Wprowadzenie do sztucznej inteligencji | cw. 6

Damian D'Souza

Założenia wstępne

Wartości parametrów początkowych:

- Współczynnik uczenia - 0,3
- Współczynnik dyskontowy - 0,97

Parametry zostały dobrane tak aby algorytm dawał rozsądne wyniki, nie szukałem ich optymalnych wartości.

W początkowych epizodach agent jest bardziej skłonny do eksploracji – wartość współczynnika epsilon jest początkowo ustawiona na 1, a w każdym kolejnym epizodzie maleje wykładniczo, aż osiągnie minimalną wartość 0,01.

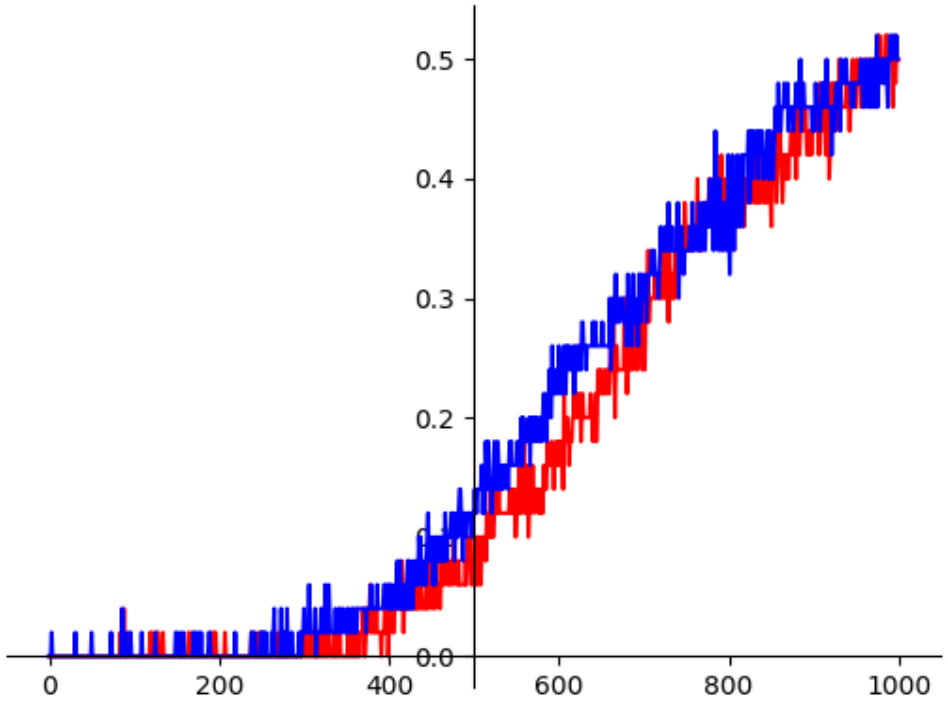
Systemy nagradzania:

W eksperymentach zastosowałem następujące systemy nagradzania:

- Domyślny – oferowany przez bibliotekę, 1 za dojście do celu, 0 w przeciwnym wypadku.
- Karanie za wpadnięcie w dziurę – -1 za wpadnięcie w dziurę, pozostałe nagrody tak jak w domyślnym
- Karanie za wpadnięcie w dziurę i stagnację – -1 za wpadnięcie w dziurę i stanie w miejscu, pozostałe nagrody tak jak w domyślnym

