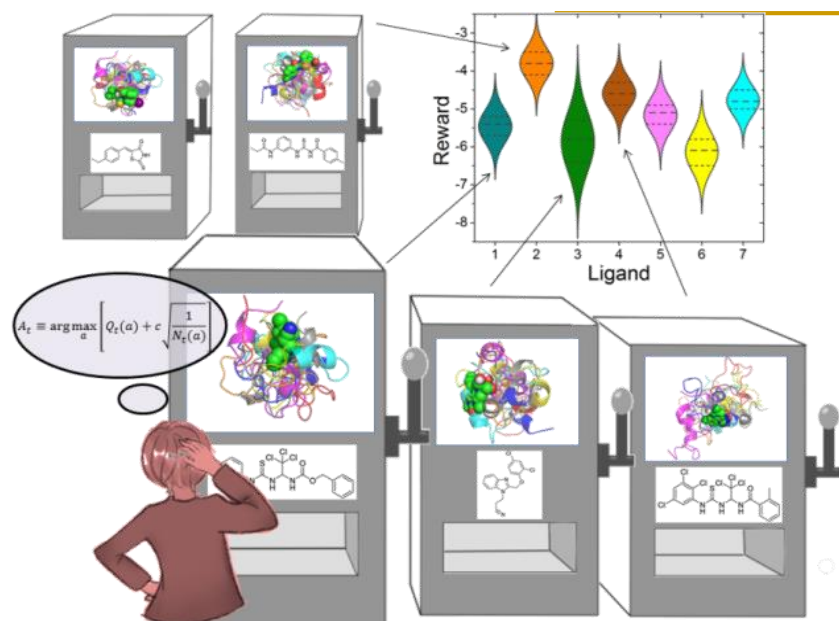


《机器学习及其在化学中的应用》2025年课程

Sec. 11



强化学习1：多臂老虎机



刘志荣 (LiuZhiRong@pku.edu.cn)

北京大学化学学院

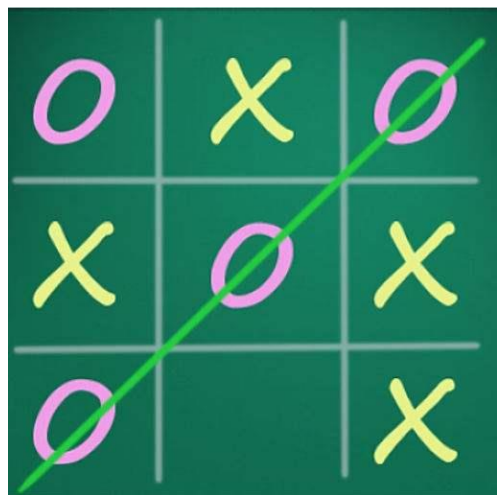
2025.12.1

内容提要

- 强化学习引言
- 多臂老虎机

1. 强化学习引言

(Reinforcement learning)



背景：学习与强化学习

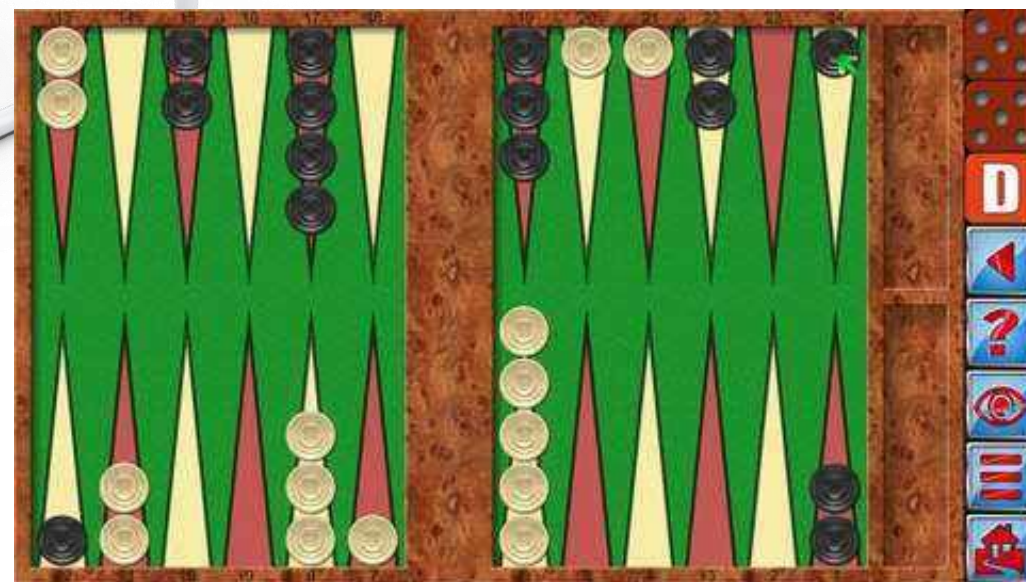
- Learning from interaction. 我们是通过与环境互动来学习的。
 - 例如：婴儿。
- 强化学习（reinforcement learning）：通过计算来（实现）从互动中学习。
 - 目标导向的互动学习（Goal-directed learning from interaction）
- 怎样做，使回报最大化？
 - what to so as to maximize a numerical reward.
- 两个显著特征：试错（trial-and-error）与延迟回报（delayed reward）。

From Sutton's book:

The idea that we learn by interacting with our environment is probably the first to occur to us when we think about the nature of learning. Learning from interaction is a foundational idea underlying nearly all theories of learning and intelligence. When an infant plays, waves its arms, or looks about, it has no explicit teacher, but it does have a direct sensorimotor connection to its environment. Exercising this connection produces a wealth of information about cause and effect, about the consequences of actions, and about what to do in order to achieve goals. Throughout our lives, such interactions are undoubtedly a major source of knowledge about our environment and ourselves. Whether we are learning to drive a car or to hold a conversation, we are acutely aware of how our environment responds to what we do, and we seek to influence what happens through our behavior.

例子

- 无人机控制；
- 棋类游戏；
- 仿人机器人行走；
- 投资组合管理；
- 电子游戏；



强化学习的一些概念与特点

- Agent（智能体、个体、本体、代理、学习者、决策者）
- 感知（sensation）、行动（action）、目标（goal）。
- 与监督学习的区别：没有（现成的）数据集。
- 与无监督学习的区别：极大化目标，而非寻找潜在结构。
- 强调探索与利用（explore-exploit）之间的平衡。
- 强调与未知环境（uncertain environment）的互动。
- Time really matters !
 - 时间序列信息；
 - 计算速度

强化学习四要素

- 策略/政策 (policy): 从（感知到的）状态（state）到（采取的）行动/动作（action）的映射。
 - 状态 s ，行动 a
 - 策略 $\pi: s \rightarrow a$
- 回报（奖励）信号（reward signal）：行动后收到的即时回报
 - 回报 r
- 价值函数（value function）：状态在未来的总体（长期）回报。即对最终目标的贡献。
- 环境模型 (a model of environment): 模拟/描述环境的行为。

例子1: Tic-Tac-Toe

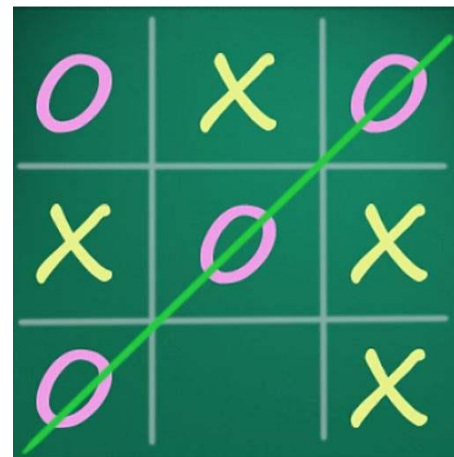
■ 初始价值函数:

$$V(S) = \begin{cases} 1, & \text{if } S = \text{赢} \\ 0, & \text{if } S = \text{输} \\ 0.5, & \text{otherwise} \end{cases}$$

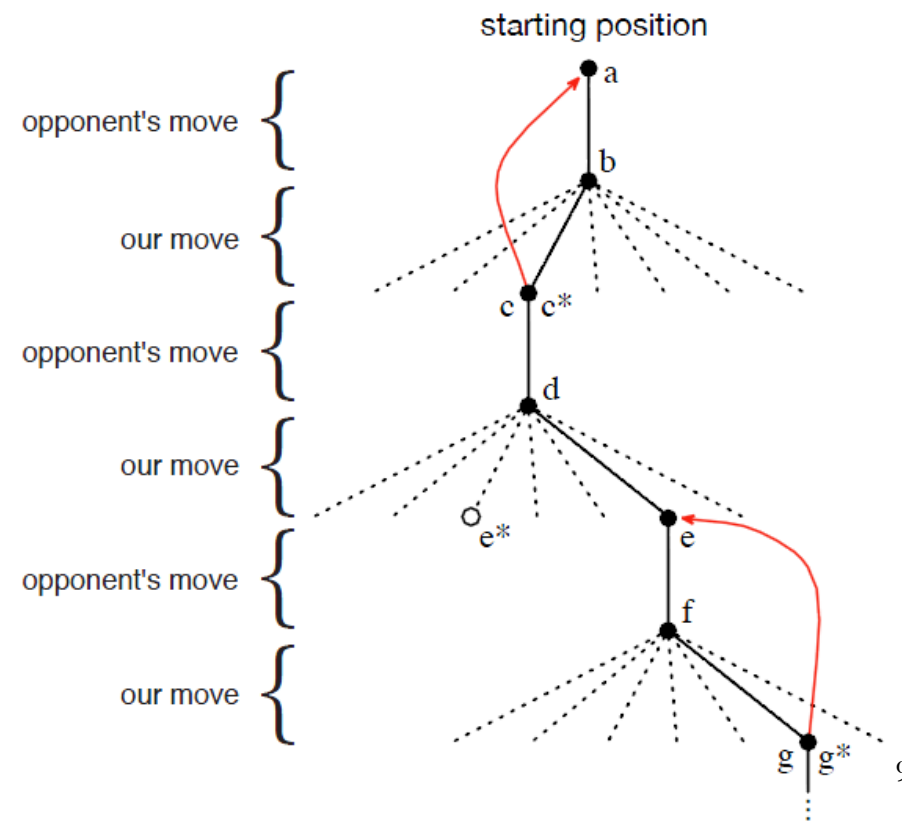
■ 与对手多次对弈

- 一般采用贪心（贪婪）走法，偶尔随机探索；
- 一边走一边更新走过状态的价值函数：

$$V(S_t) \leftarrow V(S_t) + \alpha[V(S_{t+1}) - V(S_t)]$$

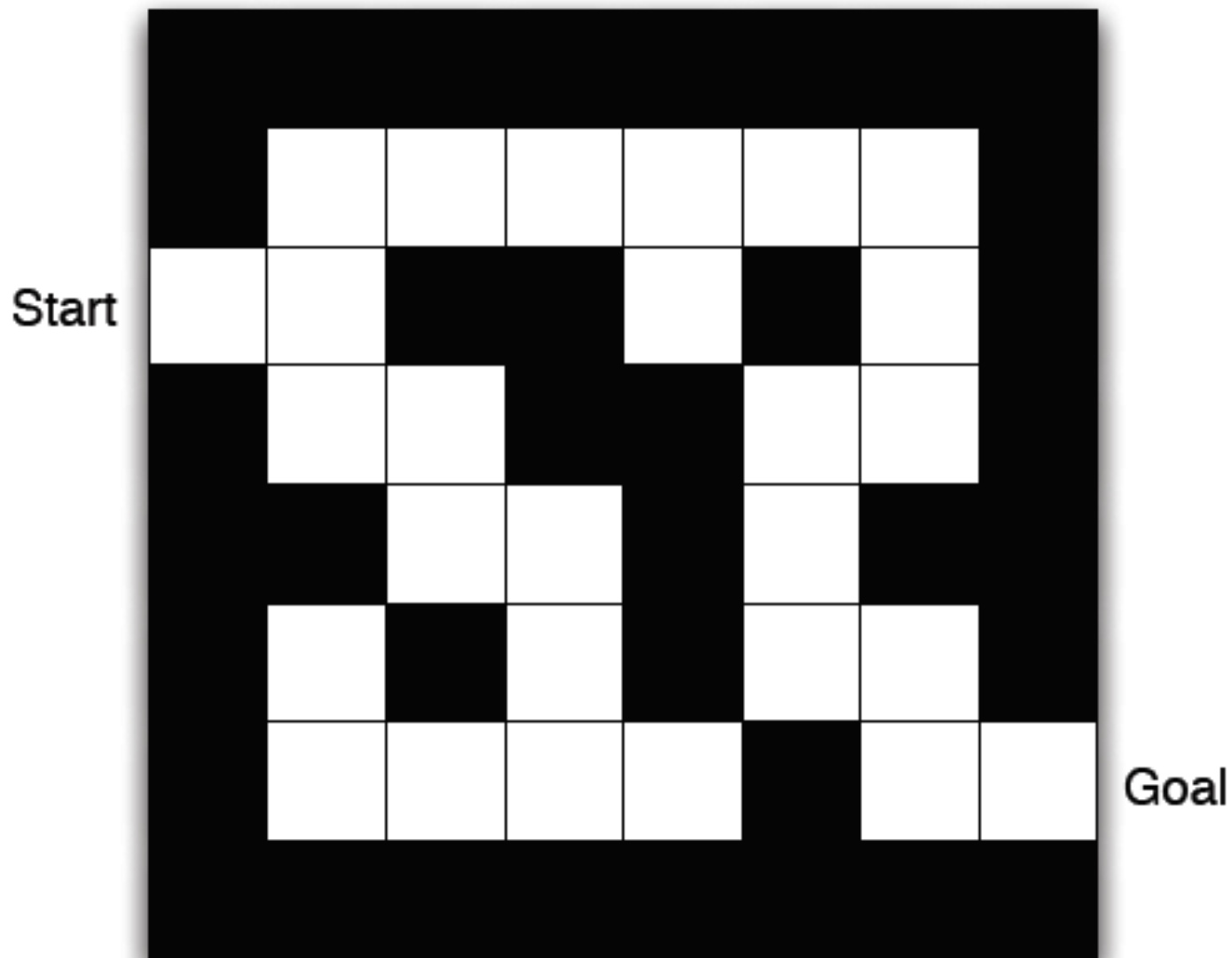


井字棋

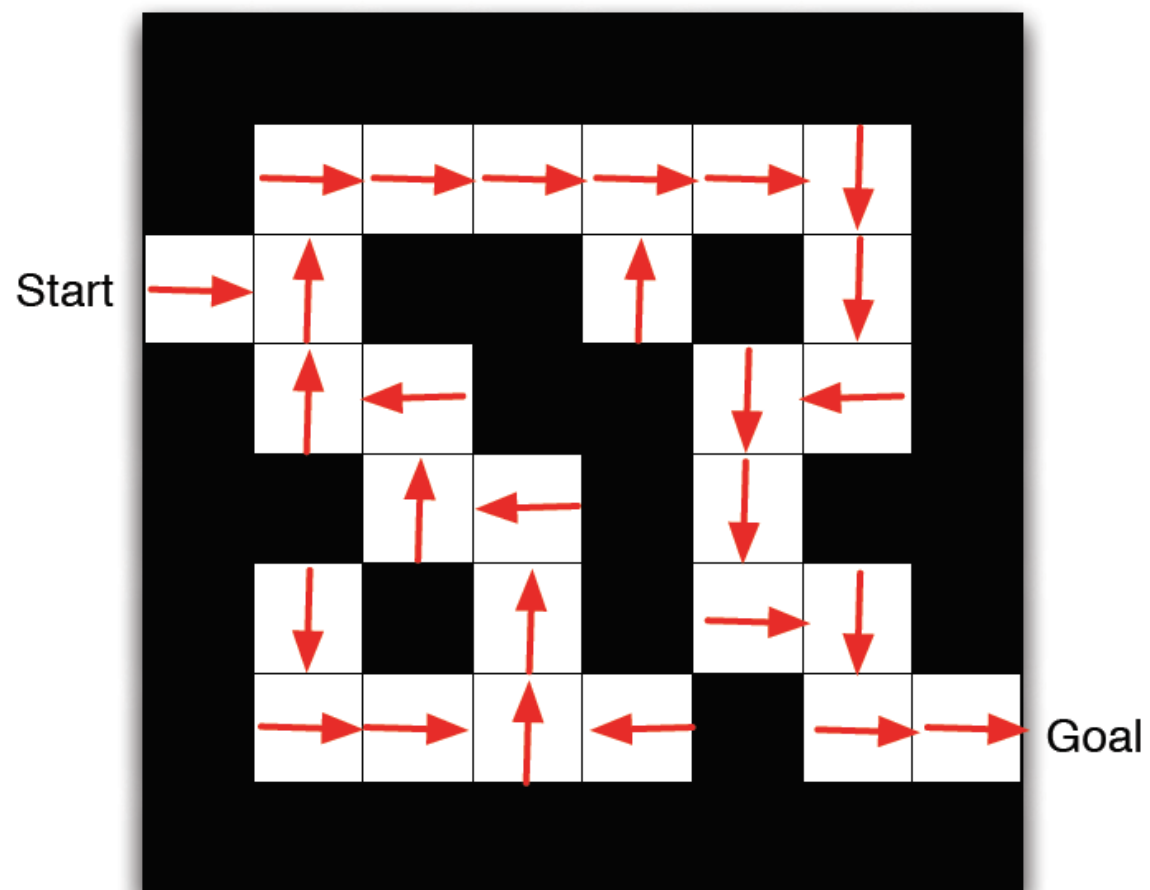


例子2：迷宫

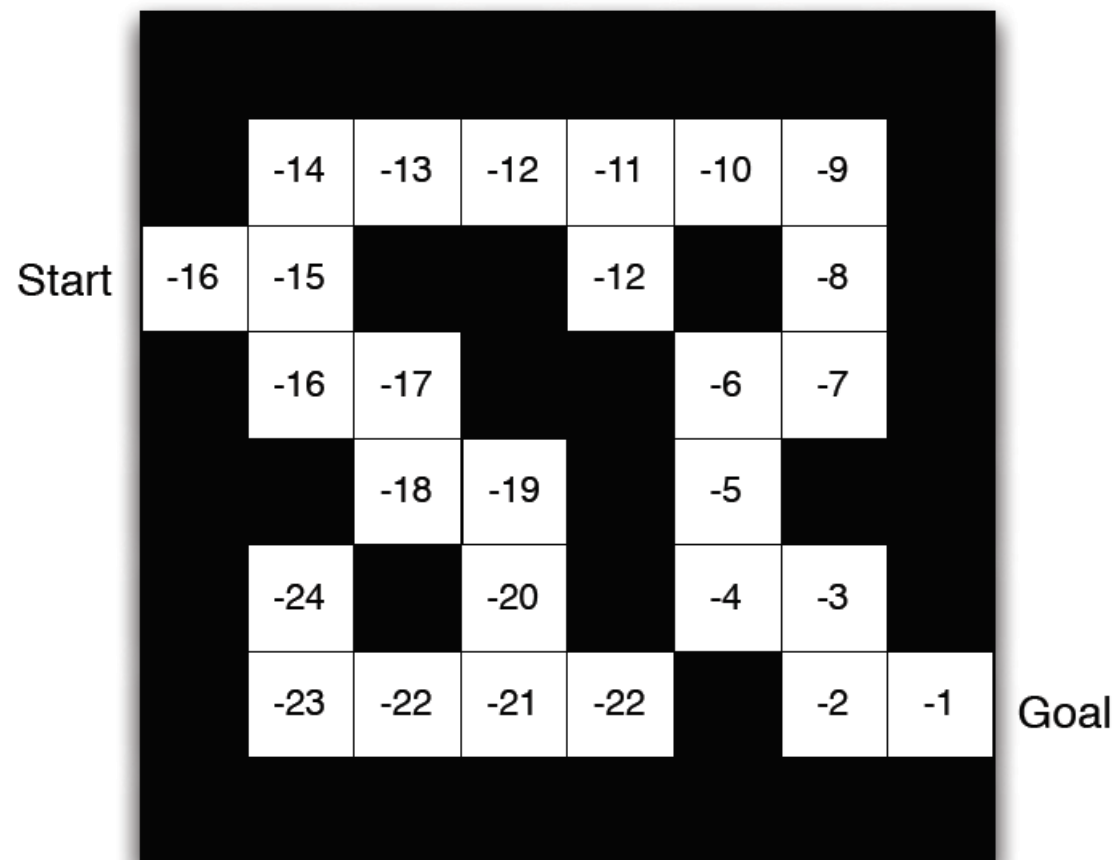
- 回报：每步-1
- 状态：当前位置
- 行动：上下左右



最佳策略

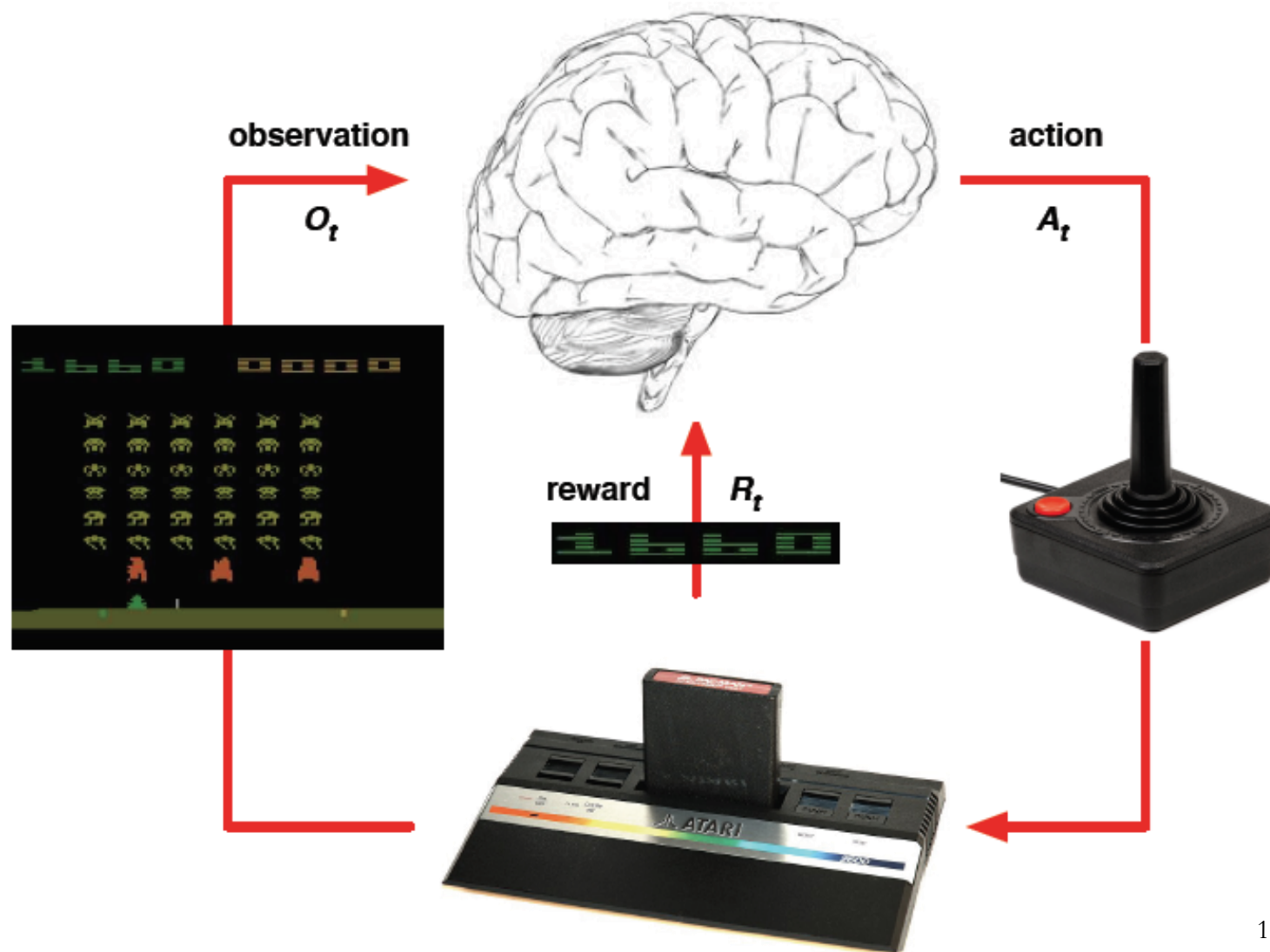


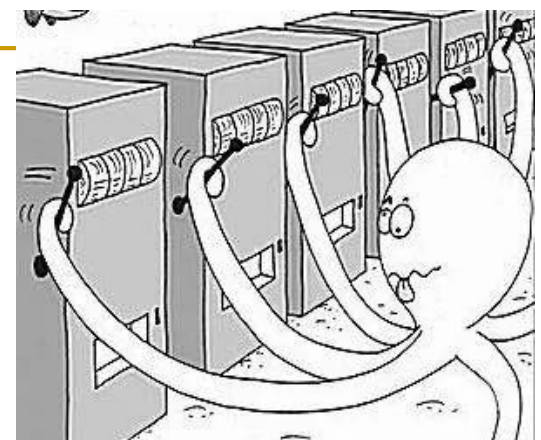
价值函数



例子3：电子游戏

- 规则未知。
- 直接打！
- 控制手柄，
看像素与分数输出。
- 其它：
 - Chem-is-try
 - Re-search

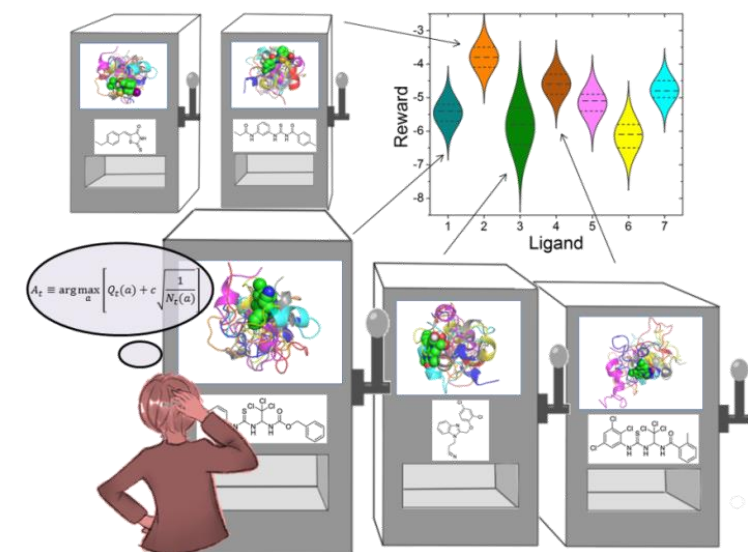




2. 多臂老虎机

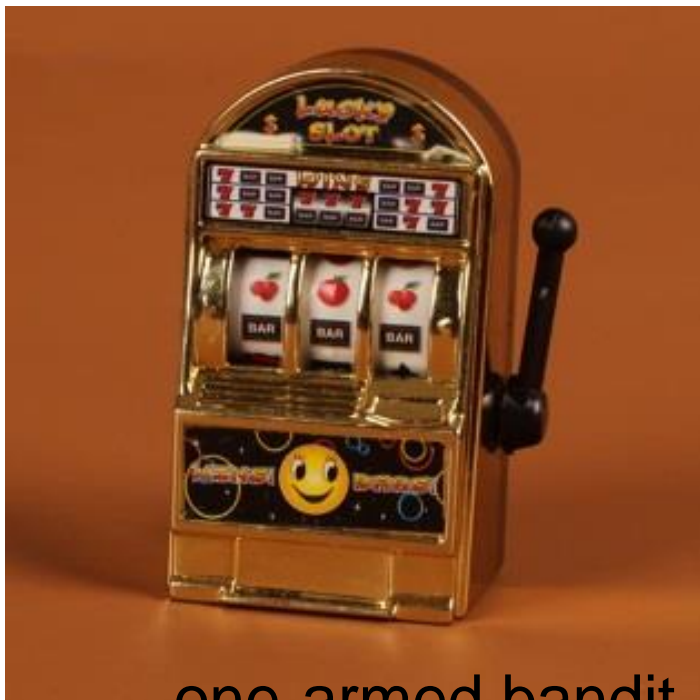
—探索与利用的平衡

explore-exploit



多臂老虎机

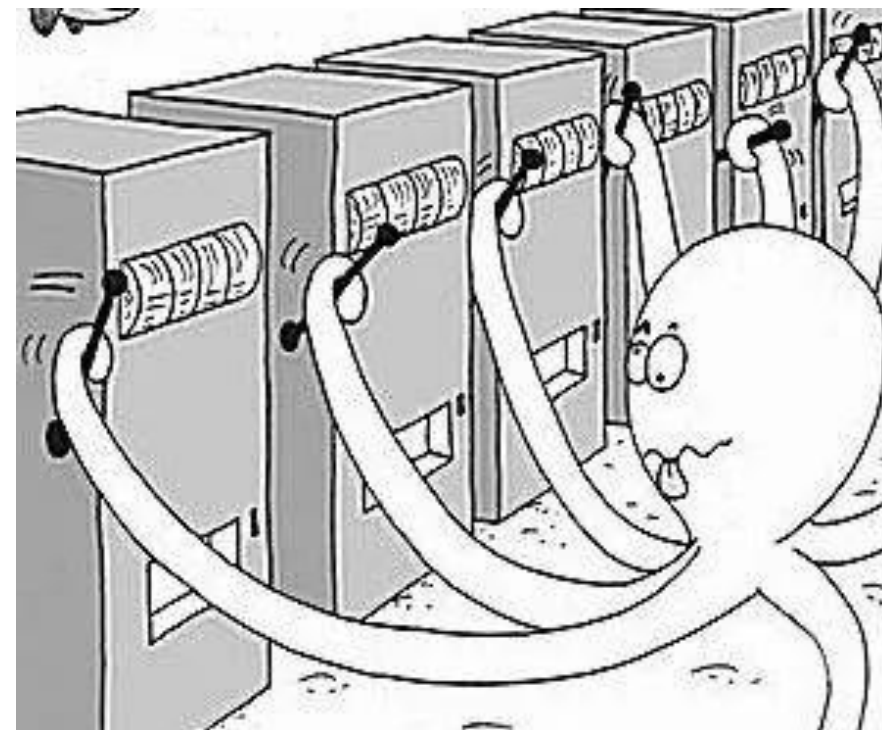
- Multi-armed bandit



one-armed bandit



大数据文摘



- 问题：如何操作多台不同的老虎机，以实现最高的可能获利？

数学描述

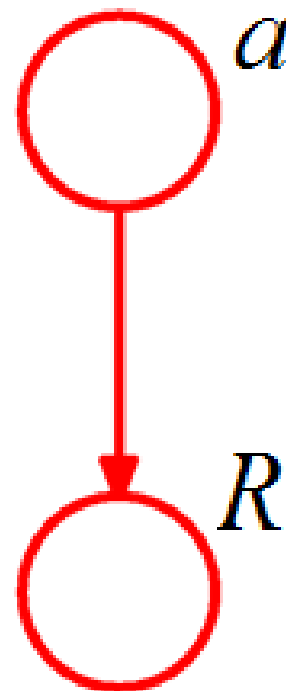
- 每次面临着有 k 个选项的选择，当选择了一个选项（行动） a 后，将得到回报 R ， R 的静态分布取决于 a ，即 $p(R|a)$ 。
- 问题：怎样逐步选择 N 个行动 A_t （ $t = 1, 2, \dots, N$ ），使获得的总回报 $\sum_{t=1}^N R_t$ 最大。
- 符号理解：

$$a \Leftrightarrow \mathbf{x}$$

（更准确地说，代表 \mathbf{x} 的 k 个可能取值的任一个）

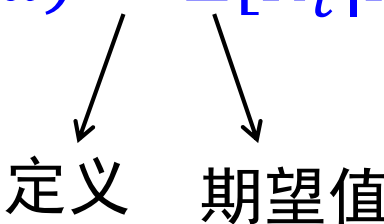
$$A_t \Leftrightarrow \mathbf{x}_n$$

$$R_t \Leftrightarrow t_n$$



分析

- 行动 a 的价值（value）定义为

$$q_*(a) \doteq \mathbb{E}[R_t | A_t = a]$$


定义 期望值

- 如果 $q_*(a)$ 已知，则问题很简单，每步只需要选择 $q_*(a)$ 最大的 a 。
- 实际上 $q_*(a)$ 未知，在第 t 步做选择时只能根据之前的结果计算估计值 $Q_t(a)$
- 探索 vs. 利用（Exploration and Exploitation）。

生活中的例子

- 在单位工作了好几年，应该跳槽吗？
- 附近有个餐馆，你已经去过十五次，九次的体验非常好，六次不怎么好。明晚你打算出去吃饭，是否应该尝试新餐馆？
- 有个老作者的书，我读过五本了，三本写得挺好，两本比较差；另一个新作者，我读过他两本书，一本比较好，一本比较差。那下一本我该买谁的书呢？
- 医生给重症病人选择的治疗方案？
- 制药公司的研发节奏？
- 石油钻探。

行动-价值方法

- 估计行动 a 的价值，并据之采取行动。

- 一种估计方法：

$$Q_t(a) = \frac{\sum_{i=1}^{t-1} R_i I(A_i = a)}{\sum_{i=1}^{t-1} I(A_i = a)}$$

$A_i = a$ 时等于1，否则0

- 或采用贝叶斯方法来估计。 a 无采样点时用缺省值（0或平均值）

- 贪心算法（greedy algorithm）：

$$A_t = \arg \max_a Q_t(a)$$

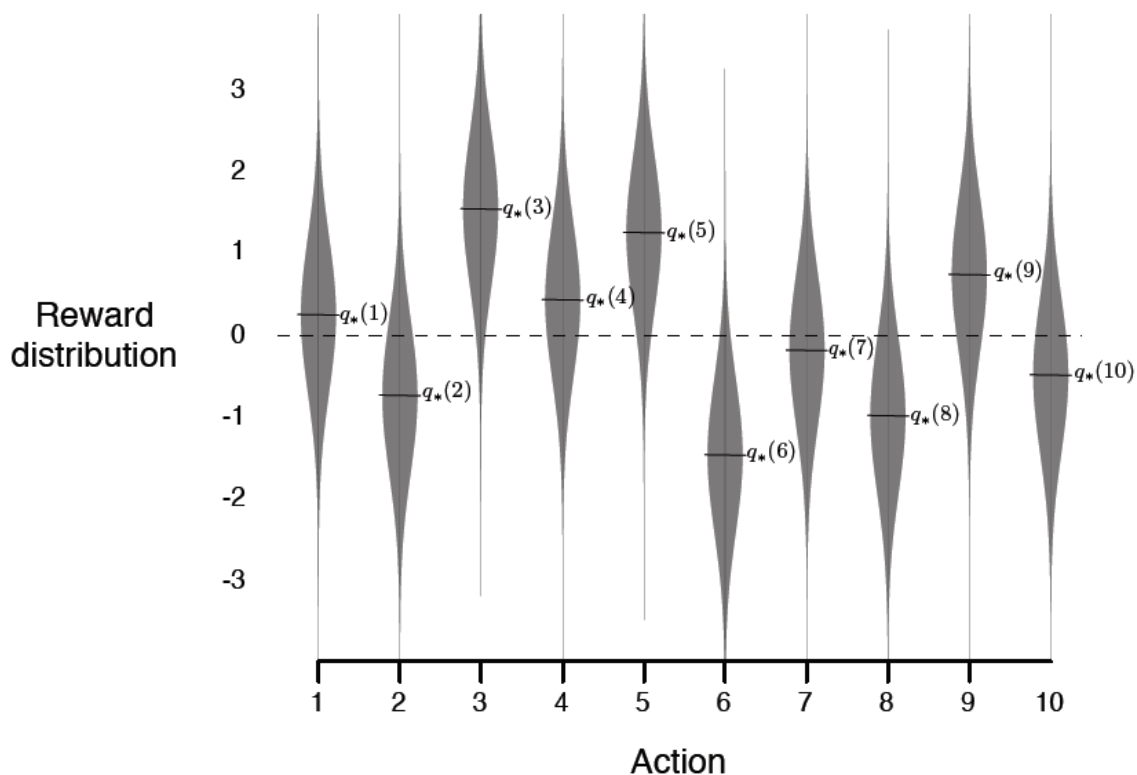
函数极大值所对应的参数(a)

- ϵ -贪心算法：有小概率 ϵ 随机选择各种行动。

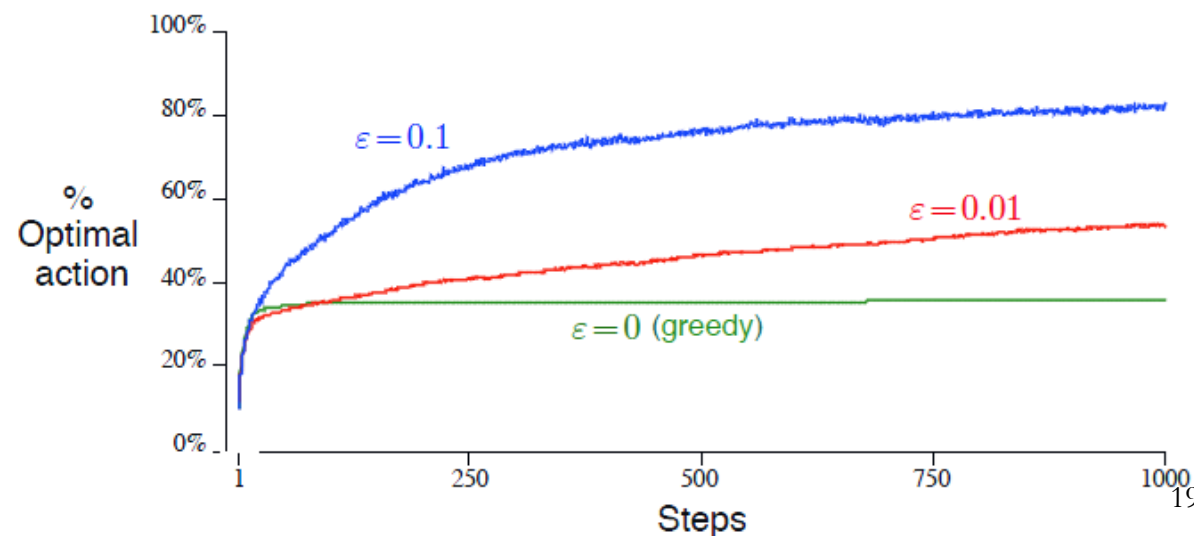
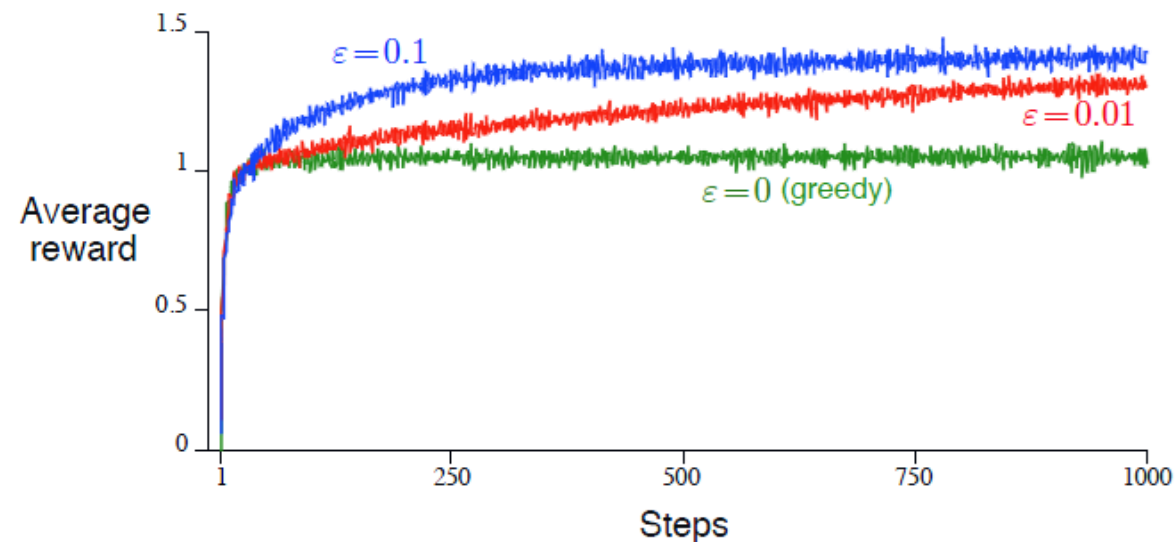
例子：10臂老虎机...

$$p[q_*(a)] = \mathcal{N}(0,1)$$

$$p(R|a) = \mathcal{N}(q_*(a), 1)$$



2000次模拟的平均



在线学习 (增量式实现, Incremental Implementation)

- 假设某个行动 a 被采纳了 n 次, 其价值函数的新的估计值是

$$\begin{aligned} Q_{n+1} &= \frac{1}{n} \sum_{i=1}^n R_i = \frac{1}{n} (R_n + \sum_{i=1}^{n-1} R_i) = \frac{1}{n} \left(R_n + \frac{n-1}{n-1} \sum_{i=1}^{n-1} R_i \right) \\ &= \frac{1}{n} (R_n + (n-1)Q_n) = Q_n + \frac{1}{n} (R_n - Q_n) \end{aligned}$$

- 因此

$$Q_{n+1} = Q_n + \eta(R_n - Q_n)$$

- 此处学习率

$$\eta = \frac{1}{n}$$

非静态分布

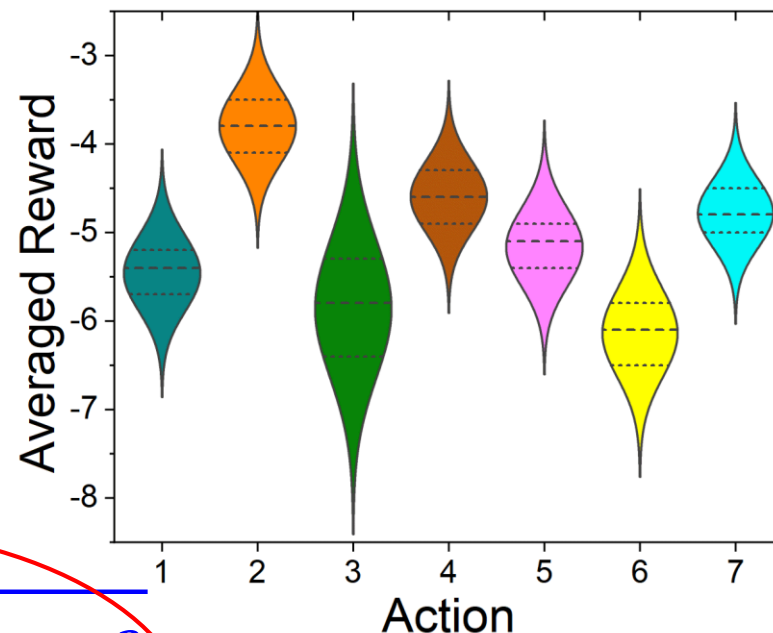
- 如果 $p(R|a)$ 随时间变化，合理的做法是给更近的样本赋予更大的权重。
- 例如，将 η 设为常数：

$$\begin{aligned} Q_{n+1} &= Q_n + \eta(R_n - Q_n) = \eta R_n + (1 - \eta)Q_n \\ &= \eta R_n + (1 - \eta)[\eta R_{n-1} + (1 - \eta)Q_{n-1}] \\ &= \dots = (1 - \eta)^n Q_1 + \sum_{i=1}^n \eta(1 - \eta)^{n-i} R_i \end{aligned}$$

这是一种指数近因加权平均（exponential recency-weighted average）。

置信区间上界算法

- Upper-Confidence-Bound (UCB)
- $Q_t(a)$ 是对 $q_*(a)$ 的估计。
- 对 a 的采样样本越多，估计越准确，变化范围（误差）与 $\sqrt{1/N_t(a)}$ 成正比。



测量平均值

真实平均值

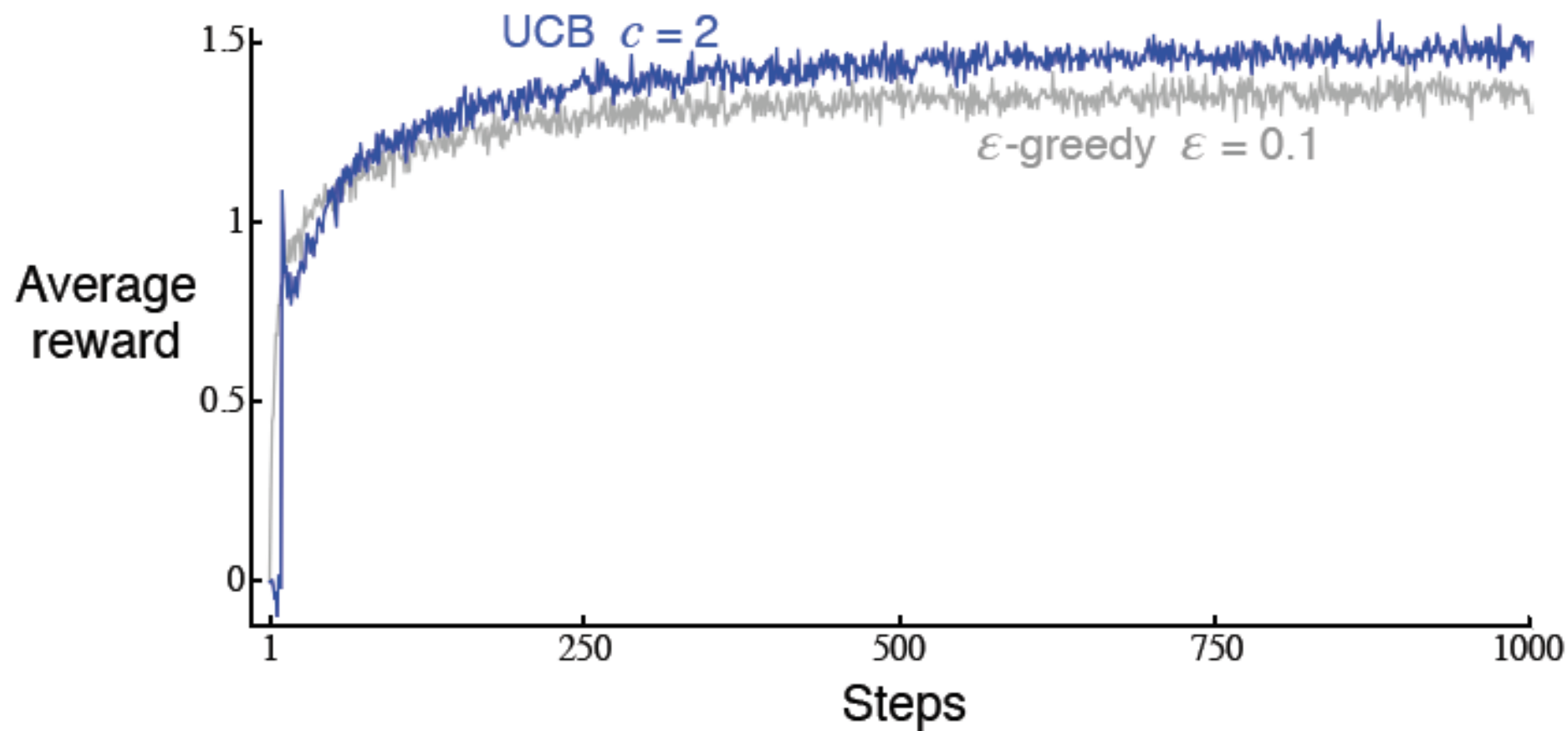
$$\bar{x} \approx \frac{\sum_{i=1}^n x_i}{n} \pm \frac{1}{\sqrt{n}} \sqrt{\frac{\sum_{i=1}^n \left[x_i - \left(\frac{\sum_{j=1}^n x_j}{n} \right) \right]^2}{n-1}}$$

测量方差

- 基于面对不确定性时的乐观原则，以期望值置信区间上界为选择行动的标准，一种UCB算法是（过程复杂，略）：

$$A_t = \arg \max_a \left[Q_t(a) + c \sigma_t(a) \sqrt{\frac{\ln t}{N_t(a)}} \right]$$

例子：10臂老虎机...

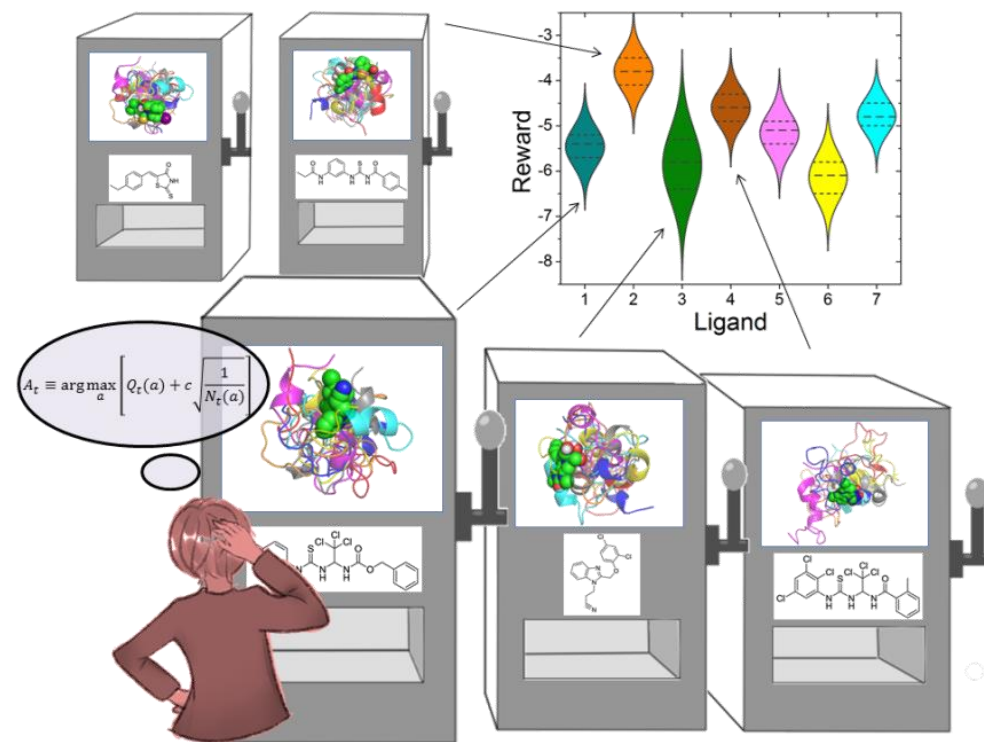


题外：天然无序蛋白质的药物虚拟筛选...

$$K_a^{(\text{app})} = \frac{1}{N} \sum_{\text{构象 } i} e^{-\frac{\Delta G_i}{RT}}$$

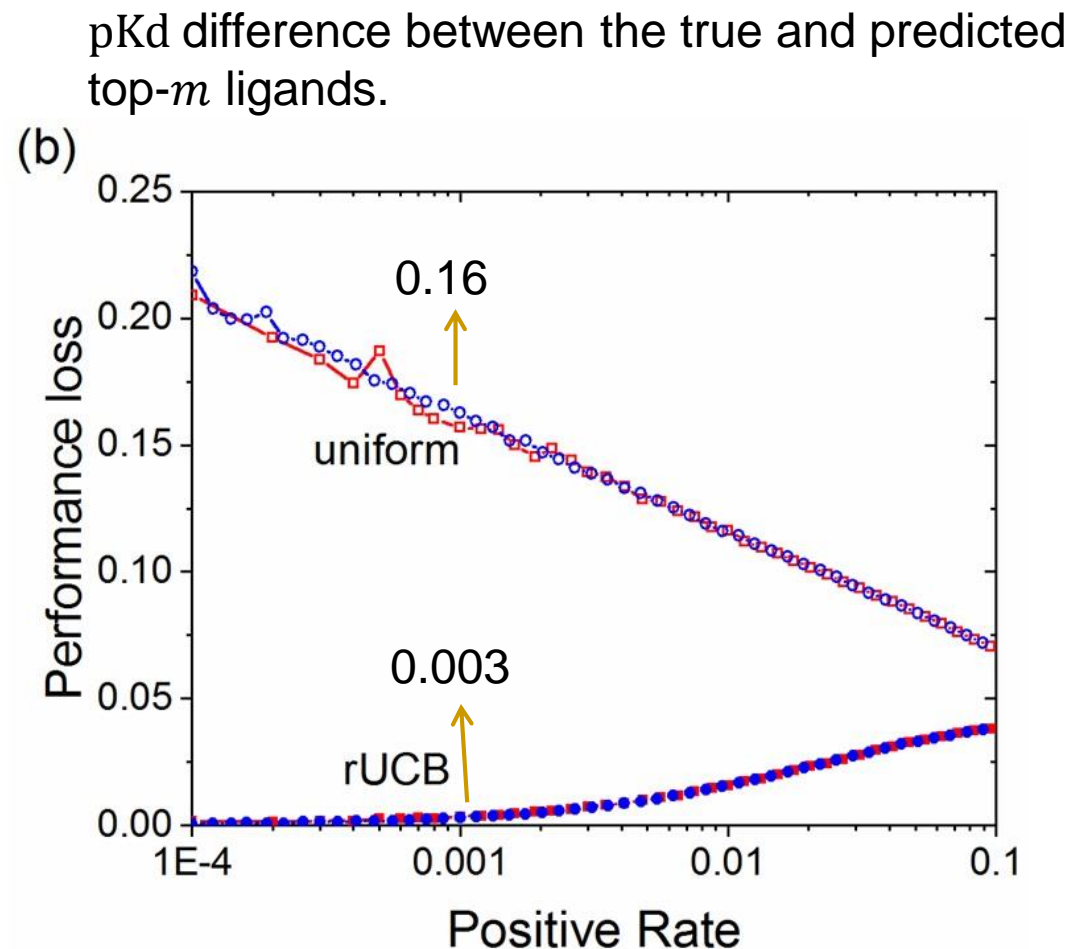
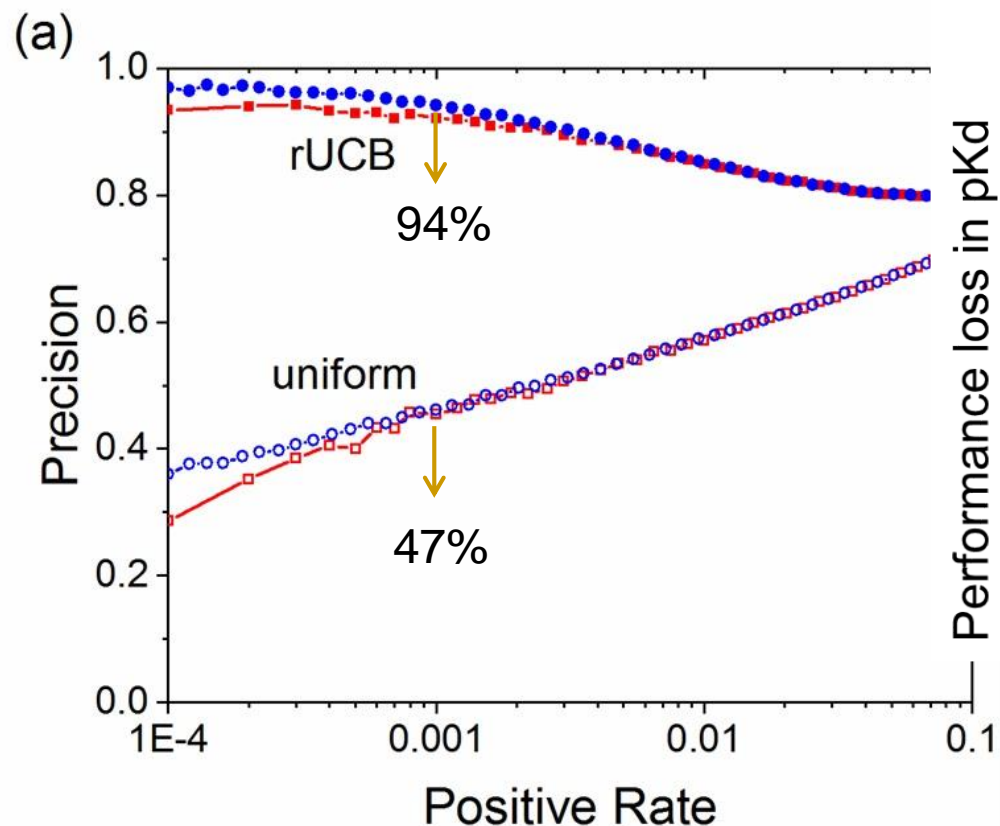
- 无序蛋白的构象不固定，需要考虑构象系综中所有构象的（分子对接）贡献。

	Bandits	IDPs screening
Arm number	~10	~100000 (ligands)
Top bids	1	~100
Total runs	~10000	~10 ⁶ -10 ⁷
Runs/arm	~1000	~10-100??



Bin Chong (崇滨), ..., Zhirong Liu*. Reinforcement learning to boost molecular docking upon protein conformational ensemble. *Phys. Chem. Chem. Phys.* **23**, 6800 (2021)

r UCB结果

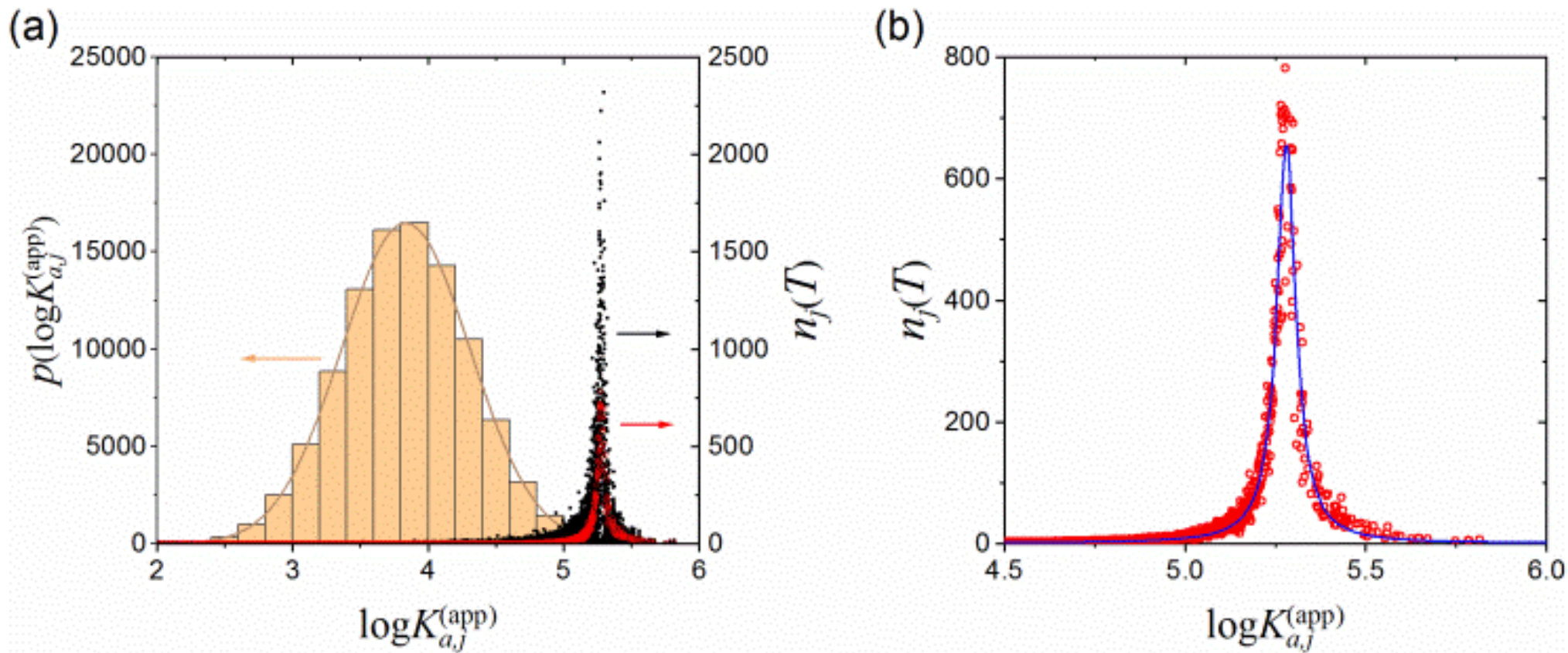


Effects on synthetic datasets with 10^4 ligands (red) or 10^5 ligands (blue).

平均每个配体只需要对接两次。

$rUCB$: 原因分析

- 对于分界线附件的配体，对接次数很多，以降低其误差



Distribution of docking times for ligands under rUCB (10 simulation)

题外：基廷斯指数 (*Gittins Index*)

- 1970s，联合利华公司请约翰.基廷斯帮助他们优化药物试验。令人意想不到的是，基廷斯解出了多臂老虎机的一类最佳策略。
- 引入贬值（折扣、贴现）的观点，考虑目标函数：

$$\sum_{t=1}^{+\infty} R_t \gamma^t$$

- $\gamma < 1$ ，是折扣因子
- 基廷斯的工作表明：每个状态有一个对应的指数（index）；最佳策略是每次选取指数最高的状态所对应的把手。

例子：餐馆的选择...

- 或：抛硬币
- $\gamma = 0.9$

基廷斯指数

		Wins									
		0	1	2	3	4	5	6	7	8	9
Losses	0	.7029	.8001	.8452	.8723	.8905	.9039	.9141	.9221	.9287	.9342
	1	.5001	.6346	.7072	.7539	.7869	.8115	.8307	.8461	.8588	.8695
	2	.3796	.5163	.6010	.6579	.6996	.7318	.7573	.7782	.7956	.8103
	3	.3021	.4342	.5184	.5809	.6276	.6642	.6940	.7187	.7396	.7573
	4	.2488	.3720	.4561	.5179	.5676	.6071	.6395	.6666	.6899	.7101
	5	.2103	.3245	.4058	.4677	.5168	.5581	.5923	.6212	.6461	.6677
	6	.1815	.2871	.3647	.4257	.4748	.5156	.5510	.5811	.6071	.6300
	7	.1591	.2569	.3308	.3900	.4387	.4795	.5144	.5454	.5723	.5960
	8	.1413	.2323	.3025	.3595	.4073	.4479	.4828	.5134	.5409	.5652
	9	.1269	.2116	.2784	.3332	.3799	.4200	.4548	.4853	.5125	.5373

自适应设计提升肿瘤药物临床试验成功率

<http://www.xinyaohui.com/news/201503/06/5283.html>

- 传统的设计导致高失败率和成本不断上升，因为关键研究问题的答案只能在试验结束时获得。自适应设计充分利用积累的数据进行设计修改，为每一个顺序步骤做出更好的决策。
- 传统的设计使用概率统计方法，对用量、随机化和样本量提前制定，通常整个试验过程没有改变。而自适应设计采纳累积的信息，数据被连续地分析，阶段性结果被用来塑造后续设计参数，如剂量、疾病的适应症或种群。使用这种灵活的方法，试验变成了学习的工具，不断变化的知识推动后续决定。

题外：强化学习...

■ <https://www.huxiu.com/article/385964.html>

看完这部纪录片，我只想赶紧扔下手机逃跑



- ❑ 在斯坦福“劝服技术实验室（Persuasive Technology Lab）”，学者们主要研究如何利用所知道的一切心理学知识来改变他人的行为。
- ❑ 很多看似“无意识”的设计，都是为了潜移默化地改变用户的行为，培养用户刷手机的习惯。比如说最常见的下拉刷新的设计，**其实和赌博老虎机的原理一样，都是一种正强化的过程**（Positive Reinforcement）。用户每次下拉软件就像拉动老虎机的把手，每次都能获得未知的新奖励，久而久之，用户就形成了无意识刷手机的反射，哪怕明知道没有新信息，也会忍不住下拉刷新。

强化学习 Toolkit



<http://incompleteideas.net/rlai.cs.ualberta.ca/RLAI/RLtoolkit/RLtoolkit1.0.html>

■ Reinforcement Learning Toolkit

- The ambition is to provide source, documentation, and updates for the Reinforcement Learning (RL) toolkit. This toolkit is a collection of utilities and demos developed by the RLAI group of Sutton which may be useful for anyone trying to learn, teach or use reinforcement learning. The tools are suitable for a range of users, from new users who have never used RL before, to very experienced users.

强化学习 Open AI Gym



<https://gym.openai.com/>

■ Open AI Gym

- ❑ Gym is a toolkit for developing and comparing reinforcement learning algorithms. It supports teaching agents everything from walking to playing games like Pong or Pinball.
- ❑ 不少课程采用。

小结

■ 强化学习

- 通过与环境互动来学习，使回报最大化。
- 试错与延迟回报。
- 探索与利用（explore-exploit）之间的平衡。
- 四要素：策略 (policy)、回报 (reward)、价值 (value)、环境

■ 多臂老虎机

- 问题：如何操作多台不同回报概率（未知）的老虎机，以实现最高获利。
- 贪心算法与 ϵ -贪心算法
- UCB算法：基于乐观原则，以期望值置信区间上界为选择标准。

■ Reference:

- Sutton 1-2;

■ 扩展阅读:

- [算法之美-探索与利用.pdf](#)

- <https://www.jianshu.com/p/b9ac7b91a154>

- [你生活中的每个选择，数学都能计算给你最正确的选择吗？.mht](#)

- <https://www.jiqizhixin.com/articles/2019-02-20-8>

- [强化学习之原理与应用.mht](#)

- <https://www.jiqizhixin.com/articles/2019-08-28-5>

- [DeepMind开源强化学习游戏框架，25款线上游戏等你来挑战.mht](#)

- <https://www.huxiu.com/article/385964.html>

- [看完这部纪录片，我只想赶紧扔下手机逃跑.pdf](#)

- <https://www.chem.pku.edu.cn/kyjz/128627.htm>

- [AI助力无序蛋白药物虚拟筛选，多臂老虎机再显神通.mht](#)

- [Chongbin3.pdf](#)

谢谢大家!