强化学习奖励, 探索利用困境,观测与状态, 动作空间

奖励：标量、反馈信号

强化学习目标：最大化奖励

输？赢？

单步棋没有奖励

每往前多走一米就有奖励

分数就是奖励

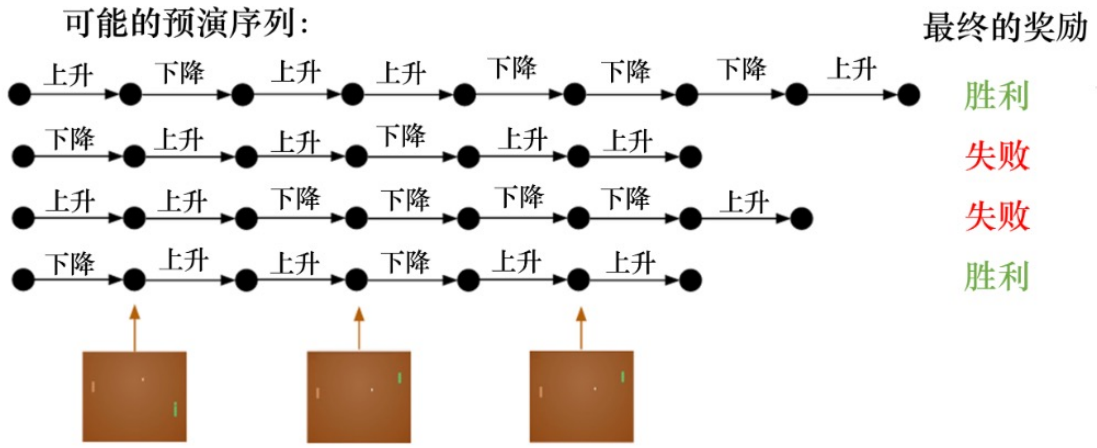


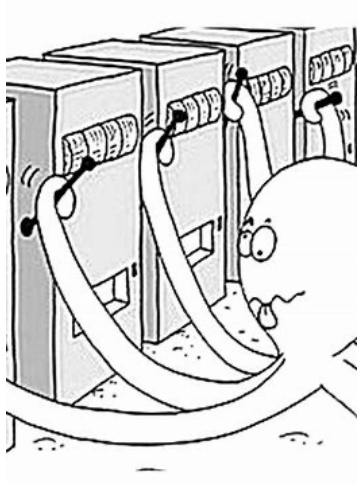
图 1.8 可能的预演序列

单步动作并不产生奖励，最终才会给出奖励（延迟奖励）

怎么办？

（提示：我们可以从一个状态出发，第一步采取向上，后续动作随机。重复10000轮检查输赢情况。再在第一步采取向下，重复10000轮。比较两个动作的输赢差异。）

平衡近期、远期奖励：K臂赌博机



摇臂	1	2	3	4	5	6
奖励	100	30	0	20	5	45

仅探索的期望奖励：（100+30+0+20+5+45）/6*n

仅利用的期望奖励：取决于赌徒知道什么

如果幸运知道1对应100，可以最大。但如果仅知道2和3，他就会不停使用2，从而得不到最大奖励。

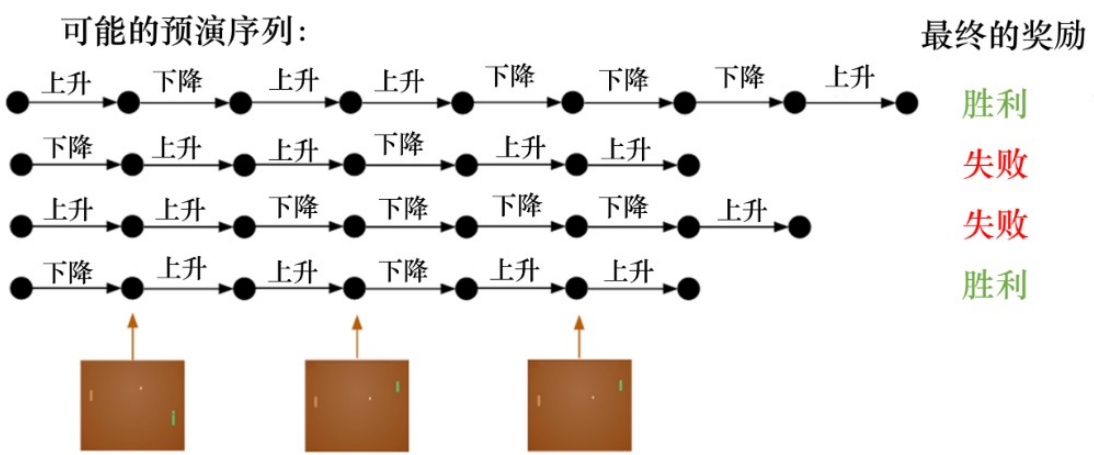


图 1.8 可能的预演序列

只看到图像
不知道内部状态
部分可观测

