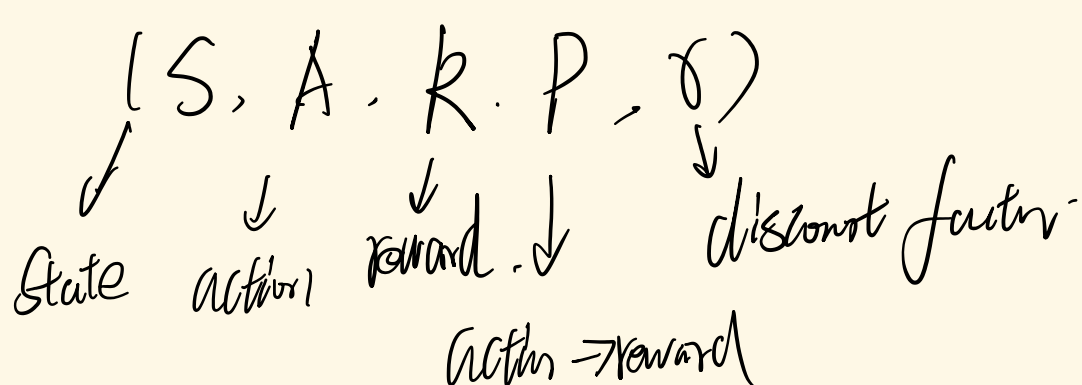


Markov Decision Process



Policy π : $V^\pi = \sum_t \gamma^t r$ $V^* := \max_\pi \sum_t \gamma^t r$

$$V^* = \arg \max_{\pi} E \left[\sum_{t=0}^{\infty} \gamma^t r_t | \pi \right]$$

Bellman Equation.

$$Q^*(s, a) = E_{s' \sim p} [r + \gamma \max_{a'} Q^*(s', a') | s, a]$$

Value iteration:

$$Q_{i+1}(s, a) = E [r + \gamma \max_{a'} Q_i(s', a') | s, a]$$

$$Q(s, a; \theta) \approx Q^*(s, a)$$

\downarrow

approximator.

forward

$$L(\theta_i) = E_{s, a \sim p(\cdot)} [y_i - Q(s, a; \theta_i)]^2$$

where: $y_i = E_{s' \sim p} [r + \gamma \max_{a'} Q(s', a'; \theta_{i-1}) | s, a]$

backward:

$$\nabla_{\theta} L(\theta_i) = E_{s, a \sim p(\cdot)} [s' \sim p [r + \gamma \max_{a'} Q(s', a'; \theta_{i-1}) - Q(s, a; \theta_i)]]$$

parametrized policies $\Pi = \{\pi_{\theta}, \theta \in \mathbb{R}^m\}$

for each policy: $J(\theta) = E_{\tau \sim p(\tau; \theta)} [r(\tau)]$

$$= \int_{\tau} r(\tau) p(\tau; \theta) d\tau$$

$r(\tau)$ is the reward of a trajectory $\tau = (s_0, a_0, r_0, s_1, \dots)$

trick $\nabla_{\theta} p(\tau; \theta) = p(\tau; \theta) \frac{\nabla_{\theta} p(\tau; \theta)}{p(\tau; \theta)} = p(\tau; \theta) \cdot \nabla_{\theta} \log p(\tau; \theta)$

$$\nabla_{\theta} J(\theta) = \int_{\tau} (r(\tau) \nabla_{\theta} \log p(\tau; \theta)) p(\tau; \theta) d\tau$$

$$= E_{\tau \sim p(\tau; \theta)} [r(\tau) \nabla_{\theta} \log p(\tau; \theta)] \quad \text{Monte Carlo.}$$

$$p(\tau; \theta) = \prod_{t=0}^{\infty} p(s_{t+1} | s_t, a_t) \pi_{\theta}(a_t | s_t)$$

Thus: $\log p(\tau; \theta) = \sum_{t=0}^{\infty} \log p(s_{t+1} | s_t, a_t) + \log \pi_{\theta}(a_t | s_t)$

$$\nabla_{\theta} \log p(\tau; \theta) = \sum_{t=0}^{\infty} \nabla_{\theta} \log \pi_{\theta}(a_t | s_t)$$

$$\nabla_{\theta} J(\theta) \approx \sum_{t=0}^{\infty} r(\tau) \nabla_{\theta} \log \pi_{\theta}(a_t | s_t)$$

baseline:

$$\nabla_{\theta} J(\theta) \approx \sum_{t=0}^{\infty} \left(\sum_{\tau \sim p} \gamma^{t-t'} r_{t'} - b(s_t) \right) \nabla_{\theta} \log \pi_{\theta}(a_t | s_t)$$

Q function as baseline.

$$\nabla_{\theta} J(\theta) \approx \sum_{t=0}^{\infty} (Q^{\pi_{\theta}}(s_t, a_t) - V^{\pi_{\theta}}(s_t)) \nabla_{\theta} \log \pi_{\theta}(a_t | s_t)$$

Universal approximation theorem.