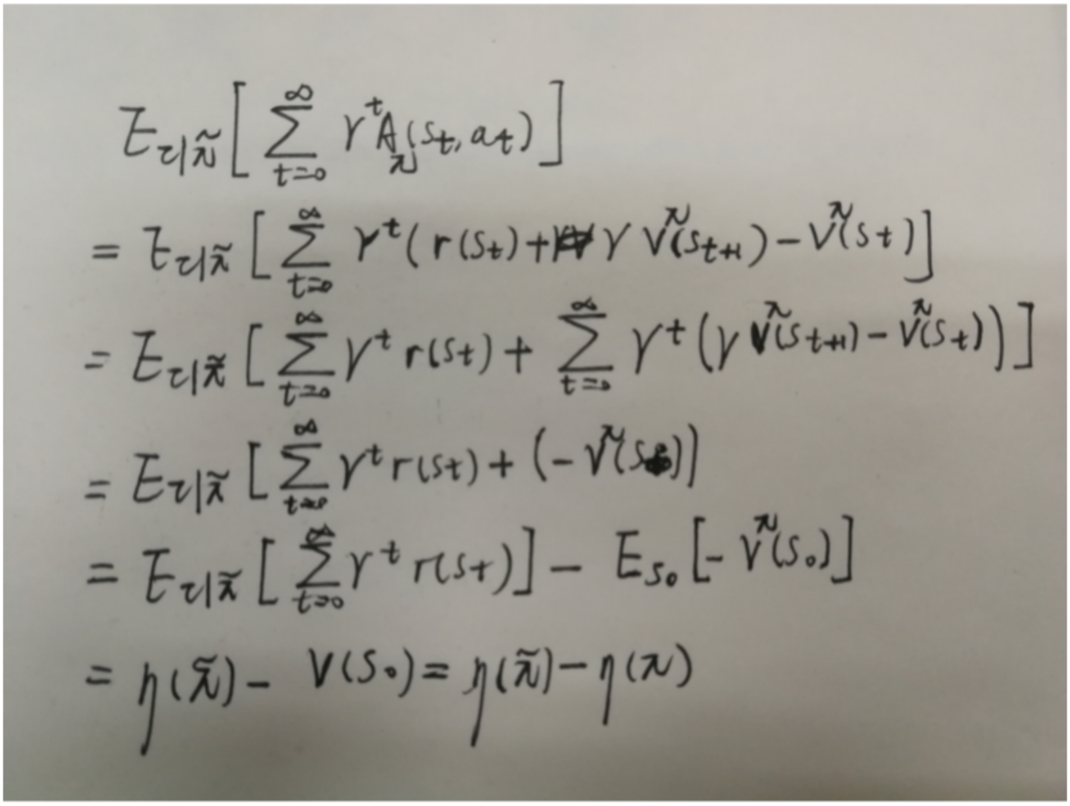
强化学习作业/第8次/个人作业/TRPO三个等式的证明

1 TRPO起点等式的证明

$$\eta(\tilde{\pi}) = \eta(\pi) + E_{\tau|\tilde{\pi}}[\sum_{t=0}^{\infty} \gamma^t (A_{\pi}(s_t, a_t))]$$

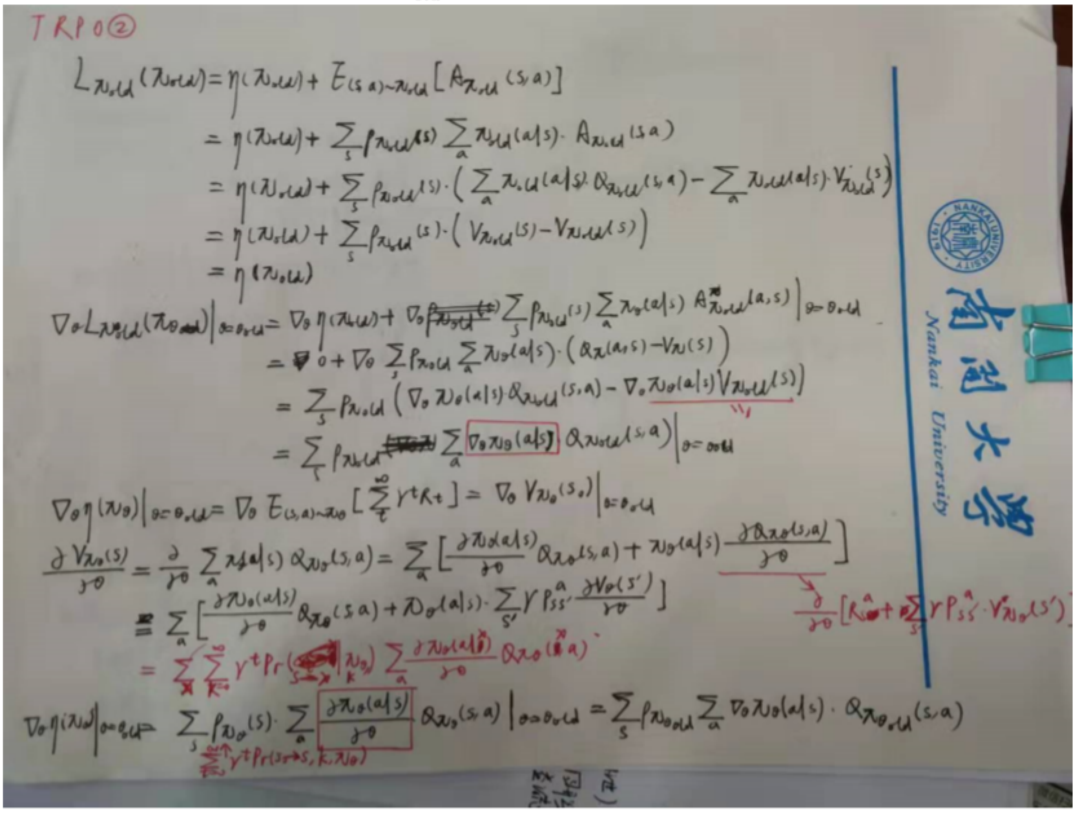
- 证明过程



$$E_{\tau|\tilde{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t A_{\pi}(s_t, a_t)\right]$$
$$= E_{\tau|\tilde{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t (r(s_t) + \gamma V^{\pi}(s_{t+1}) - V^{\pi}(s_t))\right]$$
$$= E_{\tau|\tilde{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t) + \sum_{t=0}^{\infty} \gamma^t (\gamma V^{\pi}(s_{t+1}) - V^{\pi}(s_t))\right]$$
$$= E_{\tau|\tilde{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t) + (-V^{\pi}(s_0))\right]$$
$$= E_{\tau|\tilde{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t)\right] - E_{s_0}\left[-V^{\pi}(s_0)\right]$$
$$= \eta(\tilde{\pi}) - V(s_0) = \eta(\tilde{\pi}) - \eta(\pi)$$

2 替代回报函数和原回报函数在$\theta = \theta_{old}$处相等、一阶近似

- 替代回报函数和原回报函数在策略π_{old}处一阶近似的**证明**



TRPO②

$$L_{\pi_{old}}(\pi_{old}) = \eta(\pi_{old}) + E_{(s,a)\sim\pi_{old}}[A_{\pi_{old}}(s,a)]$$
$$= \eta(\pi_{old}) + \sum_s \rho_{\pi_{old}}(s)\sum_a \pi_{old}(a|s)\cdot A_{\pi_{old}}(s,a)$$
$$= \eta(\pi_{old}) + \sum_s \rho_{\pi_{old}}(s)\cdot\left(\sum_a \pi_{old}(a|s) Q_{\pi_{old}}(s,a) - \sum_a \pi_{old}(a|s) V_{\pi_{old}}(s)\right)$$
$$= \eta(\pi_{old}) + \sum_s \rho_{\pi_{old}}(s)\cdot\left(V_{\pi_{old}}(s) - V_{\pi_{old}}(s)\right)$$
$$= \eta(\pi_{old})$$

$$\nabla_\theta L_{\pi_{old}}(\pi_\theta)|_{\theta=\theta_{old}} = \nabla_\theta \eta(\pi_{old}) + \nabla_\theta \sum_s \rho_{\pi_{old}}(s)\sum_a \pi_\theta(a|s) A_{\pi_{old}}(a,s)|_{\theta=\theta_{old}}$$
$$= 0 + \nabla_\theta \sum_s \rho_{\pi_{old}} \sum_a \pi_\theta(a|s)\cdot\left(Q_\pi(a,s) - V_\pi(s)\right)$$
$$= \sum_s \rho_{\pi_{old}}\left(\nabla_\theta \pi_\theta(a|s) Q_{\pi_{old}}(s,a) - \nabla_\theta \pi_\theta(a|s) V_{\pi_{old}}(s)\right)$$
$$= \sum_s \rho_{\pi_{old}} \sum_a \boxed{\nabla_\theta \pi_\theta(a|s)}\, Q_{\pi_{old}}(s,a)|_{\theta=\theta_{old}}$$

$$\nabla_\theta \eta(\pi_\theta)|_{\theta=\theta_{old}} = \nabla_\theta E_{(s,a)\sim\pi_\theta}\left[\sum_t^\infty \gamma^t R_t\right] = \nabla_\theta V_{\pi_\theta}(s_0)|_{\theta=\theta_{old}}$$

$$\frac{\partial V_{\pi_\theta}(s)}{\partial\theta} = \frac{\partial}{\partial\theta}\sum_a \pi_\theta(a|s) Q_{\pi_\theta}(s,a) = \sum_a\left[\frac{\partial \pi_\theta(a|s)}{\partial\theta} Q_{\pi_\theta}(s,a) + \pi_\theta(a|s)\frac{\partial Q_{\pi_\theta}(s,a)}{\partial\theta}\right]$$

$$\frac{\partial}{\partial\theta}\left[R_s^a + \sum_{s'}\gamma P_{ss'}^a\cdot V_{\pi_\theta}(s')\right]$$

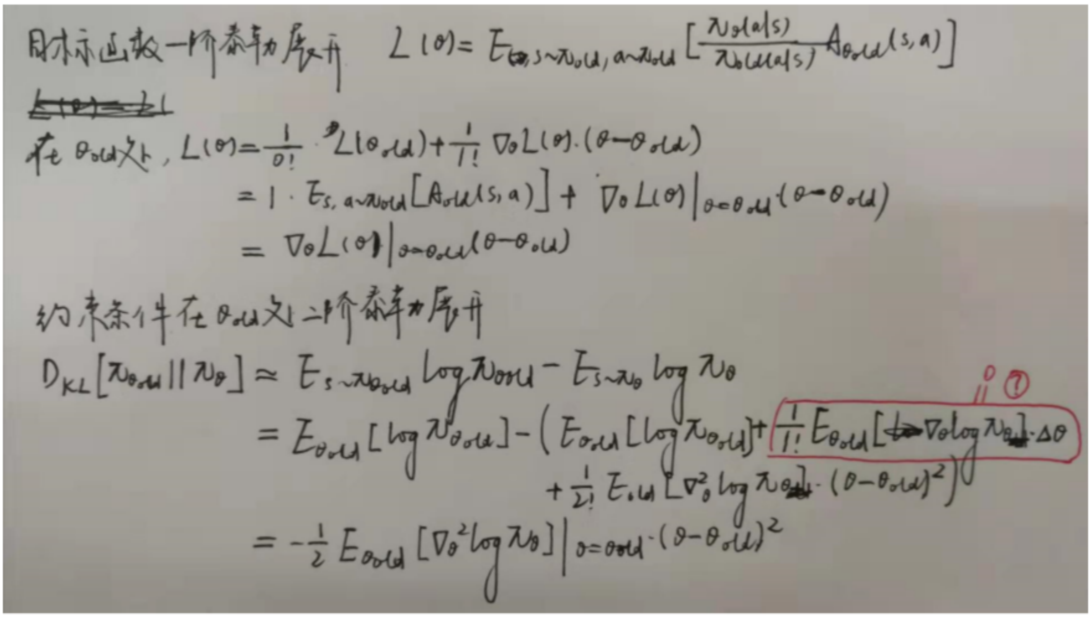
$$= \sum_a\left[\frac{\partial \pi_\theta(a|s)}{\partial\theta} Q_{\pi_\theta}(s,a) + \pi_\theta(a|s)\cdot\sum_{s'}\gamma P_{ss'}^a \frac{\partial V_\theta(s')}{\partial\theta}\right]$$
$$= \sum_x \sum_k^\infty \gamma^k Pr(s\to x, k, \pi_\theta) \sum_a \frac{\partial \pi_\theta(a|x)}{\partial\theta} Q_{\pi_\theta}(x,a)$$

$$\nabla_\theta \eta(\pi_\theta)|_{\theta=\theta_{old}} = \sum_s \underbrace{\rho_{\pi_\theta}(s)}_{\sum_{k=0}^\infty \gamma^k Pr(s_0\to s, k, \pi_\theta)} \sum_a \boxed{\frac{\partial \pi_\theta(a|s)}{\partial\theta}}\, Q_{\pi_\theta}(s,a)|_{\theta=\theta_{old}} = \sum_s \rho_{\pi_{old}} \sum_a \nabla_\theta \pi_\theta(a|s)\cdot Q_{\pi_{old}}(s,a)$$

- 上图中红色一行推导的理解：上一行表示对于$V_{\pi_\theta}(s)$的梯度，可以转化成对策略梯度 * $Q_{\pi_\theta}(s,a)$+所有可能的下一状态的$V_\pi(s')$的梯度 * 这一状态s'出现的概率，因此是一种迭代的计算方法，继续迭代下去，会遍历到所有可能的状态与行为；因此，在红色行中，计算**策略梯度 * 行为值函数**的期望：首先对行为空间进行积分，之后对状态s进行积分；对状态s积分时，概率是从s经过任意步数转变为x的概率的和，因为是要计算准确的值函数梯度，必须要计算到结束。

3 TRPO优化目标的一阶近似、约束条件的二阶近似

- 对TRPO目标函数一阶逼近、对约束条件二阶逼近



目标函数一阶泰勒展开 $L(\theta) = E_{s\sim\pi_{old}, a\sim\pi_{old}}\left[\frac{\pi_\theta(a|s)}{\pi_{\theta_{old}}(a|s)} A_{\theta_{old}}(s,a)\right]$

在 θ_{old} 处，$L(\theta) = \frac{1}{0!} L(\theta_{old}) + \frac{1}{1!}\nabla_\theta L(\theta)\cdot(\theta-\theta_{old})$

$$= 1\cdot E_{s,a\sim\pi_{old}}[A_{old}(s,a)] + \nabla_\theta L(\theta)|_{\theta=\theta_{old}}(\theta-\theta_{old})$$
$$= \nabla_\theta L(\theta)|_{\theta=\theta_{old}}(\theta-\theta_{old})$$

约束条件在 θ_{old} 处二阶泰勒展开

$$D_{KL}[\pi_{old}||\pi_\theta] \simeq E_{s\sim\pi_{old}}\log\pi_{old} - E_{s\sim\pi_\theta}\log\pi_\theta$$
$$= E_{old}[\log\pi_{\theta_{old}}] - \left(E_{old}[\log\pi_{old}] + \underbrace{\frac{1}{1!}E_{old}[\nabla_\theta\log\pi_\theta]\cdot\Delta\theta}_{=0\ ①} + \frac{1}{2!}E_{old}[\nabla_\theta^2\log\pi_\theta]\cdot(\theta-\theta_{old})^2\right)$$
$$= -\frac{1}{2}E_{old}[\nabla_\theta^2\log\pi_\theta]|_{\theta=\theta_{old}}\cdot(\theta-\theta_{old})^2$$

- KL散度非负（非负性证明见教材P155），在$\theta=\theta_{old}$处，KL散度=0，所以θ_{old}是KL散度函数(θ)的一个极值点，因此KL散度函数一阶导数=0