彪哥带你学强化学习

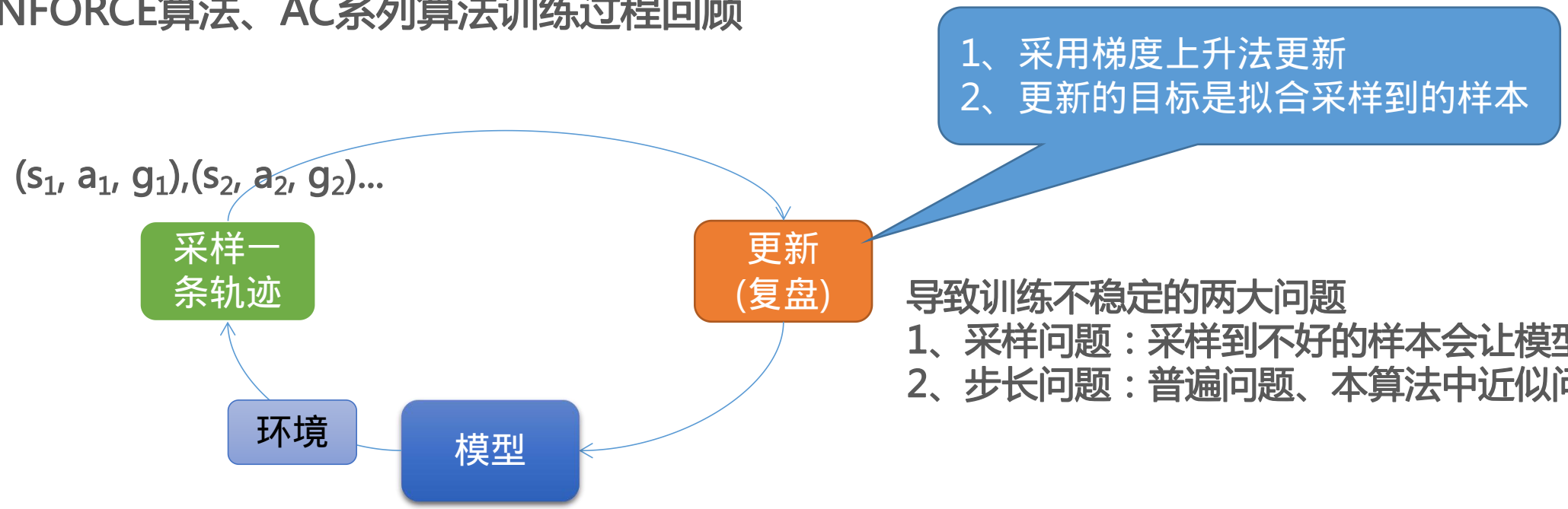
14.深入理解TRPO算法(1)

DEEPLY UNDERSTAND REINFORCEMENT LEARNING

讲师：韩路彪

TRPO解决什么问题

REINFORCE算法、AC系列算法训练过程回顾



导致训练不稳定的两大问题
1、采样问题：采样到不好的样本会让模型变差
2、步长问题：普遍问题、本算法中近似问题

采样问题的本质：采样的轨迹不是最优，导致回报的评估不准确

TRPO算法

步骤0的折扣回报

思路：每训练一步，整体回报g都有提升。整体回报用η符号来表示

$$\eta(\pi) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t)\right]$$

对于一个新策略 $\tilde{\pi}$ 有

$$\eta(\tilde{\pi}) = \eta(\pi) + \underbrace{\mathbb{E}_{\tau\sim\tilde{\pi}}[\sum_{t=0}^{\infty}\gamma^t A_\pi(s_t,a_t)]}$$

