Task 7.1D: Function approximation implementation

Introduction

The following task is another extension to the environment we have seen from Task 1.1P, but this time it implements two algorithms that incorporate Function Approximation. We implemented these two algorithms on the **'Pendulum-v0'** Environment. Please refer to this link for further details about the environment: Pendulum - Gym Documentation (gymlibrary.dev).

Our report aims to talk in detail about the environment and the results we have received from it while creating the environment and the algorithms from scratch. These algorithms are as follows:

- **Semi-Gradient SARSA(0)**
- **Semi-Gradient TD(Lambda)**

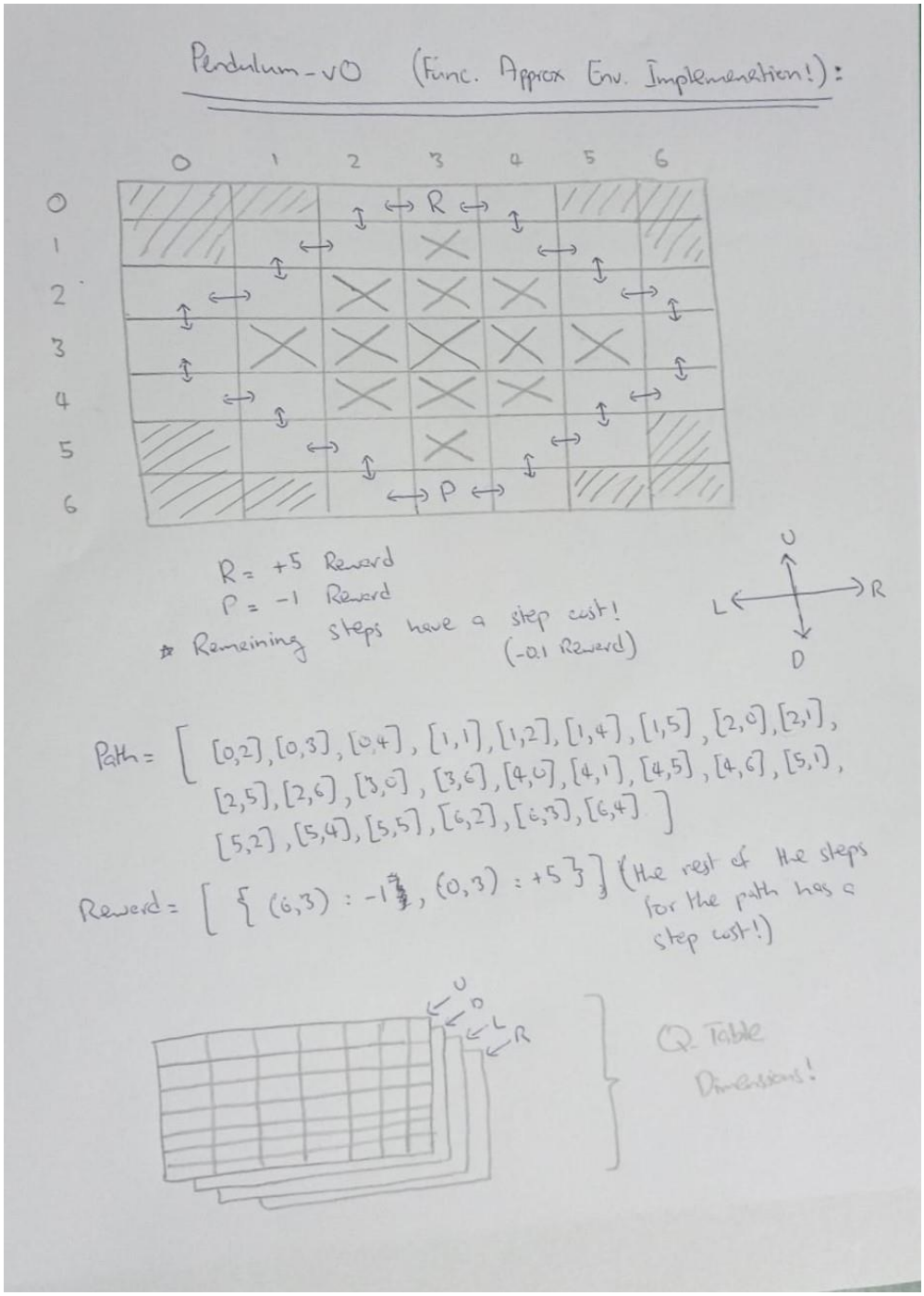
We have also appended our Python Jupyter Notebook with this report. Please refer to the end of this report.

