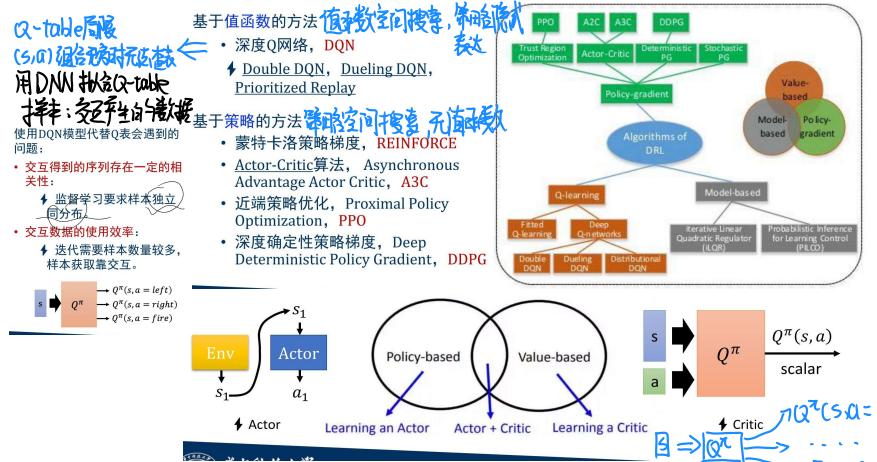
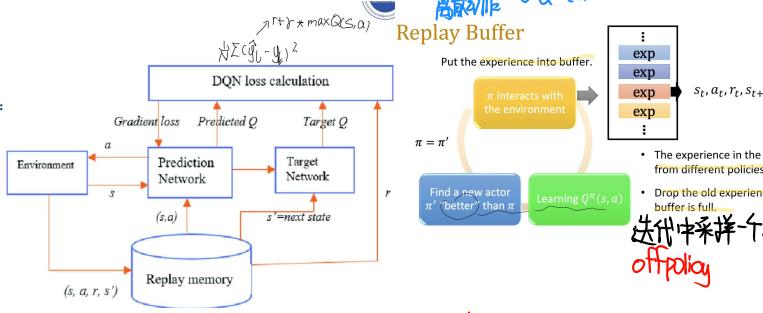


DRL  $\leftarrow$  DL  $\Rightarrow$  表达能力  $\Rightarrow$  对策略函数与值函数建模  
DRL  $\nwarrow$  RL  $\Rightarrow$  决策能力  $\Rightarrow$  定义问题与优化目标

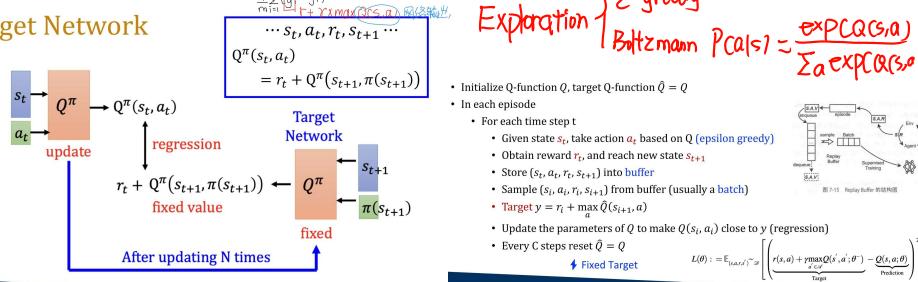


## Nature DQN

- 问题**
  - 交互得到的序列存在一定的相关性: 监督学习要求样本独立同分布。
  - 交互数据的使用效率: 迭代需要样本数量较多, 样本获取靠交互。
- 特点:**
  - 经验回放 (Experience replay)
  - 固定 target Q 值



## Target Network



## Double DQN

- 用当前Q网络计算最大Q值对应的动作, 用目标Q网络计算这个最大动作对应的目标Q值, 进而消除贪婪法带来的偏差。

## Prioritized Replay DQN

- 对DQN的经验回放池按权重采样
- 根据每个样本的TD误差绝对值  $|\delta(t)|$ , 给定该样本的优先级正比于  $|\delta(t)|$ , 将这个优先级的值存入经验回放池

## Dueling DQN

- 通过优化神经网络的结构来优化算法, 把网络输出: 状态动作值函数  $Q(s, a)$ , 分为优势函数  $A(s, a)$  和状态值函数  $V(s)$

## 基于值函数强化学习方法的共同问题:

- 无法表示随机策略
- DQN在实现时采用了贪婪策略, 无法实现按照概率执行各种候选动作的要求。
- 无法表示连续动作。DQN要求动作空间是离散的, 且只能是有限个:
  - 某些问题中, 动作是连续的, 例如要控制在x y z方向的速度、加速度, 这些值显然是连续的。
- 对受限状态下的问题处理能力不足
  - 真实环境下不同的状态由相同的特征表达。
- DQN输出值 (各个动作的Q值) 的微小改变会导致某一动作被选中或不选中, 这种不连续的变化会影响算法的收敛

