Kacper Ptaszek 331 427

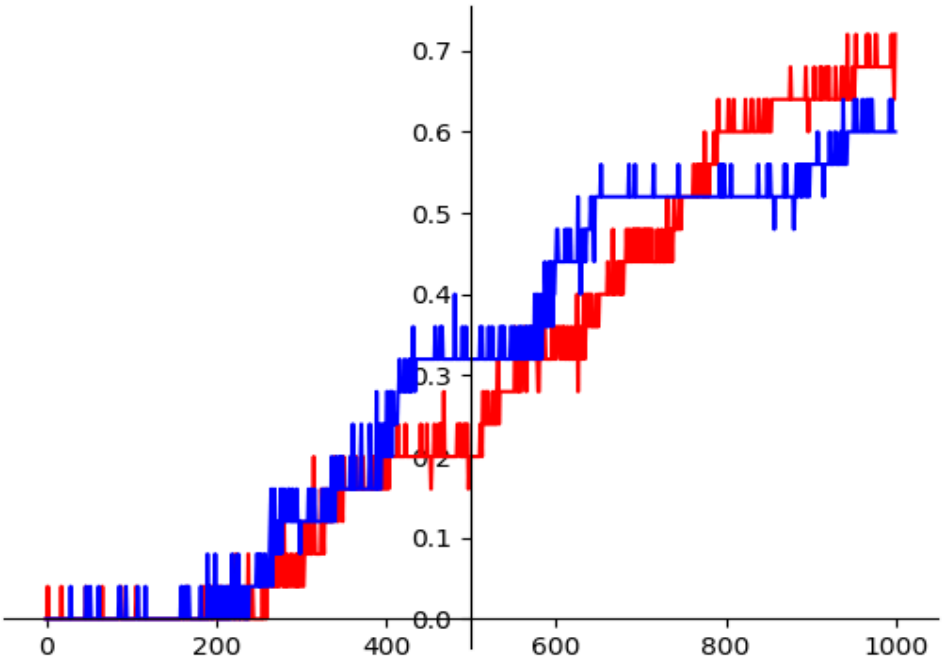
Sprawozdanie WSI ćwiczenie 6.

Treść zadania:
Proszę zaimplementować algorytm Q-Learning i użyć go do wyznaczenia polityki decyzyjnej dla problemu FrozenLake8x8. W problemie tym celem agenta jest przedostanie się przez zamarznięte jezioro z domu do celu, unikając dziur (zawsze rozpoczynamy epizod z górnego lewego rogu mapy, który ma współrzędne 0).

Na początku agent skupia się na eksploracji mapy, co opisuje współczynnik epsilon. Przy wyborze akcji przez agenta losowana jest liczba z rozkładu jednostajnego z przedziału [0;1), jeśli liczba ta jest mniejsza niż epsilon to agent wybiera losową akcję, aby eksplorować mapę. A w przeciwnym wypadku opiera swoją decyzję na wartościach z tabeli Q[stan, akcja]. Na początku epsilon wynosi 1 i z każdym kolejnym epizodem jest on odpowiednio zmniejszony, aby z kolejnymi epizodami agent kładł coraz mniejszy nacisk na eksplorację, a coraz bardziej skupiał się na eksploatacji.

