Adapting to a Stochastically Changing Environment: The Dynamic Multi-Armed Bandits Problem

Aleksandrs Slivkins and Eli Upfal

Department of Computer Science
Brown University
Providence, Rhode Island 02912

CS-07-05
May 2007

Adapting to a Stochastically Changing Environment: The Dynamic Multi-Armed Bandits Problem

Aleksandrs Slivkins* Eli Upfal*

April 22, 2007

Abstract

We introduce and study a *dynamic* version of the multi-armed bandit problem. In a multi-armed bandit problem there are k reward distribution functions associated with the rewards of playing each of 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 an algorithm is to maximize reward by balancing the use of acquired information (exploitation) with learning new information (exploration).

In the basic (*static*) multi-armed bandit problem the reward functions are time-invariant. An algorithm can eventually obtain (almost) perfect information of the reward functions and can adapt its strategy so its reward asymptotically converges to that of an optimal strategy. The quality of an algorithm is therefore measured by the expected cost (or *regret*) incurred during an initial finite time interval. In the *dynamic* multi-armed problem the reward functions stochastically and gradually change in time. In this setting a player has to continuously balance explore and exploit steps to adapt to the dynamic reward functions. In analyzing algorithms for the dynamic case we consider an infinite time horizon and the goal here is to minimize the average cost per step, where the cost per step is the difference between the reward of the algorithm and the reward of an (hypothetical) algorithm that at every step plays the arm with the maximum expectation at that step (this definition is stronger than the standard regret function which compares the reward of an algorithm to the reward of the best *time-invariant* strategy). Our goal here is to characterize the steady state cost of learning and adapting to changing reward functions, as a function of the stochastic rate of the change.

The dynamic multi-armed bandit problem models a variety of practical online, game-against- nature type optimization settings. While building on the work done for the static and adversarial versions of the problem, algorithms and analysis for the dynamic setting require a novel approach based on different stochastic tools.

1 Introduction

The multi-armed bandit problem has been studied extensively for over 50 years [18, 5, 9], 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. [1]

In the basic ("static") multi armed bandit setting 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 multi-armed bandit algorithm is to optimize

*Computer Science Department, Brown University, Providence, RI 02912. E-mail: {slivkins, eli}@cs.brown.edu. Supported in part by NSF awards CCR-0121154 and DMI-0600384, and ONR Award N000140610607.

[1] Due to the volume of prior work on the multi-armed bandits, giving a proper introduction into this problem is beyond the scope of this paper; we refer the reader to [12] for a short recent survey. Here we can only mention a few most relevant papers.

the total reward[2] by balancing the use of acquired information (exploitation) with learning new information (exploration). For several algorithms in the literature (e.g. see [5, 2]) as the number of rounds goes to infinity the expected total reward asymptotically approaches that of playing an optimal strategy (a strategy with the highest expected reward). The quality of an algorithm for the static multi-armed bandit 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. (Note that in the static case there is always a *time-invariant* optimal strategy).

The multi-armed bandit 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 [4, 10]). The router must try various routes in order to learn about the delay on that route, and the delay for one packet on a route is a random value drawn from the distribution of the delays on that route. 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 a number of 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 sells. Similarly, a web search engine tries to optimize its revenue by displaying advertisements that would bring the most number of clicks with the associated 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 this experiments is the loss of advertisement clicks when trying unsuccessful matches [17].

The above examples demonstrate the practical applications in online optimization of the "explore and exploit" paradigm captured in the multi-armed bandit 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 changes over time. A store cannot rely on information collected in previous season to optimize the next one. Similarly, a web search engine continually updates their content matching strategies to account for changing customers' response. To address this issue a number of papers [3, 1, 14, 11, 16, 8, 7, 13, 6] remove the assumption that the reward distributions are time-independent, allowing them to change arbitrarily in time. Nevertheless, these papers focus on the standard definition of regret, comparing the total reward of the algorithm with that of the best *time-invariant* policy.[3]

We propose and study here a somewhat different approach to addressing the dynamic nature of online optimization. We note that in a variety of practical applications the optimization parameters, and therefore the reward distributions, change gradually and stochastically in time. (Obvious examples are price, supply and demand in economics, load and delay in networks, etc.). In the simplest and perhaps the most illuminating case, the reward distributions change independently, following similar stochastic processes with possibly different parameters. Our goal here is to characterize the steady state cost of adapting to such changing environment, as a function of the stochastic rate of change. In these settings one can optimize with respect to a more demanding benchmark than the best time invariant policy. In fact, in this setting the standard notion of regret with respect to the best time-invariant policy is meaningless because when the expectations of all reward functions have similar distributions, all strategies are equivalent in the long term. Instead, we use a stronger and more natural benchmark – a policy that at each step plays an arm with the currently maximal expected reward.

We describe the changing reward distribution of a given arm via the evolution of its expected reward. We model the latter (as a function of time) as an ergodic Markov chain. For concreteness we focus on the case in which the expected reward follows a random walk in some bounded discrete interval. (We comment at the end

[2] In this paper the *total reward* is simply the sum of the rewards, following the line of work in [15, 2, 3] and many other papers. Alternatively, many papers consider the *time-discounted* sum of rewards, e.g. see [5, 9, 19] and references therein.

[3] It should be noted that [3, 1] achieve non-trivial guarantees with respect to any policy that changes strategies a limited number of times; however, these guarantees are too weak to be useful in the setting that we consider in this paper.

on extending our work to more general settings.) The actual reward at a given time is an independent random sample from the corresponding reward distribution. The paradigmatic example is a 0-1 random variable with a given expectation. Following the literature, we assume that the actual rewards are bounded. Our analysis holds for any such reward distribution.

We design and analyze algorithms for a number of variants of the *dynamic multi-armed bandit problem* described above. We first consider the case of *deterministic* reward distributions that always equal to their expected value. Since the expected reward follows a random walk, the study of this special case gives us significant insight for the general case, and already develops some of the tools used for the general case. Another distinction we study is between the homogeneous case, in which the expected rewards of all strategies vary at the same rate and in the same interval, and an heterogeneous case in which they vary at different rates in different (overlapping) intervals.

1.1 Basic Definitions

In general, a *multi-armed bandit* is a collection of k stochastic processes $\mathcal{X}_t^{(i)}$, $i \in [k]$, where $\mathcal{X}_t^{(i)}$ is the reward from playing strategy (a.k.a. *arm*) i at time t. A *multi-armed bandit algorithm* is a protocol that looks at the rewards received in the previous steps, and decides which arm to play next.

To simplify notation we assume that all rewards are in the range $[0, 1]$. In a *dynamic multi-armed bandit*, the expected reward of each strategy follows an independent random walk on some subinterval $[a; b] \subset [0, 1]$. Specifically, the state space of such random walk is

$$[a; b]_n := \{a + (b - a)i/n : \ 0 \leq i \leq n\},$$

a one-dimensional grid that partitions the interval $[a; b]$ into n equal-size subintervals. Consider an undirected graph on $[a; b]_n$, with a self-loop at each end-point. A random walk on this graph is called a *random walk on* $[a; b]_n$. Note that such random walk has a uniform stationary distribution. The increment of this random walk is $\pm\frac{b-a}{n}$ and its mixing time is $\Theta(n^2)$, so n is a parameter that governs the rate of change in the interval.

Definition 1.1. Let $X^{(i)}$, $i \in [k]$ be independent random walks on $G_i = [a_i; b_i]_{n_i}$ such that the initial value of each random walk is drawn independently from the corresponding stationary distribution. Let $\{\mathcal{D}_\mu : \ \mu \in [0; 1]\}$ be a family of mean-μ probability distributions on $[0; 1]$. A *dynamic multi-armed bandit* on $\{G_i\}_{i=1}^k$ is a k-armed bandit such that the reward from playing a strategy $i \in [k]$ at time t is an independent random sample from distribution $\mathcal{D}_\mu$, $\mu = X_t^{(i)}$, where $X_t^{(i)}$ is the t-th step of $X^{(i)}$.

A dynamic multi-armed bandit is called *deterministic* if each distribution $\mathcal{D}_\mu$ is deterministically equal to μ. It is called *homogeneous on* $[a; b]_n$ if each G_i is equal to $[a; b]_n$. It is called *heterogeneous* if it is not homogeneous.

We measure the performance of dynamic multi-armed bandits algorithms with respect to an optimal policy that at every step chooses a strategy with the highest expected reward. The optimal policy changes in time, and thus, this measure is a more demanding criterion than the standard, *time-invariant regret* that is often used in the multi-armed bandit literature.

Definition 1.2. Let $\mathcal{B}$ be a dynamic multi-armed bandit. Let $\mathcal{A}$ be a multi-armed bandit algorithm. Let $W_{\mathcal{A},\mathcal{B}}(t)$ be the reward received by algorithm $\mathcal{A}$ at step t. Let $\emptyset$ be an algorithm that at every step chooses a strategy with the highest expected reward. The *strong regret* at step t is

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

The *average strong regret* $\overline{R}_{\mathcal{A},\mathcal{B}}(t_1, t_2)$ is the strong regret averaged over a time interval $[t_1; t_2]$.

We state our theorems in terms of $\overline{R}_{\mathcal{A},\mathcal{B}}(t) := \overline{R}_{\mathcal{A},\mathcal{B}}(0, t)$, but they easily extend to an arbitrary starting time. Since we are interested in the average steady-state performance (and thus can ignore the initial learning period), we assume that the parameters of the problem are known to the algorithm .

1.2 Main Results and Techniques

Throughout this paper, we will prove uniform upper bounds on the expected average strong regret over any sufficiently large interval. Our expected guarantees on the average strong regret easily translate into the corresponding high-probability guarantees. We are mostly interested in the dependence on the range and rate of change parameters. Implicitely, we assume that k is small compared to the n_i's.

