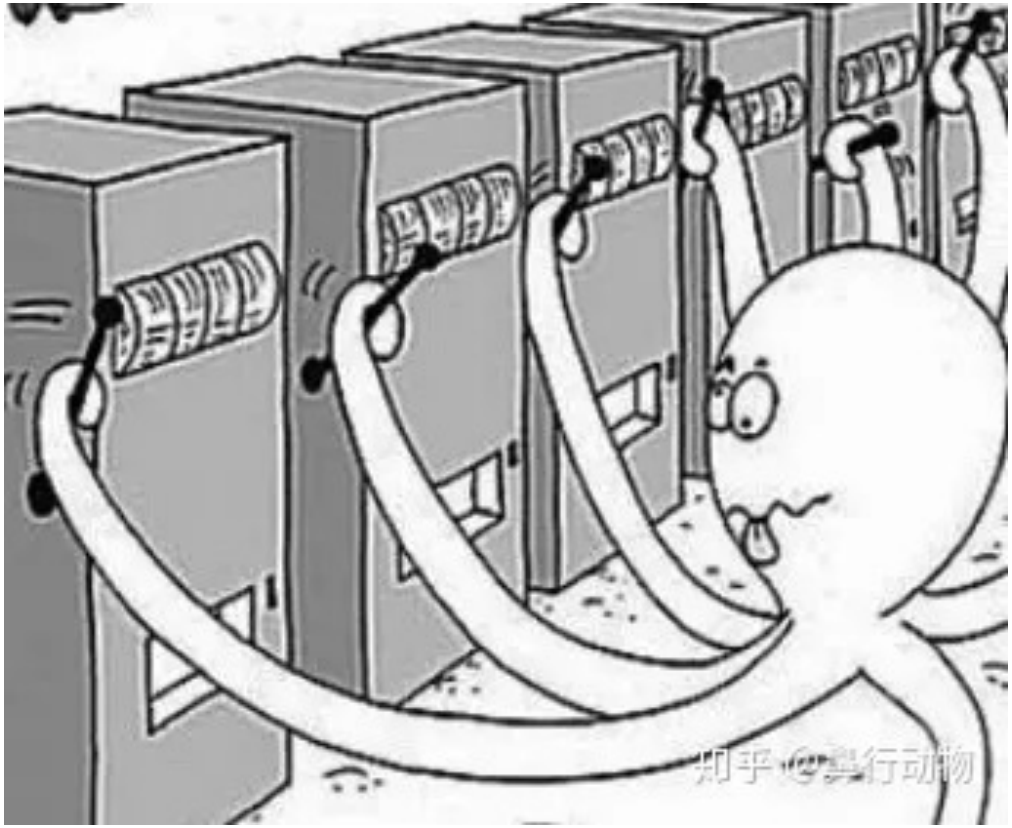
在线优化算法

张伯雷

南京邮电大学 计算机学院、通达学院

https://bolei-zhang.github.io/course/opt.html

多臂赌博机



一个赌徒面前有N个老虎机,事先他不知道每台老虎机的真实盈利情况,他如何根据每次玩老虎机的结果来选择下次拉哪台或者是否停止赌博,来最大化自己的从头到尾的收益.

