Dawid Woźniak
Informatyka - Uczenie Maszynowe I Sztuczna Inteligencja

Uczenie ze Wzmocnieniem

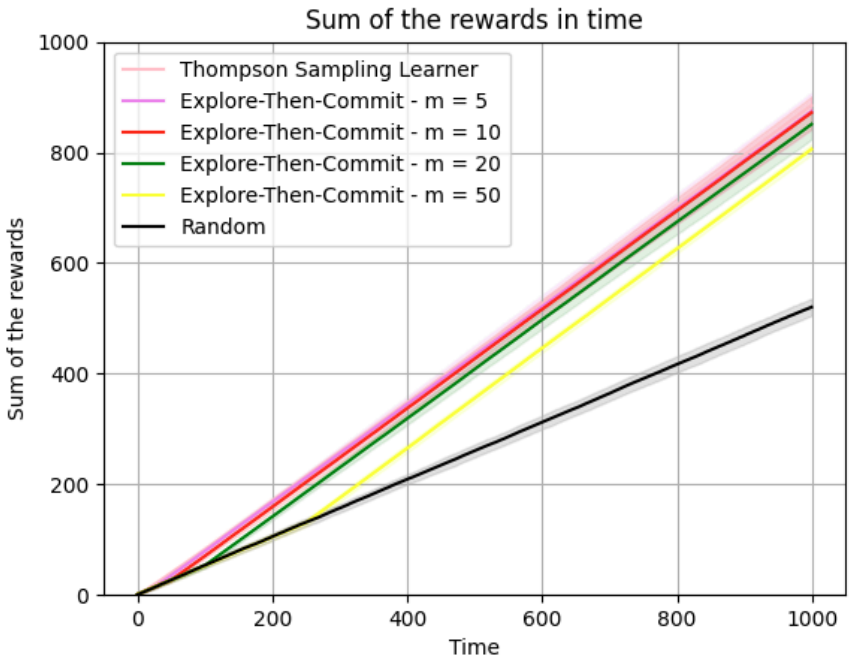
Problem K-rękiego bandyty

1. Opis ćwiczenia.

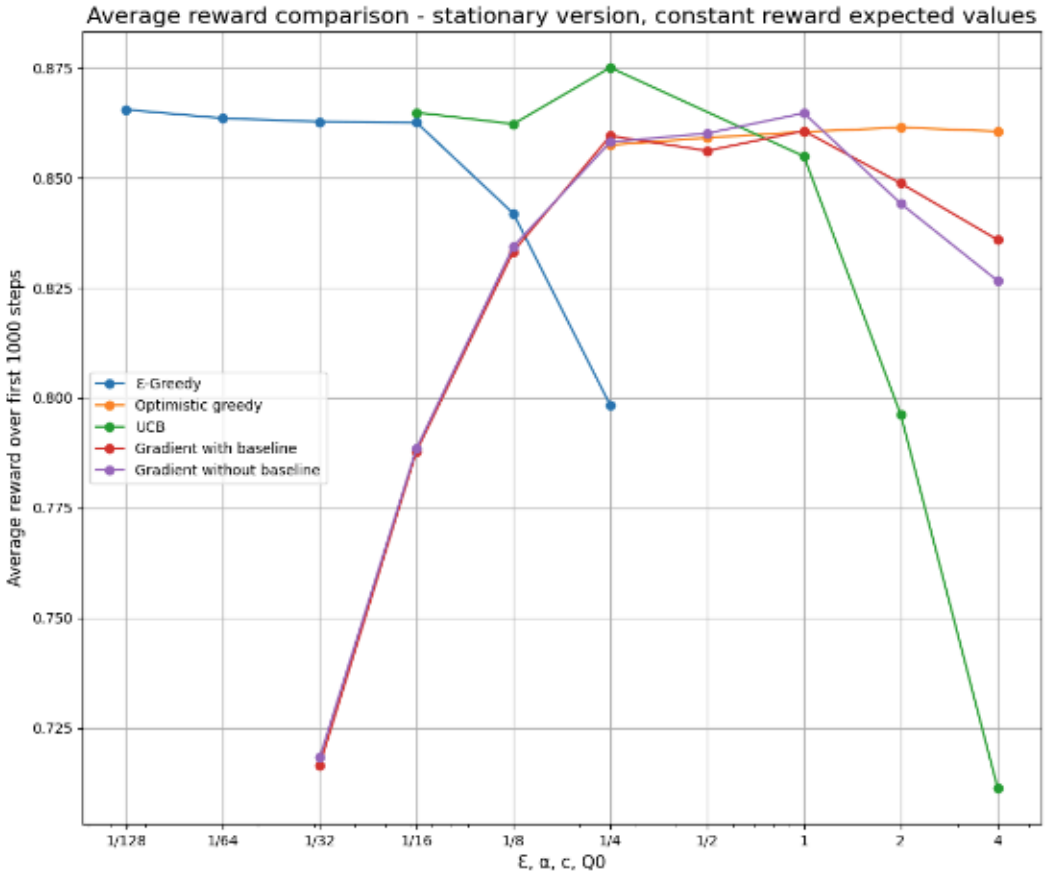
Celem laboratorium było zaimplementowanie i przetestowanie kilku algorytmów służących do rozwiązania problemu k-rękiego bandyty. Należało rozważyć wersję problemu stacjonarną jak i niestacjonarną oraz losowe wartości oczekiwane dla każdego problemu.

