Reinforcement Learning - Homework 7
Vishal I B

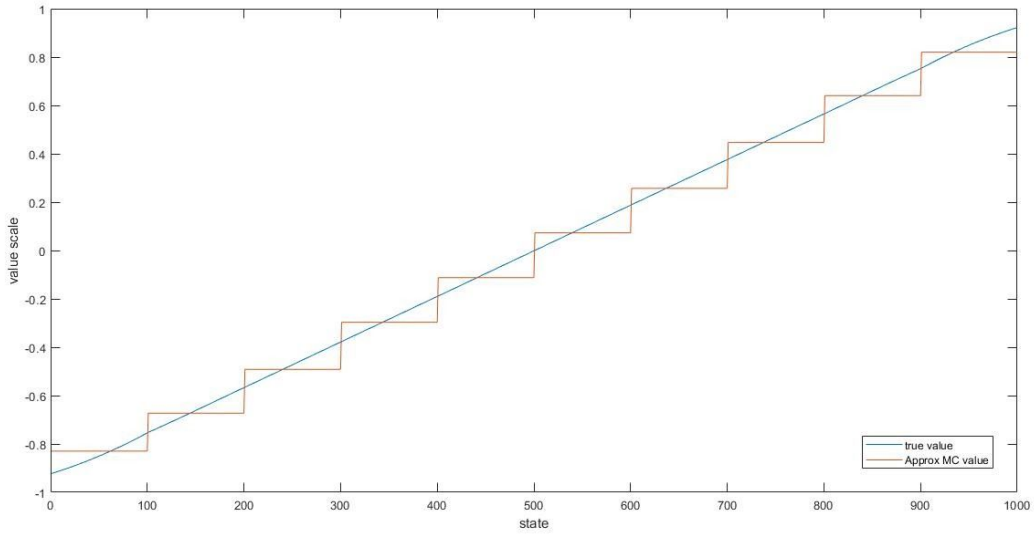
The states are numbered from 1 to 1002 with 1 as the terminal state on the left and 1002 as the terminal state on the right. The rewards are distributed such that the reward in every state is zero except for the terminal states. The reward for the terminal state on the left is -1, while for the one on the right is 1.

For the policy evaluation to compute the true value, the probability matrix was created in such a way that the terminal state received the probability of the missing states, when the number of states either on the left or the right less than 100. Vpi is calculated.

For the Gradient Monte-Carlo method, the episodes are generated using the probability matrix that was created, choosing the next step based on the probabilities. The approximate value, which is equal to the dot product of **W** and **X**, where W and X are the weights and the features. Here, I have chosen the feature to be linear and equal to one for every state. Hence, the approximated value is equal to W. The weights are aggregated based on the states, that is, each state has the same value for W with the group.

The implementation of TD(0) was similar to gradient MC, but without the episode generation. It has bootstrapping values, unlike MC. Both the algorithms were run for 100000 episodes and the graphs were plotted.

Graph 1: Gradient MC



Graph 2: Gradient MC and semi-gradient TD(0)