Best Arm Identification in Bandits with Limited Precision Sampling

Kota Srinivas Reddy*, P. N. Karthik†, Nikhil Karamchandani‡, and Jayakrishnan Nair‡
*Indian Institute of Technology Madras, † National University of Singapore, ‡ Indian Institute of Technology Bombay
Emails: ksvr1532@gmail.com, karthik@nus.edu.sg, nikhilk@ee.iitb.ac.in, jayakrishnan.nair@ee.iitb.ac.in

arXiv:2305.06082v1 [cs.LG] 10 May 2023

摘要——我们研究了多臂老虎机问题的一种变体中的最佳臂识别问题，其中学习者在臂选择方面的精度受到限制。学习者只能通过特定的探索组合（我们称之为盒子）对臂进行采样。具体而言，在每个采样时刻，学习者选择一个盒子，该盒子随后按照其特定的概率分布决定一个臂被拉取。被拉取的臂及其即时奖励将被揭示给学习者。学习者的目标是在错误概率不超过给定上界的约束下，通过最小化期望停止时间来找到最佳臂。我们提出了期望停止时间的一个渐近下界，该下界在错误概率趋于零时仍然成立。我们证明，由该下界所 ಸೂಚ示的最优分配通常并不唯一，因此难以进行跟踪。为处理最优分配不唯一的问题，我们提出了一种改进的基于跟踪的算法，并证明该算法具有渐近最优性。此外，我们还针对一种更简单的情形，给出了停止时间的非渐近下界和上界；在该情形下，一个盒子所能访问的臂与其他盒子所能访问的臂互不重叠。

I. Introduction

In this paper, we study best arm identification in a multi-armed bandit setting with K arms, where the learner has limited precision in sampling arms. In particular, the learner cannot directly sample individual arms, but can instead sample only certain exploration bundles, which we refer to as *boxes*. Each box is associated with a probability distribution over the arms; upon selecting a box, an arm is pulled randomly according to its corresponding probability distribution. The learner sees the pulled arm and its instantaneous reward. The learner's goal is to find the *best arm*, defined as the arm with the largest mean reward, while minimising the expected stopping time subject to an upper bound on the error probability (i.e., *fixed-confidence* regime). For the *boxed-bandit best arm identification* problem described above, our objective is to design sound algorithms and benchmark their performance against information theoretic lower bounds, when the arm reward distributions and the arm selection probabilities of the various boxes are apriori unknown.

A. Motivation

我们模型的一个关键特征是，学习者无法直接访问各个臂。相反，它必须通过某些中介（盒子）来进行探索，而这些中介在臂的选择上有各自的偏好、偏差或约束。举一个当代的例子，假设学习者的目标是通过.....识别某个大型群体（例如一个国家）中传染性最强的病毒或病原体毒株，在不同的本地检测机构进行排序测试。每个检测机构又通过对其本地周边地区或管辖范围内的个体进行抽样来开展测试；遇到不同菌株的可能性取决于检测机构所在的位置。

对我们的模型的另一种解释是，它刻画了臂选择中的噪声。例如，当学习者试图拉取某个臂时，只有以 $1-\eta$ 的概率才能成功拉取；以 η 的概率，要么没有臂被拉取（即发生擦除），要么拉取一个随机臂（这就是 [1] 的颤抖之手模型）。

最后，我们的模型也可以解释为学习者所进行的一种隐私保护实践。学习者通过盒子的非自适应选择配置进行探索，从而向其他观察者隐藏自身的偏好。当然，这种隐藏是以增加采样复杂度为代价的。在这种替代性解释下，假设学习者知道各个盒子的臂选择概率或许是合理的；我们的算法可自然地简化为这一特殊情形。

B. 分析挑战

请注意，在我们的问题设定中，学习者只能部分控制臂（通过盒子）。这不同于经典的最佳臂识别问题 [2]–[4]，在后者中，学习者可以在每个时刻完全控制要拉取哪个臂。例如，在 [5] 的逐步淘汰算法或 [6]，的 LUCB 算法中，学习者会在每个时刻拉取一个或多个臂，并通过淘汰次优臂，或即时判定最佳臂与次佳臂，最终找到最佳臂。在我们的设定中，由于一个臂可能通过多个盒子访问，而且事先并不知道各个盒子的臂选择概率，因此一开始并不清楚应该更频繁地选择哪个盒子，才能最大限度地提高拉取某个臂的机会。事实上，如果每个臂都属于每个盒子，且各个盒子的臂选择概率完全相同，那么所有随机化的盒子选择规则都会得到相同的期望停止时间。因此，淘汰类算法或 LUCB 类算法不能直接套用于我们的设定。

我们还注意到， [3], [4] 中出现的下界具有唯一的最优解（或最优分配），而这些研究中的最佳臂识别算法的一个关键方面是跟踪经验臂选择频率向最优分配的收敛情况。相比之下，我们证明，在我们的设定中，最优分配通常并不唯一。

在这种情况下，经验频率可能会在两个或多个最优分配之间交替变化，而长期来看并不收敛于任何一个最优分配。这凸显了改进现有基于跟踪的 [3], [4] 算法的必要性，使其能够处理非唯一的最优分配。

C. Contributions

我们推导了期望停止时间增长速率的渐近下界，其中渐近分析是在错误概率趋于零的情况下进行的。我们证明，该增长速率可由一个上确界-下确界优化问题刻画，而其最优（达到上确界的）解（或分配）可能并不唯一。受 Jedra 等人 [7], 的分析技术启发，我们提出了一种基于跟踪的算法，并对其进行了改进，使其能够在每个时间步处理非唯一分配的情况，同时证明了该算法能够渐近地达到这一下界。在可达性分析中，我们跟踪所有过去分配的经验平均的长期行为，并证明均值分配最终会趋近于正确的最优分配集合。我们的可达性分析有望应用于更一般的问题设置，例如像我们的工作中那样自然出现非唯一分配的情况，或者难以证明最优分配唯一性的情况；例如，参见 [1], [8] 中的相关说明。

最后，对于臂被划分到不同盒子中的特殊情形，即一个盒子可访问的臂集合与其他盒子可访问的臂集合互不重叠的情况，我们给出了连续消除算法的一种变体的非渐近保证。我们证明，该算法的期望停止时间满足一个上界，且该上界在未知实例参数方面是紧的。

D. Related Works

我们的设置与 [9], 中的设置非常相似，其中臂被划分为若干子集，这与我们的工作相同，并且这些集合之间可能存在重叠。然而，关键区别在于，在 [9], 中，学习者能够完全控制臂（而我们的工作中学习者只能部分控制臂）。此外，在 [9], 中，目标是在每个子集中找到最佳臂，而在我们的工作中，目标是找到整体最佳臂。论文 [10] 针对社区模式估计问题考虑了与我们类似的设置，其关键区别在于， [10] 中的分析和结果针对固定预算情形，而我们的工作针对固定置信度情形；关于这两种情形的比较，请参见 [3] 。当盒子的数量等于臂的数量，且每个盒子仅包含一个臂时，我们的设置便可特化为 [2], [3], [5], [6] 中的设置。我们的设置也可以特化为 [1] 中的颤抖之手模型；在该模型中，当学习者尝试拉取臂 k 时，该臂实际上以概率 $1-\eta$ 被拉取，而一个均匀随机选取的臂则以概率 η. 被拉取。

II. 问题表述与预备知识

我们考虑一个具有 K 个臂的多臂老虎机问题，其中各臂的标签为 $1, 2, \ldots, K$。臂 $k \in [K]$[1] 与一个奖励相关联。概率分布 $\nu_k \in \mathcal{G}$，其中 $\mathcal{G}$ 是已知的臂分布类别。令 μ_k 表示臂 k. 的平均奖励。学习者的目标是通过顺序采样识别最优臂，即平均奖励最高的臂。然而，与经典的多臂老虎机（MAB）设置不同，学习者无法直接对单个臂进行采样（又称拉取）。相反，在每个阶段，学习者会从有限个 M 盒子中选择一个盒子，这会导致某个臂按照依赖于盒子的概率分布被随机拉取。形式上，选择盒子 $m \in [M]$ 会使臂 $k \in \mathcal{A}_m \subseteq [K]$ 以概率$q_{m,k}$.被拉取。这里， $\mathcal{A}_m$ 表示使用盒子 m 可“访问”的臂集合（即 $q_{m,k} > 0$ 对于 $k \in \mathcal{A}_m$ 和 $\sum_{k\in\mathcal{A}_m} q_{m,k} = 1$）。重要的是， $\boldsymbol{q} \coloneqq \{q_{m,k} \colon k \in \mathcal{A}_m, m \in [M]\}$ 对学习者而言是先验未知的。注意， $\boldsymbol{q}$ 描述了学习者拉取特定臂的能力所存在的不精确性。事实上，刻画这种采样不精确性对最佳臂识别复杂度的影响，正是本文的主要关注点。

元组 $C = (\boldsymbol{q}, \boldsymbol{\nu})$ 完全确定了一个问题实例，其中 $\boldsymbol{\nu} = (\nu_k \colon k \in [K])$ 是臂分布向量。该问题实例对应的最优臂记为 $a^\star(C) = a^\star(\boldsymbol{\mu}) = \arg\max^{k\in[K]} \mu_k$，其中 $\boldsymbol{\mu} = (\mu_k \colon k \in [K])$。假设每个问题实例都有唯一确定的最佳臂。我们用 $\mathrm{ALT}(\boldsymbol{\mu})$ 表示与 $\boldsymbol{\mu}$ 不同的实例集合，即其最佳臂与 $a^\star(\boldsymbol{\mu})$ 不同的实例。当不存在歧义时，我们用 $C = (\boldsymbol{q}, \boldsymbol{\mu})$ 代替 $C = (\boldsymbol{q}, \boldsymbol{\nu})$。

对于 $t \in \{1, 2, \ldots\}$，令 B_t 表示学习者在时刻 t 选择的盒子。选择盒子 $B_t = m$ 后，以概率$q_{m,\ k}$拉取臂 $A_t = k$。学习者会观察到 A_t （即知道拉取了哪个臂）以及臂 A_t 的奖励 X_t。令 $(B_{1:t}, A_{1:t}, X_{1:t}) \coloneqq (B_1, A_1, X_1, \ldots, B_t, A_t, X_t)$ 表示截至时刻 t 所经历的盒子选择、臂拉取和观测历史记录。给定 $\delta \in (0, 1)$，学习者的目标是以最少的期望盒子选择次数（亦称期望停止时间）找出最佳臂，同时将停止错误概率控制在 δ 以下。