$\tau = (s_0, a_0, s_1, a_1, \ldots)$ **：轨迹**

$H_t = o_1, a_1, r_1, \ldots, o_t, a_t, r_t$ **：历史**

o和s有什么区别？（观测和状态有什么区别）

状态是对世界的完整描述，不会隐藏世界的信息。**观测**是对状态的部分描述，可能会遗漏一些信息。

完全可观测（fully observed）：马尔可夫决策过程（Markov decision process，MDP）

部分可观测（partially observed）：**部分可观测马尔可夫决策过程（partially observable Markov decision process）**

强化学习讨论的问题是智能体怎么在复杂、不确定的环境中最大化它能获得的奖励。

这一帧：
向上移动期望奖励：100 √
向下移动期望奖励：60

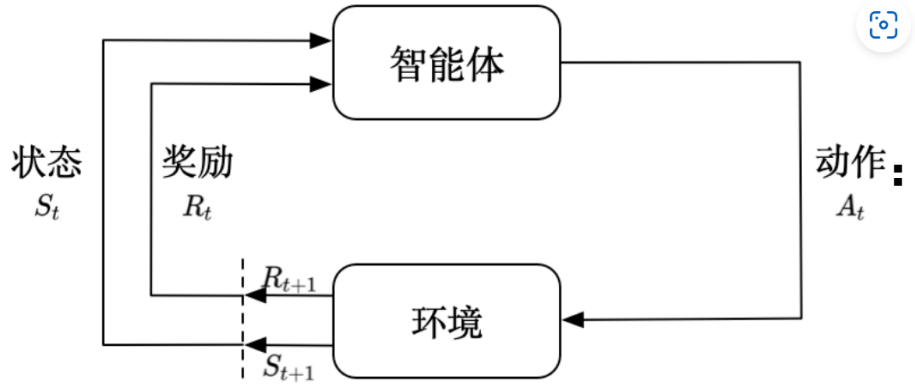


图 1.1 强化学习示意

：决策

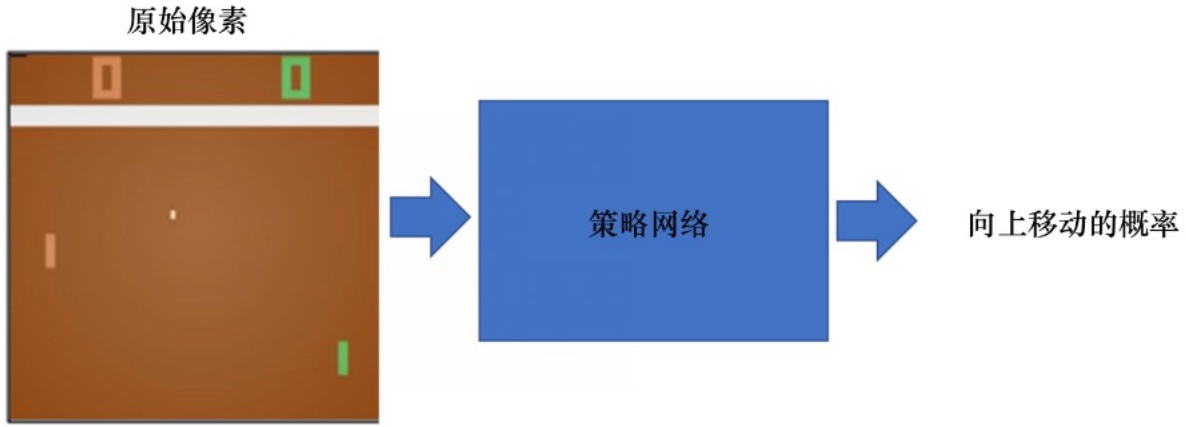
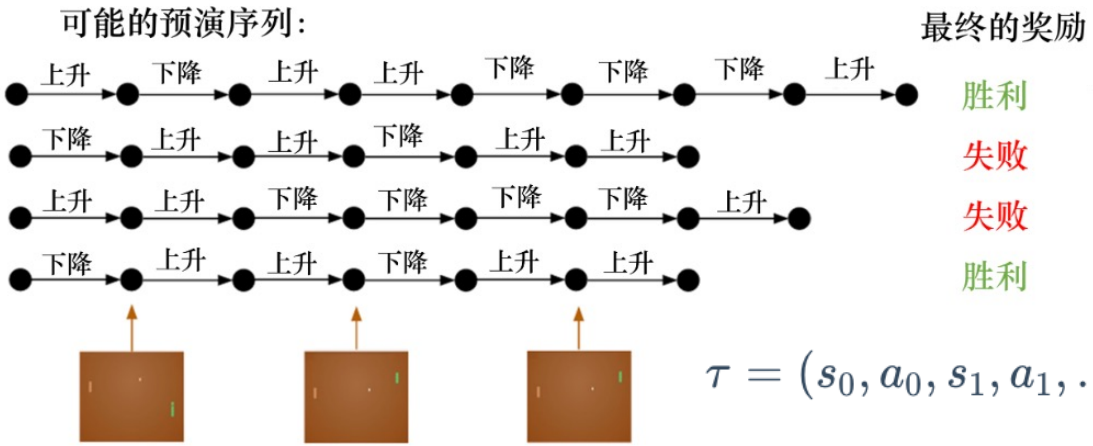


