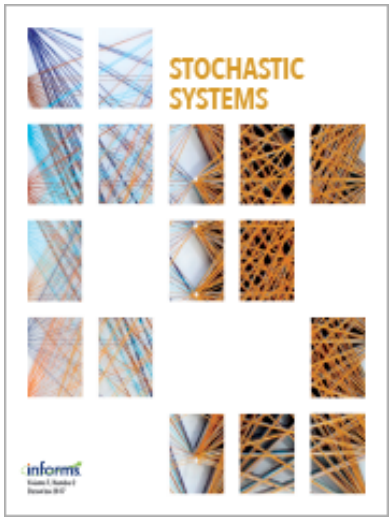



Stochastic Systems

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

# Optimal Exploration—Exploitation in a Multi-armed Bandit Problem with Non-stationary Rewards

Omar Besbes, Yonatan Gur, Assaf Zeevi

To cite this article:
Omar Besbes, Yonatan Gur, Assaf Zeevi (2019) Optimal Exploration—Exploitation in a Multi-armed Bandit Problem with Non-stationary Rewards. Stochastic Systems 9(4):319-337. https://doi.org/10.1287/stsy.2019.0033

http://pubsonline.informs.org/journal/stsy

STOCHASTIC SYSTEMS
Vol. 9, No. 4, December 2019, pp. 319–337
ISSN 1946-5238 (online)

# Optimal Exploration–Exploitation in a Multi-armed Bandit Problem with Non-stationary Rewards

**Omar Besbes,[a] Yonatan Gur,[b] Assaf Zeevi[a]**

[a] Graduate School of Business, Columbia University, New York, New York 10027; [b] Graduate School of Business, Stanford University, Stanford, California 94305

**Contact:** ob2105@gsb.columbia.edu (OB); ygur@stanford.edu, http://orcid.org/0000-0003-0764-3570 (YG); assaf@gsb.columbia.edu (AZ)

**Received:** April 26, 2018
**Revised:** November 21, 2018
**Accepted:** March 11, 2019
**Published Online in Articles in Advance:** October 31, 2019

**MSC2010 Subject Classification:** Primary: 93E35

**https://doi.org/10.1287/stsy.2019.0033**



**Abstract.** In a multi-armed bandit problem, a gambler needs to choose at each round one of $K$ arms, each characterized by an unknown reward distribution. The objective is to maximize cumulative expected earnings over a planning horizon of length $T$, and performance is measured in terms of *regret* relative to a (static) oracle that *knows* the identity of the best arm a priori. This problem has been studied extensively when the reward distributions do not change over time, and uncertainty essentially amounts to identifying the optimal arm. We complement this literature by developing a flexible non-parametric model for temporal uncertainty in the rewards. The extent of temporal uncertainty is measured via the cumulative mean change in the rewards over the horizon, a metric we refer to as *temporal variation*, and regret is measured relative to a (dynamic) oracle that plays the *pointwise* optimal action at each period. Assuming that nature can choose any sequence of mean rewards such that their temporal variation does not exceed $V$ (a temporal uncertainty budget), we characterize the complexity of this problem via the *minimax regret*, which depends on $V$ (the hardness of the problem), the horizon length $T$, and the number of arms $K$.

**History:** Former designation of this paper was SSy-2018-015.R1.



**Funding:** This work is supported by the National Science Foundation [Grant 0964170] and the U.S.–Israel Binational Science Foundation [Grant 2010466].

**Keywords:** multi-armed bandit • exploration/exploitation • nonstationary • dynamic oracle • minimax regret • dynamic regret

## 1. Introduction

### 1.1. Background and Motivation

In the prototypical multi-armed bandit (MAB) problem, a gambler needs to choose at each round of play $t = 1, \ldots, T$ one of $K$ arms, each characterized by an unknown reward distribution. Reward realizations are only observed when an arm is selected, and the gambler's objective is to maximize cumulative expected earnings over the planning horizon. To achieve this goal, the gambler needs to experiment with multiple actions (pulling arms) in an attempt to identify the optimal choice while simultaneously taking advantage of the information available at each step of the game to optimize immediate rewards. This trade-off between information acquisition via exploration (which is forward looking) and the exploitation of the latter for immediate reward optimization (which is more myopic in nature) is fundamental in many problem areas; examples include clinical trials (Zelen 1969), strategic pricing (Bergemann and Valimaki 1996), investment in innovation (Bergemann and Hege 2005), online auctions (Kleinberg and Leighton 2003), assortment selection (Caro and Gallien 2007), and online advertising (Pandey et al. 2007), to name but a few. The broad applicability of this class of problems is one of the main reasons MAB problems have been so widely studied, since their inception in the seminal papers of Thompson (1933) and Robbins (1952).

In the MAB problem, it has become commonplace to measure the performance of a policy relative to an oracle that *knows* the identity of the best arm a priori, with the gap between the two referred to as the *regret*. If the regret is sublinear in $T$, this is tantamount to the policy being (asymptotically) long-run-average optimal, a first-order measure of "good" performance. In their seminal paper, Lai and Robbins (1985) went further and identified a sharp characterization of the regret growth rate: no policy can achieve (uniformly) a regret that is smaller than order $\log T$, and there exists a class of policies, predicated on the concept of upper confidence bounds (UCBs), that achieve said growth rate of regret and hence are (second-order) asymptotically optimal.

The premultiplier in the growth rate of regret encodes the "complexity" of the MAB instance in terms of problem primitives; in essence, it is proportional to the number of arms $K$ and inversely proportional to a term that measures the "distinguishability" of the arms—roughly speaking, the closer the mean rewards are, the harder it is to differentiate the arms and the larger this term is. When the aforementioned gap between the arms' mean reward can be arbitrarily small, the complexity of the MAB problem, as measured by the growth rate of regret, is of order $\sqrt{T}$ (see Auer et al. 2002a). For overviews and further references, the reader is referred to the monographs by Berry and Fristedt (1985), Gittins (1989) for Bayesian/dynamic programming formulations, and Bubeck and Cesa-Bianchi (2012), which covers the machine learning literature and the so-called adversarial setting.

The parameter uncertainty in the MAB problem described is purely *spatial*, and the difficulty in the problem essentially amounts to uncovering the identity of the optimal arm with "minimal" exploration. However, in many application domains, *temporal changes* in the reward distribution structure are an intrinsic characteristic of the problem, and several attempts have been made to incorporate this into a stochastic MAB formulation. The origin of this line of work can be traced back to Gittins and Jones (1974), who considered a case in which only the state of the chosen arm can change, giving rise to a rich line of work (see, e.g., Gittins (1979) and Whittle (1981) as well as references therein). In particular, Whittle (1988) introduced the term *restless bandits*, a model in which the states (associated with reward distributions) of arms change in each step according to an arbitrary yet known stochastic process. Considered a "hard" class of problems (cf. Papadimitriou and Tsitsiklis 1994), this line of work has led to various approximations (see, e.g., Bertsimas and Nino-Mora 2000), relaxations (see, e.g., Guha and Munagala 2007), and restrictions of the state transition mechanism (see, e.g., Ortner et al. (2014) for irreducible Markov processes and Azar et al. (2014) for a class of history-dependent rewards).

An alternative and more pessimistic approach views the MAB problem as a game between the policy designer (gambler) and nature (adversary) in which the latter can change the reward distribution of the arms at every instance of play. These ideas date back to the work of Blackwell (1956) and Hannan (1957) and have since seen significant development; Foster and Vohra (1999), Cesa-Bianchi and Lugosi (2006), and Bubeck and Cesa-Bianchi (2012) provide reviews of this line of research. Within this so-called adversarial formulation, the efficacy of a policy over a given time horizon $T$ is measured relative to a benchmark that is defined by the *single best action in hindsight*, the best action that could have been taken after seeing all reward realizations. The single best action represents a *static oracle*, and the regret in this formulation uses that as a benchmark. For obvious reasons, this static oracle can perform quite poorly relative to a *dynamic oracle* that follows the dynamic optimal sequence of actions because the latter optimizes the (expected) reward at each time instant.[1] Thus, a potential limitation of the adversarial framework is that even if a policy exhibits a "small" regret relative to the static oracle, there is no guarantee that it will perform well with respect to the more stringent dynamic oracle.

### 1.2. Main Contributions

In this paper, we provide a non-parametric formulation that is useful for modeling non-stationary rewards and allows benchmarking performance against a dynamic oracle and yet is tractable from an analytical and computational standpoint, allowing for a sharp characterization of problem complexity. Specifically, our contributions are as follows.

**1.2.1. Modeling.** We introduce a non-parametric modeling paradigm for non-stationary reward environments, which we demonstrate to be tractable for analysis purposes yet extremely flexible and broad in the scope of problem settings to which it can be applied. The key construct in this modeling framework is that of a budget of *temporal uncertainty*, which is measured in the total variation metric with respect to the cumulative mean reward changes over the horizon $T$. One can think of this as a temporal uncertainty set with a certain prescribed "radius" $V_T$ such that all mean reward sequences that reside in this set have temporal variation that is less than this value. (These concepts are related to the ones introduced in Besbes et al. (2015) in the context of non-stationary stochastic approximation.) In particular, $V_T$ plays a central role in providing a sharp characterization of the MAB problem complexity (in conjunction with the time horizon $T$ and the number of arms $K$). In this manner, the paper advances our understanding of MAB problems and complements existing approaches to modeling and analyzing non-stationary environments. In particular, the non-parametric formulation we propose allows for very general temporal evolutions, extending most of the treatment in the non-stationary stochastic MAB literature, which mainly focuses on a finite number of changes in the mean rewards (see, e.g., Garivier and Moulines 2011). Concomitantly, as indicated, the framework allows for the more stringent dynamic oracle to be used as a benchmark as opposed to the static benchmark used in the adversarial setting. (We further discuss these connections in Section 3.2.)

**1.2.2. Minimax Regret Characterization.** When the cumulative temporal variation over the decision horizon $T$ is known to be bounded ex ante by $V_T$, we characterize the order of the minimax regret and hence the complexity of the learning problem. It is worthwhile emphasizing that the regret is measured here with respect to a dynamic oracle that knows ex ante the mean reward evolution and hence the *dynamic sequence of best actions*. This is a marked departure from the weaker benchmark associated with the best single action in hindsight, which is typically used in the adversarial literature (exceptions noted herein). First, we establish lower bounds on the performance of *any* non-anticipating policy relative to the aforementioned dynamic oracle and then show that the order of this bound can be achieved uniformly over the class of temporally varying reward sequences by a suitably constructed policy. In particular, the minimax regret is shown be of the order of $(KV_T)^{1/3}T^{2/3}$. The reader will note that this result is quite distinct from the traditional stochastic MAB problem. In particular, in that problem, the regret is either of order $K \log T$ (in the case of "well separated" mean rewards) or of order $K\sqrt{T}$ (the minimax formulation). In contrast, in the non-stationary MAB problem, the minimax complexity exhibits a very different dependence on problem primitives. In particular, even if the variation budget $V_T$ is a constant independent of $T$, then asymptotically the regret grows significantly faster, order $T^{2/3}$, compared with the stationary case, and when $V_T$ itself grows with $T$, the problem exhibits complexity on a spectrum of scales. Ultimately, if the variation budget is such that $V_T$ is of the same order as the time horizon $T$, then the regret is linear (and no policy can achieve sublinear regret).

**1.2.3. Elucidating Exploration–Exploitation Trade-Offs in the Presence of Non-stationary Rewards.** Unlike the traditional stochastic MAB problem in which the key trade-off is between the information acquired through exploration of the action space and the immediate reward obtained by exploiting said information, the non-stationary MAB problem has further subtlety. In particular, although the policy we propose accounts for the explore–exploit tension, it highlights an additional consideration that concerns *memory*. More broadly, changes in the reward distribution that are inherent in the non-stationary stochastic setting require that a policy also "forgets" the acquired information at a suitable rate.

**1.2.4. Toward Adaptive Policies.** Our work provides a sharp characterization of the minimax complexity of the non-stationary MAB via a policy that knows a priori a bound $V_T$ on the mean reward temporal variation. This leaves open the question of online adaptation to said temporal variation. Namely, are there policies that are minimax or nearly minimax optimal (in order) that do not require said knowledge and hence can *adapt* to the nature of the changing mean reward sequences on the fly. This adaption means that a policy can achieve ex post performance that is as good as (or nearly as good as) the one achievable under ex ante knowledge of the temporal variation budget. Section 5 of this paper lays out the key challenges associated with adapting to unknown variation budgets. Although we are not able to provide an answer to the question, we propose a potential solution methodology in the form of an "envelope" policy that employs several subordinate policies, each of which is constructed under a different assumption on the temporal variation budget. The subordinate policies each represent a "guess" of the unknown temporal parameter, and the "master" envelope policy switches among these policies based on observed feedback, hence learning the changes in variation as they manifest. Although we have no proof for optimality or strong theoretical indication to believe an optimal adaptive policy is to be found in this family, numerical results indicate that such a conjecture is plausible. A full theoretical analysis of adaptive policies is left as a direction for future research.

