

CS234: Reinforcement Learning – Problem Session #6

Spring 2023-2024

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 $\pi : \mathcal{S} \rightarrow \Delta(\mathcal{A})$, 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 $\gamma_0, \gamma_1, \dots, \gamma_Z$ where the final discount factor matches that of the MDP \mathcal{M} , $\gamma_Z = \gamma$. Letting $[Z] \triangleq \{1, 2, \dots, Z\}$ denote the index set, we define the following functions for any policy π :

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

where $W_0 = V_{\gamma_0}^\pi$.

- (a) For any $z \in [Z]$; any policy $\pi : \mathcal{S} \rightarrow \Delta(\mathcal{A})$; and any $s \in \mathcal{S}$, write an expression for $V_{\gamma_z}^\pi(s)$ exclusively in terms of $\{W_0^\pi, W_1^\pi, \dots, W_Z^\pi\}$.

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

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

2. Let $\gamma, \beta \in [0, 1)$ be two discount factors such that $\beta \leq \gamma$. Let $\pi : \mathcal{S} \rightarrow \Delta(\mathcal{A})$ 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

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

With the reward upper bound $R_{\text{MAX}} = \max_{(s, a) \in \mathcal{S} \times \mathcal{A}} \mathcal{R}(s, a)$, 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 $\mathcal{T}' : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$ defined for $\lambda \in [0, 1]$ as

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

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 :

$$V_{\mathcal{M}}^{\pi} = (I - \gamma\mathcal{T}^{\pi})^{-1} \mathcal{R}^{\pi} \quad V_{\mathcal{M}'}^{\pi} = (I - \alpha\mathcal{T}'^{\pi})^{-1} \mathcal{R}^{\pi},$$

where

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

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

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

- (b) If π^* is the optimal policy of MDP \mathcal{M} , prove that π^* is also optimal in \mathcal{M}' .