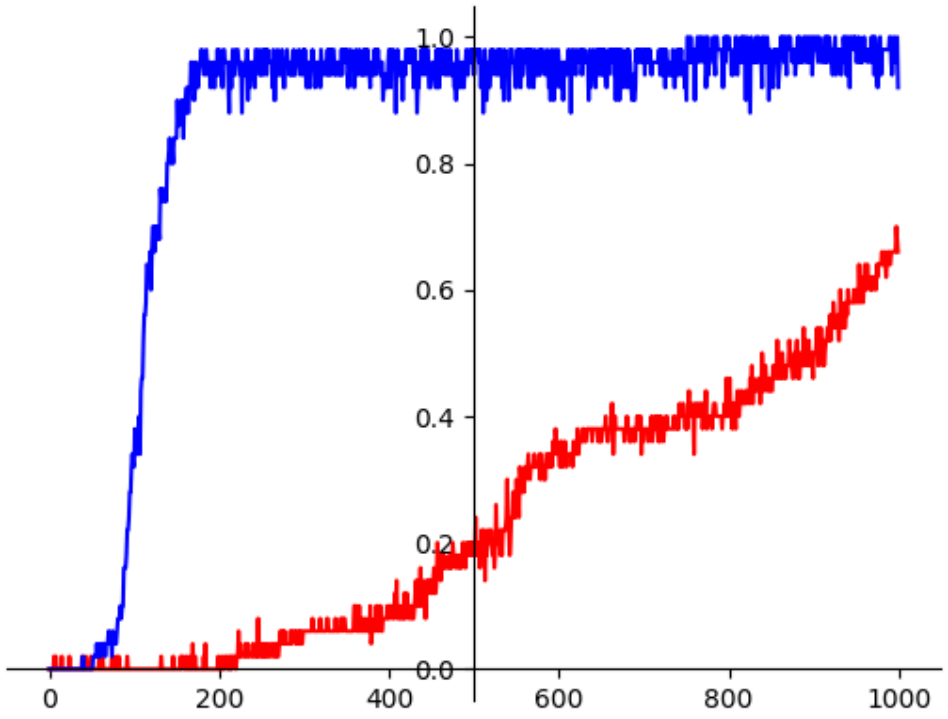
Wyniki

1. Pierwszy eksperyment (system domyślny)



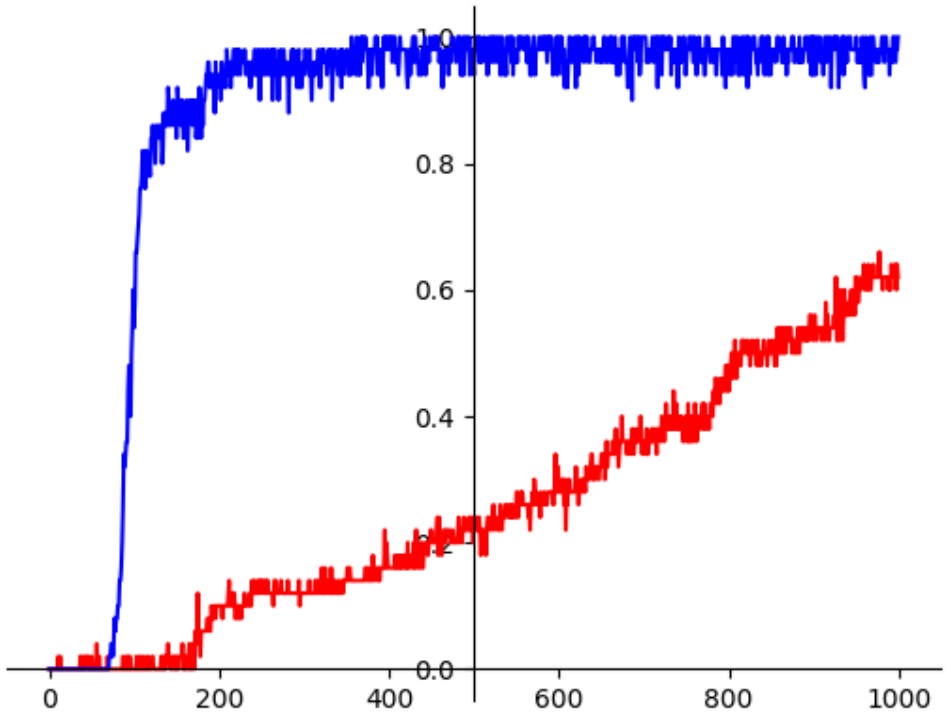
Zaobserwowano niewielkie różnice wynikające z losowości oraz to, że wartość średniej nagrody osiągnęła maksymalny poziom w okolicach 0,5. W kolejnych eksperymentach zmodyfikuję system nagradzania dla agenta o kolorze niebieskim.

