# 第 3 章 在线学习与探索利用平衡

**内容提要**

- 多臂老虎机
- 懊悔
- 先探索后利用算法
- $\varepsilon$-Greedy 算法
- 置信上界算法
- 汤普森采样
- 懊悔下界

## 3.1 在线学习的学习目标

在线学习是一种在数据不断产生过程中实时更新模型的学习范式，与传统的一次性处理固定数据集的批量学习方式有所不同。在线学习的特点使其特别适用于动态变化的场景，例如金融市场交易、推荐系统和机器人决策等。这些场景的一个共同特征是环境的复杂性和信息的不确定性，要求智能体能够实时适应变化。

在强化学习中，智能体在实时决策中往往面临一项核心挑战：如何在探索新行为（即尝试未知策略以获得更多信息）与利用已有经验（即选择当前被认为最优的策略以获取较高回报）之间取得平衡。如果仅依赖已有经验，可能会错失潜在的更优策略；而过度探索则可能导致短期内收益较低且策略收敛变慢。本章引入经典的多臂老虎机问题作为分析框架，并通过先探索后利用、$\varepsilon$-Greedy、置信上界、汤普森采样等算法，深入探讨探索利用平衡的实现机制和理论分析。

本章的学习目标主要包括：

1. 深入理解在线学习的概念、特点及其与传统学习方式的区别。尤其要理解多臂老虎机这一在线学习的经典场景，准确分析智能体在探索新行为和利用已有经验时面临的困境。
2. 熟练掌握先探索后利用、$\varepsilon$-Greedy、置信上界、汤普森采样等算法在多臂老虎机问题中的工作原理、优势与局限性，能够根据不同场景选择合适的算法。
3. 掌握对算法进行懊悔值分析的方法，初步具备运用在线学习与探索利用平衡相关知识解决实际问题的能力，能够将理论知识迁移到具体的项目或研究中。

## 3.2 先探索后利用算法

为了系统地探讨探索利用平衡问题，我们首先引入经典的**随机多臂老虎机**（Stochastic Multi-Armed Bandit）模型。

**定义 3.1 (随机多臂老虎机)**

随机多臂老虎机可以表示成一个三元组 $< K, \mathcal{A}, T >$，其中 $K$ 代表玩家可以拉动 $K$ 个不同的手臂，每个手臂产生的奖励是独立同分布（i.i.d.）的，$\mathcal{A}$ 则表示动作空间 $\{1, \ldots, K\}$，$T$ 表示该问题的时间长度。在每一时刻 $t$：

1. 玩家选择一个手臂 $k \in \mathcal{A}$，记 $\pi_t := k$；
2. 环境根据玩家的动作产生一个奖励 $X_k(t) \sim D_k$ 并把它展示给玩家；

这里的 $D_k$ 是臂 $k$ 的奖励分布，该奖励分布的期望为 $\mu_k$。 ♣

定义最优的奖励期望 $\mu^* := \max_{k \in \mathcal{A}} \mu_k$。玩家在随机多臂老虎机问题中的目标为最大化在时间长度 $T$ 内收集到的奖励和，即 $\sum_{t=1}^{T} X_{\pi_t}(t)$。我们引入懊悔（Regret）这一概念，来衡量当前收到的奖励与最优奖励的差距，定义为：

$$R_T := T\mu^* - \sum_{t=1}^{T} X_{\pi_t}$$

因此，最大化奖励和等价于最小化在时间长度 $T$ 内的懊悔 $R_T$。需要注意的是，$R_T$ 是一个随机变量，它依赖于每一时刻选择的动作，并可能受到奖励分布和算法随机性的影响。因此，在分析算法的**懊悔界**（Regret Bound）时，我们关注的是懊悔的期望值，即 $\mathbb{E}[R_T]$。

在多臂老虎机问题中，**先探索后利用**（Explore-Then-Commit, ETC）是最基本的一类探索-利用算法。其核心思想是，在初始阶段进行一段时间的探索，充分收集各个手臂的奖励信息；然后，在剩余时间里选择当前估计最优的手臂并始终使用它，从而最大化收益。

### 3.2.1 ETC 算法

ETC 算法分为探索和利用两个阶段。定义 $N_k(t) := \sum_{\tau=1}^{t} \mathbb{1}\{\pi_\tau = k\}$ 为手臂 $k$ 到 $t$ 时刻为止被选择的总次数，定义 $S_k(t) := \sum_{\tau=1}^{t} X_k(\tau)$ 为手臂 $k$ 到 $t$ 时刻为止的累计奖励，定义 $\hat{\mu}_k := \frac{S_k(t)}{N_k(t)}$ 为估计的奖励期望。算法1具体描述了此过程：在前 $NK$ 轮，算法尝试所有 $K$ 个手臂，并对每个手臂进行若干次采样，以估计其期望奖励。在剩余时间内，算法选择经验平均奖励最高的手臂，并在接下来的所有时间步中始终选择该手臂。

**Algorithm 1** 先探索后利用 (ETC)

**Input:** 手臂数 $K$, 总时间步长 $T$, 探索轮数 $m$

**阶段一：前 $m$ 轮探索**

**for** $t \leftarrow 1$ **to** $m$ **do**

  选择手臂 $k \leftarrow (t \bmod K) + 1$，并收到奖励 $X_k(t)$

  更新累计奖励 $S_k(t) \leftarrow S_k(t) + X_k(t)$，更新手臂 $k$ 的选择次数 $N_k(t) \leftarrow N_k(t) + 1$

  计算估计的奖励期望 $\hat{\mu}_k \leftarrow \frac{S_k(t)}{N_k(t)}$

**end**

**阶段二：剩余 $(T - m)$ 轮利用**

选择最优手臂：$\hat{k} \leftarrow \arg\max_k \hat{\mu}_k$

**for** $t \leftarrow m + 1$ **to** $T$ **do**

  选择手臂 $\hat{k}$

**end**

我们不失一般性地假设任意手臂 $k$ 在时刻 $t$ 产生的奖励 $X_k(t)$ 被限制在区间 $[0, 1]$ 之内。为

了衡量 $\hat{\mu}_k$ 和真实的 $\mu_k$ 之间的误差，我们引入霍夫丁不等式。

**引理 3.1 (霍夫丁不等式)**

令 $X_1, \ldots, X_n$ 为独立同分布的随机变量，$\mathbb{E}[X_i] = \mu$，且 $X_i \in [0,1]$, $\forall i = 1, \ldots, n$。那么

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geqslant \sigma\right) \leqslant 2\exp\left(-2n\sigma^2\right)$$

♡

下面给出算法1的懊悔界定理。

**定理 3.1**

对于 $1 \leq m \leq T$，算法1的懊悔界满足

$$\mathbb{E}[R_T] \leqslant \frac{m}{K}\sum_{k \leqslant K} \Delta_k + (T-m)\sum_{k \leqslant K} \Delta_k \exp\left(-\frac{m\Delta_k^2}{4K}\right)$$

