EECE 5614: Reinforcement Learning

Problem 1:

Given:

𝑝 = 0.025, 𝛾 = 0.96, 𝛼 = 0.25, 𝜖 = 0.1, the total number of episodes = 1000, the maximum length of each episode = 1000, 𝛽 = 0.05

Goal:

To analyze the performance of the temporal difference learning algorithms:

1. Q-Learning
2. SARSA
3. Tabular Actor-Critic

Analysis:

1. Q-Learning

10 out of the 10 independent runs agent reached the goal state.

Optimal Policy:

A game of a crossword

AI-generated content may be incorrect.

Optimal Path:

A game of a maze

AI-generated content may be incorrect.

Average Accumulated Rewards over the 1000 episodes for the 10 independent runs:

A graph showing a number of people

AI-generated content may be incorrect.

From the above plot, it can be observed that during initial episode numbers, say, from 0-150, there are larger negative spikes indicating that when the agent starts to train, it often ends up passing through bump or oil states or hitting walls, leading to significant penalties. But it can also be observed that during this same time, the average rewards quickly rise, indicating that the agent is beginning to learn a better policy, with lesser and lesser penalties to reach the goal state. Past the 180–200-episode number, the average accumulated reward curve plateaus around a higher positive reward, converging to an optimal policy to navigate the maze.

The plot, hence, supports the theoretical explanation for Q-learning algorithm, an off-policy method which tends to find the optimal path to the goal state given enough exploration and suitable learning rate.

1. SARSA

9 out of the 10 independent runs agent reached the goal state.

Optimal Policy:

A game with a maze

AI-generated content may be incorrect.

Optimal Path:

A maze with different colored lines

AI-generated content may be incorrect.

Average Accumulated Rewards over the 1000 episodes for the 10 independent runs:

A graph showing a number of episodes

AI-generated content may be incorrect.

It can be observed that the SARSA agent learns very similar to the Q-learning agent. Initial negative reward spikes, gradual rise in the average rewards and convergence to optimal policy to navigate through the maze while the parameters set are comparable.

1. Tabular Actor -Critic

2 out of the 10 independent runs agent reached the goal state.

Optimal Policy:

A game with arrows and different colored squares

AI-generated content may be incorrect.

Optimal Path:

A game of a maze

AI-generated content may be incorrect.

Average Accumulated Rewards over the 1000 episodes for the 10 independent runs:

A screen shot of a graph

AI-generated content may be incorrect.

Unlike the Q-learning agent and the SARSA agent, the Actor-critic agent learning fluctuates a lot and has a lot of sudden spikes and comparatively remain negative. Out of the 10 independent runs, only two of them make it to the goal state, indicating that the gradient updates are quite noisy. This can mean that the algorithm is pretty sensitive to the set parameters like the step size (𝛼) and the policy update rate (𝛽), hence, fine-tuning them is very important. As learnt in the class, the actor-critic method updates the agent’s policy via softmax over preferences. So the policy can change in unstable ways if the learning rate is not well-balanced.

Average Accumulated Rewards w.r.t. the episode number over all the three algorithms tested above:

A graph of a graph

AI-generated content may be incorrect.

With the current parameters, both Q-learning and SARSA handle this maze environment effectively, quickly learning to navigate to the goal and maintaining high reward per episode, while the Actor-Critic approach remains unstable.

Problem 2:

Given:

p53-MDM2 network

Goal:

Implement Q-Learning, SARSA, SARSA−𝝀 (𝜆 = 0.95) and tabular actor-critic (𝛽 = 0.05) algorithms. For each algorithm, you need to run the algorithm for 10 independent runs.

1. Show the optimal policy for all 10 independent runs (10 vectors of size 16, consisting of 𝒂1, 𝒂2 or 𝒂3)
2. Show the average accumulated reward (in 10 independent runs) with respect to the episode number.
3. Plot the average accumulated reward with respect to the episode number for all algorithms in a single plot. Describe your findings.

Analysis:

A table of numbers and words

AI-generated content may be incorrect.

It can be observed that Q-learning and SARSA and the Actor-Critic method return the exact same policies. SARSA- λ still needs a little more adjustment to the parameters to get the optimal policy. Actions a2 and a3 have a very high frequency that have a cost of 1 and help turn on more genes at a time, which is our goal in this problem, to keep majority of the genes on for majority of the time.

Average Accumulated Reward across all the algorithms:

A graph of different colored lines

AI-generated content may be incorrect.

From the above plot, it can be observed that Actor-Critic method is performing the best with the set parameters, suggesting that a policy-gradient approach is well-suited for this problem structure where number of states is lesser. Q-learning is close behind, showing that off-policy method can still learn near-optimal control. SARSA converges more slowly or to a slightly lower reward, consistent with it being on-policy and somewhat conservative in exploring. SARSA-λ is underperforming in this case, likely due to parameter with eligibility traces, so tuning may help get better results with SARSA-λ.