Reinforcement Learning Notes

Learn from trying!

Author: occupymars

Date: June. 20, 2025

Version: 0.1

Contents

Chapter 1 Reinforcement Learning Basics

Introduction

- ❏ *Markov Decision Process*
- ❏ *Value Function*
- ❏ *Solving Value Function*
- ❏ *Action Value Function*
- ❏ *Bellman Optimality Equation*

1.1 Markov Decision Process

State and ***Action*** can describe a robot state respect to the enviroment and actions to move around, $\mathcal{S}, \mathcal{A}$ are states and actions a robot can take, when taking an action, state after may not be deterministic, it has a probability. We use a transition function $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ to denote this, $T(s, a, s') = P(s' \mid s, a)$ is the probability of reaching s' given s and a. For $\forall s \in \mathcal{S}$ and $\forall a \in \mathcal{A}$, $\sum_{s' \in S} T(s, a, s') = 1$.

Reward $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, $r(s, a)$ depends on current state and action. And the reward may also be stochastic, given state and action, the reward has probability $p(r \mid s, a)$.

Policy $\pi(a \mid s)$ tells agent which actions to take at every state, $\sum_a \pi(a \mid s) = 1$.

This can build a Markov Decision Process, $(\mathcal{S}, \mathcal{A}, \mathcal{T}, r)$ from the ***Trajectory*** $\tau = (s_0, a_0, r_0, s_1, a_1, r_1, s_2, a_2, r_2, \ldots)$, which has probability of:

$$P(\tau) = \pi(a_0 \mid s_0) \cdot P(s_1 \mid s_0, a_0) \cdot \pi(a_1 \mid s_1) \cdot P(s_2 \mid s_1, a_1) \cdots$$

We then define ***Return*** as the total reward $R(\tau) = \sum_t r_t$, the goal of reinforcement learning is to find a trajectory that has the largest return. The trajectory might be infinite, so in order for a meaningful formular of its return, we introduce a discount factor $\gamma < 1$, $R(\tau) = \sum_{t=0}^{\infty} \gamma^t r_t$. For large γ, the robot is encouraged to explore, for small one to take a short trajetory to goal.

Markov system only depend on current state and action, not the history one (but we can always augment the system).

1.2 Value Function

Value Function is the value of a state, from that state, the expected sum reward (return).

The formular of value function is:

$$V^{\pi}(s_0) = \mathbb{E}_{a_t \sim \pi(s_t)}[R(\tau)] = \mathbb{E}_{a_t \sim \pi(s_t)} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right] \tag{1.1}$$

If we divede the trajectory into two parts, s_0 and τ', we get the return:

$$R(\tau) = r(s_0, a_0) + \gamma \sum_{t=1}^{\infty} \gamma^{t-1} r(s_t, a_t) = r(s_0, a_0) + \gamma R(\tau')$$

Put it back into the value function, using law of total expectation:

$$\mathbb{E}[X] = \sum_a \mathbb{E}[X \mid A = a] p(a) = \mathbb{E}_a \left[\mathbb{E}[X \mid A = a] \right]$$

we get:

$$\begin{aligned} V^{\pi}(s_0) &= \mathbb{E}_{a_t \sim \pi(s_t)}[r(s_0, a_0) + \gamma R(\tau')] \\ &= \mathbb{E}_{a_0 \sim \pi(s_0)}[r(s_0, a_0)] + \gamma \mathbb{E}_{a_t \sim \pi(s_t)}[R(\tau')] \\ &= \mathbb{E}_{a_0 \sim \pi(s_0)}[r(s_0, a_0)] + \gamma \mathbb{E}_{a_0 \sim \pi(s_0)} \left[\mathbb{E}_{s_1 \sim P(s_1 \mid a_0, s_0)} [\mathbb{E}_{a_t \sim \pi(s_t)}[R(\tau') \mid s_1, a_0]] \right] \\ &= \mathbb{E}_{a_0 \sim \pi(s_0)}[r(s_0, a_0)] + \gamma \mathbb{E}_{a_0 \sim \pi(s_0)} \left[\mathbb{E}_{s_1 \sim P(s_1 \mid a_0, s_0)} [V^{\pi}(s_1)] \right] \\ &= \mathbb{E}_{a \sim \pi(s)} \left[r(s_0, a_0) + \gamma \mathbb{E}_{s_1 \sim P(s_1 \mid a_0, s_0)} [V^{\pi}(s_1)] \right] \end{aligned} \tag{1.2}$$

before we put s_1 to the right as the condition, it is stochastic, inside the $E_{s_1 \sim p(s_1 \mid s_0, a_0)}$ scope it is deterministic, then we can get $V^{\pi}(s_1)$, as it needs the state to be deterministic.

The discrete formular is (get rid of the notation of time) so called ***Bellman Equation***:

$$V^{\pi}(s) = \sum_{a \in \mathcal{A}} \pi(a \mid s) \left[r(s,a) + \gamma \sum_{s'} P(s' \mid s,a) V^{\pi}(s') \right], \forall s \in S \tag{1.3}$$

And if we write $r(s,a)$ as $\sum_r p(r \mid s,a) r$, then

$$p(r \mid s,a) = \sum_{s' \in \mathcal{S}} p(s', r \mid s,a)$$

We can also get

$$p(s' \mid s,a) = \sum_{r \in \mathcal{R}} p(s', r \mid s,a)$$

combined we get

$$V^{\pi}(s) = \sum_{a \in \mathcal{A}} \pi(a \mid s) \sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s', r \mid s,a) \left[r + \gamma V^{\pi}(s') \right] \tag{1.4}$$

If the reward depend solely on the next state s', then

$$V^{\pi}(s) = \sum_{a \in \mathcal{A}} \pi(a \mid s) \sum_{s' \in \mathcal{S}} P(s' \mid s,a) \left[r(s') + \gamma V^{\pi}(s') \right] \tag{1.5}$$