♡

**证明** 对于 ETC 算法，由于其结构固定（前 $m$ 轮进行探索，之后完全利用最优估计手臂），我们可以将其懊悔拆分为两部分：

1. 探索阶段的懊悔：由于前 $m$ 轮探索时，算法未必选择了最优手臂，因此会产生一定的损失。
2. 利用阶段的懊悔：在 $m$ 轮探索后，算法可能选择了次优手臂，导致在剩余 $T-m$ 轮中积累一定的损失。

定义 $\Delta_k := \mu^* - \mu_k$ 为手臂 $k$ 相较于最优手臂 $k^*$ 的期望奖励差距。在探索阶段，ETC 算法会均匀地尝试所有 $K$ 个手臂，每个手臂被选中的次数为 $m/K$。因此，在此阶段产生的懊悔为：

$$\begin{aligned}\mathbb{E}[R_{\text{explore}}] &= m\mu^* - \sum_{k=1}^{K}\frac{m}{K}\mu_k \\ &= m\left(\mu^* - \frac{1}{K}\sum_{k=1}^{K}\mu_k\right) \\ &\leqslant \frac{m}{K}\sum_{k \leqslant K}\Delta_k\end{aligned}$$

在探索阶段结束后，ETC 选择经验平均奖励最高的手臂 $\hat{k}$. 但 $\hat{k}$ 可能不是最优手臂 $k^*$，如果 $\hat{k} \neq k^*$，则会在剩余的 $T-m$ 轮中产生额外的懊悔：

$$R_{\text{exploit}} = (T-m)(\mu^* - \mu_{\hat{k}})$$

为了评估 $\mu_{\hat{k}}$ 的偏差，我们需要分析 ETC 算法选择次优手臂的概率。由于 ETC 在探索阶段对每个手臂 $k$ 进行了 $\frac{m}{K}$ 次采样，其经验平均奖励 $\hat{\mu}_k$ 满足：

$$P(|\hat{\mu}_k - \mu_k| \geqslant \sigma) \leqslant 2\exp\left(-2\frac{m}{K}\sigma^2\right)$$

定义 $\hat{k} := \arg\max_k \hat{\mu}_k$ 为 ETC 选择的手臂。那么如果 $\hat{k} \neq k^*$，则意味着至少存在一个次优手臂 $k \neq k^*$ 使得 $\hat{\mu}_k > \hat{\mu}_{k^*}$。根据霍夫丁不等式，$\hat{\mu}_k$ 和 $\hat{\mu}_{k^*}$ 的误差概率分别为：

$$P(\hat{\mu}_{k^*} < \mu^* - \sigma) \leqslant \exp\left(-2\frac{m}{K}\sigma^2\right)$$

$$P(\hat{\mu}_k > \mu_k + \sigma) \leqslant \exp\left(-2\frac{m}{K}\sigma^2\right)$$

为了使得某个次优手臂 $k$ 的 $\hat{\mu}_k$ 超过 $\hat{\mu}_{k^*}$，至少需要：

$$\mu^* - \sigma < \mu_k + \sigma \tag{3.1}$$

即 $\sigma > \frac{\Delta_k}{2}$。因此，错误选择次优手臂的概率为：

$$P(\hat{k} \neq k^*) \leqslant \sum_{k \neq k^*} \exp\left(-\frac{m\Delta_k^2}{4K}\right)$$

由于利用阶段的懊悔 $R_{\text{exploit}}$ 仅在选择错误手臂时产生，我们可以得到：

$$\begin{aligned}\mathbb{E}[R_{\text{exploit}}] &\leqslant (T-m)\sum_{k \neq k^*} \Delta_k P(\hat{k} = k) \\ &\leqslant (T-m)\sum_{k \leqslant K} \Delta_k \exp\left(-\frac{m\Delta_k^2}{4K}\right)\end{aligned}$$

结合探索阶段和利用阶段的懊悔，我们有：

$$\begin{aligned}\mathbb{E}[R_T] &= \mathbb{E}[R_{\text{explore}}] + \mathbb{E}[R_{\text{exploit}}] \\ &\leqslant \frac{m}{K}\sum_{k \leqslant K} \Delta_k + (T-m)\sum_{k \leqslant K} \Delta_k \exp\left(-\frac{m\Delta_k^2}{4K}\right)\end{aligned} \tag{3.2}$$

类似于式 (3.2) 的结果被称为问题依赖型（problem-dependent）懊悔界，因为其上界显式地依赖于手臂之间的奖励差距 $\Delta_k$，即不同问题实例的具体结构会影响算法的懊悔表现。相比之下，问题无关型（problem-independent）懊悔界不依赖于特定的奖励分布，而是给出对所有可能问题都成立的通用上界。对于 ETC 算法，其问题无关型懊悔界为

$$O(T^{\frac{2}{3}}((K\log T)^{\frac{1}{3}}))$$

读者可以在习题中尝试推导该结果。

尽管 ETC 在理论上提供了明确的懊悔界，并且易于分析和实现，但其主要缺陷在于探索步数 $m$ 需要事先指定，难以适应不同的环境，尤其在奖励分布未知或动态变化的情况下表现较差。此外，与自适应策略相比，ETC 在大规模问题中往往表现不佳，因为它无法根据历史数据动态调整探索策略。总体而言，ETC 适用于较简单的决策问题，但在实际应用中，通常更倾向于使用更具灵活性的自适应探索方法。

## 3.3 $\varepsilon$-Greedy 算法

**$\varepsilon$-Greedy 算法**（$\varepsilon$-Greedy Algorithm）是一种最简单且直观的探索-利用策略，在**多臂老虎机问题**（Multi-Armed Bandit, MAB）中广泛应用。与 ETC（Explore-Then-Commit）不同，$\varepsilon$-Greedy 算法在整个决策过程中持续进行探索，而非仅限于前几轮。这使得它更具适应性，能够在动态环境下进行调整。

定义 $[K] := \{1, ..., K\}$ 为所有手臂的集合。算法2描述了 $\varepsilon$-Greedy 算法。其基本思想是，在每个时刻 $t$，以概率 $1-\varepsilon_t$ 选择当前估计奖励最高的手臂，以概率 $\varepsilon_t$ 从 $[K]$ 中以均匀分布随机选择一个手臂进行探索。其中，$\varepsilon_t$ 是一个超参数，通常取 $0 < \varepsilon_t < 1$，用于控制探索与利用的

平衡。当 $\varepsilon_t$ 取值较大时，算法会更多地探索，而当 $\varepsilon_t$ 取值较小时，算法更倾向于利用当前的最优估计。

**Algorithm 2** $\varepsilon$-Greedy

**Input:** 手臂数 $K$, 总时间步长 $T$, $\{\varepsilon_t\}_T$

**for** $t \leqslant T$ **do**

以 $\varepsilon_t$ 的概率选择手臂 $k \sim \text{Uniform}([K])$ // 随机探索（exploration）

以 $1-\varepsilon_t$ 的概率选择手臂 $\hat{k} = \arg\max_k \hat{\mu}_k$ // 贪心选择（exploitation）

