

# Reinforcement Learning

value-based

一般方法

- 目标函数  $J(\theta) = E_S[V_\pi(S)]$
- 策略梯度定理  $\frac{\partial J(\theta)}{\partial \theta} = \nabla_\theta J(\theta) = E_S[\mathbb{E}_{A \sim \pi(\cdot|S;\theta)}[\frac{\partial \ln \pi(A|S;\theta)}{\partial \theta} \cdot Q_\pi(S, A)]]$
- 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  $\nabla_\theta J(\theta) = E_S[\mathbb{E}_{A \sim \pi(\cdot|S;\theta)}[(Q_\pi(S, A) - b) \cdot \nabla_\theta \ln \pi(A|S; \theta)]]$
- 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  
 $Bias = E_{S,A}[g_b(s, a; \theta)] - \nabla_\theta J(\theta) = 0$   
 $b$  的取值对  $g_b(s, a; \theta)$  是有影响的  
 $Var = E_{S,A}[||g_b(s, a; \theta) - \nabla_\theta J(\theta)||^2]$   
 $b$  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  $Q_\pi(s_t, a_t) = E[U_t | S_t = s_t, A_t = a_t]$
- 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

- 策略网络
  - 折扣回报  $u_t = \sum_{k=t}^n \gamma^{k-t} \cdot r_k$
  - 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  $\theta \leftarrow \theta + \beta \cdot \tilde{g}(s_t, a_t; \theta)$

价值网络

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

Actor-Critic

- 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  $\theta \leftarrow \theta + \beta \cdot \hat{g}(s, a; \theta)$
- 价值网络
  - TD目标  $\hat{y}_t \triangleq r_t + \gamma \cdot q(s_{t+1}, a_{t+1}; \omega)$
  - 损失函数  $L(\omega) \triangleq \frac{1}{2} [q(s_t, a_t; \omega) - \hat{y}_t]^2$
  - 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  $\omega \leftarrow \omega - \alpha \cdot \nabla_\omega L(\omega)$

价值网络直接参与策略提升

Advantage Actor-Critic (A2C)

- 策略网络
  - 贝尔曼公式  $Q_\pi(s_t, a_t) = E_{S_{t+1} \sim p(\cdot|s_t, a_t)}[R_t + \gamma \cdot V_\pi(S_{t+1})]$
  - 优势函数(Advantage function)  $Q_\pi(s, a) - V_\pi(s)$
  - 近似策略梯度  $g(s_t, a_t; \theta) = [Q_\pi(s_t, a_t) - V_\pi(s_t)] \cdot \nabla_\theta \ln \pi(a_t | s_t; \theta)$   
 $= [E_{S_{t+1}}[R_t + \gamma \cdot V_\pi(S_{t+1})] - V_\pi(s_t)] \cdot \nabla_\theta \ln \pi(a_t | s_t; \theta)$   
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  $V_\pi(s)$   
 $\tilde{g}(s_t, a_t; \theta) \triangleq [r_t + \gamma \cdot v(s_{t+1}; \omega)] - v(s_t; \omega) \cdot \nabla_\theta \ln \pi(a_t | s_t; \theta)$
  - 策略网络提升  $\theta \leftarrow \theta + \beta \cdot \tilde{g}(s, a; \theta)$

Advantage Actor-Critic (A2C)

- 价值网络
  - 贝尔曼公式  $V_\pi(s_t) = E_{A_t \sim \pi(\cdot|s_t; \theta)}[E_{S_{t+1} \sim p(\cdot|s_t, A_t)}[R_t + \gamma \cdot V_\pi(S_{t+1})]]$
  - TD目标  $\hat{y}_t \triangleq r_t + \gamma \cdot v(s_{t+1}; \omega)$
  - 损失函数  $L(\omega) \triangleq \frac{1}{2} [v(s_t; \omega) - \hat{y}_t]^2$
  - 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  $\pi(A^i | S; \theta^i)$

- 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  $\nabla_{\theta^i} J(\theta^1, \dots, \theta^m) = E_{S,A}[(Q_\pi(S, A) - b) \cdot \nabla_{\theta^i} \ln \pi(A^i|S; \theta^i)]$
- 随机梯度——策略梯度的无偏估计  $g^i(s, a; \theta^i) \triangleq (Q_\pi(s, a) - V_\pi(s)) \cdot \nabla_{\theta^i} \ln \pi(a^i|s; \theta^i)$
- 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  $V_\pi(s_t)$   
 $\tilde{g}^i(s_t, a_t^i; \theta^i) \triangleq (r_t + \gamma \cdot v(s_{t+1}; \omega) - v(s_t; \omega)) \cdot \nabla_{\theta^i} \ln \pi(a_t^i | s_t; \theta^i)$
- 策略网络提升  $\theta^i \leftarrow \theta^i + \beta \cdot \tilde{g}^i(s_t, a_t^i; \theta^i)$

智能体不能独立做决策, 需要所有智能体的观测

价值网络  $v(s_t, \omega)$

- 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  $L(\omega) \triangleq \frac{1}{2} [v(s_t; \omega) - \hat{y}_t]^2$
- 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  $\omega \leftarrow \omega - \alpha \cdot \nabla_\omega L(\omega)$

