WSI 23Z LAB6 - Uczenie ze wzmocnieniem

Maksym Bieńkowski

Szczegóły implementacyjne

W tym zadaniu tych interesujących szczegółów nie ma za wiele. Zaimplementowany został algorytm Q-Learning w wariancie epsilon-zachłannym, gdzie epsilon ulega zmniejszeniu przez pierwsze n epok treningowych, a następnie pozostaje na wartości minimalnej, skupiając algorytm na eksploatacji. Model może obliczać nagrodę na podstawie różnych modeli poprzez opakowanie nagrody pochodzącej ze środowiska w funkcję `calculate_reward`. Zależnie od obranego systemu nagród, agent różnie radził sobie z zadanym problemem.

Eksperymenty i wyniki

Porównianie różnych systemów nagród

Systemy nagród porównywane są następująco:

- Agent uruchamiany jest z parametrami które okazały się sensowne we wcześniejszych ekspertymentach
- Po wytrenowaniu na 100 000 epok, agenci są testowani w 1000 epok w swoim środowisku
- Porównujemy współczynnik sukcesu epizodu między różnymi systemami nagród

Systemy nagród testowane są na następującej kombinacji parametrów:

Parametr	**Wartość**
Learning rate	0.01
Liczba epizodów	100000
Współczynnik dyskonta	0.9
Minimalny epsilon	0.03

Parametr	Wartość
Po jakiej części treningu epsilon się minimalizuje	0.5

Dla klasycznego systemu nagród (1 za prezent, 0 w przeciwnym przypadku) współczynnik sukcesu wynosił około 0.5. Rozważmy systemy z karą:

Nagroda za prezent	Kara za dziurę	Współczynnik sukcesu
1	-1	0.05
1	-0.5	0.35
1	-0.2	0.47
1	-0.15	0.5
1	-0.1	0.55

Z powyższych eksperymentów wynika, że mocne karanie bota za dziurę nie działa idealnie przy włączonym poślizgu o tak wysokim prawdopodobieństwie (2/3). Moim zdaniem wynika to z niedeterministycznej natury środowiska. Zastosowanie małej kary natomiast wydaje się minimalnie poprawiać wyniki.

Ile epizodów wystarcza?

Rozpatrzmy różne liczby epizodów. Przy jak małej ich liczbie model nie zdąży się nauczyć? (Rozpatrywane dla parametrów, na których sprawdzane też były systemy nagród oraz kary -0.2)

Liczba epizodów	Wsp. sukcesów
500 000	0.52
100 000	0.47
50 000	0.46
25 000	0.39
10 000	0.15
5 000	0.12
1 000	0.02

Jak widać, spadek pomiędzy 500 000 a 50 000 epizodów nie jest znaczny, dopiero poniżej 25 000 epizodów wyniki przestają być akceptowalne.