收到奖励并更新 $\hat{\mu}_k, \forall k \leqslant K$

$t \leftarrow t+1$

**end**

$\varepsilon$-Greedy 算法的主要优点在于其持续探索机制，使其能够适应动态变化的奖励分布，不像 ETC 那样在利用阶段完全停止探索。此外，该算法实现简单，只需在贪心选择和随机探索之间进行概率选择，因此计算开销较低。同时，$\varepsilon$-Greedy 能够适应非平稳环境，特别是当 $\varepsilon$ 逐渐衰减（如 $\varepsilon_t = \frac{1}{t}$）时，可以更有效地应对奖励分布的变化。下面我们给出 $\varepsilon$-Greedy 算法的懊悔界定理。

**定理 3.2**

对于在时刻 $i$ 以 $\varepsilon_i$ 概率进行探索的 $\varepsilon$-greedy 算法，其懊悔界满足

$$\mathbb{E}[R_T] \leqslant \sum_{t=1}^{T} \varepsilon_t + 2T\sqrt{\frac{K\log(2T)}{2\sum_{i=1}^{T}\varepsilon_t}}$$

♡

**证明** 在前 $T$ 时刻的探索阶段中共有 $\sum_{t=1}^{T}\varepsilon_t$ 次探索行为出现，其中每个手臂被选中的平均次数为 $\frac{\sum_{t=1}^{T}\varepsilon_t}{K}$。定义一个好的事件 $\mathcal{G}_1$ 为

$$\mathcal{G}_1 := \left\{\forall k \in [K], |\hat{\mu}_k - \mu_k| \leqslant \sqrt{\frac{K\log(2T)}{2\sum_{t=1}^{T}\varepsilon_t}}\right\}$$

我们选择 $\sigma = \sqrt{\frac{K\log(2T)}{2\sum_{t=1}^{T}\epsilon_t}}$，根据引理3.1有

$$\begin{aligned} P(\mathcal{G}_1) &= P\left(|\hat{\mu}_k - \mu_k| \leqslant \sqrt{\frac{K\log(2T)}{2\sum_{t=1}^{T}\varepsilon_t}}\right) \\ &\geqslant 1 - \frac{1}{T} \end{aligned} \tag{3.3}$$

当 $T$ 较大时，$\mathcal{G}_1$ 会以较大概率发生。接下来我们只考虑 $\mathcal{G}_1$ 发生的情况下的懊悔，因为当 $\mathcal{G}_1$ 不发生时，懊悔界为 $O(\frac{1}{T}\cdot T)$。因此，$\neg\mathcal{G}_1$ 的懊悔界是 1，不会影响整体懊悔界的计算。

用 $k^*$ 表示奖励期望为 $\mu^*$ 的手臂，可能存在其他手臂 $k \neq k^*$，使得在探索阶段得到的平均奖励满足 $\hat{\mu}_k > \hat{\mu}_{k^*}$。因此，手臂 $k$ 在利用阶段被选择。由(3.3)可得：

$$\mu_k + \sqrt{\frac{K\log(2T)}{2\sum_{t=1}^{T}\varepsilon_t}} \geqslant \hat{\mu}_k > \hat{\mu}_{k^*} \geqslant \mu_{k^*} - \sqrt{\frac{K\log(2T)}{2\sum_{t=1}^{T}\varepsilon_t}}$$

于是，$\mu_{k^*} - \mu_k \leqslant 2\sqrt{\frac{K\log(2T)}{2\sum_{t=1}^{T}\epsilon_t}}$，那么在 $T$ 时刻的期望懊悔界满足：

$$\begin{aligned}
\mathbb{E}[R(t)] &\leqslant \mathbb{E}\left[\sum_{t=1}^{T}\left(\varepsilon_t + (1-\varepsilon_t)\times 2\sqrt{\frac{K\log(2T)}{2\sum_{t=1}^{T}\varepsilon_t}}\right)\right] \\
&\leqslant \mathbb{E}\left[\sum_{t=1}^{T}\left(\varepsilon_t + 2\sqrt{\frac{K\log(2T)}{2\sum_{t=1}^{T}\varepsilon_t}}\right)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\left(\varepsilon_t + 2\sqrt{\frac{K\log(2T)}{2\sum_{t=1}^{T}\varepsilon_t}}\right)\right] \\
&= \sum_{t=1}^{T}\varepsilon_t + 2T\sqrt{\frac{K\log(2T)}{2\sum_{t=1}^{T}\varepsilon_t}}
\end{aligned} \tag{3.4}$$

注意到(3.4)中 $\sum_{t=1}^{T}\varepsilon_t$ 与第一项正相关，与第二项负相关，可以令

$$\sum_{t=1}^{T}\varepsilon_t = 2^{\frac{1}{3}}T^{\frac{2}{3}}(K\log(2T))^{\frac{1}{3}}$$

得到最小化的懊悔界为 $O(T^{\frac{2}{3}}(K\log(T))^{\frac{1}{3}})$。

尽管 $\varepsilon$-Greedy 算法能够在整个决策过程中保持探索，并能通过衰减 $\varepsilon_t$ 改进其性能，但其探索策略较为固定，未能根据手臂的不确定性动态调整探索频率，导致探索效率较低。相比之下，下一章介绍的置信上界算法利用置信区间构建了一种基于不确定性驱动的自适应探索机制，从而能够更有效地平衡探索与利用。在下一章，我们将介绍置信上界算法，并分析其更优的懊悔界。

## 3.4 置信上界算法

**置信上界**（Upper Confidence Bound, UCB）算法是一种基于优化探索与利用平衡的多臂老虎机算法。与 $\varepsilon$-Greedy 算法不同，UCB 算法利用统计推断理论构建置信区间，以动态调整探索力度，从而更有效地平衡探索与利用。UCB 算法基于“乐观估计”的思想，即智能体在选择动作时，应当倾向于选择那些当前不确定性较大的手臂，而不仅仅是选择经验上回报最高的手臂。这种方法确保了智能体能够探索足够多次不同的手臂，以获得较精确的估计。

### 3.4.1 UCB1 算法

算法3具体描述了 UCB1 算法，其核心思想是在经验平均奖励的基础上添加一个置信界限，以补偿样本数量较少时的不确定性。手臂 $k$ 在时刻 $t$ 的置信上界定义为

$$\mathrm{UCB}_k(t) := \hat{\mu}_k + \sqrt{\frac{2\log T}{N_k(t)}}$$

其中，$\hat{\mu}_k$ 是手臂 $k$ 在过去选择过程中得到的平均奖励，而 $\sqrt{\frac{2\log T}{N_k(t)}}$ 是探索项（confidence bound），用于度量当前奖励估计的不确定性。当 $N_k(t)$ 较小时，置信界限较大，算法倾向于选择该手臂进行探索；随着 $N_k(t)$ 增大，置信界限逐渐减小，算法更多地利用经验估计选择最优的手臂。

