Uczenie ze wzmocnieniem

Mateusz Plinta

Wstęp

Wykonano po 5 symulacji dla każdego z 9-ciu setów parametrów definiujących przebieg symulacji. Krótki opis parametrów:

- **alpha** - stosunek pomiędzy dotychczasową wiedzą a wnioskami z nowego doświadczenia
- **gamma** - współczynnik określający stosunek nagrody już otrzymanej oraz przewidywanych nagród w kolejnych krokach
- **exploration_rate** - Szansa na wykonanie losowego wyboru(eksperymentu)
- **exploration_min** - minimalna wartość **exploration_rate**
- **exploration_decay** - współczynnik degradacji **exploration_rate** po każdej iteracji
- **exploration_degrade** - <*boolean*> definiuje czy degradować **exploration_rate** czy nie

Przy nauczaniu w słowniku *knowledge* wykorzystywano 0 *reward* gdy nie był to koniec iteracji, oraz ujemny -1, gdy klocek spadał. Dzięki temu uzyskano system który w miarę możliwości szybko się uczy.

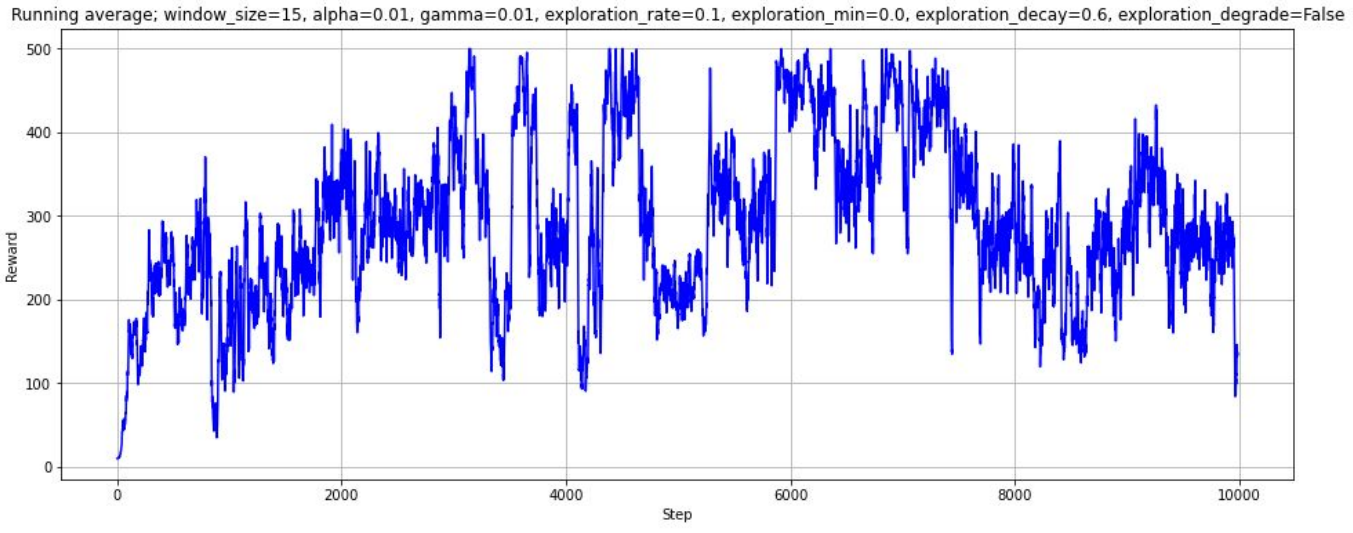
Zad. 1

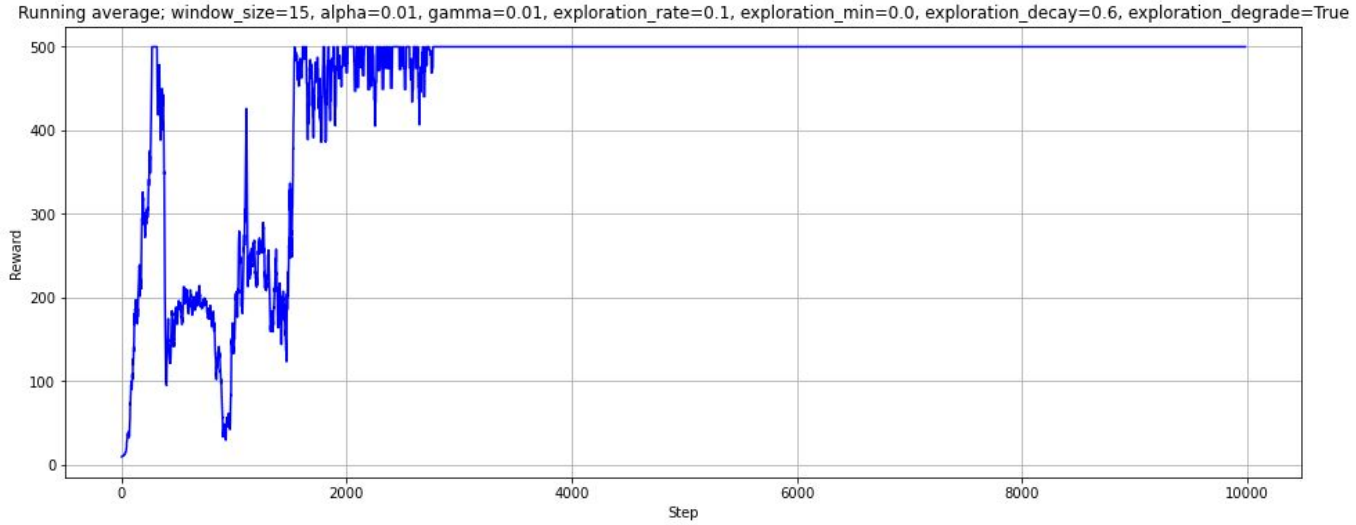
- kod w załączniku

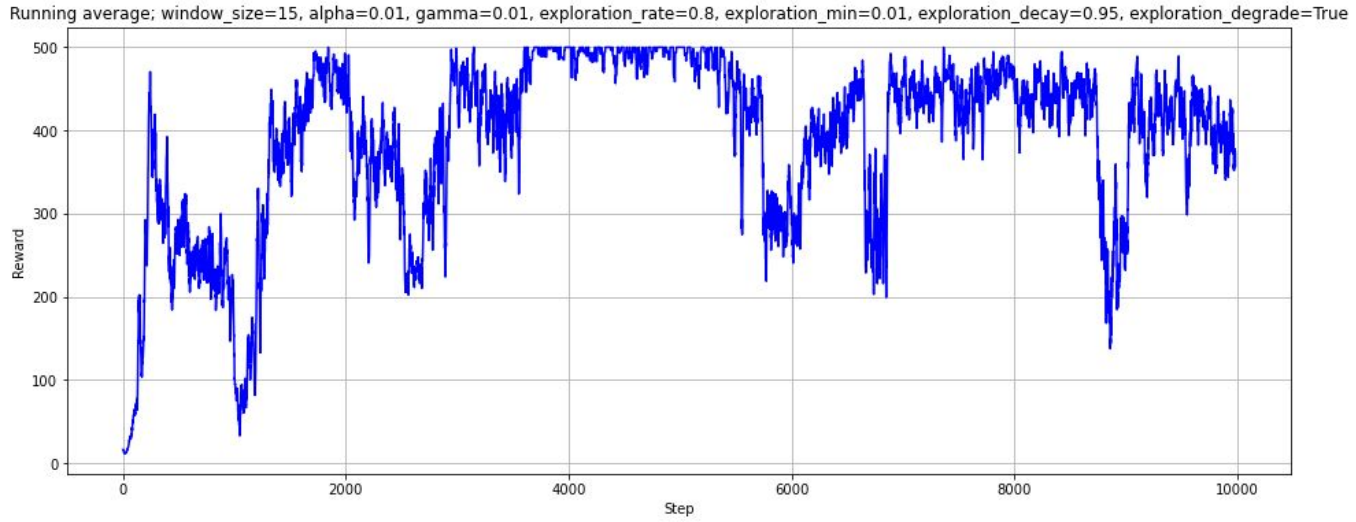
Zad. 2

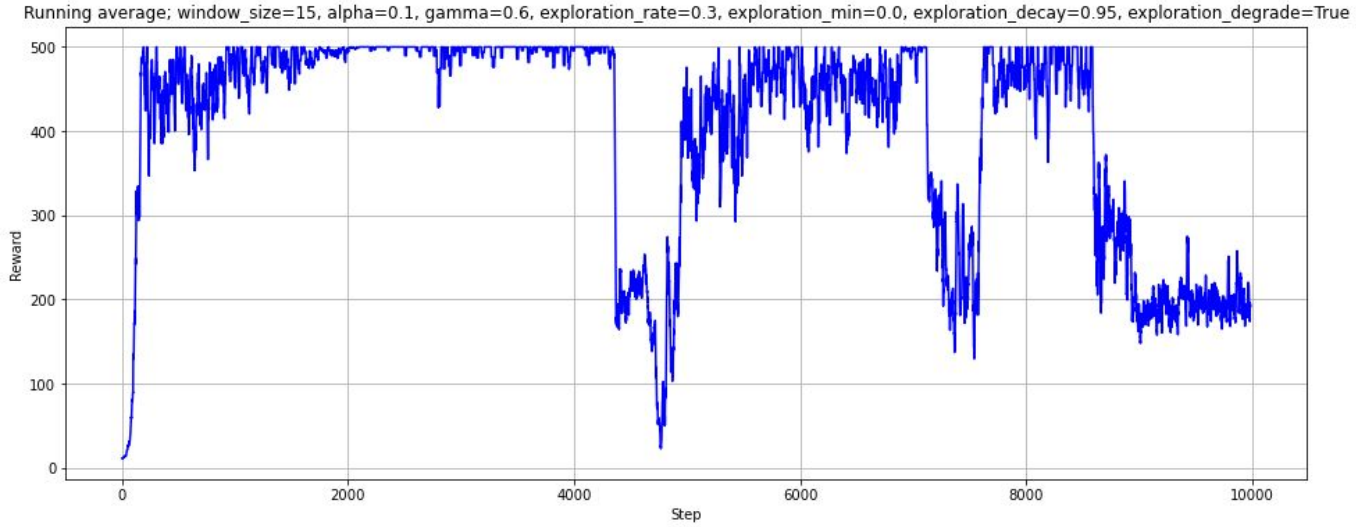
Każda symulacja trwała 10000 epok. W przypadku poniższych wykresów przedstawiających średnią kroczącą, wybrano okno szerokości 15 iteracji do uśredniania wyników. W tytule każdego wykresu znajdują się wszystkie parametry symulacji.

