Bandits with Dynamic Arm-acquisition Costs*

* A longer version of this paper is available at https://arxiv.org/abs/2110.12118

Anand Kalvit  
Columbia University  
akalvit22@gsb.columbia.edu

Assaf Zeevi  
Columbia University  
assaf@gsb.columbia.edu

Abstract—We consider a bandit problem where at any time, the decision maker can add new arms to her consideration set. A new arm is queried at a cost from an arm-reservoir containing finitely many arm-types, each characterized by a distinct mean reward. The cost of query reflects in a diminishing probability of the returned arm being optimal, unbeknown to the decision maker; this feature encapsulates defining characteristics of a broad class of operations-inspired online learning problems, e.g., those arising in markets with churn, or those involving allocations subject to costly resource acquisition. The decision maker's goal is to maximize her cumulative expected payoffs over a sequence of $n$ pulls, oblivious to the statistical properties as well as types of the queried arms. We study two natural modes of endogeneity in the reservoir distribution, and characterize a necessary condition for achievability of sub-linear regret in the problem. We also discuss a UCB-inspired adaptive algorithm that is long-run-average optimal whenever said condition is satisfied, thereby establishing its tightness.

Index Terms—Many-armed bandits, arm-reservoir, reservoir distribution, non-stationarity, endogeneity, regret minimization

I. INTRODUCTION

Background. The stochastic multi-armed bandit (MAB) problem [1] is a classical machine learning paradigm that provides a succinct abstraction of the exploration-exploitation trade-offs inherent to sequential decision making under uncertainty. This problem has guided theory and practice over the years, and appears as a basal motif in several complex data-driven applications; see [2] for a survey of well-studied formulations. In its simplest formulation, the decision maker must pull at each time instant $t \in \{1,2,\ldots\}$ one out of a set of $K$ possible arms with unknown reward distributions (and any statistical property thereof) in a way that maximizes her cumulative expected payoffs over a sequence of $n$ pulls. A natural metric to assess the performance of an algorithm in this setting is its regret w.r.t. an oracle that recommends an optimal arm at each step. The seminal work [1] was the first to show that the best achievable regret in “well-separated” instances of the problem is $\Theta(\log n)$.

Towards countably many arms. In several applications of MAB, the complexity of decision making resides not only in identifying an optimal arm among $K$ competing alternatives, but is also attributable to the choice of allocating effort across said alternatives and certain “outside options;” this characteristic is usually referred to as choice overload in decision theory.

One natural approach in the literature for modeling choice overload in MAB involves endowing the problem with an arm-reservoir that returns upon query “new” arms sampled from a distribution over arm-types, see. e.g., [3]. Among problems of this nature, a simple yet instructional setting is one where the reservoir only holds finitely many arm-types [4]. Such finite-typed models appear frequently also in learning problems in the broader operations research literature, see, e.g., [5].

Cost of reservoir access. In MAB settings where choice overload is a defining characteristic, the decision maker may experience elevated costs of acquiring new arms as time progresses. This can be viewed in light of plausible deterioration in the “quality” of the reservoir either with increasing number of queries or with time. The former can be interpreted as a Lagrangian relaxation to an optimization problem with capacity constraints on reservoir queries; a setting potentially of interest to online resource allocation problems under costly resource acquisition. The latter aspect has connotations related to market churn. For example, the relative abundance of agents of different types in a market may undergo temporal variations over the platform’s planning horizon. This is the case, for example, in online service-platforms, where arms may represent agents capable of abandoning the platform if kept idle for protracted durations. These effects may catalyze an agent-departure process and a temporal non-stationarity that may seriously hinder the decision maker’s ability to discern “good” arms from “inferior” ones. There has been significant recent interest in non-stationary bandit models, however, most of the literature is limited largely to non-stationarities in rewards, and antecedents on non-stationarities in arm-reservoirs are markedly absent (see the literature review in §I-A). In this paper, we provide the first systematic treatment of this aspect by investigating the statistical limits of learning under endogenous variations in the arm-reservoir distribution.

Stylized model. To distill key insights, it will be convenient to focus on the paradigmatic case of exactly two arm-types in the reservoir; this serves to highlight statistical idiosyncrasies of the problem without unnecessary mathematical detail. The complexity of the static version of the problem is driven by three primitives: (i) the gap $\Delta$ between the mean rewards associated with the two arm-types; (ii) the fraction $\alpha$ of the “optimal” arm-type in the reservoir; and (iii) the horizon of play $n$. The feature of costly arm-acquisitions is incorporated...
by endogenizing $\alpha$; consequently, its evolution is given by the stochastic process $\alpha(t) : t = 1, 2, \ldots$. Thus, at any time $t$, the probability that a query of the arm-reservoir returns an arm of the optimal type is given by (the sample path-dependent random variable) $\alpha(t)$. As we shall later see, the interesting regime is where $\alpha(t)$ becomes vanishingly small as $t$ grows; this captures departure of optimal (unexplored) arms from the reservoir, and abstracts out key characteristics of many online platforms that serve a large population of impatient agents.