**Algorithm 3** UCB1

**Input:** 手臂数 $K$，总时间步长 $T$

**for** $k \leftarrow 1$ **to** $K$ **do**
  选择手臂 $k$，收到奖励 $X_k(t)$
  $t \leftarrow t+1$
**end**
**for** $t \leftarrow K+1$ **to** $T$ **do**
  计算每个手臂的置信上界 $\mathrm{UCB}_k(t) \leftarrow \hat{\mu}_k + \sqrt{\frac{2\log T}{N_k(t)}}$, $\forall k \leqslant K$
  选择手臂 $k_t = \arg\max_k \mathrm{UCB}_k$
  收到奖励 $X_{k_t}(t)$ 并更新 $\hat{\mu}_{k_t}$ 和 $N_{k_t}(t)$
**end**

UCB1 算法的理论保证之一是其次线性（sublinear）懊悔界，这意味着随着时间步长 $T$ 的增加，其累积懊悔不会以线性增长，而是远小于 $O(T)$。该结果使 UCB1 成为理论上较优的多臂老虎机算法之一。

**引理 3.2 (懊悔分解引理 [91])**

对于任意策略 $\pi$ 和具有有限或可数动作集 $\mathcal{A}$ 的随机多臂老虎机环境 $\nu$，在时间跨度 $T \in \mathbb{N}$ 内，该策略的累积懊悔 $R_n$ 满足以下关系式：

$$\mathbb{E}[R_T] = \sum_{k \in \mathcal{A}} \Delta_k \mathbb{E}[N_k(T)]$$

♡

该引理将累积懊悔表示为各个手臂所造成的损失之和。具体来说，累积懊悔可以看作是每个手臂的次优间隙（sub-optimal gap）$\Delta_a$ 与该手臂被选择的期望次数 $\mathbb{E}[T_a(n)]$ 的加权和。该引理的证明请参考 [91]。下面我们给出 UCB1 算法的懊悔界定理。

**定理 3.3**

在随机多臂老虎机环境下，算法3的累积懊悔满足

$$\mathbb{E}[R_T] \leqslant \sum_{k:\mu_k<\mu^*} \frac{8\log T}{\Delta_k} + 3\sum_{k:\mu_k<\mu^*} \Delta_k$$

其中 $\Delta_k = \mu^* - \mu_k$ 表示手臂 $k$ 的次优间隙。

♡

**证明** 算法3 先轮流选择每个手臂一次，然后在剩余时间内选择 UCB 最高的手臂。定义一个好的事件 $\mathcal{G}_2$ 为

$$\mathcal{G}_2 := \left\{ \forall k \in [K],\ |\hat{\mu}_k - \mu_k| \leqslant \sqrt{\frac{2\log(T)}{N_k(t)}} \right\}$$

根据引理3.1有

$$P(\mathcal{G}_2) = P\left(|\hat{\mu}_k - \mu_k| \leqslant \sqrt{\frac{2\log(T)}{N_k(t)}}\right) \geqslant 1 - \frac{2}{T^4}$$

接下来我们考虑 $\mathcal{G}_1$ 发生的情况。

由于懊悔仅在选择次优手臂时产生，我们分析某个次优手臂 $k$ 最后一次被选择的时刻 $t$。在该时刻，我们有：

$$\mathrm{UCB}_k(t) \geqslant \mathrm{UCB}_{k^*}(t)$$

根据置信界的定义，$\mathrm{UCB}_{k^*}(t) \geqslant \mu_{k^*}$，即

$$\hat{\mu}_k + \sqrt{\frac{2\log T}{N_k(t)}} \geqslant \hat{\mu}_{k^*} + \sqrt{\frac{2\log T}{N_{k^*}(t)}} \geqslant \mu_{k^*} \tag{3.5}$$

我们定义手臂 $k$ 在时刻 $t$ 的置信下界为

$$\mathrm{LCB}_k(t) := \hat{\mu}_k - \sqrt{\frac{2\log T}{N_k(t)}}$$

那么 $\mu_k \geqslant \mathrm{LCB}_k(t)$，因此

$$\mu_k + 2\sqrt{\frac{2\log(T)}{N_k(t)}} \geqslant \hat{\mu}_k + \sqrt{\frac{2\log(T)}{N_k(t)}} \tag{3.6}$$

结合式(3.5)和式(3.6)，

$$\mu_k + 2\sqrt{\frac{2\log(T)}{N_k(t)}} \geqslant \mu_{k^*}$$

根据 $\Delta_k$ 的定义，

$$N_k(t) \leqslant \frac{8\log(T)}{\Delta_k^2} + 1$$

由引理3.2有：

$$\begin{aligned}
\mathbb{E}[R_T] &= \sum_{k \leqslant K} \Delta_k \mathbb{E}[N_k(T)] \\
&= \sum_{k \leqslant K} \Delta_k \mathbb{E}[N_k(T)] P(\mathcal{G}_1) + \sum_{k \leqslant K} \Delta_k \mathbb{E}[N_k(T)] P(\neg \mathcal{G}_1) \\
&\leqslant \sum_{k \leqslant K} \Delta_k \mathbb{E}[N_k(T)] + \sum_{k \leqslant K} \Delta_k T \frac{2}{T^4} \\
&\leqslant \sum_{k:\mu_k < \mu^*} \left(\frac{8\log T}{\Delta_k} + \Delta_k\right) + \sum_{k \leqslant K} \frac{2\Delta_k}{T^3} \\
&\leqslant \sum_{k:\mu_k < \mu^*} \frac{8\log T}{\Delta_k} + 3 \sum_{k:\mu_k < \mu^*} \Delta_k
\end{aligned}$$

### 3.4.2 UCB1 的优化变种

UCB1 可以进一步优化，以适应不同的环境，使得探索与利用的平衡更加精细化。在 UCB2 算法中，置信界限的增长速度被调整得更慢，使得算法在早期阶段能够更频繁地探索，但在后续阶段更快地聚焦于表现最优的手臂。这种调整可以提高利用阶段的效率，使得算法在长时间运行时的表现更加稳定。UCB-Tuned 进一步优化了 UCB1 通过自适应调整探索项来适应非平稳环境。传统的 UCB1 假设奖励分布是静态的，而 UCB-Tuned 通过估计奖励的方差来调整置信界限，使得算法在动态变化的环境中表现更优。另一种重要的优化变种是 KL-UCB，它利用 Kullback-Leibler（KL）散度来度量置信界限，从而提高算法的收敛速度。相比于标准的 UCB1，KL-UCB 通过更精确地建模奖励分布的不确定性，使得算法在信息获取与利用之间达到更优的权衡，在某些问题上能够获得更紧的懊悔界。

### 3.4.3 UCB 算法的优势与局限性

UCB 算法相较于 $\varepsilon$-Greedy 具有更好的探索策略，能够自适应地调整探索强度，从而在大多数情境下实现更低的累积懊悔。由于 UCB 通过置信界限调整探索行为，其探索策略是基于不确定性的，因而能够更有效地减少不必要的探索，使得算法能够更快地收敛到最优策略。然而，UCB 也存在一定的局限性。首先，UCB 计算量相对较高，每个时间步需要计算所有手臂的置信界限，这在大规模问题中可能会带来较大的计算开销。其次，UCB 主要适用于静态环境，即假设奖励分布不会随时间变化。然而，在许多实际应用中，例如在线广告推荐和金融投资，奖励分布可能是动态变化的。在这样的环境下，UCB 可能表现不佳，需要额外的机制来适应变化。因此，在实际应用中，UCB 及其优化变种通常需要结合具体场景进行调整，以适应不同的应用需求。

## 3.5 汤姆森采样法

**汤姆森采样**（Thompson Sampling, TS）是一种基于贝叶斯推理的多臂老虎机算法，能够在探索和利用之间取得良好的平衡。与 UCB 算法不同，汤姆森采样通过维护每个手臂的后验分布，并基于该分布进行采样，以决定下一步的选择。其基本思想是，每次决策时，根据当前的后验分布对每个手臂的潜在回报进行抽样，并选择回报最高的手臂进行拉动。这种方法天然地结合了探索与利用，使得探索的程度随着数据的积累而逐渐减少。

汤姆森采样的核心在于对每个手臂的奖励分布进行建模，通常假设奖励服从某种参数化分布，并通过贝叶斯更新规则不断调整参数。我们考虑**伯努利多臂老虎机**（Bernoulli bandits）。在开始具体介绍汤姆森采样算法之前，我们定义：

- $S_k(t)$：在时刻 $t-1$ 之前，对手臂 $k$ 拉动得到奖励 1 的次数（**成功次数**）。
- $F_k(t)$：在时刻 $t-1$ 之前，对手臂 $k$ 拉动得到奖励 0 的次数（**失败次数**）。

因此，对手臂 $k$ 的 Beta 共轭后验分布为

$$\text{Beta}\big(S_k(t)+1,\ F_k(t)+1\big),$$

在伯努利分布的情形下，Beta 分布被选为其共轭先验分布，因为 Beta 分布与伯努利分布的后验分布形式保持一致。假设某个手臂的奖励服从伯努利分布，其成功次数和失败次数分别记为 $S_k$ 和 $F_k$，那么该手臂的后验分布为 Beta($S_k+1, F_k+1$)。每次决策时，从该 Beta 分布中采样一个值，并选择具有最大采样值的手臂进行拉动。算法4具体描述了汤姆森采样的流程。

**Algorithm 4** 汤姆森采样

**Input:** 手臂数 $K$，总时间步长 $T$

**for** $t \leftarrow 1$ **to** $T$ **do**

  **for** 每个手臂 $i$ **do**

    从后验分布中采样 $\theta_k(t) \sim \text{Beta}(S_k(t)+1, F_k(t)+1)$

  **end**

  选择手臂 $k = \arg\max_\ell \theta_\ell(t)$ 并收到奖励 $X_k(t)$

  **if** $X_k(t) = 1$ **then**

    $S_k(t) \leftarrow S_k(t)+1$

  **else**

    $F_k(t) \leftarrow F_k(t)+1$

  **end**

**end**

在理论上，汤姆森采样能够实现次线性的累积懊悔界。我们不失一般性地假设 $\mu_1 > \ldots > \mu_K$。下面我们给出在伯努利多臂老虎机问题中汤姆森采样的懊悔界定理以及频率学派懊悔分析[4]。

**定理 3.4 ([4])**

对于 $K$ 臂随机多臂老虎机问题，汤姆森采样算法的期望懊悔为

$$\mathbb{E}[R_T] \leqslant (1+\epsilon)\sum_{k=2}^{K}\frac{\ln T}{d(\mu_k,\mu_1)}\Delta_k + O\left(\frac{K}{\epsilon^2}\right)$$

其中 $d(\mu_i,\mu_1) = \mu_i \log\frac{\mu_i}{\mu_1} + (1-\mu_i)\log\frac{(1-\mu_i)}{(1-\mu_1)}$。

**证明** 我们给出汤普森采样懊悔界的简要证明，更详细的证明请参考 [4, 72]。

定义 $n_k(t)$ 为手臂 $k$ 在时间 $t-1$ 之前的拉动次数，$k(t)$ 表示在时间 $t$ 拉动的手臂。对于每个手臂 $k$，我们将选择两个阈值 $x_k$ 和 $y_k$，使得 $\mu_k < x_k < y_k < \mu_1$。定义 $L_k(T) := \frac{\ln T}{d(x_k,y_k)}$，并且定义 $\hat{\mu}_k(t) := \frac{S_k(t)}{n_k(t)}$（当 $n_k(t)=0$ 时，定义 $\hat{\mu}_k(t)=1$）。定义 $E_k^{\mu}(t)$ 为事件 $\hat{\mu}_k(t) \leq x_k$，定义 $E_k^{\theta}(t)$ 为事件 $\theta_k(t) \leq y_k$。定义滤波 $\mathcal{F}_{t-1}$ 为时间 $t-1$ 之前的历史记录，即

$$\mathcal{F}_{t-1} = \{k(t), X_{k(t)}(w), k=1,\ldots,K, w=1,\ldots,t-1\},$$

其中 $X_k(t)$ 表示时间 $t$ 观察到的手臂 $k$ 的奖励。定义 $p_{k,t} := P(\theta_1(t) > y_k \mid \mathcal{F}_{t-1})$。

为了证明定理(3.4)，我们需要以下引理。

**引理 3.3 ([4])**

对于任意 $t \in [1, T]$ 且 $k \neq 1$,

$$P\left(k(t)=k, E_k^{\mu}(t), E_k^{\theta}(t) \mid \mathcal{F}_{t-1}\right) \leqslant \frac{(1-p_{k,t})}{p_{k,t}} P\left(k(t)=1, E_k^{\mu}(t), E_k^{\theta}(t) \mid \mathcal{F}_{t-1}\right)$$

♡

引理3.3证明了在任何先前执行历史的条件下，在当前步骤中选择任意次优手臂 $k$ 的概率可以被在当前步骤中选择最优臂的概率的线性函数来限定。

**引理 3.4 ([4])**

$$\sum_{t=1}^{T} P\left(k(t)=k, \overline{E_k^{\mu}(t)}\right) \leqslant \frac{1}{d(x_k, \mu_k)}+1$$

♡

引理3.4给出了事件 $E_k^{\mu}(t)$ 发生时，即经验均值 $\hat{\mu}_k(t)$ 低于阈值 $x_k$ 的条件下，算法选择次优臂 $k$ 的次数上界。该结果说明当估计值已经明显偏低时，算法误选次优臂的频率是受到严格限制的。

**引理 3.5 ([4])**

$$\sum_{t=1}^{T} P\left(k(t)=k, \overline{E_k^{\theta}(t)}, E_k^{\mu}(t)\right) \leqslant L_k(T)+1$$

♡

引理3.5给出了在经验均值和采样值均低于阈值的情况下，算法仍然选择次优臂 $k$ 的次数上界。该结果反映了当手臂的历史观测数量较多时，其采样值将集中于较低的均值附近，从而进一步降低被选择的可能性。

