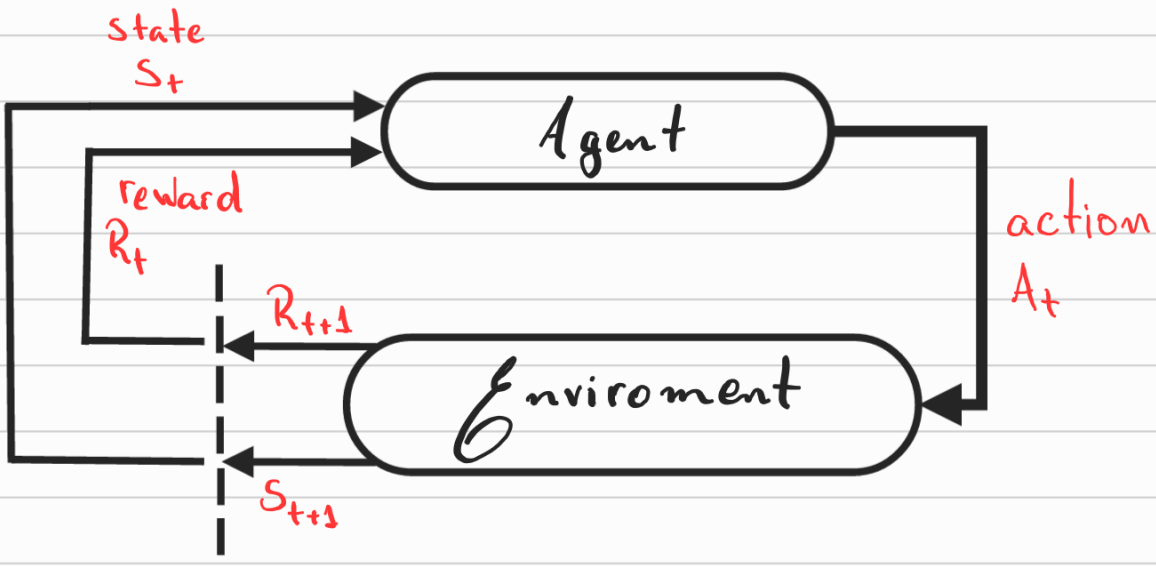
Finite Markov Decision Processes

1. The Agent-Enviroment Interface



- The agent and enviroment interact at each of a seguence of discrete time steps.
- At each time step t, the agent receives some representation of the enviroment's state S_t, and on that basis selects an action, A_t.
- One step later, the agent receives a numerical reward, R_{t+1}, and finds itself in a new state, S_{t+1}.

- The MDP and agent together thereby give rise to a sequence or trajectory:

$$S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, S_3, \ldots$$

- In a finite MDP, the sets of states, actions, and rewards all have a finite number of elements

- In this case, the random variables R_t and S_t have well defined discrete probability distributions dependent only on the preceding state and action:

$$p(s', r|s,a) \doteq Pr\{S_t = s', R_t = r | S_{t-1} = s, A_{t-1} = a\}$$

and

$$\sum_{s' \in S} \sum_{r \in R} p(s', r|s,a) = 1, \quad \forall\, s \in S, a \in A$$

- In a Markov decision process, the probability of each possible value for S_t and R_t depends on the immediatly preceding state and action, S_{t-1} and A_{t-1}, and, given them, not at all on earlier states and actions.

  → The state must include information about all aspects of the past agent-enviroment interaction that make a difference for the future.

- State-transition probabilities:

$$p(s'|s,a) \doteq Pr\{S_t = s' | S_{t-1} = s, A_{t-1} = a\}$$

- Expected rewards for state-action pairs:

$$r(s,a) \doteq \mathbb{E}[R_t | S_{t-1} = s, A_{t-1} = a] = \sum_{r \in R} r \sum_{s' \in S} p(s', r|s,a)$$

- Expected rewards for state-action-next-state triples

$$r(s,a,s') \doteq \mathbb{E}[R_t | S_{t-1} = s, A_{t-1} = a, S_t = s'] = \sum_{r \in R} r \frac{p(s', r|s,a)}{p(s'|s,a)}$$

2. Returns and Episodes

- We seek to maximize the expected return, where the return, denoted G_t, is defined as some specific function of the reward sequence. In the simplest case:

$$G_t \doteq R_{t+1} + R_{t+2} + \cdots + R_T$$

where T is a final step.

- This approach makes sense in applications in which there is a natural notion of final time step, that is, when the agent-environment interaction breaks naturally into subsequences, which we call episodes.

- Each episode ends in a special state called the **terminal state**, followed by a reset to a standard starting state or to a sample from a standard distribution of starting states.

- The additional concept needed is that of **discouting**. According to this approach, the agent tries to select actions so that the sum of the discounted rewards it receives over the future is maximized. In particular, it chooses A_t to maximize the expected **discounted return**:

$$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{K=0}^{\infty} \gamma^K R_{t+K+1}$$

where γ is the discount rate, $0 \leq \gamma \leq 1$.

- The discount rate determines the present value of future rewards

- Returns at successive time steps:

$$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots$$

$$= R_{t+1} + \gamma (R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} + \cdots)$$

$$\therefore \quad G_t = R_{t+1} + \gamma G_{t+1}$$

- Expected discounted return will be always finite. Once we have $0 < \gamma < 1$, let R_{max} be the maximum reward:

$$G_t = \sum_{K=1}^{\infty} \gamma^{K-1} R_{t+K} \leq \sum_{K=1}^{\infty} \gamma^{K-1} R_{max} = R_{max} \frac{1}{1-\gamma} < \infty$$

4. Policies and Value Functions

- Almost all reinforcement learning algorithms involve estimating value functions that estimate how good it is for the agent to be in a given state (or how good it is to perform a given action in a given state).
- A **policy** is a mapping from states to probabilities of selecting each possible action.
- The **value function** of a state s under a policy π, denoted $v_\pi(s)$, is the expected return when starting in s and following π thereafter.
- For MDPs, we can define v_π formally by:

$$v_\pi(s) \doteq \mathbb{E}_\pi[G_t | S_t = s] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s\right], \forall s \in \mathcal{S}$$

where $\mathbb{E}_\pi[\cdot]$ denotes the expected value of a random variable given that the agent follows policy π

- We call the function v_π the **state-value function** for policy π.

- The value of taking action a in state s under a policy π, as the expected return starting from s, taking the action a, and thereafter following policy π:

$$q_\pi(s,a) \doteq \mathbb{E}_\pi[G_t | S_t = s, A_t = a] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s, A_t = a\right]$$

- We call q_π the **action-value function** for policy π

- Recursive relationship between the value of s and the value of its possible successor states in the state-value function:

$$v_\pi(s) \doteq \mathbb{E}_\pi[G_t \mid S_t = s]$$

$$= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s]$$

$$= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r \mid s, a)\left[r + \gamma \mathbb{E}_\pi[G_{t+1} \mid S_{t+1} = s']\right]$$

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s', r \mid s, a)\left[r + \gamma v_\pi(s')\right], \quad \forall s \in S$$

↳ Bellman equation for v_π

5. Optimal Policies and Optimal Value functions

- A policy π is defined to be better than or equal to a policy π' if its expected return is greater than or equal to that of π' for all states.
- In other words, $\pi \geq \pi' \leftrightarrow v_\pi(s) \geq v_{\pi'}(s), \forall s \in S$.
- We denote all the optimal policies by π_*, and they share the same state-value function, called optimal state-value function, denoted v_* and defined as:

$$v_*(s) \doteq \max_\pi v_\pi(s), \quad \forall s \in S$$

- Optimal policies also share the same optimal action-value function, denoted q_*, and defined as:

$$q_*(s,a) \doteq \max_\pi q_\pi(s,a), \forall s \in S$$

- In terms of v_*:

$$q_*(s,a) = \mathbb{E}[R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a]$$

- Bellman optimality equation:

$$v_*(s) = \max_{a \in A(s)} q_{\pi_*}(s,a)$$

$$= \max_a \mathbb{E}[G_t \mid S_t = s, A_t = a]$$

$$= \max_a \mathbb{E}[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a]$$

$$= \max_a \mathbb{E}[R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a]$$

$$v_*(s) = \max_a \sum_{s',r} p(s',r \mid s,a)[r + \gamma v_*(s')]$$

and

$$q_*(s,a) = \mathbb{E}[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a') \mid S_t = s, A_t = a]$$

$$q_*(s,a) = \sum_{s',r} p(s',r \mid s,a)[r + \gamma \max_{a'} q_*(s',a')]$$