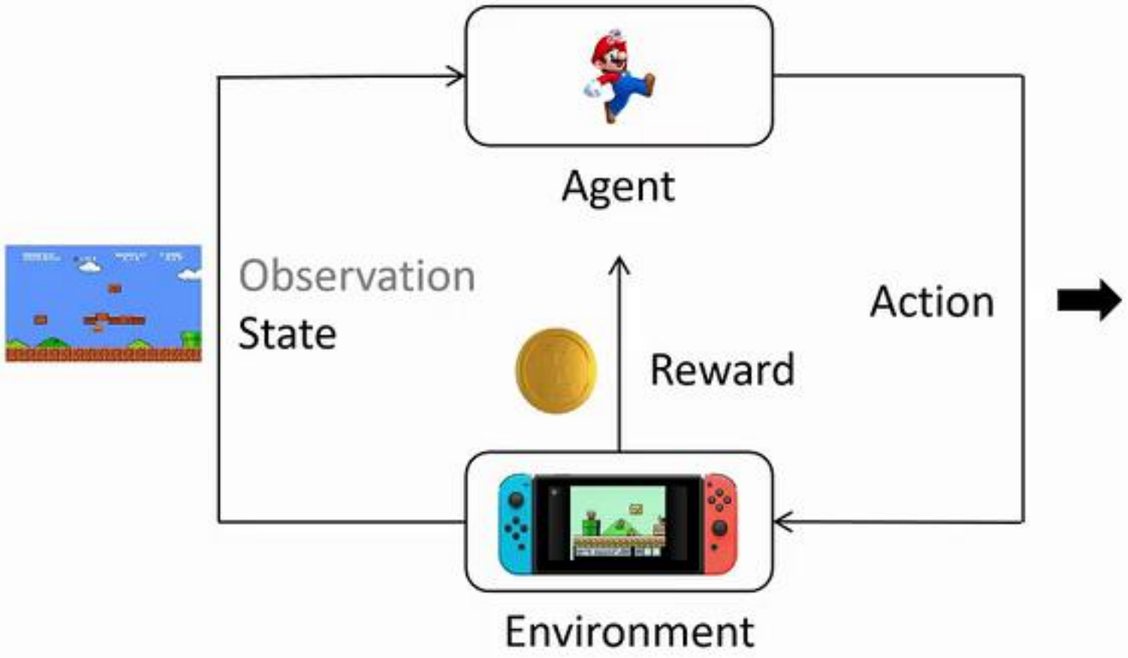
RethinkFun bilibili

零基础学习PPO


MARIO
000300
×01
WORLD
1-1
TIME
317


Agent
Observation
State
Action
Reward
Environment

Action Space: 可选择的动作，比如 {left，up，right}

Policy: 策略函数，输入State，输出Action的概率分布。一般用π表示。

$$\pi(left|s_t) = 0.1$$
$$\pi(up|s_t) = 0.2$$
$$\pi(right|s_t) = 0.7$$

Trajectory: 轨迹,用τ表示， 一连串状态和动作的序列。Episode,Rollout。 $\{s_0, a_0, s_1, a_1, \ldots\}$

$s_{t+1} = f(s_t, a_t)$ 确定

$s_{t+1} = P(\cdot|s_t, a_t)$ 随机

Return: 回报，从当前时间点到游戏结束的Reward的累积和。

老师告诉你，小明考试情况：
20%概率考80分
80%概率考90分

那么你对小明考试的期望结果是多少呢?
$0.2 * 80 + 0.8 * 90 = 88$
再考n次，统计平均成绩。

期望：每个可能结果的概率与其结果值的乘积之和

$$\mathrm{E}(\mathrm{x})_{x\sim p(x)} = \sum_{x} x * p(x) \approx \frac{1}{n}\sum_{i=1}^{n} x \quad x\sim p(x)$$

老师告诉你，小明考试情况：
20%概率考80分
80%概率考90分

那么你对小明考试的期望结果是多少呢?
$0.2 * 80 + 0.8 * 90 = 88$
再考n次，统计平均成绩。

期望：每个可能结果的概率与其结果值的乘积之和

$$\mathrm{E}(\mathrm{x})_{x\sim p(x)} = \sum_{x} x * p(x) \approx \frac{1}{n}\sum_{i=1}^{n} x \quad x\sim p(x)$$

目标：训练一个Policy神经网络π，在所有状态S下，给出相应的Action，得到Return的期望最大。

老师告诉你，小明考试情况：
20%概率考80分
80%概率考90分

那么你对小明考试的期望结果是多少呢?
$0.2 * 80 + 0.8 * 90 = 88$
再考n次，统计平均成绩。

期望：每个可能结果的概率与其结果值的乘积之和

$$\mathrm{E}(\mathrm{x})_{x\sim p(x)} = \sum_{x} x * p(x) \approx \frac{1}{n}\sum_{i=1}^{n} x \quad x\sim p(x)$$