### 1.3. Further Contrast with Related Work

The two closest papers to ours are Auer et al. (2002b), which is couched in the adversarial setting, and Garivier and Moulines (2011), which pursues the stochastic setting. In both papers, the non-stationary reward structure is constrained such that the identity of the best arm can change only a *finite* number of times. The regret in these instances is shown to be of order $\sqrt{T}$. Our analysis complements these results by treating a broader and more flexible class of temporal changes in the reward distributions. Our framework, which considers a dynamic oracle as benchmark, adds a further element that was so far absent from the adversarial formulation and provides a much more realistic comparator against which to benchmark a policy. The concept of variation budget was advanced in Besbes et al. (2015) in the context of a non-stationary stochastic approximation problem, complementing a large body of work in the online convex optimization literature. The analogous relationship can be seen between our paper and the work on adversarial MAB problems, although the techniques and analysis are quite distinct from Besbes et al. (2015), which deals with continuous (convex) optimization.

Hazan and Kale (2011) consider a non-stochastic MAB setting in which regret is measured relative to the more traditional single best action in hindsight. The regret relative to the latter depends on a quadratic variation/spread measure of the cost vectors (relative to the empirical average vector of costs). Slivkins (2014) considers a contextual bandit setting with a regret bound that depends, among other things, on a Lipschitz-type constant that limits reward differences over the joint space of arms and contexts. Other forms of variation are also considered in Jadbabaie et al. (2015) in an online convex optimization setting.

Some more recent papers have followed up on ideas presented in this paper in the context of several non-stationary sequential optimization settings. These include MAB settings in which additional structure is imposed on the non-stationarity of mean rewards (e.g., Levine et al. 2017) as well as other sequential optimization settings (see, e.g., Wei et al. (2016) in an expert advice context). A few recent papers focus on the question of how to design policies that adapt to unknown variation budgets; see, for example, Karnin and Anava (2016) and Luo et al. (2018), as well as Cao et al. (2019) and Cheung et al. (2019), in different MAB settings and Zhang et al. (2018) in a full-information online convex optimization setting.

### 1.4. Structure of the Paper

The next section introduces the formulation of a stochastic non-stationary MAB problem. In Section 3, we characterize the minimax regret using lower and upper bounds by exhibiting a family of adaptive policies that achieve rate-optimal performance. Section 4 provides numerical results. In Section 5, we lay out the challenges associated with adapting to an unknown variation budget. Proofs can be found in the appendix.

## 2. Problem Formulation

Let $\mathcal{K} = \{1, \ldots, K\}$ be a set of arms. Let $t = 1, 2, \ldots$ denote the sequence of decision epochs, in which at each $t$ the decision maker pulls one of the $K$ arms and obtains a reward $X_t^k \in [0, 1]$, where $X_t^k$ is a random variable with expectation $\mu_t^k = \mathbb{E}[X_t^k]$. We denote the best possible expected reward at decision epoch $t$ by

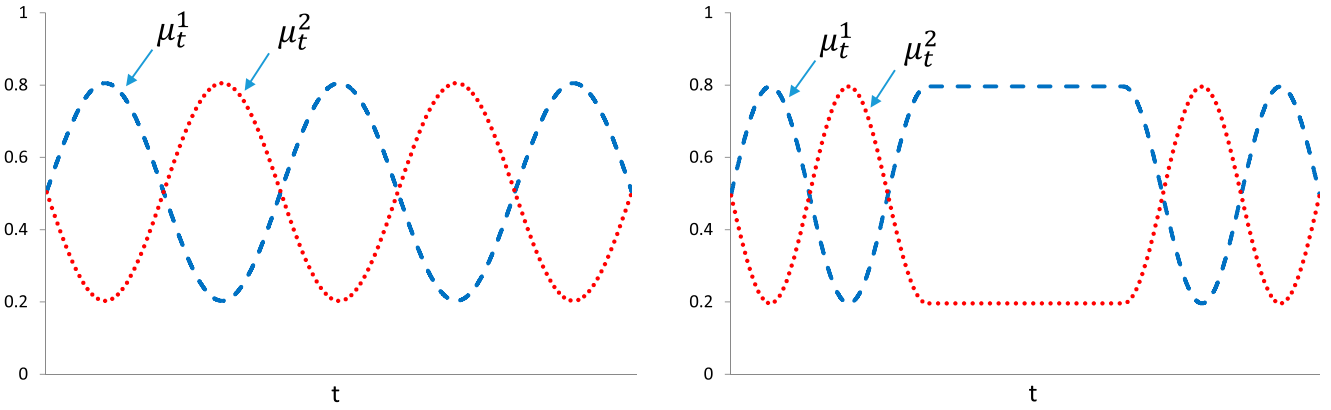
$$\mu_t^* = \max_{k \in \mathcal{K}} \{\mu_t^k\}, \quad t = 1, 2, \ldots.$$

### 2.1. Temporal Variation in the Expected Rewards

We assume that the expected reward of each arm $\mu_t^k$ may change at any decision point. We denote by $\mu^k$ the sequence of expected rewards of arm $k$: $\mu^k = \{\mu_t^k : t = 1, 2, \ldots\}$. In addition, we denote by $\mu$ the sequence of vectors of all $K$ expected rewards: $\mu = \{\mu^k : k = 1, \ldots, K\}$. We assume that the expected reward of each arm can change an arbitrary number of times and track the extent of (cumulative) *temporal variation* over a given horizon $T$ using the following metric:

$$\mathcal{V}(\mu; T) := \sum_{t=1}^{T-1} \sup_{k \in \mathcal{K}} |\mu_t^k - \mu_{t+1}^k|, \quad T = 2, 3, \ldots. \tag{1}$$

**Figure 1.** Two Instances of Temporal Changes in the Expected Rewards of Two Arms That Correspond to the Same Cumulative Variation




*Notes.* (Left) Continuous variation in which a fixed variation (which equals 3) is spread over the whole horizon. (Right) A counterpart instance in which the same variation is spent in the first and final thirds of the horizon while mean rewards are fixed in between.

As is further clarified later on, our formulation does not impose a specific structure on $\mu$, but we assume that these are independent of the realized sample path of past actions. The formulation allows for many different forms in which the mean rewards may change; for illustration, Figure 1 shows two different temporal patterns of mean reward changes that correspond to the same variation value $\mathcal{V}(\mu;T)$.

### 2.2. Admissible Policies, Performance, and Regret

Let $U$ be a random variable defined over a probability space $(\mathbb{U},\mathcal{U},\mathbf{P}_u)$. Let $\pi_1:\mathbb{U}\to\mathcal{K}$ and $\pi_t:[0,1]^{t-1}\times\mathbb{U}\to\mathcal{K}$ for $t=2,3,\ldots$ be measurable functions. With some abuse of notation, we denote by $\pi_t\in\mathcal{K}$ the action at time $t$ that is given by

$$\pi_t=\begin{cases}\pi_1(U) & t=1,\\ \pi_t\left(X_{t-1}^{\pi},\ldots,X_1^{\pi},U\right) & t=2,3,\ldots.\end{cases}$$

The mappings $\{\pi_t:\ t=1,2,\ldots\}$ together with the distribution $\mathbf{P}_u$ define the class of admissible policies. We denote this class by $\mathcal{P}$. We further denote by $\{\mathcal{H}_t,\ t=1,2,\ldots\}$ the filtration associated with a policy $\pi\in\mathcal{P}$, such that

$$\mathcal{H}_1=\sigma(U)$$

and

$$\mathcal{H}_t=\sigma\left(\left\{X_j^{\pi}\right\}_{j=1}^{t-1},U\right)$$

for all

$$t\in\{2,3,\ldots\}.$$

Note that policies in $\mathcal{P}$ are non-anticipating; that is, they depend only on the past history of actions and observations and allow for randomized strategies via their dependence on $U$.

For a given horizon $T$ and given sequence of mean reward vectors $\{\mu\}$, we define the regret under policy $\pi\in\mathcal{P}$ compared with a dynamic oracle as

$$R^{\pi}(\mu,T)=\sum_{t=1}^{T}\mu_t^*-\mathbb{E}^{\pi}\left[\sum_{t=1}^{T}\mu_t^{\pi}\right],$$

where the expectation $\mathbb{E}^{\pi}[\cdot]$ is taken with respect to the noisy rewards as well as to the policy's actions. The regret measures the difference between the expected performance of the dynamic oracle that "pulls" the arm with the highest mean reward at each epoch $t$ and that of any given policy. Note that, in a stationary setting, one recovers the typical definition of regret in which the oracle rule is constant.

### 2.3. Budget of Variation and Minimax Regret

Let $\{V_t:t=1,2,\ldots\}$ be a non-decreasing sequence of positive real numbers such that $V_1=0$, $KV_t\le t$ for all $t$, and for normalization purposes, set $V_2=2\cdot K^{-1}$. We refer to $V_T$ as the *variation budget* over time horizon $T$. For that horizon, we define the corresponding *temporal uncertainty set* as the set of mean reward vector sequences with cumulative temporal variation that is bounded by the budget $V_T$:

$$\mathcal{L}(V_T)=\left\{\mu\in[0,1]^{K\times T}:\mathcal{V}(\mu;T)\le V_T\right\}.$$

The variation budget captures the constraint imposed on the non-stationary environment faced by the decision maker. We denote by $\mathcal{R}^{\pi}(V_T,T)$ the regret guaranteed by policy $\pi$ *uniformly* over all mean rewards sequences $\{\mu\}$ residing in the temporal uncertainty set:

$$\mathcal{R}^{\pi}(V_T,T)=\sup_{\mu\in\mathcal{L}(V_T)}R^{\pi}(\mu,T).$$

In addition, we denote by $\mathcal{R}^*(V_T,T)$ the *minimax regret*, namely, the minimal worst-case regret that can be guaranteed by an admissible policy $\pi\in\mathcal{P}$:

$$\mathcal{R}^*(V_T,T)=\inf_{\pi\in\mathcal{P}}\sup_{\mu\in\mathcal{L}(V_T)}R^{\pi}(\mu,T).$$

This minimax regret formulation implies that the sequence of mean rewards is selected by a non-adaptive adversary and is only constrained to belonging to the uncertainty set $\mathscr{L}(V_T)$. Conditional on the sequence of mean rewards, our model is one of a MAB with stochastic (noisy) rewards that are generated by non-stationary distributions. A direct characterization of the minimax regret is hardly tractable. In what follows, we derive bounds on the magnitude of this quantity as a function of the horizon $T$ that elucidate the impact of the budget of temporal variation $V_T$ on achievable performance; note that $V_T$ may increase with the length of the horizon $T$.

## 3. Analysis of the Minimax Regret

### 3.1. Lower Bound

We first provide a lower bound on the best achievable performance.

**Theorem 1** (Lower Bound on Achievable Performance). *Assume that, at each time $t = 1, \ldots, T$, the rewards $X_t^k$, $k = 1, \ldots, K$, follow a Bernoulli distribution with mean $\mu_t^k$. Then, for any $T \geq 2$ and $V_T \in [K^{-1}, K^{-1}T]$, the worst-case regret for any policy $\pi \in \mathscr{P}$ is bounded below as follows:*

$$\mathscr{R}^{\pi}(V_T, T) \geq CK^{1/3}V_T^{1/3}T^{2/3},$$

*where $C$ is an absolute constant (independent of $T$ and $V_T$).*

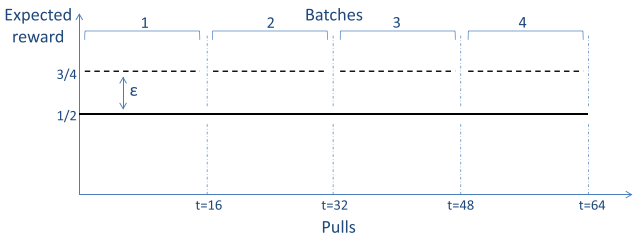
We note that when reward distributions are stationary and arm mean rewards can be arbitrarily close, there are known policies that achieve regret of order $\sqrt{T}$ up to logarithmic factors (cf. Auer et al. 2002a). Moreover, it has been established in Auer et al. (2002b) that this regret rate is still achievable even when there are changes in the mean rewards as long as the number of changes is finite and independent of the horizon length $T$. Note that such sequences belong to $\mathscr{L}(V_T)$ for the special case in which $V_T$ is a constant independent of $T$. Theorem 1 states that when the notion of temporal variation is broader, as per earlier, then it is no longer possible to achieve the aforementioned $\sqrt{T}$ performance. In particular, any policy must incur a regret of at least order $T^{2/3}$. At the extreme, when $V_T$ grows linearly with $T$, then it is no longer possible to achieve sublinear regret (and, hence, long-run-average optimality). This theorem provides a full spectrum of bounds on achievable performance that range from $T^{2/3}$ to linear regret.