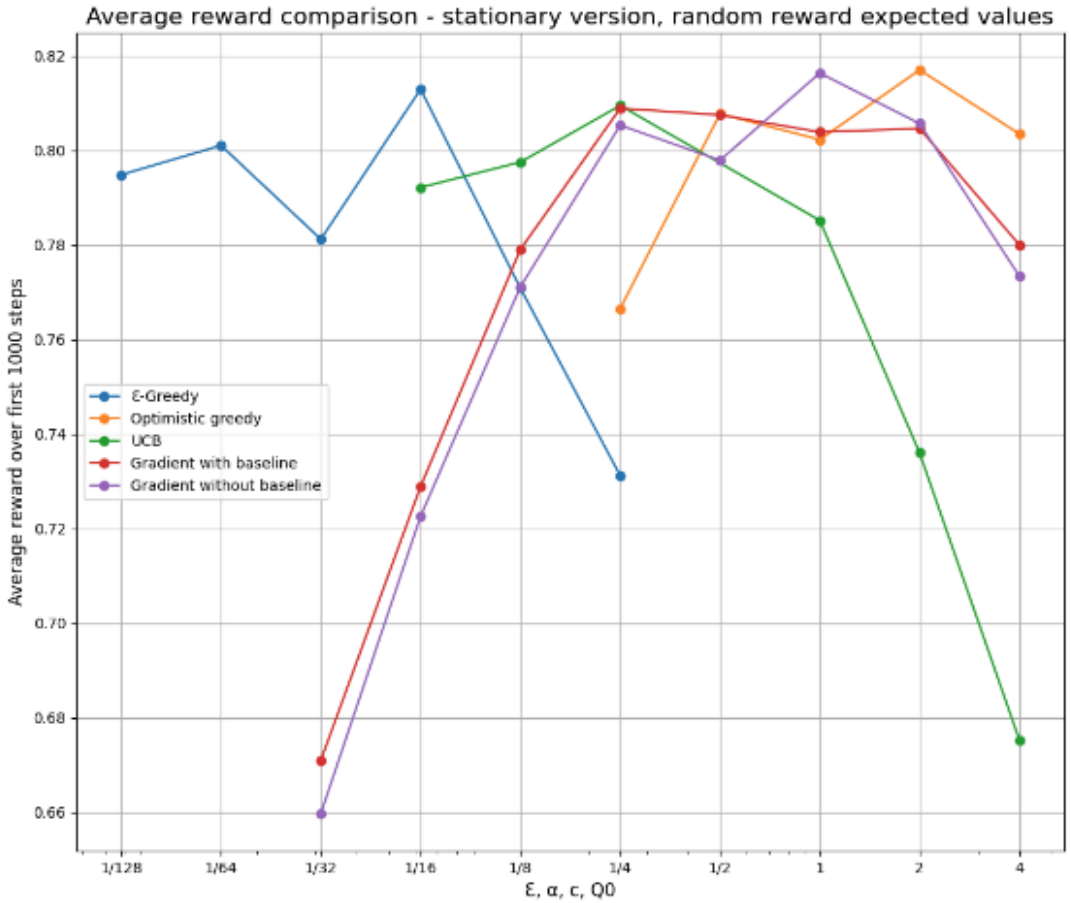
2. Wyniki.

