Q-Learning

CSE-5364 - Fall 2017

Learning to balance an inverted pendulum

Q-learning is a model-free reinforcement technique. Specifically, Q-learning can be used to find an optimal action-selection policy for any given (finite) Markov Decision Process. It works by learning an action value function that ultimately gives the expected utility of taking a given action in each state and following the optimal policy thereafter. -wikipedia

The following equation has been used for solving the problem -

$$Q^{*}_{t+1}(s, a) = (1-\alpha) \times Q^{*}_{t}(s, a) + \alpha \times (r(s) + \gamma \times \max_{b} Q^{*}_{t}(s', b))$$

where

Q_{t+1}: optimal Q value of current state

s: current state

a: current action

α: learning rate

Q_t: optimal Q value of previous state

r(s): Reward associated with current state

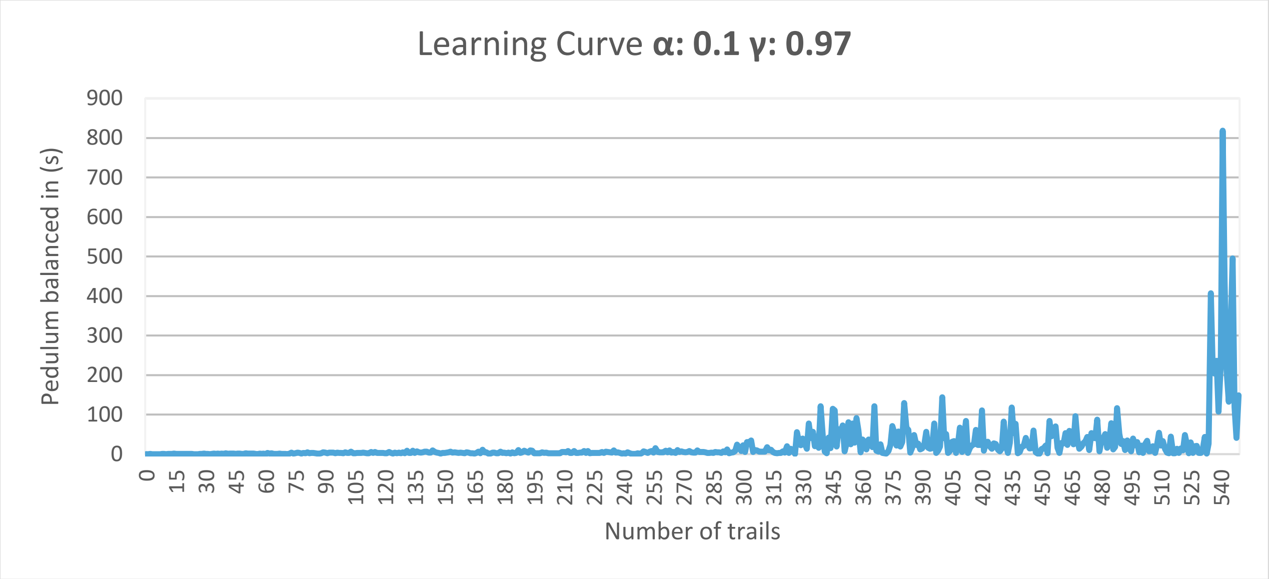
γ: Discount factor

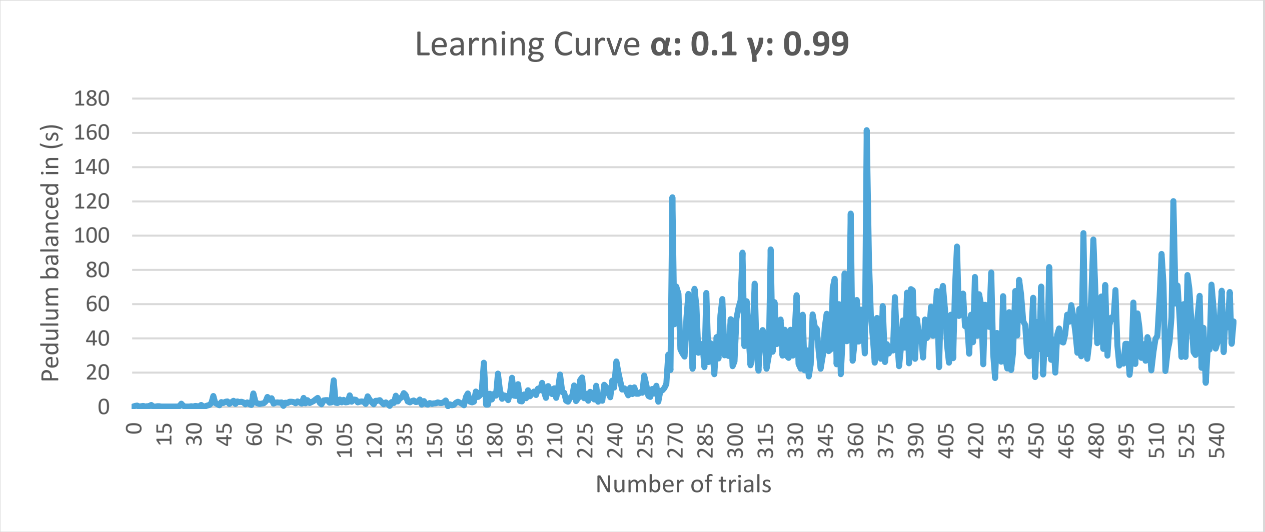
s': next state