目标：训练一个Policy神经网络π，在所有状态S下，给出相应的Action，得到Return的期望最大。

目标：训练一个Policy神经网络π，在所有的Trajectory中，得到Return的期望最大。

$$E(R(\tau))_{\tau \sim P_\theta(\tau)} = \sum_{\tau} R(\tau) P_\theta(\tau)$$

$$E(R(\tau))_{\tau\sim P_\theta(\tau)} = \sum_\tau R(\tau)P_\theta(\tau) \qquad \nabla E(R(\tau))_{\tau\sim P_\theta(\tau)} = \nabla \sum_\tau R(\tau)P_\theta(\tau)$$

$$E(R(\tau))_{\tau\sim P_\theta(\tau)} = \sum_\tau R(\tau)P_\theta(\tau)$$

$$\nabla E(R(\tau))_{\tau\sim P_\theta(\tau)} = \nabla \sum_\tau R(\tau)P_\theta(\tau)$$

$$= \sum_\tau R(\tau)\nabla P_\theta(\tau)$$

$$= \sum_\tau R(\tau)\nabla P_\theta(\tau)\frac{P_\theta(\tau)}{P_\theta(\tau)}$$

$$E(R(\tau))_{\tau\sim P_\theta(\tau)} = \sum_\tau R(\tau)P_\theta(\tau)$$

$$\nabla E(R(\tau))_{\tau\sim P_\theta(\tau)} = \nabla \sum_\tau R(\tau)P_\theta(\tau)$$

$$= \sum_\tau R(\tau)\nabla P_\theta(\tau)$$

$$= \sum_\tau R(\tau)\nabla P_\theta(\tau)\frac{P_\theta(\tau)}{P_\theta(\tau)}$$

$$= \sum_\tau P_\theta(\tau)R(\tau)\frac{\nabla P_\theta(\tau)}{P_\theta(\tau)}$$

$$E(R(\tau))_{\tau\sim P_\theta(\tau)} = \sum_\tau R(\tau)P_\theta(\tau)$$

$$\nabla E(R(\tau))_{\tau\sim P_\theta(\tau)} = \nabla \sum_\tau R(\tau)P_\theta(\tau)$$

$$= \sum_\tau R(\tau)\nabla P_\theta(\tau)$$

$$= \sum_\tau R(\tau)\nabla P_\theta(\tau)\frac{P_\theta(\tau)}{P_\theta(\tau)}$$

$$= \sum_\tau P_\theta(\tau)R(\tau)\frac{\nabla P_\theta(\tau)}{P_\theta(\tau)}$$

$$= \sum_\tau P_\theta(\tau)R(\tau)\frac{\nabla P_\theta(\tau)}{P_\theta(\tau)}$$

$$E(R(\tau))_{\tau\sim P_\theta(\tau)} = \sum_\tau R(\tau)P_\theta(\tau)$$

$$\nabla E(R(\tau))_{\tau\sim P_\theta(\tau)} = \nabla \sum_\tau R(\tau)P_\theta(\tau)$$

$$= \sum_\tau R(\tau)\nabla P_\theta(\tau)$$

$$= \sum_\tau R(\tau)\nabla P_\theta(\tau)\frac{P_\theta(\tau)}{P_\theta(\tau)}$$

$$= \sum_\tau P_\theta(\tau)R(\tau)\frac{\nabla P_\theta(\tau)}{P_\theta(\tau)}$$

$$= \sum_\tau P_\theta(\tau)R(\tau)\frac{\nabla P_\theta(\tau)}{P_\theta(\tau)}$$

$$\approx \frac{1}{N}\sum_{n=1}^{N} R(\tau^n)\frac{\nabla P_\theta(\tau^n)}{P_\theta(\tau^n)}$$

$$E(R(\tau))_{\tau\sim P_\theta(\tau)} = \sum_\tau R(\tau)P_\theta(\tau)$$

$$\nabla E(R(\tau))_{\tau\sim P_\theta(\tau)} = \nabla \sum_\tau R(\tau)P_\theta(\tau)$$

$$= \sum_\tau R(\tau)\nabla P_\theta(\tau)$$

$$= \sum_\tau R(\tau)\nabla P_\theta(\tau)\frac{P_\theta(\tau)}{P_\theta(\tau)}$$

$$= \sum_\tau P_\theta(\tau)R(\tau)\frac{\nabla P_\theta(\tau)}{P_\theta(\tau)}$$

$$= \sum_\tau P_\theta(\tau)R(\tau)\frac{\nabla P_\theta(\tau)}{P_\theta(\tau)}$$

$$\approx \frac{1}{N}\sum_{n=1}^{N} R(\tau^n)\frac{\nabla P_\theta(\tau^n)}{P_\theta(\tau^n)} \qquad \nabla \log f(x) = \frac{\nabla f(x)}{f(x)}$$

$$= \frac{1}{N}\sum_{n=1}^{N} R(\tau^n)\nabla \log P_\theta(\tau^n) \qquad \tau\sim P_\theta(\tau)$$

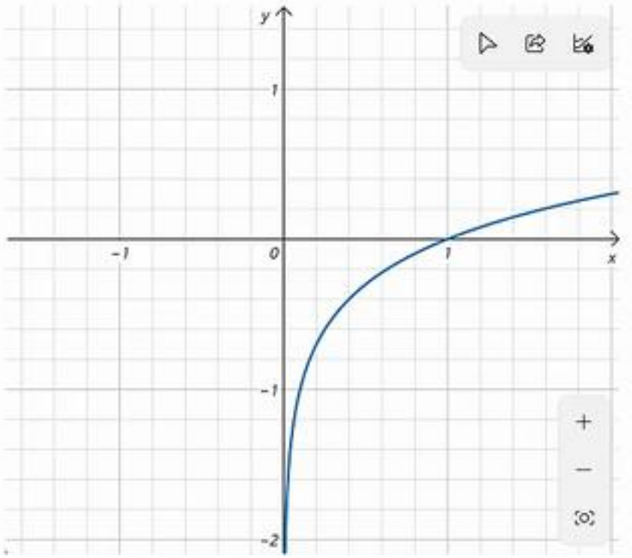
$$= \frac{1}{N}\sum_{n=1}^{N} R(\tau^n)\nabla \log P_\theta(\tau^n)$$

$$= \frac{1}{N}\sum_{n=1}^{N} R(\tau^n)\nabla \log \prod_{t=1}^{T_n} P_\theta(a_n^t|s_n^t)$$

$$= \frac{1}{N}\sum_{n=1}^{N} R(\tau^n)\sum_{t=1}^{T_n} \nabla \log P_\theta(a_n^t|s_n^t)$$

$$= \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} R(\tau^n)\nabla \log P_\theta(a_n^t|s_n^t)$$