多臂赌博机

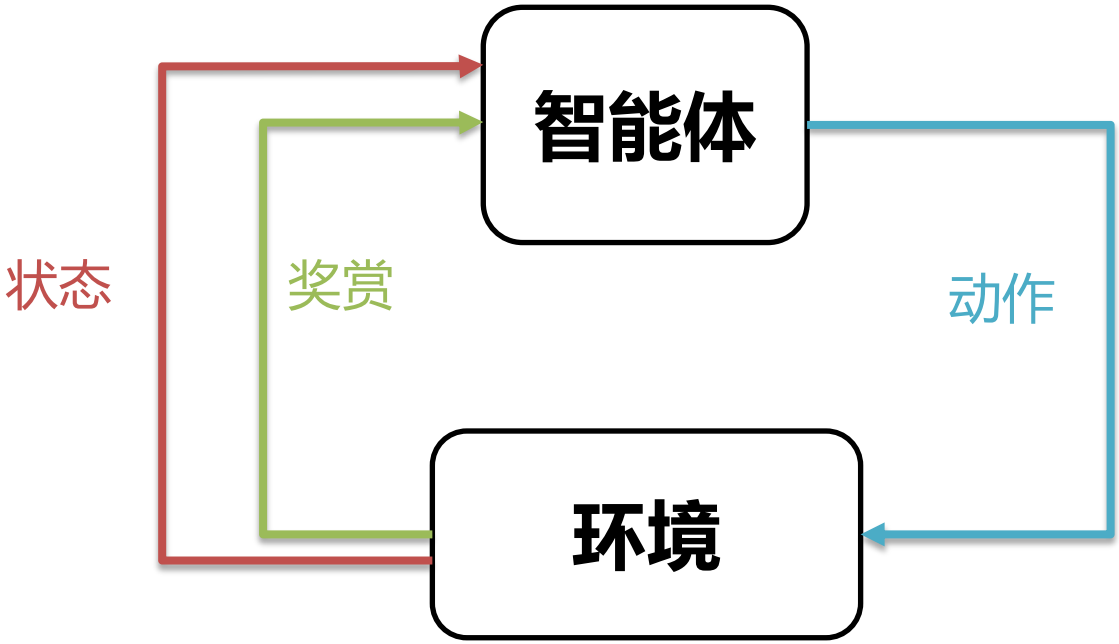
- 每台赌博机i是否有奖赏，对应一个伯努利分布

$$f(k;p_i)=\begin{cases}p_i, & if\ k=1\\ 1-p_i, & if\ k=0\end{cases}$$

- 事先不知道每台赌博机对应的概率p_i

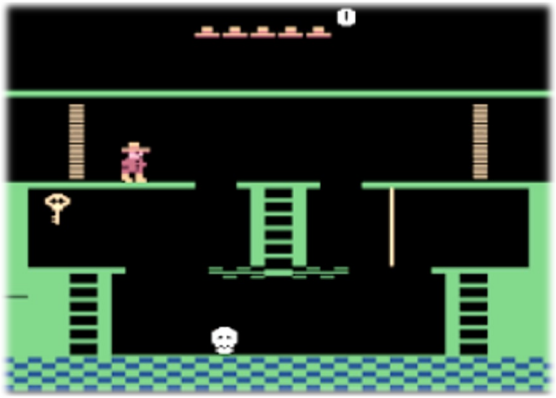
- 探索-利用（Exploration-Exploitation）：
  - 随机去摇赌博机，通过多次探索来估计p_i
- ε-greedy
  - 对开发和试探进行折中，每次**以 ε概率进行**探索，**$1-\varepsilon$ 的概率进行**利用。

