Notes on Reinforcement Learning
Updated: September 28, 2019

Reinforcement Learning:

> "A RL agent interacts with an environment over time. At each time step t, the agent receives a state s_t in a state space $\mathcal{S}$ and selects an action at from an action space $\mathcal{A}$, following a policy $\pi(a_t|s_t)$, which is the agent's behavior, i.e., a mapping from state s_t to actions a_t, receives a scalar reward r_t, and transitions to the next state s_{t+1}, according to the environment dynamics, or model, for reward function $\mathcal{R}(s,a)$ and state transition probability $\mathcal{P}(s_{t+1}|s_t, a_t)$ respectively. In an episodic problem, this process continues until the agent reaches a terminal state and then it restarts. The return $R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k}$ is the discounted, accumulated reward with the discount factor $\gamma \in (0, 1]$. The agent aims to maximize the expectation of such long term return from each state ..."

1 Markov Decision Process (MDP)

We only consider the fully observable Markov decision process with finite discrete states, with finite discrete actions, and in discrete time steps.

Notations:

- State space $\mathcal{S}$. Every element in $\mathcal{S}$ is called a state. The random process $\{S_t\}$ represents the state at time t.
- Action space $\mathcal{A}$. Every element in $\mathcal{A}$ is called an action. The random process $\{A_t\}$ represents the action taken by agent at time t.
- Reward $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. Reward is a scalar used to measure how well agent is doing by taking action a in state s. The random process $\{R_t\}$ represents the reward received by agent at time t. Moreover, we assume for all t, $R_t \leq R$ for some $R < \infty$.
- History H. The history H_t is defined as all information no later than time t.

Agent: The goal of agent is to gather rewards based on the received information from the environment; for example, we expect the behavior of an agent will maximize the expected, discounted, accumulative reward in the future. An RL agent may include one or more of these components: policy, value function, or model.

- A policy fully defines the agent's behavior.

  Definition 1.1 (Policy)**.** A *deterministic policy* π is a map from $\mathcal{S}$ to $\mathcal{A}$,
  $$\pi : s \mapsto \pi(s).$$
  A *stochastic policy* π is a distribution over actions $\mathcal{A}$ given states,
  $$\pi(a|s) = \mathbb{P}(A_t = a \mid S_t = s).$$
  Remark. We can also define a policy depending on the history, $\pi(a|H) = \mathbb{P}(A_t = a \mid H)$; it is same in the fully observable MDP case due to the Markov property of $\{S_t\}$.

- The value function is a prediction of future reward; it is used to evaluate the goodness/badness of states.

  Definition 1.2 (Value function)**.** A *return* from time-step t with the discount $\gamma \in [0, 1]$ is
  $$G_t := R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}.$$

A *state-value function* $v_\pi : \mathcal{S} \to \mathbb{R}$ w.r.t. π is defined as

$$v_\pi(s) := \mathbf{E}_\pi[G_t \mid S_t = s]$$

A *action-value function* $q_\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ w.r.t. π is defined as

$$q_\pi(s,a) := \mathbf{E}_\pi[G_t \mid S_t = s, A_t = a]$$

The *optimal state-value function* $v_* : \mathcal{S} \to \mathbb{R}$ is defined as

$$v_*(s) := \max_\pi v_\pi(s).$$

The *optimal action-value function* $q_* : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is defined as

$$q_*(s,a) := \max_\pi q_\pi(s,a).$$

Remark. Note that the value function can also be considered as a function of policy π. Therefore, we can define the optimal policy as the policy maximizing the value function.

Definition 1.3 (Optimal Policy)**.** A policy π_* is optimal in $\mathcal{D}$, if for any policy $\pi \in \mathcal{D}$ and for all $s \in \mathcal{S}$,

$$v_{\pi_*}(s) \geq v_\pi(s).$$

- A model predicts what the environment will do next.

  Definition 1.4 (Model)**.** $\mathcal{P}$ is defined as the distribution of next step,

  $$\mathcal{P}^a_{ss'} = \mathbb{P}\left[S_{t+1} = s' \mid S_t = s, A_t = a\right].$$