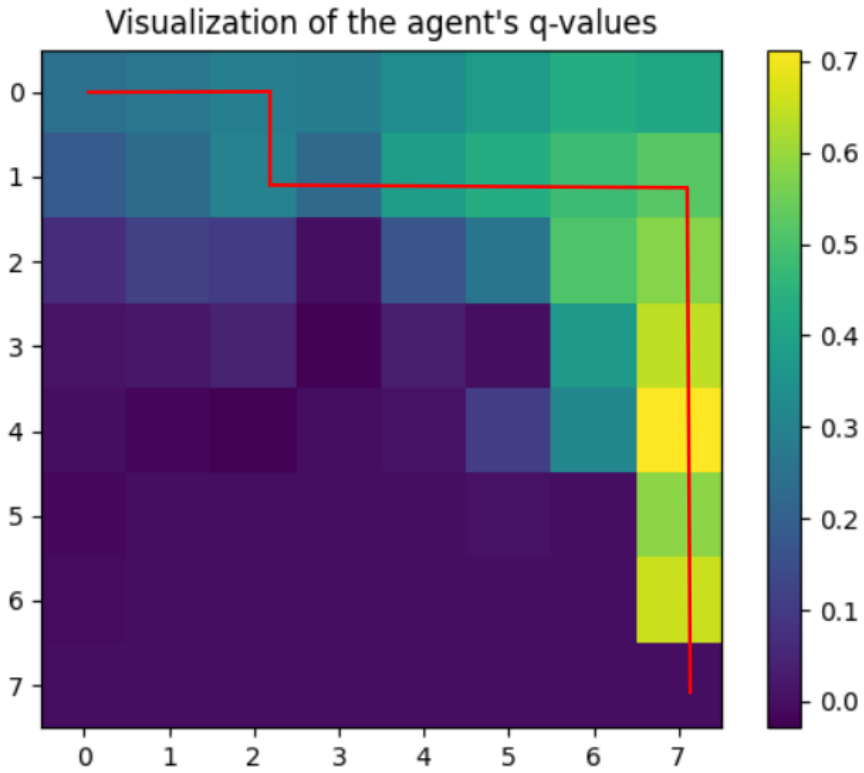
Wizualizacja wartości z tabeli q-values

Przeanalizujmy, jak wygląda średnia wartość dla akcji podjętej z każdego pola dla włączonego i wyłączonego poślizgu oraz różnych systemów nagród.

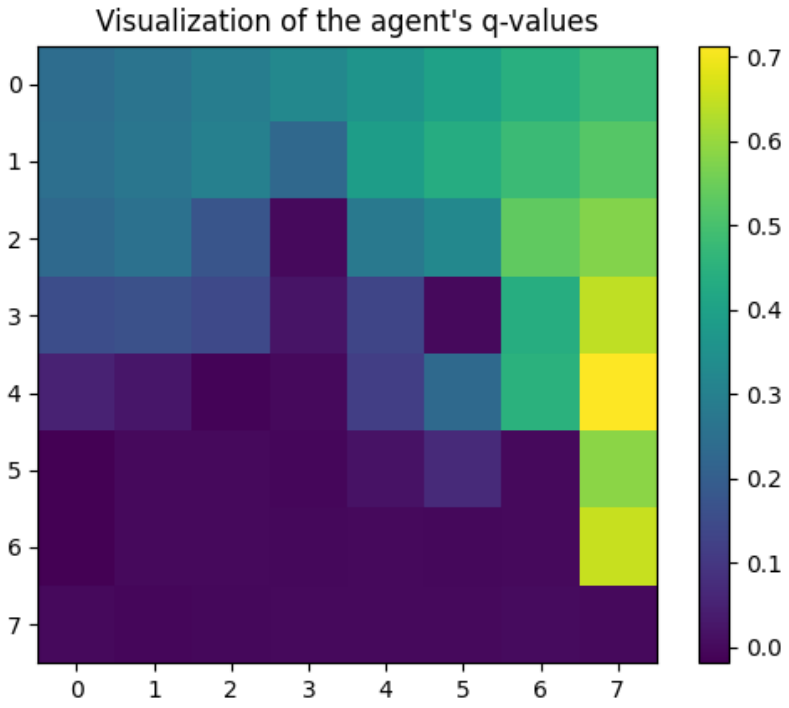
Poślizg wyłączony

Środowisko treningowe i na czerwono zaznaczona ścieżka obierana przez ageta po treningu:

Wartości z tabeli i ta sama ścieżka na tle wizualizacji oceny:

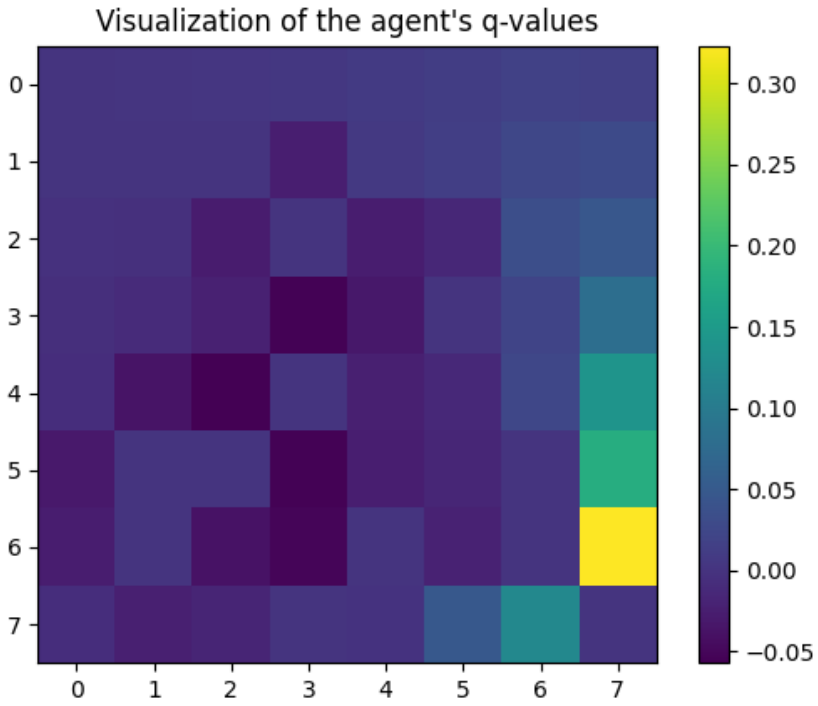


Dla wyłączonego poślizgu doskonale widać, jaką ścieżką podąża agent, możemy też zaobserwować, że dziury, na które natknął się podczas treningu są ocenione najniżej ze wszystkich pól. Warto zauważyć, że wygląda na to, że w powyższej sesji nie wyeksplorował on całej mapy, a skupił się na eksploatacji ,,górnej" ścieżki. Spójrzmy jak na rozkład wartości tabeli wpłynie zmiana parametru `epsilon decay` decydującego o tym, jak szybko kończymy eksplorować z 0.5 na 1 - epsilon zmniejszany będzie na przestrzeni całego treningu, a nie wyłącznie połowy, co wpłynie na zwiększoną eksploatację modelu.



Obserwujemy minimalnie bardziej ,,rozbudowaną" tabelę, jednak widzimy, że prawdopodobnie, po kilkukrotnym sparzeniu się na większej ilości dziur w dolnej części mapy, agent nie wyeksplorował tamtych rejonów znacznie bardziej. Droga obrana przez agenta była identyczna do poprzednio zaznaczonej na czerwono.

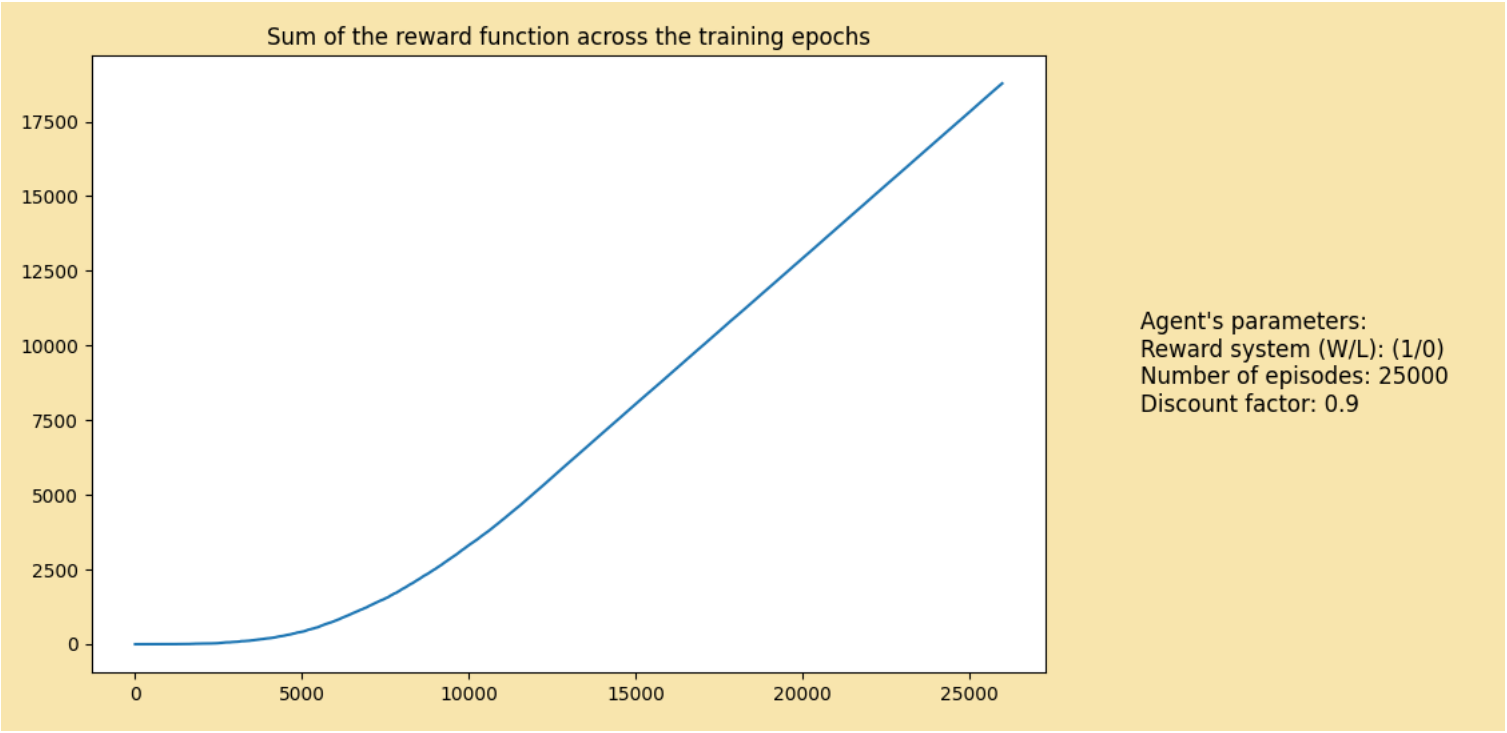
Włączony poślizg



Uruchomienie poślizgu z oczywistych względów znacząco zwiększa możliwości eksploracyjne modelu - tabela jest bardziej różnorodna, obserwujemy zaznaczoną na czerwono ścieżkę preferowaną przez agenta, zauważamy jednak jeszcze jedną istotną rzecz - pułap osiąganych q-values się drastycznie zmienił - z przedziału (0, 0.7) do (-0.05, 0.3). Spowodowane jest to ,,przypadkowymi" zboczeniami z kursu i porażkami spowodowanymi poślizgiem, co w efekcie zmniejsza wartości pól, które wcześniej byłyby uznane za bardziej obiecujące. Fakt, że ostatnie pola prowadzące do prezentu sąsiadują z dziurami sprawia, że wartość tych pól się obniża, pociągając za sobą propagację do tyłu niższych wartości, w związku z czym pole w prawym górnym rogu, uprzednio ocenione na około 0.5 ocenione jest tu tylko na 0. Warto także zauważyć, że losowość przy poślizgu sprawiła, że agent znalazł też ścieżkę do prezentu od ,,lewej", jednak przy wizualizacji testu starał się on w miarę możliwości wracać do górnej części mapy, gdyż przejście dołem obarczone jest dużym ryzykiem.

