Raport do zadania 6

Uczenie przez wzmacnianie

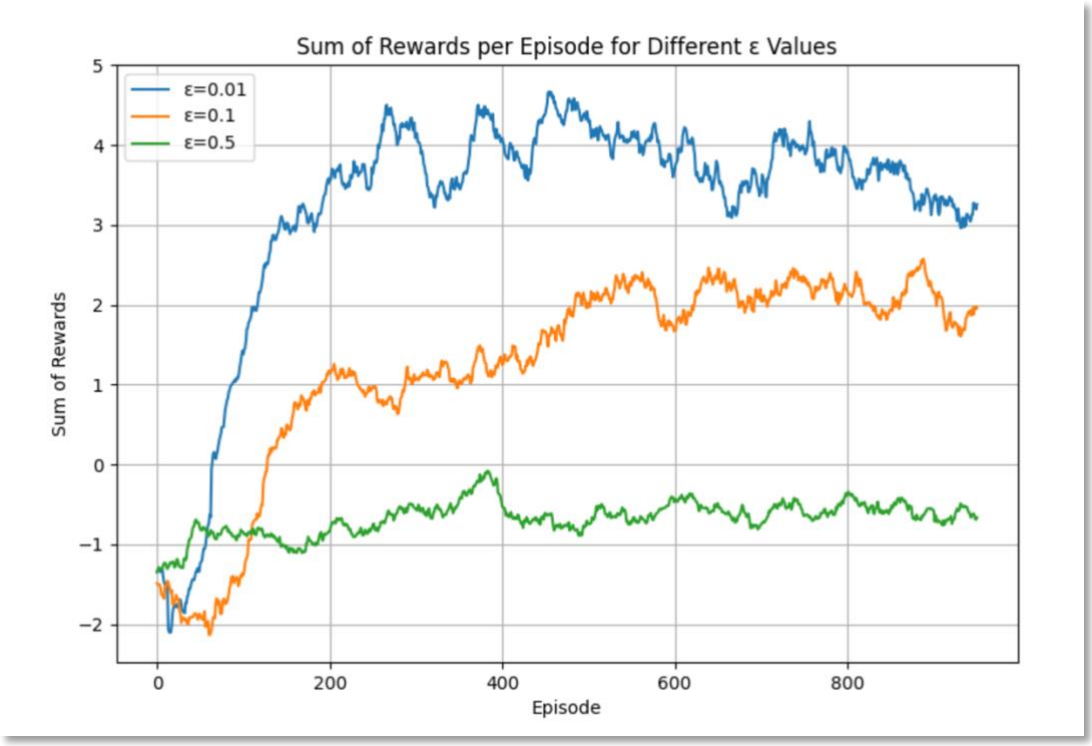
Kacper Siemionek

Numer indeksu: 331430

1. Problem balansu eksploracji i eksploatacji, wpływ wartości ε,

Dobranie odpowiedniej wartości ε bezpośrednio wpływa na balans pomiędzy eksploracją a eksploatacją.

- **Eksploracja** – agent zdobywa nową wiedzę o otoczeniu, stara się odkrywać strategie, testując losowe ruchy, tym samym podejmując pewne ryzyko.
- **Eksploatacja** – agent wykorzystuje poznaną strategię, dzięki której osiągał dobre wyniki, tym samym maksymalizuje nagrody.