令 $\pi = \{\pi_t\}_{t=1}^{\infty}$ 表示任意一个通用的最佳臂识别策略（或算法）。对于每个 $t \geq 1$， π_t 将历史记录 $(B_{1:t}, A_{1:t}, X_{1:t})$ 映射到以下动作之一:

- 根据确定性或随机化规则选择盒子 B_{t+1} 。
- 停止并宣布估计的最佳臂。

令 τ_π 表示策略 π 下的（随机）停止时间，并令$\hat{a}$表示停止时的最佳臂估计。对于每个 $\delta \in (0,1)$，我们关注由 $\Pi(\delta) \coloneqq \{\pi \colon P(\hat{a} \neq a^\star(C)) \leq \delta \quad \forall C\}$. 定义的 δ-概率正确（δ-PC）算法类。本文设计 δ-PC 算法，并将其停止时间与信息论下界进行比较。

在第三节中，我们设计并分析了一种跟踪并停止式算法，该算法将臂分布类别 $\mathcal{G}$ 设为具有已知方差的高斯分布族。该算法被证明是 δ-PC，其期望停止时间也被证明在 $\delta \downarrow 0$. 时渐近最优。接下来，在第四节中，我们考虑模型的一种特殊情况，即集合 $\{\mathcal{A}_m \colon m \in [M]\}$ 两两不相交（即，臂是

[1]令 $[n] \coloneqq \{1, \ldots, n\}$ 对任意整数 $n \geq 1$。

划分到各个盒子中）。对于这种情况，我们令 $\mathcal{G}$ 为 1-次高斯分布族，设计并分析了一种逐步淘汰式算法，该算法具有非渐近（关于 δ）的停止时间界。

III. Track & Stop Based Algorithm

我们首先研究每个臂都可能关联多个盒子的一般设置。为便于说明，我们假设来自臂 k 的观测服从均值为 μ_k 、方差为 1 的高斯分布。不失一般性，我们针对每个臂都关联所有盒子的极端设置给出结果。令底层实例由 $\boldsymbol{q}_0 = \{q^0_{m,k} : \ m \in [M], \ k \in [K]\}$ 和 $\boldsymbol{\mu}_0 = \{\mu^0_k : k \in [K]\}$ 定义。我们首先给出期望停止时间增长率的渐近下界（当 $\delta \downarrow 0$ 时）。随后，我们着重说明分析非分区设置时的核心挑战：刻画该下界的优化问题存在非唯一的最优解。接着，我们提出一种受 [7] 启发的新型基于 Track-and-Stop 的算法，并证明尽管存在上述挑战，该算法仍具有渐近最优性。

A. 逆定理：渐近下界

本节的第一个主要结果如下：当 $\delta \downarrow 0$ 时，给出了 δ-PC 算法的期望停止时间极限增长率下界。

定理 1：设 $\boldsymbol{q}_0 = \{q^0_{m,k}\}_{m,k}$, $\boldsymbol{\mu}_0 = \{\mu^0_k\}_{k=1}^K$。则，

$$\liminf_{\delta \downarrow 0} \ \inf_{\pi \in \Pi(\delta)} \ \frac{\mathbb{E}[\tau_\pi]}{\log(1/\delta)} \geq \frac{1}{T^*(\boldsymbol{q}_0, \boldsymbol{\mu}_0)}, \tag{1}$$

其中，式 (1) 中的 $T^*(\boldsymbol{q}_0, \boldsymbol{\mu}_0)$ 定义为

$$T^*(\boldsymbol{q}_0, \boldsymbol{\mu}_0) = \sup_{w \in \Sigma_M} \ \inf_{\boldsymbol{\lambda} \in \mathrm{ALT}(\boldsymbol{\mu}_0)} \sum_{m=1}^{M} \sum_{k=1}^{K} w_m \, q^0_{m,k} \, \frac{(\mu^0_k - \lambda_k)^2}{2}, \tag{2}$$

其中，Σ_M 是盒子上所有概率分布（或分配）$w = (w_1, \ldots, w_M)$ 构成的单纯形。

为简洁起见，定理 1 的证明略去。这一标准证明对多臂老虎机采用测度变换论证 [11]、[3] 的 transportation lemma 以及 Wald 恒等式。

B. Non-Uniqueness of the Optimal Allocation

Consider the following simple example with $M = 2$ boxes and $K = 4$ arms. Suppose that $\boldsymbol{\mu}_0 = \{0.5, 0.4, 0.3, 0.3\}$. Notice that arm 1 is the best arm. Let $\boldsymbol{q}_0$ be specified by the following matrix:

$$\boldsymbol{q}_0 = \begin{pmatrix} 0.3 & 0.3 & 0.3 & 0.1 \\ 0.3 & 0.3 & 0.1 & 0.3 \end{pmatrix}.$$

The first row of the above matrix represents the arm selection probabilities of box 1, and the second row that of box 2. For this example, it is easy to show that every $w = (w_1, w_2) \in \Sigma_2$ attains the supremum in (2) and hence is an optimal allocation.

上述示例表明，最优分配可能并不唯一。这与先前的工作 [3], [4] 形成对比，在那些工作中最优分配是唯一的。令 $\mathcal{W}^\star(\boldsymbol{q}_0, \boldsymbol{\mu}_0)$表示在（2）中达到上确界的所有分配构成的集合。更一般地，令 $\mathcal{W}^\star(\boldsymbol{q}, \boldsymbol{\mu})$表示与实例（$\boldsymbol{q}, \boldsymbol{\mu}$）对应的最优分配集合。

Lemma 1: The mapping $(\boldsymbol{q}, \boldsymbol{\mu}) \mapsto \mathcal{W}^\star(\boldsymbol{q}, \boldsymbol{\mu})$ is upper-hemicontinuous and compact-valued. Furthermore, $\mathcal{W}^\star(\boldsymbol{q}, \boldsymbol{\mu})$ is convex for all $(\boldsymbol{q}, \boldsymbol{\mu})$.

In particular, Lemma 1 implies that $\mathcal{W}^\star(\boldsymbol{q}_0, \boldsymbol{\mu}_0)$ is convex; this, we shall see, will play an important role in the design of an asymptotically optimal algorithm, which forms the content of the next section.

C. Achievability: Handling Non-Unique Allocations

先前研究中的最佳臂识别算法及相应的可达性分析的一个关键特征[3], [4] 是跟踪，即臂拉取的经验频率几乎必然收敛于最优分配。当最优分配并非唯一时（如本文所研究的情形），长期来看，经验频率可能在两个或多个最优分配之间交替变化，而不收敛于其中任何一个最优分配。在这种情况下，建立跟踪性质将十分困难。这凸显了改进现有基于跟踪的算法的必要性，以处理潜在的非唯一最优分配，并建立一个能够证明类似跟踪行为的框架。本节将介绍一种能够实现这一目标的技术。

令 $N(t, m, k)$ 表示截至时间 t，盒子 m 被选中且臂 k 被拉取的总次数。定义 $N(t, m) = \sum_k N(t, m, k)$ 表示截至时间 t，盒子 m 被选中的次数；$N_k(t) = \sum_m N(t, m, k)$ 表示截至时间 t，臂 k 被拉取的次数。对于所有 m, k，令

$$\hat{q}_{m,k}(t) = \frac{N(t, m, k)}{N(t, m)}, \quad \hat{\mu}_k(t) = \frac{1}{N_k(t)} \sum_{s=1}^{t} \mathbf{1}_{\{A_s = k\}} \, X_s, \tag{3}$$

be the empirical estimates of the unknown parameters at time t. The key result (inspired from [7]) that enables us to prove achievability while working with a set of optimal allocations, is stated below.

Lemma 2: Let $f(t) = \frac{\sqrt{t}}{\sqrt{M}}$. Let $\{w(t)\}_{t=1}^{\infty} \subset \Sigma_M$ be any sequence such that $w(t+1) \in \mathcal{W}^\star(\hat{\boldsymbol{q}}(t), \hat{\boldsymbol{\mu}}(t))$ for all t. Let $i_0 = 0$ and

$$i_{t+1} = (i_t \bmod M) + \mathbf{1}_{\left\{\min_{m \in [M]} N(t,m) < f(t)\right\}}, \quad t \geq 0.$$

Then, under the *modified D-tracking rule* given by

$$B_{t+1} = \begin{cases} i_t, & \min_{m \in [M]} N(t, m) < f(t), \\ b_t, & \text{otherwise}, \end{cases} \tag{4}$$

where $b_t = \arg\min_{m \in \mathrm{supp}(\sum_{s=1}^{t} w(s))} N(t, m) - \sum_{s=1}^{t} w_m(s)$, we have

$$\lim_{t \to \infty} d_\infty((N(t, m)/t)_{m \in [M]}, \ \mathcal{W}^\star(\boldsymbol{q}_0, \boldsymbol{\mu}_0)) = 0 \quad \text{a.s..} \tag{5}$$

注意，(4) 中的抽样规则在开始时会对每个盒子各抽样一次。在时间 t，如果任意一个盒子的抽样不足，即 $\min_m N(t, m) < f(t)$，则会按照轮转方式强制对盒子进行抽样，直到

Algorithm 1 Boxed-Bandit Modified Track-and-Stop

Input: $\delta \in (0,1)$, $\rho > 0$, $K \in \mathbb{N}$, and $M \in \mathbb{N}$
Output: $\hat{a} \in [K]$ – best arm
Initialisation: $t = 0$, $\hat{\mu}_a(t) = 0 \, \forall a \in [K]$, $Z(0) = 0$.
Compute $\hat{a}(t) = \arg\max_a \hat{\mu}_a(t)$.
if $Z(t) \geq \zeta(t, \delta, \rho)$ and $\min_k N_k(t) > 0$ **then**
  Stop box selections.
  return $\hat{a} = \arg\max_{a \in [K]} \hat{\mu}_a(t)$
else
  Select box B_{t+1} as per *modified D-tracking rule* (4).
  Update $\hat{\boldsymbol{q}}(t)$, $\hat{\boldsymbol{\mu}}(t)$, and $Z(t)$. Go to step 1.
end if

$N(t,m) = \Omega(\sqrt{t})$ 对所有 m 均成立。否则，根据分配 $\{w(s)$ 对一个盒子进行采样：$1 \leq s \leq t\}$. 这里，$w(s)$ 是 $\mathcal{W}^\star(\hat{\boldsymbol{q}}(s-1), \hat{\boldsymbol{\mu}}(s-1))$ 中的任意一个元素，该集合表示时刻 $s-1$. 对应于估计的实例参数的最优分配。这就是确定性等价原则。在证明中，我们证明截至时刻 t 的所有分配的经验平均值，$\bar{w}(t) = 1/t \sum_{s=1}^{t} w(s)$，当 $t \to \infty$ 时趋近于 $\mathcal{W}^\star(\boldsymbol{q}_0, \boldsymbol{\mu}_0)$。由于对各个盒子进行强制探索，$(\hat{\boldsymbol{q}}(t), \hat{\boldsymbol{\mu}}(t))$ 当 $t \to \infty$ 时几乎必然趋近于 $(\boldsymbol{q}_0, \boldsymbol{\mu}_0)$。结合引理 1 中 $\mathcal{W}^\star$ 的上半连续性，这意味着对于足够大的 t，几乎必然有 $\mathcal{W}^\star(\boldsymbol{q}(t)\ \boldsymbol{\mu}(t))$ 与 $\mathcal{W}^\star(\boldsymbol{q}_0, \boldsymbol{\mu}_0)$ "足够接近"，由此我们得到 (5)。

Let $Z_{a,b}(t)$ denote the *generalised likelihood ratio test statistic* between arms $a, b \in [K]$ up to time t, defined as

$$Z_{a,b}(t) := \log \frac{\sup_{(\boldsymbol{q},\boldsymbol{\mu}):\mu_a \geq \mu_b} P(B_{1:t}, A_{1:t}, X_{1:t})}{\sup_{(\boldsymbol{q},\boldsymbol{\mu}):\mu_a \leq \mu_b} P(B_{1:t}, A_{1:t}, X_{1:t})}. \quad (6)$$

The next result provides an explicit expression for $Z_{a,b}(t)$.

