Proximal Policy Optimization Algorithms

问题背景：为什么需要“更稳定的策略优化”

Policy Gradient Methods （策略梯度方法）

噪声大 + 更新跳跃 ⇒ 训练不稳定

梯度估计 $\hat{g} = \hat{\mathbb{E}}_t \Big[\nabla_\theta \log \pi_\theta(a_t \mid s_t) \hat{A}_t \Big]$

- ➢ 梯度方差大（噪声大），训练易震荡
- ➢ On-policy采样，样本效率低
- ➢ 更新无约束，性能易退化

代理目标 $L^{PG}(\theta) = \hat{\mathbb{E}}_t \Big[\log \pi_\theta(a_t \mid s_t) \hat{A}_t \Big].$

需要“限制更新幅度”的稳定优化

问题背景：为什么需要“更稳定的策略优化”

Trust Region Methods（信赖域方法）

代理目标函数在**策略更新受限**的情况下最大化

比率目标，控制倾向于选择什么策略

硬约束

$$\begin{aligned} \underset{\theta}{\text{maximize}} \quad & \hat{\mathbb{E}}_t\left[\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}\hat{A}_t\right] \\ \text{subject to} \quad & \boxed{\hat{\mathbb{E}}_t[\text{KL}[\pi_{\theta_{\text{old}}}(\cdot \mid s_t), \pi_\theta(\cdot \mid s_t)]] \leq \delta.} \end{aligned}$$

➢ 实现复杂，计算开销大

➢ 约束仅近似满足（采样估计 + 二阶近似）

KL散度衡量两个概率分布之间的差异，
用于保证策略更新受限

惩罚项

$$\underset{\theta}{\text{maximize}}\, \hat{\mathbb{E}}_t\left[\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}\hat{A}_t - \boxed{\beta\, \text{KL}[\pi_{\theta_{\text{old}}}(\cdot \mid s_t), \pi_\theta(\cdot \mid s_t)]}\right]$$

➢ KL 惩罚系数难以调整

PPO-Clip：用 clip 近似“信赖域”约束

- TRPO / CPI 的代理目标

$$L^{CPI}(\theta) = \hat{\mathbb{E}}_t\left[\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}\hat{A}_t\right] = \hat{\mathbb{E}}_t\left[r_t(\theta)\hat{A}_t\right]$$

将概率比例限制在[1−ϵ, 1+ϵ],
ϵ是人为设置的超参数

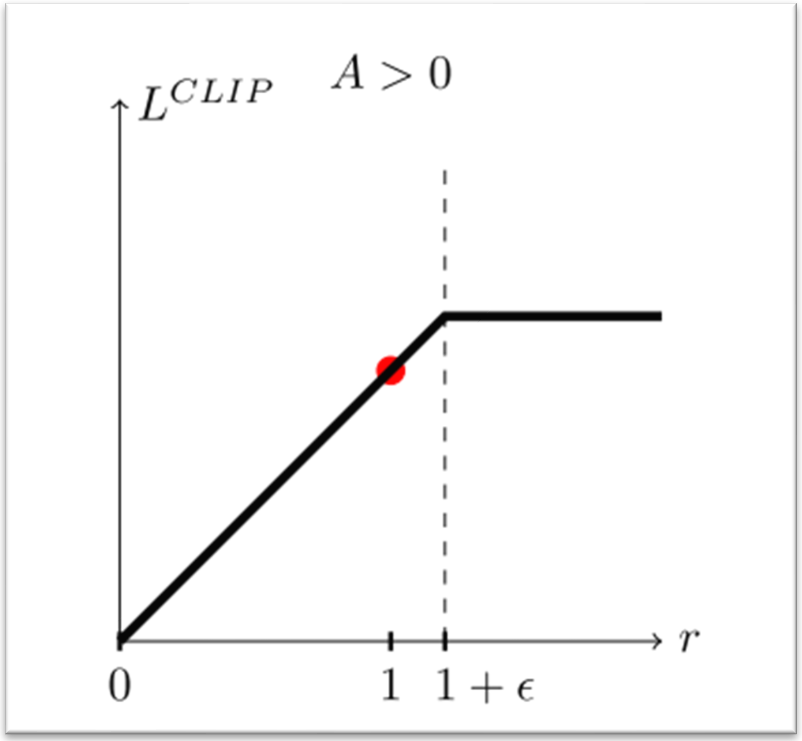
- PPO - Clipped Surrogate Objective

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t\left[\min(r_t(\theta)\hat{A}_t, \boxed{\text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)}\hat{A}_t)\right]$$

$$\text{clip}\left(r_t(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_t = \begin{cases} (1-\epsilon)\hat{A}_t, & r_t(\theta) < 1-\epsilon, \\ r_t(\theta)\hat{A}_t, & 1-\epsilon \le r_t(\theta) \le 1+\epsilon \\ (1+\epsilon)\hat{A}_t, & r_t(\theta) > 1+\epsilon \end{cases}$$

