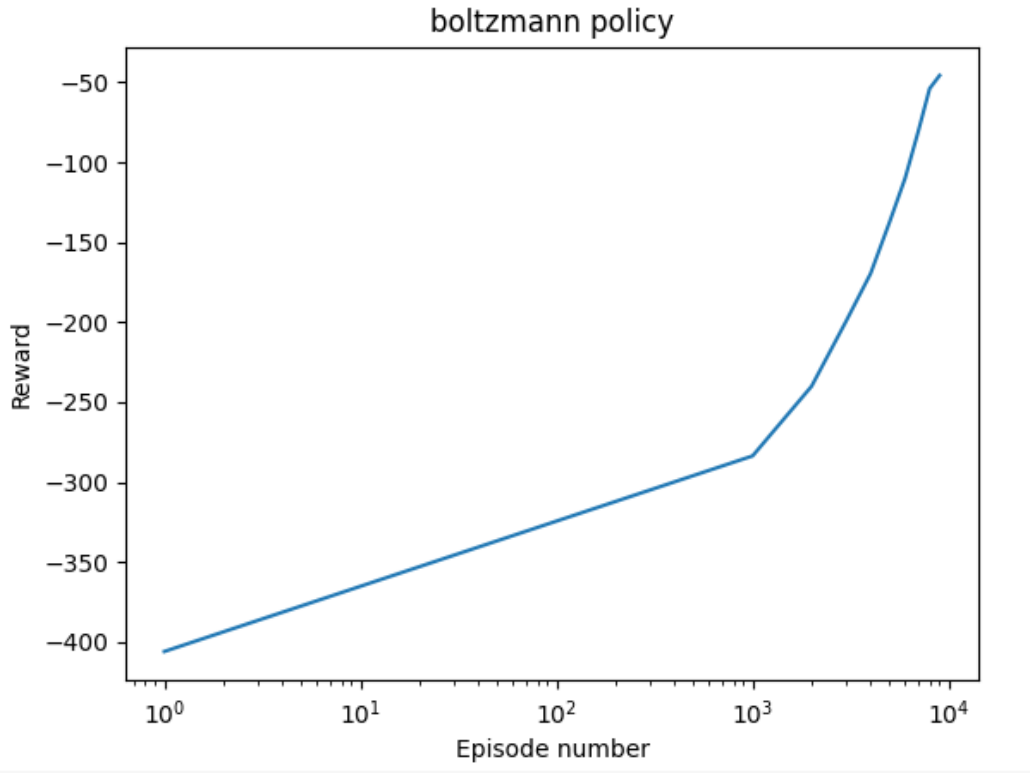


Dla T = 10



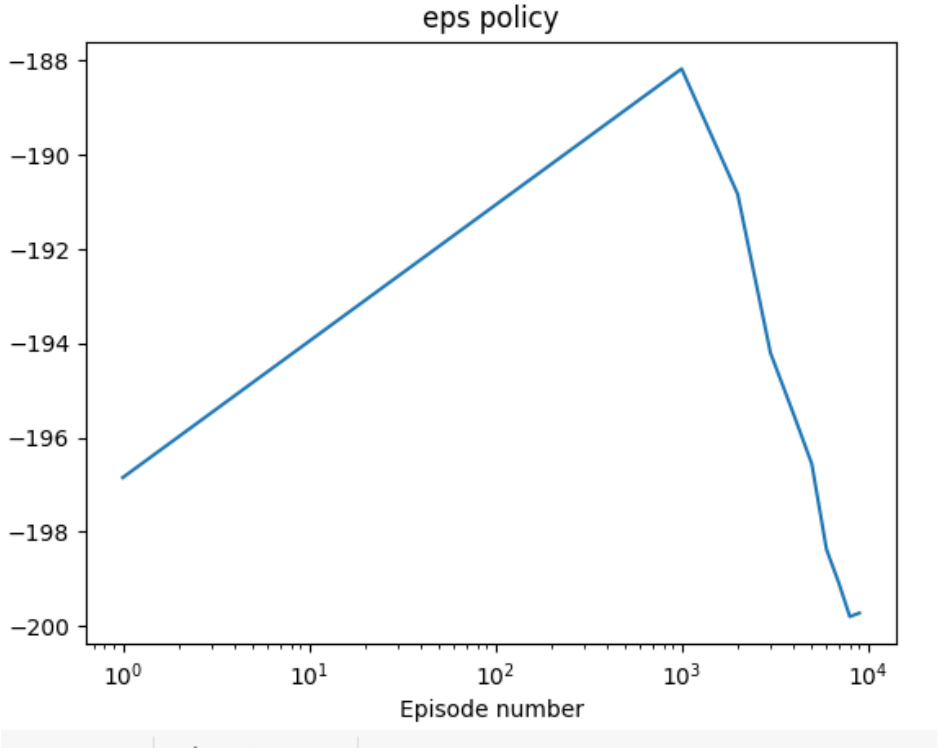
Zwiększając parametr temperatury algorytm zaczyna działać z większa losowością co jest zgodne z oczekiwaniami.

