# 强化学习基础及其在大语言模型中的应用

王榕甄

2025 年 7 月 3 日

# 提纲

- 强化学习基本概念及发展历程
- 策略梯度方法
- 策略改进方法
- 大语言模型中的强化对齐

# 强化学习的基本概念：三个随机变量、两个概率

- “强化学习是一种**从个体和环境交互的经验**中学习如何**将状态映射到动作**，以获得**最大奖励**的学习机制。”
- 状态 $s$, 动作 $a$, 奖励 $r$
- 个体 (agent)： 策略 $\pi(a|s)$
- 环境 (environment)：转移概率 $p(s', r|s, a)$

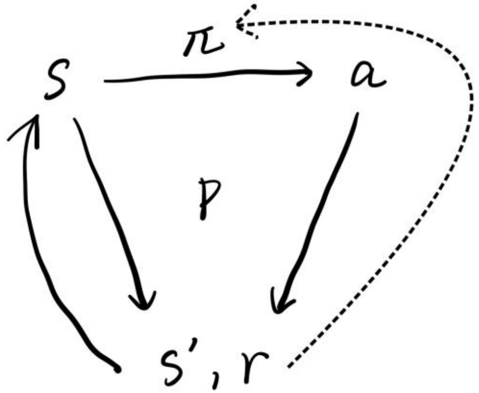
- 经过个体与环境 $t = 0, 1, \cdots, T-1$ 次交互，形成轨迹 (trajectory or episode)

$$\tau = s_0, a_0, r_1, s_1, a_1, \cdots, s_{T-1}, a_{T-1}, r_T, s_T$$

- 其中 $(s, a, r, s')$ 为一步转移 (one transition step)； 当 $t > T$ 时, $r_t = 0$

# 强化学习的问题建模

- 建模：马尔可夫决策过程 (Markov Decision Process,或随机序贯决策, 或随机动态规划)




- 目标：找到一个**好策略**以最大化**期望累积（折扣）奖励**

$$J(\pi) = \mathbb{E}_{\tau\sim p_\pi(\tau)}\left[\sum_{t=0}^{\infty}\gamma^t r_{t+1}\right], \qquad \max_\pi J(\pi)$$

为什么要折扣？

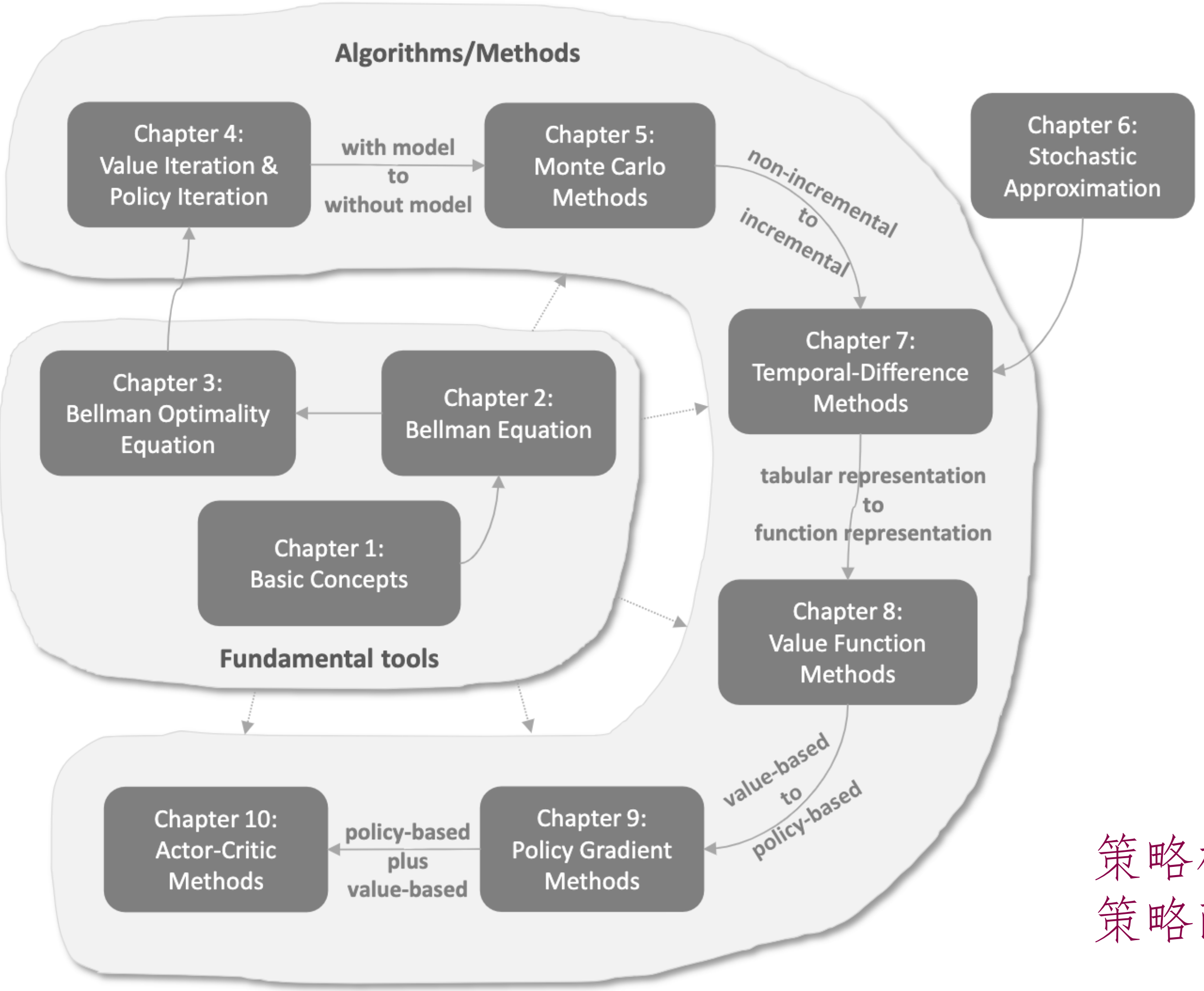
- $p_\pi(\tau) = p_\pi(s_0, a_0, r_1, s_1, \cdots) = p(s_0)\prod_{t=0}^{\infty}\pi(a_t|s_t)\,p(s_{t+1}, r_{t+1}|s_t, a_t)$

# 强化学习的基本概念：价值函数

- 如何衡量某一状态下执行某一动作的好坏？

- 动作价值函数 $Q_{\pi}(s_t, a_t) = \mathbb{E}_{r_{t+1}, s_{t+1}, \cdots \sim p_{\pi}(\cdot|s_t, a_t)}[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}]$
- 状态价值函数 $V_{\pi}(s_t) = \mathbb{E}_{a_t, r_{t+1}, s_{t+1}, \cdots \sim p_{\pi}(\cdot|s_t)}[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}] = \mathbb{E}_{a_t \sim \pi(\cdot|s_t)}[Q_{\pi}(s_t, a_t)]$
- 优势函数 $A_{\pi}(s_t, a_t) = Q_{\pi}(s_t, a_t) - V_{\pi}(s_t)$