A simple algorithm for the deterministic version of the dynamic multi-armed bandit problem is the *greedy algorithm* which partitions time into epochs of (say) m steps each. In each epoch the algorithm first probes every arm once and then plays the best arm for the remaining $m-k$ steps. The trade-off between exploration and exploitation in the greedy algorithm is clear. The epoch length m needs to balance the gain from replacing the leading arm with the cost of trying the k arms. For a homogeneous problem on $[0,1]_n$ we show that the best value of m is close to n, and we prove that for any sufficiently long time interval,

$$E[\overline{R}_{\mathcal{A},\mathcal{B}}] \leq \tilde{O}(\tfrac{k}{n}).$$

For the heterogeneous case with distinct rate-of-change parameters n_i but the same interval $[a;b]$ we obtain a similar guarantee where n is replaced by an average n_{av} given by $n_{\text{av}}^{-2} = \frac{1}{k}\sum_{i\in[k]} n_i^{-2}$. Extending this result to distinct intervals $[a_i;b_i]$ requires a careful weighted averaging of the parameters which expresses the "influence" of the corresponding arms on the maximal expected reward. The result is given in Theorem 2.1.

While the greedy algorithm is only defined for the deterministic case, its analysis and its idea of restarting after a fixed number of steps lead to our first result for the general case. Specifically, we use the general EXP3 algorithm in Auer et al. [3] and restart it every m steps; our analysis combines the result from [3] with the main lemma from the greedy algorithm. For the homogeneous problem on $[0,1]_n$ we obtain

$$E[\overline{R}_{\mathcal{A},\mathcal{B}}] \leq \tilde{O}(\tfrac{k}{n})^{2/3},$$

for any sufficiently large time interval. Theorem 3.1 gives a similar result for the heterogeneous case.

A a stronger result is obtained by building on the UCB1 algorithm in [2] which is designed for a static multi-armed bandit problem. This 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 combines the technique from the deterministic case with that from [2]. It is quite technical because the specific results from [2] do not directly apply to our setting. For the homogeneous problem on $[0,1]_n$ and any sufficiently large time interval we prove

$$E[\overline{R}_{\mathcal{A},\mathcal{B}}] \leq \tilde{O}(\tfrac{k}{n}).$$

Theorem 3.3 extends this result to the heterogeneous case.

Finally, we return to the deterministic homogeneous case and provide an improved algorithm with smarter, *adaptive* exploration: essentially, we do not explore a given arm while it is very unlikely to be better than the arm that we currently exploit. For this algorithm we obtain a guarantee for any sufficiently large time interval that is optimal in terms of n:

$$E[\overline{R}_{\mathcal{A},\mathcal{B}}] \leq \tilde{O}(k^3\, n^{-2}).$$

While the intuition for the analysis is quite simple, the details are very tricky. The analysis is not tight; we conjecture that the right bound is actually $\tilde{O}(k/n^2)$, matching our lower bound (Theorem A.1) up to poly-log factors. The algorithm is well-defined for the heterogeneous case; however, extending the analysis is beyond our current techniques.

1.3 Extensions and Further Work

High probability result: Our guarantees on expected average strong regret easily translate into similar high-probability guarantees. Indeed, suppose we have proved a uniform bound of the form $E[\overline{R}_{\mathcal{A},\mathcal{B}}(t_1,\, t_2)] < C$

for any sufficiently large time interval. For any fixed k, any of our algorithms has a finite state space, and can be easily modified to be aperiodic. The product of the state of the dynamic multi-armed bandit and the state of the (modified) algorithm defines an ergodic, finite state space, Markov chain such that the strong regret at a given time is determined by the current state of this chain. Therefore for any starting time t_1 the limit $\lim_{t\to\infty} \overline{R}_{\mathcal{A},\mathcal{B}}(t_1,\, t_1+t)$ exists and is upper-bounded by C and, moreover, the rate of convergence is independent of t_1. Thus, for any given sufficiently large time interval $[t_1; t_2]$ we have $\overline{R}_{\mathcal{A},\mathcal{B}}(t_1,\, t_2) < C$ with high probability. We conjecture that a time interval of length $\Omega(n^2)$ would be sufficient.

More general stochastic processes: The expectation of the reward functions $X_t^{(i)}$ can follow more general stochastic process than the random walk discussed here. Recall that our random walks move by at most $\sigma = \frac{b-a}{n}$ at each step. While meaningful results on strong regret seem to require some bound on the change in $X_t^{(i)}$ within a short time interval, this bound does not necessarily need to be deterministic as in the random walks analyzed here. We could consider a more general Markov chain. For instance, $X_t^{(i)}$ could be a sum of t independent zero-mean random increments with standard deviation σ. In fact, our results in Section 2 and Section 3 hold as long as the expected reward satisfies the Azuma-type inequality (7). We conjecture that having a uniform bound on the first few moments of the random increments may be sufficient for a meaningful bound on strong regret (although, to the best of our knowledge, such a bound does not imply (7)). The expected rewards, however, need to vary in a bounded interval. Otherwise the expected gaps between the expectations of different reward functions will grow in time making the problem asymptotically trivial.

Further work: We conjecture that the $\tilde{O}(k/n)$ result for the general case is in fact optimal, and the corresponding lower bound can be proved by adapting the relative entropy technique from [3]. Also, while our result in Section 4 is optimal (up to poly-log factors) in terms of the rate-of-change parameter n, our analysis is clearly not tight in terms of k. We conjecture the dependence on k can be improved. Also, we think the result can be extended to the heterogeneous case.

2 Deterministic Case: the Greedy Algorithm

We consider the following setting. The main object is a dynamic k-armed bandit $\mathcal{B}$ on

$$\{[a_i; b_i]_{n_i}\}_{i=1}^k \text{ such that } (\max a_i < \min b_i) \text{ and } (\min n_i \geq 3). \tag{1}$$

(The first assumption in (1) is w.l.o.g. since otherwise we can always ignore the arm with the smallest b_i.)

For convenience, we let $\epsilon_i = (b_i - a_i)/n_i$ be the step size for each arm i, and assume (w.l.o.g.) that $b_1 \leq \ldots \leq b_k$. We will use the average rate-of-change parameter n_{av} defined by

$$n_{\text{av}}^{-2} = \frac{\sum_{i\in[k]} w_i(b_i-a_i)\, n_i^{-2}}{\sum_{i\in[k]} w_i(b_i-a_i)} \quad \text{where } w_i = \prod_{l=i}^{k} \frac{b_i - a_l + \epsilon_l}{b_l - a_l + \epsilon_l}, \tag{2}$$

Here the weight w_i express the "importance" of the corresponding arm i: arms with larger maximal values are more important. Define the average interval length

$$d_{\text{av}} = \tfrac{1}{k} \textstyle\sum_{i\in[k]} w_i(b_i - a_i). \tag{3}$$

Note that for a homogenous dynamic bandit on $[a; b]_n$ we have $w_i = 1$, so $n_{\text{av}} = n$ and $d_{\text{av}} = b - a$.

Theorem 2.1. *Consider a deterministic dynamic bandit $\mathcal{B}$ in the setting of (1-3) and let $n = n_{av}$. Let $\mathcal{A}$ be the m-greedy algorithm, where $m = n\sqrt{(b_k - a_{min})/\log n}$ and $a_{min} = \min a_i$. Then*

$$E[\overline{R}_{\mathcal{A},\mathcal{B}}(t)] \leq O(k\, n^{-1})\, \sqrt{(b_k - a_{min})\, d_{av} \log n} \quad \textit{for any time } t \geq m. \tag{4}$$

Note that for the homogenous case on $[a;b]_n$, the right-hand side of (4) is simply $(b-a)\frac{k}{n}\sqrt{\log n}$.

For comparison, we provide an easy lower bound: for any algorithm $\mathcal{A}$ and any time t we have

$$E\left[R_{\mathcal{A},\mathcal{B}}(t)\right] \geq \Omega(k\,n^{-2}) \tag{5}$$

The idea is very simple. If at time t an algorithm knows the (expected) rewards of all arms at time $t-1$, then it chooses to play the arm with the best reward. However, there are two or more such arms, then in expectation the algorithm incurs some regret, see the proof of Theorem A.1 for the easy details. This bound seems very weak; surprisingly, in Section 4 we essentially match it in terms of the dependence on n.

The crux of the proof of Theorem 2.1 is captured by the following lemma:

Lemma 2.2. *Consider the setting (1-3). Let μ_i, $i \in [k]$ be independent random variables, each distributed uniformly on $S_i = [a_i;b_i]_{n_i}$. Let μ^* be their maximum, and let $i^* \in \operatorname{argmax}\mu_i$, ties broken arbitrarily. Let $\{Y_t\}_{t=0}^{\infty}$ be an independent random walk on S_{i^*} such that $Y_0 = \mu^*$. Then for any time t and any $m \geq t$*

$$E[\mu^* - Y_t] \leq O(k\,d_{av})(m^{-4} + n_{av}^{-2}\,m\log m). \tag{6}$$

Remark. In expectation, Y_t drifts down from its initial value μ^* due to the proximity of the upper boundary. This lemma bounds such drift. We will also use it in Section 3 for the analysis of the general case.

Proof of Theorem 2.1: Fix any $m > k$ and consider a single epoch of the m-greedy algorithm. Let $\overline{W}$ be the average winnings in this epoch. Assume without loss of generality that in the first phase (*exploration*) our algorithm plays arm i in step i. Let $\mu_i = X_i^{(i)}$ be the corresponding rewards, and let μ^* be the largest of them. Note that the winnings from this phase are at least $k\,a_{\min}$.

Recall that for the second phase (*exploitation*) the algorithm chooses arm $i^* \in \operatorname{argmax}_{i\in[k]}\mu_i$. The winnings from the second phase are $W_2 := \sum_{t=1}^{m-k} Y_t$, where $Y_t = X_{k+t}^{(i^*)}$. For convenience, let z be the right-hand side of (6). By Lemma 2.2 we have $E[Y_t] \geq E[\mu^*] - z$, so

$$\begin{aligned}
E[\overline{W}] &\geq \tfrac{k}{m}\,a_{\min} + \tfrac{1}{m}E[W_2] \geq \tfrac{k}{m}\,a_{\min} + \tfrac{m-k}{m}\,(E[\mu^*] - z)\\
E[\mu^* - \overline{W}] &\leq z + \tfrac{k}{m}\,E[\mu^* - a_{\min}] \leq O(km\log m)(d_{\text{av}}\,n^{-2}) + \tfrac{k}{m}(1+\tfrac{1}{m})\,(b_k - a_{\min})\\
&= O(k\,n^{-1})\,\sqrt{(b_k - a_{\min})\,d_{\text{av}}\log n} \quad \text{for } m = n\,\sqrt{(b_k - a_{\min})/\log n}.
\end{aligned}$$

Finally, note that $E[\overline{R}_{\mathcal{A},\mathcal{B}}(t)] \leq 2\,E[\mu^* - \overline{W}]$ for any $t \geq m$. □

In the rest of this section we prove the following lemma which implies Lemma 2.2:

Lemma 2.3. *In the setting of Lemma 2.2, for every $m \in \mathbb{N}$ there exists a random variable Z_m such that:*

(a) $E[Z_m - Y_t] \leq m^{-3}k\,d_{av}$ *for each $t \in [m]$.*

(b) $E[\mu^* - Z_m] \leq O(km\log m)\,(d_{av}\,n_{av}^{-2})$.