**引理 3.6 ([4])**

$\tau_i$ 表示第 $i$ 次选择手臂 1 的时间点，那么

$$\mathbb{E}\left[\frac{1}{p_{k,\tau_i+1}}\right] \leqslant \begin{cases} 1+\frac{3}{\Delta_k'}, & i<\frac{8}{\Delta_k'} \\ 1+\Theta\left(e^{-\Delta_k'^2 \frac{i}{2}}+\frac{1}{(i+1)\Delta_k'^2} e^{-D_k i}+\frac{1}{e^{\Delta_k'^2 i/4}-1}\right), & i \geqslant \frac{8}{\Delta_k'} \end{cases}$$

其中 $\Delta_k'=\mu_1-x_1, D_k=x_1 \log \frac{x_1}{\mu_1}+(1-x_1) \log \frac{1-x_1}{1-\mu_1}$ 。

♡

引理3.6证明了 $p_{k,\tau_i+1}$ 的期望的上界。注意到

$$\begin{aligned} \sum_{t=1}^{T} P(k(t)=k, \mathbb{E}_k^{\theta}(t), \mathbb{E}_k^{\mu}(t)) &\leqslant \sum_{t=1}^{T} \mathbb{E}\left[\frac{1-p_{k,t}}{p_{k,t}} \mathbb{1}\{k(t)=1, \mathbb{E}_k^{\theta}(t), \mathbb{E}_k^{\mu}(t)\}\right] \\ &\leqslant \sum_{t=0}^{T-1} \mathbb{E}\left[\frac{1-p_{k,T_i+1}}{p_{k,T_i+1}} \sum_{t=\tau_i+1}^{\tau_i+1} \mathbb{1}\{k(t)=1\}\right] \\ &\leqslant \mathbb{E}\left[\frac{1}{p_{k,\tau_i+1}}-1\right] \end{aligned}$$

因此，玩家选择手臂 $k$ 的总次数可以分解为

$$\begin{aligned}\mathbb{E}[N_k(T+1)] &= \sum_{t=1}^{T} P(k(t)=k) \\ &= \sum_{t=1}^{T} P(k(t)=k, \mathbb{E}_k^{\theta}(t), \mathbb{E}_k^{\mu}(t)) + P(k(t)=k, \overline{\mathbb{E}_k^{\theta}(t)}, \mathbb{E}_k^{\mu}(t)) + P(k(t)=k, \overline{\mathbb{E}_k^{\mu}(t)}) \\ &\leqslant \sum_{i=0}^{T-1} \mathbb{E}\left[\frac{1}{p_{k,\tau_{i+1}}} - 1\right] + L_k(T) + 1 + \frac{1}{d(x_k,\mu_k)} + 1 \\ &\leqslant \frac{24}{\Delta_k'^2} + \sum_{i=0}^{T-1} \Theta\left(e^{-\Delta_k'^2 \frac{i}{2}} + \frac{1}{(i+1)\Delta_k'^2} e^{-D_k i} + \frac{1}{e^{\Delta_k'^2 \frac{i}{4}} - 1}\right) + L_k(T) + \frac{1}{d(x_k,\mu_k)} + 2\end{aligned}$$

最后，通过引理3.2可以得到最终结果。

与置信上界算法相比，汤姆森采样通过对每个手臂的后验分布进行抽样，使得探索的程度能够根据历史数据的积累而自适应地调整，避免了人为设定探索参数的问题。由于汤姆森采样采用自适应探索，探索行为随着数据积累逐渐减少，而置信上界算法的探索策略则依赖于置信界的数学构造。这种自适应特性使得汤姆森采样在实践中更具优势，特别是在奖励分布未知或具有复杂结构的场景，例如推荐系统、医疗试验、广告投放和金融交易等领域。在推荐系统中，汤姆森采样可以根据用户的历史行为动态调整推荐内容，从而提高点击率和用户体验。在医疗试验中，它能够在保障患者安全的同时优化实验方案，使更多的试验组接受更有效的治疗。在广告投放领域，汤姆森采样可用于优化广告展示策略，动态调整广告曝光，以提升整体收益。此外，在金融交易和投资组合优化中，该方法通过贝叶斯更新进行自适应决策，提高投资回报。

## 3.6 在线学习的问题下界

在前几小节中，我们介绍了多种经典的多臂老虎机算法，包括先探索后利用、$\varepsilon$-Greedy、置信上界以及汤姆森采样。这些算法在不同环境下表现良好，并在理论上具有可证明的次线性累积懊悔界。然而，一个核心问题是，现有算法的性能是否已经达到最优，或者是否存在一个普适的下界，使得任何算法都无法突破。为了解答这一问题，我们需要研究在线学习的**最小懊悔界**（Lower Bound），即在任何策略下都必须满足的累积懊悔的理论极限。

在讨论下界之前，我们需要引入一些基础概念，这些概念不仅在下界分析中至关重要，同时也是信息论和统计学习中的核心工具。

**定义 3.2 (全变差距离 (Total Variation Distance))**

对于定义在样本空间 $\Omega$ 上的两个概率分布 $p, q$，它们的全变差距离（Total Variation Distance, TV 距离）定义为

$$d_{\mathrm{TV}}(p,q) = \sup_{A \subseteq \Omega} |p(A) - q(A)| \in [0,1]$$

**定义 3.3 (KL 散度 (Kullback-Leibler divergence))**

对于定义在样本空间 $\Omega$ 上的两个概率分布 $p, q$，它们的 Kullback-Leibler（KL）散度定义为

$$\mathrm{KL}(p\|q) = \sum_{x\in\Omega} p(x)\log\frac{p(x)}{q(x)}$$

♣

**引理 3.7 (Pinsker 不等式)**

对于定义在样本空间 $\Omega$ 上的两个概率分布 $p, q$，

$$2d_{\mathrm{TV}}(p,q)^2 \leqslant \mathrm{KL}(p\|q)$$

♡

为了简化分析，我们只考虑**高斯老虎机**（Gaussian Bandits），即所有手臂的奖励均服从均值为 $\mu_k$、方差为 1 的高斯分布 $\mathcal{N}(\mu_k, 1)$，其中 $k \in [K]$。在前文中，我们已经讨论了问题依赖型（instance-dependent）和问题无关型（instance-independent）的上界分析。接下来，我们将分别给出这两类情境下的懊悔下界，以探究在在线学习框架下，所有策略在最优情况下所能达到的性能极限。

### 3.6.1 问题独立型下界

我们有如下的**极小极大下界**（Minimax Lower Bound）定理。

**定理 3.5 (极小极大下界)**

设 $K > 1$ 且 $T \geqslant K-1$，对于任意策略 $\pi$，存在一个均值向量 $\mu = [\mu_k]_{1\leqslant k\leqslant K} \in [0,1]^K$ 使得

$$R_T \geqslant \frac{1}{27}\sqrt{(K-1)T}$$

♡

这一结果表明，任何在线学习算法都无法获得比 $\Omega(\sqrt{T})$ 更优的懊悔下界。我们继续给出定理3.5的简要证明，更详细的版本请参考 [32, 91]。