# 强化学习的发展历程




策略梯度方法
策略改进方法

# 策略梯度方法 (Policy Gradient Methods)

- 回顾： $\max_\pi J(\pi) = \mathbb{E}_{\tau\sim p_\pi(\tau)}[\sum_{t=0}^{\infty}\gamma^t r_{t+1}]$
- 参数化策略函数： $\pi_\theta$
- 目标：

$$\max_\theta J(\theta) = \mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}\left[\sum_{t=0}^{\infty}\gamma^t r_{t+1}\right]$$

- 策略梯度方法： $\theta \leftarrow \theta + \alpha\nabla_\theta J(\theta)$

# 策略梯度

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}\left[\left(\sum_{t=0}^{\infty}\nabla_\theta \log \pi_\theta(a_t|s_t)\right)\left(\sum_{t=0}^{\infty}\gamma^t r_{t+1}\right)\right]$$

## **Step 1:** Log-Derivative Trick

- $J(\theta) = \mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}[\sum_{t=0}^{\infty}\gamma^t r_{t+1}]$
- 令 $f(\tau) \triangleq \sum_{t=0}^{\infty}\gamma^t r_{t+1}$
- 则

## **Step 2:** Unrolling Log Probability

- $p_\pi(\tau) = p(s_0)\prod_{t=0}^{\infty}\pi(a_t|s_t)\, p(s_{t+1}, r_{t+1}|s_t, a_t)$
- 则

Seita's Place: Going Deeper Into Reinforcement Learning: Fundamentals of Policy Gradients

注：策略梯度定理有另一种表述，参见：Richard S. Sutton and Andrew G. Barto, 2020, Reinforcement Learning: An Introduction

$\max_\pi J(\pi) = \mathbb{E}_{\tau \sim p_\pi(\tau)}[\sum_{t=0}^{\infty} \gamma^t r_{t+1}]$

# REINFORCE 算法

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim p_{\pi_\theta}(\tau)}\left[\left(\sum_{t=0}^{\infty} \nabla_\theta \log \pi_\theta(a_t|s_t)\right)\left(\sum_{t=0}^{\infty} \gamma^t r_{t+1}\right)\right]$$

- REINFORCE algorithm:

1. sample $\{\tau^i\}$ from $\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)$ (run the policy)
2. $\nabla_\theta J(\theta) \approx \sum_i \left(\sum_t \nabla_\theta \log \pi_\theta(\mathbf{a}_t^i|\mathbf{s}_t^i)\right)\left(\sum_t r(\mathbf{s}_t^i, \mathbf{a}_t^i)\right)$
3. $\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$

缺点：方差高、更新慢

- REINFORCE 的变体

$$V_\pi(s_t) = \mathbb{E}_{a_t, r_{t+1}, s_{t+1}, \cdots \sim p_\pi(\cdot|s_t)}[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}]$$

  - REINFORCE with Baseline： $\sum_{t=0}^{\infty} \gamma^t r_{t+1} - V_\pi^\phi(s_0)$ 无偏，优势函数
  - Advantage Actor-Critic (A2C)： $r_1 + \gamma V_\pi^\phi(s_1) - V_\pi^\phi(s_0)$ 很可能有偏
  - Actor-Critic with GAE 结合以上二者

Figure: Sergey Levine, CS 285 UC Berkeley, Deep Reinforcement Learning, 2023
Richard S. Sutton and Andrew G. Barto, 2020, Reinforcement Learning: An Introduction

# 广义优势估计 (General Advantage Estimation, GAE)

- $\sum_{k=0}^{\infty}\gamma^{k+t}r_{t+k+1} - V_\pi(s_t)$
- $r_{t+1} + \gamma V_\pi(s_{t+1}) - V_\pi(s_t)$

$$\boxed{\delta_t^V = r_t + \gamma V(s_{t+1}) - V(s_t)}$$

- 观察：

$$\begin{aligned}
\hat{A}_t^{(1)} &= -V(s_t) + r_t + \gamma V(s_{t+1}) &&= \delta_t^V \\
\hat{A}_t^{(2)} &= -V(s_t) + r_t + \gamma r_{t+1} + \gamma^2 V(s_{t+2}) &&= \delta_t^V + \gamma\delta_{t+1}^V \\
\hat{A}_t^{(3)} &= -V(s_t) + r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 V(s_{t+3}) &&= \delta_t^V + \gamma\delta_{t+1}^V + \gamma^2\delta_{t+2}^V
\end{aligned}$$

- 加权平均：

$$\begin{aligned}
\hat{A}_t^{\mathrm{GAE}(\gamma,\lambda)} &:= (1-\lambda)\left(\hat{A}_t^{(1)} + \lambda\hat{A}_t^{(2)} + \lambda^2\hat{A}_t^{(3)} + \ldots\right) \\
&= (1-\lambda)(\delta_t^V + \lambda(\delta_t^V + \gamma\delta_{t+1}^V) + \lambda^2(\delta_t^V + \gamma\delta_{t+1}^V + \gamma^2\delta_{t+2}^V) + \ldots) \\
&= (1-\lambda)(\delta_t^V(1 + \lambda + \lambda^2 + \ldots) + \gamma\delta_{t+1}^V(\lambda + \lambda^2 + \lambda^3 + \ldots) \\
&\qquad + \gamma^2\delta_{t+2}^V(\lambda^2 + \lambda^3 + \lambda^4 + \ldots) + \ldots) \\
&= (1-\lambda)\left(\delta_t^V\left(\frac{1}{1-\lambda}\right) + \gamma\delta_{t+1}^V\left(\frac{\lambda}{1-\lambda}\right) + \gamma^2\delta_{t+2}^V\left(\frac{\lambda^2}{1-\lambda}\right) + \ldots\right) \\
&= \sum_{l=0}^{\infty}(\gamma\lambda)^l\delta_{t+l}^V
\end{aligned}$$

# 广义优势估计 (General Advantage Estimation, GAE)

- GAE：

$$\hat{A}_t^{\mathrm{GAE}(\gamma,\lambda)} := \sum_{l=0}^{\infty}(\gamma\lambda)^l \delta_{t+l}^V$$

$$\mathrm{GAE}(\gamma,0): \quad \hat{A}_t := \delta_t \qquad = r_t + \gamma V(s_{t+1}) - V(s_t)$$

$$\mathrm{GAE}(\gamma,1): \quad \hat{A}_t := \sum_{l=0}^{\infty}\gamma^l \delta_{t+l} = \sum_{l=0}^{\infty}\gamma^l r_{t+l} - V(s_t)$$

- 价值函数的估计与优化：
  - 参数化： $V_\pi(s_t) \approx V_\pi^\phi(s_t)$
  - 优化： $\min_\phi \frac{1}{N|\tau^i|}\sum_{i=1}^{N}\left\|V_\pi^\phi(s_t^i) - \hat{R}_t^i\right\|_2^2, \hat{R}_t = \sum_{k=0}^{T-t-1}\gamma^{k+t} r_{t+k+1}$

Schulman, J., Moritz, P., Levine, S., Jordan, M., & Abbeel, P. (2015). High-Dimensional Continuous Control Using Generalized Advantage Estimation.