1. Scenariusz bez poślizgu

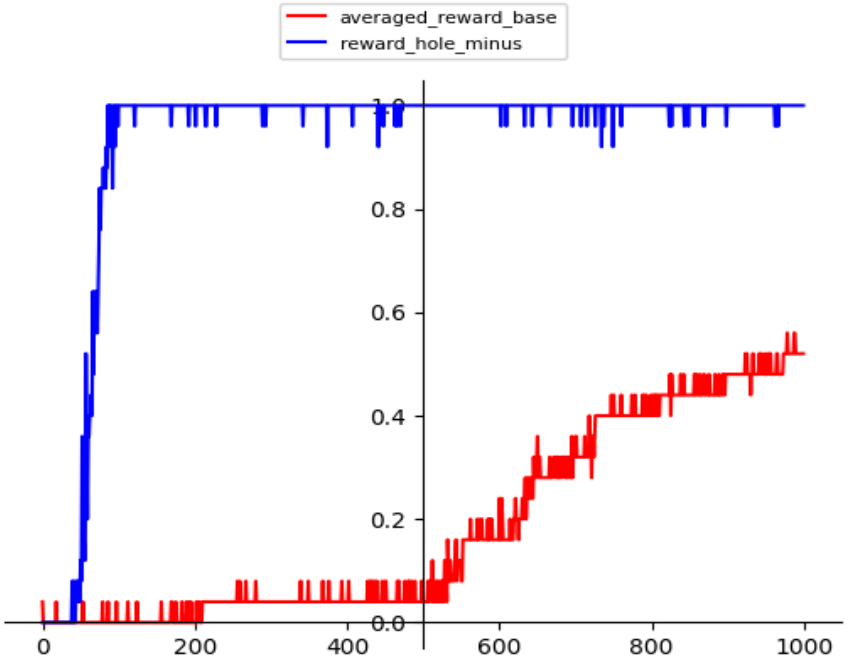
Wykres średniej ilości sukcesów w zależności od epizodu, dla 25 prób.
Obie krzywe są utworzone dla takich samych parametrów algorytmu.
Parametry: współczynnik uczenia = 0,3 ; współczynnik dyskontowania = 0,9



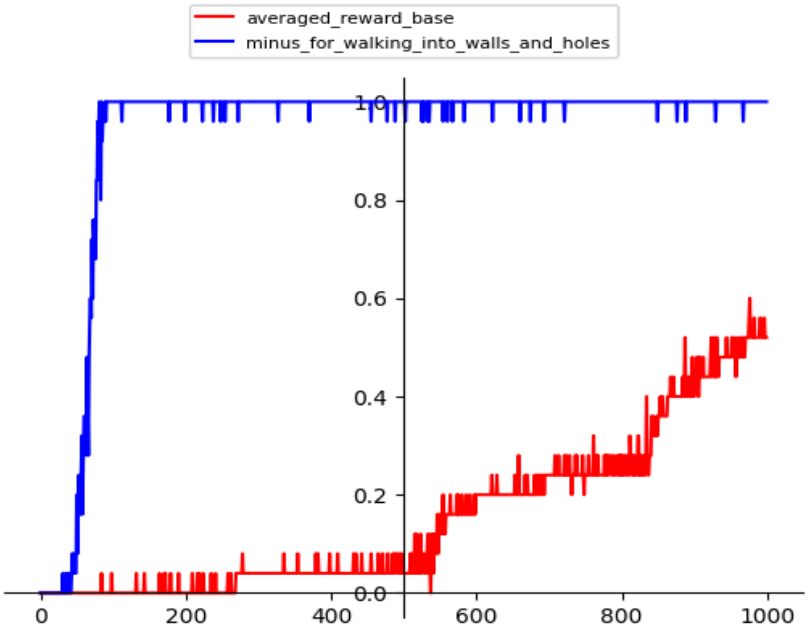
Widać że losowość ma duży wpływ na działanie algorytmu i mimo takich samych parametrów algorytmu agent może uczyć się w różnym tempie.

Wpływ zmiany systemu nagród na agenta

a) Pierwszy zaproponowany przeze mnie system nagród to: -1 za wpadnięcie do dziury i +100 za dotarcie do prezentu
Parametry algorytmu podczas testowania:
- współczynnik uczenia = 0,3
- współczynnik dyskontowania = 0,9
- liczba epizodów = 1000
- maksymalna liczba kroków agenta = 200



