Reinforcement Learning

Study Notes: From the basics to PPO and GRPO

- What is the reinforcement learning?
- What are PPO and GRPO?
- How is RL utilized in LLMs?

Xin Zhang

Reinforcement Learning

2025 Turing Award

deepseek

DeepSeek Homepage | Chat DeepSeek R1 | Hugging Face DeepSeek AI

Discord DeepSeek AI | WeChat DeepSeek AI | Twitter deepseek ai

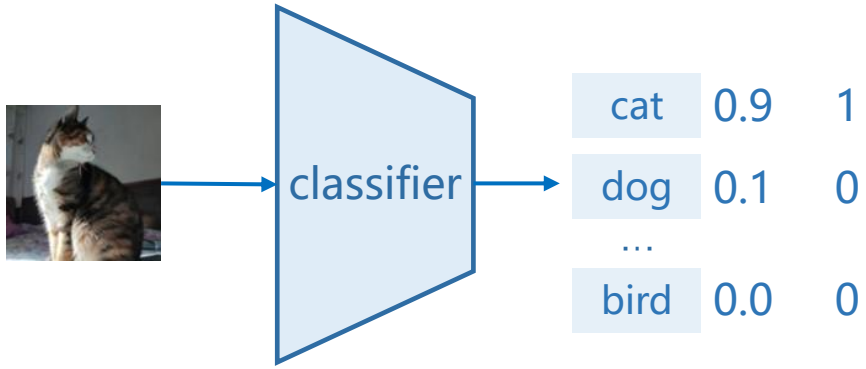
License MIT

Paper Link

2025 deepseek-R1

2016 AlphaGo

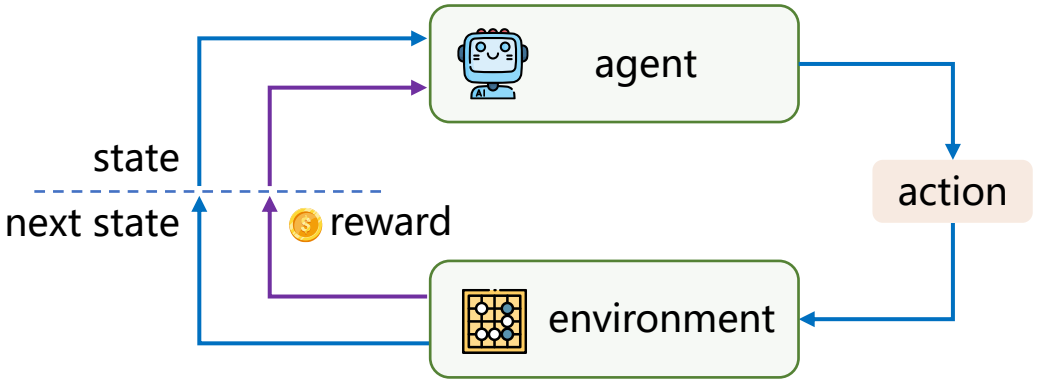
What is Reinforcement Learning



Supervised Learning

labeled data

Find a model to fit the mapping.

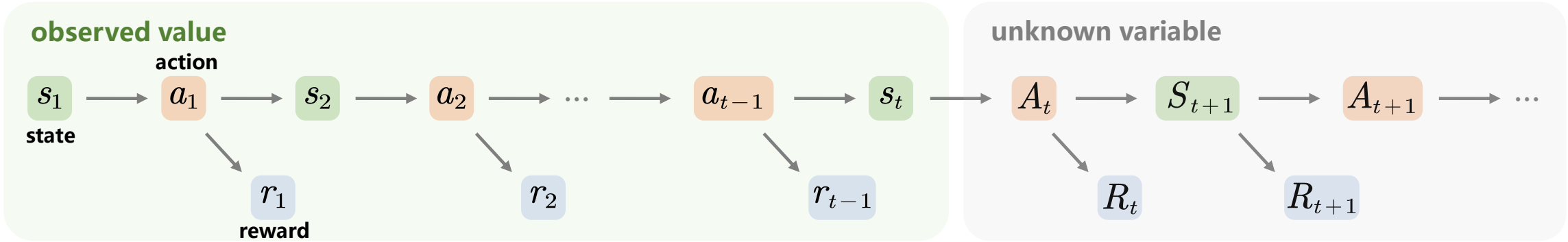


Reinforce Learning

Interaction

Find a policy to maximize the reward.

Markov Decision Process



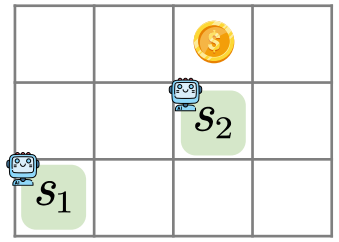
Return $U_1 = R_1 + \gamma R_2 + \cdots + \gamma^{t-2} R_{t-1} + \gamma^{t-1} R_t + \gamma^t R_{t+1} + \cdots + \gamma^{n-1} R_n$

Return (next state) $U_2 = R_2 + \cdots + \gamma^{t-3} R_{t-1} + \gamma^{t-2} R_t + \gamma^{t-1} R_{t+1} + \cdots + \gamma^{n-2} R_n$

Discounted Return

Immediate rewards | Future rewards

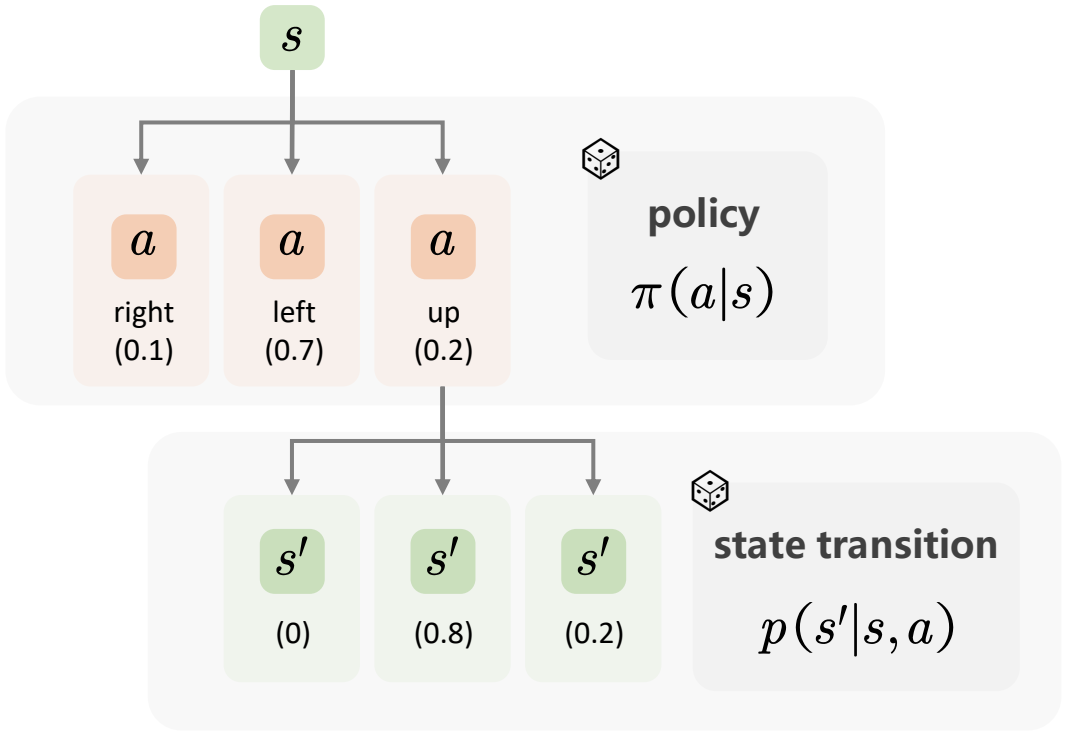
$$U_t = \sum_{k=t}^{n} \gamma^{k-t} R_k \qquad U_t = R_t + \gamma U_{t+1}$$



$$V(s_2) > V(s_1)$$

$$Q(s_2, up) > Q(s_2, down)$$

Randomness and Expectation



Expectation

$$\mathbb{E}_{x\sim \mathrm{p}(x)}[h(x)] = \sum_{x\in\mathcal{X}} p\,(x)\cdot h(x)$$

state value Bellman Equation

$$\begin{aligned}
V_\pi(s) &= \mathbb{E}_{A_t,S_{t+1}\ldots}[U_t|S_t=s] \\
&= \mathbb{E}_{A_t,S_{t+1}\ldots}[R_t + \gamma U_{t+1}|S_t=s] \\
&= \mathbb{E}_{A_t,S_{t+1}\ldots}[R_t + \gamma V_\pi(S_{t+1})|S_t=s] \\
&= \sum_{a\in A}\pi\,(a\mid s)\left(r(s,a) + \gamma\sum_{s'\in S} p\,(s'\mid s,a)V^\pi(s')\right)
\end{aligned}$$