2. Eksperyment z karaniem za wpadnięcie w dziurę



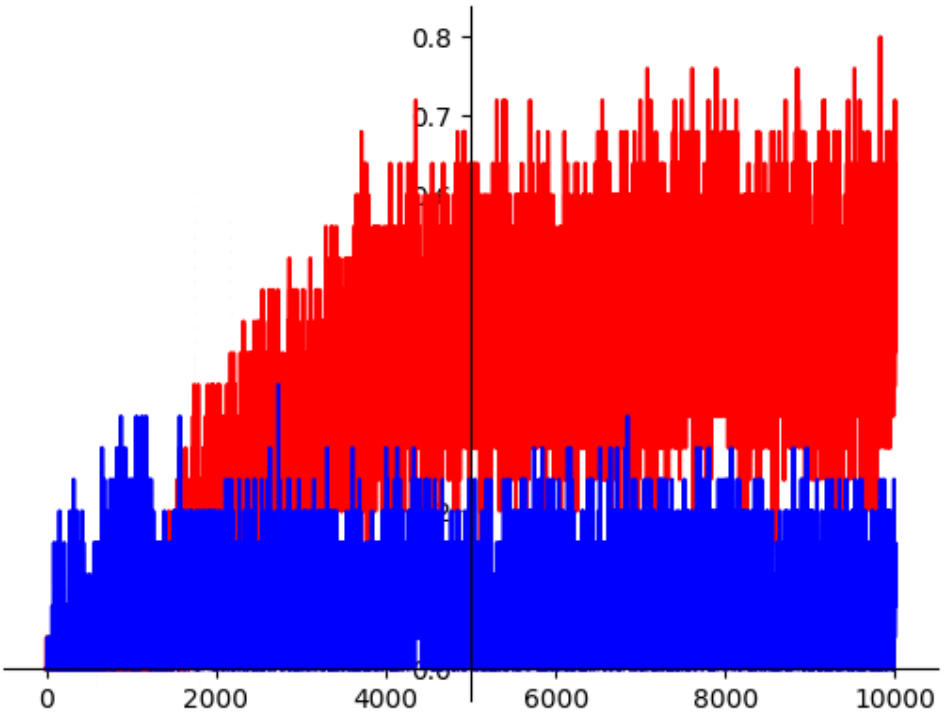
W tym eksperymencie agent był karany wyłącznie za wpadnięcie w dziurę. Średnia wartość nagrody znacząco wzrosła, osiągając około 1 już po 200 epizodach i utrzymując się na tym poziomie. Agent znacznie częściej i szybciej dochodził do celu, co stanowi wyraźną poprawę w porównaniu z systemem domyślnym.

3. Eksperyment z karaniem za wpadnięcie w dziurę i stagnację



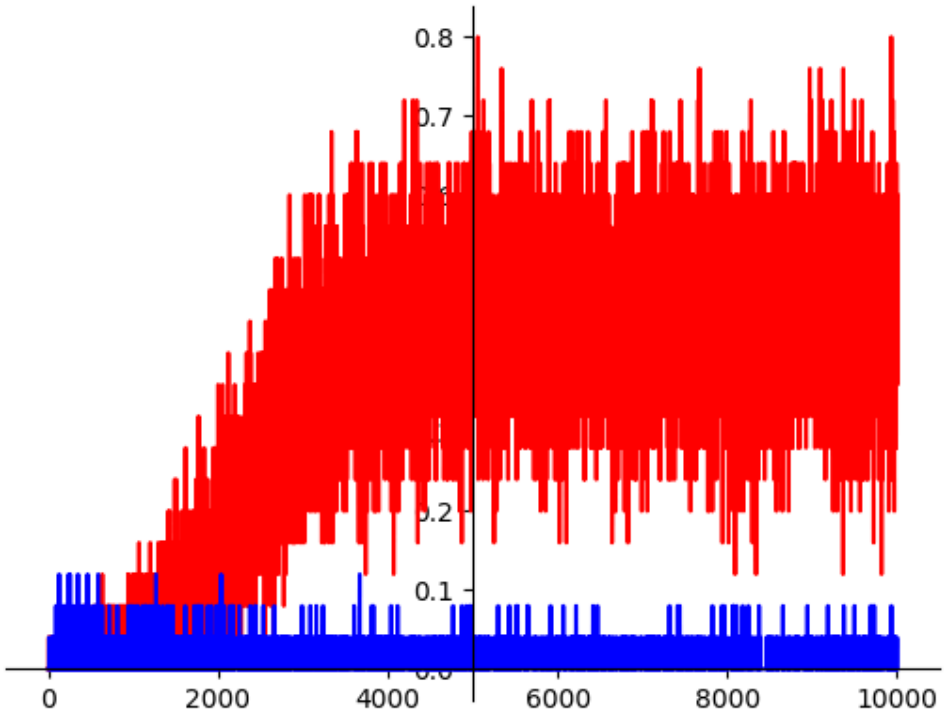
Po wprowadzeniu systemu nagród, w którym agent był dodatkowo karany za stanie w miejscu, zaobserwowano znaczącą poprawę w porównaniu z systemem domyślnym. W porównaniu z poprzednim systemem nagród zmiana nie była już tak wyraźna, ale można zauważyć marginalny wzrost średniej wartości nagród. Co więcej, średnia wartość nagrody osiągnęła poziom 1 szybciej niż w poprzednim eksperymencie. Podobnie jak wcześniej, agent częściej i szybciej dochodził do celu, choć w tym przypadku poprawa była marginalna.

