Q-Learning

By: margu424 & axega544

Question 1 (theory):

a) describe your choices of state and reward functions answer: We decided to go with a simple implementation of states consisting of states up, left, right, hard_left, hard_right. We did this distinction between hard_left and left as we felt a greater lean angle would require more drastic actions. We therefore made the reward for being in hard_left or hard_right -5 which then means that this state is much worse to be in than the states left or right which were given the reward -1. The up state was then rewarded with +1 as this was the desired state for the rocket.

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? answer:

```
Qtable.get(prev_stateaction) + alpha(num_tested) *
(previous_reward + GAMMA_DISCOUNT_FACTOR * getMaxActionQValue(new_state) - Qtable.get(prev_stateaction));
```

Qtable.get(prev_stateaction) : previous Q-value for this state and action alpha(num_tested) : Computes the learning rate which is used to make the same action and state multiple times mean less in term of the Q-value. previous_reward : reward we got for doing the previous action getMaxActionQValue(new_state): finds the action that give the highest Q-value and returns the Q-value. Used to say in the best case what Q-value will we get in this state.

This all then means that the Q-values are the expected value of being in a state and action. So if we are in a state and action we are expected to get the value of the Q-value as a reward if that action is chosen.

Question 2:

Try turning off exploration from the start before learning. What tends to happen? Explain why this happens in your report. answer: If we turned off exploration from the start the agent never learned to use any rockets which led to it just dropping to the ground. Then when we but the agent in a bad situation it eventually recovered but it took a lot longer and once it had recovered it went back to not using any engines which is what it had been training to do. This happens because when we never explore it will always maximize the Q-value and as it knows it can get a reward for not doing anything it will keep doing it as it will maximize the Q-value.