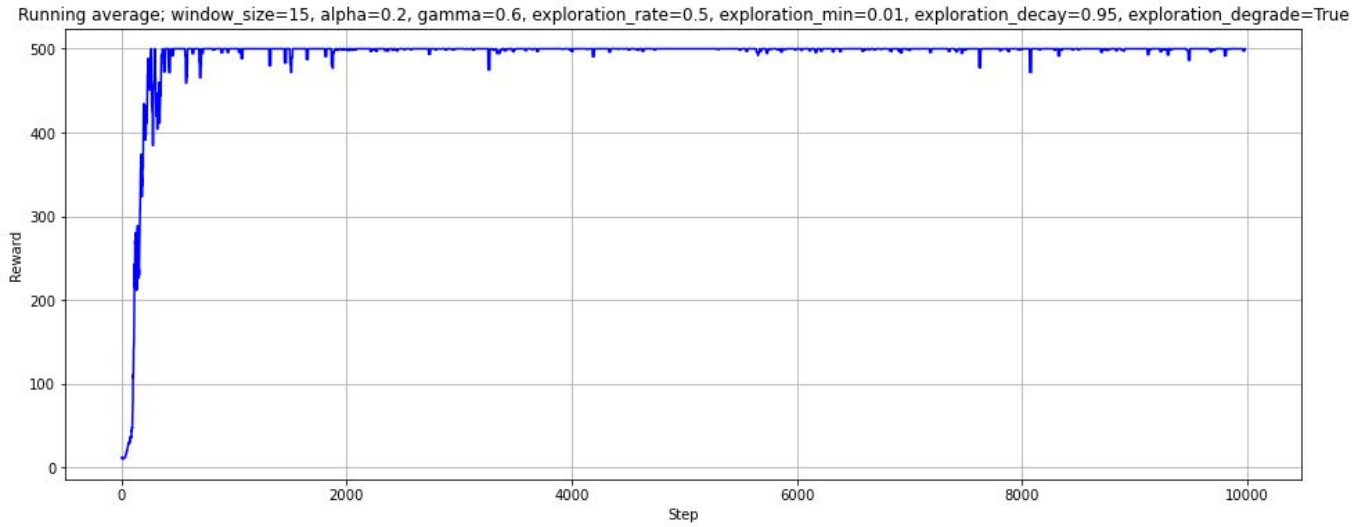
Średnia krocząca

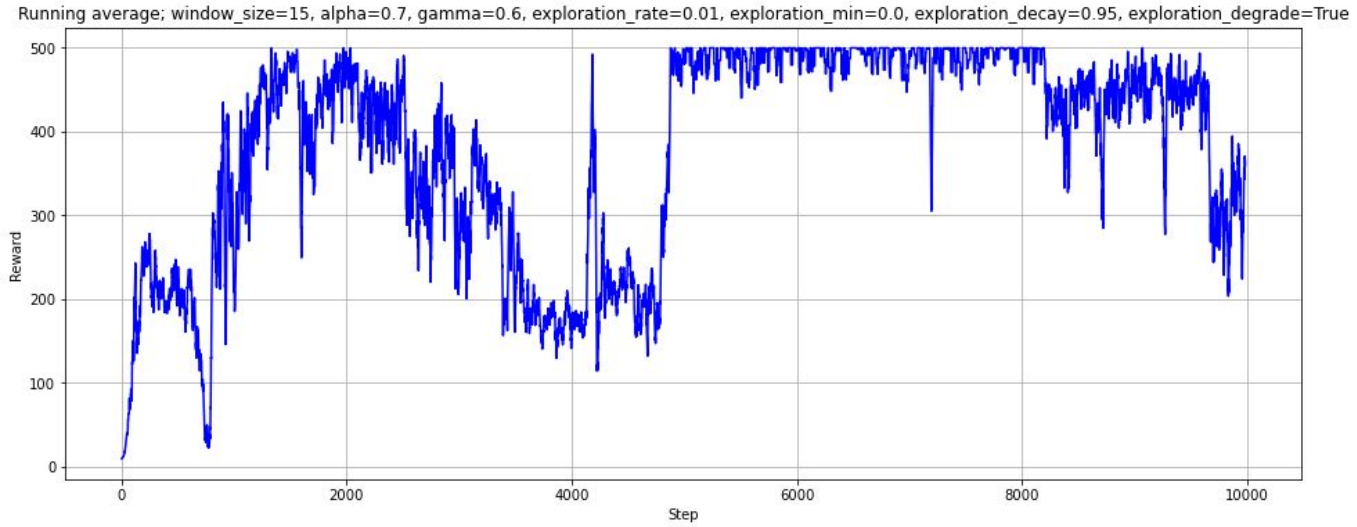


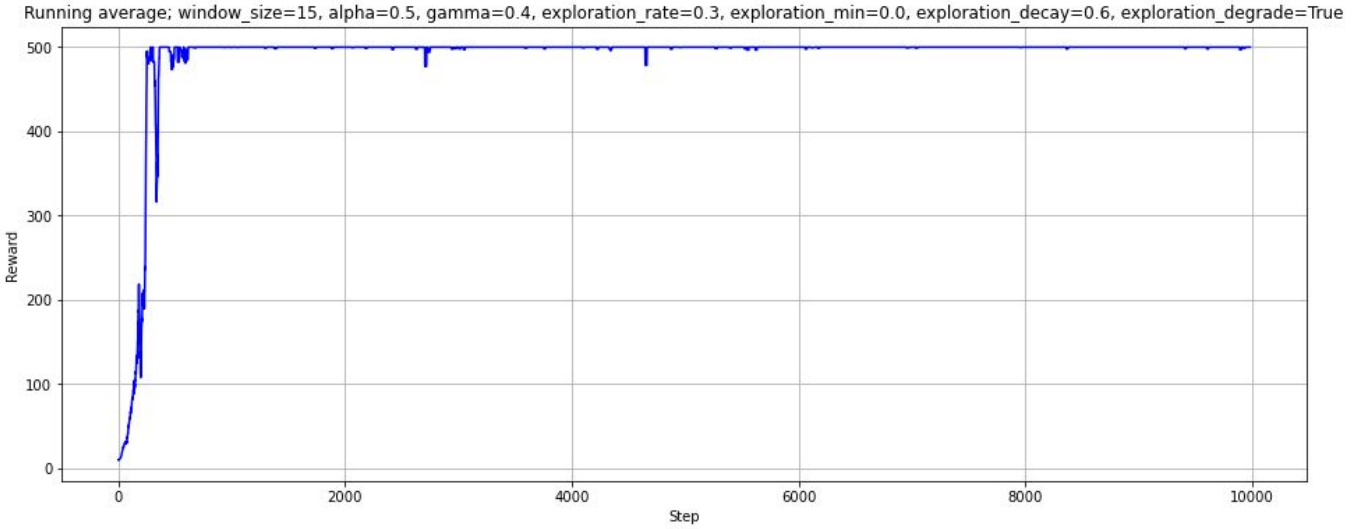

Running average; window_size=15, alpha=0.01, gamma=0.01, exploration_rate=0.1, exploration_min=0.0, exploration_decay=0.6, exploration_degrade=True
Reward
Step
0
100
200
300
400
500
0
2000
4000
6000
8000
10000


Running average; window_size=15, alpha=0.01, gamma=0.01, exploration_rate=0.8, exploration_min=0.01, exploration_decay=0.95, exploration_degrade=True
Reward
Step
0
100
200
300
400
500
0
2000
4000
6000
8000
10000