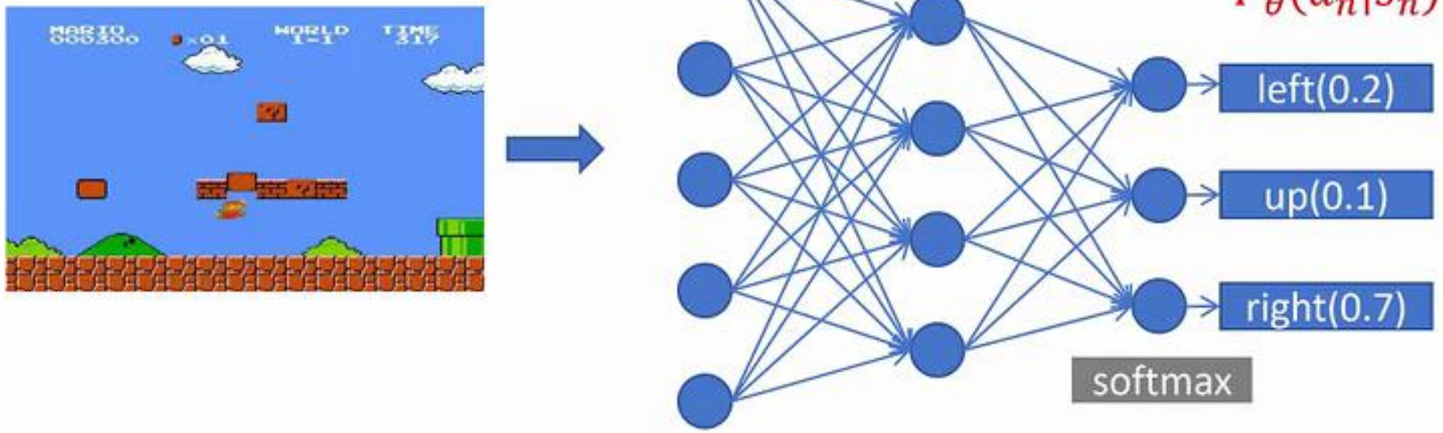
$$\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} R(\tau^n)\log P_\theta(a_n^t|s_n^t)$$



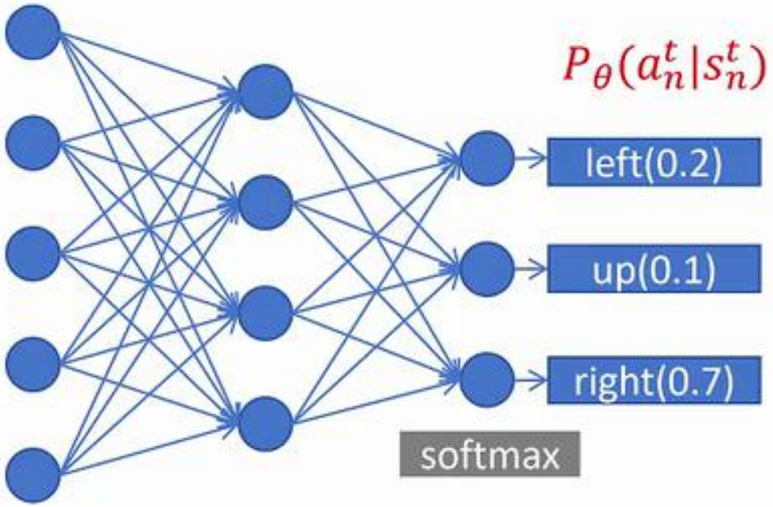
Policy gradient

$$\text{Loss}=-\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} R(\tau^n)\log P_\theta(a_n^t|s_n^t)$$

$$\text{Loss}=-\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} R(\tau^n)\log P_\theta(a_n^t|s_n^t)$$



$$\text{Loss}=-\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} R(\tau^n)\log P_\theta(a_n^t|s_n^t)$$



$R(\tau^n)$

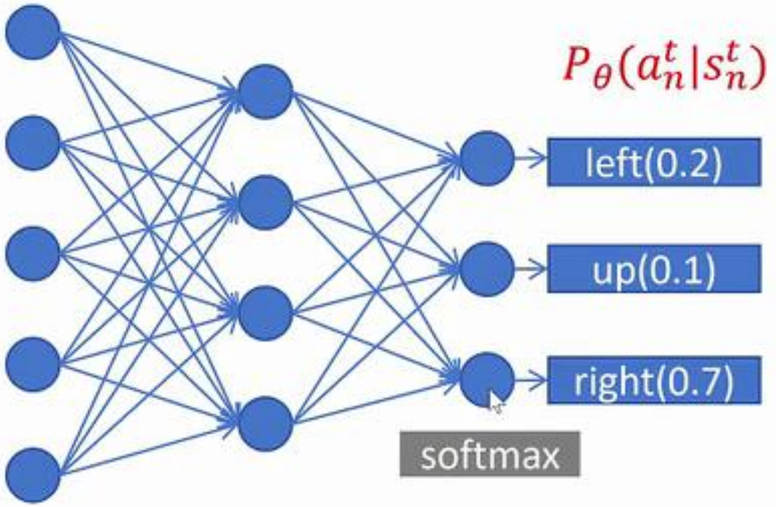
$\tau^1 \rightarrow R(\tau^1)$

$\tau^2 \rightarrow R(\tau^2)$

$\vdots$

$\tau^n \rightarrow R(\tau^n)$

$$\text{Loss}=-\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} R(\tau^n)\log P_\theta(a_n^t|s_n^t)$$



