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 0

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

Timestep 1

$$\begin{aligned}
q_\pi(X, U2) &= q_\pi(X, U2) + \alpha\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 q_\pi(S_{t+2}, A_{t+2}) - q_\pi(X, U2)\right] \\
&= 0 + 0.1\left[12 + 0.5 \cdot 24 + 0.5^2 \cdot q_\pi(X, U1) - 0\right] \\
&= 0 + 0.1\left[12 + 12 - 2.2\right] \\
&= 0 + 0.1\left[24 - 2.2\right] \\
&= 0 + 0.1\left[21.8\right] \\
&= 2.18
\end{aligned}$$

Timestep 2

$$\begin{aligned}
q_\pi(X, U1) &= q_\pi(X, U1) + \alpha\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 q_\pi(S_{t+2}, A_{t+2}) - q_\pi(X, U1)\right] \\
&= 2.2 + 0.1\left[24 + 0.5 \cdot 16 + 0.5^2 \cdot 0 - 2.2\right] \\
&= 2.2 + 0.1\left[24 + 8 - 2.2\right] \\
&= 2.2 + 0.1\left[30.8\right] \\
&= 5.18
\end{aligned}$$

Therefore, the estimate of $q_\pi(X, U1)$ is 5.18 and the estimate of $q_\pi(X, U2)$ is 2.18.

Problem 2

What is the purpose of introducing Control Variates in per-decision importance sampling?

Answer

The purpose of introducing Control Variates in per-decision importance sampling is to reduce the variance of G. This is done by using a linear combination of the original estimate and a control variate term. The updated equation is shown below:

$$G_{t:h} = \rho_t(R_{t+1} + \gamma G_{t+1:h}) + (1 - \rho_t)V_{h-1}(S_t)$$

Where the second term is the control variate term.

Problem 3

In off-policy learning, what are the pros and cons of the Tree-Backup algorithm versus off-policy SARSA (comment on the complexity, exploration, variance, and bias, and others)?

Answer

The pros and cons of the Tree-Backup algorithm versus off-policy SARSA are as follows:

- **Complexity:** The complexity of the Tree-Backup algorithm is $O(n)$, where n is the number of steps. The complexity of off-policy SARSA is $O(1)$.
- **Exploration:** The Tree-Backup algorithm explores the environment by following the policy π and then following the behavior policy μ for the remaining steps. Off-policy SARSA explores the environment by following the behavior policy μ for all steps.
- **Variance:** The variance of the Tree-Backup algorithm is lower than that of off-policy SARSA. This is because the Tree-Backup algorithm uses a control variate term to reduce the variance of the estimate.
- **Bias:** The bias of the Tree-Backup algorithm is higher than that of off-policy SARSA. This is because the Tree-Backup algorithm uses a control variate term to reduce the variance of the estimate.
- **Others:** The Tree-Backup algorithm is an on-policy algorithm, while off-policy SARSA is an off-policy algorithm.

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.