Running average; window_size=15, alpha=0.1, gamma=0.6, exploration_rate=0.3, exploration_min=0.0, exploration_decay=0.95, exploration_degrade=True
Reward
Step
0
100
200
300
400
500
0
2000
4000
6000
8000
10000

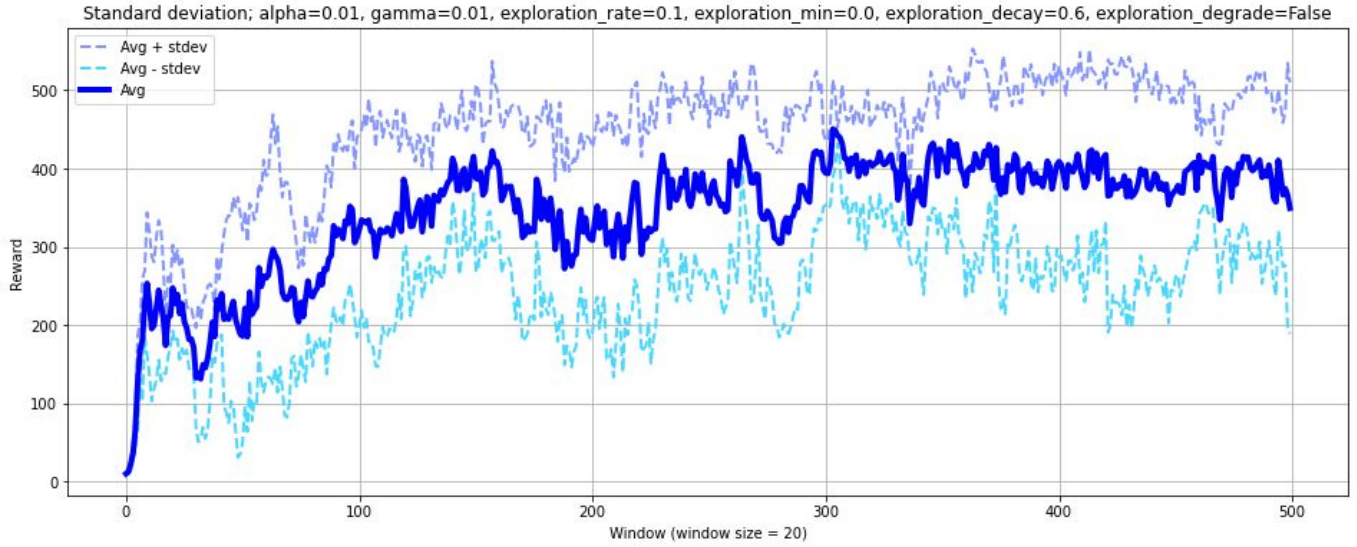