We start with a simple but very useful fact about random walks on $[a;b]_n$. Essentially, we want to argue that in m steps the expected value of such random walk is (almost) as large as the initial value. The proof follows from the Azuma's inequality; we defer it to Appendix A.

Claim 2.4. *Let $\{Y_t\}_{t=0}^{\infty}$ be a random walk on $[a;b]_n$ such that $Y_0 = \mu$. Fix $m \in \mathbb{N}$ and $t \in [m]$. Then:*

$$\Pr[\,|Y_t - \mu| > \delta] < m^{-3}, \quad \text{where } \delta = n^{-1}(b-a)\sqrt{6m\log m}. \tag{7}$$

$$E[Y_t] \geq \min(\mu,\, b-\delta) - (b-a)\,m^{-2} \tag{8}$$

First we prove Lemma 2.3 for a much cleaner homogenous case on $[a;b]_n$. Define $Z_m = \min(\mu^*, b-\delta)$, where δ is from (7). Then part (a) of the lemma follows immediately from (8). For part (b) it suffices to show that $E[\mu^* - Z_m] \leq \frac{1}{b-a}\, k\delta^2/2$, which we prove by evaluating the corresponding integral (see Lemma A.2).

For the general case of Lemma 2.3, fix $m \in \mathbb{N}$, define

$$\rho_i := \min(\mu_i;\ b_i - \delta_i), \quad \text{where } \delta_i = (b_i - a_i)\, n_i^{-1} \sqrt{6m \log m} \tag{9}$$

and $Z_m = \rho_{i^*}$. We tame the heterogeneity by claiming the following (see Claim A.3 in Appendix A):

$$\begin{aligned} E[b_{i^*} - a_{i^*}] &\leq k\, d_{\text{av}} && (10)\\ E\left[1_{\{\mu_i = \mu^*\}}\, (\mu^* - \rho_i)\right] &\leq O(w_i\, \delta_i^2)/(b_i - a_i). && (11) \end{aligned}$$

As in the homogenous case, Lemma 2.3(a) follows from (8), but here we also use (10)

$$\begin{aligned} E[Y_t|\, i^*, \mu^*] &\geq Z_m - (b_{i^*} - a_{i^*})\, m^{-2} \\ E[Y_t] &\geq E[Z_m] - E[b_{i^*} - a_{i^*}]\, m^{-2} \geq E[Z_m] - m^{-2}\, k\, d_{\text{av}}. \end{aligned}$$

For Lemma 2.3(b), we use (10) and plug in the definitions (9) and (2-3) to obtain

$$E[\mu^* - Z_m] \leq \sum_{i\in[k]} E\left[1_{\{\mu_i=\mu^*\}}\, (\mu^* - \rho_i)\right] \leq \sum_{i\in[k]} \frac{O(w_i\, \delta_i^2)}{b_i - a_i} = O(km \log m)(d_{\text{av}}\, n_{\text{av}}^{-2}). \tag{12}$$

3 The General Dynamic Multi-Armed Bandits

Our first algorithm for general dynamic bandits[4] combines the technique from Section 2 with a very general result from Auer et al. [3]. For simplicity, here we only state this result for the setting of dynamic bandits; in what follows let $\overline{W}_{\mathcal{A},\mathcal{B}}(t)$ be the average reward collected by algorithm $\mathcal{A}$ during the time interval $[1;t]$.

Theorem 3.1 ([3]). *Consider a k-armed dynamic bandit $\mathcal{B}$. Let $\mathcal{A}_i$ be an algorithm that plays arm i at every step. Then there is an algorithm* EXP3 *such that for any arm i and any time $t \in \mathbb{N}$*

$$E[\overline{W}_{\text{EXP3},\mathcal{B}}(t)] \geq E(\overline{W}_{\mathcal{A}_i,\mathcal{B}}(t)) - O(\tfrac{k}{t}\log t)^{1/2}.$$

Theorem 3.2. *Consider a k-armed dynamic bandit $\mathcal{B}$ in the setting of (1-3). Then there is an algorithm $\mathcal{A}$ and time $m \in \mathbb{N}$ such that for any time $t \geq m$ we have*

$$E[\overline{R}_{\mathcal{A},\mathcal{B}}(t)] \leq O(d_{\text{av}}^{1/3})\, (n_{\text{av}}^{-1}\, k \log n_{\text{av}})^{2/3}.$$

Proof. Fix $m \in \mathbb{N}$. We use algorithm EXP3 from Theorem 3.1 which we restart every m steps; let $\mathcal{A}$ be the resulting algorithm. 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 $Y_t = X_t^{(i^*)}$. Then by Lemma 2.2 we have $E[Y_t] \geq E[\mu^*] - z_m$, where z_m is the right-hand side of (6). Therefore:

$$\begin{aligned} E[\overline{W}_{\mathcal{A}^*,\mathcal{B}}(m)] &= E\left[\, E[\overline{W}_{\mathcal{A}^*,\mathcal{B}}(m)\,|\, Y_1,\ \ldots\ , Y_m]\,\right] \\ &= \tfrac{1}{m}\, E[\textstyle\sum_{t=1}^m Y_t] \geq \mu^* - z_m \\ E[\overline{R}_{\mathcal{A},\mathcal{B}}(m)] &= E\left[\mu^* - \overline{W}_{\mathcal{A}^*,\mathcal{B}}(m)\right] + E\left[\overline{W}_{\mathcal{A}^*,\mathcal{B}}(m) - \overline{W}_{\mathcal{A},\mathcal{B}}(m)\right] \\ &\leq z_m + O(\tfrac{k}{m}\log m)^{1/2}. \end{aligned} \tag{13}$$

We used Theorem 3.1 to obtain (13). The theorem follows by minimizing (13) with respect to m. □

[4] We note in passing that we can also get non-trivial guarantees by adapting the greedy algorithm (probe every arm a few times in the exploration phase) and analyzing it using Lemma 2.2. However, the guarantees are much weaker than the ones in Theorem 3.2.

We proceed to the stronger result where we replace the algorithm from [3] with a new algorithm tailored to dynamic bandits, which is an extension of the multi-armed bandit algorithm from [2]. For the proof, we combine the technique from Section 2 with that from [2]; the argument is quite technical because the specific results from [2] do not directly apply to our setting. We consider k-armed dynamic bandit on

$$\{[a;b]_{n_i}\}_{i=1}^k \text{ with an average given by } n_{\text{av}}^{-2} = \tfrac{1}{k}\sum_{i\in[k]} n_i^{-2}. \tag{14}$$

Note that the average n_{av} here coincides with the one defined as in (2).

Theorem 3.3. *Consider a k-armed dynamic bandit $\mathcal{B}$ in the setting of (14). Then there is an algorithm $\mathcal{A}$ such that for any time $t \geq n_{av}\sqrt{\log n_{av}}$ we have*

$$E[\overline{R}_{\mathcal{A},\mathcal{B}}(t)] \leq O(k\, n_{av}^{-1})(\log^{3/2} n_{av})\,(1+2^{-k}\log n_{av}). \tag{15}$$

We define the algorithm using the following notation. For a given multi-armed bandit and a given algorithm, let $T_i(t)$ be the number of times arm i has been played in the first t steps, and let $\bar{X}_i(t_i)$ be the average reward from playing arm i the first t_i times. Define $\bar{X}_i(0) = 0$.

Definition 3.4 (padding). A *padding* is a function $f : [k]\times\mathbb{N}\times\mathbb{N}\to\mathbb{R}_+$; such that for each $i\in[k]$ the function $f_i(t,t_i) := f(i,t,t_i)$ is increasing in t and decreasing in t_i. It is called a *$\mathcal{B}$-padding* for a given multi-armed bandit $\mathcal{B}$ if for each arm $i\in[k]$ there exists a number μ_i such that for any algorithm we have

$$\Pr\left[\,|\bar{X}_i(t_i)-\mu_i| > f_i(t,\,t_i)\,\right] < O(t^{-4}) \text{ for any time } t\in\mathbb{N} \text{ and any } t_i < t. \tag{16}$$

For any padding f we define the following simple algorithm, which we call UCB_f:

Algorithm 3.5 (UCB_f). *On each step t, play any arm $i\in\operatorname{argmax}[\bar{X}_i(t_i)+f_i(t,\,t_i)]$, where $t_i = T_i(t)$.*

Claim 3.6 (Auer et al. [2]). *Consider a multi-armed bandit $\mathcal{B}$ and algorithm UCB_f such that f is a $\mathcal{B}$-padding. Let $\mu^* = \max\mu_i$ where the μ_i's are from (16). Then for each arm i we have*

$$f_i(m,t_i) \leq \tfrac{1}{2}(\mu^*-\mu_i) \;\Rightarrow\; E[T_i(m)] \leq t_i + O(1). \tag{17}$$

This claim is implicit in [2], where it is the crux of the main proof. That proof considers a specific padding f, and in fact does not explicify the notion of padding. Here we define padding in order to extend the idea from [2] to our setting.

Consider a dynamic bandit $\mathcal{B}$ in the setting of (14). We will use algorithm UCB_f where f is defined by

$$f_i(t,t_i) = \sqrt{2\ln(t)/t_i} + \delta_i(t), \;\text{ where }\; \delta_i(t) = (b-a)\,n_i^{-1}\sqrt{8t\log t}. \tag{18}$$

To show that this is a $\mathcal{B}$-padding, we need to prove (16). The first summand in (18) is tuned for an application of Chernoff-Hoeffding Bounds, whereas the second one corrects for the drift; see Claim B.1 for details.