# 策略改进方法 (Policy Improvement Methods)

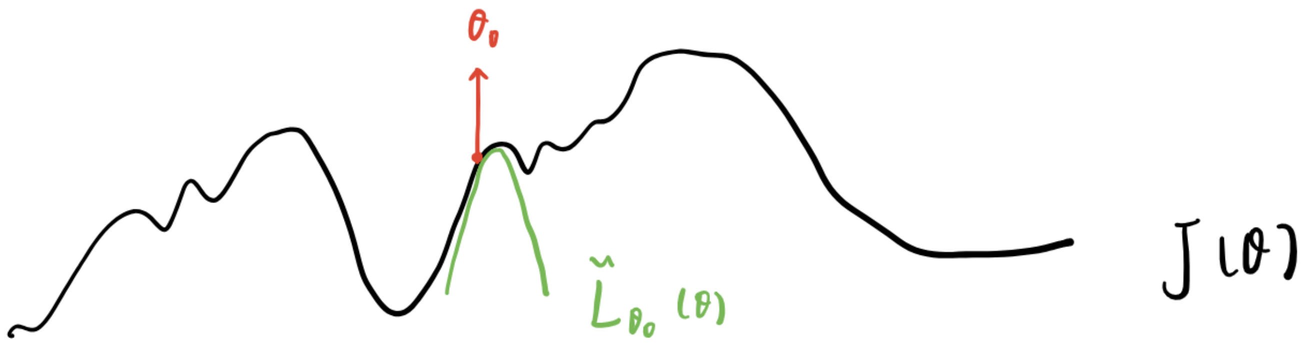
- 策略梯度方法：梯度上升，用一阶泰勒展开近似指导目标函数上升

$$\theta_1 \leftarrow \theta_0 + \alpha\nabla_\theta J(\theta)|_{\theta=\theta_0} \qquad \Rightarrow \qquad J(\theta_1) \geq J(\theta_0)\,?$$

- 如何能保证目标函数上升？

- 考虑一个替代函数 $\tilde{L}(\theta)$，满足 $\tilde{L}(\theta) \leq J(\theta)$ 且 $\tilde{L}(\theta_0) = J(\theta_0)$

$$\theta_1 \leftarrow \mathrm{argmax}_\theta \tilde{L}(\theta) \qquad \Rightarrow \qquad J(\theta_1) \geq \tilde{L}(\theta_1) \geq \tilde{L}(\theta_0) = J(\theta_0)$$



# 替代函数的构造

- **第一步：**基于 $J(\theta_0)$ 的 $J(\theta)$

$$J(\theta) = J(\theta_0) + \mathbb{E}_{s\sim\rho_{\pi_\theta}}\left[\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[A_{\pi_{\theta_0}}(s,a)\right]\right]$$

- 其中 $\rho_\pi(s) = \Pr(s_0 = s|\pi,p) + \gamma\Pr(s_1 = s|\pi,p) + \gamma^2\Pr(s_2 = s|\pi,p) + \cdots$
- 推导：
  - **Step 1** ：$J(\theta) = J(\theta_0) + \mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}\left[\sum_{t=0}^{\infty}\gamma^t A_{\pi_{\theta_0}}(s_t,a_t)\right]$
  - **Step 2** ：$\mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}\left[\sum_{t=0}^{\infty}\gamma^t A_{\pi_{\theta_0}}(s_t,a_t)\right] = \mathbb{E}_{s\sim\rho_{\pi_\theta}}\left[\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[A_{\pi_{\theta_0}}(s,a)\right]\right]$

# 替代函数的构造

**Step 1.** $J(\theta) = J(\theta_0) + E_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^{t} A_{\pi_{\theta_0}}(s_t,a_t)\right]$

证明:

- $A_{\pi_{\theta_0}}(s_t,a_t) = Q_{\pi_{\theta_0}}(s_t,a_t) - V_{\pi_{\theta_0}}(s_t)$

$$= E_{r_{t+1},s_{t+1},\cdots\sim\pi_{\theta_0}|s_t,a_t}\left[\sum_{k=0}^{\infty}\gamma^k r_{t+k+1}\right] - V_{\pi_{\theta_0}}(s_t)$$

$$= E_{r_{t+1},s_{t+1}\sim p(r_{t+1},s_{t+1}|s_t,a_t)}\left[r_{t+1}+\gamma V_{\pi_{\theta_0}}(s_{t+1})\right] - V_{\pi_{\theta_0}}(s_t)$$

- $E_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t A_{\pi_{\theta_0}}(s_t,a_t)\right] = \sum_{t=0}^{\infty}\overset{\gamma^t}{E}_{\tau\sim\pi_\theta}\left[E_{r_{t+1},s_{t+1}\sim p(r_{t+1},s_{t+1}|s_t,a_t)}[r_{t+1}+\gamma V_{\pi_{\theta_0}}(s_{t+1})] - V_{\pi_{\theta_0}}(s_t)\right]$

$$= \sum_{t=0}^{\infty}\gamma^t E_{s_0,\cdots,s_t,a_t,r_{t+1},s_{t+1}\sim\pi_\theta}\left[r_{t+1}+\gamma V_{\pi_{\theta_0}}(s_{t+1}) - V_{\pi_{\theta_0}}(s_t)\right]$$

$$= \sum_{t=0}^{\infty}\gamma^t E_{\tau\sim\pi_\theta}\left[r_{t+1}+\gamma V_{\pi_{\theta_0}}(s_{t+1}) - V_{\pi_{\theta_0}}(s_t)\right]$$

$$= E_{\tau\sim\pi_\theta}\big[r_1 + \cancel{\gamma V_{\pi_{\theta_0}}(s_1)} - V_{\pi_{\theta_0}}(s_0)$$
$$+\gamma r_2 + \cancel{\gamma^2 V_{\pi_{\theta_0}}(s_2)} - \cancel{\gamma V_{\pi_{\theta_0}}(s_1)}$$
$$+\cdots$$

$$= E_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t r_{t+1} - V_{\pi_{\theta_0}}(s_0)\right]$$

$$= E_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t r_{t+1}\right] - E_{s_0\sim p(s_0)}\left[V_{\pi_{\theta_0}}(s_0)\right]$$

$$= J(\theta) - J(\theta_0)$$

**Step 2.** $E_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t A_{\pi_{\theta_0}}(s_t,a_t)\right] = E_{s\sim\rho_{\pi_\theta}} E_{a\sim\pi_\theta(\cdot|s)}\left[A_{\pi_{\theta_0}}(s,a)\right]$

证明:

- $E_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t A_{\pi_{\theta_0}}(s_t,a_t)\right] = \sum_{t=0}^{\infty}\gamma^t E_{\underbrace{s_0,\cdots,}_{\tau_{<t}}s_t,a_t,\cdots\sim\pi_\theta}\left[A_{\pi_{\theta_0}}(s_t,a_t)\right]$

$$= \sum_{t=0}^{\infty}\gamma^t\sum_{\tau_{<t}}\sum_{s_t,a_t} P_{\pi_\theta}(\tau_{<t},s_t,a_t)\left[A_{\pi_{\theta_0}}(s_t,a_t)\right]$$

$$= \sum_{t=0}^{\infty}\gamma^t\sum_{s_t,a_t} P_{\pi_\theta}(s_t,a_t)\left[A_{\pi_{\theta_0}}(s_t,a_t)\right]$$

$$= \sum_{t=0}^{\infty}\gamma^t\sum_{s}\sum_{a} P_{\pi_\theta}(s_t=s,a_t=a)\left[A_{\pi_{\theta_0}}(s,a)\right]$$

$$= \sum_{t=0}^{\infty}\gamma^t\sum_{s}\sum_{a} P_{\pi_\theta}(s_t=s)\,\pi_\theta(s|a)\left[A_{\pi_{\theta_0}}(s,a)\right]$$

$\pi_\theta$ 和 $A_{\pi_\theta}$ 不区分时刻

$$= \sum_{s}\underbrace{\left(\sum_{t=0}^{\infty}\gamma^t P_{\pi_\theta}(s_t=s)\right)}_{\triangleq\,\rho_{\pi_\theta}(s)}\sum_a\pi_\theta(s|a)\left[A_{\pi_{\theta_0}}(s,a)\right]$$

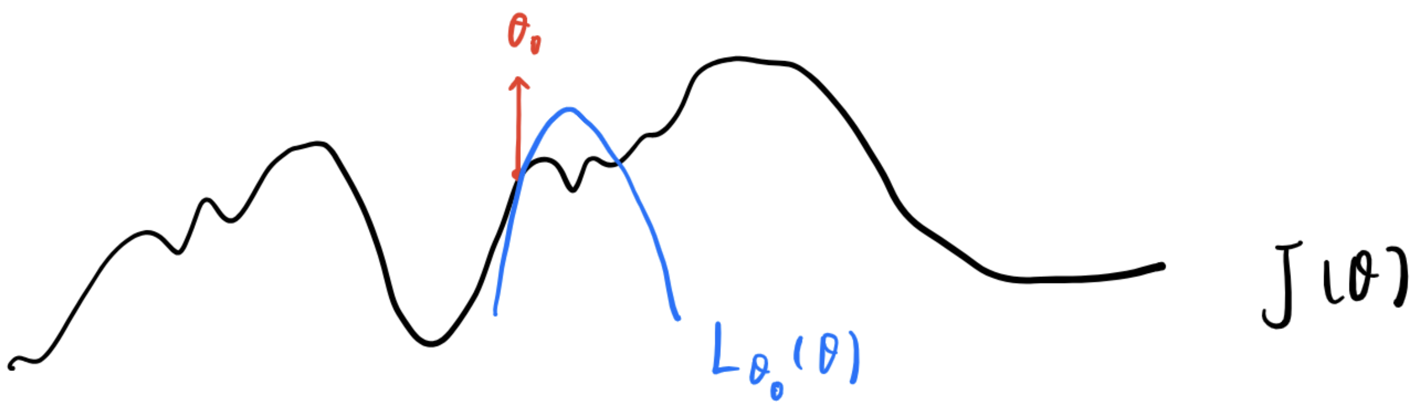
$$= E_{s\sim\rho_{\pi_\theta}} E_{a\sim\pi_\theta(\cdot|s)}\left[A_{\pi_{\theta_0}}(s,a)\right]$$

# 替代函数的构造

- **第二步：**近似

$$J(\theta) \approx J(\theta_0) + \mathbb{E}_{s\sim\rho_{\pi_{\theta_0}}}\left[\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[A_{\pi_{\theta_0}}(s,a)\right]\right] = L(\theta)$$

- 满足条件：$L(\theta_0) = J(\theta_0)$ 且 $\nabla_\theta L(\theta)|_{\theta=\theta_0} = \nabla_\theta J(\theta)|_{\theta=\theta_0}$



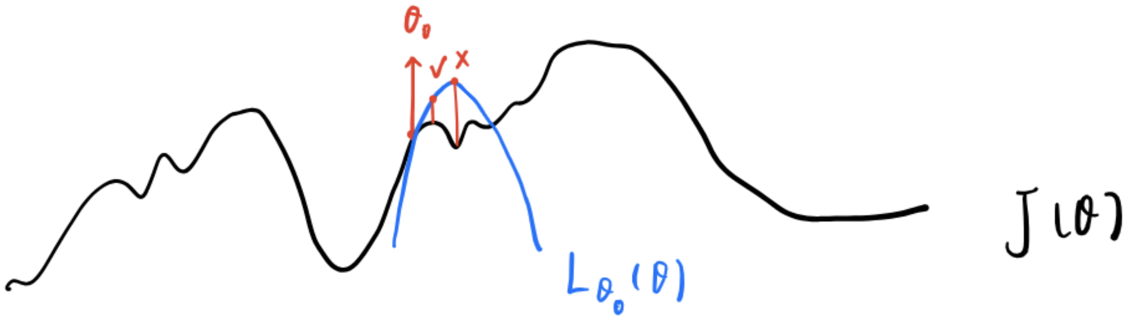
# 替代函数的构造

- **第三步：** 下界定理

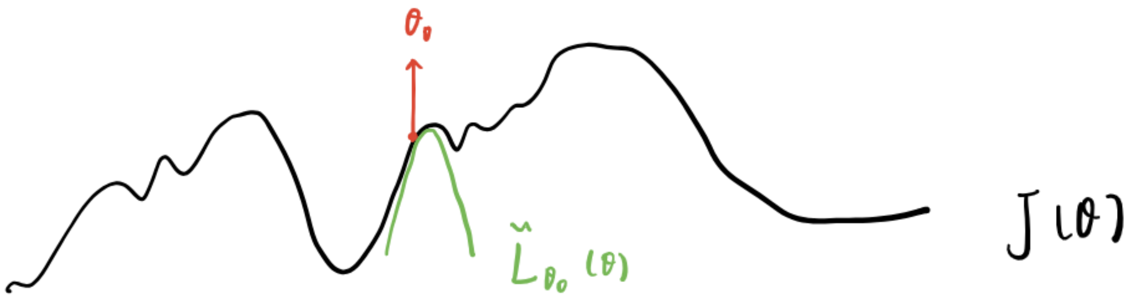
$$J(\theta) \geq L(\theta) - C \max_s D_{\mathrm{KL}}\left(\pi_{\theta_0}(\cdot|s) \| \pi_\theta(\cdot|s)\right) = \tilde{L}(\theta)$$

- $L(\theta) \triangleq J(\theta_0) + \mathbb{E}_{s\sim\rho_{\pi_\theta}}\left[\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[A_{\pi_{\theta_0}}(s,a)\right]\right], \quad C \triangleq \frac{4\gamma \max_{s,a}\left|A_{\pi_{\theta_0}}(s,a)\right|}{(1-\gamma)^2}$
- $\rho_\pi(s) \triangleq \Pr(s_0 = s|\pi,p) + \gamma\Pr(s_1 = s|\pi,p) + \gamma^2\Pr(s_2 = s|\pi,p) + \cdots$