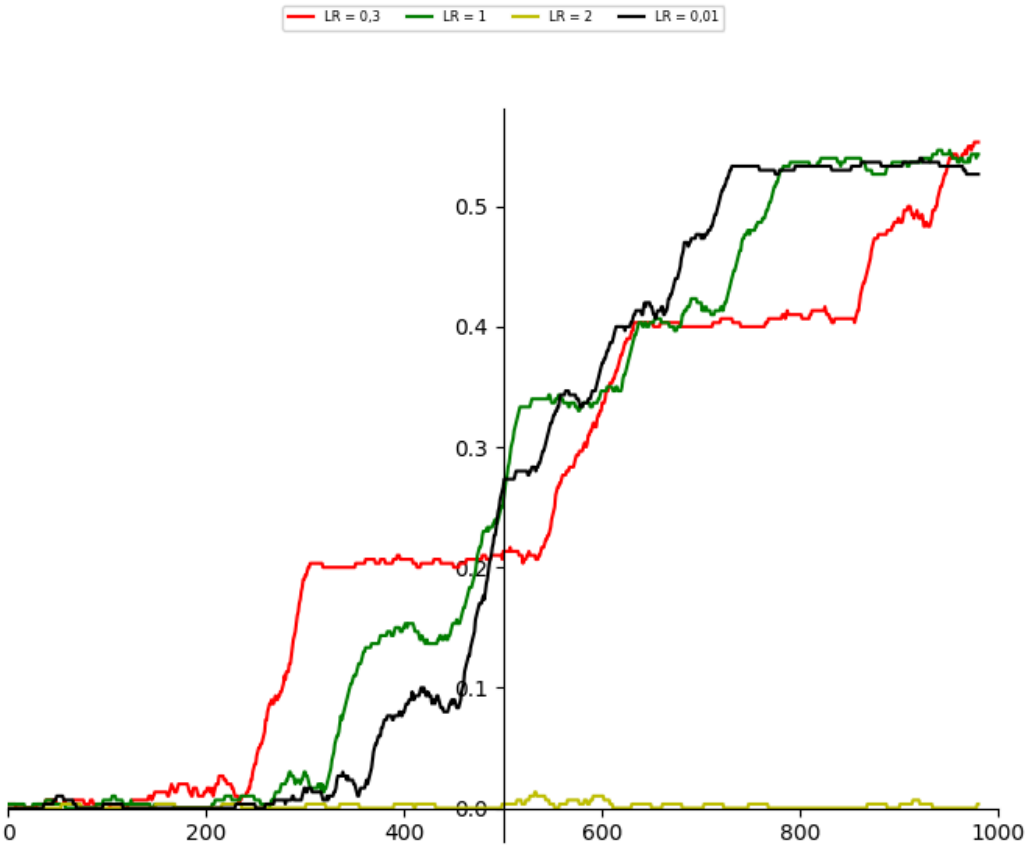
b) Drugi zaproponowany system nagród to: -2 za wpadnięcie do dziury, -2 za zostanie w tym samym stanie, +10 za dotarcie do prezentu.



Jak widać na wykresach domyślny system nagród jest w stanie zagwarantować skuteczność 50-60% dojścia do prezentu po 1000 epizodach nauki. Inne dwa zaproponowane systemy nagród uczą agenta, aby unikać dziur i to okazuje się kluczem do szybkiej nauki ścieżki do prezentu, gdyż agent dociera tam już po 100 epizodach uczenia się prawie za każdym razem.

