TRPO知识总结

- 随机策略梯度方法的问题：不好选取合适的步长；若步长太长，策略很容易发散；若步长太短，策略收敛速度很慢
- TRPO解决的问题：选取合适的步长，使新的策略的回报函数的值比旧策略大
- 定义策略的回报函数

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

  - τ： $s_0, u_0, \dots, s_H, u_H$
- 将回报函数写成老策略的回报+某一大于零的项

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

  - **证明过程**

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

- 将优势函数的期望展开，写成对分别对状态空间和动作空间的积分

$$\eta(\tilde{\pi}) = \eta(\pi) + \sum_s \rho_{\tilde{\pi}}(s)\sum_a \tilde{\pi}(a|s)A_\pi(s, a)$$

- TRPO技巧1：忽略状态分布的变化，消除状态分布对新策略的依赖 - 对代价函数的近似1

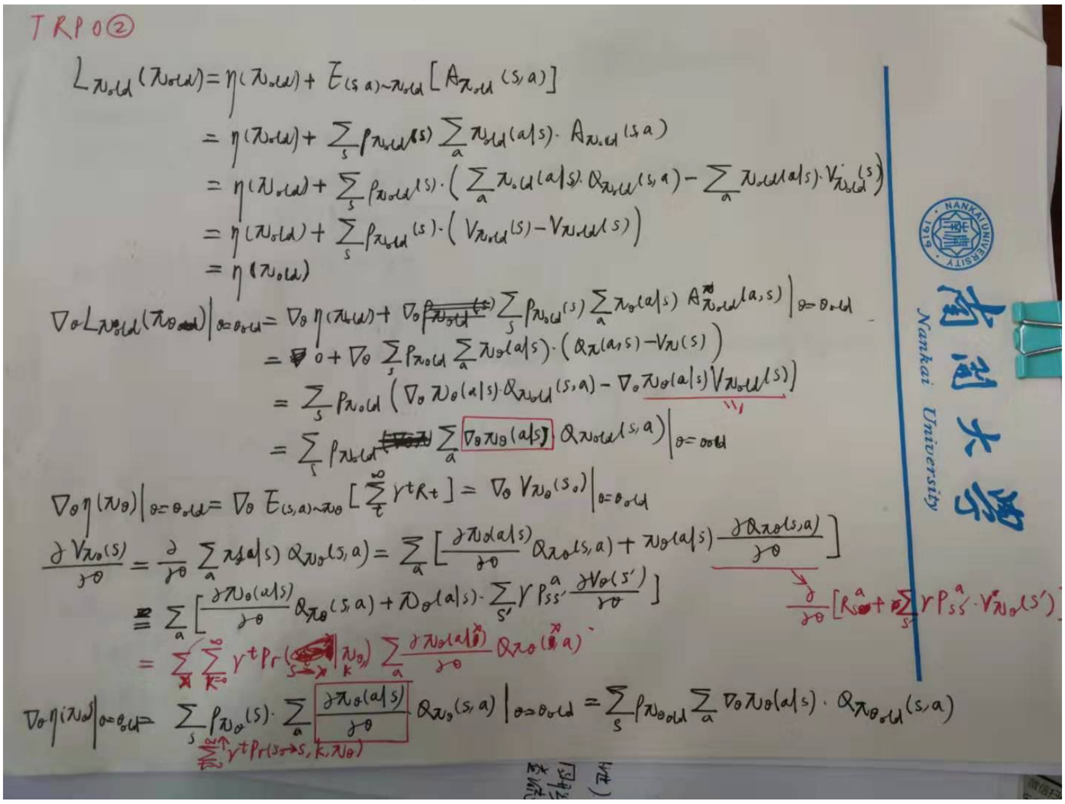
$$L_\pi(\tilde{\pi}) = \eta(\pi) + \sum_s \rho_{\pi}(s)\sum_a \tilde{\pi}(a|s)A_\pi(s, a)$$