#### 3.1.1. Key Ideas in the Proof of Theorem 1.

We define a subset of vector sequences $\mathscr{L}' \subset \mathscr{L}(V_T)$ and show that when $\mu$ is drawn randomly from $\mathscr{L}'$, any admissible policy must incur regret of order $V_T^{1/3}T^{2/3}$. We define a partition of the decision horizon $\mathscr{T} = \{1, \ldots, T\}$ into batches $\mathscr{T}_1, \ldots, \mathscr{T}_m$ of size $\Delta$ each (except possibly the last batch). In $\mathscr{L}'$, every batch contains exactly one good arm with expected reward $1/2 + \varepsilon$ for some $0 < \varepsilon \leq 1/4$, and all the other arms have expected reward $1/2$. The good arm is drawn independently in the beginning of each batch according to a discrete uniform distribution over $\mathscr{K}$. Thus the identity of the good arm can change only between batches. See Figure 2 for an example of possible realizations of a sequence $\mu$ that is randomly drawn from $\mathscr{L}'$. By selecting $\varepsilon$ such that $\varepsilon T/\Delta \leq V_T$, any $\mu \in \mathscr{L}'$ is composed of expected reward sequences with a variation of at most $V_T$, and therefore, $\mathscr{L}' \subset \mathscr{L}(V_T)$. Given the draws under which expected reward sequences are generated, nature prevents any accumulation of information from one batch to another because, at the beginning of each batch, a new good arm is drawn independently of the history.

Using ideas from the analysis of Auer et al. (2002b, proof of theorem 5.1), we establish that no admissible policy can identify the good arm with high probability within a batch. Because there are $\Delta$ epochs in each batch, the regret that any policy must incur along a batch is of order $\Delta \cdot \varepsilon \approx \sqrt{\Delta}$, which yields a regret of order

**Figure 2.** Drawing a Sequence from $\mathscr{L}'$




*Notes.* A numerical example of possible realizations of expected rewards. Here $T = 64$, and we have four decision batches, each containing 16 pulls. We have $K^4$ equally probable realizations of reward sequences. In every batch, one arm is randomly and independently drawn to have an expected reward of $1/2 + \varepsilon$, and in this example, $\varepsilon = 1/4$. This example corresponds to a variation budget of $V_T = \varepsilon\Delta = 1$.

$\sqrt{\Delta} \cdot T/\Delta \approx T/\sqrt{\Delta}$ throughout the whole horizon. Theorem 1 then follows from selecting $\Delta \approx (T/V_T)^{2/3}$, the smallest feasible $\Delta$ that satisfies the variation budget constraint, yielding regret of order $V_T^{1/3}T^{2/3}$.

### 3.2. Rate-Optimal Policy

We now show that the order of the bound in Theorem 1 can be achieved. To that end, we introduce the following policy, referred to as Rexp3.

**Rexp3.** Inputs: a positive number $\gamma$ and a batch size $\Delta$.

1. Set batch index $j = 1$.
2. Repeat while $j \le \lceil T/\Delta \rceil$:
   a. Set $\tau = (j-1)\Delta$.
   b. Initialization: For any $k \in \mathcal{K}$, set $w_t^k = 1$.
   c. Repeat for $t = \tau+1, \ldots, \min\{T, \tau+\Delta\}$:
   - For each $k \in \mathcal{K}$, set

$$p_t^k = (1-\gamma)\frac{w_t^k}{\sum_{k'=1}^K w_t^{k'}} + \frac{\gamma}{K}.$$

   - Draw an arm $k'$ from $\mathcal{K}$ according to the distribution $\{p_t^k\}_{k=1}^K$ and receive a reward $X_t^{k'}$.
   - For $k'$, set $\hat{X}_t^{k'} = X_t^{k'}/p_t^{k'}$, and for any $k \ne k'$, set $\hat{X}_t^k = 0$. For all $k \in \mathcal{K}$, update:

$$w_{t+1}^k = w_t^k \exp\left\{\frac{\gamma \hat{X}_t^k}{K}\right\}.$$

   d. Set $j = j+1$ and return to the beginning of step 2.

Clearly, $\pi \in \mathcal{P}$. The Rexp3 policy uses Exp3, a policy introduced by Freund and Schapire (1997) for solving a worst-case sequential allocation problem as a subroutine, restarting it every $\Delta$ epochs. At each epoch, with probability $\gamma$, the policy explores by sampling an arm from a discrete uniform distribution over the set $\mathcal{K}$. With probability $(1-\gamma)$, the policy exploits information gathered thus far by drawing an arm according to weights that are updated exponentially based on observed rewards. Therefore, $\pi$ balances exploration and exploitation by a proper selection of an *exploration rate* $\gamma$ and a batch size $\Delta$. At a high level, the sampling probability $\{p_t^k\}$ represents the current "certainty" of the policy regarding the identity of the best arm.

**Theorem 2** (Rate Optimality). *Let $\pi$ be the Rexp3 policy with a batch size $\Delta = \lceil (K\log K)^{1/3}(T/V_T)^{2/3} \rceil$ and with*

$$\gamma = \min\left\{1\ ,\ \sqrt{\frac{K\log K}{(e-1)\Delta}}\right\}.$$

*Then, for every $T \ge 2$ and $V_T \in [K^{-1}, K^{-1}T]$, the worst-case regret for this policy is bounded from above as follows:*

$$\mathcal{R}^\pi(V_T, T) \le \bar{C}(K\log K)^{1/3}V_T^{1/3}T^{2/3},$$

*where $\bar{C}$ is an absolute constant independent of $T$ and $V_T$.*

This theorem (in conjunction with the lower bound in Theorem 1) establishes the order of the minimax regret, namely

$$\mathcal{R}^*(V_T, T) \asymp (V_T)^{1/3}T^{2/3}.$$

**Remark 1** (Dependence on the Number of Arms). Our proposed policy, Rexp3, is driven primarily by its simplicity and the ability to elucidate key trade-offs in exploration–exploitation in the non-stationary problem setting. We note that there is a minor gap between the lower and upper bounds insofar as their dependence on $K$ is concerned. In particular, the logarithmic term in $K$ in the upper bound on the minimax regret can be removed by adapting other known policies that are designed for adversarial settings, for example, the INF policy that is described in Audibert and Bubeck (2009).

### 3.3. Further Extensions and Discussion

**3.3.1. A Continuous-Update Near-Optimal Policy.** The Rexp3 policy is particularly simple in its structure and lends itself to elucidating the exploration–exploitation trade-offs that exist in the non-stationary stochastic

model. It is, however, somewhat clunky in the manner in which it addresses the "remembering–forgetting" balance via the batching and abrupt restarts. The following policy, introduced in Auer et al. (2002b), can be modified to present a "smoother" counterpart to Rexp3.

**Exp3.S.** Inputs: positive numbers $\gamma$ and $\alpha$.

1. Initialization: for $t = 1$, for any $k \in \mathcal{K}$, set $w_t^k = 1$.
2. For each $t = 1, 2, \ldots,$
   - For each $k \in \mathcal{K}$, set
$$p_t^k = (1-\gamma)\frac{w_t^k}{\sum_{k'=1}^K w_t^{k'}} + \frac{\gamma}{K}.$$
   - Draw an arm $k'$ from $\mathcal{K}$ according to the distribution $\{p_t^k\}_{k=1}^K$, and receive a reward $X_t^{k'}$.
   - For $k'$, set $\hat{X}_t^{k'} = X_t^{k'}/p_t^{k'}$, and for any $k \neq k'$, set $\hat{X}_t^k = 0$.
   - For all $k \in \mathcal{K}$, update
$$w_{t+1}^k = w_t^k \exp\left\{\frac{\gamma \hat{X}_t^k}{K}\right\} + \frac{e\alpha}{K}\sum_{k'=1}^K w_t^{k'}.$$

The performance of Exp3.S was analyzed in Auer et al. (2002b) under a variant of the adversarial MAB formulation in which the *number of switches s* in the identity of the best arm is finite. The next result shows that by selecting appropriate tuning parameters, Exp3.S guarantees near-optimal performance in the much broader non-stationary stochastic setting we consider here.

**Theorem 3** (Near Rate Optimality for Continuous Updating). *Let $\pi$ be the Exp3.S policy with the parameters $\alpha = \frac{1}{T}$ and*
$$\gamma = \min\left\{1, \left(\frac{4V_T K \log(KT)}{(e-1)^2 T}\right)^{1/3}\right\}.$$
*Then, for every $T \geq 2$ and $V_T \in [K^{-1}, K^{-1}T]$, the worst-case regret is bounded from above as follows:*
$$\mathcal{R}^\pi(\mathcal{V}, T) \leq \bar{C}(K\log K)^{1/3}(V_T \log T)^{1/3} \cdot T^{2/3},$$
*where $\bar{C}$ is an absolute constant independent of $T$ and $V_T$.*

**3.3.2. Minimax Regret and Relation to Traditional (Stationary) MAB Problems.** The minimax regret should be contrasted with the stationary MAB problem in which $\sqrt{T}$ is the order of the minimax regret (see Auer et al. 2002a); if the arms are well separated, then the order is $\log T$ (see Lai and Robbins 1985). To illustrate the type of regret performance driven by the non-stationary environment, consider the case $V_T = C \cdot T^\beta$ for some $C > 0$ and $0 \leq \beta < 1$, in which the minimax regret is of order $T^{(2+\beta)/3}$. The driver of the change is the optimal exploration–exploitation balance. Beyond the "classical" exploration–exploitation trade-off, an additional key element of our problem is the non-stationary nature of the environment. In this context, there is an additional tension between remembering and forgetting. Specifically, keeping track of more observations may decrease the variance of the mean reward estimates, but "older" information is potentially less useful and might bias our estimates. (See also discussion along these lines for UCB-type policies, although in a different setup, in Garivier and Moulines (2011).) The design of Rexp3 reflects these considerations. An exploration rate $\gamma \approx \Delta^{-1/2} \approx (V_T/T)^{-1/3}$ leads to an order of $\gamma T \approx V_T^{1/3} T^{2/3}$ exploration periods, significantly more than the order of $T^{1/2}$ explorations that is optimal in the stationary setting (with non-separated rewards).

**3.3.3. Relation to Other Non-stationary MAB Instances.** The class of MAB problems with non-stationary rewards formulated in this paper extends to other MAB formulations that allow rewards to change in a more restricted manner. As mentioned earlier, when the variation budget grows linearly with the time horizon, the regret must grow linearly. This also implies the observation of Slivkins and Upfal (2008) in a setting in which rewards evolve according to a Brownian motion, and hence, the regret is linear in $T$. Our results can also be positioned relative to those of Garivier and Moulines (2011), who study stochastic MAB problems in which expected rewards may change a finite number of times, and Auer et al. (2002b), who formulate an adversarial MAB problem in which the identity of the best arm may change a finite number of times (independent of $T$). Both studies suggest policies that, using prior knowledge that the number of changes must be finite, achieve regret of order $\sqrt{T}$ relative to the best sequence of actions. As noted earlier, in our problem formulation, this

set of reward sequence instances would fall into the case in which the variation budget $V_T$ is fixed and independent of $T$. In that case, our results establish that the regret must be at least of order $T^{2/3}$. Moreover, a careful look at the proof of Theorem 1 clearly identifies the hard set of sequences as those that have a "large" number of changes in the expected rewards; hence, complexity in our problem setting is markedly different from that in the aforementioned studies. It is also worthwhile to note that in Auer et al. (2002b), the performance of Exp3.S is measured relative to a benchmark that resembles the dynamic oracle discussed in this paper, but even though the latter can switch arms in every epoch, the benchmark in Auer et al. (2002b) is limited to be a sequence of $s+1$ actions (each of them is ex post optimal in a different segment of the decision horizon). This represents an extension of the single-best-action-in-hindsight benchmark but is still far more restrictive than the dynamic oracle formulation we develop and pursue in this paper.

## 4. Numerical Results

We illustrate our main results for the near-optimal policy with continuous updating detailed in Section 3.3. This policy, unlike Rexp3, exhibits much smoother performance and is more conducive for illustrative purposes.

