A Short Note on Reinforcement Learning

Tianyu Du

November 25, 2019

1 Notations

- Y^X: The space of all functions $f : X \to Y$.
- $f(\ \cdot\ )$: functions.
- $F[\ \cdot\ ]$: Functionals.
- $\Delta(X)$: The space of all probability distributions over X.

2 Setup

Definition 2.1. A **Markov decision process** is a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$.

- Space of states $\mathcal{S}$;
- Space of actions $\mathcal{A}$, $\mathcal{A}$ is assumed to be finite;
- Transition probability: $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$;
- Immediate reward distribution: $R_t(S_t, S_{t+1}, A) \sim \mathcal{R} : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to \Delta(\mathbb{R})$
- Discount factor: $\gamma \in [0, 1]$.

Definition 2.2. A **policy** π is a mapping from $\mathcal{S}$ to $\mathcal{A}$ (pure strategies) or $\Delta(\mathcal{A})$ (mixed strategies). Let $\Pi \equiv \Delta(\mathcal{A})^{\mathcal{S}}$ denote the collection of all policies.

3 Value Functions

Definition 3.1. The **value function** $V[\cdot] : \Pi \to \mathbb{R}^{\mathcal{S}}$ is defined as

$$V[\pi](s) := \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} | S_t = s \right] \tag{1}$$

Assumption 3.1. Assume $\mathcal{R}$ is deterministic and depends on (S_t, A_t) only:

$$\mathbb{P}[\mathcal{R}(s, S_{t+1}, a) = r(s, a)] = 1\ \forall S_{t+1} \in \mathcal{S} \tag{2}$$

Assumption 3.2 (Markov property)**.** The future depends on the past only through the current state. That is,

$$\mathcal{P}(S_t, A_t) \perp A_t \tag{3}$$

Apply the law of total expectation (LTE),

$$V[\pi](s) := \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} \middle| S_t = s \right] \tag{4}$$

$$= \sum_{a \in \mathcal{A}} \pi(a|s) \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} \middle| S_t = s, A_t = a \right] \text{ (LTE)} \tag{5}$$

$$= \sum_{a \in \mathcal{A}} \pi(a|s) \int_{\mathcal{S}} \mathcal{P}(s'|s,a) \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} \middle| S_t = s, A_t = a, S_{t+1} = s' \right] ds' \text{ (LTE)} \tag{6}$$

$$= \sum_{a \in \mathcal{A}} \pi(a|s) \int_{\mathcal{S}} \mathcal{P}(s'|s,a) \mathbb{E}_\pi \left[r(s,a) + \sum_{k=1}^{\infty} \gamma^k R_{t+k} \middle| S_t = s, A_t = a, S_{t+1} = s' \right] ds' \tag{7}$$

$$= \sum_{a \in \mathcal{A}} \pi(a|s) \int_{\mathcal{S}} \mathcal{P}(s'|s,a) \mathbb{E}_\pi \left[r(s,a) + \gamma \sum_{k=0}^{\infty} \gamma^k R_{t+1+k} \middle| S_{t+1} = s' \right] ds' \tag{8}$$

$$= \sum_{a \in \mathcal{A}} \pi(a|s) \int_{\mathcal{S}} \mathcal{P}(s'|s,a) \left\{ r(s,a) + \gamma \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+1+k} \middle| S_{t+1} = s' \right] \right\} ds' \tag{9}$$

$$= \sum_{a \in \mathcal{A}} \pi(a|s) \left[r(s,a) + \gamma \int_{\mathcal{S}} \mathcal{P}(s'|s,a) V[\pi](s') ds' \right] \tag{10}$$

Similarly, conditioning $V[\pi](s)$ on $A_t = a$ gives

$$Q[\pi](s,a) := \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} \middle| S_t = s, A_t = a \right] \tag{11}$$

$$= \int_{\mathcal{S}} \mathcal{P}(s'|s,a) \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} \middle| S_t = s, A_t = a, S_{t+1} = s' \right] ds' \text{ (LTE)} \tag{12}$$

$$= \int_{\mathcal{S}} \mathcal{P}(s'|s,a) \mathbb{E}_\pi \left[r(s,a) + \gamma \sum_{k=1}^{\infty} \gamma^k R_{t+1+k} \middle| S_t = s, A_t = a, S_{t+1} = s' \right] ds' \tag{13}$$

$$= \int_{\mathcal{S}} \mathcal{P}(s'|s,a) \left\{ r(s,a) + \gamma \mathbb{E}_\pi \left[\sum_{k=1}^{\infty} \gamma^k R_{t+1+k} \middle| S_{t+1} = s', A_{t+1} = \pi(s') \right] \right\} ds' \tag{14}$$

$$= \int_{\mathcal{S}} \mathcal{P}(s'|s,a) \left\{ r(s,a) + \gamma Q[\pi](s', \pi(s')) \right\} ds' \tag{15}$$

$$= r(s,a) + \gamma \int_{\mathcal{S}} \mathcal{P}(s'|s,a) Q[\pi](s', \pi(s')) ds' \tag{16}$$

4 Bellman Operators

Definition 4.1. Define the **Bellman operator** $T[Q,\theta] : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ where $Q \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}$ and $\theta \in \Pi \equiv \Delta(\mathcal{A})^{\mathcal{S}}$ as

$$T[Q,\theta](s,a) := r(s,a) + \gamma \int_{\mathcal{S}} \mathcal{P}(s'|s,a) Q(s', \theta(s')) ds' \tag{17}$$

Definition 4.2. Define the **optimal (pure) policy** $\pi^*(\cdot)$ as

$$\forall s \in \mathcal{S},\ \pi^*(s) \in \operatorname*{argmax}_{a \in \mathcal{A}} Q(s,a) \tag{18}$$

Definition 4.3. Define the **Bellman optimality operator** $T^*[Q]$ as

$$
\begin{aligned}
T^*[Q] &:= T[Q, \pi^*] && (19)\\
&= r(s,a) + \gamma \int_{\mathcal{S}} \mathcal{P}(s'|s,a) Q(s', \pi^*(s')) ds' && (20)\\
&= r(s,a) + \gamma \int_{\mathcal{S}} \mathcal{P}(s'|s,a) \max_{a' \in \mathcal{A}} Q(s', a') ds' && (21)
\end{aligned}
$$

Proposition 4.1. For every $\pi \in \Delta(\mathcal{A})^{\mathcal{S}}$,

$$
T[Q[\pi](\cdot); \pi](s,a) = Q[\pi](s,a) \ \forall (s,a) \in \mathcal{S} \times \mathcal{A} \tag{22}
$$

That is, given each π, $Q[\pi]$ is the fixed point for the corresponding Bellman operator $T[\,\cdot\,;\pi]$.

Proof. Follows the definition of Bellman operator. ■

Theorem 4.1. For every $\pi \in \Delta(\mathcal{A})^{\mathcal{S}}$, the fixed point of the corresponding Bellman operator is unique.

Corollary 4.1. In particular, take $\pi = \pi^*$, then $Q[\pi^*]$ is the unique fixed point for Bellman optimality operator.

Consequently, finding the unique $Q[\pi^*]$ from $T^*[\,\cdot\,]$ would provide sufficient evidence to identify the optimal policy π^*.

Theorem 4.2. For every π, the corresponding Bellman operator is a contraction mapping.

5 Value Iteration

6 Q-Learning: Exploration vs. Exploitation