Reinforcement Learning (RBE595)

Programming Assignment 3:

Monte Carlo

02/19/2023

By: Swapneel Dhananjay Wagholikar (WPI ID: 257598983)

Bhaavin Jogeshwar (WPI ID: 888602054)

Chinmayee Prabhakar (WPI ID: 101177306)

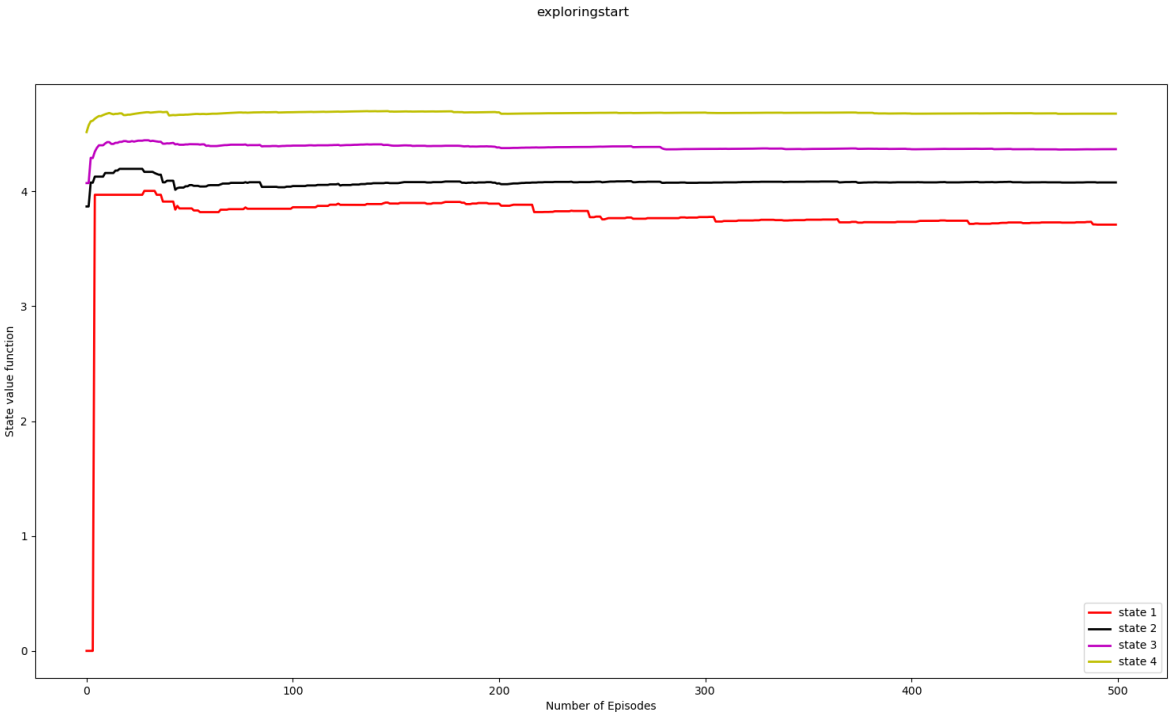
We have implemented the on-policy first visit Monte-Carlo and exploring starts policy in this assignment. The first action is decided by the policy which is then given to the stochastic transition dynamics. We calculate the rewards of each of the states per episode and obtain the action-value function and the state-value function. Using the action-value function, we update the policy to eventually take the best action possible. This has been run for 500 episodes. The results obtained are as follows:

Outputs:

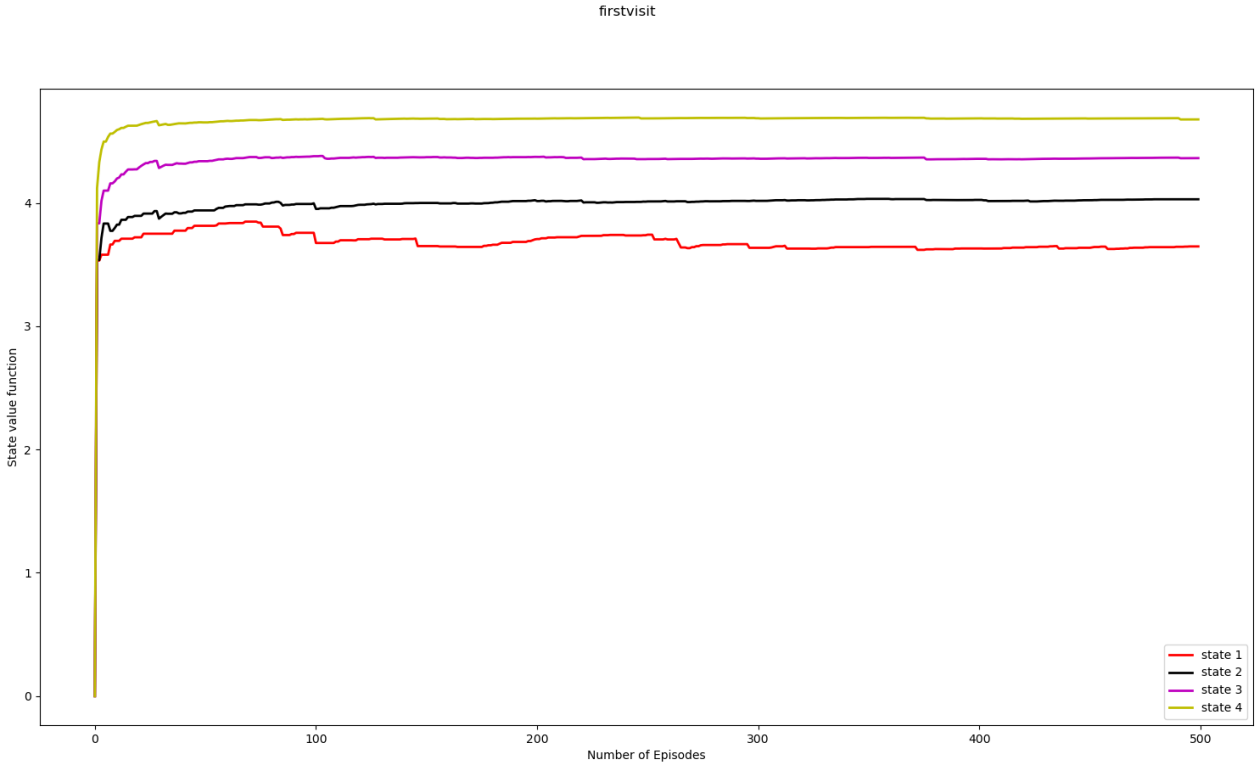
The optimal action value functions and optimal policies using the Exploring starts and First visit Monte-Carlo:

```
(base) swapneel@swapneel:~/rbe595/prog_assi_3$  cd /home/swapneel/rbe595/prog_assi_3 ; /usr/bin/env /bin/python3 /home/swapneel/.vscode/extensions/ms-python.
python-2023.2.0/pythonFiles/lib/python/debugpy/adapter/../../debugpy/launcher 42475 -- /home/swapneel/rbe595/prog_assi_3/montecarlo_main.py
Exploring Starts

optimal action value function:
{(0, 'T'): 1.0, (1, 'L'): 1.0971000235787456, (1, 'R'): 0, (2, 'L'): 0, (2, 'R'): 3.951684881802266, (3, 'L'): 0, (3, 'R'): 4.3404052300376375, (4, 'L'): 0,
(4, 'R'): 4.631602677943441, (5, 'T'): 5.0}

optimal policy:
[[0, 'T'], [1, 'L'], [2, 'R'], [3, 'R'], [4, 'R'], [5, 'T']]

First Visit

optimal action value function:
{(0, 'T'): 1.0, (1, 'L'): 0.9342623761792461, (1, 'R'): 3.4795165790538634, (2, 'L'): 1.1901550609240417, (2, 'R'): 3.9609757341097267, (3, 'L'): 3.647510156
2500005, (3, 'R'): 4.246015150429247, (4, 'L'): 3.3575073587955617, (4, 'R'): 4.636688191878259, (5, 'T'): 5.0}

optimal policy:
[[0, 'T'], [1, 'R'], [2, 'R'], [3, 'R'], [4, 'R'], [5, 'T']]
```

The estimate of state-value function for each state as a function of number of episodes in Monte-Carlo with exploring start:



The estimate of state-value function for each state as a function of number of episodes in On-policy first visit Monte-Carlo:



Convergence:

By simulating multiple episodes of an agent interacting with an environment and then averaging the returns for each state across all the episodes, the first-visit Monte Carlo method can estimate the state value function. The state values converge to their true values as the number of episodes increases, improving the estimates of the state values. The law of large numbers guarantees that the state value function will converge under the first-visit Monte Carlo policy. The sample mean (estimated state value) will approach the true mean (actual state value) of the distribution from which the samples are taken as the number of episodes rises.

In the first-visit Monte Carlo policy, the state value function convergence can be slower, particularly for large state spaces. This is due to the fact that the accuracy of the state value estimations directly relates to how frequently each state is visited, and certain states may only be visited infrequently.

In contrast to the first-visit Monte Carlo approach, the method guarantees that all state-action pairs are visited with a non-zero probability, which could result in a quicker convergence of the state value function.