RBE 595 — Reinforcement Learning
Deep Reinforcement Learning Assignment

Arjan Gupta

Problem 1

What are the two sources of error in Deep RL with function approximation?

Answer

The two sources of error in Deep RL with function approximation are as follows:

- **Bootstrapping error** — This is the error that arises due to the use of bootstrapping. Bootstrapping is the process of using the value of a successor state to update the value of a state. This is done in TD methods. The error is the difference between the target value and the current estimate value.
- **Approximation error** — This is the error defined as the difference between the true value function and the approximate value function. This error arises due to the use of function approximation itself.

Problem 2

In TD learning with a neural network what are we trying to minimize? What are we trying to maximize?

Answer

In general with any deep TD learning method, we aim to minimize the error between the target value and the current estimate Q-value. We try to maximize the value of the current state by choosing the action that maximizes the value of the next state.

We can use the example of gradient Q-learning to answer this question more specifically. In gradient Q-learning, we are trying to minimize the error given by the loss function as follows:

$$e(w) = \frac{1}{2}\left[Q_w(s,a) - \left(r + \gamma \max_{a'} Q_{w'}(s',a')\right)\right]^2$$

We are trying to maximize the value of the current state by choosing the action that maximizes the value of the next state. This is done by updating the weights of the neural network.

Problem 3

What are the similarities of TD and MC?

Answer

The similarities between TD and MC are as follows:

- Both TD and MC are model-free, i.e. they do not require a model of the environment.
- Both TD and MC are *sample updates*, i.e., they involve looking ahead at a sample successor state (or state-action pair), using the value of that state to compute a backed-up value, and then updating the value of the original state (or state-action pair) accordingly.

Problem 4

Assume that we have two states x and y with the current value of $V(x) = 10$, $V(y) = 1$. We run an episode of $\{x, 3, y, 0, y, 5, T\}$. What's the new estimate of $V(x)$, $V(y)$ using TD (assume step size $\alpha = 0.1$ and discount rate $\gamma = 0.9$).

Answer

The new estimate of $V(x)$ is as follows:

$$\begin{aligned} V(x) &= V(x) + \alpha\left[R_{t+1} + \gamma V(S_{t+1}) - V(x)\right] \\ &= 10 + 0.1\left[3 + 0.9 \cdot 1 - 10\right] \\ &= 10 + 0.1\left[3.9 - 10\right] \\ &= 10 + 0.1\left[-6.1\right] \\ &= 10 - 0.61 \\ &= 9.39 \end{aligned}$$

However, $V(y)$ gets updated twice in this episode. The first update is as follows:

$$\begin{aligned} V(y) &= V(y) + \alpha\left[R_{t+1} + \gamma V(S_{t+1}) - V(y)\right] \\ &= 1 + 0.1\left[0 + 0.9 \cdot 1 - 1\right] \\ &= 1 + 0.1\left[0.9 - 1\right] \\ &= 1 + 0.1\left[-0.1\right] \\ &= 1 - 0.01 \\ &= 0.99 \end{aligned}$$

The second update is as follows:

$$\begin{aligned} V(y) &= V(y) + \alpha\left[R_{t+1} + \gamma V(S_{t+1}) - V(y)\right] \\ &= 0.99 + 0.1\left[5 + 0.9 \cdot 0 - 0.99\right] \\ &= 0.99 + 0.1\left[5 - 0.99\right] \\ &= 0.99 + 0.1\left[4.01\right] \\ &= 0.99 + 0.401 \\ &= 1.391 \end{aligned}$$

Therefore, the new estimate of $V(x)$ is 9.39 and the new estimate of $V(y)$ is 1.391.

Problem 5

Can we consider TD an online (real-time) method and MC an offline method? Why?

Answer

Yes, we can consider TD an online (real-time) method and MC an offline method. This is because TD learns during the episode, whereas MC learns after the episode has ended. Specifically, TD updates the value of a state based on the value of the next state (during the episode), whereas MC updates the value of a state based on successive returns (after the episode has ended).

Problem 6

Does Q-learning learn the outcome of exploratory actions? (Refer to the Cliff walking example).

Answer

Yes, Q-learning learns the outcome of exploratory actions. This is because Q-learning is an off-policy TD control algorithm. In the context of the Cliff walking example, this means that Q-learning learns the optimal action-value function, q_*, which is closest to the cliff. The behavior policy takes exploratory actions, and the target policy is greedy. This causes makes it so that initially, the agent falls off the cliff a lot, but eventually, the agent learns to avoid the cliff and converges to the optimal action-value function, q_*. This is why the graph of the sum of rewards per episode for Q-learning is initially very low, but eventually converges to a high value. The graph is shown below:

Problem 7

What is the advantage of Double Q-learning over Q-learning?

Answer

The advantage of Double Q-learning over Q-learning is that Double Q-learning is less prone to bias than Q-learning. Specifically, Q-learning is biased toward the maximum value action, which is also known as maximization bias. This is because Q-learning uses the maximum value action to update the value of a state.

In contrast, Double Q-learning is not biased toward the maximum value action. This is because Double Q-learning uses two action-value functions, Q_1 and Q_2, to update the value of a state. With probability 0.5, Double Q-learning uses either Q_1 or Q_2 to update the value of a state. Now, for example, the action is picked as per Q_1, then that action is probably not the maximum value action as per Q_2. This way, for a given state, we have two estimates of the value of the maximum value action. This reduces the bias of Double Q-learning and is hence the advantage over Q-learning.