### 4.1. Setup

We consider instances in which two arms are available: $\mathcal{K} = \{1,2\}$. The reward $X_t^k$ associated with arm $k$ at epoch $t$ has a Bernoulli distribution with a changing expectation $\mu_t^k$:

$$X_t^k = \begin{cases} 1 & \text{w.p. } \mu_t^k, \\ 0 & \text{w.p. } 1-\mu_t^k, \end{cases}$$

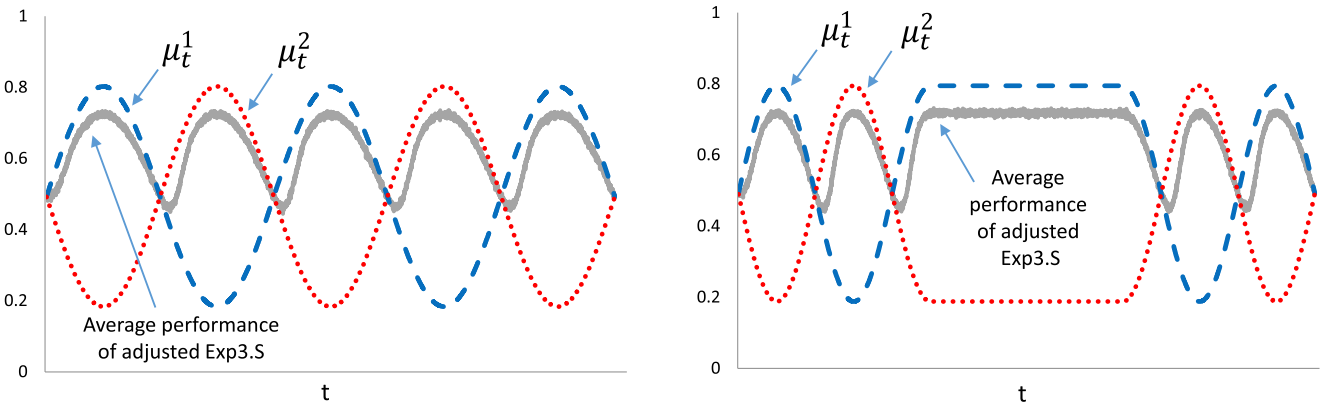
for all $t \in \mathcal{T} = \{1,\dots,T\}$ and for any arm $k \in \mathcal{K}$. The evolution patterns of $\mu_t^k$, $k \in \mathcal{K}$, are specified as follows. At each epoch $t \in \mathcal{T}$, the policy selects an arm $k \in \mathcal{K}$. Then the reward $X_t^k$ is realized and observed. The mean loss at epoch $t$ is $\mu_t^* - \mu_t^\pi$. Summing over the horizon and replicating, the average of the empirical cumulative mean loss approximates the expected regret compared with the dynamic oracle.

### 4.2. Experiment 1: Fixed Variation and Sensitivity to Time Horizon

The objective is to measure the growth rate of the regret as a function of the horizon length under a fixed variation budget. We use two basic instances. In the first instance (displayed on the left side of Figure 1) the expected rewards are sinusoidal:

$$\mu_t^1 = \frac{1}{2} + \frac{3}{10}\sin\left(\frac{5V_T\pi t}{3T}\right), \qquad \mu_t^2 = \frac{1}{2} + \frac{3}{10}\sin\left(\frac{5V_T\pi t}{3T} + \pi\right), \tag{2}$$

**Figure 3.** Numerical Simulation of the Average Performance Trajectory of the Adjusted Exp3.S Policy in Two Complementary Non-stationary Mean-Reward Instances




*Notes.* (Left) An instance with sinusoidal expected rewards with a fixed variation budget $V_T = 3$. (Right) An instance in which similar sinusoidal evolution is limited to the first and last thirds of the horizon. In both of the instances, the average performance trajectory of the policy is generated along $T = 1,500,000$ epochs.

for all $t \in \mathcal{T}$. In the second instance (depicted on the right side of Figure 1), a similar sinusoidal evolution is limited to the first and last thirds of the horizon, and in the middle third, mean rewards are constant:

$$\mu_t^1 = \begin{cases} \frac{1}{2} + \frac{3}{10}\sin\left(\frac{15V_T\pi t}{2T}\right) & \text{if } t < \frac{T}{3}, \\ \frac{4}{5} & \text{if } \frac{T}{3} \le t \le \frac{2T}{3}, \\ \frac{1}{2} + \frac{3}{10}\sin\left(\frac{15V_T\pi(t-T/3)}{2T}\right) & \text{if } \frac{2T}{3} < t \le T, \end{cases} \quad \mu_t^2 = \begin{cases} \frac{1}{2} + \frac{3}{10}\sin\left(\frac{15V_T\pi t}{2T} + \pi\right) & \text{if } t < \frac{T}{3}, \\ \frac{1}{5} & \text{if } \frac{T}{3} \le t \le \frac{2T}{3}, \\ \frac{1}{2} + \frac{3}{10}\sin\left(\frac{15V_T\pi(t-T/3)}{2T} + \pi\right) & \text{if } \frac{2T}{3} < t \le T, \end{cases}$$

for all $t \in \mathcal{T}$. Both instances describe different changing environments under the same fixed variation budget $V_T = 3$. In the first instance, the variation budget is "spent" more evenly throughout the horizon, and in the second instance, that budget is spent only over the first third of the horizon. For different values of $T$ up to $3 \cdot 10^8$, we use 100 replications to estimate the expected regret. The average performance trajectory (as a function of time) of the adjusted Exp3.S policy for $T = 1.5 \cdot 10^6$ is depicted using the solid curves in Figure 3 (the dotted and dashed curves correspond to the mean reward paths for each arm, respectively).

The first simulation experiment depicts the average performance of the policy (as a function of time) and illustrates the balance between exploration, exploitation, and the degree to which the policy needs to forget "stale" information. The policy selects the arm with the highest expected reward with higher probability. Of note are the delays in identifying the crossover points, which are evidently minor; the speed at which the policy switches to the new optimal action; and the fact that the policy keeps experimenting with the sub-optimal arm. These imply that the performance does not match the one of the dynamic oracle but rather falls short.

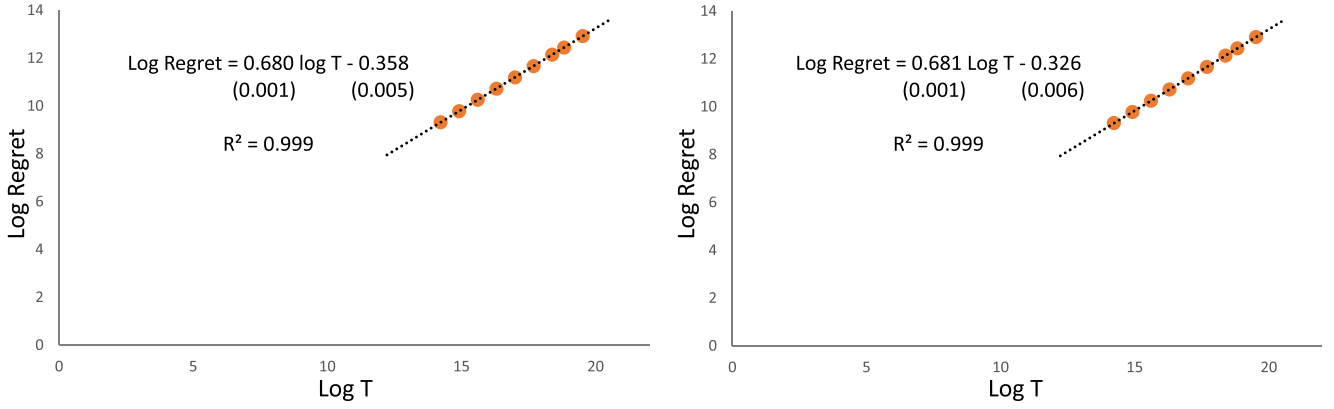
The two graphs in Figure 4 depict log–log plots of the mean regret as a function of the horizon. We observe that the slope is close to 2/3 and hence consistent with the result of Theorem 2 applied to a constant variation. The standard errors for the slope and intercept estimates are in parentheses.

### 4.3. Experiment 2: Fixed Time Horizon and Sensitivity to Variation

The objective of the second experiment is to measure how the regret growth rate (as a function of $T$) depends on the variation budget. For this purpose, we take $V_T = 3T^{\beta}$, and we explore the dependence of the regret on $\beta$. Under the sinusoidal variation instance in (2), Table 1 reports the estimates of regret rates obtained via slopes of the log–log plots for values of $\beta$ between zero (constant variation, simulated in Experiment 1) and 0.5.[2]

This simulation experiment illustrates how the variation level affects the policy's performance. The slopes of log–log dependencies of the regret as a function of the horizon length were estimated for the various $\beta$ values and are summarized in Table 1 along with standard errors. This is contrasted with the theoretical rates for the minimax regret obtained in previous results (Theorems 1 and 2). The estimated slopes illustrate the growth of regret as variation increases, and in that sense, Table 1 emphasizes the spectrum of minimax regret rates (of

**Figure 4.** Log–Log Plots of the Averaged Regret Incurred by the Adjusted Exp3.S as a Function of the Horizon Length $T$ with the Resulting Linear Relationship (Slope and Intercept) Estimates Under the Two Instances That Appear in Figure 3




*Notes.* (Left) Instance with sinusoidal expected rewards with a fixed variation budget $V_T = 3$. (Right) An instance in which similar sinusoidal evolution is limited to the first and last thirds of the horizon. Standard errors appear in parentheses.

**Table 1.** Estimated Log–Log Slopes for Growing Variation Budgets of the Structure $V_T = 3T^{\beta}$ (Standard Errors Appear in Parentheses) Contrasted with the Slopes ($T$ Dependence) for the Theoretical Minimax Regret Rates

| $\beta$-value | Theoretical slope $(2+\beta)/3$ | Estimated slope |
|---|---|---|
| 0.0 | 0.67 | 0.680 (0.001) |
| 0.1 | 0.70 | 0.710 (0.001) |
| 0.2 | 0.73 | 0.730 (0.001) |
| 0.3 | 0.77 | 0.766 (0.001) |
| 0.4 | 0.80 | 0.769 (0.001) |
| 0.5 | 0.83 | 0.812 (0.001) |

order $V_T^{1/3}T^{2/3}$) that are obtained for different variation levels. The numerical performance of the proposed policy achieves a regret rate that matches quite closely the theoretical values established in previous theorems.

**Remark 2.** As mentioned earlier, the Exp3.S policy was originally designed for an adversarial setting that considers a finite number of changes in the identity of the best arm and does not have performance guarantees under the framework of this paper, which allows for growing (horizon-dependent) changes in the values of the mean rewards as well as in the identity of the best arm. In particular, when there are $H$ changes in the identity of the best arm, the upper bound that is obtained for the performance of Exp3.S in Auer et al. (2002b) is of order $H\sqrt{T}$ (excluding logarithmic terms). Notably, our experiment illustrates a case in which the number of changes of the identity of the best arm is growing with $T$. In particular, when $\beta = 0.5$, the number of said changes is of order $\sqrt{T}$; thus the upper bound in Auer et al. (2002b) is of order $T$ and does not guarantee sublinear regret anymore in contrast with the upper bound of order $T^{5/6}$ that is established in this paper. Additional numerical experiments implied that at settings with high variation levels, the empirical performance of Exp3.S is dominated by the one achieved by the policies described in this paper. For example, in the setting of Experiment 2, for $\beta = 0.5$, the estimated slope for Exp3.S was 1 (implying linear regret) with standard error of 0.001; similar results were obtained in other instances of high variation and frequent switches in the identity of the best arm.

## 5. Adapting to Unknown Variation

### 5.1. Motivation and Overview

In previous sections, we established the minimax regret rates and have put forward rate-optimal policies that tune parameters using knowledge of the variation budget $V_T$. This leaves open the question of designing policies that can adapt to the variation "on the fly" without such a priori knowledge. A further issue, pertinent to this question, is the behavior of the current policy that is tuned to the "worst case" variation. Specifically, the proposed Rexp3 policy guarantees rate optimality by countering the worst-case sequence of mean rewards in $\mathcal{L}(V_T)$. In so doing, it ends up "overexploring" whether the realized environment is more benign. By contrast, if the actual variation turns out to exceed the variation budget through which the policy is tuned, the policy should be expected to incur additional regret as well. To formalize these observations, one can modify the analysis of Theorem 2 to represent the regret bound in terms of both the input $V_T$ and the actual variation $\mathcal{V}(\mu;T)$ to establish that the worst-case regret for Rexp3 is bounded from above as follows:

$$\mathcal{R}^{\pi}(V_T,T) \le \bar{C}(K\log K)^{1/3}T^{2/3}\cdot\max\left\{V_T^{1/3},\frac{\mathcal{V}(\mu;T)}{V_T^{2/3}}\right\}, \tag{3}$$

where $\bar{C}$ is the same absolute constant that appears in Theorem 2. A similar result can be obtained for the adjusted Exp3.S policy, respectively adapting the proof of Theorem 3.