**Challenges due to costly reservoir access.** It is non-trivial to design “good” policies that are reservoir distribution-agnostic as well as gap-adaptive. To see this, consider the simplest scenario where $\alpha(t) = c$ (constant), and the decision maker is endowed with ex ante knowledge of $c$ as well as the horizon $n$. A natural heuristic in this setting is to query $\Omega(\log n)$ arms upfront (this would guarantee with probability $\Omega(1/n)$ the existence of at least one optimal arm among the queried ones), and subsequently deploy a conventional bandit algorithm such as Thompson Sampling [6] or UCB [7] on the collected set of arms. One can quite easily show that such an approach will incur poly-logarithmic regret in $n$. On the other hand, it is possible to achieve $O((c\Delta)^{-1} \log n)$ regret (up to poly-logarithmic factors in $\Delta^{-1}$) using a more sophisticated policy [8]. Regardless, none of these approaches utilize the fact that there exist exactly $K$ arm-types in the reservoir. By leveraging this knowledge, $c$-dependence of the aforementioned upper bound can, in fact, be relegated to sub-logarithmic terms, surprisingly, in a manner that is adaptive to $c$ [4]. It is noteworthy, however, that the performance of such approaches is fragile w.r.t. the premise that $\alpha(t)$ remains bounded away from 0 by some problem-independent constant $c$ (which may or may not be known a priori) at all times. Naturally, endogenous variations potentially causing $\alpha(t)$ to vanish in $t$ will only exacerbate the problem, and it remains unclear if it is even possible in this setting to achieve sub-linear regret relative to the classical full-information benchmark that prescribes pulling an optimal arm at each $t$.

**Contributions.** We first derive a necessary condition for “complete learning” when the evolution of $\alpha(t)$ is independent of the decision maker’s actions; specifically, $\sum_{t=1}^{\infty} \alpha(t) = \infty$ is necessary for achieving sub-linear regret relative to an oracle that knows the identities of optimal arms ex ante (Theorem 1). We also establish its near-tightness in that a slightly stronger version of said condition is sufficient for a gap-aware policy based on the Explore-then-Commit principle to achieve poly-logarithmic instance-dependent regret in the problem (Theorem 2). In addition, we discuss a novel gap-adaptive policy based on the UCB principle that achieves a polynomial regret in the same regime (Theorem 3). We then consider the setting where $\alpha(t)$ is endogenous (policy-dependent), and characterize matching necessary and sufficient conditions (up to leading order terms) for asymptotic-optimality of aforementioned policies.

Before proceeding with a formal description of our model and discussion of results, we provide a brief overview of related literature below.

**A. Related literature**

**Bandits with state-dependent rewards.** The earliest work on MAB problems involving endogenous arm-reward distributions dates back to the seminal work [9] which studied finite-state Markovian bandits. In this model, the state of an arm only changes upon execution of a pull while remaining unchanged otherwise, thus prompting the name resting bandits; the celebrated Gittins index policy is well-known to maximize the infinite-horizon discounted cumulative expected reward in this setting. In contrast, in the so-called restless bandits formulation [10], the states of all the arms may change simultaneously irrespective of which arms are pulled; in addition, this formulation permits pulling any fixed number of arms in each period. Subsequent works such as [11] focus on heuristics that are optimal in an asymptotic regime where the number of arms pulled in each period scales linearly with the total number of arms. More recently, a finite-horizon variant of the restless bandits problem was studied in [12] under a similar scaling; see [13]–[15], etc., for a survey of other well-studied variations. Our work is quite distinct from this strand of literature: (i) asymptotic analyses in aforementioned papers are w.r.t. the number of arms, not the horizon of play; (ii) the number of arms pulled in each period is fixed at 1 in our problem setting and does not scale with the total number of arms; (iii) most importantly, cited references formulate the problem as a Markov decision process assuming full knowledge of the transition kernels. In contrast, we consider a learning theoretic formulation where the decision maker is oblivious to the statistical properties of reward distributions as well as the nature of endogeneity in the arm-reservoir.

**Bandits with non-stationary rewards.** This line of work focuses on policies that minimize the expected cumulative regret relative to a dynamic oracle that plays at each time $t$ an arm with the highest mean reward at $t$. Some of the early work in this area is premised on a formulation in which the identity of the best arm may change a finite number of times adversarially during the horizon of play, see, e.g., [16]. While other works such as [17] study specific models of temporal variation where, for example, rewards evolve according to a Brownian motion, much of the traditional literature is limited, by and large, to a finite number of changes in the mean rewards; see [18] and references therein. Subsequently, a unified framework for studying aforementioned problem classes was provided in [19] by introducing a variation budget to bound the evolution of mean rewards over the horizon of play. Several other forms of non-stationarity have also been studied in the literature; these include formulations with rotting [20], recharging [21], and delay-dependent rewards [22], among others (see [23] for a survey).
Aforementioned works are largely limited to the study of finite-armed bandit problems where non-stationarity can be ascribed to changes in arm-means. In contrast, our work differs fundamentally in that it is premised on an infinite-armed formulation with non-stationarity attributable to an endogenous arm-reservoir. In a nutshell, while preceding work focuses on distributional shifts in rewards given a fixed set of arms, we propose a new paradigm where the arm-reservoir itself undergoes distributional shifts owing to possible “leakages,” which is functionally a very different concept.