Lemma 3: Fix $a, b \in [K]$ and a policy π. Fix t such that $\min_{k \in [K]} N_k(t) > 0$ a.s.. If $\hat{\mu}_a(t) \geq \hat{\mu}_b(t)$, then

$$Z^{\pi}_{a,b}(t) = N_a(t) \frac{(\hat{\mu}_a(t) - \hat{\mu}_{a,b}(t))^2}{2} + N_b(t) \frac{(\hat{\mu}_b(t) - \hat{\mu}_{a,b}(t))^2}{2}, \quad (7)$$

where $\hat{\mu}_{a,b}(t)$ is defined as

$$\hat{\mu}_{a,b}(t) := \frac{N_a(t)}{N_a(t) + N_b(t)} \hat{\mu}_a(t) + \frac{N_b(t)}{N_a(t) + N_b(t)} \hat{\mu}_b(t). \quad (8)$$

If $\hat{\mu}_a(t) \leq \hat{\mu}_b(t)$, then $Z^{\pi}_{a,b}(t) = -Z^{\pi}_{b,a}(t)$. Thus, $Z^{\pi}_{a,b}(t) \geq 0$ if and only if $\hat{\mu}_a(t) \geq \hat{\mu}_b(t)$.

令 $Z(t) = \max_a \min_{b \neq a} Z_{a,b}(t)$。对于固定的 $\delta \in (0, 1)$ 和 $\rho > 0$，令 $\zeta(t, \delta, \rho) := \log(C\, t^{1+\rho}/\delta)$，其中 C 是一个满足 $\sum_{t=1}^{\infty} \frac{e^{K+1}}{K^K} (\log^2(C\, t^{1+\rho}) \log \overline{t^K} \leq C)$ 的常数。

最佳臂识别算法：我们提出了 Garivier 等人的跟踪并停止算法的一种变体，称为盒式多臂赌博机修正追踪与停止算法（BBMTS），该算法经过改进，可在每个时间步处理一组分配。我们的算法以两个参数作为输入：$\delta \in (0, 1)$ 和 $\rho > 0$。在每个时刻 t，算法都会维护对最佳臂的估计值 $\hat{a}(t) \in \arg\max_a \hat{\mu}_a(t)$，当出现并列情况时，采用均匀随机方式进行决胜。引理 3 表明，$\hat{a}(t) \in \arg\max_a \min_{b \neq a} Z_{a,b}(t)$。随后，算法检查若 $\min_{b \neq \hat{a}(t)} Z_{\hat{a}(t),b}(t) \geq \zeta(t, \delta, \rho)$，则算法有足够的把握认为 $\hat{a}(t)$ 是最佳臂；此时算法停止，并输出 $\hat{a}(t)$ 作为最佳臂。否则，算法将按照前文所述的修正 D-追踪规则 (4) 对盒子 B_{t+1} 进行采样。伪代码见算法 1。

Performance analysis of BBMTS: Let $\pi_{\text{BBMTS}}(\delta, \rho)$ symbolically represent the BBMTS algorithm with parameters δ, ρ. The below result characterises its performance.

Theorem 2: The BBMTS algorithm meets the following performance criteria.

1) $\pi_{\text{BBMTS}}(\delta, \rho) \in \Pi(\delta)$ for each $\delta \in (0,1)$ and $\rho > 0$.
2) For each $\rho > 0$, the stopping time of $\pi_{\text{BBMTS}}(\delta, \rho)$ satisfies
$$\limsup_{\delta \downarrow 0} \frac{\tau_{\pi_{\text{BBMTS}}(\delta,\rho)}}{\log(1/\delta)} \leq \frac{1+\rho}{T^*(\boldsymbol{q}_0, \boldsymbol{\mu}_0)} \quad \text{a.s..} \quad (9)$$
3) For each $\rho > 0$, the quantity $\mathbb{E}[\tau_{\pi_{\text{BBMTS}}(\delta,\rho)}]$ satisfies
$$\limsup_{\delta \downarrow 0} \frac{\mathbb{E}[\tau_{\pi_{\text{BBMTS}}(\delta,\rho)}]}{\log(1/\delta)} \leq \frac{1+\rho}{T^*(\boldsymbol{q}_0, \boldsymbol{\mu}_0)}. \quad (10)$$

请注意，ρ 是一个可调参数，可以通过设置该参数使 (10) 中的上界尽可能接近 (1)。显然，当 $\rho \downarrow 0$ 时，(10) 的右端与 (1) 中的下界相吻合。因此，定理 1 和定理 2 共同表明，$1/T^*(\boldsymbol{q}_0, \boldsymbol{\mu}_0)$ 是与实例 $(\boldsymbol{q}_0, \boldsymbol{\mu}_0)$ 对应的期望停止时间的最优渐近增长率。

IV. Partition Setting

In this section, we analyze the simpler setting when the K arms are *partitioned* among the M boxes and the learner knows $\{\mathcal{A}_m : m \in [M]\}$ but not the arm selection probabilities of the boxes. For this setting, we present an algorithm based on successive elimination, and provide a non-asymptotic, high-probability upper bound on its stopping time. We show that the upper bound is tight in the instance-specific parameters.

In the partition setting, we find it convenient to index an arm by the box it is associated with; the kth arm in box m is denoted by $A_{m,k}$ and its mean reward is denoted by $\mu_{m,k}$. All arms are assumed to yield 1-sub-Gaussian rewards. Let $\boldsymbol{q}_0 = \{q^0_{m,k} : k \in \mathcal{A}_m, m \in [M]\}$ and $\boldsymbol{\mu}_0 = \{\mu^0_{m,k} : k \in \mathcal{A}_m, m \in [M]\}$ define the underlying instance. Without loss of generality, let $A_{1,1}$ be the best arm in this instance. We let $\Delta_{m,k} := \mu^0_{1,1} - \mu^0_{m,k}$ for all $(m,k) \neq (1,1)$, and $\Delta_{1,1} := \min_{(m,k) \neq (1,1)} \Delta_{m,k}$ denote the arm sub-optimality gaps.

A. Non-Asymptotic Analysis: Achievability

现在，我们提出一种基于逐步淘汰的算法（称为盒式老虎机逐步淘汰算法，即 BBSEA），并分析其性能。在介绍该算法之前，我们先引入一些符号。我们的算法按轮次执行；我们用 n 表示轮次编号，用 t 表示运行时间。令 $t_{m,k}(n)$ 表示截至第 n 轮，臂 $A_{m,k}$ 被拉取的次数；令 $\hat{\mu}_{m,k}(n)$ 表示臂 $A_{m,k}$ 在被拉取 $t_{m,k}(n)$ 次后的经验均值。在任意给定的轮次 n 中，令 S 表示活跃臂（候选最佳臂）的集合，S_m 表示活跃的集合。

算法 2 盒式多臂老虎机逐次消除算法 输入： K, M, $\delta > 0$, $\mathcal{A}_m$， 用于 $m \in [M]$ 输出：$\hat{a}_{\text{BBSEA}} \in [K]$ （最佳臂）。初始化： $B = [M]$, $n = 0$, $t = 0$,$S_m = \mathcal{A}_m$ $\forall m$, $S = \bigcup_m S_m$, $\hat{}\mu_{m,k}(0) = 0$ $\forall m, k$。 1: 当 $|S| > 1$ 时执行 2: $n \leftarrow n + 1$ 3: 对于每个 $m \in B$, 选择盒子 m ，直到每个活跃臂$A_{m,k}$ 在盒子 m 中至少被拉取 n 次。每次选择盒子时，将 t 增加 1。 4: 更新所有活跃臂的 $t_{m,k}(n)$、$\hat{}\mu_{m,k}(n)$、$\text{UCB}_{m,k}(n)$ 和 $\text{LCB}_{m,k}(n)$。 5: 如果存在 $\exists A_{m'}$, $_{k'} \in S$ ，使得 $\text{UCB}_{m,k}(n) < \text{LCB}_{m',k'}(n)$，则 6: $S_m \leftarrow S_m \backslash A_{m,k}$, $S \leftarrow \bigcup_{m\in[M]} S_m$, 7: $B \leftarrow \{m : S_m \neq \emptyset\}$。 8: 结束 if 9: 如果 $|S| = 1$ ，则 10: $\hat{}$ $a_{\text{BBSEA}} \leftarrow a \in S$, $S \leftarrow \emptyset$, $B \leftarrow \emptyset$。 11: 结束 if 12: 结束 while 13: 返回 $\hat{a}_{\text{BBSEA}}$。

与盒子 m 相关联的臂，以及 B 活跃盒子集合（至少与一个活跃臂相关联的盒子）。~~对于固定的~~ $\delta \in$ （0, 1），令 α_δ（x）$:= \sqrt{^2_{\log(8Kx^2/\delta)}x}$。~~令 $\text{UCB}_{m,k}$~~（n）和 $\text{LCB}_{m,k}$（n）分别表示 $\hat{}\mu_{m,k}$（n）的上置信界和下置信界，其置信区间长度为 α_δ（$t_{m,k}$（n））， 即,

$$\text{UCB}_{m,k}(n) = \hat{\mu}_{m,k}(n) + \alpha_\delta(t_{m,k}(n)), \tag{11}$$
$$\text{LCB}_{m,k}(n) = \hat{\mu}_{m,k}(n) - \alpha_\delta(t_{m,k}(n)). \tag{12}$$

