

# 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  $\pi : \mathcal{S} \rightarrow \mathcal{A}$ .

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

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  $\bar{\mathcal{R}}(s,a) = f(\mathcal{R}(s,a))$  such that **(1)** the corresponding MDP is guaranteed to have  $0 \leq V^\pi(s) \leq 1$ , for any policy  $\pi$ , 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).

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,  $\mathcal{R} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$ . 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  $\mathcal{F} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$  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  $\phi : \mathcal{S} \rightarrow \mathbb{R}$  and  $\forall (s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ . Let  $Q_{\mathcal{M}}^*, Q_{\mathcal{M}'}^*$  denote the optimal action-value functions of MDPs  $\mathcal{M}$  and  $\mathcal{M}'$ , respectively. Using the Bellman equation, prove that  $Q_{\mathcal{M}}^*(s, a) - \phi(s) = Q_{\mathcal{M}'}^*(s, a)$  and then use this fact to conclude that  $\pi_{\mathcal{M}'}^*(s) = \pi_{\mathcal{M}}^*(s), \forall s \in \mathcal{S}$ .