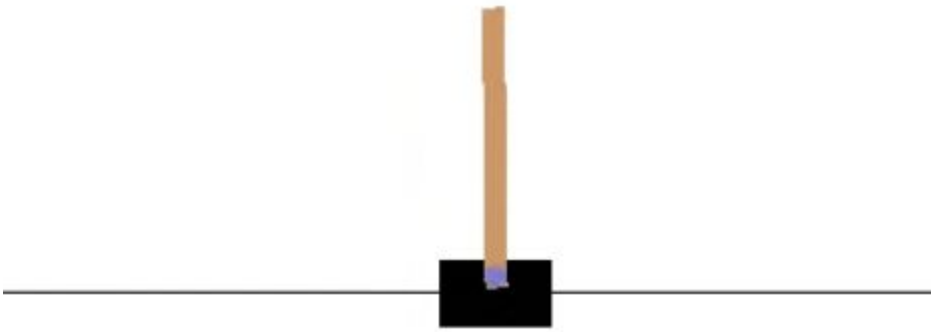
Roboty Mobilne

Uczenie ze wzmocnieniem
Problem odwróconego wahadła



Wprowadzenie

Zagadnienie odwróconego wahadła jest akademicki przykładem, rozważań sterowania w przestrzeni stanów. W klasycznym podejściu często korzysta się z odpowiednio dostrojonego sterownika PID, którego wzmocnienia zostają wyliczone m.in. na podstawie wyprowadzeń zależności fizycznych. W nowoczesnych metodach sterowania do rozwiązania danego zagadnienia wykorzystuje się szeroko pojęte uczenie maszynowe. W rozważaniach został przyjęty model odwróconego wahadła matematycznego zaimplementowany z biblioteki “gym” w środowisku Phyton. Celem projektu było ustabilizowanie wahadła z wykorzystaniem algorytmu Q-learning.

Opis

1. Algorytm

   Do realizacji zagadnienia wykorzystano następujący algorytm Q-learning:

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha\left[r + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)\right]$$

Gdzie:

Q(s,a) - macierz wzmocnień której wymiar bezpośrednio wynika z długości wektora stanu oraz wektora akcji zgodnie z następującym wzorem:

$$Q : S \times A \to \mathbb{R}\,.$$

∝ - współczynnik uczenia

γ- ocena otrzymanych nagród [0-1]

r - nagroda

W celu dyskretyzacji z przedziałów ciągłych opisujących stany skorzystano z następującej zależności przeskalowywującej:

$$V' = \frac{(V - \min)}{\max - \min} * (new\_\max - new\_\min) + new\_\min$$

Dana funkcja pozwala na przeskalowanie zmiennej ze zbioru pierwszego na jej reprezentację w zbiorze drugim.

2. Środowisko

   Biblioteka “gym” zawiera w sobie wiele przydatnych funkcji pozwalających na szybką obróbkę informacji dotyczących otoczenia. Wykorzystana funkcja:

   obserwacja, nagroda, ukończenie, błędy = env.step(akcja)

   obserwacja - zawiera aktualny stan w skład, którego wchodzą kolejno:

   1. Położenie wózka [-2.4, 2.4]
   2. Prędkość wózka [inf]
   3. Kąt odchylenia wahadła [-18°, 18°]
   4. Prędkość kątową wahadła [inf]

   nagroda - nagroda za uniknięcie resetu środowiska ustalona stale na 1

   ukończenie - zmienna boolowska informująca o zakończeniu jednego cyklu symulacyjnego na skutek przekroczenia dopuszczalnego kąta odchylenia przez wahadło, wykroczenie wózka poza zakres tudzież przekroczenie 200 ruchów

   błędy - nie zostały wykorzystane w algorytmie

3. Implementacja

   Przestrzeń stanu została zdefiniowana jako:

   s= [kąt wahadła, prędkość kątowa wahadła]

   Przy czym przedziały ciągłe zostały poddane dyskretyzacji, poprzez skalowanie wartości do nowych zakresów a następnie zaokrąglano wynik do liczby całkowitej i odejmowano jeden. Dzięki czemu uzyskano zdyskretyzowaną ilość stanów. Długości wektorów stanu zostały dobrane doświadczalnie: 65, 42, kolejno dla kąta odchylenia i prędkości kątowej.

Wektor akcji jest zdefiniowany w środowisku jako:

a = [0,1]

Zatem deklaracja macierzy jest następująca: Q[[s],[a]], gdzie wymiar przestrzeni stanu 65x42 a wektora akcji 2. Wymiar zadeklarowany w środowisku: (1,1,62,42,2), deklaracja umożliwia wykorzystanie pozostałych zmiennych stanu jakim jest prędkość i przemieszczenie wózka.