Let

$$r^{\pi}(s) = \sum_{a \in \mathcal{A}} \sum_{r} p(r \mid s,a) r$$

$$p^{\pi}(s' \mid s) = \sum_{a \in \mathcal{A}} p(s' \mid s,a)$$

rewirte 1.3 into the vector form:

$$V^{\pi} = r^{\pi} + \gamma P^{\pi} V^{\pi} \tag{1.6}$$

where $V^{\pi} = [V^{\pi}(s_1), \ldots, V^{\pi}(s_n)]^{\top} \in \mathbb{R}^n$, $r^{\pi} = [r^{\pi}(s_1), \ldots, r^{\pi}(s_n)]^{\top} \in \mathbb{R}^n$, and $P^{\pi} \in \mathbb{R}^{n \times n}$ with $P^{\pi}_{ij} = p^{\pi}(s_j \mid s_i)$.

1.3 Solving Value Function

Next, we need to solve the value function, first way is closed-form solution:

$$V^{\pi} = (I - \gamma P^{\pi})^{-1} r^{\pi}$$

Some properties: $I - \gamma P^{\pi}$ is invertible, $(I - \gamma P^{\pi})^{-1} \geq I$ which means every element of this inverse is nonnegative. For every vector $r \geq 0$, it holds that $(I - \gamma P^{\pi})^{-1} r^{\pi} \geq r \geq 0$, so if $r_1 \geq r_2$, $(I - \gamma P^{\pi})^{-1} r_1^{\pi} \geq (I - \gamma P^{\pi})^{-1} r_2^{\pi}$

However, this method need to calculate the inverse of the matrix, that need some numerical algorithms. We can use a iterative solution:

$$V_{k+1} = r^{\pi} + \gamma P^{\pi} V_k$$

as $k \to \infty$, $V_k \to V^{\pi} = (I - \gamma P^{\pi})^{-1} r^{\pi}$.

1.4 Action Value Function

Similarly to value funtion, ***Action Value Function*** is the value of an action at state s, from that state, take that action, the expected sum reward (return). We use $V^{\pi}(s)$ to denote value function, and $Q^{\pi}(s,a)$ to denote action value, their connection is:

$$V^{\pi}(s) = \sum_{a \in \mathcal{A}} \pi(a \mid s) Q^{\pi}(s,a) \tag{1.7}$$

The action value function is given as:

$$\begin{aligned} Q^{\pi}(s_0, a_0) &= r(s_0, a_0) + \mathbb{E}_{a_t \sim \pi(s_t)} \left[\sum_{t=1}^{\infty} \gamma^t r(s_t, a_t)\right] \\ &= r(s_0, a_0) + \gamma \mathbb{E}_{a_t \sim \pi(s_t)} \left[\sum_{t=1}^{\infty} \gamma^{t-1} r(s_t, a_t)\right] \\ &= r(s_0, a_0) + \gamma \mathbb{E}_{a_t \sim \pi(s_t)}[R(\tau')] \\ &= r(s_0, a_0) + \gamma \mathbb{E}_{s_1 \sim p(s_1 | s_0, a_0)} \left[\mathbb{E}_{a_t \sim \pi(s_t)}[R(\tau') \mid s_1]\right] \\ &= r(s_0, a_0) + \gamma \mathbb{E}_{s_1 \sim p(s_1 | s_0, a_0)}[V^{\pi}(s_1)] \\ &= r(s_0, a_0) + \gamma \mathbb{E}_{s_1 \sim p(s_1 | s_0, a_0)} \left[\sum_{a_1 \in \mathcal{A}} \pi(a_1 \mid s_1) Q^{\pi}(s_1, a_1)\right] \end{aligned} \tag{1.8}$$

Then the bellman equation of action value is:

$$\begin{aligned} Q^{\pi}(s, a) &= r(s, a) + \gamma \sum_{s'} P(s' \mid s, a) V^{\pi}(s') \\ &= r(s, a) + \gamma \sum_{s'} P(s' \mid s, a) \sum_{a' \in \mathcal{A}} \pi(a' \mid s') Q^{\pi}(s', a') \end{aligned} \tag{1.9}$$

Note that we can always write $r(s, a)$ as $\sum_r p(r \mid s, a) r$ if it is stochastic, and it follows the same notation in the book *Math of Reinforcement Learning*.

Rewrite 1.9 into vector form:

$$Q^{\pi} = \tilde{r}^{\pi} + \gamma P^{\pi} \Pi^{\pi} Q^{\pi} \tag{1.10}$$

where $\tilde{r}^{\pi}_{(s,a)} = \sum_r p(r \mid s, a) r$, $P^{\pi}_{(s,a),s'} = p(s' \mid s, a)$, $\Pi_{s',(s',a')} = \pi(a' \mid s')$.

1.5 Bellman Optimality Equation

Bellman Optimality Equation is given by:

$$\begin{aligned} V(s) &= \max_{\pi(s) \in \Pi(s)} \sum_{a \in \mathcal{A}} \pi(a \mid s) \left[\sum_r p(r \mid s, a) r + \gamma \sum_{s'} P(s' \mid s, a) V^{\pi}(s')\right] \\ &= \max_{\pi(s) \in \Pi(s)} \sum_{a \in \mathcal{A}} \pi(a \mid s) Q(s, a) \end{aligned} \tag{1.11}$$

Chapter 2 From LQR to RL

Introduction

- ❏ *LQR Problem*
- ❏ *iLQR and DDP*
- ❏ *Reinforcement Learning*

2.1 LQR and Value function

Given a linear model $x_{k+1} = A_k x_k + B_k u_k + C_k$.