COMP 767 Reinforcement Learning
Assignment 2

Ali Rezagholizadeh and Mohammad Nikou Sefat

March 18, 2019

1 Question 1 (a).

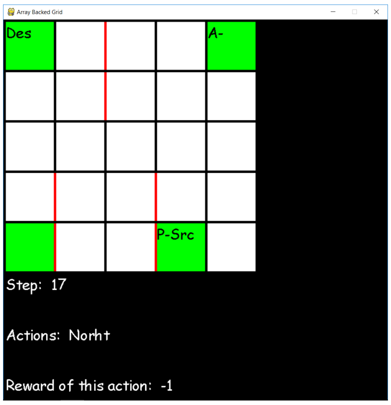
In this question, the aim is to implement taxi delivery question proposed in the *[Dietterich2000]*. There is a 5-by-5 grid world which passenger can be in one of four locations as a source, as in the Figure 1, and request for a taxi to be transported to the destination (one of such four locations). The taxi should find a path to reach and pick up the passenger, then deliver him/her to the destination. So the problem is to find a policy for contributing the taxi (agent) to find its way to do this task.

For solving this problem using Reinforcement Learning, we determined the environment:

- Shape of the environment: As it is shown in the Figure 1, the taxi can move through this 5-by-5 grid world which there are some blocks shown in red lines and four fixed positions in green color.
- Reaction of and feedback from the environment: environment gets a reward of -1 for each movement of the taxi (four directions of Up, Dawn, Left, and Right). This reward is the same when a movement takes the taxi to the out of the grid or to the wall (block), but it can't move. If the taxi correctly picks up the passenger at the source or put down the passenger at the destination, the environment gives +20 reward. otherwise, by picking up or taking off actions in other positions in the grid world would result in getting a reward of -10.

The Agent can perceive its position, passenger's position, the destination's position in which the passenger is going to be put down. In each of these perceived states, the agent (taxi) can take four movement actions and two picks up and put down actions.

The agent is supposed to learn a suitable policy using three RL methods: SARSA, Expected SARSA, and Q_learning. The exploration of all these methods is softmax. The experiment was done with three parameters for



Step: 17

Actions: Norht

Reward of this action: -1

Figure 1: The taxi (agent), passenger, source, and destination is shown with the label of A, P, Src, and Des respectively. The red lines are blocks which the taxi can't pass over it.

learning rate (0.2, 0.5, 0.9) and three parameters for temperature (1, 2, 3), all with discounting of 1.

The result of these experiments are shown in Figure 2. These results obtained after averaging of ten runs, each after doing 99 segments of 10 training episodes.

What is apparent from most of these results, in Figure 2, is that whatever the temperature increases, the average reward increases and whatever the learning rate is increasing the average reward is decreasing.

The result of test after 100 training of such segments, for all methods, are rather similar. That is, about all of them was failed within 4000 steps limitation and just one or two tests in each method were successful in picking up the passenger (the limitation was done because the agent often sticks in a place after some steps of movement and then takes one wrong action again and again, to the wall or to the out of the grid or one of two pick up and put down actions). But these failure happened because of the nature of the method used in explorations, Softmax, and the number of learning episodes.

Softmax tries to explore all actions, so the action value function can roughly converge to the optimal values after large time steps. According to the results of the training, it can be estimated that using a suitable parameter (ex. 0.2 in learning rate and 3 in temperature) and taking more segments, the agent can see more successful tests.

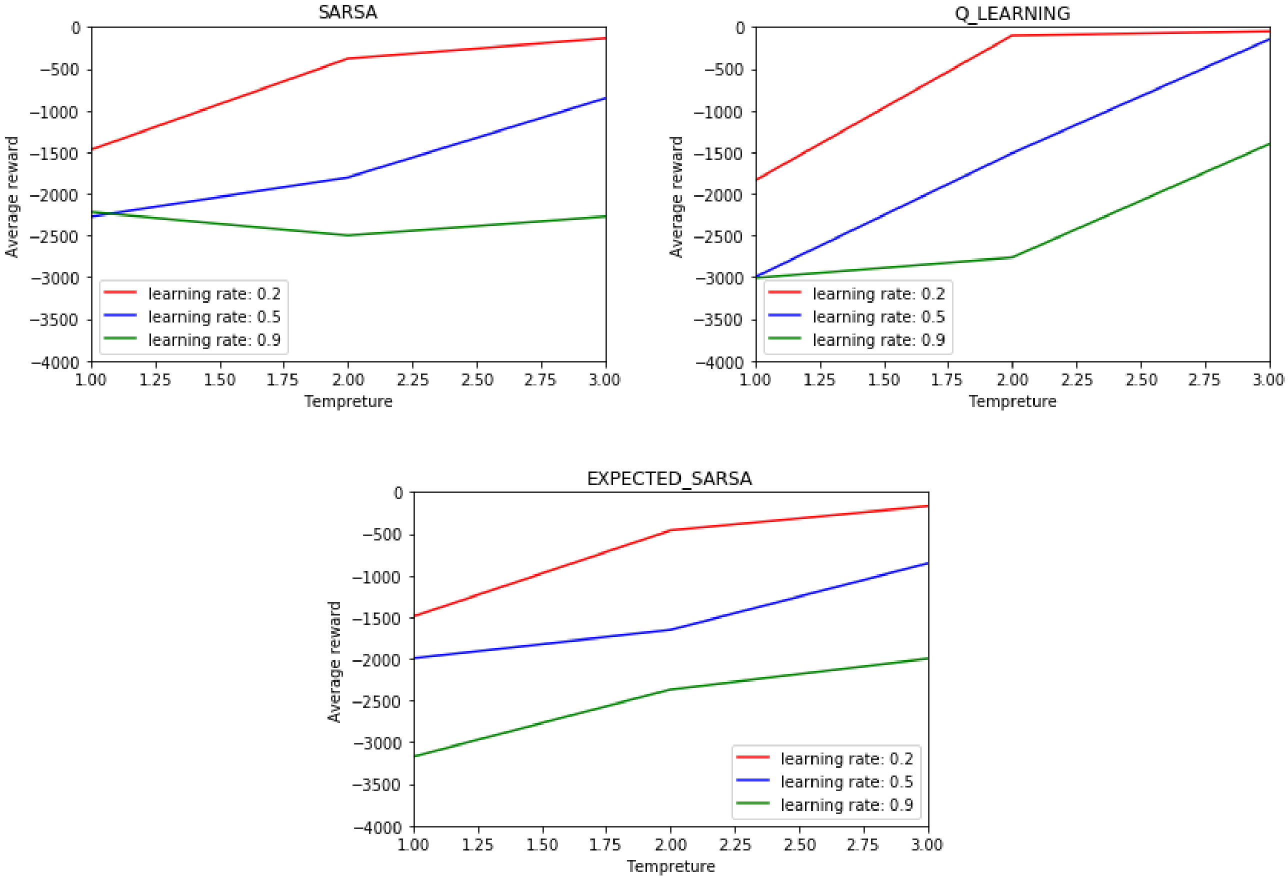


Figure 2: Average reward of training 100 segments of ten episodes using SARSA, Q_learning, and Expected SARSA. These results obtained after averaging of 10 runs.