**Bandits with infinitely many arms.** These problems involve settings where an unlimited supply of arms is governed by some fixed distribution over an uncountable set of arm-types (possible mean rewards); a common reward statistic (usually the mean) uniquely characterizes each arm-type. The infinite-armed bandit problem traces its roots to [3] where it was studied under the Bernoulli reward setting with means distributed Uniformly on [0, 1]. More general reward and reservoir distributions on [0, 1] have also been studied in subsequent works, see, e.g., [24]–[27]. Our model differs from these in that we only assume the reservoir to be finitely supported and posit no distributional knowledge thereof, unlike cited references. Furthermore, the distribution of arm-types is allowed to vary endogenously over the problem horizon in a manner that may be unknown ex ante.

**B. Outline of the paper**

A formal description of the model is provided in §II. §III discusses reservoir distribution-agnostic algorithms; two natural modes of endogeneity in the reservoir distribution are discussed in §IV along with corresponding guarantees on achievable performance. §V provides concluding remarks. Proofs and ancillary results are delegated to the online appendix.

**II. PROBLEM FORMULATION**

**Primitves.** There are finitely many possible arm-types in the reservoir denoted by the collection $\mathcal{K}$; the vector $\mu := (\mu_k : k \in \mathcal{K})$ characterizes the mean reward (pairwise distinct) associated with each. The decision maker (DM) only knows the cardinality of $\mathcal{K}$; for simplicity of exposition, we assume $|\mathcal{K}| = 2$ in this work and index the two arm-types by “1” and “2,” i.e., $\mathcal{K} = \{1, 2\}$. Without loss of generality, arm-type 1 is assumed “optimal” with a gap (or separation) of $\Delta := \mu_1 - \mu_2 > 0$ from the “inferior” type (arm-type 2); as we shall later see, $\Delta$ is an important driver of the problem’s statistical complexity. DM must play one arm at each time $t \in \{1, \ldots, n\}$, where $n$ denotes the horizon of play.

**Reservoir access.** The collection of arms to have been played at least once until time $t$ (inclusive) is denoted by $\mathcal{I}_t$ (with $\mathcal{I}_0 := \emptyset$). The set of available actions at $t$ is given by $\mathcal{A}_t = \mathcal{I}_{t-1} \cup \{\text{new}_{t}\}$; DM must either play an arm from $\mathcal{I}_{t-1}$ or select the action “new,” which corresponds to playing a new arm queried from the reservoir. A newly queried arm at time $t$ is optimal-typed with probability $\alpha(t)$ and inferior-typed otherwise ($\alpha(t)$ may potentially be random and endogenous, unbeknown to DM; a detailed discussion is deferred to §IV). Arm-types are private attributes and remain unobservable throughout. Upon pulling an arm labeled $i$ (henceforth referred to as arm $i$) for the $t^{th}$ time, DM observes a $[0, 1]$-valued stochastic reward denoted by $X_{i,t}$. The realized rewards are independent across arms and time, and mean-preserving in time (not necessarily identically distributed) keeping the arm fixed.

**Admissible controls.** A policy $\pi := (\pi_1, \pi_2, \ldots)$ is an adaptive allocation rule that prescribes at time $t$ an action $\pi_t$ (possibly randomized) from $\mathcal{A}_t$. The collection of all observable information until $t$ is given by the natural filtration $\mathcal{F}_t := \sigma \left\{ (\pi_s)_{1 \leq s \leq t}, \{ (X_{i,j})_{1 \leq j \leq N_i(t) : i \in \mathcal{I}_t} \} \right\}$ (with $\mathcal{F}_0 := \emptyset$), where $N_i(t)$ indicates the number of times arm $i$ is pulled until $t$. The cumulative regret of $\pi$ after $n$ plays is given by $R_n := \sum_{t=1}^n (\mu_1 - X_{\pi_t}N_{\pi_t}(t))$ and its cumulative pseudo-regret by $\hat{R}_n := \sum_{t=1}^n (\mu_1 - \mu_{\pi_t}(\pi_t))$, where $\pi_t : K \rightarrow \{1 \}$ encodes the type of arm $\pi_t$; note that both regret as well as pseudo-regret are sample path-dependent by definition. DM is interested in the following stochastic minimization problem

