SANTA

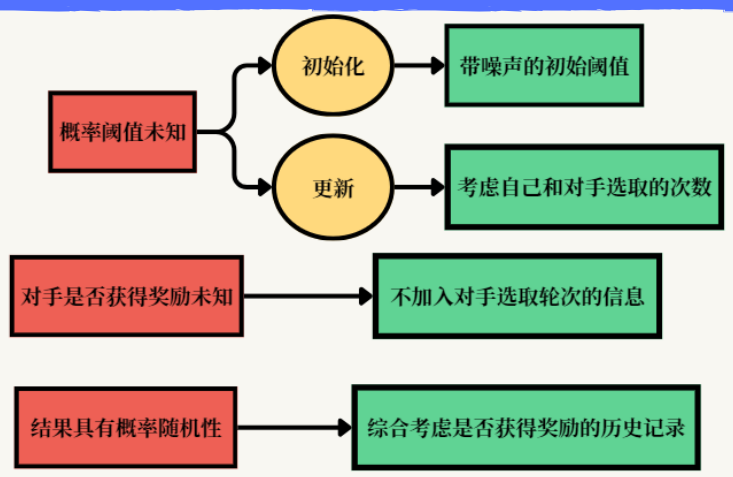
基于强化学习的多臂老虎机问题解决方案

郑翔 吴嘉骜 程子豪

问题背景与挑战

- **背景：** 强化学习的经典问题——多臂老虎机
- **挑战一：** 各售货机提供糖果的概率阈值未知
- **挑战二：** 对手是否获得奖励的情况未知
- **挑战三：** 选择某一售货机时得到的结果具有概率随机性
- **挑战四：** 游戏过程中平衡探索与利用

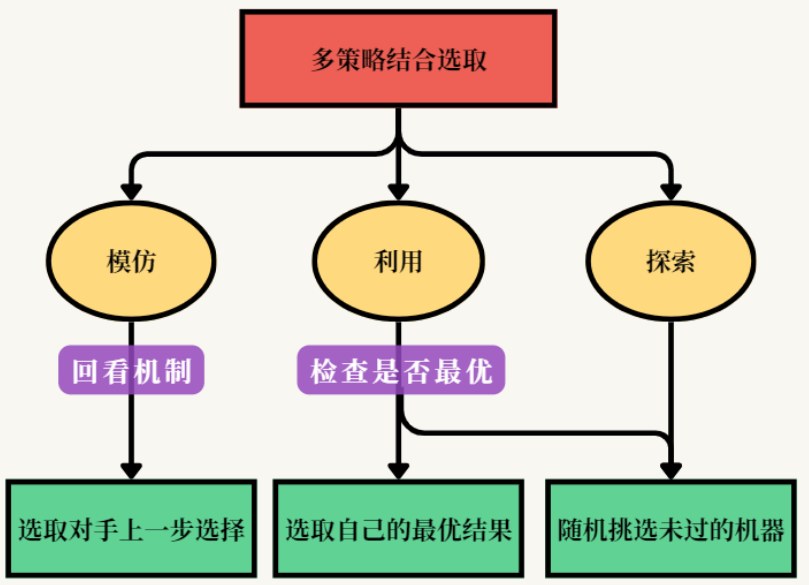
概率的估计和更新



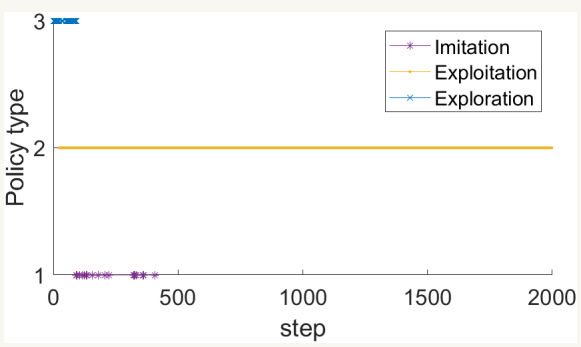
- **阈值的初始化**
  - 带随机均匀噪声的初始阈值—— 0.5 + random(0, 1) * 0.001
  - 利用高斯分布初始化阈值—— N(0.5, 0.005)
- **阈值的更新**
  - 考虑自己和对手在该售货机上总共选取的次数
  - 考虑自己选取该售货机是否得到奖励的历史记录
  - 由于对手是否获得奖励情况未知，因此不加入对手选取轮次的信息

多策略结合选取

- **模仿**
  - 如果对手上一轮次选取的选取的售货机选取次数大于1，自己选取的次数小于2，则选取该售货机
  - 回看对手前60步选取，若选取上一轮次的售货机次数大于1，自己选取的总次数小于5，则进行选取
- **利用**
  - 先判断自己的“最优”选择是否已经达到最优，若是，则继续
    - 对手的选取在自己的最优选择中，返回对手的上一轮此选择
    - 否，且存在还未选取的售货机，则进行探索
  - 否，则在自己的“最优”选择中选取最远一次选择的售货机
- **探索**
  - 随机挑选未选择过的售货机



实验结果及创新



version	characteristic	score
agent-test	UCB	574
agent-test-2	ϵ-greedy	1071.3
agent-2	traditional RL	1255.6
agent-5	without look-back	1343.2
agent-7	RL with imitation	2824.1 rank 2!

- 多策略结合利用游戏不同阶段的特点
  - 前100步智能体进行探索
  - 第100-400步智能体倾向借鉴对手选择
  - 第400步之后智能体基本根据自己对概率的估计进行选择
- 加入检查机制避免较弱智能体误导
- 利用对手的选取情况修正智能体