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

$$\left\{\begin{array}{rcll} \mu_i &=& X_0^{(i)} & \text{for each } i\in[k] \\ S(m) &=& \{i\in[k] : \mu^*-\mu_i \geq 4\delta_i(m)\}, & \text{where } \mu^* = \max_{i\in[k]}\mu_i \\ \rho_i(m) &=& \min(\mu_i,\, b-\delta_i(m)) & \end{array}\right. \tag{19}$$

We will write S and ρ_i when the corresponding m is clear.

Lemma 3.7. *Consider a k-armed dynamic bandit $\mathcal{B}$ and algorithm UCB_f in the setting of (14) and (18-19). Fix time $m\geq k$ and let $T_i = T_i(m)$. Then conditioning on $\mu_1\ldots\mu_k$ we have*

$$E\left[T_i\,\bar{X}_i(T_i)\right] \;\geq\; \rho_i\,E[T_i] - m^{-2}(b-a) \qquad \text{for each arm } i\in[k] \tag{20}$$

$$\mu^* - E[\overline{W}_{\text{UCB}_f}(m)] \;\leq\; \frac{k(b-a)}{m^2} + O(1)\left[\sum_{i\notin S}\mu^*-\rho_i\right] + O(\tfrac{1}{m}\log m)\left[\sum_{i\in S}\frac{1}{\mu^*-\mu_i}\right]. \tag{21}$$

Integrating both sides of (21) with respect to $\mu_1 \dots \mu_k$ and letting $\gamma = 2^{-k} \log n_{av}$ we obtain

$$E[\overline{R}_{\text{UCB}_f}](m) \quad \leq \quad O(k \log m)(1+\gamma) \left[m(b-a)(m^{-3} + n_{av}^{-2}) + \tfrac{1}{m(b-a)} \log(n_{av}) \right]. \tag{22}$$

Proof Sketch. The left-hand side of (20) is just the total winnings collected by arm i up to time m. If the bandit algorithm always plays arm i, then $T_i(m) = m$ and the left-hand side of (20) is simply equal to $\sum_{t \in [m]} E[X_t^{(i)}]$, so the right-hand side follows from Claim 2.4(b). Therefore one can view (20) as an extension of that claim. While the intuition from Claim 2.4(b) is relatively simple, the actual proof of (20) is a rather intricate exercise in conditional expectations and martingales. We note in passing that (20) in fact holds for any bandit algorithm, not just UCB_f.

Denote $\bar{X}_i = \bar{X}_i(T_i)$ and $\Delta_i = \mu^* - \mu_i$ for each arm i. Conditioning on $\mu_1 \dots \mu_k$ and using (20) we write

$$m\mu^* - E[W_{\text{UCB}_f}(m)] = \textstyle\sum_{i \in [k]} E[(\mu^* - \bar{X}_i) T_i] \leq \sum_{i \in [k]} E[T_i] \, (\mu^* - \rho_i) + m^{-2} \, (b-a).$$

We obtain (21) since for each $i \in S(m)$ we have $\mu^* - \rho_i \leq 2\Delta_i$ and (by Claim 3.6)

$$E[T_i] \leq t_i + O(1) \text{ where } t_i = 32 \, \ln(m)/\Delta_i^2.$$

To obtain (22), we need to integrate the right-hand side of (21) with respect to $\mu_1 \dots \mu_k$. We will take this expectation conditional on μ^*. We partition the arms into three sets and bound the three corresponding sums separately. One of these three sets is S. The other two, denoted S^* and S^+, are the sets of all arms $i \in [k]$ such that, respectively, $\Delta_i = 0$ and $0 < \Delta_i < 4\delta_i$. We show that

$$\begin{aligned}
E\left[\textstyle\sum_{i \in S} \Delta_i^{-1} \,|\, \mu^*\right] &\leq \alpha(\mu^*) \times O(k \log n_{\text{av}}) \, (b-a)^{-1} \\
E\left[\textstyle\sum_{i \in S^+} \mu^* - \rho_i \,|\, \mu^*\right] &\leq \alpha(\mu^*) \times O(k \, n_{\text{av}}^{-2} \, m \log m)(b-a)^{-1} \\
E\left[\textstyle\sum_{i \in S^*} \mu^* - \rho_i\right] &\leq O(b-a)(k \, n_{\text{av}}^{-2})(m \log m).
\end{aligned}$$

where $\alpha(\mu^*) = (b-a)/(\mu^* - a)$. The third inequality is yet another case when we invoke Lemma 2.3(b). Finally, to obtain (22) we show that $E[\alpha(\mu^*)] \leq O(1+\gamma)$. □

To obtain Theorem 3.3 from (22), restart algorithm UCB_f after every $m = n_{\text{av}} \, (b-a)^{-1} \sqrt{\log n_{\text{av}}}$ steps.

4 Deterministic Case: Adaptive Exploration

We return to the deterministic dynamic bandits and provide an improved algorithm that matches the lower bound (5) in terms of n. We focus on the homogenous case on $[0;1]_n$.

Definition 4.1 (notation for the algorithm). At each time t some arm $L(t)$ is called the *leader*. Suppose before a given time t each arm i was last played t_i steps ago and at that time its reward was y_i. Let $y^*(t) = y_{L(t)}$ be the reward returned by the leader the last time it was played. Arm i is called *suspicious* at time t if $y^*(t) - y_i \leq c_{\mathcal{A}} \, n^{-1} \sqrt{t_i}$ for some $c_{\mathcal{A}} = \Theta(\log n)$. When an arm becomes suspicious, it is called *active* until it is played. Arm i is called *max-active* at time t if its t_i is the largest among the active arms.

The idea is that if an arm i is not suspicious at time t, then we can prove that with high probability its current reward is less than $y^*(t)$. The intuition behind our algorithm is that if no arm is suspicious then the best bet is to play the current leader.

Algorithm 4.2 (adaptive exploration). *Consider time slot t and assume that the current leader $L(t)$ is chosen. Set $L(t+1) \leftarrow L(t)$. If t is even, play the current leader $L(t)$. If t is odd, play the max-active arm, call it i. If $X_i(t) > y^*(t)$, change the leader: $L(t+1) \leftarrow i$. For bootstrapping, in the first k steps each arm is played once; for $t = k+1$ the leader is any arm that returned the maximal reward.*

It is crucial that if an arm becomes suspicious, it is played within the next $2k$ time steps. Note that there always is at least one active arm, namely the current leader.

Theorem 4.3. *Let $\mathcal{B}$ be a deterministic dynamic bandit on $[0;1]_n$. Let $\mathcal{A}$ be Algorithm 4.2. Then*

$$E[\overline{R}_{\mathcal{A},\mathcal{B}}(t)] \leq O(k\,n^{-2})(k^2+\log^2 n) \text{ for any } t > n^2.$$

In the remainder of this section we will outline the proof of Theorem 4.3. Proofs of all lemmas are moved to Appendix C. We will denote the reward of arm i at time t by $X_i(t)$ instead of $X_t^{(i)}$.

Let $X^*(t) = \max X_i(t)$ be the maximal reward at time t, and let $Y^*(t) = X_{L(t)}(t)$ be the current leader's reward. Let us say that a real-valued function f is *well-behaved* on an interval $[t_1;t_2]$ if

$$|f(t)-f(t')| < c_f\, n^{-1}\,\sqrt{t-t'} \quad \text{whenever } t_1 \leq t' \leq t \leq t_2., \tag{23}$$

where $c_f = \Theta(\log n)$ will be set later. The bandit $\mathcal{B}$ is *well-behaved* on an interval if all reward functions X_i are well-behaved. Say that the bandit is *well-behaved around time* t if it is well-behaved on $[t-3n^2;\, t+n^2]$. Choose c_f so that for any fixed t the probability of this event is at least $1-n^{-3}$. We define the constant $c_{\mathcal{A}}$ in Definition 4.1 to be $c_{\mathcal{A}} = 5\,c_f$.

It is easy to see that if the bandit is well-behaved on an interval then the maximal reward X^* is well-behaved, too. Ideally, we want the leader's reward Y^* to be well-behaved and very close to X^*. In fact:

Lemma 4.4. *Suppose the bandit is well-behaved on the time interval $[t_1-n^2-2k;\, t_2]$. Then for any time $t \in [t_1;t_2]$ we have $X^*(t)-Y^*(t) \leq O(k/n)$.*

Remark. Our analysis is not tight. We conjecture that the right bound for this lemma is $O(1/n)$. Once such bound is established, it is possible to modify the algorithm and analysis to improve the bound in Theorem 4.3 to $O(kn^{-2}\log n)$. We omit the details from this version.

For a given time t and arm i, let $T_i(t)$ be the set of times $t' \leq t$ when this arm was played. Let us split the average regret into that of every arm i with respect to the leader (denoted $\overline{R}_i$), and that of the leader with respect to the true maximum X^* (denoted $\overline{R}^*$):

$$\overline{R}_{\mathcal{A},\mathcal{B}}(t) = \overline{R}^*(t) + \sum_{i\in[k]} \overline{R}_i(t), \text{ where } \left\{ \begin{array}{rcll} \overline{R}_i(t) &=& \frac{1}{t}\sum_{\tau\in T_i(t)} & Y^*(\tau)-X_i(\tau) \\ \overline{R}^*(t) &=& \frac{1}{t}\sum_{\tau\in[t]} & X^*(\tau)-Y^*(\tau) \end{array} \right.$$

We will bound the expectations of $\overline{R}^*$ and each $\overline{R}_i$ separately. First we use Lemma 4.4 to bound $\overline{R}_i$.