(R_t: A_t; U_{t+1}: A_t, S_{t+1})

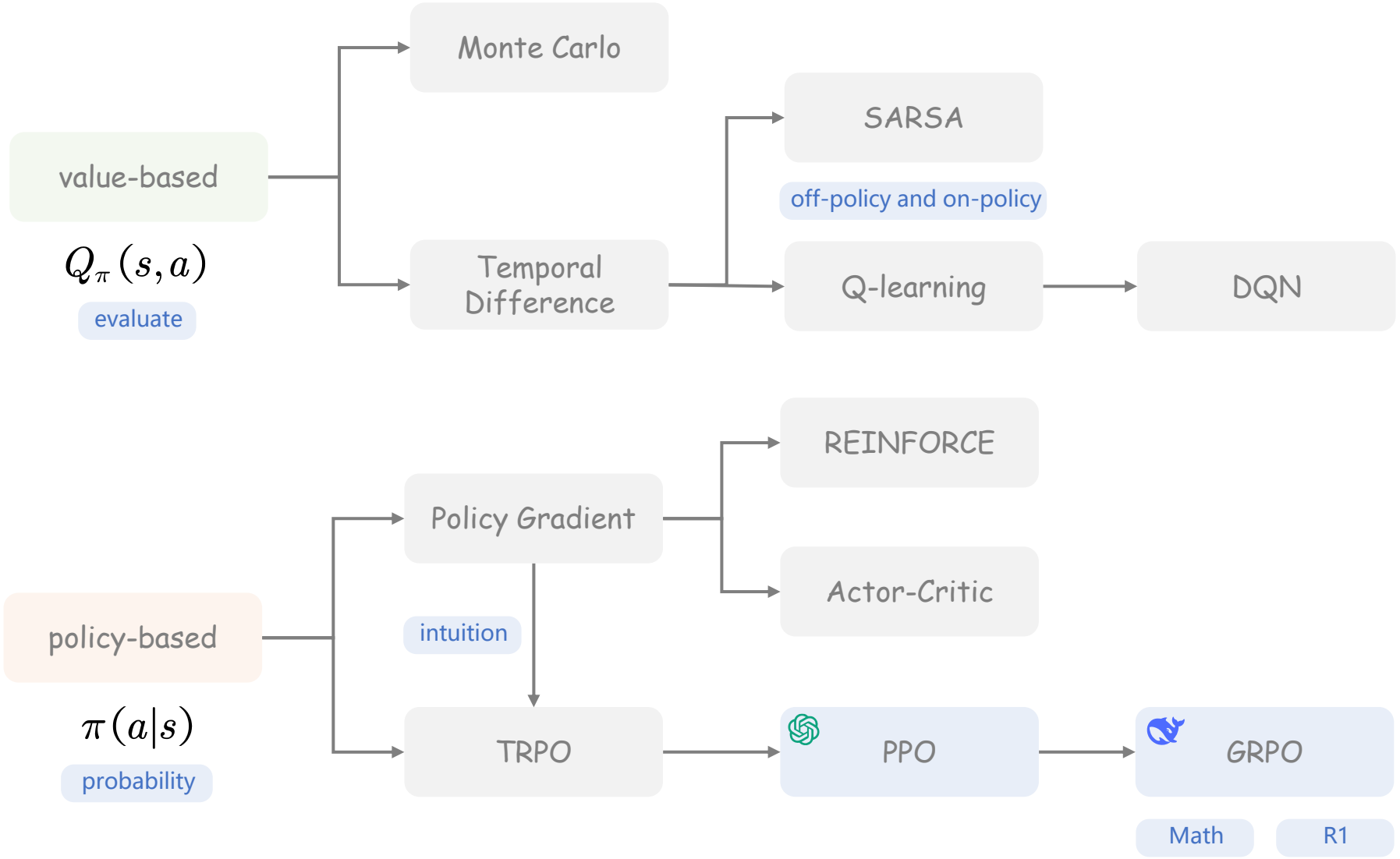
action value Bellman Equation

$$\begin{aligned}
Q_\pi(s,a) &= \mathbb{E}_{S_{t+1}\ldots}[U_t|S_t=s,A_t=a] \\
&= \mathbb{E}_{S_{t+1}\ldots}[R_t + \gamma U_{t+1})|S_t=s,A_t=a] \\
&= \mathbb{E}_{S_{t+1}\ldots}[R_t + \gamma Q_\pi(S_{t+1},A_{t+1})\mid S_t=s,A_t=a] \\
&= r(s,a) + \gamma\sum_{s'\in S} p\,(s'\mid s,a)\sum_{a'\in A}\pi\,(a'\mid s')Q_\pi(s',a')
\end{aligned}$$

(U_{t+1}: S_{t+1}, A_{t+1})

relationship $V_\pi(s) = \sum_{a\in A}\pi\,(a|s)Q_\pi(s,a)$

The Top-Down Framework

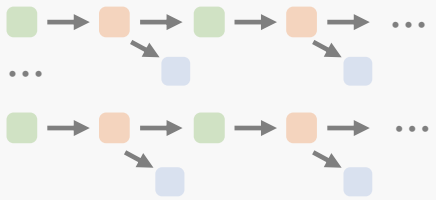


Temporal Difference

How to estimate $Q_\pi(s,a)$

$$Q_\pi(s,a) = \mathbb{E}_{S_{t+1}\cdots}[R_t + \gamma Q_\pi(S_{t+1}, A_{t+1}) \mid S_t = s, A_t = a]$$