$L(\theta_0) = J(\theta_0)$ 且 $\nabla_\theta L(\theta)|_{\theta=\theta_0} = \nabla_\theta J(\theta)|_{\theta=\theta_0}$




$\tilde{L}(\theta) \leq J(\theta)$ 且 $\tilde{L}(\theta_0) = J(\theta_0)$



# 信赖域策略优化 (Trust Region Policy Optimization, TRPO)

- 有保证的目标： $\max_\theta L(\theta) - C \max_s D_{\mathrm{KL}}\left(\pi_{\theta_0}(\cdot|s) \| \pi_\theta(\cdot|s)\right)$ >> 步长过小！
- **替代地**，我们优化：

信赖域约束

$$\max_\theta L(\theta), \quad \text{subject to} \quad \max_s D_{\mathrm{KL}}\left(\pi_{\theta_0}(\cdot|s) \| \pi_\theta(\cdot|s)\right) \le \delta$$

- **进一步替代地**，我们优化：

$$\max_\theta L(\theta), \quad \text{subject to} \quad \mathbb{E}_{s\sim\rho_{\pi_{\theta_0}}}\left[D_{\mathrm{KL}}\left(\pi_{\theta_0}(\cdot|s) \| \pi_\theta(\cdot|s)\right)\right] \le \delta$$

Schulman, J., EDU, B., Levine, S., Moritz, P., Jordan, M., & Abbeel, P. (2015). Trust Region Policy Optimization.

# 采样估计

- 目标函数：$\max_\theta L(\theta) = \max_\theta J(\theta_0) + \mathbb{E}_{s\sim\rho_{\pi_{\theta_0}}}\left[\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[A_{\pi_{\theta_0}}(s,a)\right]\right]$

$$= \max_\theta \mathbb{E}_{s\sim\rho_{\pi_{\theta_0}}}\left[\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[A_{\pi_{\theta_0}}(s,a)\right]\right]$$

$$= \max_\theta \mathbb{E}_{s\sim\rho_{\pi_{\theta_0}}}\left[\mathbb{E}_{a\sim\pi_{\theta_0}(\cdot|s)}\left[\frac{\pi_\theta(a|s)}{\pi_{\theta_0}(a|s)} A_{\pi_{\theta_0}}(s,a)\right]\right]$$

重要性采样 (Importance Sampling) 技巧

其中 $\rho_\pi(s) = \Pr(s_0 = s|\pi,p) + \gamma\Pr(s_1 = s|\pi,p) + \gamma^2\Pr(s_2 = s|\pi,p) + \cdots$

$$= \max_\theta \sum_{t=0}^{\infty} \mathbb{E}_{s_t,a_t\sim\pi_{\theta_0}}\left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_0}(a_t|s_t)}\gamma^t A_{\pi_{\theta_0}}(s_t,a_t)\right]$$

蒙特卡洛采样

$$\approx \max_\theta \frac{1}{N}\sum_{i=1}^{N}\sum_{t=0}^{|\tau^i|-1}\left[\frac{\pi_\theta(a_t^i|s_t^i)}{\pi_{\theta_0}(a_t^i|s_t^i)}\gamma^t A_{\pi_{\theta_0}}(s_t^i,a_t^i)\right]$$

- 约束：$\mathbb{E}_{s\sim\rho_{\pi_{\theta_0}}}\left[D_{\mathrm{KL}}\left(\pi_{\theta_0}(\cdot|s)\|\pi_\theta(\cdot|s)\right)\right] \approx \frac{1}{N}\sum_{i=1}^{N}\sum_{t=0}^{|\tau^i|-1}\left[\gamma^t \widehat{D}_{\mathrm{KL}}\left(\pi_{\theta_0}(\cdot|s_t^i)\|\pi_\theta(\cdot|s_t^i)\right)\right]$
- 优势：$\hat{A}_t^{\mathrm{GAE}(\gamma,\lambda)} = \sum_{l=0}^{\infty}(\gamma\lambda)^l \delta_{t+l}^V \qquad \delta_t^V = r_t + \gamma V(s_{t+1}) - V(s_t)$

GAE

➢ $max_\pi J(\pi) = \mathbb{E}_{\tau\sim p_\pi(\tau)}[\sum_{t=0}^{\infty}\gamma^t r_{t+1}]$

# 实际算法

**Algorithm 1** Trust Region Policy Optimization

1: Input: initial policy parameters $\theta_0$, initial value function parameters $\phi_0$

2: Hyperparameters: KL-divergence limit $\delta$, backtracking coefficient $\alpha$, maximum number of backtracking steps $K$

3: **for** $k = 0, 1, 2, \ldots$ **do**

4: Collect set of trajectories $\mathcal{D}_k = \{\tau_i\}$ by running policy $\pi_k = \pi(\theta_k)$ in the environment.

5: Compute rewards-to-go $\hat{R}_t$.

6: Compute advantage estimates, $\hat{A}_t$ (using any method of advantage estimation) based on the current value function $V_{\phi_k}$.

1. **做近似**——构造函数 $\tilde{L}$ 近似目标函数 $J(\boldsymbol{\theta})$：

(a). 设当前策略网络参数是 $\boldsymbol{\theta}_{\text{now}}$。用策略网络 $\pi(a\,|\,s;\,\boldsymbol{\theta}_{\text{now}})$ 控制智能体与环境交互，玩完一局游戏，记录下轨迹：

$$s_1,\ a_1,\ r_1,\quad s_2,\ a_2,\ r_2,\quad \cdots,\quad s_n,\ a_n,\ r_n.$$

(b). 对于所有的 $t = 1, \cdots, n$，计算折扣回报 $u_t = \sum_{k=t}^{n}\gamma^{k-t}\cdot r_k$。

(c). 得出近似函数：

$$\tilde{L}(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}_{\text{now}}) = \sum_{t=1}^{n}\frac{\pi(a_t\,|\,s_t;\,\boldsymbol{\theta})}{\pi(a_t\,|\,s_t;\boldsymbol{\theta}_{\text{now}})}\cdot u_t.$$

2. **最大化**——用某种数值算法求解带约束的最大化问题：

$$\boldsymbol{\theta}_{\text{new}} = \underset{\boldsymbol{\theta}}{\text{argmax}}\ \tilde{L}(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}_{\text{now}}); \qquad \text{s.t.}\ \left\|\boldsymbol{\theta}-\boldsymbol{\theta}_{\text{now}}\right\|_2 \le \Delta.$$

此处的约束条件是二范数距离。可以把它替换成 KL 散度，即公式 (9.10)。

$$\max_\theta L(\theta)$$

$$= \max_\theta \sum_{t=0}^{\infty}\mathbb{E}_{s_t,a_t\sim\pi_{\theta_0}}\left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_0}(a_t|s_t)}A_{\pi_{\theta_0}}(s_t,a_t)\right]$$

$$\approx \max_\theta \frac{1}{N}\sum_{i=1}^{N}\sum_{t=0}^{|\tau^i|-1}\left[\frac{\pi_\theta(a_t^i|s_t^i)}{\pi_{\theta_0}(a_t^i|s_t^i)}\gamma^t A_{\pi_{\theta_0}}(s_t^i,a_t^i)\right]$$

OpenAI, Welcome to Spinning Up in Deep RL!
王树森 张志华, 2022, 深度强化学习（初稿）

# 实际算法

**Algorithm 1** Trust Region Policy Optimization

$\max_\pi J(\pi) = \mathbb{E}_{\tau\sim p_\pi(\tau)}[\sum_{t=0}^{\infty}\gamma^t r_{t+1}]$

1: Input: initial policy parameters $\theta_0$, initial value function parameters $\phi_0$

2: Hyperparameters: KL-divergence limit $\delta$, backtracking coefficient $\alpha$, maximum number of backtracking steps $K$

3: **for** $k = 0, 1, 2, \ldots$ **do**

4: Collect set of trajectories $\mathcal{D}_k = \{\tau_i\}$ by running policy $\pi_k = \pi(\theta_k)$ in the environment.