Lemma 4.5. *For any given t we have $E[X^*(t)-Y^*(t)] \leq O(k)(k/n)^2$.*

Proof. Fix time t. If the bandit is well-behaved around time t, then by Lemma 4.4 either $X^*(t) = Y^*(t)$, or the second largest reward $X_i(t)$ is within at most $O(k/n)$ from $X^*(t)$. Denote the latter event by F. It is easy to see that $\Pr[F] \leq O(k/n^2)$. For convenience, denote $Z^* = X^*(t)-Y^*(t)$. Then

$$E[Z^*] \leq E[Z^*|E_t] + \Pr[\bar{E}_t] \leq E[Z^*|E_t\cap F]\,\Pr[F|E_t] + \Pr[\bar{E}_t] \leq O(k^3/n^2). \quad \square$$

To upper-bound the expectation of $\overline{R}_i$ for a fixed arm i, we spread this regret over all time slots as follows. Let $\tau_i(t)$ be the last time $t' \leq t$ when arm i was played before time t. Let $\sigma_i(t)$ be the number of time steps between $\tau(t)$ and the next time arm i is played. Define the *contribution* of the time slot t as

$$c_i(t) = \tfrac{Z_i(\tau_i(t))}{\sigma_i(t)}, \text{ where } Z_i(\tau) = Y^*(\tau)-X_i(\tau).$$

Then the regret w.r.t. the leader is just the mean of these contributions: $\overline{R}_i(t) = \frac{1}{t}\sum_{t'\in[t]} c_i(t')$.

We consider each $E[c_i(t)]$ separately. Letting $\epsilon = \frac{1}{n}$, we claim that for any time t we have

$$E[\, c_i(t) \mid Z_i(t) = \delta \text{ and } X_i(t) < X^*(t)] = \begin{cases} O(\epsilon^2/\delta)(\log n) & \text{if } \delta \geq \Theta(k\epsilon), \\ O(k\epsilon) & \text{otherwise.} \end{cases} \quad (24)$$

$$\Pr[\, Z_i(t) < \delta \mid X_i(t) < X^*(t)] \leq O(\delta + k\epsilon) \quad (25)$$

$$E[\, c_i(t) \mid X_i(t) = X^*(t)] \leq O(k\epsilon)^2 \quad (26)$$

$$E[c_i(t)] \leq O(\epsilon^2)(k^2 + \log n) \quad (27)$$

The proof of (24-27) is very detailed, see Appendix C. For (24), the intuition is that function $Z_i(t)$ is not likely to change much from the last time arm i was played, so $Z_i(\tau_i(t)) = \Theta(\delta)$ and consequently arm i will not be played for at least $\Omega(\delta/\epsilon)^2$ rounds. The most troublesome case is $\delta \sim k\epsilon$ when this intuition falters.

The reason we condition on $X_i(t) < X^*(t)$ is to obtain a clean bound (25). For the exception case $X_i(t) = X^*(t)$ we obtain (26) as follows. First we show that either $c_i(t) = 0$ or we have both $c_i(t) \leq 2\epsilon$ and $i \neq L(t)$. Then by Lemma 4.4 some arm (namely, the leader) is very close to the maximal arm i, which is very unlikely. Finally, we put together (24-26) and obtain (27). This completes the proof of Theorem 4.3.

References

[1] P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *J. Machine Learning Research*, 3:397–422, 2002. Preliminary version in *41st IEEE FOCS*, 2000.

[2] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002. Preliminary version in *15th ICML*, 1998.

[3] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002. Preliminary version in *36th IEEE FOCS*, 1995.

[4] B. Awerbuch and R. D. Kleinberg. Adaptive routing with end-to-end feedback: distributed learning and geometric approaches. In *36th ACM Symp. on Theory of Computing (STOC)*, pages 45–53, 2004.

[5] D. A. Berry and B. Fristedt. *Bandit problems: sequential allocation of experiments*. Chapman and Hall, 1985.

[6] V. Dani and T. P. Hayes. How to beat the adaptive multi-armed bandit. Technical report. Available from arXiv at `http://arxiv.org/cs.DS/0602053`, 2006.

[7] V. Dani and T. 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, 2006.

[8] A. Flaxman, A. Kalai, and H. B. 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.

[9] J. C. Gittins. *Multi-Armed Bandit Allocation Indices*. John Wiley & Sons, 1989.

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

[11] R. D. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *18th Advances in Neural Information Processing Systems (NIPS)*, 2004. Full version appeared as Chapters 4-5 in [12].

[12] R. D. Kleinberg. *Online Decision Problems with Large Strategy Sets*. PhD thesis, MIT, Boston, MA, 2005.

[13] R. D. Kleinberg. Anytime algorithms for multi-armed bandit problems. In *17th ACM-SIAM Symp. on Discrete Algorithms (SODA)*, pages 928–936, 2006. Full version appeared as Chapter 6 in [12].

[14] R. D. Kleinberg and F. T. Leighton. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *44th Symp. on Foundations of Computer Science (FOCS)*, pages 594–605, 2003.

[15] T. Lai and H. Robbins. Asymptotically efficient Adaptive Allocation Rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

[16] H. B. McMahan and A. Blum. Online Geometric Optimization in the Bandit Setting Against an Adaptive Adversary. In *17th Conference on Learning Theory (COLT)*, pages 109–123, 2004.

[17] S. Pandey, D. Agarwal, D. Chakrabarti, and V. Josifovski. Bandits for Taxonomies: A Model-based Approach. In *SIAM Intl. Conf. on Data Mining (SDM)*, 2007.

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

[19] R. K. Sundaram. Generalized Bandit Problems. In D. Austen-Smith and J. 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.

A Proofs from Section 2 (the greedy algorithm)

We start with a simple lower bound (5) on strong regret. We state it for the homogenous case, but it easily extends to the heterogenous setting of (14).

Theorem A.1. *Consider a k-armed RW-bandit $\mathcal{B}$ on $[0;1]_n$. Then for any algorithm $\mathcal{A}$ and any time t*

$$E\left[R_{\mathcal{A},\mathcal{B}}(t)\right] \geq \Omega(k\,n^{-2}).$$

Proof. Let $\epsilon = 1/n$. If at time t an algorithm knows the (expected) rewards of all arms at time $t-1$, then it chooses to play the arm with the best reward. However, there are two or more such arms, then in expectation the algorithm incurs regret $\Omega(\epsilon)$.

We claim that the probability of such a "tie" is $\Omega(\epsilon k)$. Indeed, let A_{ijx} be the event when $X_0^{(i)} = X_0^{(j)} = x$ and all other arms are less than x. These events are disjoint, and $\Pr[A_{ijx}] = \epsilon^2(\frac{x\,n}{n+1})^{k-2}$. Therefore the near-tie happens with probability at least

$$\textstyle\sum_{ij}\sum_x \Pr[A_{ijx}] \geq \sum_{ij}\Omega(\epsilon)\int_0^1 x^{k-2}dx \geq \sum_{ij}\Omega(\epsilon/k) = \Omega(\epsilon k). \quad \square$$

Proof of Claim 2.4: Let us couple the bounded random walk in question with the corresponding *unbounded* random walk. Specifically, let $\{\widehat{Y}_t\}_{t=0}^{\infty}$ be the unbounded symmetric random walk with step $\epsilon = \frac{b-a}{n}$ and $\widehat{Y}_0 = \mu$ such that $Y_t = f(\widehat{Y}_t)$ for some deterministic function f (we omit the details). Then $|Y_t - \mu| > \delta$ only if $|\widehat{Y}_t - \mu| > \delta$. We can apply the well-known Azuma's inequality:

$$\Pr[\,|\widehat{Y}_t - \mu| > c\,\epsilon\,\sqrt{t}] < 2e^{-c^2/2}, \quad \text{for any } c > 1.$$

We obtain (7) setting $c = \sqrt{6\log m}$.

Now let us prove (8). Let $w = \frac{b-a}{2}$ and note that if $\mu < w$ then $E[Y_t] > \mu$. Also, note that the function $f(\mu) = E[Y_t|\mu]$ is increasing and $f(w) = w$ by symmetry. Therefore, it suffices to prove (8) under the assumption that $w < \mu \leq b - \delta$.

Let T^* be the first time t such that $Y_t \in \{a, b\}$. Fix $t' \in [m]$ and consider $T = \min(t', T^*)$. Then $Z_t = Y_{\min(t,T)}$ is a martingale for which T is a bounded stopping time, so by the Optional Stopping Theorem we have $Z_T = Z_0 = \mu$. Finally, note that by (7) we have $T^* \geq m$ with probability at least $1 - m^{-2}$, in which case $T = t'$ and $\mu = Z_T = Y_{t'}$. $\square$

Lemma A.2. *Let μ^* be the maximum of k independent random variables distributed uniformly on $[a;b]_n$, where $k \leq n/2$. Then:*

(a) $E[\mu^*] = b - (b-a)(\frac{1}{k+1} + r)$, *where* $-\frac{1}{n} < r < \frac{1}{n}\,\frac{1}{k+1}$.

(b) $E[\mu^* - Z_\delta] \leq \frac{1}{b-a}\,k\delta^2/2$, *where* $\delta > \frac{b-a}{n}$ *and* $Z_\delta = \min(\mu^*, b-\delta)$.

Proof. For better typesetting, for this proof we rename $Z = \mu^*$.

For simplicity, let us first prove the lemma for the special case $[a;b] = [0;1]$, and then extend to the general case by scaling. Let $\epsilon = \frac{1}{n}$, and denote $F(x) = (1+\epsilon)^k\, x^k$ and $I_\Delta(x) = \int_x^{x+\Delta} F(x)dx$. Then:

$$\begin{aligned}
\Pr[Z < i\epsilon] &= (\tfrac{i}{n+1})^k = (\tfrac{n}{n+1})^k\,(i\epsilon)^k = F(i\epsilon), \quad \text{for each } i \in [n] \\
1 - E[Z] &= 1 - \sum_{i=1}^{n} \epsilon \Pr[Z \geq i\epsilon] = \sum_{i=1}^{n} \epsilon\, F(i\epsilon) \in [I_1(0);\ I_1(\epsilon)] = \left[\frac{1-k\epsilon}{k+1};\ \frac{1+\epsilon}{k+1}\right]. \qquad (28)
\end{aligned}$$

For the last inequality, we evaluate and bound the corresponding integrals in a straightforward way. This proves part (a). For part (b), let $\alpha = 1-\delta$, and let $i_0 = \alpha/\epsilon + 1$. Then:

$$E(Z - Z_\delta) = \sum_{i=i_0}^{n} \epsilon\, \Pr[Z \geq i\epsilon] = \sum_{i=i_0}^{n} \epsilon\,(1 - F(i\epsilon)) = \delta - \sum_{i=i_0}^{n} \epsilon\, F(i\epsilon) < \delta - I_\alpha(\delta).$$

Now if $\delta(k+1) < \frac{2}{3}$ then we have

$$\begin{aligned}
\alpha^{k+1} &= (1-\delta)^{k+1} < 1 - (k+1)\delta + O(k^2\delta^2), \\
I_\alpha(\delta) &= \left(\frac{n}{n+1}\right)^k \frac{1-\alpha^{k+1}}{k+1} > \left(1 - \frac{k}{n+1}\right) \frac{(k+1)\delta - O(k^2\delta^2)}{k+1} = \delta - O(k\delta^2),
\end{aligned}$$

so $E(Z - Z_\delta) < O(k\delta^2)$. Otherwise we trivially have $E(Z - Z_\delta) < \delta < O(k\delta^2)$. This completes the proof for the special case.