PPO-Clip：用 clip 近似“信赖域”约束

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t\left[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)\right]$$

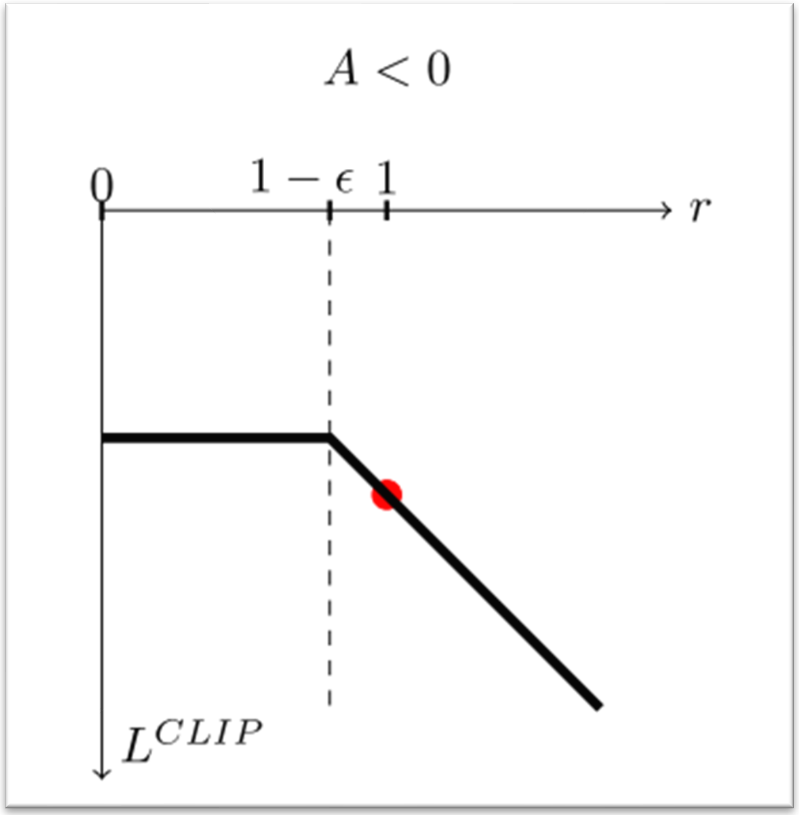


$$\ell_t^{\text{CLIP}}(\theta) = \begin{cases} r_t(\theta)\hat{A}_t, & r_t(\theta) \le 1+\epsilon, \\ (1+\epsilon)\hat{A}_t, & r_t(\theta) > 1+\epsilon. \end{cases}$$

A>0：动作“值得鼓励” → 增加概率（但不让增太多）

PPO-Clip：用 clip 近似“信赖域”约束

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t\Big[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)\Big]$$



$$\ell_t^{\text{CLIP}}(\theta) = \begin{cases} (1-\epsilon)\hat{A}_t, & r_t(\theta) < 1-\epsilon, \\ r_t(\theta)\hat{A}_t, & r_t(\theta) \geq 1-\epsilon. \end{cases}$$

A<0：动作“不值得” → 减少概率（但不让减太多）

PPO-Clip 的核心：悲观下界，限制过大策略更新

$$\ell_t^{CPI}(\theta) = r_t(\theta)\hat{A}_t, \qquad \ell_t^{CLIP}(\theta) = \min\left(\ell_t^{CPI}(\theta), \mathrm{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\right).$$

