Sprawozdanie Laboratorium 6 Uczenie ze wzmocnieniem

Piotr Gierżatowicz-Sierpień 331376

Cel zadania

Należało zaimplementować algorytm qlearning, który miał wskazać optymalną politykę dla agenta zarządzającego stanem magazynowym sklepu. Agent otrzymuje nagrodę = 1 przy akcji sprzedaży gdy stan > 0, oraz nagrodę = 100 przy akcji kupienia, gdy stan = 9. Należało również przeprowadzić wpływ parametrów algorytmu na znajdowanie optymalnej polityki.

Implementacja

Parametry wykorzystane w badaniu zachowania algorytmu:

Alpha = 0.0005, 0.001, 0.005, 0.01, 0.02, 0.05, 0.1, 0.15, 0.2, 0.3 <- learning_rate

Gamma = 1.0

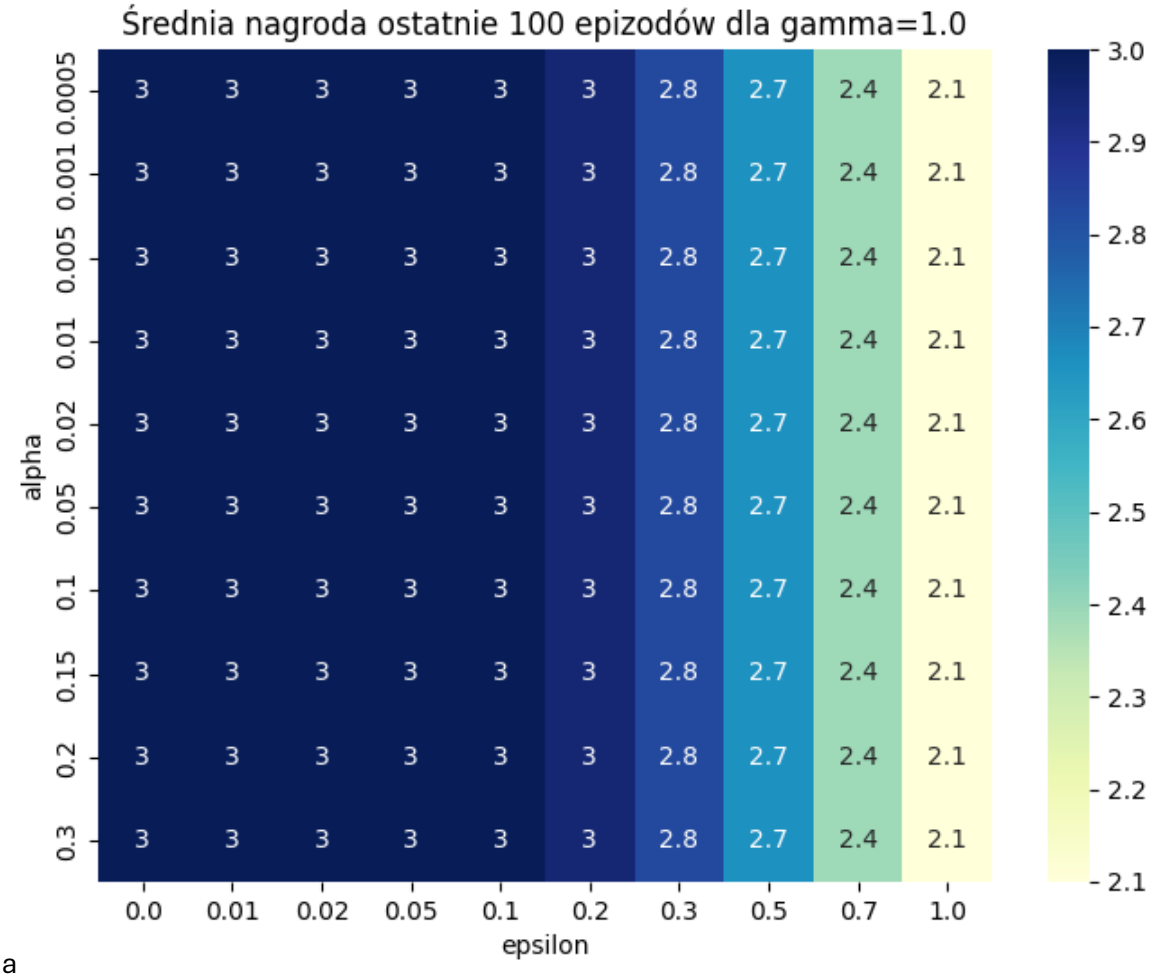
Epsilon = 0.0, 0.01, 0.02, 0.05, 0.1, 0.2, 0.3, 0.5, 0.7, 1.0 <- współczynnik eksploracji