Consider the general case. Let Z' and Z'_δ be the corresponding random variables for the special case. Then without changing the distribution of Z we can write $Z = a + (b-a)Z'$. Part (a) follows immediately from (28). Part (b) follows since

$$\begin{aligned}
Z_{\delta(b-a)} - a &= \min(Z - a,\ (b-a) - \delta(b-a)) = (b-a)Z'_\delta. \\
E[Z - Z_{\delta(b-a)}] &= E[Z-a] - E[Z_{\delta(b-a)} - a] \\
&= (b-a)\,E[Z] - (b-a)\,E[Z_\delta] \leq (b-a)\,k\delta^2/2. \quad \square
\end{aligned}$$

Claim A.3. *Consider the setting in Lemma 2.2. Then:*

(a) $\Pr[\mu^* = \mu_i = y] \leq w_i/(1+n_i)$ *for any* $i \in [k]$ *and* $y \in \mathbb{R}$.

(b) $E[b_{i^*} - a_{i^*}] \leq k\, d_{\mathsf{av}}$.

(c) $E\left[1_{\{\mu_i = \mu^*\}}\,(\mu^* - \rho_i)\right] \leq O(w_i\,\delta_i^2)/(b_i - a_i)$.

Proof. Let $S_i = [a_i;\ b_i]_{n_i}$ and $a^* = \max_l a_l$. If $y \notin [a^*; b_i]$ then the probability in part (a) is zero. Otherwise

$$\begin{aligned}
\Pr[\mu_l \leq y] &= \frac{1 + \lfloor (y - a_l)/\epsilon_l \rfloor}{1+n_l} \leq \frac{1 + (y-a_l)/\epsilon_l}{1+n_l} = \frac{y - a_l + \epsilon_l}{b_l - a_l + \epsilon_l} \\
\Pr[\mu^* = \mu_i = y] &= \Pr[\mu_i = y] \prod_{l \neq i} \Pr[\mu_l \leq y] \leq \frac{1}{1+n_i} \prod_{l > i} \left(\frac{b_i - a_l + \epsilon_l}{b_l - a_l + \epsilon_l}\right) = \frac{w_i}{1+n_i} \\
E[b_{i^*} - a_{i^*}] &= \sum_{i \in [k]} \sum_{y \in S_i} \Pr[\mu^* = y \text{ and } i = i^*]\,(b_i - a_i) \leq \sum_{i \in [k]} w_i (b_i - a_i) = k\, d_{\mathsf{av}}.
\end{aligned}$$

This proves parts (a) and (b). For part (c), denote $z_i = b_i - \delta_i$, and define the set

$$S_i = \{x \geq 0 :\ x + z_i \in [a_i; b_i]_{n_i}\}.$$

Then, letting $k_i = \lceil (z_i - a_i)/\epsilon_i \rceil$, we have

$$\begin{aligned} n_i - k_i &\leq n_i - \frac{z_i - a_i}{\epsilon_i} = \frac{(b_i - a_i) - (z_i - a_i)}{\epsilon_i} = \frac{b_i - z_i}{\epsilon_i} = \frac{\delta_i}{\epsilon_i}. \\ \sum_{x \in S_i} x &= \sum_{j=k_i}^{n_i} (a_i + j\epsilon_i - z_i) \leq \sum_{j=0}^{n_i - k_i} j\epsilon_i = O(\epsilon_i)(n_i - k_i)^2 = O(\delta^2/\epsilon_i). \end{aligned} \tag{29}$$

Now using part (a) and (29) we obtain

$$\begin{aligned} E\left[1_{\{\mu_i = \mu^*\}} (\mu^* - \rho_i)\right] &= \sum_{x \in S_i} x \Pr[\mu^* = \mu_i = z_i + x] \\ &\leq \frac{w_i}{n_i} \sum_{x \in \Omega_i} x \leq \frac{w_i}{n_i} \frac{O(\delta_i^2)}{\epsilon_i} = \frac{O(w_i\, \delta_i^2)}{b_i - a_i}. \quad \square \end{aligned}$$

B Proofs from Section 3 (the general case)

Claim B.1. *Consider a dynamic bandit $\mathcal{B}$ in the setting of (14). Then (18) gives a $\mathcal{B}$-padding.*

Proof. We need to prove (16). Let us prove it for a fixed arm $i \in [k]$ and time $t = T$. Let us restate the problem in the notation which is more convenient for this proof. For every time t, let $Y_t = X_t^{(i)}$ and let Y_t^* be the winnings from arm i at time t. Let $\{t_j\}_{j=1}^{\infty}$ be the enumeration of all times when arm i is played. Let $Z_j = Y_{t_j}$ and $Z_j^* = Y_{t_j}^*$. Let us define the sums $S = \sum_{j \in [m]} Z_j$ and $S^* = \sum_{j \in [m]} Z_j^*$, where $m = t_i$. Letting $\mu = \mu_i$ and $\delta = \delta_i(m)$, we can rewrite (16) as follows:

$$\Pr[|S^* - \mu\, m| > \sqrt{2m \ln T} + \delta m] < O(T^{-4}). \tag{30}$$

Let F be the *failure event* when $|Y_t - \mu| > \delta$ for some $t \in [T]$. Recall that by Claim 2.4(a) the probability of F is at most T^{-4}. In the probability space induced by conditioning on $Z_1^* \ldots Z_{j-1}^*$ and the event $\bar{F}$, we have

$$E[Z_j^*] = E\left[\, E[Z_j^* \,|t_j, Z_j]\,\right] = E\left[\, E[Z_j \,|t_j]\,\right] = E[Z_j] \in [\mu - \delta, \mu + \delta].$$

Going back to the original probability space, we've just proved that

$$E[Z_j^* |\, Z_1^* \ldots Z_{j-1}^*, \bar{F}] \in [\mu - \delta, \mu + \delta]. \tag{31}$$

The Chernoff-Hoeffding bounds (applied to the probability space induced by conditioning on $\bar{F}$) say precisely that the condition (31) implies the following tail inequality:

$$\Pr\left[\, |S^* - \mu\, m| > \delta m + a \mid \bar{F}\,\right] \leq 2\, e^{-2a^2/m} \text{ for any } a \leq 0.$$

We obtain (30) by taking $a = \sqrt{2m \ln T}$. $\square$

Proof of Lemma 3.7, Eqn. (20): We will prove (20) for any given bandit algorithm. Fix arm i and let us introduce a convenient notation that gets rid of the subscript i. Specifically, let $\mu = \mu_i$, $\delta = \delta_i(m)$, and $J = T_i(m)$. For every time t, let $Y_t = X_t^{(i)}$, and let Y_t^* be the winnings from arm i at time t. Let ζ_t be equal to 1 if arm i is played at time t, and 0 otherwise. To prove (20), we will show that

$$E\left[\textstyle\sum_{t \in [m]} \zeta_t\, Y_t^*\right] = E\left[\textstyle\sum_{t \in [m]} \zeta_t\, Y_t\right] \geq \min(\mu, b - \delta)\, E[J] + m^{-2}(b - a). \tag{32}$$

Note that ζ_t and Y_t^* are conditionally independent given Y_t. It follows that

$$E\left[\zeta_t\, Y_t^* \,|Y_t\right] = E\left[\zeta_t \,|Y_t\right]\ E\left[Y_t^* \,|Y_t\right] = E\left[\zeta_t \,|Y_t\right]\ Y_t = E\left[\zeta_t\, Y_t \,|Y_t\right].$$

Taking expectations on both sides, we obtain $E[\zeta_t Y_t^*] = E[\zeta_t Y_t]$, which proves the equality in (32).

Let $S = \sum_{t\in[m]} \zeta_t Y_t$. Let us consider the case $\mu \leq b - \delta$. We claim that

$$\text{if } \mu \leq b-\delta \text{ then } E[S] \geq \mu\, E[J] - m^{-2}\,(b-a). \tag{33}$$

Couple $\{Y_t\}_{t=0}^{\infty}$ with the corresponding *unbounded* random walk with step $\epsilon = \frac{b-a}{n}$, call it $\{\widehat{Y}_t\}_{t=0}^{\infty}$. Define the corresponding shorthand $\widehat{S} = \sum_{t\in[m]} \zeta_t \widehat{Y}_t$. Let F be the *failure event* when $\widehat{Y}_t = b$ for some $t \leq m$. Note that if this event does not occur, then $Y_t \geq \widehat{Y}_t$ for every time $t \in [m]$ and therefore $S \geq \widehat{S}$. We use this observation to express $E[S]$ in terms of $E[\widehat{S}]$. Let $p := \Pr[F]$ and note that it is at most m^{-4} by Azuma's inequality. Then:

$$\begin{aligned}
E[\widehat{S}] &= (1-p)\, E[\widehat{S} \mid \text{not } F] + p\, E[\widehat{S}\,|F] \\
&\leq (1-p)\, E[\widehat{S} \mid \text{not } F] + p(\mu + m\epsilon) \\
E[S] &\geq (1-p)\, E[S \mid \text{not } F] + p\, E[S\,|F] \\
&\geq (1-p)\, E[\widehat{S} \mid \text{not } F] + pa \\
&\geq E[\widehat{S}] - pm\epsilon - p(b-a).
\end{aligned}$$

To prove (33), it remains to bound $E[\widehat{S}]$.