5: Compute rewards-to-go $\hat{R}_t$.

6: Compute advantage estimates, $\hat{A}_t$ (using any method of advantage estimation) based on the current value function $V_{\phi_k}$.

7: Estimate policy gradient as

$$\hat{g}_k = \frac{1}{|\mathcal{D}_k|}\sum_{\tau\in\mathcal{D}_k}\sum_{t=0}^{T}\nabla_\theta \log \pi_\theta(a_t|s_t)|_{\theta_k}\hat{A}_t.$$

$$\max_\theta L(\theta)$$
$$= \max_\theta \sum_{t=0}^{\infty}\mathbb{E}_{s_t,a_t\sim\pi_{\theta_0}}\left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_0}(a_t|s_t)}A_{\pi_{\theta_0}}(s_t,a_t)\right]$$
$$\approx \max_\theta \frac{1}{N}\sum_{i=1}^{N}\sum_{t=0}^{|\tau^i|-1}\left[\frac{\pi_\theta(a_t^i|s_t^i)}{\pi_{\theta_0}(a_t^i|s_t^i)}\gamma^t A_{\pi_{\theta_0}}(s_t^i,a_t^i)\right]$$

8: Use the conjugate gradient algorithm to compute

$$\hat{x}_k \approx \hat{H}_k^{-1}\hat{g}_k,$$

where $\hat{H}_k$ is the Hessian of the sample average KL-divergence.

9: Update the policy by backtracking line search with

$$\theta_{k+1} = \theta_k + \alpha^j\sqrt{\frac{2\delta}{\hat{x}_k^T\hat{H}_k\hat{x}_k}}\hat{x}_k,$$

where $j \in \{0, 1, 2, \ldots K\}$ is the smallest value which improves the sample loss and satisfies the sample KL-divergence constraint.

10: Fit value function by regression on mean-squared error:

$$\phi_{k+1} = \arg\min_\phi \frac{1}{|\mathcal{D}_k|T}\sum_{\tau\in\mathcal{D}_k}\sum_{t=0}^{T}\left(V_\phi(s_t) - \hat{R}_t\right)^2,$$

typically via some gradient descent algorithm.

11: **end for**

OpenAI, Welcome to Spinning Up in Deep RL!

# 近端策略优化 (Proximal Policy Optimization, PPO)

- 回顾：
  - 下界定理：$J(\theta) \geq L(\theta) - C D_{\mathrm{KL}}^{\max}(\pi_{\theta_0}, \pi_\theta)$， $C = \frac{4\gamma \max_{s,a}\left|A_{\pi_{\theta_0}}(s,a)\right|}{(1-\gamma)^2}$
  - $\max_\theta L(\theta) = \max_\theta \sum_{t=0}^{\infty} \mathbb{E}_{s_t,a_t\sim\pi_{\theta_0}}\left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_0}(a_t|s_t)} A_{\pi_{\theta_0}}(s_t,a_t)\right]$ s.t. $\max_s D_{\mathrm{KL}}\left(\pi_{\theta_0}(\cdot|s)\|\pi_\theta(\cdot|s)\right) \leq \delta$
- PPO-Penalty： $\max_\theta L(\theta) - \beta D_{\mathrm{KL}}^{\max}(\pi_{\theta_0}, \pi_\theta)$
- PPO-Clip：

$$\max_\theta \sum_{t=0}^{\infty} \mathbb{E}_{s_t,a_t\sim\pi_{\theta_0}}\left[\min\left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_0}(a_t|s_t)} A_{\pi_{\theta_0}}(s_t,a_t), \mathrm{clip}\left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_0}(a_t|s_t)}, 1-\epsilon, 1+\epsilon\right) A_{\pi_{\theta_0}}(s_t,a_t)\right)\right]$$

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal Policy Optimization Algorithms.

# PPO-Clip

**优势为正**：假设该状态-动作对的优势为正，在这种情况下，其对目标的贡献减少为

$$L(s,a,\theta_k,\theta) = \min\left(\frac{\pi_\theta(a|s)}{\pi_{\theta_k}(a|s)}, (1+\epsilon)\right) A^{\pi_{\theta_k}}(s,a).$$

由于优势为正，如果行动的可能性增大（即$\pi_\theta(a|s)$增加），目标也会随之增加。但该项中的最小值限制了目标的增幅。一旦，最小值就会生效，该项就会达到 的上限。因此：*远离旧策略 并不会给新策略带来好处*。$\pi_\theta(a|s) > (1+\epsilon)\pi_{\theta_k}(a|s)(1+\epsilon)A^{\pi_{\theta_k}}(s,a)$

**优势为负**：假设该状态-动作对的优势为负，在这种情况下，其对目标的贡献减少为

$$L(s,a,\theta_k,\theta) = \max\left(\frac{\pi_\theta(a|s)}{\pi_{\theta_k}(a|s)}, (1-\epsilon)\right) A^{\pi_{\theta_k}}(s,a).$$

由于优势为负，如果行动发生的可能性降低（即$\pi_\theta(a|s)$减小），目标就会增加。但该项中的最大值限制了目标的增幅。一旦，最大值就会生效，该项就会达到 的上限。因此，再次强调：*远离旧策略并不会给新策略带来好处*。$\pi_\theta(a|s) < (1-\epsilon)\pi_{\theta_k}(a|s)(1-\epsilon)A^{\pi_{\theta_k}}(s,a)$

到目前为止，我们所看到的是，裁剪通过消除政策发生剧烈变化的动机，起到了正则化的作用，而超参数$\epsilon$对应于新政策与旧政策相差多远，同时仍然有利于实现目标。

OpenAI, Welcome to Spinning Up in Deep RL!