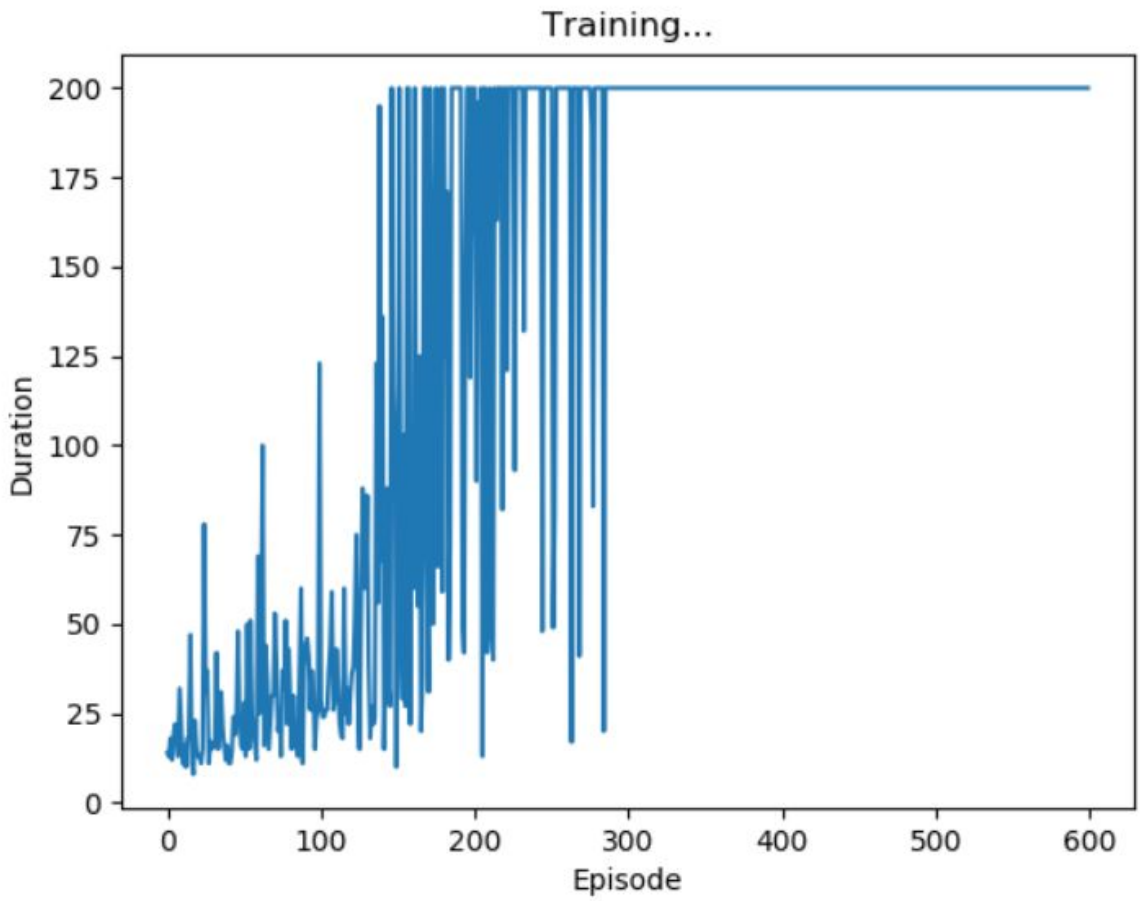
Funkcja kary każdorazowa za ukończenie gry przed czasem wynosi -60, za opuszczenie przestrzeni w pobliżu punktu środkowego wynosi -10, nagroda za utrzymanie wahadła w danej chwili wynosi 1.

Współczynnik nauczania jest wygaszana w czasie poprzez wykorzystanie logarytmu przy podstawie 10.

W analogiczny sposób wygaszany jest epsilon, współczynnik eksploracji w celu uniknięcia zatrzymania się w lokalnym minimum nie może być zbyt mały jednakże z biegiem czasu należy zmniejszać jego wartość. Wartość zostaje zmniejszona za wykorzystaniem funkcji logarytmicznej następnie przypisana jest minimalna wartość współczynnika, kolejno jest on zerowany tuż przed końcem nauczania. W ten sposób uzyskaliśmy najlepszą odpowiedź układu.

Współczynnik gamma został określony na 1 co zapewnia “dalekowzroczność” algorytmu podczas dobierania najbardziej korzystnej ścieżki rozwiązania.

Obserwacje



Rys. 1 Symulacja pracy algorytmu

Wnioski:

Dostrojenie algorytmu wymaga przeprowadzenia wielu prób testowych. Na poprawę jakości sterowania duży wpływ ma wymiar macierzy Q, czas treningu, oraz sposób w jaki są wygaszane poszczególne współczynniki. Dużą rolę odgrywa funkcja kary, która została dobrana w sposób doświadczalny. Największym problemem podczas realizacji były sytuację “wpadania w lokalne minimum” tudzież zbyt późne wygaszenie współczynnika epsilon.

Realizacja:

Marcin Januszewicz 160622

Jan Glinko 161259

Michał Topka 160836

Dawid Łukaszewski 160539