Machine Learning and Neural Computation
Assessed Coursework

Beatriz Isabel Lopez Andrade (bil14)

November 5, 2014

Section A

Question 1

The function `GetTrace` generates a random episode from a given MDP. Each row in the output has three columns, which represents the reward for a transition from the previous state to the current one having taken a particular action, the current state, and the action taken in the current state to move to the following one.

In general, the function `GetTrace` generates a random sample of the current state taken into account the previous state and the chosen action in that state. After that, it gets the reward corresponding to that transition and action. If the current state is not an absorbing one, the function generates a random sample of the action in that state. However, if the current state is an absorbing one, none further actions are taken and therefore, the trace ends. The row corresponding to this state is formed by the reward, the state itself, and the action taken in the current state.

Reward	Status	Action
1	1	L
2	2	R

Table 1: Trace corresponding to the Stair Climbing MDP.

For the particular trace of the **Stair Climbing MDP** (Table 1), the first state is four, because the function `StairClimbingMDP` assigns a probability of one to this state as the first state. The action in this state is ?Left?, but it could have been ?Right?, as the unbiased policy is used. In the following row, the state is three, since the probability of going from state four to three having taken action ?Left? is one. The reward for this transition is one and the next action ?Right?, but it could have been ?Left?. The following rows are computed in the same way, until an absorbing state is reached.

In the case of the first row, there is no transition to get to the current state, i.e. it is the first state. Hence, there is no reward for getting to that particular state. That is the reason why a dummy value is assigned to the first reward.

In an absorbing state, the state where the trace ends, no further actions are taken. Therefore, the value for the last action is a dummy one.

Question 2

In the case of `GridWorld1`, the number of traces needed to get an accurate estimation of the state-value function is about 16250. With this number of traces, the maximum difference between the values of the state-value function between two consecutive iterations is less than 0.001, and therefore, convergence is assumed. The number of batches needed until converge to the optimal policy is 5.