About our Environment

Our Environment is now a resemblance to the Grid World Environment we have working through the workshops in the past couple of weeks. In fact, I have attached an Image and the Jupyter Notebook File of the 'Grid World' Environment in the GitHub Repository mentioned below under this sub-heading. Our observation space comprises a 3D Array this time, where we have the x and y coordinates of the Grid Environment along with the actions to be taken within the Grid. This is shown below in the following picture below.

I have happened to also mention the actions and rewards per state in the figure. Please note three important aspects of the environment. Firstly, the states we have mentioned in the environment below happened to resemble the circular path in which our pendulum moves around, in order to balance itself from the fixed end. They are enlisted as shown in the picture below. Secondly, each state comprises a pair of actions pointing outwards (although the actions I should like a double arrow). So, for example, the actions of the initial state S(6,3) are left (L) and right(R). please also know that I have shown the directions within the picture, in initials. All of the actions for each state are mentioned within the Jupyter Notebook. And thirdly, each state also has a reward associated with it. So for the starting state S(6,3) and the terminal state S(0,3), the rewards are -1 and +5 respectively. And as for the remaining states, they have a step cost of -0.1. Finally, the Q-Table derived for this environment is a 3D environment, where the x, y, and z coordinates are the x and y coordinates of the 7x7 grid along with the number of all possible actions, which are 4. The third coordinate of the environment is exempted from Semi-gradient TD(Lambda).

In order to compute the results, we have taken 200 timesteps per episode. In the later heading, we computed a graph of average rewards received by the agent per episode. We have computed two trends with different colors in order to get the results we need.



Results

The following are some of the results we have taken from our Notebook. We have taken the mean value of the reward after 100 episodes, and the final Q-Table after the 100th episode, to justify the working of our Model. These are how the results look like:

(Semi-Gradient SARSA(0))

```
Q-Table in the 100th Episode:
 [[[ 0.          0.          0.          0.        ]
  [ 0.          0.          0.          0.        ]
  [ 0.          0.          0.          0.        ]
  [ 0.          0.          0.          0.        ]
  [ 0.          0.          0.          0.        ]
  [ 0.          0.          0.          0.        ]
  [ 0.          0.          0.          0.        ]]

 [[ 0.          0.          0.          0.        ]
  [ 0.          0.          0.          0.        ]
  [ 0.          0.          0.          0.        ]
  [ 0.          0.          0.          0.        ]
  [ 0.          0.          0.          0.        ]
  [ 0.          0.          0.          0.        ]
  [ 0.          0.          0.          0.        ]]

 [[ 0.          0.          0.          0.        ]
  [ 0.          0.          0.          0.        ]
  [ 0.          0.          0.          0.        ]
  [ 0.          0.          0.          0.        ]
  [ 0.          0.          0.          0.        ]
  [ 0.          0.          0.          0.        ]
  [ 0.         -0.06894811  0.          0.        ]]

 [[ 0.          0.          0.          0.        ]
  [ 0.          0.          0.          0.        ]
  [ 0.          0.          0.          0.        ]
  [ 0.          0.          0.          0.        ]
  [ 0.          0.          0.          0.        ]
  [ 0.          0.          0.          0.        ]
  [-0.02053715 -0.51062567 -0.04247554 -0.07685514]]

 [[ 0.          0.          0.          0.        ]
  [ 0.          0.          0.          0.        ]
  [ 0.          0.          0.          0.        ]
  [ 0.          0.          0.          0.        ]
  [ 0.          0.          0.          0.        ]
  [-0.01015181 -0.23219231  0.         -0.07250632]
  [-0.54116897  0.         -0.01811211  0.        ]]

 [[ 0.          0.          0.          0.        ]
  [ 0.          0.          0.          0.        ]
  [ 0.          0.          0.          0.        ]
  [ 0.          0.          0.          0.        ]
  [ 0.          0.          0.         -0.01      ]
  [-0.21442676 -0.00611702  0.          0.        ]
  [ 0.          0.          0.          0.        ]]

 [[ 0.          0.          0.          0.        ]
  [ 0.          0.          0.          0.        ]
  [ 0.          0.          0.          0.        ]
  [ 0.          0.          0.         -0.01      ]
  [-0.01        0.          0.          0.        ]
  [ 0.          0.          0.          0.        ]
  [ 0.          0.          0.          0.        ]]]
```