# 实际算法

**Algorithm 1** PPO-Clip

Input: initial policy parameters $\theta_0$, initial value function parameters $\phi_0$
**for** $k = 0, 1, 2, \ldots$ **do**
    Collect set of trajectories $\mathcal{D}_k = \{\tau_i\}$ by running policy $\pi_k = \pi(\theta_k)$ in the environment.
    Compute rewards-to-go $\hat{R}_t$.
    Compute advantage estimates, $\hat{A}_t$ (using any method of advantage estimation) based on the current value function $V_{\phi_k}$.
    Update the policy by maximizing the PPO-Clip objective:

$$\theta_{k+1} = \arg\max_{\theta} \frac{1}{|\mathcal{D}_k|T} \sum_{\tau\in\mathcal{D}_k} \sum_{t=0}^{T} \min\left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_k}(a_t|s_t)} A^{\pi_{\theta_k}}(s_t, a_t),\ \ g(\epsilon, A^{\pi_{\theta_k}}(s_t, a_t))\right),$$

    typically via stochastic gradient ascent with Adam.
    Fit value function by regression on mean-squared error:

$$\phi_{k+1} = \arg\min_{\phi} \frac{1}{|\mathcal{D}_k|T} \sum_{\tau\in\mathcal{D}_k} \sum_{t=0}^{T} \left(V_\phi(s_t) - \hat{R}_t\right)^2,$$

    typically via some gradient descent algorithm.
**end for**

OpenAI, Welcome to Spinning Up in Deep RL!

# 大语言模型中的强化学习

# 大语言模型中的强化学习：从 RLHF 说起

- 问题建模：单步策略，学习 question 对应的 output

$$J(\theta) = \mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}\left[\sum_{t=0}^{1}\gamma^t r_{t+1}\right] = \mathbb{E}_{s_0\sim p(s_0),a_0\sim\pi_\theta(\cdot|s_0)}[r_1] = \mathbb{E}_{q\sim p(q),o\sim\pi_\theta(\cdot||q)}[r_1]$$

- 奖励模型：
    - Rule-based (or verifiable, RLVR)
    - Model-based：结果奖励模型 (outcome RM)、过程奖励模型 (process RM)

➢ 结果奖励模型 $r_\psi(q,o)$，BT 公式+交叉熵损失

$$\mathcal{L}_R(r_\phi,\mathcal{D}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma(r_\phi(x,y_w) - r_\phi(x,y_l))\right] \qquad \mathcal{D} = \left\{x^{(i)},y_w^{(i)},y_l^{(i)}\right\}_{i=1}^{N}$$

➢ $r_1 = r(q,o) = r_\psi(q,o) - \beta\log\frac{\pi_\theta(\cdot|q)}{\pi_{\text{ref}}(\cdot|q)}$

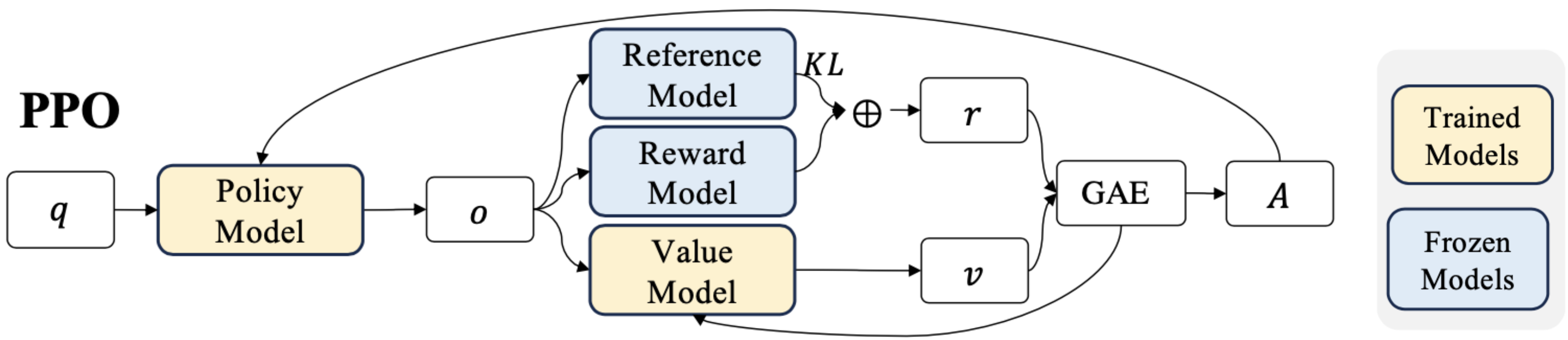
Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback.

# 大语言模型中的强化学习：从 RLHF 说起

- 求解算法：PPO

$$J_{\text{PPO}}(\theta) = \mathbb{E}_{q\sim p(q),o\sim\pi_{\theta_0}(\cdot|q)}\left[\min\left(\frac{\pi_\theta(o|q)}{\pi_{\theta_0}(o|q)}A^{\psi,\phi}_{\pi_{\theta_0}}(q,o), \text{clip}\left(\frac{\pi_\theta(o|q)}{\pi_{\theta_0}(o|q)}, 1-\epsilon, 1+\epsilon\right)A^{\psi,\phi}_{\pi_{\theta_0}}(q,o)\right)\right]$$

$$\hat{A}_t^{\text{GAE}(\gamma,\lambda)} := \sum_{l=0}^{\infty}(\gamma\lambda)^l\delta^V_{t+l} \qquad \delta^V_t = r_t + \gamma V(s_{t+1}) - V(s_t)$$




奖励模型 $r_\psi(q,o)$ 6B

价值模型 $V_\pi(s_t)$ 6B

策略模型 $\pi_\theta(o|q)$ 175B

# Group Relative Policy Optimization (GRPO)

- 价值模型占用内存、耗费计算量 **>>** 不使用价值模型估计优势
- 对给定 $q$，从 $\pi_{\theta_0}$ 中采样一组回复 $\{o^1, o^2, \cdots, o^G\}$，打分 $\boldsymbol{r} = \{r^1, r^2, \cdots, r^G\}$

$$\tilde{r}^i = \frac{r^i - \mathrm{mean}(\boldsymbol{r})}{\mathrm{std}(\boldsymbol{r})}, \qquad A^{\psi}_{\pi_{\theta_0}}(q, o^i) = \tilde{r}^i$$

- $J_{\mathrm{GRPO}}(\theta) = \mathbb{E}_{q \sim p(q), \{o^1, o^2, \cdots, o^G\} \sim \pi_{\theta_0}(\cdot|q)}$

$$\left[\frac{1}{G}\sum_{i=1}^{G} \min\left(\frac{\pi_\theta(o^i|q)}{\pi_{\theta_0}(o^i|q)} A^{\psi}_{\pi_{\theta_0}}(q, o^i), \mathrm{clip}\left(\frac{\pi_\theta(o^i|q)}{\pi_{\theta_0}(o^i|q)}, 1-\epsilon, 1+\epsilon\right) A^{\psi}_{\pi_{\theta_0}}(q, o^i)\right) - \beta D_{\mathrm{KL}}(\pi_\theta(\cdot|q) \| \pi_{\mathrm{ref}}(\cdot|q))\right]$$

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., ... & Guo, D. (2024).
Deepseekmath: Pushing the limits of mathematical reasoning in open language models.