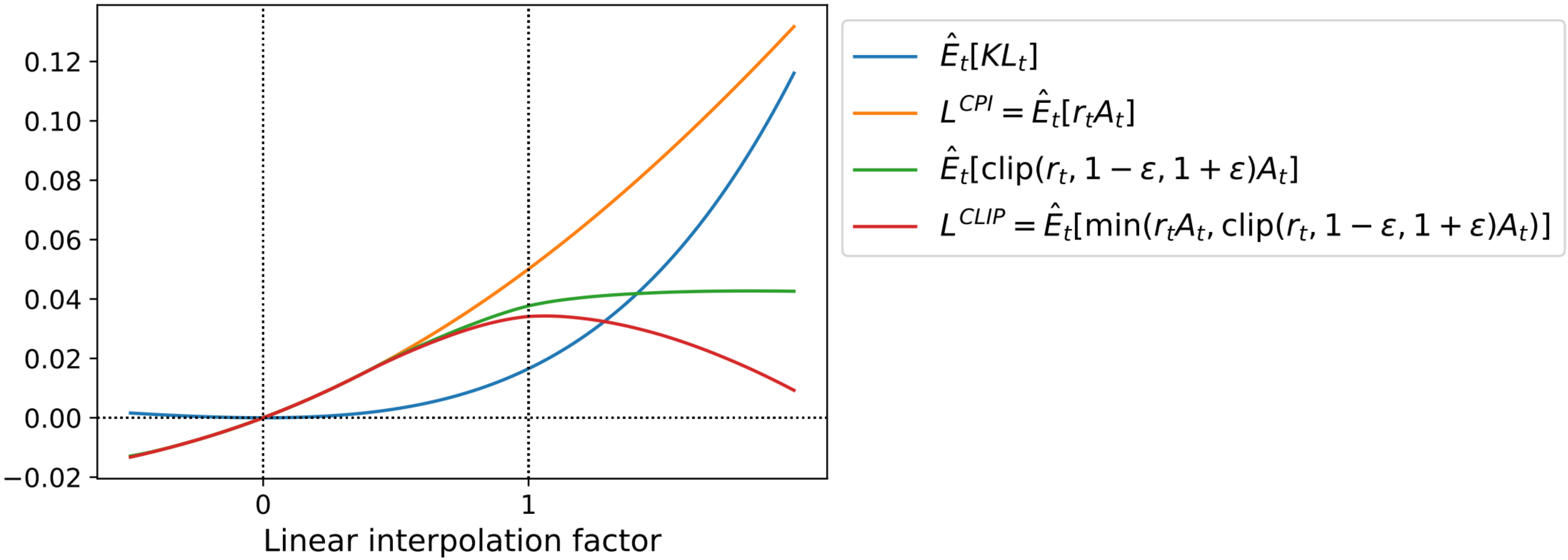
由于对于任意的实数x,y 都有min(x,y) ≤ x

$$\ell_t^{CLIP}(\theta) \leq \ell_t^{CPI}(\theta)$$

取期望后

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t\left[\ell_t^{CLIP}(\theta)\right] \leq \hat{\mathbb{E}}_t\left[\ell_t^{CPI}(\theta)\right] = L^{CPI}(\theta)$$

PPO-Clip 的核心：悲观下界，限制过大策略更新



- 橙线：不限制步长，越走越“乐观”
- 绿线：把比率截断，收益封顶
- 红线：对每个样本取更保守的策略，自动抑制过大更新
- 蓝线：度量策略偏移

Adaptive KL Penalty Coefficient （自适应KL惩罚系数）

- TRPO的惩罚项函数

惩罚项

$$\underset{\theta}{\text{maximize}}\,\hat{\mathbb{E}}_t\left[\frac{\pi_\theta(a_t\mid s_t)}{\pi_{\theta_{\text{old}}}(a_t\mid s_t)}\hat{A}_t-\boxed{\beta\,\text{KL}[\pi_{\theta_{\text{old}}}(\cdot\mid s_t),\pi_\theta(\cdot\mid s_t)]}\right]$$