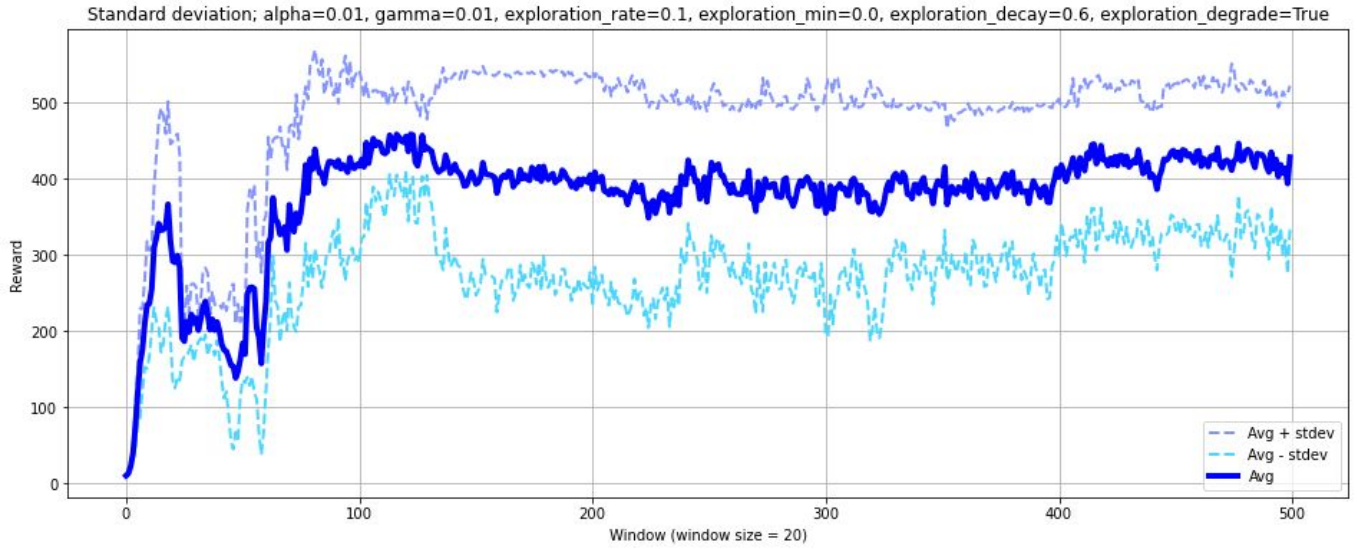


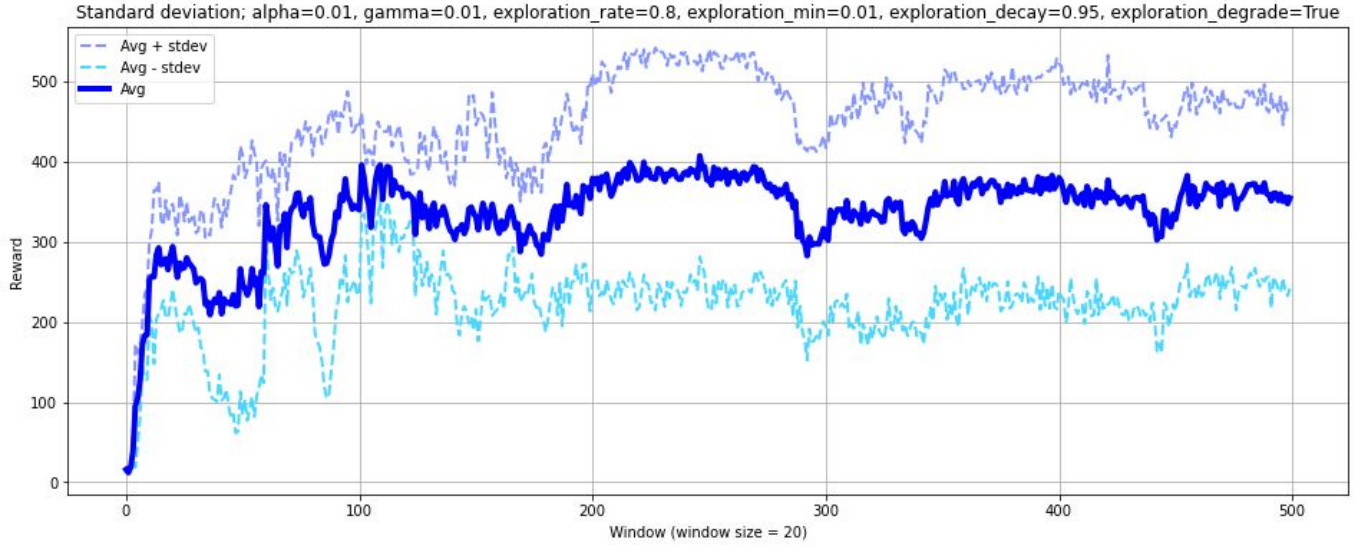


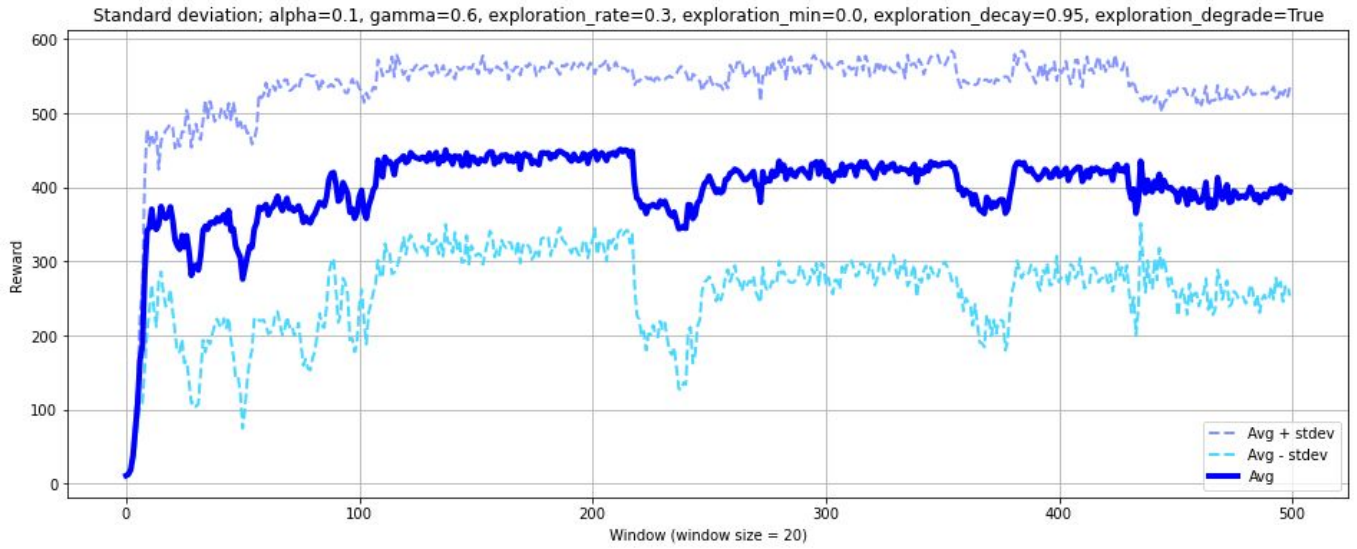
Powyższe wykresy są wyznaczone na podstawie pierwszych ze zbioru pięciu prób symulacji o tych samych parametrach.

