Adapting to a Changing Environment: the Brownian Restless Bandits*

Aleksandrs Slivkins[†] Eli Upfal[‡]

First version: 2007
Conference version: COLT 2008
This version: March 2023

Abstract

In the multi-armed bandit (MAB) problem there are k distributions associated with the rewards of playing each of k strategies (slot machine arms). The reward distributions are initially unknown to the player. The player iteratively plays one strategy per round, observes the associated reward, and decides on the strategy for the next iteration. The goal is to maximize the reward by balancing *exploitation*: the use of acquired information, with *exploration*: learning new information.

We introduce and study a *dynamic* MAB problem in which the reward functions stochastically and gradually change in time. Specifically, the expected reward of each arm follows a Brownian motion, a discrete random walk, or similar processes. In this setting a player has to continuously keep exploring in order to adapt to the changing environment. Our formulation is (roughly) a special case of the notoriously intractable *restless MAB problem*.

Our goal here is to characterize the cost of learning and adapting to the changing environment, in terms of the stochastic rate of the change. We consider an infinite time horizon, and strive to minimize the average cost per step which we define with respect to a hypothetical algorithm that at every step plays the arm with the maximum expected reward at this step. A related line of work on the *adversarial MAB problem* used a significantly weaker benchmark, the best *time-invariant* policy.

The dynamic MAB problem models a variety of practical online, game-against- nature type optimization settings. While building on prior work, algorithms and steady-state analysis for the dynamic setting require a novel approach based on different stochastic tools.

---

*This version is a bug-fix for the COLT 2008 paper, resolving an issue in the proof of Theorem 2.7. Specifically, we add the two cases in Lemma 2.11(b,c) (the proof of this Lemma remains unchanged), and we use them to prove Eq. (2.11). In the previous version, Eq. (2.11) was implied but did not follow from what was written. The authors are grateful to Yukuan Jia for pointing out the issue. There were no other changes compared to the version from 2008, other than reformatting and very minor edits.

[†]Microsoft Research. E-mail: `slivkins at microsoft.com`. Parts of this work has been completed while A. Slivkins was a postdoc at Brown University.

[‡]Computer Science Department, Brown University, Providence RI. E-mail: `eli at cs.brown.edu`. Supported in part by NSF awards CCR-0121154 and DMI-0600384, and ONR Award N000140610607.

1 Introduction

The multi-armed bandit (MAB) problem (Robbins, 1952; Berry and Fristedt, 1985; Gittins, 1989) has been studied extensively for over 50 years in Operations Research, Economics and Computer Science literature, modeling online decisions under uncertainty in a setting in which an agent simultaneously attempts to acquire new knowledge and to optimize its decisions based on the existing knowledge. In the basic MAB setting, which we term the *static MAB problem*, there are k time-invariant probability distributions associated with the rewards of playing each of the k strategies (slot machine arms). The distributions are initially unknown to the player. The player iteratively plays one strategy per round, observes the associated reward, and decides on the strategy for the next iteration. The goal of a MAB algorithm is to optimize the total reward[1] by balancing *exploitation*: the use of acquired information, with *exploration*: learning new information. For several algorithms in the literature (*e.g.,* Berry and Fristedt, 1985; Auer et al., 2002a) as the number of rounds goes to infinity the expected total reward asymptotically approaches that of playing a strategy with the highest expected reward. The quality of an algorithm for the static MAB problem is therefore measured by the expected cost, or *regret*, incurred during an initial finite time interval. The regret in the first t steps is defined as the expected gap between the total reward collected by the algorithm and that collected by playing an optimal strategy in these t steps.

The MAB problem models a variety of practical online optimization problems. As an example consider a packet routing network where a router learns about delays on routes by measuring the time to receive an acknowledgment for a packet sent on that route (Awerbuch and Kleinberg, 2008; Heidari et al., 2007). The delay for one packet on a given route is a random value drawn from some distribution. The router must try various routes in order to learn about the delays. Trying a loaded route adds unnecessary delay to the routing of one packet, while discovering a route with low delay can improve the routing of the future packets.

Another application is in marketing and advertising. A store would like to display and advertise the products that sell best, but it needs to display and advertise various products to learn how good they sell. Similarly, a web search engine tries to optimize its revenue by displaying advertisements that would bring the largest number of clicks for a given web content. The company needs to experiment with various combinations of advertisements and page contents in order to find the best matches. The cost of these experiments is the loss of advertisement clicks when trying unsuccessful matches (Pandey et al., 2007).

The above examples demonstrate the practical applications of the "explore and exploit" paradigm captured in the MAB model. These examples also point out the limitation of the static approach to the problem. The delay on a route is gradually changing over time, and the router needs to continuously adapt its routing strategy to the changes in route delays. Taste and fashion change over time. A store cannot completely rely on information collected in the previous season to optimize for the next one. Similarly, a web search engine continually updates their content matching strategies to account for the changing customers' response.

A number of models have been proposed for capturing the dynamic aspect of the MAB problem. Motivated by task scheduling, Gittins and Jones (1974) considered the case where only the state of the active arm (the arm currently being played) can change in a given step, giving an optimal

[1] In this paper the *total reward* is simply the sum of the rewards, following the line of work in (Lai and Robbins, 1985; Auer et al., 2002a,b) and many other papers. Alternatively, many papers consider the *time-discounted* sum of rewards (Berry and Fristedt, 1985; Gittins, 1989; Sundaram, 2005, *e.g.,*) and references therein.

policy for the Baysean formulation with time discounting. This seminal result gave rise to a rich line of work, see Gittins et al. (2011) for an overview. In particular, Whittle (1988) introduced an extension termed *restless bandits* (Whittle, 1988; Bertsimas and Nino-Mora, 2000; Nino-Mora, 2001), where the states of all arms can change in each step according to a known (but arbitrary) stochastic transition function. Restless bandits are notoriously intractable: *e.g.,* even with deterministic transitions the problem of computing an (approximately) optimal strategy is PSPACE-hard (Papadimitriou and Tsitsiklis, 1994). Guha and Munagala (2007); Guha et al. (2009) have recently made a progress on some tractable special cases of the restless MAB problem.[2] However, their motivations, the actual problems they considered, and the techniques they used, are very different from ours.

Auer et al. (2002b) adopted an adversarial approach: they defined the *adversarial MAB problem* where the reward distributions are allowed to change arbitrarily in time, and the goal is to approach the performance of the best *time-invariant* policy. This formulation has been further studied in Auer (2002); Kleinberg and Leighton (2003); Kleinberg (2004); McMahan and Blum (2004); Flaxman et al. (2005); Dani and Hayes (2006b); Kleinberg (2006); Dani and Hayes (2006a). Auer et al. (2002b); Auer (2002) also considered a more general definition of regret, where the comparison is to the best policy that can change arms a limited number of times. Due to the overwhelming strength of the adversary, the guarantees obtained in this line of work are relatively weak when applied to the setting that we consider in this paper.

We propose and study here a somewhat different approach to addressing the dynamic nature of the MAB problem. We note that in a variety of practical applications the time evolution of the system, in particular of the reward functions, is *gradual*. Obvious examples are price, supply and demand in economics, load and delay in networks, etc. A gradual stochastic evolution is traditionally modeled via a random walk or a Brownian motion; for instance, in Mathematical Finance the (geometric) Brownian motion (Wiener process) is the standard model for continuous-time evolution of a stock price. In line with this approach, we describe the *state* of each arm – its expected reward at time t – via a Brownian motion.[3] The actual reward at a given time is an independent random sample from the reward distribution parameterized by the current state of this arm, *e.g.,* a 0-1 random variable with an expectation given by the state of the arm (in the web advertising setting this corresponds to a user clicking or not clicking on an ad).

We are interested in systems that exhibit a stationary, steady-state behavior. For this reason instead of the usual Brownian motion on a real line (which diverges to infinity) we consider a Brownian motion on an interval with reflecting bounds. Following the bulk of the stochastic MAB literature, we assume that the evolution of each arm is independent (in fact, we conjecture that regret is maximized in the case of independently evolving arms).

Our goal here is to characterize the long-term average cost of adapting to such changing environment in terms of the stochastic rate of change – the *volatility* of Brownian motion. The paradigmatic setting for us is one in which each arm's state has the same stationary distribution and, therefore, all arms are essentially equivalent in the long term. In such setting the standard benchmark – the *best* time-invariant policy – is uninformative. Instead, we optimize with respect to a more demanding (and also more natural) benchmark – a policy that at each step plays an arm with the currently maximal expected reward.

[2] These papers were published after the initial technical report version of this paper appeared.

[3] As we only sample arms at integer time points, we can equivalently describe the state as a sum of t i.i.d. normal increments. In fact, we allow the increments to come from a somewhat more general class of distributions.

We consider two versions of the *dynamic MAB problem* described above. In the *state-informed* version an algorithm not only receives a reward of the chosen arm but also finds out the current state of this arm. This is the setting in the restless MAB problem as defined in Whittle Whittle (1988) and the follow-up literature. In the second, *state-oblivious*, version an algorithm receives its reward and no other information. This formulation generalizes the static MAB problem to stochastically changing expected rewards.

1.1 The Dynamic MAB problem

Let $\{\mathcal{D}(\mu) : \mu \in [0;1]\}$ be a fixed family of probability distributions on $[0;1]$ such that $\mathcal{D}(\mu)$ has expectation μ. Time proceeds in rounds. Each arm i at each round t has a *state* $\mu_i(t) \in [0;1]$ such that the reward from playing arm i in round t is an independent random sample from $\mathcal{D}(\mu_i(t))$. At each round t an algorithm chooses one of the k alternative strategies ("arms") and receives a reward. In the *state-oblivious* version, the reward is the only information that the algorithm receives in a given round. In the *state-informed* version, the algorithm also finds out the current state of the arm that it has chosen. The distributions $\mathcal{D}(\cdot)$ are not revealed to the algorithm (and are not essential to the analysis).

The state $\mu_i(\cdot)$ varies in an interval with reflecting boundaries. To clarify the concept of reflecting boundaries, consider an object that starts moving on an interval $I = [0;1]$, reversing direction every time it hits a boundary. If the object starts at 0 and traverses distance $x \geq 0$, its position is

$$f_I(x) = \begin{cases} x', & x' \leq 1 \\ 1 - (x' - 1), & x' > 1, \end{cases} \tag{1.1}$$

where $x' = x \pmod 2 = x - 2\lfloor x/2 \rfloor$. Similarly, we define $f_I(x)$, $x < 0$ as the position of an object that starts moving from 1 and traverses distance $|x|$.

For concreteness we focus here on the case when each arm's state follows a Brownian motion. Similar results hold for related stochastic processes such as discrete random walks (see the Extensions Section).

