CS285: Deep Reinforcement Learning
Assignment 3
Written Report

Alan Sorani

May 29, 2025

1 Multistep Q-Learning

1.1 TD-Learning Bias

Assume that $\hat{Q}$ is a noisy unbiased estimate for Q. Then the Bellman backup $\mathscr{B}\hat{Q} \coloneqq r(s,a) + \gamma \max_{a'} \hat{Q}(s', a')$ is a **biased estimate** of $\mathscr{B}Q$. We have

$$\begin{aligned}\mathbb{E}_{\tau\sim p_\theta}\left[\mathscr{B}\hat{Q}\right] &= \mathbb{E}_{\tau\sim p_\theta}\left[r(s,a) + \gamma \max_{a'} \hat{Q}(s',a')\right] \\ &= \mathbb{E}_{\tau\sim p_\theta}\left[r(s,a)\right] + \gamma \mathbb{E}_{\tau\sim p_\theta}\left[\max_{a'} \hat{Q}(s',a')\right],\end{aligned}$$

and similarly

$$\mathbb{E}_{\tau\sim p_\theta}\left[\mathscr{B}Q\right] = \mathbb{E}_{\tau\sim p_\theta}\left[r(s,a)\right] + \gamma \mathbb{E}_{\tau\sim p_\theta}\left[\max_{a'} Q(s',a')\right],$$

so $\mathscr{B}\hat{Q}$ is an unbiased estimate of $\mathscr{B}Q$ if and only if

$$\mathbb{E}_{\tau\sim p_\theta}\left[\max_{a'} \hat{Q}(s',a')\right] = \mathbb{E}_{\tau\sim p_\theta}\left[\max_{a'} Q(s',a')\right].$$

This is not true. Consider an MDP with the action space $\mathscr{A} = \mathbb{R}$. Then we can have, $Q(s,\cdot) = 0$ and $Q(s,\cdot) \sim \mathscr{N}(0,1)$ where the latter is a Gaussian distribution of mean 0 and variance 1. Then, for every state s,

$$\mathbb{E}_{a\in\mathscr{A}}\left[\hat{Q}(s,a)\right] = 0 = \mathbb{E}_{a\in\mathscr{A}}\left[Q(s,a)\right]$$

so $\hat{Q}$ is an unbiased estimator of Q, but

$$\mathbb{E}_{\tau\sim p_\theta}\left[\max_{a'} \hat{Q}(s',a')\right] > 0, \mathbb{E}_{\tau\sim p_\theta}\left[\max_{a'} Q(s',a')\right] = \mathbb{E}_{\tau\sim p_\theta}\left[0\right] = 0,$$

as the expected value of the maximum of samples from Gaussian distribution is clearly positive. We get that $\mathscr{B}\hat{Q}$ is not an unbiased estimate for $\mathscr{B}Q$.

1.2 Tabular Learning

1.3 Variance of Q Estimates

For $N = 1$, the target value

$$y_{i,t} = r_{i,t} + \gamma^N \max_{a_{i,t+1}} Q_{\phi_k}(s_{i,t+1}, a_{i,t+1})$$

is an approximation for $Q(s_{i,t}, a_{i,t})$. As we increase N, the target value takes more actions which do not maximize the Q-value, before eventually maximizing the Q-value on the N^{th} step. We expect therefore that as N increases, the model would more closely fit the data, which would result in increased variance. Therefore, the minimal variance would be given when $N = 1$ and the maximal one as $N \to \infty$.

1.4 Function Approximation

1.5 Multistep Importance Sampling