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. There are two types of policies: deterministic policy and stochastic policy. The stochastic policy is often written as a probabilty function $\pi(s) = P(A_t|S_t = s)$, while the deterministic policy is a mapping $\pi(S_t = s) = A_t$.

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**
- **Transition Matrix**$P_{s,s'} = \mathbf{P}(S_{t+1} = s^{'}|S_t = s)$

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 the states transition 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 takes the form of expectation, $R_s = E[R_{t+1}|S_t = s]$, and the use of the subscipt of s means that the reward is only dependent on the state the agent is in.
- **Return:** 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

Notes:

3 Markov Decision Process

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:** $\pi(s)$
- **Transition Matrix:** $P^a_{s,s'}$
- **Reward Function:** R^a_s
- **Action-Value Function:** $Q^\pi(s, a)$
- **Value Function:** $V^\pi(s)$
- **Bellman Expectation Equation:**

Notes: