Reinforcement Learning Fundamentals: MDPs and Policies

Last Updated April 22, 2025

Markov Decision Process (MDP)

An MDP models a sequential decision problem under uncertainty.

Definition: An MDP is a tuple $(\mathcal{S}, \mathcal{A}, p, r, \gamma)$.

- $\mathcal{S}$: Set of states s.
- $\mathcal{A}$: Set of actions a.
- $p(s'|s,a)$: Transition probability kernel.

$$p(s'|s,a) = \mathbb{P}(S_{t+1} = s'|S_t = s, A_t = a)$$

- $r(s,a)$ or $r(s,a,s')$: Reward function.

$$r(s,a) = \mathbb{E}[R_t|S_t = s, A_t = a]$$

$$r(s,a,s') = \mathbb{E}[R_t|S_t = s, A_t = a, S_{t+1} = s']$$

  Note: $r(s,a) = \sum_{s'} p(s'|s,a) r(s,a,s')$.
- $\gamma \in [0,1)$: Discount factor.

Markov Property: The future depends only on the current state and action, not the past history.

$$\mathbb{P}(S_{t+1}, R_t|S_t, A_t, S_{t-1}, \dots) = \mathbb{P}(S_{t+1}, R_t|S_t, A_t)$$

Policies

A policy specifies how an agent selects actions. **Definition**: A policy π is a sequence of decision rules π_t. **Decision Rule** π_t: Determines the distribution of action A_t given the history $H_t = (S_0, A_0, \dots, S_t)$.

$$A_t \sim \pi_t(\cdot|H_t)$$

Types of Policies

- **History-dependent**: $\pi_t(\cdot|H_t)$.
- **Markovian** (Memoryless): Depends only on the current state S_t.

$$\pi_t(\cdot|H_t) = \pi_t(\cdot|S_t)$$

  Often written as $\pi_t(a|s)$.
- **Stationary**: The decision rule is time-independent.

$$\pi(\cdot|s) = \pi_t(\cdot|s) \quad \forall t$$

  Often written as $\pi(a|s)$.
- **Deterministic**: Maps each state (or history) to a single action.

$$\pi(s) = a$$

  Or $\pi_t(H_t) = a$.

Induced Markov Chain: Given an MDP and a fixed stationary policy π, the state sequence (S_t) forms a Markov chain with transition kernel $p^\pi(s'|s)$:

$$p^\pi(s'|s) = \sum_{a \in \mathcal{A}} \pi(a|s) p(s'|s,a)$$

If π is deterministic, $p^\pi(s'|s) = p(s'|s,\pi(s))$.

Value Functions & Optimality

Evaluating how good states and policies are.

Return (Discounted Sum of Rewards): Total discounted reward from time t.

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

This is a random variable depending on the policy and system dynamics.

State Value Function $v^\pi(s)$: Expected return starting from state s and following policy π.

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

$$v^\pi(s) = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} \middle| S_t = s\right]$$

Policy Value / Objective Function $J(\pi)$: Expected value starting from an initial state distribution ρ_0.

$$J(\pi) = \mathbb{E}_{S_0 \sim \rho_0}[v^\pi(S_0)] = \mathbb{E}_{S_0 \sim \rho_0, \pi}[G_0]$$

Optimal Value Function $v^*(s)$: Maximum possible expected return from state s.

$$v^*(s) = \max_\pi v^\pi(s)$$

Optimal Policy π^*: A policy achieving the optimal value function for all states.

$$\pi^* \text{ is optimal} \iff v^{\pi^*}(s) = v^*(s) \quad \forall s \in \mathcal{S}$$

Equivalently:

$$\pi^* \text{ is optimal} \iff v^{\pi^*}(s) \geq v^\pi(s) \quad \forall s \in \mathcal{S}, \forall \pi$$

Policy Optimization Problem: Find π^* maximizing $J(\pi)$.

$$\pi^* \in \arg\max_\pi J(\pi)$$

If $\rho_0(s) > 0$ for all s, solving $\max_\pi J(\pi)$ is equivalent to finding a π^* such that $v^{\pi^*}(s) = v^*(s)$ for all s.

Existence of Optimal Policies

Theorem: For any MDP with a γ-discounted criterion ($\gamma < 1$) and infinite horizon, there exists at least one optimal policy π^* that is:

- Stationary
- Deterministic
- Memoryless (Markovian)

This means we can search for optimal policies of the form $\pi : \mathcal{S} \to \mathcal{A}$.

State Occupancy Measure

Alternative view of policy value based on state visitation frequency.

Expected Reward under π:

$$r^\pi(s) = \mathbb{E}_{a \sim \pi(\cdot|s), s' \sim p(\cdot|s,a)}[r(s,a,s')]$$

$$r^\pi(s) = \sum_a \pi(a|s) \sum_{s'} p(s'|s,a) r(s,a,s')$$

State Visitation Probability: $p(S_t = s|S_0 \sim \rho_0, \pi)$ is the probability of being in state s at time t. For finite states, if ρ_0 is a row vector, this is the s-th element of $\rho_0(p^\pi)^t$.

Discounted State Occupancy Measure $\rho^\pi_{\rho_0}(s)$: Expected total discounted time spent in state s.

$$\rho^\pi_{\rho_0}(s) = \sum_{t=0}^{\infty} \gamma^t p(S_t = s|S_0 \sim \rho_0, \pi)$$

For finite states, $\rho^\pi_{\rho_0} = \rho_0 \sum_{t=0}^\infty (\gamma p^\pi)^t = \rho_0(I - \gamma p^\pi)^{-1}$.

Policy Value via Occupancy:

$$J(\pi) = \sum_{s \in \mathcal{S}} \rho^\pi_{\rho_0}(s) r^\pi(s) = \langle \rho^\pi_{\rho_0}, r^\pi \rangle$$

Total Occupancy: Summing over all states:

$$\sum_{s \in \mathcal{S}} \rho^\pi_{\rho_0}(s) = \sum_{t=0}^{\infty} \gamma^t \sum_{s \in \mathcal{S}} p(S_t = s|\dots) = \sum_{t=0}^{\infty} \gamma^t = \frac{1}{1-\gamma}$$

Normalized Occupancy Distribution:

$$d^\pi_{\rho_0}(s) = (1-\gamma)\rho^\pi_{\rho_0}(s)$$

This is a proper probability distribution ($\sum_s d^\pi_{\rho_0}(s) = 1$).

$$J(\pi) = \frac{1}{1-\gamma} \sum_s d^\pi_{\rho_0}(s) r^\pi(s) = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d^\pi_{\rho_0}}[r^\pi(s)]$$

Interpretation of Discount Factor γ

γ can be seen as the probability of continuing the process at each step.

- Consider an MDP where each transition has a probability $1-\gamma$ of ending in a terminal absorbing state (with 0 reward) and γ of continuing according to p.
- The probability of a trajectory lasting exactly h steps is $(1-\gamma)\gamma^{h-1}$ (for $h \geq 1$).
- The expected length of a trajectory (effective horizon) is $\frac{1}{1-\gamma}$.
- The value $v^\pi_\gamma(s)$ in the original MDP (discounted) is related to the value $v^{\pi'}(s)$ in the modified MDP (total reward) by $v^{\pi'}(s) \approx \gamma v^\pi_\gamma(s)$.