$$
\inf_{\pi \in \Pi} \mathbb{E} R_n^{\pi} = \inf_{\pi \in \Pi} \mathbb{E} \left[ \sum_{t=1}^n (\mu_1 - X_{\pi_t}N_{\pi_t}(t)) \right] = \inf_{\pi \in \Pi} \mathbb{E} \left[ \sum_{t=1}^n (\mu_1 - \mu_{\pi_t}(\pi_t)) \right]
$$

where the infimum is over the class $\Pi$ of non-anticipating policies, i.e., $\pi_t : \mathcal{F}_{t-1} \rightarrow \mathcal{P}(\mathcal{A}_t); t\in\{1,2,\ldots\}$ ($\mathcal{P}(\mathcal{A}_t)$ denotes the probability simplex on $\mathcal{A}_t$), the expectations are w.r.t. all the possible sources of stochasticity in the problem (rewards, policy, and the reservoir distribution), and (†) holds since cumulative regret and pseudo-regret are equal in expectation in the stochastic bandits setting (follows from the Tower property of expectations).

**III. DISTRIBUTION-AGNOSTIC POLICIES**

As discussed in the introduction, our goal in this work is to investigate achievable regret under algorithms that are agnostic to the reservoir distribution. To this end, we discuss two such algorithms; one based on the forced-exploration principle, and the other on optimism-under-uncertainty.

**A. A fixed-design policy based on forced exploration**

The non-adaptive ETC (Explore-then-Commit) approach outlined below is predicated on ex ante knowledge of the problem horizon $n$ (this is not a constraining factor since the exponential doubling trick [28] can be used to make the algorithm horizon-free) and a gap parameter $\delta \in (0, \Delta]$. In what follows, a new arm refers to one that is freshly queried from the arm-reservoir.
**Algorithm 1 ALG1(n, δ) (Non-adaptive ETC)**

1. **Input:** Horizon of play $n$, gap parameter $δ$.
2. Set $m = \lceil 2δ^{-2} \log n \rceil$.
3. **New epoch:** Play $K = 2$ new arms (call $A = \{1, 2\}$) from the reservoir; observe rewards ($X_{1,1}, X_{2,1}$).
4. Play each arm $m - 1$ times more; observe rewards ($X_{1,j}, X_{2,j} : j = 2, ..., m$).
5. If $\sum_{j=1}^{m} (X_{1,j} - X_{2,j}) < δm$ then
6. Permanently discard $A$ and repeat from step (3).
7. else
8. Commit to arm $i^* \in \arg \max_{i \in A} \{\sum_{j=1}^{m} X_{i,j}\}$.

**Policy dynamics.** The horizon is divided into epochs of length $2m = \Theta(\log n)$ each. In each epoch, the algorithm re-initializes by querying the arm-reservoir for a pair of new arms, and playing them $m$ times each. Subsequently, the pair is classified as either “distinct” or “identical”-typed via a hypothesis test (step 5 of Algorithm 1). If classified as distinct, the algorithm commits the residual budget of play to the empirically superior arm among the two (with ties broken arbitrarily). On the other hand, if the pair is classified as identical, the algorithm discards it permanently and ushers in a new epoch. The entire process repeats until a distinct-typed pair is identified. ALG1(n, δ), albeit non-adaptive to $\Delta$, serves as an insightful basal motif for algorithm design and its operating principle will guide the development of the $\Delta$-adaptive algorithm discussed next.

**B. A UCB-based approach with adaptive resampling**

In this section, we revisit the UCB-based adaptive policy proposed in [4] for the static version of our problem where $\alpha(t)$ is constant. This policy is restated as ALG2 after suitable modifications for reasons discussed next. The original policy (Algorithm 2 in cited reference) suffers a limitation through its dependence on ex ante knowledge of the support of reward distributions. In particular, the algorithm requires reward distributions associated with the arm-types to be “maximally supported” on $[0, 1]$, e.g., only distributions such as Bernoulli$(i)$, Uniform on $[0, 1]$. Beta$(\cdot, \cdot)$, etc., are amenable to its performance guarantees; Uniform on $[0, 0.5]$, on the other hand, is not. We identify a simple fix to this issue and propose adding a centered Gaussian noise term to the cumulative-difference-of-reward statistic (see step (5) of Algorithm 2) to essentially create an unbounded support. This rids the algorithm of its dependence on assumptions pertaining to the support, and at the same time also preserves its regret guarantees in the static setting. In fact, our modifications lead to a sharper characterization of the scaling behavior of regret w.r.t. $\Delta$ in the static setting (see Theorem 3 and 5).

**Remark.** The original policy (Algorithm 2 in cited reference) uses a threshold that is distinct from $4\sqrt{m \log m}$ (see step (5) of Algorithm 2); the choice of $4\sqrt{m \log m}$ here aims to simplify the exposition of key technical innovations and facilitate an easy comparison between various upper bounds.