From the expression in (3), it is possible to tease out conditions under which long-run-average optimality is ensured under an inaccurate input $V_T$, which does not match the actual variation $\mathcal{V}(\mu;T)$. This also quantifies the loss associated with using an inaccurate input $V_T$ as opposed to tuning the policy using the actual variation $\mathcal{V}(\mu;T)$. On the one hand, Equation (3) implies that whenever the actual variation does not exceed the input $V_T$, long-run-average optimality is guaranteed. For example, if the policy is tuned using $V_T = T^{\alpha}$ for some $\alpha \in (0,1)$, and the variation on mean rewards is zero, which is an admissible sequence contained in $\mathcal{L}(V_T)$, then the policy incurs a regret of order $T^{(2+\alpha)/3}$. Although this is long-run-average optimal for any $\alpha \in (0,1)$, it also includes a regret rate "penalty" relative to the optimal regret rate of order $T^{1/2}$ that would have been achieved if it was known a priori that there would be no variation in the mean rewards.

On the other hand, Equation (3) also implies that long-run-average optimality can be guaranteed when the actual variation *exceeds* the input $V_T$ as long as $V_T$ is "close enough" to the actual variation. In particular, when the policy is tuned by $V_T = T^\alpha$ for some $\alpha \in (0,1)$, it guarantees long-run-average optimality as long as the actual variation $\mathcal{V}(\mu;T)$ is at most of order $T^{\alpha+\delta}$ for some $\delta < (1-\alpha)/3$. For example, if $\alpha = 0$ and $\delta = 1/4$, Rexp3 guarantees sublinear regret of order $T^{11/12}$ (accurate tuning would have guaranteed order $T^{3/4}$).

Because there are no restrictions on the rate at which the variation budget can be spent, an interesting and challenging open problem is to delineate to what extent it is possible to design adaptive policies that do not use prior knowledge of $V_T$ yet guarantee good performance. In the rest of this section, we indicate a possible approach to address this issue. Ideally, one would like the regret of adaptive policies to scale with the actual variation $\mathcal{V}(\mu;T)$ rather than with the upper bound $V_T$. Our approach uses an envelope policy subordinate to which are multiple primitive restart Exp3-type policies. Each of the latter is tuned to a different guess of the variation budget, and the master policy switches among the subordinates based on realized rewards. Although we have no proof of optimality or strong theoretical indication to believe an optimal adaptive policy is to be found in this family, we believe that the main ideas presented in our approach may be useful for deriving fully adaptive rate-optimal policies.

### 5.2. An Envelope Policy

We introduce an envelope policy $\bar{\pi}$ that judiciously applies $M$ subordinate MAB policies $\{\pi^m : m = 1,\dots,M\}$. At each epoch, only one of these "virtual" policies may be used, but the collected information can be shared among all subordinate policies (concrete subordinate MAB policies are suggested in Section 5.3).

**Envelope Policy** ($\bar{\pi}$).

Inputs: $M$ admissible policies $\{\pi^m\}_{m=1}^M$ and a number $\bar{\gamma} \in (0,1]$.

1. Initialization: for $t = 1$, for any $m \in \{1,\dots,M\}$, set $v_t^m = 1$.
2. For each $t = 1,2,\dots$, do
   - For each $m \in \{1,\dots,M\}$, set
   $$q_t^m = (1-\bar{\gamma})\frac{v_t^m}{\sum_{m'=1}^M v_t^{m'}} + \frac{\bar{\gamma}}{M}.$$
   - Draw $m'$ from $\{1,\dots,M\}$ according to the distribution $\{q_t^m\}_{m=1}^M$.
   - Select the arm $\hat{k} = \pi_t^{m'}$ from the set $\{1,\dots,K\}$ and receive a reward $X_t^{\hat{k}}$.
   - Set $\hat{Y}_t^{m'} = X_t^{\hat{k}}/q_t^{m'}$, and for any $m \neq m'$, set $\hat{Y}_t^m = 0$.
   - For all $m \in \{1,\dots,M\}$, update:
   $$v_{t+1}^m = v_t^m \exp\left\{\frac{\bar{\gamma}\hat{Y}_t^m}{M}\right\}.$$
   - Update subordinate policies $\{\pi^m : m = 1,\dots,M\}$ (details of subordinate policy structure follow).

The envelope policy operates as follows. At each epoch, a subordinate policy is drawn from a distribution $\{q_t^m\}_{m=1}^M$ that is updated every epoch based on the observed rewards. This distribution is endowed with an exploration rate $\bar{\gamma}$ that is used to balance exploration and exploitation over the subordinate policies.

### 5.3. Structure of the Subordinate Policies

In Section 3.2, we identified classes of candidate Rexp3 and adjusted Exp3.S policies that were shown to achieve rate optimality when *accurately tuned* ex ante using the bound $V_T$. We therefore take $(\pi^1,\dots,\pi^m)$, the subordinate policies, to be the proposed adjusted Exp3.S policies tuned by exploration rates $(\gamma_1\dots,\gamma_M)$ to be specified. We denote by $\bar{\pi}_t \in \{1,\dots,M\}$ the action of the envelope policy at time $t$ and by $\pi_t^m \in \{1,\dots,K\}$ the action of policy $\pi^m$ at the same epoch. We denote by $\{p_t^{k,m} : k = 1,\dots,K\}$ the distribution from which $\pi_t^m$ is drawn and by $\{w_t^{k,m} : k = 1,\dots,K\}$ the weights associated with this distribution according to the Rexp3.S structure. We adjust the Rexp3.S description that is given in Section 3.3 by defining the following update rule. At each epoch $t$ and for each arm $k$, each subordinate Rexp3.S policy $\pi^m$ selects an arm in $\{1,\dots,K\}$ according to the distribution $\{p_t^{k,m}\}$. However, rather than updating the weights $\{w_t^{k,m}\}$ using the observation from that

arm, each subordinate policy uses the update rule

$$\hat{X}_t^k = \sum_{m=1}^{M} \frac{X_t^k}{p_t^{k,m}} \mathbb{1}\{\bar{\pi}_t = m\} \mathbb{1}\{\pi_t^m = k\}, \quad k = 1, \ldots, K,$$

and then updates weights as in the original Rexp3.S description

$$w_{t+1}^{k,m} = w_t^{k,m} \exp\left\{\frac{\gamma_m \hat{X}_t^k}{K}\right\} + \frac{e\alpha}{K} \sum_{k'=1}^{K} w_t^{k',m}.$$

We note that at each time step, the variables $\hat{X}_t^k$ are updated in an identical manner by all the subordinate Rexp3.S policies (independent of $m$). This update rule implies that information is shared between subordinate policies.

### 5.4. Numerical Analysis

We consider a case in which there are three possible levels for the realized variation $\mathcal{V}(\mu; T)$. With slight abuse of notation, we write $\mathcal{V}(\mu; T) \in \{V_{T,1}, V_{T,2}, V_{T,3}\}$ with $V_{T,1} = 0$ (no variation), $V_{T,2} = 3$ (fixed variation), and $V_{T,3} = 3T^{0.2}$ (increasing variation). The envelope policy does not know which of the latter variation levels describes the actual variation. We measured the performance of the envelope policy, tuned by

$$\bar{\gamma} = \min\left\{1, \sqrt{M \log(M) \cdot (e-1)^{-1} \cdot T^{-1}}\right\},$$

and with three input policies that guess $V_{T,1} = 0$, $V_{T,2} = 3$, and $V_{T,3} = 3T^{0.2}$, respectively. More formally, the input policies are selected as follows: $\pi_1$ being the Exp3.S policy with

$$\alpha = 1/T$$

and

$$\gamma_1 = \min\left\{1, \sqrt{K \log(KT) \cdot T^{-1}}\right\}$$

(the parametric values that appear in Auer et al. (2002b)); $\pi_2$ and $\pi_3$ are both Exp3.S policies with

$$\alpha = 1/T$$

and

$$\gamma_m = \min\left\{1, \left(\frac{2V_{T,m} K \log(KT)}{(e-1)^2 T}\right)^{1/3}\right\}$$

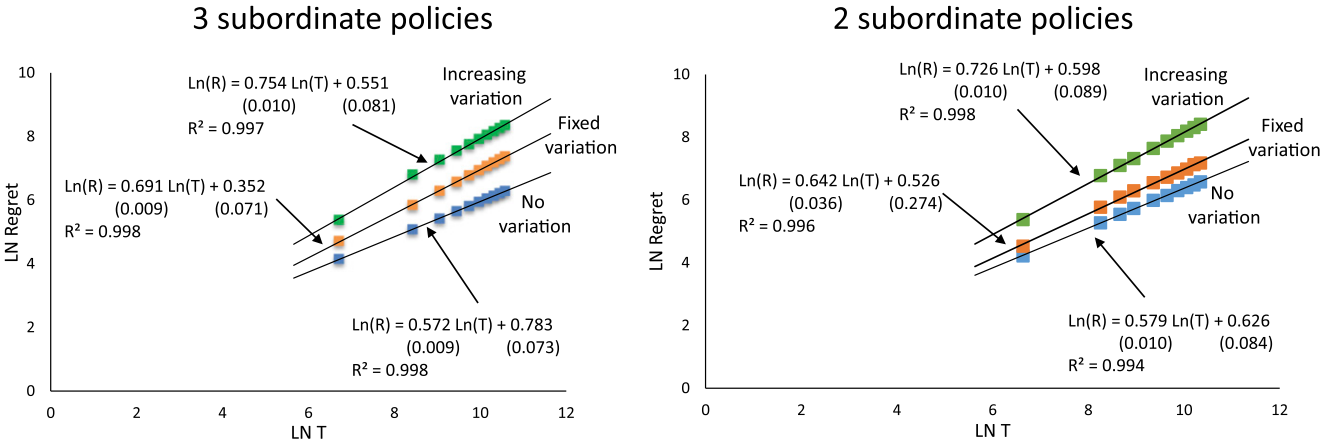
with the respective $V_{T,m}$ values. Each subordinate Exp3.S policy uses the update rule described in the preceding paragraph.

We measure the performance of the envelope policy in three different settings corresponding to the different values of the variation $\mathcal{V}(\mu; T)$ specified. When the realized variation is $\mathcal{V}(\mu; T) = V_{T,1} = 0$ (no variation), the mean rewards were $\mu_t^1 = 0.2$ and $\mu_t^2 = 0.8$ for all $t = 1, \ldots, T$. When the variation is $\mathcal{V}(\mu; T) = V_{T,2} = 3$ (fixed variation), the mean rewards followed the sinusoidal variation instance in (2). When the variation is $\mathcal{V}(\mu; T) = V_{T,3} = 3T^{0.2}$ (increasing variation), the mean rewards follow the same sinusoidal pattern in which variation increased with the horizon length. We then repeat the experiment while considering an envelope policy with only two subordinate policies, one that guesses $V_{T,1} = 0$ (no variation) and one that guesses $V_{T,3} = 3T^{0.2}$ (increasing variation), but measuring its performance under each of the cases.

### 5.5. Discussion

The left side of Figure 5 depicts the three log–log plots obtained for the three scenarios. The slopes illustrate that the envelope policy seems to adapt to different realized variation levels. In particular, the slopes appear close to those that would have been obtained with prior knowledge of the variation level. However, the uncertainty regarding the realized variation may cause a larger multiplicative constant in the regret expression; this is demonstrated by the higher intercept in the log–log plot of the fixed variation case (0.352) relative to the intercept obtained when the same variation level is known a priori (–0.358); see left part of Figure 4.

**Figure 5.** Adapting to Unknown Variation: Log–Log Plots of Regret as a Function of the Horizon Length $T$ Obtained Under Three Different Variation Levels: No Variation, Fixed Variation, and Increasing Variation of the Form $V_T = 3T^{0.2}$




*Notes.* (Left) Performance of an envelope policy with three subordinate policies, each corresponding to one of the former variation levels. (Right) Performance of the envelope policy with two subordinate policies: one corresponding to no variation and one corresponding to the increasing variation case (standard errors appear in parentheses).

The envelope policy seems effective when the set of possible variations is known a priori. The second part of the experiment suggests that a similar approach may be effective even when one cannot limit a priori the set of realized variations. The right side of Figure 5 depicts the three log–log plots obtained under the three cases when the envelope policy uses only two subordinate policies, one that guesses no variation and one that guesses an increasing variation. Although performance was sensitive to the initial amplitude of the sinusoidal functions, the results suggest that, on average, the envelope policy achieved comparable performance to the one using three subordinate policies, which include the well-specified fixed variation case.

## Acknowledgment

An initial version of this work, including some preliminary results, appeared in Besbes et al. (2014).