1 Monte Carlo multi-step variance ↑ bias ↓

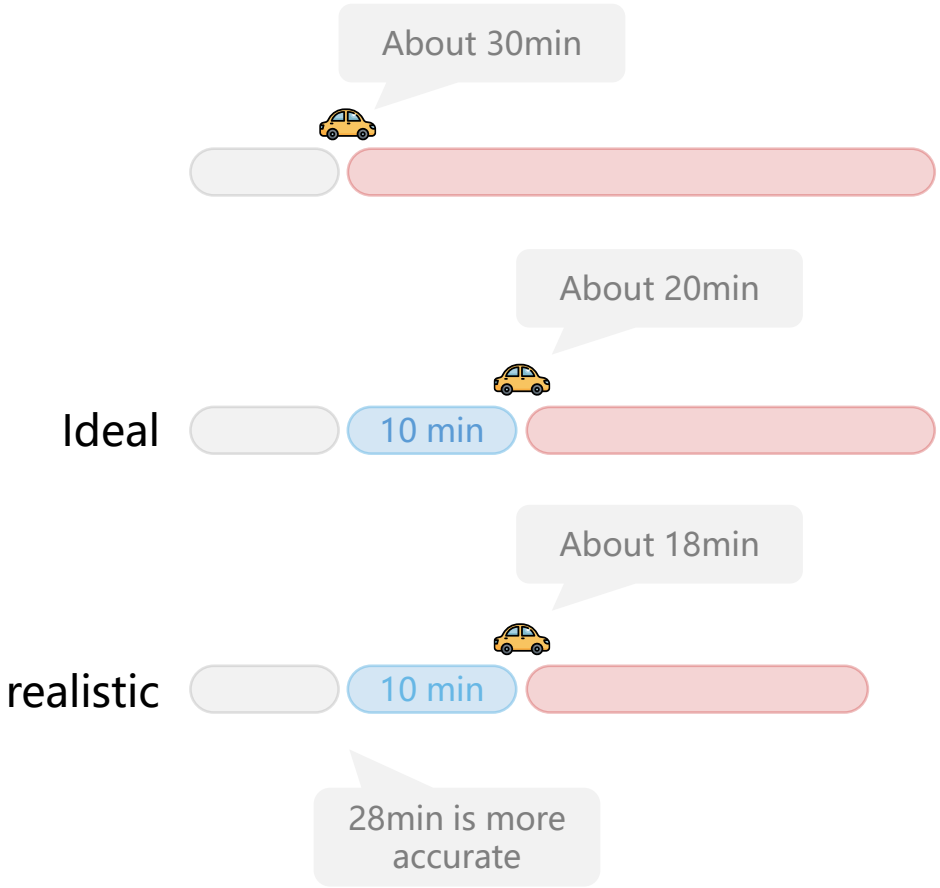


$$Q_\pi(s,a) \approx \frac{1}{N}\sum_{i=1}^{N} U^{(i)}$$

2 Temporal Difference one-step variance ↓ bias ↑

$$Q_\pi(s_t,a_t) \approx \underbrace{r_t + \gamma Q_\pi(s_{t+1}, a_{t+1})}_{\textbf{TD target}}$$

$$Q_\pi(s_t,a_t) \leftarrow Q_\pi(s_t,a_t) + \alpha[r_t + \gamma Q_\pi(s_{t+1},a_{t+1}) - Q_\pi(s_t,a_t)]$$



SARSA and Q-learning

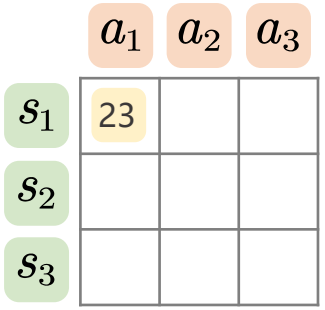
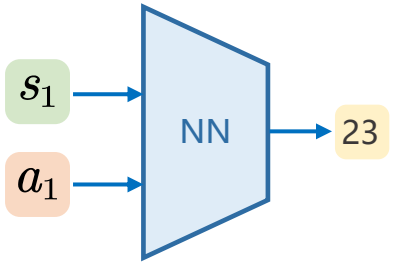


Table Query

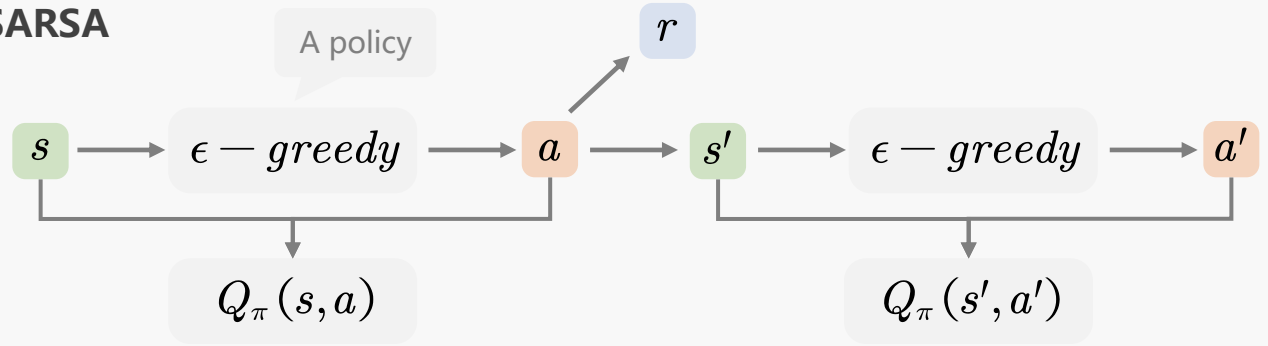
(SARSA, Q-Learning)



Neural Network

(DQN)

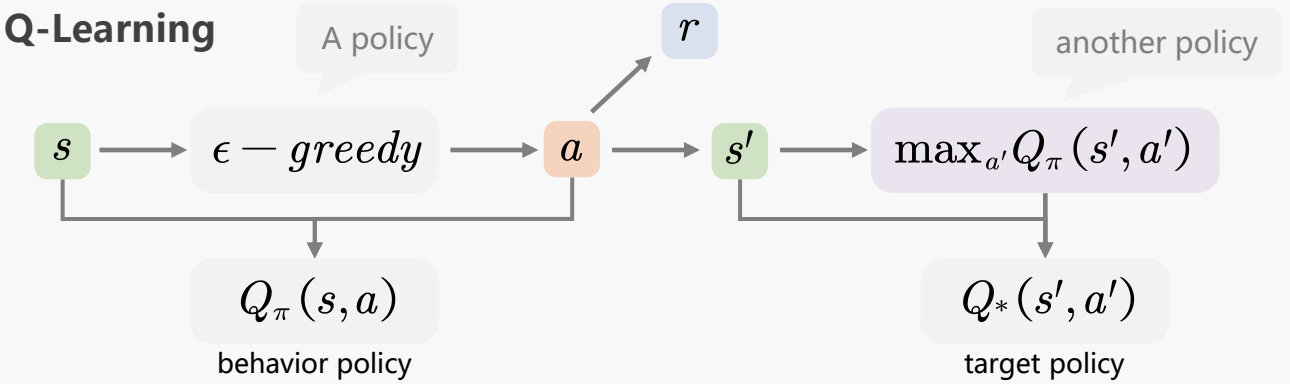
SARSA



obtain $Q_\pi(s,a)$

on-policy

Q-Learning



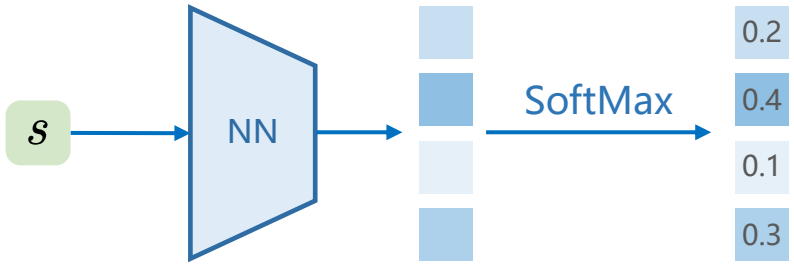
obtain $Q_*(s,a)$

off-policy

Table Update: $Q_\pi(s_t,a_t) \leftarrow Q_\pi(s_t,a_t) + \alpha[r_t + \gamma Q_\pi(s_{t+1},a_{t+1}) - Q_\pi(s_t,a_t)]$

NN Update: $Loss = \frac{1}{2}(r + \gamma Q_\pi(s',a') - Q_\pi(s,a))^2$

Policy Gradient



$$\pi(\cdot|s;\theta)$$

Maximize the return

$$J(\boldsymbol{\theta}) = \mathbb{E}_S[V_\pi(S)] = \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)}\left[\sum_t \gamma^t r(s_t, a_t)\right]$$

Optimization $\max_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$

Gradient Ascent $\boldsymbol{\theta}_{new} \leftarrow \boldsymbol{\theta}_{old} + \beta \cdot \nabla_{\boldsymbol{\theta}} J(\theta_{old})$