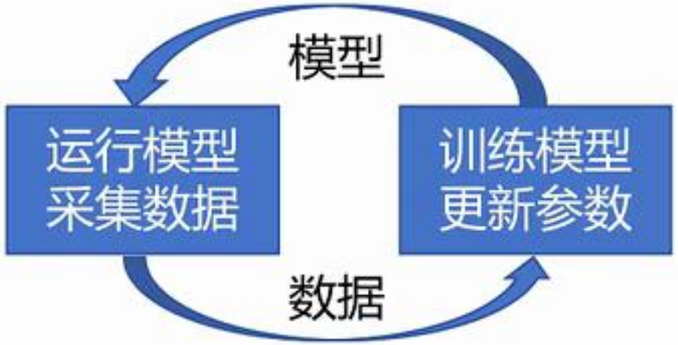
$$R(\tau^n)$$

$$\tau^1 -> R(\tau^1)$$

$$\tau^2 -> R(\tau^2)$$

$$\vdots$$

$$\tau^n -> R(\tau^n)$$



On Policy

$$= \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} R(\tau^n)\nabla \log P_\theta(a_n^t|s_n^t)$$

$$= \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} R(\tau^n)\nabla \log P_\theta(a_n^t|s_n^t) \qquad R(\tau^n) \to \sum_{t'=t}^{T_n} \gamma^{t'-t} r_{t'}^n = R_t^n$$

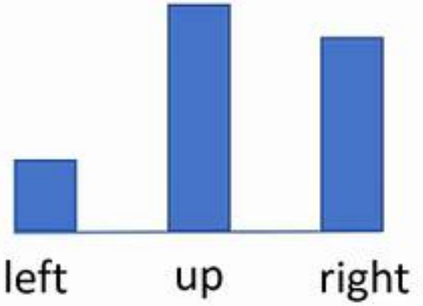
$$= \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} R(\tau^n)\nabla \log P_\theta(a_n^t|s_n^t) \qquad R(\tau^n) \to \sum_{t'=t}^{T_n} \gamma^{t'-t} r_{t'}^n = R_t^n$$

$$= \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} R_t^n \nabla \log P_\theta(a_n^t|s_n^t)$$

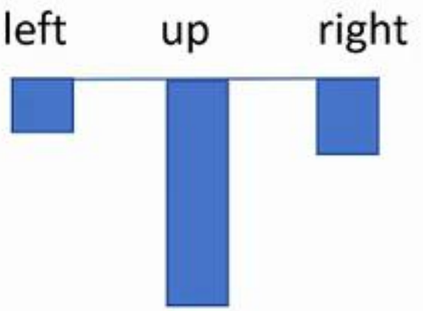
$$= \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} R(\tau^n)\nabla \log P_\theta(a_n^t|s_n^t)$$

$$= \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} R_t^n \nabla \log P_\theta(a_n^t|s_n^t)$$

$$R(\tau^n) \rightarrow \sum_{t'=t}^{T_n} \gamma^{t'-t} r_{t'}^n = R_t^n$$



好的局势

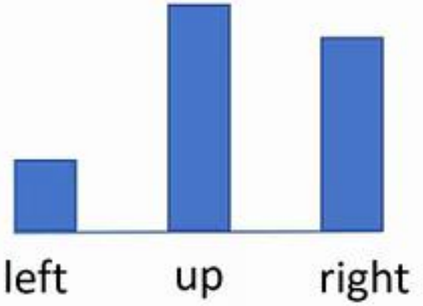


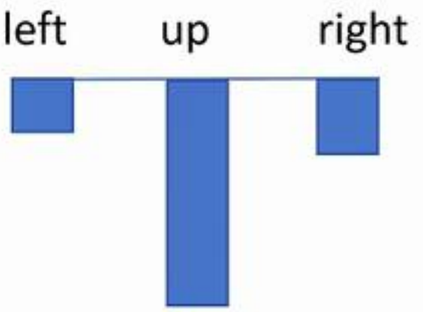
坏的局势

$$= \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} R(\tau^n)\nabla \log P_\theta(a_n^t|s_n^t)$$

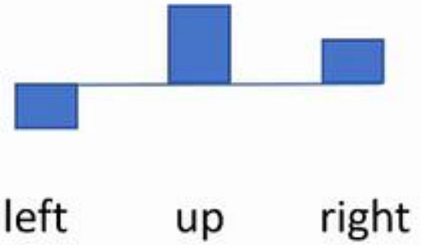
$$R(\tau^n) \rightarrow \sum_{t'=t}^{T_n} \gamma^{t'-t} r_{t'}^n = R_t^n$$

$$= \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} R_t^n \nabla \log P_\theta(a_n^t|s_n^t)$$

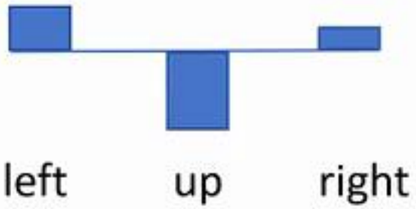




好的局势



坏的局势

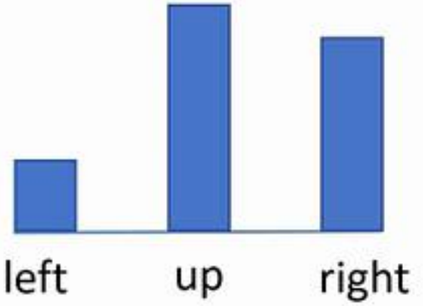


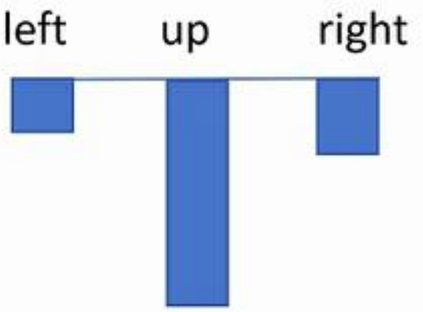
$$= \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} R(\tau^n)\nabla \log P_\theta(a_n^t|s_n^t)$$