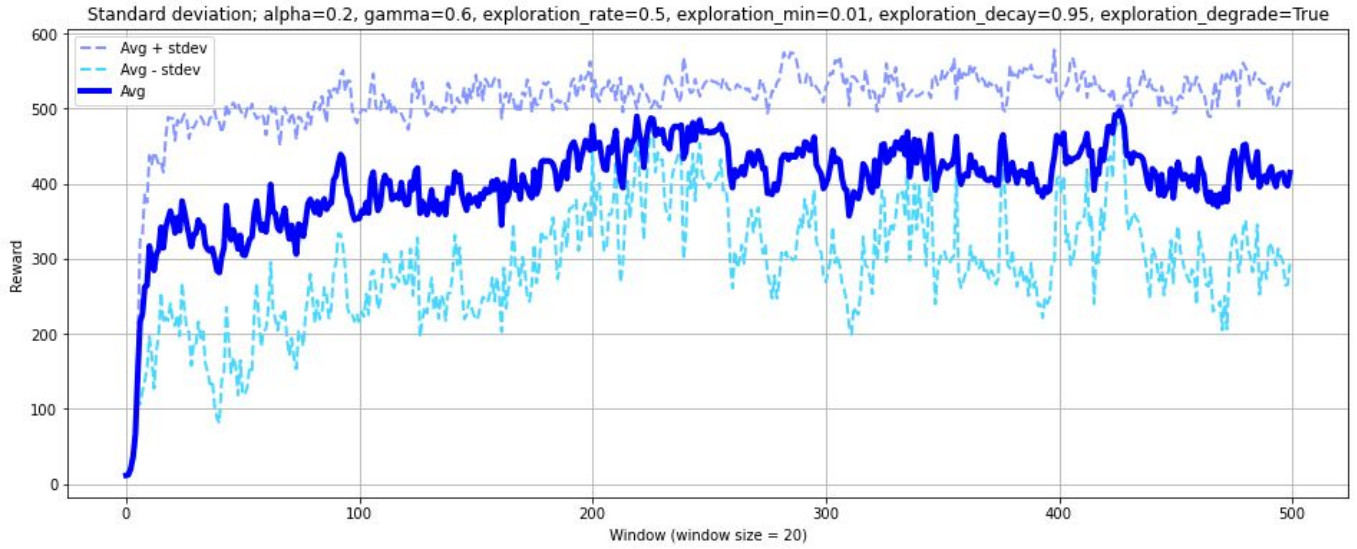
Odchylenie standardowe

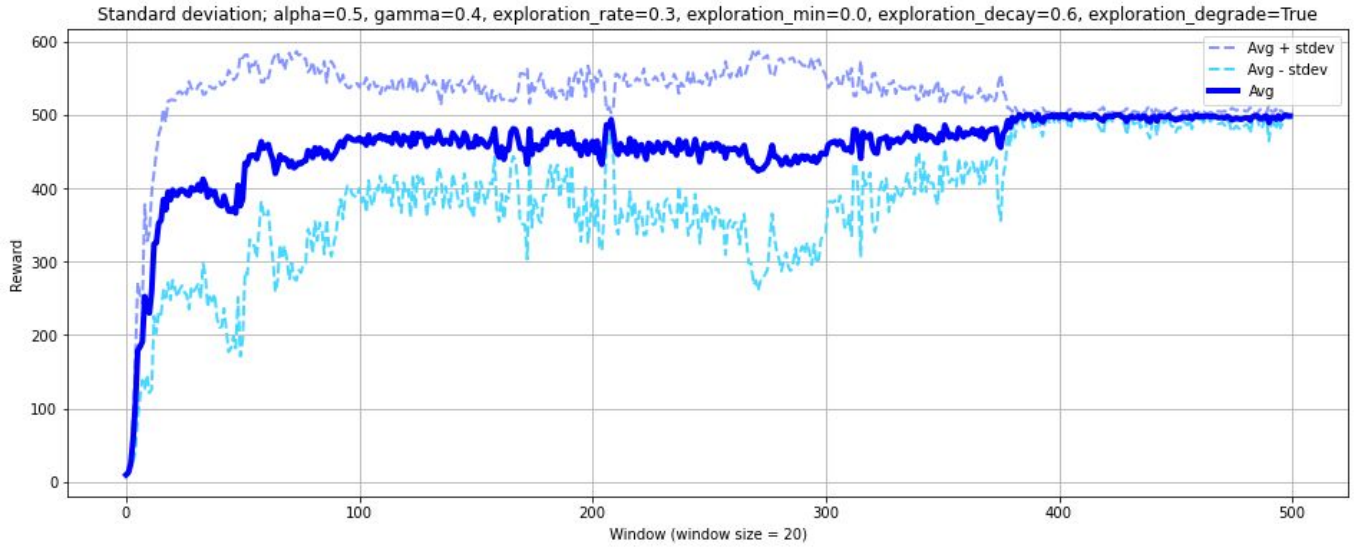


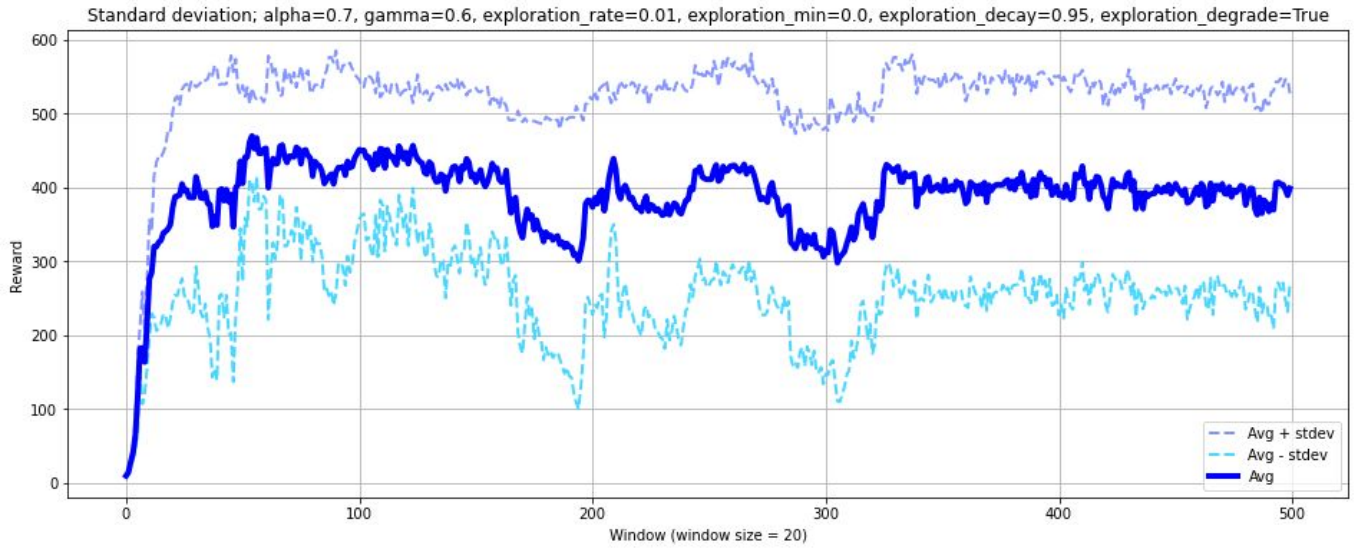


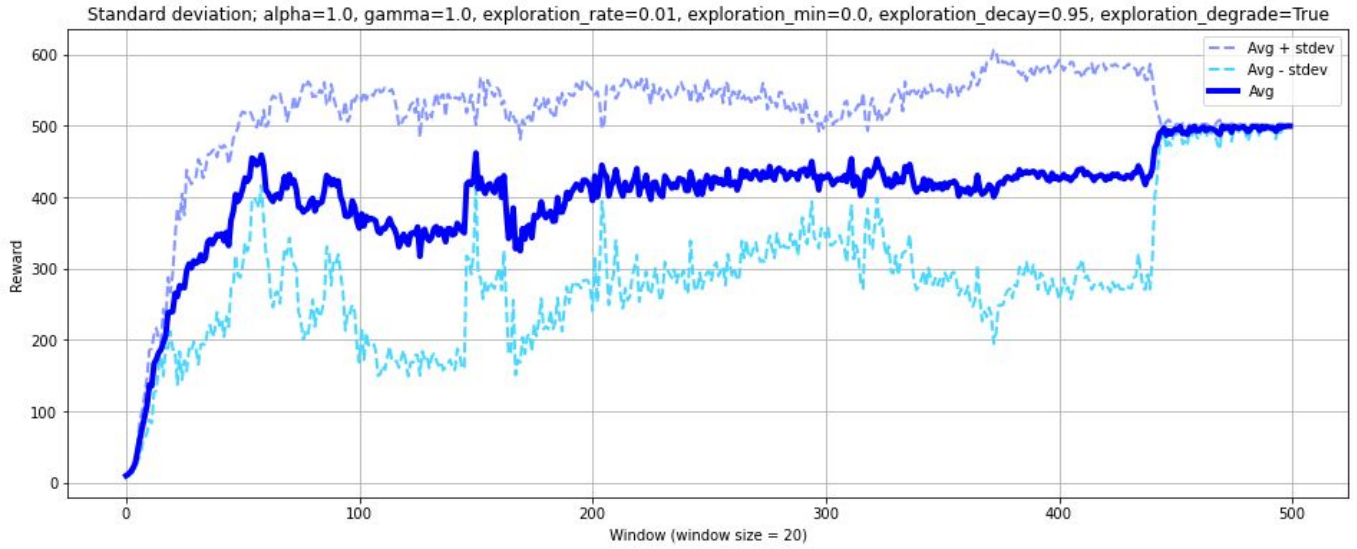


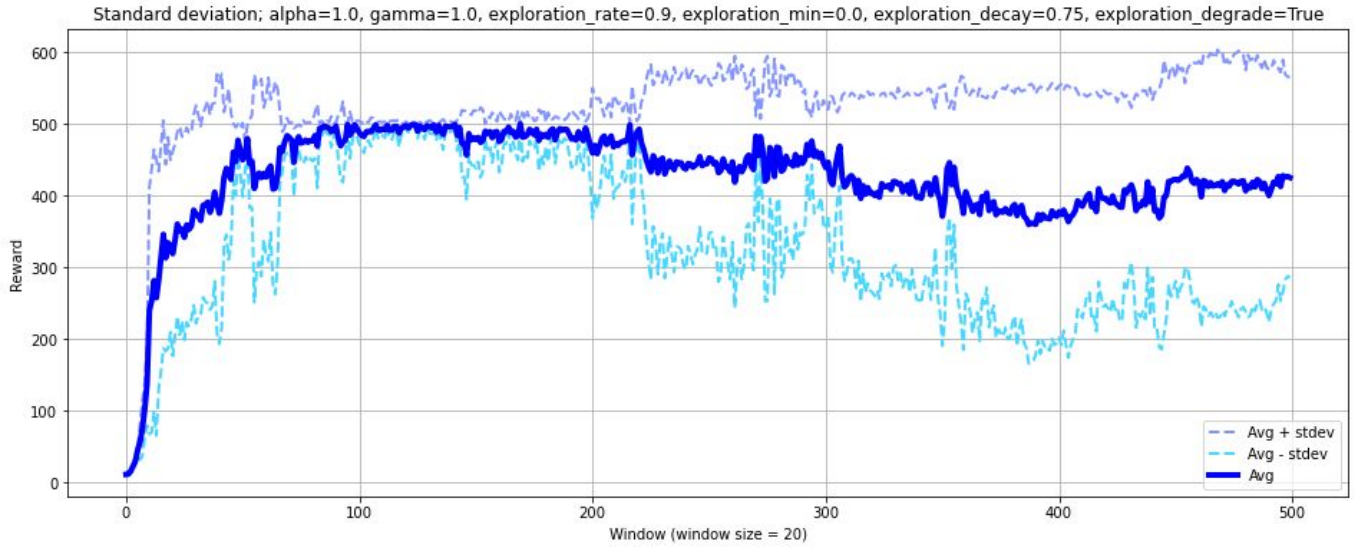

Standard deviation; alpha=0.1, gamma=0.6, exploration_rate=0.3, exploration_min=0.0, exploration_decay=0.95, exploration_degrade=True
Reward
Window (window size = 20)
Avg + stdev
Avg - stdev
Avg


Standard deviation; alpha=0.2, gamma=0.6, exploration_rate=0.5, exploration_min=0.01, exploration_decay=0.95, exploration_degrade=True
Reward
Window (window size = 20)
Avg + stdev
Avg - stdev
Avg


Standard deviation; alpha=0.5, gamma=0.4, exploration_rate=0.3, exploration_min=0.0, exploration_decay=0.6, exploration_degrade=True
Reward
Window (window size = 20)
Avg + stdev
Avg - stdev
Avg







Powyższe wykresy uśredniono zakładając okno wielkości 20 iteracji. Linie przerywane przedstawiają wartości wykresu +/- odchylenie standardowe wartości w danym oknie.

Wnioski

Pierwszy wykres przedstawia symulacje z wyłączoną degradacją **exploration_rate**. Wykres poniżej zakłada te same parametry, lecz z włączonym tym parametrem. Widać wyraźnie zwiększone zaszumienie pomiarów przy wyłączonym **exploration_degrade**, spowodowane nie zmniejszaniem się wartości **exploration_rate**. Mimo tego jednak wygląda na to że z wyłączonym parametrem **exploration_degrade** system uczy się wolniej, lecz stabilniej. Ostatecznie jednak oba wykresy kończą na średnio podobnych wynikach ~400 score.

Najlepsze wyniki uzyskano dla parametrów:

```
alpha=0.5, gamma=0.4, exploration_rate=0.3, exploration_min=0.0,
exploration_decay=0.6, exploration_degrade=True
```

gdzie dane rosną bardzo szybko do ~450 score, w miarę stabilnie. Po ok. 400 oknach system stabilizuje się na 500. Mogły na to mieć wpływ stosunkowo duże wartości parametrów **alpha** i **gamma oraz exploration_rate**, który stosunkowo duży na początku pozwala na eksplorację potencjalnie lepszych możliwości.