Differentiate a composite function

$$\nabla \ln f(x) = \frac{\nabla f(x)}{f(x)}$$

$$\nabla_{\boldsymbol{\theta}} V_\pi(s) = \nabla_{\boldsymbol{\theta}} \sum_{a \in \mathcal{A}} \pi(a|s;\boldsymbol{\theta}) \cdot Q_\pi(s,a)$$

$$\approx \sum_{a \in \mathcal{A}} (\nabla_{\boldsymbol{\theta}} \pi(a|s;\boldsymbol{\theta}) \cdot Q_\pi(s,a))$$ Ignore one item

$$= \sum_{a \in \mathcal{A}} (\pi(a|s;\boldsymbol{\theta}) \cdot \nabla_{\boldsymbol{\theta}} \ln \pi(a|s;\boldsymbol{\theta}) \cdot Q_\pi(s,a))$$

$$= \mathbb{E}_{A \sim \pi(\cdot|s;\boldsymbol{\theta})}[\nabla_{\boldsymbol{\theta}} \ln \pi(A|s;\boldsymbol{\theta}) \cdot Q_\pi(s,A)]$$

random gradient

REINFORCE and Actor-Critic

Optimization $\max_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$

Gradient Ascent $\boldsymbol{\theta}_{new} \leftarrow \boldsymbol{\theta}_{old} + \beta \cdot \nabla_{\boldsymbol{\theta}} J(\theta_{old})$

$$\boldsymbol{\theta}_{new} \leftarrow \boldsymbol{\theta}_{old} + \beta \cdot Q_\pi(s,a) \cdot \nabla_{\boldsymbol{\theta}} \ln \pi(a|s;\boldsymbol{\theta}_{old})$$

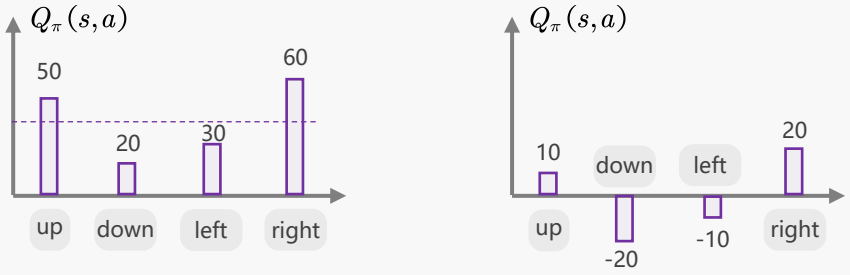
REINFORCE

$$u_t = \sum_{k=t}^{n} \gamma^{k-t} \cdot r_k$$

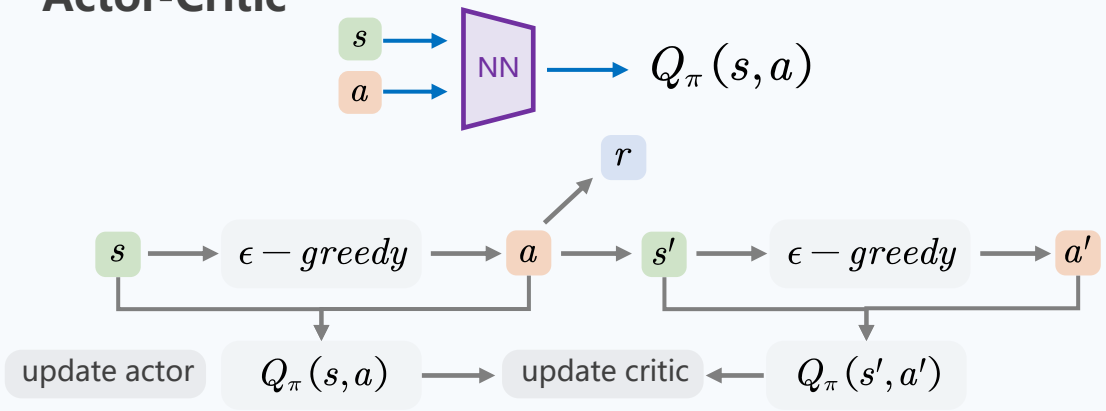
$$\boldsymbol{\theta}_{new} \leftarrow \boldsymbol{\theta}_{old} + \beta \cdot \sum_{t=1}^{n} \gamma^{t-1} \cdot u_t \cdot \nabla_{\boldsymbol{\theta}} \ln \pi(a_t|s_t;\boldsymbol{\theta}_{old})$$

Baseline

$$\mathbb{E}_S\left[_{A\sim\pi(\cdot|S;\boldsymbol{\theta})}\left[(Q_\pi(S,A) - b) \cdot \nabla_{\boldsymbol{\theta}} \ln \pi(A|S;\boldsymbol{\theta})\right]\right]$$



Actor-Critic



The Problem of Policy Gradient

Difficulty in setting the step size for updates.

Trust region

$$\|\boldsymbol{\theta}_{new} - \boldsymbol{\theta}_{old}\| \leqslant \Delta$$

A new policy is better

$$J(\boldsymbol{\theta}_{new}) - J(\boldsymbol{\theta}_{old}) > 0$$

Trust Region Policy Optimization

$J(\boldsymbol{\theta'}) - J(\boldsymbol{\theta})$ new policy old policy

$= J(\boldsymbol{\theta'}) - \mathbb{E}_{s_0 \sim p(s_0)}[V_{\pi_{\boldsymbol{\theta}}}(s_0)]$

$= J(\boldsymbol{\theta'}) - \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}'}(\tau)}[V_{\pi_{\boldsymbol{\theta}}}(s_0)]$ initial state unchanged

$= J(\boldsymbol{\theta'}) - \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}'}(\tau)}\left[\sum_{t=0}^{\infty} \gamma^t V_{\pi_{\boldsymbol{\theta}}}(s_t) - \sum_{t=0}^{\infty} \gamma^{t+1} V_{\pi_{\boldsymbol{\theta}}}(s_{t+1})\right]$

$= \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}'}(\tau)}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right] + \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}'}(\tau)}\left[\sum_{t=0}^{\infty} \gamma^t \left(\gamma V_{\pi_{\boldsymbol{\theta}}}(s_{t+1}) - V_{\pi_{\boldsymbol{\theta}}}(s_t)\right)\right]$

$= \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}'}(\tau)}\left[\sum_{t=0}^{\infty} \gamma^t \left[r(s_t, a_t) + \gamma V_{\pi_{\boldsymbol{\theta}}}(s_{t+1}) - V_{\pi_{\boldsymbol{\theta}}}(s_t)\right]\right]$

$= \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}'}(\tau)}\left[\sum_{t=0}^{\infty} \gamma^t \left[Q_{\pi_{\boldsymbol{\theta}}}(s_t, a_t) - V_{\pi_{\boldsymbol{\theta}}}(s_t)\right]\right]$

$= \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}'}(\tau)}\left[\sum_{t=0}^{\infty} \gamma^t A_{\pi_{\boldsymbol{\theta}}}(s_t, a_t)\right]$

Advantage function

Importance Sampling

$\mathbb{E}_{x \sim p(x)}[f(x)]$

$= \int_x q(x) \frac{p(x)}{q(x)} f(x) dx = \mathbb{E}_{x \sim q(x)}\left[\frac{p(x)}{q(x)} f(x)\right]$

$\mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}'}(\tau)}\left[\sum_{t=0}^{\infty} \gamma^t \cdot A_{\pi_{\boldsymbol{\theta}}}(s_t, a_t)\right]$

$= \mathbb{E}_{s_t \sim p_{\boldsymbol{\theta}'}(s_t)}\left[\mathbb{E}_{a_t \sim \pi_{\boldsymbol{\theta}'}(a_t|s_t)}\left[\gamma^t \cdot A_{\pi_{\boldsymbol{\theta}}}(s_t, a_t)\right]\right]$

Approximation 1: Ignoring the changes in the state distribution.

