

Reinforcement Learning

MDP

马尔科夫过程 $\langle S, P \rangle$

马尔科夫奖励过程 $\langle S, P, r, \gamma \rangle$

回报 $G_t = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \dots = \sum_{k=0}^{\infty} \gamma^k R_{t+k}$

价值函数 $V(s) = \mathbb{E}[G_t | S_t = s]$
 $= \mathbb{E}[R_t + \gamma V(S_{t+1}) | S_t = s]$

贝尔曼方程：
$$= r(s) + \gamma \sum_{s' \in S} P(s'|s) V(s')$$

策略 $\pi(a|s) = P(A_t = a | S_t = s)$

状态价值函数 $V^\pi(s) = \mathbb{E}_\pi[G_t | S_t = s]$

动作价值函数 $Q^\pi(s, a) = \mathbb{E}_\pi[G_t | S_t = s, A_t = a]$

贝尔曼期望方程
$$Q^\pi(s, a) = \mathbb{E}_\pi[R_t + \gamma Q^\pi(s', a') | S_t = s, A_t = a]$$
$$= r(s, a) + \gamma \sum_{s' \in S} P(s'|s, a) \sum_{a' \in A} \pi(a'|s') Q^\pi(s', a')$$

贝尔曼最优方程
$$V^*(s) = \max_{\pi} V^\pi(s) \quad \forall s \in S$$
$$= \max_{a \in A} Q^*(s, a)$$
$$= \max_{a \in A} \{r(s, a) + \gamma \sum_{s' \in S} P(s'|s, a) V^*(s')\}$$
$$Q^*(s, a) = \max_{\pi} Q^\pi(s, a) \quad \forall s \in S, a \in A$$
$$= r(s, a) + \gamma \sum_{s' \in S} P(s'|s, a) \max_{a' \in A} Q^*(s', a')$$

$$V^\pi = \sum_{a \in A} \pi(a|s) Q^\pi(s, a)$$
$$Q^\pi = r(s, a) + \gamma \sum_{s' \in S} P(s'|s, a) V^\pi(s')$$

value-based

目标函数 $J(\theta) = E_S[V_\pi(S)]$

策略梯度定理
$$\frac{\partial J(\theta)}{\partial \theta} = \nabla_\theta J(\theta) = \mathbb{E}_S[\mathbb{E}_{A \sim \pi(\cdot|S; \theta)}[\frac{\partial \ln \pi(A|S; \theta)}{\partial \theta} \cdot Q_\pi(S, A)]]$$

随机梯度——策略梯度的无偏估计
$$g(s, a; \theta) \triangleq Q_\pi(s, a) \cdot \nabla_\theta \ln \pi(a|s; \theta)$$

策略网络提升
$$\theta \leftarrow \theta + \beta \cdot g(s, a; \theta)$$

带基线的策略梯度方法

带基线的策略梯度定理——b是不依赖于A的任意函数
$$\nabla_\theta J(\theta) = \mathbb{E}_S[\mathbb{E}_{A \sim \pi(\cdot|S; \theta)}[(Q_\pi(S, A) - b) \cdot \nabla_\theta \ln \pi(A|S; \theta)]]$$

随机梯度
$$g_b(s, a; \theta) = [Q_\pi(S, A) - b] \cdot \nabla_\theta \ln \pi(A|S; \theta)$$

$g_b(s, a; \theta)$ 是 $\nabla_\theta J(\theta)$ 的无偏估计

$Bias = \mathbb{E}_{S, A}[g_b(s, a; \theta)] - \nabla_\theta J(\theta) = 0$

b 的取值对 $g_b(s, a; \theta)$ 是有影响的

$Var = \mathbb{E}_{S, A}[||g_b(s, a; \theta) - \nabla_\theta J(\theta)||^2]$

b 接近 $Q_\pi(s, a)$ 关于 a 的均值, 方差会比较小

$\Rightarrow b = V_\pi(s)$ 是很好的基线

REINFORCE

折扣回报
$$U_t = \sum_{k=t}^n \gamma^{k-t} \cdot R_k$$

动作价值是折扣回报的期望
$$Q_\pi(s_t, a_t) = \mathbb{E}[U_t | S_t = s_t, A_t = a_t]$$

用折扣回报的观测值蒙特卡洛近似动作价值
$$\tilde{g}(s_t, a_t; \theta) = u_t \cdot \nabla_\theta \ln \pi(a_t | s_t; \theta)$$

策略网络提升
$$\theta_{new} \leftarrow \theta_{now} + \beta \cdot \sum_{t=1}^n \gamma^{t-1} \cdot \tilde{g}(s_t, a_t; \theta_{now})$$

带基线的REINFORCE

折扣回报
$$u_t = \sum_{k=t}^n \gamma^{k-t} \cdot r_k$$

策略网络

基线——价值网络做出的预测
$$\hat{v}_t = v(s_t; \omega)$$

带基线的策略梯度
$$\tilde{g}(s_t, a_t; \theta) = (u_t - \hat{v}_t) \cdot \nabla_\theta \ln \pi(a_t | s_t; \theta)$$

梯度上升
$$\theta \leftarrow \theta + \beta \cdot \tilde{g}(s_t, a_t; \theta)$$

损失函数
$$L(\omega) = \frac{1}{2n} \sum_{t=1}^n [v(s_t; \omega) - u_t]^2$$

价值网络

损失函数的梯度
$$\nabla_\omega L(\omega) = \frac{1}{n} \sum_{t=1}^n [v(s_t; \omega) - u_t] \cdot \nabla_\omega v(s_t; \omega)$$