Liczba epizodów w każdej symulacji = 100000

Eksperymenty

Badanie wpływu parametru H – horyzontu

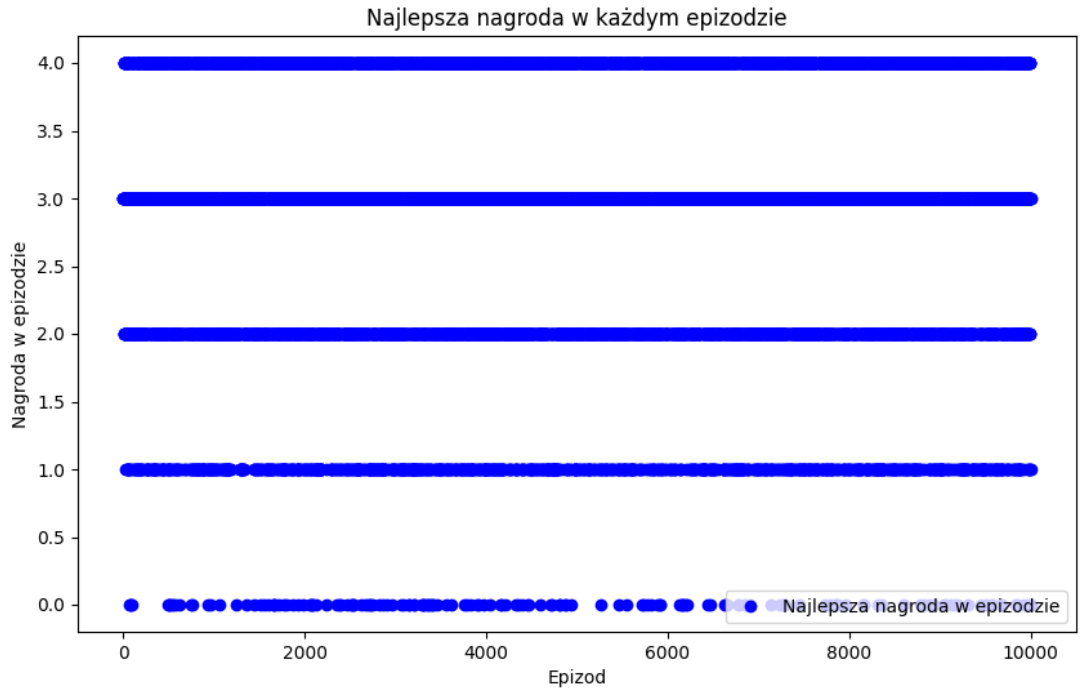
H = 4

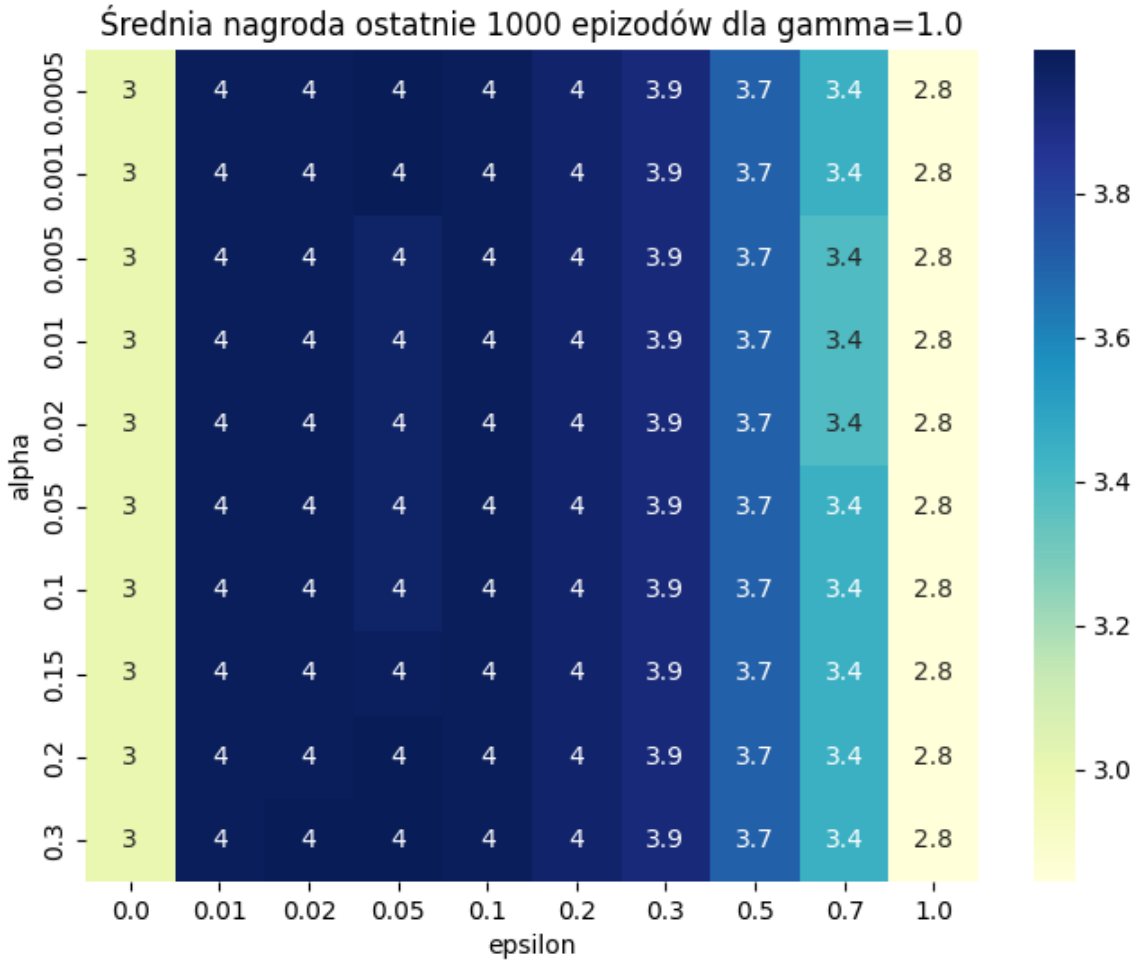


a

Na powyższej heatmapie, widzimy jakie wartości osiąga agent dla H = 4, widać tendencję przy wartościach epsilon bliskich 0, że agent wybiera szybki zysk, zamiast eksploracji i dzięki temu przy małym horyzoncie osiąga najwyższe w tym przypadku nagrody. Z drugiej strony widzimy, że agenci z wyższą tendencją do eksploracji, osiągają średnio gorsze wyniki.