令 \(\hat{a}_{\text{BBSEA}}\) 表示 BBSEA 输出的最佳臂。对于所有 m, 我们设置初始值 $t = 0$、 $n = 0$、 $S = [K]$、 $B = [M]$ 和 $S_m = \mathcal{A}_m$。在每一轮 n 中，BBSEA 执行以下一系列动作: (i) 选择每个盒子 $m \in B$， 直至与盒子 m 相关联的每个活跃臂至少被拉取 n 次；对于每次盒子选择，时间 t 增加 1。(ii) 每次进行盒子选择和臂拉取时，更新所有活跃臂的 t_{m}、$_{k}(n)$、\(\hat{\mu}_{m}\)、$_{k}(n)$、UCB_{m}、$_{k}(n)$ 以及 LCB_{m}、$_{k}(n)$ 的值。(iii) 如果 UCB_{m}、$_{k}(n) \leq \max_{m'}$、$_{k'}$ $\text{LCB}_{m'}$、$_{k'}(n)$，则在第 n 轮中将臂 A_{m}、$_{k}$ 从 S 中淘汰。上述动作序列不断重复，直至 S 中仅剩一个臂，此时算法停止，并输出 S 中的唯一臂作为最佳臂。由于每次盒子选择都会使 t 增加，因此停止时间等于盒子选择的总次数。BBSEA 的伪代码见算法 2。

BBSEA 的性能分析：在介绍 BBSEA 的性能结果之前，我们先引入一些有用的符号。对于每一对 (m, k) 和 $\delta \in (0, 1)$，令 $\alpha_{m,k} = 1 + \frac{102}{\Delta^2_{m,k}} \log\left(\frac{64\sqrt{\frac{8K}{\delta}}}{\Delta^2_{m,k}}\right)$，并令 $T_{m,k}$ 表示臂 $A_{m,k}$ 从 S（集合中的轮次。

active arms). Furthermore, let $\beta_{m,k} = \frac{1}{q^0_{m,k}}\left[\alpha_{m,k} + 2\log\frac{2K}{\delta} + 2\sqrt{\log\frac{2K}{\delta}\left(\log\frac{2K}{\delta} + \alpha_{m,k}\right)}\right]$, and let $\beta_m = \max_{k\in\mathcal{A}_m} \beta_{m,k}$.

有了上述记号，我们现在可以陈述本节关于 BBSEA 性能的主要结果。

Theorem 3: Fix $\delta \in (0, 1)$, The following hold with probability greater than $1 - \delta$.

1) BBSEA outputs the best arm correctly.
2) The stopping time of BBSEA is $\leq \sum_{m=1}^{M} \beta_m$.

Furthermore, for any $\pi \in \Pi(\delta)$,

$$\mathbb{E}[\tau_\pi] \geq \log\left(\frac{1}{2.4\,\delta}\right) \cdot \sum_{m=1}^{M} \max_{k\in\mathcal{A}_m} \frac{1}{q^0_{m,k}\,\Delta^2_{m,k}}. \tag{13}$$

Notice that

$$\beta_{m,k} = O\left(\frac{1}{q^0_{m,k}\,\Delta^2_{m,k}} \log\left(\frac{K}{\delta\Delta_{m,k}}\right)\right), \tag{14}$$

$$\sum_{m=1}^{M} \beta_m = \sum_{m=1}^{M} O\left(\max_{k\in\mathcal{A}_m} \frac{1}{q^0_{m,k}\,\Delta^2_{m,k}} \log\left(\frac{K}{\delta\Delta_{m,k}}\right)\right). \tag{15}$$

It is easy to see that the high-probability, instance-dependent upper bound on the stopping time of BBSEA given by (15) matches with the instance-dependent lower bound in (13) order-wise in δ and the instance-specific parameters $(\boldsymbol{q}_0, \boldsymbol{\mu}_0)$. These bounds are hence tight up to multiplicative factors.

V. Concluding Remarks

我们研究了多臂老虎机中的最佳臂识别问题。在该问题中，学习者通过我们称为盒子的探索中介来访问各个臂，而盒子的臂访问概率对学习者而言是未知的。我们在分析中解决的关键挑战是信息论下界中最优分配的不唯一性。我们证明，通过跟踪与问题实例的运行估计相对应的、任意选取的最优分配的经验平均值，可以实现渐近最优性。未来一个值得研究的方向是开发能够获得非渐近上界的高效算法。在本文中，我们仅在各个臂被划分到不同盒子的特殊情况下实现了这一目标。这里的主要挑战在于，如何利用不同盒子的臂选择概率估计来指导盒子选择。鉴于任意特定的臂都可能通过多个盒子被访问（且这些盒子的访问概率可能不同），因此在一开始并不清楚应该选择哪个盒子，才能最大限度地提高拉取给定臂的概率。因此，像 LUCB 这样的算法无法直接推广到我们的研究场景中。

Acknowledgements: Kota Srinivas Reddy was supported by the Department of Science and Technology (DST), Govt. of India, through the INSPIRE faculty fellowship.
P. N. Karthik was supported by the National University of Singapore via grant A-0005994-01-00.
Nikhil Karamchandani and Jayakrishnan Nair acknowledge support from DST via grant CRG/2021/002923 and two SERB MATRICS grants.

References

[1] P. N. Karthik 和 R. Sundaresan，“通过颤抖的手检测奇异非静止马尔可夫臂”，IEEE 信息论汇刊，第 67 卷，第 8 期，第 5230–5258 页，2021年。[2] S. Kalyanakrishnan、A. Tewari、P. Auer 和 P. Stone，“随机多臂老虎机中的 PAC 子集选择”，2012年。[3] E. Kaufmann、O. Cappé 和 A. Garivier，“关于多臂老虎机模型中最佳臂识别的复杂度”，机器学习研究期刊，第 17 卷，第 1 期，第 1–42 页，2016年。[4] A. Garivier 和 E. Kaufmann，“固定置信度下的最优最佳臂识别”，载于《学习理论会议》。PMLR，2016年，第 998–1027 页。[5] E. Even-Dar、S. Mannor 和 Y. Mansour，“多臂老虎机与强化学习问题中的动作消除和停止条件”，《机器学习研究期刊》，第 7 卷，第 6 月号，第 1079–1105 页，2006年。[6] K. Jamieson 和 R. Nowak，“固定置信度设置下多臂老虎机的最佳臂识别算法”，载于《2014年第48届信息科学与系统年会（CISS）》。IEEE，2014年，第 1–6 页。[7] Y. Jedra 和 A. Proutiere，“线性老虎机中的最优最佳臂识别”，《神经信息处理系统进展》，第 33 卷，第 10 007–10 017 页，2020年。[8] Z. Chen、P. N. Karthik、V. Y. F. Tan 和 Y. M. Chee，“具有异构客户端的联邦最佳臂识别”，arXiv 预印本 arXiv:2210.07780，2022年。[9] J. Scarlett、I. Bogunovic 和 V. Cevher，“重叠式多老虎机最佳臂识别”，载于《2019年IEEE国际信息论研讨会（ISIT）》。IEEE，2019年，第 2544–2548 页。[10] S. A. Jain、S. Goenka、D. Bapna、N. Karamchandani 和 J. Nair，“序贯社区众数估计”，《性能评估》，第 152 卷，第 102247 页，2021年。[11] T. Lai 和 H. Robbins，“渐近高效的自适应分配规则”，《应用数学进展》，第 6 卷，第 1 期，第 4–22 页，1985年3月。 [在线]。可查阅：https://doi.org/

10 1016 0196 8858 85 90002 8 。/ - ( ) - S.

[12] Hu 和 N. S. Papageorgiou，“多值分析手册。第一卷：理论，第 419 卷”，《数学及其应用》，1997年。[13] S. Magureanu、R. Combes 和 A. Proutiere，“Lipschitz老虎机：遗憾下界与最优算法”，载于《学习理论会议》。PMLR，2014年，第 975–999 页。[14] S. Boucheron、G. Lugosi 和 P. Massart，《集中不等式：独立性的非渐近理论》。牛津大学出版社，2013年。[15] P. Natalini 和 B. Palumbo，“不完全伽马函数的不等式”，《数学不等式与应用》，第 3 卷，第 1 期，第 69–77 页，2000年。[16] J. M. Borwein、O.-Y. Chan 等，“余不完全伽马函数的一致界”，《数学不等式与应用》，第 12 卷，第 115–121 页，2009年。[17] K. S. Reddy、P. N. Karthik 和 V. Y. F. Tan，“联邦最佳臂识别中的近乎零成本通信”，arXiv 预印本 arXiv:2208.09215，2022年。

Appendix

给定两个大小相同的实值向量 u, v 以及一个紧致凸集 C，以下距离度量将在我们的证明中用到：$d_\infty(u,v) \coloneqq \max_i |u_i - v_i|$，$d_\infty(u, C) \coloneqq \min_{v\in C} d_\infty(u,v)$，以及 $\|u - v\| \coloneqq \sqrt{\sum_i (u_i - v_i)^2}$。

Appendix A
Proof of Lemma 1

我们首先注意到，对于所有 $w \in \Sigma_M$，

$$\psi(\boldsymbol{q}_0, \boldsymbol{\mu}_0, w) = \inf_{\boldsymbol{\lambda}\in \mathrm{ALT}(\boldsymbol{\mu}_0)} \sum_{m=1}^{M}\sum_{k=1}^{K} w_m\, q^0_{m,k}\, \frac{(\mu^0_k - \lambda_k)^2}{2} \tag{16}$$

$$= \min_{k\neq a*(\boldsymbol{\mu}_0)} \frac{w^A_k\, w^A_{a^*(\boldsymbol{\mu}_0)}}{w^A_k + w^A_{a^*(\boldsymbol{\mu}_0)}} \frac{(\mu^0_k - \mu^0_{a^*(\boldsymbol{\mu}_0)})^2}{2}, \tag{17}$$

其中 $w^A_a = \sum_{m=1}^M w_m\, q^0_{m,\,a}$ 对于所有 $a \in [K]$。证明只需沿用 [3, 引理 3]的思路即可。由上述内容可知，$\psi(\boldsymbol{q}_0, \boldsymbol{\mu}_0, \cdot)$ 是连续的。根据 [12, 第 84] 页的 Berge最大值定理，可立即得到，对于所有满足 $a^\star(\boldsymbol{\mu})$ 唯一的 $(\boldsymbol{q}, \boldsymbol{\mu})$，$(q, \mu) \mapsto \mathcal{W}^\star(\boldsymbol{q}, \boldsymbol{\mu})$ 是上半连续的。

