CS234: Reinforcement Learning – Problem Session #2

Winter 2021-2022

Problem 1

Let's say we have an infinite-horizon MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \gamma \rangle$ and consider deterministic policies of the form $\pi : \mathcal{S} \to \mathcal{A}$.

1. Suppose we have an upper bound on the reward received at any timestep, $R_{\text{MAX}} \triangleq \max_{s,a} \mathcal{R}(s,a)$. Prove that for any state $s \in \mathcal{S}$ and any policy π, $V^\pi(s) \leq \frac{R_{\text{MAX}}}{(1-\gamma)}$.

   Solution:

$$\begin{aligned} V^\pi(s) &= \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, \pi(s_t)) \Big| s_0 = s\right] \\ &\leq \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_{\text{MAX}} \Big| s_0 = s\right] \\ &= \mathbb{E}\left[\frac{R_{\text{MAX}}}{(1-\gamma)} \Big| s_0 = s\right] \\ &= \frac{R_{\text{MAX}}}{(1-\gamma)} \end{aligned}$$

2. Now further assume that you also have a lower bound of 0 on rewards: $0 \leq \mathcal{R}(s,a) \leq R_{\text{MAX}}, \forall s \in \mathcal{S}, a \in \mathcal{A}$. Provide a simple modification to the reward function $\overline{\mathcal{R}}(s,a) = f(\mathcal{R}(s,a))$ such that **(1)** the corresponding MDP is guaranteed to have $0 \leq V^\pi(s) \leq 1$, for any policy π, state $s \in \mathcal{S}$ and **(2)** the optimal policy in the new MDP is the same as that of $\mathcal{M}$ (you should need at most two sentences to justify that the optimal policy of this new MDP and that of $\mathcal{M}$ are identical).

   Solution: Take $\overline{\mathcal{R}}(s,a) = \frac{(1-\gamma)}{R_{\text{MAX}}} \mathcal{R}(s,a)$ and repeat the calculations from the previous part. All we have done is scale our rewards by a fixed, positive constant $\frac{(1-\gamma)}{R_{\text{MAX}}}$; therefore, the optimal policy remains unchanged.

3. In the previous part, you made a modification to the reward function and assessed its effect on the optimal policy of the original MDP. Let's do this again for a different kind of reward function manipulation but actually prove that it preserves the optimal policy. For this problem, we will work with a reward function operating on transitions, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$. While our original MDP is defined with reward function $\mathcal{R}$, 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 $\forall (s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$. 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].

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.