

### 1. 请写出强化学习model-base和model-free方法在思想上的差异

强化学习较之于 MDP，就少了转移概率和奖励，那么想办法把转移概率和奖励计算出来，问题得解。

model-base 方法：

使用蒙特卡洛方法，以出现频率代替概率。所谓模型，就是所有的转移概率和奖励构成的集合。

model-free 方法：

model-free 方法直接从数据或经验中学习策略或价值函数，不依赖于环境模型的构建。

仅从数据中来估计  $u_t$ ：

$$u_t = r_t + \lambda r_{t+1} + \lambda^2 r_{t+2} + \dots,$$

$$Q_\pi(s, a) = (1 - \xi)Q_\pi(s, a) + \xi u_t$$

从以前的积累  $Q_\pi(s', a')$  和新数据中估计  $u_t$ ：

自助法  $\implies$  SARSA 算法

$$u_t = r_t + \lambda Q_\pi(s', a'),$$

$$Q_\pi(s, a) = (1 - \xi)Q_\pi(s, a) + \xi(r_t + \lambda Q_\pi(s', a'))$$

### 2. 强化学习和MDP的差异

强化学习与 MDP 的差异在于已知条件：

强化学习，已知样本数据序列，可能不止一个序列；

MDP，已知转移概率和奖励的全部信息。

强化学习和 MDP 所要求目标是一致的。

### 3. 长期回报 $u_t$ 和单步奖励 $r_t$ 之间的关系。强化学习追求谁的最大化

$$u_t = r_t + \lambda r_{t+1} + \lambda^2 r_{t+2} + \dots$$

强化学习追求的是长期回报  $u_t$  的最大化