5B_Reinforcement_Learning_Assignment

May 12, 2021

1 PyTorch Assignment: Reinforcement Learning (RL)

Duke Community Standard: By typing your name below, you are certifying that you have adhered to the Duke Community Standard in completing this assignment.

Name:

1.0.1 Short answer

1. One of the fundamental challenges of reinforcement learning is balancing *exploration* versus *exploitation*. What do these two terms mean, and why do they present a challenge?

`[Your answer here]`

2. Another fundamental reinforcement learning challenge is what is known as the *credit assignment problem*, especially when rewards are sparse. What do we mean by the phrase, and why does it make learning especially difficult?

`[Your answer here]`

1.0.2 Deep SARSA Cart Pole

SARSA (state-action-reward-state-action) is another Q value algorithm that resembles Q-learning quite closely:

Q-learning update rule:

$$Q_\pi(s_t, a_t) \leftarrow (1-\alpha) \cdot Q_\pi(s_t, a_t) + \alpha \cdot \left(r_t + \gamma \max_a Q_\pi(s_{t+1}, a)\right) \tag{1}$$

SARSA update rule:

$$Q_\pi(s_t, a_t) \leftarrow (1-\alpha) \cdot Q_\pi(s_t, a_t) + \alpha \cdot \left(r_t + \gamma Q_\pi(s_{t+1}, a_{t+1})\right) \tag{2}$$

Unlike Q-learning, which is considered an *off-policy* network, SARSA is an *on-policy* algorithm. When Q-learning calculates the estimated future reward, it must "guess" the future, starting with the next action the agent will take. In Q-learning, we assume the agent will take the best possible action: $\max_a Q_\pi(s_{t+1}, a)$. SARSA, on the other hand, uses the action that was actually taken next in the episode we are learning from: $Q_\pi(s_{t+1}, a_{t+1})$. In other words, SARSA learns from the next

action he actually took (on policy), as opposed to what the max possible Q value for the next state was (off policy).

Build an RL agent that uses SARSA to solve the Cart Pole problem.

Hint: You can and should reuse the Q-Learning agent we went over earlier. In fact, if you know what you're doing, it's possible to finish this assignment in about 30 seconds.

[ ]:
```
### YOUR CODE HERE ###
```