- KL 惩罚系数难以调整

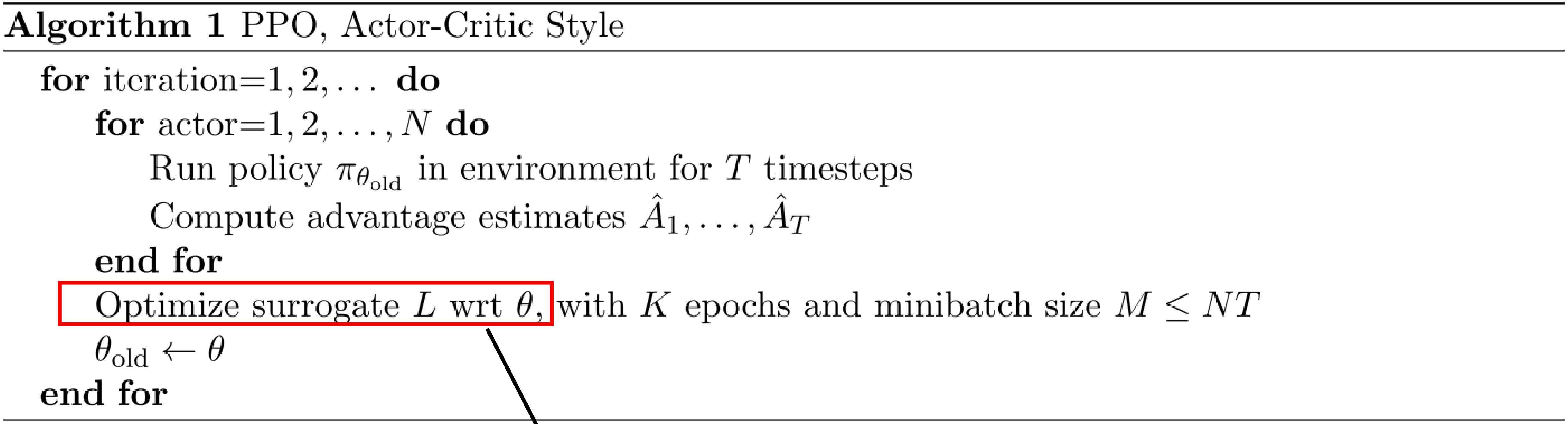
- PPO的自适应KL惩罚系数

$$L^{KLPEN}(\theta)=\hat{\mathbb{E}}_t\left[\frac{\pi_\theta(a_t\mid s_t)}{\pi_{\theta_{\text{old}}}(a_t\mid s_t)}\hat{A}_t-\beta\,\text{KL}[\pi_{\theta_{\text{old}}}(\cdot\mid s_t),\pi_\theta(\cdot\mid s_t)]\right]$$

$$d=\hat{\mathbb{E}}_t\left[\text{KL}\left(\pi_{\theta_{\text{old}}}(\cdot\mid s_t),\pi_\theta(\cdot\mid s_t)\right)\right].$$

$$\beta\leftarrow\begin{cases}\beta/2, & d<d_{\text{targ}}/1.5,\\ 2\beta, & d>1.5d_{\text{targ}},\\ \beta, & \text{otherwise}.\end{cases}$$

PPO算法总览



Algorithm 1 PPO, Actor-Critic Style

```
for iteration=$1, 2, \ldots$ do
    for actor=$1, 2, \ldots, N$ do
        Run policy $\pi_{\theta_{\text{old}}}$ in environment for $T$ timesteps
        Compute advantage estimates $\hat{A}_1, \ldots, \hat{A}_T$
    end for
    Optimize surrogate $L$ wrt $\theta$, with $K$ epochs and minibatch size $M \le NT$
    $\theta_{\text{old}} \leftarrow \theta$
end for
```

$$L_t^{CLIP+VF+S}(\theta) = \hat{\mathbb{E}}_t\left[\underbrace{L_t^{CLIP}(\theta)}_{\text{策略}} - \underbrace{c_1 L_t^{VF}(\theta)}_{\text{价值}} + \underbrace{c_2 S[\pi_\theta](s_t)}_{\text{熵奖励（鼓励探索）}}\right],$$

PPO 的优势估计

优势估计 =（片段内折扣回报 + 末端 补偿）− 价值基线

$$\hat{A}_t = \underbrace{-V(s_t)}_{\text{价值基线}} + \underbrace{r_t + \gamma r_{t+1} + \cdots + \gamma^{T-t+1} r_{T-1}}_{\text{片段内折扣回报}} \underbrace{+ \gamma^{T-t} V(s_T)}_{\text{末端补偿}}$$

