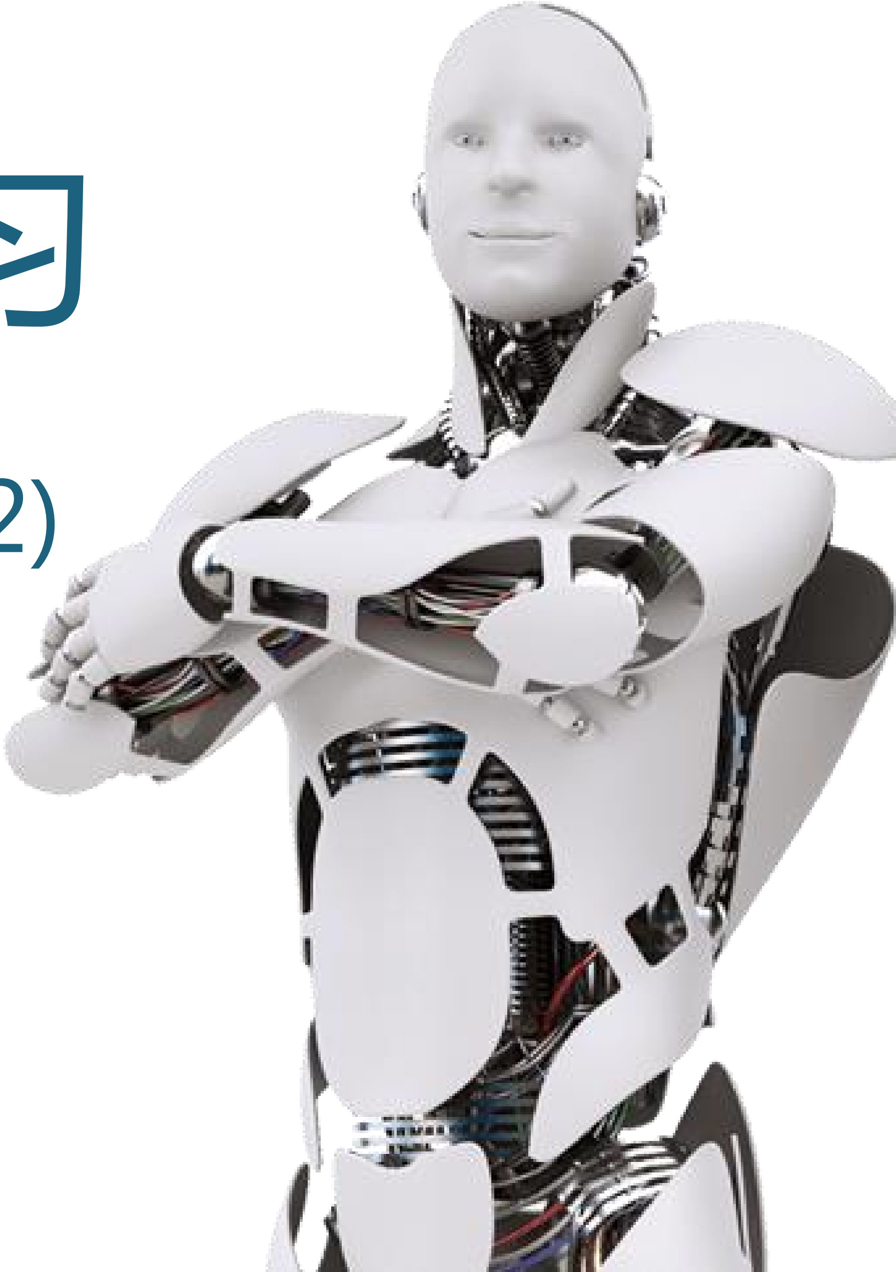
彪哥带你学强化学习

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 $|L_\pi - \eta| \le X$， 就能在$\nabla L_\pi \ge X$ 前提下做到 $\nabla\eta \ge 0$

