误差函数梯度

DQN误差函数

经验池

行为策略网络

目标策略网络

环境

(s,a,r,s\_)四元组

最大当前Q值

当前状态s

下一个状态的最大Q值

奖励r

当前Q值

(s,a)

按一定学习间隔更新目标

下一个状态s’

经验池

目标策略: 生成若干个估计值Q

行为策略：生成若干个估计值Q

损失函数

L=(r+γmaxQ(a’)-Q(a\_))²