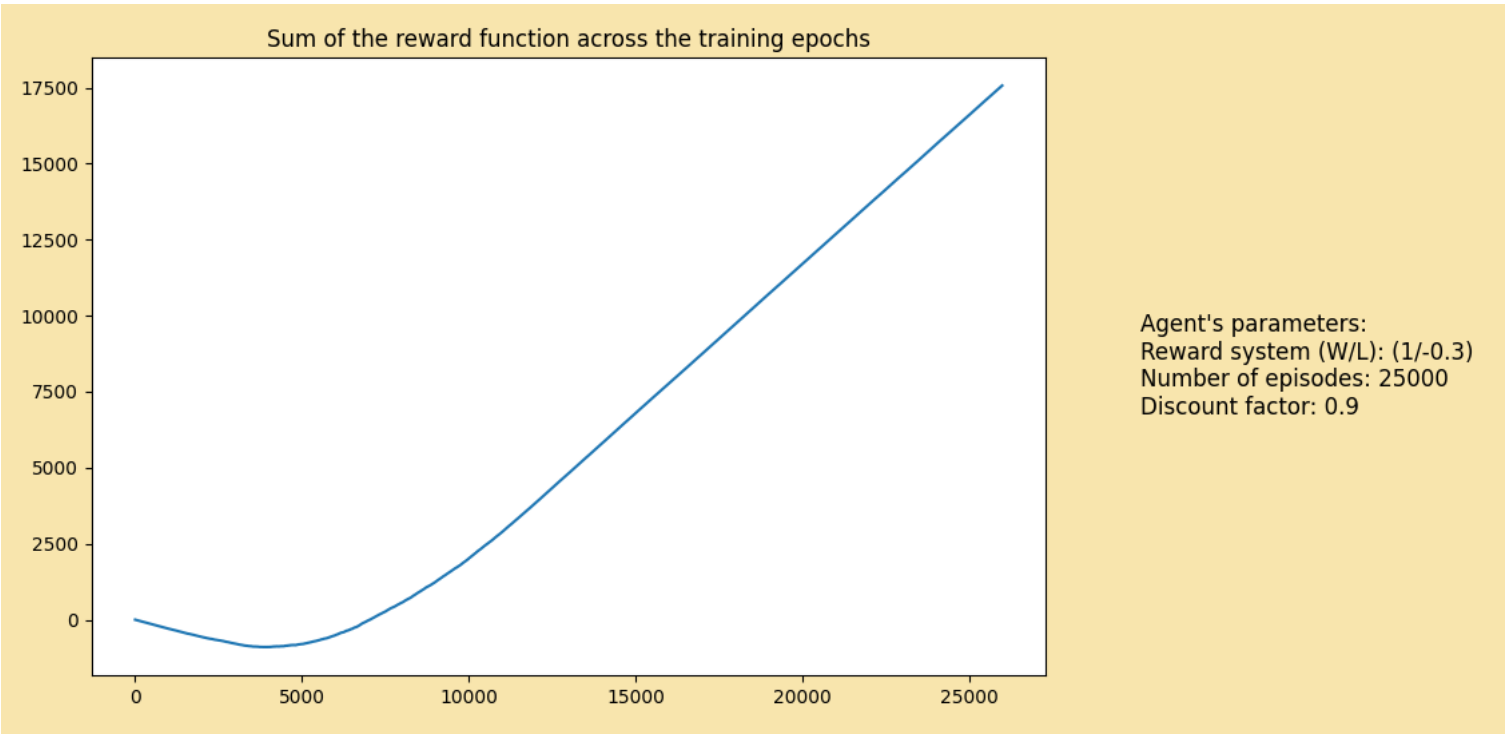
Zmiana sumy nagród na przestrzeni czasu w zależności od systemu nagród i innych parametrów

Wyłączony poślizg, klasyczny system 1/0



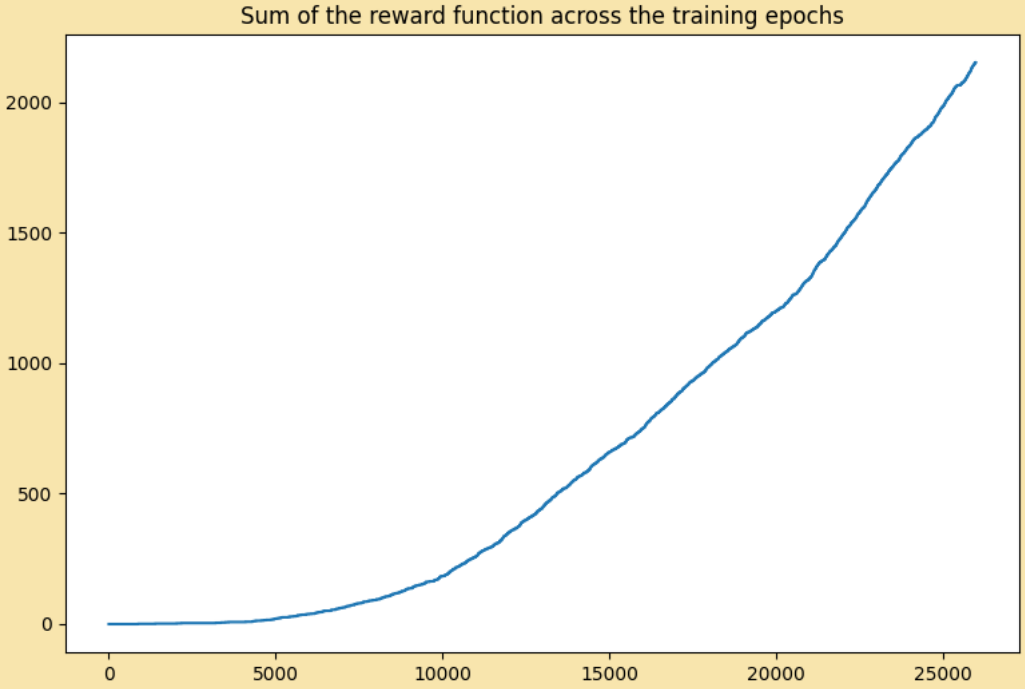
Dla klasycznego systemu nagród, obserwujemy początkową stagnację na poziomie 0 przed pierwszym znalezieniem prezentu i dość szybkie przejście w liniowy wzrost sumy na przestrzeni epok.