只要这堆大于0就有提升

$$\begin{aligned}
&\mathrm{E}_{\tau\sim\tilde{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t A_\pi(s_t,a_t)\right]\\
=&\mathrm{E}_{\tau\sim\tilde{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t (Q_\pi(s_t,a_t)-V_\pi(s_t))\right]\\
=&\mathrm{E}_{\tau\sim\tilde{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t (r(s_t)+\gamma V_\pi(s_{t+1})-V_\pi(s_t))\right]\\
=&\mathrm{E}_{\tau\sim\tilde{\pi}}\left[-V_\pi(s_0)+\sum_{t=0}^{\infty}\gamma^t r(s_t)\right]\\
=&-\mathrm{E}_{s_0}[V_\pi(s_0)]+\mathrm{E}_{\tau\sim\tilde{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t)\right]\\
=&-\eta(\pi)+\eta(\tilde{\pi})
\end{aligned}$$

TRPO算法

$$\eta(\tilde{\pi}) = \eta(\pi) + \underbrace{\mathbb{E}_{\tau\sim\tilde{\pi}}[\sum_{t=0}^{\infty}\gamma^t A_\pi(s_t,a_t)]}_{\text{不好找优化方向}}$$

$$\begin{aligned}
&\mathbb{E}_{\tau\sim\tilde{\pi}}[\sum_{t=0}^{\infty}\gamma^t A_\pi(s_t,a_t)] \\
&= \sum_{t=0}^{\infty}\sum_{s} P(s_t=s|\tilde{\pi})\sum_{a}\tilde{\pi}(a|s)\gamma^t A_\pi(s,a) \\
&= \sum_{s}\sum_{t=0}^{\infty}\gamma^t P(s_t=s|\tilde{\pi})\sum_{a}\tilde{\pi}(a|s) A_\pi(s,a) \\
&= \sum_{s}\rho_{\tilde{\pi}}(s)\sum_{a}\tilde{\pi}(a|s) A_\pi(s,a)
\end{aligned}$$

$$\rho_\pi(s) = P(s_0=s) + \gamma P(s_1=s) + \gamma^2 P(s_2=s) + \dots$$

$$\eta(\tilde{\pi}) = \eta(\pi) + \sum_{s}\rho_{\tilde{\pi}}(s)\sum_{a}\tilde{\pi}(a\mid s)A_\pi(s,a)$$

TRPO算法

$$\eta(\tilde{\pi}) = \eta(\pi) + \sum_s \rho_{\tilde{\pi}}(s) \sum_a \tilde{\pi}(a \mid s) A_\pi(s,a)$$

只要这堆大于0就有提升

$$L_\pi(\tilde{\pi}) = \eta(\pi) + \sum_s \rho_{\pi}(s) \sum_a \tilde{\pi}(a \mid s) A_\pi(s,a)$$

近似的目的：可以直接用原策略计算采样概率，从而去掉对未知策略对依赖

TRPO算法

L_π 和 η 的具体含义是什么

$$\eta(\tilde{\pi}) = \eta(\pi) + \sum_s \rho_{\tilde{\pi}}(s) \sum_a \tilde{\pi}(a \mid s) A_\pi(s,a)$$

$$L_\pi(\tilde{\pi}) = \eta(\pi) + \sum_s \rho_\pi(s) \sum_a \tilde{\pi}(a \mid s) A_\pi(s,a)$$