- TRPO技巧2：利用重要性采样处理动作分布，克服新的策略（动作分布）未知的问题

  采样策略：旧策略

$$L_\pi(\tilde{\pi}) = \eta(\pi) + E_{(s,a)\sim\pi}[\frac{\tilde{\pi}(a|s)}{\pi(a|s)}A_\pi(s, a)]$$

- 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

$L_{\pi_{old}}(\pi_{old}) = \eta(\pi_{old}) + E_{(s,a)\sim\pi_{old}}[A_{\pi_{old}}(s,a)]$

$= \eta(\pi_{old}) + \sum_s \rho_{\pi_{old}}(s)\sum_a \pi_{old}(a|s)\cdot A_{\pi_{old}}(s,a)$

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

$= \eta(\pi_{old}) + \sum_s \rho_{\pi_{old}}(s)\cdot\left(V_{\pi_{old}}(s) - V_{\pi_{old}}(s)\right)$

$= \eta(\pi_{old})$

$\nabla_\theta L_{\pi_{old}}(\pi_\theta)|_{\theta=\theta_{old}} = \nabla_\theta \eta(\pi_{old}) + \nabla_\theta \sum_s \rho_{\pi_{old}}(s)\sum_a \pi_\theta(a|s) A_{\pi_{old}}(a,s)|_{\theta=\theta_{old}}$

$= 0 + \nabla_\theta \sum_s \rho_{\pi_{old}} \sum_a \pi_\theta(a|s)\cdot(Q_\pi(a,s) - V_\pi(s))$

$= \sum_s \rho_{\pi_{old}}\left(\nabla_\theta \pi_\theta(a|s)\, Q_{\pi_{old}}(s,a) - \nabla_\theta \pi_\theta(a|s)\, V_{\pi_{old}}(s)\right)$

$= \sum_s \rho_{\pi_{old}} \sum_a \boxed{\nabla_\theta \pi_\theta(a|s)}\, Q_{\pi_{old}}(s,a)|_{\theta=\theta_{old}}$

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

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

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

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

$= \sum_x\sum_{k=0}^\infty \gamma^t Pr(s\to x, k, \pi_\theta)\sum_a \frac{\partial \pi_\theta(a|x)}{\partial\theta} Q_{\pi_\theta}(x,a)$

$\nabla_\theta \eta(\pi_\theta)|_{\theta=\theta_{old}} = \sum_s \rho_{\pi_\theta}(s)\cdot\sum_a \boxed{\frac{\partial \pi_\theta(a|s)}{\partial\theta}}\, Q_{\pi_\theta}(s,a)|_{\theta=\theta_{old}} = \sum_s \rho_{\pi_{\theta_{old}}}\sum_a \nabla_\theta\pi_\theta(a|s)\cdot Q_{\pi_{\theta_{old}}}(s,a)$

$\sum_{k=0}^\infty \gamma^t Pr(s_0\to s, k, \pi_\theta)$

- 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。
- 单调性的证明：书上不等式（8.8）
- 为决定更新步长，问题转化为

$$\underset{\theta}{maximize}\, E_{s\sim\rho_{\theta_{old}},\, a\sim\pi_{\theta_{old}}}\left[\frac{\pi_\theta(a|s)}{\pi_{\theta_{old}}(a|s)}A_{\theta_{old}}(s,a)\right]$$