$$R(\tau^n) \rightarrow \sum_{t'=t}^{T_n} \gamma^{t'-t} r_{t'}^n = R_t^n$$

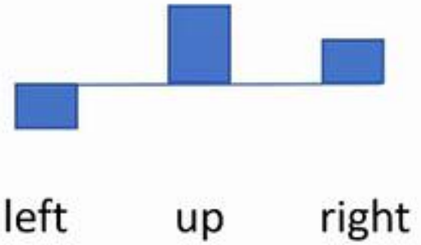
$$= \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} R_t^n \nabla \log P_\theta(a_n^t|s_n^t)$$

$$= \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} (R_t^n - B(s_n^t))\nabla \log P_\theta(a_n^t|s_n^t)$$





好的局势



坏的局势

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n}(R_t^n - B(s_n^t))\nabla \log P_\theta(a_n^t|s_n^t)$$

Action-Value Function

R_t^n 每次都是一次随机采样，方差很大，训练不稳定。

$Q_\theta(s,a)$在state s下，做出Action a，期望的回报。动作价值函数。

State-Value Function

$V_\theta(s)$在state s下, 期望的回报。状态价值函数。

Advantage Function

$A_\theta(s,a) = Q_\theta(s,a) - V_\theta(s)$ 在state s下，做出Action a，比其他动作能带来多少优势。

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n}A_\theta(s_n^t, a_n^t)\nabla \log P_\theta(a_n^t|s_n^t)$$

$A_\theta(s,a) = Q_\theta(s,a) - V_\theta(s)$

$Q_\theta(s,a)$在state s下，做出Action a，期望的回报。动作价值函数。

⋮

$$A_\theta(s,a) = Q_\theta(s,a) - V_\theta(s)$$

$Q_\theta(s,a)$在state s下，做出Action a，期望的回报。动作价值函数。

$$Q_\theta(s_t,a) = r_t + \gamma * V_\theta(s_{t+1})$$

⋮

$$A_\theta(s,a) = Q_\theta(s,a) - V_\theta(s)$$

$Q_\theta(s,a)$在state s下，做出Action a，期望的回报。动作价值函数。

$$Q_\theta(s_t,a) = r_t + \gamma * V_\theta(s_{t+1})$$

$$A_\theta(s_t,a) = r_t + \gamma * V_\theta(s_{t+1}) - V_\theta(s_t)$$

$$A_\theta(s,a) = Q_\theta(s,a) - V_\theta(s)$$

$Q_\theta(s,a)$在state s下，做出Action a，期望的回报。动作价值函数。

$$Q_\theta(s_t,a) = r_t + \gamma * V_\theta(s_{t+1})$$

$$A_\theta(s_t,a) = r_t + \gamma * V_\theta(s_{t+1}) - V_\theta(s_t)$$

$$V_\theta(s_{t+1}) \approx r_{t+1} + \gamma * V_\theta(s_{t+2})$$

⋮

$$A_\theta(s,a) = Q_\theta(s,a) - V_\theta(s)$$

$Q_\theta(s,a)$在state s下，做出Action a，期望的回报。动作价值函数。

$$Q_\theta(s_t,a) = r_t + \gamma * V_\theta(s_{t+1})$$

$$A_\theta(s_t,a) = r_t + \gamma * V_\theta(s_{t+1}) - V_\theta(s_t) \qquad A_\theta^1(s_t,a) = r_t + \gamma * V_\theta(s_{t+1}) - V_\theta(s_t)$$

$$V_\theta(s_{t+1}) \approx r_{t+1} + \gamma * V_\theta(s_{t+2})$$

⋮

$$A_\theta(s,a) = Q_\theta(s,a) - V_\theta(s)$$

$Q_\theta(s,a)$在state s下，做出Action a，期望的回报。动作价值函数。

$$Q_\theta(s_t,a) = r_t + \gamma * V_\theta(s_{t+1})$$

$$A_\theta(s_t,a) = r_t + \gamma * V_\theta(s_{t+1}) - V_\theta(s_t)$$

$$V_\theta(s_{t+1}) \approx r_{t+1} + \gamma * V_\theta(s_{t+2})$$

$$A_\theta^1(s_t,a) = r_t + \gamma * V_\theta(s_{t+1}) - V_\theta(s_t)$$

$$A_\theta^2(s_t,a) = r_t + \gamma * r_{t+1} + \gamma^2 * V_\theta(s_{t+2}) - V_\theta(s_t)$$

$$A_\theta^3(s_t,a) = r_t + \gamma * r_{t+1} + \gamma^2 * r_{t+2} + \gamma^3 V_\theta(s_{t+3}) - V_\theta(s_t)$$

$$\vdots$$

$$A_\theta(s,a) = Q_\theta(s,a) - V_\theta(s)$$

$Q_\theta(s,a)$在state s下，做出Action a，期望的回报。动作价值函数。

$$Q_\theta(s_t,a) = r_t + \gamma * V_\theta(s_{t+1})$$

$$A_\theta(s_t,a) = r_t + \gamma * V_\theta(s_{t+1}) - V_\theta(s_t)$$

$$V_\theta(s_{t+1}) \approx r_{t+1} + \gamma * V_\theta(s_{t+2})$$

$$A_\theta^1(s_t,a) = r_t + \gamma * V_\theta(s_{t+1}) - V_\theta(s_t)$$

$$A_\theta^2(s_t,a) = r_t + \gamma * r_{t+1} + \gamma^2 * V_\theta(s_{t+2}) - V_\theta(s_t)$$