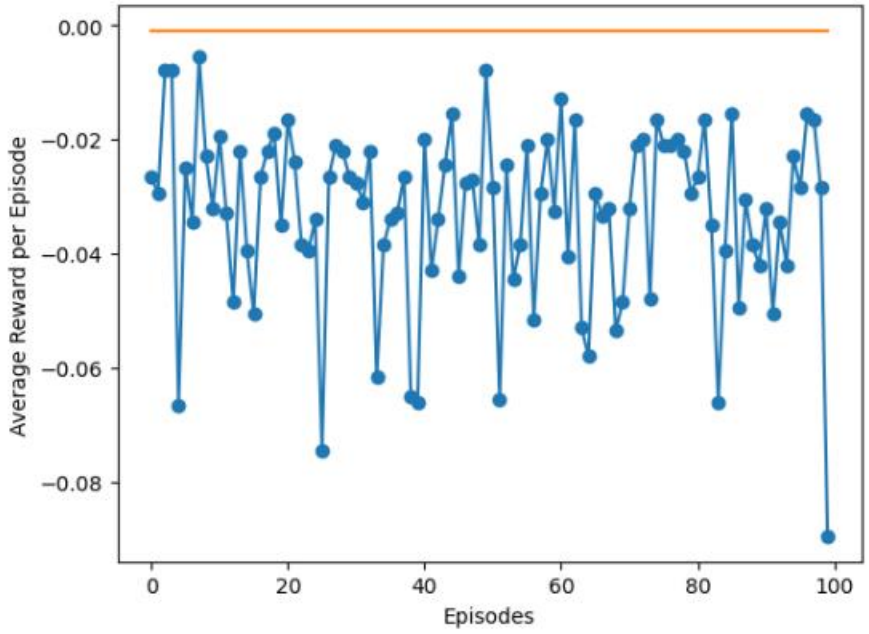
```
Episode --- [50/100]  | Reward ---- -1.6
Episode --- [51/100]  | Reward ---- -5.699999999999998
Episode --- [52/100]  | Reward ---- -13.09999999999997
Episode --- [53/100]  | Reward ---- -4.8999999999999995
Episode --- [54/100]  | Reward ---- -8.899999999999986
Episode --- [55/100]  | Reward ---- -7.6999999999999975
Episode --- [56/100]  | Reward ---- -4.199999999999999
Episode --- [57/100]  | Reward ---- -10.299999999999997
Episode --- [58/100]  | Reward ---- -5.899999999999998
Episode --- [59/100]  | Reward ---- -4.0
Episode --- [60/100]  | Reward ---- -6.499999999999998
Episode --- [61/100]  | Reward ---- -2.6
Episode --- [62/100]  | Reward ---- -8.099999999999998
Episode --- [63/100]  | Reward ---- -3.3000000000000003
Episode --- [64/100]  | Reward ---- -10.599999999999978
Episode --- [65/100]  | Reward ---- -11.599999999999996
Episode --- [66/100]  | Reward ---- -5.899999999999999
Episode --- [67/100]  | Reward ---- -6.699999999999999
Episode --- [68/100]  | Reward ---- -6.399999999999999
Episode --- [69/100]  | Reward ---- -10.699999999999998
Episode --- [70/100]  | Reward ---- -9.699999999999998
Episode --- [71/100]  | Reward ---- -6.399999999999998
Episode --- [72/100]  | Reward ---- -4.2
Episode --- [73/100]  | Reward ---- -4.0
Episode --- [74/100]  | Reward ---- -9.599999999999996
Episode --- [75/100]  | Reward ---- -3.300000000000001
Episode --- [76/100]  | Reward ---- -4.2
Episode --- [77/100]  | Reward ---- -4.2
Episode --- [78/100]  | Reward ---- -4.000000000000001
Episode --- [79/100]  | Reward ---- -4.4
Episode --- [80/100]  | Reward ---- -5.899999999999999
Episode --- [81/100]  | Reward ---- -5.300000000000001
Episode --- [82/100]  | Reward ---- -3.3000000000000007
Episode --- [83/100]  | Reward ---- -6.999999999999998
Episode --- [84/100]  | Reward ---- -13.199999999999974
Episode --- [85/100]  | Reward ---- -7.899999999999998
Episode --- [86/100]  | Reward ---- -3.100000000000001
Episode --- [87/100]  | Reward ---- -9.899999999999997
Episode --- [88/100]  | Reward ---- -6.1
Episode --- [89/100]  | Reward ---- -7.6999999999999975
Episode --- [90/100]  | Reward ---- -8.399999999999997
Episode --- [91/100]  | Reward ---- -6.399999999999999
Episode --- [92/100]  | Reward ---- -10.099999999999998
Episode --- [93/100]  | Reward ---- -6.8999999999999995
Episode --- [94/100]  | Reward ---- -8.399999999999988
Episode --- [95/100]  | Reward ---- -4.6
Episode --- [96/100]  | Reward ---- -5.6999999999999975
Episode --- [97/100]  | Reward ---- -3.1000000000000005
Episode --- [98/100]  | Reward ---- -3.3000000000000003
Episode --- [99/100]  | Reward ---- -5.6999999999999975
Episode --- [100/100]  | Reward ---- -17.899999999999984
Average Reward after 100 Episodes:  -6.577999999999998
```

