PPO

The PPO loss

$$L_{\text{POLICY}} = \min(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}\hat{A}_t, clip(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}, 1-\epsilon, 1+\epsilon)\hat{A}_t)$$

$$L_{\text{VF}} = \frac{1}{2}\left\| V_{\theta(s)} - \left(\sum_{t=0}^{T} \gamma^t r_t \mid s_0 = s\right)\right\|_2^2$$

$$L_{\text{ENTROPY}} = -\sum_x p(x)\log p(x)$$

$$L_{PPO} = L_{\text{POLICY}} + c_1 L_{\text{VF}} + c_2 L_{\text{ENTROPY}}$$

四个模型：actor, reference, reward, critic

第一个LOSS，actor 的 loss，新老策略比值*优势函数，优势函数需要V

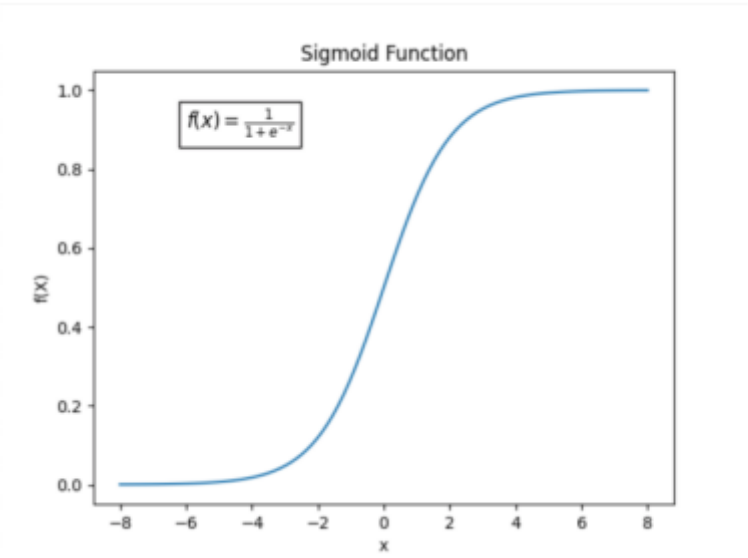
V由critic算

第二个Loss, critic的loss,衡量估计的价值，与真实即时奖励绝对值二范数

第三个Loss，防止集中在某一策略

KL散度隐藏在reward中

训练rewardmodel：正负概率差值取-logsigmod



DPO

推公式把奖励模型隐式消除了

$$\max_{\pi_\theta} \left\{ \mathbb{E}_{(x, y_{\text{win}}, y_{\text{lose}}) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_{\text{win}}|x)}{\pi_{\text{ref}}(y_{\text{win}}|x)} - \beta \log \frac{\pi_\theta(y_{\text{lose}}|x)}{\pi_{\text{ref}}(y_{\text{lose}}|x)} \right) \right] \right\}$$

$$L_{DPO}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim D} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right) \right]$$

https://www.mlpod.com/705.html