We now show that $\mathcal{W}^\star(\boldsymbol{q}, \boldsymbol{\mu})$ is convex for all $(\boldsymbol{q}, \boldsymbol{\mu})$. Fix $(\boldsymbol{q}, \boldsymbol{\mu})$ and pick $w^{(1)}, w^{(2)} \in \mathcal{W}^\star(\boldsymbol{q}, \boldsymbol{\mu})$ and $\alpha \in [0,1]$ arbitrarily. By definition,

$$\psi(\boldsymbol{q}, \boldsymbol{\mu}, w^{(1)}) = T^\star(\boldsymbol{q}, \boldsymbol{\mu}), \quad \psi(\boldsymbol{q}, \boldsymbol{\mu}) = T^\star(\boldsymbol{q}, \boldsymbol{\mu}).$$

We then have

$$\begin{aligned}
&\psi(\boldsymbol{q}, \boldsymbol{\mu}, \alpha w^{(1)} + (1-\alpha) w^{(2)}) \\
&= \inf_{\boldsymbol{\lambda}\in \mathrm{ALT}(\boldsymbol{\mu})} \sum_{m=1}^{M}\sum_{k=1}^{K} (\alpha w^{(1)}_m + (1-\alpha) w^{(2)}_m)\, q_{m,k}\, \frac{(\mu_k - \lambda_k)^2}{2} \\
&\geq \alpha \inf_{\boldsymbol{\lambda}\in \mathrm{ALT}(\boldsymbol{\mu})} \sum_{m=1}^{M}\sum_{k=1}^{K} w^{(1)}_m\, q_{m,k}\, \frac{(\mu_k - \lambda_k)^2}{2} \\
&+ (1-\alpha) \inf_{\boldsymbol{\lambda}\in \mathrm{ALT}(\boldsymbol{\mu})} \sum_{m=1}^{M}\sum_{k=1}^{K} w^{(2)}_m\, q_{m,k}\, \frac{(\mu_k - \lambda_k)^2}{2} \\
&= \alpha\, T^\star(\boldsymbol{q}, \boldsymbol{\mu}) + (1-\alpha)\, T^\star(\boldsymbol{q}, \boldsymbol{\mu}) \\
&= T^\star(\boldsymbol{q}, \boldsymbol{\mu}).
\end{aligned}$$

另一方面，由于 Σ_M 是凸的，$\alpha w^{(1)} + (1-\alpha) w^{(2)} \in \Sigma_M$，因此 $\psi(\boldsymbol{q}, \boldsymbol{\mu}, \alpha w^{(1)} + (1-\alpha) w^{(2)}) \leq T^\star(\boldsymbol{q}, \boldsymbol{\mu})$，这意味着 $\psi(\boldsymbol{q}, \boldsymbol{\mu}, \alpha w^{(1)} + (1-\alpha) w^{(2)}) = T^\star(\boldsymbol{q}, \boldsymbol{\mu})$。这证明了 $\alpha w^{(1)} + (1-\alpha) w^{(2)} \in \mathcal{W}^\star(\boldsymbol{q}, \boldsymbol{\mu})$。注意到 $w^{(1)}$、$w^{(2)} \in \mathcal{W}^\star(\boldsymbol{q}, \boldsymbol{\mu})$ 和 $\alpha \in [0,1]$ 都是任意的，我们便得到了所需结果。

附录B 引理2的证明

令 $C \coloneqq \mathcal{W}^\star(\boldsymbol{q}_0, \boldsymbol{\mu}_0)$。证明的核心思路是跟踪平均值 $\bar{w}(t) := \frac{1}{t}\sum_{s=1}^t w(s)$，并证明它最终会落在 C 的某个邻域内。这里我们只给出高层次的论证。任取一个 $\varepsilon > 0$。根据式 (4) 中的抽样规则，对于所有 m，我们有 $N(t,m) = \Omega(\sqrt{t})$ 几乎必然成立。由强大数定律可知，$\|\hat{\boldsymbol{q}}(t) - \boldsymbol{q}_0\| \to 0$ 几乎必然成立，$\|\hat{\boldsymbol{\mu}}(t) - \boldsymbol{\mu}_0\| \to 0$ 几乎必然成立，当 $t \to \infty$ 时。由于对于所有 t，$w(t+1) \in \mathcal{W}(\hat{\boldsymbol{q}}(t), \hat{\boldsymbol{\mu}}(t))$，且对于所有足够大的 t，$\mathcal{W}(\hat{\boldsymbol{q}}(t), \hat{\boldsymbol{\mu}}(t))$ 与 C “接近”（这得益于映射 $(\boldsymbol{q}, \boldsymbol{\mu}) \mapsto \mathcal{W}^\star(\boldsymbol{q}, \boldsymbol{\mu})$ 的上半连续性），因此存在 $t_0 = t_0(\varepsilon)$，使得对于所有 $t \geq t_0$，$d_\infty(w(t), C) \leq \varepsilon$。若 $v(t) \coloneqq \arg\min_{w\in C} d_\infty(w(t), w)$ 表示将 $w(t)$ 投影到 C 上，则对于所有 $t \geq t_0$，有 $d_\infty(w(t), v(t)) \leq \varepsilon$。令 $\bar{v}(t) \coloneqq \frac{1}{t}\sum_{s=1}^t v(s)$，容易证明，对于所有 $t \geq t_0' = t_0/(8\varepsilon)$，$\|\bar{w}(t) - \bar{v}(t)\| \leq 9\varepsilon$。令 $\hat{w}(t) \coloneqq \arg\min_{w\in C} d_\infty(\bar{w}(t), w)$，于是可得，对于所有 $t \geq t_0'$，$d_\infty(\bar{w}(t), \hat{w}(t)) \leq 9\varepsilon$。

假设对于所有 m、t，$\varepsilon_m(t) \coloneqq N(t,m) - t\,\hat{w}_m(t)$。我们将在下文证明，存在 $t_0'' = t_0''(\varepsilon)$，使得对于所有 $t \geq t_0''$，$\{B_{t+1} = m\} \subset \{\varepsilon_m(t) \leq 9t\varepsilon\}$。假设这一结论成立，注意到 $\varepsilon_m(t)$ 满足递归关系 $\varepsilon_m(t+1) = \varepsilon_m(t) + \mathbf{1}_{\{B_{t+1}=m\}} - \hat{w}_m(t+1)$。对于所有 $t \geq t_0''$，我们有

$\varepsilon_m(t+1) \le \varepsilon_m(t) + \mathbf{1}_{\{\varepsilon_m(t)\le 9t\varepsilon\}} - \hat{w}_m(t+1)$。现在，我们通过归纳法证明，对于所有 $t \ge t_0''$ 和 m，均有 $\varepsilon_m(t) \le \max\{\varepsilon_m(t_0''),\ 9t\varepsilon+1\}$。显然，当 $t = t_0''$ 时，该结论成立。若 $\varepsilon_m(t) \le 9t\varepsilon$，则

$$\begin{aligned}
\varepsilon_m(t+1) &\le \varepsilon_m(t) + \mathbf{1}_{\{\varepsilon_m(t)\le 9t\varepsilon\}} - \hat{w}_m(t+1) \\
&\le 9t\varepsilon + 1 - \hat{w}_m(t+1) \\
&\le 9t\varepsilon + 1 \\
&\le \max\{\varepsilon_m(t_0''),\ 9t\varepsilon+1\} \\
&\le \max\{\varepsilon_m(t_0''),\ 9(t+1)\varepsilon+1\} \quad \forall m \in [M],
\end{aligned}$$

whereas if $\varepsilon_m(t) > 9t\varepsilon$, then

$$\begin{aligned}
\varepsilon_m(t+1) &\le \varepsilon_m(t) - \hat{w}_m(t+1) \\
&\le \max\{\varepsilon_m(t_0''),\ 9t\varepsilon+1\} - \hat{w}_m(t+1) \\
&\le \max\{\varepsilon_m(t_0''),\ 9t\varepsilon+1\} \\
&\le \max\{\varepsilon_m(t_0''),\ 9(t+1)\varepsilon+1\} \quad \forall m \in [M].
\end{aligned}$$

Because $\sum_m \varepsilon_m(t) = 0$, we have

$$\begin{aligned}
\varepsilon_m(t) &= -\sum_{m'\ne m} \varepsilon_{m'}(t) \\
&\ge -(M-1)\max_{m'} \max\{\varepsilon_{m'}(t_0''),\ 9(t+1)\varepsilon+1\}
\end{aligned}$$

for all m. Thus, it follows that

$$\begin{aligned}
|\varepsilon_m(t)| &\le (M-1)\max_{m'} \max\{\varepsilon_{m'}(t_0''),\ 9t\varepsilon+1\} \\
&\le (M-1)\max\{t_0'',\ 9t\varepsilon+1\} \\
&= (M-1)\,t\,\max\left\{\frac{t_0''}{t},\ 9\varepsilon + \frac{1}{t}\right\} \\
&\le 10(M-1)\,t\,\varepsilon
\end{aligned}$$

for all $t \ge t_1 = \frac{1}{\varepsilon}\max\{t_0'', 10\}$, which in turn yields

$$\begin{aligned}
d_\infty((N(t,m)/t)_{m\in M}, C) &\le d_\infty((N(t,m)/t)_{m\in M}, \hat{w}(t)) \\
&= \max_m \left| \frac{N(t,m)}{n} - \hat{w}_m(t) \right| \\
&= \max_m \frac{|\varepsilon_m(t)|}{t} \\
&\le 10(M-1)\,\varepsilon.
\end{aligned}$$

Noting that ε is arbitrary, we arrive at the desired result.