强化学习



目标：通过采取最优的动作，使得累积奖赏达到最大

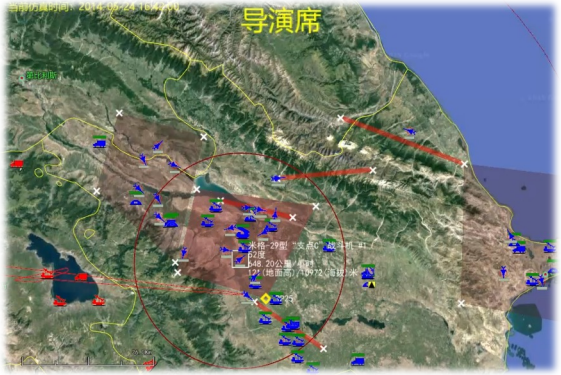
相关应用



游戏AI

无人驾驶

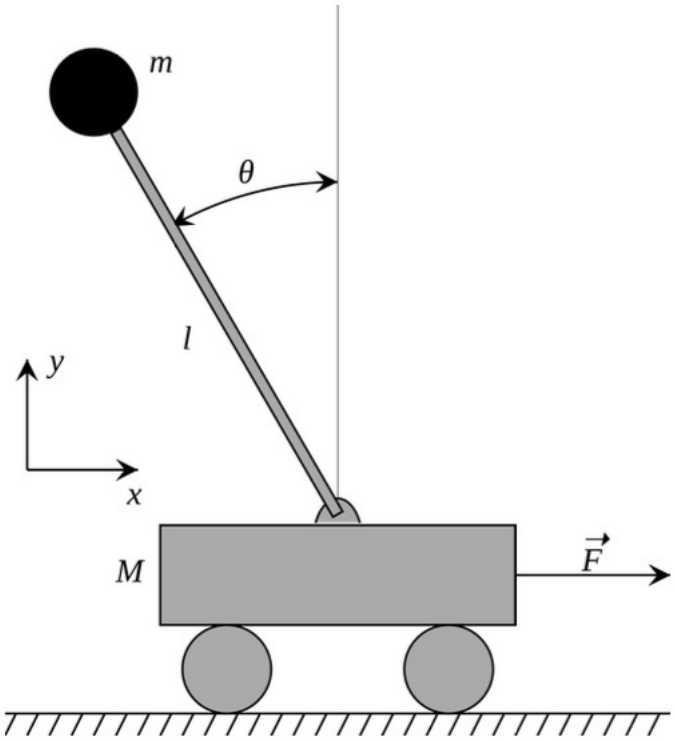
工业控制



JS指控

强化学习是人工智能中重要的建模、求解与优化技术，近年来随着深度学习和智能需求的发展成为当前研究热点。

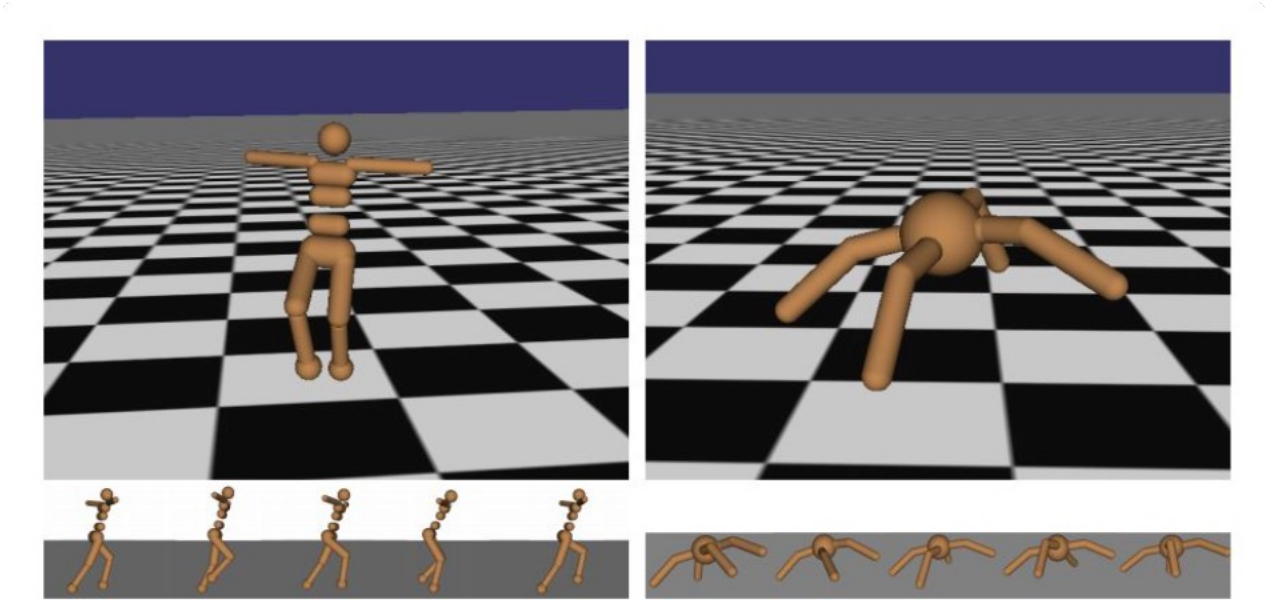
Cart-Pole



- 目标：在推车顶部平衡一根杆子

- 状态：角度、角速度、位置、水平速度
- 动作：施加在小车上的水平力
- 奖励：如果杆是直立的，则+1

Robot Locomotion



- 目的：使机器人向前移动

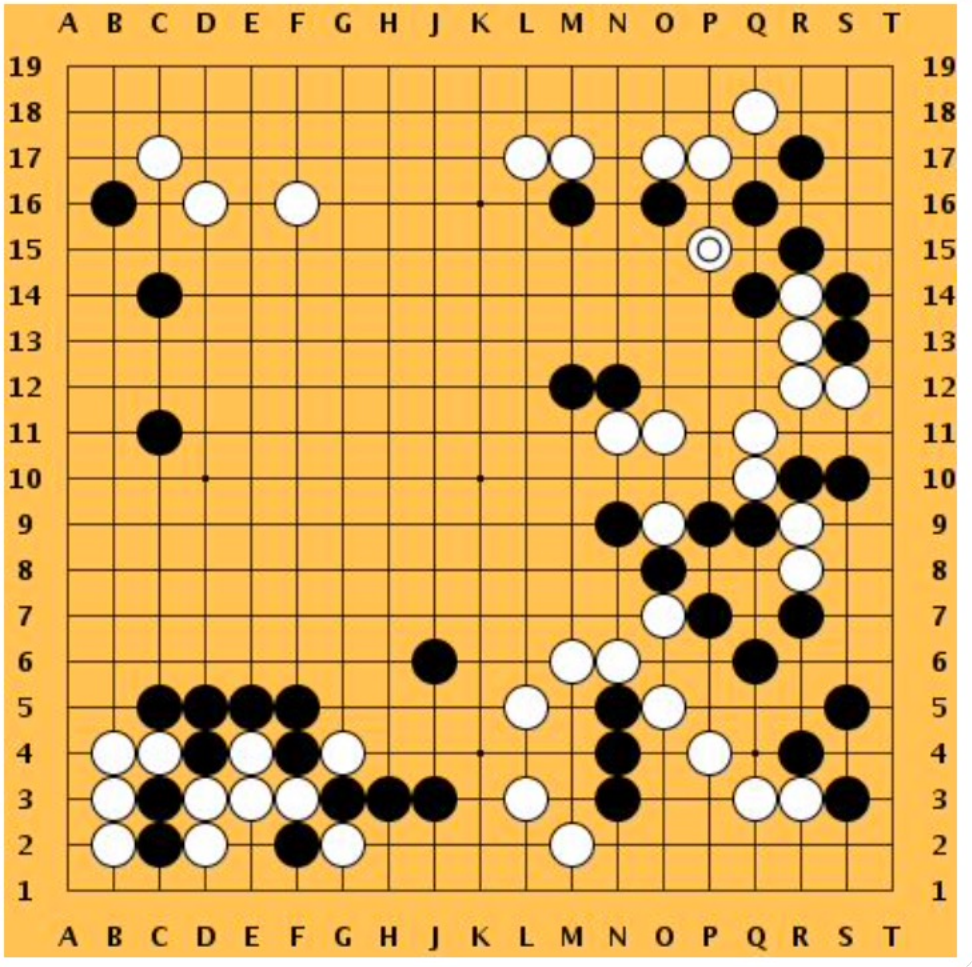
- 状态：关节的角度和位置
- 动作：施加在关节上的扭矩
- 奖励：每次直立向前移动1步+1

Atari Games

- 目标：以最高分完成游戏

- 状态：游戏状态的原始像素输入
- 动作：游戏控制，例如 左、右、上、下
- 奖励：分数增加/减少

Go



- 目标：赢得比赛

- 状态：所有棋子的位置
- 动作：下一块放在哪里
- 奖励：比赛结束时获胜则为1，否则为0

马尔科夫决策过程

- 马尔可夫性质：环境的下一个状态，之和当前状态与动作相关，与更早的状态无关

- 被定义为（S, A, R, T, γ）
  - S：一组可能的状态
  - A：一组可能的操作
  - R：给定（状态，动作）对的奖励分配
  - T：转移概率，即给定（状态，动作）对的下一个状态的分布
  - γ：折扣因子

马尔科夫决策过程

- 在$t=0$时，环境采样初始状态 $s_0 \sim p(s_0)$
- 从$t=0$ 开始直到完成：
  - 智能体选择操作a_t
  - 环境样本奖励 $r_t \sim R(\cdot|s_t, a_t)$
  - 环境采样下一个状态 $s_{t+1} \sim T(\cdot|s_t, a_t)$
  - 智能体收到奖励r_t和下一个状态 s_{t+1}