## Appendix. Proofs of Main Results

### Proof of Theorem 1

At a high level, the proof adapts a general approach of identifying a worst-case nature "strategy" (see proof of theorem 5.1 in Auer et al. (2002b), which analyzes the worst-case regret relative to a single best action benchmark in a fully adversarial environment), extending these ideas appropriately to our setting. Fix $T \geq 1$, $K \geq 2$, and $V_T \in [K^{-1}, K^{-1}T]$. In what follows, we restrict nature to the class $\mathcal{V}' \subseteq \mathcal{V}$ that was described in Section 3 and show that when $\mu$ is drawn randomly from $\mathcal{V}'$, any policy in $\mathcal{P}$ must incur regret of order $(KV_T)^{1/3}T^{2/3}$.

**Step 1** (Preliminaries). Define a partition of the decision horizon $\mathcal{T}$ to $m = \lceil \frac{T}{\Delta} \rceil$ batches $\mathcal{T}_1, \ldots, \mathcal{T}_m$ of size $\Delta$ each (except perhaps $\mathcal{T}_m$), according to $\mathcal{T}_j = \{t : (j-1)\Delta + 1 \leq t \leq \min\{j\Delta, T\}\}$ for all $j = 1, \ldots, m$, where $m = \lceil T/\Delta \rceil$ is the number of batches. For some $\varepsilon > 0$ that is specified shortly, define $\mathcal{V}'$ to be the set of reward vector sequences $\mu$ such that

- $\mu_t^k \in \{1/2, 1/2 + \varepsilon\}$ for all $k \in \mathcal{K}$, $t \in \mathcal{T}$;
- $\sum_{k \in \mathcal{K}} \mu_t^k = K/2 + \varepsilon$ for all $t \in \mathcal{T}$;
- $\mu_t^k = \mu_{t+1}^k$ for any $(j-1)\Delta + 1 \leq t \leq \min\{j\Delta, T\} - 1$, $j = 1, \ldots, m$, for all $k \in \mathcal{K}$.

For each sequence in $\mathcal{V}'$ in any epoch, there is exactly one arm with expected reward $1/2 + \varepsilon$, and the rest of the arms have expected reward 1/2, and expected rewards cannot change within a batch. Let $\varepsilon = \min\{\frac{1}{4} \cdot \sqrt{K/\Delta}, V_T\Delta/T\}$. Then, for any $\mu \in \mathcal{V}'$, one has

$$\sum_{t=1}^{T-1} \sup_{k \in \mathcal{K}} |\mu_t^k - \mu_{t+1}^k| \leq \sum_{j=1}^{m-1} \varepsilon = \left( \left\lceil \frac{T}{\Delta} \right\rceil - 1 \right) \cdot \varepsilon \leq \frac{T\varepsilon}{\Delta_T} \leq V_T,$$

where the first inequality follows from the structure of $\mathcal{V}'$. Therefore, $\mathcal{V}' \subset \mathcal{V}$.

**Step 2** (Single-Batch Analysis). Fix some policy $\pi \in \mathcal{P}$, and fix a batch $j \in \{1, \ldots, m\}$. Let $k_j$ denote the good arm of batch $j$. We denote by $\mathbb{P}_{k_j}^j$ the probability distribution conditioned on arm $k_j$ being the good arm in batch $j$ and by $\mathbb{P}_0$ the probability distribution with respect to random rewards (i.e., expected reward 1/2) for each arm. We further denote by $\mathbb{E}_{k_j}^j[\cdot]$ and $\mathbb{E}_0[\cdot]$ the respective expectations. Assuming binary rewards, we let $X$ denote a vector of $|\mathcal{T}_j|$ rewards; that is, $X \in \{0,1\}^{|\mathcal{T}_j|}$. We

denote by $N_k^j$ the number of times arm $k$ was selected in batch $j$. In the proof we use lemma A.1 from Auer et al. (2002b), which characterizes the difference between the two different expectations of some function of the observed rewards vector.

**Lemma A.1** (lemma A.1 from Auer et al. 2002b). *Let $f : \{0,1\}^{|\mathcal{T}_j|} \to [0,M]$ be a bounded real function. Then, for any $k \in \mathcal{H}$,*

$$\mathbb{E}_k^j[f(X)] - \mathbb{E}_0[f(X)] \le \frac{M}{2}\sqrt{-\mathbb{E}_0\left[N_k^j\right]\log(1-4\varepsilon^2)}.$$

Recalling that $k_j$ denotes the good arm of batch $j$, one has

$$\mathbb{E}_{k_j}^j[\mu_t^\pi] = \left(\frac{1}{2}+\varepsilon\right)\mathbb{P}_{k_j}^j\{\pi_t = k_j\} + \frac{1}{2}\mathbb{P}_{k_j}^j\{\pi_t \ne k_j\} \;=\; \frac{1}{2} + \varepsilon\mathbb{P}_{k_j}^j\{\pi_t = k_j\},$$

and therefore,

$$\mathbb{E}_{k_j}^j\left[\sum_{t\in\mathcal{T}_j}\mu_t^\pi\right] = \frac{|\mathcal{T}_j|}{2} + \sum_{t\in\mathcal{T}_j}\varepsilon\mathbb{P}_{k_j}^j\{\pi_t = k_j\} = \frac{|\mathcal{T}_j|}{2} + \varepsilon\mathbb{E}_{k_j}^j\left[N_{k_j}^j\right]. \tag{A.1}$$

In addition, applying Lemma A.1 with $f(X) = N_{k_j}^j$ (clearly $N_{k_j}^j \in \{0,\dots,|\mathcal{T}_j|\}$), we have

$$\mathbb{E}_{k_j}^j\left[N_{k_j}^j\right] \le \mathbb{E}_0\left[N_{k_j}^j\right] + \frac{|\mathcal{T}_j|}{2}\sqrt{-\mathbb{E}_0\left[N_{k_j}^j\right]\log(1-4\varepsilon^2)}.$$

Summing over arms, one has

$$\begin{aligned}\sum_{k_j=1}^K \mathbb{E}_{k_j}^j\left[N_{k_j}^j\right] &\le \sum_{k_j=1}^K \mathbb{E}_0\left[N_{k_j}^j\right] + \sum_{k_j=1}^K \frac{|\mathcal{T}_j|}{2}\sqrt{-\mathbb{E}_0\left[N_{k_j}^j\right]\log(1-4\varepsilon^2)} \\ &\le |\mathcal{T}_j| + \frac{|\mathcal{T}_j|}{2}\sqrt{-\log(1-4\varepsilon^2)|\mathcal{T}_j|K},\end{aligned} \tag{A.2}$$

for any $j \in \{1,\dots,m\}$, where the last inequality holds because $\sum_{k_j=1}^K \mathbb{E}_0[N_{k_j}^j] = |\mathcal{T}_j|$, and thus, by Cauchy–Schwarz inequality,

$$\sum_{k_j=1}^K \sqrt{\mathbb{E}_0\left[N_{k_j}^j\right]} \le \sqrt{|\mathcal{T}_j|K}.$$

**Step 3** (Regret Along the Horizon). Let $\tilde{\mu}$ be a random sequence of expected rewards vectors in which in every batch the good arm is drawn according to an independent uniform distribution over the set $\mathcal{H}$. Clearly, every realization of $\tilde{\mu}$ is in $\mathcal{V}'$. In particular, taking expectation over $\tilde{\mu}$, one has

$$\begin{aligned}\mathcal{R}^\pi(\mathcal{V}',T) &= \sup_{\mu\in\mathcal{V}'}\left\{\sum_{t=1}^T \mu_t^* - \mathbb{E}^\pi\left[\sum_{t=1}^T \mu_t^\pi\right]\right\} \ge \mathbb{E}^{\tilde{\mu}}\left[\sum_{t=1}^T \tilde{\mu}_t^* - \mathbb{E}^\pi\left[\sum_{t=1}^T \tilde{\mu}_t^\pi\right]\right] \ge \sum_{j=1}^m\left(\sum_{t\in\mathcal{T}_j}\left(\frac{1}{2}+\varepsilon\right) - \frac{1}{K}\sum_{k_j=1}^K \mathbb{E}^\pi\mathbb{E}_{k_j}^j\left[\sum_{t\in\mathcal{T}_j}\tilde{\mu}_t^\pi\right]\right) \\ &\overset{(a)}{\ge} \sum_{j=1}^m\left(\sum_{t\in\mathcal{T}_j}\left(\frac{1}{2}+\varepsilon\right) - \frac{1}{K}\sum_{k_j=1}^K\left(\frac{|\mathcal{T}_j|}{2} + \varepsilon\mathbb{E}^\pi\mathbb{E}_{k_j}^j\left[N_{k_j}^j\right]\right)\right) \ge \sum_{j=1}^m\left(\sum_{t\in\mathcal{T}_j}\left(\frac{1}{2}+\varepsilon\right) - \frac{|\mathcal{T}_j|}{2} - \frac{\varepsilon}{K}\mathbb{E}^\pi\sum_{k_j=1}^K\mathbb{E}_{k_j}^j\left[N_{k_j}^j\right]\right) \\ &\overset{(b)}{\ge} T\varepsilon - \frac{T\varepsilon}{K} - \frac{T\varepsilon}{2K}\sqrt{-\log(1-4\varepsilon^2)\Delta K} \\ &\overset{(c)}{\ge} \frac{T\varepsilon}{2} - \frac{T\varepsilon^2}{K}\sqrt{\log(4/3)\Delta K},\end{aligned}$$

where $(a)$ holds by (A.1); $(b)$ holds by (A.2) because $\sum_{j=1}^m |\mathcal{T}_j| = T$, because $m \ge T/\Delta$, and because $|\mathcal{T}_j| \le \Delta$ for all $j \in \{1,\dots,m\}$; and $(c)$ holds by $4\varepsilon^2 \le 1/4$, and $-\log(1-x) \le 4\log(4/3)x$ for all $x \in [0,1/4]$ and because $K \ge 2$. Set $\Delta = \lceil K^{1/3}(\frac{T}{V_T})^{2/3}\rceil$. Because $\varepsilon = \min\{\frac{1}{4}\cdot\sqrt{K/\Delta}_T, V_T\Delta/T\}$, one has

$$\begin{aligned}\mathcal{R}^\pi(\mathcal{V}',T) &\ge T\varepsilon\left(\frac{1}{2} - \varepsilon\sqrt{\frac{\Delta_T\log(4/3)}{K}}\right) \ge T\varepsilon\left(\frac{1}{2} - \frac{\sqrt{\log(4/3)}}{4}\right) \ge \frac{1}{4}\cdot\min\left\{\frac{T}{4}\cdot\sqrt{\frac{K}{\Delta}}, V_T\Delta\right\} \\ &\ge \frac{1}{4}\cdot\min\left\{\frac{T}{4}\cdot\sqrt{\frac{K}{2K^{1/3}(T/V_T)^{2/3}}}, (KV_T)^{1/3}T^{2/3}\right\} \ge \frac{1}{4\sqrt{2}}\cdot(KV_T)^{1/3}T^{2/3}.\end{aligned}$$

This concludes the proof. □

### Proof of Theorem 2

The structure of the proof is as follows. First, we break the horizon into a sequence of batches of size $\Delta$ each and analyze the performance gap between the single best action and the dynamic oracle in each batch. Then we plug in a known performance guarantee for Exp3 relative to the single best action and sum over batches to establish the regret of Rexp3 relative to the dynamic oracle.

**Step 1** (Preliminaries). Fix $T \geq 1$, $K \geq 2$, and $V_T \in [K^{-1}, K^{-1}T]$. Let $\pi$ be the Rexp3 policy, tuned by

$$\gamma = \min\left\{1\ ,\ \sqrt{\frac{K\log K}{(e-1)\Delta}}\right\}$$

and $\Delta \in \{1,\dots,T\}$ (to be specified later). Define a partition of the decision horizon $\mathcal{T}$ to $m = \lceil \frac{T}{\Delta} \rceil$ batches $\mathcal{T}_1,\dots,\mathcal{T}_m$ of size $\Delta$ each (except perhaps $\mathcal{T}_m$), according to $\mathcal{T}_j = \{t : (j-1)\Delta + 1 \leq t \leq \min\{j\Delta, T\}\}$ for all $j = 1,\dots,m$, where $m = \lceil T/\Delta \rceil$ is the number of batches. Let $\mu \in \mathcal{V}$, and fix $j \in \{1,\dots,m\}$. We decompose the regret in batch $j$:

$$\mathbb{E}^\pi\left[\sum_{t\in\mathcal{T}_j}(\mu_t^* - \mu_t^\pi)\right] = \underbrace{\sum_{t\in\mathcal{T}_j}\mu_t^* - \mathbb{E}\left[\max_{k\in\mathcal{K}}\left\{\sum_{t\in\mathcal{T}_j}X_t^k\right\}\right]}_{J_{1,j}} + \underbrace{\mathbb{E}\left[\max_{k\in\mathcal{K}}\left\{\sum_{t\in\mathcal{T}_j}X_t^k\right\}\right] - \mathbb{E}^\pi\left[\sum_{t\in\mathcal{T}_j}\mu_t^\pi\right]}_{J_{2,j}}. \tag{A.3}$$

The first component, $J_{1,j}$, is the expected loss associated with using a single action over batch $j$. The second component, $J_{2,j}$, is the expected regret relative to the best static action in batch $j$.

**Step 2** (Analysis of $J_{1,j}$ and $J_{2,j}$). Defining $\mu_{T+1}^k = \mu_T^k$ for all $k \in \mathcal{K}$, we denote the variation in expected rewards along batch $\mathcal{T}_j$ by $V_j = \sum_{t\in\mathcal{T}_j}\max_{k\in\mathcal{K}}|\mu_{t+1}^k - \mu_t^k|$. We note that

$$\sum_{j=1}^m V_j = \sum_{j=1}^m\sum_{t\in\mathcal{T}_j}\max_{k\in\mathcal{K}}|\mu_{t+1}^k - \mu_t^k| \leq V_T. \tag{A.4}$$

Let $k_0$ be an arm with best expected performance over $\mathcal{T}_j$: $k_0 \in \arg\max_{k\in\mathcal{K}}\{\sum_{t\in\mathcal{T}_j}\mu_t^k\}$. Then

$$\max_{k\in\mathcal{K}}\left\{\sum_{t\in\mathcal{T}_j}\mu_t^k\right\} = \sum_{t\in\mathcal{T}_j}\mu_t^{k_0} = \mathbb{E}\left[\sum_{t\in\mathcal{T}_j}X_t^{k_0}\right] \leq \mathbb{E}\left[\max_{k\in\mathcal{K}}\left\{\sum_{t\in\mathcal{T}_j}X_t^k\right\}\right], \tag{A.5}$$

and therefore, one has

$$J_{1,j} = \sum_{t\in\mathcal{T}_j}\mu_t^* - \mathbb{E}\left[\max_{k\in\mathcal{K}}\left\{\sum_{t\in\mathcal{T}_j}X_t^k\right\}\right] \overset{(a)}{\leq} \sum_{t\in\mathcal{T}_j}\left(\mu_t^* - \mu_t^{k_0}\right) \leq \Delta\max_{t\in\mathcal{T}_j}\left\{\mu_t^* - \mu_t^{k_0}\right\} \overset{(b)}{\leq} 2V_j\Delta, \tag{A.6}$$

for any $\mu \in \mathcal{V}$ and $j \in \{1,\dots,m\}$, where $(a)$ holds by (A.5) and $(b)$ holds by the following argument: otherwise, there is an epoch $t_0 \in \mathcal{T}_j$ for which $\mu_{t_0}^* - \mu_{t_0}^{k_0} > 2V_j$. Indeed, suppose that $k_1 = \arg\max_{k\in\mathcal{K}}\mu_{t_0}^k$. In such a case, for all $t \in \mathcal{T}_j$, one has $\mu_t^{k_1} \geq \mu_{t_0}^{k_1} - V_j > \mu_{t_0}^{k_0} + V_j \geq \mu_t^{k_0}$ because $V_j$ is the maximal variation in batch $\mathcal{T}_j$. This, however, contradicts the optimality of $k_0$ at epoch $t$, and thus, (A.6) holds. In addition, corollary 3.2 in Auer et al. (2002b) points out that the regret incurred by Exp3 (tuned by $\gamma = \min\{1\ ,\ \sqrt{\frac{K\log K}{(e-1)\Delta}}\}$) along $\Delta$ epochs, relative to the single best action, is bounded by $2\sqrt{e-1}\sqrt{\Delta K\log K}$. Therefore, for each $j \in \{1,\dots,m\}$, one has

$$J_{2,j} = \mathbb{E}\left[\max_{k\in\mathcal{K}}\left\{\sum_{t\in\mathcal{T}_j}X_t^k\right\}\right] - \mathbb{E}^\pi\left[\sum_{t\in\mathcal{T}_j}\mu_t^\pi\right] \overset{(a)}{\leq} 2\sqrt{e-1}\sqrt{\Delta K\log K}, \tag{A.7}$$

for any $\mu \in \mathcal{V}$, where $(a)$ holds because within each batch arms are pulled according to Exp3($\gamma$).

**Step 3** (Regret Along the Horizon). Summing over $m = \lceil T/\Delta \rceil$ batches, we have

$$\begin{aligned}\mathcal{R}^\pi(\mathcal{V},T) &= \sup_{\mu\in\mathcal{V}}\left\{\sum_{t=1}^T\mu_t^* - \mathbb{E}^\pi\left[\sum_{t=1}^T\mu_t^\pi\right]\right\} \overset{(a)}{\leq} \sum_{j=1}^m\left(2\sqrt{e-1}\sqrt{\Delta K\log K} + 2V_j\Delta\right)\\ &\overset{(b)}{\leq} \left(\frac{T}{\Delta}+1\right)\cdot 2\sqrt{e-1}\sqrt{\Delta K\log K} + 2\Delta V_T = \frac{2\sqrt{e-1}\sqrt{K\log K}\cdot T}{\sqrt{\Delta}} + 2\sqrt{e-1}\sqrt{\Delta K\log K} + 2\Delta V_T,\end{aligned} \tag{A.8}$$

where $(a)$ holds by (A.3), (A.6), and (A.7), and $(b)$ follows from (A.4). Selecting $\Delta = \lceil (K\log K)^{1/3}(T/V_T)^{2/3} \rceil$, one has

$$\begin{aligned}\mathcal{R}^{\pi}(\mathcal{V},T) &\le 2\sqrt{e-1}\left(K\log K\cdot V_T\right)^{1/3}T^{2/3} + 2\sqrt{e-1}\sqrt{\left((K\log K)^{1/3}(T/V_T)^{2/3}+1\right)K\log K} \\ &\quad + 2\left((K\log K)^{1/3}(T/V_T)^{2/3}+1\right)V_T \\ &\overset{(a)}{\le} \left(\left(2+2\sqrt{2}\right)\sqrt{e-1}+4\right)\left(K\log K\cdot V_T\right)^{1/3}T^{2/3},\end{aligned}$$

where $(a)$ follows from $T \ge K \ge 2$, and $V_T \in [K^{-1}, K^{-1}T]$. This concludes the proof. □

**Proof of Theorem 3**

We prove that by selecting the tuning parameters to be $\alpha = \frac{1}{T}$ and

$$\gamma = \min\left\{1, \left(\frac{4V_T K\log(KT)}{(e-1)^2 T}\right)^{1/3}\right\},$$

Exp3.S achieves near-optimal performance in the non-stationary stochastic setting. The structure of the proof is as follows. First, we break the decision horizon into a sequence of decision batches and analyze the difference in performance between the sequence of single best actions and the performance of the dynamic oracle. Then we analyze the regret of Exp3.S relative to a sequence composed of the single best actions of each batch (this part of the proof roughly follows the proof lines of theorem 8.1 in Auer et al. (2002b) while considering a possibly infinite number of changes in the identity of the best arm). Finally, we select tuning parameters that minimize the overall regret.

**Step 1** (Preliminaries). Fix $T \ge 1$, $K \ge 2$, and $V_T \in [K^{-1}, K^{-1}T]$. Let $\pi$ be the Exp3.S policy described in Section 3.2, tuned by $\alpha = \frac{1}{T}$ and

$$\gamma = \min\left\{1, \left(\frac{4V_T K\log(KT)}{(e-1)^2 T}\right)^{1/3}\right\},$$

and let $\Delta \in \{1,\dots,T\}$ be a batch size (to be specified later). We break the horizon $\mathcal{T}$ into a sequence of batches $\mathcal{T}_1,\dots,\mathcal{T}_m$ of size $\Delta$ each (except possibly $\mathcal{T}_m$) according to $\mathcal{T}_j = \{t : (j-1)\Delta + 1 \le t \le \min\{j\Delta, T\}\}$, $j = 1,\dots,m$. Let $\mu \in \mathcal{V}$, and fix $j \in \{1,\dots,m\}$. We decompose the regret in batch $j$:

$$\mathbb{E}^{\pi}\left[\sum_{t\in\mathcal{T}_j}\left(\mu_t^* - \mu_t^{\pi}\right)\right] = \underbrace{\sum_{t\in\mathcal{T}_j}\mu_t^* - \max_{k\in\mathcal{K}}\left\{\sum_{t\in\mathcal{T}_j}\mu_t^k\right\}}_{J_{1,j}} + \underbrace{\max_{k\in\mathcal{K}}\left\{\sum_{t\in\mathcal{T}_j}\mu_t^k\right\} - \mathbb{E}^{\pi}\left[\sum_{t\in\mathcal{T}_j}\mu_t^{\pi}\right]}_{J_{2,j}}. \tag{A.9}$$

The first component, $J_{1,j}$, corresponds to the expected loss associated with using a single action over batch $j$. The second component, $J_{2,j}$, is the regret relative to the best static action in batch $j$.

**Step 2** (Analysis of $J_{1,j}$). Define $\mu_{T+1}^k = \mu_T^k$ for all $k \in \mathcal{K}$, and denote $V_j = \sum_{t\in\mathcal{T}_j}\max_{k\in\mathcal{K}}|\mu_{t+1}^k - \mu_t^k|$. We note that

$$\sum_{j=1}^{m} V_j = \sum_{j=1}^{m}\sum_{t\in\mathcal{T}_j}\max_{k\in\mathcal{K}}|\mu_{t+1}^k - \mu_t^k| \le V_T. \tag{A.10}$$

Letting $k_0 \in \arg\max_{k\in\mathcal{K}}\{\sum_{t\in\mathcal{T}_j}\mu_t^k\}$, we follow Step 2 in the proof of Theorem 2 to establish for any $\mu \in \mathcal{V}$ and $j \in \{1,\dots,m\}$,

$$J_{1,j} = \sum_{t\in\mathcal{T}_j}\left(\mu_t^* - \mu_t^{k_0}\right) \le \Delta\max_{t\in\mathcal{T}_j}\left\{\mu_t^* - \mu_t^{k_0}\right\} \le 2V_j\Delta. \tag{A.11}$$

**Step 3** (Analysis of $J_{2,j}$.). We next bound $J_{2,j}$, the difference between the performance of the single best action in $\mathcal{T}_j$ and that of the policy, throughout $\mathcal{T}_j$. Let $t_j$ denote the first decision index of batch $j$, that is, $t_j = (j-1)\Delta + 1$. We denote by $W_t$ the sum of all weights at decision $t$: $W_t = \sum_{k=1}^{K} w_t^k$. Following the proof of theorem 8.1 in Auer et al. (2002b), one has

$$\frac{W_{t+1}}{W_t} \le 1 + \frac{\gamma/K}{1-\gamma}X_t^{\pi} + \frac{(e-2)(\gamma/K)^2}{1-\gamma}\sum_{k=1}^{K}\hat{X}_t^k + e\alpha. \tag{A.12}$$

Taking logarithms on both sides of (A.12) and summing over all $t \in \mathcal{T}_j$, we get

$$\log\left(\frac{W_{t_{j+1}}}{W_{t_j}}\right) \le \frac{\gamma/K}{1-\gamma}\sum_{t\in\mathcal{T}_j}X_t^{\pi} + \frac{(e-2)(\gamma/K)^2}{1-\gamma}\sum_{t\in\mathcal{T}_j}\sum_{k=1}^{K}\hat{X}_t^k + e\alpha|\mathcal{T}_j|, \tag{A.13}$$

where, for $\mathcal{T}_m$, set $W_{t_{m+1}} = W_T$. Let $k_j$ be the best action in $\mathcal{T}_j$: $k_j \in \arg\max_{k\in\mathcal{K}}\{\sum_{t\in\mathcal{T}_j} X_t^k\}$. Then

$$w_{t_{j+1}}^{k_j} \geq w_{t_j+1}^{k_j} \exp\left\{\frac{\gamma}{K}\sum_{t_j+1}^{t_{j+1}-1}\hat{X}_t^{k_j}\right\} \geq \frac{e\alpha}{K} W_{t_j} \exp\left\{\frac{\gamma}{K}\sum_{t_j+1}^{t_{j+1}-1}\hat{X}_t^{k_j}\right\} \geq \frac{\alpha}{K} W_{t_j} \exp\left\{\frac{\gamma}{K}\sum_{t\in\mathcal{T}_j}\hat{X}_t^{k_j}\right\},$$

where the last inequality holds because $\gamma\hat{X}_t^{k_j}/K \leq 1$. Therefore,

$$\log\left(\frac{W_{t_{j+1}}}{W_{t_j}}\right) \geq \log\left(\frac{w_{t_{j+1}}^{k_j}}{W_{t_j}}\right) \geq \log\left(\frac{\alpha}{K}\right) + \frac{\gamma}{K}\sum_{t\in\mathcal{T}_j} X_t^\pi. \tag{A.14}$$

