PPO算法.

目标：训练一个 Policy π. 使其在所有state和action下. 找一条运行trajectory τ. 其reward $R(\tau)$最大.

推导： $\max E(R(\tau))_{\tau \sim P_\theta(\tau)} = \sum_{\tau} R(\tau) P_\theta(\tau)$.

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

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

$$\approx \frac{1}{N} \sum_{n=1}^{N} R(\tau_n) \cdot \nabla \log P_\theta(\tau_n).$$

$$= \frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T_n} R(\tau_n) \nabla \log P_\theta(a_t^n | s_t^n).$$

$\Rightarrow$ 损失函数： $-\frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T_n} R(\tau_n) \log P_\theta(a_t^n | s_t^n)$.

$\Rightarrow$ $R(\tau_n)$影响整个action链. 所有action被赋予相同reward.

$$\Rightarrow R(\tau_n) \rightsquigarrow R_t^n = \sum_{t'=t}^{T_n} \gamma^{t'-t} \cdot r_{t'}^n$$

$\Rightarrow -\frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T_n} (R_t^n - B(s_t^n)) \cdot \log P_\theta(a_t^n | s_t^n)$ $\Rightarrow$ $B(s_t^n)$为平均价值. 给出多个"好"action中最"好"的

训练中 $R_t^n - B(s_t^n)$ 用 $Q(s,a) - V_\theta(s)$ 代替.

定义：① $Q(s,a)$为动作价值函数. $\Rightarrow$ 判断s下a的价值.

② $V_\theta(s)$为状态价值函数.

③ $A_\theta(s,a)$为优势函数 $\Rightarrow$ $A_\theta(s,a) = Q(s,a) - V_\theta(s)$.

关系： $Q_\theta(s_t, a) = r_t + \gamma \cdot V_\theta(s_{t+1})$. $V_\theta(s_{t+1}) \approx r_{t+1} + \gamma \cdot V_\theta(s_{t+2})$

目前损失函数： $-\frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T_n} A_\theta(s_t^n, a_t^n) \cdot \log P_\theta(a_t^n | s_t^n)$

$A_\theta(s,a)$ 多步代入

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

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

偏差↑ ↓ ↑ 方差↑

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

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

$$\Rightarrow A_\theta^1(s_t,a) = \delta_t^V, \quad A_\theta^2(s_t,a) = \delta_t^V + \gamma\delta_{t+1}^V \quad \cdots \quad A_\theta^T(s_t,a) = \sum_{k=0}^{T-1}\gamma^k \delta_{t+k}^V$$

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

$$= \sum_{i=0}^{\infty}(\gamma\lambda)^i \delta_{t+i}^V$$

$$\Rightarrow Loss = -\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} A_\theta^{GAE}(s_t^n, a_t^n)\cdot \log P_\theta(a_t^n|s_t^n).$$

其中，A_θ^{GAE} 中的 $V_\theta(s)$ 由神经网络拟合，label 为 $\sum_{t'=t}^{T_n}\gamma^{t'-t} r_{t'}^n$.

重要性采样：

$$E(f(x))_{x\sim p(x)} = \sum_x p(x)f(x) = \sum_x q(x)\frac{p(x)}{q(x)}f(x) = E\left(\frac{p(x)}{q(x)}f(x)\right)_{x\sim q(x)}.$$

off-Policy：从 $q(x)$ 中采样，训练 $p(x)$，即可以从一个固定策略中采样，

用其从环境中取得的反馈训练新策略.

$$Loss = -\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} A_{\theta'}^{GAE}(s_t^n, a_t^n)\cdot\frac{P_\theta(a_t^n|s_t^n)}{P_{\theta'}(a_t^n|s_t^n)}\cdot \log P_\theta(a_t^n|s_t^n).$$

求导更新后等价于 $-\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} A_{\theta'}^{GAE}(s_t^n, a_t^n)\frac{P_\theta(a_t^n|s_t^n)}{P_{\theta'}(a_t^n|s_t^n)} + \beta KL(P_\theta, P_{\theta'})$

截断策略：$-\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n}\min\left(A_{\theta'}^{GAE}(s_t^n, a_t^n)\frac{P_\theta(a_t^n|s_t^n)}{P_{\theta'}(a_t^n|s_t^n)}, clip\left(\frac{P_\theta(a_t^n|s_t^n)}{P_{\theta'}(a_t^n|s_t^n)}, 1-\varepsilon, 1+\varepsilon\right)A_{\theta'}^{GAE}(s_t^n, a_t^n)\right)$

均限制 θ 与 θ' 分布差距不应过大.