Derivation of optimality condition in Inverse reinforcement learning algorithm

Professor Vwani Roychowdhury

May 11, 2018

In this set of notes, we will derive the feasibility condition that the reward vectors R needs to satisfy in the Inverse reinforcement learning algorithm. To facilitate the derivation, we introduce the following notation

$$\mathbf{V}^* = [V^{\pi^*}(s_1), V^{\pi^*}(s_2), V^{\pi^*}(s_3), \cdots, V^{\pi^*}(s_{|\mathcal{S}|})]^T$$
$$\mathbf{R} = [R(s_1), R(s_2), R(s_3), \cdots, R(s_{|\mathcal{S}|})]^T$$

where π^* is the optimal policy. We also rename the action set $\mathcal{A}$ in a manner such that the optimal action at each state is always denoted by a_1

$$a_1 = \arg \max_{a \in \mathcal{A}} Q^{\pi^*}(s, a), \quad \forall s \in \mathcal{S} \tag{1}$$

Then with the above notation, the bellman optimality equation becomes

$$V^*(s) = \sum_{s'} \mathcal{P}^a_{ss'}[R(s') + \gamma V^*(s')] \tag{2}$$

Writing equation 2 in matrix form and doing some manipulation, we get

$$\begin{aligned} \mathbf{V}^* &= \mathbf{R} + \gamma \mathbf{P}_{a_1} \mathbf{V}^* \\ \Rightarrow (I - \gamma \mathbf{P}_{a_1})\mathbf{V}^* &= \mathbf{R} \\ \Rightarrow \mathbf{V}^* &= (I - \gamma \mathbf{P}_{a_1})^{-1}\mathbf{R} \end{aligned} \tag{3}$$

From equation 1, we have

$$Q^*(s, a_1) \geq Q^*(s, a_j),\ j = 2, \cdots, k,\ \forall s \tag{4}$$

The inequality given by 4 can be written in matrix form as

$$\mathbf{R} + \gamma \mathbf{P}_{a_1} \mathbf{V}^* \geq \mathbf{R} + \gamma \mathbf{P}_a \mathbf{V}^*,\ \forall a \in \mathcal{A} \setminus a_1 \tag{5}$$

Simplifying the above inequality, we get

$$(\mathbf{P}_{a_1} - \mathbf{P}_a)\mathbf{V}^* \geq 0,\ \forall a \in \mathcal{A} \setminus a_1 \tag{6}$$

Plugging in the expression for $\mathbf{V}^*$ into the inequality given by 6, we get the desired result

$$(\mathbf{P}_{a_1} - \mathbf{P}_a)(I - \gamma \mathbf{P}_{a_1})^{-1}\mathbf{R} \geq 0,\ \forall a \in \mathcal{A} \setminus a_1 \tag{7}$$