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

$$\nabla_\theta J(\theta)=E_{s\sim\rho^\pi,a\sim\pi_\theta}[\nabla\log\pi_\theta(a|s)Q^\pi(s,a)]$$

其中，$Q^\pi(s,a)$表示策略π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:

$$Q^w(s,a)=\nabla_\theta\log\pi_\theta(a|s)^Tw$$
$$w^*=\arg\min_w E_{(s,a)\sim\pi}[(Q^w(s,a)-Q^\pi(s,a))^2]$$

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：

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

带入相容条件（1），有

$$\nabla_\theta\log\pi_\theta(a|s)^T[Q^w(s,a)-Q^\pi(s,a)]=0$$

因此，有

$$\nabla_\theta\log\pi_\theta(a|s)^TQ^w(s,a)=\nabla_\theta\log\pi_\theta(a|s)^TQ^\pi(s,a)$$

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

$$\nabla_\theta J(\mu_\theta)=E_{s\sim\rho^\mu}[\nabla_\theta\mu_\theta(s)\nabla_aQ^\mu(s,a)|_{a=\mu_\theta(s)}]$$

用$Q^w(s,a)$无偏差地逼近$Q^\pi(s,a)$，相容条件为

$$\nabla_aQ^w(s,a)|_{a=\mu_\theta(s)}=\nabla_\theta\mu_\theta(s)^Tw$$
$$w^*=\arg\min_w E[\epsilon(s;\theta,w)^T\epsilon(s;\theta,w)]$$
$$\epsilon(s;\theta,w)=\nabla_aQ^w(s,a)|_{a=\mu_\theta(s)}-\nabla_aQ^\mu(s,a)|_{a=\mu_\theta(s)}$$

证明:

仍然对相容条件（2）求w的偏导数，得到

$$[\nabla_aQ^w(s,a)|_{a=\mu_\theta(s)}-\nabla_aQ^\mu(s,a)|_{a=\mu_\theta(s)}]\frac{\partial\nabla_aQ^w(s,a)}{\partial w}=0$$

带入相容条件（1），有

$$[\nabla_aQ^w(s,a)|_{a=\mu_\theta(s)}-\nabla_aQ^\mu(s,a)|_{a=\mu_\theta(s)}]\nabla_\theta\mu_\theta(s)=0$$

因此，有

$$\nabla_a Q^w(s,a)|_{a=\mu_\theta(s)} \nabla_\theta \mu_\theta(s) = \nabla_a Q^\mu(s,a)|_{a=\mu_\theta(s)} \nabla_\theta \mu_\theta(s)$$