Reinforcement Learning

Value Function

Dr. Alireza Aghamohammadi

Policies

- In reinforcement learning, the goal of the agent is to learn a **policy**.
- A policy, denoted as π, is a strategy or plan that determines the action the agent takes based on its current state.
- Policies are comprehensive plans covering all possible states:
  - **Deterministic Policy**: Returns a specific action for each state.
  - **Stochastic Policy**: Returns a probability distribution over possible actions for each state.
- The agent aims to find an optimal policy that prescribes the best actions for all non-terminal states.

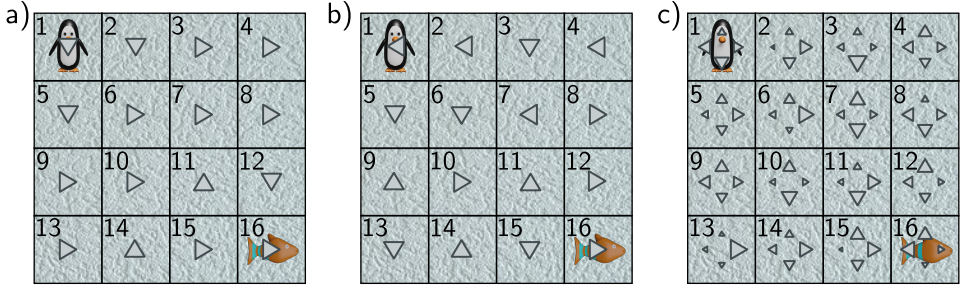

Image credit: Understanding Deep Learning

The Agent-Environment Interaction Loop

- At time step t, the agent observes the current state s_t and selects an action a_t based on the policy $\pi\left[a_t \mid s_t\right]$.
- The environment then transitions to a new state s_{t+1} according to the state transition function and generates a reward r_{t+1} based on the reward function.
- The new state s_{t+1} and reward r_{t+1} are fed back to the agent, which uses them to decide the next action.

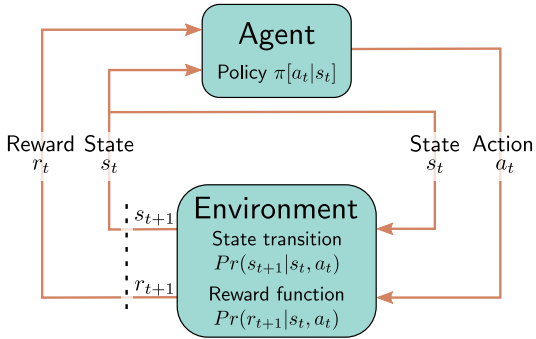

Image credit: Understanding Deep Learning

The Value Function

- An important question to ask when analyzing a policy is: **How good is this policy?**
- If we can assign a numerical value to policies, we can compare how much better one policy is compared to another.
- Given a policy π and the Markov Decision Process (MDP), we can compute the expected return starting from any state.
- We define the value of a state s under a policy π: It is the expected return when the agent starts from state s and follows policy π thereafter.
- The value function $V_\pi(s)$ is defined as:

$$\begin{aligned} V_\pi(s) &= \mathbb{E}_\pi \left[G_t \mid S_t = s \right] \\ &= \mathbb{E}_\pi \left[R_{t+1} + \gamma G_{t+1} \mid S_t = s \right] \\ &= \sum_a \pi \left[a \mid s \right] \sum_{s', r} P \left[s', r \mid s, a \right] \left[r + \gamma V_\pi(s') \right], \qquad \forall s \in S \end{aligned}$$

- This is known as the **Bellman equation**.
- It describes the expected long-term reward when starting in a given state and following the specified policy.