(Semi-Gradient TD(Lambda))

```
Q-Table in the 100th Episode:
 [[[ 0.     0.     0.     0.    ]
  [ 0.     0.     0.     0.    ]
  [ 0.     0.     0.     0.    ]
  [ 0.     0.     0.     0.    ]
  [ 0.     0.     0.     0.    ]
  [ 0.     0.     0.     0.    ]
  [ 0.     0.     0.     0.    ]]

 [[ 0.     0.     0.     0.    ]
  [ 0.     0.     0.     0.    ]
  [ 0.     0.     0.     0.    ]
  [ 0.     0.     0.     0.    ]
  [ 0.     0.     0.     0.    ]
  [ 0.     0.     0.     0.    ]
  [ 0.     0.     0.     0.    ]]

 [[ 0.     0.     0.     0.    ]
  [ 0.     0.     0.     0.    ]
  [ 0.     0.     0.     0.    ]
  [ 0.     0.     0.     0.    ]
  [ 0.     0.     0.     0.    ]
  [ 0.     0.     0.     0.    ]
  [ 0.     0.     0.     0.    ]]

 [[ 0.     0.     0.     0.    ]
  [ 0.     0.     0.     0.    ]
  [ 0.     0.     0.     0.    ]
  [ 0.     0.     0.     0.    ]
  [ 0.     0.     0.     0.    ]
  [ 0.     0.     0.     0.    ]
  [ 0.     0.     0.     0.    ]]

 [[ 0.     0.     0.     0.    ]
  [ 0.     0.     0.     0.    ]
  [ 0.     0.     0.     0.    ]
  [ 0.     0.     0.     0.    ]
  [ 0.     0.     0.     0.    ]
  [ 0.     0.     0.     0.    ]
  [ 0.     0.     0.     0.    ]]

 [[ 0.     0.     0.     0.    ]
  [ 0.     0.     0.     0.    ]
  [ 0.     0.     0.     0.    ]
  [ 0.     0.     0.     0.    ]
  [ 0.     0.     0.     0.    ]
  [ 0.     0.     0.     0.    ]
  [ 0.     0.     0.     0.    ]]

 [[ 0.     0.     0.     0.    ]
  [ 0.     0.     0.     0.    ]
  [ 0.     0.     0.     0.    ]
  [ 0.     0.     0.    -0.015]
  [-0.01   0.     0.     0.    ]
  [ 0.     0.     0.     0.    ]
  [ 0.     0.     0.     0.    ]]]
```

```
Episode --- [49/100]  | Reward ---- -0.2
Episode --- [50/100]  | Reward ---- -0.2
Episode --- [51/100]  | Reward ---- -0.2
Episode --- [52/100]  | Reward ---- -0.2
Episode --- [53/100]  | Reward ---- -0.2
Episode --- [54/100]  | Reward ---- -0.2
Episode --- [55/100]  | Reward ---- -0.2
Episode --- [56/100]  | Reward ---- -0.2
Episode --- [57/100]  | Reward ---- -0.2
Episode --- [58/100]  | Reward ---- -0.2
Episode --- [59/100]  | Reward ---- -0.2
Episode --- [60/100]  | Reward ---- -0.2
Episode --- [61/100]  | Reward ---- -0.2
Episode --- [62/100]  | Reward ---- -0.2
Episode --- [63/100]  | Reward ---- -0.2
Episode --- [64/100]  | Reward ---- -0.2
Episode --- [65/100]  | Reward ---- -0.2
Episode --- [66/100]  | Reward ---- -0.2
Episode --- [67/100]  | Reward ---- -0.2
Episode --- [68/100]  | Reward ---- -0.2
Episode --- [69/100]  | Reward ---- -0.2
Episode --- [70/100]  | Reward ---- -0.2
Episode --- [71/100]  | Reward ---- -0.2
Episode --- [72/100]  | Reward ---- -0.2
Episode --- [73/100]  | Reward ---- -0.2
Episode --- [74/100]  | Reward ---- -0.2
Episode --- [75/100]  | Reward ---- -0.2
Episode --- [76/100]  | Reward ---- -0.2
Episode --- [77/100]  | Reward ---- -0.2
Episode --- [78/100]  | Reward ---- -0.2
Episode --- [79/100]  | Reward ---- -0.2
Episode --- [80/100]  | Reward ---- -0.2
Episode --- [81/100]  | Reward ---- -0.2
Episode --- [82/100]  | Reward ---- -0.2
Episode --- [83/100]  | Reward ---- -0.2
Episode --- [84/100]  | Reward ---- -0.2
Episode --- [85/100]  | Reward ---- -0.2
Episode --- [86/100]  | Reward ---- -0.2
Episode --- [87/100]  | Reward ---- -0.2
Episode --- [88/100]  | Reward ---- -0.2
Episode --- [89/100]  | Reward ---- -0.2
Episode --- [90/100]  | Reward ---- -0.2
Episode --- [91/100]  | Reward ---- -0.2
Episode --- [92/100]  | Reward ---- -0.2
Episode --- [93/100]  | Reward ---- -0.2
Episode --- [94/100]  | Reward ---- -0.2
Episode --- [95/100]  | Reward ---- -0.2
Episode --- [96/100]  | Reward ---- -0.2
Episode --- [97/100]  | Reward ---- -0.2
Episode --- [98/100]  | Reward ---- -0.2
Episode --- [99/100]  | Reward ---- -0.2
Episode --- [100/100]  | Reward ---- -0.2
Average Reward after 100 Episodes:  -0.19999999999999996
```

(Semi-Gradient SARSA(0) vs Semi-Gradient TD(Lambda))



Our final results apparently show for now that the Semi-gradient TD(Lambda) outperforms its counterpart, given that it overcomes the effect of the step cost from the circular path in the environment. The average rewards from the second algorithm always seem to be constant compared to the former, whose average rewards happen to be almost random and sporadic by nature.

References

- https://www.gymlibrary.dev/environments/classic_control/
- https://www.gymlibrary.dev/environments/classic_control/pendulum/
- https://numpy.org/doc/stable/reference/
- https://www.learndatasci.com/tutorials/reinforcement-q-learning-scratch-python-openai-gym/
- Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction. MIT Press. Semi-Gradient SARSA:
- https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf (p. 152-154)
- Class Slides (Week 7, Week07_01_Function_Approximation1, Slides 9,14)