The state of each arm i undergoes an independent Brownian motion on an interval with reflecting boundaries. Specifically, we define $\mu_i(t) = f_I(B_i(t))$ where $I = [0;1]$ is the *fundamental interval* and B_i is an independent Brownian motion with volatility σ_i. Since we only sample $\mu_i(\cdot)$ at integer times, we can also define it as a Markov chain:

$$\mu_i(t) = f_I\,(\,\mu_i(t-1) + X_i(t)\,)\,, \tag{1.2}$$

where each $X_i(t)$ is an i.i.d. sample from $\mathcal{N}(0, \sigma_i)$. The stochastic rate of change is thus given by σ_i, which we term the *volatility* of arm i.

We assume that for each arm i the initial state $\mu_i(0)$ is an independent uniformly random sample from I. This is a reasonable assumption given our goal to study the stationary behavior of the system. Indeed, the uniform distribution on I is the stationary distribution of the Markov chain 1.2 to which this Markov chain eventually converges.[4]

[4]The convergence follows from the ergodic theorem. It should be noted that the *rate* of convergence for Markov chains with infinite state spaces is a rather delicate matter, *e.g.,* see Rosenthal **?**. In this paper the rate of convergence is non-essential. Moreover, the convergence itself does not appear in the proofs: it is used only as intuition and an (additional) justification for assuming the uniform distribution of the initial state.

In the dynamic MAB problem, we measure the performance of a MAB algorithm with respect to a policy that at every step chooses a strategy with the highest expected reward. This policy changes in time, and thus it is a more demanding benchmark than the *time-invariant regret* that is often used in the MAB literature.

Definition 1.1. Consider an instance of the dynamic MAB problem. For a given MAB algorithm $\mathcal{A}$, let $W_{\mathcal{A}}(t)$ be the reward received by algorithm $\mathcal{A}$ in round t. Let $\emptyset$ be an algorithm that in every round chooses a strategy with the highest expected reward. The *dynamic regret* in round t is

$$R_{\mathcal{A}}(t) = W_{\emptyset}(t) - W_{\mathcal{A}}(t).$$

Define the *steady-state regret* as

$$\bar{R}_{\mathcal{A}} = \limsup_{t} \sup_{t_0} E\left[\frac{1}{t}\sum_{s=t_0+1}^{t_0+t} R_{\mathcal{A}}(s)\right]. \tag{1.3}$$

Thus, for any fixed $R > \bar{R}_{\mathcal{A}}$ the expected average dynamic regret of algorithm $\mathcal{A}$ over any sufficiently large interval is at most R, and it is the best possible upper bound of this form. Our goal is to bound $\bar{R}_{\mathcal{A}}$ in terms of the arms' volatility.

We use the following notation throughout the paper. The state of arm i at time t is $\mu_i(t)$. The maximal state at time t is $\mu^*(t) = \max_{i\in[k]} \mu_i(t)$. An arm i is *maximal* in round t if $\mu_i(t) = \mu^*(t)$.

1.2 Results: the state-informed case

We present an algorithm whose steady-state regret is optimal up to a poly-log factor.

Theorem 1.2. *Consider the state-informed dynamic MAB problem with k arms, each with volatility at most σ. Assume that $k < \sigma^{-\gamma}$ for some $\gamma < \frac{1}{2}$. Then there exists a MAB algorithm whose steady-state regret is at most $\tilde{O}(k\sigma^2)$.*

The algorithm is very intuitive. An arm with the highest last-observed state is called a *leader* and is played often, *e.g.,* at least every other round. Suppose the last time some other arm i was observed was t rounds ago. By Azuma inequality the state of this arm changed by at most $\Delta\mu = \tilde{O}(\sigma\sqrt{t})$ since then, with high probability. If $\mu_i(t) + \Delta\mu$ is smaller than the state of the leader, then there is no point yet in trying arm i again. Else, we mark this arm *suspicious* and enqueue it to be played soon.

The main technical contribution here is the analysis, which is quite delicate since we need to deal with the complicated dependencies in the algorithm's behavior induced by the stochastically changing environment. Essentially, we manage to reduce the stochastic aspect of the problem to simple events in the state space. We achieve it as follows. Every time each arm is played, we spread the corresponding dynamic regret evenly over the corresponding idle time. This way we express the cumulative dynamic regret as a sum over the contributions of each arm in each round. We prove a uniform bound on the expectation of each such contribution. To this end, we identify a useful high-probability behavior of the system, derive deterministic guarantees conditional on this behavior (which is the tricky part), and then argue in terms of the corresponding conditional expectations.

Surprisingly, the steady-state regret of our algorithm essentially matches a lower bound based on a very simple idea: if in a given round the states of the best two arms are within $\frac{\sigma}{4}$ from one another, then in the next round with constant probability either one of them can be $\frac{\sigma}{4}$ above another, so any algorithm incurs expected dynamic regret $\Omega(\sigma)$.[5]

Theorem 1.3. *Consider the state-informed dynamic MAB problem with k arms of volatility σ. Then the steady-state regret of any MAB algorithm is at least $\Omega(k\sigma^2)$.*

1.3 Results: the state-oblivious case

Our algorithm for the state-oblivious case builds on an algorithm from Auer et al. (2002a) for the static MAB problem. That algorithm implicitly uses a simple "padding" function that for a given arm bounds the drift of an average reward from its (static) expected value. We design a new algorithm UCB_f which relies on a novel "padding" function f that accounts for the changing expected rewards. The analysis is quite technical: the specific results from Auer et al. (2002a) do not directly apply to our setting; instead, we need to "open up the hood" and combine the technique from Auer et al. (2002a) with some new ideas.

Theorem 1.4. *Consider the state-oblivious dynamic MAB problem with k arms such that each arm i has volatility at most σ_i. Then there exists a MAB algorithm whose steady-state regret is $\tilde{O}(k\,\sigma_{av})$, where $\sigma_{av}^2 = \frac{1}{k}\sum_{i=1}^{k}\sigma_i^2$.*

Note that (unlike the guarantee in Theorem 1.2), the guarantee here is in terms of an average volatility rather than the maximal one.

1.4 Using off-the-shelf MAB algorithms?

We ask whether similar results can be obtained using off-the-shelf MAB algorithms. Specifically, we investigate the following idea: take an off-the-shelf algorithm, run it and restart it every fixed number of rounds.

For the state-informed version we consider the obvious "greedy" approach: probe each arm, choose the best one, play it for a fixed number m of rounds, restart. The greedy algorithm is parameterized by the *phase length* m which can be tuned depending on the number of arms and their volatility. We show that the greedy algorithm is indeed suboptimal as compared to Theorem 1.2: the dependence on volatility (which is smaller than one) is linear rather than quadratic; we provide both upper and lower bounds.

For the state-oblivious version one can leverage on the existing work for the adversarial MAB problem Auer et al. (2002b). This work assumes no restrictions on the state evolution, but provides guarantees only with respect to the best time-invariant policy, or a policy that switches arms a bounded number of times. We consider the following algorithm: run a fresh instance of algorithm EXP3 from Auer et al. (2002b) for a fixed number m of rounds, then restart. Using the off-the-shelf performance guarantees for EXP3 and fine-tuning m, one can (only) bound the steady-state regret by $\tilde{O}((k\sigma_{\text{av}})^{2/3})$, which is inferior to the result in Theorem 1.4. It is an open question whether one can obtain improved guarantees by tailoring the analysis in Auer et al. (2002b) to our setting.

[5]The former event happens with probability $\Omega(k\sigma)$, so the steady-state regret is $\Omega(k\sigma^2)$. This is the entire proof!

1.5 Extensions and open questions

We extend our results in several directions. First, we generalize the Markov-chain formulation Eq. (1.2) to allow the random increments $X_i(t)$ to come from other distributions which has a certain "light-tailed" property, such as the discrete random walk. Second, we consider the setting in which each arm has a distinct fundamental interval. Third, we relax the assumption that the upper bound(s) on volatilities are known to the algorithm.

The main question left open by this paper is to close the gap between the upper and lower bounds for the state-oblivious dynamic MAB problem. The only lower bound we have is Theorem 1.3. We conjecture that one may obtain a better bound based on the relative entropy-based technique from Auer et al. (2002b). It is also possible that the algorithmic result can be improved, possibly via a more refined mechanism for discounting information with time.

Another open question is whether one can obtain the optimal $\tilde{O}(k\sigma^2)$ steady-state regret for the state-informed version in the case when $k \geq \sigma^{-1/2}$. Note that the greedy algorithm mentioned in Section 1.4 achieves steady-state regret $\tilde{O}(k\sigma)$ which is non-trivial for any $k \leq \sigma^{-1}$.

1.6 Organization of the paper

In Sections 2 and 3 we present our main results for the state-informed and the state-oblivious versions, respectively. Section 4 discusses using off-the-shelf MAB algorithms. Section 5 covers the extensions.

2 The state-informed dynamic MAB problem

We consider the state-informed dynamic MAB problem where the volatility of each arm is at most σ. Recall that the state of arm i at time t is denoted $\mu_i(t)$.

For arm i and time t, the *last-seen time* $\tau_i(t)$ is the last time this arm has been played strictly before time t; the *last-seen state* $\nu_i(t) = \mu_i(\tau_i(t))$ is the corresponding state.

Definition 2.1. The *leader* in round t is the arm with a larger last-seen state, among the arms played in rounds $t-1$ and $t-2$; break ties in favor of the arm played in round $t-1$.

In our algorithm, the leader is our running estimate for an arm with the maximal state. We alternate rounds in which we always *exploit* – play the leader, with rounds in which we may explore other options. Since we define the leader in terms of the last two rounds only, our knowledge of its state is essentially up-to-date.

Let $\nu^*(t)$ be the last-seen state of the leader in round t. Let $c_{\text{susp}} = \Theta(\log \frac{1}{\sigma})^{1/2}$ be the factor to be defined later.

Definition 2.2. An arm i is called *suspicious* at time t if

$$\nu^*(t) - \nu_i(t) \leq c_{\text{susp}}\, \sigma \sqrt{t - \tau_i(t)}. \tag{2.1}$$

If an arm i is not suspicious at time t, then with high probability its current reward is less than $\nu^*(t)$. If no arm is suspicious then, intuitively, the best bet is to play the leader. Roughly, our algorithm behaves as follows: if the time is even it plays the current leader, and if the time is odd it plays a suspicious arm if one exists, and the leader otherwise. To complete the description of the algorithm, we need to specify what it does when there are multiple suspicious arms. In particular, we need to guarantee that after an arm becomes suspicious, it is played eventually.

Definition 2.3. An arm i is *active* at time t if it is not the leader and it has been suspicious at some time $t' > \tau_i(t)$. The *activation time* $\tau_i^{\text{act}}(t)$ is the earliest such time t'.