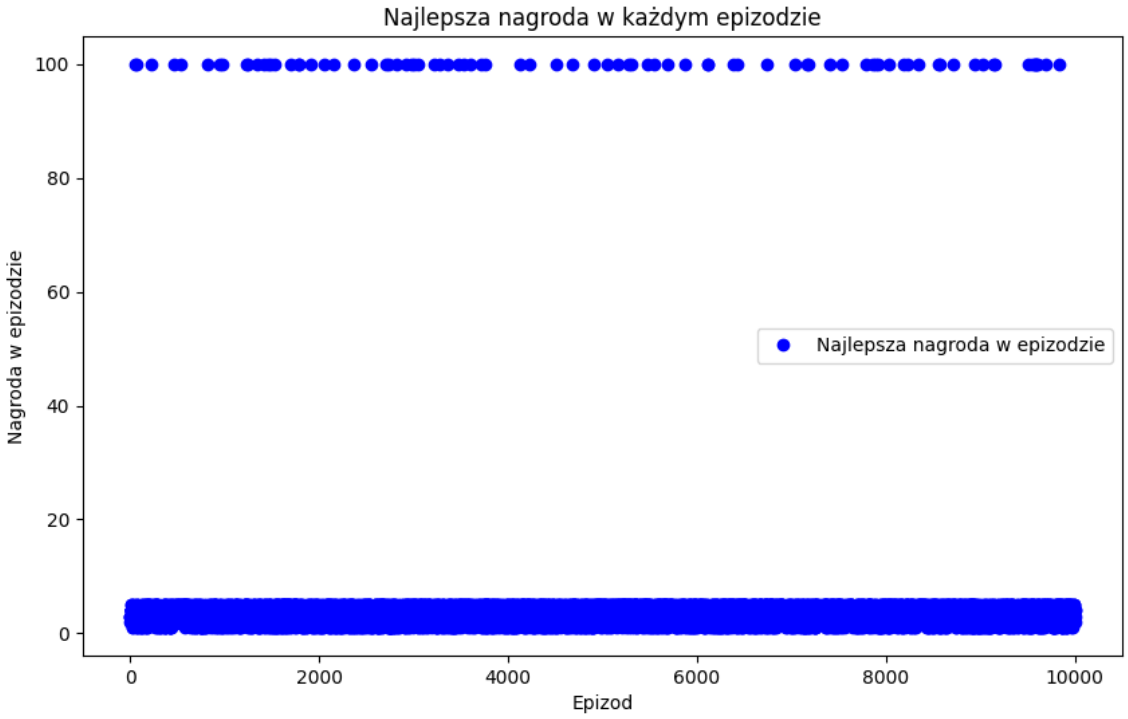
H = 6

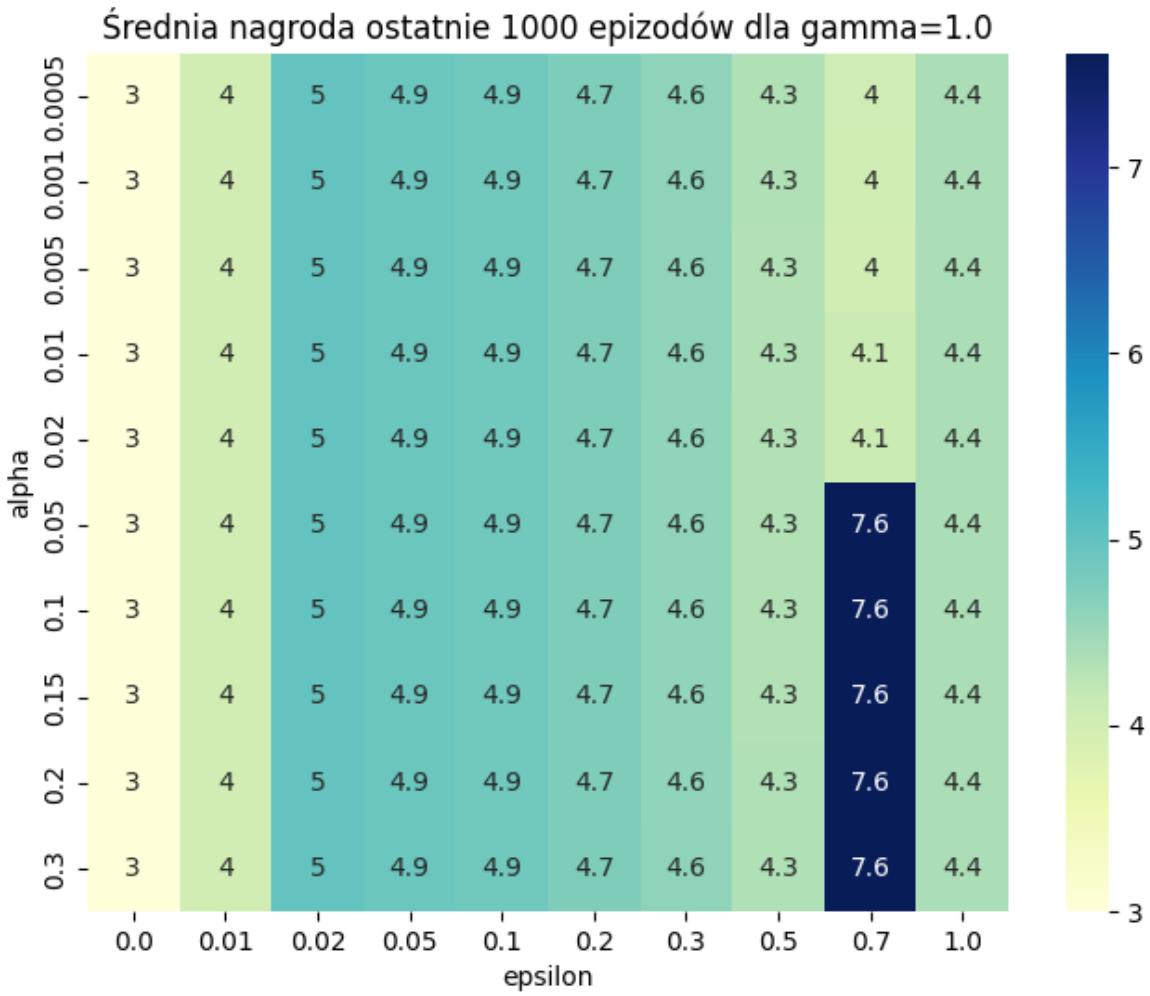

Najlepsza nagroda w każdym epizodzie
Nagroda w epizodzie
Epizod
Najlepsza nagroda w epizodzie
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
0
2000
4000
6000
8000
10000