It then remains to show that $\{B_{t+1} = m\} \subset \{\varepsilon_m(t) \le 9t\varepsilon\}$ for all m and large t. From the sampling rule in (4), it is clear that $\{B_{t+1} = m\} \subset \mathcal{E}_1(t) \cup \mathcal{E}_2(t)$, where

$$\mathcal{E}_1(t) = \left\{ m = \arg \min_{m' \in \mathrm{supp}\left(\sum_{s=1}^t w(s)\right)} N(t,m') - \sum_{s=1}^t w_{m'}(s) \right\},$$

$$\mathcal{E}_2(t) = \left\{ \min_{m' \in [M]} N(t,m') < f(t) \text{ and } m = i_t \right\}.$$

假设 $\mathcal{E}_1(t)$ 成立。那么，对于所有 $t \ge t_0'$ （其中 t_0' 的定义与证明前文相同），我们有

$$\begin{aligned}
\varepsilon_m(t) &= N(t,m) - t\,\hat{w}_m(t) \\
&= N(t,m) - n\,\bar{w}_m(t) + t(\bar{w}_m(t) - \hat{w}_m(t)) \\
&\le N(t,m) - n\,\bar{w}_m(t) + t d_\infty(\bar{w}(t), \hat{w}(t)) \\
&\le N(t,m) - t\,\bar{w}_m(t) + 9t\varepsilon \\
&\overset{(a)}{=} \min_{m' \in \mathrm{supp}\left(\sum_{s=1}^t w(s)\right)} N(t,m') - t\,\bar{w}_{m'}(t) + 9t\varepsilon \\
&\le 9t\varepsilon,
\end{aligned}$$

其中，上述 (a) 是由注意到 $\mathcal{E}_1(t)$ 成立而得，最后一个不等式则由注意到以下事实而得：

$$\sum_{m' \in \mathrm{supp}\left(\sum_{s=1}^t w(s)\right)} N(t,m') - t\,\bar{w}_{m'}(t) \le 0.$$

按照上述类似思路，可以证明，只要 $\mathcal{E}_2(t)$ 成立，就有 $\varepsilon_m(t) \le 9t\varepsilon$。证明至此完成。

Appendix C
Proof of Lemma 3

给定 $(\boldsymbol{q}, \boldsymbol{\mu})$ 和 $t \ge 1$，在问题实例 $C = (\boldsymbol{q}, \boldsymbol{\mu})$ 下，直到 t 为止所有盒子选择、臂拉取以及臂的观测值的对数似然为

$$\begin{aligned}
Z(t) &= \log P(B_1) + \sum_{s=2}^t \log P(B_s | \mathcal{F}_{s-1}) \\
&\quad + \sum_{m=1}^M \sum_{k=1}^K N(t,m,k) \log q_{m,k} \\
&\quad + \sum_{m=1}^M \sum_{k=1}^K \sum_{s=1}^t \mathbf{1}_{\{A_s = k\}} \log\left(\beta \exp\left(-\frac{(X_s - \mu_k)^2}{2} \right) \right),
\end{aligned} \tag{18}$$

其中 β 是方差为 1 的高斯分布的归一化常数。注意到式 (18) 中的前两项不依赖于 C（因为任何策略 π 都不受对底层问题实例的了解影响），由式 (18) 可得

$$\begin{aligned}
&\sup_{C=(\boldsymbol{q},\boldsymbol{\mu}):\mu_a \ge \mu_b} \exp(Z_C(t)) \\
&\propto \sup_{C=(\boldsymbol{q},\boldsymbol{\mu}):\mu_a \ge \mu_b} \exp\left(\sum_{m=1}^M \sum_{k=1}^K N(t,m,k) \log q_{m,k} \right. \\
&\qquad \left. - \sum_{k=1}^K \sum_{s=1}^t \mathbf{1}_{\{A_s=k\}} \frac{(X_s - \mu_k)^2}{2} \right) \\
&= \exp\left(\sum_{m=1}^M \sum_{k=1}^K N(t,m,k) \log \hat{q}_{m,k}(t) \right. \\
&\qquad \left. - \inf_{\mu:\mu_a \ge \mu_b} \sum_{k=1}^K \sum_{s=1}^t \mathbf{1}_{\{A_s=k\}} \frac{(X_s - \mu_k)^2}{2} \right),
\end{aligned} \tag{19}$$

其中 $\{\hat{q}_{m,\,k}(t)\}_{m,\,k}$ 是时刻 t 盒子概率的经验估计。现在注意到，最小化

$$\inf_{\boldsymbol{\mu}:\mu_a \ge \mu_b} \sum_{k=1}^K \sum_{s=1}^t \mathbf{1}_{\{A_s=k\}} \frac{(X_s - \mu_k)^2}{2} \tag{20}$$

是一个简单的凸规划，其最优解可以表示为

$$\mu_k' = \begin{cases} \hat{\mu}_k(t), & \text{if } \hat{\mu}_a(t) > \hat{\mu}_b(t), \\ \hat{\mu}_{a,b}(t), & \text{otherwise} \end{cases} \quad \text{for } k = a, b, \tag{21}$$

以及对于所有 $k \neq a$、b，$\mu'_k = \hat{\mu}_k(t)$。类似地，可以证明，.....的最优解是

$$\inf_{\boldsymbol{\mu}:\mu_a \leq \mu_b} \sum_{k=1}^{K} \sum_{s=1}^{t} \mathbf{1}_{\{A_s=k\}} \frac{(X_s - \mu_k)^2}{2},$$

say $\mu'' = (\mu''_k : k \in [K])$, is

$$\mu''_k = \begin{cases} \hat{\mu}_k(t), & \text{if } \hat{\mu}_a(t) < \hat{\mu}_b(t), \\ \hat{\mu}_{a,b}(t), & \text{otherwise} \end{cases} \quad \text{for } k = a, b, \tag{22}$$

对于所有 $k \neq a, b$，以及 $\mu''_k = \hat{\mu}_k(t)$。现在我们证明引理 3。假设 $\hat{\mu}_a(t) \geq \hat{\mu}_b(t)$。然后，注意到式 (19) 中指数函数内的第一项在广义对数似然比 (6) 的分子项和分母项中是相同的（因此会相互抵消），我们得到

$$\begin{aligned} Z_{a,b}(t) &= -\sum_{k=1}^{K} \sum_{s=1}^{t} \mathbf{1}_{\{A_s=k\}} \frac{(X_s - \mu'_k)^2}{2} \\ &\quad + \sum_{k=1}^{K} \sum_{s=1}^{t} \mathbf{1}_{\{A_s=k\}} \frac{(X_s - \mu''_k)^2}{2} \\ &= \sum_{s=1}^{t} \mathbf{1}_{\{A_s=a\}} \frac{(X_s - \hat{\mu}_{a,b}(t))^2 - (X_s - \hat{\mu}_a(t))^2}{2} \\ &\quad + \sum_{s=1}^{t} \mathbf{1}_{\{A_s=b\}} \frac{(X_s - \hat{\mu}_{a,b}(t))^2 - (X_s - \hat{\mu}_b(t))^2}{2} \\ &= N_a(t) \frac{\left(\hat{\mu}_a(t) - \hat{\mu}_{a,b}(t)\right)^2}{2} \\ &\quad + N_b(t) \frac{\left(\hat{\mu}_b(t) - \hat{\mu}_{a,b}(t)\right)^2}{2}. \end{aligned} \tag{23}$$

When $\hat{\mu}_a(t) \leq \hat{\mu}_b(t)$, it can be shown along similar lines as above that $Z_{a,b}(t) = -Z_{b,a}(t)$, where $Z_{b,a}(t)$ is as in (23) with the roles of a and b interchanged.

Appendix D
Proof of Theorem 2

Throughout the proof, we use the symbol π to denote $\pi_{\text{BBMTS}(\delta,\rho)}$ for fixed parameters δ, ρ. Let $\hat{a}$ denote the best arm ouput by BBMTS at stoppage.

Proof of part 1: Let

$$\mathcal{E}_1 = \left\{ \exists t \geq 1 : Z(t) \geq \zeta(t,\delta,\rho) \text{ and } \min_{k\in[K]} N_k(t) > 0 \right\},$$
$$\mathcal{E}_2 = \{\mu_{\hat{a}} < \mu_{a^\star(\boldsymbol{\mu}_0)}\}.$$

Recall that $\hat{a}(t) \in \arg\max_{a\in[K]} \hat{\mu}_a(t)$ denotes the estimate of the best arm at time t, and $\hat{a}(\tau_\pi) = \hat{a}$. We then have

$$\begin{aligned} &P(\mathcal{E}_1 \cap \mathcal{E}_2) \\ &\leq \sum_{t=1}^{\infty} P\left(Z(t) \geq \zeta(t,\delta,\rho),\, \min_k N_k(t) > 0,\, \mu_{\hat{a}(t)} < \mu_{a^\star(\boldsymbol{\mu}_0)} \right) \\ &\leq \sum_{t=1}^{\infty} P\left(Z_{\hat{a}(t),a^\star(\boldsymbol{\mu}_0)} \geq \zeta(t,\delta,\rho),\, \min_k N_k(t) > 0 \right) \\ &= \sum_{t=1}^{\infty} P\left(\sum_{k=1}^{K} N_k(t) \frac{(\hat{\mu}_k(t) - \mu_k)^2}{2} \geq \zeta(t,\delta,\rho) \right) \\ &\overset{(a)}{\leq} \sum_{t=1}^{\infty} e^{K+1} \left(\frac{\zeta(t,\delta,\rho)^2 \log t}{K} \right)^K e^{-\zeta(t,\delta,\rho)} \\ &\leq \delta, \end{aligned} \tag{24}$$

where (a) above is due to [13, Theorem 2] and the last line follows from the definition of $\zeta(t,\delta,\rho)$. This completes the proof of part 1.

Proof of part 2: We use the below technical lemma from [4] in the proof.

Lemma 4: [4, Lemma 18] For any two constants $c_1 > 0, c_2 > 0$ such that $c_2/c_1 \geq 1$ and $\alpha \geq 1$,

$$\inf\{t \geq 1 : tc_1 \geq \log(c_2 t^\alpha)\} \leq \frac{\alpha}{c_1} \left(\log \frac{c_2 e}{c_1^\alpha} + \log\log \frac{c_2}{c_1^\alpha} \right).$$

Consider the event

$$\mathcal{E} = \Big\{ d_\infty((N(t,m)/t)_{m\in[M]},\ \mathcal{W}^\star(\boldsymbol{q}_0, \boldsymbol{\mu}_0)) \overset{t\to\infty}{\longrightarrow} 0, \\ \hat{\boldsymbol{\mu}}(t) \overset{t\to\infty}{\longrightarrow} \boldsymbol{\mu},\ \hat{\boldsymbol{q}}(t) \overset{t\to\infty}{\longrightarrow} \boldsymbol{q}_0 \Big\}.$$

根据引理2，我们知道在BBMTS下，$\mathcal{E}$ 以概率1发生。任意固定 $\varepsilon > 0$ 和 $w^\star \in \mathcal{W}^\star(q, \mu)$。根据附录A中所确立的 ψ（如式(17)所定义）的连续性，我们知道存在 $\{\boldsymbol{q}_0\} \times \{\boldsymbol{\mu}_0\} \times \mathcal{W}^\star$（$\boldsymbol{q}_0$，$\boldsymbol{\mu}_0$）的一个开邻域，记为 $\mathcal{U} = \mathcal{U}(\varepsilon)$，使得

$$\psi(\boldsymbol{q}, \boldsymbol{\mu}, w') \geq (1-\varepsilon)\, \psi(\boldsymbol{q}_0, \boldsymbol{\mu}_0, w^\star) = (1-\varepsilon)\, T^\star(\boldsymbol{q}_0, \boldsymbol{\mu}_0). \tag{25}$$