Wyłączony poślizg, system 1/-0.3



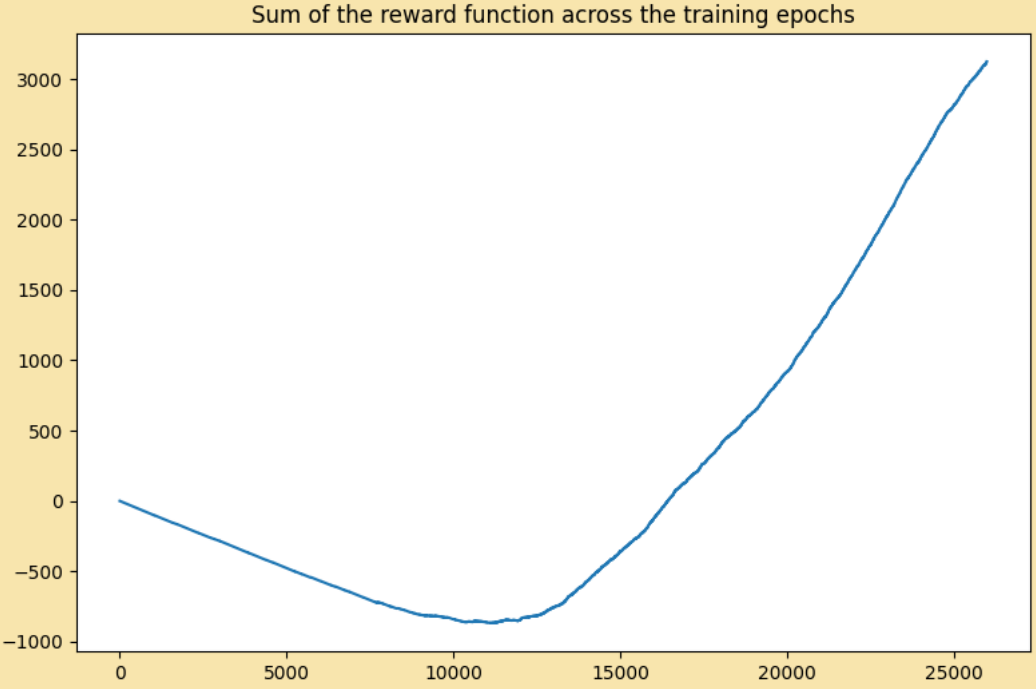
Wykresy mają w ogólności podobny kształt i końcowo rosną liniowo, obserwujemy jednak, że przed znalezieniem drogi, suma konsekwentnie spada poniżej 0.