Średnia nagroda ostatnie 1000 epizodów dla gamma=1.0
alpha
epsilon
0.0005
0.001
0.005
0.01
0.02
0.05
0.1
0.15
0.2
0.3
0.0
0.01
0.02
0.05
0.1
0.2
0.3
0.5
0.7
1.0
3 4 4 4 4 4 3.9 3.7 3.4 2.8
3.8
3.6
3.4
3.2
3.0

W przypadku H = 6, widzimy, że agenci nie są w stanie osiągać stanu 10, przez co nadal najlepsze nagrody, daje strategia względnie negatywnie nastawiona do eksploracji. Widzimy, że im agent chętnie eksploruje, tym niższe wyniki osiąga.

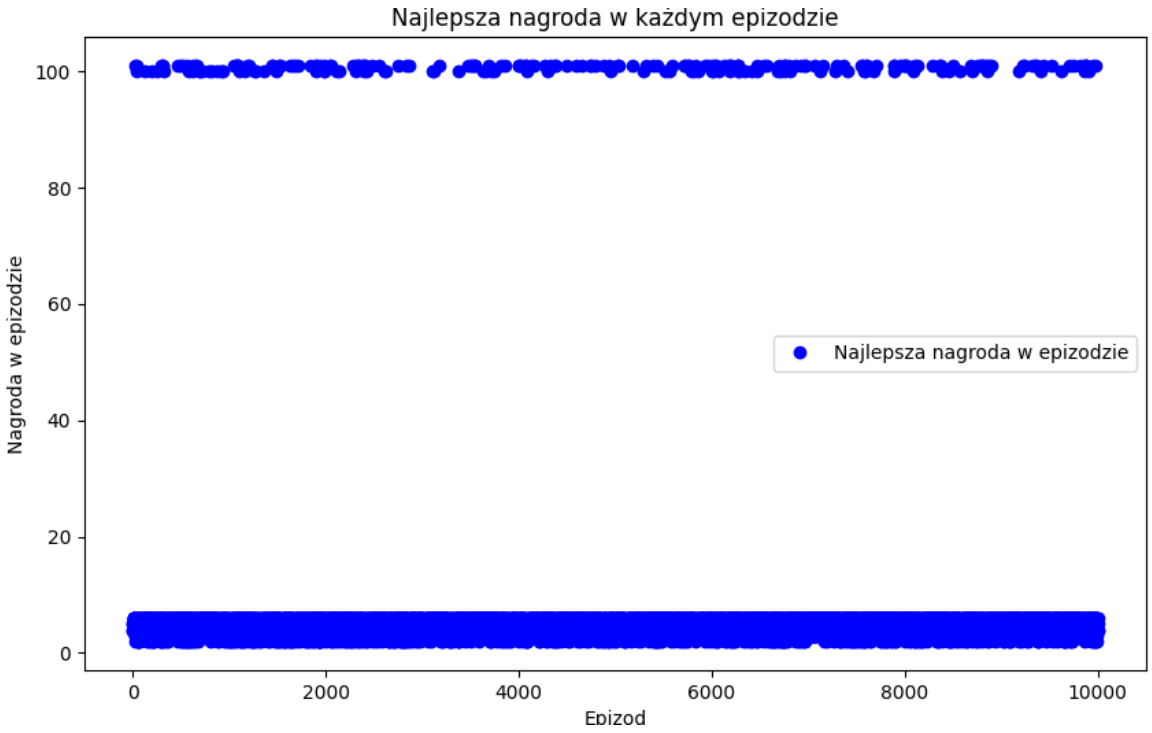
H = 7