Wpływ zmiany współczynnika uczenia się

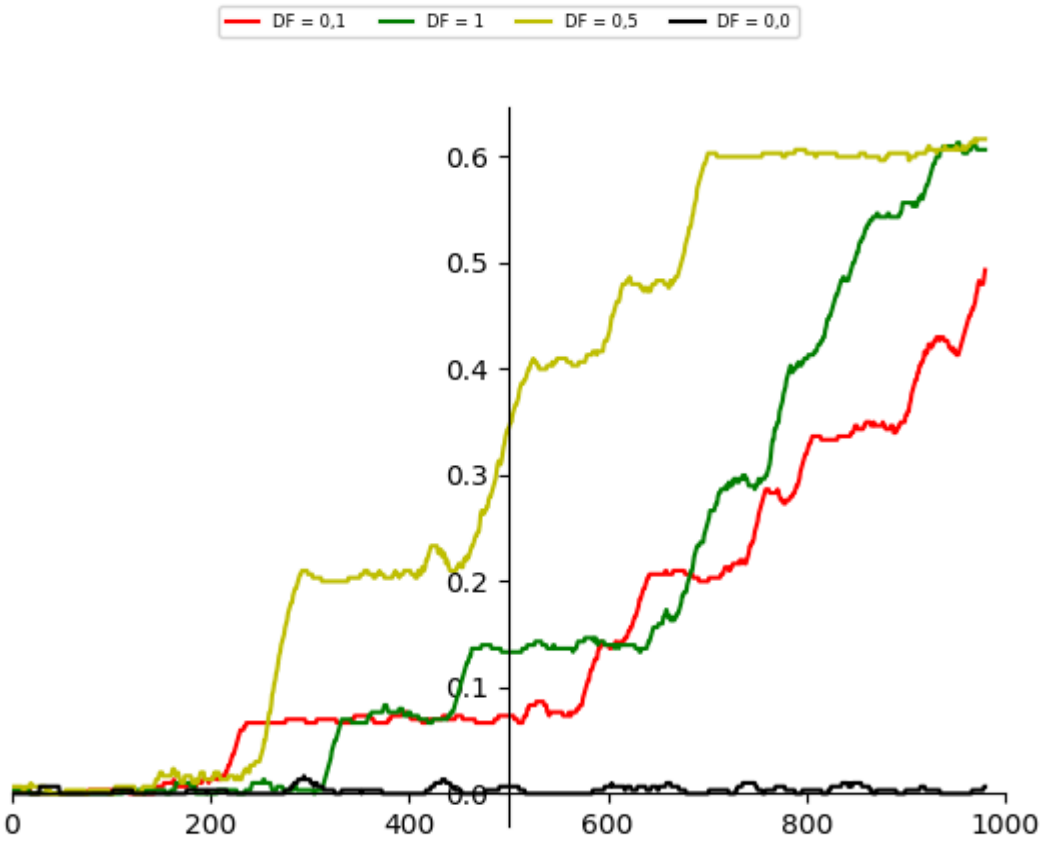
Wykonałem testy nauki agenta dla 4 różnych współczynników uczenia się, uzyskane wartości sukcesów zostały obliczone dla średnich z 25 uruchomień, a przy testach używany był domyślny system nagród oraz współczynnik dyskontowania=0,9:



Przy wyższym współczynniku uczenia się agent szybko aktualizuje swoją tabelę Q, zatem jego nowe doświadczenia mają większy wpływ. Z drugiej strony mały współczynnik uczenia się powoduje, że adaptacja do nowych informacji staje się znacznie wolniejsza. Mimo tych różnic wyniki dla współczynników uczenia się równych 0,3; 1 lub 0,01 są podobne. Jedynie współczynnik uczenia się równy 2 okazał się złym wyborem, gdyż agent kładł za duży nacisk na nowe doświadczenia i tym samym bardzo rzadko udawało mu się dotrzeć do prezentu.

Wpływ zmiany współczynnika dyskontowania

Wykonałem testy nauki agenta dla 4 różnych współczynników dyskontowania, uzyskane wartości sukcesów zostały obliczone dla średnich z 25 uruchomień, a przy testach używany był domyślny system nagród oraz współczynnik uczenia=0,3:

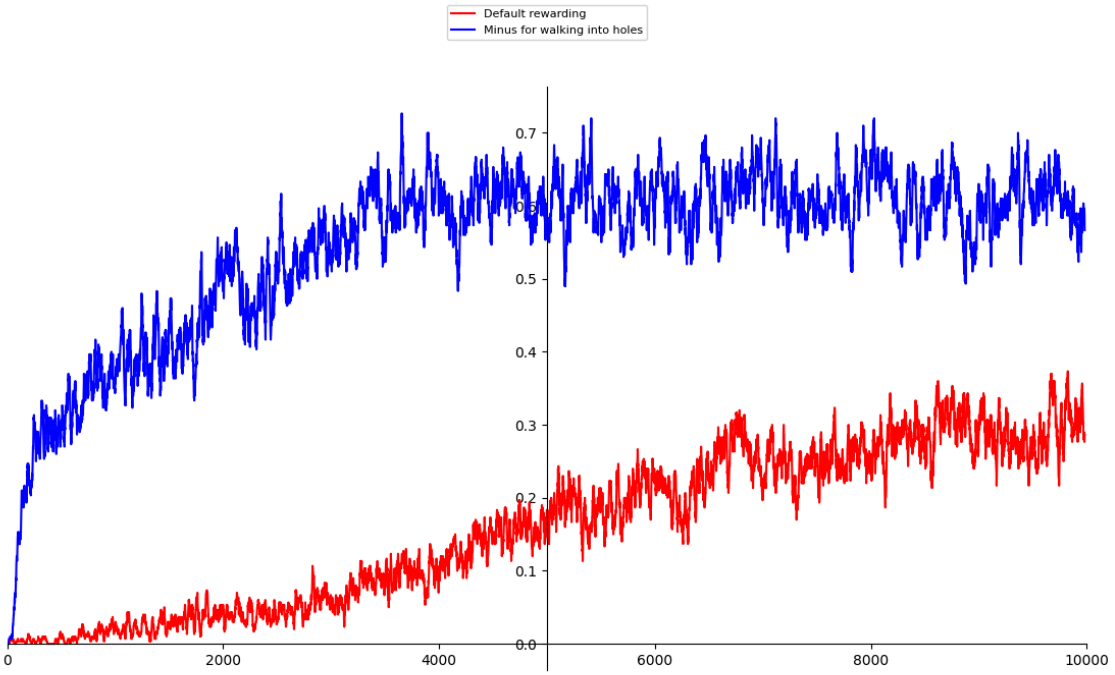


Uzyskane wyniki dla większych współczynników dyskontowania są lepsze niż te dla mniejszych współczynników. Gdyż dla wyższych współczynników dyskontowania agent bardziej ceni nagrody możliwe do uzyskania w przyszłości niż obecnie uzyskane. Gdy współczynnik ten jest równy 0 agent skupia się wyłącznie na nagrodach natychmiastowych i ignoruje nagrody przyszłe. Jak widzimy takie podejście nie pozwoliło agentowi dobrze nauczyć się drogi do prezentu.