4. Eksperyment z karaniem za wpadnięcie w dziurę w środowisku z poślizgiem



W przypadku środowiska z poślizgiem widać znacznie większy rozrzut średnich wartości nagród. Można również zaobserwować, że agent posługujący się zmienionym systemem nagród osiągnął zauważalnie gorsze wyniki, podczas gdy drugi agent, korzystający z domyślnego systemu nagród, już w okolicy 4 tys. epizodów zaczął osiągać średnią wartość nagród około 0,7. Większy rozrzut oraz gorsze wyniki agenta niebieskiego mogą wynikać z niedeterministycznej natury środowiska, w którym każdy ruch ma jedynie ⅓ szans na wykonanie zgodnie z planem. W takich warunkach trudniej jest nauczyć się odpowiedniej strategii, a agent może naliczać ujemną nagrodę w wyniku nieplanowanego ruchu.

5. Eksperyment z karaniem za wpadnięcie w dziurę i stagnację w środowisku z poślizgiem



Na wykresie widać, że dodanie kary za stagnację dodatkowo pogorszyło wyniki agenta niebieskiego, którego średnia wartość nagród oscyluje wokół zera, co wskazuje na trudności w nauce skutecznej strategii w niedeterministycznym środowisku. Tak jak wcześniej, losowość ruchów prowadzi do częstszego naliczania kar, co utrudnia adaptację. Agent czerwony utrzymuje wyniki na poziomie około 0,7, co potwierdza wyższą skuteczność domyślnego systemu nagradzania.

Wnioski

Na podstawie przeprowadzonych eksperymentów można stwierdzić, że algorytm Q-Learning osiąga różne wyniki w zależności od charakterystyki środowiska i zastosowanego systemu nagradzania.

W środowisku deterministycznym (bez poślizgu) agent wykazywał szybki proces uczenia, stabilizując swoje wyniki w mniej niż 1000 epizodów. Systemy nagród uwzględniające dodatkowe kary za działania niepożądane, takie jak stanie w miejscu czy wpadnięcie w dziurę, przynosiły najlepsze rezultaty. Było to spowodowane tym, że dodatkowe informacje o skutkach działań agenta

pozwalały mu skuteczniej unikać błędów i szybciej opracowywać optymalne strategie.

W środowisku niedeterministycznym (z poślizgiem) wyniki były bardziej rozproszone, a proces uczenia wymagał znacznie większej liczby epizodów. Najlepiej sprawdził się najprostszy, standardowy system nagród, który okazał się bardziej odporny na losowe zakłócenia. Dodatkowe kary za stanie w miejscu lub wpadnięcie w dziurę wprowadzały trudności w procesie uczenia, gdyż agent często był karany za działania, które wynikały z losowości, a nie z jego rzeczywistych decyzji. To prowadziło do błędnych aktualizacji tablicy Q i suboptymalnych strategii.

Ostatecznie, w środowiskach z większym poziomem nieprzewidywalności, prostsze systemy nagradzania wydają się bardziej skuteczne, podczas gdy w środowiskach deterministycznych bardziej złożone podejścia lepiej wspierają proces uczenia.