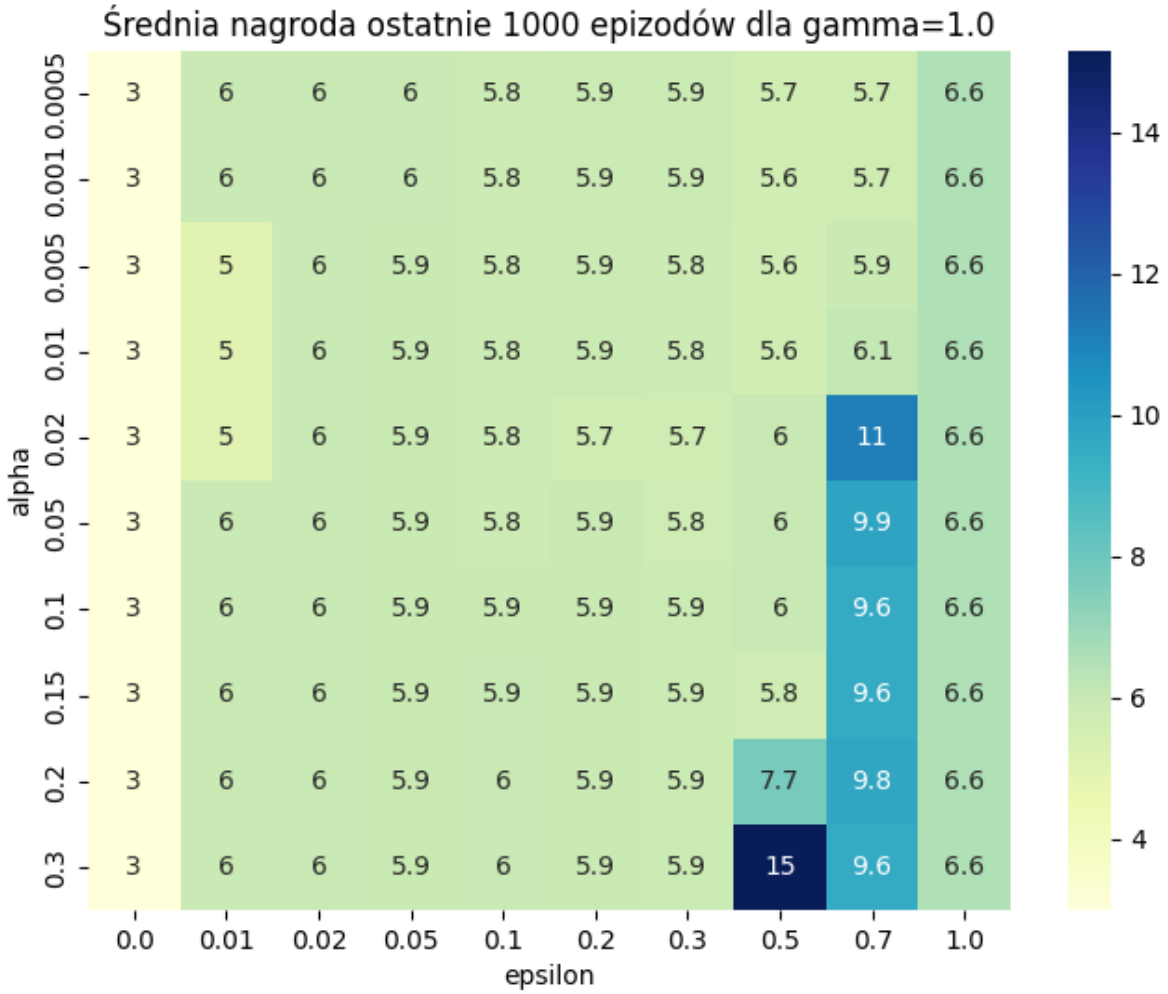




Dla H = 7 widzimy, że jest to horyzont przełomowy. Pojedynczy agenci są w stanie osiągnąć wartości nagród > 100, jednak zdecydowana większość ma nagrody w zakresie 4-5. Na heatmapie, widzimy, że najlepsze wyniki osiągane są dla epsilon = 0.7 i alpha 0.3-0.05. Jednak przy horyzoncie = 7, nagroda 7 oznacza 7 akcji sprzedaży co jest niemożliwe, bez żadnych akcji kupno, przez co wiemy, że średnia jest zawyżona przez agentów, którzy osiągnęli nagrodę za wejście do stanu 10.