$= \mathbb{E}_{s_t \sim p_{\boldsymbol{\theta}}(s_t)}\left[\mathbb{E}_{a_t \sim \pi_{\boldsymbol{\theta}'}(a_t|s_t)}\left[\gamma^t \cdot A_{\pi_{\boldsymbol{\theta}}}(s_t, a_t)\right]\right]$

Approximation 2: Importance sampling.

$= \mathbb{E}_{s_t \sim p_{\theta}(s_t)}\left[\mathbb{E}_{a_t \sim \pi_{\theta}(a_t|s_t)}\left[\frac{\pi_{\theta'}(a_t|s_t)}{\pi_{\theta}(a_t|s_t)} \cdot \gamma^t \cdot A_{\pi_{\theta}}(s_t, a_t)\right]\right]$

TRPO

$\underset{\boldsymbol{\theta'}}{\operatorname{argmax}} \mathbb{E}_{s \sim v_{\boldsymbol{\theta}}, a \sim \pi_{\boldsymbol{\theta}}(\cdot|s)}\left[\frac{\pi_{\boldsymbol{\theta'}}(a, s)}{\pi_{\boldsymbol{\theta}}(a, s)} \cdot A_{\pi_{\boldsymbol{\theta}}}(s, a)\right]$

$s.t.\ D_{KL}(\pi_{\boldsymbol{\theta}}(\cdot|s) \ || \ \pi_{\boldsymbol{\theta'}}(\cdot|s))$

Proximal Policy Optimization

TRPO

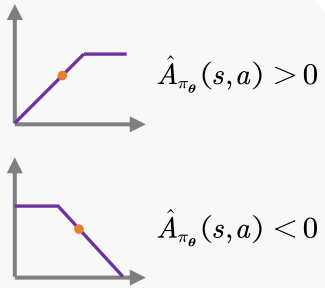
$$\underset{\theta'}{\operatorname{argmax}} \mathbb{E}_{s\sim v_\theta, a\sim\pi_\theta(\cdot|s)}\left[\frac{\pi_{\theta'}(a,s)}{\pi_\theta(a,s)}\cdot A_{\pi_\theta}(s,a)\right] \qquad s.t.\ D_{KL}(\pi_{\boldsymbol{\theta}}(\cdot|s)\ ||\ \pi_{\boldsymbol{\theta}'}(\cdot|s))$$

PPO-penalty

$$\underset{\boldsymbol{\theta}'}{\operatorname{argmax}} \mathbb{E}_{s\sim v_{\boldsymbol{\theta}}} \mathbb{E}_{a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)}\left[\frac{\pi_{\boldsymbol{\theta}'}(a|s)}{\pi_{\boldsymbol{\theta}}(a|s)}\hat{A}_{\pi_{\boldsymbol{\theta}}}(s,a) - \beta D_{KL}[\pi_{\boldsymbol{\theta}}(\cdot|s)||\pi_{\boldsymbol{\theta}'}(\cdot|s)]\right] \qquad \begin{cases}\beta\leftarrow\beta/2 & if\ D_{KL}<\delta/1.5\\ \beta\leftarrow\beta\times 2 & if\ D_{KD}>\delta\times 1.5\end{cases}$$

PPO-clip

$$\underset{\boldsymbol{\theta}'}{\operatorname{argmax}} \mathbb{E}_{s\sim v_{\boldsymbol{\theta}}} \mathbb{E}_{a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)}\left[\min\left(\frac{\pi_{\boldsymbol{\theta}'}(a|s)}{\pi_{\boldsymbol{\theta}}(a|s)}\hat{A}_{\pi_{\boldsymbol{\theta}}}(s,a),\ clip\left(\frac{\pi_{\boldsymbol{\theta}'}(a|s)}{\pi_{\boldsymbol{\theta}}(a|s)}, 1-\epsilon, 1+\epsilon\right)\hat{A}_{\pi_{\boldsymbol{\theta}}}(s,a)\right)\right]$$



Generalized Advantage Estimation

1 Monte Carlo multi-step variance ↑ bias ↓

2 Temporal Difference one-step variance ↓ bias ↑

➡ **Multi-Step Temporal Difference**

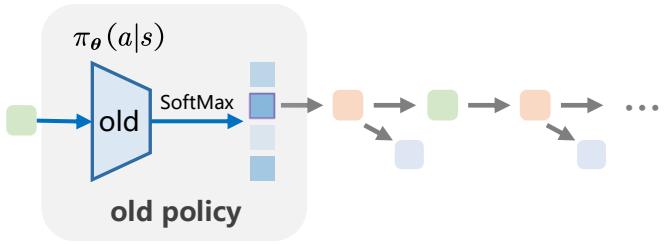
$$A_t^{(1)} = r_t + \gamma V(s_{t+1}) - V(s_t) \Rightarrow \delta_t \qquad A_{t+1}^{(1)} = r_{t+1} + \gamma V(s_{t+2}) - V(s_{t+1}) \Rightarrow \delta_{t+1}$$

$$A_t^{(2)} = r_t + \gamma r_{t+1} + \gamma^2 V(s_{t+2}) - V(s_t) \Rightarrow \delta_t + \gamma\, \delta_{t+1}$$

$$A_t^{(k)} = r_t + \gamma r_{t+1} + \cdots + \gamma^{k-1} r_{t+k-1} + \gamma^k V(s_k) - V(s_t) \Rightarrow \delta_t + \gamma\, \delta_{t+1} + \cdots + \gamma^{k-1}\, \delta_{t+k-1}$$

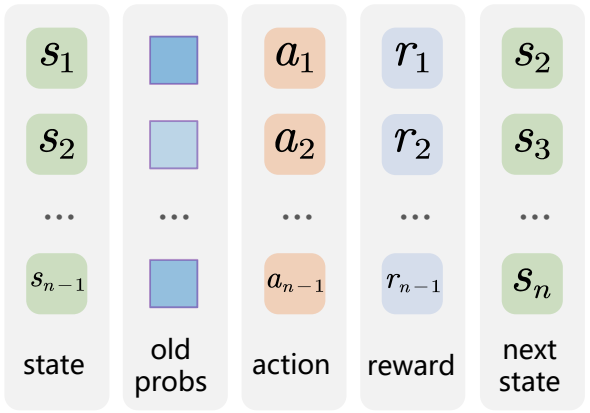
$$\begin{aligned} A_t^{GAE} &= (1-\lambda)\left(A_t^{(1)} + \lambda A_t^{(2)} + \lambda^2 A_t^{(3)} + \cdots\right) \\ &= (1-\lambda)\left(\delta_t(1+\lambda+\lambda^2+\cdots) + \gamma\delta_{t+1}(\lambda+\lambda^2+\cdots) + \cdots\right) \\ &= (1-\lambda)\left(\delta_t \frac{1}{1-\lambda} + \gamma\delta_{t+1}\frac{\lambda}{1-\lambda} + \gamma^2\delta_{t+2}\frac{\lambda^2}{1-\lambda} + \cdots\right) = \sum_{l=0} (\gamma\lambda)^l \delta_{t+l} \end{aligned}$$