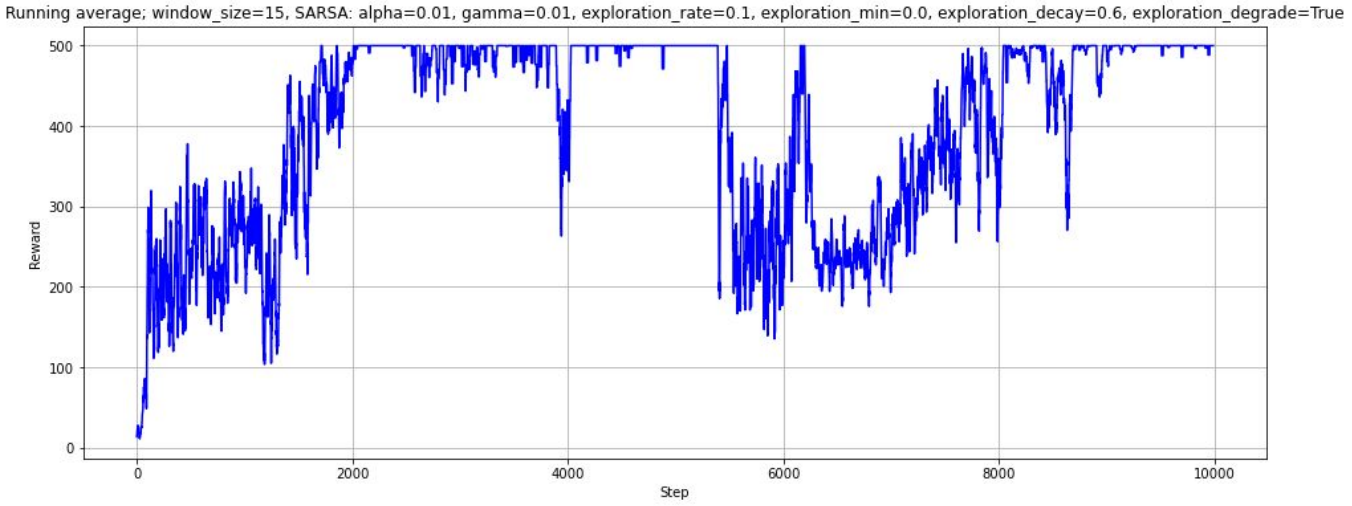
Dosyć dobre rezultaty uzyskano również wykorzystując poniższy zestaw parametrów:

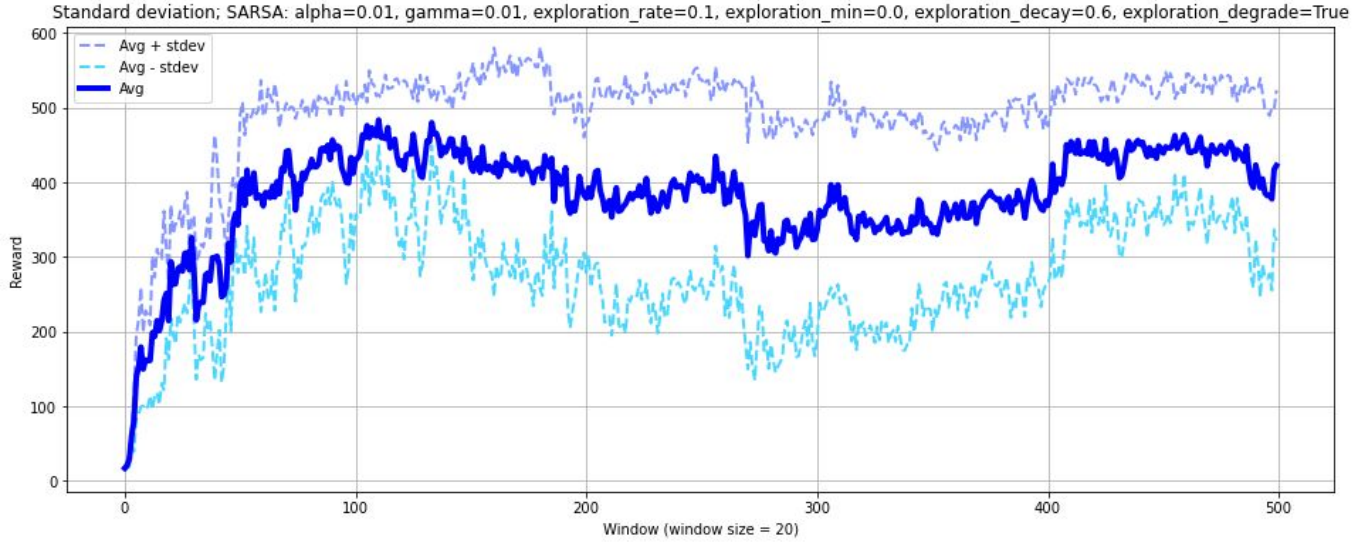
```
alpha=1.0, gamma=1.0, exploration_rate=0.01, exploration_min=0.0,
exploration_decay=0.95, exploration_degrade=True
```

gdzie score również szybko rośnie do wartości 400, jednakże zdecydowanie mniej stabilnie niż w poprzednim przypadku. Prawdopodobnie spowodowane jest to przez zbyt duże wartości **alpha** oraz **gamma**. Ostatecznie wynik stabilizuje się na 500 w okolicach 450 okna.

Pozostałe zbiory wartości parametrów powodują końcowo, średnio wyniki rzędu wartości ~400.

Zad. 3 - SARSA





Poprzednia metoda vs SARSA

Porównując wykresy dla tych samych wartości parametrów dla obu metod, można stwierdzić iż metoda SARSA w moim przypadku powoduje wzrost score'u w podobnym czasie, jednak stan ten się nie utrzymuje - jest o wiele mniej stabilny, przez co w okolicy 200 okna następuje degradacja wyniku. Jest on później odbudowany. W przypadku poprzedniej metody uczenia ze wzmocnieniem obserwujemy większą stabilność wyniku.