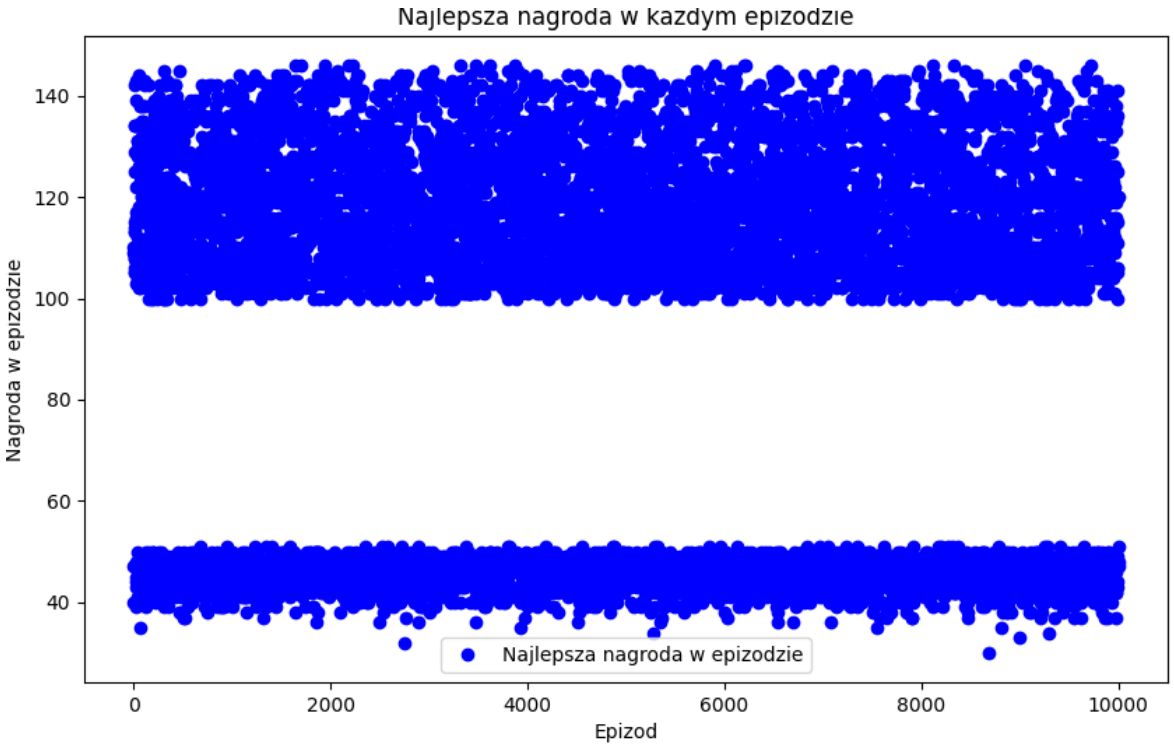
H = 10

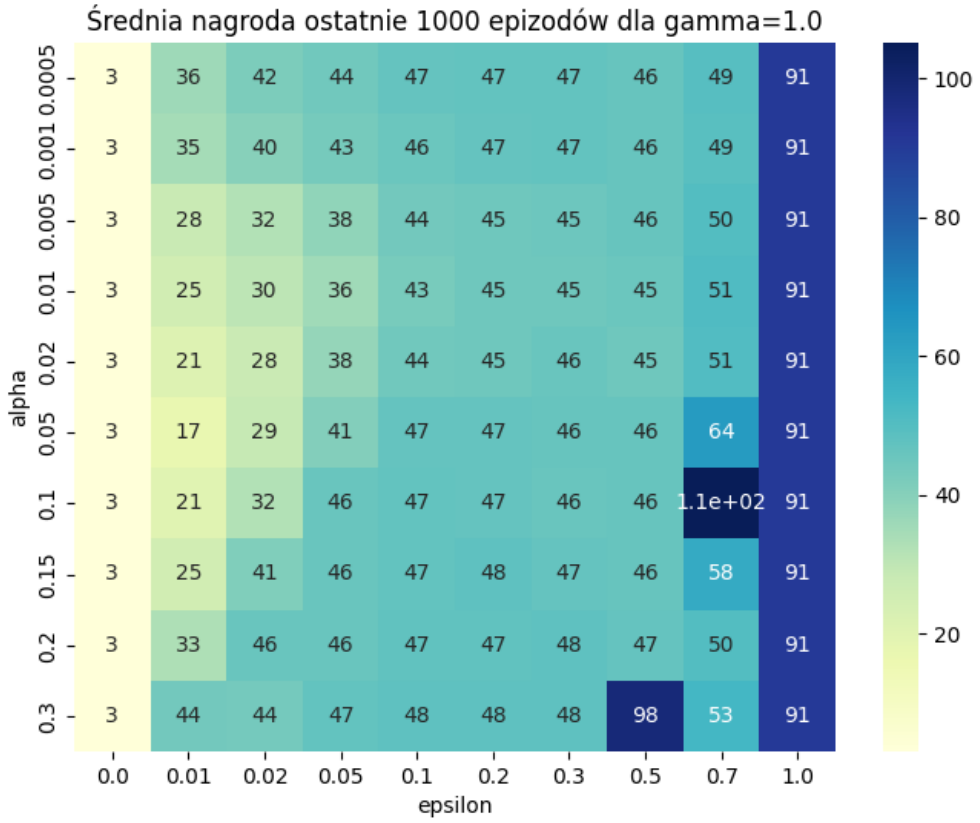

Najlepsza nagroda w każdym epizodzie
Nagroda w epizodzie
Epizod
Najlepsza nagroda w epizodzie
0
20
40
60
80
100
0
2000
4000
6000
8000
10000


Średnia nagroda ostatnie 1000 epizodów dla gamma=1.0
alpha
epsilon
0.0005 3 6 6 6 5.8 5.9 5.9 5.7 5.7 6.6
0.001 3 6 6 6 5.8 5.9 5.9 5.6 5.7 6.6
0.005 3 5 6 5.9 5.8 5.9 5.8 5.6 5.9 6.6
0.01 3 5 6 5.9 5.8 5.9 5.8 5.6 6.1 6.6
0.02 3 5 6 5.9 5.8 5.7 5.7 6 11 6.6
0.05 3 6 6 5.9 5.8 5.9 5.8 6 9.9 6.6
0.1 3 6 6 5.9 5.9 5.9 5.9 6 9.6 6.6
0.15 3 6 6 5.9 5.9 5.9 5.9 5.8 9.6 6.6
0.2 3 6 6 5.9 6 5.9 5.9 7.7 9.8 6.6
0.3 3 6 6 5.9 6 5.9 5.9 15 9.6 6.6
0.0 0.01 0.02 0.05 0.1 0.2 0.3 0.5 0.7 1.0
4
6
8
10
12
14

W przypadku horyzontu = 10, widzimy większy przekrój wartości nagród osiąganych przez agenta. Teraz tendencja związana z parametrem jest odwrotna, agent stawiający na szybką nagrodę osiąga niskie wyniki. Za to najlepsze wyniki są dla parametru e = 0.5 i alpha = 0.3. Co wskazuje na całkiem wysoką tendencję do eksploracji. Jednak mimo większego horyzontu, średnia jest niska, bo osiągnięcie stanu 10 związanego z nagrodą wymaga 7 akcji kupna. Jednak widzimy, że pojedynczy agenci osiągają rezultaty > 100, jednak jest to zbyt mało by wpłynąć znacząco na średnią.

