CSC 696H: Topics in Reinforcement Learning Theory **Fall 2021**

Lecture 3: Bellman consistency equation; MRPs; Optimal value function

Lecturer: Chicheng Zhang *Scribe: Yinan Li*

1 Questions from last class

Recall that last time, we defined the value function of π, $V^\pi : \mathcal{S} \to \mathbb{R}$, which is the expected discounted return of a policy, conditioned on the starting state:

$$V^\pi(s) = \mathbb{E}\left[G_0 \mid s_0 = s, \pi\right].$$

We also defined the action-value function of π, $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, which is conditioned additionally on the starting action:

$$Q^\pi(s,a) = \mathbb{E}\left[G_0 \mid s_0 = s, a_0 = a, \pi\right].$$

Using *Bellman consistency equation*, and eliminating variables $V^\pi(s)$, we have:

$$\forall s \in \mathcal{S}, a \in \mathcal{A} : Q^\pi(s,a) = r(s,a) + \gamma \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} P(s' \mid s,a)\pi(a' \mid s')Q^\pi(s',a').$$

Note that this is a system of linear equations, where the unknown variables are $Q^\pi(s,a)$'s, and the known variables are immediate rewards $r(s,a)$'s and environmental dynamics.

The questions left from the last time are:

1. Does this system have solutions?
2. If so, is the solution unique?

In other words, we would like to find a plausible choice of f, that satisfies:

$$f(s,a) = r(s,a) + \gamma \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} P(s' \mid s,a)\pi(a' \mid s')f(s',a'). \tag{1}$$

For the first question, we observe that $f(s,a) = Q^\pi(s,a)$ is a valid solution to (1). We will answer the second question in this lecture.

2 Markov Reward Processes (MRPs)

Markov Reward Processes (MRPs) are slight variants of MDPs, where the agent does not have controls on the environment. Formally, an MRP M is described as:

$$M = (\mathcal{S}, \mathbb{P}, r, \gamma, \mu),$$

where, different from those of MDPs,

1. Transition probability $\mathbb{P} : \mathcal{S} \to \Delta(\mathcal{S})$;
2. Reward function $r : \mathcal{S} \to [0,1]$.

Algorithm 1 Agent-environment interaction protocol with an MRP

Initial state $z_0 \sim \mu$
for time steps $t = 0, 1, \ldots,$ **do**
  Agent receives reward $r_t = r(z_t)$
  Agent observes next state $z_{t+1} \sim \mathbb{P}(\cdot \mid z_t)$
end for

Similar to MDP setting, we could define the value function in MRP context:

$$V(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s\right].$$

Exercise 1. Give a Bellman consistency equation for V.

Exercise 2. Suppose we execute a stationary policy π in a given MDP M. Can the interaction process be seen as an MRP? Here are two possible answers to this exercise.

Answer 1. MRP has the same state space as MDP, and we could let $z_t = s_t$, and

$$\mathbb{P}(s_{t+1} = s' \mid s_t = s) = \sum_{a \in \mathcal{A}} \pi(a \mid s)\mathbb{P}(s' \mid s, a).$$

Note that there is some subtlety, in that r_t is not a deterministic function of s_t, but still, we could define the reward function in expectation:

$$\mathbb{E}\left[r_t \mid s_t = s\right] = \sum_{a \in \mathcal{A}} \pi(a \mid s) r(s, a).$$

Answer 2. MRP has the state space $\mathcal{S} \times \mathcal{A}$, namely $z_t = (s_t, a_t)$, and

$$\mathbb{P}\left[(s_{t+1}, a_{t+1}) = (s', a') \mid (s_t, a_t) = (s, a)\right] = \mathbb{P}(s' \mid s, a)\pi(a' \mid s').$$

MRP has the same reward function as MDP.

3 Back to Bellman consistency equation

Recall that,

$$\forall s \in \mathcal{S}, a \in \mathcal{A} : Q^\pi(s, a) = r(s, a) + \gamma \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} P(s' \mid s, a)\pi(a' \mid s')Q^\pi(s', a').$$

and it helps to look at it from linear algebra perspective.

Denote by $S = |\mathcal{S}|$, and $A = |\mathcal{A}|$,

$$Q^\pi = \begin{bmatrix} \vdots \\ Q^\pi(s, a) \\ \vdots \end{bmatrix}_{SA \times 1},$$

$$r = \begin{bmatrix} \vdots \\ r(s, a) \\ \vdots \end{bmatrix}_{SA \times 1},$$

$$P^\pi = \begin{bmatrix} & \vdots & \\ \pi(a' \mid s') \cdot P(s' \mid s, a) & & \\ & \vdots & \end{bmatrix}_{SA \times SA}.$$

Now we are ready to rewrite Bellman consistency equation in the matrix notation,

$$Q^\pi = r + \gamma \cdot P^\pi \cdot Q^\pi.$$

To solve it, rewrite it as:

$$(I - \gamma \cdot P^\pi) \cdot Q^\pi = r.$$

Claim 1. *$I - \gamma \cdot P^\pi$ is invertible.*

Proof. It suffices to show, $\forall x \neq 0$, $(I - \gamma \cdot P^\pi) \cdot x \neq 0$. To see this, we analyse the ℓ_∞ norm.

$$\begin{aligned} \|(I - \gamma \cdot P^\pi) \cdot x\|_\infty =& \|x - \gamma \cdot P^\pi \cdot x\|_\infty \\ \geq& \|x\|_\infty - \|\gamma \cdot P^\pi \cdot x\|_\infty \text{ (by triangle ineq.)} \end{aligned}$$