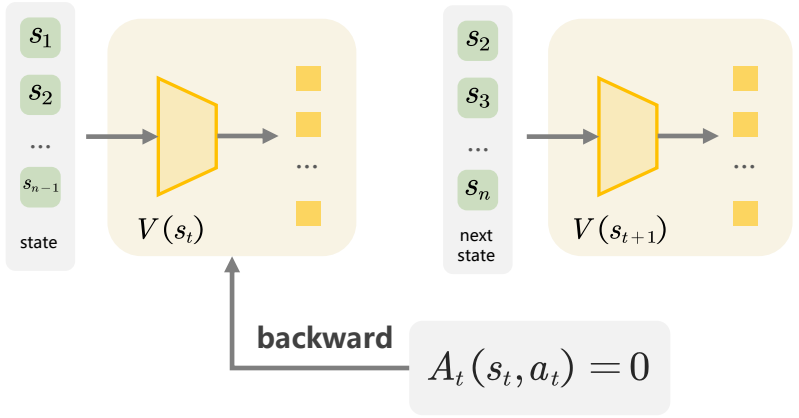
Training Process of PPO



PPO-clip

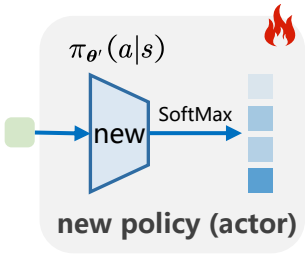
$$\underset{\theta'}{\operatorname{argmax}}\, \mathbb{E}_{s\sim v_\theta}\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[\min\left(\frac{\pi_{\theta'}(a|s)}{\pi_\theta(a|s)}A^{GAE}{}_{\pi_\theta}(s,a),\ clip\left(\frac{\pi_{\theta'}(a|s)}{\pi_\theta(a|s)},1-\epsilon,1+\epsilon\right)A^{GAE}{}_{\pi_\theta}(s,a)\right)\right]$$

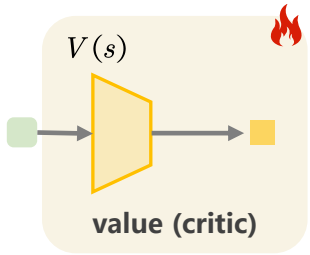


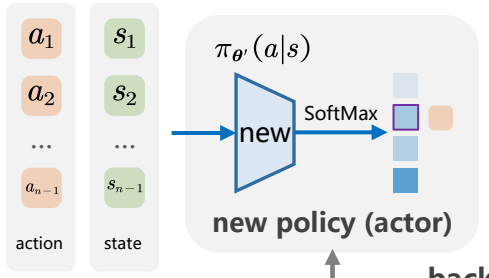


$$A_t(s_t,a_t) = r_1 + r_2 + \ldots + r_{n-1} \text{ (reward)} + \gamma V(s_{t+1}) - V(s_t)$$

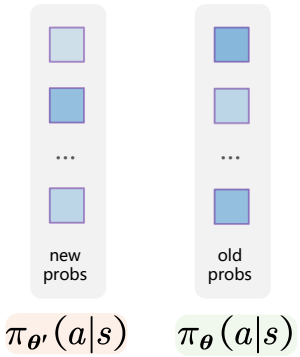
$$A^{GAE}{}_t(s_t,a_t) = \sum_{l=0}(\gamma\lambda)^l\delta_{t+l}$$