对于所有 $(\boldsymbol{q}, \boldsymbol{\mu}, w') \in \mathcal{U}$。在事件 $\mathcal{E}$ 发生的条件下，存在 $n_0 = n_0(\varepsilon)$，使得 $(\hat{\boldsymbol{\mu}}(t), \hat{\boldsymbol{q}}(t), (N(t,m)/t)_{m\in[M]}) \in \mathcal{U}$对于所有$t \geq n_0$均成立，因此

$$\psi(\hat{\boldsymbol{q}}(t), \hat{\boldsymbol{\mu}}(t), (N(t,m)/t)_{m\in[M]}) \geq (1-\varepsilon)\, T^\star(\boldsymbol{q}_0, \boldsymbol{\mu}_0) \tag{26}$$

对所有 $t \geq n_0$ 均成立。由于 $\|\hat{\boldsymbol{\mu}}(t) - \boldsymbol{\mu}_0\| \overset{t\to\infty}{\longrightarrow} 0$ 在 $\mathcal{E}$ 下成立，且 $a^\star(\boldsymbol{\mu}_0)$ 是唯一的，因此存在足够大的 n_1，使得 $\hat{a}(t) = \arg\max_{a\in[K]} \hat{\mu}_a(t)$ 是唯一的 $\forall t \geq n_1$。在这种情况下，我们可以将检验统计量 $Z(t)$ 表示为 $Z(t) = \min_{b\neq\hat{a}(t)} Z_{\hat{a}(t),b}(t) = t\, \psi(\hat{\boldsymbol{q}}(t), \hat{\boldsymbol{\mu}}(t), (N(t,m)/t)_{m\in[M]})$。根据抽样规则 (4)，对于所有足够大的 t，几乎必然有 $\min_{m\in[M]} N(t,m) = \Omega(\sqrt{t}) > 0$。结合在事件 $\mathcal{E}$ 下 $\|\hat{\boldsymbol{q}}(t) - \boldsymbol{q}_0\| \overset{t\to\infty}{\longrightarrow} 0$ 的事实可知，存在 $n_2 \geq 1$，使得对于所有 $t \geq n_2$，均有 $\min_{k\in[K]} N_k(t) > 0$。因此，在 $\mathcal{E}$ 下，对于所有 $t \geq \max\{n_0, n_1, n_2\}$，我们有

$$Z(t) \geq t(1-\varepsilon)\, T^\star(\boldsymbol{q}_0, \boldsymbol{\mu}_0), \quad \min_{k\in[K]} N_k(t) > 0. \tag{27}$$

接下来，我们有（从此处开始采用的格式）

$$\begin{aligned} \tau_\pi &= \inf\left\{ t \geq 1 : Z(t) \geq \zeta(t,\delta,\rho),\ \min_{k\in[K]} N_k(t) > 0 \right\} \\ &\leq \inf\{ t \geq \max\{n_0, n_1, n_2\} : Z(t) \geq \zeta(t,\delta,\rho) \} \\ &\leq \inf\left\{ t \geq \max\{n_0, n_1, n_2\} : t(1-\varepsilon)\, T^\star(\boldsymbol{q}_0, \boldsymbol{\mu}_0) \right. \end{aligned} \tag{28}$$

$$\geq \log \frac{C\, t^{1+\rho}}{\delta}\Big\}$$

$\leq$ 最大{最大$\{n_0, n_1, n_2\}$, 下确界$\{t \geq 1 :\ tc_1 \geq \log(c_2 t^{1+\rho})\}\}$

$$\leq \max\left\{n_0, n_1, n_2, \frac{1}{c_1}\left(\log \frac{c_2 e}{c_1^{1+\rho}} + \log\log \frac{c_2}{c_1^{1+\rho}}\right)\right\}, \quad (29)$$

for all δ sufficiently small, where the last line above follows from the application of Lemma 4 with

$$c_1 = \frac{T^\star(q,\mu)(1-\varepsilon)}{1+\rho}, \quad c_2 = \frac{1}{\delta}, \quad \alpha = 1+\rho. \quad (30)$$

Eq. (29) immediately yields $\tau_\pi < \infty$ almost surely. Dividing by $\log(1/\delta)$ and taking limits as $\delta \downarrow 0$ in (29), we get

$$\limsup_{\delta \downarrow 0} \frac{\tau_\pi}{\log(1/\delta)} \leq \frac{1+\rho}{T^\star(q,\mu)(1-\varepsilon)} \quad a.s.. \quad (31)$$

Noting that (31) holds for all $\varepsilon > 0$, and taking limits as $\varepsilon \downarrow 0$, we arrive at the desired result.

Proof of part 3: Fix $\varepsilon > 0$ and $w^\star \in \mathcal{W}^\star(\boldsymbol{q}_0, \boldsymbol{\mu}_0)$. By virtue of the continuity of ψ, there exists $\xi_1(\varepsilon) > 0$ such that for all $(\boldsymbol{q}, \boldsymbol{\mu}, w')$ satisfying

$$\|\boldsymbol{q} - \boldsymbol{q}_0\| \leq \xi_1(\varepsilon),\ \|\boldsymbol{\mu} - \boldsymbol{\mu}_0\| \leq \xi_1(\varepsilon),\ d_\infty(w', w^\star) \leq \xi_1(\varepsilon), \quad (32)$$

我们有 $|\psi(\boldsymbol{q},\boldsymbol{\mu},w') - \psi(\boldsymbol{q}_0,\boldsymbol{\mu}_0,w^\star)| \leq \varepsilon\, \psi(\boldsymbol{q}_0,\boldsymbol{\mu}_0,w^\star) = \varepsilon T^\star(\boldsymbol{q}_0, \boldsymbol{\mu}_0)$。此外，根据 $(\boldsymbol{q},\boldsymbol{\mu}) \mapsto \mathcal{W}^\star(\boldsymbol{q},\boldsymbol{\mu})$ 的上半连续性（引理 1），存在 $\xi_2(\varepsilon) > 0$，使得

$$\|\boldsymbol{q} - \boldsymbol{q}_0\| \leq \xi_2(\varepsilon),\ \|\boldsymbol{\mu} - \boldsymbol{\mu}_0\| \leq \xi_2(\varepsilon)$$
$$\implies d_\infty(w', \mathcal{W}^\star(\boldsymbol{q}_0, \boldsymbol{\mu}_0)) \leq \frac{\xi_1(\varepsilon)}{10(M-1)}\ \forall w' \in \mathcal{W}^\star(\boldsymbol{q}, \boldsymbol{\mu}).$$

Set $\xi(\varepsilon) = \min\{\xi_1(\varepsilon)\ \xi_2(\varepsilon)\}$. For $N > 1$, let

$$\mathcal{G}_N = \bigcap_{t=N-1}^{\infty} \{\|\boldsymbol{q} - \hat{\boldsymbol{q}}(t)\| \leq \xi(\varepsilon), \|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(t)\| \leq \xi(\varepsilon)\}. \quad (33)$$

Under the event $\mathcal{G}_N$, we note that for all $t \geq N$,

$$d_\infty(w(t), \mathcal{W}^\star(\boldsymbol{q}, \boldsymbol{\mu})) \quad (34)$$
$$\leq \max_{w' \in \mathcal{W}^\star(\hat{q}(t-1), \hat{\mu}(t-1))} d_\infty(w', \mathcal{W}^\star(\boldsymbol{q}, \boldsymbol{\mu}))$$
$$\leq \frac{\xi_1(\varepsilon)}{10(M-1)}. \quad (35)$$

Furthermore, defining $\varepsilon_1 := \xi_1(\varepsilon)/(10(M-1))$, it follows from Lemma 2 that there exists $n_1 = n_1(\varepsilon_1)$ such that for all $t \geq n_1$, a.s.,

$$d_\infty((N(t,m)/t)_{m\in[M]},\ \mathcal{W}^\star(\boldsymbol{q},\boldsymbol{\mu})) \leq 10(M-1)\varepsilon_1 \leq \xi_1(\varepsilon). \quad (36)$$

Setting $N_0 = n_1$, and using (32) and (36), we see that $\mathcal{G}_N$ is a subset of the event

$$\bigcap_{t=N}^{\infty} \left\{\psi(\hat{\boldsymbol{q}}(t), \hat{\boldsymbol{\mu}}(t), (N(t,m)/t)_{m\in[M]}) \geq (1-\varepsilon)T^\star(\boldsymbol{q},\boldsymbol{\mu})\right\}$$

for all $N \geq N_0$. Let $N_1 = \inf\{t \geq 1 : \min_{k\in[K]} N_k(t) > 0\}$. Recall that $Z(t) = t\psi(\hat{\boldsymbol{q}}(t), \hat{\boldsymbol{\mu}}(t), (N(t,m)/t)_{m\in[M]})$ for all $t \geq N_1$. Then, for all $N \geq \max\{N_0, N_1\}$, we have under $\mathcal{G}_N$

$$\tau_\pi = \inf\left\{t \geq 1 : Z(t) \geq \zeta(t,\delta,\rho),\ \min_{k\in[K]} N_k(t) > 0\right\}$$
$$\leq \inf\{t \geq N : Z(t) \geq \zeta(t,\delta,\rho)\}$$
$$\leq \inf\{t \geq N : t\, c_1 \geq \log(c_2\, t)\}$$
$$\leq \max\{N,\ \inf\{t \geq 1 : c_1 t \geq \log(c_2 t^{1+\rho})\}\}$$
$$\leq \max\left\{N,\ \frac{1+\rho}{c_1}\left(\log \frac{c_2 e}{c_1^{1+\rho}} + \log\log \frac{c_2}{c_1^{1+\rho}}\right)\right\}, \quad (37)$$

where the last line above follows from Lemma 4, with c_1, c_2, and α as in (30). Also, in (37), we have $\log\log \frac{c_2}{c_1} = o(\log(1/\delta))$. Letting

$$N_2(\delta,\rho) := \frac{1+\rho}{(1-\varepsilon)\, T^\star(q,\mu)} \log\frac{1}{\delta} + o(\log(1/\delta)),$$

it follows from (37) that

$$\mathcal{G}_N \subset \{\tau_\pi \leq N\} \quad \forall N \geq \max\{N_0,\, N_1,\, N_2(\delta,\rho)\}. \quad (38)$$

