Chapter 1 Basic Concepts

State, Action, State Transition, Reward, Policy, Trajectory, Return, Episode, Discounted Return, Model, Markov Property

1. Explain the process the agent interacting with the environment. Should use concepts including state, action, state transition and reward.
2. Is the action space the same in different states?
3. What attributes should be designed by human in a RL prob?
4. Explain how State Transition can be stochastic.
5. Explain the difference between Policy and State Transition.
6. What concepts can be described in a stochastic math format?
7. Why we introduce Return?
8. Why should we use Discounted Return in an infinite prob?
9. How discounted rate influent the behavior of the agent? Can we just set discounted rate very close to 1 so that the agent can consider very long-run benefit?
10. How to convert finite process to infinite process?
11. Explain the Markov property in natural languange and math format.
12. What is model?
13. What really counts when design the Reward? Is it possible if all rewards are set negative?
14. What determine Reward? Is Reward also influenced by s', which is the next state after taking an action? What is the expression of Reward if we forcibly put s' into consider?

1. From state s, take action a according to the policy pi, get a reward r according to s&a, then step into s' according to the state transition.
2. Different states may have different action space.
3. The elements can be divided into 3 categories, agent-related, env-related and model-related. We should design env-related and model-related elements including state space, action space in different state, state transition probabilities, reward probabilities. (First 2 env-related, last 2 model-related)
4. From the same (s, a), may end into different s'.
5. State Transition is static, is the probabilities to show which state can go, is the map from state, action to state. Policy is the probabilities to define the action to take in this state, is the map from state to action. The relationship is, considering policy first to take an action, then consider State Transition probabilities to define which s' to go to.
6. Action(described by policy), State Transition, Reward(described by condition probablities)
7. Return evaluates whether the policy is good.
8. First, infinite prob may result into diverged return(infinite return). Second, discounted rate can make the agent more short-sighted(->0) or long-sighted(->1). Third, it can help agent avoid meaningless detour(see details in chapter3).

9. The smaller γ is, the short-sighted it is. But we cannot just set γ very very close to 1 because it will render a very slow convergence rate.
10. The first method is continuously stay in the terminal state. The second method is treat terminal state as a normal state but as terminal state has larger return the agent will eventually stay in it either.
11. Which s' to go and r to get only depend on current state and action, instead of considering previous states and actions.

$$p(s_{t+1}|s_t, a_t, s_{t-1}, a_{t-1}, \ldots, s_0, a_0) = p(s_{t+1}|s_t, a_t),$$
$$p(r_{t+1}|s_t, a_t, s_{t-1}, a_{t-1}, \ldots, s_0, a_0) = p(r_{t+1}|s_t, a_t),$$

12. p(s'|s, a) p(r'|s, a) consist of the model.
13. The relative value counts.
14. Only and just s, a that determine the return, not s'. If consider s', then

$$p(r|s,a) = \sum_{s'} p(r|s,a,s')p(s'|s,a)$$