Let $\{t_j\}_{j=1}^{\infty}$ be the enumeration of all times when arm i is played. Note that $J = \max\{j : t_j \leq m\}$. Define $\widehat{Z}_j = \widehat{Y}_{t_j}$ for each j. We would like to argue that $\{\widehat{Z}_j\}_{j=1}^{\infty}$ is a martingale and J 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 t_{j+1} and all tuples $(t_l, Z_l, Z_l^*, \widehat{Z}_l)$ such that $l \leq j$. Now using the Optional Stopping Theorem one can show that

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

which proves (33) since $\widehat{Z}_0 = \mu$.

Now consider the case $\mu > b - \delta$. Specifically, we claim that

$$\text{if } \mu > b-\delta \text{ then } E[S] \geq (b-\delta)\, E[J] - m^{-2}\,(b-a). \tag{34}$$

Let T be the smallest time t such that $Y_t = b - \delta$, and let N be the largest index j such that $t_j \leq T$. Conditioning on T and N, consider the entire problem starting from time $T+1$. Then by (33) we have:

$$E\left[\textstyle\sum_{t=T+1}^{m} \zeta_t Y_t \,|T, N\right] \geq (b-\delta)(E[J] - N) - m^{-2}\,(b-a).$$

Let $S_T = \sum_{t=T+1}^{m} \zeta_t Y_t$. It follows that

$$\begin{aligned}
S &= S_T + \textstyle\sum_{t\in[T]} \zeta_t Y_t \geq S_T + (b-\delta)\, N \\
E[S] &= E[S_T] + (b-\delta)E[N] \\
&\geq (b-\delta)E[J-N] - m^{-2}\,(b-a) + (b-\delta)E[N] \\
&\geq (b-\delta)E[J] - m^{-2}\,(b-a).
\end{aligned}$$

This proves (34) and completes the proof of (32). □

Proof of Lemma 3.7, Eqn. (22): Denote $\Delta_i = \mu^* - \mu_i$ for each arm i. We need to integrate the right-hand side of (21) with respect to $\mu_1 \ldots \mu_k$. To this end, we will condition on μ^*. From here on we will work in the probability space induced by this conditioning. We partition the arms into three sets and bound the three corresponding sums separately. One of these three sets is S. The other two, denoted S^* and S^+, are the sets of all arms $i \in [k]$ such that, respectively, $\Delta_i = 0$ and $0 < \Delta_i < 4\delta_i$.

Firstly, for any $\gamma < \mu^*$ let us define a discrete interval $\mathcal{Y}_i(\gamma) = [a;b]_{n_i} \cap [a; \mu^* - \gamma)$ and observe that given the event $\{\Delta_i > \gamma\} = \{\mu_i < \mu^* - \gamma\}$, the random variable μ_i is distributed uniformly on $\mathcal{Y}_i(\gamma)$. We claim that

$$E[\Delta_i^{-1} | \Delta_i > \epsilon_i] \leq O(\log n_i)\,(\mu^* - a)^{-1}. \tag{35}$$

Indeed, let $\{y_j\}_{j=1}^n$ be a decreasing enumeration of $\mathcal{Y}_i(\epsilon_i)$. Then

$$n\,E[\Delta^{-1} | \Delta_i > \epsilon_i] = \sum_{y \in \mathcal{Y}_i(\epsilon_i)} \frac{1}{\mu^* - y} = \sum_{j \in [n]} \frac{1}{\mu^* - y_j} = \sum_{j \in [n]} \frac{1}{j\epsilon_i} = \frac{O(n_i \log n)}{b-a},$$

which implies (35) since $\frac{n}{n_i} = O(\frac{\mu^* - a}{b-a})$ by definition of n. Noting that $n_i \leq \max n_i \leq k\,n_{\text{av}}$, we have

$$\begin{aligned} E\left[\sum_{i \in S} \Delta_i^{-1}\right] &\leq E\left[\sum_{i:\,\Delta_i > \epsilon_i} \Delta_i^{-1}\right] \leq \sum_{i \in [k]} E\left[\Delta_i^{-1} |\ \Delta_i > \epsilon_i\right] \leq O(k \log n_{\text{av}})\,(\mu^* - a)^{-1} \\ &= \alpha(\mu^*) \times O(k \log n_{\text{av}})\,(b-a)^{-1}, \end{aligned} \tag{36}$$

where $\alpha(\mu^*) = (b-a)/(\mu^* - a)$.

Secondly, recall that given $\{\mu_i < \mu^*\}$, the random variable μ_i is distributed uniformly $\mathcal{Y}_i(0)$. Therefore

$$\Pr[i \in S^+ |\, \mu_i < \mu^*] = \Pr[\mu_i > \mu^* - 4\delta_i |\, \mu_i < \mu^*] = O(\delta_i)/(\mu^* - a).$$

Since for any $i \in S^+$ we have $\mu^* - \rho_i \leq O(\delta_i)$, it follows that

$$\begin{aligned} E\left[\sum_{i \in S^+} \mu^* - \rho_i\right] &\leq \sum_{i \in [k]} O(\delta_i)\ \Pr[i \notin S |\, \mu_i < \mu^*] \leq \sum_{i \in [k]} \frac{O(\delta_i^2)}{\mu^* - a} \\ &= \alpha(\mu^*) \times O(k\,n_{\text{av}}^{-2}\,m \log m)(b-a)^{-1}. \end{aligned} \tag{37}$$

Thirdly, in the proof of Lemma 2.3(b) we have actually shown a slightly stronger statement (12) which in the present setting can we rewritten as an unconditional expectation of the sum over S^*:

$$E\left[\sum_{i \in S^*} \mu^* - \rho_i\right] \leq O(b-a)(k\,n_{\text{av}}^{-2})(m \log m). \tag{38}$$

To obtain (22) we observe that right-hand side of (21) is the sum of the left-hand sides of (36), (37) and (38). Finally, it is straightforward to upper-bound the expectation of $\alpha(\mu^*)$ by $O(1+\gamma)$. □

C Proofs from Section 4 (adaptive exploration)

In this section we prove Lemma 4.4 and Eqns. (24-27). We restate the latter as the following lemma:

Lemma C.1. *there is $\gamma = \Theta(k\epsilon)$ such that for any time t we have*

$$E[\,c_i(t) \mid Z_i(t) = \delta \text{ and } X_i(t) < X^*(t)] = \begin{cases} O(\epsilon^2/\delta)(\log n) & \text{if } \delta \geq \gamma, \\ O(k\epsilon) & \text{otherwise.} \end{cases} \tag{39}$$

$$\Pr[\,Z_i(t) < \delta \mid X_i(t) < X^*(t)] \leq O(\delta + k\epsilon) \tag{40}$$

$$E[\,c_i(t) \mid X_i(t) = X^*(t)] \leq O(k\epsilon)^2 \tag{41}$$

$$E[c_i(t)] \leq O(\epsilon^2)(k^2 + \log n) \tag{42}$$

Firstly, we need to argue that Y^* cannot decrease too fast:

Claim C.2. $Y^*(t+1) \geq Y^*(t) - \epsilon$ *for any time* t.

Proof. This is trivially true when $L(t) = L(t+1)$, i.e. there is no change of leader after time step t. Now suppose there is. Let $L = L(t)$ be the current leader at time t, and let $i \neq L$ be the arm played at time t. Recall that the change of leader after time t happens if and only if $X_i(t) > X_L(t-1)$. Therefore,

$$Y^*(t) = X_L(t) \leq X_L(t-1) + \epsilon \leq X_i(t) \leq X_i(t+1) + \epsilon = Y^*(t+1) + \epsilon. \quad \square$$

Secondly, we establish a very intuitive and desirable property that any maximal arm is suspicious.

Claim C.3. *Fix time* t *and let* $\mathcal{B}$ *be a dynamic* k*-armed bandit on* $[0;1]_n$ *which is well-behaved on a time interval* $[t-n^2;\, t]$*. Let* i *be an arm such that* $X_i(t) = X^*(t)$*. Then in any run of Algorithm 4.2 on* $\mathcal{B}$ *this arm is suspicious at time* t*.*

Proof. For simplicity assume that t is odd (if t is even, the proof is similar). Let τ be the last time arm i was played before time t. If $t - \tau > n^2$ then the claim is trivially true. Assume $t - \tau \leq n^2$. The current leader $L(t)$ was played at the even time $t-1$, so its reward satisfies

$$y^*(t) = X_{L(t)}(t-1) \leq X^*(t-1) \leq X_i(t) + \epsilon.$$

Then since X_i is well-behaved, we have

$$y^*(t) - X_i(\tau) \leq X_i(t) + \epsilon - X_i(\tau) \leq \epsilon + \alpha_{\mathcal{A}}\, \epsilon\sqrt{t-\tau} \leq c_{\mathcal{A}}\, \epsilon\sqrt{t-\tau}. \quad \square$$

Now we are ready to prove Lemma 4.4.

Proof of Lemma 4.4: Let $\epsilon = \frac{1}{n}$ and fix time $t \in [t_1; t_2]$. Let $t_0 = t - 2k$ and consider an arm i such that $X_i(t_0) = X^*(t_0)$. Then by Claim C.3 this arm is suspicious at time t, so it is played at some time $t' \in [t_0; t]$. Then by definition of the algorithm either $X_i(t') \leq y^*(t')$, or for the next step the leader is reset to i. In both cases we have $Y^*(t'+1) \geq X_i(t') - O(\epsilon)$. Therefore due to Claim C.2 we have

$$\begin{aligned} Y^*(t) &\geq Y^*(t'+1) - O(k\epsilon) \geq X_i(t') - O(k\epsilon) \\ &\geq X_i(t_0) - O(k\epsilon) \geq X^*(t_0) - O(k\epsilon) \geq X^*(t) - O(k\epsilon). \quad \square \end{aligned}$$

The crux of the proof of Lemma C.1 lies in the following two claims. Once these claims are proved, the rest of the proof is a relatively straightforward (although quite tedious) exercise in conditional expectations.