Rysunek 1.1 Suma nagród w epizodzie dla różnych wartości ε

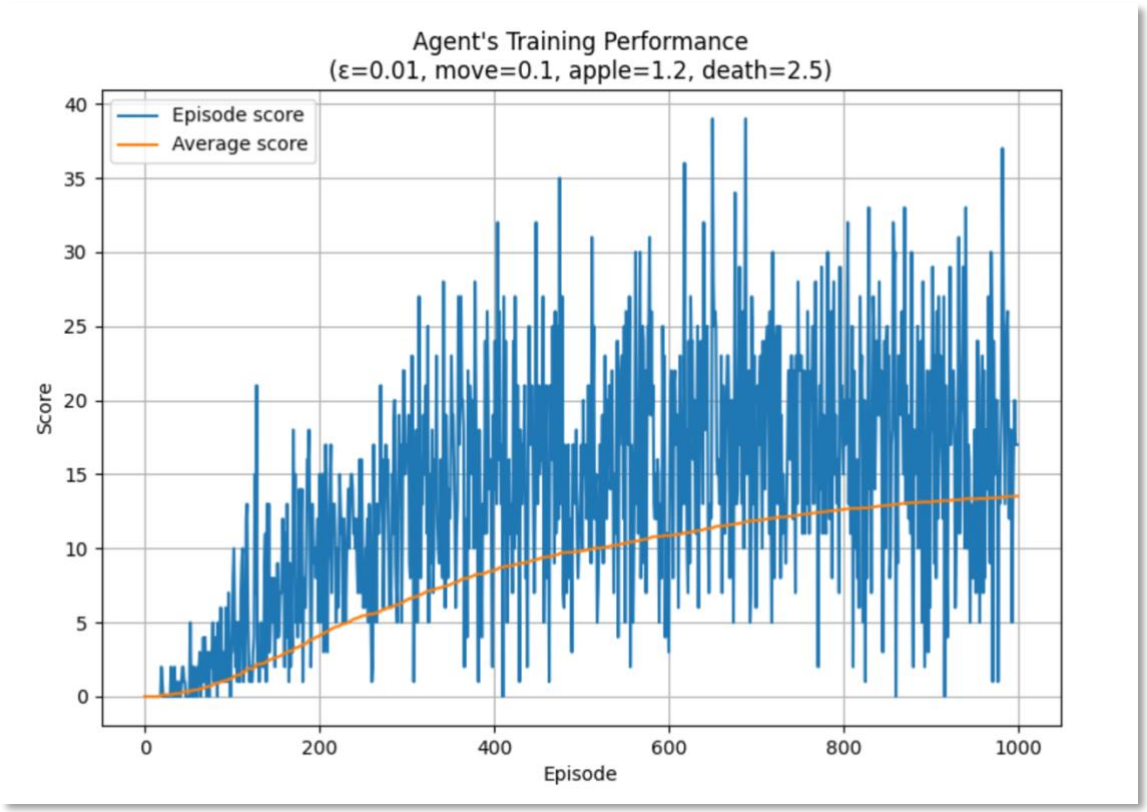
Z powyższego wykresu możemy wywnioskować, że najlepiej poradził sobie model z najniższą wartością ε. Agent w takiej sytuacji stosunkowo szybko przeszedł do eksploatacji poznanej strategii, co poskutkowało w osiąganiu lepszych wyników. Dla $\varepsilon = 0.1$ eksploracja jest bardziej widoczna, agent odkrywa więcej strategii, dlatego jego wyniki rosną wolniej. Wartość $\varepsilon = 0.5$ zaburzyła balans pomiędzy eksploatacją i eksploracją, agent zbyt często wykonywał losowe ruchy, zamiast trzymać się ustalonej strategii.

W tym przypadku mała wartość ε miała pozytywny wpływ na wyniki, jednak szybka eksploatacja nie zawsze jest dobrym rozwiązaniem, agent może utknąć w maksimum lokalnym i nie odkrywać nowych ruchów.

2. Najlepsza Q-funkcja

Agent osiągający najlepsze wyniki został wytrenowany przy użyciu następujących parametrów:

- Epsilon (ε): 0.01,
- Nagroda za znalezienie jabłka: 1.2,
- Kara za ruch: 0.1,
- Kara za śmierć: 2.5,
- Liczba epizodów: 1000.



Rysunek 2.1 Wykres punktów osiąganych w poszczególnych epizodach

Podczas kilku testów składających się ze 100 rozgrywek ($\varepsilon = 0$), średnie wartości osiągane przez model wahały się w okolicach $19 - 20$ punktów. Aby osiągnąć bardziej rzetelny wynik, uruchomiono 1000 rozgrywek, w których agent osiągnął średnio 19.189 punktów.

3. Poprawienie działania agenta

Jak poprawić działanie agenta?

- **Zwiększenie liczby epizodów podczas treningu** – skuteczność agenta może się znacznie poprawić, kiedy będzie miał więcej czasu na poszukiwanie najlepszej strategii i pogłębianie jej.
- **Dynamiczna wartość ε** – początkowo wyższe wartości, agent bardziej skupia się na eksploracji, poszukując strategie, zmniejszanie wartości z każdym postępem, aby mógł eksploatować
- **Zmiana nagród i kar** – zwiększając nagrodę za jabłko, zachęcamy agenta do realizacji celu, zmniejszając karę za ruch, zniechęcamy go do ryzykownych ruchów, a zwiększając karę za śmierć, odciągamy go od częstych pomyłek, jednak złe parametry mogą doprowadzić do zapętlenia ruchów agenta.