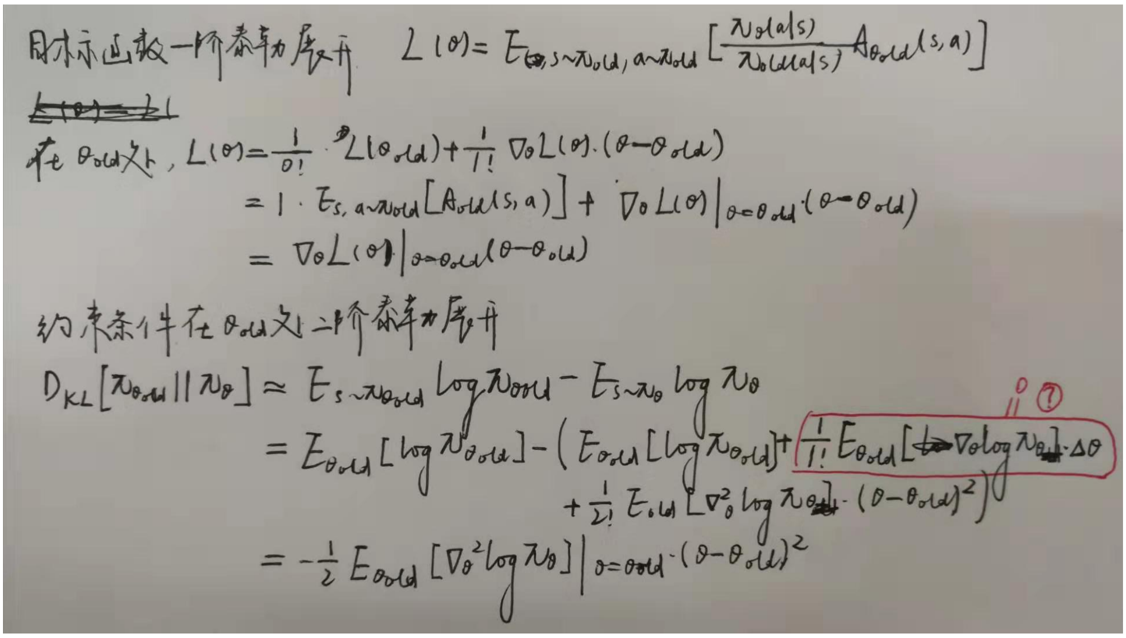
$$subject\ to\quad D_{KL}^{\max}(\theta_{old},\theta) \leq \delta$$

- TRPO第三个技巧：利用平均KL散度代替最大KL散度
  - 最大：遍历所有s，对每个s都有一个KL散度，找最大的那个
  - 平均：s分布服从ρ，对s的分布求期望，概率是$\rho(s)$，值是这个s下的KL散度
- TRPO第四个技巧：$s \sim \rho_{\theta_{old}} \to s \sim \pi_{\theta_{old}}$
  - 理解：忽略折扣因子？
- 因此，平均KL散度的期望中，s服从的分布也变为$\pi_{\theta_{old}}$
  - 约束条件变为

$$E_{s\sim\pi_{old}}[D_{KL}(\pi_{old}||\pi)]$$

    - 采样，利用样本均值代替期望 - ppt44

- 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 $L(\theta) = E_{s\sim\pi_{old}, a\sim\pi_{old}}\left[\frac{\pi_\theta(a|s)}{\pi_{old}(a|s)} A_{\theta_{old}}(s,a)\right]$

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

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

$= \nabla_\theta L(\theta)\big|_{\theta=\theta_{old}} (\theta - \theta_{old})$

约束条件在 θ_{old} 处二阶泰勒展开

$D_{KL}[\pi_{\theta_{old}} \| \pi_\theta] \simeq E_{s\sim\pi_{\theta_{old}}} \log \pi_{\theta_{old}} - E_{s\sim\pi_\theta} \log \pi_\theta$

$= E_{\theta_{old}}[\log \pi_{\theta_{old}}] - \Big(E_{\theta_{old}}[\log \pi_{\theta_{old}}] + \frac{1}{1!} E_{\theta_{old}}[\nabla_\theta \log \pi_\theta] \cdot \Delta\theta$ ①

$+ \frac{1}{2!} E_{old}[\nabla_\theta^2 \log \pi_\theta] \cdot (\theta - \theta_{old})^2\Big)$

$= -\frac{1}{2} E_{\theta_{old}}[\nabla_\theta^2 \log \pi_\theta]\big|_{\theta=\theta_{old}} \cdot (\theta - \theta_{old})^2$

  - 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