# 拓展到多步策略

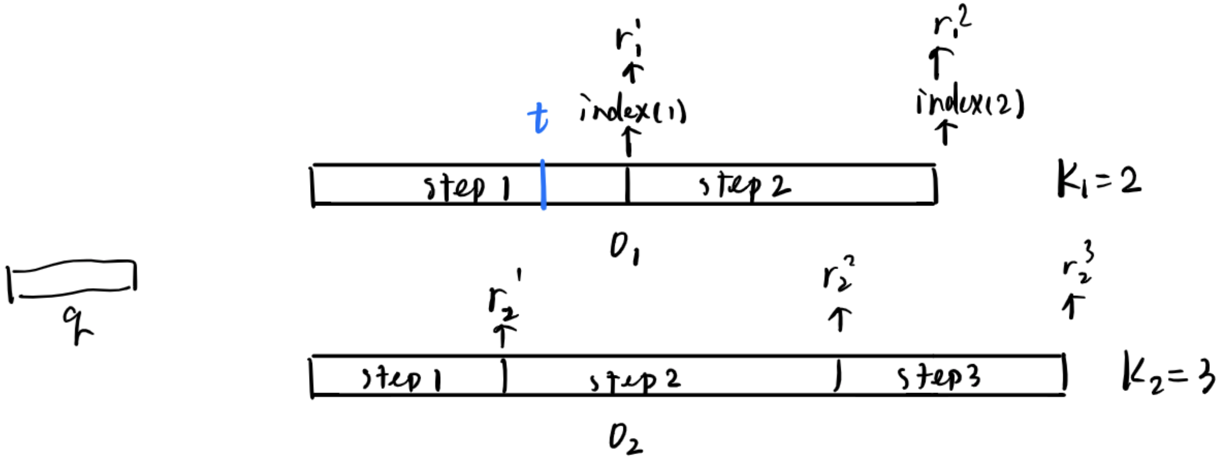
- PPO

$$\mathcal{J}_{PPO}(\theta) = \mathbb{E}[q \sim P(Q), o \sim \pi_{\theta_{old}}(O|q)] \frac{1}{|o|}\sum_{t=1}^{|o|} \min\left[\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{\theta_{old}}(o_t|q,o_{<t})}A_t, \text{clip}\left(\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{\theta_{old}}(o_t|q,o_{<t})}, 1-\varepsilon, 1+\varepsilon\right)A_t\right].$$

- GRPO

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{old}}(O|q)]$$
$$\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left\{\min\left[\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})}\hat{A}_{i,t}, \text{clip}\left(\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})}, 1-\varepsilon, 1+\varepsilon\right)\hat{A}_{i,t}\right] - \beta\mathbb{D}_{KL}\left[\pi_\theta||\pi_{ref}\right]\right\}$$

- 过程奖励模型

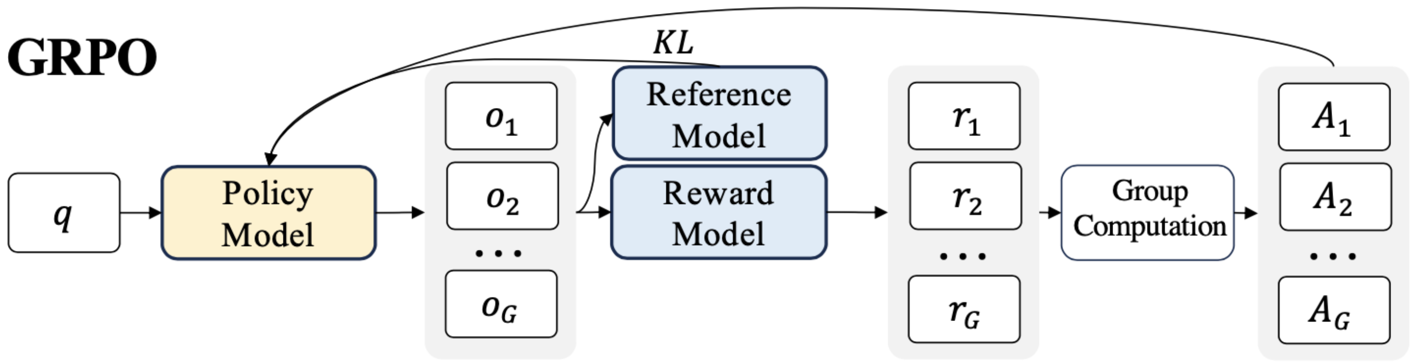



$$\tilde{r}_1^{index(1)} = \frac{r_1^1 - mean(r_1^1, r_1^2, r_2^1, r_2^2, r_2^3)}{std(r \dots)}$$

$$\Rightarrow \quad \hat{A}^P(q, o_{1,\le t}) = \tilde{r}_1^{index(1)} + \tilde{r}_1^{index(2)}$$

# 实际算法

## GRPO




---

**Algorithm 1** Iterative Group Relative Policy Optimization

---

**Input** initial policy model $\pi_{\theta_{\text{init}}}$; reward models $r_\varphi$; task prompts $\mathcal{D}$; hyperparameters $\varepsilon, \beta, \mu$

policy model $\pi_\theta \leftarrow \pi_{\theta_{\text{init}}}$
**for** iteration = 1, …, I **do**
  reference model $\pi_{ref} \leftarrow \pi_\theta$
  **for** step = 1, …, M **do**
    Sample a batch $\mathcal{D}_b$ from $\mathcal{D}$
    Update the old policy model $\pi_{\theta_{old}} \leftarrow \pi_\theta$
    Sample $G$ outputs $\{o_i\}_{i=1}^G \sim \pi_{\theta_{old}}(\cdot \mid q)$ for each question $q \in \mathcal{D}_b$
    Compute rewards $\{r_i\}_{i=1}^G$ for each sampled output $o_i$ by running $r_\varphi$
    Compute $\hat{A}_{i,t}$ for the $t$-th token of $o_i$ through group relative advantage estimation.
    **for** GRPO iteration = 1, …, $\mu$ **do**
      Update the policy model $\pi_\theta$ by maximizing the GRPO objective (Equation 21)
  Update $r_\varphi$ through continuous training using a replay mechanism.

**Output** $\pi_\theta$

# 总结

- 强化学习系统性很强，一环扣一环
- 目前大语言模型中的强化学习：

  本质吗？是最好的吗？——算法还很粗糙、应用还很初步

- 学习资料：
  - 教材：
    - **Richard S. Sutton and Andrew G. Barto, 2020, Reinforcement Learning: An Introduction**
    - 王树森 张志华, 2022, 深度强化学习（初稿）
  - 课程：
    - 西湖大学赵世钰【【强化学习的数学原理】课程：从零开始到透彻理解（完结）】
    - **OpenAI, Welcome to Spinning Up in Deep RL!**

# 总结

- 其它资料：
  - 博客： Seita's Place: Going Deeper Into Reinforcement Learning: Fundamentals of Policy Gradients
  - 博客：Seita's Place: Notes on the Generalized Advantage Estimation Paper
  - 博客：Weng Lilian, A (Long) Peek into Reinforcement Learning, 2018
  - 博客：Weng Lilian, Policy Gradient Algorithms, 2018
  - 课程：Sergey Levine, CS 285 UC Berkeley, Deep Reinforcement Learning, 2023
  - 教材：Kevin Murphy, Reinforcement Learning: An Overview
  - 课程：David Silver, UCL, Reinforcement Learning, 2015

**Thank you for listening!**