梯度下降
$$\omega \leftarrow \omega - \alpha \cdot \nabla_\omega L(\omega)$$

价值网络只起到基线的作用

policy-based

策略网络

用价值网络近似动作价值
$$\hat{g}(s, a; \theta) \triangleq q(s, a; \omega) \cdot \nabla_\theta \ln \pi(a|s; \theta)$$

策略网络提升
$$\theta \leftarrow \theta + \beta \cdot \hat{g}(s, a; \theta)$$

Actor-Critic

TD目标
$$\hat{y}_t \triangleq r_t + \gamma \cdot q(s_{t+1}, a_{t+1}; \omega)$$

损失函数
$$L(\omega) \triangleq \frac{1}{2} [q(s_t, a_t; \omega) - \hat{y}_t]^2$$

价值网络

损失函数梯度
$$\nabla_\omega L(\omega) = [q(s_t, a_t; \omega) - \hat{y}_t] \cdot \nabla_\omega q(s_t, a_t; \omega)$$

梯度下降
$$\omega \leftarrow \omega - \alpha \cdot \nabla_\omega L(\omega)$$

Advantage Actor-Critic (A2C)

贝尔曼公式
$$Q_\pi(s_t, a_t) = \mathbb{E}_{S_{t+1} \sim p(\cdot|s_t, a_t)}[R_t + \gamma \cdot V_\pi(S_{t+1})]$$

优势函数(Advantage function)
$$Q_\pi(s, a) - V_\pi(s)$$

策略网络

近似策略梯度
$$g(s, a; \theta) = [Q_\pi(s, a) - V_\pi(s)] \cdot \nabla_\theta \ln \pi(a|s; \theta)$$
$$= [\mathbb{E}_{S_{t+1}}[R_t + \gamma \cdot V_\pi(S_{t+1})] - V_\pi(s_t)] \cdot \nabla_\theta \ln \pi(a_t | s_t; \theta)$$

蒙特卡洛近似
$$\approx [r_t + \gamma \cdot V_\pi(s_{t+1}) - V_\pi(s_t)] \cdot \nabla_\theta \ln \pi(a_t | s_t; \theta)$$

用价值网络 $v(s; \omega)$ 替换状态价值函数 $V_\pi(s)$
$$\hat{g}(s, a; \theta) \triangleq [r_t + \gamma \cdot v_\pi(s_{t+1}; \omega) - v_\pi(s_t; \omega)] \cdot \nabla_\theta \ln \pi(a_t | s_t; \theta)$$

策略网络提升
$$\theta \leftarrow \theta + \beta \cdot \hat{g}(s, a; \theta)$$

策略目标

TRPO

策略网络

贝尔曼公式
$$V_\pi(s_t) = \mathbb{E}_{A_t \sim \pi(\cdot|s_t; \theta)}[\mathbb{E}_{S_{t+1} \sim p(\cdot|s_t, A_t)}[R_t + \gamma \cdot V_\pi(S_{t+1})]]$$

TD目标
$$\hat{y}_t \triangleq r_t + \gamma \cdot v(s_{t+1}; \omega)$$

损失函数
$$L(\omega) \triangleq \frac{1}{2} [v(s_t; \omega) - \hat{y}_t]^2$$

价值网络

损失函数的梯度
$$\nabla_\omega L(\omega) = [v(s_t; \omega) - \hat{y}_t] \cdot \nabla_\omega v(s_t; \omega)$$

梯度下降
$$\omega \leftarrow \omega - \alpha \cdot \nabla_\omega L(\omega)$$

Multi-agent

完全合作关系 multi-agent cooperative A2C(MAC-A2C)

策略网络

动作A的概率密度函数
$$\pi(A|S; \theta^1, \dots, \theta^m) \triangleq \pi(A^1|S; \theta^1) \times \dots \times \pi(A^m|S; \theta^m)$$

合作关系 MARL 的策略梯度定理
$$\nabla_\theta J(\theta^1, \dots, \theta^m) = \mathbb{E}_{S, A}[(Q_\pi(S, A) - b) \cdot \nabla_\theta \ln \pi(A^i|S; \theta^i)]$$

策略网络

策略网络提升
$$\theta^i \leftarrow \theta^i + \beta \cdot \tilde{g}^i(s_t, a_t^i; \theta^i)$$

策略网络

策略网络提升
$$\theta^i \leftarrow \theta^i + \beta \cdot \tilde{g}^i(s_t, a_t^i; \theta^i)$$

观测
$$s = [o_1, \dots, o_m]$$

TD目标
$$\hat{y}_t \triangleq r_t + \gamma \cdot v(s_{t+1}; \omega)$$

损失函数
$$L(\omega) \triangleq \frac{1}{2} [v(s_t; \omega) - \hat{y}_t]^2$$

价值网络

损失函数的梯度
$$\nabla_\omega L(\omega) = [v(s_t; \omega) - \hat{y}_t] \cdot \nabla_\omega v(s_t; \omega)$$

梯度下降
$$\omega \leftarrow \omega - \alpha \cdot \nabla_\omega L(\omega)$$

智能体不能独立做决策, 需要所有智能体的观测

策略网络

策略网络提升
$$\tilde{g}^i(s, a; \theta^i) \triangleq (Q_\pi(s, a) - V_\pi(s)) \cdot \nabla_\theta \ln \pi(a^i|s; \theta^i)$$

用 $r_t + \gamma \cdot v(s_{t+1}; \omega)$ 近似 $Q_\pi(s_t, a_t)$,
用 $v(s_t; \omega)$ 近似 $V_\pi(s_t)$

策略网络提升
$$\tilde{g}^i(s_t, a_t^i; \theta^i) \triangleq (r_t + \gamma \cdot v(s_{t+1}; \omega) - v(s_t; \omega)) \cdot \nabla_\theta \ln \pi(a_t^i|s_t; \theta^i)$$

