Markov Decision Process

$$\langle \mathcal{S}, \mathcal{A}, \mathcal{P}^a_{ss'}, \mathcal{R}^a_s, \gamma \rangle$$

Discount Factor: $0 \leq \gamma \leq 1$

State Transition Matrix

$$\mathcal{P}^a = [\mathcal{P}^a_{ss'}] \qquad \mathcal{P}^a_{ss'} = \mathbb{P}[S_{t+1} = s' \mid S_t = s, A_t = a] \qquad s, s' \in \mathcal{S}$$

Policy

$$\pi(a \mid s) = \mathbb{P}[A_t = a \mid S_t = s]$$

Reward Function

$$\mathcal{R}^a_s = \mathbb{E}[R_{t+1} \mid S_t = s, A_t = a] = \sum_{s' \in \mathcal{S}} \mathcal{P}^a_{ss'} \mathcal{R}^a_{ss'}$$

$$\mathcal{R}^a = \begin{bmatrix} \mathcal{R}^a_{s_1} & \mathcal{R}^a_{s_2} & \cdots & \mathcal{R}^a_{s_{|\mathcal{S}|}} \end{bmatrix}^\top \qquad \mathcal{R}^a_{s_i} \in \mathbb{R}$$

Return

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

State-Value Function

$$v_\pi(s) = \mathbb{E}_\pi[G_t \mid S_t = s] = \sum_{a \in \mathcal{A}} \pi(a \mid s) \sum_{s' \in \mathcal{S}} \mathcal{P}^a_{ss'} \left[\mathcal{R}^a_{ss'} + \gamma v_\pi(s')\right]$$

$$v_\pi = \begin{bmatrix} v_\pi(s_1) & v_\pi(s_2) & \cdots & v_\pi(s_{|\mathcal{S}|}) \end{bmatrix}^\top \qquad v_\pi(s_i) \in \mathbb{R}$$

Action-Value Function

$$q_\pi(s, a) = \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a]$$

$$q_\pi(s, a) = \sum_{s'} \mathcal{P}^a_{ss'} \left[\mathcal{R}^a_{ss'} + \gamma v_\pi(s')\right] \qquad v_\pi(s) = \sum_{a} \pi(a \mid s) q_\pi(s, a)$$

Reduction: MDP $\longrightarrow$ Markov Reward Process (MRP)

Applying a *fixed policy* π defines an induced *Markov Reward Process* (MRP): $\langle \mathcal{S}, \mathcal{P}^\pi_{ss'}, \mathcal{R}^\pi_s, \gamma \rangle$

$$\mathcal{P}^\pi_{ss'} = \sum_{a \in \mathcal{A}} \pi(a \mid s) \mathcal{P}^a_{ss'} \qquad \mathcal{R}^\pi_s = \sum_{a \in \mathcal{A}} \pi(a \mid s) \sum_{s' \in \mathcal{S}} \mathcal{P}^a_{ss'} \mathcal{R}^a_{ss'}$$

$$\mathcal{P}^\pi = [\mathcal{P}^\pi_{ss'}] \qquad \mathcal{R}^\pi = \begin{bmatrix} \mathcal{R}^\pi_{s_1} & \mathcal{R}^\pi_{s_2} & \cdots & \mathcal{R}^\pi_{s_{|\mathcal{S}|}} \end{bmatrix}^\top \qquad \mathcal{R}^\pi_{s_i} \in \mathbb{R}$$

Bellman Equation

$$v_\pi = \mathcal{R}^\pi + \gamma \mathcal{P}^\pi v_\pi \quad \Longrightarrow \quad v_\pi = (I - \gamma \mathcal{P}^\pi)^{-1} \mathcal{R}^\pi$$

Dynamic Programming

Norms: For a vector $v \in \mathbb{R}^{|\mathcal{S}|}$, the L_p norm is

$$\|v\|_p = \left(\sum_i |v_i|^p\right)^{1/p}, \quad \|v\|_\infty = \max_i |v_i|$$

Contraction: An operator T is a γ-contraction if

$$\|T(u) - T(v)\|_\infty \leq \gamma \|u - v\|_\infty, \quad 0 \leq \gamma < 1$$

Fixed Point: v is a fixed point of T if

$$T(v) = v$$

Bellman Expectation Operator

$$T^\pi(v) = \mathcal{R}^\pi + \gamma \mathcal{P}^\pi v \qquad \|T^\pi(u) - T^\pi(v)\|_\infty \leq \gamma \|u - v\|_\infty$$

Optimal Value Functions

$$v_*(s) = \max_\pi v_\pi(s) = \max_{a \in \mathcal{A}} q_*(s, a) \qquad v_* = \max_{a \in \mathcal{A}} \left[\mathcal{R}^a + \gamma \mathcal{P}^a v_*\right] \quad \text{(element-wise)}$$

$$q_*(s, a) = \max_\pi q_\pi(s, a) = \sum_{s'} \mathcal{P}^a_{ss'} \left[\mathcal{R}^a_{ss'} + \gamma \max_{a'} q_*(s', a')\right] = \sum_{s'} \mathcal{P}^a_{ss'} \left[\mathcal{R}^a_{ss'} + \gamma v_*(s')\right]$$

Optimal Policy

$$\pi_*(a \mid s) = \begin{cases} 1, & a \in \arg\max a' \in \mathcal{A} q_*(s, a') \\ 0, & \text{otherwise} \end{cases}$$

Bellman Optimality Operator

$$T^*(v) = \max_{a \in \mathcal{A}} \left[\mathcal{R}^a + \gamma \mathcal{P}^a v\right] \quad \text{(element-wise)} \qquad \|T^*(u) - T^*(v)\|_\infty \leq \gamma \|u - v\|_\infty$$

Contraction Mapping Theorem

A γ-contraction T on a complete metric space has a *unique fixed point* v_e, and the iterates $v_{n+1} = T(v_n)$ converge to v_e at a linear rate γ: $\|v_{n+1} - v_e\| \leq \gamma \|v_n - v_e\|$.

$$T^\pi(v_\pi) = v_\pi \quad (\textbf{iterative policy evaluation})$$

$$T^*(v_*) = v_* \quad (\textbf{value iteration})$$

Policy Iteration

$$v_{\pi_k} = T^{\pi_k}(v_{\pi_k}) \quad \text{(evaluate current policy)}$$

$$\pi_{k+1}(s) = \arg\max_{a \in \mathcal{A}} \sum_{s'} \mathcal{P}^a_{ss'} \left[\mathcal{R}^a_{ss'} + \gamma v_{\pi_k}(s')\right] \quad \text{(improve policy)}$$

$$\lim_{k \to \infty} \pi_k = \pi_* \qquad \lim_{k \to \infty} v_{\pi_k} = v_*$$

Model-Free Prediction

Monte Carlo Estimation

Suppose a random variable X with unknown expected value $\mu = \mathbb{E}[X]$, the Monte Carlo estimate using N independent samples $X_1, \ldots, X_N$ is:

$$\hat{\mu}_N = \frac{1}{N}\sum_{i=1}^{N} X_i$$

$$\mathbb{E}[\hat{\mu}_N] = \mu \qquad \hat{\mu}_N \to \mu \text{ as } N \to \infty$$

Monte Carlo Prediction in RL

Estimate $v_\pi(s)$ using returns G_t over many episodes (samples):

$$G_t = \sum_{k=0}^{T-t-1} \gamma^k R_{t+k+1} \quad \text{(where } T \text{ is the terminal time)}$$

$$v_\pi(s) \approx \frac{1}{N(s)}\sum_{i=1}^{N(s)} G_t^{(i)} \qquad (N(s)\text{: visit count for state } s)$$

Incremental form:

$$V(S_t) \leftarrow V(S_t) + \frac{1}{N(S_t)}\big(G_t - V(S_t)\big)$$

$$V(S_t) \leftarrow V(S_t) + \alpha\big(\text{Target} - V(S_t)\big)$$

Temporal Difference Learning

Use bootstrapped target for one next step instead of full episode:

$$V(S_t) \leftarrow V(S_t) + \alpha(\underbrace{R_{t+1} + \gamma V(S_{t+1})}_{\text{TD Target}} - V(S_t))$$

The term $\delta_t = R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$ is the **TD error**.

n-Step TD and TD(λ)

n-step target:

$$G_t^{(n)} = R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{n-1} R_{t+n} + \gamma^n V(S_{t+n}) \qquad (G_t^{(\infty)} = G_t)$$

TD(λ) combines all n-step targets:

$$G_t^\lambda = (1-\lambda)\sum_{n=1}^{\infty} \lambda^{n-1} G_t^{(n)} \qquad V(S_t) \leftarrow V(S_t) + \alpha\big(G_t^\lambda - V(S_t)\big)$$

$\lambda = 0$: TD(0) 1-step update $\qquad$ $\lambda = 1$: TD(1) Monte Carlo update