$$A_\theta^3(s_t,a) = r_t + \gamma * r_{t+1} + \gamma^2 * r_{t+2} + \gamma^3 V_\theta(s_{t+3}) - V_\theta(s_t)$$

$$\vdots$$

$$A_\theta^T(s_t,a) = r_t + \gamma * r_{t+1} + \gamma^2 * r_{t+2} + \gamma^3 * r_{t+3} + \cdots + \gamma^T * r_T - V_\theta(s_t)$$

$$A_\theta(s,a) = Q_\theta(s,a) - V_\theta(s)$$

$Q_\theta(s,a)$在state s下，做出Action a，期望的回报。动作价值函数。

$$Q_\theta(s_t,a) = r_t + \gamma * V_\theta(s_{t+1})$$

$$A_\theta(s_t,a) = r_t + \gamma * V_\theta(s_{t+1}) - V_\theta(s_t)$$

$$V_\theta(s_{t+1}) \approx r_{t+1} + \gamma * V_\theta(s_{t+2})$$

$$A_\theta^1(s_t,a) = r_t + \gamma * V_\theta(s_{t+1}) - V_\theta(s_t)$$

$$A_\theta^2(s_t,a) = r_t + \gamma * r_{t+1} + \gamma^2 * V_\theta(s_{t+2}) - V_\theta(s_t)$$

$$A_\theta^3(s_t,a) = r_t + \gamma * r_{t+1} + \gamma^2 * r_{t+2} + \gamma^3 V_\theta(s_{t+3}) - V_\theta(s_t)$$

$$\vdots$$

$$A_\theta^T(s_t,a) = r_t + \gamma * r_{t+1} + \gamma^2 * r_{t+2} + \gamma^3 * r_{t+3} + \cdots + \gamma^T * r_T - V_\theta(s_t)$$

$$A_\theta^1(s_t,a) = r_t + \gamma * V_\theta(s_{t+1}) - V_\theta(s_t)$$

$$A_\theta^2(s_t,a) = r_t + \gamma * r_{t+1} + \gamma^2 * V_\theta(s_{t+2}) - V_\theta(s_t)$$

$$A_\theta^3(s_t,a) = r_t + \gamma * r_{t+1} + \gamma^2 * r_{t+2} + \gamma^3 V_\theta(s_{t+3}) - V_\theta(s_t)$$

$$\vdots$$

$$A_\theta^T(s_t,a) = r_t + \gamma * r_{t+1} + \gamma^2 * r_{t+2} + \gamma^3 * r_{t+3} + \cdots + \gamma^T * r_T - V_\theta(s_t)$$

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

$$\delta_{t+1}^V = r_{t+1} + \gamma * V_\theta(s_{t+2}) - V_\theta(s_{t+1})$$

$$A_\theta^1(s_t,a) = \delta_t^V$$

$$A_\theta^2(s_t,a) = \delta_t^V + \gamma\delta_{t+1}^V$$

$$A_\theta^3(s_t,a) = \delta_t^V + \gamma\delta_{t+1}^V + \gamma^2\delta_{t+2}^V$$

$$\vdots$$

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

$$A_\theta^1(s_t, a) = \delta_t^V$$

$$A_\theta^2(s_t, a) = \delta_t^V + \gamma\delta_{t+1}^V$$

$$A_\theta^3(s_t, a) = \delta_t^V + \gamma\delta_{t+1}^V + \gamma^2\delta_{t+2}^V$$

$$\vdots$$

Generalized Advantage Estimation (GAE)

$$A_\theta^{GAE}(s_t, a) = (1-\lambda)(A_\theta^1 + \lambda * A_\theta^2 + \lambda^2 A_\theta^3 + \cdots)$$

$\lambda = 0.9$:

$$A_\theta^{GAE} = 0.1A_\theta^1 + 0.09A_\theta^2 + 0.081A_\theta^3 + \cdots$$

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

$$A_\theta^1(s_t, a) = \delta_t^V$$

$$A_\theta^2(s_t, a) = \delta_t^V + \gamma\delta_{t+1}^V$$

$$A_\theta^3(s_t, a) = \delta_t^V + \gamma\delta_{t+1}^V + \gamma^2\delta_{t+2}^V$$

$$\vdots$$

Generalized Advantage Estimation (GAE)

$$A_\theta^{GAE}(s_t, a) = (1-\lambda)(A_\theta^1 + \lambda * A_\theta^2 + \lambda^2 A_\theta^3 + \cdots) \qquad \lambda = 0.9: \qquad A_\theta^{GAE} = 0.1A_\theta^1 + 0.09A_\theta^2 + 0.081A_\theta^3 + \cdots$$

$$= (1-\lambda)(\delta_t^V + \lambda * (\delta_t^V + \gamma\delta_{t+1}^V) + \lambda^2(\delta_t^V + \gamma\delta_{t+1}^V + \gamma^2\delta_{t+2}^V) + \cdots)$$

