WSI Laboratorium 6
Algorytm Q-learning

Filip Misztal 310276

1 marca 2024

1 Opis Implementowanych Algorytmów

Celem tego ćwiczenia było zaimplementowanie algorytmu Q-learningu, który będzie rozwiązywał zadanie ze środowiska "Taxi-v3"pochodzącego z biblioteki Gymnasium.

Uczenie ze wzmocnieniem - algorytm Q-learning (pseudokod)

1. Inicjalizuj Q-wartości dla wszystkich stanów i akcji $Q(s, a)$
2. Dla każdej epoki (episode):
   (a) Zainicjalizuj stan początkowy (s)
   (b) Powtarzaj, dopóki nie osiągniesz stanu końcowego:
      i. Wybierz akcję a na podstawie strategii epsilon-zachłannej (może być losowe eksplorowanie z prawdopodobieństwem epsilon)
      ii. Wykonaj akcję a, przechodząc do nowego stanu s' i otrzymując nagrodę r
      iii. Oblicz nową Q-wartość dla pary (s, a) za pomocą równania Q-learning: $Q(s, a) = Q(s, a) + \alpha * [r + \gamma * max(Q(s', a')) - Q(s, a)]$
      iv. Przejdź do nowego stanu s'
   (c) Powtarzaj powyższe kroki przez ustaloną liczbę epok
3. Zwróć wytrenowane wartości Q

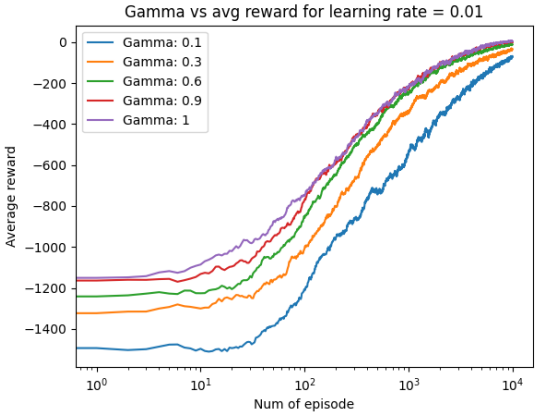
2 Opis planowanych eksperymentów numerycznych

W celu przetestowania możliwości zaimplementowanego algorytmu przeprowadziłem test badający osiągi wytrenowanych wartości Q przy zastosowaniu różnych wartości learnign rate'u i parametru γ. Pozostałe hiperparametry ustawiłem na wartości:

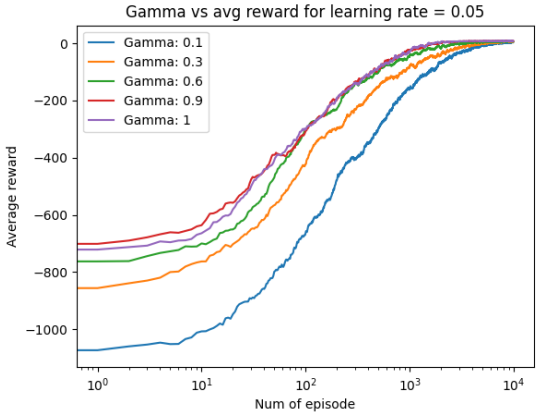
1. Początkowa wartość ϵ: 1
2. Redukcja ϵ: 0.99
3. Minimalna wartość ϵ: 0.005

3 Opis uzyskanych Wyników

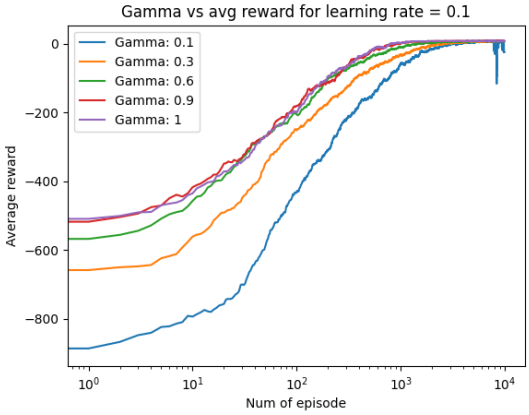
Poniższa analiza jest przeprowadzona na podstawie średnich wyników jakie osiągnął algorytm. Średnie są wyliczane w ten sposób, że w każdym epizodzie do kolejki o ustalonej długości dodawana jest końcowa średnia nagroda zdobyta w danej iteracji. Po 100 iteracjach do drugiej kolejki zaczynam dodawać średnią wartość nagrody wyciągniętą z wartości pierwszej kolejki.



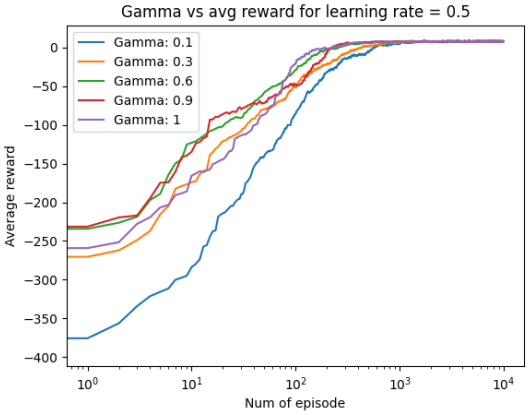
(a)



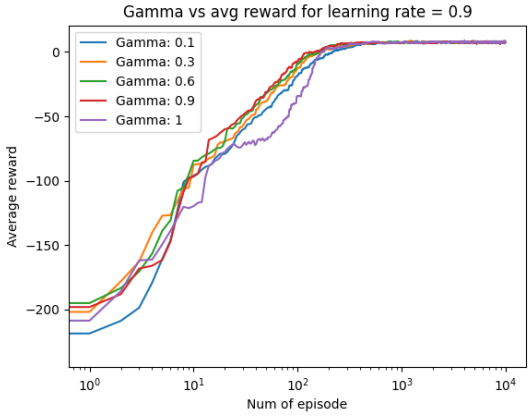
(b)



(c)



(d)



(e)

Rysunek 1: Porównanie wpływu wartości hiperparametrów *learnignrate* oraz γ na średnią wysokość nagrody osiąganej przez algorytm

4 Wnioski z przeprowadzonych badań

1. Na podstawie przeprowadzonych badań można wyciągnąć wniosek, że w tym przypadku najoptymalniejsze jest użycie wartośći γ bliskiej 1. Jest to związane z tym, że wartość tego parametru wpływa na sposób, w jaki algorytm rozpatruje nagrody (gdy $\gamma = 0$ - algorytm zaniedbuje przyszłe nagrody skupiając się tylko na natychmiastowych, $\gamma = 1$ - algorytm bierze pod uwagę wszystkie przyszłe

nagrody równie ważnie)

2. Epsilon to parametr odpowiadający za to jak dużą skłonność do eksploracji i eksploatacji ma algorytm. Zdecydowałem się go stopniowo redukować w trakcie trenowania aż do momentu osiągnięcia ustalonej minimalnej wartości, ponieważ waga czynnika eksploracji maleje z czasem i lepsze efekty daje poleganie na eksploatacji

3. Współczynnik uczenia pozytywnie wpływał na działanie algorytmu dla dosyć dużych jego wartości, jak np. 0.8. Niższe wartości znacząco zmniejszały prędkość zbiegania algorytmu do optymalnej średniej nagrody, a także czasami skutkowały obniżeniem stabliności procesu uczenia