An arm becomes active when it becomes suspicious. It stays active until it is played. The idea is to play an active arm with the earliest activation time.

Algorithm 2.4. *For bootstrapping, each arm is played once. At any later time t do the following. If t is even, play the current leader. If t is odd play an active arm (with the earliest activation time) if one exists, else play the leader.*

We will use a slightly more refined algorithm which allows for a more efficient analysis. Essentially, we give priority to arms whose state is close to the leader's.

Definition 2.5. Arm i is *high-priority* at time t if it is active at this time and moreover $\tau_i^{\text{act}}(t) - \tau_i(t) \leq 4k$.

Algorithm 2.6. *For bootstrapping, each arm is played once. At any later time t do the following. If t is even, play the current leader. If t is odd play an active arm if one exists, else play the leader. If there are multiple active arms:*

- *if $t \equiv 1 \pmod 4$ then play an active arm with the earliest activation time; break ties arbitrarily*
- *if $t \equiv 3 \pmod 4$ then play a high-priority arm with the earliest activation time if one exists; else, play any active arm; break ties arbitrarily.*

The analysis of these two algorithms are very similar, except that Algorithm 2.4 has inefficiencies which lead to an extra k^2 factor in its regret. We focus on Algorithm 2.6.

Theorem 2.7. *Consider the state-informed dynamic MAB problem with k arms, each with volatility at most σ. Assume that $k < \sigma^{-\gamma}$ for some $\gamma < \frac{1}{2}$. Then Algorithm 2.6 achieves steady-state regret $O(k\,\sigma^2\ \log^2 1/\sigma)$.*

In the rest of this section we prove Theorem 2.7.

Let $\bar{R}_{\mathcal{A}}(t)$ be the average dynamic regret up to time t. Then, letting $T_i(t)$ be the set of times arm i was played before and including time t, we have

$$E\left[\bar{R}_{\mathcal{A}}(t)\right] = \frac{1}{t} \sum_{i\in[k]} \sum_{t'\in T_i(t)} E\left[\mu^*(t') - \mu_i(t')\right]. \tag{2.2}$$

Let us spread contributions of individual arms evenly over the corresponding idle time. Let us define

$$\begin{aligned} \boldsymbol{\Delta}\mu_i(t) &= \mu^*(t) - \mu_i(t), \\ \boldsymbol{\Delta}\tau_i(t) &= \tau_i^+(t) - \tau_i(t), \end{aligned}$$

where $\tau_i^+(t)$ is the next time arm i is played after time $\tau_i(t)$. In other words, $\tau_i(t)$ and $\tau_i^+(t)$ are the two consecutive times arm i is played such that $\tau_i(t) < t \leq \tau_i^+(t)$. We re-write Eq. (2.2) as

$$E\left[\bar{R}_{\mathcal{A}}(t)\right] = \frac{1}{t} \sum_{i\in[k]} \sum_{t'\in[t]} E\left[\frac{\boldsymbol{\Delta}\mu_i(\tau_i(t'))}{\boldsymbol{\Delta}\tau_i(t')}\right]. \tag{2.3}$$

We define the *contribution* of arm i in round t as

$$C_i(t) = \frac{\Delta\mu_i(\tau_i(t))}{\Delta\tau_i(t)}.$$

A crucial idea is that we upper-bound $\mathbb{E}[C_i(t)]$ for each round t separately. Namely, we prove that

$$\mathbb{E}[C_i(t)] < O(\sigma^2 \log^2 \tfrac{1}{\sigma}). \tag{2.4}$$

We identify the high-probability behavior of the processes $\{\mu_i(t)\}_{i\in[k]}$. Specifically, we consider the $\tilde{O}(\sqrt{t})$ bound on deviations, and an $O(1)$ bound on the number of near-optimal arms. A large portion of our analysis is deterministic conditional on such behavior.

Definition 2.8. A real-valued function f is *well-behaved* on an interval $[t_1; t_2]$ if for any t, $t + \Delta t \in [t_1; t_2]$ we have

$$|f(t+\Delta t) - f(t)| < c_{\text{well}}\, \sigma \sqrt{\Delta t}. \tag{2.5}$$

where $c_{\text{well}} = \Theta(\log \frac{1}{\sigma})^{1/2}$ will be chosen later.

Definition 2.9. An instance of the dynamic MAB problem is *well-behaved* on a time interval I if

(i) functions $\mu_1(t), \ldots, \mu_k(t)$ are well-behaved on I;

(ii) at each time $t \in I$ there are at most $c_{\text{near}} = O(1)$ arms i such that

$$\Delta\mu_i(t) < (8k + 15\sqrt{k})\, c_{\text{well}}\, \sigma. \tag{2.6}$$

Finally, a problem instance is called *well-behaved near time* t it is well-behaved on the time interval $[t - 3\sigma^{-2};\, t + \sigma^{-2}]$.

Remark 2.10. The expression in Eq. (2.6) is tailored to Eq. (2.14) in the subsequent analysis. The event in question happens with probability at least $1 - O(c_{\text{well}}\, k^2\, \sigma)^{c_{\text{near}}}$. Recall that we assume $k < \sigma^{-\gamma}$ for some $\gamma < \frac{1}{2}$. Thus, a (large enough) constant c_{near} suffices to guarantee a sufficiently low failure probability.

Choosing the parameters. The factors c_{well} and c_{near} are chosen so that for any fixed t a problem instance is well-behaved near time t with probability at least $1 - \sigma^{-3}$. In Definition 2.1, define $c_{\text{susp}} = 5\, c_{\text{well}}$.

Our conditionally deterministic guarantees (conditional on the problem instance being well-behaved) are expressed by the following lemma.

Lemma 2.11 (The Deterministic Lemma)**.** *Suppose a problem instance is well-behaved near time t. Fix arm i and let $\delta = \Delta\mu_i(t)$. Then:*

(a) If $\delta = 0$ and $C_i(t) > 0$ then

$$C_i(t) \le O(\sigma \log \tfrac{1}{\sigma}) / \sqrt{t - \tau_i(t)}, \tag{2.7}$$

and moreover for some arm $j \neq i$ we have

$$\Delta\mu_j(t) < O(\sigma \log \tfrac{1}{\sigma})\, \sqrt{t - \tau_i(t)}. \tag{2.8}$$

(b) If $\delta \in \left(0, \sigma \log \frac{1}{\sigma}\right)$ *then* $C_i(t) \leq O\left(\sigma \log \frac{1}{\sigma}\right)$.

(c) If $\delta \geq \sigma \log \frac{1}{\sigma}$ *then* $C_i(t) \leq O(\sigma^2/\delta)\, \log \frac{1}{\sigma}$.

Let us use Lemma 2.11 to derive the main result.

Proof of Theorem 2.7: It suffices to prove Eq. (2.4). Let $\mathcal{E}(t)$ denote the event that the problem instance is well-behaved near time t. By Lemma 2.11(a), letting $x = \sqrt{t - \tau_i(t)}$ and suppressing the $\log \frac{1}{\sigma}$ factors under the $\tilde{O}(\cdot)$ notation,

$$\begin{aligned} \mathbb{E}[C_i(t) \mid \boldsymbol{\Delta}\mu_i(t) = 0,\ \mathcal{E}(t)] &\leq \tilde{O}(\sigma/x)\ \Pr[\exists j \neq i :\ \boldsymbol{\Delta}\mu_j(t) < \tilde{O}(\sigma x)] \qquad (2.9)\\ &\leq \tilde{O}(\sigma^2). \qquad (2.10) \end{aligned}$$

We use Lemma 2.11(b,c) to prove that

$$\mathbb{E}\left[\, C_i(t) \mid \boldsymbol{\Delta}\mu_i(t) > 0,\ \mathcal{E}(t)\,\right] \leq O\left(\sigma^2 \log \tfrac{1}{\sigma}\right). \qquad (2.11)$$

For brevity, fix arm i and round t. Consider events

$$\begin{aligned} \mathcal{E}_+ &= \{\, \mathcal{E}(t) \text{ and } \boldsymbol{\Delta}\mu_i(t) > 0 \,\}\\ \mathcal{E}_0 &= \{\, \mathcal{E}(t) \text{ and } \sigma > \boldsymbol{\Delta}\mu_i(t) > 0 \,\}\\ \mathcal{E}_j &= \left\{\, \mathcal{E}(t) \text{ and } \sigma \cdot 2^j > \boldsymbol{\Delta}\mu_i(t) \geq \sigma \cdot 2^{j-1} \,\right\}, \quad \forall j \in [\,\bar{j}\,], \end{aligned}$$

where $\bar{j} = \lceil \log \frac{1}{\sigma} \rceil$. With this notation,

$$\Pr\left[\, \boldsymbol{\Delta}\mu_i(t) \leq \delta \mid \boldsymbol{\Delta}\mu_i(t) > 0,\ \mathcal{E}(t)\,\right] \leq \tilde{O}(\delta).$$

$$\begin{aligned} \mathbb{E}\left[\, C_i(t) \mid \mathcal{E}_0\,\right] \cdot \Pr\left[\, \mathcal{E}_0 \mid \mathcal{E}_+\,\right] &\leq (\sigma \log \tfrac{1}{\sigma}) \cdot O(\sigma) && \textit{(by Lemma 2.11(a))}\\ &\leq O(\sigma^2 \log \tfrac{1}{\sigma}).\\ \mathbb{E}\left[\, C_i(t) \mid \mathcal{E}_j\,\right] \cdot \Pr\left[\, \mathcal{E}_j \mid \mathcal{E}_+\,\right] &\leq O\left(\frac{\sigma^2 \log \frac{1}{\sigma}}{\sigma\, 2^j} \cdot \sigma\, 2^{j+1} \right) && \textit{(by Lemma 2.11(b))}\\ &\leq O(\sigma^2 \log \tfrac{1}{\sigma}), \quad \forall j \in [\,\bar{j}\,].\\ \mathbb{E}\left[\, C_i(t) \mid \mathcal{E}_+\,\right] &= \sum_{j=0}^{\bar{j}} \mathbb{E}\left[\, C_i(t) \mid \mathcal{E}_j\,\right] \cdot \Pr\left[\, \mathcal{E}_j \mid \mathcal{E}_+\,\right]\\ &\leq O(\sigma^2 \log^2 \tfrac{1}{\sigma}). \end{aligned}$$

This completes the proof of Eq. (2.11), and accordingly of Eq. (2.4). □

In the rest of this section we prove Lemma 2.11.

2.1 Deterministic bounds for the leader

We will argue deterministically assuming that the problem instance is well-behaved. We split our argument into a chain of claims and lemmas. The proofs are quite detailed; one can skip them for the first reading. For shorthand, let $\mathcal{E}[t_1; t_2]$ denote the event that the (fixed) problem instance is well-behaved on the time interval $[t_1; t_2]$.

First, we argue that the leader's last-seen state, $\nu^*(\cdot)$, does not decrease too much in one round.

Claim 2.12. *If $\mathcal{E}[t-2;\, t]$ then $\nu^*(t+1) \geq \nu^*(t) - 2\, c_{well}\, \sigma$.*

Proof. Assume that t is even (if t is odd the proof proceeds similarly). Recall that the leader in round t is some arm i played in one of the previous two rounds. It follows that

$$\nu^*(t) = \nu_i(t) \leq \mu_i(t) + 2\, c_{\text{well}}\, \sigma.$$

Moreover, the leader (*i.e.,* arm i) is played in round t and therefore $\nu^*(t+1) \geq \mu_i(t)$. □

Second, each arm becomes active eventually.

Claim 2.13. *Any arm i becomes active at most σ^{-2} rounds after it is played: $\tau_i^{act}(t) - \tau_i(t) \leq \sigma^{-2}$ for any time t.*

Proof. If $t - \tau_i(t) \geq \sigma^{-2}$ then Eq. (2.1) is trivially true. □

Third, we show that a currently maximal arm has been activated within $4k$ rounds from its last-seen time, and therefore it has been played in the previous $8k$ rounds. The proof of this lemma is one of the crucial arguments in our analysis.

Lemma 2.14. *Suppose $\mathcal{E}[t-\sigma^{-2};\, t]$ and arm i is maximal at time t. Then*

$$\tau_i^{act}(t) - \tau_i(t) \leq t - \tau_i^{act}(t) \leq 4k.$$

Proof. Note that $t - \tau_i^{\text{act}}(t) \leq 4k$, since otherwise after becoming active at time $\tau_i^{\text{act}}(t)$ arm i would have been played strictly before round t, contradiction.

Let $\tau = \tau_i(t)$. For the sake of contradiction assume that

$$\tau_i^{\text{act}}(t) - \tau > t - \tau_i^{\text{act}}(t). \tag{2.12}$$

Since arm i is not suspicious at time $t' = \tau_i^{\text{act}}(t) - 1$, by Definition 2.2 we have

$$\nu^*(t') - \nu_i(t') \geq c_{\text{susp}}\, \sigma \sqrt{t' - \tau}. \tag{2.13}$$

By Claim 2.13 the problem instance is well-behaved on $[\tau; t]$. It follows that

$$\nu_i(t') = \mu_i(\tau) \geq \mu_i(t) - c_{\text{well}}\, \sigma\, \sqrt{t-\tau}$$
$$\nu^*(t') = \mu_j(t'') \leq \mu_j(t) + c_{\text{well}}\, \sigma\, \sqrt{t - t''},$$

where arm j is the leader in round t', and t'' is one of the two rounds preceding t'. Plugging this into Eq. (2.13) and using Eq. (2.12), we see that $\mu_j(t) > \mu_i(t)$, contradiction. □

Fourth, we show that the leader's last-seen state is not much worse than the maximal state.

Claim 2.15. *If $\mathcal{E}[t-\sigma^{-2};\, t]$ then*

$$\mu^*(t) - \nu^*(t) \leq (8k + \sqrt{8k})\, c_{well}\, \sigma.$$

Proof. Let $\mu^*(t) = \mu_i(t)$ for some arm i, and let $\tau = \tau_i(t)$ be the last time this arm was played. By Lemma 2.14 we have $t - \tau \leq 8k$. Therefore

$$\nu^*(\tau+1) \geq \mu_i(\tau) \geq \mu_i(t) - c_{\text{well}}\, \sigma \sqrt{8k},$$

and the claim follows by Claim 2.12. □

Fifth, we show that high-priority arms are played very soon after they become active.

Claim 2.16. *Suppose arm i is a high-priority active arm at time t. Assume $\mathcal{E}[t-\sigma^{-2};\, t]$. Then $t-\tau_i^{act}(t) \leq 4\, c_{near}$.*

Proof. Fix time t and let $t' = \tau_i^{\text{act}}(t)$ be the activation time of arm i. Then by Definition 2.4 and Definition 2.3

$$\nu^*(t') - \nu_i(t') \leq c_{\text{susp}}\, \sigma\, \sqrt{t-t'} \leq c_{\text{susp}}\, \sigma\, \sqrt{4k}.$$

Using Claim 2.15 to relate $\nu^*(t')$ and $\mu^*(t')$, and using the fact that $\nu_i(t') = \mu_i(\tau)$ and that $\mu_i(\cdot)$ is well-behaved, we obtain

$$\boldsymbol{\Delta}\mu_i(t') \leq (8k + 15\sqrt{k})\, c_{\text{well}}\, \sigma. \tag{2.14}$$

Lemma follows by Definition 2.9(ii) which is, in fact, tailored to Eq. (2.14). □

Now we have the tools needed to prove a stronger version of Claim 2.15: $\mu^*(t) - \nu^*(t) \leq \tilde{O}(\sigma)$.

Lemma 2.17. *If the problem instance is well-behaved on $[t-\sigma^{-2};\, t]$ then $\mu^*(t) - \nu^*(t) \leq O(c_{well}\, \sigma)$.*

Proof. Let i be an active arm at time t. By Lemma 2.14 $\tau_i^{\text{act}}(t) - \tau_i(t) \leq 4k$, so at time $\tau_i^{\text{act}}(t)$ arm i is a high-priority active arm. By Claim 2.16 $t - \tau_i^{\text{act}}(t) \leq 4\, c_{\text{near}} = O(1)$. By Lemma 2.14 it follows that $t - \tau_i(t) \leq O(1)$.

Now $\nu^*(\tau+1) \geq \mu_i(\tau)$ by definition of the leader; $\nu^*(t) \geq \nu^*(\tau+1) - O(c_{\text{well}}\, \sigma)$ by Claim 2.12; and also $\mu^*(t) \leq \mu^*(\tau) + O(c_{\text{well}}\, \sigma)$ since the problem instance is well-behaved. Putting it together, we obtain the lemma. □

2.2 Proof of *The Deterministic Lemma*

Let $\tau = \tau_i(t)$ and recall that we denote $\delta = \boldsymbol{\Delta}\mu_i(t)$.

By Lemma 2.14 we have $t - \tau \leq 8k$. Since the problem instance is well-behaved on $[t-8k; t]$, it follows that $\mu^*(\cdot)$ is well-behaved, too, and therefore

$$|\boldsymbol{\Delta}\mu_i(t) - \boldsymbol{\Delta}\mu_i(\tau)| \leq 2\, c_{\text{well}}\, \sigma\, \sqrt{t-\tau}, \tag{2.15}$$

which immediately implies Eq. (2.7). To obtain Eq. (2.8) note that Eq. (2.15) in fact applies to any arm j, in particular to an arm j that is maximal at time τ.

Parts (b,c) of Lemma 2.11 follow from these two inequalities:

$$\boldsymbol{\Delta}\tau_i(t) \geq \Omega(\delta/\sigma)^2/\, \log\tfrac{1}{\sigma}, \tag{2.16}$$

$$\boldsymbol{\Delta}\mu_i(\tau) \leq O(\delta + \sigma \log\tfrac{1}{\sigma}). \tag{2.17}$$

(For part (b), we note that $\boldsymbol{\Delta}\tau_i(t) \geq 1$.)

Proof of Eq. (2.16): We consider two cases.

First, if we have $\boldsymbol{\Delta}\mu_i(\tau) < \delta/2$ then by Eq. (2.15) we obtain

$$2\, c_{\text{well}}\, \sigma\, \sqrt{t-\tau} \geq |\boldsymbol{\Delta}\mu_i(t) - \boldsymbol{\Delta}\mu_i(\tau)| \geq \delta/2,$$

and Eq. (2.16) follows since $\boldsymbol{\Delta}\tau_i(t) \geq t - \tau$.

Second, assume $\boldsymbol{\Delta}\mu_i(\tau) \geq \delta/2$. Then by Lemma 2.17 for any time $t' \in (\tau; t+\sigma^{-2})$ we have

$$\begin{aligned}\nu^*(t') - \mu_i(\tau) &\geq \mu^*(t') - O(c_{\text{well}}\,\sigma) + \boldsymbol{\Delta}\mu_i(\tau) - \mu^*(\tau) \\ &\geq \delta/2 - c_{\text{well}}\,\sigma\,\sqrt{t'-\tau+O(1)}.\end{aligned}$$

This is at least $\geq c_{\text{susp}}\,\sigma\,\sqrt{t'-\tau}$ as long as it is the case that $t'-\tau \leq (12\,c_{\text{well}}\,\sigma/\delta)^{-2}$. So for any such t' arm i is not suspicious, proving Eq. (2.16). □

Proof of Eq. (2.17): First, note that if $\tau_i^{\text{act}}(t) - \tau_i(t) \leq 4k$ then by Definition 2.5 arm i is a high-priority active arm at time $\tau_i^{\text{act}}(t)$, so by Claim 2.16 we have $t - \tau_i^{\text{act}}(t) \leq O(1)$ and so $t - \tau_i(t) \leq O(1)$ by Lemma 2.14. It follows by Eq. (2.15) that

$$\boldsymbol{\Delta}\mu_i(\tau) \leq \boldsymbol{\Delta}\mu_i(t) + O(\sigma),$$

and we are done. In what follows we will assume that

$$\tau_i^{\text{act}}(t) - \tau_i(t) > 4k. \tag{2.18}$$

Note that for any time t' we have

$$\begin{aligned}\nu^*(t') &\leq \max(\mu^*(t'-1),\ \mu^*(t'-2)) \\ &\leq \mu^*(t') + 2\,c_{\text{well}}\,\sigma.\end{aligned}$$

Let $t' = \tau_i^{\text{act}}(t) - 1$ be the round immediately preceding the activation time. Since arm i is not suspicious at time t',

$$\begin{aligned}c_{\text{susp}}\,\sigma\sqrt{t'-\tau} &\leq \nu^*(t') - \mu_i(\tau) \\ &\leq \mu^*(t') - \mu_i(\tau) + 2\,c_{\text{well}}\,\sigma. \\ &\leq \boldsymbol{\Delta}\mu_i(t') + c_{\text{well}}\,\sigma(2+\sqrt{t'-\tau}).\end{aligned}$$

Since $c_{\text{susp}} = 5\,c_{\text{well}}$, it follows that

$$\boldsymbol{\Delta}\mu_i(t') + 2\,c_{\text{well}}\,\sigma \geq 4\,c_{\text{well}}\,\sigma\sqrt{t'-\tau}. \tag{2.19}$$

Combining Eq. (2.15) and Eq. (2.19), we obtain

$$\begin{aligned}\boldsymbol{\Delta}\mu_i(\tau) &\leq \boldsymbol{\Delta}\mu_i(t') + 2\,c_{\text{well}}\,\sigma\,\sqrt{t'-\tau} \\ &\leq \tfrac{3}{2}\,\boldsymbol{\Delta}\mu_i(t') + 2\,c_{\text{well}}\,\sigma.\end{aligned}$$

Finally, by Eq. (2.15), Eq. (2.18) and Eq. (2.19) we obtain

$$\begin{aligned}\boldsymbol{\Delta}\mu_i(t') &\leq \boldsymbol{\Delta}\mu_i(t) + 2\,c_{\text{well}}\,\sigma\,\sqrt{t-t'} \\ &\leq \boldsymbol{\Delta}\mu_i(t) + \tfrac{1}{2}\,\boldsymbol{\Delta}\mu_i(t') + 2\,c_{\text{well}}\,\sigma. \\ \boldsymbol{\Delta}\mu_i(t') &\leq 2\,\boldsymbol{\Delta}\mu_i(t) + 4\,c_{\text{well}}\,\sigma. \\ \boldsymbol{\Delta}\mu_i(\tau') &\leq 3\,\boldsymbol{\Delta}\mu_i(t) + 6\,c_{\text{well}}\,\sigma.\end{aligned}$$

□

3 The state-oblivious dynamic MAB problem

We consider the state-oblivious dynamic MAB problem with k arms where the volatility of each arm i is at most σ_i.

