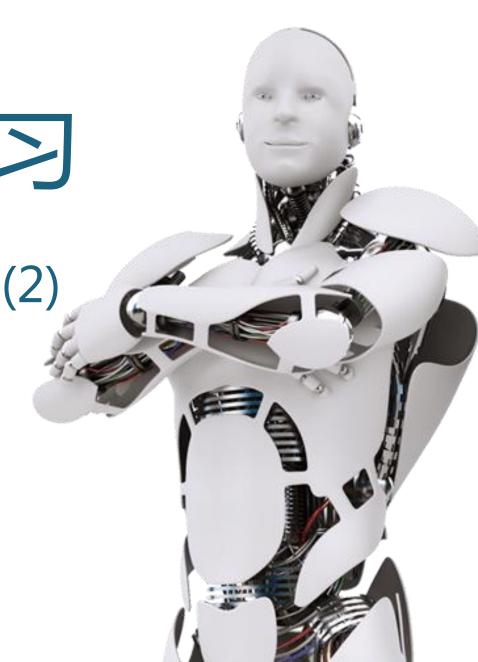
彪哥带你学强化学习

15.深入理解TRPO算法(2)

DEEPLY UNDERSTAND REINFORCEMENT LEARNING

讲师: 韩路彪



# TRPO算法 —— 整体结构

### 理论

- 每训练一步,整体回报都有提升
- 目标函数η,近似函数L
- 如果能保证 | L<sub>π</sub> η | <= X , 就能在 ▽ L<sub>π</sub> >= X 前提下做到 ▽η >=0
   理论实现
- 给出 | L<sub>π</sub> η | <= X 里边的上限X

#### 工程实现

• 实现每训练一步η都有提升

#### TRPO算法 —— 理论实现

$$egin{aligned} |L_{\pi}( ilde{\pi}) - \eta( ilde{\pi})| &\leq rac{4lpha^2\gamma\epsilon}{(1-\gamma)^2} \ lpha &= D_{TV}^{max}ig(\pi_{old},\pi_{new}ig) \ D_{TV}(p||q) = rac{1}{2}\sum_i |p_i - q_i| \ D_{TV}^{max}(\pi|| ilde{\pi}) &= \max_s D_{TV}(\pi(.\ket{s}), ilde{\pi}(.\ket{s})) \ \epsilon &= \max_{s,a} |A_{\pi}(s,a)| \end{aligned}$$

$$\eta( ilde{\pi}) \geq L_{\pi}( ilde{\pi}) - rac{4lpha^2\gamma\epsilon}{(1-\gamma)^2}$$

## TRPO算法 —— 理论实现

证明 
$$|L_{\pi}( ilde{\pi}) - \eta( ilde{\pi})| \leq rac{4lpha^2\gamma\epsilon}{(1-\gamma)^2}$$

$$|L_{\pi}( ilde{\pi}) - \eta( ilde{\pi})|$$

$$=\sum_{t=0}^{\infty} \gamma^t |\mathbb{E}_{ au \sim ilde{\pi}}[ar{A}(s_t)] - \mathbb{E}_{ au \sim \pi}[ar{A}(s_t)]|^2$$

$$=\sum_{t=0}^{\infty} \gamma^t |P(n_t=0) \mathbb{E}_{s_t \sim ilde{\pi}|n_t=0}[ar{A}(s_t)] + P(n_t>0) \mathbb{E}_{s_t \sim ilde{\pi}|n_t>0}[ar{A}(s_t)] |$$

$$-\left. (P(n_t=0)\mathbb{E}_{s_t\sim \pi|n_t=0}[ar{A}(s_t)] + P(n_t>0)\mathbb{E}_{s_t\sim \pi|n_t>0}[ar{A}(s_t)] 
ight) |$$

$$=\sum_{t=0}^{\infty} \gamma^t |P(n_t>0) \mathbb{E}_{s_t \sim ilde{\pi}|n_t>0}[ar{A}(s_t)] - P(n_t>0) \mathbb{E}_{s_t \sim \pi|n_t>0}[ar{A}(s_t)]|^2$$

$$0 \leq \sum_{t=0}^{\infty} \gamma^t P(n_t > 0) (|\mathbb{E}_{s_t \sim ilde{\pi}|n_t > 0}[ar{A}(s_t)]| + |\mathbb{E}_{s_t \sim \pi|n_t > 0}[ar{A}(s_t)]|)^{-1}$$

nt: 前n步,有几步不一样



$$|L_{\pi}( ilde{\pi}) - \eta( ilde{\pi})|$$

对任意 
$$(a, \tilde{a})|s \sim (\pi, \tilde{\pi})$$
 有  $P(a \neq \tilde{a}|s) \leq \alpha$ 

$$\leq \sum_{t=0}^{\infty} \gamma^t P(n_t>0)(|\mathbb{E}_{s_t\sim ilde{\pi}|n_t>0}[ar{A}(s_t)]|+|\mathbb{E}_{s_t\sim\pi|n_t>0}[ar{A}(s_t)]|)$$

$$\leq \sum_{t=0}^{\infty} \gamma^t 2(1-(1-lpha)^t) 2lpha \max_{s,a} |A(s,a)|^t$$

$$=4\epsilonlpha\sum_{t=0}^{\infty}\gamma^{t}(1-(1-lpha)^{t})$$

$$=4\epsilon lpha (rac{1}{1-\gamma} - rac{1}{1-\gamma(1-lpha)})$$

$$=rac{4\epsilonlpha^2\gamma}{(1-\gamma)(1-\gamma(1-lpha))}$$

$$\leq \! \frac{4\epsilon\alpha^2\gamma}{(1-\gamma)^2}$$

$$\epsilon = \max_{s,a} |A(s,a)|$$

#### 等比级数