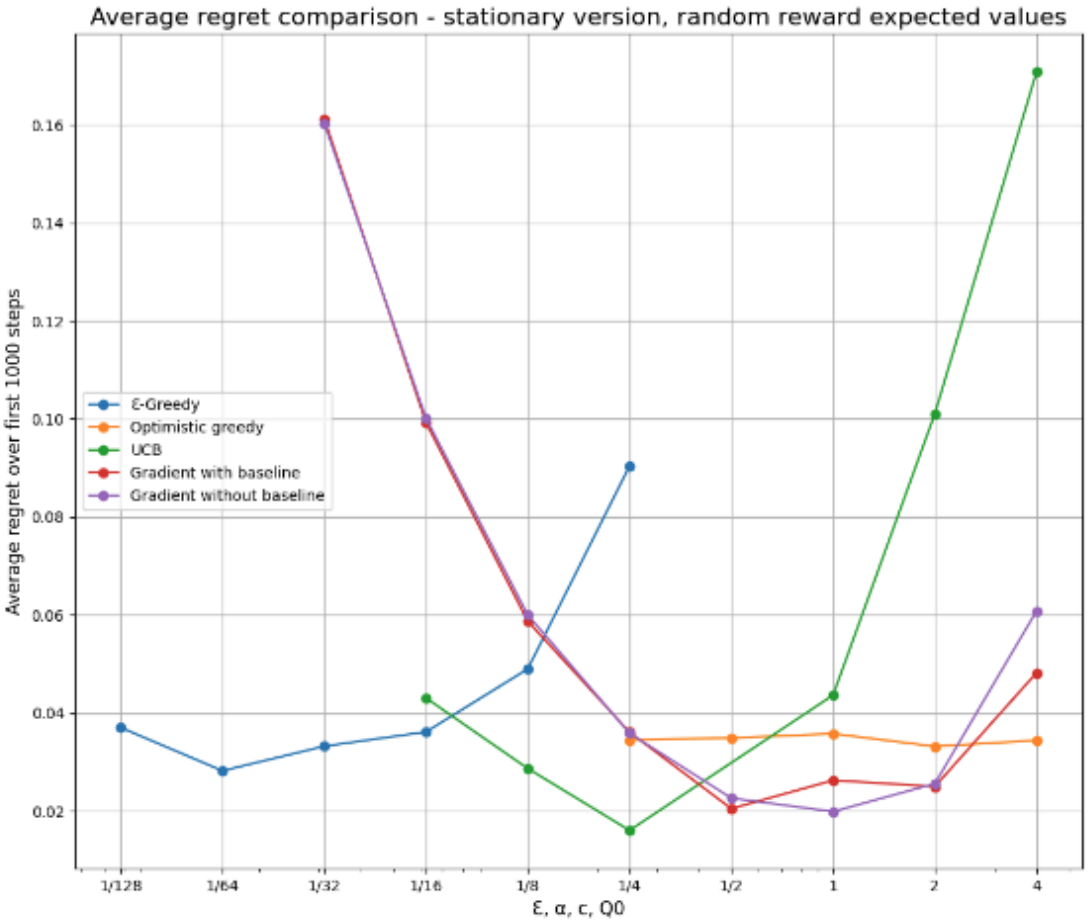
2.1. Porównanie Próbkowania Thompsona, Explore-Then-Commit oraz Random.