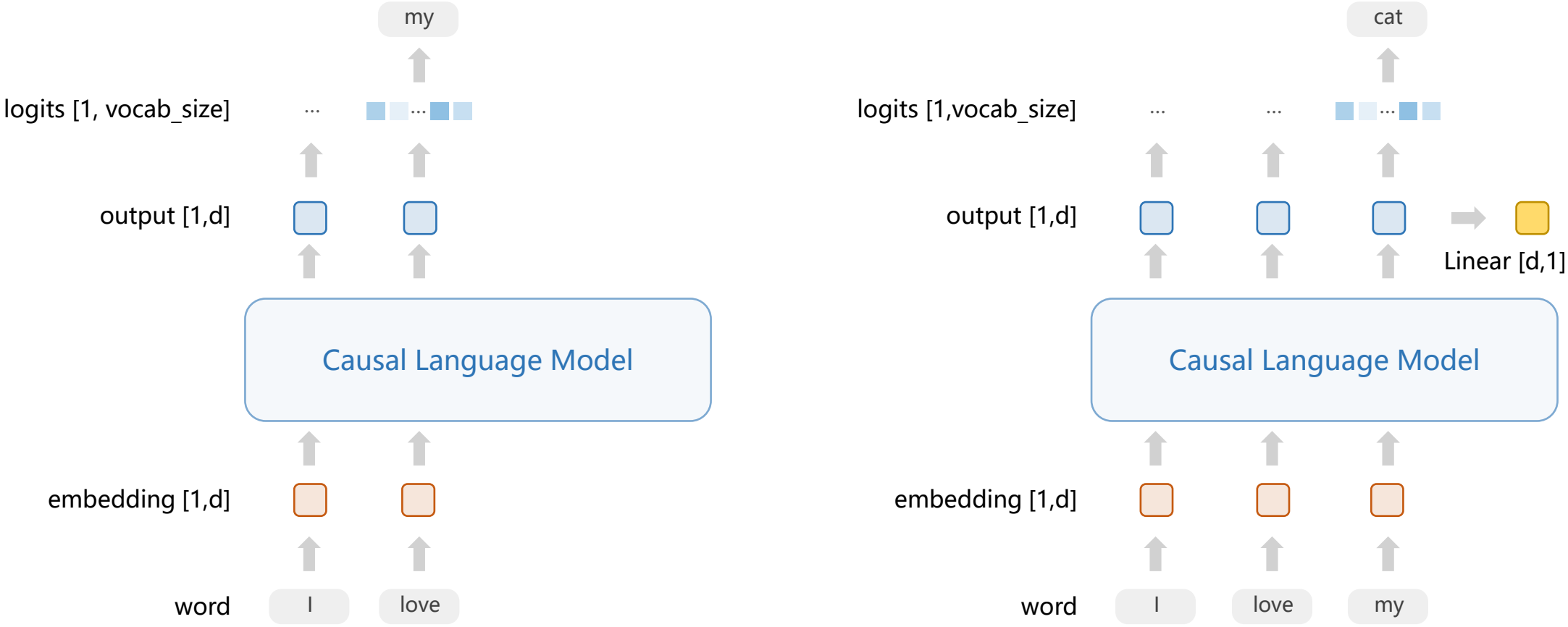




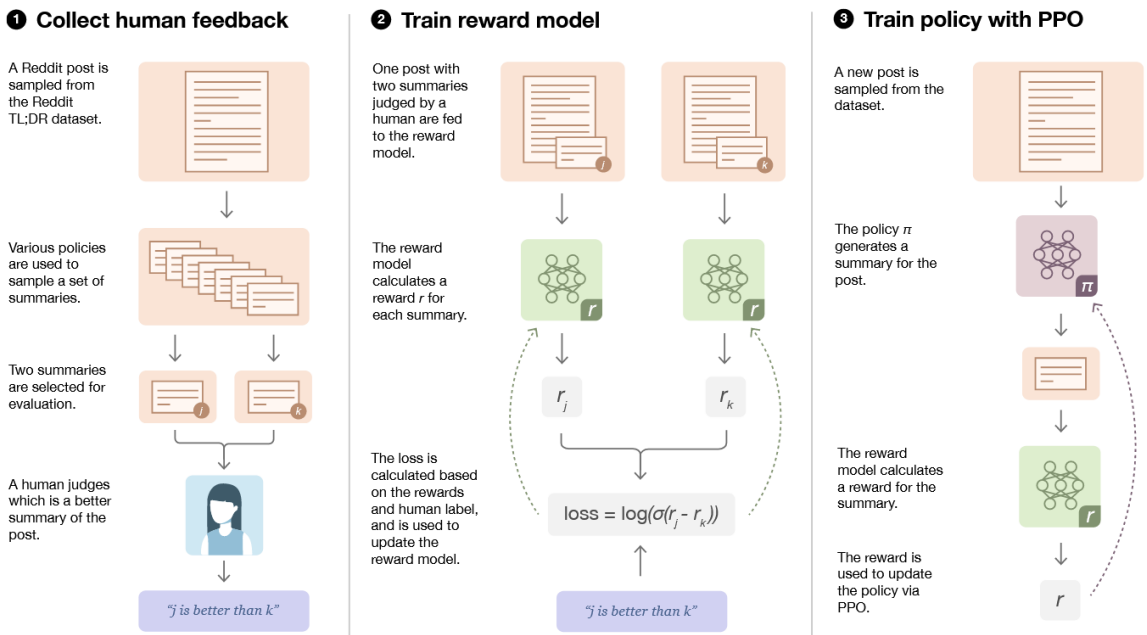
PPO

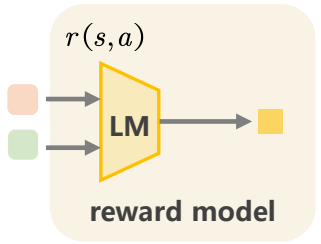


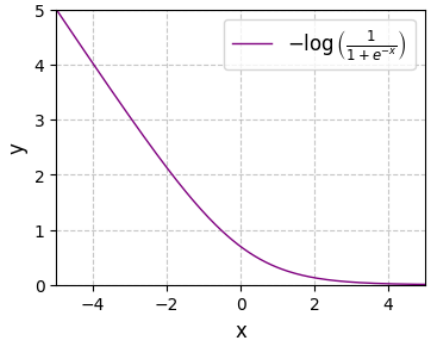
Large Language Model



How To Train A Reward Model







Bradley-Terry

	win	lose
A vs. B	8	4
A vs. C	3	5
B vs. C	2	3

$A:\pi_a$

$B:\pi_b$

$C:\pi_c$

$$P_{A>B}=\frac{\pi_a}{\pi_a+\pi_b}$$

Maximum Likelihood Estimation

$$L=\prod_{i\neq j}\left(\frac{\pi_i}{\pi_i+\pi_j}\right)^{n_{ij}}$$

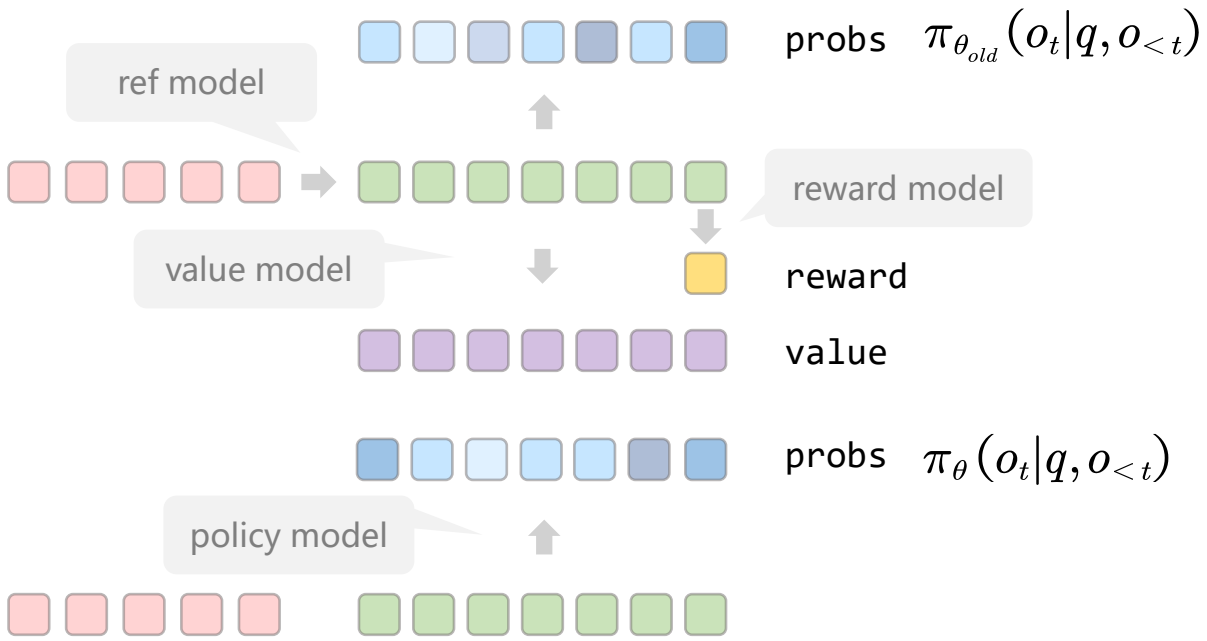
$$L=\prod\left(\frac{\exp(r_{\boldsymbol{\theta}}(x,yes))}{\exp(r_{\boldsymbol{\theta}}(x,yes))+\exp(r_{\boldsymbol{\theta}}(x,no))}\right)$$

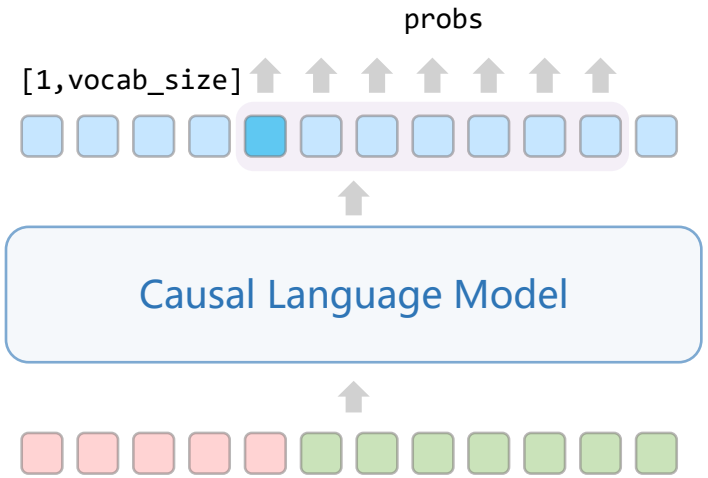
$$\ln L=\ln\sum\frac{1}{1+\exp(r_{\boldsymbol{\theta}}(x,no)-r_{\boldsymbol{\theta}}(x,yes))}\quad\sigma=\frac{1}{1+e^{-x}}$$

$$Loss=\mathbb{E}\left[\ln\sigma(r_{\boldsymbol{\theta}}(x,yes)-r_{\boldsymbol{\theta}}(x,no))\right]$$

