EE 290 Theory of Multi-armed Bandits and Reinforcement Learning Lecture 15 - 3/9/2021

Lecture 15: Interpretations of Bellman Equation, Optimality Conditions

Lecturer: Jiantao Jiao *Scribe: Koulik Khamaru, Yiling You*

In this lecture, we continue the discussions on Markov decision processes (MDPs) and basics of reinforcement learning (RL).

1 Remarks

- In the old days of reinforcement learning, people spent lots of time studying the planning algorithms, which is still an active area of research today.
- Nowadays, many statistics and machine learning people start to learn the RL problems, introducing statistical and learning aspects to the story.

2 Recap

Last time we saw the Bellman equation for policy evaluation:

$$V^{\pi}(s) = \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \middle| s_1 = s, \pi\right] \tag{1}$$

where π is the policy, $\gamma \in [0,1)$ is the discounted factor, and $\{r_t\}_{t\in\mathbb{N}}$ is the reward. The $V^{\pi}(s)$ tells us the expected accumulated discounted reward given a policy π, which quantifies how good the policy π is. In the vector notation, assuming temporally π is deterministic (for the sake of simplicity), if we write

$$V^{\pi} = [V^{\pi}(s)]_{s\in\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}|\times 1}, \tag{2}$$

$$R^{\pi} = [R^{\pi}(s, \pi(s))]_{s\in\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}|\times 1}, \tag{3}$$

$$P^{\pi} = [p(s'|s, \pi(s))]_{s'\in\mathcal{S}, s\in\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}, \tag{4}$$

we can re-state the Bellman equation for policy evaluation in eq. (1) as

$$V^{\pi} = R^{\pi} + \gamma P^{\pi} V^{\pi} \tag{5}$$

or equivalently,

$$V^{\pi} = R^{\pi}. \tag{6}$$

3 Interpretation of the Bellman Equation eq. (6)

The eq. (6) is quite interesting: the R^{π} is the instantaneous reward, which is designed to capture the structure of the problem. When multiplied with the matrix $(I - \gamma P^{\pi})^{-1}$, the expression gives rise to the value of the policy. When the reward is not dependent on π, i.e., $R^{\pi}(s, \pi(s)) = R(s)$, eq. (6) shows the value function of π is a linear function of the reward R. This linear transformation has a beautiful interpretation called *discounted occupancy measure*, which is an important concept in the discounted infinite horizon MDP.

3.1 Interpretation of $(I - \gamma P^\pi)^{-1}$

We aim to prove the following result: each row of the matrix $(I - \gamma P^\pi)^{-1}$, indexed by $s \in \mathcal{S}$, is the unnormalized discounted state occupancy, whose s'-th entry is

$$\frac{d_s^\pi(s')}{1-\gamma} = \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} \mathbb{1}(s_t = s') \middle| s_1 = s, \pi\right]. \tag{7}$$

- The $d_s^\pi(s')$ in the numerator of the LHS of eq. (7) is a probability distribution supported on $\mathcal{S}$, indexed by s', because we will check

$$d_s^\pi(s') \geq 0, s' \in \mathcal{S}, \tag{8}$$

$$\sum_{s' \in \mathcal{S}} d_s^\pi(s') = 1. \tag{9}$$

  We call $d_s^\pi(s')$ the *normalized discounted occupancy*.

- On the RHS of eq. (7), we observe the correct interpretation of state occupancy: imagine we rollout the policy π starting from state $s_1 = s$ and check at every time step t whether we are in the state s'. We prefer early visits of state s' and therefore we give late visits large discounting factors. Overall, the RHS is the expected total discounted visits of a particular state s' over the trajectory.

- The $1 - \gamma$ in the denominator of the LHS of eq. (7) is due to

$$\sum_{t=1}^{\infty} \gamma^{t-1} = \frac{1}{1-\gamma}. \tag{10}$$

3.1.1 Deeper dive into eq. (7)

Note that thanks to the fact that the terms in the expectation are non-negative and by the bounded convergence theorem, we can swap the expectation $\mathbb{E}[\cdot]$ and the infinite sum $\sum_{t=1}^{\infty}$,

$$\frac{d_s^\pi(s')}{1-\gamma} = \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} \mathbb{1}(s_t = s') \middle| s_1 = s, \pi\right] \tag{11}$$

$$= \sum_{t=1}^{\infty} \gamma^{t-1} \mathbb{E}\left[\mathbb{1}(s_t = s') | s_1 = s, \pi\right] \tag{12}$$

$$= \sum_{t=1}^{\infty} \gamma^{t-1} \mathbb{P}\left[s_t = s' | s_1 = s, \pi\right]. \tag{13}$$

We further introduce

$$d_{s,t}^\pi(s') = \mathbb{P}\left[s_t = s' | s_1 = s, \pi\right], \tag{14}$$

i.e., the step t state distribution conditioned on the initial state $s_1 = s$ and policy π. eq. (14) is closely related to the Markov chain theory: after specifying the policy π, the MDP reduces to a Markov chain, and eq. (14) quantifies the marginal distribution of the state s_t. Combining eq. (13) and eq. (14) and using the vector notations $d_s^\pi = [d_s^\pi(s')]_{s' \in \mathcal{S}} \in \mathbb{R}^{|\mathcal{S}| \times 1}, d_{s,t}^\pi = \left[d_{s,t}^\pi(s')\right]_{s' \in \mathcal{S}} \in \mathbb{R}^{|\mathcal{S}| \times 1}$, we can easily conclude

$$d_s^\pi = (1-\gamma) \sum_{t=1}^{\infty} \gamma^{t-1} d_{s,t}^\pi. \tag{15}$$

Hence we see the distribution d_s^π is exactly a weighted distribution of all marginals $d_{s,t}^\pi$, where the initial steps marginals are assigned more weights.

3.1.2 Extract the rows of $(I - \gamma P^\pi)^{-1}$

In order to give a more clear interpretation, let's extract the rows of $(I - \gamma P^\pi)^{-1}$ explicitly. This can be done by introducing the row vectors

$$e_s = [0, \ldots, 0, 1, 0, \ldots, 0] \in \mathbb{R}^{1 \times |\mathcal{S}|} \tag{16}$$

where the s-th entry is 1, and using the infinite series expansion of $(I - \gamma P^\pi)^{-1}$,

$$e_s (I - \gamma P^\pi)^{-1} = e_s \left[\sum_{t=1}^{\infty} \gamma^{t-1} (P^\pi)^{t-1} \right] \tag{17}$$

$$= \sum_{t=1}^{\infty} \gamma^{t-1} \left[e_s (P^\pi)^{t-1} \right] \tag{18}$$

Now by the Markov chain theory, we immediately see $e_s (P^\pi)^{t-1}$ describes the distribution of states at time t with the initial state s. We hence build a solid connection between the Bellman equation and classical Markov chain theory by the matrix $(I - \gamma P^\pi)^{-1}$.

One last observation is that our computation in eq. (17) implies

$$V^\pi(s) = e_s (I - \gamma P^\pi)^{-1} R_\pi \tag{19}$$

$$= \sum_{t=1}^{\infty} \gamma^{t-1} \left\langle d_{s,t}^\pi, R^\pi \right\rangle \tag{20}$$

$$= \sum_{t=1}^{\infty} \gamma^{t-1} \mathbb{E}_{s' \sim d_{s,t}^\pi} \left[R(s', \pi(s')) \right]. \tag{21}$$

4 Optimal policies and the Q-functions

Theorem 1. *For infinite horizon discounted MDPs, there exists a stationary and deterministic policy such that it is optimal for all starting state simultaneously.*

Proof See the Theorem 6.2 in the book [1] for details. □

A few comments regarding Theorem 1 are in order. The theorem posits that there exists an optimal policy (say $\pi^\star$) which is *stationary* and *deterministic*. Recall that a policy is *stationary* if it does not depend on the time t; a policy π is deterministic if for each $s \in \mathcal{S}$, the $\pi(s)$ is a fixed element in the action space $\mathcal{A}$, as opposed to a distribution over the action space $\mathcal{A}$ (for a random policy). Combining the last two statements we deduce that while searching for an optimal policy of an infinite horizon discounted MDP, it suffices to search over stationary and deterministic policies.

Caution: We point out that Theorem 1 *does not* claim that optimal policy is *unique*. In fact, there might exists optimal policies which are both non-stationary and non-deterministic.

