CS234: Reinforcement Learning – Problem Session #2

Spring 2023-2024

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 $\pi^\star_{\mathcal{M}'}(s) = \pi^\star_{\mathcal{M}}(s), \forall s \in \mathcal{S}$.

   Solution:

$$
\begin{aligned}
Q^\star_{\mathcal{M}}(s,a) - \phi(s) &= \mathbb{E}_{s' \sim \mathcal{T}(\cdot|s,a)} \left[\mathcal{R}(s,a,s') + \gamma \max_{a' \in \mathcal{A}} Q^\star_{\mathcal{M}}(s',a') \right] - \phi(s) \\
&= \mathbb{E}_{s' \sim \mathcal{T}(\cdot|s,a)} \left[\mathcal{R}(s,a,s') - \phi(s) + \gamma \max_{a' \in \mathcal{A}} Q^\star_{\mathcal{M}} s', a') \right] \\
&= \mathbb{E}_{s' \sim \mathcal{T}(\cdot|s,a)} \left[\mathcal{R}(s,a,s') + \gamma\phi(s') - \gamma\phi(s') - \phi(s) + \gamma \max_{a' \in \mathcal{A}} Q^\star_{\mathcal{M}}(s',a') \right] \\
&= \mathbb{E}_{s' \sim \mathcal{T}(\cdot|s,a)} \left[\mathcal{R}(s,a,s') + \gamma\phi(s') - \phi(s) + \gamma \max_{a' \in \mathcal{A}} (Q^\star_{\mathcal{M}}(s',a') - \phi(s')) \right] \\
&= \mathbb{E}_{s' \sim \mathcal{T}(\cdot|s,a)} \left[\mathcal{R}(s,a,s') + \mathcal{F}(s,a,s') + \gamma \max_{a' \in \mathcal{A}} (Q^\star_{\mathcal{M}}(s',a') - \phi(s')) \right] \\
&= \mathbb{E}_{s' \sim \mathcal{T}(\cdot|s,a)} \left[\mathcal{R}'(s,a,s') + \gamma \max_{a' \in \mathcal{A}} (Q^\star_{\mathcal{M}}(s',a') - \phi(s')) \right] \\
\implies Q_{\mathcal{M}'}(s,a) &= \mathbb{E}_{s' \sim \mathcal{T}(\cdot|s,a)} \left[\mathcal{R}'(s,a,s') + \gamma \max_{a' \in \mathcal{A}} Q_{\mathcal{M}'}(s',a') \right] \\
\implies Q^\star_{\mathcal{M}'}(s,a) &= \mathbb{E}_{s' \sim \mathcal{T}(\cdot|s,a)} \left[\mathcal{R}'(s,a,s') + \gamma \max_{a' \in \mathcal{A}} Q^\star_{\mathcal{M}'}(s',a') \right].
\end{aligned}
$$

   Note that in the second-to-last line, we recognize that the equation we have corresponds to *some* action-value function of MDP $\mathcal{M}'$. In the final line, we acknowledge that this is the Bellman optimality

equation, which only holds for the optimal action-value function of $\mathcal{M}'$, $Q^\star_{\mathcal{M}'}$.

$$\begin{aligned}\pi^\star_{\mathcal{M}'}(s) &= \arg\max_{a\in\mathcal{A}} Q^\star_{\mathcal{M}'}(s,a)\\ &= \arg\max_{a\in\mathcal{A}} Q^\star_{\mathcal{M}}(s,a) - \phi(s)\\ &= \arg\max_{a\in\mathcal{A}} Q^\star_{\mathcal{M}}(s,a)\\ &= \pi^\star_{\mathcal{M}}(s)\end{aligned}$$

The general technique shown here for modifying the reward function is known as reward shaping. When $\mathcal{F} : \mathcal{S}\times\mathcal{A}\times\mathcal{S}\to\mathbb{R}$ is defined as described in this problem, this is known as potential-based reward shaping [Ng et al., 1999].

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.

Solution: We can expand the Q-learning update rule for the agent that does not use reward shaping as

$$\begin{aligned}&\mathcal{R}(s,a,s') + \gamma\max_{a'\in\mathcal{A}} Q_{\mathcal{M}}(s',a') - Q_{\mathcal{M}}(s,a) = \mathcal{R}(s,a,s') + \gamma\max_{a'\in\mathcal{A}} Q_{\mathcal{M}}(s',a') - Q^0_{\mathcal{M}}(s,a) - \Delta Q_{\mathcal{M}}(s,a)\\ &= \mathcal{R}(s,a,s') + \gamma\max_{a'\in\mathcal{A}} \left(Q^0_{\mathcal{M}}(s',a') + \Delta Q_{\mathcal{M}}(s',a')\right) - Q^0_{\mathcal{M}}(s,a) - \Delta Q_{\mathcal{M}}(s,a)\\ &= \mathcal{R}(s,a,s') + \gamma\max_{a'\in\mathcal{A}} \left(q_{\text{init}} + \phi(s') + \Delta Q_{\mathcal{M}}(s',a')\right) - q_{\text{init}} - \phi(s) - \Delta Q_{\mathcal{M}}(s,a)\\ &= \mathcal{R}(s,a,s') + \gamma\phi(s') - \phi(s) + \gamma\max_{a'\in\mathcal{A}} \left(q_{\text{init}} + \Delta Q_{\mathcal{M}}(s',a')\right) - q_{\text{init}} - \Delta Q_{\mathcal{M}}(s,a)\\ &= \mathcal{R}(s,a,s') + \mathcal{F}(s,a,s') + \gamma\max_{a'\in\mathcal{A}} \left(q_{\text{init}} + \Delta Q_{\mathcal{M}}(s',a')\right) - q_{\text{init}} - \Delta Q_{\mathcal{M}}(s,a)\\ &= \mathcal{R}'(s,a,s') + \gamma\max_{a'\in\mathcal{A}} \left(Q^0_{\mathcal{M}'}(s',a') + \Delta Q_{\mathcal{M}'}(s',a')\right) - Q^0_{\mathcal{M}'}(s,a) - \Delta Q_{\mathcal{M}'}(s,a)\\ &= \mathcal{R}'(s,a,s') + \gamma\max_{a'\in\mathcal{A}} Q_{\mathcal{M}'}(s',a') - Q_{\mathcal{M}'}(s,a),\end{aligned}$$

where the final equation is the Q-learning update for MDP $\mathcal{M}'$. This result shows that potential-based reward shaping as described in the previous part is equivalent to a particular Q-value initialization based on the potential function ϕ [Wiewiora, 2003].

References

Andrew Y Ng, Daishi Harada, and Stuart J Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 278–287. Morgan Kaufmann Publishers Inc., 1999.

Eric Wiewiora. Potential-based shaping and Q-value initialization are equivalent. *Journal of Artificial Intelligence Research*, 19:205–208, 2003.