Włączony poślizg, system 1/0



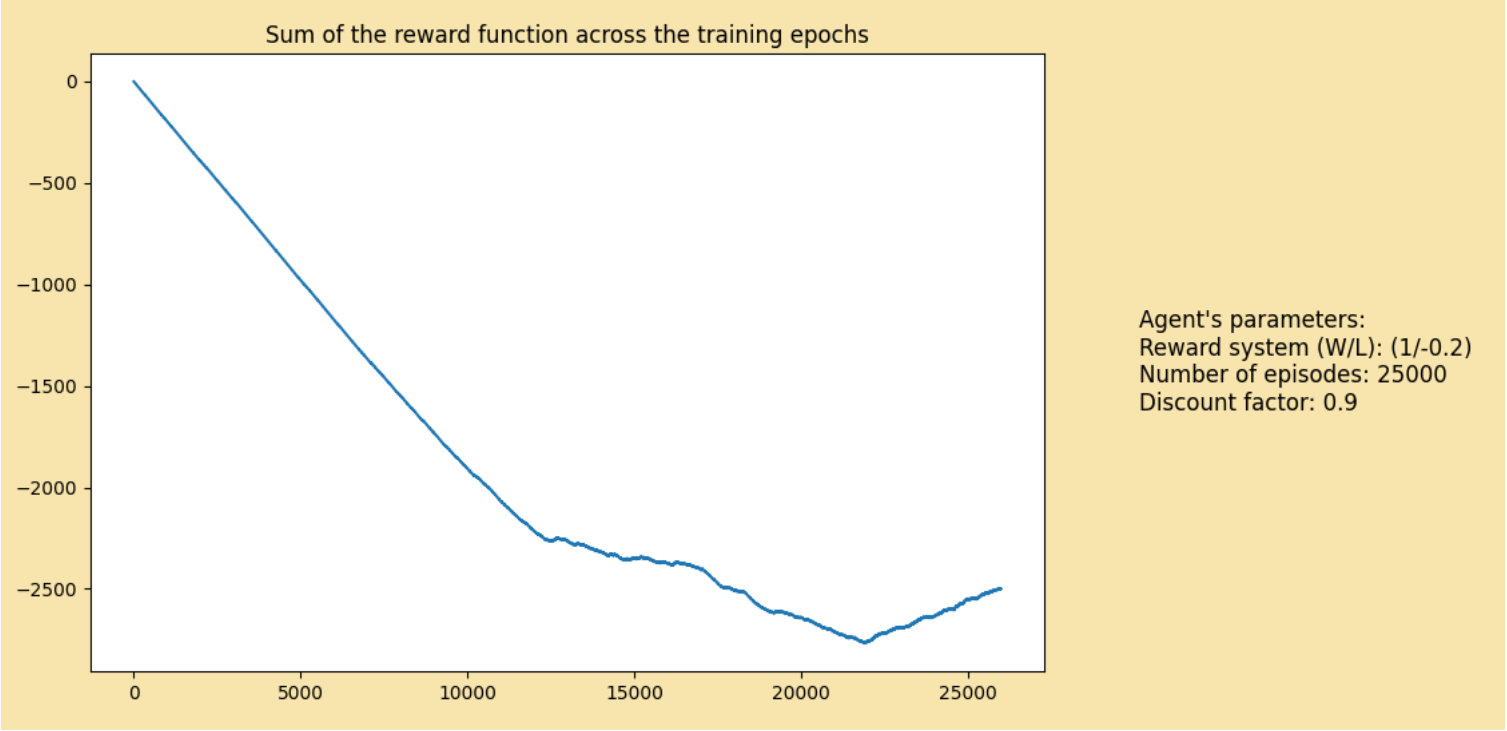
Agent's parameters:
Reward system (W/L): (1/0)
Number of episodes: 25000
Discount factor: 0.9

Wykres przypomina ten z wyłączonym poślizgiem, obserwujemy jednak różnicę - suma nie rośnie ściśle liniowo (zwróćmy uwagę na rząd jej wielkości w porównaniu do tamtego!), nawet po nauczeniu się optymalnej polityki agentowi zdarzają się porażki spowodowane niedeterministycznością środowiska, więc suma rośnie znacznie wolniej. Warto jednak zauważyć, że dla włączonego poślizgu nie jest to optymalna kara i możemy poprawić tempo wzrostu po ukaraniu agenta:

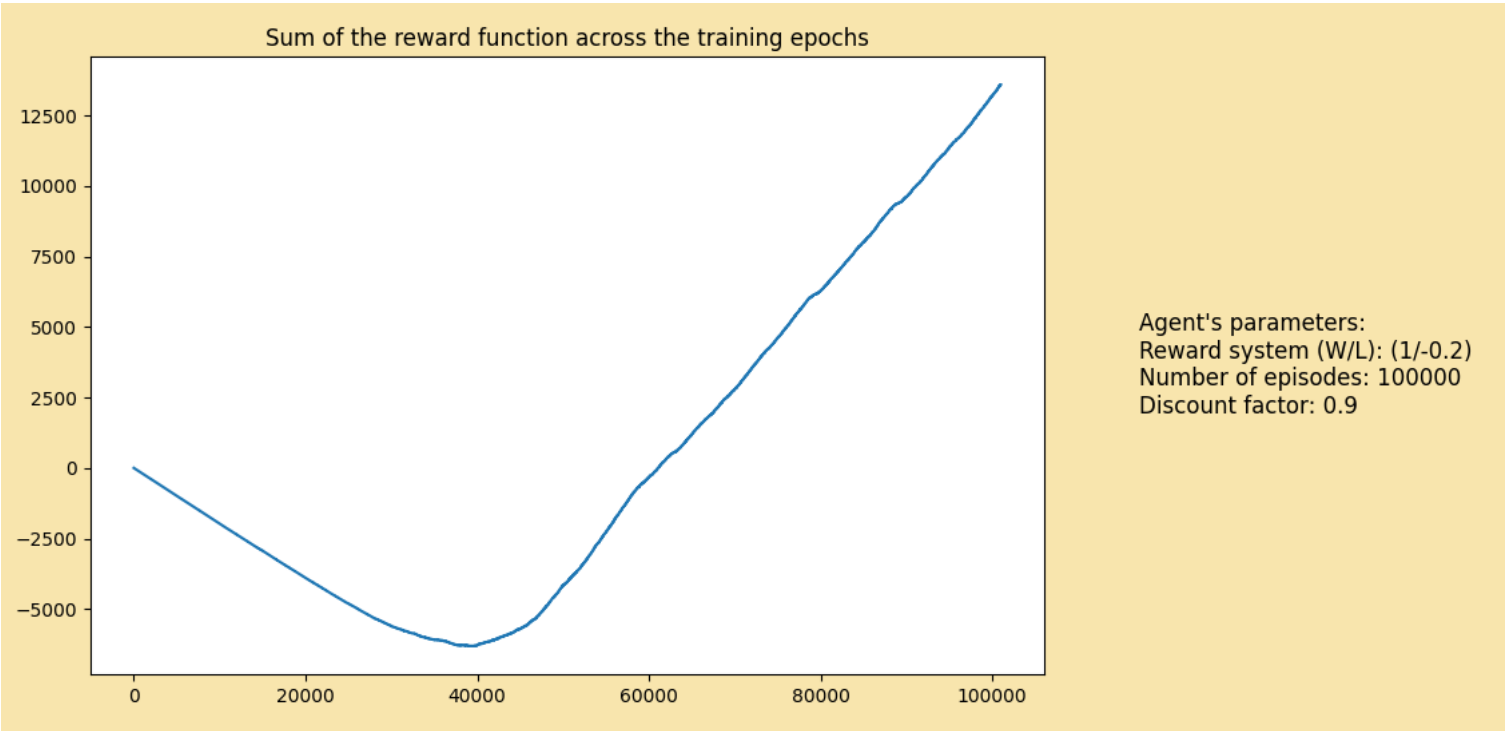


Agent's parameters:
Reward system (W/L): (1/-0.1)
Number of episodes: 25000
Discount factor: 0.9

Obserwujemy dłuższy okres stagnacji, którego wynikiem jest lepsza polityka, w związku z czym końcowo osiągnięta suma jest ostatecznie wyższa, agent osiąga około 50% sukcesów. Spójrzmy, jak wygląda wykres dla jeszcze wyższej kary, która jest już suboptymalna w porównaniu do -0.1, co zostało omówione na początku sprawozdania:



Dwukrotne zwiększenie kary drastycznie zmienia ostateczną sumę nagród - agent kończy z ujemną sumą! Początkowo obserwujemy stały spadek, następnie agent trafia na prezent, wykres zaczyna się wypłaszczać i rosnąć, jednak brakuje mu epok na poprawienie sumy. Ukończył w tym wypadku z sukcesem 20% epok. Gdy damy mu wyższy budżet obserwujemy, że faktycznie jest w stanie poprawić wynik i osiąga 50% sukcesów z takim wykresem:



Przemyślenia

Napisany przeze mnie model agenta można z małym nakładem pracy uogólnić dla innych problemów z biblioteki gymnasium, na pewno sprawdzę jeszcze, jak sobie radzi przy bardziej skomplikowanych środowiskach i co sprawdza się tam.