Setting $N_3(\delta,\rho) = \max\{N_0,\, N_1,\, N_2(\delta,\rho)\}$, we then have

$$\tau_\pi = \tau_\pi\, \mathbf{1}_{\{\tau_\pi \leq N_3(\delta,\rho)\}} + \tau_\pi\, \mathbf{1}_{\{\tau_\pi > N_3(\delta,\rho)\}}$$
$$\leq N_3(\delta,\rho) + \max\{N_3(\delta,\rho),\, \tau_\pi\}$$
$$\leq N_2(\delta,\rho) + \max\{N_0,\, N_1\} + \tau, \quad (39)$$

where in writing (39), we use the relation $\max\{a,b,c\} \leq c + \max\{a,b\}$, and $\tau := \max\{N_3(\delta,\rho)\ \ \tau_\pi\}$. Taking expectations on both sides of (39), we get

$$\mathbb{E}[\tau_\pi] \leq N_2(\delta,\rho) + \max\{N_0,\, N_1\} + \mathbb{E}[\tau]. \quad (40)$$

We now note that

$$\mathbb{E}[\tau] = \sum_{N=0}^{\infty} P(\max\{N_3(\delta,\rho),\, \tau_\pi\} > N)$$
$$= \sum_{N=N_3(\delta,\rho)}^{\infty} P(\tau_\pi > N)$$
$$\leq \sum_{N=N_3(\delta,\rho)}^{\infty} P(\mathcal{G}_N^{\mathsf{c}})$$
$$\leq \sum_{N=\max\{N_0,\, N_1\}}^{\infty} P(\mathcal{G}_N^{\mathsf{c}}), \quad (41)$$

where the penultimate line is due to (38).

下面我们将证明，(41) 中的求和是有限的，并且可以用一个与 δ 无关的常数作为上界。事实上，由 (33) 可得

$$P(\mathcal{G}_N^{\mathsf{c}}) \leq \sum_{t=N-1}^{\infty} P(\|\hat{\boldsymbol{q}}(t) - \boldsymbol{q}_0\| > \xi(\varepsilon))$$
$$+ \sum_{t=N-1}^{\infty} P(\|\hat{\boldsymbol{\mu}}(t) - \boldsymbol{\mu}_0\| > \xi(\varepsilon))$$

$$
\begin{aligned}
&\overset{(a)}{\leq} \sum_{t=N-1}^{\infty} c_1\, t^{M/4} \exp(-c_2 \xi(\varepsilon)^2 \sqrt{t}) \\
&\qquad + \sum_{t=N-1}^{\infty} c_3\, t^{K/4} \exp(-c_4 \xi(\varepsilon)^2 \sqrt{t}) \\
&\leq \int_{N-2}^{\infty} c_1\, s^{M/4} \exp(-c_2 \xi(\varepsilon)^2 \sqrt{s})\, ds \\
&\qquad + \int_{N-2}^{\infty} c_3\, s^{K/4} \exp(-c_4 \xi(\varepsilon)^2 \sqrt{s})\, ds \\
&= O\bigg(\frac{(N-2)^{M/2+1}}{\xi(\varepsilon)^4\, \exp(c_2\, \xi(\varepsilon)^2\, \sqrt{N-2})}\bigg) \\
&\qquad + O\bigg(\frac{(N-2)^{K/2+1}}{\xi(\varepsilon)^4\, \exp(c_4\, \xi(\varepsilon)^2\, \sqrt{N-2})}\bigg), \quad (42)
\end{aligned}
$$

其中，(a) 是由关于次高斯随机变量 [14]，的标准集中不等式得到的；最后一行则是通过用不完全伽马函数对倒数第二行中的积分取上界，并利用一些经典不等式 [15], [16] 对这些函数取上界而得。注意到，对 $N \geq \max\{N_0,\, N_1\}$ 求和，(42) 右侧的结果是有限的（记为 b），且与 δ 无关。将 (41) 和 (42) 代入 (40)，可得

$$
\begin{aligned}
&\frac{\mathbb{E}[\tau_\pi]}{\log(1/\delta)} \leq \frac{N_2(\delta,\rho)}{\log(1/\delta)} + \frac{\max\{N_0\, \varepsilon_1^3,\, N_1\} + b}{\log(1/\delta)} \\
&= \frac{1+\rho}{(1-\varepsilon)\, T^\star(q,\mu)} + \frac{o(\log(1/\delta)) + \max\{N_0\, \varepsilon_1^3,\, N_1\} + b}{\log(1/\delta)}. \quad (43)
\end{aligned}
$$

Taking limits as $\delta \downarrow 0$ followed by $\varepsilon \downarrow 0$, we arrive at (10).

Appendix E
Proof of Theorem 3

To prove Theorem 3, we use the following concentration inequalities. For the proofs, we refer the reader to [14].

Lemma 5: Let $X_1, X_2, \ldots, X_n$ be n i.i.d. sub-Gaussian random variables with mean μ and variance σ^2. Then, for any $\epsilon > 0$, we have

$$
P\bigg(\bigg|\frac{\sum_{i=1}^n X_i}{n} - \mu\bigg| > \epsilon\bigg) \leq 2\exp\bigg\{-\frac{n\epsilon^2}{2\sigma^2}\bigg\}. \quad (44)
$$

Lemma 6: Let $X_1, X_2, \ldots, X_n$ be n i.i.d. Bernoulli random variables with mean p. Then, for any $0 < \gamma < 1$, we have

$$
P\bigg(\sum_{i=1}^n X_i \leq (1-\gamma)np\bigg) \leq \exp\bigg(-\frac{np\gamma^2}{2}\bigg). \quad (45)
$$

The following result formalises that for all (m,k) pairs, the empirical mean $\hat{\mu}_{m,k}(n)$ of the arm $A_{m,k}$ lies between its upper and lower confidence bounds defined in Section IV with high probability.

Lemma 7: Let $\mathcal{E}_1$ be the following event:

$$
\mathcal{E}_1 \coloneqq \bigcap_{n\in\mathbb{N}} \bigcap_{m=1}^{M} \bigcap_{k\in\mathcal{A}_m} \{\mathrm{LCB}_{m,k}(n) \leq \mu_{m,k}(n) \leq \mathrm{UCB}_{m,k}(n)\}.
$$

Then, $P(\mathcal{E}_1) \geq 1 - \frac{\delta}{2}$.

The proof of Lemma 7 uses Lemma 5 and is quite standard; see, for instance, [17, Lemma 2] for a proof template.

Lemma 8: Conditioned on $\mathcal{E}_1$, the following hold w.p.1.

- BBSEA outputs the correct best arm.
- $T_{m,k} \leq \alpha_{m,k}$ for all $m \in [M]$ and $k \in \mathcal{A}_m$.

Lemma 8 informs that (a) as long as the arm means lie within their respective confidence intervals, BBSEA gives correct output, and (b) arm $A_{m,k}$ is identified as optimal or sub-optimal in at most $\alpha_{m,k}$ rounds of BBSEA. Again, the proof of Lemma 8 is quite standard and omitted; see, for instance, [17, Theorem 1] and [17, Theorem 2] for a proof template.

Lemma 9: Let $\mathcal{E}_2$ be the event that $\beta_{m,k}$ selections of box m gives rise to at least $\alpha_{m,k}$ pulls of arm $A_{m,k}$ for all $m \in [M]$ and $k \in \mathcal{A}_m$. Then, $P(\mathcal{E}_2) \geq 1 - \frac{\delta}{2}$.

Proof of Lemma 9: Recall from our problem setup that when the underlying instance is $(\boldsymbol{q}_0, \boldsymbol{\mu}_0)$, selection of box m leads to the pulling of arm $A_{m,k}$ with probability $q^0_{m,k}$. Fix (m,k), and define

$$
X_i = \begin{cases} 1, & A_{m,k} \text{ is pulled at the } i\text{th selection of box } m, \\ 0, & A_{m,k} \text{ is not pulled at the } i\text{th selection of box } m. \end{cases}
$$

Clearly, $X_1, X_2, \ldots$ are i.i.d. and follow Bernoulli distribution with mean $q^0_{m,k}$. Applying Lemma 6 with $n = \beta_{m,k}$,

$$
\gamma = \frac{2\sqrt{\ln\big(\frac{2K}{\delta}\big)}}{\sqrt{\ln\big(\frac{2K}{\delta}\big) + \alpha_{m,k}} + \sqrt{\ln\big(\frac{2K}{\delta}\big)}},
$$

以及 $p = q^0_{m,k}$，我们可以得到 $\beta_{m,k}$ 次盒子 m 的选择以大于 $\alpha_{m,k}$ 的概率产生了对臂 $A_{m,k}$ 的拉取。通过对所有 (m,k) 对应用并集界，并注意到 $\sum_{m=1}^M |\mathcal{A}_m| = K$，即可完成证明。 ■

有了上述准备工作，我们接下来证明定理 3。注意，$P(\mathcal{E}_1 \cap \mathcal{E}_2) \geq 1 - \delta$. 定理 3 的第一部分可直接由引理 8 得出。为了证明第二部分，令 $U'_{m,k}$ 表示为了至少 $T_{m,k}$ 次拉取臂 $A_{m,k}$ 而需要对盒子 m进行选择的随机数量。令$U'_m \coloneqq \max_{k\in\mathcal{A}_m} U'_{m,k}$。

Lemma 10: On the event $\mathcal{E}_1 \cap \mathcal{E}_2$, the following hold for all $m \in [M]$ and $k \in \mathcal{A}_m$ w.p.1.

1) $U'_{m,k} \leq \beta_{m,k}$.
2) $U'_m \leq \beta_m$.

引理 10 的证明：回顾可知，在事件 $\mathcal{E}_2$, $\beta_{m,k}$ 下，选择盒子 m 会产生 $\alpha_{m,k}$ 次对臂 $A_{m,k}$ 的拉取，对所有 (m,k) 均成立。利用这一点以及在 $\mathcal{E}_1$ 下（引理 8）对所有 (m,k) 均有 $T_{m,k} \leq \alpha_{m,k}$ 这一事实，我们可以推导出 $U'_{m,k} \leq \beta_{m,k}$，从而证明第 1 部分。第 2 部分只需对第 2 部分的结果关于 $k \in \mathcal{A}_m$ 取最大值，并利用每个 $m \in [M]$ 对应的 U'_m 和 β_m 的定义即可得出。 ■

定理 3 的第 2 部分可由引理 10 以及 BBSEA 的停止时间（即 BBSEA 下盒子选择的总次数）等于 $\sum_{m=1}^M U'_m$ 立即得出。定理 3 的第 3 部分可由多臂老虎机的标准测度变换论证 [3,引理 1] 得出。证明完毕。