- ➢ 把更多未来信息并入优势计算，信息更多但噪声更大
- ➢ 末端 bootstrap 误差会前传，影响优势估计
- ➢ 缺少 λ做偏差–方差折衷

PPO 的优势估计

TD 残差（Temporal-Difference error）

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

一步回报+bootstrap

广义优势估计的一般形式

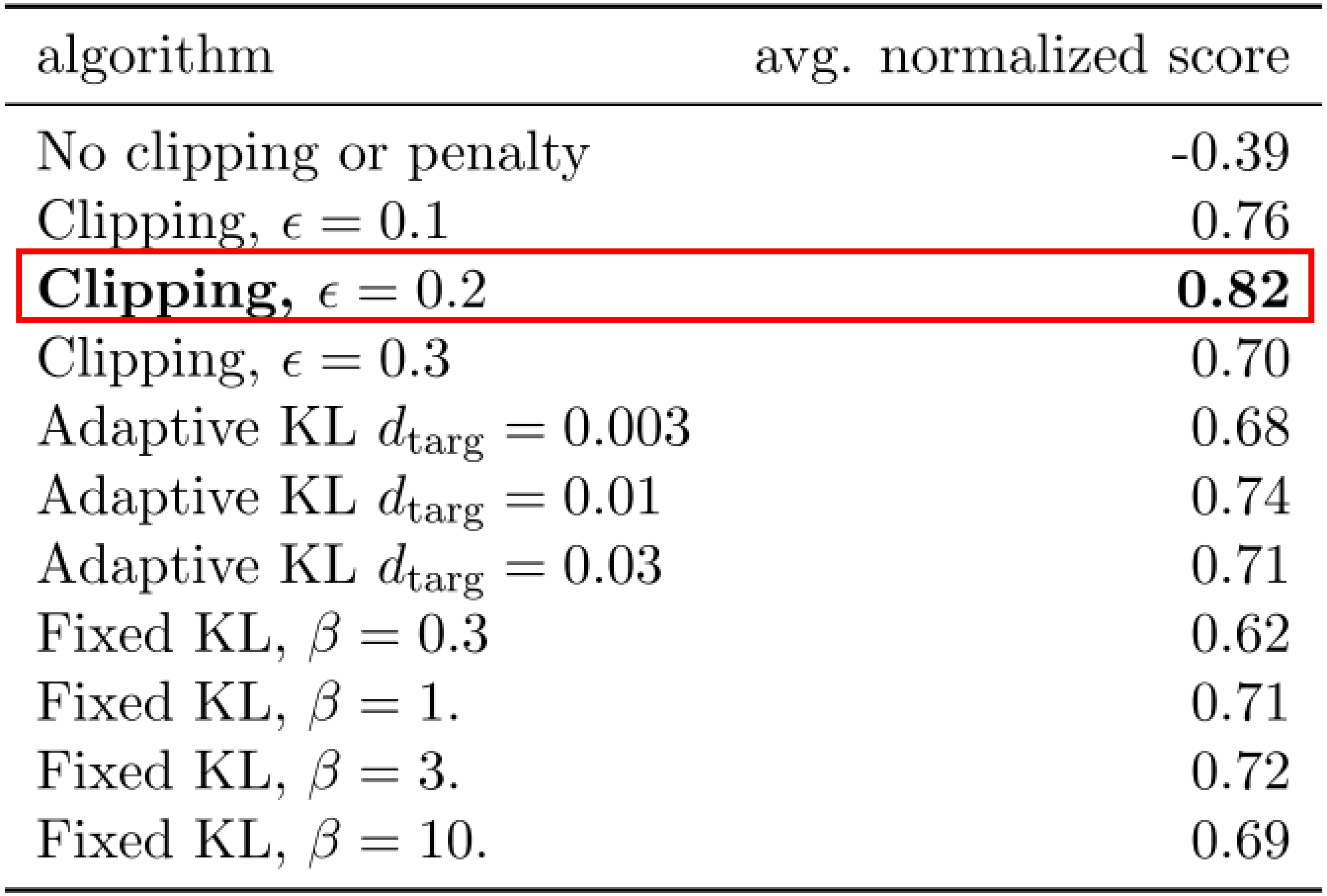
λ 控制偏差- 方差权衡

$$\hat{A}_t = \delta_t + (\gamma\lambda)\delta_{t+1} + \cdots + \cdots + (\gamma\lambda)^{T-t+1}\delta_{T-1}$$

当$\lambda = 1$时

$$\hat{A}_t = -V(s_t) + r_t + \gamma r_{t+1} + \cdots + \gamma^{T-t+1} r_{T-1} + \gamma^{T-t} V(s_T)$$

策略对比



algorithm	avg. normalized score
No clipping or penalty	-0.39
Clipping, $\epsilon = 0.1$	0.76
Clipping, $\epsilon = 0.2$	**0.82**
Clipping, $\epsilon = 0.3$	0.70
Adaptive KL $d_{\text{targ}} = 0.003$	0.68
Adaptive KL $d_{\text{targ}} = 0.01$	0.74
Adaptive KL $d_{\text{targ}} = 0.03$	0.71
Fixed KL, $\beta = 0.3$	0.62
Fixed KL, $\beta = 1.$	0.71
Fixed KL, $\beta = 3.$	0.72
Fixed KL, $\beta = 10.$	0.69

- ➢ 无约束更新易崩溃
- ➢ Clipping 最稳且效果最好
- ➢ KL 惩罚可用但更依赖超参

每个环境：随机策略=0，最佳=1；表中为 7 个环境 × 21 次运行的平均 normalized score