COMP 690 Human-Centric ML Assignment 2

Li Chen

February 28, 2024

Question 1.

The learning rate is set to be 0.01. Since we are dealing with a small question with limited demos and state spaces, a wide range of learning rate selections can perform well. A learning rate of 0.01 is appropriate for this task.

Question 2.

The shape of reward function is similar to a funnel type, where the states on the edges of the state spaces has have higher reward, while the states in the central have lower rewards. The graph reveals that the policy encourages the agent to reach the goal via passing through the states on the edges. The state 12 is sensitive and might shows a wrong reward because the state 12 is actually on one optimal path to the goal state.

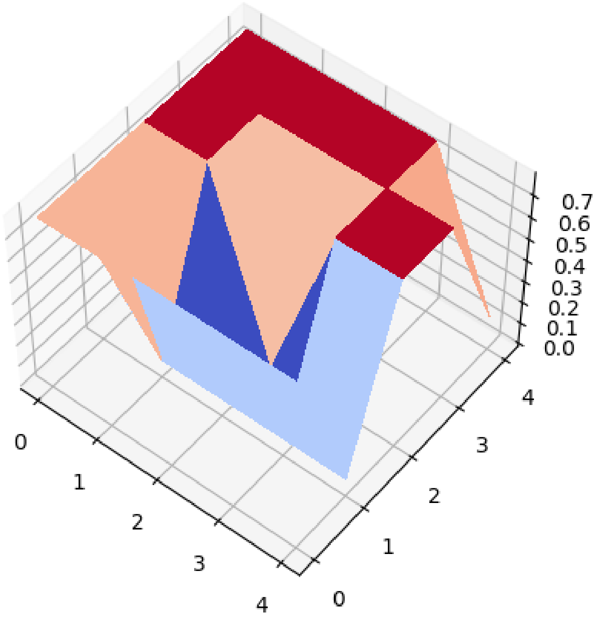

Figure 1: Graph 1 of reward function

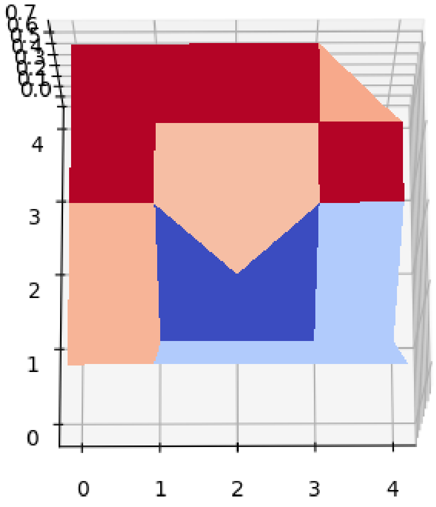

Figure 2: Graph 2 of reward function

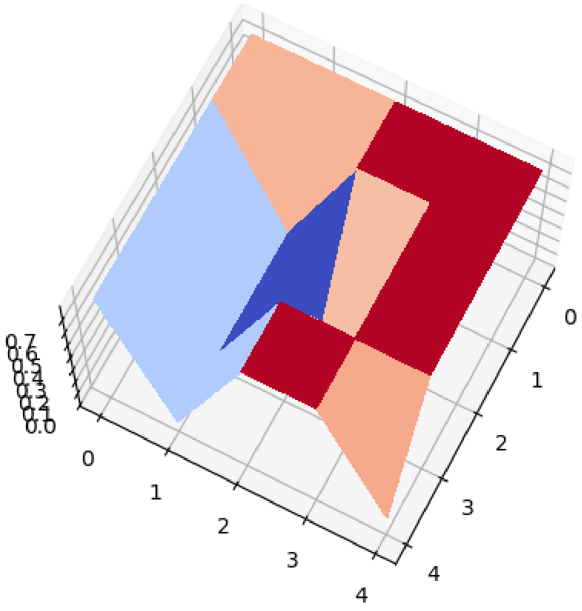

Figure 3: Graph 3 of reward function

Question 3.

We write the loss function $L(\theta)$ as

$$L(\theta) = \log P(\tau|\theta) = \log \left[\frac{e^{\theta^T f_\tau}}{Z(\theta)} \right] = \log \left[\frac{e^{\theta^T f_\tau}}{\sum_\zeta e^{\theta^T f_\zeta}} \right]. \tag{1}$$

We then calculate the gradient of Eq.1 with the respect of θ.

$$
\begin{aligned}
\nabla L(\theta) &= \frac{\partial L(\theta)}{\partial \theta} \\
&= \frac{\partial}{\partial \theta}\left(\theta^T f_\tau - \log \sum_\zeta e^{\theta^T f_\zeta}\right) \\
&= f_\tau - \frac{\sum_\zeta \left(e^{\theta^T f_\zeta} \cdot f_\zeta\right)}{\sum_\zeta e^{\theta^T f_\zeta}} \\
&= f_\tau - \sum_\zeta \frac{e^{\theta^T f_\zeta}}{\sum_{\zeta'} e^{\theta^T f_{\zeta'}}} f_\zeta \\
&= f_\tau - \sum_\zeta P(\zeta|\theta)\mathbf{f}_\zeta \\
&= \tilde{\mathbf{f}} - \sum_\zeta P(\zeta|\theta)\mathbf{f}_\zeta \\
&= \tilde{\mathbf{f}} - \sum_{s_i} D_{s_i}\mathbf{f}_{s_i}
\end{aligned}
\tag{2}
$$

Question 4.

The action-based distribution only weights the paths locally at the action level. The actions of branching at earlier paths are not taken into consideration. The policy generated from this weighting will prefer the path with less branching factor. Meanwhile, MaxEnt policy weights the paths based on the trajectories, which involves all the actions along the trajectories.

Question 5.

Z_{s_i} represents state partition function. It means, given a set of parameter θ, we calculate the probability of observing a state or a trajectory. Z_{s_i} is a normalization factor ensures that the sum of probabilities of all states/trajectories equals to 1.

$Z_{a_{i,j}}$ represents state-action partition function. Given a set of parameters θ, we consider the probabilities of both states and actions. $Z_{a_{i,j}}$ is a normalization factor here.

Question 6.

Adding 1 to the terminal state of Z_s can ensure that the terminal state can contribute positively to the partition function, which means the terminal state can indeed end the trajectory. Also, if the terminal state is set to be 0, by the implementation, the zero will be backward passed to other states and therefore generate a invalid policy.