2.2. Studium parametryczne - wersja stacjonarna

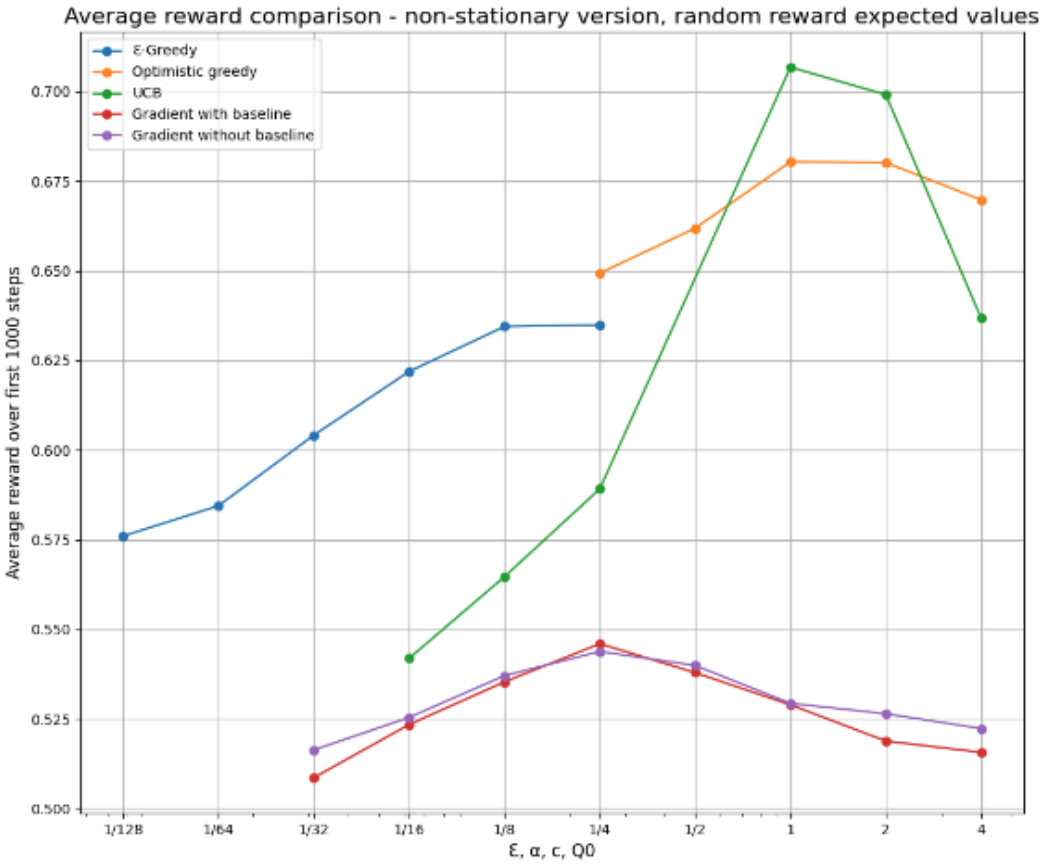


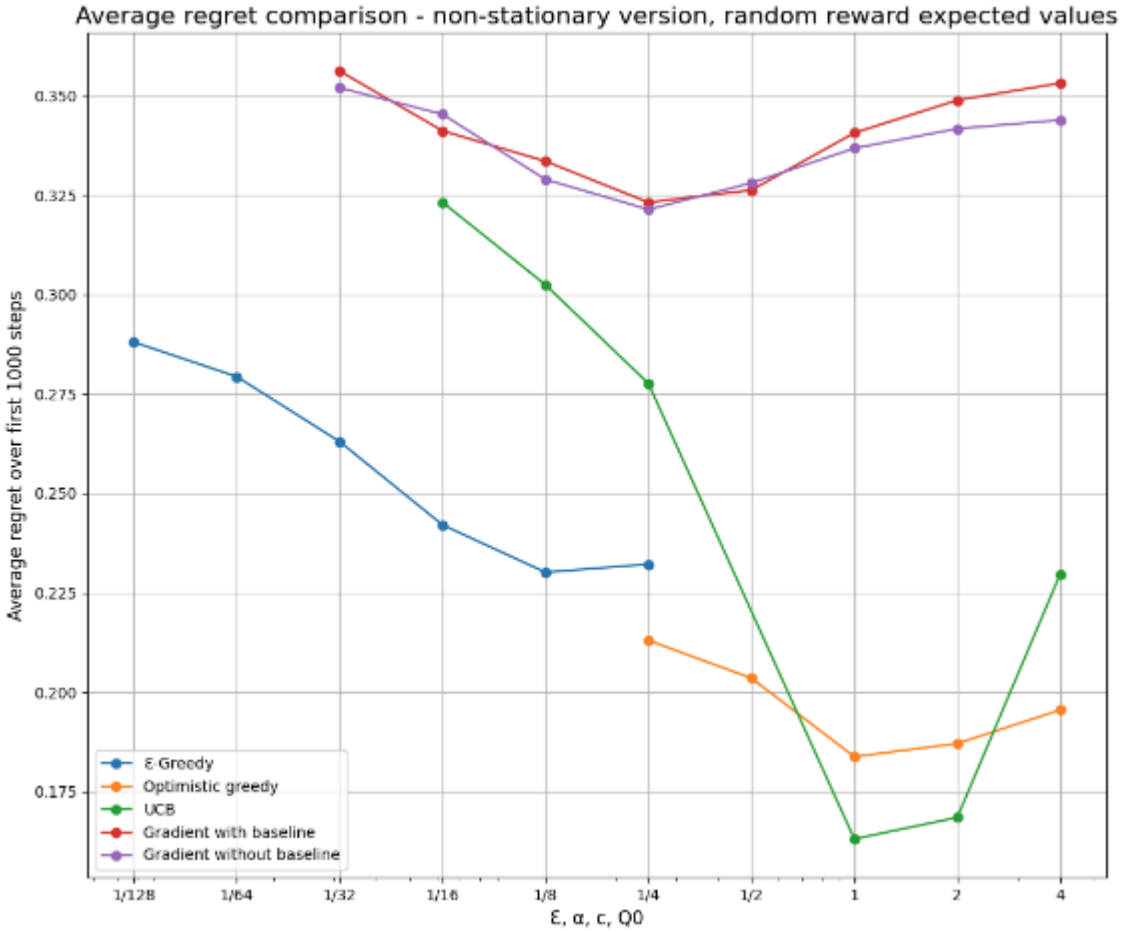




2.3 Studium parametryczne - wersja niestacjonarna

Dla wersji niestacjonarnej, w każdym kroku wartości nagród zostały uaktualnione o wartość wylosowaną z rozkładu normalnego podzieloną przez 10. Jeżeli po aktualizacji wartości nie mieściły się w przedziale [0,1], została im przypisana wartość krańcowa przedziału. Dla algorytmu zachłannego została zrealizowana średnia wykładniczo ważona aktualnością z parametrem learning rate równym 0.1.


Average reward comparison - non-stationary version, random reward expected values
ε-Greedy
Optimistic greedy
UCB
Gradient with baseline
Gradient without baseline
Average reward over first 1000 steps
0.700
0.675
0.650
0.625
0.600
0.575
0.550
0.525
0.500
1/128
1/64
1/32
1/16
1/8
1/4
1/2
1
2
4
ε, α, c, Q0


Average regret comparison - non-stationary version, random reward expected values
Average regret over first 1000 steps
0.350
0.325
0.300
0.275
0.250
0.225
0.200
0.175
ε-Greedy
Optimistic greedy
UCB
Gradient with baseline
Gradient without baseline
1/128
1/64
1/32
1/16
1/8
1/4
1/2
1
2
4
ε, α, c, Q0

3. Wnioski

