PPO (Proximal Policy Optimization)

定义强化学习的优化目标：

$$J(\pi) = \mathbb{E}_{s_0 \sim \rho_0}[V_\pi(s_0)] \tag{1}$$

上式中，s_0 为初始状态，ρ_0 为初始状态分布，$V_\pi(s_0)$ 为在策略 π 下 s_0 的状态价值函数。每次更新策略时，新策略比旧策略优化在：

$$J(\pi_{new}) = J(\pi_{old}) + \mathbb{E}_{s_0, a_0, s_1, a_1, \ldots}\left[\sum_{t=0}^{\infty} \gamma^t A_{\pi_{old}}(s_t, a_t)\right] \tag{2}$$

其中，$s_0 \sim \rho_0$，$a_t \sim \pi_{new}(\cdot|s_t)$，$s_{t+1} \sim P(s_{t+1}|s_t, a_t)$， $P(s_{t+1}|s_t, a_t)$ 为环境转移概率，γ 为折扣因子，$A_\pi(s_t, a_t)$ 是优势函数，其定义为：

$$A_\pi(s_t, a_t) = \mathbb{E}_{s_{t+1} \sim P(s_{t+1}|s_t, a_t)}\left[r(s_t) + \gamma V_\pi(s_{t+1}) - V_\pi(s_t)\right] \tag{3}$$

公式 （2）直观地展示了每次策略的提升是**新策略生成的轨迹上，每一个 状态-动作对 折扣优势函数期望之和**。每次策略更新时，只要让 π_{new} 选择 $A_{\pi_{old}} \geq 0$ 的动作，则 π_{new} 一定比 π_{old} 好（否则策略已经收敛至最优策略），因为这种情况下：

$$J(\pi_{new}) - J(\pi_{old}) = \mathbb{E}_{s_0, a_0, s_1, a_1, \ldots}\left[\sum_{t=0}^{\infty} \gamma^t A_{\pi_{old}}(s_t, a_t)\right] \geq 0 \tag{4}$$

TRPO（Trust Region Policy Optimization）

公式 （2）是在时间层面对 $\gamma^t A_{\pi_{old}}(s_t, a_t)$ 进行求和的。同样也可以在状态空间层面对其进行求和：

$$J(\pi_{new}) = J(\pi_{old}) + \sum_s \rho_{\pi_{new}}(s) \sum_a \pi_{new}(a|s) A_{\pi_{old}}(s, a) \tag{5}$$

TRPO 首先构造了 $J(\pi_{new})$ 在初始点 π_{old} 处的局部拟合值：

$$L_{\pi_{old}}(\pi_{new}) = J(\pi_{new}) + \sum_s \rho_{\pi_{old}}(s) \sum_a \pi_{new}(a|s) A_{\pi_{old}}(s, a) \tag{6}$$

如果策略是参数化的 π_θ，那么 L_π 和 J 一阶近似等价，即：

$$\begin{gathered} L_{\pi_{\theta_0}}(\pi_{\theta_0}) = J(\pi_{\theta_0}) \\ \nabla_\theta L_{\pi_\theta}(\pi_\theta)|_{\theta=\theta_0} = \nabla_\theta J(\pi_\theta)|_{\theta=\theta_0} \end{gathered} \tag{7}$$

对于一次足够小的更新 $\pi_{\theta_0} \rightarrow \pi_{new}$，如果能够提升 L_π，则同样能够提升 J。

以 π_{old} 表示当前策略，如果能够解决 $\pi' = argmax_{\pi'} L_{\pi_{old}}(\pi')$ 即 $\pi'(a|s) = argmax_a A_{\pi_{old}}(s, a)$，则新策略为以下混合策略：

$$\pi_{new}(a|s) = (1-\alpha)\pi_{old}(a|s) + \alpha\pi'(a|s) \tag{8}$$

那么

$$J(\pi_{new}) \geq L_{\pi_{old}}(\pi_{new}) - \frac{2\epsilon\gamma}{(1-\gamma(1-\alpha))(1-\gamma)}\alpha^2 \tag{9}$$

其中，$\epsilon = max_s|\mathbb{E}_{a \sim \pi'(a|s)}[A_{\pi_{old}}(s, a)]|$。

TRPO的核心思想：

使用 π_{old} 与环境互动，获得轨迹 τ_{old} (该轨迹中的每一个状态服从 $\rho_{\pi_{old}}$ 分布)，并利用 τ_{old} 将策略改进为 π_{new}，同时保证 τ_{old} 中所有状态的平均 KL 小于某个阈值 δ。

