Markov Decision Process (with reward)

$$\mathcal{M} = (\mathcal{S}, \mu_0, \mathcal{A}, P, R, \gamma)$$

$\mathcal{S}$: state space, set of all possible states

μ_0 : initial state distribution ; $\mu_0 \in \Delta(\mathcal{S})$

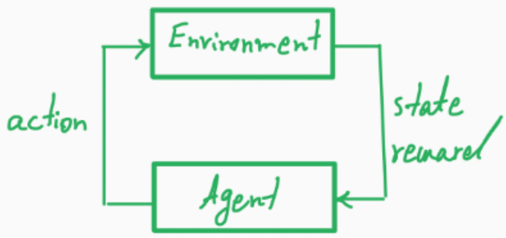
$\mathcal{A}$: action space, set of all possible actions

P : transition function ; $P : \mathcal{S} \times \mathcal{A} \longrightarrow \Delta(\mathcal{S})$

R : reward function ; $R : \mathcal{S} \times \mathcal{A} \longrightarrow \mathbb{R}$

γ : discount factor ; $\gamma \in [0, 1]$

Agent's interaction with the environment:



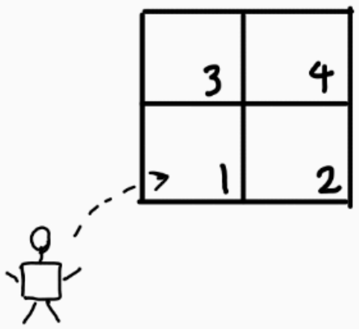
- The interaction starts at time step $t=0$ and state $s_0 \sim \mu_0$.

- At time step $t \in \mathbb{N}_0$ and state $s_t \in \mathcal{S}$, the agent takes action $A_t \in \mathcal{A}$, observes the next state $s_{t+1} \sim P(\cdot \mid s_t, A_t)$, and receives the immediate reward $R_t = R(s_t, A_t) \in \mathbb{R}$.

- The history of interaction at time t is called a trajectory $\tau_t \in \mathcal{T}$, where

$$\tau_t = (\underbrace{s_0, a_0, r_0}_{t=0}, \underbrace{s_1, a_1, r_1}_{t=1}, \dots, \underbrace{s_t, a_t, r_t}_{t=t})$$

Example:

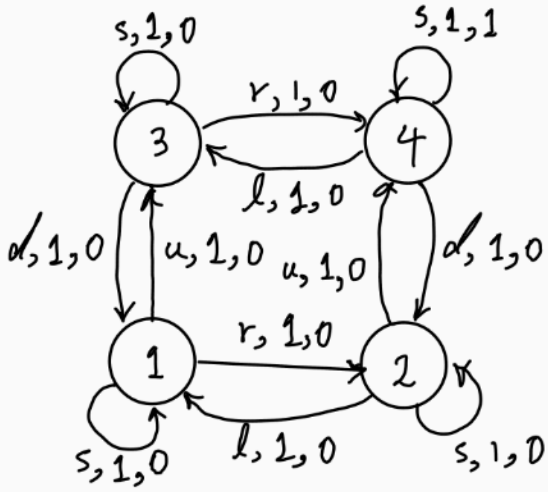


$S = \{1, 2, 3, 4\}$

$A = \{r, l, u, d, s\}$

$\mu_0 = [1, 0, 0, 0]$

MDP



edge label: (action, transition probability, reward)

sample trajectories

$(s_0 = 1, a_0 = u, r_0 = 0, s_1 = 3, a_1 = s, r_1 = 0, s_2 = 3 \dots)$

$(s_0 = 1, a_0 = r, r_0 = 0, s_1 = 2, a_1 = u, r_1 = 0, s_2 = 4, a_2 = s, r_2 = 1, \dots)$

Policy: A possibly randomized mapping from a trajectory to actions;

$$\pi : \underbrace{\mathcal{T}}_{\text{set of all possible trajectories}} \longrightarrow \Delta(A)$$

$(\underbrace{s_0}_{n \text{ possibilities}}, \overbrace{a_0}^{m \text{ possibilities}}, \underbrace{r_0}_{\text{deterministic}}, s_1, a_1, r_1, \dots, s_{H-1}, a_{H-1}, r_{H-1})$ — $n^H m^H$ possibilities

- stationary policy: A policy where the action only depends on the current state;