  $\mathcal{R}$ is defined as the next expected reward,

  $$\mathcal{R}^a_s = \mathbf{E}[R_{t+1} \mid S_t = s, A_t = a].$$

Probability Review: A random process $\{S_t\}_{t\in\mathbb{N}}$ with a finite state space $\mathcal{S}$ is called a Markov process if for every $s, s_1, \ldots, s_t \in \mathcal{S}$,

$$\mathbb{P}(S_{t+1} = s \mid S_1 = s_1, \ldots, S_t = s_t) = \mathbb{P}(S_{t+1} = s \mid S_t = s_t);$$

or equivalently, it could be defined relative to a filtration $\mathcal{F}$,

$$\mathbf{E}(f(S_{t+1}) \mid \mathcal{F}_t) = \mathbf{E}(f(S_{t+1}) \mid S_t)$$

for any measurable function $f : \mathcal{S} \to \mathbb{R}$.
Given a Markov process $\{S_t\}_{t\in\mathbb{N}}$, the state transition probability from s to s' is written as

$$\mathcal{P}_{ss'} = \mathbb{P}\left(S_{t+1} = s' \mid S_t = s\right);$$

It forms the state transition matrix

$$\mathcal{P} = \begin{pmatrix} \mathcal{P}_{11} & \ldots & \mathcal{P}_{1n} \\ \vdots & & \\ \mathcal{P}_{n1} & \ldots & \mathcal{P}_{nn} \end{pmatrix}.$$

Markov Decision Process

There are two similar concepts: Markov reward process (MRP) and Markov decision process (MDP). In the MRP model, the history is considered as the filtration generated by $\{S_t\}$, while in the MDP model, the history is considered as the filtration generated by $\{S_t\}$ and $\{A_t\}$.

Definition 1.5. A *Markov Reward process* is a four tuple $\langle \mathcal{S}, \mathcal{P}, \mathcal{R}, \gamma \rangle$:

- A finite state space $\mathcal{S}$.
- A transition matrix $\mathcal{P}$; that is,

$$\mathcal{P}_{ss'} = \mathbb{P}\left(S_{t+1} = s' \mid S_t = s\right).$$

- A reward function $\mathcal{R} : \mathcal{S} \to \mathbb{R}$ defined as

$$\mathcal{R} : s \mapsto \mathbf{E}[R_{t+1} \mid S_t = s].$$

  where R_{t+1} is the reward at time $t+1$.
- A discount factor $\gamma \in [0, 1]$.

A *Markov decision process* is a five tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$:

- A finite state space $\mathcal{S}$.
- A finite action space $\mathcal{A}$.
- A transition matrix $\mathcal{P}$; that is,

$$\mathcal{P}_{ss'}^{a} = \mathbb{P}\left(S_{t+1} = s' \mid S_t = s, A_t = a\right).$$

- A reward function $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ defined as

$$\mathcal{R} : (s, a) \mapsto \mathbf{E}[R_{t+1} \mid S_t = s, A_t = a].$$

  where R_{t+1} is the reward at time $t+1$.
- A discount factor $\gamma \in [0, 1]$.

Example 1.6 (Policies in MDP)**.** Given a MDP $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$ with policy π, the agent's action will be leaded as follow:

1) Start from time t with an initial state $S_t = s$.
2) Take an action based on the policy: $A_t \sim \pi(\cdot | S_t = s)$.
3) Compute the reward: $(s, a) \mapsto \mathcal{R}(s, a)$.
4) Move to the next state based on the transition kernel: $S_{t+1} \sim \mathbb{P}(\cdot \mid S_t = s, A_t = a)$.

Theorem 1.7 (Bellman Equation)**.** *The state-value function can be decomposed into the sum of immediate reward and discounted value of successor state,*

$$v_\pi(s) = \mathbf{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s].$$

The action-value function can similarly be decomposed,

$$q_\pi = \mathbf{E}_\pi[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1}) \mid S_t = s, A_t = a].$$

Proof. Directly by the definition of $v_\pi(s)$:

$$
\begin{aligned}
v_\pi(s) &= \mathbf{E}_\pi[R_{t+1} + \gamma \sum_{k=1}^{\infty} \gamma^{k-1} R_{t+k+1} \mid S_t = s] \\
&= \mathbf{E}_\pi \left[R_{t+1} + \gamma \mathbf{E}_\pi[\sum_{k=1}^{\infty} \gamma^{k-1} R_{t+k+1} \mid S_{t+1}] \middle| S_t = s \right] \\
&= \mathbf{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s]
\end{aligned}
$$

The action-value case is omitted. □

Remark. There are two other equivalent representations of Bellman expectation equation,

1) The Bellman expectation equation can be written as the matrix form as

$$q_\pi = \mathcal{R}^\pi + \gamma \mathcal{P}^\pi q_\pi$$

with direct solution

$$q_\pi = (I - \gamma \mathcal{P}^\pi)^{-1} \mathcal{R}^\pi.$$

2) Or we can write it more explicitly,

$$q_\pi(s, a) = \mathcal{R}^\pi(s, a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}^a_{ss'} q_\pi(s', a).$$

Optimal Policy

In this part, we will show the existence of optimal value function. Let $\mathcal{D}^{\text{MD}}$ be the space of all deterministic policies and $\mathcal{D}^{\text{MR}}$ be the space of all stochastic polices (of course, $\mathcal{D}^{\text{MD}} \subset \mathcal{D}^{\text{MR}}$). First, we begin from the Bellman optimality equation.

