start.

PPO. 2023.10.13.

)): Pre-Request: Policy Graident.

目标：最大化在策略 θ 下的期望回报（$R(\tau)$. return on τ）.

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

故，对其使用 梯度上升 → maximize here.

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

找 θ 变化的方向.

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

$R(\tau)$ 与 θ 无关. $P_\theta(\tau)$ 难以直接得到.

$$= \sum_\tau R(\tau)\cdot P_\theta(\tau)\cdot \frac{\nabla P_\theta(\tau)}{P_\theta(\tau)} \quad (\text{数学变换}).$$

Δtip. 梯度 ∇.

e.g. $\nabla_\theta \log P_\theta(a_t|s_t)$

指向 θ 空间内 (s_t,a_t) 处使 $\log P_\theta(a_t|s_t)$ 增大最快的方向.

(需要回顾∇).

10.13.

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

$$= E_{\tau\sim P_\theta(\tau)}\left(R(\tau)\frac{\nabla P_\theta(\tau)}{P_\theta(\tau)}\right).$$

由 Monte-Carlo → 采样近似.

$$\approx \frac{1}{N}\cdot\sum_{n=1}^{N} R(\tau^n)\cdot\frac{\nabla P_\theta(\tau^n)}{P_\theta(\tau^n)}.$$ 又：$\nabla \log f(x) = \frac{\nabla f(x)}{f(x)}$ ↓ 类似公式.

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

eg: $\tau: \{s_1, a_1, s_2, a_2, \cdots, s_n, a_n, s_{n+1}\}$.

接上.

$$\nabla E_{\tau\sim p_\theta(\tau)}(R(\tau)) \approx \frac{1}{N}\sum_{n=1}^{N} R(\tau^n)\cdot\nabla\log p_\theta(\tau^n) \quad (1).$$

（$p_\theta(\tau^n)$ 仍需处理.）

这里认为，状态转移仅由当前 s, a 决定.

即: $p(s_{t+1}|s_t, a_t)$.（即. 马尔可夫过程）.

此时，$p_\theta(\tau)$ 可表示为: $p_\theta(\tau) = p(s_1)\cdot\prod_{t=1}^{T} p_\theta(a_t|s_t)\cdot p(s_{t+1}|s_t, a_t)$.

$$p_\theta(\tau) = p(s_1)\cdot p_\theta(a_1|s_1)\, p(s_2|s_1, a_1)\cdot p_\theta(a_2|s_2)\cdots p_\theta(a_{t-1}|s_{t-1})\cdot p(s_t|s_{t-1}, a_{t-1})$$

初始 state 和 状态转移 由 Environment 决定 与θ无关

模型θ进行决策的过程.

故有:

$$\nabla_\theta \log p_\theta(\tau) = \nabla_\theta\Big(\log p(s_1) + \sum_{t=1}^{T-1}\log(p_\theta(a_t|s_t)) + \sum_{t=1}^{T}\log(p(s_t|s_{t-1}, a_{t-1}))\Big)$$

（$\log p(s_1)$: θ无关 → 0；$\sum\log p(s_t|s_{t-1},a_{t-1})$: θ无关 → 0）

$$= \sum_{t=1}^{T}\nabla\log p_\theta(a_t|s_t).$$

故. (1) 化为:

$$= \frac{1}{N}\sum_{n=1}^{N} R(\tau^n)\cdot\sum_{t=1}^{T}\nabla\log p_\theta(a_t^n|s_t^n).$$

$$= \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T} R(\tau^n)\,\nabla_\theta\log p_\theta(a_t^n|s_t^n). \quad (2)$$

理解:

当 $R(\tau^n) > 0$,

算法会将 τ^n 内所有 $p_\theta(a_t^n|s_t^n)$ 的概率都加大.

（认为既然 $R(\tau^n) > 0$，则所有动作皆优）.

显然，需优化，

现在的问题是:

直接用 $R(\tau^n)$ 显然不合适.

如何优化?

$\log p_\theta(a_t^n|s_t^n)$: 模型真正干的事: 由 state 做决策.

相较于 $p_\theta(\tau^n)$,

好求了许多.

10-13. end.

接上.