PPO

由公式（2)，可以得到：

$$J(\pi_{new})-J(\pi_{old})\geq\frac{1}{1-\gamma}\mathbb{E}_{\substack{s\sim D^{\pi_{old}}\\a\sim\pi_{old}}}\left[\frac{\pi_{new}(a|s)}{\pi_{old}(a|s)}A_{\pi_{old}}(s,a)\right]-\frac{\gamma\epsilon}{(1-\gamma)^2}\mathbb{E}_{\substack{s\sim D_{\pi_{old}}\\a\sim\pi_{old}}}\left[\left|\frac{\pi_{new}(a|s)}{\pi_{old}(a|s)}-1\right|\right]\quad(10)$$

不等式右侧部分被称作策略改进的下界(Policy improvement lower bound；PILB)，第一项为代理目标(surrogate objective；SO)，第二项为惩罚项(penalty term；PT)。每次策略改进时，只要保证每次策略改进的下界为正，即PILB = SO - PT ≥ 0，则可以保证新策略优于旧策略。

通过增大执行 $A_{\pi_{old}}(s,a)>0$ 动作的概率，减小执行 $A_{\pi_{old}}(s,a)<0$ 动作的概率，可以增大SO的数值，但如果新旧策略差异太大（即 $\left|\frac{\pi_{new}(a|s)}{\pi_{old}(a|s)}-1\right|\gg 0$ ），PT的数值也会陡增。

证明

公式（2）证明：

$$\begin{aligned}
&\mathbb{E}_{s_0,a_0,s_1,a_1,\ldots}\left[\sum_{t=0}^{\infty}\gamma^t A_{\pi_{old}}(s_t,a_t)\right]\\
&=\mathbb{E}_{s_0,a_0,s_1,a_1,\ldots}\left[\sum_{t=0}^{\infty}\gamma^t\left(r(s_t)+\gamma V_{\pi_{old}}(s_{t+1})-V_{\pi_{old}}(s_t)\right)\right]\\
&=\mathbb{E}_{s_0,a_0,s_1,a_1,\ldots}\left[-V_{\pi_{old}}(s_0)+\sum_{t=0}^{\infty}\gamma^t r(s_t)\right]\\
&=-\mathbb{E}_{s_0}\left[V_{\pi_{old}}(S_0)\right]+\mathbb{E}_{s_0,a_0,s_1,a_1,\ldots}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t)\right]\\
&=-J(\pi_{old})+J(\pi_{new})
\end{aligned}$$

公式（5）证明：

$$\begin{aligned}
J(\pi_{new})&=J(\pi_{old})+\mathbb{E}_{s_0,a_0,s_1,a_1,\ldots}\left[\sum_{t=0}^{\infty}\gamma^t A_{\pi_{old}}(s_t,a_t)\right]\\
&=J(\pi_{old})+\sum_{t=0}^{\infty}\gamma^t\sum_{s_t}\Pr(s_t|\pi_{new})\sum_{a_t}\pi_{new}(a_t|s_t)A_{\pi_{old}}(s_t,a_t)\\
&=J(\pi_{old})+\sum_{s}\sum_{t=0}^{\infty}\gamma^t\Pr(s_t=s|\pi_{new})\sum_{a}\pi_{new}(a|s)A_{\pi_{old}}(s,a)\\
&=J(\pi_{old})+\sum_{s}\rho_{\pi_{new}}(s)\sum_{a}\pi_{new}(a|s)A_{\pi_{old}}(s,a)
\end{aligned}$$

公式（10）证明：

$$J(\pi_{new})-J(\pi_{old})=\sum_{s}\sum_{t=0}^{\infty}\gamma^t\Pr(s_t=s|\pi_{new})\sum_{a}\pi_{new}(a|s)A_{\pi_{old}}(s,a)$$

由于

$$
\begin{aligned}
\sum_s \sum_{t=0}^{\infty} \gamma^t \Pr(s_t = s|\pi_{new}) &= \sum_{t=0}^{\infty} \gamma^t \sum_{s_t} \Pr(s_t|\pi_{new}) \\
&= \sum_{t=0}^{\infty} \gamma^t \\
&= \frac{1}{1-\gamma} \neq 1
\end{aligned}
$$