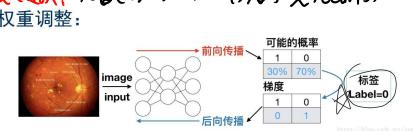
贪婪策略的缺陷

## 基于策略梯度直接更新策略梯度

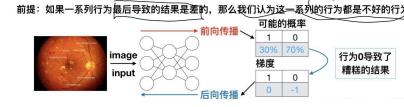
连续动作, 策略随机

策略函数形式:  $\pi_\theta(s, a) = \text{PCA}(s, \theta) \propto \pi_\text{CA}(s)$

优化

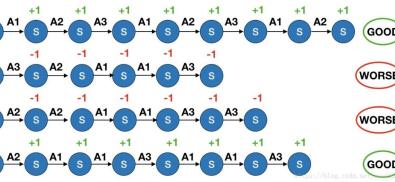


## 策略梯度强化学习的权重调整



## 以游戏为例

- 所有获胜对局中的动作都认为是好的  $\rightarrow$  正向更新
- 所有失败对局中的动作都认为是不好的  $\rightarrow$  负向更新



## 策略梯度定理

策略梯度定理: 对于任意MDP, 不论是优化平均奖励还是初始状态奖励, 目标对参数  $\theta$  求梯度的形式都可以表示为:

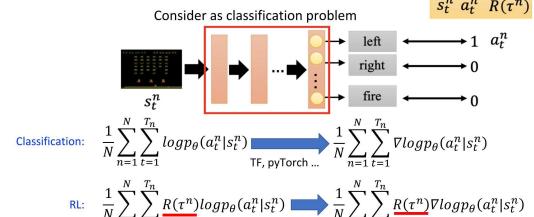
$$\nabla_\theta J(\theta) = E_{s, \pi_\theta} [\nabla_\theta \log \pi_\theta(s, a) R_\pi(s, a)]$$

分值函数 score function	累积收益
---------------------	------

## 可优化的函数目标:

- 优化初始状态收获的期望:  $J_1(\theta) = V_{\pi_\theta}(s_1) = E_{\pi_\theta}(G_1)$
- 优化平均价值:  $J_{avg}(\theta) = \sum_s d_{\pi_\theta}(s) V_{\pi_\theta}(s)$
- 优化每一时间步的平均奖励:  $J_{avg}(\theta) = \sum_s d_{\pi_\theta}(s) \sum_a \pi_\theta(s, a) R_\pi^{sa}$

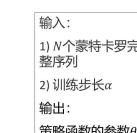
## 分值函数



## 估计累积收益 $R_\pi(s, a)$

### 蒙特卡洛策略梯度算法 REINFORCE

- 使用价值函数  $v(s)$  近似代替  $G_\pi(s, a)$
- 蒙特卡洛方法估计  $v(s)$

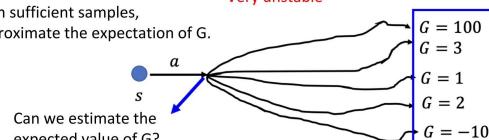


## 蒙特卡洛策略梯度的局限

$$\nabla \bar{R}_\theta \approx \frac{1}{N} \sum_{n=1}^N \sum_{t=1}^{T_n} \left( \sum_{t'=t}^{T_n} \gamma^{t'-t} r_{t'}^n - b \right) \nabla \log p_\theta(a_t^n | s_t^n)$$

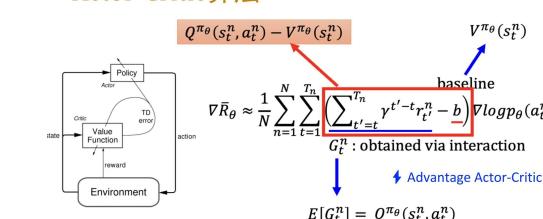
baseline  
 $G_t^n$ : obtained via interaction  
Very unstable

With sufficient samples, approximate the expectation of G.



## 结合两者

## Actor-Critic 算法



### 基于状态价值 (蒙特卡洛策略梯度)

$$\theta \leftarrow \theta + \alpha \nabla_\theta \log \pi_\theta(s, a) V(s, \omega)$$

### 基于动作价值 (DQN)

$$\theta \leftarrow \theta + \alpha \nabla_\theta \log \pi_\theta(s, a) Q(s, a, \omega)$$

### 基于时序差分误差 (时序差分学习)

$$\theta \leftarrow \theta + \alpha \nabla_\theta \log \pi_\theta(s, a) \delta(t)$$

### 基于TD( $\lambda$ )误差 (时序差分学习)

$$\theta \leftarrow \theta + \alpha \nabla_\theta \log \pi_\theta(s, a) \delta(t) E_r(t)$$

### 基于优势函数 (Dueling DQN)

$$\theta \leftarrow \theta + \alpha \nabla_\theta \log \pi_\theta(s, a) A(S, A, \omega, \beta)$$