Reinforcement Learning

Mathematical Foundation

Dr. Alireza Aghamohammadi

Transition Function

- ❖ We need a way to model the environment and its dynamics.
- ❖ The agent interacts with the environment through actions, which may cause the environment to change states.
- ❖ The function that describes this state change is called the **transition function** $T(s, a, s')$.
- ❖ The transition function takes a current state s, an action a, and a next state s', returning the probability of moving from s to s' after taking action a:

$$T : S \times A \times S \to [0, 1]$$

- ❖ Formally, the transition function is defined as the probability of reaching state s' at time t, given the state s and action a at time $t-1$:

$$P(S_t = s' \mid S_{t-1} = s, A_{t-1} = a)$$

- ❖ In many Reinforcement Learning (RL) algorithms, we assume this transition probability is **stationary**, meaning it does not change during training or evaluation.

Reward Function

- The environment may provide a reward signal as feedback to the agent's actions.
- The **reward function** $R(s, a)$ assigns a numerical value indicating the desirability of a state-action pair:

$$R : S \times A \to \mathbb{R}$$

- The reward function can be defined as the expected reward at time step t, given the state s and action a at time $t - 1$:

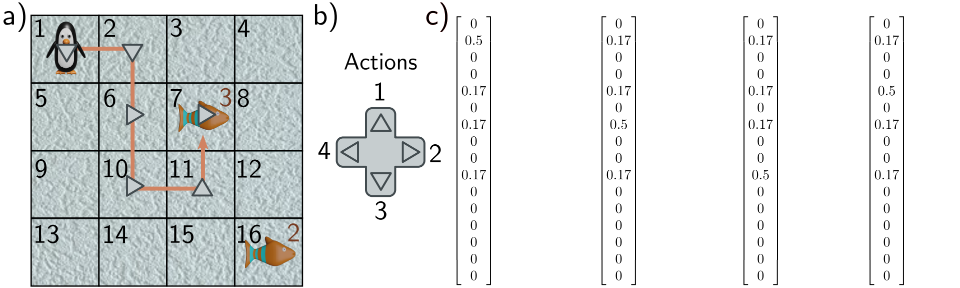
$$r(s, a) = \mathbb{E}\left[R_t \mid S_{t-1} = s, A_{t-1} = a\right]$$

- Together, the transition function and the reward function form the **model of the environment**.

Markov Decision Process (MDP)

- In RL, we often model the environment as a **Markov Decision Process (MDP)**.
- An MDP assumes that the probability of transitioning to the next state depends only on the current state and action, not on the full history of past states and actions. This is known as the **Markov Property**:

$$P(S_{t+1} \mid S_t, A_t) = P(S_{t+1} \mid S_t, A_t, S_{t-1}, A_{t-1}, \dots)$$



$$\begin{bmatrix}0\\0.5\\0\\0\\0.17\\0\\0.17\\0\\0\\0.17\\0\\0\\0\\0\\0\\0\end{bmatrix} \quad \begin{bmatrix}0\\0.17\\0\\0\\0.17\\0\\0.5\\0\\0\\0.17\\0\\0\\0\\0\\0\\0\end{bmatrix} \quad \begin{bmatrix}0\\0.17\\0\\0\\0.17\\0\\0.17\\0\\0\\0.5\\0\\0\\0\\0\\0\\0\end{bmatrix} \quad \begin{bmatrix}0\\0.17\\0\\0\\0.5\\0\\0.17\\0\\0\\0.17\\0\\0\\0\\0\\0\\0\end{bmatrix}$$

$Pr(s_{t+1}|s_t\!=\!6, a_t\!=\!1)$ $Pr(s_{t+1}|s_t\!=\!6, a_t\!=\!2)$ $Pr(s_{t+1}|s_t\!=\!6, a_t\!=\!3)$ $Pr(s_{t+1}|s_t\!=\!6, a_t\!=\!4)$

$$\boldsymbol{\tau} = [\overset{s_1}{1}, \overset{a_1}{3}, \overset{r_2}{0}, \overset{s_2}{2}, \overset{a_2}{3}, \overset{r_3}{0}, \overset{s_3}{6}, \overset{a_3}{2}, \overset{r_4}{0}, \overset{s_4}{10}, \overset{a_4}{2}, \overset{r_5}{0}, \overset{s_5}{11}, \overset{a_5}{1}, \overset{r_6}{0}, \overset{s_6}{7}, \overset{a_6}{2}, \overset{r_7}{3}]$$

$$\boldsymbol{\tau} = [\overset{s_1}{1}, \overset{r_1}{0}, \overset{a_1}{3}, \overset{s_2}{2}, \overset{r_2}{0}, \overset{a_2}{3}, \overset{s_3}{6}, \overset{r_3}{0}, \overset{a_3}{2}, \overset{s_4}{10}, \overset{r_4}{0}, \overset{a_4}{2}, \overset{s_5}{11}, \overset{r_5}{0}, \overset{a_5}{1}, \overset{s_6}{7}, \overset{r_6}{1}, \overset{a_6}{2}]$$

Image credit: Understanding Deep Learning

Discount Factor

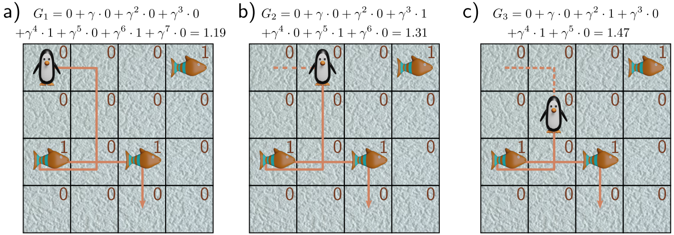
- The **discount factor** $\gamma \in [0, 1)$ determines the importance of future rewards.
- We can use the discount factor to compute the **discounted return**, which gives a measure of the total reward accumulated over time:

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots + \gamma^{T-t-1} R_T$$

- This can be simplified and generalized:

$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

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


a) $G_1 = 0 + \gamma \cdot 0 + \gamma^2 \cdot 0 + \gamma^3 \cdot 0$
$+\gamma^4 \cdot 1 + \gamma^5 \cdot 0 + \gamma^6 \cdot 1 + \gamma^7 \cdot 0 = 1.19$
b) $G_2 = 0 + \gamma \cdot 0 + \gamma^2 \cdot 0 + \gamma^3 \cdot 1$
$+\gamma^4 \cdot 0 + \gamma^5 \cdot 1 + \gamma^6 \cdot 0 = 1.31$
c) $G_3 = 0 + \gamma \cdot 0 + \gamma^2 \cdot 1 + \gamma^3 \cdot 0$
$+\gamma^4 \cdot 1 + \gamma^5 \cdot 0 = 1.47$
0 0 0 1
0 0 0 0
1 0 1 0
0 0 0 0

$$\begin{array}{c} s_1 \; r_2 \; s_2 \; r_3 \; s_3 \; r_4 \; s_4 \; r_5 \; s_5 \; r_6 \; s_6 \; r_7 \; s_7 \; r_8 \; s_8 \; r_9 \\ \boldsymbol{\tau} = [1, 0, 2, 0, 6, 0, 10, 0, 9, 1, 10, 0, 11, 1, 15, 0] \\ s_1, r_1, s_2, r_2, s_3, r_3, s_4, r_4, s_5, r_5, s_6, r_6, s_7, r_7, s_8, r_8 \end{array}$$