**深 圳 大 学 实 验 报 告**

|  |
| --- |
| **课程名称 人工智能导论**  **项目名称 实验三：强化学习实践**  **学 院 计算机与软件学院**  **专 业 计算机科学与技术**  **指导教师 高 灿**  **报 告 人 简校滨 学号 2022150176**  **实验时间 2024年11月24日至2024年12月21日**  **实验报告提交时间 2024年12月19日** |

**教务处制**

# 一、实验目的与要求

**实验目的：**

1．熟悉强化学习相关概念；

2．了解表格解决算法；

3．了解探索与利用的平衡策略，运用强化学习解决问题；

**实验要求：**

1. 实验提交文件为实验报告和相关程序代码，以压缩包的形式提交，命名规则为“学号数字+姓名+Task3”，如2020154099张三Task3；

2. 所有素材和参考材料需列明出处，实验报告中的图片和程序代码建议标注个人水印或标识信息：姓名，班级，学号信息；

3．实验报告内容原则上控制在10页之内。

# 二、实验内容与方法

**实验内容（三选一）：**

1. 使用贪心算法和ϵ-贪心算法解决多臂老虎机问题；

2. 使用价值迭代算法完成网格世界问题；具体细节参考文件“RL指引.doc”；

3. 自选其它强化学习案例并实现；

本实验选择解决多臂老虎机问题作为实验内容。

# 三、实验步骤与过程

1. **问题定义**

在多臂老虎机（*multi-armed bandit*，*MAB*）问题中，有一个拥有 *K* 根拉杆的老虎机，拉动每一根拉杆都对应一个关于奖励的概率分布 。我们每次拉动其中一根拉杆，就可以从该拉杆对应的奖励概率分布中获得一个奖励。我们在各根拉杆的奖励概率分布未知的情况下，从头开始尝试，目标是在操作 *T* 次拉杆后获得尽可能高的累积奖励。由于奖励的概率分布是未知的，因此我们需要在“探索拉杆的获奖概率”和“根据经验选择获奖最多的拉杆”中进行权衡。“采用怎样的操作策略才能使获得的累积奖励最高”便是多臂老虎机问题。

1. **问题数学化定义**

假设我们面对 *K* 台老虎机，每台老虎机的奖励分布都是未知的。目标是通过多个回合的选择，最大化累积奖励。具体数学化定义如下：

* 输入参数：
  + *K*：老虎机的数量。
  + *T*：允许的总拉杆次数。
  + ~*Reward Distribution*：第 *i* 台老虎机的奖励分布，通常假设 是独立的随机分布。
* 目标，最大化累积奖励：

其中：

* + ：表示第 *t* 次拉杆时选择的老虎机。
  + ：第 *t* 次操作的老虎机的奖励。
* 约束条件：
  + 每次操作只能选择一个拉杆。
* 其他重要变量：
  + 估计期望奖励

对于每台老虎机 *i*，它的奖励分布 是未知的，但可以通过多次拉动拉杆 *i* 来估计其期望奖励。每次操作拉动一根拉杆，需要更新对应老虎机的估计期望奖励。初始时，每台老虎机的估计期望奖励为，每拉动一次拉杆，对应的第 *i* 台老虎机的估计期望奖励为：

其中， 为第 *i* 台老虎机的估计期望奖励， 为第 *t* 次拉动第 *i* 台老虎机的奖励， 为第 *i* 台老虎机被操作总次数。

* + 累积懊悔

懊悔定义为拉动当前拉杆的动作 与最优拉杆的期望奖励差，累积懊悔即操作 *T* 次拉杆后累积的懊悔总量。优秀的算法通常使累积懊悔呈次线性增长，这表明随着时间增加，算法能够逐渐学习并稳定选择最优行动，从而减少平均懊悔。累积懊悔越低，说明算法在探索与利用的权衡中表现越优，学习效率越高，其决策质量也就越好。

* + 最佳行动占比

最佳行动占比用于评估策略在多次操作中选择最佳老虎机的频率，最佳行动占比越高，说明算法能够更频繁地识别并选择最优行动，体现出较强的学习能力和对环境的适应能力。理想情况下，随着时间的推移，优秀的算法会逐渐提高最佳行动占比，并趋近于 100%。

1. **解决方案**
2. 老虎机和解决方案的基类

定义了一个多臂老虎机，拉动其每根拉杆的奖励服从伯努利分布，即每次拉下拉杆有的概率获得的奖励为 1，有的概率获得的奖励为 0。奖励为 1 代表获奖，奖励为 0 代表没有获奖。

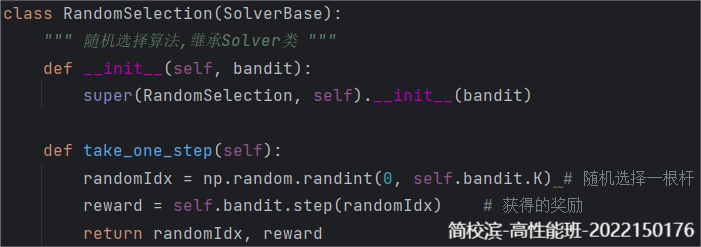
同时定义了一个多臂老虎机解决方案的基础框架，通过记录累积奖励、累积懊悔和最佳行动占比，评估算法在决策过程中的性能表现。

# 



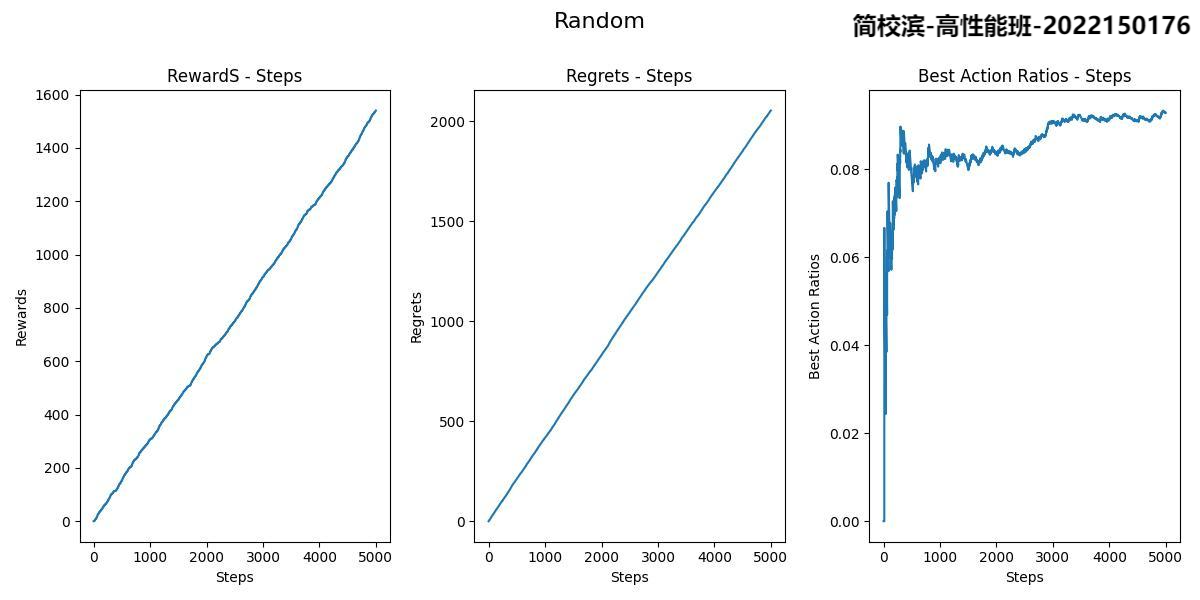
1. 随机选择算法

随机选择算法的思路是每次从所有拉杆中等概率随机选择一个进行操作，不依赖历史奖励信息或探索规律，仅通过纯随机的方式决策。



测试结果：

通过累积懊悔和最佳行动占比我们不难看出，随机选择算法未能准确识别并集中选择最优拉杆，因此取得了不佳的结果。



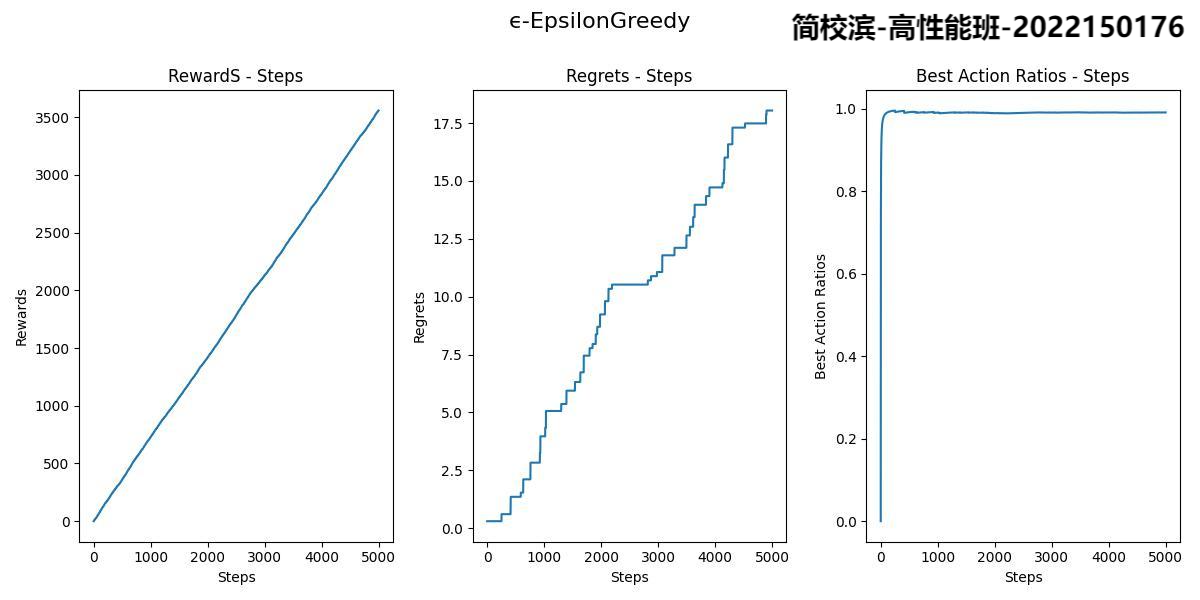
1. ϵ-贪心算法

ϵ-贪心算法为每次操作时，会选择拉动以往经验中期望奖励估值最大的拉杆，同时在选择的过程中加入扰动，以概率 ϵ 随机选择一根拉杆。具体而言，通过以一定概率 ϵ 随机选择拉杆（探索），尽可能尝试所有的拉杆，避免过早陷入局部最优；同时，以概率 1−ϵ 选择当前奖励估计最高的拉杆（利用），能够利用已有的知识获取更高的奖励。

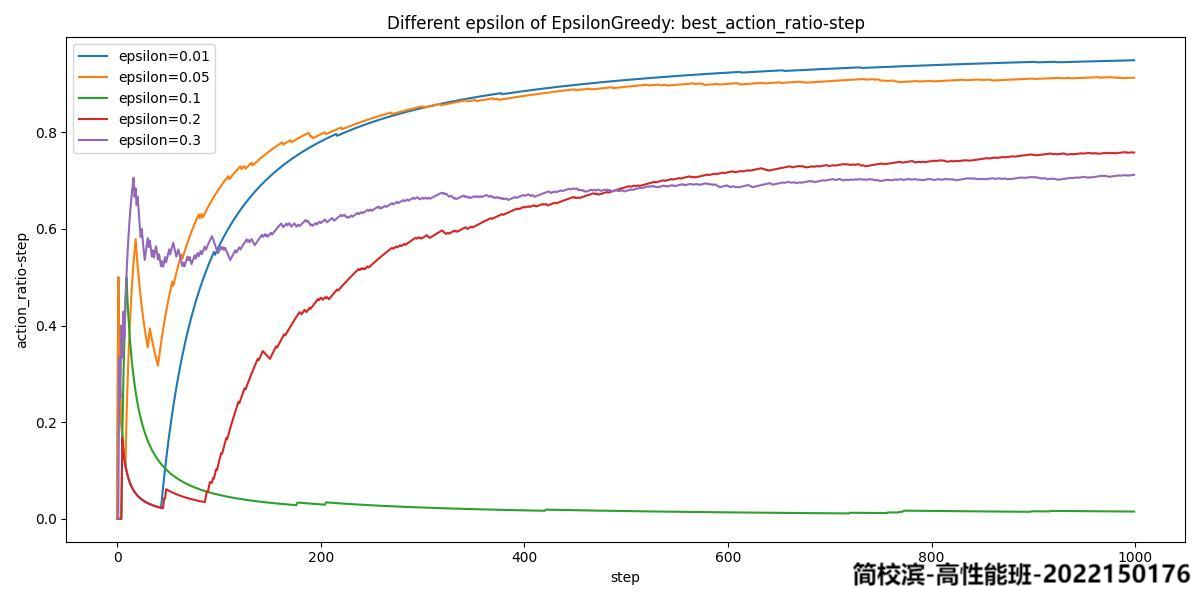


测试结果：

以下是ϵ=0.01的ϵ-贪心算法运行结果。结果表明，累积奖励随步骤稳定增长，说明ϵ-贪心算法能够有效选择高奖励拉杆；累积懊悔增长缓慢且呈阶梯状，表明随着时间推移，错误选择逐渐减少；最佳行动比例迅速趋近于1，表示ϵ-贪心算法在大多数情况下选择了最佳拉杆，体现了较强的收敛性和性能优越性。



这是不同的概率 ϵ 下最佳行动比例随操作步数的变化情况，可以观察到，较小的 ϵ（如 ϵ=0.01 和 ϵ=0.05）在初期迅速接近最佳行动，且最终稳定在较高的最佳行动比例，表明探索较少时更容易选择最优拉杆；较大的 ϵ（如 ϵ=0.2 和 ϵ=0.3）在初期因较多的探索导致最佳行动比例增长较慢，同时随机探索的概率大，不会优先选择收益最大拉杆，最终导致收敛到较低水平，说明过多的随机探索会影响选择最优拉杆的效率和稳定性。



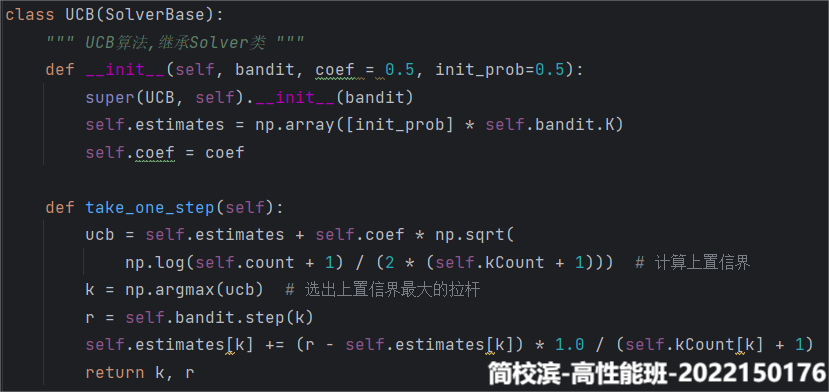
1. 上置信界算法

ϵ-贪心算法在进行尝试探索时是盲目地选择，因为它不大会选择接近贪心或者不确定性特别大的动作。在非贪心动作中，最好是根据它们的潜力来选择可能事实上是最优的动作，这要考虑它们的估计有多接近最大值，以及这些估计的不确定性。

上置信界算法（UCB）思想的选择动作依据如下：

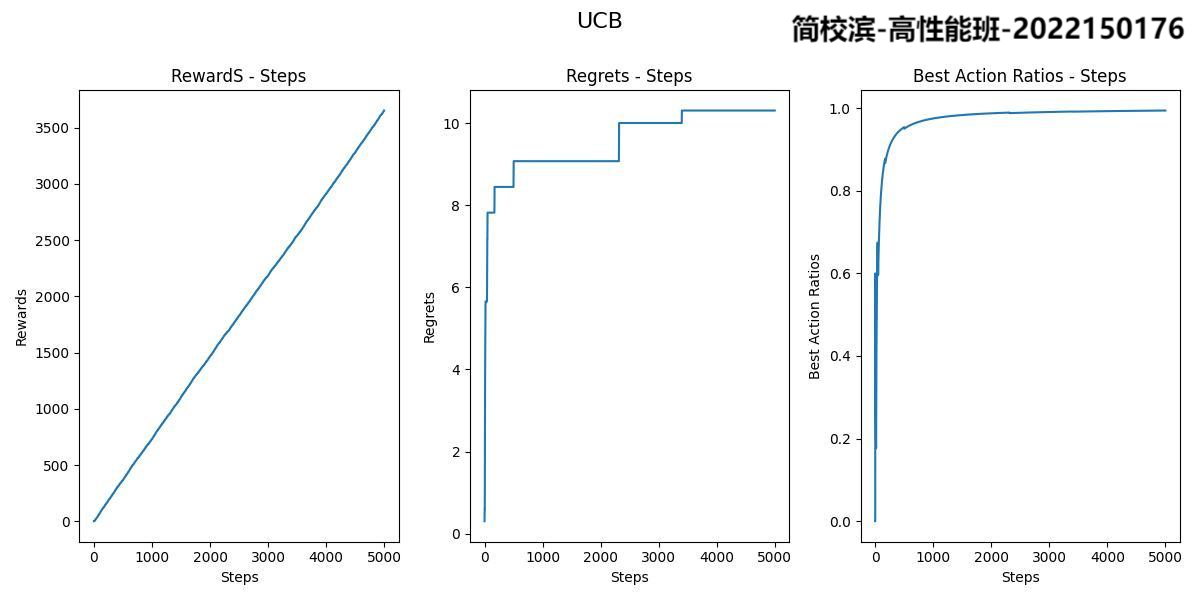
其中， 表示在时刻 *t* 之前动作 被选择的次数，*c* 用于控制试探的程度，*c*过大可能会过度探索，*c*过小可能会过度利用。该式子表明，通过综合考虑动作 的当前平均奖励和不确定性（探索项），选择具有最高上置信界值的动作。

更直观地说，上置信界算法在每次选择拉杆前，先估计每根拉杆的期望奖励的上界，使得拉动每根拉杆的期望奖励只有一个较小的概率超过这个上界，接着选出期望奖励上界最大的拉杆，从而选择最有可能获得最大期望奖励的拉杆。



测试结果：

以下是*c=*0.5的上置信界算法运行结果。结果表明，上置信界算法也能快速找到最有拉杆，以获得较高的累积奖励和较低的累积懊悔。



1. 汤普森采样算法

汤普森采样的核心思想是每次选择一个拉杆时，基于该拉杆的当前估计奖励分布进行采样，其在不完全了解环境的情况下可以快速找到最优解。具体而言：

* 首先为每个拉杆假设一个先验分布，这里采用 *Beta* 分布：
* 每次选择一个拉杆时，汤普森采样会从每个拉杆的奖励分布中进行一次采样，选择具有最大采样值的拉杆。

表示从拉杆 的奖励分布中采样得到的值。

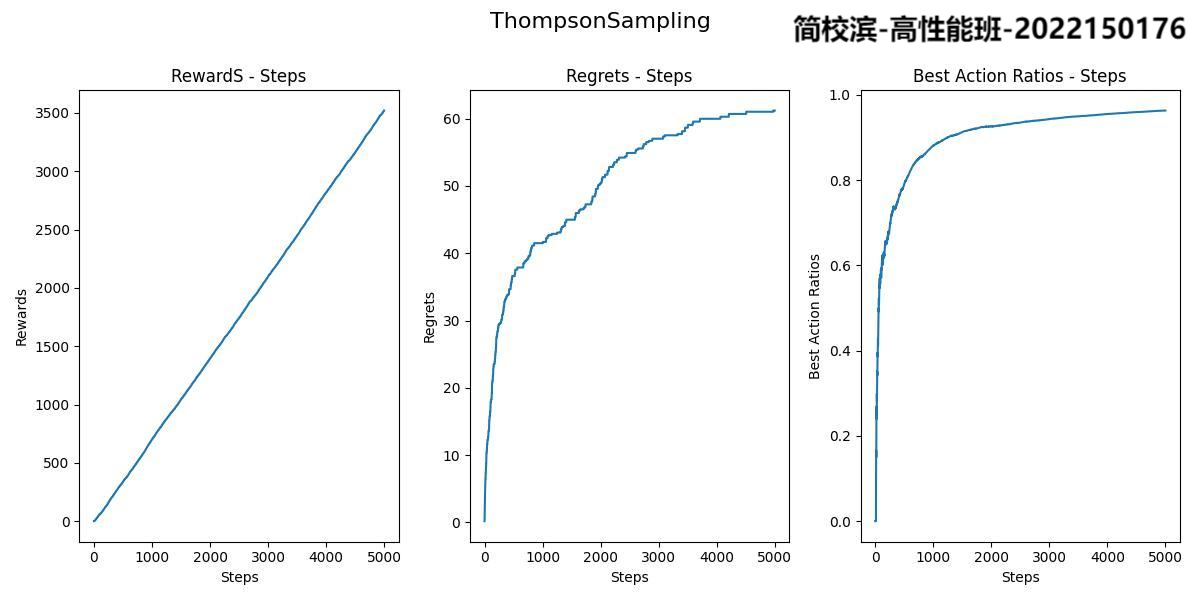
* 一旦选择了拉杆 ，更新其奖励分布：

通俗来讲，每次选择时，汤普森采样算法根据当前的奖励数据从每个选项的概率分布中随机采样一个值，然后选择采样值最大的选项。这样既能利用已有数据选择可能最优的选项，又能通过随机性探索不常被选择的选项，从而在长期内找到最优解并最大化累积奖励。



测试结果：

以下是使用 *Beta* 分布的汤普森采样算法运行结果。结果表明，汤普森采样算法也能快速找到最有拉杆，以获得较高的累积奖励和较低的累积懊悔。但相对于前两个算法，汤普森采样算法会有较高的累积懊悔，这是因为其在早期需要花费更多的步数探索各个拉杆的奖励分布，收集足够的数据来进行后续决策。



1. 梯度赌博机算法

梯度赌博机算法是一种通过优化策略选择拉杆的算法，它的核心思想是直接学习每个拉杆被选中的概率（策略概率），并根据奖励的反馈不断调整这些概率。具体来说，算法会使用梯度上升的方法，让表现好的拉杆的概率逐渐增加，而表现差的拉杆的概率逐渐减少。具体而言：

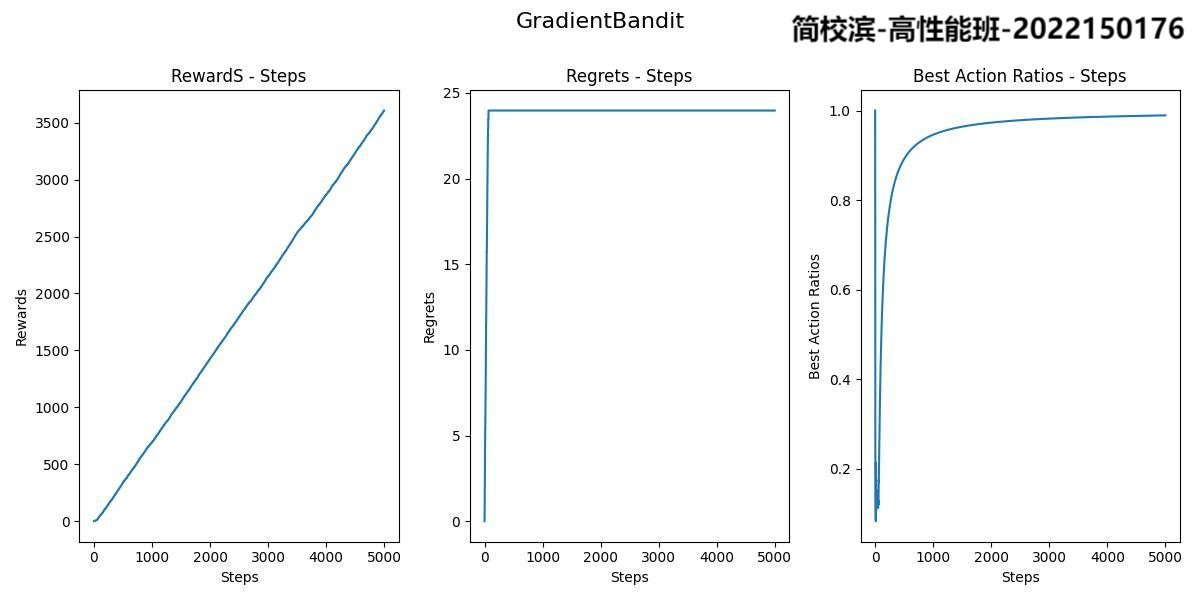
* 对每个拉杆初都有一个偏好值，通过 *Softmax* 函数根据偏好值计算每个拉杆的选择概率：
* 然后根据每个拉杆的选择概率选择一个拉杆进行尝试。
* 通过选择的拉杆获得奖励 *r* ，如果选择的拉杆 *k* 是获得奖励的拉杆且获得的奖励为1，则增加其偏好值，对其他未选择的拉杆，则减少其偏好值：

其中， 是学习率， 是获得奖励的拉杆， 是当前所有拉杆的平均奖励。



测试结果：

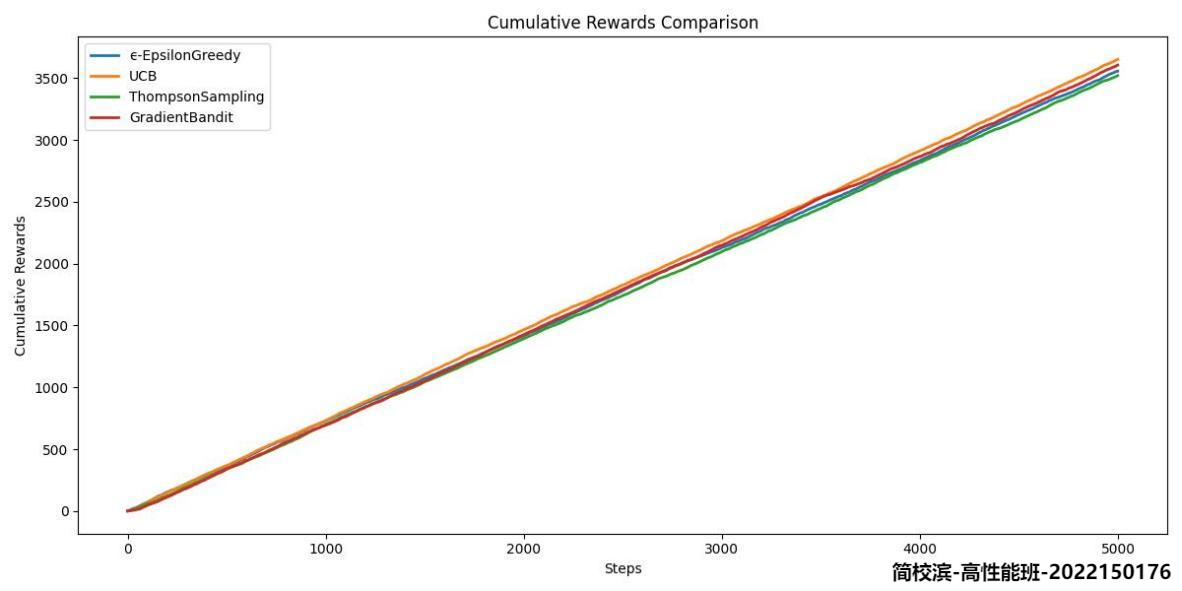
以下是使用 梯度赌博机算法运行结果。结果表明，梯度赌博机算法也能快速找到最有拉杆，以获得较高的累积奖励和较低的累积懊悔。

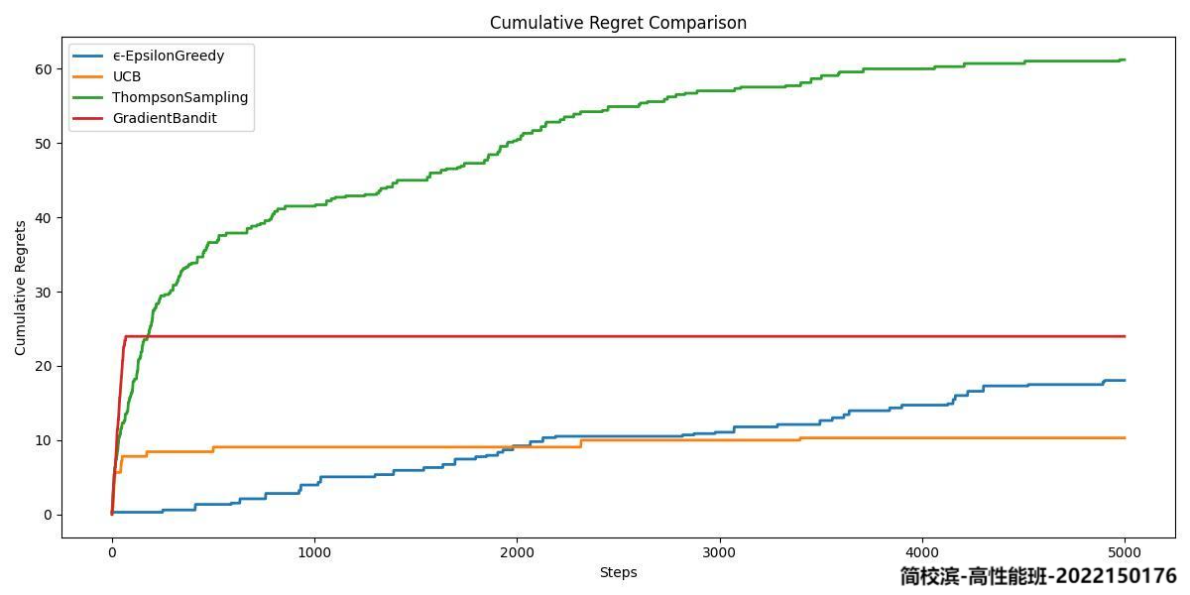


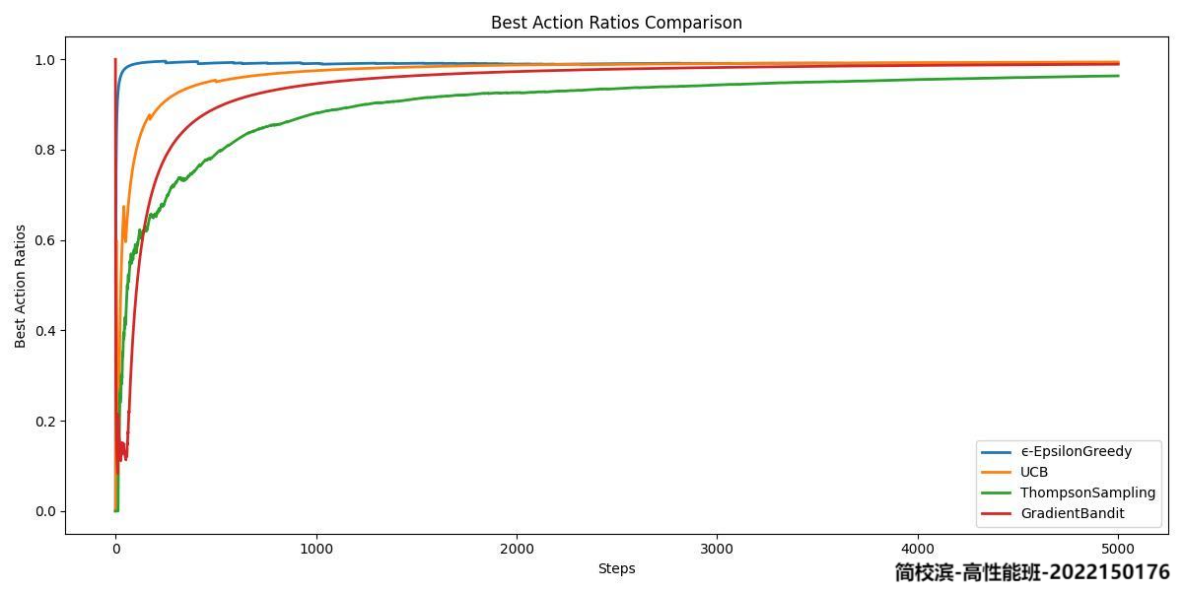
1. 各算法之间的对比

以下是ϵ-贪心算法、上置信界算法、汤普森采样算法、梯度赌博机算法在累积奖励、累积懊悔、最佳行动占比随步数变化的对比。

结果表明，各算法最终累积奖励基本一致，上置信界算法的累积懊悔最少，ϵ-贪心算法能最快找到最佳拉杆。







# 四、实验结论或体会

通过本次实验，我深入理解并实践了多臂老虎机问题的多种强化学习算法。实验结果表明，尽管所有算法在累积奖励上表现相近，但它们在累积懊悔和最佳行动占比上展现了不同的性能特点：

* 上置信界算法在累积懊悔方面表现最佳。
* ϵ-贪心算法在较小的ϵ值下能够迅速识别并选择最佳拉杆，显示出较强的学习能力。
* 汤普森采样算法则在平衡探索与利用方面表现出色，且性能稳定。
* 梯度赌博机算法虽然初期探索较为随机，但随着学习过程的推进，其选择最优拉杆的概率逐渐提高。

这些发现加深了我对强化学习中探索与利用权衡的理解，同时也让我体会到在不同情境下选择合适算法的必要性。

|  |
| --- |
| 指导教师批阅意见：  成绩评定：  指导教师签字：高灿  2024 年 12 月 24 日 |
| 备注： |

注：1、报告内的项目或内容设置，可根据实际情况加以调整和补充。

参考资料：

1. <https://hrl.boyuai.com/chapter/1/%E5%A4%9A%E8%87%82%E8%80%81%E8%99%8E%E6%9C%BA>
2. <https://leovan.me/cn/2020/05/multi-armed-bandit/>