图 1.6 强化学习玩 Pong



$\tau = (s_0, a_0, s_1, a_1, \ldots)$ **：轨迹**

图 1.8 可能的预演序列

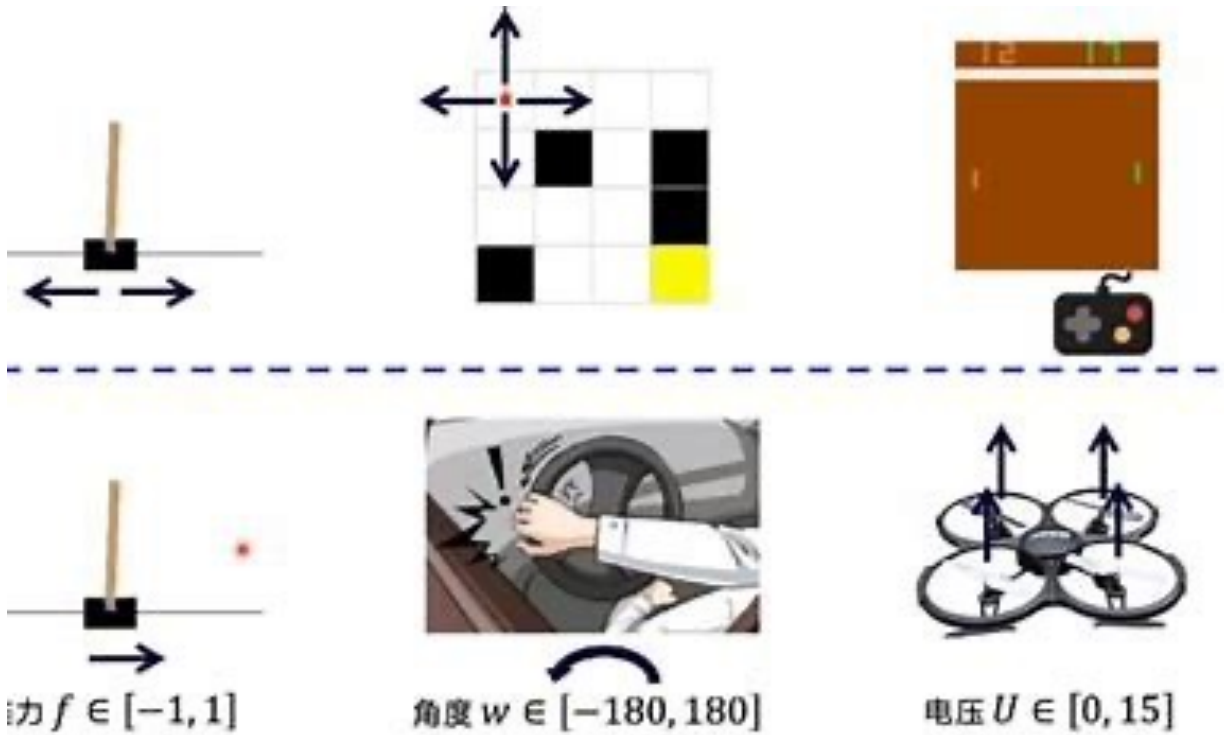
强化学习的目标

输入：观测/状态

输出：当前采取动作

学习方法：利用轨迹与奖励

动作空间



离散

连续

本章小结

- 什么是奖励？
- 什么是k-臂赌博机问题？
- 什么是探索利用困境？
- 观测和状态有什么区别？
- 动作空间有哪两种？

下一章：强化学习智能体的组成成分和类型

Credit goes to: EasyRL