Theorem 1.8 (Bellman Optimality Equation)**.** *Let v_*, q_* be the optimal state-value function and the optimal action-value function, then*

$$
\begin{aligned}
v_*(s) &= \max_a q_*(s, a) \\
q_*(s, a) &= \mathcal{R}^a_s + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}^a_{ss'} v_*(s').
\end{aligned}
$$

Proof. First, we notice that

$$v_\pi(s) = \sum_{a \in \mathcal{A}} q_\pi(s, a) \pi(a|s).$$

If q_* is the optimal value function, then

$$\pi_*(a|s) = \begin{cases} 1 & \text{if } s = \arg\max_{a \in \mathcal{A}} q_*(s, a) \\ 0 & o.w. \end{cases}$$

is the optimal policy (it could be shown by the definition of optimal policy). Then

$$\max_\pi v_\pi(s) = \max_{a \in \mathcal{A}} q_*(s, a).$$

Plug the optimal policy into the Bellman expectation equation, then we can get the second equation. □

Remark. We can also write it as below:

$$v_*(s) = \max_a \Big[\mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v_*(s')\Big];$$
$$q_*(s,a) = \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a \max_{a'} q_*(s',a').$$

This form gives us an anther perspective of the optimal value function; if we define an operator

$$L : v \mapsto \max_a \Big[\mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v(s')\Big],$$

then v_* is a stationary point of this operator (that is, $Lv_* = v_*$). The existence of optimal value function will be immediately implied by the contraction of L.

Let $\mathcal{V}$ be the space of all bounded functionals on $\mathcal{S}$; it is a Banach space with the supremum norm

$$\|v\| := \max_{s \in \mathcal{S}} v(s).$$

The Bellman optimality equation defines an operator $L : \mathcal{V} \to \mathcal{V}$ as

$$L : \ v \mapsto \max_{\pi \in \mathcal{D}^{\mathrm{MD}}} \{\mathcal{R}^\pi + \gamma \mathcal{P}^\pi v\}.$$

More explicitly,

$$\begin{aligned} Lv(s) &= \max_{\pi \in \mathcal{D}^{\mathrm{MD}}} \{\mathcal{R}^\pi + \gamma \mathcal{P}^\pi v\}(s) \\ &= \max_{a \in \mathcal{A}} \{\mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v(s')\}, \end{aligned}$$

where the maximum is attained at

$$a_*^s := \arg\max_{a \in \mathcal{A}} \{\mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v(s')\}.$$

Lemma 1.9. *Suppose that the discount $\gamma \in (0,1)$. Then $L : \mathcal{V} \to \mathcal{V}$ is a contraction operator.*

Proof. Let $u, v \in \mathcal{V}$. Without loss of generality, fix $s \in \mathcal{S}$ such that $Lv(s) \geq Lu(s)$. Then

$$\begin{aligned} Lv(s) - Lu(s) &= \left[\mathcal{R}(s, a_*^s) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^{a_*^s} v(s')\right] - \left[\mathcal{R}(s, a_*^s) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^{a_*^s} u(s')\right] \\ &= \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^{a_*^s} [v(s') - u(s')] \\ &\leq \gamma \|v - u\|. \end{aligned}$$

Therefore, for all $s \in \mathcal{S}$,

$$|Lv(s) - Lu(s)| \leq \gamma \|v - u\|;$$

it implies $\|Lv - Lu\| \leq \gamma \|v - u\|$. ☐

Remark. Under additional technical conditions, this result also holds for more general state spaces.

Now we can prove the fundamental result related to MDP.

Theorem 1.10. *There is always a deterministic optimal policy for the MDP defined in Definition 1.5 with discount $\gamma \in (0,1)$.*

Proof. The proof would be divided into three part:

- Existence and uniqueness of v_*: By Lemma 1.9 and Banach fixed-point theorem, there exists unique $v_* \in \mathcal{V}$ such that
$$Lv_* = v_*;$$
moreover, for any $v \in \mathcal{V}$, the sequence $v_n := L^n v$ converges to v_* in norm.

- Construction of q_*: By Theorem 1.8, the optimal action-value function is
$$q_*(s,a) = \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v_*(s').$$

- Construction of π_*: Define
$$\pi_* : s \mapsto \arg\max_{a \in \mathcal{A}} q_*(s,a);$$
it is the deterministic optimal policy for the given MDP.

$\square$