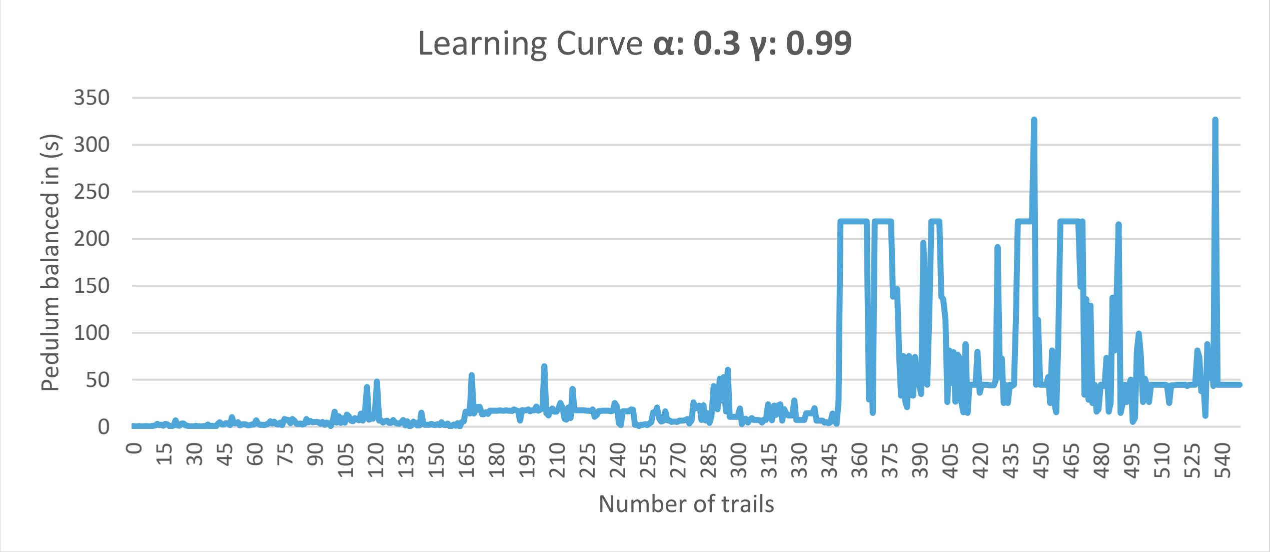
b: all possible actions

The code is attached alongside.

Below are the results of the learning curve obtained after implementing the above equation


Learning Curve α: 0.1 γ: 0.97
Pedulum balanced in (s)
900
800
700
600
500
400
300
200
100
0
0 15 30 45 60 75 90 105 120 135 150 165 180 195 210 225 240 255 270 285 300 315 330 345 360 375 390 405 420 435 450 465 480 495 510 525 540
Number of trails


Learning Curve α: 0.1 γ: 0.99
Pedulum balanced in (s)
180
160
140
120
100
80
60
40
20
0
0 15 30 45 60 75 90 105 120 135 150 165 180 195 210 225 240 255 270 285 300 315 330 345 360 375 390 405 420 435 450 465 480 495 510 525 540
Number of trials


Learning Curve α: 0.3 γ: 0.99
Pedulum balanced in (s)
350
300
250
200
150
100
50
0
0 15 30 45 60 75 90 105 120 135 150 165 180 195 210 225 240 255 270 285 300 315 330 345 360 375 390 405 420 435 450 465 480 495 510 525 540
Number of trails