Taking (A.13) and (A.14) together, one has

$$\sum_{t\in\mathcal{T}_j} X_t^\pi \geq (1-\gamma)\sum_{t\in\mathcal{T}_j}\hat{X}_t^{k_j} - \frac{K\log(K/\alpha)}{\gamma} - (e-2)\frac{\gamma}{K}\sum_{t\in\mathcal{T}_j}\sum_{k=1}^{K}\hat{X}_t^k - \frac{e\alpha K|\mathcal{T}_j|}{\gamma}.$$

Taking expectation with respect to the noisy rewards and the actions of Exp3.S, we have

$$\begin{aligned}
J_{2,j} &= \max_{k\in\mathcal{K}}\left\{\sum_{t\in\mathcal{T}_j}\mu_t^k\right\} - \mathbb{E}^\pi\left[\sum_{t\in\mathcal{T}_j}\mu_t^\pi\right] \\
&\leq \sum_{t\in\mathcal{T}_j}\mu_t^{k_j} + \frac{K\log(K/\alpha)}{\gamma} + (e-2)\frac{\gamma}{K}\sum_{t\in\mathcal{T}_j}\sum_{k=1}^{K}\mu_t^k + \frac{e\alpha K|\mathcal{T}_j|}{\gamma} - (1-\gamma)\sum_{t\in\mathcal{T}_j}\mu_t^{k_j} \\
&= \gamma\sum_{t\in\mathcal{T}_j}\mu_t^{k_j} + \frac{K\log(K/\alpha)}{\gamma} + (e-2)\frac{\gamma}{K}\sum_{t\in\mathcal{T}_j}\sum_{k=1}^{K}\mu_t^k + \frac{e\alpha K|\mathcal{T}_j|}{\gamma} \\
&\overset{(a)}{\leq} (e-1)\gamma|\mathcal{T}_j| + \frac{K\log(K/\alpha)}{\gamma} + \frac{e\alpha K|\mathcal{T}_j|}{\gamma},
\end{aligned} \tag{A.15}$$

for every batch $1 \leq j \leq m$, where $(a)$ holds because $\sum_{t\in\mathcal{T}_j}\mu_t^{k_j} \leq |\mathcal{T}_j|$ and $\sum_{t\in\mathcal{T}_j}\sum_{k=1}^{K}\mu_t^k \leq K|\mathcal{T}_j|$.

**Step 4** (Regret Throughout the Horizon). Taking (A.11) together with (A.15) and summing over $m = \lceil T/\Delta\rceil$ batches, we have

$$\begin{aligned}
\mathcal{R}^\pi(\mathcal{V},T) &\leq \sum_{j=1}^{m}\left((e-1)\gamma|\mathcal{T}_j| + \frac{K\log(K/\alpha)}{\gamma} + \frac{e\alpha K|\mathcal{T}_j|}{\gamma} + 2V_j\Delta\right) \\
&\leq (e-1)\gamma T + \frac{e\alpha KT}{\gamma} + \left(\frac{T}{\Delta}+1\right)\frac{K\log(K/\alpha)}{\gamma} + 2V_T\Delta \\
&\leq (e-1)\gamma T + \frac{e\alpha KT}{\gamma} + \frac{2KT\log(K/\alpha)}{\gamma\Delta} + 2V_T\Delta,
\end{aligned} \tag{A.16}$$

for any $\Delta \in \{1,\ldots,T\}$. Setting the tuning parameters to be $\alpha = \frac{1}{T}$ and $\gamma = \min\{1, \left(\frac{4V_T K\log(KT)}{(e-1)^2 T}\right)^{1/3}\}$, and selecting a batch size $\Delta = \lceil(\frac{e-1}{2})^{1/3}\cdot(K\log(KT))^{1/3}\cdot(\frac{T}{V_T})^{2/3}\rceil$, one has

$$\begin{aligned}
\mathcal{R}^\pi(\mathcal{V},T) &\leq e\cdot\left(\frac{e-1}{2}\right)^{2/3}\frac{K^{2/3}T^{1/3}}{(V_T\log(KT))^{1/3}} + 3\cdot 2^{2/3}(e-1)^{2/3}(KV_T\log(KT))^{1/3}T^{2/3} \\
&\leq \left(e\cdot\left(\frac{e-1}{2}\right)^{2/3} + 4\cdot 2^{2/3}(e-1)^{2/3}\right)\cdot(KV_T\log(KT))^{1/3}T^{2/3},
\end{aligned}$$

where the last inequality holds by recalling $K < T$. Whenever $T$ is unknown, we can use Exp3.S as a subroutine over exponentially increasing pulls epochs $T_\ell = 2^\ell$, $\ell = 0,1,2,\ldots$, in a manner that is similar to the one described in corollary 8.4 in Auer et al. (2002b) to show that because, for any $\ell$, the regret incurred during $T_\ell$ is at most $C(KV_T\log(KT_\ell))^{1/3}\cdot T_\ell^{2/3}$ (by tuning $\alpha$ and $\gamma$ according to $T_\ell$ in each epoch $\ell$), and for some absolute constant $\tilde{C}$, we get that $\mathcal{R}^\pi(\mathcal{V},T) \leq \tilde{C}(\log(KT))^{1/3}\cdot(KV_T)^{1/3}T^{2/3}$. This concludes the proof. □

## Endnotes

[1] Under a non-stationary reward structure, it is immediate that the single best action may be suboptimal in a large number of decision epochs, and the gap between the performance of the static and the dynamic oracles can grow linearly with $T$.

[2] We focus here on $\beta$ values up to 0.5 because for high values of $\beta$, observing the asymptotic regret requires a horizon significantly longer than the $3\cdot 10^8$ that is used here.

## References

Audibert J, Bubeck S (2009) Minimax policies for adversarial and stochastic bandits. *Proc. 22nd Annual Conf. Learn. Theory (COLT), Montreal.*
Auer P, Cesa-Bianchi N, Fischer P (2002a) Finite-time analysis of the multiarmed bandit problem. *Machine Learn.* 47(2–3):235–246.
Auer P, Cesa-Bianchi N, Freund Y, Schapire RE (2002b) The non-stochastic multi-armed bandit problem. *SIAM J. Comput.* 32(1):48–77.
Azar MG, Lazaric A, Brunskill E (2014). Online stochastic optimization under correlated bandit feedback. *Proc. 31st Internat. Conf. Machine Learn., Beijing*.
Bergemann D, Valimaki J (1996) Learning and strategic pricing. *Econometrica: J. Econometric Soc.* 64(5):1125–1149.
Bergemann D, Hege U (2005) The financing of innovation: Learning and stopping. *RAND J. Econom.* 36(4):719–752.
Berry DA, Fristedt B (1985) *Bandit Problems: Sequential Allocation of Experiments* (Chapman and Hall, London).
Bertsimas D, Nino-Mora J (2000) Restless bandits, linear programming relaxations, and primal dual index heuristic. *Oper. Res.* 48(1):80–90.
Besbes O, Gur Y, Zeevi A (2014) Stochastic multi-armed-bandit problem with non-stationary rewards. Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ, eds. *Proc. 27th Internat. Conf. Neural Inform. Processing Systems*, vol. 1 (MIT Press, Cambridge, MA), 199–207.
Besbes O, Gur Y, Zeevi A (2015) Non-stationary stochastic optimization. *Oper. Res.* 63(5):1227–1244.
Blackwell D (1956) An analog of the minimax theorem for vector payoffs. *Pacific J. Math.* 6(1):1–8.
Bubeck S, Cesa-Bianchi N (2012) Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations Trends Machine Learn.* 5(1):1–122.
Cao Y, Zheng W, Kveton B, Xie Y (2019) Nearly optimal adaptive procedure for piecewise-stationary bandit: A change-point detection approach. *22nd Internat. Conf. Artificial Intelligence Statist., Okinawa, Japan*.
Caro F, Gallien J (2007) Dynamic assortment with demand learning for seasonal consumer goods. *Management Sci.* 53(2):276–292.
Cesa-Bianchi N, Lugosi G (2006) *Prediction, Learning, and Games* (Cambridge University Press, Cambridge, UK).
Cheung WC, Simchi-Levi D, Zhu R (2019). Learning to optimize under non-stationarity. *22nd Internat. Conf. Artificial Intelligence Statist., Okinawa, Japan*.
Foster DP, Vohra RV (1999) Regret in the on-line decision problem. *Games Econom. Behav.* 29(1–2):7–35.
Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. System Sci.* 55(1): 119–139.
Garivier A, Moulines E (2011) On upper-confidence bound policies for switching bandit problems. Kivinen J, Szepesvari C, Ukkonen E, Zeugmann T, eds. *Algorithmic learning theory*, Lecture Notes in Computer Science, vol. 6925 (Springer, Berlin, Heidelberg), 174–188.
Gittins JC (1979) Bandit processes and dynamic allocation indices (with discussion). *J. Roy. Statist. Soc. B*. 41(2):148–177.
Gittins JC (1989) *Multi-Armed Bandit Allocation Indices* (John Wiley & Sons, New York).
Gittins JC, Jones DM (1974) A dynamic allocation index for the sequential design of experiments. Gani J, Sarkadi K, Vincze I, eds. *Progress in Statistics* (North-Holland, Amsterdam), 241–266.
Guha S, Munagala K (2007) Approximation algorithms for partial-information based stochastic control with markovian rewards. *48th Annual IEEE Sympos. Foundations Comput. Sci.* (IEEE, Piscataway, NJ), 483–493.
Hannan J (1957) *Approximation to Bayes Risk in Repeated Play, Contributions to the Theory of Games*, vol. 3. (Princeton University Press, Princeton, NJ).
Hazan E, Kale S (2011) Better algorithms for benign bandits. *J. Machine Learn. Res.* 12:1287–1311.
Jadbabaie A, Rakhlin A, Shahrampour S, Sridharan K (2015) Online optimization: Competing with dynamic comparators. Lebanon G, Vishwanathan SVN, eds. *18th Internat. Conf. Artificial Intelligence Statist., San Diego*.
Karnin Z, Anava O (2016) Multi-armed bandits: Competing with optimal sequences. *Proc. Adv. Neural Inform. Processing Systems*, 199–207.
Kleinberg R, Leighton T (2003) The value of knowing a demand curve: Bounds on regret for online posted-price auctions. *Proc. 44th Annual IEEE Sympos. Foundations Comput. Sci.* (IEEE, Piscataway, NJ), 594–605.
Lai TL, Robbins H (1985) Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.* 6(1):4–22.
Levine N, Crammer K, Mannor S (2017) Rotting bandits. *Proc. Adv. Neural Inform. Processing Systems*, 3074–3083.
Luo H, Wei C, Agarwal A, Langford J (2018). Efficient contextual bandits in nonstationary worlds. *31st Conf Learn Theory.*
Ortner R, Ryabko D, Auer P, Munos R (2014) Regret bounds for restless markov bandits. *Theoretical Comp. Sci.* 558(November):62–76.
Pandey S, Agarwal D, Chakrabarti D, Josifovski V (2007) Bandits for taxonomies: A model-based approach. *Proc. 2007 SIAM Internat. Conf. Data Mining, Minneapolis.*
Papadimitriou CH, Tsitsiklis JN (1994). The complexity of optimal queueing network control. *Structure Complexity Theory Conf.*, 318–322.
Robbins H (1952) Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc. (N.S.).* 58(5):527–535.
Slivkins A (2014) Contextual bandits with similarity information. *J. Machine Learn. Res.* 15(1):2533–2568.
Slivkins A, Upfal E (2008). Adapting to a changing environment: The Brownian restless bandits. *Proc. 21st Annual Conf. Learn. Theory*, 343–354.
Thompson WR (1933) On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*. 25(3):285–294.
Wei C, Hong Y, Lu C (2016) Tracking the best expert in non-stationary stochastic environments. *Proc. Adv. Neural Inform. Processing Systems*, 3972–3980.
Whittle P (1981) Arm acquiring bandits. *Ann. Probab.* 9(2):284–292.
Whittle P (1988) Restless bandits: Activity allocation in a changing world. *J. Appl. Probab.* 25A:287–298.
Zelen M (1969) Play the winner rule and the controlled clinical trials. *J. Amer. Statist. Assoc.* 64(325):131–146.
Zhang L, Yang T, Zhou Z (2018). Dynamic regret of strongly adaptive methods. *Proc. 35th Internat. Conf. Machine Learn.*, 5877–5886.