理论实现

- 给出 $|L_\pi - \eta| \le X$ 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

$$|L_\pi(\tilde{\pi}) - \eta(\tilde{\pi})| \leq \frac{4\alpha^2\gamma\epsilon}{(1-\gamma)^2}$$

$$\alpha = D_{TV}^{max}(\pi_{old}, \pi_{new})$$

$$D_{TV}(p||q) = \frac{1}{2}\sum_i |p_i - q_i|$$

$$D_{TV}^{max}(\pi||\tilde{\pi}) = \max_s D_{TV}(\pi(.|s), \tilde{\pi}(.|s))$$

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

$$\eta(\tilde{\pi}) \geq L_\pi(\tilde{\pi}) - \frac{4\alpha^2\gamma\epsilon}{(1-\gamma)^2}$$

TRPO算法 —— 理论实现

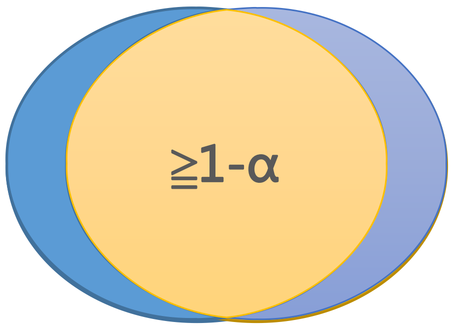
证明 $\left|L_\pi(\tilde{\pi})-\eta(\tilde{\pi})\right| \leq \frac{4\alpha^2\gamma\epsilon}{(1-\gamma)^2}$

$$
\begin{aligned}
&\left|L_\pi(\tilde{\pi})-\eta(\tilde{\pi})\right| \\
=&\sum_{t=0}^{\infty}\gamma^t\left|\mathbb{E}_{\boldsymbol{\tau}\sim\tilde{\pi}}[\bar{A}(s_t)]-\mathbb{E}_{\boldsymbol{\tau}\sim\pi}[\bar{A}(s_t)]\right| \\
=&\sum_{t=0}^{\infty}\gamma^t\left|P(n_t=0)\mathbb{E}_{s_t\sim\tilde{\pi}|n_t=0}[\bar{A}(s_t)]+P(n_t>0)\mathbb{E}_{s_t\sim\tilde{\pi}|n_t>0}[\bar{A}(s_t)]\right. \\
&\left.-\left(P(n_t=0)\mathbb{E}_{s_t\sim\pi|n_t=0}[\bar{A}(s_t)]+P(n_t>0)\mathbb{E}_{s_t\sim\pi|n_t>0}[\bar{A}(s_t)]\right)\right| \\
=&\sum_{t=0}^{\infty}\gamma^t\left|P(n_t>0)\mathbb{E}_{s_t\sim\tilde{\pi}|n_t>0}[\bar{A}(s_t)]-P(n_t>0)\mathbb{E}_{s_t\sim\pi|n_t>0}[\bar{A}(s_t)]\right| \\
\leq&\sum_{t=0}^{\infty}\gamma^t P(n_t>0)\left(\left|\mathbb{E}_{s_t\sim\tilde{\pi}|n_t>0}[\bar{A}(s_t)]\right|+\left|\mathbb{E}_{s_t\sim\pi|n_t>0}[\bar{A}(s_t)]\right|\right)
\end{aligned}
$$

n_t: 前n步，有几步不一样

TRPO算法 —— 理论实现

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



$$|L_\pi(\tilde{\pi}) - \eta(\tilde{\pi})|$$
$$\leq \sum_{t=0}^{\infty} \gamma^t P(n_t > 0)(|\mathbb{E}_{s_t\sim\tilde{\pi}|n_t>0}[\bar{A}(s_t)]| + |\mathbb{E}_{s_t\sim\pi|n_t>0}[\bar{A}(s_t)]|)$$
$$\leq \sum_{t=0}^{\infty} \gamma^t 2(1-(1-\alpha)^t)2\alpha \max_{s,a}|A(s,a)|$$
$$=4\epsilon\alpha \sum_{t=0}^{\infty} \gamma^t (1-(1-\alpha)^t)$$
$$=4\epsilon\alpha(\frac{1}{1-\gamma} - \frac{1}{1-\gamma(1-\alpha)})$$
$$=\frac{4\epsilon\alpha^2\gamma}{(1-\gamma)(1-\gamma(1-\alpha))}$$
$$\leq \frac{4\epsilon\alpha^2\gamma}{(1-\gamma)^2}$$

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

等比级数