**Algorithm 2 ALG2 (Nested UCB)**

1. **New epoch** ($t \leftarrow 0$): Play $K = 2$ new arms (call $A = \{1, 2\}$) from the reservoir; observe rewards ($X_{1,1}, X_{2,1}$).
2. Minimum per-arm sample count $m \leftarrow 1$.
3. Generate an independent sample $Z$ of a standard Gaussian.
4. for $t \in \{3, 4, \ldots\}$ do
5. if $\left| Z + \sum_{j=1}^{m} (X_{1,j} - X_{2,j}) \right| < 4\sqrt{m \log m}$ then
6. Permanently discard $A$ and repeat from step (1).
7. else
8. Play $i_t \in \arg \max_{i \in A} \left\{ \frac{\sum_{j=1}^{N_i(t-1)} X_{i,j}}{N_i(t-1)} + \sqrt{\frac{2 \log(t-1)}{N_i(t-1)}} \right\}$.
9. Observe reward $X_{i_t, N_i(t)}$.
10. if $m < \min_{i \in A} N_i(t)$ then
11. $m \leftarrow m + 1$.

**Policy dynamics.** ALG2 also has an episodic dynamic with exactly one pair of arms played per episode. It is noteworthy that ALG2 plays arms according to UCB1 [7] in every episode as opposed to playing them equally often until committing to the empirically superior one à la ALG1(n, δ). Secondly, ALG2 never “commits” to an arm (or a consideration set); the implication is that the algorithm will keep querying new consideration sets from the reservoir throughout the horizon of play. Aforementioned adaptive resampling property is at the core of its horizon-free nature.

**Operating principle.** At any time, ALG2 computes a threshold of $O(\sqrt{m \log m})$ for the length-$m$ cumulative-difference-of-reward process, where $m$ denotes the minimum sample count among the two arms. If the envelope of said process is dominated by $O(\sqrt{m \log m})$, the arms are likely to belong to the same type (simultaneously optimal or inferior). The explanation stems from the Law of the Iterated Logarithm (see [29], Theorem 8.5.2): a zero-drift length-$m$ random walk process has its envelope bounded by $O(\sqrt{m \log m \log \log m})$. In the aforementioned scenario, the algorithm discards the consideration set and subsequently, a new epoch is ushered in. This is done to avoid the possibility of incurring linear regret in the event that the two arms are inferior, since it is statistically impossible to distinguish a simultaneous-inferior consideration set from one where both arms are optimal, in the absence of any auxiliary information such as the mean rewards associated with the two types. On the other hand, if the cumulative-difference-of-reward dominates $O(\sqrt{m \log m})$, the consideration set is likely to contain arms of distinct types and ALG2 continues to run UCB1 on this set until the $O(\sqrt{m \log m})$ threshold is breached again.

**Reason for introducing the Gaussian corruption.** Centered Gaussian noise is added to the cumulative-difference-of-reward process in step (5) of ALG2 to avoid the possibility of incurring linear regret should the support of the reward distributions be a “very small” subset of $[0, 1]$. To illustrate this point, suppose that the rewards associated with the types are deterministic with $\Delta < 2\sqrt{2 \log 2}$. Then, as soon as the algorithm queries a heterogeneous consideration set (one arm
optimal and the other inferior) and the per-arm sample count reaches 2, the cumulative-difference-of-reward statistic will satisfy
\[ \sum_{j=1}^{2} (X_{1,j} - X_{2,j}) = 2\Delta < 4\sqrt{2} \log 2, \]
resulting in the consideration set getting discarded. On the other hand, if the consideration set is homogeneous (both arms optimal or inferior), the algorithm will still re-initialize within a finite number of samples in expectation (again, owing to the Law of the Iterated Logarithm). This will force the algorithm to keep querying new arms from the reservoir at rate that is linear in time, which is tantamount to incurring linear regret in the horizon. The addition of centered Gaussian noise hedges against this risk by guaranteeing that the cumulative-difference-of-reward process essentially has an infinite support at all times (even when the reward distributions might be degenerate). This rids the regret performance of its fragility w.r.t. ex ante knowledge of the support of reward distributions. The next proposition crystallizes this discussion.

**Proposition 1 (Persistence of heterogeneous consideration sets):** Let \( \{X_{i,j} : j = 1, 2, \ldots\} \) be a collection of independent samples from an arm of type \( i \) (implying \( E[X_{i,j} - X_{2,j}] = \Delta \) \( \forall \ j = 1, 2, \ldots \)). Let \( Z \) be an independently generated standard Gaussian random variable. Then,
\[
P \left( \bigcap_{m \geq 1} \left\{ \left| Z + \sum_{j=1}^{m} (X_{1,j} - X_{2,j}) \right| \geq 4\sqrt{m \log m} \right\} \right) > \frac{\bar{\Phi}(f(T_0))}{2} =: \beta_\Delta > 0, \tag{2} \]
where \( \bar{\Phi}(v) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{v} e^{-u^2/2} du \) \( \forall \ v > 0 \) is the right-tail of the standard Gaussian CDF, \( T_0 := \left[ (64/\Delta^2) \log^2 (64/\Delta^2) \right] \) and \( f(x) := x + 4\sqrt{x \log x} \) \( \forall \ x \geq 1 \).