Definition 3.1. For each arm i, $N_i(t)$ is the number of times it has been played in the first $t-1$ rounds, and $\overline{W}_i(t)$ is the corresponding average reward. Let $\overline{W}_i(0) = 0$ if $N_i(t) = 0$. For shorthand, let $\mu_i = \mu_i(0)$ be the initial state.

Definition 3.2. Consider an instance of the state-oblivious dynamic MAB problem. A function $f_i : \mathbb{N} \times \mathbb{N} \to \mathbb{R}_+$ is a *padding* for arm i if the following two properties hold:

- $f_i(t, t_i)$ is increasing in t and decreasing in t_i,
- for any time t, letting $t_i = N_i(t)$ we have

$$\Pr\left[\, |\overline{W}_i(t) - \mu_i(0)| > f_i(t,\, t_i) \,\right] < O(t^{-4}). \tag{3.1}$$

The family $\{f_i\}_{i\in[k]}$ is a *padding* for the problem instance.

We build on an algorithm UCB1 from Auer et al. (2002a) for the static MAB problem. We define a generalization of UCB1, which we call UCB_f, which is parameterized by a padding $f = \{f_i\}_{i\in[k]}$.

Algorithm 3.3 (UCB_f)**.** *In each round t play any arm*

$$i \in \operatorname*{argmax}_{i\in[k]} \left[\, \overline{W}_i(t) + f_i(t,\, N_i(t)) \,\right].$$

The original UCB1 algorithm is defined for a specific padding f, and in fact does not explicitly uses the notion of a padding. We introduce this notion here in order to extend the ideas from Auer et al. (2002a) to our setting.

We incorporate the analysis from Auer et al. (2002a) via the following lemma which, essentially, bounds the number of times a suboptimal arm is played by the algorithm.

Lemma 3.4 (Auer et al. Auer et al. (2002a))**.** *Consider an instance of the state-oblivious MAB problem with a padding $f = \{f_i\}_{i\in[k]}$. Consider the behavior of algorithm UCB_f in the first t rounds. Then for each arm i and any $t_i < t$ we have*

$$f_i(t, t_i) \le \tfrac{1}{2}\, \boldsymbol{\Delta}\mu_i(t) \;\Rightarrow\; \mathbb{E}[N_i(t)] \le t_i + O(1). \tag{3.2}$$

This lemma is implicit in Auer et al. Auer et al. (2002a), where it is the crux of the main proof. That proof considers the static MAB problem and (implicitly) a specific padding f.

We will use UCB_f where $f = \{f_i\}_{i\in[k]}$ is defined by

$$f_i(t, t_i) = \sqrt{2\ln(t)/t_i} + \sigma_i\,\sqrt{8t\log t}. \tag{3.3}$$

Define the *average dynamic regret* of an algorithm $\mathcal{A}$ $\bar{R}_{\mathcal{A}}(t) = \frac{1}{t}\sum_{s\in[t]} R_{\mathcal{A}}(s)$. We prove the following guarantee for algorithm UCB_f:

Theorem 3.5. *Consider the state-oblivious dynamic MAB problem with k arms. Suppose the volatility of each arm i is at most σ_i. Then there exists time t_0 such that*

$$\mathbb{E}[\bar{R}_{\text{UCB}_f}(t_0)] \leq O(k\,\sigma_{av})\,\log^{3/2}(\sigma_{av}^{-1}), \tag{3.4}$$

where $\sigma_{av}^2 = \frac{1}{k}\sum_{i=1}^k \sigma_i^2$.

To obtain Theorem 1.4 from Theorem 3.5, we start a fresh instance of algorithm UCB_f after every t_0 steps. We take advantage of the facts that (i) the "restarting times" are deterministic and, in particular, independent of the past history, and (ii) in any fixed round each $\mu_i(t)$ is distributed independently and uniformly on $[0;1]$.

In the rest of this section we prove Theorem 3.5. We start with a very useful fact about the state evolution $\mu_i(t)$. In general, if $\mu_i(0) > \frac{1}{2}$ then due to the influence of the upper boundary the expected state $\mathbb{E}[\mu_i(\cdot)]$ drifts down from its initial value. The following claim upper-bounds such drift.

Let us use a shorthand for the second summand in Eq. (3.3):

$$\delta_i(t) = \sigma_i\sqrt{8t\log t}.$$

Claim 3.6. *Fix arm i and integer times $t \leq t_*$. Then*

$$\Pr[\,|\mu_i(t) - \mu_i\,| > \delta_i] < t_*^{-3} \tag{3.5}$$

where $\mu_i = \mu_i(0)$ and $\delta_i = \delta_i(t_*)$, and therefore

$$\mathbb{E}[\mu_i(t) \mid \mu_i] \geq \min(\mu_i,\, 1-\delta_i) - t_*^{-2}. \tag{3.6}$$

Proof. Recall that the state $\mu_i(t)$ is defined as $f_I(B_i(t))$ where B_i is a Brownian motion with volatility σ_i, and f_I is the "projection" Eq. (1.1) into the interval $I = [0;1]$ with reflective boundaries. Note that $\mu_i = \mathcal{B}_i(0)$.

It follows that $|\mu_i(t) - \mu_i| > \delta_i$ only if $|\mathcal{B}_i(t) - \mu_i| > \delta_i$. We know that for any $c > 1$ we have

$$\Pr[\,|B_i(t) - \mu_i| > c\,\sigma_i\,\sqrt{t}] < 2\,e^{-c^2/2}.$$

We obtain Eq. (3.5) setting $c = \sqrt{6\log t_*}$.

Now let us prove Eq. (3.6). Define

$$f(\mu) = \mathbb{E}[\,\mu_i(t)\,|\,\mu_i].$$

Note that if $\mu < \frac{1}{2}$ then $f(\mu) > \mu$. Also, note that $f(\mu)$ is increasing and $f(\frac{1}{2}) = \frac{1}{2}$ by symmetry. Therefore, it suffices to prove Eq. (3.6) under the assumption that $\frac{1}{2} < \mu_i \leq 1 - \delta_i$.

Consider $T = \min(t, T_{\text{B}})$, where

$$T_{\text{B}} = \min\{s \in \mathbb{N} : B_i(s) \notin (0;1)\}.$$

Then $Z_s = \mu_i(\min(s,T))$ is a martingale such that $Z_0 = \mu$ and T is a bounded stopping time. By the Optional Stopping Theorem it follows that $\mathbb{E}[Z_T] = \mu$. By Eq. (3.5) we have $T_{\text{B}} \geq t$ with probability at least $1 - t_*^{-2}$, in which case $T = t$ and $\mu = Z_T = \mu_i(t)$. Thus Eq. (3.6) follows. □

Using Claim 3.6, let us argue that Eq. (3.3) is indeed a padding. Essentially, the first summand in Eq. (3.3) is tuned for an application of Chernoff-Hoeffding Bounds, whereas the second one corrects for the drift.

Lemma 3.7. *The family f defined by Eq. (3.3) is a padding.*

Proof. We need to prove Eq. (3.1). Fix arm i and time t. Let $\{t_j\}_{j=1}^{\infty}$ be the enumeration of all times when arm i is played. Let $X_j = \mu_i(t_j)$ be the state of arm i in round t. Let $\hat{X}_j$ be the actual reward collected by the algorithm from arm i in round t_j. Let us define the sums $S = \sum_{j\in[n]} X_j$ and $S^* = \sum_{j\in[n]} \hat{X}_j$, where $n = N_i(t)$ is the number of times arm i is played before time t. Let $\mu = \mu_i(0)$ and $\delta = \delta_i(t)$.

We can rewrite Eq. (3.1) as follows:

$$\Pr[|S^* - \mu\, n| > \sqrt{2n\ln t} + \delta n] < O(t^{-4}). \tag{3.7}$$

Let F be the *failure event* when $|\mu_i(s) - \mu| > \delta$ for some $s \in [t]$. Recall that by Claim 3.6 the probability of F is at most t^{-4}. In the probability space induced by conditioning on $\hat{X}_1, \dots, \hat{X}_{j-1}$ and the event $\bar{F}$, we have

$$\begin{aligned} \mathbb{E}[\hat{X}_j] &= E\left[\mathbb{E}[\hat{X}_j\,|t_j, X_j]\right] = E\left[\mathbb{E}[X_j\,|t_j]\right] \\ &= \mathbb{E}[X_j] \in [\mu-\delta, \mu+\delta]. \end{aligned}$$

Going back to the original probability space,

$$\mathbb{E}[\hat{X}_j|\,\hat{X}_1 \dots \hat{X}_{j-1}, \bar{F}] \in [\mu-\delta, \mu+\delta]. \tag{3.8}$$

The Chernoff-Hoeffding bounds (applied to the probability space induced by conditioning on $\bar{F}$) say precisely that the condition Eq. (3.8) implies the following tail inequality:

$$\Pr\left[\,|\hat{S} - \mu\, m| > \delta m + a \mid \bar{F}\,\right] \le 2\, e^{-2a^2/m}$$

for any $a \le 0$. We obtain Eq. (3.7) by taking $a = \sqrt{2m\ln T}$. $\square$

To argue about algorithm UCB_f, we will use the following notation:

Definition 3.8. We will use the following notation:

$$\begin{cases} \rho_i(t) &= \min(\mu_i,\, 1-\delta_i(t)), & \mu_i = \mu_i(0), \\ \Delta_i &= \mu^* - \mu_i, & \mu^* = \mu^*(0) \\ S(t) &= \{\text{arms } i : \Delta_i \ge 4\delta_i(t)\}. \end{cases}$$

Lemma 3.9. *Consider any algorithm for the state-oblivious dynamic MAB problem. Then for each arm i and time $t \ge k$*

$$E\left[N_i(t)\,\overline{W}_i(t) \mid \mu_i\right] \ge \rho_i(t)\;\mathbb{E}[N_i(t)] - t^{-2}. \tag{3.9}$$

The left-hand side of Eq. (3.9) is the total winnings collected by arm i up to time t. If the bandit algorithm always plays arm i, then $N_i(t) = t$ and the left-hand side of Eq. (3.9) is simply equal to $\sum_s \mathbb{E}[\mu_i(s)]$, so the lemma follows from Claim 3.6. In this sense, Lemma 3.9 is an extension of Claim 3.6. The proof of Eq. (3.9) is a rather intricate exercise in conditional expectations and martingales. We defer it to Section 3.1.

We combine Lemma 3.9 and Lemma 3.4 to derive a conditional bound on $\bar{R}_{\text{UCB}_f}(t)$:

Corollary 3.10. *For any time t we have*

$$E\left[\bar{R}_{\mathrm{UCB}_f}(t) \mid \mu_1,\, \ldots,\, \mu_k\right] \le \frac{k}{t^2} + O(1)\left[\sum_{i\notin S(t)} \mu^* - \rho_i(t)\right] + O(\tfrac{1}{t}\,\log t)\left[\sum_{i\in S(t)} \frac{1}{\Delta_i}\right]. \tag{3.10}$$

Proof. Fix time t and let $\overline{W}_i = \overline{W}_i(t)$, $\rho_i = \rho_i(t)$ and $N_i = N_i(t)$. Let $R(t)$ be the left-hand side of Eq. (3.10). Using Eq. (3.9),

$$\begin{aligned} t\,R(t) &= \sum_{i\in[k]} \mathbb{E}[(\mu^* - \overline{W}_i)\, N_i] \\ &\le \sum_{i\in[k]} \mathbb{E}[N_i]\,(\mu^* - \rho_i) + t^{-2}. \end{aligned}$$

For each $i \in S(t)$ we have $\mu^* - \rho_i \le 2\Delta_i$ and by Lemma 3.4

$$\mathbb{E}[N_i(t)] \le 32\, \ln(m)/\Delta_i^2 + O(1). \quad \square$$

We obtain Theorem 3.5 by integrating both sides of Eq. (3.10) with respect to $\mu_1 \ldots \mu_k$.

Proof of Theorem 3.5: Fix time t and let $\delta_i = \delta_i(t)$ and $\rho_i = \rho_i(t)$. Note that Eq. (3.10) is, essentially, the sum over all arms. We partition the arms into three sets and bound the three corresponding sums separately.

Note that the following holds for any fixed $\gamma > 0$: given μ^* and the event $\{\Delta_i > \gamma\}$, the random variable μ_i is distributed uniformly on the interval $[0; \mu^* - \gamma)$. We will use this property in the forthcoming integrations.

First, we consider the set $S = S(t)$. Conditional on μ^*,

$$\begin{aligned} E\left[\sum_{i\in S} \Delta_i^{-1}\right] &= \sum_{i\in[k]} E\left[\Delta_i^{-1} \mid \Delta_i > 4\delta_i\right] \Pr[\Delta_i > 4\delta_i] \\ &\le \sum_{i\in[k]} \ln \sigma_i^{-1} \le O(k \ln \sigma_{\mathrm{av}}^{-1}). \end{aligned} \tag{3.11}$$

Second, let us consider the set S^+ of all arms i such that $0 < \Delta_i < 4\delta_i$. Conditional on μ^*, we obtain

$$\begin{aligned} E\left[\sum_{i\in S^+} \mu^* - \rho_i\right] &\le \sum_{i\in[k]} O(\delta_i)\; \Pr[\Delta_i < 4\delta_i \mid \Delta_i > 0] \\ &\le \textstyle\sum_{i\in[k]} O(\delta_i) \min(1, \delta_i/\mu^*). \end{aligned}$$

Integrating over μ^*, we obtain

$$\begin{aligned} E\left[\textstyle\sum_{i\in S^+} \mu^* - \rho_i\right] &\le \textstyle\sum_{i\in[k]} O(\delta_i^2) \\ &\le O(k\, \sigma_{\mathrm{av}}^2\, t \log t). \end{aligned} \tag{3.12}$$

Third, we consider the set S^* of all maximal arms, *i.e.,* the set of all arms i such that $\Delta_i = 0$. We show the main steps of the argument, omitting the details of some straightforward integrations:

$$Z_i := \mathbb{I}_{\{\Delta_i=0\}}\,(\mu^* - \rho_i)$$
$$\mathbb{E}[Z_i] = \mathbb{E}[\mathbb{E}[Z_i|\mu^*]] = \tfrac{1}{k}\;\mathbb{E}[\mu^* - \rho_i] = O(\delta_i^2)$$

$$E\left[\sum_{i\in S^*} \mu^* - \rho_i\right] = \sum_{i\in[k]} \mathbb{E}[Z_i] \leq O(k\,\sigma_{\text{av}}^2)(t\log t). \tag{3.13}$$

Finally, using(3.11-3.13), we take expectations in Eq. (3.10):

$$E\left[\bar{R}_{\text{UCB}_f}(t)\right] = O(\tfrac{k}{t}\;\log t)((\sigma_{\text{av}}\,t)^2 + \log\sigma_{\text{av}}^{-1}).$$

The theorem follows if we take $t_0 = \sigma_{\text{av}}\sqrt{\log\sigma_{\text{av}}^{-1}}$. □

3.1 Proof of Lemma 3.9: conditional expectations

Fix arm i and time t. Let us introduce a more concise notation which gets rid of the subscript i. Let $\mu = \mu_i(0)$ and $\delta = \delta_i(t)$, and denote $N = N_i(t)$. For every time s, let $Y_s = \mu_i(s)$, and let X_s be the winnings from arm i at time s if it is played by the algorithm.[6] Let ζ_s be equal to 1 if arm i is played at time s, and 0 otherwise.

To prove Eq. (3.9), we will show that

$$E\Big[\textstyle\sum_{s\in[t]}\zeta_s\,X_s\Big] = E\Big[\textstyle\sum_{s\in[t]}\zeta_s\,Y_s\Big] \tag{3.14}$$
$$\geq \min(\mu, 1-\delta)\;\mathbb{E}[N] + t^{-2}. \tag{3.15}$$

Note that ζ_s and X_s are conditionally independent given Y_s. It follows that

$$\begin{aligned} E\left[\zeta_s\,X_s\,|Y_s\right] &= E\left[\zeta_s\,|Y_s\right]\;E\left[X_s\,|Y_s\right] = E\left[\zeta_s\,|Y_s\right]\;Y_s \\ &= E\left[\zeta_s\,Y_s\,|Y_s\right]. \end{aligned}$$

Taking expectations on both sides, we obtain

$$E[\zeta_s\,X_s] = E[\zeta_s\,Y_s],$$

which proves Eq. (3.14). Going from Eq. (3.14) to Eq. (3.15) is somewhat more complicated. In what follows we denote $S = \sum_{t\in[m]}\zeta_s\,Y_s$.

Claim 3.11. *If $\mu \leq 1-\delta$ then $\mathbb{E}[S] \geq \mu\;\mathbb{E}[N] - t^{-2}$.*

Proof. As in Claim 3.6, we recall the definition $\mu_i(s) = f_I(B_i(s))$ where B_i is a Brownian motion with volatility σ_i, and f_I is the "projection" Eq. (1.1) into the interval $I = [0;1]$ with reflective boundaries. Note that $\mu_i = \mathcal{B}_i(0)$.

For brevity, denote $\widehat{Y}_s = B_i(s)$, and define the corresponding shorthand $\widehat{S} = \sum_{s\in[t]}\zeta_s\,\widehat{Y}_s$. Let F be the *failure event* when $\widehat{Y}_s \geq 1$ for some $t \leq m$. Note that if this event does not occur, then

[6] That is, X_s is an independent random sample from distribution $\mathcal{D}(Y_s)$, as defined in Section 1.1.

$Y_s \geq \widehat{Y}_s$ for every time $t \in [m]$ and therefore $S \geq \widehat{S}$. We use this observation to express $\mathbb{E}[S]$ in terms of $\mathbb{E}[\widehat{S}]$. Let $p := \Pr[F]$ and note that it is at most m^{-4}. Then:

$$\begin{aligned}
\mathbb{E}[\widehat{S}] &= (1-p)\ \mathbb{E}[\widehat{S} \mid \text{not } F] + p\ \mathbb{E}[\widehat{S}\,|F] \\
&\leq (1-p)\ \mathbb{E}[\widehat{S} \mid \text{not } F] + p(\mu + t\sigma_i) \\
\mathbb{E}[S] &\geq (1-p)\ \mathbb{E}[S \mid \text{not } F] + p\ \mathbb{E}[S\,|F] \\
&\geq (1-p)\ \mathbb{E}[\widehat{S} \mid \text{not } F] \\
&\geq \mathbb{E}[\widehat{S}] - pt\sigma_i - p.
\end{aligned}$$

To prove the claim, it remains to bound $\mathbb{E}[\widehat{S}]$.

Let $\{s_j\}_{j=1}^{\infty}$ be the enumeration of all times when arm i is played. Note that $N = \max\{j : s_j \leq t\}$. Define $\widehat{Z}_j = \widehat{Y}_{s_j}$ for each j. We would like to argue that $\{\widehat{Z}_j\}_{j=1}^{\infty}$ is a martingale and N is a stopping time. More precisely, claim that this is true for some common filtration. Indeed, one way to define such filtration $\{\mathcal{F}_j\}_{j=1}^{\infty}$ is to define $\mathcal{F}_j$ as the σ-algebra generated by s_{j+1} and all tuples $(s_l, Z_l,\, Z_l^*,\, \widehat{Z}_l)$ such that $l \leq j$. Now using the Optional Stopping Theorem one can show that

$$\mathbb{E}[\widehat{S}] = \textstyle\sum_{j\in[N]} Z_j = \mathbb{E}[N]\ \mathbb{E}[\widehat{Z}_0],$$

which proves the claim since $\widehat{Z}_0 = \mu$. □

To prove Eq. (3.15), it remains to consider the case $\mu > 1-\delta$.

Claim 3.12. *if* $\mu > 1-\delta$ *then*

$$\mathbb{E}[S] \geq (1-\delta)\ \mathbb{E}[N] - t^{-2}.$$

Proof. Let T be the smallest time s such that $Y_s \leq 1-\delta$. Let $\{s_j\}_{j=1}^{\infty}$ be the enumeration of all times when arm i is played, and let $J = \max j : t_j \leq T$. Conditioning on T and J, consider the entire problem starting from time $T+1$. Then by Claim 3.11 we have:

$$E\left[\sum_{s=T+1}^{m} \zeta_s\, Y_s\,|T, J\right] \geq (1-\delta)(\mathbb{E}[N] - J) - t^{-2}.$$

Let $S_T = \sum_{s=T+1}^{t} \zeta_s\, Y_s$. It follows that

$$\begin{aligned}
S &= S_T + \textstyle\sum_{t\in[T]} \zeta_s\, Y_s \geq S_T + (1-\delta)\, J \\
\mathbb{E}[S] &= \mathbb{E}[S_T] + (1-\delta)\, \mathbb{E}[J] \\
&\geq (1-\delta)\ \mathbb{E}[N-J] - t^{-2} + (1-\delta)\ \mathbb{E}[J] \\
&\geq (1-\delta)\ \mathbb{E}[N] - t^{-2}. \quad \square
\end{aligned}$$

4 Using off-the-shelf algorithms

In this section we investigate the following idea: take an off-the-shelf MAB algorithm, run it, and restart it every fixed number of rounds. We consider both the state-informed and state-oblivious versions of the dynamic MAB problem.

We use the following notation: there are k arms, each arm i has volatility σ_i, and the average volatility σ_{av} is defined by $\sigma_{\text{av}}^2 = \frac{1}{k}\sum_{i=1}^{k}\sigma_i^2$. We rely on the following lemma:

Lemma 4.1. *Let $\mu^* = \mu^*(0)$ and let $i^* \in \operatorname{argmax} \mu_i(0)$, ties broken arbitrarily. Then for any times $t \le t_*$*

$$\mathbb{E}[\mu^* - \mu_{i^*}(t)] \le O(k)(t_*^{-4} + \sigma_{av}^2\, t_* \log t_*). \tag{4.1}$$

More generally, we can consider arbitrary fixed times

$$0 \le t_1 \le t_2 \le \ldots \le t_k \le t \le t_*$$

and define $\mu^* = \max \mu_i(t_i)$ and $i^* \in \operatorname{argmax} \mu_i(t_i)$.

The lemma is obtained by combining Claim 3.6 and Eq. (3.13); we omit the details.

Remark. The intuition is that each arm i is probed in round t_i, so that $\mu_i(t_i)$ is the expected value of the corresponding probe. This lemma is similar to Claim 3.6 in that it bounds the downwards drift of $\mathbb{E}[\mu_i(\cdot)]$ which is caused by the proximity of the upper boundary. The difference is that here we specifically consider a "maximal" arm, *e.g.,* when $t_i \equiv 0$ we consider an arm which is maximal at time 0.

4.1 State-informed version: greedy algorithm

For the state-informed version we consider a very simple, "greedy" approach: probe each arm once, choose one with the largest state, play it for a fixed number $m - k$ of rounds, restart. Call this a *greedy algorithm* with phase length m.

Theorem 4.2. *Consider the state-informed dynamic MAB problem with k arms such that the volatility of each arm i is σ_i. With phase length $m = \sigma_{av} \sqrt{\log \sigma_{av}^{-1}}$, the steady-state regret of the greedy algorithm is at most*

$$O(k\, \sigma_{av} \log \sigma_{av}^{-1}), \text{ where } \sigma_{av}^2 = \tfrac{1}{k} \textstyle\sum_{i=1}^k \sigma_i^2.$$

Proof. For the algorithmic result, fix phase length $m > k$ and consider a single phase of the greedy algorithm. Assume without loss of generality that in the first k rounds of the phase our algorithm plays arm i in step i. Let $\mu_i = \mu_i(i)$ be the corresponding rewards, and let μ^* be the largest of them. Then the greedy algorithm chooses arm $i^* \in \operatorname{argmax}_{i \in [k]} \mu_i$ and plays it for $m - k$ rounds. Consider the t-th of these $m-k$ rounds and let $Y_t = \mu_{i^*}(t+k)$ be the state of arm i^* in this round. By Lemma 4.1 we have $\mathbb{E}[Y_t] \ge \mathbb{E}[\mu^*] - z$, where z is the right-hand side of Eq. (4.1). Therefore, letting $\overline{W}$ be the per-round average reward in this phase, we have

$$\begin{aligned} \mathbb{E}[\overline{W}] &\ge \tfrac{1}{m} \textstyle\sum_{t=1}^{m-k} Y_t \ge \tfrac{m-k}{m} (\mathbb{E}[\mu^*] - z) \\ \mathbb{E}[\mu^* - \overline{W}] &\le z + \tfrac{k}{m}\, \mathbb{E}[\mu^*] \\ &\le O(km\, \sigma_{\text{av}}^2 \log m) + \tfrac{k}{m}(1 + \tfrac{1}{m}) \\ &= O(k\, \sigma_{\text{av}}) \sqrt{\log\, \sigma_{\text{av}}^{-1}} \end{aligned}$$

for $m = \sigma_{\text{av}} \sqrt{\log\, \sigma_{\text{av}}^{-1}}$. □

We provide a matching lower bound.

Theorem 4.3. *Consider the setting in Theorem 4.2. Then the steady-state regret of the greedy algorithm is $\tilde{\Omega}(k\, \sigma_{av})$.*

Proof Sketch. For simplicity assume $\sigma_i \equiv \sigma$. It is known that in time t a Brownian motion with volatility σ drifts by at least $\Delta = \tilde{\Omega}(\sigma\sqrt{t})$ with high probability. Thus for each arm i with high probability $\mu_i(t) \leq 1 - \Delta/2$, regardless of the initial value $\mu_i(0)$. Now we can obtain a lower bound that corresponds to Lemma 4.1: letting $\mu^* = \max \mu_i(i)$ and $i^* \in \operatorname{argmax} \mu_i(i)$ be the arm chosen by the greedy algorithm,

$$\mathbb{E}[\mu^* - \mu_{i^*}(t)] \geq \tilde{\Omega}(k\,\sigma^2 t), \tag{4.2}$$

for any $t > k$. Now consider a given phase of the greedy algorithm. In the first k rounds the algorithm accumulates regret $\Omega(k)$, and in each subsequent round t the regret is the left-hand side of Eq. (4.2). The theorem follows easily. □

4.2 State-oblivious version via adversarial MAB

For the state-oblivious dynamic MAB problem, we use a very general result of Auer et al. (2002b) for the adversarial MAB problem. For simplicity, here we only state this result in terms of the present setting.

Let $\overline{W}_{\mathcal{A}}(t)$ be the average reward collected by algorithm $\mathcal{A}$ during the time interval $[1;t]$.

Theorem 4.4 (Auer et al. (2002b)). *Consider the state-oblivious dynamic MAB problem with k arms. Let $\mathcal{A}_i$ be an algorithm that plays arm i at every step. Then there exists an algorithm, call it* Exp3*, such that for any arm i and any time t*

$$\mathbb{E}[\overline{W}_{\text{Exp3}}(t)] \geq E(\overline{W}_{\mathcal{A}_i}(t)) - O(\tfrac{k}{t}\log t)^{1/2}.$$

For our problem, we restart EXP3 every m steps, for some fixed m; call this algorithm EXP3(m).

Theorem 4.5. *Consider the state-informed dynamic MAB problem with k arms such that the volatility of each arm i is at most σ_i. Then there exists m such that algorithm* EXP3(m) *has steady-state regret*

$$O(k\,\sigma_{av}\log\sigma_{av}^{-1})^{2/3},\ \textit{where}\ \sigma_{av}^2 = \tfrac{1}{k}\textstyle\sum_{i=1}^{k}\sigma_i^2.$$

Proof. Let use shorthand $\mathcal{A} =$ EXP3(m). Let μ^* be the maximal expected reward at time 0, and suppose it is achieved by some arm i^*. Let $\mathcal{A}^*$ be the algorithm that plays this arm at every step. Let $Y_t = \mu_{i^*}(t)$ the state of arm i^* in round t. Then by Lemma 4.1 we have $\mathbb{E}[Y_t] \geq \mathbb{E}[\mu^*] - z_m$, where $z(m)$ is the right-hand side of Eq. (4.1). Therefore:

$$\begin{aligned}
\mathbb{E}[\overline{W}_{\mathcal{A}^*}(m)] &= E\left[\,\mathbb{E}[\overline{W}_{\mathcal{A}^*}(m)\,|\,Y_1,\ \ldots\ ,Y_m]\,\right] \\
&= \tfrac{1}{m}\,E\left[\,\textstyle\sum_{t=1}^{m} Y_t\,\right] \\
&\geq \mu^* - z(m) \\
\mathbb{E}[\bar{R}_{\mathcal{A}}(m)] &= E\left[\mu^* - \overline{W}_{\mathcal{A}^*}(m)\right] \\
&\quad +\ E\left[\overline{W}_{\mathcal{A}^*}(m) - \overline{W}_{\mathcal{A}}(m)\right]
\end{aligned} \tag{4.3}$$

Now using Eq. (4.3) and Theorem 4.4 we obtain

$$\mathbb{E}[\bar{R}_{\mathcal{A}}(m)] \leq z(m) + O(\tfrac{k}{m}\log m)^{1/2}. \tag{4.4}$$

We choose m that minimizes the right-hand side of Eq. (4.4). □

We note in passing that we can also get non-trivial (but worse) guarantees for the state-oblivious dynamic MAB problem using two other off-the-shelf approaches:

- a version of the greedy algorithm which probes each arm a few times in the beginning of each phase,
- a version of Theorem 4.4 in which the benchmark algorithm is allowed to switch arms a few times (Auer et al., 2002b).

Essentially, the first approach is too primitive, while the second one makes overly pessimistic assumptions about the environment. In both cases we obtain guarantees of the form $\tilde{O}(k\sigma_{\text{av}})^{\gamma}$, $\gamma < \frac{2}{3}$, which are inferior to Theorem 4.5.

5 Extensions

Recall that the state evolution of arm i in the dynamic MAB problem is described by Eq. (1.2), where the i.i.d. increments $\mu_i(t)$ are distributed with respect to some fixed distribution $\mathcal{X}_i$. Can we relax the assumption that $\mathcal{X}_i$ is normal?

Definition 5.1. Random variable X is *stochastically* (ρ, σ)*-bounded* if its moment-generating function satisfies

$$\mathbb{E}[e^{r(X-\mathbb{E}[x])}] \leq e^{r^2\sigma^2/2} \quad \text{for } |r| \leq \rho.$$

This is precisely the condition needed to establish an Azuma-type inequality: if S is the sum of t independent stochastically (ρ, σ)-bounded random variables with zero mean, then with high probability $S \leq \tilde{O}(\sigma\sqrt{t})$. Specifically, for any $\lambda \leq \frac{1}{2}\,\rho\,\sigma\sqrt{t}$ we have

$$\Pr\left[S > \lambda\sigma\sqrt{t}\right] \leq \exp(-\lambda^2/2). \tag{5.1}$$

Note that a normal distribution $\mathcal{N}(0,\sigma)$ is (∞,σ)-bounded, and any distribution with support $[-\sigma,\sigma]$ is $(1,\sigma)$-bounded.

We can recover all of our algorithmic results if we assume that each distribution $\mathcal{X}_i$ has zero mean and is stochastically (ρ,σ_i)-bounded for some σ_i, where $\rho > 0$ is a fixed absolute constant. We re-define the *volatility* of arm i as the infimum of all σ such that $\mathcal{X}_i$ is (ρ,σ)-bounded.

It is appealing to tackle a more general setting when the only restriction on each distribution $\mathcal{X}_i$ is that it has mean 0 and variance σ_i^2. We can extend our analysis (at the cost of somewhat weaker guarantees) if we further assume that, essentially, the absolute third moment of $\mathcal{X}_i$ is comparable to σ_i^3. Then instead of Eq. (5.1) we can use a weaker inequality called the *non-uniform Berry-Esseen theorem* (Neammanee, 2005):

$$\Pr\left[\textstyle\sum_{s=1}^{t}\mu_i(s) > \sigma_i\, t^{\gamma}\right] \leq O\left((\tfrac{\rho_i}{\sigma_i})^3\; t^{1-3\gamma}\right), \tag{5.2}$$

for any $\gamma > 1/2$, where $\rho_i^3 = \mathbb{E}[\,|\mu_i(s)|^3\,]$. We omit further discussion of this extension from the present version.

Let us discuss one other direction in which our setting can be generalized. Recall that in the dynamic MAB problem the state of each arm evolves on the same interval $I = [0;1]$ (see Section 1.1) which we term the *fundamental interval*. What if we allow each arm to have a distinct fundamental interval? All our algorithms fit this extended setting with little or no modification. The performance guarantees should look like a weighted sum of contributions from different arms, where the weights depend (perhaps in rather complicated way) on the respective fundamental intervals. To illustrate this point, we worked out the guarantees for the two algorithms discussed in Section 4, see Appendix A for details. It is an open question to derive similar closed-form guarantees for the other algorithms in this paper.

Recall that in all our results we assumed that the volatilities are known to the algorithm. In fact, this assumption is not necessary: we are interested in the stationary performance of our algorithms and, as it turns out, we can afford to learn the static parameters of the model. Roughly, the argument goes as follows. It suffices for our analysis if for each arm an algorithm knows a 2-approximate upper bound on volatility σ_i, rather than the exact value. One can learn such bound by playing arm i for $O(\log^2 \sigma_i)$ rounds, with failure probability as low as $O(\sigma_i^{-10})$, and repeat this learning phase every σ_i^{-1} rounds (we omit the details).

Acknowledgments. The first author would like to thank Bobby Kleinberg for many stimulating conversations about multi-armed bandits.

References

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *J. of Machine Learning Research (JMLR)*, 3:397–422, 2002. Preliminary version in *41st IEEE FOCS*, 2000.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002a.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002b. Preliminary version in *36th IEEE FOCS*, 1995.

Baruch Awerbuch and Robert Kleinberg. Online linear optimization and adaptive routing. *J. of Computer and System Sciences*, 74(1):97–114, February 2008. Preliminary version in *36th ACM STOC*, 2004.

Donald A. Berry and Bert Fristedt. *Bandit problems: sequential allocation of experiments.* Springer, Heidelberg, Germany, 1985.

D. Bertsimas and J. Nino-Mora. Restless bandits, linear programming relaxations, and a primal-dual index heuristic. *Operations Research*, 48(1):80–90, 2000.

Varsha Dani and Thomas P. Hayes. How to beat the adaptive multi-armed bandit. Technical report. Available from arXiv at `http://arxiv.org/cs.DS/0602053`, 2006a.

Varsha Dani and Thomas P. Hayes. Robbing the bandit: less regret in online geometric optimization against an adaptive adversary. In *17th ACM-SIAM Symp. on Discrete Algorithms (SODA)*, pages 937–943, 2006b.

Abraham Flaxman, Adam Kalai, and H. Brendan McMahan. Online Convex Optimization in the Bandit Setting: Gradient Descent without a Gradient. In *16th ACM-SIAM Symp. on Discrete Algorithms (SODA)*, pages 385–394, 2005.

J. C. Gittins and D. M. Jones. A dynamic allocation index for the sequential design of experiments. In J. Gani et al., editor, *Progress in Statistics*, pages 241–266. North-Holland, 1974.

John Gittins, Kevin Glazebrook, and Richard Weber. *Multi-Armed Bandit Allocation Indices.* John Wiley & Sons, Hoboken, NJ, USA, 2nd edition, 2011.

John C. Gittins. *Multi-Armed Bandit Allocation Indices.* John Wiley & Sons, Hoboken, NJ, USA, 1989.

Sudipta Guha and Kamesh Munagala. Approximation algorithms for partial-information based stochastic control with Markovian rewards. In *48th IEEE Symp. on Foundations of Computer Science (FOCS)*, pages 483–493, 2007.

Sudipta Guha, Kamesh Munagala, and Peng Shi. Approximation algorithms for restless bandit problems. In *20th ACM-SIAM Symp. on Discrete Algorithms (SODA)*, pages 28–37, 2009.

F. Heidari, S. Mannor, and L.G. Mason. Reinforcement learning-based load shared sequential routing. In *IFIP Networking*, 2007.

Robert Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *18th Advances in Neural Information Processing Systems (NIPS)*, 2004.

Robert Kleinberg. Anytime algorithms for multi-armed bandit problems. In *17th ACM-SIAM Symp. on Discrete Algorithms (SODA)*, pages 928–936, 2006.

Robert D. Kleinberg and Frank T. Leighton. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *IEEE Symp. on Foundations of Computer Science (FOCS)*, pages 594–605, 2003.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient Adaptive Allocation Rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

H. Brendan McMahan and Avrim Blum. Online Geometric Optimization in the Bandit Setting Against an Adaptive Adversary. In *17th Conf. on Learning Theory (COLT)*, pages 109–123, 2004.

K. Neammanee. On the constant in the nonuniform version of the Berry-Esseen theorem. *Intl. J. of Mathematics and Mathematical Sciences*, 2005:12:1951–1967, 2005.

J. Nino-Mora. Restless bandits, partial conservation laws and indexability. *Advances in Applied Probability*, 33:76–98, 2001.

Sandeep Pandey, Deepak Agarwal, Deepayan Chakrabarti, and Vanja Josifovski. Bandits for Taxonomies: A Model-based Approach. In *SIAM Intl. Conf. on Data Mining (SDM)*, 2007.

Christos H. Papadimitriou and John N. Tsitsiklis. The complexity of optimal queueing network control. In *Structure in Complexity Theory Conference*, pages 318–322, 1994.

Herbert Robbins. Some Aspects of the Sequential Design of Experiments. *Bull. Amer. Math. Soc.*, 58: 527–535, 1952.

Rangarajan K. Sundaram. Generalized Bandit Problems. In David Austen-Smith and John Duggan, editors, *Social Choice and Strategic Decisions: Essays in Honor of Jeffrey S. Banks (Studies in Choice and Welfare)*, pages 131–162. Springer, 2005. First appeared as *Working Paper, Stern School of Business*, 2003.

P. Whittle. Restless bandits: Activity allocation in a changing world. *J. of Appl. Prob.*, 25A:287–298, 1988.

A Distinct fundamental intervals

Recall that in the dynamic MAB problem the state of each arm evolves on the same interval $I = [0; 1]$ (see Section 1.1) which we term the *fundamental interval.* In this section we consider a generalization in which we allow each arm to have a distinct fundamental interval. We work out the guarantees for the two algorithms discussed in Section1.4.

The main contribution of this appendix is that we find a way to upper-bound the steady-state regret of the respective algorithms in terms of reasonably defined averages of the arms' properties. The actual derivations are rather tedious but not that illuminating; we omit them from this version.

A.1 The setting and notation

We consider the following setting. There are k arms. Each arm has volatility σ_i and fundamental interval $[a_i; b_i]$. Without loss of generality we assume that $b_1 \leq \ldots b_k$ and that $\max a_i < \min b_i$. (If the latter fails then we can always ignore the arm with the smallest upper boundary b_i.) To simplify the derivation we assume that $\max \sigma_i \leq \frac{1}{3}$.

Define the *weight* of arm i as

$$w_i = \prod_{l=i}^{k} \frac{b_i - a_l}{b_l - a_l},$$

Define the *average volatility* σ_{av} and the *average length* as

$$\sigma_{\text{av}}^2 = \frac{\sum_{i\in[k]} w_i (b_i - a_i)\, \sigma_i^2}{\sum_{i\in[k]} w_i (b_i - a_i)}$$
$$d_{\text{av}} = \tfrac{1}{k} \textstyle\sum_{i\in[k]} w_i (b_i - a_i).$$

To see that the quantities we defined above are reasonable as *averages*, note that if all arms have the same fundamental interval $[a; b]$ then all weights are 1 and $d_{\text{av}} = b - a$ and, moreover, the average volatility σ_{av} coincides with the one defined in the body of the paper.

A.2 Results

We present two results that extend, respectively, Theorem 4.2 and Theorem 4.5 to the setting from Section A.1. In both cases the algorithms are exactly the same. The main tool is a version of Lemma 4.1, where the guarantee Eq. (4.1) looks exactly the same in our notation, except the right-hand side is multiplied by d_{av}.

Theorem A.1. *Consider the deterministic dynamic MAB problem in the setting from Section A.1. Let $a_{min} = \min a_i$. Then for phase length*

$$m = \sigma_{av}^{-1} \sqrt{(b_k - a_{min}) / \log \sigma_{av}^{-1}}$$

the greedy algorithm has steady-state regret

$$O(k\, \sigma_{av}) \sqrt{(b_k - a_{min})\; d_{av}\; \log \sigma_{av}^{-1}}.$$

Theorem A.2. *Consider the state-informed dynamic MAB problem in the setting from Section A.1. Then there exists m such that algorithm* EXP3(m) *has steady-state regret*

$$O(d_{av})^{1/3} (k\, \sigma_{av}\, \log \sigma_{av}^{-1})^{2/3}.$$