Specifying policies

A policy is a strategy used by an agent to decide what action to take in the given state. It essentially defines the agent's behavior at any point of time

A policy is a function that maps states to actions:

- Deterministic policy: $\pi(s) = a$, where for every state it returns a specific action a
- Stochastic policy: $\pi(a|s) = P(a|s)$, a probability distribution over actions given a state

Policies **only depend on the current state, not on other factors** like time or the previous states.

Value functions

A state-value function is the future reward that an agent can expect starting from a particular state.

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

An action-value function describes what happens when an agent first selects the particular action.

$$q_\pi(s,a) \doteq \mathbb{E}_\pi[G_t|S_t = s, A_t = a]$$

Value functions are crucial for reinforced learning because they allow the agent to query the quality of the current situations instead of waiting to observe the long-term outcome:

- The return is not immediately available
- The return can be random due to stochasticity in both the policy and environment dynamics

Bellman Equation

1. **State-value Bellman equation**

$$\begin{aligned} v_\pi(s) &\doteq \mathbb{E}_\pi[G_t|S_t = s] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1}|S_t = s] \\ &= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s',r|s,a)\left[r + \gamma \mathbb{E}[G_{t+1}|S_{t+1} = s']\right] \\ &= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s',r|s,a)\left[r + \gamma v_\pi(s')\right] \end{aligned}$$

2. **Action-value Bellman equation**

$$\begin{aligned} q_\pi(s,a) &\doteq \mathbb{E}[G_t | S_t = s, A_t = a] \\ &= \sum_{s'} \sum_{r} p(s',r|s,a) \big[r + \gamma \mathbb{E}_\pi [G_{t+1} | S_{t+1} = s'] \big] \\ &= \sum_{s'} \sum_{r} p(s',r|s,a) \left[r + \gamma \sum_{a'} \pi(a'|s') \, \mathbb{E}_\pi [G_{t+1} | S_{t+1} = s', A_{t+1} = a'] \right] \\ &= \sum_{s'} \sum_{r} p(s',r|s,a) \left[r + \gamma \sum_{a'} \pi(a'|s') \, q_\pi(s',a') \right] \end{aligned}$$

Optimal policies

An optimal policy π_* is as good as or better than all the other policies

To be an optimal policy, it must have its value as high as or higher than the other policies all the time

The value function of the state under the policy is the expected return when starting in s and following π thereafter

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

The value function of taking action a in a state s under a policy π is an expected return starting from s, taking the action a, and thereafter the policy π

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

Optimal Value Functions

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

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