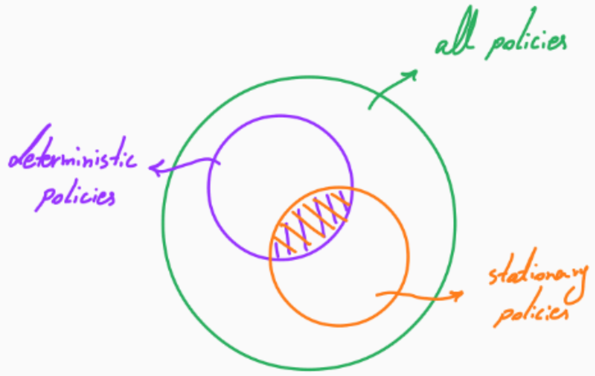
$$\pi : S \longrightarrow \Delta(A)$$

- Deterministic policy: A policy where the action is chosen deterministically;

$$\pi : \mathcal{T} \longrightarrow A$$

- Deterministic, stationary policy:

$$\pi : S \longrightarrow A$$



Goal: The learning agent aims to find a policy π that maximizes the expected (discounted) cumulative reward

$$\pi^* \in \max_{\pi \in \Pi} \; \mathbb{E}_{\substack{S_0 \sim \mu_0 \\ A_t \sim \pi(\tau_t) \\ S_{t+1} \sim P(\cdot \mid S_t, A_t)}} \left[\overbrace{\sum_{t=0}^{T} \gamma^t R(S_t, A_t)}^{\text{return}} \right]$$

for stochastic reward $\longrightarrow R(S_t, A_t) \sim \mathcal{R}(\cdot, \cdot)$

$T \longrightarrow T = \infty$ infinite horizon

$T \longrightarrow T < \infty$ finite horizon

$\gamma \longrightarrow \gamma = 1$ undiscounted

$\gamma \longrightarrow 0 < \gamma < 1$ discounted

$\gamma \longrightarrow \gamma = 0$ greedy / myopic

* We will start with the infinite-horizon discounted setting.

Value functions

- (state) value function of a policy π, $V^\pi : S \longrightarrow \mathbb{R}$, is defined as

$$\boxed{V^\pi(s) = \mathbb{E}_{\substack{A_t \sim \pi(\tau_t) \\ S_{t+1} \sim P(\cdot \mid S_t, A_t)}} \left[\sum_{t=0}^{\infty} \gamma^t R(S_t, A_t) \,\middle|\, S_0 = s \right] \qquad \forall s \in S}$$

- (state-) action value function of a policy π, $Q^\pi : S \times A \longrightarrow \mathbb{R}$ is defined as

$$\boxed{Q^\pi(s,a) = \mathbb{E}_{\substack{A_t \sim \pi(\tau_t) \\ S_{t+1} \sim P(\cdot \mid S_t, A_t)}} \left[\sum_{t=0}^{\infty} \gamma^t R(S_t, A_t) \,\middle|\, S_0 = s, A_0 = a \right] \quad \forall s \in S}$$

$t \geq 1$

Bellman (consistency) Equations:

Let π be a stationary policy. The value functions under π, i.e., V^π and Q^π, satisfy the following equations:

$$V^\pi(s) = \mathbb{E}_{a\sim\pi(s)}\Big[\overbrace{R(s,a)+\gamma\,\mathbb{E}_{s'\sim P(\cdot|s,a)}\big[V^\pi(s')\big]}^{Q^\pi(s,a)}\Big]$$

$$Q^\pi(s,a) = R(s,a)+\gamma\,\mathbb{E}_{s'\sim P(\cdot|s,a)}\big[V^\pi(s')\big]$$

$$Q^\pi(s,a) = \mathbb{E}_{\substack{A_t\sim\pi(\underline{s}) \\ s_{t+1}\sim P(\cdot|s_t,A_t)}}\Big[\sum_{t=0}^{\infty}\gamma^t R(s_t,A_t)\,\Big|\, s_0=s, A_0=a\Big] \quad \forall s\in S$$

$t \geq 1$

$$= \mathbb{E}_{\substack{A_t \\ s_{t+1}}}\Big[\underbrace{R(s_0,A_0)}_{R(s,a)} + \sum_{t=1}^{\infty}\gamma^t R(s_t,A_t)\,\Big|\, s_0=s, A_0=a\Big]$$

$$= R(s,a) + \gamma\,\underbrace{\mathbb{E}_{\substack{A_t \\ s_{t+1}}}\underbrace{\Big[\sum_{t=1}^{\infty}\gamma^{t-1} R(s_t,A_t)\,\Big|\, s_0=s, A_0=a\Big]}_{\Big[\sum_{t'=0}^{\infty}\gamma^{t'} R(s_{t'+1},A_{t'+1})\,\Big|\, s_0=s, A_0=a\Big]}}_{\mathbb{E}_{s'\sim P(\cdot|s,a)}\underbrace{\Big[\sum_{t'=0}^{\infty}\gamma^{t'} R(s_{t'+1},A_{t'+1})\,\Big|\, s_{t'=0}=s'\Big]}_{V^\pi(s')}}$$