**证明** 首先，我们需要两个引理。

**引理 3.8 (散度分解)**

设 $\nu = (\mathbb{P}_1, \ldots, \mathbb{P}_K)$ 为一个 $K$ 臂老虎机模型的奖励分布，$\nu' = (\mathbb{P}'_1, \ldots, \mathbb{P}'_K)$ 为另一个 $K$ 臂老虎机模型的奖励分布。固定某个策略 $\pi$，令 $\mathbb{P}_\nu = \mathbb{P}_{\nu,\pi}$，$\mathbb{P}_{\nu'} = \mathbb{P}_{\nu',\pi}$，表示在策略 $\pi$ 下，由老虎机模型 $\nu$ 和 $\nu'$ 诱导出的 $T$ 轮交互概率测度。那么

$$\mathrm{KL}(\mathbb{P}_\nu\|\mathbb{P}_{\nu'}) = \sum_{k=1}^{K}\mathbb{E}_\nu[N_k(t)]\mathrm{KL}(\mathbb{P}_k\|\mathbb{P}'_k)$$

♡

**引理 3.9 (Bretagnolle-Huber 不等式)**

对于定义在样本空间 $\Omega$ 上的两个概率分布 $p, q$，

$$d_{\mathrm{TV}}(p,q) \leqslant \sqrt{1-e^{-\mathrm{KL}(p\|q)}} \leqslant 1-\frac{1}{2}e^{-\mathrm{KL}(p\|q)}$$

♡

考虑两个方差均为 1 的高斯老虎机 $\nu$ 和 $\nu'$，即 $\mathbb{P}_k = \mathcal{N}(\mu_k, 1)$，同时不失一般性地假设 $\mu_1^* > \mu_1 \geqslant \cdots \geqslant \mu_K$。$\Delta_k$ 为老虎机 $\nu$ 中手臂 $k$ 的次优间隙，$\Delta_k'$ 为老虎机 $\nu'$ 中手臂 $k$ 的次优间隙。设 $\mathbb{P}_\nu$ 为 $T$ 轮交互后得到的概率测度。根据引理3.8，我们定义 $i := \arg\min_{k>1} \mathbb{E}_\nu[T_k, T]$ 为选择次数最少的手臂。然后我们构造第二个老虎机 $\nu'$，使得

$$\mathbb{P}_k' = \begin{cases} \mathbb{P}_k, & k \neq i \\ \mathcal{N}(\mu_i + \lambda, 1), & k = i \end{cases},$$

其中 $\lambda > \Delta_i$。手臂 $i$ 在老虎机 $\nu'$ 中是最优手臂，设 $\mathbb{P}_{\nu'}$ 为相应的概率测度。

我们有

$$\begin{aligned} \mathrm{KL}(\mathbb{P}_\nu \| \mathbb{P}_{\nu'}) &= \mathbb{E}_\nu[N_k(T)] \cdot \mathrm{KL}(\mathbb{P}_i \| \mathbb{P}_i') \\ &\overset{(a)}{\leqslant} \frac{1}{K-1}\sum_{k=1}^{K} \mathbb{E}_\nu[N_k(T)] \cdot \mathrm{KL}(\mathcal{N}(\mu_i,1) \| \mathcal{N}(\mu_i+\lambda,1)) \\ &\leqslant \frac{1}{K-1} T \cdot \frac{\lambda^2}{2} \\ &\leqslant \frac{T\lambda^2}{2(K-1)} \end{aligned}$$

其中 (a) 是由于引理3.8。根据引理3.2，对于老虎机 $\nu$，由于 $i$ 是次优手臂，

$$\begin{aligned} R_T &= \sum_{k \neq 1} \Delta_k \mathbb{E}_\nu[N_k(T)] \\ &\geqslant \Delta_i \mathbb{E}_\nu[N_i(T)] \\ &\geqslant \frac{T\Delta_i}{2} \mathbb{P}_\nu(N_i(T) \leqslant \frac{T}{2}) \end{aligned} \tag{3.7}$$

对于老虎机 $\nu'$，由于 $i$ 是最优手臂，对于任意 $k \neq i$，

$$\begin{aligned} \Delta_k' &= \mu_i + \lambda - \mu_k \\ &= \lambda - (\mu_k - \mu_i) \\ &\geqslant \lambda - \Delta_i \end{aligned}$$

于是

$$\begin{aligned} R_T' &= \sum_{k \neq i} \Delta_k' \mathbb{E}_{\nu'}[N_k(T)] \\ &\geqslant \frac{T(\lambda - \Delta_i)}{2} \mathbb{P}_{\nu'}(N_i(T) < \frac{T}{2}) \end{aligned} \tag{3.8}$$

定义 $A := \{N_i(T) < \frac{T}{2}\}$，$\lambda := 2\Delta_i$，相加式(3.7)和式(3.8)有

$$\begin{aligned} R_T + R'_T &\geqslant \frac{T}{2}\min\{\Delta_i, \lambda - \Delta_i\}\left[\mathbb{P}_\nu(A) + \mathbb{P}_{\nu'}(A^c)\right] \\ &\overset{(b)}{\geqslant} \frac{T}{2}\min\{\Delta_i, \lambda - \Delta_i\}\frac{1}{2}e^{-\mathrm{KL}(\mathbb{P}_\nu \| \mathbb{P}_{\nu'})} \\ &\overset{(c)}{\geqslant} \frac{T}{2}\min\{\Delta_i, \lambda - \Delta_i\}\frac{1}{2}\exp\left(-\frac{2T\lambda^2}{(K-1)}\right) \\ &\geqslant \frac{T}{4}\Delta_i \exp\left(-\frac{2T\Delta_i^2}{(K-1)}\right) \end{aligned}$$

其中 (b) 和 (c) 是由于引理3.9。令 $\mu^* = \mu_1 = \Delta$ 且 $\mu_2, \ldots, \mu_n = 0$，令 $\Delta_i = \Delta = \sqrt{\frac{K-1}{4T}} \leqslant \frac{1}{2}$，于是

$$\begin{aligned} \max\{R_T, R'_T\} &\geqslant \frac{1}{2}(R_T + R'_T) \\ &\geqslant \frac{e^{-\frac{1}{2}}}{16}\sqrt{(K-1)T} \end{aligned}$$

### 3.6.2 问题依赖型下界

**定理 3.6 (问题依赖型下界 [87])**

考虑一个策略 $\pi$，如果对任意奖励分布集合 $\{\mathbb{P}_k\}_{1\leqslant k\leqslant K}$，且这些分布由一个实数参数索引，并且对任何具有次优间隙（sub-optimality gap）$\Delta_k > 0$ 的手臂 $k$ 以及任意 $\alpha > 0$，满足 $\mathbb{E}[R_T] = o(T^\alpha)$，则有

$$\liminf_{T\to\infty} \frac{R_T}{\log T} \geqslant \sum_{k:\Delta_k>0} \frac{\Delta_k}{\mathrm{KL}(\mathbb{P}_k \| \mathbb{P}^\star)}$$

