CS234: Reinforcement Learning – Problem Session #5

Spring 2023-2024

Problem 1

Consider an infinite-horizon, discounted MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \gamma \rangle$. Define the maximal reward $R_{\text{MAX}} = \max\limits_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mathcal{R}(s,a)$. 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

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