Claim C.4 (implies (39)). *Let* $\mathcal{B}$ *be a dynamic* k*-armed bandit on* $[0;1]_n$ *which is well-behaved around a given time* t*. Consider a run of Algorithm 4.2 on* $\mathcal{B}$*. Fix arm* i*, let* $\delta = Z_i(t)$ *and* $\tau = \tau_i(t)$ *and assume that* $\tau > t - n^2$*. Let* $\epsilon = \frac{1}{n}$*. Then:*

(a) if $\delta \geq \Omega(k\epsilon)$ *then* $\sigma_i(t) \geq \Omega(\delta/\epsilon)^2\, \log^{-2} n$*.*

(b) If arm i *is not active at any time* $t' \in (\tau; t]$ *then* $Z_i(\tau) \leq O(k\epsilon + Z_i(t'))$*.*

(c) $Z_i(\tau) \leq 2\delta + O(k\epsilon)$*.*

Proof. Let $\epsilon = \frac{1}{n}$. For convenience we will omit subscript i for $X_i(\cdot)$, $Z_i(\cdot)$ and $\sigma_i(\cdot)$.

Since the bandit is well-behaved around time t, for any $t' \in [t-n^2;\, t+n^2]$ we have

$$\begin{cases} |X(t') - X(\tau)| &\leq c_f\, \epsilon\, |t'-\tau|^{1/2} \\ |Y^*(t') - Y^*(\tau)| &\leq c_f\, \epsilon\, |t'-\tau|^{1/2} + O(k\epsilon), \end{cases} \tag{43}$$

where $c_f = \Theta(\log n)$ is from (23). Adding up the two inequalities in (43), we obtain

$$|Z(t') - Z(\tau)| \leq 2\, c_f\, \epsilon\, |t' - \tau|^{1/2} + O(k\epsilon). \tag{44}$$

For part (a) we consider two cases. Firstly, if $Z(\tau) < \delta/2$ then by (44) we have

$$2c_f\, \epsilon\sqrt{t-\tau} \geq |Z(t) - Z(\tau)| - O(k\epsilon) \geq \delta/4$$

and therefore $\sigma(t) \geq t - \tau \geq \Omega(\delta/\epsilon)^2$. Secondly, if $Z(\tau) \geq \delta/2$ then for any time $t' \in (\tau;\, t + n^2)$ we have

$$Y^*(t') - X(\tau) = Y^*(t') - Y^*(\tau) + Z(\tau) \geq \delta/2 - c_f\, \epsilon\sqrt{t' - \tau} - O(k\epsilon).$$

This is at least $c_{\mathcal{A}}\, \epsilon\sqrt{t' - \tau}$ as long as $t' - t \leq O(\delta n/c_f)^2$ Therefore arm i is not suspicious at any such time t'. It follows that $\sigma(t) \geq \Omega(\delta n/c_f)^2$. This completes part (a).

For part (b), note that since arm i is not suspicious at time t', we have

$$c_{\mathcal{A}}\, \epsilon\sqrt{t' - \tau} \leq Y^*(t') - X(\tau) = Z(t') + X(t') - X(\tau) \leq Z(t') + c_f\, \epsilon\sqrt{t' - \tau}.$$

Since $c_{\mathcal{A}} = 5c_f$ and using (44), it follows that

$$|Z(t') - Z(\tau)| \leq 2c_f\, \epsilon\, |t' - \tau|^{1/2} + O(k\epsilon) \leq \tfrac{1}{2} Z(t') + O(k\epsilon), \tag{45}$$

proving part (b).

For part (c), note that if $t - \tau \leq 4k$ then the bound is trivial. Assume that $t - \tau > 4k$. Then in the interval $(\tau, t]$ there exists a time when arm i is not suspicious. Let t' be the largest such time. By part (b) we have $Z(\tau) \leq Z(t') + O(k\epsilon)$. Since arm i is active during the time interval $[t' + 1, t]$ but has not been played before time t, it follows that $t - t' \leq 2k$. Therefore $t' - \tau \geq 2k \geq t - t'$. Now using (44) and (45) we have

$$Z(t') - Z(t) \leq 2c_f\, \epsilon\, \sqrt{t - t'} + O(k\epsilon) \leq 2c_f\, \epsilon\, \sqrt{t' - \tau} + O(k\epsilon) \leq \tfrac{1}{2} Z(t') + O(k\epsilon),$$

so $Z(t') \leq 2\delta + O(k\epsilon)$ and part (c) follows. □

Claim C.5 (implies (40)). *Let $\mathcal{B}$ be a dynamic k-armed bandit on $[0;1]_n$. Consider a run of Algorithm 4.2 on $\mathcal{B}$. Let i be an arm such that $X_i(t) = X^*(t)$. Then either $c_i(t) = 0$ or we have $c_i(t) \leq 2\epsilon$ and $i \neq L(t)$.*

Proof. Let τ be the last time arm i was played before time t. If $i = L(\tau)$ then $c_i(t) = 0$ by definition. If $i = L(t) \neq L(\tau)$ then by the specification of the algorithm there was a change of leader after time τ, which implies $c_i(t) < 0$. (By the way, it also follows that $\tau = t - 1$ and t is even.)

Finally, assume that $i \neq L(t)$. Recall that $c_i(t) \leq Z_i(\tau)/(t - \tau)$ where $Z_i(\tau) := Y^*(\tau) - X_i(\tau)$. By Claim C.2 on every time step the function Z_i decreases by at most 2ϵ. Therefore, $Z_i(t) \geq Z_i(\tau) - 2\epsilon(t - \tau)$. By the choice of arm i we have $Z_i(t) \leq 0$, which implies $c(t) \leq 2\epsilon$. □

Now we are ready for the final push.

Proof of Lemma C.1: Let $\tau = \tau_i(t)$ and recall that $c_i(t) = Z_i(\tau)/\sigma_i(t)$.

Let E_t be the event that the bandit is well-behaved around time t. Suppose this event holds. If $\tau \leq t - n^2$ then $\sigma(t) \geq n^2$ and so $c(t) \leq \epsilon^2$. If $\tau > t - n^2$ then by Claim C.4(ac) $c(t)$ is equal to the right-hand side of (39) for some $\gamma = \Theta(k\epsilon)$.

Proof of (39). Let Γ be the event that $Z(t) = \delta$ and $X_i(t) < X^*(t)$. We just proved that $E[c_i(t) \mid \Gamma \cap E_t]$ equals the right-hand side of (39). Since $c_i(t) \leq 1$ and $\Pr[\Gamma] \geq \Omega(\epsilon)$, we have

$$\begin{aligned} E[c(t) \mid \Gamma] &\leq E[c(t) \mid \Gamma \cap E_t] + \Pr[\bar{E}_t \mid \Gamma], \\ \Pr[\bar{E}_t \mid \Gamma] &\leq \Pr[\bar{E}_t] / \Pr[\Gamma] \leq \epsilon^2. \end{aligned}$$

Proof of (40). Note that for any $\delta, x^* \in [0;1]_n$ we have:

$$\begin{aligned}
p(\delta, x^*) &:= \Pr[Z(t) \leq \delta \mid X(t) < X^*(t) = x^*] \\
&\leq \Pr[x^* - X(t) \leq \delta + O(k\epsilon) \mid X(t) < x^*] \\
&\leq O(\delta + k\epsilon)/x^*. \\
\Pr[Z(t) \leq \delta \mid X(t) < X^*(t)] &= \sum_{x^* \in [0;1]_n} \Pr[X^*(t) = x^*] \;\; p(\delta, x^*) \\
&\leq O(\delta + k\epsilon) \; E[1/X^*(t)] \\
&\leq O(\delta + k\epsilon)
\end{aligned}$$

Proof of (41). For a more efficient notation, let us work in the probability space induced by conditioning on the event $\{X_i(t) = X^*(t)\}$. Let F be the event that there exists some arm $j \neq i$ such that $X^*(t) - X_j(t) < O(k\epsilon)$. In particular, by Lemma 4.4 this event happens (for $j = L(t)$) if the bandit is well-behaved around time t and $i \neq L(t)$. Then

$$\begin{aligned}
\Pr[i \neq L(t) \mid E_t] &\leq \Pr[F|E_t] \leq O(\Pr[F]) \leq O(k^2\epsilon) \\
\Pr[i \neq L(t)] &\leq \Pr[i \neq L(t) \mid E_t] + \Pr[\bar{E}_t] \leq O(k^2\epsilon).
\end{aligned}$$

Restating Claim C.5, we have $c(t) \leq 2\epsilon$ and $E[c(t)|i = L(t)] = 0$. Therefore,

$$E[c(t)] \leq \; E[c(t)|i \neq L(t)] \; \Pr[i \neq L(t)] \; + \; E[c(t)|i = L(t)] \leq O(k\epsilon)^2.$$

Proof of (42). Let $f(\delta)$ and $p(\delta)$ be the left-hand sides of (39) and (40), respectively. Then

$$\begin{aligned}
f(\delta) &:= E[\, c(t) \mid Z(t) = \delta \text{ and } X(t) < X^*(t)] \\
p(\delta) &:= \Pr[Z(t) \leq \delta \mid X(t) < X^*(t)] \leq O(\delta + k\epsilon) \\
p(\delta)\, f(\delta) &\leq \begin{cases} O(k\epsilon)^2 & \text{if } \delta \leq \gamma \\ O(\epsilon^2 \log n) & \text{otherwise.} \end{cases}
\end{aligned}$$

for any $\delta \in [0;1]_n$. Now it is straightforward to derive (42):

$$\begin{aligned}
E[c(t) \mid X(t) < X^*(t)] &= \sum_{\delta \in [0;1]_n} f(\delta) \; \Pr[Z(t) = \delta \mid X(t) < X^*(t)] \\
&\leq f(\gamma)\, p(\gamma) + \sum_{j \in [\log n]} f(\gamma 2^j)\, p(\gamma 2^j) \\
&\leq O(\epsilon^2)(k^2 + \log n) \\
E[c(t)] &\leq E[c(t) \mid X(t) = X^*(t)] \; \Pr[X(t) = X^*(t)] + E[c(t) \mid X(t) < X^*(t)] \\
&\leq O(\epsilon^2)(k^2 + \log n). \quad \square
\end{aligned}$$