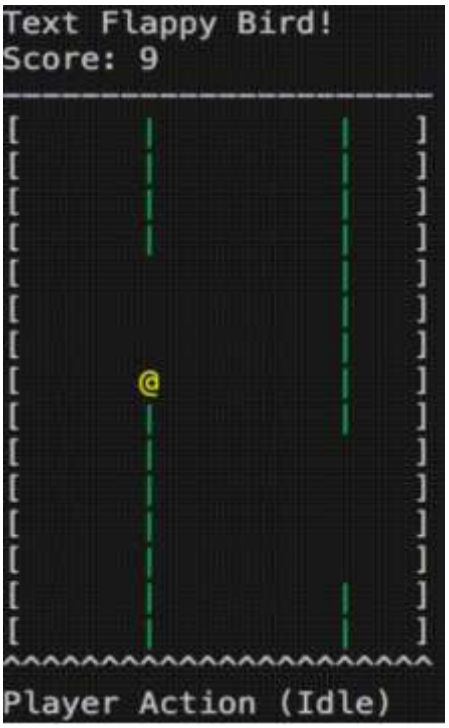
Reinforcement Learning Individual Assignment

Raghuwansh Raj (raghuwansh.raj@student-cs.fr)

1. Description:

The goal of this assignment is to apply reinforcement learning methods to a simple game called Text Flappy Bird (TFB). The game is a variation to the well know Flappy Bird in which the player is made with a simple unit-element character as can be seen in Figure 1.



2.Two Agents that are employed to solve the flappy bird using reinforcement learning:

Number of action_states:2

[0,1] flip and idle

1. Q-Learning(off policy RL Technique)

Q-learning is a model-free reinforcement learning algorithm to learn the value of an action in a particular state. It does not require a model of the environment (hence "model-free"), and it can handle problems with stochastic transitions and rewards without requiring adaptations. Q-learning can identify an optimal action-selection policy for any given FMDP, given infinite exploration time and a partly-random policy.

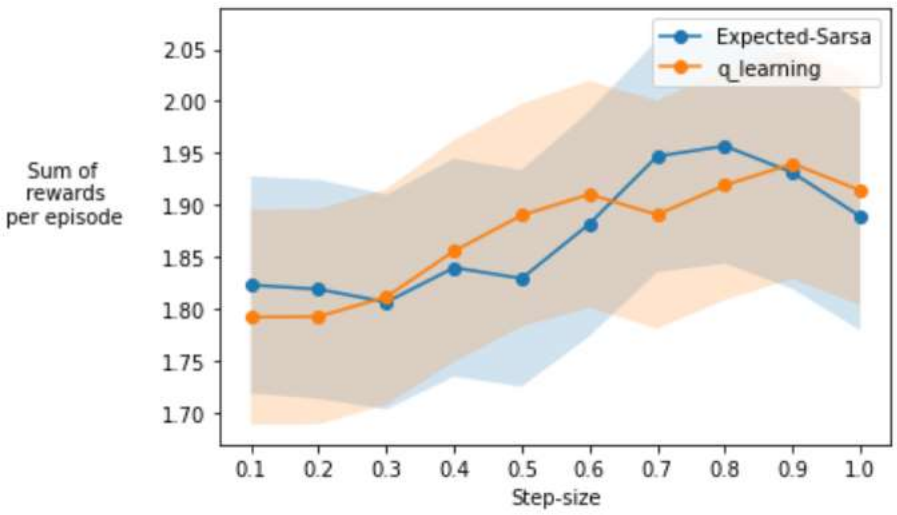
$$Q^{new}(s_t, a_t) \leftarrow \underbrace{Q(s_t, a_t)}_{\text{old value}} + \underbrace{\alpha}_{\text{learning rate}} \cdot \overbrace{\Big(\underbrace{r_t}_{\text{reward}} + \underbrace{\gamma}_{\text{discount factor}} \cdot \underbrace{\max_a Q(s_{t+1}, a)}_{\text{estimate of optimal future value}}}^{\text{new value (temporal difference target)}} - \underbrace{Q(s_t, a_t)}_{\text{old value}}\Big)}^{\text{temporal difference}}$$

2. Sarsa-Learning(on-policy RL technique)

A SARSA agent interacts with the environment and updates the policy based on actions taken, hence this is known as an on-policy learning algorithm. The Q value for a state-action is updated by an error, adjusted bythe learning rate. Q values represent the possible reward received in the next time step for taking action a in state s, plus the discounted future reward received from the next state-action observation

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\,[r_t + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)]$$

3.Sensitivity towards parameters



Our, on-policy, SARSA agent views the pole riskier because it chooses and updates actions subject to its stochastic policy. That means it has learned it has a high likelihood of hitting off the pole and receiving a high negative reward.