π

t	1	2	3	…
s_1	0.1	0.2	0.2	
s_2	0.2	0.1	0.4	
s_3	0.3	0.1	0.1	
…				

$\tilde{\pi}$

t	1	2	3	…
s_1	0.1	0.1	0.2	
s_2	0.2	0.2	0.4	
s_3	0.3	0.1	0.1	
…				

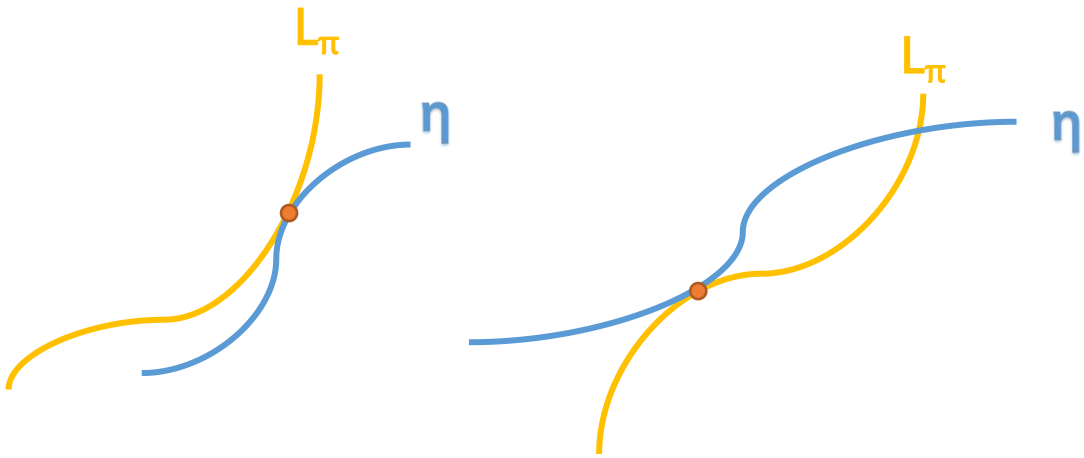
TRPO算法

η 近似成 L_π 的可行性

如果π(θ)可导，则有

$$L_{\pi_{\theta_0}}(\pi_{\theta_0}) = \eta(\pi_{\theta_0})$$

$$\nabla_\theta L_{\pi_{\theta_0}}(\pi_{\theta_0})|_{\theta=\theta_0} = \nabla_\theta \eta(\pi_{\theta_0})|_{\theta=\theta_0}$$



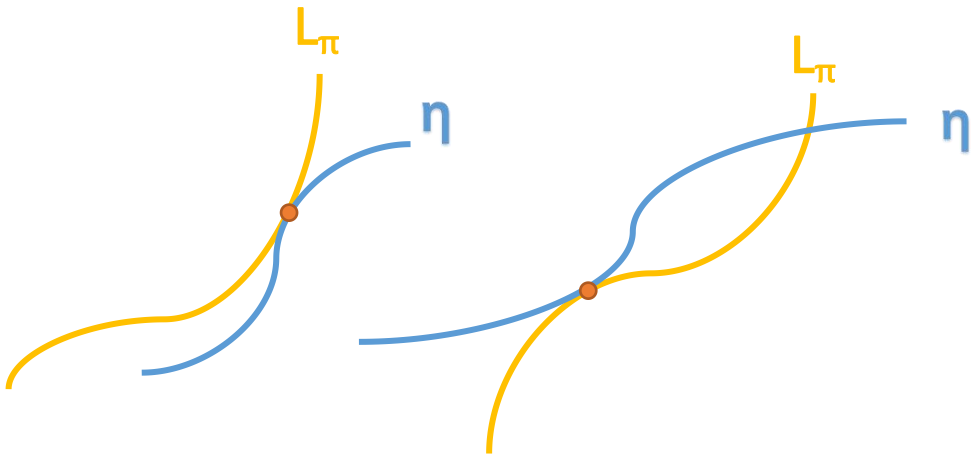
1、L_π 和 η 不能确定谁比谁大
2、在策略π变化不大的前提下两者差距不大
3、对 L_π 在梯度上升方向更新足够小的参数，η也会提升

TRPO算法

如何保证单调改进策略？

如果能找出 L_π 和 η 差距的极限，就有办法做到单调改进策略

假设在更新 L_π 后，可以确保 L_π 和 η 差距不超过 X(>0)
那么，在 L_π 增长的前提下，η 最多减少-X
如果能保证 L_π 增长量至少超过X，就能保证η 一定会增长



假设 $|L_\pi - \eta| <= 5$

如果 L_π 增长了6，η至少增长1（最多增长11）