CS234: Reinforcement Learning – Problem Session #3

Winter 2022-2023

Problem 1

For this problem, we will work with a reward function operating on transitions, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$. We are given an infinite-horizon, discounted MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \gamma \rangle$ but we will actually solve a MDP $\mathcal{M}'$ with an augmented reward function $\mathcal{M}' = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}', \mathcal{T}, \gamma \rangle$ where $\mathcal{R}'(s, a, s') = \mathcal{R}(s, a, s') + \mathcal{F}(s, a, s')$. To provide some motivation, think of a scenario where $\mathcal{R}$ produces values of 0 for most transitions; a bonus reward function $\mathcal{F} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ that produces non-zero values could provide us more immediate feedback and help accelerate the learning speed of our agent. In this problem, we will focus on a particular type of reward bonus $\mathcal{F}(s, a, s') = \gamma\phi(s') - \phi(s)$, for some arbitrary function $\phi : \mathcal{S} \to \mathbb{R}$ and $\forall (s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$.

1. Let $Q^\star_{\mathcal{M}}, Q^\star_{\mathcal{M}'}$ denote the optimal action-value functions of MDPs $\mathcal{M}$ and $\mathcal{M}'$, respectively. Using the Bellman equation, prove that $Q^\star_{\mathcal{M}}(s, a) - \phi(s) = Q^\star_{\mathcal{M}'}(s, a)$ and then use this fact to conclude that $\pi^\star_{\mathcal{M}'}(s) = \pi^\star_{\mathcal{M}}(s), \forall s \in \mathcal{S}$.

2. Consider running Q-learning in each MDP $\mathcal{M}$ and $\mathcal{M}'$ which requires, for each MDP, initial values $Q^0_{\mathcal{M}}(s,a)$ and $Q^0_{\mathcal{M}'}(s,a)$. Let $q_{\text{init}} \in \mathbb{R}$ be a real value such that

$$Q^0_{\mathcal{M}}(s,a) = q_{\text{init}} + \phi(s), \qquad Q^0_{\mathcal{M}'}(s,a) = q_{\text{init}}.$$

At any moment in time, the current Q-value of any state-action pair is always equal to its initial value plus some Δ value denoting the total change in the Q-value across all updates:

$$Q_{\mathcal{M}}(s,a) = Q^0_{\mathcal{M}}(s,a) + \Delta Q_{\mathcal{M}}(s,a), \qquad Q_{\mathcal{M}'}(s,a) = Q^0_{\mathcal{M}'}(s,a) + \Delta Q_{\mathcal{M}'}(s,a).$$

Show that if $\Delta Q_{\mathcal{M}}(s,a) = \Delta Q_{\mathcal{M}'}(s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, then show that these two Q-learning agents yield identical updates for any state-action pair.