Strategia boltzmana zachowuje się podobno do strategi zachłannego epsilona jeżeli chodzi o wpływ learning rate na wykres nagrody od ilości epizodów. Dla małego learning rate = 0.01

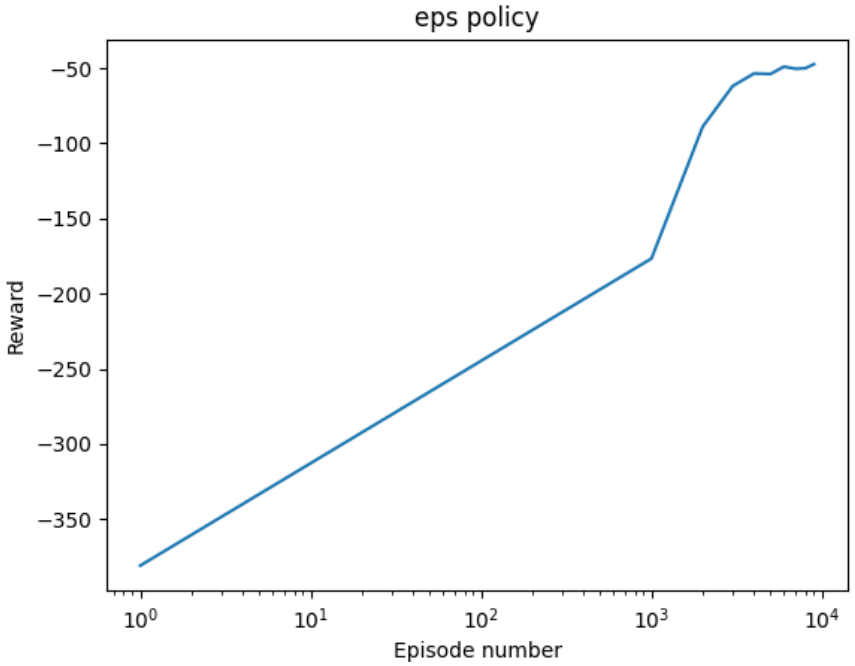


Wpływ wartości discount factor:

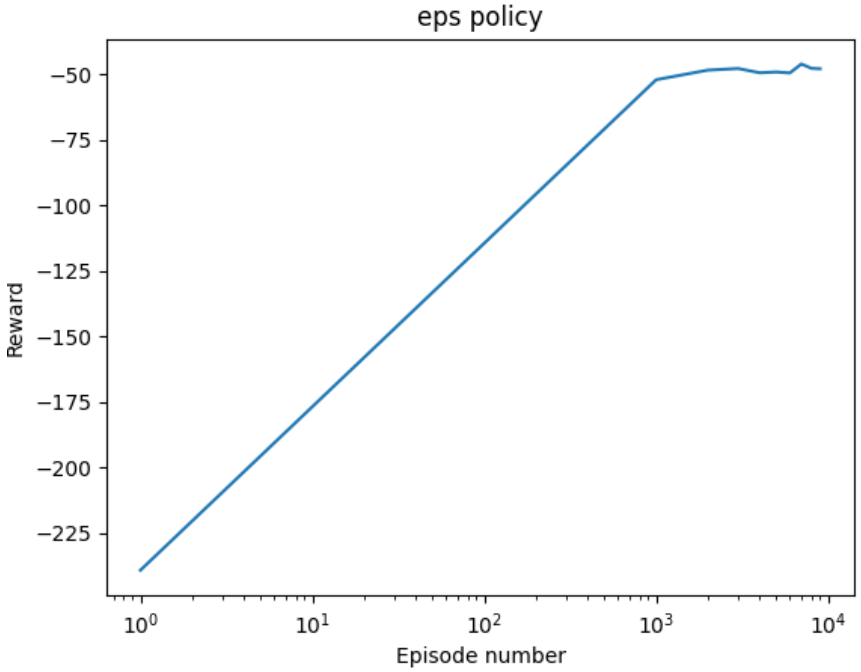
W przypadku ustawienia wartości discount factor na 0, algorytm nie będzie patrzył ‘w przód’ i będzie zwracał uwagę jedynie na natychmiastowe nagrody co spowoduje brak otrzymania prawidłowego rozwiązania zadania.



Natomiast przy tym parametrze ustawionym na wartość 0.3 przy epsilonie = 0.5



discount factor = 1 przy eps = 0.5



Wnioski:

Algorytm działa zgodnie z założeniami. Przy doborze odpowiednich parametrów zarówno strategia boltzmanna jak i zachłannego epsilona radzi sobie z uczeniem i zrealizowaniem zadania taxi. Najbardziej znaczącymi parametrami jest epsilon i temperatura, od których zależy skuteczność/prawdopodobieństwo wybrania losowego. Discount factor również nie może być zbyt mały ponieważ algorytm będzie brał pod uwagę jedynie natychmiastowe nagrody