CS234: Reinforcement Learning – Problem Session #2

Winter 2022-2023

Problem 1

Consider an infinite-horizon, discounted MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \gamma \rangle$.

1. Define the maximal reward $R_{\text{MAX}} = \max\limits_{(s,a) \in \mathcal{S} \times \mathcal{A}} \mathcal{R}(s,a)$ and show that, for any policy $\pi : \mathcal{S} \to \mathcal{A}$,

$$V^\pi(s) \leq \frac{R_{\text{MAX}}}{1-\gamma}, \qquad \forall s \in \mathcal{S}.$$

Solution:

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

2. Consider a second MDP $\widehat{\mathcal{M}} = \langle \mathcal{S}, \mathcal{A}, \widehat{\mathcal{R}}, \widehat{\mathcal{T}}, \gamma \rangle$ and define the constant $V_{\text{MAX}} = \frac{R_{\text{MAX}}}{1-\gamma}$. We will use subscripts to distinguish between arbitrary value functions $V_{\mathcal{M}}$ and $V_{\widehat{\mathcal{M}}}$ of MDPs $\mathcal{M}$ and $\widehat{\mathcal{M}}$, respectively. Suppose we have two constants $\varepsilon_1, \varepsilon_2 > 0$ such that

$$\max_{s,a \in \mathcal{S} \times \mathcal{A}} |\mathcal{R}(s,a) - \widehat{\mathcal{R}}(s,a)| \leq \varepsilon_1 \qquad \max_{s,a \in \mathcal{S} \times \mathcal{A}} \sum_{s' \in \mathcal{S}} |\mathcal{T}(s'|s,a) - \widehat{\mathcal{T}}(s'|s,a)| \leq \varepsilon_2.$$

For any policy $\pi : \mathcal{S} \to \mathcal{A}$, show that

$$||V^\pi_{\mathcal{M}} - V^\pi_{\widehat{\mathcal{M}}}||_\infty \leq \frac{\varepsilon_1 + \gamma \varepsilon_2 V_{\text{MAX}}}{(1-\gamma)}.$$

Solution:

$$
\begin{aligned}
||V^{\pi}_{\mathcal{M}} - V^{\pi}_{\widehat{\mathcal{M}}}||_{\infty} &= \max_{s\in\mathcal{S}} \Big| V^{\pi}_{\mathcal{M}}(s) - V^{\pi}_{\widehat{\mathcal{M}}}(s) \Big| \\
&= \max_{s\in\mathcal{S}} \Big| \mathcal{R}(s,\pi(s)) + \gamma \sum_{s'\in\mathcal{S}} \mathcal{T}(s'|s,\pi(s)) V^{\pi}_{\mathcal{M}}(s') - \widehat{\mathcal{R}}(s,\pi(s)) + \gamma \sum_{s'\in\mathcal{S}} \widehat{\mathcal{T}}(s'|s,\pi(s)) V^{\pi}_{\widehat{\mathcal{M}}}(s') \Big| \\
&\leq \max_{s\in\mathcal{S}} \Big| \mathcal{R}(s,\pi(s)) - \widehat{\mathcal{R}}(s,\pi(s)) \Big| + \Big| \gamma \sum_{s'\in\mathcal{S}} \mathcal{T}(s'|s,\pi(s)) V^{\pi}_{\mathcal{M}}(s') - \gamma \sum_{s'\in\mathcal{S}} \widehat{\mathcal{T}}(s'|s,\pi(s)) V^{\pi}_{\widehat{\mathcal{M}}}(s') \Big| \\
&\leq \varepsilon_1 + \max_{s\in\mathcal{S}} \gamma \Big| \sum_{s'\in\mathcal{S}} \mathcal{T}(s'|s,\pi(s)) V^{\pi}_{\mathcal{M}}(s') - \sum_{s'\in\mathcal{S}} \widehat{\mathcal{T}}(s'|s,\pi(s)) V^{\pi}_{\widehat{\mathcal{M}}}(s') \Big| \\
&= \varepsilon_1 + \max_{s\in\mathcal{S}} \gamma \Big| \sum_{s'\in\mathcal{S}} \mathcal{T}(s'|s,\pi(s)) V^{\pi}_{\mathcal{M}}(s') - \sum_{s'\in\mathcal{S}} \mathcal{T}(s'|s,\pi(s)) V^{\pi}_{\widehat{\mathcal{M}}}(s') + \sum_{s'\in\mathcal{S}} \mathcal{T}(s'|s,\pi(s)) V^{\pi}_{\widehat{\mathcal{M}}}(s') - \sum_{s'\in\mathcal{S}} \widehat{\mathcal{T}}(s'| \\
&\leq \varepsilon_1 + \max_{s\in\mathcal{S}} \gamma \Big| \sum_{s'\in\mathcal{S}} \mathcal{T}(s'|s,\pi(s)) \left(V^{\pi}_{\mathcal{M}}(s') - V^{\pi}_{\widehat{\mathcal{M}}}(s') \right) \Big| + \gamma \Big| \sum_{s'\in\mathcal{S}} V^{\pi}_{\widehat{\mathcal{M}}}(s') \left(\mathcal{T}(s'|s,\pi(s)) - \widehat{\mathcal{T}}(s'|s,\pi(s)) \right) \Big| \\
&\leq \varepsilon_1 + \max_{s\in\mathcal{S}} \gamma \Big| \sum_{s'\in\mathcal{S}} \mathcal{T}(s'|s,\pi(s)) ||V^{\pi}_{\mathcal{M}} - V^{\pi}_{\widehat{\mathcal{M}}}||_{\infty} \Big| + \gamma \Big| \sum_{s'\in\mathcal{S}} V^{\pi}_{\widehat{\mathcal{M}}}(s') \left(\mathcal{T}(s'|s,\pi(s)) - \widehat{\mathcal{T}}(s'|s,\pi(s)) \right) \Big| \\
&= \varepsilon_1 + \max_{s\in\mathcal{S}} \gamma ||V^{\pi}_{\mathcal{M}} - V^{\pi}_{\widehat{\mathcal{M}}}||_{\infty} \Big| \sum_{s'\in\mathcal{S}} \mathcal{T}(s'|s,\pi(s)) \Big| + \gamma \Big| \sum_{s'\in\mathcal{S}} V^{\pi}_{\widehat{\mathcal{M}}}(s') \left(\mathcal{T}(s'|s,\pi(s)) - \widehat{\mathcal{T}}(s'|s,\pi(s)) \right) \Big| \\
&= \varepsilon_1 + \gamma ||V^{\pi}_{\mathcal{M}} - V^{\pi}_{\widehat{\mathcal{M}}}||_{\infty} \Big| 1 \Big| + \max_{s\in\mathcal{S}} \gamma \Big| \sum_{s'\in\mathcal{S}} V^{\pi}_{\widehat{\mathcal{M}}}(s') \left(\mathcal{T}(s'|s,\pi(s)) - \widehat{\mathcal{T}}(s'|s,\pi(s)) \right) \Big| \\
&= \varepsilon_1 + \gamma ||V^{\pi}_{\mathcal{M}} - V^{\pi}_{\widehat{\mathcal{M}}}||_{\infty} + \max_{s\in\mathcal{S}} \gamma \Big| \sum_{s'\in\mathcal{S}} V^{\pi}_{\widehat{\mathcal{M}}}(s') \left(\mathcal{T}(s'|s,\pi(s)) - \widehat{\mathcal{T}}(s'|s,\pi(s)) \right) \Big| \\
&\leq \varepsilon_1 + \max_{s\in\mathcal{S}} \gamma ||V^{\pi}_{\mathcal{M}} - V^{\pi}_{\widehat{\mathcal{M}}}||_{\infty} + \max_{s\in\mathcal{S}} \gamma \Big| \sum_{s'\in\mathcal{S}} (V_{\mathrm{MAX}}) \left(\mathcal{T}(s'|s,\pi(s)) - \widehat{\mathcal{T}}(s'|s,\pi(s)) \right) \Big| \\
&= \varepsilon_1 + \gamma ||V^{\pi}_{\mathcal{M}} - V^{\pi}_{\widehat{\mathcal{M}}}||_{\infty} + \max_{s\in\mathcal{S}} \gamma V_{\mathrm{MAX}} \Big| \sum_{s'\in\mathcal{S}} \left(\mathcal{T}(s'|s,\pi(s)) - \widehat{\mathcal{T}}(s'|s,\pi(s)) \right) \Big| \\
&\leq \varepsilon_1 + \gamma ||V^{\pi}_{\mathcal{M}} - V^{\pi}_{\widehat{\mathcal{M}}}||_{\infty} + \gamma \varepsilon_2 V_{\mathrm{MAX}}
\end{aligned}
$$

$$
\begin{aligned}
\implies ||V^{\pi}_{\mathcal{M}} - V^{\pi}_{\widehat{\mathcal{M}}}||_{\infty} - \gamma ||V^{\pi}_{\mathcal{M}} - V^{\pi}_{\widehat{\mathcal{M}}}||_{\infty} &\leq \varepsilon_1 + \gamma \varepsilon_2 V_{\mathrm{MAX}} \\
(1-\gamma) ||V^{\pi}_{\mathcal{M}} - V^{\pi}_{\widehat{\mathcal{M}}}||_{\infty} &\leq \varepsilon_1 + \gamma \varepsilon_2 V_{\mathrm{MAX}} \\
||V^{\pi}_{\mathcal{M}} - V^{\pi}_{\widehat{\mathcal{M}}}||_{\infty} &\leq \frac{\varepsilon_1 + \gamma \varepsilon_2 V_{\mathrm{MAX}}}{(1-\gamma)}
\end{aligned}
$$

The cited result is known as the simulation lemma [Kearns and Singh, 2002]. The simulation lemma tells us that when we can recover an accurate approximation to the true reward function and transition function of our MDP, the return obtained by policies in the approximate MDP will be close to those of the original MDP.

References

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2-3):209–232, 2002.