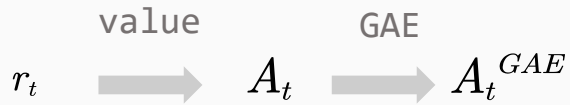
Training Process of PPO in LLM

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



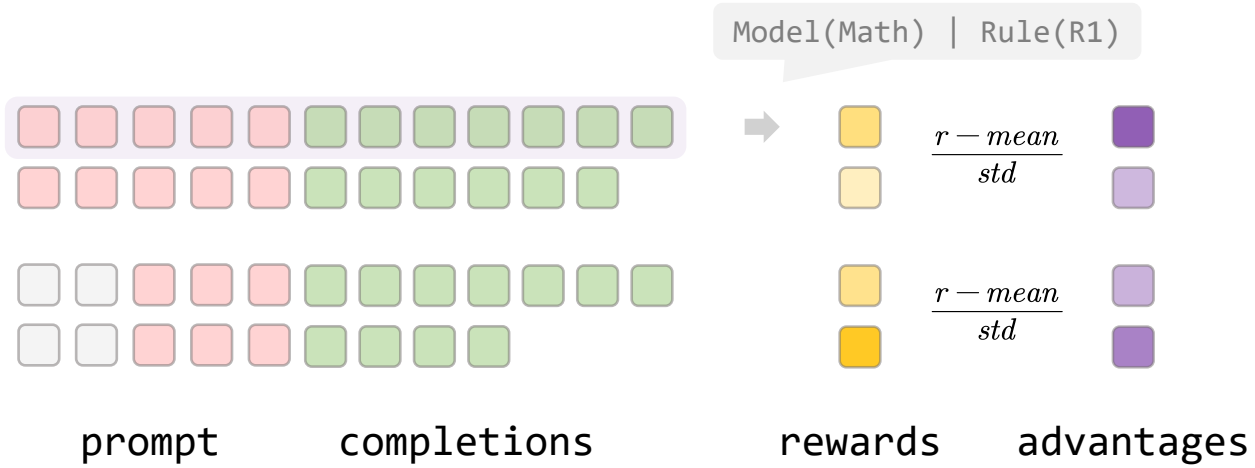


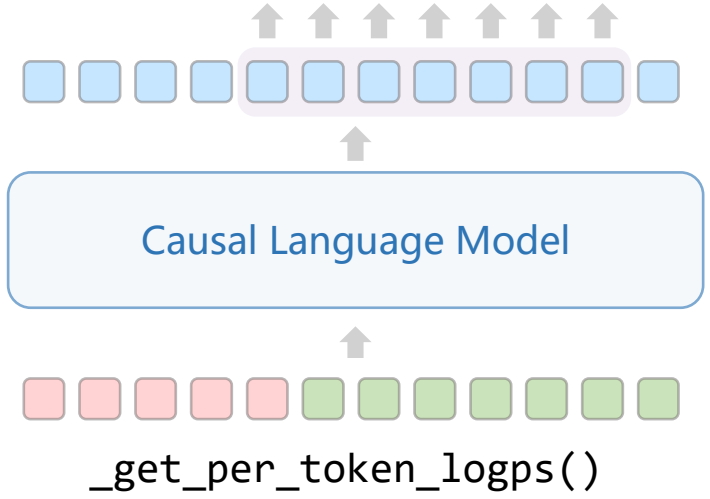
$$r_t = \underbrace{r_\varphi(q, o_{\leqslant} t)}_{\text{reward}} - \underbrace{\beta \ln \frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{\theta_{old}}(o_t|q,o_{<t})}}_{\text{KL}}$$

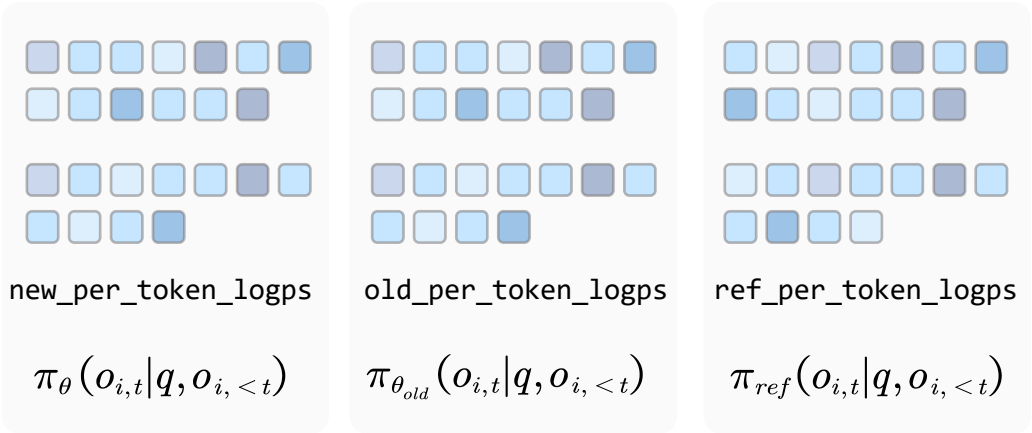


$$A_t = r_t + \gamma V_{t-1} - V_t$$

Group Relative Policy Optimization





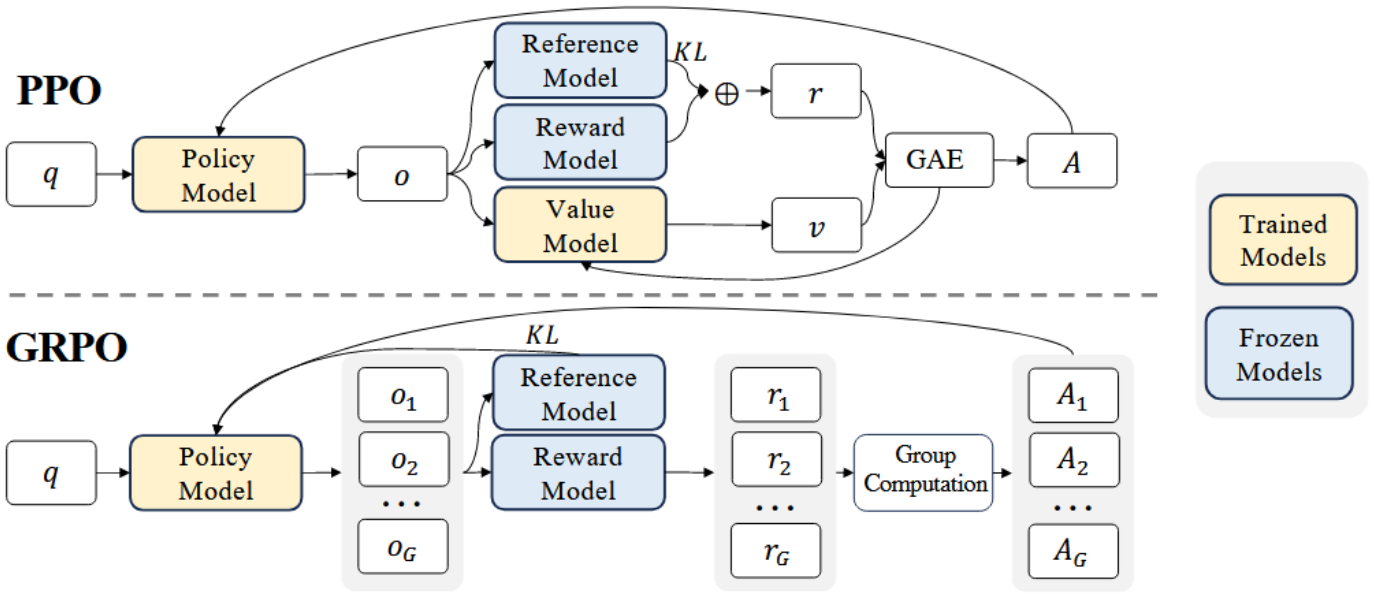


$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}\left[q \sim \mathrm{P}(Q), \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{old}}(O|q)\right]$$

$$\frac{1}{G}\sum_{i=i}^{G}\frac{1}{|o_i|}\sum_{t=1}^{o_i}\left\{\min\left[\frac{\pi_\theta}{\pi_{\theta_{old}}}\hat{A}_{i,t}, clip\left(\frac{\pi_\theta}{\pi_{\theta_{old}}}, 1-\epsilon, 1+\epsilon\right)\hat{A}_{i,t}\right] - \beta\mathbb{D}_{KL}\left[\pi_\theta||\pi_{ref}\right]\right\}$$

$$\mathbb{D}_{KL}\left[\pi_\theta||\pi_{ref}\right] = \frac{\pi_{ref}}{\pi_\theta} - \log\frac{\pi_{ref}}{\pi_\theta} - 1$$

Review the PPO and GRPO



PPO

policy model	ref model
reward model	value model

GRPO

policy model	ref model
reward model	

Reference

1. 王树森,黎彧君,张志华. 深度强化学习. 人民邮电出版社, 2022. https://github.com/wangshusen/DRL
2. 张伟楠,沈键,俞勇.动手学强化学习. 人民邮电出版社, 2022. https://github.com/boyu-ai/Hands-on-RL
3. Sutton R S, Barto A G. Reinforcement learning: An introduction. Cambridge: MIT press, 1998.
4. Shao Z, Wang P, Zhu Q, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv, 2024.
5. Guo D, Yang D, Zhang H, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv, 2025.
6. Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms. arXiv, 2017.
7. Schulman J, Levine S, Abbeel P, et al. Trust region policy optimization. ICML, 2015.
8. Stiennon N, Ouyang L, Wu J, et al. Learning to summarize with human feedback. NeurIPS, 2020.
9. https://github.com/huggingface/trl


踏上取经路，
比抵达灵山更重要。
——《黑神话：悟空》制作人 冯骥