CMPUT 609 Assignment 7

Shaurya Seth

March 28, 2023

Question 1

Not attempted

Question 2

In this case, π would be the question "how to move towards an object", γ would be "bumper being activated", y would be the "distance from object" and z would be "distance at object" which would be 0.

Question 3

In this case, π would be the question "how to move towards the charging station", γ would be "charging has begun", y would be the "probability of reaching charging station" and z would be "probability of having reached charging station" which would be 1.

Question 4

Dyna would probably do better than multi-step bootstrapping because the latter can only update values in the current trajectory while the former can update other states that are present in the model but were not visited during the current episode.

Question 5

In the first phase, the Dyna-Q+ agent appears to do better (higher cumulative reward) because the Dyna-Q+ algorithm gives it additional rewards to incentivize exploration. Whether or not it is actually doing better or worse is not clear because of this. My guess would be that it's doing slightly worse because of the need to explore.

In the second phase, the Dyna-Q+ agent actually does better as it move a lot higher than Dyna-Q because it is able to find the shortcut because of the exploration incentive. Meanwhile the Dyna-Q agent does not even notice the shortcut.

Question 6

The difference narrowed slightly because Dyna-Q found the optimal policy for the first phase and was able to exploit it while the Dyna-Q+ had some exploring to do along the way.

Question 7

We would have to use a stochastic model. We would need to figure out the probability distribution of the next reward and state and use that to update the model in part e) of the algorithm.

This could perform poorly in a changing environment because the change is deterministic (the deterministic model would learn it in first go) whereas the stochastic model would have to gradually update the expectation.

We could modify the algorithm to be more sensitive to recent changes and add an exploration incentive like in Dyna-Q+.

Question 8

A highly skewed distribution would strengthen the case for sample updates since the more likely states would be sampled more often. Trying to get the expectation would be a waste of computation since the sampled states will already be close to the expected states.

Question 9

Not attempted

Question 10

The GVF makes predictions or forecasts on signals while γ sets the horizon for the predictions. It might make sense to make the termination a function of earlier states, actions and rewards if the same signal might have a different meaning based on the context the agent is in.

For example, if receiving signal from a thermostat, while entering and exiting a room the agent would perceive the same temperature but the horizon of the predictions would be different.

However, if the MDP we are dealing with is Markov, it is sufficient for the termination function to just be a function of the current state.

Question 11

a) The environment can be in four states: up, down, right and left. But there are only two observations: 0 and 1. This makes it partially observable.

b) No, because if the history ends with an observation of 1, the action-conditional futures have different probability distributions.

c) The six other equivalence states are:

$$*0c1$$
$$*0a1$$
$$*0c1c1$$
$$*0a1a1$$
$$*1c1c1$$
$$*1a1a1$$

Question 12

There are four environment states: up, down, right and left. There are no information states.

Question 13

There are $4 \times 3 = 12$ environment states. There are six information states:

```
∗0c1c1
∗0a1a1
∗1a1a0
∗1c1c0
  ∗0f1
∗0f0f0
```

Question 14

There are two information states:

```
∗1check1
∗2check2
```