Reinforcement Learning

Muchang Bahng

Spring 2024

Contents

The following sources were used to make these notes.

1. David Silver's Google Deepmind Lectures on Reinforcement Learning

1 Utility Theory

Axioms of utility theory. $\prec, \succ$ indicates a preference for.

1. *Orderability.*

2 Markov Decision Processes

MDPs are the mathematical relation of decision theory. Reinforcement learning techniques to challenges of MDPs with numerous or unknown parameters. Though some researchers talk about reinforcement learning as a bigger thing.

Definition 2.1 (State Space)

The formulations for reinforcement learning is very different from that of supervised learning. It is formulated as follows.

1. We start off with an agent, who's state lives in a finite set $\mathcal{S}$ of states.
2. At each discrete time step t, the agent can take some action $a_t \in \mathcal{A}$. Let $\mathcal{A}$ be the set of all actions and $\mathcal{A}(s_t)$ be the set of all actions available in state s_t.
3. There are probabilities $P_{a,s,s'}$ that determine the probability of transitioning to state s' from state s after taking action a. It must be normalized, so the transition probabilities satisfy
$$\sum_{a \in \mathcal{A}(s), s'} P_{a,s,s'} = 1 \text{ for all } s \in \mathcal{S} \tag{1}$$
4. There is also a reward function $R(s,a) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ for taking action a in state s.

Note that this is really just a directed graph with some transition probability matrix, and therefore the tools of Markov chains can be applied here. We restate some common definitions. Sometimes, the reward function can output to $\mathbb{R}^n$, which corresponds to a *multi-objective* function that tries to maximize multiple rewards.

Example 2.1 (Autonomous Helicopter Flight)

To define an MDP for autonomous helicopter flight, we have

1. The state space $\mathcal{S}$ is the set of all possible positions and orientations of the helicopter, which is isomorphic to $\text{Tran}\mathbb{R}^3 \times \text{SO}(3)$.
2. The set of actions A represents the set of all possible configurations of the helicopter. Assuming that we have 3 switches that we can flick on and off, 2 meters that can be moved from 0 to 10, and two control sticks that can move 360 degrees, the configuration of all controls is really the space:
$$\mathcal{A} = \{0,1\}^3 \times [0,10]^2 \times \text{SO}(2)^2 \tag{2}$$

Definition 2.2 (Terminal State)

A state is **terminal** if it only transitions to itself and yield 0 reward.

Definition 2.3 (Trajectory)

At each time step t, the agent observes state $s_t \in \mathcal{S}$, takes action $a_t \in \mathcal{A}(s_t)$, and receives a reward $r_t = R(s_t, a_t) \in \mathbb{R}$. The environment, in turn, transitions to s_{t+1} with probability $P_{a_t, s_t, s_{t+1}}$. This sequence of states, actions, and rewards is called a **trajectory**.

$$\tau = (s_1, a_1, r_1), (s_2, a_2, r_2), \dots, (s_T, a_T, r_T) \tag{3}$$

A trajectory that begins in a starting state s_1 and ends in a terminal state is called an **episode**.

Definition 2.4 (Return)

The return should be defined as the sum of all rewards that the agent gets within an episode, but there a discount factor that comes with time. Therefore, given a discounting factor $\gamma \in [0, 1)$, the **discounted cumulative return** of a trajectory is the vector

$$G(\tau) = \sum_{t=1}^{\infty} \gamma^{t-1} R(s_t, a_t) \tag{4}$$

Since the reward at timestep t is discounted by a factor of γ^t, we would like to accrue positive rewards as soon as possible.

So far, we have not defined how the agent should act given that it is on state s_t. This is defined by the *policy*, which can be arbitrary defined stochastic process (i.e. a sequence of random variables) and does not need to be Markov. It just needs to depend on the trajectory up until that state.

Definition 2.5 (Policy)

A **policy** is a family of conditional measures over $\mathcal{A} \times \mathcal{S}^\infty$ that tells the agent what action to take given its current state S_t and the trajectory τ (the history prior to reaching state s).

$$\mathbb{P}_\pi(A_t = a \mid S_t = s, \tau) = \mathbb{P}_\pi(A_t = a \mid S_t = s, S_{t-1} = s_{t-1}, \dots, S_1 = s_1) \tag{5}$$

Definition 2.6 (Stationary Policy)

However, if it is Markov, then it is called a *stationary policy*, which depends only on the current state. Therefore, π is defined over $\mathcal{A} \times \mathcal{S}$.

$$\mathbb{P}_\pi(A_t = a \mid S_t = s) \tag{6}$$

With this policy in place, we should define some sort of total reward function.

Definition 2.7 (Value Function)

The **value function** is the expected total reward starting from s and following policy π.

$$V_\pi(s) \coloneqq \mathbb{E}_{\tau \sim \pi}[G(\tau) \mid S_1 = s] \tag{7}$$

The **action value function** is the same thing but we fix an action $A_1 = a$.

$$Q_\pi(s, a) \coloneqq \mathbb{E}_{\tau \sim \pi}[G(\tau) \mid S_1 = s, A_1 = a] \tag{8}$$

Therefore,

$$V_\pi(s) = \mathbb{E}_{A_1}[Q_\pi] = \sum_{a \in \mathcal{A}(s)} \pi(a \mid s) Q_\pi(s, a) \tag{9}$$

3 Multi Armed Bandits