



- 时序差分法 Temporal-Difference 单步更新 Policy
  - 在线策略 On-Policy 将每次采样的数据存入经验池，使用经验回放更新 Policy
  - 离线策略 Off-Policy 使用在当前 Policy 下采样的数据进行学习并更新 Policy
    - Sarsa On-Policy
    - $\epsilon$ -学习  $\epsilon$ -Learning off-Policy
- 多步自举法 n-step Temporal-Difference
  - TD 和 MC 都有中立法 (MC 高 Variance 低 Bias, TD 低 Variance 高 Bias)
  - n 步 sarsa

## 深度强化学习 Deep Reinforcement Learning

### 基于值的方法 (Value-Based)

- DQN 引入 NN 代替 Q-learning 中的 Q(s,a) 表格
- Double DQN 两个双网络解决 DQN 中对 Q 值估计过高问题
- Duel DQN 将 DQN 的状态价值函数和动作优势函数分别建模

### 基于策略的方法 (Policy-Based.)

- 策略梯度 Policy Gradient

→ REINFORCE

### 行动器-评判器 Actor-Critic

Actor 是 Policy Based, Critic 是 Value Based

- Actor-Critic Actor 学习策略  $a = \pi(s)$ , Critic 学习价值函数  $V^\pi(s)$  或  $\hat{Q}^\pi(s,a)$
- A2C & A3C A2C 引入 Advantage, A3C 引入 Asynchronous
- TRPO 在 Policy Gradient 上加入了 Trust Region，在此区域内做策略优化，更加安全
- PPO 改进版 TRPO，引入 PPO-Penalty, PPO-Clip 两种方式，使求解更简单有效
- DDPG
- TD3
- SAC

## 强化学习前沿

- 多智能体 Multi-Agent RL
  - 去中心化
    - IMPALA
  - 中心化
    - QMIX
    - MADDPG
- 分层强化学习 HRL
- 离线强化学习 Offline RL
- 模仿学习 Imitation Learning
  - 行为克隆 Behavior Cloning
  - 逆强化学习 Inverse RL
  - 生成对抗模仿学习 Generative Adversarial IL
- 模型预测控制 MPC
  - PETR
- 基于模型的策略优化
  - MBPO