H = 100





W przypadku dużego horyzontu np. 100. Widzimy znaczny wzrost średnich wartości nagród osiąganych przez agentów. Większość agentów osiąga wartości > 100, z czego największe wartości to około 140-150. Widzimy, też że średnio najbardziej optymalnym parametrem, jest epsilon = 1.0, pozwalający osiągać największe nagrody. Obserwujemy też znikomy wpływ parametru alpha.

Wnioski

Horyzont ma duży wpływ na to jakie parametry pozwalają agentom osiągać najlepsze wyniki. Mały horyzont będzie premiował agentów z małym parametrem epsilon, przez co agenci będą stawiać na krótkoterminowe zyski. Z drugiej strony, duży horyzont premiuje agentów z dużym epsilonem, którzy stawiają na eksplorację, dzięki czemu wiedzą o nagrodzie za s = 10.

Przez cały eksperyment widzimy, znikomy wpływ parametru alpha, mówiącym o learing rate'cie algorytmu. Dla badanych wartości osiągane wyniki są zbliżone, z niewielkimi odchyleniami.

Możemy zatem stwierdzić, że w przypadku tego problemu, w celu znalezienia optymalnej strategii dla agenta zarządzającego stanem sklepu, warto dla mniejszych wartości horyzontu ustawić mniejszy epsilon, lecz w celu osiągnięcia większego zadowolenia, parametr epsilon powinien mieć wartość bliską 1.0.