TDDC17 Lab 5 Report

Alim Amanzholov (aliam864)

Part 2

Question 1 (theory): In the report, a) describe your choices of state and reward functions, and b) describe in your own words the purpose of the different components in the Q-learning update that you implemented. In particular, what are the Q-values?

a) I discretized the angle uniformly into 12 states. For a reward for the angle I chose function **8-abs(angle)** since angle 0 is desirable and getting away from 0 is not desirable. Hence from my choice of the reward function, the agent gets the maximum reward if it is at 0 or close to that value.

b) The components of Q-learning update are Q-value of the current state, learning rate alpha (which decreases over time), the reward for the current state, gamma discount factor (low discount factor makes the agent short-sighted and greedy and high discount factor makes the agent patient), highest Q-value of any action in the current state and the old Q-value of the current state. Q-values are previous states of the environment and they help to identify what is the best action to take from the current state.

Question 2: Try turning off exploration from the start before learning. What tends to happen? Explain why this happens in your report.

If exploration is turned off from the start before agent gets to learn, the agent will perform poorly since it has visited only a few states. The agent will stick to the local optimum that it got from the states it had the chance to visit.

Part 3

There are 10 states for the angle, 8 states for vx and 6 states for vy. We also have 8 possible actions. So the number of possible states and actions is 10*8*6*8 = 3840. For hover reward, I just summed the individual rewards for angle, vx, and vy. I thought that closer the values of vx and vy to 0 the better and therefore used similar reward function for these values as for angle in part 2. (I used **4-abs(vx)** and **4-abs(vy)** for rewards)