Chapter 3 - Markov Process

David Ding

October 24, 2024

1 Markov Process

In previous sections, the idea of state, reward, and policy has already been introduced into the scope of the RL learning. Here is a quick review:

- **State:** A state is kind of like the depiction of all the information the agent and environment carrys. The environment state is the private representation of environment like how much rewards will be given to the agent or what might be observed in the agent next observation, often denoted as S_t^e. The agent state is the information the agent has about the environment, denoted as S_t^a. Generally, the state is A WAY that captures all the history including the environment or its actions or its received rewards, $S_t = f(H_t)$. In Markov process, we would like to draw the equivalence between the state and the history, $S_t^a = S_t^e = S_t$.
- **Reward:** Rewards are the direct feedback from the environment to the agent. When the agent did an action A_t in a state S_t, it will receive a reward R_{t+1} from the environment at time $t+1$. Most of the time, we would not simply consider the immediate reward, but the **return**, which is a exponentially discounted cumulative rewards updating from the time when this action is done to the end of the episode.
- **Policy:** A policy π is a preference of choosing a certain action A_t depending ONLY on the certain state S_t where the agent stands. Policy could be divided mainly into two types: deterministic policy and stochastic policy.

A **Markov process**, aka Markov chain, is a memoryless random process. It is a tuple of $< S, P >$, where S is a finite set of states that satisfies the Markov property, and $P_{s,s'} = \mathbf{P}(S_{t+1} = s^{'} | S_t = s)$ is the probability transitioning matrix from state s to state $s^{'}$.

- **Markov Property:** The Markov Property actually means that the future is **independent** of the past given present, and its mathematical expression is:

$$P[S_{t+1}|S_t] = P[S_{t+1}|S_1, S_2, ..., S_t]$$

- **Transition Matrix($\mathbf{P}_{s,s'}$):** The transition matrix is a matrix that indicate how the current state S_t would like to transform into a new stateS_{t+1}. Because the markov process does not has any agent interaction in the whole process, it could be seen as how the **environment** would like to move after a time step.

$$\mathbf{P}_{s,s'} = P(S_{t+1} = s^{'} | S_t = s)$$

Notes: Given a Markov Process, we could easily sample some trajectories or episodes using the transition matrix.

While the markov process has nothing to do with the rewards(Added in MRP), and the agent's decision making(Policy-Added in MDP), thus it is just a description of how states transit ONLY.

2 Markov Reward Process

A **Markov Reward Process** is a tuple of $< S, P, R, \gamma >$, where S is a finite set of states, P is the transition matrix, R is the reward function given a certain state the agent enters into, and γ is the discount factor that is vital in calculating the long-turn **return**.

- **Rewards(R_s):** Rewards here are the IMMEDIATE feedback from the environment to the agent. It only depends on the state the agent enters into, $R_s = E[R_{t+1}|S_t = s]$.
- **Return(G_t):** The return G_t is the total discounted rewards from time-step t.

  $$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + ... = \sum_{k=1}^{\text{End of Episode}} \gamma^k R_{t+k}$$

  The colored part is the balanced future rewards with immediate rewards, and the discounted factor is often used to prevent inflation of the return.

- **Value Function:** The value function $v(s)$ is the **expected return** starting from state s, $v(s) = E[G_t|S_t = s]$. The value function is a way to evaluate the goodness of a state.
- **Bellman Equation:** The Bellman equation is a recursive equation that could be used to calculate the value function starting from a immediate reward received at the starting state and average all the possible future rewards(How the wind of the environment blows you to the future). The Bellman equation for MRP is:

  $$\begin{aligned} v(s) &= E[G(t)|S_t = s] \\ v(s) &= E[R_{t+1} + \gamma G_{t+1}|S_t = s] \\ v(s) &= E[R_{t+1} + \gamma E[G_{t+1}]|S_t = s] \\ v(s) &= E[R_{t+1} + \gamma v(S_{t+1})|S_t = s] \\ v(s) &= R_s + \gamma \sum_{s' \in S} v(s^{'}) P^a_{s,s'} \end{aligned}$$

  And the matrix form of the Bellman equation is:

  $$v = R + \gamma \mathbf{P} v \qquad (1)$$

Notes: The Markov Reward Process adds the concept of rewards into the Markov process and derives some elements like return based on the rewards.

While, the key point of MRP remains same with MRP that the agent's decision making is not considered and what we calculate for(Return or Value function) is just showing the general **goodness of a state** when the agent is exposed to the wind of the environment.

3 Markov Decision Process

3.1 Part1: Expectation MDP

A **Markov Decision Process** is a tuple of $< S, A, P, R, \gamma >$, where S is a finite set of states, A is a finite set of actions, P is the transition matrix, R is the reward function, and γ is the discount factor.

- **Policy and Action($\pi(s)$):** Policy reflects how the agent interacts with the environment given a state in a finite set of states. There are two types of policies: deterministic policy and stochastic policy.

  The stochastic policy is often written as a probabilty function$\pi(a|s) = P(A_t = a|S_t = s)$, while the deterministic policy is a mapping $\pi(S_t = s) = A_t$.

- **Transition Matrix:** $P^{a}_{s,s'}$ The transition matrix in MDP is slightly different from the previous Markov Process or MRP, as it is now being influcenced by the agent's action. Each action of the agent would lead to a different transition matrix, moving the agent into another state.

- **Reward Function:** $R(s,a)$ Similar to the Transition matrix, the reward function of the MDP has considered the the cost of the agent action and the direct rewards from the environment. Thus, reward takes the form of $R(s,a) = E[R_{t+1}|S_t = s, A_t = a]$
- **Action-Value Function($Q^\pi(s,a)$):** The action value function is the expected return starting not only from the state the agent lives in, but the action the agent has taken. Also, due to the different policy the agent choose, the expected return diversed. The formula could be expressed as:

  $$q^\pi(s,a) = R(s,a) + \gamma \sum_{s^{'} \in S} P^\pi(s^{'}|s,a)V^\pi(s^{'})$$

  It is the sum of the immediate reward from the reward function and the future expected return after the agent enters into the next state with the action it has made [**averaging through $\mathbf{P}^\pi(s,a)$**].
- **Value Function($V^\pi(s)$):** $V^\pi(s)$ is then the expected return averaging through all actions the agent might take from its policy.

  $$V^\pi(s) = \sum_{a \in A} \pi(a|s)q^\pi(s,a)$$
- **Bellman Expectation Equation:** The Bellman expectation equation without demanding the agent to take the best action could be represented as a recursive form taking account of the policy, the transition matrix and the possible rewards when the agent takes the action.

  $$\begin{cases} V^\pi(s) = \sum_{a \in A} \pi(a|s)(R(s,a) + \gamma \sum_{s^{'} \in S} P^\pi(s^{'}|s,a)V^\pi(s^{'})) \\ q^\pi(s,a) = \sum_{s^{'} \in S}(R(s,a) + \gamma \sum_{a \in A} \pi(a^{'}|s^{'})q^\pi(s^{'},a^{'})) \end{cases} \tag{2}$$

3.2 Part2: Optimal MDP

Notes: