RBE 595 — Reinforcement Learning
Chapter #7 Assignment
n-step Bootstrapping

Arjan Gupta

Problem 1

The first episode of an agent interacting with an environment under policy π is as follows:

Timestep	Reward	State	Action
0		X	U1
1	16	X	U2
2	12	X	U1
3	24	X	U1
4	16	T	

Assume discount factor, $\gamma = 0.5$, step size $\alpha = 0.1$ and q_π is initially zero. What are the estimates of $q_\pi(X, U1)$ and $q_\pi(X, U2)$ using 2-step SARSA?

Answer

The estimates of $q_\pi(X, U1)$ and $q_\pi(X, U2)$ using 2-step SARSA are as follows:

Timestep 1

$$\begin{aligned}
q_\pi(X, U1) &= q_\pi(X, U1) + \alpha \left[R_{t+1} + \gamma R_{t+2} + \gamma^2 q_\pi(S_{t+2}, A_{t+2}) - q_\pi(X, U1)\right] \\
&= 0 + 0.1 \left[16 + 0.5 \cdot 12 + 0.5^2 \cdot 0 - 0\right] \\
&= 0 + 0.1 \left[16 + 6 - 0\right] \\
&= 0 + 0.1 \left[22\right] \\
&= 0 + 2.2 \\
&= 2.2
\end{aligned}$$

Timestep 2

$$\begin{aligned}
q_\pi(X, U2) &= q_\pi(X, U2) + \alpha \left[R_{t+1} + \gamma R_{t+2} + \gamma^2 q_\pi(S_{t+2}, A_{t+2}) - q_\pi(X, U2)\right] \\
&= 0 + 0.1 \left[12 + 0.5 \cdot 24 + 0.5^2 \cdot 0 - 0\right] \\
&= 0 + 0.1 \left[12 + 12 - 0\right] \\
&= 0 + 0.1 \left[24\right] \\
&= 0 + 2.4 \\
&= 2.4
\end{aligned}$$

Timestep 3

$$\begin{aligned}
q_\pi(X, U1) &= q_\pi(X, U1) + \alpha \left[R_{t+1} + \gamma R_{t+2} + \gamma^2 q_\pi(S_{t+2}, A_{t+2}) - q_\pi(X, U1)\right] \\
&= 2.2 + 0.1 \left[24 + 0.5 \cdot 16 + 0.5^2 \cdot 0 - 2.2\right] \\
&= 2.2 + 0.1 \left[24 + 8 - 2.2\right] \\
&= 2.2 + 0.1 \left[30.8\right] \\
&= 2.2 + 3.08 \\
&= 5.28
\end{aligned}$$

Problem 2

We mentioned that the target value for TD is $[R_{t+1} + \gamma V(s_{t+1})]$. What is the target value for Monte-carlo, Q-learning, SARSA and Expected-SARSA?

Answer

The Target is shown as part of the following equation:

$$NewEstimate \leftarrow OldEstimate + StepSize\,[Target - OldEstimate]$$

- **Monte-Carlo** (MC) does not use bootstrapping. Its target value is the actual return value, G_t.
- **Q-Learning** — As given in the algorithm, the target value is $R_{t+1} + \gamma \max_a Q(S_{t+1}, a)$.
- **SARSA** — As shown in the algorithm, the target value is $R_{t+1} + \gamma Q(S_{t+1}, A_{t+1})$.
- **Expected-SARSA** — As described in the book, the target value is $R_{t+1} + \gamma \mathbb{E}_\pi\,[Q(S_{t+1}, A_{t+1}) \mid S_{t+1}]$.

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.