其中 $\mathbb{P}^\star$ 是最优手臂的奖励分布。 ♡

因此，问题依赖型的懊悔下界为 $\Omega(\log T)$。

**证明** 我们仅考虑高斯老虎机。设定一个固定的次优手臂 $i \neq k^*$。我们构造第二个老虎机 $\nu'$，奖励分布为：

$$\mathbb{P}'_k = \begin{cases} \mathbb{P}_k, & k \neq i \\ \mathcal{N}(\mu_k + \lambda, 1), & k = i \end{cases}$$

其中 $\lambda > \Delta_i$，并且手臂 $i$ 在新的老虎机 $\nu'$ 中是最优手臂。

根据引理3.8，

$$\begin{aligned} R_T + R'_T &\geqslant \frac{T\min\{\Delta_i, \lambda - \Delta_i\}}{4}e^{-\mathrm{KL}(\mathbb{P}_\nu \| \mathbb{P}_{\nu'})} \\ &\geqslant \frac{T\min\{\Delta_i, \lambda - \Delta_i\}}{4}e^{\frac{-\lambda^2\mathbb{E}_\nu[N_i(T)]}{2}} \end{aligned} \tag{3.9}$$

由于证明问题依赖型下界只需对任意次优手臂 $k$ 证明

$$\liminf_{T\to\infty} \frac{\mathbb{E}_\nu[N_k(T)]}{\log T} \geqslant \frac{2}{\Delta_k^2}$$

整理(3.9)有，

$$\frac{\lambda^2}{2}\frac{\mathbb{E}_\nu[N_i(T)]}{\log T} \geqslant 1 + \frac{\log\min\{\Delta_i, \lambda-\Delta_i\}}{4\log T} - \frac{\log(R_T + R'_T)}{\log T}$$

由于对于任意 $\alpha > 0$，存在常数 $C_\alpha > 0$，

$$R_T + R'_T \leqslant C_\alpha T^\alpha$$

对所有 $T$ 取极限，

$$\begin{aligned}\limsup_{T\to\infty}\frac{\log(R_T + R'_T)}{\log T} &\leqslant \limsup_{T\to\infty}\frac{\alpha\log T + \log C_\alpha}{\log T}\\ &= \alpha\end{aligned}$$

由于该式对任意 $\alpha > 0$ 成立，因此：

$$\limsup_{T\to\infty}\frac{\log(R_T + R'_T)}{\log T} = 0$$

最终可以得到：

$$\liminf_{T\to\infty}\frac{\lambda^2}{2}\frac{\mathbb{E}_\nu[N_i(T)]}{\log T} \geqslant 1$$

取 $\lambda > \Delta_i$ 的下确界，即令 $\lambda$ 逼近 $\Delta_i$，从而得到最终的不等式。这表明，对于任何次优手臂 $i$，其期望选择次数的增长速度至少为 $\Omega(\log T)$。

## 3.7 本章小结

本章介绍了在线学习中的核心挑战之一探索与利用的平衡，并围绕多臂老虎机问题探讨了几种经典的在线学习算法，包括先探索后利用（ETC）、$\varepsilon$-Greedy、置信上界（UCB）以及汤姆森采样（Thompson Sampling）。这些算法各具特点，在不同的场景下表现出不同的优劣势。ETC 采用固定的探索阶段，虽然概念简单且易于分析，但无法适应环境的动态变化；$\varepsilon$-Greedy 在整个学习过程中持续进行探索，避免了 ETC 策略的过早收敛问题，但其探索策略较为固定，可能导致不必要的探索开销。相比之下，UCB 利用置信界的思想，使得探索策略可以自适应地调整，从而有效降低累积懊悔；汤姆森采样则基于贝叶斯推断，通过后验分布的更新实现探索与利用的动态平衡。

在分析在线学习算法的理论性能时，懊悔（regret）被用作衡量算法性能的核心指标。本章讨论了问题依赖型和问题无关型的懊悔上界，并进一步探讨了在线学习的最优懊悔下界。我们通过信息论工具，如 KL 散度、全变差距离及 Pinsker 不等式，推导出多臂老虎机问题的最优懊悔下界，并证明了极小极大懊悔下界为 $\Omega(\sqrt{T})$，而问题依赖型懊悔下界为 $\Omega(\log T)$。这一结果表明，所有策略在理论上都无法突破该下界，从而为多臂老虎机问题设定了性能的理论极限。

在实际应用中，探索与利用的权衡仍然是一个至关重要的问题。例如，在推荐系统中，如何在推广新内容和展示用户最喜爱的内容之间找到平衡；在医疗试验中，如何在探索新药的有效性和确保病人安全之间作出权衡；在金融投资中，如何在发现新的投资机会与最大化收益之间进行决策。多臂老虎机问题及其相关算法为这些问题提供了理论指导，而在线学习的进一步发展仍然需要结合具体应用，优化探索策略，以应对更加复杂的环境。

## 3.8 习题

1. 证明先探索后利用（Explore-Then-Commit, ETC）算法在随机多臂老虎机（Stochastic Multi-Armed Bandit）问题中的问题独立型懊悔界为 $O(T^{2/3}(K\log T)^{1/3})$
2. 实现本章介绍的先探索后利用（ETC）、$\varepsilon$-Greedy、置信上界（UCB1）、汤普森采样（Thompson Sampling）算法的 Python 代码，绘制懊悔增长曲线，比较不同算法的懊悔增长趋势，并讨论哪种算法在不同时间范围内表现较优。
3. 随机多臂老虎机问题中，Successive Elimination（SE）算法采用逐步淘汰的方法，在时间 $T$ 内迭代剔除置信下界（LCB）比其他任意手臂的置信上界（UCB）都小的手臂，最终收敛到最优手臂。
   (A). 尝试写出 SE 算法的伪代码，并给出懊悔界分析。
   (B). 实现 SE 算法的 Python 代码，并与 UCB1 算法对比。
   (C). 结合理论分析与实验结果，讨论 SE 和 UCB1 各自的优缺点，以及在不同应用场景下如何选择合适的算法。
4. 为什么 UCB 算法不依赖于具体的 reward 分布，而 Thompson Sampling 需要设定先验？这在实际中会带来哪些挑战？
5. 在前述算法中，我们均假设每个手臂的奖励分布是静态的，即在整个时间范围内保持不变。请思考：如果奖励分布随时间变化（non-stationary），现有的 UCB 算法可能失效。你将如何设计一个改进算法以适应这一情形？请说明你的设计思路。
6. 在本章中，我们讨论了汤普森采样（Thompson Sampling）算法的频率学派（Frequentist）懊悔界证明。而贝叶斯学派（Bayesian）采用后验更新的方式，通过假设奖励服从某个先验分布，并在每次交互后依据贝叶斯公式更新后验分布，从而对未知奖励进行估计。尝试证明汤普森采样在贝叶斯学派下的问题独立型懊悔界。