Note that each row in P^π sum up to 1, thus we have

$$|(P^\pi \cdot x)_i| = |\sum_i w_i x_i| \leq \sum_i w_i |x_i| \leq \|x\|_\infty.$$

Therefore,

$$\|(I - \gamma \cdot P^\pi) \cdot x\|_\infty \geq (1 - \gamma)\|x\|_\infty > 0.$$

□

Since $I - \gamma \cdot P^\pi$ is invertible, we are able to claim that

$$Q^\pi = (I - \gamma \cdot P^\pi)^{-1} \cdot r.$$

Now we take a closer look at entries of $(I - \gamma \cdot P^\pi)^{-1}$. Analogous to the algebraic fact that $\frac{1}{1-\gamma} = \sum_{t=0}^\infty \gamma^t$, we also have,

$$\begin{aligned} (I - \gamma \cdot P^\pi)^{-1}_{(s,a),(s',a')} =& \sum_{t=0}^\infty \gamma^t (P^\pi)^t_{(s,a),(s',a')} \\ =& \sum_{t=0}^\infty \gamma^t P\left[(s_t, a_t) = (s', a') \mid (s_0, a_0) = (s, a)\right] \geq 0. \end{aligned}$$

This is called discounted state-action visitation (occupancy).

To see the second equality, we analyse the case where $t = 2$, as an example; this can be easily generalized to interpret $(P^\pi)^t_{(s,a),(s',a')}$ as $P\left[(s_t, a_t) = (s', a') \mid (s_0, a_0) = (s, a)\right]$.

$$\begin{aligned} (P^\pi)^2_{(s,a),(s',a')} =& \sum_{\tilde{s},\tilde{a}} (P^\pi)_{(s,a),(\tilde{s},\tilde{a})} \cdot (P^\pi)_{(\tilde{s},\tilde{a}),(s',a')} \\ =& \sum_{\tilde{s},\tilde{a}} P\left[(s_1, a_1) = (\tilde{s}, \tilde{a}) \mid (s_0, a_0) = (s, a)\right] \cdot P\left[(s_2, a_2) = (s', a') \mid (s_1, a_1) = (\tilde{s}, \tilde{a})\right] \\ =& P\left[(s_2, a_2) = (s', a') \mid (s_0, a_0) = (s, a)\right] \end{aligned}$$

where the first equality is by matrix multiplication formula, the second equality is by the definition of P^π, and the third is by Markov property and total law of probability.

4 Optimal value function

Optimal value function could be thought of as a more objective measure of how advantage a state is. Previously, we defined the value function of a state, but that depends on the policy the agent executes. For example, think about a state that is very close to the goal state, and it is promising to get high reward. But the policy may not be performing so well, and it is actually directing away from the goal state. In this case, the value function of the state may be bad, although the state itself is good. Thus we would like a more objective measure of how "good" a state is.

The way we formulate this idea is to define the optimal value function and optimal action-value function as follows:

$$V^\star(s) = \max_{\pi \in \Pi} V^\pi(s),$$
$$Q^\star(s,a) = \max_{\pi \in \Pi} Q^\pi(s,a),$$

here, Π is the collection of all policies (including nonstationary ones).

Representing $V^\star$ with $Q^\star$:

$$\begin{aligned} V^\star(s) =& \max_\pi \mathbb{E}\left[G_0 \mid s_0 = s, \pi\right] \\ =& \max_\pi \sum_{a \in \mathcal{A}} \pi(a \mid s) \mathbb{E}\left[G_0 \mid s_0 = s, a_0 = a, \pi\right] \\ \leq& \max_\pi \sum_{a \in \mathcal{A}} \pi(a \mid s) \max_{\pi'} \mathbb{E}\left[G_0 \mid s_0 = s, a_0 = a, \pi'\right] \\ =& \max_\pi \sum_{a \in \mathcal{A}} \pi(a \mid s) Q^\star(s,a). \end{aligned} \tag{2}$$

Note that the optimal π will be: $\pi(a \mid s) = \begin{cases} 1 & a = \text{argmax}_{a'} Q^\star(s,a') \\ 0 & \text{otherwise} \end{cases}$,

then we further have,

$$V^\star(s) \leq \max_a Q^\star(s,a).$$

Representing $Q^\star$ using $V^\star$:

$$\begin{aligned} Q^\star(s,a) =& \max_\pi \mathbb{E}\left[G_0 \mid s_0 = s, a_0 = a, \pi\right] \\ =& r(s,a) + \gamma \cdot \max_\pi \mathbb{E}\left[G_1 \mid s_0 = s, a_0 = a, \pi\right] \\ =& r(s,a) + \gamma \cdot \max_\pi \sum_{s' \in \mathcal{S}} P(s' \mid s,a) \mathbb{E}\left[G_1 \mid s_0 = s, a_0 = a, s_1 = s', \pi\right] \\ \leq& r(s,a) + \gamma \cdot \sum_{s' \in \mathcal{S}} P(s' \mid s,a) \max_\pi \mathbb{E}\left[G_1 \mid s_0 = s, a_0 = a, s_1 = s', \pi\right] \\ \leq& r(s,a) + \gamma \cdot \sum_{s' \in \mathcal{S}} P(s' \mid s,a) V^\star(s') \text{ (Next class)} \end{aligned} \tag{3}$$

We will show in next class that, combining equations (2) and (3), we'll get:

$$Q^\star(s,a) \leq r(s,a) + \gamma \cdot \sum_{s' \in \mathcal{S}} P(s' \mid s,a) \max_{a'} Q^\star(s',a').$$