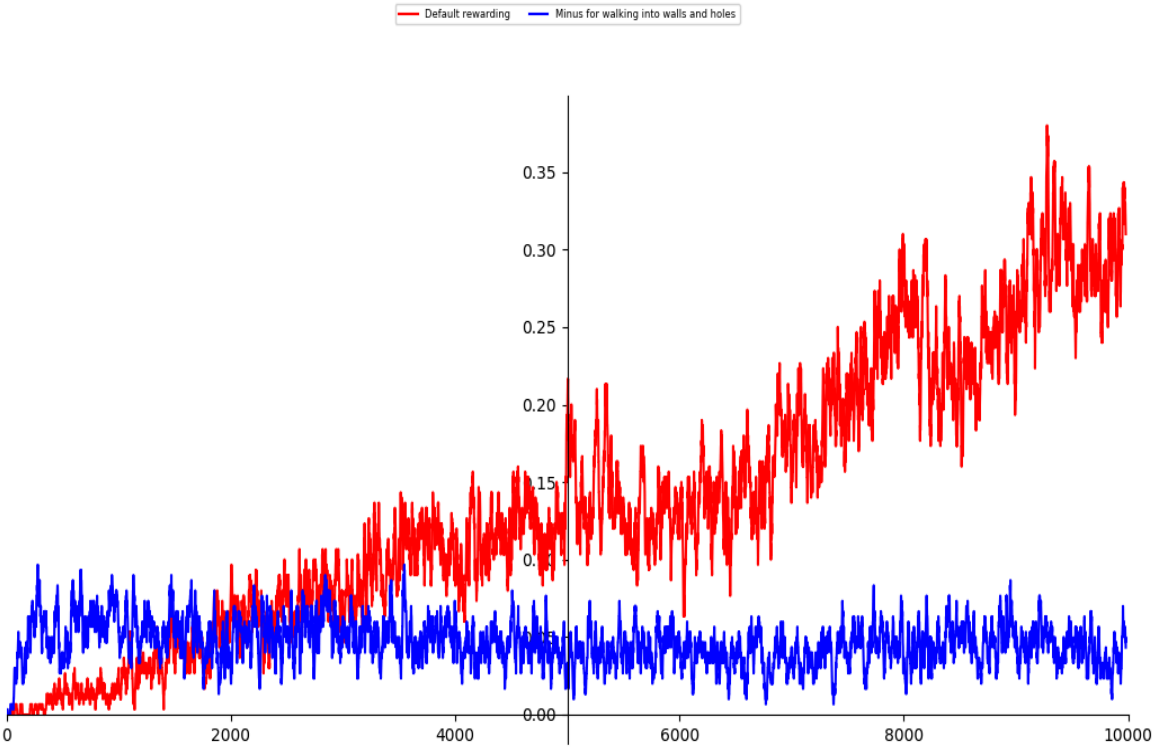
2. Scenariusz z poślizgiem

Dla wersji z poślizgiem wykonywałem testy przy 10000 epizodach uczenia się. Gdyż ta wersja środowiska jest znacznie cięższa do nauki. Dane uzyskałem z 25 niezależnych uruchomień algorytmu uczenia się. Wykres dla dwóch różnych systemów nagrody:

a) domyślny system nagród i pierwszy zaproponowany przeze mnie system nagród (-1 za wpadnięcie do dziury i +100 za dotarcie do prezentu)



b) domyślny system nagród i drugi zaproponowany system nagród (-2 za wpadnięcie do dziury, -2 za zostanie w tym samym stanie, +10 za dotarcie do prezentu).



Domyślny system nagród nawet dla zwiększonej liczby epizodów do 10000 osiąga zaledwie 30% sukcesów. Jest to spowodowane tym, że wybrana akcja przez agenta jest wykonywana jedynie w ⅓ liczbie przypadków. Na obu wykresach widać stały

wzrost sukcesów dla domyślnego systemu nagród wraz z kolejnymi epizodami. Za to pierwszy zaproponowany system nagród lepiej wpływa na naukę agenta, gdyż agentowi udaje się dotrzeć do prezentu nawet w 60% prób mimo poślizgu. A drugi system, który każe dodatkowo agenta za wejście w ścianę powoduje znacząco gorsze wyniki niż w wersji bez poślizgu. Jest to spowodowane większym unikaniem krawędzi planszy przez agenta i tym samym przypadkowym wpadaniem do dziur z powodu poślizgnięcia.

Wnioski końcowe:

W wersji środowiska bez poślizgu domyślna funkcja oceny jest znacznie gorsza od zaproponowanych przeze mnie systemów nagrody. Potwierdza to, że dodatkowa informacja, którą uzyskuje agent jest dla niego znacznie pomocna w dotarciu do prezentu. Obie alternatywne funkcje oceny wypadają bardzo podobnie i ciężko jest wskazać lepszą z nich, gdyż obie już po około 100 epizodach gwarantują sukces w większości prób. Za to w wersji z poślizgiem drugi alternatywny system nagród wypada znacznie gorzej (10% sukcesów) od mojej pierwszej funkcji oceny (60% sukcesów). A z domyślną funkcją oceny agent potrzebuje dużej liczby epizodów w porównaniu do wersji bez poślizgu, aby nauczyć się środowiska przynajmniej przeciętnie.