Lecture 6: 多臂老虎机 (MAB) 可能考点整理

一、基本概念与核心问题 (简答题)

1. **多臂老虎机 (MAB) 问题是什么?**
    - **概念**: 多臂老虎机是一个经典的**在线决策问题**。一个决策者（玩家）面对 K 个选项（“臂”），在 T 轮试验中，每轮需要选择一个臂来拉动，并会获得一个随机的奖励。决策者不知道每个臂的真实奖励分布，其目标是通过一系列的选择，来最大化累积的总奖励。
    - **核心**: 问题的核心在于处理**不确定性**，并在**探索（Exploration）**和**利用（Exploitation）**之间做出权衡。
2. **什么是探索与利用的困境 (Explore-Exploit Dilemma)?**
    - **概念**: 这是 MAB 问题的根本性矛盾。
        - **探索 (Exploration)**: 尝试不同的、信息不足的臂，以期发现可能比当前最优臂更好的臂。这可能会牺牲短期收益，但有助于收集信息，避免陷入局部最优。
        - **利用 (Exploitation)**: 选择当前已知历史表现最好的臂，以最大化眼前的即时收益。过度利用可能导致错过真正的最优臂。
    - **结论**: 一个好的 MAB 算法必须能够有效地平衡这两者。
3. **MAB 问题的反馈类型有哪些?**
    - **性质**: 根据玩家在选择一个臂后能观察到的信息量，反馈类型可分为：
        - **完全反馈 (Full Feedback)**: 选择一个臂后，可以看到**所有臂**在本轮的奖励。例如，投资组合中，你可以看到所有股票当天的涨跌。
        - **部分反馈 (Partial Feedback)**: 选择一个臂后，除了该臂的奖励，还能看到**部分其他臂**的奖励信息。例如，动态定价中，若你出价 p 并成交，你也就知道了任何低于 p 的价格也都能成交。
        - **老虎机反馈 (Bandit Feedback)**: 最常见的类型。选择一个臂后，**只能看到被选择的那个臂**的奖励，对其他未被选择的臂一无所知。例如，新闻推荐中，你只知道用户是否点击了你推荐的那条新闻。
4. **如何衡量 MAB 算法的性能？什么是遗憾 (Regret)?**
    - **概念**: 遗憾是衡量 MAB 算法性能的核心指标，它量化了算法的决策与“事后诸葛亮”的最优决策之间的差距。
    - **定义 (期望遗憾)**: 算法在 T 轮中的期望遗憾 `R(T)` 定义为：在 T 轮中始终选择最优臂所能获得的总期望奖励，与算法实际获得的总期望奖励之差。

      $$E[R(T)] = T \cdot \mu^* - E[\sum_{t=1}^{T} \mu(a_t)]$$

      其中 μ^* 是最优臂的平均奖励，$\mu(a_t)$ 是在第 t 轮选择的臂 a_t 的平均奖励。
    - **结论/性质**:
        - 一个好的算法应该具有**次线性 (sub-linear)** 的遗憾，即 $\lim_{T\to\infty} \frac{E[R(T)]}{T} = 0$。这意味着随着时间的推移，算法的平均每轮遗憾会趋向于0，表明算法正在逐渐学习并收敛到最优选择。

二、随机多臂老虎机 (Stochastic MAB) 核心算法与结论

在随机 MAB 中，每个臂的奖励服从一个固定的、未知的概率分布。

1. **贪心算法 (Greedy Algorithm)**
   - **核心思想**: 极度侧重“利用”。在初始探索阶段（每个臂各尝试 N 次）之后，永久性地选择当前采样平均奖励最高的那个臂。
   - **结论/性质**: 这种算法有很大概率会陷入**局部最优**。如果初始探索阶段没能准确识别出真正的最优臂，它将永远无法纠正错误，从而导致**线性级别的遗憾**，是一种性能很差的算法。
2. **ε-贪心算法 (ε-Greedy Algorithm)**
   - **核心思想**: 对贪心算法的简单修正，引入了“探索”的概率。
     - 以 $1-\epsilon$ 的概率进行**利用**：选择当前采样平均奖励最高的臂。
     - 以 ϵ 的概率进行**探索**：从所有 K 个臂中随机选择一个。
   - **结论/性质**:
     - 通过参数 ϵ (一个 0 到 1 之间的小数) 来平衡探索和利用。
     - 它保证了每个臂都有持续被探索的机会，避免了永久陷入局部最优的问题。
     - **缺点**: 探索是“盲目的”，它以同样的概率探索所有臂，包括那些明显表现很差的臂，这造成了不必要的遗憾。
3. **上置信界算法 (Upper Confidence Bound, UCB)**
   - **核心思想**: **“面对不确定性时的乐观主义” (Optimism in the face of uncertainty)**。它为每个臂的平均奖励估计一个置信区间，并总是选择那个**置信区间上界**最高的臂。
   - **关键公式**: 在第 t 轮，为每个臂 a 计算一个 UCB 分数，并选择分数最高的臂：

$$\mathrm{UCB}_t(a) = \underbrace{Q_t(a)}_{\text{利用项}} + \underbrace{\sqrt{\frac{2\ln t}{N_t(a)}}}_{\text{探索项}}$$

     - $Q_t(a)$: 臂 a 在前 t-1 轮的**采样平均奖励**，代表已知的性能（利用）。
     - $N_t(a)$: 臂 a 在前 t-1 轮被**选择的次数**。
   - **结论/性质**:
     - 这是一个巧妙的平衡策略：如果一个臂的已知奖励 $Q_t(a)$ 很高，它的分数会高；如果一个臂被选择的次数 $N_t(a)$ 很少，分母变小，探索项就会变大，导致它的分数也可能很高，从而被“乐观地”选择。
     - UCB 实现了**智能探索**，它更倾向于探索那些有潜力成为最优臂（即不确定性大）的臂。
     - 理论上，UCB 具有非常好的遗憾上界，约为 $O(\sqrt{KT\log T})$。
4. **汤普森采样 (Thompson Sampling)**
   - **核心思想**: 一种**贝叶斯方法**，它将每个臂的平均奖励 $\mu(a)$ 视为一个随机变量，并为其维护一个**后验概率分布**。
   - **执行步骤**:
     1. 为每个臂的奖励概率维护一个先验分布（通常是 Beta 分布）。
     2. 在每一轮，从每个臂的**当前后验分布**中各抽取一个随机样本。
     3. 选择样本值最大的那个臂。

4. 观察获得的奖励，并使用贝叶斯法则更新被选择臂的后验分布。

- **结论/性质**:
  - **概率匹配 (Probability Matching)**: 一个臂被选中的概率，与其是当前最优臂的后验概率相匹配。
  - **实现方式**: 对于奖励为 0 或 1 的情况（伯努利分布），使用 Beta 分布作为后验。若一个臂有 S 次成功和 F 次失败，其后验分布就是 $\text{Beta}(S+1, F+1)$。
  - 在实践中，汤普森采样通常表现得比 UCB 更好或相当，且具有同样优秀的理论遗憾保证。

三、对抗性多臂老虎机 (Adversarial MAB)

在对抗性 MAB 中，奖励不再服从固定分布，而是由一个“对手”在每一轮动态生成，对手可能试图最大化玩家的损失。

1. **跟风算法 (Follow-The-Leader, FTL)**
   - **核心思想**: 在每一轮，选择到目前为止**历史累积代价最小**的臂。
   - **结论/性质**: 这是一个非常直观但**有缺陷**的算法。在一个聪明的对手面前，FTL 很容易被“戏耍”。例如，对手可以预测到 FTL 下一轮的选择，并为该选择设置高代价，导致 FTL 性能很差，产生线性遗憾。
2. **乘性权重更新算法 (Multiplicative Weights Update, MWU)**
   - **核心思想**: 维护一组权重，每个臂对应一个权重，所有权重的总和是固定的。在每一轮，根据臂的代价来**乘性地减少**其权重。表现好的臂（代价低）权重减少得慢，表现差的臂（代价高）权重减少得快。
   - **关键公式**: 权重更新规则：

     $$w_{t+1}(i) = w_t(i) \cdot (1 - \epsilon \cdot c_t(i))$$

     其中 $w_t(i)$ 是臂 i 在第 t 轮的权重，$c_t(i)$ 是其代价，ϵ 是学习率。
   - **结论/性质**:
     - MWU 是一种非常强大且通用的**无憾 (no-regret)** 算法，广泛应用于在线学习、博弈论等领域。
     - 它能保证在对抗环境下，其总代价与最好的单个固定策略（即始终选择某个臂）的总代价相比，遗憾是次线性的，约为 $O(\sqrt{T \ln K})$。