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

$$A_\theta^1(s_t, a) = \delta_t^V$$

$$A_\theta^2(s_t, a) = \delta_t^V + \gamma\delta_{t+1}^V$$

$$A_\theta^3(s_t, a) = \delta_t^V + \gamma\delta_{t+1}^V + \gamma^2\delta_{t+2}^V$$

$$\vdots$$

Generalized Advantage Estimation (GAE)

$$A_\theta^{GAE}(s_t, a) = (1-\lambda)(A_\theta^1 + \lambda * A_\theta^2 + \lambda^2 A_\theta^3 + \cdots) \qquad \lambda = 0.9: \qquad A_\theta^{GAE} = 0.1A_\theta^1 + 0.09A_\theta^2 + 0.081A_\theta^3 + \cdots$$

$$= (1-\lambda)(\delta_t^V + \lambda * (\delta_t^V + \gamma\delta_{t+1}^V) + \lambda^2(\delta_t^V + \gamma\delta_{t+1}^V + \gamma^2\delta_{t+2}^V) + \cdots)$$

$$= (1-\lambda)(\delta_t^V(1 + \lambda + \lambda^2 + \cdots) + \gamma\delta_{t+1}^V * (\lambda + \lambda^2 + \cdots) + \cdots)$$

$$= (1-\lambda)(\delta_t^V \frac{1}{1-\lambda} + \gamma\delta_{t+1}^V \frac{\lambda}{1-\lambda} + \cdots)$$

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

$$A_\theta^{GAE}(s_t, a) = \sum_{b=0}^{\infty} (\gamma\lambda)^b \, \delta_{t+b}^V$$

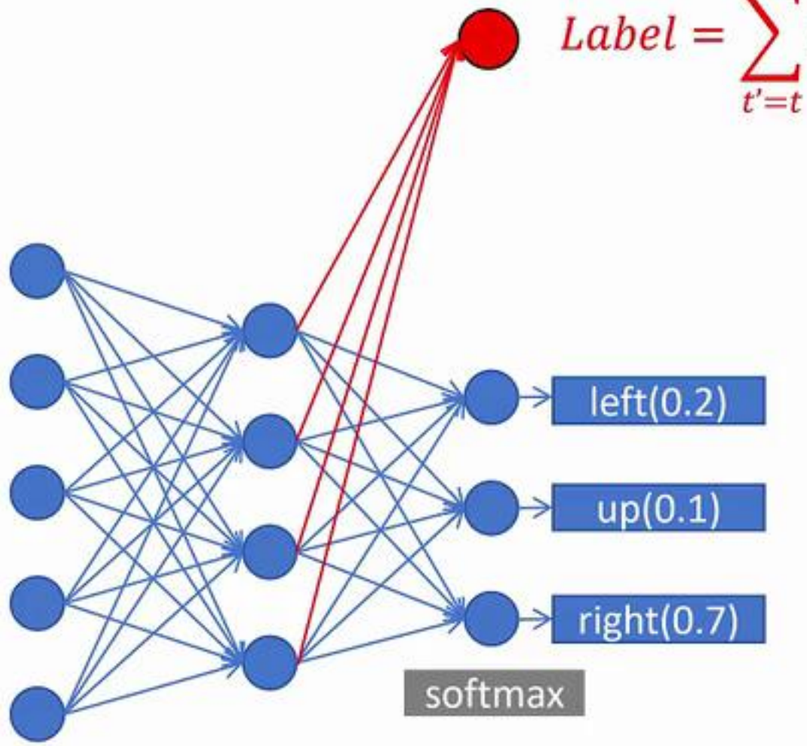
$$\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} A_\theta^{GAE}(s_n^t, a_n^t)\nabla \log P_\theta(a_n^t|s_n^t)$$

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

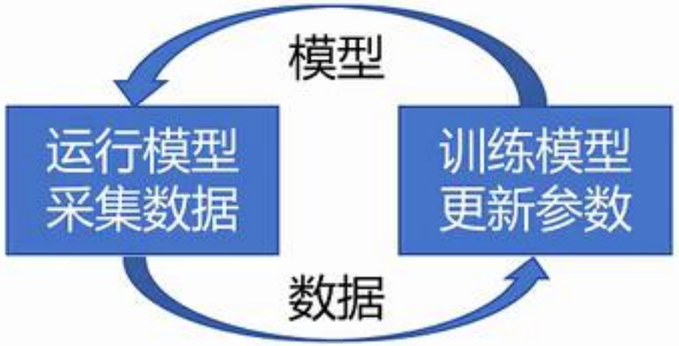
$$A_\theta^{GAE}(s_t, a) = \sum_{b=0}^{\infty} (\gamma\lambda)^b \, \delta_{t+b}^V$$

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} A_\theta^{GAE}(s_n^t, a_n^t)\nabla \log P_\theta(a_n^t|s_n^t)$$

$$Label = \sum_{t'=t}^{T_n} \gamma^{t'-t} r_{t'}^n$$

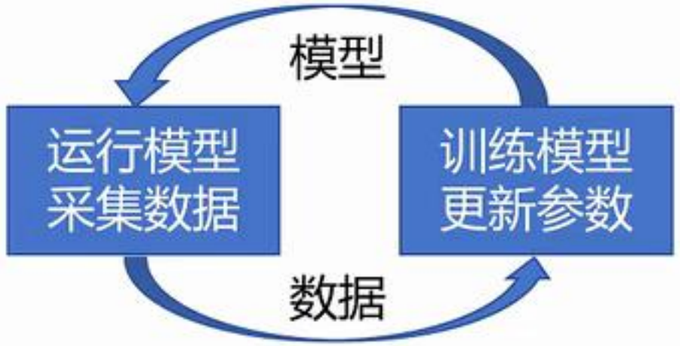


Proximal Policy Optimization(PPO) 邻近策略优化



On Policy

Off Policy



On Policy

Off Policy

重要性采样

重要性采样

$$\mathrm{E}(f(x))_{x\sim p(x)} = \sum_{x} f(x) * p(x)$$

重要性采样

$$\mathrm{E}(f(x))_{x\sim p(x)} = \sum_x f(x) * p(x)$$

$$= \sum_x f(x) * p(x) \frac{q(x)}{q(x)}$$

重要性采样

$$\mathrm{E}(f(x))_{x\sim p(x)} = \sum_x f(x) * p(x)$$

$$= \sum_x f(x) * p(x) \frac{q(x)}{q(x)}$$

$$= \sum_x f(x) \frac{p(x)}{q(x)} * q(x)$$

重要性采样

$$
\begin{aligned}
\mathrm{E}(f(x))_{x\sim p(x)} &= \sum_x f(x) * p(x) \\
&= \sum_x f(x) * p(x) \frac{q(x)}{q(x)} \\
&= \sum_x f(x) \frac{p(x)}{q(x)} * q(x) \\
&= \mathrm{E}(f(x) \frac{p(x)}{q(x)})_{x\sim q(x)}
\end{aligned}
$$

重要性采样

$$
\begin{aligned}
\mathrm{E}(f(x))_{x\sim p(x)} &= \sum_x f(x) * p(x) \\
&= \sum_x f(x) * p(x)\frac{q(x)}{q(x)} \\
&= \sum_x f(x)\frac{p(x)}{q(x)} * q(x) \\
&= \mathrm{E}(f(x)\frac{p(x)}{q(x)})_{x\sim q(x)} \\
&\approx \frac{1}{N}\sum_{n=1}^{N} f(x)\frac{p(x)}{q(x)}_{x\sim q(x)}
\end{aligned}
$$

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} A_{\theta}^{GAE}(s_n^t, a_n^t)\nabla \log P_{\theta}(a_n^t|s_n^t)$$

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} A_{\theta}^{GAE}(s_n^t, a_n^t)\nabla \log P_{\theta}(a_n^t|s_n^t)$$

$$= \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} A_{\theta'}^{GAE}(s_n^t, a_n^t)\frac{P_{\theta}(a_n^t|s_n^t)}{P_{\theta'}(a_n^t|s_n^t)}\nabla \log P_{\theta}(a_n^t|s_n^t)$$

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} A_{\theta}^{GAE}(s_n^t, a_n^t)\nabla \log P_{\theta}(a_n^t|s_n^t)$$

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

$$= \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} A_{\theta'}^{GAE}(s_n^t, a_n^t)\frac{P_{\theta}(a_n^t|s_n^t)}{P_{\theta'}(a_n^t|s_n^t)}\nabla \log P_{\theta}(a_n^t|s_n^t)$$

$$= \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} A_{\theta'}^{GAE}(s_n^t, a_n^t)\frac{\cancel{P_{\theta}(a_n^t|s_n^t)}}{P_{\theta'}(a_n^t|s_n^t)}\frac{\nabla P_{\theta}(a_n^t|s_n^t)}{\cancel{P_{\theta}(a_n^t|s_n^t)}}$$

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} A_{\theta}^{GAE}(s_n^t, a_n^t)\nabla \log P_{\theta}(a_n^t|s_n^t)$$

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

$$= \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} A_{\theta'}^{GAE}(s_n^t, a_n^t)\frac{P_{\theta}(a_n^t|s_n^t)}{P_{\theta'}(a_n^t|s_n^t)}\nabla \log P_{\theta}(a_n^t|s_n^t)$$

$$= \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} A_{\theta'}^{GAE}(s_n^t, a_n^t)\frac{\cancel{P_{\theta}(a_n^t|s_n^t)}}{P_{\theta'}(a_n^t|s_n^t)}\frac{\nabla P_{\theta}(a_n^t|s_n^t)}{\cancel{P_{\theta}(a_n^t|s_n^t)}}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} A_{\theta'}^{GAE}(s_n^t, a_n^t)\frac{\nabla P_{\theta}(a_n^t|s_n^t)}{P_{\theta'}(a_n^t|s_n^t)}$$

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} A_{\theta}^{GAE}(s_n^t, a_n^t)\nabla \log P_{\theta}(a_n^t|s_n^t)$$

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

$$= \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} A_{\theta'}^{GAE}(s_n^t, a_n^t)\frac{P_{\theta}(a_n^t|s_n^t)}{P_{\theta'}(a_n^t|s_n^t)}\nabla \log P_{\theta}(a_n^t|s_n^t)$$

$$= \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} A_{\theta'}^{GAE}(s_n^t, a_n^t)\frac{\cancel{P_{\theta}(a_n^t|s_n^t)}}{P_{\theta'}(a_n^t|s_n^t)}\frac{\nabla P_{\theta}(a_n^t|s_n^t)}{\cancel{P_{\theta}(a_n^t|s_n^t)}}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} A_{\theta'}^{GAE}(s_n^t, a_n^t)\frac{\nabla P_{\theta}(a_n^t|s_n^t)}{P_{\theta'}(a_n^t|s_n^t)}$$

$$\text{Loss} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} A_{\theta'}^{GAE}(s_n^t, a_n^t)\frac{P_{\theta}(a_n^t|s_n^t)}{P_{\theta'}(a_n^t|s_n^t)}$$

这个学生不能和你差距太大。

不然你很难学到对你有用的经验和教训。

$$Loss_{ppo} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} A_{\theta'}^{GAE}(s_n^t, a_n^t)\frac{P_\theta(a_n^t|s_n^t)}{P_{\theta'}(a_n^t|s_n^t)} + \beta KL(P_\theta, P_{\theta'})$$

$$Loss_{ppo2} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} min(A_{\theta'}^{GAE}(s_n^t, a_n^t)\frac{P_\theta(a_n^t|s_n^t)}{P_{\theta'}(a_n^t|s_n^t)}, clip(\frac{P_\theta(a_n^t|s_n^t)}{P_{\theta'}(a_n^t|s_n^t)}, 1-\varepsilon, 1+\varepsilon)A_{\theta'}^{GAE}(s_n^t, a_n^t))$$