## 强化学习个人作业 - AC - 1911475 王禹

## 证明1: 随机策略AC方法相容条件的证明

AC算法中Critic策略利用函数逼近的方法估计值函数。随机策略的梯度为

$$abla_{ heta}J( heta) = E_{s\sim
ho^\pi,a\sim\pi_ heta}[
abla\log\pi_ heta(a|s)Q^\pi(s,a)]$$

其中, $Q^\pi(s,a)$ 表示策略 $\pi$ 下真实的行为值函数。在AC中,用来逼近的值函数表示为 $Q^w(s,a)$ ,w为待逼近的参数。根据相容条件:

$$egin{aligned} Q^w(s,a) &= 
abla_ heta \log \pi_ heta(a|s)^T w \ w^* &= rg \min_w E_{(s,a) \sim \pi}[(Q^w(s,a) - Q^\pi(s,a))^2] \end{aligned}$$

可以保证 $Q^w(s,a)$ 无偏差于 $Q^{\pi}(s,a)$ 

证明:

相容条件 (2) 表明, w取 $w^*$ 时,  $[(Q^w(s,a)-Q^\pi(s,a))^2]$ 为极小值, 因此对此式求w的偏导为0:

$$rac{\partial Q^w(s,a)}{\partial w}[Q^w(s,a)-Q^\pi(s,a)]=0$$

带入相容条件(1),有

$$abla_{ heta} \log \pi_{ heta}(a|s)^T [Q^w(s,a) - Q^{\pi}(s,a)] = 0$$

因此,有

$$abla_{ heta} \log \pi_{ heta}(a|s)^T Q^w(s,a) = 
abla_{ heta} \log \pi_{ heta}(a|s)^T Q^{\pi}(s,a)$$

即 $Q^w(s,a)$ 无偏差于 $Q^{\pi}(s,a)$ 。

## 证明2: 确定性策略AC方法相容条件的证明

确定性策略梯度为

$$abla_{ heta}J(\mu_{ heta}) = E_{s\sim
ho^{\mu}}[
abla_{ heta}\mu_{ heta}(s)
abla_{a}Q^{\mu}(s,a)|_{a=\mu_{a}(s)}]$$

用 $Q^w(s,a)$ 无偏差地逼近 $Q^\pi(s,a)$ ,相容条件为

$$egin{aligned} \left. 
abla_a Q^w(s,a) 
ight|_{a=\mu_{ heta}(s)} &= 
abla_{ heta} \mu_{ heta}(s)^T w \ & w^* = rg \min_w E[\epsilon(s; heta,w)^T \epsilon(s; heta,w)] \ & \epsilon(s; heta,w) &= 
abla_a Q^w(s,a) 
vert_{a=\mu_{ heta}(s)} &- 
abla_a Q^\mu(s,a) 
vert_{a=\mu_{ heta}(s)} \end{aligned}$$

证明:

仍然对相容条件(2) 求w的偏导数,得到

$$[
abla_a Q^w(s,a)|_{a=\mu_{ heta}(s)} - 
abla_a Q^\mu(s,a)|_{a=\mu_{ heta}(s)}] rac{\partial 
abla_a Q^w(s,a)}{\partial w} = 0$$

带入相容条件(1),有

$$[
abla_a Q^w(s,a)|_{a=\mu_{ heta}(s)} - 
abla_a Q^\mu(s,a)|_{a=\mu_{ heta}(s)}]
abla_ heta \mu_{ heta}(s) = 0$$

$$\left. 
abla_a Q^w(s,a) 
ight|_{a=\mu_ heta(s)} 
abla_ heta \mu_ heta(s) = 
abla_a Q^\mu(s,a) 
ight|_{a=\mu_ heta(s)} 
abla_ heta \mu_ heta(s)$$