In order to avoid this non-uniqueness problem of optimal policies, we often calculate the value function and Q function associated with optimal policies, also known as the optimal value function and optimal Q function, respectively. While optimal policies may not be unique, both the optimal value function and the optimal Q-function are unique. Moreover, once we have calculated the optimal Q-function, we can find an optimal policy from the optimal Q-function. In the rest of the lecture, we discuss how to calculate the optimal value function and the optimal Q function.

4.1 Q-function

Before introducing the optimal Q-function associated with an MDP, we need to introduce the notion of Q-function associated with a policy π. Formally, for a given policy π, the Q-function Q^π is a map from the state-action space $\mathcal{S} \times \mathcal{A}$ to the real numbers, and it is defined as follows:

$$Q^\pi(s,a) := \mathbb{E}[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \mid s_1 = s, a_1 = a, \pi].$$

In words, $Q^\pi(s,a)$ is the expected long-term discounted reward when we start at state $s_1 = s$ with an initial action $a_1 = a$, and in the subsequent states we take actions according to policy π. In terms of the above notation, we have the following relation between the Q-function Q^π and the value function V^π:

$$V^\pi(s) := \mathbb{E}_{a \sim \pi(\cdot|s)}[Q^\pi(s,a)].$$

Let $Q^{\pi^\star}$ be the Q-function associated with an optimal policy $\pi^\star$, then, in the later lectures, we show that $Q^{\pi^\star}$ is same for any optimal policy $\pi^\star$ and we use $Q^\star$ to denote this (unique) Q-function. We also call it the optimal Q function.

4.2 Bellman equations

The Q-functions Q^π and $Q^\star$ follows a set of equations which allows us to compute these functions easily. These equations are called the "Bellman equations":

$$Q^\pi(s,a) = R(s,a) + \gamma \mathbb{E}_{a \sim \pi(\cdot|s), s' \sim p(\cdot|s,a)}[Q^\pi(s',a)] \qquad \text{for all} \quad (s,a) \in \mathcal{S} \times \mathcal{A}. \tag{22}$$

$$Q^{\pi^\star}(s,a) = R(s,a) + \gamma \mathbb{E}_{s' \sim p(\cdot|s,a)}[\max_a Q^{\pi^\star}(s',a)] \quad \text{for all} \quad (s,a) \in \mathcal{S} \times \mathcal{A}. \tag{23}$$

To understand why these Bellman equations are useful, let us assume that the state space and the action space are finite, and the transition dynamics $p(\cdot \mid s,a)$ is known for all state action pairs (s,a). Then,the set of equations (22) is a system is linear equations, and the Q-function Q^π can be obtained as the (unique[1]) solution of this linear system of equations. The Bellman equation (23) for the optimal Q function Q^* is a system of non-linear equations, and we need slightly more involved algorithms to solve them. We will discuss relevant algorithms in future lectures.

Similar to the Q-functions, the optimal value function V^* also satisfies the following Bellman equations:

$$V^*(s) = \max_{a \in \mathcal{A}} \ \left\{ R(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^*(s')] \right\} \quad \text{for all} \quad s \in \mathcal{S}.$$

Comparing trajectories of the value functions and Q-functions

Before we end today's lecture, we provide an interpretion of the Q-function and value function in terms of the actions taken (trajectories) at different times. Both these type of functions calculate expected discounted cumulative reward, but they follow different trajectories:

$$V^\pi(s) = a_1 \sim \pi(s_1) \ \ a_2 \sim \pi(s_2) \ldots \pi \ldots \tag{24}$$

$$Q^\pi(s,a) = a_1 \sim a \ \ a_2 \sim \pi(s_2) \ \ \ldots \pi \ldots \tag{25}$$

$$V^\star(s) = a_1 \sim \pi^\star(s_1) \ \ a_2 \sim \pi^\star(s_2) \ldots \pi^\star \ldots \tag{26}$$

$$Q^*(s,a) = a_1 =\sim a \ \ a_2 \sim \pi^\star(s_2) \ \ \ldots \pi^\star \ldots \tag{27}$$

Here, $\pi^\star$ is an optimal policy.

[1]The uniqueness part will be clear in subsequent lectures

References

[1] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming.* John Wiley & Sons, 2014.