AID 554: Reinforcement Learning

EndSem Exam - Spring 2022

Instructor:	Manu K. Gupta	**Due date:**	2nd May
Email:	manu.gupta@mfs.iitr.ac.in		End of the day

Note: Provide justifications/calculations/steps along with each answer to illustrate how you arrived at the answer. State your arguments clearly, in logical sequence. You will not receive credit for giving an answer without sufficient explanation.

Submission: You can either write down your answer by hand or type it in. Be sure to mention your name and roll number. Upload to Moodle as a **pdf** file along with the your python codes.

Questions

Q1: **Value function prediction problem (20 points)**

In this question, you will implement an algorithm for estimating the value function of a policy for a given MDP from a trajectory of the form state, action, reward, state, action, reward, ...

```
Number of states
Number of actions
Discount factor
state1 action1 reward1
state2 action2 reward2
state3 action3 reward3
.
.
.
stateN actionN rewardN
stateN+1
```

The number of states S and the number of actions A will be integers greater than 0. Assume that the states are numbered $0, 1, ..., S-1$, and the actions numbered $0, 1, ..., A-1$. The discount factor will lie between 0 and 1. The trajectory over time will be long enough, and the dynamics of the underlying MDP such, that there is at least one outgoing transition from each state in the MDP. Note that the MDP is not episodic: that is, `stateN+1` is not a terminal state (and can occur within the trajectory multiple times). The trajectory is merely a finite sequence generated according to the underlying transition and reward functions, and terminated at some arbitrary time step.

You can assume that S and A will not exceed 50, and N, the total number of transitions in the trajectory, will not exceed 500,000. In data directory, you will find two sample data files (`d1.txt` and `d2.txt`).

Your evaluator must estimate the value function V under the policy being followed. The output, written to standard output, must be in the following format (`Est-V` is your estimate of V).

```
Est-V(0)
Est-V(1)
.
.
.
Est-V(S - 1)
```

In the data directory enclosed, you will find output files corresponding to the two data files, which have solutions in the format above. The values mentioned in these output files are indeed the true values (under the same policy) from the MDP being sampled. Naturally, as you will have to estimate values based on samples alone, your estimates cannot be expected to match the true values perfectly.

Notice that since this is a prediction problem, wherein a fixed policy is being followed, the actual names of the actions taken do not matter. Nor does it matter if the policy being followed is deterministic or stochastic. Your logic only needs to consider the state, reward, and next state associated with each transition.

`Question:` Generate trajectories from different MDPs (of your choice) and policies. Your task is to ensure that it prints out a good estimate of the true value function in each case. Performance will be quantified based on the (unweighted) squared distance between your estimate `Est-V` and the true value function V: that is,

$$Error = \sum_{s \in S} (V(s) - \hat{V}(s))^2$$

where $\hat{V}(s)$ is `Est-V(s)`. Report the performance for several instances. Be sure to describe your approach and explain why you chose it over alternative approaches.

Note: *It is okay to use libraries for data structures and for operations such as sorting. However, the logic used for value prediction must entirely be code that you have written.*

Q2: **DQN algorithm for reinforcement learning (15 points)**

This is an open ended question on Deep Q-network (DQN) which is one of the popular architecture for modern reinforcement learning.

(a) Use internet resources to understand DQN and write one page summary of DQN in your own words.

(b) Implement basic DQN for a problem of your choice. Please be sure to present the full description of the problem context and your implementation.