**Interpretation of \( \beta_\Delta \).** First of all, note that \( \beta_\Delta \) admits a closed-form characterization in terms of standard functions and satisfies \( \beta_\Delta > 0 \) when \( \Delta > 0 \) with \( \lim_{\Delta \to 0} \beta_\Delta = 0 \). Secondly, \( \beta_\Delta \) depends exclusively on \( \Delta \), and represents a lower bound on the probability that \( \Delta \log 2 \) will never discard a consideration set containing arms of distinct types. This meta-result will be key to the upper bounds stated in forthcoming sections.

**IV. NATURAL MODELS OF THE ARM-RESERVOIR**

The probability \( \alpha(t) \) of a newly queried arm at time \( t \) being optimal-typed will likely vary over the horizon of play in realistic settings. For example, in the context of crowdsourcing applications, the availability of “high quality” workers for a given task may depend on the prevailing population-level perception of the platform. This could plausibly be a function of the age \( t \) of the platform. Alternatively, the reservoir may react to a query at time \( t \) through its cumulative query count \( J_t \); this model may be suited to settings where, for example, acquiring a new resource is costly and yields diminishing returns. We capture these aspects through two reservoir models described next.

**Model 1 (Exogenous reservoir):** \( \alpha(t) : t = 1, 2, \ldots \) is a non-increasing deterministic process with \( \alpha(1) = c \in (0, 1) \), evolving independently of the decision maker’s actions.

**Model 2 (Endogenous reservoir):** \( \alpha(t) : t = 1, 2, \ldots \) evolves as \( \alpha(t) = g(J_{t-1}) \), where \( g : \mathbb{N} \cup \{0\} \to (0, c] \) is a non-increasing deterministic mapping with \( g(0) = c \in (0, 1) \), and \( J_t \) denotes the number of reservoir queries until time \( t \) (inclusive) with \( J_0 := 0 \).

Our first result below states a necessary condition for achievability of sub-linear regret in the two reservoir models.

**Theorem 1 (Necessary conditions for “complete learning” in the two reservoir models):**

1) Under Model 1, if \( \sum_{t=1}^{\infty} \alpha(t) < \infty \), the expected cumulative regret of any policy \( \pi \) grows linearly in the horizon of play, i.e., \( \mathbb{E}R^n_t = \Omega(\Delta n) \), where the \( \Omega(\cdot) \) only hides multiplicative constants independent of \( \Delta \) and \( n \). Equivalently, a necessary condition for achieving sub-linear regret in the problem is \( \sum_{t=1}^{\infty} \alpha(t) = \infty \).

2) Under Model 2, if \( \sum_{t=1}^{\infty} g(t) < \infty \), the expected cumulative regret of any policy \( \pi \) grows linearly in the horizon of play, i.e., \( \mathbb{E}R^n_t = \Omega(\Delta n) \), where the \( \Omega(\cdot) \) only hides multiplicative constants independent of \( \Delta \) and \( n \). Equivalently, a necessary condition for achieving sub-linear regret in the problem is \( \sum_{t=1}^{\infty} g(t) = \infty \).

The proof relies essentially on reduction to a full-information setting where the decision maker observes the true mean of an arm immediately upon play (see the online appendix for details). The optimal policy is this setting will keep querying the reservoir for new arms until it finds one with the optimal mean, which it will subsequently commit the rest of its sampling budget to. The conditions stated in Theorem 1 are necessary for this policy to find an optimal arm within its lifetime. It is only natural that the same condition is necessary for achievability of sub-linear regret in the general setting where the decision maker only observes a noisy version of the true means. Surprisingly, however, these conditions are also almost-sufficient for sub-linear regret (as forthcoming results will show), and are therefore nearly-tight.

**A. Exogenous reservoirs**

The focus here will be on settings specified by Model 1. Theorem 1 establishes a necessary condition of \( \sum_{t=1}^{\infty} \alpha(t) = \infty \) for achievability of sub-linear instance-dependent regret in the problem. In what follows, we show a slightly more refined characteristic: \( \tilde{\Theta}(t^{-1}) \) is, in fact, a critical rate for “complete learning” in that policies achieving sub-linear regret exist if \( \alpha(t) = \omega((\log t)/t) \).

To elucidate the criticality of the \( \tilde{\Theta}(t^{-1}) \) rate, it will be convenient to consider a parameterization of \( \alpha(t) \) given by \( \alpha(t) = ct^{-\gamma} \), where \( c \) is as specified in Model 1 and \( \gamma \in [0, 1) \) is a fixed parameter. This parameterization offers meaningful insights as to the statistical complexity of the problem w.r.t.
\(\gamma\) and facilitates an easy comparison between the regret guarantees of the algorithms discussed in §III. We will begin with an upper bound on the performance of Algorithm 1.