W badanych przeze mnie przypadkach, zarówno w wersji stacjonarnej jak i nie stacjonarnej najepsze wyniki osiągnął algorytm Upper Confidence Bound, z parametrem 0.4 dla wersji stacjonarnej i 1 dla niestacjonarnej. Wyższa wartość dla wersji niestacjonarnej bierze się stąd, że jeżeli rozkład nagród zmienia się w czasie to powinniśmy zwiększać nasz rejon eksploracji. Algorytm ten zakłada mniejszy rozmiar eksploracji dla akcji, które odwiedziliśmy dużo razy. Gdy rozkład zmienia się w czasie, nie odwiedzali byśmy akcji bazując na poprzednich obserwacjach, które po zmianach rozkładu mogą być nieaktualne. Dlatego większa wartość parametru c, odpowiadająca wielkość przedziału ufności daje lepsze wyniki dla wersji niestacjonarnej. Podobną sytuację mamy dla algorytmu zachłannego. W wersji stacjonarnej, od wartości parametru epsilon równej około 1/16, jakość algorytmu spada. Dla wersji niestacjonarnej, jakość zwiększa się wraz z wzrostem parametru epsilon - więcej eksplorujemy więc możemy poznać nowy rozkład. W obu przypadkach, algorytm zachłanny z optymistyczną wartością początkową zadziałał lepiej niż wersja domyślna. Zaprezentowanie studium parametrycznego na jednym wykresie zależnym od poszczególnych parametrów pozwala w sposób efektywny porównać algorytmy. Wszystkie algorytmy zadziałały gorzej dla problemu niestacjonarnego. Jest to zagadnienie bliższe realnym problemom.