1. MDP:

① Markov-chain $= \{S, T\}$

S: state-space

T: state-transform-operator

② $\mu_{t,i} = p(S_t = i)$

$T_{ij} = p(S_{t+1} = i \mid S_t = j)$

$\mu_{t+1} = T\mu_t$

$$\mu_t = \begin{bmatrix} p(S_t = i) \\ p(S_t = j) \\ p(S_t = k) \end{bmatrix} \quad T = \begin{bmatrix} p(S_{t+1} = i \mid S_t = i) & \vdots \\ p(S_{t+1} = i \mid S_t = j) & \vdots \\ p(S_{t+1} = i \mid S_t = k) & \vdots \end{bmatrix}$$

③ Markov-Decision-Process:

$M = \{S, A, T, r\}$

S: ...

A: action-space

T: Matrix $\rightarrow$ Tensor

Match the model in Reinforce-Learning

$T_{i,j,k} = p(S_{t+1} = i \mid S_t = j, a_t = k)$

$$\boxed{\mu_{t,j} = p(S_t = j),\ \xi_{t,k} = p(a_t = k)}$$

$\mu_{t+1,i} = \sum_{j,k} T_{i,j,k} \cdot \mu_{t,j} \cdot \xi_{t,k}$

$$T = \begin{bmatrix} \begin{bmatrix} p(s_{t+1}=i_0 \mid s_t=j_0, a_t=k_0) & \cdots \\ \vdots & \\ p(s_{t+1}=i_n \mid s_t=j_n, a_t=k_0) & \cdots \end{bmatrix} \\ \vdots \\ \begin{bmatrix} p(s_{t+1}=i_n \mid s_t=j_0, a_t=k_0) & \cdots \\ \vdots & \\ p(s_{t+1}=i_n \mid s_t=j_n, a_t=k_0) & \cdots \end{bmatrix} \end{bmatrix}$$

$r \rightarrow r(s,a) \rightarrow S \times A \rightarrow R$

④ Partially-Observed-Markov-Decision-Process

$\mathcal{M} = \{S, A, O, T, \mathcal{E}, r\}$

$\mathcal{E} \longrightarrow$ emission-probability.

$\longrightarrow$ decide $\boxed{p(O_t \mid S_t)}$

⑤ trick: $\mu \cdot \xi \longrightarrow (s, a)$

将 state-action 视为一个联合状态，简化了分析

$P((s_{t+1}, a_{t+1}) | (s_t, a_t)) = P(s_{t+1} | s_t, a_t) \cdot \pi_\theta(a_{t+1}|$

⑥平稳分布：经过一次转移后不发生改变的分布.

$\boxed{\mu = T\mu}$ → μ是T特征值为1的特征向量

2. Expectation of Reinforce-Learning

① Original Reward / Cost Function 是不平滑的，无法直接求gradient，然后backward

② 平稳分布下的Expectation-of-Reward是光滑的

$\boxed{E_{(s,a) \sim p_\theta(s,a)} r(s,a)}$ ✓

* θ收敛是μ达到平稳分布的必要条件.