**Theorem 2 (Upper bound for \textit{ALG}1(n, \delta))**: Under Model 1 with \(\alpha(t) = ct^{-\gamma}\), where \(\gamma \in [0, 1)\), the expected cumulative regret of the policy \(\pi\) given by Algorithm 1 satisfies

\[
\limsup_{n \to \infty} \frac{\mathbb{E} R_n^\pi}{\log n} \leq 24\Delta \left( \frac{8}{\delta^2 c} \right)^{\frac{1}{1-\gamma}} \hat{\delta} \left( \left\lceil \frac{\gamma}{1-\gamma} \right\rceil \right),
\]

where \(\hat{\delta}(\cdot)\) denotes the Factorial function.

The proof of Theorem 2 involves technical details beyond the scope of a succinct discussion here (see the online appendix for details). The above result establishes that the inflation in regret as a result of asymptotically vanishing \(\alpha(t)\) is at most poly-logarithmic in the horizon for “slowly decaying” \(\alpha(t)\), until a sharp phase transition to linear regret occurs around the \(\alpha(t) = \Theta\left(t^{-1}\right)\) “critical rate” (see Theorem 1.1).

**Remarks.** (i) **Logarithmic regret.** Theorem 2 implies that \(\mathbb{E} R_n^\pi = O(\log n)\) when \(\delta = \Delta\) and the reservoir has no “leakage” (\(\gamma = 0\)), consistent with known results for the static version of the problem [4]. (ii) **Improving sample-efficiency.** Instead of discarding both arms after step 6 of Algorithm 1, one can, in principle, discard only one arm, and query only one new arm from the reservoir as replacement. The regret incurred by this modified policy will differ only in absolute constants. The given design only intends to unify Algorithm 1 structurally with the other algorithm discussed in §III so as to facilitate an easy comparison between the performance guarantees of the two algorithms.

**Theorem 3 (Upper bound for \textit{ALG}2):** Under Model 1, the expected cumulative regret of the policy \(\pi\) given by Algorithm 2 satisfies

\[
\limsup_{n \to \infty} \frac{\alpha(n)\mathbb{E} R_n^\pi}{\log n} \leq \frac{8}{\Delta \beta_{\Delta}},
\]

where \(\beta_{\Delta}\) is as defined in (2).

Proof is provided in the online appendix.

**Discussion.** It follows directly from the above result that \(\alpha(n) = \omega(\log n/n)\) is sufficient for \textit{ALG}2 to be first-order optimal. On the other hand, we have already identified \(\alpha(n) = \omega(1/n)\) as a necessary condition for the existence of a first-order optimal policy (Theorem 1.1). Thus, the characterization of \(t^{-1}\) as a critical rate in Theorem 1.1 is sharp up to a logarithmic scaling term.

**Remark.** The scaling factor \(\beta_{\Delta}\) in Theorem 3 can, in fact, be shaved off entirely by introducing in \textit{ALG}2 an initial “burn-in” phase (sub-linear and coercive in the horizon) during each epoch à la \textit{ALG}1(n, \delta). This will, however, be achieved at the expense of \textit{ALG}2’s \textit{anytime} property. Horizon-independence can then be restored by the use of a standard exponential doubling trick, see, e.g., [28]. The resulting algorithm is not discussed in this paper for brevity.

**Comparison of theoretical performance.** To facilitate a direct comparison between the two upper bounds, it is conducive to consider \(\alpha(n) = cn^{-\gamma}\) with \(\gamma < 1\) in Theorem 3. Evidently, \textit{ALG}2 pays a heavy price for adaptivity to \(\Delta\) which reflects in a polynomial \(O\left(n^\gamma\right)\) regret as compared to the poly-logarithmic \(O\left((\log n)^{\frac{1}{1-\gamma}}\right)\) regret achievable under \textit{ALG}1(n, \delta). As to whether a performance gap between \(\Delta\)-aware and \(\Delta\)-adaptive algorithms is fundamental in Model 1 remains an open problem at the moment.

**B. Endogenous reservoirs**

The focus here will be on settings specified by Model 2. We will begin with an upper bound on the theoretical performance of Algorithm 1. As before, it will be conducive to pivot to a parametric family of mappings \(g(\cdot)\).

**Theorem 4 (Upper bound for \textit{ALG}1(n, \delta)):** Under Model 2 with \(g(u) = c(u + 1)^{-\gamma}\) for \(u \geq 0\), where \(\gamma \in [0, 1)\) is a fixed parameter, the expected cumulative regret of the policy \(\pi\) given by Algorithm 1 satisfies

\[
\limsup_{n \to \infty} \frac{\mathbb{E} R_n^\pi}{\log n} \leq \left( \frac{48\Delta}{\delta^2} \right) \left( \frac{1}{c} \right)^{\frac{1}{1-\gamma}} \hat{\delta} \left( \left\lceil \frac{\gamma}{1-\gamma} \right\rceil \right),
\]

where \(\hat{\delta}(\cdot)\) denotes the Factorial function.

