CS234: Reinforcement Learning – Problem Session #1

Winter 2022-2023

Problem 1

Consider an infinite-horizon, discounted MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \gamma \rangle$. As usual, for any policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, the value function induced by π is defined as

$$V^\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t) \mid s_0 = s, \pi\right].$$

1. For an arbitrary $Z \in \mathbb{N}$, consider learning with $Z+1$ distinct discount factors $\gamma_0, \gamma_1, \ldots, \gamma_Z$ where the final discount factor matches that of the MDP $\mathcal{M}$, $\gamma_Z = \gamma$. Letting $[Z] \triangleq \{1, 2, \ldots, Z\}$ denote the index set, we define the following functions for any policy π:

$$V^\pi_{\gamma_z} = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma_z^t \mathcal{R}(s_t, a_t) \mid s_0 = s, \pi\right] \qquad W^\pi_z = V^\pi_{\gamma_z} - V^\pi_{\gamma_{z-1}}, \qquad \forall z \in [Z]$$

where $W_0 = V^\pi_{\gamma_0}$.

(a) For any $z \in [Z]$; any policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$; and any $s \in \mathcal{S}$, write an expression for $V^\pi_{\gamma_z}(s)$ exclusively in terms of $\{W^\pi_0, W^\pi_1, \ldots, W^\pi_Z\}$.

(b) Show that W^π_z obeys the following Bellman equation for any $z \in [Z]$ and $s \in \mathcal{S}$:

$$W^\pi_z(s) = \mathbb{E}_{\substack{a \sim \pi(\cdot|s) \\ s' \sim \mathcal{T}(\cdot|s,a)}}\left[(\gamma_z - \gamma_{z-1}) V^\pi_{\gamma_{z-1}}(s') + \gamma_z W_z(s')\right]$$

2. Let $\gamma, \beta \in [0, 1)$ be two discount factors such that $\beta \leq \gamma$. Let $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ be an arbitrary policy that induces value functions V_γ^π and V_β^π under the two discount factors, respectively. Similarly, define the Bellman operators

$$\mathcal{B}_\gamma^\pi V(s) = \mathbb{E}_{a \sim \pi(\cdot|s)} \left[\mathcal{R}(s, a) + \gamma \mathbb{E}_{s' \sim \mathcal{T}(\cdot|s,a)} \left[V(s')\right]\right]$$
$$\mathcal{B}_\beta^\pi V(s) = \mathbb{E}_{a \sim \pi(\cdot|s)} \left[\mathcal{R}(s, a) + \beta \mathbb{E}_{s' \sim \mathcal{T}(\cdot|s,a)} \left[V(s')\right]\right].$$

With the reward upper bound $R_{\text{MAX}} = \max\limits_{(s,a) \in \mathcal{S} \times \mathcal{A}} \mathcal{R}(s, a)$, prove that

$$||V_\gamma^\pi - V_\beta^\pi||_\infty \leq \frac{(\gamma - \beta) R_{\text{MAX}}}{(1 - \gamma)(1 - \beta)}.$$

3. Let $\alpha, \gamma \in [0, 1)$ be two discount factors such that $\gamma \leq \alpha$. Consider a new MDP $\mathcal{M}' = \langle \mathcal{S}, \mathcal{A}, \mathcal{T}', \mathcal{R}, \alpha \rangle$ with a different transition function $\mathcal{T}' : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ defined for $\lambda \in [0, 1]$ as

$$\mathcal{T}'(s' \mid s, a) = (1 - \lambda)\mathcal{T}(s' \mid s, a) + \lambda \mathbb{1}(s = s'), \qquad \forall (s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}.$$

In words, the new transition function $\mathcal{T}'$ follows the transitions of the original MDP $\mathcal{T}$ with probability $(1 - \lambda)$ and takes a self-looping transition with probability λ. We will use subscripts to distinguish between value functions of $\mathcal{M}$ versus those of $\mathcal{M}'$.

Assuming that both $\mathcal{M}$ and $\mathcal{M}'$ are tabular, recall the matrix form of the Bellman equations for any policy π:

$$V^{\pi}_{\mathcal{M}} = \left(I - \gamma \mathcal{T}^{\pi}\right)^{-1} \mathcal{R}^{\pi} \qquad V^{\pi}_{\mathcal{M}'} = \left(I - \alpha \mathcal{T}'^{\pi}\right)^{-1} \mathcal{R}^{\pi},$$

where

$$\mathcal{R}^{\pi}(s) = \mathbb{E}_{a \sim \pi(\cdot|s)}\left[\mathcal{R}(s, a)\right] \qquad \mathcal{T}^{\pi}(s' \mid s) = \mathbb{E}_{a \sim \pi(\cdot|s)}\left[\mathcal{T}(s' \mid s, a)\right] \qquad \mathcal{T}'^{\pi}(s' \mid s) = \mathbb{E}_{a \sim \pi(\cdot|s)}\left[\mathcal{T}'(s' \mid s, a)\right]$$

(a) Give a value of λ such that, for any policy π,

$$V^{\pi}_{\mathcal{M}'} = \frac{1 - \gamma}{1 - \alpha} \cdot V^{\pi}_{\mathcal{M}}.$$

(b) If $\pi^{\star}$ is the optimal policy of MDP $\mathcal{M}$, prove that $\pi^{\star}$ is also optimal in $\mathcal{M}'$.