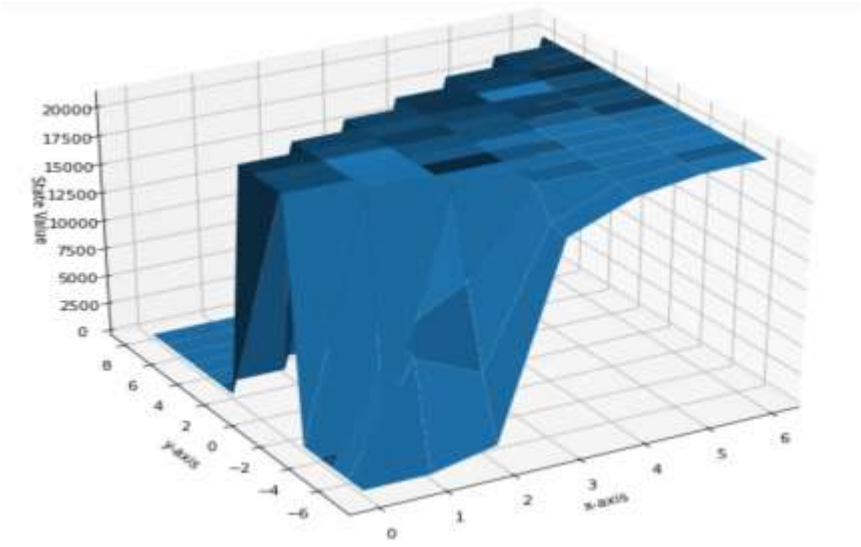
Our Q-learning agent by contrast has learned its policy based on the optimal policy which always chooses the action with the highest Q-value. It is more confident in its ability to dodge the pole.

The total mean reward for sarsa-learning agent is **60.4292**

The total mean reward for Q-learning agent is **80.09748**

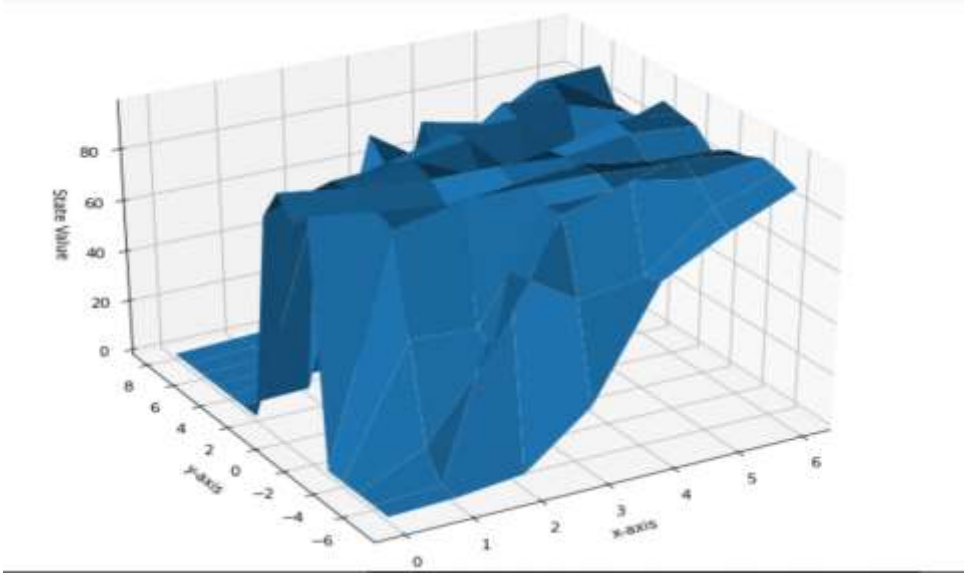
4.Plot State Values

The state value represents the total reward that can be obtained from a state. As we've seen, this is calculated as the sum of all the rewards that will be obtained, starting in the state and then following the policy thereafter.



State Value function using the q-table of Q-learning

i) Sarsa learning



State Value function using the q-table of Sarsa-learning