Proof is provided in the online appendix. Evidently, unlike Theorem 2, only the multiplicative factors of the logarithmic term in Theorem 4 blow up as \(\gamma\) approaches 1. Thus, under Model 2, a sharper phase transition from logarithmic to linear regret occurs at the critical reservoir-depletion rate of \(\Theta(u^{-1})\), where \(u\) is the cumulative query count. We next look at the performance guarantee of \textit{ALG}2 under Model 2.

**Theorem 5 (Upper bound for \textit{ALG}2):** Under Model 2, the expected cumulative regret of the policy \(\pi\) given by Algorithm 2 satisfies

\[
\limsup_{n \to \infty} \frac{\mathbb{E} R_n^\pi}{\log n} \leq \left( \frac{16c}{\Delta} \right) \sum_{k=0}^{\infty} \exp \left( -\beta_{\Delta} \sum_{j=0}^{k-1} g(2j) \right),
\]

where \(\beta_{\Delta}\) is as defined in (2).

Proof is provided in the online appendix.

**Remark.** If the function \(g(\cdot)\) is constant (equal to \(c\)), the upper bound in Theorem 5 is bounded above by \(32/(\Delta \beta_{\Delta})\), which is independent of \(c\). The implication is that achievable regret in the problem depends on the probability \(c\) of sampling optimal-typed arms from the reservoir, surprisingly, only through sub-logarithmic terms when the reservoir is “static.” This is fundamentally distinct from the upper bound in Theorem 4 which scales inversely with \(c\) when \(\gamma = 0\).

**Comparison of theoretical performance.** Akin to Theorem 3, the \(\beta_{\Delta}\) factor in Theorem 5 may also be shaved off, albeit at the expense of \textit{ALG}2’s anytime property. Then, for \(g(u) = c(u + 1)^{-\gamma}\), the upper bound in Theorem 5 can be shown to be bounded above by \((c/\Delta)(4/c)^{1/(1-\gamma)} \hat{\delta} \left( (\gamma/(1-\gamma)) \right).\)
This order matches (up to numerical factors) the upper bound in Theorem 4. However, as $\gamma$ approaches 0, the upper bound in Theorem 4 approaches a scaling of $c^{-1}$ while that in Theorem 5 is independent of $c$ in the limit. This observation suggests that there is merit to using ALG2 when the reservoir is static or nearly-static, i.e., when it suffers a negligible leakage (or loss) of optimal arms over time (or with increasing number of queries).

In a nutshell, the upper bounds in Theorem 4 and 5, combined with the lower bound in Theorem 1.2, underscore the criticality of queries.

This work attempts to develop a principled approach to understanding the impact of endogenous variations in countably many-armed bandit problems through stylized arm-reservoir models. While this paper provides a sharp characterization of critical reservoir-decay rates for achieving sub-linear regret in two natural arm-reservoir models, an important outstanding challenge is to identify more “reasonable” models of endogeneity, and the design of “robust” algorithms for such settings. To our best knowledge, the modeling and methodological approaches adopted in this paper are novel; we hope these may guide future research on some of the interesting directions emanating out of this work.

V. CONCLUDING REMARKS

REFERENCES

[1] T. L. Lai and H. Robbins, “Asymptotically efficient adaptive allocation rules,” *Advances in applied mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[2] S. Bubeck, N. Cesa-Bianchi et al., “Regret analysis of stochastic and nonstochastic multi-armed bandit problems,” *Foundations and Trends® in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012.

[3] D. A. Berry, R. W. Chen, A. Zame, D. C. Heath, L. A. Shepp et al., “Bandit problems with infinitely many arms,” *The Annals of Statistics*, vol. 25, no. 5, pp. 2103–2116, 1997.

[4] A. Kalvit and A. Zeevi, “From finite to countable-armed bandits,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 8259–8269, 2020.

[5] R. Johari, V. Kamble, and Y. Kanoria, “Matching while learning,” *Operations Research*, vol. 69, no. 2, pp. 655–681, 2021.

[6] S. Agrawal and N. Goyal, “Analysis of thompson sampling for the multi-armed bandit problem,” in *Conference on learning theory*, 2012, pp. 39–1.

[7] P. Auer, N. Cesa-Bianchi, and P. Fischer, “Finite-time analysis of the multiarmed bandit problem,” *Machine learning*, vol. 47, no. 2-3, pp. 235–256, 2002.

[8] R. de Heide, J. Cheshire, P. Ménard, and A. Carpentier, “Bandits with many optimal arms,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 22 457–22 469, 2021.

[9] J. Gittins and D. Jones, “A dynamic allocation index for the sequential design of experiments,” *Progress in statistics*, pp. 241–266, 1974.

[10] P. Whittle, “Restless bandits: Activity allocation in a changing world,” *Journal of applied probability*, pp. 287–298, 1988.

[11] R. R. Weber and G. Weiss, “On an index policy for restless bandits,” *Journal of applied probability*, pp. 637–648, 1990.