- 策略π是从 S 到 A 的函数，即在每个状态下采取什么操作
- 目标：找到使累计折扣奖励最大化的政策π^*：

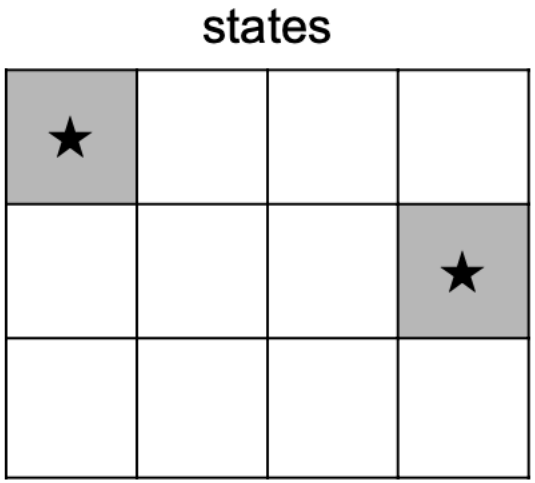
$$\sum_{t\geq 0} \gamma^t r_t$$

Grid World

actions = {

1. right
2. left
3. up
4. down

}

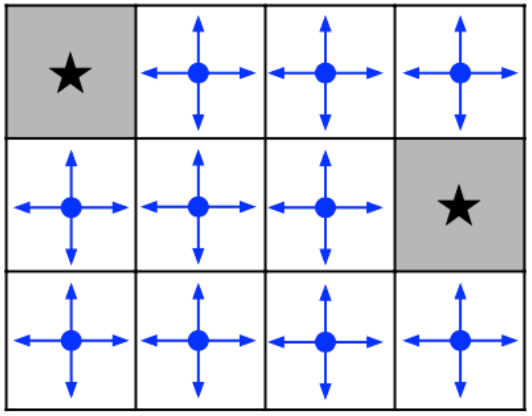


确定转移函数

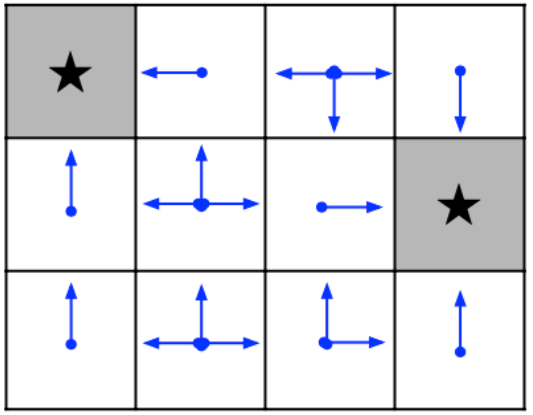
目标：在最少得步数内达到目标状态

如何设计奖赏函数?

Grid World



随机策略



最优策略

估值函数与Q-函数

- 遵循策略会产生样本轨迹（或路径）s_0、a_0、r_0、s_1、a_1、r_1，…
- 一个状态有多好?
- 状态 s 的价值函数是遵循状态 s 的策略的预期累积奖励:

$$V^{\pi}(s) = \mathbb{E}\left[\sum_{t\geq 0}\gamma^t r_t | s_0 = s, \pi\right]$$

- 状态-行动对有多好?
- 状态 s 和动作 a 的 Q 值函数是在状态 s 中采取动作 a 然后遵循策略的预期累积奖励:

$$Q^{\pi}(s,a) = \mathbb{E}\left[\sum_{t\geq 0}\gamma^t r_t | s_0 = s, a_0 = a, \pi\right]$$

贝尔曼方程

- 最优 Q 值函数 Q^* 是给定状态-动作下可实现的最大预期累积奖励：

$$Q^*(s,a) = \max_{\pi} \mathbb{E}\left[\sum_{t\geq 0} \gamma^t r_t | s_0 = s, a_0 = a, \pi\right]$$

- Q^* 满足以下贝尔曼方程：

$$Q^*(s,a) = \mathbb{E}_{s'\sim\mathcal{E}}\left[r + \gamma \max_{a'} Q^*(s',a') | s, a\right]$$

- Intuition：如果下一个时间步 $Q^*(s',a')$ 的最优状态动作值已知，那么最优策略就是采取使期望值最大化的动作

谢谢！！